
A Guide to SPSS for Information Science

SPSS version: 19.0

Department of Information Science
Loughborough University

Contact: Professor Anne Morris
a.morris@Lboro.ac.uk



Acknowledgements

I would like to thank the Subject Centre for Information and Computer Sciences (Higher Education Academy) for funding the development of this *SPSS* Guide and supporting tutorials.

I would also like to thank David Green, Mathematics Education Centre, Loughborough University, for the many hours of labour, much of it unpaid, he spent on the project. Without his support, dedication and enthusiasm the outcome would have been very different.

In addition, I would like to acknowledge the use of the datasets listed below.

Professor Anne Morris
Department of Information Science
Loughborough University
1st March 2012

A Guide to SPSS for Information Science

This Guide is freely available for use (with acknowledgement) in non-commercial UK organisations. Comments, suggestions and corrections are welcome.

Acknowledgement of Data sources

Full details of sources are found on the last pages of the Appendix to this Guide.

Population, Gross National Income (GNI) and Gross Domestic Product (GDP) for Countries

The World Bank
The International Monetary Fund

Land area of Countries

Internet World Stats

World Countries and Territories by Region

United Nations Statistics Division

Social Class Population Profiles for UK

businessballs.co.uk

100 top-selling books of all time (Nielsen, 1998-2010)

guardian.co.uk

RCUK National Survey

LISU (Loughborough University) and SQWconsulting.
Project funded by Research Councils UK.

Facebook Users

Internet World Stats

Internet Users and Usage

Internet World Stats
Europa

IT Piracy Rates and Values

Business Software Alliance

World Adult Literacy Rates

UNESCO Institute of Statistics

Contents

PART 1 – REFERENCE SECTIONS

Section	Page
1 Introduction	1
1.1 What are SPSS and PASW?	1
1.2 Purpose of this Guide	1
1.3 Prerequisites for using this Guide	1
1.4 How to use this Guide	1
2 The main SPSS windows, organisation and terminology	2
2.1 The main SPSS windows	2
2.2 The Data Editor window – overview	2
2.3 The Data Editor window – Data View	4
2.3.1 Cells	4
2.3.2 Cases	4
2.3.3 Variables	4
2.4 The Data Editor window – Variable View	4
2.4.1 Variables	4
2.4.2 Variable labels	5
2.4.3 Variable types and Data types	5
2.4.4 Values	6
2.4.5 Value labels	6
2.4.6 Missing and invalid data	6
2.4.7 Viewing coded data with value labels – two choices	7
2.5 Edit Options	8
2.5.1 Viewing lists of variables which have labels	8
2.5.2 Format options for new numeric variables	9
2.5.3 Format options for currency	9
2.5.4 Display options for output labels in pivot tables	9

2.5.5	Display options for output labels in outlines	9
2.6	The Viewer window – displaying and handling output	10
2.6.1	The Viewer window	10
2.6.2	Exporting output to Microsoft Word	10
2.6.3	Saving output in a standard SPSS output file	11
2.6.4	Printing output	11
2.6.5	Exiting SPSS	11
3	Getting on-screen help	12
3.1	Toolbar	12
3.2	Help menu	12
3.2.1	Topics	12
3.2.2	Tutorial	13
3.2.3	Case Studies	14
3.2.4	Statistics Coach	14
3.2.5	Dialog Box help	14
4	Creating variables and entering data	15
4.1	Variables	15
4.2	General instructions for creating variables	15
5	Loading a data file and Editing data in a data file	18
5.1	Loading a data file	18
5.2	Changing data in a cell	19
5.3	Copying or moving data in a row, column or block of cells	19
5.4	Inserting a new case (row)	19
5.5	Deleting a case (row)	20
5.6	Inserting a new variable	20
5.7	Duplicating data in a variable	20
5.8	Deleting a variable	20
6	Saving a data file	21

7	Importing and Exporting data in Microsoft Excel format	22
7.1	Importing data in Microsoft Excel format	22
7.2	Exporting data into Microsoft Excel format	24
8	Sorting cases and Selecting cases	25
8.1	Sorting cases	25
8.2	Selecting cases	26
9	Computing variables and Recoding variables	28
9.1	Computing variables	28
9.2	Recoding variables	29
10	Introduction to Charts and Graphs	33
10.1	General points about Charts and Graphs in SPSS	33
10.2	Chart Editor Menus and Toolbars	33
10.3	FORMAT Toolbar	34
10.4	OPTIONS Toolbar	35
10.5	EDIT Toolbar	35
10.6	ELEMENTS Toolbar	35
11	Supplied SPSS data files for use with this Guide	36

PART 2 – TUTORIALS

TUTORIAL T1: Starting the SPSS program	37
TUTORIAL T2: Loading a saved SPSS data file	38
TUTORIAL T3: Analysing data using Frequencies	39
TUTORIAL T4: Creating a new data file – inputting data	43
T4.1 Defining the variables	43
T4.2 Moving around the Data Editor window	48
T4.3 Entering and Saving the data	49
T4.4 Analyzing the entered data	50
T4.5 Analyzing the complete data file	51
TUTORIAL T5: Checking Data – using Case Summaries	52
TUTORIAL T6: One-variable Frequency Tables	54
T6.1 One-variable Frequency Table for scale data – Mean	54
T6.2 One-variable Frequency Table for nominal data – Count and Total	56
T6.3 One-variable Frequency Table for ordinal data – Count with Subtotals	57
T6.4 One-variable Frequency Table for nominal data – Count with order sorted	59
TUTORIAL T7: Two-variable Frequency Tables	60
T7.1 Two-variable Two-way Frequency Table for scale and nominal data – Count, Max, Min, Median	60
T7.2 Two-variable Nested Frequency Table for nominal data – Count & Col%	62
T7.3 Two-variable Two-way Frequency Table for nominal data – Count & Col%	63
T7.4 Two-variable Two-way Frequency Table for nominal and ordinal data – interchanging rows and columns – Row%	64
T7.5 Two-variable Nested and Stacked Frequency Tables for nominal and scale data – Count	66
TUTORIAL T8: Three- and Four-variable Frequency Tables	68
T8.1 Three-variable Frequency Table for two nominal variables and one scale variable – Mean	68
T8.2 Three-variable Frequency Table for three nominal variables – Count	69
T8.3 Four-variable Frequency Table for three nominal variables and one scale variable – Median and Mode	70

TUTORIAL T9: Descriptive Statistics	71
TUTORIAL T10: Exploratory Data Analysis	72
TUTORIAL T11: Simple Bar Chart – basic building	73
TUTORIAL T12: Simple Bar Chart – basic editing	75
TUTORIAL T13: Simple Bar Chart – advanced	76
T13.1 Building a simple bar chart	76
T13.2 Editing possibilities	76
T13.3 Editing the simple bar chart	77
TUTORIAL T14: Clustered Bar Chart	83
T14.1 Clustered Bar Chart – two variables	83
T14.2 Clustered Bar Chart – three variables	88
TUTORIAL T15: Stacked Bar Chart	89
T15.1 Stacked Bar Chart – basics	89
T15.2 Stacked Bar Chart – advanced	90
TUTORIAL T16: Histogram	95
T16.1 Simple Histogram	95
T16.2 Stacked Histogram	98
TUTORIAL T17: Frequency Polygon and Population Pyramid	99
T17.1 Frequency Polygon	99
T17.2 Population Pyramid	99
TUTORIAL T18: Pie Chart	101
TUTORIAL T19: Line Chart	104
T19.1 Simple Line Chart	104
T19.2 Multiple Line Chart	105
TUTORIAL T20: Scatterplot	107
T20.1 Scatterplot – basics	107
T20.2 Scatterplot – further explorations for the intrepid	109

TUTORIAL T21: Boxplot	111
T21.1 Simple Boxplot – one variable	111
T21.2 Simple Boxplot – two variables	112
T21.3 Multiple Boxplot – two variables	113
TUTORIAL T22: Means	114
TUTORIAL T23: Correlation	115
T23.1 Pearson Correlation (parametric)	115
T23.2 Spearman Correlation (nonparametric)	117
TUTORIAL T24: Crosstabs and the Chi-square Test	118
T24.1 Crosstabs – introduction	118
T24.2 Crosstabs and the Chi-square Test option	118
T24.3 Notes on the Chi-square Test – Adjusted Residuals option	120
T24.4 Notes on the Chi-square Test – Significance level	120
T24.5 Notes on the Chi-square Test – Criteria for validity	121
T24.6 Crosstabs and the Chi-square Test – Recoding data	121
T24.7 Notes on Measures of Strength of Association	127
TUTORIAL T25: Chi-square Test for Frequency Table data	128
TUTORIAL T26: One-sample Chi-square Test – Goodness of Fit	132
T26.1 The One-sample Chi-square Test – introduction	132
T26.2 The One-sample Chi-square Test – expected category values equal	132
T26.3 The One-sample Chi-square Test – expected category values entered	135
TUTORIAL T27: The <i>t</i> Test	137
T27.1 The <i>t</i> Test formats and criteria for validity	137
T27.2 The Independent-Samples <i>t</i> Test	138
T27.3 The One-Sample <i>t</i> Test	140
T27.4 The Paired-Samples <i>t</i> Test – scale data	141
T27.5 The Paired-Samples <i>t</i> Test – ordinal data	142

TUTORIAL T28: Nonparametric alternatives to the <i>t</i> Test	145
T28.1 The Mann-Whitney Rank-sum Test – for independent samples	145
T28.2 The Wilcoxon Matched-pairs Signed-ranks Test – for paired samples	146
TUTORIAL T29: Analysis of Variance (ANOVA)	147
T29.1 Introduction to Analysis of Variance	147
T29.2 One-Way between-subjects ANOVA (independent measures) – Post Hoc	148
T29.3 One-Way between-subjects ANOVA (independent measures) – Contrasts	152
T29.4 One-Way between-subjects ANOVA (independent measures) – Kruskal-Wallis	155
T29.5 One-Way within-subjects ANOVA (repeated measures)	156
T29.6 One-Way within-subjects ANOVA (repeated measures) – Friedman	162
T29.7 Two-Way between-subjects ANOVA (independent measures)	163
T29.8 Two-Way within-subjects ANOVA (repeated measures)	166
TUTORIAL T30: Kolmogorov-Smirnov One-sample Test	170
T30.1 Kolmogorov-Smirnov One-sample Test – Normality test: Example 1	170
T30.2 Kolmogorov-Smirnov One-sample Test – Normality test: Example 2	172
TUTORIAL T31: Linear Regression	173
T31.1 Simple Linear Regression	173
T31.2 Multiple Linear Regression – using Entry method ‘Enter’	177
T31.3 Multiple Linear Regression – using Entry method ‘Stepwise’	179
TUTORIAL T32: Logistic Regression	182
T32.1 Logistic Regression – using Entry method ‘Forward LR’	182
T32.2 Logistic Regression – using Entry method ‘Enter’	189
TUTORIAL T33: Reliability Analysis	191
T33.1 Reliability Analysis – Introduction	191
T33.2 Cronbach’s Alpha method – Example 1	192
T33.3 Cronbach’s Alpha method – Example 2	195
TUTORIAL T34: Factor Analysis	199
T34.1 Factor Analysis – Example 1 – 10 variables	199
T34.2 Factor Analysis – Example 2 – 36 variables	204

APPENDIX

DATA SET 1 – 100 Top-selling Books 1998-2010	209
DATA SET 2 – VLE Questionnaire	210
VLE QUESTIONNAIRE	211
RCUK OPEN ACCESS SURVEY – Introduction	212
DATA SET 3 – RCUK Survey on Open Access – General	213
DATA SET 4 – RCUK Survey on Open Access – Institutions	214
DATA SET 5 – RCUK Survey on Open Access – Researchers	217
RCUK SURVEY – RESEARCHER QUESTIONNAIRE	220
RCUK OPEN ACCESS SURVEY – INSTITUTIONAL QUESTIONNAIRE	223
DATA SET 6 – School Maths Research Project (NCETM)	225
SCHOOL MATHEMATICS RESEARCH PROJECT (NCETM) – QUESTIONNAIRE	229
DATA SET 7 – IT Piracy Worldwide	233
DATA SET 8 – Facebook Users Worldwide	234
DATA SET 9 – Internet Users in Europe	235
DATA SET 10 – Demographics Worldwide	237
DATA SET 11 – Internet Users Worldwide	238

PART 1 – REFERENCE SECTIONS

1 Introduction

1.1 What are *SPSS and PASW*?

SPSS is a program designed to be used for statistics data presentation and analysis. As such it is a powerful program which can manipulate and display data and perform a wide range of statistical operations. It has its origins as long ago as 1968 when the innovative software package *SPSS* (Statistical Package for the Social Sciences) was launched. *SPSS* continued under that name until 2010 when it was acquired by IBM. Initially the name became *PASW* (Predictive Analytics Software) but with copyright issues settled, the latest version is known as *IBM SPSS Statistics 19.0*. So *SPSS* lives on ...

This Guide was originally written for *PASW 18.0* and has been adapted for *IBM SPSS Statistics 19.0*. Some *PASW* screenshots have been retained.

To simplify matters we simply refer to the software as '*SPSS*'.

1.2 Purpose of this Guide

This Guide presents key aspects and terminology relating to *SPSS* and is primarily for students of Information Science. Although it aims to require no knowledge of statistical methods beyond that met in GCSE Mathematics, it is not a substitute for a statistics text.

The purpose of this Guide is to enable the reader to perform many of the fundamental operations of *SPSS*. The intention is that the reader will then be able to investigate further capabilities of *SPSS*, as required. Not every facility available, is detailed, or even mentioned, here. *SPSS* itself has an extensive built-in TUTORIAL system to aid further exploration.

1.3 Prerequisites for using this Guide

This Guide assumes that the reader has used the Microsoft Windows operating system and Microsoft Windows based programs (e.g. Word), and spreadsheet programs (e.g. *Excel*). *SPSS* is available for the Macintosh OS, but this Guide only deals with the Windows version.

In order to use this Guide you will need to understand how to carry out basic Windows operations using a mouse, to open menus and make menu selections, and re-size windows.

You will also need to understand some basic Windows terminology such as 'menu bars', 'toolbars', 'panes', 'windows' and 'drop down menus'. It would be helpful to know the basic concepts of a spreadsheet (although *SPSS* operates somewhat differently).

1.4 How to use this Guide

The Guide is split into two parts:

- Part 1: Sections 1 to 10 are primarily for reference.
- Part 2: TUTORIALS 1 to 30 are primarily for user activities
 - They consist of sets of numbered step-by-step instructions.
 - They have explanations and comments signified by a ► symbol.

Main windows, top level menu names and options are shown in Arial Black ...	like this
All titles, options, buttons, variable names and labels are shown in Arial bold ...	like this
Variable values and codes for them are show within single quote in Arial font ...	'like this'
Main text in this Guide is printed in Arial font ...	like this

Combinations of *SPSS* procedures are shown thus: **File → Open → Data**

The screenshots provided throughout, were obtained using *Snagit 10* software.

2 The main SPSS windows, organisation and terminology

2.1 The main SPSS windows

SPSS has a number of windows, and the main menus and buttons are accessible from all of them.

The two primary windows are:

- **Data Editor.** This enables you to insert, view or amend data, and to create or edit data files. It has two formats: Data View and Variable View.
- **Viewer.** This displays all statistical results, tables and charts, which can be edited and saved for later use. It opens automatically when you first ask the system to generate output.

Additional windows are:

- **Chart Editor.** This editor enables you to modify chart and plots. It is activated by double-clicking on a previously created chart.
- **Text Output Editor.** This enables you to edit text that is not displayed in pivot tables.
- **Pivot Table Editor.** This enables you to edit pivot tables, such as transposing rows and columns and showing/hiding parts of tables. [This is beyond the scope of this Guide. See: Online Help → Contents → Pivot Tables → Manipulating a Pivot Table.]
- **Syntax Editor.** This advanced feature enables you to create and edit command syntax. [This is beyond the scope of this Guide.]

2.2 The Data Editor window - overview

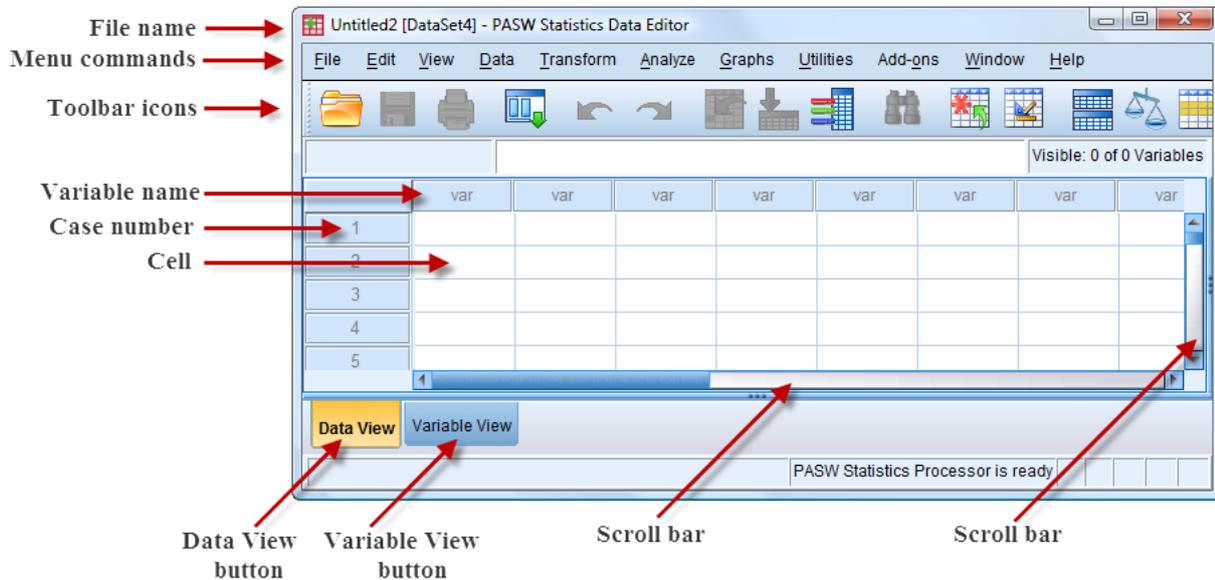
There are some distinctions between the way spreadsheets operate and the way SPSS data is organised and displayed in the **Data Editor** window.

The **Data Editor** window displays the content of the active SPSS file in either of two formats: **Data View** and **Variable View**. The window would typically have the title:

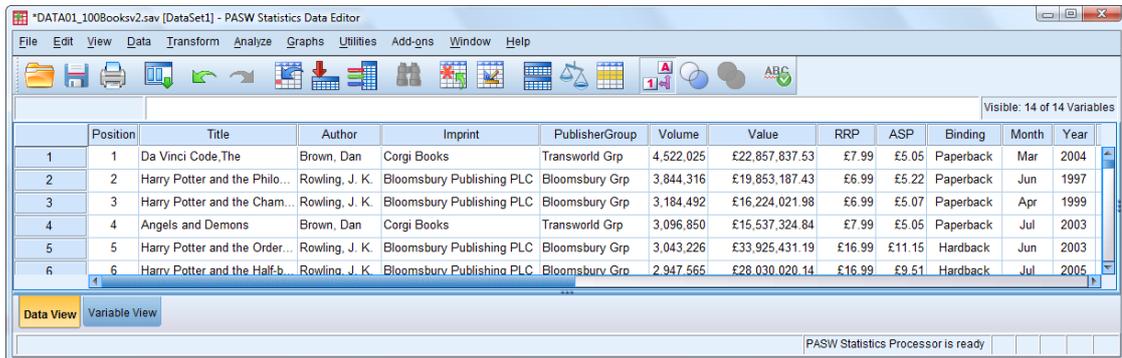
Filename.sav [DataSet1]- PASW Statistics Data Editor

signifying that the source of the SPSS data displayed is a file with extension **sav** called **Filename**.

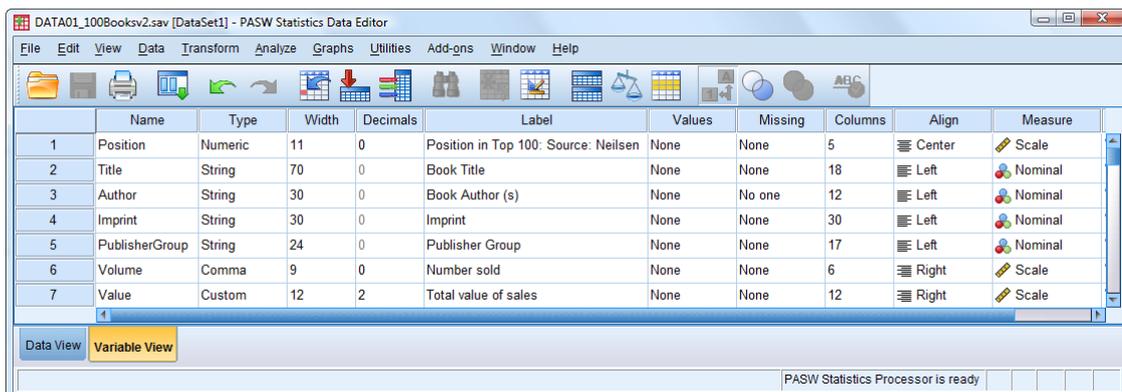
For reference, an annotated **Data Editor** window (in **Data View** mode) is shown below. All Menu commands and Toolbar icons are set out exactly the same in both **View** modes.



In **Data View** the data is displayed in a spreadsheet format of rows (representing cases e.g. the 100 best-selling books in GB) and columns (representing variables e.g. **position**, **title**, **author**, **imprint**, **publisher**, **volume_of_sales**, **value**, etc.).



In **Variable View** the data is displayed quite differently – each row represents one of the variables and each column contains information about an attribute of that variable or how it is to be displayed on screen (e.g. variable name, type (numeric or character string usually), width allowed for data entry, number of decimal places etc.).



2.3 The Data Editor window – Data View

2.3.1 Cells

In **Data View** data is displayed in **cells**, each item of data in a cell being known as a **value**. Each cell contains a single value of one variable for a particular case. Unlike a spreadsheet, cells in *SPSS* cannot contain formulas.

Bloomsbury Grp	3,844,316	£19,853,187.43
Bloomsbury Grp	3,184,492	£16,224,021.98
Transworld Grp	3,096,850	£15,537,324.84

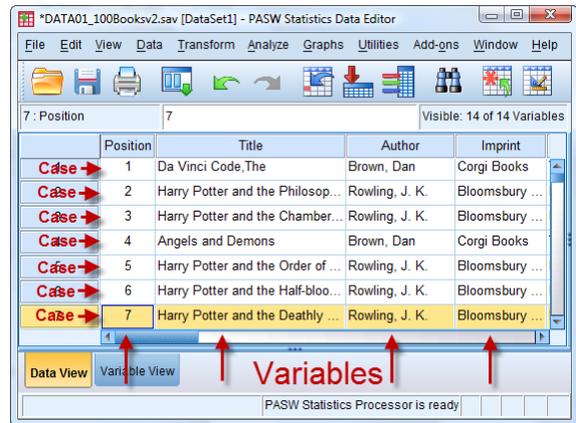
Value Cells Value

In **Data View** the data file is displayed as a rectangular array of cells whose dimensions are determined by the number of cases (rows) and the number of variables (columns).

2.3.2 Cases

In **Data View** rows are **cases**. Each row represents a different **case**. A case is a set of observations about one person, one country, one object, one experiment, etc.

For example, all the information for each individual completing a questionnaire is a case; information about each book in a library catalogue is a case; records concerning each student on a course make a case.



As in a spreadsheet, *SPSS* numbers each row but this is not tied to, or part of, the case. Often a unique ID number is provided for each case which is tied to the case (being a variable), as in the example here.

2.3.3 Variables

In **Data View** columns are **variables**. Each column represents a different **variable**. A **variable** is a measure of a characteristic or outcome that is being observed, measured or generated. A variable can take different values. For example, the response to each item on a multiple choice questionnaire would be a separate variable (which could take different values). A name must be provided for each variable (e.g. **book_title**).

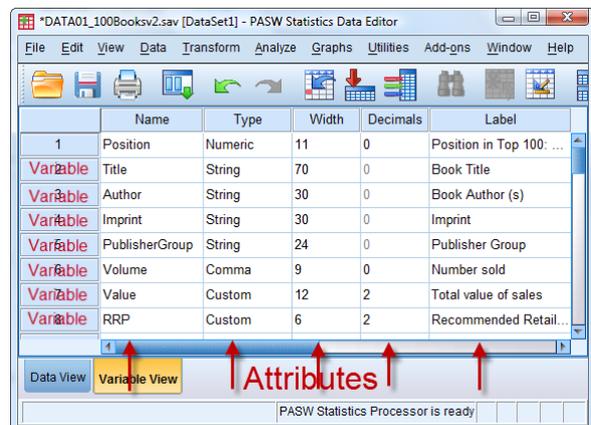
2.4 The Data Editor window – Variable View

2.4.1 Variables

Variables are usually created and their attributes defined in **Variable View**.

When creating a data file in the **Data Editor** it is normal to define the variables first (in **Variable View**) before entering the data (in **Data View**).

These processes are described in TUTORIAL T4.

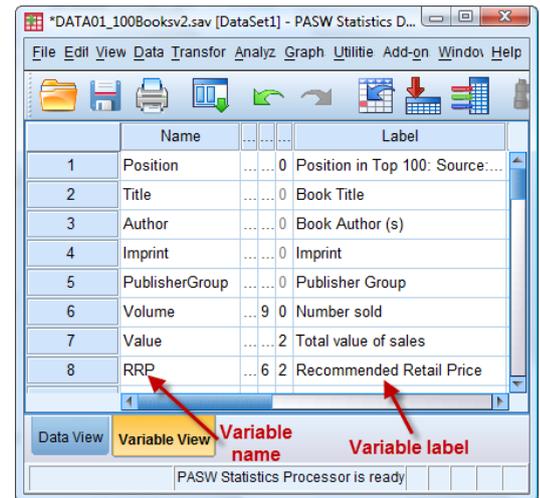


2.4.2 Variable labels

When creating a new variable, a unique (usually short) **variable name** must be provided, and, in addition, a **variable label** can be provided to give an explanation of the variable name i.e. what the variable is representing.

E.g. if the variable name was **age** then the label might be 'age of the respondent in years on 1 January 2012'; if the variable name was **bkdate** then the label might be 'date of publication of the first edition'.

The label is used to avoid confusion over exactly what a variable name might mean. It must not exceed 255 characters including spaces.



2.4.3 Variable types and Data types

In statistical textbooks variables and data (representing values of those variables) are often classified into four types:

1. **Nominal** Ex 1: variable colour might take values red, green, etc.
Ex 2: variable party might take values tory, labour, libdem, monster, etc.

These values are just 'names' or 'categories' with no specific way to order or measure them. The corresponding variables are classified as **string** variables as they are just character string. String variables can use digits as well as alphabetic characters (e.g. case numbers, bank account numbers).

Some statistics books call this type of data **categorical**, and SPSS uses the term **category** for the axis of a chart of a nominal variable.

2. **Ordinal** Ex 3: variable **age_group** might take values 'infant', 'child', 'youth'.
Ex 4: variable **age_group** might take values '0-5', '6-12', '13-18'.
Ex 5: variable **friendliness** might take value codes on a five point scale: '1' (very unfriendly), '2', '3', '4', '5' (very friendly).

These values are more than just 'names' or 'categories' because there is an obvious way to order them. However, they cannot really be measured mathematically.

In Ex 3 and Ex 4 you can meaningfully say 'infant' is less than 'child' and that '0-5' is less than '6-12' but you cannot say that one is (say) 3 more than the other, or one is half the other.

Even when there is a single number (a numerical code) to signify each value (as in Ex 5) it does not mean the rules of arithmetic apply. It is true that '4' signifies being friendlier than someone who is classified as '2' but that does not mean twice as friendly! Furthermore, it is not meaningful to say the difference in friendliness between '5' and '3' is the same as the difference between '3' and '1'.

N.B. There is a complication with the typical five point scale such as in Ex 5 if there is a further code or codes (e.g. '6' for 'Don't know'). Unless this extra code is treated as a missing value (see Section 2.4.6) and *excluded from most statistical analyses* you cannot really claim to have **ordinal** data, and it should be considered **nominal**.

3. Scale - Interval Ex 6: variable **temperature** (Celsius).

Here the difference between '95' and '96' is the same as the difference between '100' and 101'. However, it is not meaningful to say that what is measured as '100' is twice '50'. This is because there is a false origin (zero position) and temperatures can be below zero.

SPSS calls this type of data **scale**.

4. Scale - Ratio Ex 7: variable **distance** measured in millimetres. Ex 8: variable **time** measured in seconds. Ex 9: variable **monetary_wealth** measured in £.

In all these cases the rules of arithmetic do apply – '100' is twice '50', and the difference between '95' and '96' is the same as the difference between '100' and 101'.

SPSS calls this type of data **scale** too. So it does not differentiate between Interval and Ratio. This is because statistical procedures which apply to one of these will also apply to the other.

In summary, there are three SPSS variable types (and three corresponding data types) with associated icons:



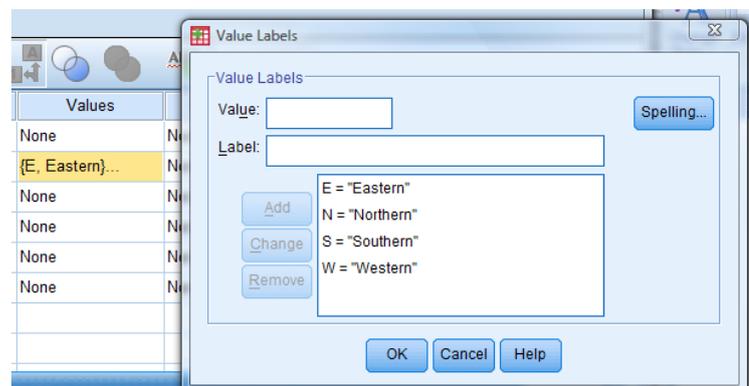
2.4.4 Values

Values for variables can be the actual data, e.g. if the variable is **age** then the values could be the actual ages of respondents in complete years. The values could instead be codes representative of the actual data. For example, for the variable **gender** the values could be '1' representing 'male' and '2' representing 'female'. Alternatively the full words could be entered, or abbreviations used such as 'M' and 'F'. It is normal to use numerical codes rather than strings as they are more amenable to statistical procedures.

2.4.5 Value labels

When creating a new variable, value labels can be used to explain what the different coded values of the variable represent. For example, if the age of a respondent has been coded into three age categories using the numbers '1', '2' and '3', then the three corresponding value labels might be: 'aged 0 – 12', 'aged 13 – 17' and 'aged 18 and over', or instead they could be 'children', 'youths', 'adults'.

In this screenshot the variable is **European_region** and the value codes used are 'E', 'W', 'N', 'S' which are defined in the **Value Labels** window.



2.4.6 Missing and invalid data

Missing data cannot be entered, of course, and the cell for the missing value can either be left blank or a special code (of one's choice) may be entered. For numeric variables, blank cells are automatically converted to the **system missing value** which SPSS represents by a full-stop. Users can define their own missing value codes: for example, the number '0' could be used to represent a *no-response* for the variable **gender** where '1' represents 'male' and '2' represents 'female'. The number '0' would be a **user-defined missing value**.

For string variables, a blank or series of blanks is considered a valid value and so is *not* interpreted as signifying missing data *unless explicitly declared as such*. There is, therefore, no **system missing value** for string variables.

Respondents to a questionnaire may answer a question but not provide a *valid* measure for the variable (e.g. claiming an age of 200). When this occurs the value could be entered but then explicitly declared a missing value or, more likely, an invalid-response code could be entered instead. For example one might use the number '999' to represent an invalid response for the variable **age**. The number '999' would be a **user-defined missing value**. A different code could be used to signify a *no-response*.

The number used for a missing value must, of course, be one that could not possibly occur for the variable in question. E.g. the code for signifying *no-response* to **number_of_children** could not be '0' or '9' but it could be '99'.

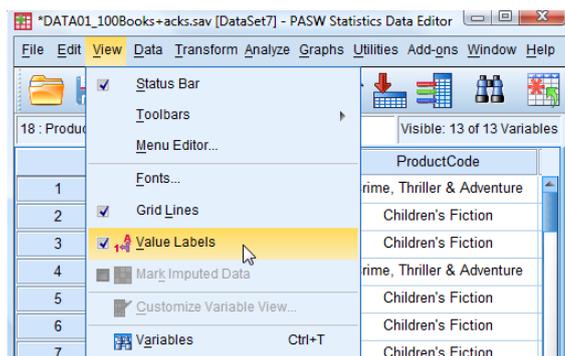
2.4.7 Viewing coded data with value labels – two choices

If data is coded and value labels have been defined then there are two ways to display the data – either show the codes themselves or show the labels which explain the codes. As an example, a data file containing details of the UKs 'all time' 100 best-selling Books (published by Nielsen, Dec 2010) has codes for month of publication ('1', ... '12') and for Product Code ('F1.1', ... 'Y2.2', ...). Here are the **Data View** screens for the top 15 books showing the codes (left) and showing the labels (right):

	Binding	Month	Year	ProductCode
1	Paperback	3	2004	F2.1
2	Paperback	6	1997	Y2.1
3	Paperback	4	1999	Y2.1
4	Paperback	7	2003	F2.1
5	Hardback	6	2003	Y2.1
6	Hardback	7	2005	Y2.1
7	Hardback	7	2007	Y2.1
8	Paperback	4	2000	Y2.1
9	Paperback	3	2007	Y2.2
10	Paperback	7	2001	Y2.1
11	Paperback	5	2004	F2.1
12	Paperback	9	2007	Y2.2
13	Paperback	12	2009	F1.1
14	Paperback	3	2004	F2.1
15	Paperback	4	2004	F1.1

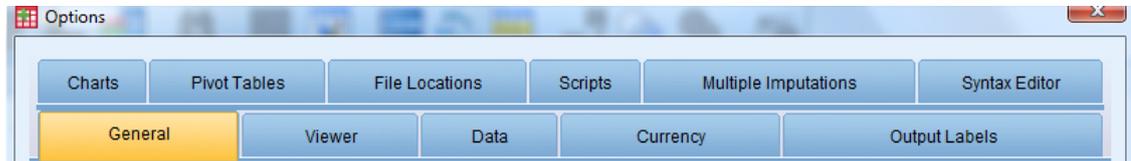
	Binding	Month	Year	ProductCode
1	Paperback	Mar	2004	Crime, Thriller & Adventure
2	Paperback	Jun	1997	Children's Fiction
3	Paperback	Apr	1999	Children's Fiction
4	Paperback	Jul	2003	Crime, Thriller & Adventure
5	Hardback	Jun	2003	Children's Fiction
6	Hardback	Jul	2005	Children's Fiction
7	Hardback	Jul	2007	Children's Fiction
8	Paperback	Apr	2000	Children's Fiction
9	Paperback	Mar	2007	Young Adult Fiction
10	Paperback	Jul	2001	Children's Fiction
11	Paperback	May	2004	Crime, Thriller & Adventure
12	Paperback	Sep	2007	Young Adult Fiction
13	Paperback	Dec	2009	General & Literary Fiction
14	Paperback	Mar	2004	Crime, Thriller & Adventure
15	Paperback	Apr	2004	General & Literary Fiction

To toggle between the two use **View → Value Labels**.



2.5 Edit Options

The command **Edit** → **Options** reveals a whole world of choices. Just a few of the most useful of these will be mentioned in this section. If you have a few spare months then you can investigate the rest. Below is the set of menu choices available. Bizarrely, when you select an option from the upper row the two rows swap round. Also, the layout is affected by how wide your window is. So your screen might look different!



2.5.1 Viewing lists of variables which have labels

In statistical procedures lists of variables for analysis are often provided. They can be displayed in alphabetical order or in the order they exist in the data file. They can be displayed by their variable name or by their variable label (if defined). If displayed by label then the name is included in square brackets at the end. SPSS remembers which format you have chosen – the choice is not associated with the dataset itself. If in this Guide a screenshot shows a list displayed in a format different from yours on screen that is probably the reason.

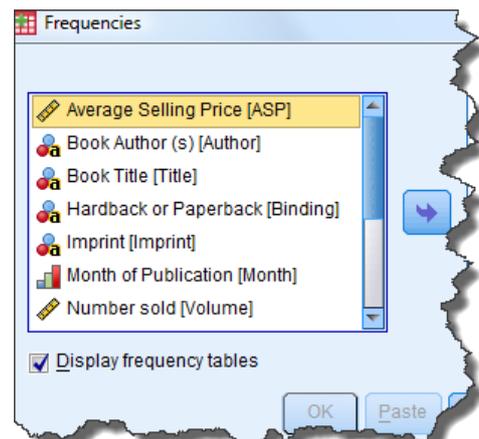
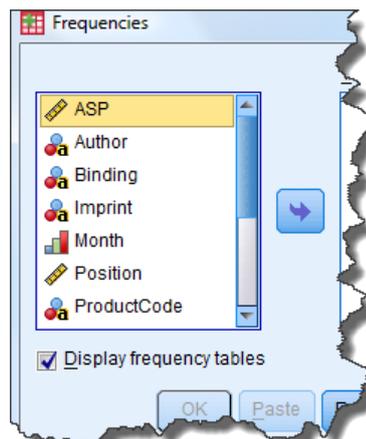
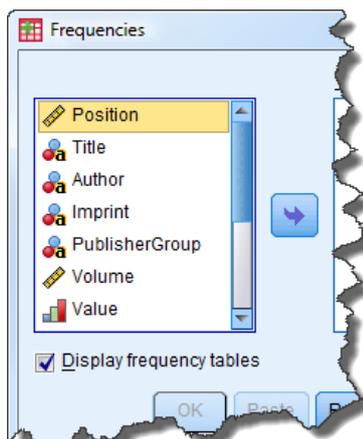
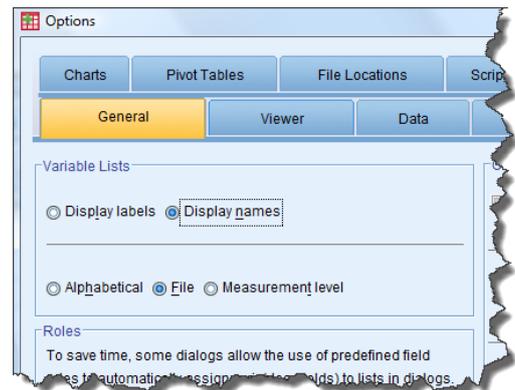
Alphabetical order is certainly better than file order if there are a lot of variables to search through. Names or labels will depend on whether the names are clear and distinct enough or not. Below are some screens showing what one gets. The example is for the **Frequencies** procedure which will be introduced in TUTORIAL T3.

To change the format use **Edit** → **Options** and select **General** (if not already selected). Then click the radio buttons you want (see the screen on the right which shows a possible selection).

Click **Apply** then **OK**.

Three possible formats are illustrated below:

- The left screen below has **Display names** in **File** order.
- The middle screen below has **Display names** in **Alphabetical** order.
- The right screen below has **Display labels** in **Alphabetical** order (note that the name is included in square brackets after the label).
- NOTE: You can right-click on a variable list and choose the display format you want without needing to use **Edit** → **Options**.



2.5.2 Format options for new numeric variables

SPSS assumes that whenever you create a variable it will be numeric with 2 decimal places (d.p.). If this annoys you, the d.p. level can be set to some other value – most likely zero.

To change the d.p. level use **Edit** → **Options** and select **Data**. Then click the up/down arrows to achieve whatever d.p.level you want for all future new variables you create. Don't forget to click **Apply** before finishing with **OK**.

2.5.3 Format options for currency

SPSS assumes that in output any currency will be US\$, but SPSS does provide do-it-yourself Custom Currency Formats (CCA to CCE) which can be used to produce £Sterling and other currency formats.

To do so use **Edit** → **Options** and select **Currency**. Then click on CCA (or another) and enter the required prefix (e.g. £) and choose period (i.e. full-stop) as the decimal separator. Don't forget to click **Apply** before clicking **OK**.

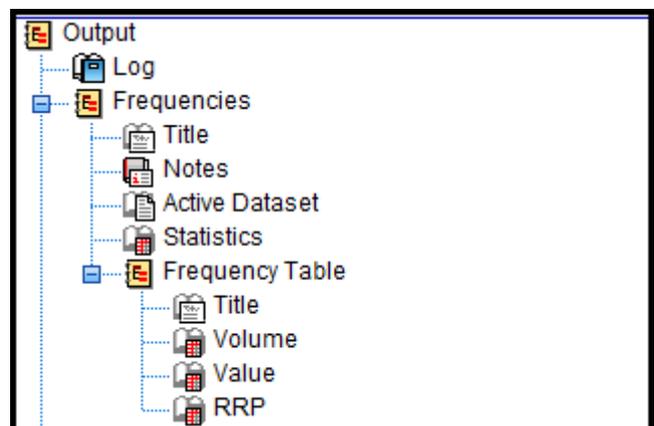
2.5.4 Display options for output labels in pivot tables

SPSS assumes that in output tables (called pivot tables) both variables and variable values will be shown by their labels (if defined). There are other choices. Variables can be shown by their names or by both names and labels. Variable values can be shown by their names or by both values and labels.

To make changes use **Edit** → **Options** and select **Output Labels**. Then for **Pivot Table Labelling** use the drop-down menus to select what you want for **Variables in labels shown as** and for **Variable values in labels shown as**. Then click **Apply** and finish with **OK**.

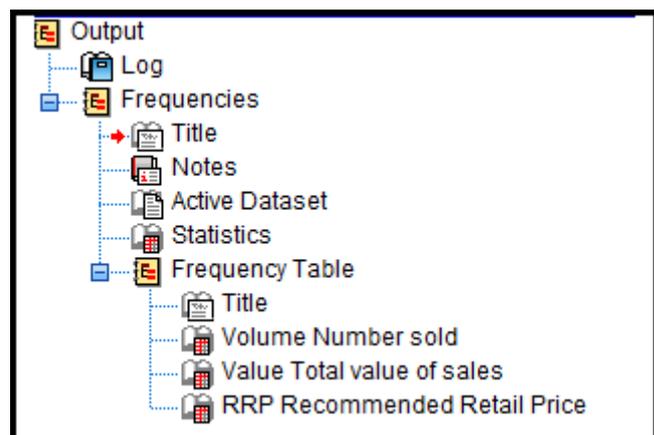
2.5.5 Display options for output labels in outlines

The SPSS output window (called the **Viewer**) has two 'panes' dividing the whole window vertically in two. The right pane contains the actual output e.g. a chart or table. The left pane contains the 'outline' which is a list of the items ("objects") in the right pane. On the right is an example of a left pane for a **Frequencies** procedure. The bottom five lines it shows that there is a Frequency Table with a Title and frequencies for three variables (Volume, Value, RRP – those are the variables' names). It is possible to specify that the outline should contain the variables' labels instead of, or as well as, their names.



To make changes use **Edit** → **Options** and select **Output Labels**. Then for **Outline Labeling** use the drop-down menu to select what you want for **Variables in item labels shown as**. Then click **Apply** and finish with **OK**.

To the right is the outline for the same procedure as above, but asking for both names and labels.



2.6 The Viewer window – displaying and handling output

2.6.1 The Viewer window

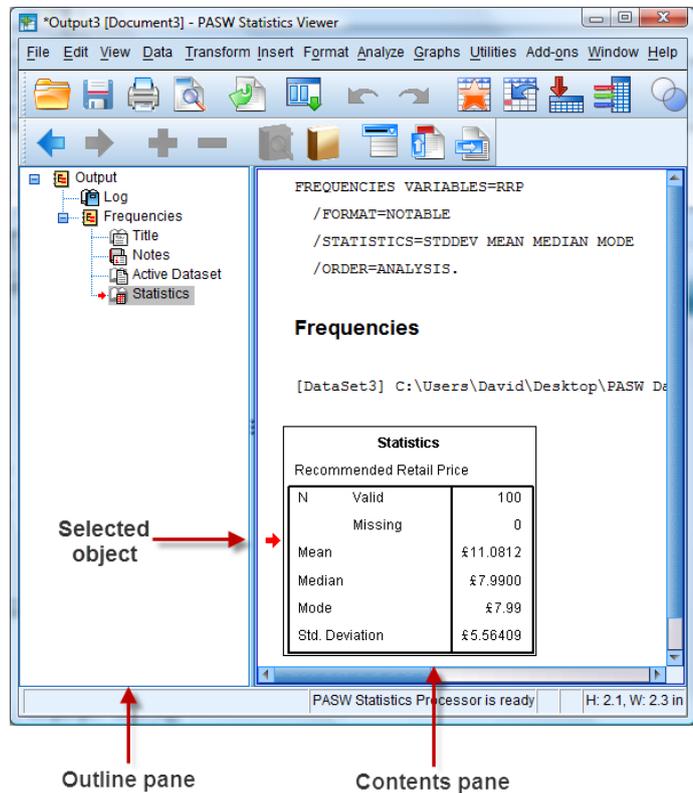
When SPSS performs an operation it creates output which is placed in an output file. This is displayed on the computer screen in the **Viewer** window, which would typically have the title:

***Output1 [Document1] – IBM SPSS Statistics Viewer**

The **Viewer** window displays the results from data analysis and statistical operations. It consists of two panes. In the left pane – called the **outline** pane – is a diagrammatic tree representation of the 'objects' in the right pane – called the **contents** pane, visible by scrolling if necessary. (See right.) These objects include titles, report of actions taken, name of the data file, and the output proper – i.e. statistical tables and charts.

Many actions in the outline pane have a corresponding effect on the contents pane.

- Clicking on an object's name in the outline pane or on the object in the right pane selects (or deselects) that object (shown by a very small red arrow on the left).
- Selecting an item in the outline pane brings the object itself into view in the contents pane.
- Moving an item in the outline pane moves the corresponding item in the contents pane.
- Selecting an object makes it available for editing, printing or exporting (e.g. to a Microsoft Word document).
- Double-clicking an object makes it amenable to editing: text and numbers can be changed, and the width of columns varied by dragging with the cursor.



2.6.2 Exporting output to Microsoft Word

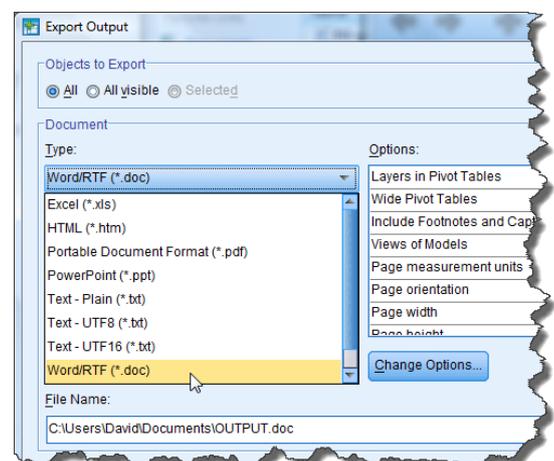
The whole or part of the output can be exported to a Microsoft Word document. This is very useful when writing reports. This is achieved by: **File** → **Export** which brings up the **Export Output** window →.

The default output **Type** is **WordRTF (*.doc)**.

See right for the full list of output choices →

The **Objects to Export** choices are controlled by three radio buttons (All; All visible; Selected):

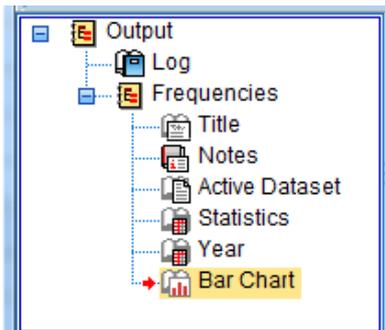
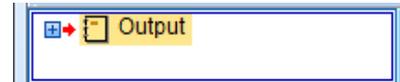
- **All** exports all the output plus hidden SPSS commands not visible on screen, so probably not wanted.



- **All visible** exports all the output apart from the hidden commands (N.B. it may not actually be showing on screen, depending on what you have selected)
- **Selected** exports only selected objects.

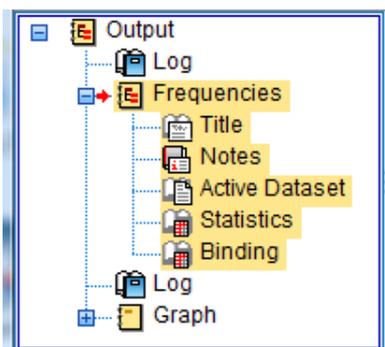
The selection can be just one object whose selection is shown by a short red pointer in the outline pane and in the contents pane (though it may be hidden unless you widen the window) or a whole group selected in the outline pane

Exports all the output if **All visible** chosen →



← Exports all the output if **All visible** chosen.

← Exports only the selected object (Bar Chart) if **Selected** chosen.



← Exports all the output if **All visible** chosen.

← Exports only the selected objects (all the Frequencies outputs) if **Selected** chosen.

2.6.3 Saving output in a standard SPSS output file

Output appearing in the **Viewer** window can be saved as an SPSS document with extension **.spv**. Use the normal **File** → **Save** or **File** → **Save As...** commands. It can then be retrieved using **File** → **Open** → **Output...**

2.6.4 Printing output

Output which appears in the **Viewer** window can be printed. The whole output will be printed unless just some of it has been selected (as indicated in 2.6.2).

Each object shown in the outline pane as an open book icon will print. Each object shown as a closed book icon will not print.

Opening and closing can be achieved by double-clicking or by clicking on the + or – sign.

Use the standard **File** → **Print** command.

2.6.5 Exiting SPSS

Use the standard **File** → **Exit** command.

If any open data files or output files have not been saved, the user is alerted.

3 Getting on-screen help

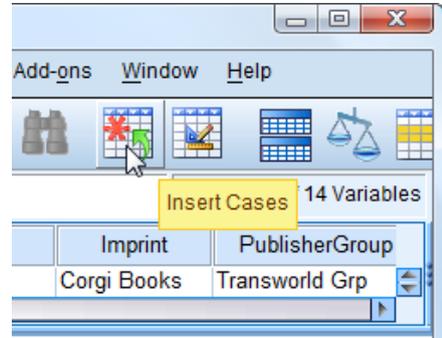
Once you have launched SPSS there are a number of ways in which you can obtain help on the screen. The information below is provided for reference; you are likely to find it useful later.

3.1 Toolbar

Move the arrow cursor across the toolbar and position it over any one of the icons.

Text describing the function will appear just beneath the cursor (see example here, available when in **Data View** mode) →.

The descriptive text also appears at the bottom of the window.

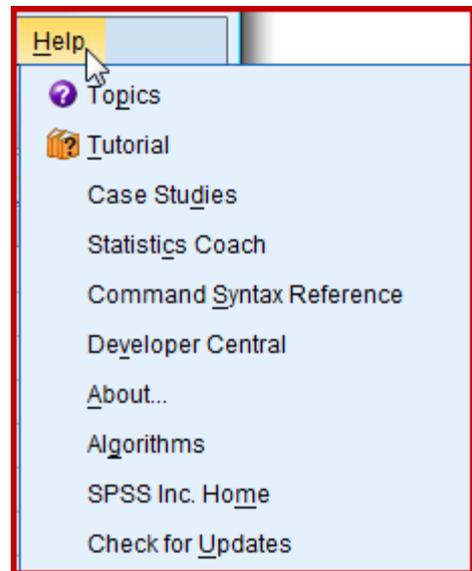


3.2 Help menu

Click on **Help** on the menu bar.

A drop down menu will appear. The four most useful menu options are:

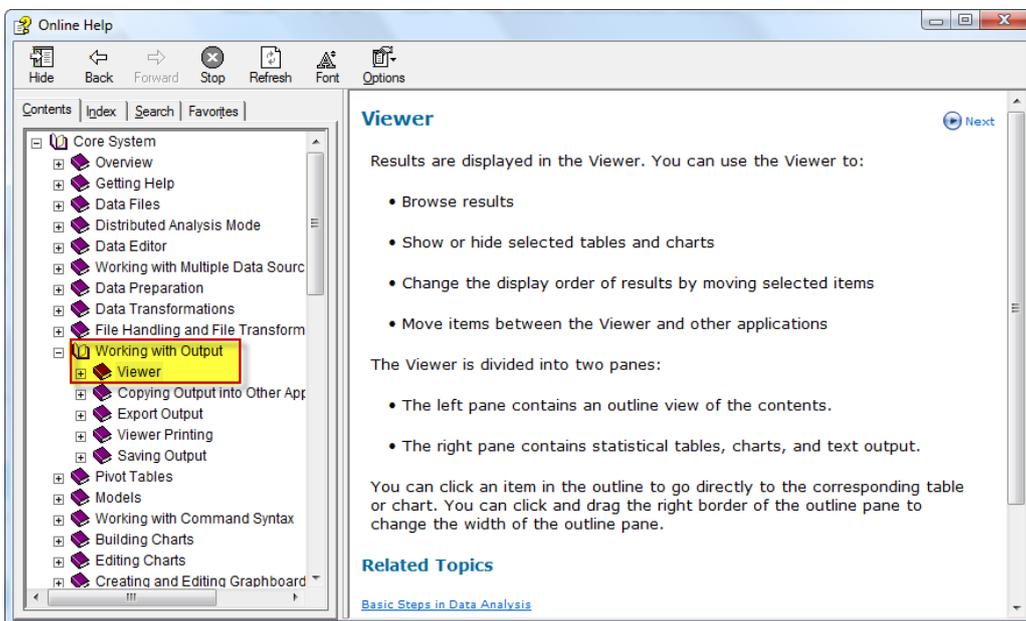
- **Topics**
- **Tutorial**
- **Case Studies**
- **Statistics Coach**



3.2.1 Topics

Click on **Topics** on the **Help** menu.

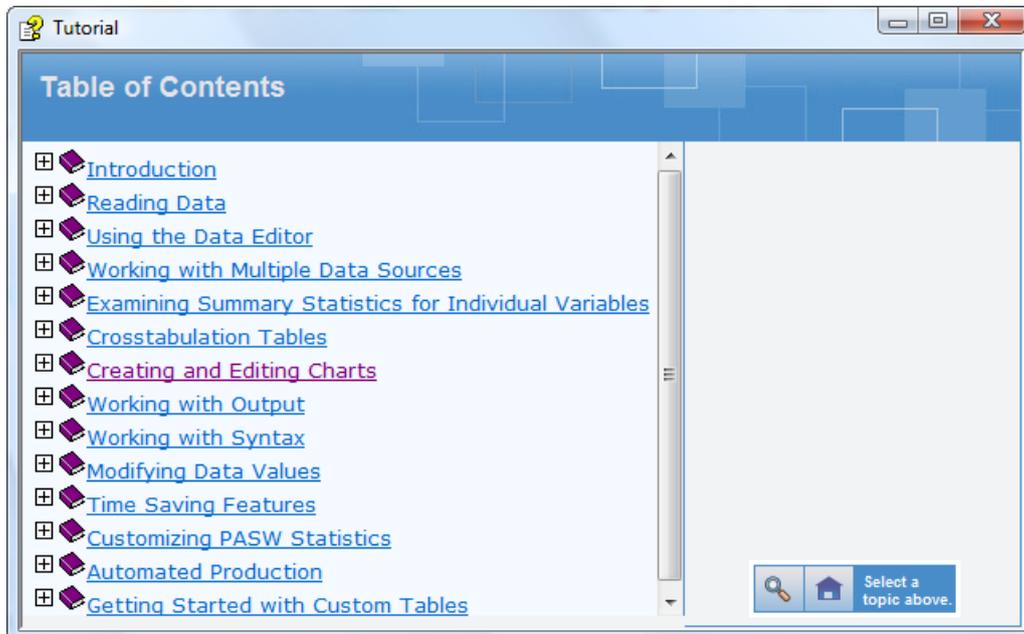
The Contents list appears from which a topic can be selected (see example below). Also available are an Index and a Search facility, into which one can type a keyword.



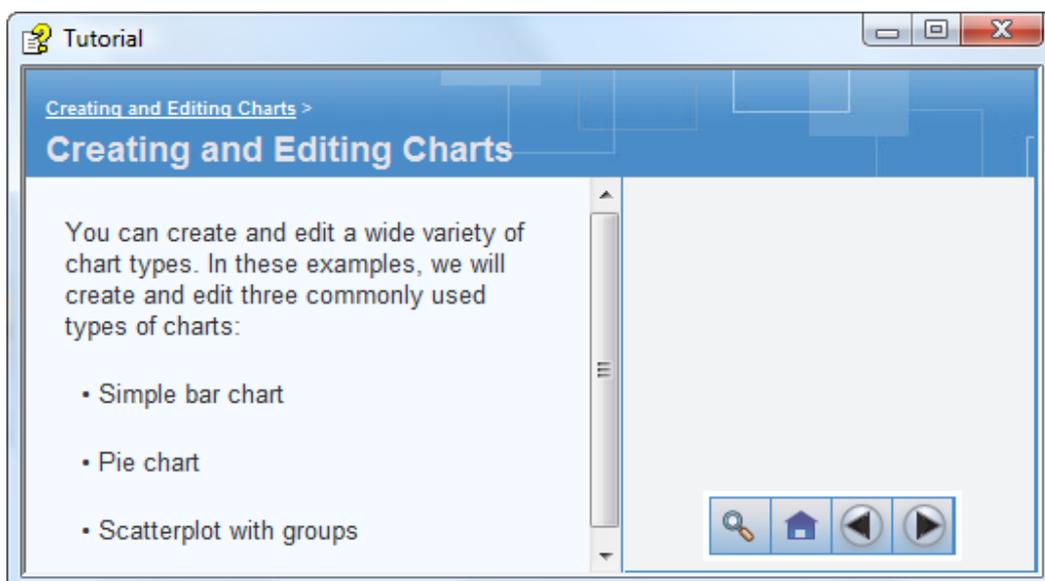
3.2.2 Tutorial

Click on **Tutorial** on the **Help** menu.

This presents a Table of Contents which one can browse to find illustrated, step-by-step instructions on many basic *SPSS* features.



Some of the tutorials use demo data files (see example below).



3.2.3 Case Studies

Click on **Case Studies** in the **Help** menu.

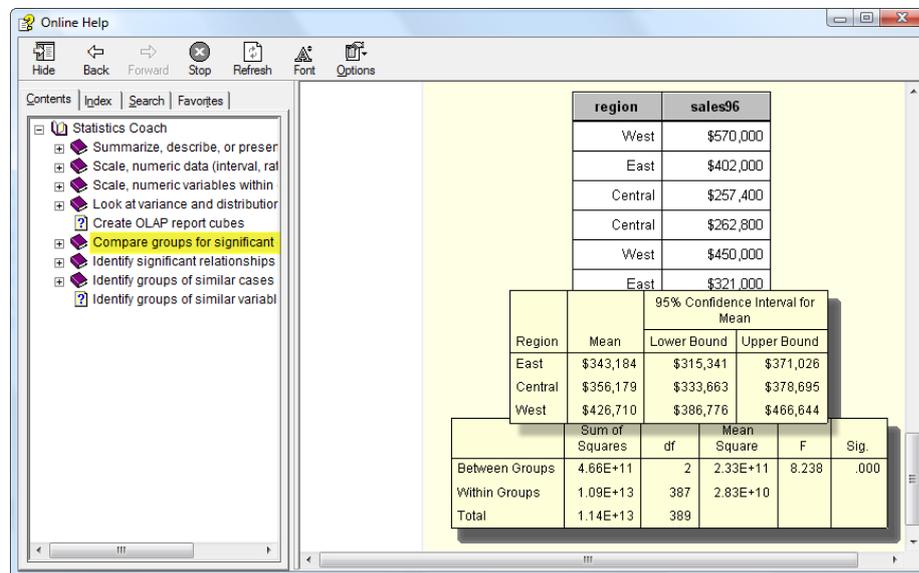
This displays a Table of Contents, the most useful options being:

- **Statistics Base**
- **Advanced Statistics Option**

3.2.4 Statistics Coach

Click on **Statistics Coach** in the **Help** menu.

A tutorial appears asking the user 'What do you want to do?' and presents information to help the user to select the appropriate presentation method or statistical test to employ. (See below.)



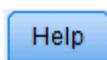
The screenshot shows the 'Online Help' window with the 'Statistics Coach' section active. The left-hand 'Contents' pane lists several tasks, with 'Compare groups for significant' highlighted in yellow. The main content area displays two tables:

region	sales96
West	\$570,000
East	\$402,000
Central	\$257,400
Central	\$262,800
West	\$450,000
East	\$321,000

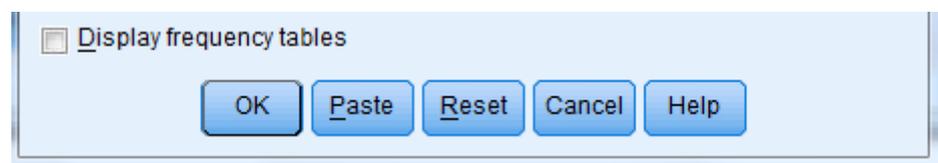
Region	Mean	95% Confidence Interval for Mean	
		Lower Bound	Upper Bound
East	\$343,184	\$315,341	\$371,026
Central	\$356,179	\$333,663	\$378,695
West	\$426,710	\$386,776	\$466,644

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.66E+11	2	2.33E+11	8.238	.000
Within Groups	1.09E+13	387	2.83E+10		
Total	1.14E+13	389			

3.2.5 Dialog Box help



If a Help button can be seen in the current window in which you are working, then you can click on it to obtain context-sensitive help.



A window titled **Online Help** will appear containing text specific to the task in hand. I.e. It takes you straight to the relevant Help page to save you having to look for it.

4 Creating variables and entering data

4.1 Variables

A **variable** represents a category of data collected, e.g. gender, age, nationality, response to a multiple choice question.

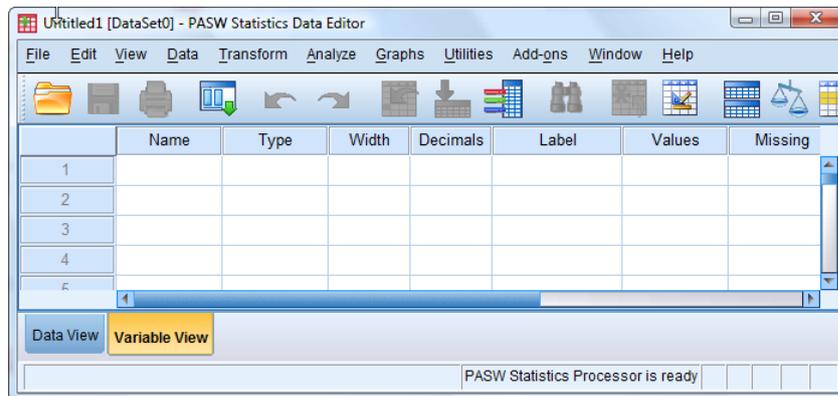
A **column** in the **Data Editor** window of *SPSS* stores all the values for one variable.

It is advisable, though not necessary, to create the variables in the data file *before* the data values are keyed in. If the column of data is entered first then *SPSS* creates a dummy name (**VAR0001**, and so on) which can be edited subsequently.

4.2 General instructions for creating variables

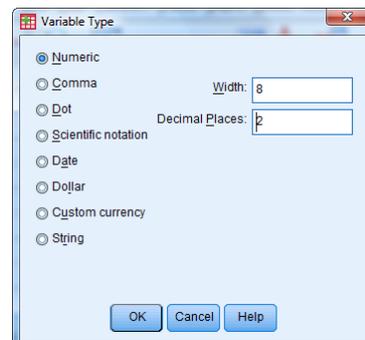
Note: a specific example is the subject of TUTORIAL T4.

1. Click on the tab **Variable View** at the bottom of the **Data Editor** window.
 - A window appears which looks like this:



2. Type in the first variable name into the first cell of the column headed **Name** and press ↓.
 - Maximum of 64 characters – upper and lower case letters, digits and some symbols
 - Name must begin with a letter or @ symbol
 - No spaces allowed (underscore is often used in its place)
 - Punctuation marks are allowed (e.g. full stop)
 - Cannot be one of the few reserved keywords (e.g. NOT)
 - Default attributes of the variable appear in columns to the right of **Name**. These are:

- **Type** – Numeric
- **Width** – 8
- **Decimals** – 2 d.p.
- **Label** – blank
- **Values** – None
- **Missing** – None
- **Columns** – 8
- **Align** – Right
- **Measure** – Unknown
- **Role** – Input



All these attributes can be edited, as explained below.

It is best to have a short variable name that is easily recognised as related to the underlying variable or to the source (e.g. question number).

3. Type

To change the variable type, click the appropriate cell in the column titled **Type** (which will normally contain the default **Numeric**), click on the three dots shaded in grey to the right of **Numeric** and click again to open the **Variable Type** dialog box.

Make your selection and click on **OK**.

4. Width

For a string variable, the width determines the maximum number of characters allowed in the string. For example if a string has width set to 3 then 'CAT' can be entered but 'MOUSE' cannot. It can be useful when entering string data to prevent certain input mistakes. A string can have a maximum 32767 characters! Any string shorter than the width is 'padded' on the right with blanks.

For all other variables (numeric), the width specifies the expected maximum width for the number (not the maximum number of digits allowed). A numeric can have a maximum 40 digits (maximum 16 decimal places)

To change the width of a variable click the appropriate cell in the column titled **Width**, and then type in the value you want or use the up and down scroller. (Alternatively, you could edit the **Width** within the **Variable Type** box discussed in 3 above.)

5. Decimals

To change the number of decimal places of a numeric variable displayed, click the appropriate cell in the column titled **Decimals**, and type in the desired number or use the up and down scroller. The maximum is 16. This does not affect the actual number of decimals in the variable. (Alternatively, you could edit the **Decimal Places** within the Variable Type box discussed in 3 above.)

6. Label

To enter a variable label, move the cursor to the **Label** column and click the appropriate cell, and then type in the label of your choice.

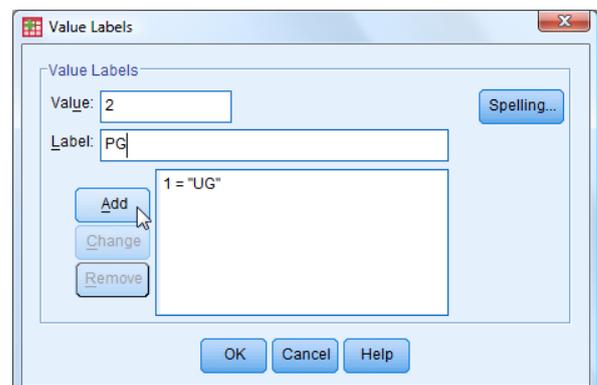
- The variable label is an explanation of what the variable is, e.g. if the variable name was **sex** then the label might be 'gender of the respondent'.
- The label must not exceed 256 characters including spaces.

7. Values

To enter values and value labels, move the cursor to the **Values** column and then click it. Click on the three highlighted dots just to the right of **None** to open the **Value Labels** dialog box.

Use VLE	{1, Yes}...	None
No of modules accessed	None	None
Frequency of VLE use	{1, Several t...	None

- The **Value Labels** dialog box appears.
- Values are numbers or strings (sets of characters) used to represent (codify) data.
E.g. for the variable **sex** it might be
1 = 'male', 2 = 'female' or
M = 'male', F = 'female'.
- The maximum length is 120 characters.



By way of example, suppose a question can be answered ‘Yes’, ‘No’, or ‘Don’t know’ which are coded ‘1’, ‘2’, ‘3’ respectively. To enter values and value labels proceed as follows:

First, type in the first value that can represent the variable – in this case ‘1’.

Second, click on the **Value Label:** field of the **Value Labels** window and type in what the value represents – in this case ‘male’.

Third, click on the **Add** button of the **Define Labels** window.

- The value and its meaning now appear in a box next to the **Add** button.

Repeat the process to insert the ‘No’ and ‘Don’t know’ values.

When completed click on **OK**.

8. Missing

To define missing values, move the cursor to the **Missing** column and then click the appropriate cell. Click on the three highlighted dots to the right of **None** to open the **Missing Values** dialog box.

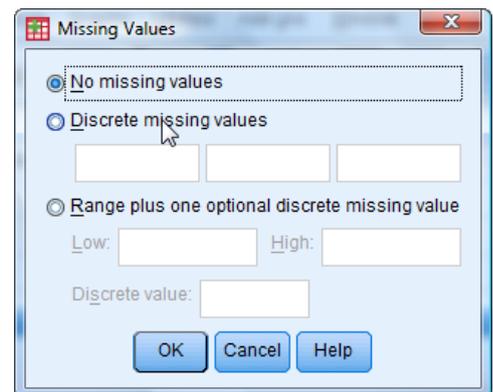
A missing value is where there is no valid response entered for a variable.

The purpose of defining missing values is to prevent SPSS including them when doing calculations (e.g. finding the mean of a set of numbers).

- A window titled **Missing Values** appears.

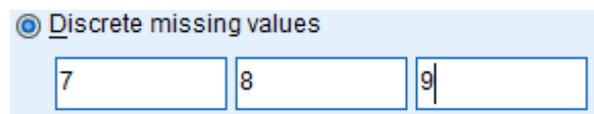
It is common to have just one discrete code to signify a missing value (e.g. ‘0’ in cases where ‘1’ represents ‘male’ and ‘2’ represents ‘female’), but SPSS will allow three different codes or one single code together with a specified range of codes.

Click on the **Discrete Missing Values** button of the **Missing Values** window.



Type in the number representing the missing value response.

- E.g. ‘9’ to signify no response to an MCQ which has allowable choices coded ‘1’ to ‘5’.
- A maximum of three discrete missing values can be typed in: e.g.

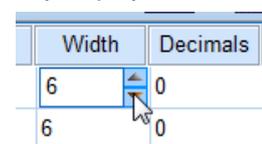


- ‘9’ for ‘no response’,
- ‘8’ for ‘selected more than one answer’,
- ‘7’ for ‘no choice selected but wrote a comment’.

Click on the **OK** button of the **Missing Values** window.

9. Column

The width of a column displayed in the **Data Editor** in **Data View** mode can be set using this. It does not affect the underlying variable, only what is actually displayed. In practice, this is little used as the widths of columns in **Data View** can be over-riden by dragging. To change the declared width, in **Variable View** type in the desired number or use the up and down scroller to vary the current entry.



10. Align

The default alignment of numeric data in the **Data Editor** in **Data View** is the left margin, and the default alignment for string data is the right margin. Sometimes it is preferable to centre data or even align the data to the opposite margin. This column enables this to be done by selecting the variable's cell in the **Align** column and clicking on the downward arrow to select one of the alternatives **Left**, **Right** or **Center**.

11. Measure

Initially, **Measure** is set to 'Unknown'.

For all **Types** except 'String' you need to click on **Measure** and choose from 'Nominal', 'Ordinal', 'Scale'.

If you set **Type** as 'String' then **Measure** is automatically set to the default 'Nominal'. String variables can be designated 'Nominal' or 'Ordinal'.

For some chart drawing procedures it is important to be able to specify if measurement is 'Nominal' or 'Ordinal'.

See Section 2.4.3 for more information on data types.

12. Role

This is a new advanced feature, beyond this Guide's scope. (Default is: **Input**.)

5 Loading a data file and Editing data in a data file

5.1 Loading a data file

1. If you are opening a data file on a USB memory device, insert it.

2. Select **File** → **Open** → **Data**.

- ▶ Note that the **Files of type** box contains the default: **SPSS Statistics (*.sav)**.

4. Click on the down arrow next to the **Look in** box.

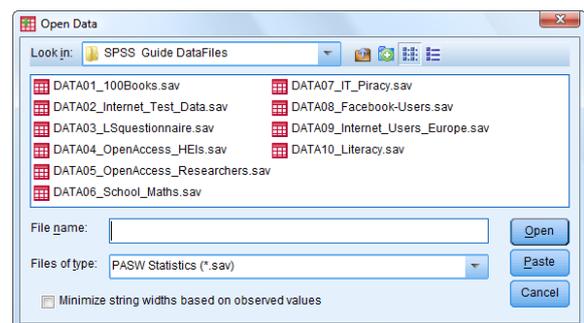
- ▶ By scrolling and clicking as required you can reveal the names of folders, and any files not in folders (which are in **.sav** format), on your PC desktop or USB memory stick.

- ▶ A horizontal scroll bar will appear below if there are too many entries to fit in the window.

5. Locate and select the folder or file required. Double-click a folder name to reveal the list of files (which are in **.sav** format) and click on the required file.

6. Click on the **Open** button to import the file into *SPSS*.

- ▶ The selected file will now be loaded into a **Data Editor** window in **Data View** mode.



Data stored in the cells displayed in the **Data Editor** window can be edited in several ways.

5.2 Changing data in a cell

1. To change an entry, click on the cell which will highlight the data, and key in the new data and press the **Enter** key.
2. To delete an entry, click on the cell and press **Backspace** key or use **Edit** → **Clear**.

5.3 Copying or moving data in a row, column or block of cells

1. Click *and hold down the mouse button* on the first (top left) cell that is to be copied or moved.
2. Drag the mouse pointer to the last (bottom right) cell to highlight the block of cells that are to be copied or moved, and release the mouse button.

Alternatively to 1 and 2, click the top-left cell and then shift-click the bottom-right cell.

3. Click on **E**dit on the menu bar.
4. If the cells are to be copied click on **C**opy on the drop down menu, and then go to step 6.
5. If the highlighted cells are to be moved click on **C**ut on the drop down menu.
 - The highlighted cells will disappear.
6. Click on the first destination cell to which the copied cells are to be moved or copied.
7. Click on **E**dit on the menu bar.
8. Click on **P**aste on the drop down menu.
 - The copied or moved cells will appear in the new location.

5.4 Inserting a new case (row)

1. Ensure **Data View** is selected (by clicking on the tab at the bottom of the **Data Editor** window if necessary).
 2. Then click on the case number immediately below where you want the new case inserted, which selects that case (row).
 3. Click on **E**dit on the menu bar.
 4. Click on **I**nsert Cases on the drop down menu.
 - The new case will be inserted immediately above the selected row.
 5. Multiple new cases can be inserted at once by highlighting multiple rows.
-

5.5 Deleting a case (row)

1. Ensure the **Data View** is selected (by clicking on the tab at the bottom of the **Data Editor** window if necessary).
2. Click on the number of the case to be deleted (this action selects the case).
3. Click on **E**dit on the menu bar.
4. Click on **C**lear on the drop down menu
 - *Alternatively* after step 2 simply press **Delete** or **Del**.

5.6 Inserting a new variable

1. If **Variable View** is selected, click on the number of an existing variable (or blank row) where the new variable is to be inserted (this selects the row for that variable).

If **Data View** is selected, click on the name of an existing variable (or blank column) where the new variable is to be inserted (this selects the column for that variable).
2. Click on **E**dit on the menu bar
3. Click on **I**nsert Variable on the drop down menu.
 - The new variable will be inserted (it will be named **VAR00001** or similar, which you can edit).

5.7 Duplicating data in a variable

It can be useful to have a variable with data identical to another. The data can be placed in an existing variable (whether empty or not) or in a new position, creating a new variable.

1. Ensure **Data View** is selected.
2. Click on the variable name at the top of the column to be copied. This highlights the column.
3. Click on **E**dit on the menu bar.
4. Click on **C**opy on the drop down menu.
5. Click on the column into which the data is to be inserted.
6. Click on **P**aste on the drop down menu.
 - All the data will be placed in the new variable.

5.8 Deleting a variable

1. If **Data View** is selected, click on the variable name at the top of the column.
If **Variable View** is selected, click on the variable number at the left of the row
 - These actions highlight the selected variable.
 2. Click on **E**dit on the menu bar.
 3. Click on **C**lear on the drop down menu.
-

6 Saving a data file

Saving the data that has been keyed in or amended is very important. This ensures that if a computer malfunction or human error occurs then the data is retrievable.

If data is not saved regularly there is the danger of losing a great deal of work.

Data should be saved on a regular basis and especially after:

- existing data has been significantly amended;
- several new items of data have been keyed in.

It is wise to save the data into a new file each time rather than overwriting the existing file, using a progressive numbering system.

1. If you are saving to a USB drive, insert it as normal.
2. Click on the **Data Editor** window to make sure it is active.
3. If the data is to be stored overwriting the current file (not normally advised!):

Click on **File** on the menu bar and **Save** on the drop down menu.

4. If the data is to be saved in a new file use the follow steps:

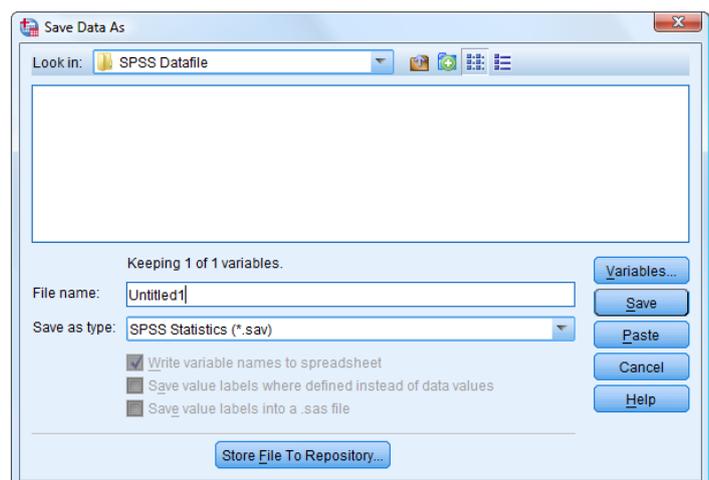
Click on **File** on the menu bar, and **Save As** on the drop down menu.

- The **Save Data As** window appears.

5. Click on the down arrow next to the **Save In** box and locate and select the USB drive, or your Central File Store, or wherever else you wish to store the data.

6. Click on the **File name** field of the **Save Data As** window and type in the desired file name.

- maximum length is 260 characters (including any path name)
- spaces are not allowed but underscore is allowed
- none of the following characters are allowed:
\ | / < > : " ? *



- Note: the dot symbol should be used in a name with caution. If the dot is followed by exactly three letters SPSS will think it is an extension name (determining the file type) and will not save the file as a .sav file in which case it will not be recognised by SPSS when you try to locate it to load in again.

7. Click on the **Save** button.

7 Importing and Exporting data in Microsoft Excel format

7.1 Importing data in Microsoft Excel format

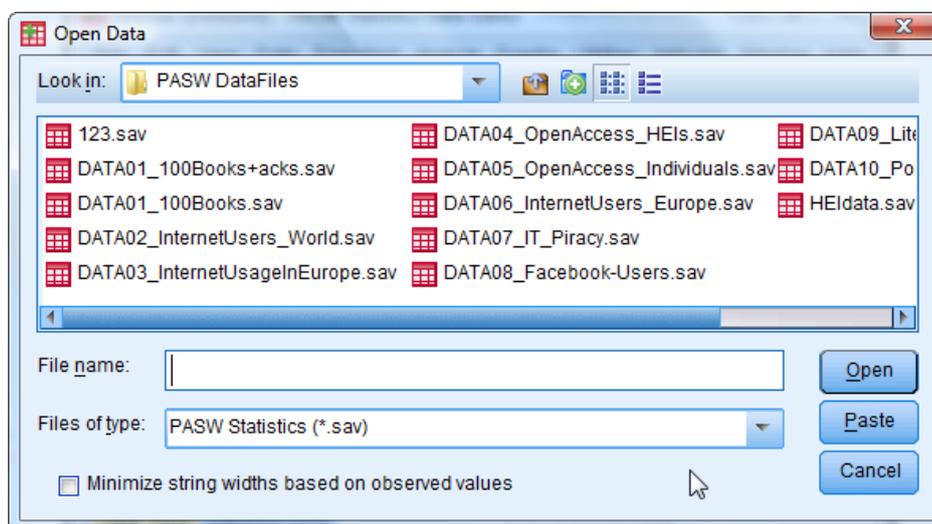
Data can be imported from a wide range of spreadsheets, databases and text files. The most common is Microsoft *Excel* spreadsheet, which is the subject of this section. (For other valid sources and how to import then see the Online Tutorial 'Reading Data'.)

It can be easier to first enter one's raw data in *Excel* and only when satisfied with it to transfer it to *SPSS*. Sometimes secondary data will already be available in *Excel*, which can be copied straight across to *SPSS*.

It is important that the *Excel* data is set out with rows for the cases and columns for the variables.

Locating the required *Excel* data file can be a challenge, as the following will indicate.

Selecting **File** → **Open** → **Data** will produce an **Open Data** window something like this:



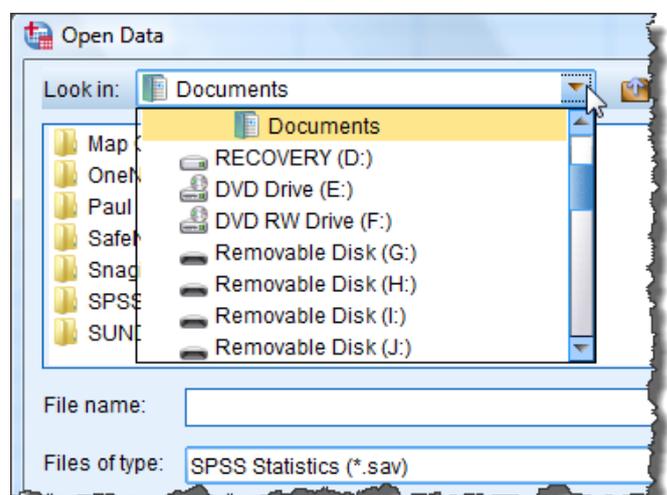
The **Look in** entry specifies whereabouts (in computer memory, USB sticks, etc.) data files and folders are being looked for.

This can be changed by clicking on the down arrow and navigating through memory.

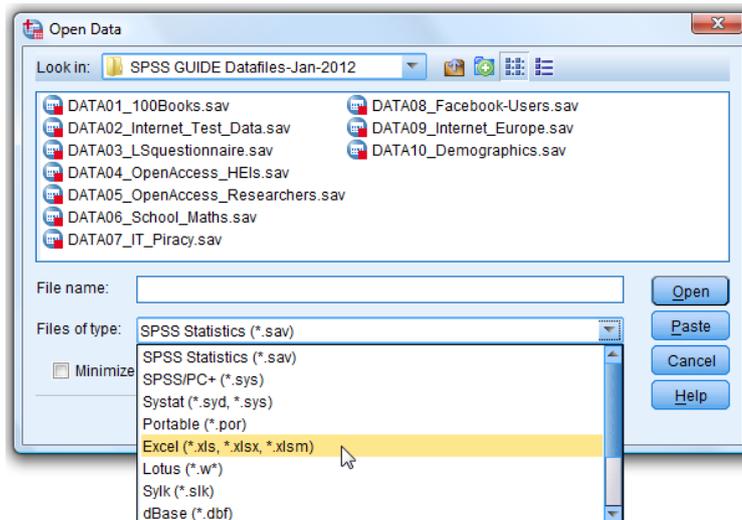
Note that the **Files of type** choice is **SPSSStatistics (*.sav)**. This means that the only *SPSS* **.sav** files (which are not hidden in folders) will appear listed. Folders which may contain such files will also be listed.

The next step, then, is to make *Excel* files visible.

This requires changing the **Files of type** choice to **Excel (*.xls, *.xlsx, *.xlsm)**.



This is done by opening the **Files of type** box down and selecting the format required, like this:

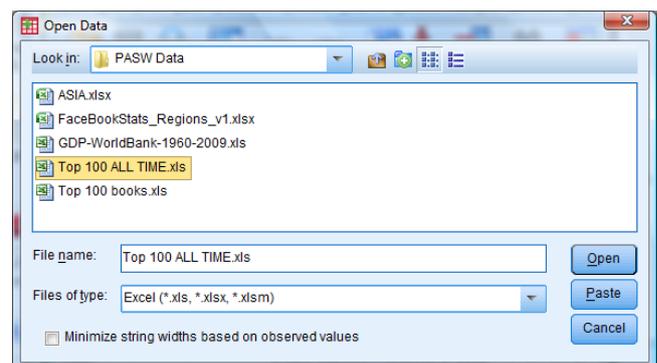


There is an option to display all types of files – **All Files (*.*)** – which can be useful sometimes. (It is hidden at the bottom of the list of formats, so scroll down to reveal it.)

The next step is to search through memory to locate the required *Excel* file (or the folder which contains it).

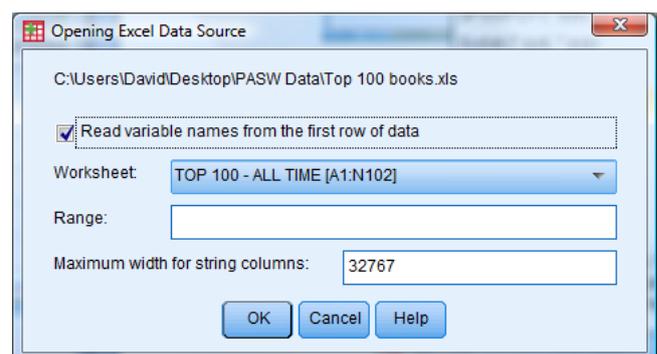
You can double-click any folder to open it to see what is inside. With any luck you will find the *Excel* file you want.

The **Open Data** window will appear, looking something like this →



Having located and selected the required *Excel* file, the next step is to load the data into *SPSS*. This is achieved by clicking on the **Open Data** window's **Open** button which opens another window like this →

It names the *Excel*/Worksheet file and specifies the range of cells in which it has found data.



An important matter is whether the first *Excel* row contains headings (which can become the variable names) or whether there are no headings and the first row represents data for the first case.

SPSS assumes the first row *are* headings (indicated by a tick in the 'Read variable names from the first row of data' tick box – see above). It is essential to indicate if this is *not* true by removing the tick, otherwise the first case will be lost.

If no *Excel* headings are indicated, then *SPSS* invents variable names (**VAR0001**, **VAR0002** and so on, which can be edited).

Excel headings, if present, may not conform to *SPSS* requirements (e.g. spaces are not allowed in *SPSS* variable names; names must begin with a letter or @ symbol). In such cases *SPSS* amends the variable names to conform.

The *SPSS* default is to import *all* of the spreadsheet but the user can import just *part* of the spreadsheet by specifying a reduced range (e.g. A1:G50). In fact, it is a good idea to specify the correct range, even when you want the whole spreadsheet, as *SPSS* has a habit of loading in superfluous cases and variables and filling them with the system missing value symbol (a full stop). (This can happen, for example, if there is a spurious entry hidden away outside the correct range.) These phantom cases and variables would have to be deleted before analysing the data.

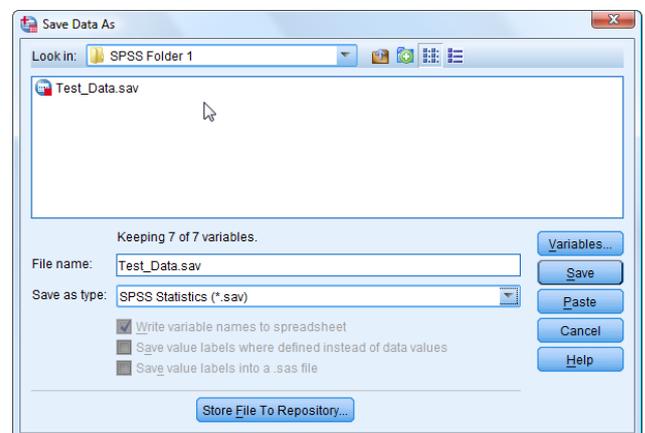
7.2 Exporting data into Microsoft Excel format

SPSS data files can be exported into a wide range of file format. The most commonly used is Microsoft *Excel*, which is the subject of this Section. (The procedure is very similar for other formats.)

Having created an *SPSS* data file (e.g. Test_Data.sav) one proceeds as follows:

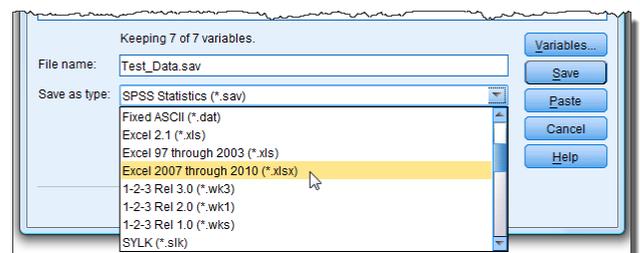
1. **File** → **Save As**

- ▶ This will open the **Save Data As** window giving access to the PC's desktop and folders (and also to connected storage media such as USB sticks).



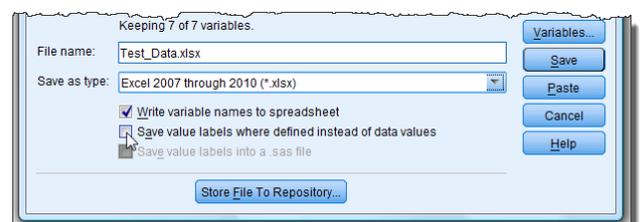
2. Once the desired location in which to save the file has been found, open the **Save as type** drop-down-menu – the default is PASW Statistics (*.sav) which must be changed.

- ▶ The menu will list all the possible formats, including three for different versions of *Excel* →

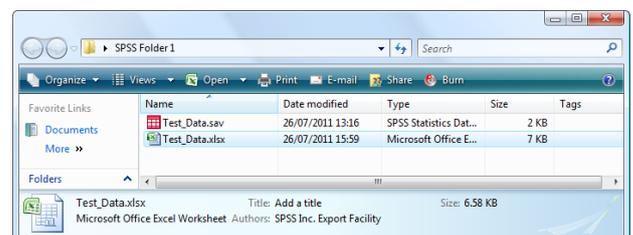


3. Select the format required (scrolling through the list if necessary) →

4. Choose whether or not to save variable names (as column headings in *Excel*) → and whether or not to save value labels → (e.g. 'Male' and 'Female') rather than raw data values (e.g. '1' and '2').



- ▶ Note that unless you change it, the *Excel* file will have the same name as the corresponding *SPSS* file, as illustrated here →



- ▶ The *Excel* file will, of course, have a different extension (.xls or .xlsx) so the *SPSS* and *Excel* versions will be quite distinct.

8 Sorting cases and Selecting cases

8.1 Sorting cases

It can be useful to change the order of cases, and this is easily achieved. For example, the data set of the 100 top-selling UK books published by Nielsen in December 2010 quite naturally lists them in position based on volume. Below we can see that *The Da Vinci Code* tops the list (i.e. **Case 1** has variable **Position** with value 1) with 4,522,025 sales.

	Position	Title	Author	Imprint	PublisherGroup	Volume	Value	RRP	ASP	Binding	Month	Year
1	1	Da Vinci Code,The	Brown, Dan	Corgi Books	Transworld Grp	4,522,025	£22,857,837.53	£7.99	£5.05	Paperback	3	2003
2	2	Harry Potter and the Philosopher...	Rowling, J. K.	Bloomsbury ...	Bloomsbury Grp	3,844,316	£19,853,187.43	£6.99	£5.22	Paperback	6	1997
3	3	Harry Potter and the Chamber of ...	Rowling, J. K.	Bloomsbury ...	Bloomsbury Grp	3,184,492	£16,224,021.98	£6.99	£5.07	Paperback	4	1998

It could be interesting to see if that or some other book earned the most money. There is a variable for that – **Value** – so we can sort on it, as follows:

Data → **Sort Cases** produces the **Sort Cases** window below left. Selecting **Value** and moving it using the blue arrow into the **Sort by** box, clicking on the radio button **Descending** produces the window below right. Clicking **OK** then initiates the sorting. The '(D)' after the variable name indicates that descending order sorting has been chosen.

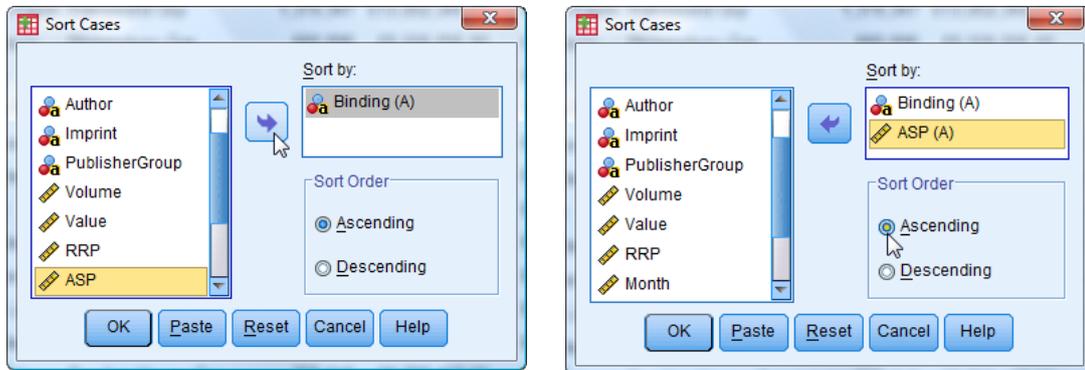


The result is shown below:

	Position	Title	Author	Imprint	PublisherGroup	Volume	Value	RRP	ASP	Binding	Month	Year
1	5	Harry Potter and the Order of the...	Rowling, J. K.	Bloomsbury ...	Bloomsbury Grp	3,043,226	£33,925,431.19	£16.99	£11.15	Hardback	6	2003
2	6	Harry Potter and the Half-blood P...	Rowling, J. K.	Bloomsbury ...	Bloomsbury Grp	2,947,565	£28,030,020.14	£16.99	£9.51	Hardback	7	2005
3	7	Harry Potter and the Deathly Hall...	Rowling, J. K.	Bloomsbury ...	Bloomsbury Grp	2,842,059	£25,028,396.30	£17.99	£8.81	Hardback	7	2007
4	1	Da Vinci Code,The	Brown, Dan	Corgi Books	Transworld Grp	4,522,025	£22,857,837.53	£7.99	£5.05	Paperback	3	2003

This shows that *The Da Vinci Code* is relegated to fourth place and *Harry Potter* books occupy the first three places. Note that when SPSS sorts cases it always moves the whole row, i.e. it preserves the case intact. (This does not happen with Microsoft Excel unless the whole spreadsheet is selected.)

This example sorted on one criterion – **Value**. Multiple criteria sorting is possible by moving more than one variable into the **Sort by** box (in the correct order). For example, to find the cheapest hardback book in the Top 100 one would need to sort first on **Binding** (ascending order – alphabetical) and then on **ASP** (ascending order – numerical). Some of the stages for this are illustrated below.



The result, below, shows that *The Tales of Beedle the Bard* gave JK Rowling another first place, at a remarkable ASP of only £4.03, with *The Very Hungry Caterpillar* not far behind.

	Position	Title	Author	Imprint	PublisherGroup	Volume	Value	RRP	ASP	Binding	Month	Year
1	37	Tales of Beedle the Bard, The	Rowling, J. K.	Bloomsbury ...	Bloomsbury Grp	1,039,823	£4,185,291.09	£6.99	£4.03	Hardback	12	2003
2	54	Very Hungry Caterpillar, The	Carle, Eric	Puffin Books	Penguin Grp	855,920	£4,101,816.11	£5.99	£4.79	Hardback	9	195
3	7	Harry Potter and the Deathly Hall...	Rowling, J. K.	Bloomsbury ...	Bloomsbury Grp	2,842,059	£25,028,396.30	£17.99	£8.81	Hardback	7	200

8.2 Selecting cases

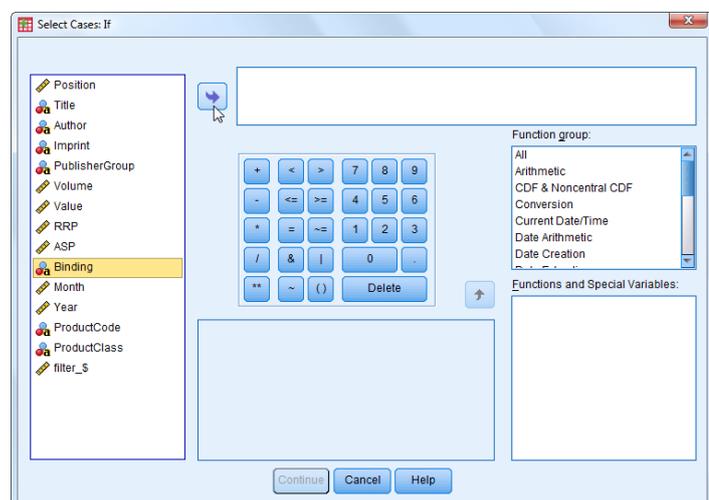
Selecting cases is quite different from sorting cases. **Select Cases** is used to choose cases for analysis based on one or more criteria, and so to eliminate unwanted cases from the analysis. (In contrast, sorting cases is really a cosmetic exercise as far as statistical analysis is concerned – the order does not matter and all cases will be included in any analysis unless **Select Cases** is employed.)

As an example, suppose we wished to select just those books among the top 20 best-selling Paperback books with RRP under £10 first published in either of the two years 2003 and 2004 this could be achieved as follows. First the criteria need to be put into terms SPSS will be able to interpret. They are:

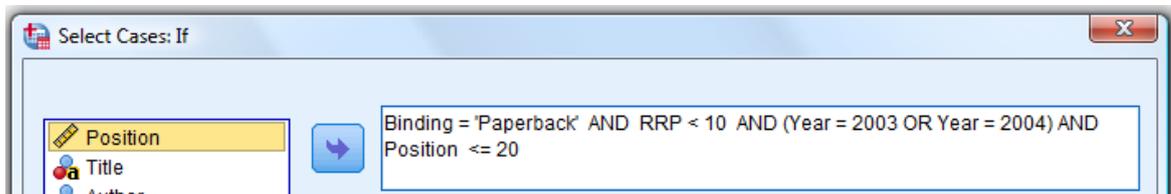
- Criterion 1: **Binding** = Paperback
- Criterion 2: **Value** < 10
- Criterion 3: **Year** = 2003 OR
Year = 2004
- Criterion 4: **Position** <= 20

The next step is to enter the formula required:

Data → **Select Cases** followed by clicking on the radio button for **If condition is satisfied** followed by clicking on the **If** button will produce the **Select Cases:If** window (shown right).



By a combination of moving variable names into the formula box using the blue arrow and typing in numbers and arithmetic symbols (using the supplied on-screen keypad or the keyboard) the following formula is inserted:



SPSS is very fussy about how the formula is constructed so it is wise to write it out on paper first – and think about it!

- Note the use of double-quotes round the string variable name.
- Note the use of 'OR' to specify the years required – using 'AND' would be wrong because no book could be first published in BOTH years – it's one or the other.
- Note the use of brackets – essential when both 'OR's and 'AND's occur to clarify meaning.
- Note that there are always alternative ways to set out a formula.

For further guidance on using **Select Cases** and in particular constructing formulae see the online Help (**Help** → **Topics** then **Search** for 'select cases').

The result of applying the formula is shown here. Six cases are selected (highlighted with arrows).

SPSS puts diagonal lines through row numbers of excluded case. There are just six without a line through that have been selected as satisfying the criteria.

It is wise to check one has the right outcome – getting these formulae right can be a challenge.

	Position	Title	Author	RRP	ASP	Binding	Month	Year
→	1	Da Vinci Code,The	Brown, Dan	£7.99	£5.05	Paperback	3	2004
/	2	Harry Potter and the Philosopher...	Rowling, J. K.	£6.99	£5.22	Paperback	6	1997
/	3	Harry Potter and the Chamber of ...	Rowling, J. K.	£6.99	£5.07	Paperback	4	1999
→	4	Angels and Demons	Brown, Dan	£7.99	£5.05	Paperback	7	2003
/	5	Harry Potter and the Order of the...	Rowling, J. K.	£16.99	£11.15	Hardback	6	2003
/	6	Harry Potter and the Half-blood P...	Rowling, J. K.	£16.99	£9.51	Hardback	7	2005
/	7	Harry Potter and the Deathly Hall...	Rowling, J. K.	£17.99	£8.81	Hardback	7	2007
/	8	Harry Potter and the Prisoner of ...	Rowling, J. K.	£6.99	£5.05	Paperback	4	2000
/	9	Twilight	Meyer, Stephenie	£7.99	£4.80	Paperback	3	2007
/	10	Harry Potter and the Goblet of Fire	Rowling, J. K.	£8.99	£5.22	Paperback	7	2001
→	11	Deception Point	Brown, Dan	£7.99	£4.95	Paperback	5	2004
/	12	New Moon	Meyer, Stephenie	£7.99	£4.83	Paperback	9	2007
/	13	Lovely Bones,The	Sebold, Alice	£7.99	£5.24	Paperback	12	2009
→	14	Digital Fortress	Brown, Dan	£7.99	£5.02	Paperback	3	2004
→	15	Curious Incident of the Dog in th...	Haddon, Mark	£7.99	£5.40	Paperback	4	2004
/	16	Eclipse	Meyer, Stephenie	£7.99	£4.86	Paperback	7	2008
/	17	Girl with the Dragon Tattoo,The: ...	Larsson, Stieg	£7.99	£5.51	Paperback	7	2008
/	18	Kite Runner,The	Hosseini, Khaled	£7.99	£6.08	Paperback	6	2004
→	19	Time Traveler's Wife,The	Niffenegger, Audrey	£7.99	£5.50	Paperback	5	2004
/	20	World According to Clarkson,The	Clarkson, Jeremy	£7.99	£5.18	Paperback	5	2005

- There are many other options in the **Select cases** window – selection can be randomly, by row number, by date, by variable values and ranges, etc.

9 Computing variables and Recoding variables

9.1 Computing variables

The dataset shown here contains the population and numbers of internet users for each region of the world for 2011.

We wish to calculate the 'internet penetration rate' for each region and save the values in a new variable.

The calculation for this is:

$$\text{Internet_Users} / \text{Population} * 100.$$

This is easily accomplished in SPSS, as follows. (The reader might like to load the file **DATA11_Internet_Users_Worldwide** and carry out the following actions.)

Transform → **Compute Variable...** produces the **Compute Variable** window. Inside this window, in the **Target Variable** box, must be typed a name for the new variable (e.g. **Penetration**).

By a combination of moving variable names into the **Numeric Expression** box, using the blue arrow, and typing in numbers and arithmetic symbols (using the supplied on-screen keypad or the computer keyboard) the following calculation must be inserted:

$$\text{Internet_Users} / \text{Population} * 100.$$

(See screenshot on the right.)

This is very similar to Reference Section 8.2 – Selecting Cases.)

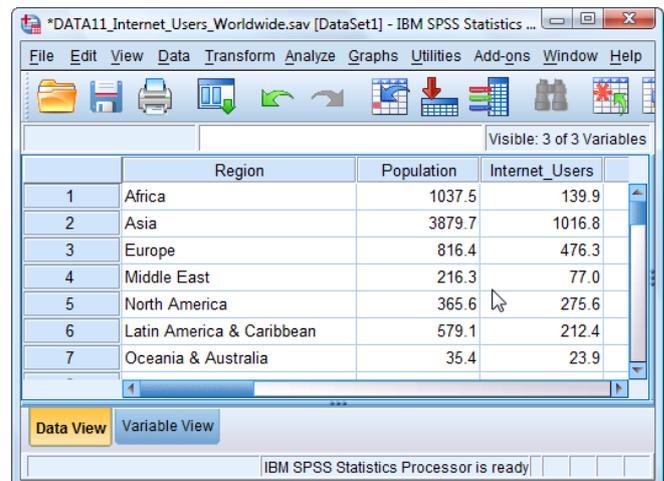
Finally click **OK**.

The new variable is created and inserted in the data file (see right).

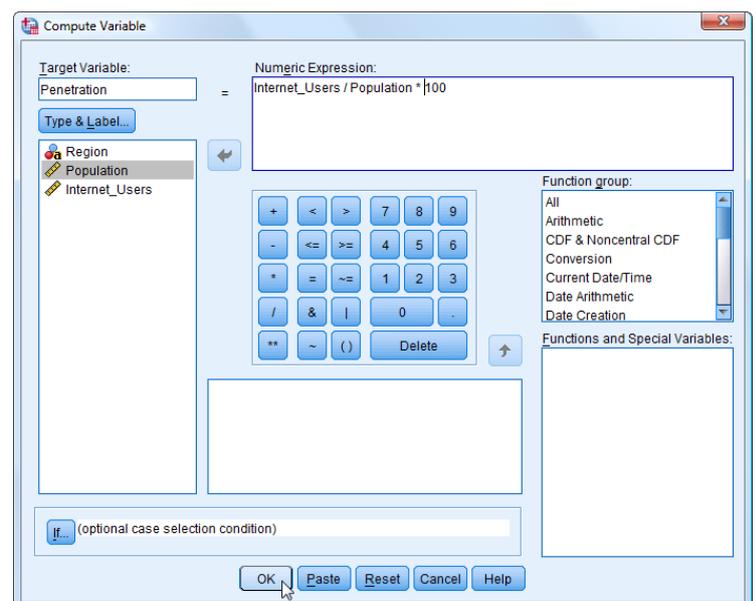
Its attributes can be modified, if necessary, in **Variable View**.

It can be used in statistical analysis or further calculations using **Compute Variable**.

Unless the data file is saved (preferably with a different name) the updated file will be lost when SPSS is closed.



	Region	Population	Internet_Users
1	Africa	1037.5	139.9
2	Asia	3879.7	1016.8
3	Europe	816.4	476.3
4	Middle East	216.3	77.0
5	North America	365.6	275.6
6	Latin America & Caribbean	579.1	212.4
7	Oceania & Australia	35.4	23.9



Compute Variable dialog box showing the configuration for creating the variable 'Penetration'.

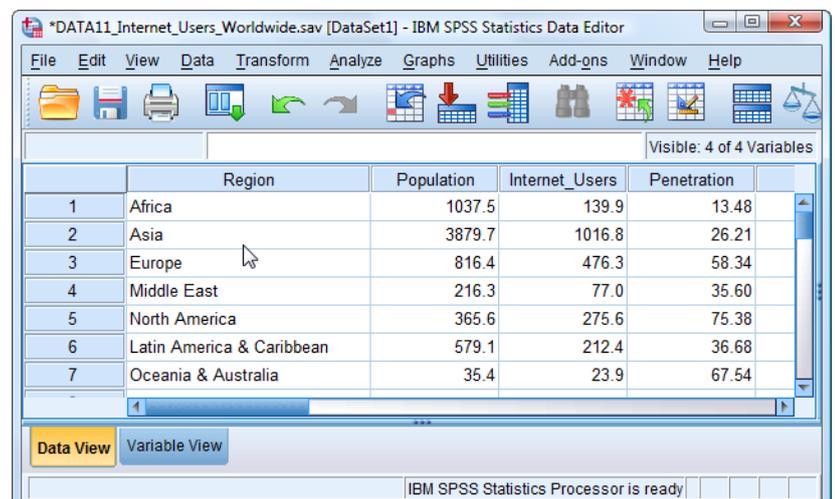
Target Variable: Penetration

Numeric Expression: Internet_Users / Population * 100

Function group: All

Functions and Special Variables:

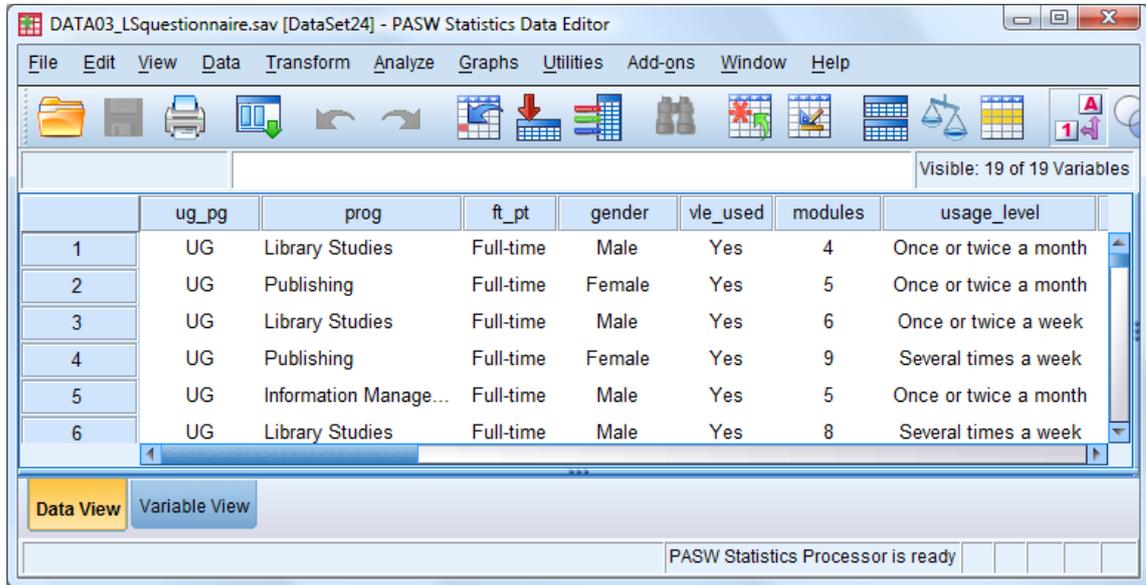
Buttons: OK, Paste, Reset, Cancel, Help



	Region	Population	Internet_Users	Penetration
1	Africa	1037.5	139.9	13.48
2	Asia	3879.7	1016.8	26.21
3	Europe	816.4	476.3	58.34
4	Middle East	216.3	77.0	35.60
5	North America	365.6	275.6	75.38
6	Latin America & Caribbean	579.1	212.4	36.68
7	Oceania & Australia	35.4	23.9	67.54

9.2 Recoding variables

The dataset partially shown here contains responses to a questionnaire given to 150 undergraduate and postgraduate Information Science students. One question asked was for how many modules (out of 12) did the students make use of the University's VLE. The answers given were recorded in a variable called **modules** (see below).



One question of interest is “Is there any gender difference?” A frequency table (below) has too many cells with small numbers in for a common statistical test called Chi-square to be valid.

No of modules accessed * Male or Female Crosstabulation

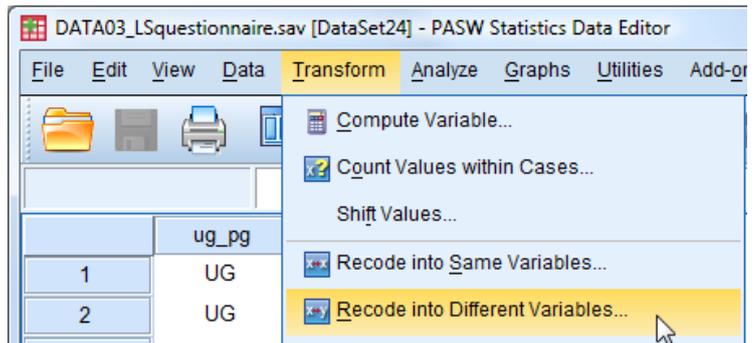
Count		Male or Female		Total
		Male	Female	
No of modules accessed	0	1	0	1
	1	2	0	2
	2	4	1	5
	3	4	4	8
	4	6	11	17
	5	6	15	21
	6	16	13	29
	7	15	10	25
	8	7	9	16
	9	10	1	11
	10	2	4	6
	11	2	3	5
	12	1	3	4
Total		76	74	150

In order to overcome this, the **Recode** command is needed to combine classes. The choice made is to reduce the 13 classes down to 7 classes (with approximately equal frequencies for each) as follows: 0 to 3 → 1, 4 → 2, 5 → 3, 6 → 4, 7 → 5, 8 → 6, 9 to 12 → 7.

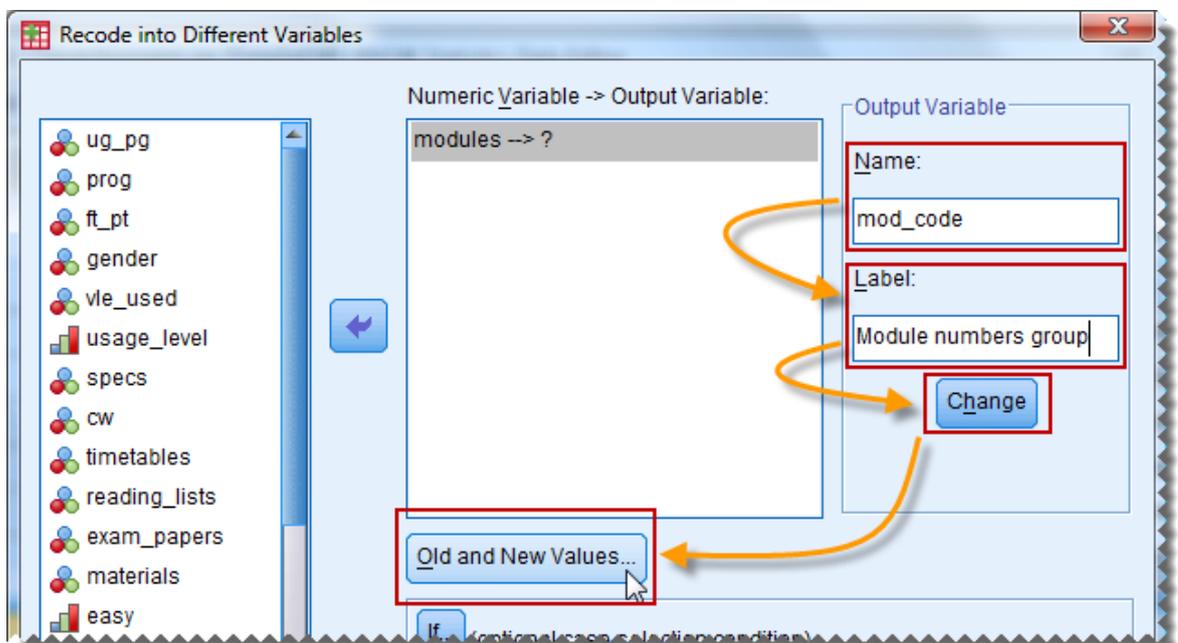
The procedure is quite lengthy and intricate, but commonly needed, so worth mastering.

A new variable will be created for this codification as follows. The command sequence

Transform → Recode into Different Variables...

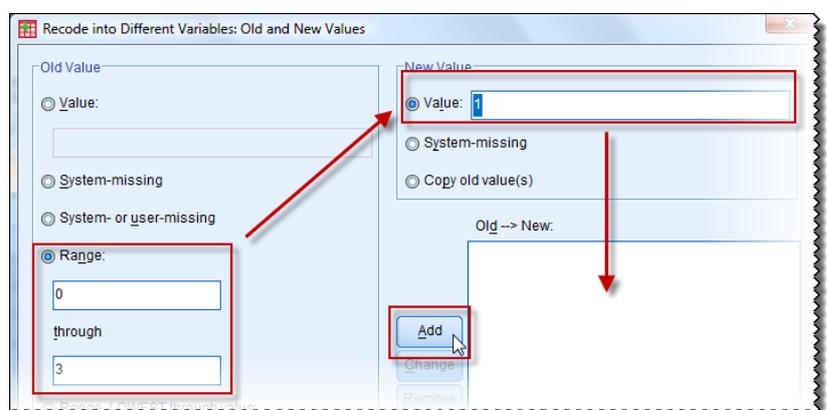


opens the **Recode into Different Variables** window (shown below). First the existing variable to be recoded must be passed from the list on the left into the **Input Variable → Output Variable** box on the right. Then a name for the new variable must be typed in (**mod_code**) and, optionally, a suitable label for it entered (**Module numbers group**). To confirm this **Change** is clicked.

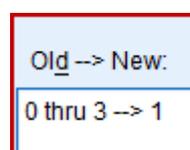


The next stage is to enter all the old and new codes to effect the recoding. To do this **Old and New Values...** is clicked. This opens up the **Recode into Different Variables** window.

To recode the range 0 to 3 → 1 the **Range** radio button is selected and the numbers 0 and 3 entered, and 1 is placed in the **New Value** box.

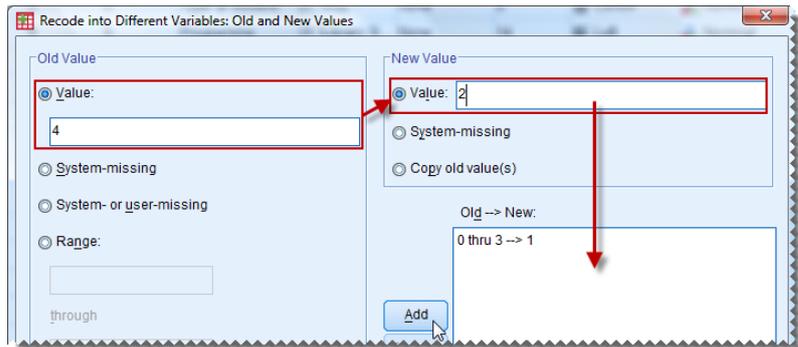


Once **Add** is clicked, the first recoding command is moved into the **Old → New** box and the entries are cleared.



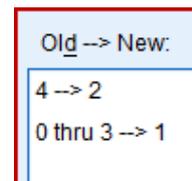
The next step is to recode 4 → 2.

This is done by selecting the **Value** radio button and entering 4 and then entering 2 in the **New Value** box.

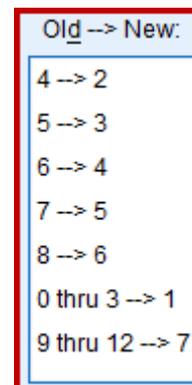


Once **Add** is clicked, the second recoding command is moved into the **Old → New** box and the **Old** and **New Value** entries are cleared.

In a similar way the recodings 5 → 3, 6 → 4, 7 → 5, 8 → 6 are performed, and recoding 9 to 12 → 7 can be done much as the first, specifying the range.



Note: There is no need to specify what the highest or lowest values are (they might not be readily known in a very large data set). Instead one could use the **Range, LOWEST through value** radio button for 0 → 3 and the **Range, value through HIGHEST** radio button for 9 → 12.



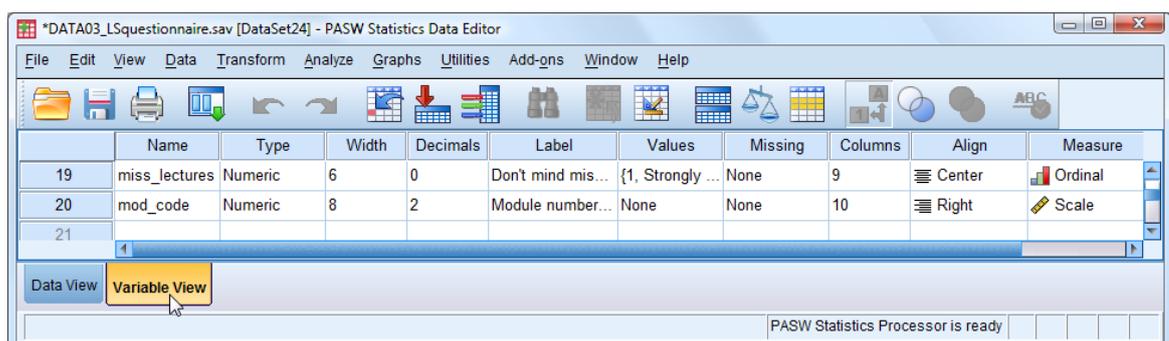
The complete **Old → New** list is shown here →

Finally, click **Continue** and **OK**.

This will produce a report in the **Viewer** Output window as below. Normally these are of limited interest but this one is worth taking notice of because it reports exactly what recoding has been done. (It is very easy to not do what one intended or to forget what one did.)

```
RECODE modules (4=2) (5=3) (6=4) (7=5) (8=6) (0 thru 3=1) (9 thru 12=7) INTO mod_code.
VARIABLE LABELS mod_code 'Module numbers group'.
EXECUTE.
```

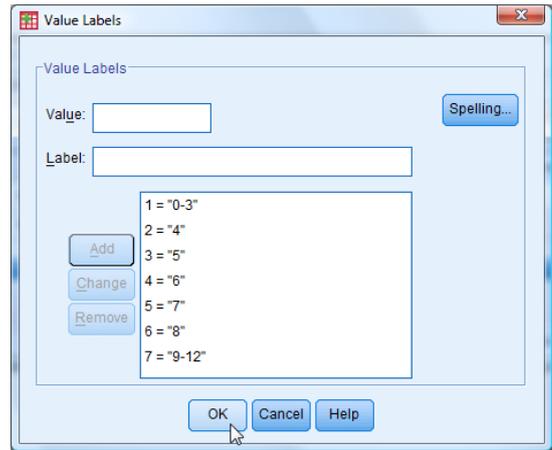
There is one final matter to attend to which is to check that the newly created variables' attributes are what they should be. SPSS provides defaults and it doesn't always get it right. It is necessary to change the **Data Editor** window to **Variable View**. Below is SPSS's defaults for **mod_code**:



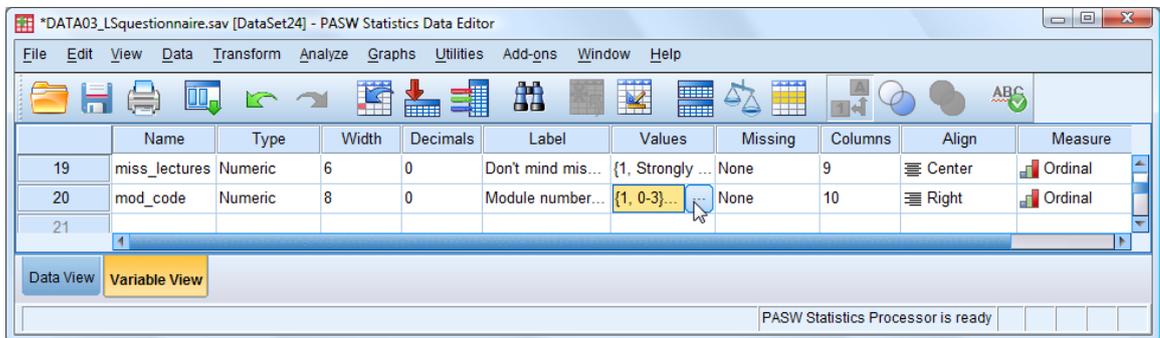
The following need changing:

- It is numeric but should have zero **Decimals**.
- It has no variable **Values** to explain the coding.
- Its **Measure** is shown as **Scale** but it should be **Ordinal**.

These can all be rectified. The appropriate **Value labels** window is shown here:



The revised **Variable View** is shown below.



Finally, the new frequency table is shown below for comparison with that obtained originally.

The data is now in a form which satisfies the conditions of the statistical test.

Module numbers grouped

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0-3	16	10.7	10.7	10.7
	4	17	11.3	11.3	22.0
	5	21	14.0	14.0	36.0
	6	29	19.3	19.3	55.3
	7	25	16.7	16.7	72.0
	8	16	10.7	10.7	82.7
	9-12	26	17.3	17.3	100.0
	Total	150	100.0	100.0	

10: Introduction to Charts and Graphs

10.1 General points about Charts and Graphs in SPSS

Graphs and charts take many forms and can be obtained in many different ways in *SPSS*. The two main procedures have huge choices of charts and ways to edit them. This can make it a frustrating process to find out how to get exactly what you want.

Here are some basic points:

- Numeric data can be used for nominal (categorical) variables, ordinal variables or scale (interval and ratio) variables.
- String data is only used for nominal variables.
- Different kinds of chart are appropriate for different types of variable (nominal, ordinal, scale).
- If you try to generate a chart not suitable for the type of variable chosen then a warning will be given. You can either choose another method or (sometimes) change the type (**Measure**) of the variable.

There are two fundamental ways to obtain charts and graphs in *SPSS*:

- **Graphs** → **Chart Builder...** The step-by-step drag-and-drop method introduced for recent *SPSS* versions.
- **Graphs** → **Legacy dialogs** A method used in earlier *SPSS* versions (hence the word 'legacy') which some users prefer.

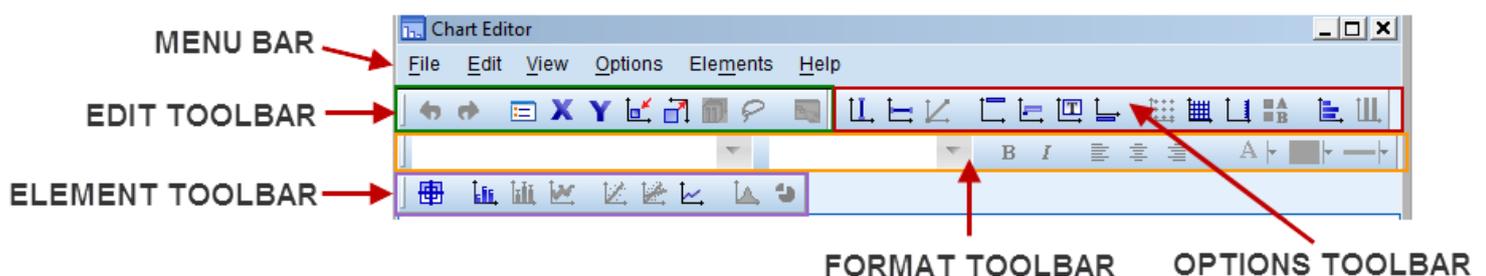
The Graphs and Charts TUTORIALS later in this Guide will concentrate on **Chart Builder...**

In the next five sections we introduce for reference the **Chart Editor** menu bar and its four associated toolbars for producing graphs and charts.

10.2 Chart Editor Menus and Toolbars

The **Chart Editor Menus and Toolbars** can only be described as complex and comprehensive. There is a great deal of overlap between what is available (a) through the **Menu** options, (b) selecting icons on the **Toolbars** and (c) by double-clicking on the regions and elements of the Chart itself. For reference, below are details of the Toolbars. The **Format** toolbar will be familiar to all Windows users.

CHART EDITOR MENU and TOOLBARS



The **Edit**, **Options** and **Elements** menus include all the items in the equivalent **Toolbars**, plus a few more. The advantage of using the **Menu Bar** rather than **Toolbars** is that the menus have descriptive words as well as icons. The **Toolbar** icons are, by default, always visible and quicker to use. The icons may not be so easy to recognise, but their descriptions appear when hovered over with the mouse cursor. **Toolbars** have an advantage over **Menus**, as the correct menu has to be opened to find the procedure sought.

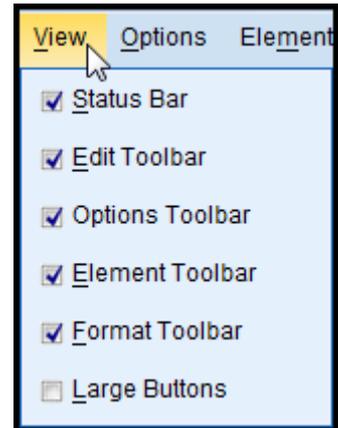
The positioning of the **Toolbars** will depend on the width selected for the **Chart Editor** window.

The **View** menu has options to turn on and off the **Toolbars**.

In addition, it controls the **Status Bar** which is at the bottom of the window and displays brief messages about the current activity.

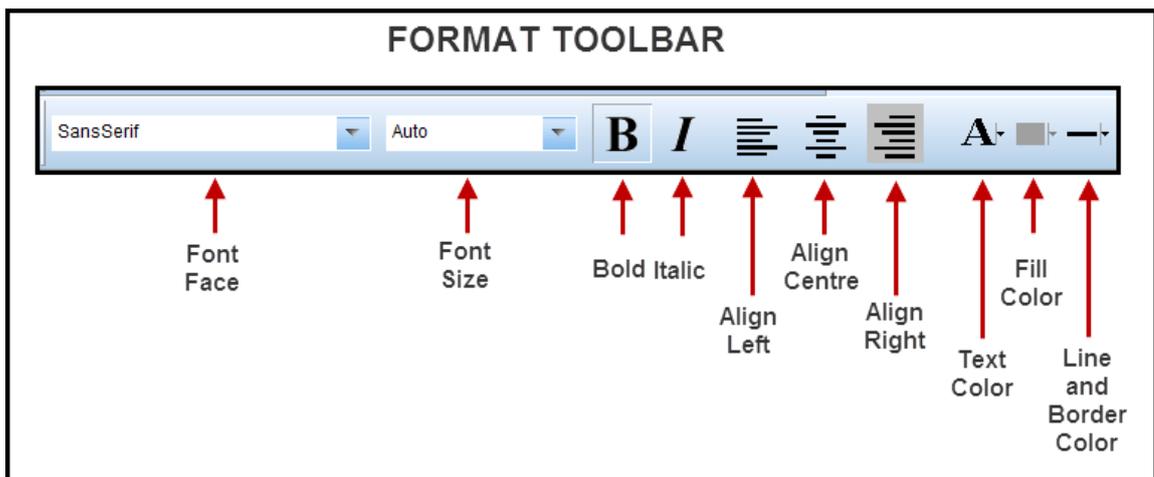
The last item in the **View** menu (**Large Buttons**) enlarges all the icons displayed in the selected **Toolbars** which can be a helpful aid to recognition.

On the next page, for reference, are the four **Toolbars** with annotations.



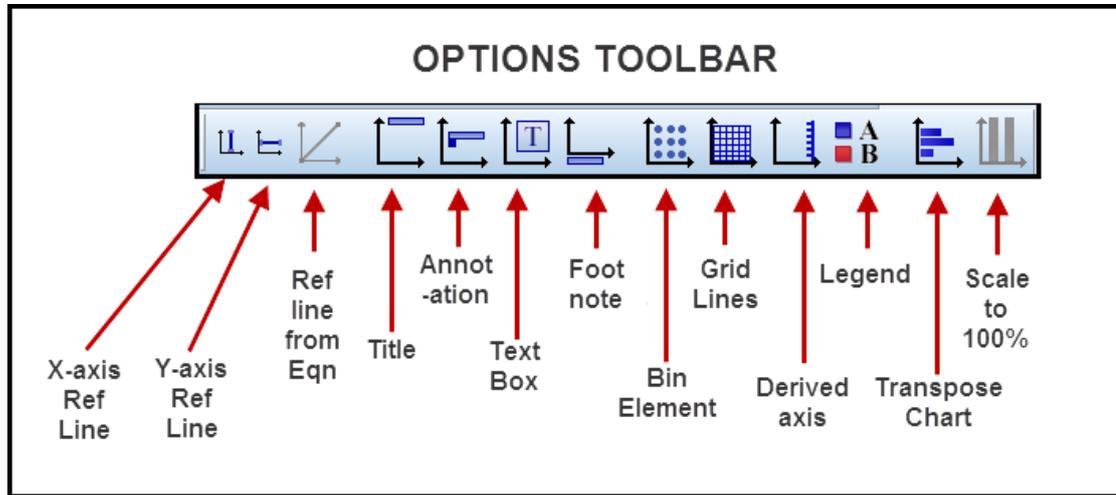
10.3 FORMAT Toolbar

To check what each **FORMAT** icon represents you can hover over each with the mouse to reveal a description of the action, as shown in the diagram below.



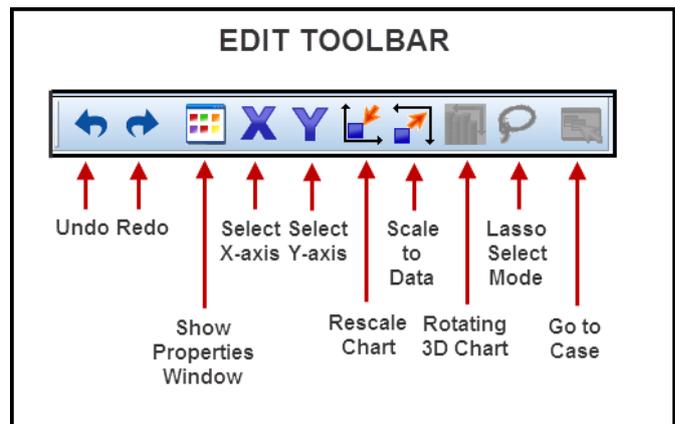
10.4 OPTIONS Toolbar

To check what each **OPTIONS** icon represents you can hover over each with the mouse to reveal a description of the action, as shown in the diagram below.



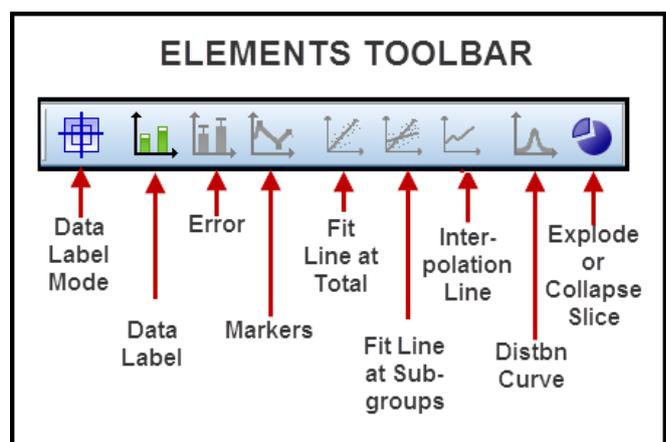
10.5 EDIT Toolbar

To check what each **EDIT** icon represents you can hover over each with the mouse to reveal a description of the action, as shown in the diagram here.



10.6 ELEMENTS Toolbar

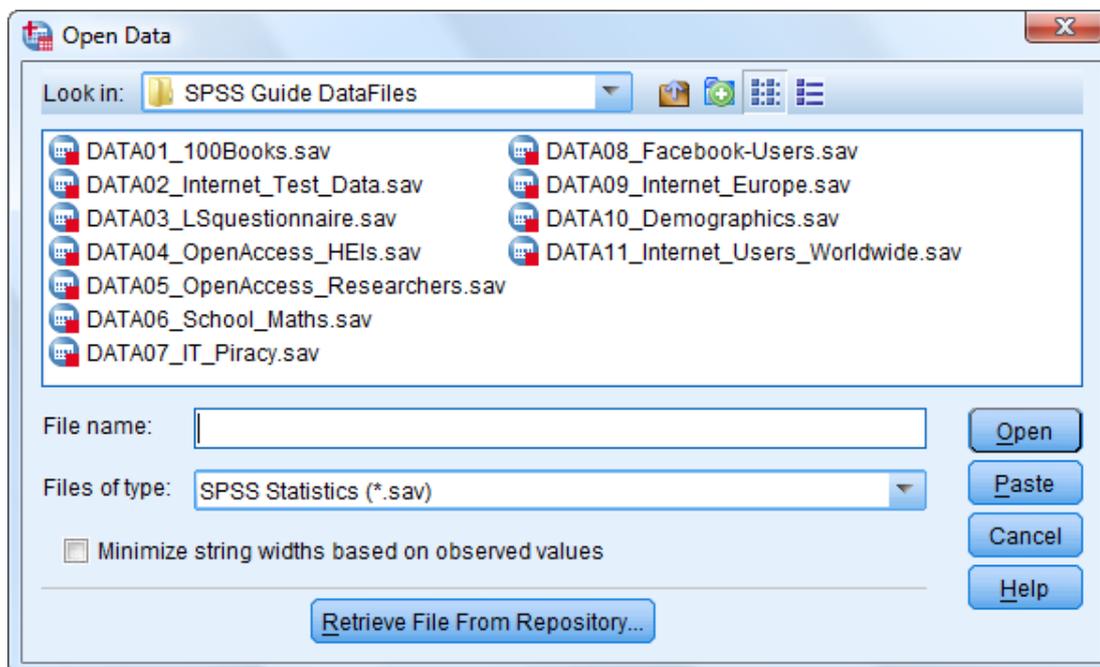
To check what each **ELEMENTS** icon represents you can hover over each with the mouse to reveal a description of the action, as shown in the diagram here.



11 Supplied SPSS data files for use with this Guide

There are several types of file that *SPSS* can use for storing and retrieving data. The most commonly used file type is the standard *SPSS* data file with filename extension *.sav*. Another common format (Microsoft *Excel*) will be introduced in a later section.

Eleven prepared *.sav* data files are available for working with this Guide, which should appear as:



The contents of these data files are indicated below. Further details are given in the Appendix.

Dataset	Description	Cases	Used in GUIDE	TUTORIALS or REFERENCE	APPENDIX QUESTIONS
DATA01	100 best-selling books 1989 to 2010 (Nielsen) - UK	100	✓	2, 3, 19, 20, 27, 28, 30	✓
DATA02	Internet usage by age / gender – Europe	10	✓	5	✗
DATA03	University students' responses to a VLE questionnaire - UK	150	✓	8-10, 12-18, 21, 24, 26-30	✓
DATA04	Open Access policies in HEIs - UK	39	✓	4, 6, 7, 8	✓
DATA05	Open Access practices of researchers - UK	418	✓	15	✓
DATA06	School students' attitudes to mathematics - UK	180	✓	29, 31-34	✓
DATA07	IT Piracy rates, Population, GDP and GNI - Worldwide	109	✓	22, 23	✓
DATA08	Facebook and Internet Users - Worldwide	157	✗	None	✓
DATA09	Internet Users - Europe	31	✗	None	✓
DATA10	Literacy rate, Population, Land area, GDP and GNI - Worldwide	155	✗	None	✓
DATA11	Internet Users by World Region - Worldwide	7	✓	Reference Section 9.1	✗

PART 2 - TUTORIALS

TUTORIAL T1: Starting the SPSS program

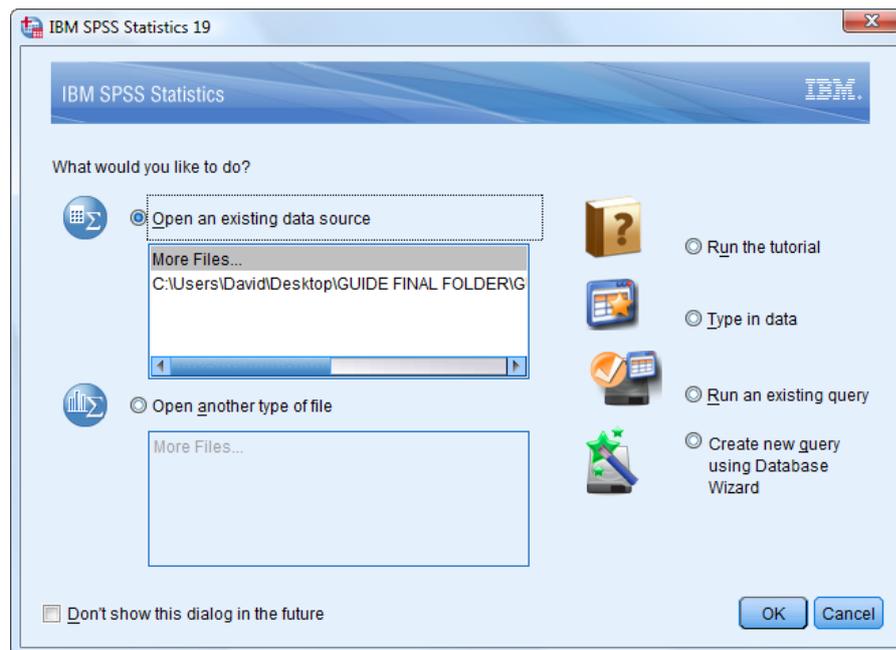
For instructions on how to load *SPSS* and locate data files on the computers you are using see separate documentation or consult the appropriate Appendix (if provided).

Follow the instructions below:

1. Switch on the PC and screen and log in as usual.
2. Click on the start icon located in the bottom left corner of the screen
3. Locate the list of programs (select **All Programs** if necessary) and find the *SPSS* program (e.g. IBM SPSS Statistics 19).
4. Click to start *SPSS*.



- ▶ A dialog box *will probably* appear:



- ▶ If the dialog box appears, it will present various button choices. In practice, users often find it easier not to use this dialog box and instead use the drop-down menu options from the main **Data Editor** window. (If that is preferred, then click in the 'Don't show this dialog in the future' box at the bottom of the window before closing the window.)

5. Click the dialog's close box.
 - ▶ The **Data Editor** window will appear, and all its menus and tools will be displayed.
 - ▶ At this point you have a choice: either enter data yourself (Reference Section 4 and Tutorial 4) or load a prepared data file (Reference Section 5, and all other Tutorials).

TUTORIAL T2: Loading a saved SPSS data file

1. If you are opening a data file on a USB memory device, insert it.

2. Click on **File** on the menu bar.

3. Click on **Open** and select **Data** on the drop down menu.

▶ An **Open Data** window will appear, displaying the contents of **Documents** on your computer which might contain SPSS data files.

▶ Note that the **Files of type** box contains the default: **SPSS Statistics (*.sav)**.

4. Click on the down arrow next to the **Look in** box, to gain access to all data on the computer and attached drives.

▶ This tree-structure list can be scrolled through to reveal the names of external drives and folders on your PC desktop and any files (not in folders) which are in **.sav** format).

5. Locate and select the folder required, which for this TUTORIAL is **SPSS Guide DataFiles** (unless otherwise advised).

▶ Double-click to reveal the list of files.

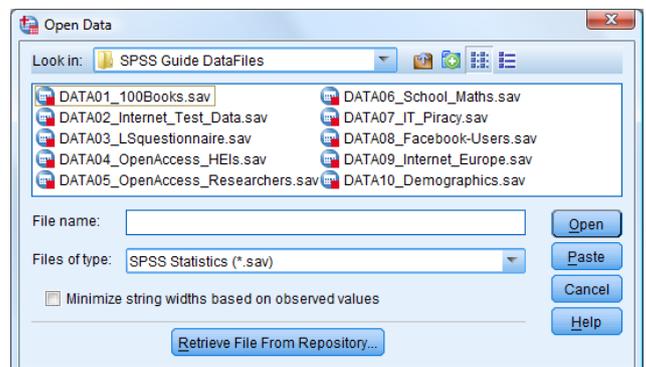
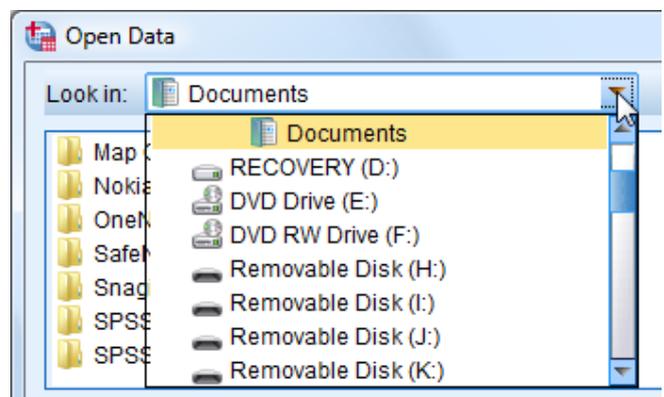
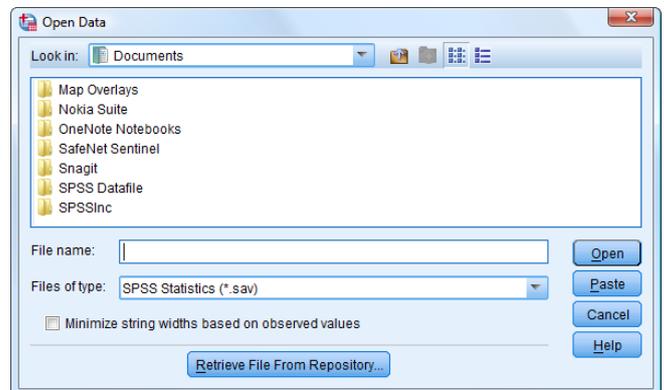
▶ A horizontal scroll bar will appear if there are too many entries to fit in the window.

6. Click on **DATA01_100Books.sav**.

▶ The name will appear in the **File name** field of the **Open Data** window.

7. Click on the **Open** button to import the file.

▶ The selected file will now be loaded into a **Data Editor** window, which may be in **Data View** mode (illustrated below) or **Variable View** mode.



	Position	Title	Author	
1	1	Da Vinci Code,The	Brown, Dan	Corgi Books
2	2	Harry Potter and the Philosopher's St...	Rowling, J. K.	Bloomsbury

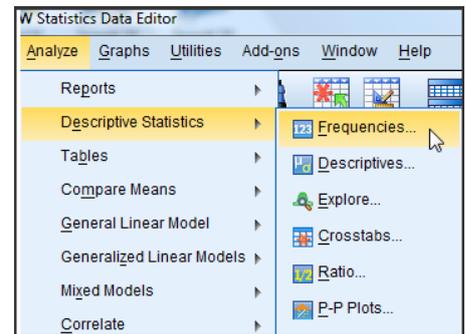
8. If necessary, change to **Data View** (by clicking the bottom left button). Scroll round the data set to view all the variables (columns) and to check that there are 100 cases (rows).

TUTORIAL T3: Analysing data using Frequencies

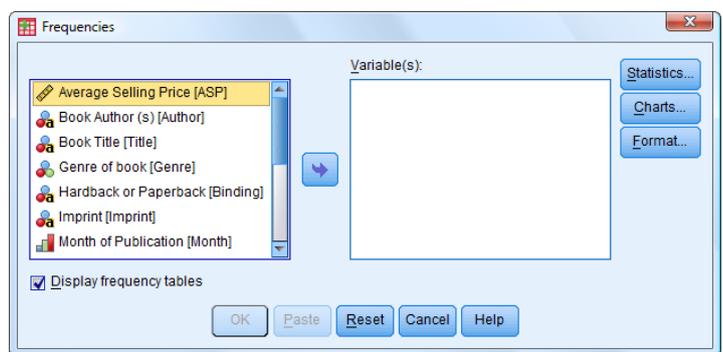
This uses the data file loaded in TUTORIAL T2.

1. To perform simple analyses to obtain frequency tables for some of the Top 100 books variables select

Analyze → Descriptive Statistics → Frequencies



- ▶ This opens up a **Frequencies** window listing all the variables which can be analysed (below).
- ▶ You can enlarge the window to reveal more information, and scroll also.



- ▶ Here the variables are shown in alphabetical order of variable label, with the variable name shown in square brackets.
- ▶ If your list is different then you need to change the display format.
- ▶ If your list is the same, you can skip to Step 2.

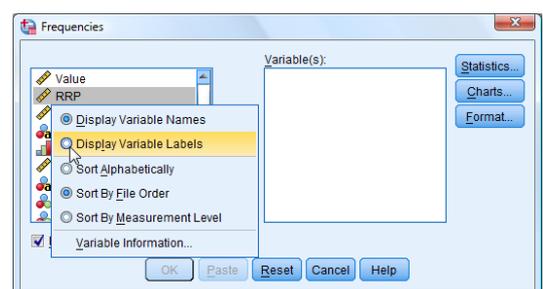
- ▶ The display format can be changed using **Edit → Options** and selecting the **General** tab and choosing **Variable Lists** options **Display labels** and **Alphabetical** → (see Reference Section 2.5.1).



- ▶ See below for an alternative method to achieve this, but which only affects the current list and is undone if **Reset** is clicked.

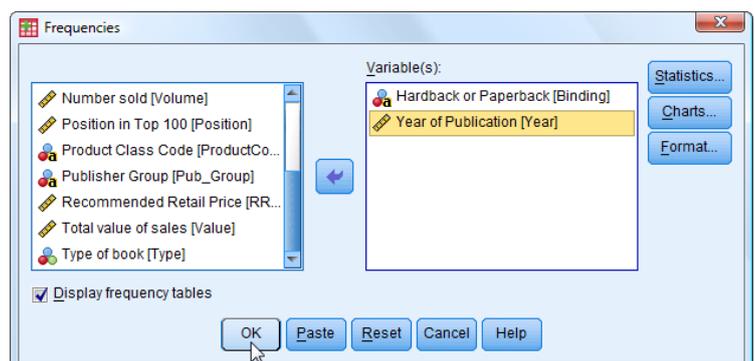
- ▶ To alter the format of the current displayed list right-click on the variable list itself and choose the display format you want →

Select **Display Variable Labels** and **Sort Alphabetically**.



2. Select **Hardback or Paperback [Binding]** and click on the right-pointing blue arrow to move it into the **Variable(s)** box for processing. Repeat for **Year of Publication [Year]**.

- ▶ This selects the two variables whose frequency tables are required →
- ▶ Note that the **Display frequency tables** box should be ticked → (if not, click it.)



3. Click **OK** to produce Output displayed in the **Viewer** window, part of which is shown below.

→ **Frequencies**

[DataSet3] C:\Users\David\Desktop\PASW DataFiles\DATA01_10

Statistics

		Hardback or Paperback	Year of Publication
N	Valid	100	100
	Missing	0	0

Frequency Table

Hardback or Paperback

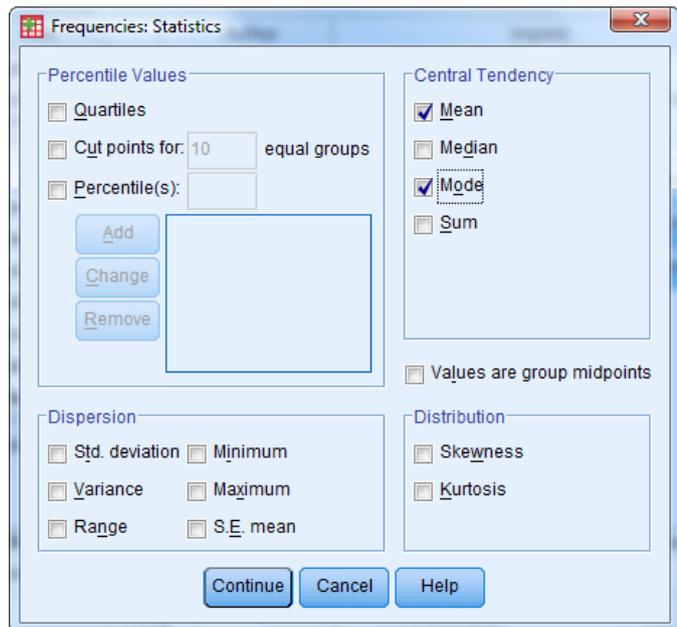
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Hardback	28	28.0	28.0	28.0
	Paperback	72	72.0	72.0	100.0
Total		100	100.0	100.0	

Year of Publication

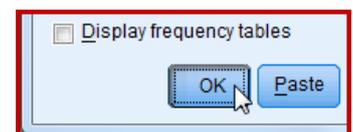
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1989	1	1.0	1.0	1.0
	1994	3	3.0	3.0	4.0
	1995	1	1.0	1.0	5.0

- ▶ Firstly, this shows that there were two variables analysed with 100 valid cases for each, and no data missing for either.
 - ▶ Secondly, this shows that there were 28 Hardbacks and 72 Paperbacks in the Top 100 list. This variable – **Binding** – takes two values which are simply names. It is a **nominal variable** producing **nominal data**.
 - ▶ Thirdly, this shows the Year of Publication numbers of Top 100 books (e.g. 3 in 1994). This variable – **Year** – takes many discrete values which have an order to them. It is an **ordinal variable** producing **ordinal data**.
4. We now return to the Frequencies procedure and explore the Statistics options. Select **Analyze** → **Descriptive Statistics** → **Frequencies** again.
 5. The same two variables should still be in the **Variable(s)** box on the right. Click **Reset** to remove them.
 6. Select **Recommended Retail Price [RRP]** and **Number Sold [Volume]**.
 - ▶ If necessary, you can alter the format of the current displayed list by right-clicking on the list and choosing the display format you want: **Display Variable Labels** and **Sort Alphabetically**.

7. Open the **Statistics** options window and select **Mean** and **Mode**.
8. Click **Continue**.



9. Deselect the **Display frequency tables** box.
10. Click **OK**.



- ▶ This generates further output in the same **Viewer** output window as before. Scroll down to see it.
- ▶ The second column shows that across the Top 100 books the mean RRP was £11.08 and the most common RRP was £7.99 (it actually occurred 49 times!).
- ▶ The third column shows that the mean number of copies sold was 1.17 million.

Statistics

		Recommended Retail Price	Number sold
N	Valid	100	100
	Missing	0	0
Mean		£11.0812	1,169,799.01
Mode		£7.99	643,636 ^a

a. Multiple modes exist. The smallest value is shown

- ▶ Note the warning below the Statistics box stating that there are other modes for **Number sold**. That is because each book sold a different number of copies so every 'sold' number has frequency 1 and therefore they are *all* technically modes (the most common). So here 'mode' is a useless statistic! Mode is only of interest when there is one or two of them.

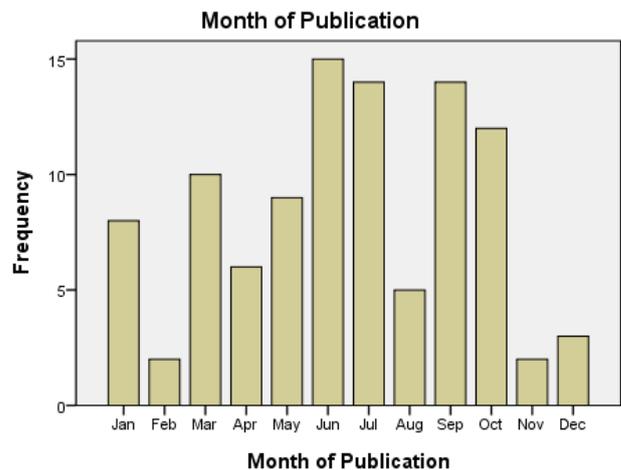
11. We return to the **Frequencies** procedure and explore its **Chart** options.

Select **Analyze** → **Descriptive Statistics** → **Frequencies** again.

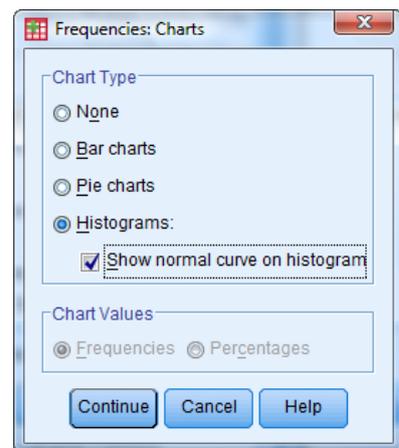
Click **Reset** to remove any selected variables and select **Month of Publication [Month]**.

12. Open the **Charts** option and click **Bar charts**. Leave **Chart Values** as **Frequencies**.
13. Click **Continue**.
14. Click **OK** to obtain the chart in the Output window (scroll down to see it, if necessary).

- ▶ The bar chart shown here is obtained. Note that the bars are all separate – because the categories (months) are distinct entities (discrete) which have an obvious order. This is called **ordinal** data.
- ▶ The chart shows that winter months are generally less popular. This could be investigated further.



15. We return once again to the **Frequencies** procedure and explore **Chart** options further. Select **Analyze** → **Descriptive Statistics** → **Frequencies** again. Click **Reset** to remove any selected variables and select **Number Sold [Volume]**.

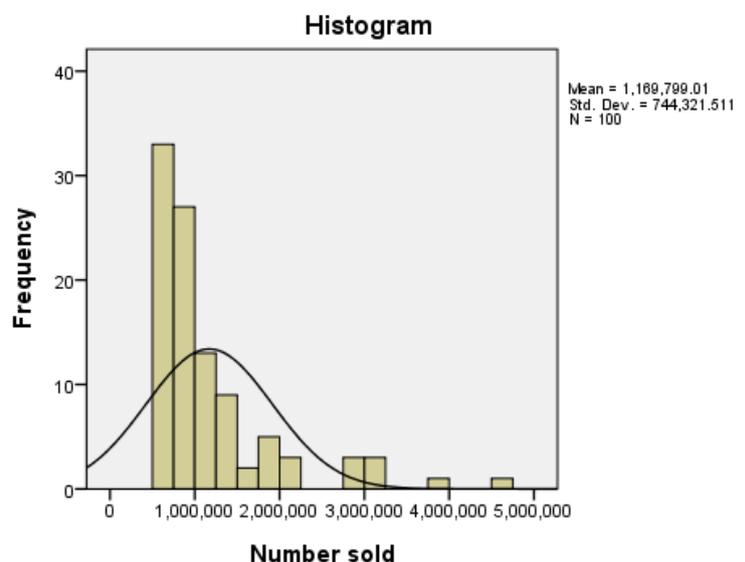


16. Open **Charts** and select **Histograms**. Also select **Show normal curve on histogram**.

17. Deselect the **Display frequency tables** box and then click **OK**.

18. Click **Continue** and then click **OK** to obtain the chart.

- ▶ The histogram shown here is obtained. Note that the bars are contiguous not separated. This is because the variable **Number sold** can take very many different values, and is considered to be a *continuous* rather than a *discrete* variable. In *SPSS* this is called a **scale** variable which gives rise to **scale** data. Each bar represents a range of values.
- ▶ The superimposed normal curve clearly shows that these data are *not* normally distributed. The distribution is said to be positively skewed.



The above has provided a short introduction to one of the simpler *SPSS* procedures. Even so, there is has been a lot to learn. And *SPSS* has many other procedures: **Analyze** is just one of four main procedural menus; the **Analyze** menu has 22 menu options (of which about seven are likely to be of use to most Information Scientists); of these seven, the **Descriptive Statistics** menu option itself has seven menu options (of which three are mostly used).

Fortunately, learning to use **Frequencies** provides a good basis for using many other procedures.

It is a good idea to close the Output window so that a new one is created for the next Tutorial.

TUTORIAL T4: Creating a new data file – inputting data

These instructions are for creating a new data file in *SPSS* to store a new set of data (e.g. entering responses to a questionnaire, entering details of electronic journals subscribed to).

This Tutorial can be skipped if you are not interested in creating your own data files.

New data files are created by keying data into a new **Data Editor** window whose title will begin **Untitled...**

T4.1 Defining the variables

Below is a small amount of real data selected from the supplied data file **DATA04** containing the results of a survey of UK HEIs about their Open Access Policies. The full dataset has 39 cases with 26 variables and will be used later in this Guide.

The data here, which has 5 cases and 5 variables, is for you to practise entering and formatting data. The five cases are:

Case	INSTITUTION	SECTOR	Q1	Q8c	Q9e
1	02	1	1	2385	£77000
2	03	2	3	464	£27000
3	13	2	4		
4	19	4	4	0	£9000
5	32	3	4	5	£0

There are two systematic ways you can enter the above data in **Data View**:

- (1) You can go across entering data from variable to variable.
- (2) You can go down entering data from case to case.

It is a matter of preference. Provided you have already entered the variables information first, the advantage of going across is that you can use **Tab** to move from cell to cell and *SPSS* will automatically take you to the first cell of the next case when you have finished entering all the data for one case.

The five variables are:

Variable	Name	Description / Question	Values
1	INSTITUTION	A string code to identify the individual HEI.	A string (digits)
2	SECTOR	A numeric code identifying the type of HEI.	1 = RLUK member 2 = Other Pre-1992 university 3 = Post-1992 university 4 = HE College
3	Q1	“Does your HEI have an Open Access written policy?”	1 = Yes 2 = No – planned 3 = No – rejected 4 = No – not considered 5 = Don’t know
4	Q8c	“How many Open Access items are currently held?”	A number.
5	Q9e	“What are the annual running costs?”	A number in £.

There are two systematic ways you can enter the above variable information in **Variable View**:

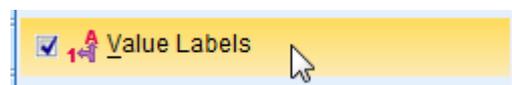
- (1) You can go across entering each variable name and its attributes and then move onto the next variable.
- (2) You can go down entering all the variable names and then go back and edit all the attributes.

It is a matter of preference, which may depend on how many cases there are, how many variables there are, and how much variable information has to be added.

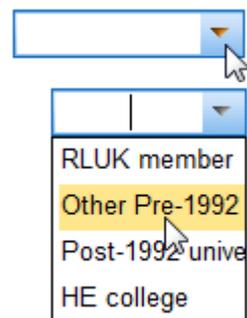
In this TUTORIAL we use case by case entry for data and variable by variable entry for variable information.

It is possible to enter the data before entering all the variable information (*SPSS* will use dummy variable names **VAR0001** etc. which you can later edit) but it is preferable to enter the variable information first – especially if there are any **Values** labels.

The reason is that when you come to enter the data in a cell, and have **Value labels** selected in the **View** menu →



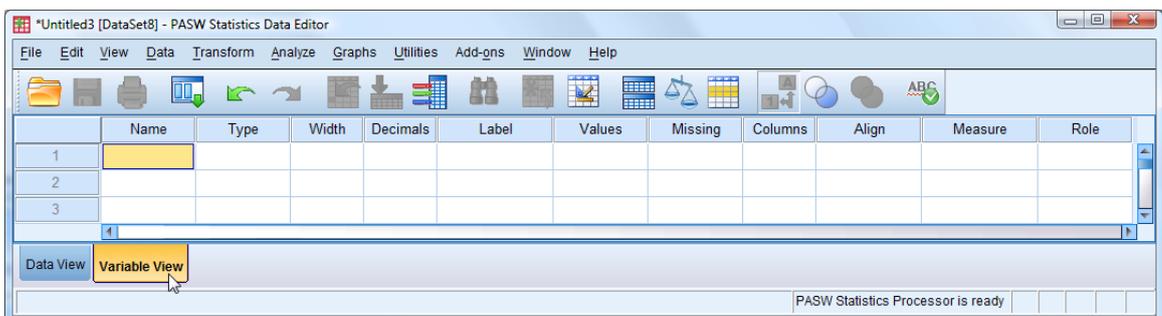
then when you click (or double-click) in the cell a drop-down menu will be made available (by clicking on the down arrow) →



and you can select the entry you want rather than needing to type it in →

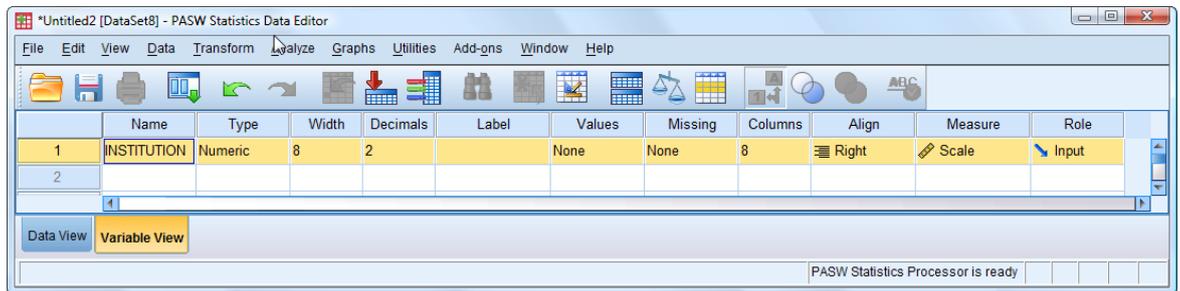
1. If the **Data Editor** window is already visible and is empty (as it will be if just starting *SPSS*) then you can proceed straight to step 5. Otherwise, first create a file as follows:
2. Load *SPSS* and click on **F**ile on the menu bar.
3. Click on **N**ew from the drop down menu.
4. Click on **D**ata from the drop down menu.

► A **Data Editor** window with title beginning **Untitled** will appear (see below).



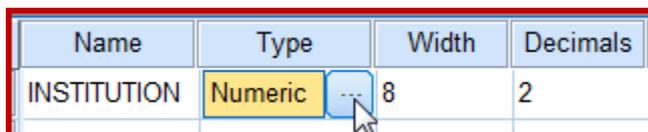
5. Make sure you are in **Variable View** by clicking the button at the bottom left corner.

- In row 1 type in the first variable name **INSTITUTION** into the first cell of the column headed **Name** and press **Enter** or **→** or **Tab** on your keyboard.

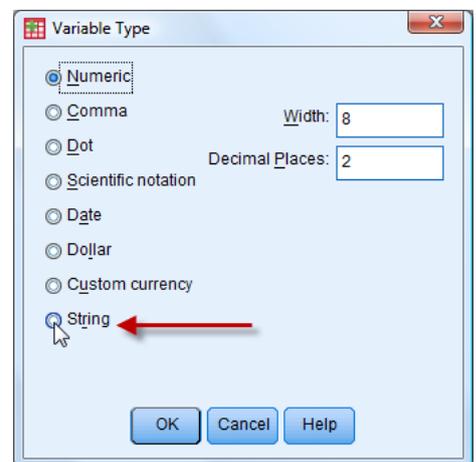


- ▶ SPSS has filled in defaults for many of the variable's attributes. Most are OK but some of these need changing or more information inserted.
- ▶ **INSTITUTION's Type** is set as 'Numeric' and is to be changed to 'String'.

- Click in **INSTITUTION's Type** cell and click on the dots.



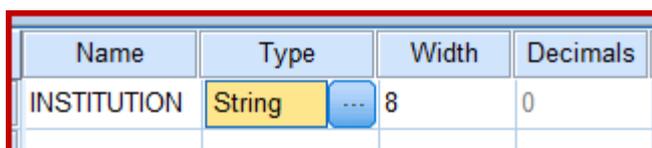
- ▶ The **Variable Type** dialog box opens →



- Select 'String'.

- Click **OK**.

- ▶ SPSS automatically sets **INSTITUTION's Decimals** cell entry to '0' as it is a string variable:



- Click in **INSTITUTION's Label** cell and type in 'Identifier code'.

- ▶ SPSS automatically sets **INSTITUTION's Measure** cell entry to →



- ▶ All the other SPSS default entries for **INSTITUTION** are OK.
- ▶ The information for the remaining variables must now be entered in a similar way.

- In row 2 type in the **Name** cell 'SECTOR'.

- Change the **Type** cell to 'String'.

- Type in the **Label** cell 'HEI group'.

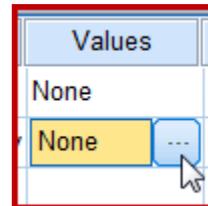
- Set the **Align** cell to 'Center'.

15. Insert **Values** as in this table →

Do this as follows.

Click on the **Values** cell then click on the three dots which appear:

- 1 = RLUK member
- 2 = Other Pre-1992 university
- 3 = Post-1992 university
- 4 = HE college



► This will open a **Value Labels** window.

16. Insert '1' for the **Value** and 'RLUK member' for the **Label** and click **Add**.

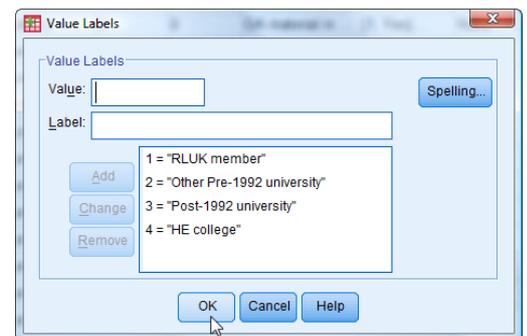
► This inserts the value label as shown here →



17. Insert '2' for the **Value** and 'Other Pre-1992 university' for the **Label** and click **Add**.

18. Repeat the process for the other two labels and click **OK**.

► Clicking **OK** confirms all four value labels as here →

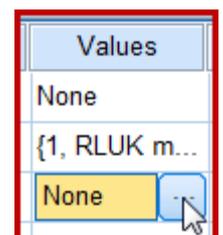


19. In row 3 type in the **Name** 'Q1' and set the **Decimals** to '0' and type in the **Label** 'Written OA policy', set **Align** to 'Left' and set the **Measure** to 'Nominal'.

20. Insert **Values** as in this table.

To do this click on the **Values** cell then click on the three dots which appear:

- 1 = Yes
- 2 = No – planned
- 3 = No – rejected
- 4 = No – not considered
- 5 = Don't know



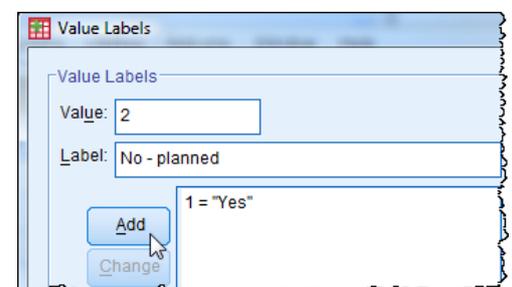
► This will open a **Value Labels** window.

21. Insert '1' for the **Value** and 'Yes' for the **Label**.

22. Click **Add**.

23. Insert '2' for the **Value** and 'No - planned' for the **Label**.

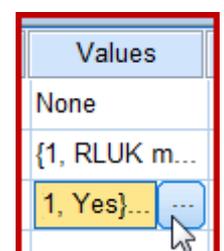
► At this point the window will be as shown here →



24. Click **Add**, repeat the process for the other three labels, then click **OK**.

► This inserts all the value labels into the data file.

► The **Values** entry for **Q1** in the **Data Editor** will now show the first entry (all the value labels can be seen by clicking on the three dots).

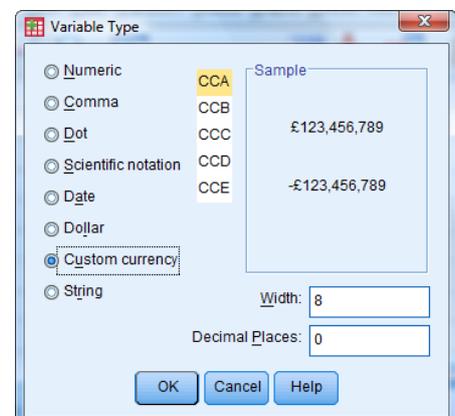


25. In row 4 type in the **Name** 'Q8c'.
26. Change **Type** from 'Numeric' to 'Comma', and set **Decimals** to '0'.
 - ▶ 'Comma' is the same as 'Numeric' but inserts commas to mark off every three digits before the decimal point, to aid legibility.
 - ▶ Setting the number of decimal places can be done within the **Variable Type** dialog box at the time 'Comma' is selected, or done separately.
27. Type in the **Label** 'Total items held (June 2008)'.
28. Change **Measure** to 'Scale'.
 - ▶ All the other defaults are OK.
29. For the next variable we will want currency in £ not \$.

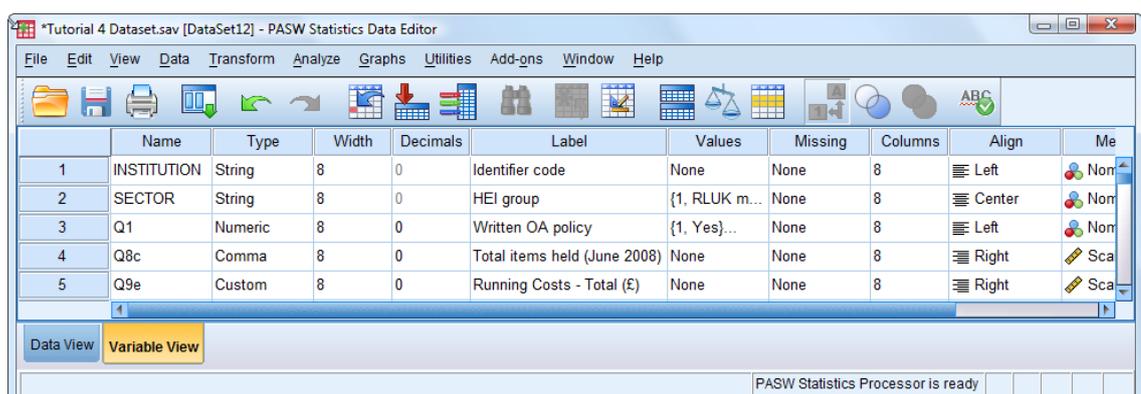
To do this use **Edit** → **Options** → **Currency** and select **CCA** and enter '£' in the prefix box (if not already there).

- ▶ This establishes the Custom Currency type CCA as £ which can be selected as required.

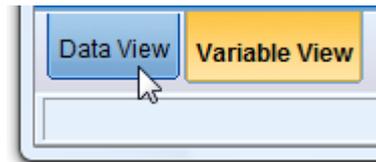
30. Click **Apply**.
31. Click **OK**.
32. In row 5 type in the **Name** 'Q9e'.
33. Because this is to be currency in £ (the default is \$US) change the **Type** to 'Custom currency' and choose **CCA**, and set the **Decimal Places** to '0'.



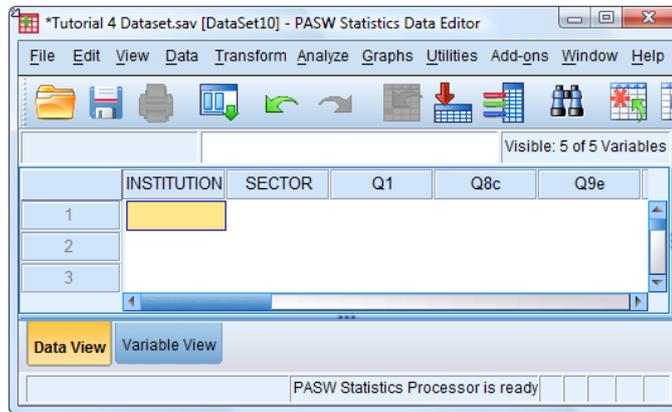
34. Click **OK**.
 - ▶ This will insert a £ sign before each value and commas marking off every three digits before the decimal point.
35. Type in the **Label** 'Running Costs – Total (£)'.
36. Change **Measure** to 'Scale'.
 - ▶ All the other defaults are OK.
 - ▶ The **Data Editor** at this stage should look like this:



37. Now the data itself must be entered. Begin by switching to **Data View**. Do this by clicking in the bottom left corner of the **Data Editor** window.



- This presents an empty array of cells with the five variable headings at the top.



- Before proceeding to entering the data, a summary of ways to move around the **Data Editor** window is presented, for easy reference.

T4.2 Moving around the Data Editor window

Moving around the Data Editor window

To move around the **Data Editor** window in **Data View** or **Variable View** modes:

- **Tab** press moves the cursor from one data cell to the next cell to the right until the last variable is reached at which point the next TAB press moves the cursor to the first cell of the next row. This is particularly useful when in **Data View** for entering data case by case.
- Vertical and horizontal scroll bars are located at the right hand side and bottom and can be used to move around the window.
- ← → ↑ ↓ keys can be used to move from cell to cell, instead of 'point and click'.
- **Control + Home** key combination takes you to the first cell (top left) in the window.
- **Control + End** key combination takes you to the last cell (bottom right) in the window.

To move around the **Data Editor** window in **Data View** mode only:

- **Control + ↑** key combination takes you to the first cell in the current column.
- **Control + ↓** key combination takes you to the last cell in the current column.
- **Control + ←** key combination takes you to the first cell in the current row.
- **Control + →** key combination takes you to the last cell in the current row.

T4.3 Entering and Saving the data

34. Enter the data shown in this table, but first read the notes below.

INSTITUTION	SECTOR	Q1	Q8c	Q9e
02	1	1	2385	£77000
03	2	3	464	£27000
13	2	4		
19	4	4	0	£9000
32	3	4	5	£0

- ▶ First deselect **Value Labels** so that the actual (numeric) codes rather than the corresponding labels are shown. Do this by opening the **View** menu and making sure **Value Labels** does not have a tick. Do that now.



- ▶ Note that there are some blanks and some zeros.
- ▶ Do not enter the £ signs (they are provided by the **Custom Currency Type** attribute).
- ▶ The best way to enter the data is to type in the first entry then press Tab to move across to the next cell. When at the end of the row Tab automatically takes you onto the first cell in the next line.

- ▶ Enter all the data now.
- ▶ What you should get is something very similar to this →

 A screenshot of the SPSS Data Editor window. The data table is displayed with columns: INSTITUTION, SECTOR, Q1, Q8c, and Q9e. The rows correspond to the data in the table above. The Q1 column has values 1, 3, 4, 4, 4. The Q8c column has values 2,385, 464, ., 0. The Q9e column has values £77,000, £27,000, ., £9,000, £0. A dot is used for missing values.

- ▶ Note that SPSS has inserted a dot where a value was missing. The dot is the SPSS **missing values** symbol.
- ▶ The alignment for the variables **INSTITUTION** and **SECTOR** could be better centred. SPSS's default is **Left** alignment for **String** and **Right** alignment for everything else.
- ▶ If **View** → **Value Labels** is selected the value labels and not the numeric codes will be shown (try it).

 A screenshot of the SPSS Data Editor window showing the same data table as above, but with value labels for the SECTOR and Q1 variables. For example, SECTOR 1 is labeled 'RLUK me...', SECTOR 2 is 'Other Pre-...', and Q1 1 is 'Yes'. The 'Data View' tab is selected at the bottom.

35. Save the data using **File** → **Save As** and type in a name of your choice to a destination of your choice. (See Reference Section 6 for more details.)

T4.4 Analyzing the entered data

When data has been entered it needs to be checked. For a large data set this is an important and considerable task. SPSS has a special procedure (**Case Summaries**) designed for this which will be introduced in TUTORIAL T5. For a small data set various other SPSS procedures can be just as – or even more – effective in looking for mistakes, such as wrong values, missing values, typing errors in labels and names.

Entering this small data set you probably have made no errors. As a simple check we explore the data using the **Frequencies** procedures (met in TUTORIAL T3).

We try to find the frequency tables, mean values and bar charts for all five variables all in one go, proceeding as follows (you can skip step 1 if continuing straight on from T4.3).

1. Load data file: **File** → **Open** → **Data** → **your_file_name.sav**
2. Select **Analyze** → **Descriptive Statistics** → **Frequencies**.
3. Click **Reset** to remove any selected variables (if necessary).
4. Move all five variables into the **Variable(s)** box.
5. Open the **Statistics** options window and select **Mean**, **Median**, **Mode**, **Minimum**, **Maximum**.
6. Click **Continue**.
7. Click **OK** to obtain the output.

► Just the **Statistics** table is shown here:

		Identifier code	HEI group	Written OA policy	Total items held (June 2008)	Running Costs - Total (£)
N	Valid	5	5	5	4	4
	Missing	0	0	0	1	1
Mean				3.20	713.50	£28,250.00
Median				4.00	234.50	£18,000.00
Mode				4	0 ^a	£0 ^a
Minimum				1	0	£0
Maximum				4	2,385	£77,000

a. Multiple modes exist. The smallest value is shown

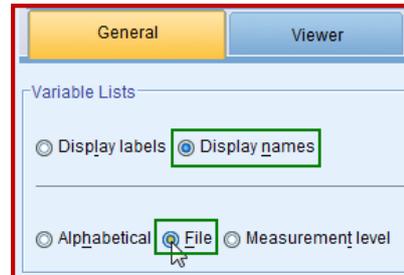
- Note that when there is missing data it is excluded from calculations (for the mean etc.).
- Note the difference when the value is not missing but is zero – it is included in calculations.
- Note that the mean is calculated for **Q1** [Written OA policy] – although it is meaningless, because **Q1** is defined as a **Numeric** variable.
- Note that the mean is not calculated for **INSTITUTION** [Identifier code] – because it is defined as a **String** variable.
- Note that the variable labels and not the variable names are shown in this table. This is a matter of choice, controlled by **Edit** → **Options** – see Section 2.5.4 on page 9.

T4.5 Analyzing the complete data file

It would be a long and tedious and rather pointless task to enter all the data for **DATA04_OpenAccess_HEIs** with its 39 cases each with 26 variables! Fortunately you are spared this as the dataset is provided for use with this Guide. Here you are encouraged to repeat the same analysis as was done in T4.4, this time on the full dataset.

1. Load data file: **File** → **Open** → **Data** → **DATA04_OpenAccess_HEIs.sav**

2. Select **Edit** → **Options** and click on the **General** tab (if not already highlighted) and in the **Variable Lists** section click on **Display names** and **File**. This will ensure that the variables are listed in the order they appear in the **Data Editor** and their names (not their labels) are shown.



3. Click **Apply**.

4. Click **OK**.

5. Select **Analyze** → **Descriptive Statistics** → **Frequencies**.

6. Locate and move **INSTITUTION** into the **Variable(s)** box.

7. Locate and move **SECTOR** into the **Variable(s)** box.

8. Locate and move **Q1** into the **Variable(s)** box.

9. Locate and move **Q8c** into the **Variable(s)** box.

10. Locate and move **Q9e** into the **Variable(s)** box.

11. Open the **Statistics** options window and select **Mean**, **Median**, **Mode**, **Minimum** and **Maximum**.

12. Click **Continue**.

13. Deselect **Display frequency tables** in the **Frequencies** window.

14. Click **OK** to obtain the output below:

Statistics

	Identifier code	HEI group	Written OA policy	Total items held (June 2008)	Running costs - Total (£)
N Valid	39	39	39	21	13
Missing	0	0	0	18	26
Mean		2.13	2.59	2047.05	£27,038.46
Median		2.00	2.00	1089.00	£20,000.00
Mode		3	4	0 ^a	£0
Minimum		1	1	0	£0
Maximum		4	4	9800	£77,000

a. Multiple modes exist. The smallest value is shown

TUTORIAL T5: Checking Data - using Case Summaries

This is very much an *optional* TUTORIAL which can be returned to at later time.

It is important to check that data is entered correctly. Any of the various procedures which produce tables can be used. For example:

(a) **Analyze** → **Descriptive Statistics** → **Frequencies**

(b) **Analyze** → **Descriptive Statistics** → **Descriptives**

(c) **Analyze** → **Descriptive Statistics** → **Explore**

(d) **Analyze** → **Tables** → **Custom Tables**

However, there is a special procedure for the purpose:

(e) **Analyze** → **Reports** → **Case Summaries**

In this TUTORIAL you will load a specially constructed small data file and generate a list of values for variables to look for anomalies.

1. Load data file: **File** → **Open** → **Data** → **DATA02_Internet_Test_Data.sav**
2. Select **Analyze** → **Reports** → **Case Summaries** to open the **Summarize Cases** window.

3. Transfer all the variables from the left box to the **Variables** box using the blue arrow (click the top variable and shift-click the bottom variable to do it quickly all in one go).

► If the list of variables is displayed differently then, if you wish, select **Edit** → **Options**, click on the **General** tab and in the **Variable Lists** section click on **Display names** and **File**, and click **OK**.

4. Deselect **Limit cases to first** removing the tick.

► Note: Doing this is not essential here as we only have 10 cases, but it will eliminate an annoying footnote in the output!

5. Deselect **Show only valid cases**.

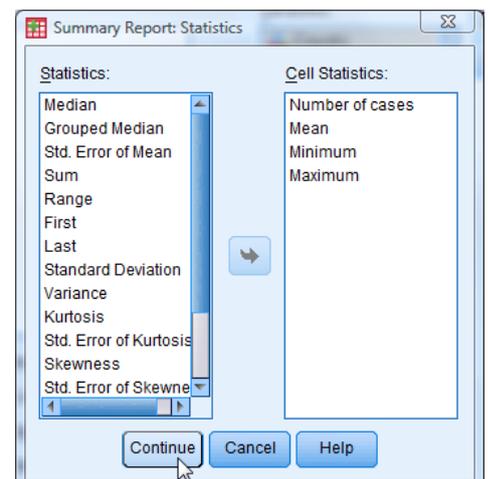
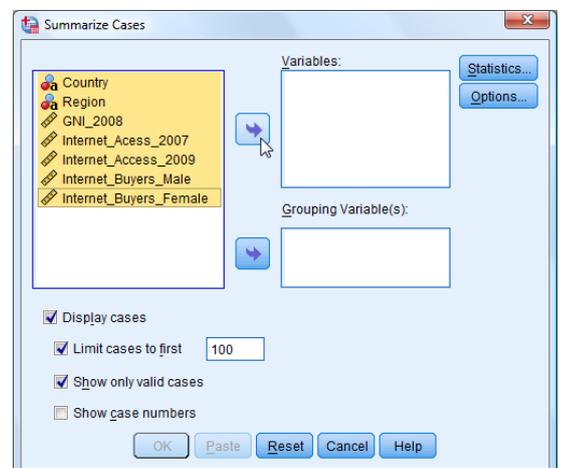
6. Select **Show case numbers**.

7. Open the **Statistics...** window.

► Note that 'Number of cases' will be in the **Cell Statistics** box already.

8. Move '**Mean**', '**Minimum**' and '**Maximum**' into the **Cell Statistics** box using the blue arrow.

9. Click **Continue**.



10. Click **OK** to produce two tables in the **Viewer** Output window, shown below:

- ▶ This shows that for some reason the fourth variable is excluded from two cases.

Case Processing Summary						
	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
European Sovereign State	10	100.0%	0	.0%	10	100.0%
European Region	10	100.0%	0	.0%	10	100.0%
Gross National Income per annum per capita (US\$)	10	100.0%	0	.0%	10	100.0%
Households with Internet Access 2007 (%)	8	80.0%	2	20.0%	10	100.0%

- ▶ The table below (shown partially) lists the values of all seven selected variables for all 10 cases. The bottom of the table provides the statistics asked for (Number, Mean, Min, Max).

Case Summaries								
	Case Number	European Sovereign State	European Region	Gross National Income per annum per capita (US\$)	Households with Internet Access 2007 (%)	Households with Internet Access 2009 (%)	Males 16-74 who have bought over internet in 2009 (%)	Females 16-74 who have bought over internet in 2009 (%)
1	1	Austria	Western Europe	\$37,680.00	60	7	46	36
2	2	Belgium	Western Europe	\$34,760.00	60	67	41	30
Total	N	10	10	10	8	10	10	10
	Minimum	Austria	Eastern Europe	\$359.40	41	7	5	4
	Maximum	Germany	Western Europe	\$356,660.00	78	83	111	61
	Mean			\$57,357.9400	60.13	51.40	47.20	32.70

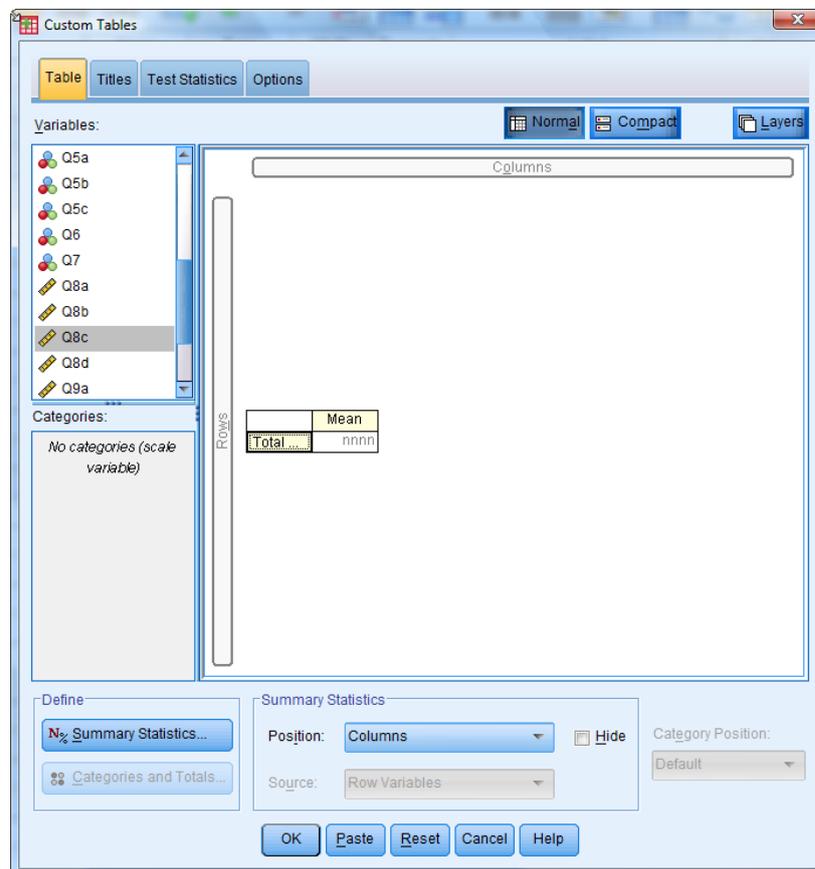
- ▶ The Number, Mean, Min, Max statistics – together with applied common sense – are very useful for spotting anomalies and errors, but will not find them all, depending on their nature.
- ▶ The reader should do some detective work looking at the full output to see what errors there appear to be, in conjunction with viewing the actual dataset.
- ▶ Finally look at the notes in the table below to see if you spotted all the anomalies.

Case	Country	Error	Comment
1	Austria	Clearly the 7 for 2009 is too low.	Typing error (70)
3	Bulgaria	Access data missing	Could be N/A or just missed (19)
4	Croatia	Male % cannot exceed 100	Digit repeated (11)
5	Czech	Access data missing	Could be N/A or just missed (35)
8	Finland	Too rich!	Digit repeated (£35660)
9	France	Region is 'Western' not 'West'	Could see this error if used Frequencies
10	Germany	Too poor!	Decimal point wrongly positioned (£35940)

TUTORIAL T6: One-variable Frequency Tables

T6.1 One-variable Frequency Table for scale data – Mean

1. Load data file: **File** → **Open** → **Data** → **DATA04_OpenAccess_HEIs.sav** (if not loaded).
 - ▶ Use **Edit** → **Options** to check that in the **General** tab window the **Variable Lists** choices are **Display names** and **File**, to match the listing format used in this tutorial. Change if needed, then click **Apply**, then **OK** to close the message box which appears, and **OK** again to exit.
2. Select **Analyze** → **Tables** → **Custom Tables**
 - ▶ This opens the **Custom Tables** window shown below.
3. Drag the scale variable **Q8a** [Number of items deposited ...] to **Rows**.



4. Click **OK**.
 - ▶ This produces a single number – the mean of all the values for that variable. It is displayed with no decimal places (the Auto default setting in this case):

	Mean
Number of items deposited in 2006-7	1163

5. To override the **Auto** setting, proceed as follows:

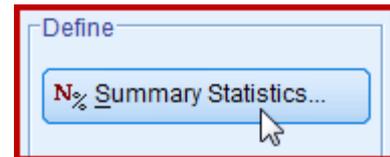
Select **Analyze** → **Tables** → **Custom Tables**

► You can close the annoying Custom Tables pop-up dialog – if you first click on the ‘Don’t show this dialog again’ box it won’t appear again)

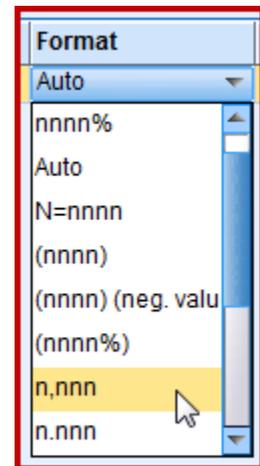
6. Click on the label ‘Number’ to ensure it is highlighted →

	Mean
Number ..	nnnn

and click on the **N% Summary Statistics** button →



7. Open the **Format** drop-down menu → (which will probably be displaying ‘Auto’).



and carefully select ‘n,nnn’ (with a comma not a full stop!) →

8. Change the Decimals to ‘2’.

► Note that the **Format** (see below) has changed to ‘n,nnn.nn’:

Statistics	Label	Format	Decima...
Mean	Mean	n,nnn,nn	2

9. Click **Apply to Selection**.

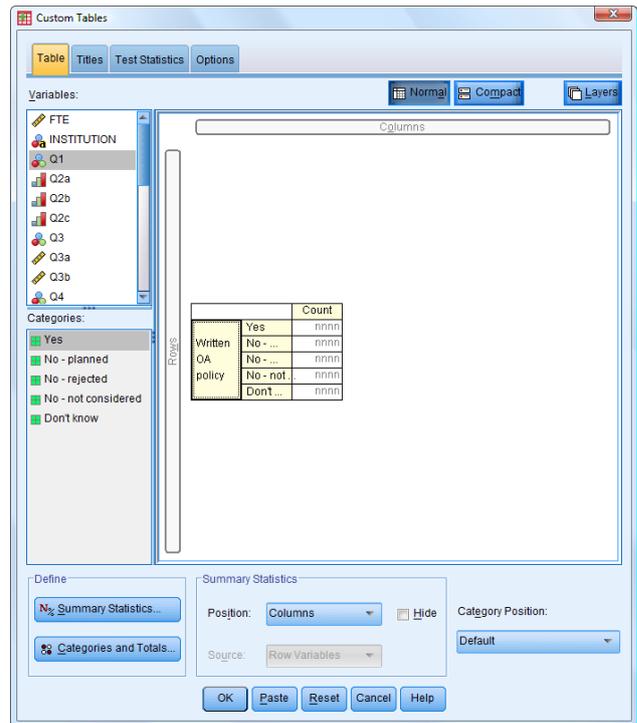
10. Click **OK**.

► This produces the mean of all the values for that variable, but now it is displayed with 2 decimal places:

	Mean
Number of items deposited in 2006-7	1,163.44

T6.2 One-variable Frequency Table for nominal data – Count and Total

1. Load data file: **File** → **Open** → **Data** → **DATA04_OpenAccess_HEIs.sav** (if not loaded).
2. Select **Analyze** → **Tables** → **Custom Tables**.
3. Drag the nominal variable **Q1** [Written OA policy] to the **Rows** label rectangle (it will change to red until the mouse button is released).
4. The **Custom Tables** window will appear as shown here with the table template →
5. Click on **Categories and Totals...**
6. Click on **Total** in the **Show** box.
7. Click **Apply**.

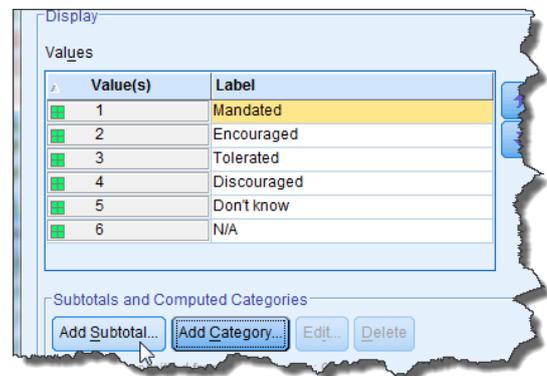
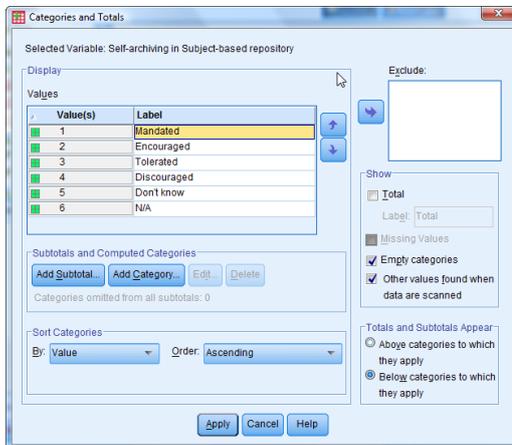


8. Click **OK** to produce the table in the **Viewer** Output window as shown below.
- Note that by default this procedure produces a Count of the different values for the variable and we have additionally asked for the Total which appears at the bottom.

		Count
Written OA policy	Yes	9
	No - planned	13
	No - rejected	2
	No - not considered	15
	Don't know	0
	Total	39

T6.3 One-variable Frequency Table for ordinal data – Count with Subtotals

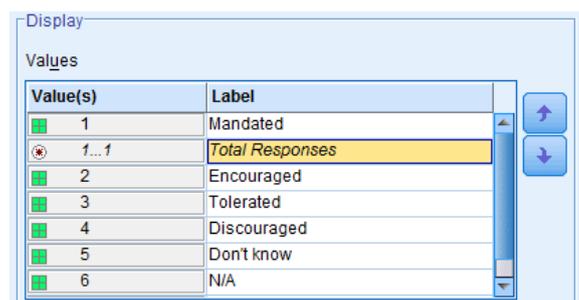
1. Load data file: **File** → **Open** → **Data** → DATA04_OpenAccess_HEIs.sav (if not loaded).
2. Select **Analyze** → **Tables** → **Custom Tables**.
3. Click **Reset**.
4. Click **All Tabs**.
5. Drag variable **Q2b** [Self-archiving in Subject-based repository] to the **Rows** label.
 - ▶ This variable has four codes (1 to 4) for a range of responses to a question plus two codes for Don't know (5) and N/A (6). It could be useful to have separate subtotals for the first four and the other two. This is how it can be achieved.
6. In the **Define** box click on **Categories and Totals...** which will show the six responses:



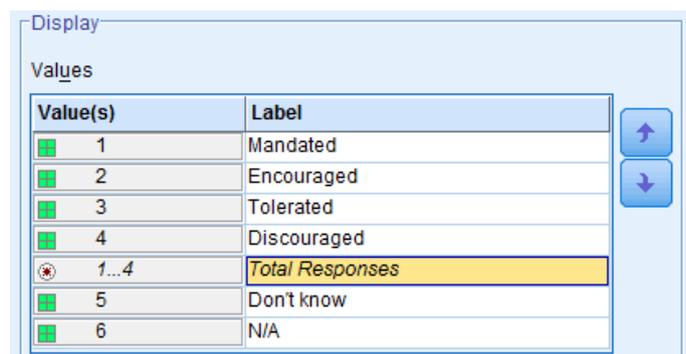
7. Click on **Add Subtotal...** in the **Subtotals and Computed Categories** box.
8. In the **Define Subtotal** window change the Label from **Subtotal** to **Total Responses**.
9. Click **Continue**.

▶ This will insert **Total Responses** below the first value label ('Mandated').

10. Ensuring that **Total Responses** is still highlighted, click the blue down arrow three times to move **Total Responses** to beneath the fourth value label ('Discouraged').



▶ Note that the **Value(s)** column shows the range for **Total Responses** to be **1...4**.



11. Click on the next row down ('Don't know').
12. Click on **Add Subtotal...** in the **Subtotals and Computed Categories** box.
13. In the **Define Subtotal** window change the Label to **Total Non-Responses**.
14. Click **Continue**.
 - ▶ This will insert **Total Non-Responses** below the value label ('Don't know').
 - ▶ You may need to enlarge the whole **Categories and Totals** window (or scroll down in the **Value(s)** box) to see this.
15. Ensuring that **Total Non-Responses** is still highlighted, click the blue down arrow once to move **Total Non-Responses** to beneath the last value label ('N/A').

Value(s)	Label
1	Mandated
2	Encouraged
3	Tolerated
4	Discouraged
1...4	Total Responses
5	Don't know
6	N/A
5...6	Total Non- Responses

- ▶ Note that the **Value(s)** column shows the range for **Total Non-Responses** to be 5...6.

16. Click **Apply**.
17. Click **OK**.

- ▶ This produces a stacked table with two subtotals:

		Count
Self-archiving in Subject-based repository	Mandated	0
	Encouraged	15
	Tolerated	9
	Discouraged	0
	Total Responses	24
	Don't know	6
	N/A	5
	Total Non-Responses	11

T6.4 One-variable Frequency Table for nominal data – Count with order sorted

1. Load data file: **File** → **Open** → **Data** → DATA04_OpenAccess_HEIs.sav (if not loaded).
2. Select **Analyze** → **Tables** → **Custom Tables**
3. Click **Reset**.
4. Click **All Tabs**.
5. Drag nominal variable **SECTOR** [HEI group] to the **Rows** label.

		Count
HEI group	RLUK member	12
	Other Pre-1992 university	12
	Post-1992 university	13
	HE college	2

6. Click **OK** to produce a one-way table → which lists the variable in ascending value order (i.e. in the codes order – ‘1’ is the code for ‘RLUK member’, and so on).
7. Then repeat **Analyze** → **Tables** → **Custom Tables**
8. Click on **Categories and Totals...** in the **Define** box.
9. In the **Sort Categories** drop down **By** menu change **Value** to **Count**.
10. Select **Descending** in the **Sort Categories** drop down **Order** menu.
11. Click **Apply**.

		Count
HEI group	Post-1992 university	13
	Other Pre-1992 university	12
	RLUK member	12
	HE college	2

12. Click **OK** to produce a one-way table → which lists the variable in descending Count order (i.e. frequency order).
13. Then repeat **Analyze** → **Tables** → **Custom Tables**
14. Click on **Categories and Totals...** in the **Define** box.
15. In the **Sort Categories** **By** drop down menu select **Label**.
16. In the **Sort Categories** **Order** drop down menu select **Ascending**.
17. Click **Apply**.
18. Click **OK** to produce a one-way table → which lists the variable in alphabetical order of **Label**.

		Count
HEI group	HE college	2
	Other Pre-1992 university	12
	Post-1992 university	13
	RLUK member	12

TUTORIAL T7: Two-variable Frequency Tables

T7.1 Two-variable Two-way Frequency Table for scale & nominal data – Count , Max, Min, Median

1. Load data file: **File** → **Open** → **Data** → DATA04_OpenAccess_HEIs.sav (if not loaded).
2. Select **Analyze** → **Tables** → **Custom Tables**
3. Click **Reset**.
4. Click **All Tabs**.

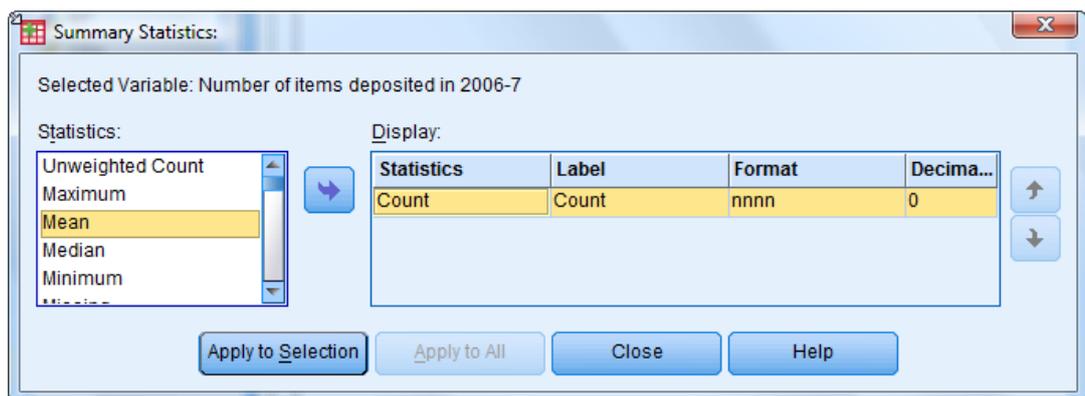
5. Drag the **nominal** variable **SECTOR** [HEI group] to the **Rows** label to produce this table template:

		Count
HEI group	RLUK ...	nnnn
	Other ...	nnnn
	Post-1...	nnnn
	HE ...	nnnn

6. Drag **scale** variable **Q8a** [Number of items deposited] to the **Columns** label to produce this table template →

		Number .
		Mean
HEI group	RLUK ...	nnnn
	Other ...	nnnn
	Post-1...	nnnn
	HE ...	nnnn

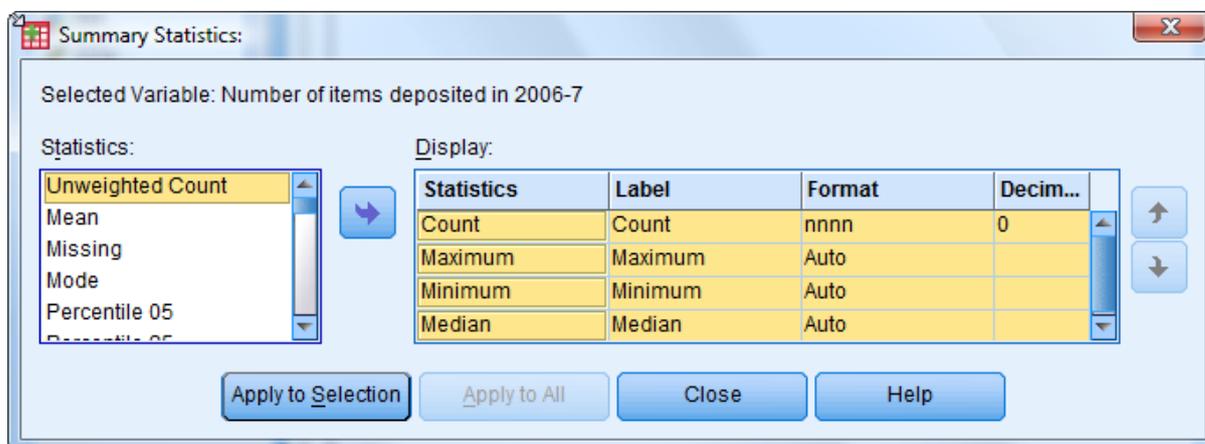
7. Click on **N% Summary Statistics**.
8. Click on **Mean** in the **Summary Statistics** window **Display** box.
9. Click on the left pointing arrow, to remove **Mean**.
10. Click on **Count** in the **Statistics** box (it will be the top entry).
11. Click on the right pointing arrow to insert **Count** into the **Display** box.



12. Find and click on **Maximum** in the **Statistics** box.
13. Click on the right pointing arrow to insert **Maximum** into the **Display** box.
14. Find and click on **Minimum** in the **Statistics** box.
15. Click on the right pointing arrow to insert **Minimum** into the **Display** box.

16. Find and click on **Median** in the **Statistics** box.
17. Click on the right pointing arrow to insert **Median** into the **Display** box.

► The **Summary Statistics** window will look something like this:



18. Click **Apply to Selection**.

19. Click **OK**.

► This produces a set of four statistics for the four values for the HEI group:

		Number of items deposited in 2006-7			
		Count	Maximum	Minimum	Median
HEI group	RLUK member	12	6742	13	535
	Other Pre-1992 university	12	400	24	100
	Post-1992 university	13	9000	0	0
	HE college	2	0	0	0

- The order in which the variables appear is determined by their order in the **Display** box, which can be changed using the up/down arrows. (Repeat **Analyze** → **Tables** → **Custom Tables** and click on **N% Summary Statistics** if interested.)
- 'RLUK' stands for 'Research Libraries UK' - this variable indicates whether a library is a member of this organisation.
- Most 'RLUK' members are universities (all those pre-1992 universities)
- In 1992 many former polytechnics gained university status. Subsequently some other HE colleges have gained university status. 'Post-1992' includes the former polytechnics and those former HE colleges.
- Some HE colleges provide degree level courses, particularly specialist colleges (e.g. Art, Music).

T7.2 Two-variable Nested Frequency Table for nominal data – Count & Col%

1. Load data file: **File** → **Open** → **Data** → DATA04_OpenAccess_HEIs.sav (if not loaded).
2. Select **Analyze** → **Tables** → **Custom Tables**
3. Click **Reset**.
4. Click **All Tabs**.
5. Drag nominal variable **Q3** [OA material in Library Catalogue] to the **Rows** label.
6. Drag nominal variable **SECTOR** [HEI group] to the **Rows** label.
7. Click on the column heading 'OA material in ...' in the display icon to select that variable.

				Count	Colum...
HEI group	RLUK member	OA material in ...	Yes	nnnn	nnnn.n%
			No	nnnn	nnnn.n%
			Don't...	nnnn	nnnn.n%
	Other Pre-1992 ...	OA material in ...	Yes	nnnn	nnnn.n%
			No	nnnn	nnnn.n%
			Don't...	nnnn	nnnn.n%
	Post-1992 universit.	OA material in ...	Yes	nnnn	nnnn.n%
			No	nnnn	nnnn.n%
			Don't...	nnnn	nnnn.n%
	HE college	OA material in ...	Yes	nnnn	nnnn.n%
			No	nnnn	nnnn.n%
			Don't...	nnnn	nnnn.n%

8. Click on **N% Summary Statistics ...** in the **Define** box.
 - ▶ 'Count' should already appear in the **Display** box.
9. Click on **Col N %** in the **Statistics** box.
10. Click on the right pointing blue arrow to move **Col N %** into the **Display** box.
11. Click **Apply to Selection**.
12. Click **OK** to produce the table below in the **Viewer** window:

				Count	Column N %
HEI group	RLUK member	OA material in Library Catalogue	Yes	8	80.0%
			No	2	20.0%
			Don't know	0	.0%
	Other Pre-1992 university	OA material in Library Catalogue	Yes	7	63.6%
			No	3	27.3%
			Don't know	1	9.1%
	Post-1992 university	OA material in Library Catalogue	Yes	8	66.7%
			No	4	33.3%
			Don't know	0	.0%
HE college	OA material in Library Catalogue	Yes	1	50.0%	
		No	1	50.0%	
		Don't know	0	.0%	

T7.3 Two-variable Two-way Frequency Table for nominal data – Count & Col%

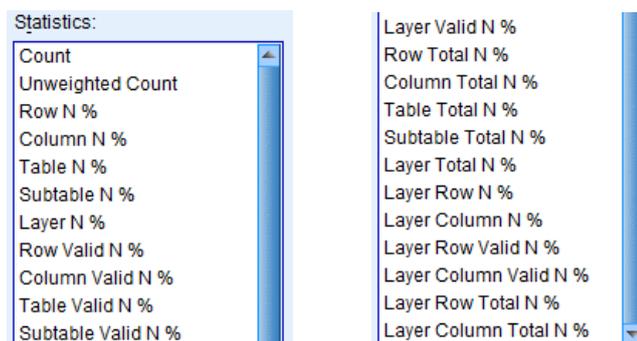
This is a repeat of T7.2 but setting out the data differently.

1. Load data file: **File** → **Open** → **Data** → DATA04_OpenAccess_HEIs.sav (if not loaded).
2. Select **Analyze** → **Tables** → **Custom Tables**
3. Click **Reset**.
4. Click **All Tabs**.
5. Drag nominal variable **Q3** [OA material in Library Catalogue] to the **Columns** label.
[This is the difference from T7.2]
6. Drag nominal variable **SECTOR** [HEI group] to the **Rows** label.
7. Click on the left hand side of the display icon ('OA material ...') to select that variable.
8. Click on **N% Summary Statistics ...** in the **Define** box.
9. Click on **Row N%** in the **Statistics** box.
10. Click on the right pointing blue arrow to insert **Row N%** into the **Display** box.
11. Click **Apply to Selection**.
12. Click **OK** to produce the table below in the **Viewer** window:

		OA material in Library ...		
		Yes	No	Don't ...
		Count	Count	Count
HEI group	RLUK ...	nnnn	nnnn	nnnn
	Other ...	nnnn	nnnn	nnnn
	Post-...	nnnn	nnnn	nnnn
	HE ...	nnnn	nnnn	nnnn

		OA material in Library Catalogue					
		Yes		No		Don't know	
		Count	Row N %	Count	Row N %	Count	Row N %
HEI group	RLUK member	8	80.0%	2	20.0%	0	.0%
	Other Pre-1992 university	7	63.6%	3	27.3%	1	9.1%
	Post-1992 university	8	66.7%	4	33.3%	0	.0%
	HE college	1	50.0%	1	50.0%	0	.0%

- ▶ Although this table contains exactly the same data as that in T7.2 you may find this one much easier to understand.
- ▶ There is a lot to choosing the right format for tables for the given purpose.
- ▶ There are many types of percentages which can be applied to tables. If you have the time and inclination it can be an interesting challenge to explore them further. Below is the complete list found in the **Statistics** box!



T7.4 Two-variable Two-way Frequency Table for nominal and ordinal data – interchanging rows and columns – Row%

1. Load data file: **File** → **Open** → **Data** → DATA04_OpenAccess_HEIs.sav (if not loaded).
2. Select **Analyze** → **Tables** → **Custom Tables**
3. Click **Reset**.
4. Click **All Tabs**.
5. Drag nominal variable **RLUK** [Research Libraries UK member] to the **Rows** label.
6. Drag ordinal variable **Q2a** [Self-archiving in HEI's repository] to the **Columns** label to produce this table template:

		Self-archiving in HEI's repository			
		Mandate.	Encoura.	Tolerated	Discoura.
		Count	Count	Count	Count
RLUK member	No	nnnn	nnnn	nnnn	nnnn
	Yes	nnnn	nnnn	nnnn	nnnn

7. Click on the display icon 'RLUK member' to select that variable.
8. Click on **N% Summary Statistics ...** in the **Define** box.
9. Click on **Count** in the **Display** box.
10. Click on the left pointing blue arrow, to remove **Count**.

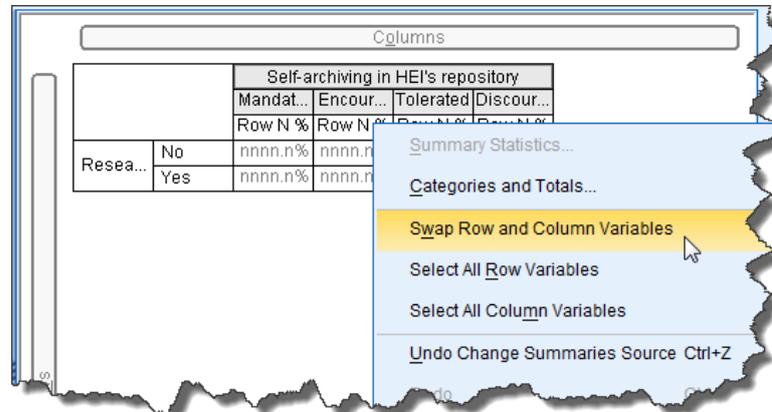


11. Click on **Row N%** in the **Statistics** box.
12. Click on the right pointing blue arrow to insert **Row N%** into the **Display** box.
 - Make sure the **Decimals** entry for **Row N%** is set to '1'
13. Click **Apply to Selection**.
14. Click **OK** to produce the table below in the **Viewer** Output window:

		Self-archiving in HEI's repository			
		Mandated	Encouraged	Tolerated	Discouraged
		Row N %	Row N %	Row N %	Row N %
Research Libraries UK member	No	26.3%	73.7%	.0%	.0%
	Yes	25.0%	66.7%	8.3%	.0%

15. Now once again select **Analyze** → **Tables** → **Custom Tables**

16. Right click on the output table template to produce this drop down menu:



17. Click **Swap Row and Column variables** to produce the reorganized table template:

		RLUK member	
		No	Yes
		Count	Count
Self-archiving in HEI's ...	Mandate...	nnnn	nnnn
	Encoura..	nnnn	nnnn
	Tolerate..	nnnn	nnnn
	Discour..	nnnn	nnnn

18. Click **Apply to Selection**.

19. Click **OK** to produce the table below in the **Viewer** Output window:

		RLUK member	
		No	Yes
		Count	Count
Self-archiving in HEI's repository	Mandated	5	3
	Encouraged	14	8
	Tolerated	0	1
	Discouraged	0	0

► This is probably not what you would expect! The rows and columns have indeed been swapped (compare to the table below step 14). However, the **Row N%** has been replaced by the default **Count**. Presumably this is because **Row N%** might not be what was now wanted (maybe **Column N%**?) and *SPSS* cannot decide.

► You can, of course go back and choose whatever you require, by selecting **Analyze** → **Tables** → **Custom Tables** and clicking on **N% Summary Statistics**.

T7.5 Two-variable Nested & Stacked Frequency Tables for nominal and scale data – Count

1. Load data file: **File** → **Open** → **Data** → DATA04_OpenAccess_HEIs.sav (if not loaded)
2. Select **Analyze** → **Tables** → **Custom Tables**
3. Click **Reset**.
4. Click **All Tabs**.
5. Drag nominal variable **RLUK** [RLUK member] to the **Rows** label.
6. Drag scale variable **HE_TYPE** [Type of HEI] into the table template box and drag it around over **RLUK**'s box and note that it can produce the following:
 - a thin red rectangle at the top of **RLUK**'s box
 - a thin red rectangle at the bottom of **RLUK**'s box
 - a thin red rectangle at the right of **RLUK**'s box
 - a thin red rectangle at the left of **RLUK**'s box
 - a large red box covering all of **RLUK**'s box.
7. Investigate in turn what these five actions produce. Below are the results you should obtain.

► Thin red rectangle to the left:

		Count
RLUK member	No	nnnn
RLUK member	Yes	nnnn

► When the mouse is released produces this table template:

			Count	
Type of HEI	Pre-19...	RLUK member	No	nnnn
			Yes	nnnn
	Post-1...	RLUK member	No	nnnn
			Yes	nnnn

► Which produces a **nested table** with **HE_TYPE** on the left and **RLUK** on the right.

				Count
Type of HEI	Pre-1992	RLUK member	No	12
			Yes	12
	Post-1992	RLUK member	No	15
			Yes	0

► Thin red rectangle to the right:

		Count
RLUK member	No	nnnn
RLUK member	Yes	nnnn

► Which produces a **nested table** with **RLUK** on the left and **HE_TYPE** on the right:

				Count
RLUK member	No	Type of HEI	Pre-1992	12
			Post-1992	15
	Yes	Type of HEI	Pre-1992	12
			Post-1992	0

- ▶ Thin red rectangle at the top:
- ▶ Which produces a **stacked table** with **HE_TYPE** above **RLUK**.

		HE_TYPE	Count
RLUK member	No		nnnn
	Yes		nnnn

		Count
Type of HEI	Pre-1992	24
	Post-1992	15
RLUK member	No	27
	Yes	12

- ▶ Thin red rectangle at the bottom – produces a **stacked table** with **RLUK** above **HE_TYPE**.

		Count
RLUK member	No	nnnn
	Yes	nnnn

HE_TYPE

		Count
RLUK member	No	27
	Yes	12
Type of HEI	Pre-1992	24
	Post-1992	15

- ▶ Large red rectangle over the whole variable 1 table – *replaces* **RLUK**'s table with **HE_TYPE**'s table (probably not what you want !).

		Count
RLUK member	No	nnnn
	Yes	nnnn

		Count
Type of HEI	Pre-1992	24
	Post-1992	15

TUTORIAL T8: Three- and Four-variable Frequency Tables

T8.1 Three-variable Frequency Table for two nominal variables and one scale variable – Mean

1. Load data file: **File** → **Open** → **Data** → DATA04_OpenAccess_HEIs.sav (if not loaded).
2. Select **Analyze** → **Tables** → **Custom Tables**
3. Click **Reset**.
4. Click **All Tabs**.
5. Drag nominal variable **SECTOR** [HEI group] to the **Rows** label.
6. Drag nominal variable **Q1** [Written OA policy] to the **Columns** label.
7. Drag scale variable **Q9e** [Running costs - Total (£)] to the **Columns** label.
8. Click **OK** to produce the table below:

		Running costs - Total (£)				
		Written OA policy				
		Yes	No - planned	No - rejected	No - not considered	Don't know
		Mean	Mean	Mean	Mean	Mean
HEI group	RLUK member	£68,500	£6,833	.	£15,000	.
	Other Pre-1992 university	£46,000	£23,000	£27,000	£.	.
	Post-1992 university	£51,000	£.	.	£0	.
	HE college	.	.	.	£9,000	.

- ▶ Reminder: 'RLUK' stands for 'Research Libraries UK' - this variable indicates whether a library is a member of this organisation.
- ▶ There are empty cells because there were no entries in some categories. However, the entry of zero for the Post-1992 universities is different – it shows that there was at least one entry but the mean score was £0.
- ▶ The RLUK members with written OA policies spent much more on average than RLUK members without OA written policies.
- ▶ The RLUK members with OA written policies spent more on average than other Pre-1992 universities and Post-1992 universities with OA written policies.
- ▶ It is seen that in this sample the Post-1992 universities spent a lot more than the specialist HE colleges.
- ▶ Institution size will be a factor in these findings.
- ▶ The order in which the three variables are organised, and whether the table is nested or stacked or a mixture is controlled by the order in which the variables are dragged to the **Rows** label or **Columns** label or dragged *within* the table template display box. There is plenty of room for exploration here!

T8.2 Three-variable Frequency Table for three nominal variables – Count

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav**.
 - ▶ Use **Edit** → **Options** to check that in the **General** tab window the **Variable Lists** options selected are **Display names** and **File**, to match the variable list format used in this tutorial.
 - ▶ Use **Edit** → **Options** to check that in the **Output Labels** tab window the **Pivot Table Labeling** option selected for **Variables in labels shown as** is **Labels**, and the option selected for **Variable values in labels shown as** is **Labels**, to match the output display formats used in this tutorial.
 - ▶ Be sure to click **Apply** if you need make any changes above, before clicking **OK** to exit.
2. Select **Analyze** → **Tables** → **Custom Tables**
3. Drag nominal variable **gender** [Male or Female] to the **Rows** label.
4. Drag nominal variable **ft_pt** [Full Time or Part time] to the **Columns** label.
5. Drag nominal variable **prog** [Programme] to the **Rows** label.
6. Click **OK** to produce the table below:

				Study mode	
				Full-time	Part-time
				Count	Count
Programme	Library Studies	Gender	Male	16	0
			Female	12	1
	Information Management	Gender	Male	31	1
			Female	4	0
	Publishing	Gender	Male	7	0
			Female	13	2
	Information & Know. Man.	Gender	Male	11	1
			Female	18	1
	Information & Library Man.	Gender	Male	6	1
			Female	19	1
	Electronic Publishing	Gender	Male	2	0
			Female	3	0

- ▶ Varying the order of dragging in the variables (or positioning them in the rows and columns or within the table template) will produce different presentations of the data (for the interested reader to investigate.)
- ▶ Using **Edit** → **Options** to change the **Output Labels** tab window **Pivot Table Labeling** options for **Variables in labels ...** and for **Variable values in labels ...** will produce different output display formats (for the interested reader to investigate.)

T8.3 Four-variable Frequency Table for three nominal variables and one scale variable – Median and Mode

1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav
 - ▶ Use **Edit** → **Options** to ensure that in the **Output Labels** tab window the **Pivot Table Labeling** option for **Variables in labels shown as** is **Labels**, and the option for **Variable values in labels shown as** is **Labels**, to match the output display formats used here.
2. In **Variable View** change the **Decimals** attribute for **modules** to 1 d.p.
3. Select **Analyze** → **Tables** → **Custom Tables**
4. If variables are already selected click **Reset** and click **All Tabs**.
5. Locate and drag scale variable **modules** [Number of modules] to the **Columns** label.
6. Click on **N% Summary Statistics** in the **Define** box and remove **Mean** from the **Display** box (by selecting **Mean** and using the blue arrow) and replace it with **Median** and **Mode** (along the lines of what was done in TUTORIAL T7.1).

7. Click **Apply to Selection**. This table template should appear:

No of modules ...	
Median	Mode
nnnn.n	nnnn.n

8. Drag nominal variable **gender** [Male or Female] to the **Columns** label.

▶ The table template should change to this:

Male or Female			
Male		Female	
No of modules ...		No of modules ...	
Median	Mode	Median	Mode
nnnn.n	nnnn.n	nnnn.n	nnnn.n

9. Drag nominal variable **ug_pg** [UG or PG] to the **Rows** label.

10. Drag nominal variable **ft_pt** [Full time or Part time] to the **Rows** label.

▶ This table template should now look like this:

				Gender			
				Male		Female	
				No of modules ...		No of modules ...	
				Median	Mode	Median	Mode
Study mode	Full-time	Type of student	Undergr..	nnnn.n	nnnn.n	nnnn.n	nnnn.n
			Postgra..	nnnn.n	nnnn.n	nnnn.n	nnnn.n
	Part-time	Type of student	Undergr..	nnnn.n	nnnn.n	nnnn.n	nnnn.n
			Postgra..	nnnn.n	nnnn.n	nnnn.n	nnnn.n

11. Click **OK** to produce the output table below:

				Gender			
				Male		Female	
				No of modules accessed		No of modules accessed	
				Median	Mode	Median	Mode
Study mode	Full-time	Type of student	Undergraduate	6.5	6.0	6.0	5.0
			Postgraduate	6.0	6.0	6.0	6.0
	Part-time	Type of student	Undergraduate	8.0	8.0	5.0	5.0
			Postgraduate	5.5	2.0	9.0	8.0

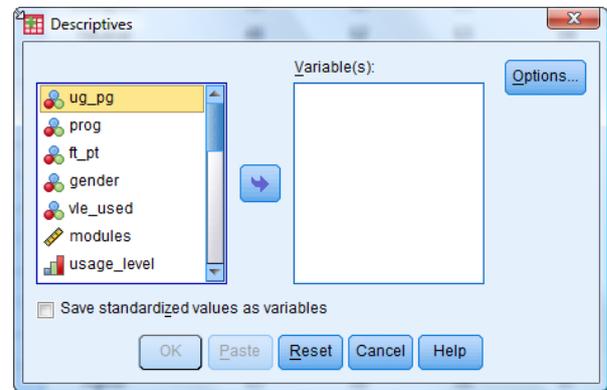
TUTORIAL T9: Descriptive Statistics

The **Descriptives** procedure provides basic statistical measures of location and dispersion ('average' and 'spread') for ordinal and scale data similar to that available in **Frequencies**.

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded)

2. Select **Analyze** → **Descriptive Statistics** → **Descriptives**

▶ This opens the **Descriptives** window →



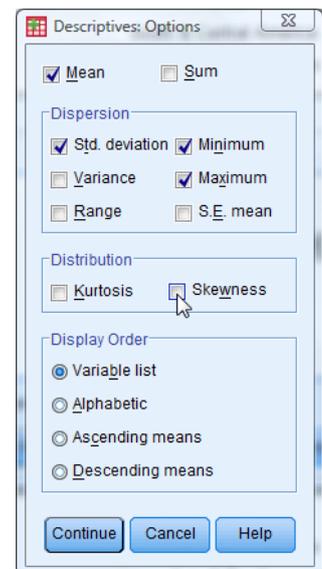
3. Using the blue arrow (or by double-clicking each) move the following variables into the Variable(s) box:

Marks_modA, Marks_modB, Marks_modC, Marks_modD.

4. Click on **Options**.

▶ This opens the **Descriptives Options** window shown here →

▶ This shows the default statistics to be:
Mean, Standard deviation, Minimum and Maximum →



5. Select **Skewness** →
(this is a measure of how 'flattened' to one side is the distribution of the data. N.B. positively skewed means having a longer tail to the right).

6. Click on **Continue** and click **OK** to produce the output below:

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Module A marks	150	43	78	63.47	6.137	-.830	.198
Module B marks	150	42	83	63.40	6.364	-.268	.198
Module C marks	149	48	75	64.29	5.440	-.443	.199
Module D marks	150	18	92	52.65	14.512	.236	.198
Valid N (listwise)	149						

- ▶ The N column indicates that there is one missing value for Module C, i.e. 149 valid entries.
- ▶ Module D marks are very different from the others – with a much lower mean, a much larger standard deviation and a much wider range (Max – Min).
- ▶ The standard deviations are similar for Modules A, B and C, so they could be combined, and a parametric test such as a *t* Test (or ANOVA) could be used to compare them.
- ▶ Significant skewness occurs when the Skewness Statistic lies outside +/- 2 x Std. Error.

TUTORIAL T10: Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach in which relatively simple but insightful analysis of data is undertaken before any serious statistical tests of hypotheses are made. It gives a 'feel' for the data. It was pioneered by John Tukey in his 1977 classic textbook *Exploratory Data Analysis*. SPSS's **Explore** procedure introduces these techniques.

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav**
2. Select **Analyze** → **Descriptive Statistics** → **Explore...**
3. Use the blue arrow to move the variable **modules** into the **Dependent List** box.
 - ▶ This variable records the number of modules (out of a maximum of 12) about which each student accessed information on the institution's VLE.
4. Click on **Plots...** and if **Stem-and-leaf** is not already selected, select it.
 - ▶ Histogram is available as an option – choose that too if you wish.
5. Click **Continue**.
6. Click **OK** to produce the following outputs:
 - ▶ The table contains many basic statistics about the chosen variable:

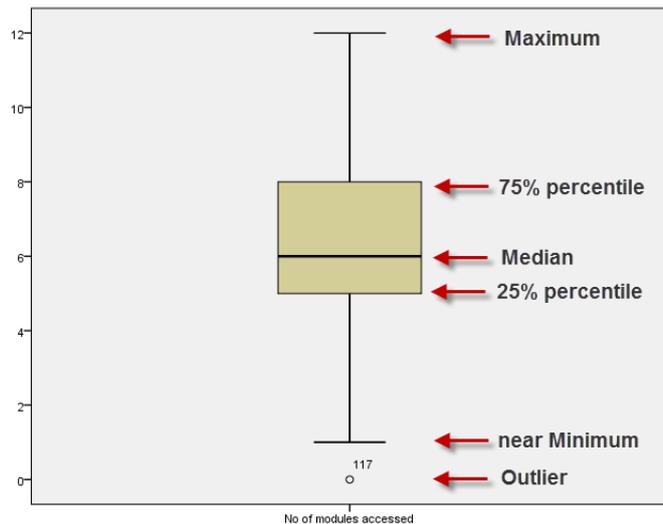
Descriptives			Statistic	Std. Error
No of modules accessed	Mean		6.32	.197
	95% Confidence Interval for Mean	Lower Bound	5.93	
		Upper Bound	6.71	
	5% Trimmed Mean		6.30	
	Median		6.00	
	Variance		5.803	
	Std. Deviation		2.409	
	Minimum		0	
	Maximum		12	
	Range		12	
	Interquartile Range		3	
	Skewness		.160	.198
	Kurtosis		.016	.394

- ▶ The stem-and-leaf plot (below left) is a cross between a table and a chart.
- ▶ The box plot (below right) illustrates the spread of the data – the box itself contains the middle 50%.

No of modules accessed Stem-and-Leaf Plot

Frequency	Stem	Leaf
1.00	Extremes	(=<.0)
2.00	1 .	00
5.00	2 .	00000
8.00	3 .	00000000
17.00	4 .	000000000000000000
21.00	5 .	00000000000000000000
29.00	6 .	000000000000000000000000
25.00	7 .	000000000000000000000000
16.00	8 .	000000000000000000
11.00	9 .	000000000000
6.00	10 .	000000
5.00	11 .	00000
4.00	12 .	0000

Stem width: 1
Each leaf: 1 case(s)

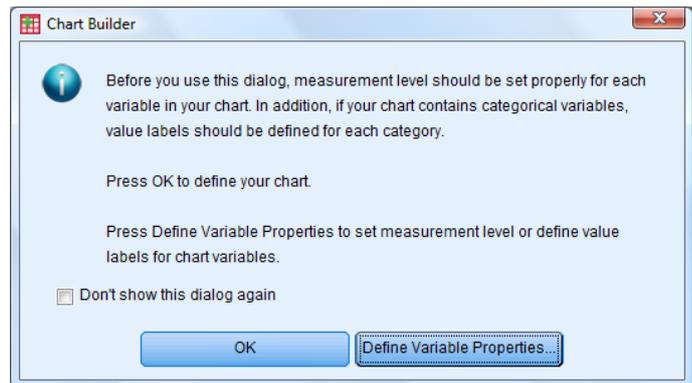


TUTORIAL T11: Simple Bar Chart – basic building

Here you will learn how to create and edit a simple bar chart. Many of the commands needed for this are common to the other types of chart.

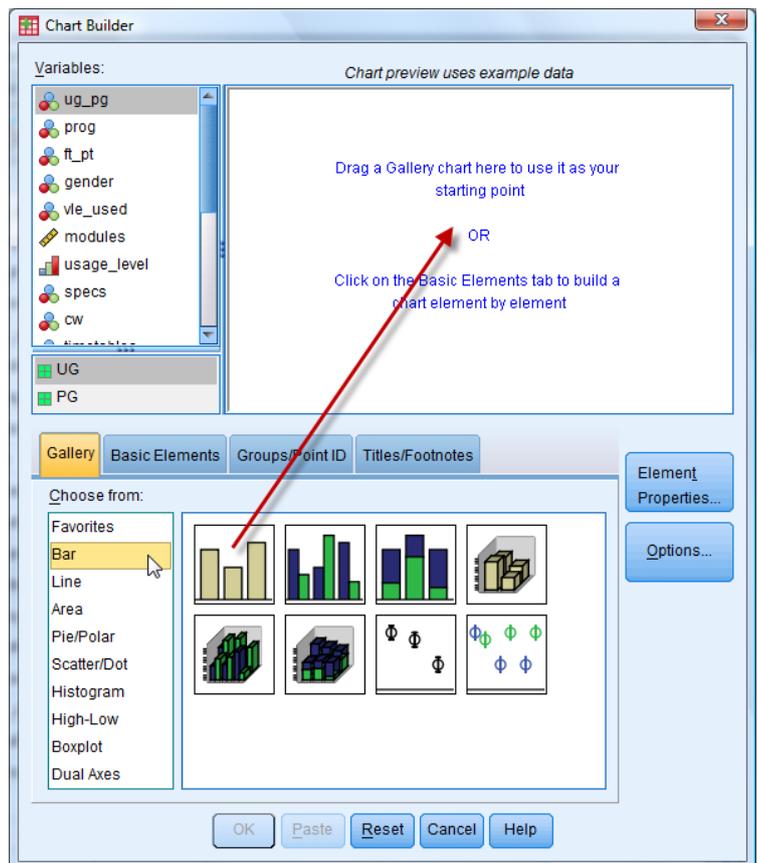
1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...**

- ▶ This message may appear →
- ▶ It provides a warning about categorical variables (i.e. nominal).
- ▶ You can check the tick box to stop this possibly helpful but probably annoying message popping up in the future.



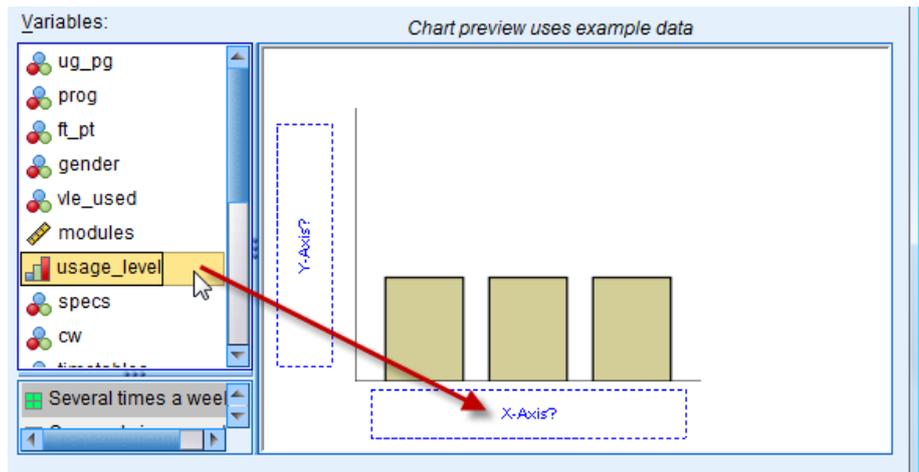
3. If necessary click **OK** to start building.

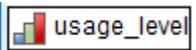
- ▶ The Chart Builder window will appear →
- ▶ Click on **Reset** if any chart is already selected.
- ▶ **Gallery** should be highlighted (if not, then select it).



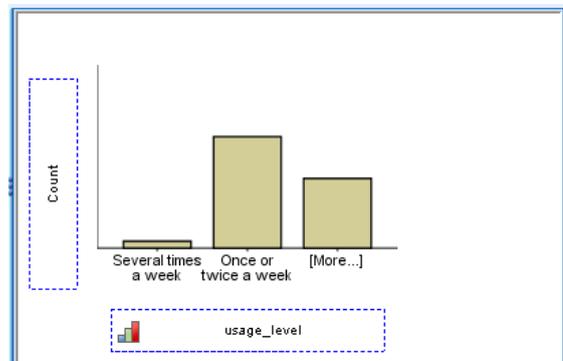
4. Click on **Bar** in the **Choose from** menu →
 - ▶ Hovering the cursor pointer over a **Gallery** icon reveals the name of the type of chart whose icon is shown (try it).
5. Drag the top-left icon ('Simple Bar') from the **Gallery** into the **Chart Preview** box.

6. Close (or just ignore) the **Element Properties** window which appears.
 - ▶ **Element Properties...** is available using a button in the **Chart Builder** window located on the far right below the **Chart Preview** box.
 - ▶ At this point **Chart Preview** will show a generic bar chart (as illustrated below).
7. Locate **usage_level** in the **Variables** list and drag it across to the X axis box whose label 'X-Axis?' has a questionable question mark.

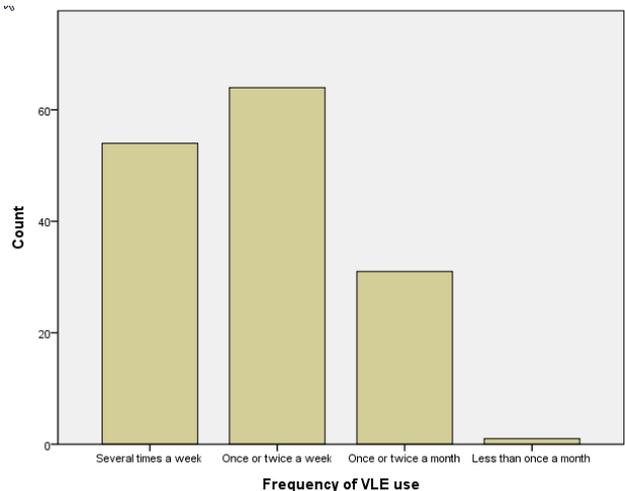


- ▶ The icon next to **usage_level** indicates that it is an ordinal variable → 

- ▶ 'Count' comes up automatically as the default for the Y-axis
- ▶ **Chart Preview** will show a schematic bar chart for the variable **usage_level**. Actually it is nothing like the shape of the distribution you will actually get, it's just some random data.



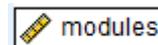
8. Click **OK** to generate the actual Simple Bar Chart shown here →



TUTORIAL T12: Simple Bar Chart – basic editing

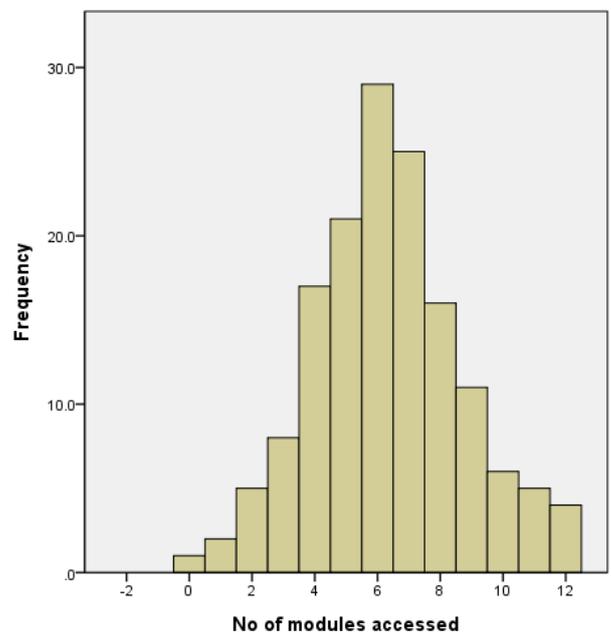
1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...**
 - ▶ Click on **Reset** if any chart is already selected.
3. Click on **Bar** and drag the 'Simple Bar' icon from the **Gallery** into the **Chart Preview** box.
4. Close the **Element Properties** window (if it opens).
5. Locate **modules** in the **Variables** list and drag it across to **X-Axis?**.

- ▶ The icon next to **modules** indicates that it is a scale variable →



6. Click **OK** to generate ... not a Bar Chart but a Histogram!

- ▶ The reason the data is presented as a Histogram rather than a Bar Chart (with bars separated) is because the data is defined as scale and *SPSS* assumes there may be many different values and therefore a Histogram would be more appropriate.
- ▶ This can be investigated by entering **Variable View** mode and changing the **Measure** for **modules**, as follows:



7. Click on **Variable View**.

8. Locate the **modules** row and scan across to the **Measure** field and change it from **Scale** to **Ordinal**.

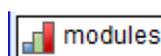
9. Select **Graphs** → **Chart Builder...**

10. Click **Reset** (otherwise *SPSS* will still think the variable is scale).

11. Click on **Bar** and drag the 'Simple Bar' icon from the **Gallery** into the **Chart Preview** box.

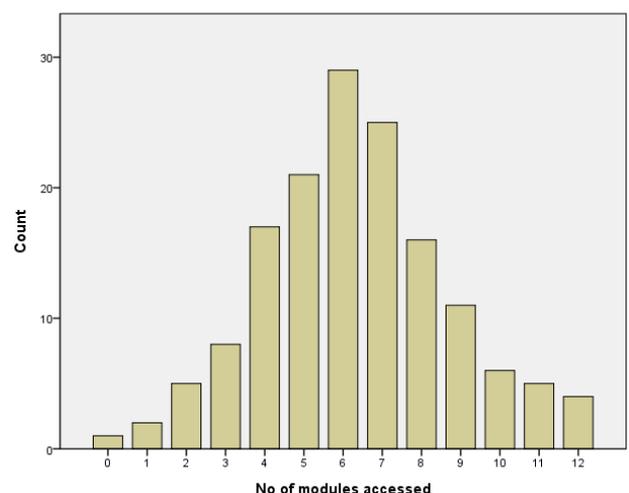
12. Locate **modules** in the **Variables** list and drag it across to **X-Axis?**.

- ▶ The icon next to **modules** now indicates it is an ordinal variable:



13. Click **OK** to generate ... a Bar Chart.

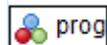
14. Finally, in **Variable View** change the **Measure** field for **modules** back to **Scale**.

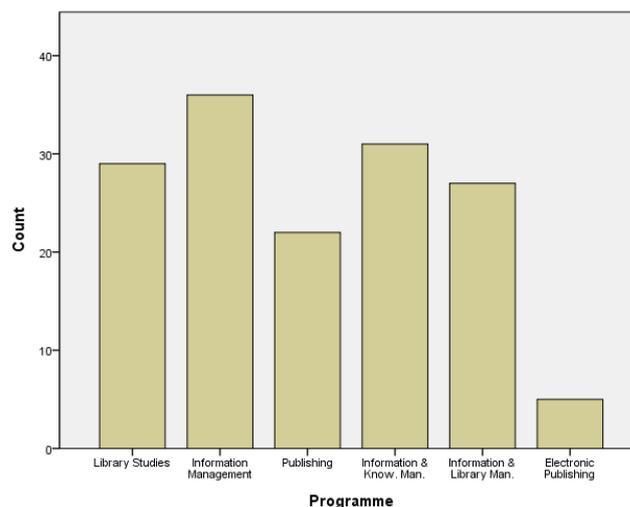


TUTORIAL T13: Simple Bar Chart – advanced

Here we build a basic chart (as before) and edit it in various ways. Many of the procedures illustrated apply equally to other types of chart, and will not be shown again. This is a somewhat lengthy and perhaps even tedious section (patience and care is needed to follow the steps carefully) but it is worth exploring the possibilities *SPSS* offers, and it should be useful to have this detailed example to refer to later.

T13.1 Building a simple bar chart

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav**
 - ▶ Use **Edit** → **Options** to check that in the **General** tab window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.
2. Select **Graphs** → **Chart Builder...** and click on **Reset** if a chart is already selected.
3. Click on **Gallery** (if not selected) and click on **Bar**.
4. Drag the 'Simple Bar' icon from the **Gallery** into the **Chart Preview** box.
5. Close the **Element Properties** window.
6. Locate **prog** in the **Variables** list and drag it across to **X-Axis?**
 - ▶ The icon next to **prog** indicates that it is a nominal variable → 
7. Click **OK** to generate the Bar Chart below.
 - ▶ Note that the Y axis label 'Count' is quite small (we will enlarge it a bit later).



T13.2 Editing possibilities

Every element and aspect of this chart can be edited and further elements can be added:

Some of the editing possibilities are:

- X-axis and Y-axis labels and their fonts, font sizes, font styles, colours.
- Y-axis increments, tick positions, font sizes.
- Bar colour.
- Bar label wording, font, font size, font style, colour.
- Chart background colour.
- Orientation.

Some of the possible additions which can be made to a chart are:

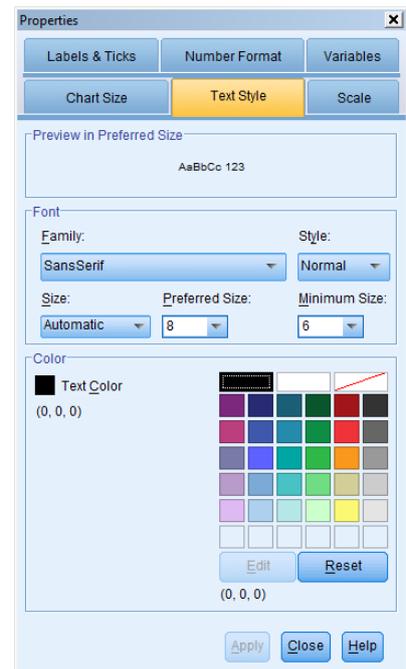
- Title, Subtitle, Text box, Footnote.
- Annotations to place within bars (mini-titles).
- Show numerical values within bars (frequency, percentage).
- Reference line (e.g. pass mark, average score, sales target).

More advanced bar charts have additional possibilities.

To edit a chart double-click on it. This opens the **Chart Editor** window. Once in **Chart Editor** there are several ways to set about editing an element of a chart. The two most used are:

- (1) Double-click on the element (needs to be done accurately!) to open a context-dependent **Properties** window →
- (2) Click on an element so that a light yellow box surrounds it, then do one of the following:
 - (a) Click on an icon in the one of the **Toolbars**.
 - (b) Choose an option from the **Edit** menu.
 - (c) Choose an option from the **Options** menu.
 - (d) Choose an option from the **Elements** menu.
 - (e) Choose an option from the **Format** menu.

Many of these actions open a context-dependent **Properties** window (examples on the next page).

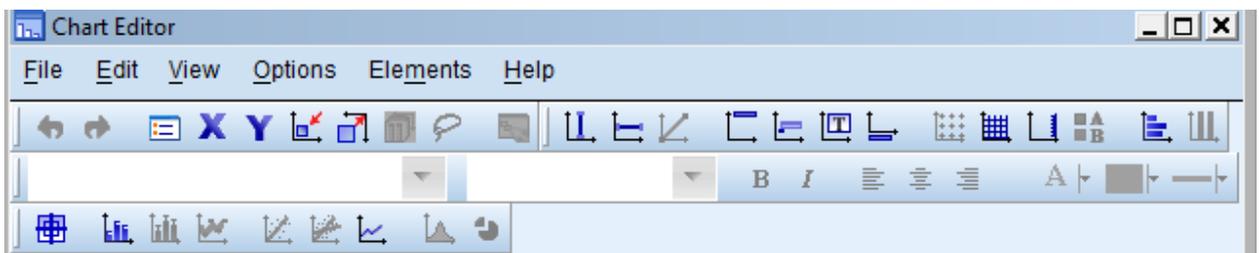


We now proceed to edit the bar chart

T13.3 Editing the simple bar chart

1. Double-click anywhere on the chart.

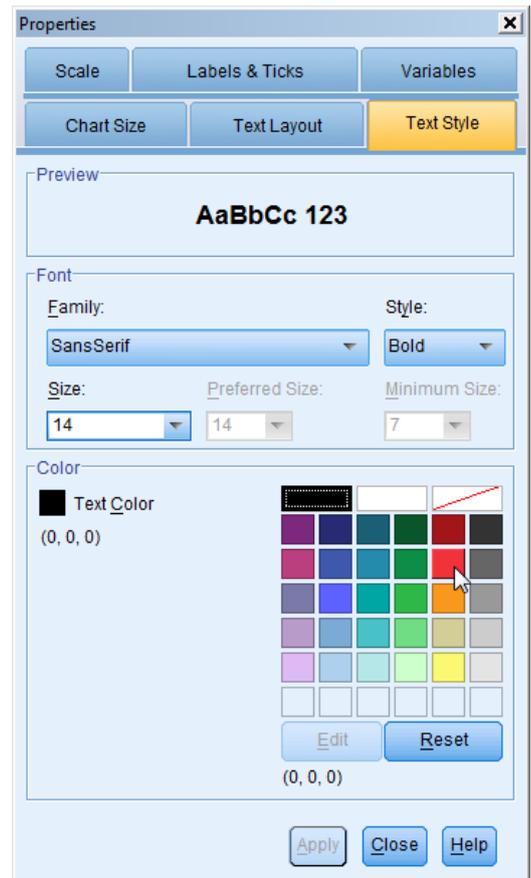
► This opens the **Chart Editor** window and may open a **Properties** menu (if so, close it).



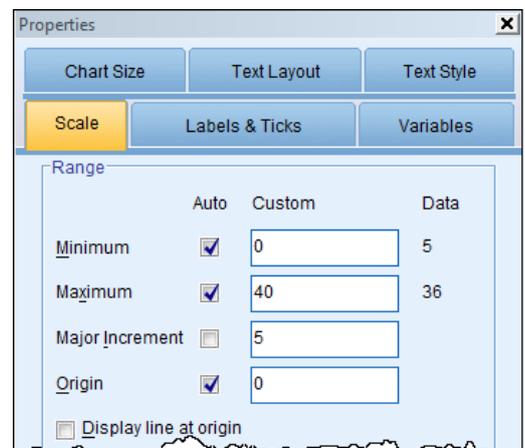
2. Click on the **Y-axis** label ('Count') to select it.

► 'Count' should be surrounded by a pale yellow rectangle. If not, try again.

3. **Edit** → **Properties** to open this window →
 - ▶ The **Text Style** menu should be selected (as illustrated). If not, select it before continuing.
4. To change the font size of the label click on the **Size** drop-down menu (default is 'Automatic') and choose '16'.
 - ▶ Note: This is not quite the same as changing the **Preferred Size** entry to 16 which leaves SPSS to make the final decision!
5. To check or change the style of the **Y-axis** label click on **Style** drop-down menu (default is 'Bold').
6. To change the colour of the **Y-axis** label (default is black) click on a colour in the **Color** palette (e.g. bright red).
7. Click **Apply** to confirm these changes in the **Text Style** menu.
 - ▶ This **Properties** window remains open – you can drag it aside if it obscures what you have done.



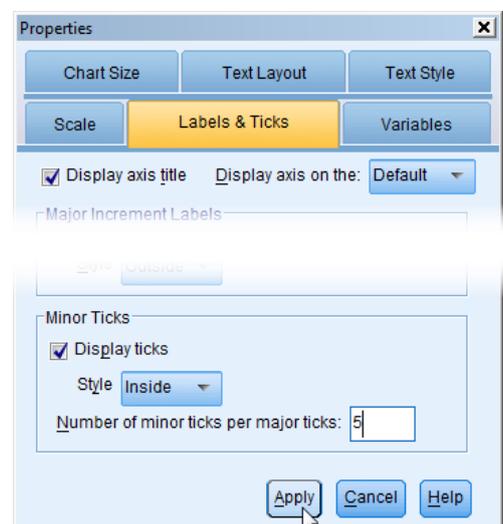
8. Select the **Scale** menu →
9. To change the number of values on the **Y-axis** change the **Major increment** to '5' → (the default is '10').



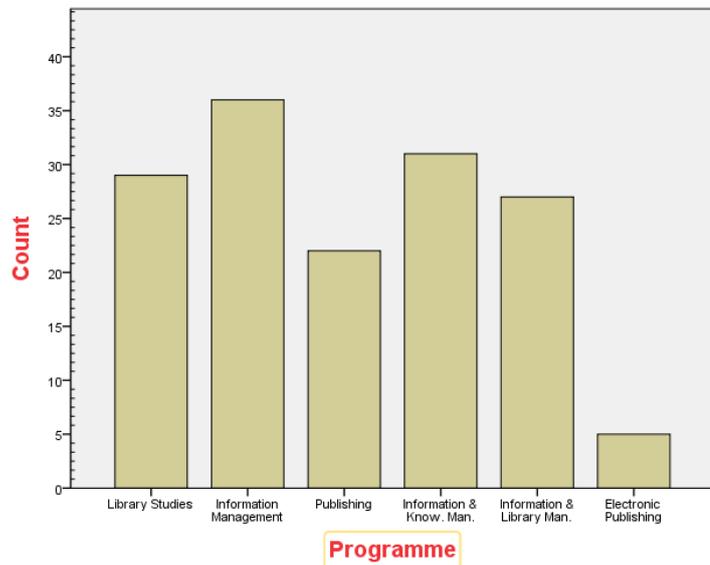
10. Select the **Labels & Ticks** menu →

[WARNING CONCERNING STEP 11:
 INSERTING MINOR TICKS DOES NOT SEEM TO WORK
 IN SPSS 19.0 ALTHOUGH IT DID IN SPSS 18.0.
 IT IS INCLUDED HERE FOR COMPLETION –TRY IT – THE
 BUG MAY BE FIXED SOON!]

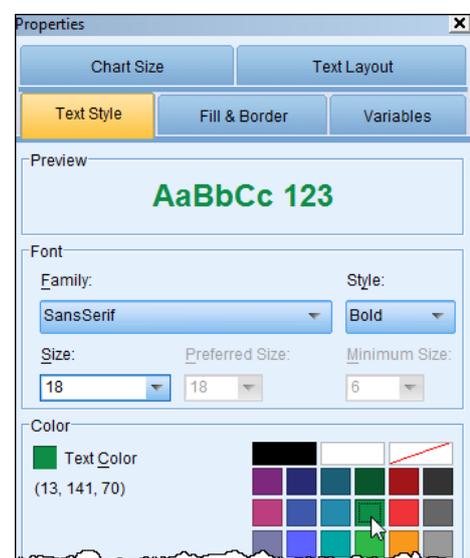
11. To insert extra ticks in the **Y-axis** select the **Minor Ticks Display ticks** box →
 - Change the **Style** to 'Inside' →
 - Set **Number of ticks per major tick** to '5' →
12. Click **Apply** to confirm these changes →



13. To edit the X-axis label ('Programme') in the same way: click on the label to select it.
 - ▶ 'Programme' should be surrounded by a pale yellow rectangle.
 - ▶ The relevant **Properties** window should still be open (if not, use **Edit → Properties**).
14. Select the **Text Style** menu.
15. To change the font size of the label, click on the **Size** drop-down menu and choose '16'.
16. To check the style of the label, click on **Style** drop-down menu (default 'Bold').
17. To change the colour of the label, click on a colour in the **Color** palette (e.g. bright red).
18. Click **Apply** to confirm these changes.
 - ▶ The chart should by now look like this →
 - ▶ Note the minor inside ticks which you may not be able to achieve with SPSS 19.0



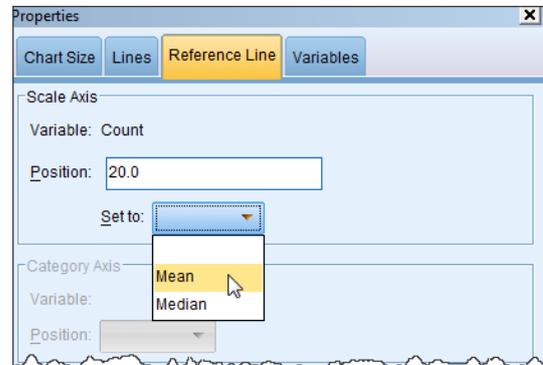
19. Select **Options → Title**
 - ▶ The word 'Title' should appear centrally above the chart (drag away the **Properties** window to see it, if necessary).
20. Select the word 'Title' by dragging across it and change it to 'Students on Programmes'.
21. Double-click on the title, if necessary, to open the **Properties** window shown below.
21. Using the **Properties** option **Text Style** to change the font size from 'Automatic' to '18' and colour to bright green.



22. Select **Options** → **Y-axis reference line**.

- ▶ This inserts a horizontal line through half way up the scale.
- ▶ The inserted line should still be highlighted (yellow rectangle round it), if not click on it.

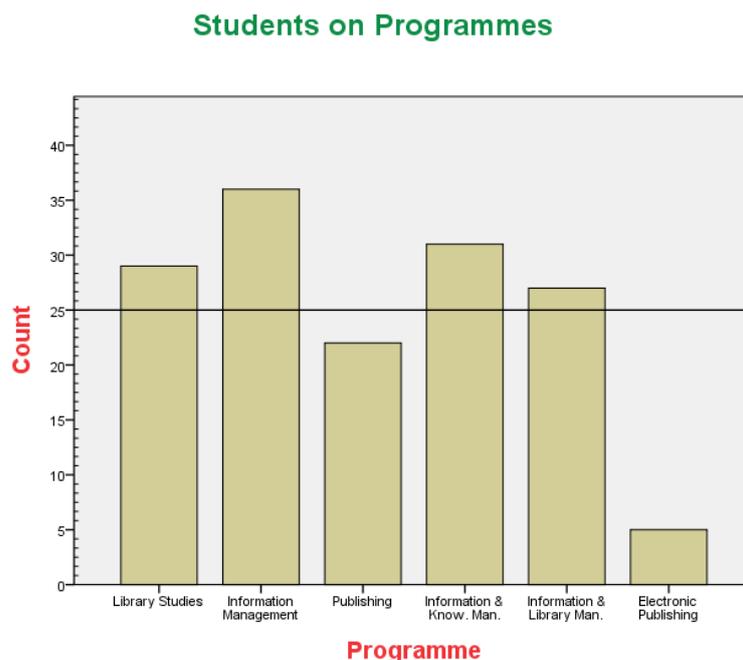
- ▶ The **Properties** window should appear like this →



- ▶ The reference line **Position** is 20.0 →

23. To move the reference line to indicate the mean use **Set to** 'Mean' as here →

24. Click **Apply** to produce the chart below:



25. To add a small text box to the explain what the reference line represents use **Options** → **Text Box**

- ▶ The word 'Textbox' will appear in a yellow rectangle.

26. Drag the yellow box down lower if it is obscured by the title. Select the word 'Textbox' and change it to 'MEAN'.

27. Reduce the size of the box using the 'handles':



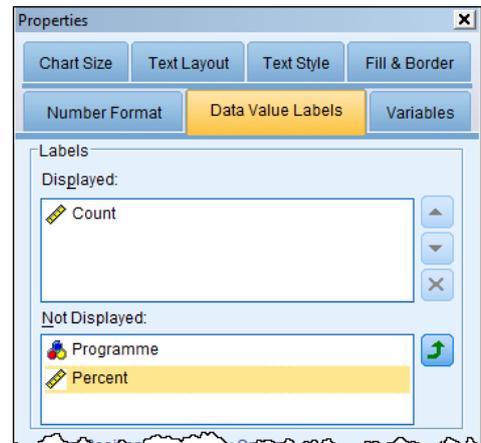
28. Drag it into position on the reference line and click elsewhere to embed it.

29. To add markers to the individual bars to show their values, click on a bar (so that all the bars are highlighted with pale yellow rectangles) and select **Elements** → **Show Data Labels**.

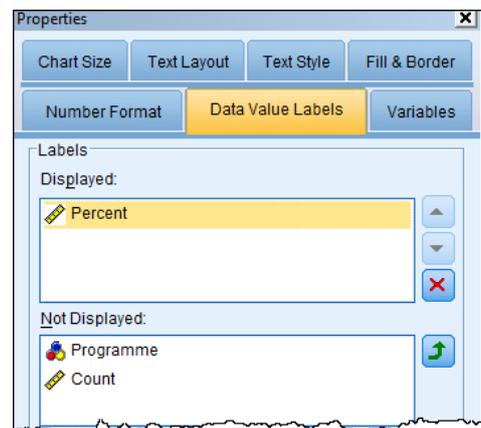
▶ A new Properties window will appear →

▶ The default is 'Count' →

30. To have bar markers showing 'Percent' rather than 'Count' in the **Properties** window **Data Value Labels** view select 'Percent' →

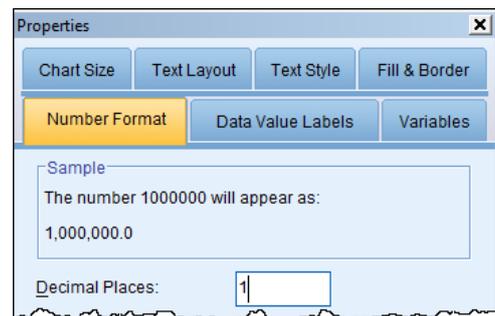


31. Move 'Percent' up to the **Displayed** box (use the green arrow) and remove 'Count' (use the red cross) to produce this →

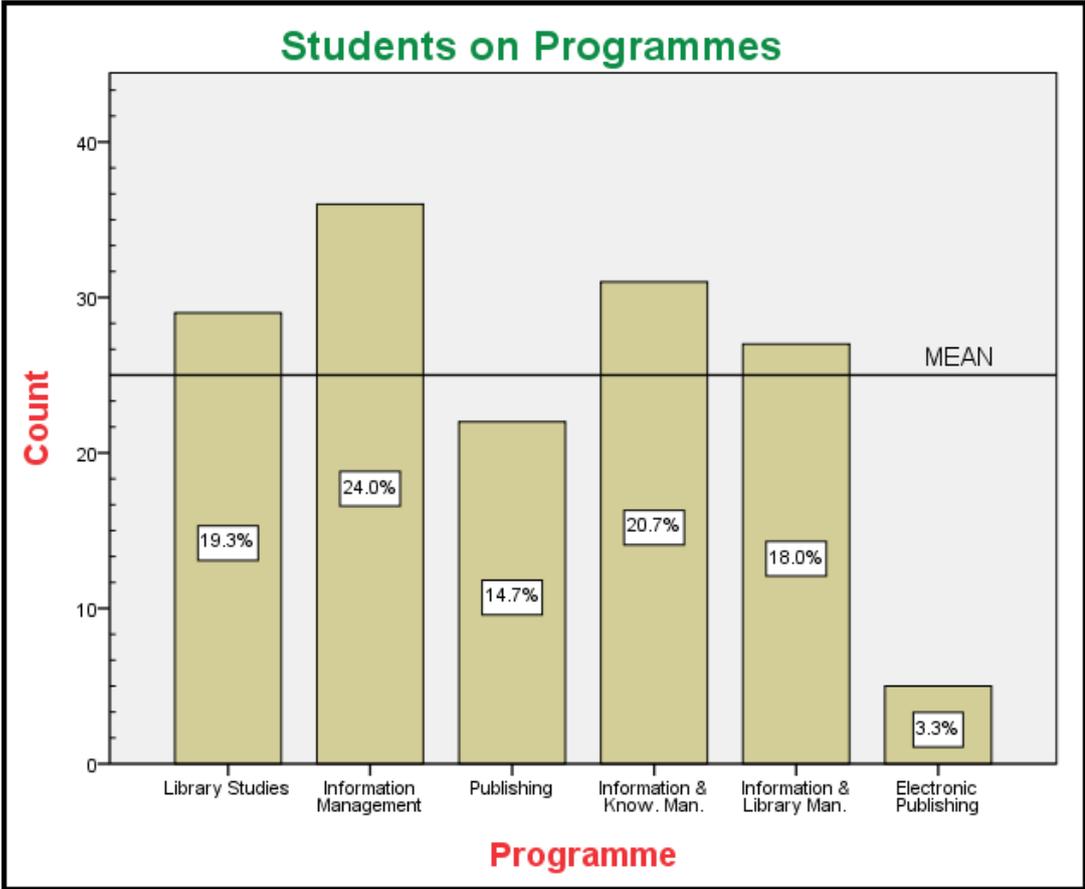


32. As the default decimal places setting will be '2', which is too many, change it to '1' by selecting the **Properties** window **Number Format** view and entering '1' in the Decimal places box (Note: it will initially be blank rather than showing '2').

33. Click **Apply**

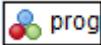


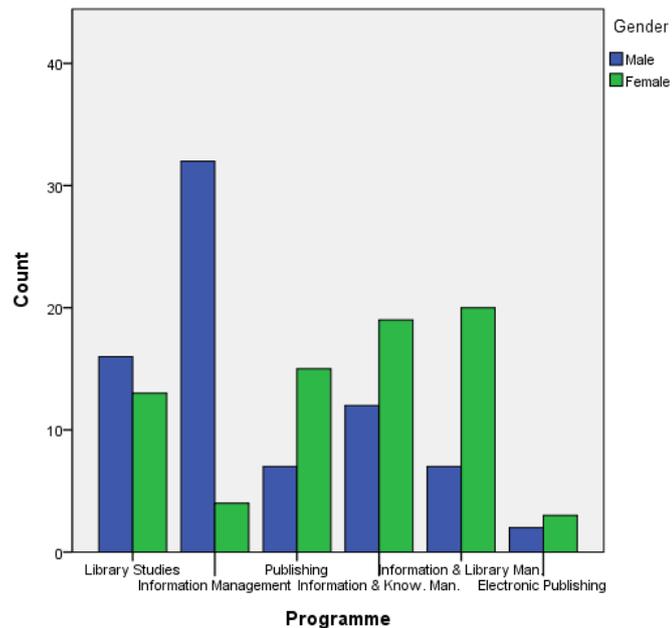
34. Finally, click on the **Chart Editor** window's close box to shut the **Chart Editor** window and embed the chart into the **Viewer** Output window. The next page shows how it should look.

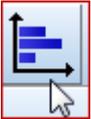


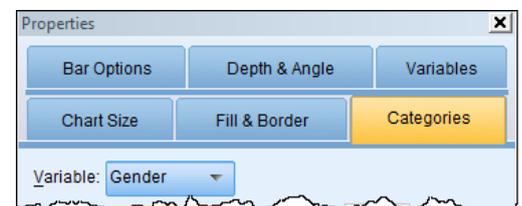
TUTORIAL T14: Clustered Bar Chart

T14.1 Clustered Bar Chart – two variables

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...** and click on **Reset** if any chart is already selected.
3. Click on **Gallery** (if not selected) and click on **Bar**.
4. Drag the 'Clustered Bar' (second icon across) from the **Gallery** into the **Chart Preview** box.
5. Close the **Element Properties** window (if it opens).
6. Locate **prog** in the **Variables** list and drag it across to **X-Axis?** 
7. Locate **gender** in the **Variables** list and drag it across to **Cluster on X: set color** (top right corner).
8. Click **OK** to generate a Clustered Bar Chart →



9. Double-click anywhere on the chart to open the **Chart Editor** window.
 - ▶ This may also open the **Properties** window. If not, select **Edit** → **Properties**.
 - ▶ Note that the labels for the programmes are a bit jumbled because they are so long.
 - ▶ This could be overcome by transposing the chart by clicking the Toolbar icon →  but this will not be done just now (you can try it, if so click again to undo it).
10. Click on one of the bars to select all the bars.
 - ▶ A pale yellow rectangle will surround each bar.
 - ▶ The **Properties** window will change to this →
11. Select the **Bar Options** view and change the width of Bars to '80' and the width of Clusters to '75'. Either use the sliders or type in the numbers.



12. Click **Apply** if using the sliders, or if typing in the numbers you can press **Enter** instead.

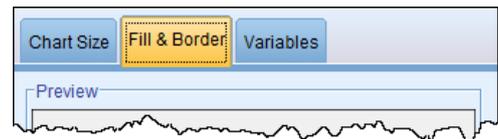
▶ This will slim down the bars and slightly separate the bars for the males and females.

13. Click on the blank area inside the chart rectangle to select the whole 'inner frame'.

▶ A pale yellow rectangle will surround the 'inner frame' containing the bars, the axis labels and the legend.

14. Select **Edit** → **Properties** (if not open already).

▶ The context-specific **Properties** window like this will open →

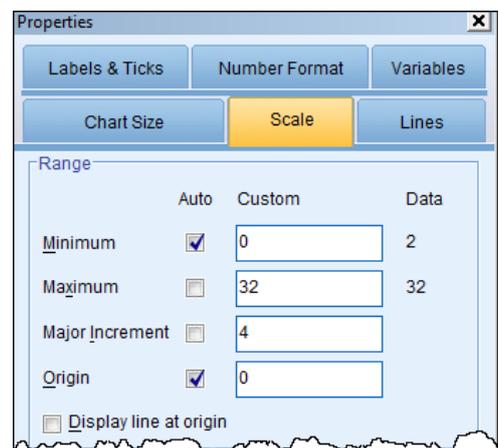


15. Ensure **Fill & Border** view is selected and click on the small **Color Fill** box and then click on 'yellow' for the background.

16. Click **Apply** to fill the background with yellow.

17. In the **Chart Editor** select **Edit** → **Select Y-Axis** or click on the **Y** icon in the **Edit** toolbar.

▶ This **Properties** window should appear → (If not, select **Edit** → **Properties**).



18. Select **Scale** view (if not selected already).

19. Change the **Range Maximum** to '32' and the **Major increment** to '4'.

20. Click **Apply**.

▶ This heightens the bars a little and puts more ticks on the Y-axis, to aid reading off values.

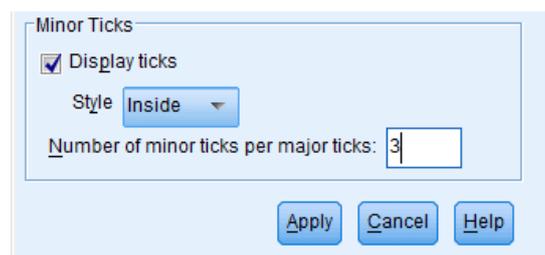
[WARNING CONCERNING STEPS 21 to 23 BELOW: INSERTING MINOR TICKS DOES NOT SEEM TO WORK IN SPSS 19.0 ALTHOUGH IT DID IN SPSS 18.0. IT IS INCLUDED HERE – TRY IT – THE BUG MAY BE FIXED SOON!]

21. Select **Labels & Ticks** view →



22. To insert extra ticks in the **Y-axis**:

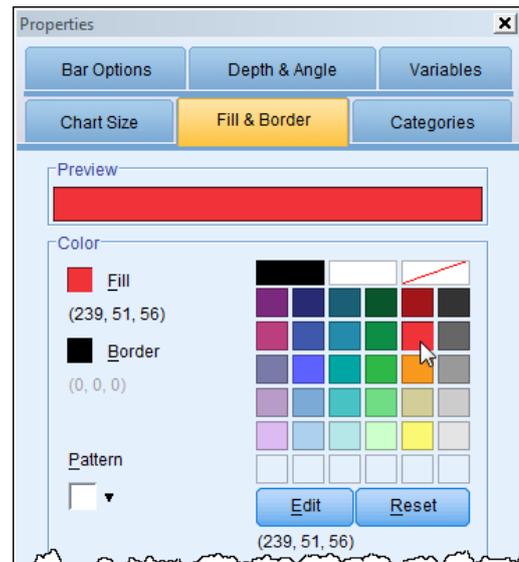
Select the **Minor Ticks Display ticks** box →
Change the **Style** to 'Inside' →
Set **Number of ticks per major tick** to '3' →
(i.e. one less than the **Major increment**).



23. Click **Apply**.

24. Click on the **Gender Legend's** Male small coloured box to select it (a pale yellow rectangle will appear round it).

25. Ensure that the **Properties** window is open, by selecting **Edit → Properties** (in the **Chart Editor**) if necessary, and choose the **Fill & Border** view which should appear like this →



26. Change the **Color Fill** to bright red.

27. Click **Apply**.

- ▶ This will change the Male bars to red.
- ▶ If the **Gender Legend** is not well-placed it can be moved by selecting it and dragging.

This is quite tricky! You need to double-click below the word 'Female' to get a yellow rectangle enclosing the whole of the Legend. (This may take a few attempts!) Then move the cursor around slowly until the drag icon appears - then drag it.

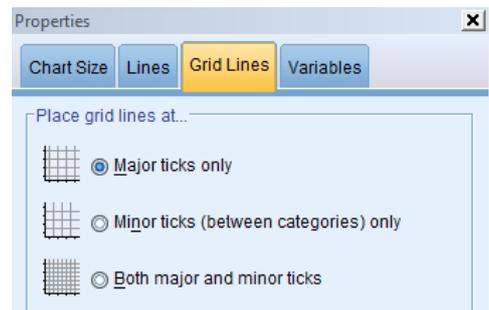
28. In the **Chart Editor** select **Edit → Select Y-Axis**.

- ▶ This highlights the Y axis again (a yellow rectangle will appear round it).

29. Select **Options → Show Grid Lines**.

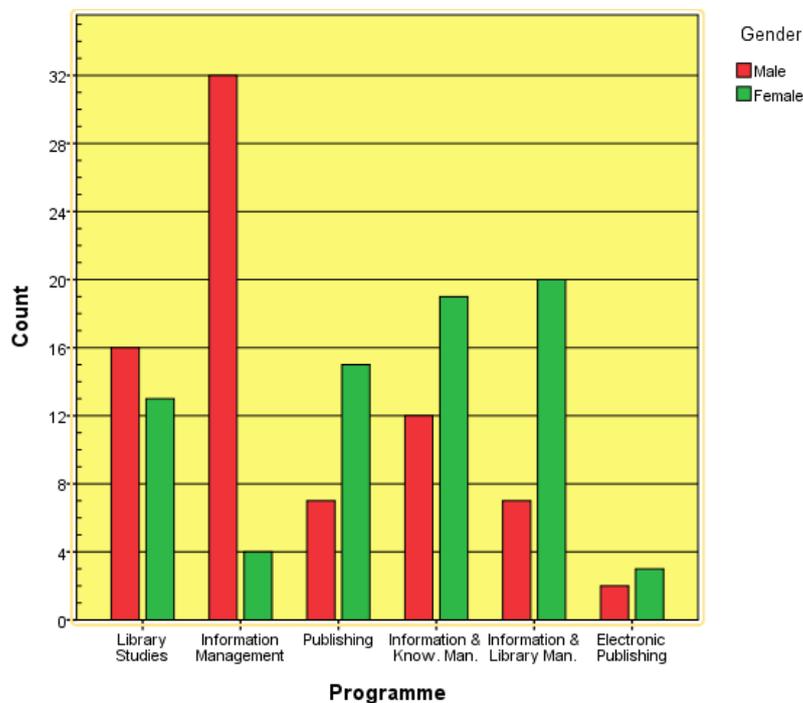
30. Select **Grid Lines** view.

- ▶ The default 'Major ticks only' is wanted → (you could try another option to see the effect)
- ▶ Just horizontal grid lines will appear as only the Y axis was selected.



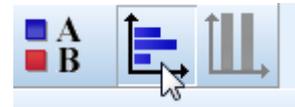
31. Click **Close**.

- ▶ The chart should appear as below:

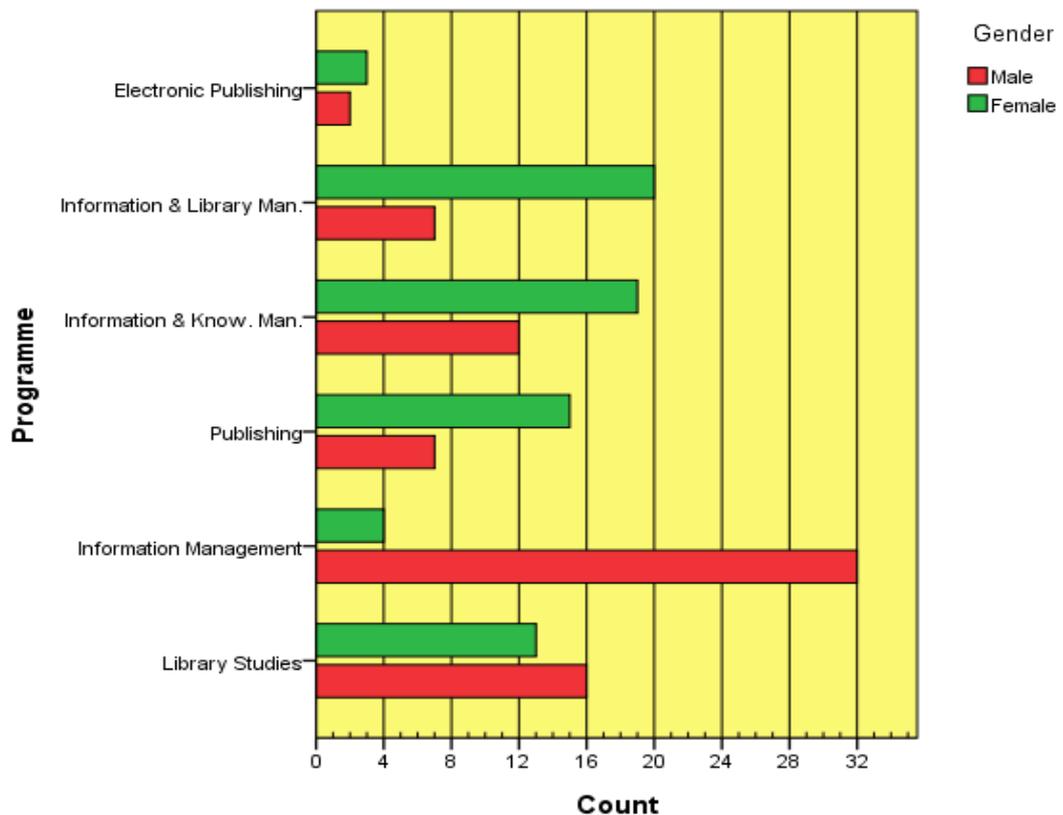


32. As the programme names are long the labels may be squashed up and if so it can be better to transpose the axes. This is very easily achieved:

- Either by selecting **Options** → **Transpose Chart**
- Or by clicking on the transpose icon in the **Options Toolbar** (to undo it just click on it again).

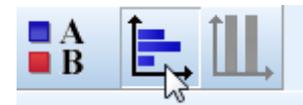


► This produces the chart, shown below:



33. There is an alternative way to improve the label legibility, which may be preferred.

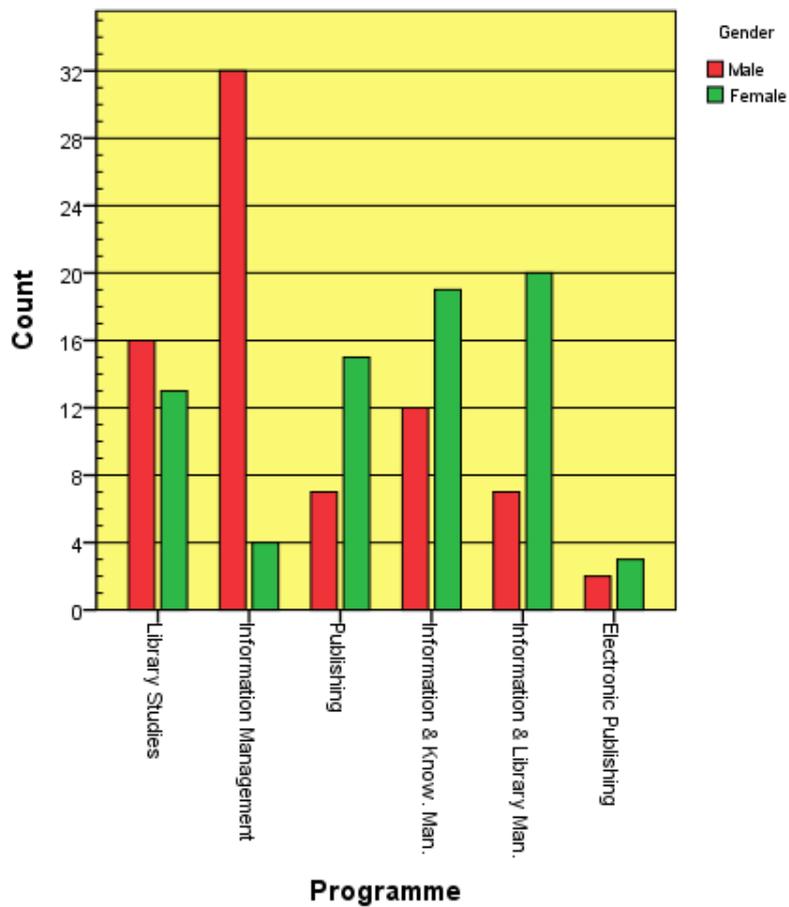
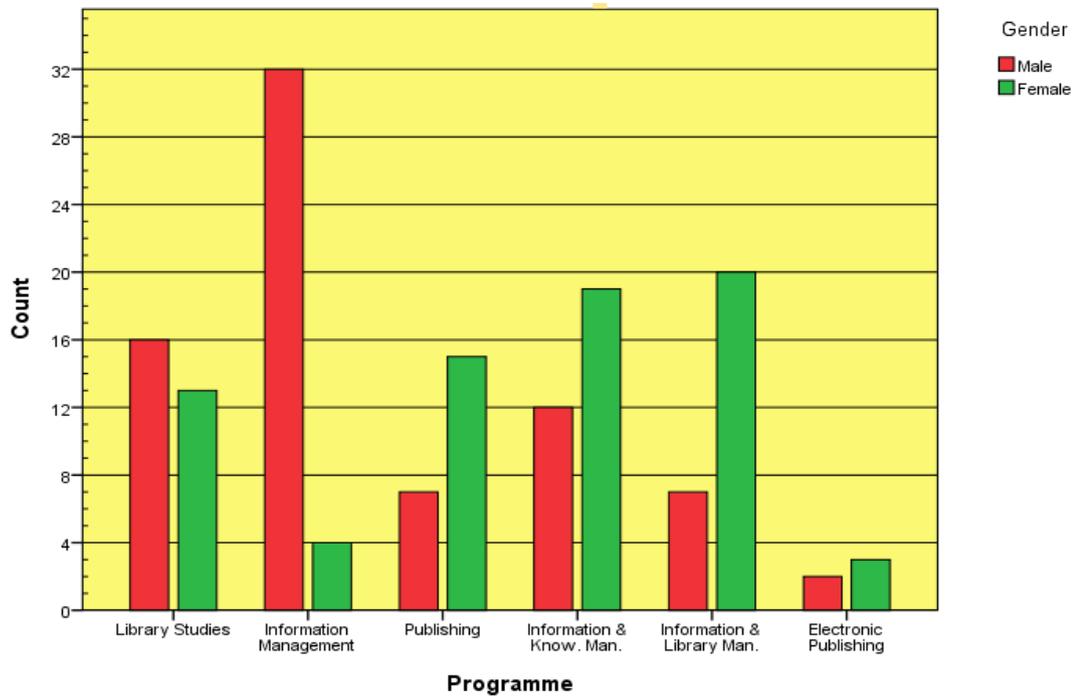
First transpose the chart to its 'upright' position by clicking on



34. Select **Edit** → **Select Chart** and close the **Properties** window (if it is still open).

- A pale yellow rectangle should enclose the chart, with corner and side 'handles' which can be dragged to resize the chart.
- Alternatively, close the **Chart Editor** to embed the chart in the **Output** window and click on the chart and resize it there.

35. Either drag a handle to the right to stretch the chart and separate the labels or drag to the left to narrow the chart and force the labelling to go sideways. The two effects are shown below:

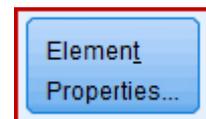


T14.2 Clustered Bar Chart – three variables

Using the data from T14.1, the question, whose responses are to be illustrated here, asked students if they would be happy to miss lectures if lecture notes were available on the VLE.

1. Select **Graphs** → **Chart Builder...** and click on **Reset**.
2. Click on **Gallery** (if not selected) and click on **Bar**.
3. Drag the 'Clustered Bar' icon from the **Gallery** into the **Chart Preview** box.
4. Locate **miss_lectures** in the **Variables** list and drag it across to the **Y-Axis**.

- ▶ The **Element Properties** window should appear, if not, open it by clicking on the button on the right side of the window →

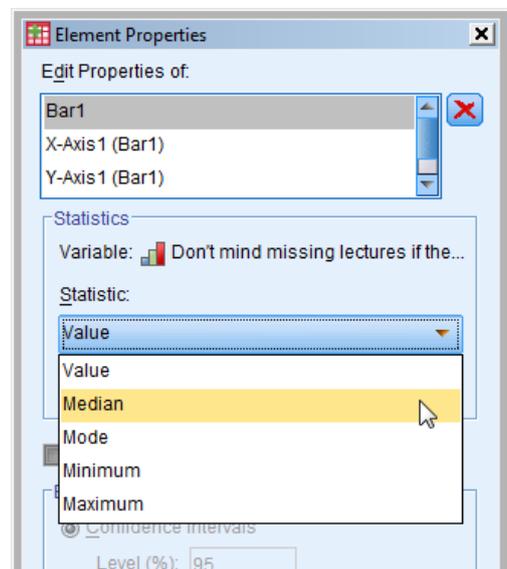


5. In the **Element Properties** window change the **Statistic** from 'Value' to 'Median'.

- ▶ Because **miss_lectures** is an ordinal variable – a 5-point question response scale – the median and not the mean is used as the 'average'.
- ▶ If **miss_lectures** were declared to be a scale variable then the mean would be available.

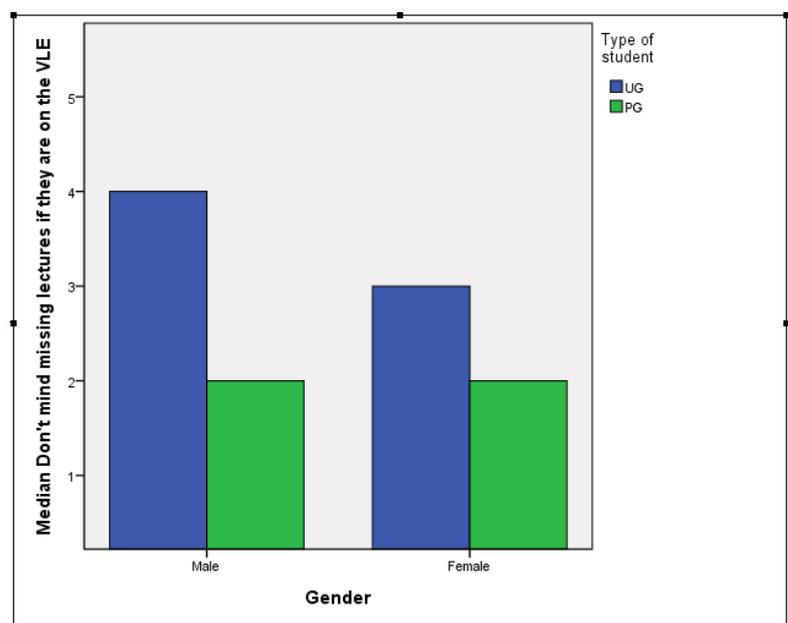
6. Click **Apply**.
7. Click **Close**.

8. Locate **gender** in the **Variables** list and drag it across to the X-axis.



9. Locate **ug_pg** in the **Variables** list and drag it across to **Cluster on X: set color**.

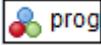
10. Click **OK** to generate the Clustered Bar Chart here which shows response differences between males and females, and between the undergraduates and postgraduates.

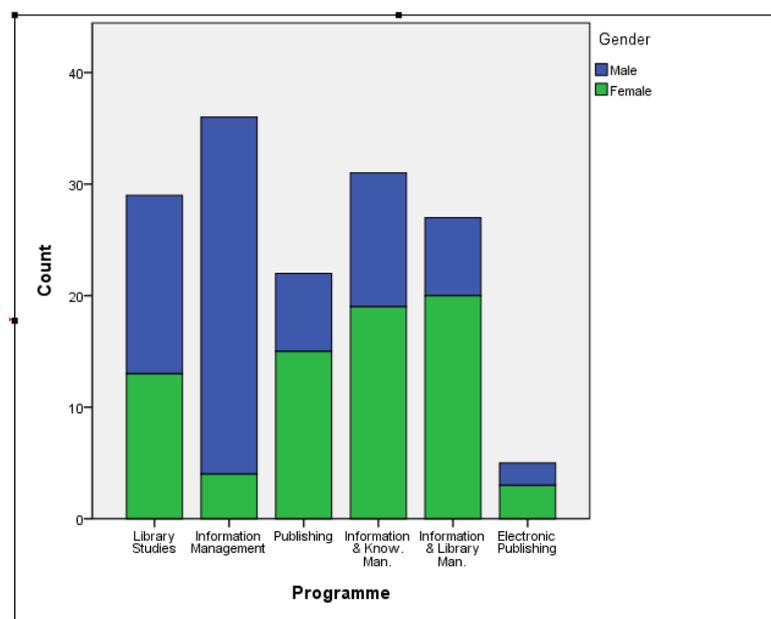


TUTORIAL T15: Stacked Bar Chart

T15.1 Stacked Bar Chart - basics

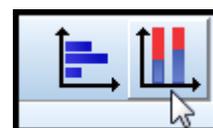
If you are continuing straight on from T14 you can skip straight to Step 2.

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...** and then click on **Reset** if continuing from T14.
3. Click on **Gallery** (if not selected) and click on **Bar**.
4. Drag the 'Stacked Bar' (third icon across) from the **Gallery** into the **Chart Preview** box.
5. Close the **Element Properties** window (if it opens).
6. Locate **prog** in the **Variables** list and drag it across to **X-Axis?**. 
7. Locate **gender** in the **Variables** list and drag it across to **Stack: set color**.
8. Click **OK** to generate a basic Stacked Bar Chart:



The changes made to the Simple Bar Chart (T12, T13) and Clustered Bar Chart (T14) apply also to the Stacked Bar Chart, which we will not illustrate here. We will only make one (new) change.

9. Double-click on the chart to open the **Chart Editor** window, then
 Either select **Options** → **Scale to 100%**
 or click the **Options Toolbar** icon →



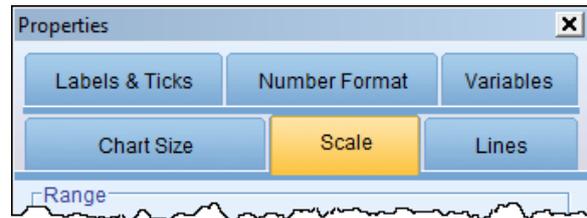
- ▶ The result will be a Percentage Stacked Bar Chart
- ▶ To undo this, and revert to the normal Stacked Bar chart, either select **Options** → **Scale by value** or click the **Options Toolbar** icon again. (Try this and then revert to the Percentage Stacked Bar Chart before continuing.)

- Before finishing, we will remove the unnecessary decimal places which will appear (by default) on the Y-axis.

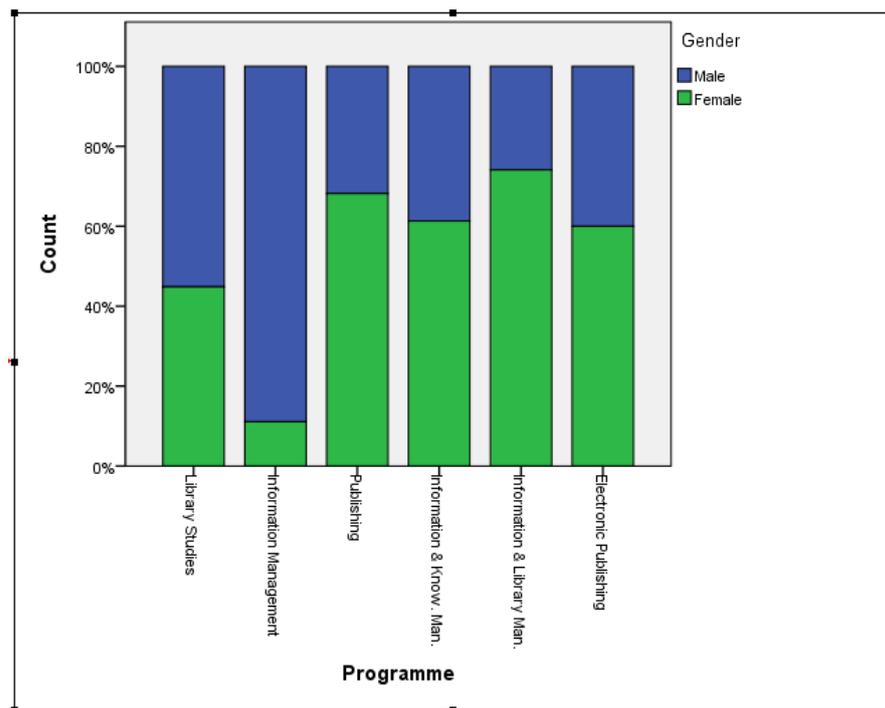
Do this by

Edit → **Select Y-Axis** which will open this **Properties** window →

and select **Number Format** view and insert a '0' in **Decimal Places**.



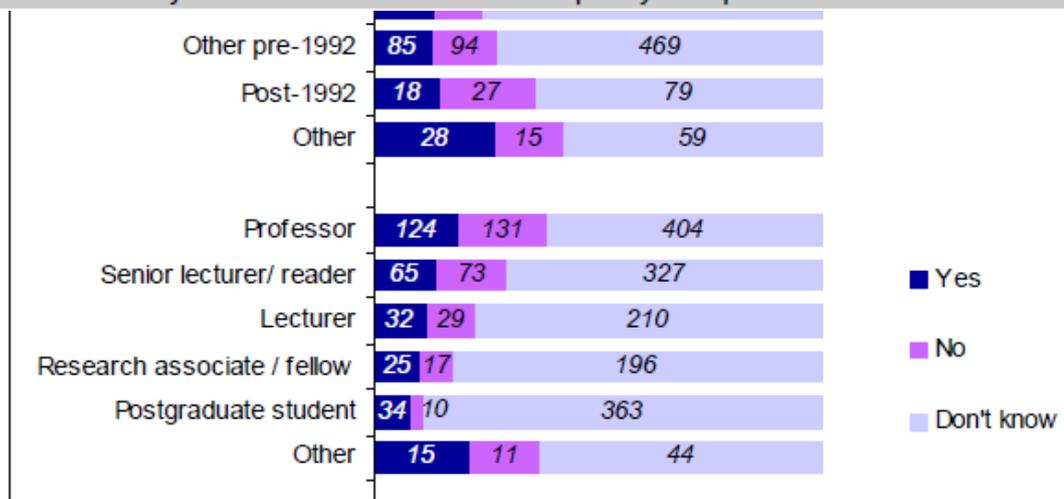
- Click **Apply** to produce this chart:



T15.2 Stacked Bar Chart - advanced

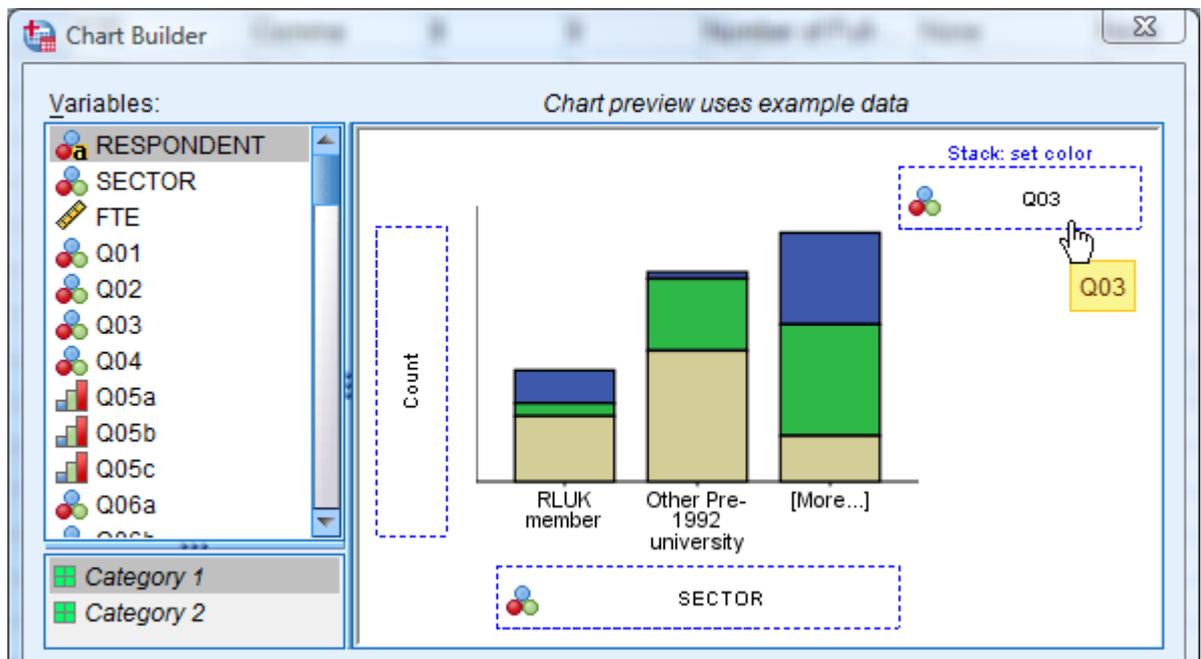
The LISU/SQW consulting Open Access Report for RCUK (see Appendix) includes this double chart:

Figure B-3 Does your institution have a written policy on open access to research outputs?



Here we will explore how to produce a chart to closely resemble one of these, using the supplied subset of the Open Access data. (As it is a subset of the data, the chart will not be exactly the same, but, provided the subset is representative, it should be similar.)

1. Load data file: **File** → **Open** → **Data** → **DATASET5_OpenAccess_Researchers.sav**
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.
2. Select **Graphs** → **Chart Builder...**
3. Click on **Gallery** and click on **Bar**.
4. Drag the 'Stacked Bar' icon from the **Gallery** into the **Chart Preview** box.
5. Close the **Element Properties** window (if it opens).
6. Locate **SECTOR** in the **Variables** list and drag it across to **X-Axis?**.
7. Locate **Q03** in the **Variables** list and drag it across to **Stack: set color**.
8. Click **OK** to generate the Stacked Bar Chart below.
 - ▶ The Y-axis will show 'Count' as the variable (i.e. the frequency).



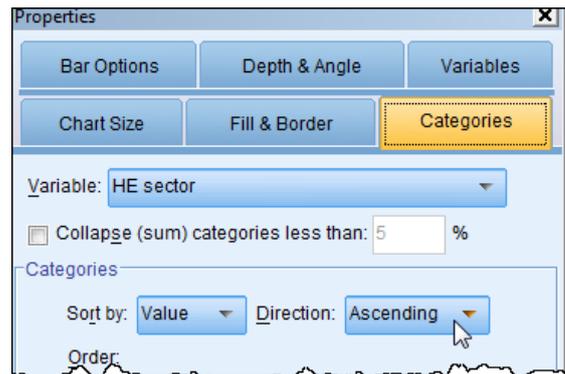
9. Click **OK**.
 - ▶ This places the chart in the **Viewer** Output window.
10. Double-click on the chart to open **Chart Editor**.
 - ▶ Close the **Properties** window which appears.

11. Click on a bar in the chart to select all the bars.

- ▶ This **Properties** window should open → but may have a different view selected.
- ▶ If the window fails to open, or is accidentally closed, first ensure that the bars are selected (pale yellow edge round them) and then use **Edit → Properties**.

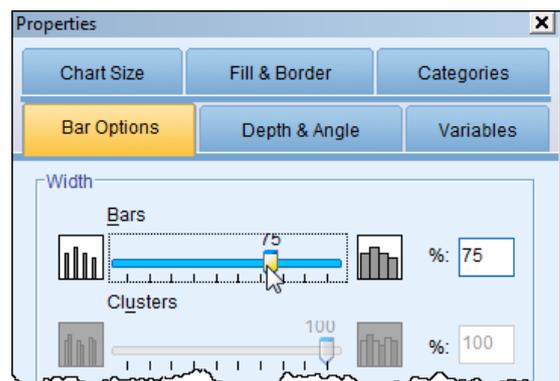


12. Select the **Categories** view →



13. Change sort **Direction** from **Ascending** to **Descending** →

14. Click **Apply** and note the effect.



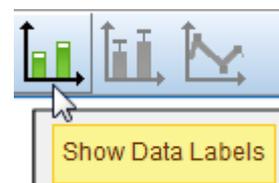
15. Select the **Bar Options** view →

16. Change **Bars** Width from '75' to '40', either using the slider or by typing in the number.

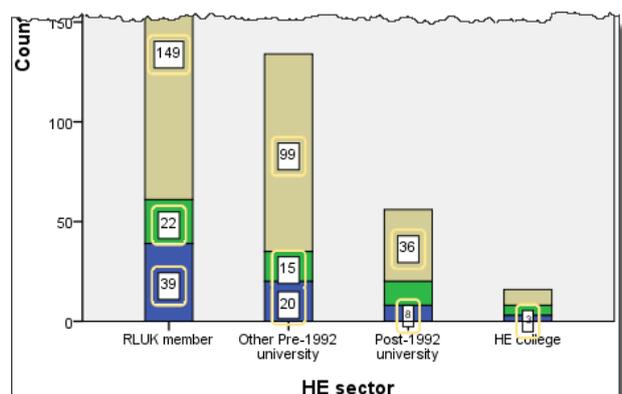
17. Click **Apply** and note the effect.

18. Click on a bar in the chart to select all the bars

19. Either **Edit → Elements → Show Data Labels** or click on →



- ▶ This produces this effect →



20. Click outside the bars to embed the frequencies in the bars of the chart.

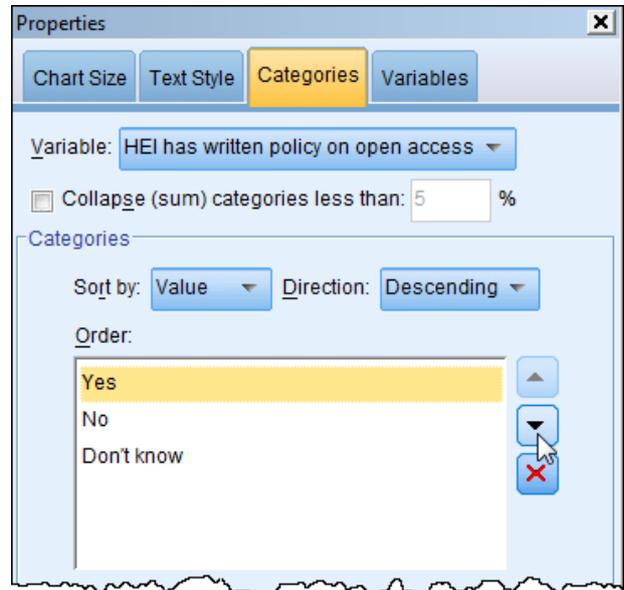
21. Click on the **Legend** to highlight the words next to the three squares →



► This opens this **Properties** window →

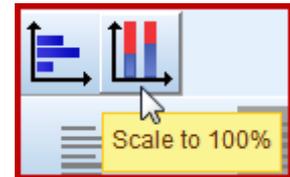
22. Ensure that the **Categories** view is selected and reverse the **Order** of the words by selecting 'Yes' and moving it down to the bottom row using the arrow, then move 'Don't know' down to the middle.

► It is tempting to try to do it all in one go by changing **Direction** from **Descending** to **Ascending** but that reverses the order of the regions within the bars.



23. Click **Apply** and note the effect.

24. Either select **Options** → **Scale to 100%** or click on the icon:

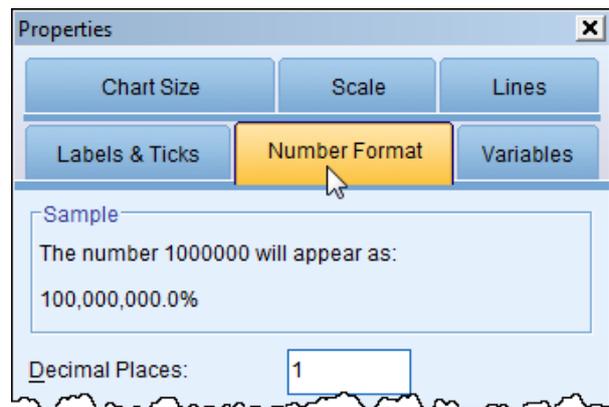


25. This converts the chart to percentages. Remove the superfluous decimals → either by selecting **Edit** → **Select Y Axis** or by clicking on the **Y** icon.

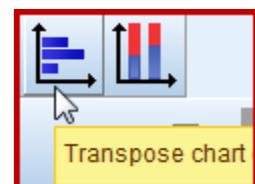
► This opens this **Properties** window →

26. Ensure that the **Number Format** view is selected and insert a '0' in the **Decimal places** box.

27. Click **Apply** and note the effect.



28. Either select **Options** → **Transpose Chart** or click on the icon:



29. Close the **Properties** window and check the chart. You will see that order of the bars is now the opposite of what it was (**HE college** was last but now is first). To rectify this either select **Edit** → **Select X Axis** or click on the **X** icon.

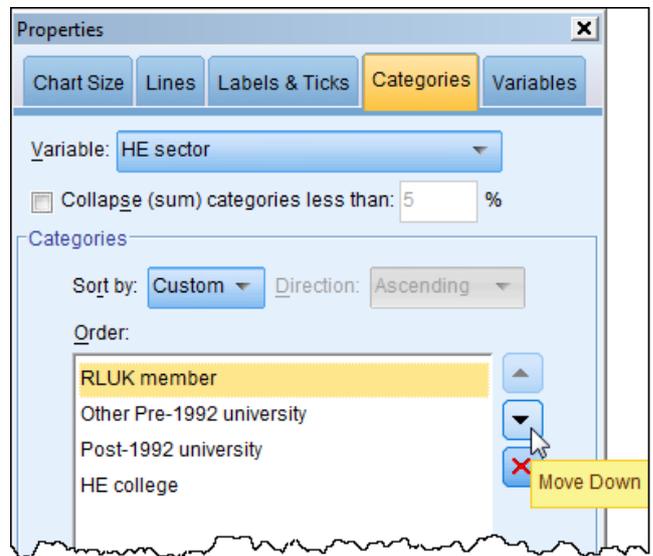
► This opens this **Properties** window →

30. Ensure that the **Categories** view is selected and reverse the **Order** of the labels by selecting 'RLUK member' and moving it down to the bottom rows using the arrow, then move the other labels down, to complete the reversal.

31. Click **Apply** and note the effect.

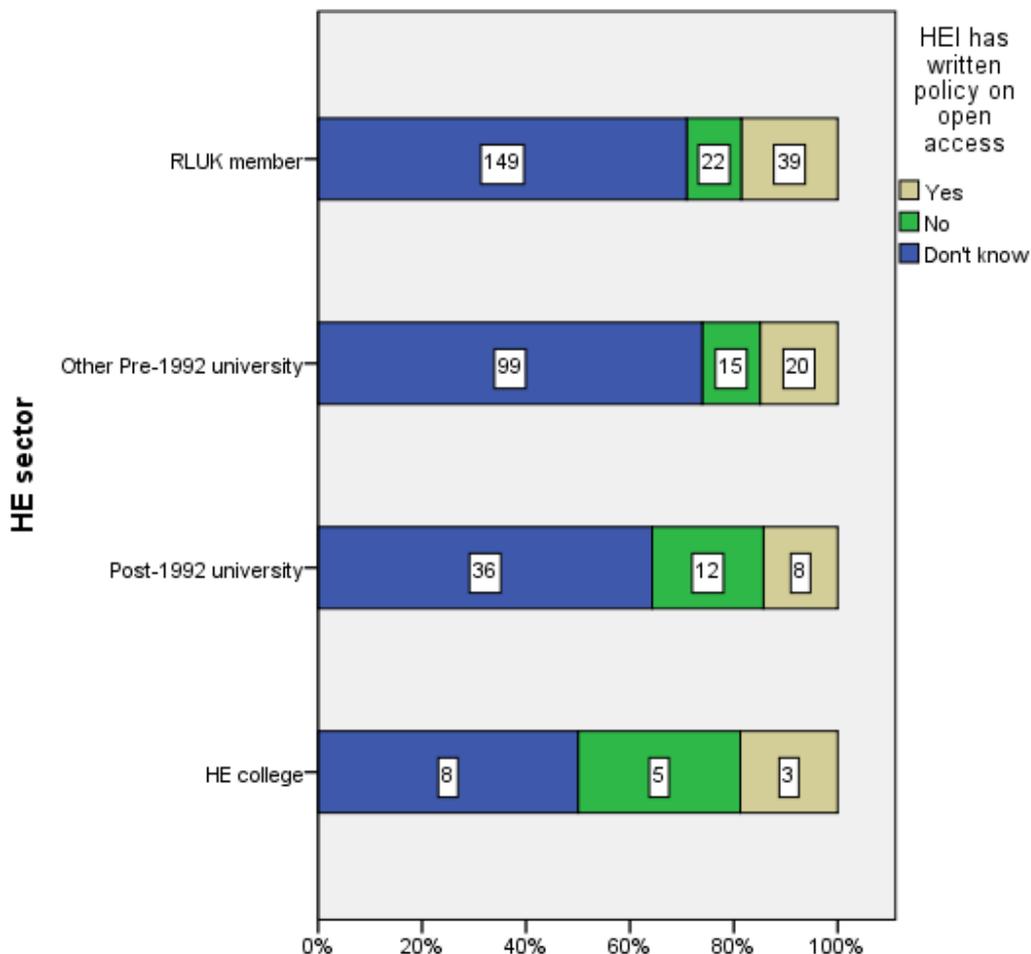
32. Click **Close**.

33. Close the **Chart Editor** to embed the chart in the **Viewer** Output window.



► This produces the chart below.

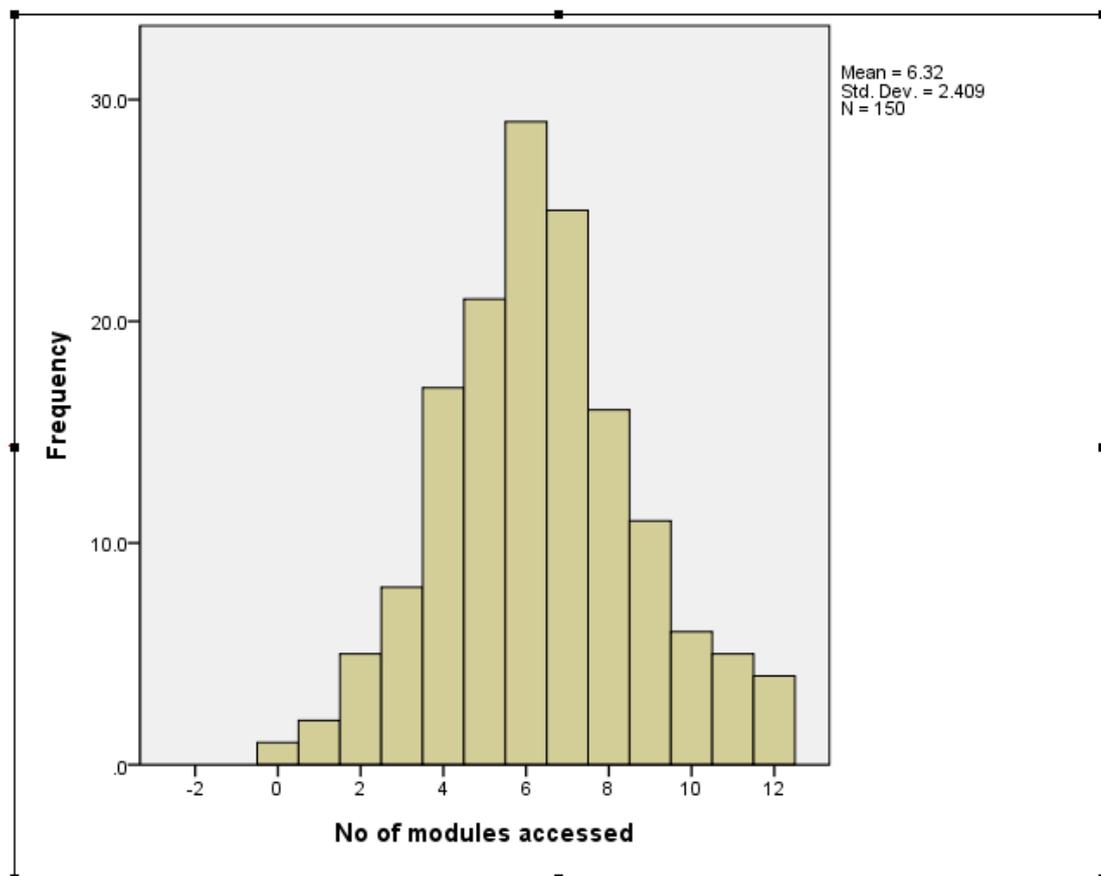
► Further editing could be done to more closely mimic the LISU/SQW consulting chart shown at the beginning.



TUTORIAL T16: Histogram

T16.1 Simple Histogram

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...** and click on **Reset**.
3. Click on **Gallery** (if no highlighted) and click on **Histogram**.
4. Drag Drag the 'Simple Histogram' (first icon) from the **Gallery** into the **Chart Preview** box.
 - ▶ Note that the **Element Properties** window which opens shows that the **Statistic** is set to the default 'Histogram' – this means that the y-axis will show frequencies. If you change this to 'Histogram Percent' then the y-axis will show percentage instead.
5. Locate **modules** in the **Variables** list and drag it across to **X-Axis**?
 - ▶ Note the icon next to **modules** indicates that it is a scale variable.  **modules**
6. Click **OK** to generate a Simple Histogram:
 - ▶ As with Bar Charts, every element and aspect of this Histogram can be edited and further elements can be added.



7. Double-click anywhere on the chart to open the **Chart Editor** window.

8. To see the frequencies for each bin (bar) select **Elements → Show Data Labels**.

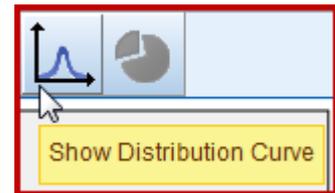
- ▶ Alternatively, click on the **Elements Toolbar** icon →
- ▶ Note that the frequencies now appear inside each bar.



9. Hide the frequencies using **Elements → Hide Data Labels**.

10. To superimpose a normal curve for comparison select **Elements → Show Distribution Curve**.

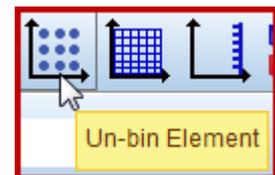
- ▶ Alternatively, click on the **Elements Toolbar** icon →



11. Hide the normal curve again using **Elements → Hide Distribution Curve**.

12. To alter the Histogram to have lines rather than bars first click on the bars and then select **Options → Un-bin Element**.

- ▶ Alternatively, click on the **Options Toolbar** icon →



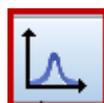
13. Re-bin the Histogram by selecting **Options → Bin Element**.

- ▶ This action should open this **Properties** window →
- ▶ If this window is not visible, click on the Histogram to highlight it and select **Edit → Properties**.

14. Ensure **Binning** view is selected, click on **Custom** → and enter '5' in **Number of intervals** →

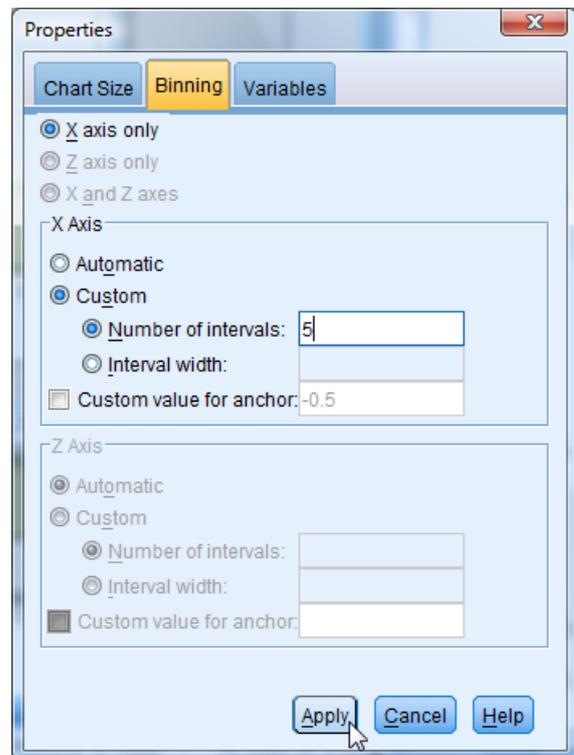
15. Click **Apply**.

16. Click on this Toolbar icon

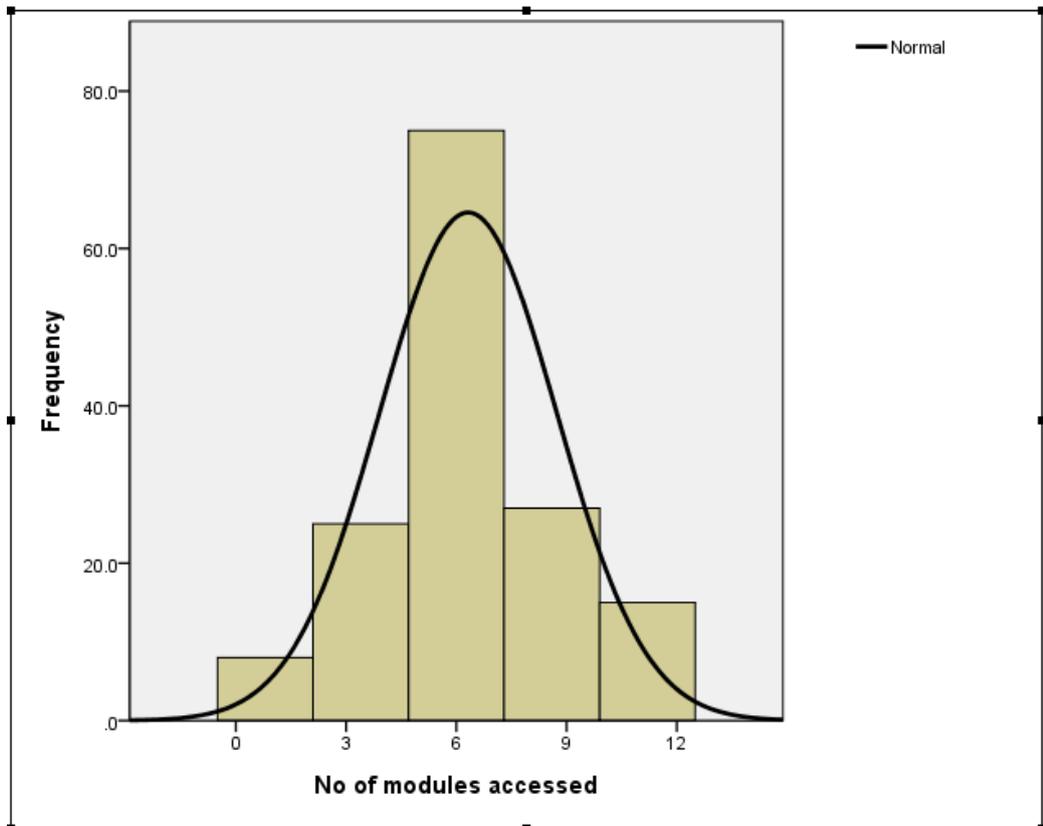


to insert a normal curve.

17. Click **Close**.



18. Close the **Chart Editor** to embed the revised Histogram in the **Viewer** Output window (see below).



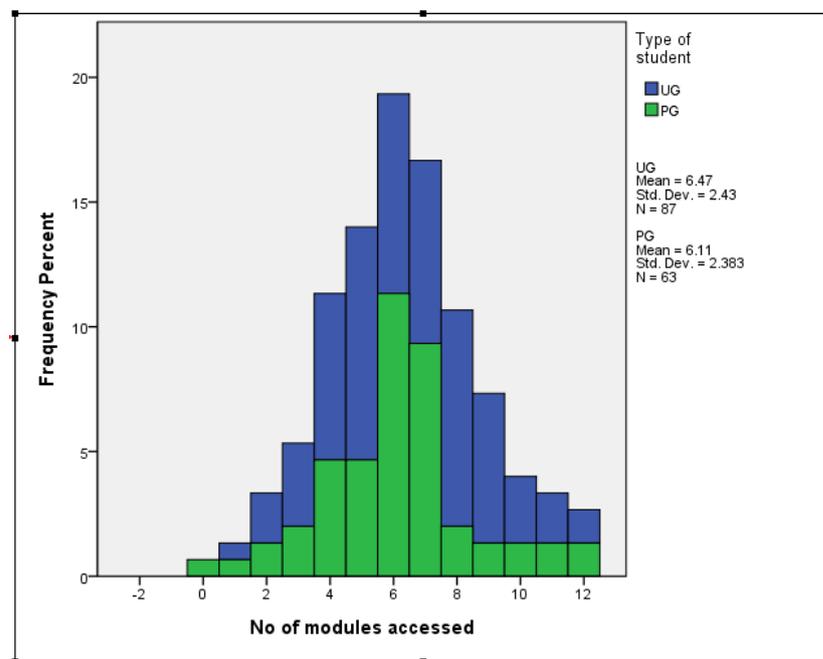
- ▶ The data analysed here has a small range of values (the integers 0 to 12). Altering the binning is more useful when there is a large range of possible values.
- ▶ If you were to enter '13' in **Number of intervals** then the result would essentially be a Bar Chart as there are 13 possible values: 0, 1, 2, ...12. (Try it.)

T16.2 Stacked Histogram

We now turn briefly to another much less well-known form of the Histogram.

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...**
3. Click **Reset**.
4. In **Gallery** click on **Histogram**.
5. Drag the 'Stacked Histogram' icon from the **Gallery** into the **Chart Preview** box.
6. In the **Element Properties** window which opens change the **Statistic** to 'Histogram Percent'.
7. Click **Apply**.
8. Locate **modules** in the **Variables** list and drag it across to **X-Axis?**.
9. Locate **ug_pg** in the **Variables** list and drag it across to the **Stack: set color** box in the **Chart Preview** box.
10. Click **OK** to produce the chart below.

► This chart shows the contributions separately for undergraduates and postgraduates.



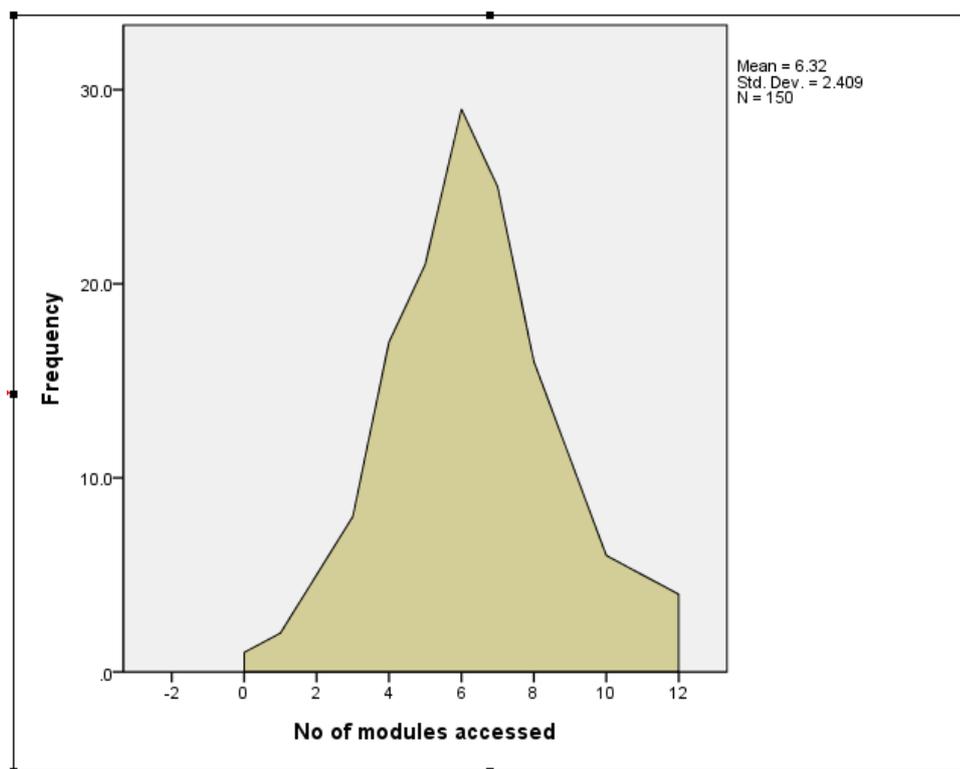
- Sometimes such a chart will reveal marked differences, but not in this case, with both types of student having broadly the same shaped distribution.
- It is equivalent to the Stacked Bar Chart for nominal and ordinal data.
- Usually the data set would have more values than in this example.

TUTORIAL T17: Frequency Polygon and Population Pyramid

T17.1 Frequency Polygon

If you like mountaineering this could be the chart for you ...

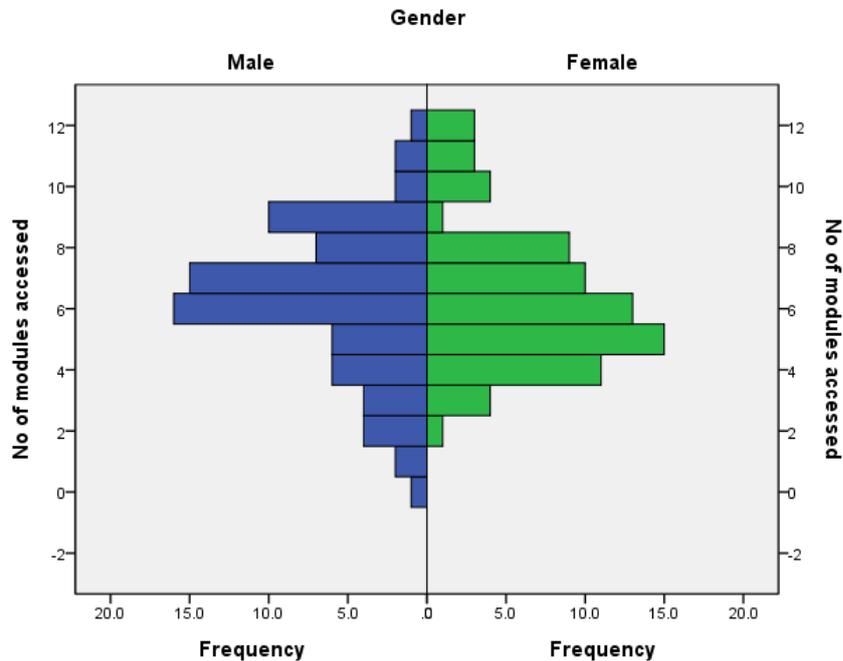
1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...**
3. Click **Reset**.
4. Click on **Gallery** (if not highlighted) and click on **Histogram**.
5. Drag the 'Frequency Polygon' icon from the **Gallery** into the **Chart Preview** box.
6. Locate **modules** in the **Variables** list and drag it across to **X-Axis?**.
7. Click **OK** to obtain the chart:



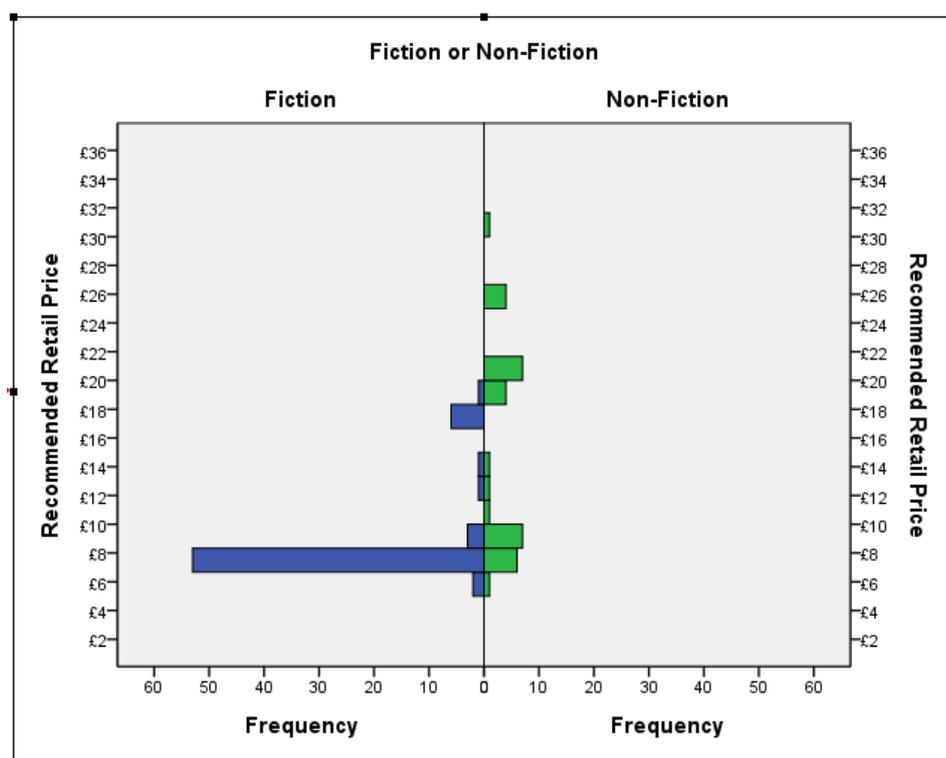
T17.2 Population Pyramid

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...** and click **Reset**.
3. Click on **Gallery** (if not highlighted) and click on **Histogram**.
4. Drag the 'Population Pyramid' icon from the **Gallery** into the **Chart Preview** box.

5. Locate **modules** in the **V**ariables list and drag it to the **D**istribution Variable? box.
6. Locate **gender** in the **V**ariables list and drag it to the **S**plit Variable? box.
7. Click **OK**.

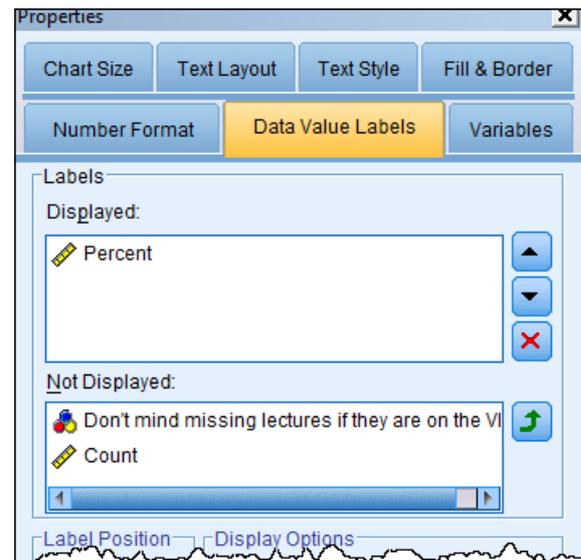


- ▶ This is essentially the Stacked Histogram met in T16.2, broken into its two constituent parts and set out sideways...
- ▶ Usually the data set would have more values than in this example.
- ▶ Below is a different looking Population Pyramid of Recommended Retail Price for the UK's 100 top-selling books, obtained from the data file: **DATA01_100Books.sav** provided to support this Guide.

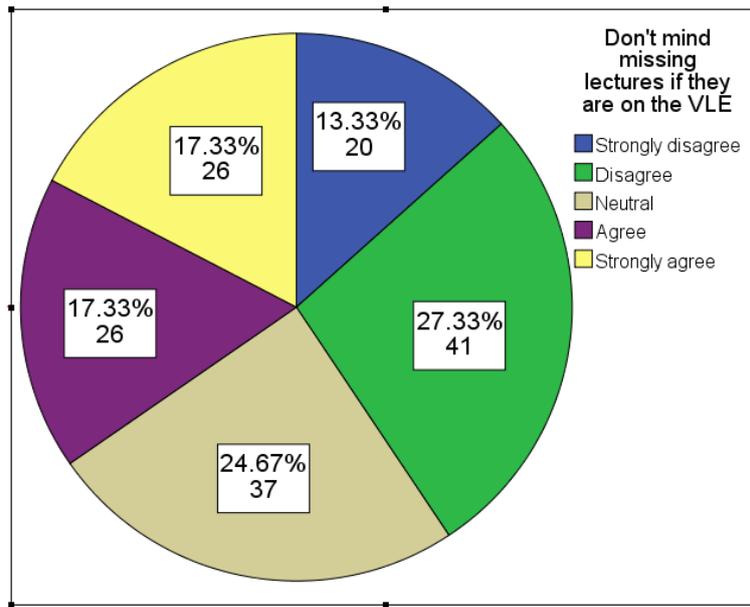


TUTORIAL T18: Pie Chart

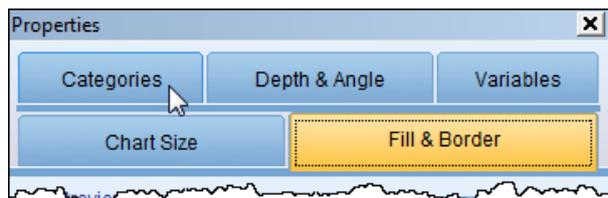
1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...** and click on **Reset**.
3. Click on **Gallery** (if not selected) and click on **Pie/Polar**.
4. Drag the 'Pie Chart' icon from the **Gallery** into the **Chart Preview** box.
5. Close the **Element Properties** window (if it opens).
6. Locate **miss_lectures** in the **Variables** list and drag it across to the **Slice by?** box.
7. Click **OK** to generate a Pie Chart:
8. Double-click the Pie Chart to enter **Chart Editor**.
9. Select **Elements** → **Show Data Labels**.
 - ▶ Percentages will appear on the slices.
10. In the **Properties** window which opens select **Data Value Labels** view.
11. Select 'Count' and move it into the **Displayed** box using the green 'up' arrow.
12. Click **Apply**.
 - ▶ Frequencies will appear on the slices.
13. Select **Number Format** view and insert '0' in **Decimal Places**.
14. Select **Text Style** view and change **Size** from 'Automatic' to '16'.
 - ▶ Scroll down to reveal '16' if hidden.
15. Click **Apply**.
16. Click on the **Legend** words (next to the coloured squares).
17. Select **Text Style** view and change **Size** from 'Automatic' to '12'.
18. Click **Apply**.
19. Click on the questionnaire words above the **Legend**.
20. Select **Text Style** view and change **Size** from 'Automatic' to '14'.
21. Click **Apply**.
22. Click to **Close** the **Properties** window.



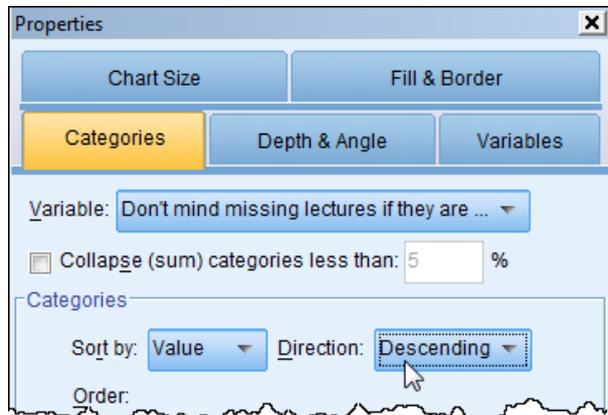
23. The Pie Chart will now appear like this:



24. To change the order of the slices select the Pie Chart by clicking on it in the middle and, if necessary, choose **Edit → Properties** to open up this window →



25. Select **Categories** view →

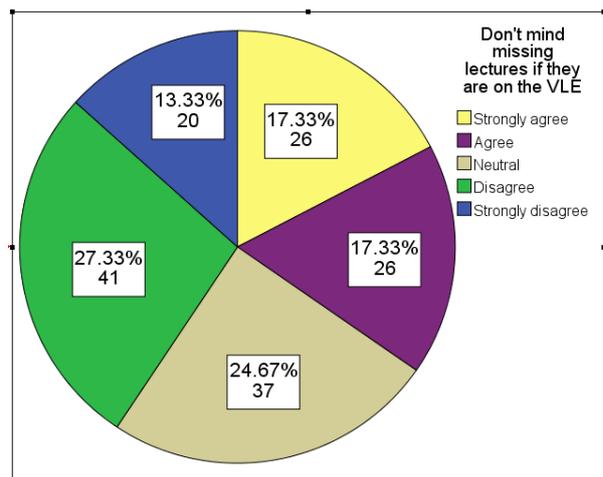


26. Select **Direction** 'Descending' →

► Note that the default **Sort by** 'Value' here means sort by the value of the codes used for the five possible responses (1 = 'Disagree strongly', etc.), not by size of slices.

27. Click **Apply**.

► This produces the chart shown here, in a more natural order for the responses.



28. It can sometimes be best to have the slices in descending order of size of slice (clockwise starting from the '12-o'clock' position), although it's not appropriate here.

Purely for illustration, to achieve this proceed as follows: in **Chart Editor** select the Pie Chart by clicking on it in the middle.

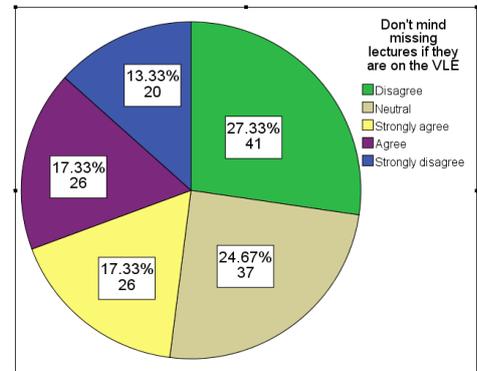
29. In the Properties window select **Categories** view.

30. Change **Sort by** from 'Value' to 'Statistic'.

▶ 'Statistic' here is 'Count'.

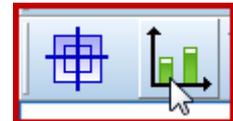
31. Click **Apply**.

▶ This produces the chart here, in an order which more easily reflects the difference in magnitudes of the slices.



32. Remove all the data labels, in **Chart Editor** select **Elements** → **Hide Data Labels**.

▶ Alternatively, click on the icon in the **Elements Toolbar** →



33. To insert individual labels click on the data label icon → in the **Elements Toolbar** and then click on individual slices.



34. To 'explode' the whole Pie Chart, first select the Chart (either click on it in the middle or use **Edit** → **Select Chart**) and then choose **Elements** → **Explode Slice**.

▶ Alternatively, click on the pie slice icon in the **Elements Toolbar** →

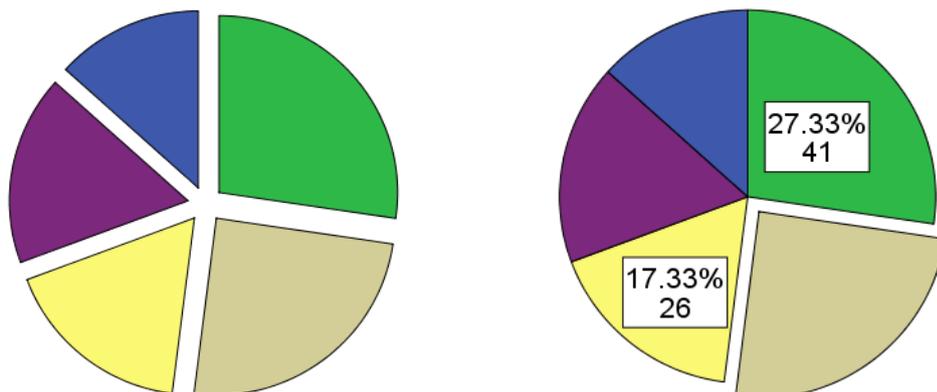


35. To put the Pie Chart back together, choose **Elements** → **Return Slice**

▶ Alternatively, click again on the pie slice icon in the **Elements Toolbar**.

36. To 'explode' one slice (e.g. the smallest), select the slice by clicking on it (ensuring it alone has a pale yellow outline round it) and choose **Elements** → **Explode Slice** (or click on the Toolbar icon).

▶ The above actions can produce effects such as those shown below:

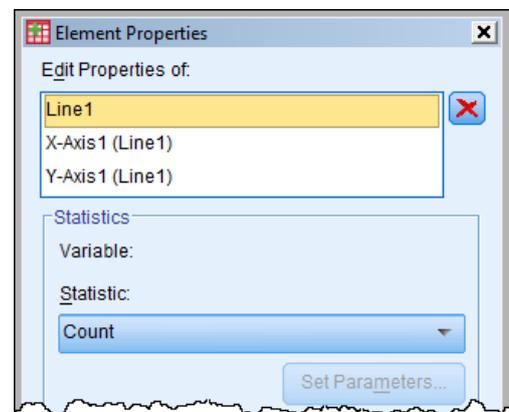


TUTORIAL T19: Line Chart

T19.1 Simple Line Chart

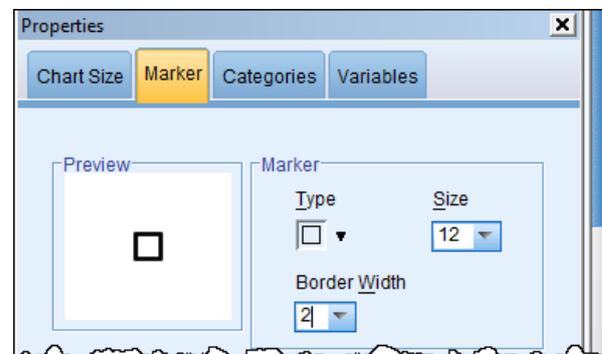
1. Load data file: **File** → **Open** → **Data** → **DATA01_100Books.sav**
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.
2. Select **Graphs** → **Chart Builder...**
3. Click on **Gallery** (if not already highlighted in yellow) and click on **Line**.
4. Drag the 'Simple Line' icon from the **Gallery** into the **Chart Preview** box.

- ▶ **Element Properties** window will open → It shows 'Count' as the default **Statistic**.



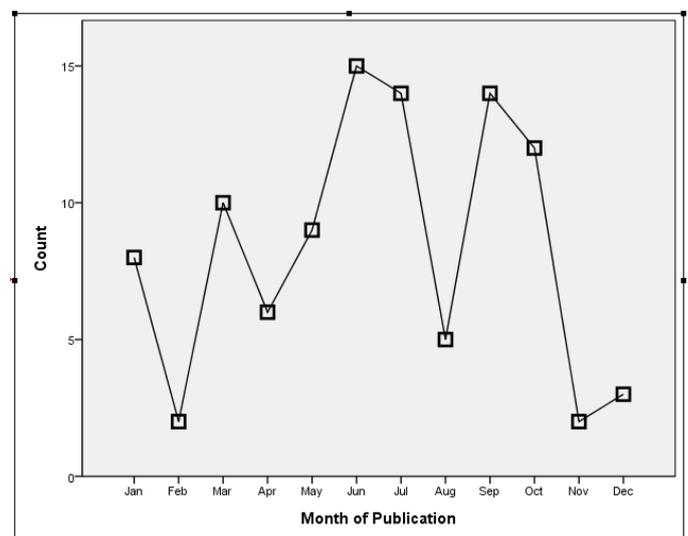
5. Close the **Element Properties** window.
6. Locate **Month** in the **Variables** list (near the bottom) and drag it across to the **X-Axis?** box.
7. Click **OK** to generate a Line Chart.
8. Double-click the Line Chart to enter **Chart Editor**.

9. Select **Elements** → **Add Markers**.
10. In the **Properties** window which opens select **Marker** view →
11. Select **Marker Type** as 'square', **Size** as '12', **Border Width** as '2'.
12. Click **Apply**.
13. Click **Close**.



14. Close the **Chart Editor** window to embed this Line Chart in the **Viewer** Output window.

- ▶ The Simple Line Chart here is essentially the same as a Bar Chart but it does somehow better emphasise the 'high and lows' across the 12 months.
- ▶ The chart suggests there are some preferable times of the year for launching a best-seller.



T19.2 Multiple Line Chart

We now refine the analysis to compare month of launch of Hardback and Paperback books. We can do this by modifying the current **Chart Builder** contents rather than starting afresh.

15. Select **Graphs** → **Chart Builder...**

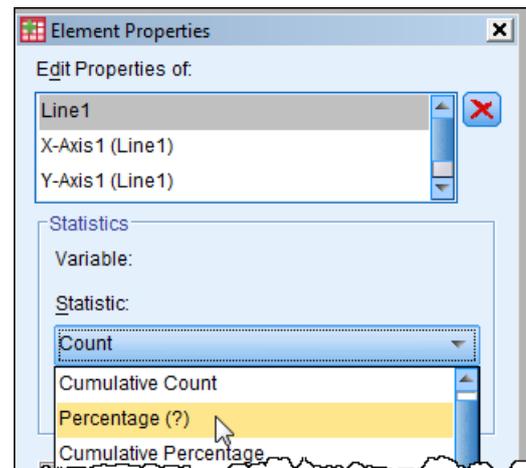
16. Drag the 'Multiple Line' icon from the **Gallery** into the **Chart Preview** box to replace the 'Simple Line'.

▶ The **Element Properties** window should open. Do not close it.

17. In **Chart Builder** locate **Binding** in the **Variables** list (near the bottom) and drag it to the **Set color** box within the **Chart Preview** box.

▶ The **Element Properties** window should still be open →

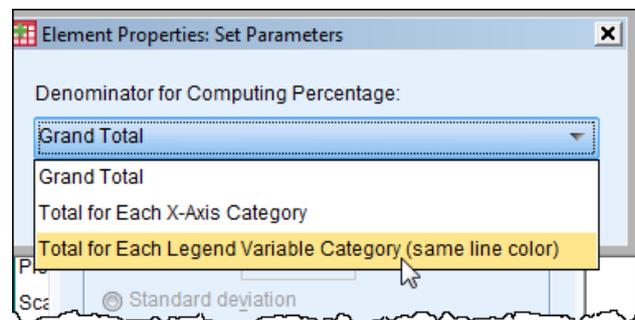
▶ If not, select it from the right side of the **Chart Builder**.



18. With Line 1 selected, change the **Statistic** from the default 'Count' to 'Percentage (?)' →

19. Click on the **Set Parameters...** button and choose ...

Total for Each Legend Variable Category (same line color) →



20. Click **Continue**.

21. Click **Apply**.

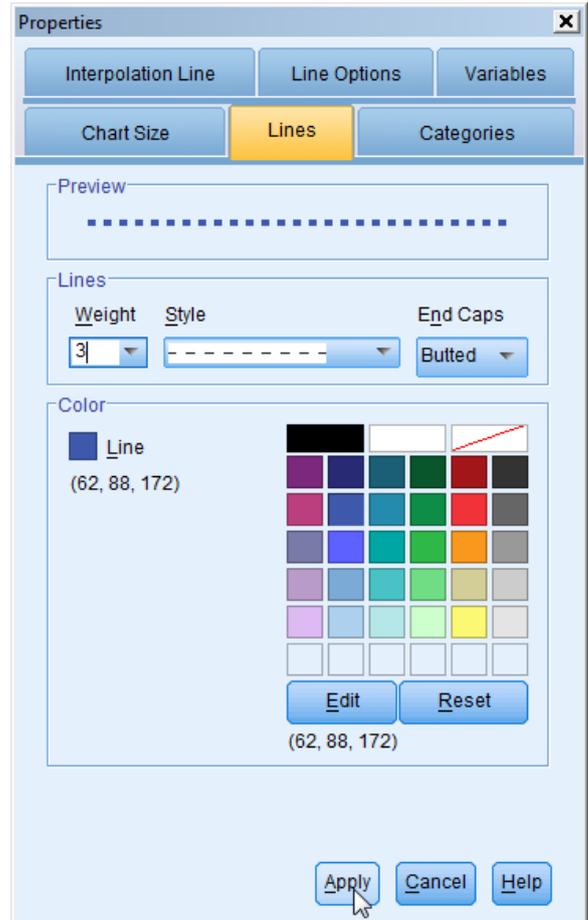
▶ This will make the number of Hardback books launched in a month into a percentage of all Hardback books for the 12 months (and likewise for Paperbacks). This 'evens out' the difference in the total numbers of Hardback and Paperback books, to allow a proper comparison.

22. Close the **Element Properties** window.

23. Click **OK** to exit **Chart Builder** and embed the Line Chart in the **Viewer** Output window.

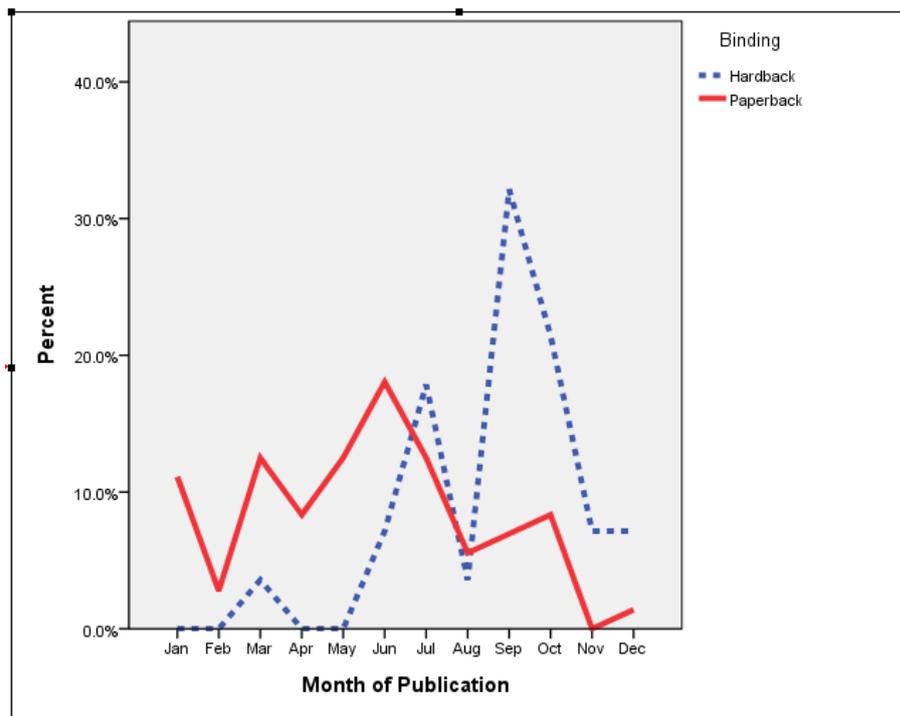
24. Double-click the Line Chart to enter **Chart Editor** again.

25. In the Legend select just the blue line next to the word 'Hardback' – make sure only that line is selected (it will have a pale yellow rectangle round it) .
26. Select **Edit Properties**.
27. In **Lines** view change the **Weight** to '3'.
28. Change the **Lines Style** to dashed – the third dashed type down.
29. Click **Apply**.
30. In the Legend select just the green line next to the word 'Paperback' – make sure only that line is selected.
31. Change the **Lines Weight** to '3'.
32. Change the **Lines Color** to bright red.
33. Click **Apply**.
34. Close the **Properties** window.



35. Close the **Chart Editor** window to embed this Line Chart in the **Viewer** Output window.

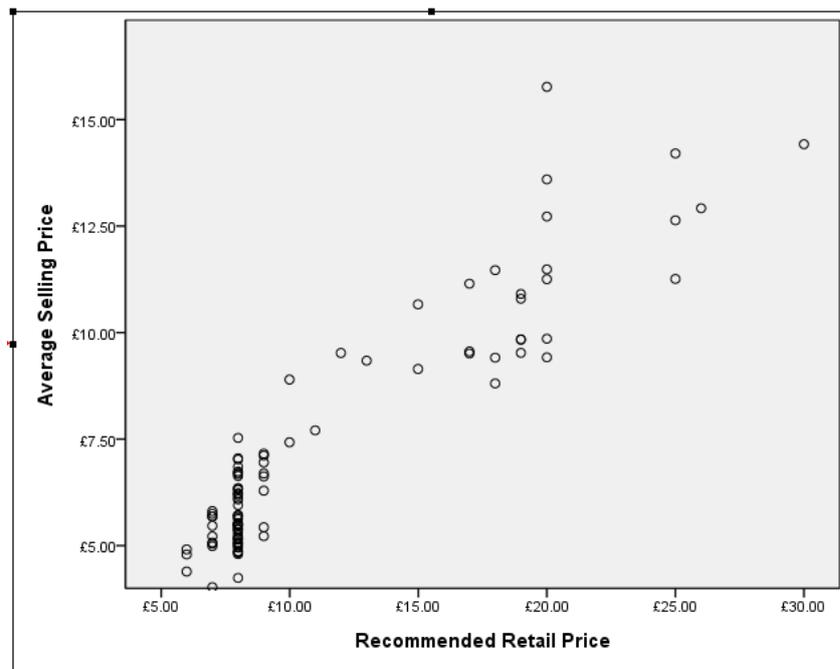
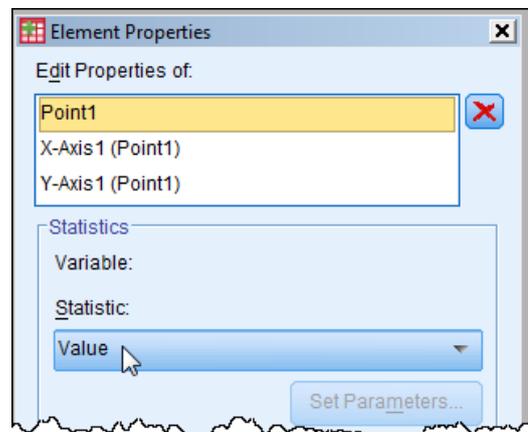
- ▶ This Multiple Line Chart (below) shows that there is a clear difference in the launch month profiles of top-100 books for the two bindings.
- ▶ Questions to ask are: Is this a coincidence? Is it true more generally across all books? Has there been a change over time? Is there a logical explanation?



TUTORIAL T20: Scatterplot

T20.1 Scatterplot – basics

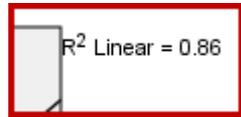
- Load data file: **File** → **Open** → **Data** → **DATA01_100Books.sav**
 - Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.
- Select **Graphs** → **Chart Builder...** and click **Reset**.
- In **Gallery** click on **Scatter/Dot**.
- Drag the 'Simple Scatter' icon from the **Gallery** into the **Chart Preview** box.
 - Element Properties** window will open → It shows 'Value' as the default **Statistic**.
- Close the **Element Properties** window.
- Locate **RRP** [Recommended Retail Price] in the **Variables** list and drag it across to the **X-Axis?** box.
- Locate **ASP** [Average Selling Price] in the **Variables** list and drag it across to the **Y-Axis?** box.
- Click **OK** to generate a Simple Scatterplot:



- Double-click the Line Chart to enter **Chart Editor**.

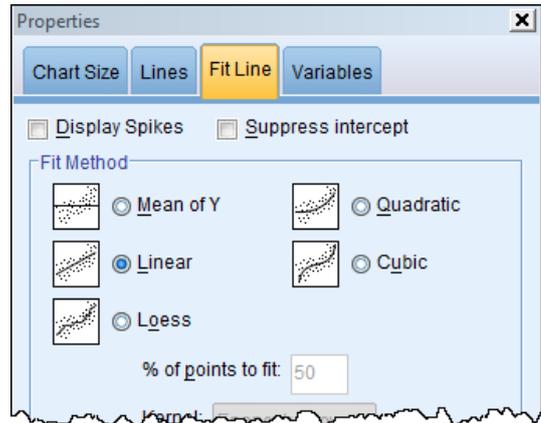
10. Select **Elements** → **Fit Line at Total** to obtain the regression line ('line of best fit').

- ▶ The value of the square of the Pearson correlation coefficient R^2 is shown to the top right of the scatterplot.



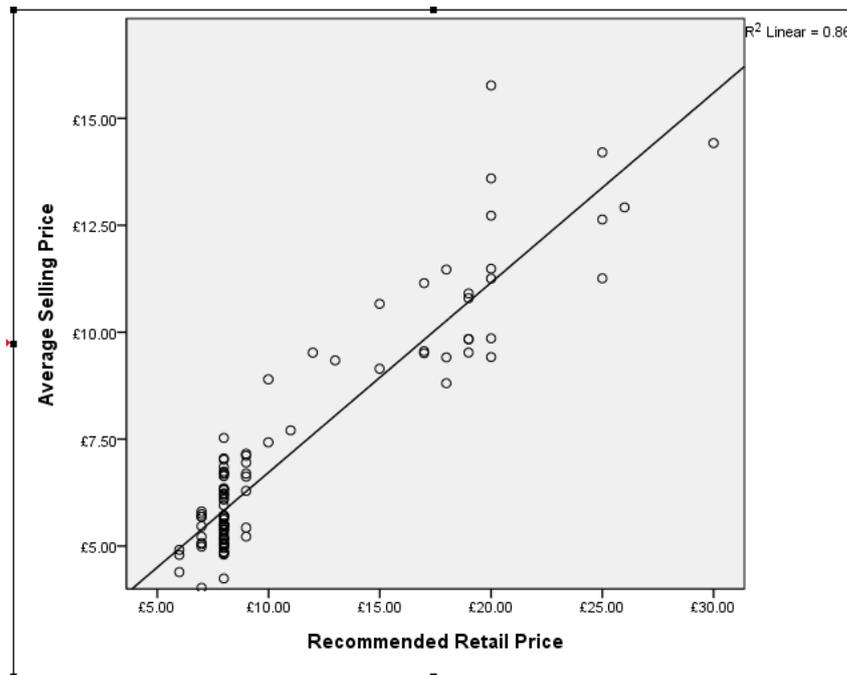
R^2 is a measure of how much one variable 'determines' or 'predicts' the value of the other variable – in this case it is very high at 86%.

- ▶ The **Properties** window opens →



- ▶ The default **Fit Method** is 'Linear' so a straight line is fitted.

11. Close the **Chart Editor** window to embed this Line Chart in the **Viewer** Output window.



We now refine this analysis to separate out the Hardback and Paperback books.

12. **Graphs** → **Chart Builder...**

13. Drag the 'Grouped Scatter' icon from the **Gallery** into the **Chart Preview** box.

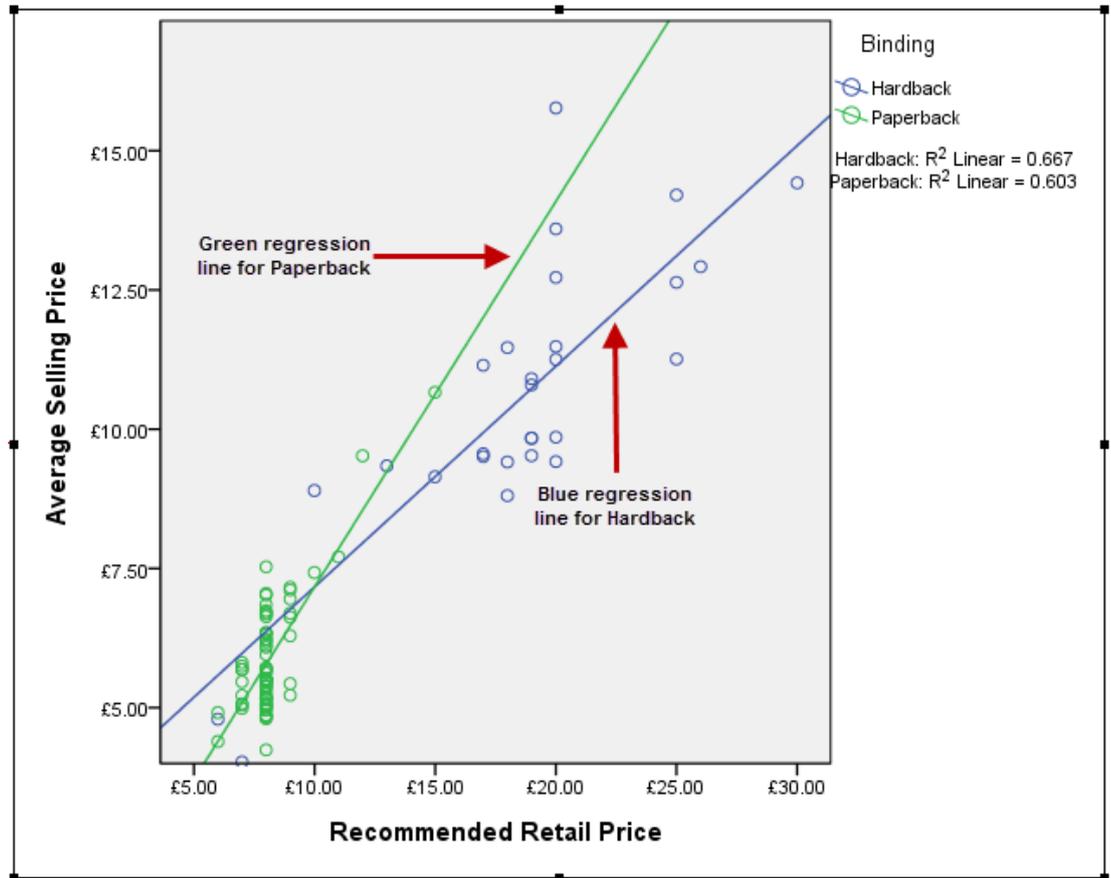
14. Close the **Element Properties** window which will have opened.

15. Locate **Binding** in the **Variables** list and drag it across to the **Set color** box.

16. Click **OK** to generate a Grouped Scatterplot.

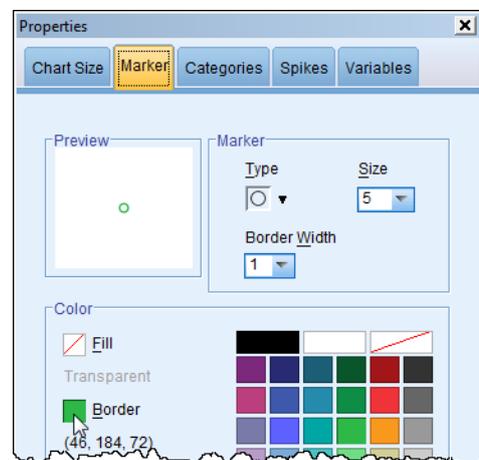
- ▶ The two bindings are now differentiated by colour.

17. Double-click the Scatterplot to enter **Chart Editor**.
18. To get separate regression lines for the two bindings select **Elements → Fit Line at Subgroups**.
 - ▶ SPSS produces the two lines as required, and R^2 values for each.



T20.2 Scatterplot – further explorations for the intrepid

19. To change the colour of the Paperback symbol click on the green circle in the **Legend** next to 'Paperback' which will highlight all the green circles in the scatterplot.
 - ▶ This **Properties** window opens → (if not, open it using **Edit Properties**) .
20. Select **Marker** view (if not highlighted) and click in the **Border** box and then click on 'red'.
21. Click **Apply**.
22. To change the Hardback symbol click on the blue circle next to 'Hardback', which will highlight all the blue circles.



23. In the **Properties** window select **Marker** view (if not selected) and change the **Type** from 'circle' to 'square' and click in the **Fill** box and then click on 'blue'.

24. Click **Apply**.

- ▶ Note that SPSS has also filled in the Paperback red circles which we did not ask for (a minor bug!). To overcome this, proceed as follows:

25. Click on the red circle next to 'Paperback' which will highlight all the red circles.

26. In the **Properties** window select **Marker** view (if not selected) and click in the **Fill** box and then click on 'white'.

27. Click **Apply**.

- ▶ This should 'unfill' the Paperback circles.

To edit the separate regression lines for the two bindings proceed as follows:

28. In the Legend click on the blue line next to 'Hardback'.

In the **Properties** window select the Lines tab and for Style choose a long dash.

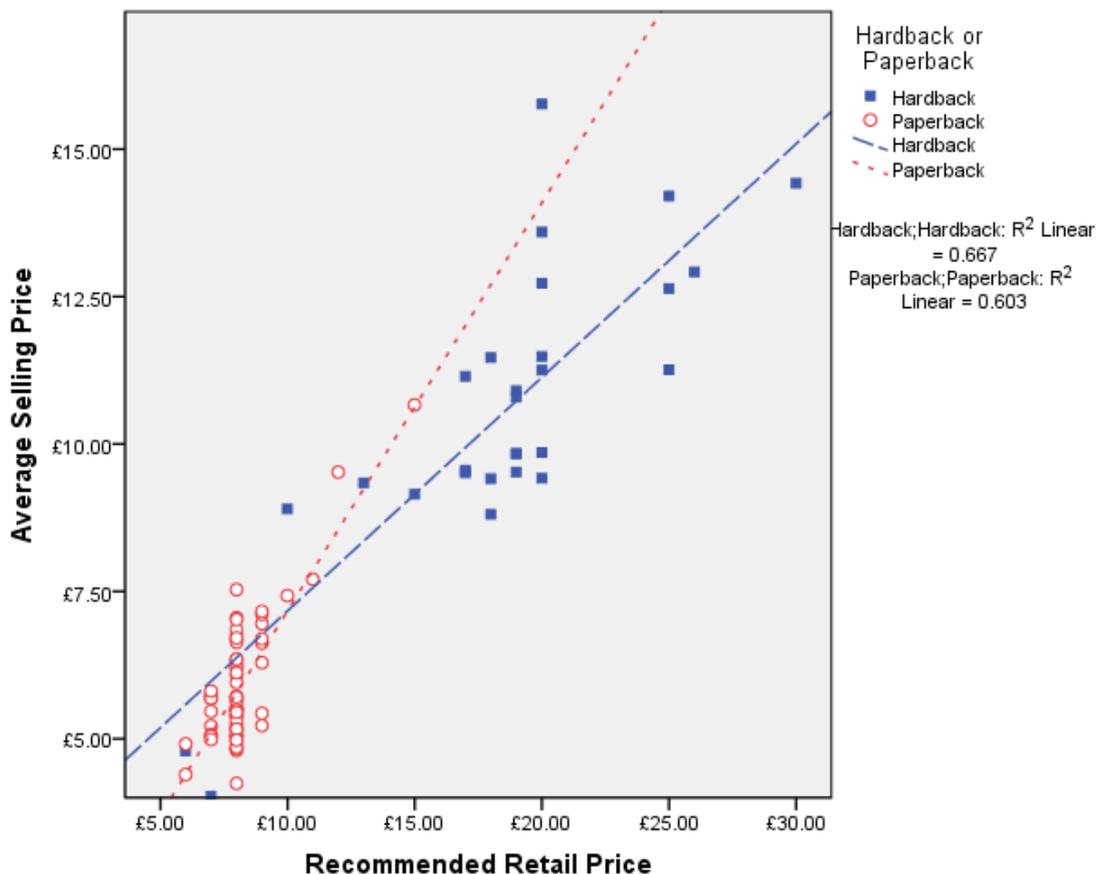
Click **Apply**.

29. In the Legend click on the red line next to 'Paperback'.

In the **Properties** window select the Lines tab and for Style choose a short dash, and for the Line Color click on 'red'.

Click **Apply**.

- ▶ The result should appear as follows:

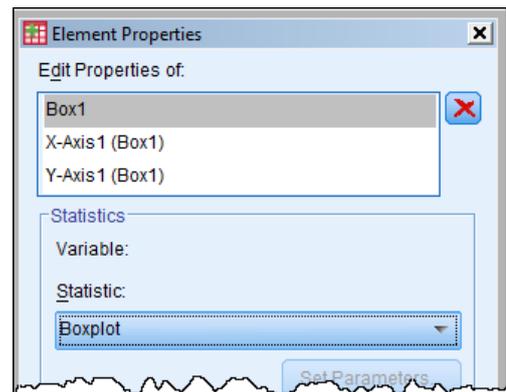


TUTORIAL T21: Boxplot

T21.1 Simple Boxplot – one variable

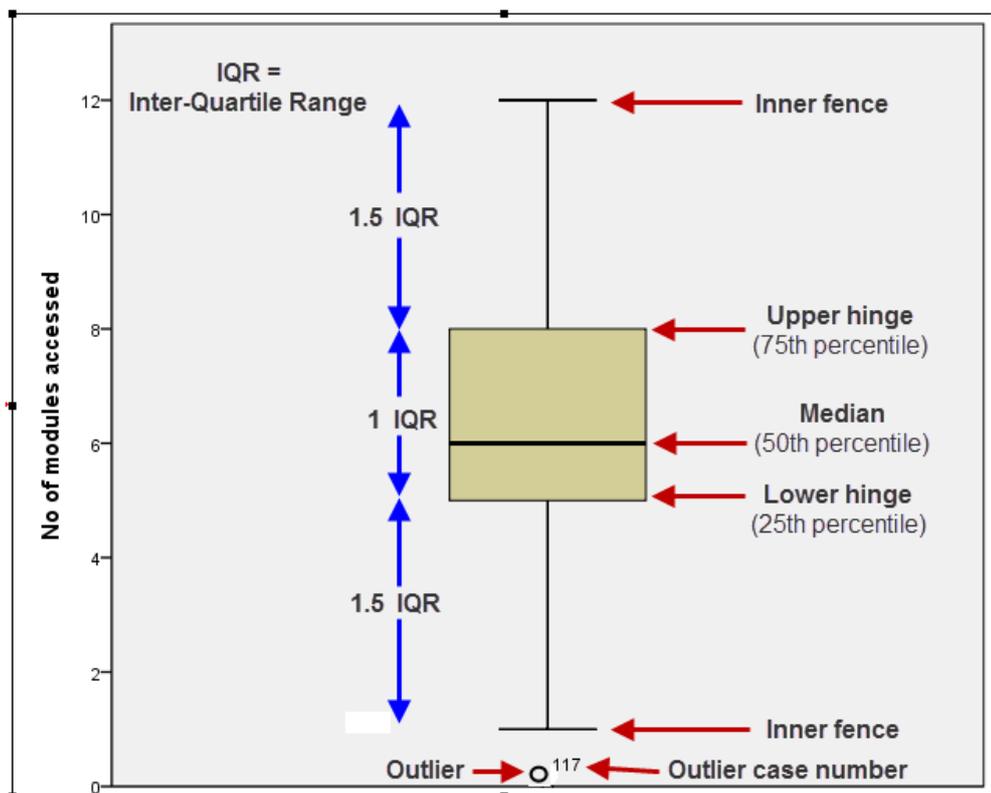
1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...** and click **Reset**.
3. In **Gallery** click on **Boxplot**.
4. Drag the 'Simple Boxplot' icon from the **Gallery** into the **Chart Preview** box.

- ▶ **Element Properties** window will open → It shows 'Boxplot' as the default **Statistic**.



5. Close the **Element Properties** window.
6. Locate **modules** in the **Variables** list and drag it across to the **Y-Axis?** box.
7. Click **OK** to generate a Simple Boxplot.

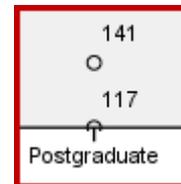
- ▶ This Simple Boxplot has been annotated to display its significant features (below).
- ▶ The 1-D Boxplot would produce the same output as this does here.
- ▶ For a normal distribution the distance from Median to Inner fence is approximately equivalent to 3 standard deviations, enclosing virtually all the values.



T21.2 Simple Boxplot – two variables

1. Load data file: **File** → **Open** → **Data** → **DATA03_LSquestionnaire.sav** (if not loaded).
2. Select **Graphs** → **Chart Builder...** and click **Reset**.
3. In **Gallery** click on **Boxplot**.
4. Drag the 'Simple Boxplot' icon from the **Gallery** into the **Chart Preview** box.
5. Close the **Element Properties** window.
6. Locate **modules** in the **Variables** list and drag it across to the **Y-Axis?** box.
7. Locate **ug_pg** in the **Variables** list and drag it across to the **X-Axis?** box.
8. Click **OK** to generate a Simple Boxplot with two boxplots.

- There are several outliers and the circle outlier 117 is right on the zero line. We will change the origin to reveal it properly.



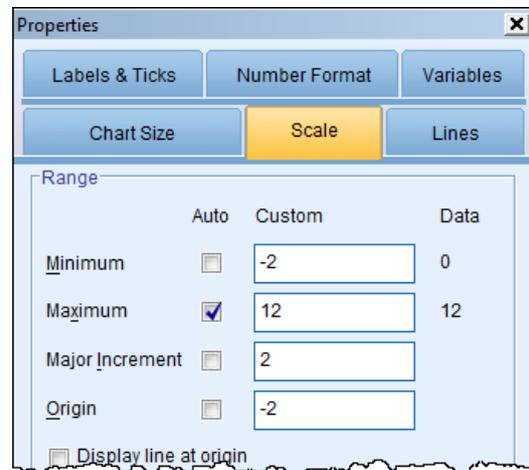
9. Double-click on the chart to open **Chart Editor**.

10. Select the Y-axis either by **Edit** → **Select Y-Axis** or by clicking here:



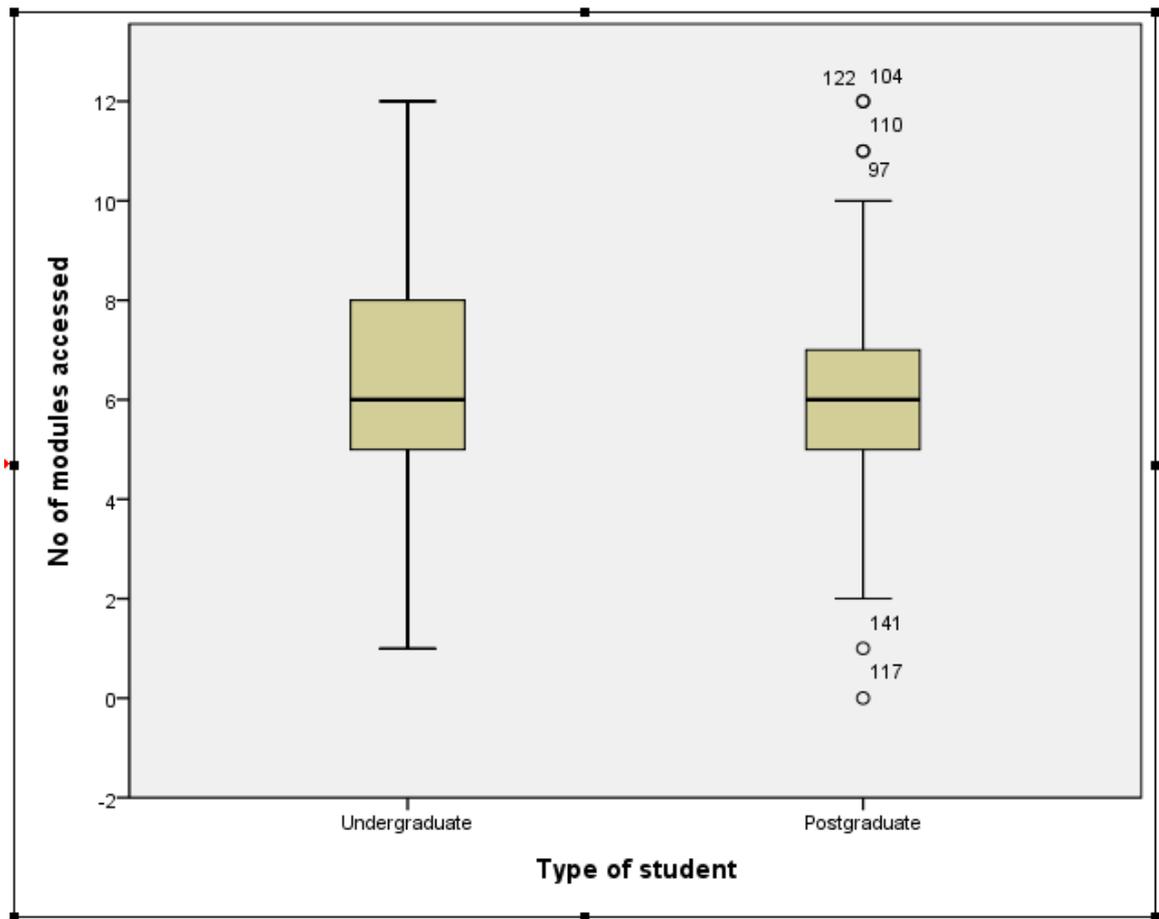
11. In the **Properties** window which opens make sure **Scale** view is selected and set:

Minimum to '-2',
Major increment to '2',
Origin to '-2'.



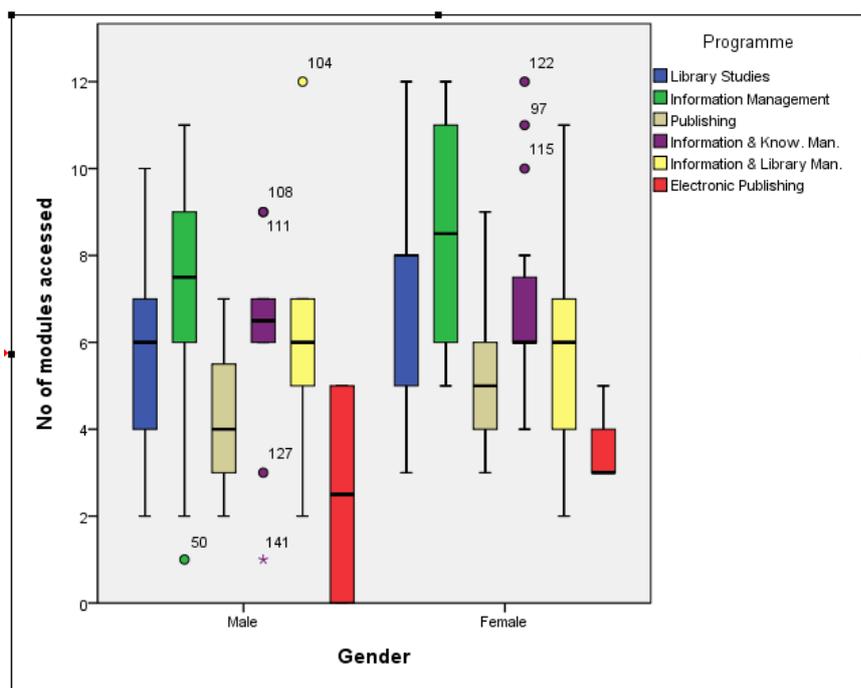
- It may seem quite unnecessary to edit the **Major increment** (which process removes the tick from its **Auto** box) but if you don't do anything then **Auto** changes the '2' to '2.5' when you click **Apply**, which spoils the neatness of the Y-axis scale. (However, instead of editing the **Major increment** you could in this case just deselect **Auto** instead.)

12. Click **Apply**.
13. Click **Close**.
14. Close **Chart Editor** to embed the chart in the **Viewer** Output window (see next page).



T21.3 Multiple Boxplot – two variables

More complex boxplots can be easily produced (as below), but this will not be pursued here.

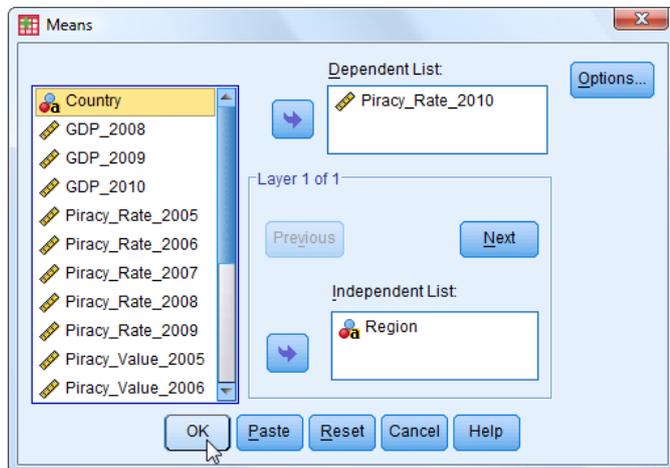


TUTORIAL T22: Means

1. Load data file: **File** → **Open** → **Data** → **DATA07_IT_Piracy.sav**
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.

2. Select **Analyze** → **Compare Means** → **Means**

- ▶ This opens the Means window shown here →



3. Move the scale variable **Piracy_Rate_2010** into the **Dependent List** and the string variable **Region** into the **Independent List**.

4. Click **OK**.

- ▶ This produces the Report below which gives the **IT Piracy** statistics broken down by **Region**, showing means and standard deviations:

Report
Piracy Rate 2010 - % of all software in use - Source: BSA

Region of World	Mean	N	Std. Deviation
Africa	74.07	15	14.577
Asia & Pacific	58.78	18	24.489
Central America & Caribbean	72.00	8	9.150
Central & Eastern Europe	66.58	24	17.275
Middle East	59.17	12	17.362
North America	24.00	2	5.657
South America	69.50	10	11.453
Western Europe	34.35	20	11.047
Total	59.48	109	21.362

- ▶ Many other statistics can be obtained by opening the **Options** window.
- ▶ If more variables are added to the **Dependent List** then the table is set out differently with one column per variable, as below:

Report

Region of World		Piracy Rate 2010 - % of all software in use - Source: BSA	Piracy Rate 2005 - % of all software in use - Source: BSA
Africa	Mean	74.07	75.57
	N	15	14
	Std. Deviation	14.577	14.179
Asia & Pacific	Mean	58.78	59.80
	N	18	15
	Std. Deviation	24.489	23.604

- ▶ If more variables are added to the **Independent List** then more tables are obtained.

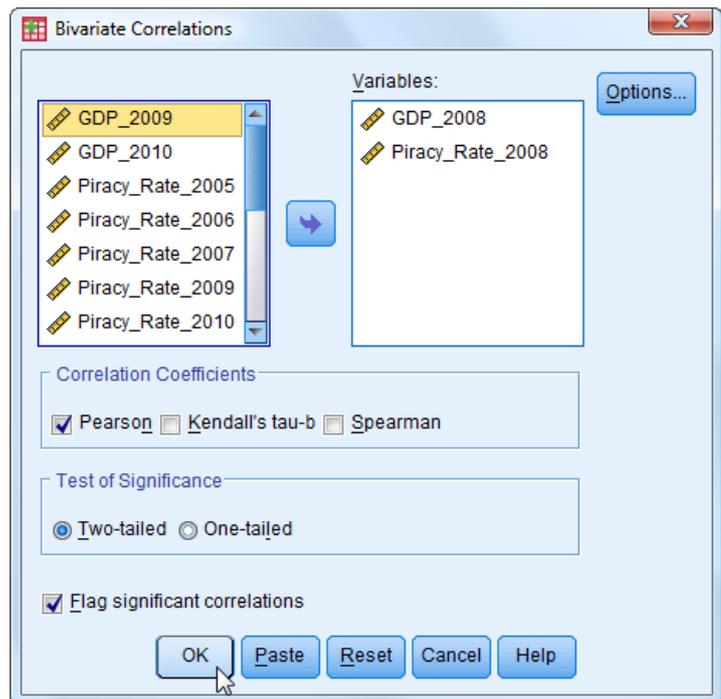
TUTORIAL T23: Correlation

T23.1 Pearson Correlation (parametric)

1. Load data file: **File** → **Open** → **Data** → **DATA07_IT_Piracy.sav**
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.
2. Select **Analyze** → **Correlate** → **Bivariate**

▶ Widen the whole window to reveal the full variable names, as necessary.

3. Move the scale variables **GDP_2008** and **Piracy_Rate_2008** into the **Variables** box.
 - ▶ Note that **Pearson** correlation is selected as default and so are **Two-tailed** test and **Flag significant correlations**.



4. Click **OK** to produce the **Pearson** correlation table below.

- ▶ The correlation obtained is reported as significant at the 1% level ($p < 0.01$). However, the relationship is not very strong.
- ▶ The correlation is negative indicating that higher **GDP** is associated with lower **IT Piracy Rate**.

Correlations

		Gross Domestic Product Total (US\$millions) 2008 - Source: World Bank	Piracy Rate 2008 - % of all software in use - Source: BSA
Gross Domestic Product Total (US\$millions) 2008 - Source: World Bank	Pearson Correlation	1	-.329**
	Sig. (2-tailed)		.000
	N	109	108
Piracy Rate 2008 - % of all software in use - Source: BSA	Pearson Correlation	-.329**	1
	Sig. (2-tailed)	.000	
	N	108	108

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

5. Select **Analyze** → **Correlate** → **Bivariate**.
6. Move the scale variables **Piracy_Value_2008** into the **Variables** box, to join the other two variables already there.
7. Click **OK** to produce the **Pearson** correlations shown below.
 - ▶ The correlations obtained are all reported as significant at the 1% level ($p < 0.01$) shown by ** next to the value, or significant at the 5% level ($p < 0.05$) shown by * next to the value.
 - ▶ The correlation for **GDP** against **Piracy Value** is very high and positive.
 - ▶ It is perhaps surprising at first that the correlation for **Piracy Rate** against **Piracy Value** is negative. We will investigate this further ...

Correlations

		Gross Domestic Product Total (US\$millions) 2008 - Source: World Bank	Piracy Rate 2008 - % of all software in use - Source: BSA	Piracy Value 2008 (US\$ millions) - Source: Business Software Alliance
Gross Domestic Product Total (US\$millions) 2008 - Source: World Bank	Pearson Correlation	1	-.329**	.900**
	Sig. (2-tailed)		.000	.000
	N	109	108	108
Piracy Rate 2008 - % of all software in use - Source: BSA	Pearson Correlation	-.329**	1	-.206*
	Sig. (2-tailed)	.000		.033
	N	108	108	108
Piracy Value 2008 (US\$ millions) - Source: Business Software Alliance	Pearson Correlation	.900**	-.206*	1
	Sig. (2-tailed)	.000	.033	
	N	108	108	108

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

* . Correlation is significant at the 0.05 level (2-tailed).

8. Select **Analyze** → **Correlate** → **Partial**.
9. Move the variables **Piracy_Rate_2008** and **Piracy_Value_2008** into the **Variables** box.
10. Move **GDP_2008** into the **Controlling for** box.
11. Click **OK** to produce the **Pearson** partial correlation:

Correlations

Control Variables			Piracy Rate 2008 - % of all software in use - Source: BSA	Piracy Value 2008 (US\$ millions) - Source: Business Software Alliance
Gross Domestic Product Total (US\$millions) 2008 - Source: World Bank	Piracy Rate 2008 - % of all software in use - Source: BSA	Correlation	1.000	.220
		Significance (2-tailed)	.	.023
		df	0	105
	Piracy Value 2008 (US\$ millions) - Source: Business Software Alliance	Correlation	.220	1.000
		Significance (2-tailed)	.023	.
		df	105	0

- ▶ Now that the effect of GDP is removed we see that the (partial) correlation of **Piracy Rate** against **Piracy Value** is positive.

T23.2 Spearman Correlation (nonparametric)

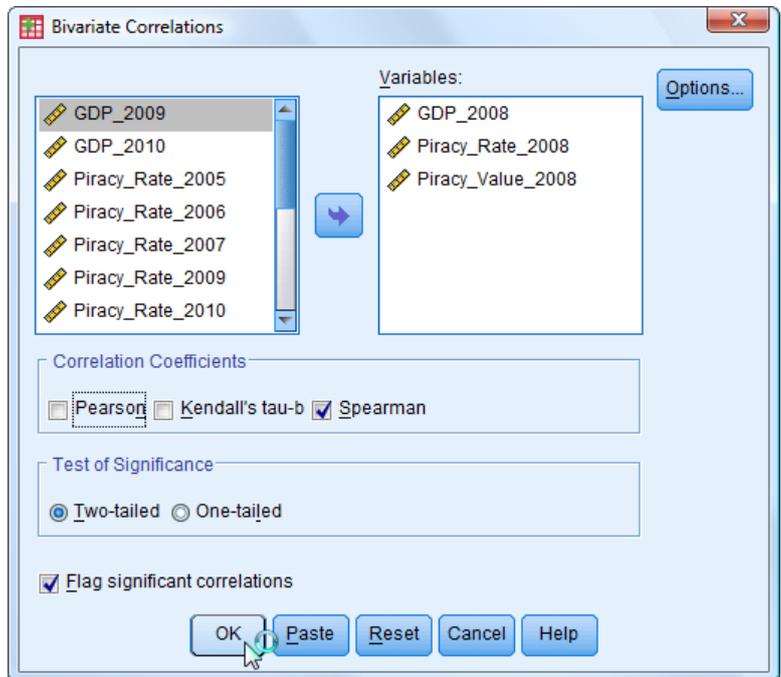
1. **Analyze** → **Correlate** → **Bivariate**

2. If not already there, move variables **GDP_2008**, **Piracy_Rate_2008** and **Piracy_Value_2008** into the **Variables** box.

3. Click on the **Pearson** box to deselect it (it's the default), shown by removing the tick.

4. Click on the **Spearman** box to select it inserting a tick.

5. Click **OK** to produce the **Spearman** correlation table below, and compare it with the **Pearson** results on the previous page.



Correlations

			Gross Domestic Product Total (US\$millions) 2008 - Source: World Bank	Piracy Rate 2008 - % of all software in use - Source: BSA	Piracy Value 2008 (US\$ millions) - Source: Business Software Alliance
Spearman's rho	Gross Domestic Product Total (US\$millions) 2008 - Source: World Bank	Correlation Coefficient	1.000	-.533**	.942**
		Sig. (2-tailed)	.	.000	.000
		N	109	108	108
	Piracy Rate 2008 - % of all software in use - Source: BSA	Correlation Coefficient	-.533**	1.000	-.405**
		Sig. (2-tailed)	.000	.	.000
		N	108	108	108
	Piracy Value 2008 (US\$ millions) - Source: Business Software Alliance	Correlation Coefficient	.942**	-.405**	1.000
		Sig. (2-tailed)	.000	.000	.
		N	108	108	108

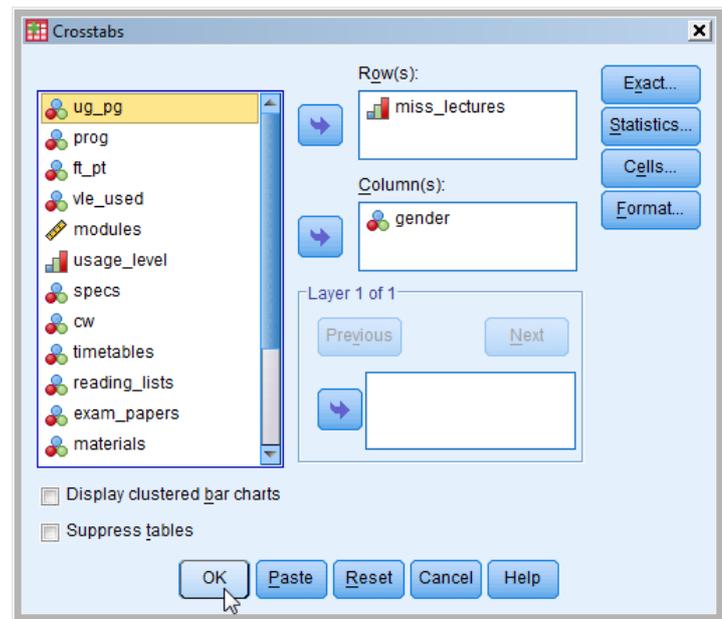
** . Correlation is significant at the 0.01 level (2-tailed).

- ▶ The correlations obtained are similar to but somewhat higher than, those for Pearson.
- ▶ Pearson has a number of conditions:
 - (a) The relationship is linear
 - (b) Points are evenly distributed along the line ('homoscedasticity')
 - (c) The data come from a normal distribution
 - (d) The data are scale (i.e. interval or ratio) from continuous distributions.
- ▶ Spearman does not have these conditions, and is a nonparametric statistic, and so is to be preferred if the Pearson conditions are not met.
- ▶ In practice, the two correlations usually lead to the same broad conclusion.

TUTORIAL T24: Crosstabs and the Chi-square Test

T24.1 Crosstabs – introduction

1. Load data file: **File** → **Open** → **Data** → DATA03_LS_questionnaire.sav
2. Select **Analyze** → **Descriptive Statistics** → **Crosstabs**
3. In the Crosstabs window which opens, move the ordinal variable **miss_lect** into the **Row(s)** box and the nominal variable **gender** into the **Column(s)** box.
4. Click **OK**.



- ▶ The crosstabulation table below appears, which compares the attitude to missing lectures of males and females, and seems to show a significant difference between the genders.

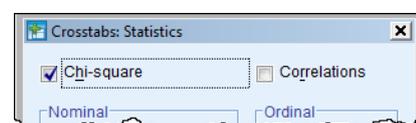
Don't mind missing lectures if they are on the VLE * Gender Crosstabulation

		Gender		Total
		Male	Female	
Don't mind missing lectures if they are on the VLE	Strongly disagree	4	16	20
	Disagree	18	23	41
	Neutral	19	18	37
	Agree	20	6	26
	Strongly agree	15	11	26
Total		76	74	150

T24.2 Crosstabs and the Chi-square Test option

Continuing from T24.1 with the same analysis ...

1. Select **Analyze** → **Descriptive Statistics** → **Crosstabs**
 - ▶ This time we will use the various options buttons (top right of the **Crosstabs** window) to investigate the differences further.
2. In the **Crosstabs** window click on **Statistics** and click in the **Chi-square** tick-box.
3. Click **Continue**.



4. Click on **Cells** and click the **Expected** tick-box in the **Counts** section and the **Adjusted standardized** tick-box in the **Residuals** section.
5. Click **Continue**.
6. Click **OK** to obtain this augmented crosstabulation table in the **Viewer** Output window below:

Don't mind missing lectures if they are on the VLE ^ Gender Crosstabulation

			Gender		Total
			Male	Female	
Don't mind missing lectures if they are on the VLE	Strongly disagree	Count	4	16	20
		Expected Count	10.1	9.9	20.0
		Adjusted Residual	-2.9	2.9	
	Disagree	Count	18	23	41
		Expected Count	20.8	20.2	41.0
		Adjusted Residual	-1.0	1.0	
	Neutral	Count	19	18	37
		Expected Count	18.7	18.3	37.0
		Adjusted Residual	.1	-.1	
	Agree	Count	20	6	26
		Expected Count	13.2	12.8	26.0
		Adjusted Residual	2.9	-2.9	
	Strongly agree	Count	15	11	26
		Expected Count	13.2	12.8	26.0
		Adjusted Residual	.8	-.8	
	Total	Count	76	74	150
		Expected Count	76.0	74.0	150.0

- ▶ The table above contains the actual Count information, as in the previous table, but also includes the Expected Count – i.e. the frequencies which one would expect if there were no difference between the genders (no ‘bias’ one might say). Clearly some differences will happen by chance and the ‘significance’ of the size of each individual difference will depend upon the size of the frequency.
- ▶ The Adjusted Residual signifies the level of importance; any value of the Adjusted Residual outside +2 to – 2 is considered ‘significant’. (N.B. Here ‘significant’ is the ordinary English usage of the word which is not the same as ‘statistically significant’ as used below.)

[See T24.3: Notes on the Chi-square Test – Adjusted Residuals option]

- ▶ The **Chi-Square Tests** table below reports indicates if overall the differences are statistically significant by giving the probability of differences this large or larger occurring by chance.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	15.967 ^a	4	.003
Likelihood Ratio	16.892	4	.002
Linear-by-Linear Association	10.314	1	.001
N of Valid Cases	150		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.87.

- ▶ The **Chi-Square Tests** table shows that in this case the probability of differences this large or larger occurring by chance is 0.003. This is much smaller than the normal 0.05 criterion level used (95% significance) so there is very strong evidence of a difference between the genders.

As $p < 0.05$, it is significant at the 95% level.

As $p < 0.01$, it is significant at the 99% level.

As $p > 0.001$ it is not significant at the 99.9% level.

[See T24.4: Notes on the Chi-square Test – Significance level]

- ▶ The ‘small print’ beneath the **Chi-Square Tests** table is important. It shows that the conditions of the Chi-square test have been met so the test is **valid**.

[See T24.5: Notes on the Chi-square Test – Criteria for validity.]

T24.3 Notes on the Chi-square Test – Adjusted Residuals option

A Chi-square Test report may confirm the existence of a statistically significant association but it will not indicate how *strong* this is or which cells are *most deviant*.

The Crosstabs table provides some evidence because the difference between a cell’s Count (*aka Observed* count) and its **Expected** count does show the deviation. But it can be hard to analyse or interpret this. By adding an extra optional line – showing the **Adjusted Residuals** – we can get some help.

Any Adjusted Residual which exceeds +2 or is less than –2 is an indicator of a marked deviation – the sign indicates the direction. The bigger this is the more important a contribution it makes.

T24.4 Notes on the Chi-square Test – Significance level

The Chi-square test allows us to determine whether or not there is a **statistically significant association** between two variables such as in a Crosstabs table.

The test provides a significance value for the association as a probability (p) – it is called **Asymp. Sig.** in SPSS Crosstabs output.

To be a significant association (and not a product of random chance) small values of p are needed. Normal choices are:

95% level: $p < 0.05$ 99% level: $p < 0.01$ 99.9% level: $p < 0.001$

The choice of level depends on how confident you want to be before declaring the existence of an association; 95% is commonly used. (It could be quite a **weak** association.)

If p is larger than the chosen significance level then the variables are said to be **statistically independent**.

The SPSS Crosstabs procedure generates a **Chi-Square Test** report, showing the significance level:

For a table larger than 2 by 2 (i.e. more than 4 cells) look for the line:

Pearson Chi-square to find the **Asymp. Sig.** value.

For a 2 by 2 table (i.e. one with 4 cells) look for the line below that:

Continuity Correction to find the **Asymp. Sig.** value.

T24.5 Notes on the Chi-square Test – Criteria for validity

Chi-square is **not valid** if the Expected cell counts are too small – caused by having too few cases or too many cells. The criteria which must be met to use Chi-square are:

For a table larger than 2 by 2 (i.e. more than 4 cells):

- No cell may have an expected count less than 1,
- No more than 20% of cells may have a count less than 5.

For a 2 by 2 table (i.e. one with 4 cells):

- No cell may have an expected count less than 5.

SPSS automatically prints out the relevant information below the **Chi-Square Tests** report table. Remember to check this before drawing any conclusions!

T24.6 Crosstabs and the Chi-square Test – Recoding of Data

1. Load data file: **File** → **Open** → **Data** → DATA03_LS_questionnaire.sav

2. Select **Analyze** → **Descriptive Statistics** → **Crosstabs**

▶ The Crosstabs window opens.

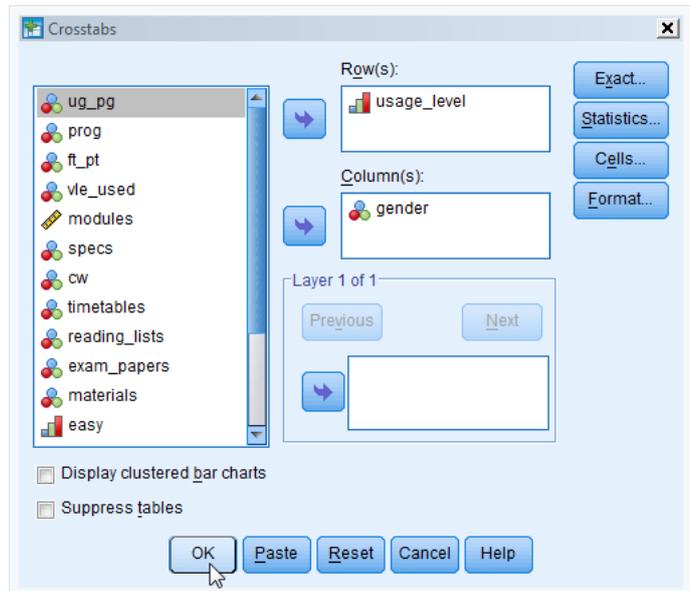
3. Click **Reset**.

4. Move the ordinal variable **usage_level** into the **Row(s)** box.

5. Move the nominal variable **gender** into the **Columns** box.

6. Click **OK**.

▶ The main output is the crosstabulation table below, which compares the VLE usage of males and females, and seems to show a significant difference between the genders.



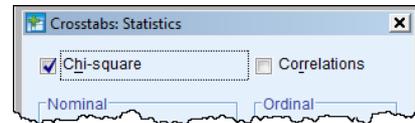
Frequency of VLE use * Gender Crosstabulation

Count		Gender		Total
		Male	Female	
Frequency of VLE use	Several times a week	19	35	54
	Once or twice a week	41	23	64
	Once or twice a month	15	16	31
	Less than once a month	1	0	1
Total		76	74	150

7. Select **Analyze** → **Descriptive Statistics** → **Crosstabs**

► We use the various options buttons to analyze this further ...

8. In the **Crosstabs** window click on **Statistics** and click in the **Chi-square** tick-box.



7. Click **Continue**.

8. Click on **Cells** and click the **Expected** tick-box in the **Counts** section and the **Adjusted standardized** tick-box in the **Residuals** section.

9. Click **Continue** and then **OK**.

► The output is the crosstabulation table below:

Frequency of VLE use * Gender Crosstabulation

			Gender		Total
			Male	Female	
Frequency of VLE use	Several times a week	Count	19	35	54
		Expected Count	27.4	26.6	54.0
		Adjusted Residual	-2.8	2.8	
	Once or twice a week	Count	41	23	64
		Expected Count	32.4	31.6	64.0
		Adjusted Residual	2.8	-2.8	
	Once or twice a month	Count	15	16	31
		Expected Count	15.7	15.3	31.0
		Adjusted Residual	-.3	.3	
	Less than once a month	Count	1	0	1
		Expected Count	.5	.5	1.0
		Adjusted Residual	1.0	-1.0	
Total		Count	76	74	150
		Expected Count	76.0	74.0	150.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10.811 ^a	3	.013
Likelihood Ratio	11.336	3	.010
Linear-by-Linear Association	3.447	1	.063
N of Valid Cases	150		

a. 2 cells (25.0%) have expected count less than 5. The minimum expected count is .49.

► The Pearson **Chi-Square** row shows the result to be significant at the 95% level since $p = 0.013$ (i.e. $p < 0.05$) BUT there is a problem because the 'small print, beneath the table shows the test to be **invalid**. As it fails to meet the criteria for larger than 2 by 2 tables:

- No cell may have an expected count less than 1 – here it is **0.49**.
- No more than 20% of cells may have a count less than 5 – here it is **25%**.

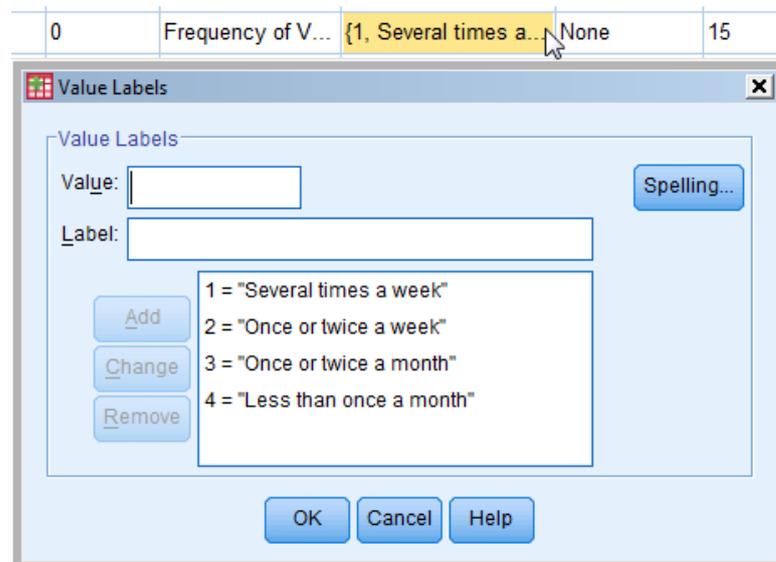
► It is clear where the problem lies – only 1 entry for 'Less than once a month'.

- The solution is to combine categories together – in this case combine it with the previous category to make a new category of 'Less than twice a month'.

	Gender		Total
	Male	Female	
Several times a week	19	35	54
Once or twice a week	41	23	64
Once or twice a month	15	16	31
Less than once a month	1	0	1

- This requires using the **Recode** procedure which was explained in Reference Section 9.2 on page 29.

10. It is important to be sure what the current codes are. Depending on the Output option settings, the labels rather than the values will appear – as in the case here. To find the codes, select **Variable View** and click on the three dots in the Values column for variable for **usage_level**. This is what will appear:



- We must recode 4 → 3 (reducing the categories by one), and can keep the remaining categories all the same. Do this as follows:

11. Close the **Value Labels** box if it has been opened.
12. Select **Transform** → **Recode into Different Variables**.
13. Within the **Recode into Different Variables** window, locate **usage_level** in the variable list on the left and move it into the **Input Variable** → **Output Variable** box on the right.

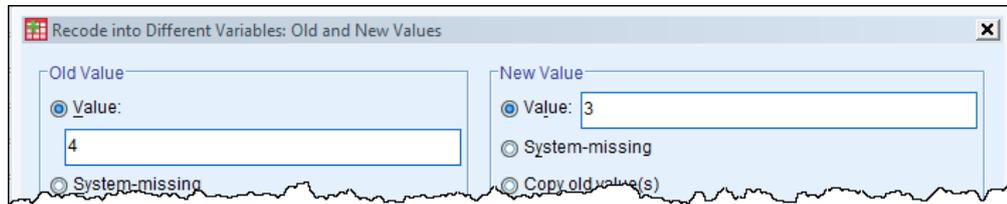
Then type into the **Name** box the name for the new variable (**usage_code**).

Then type into the **Label** box a suitable label for the new variable (**New usage categories**).

14. Click **Change**.

15. Click **Old and New Values**.

► The **Recode into Different Variables** window opens.



16. To recode 4 → 3, enter '4' in the **Value** box in the **Old Value** section and enter '3' in the **Value** box in the **New Value** section.

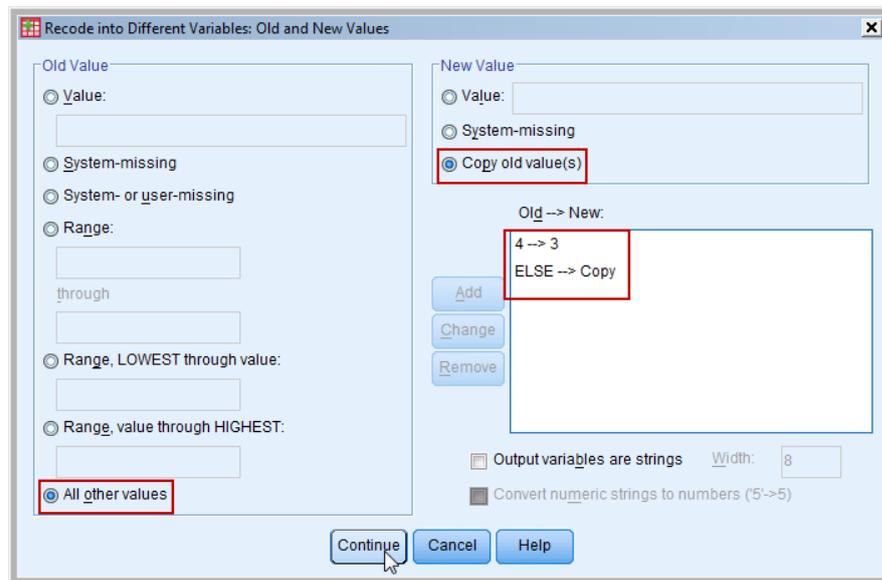
17. Click **Add**.

18. It is **VERY IMPORTANT** to tell *SPSS* to keep all the other values otherwise they will be lost. Do this as follows:

Click on the **All other values** radio button and click on the **Copy old value(s)** radio button.

19. Click **Add**.

► The window should look exactly as shown in the screenshot below.



20. Click **Continue**.

21. Click **OK**.

► The report, confirming the recoding, will appear in the **Viewer** Output window.

```
RECODE usage_level (4=3) (ELSE=Copy) INTO usage_code.
VARIABLE LABELS usage_code 'New usage categories'.
EXECUTE.
```

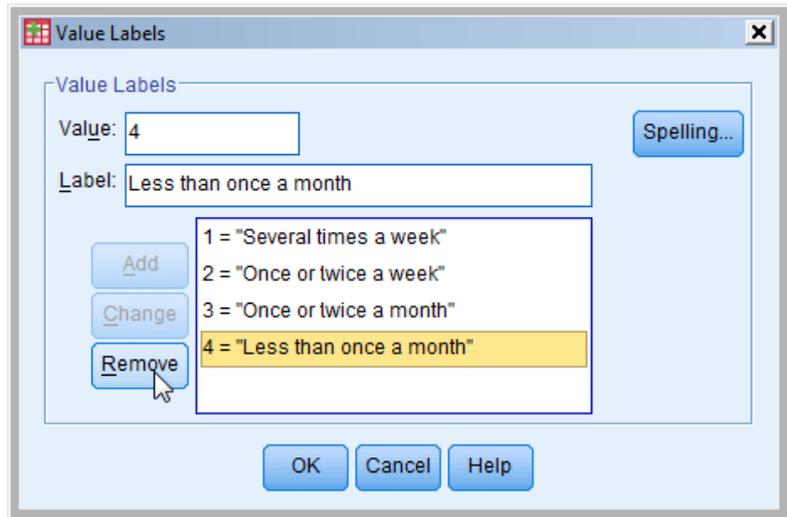
- The attributes of the the new recoded variable **usage_code** need attending to so that they match those of the **usage_level** variable from which it has been derived. Do this as follows:

22. Select **Variable View**.

23. Locate the row for **usage_code** (it will be the last variable) and change the **Decimals** to '0'.

24. Copy the **Values** entry for **usage_level** into that for **usage_code Values**.

25. Edit the **Value Labels** entry by removing '4' and rewording the label for '3' to 'Less than twice a month' and click **OK** when a warning message appears.



26. Click **OK**.

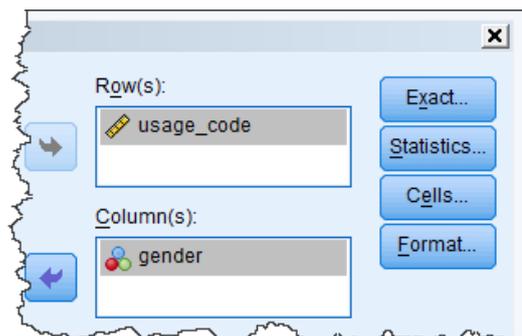
- It is now possible to repeat the **Chi-square** test using the new recoded variable.

27. Select **Analyze** → **Descriptive Statistics** → **Crosstabs**

28. Click **Reset**.

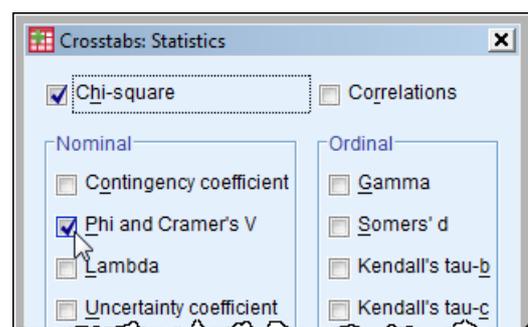
29. Locate the variable **usage_code** (look at the bottom of the list) and move it into the **Row(s)** box →

30. Locate the variable **gender** and move it into the **Column(s)** box →

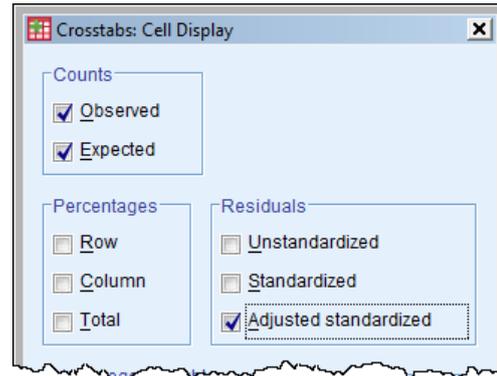


31. Click on **Statistics** and select both **Chi-square** and **Phi and Cramér's V**.

32. Click **Continue**.



33. Click on **Cells** and select **Counts Expected** and **Residuals Adjusted standardized**.



34. Click **Continue** and then **OK**.

- ▶ The crosstabulation output is shown below. There are three tables:
- ▶ In the first table (below), 4 of the 6 **Adjusted Residuals** are larger than '2', indicating that Males and Females differ markedly in reporting their level of regular usage.

New usage categories * Gender Crosstabulation

			Gender		Total
			Male	Female	
New usage categories	Several times a week	Count	19	35	54
		Expected Count	27.4	26.6	54.0
		Adjusted Residual	-2.8	2.8	
	Once or twice a week	Count	41	23	64
		Expected Count	32.4	31.6	64.0
		Adjusted Residual	2.8	-2.8	
	Less than twice a month	Count	16	16	32
		Expected Count	16.2	15.8	32.0
		Adjusted Residual	-.1	.1	
Total	Count	76	74	150	
	Expected Count	76.0	74.0	150.0	

▶ In the second table (below), the Pearson **Chi-Square** row shows the result to be significant at the 99% level since $p = 0.008$ (i.e. $p < 0.01$) and checking the 'small print' beneath the table shows the test to be **valid**. It meets the criteria for larger than 2 by 2 tables:

- No cell may have an expected count less than 1 – the minimum is **15.79**.
- No more than 20% of cells may have a count less than 5 – here it is **0%**.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9.778 ^a	2	.008
Likelihood Ratio	9.917	2	.007
Linear-by-Linear Association	3.186	1	.074
N of Valid Cases	150		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.79.

- ▶ We conclude that there is a significant difference in Male and Female's responses to this question i.e. there is an **association** between gender and 'usage' (as measured here).
- ▶ There is still one question to ask: How strong is the association between the two?
- ▶ To answer this we turn to the third table which was generated because we ticked the **Phi and Cramér's V** box.

		Value	Approx. Sig.
Nominal by Nominal	Phi	.255	.008
	Cramer's V	.255	.008
N of Valid Cases		150	

- ▶ The crosstabs table we have analysed is bigger than 2 by 2 (it was originally 4 by 2 and then recoded to 3 by 2). Therefore the relevant test of association is Cramér's V. Its **value** is a measure of strength of association. (It's **Approx, Sig.** is the same as for **Pearson Chi-square** in the second table, and can be ignored.)
- ▶ **Cramer's V** varies between 0 and 1 with 0 indicating no association and 1 indicating perfect association (very similar to correlation). Here the value is 0.255 – a moderate level.

[See T24.7 below – Notes on Measures of Strength of Association.]

T24.7 Notes on Measures of Strength of Association

To get a measure of the *strength* of any association we need another statistic:

For a table larger than 2 by 2: **Cramér's V measure of association (0 to 1).**

For a 2 by 2 table (4 cells): **Phi measure of association (-1 to +1 normally).**

N.B. Look at the **Value** (not the **sig.**).

Getting these two statistics is an option in **Crosstabs**, and produces a table entitled **Symmetric Measures**.

These (normally) have absolute values between 0 and 1, with 1 indicating perfect association and 0 signifying no association whatsoever.

Strictly speaking **Phi** and **Cramér's V** are designed for nominal (categorical) data, and other statistics are more specifically designed for ordinal data, but few take notice of that!

In summary, Chi-square tells you whether the table values could be due to chance - and the variables are independent – or, alternatively, that there is good evidence of some real association between the two variables.

Phi or Cramér's V indicates the strength of any detected association.

The larger is N, the smaller is the level of association which can be detected. So Chi-square can deliver a statistically significant verdict but it might not be a significant one!

TUTORIAL 25: Chi-square Test for Frequency Table data

So far, we have used the *SPSS Crosstabs* procedure to analyse raw data (in cases) using the Chi-square Test. In applying **Crosstabs** both a two-way table and the Chi-square results table are produced together. But what if the data is already in a two-way table and (perhaps) the raw data is not available? Although this is a common situation, surprisingly, *SPSS* does not have a procedure to directly perform a Chi-square test on a two-way table, ... but there is a way to do it, which this TUTORIAL illustrates.

We will use the following table for the analysis, and need to enter some of this data into *SPSS*.

		Accessed Coursework Outlines		Total
		Yes	No	
Gender	Male	69	7	76
	Female	59	15	74
Total		128	22	150

Before entering any data, however, it is best to get to understand the table needed (and this is not obvious!).

First we decide on numeric codes to represent Male/Female and Yes/No. This is arbitrary, and we will use:

'1' for 'Male', '2' for 'Female'
'1' for 'Yes', '0' for 'No'

[Actually, we could use strings but it is normal to use numeric codes.]

There are four cells in the above table which contain the essential information:

Cell 1: 'Male' and 'Yes' 69
Cell 2: 'Male' and 'No' 7
Cell 3: 'Female' and 'Yes' 59
Cell 4: 'Female' and 'No' 15

Using the chosen codes (values) instead of the labels this becomes:

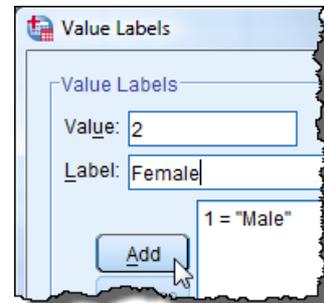
Cell 1: '1' and '1' 69
Cell 2: '1' and '0' 7
Cell 3: '2' and '1' 59
Cell 4: '2' and '0' 15

This is the set of numbers to be entered, each cell being one case. The order does not matter but it is best to be systematic.

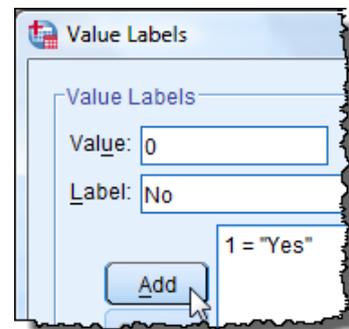
The column above containing the frequencies will be used as **weights** (to be explained).

Now we can begin – to enter the variables and the data:

1. Select **File** → **New** → **Data**
2. Select **Variable View**.
3. In row 1: for **Name** enter 'GENDER', for **Decimals** enter '0', for **Label** enter 'Gender of student'.
4. In row 1: open the **Value Labels** window and enter '1' for 'Male' and '2' for 'Female'.
5. In row 1: for **Measure** choose 'Nominal'.



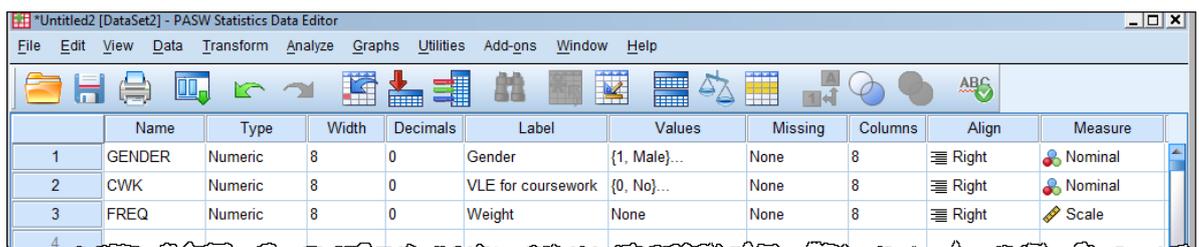
6. In row 2: for **Name** enter 'CWK', for **Decimals** enter '0', for **Label** enter 'Used VLE for coursework'.
7. In row 2: open the **Value Labels** window and enter '1' for 'Yes' and '0' for 'No'.



8. In row 2: for **Measure** choose 'Nominal'.

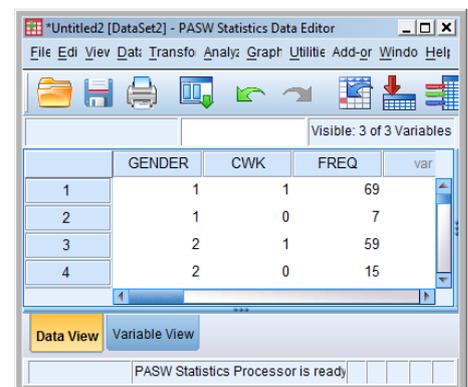
9. In row 3: for **Name** enter 'FREQ', for **Decimals** enter '0', for **Label** enter 'Weight' and for **Measure** choose 'Scale'

► The **Variable View** window should now appear like this:



10. Select **Data View**.

11. In row 1 enter: 1 1 69
12. In row 2 enter: 1 0 7
13. In row 3 enter: 2 1 59
14. In row 4 enter: 2 0 15



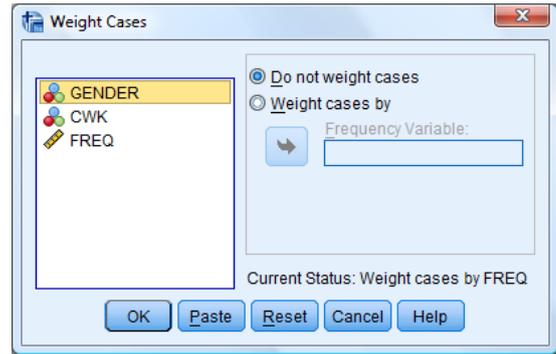
► The window should now appear like this →

► However, if the **View** menu option **Value Labels** has been selected then the labels rather than the codes will be shown (look in the **View** menu to change this if you wish).

- It just remains to tell SPSS to weight the data according to the **FREQ** variable (the last column of numbers). Do this as follows:

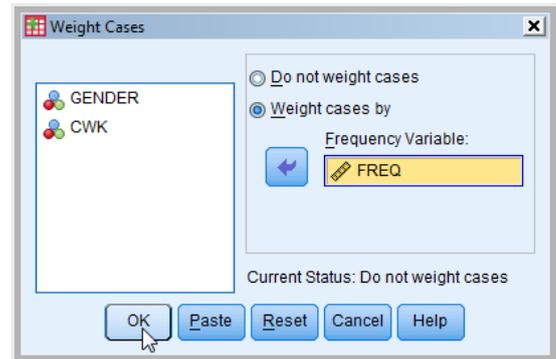
Select **Data** → **Weight Cases**.

- ▶ The window will appear like this →



- Click on the **Weight cases by** radio button and move **FREQ** into the **Frequency Variable** box.

- ▶ The window will appear like this →



- Click **OK**.

- ▶ This advisory message will appear in the Information area at the bottom of the **Data Editor** window →



- ▶ We are now ready to perform the Chi-square test.

- Analyze** → **Descriptive Statistics** → **Crosstabs**

- Move the variable **GENDER** into the **Row(s)** box.

- Move the variable **CWK** and move it into the **Column(s)** box.

- Click on **Statistics** and select **Chi-square**.

- Click **Continue**.

- Click **OK**.

- ▶ The crosstabulation and Chi-square report are generated:

- ▶ The crosstabs table below is essentially the same as that given at the start of this TUTORIAL. The difference is that here 'No' come before 'Yes' because the chosen code for 'No' is numerically less than that for 'Yes'.

Gender of student * Use VLE for coursework Crosstabulation

Count		Use VLE for coursework		Total
		No	Yes	
Gender of student	Male	7	69	76
	Female	15	59	74
Total		22	128	150

- The table below presents the Chi-square Test results.

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.664 ^a	1	.056		
Continuity Correction ^b	2.834	1	.092		
Likelihood Ratio	3.732	1	.053		
Fisher's Exact Test				.067	.045
Linear-by-Linear Association	3.640	1	.056		
N of Valid Cases	150				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.85.
 b. Computed only for a 2x2 table

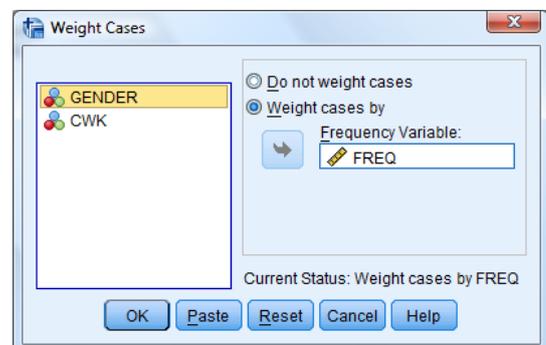
- The footnote 'a' shows that the test satisfies the criteria for validity.
- The footnote 'b' alerts us to the fact that because it is a 2 by 2 table an extra row called **Continuity Correction** has been included. In such cases it is this row which provides the correct significance level (**Asymp. Sig.**), not the first row (**Pearson Chi-Square**). The value is $p = 0.92$ which is not statistically significant because $p > 0.05$.

24. **VERY IMPORTANT** Having finished this analysis, which used weighted data, it is essential to turn off the weighting otherwise it could cause false results in future analyses.

Do this as follows:

Select **Data** → **Weight Cases**.

- The **Weight Cases** window will appear →



25. Remove **FREQ** from the **F**requency Variable box.
26. Click on the **Do not weight cases** radio button. (This step is essential, whereas step 25 is optional but done for completeness.)
27. Click **OK**.
- The 'Weight on' advisory message will disappear from the bottom of the **Data Editor** window.

TUTORIAL T26: One-sample Chi-square Test – Goodness of Fit

T26.1 The One-sample Chi-square Test – introduction

This test analyses a variable and compares its value frequencies with some predetermined frequencies to see if the proportions are more-or-less the same, or differ significantly.

For example, the ACORN profile system for SOCIAL CLASS developed by CACI Ltd has the following broad classes:

Class	Description	UK 2010
1	Wealthy Achievers	25.3%
2	Urban Prosperity	11.6%
3	Comfortably Off	26.9%
4	Modest Means	13.9%
5	Hard-Pressed	20.7%

[See details for this classification system and recent results – *per* businessballs.co.uk - in the final section of the Appendix to this Guide.]

The question we address is: 'How does the university student population compare?'

The **One-sample Chi-square Test** is found by selecting:

Analyze → Nonparametric Tests → Legacy Dialogs → Chi-square

This test can be used in two forms:

Comparing the distribution of frequencies with

- (a) a uniform distribution (i.e. all frequencies the same)
- (b) a set of frequencies typed in by the user.

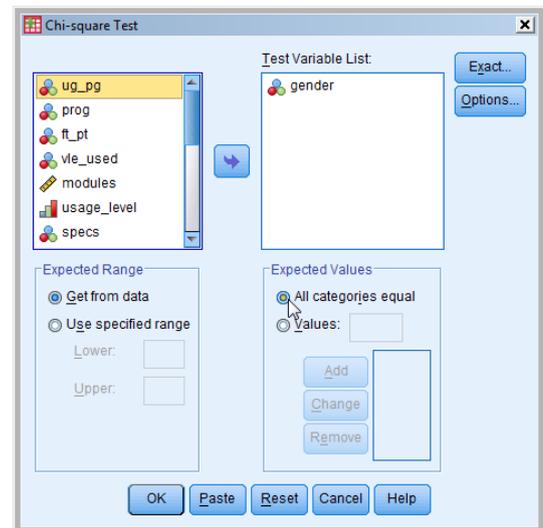
T26.2 The One-sample Chi-square Test – expected category values equal

In the data file **DATA03_LSquestionnaire.sav** are the responses by 150 students to the question named usage_level:

Approximately, how frequently do you use the VLE to access module information during Semesters?
<input type="checkbox"/> Several times a week <input type="checkbox"/> Once or twice a week <input type="checkbox"/> Once or twice a month <input type="checkbox"/> Less than once a month

Firstly, we test to see if the numbers of males and females is about equal or differs significantly.

1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.
2. Select **Analyze** → **Nonparametric Tests** → **Legacy Dialogs** → **Chi-square**
 - ▶ The **Chi-square Test** window opens →
3. Move variable **gender** into the **Test Variable List**.



4. Check that the **Expected Values** radio button for **All categories equal** is selected.
5. Click **OK**.

- ▶ The result is produced in the **Viewer** Output window →
- ▶ The proportions are remarkably close to '50-50' i.e. 75 in each in this case.

Gender			
	Observed N	Expected N	Residual
Male	76	75.0	1.0
Female	74	75.0	-1.0
Total	150		

- ▶ The reported value for **Asymp. Sig.** is 0.870 (this is 'p').
- ▶ There is no statistically significant difference, because $p > 0.05$.

Test Statistics	
	Gender
Chi-square	.027 ^a
df	1
Asymp. Sig.	.870

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 75.0.

Secondly, we test to see if the distribution of responses to the question on VLE usage is uniform (that would be about 35 to 40 for each answer – very unlikely in this case of course)

6. Select **Analyze** → **Nonparametric Tests** → **Legacy Dialogs** → **Chi-square**
 7. In the **Chi-square Test** window click **Reset**.
 8. Move variable **usage_level** into the **Test Variable List**.
 9. Check that the radio button for **All categories equal** is selected (for **Expected Values**).
 10. Click **OK**.
- ▶ The results in the **Viewer** Output window show that the proportions are very unevenly distributed (with Expected 37.5 for each) and **Asymp. Sig.** is 0.000, so $p < 0.0005$.
 - ▶ As $p < 0.05$, the result is statistically significant at the standard 95% level (in fact, because here $p < 0.001$, the result is statistically significant at the 99.9% level).

Frequency of VLE use

	Observed N	Expected N	Residual
Several times a week	54	37.5	16.5
Once or twice a week	64	37.5	26.5
Once or twice a month	31	37.5	-6.5
Less than once a month	1	37.5	-36.5
Total	150		

Test Statistics

	Frequency of VLE use
Chi-square	62.640 ^a
df	3
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 37.5.

- ▶ Hovering the pointer over the **Test Statistics** output table displays a message indicating that double-clicking will activate it (for editing) →
- ▶ Then double-clicking on the **Asymp. Sig.** value '.000' (i.e. p) reveals that it is not actually zero but

1.6033822461212044E-13

which is 0.00000000000016033822461212044.

- ▶ So, $p = 0.000$ always means: $p < 0.00005$.

Test Statistics

	Frequency of VLE use
Chi-square	62.640 ^a
df	3
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 37.5.

Double-click to activate

T26.3 The One-sample Chi-square Test – Expected category values entered

This continues on using the dataset from T26.2. We test to see if the distribution of responses to the question on VLE usage is similar to the results obtained the previous year when there were 132 student responses, as shown in the table below:

<input type="checkbox"/> Several times a week	43
<input type="checkbox"/> Once or twice a week	44
<input type="checkbox"/> Once or twice a month	28
<input type="checkbox"/> Less than once a month	7

1. Select **Analyze** → **Nonparametric Tests** → **Legacy Dialogs** → **Chi-square**

▶ The **Chi-square Test** window opens →

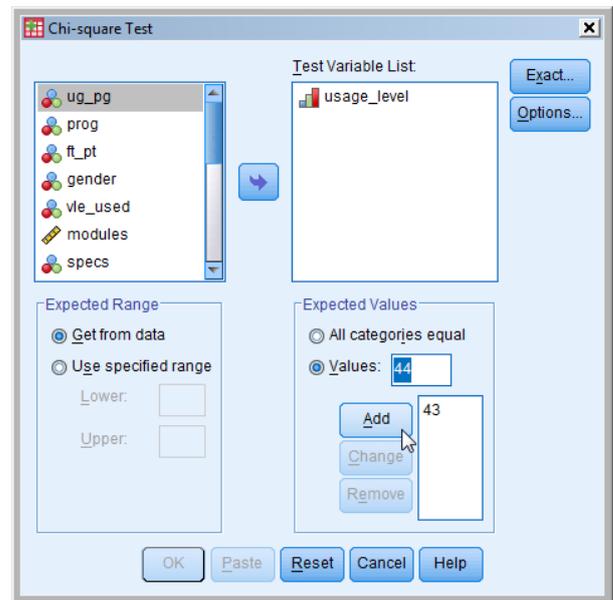
2. Check that variable **usage_level** is in the **Test Variable List**.

3. Click the **Expected Values** radio button for **Values** to select it.

4. Type in the first value which is '43'.

5. Click **Add**.

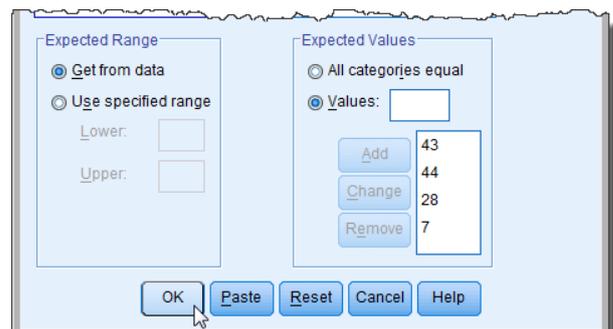
▶ It is very important that the numbers are typed in the correct order matching the order *SPSS* expects, i.e. in the order of increasing value of the codes used for **usage_level** ('1' = 'Several times a week', etc)



▶ The order be checked by entering **Variable View** and opening the **Value Labels** window for the variable **usage_level**, as shown here →



6. Add in the remaining three values '44', '28', '7'.



7. Click **OK**.

- ▶ The results appear in the **Viewer** Output window (below).

Frequency of VLE use

	Observed N	Expected N	Residual
Several times a week	54	52.9	1.1
Once or twice a week	64	54.1	9.9
Once or twice a month	31	34.4	-3.4
Less than once a month	1	8.6	-7.6
Total	150		

Test Statistics

	Frequency of VLE use
Chi-square	8.900 ^a
df	3
Asymp. Sig.	.031

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 8.6.

- ▶ The **Observed N** are the frequencies for the four possible values of **usage_level**.
- ▶ The **Expected N** are the frequencies for the values of **usage_level** if the same proportions as last year had occurred again.
- ▶ The reason why these **Expected N** are not the four numbers typed in is because they have been scaled up (each one multiplied by 150/132) so their total is also 150.
- ▶ **Asymp. Sig.** is 0.031, so $p = 0.031$.
- ▶ As $p < 0.05$ the result is statistically significant at the 95% level (but not at the 99% or 99.9% levels).
- ▶ Interestingly, if the '7' were '6' instead then the result would NOT be statistically significant at the 95% level (try it and see).

TUTORIAL T27: The t Test

T27.1 The t Test formats and criteria for validity

The t test can be used in three formats:

- (a) Comparing two sample means from two different groups to infer if the populations from which they came differ.
This is the **Independent Samples t Test**.
- (b) Comparing the sample mean taken from one group with some specified mean value to infer if the population mean differs from that specified.
This is the **One Sample t Test**.
- (c) Comparing two sample means from one group under two different circumstances ('treatments' or 'conditions') to infer if there is an underlying difference.
This is the **Paired Samples t Test**.

The **Independent Samples t Test** and the **Paired Samples t Test** are very commonly used.

Important criteria are associated with t tests.

- The samples are random and independently selected from the parent population(s).
- The parent population(s) have equal variances (for the Independent Samples Test).
- The data is scale (i.e. interval or ratio) from a continuous normal distribution.

These criteria are rarely fully met in practice!

Within the SPSS Independent Samples Test there is a statistical test to determine whether the amount of deviation from equality of variance is acceptable or not.

A test of normality can be done 'by eye' using the **Frequencies** procedure to draw a Histogram with a normal curve superimposed.

A statistical test of normality can be carried out using the **Kolmogorov-Smirnov Test** (covered in TUTORIAL T30).

Warning Note:

As was said above, the t test looks at **samples** to infer information about **populations** from which the samples are drawn. So, strictly speaking, if you have a class of students and test them on their verbal ability to see if overall the females out-perform the males then no t test is needed. Just look at the means!

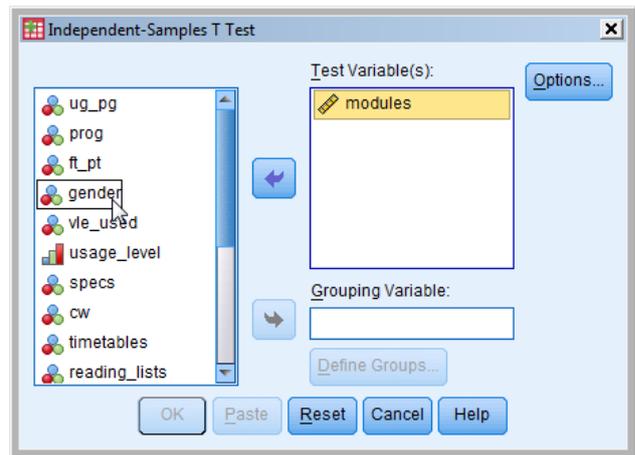
However, if you consider that your students are a representative **sample** of a (well-defined) wider **population** of then a t test could be applied, as there is an **inference** to make about the wider population.

In this TUTORIAL, to keep things simple, we are not drawing a clear distinction between sample and population. We are concentrating on the SPSS procedure.

T27.2 The Independent-Samples *t* Test

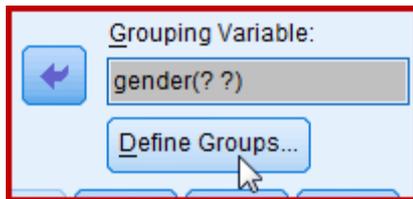
1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.
2. Select **Analyze** → **Compare Means** → **Independent-Samples T Test**

▶ The **T Test** window opens →

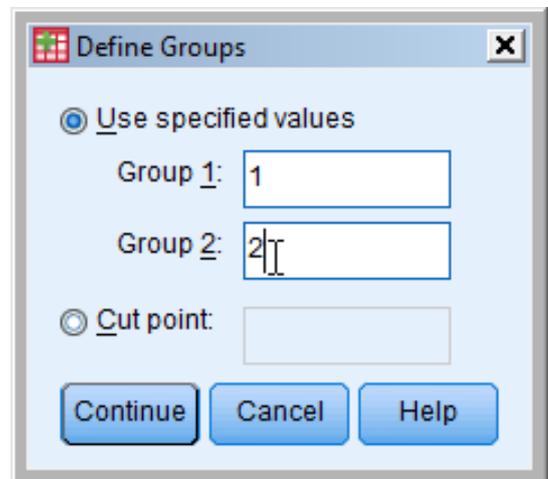


3. Move the scale variable **modules** into the **Test Variable(s)** box.
4. Move the nominal variable **gender** into the **Grouping Variable** box.

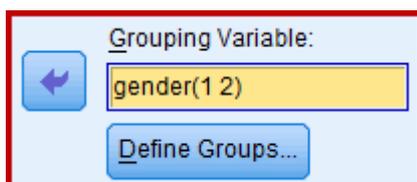
▶ The message 'gender (? ?)' appears. This is asking for the two values for the variable **gender** to be entered:



5. Click **Define Groups**
 - ▶ The **Define Groups** window opens →
6. Enter the codes for **gender**, which in this case are '1' and '2' →



7. Click **Continue**.
 - ▶ The codes for gender now appear in the **Grouping Variable** box:



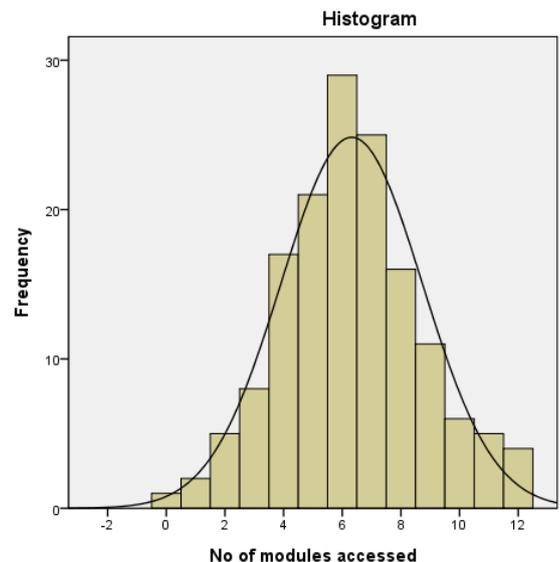
8. Click **OK**.
 - ▶ The result is produced in two tables in the **Viewer** Output window.

- ▶ The first table gives basic descriptive statistics. It shows the means differ, but not by much. It shows the two **Std. Deviation** values (which are the square roots of the two variances) to be almost the same.

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
No of modules accessed	Male	76	6.28	2.480	.284
	Female	74	6.36	2.350	.273

- ▶ There is no information here about whether the distribution of **modules** is normal.
- ▶ To explore this, use **Analyze** → **Descriptive Statistics** → **Frequencies** and, for the **modules** variable choose the **Chart Type** option **Histograms** with 'Show normal curve ...' to get this →
- ▶ It is quite clear that this is close to normal. (Of course it isn't a continuous distribution – like height – and takes just 13 values.)
- ▶ The second table provides the important information whereby to decide whether or not the difference in means is statistically significant.



Independent Samples Test

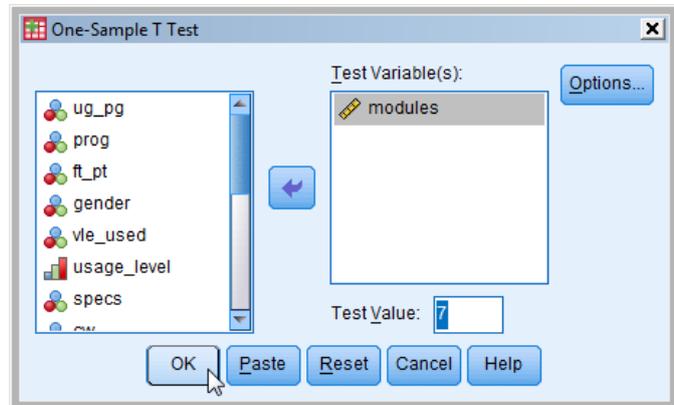
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
No of modules accessed	Equal variances assumed	.040	.841	-.224	148	.823	-.089	.395	-.868	.691
	Equal variances not assumed			-.225	147.893	.823	-.089	.394	-.868	.691

- ▶ **Levene's Test** determines whether the two variances are so different as to be unacceptable. This is shown by the **Sig.** value (what we call p). If $p < 0.05$, then the difference in variances is statistically significant and the test is invalid. Here $p = 0.841$ for the **Levene's Test** so there is no problem. If we did have $p < 0.05$ then all would not be lost – we would simply use the results in the second row 'Equal variances not assumed'
- ▶ The **t** value (-0.224) has a **Sig. (2-tailed)** value of 0.823 (what we call p). So in this case $p > 0.05$, and the difference in means is NOT statistically significant.
- ▶ The **95% Confidence interval** shows the likely range of values for the difference in means. Notice that the range includes zero, showing that no difference is a real possibility. If the 95% Confidence Interval did not include zero then there would be a statistically significant difference.
- ▶ So the conclusion is that there is no evidence for a difference in means.

T27.3 The One-Sample *t* Test

1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav (if not open)
2. Select **Analyze** → **Compare Means** → **One-Sample T Test**

▶ The T Test window opens →



3. Move the scale variable **modules** into the **Test Variable(s)** box.

4. Enter the '7' in the **Test Value** box.

▶ This is going to test whether the mean number of modules taken could be 7.

5. Click **OK**.

▶ The result is produced in two tables in the **Viewer** Output window.

▶ The first table gives basic statistics. It shows the mean is 6.32 which is close to 7.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
No of modules accessed	150	6.32	2.409	.197

▶ The second table provides a **Sig. (2-tailed)** value of 0.001 which is statistically significant ($p < 0.05$). If this were a sample then the conclusion would be that the population mean was not 7.

▶ Notice that here the 95% Confidence Interval does NOT include zero for the difference between the means.

One-Sample Test

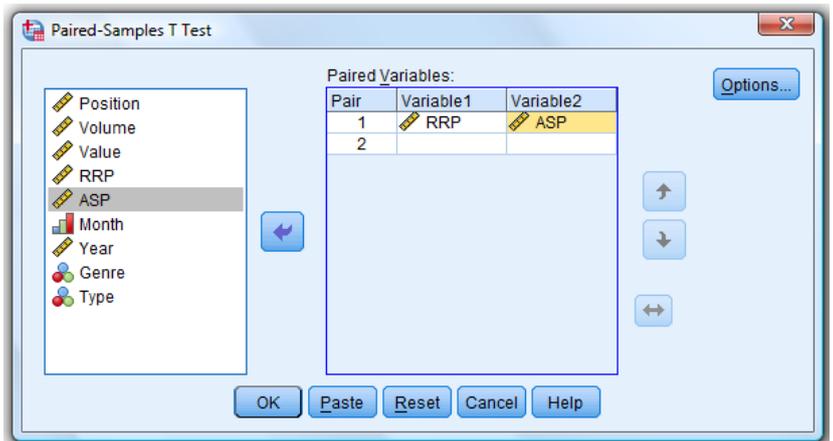
	Test Value = 7					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
No of modules accessed	-3.457	149	.001	-.680	-1.07	-.29

T27.4 The Paired-Samples t Test – scale data

Here we test if the Average Selling Price (ASP) of the Top-selling 100 books is significantly different from the Recommended Retail Price (RRP).

1. Load data file: **File** → **Open** → **Data** → DATA01_100Books.sav
2. Select **Analyze** → **Compare Means** → **Paired-Samples T Test**

► The **Paired-Samples T Test** window opens →



3. Move the scale variable **RRP** into the **Pair 1 Variable 1** box.

4. Move the scale variable **ASP** into the **Pair 1 Variable 2** box.

5. Click **OK**.

► The result is produced in three tables in the **Viewer** Output window.

► The first table shows the pairs of variables being compared and their means etc. It can be seen that the Means are quite different and Std. Deviations are very different.

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 Recommended Retail Price	£11.0812	100	£5.56409	£0.55641
Average Selling Price	£7.2031	100	£2.66229	£0.26623

► The second table is an unexpected bonus! We did not ask for the correlation but we got it all the same. We can see that the variables have an extremely high positive correlation.

► Its **Sig.** of 0.000 means $p < 0.0005$ so the correlation 'definitely' isn't zero, (or rather, expressing it more precisely, if this were a sample then the parent population's correlation couldn't be zero).

Paired Samples Correlations

	N	Correlation	Sig.
Pair 1 Recommended Retail Price & Average Selling Price	100	.927	.000

- ▶ The third table gives the results. The important value is in the **Sig. (2-tailed)** column.
- ▶ The means of ASP and RRP are ‘definitely’ different. In the table **Sig. (2-tailed) = 0.000** which means $p < 0.0005$ so $p < 0.001$ which gives 99.9% significance level.

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Recommended Retail Price - Average Selling Price	£3.87809	£3.25120	£0.32512	£3.23298	£4.52319	11.928	99	.000

- ▶ Double-clicking on the table to activate it (for editing) and then double-clicking on the

Sig. value ‘.000’ reveals that it is not exactly zero but is actually 0.00000000000000000000000731 approximately!

7.307330558572866E-21

T27.5 The Paired-Samples *t* Test – ordinal data

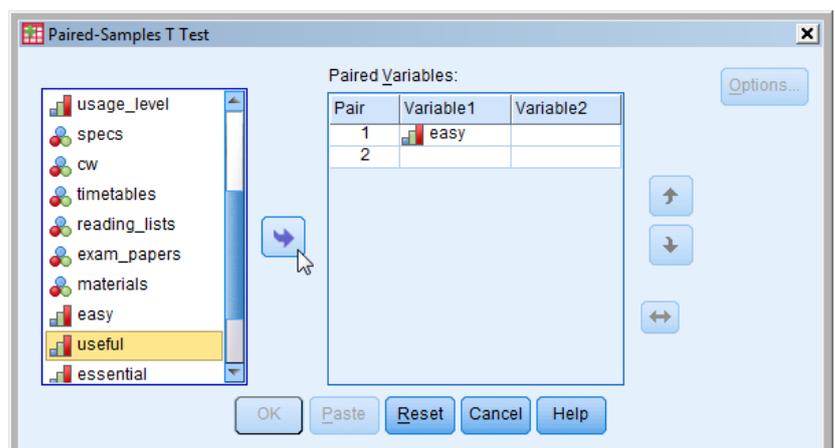
Here we will look at some ordinal data derived from responses to four questions about a university’s VLE, each answered on a 5-point scale. We should really only use scale data but it illustrates the procedure well.

1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav (if not open).
2. Select **Analyze** → **Compare Means** → **Paired-Samples T Test**

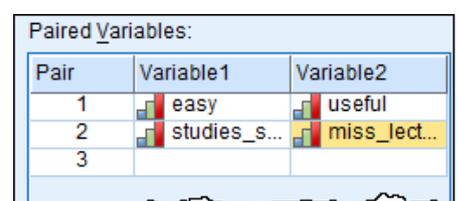
- ▶ The **T Test** window opens →

3. Move the ordinal variable **easy** into the **Pair 1 Variable 1** box.
4. Move the ordinal variable **useful** into the **Pair 1 Variable 2** box.

- ▶ We enter a second pair to test at the same time.



5. Move the ordinal variable **studies_suffer** into the **Pair 2 Variable 1** box.
6. Move the ordinal variable **miss_lectures** into the **Pair 2 Variable 2** box.



7. Click **OK**.

- ▶ The result is produced in three tables in the **Viewer** Output window.
- ▶ The first table produced (shown below) displays the pairs of questions being compared and their means.
- ▶ It can be seen that for the first pair the Means and Std. Deviations look very similar, but not so for the second pair look.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	VLE is easy to use	4.03	150	.763	.062
	VLE has useful information	4.13	150	.717	.059
Pair 2	Studies would suffer without the VLE	3.95	150	.933	.076
	Don't mind missing lectures if they're on the VLE	2.98	150	1.298	.106

- ▶ The second table provides correlations. We can see that the first pair have a moderately strong positive correlation (0.495). It's **Sig.** of 0.000 implies $p < 0.0005$ which means it 'definitely' isn't from a population with a zero correlation. In contrast the second pair have a negligible correlation – which may well be considered zero.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	VLE is easy to use & VLE has useful information	150	.495	.000
Pair 2	Studies would suffer without the VLE & Don't mind missing lectures if they're on the VLE	150	-.051	.537

- ▶ From the Pair 1 correlation (+0.495, Sig. = 0.000) we conclude that students who said the VLE was easy to use also generally said that the VLE contained useful information.

In contrast...

- ▶ From the Pair 2 correlation (−0.051, Sig. = 0.537) we conclude that students who said their studies would suffer without the VLE were not generally the same as those who said they did not mind missing lectures (N.B. there is no relationship either way as the correlation is so low).

- ▶ The third table gives the *t* Test results – the important values are in the **Sig. (2-tailed)** column.
- ▶ With the *t* Test, if the data is coded, it is important to be clear which way the coding goes.
- ▶ Here the stronger the agreement with the statement the bigger the number.

1 = "Strongly disagree"
2 = "Disagree"
3 = "Neutral"
4 = "Agree"
5 = "Strongly agree"

- ▶ Also important with the *t* Test it is to be clear what the sign of the difference of means signifies i.e. which number is being subtracted from which.
- ▶ For Pair 1 the calculation is: 'easy' – 'useful', which is 4.03 – 4.13, which is negative.

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	VLE is easy to use - VLE has useful information	-.093	.745	.061	-.214	.027	-1.534	149	.127
Pair 2	Studies would suffer without the VLE - Don't mind missing lectures if they're on the VLE	.967	1.636	.134	.703	1.231	7.236	149	.000

- ▶ The means of the first pair are not found to be significantly different as the difference is small (– 0.93) (**Sig.** = $p = 0.127$ so $p > 0.05$).
- ▶ The means of the second pair are 'definitely' different as the difference is large (+0.967) (**Sig.** $p = 0.000$ means $p < 0.005$ so $p < 0.001$).
- ▶ From the Pair 1 *t* Test (Sig. = 0.127) we conclude that students on average rated the VLE's ease of use about the same as it's having useful information (the difference in average rating is very small: i.e. only –0.093 on a 5-point scale).
- ▶ From the Pair 2 *t* Test (Sig. = 0.000) we conclude that students on average rated suffering without the VLE higher than not minding missing lectures (the difference in average rating is quite large: i.e. +0.967 on a 5-point scale).
- ▶ Warning note: This particular example illustrates the problems with having NEGATIVE statements like 'Don't mind ...' – the results can be quite hard to interpret as you may find! Avoid negative statements in questionnaires if you can as they often lead to 'double negatives' when interpreting.
- ▶ Finally, the responses were compared *in pairs*. It would be nice, say, to compare all four together. That is not possible with a *t* Test ... but that kind of analysis can be done using One-Way Analysis of Variance (ANOVA), which is the subject of T29.

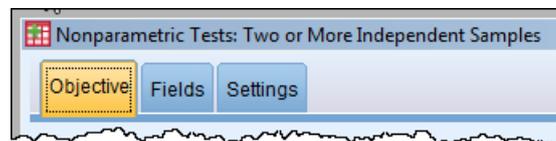
TUTORIAL T28: Nonparametric alternatives to the *t* Test

If the conditions for the validity of the *t* test are not met (or it is not known that they are) then a nonparametric alternative should be used, as illustrated here.

T28.1 The Mann-Whitney Rank-sum Test – for independent samples

1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav (if not open).
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.
2. Select **Analyze** → **Nonparametric Tests** → **Independent Samples**

▶ This window opens →



3. Click on **Fields** to open the **Nonparametric Tests** window:
4. Move the scale variable **modules** into the **Test Fields** box.
5. Move the nominal variable **gender** into the **Groups** box.
6. Click on **Settings** to open the choice of tests.
7. Select **Mann-Whitney U (2 samples)**.
8. Click **Run**.

▶ This output appears:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of No of modules accessed is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.625	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

- ▶ The result is that the difference is not significant ($p = 0.625$) so the Null Hypothesis is to be retained. I.e. there is no evidence of a difference in the number of modules males and females choose.
- ▶ This is the same conclusion as was reached using the Independent Samples *t* Test in T27.2.
- ▶ If you do not perform step 6 (to enter **Settings** and choose which test to use) then *SPSS* will make the decision for itself. It may use the Kruskal-Wallis One-Way ANOVA test (see T29.2) which gives exactly the same result.

T28.2 The Wilcoxon Matched-pairs Signed-ranks Test – for paired samples

This test requires the variables to be scale. It will work if ordinal variables are reassigned to scale, where appropriate.

Here we test if the Average Selling Price (ASP) of the Top-selling 100 books is significantly different from the Recommended Retail Price (RRP). This was done in T27.4 using a *t* Test.

1. Load data file: **File** → **Open** → **Data** → DATA01_100Books.sav
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.

2. Select **Analyze** → **Nonparametric Tests** → **Related Samples**

- ▶ This window opens →



3. Click on **Fields**.
4. Move the scale variable **RRP** into the **Test Fields** box.
5. Move the scale variable **ASP** into the **Test Fields** box.
6. Click **Run**.

- ▶ This output appears:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Recommended Retail Price and Average Selling Price equals 0.	Related-Samples Wilcoxon Signed Rank Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

- ▶ The result is that the difference is highly significant (**Sig.** = 0.000 so $p < 0.001$) so the Null Hypothesis is to be rejected.
- ▶ This is the same conclusion as was reached using the Independent Samples *t* Test in T27.4.

TUTORIAL T29: Analysis of Variance (ANOVA)

T29.1 Introduction to Analysis of Variance

Analysis of Variance (ANOVA) and its more complex forms – ANCOVA, MANOVA, and MANCOVA – provide a very powerful set of methods for comparing sample means to see if there is evidence to infer that the underlying populations from which they are derived are different. However, these methods can be complex both conceptually and procedurally. This Guide only introduces some of the more basic methods and cannot do justice to the underlying statistics. A good statistics textbook or SPSS textbook is essential.

ANOVA is a procedure for comparing sample means for one dependent variable (scale data – e.g. statistics exam mark) for one or more independent variables (categorical data, also known as nominal data – e.g. gender) to see if there is statistically significance difference from which one can infer that the populations from which these samples came themselves are different.

ANOVA is called a univariate method because it has one dependent variable (e.g. overall exam mark).

The dependent variable must always be scale (= interval or ratio). [Other criteria are that the underlying populations from which samples are drawn should be normal, variances equal, sampling random.]

It is one-way ANOVA if there is just one independent variable (e.g. GCSE English grade). It is two-way ANOVA if there are two independent variables (e.g. gender and racial group).

The independent variable must always be categorical (= nominal). It may take just two values (e.g. male or female) or several (e.g. racial group defined as caucasian / black / asian / hispanic etc.).

The *t* test is the special case of one-way ANOVA when the independent variable takes only two values.

MANOVA – Multivariate Analysis of Variance – is an extension of ANOVA when there is more than one dependent variable.

ANCOVA – Analysis of Covariance – and MANCOVA – Multivariate Analysis of Covariance are variants of the above when covariation between the variables is taken into account.

There are two basic scenarios:

Independent measures: when two or more groups of subjects undergo exactly the same ‘experience’ – e.g. male and female students take a calculus exam. Here gender is the independent variable – also called a ‘factor’ (having two levels). ANOVA can test whether *in general* males and females would have the same level of performance in the exam. This is referred to as ‘between-subjects’ as it looks at the differences found between different groups of subjects.

Repeated measures: when each subject experiences more than one level of a factor – e.g. all students on a module take a test on a topic both before and after doing a practical on that topic. ANOVA can test whether *in general* the practical would have an effect upon test performance. This is referred to as ‘within-subjects’ as it looks at the differences found within individual subjects’ performances ‘before’ and ‘after’. [Another example would be students on a programme all studying the same six modules, in which case ANOVA could test for differences in the module results.]

This Guide presents seven basic methods:

- One-way ANOVA for independent measures (‘between-subjects’) (Post Hoc) – parametric
- One-way ANOVA for independent measures (‘between-subjects’) (Contrasts) – parametric
- One-way ANOVA for independent measures (‘between-subjects’) (Kruskal-Wallis) – nonparametric
- One-way ANOVA for repeated measures (‘within-subjects’) – parametric
- One-way ANOVA for repeated measures (‘within-subjects’) (Friedman) – nonparametric
- Two-way ANOVA for independent measures (‘between-subjects’) – parametric
- Two-way ANOVA for repeated measures (‘within-subjects’) – parametric

T29.2 One-Way between-subjects ANOVA (independent measures) – Post Hoc

Here we investigate the amount to which students used their department's VLE for support. The students were on one of these six programmes:

Code	UG or PG	Programme name
1	UG	LS – Library Studies
2	UG	IM – Information Management
3	UG	PB – Publishing
4	PG	ILM – Information & Library Management
5	PG	IKM – Information & Knowledge Management
6	PG	EPB – Electronic Publishing

The question asked is: “Does the number of modules for which a student used the VLE for support vary significantly from programme to programme?”

As there is one dependent variable (modules), it is an ANOVA (not an MANOVA).

As there is just one independent variable (programme), it is a one-way ANOVA.

As different students were on different programmes, it is a between-subjects one-way ANOVA.

Note: For this first introduction to ANOVA we take the simpler approach:

Analyze → Compare Means → One-Way ANOVA.

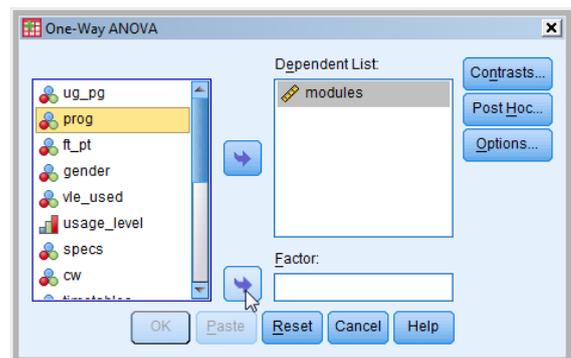
If we took the more general approach we would use

Analyze → General Linear Model → Univariate.

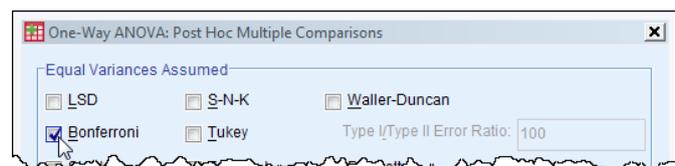
1. Load data file: **File → Open → Data → DATA03_LSquestionnaire.sav** (if not open)
2. Select **Analyze → Compare Means → One-Way ANOVA**

► The **One-Way ANOVA** window opens →

3. Move the scale variable **modules** into the **Dependent List** box.
4. Move the nominal variable **prog** into the **Factor** box.

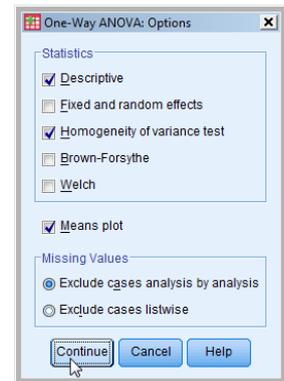


5. Click the **Post Hoc** button and select **Bonferroni**, one of the most conservative of the many possible choices (**Scheffé** is also popular).



- The **Post Hoc** option produces a comparison of all possible pairs of factors.
- The term *post hoc* literally means ‘after the fact’ and signifies that no decision is made as to what to compare before the analysis takes place. The alternative approach – deciding beforehand – is known as ‘Contrasts’ (see T29.3).

6. Click **Continue**.
7. Click the **Options** button and select
Descriptive
Homogeneity of variance test
Means plot
8. Click **Continue**.
9. Click **OK** to generate four output tables and a chart.



- ▶ Table 1 (**Descriptives**) displays the statistics Count, Mean, Std Deviation, etc. for the **modules** variable, broken down by the six categories of the **prog** variable.
- ▶ The **Descriptives** information is a good first place to look for similarities and differences of the two most important statistics – Mean and Standard Deviation (square root of Variance).

No of modules accessed

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Library Studies	29	6.59	2.486	.462	5.64	7.53	2	12
Information Management	36	7.28	2.398	.400	6.47	8.09	1	12
Publishing	22	5.00	1.718	.366	4.24	5.76	2	9
Information & Know. Man.	31	6.61	2.155	.387	5.82	7.40	1	12
Information & Library Man.	27	6.07	2.368	.456	5.14	7.01	2	12
Electronic Publishing	5	3.20	2.049	.917	.66	5.74	0	5
Total	150	6.32	2.409	.197	5.93	6.71	0	12

- ▶ Table 2 (**Test of Homogeneity of Variances**) displays the result of **Levene's Test** (which was met in the **Independent-Samples t Test** – see T27.2).
- ▶ The **Sig.** value of 0.459 (i.e. $p = 0.459$) means $p > 0.05$ so there is no problem here – equality of variances can be assumed.

Test of Homogeneity of Variances

No of modules accessed

Levene Statistic	df1	df2	Sig.
.937	5	144	.459

- ▶ Table 3 (**ANOVA**) has the most important result, the ANOVA **F** value and its **Sig.** value.
- ▶ In this case **Sig.** = 0.000 (so $p < 0.001$) and the result is significant at the 99.9% level. There is very strong evidence that the (population) means are not all the same – i.e. there is a lot of variation between the groups (programmes).

ANOVA

No of modules accessed

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	126.377	5	25.275	4.930	.000
Within Groups	738.263	144	5.127		
Total	864.640	149			

- ▶ For those interested, the **F** statistic is the ratio of **Between Groups Sum of Squares** and **Within Groups Sum of Squares**. So its magnitude measures whether most variation is between different groups or between individuals within the groups.
- ▶ Table 4 (**Multiple Comparisons**) is very large because it compares every one of the categories with all the others (twice actually!). The Bonferroni test (there are many others – consult a statistics textbook and take your pick) indicates which pairs differ significantly – i.e. where **Sig.** is less than 0.05. An asterisk against the Mean Difference value highlights them. Bonferroni is a conservative test. Some prefer LSD.
- ▶ In this case there are 15 different comparisons – that's 30 in the table.
- ▶ An asterisk against the Mean Difference indicates a significant pair – there are four such pairs here, each reported twice.

Multiple Comparisons

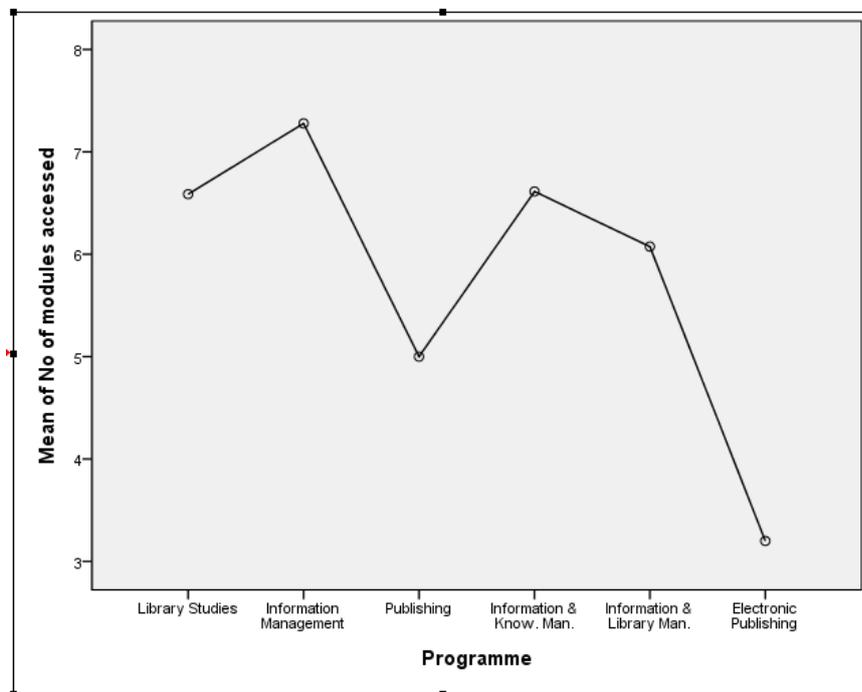
No of modules accessed
Bonferroni

(I) Programme	(J) Programme	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Library Studies	Information Management	-.692	.565	1.000	-2.38	.99
	Publishing	1.586	.640	.216	-.32	3.50
	Information & Know. Man.	-.027	.585	1.000	-1.77	1.72
	Information & Library Man.	.512	.606	1.000	-1.30	2.32
	Electronic Publishing	3.386*	1.096	.036	.11	6.66
Information Management	Library Studies	.692	.565	1.000	-.99	2.38
	Publishing	2.278*	.613	.004	.45	4.11
	Information & Know. Man.	.665	.555	1.000	-.99	2.32
	Information & Library Man.	1.204	.576	.578	-.52	2.92
	Electronic Publishing	4.078*	1.081	.004	.85	7.30
Publishing	Library Studies	-1.586	.640	.216	-3.50	.32
	Information Management	-2.278*	.613	.004	-4.11	-.45
	Information & Know. Man.	-1.613	.631	.175	-3.50	.27
	Information & Library Man.	-1.074	.650	1.000	-3.02	.87
	Electronic Publishing	1.800	1.122	1.000	-1.55	5.15
Information & Know. Man.	Library Studies	.027	.585	1.000	-1.72	1.77
	Information Management	-.665	.555	1.000	-2.32	.99
	Publishing	1.613	.631	.175	-.27	3.50
	Information & Library Man.	.539	.596	1.000	-1.24	2.32
	Electronic Publishing	3.413*	1.091	.032	.16	6.67
Information & Library Man.	Library Studies	-.512	.606	1.000	-2.32	1.30
	Information Management	-1.204	.576	.578	-2.92	.52
	Publishing	1.074	.650	1.000	-.87	3.02
	Information & Know. Man.	-.539	.596	1.000	-2.32	1.24
	Electronic Publishing	2.874	1.102	.151	-.42	6.16
Electronic Publishing	Library Studies	-3.386*	1.096	.036	-6.66	-.11
	Information Management	-4.078*	1.081	.004	-7.30	-.85
	Publishing	-1.800	1.122	1.000	-5.15	1.55
	Information & Know. Man.	-3.413*	1.091	.032	-6.67	-.16
	Information & Library Man.	-2.874	1.102	.151	-6.16	.42

*. The mean difference is significant at the 0.05 level.

- ▶ The conclusions are that:
 - (a) Electronic Publishing has a significantly different mean from three programmes – Library Studies, Information Management, Information & Knowledge Management. [Of course, there are only 5 students on that programme so it is not a very robust conclusion.]
 - (b) Publishing has a significantly different mean from one programme – Information Management.

- ▶ The fifth output is a simple graph – **Means Plots** – which is a line chart of all the categories' means. This may not seem a very appropriate choice of chart, but that's what *SPSS* provides for ANOVA.

Means Plots

T29.3 One-Way between-subjects ANOVA (independent measures) – Contrasts

Here we repeat the analysis in T29.2 but this time do not use **Post Hoc** (after the fact) comparisons but instead use **Contrasts** (comparisons decided in advance). The six programmes in this analysis are:

Code	UG or PG	Programme name
1	UG	LS – Library Studies
2	UG	IM – Information Management
3	UG	PB – Publishing
4	PG	ILM – Information & Library Management
5	PG	IKM – Information & Knowledge Management
6	PG	EPB – Electronic Publishing

In T29.2 the basic question was “Does the number of modules for which a student used the VLE for support vary significantly from programme to programme?” The **Post Hoc** test compared all possible pairs of programmes and produced answers to this question. No decision as to which pairs to look at was made until after the analysis.

Here we will ask in advance two questions, which the **Post Hoc** method did not (could not) answer.

Q1: “Is Publishing different from the two other undergraduate programmes *taken together*?”

This is called a **contrast** because we want to contrast PB with LS+IM.

For a contrast we have to assign **weights**: the same positive integer for each member of one group, and the same negative integer for each member of the contrasting group. The potentially tricky part is that the sum of all these weights must be zero.

For Q1 we can choose the following simple weights: **Group 1**: LS: +1, IM: +1 **Group 2**: PB: –2. We do not want to involve the other three programmes at all, so assign them zero weight. We have:

Q1			
Code	UG or PG	Programme name	Weight
1	UG	LS – Library Studies	1
2	UG	IM – Information Management	1
3	UG	PB – Publishing	–2
4	PG	ILM – Information & Library Management	0
5	PG	IKM – Information & Knowledge Management	0
6	PG	EPB – Electronic Publishing	0

Q2: “Are the three undergraduate programmes *taken together* different from the two postgraduate programmes (excluding EPB) *taken together*?” [We exclude EPB has only 5 students.]

We want to contrast LS+IM+PB with ILM+IKM so deciding the weights here is less obvious than in Q1.

For Q2 we can choose these weights: **Group 1**: LS: +2, IM: +2 PB +2 **Group 2**: ILM: –3, IKM: –3.

Q2			
Code	UG or PG	Programme name	Weight
1	UG	LS – Library Studies	2
2	UG	IM – Information Management	2
3	UG	PB – Publishing	2
4	PG	ILM – Information & Library Management	–3
5	PG	IKM – Information & Knowledge Management	–3
6	PG	EPB – Electronic Publishing	0

These are the simplest weight choices for Q1 and Q2, but there infinitely many equivalent possibilities.

Having decided in advance on the contrasts, we now proceed with the analysis.

1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav (if not open)
2. Select **Analyze** → **Compare Means** → **One-Way ANOVA**
3. Click **Reset** and move the scale variable **modules** into the **Dependent List** box.

4. Move the nominal variable **prog** into the **Factor** box.
5. Click the **Contrasts** to open this window →

- ▶ The weights for Q1 can now be entered for **Contrast 1**.

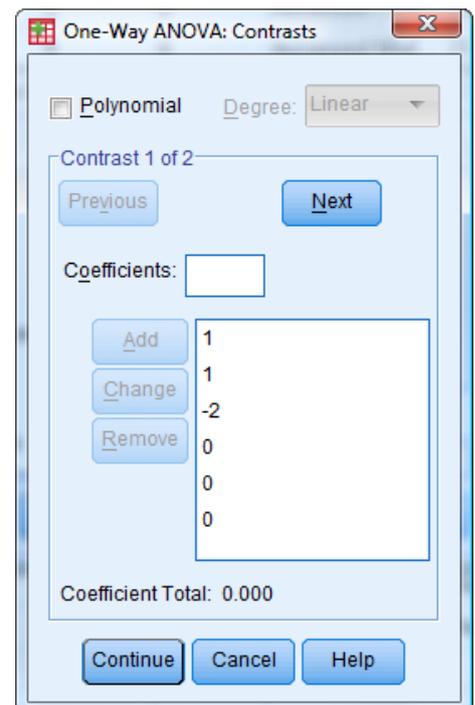
6. Type '1' in the **Coefficients** box and click **Add**.
7. Type '1' in the **Coefficients** box and click **Add**.
8. Type '-2' in the **Coefficients** box and click **Add**.
9. Type '0' in the **Coefficients** box and click **Add**.
10. Type '0' in the **Coefficients** box and click **Add**.
11. Type '0' in the **Coefficients** box and click **Add**.

- ▶ The weights for Q1 have now been entered →

- ▶ This is **Contrast 1** completed.

- ▶ Note that the Coefficient Total is zero, as required.

- ▶ Note that the association of the weights with the programmes is done by the order in which the weights are entered. This must correspond with the numeric codes (1 to 6) assigned to the programmes, as shown in the Q1 table on the previous page.

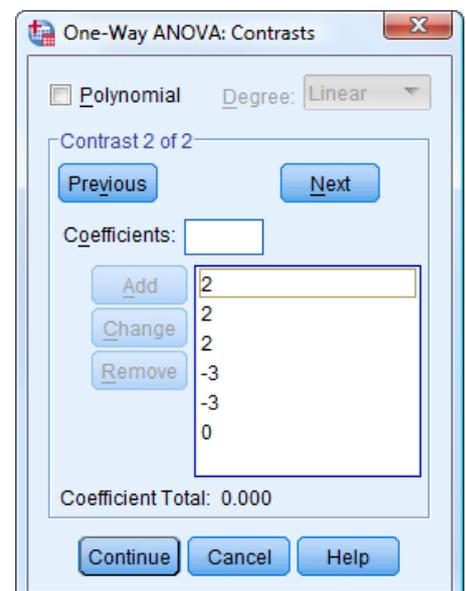


12. Click **Next** to enter the weights for **Contrast 2** →

13. Repeat steps 6 to 11, entering 2, 2, 2, -3, -3, 0 →

- ▶ Note that the Coefficient Total is zero, as required.

14. Click **Continue** →



15. Click the **Options** button and select **Homogeneity of variance test**.

16. Click **Continue**.

17. Click **OK**.

- ▶ Four tables are produced. The first is an ANOVA table (not shown) which we ignored here.
- ▶ Table 1 (**Test of Homogeneity of Variances**) reports Levene’s Test result. This determines which line of the **Contrast Tests** table we read (shown later) to find the significance.
- ▶ In this case Levene’s Test result is not significant as **Sig.** > 0.05, so equality of variances can be assumed.

Test of Homogeneity of Variances

No of modules accessed

Levene Statistic	df1	df2	Sig.
.937	5	144	.459

- ▶ Table 2 (**ANOVA** – not shown) we can ignore here.
- ▶ Table 3 (**Contrasts Coefficients**) presents the weights used in each Contrast. It is a good idea to confirm that these are what you wanted!

Contrast Coefficients

Contrast	Programme					
	Library Studies	Information Management	Publishing	Information & Know. Man.	Information & Library Man.	Electronic Publishing
1	1	1	-2	0	0	0
2	2	2	2	-3	-3	0

- ▶ Table 4 (**Contrast Tests**) is the most important and presents the results for each Contrast.
- ▶ We can assume equal variances here, so read from the top two lines.

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
No of modules accessed	Assume equal variances	1	3.86	1.119	3.454	144	.001
		2	-.33	2.325	-.143	144	.886
	Does not assume equal variances	1	3.86	.954	4.051	51.479	.000
		2	-.33	2.291	-.145	111.695	.885

- ▶ For Contrast 1 the result is significant (**Sig.** = 0.001).
So Publishing is different from the two other undergraduate programmes *taken together*.
- ▶ For Contrast 2 the result is not significant (**Sig.** = 0.886).
The three undergraduate programmes *taken together* are not different from the two postgraduate programmes (excluding EPB) *taken together*.

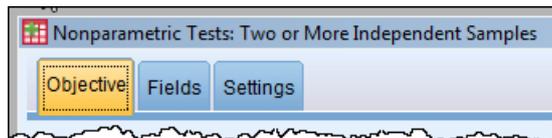
T29.4 One-Way between-subjects ANOVA – Kruskal-Wallis nonparametric test

If the criteria for the normal (parametric) ANOVA are seriously violated then a nonparametric version should be used. For the One-Way between-subjects ANOVA SPSS supplies the **Kruskal-Wallis One-Way ANOVA**. As a demonstration, we repeat the analysis just carried out in T29.2 and T29.3.

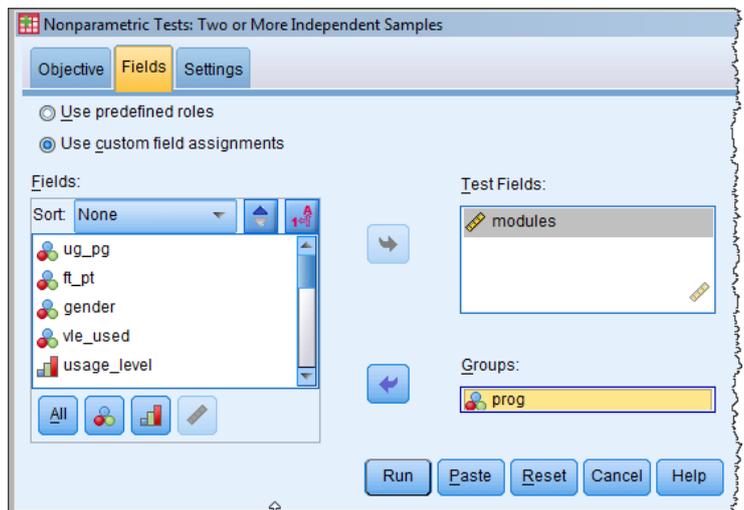
1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav (if not open).
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.

2. Select **Analyze** → **Nonparametric Tests** → **Independent Samples**

▶ This window opens →



3. Select **Fields** to reveal the **Test Fields** box:



4. Move **modules** into the **Test Fields** box →

5. Move **prog** into the **Groups** box →

6. Select **Run**.

- ▶ SPSS automatically determines which test to apply.
- ▶ Alternatively, before clicking on **Run** you could click on **Customize tests** and explicitly select **Kruskal-Wallis 1-way ANOVA (k samples)**.
- ▶ Either way, the output is as follows.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of No of modules accessed is the same across categories of Programme.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

- ▶ The conclusion is that there is a significant difference. **Sig.** = 0.000, so $p < 0.001$ and the significance level is actually much greater than the default 95% (i.e. 99.9%).

T29.5 One-Way within-subjects ANOVA (repeated measures)

Here we investigate whether students' marks are significantly different across four different modules.

As there is one dependent variable (module mark) it is an ANOVA.

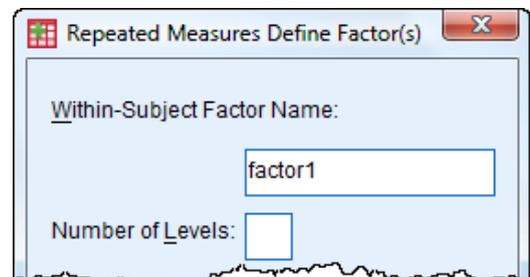
As there is just one independent variable or 'factor' (modules), it is a one-way ANOVA.

As the same students took all four modules it is a within-subjects one-way ANOVA.

1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav (if not open)
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial. Remember to click **Apply** before clicking **OK**.

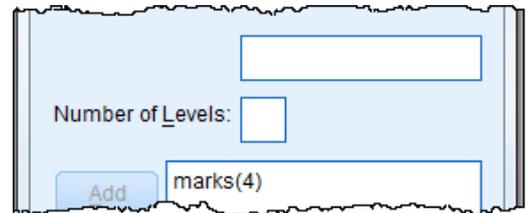
2. Select **Analyze** → **General Linear Model** → **Repeated Measures**

▶ This window opens →

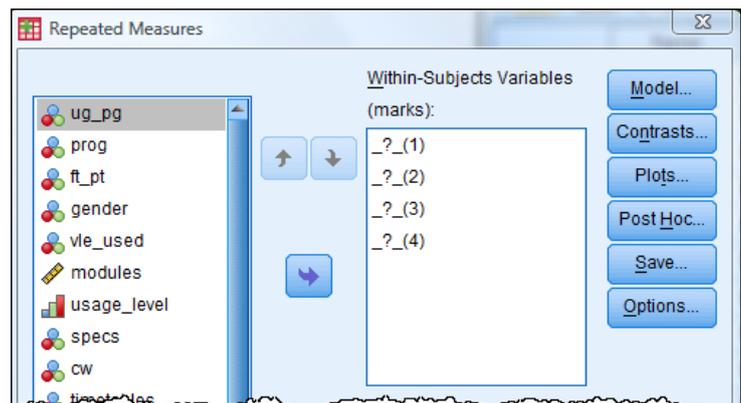


3. Change the default **Within-Subject Factor Name** from 'factor1' to 'marks' →

4. Enter '4' in the **Number of Levels** box, as there are four sets of marks →



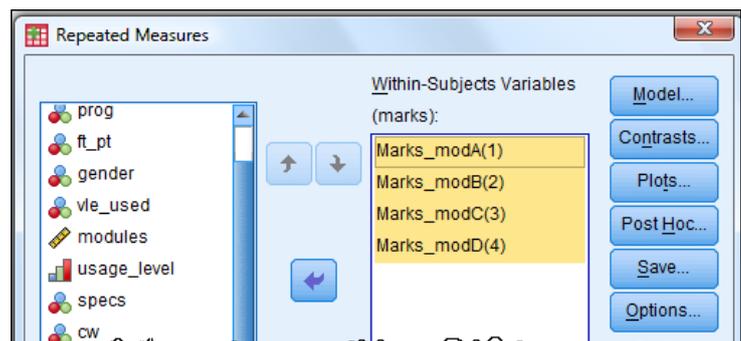
5. Click **Add** to produce this →



6. Click **Define** which produces this window →

▶ The four variables now need to be entered.

7. Scroll through the variables, list and select **Marks_modA** and move it, using the top blue arrow, into the **Within-Subjects Variables** box.



8. Repeat step 7 for **Marks_modB**, **Marks_modC**, **Marks_modD**.

▶ You could move all variables across together – selecting them by clicking on the first and shift-clicking on the last.

- ▶ You can use the blue 'up' and 'down' arrows to alter the order of the variables. The order can matter as they are referred to by number later on ...
- 9. You could now click **Plots** and move 'marks' into the **Horizontal Axis** box, and click **Add** to insert 'marks' into the **Plots** box to eventually produce a simple line plot showing the four means (similar to what was done in T29.2), but we will not do so here.
- 10. Click the **Options** button.
- 11. Move 'marks' into the **Display Means for** box. [This will generate Table 8.]
- 12. Select **Compare main effects**. [This will generate Table 9.]
- 13. Change the **Confidence interval adjustment** method to **Bonferroni**. [An option for Table 9.]
- 14. Select **Descriptive statistics**. [This will generate Table 2.]
 - ▶ This will later produce a table showing the means, sds and counts of the four variables.
 - ▶ Note that the **Significance level** is set at 0.05, so the **Confidence intervals** are 95%.
- 15. Click **Continue**.
- 16. Click **OK** to generate a series of tables.

- ▶ Table 1 (**Within-Subjects Factors**) is generated automatically. It just lists the four dependent variables.

Within-Subjects Factors

Measure:MEASURE_1

marks	Dependent Variable
1	Marks_modA
2	Marks_modB
3	Marks_modC
4	Marks_modD

- ▶ Table 2 (**Descriptive Statistics**) is optional. It displays the dependent variables' means, standard deviations and N (number of cases).
- ▶ We can see that Modules A, B and C look very similar
- ▶ Module D looks very different, having a much lower mean and much greater variability.
- ▶ N = 149 is one less than the expected 150 because one case has some missing values.

Descriptive Statistics

	Mean	Std. Deviation	N
Module A marks	63.47	6.157	149
Module B marks	63.40	6.385	149
Module C marks	64.29	5.440	149
Module D marks	52.70	14.547	149

- ▶ Table 3 (**Multivariate Tests**) is generated automatically. It is not needed unless sphericity is a problem in a multivariate analysis. This is a univariate analysis so it does not apply here. We can ignore this.

Multivariate Tests^b

Effect		Value	F	Hypothesis df	Error df	Sig.
marks	Pillai's Trace	.394	31.687 ^a	3.000	146.000	.000
	Wilks' Lambda	.606	31.687 ^a	3.000	146.000	.000
	Hotelling's Trace	.651	31.687 ^a	3.000	146.000	.000
	Roy's Largest Root	.651	31.687 ^a	3.000	146.000	.000

a. Exact statistic

b. Design: Intercept
Within Subjects Design: marks

- ▶ Table 4 (**Mauchly's Test of Sphericity**) is generated automatically and provides a very important check, being a measure of the variability of the dependent variables. It can be interpreted as testing whether the correlations between all the variables are the same. (It is equivalent to Levene's test for homoscedasticity (equality of variances) used with the *t* test.) Mauchly's test result determines which line of the next table to read significance values from.

Mauchly's Test of Sphericity^b

Measure:MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon ^a		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
marks	.097	342.793	5	.000	.435	.438	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

- ▶ Provided **Mauchly's Test** result is not significant (i.e. provided **Sig.** > 0.05) then sphericity can be assumed and the top line of the next table is used, otherwise a lower line is used.
- ▶ Table 4 (above) shows in this case that **Sig.** = 0.000 which is highly significant. This is because Module D has a very different standard deviation from the other three modules – its variation is much greater. So in this case sphericity cannot be assumed.
- ▶ We now come to the most important table ...

- ▶ Table 5 (**Tests of Within-Subjects Effects**) is generated automatically. It is the most important of all the many tables generated because it provides the evidence as to whether or not there is a significant difference in the results.
- ▶ We know from Table 4 (**Mauchly's Test**) that sphericity cannot be assumed, so we cannot use the top line of Table 5 (below). Instead we use the second line – the **Greenhouse-Geisser** line (we could choose **Huynh-Feldt** or **Lower-bound** but **Greenhouse-Geisser** is the most popular test).
- ▶ The **Greenhouse-Geisser** line has **Sig.** = 0.000 so it is highly significant ($p < 0.0005$). We conclude that there is definitely a within-subjects effect – i.e. there is a significant difference between the module marks. However, it does not indicate where the main differences lie – that comes later in Tables 8 and 9.

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
marks	Sphericity Assumed	13649.307	3	4549.769	82.435	.000
	Greenhouse-Geisser	13649.307	1.306	10450.787	82.435	.000
	Huynh-Feldt	13649.307	1.313	10395.883	82.435	.000
	Lower-bound	13649.307	1.000	13649.307	82.435	.000
Error(marks)	Sphericity Assumed	24505.443	444	55.192		
	Greenhouse-Geisser	24505.443	193.296	126.777		
	Huynh-Feldt	24505.443	194.317	126.111		
	Lower-bound	24505.443	148.000	165.577		

- ▶ You may have noticed that in Table 5 (above) all four lines have **Sig.** = 0.000, and may therefore wonder what all the fuss was about! Well that's real statistics – being cautious in coming to conclusions.
- ▶ Table 6 (**Tests of Within-Subjects Contrasts**) is generated automatically. It is of little value here – it reports that a straight line, or a quadratic, or a cubic, could all be fitted to illustrate the trend of the means – in each case **Sig.** = 0.000 so each is highly significant.

Tests of Within-Subjects Contrasts

Measure:MEASURE_1

Source	marks	Type III Sum of Squares	df	Mean Square	F	Sig.
marks	Linear	7359.224	1	7359.224	84.700	.000
	Quadratic	4946.458	1	4946.458	90.797	.000
	Cubic	1343.624	1	1343.624	55.490	.000
Error(marks)	Linear	12859.026	148	86.885		
	Quadratic	8062.792	148	54.478		
	Cubic	3583.626	148	24.214		

- Table 7 (**Tests of Between-Subjects Effects**) is generated automatically. It is of no use to us here. All it tells us is that the intercept (where the trend line crosses the y axis) is significantly different from zero (i.e. the overall mean mark is not zero). It is measuring whether all the subjects performed the same (treating the participants as a factor). We are not interested in this, only the within-subjects effects.

Tests of Between-Subjects Effects

Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	2215154.740	1	2215154.740	14354.200	.000
Error	22839.510	148	154.321		

- Tables 8 and 9 are optional. They are useful for revealing where the main differences lie.
- Table 8 (**Estimates**) provides data on the four modules' marks. The last line (corresponding to Module D) is very different from the others:
- Its **Mean** is much lower, showing that the marks are mostly lower.
 - Its **Std. Error** is much larger, showing that it has a lot more variability.
 - Its **95% Confidence Interval** is much lower (and wider), which is a consequence of the above two.

Estimates

Measure: MEASURE_1

marks	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	63.470	.504	62.473	64.467
2	63.403	.523	62.369	64.436
3	64.289	.446	63.408	65.169
4	52.698	1.192	50.343	55.053

- Table 9 (**Pairwise Comparisons**) shows where the significant differences lie. First we note that in this table:

1 = Module A
 2 = Module B
 3 = Module C
 4 = Module D

An asterisk next to the Mean Difference signifies that:

1 and 4 differ significantly
 2 and 4 differ significantly
 3 and 4 differ significantly.

Pairwise Comparisons

Measure: MEASURE_1

(I) marks	(J) marks	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1	2	.067	.370	1.000	-.923	1.057
	3	-.819	.362	.151	-1.787	.149
	4	10.772*	1.149	.000	7.698	13.845
2	1	-.067	.370	1.000	-1.057	.923
	3	-.886	.413	.201	-1.990	.218
	4	10.705*	1.131	.000	7.679	13.730
3	1	.819	.362	.151	-.149	1.787
	2	.886	.413	.201	-.218	1.990
	4	11.591*	1.186	.000	8.420	14.761
4	1	-10.772*	1.149	.000	-13.845	-7.698
	2	-10.705*	1.131	.000	-13.730	-7.679
	3	-11.591*	1.186	.000	-14.761	-8.420

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

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

- Of course, since 1 differs from 4 then 4 differs from 1, and so on, which is why the results appear twice, but they are only highlighted once above.
- Note that for the significant rows (highlighted) the Confidence Interval does not include zero as a possibility.
- Note that for the non-significant rows the Confidence Interval does include zero as a possibility.
- Table 10 is a repeat of Table 3 (**Multivariate Tests**). We can ignore this (again!).

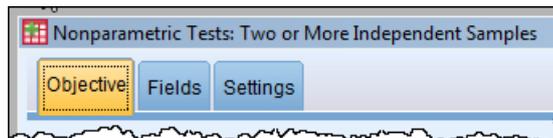
T29.6 One-Way within-subjects ANOVA – Friedman nonparametric test

If the criteria for the normal (parametric) ANOVA are seriously violated then a nonparametric version should be used. For the One-Way within-subjects ANOVA SPSS supplies the **Friedman’s ANOVA** method. As a demonstration, we repeat the analysis just carried out in T29.5.

1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav (if not open).
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.

2. Select **Analyze** → **Nonparametric Tests** → **Related Samples**

▶ This window opens →

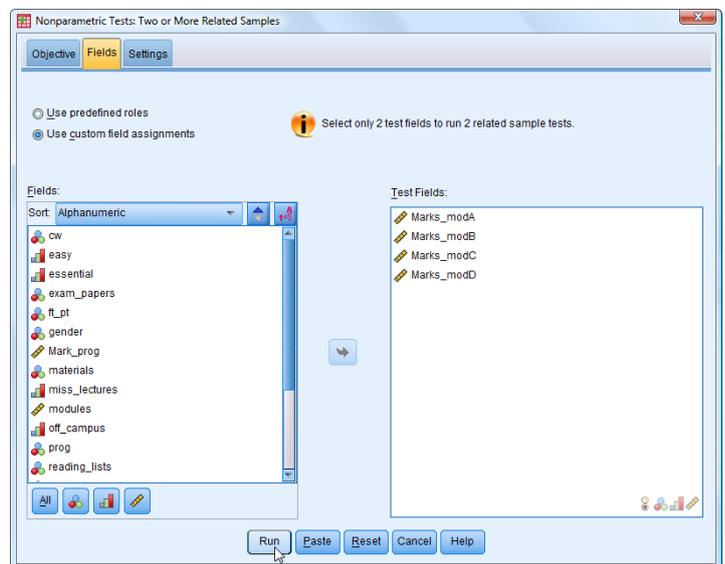


3. Select **Fields** to reveal the **Test Fields** box:

4. Move **Marks_modA, Marks_modB, Marks_modC, Marks_modD** into the **Test Fields** box →

5. Select **Run**.

- ▶ SPSS automatically determines which test to apply.
- ▶ Alternatively, before clicking on **Run** you could click on **Customize tests** and select **Friedman’s 2-way ANOVA by ranks (k samples)**.



6. The result shown below is to reject the null hypothesis (that the mean marks on the four modules are all the same). This is the same conclusion as was reached in T29.5 but with a lot less effort (but this is a test with less power – a concept not discussed in this Guide).

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Module A marks, Module B marks, Module C marks and Module D marks are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

- ▶ Note: This is really a one-way test but just to confuse you it is called a 2-way test because in a within-subjects ANOVA the participants can be considered to constitute a factor.

T29.7 Two-Way between-subjects ANOVA (independent measures)

Here we investigate whether school students' enjoyment of mathematics depends upon the student's gender, the teacher's gender, or an interaction between those two factors. The students were in Y11 – aged 15-16 years. Their level of 'enjoyment' was assessed by their responses to 12 multiple choice questions.

As there is one dependent variable (enjoyment), it is an ANOVA.

As there are two independent variables (student's gender and teacher's gender) it is a two-way ANOVA.

As all students answered the same enjoyment questions, it is a between-subjects two-way ANOVA.

1. Load data file: **File** → **Open** → **Data** → **DATA06_School Maths.sav**
2. Select **Edit** → **Options**
3. Click the **General** tab, if not already highlighted.
4. In the **Variable Lists** section select **Display names** and **Alphabetical**, to match the variable list format used in this tutorial and click **OK**.
5. Select **Analyze** → **General Linear Model** → **Univariate**

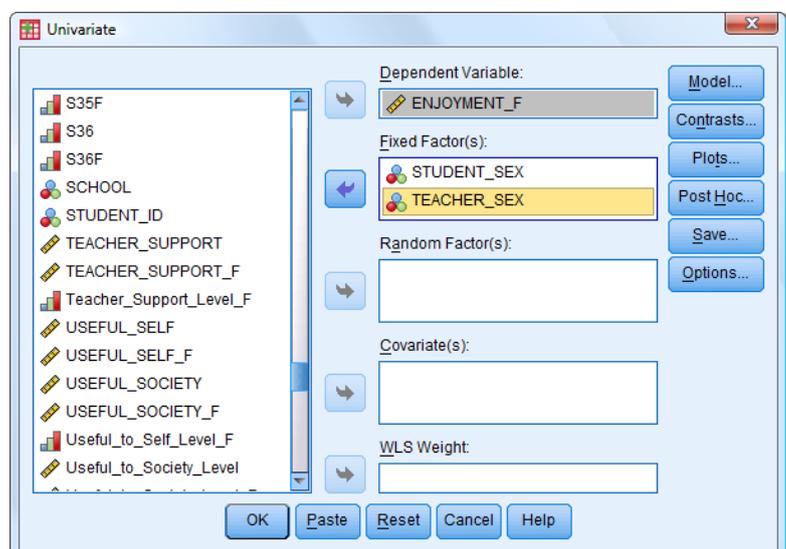
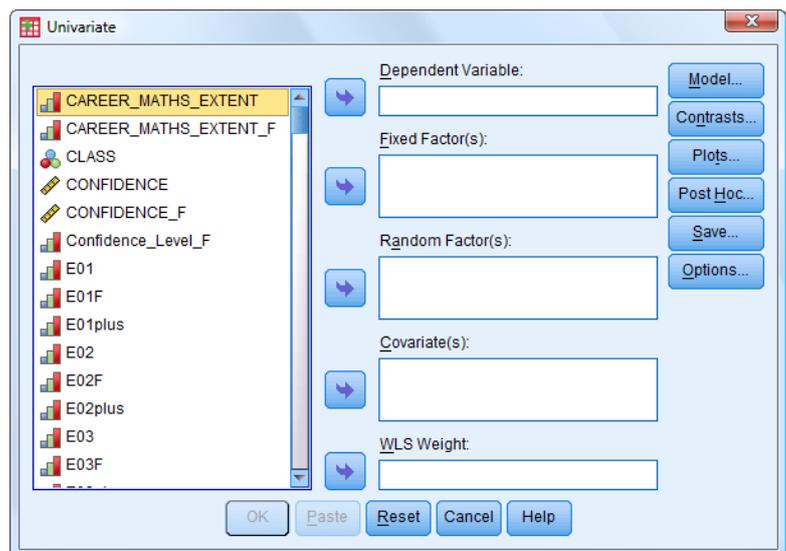
► The **Univariate** window will open →

► It is a good idea to widen the **Univariate** window so you can fully see the variable names.

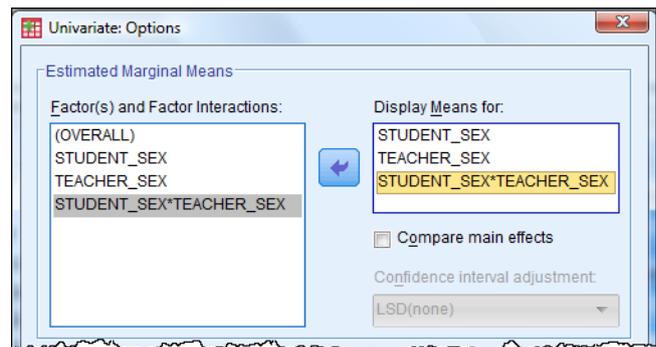
6. Scroll down the list of variables, which will be in alphabetical order, to locate **ENJOYMENT_F** and use the blue arrow to move it into the **Dependent Variable** box.
7. Scroll down the list further to locate **STUDENT_SEX** and move it into the **Fixed Factor(s)** box.
8. Move **TEACHER_SEX** into the **Fixed Factor(s)** box.

► The window will now appear like this →

9. Click on the **Options** button.



10. Move the three variables into the **Display Means for** box.
11. Click **Continue**.
12. Click **OK**
 - ▶ This produces five output tables.



- ▶ Table 1 (**Between-Subjects Factors**) simply reports on the value labels and counts for the two factors.

Between-Subjects Factors

		Value Label	N
Student's Gender	1	Female	111
	2	Male	121
Teacher's Gender	1	Female	66
	2	Male	166

- ▶ Table 2 (**Tests of Between-Subjects Effects**) is the most important. It signifies where any significant differences (effects) are found.

Tests of Between-Subjects Effects

Dependent Variable: Enjoyment of maths (0-40) [Final]

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	624.914 ^a	3	208.305	2.646	.050
Intercept	93980.523	1	93980.523	1193.701	.000
STUDENT_SEX	.287	1	.287	.004	.952
TEACHER_SEX	173.403	1	173.403	2.202	.139
STUDENT_SEX * TEACHER_SEX	359.788	1	359.788	4.570	.034
Error	17950.517	228	78.730		
Total	138832.000	232			
Corrected Total	18575.431	231			

- ▶ The STUDENT_SEX row shows that this is not a significant factor (i.e. there is not evidence that a females and males would have different level on enjoyment).
- ▶ The TEACHER_SEX row shows that this is not a significant factor (i.e. there is not evidence that the students with female teachers and male teachers would have different levels on enjoyment).
- ▶ The STUDENT_SEX *TEACHER_SEX row shows that this is a significant factor (i.e. there is evidence of an **interaction** between the two factors – although it does not say what it is. That can be discovered by examining a later table.
- ▶ The next three tables are optional and appear because we asked for them in steps 9 to 11.

- ▶ The **Student's Gender** table shows that the mean scores and variability of scores of **Female students** and **Male students** were very similar. This explains why the result was not significant.
- ▶ Table 3 (**Student's Gender**) shows that the mean scores and variability of scores of students in **Female teachers** classes and students in **Male teachers** classes were very similar. This explains why the result was not significant.
- ▶ Table 4 (**Teacher's Gender**) shows that the mean scores and variability of scores of students in **Female teachers** classes and students in **Male teachers** classes were very similar. This explains why the result was not significant.

1. Student's Gender

Dependent Variable:Enjoyment of maths (0-40) [Final]

Student's Gender	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Female	22.359	.930	20.527	24.191
Male	22.281	.897	20.513	24.049

2. Teacher's Gender

Dependent Variable:Enjoyment of maths (0-40) [Final]

Teacher's Gender	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Female	21.361	1.093	19.208	23.514
Male	23.279	.689	21.920	24.637

- ▶ Table 5 (**Student's Gender * Teacher's Gender**) is an interaction table which reveals where the differences lie.
 - For female teachers, the female students recorded higher enjoyment than the male students.
 - For male teachers, the male students recorded higher enjoyment than the females.
 - The interaction effect was more pronounced among the male students:
 - Males with female teachers recorded the lowest level (mean 19.9).
 - Males with male teachers recorded the highest level (24.6).

3. Student's Gender * Teacher's Gender

Dependent Variable:Enjoyment of maths (0-40) [Final]

Student's Gender	Teacher's Gender	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Female	Female	22.781	1.569	19.691	25.872
	Male	21.937	.998	19.970	23.904
Male	Female	19.941	1.522	16.943	22.940
	Male	24.621	.951	22.746	26.495

- ▶ A final word of caution: this was based on a small study involving only a few teachers. A much larger study replicating this finding would be needed to be able to claim it held true generally.

T29.8 Two-Way within-subjects ANOVA (repeated measures)

Here we investigate whether there are significant differences in school students' attitudes to the value of mathematics depending on two factors labeled TEST and TIME.

The students were in Y11 – aged 15-16 years. Their levels of perceived 'Value-to-Self' and 'Value-to-Society' were assessed by their responses to 22 (12 Self and 10 Society) multiple choice questions administered as part of two identical questionnaires administered at two different times, labeled 'Initial' and 'Final'.

- (1) TEST: This has two values: TEST1 = 'Self' and TEST2 = 'Society'.
- (2) TIME: This has two values: TIME1 = 'Initial' and TIME2 = 'Final'.

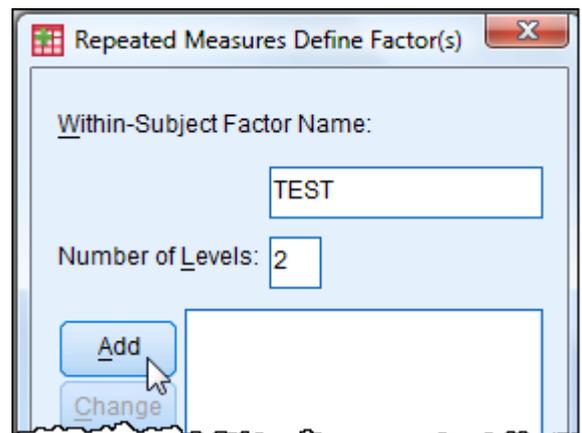
As there is one dependent variable (value) it is an ANOVA.

As there are two independent variables or 'factors' (TEST and TIME), it is a two-way ANOVA.

As the same students all completed two sets of questions (on 'Self' and 'Society') at two different times ('Initial' and 'Final') it is a within-subjects two-way ANOVA.

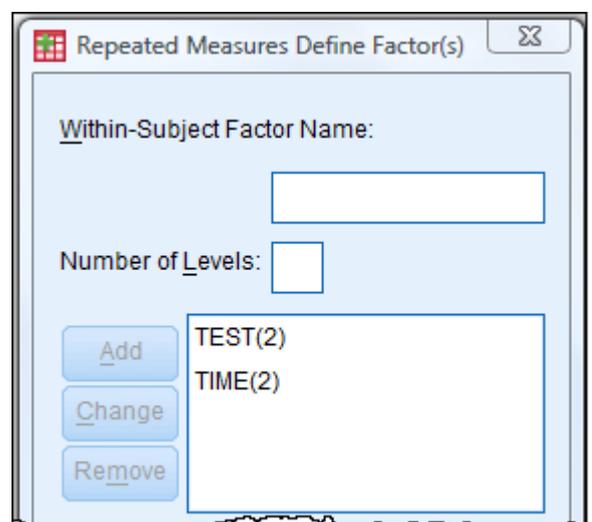
1. Load data file: **File** → **Open** → **Data** → DATA06_School_Maths.sav (if not open)
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **Alphabetical**, to match the list format used in this tutorial.
2. Select **Analyze** → **General Linear Model** → **Repeated Measures**

- ▶ The **Repeated Measures Define Factor(s)** window opens →



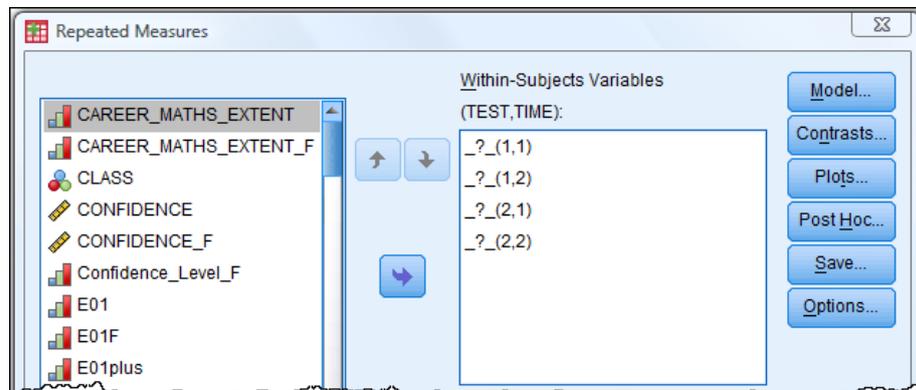
3. Change the **Within-Subjects Factor Name** from the default 'factor1' to 'TEST' →
4. Enter '2' in the **Number of Levels** box →
 - ▶ There were two tests taken – one on 'Value-to-Self' and one on 'Value-to-Society'.
5. Click **Add**.

6. Enter in the **Within-Subjects Factor Name** 'TIME'
7. Enter '2' in the **Number of Levels** box
 - ▶ There were two times that the tests was taken – 'Initial' and 'Final'.
8. Click **Add**.
 - ▶ The window now appears like this →



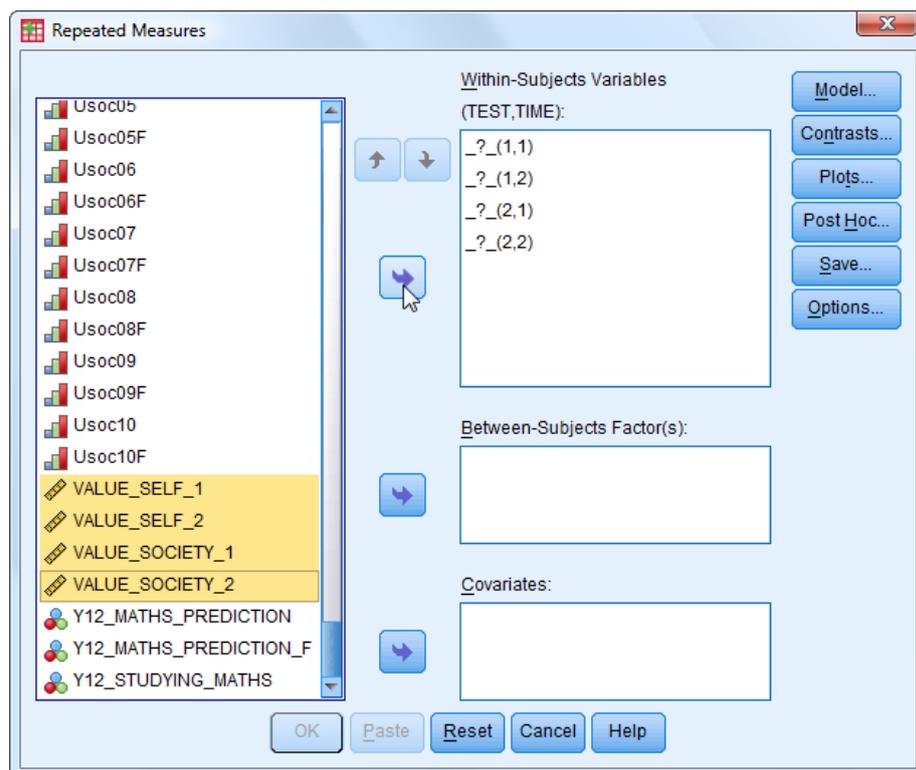
9. Click **Define**.

► The **Repeated Measures** window now appears like this:



► It is a good idea to widen the **Repeated Measures** window so you can fully see the variable names.

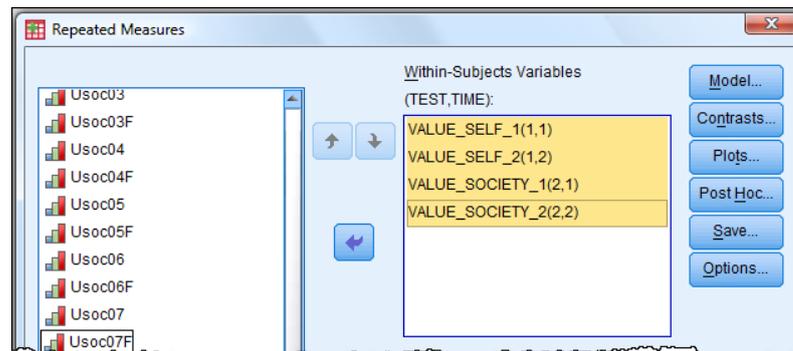
10. Scroll down the variables list (in alphabetical order) to locate and select **VALUE_SELF_1**, **VALUE_SELF_2**, **VALUE_SOCIETY_1**, **VALUE_SOCIETY_2**:



VALUE_SELF_1	contains the TEST=1, TIME=1 variable	[Self - Initial results]
VALUE_SELF_2	contains the TEST=1, TIME=2 variable	[Self - Final results]
VALUE_SOCIETY_1	contains the TEST=2, TIME=1 variable	[Society - Initial results]
VALUE_SOCIETY_2	contains the TEST=2, TIME=2 variable	[Society - Final results]

These four variables are all on a scale of 0 to 100.

11. Use the blue arrow to move all four variables into the **Within-Subjects Variables** window:



12. Click **Options**.

13. Select **Descriptive statistics**.

14. Click **Continue**.

15. Click **OK** to generate a series of seven tables.

- Table 1 (**Within-Subjects Factors**) is generated automatically. It just lists the four variables.

Within-Subjects Factors

Measure:MEASURE_1

TEST	TIME	Dependent Variable
1	1	VALUE_SELF_1
	2	VALUE_SELF_2
2	1	VALUE_SOCIETY_1
	2	VALUE_SOCIETY_2

- Table 2 (**Descriptive Statistics**) which is optional, displays the dependent variables' means, standard deviations and N (number of cases).
- We can see that all means look very similar but the 'Value to Society' standard deviations are somewhat lower than the 'Value to Self' standard deviations.

Descriptive Statistics

	Mean	Std. Deviation	N
Value to Self (0-100) [Initial]	68.74	17.259	208
Value to Self (0-100) [Final]	67.43	18.002	208
Value to Society (0-100) [Initial]	68.20	15.092	208
Value to Society (0-100) [Final]	69.64	15.836	208

- ▶ Table 3 (**Multivariate Tests** – not shown) is generated automatically. It is not relevant when the variables only take two values, as is the case here. (It is only of interest if sphericity is a problem – see Table 4.)
- ▶ Table 4 (**Mauchly's Test of Sphericity** – not shown) is generated automatically but is not relevant when the variables only take two values, as is the case here.
- ▶ Table 5 (**Tests of Within-Subjects Effects**) is generated automatically. It is the most important table because it provides the evidence as to whether or not there is a significant difference in the results.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TEST	Sphericity Assumed	144.722	1	144.722	1.259	.263
	Greenhouse-Geisser	144.722	1.000	144.722	1.259	.263
	Huynh-Feldt	144.722	1.000	144.722	1.259	.263
	Lower-bound	144.722	1.000	144.722	1.259	.263
Error(TEST)	Sphericity Assumed	23801.028	207	114.981		
	Greenhouse-Geisser	23801.028	207.000	114.981		
	Huynh-Feldt	23801.028	207.000	114.981		
	Lower-bound	23801.028	207.000	114.981		
TIME	Sphericity Assumed	1.011	1	1.011	.005	.944
	Greenhouse-Geisser	1.011	1.000	1.011	.005	.944
	Huynh-Feldt	1.011	1.000	1.011	.005	.944
	Lower-bound	1.011	1.000	1.011	.005	.944
Error(TIME)	Sphericity Assumed	42169.739	207	203.719		
	Greenhouse-Geisser	42169.739	207.000	203.719		
	Huynh-Feldt	42169.739	207.000	203.719		
	Lower-bound	42169.739	207.000	203.719		
TEST * TIME	Sphericity Assumed	391.876	1	391.876	8.035	.005
	Greenhouse-Geisser	391.876	1.000	391.876	8.035	.005
	Huynh-Feldt	391.876	1.000	391.876	8.035	.005
	Lower-bound	391.876	1.000	391.876	8.035	.005
Error(TEST*TIME)	Sphericity Assumed	10095.874	207	48.772		
	Greenhouse-Geisser	10095.874	207.000	48.772		
	Huynh-Feldt	10095.874	207.000	48.772		
	Lower-bound	10095.874	207.000	48.772		

- ▶ Neither TEST (**Sig.** = 0.263) nor TIME (**Sig.** = 0.944) has a significant effect, but the interaction TEST*TIME (**Sig.** = 0.005) does. By looking at the **Descriptive Statistics** output (Table 2 – shown earlier) we deduce that the explanation is that whereas 'Value-to-Self' declined from Initial test to Final test the opposite was true for 'Value-to-Society'.
- ▶ Two further tables are automatically produced but they are of no interest here.

TUTORIAL T30: Kolmogorov-Smirnov One-sample Test

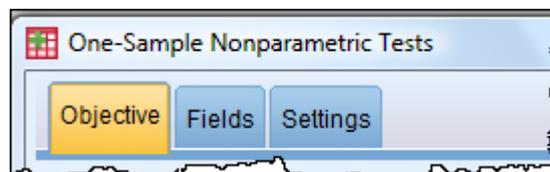
This test is designed to test whether a particular distribution differs significantly from a standard distribution – most commonly the normal distribution.

T30.1 Kolmogorov-Smirnov One-sample Test – Normality test: Example 1

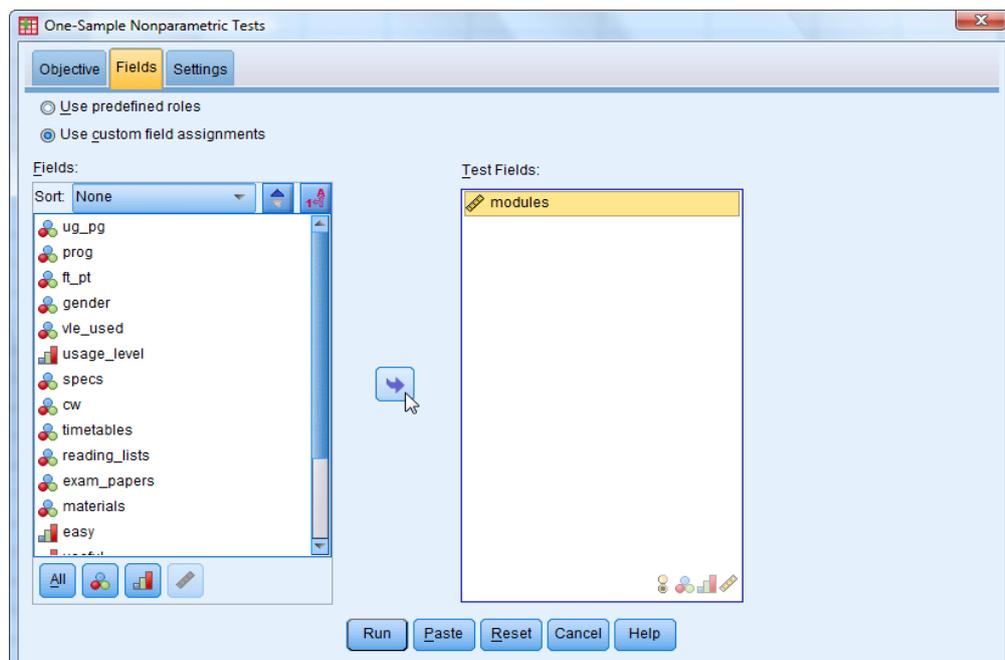
Here we test to see if the distribution of modules accessed on the VLE is normal. This is a variable with relatively few (13) values – usually this test would be applied to a distribution taking many values.

1. Load data file: **File** → **Open** → **Data** → DATA03_LSquestionnaire.sav (if not open)
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **File**, to match the list format used in this tutorial.
2. Select **Analyze** → **Nonparametric Tests** → **One Sample**

- ▶ This window opens →

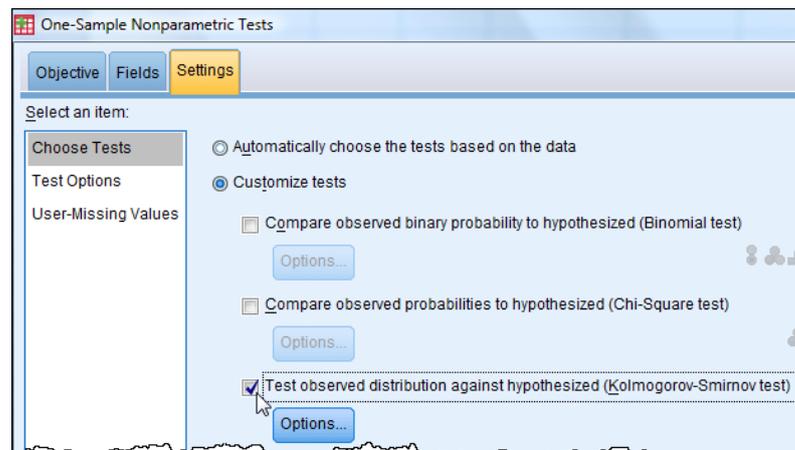


3. Click on the **Fields** button.
 - ▶ If there are any variables listed in the **Test Fields** box select them and move them out using the blue arrow.
4. Move the scale variable **modules** into the **Test Fields** box:



5. Click on the **Settings** button and highlight 'Choose Tests'.

- Click on the **Customize tests** radio button to select it.



- Select **Test observed distribution against hypothesized (Kolmogorov-Smirnov test)**
- Click the **Options** button.
- Select the 'Normal' box.
- Click **OK**.
- Click **Run**.

WARNING: IN SPSS 19 THIS DOES NOT SEEM TO WORK. HOWEVER, YOU CAN GET THE RESULT(S) YOU WANT BY OMITTING STEP 6 AND INSTEAD LET SPSS DECIDE WHICH TEST TO APPLY – i.e. SELECT 'Automatically choose the test based on the data'.

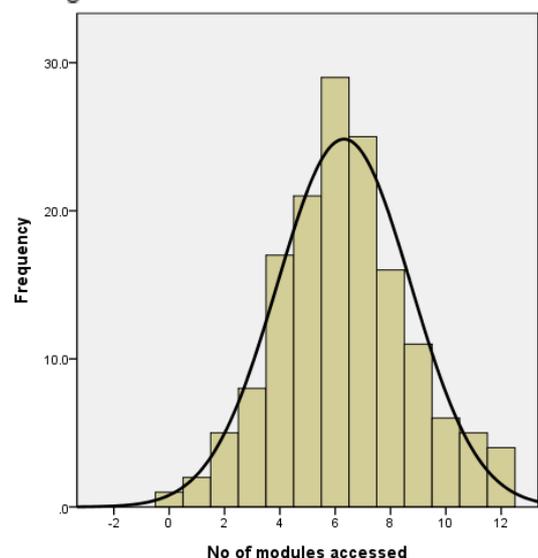
► This output appears:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of No of modules accessed is normal with mean 6.32 and standard deviation 2.409.	One-Sample Kolmogorov-Smirnov Test	.057	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

- The conclusion is to accept the Null Hypothesis which is that the distribution is normal, since $p > 0.05$, but it is marginal as $p = 0.057$.
- See the histogram here which demonstrates the normality →



T30.2 Kolmogorov-Smirnov One-sample Test – Normality test: Example 2

Here we test if the distributions of Average Selling Price (ASP) and Recommended Retail Price (RRP) of the Top-selling 100 books are normal. These variables can take many values.

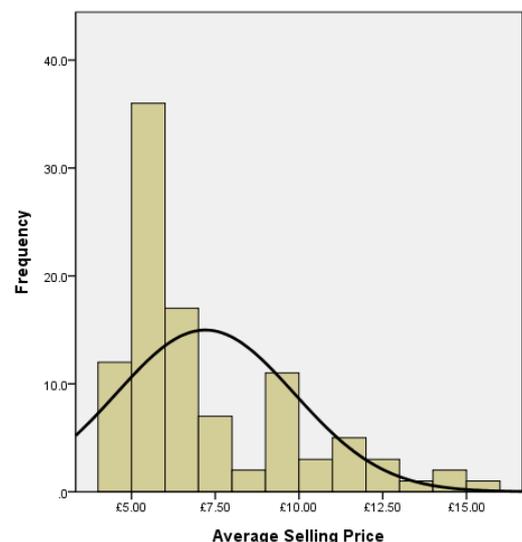
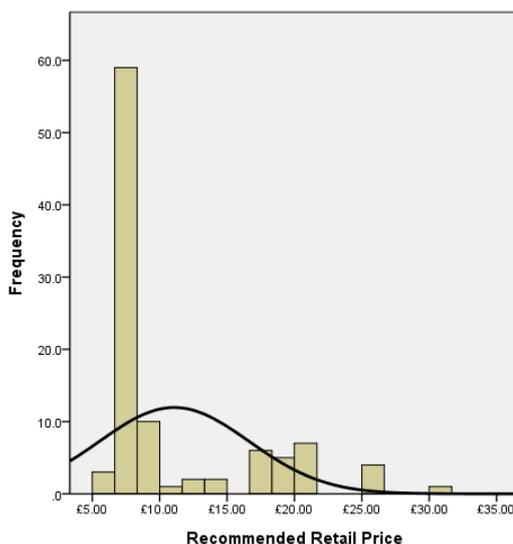
1. Load data file: **File** → **Open** → **Data** → DATA01_100Books.sav
2. Select **Analyze** → **Nonparametric Tests** → **One Sample**
3. Click on the **Fields** button.
4. If there are any variables in the **Test Fields** box remove them using the blue arrow.
5. Move the scale variables **ASP** [Average Selling Price] and **RRP** [Recommended Retail Price] into the **Test Fields** box.
6. Click on the **Settings** button.
7. Click on the **Customize tests** radio button to select it.
8. Select **Test observed distribution against hypothesized (Kolmogorov-Smirnov test)**
9. Click the **Options** button.
10. Select the 'Normal' box.
11. Click **OK**.
12. Click **Run**. **WARNING: IN SPSS 19 THIS DOES NOT SEEM TO WORK FOR THIS DATA FILE.**

► This output report appears, which clearly shows that the Null Hypothesis is rejected in both cases as neither distribution is at all close to normal (see charts beneath which confirm this).

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Recommended Retail Price is normal with mean 11.081 and standard deviation 5.564.	One-Sample Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.
2	The distribution of Average Selling Price is normal with mean 7.203 and standard deviation 2.662.	One-Sample Kolmogorov-Smirnov Test	.001	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

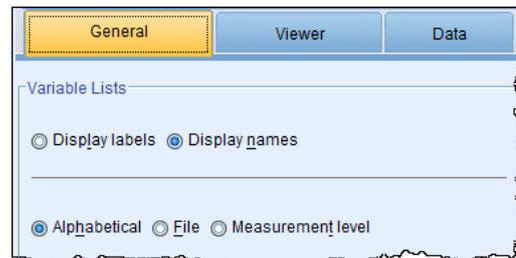


TUTORIAL T31: Linear Regression

In its simplest form, Linear Regression is used to predict values of one variable from values of another variable using a straight line (linear) equation.

Note: It must be remembered that having a regression equation does not mean that variation in the independent variable(s) causes the variation in the dependent variable. It just means there is an association (just as for correlation, to which regression is closely related).

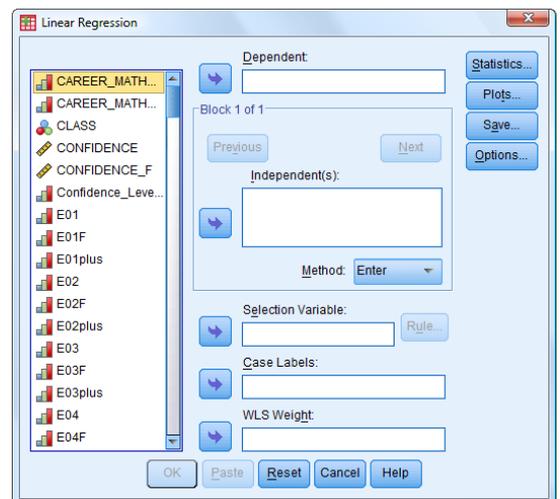
Note that with this dataset **Edit** → **Options** has been used to select the listing of variables to **Display names** in **Alphabetical** order. This makes it much easier to find variables when there are a lot of them (as is the case with this dataset). See Section 2.5.1 for more information on this.



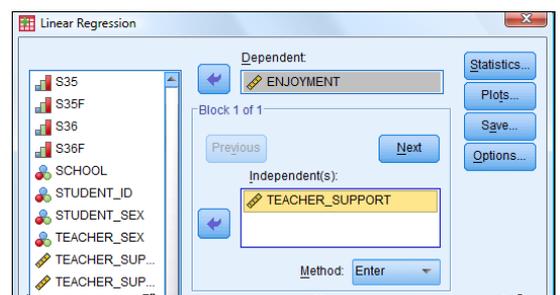
T31.1 Simple Linear Regression

Here we seek a regression equation (linear) which can be used to predict the value of a dependent variable given a value of an independent variable. This is the simplest regression model. More complex models are sometimes used to find a curvilinear equation (not a simple straight line).

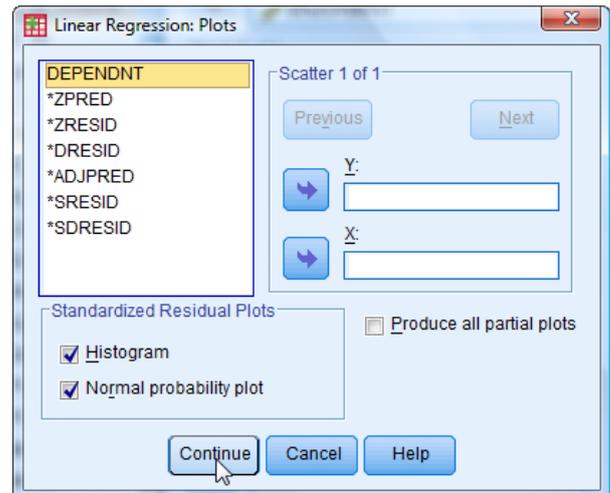
1. Load data file: **File** → **Open** → **Data** → **DATA06_School_Maths.sav**
2. Select **Edit** → **Options** and click on the **General** tab and in the **Variable Lists** section choose **Display names** and **Alphabetical**, for the most useful listing format here.
3. Click **Apply** (and **OK** to close the advisory message).
4. Click **OK**.
5. Select **Analyze** → **Regression** → **Linear**
 - ▶ The Linear Regression window opens →
 - ▶ Widen the **Linear Regression** window as necessary so you can see the whole of each variable name.
6. Scroll down and select **ENJOYMENT** and move it into the **Dependent** window.
7. Scroll down much further and select **TEACHER_SUPPORT** and move it to the **Independent(s)** window.



- ▶ The window appears like this →



8. Select **Plots** to open the **Linear Regression Plots** window and choose the following two options:



Histogram →
Normal probability plot →

9. Click **Continue**.
10. Click **OK** to produce the following five output tables and two charts.

► Table 1 (**Variables Entered/Removed**) states that the independent variable is **Teacher supportiveness** (on a scale 0–48).

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	Teacher supportiveness (0-48) [Initial] ^a	.	Enter

- a. All requested variables entered.
- b. Dependent Variable: Enjoyment of maths (0-40) [Initial]

► Table 2 (**Model Summary**) provides the Pearson correlation coefficient between the independent variable **Teacher supportiveness** and the dependent variable **Enjoyment of maths** (on a scale 0 to 40). In this case the correlation is 0.540 and so the **Adjusted R Square** is 0.289 which shows that about 30% of the variation in **Enjoyment** can be 'explained' by **Teacher supportiveness**, and the relationship is positive.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.540 ^a	.291	.289	6.698

- a. Predictors: (Constant), Teacher supportiveness (0-48) [Initial]
- b. Dependent Variable: Enjoyment of maths (0-40) [Initial]

► Table 3 (**ANOVA**) reports the ANOVA result showing the significance of the regression model. Here the **Sig.** associated with the F test is 0.000 (i.e. $p < 0.0005$) which is highly significant, confirming that the independent variable does explain a significant amount of the variation in the dependent variable.

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4797.093	1	4797.093	106.916	.000 ^a
	Residual	11665.675	260	44.868		
	Total	16462.767	261			

- a. Predictors: (Constant), Teacher supportiveness (0-48) [Initial]
- b. Dependent Variable: Enjoyment of maths (0-40) [Initial]

- ▶ The conclusion one might draw is that if a teacher is supportive then the student is more likely to enjoy mathematics. However, this is a only measure of a student's perception of teacher supportiveness, so it might be argued that if a student enjoys mathematics then they think the teacher is supportive. Cause and effect are not easily determined!
- ▶ In Table 3 (**ANOVA**) the **Mean Square** column (produced by dividing the **Sum of Squares** by the **df**) gives the variance. Here very much more of the variance is explained by the Regression line (4797.093) than by the Residual (44.868) which can be considered as unaccounted for 'error'. This reinforces the conclusion that the model is good.
- ▶ Table 4 (**Coefficients**) presents the coefficients for the regression equation.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.635	1.950		1.864	.063
	Teacher supportiveness (0-48) [Initial]	.610	.059	.540	10.340	.000

a. Dependent Variable: Enjoyment of maths (0-40) [Initial]

Here the regression equation is

$$\text{Enjoyment} = 3.635 + 0.610 \times \text{Teacher supportiveness}$$

which can be used to predict the **Enjoyment** level (0 to 40) for any given **Teacher supportiveness** level (0 to 48).

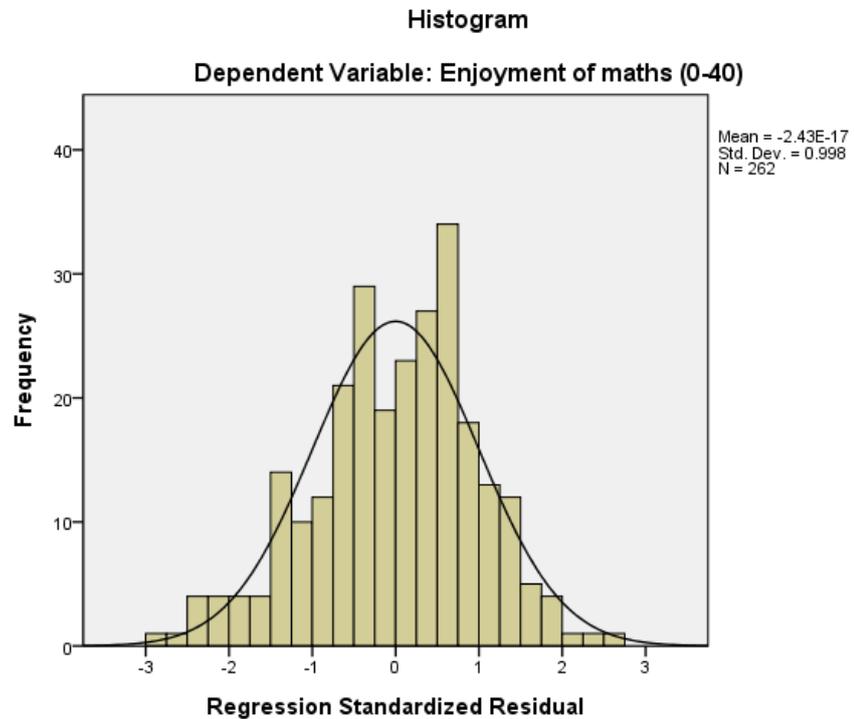
- ▶ The **Standardized Coefficient Beta** tells us the contribution the variable makes to the model. In this case there is just one variable and its contribution is 0.540 (54%), which is the Pearson correlation shown in an earlier table.
- ▶ The **t** value of 1.864 and associated **Sig.** of 0.063 (i.e. $p = 0.063$) for the **constant** is just above $p = 0.05$ so one cannot rule out the possibility that the true value of the **constant** in the equation is zero, although that is unlikely.
- ▶ The **t** value of 10.340 and associated **Sig.** of 0.000 (i.e. $p < 0.0005$) for the **Teacher supportiveness** independent variable shows that the regression is statistically significant.
- ▶ Table 5 (**Residuals Statistics**) is shown here but will not be discussed. The two plots which follow provide a more visual approach to examining residuals (what the model does not 'explain').

Residuals Statistics^a

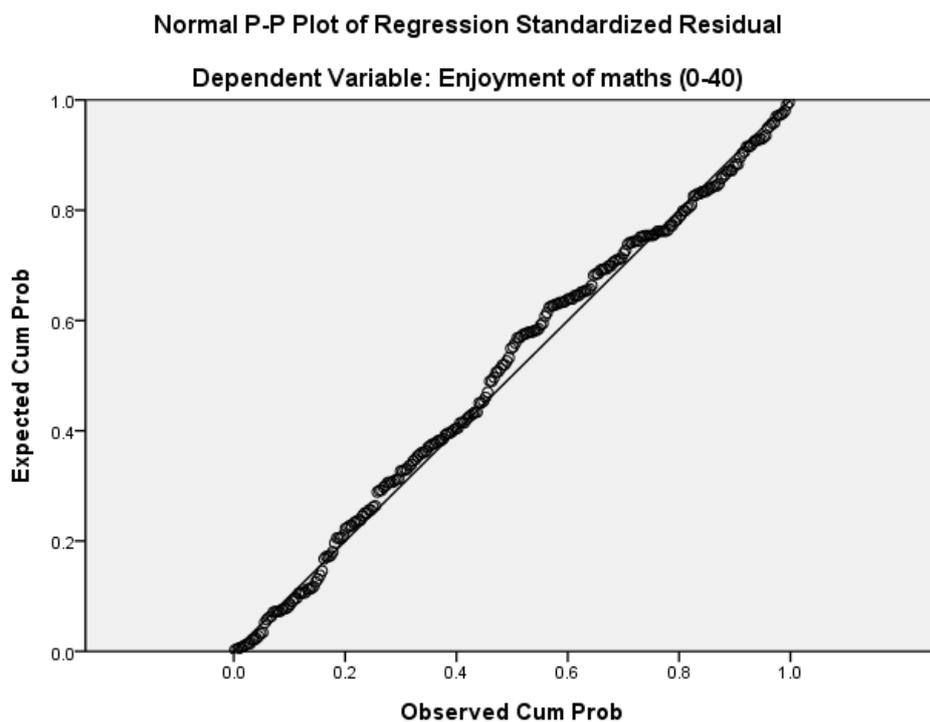
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	7.30	32.33	23.34	4.287	262
Residual	-18.559	17.598	.000	6.686	262
Std. Predicted Value	-3.742	2.096	.000	1.000	262
Std. Residual	-2.771	2.627	.000	.998	262

a. Dependent Variable: Enjoyment of maths (0-40) [Initial]

- ▶ The optional Histogram plot of the Regression Standardized Residuals, with normal curve superimposed, shows a good fit confirming that the distribution of the residuals is normal which is a condition for the validity of the linear model.



- ▶ The optional Normal P-P Plot of the Regression Standardized Residuals shows a very good fit between the expected cumulative probability and the observed cumulative probability, confirming that the distribution of the residuals (i.e. the variations from the predicted line) can be considered normal, which is a condition for the validity of the linear model.



T31.2 Multiple Linear Regression – using Entry method ‘Enter’

Here we seek a regression equation (linear) which can be used to predict the value of a dependent variable given the values of several independent variables.

1. Load data file: **File** → **Open** → **Data** → DATA06_School_Maths.sav (if not already loaded).
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **Alphabetical**, to match the list format in this tutorial.

2. Select **Analyze** → **Regression** → **Linear**

3. Click **Reset**.

4. Select **ENJOYMENT** and move it into the **Dependent** window →

5. Move **TEACHER_SUPPORT** to the **Independent(s)** window →

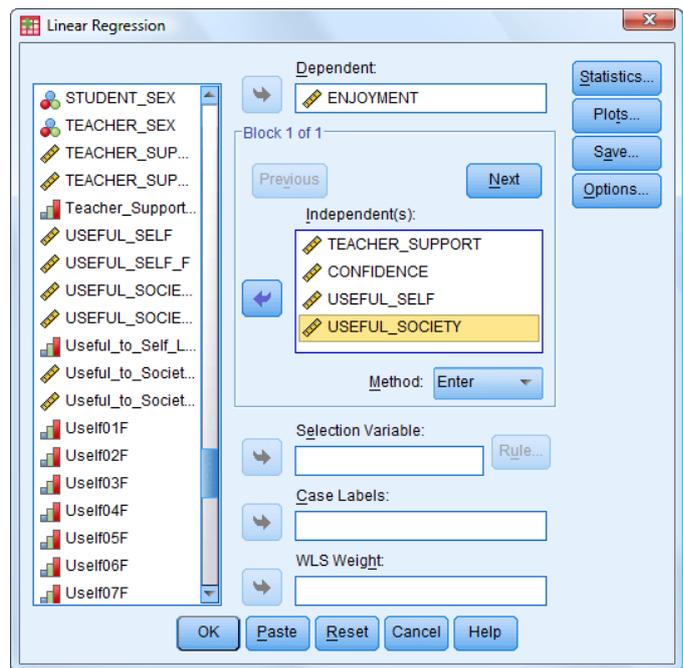
6. Move **CONFIDENCE** to the **Independent(s)** window →

7. Move **USEFUL_SELF** to the **Independent(s)** window →

8. Move **USEFUL_SOCIETY** to the **Independent(s)** window →

▶ The window appears like this →

▶ Note that the **Method** drop-down menu has **Enter** as the selected option.



9. Click **OK** to produce the following four output tables.

▶ Table 1 (**Variables Entered/Removed**) states the four independent variables and the entry method chosen ('Enter'). This entry method choice means that all four independent variables will be used in the model, even if their contribution is very small.

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	Usefulness of maths to society (0-40) [Initial], Confidence in maths (0-48) [Initial], Teacher supportiveness (0-48) [Initial], Usefulness of maths to self (0-48) [Initial] ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: Enjoyment of maths (0-40) [Initial]

- Table 2 (**Model Summary**) provides the overall Pearson correlation coefficient between the independent variables and the dependent variable. In this case the multiple correlation is 0.822 and so the **Adjusted R Square** is 0.671 which shows that about 67% of the variation in **Enjoyment** can be 'explained' by the model comprised of the four independent variables.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.822 ^a	.676	.671	4.569

a. Predictors: (Constant), Usefulness of maths to society (0-40) [Initial], Confidence in maths (0-48) [Initial], Teacher supportiveness (0-48) [Initial], Usefulness of maths to self (0-48) [Initial]

b. Dependent Variable: Enjoyment of maths (0-40) [Initial]

- Table 3 (**ANOVA**) reports the significance of the regression model. Here the **Sig.** associated with the F test is 0.000 (i.e. $p < 0.0005$) which is highly significant, which confirms that the model does explain a significant amount of the variation in the dependent variable.

- In Table 3 the **Mean Square** column shows that very much more of the variance is explained by the Regression line than by the Residual (2767.213 compared to 20.873). This reinforces the conclusion that the model is good.

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11068.852	4	2767.213	132.572	.000 ^a
	Residual	5301.820	254	20.873		
	Total	16370.672	258			

a. Predictors: (Constant), Usefulness of maths to society (0-40) [Initial], Confidence in maths (0-48) [Initial], Teacher supportiveness (0-48) [Initial], Usefulness of maths to self (0-48) [Initial]

b. Dependent Variable: Enjoyment of maths (0-40) [Initial]

- Table 4 (**Coefficients**) presents the coefficients for the regression equation, which is:

$$\begin{aligned} \text{Enjoyment} = & -9.695 + 0.450 \times \text{Confidence in maths} \\ & + 0.417 \times \text{Usefulness of maths to society} \\ & + 0.105 \times \text{Usefulness of maths to self} \\ & + 0.104 \times \text{Teacher supportiveness} \end{aligned}$$

which can be used to predict the **Enjoyment** level for any given levels of the four variables.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-9.695	1.573		-6.163	.000
	Teacher supportiveness (0-48) [Initial]	.104	.050	.092	2.096	.037
	Confidence in maths (0-48) [Initial]	.450	.041	.485	10.927	.000
	Usefulness of maths to self (0-48) [Initial]	.105	.053	.109	1.966	.050
	Usefulness of maths to society (0-40) [Initial]	.417	.073	.310	5.690	.000

- The above regression equation could be simplified, with little loss of accuracy, to:

$$\begin{aligned} \text{Enjoyment} = & -10 + 0.45 \times \text{Confidence in maths} \\ & + 0.42 \times \text{Usefulness of maths to society} \\ & + 0.1 \times \text{Usefulness of maths to self} \\ & + 0.1 \times \text{Teacher supportiveness} \end{aligned}$$

- The **Standardized Coefficient Beta** tells us the contribution each variables makes to the model (measured in standard deviation units of the target variable). In this case **Confidence** is the most important: a change of 1 SD in **Confidence** would lead to a change of 0.45 SD in **Enjoyment**.
- The **t** value of 1.864 and associated **Sig.** of 0.063 (i.e. $p = 0.063$) for the **constant** is just above $p = 0.05$ so one cannot rule out the possibility that the true value of the **constant** in the equation is zero, although that is unlikely.
- The **t** value of 10.340 and associated **Sig.** of 0.000 (i.e. $p < 0.0005$) for the **Teacher supportiveness** independent variable shows that the regression is statistically significant.

T31.3 Multiple Linear Regression – using Entry method ‘Stepwise’

In the previous section we found that the regression equation using the four independent variables was approximately

$$\begin{aligned} \text{Enjoyment} = & -10 + 0.45 \times \text{Confidence in maths} \\ & + 0.42 \times \text{Usefulness of maths to society} \\ & + 0.1 \times \text{Usefulness of maths to self} \\ & + 0.1 \times \text{Teacher supportiveness} \end{aligned}$$

This raises the question of whether it is sensible or worthwhile including the last two variables whose contribution is quite small. By changing the variable entry method *SPSS* will take care of this by only including variables which contribute significantly. This is controlled by the **Method** drop-down menu, which has five options:

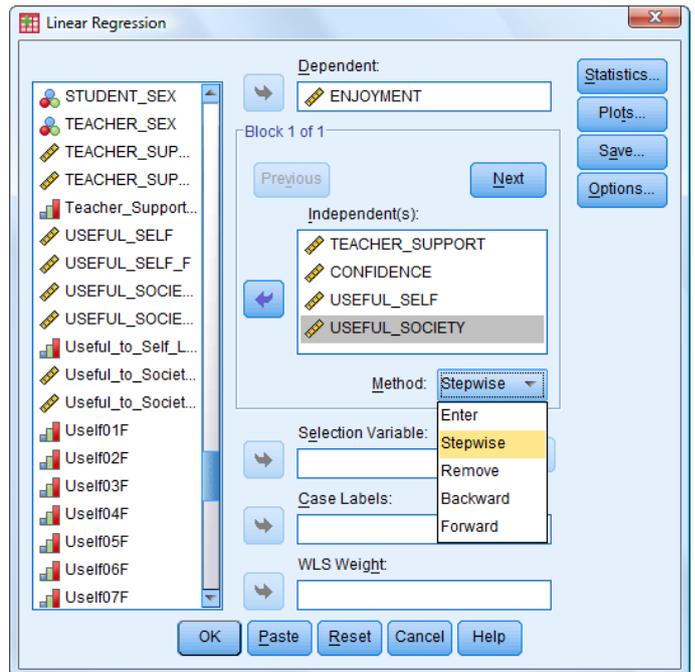
- Enter Includes all selected variables (as used in T31.2).
- Forward Enters variables one at a time (in order of importance) until no significant improvement occurs.
- Backward Enters all variables at once then removes one at a time until no significant improvement occurs.
- Stepwise A combination of Forward and Backward methods designed to ensure the final selection of variables is the best possible. [A common choice.]
- Remove Following Enter method it removes any variables the user chooses.

Here we repeat T31.2 using the **Stepwise** method of entry (which gives the same result as using the **Forward** method in this case – for the reader to verify).

If following on from T31.2, then skip straight to step 8.

1. Load data file: **File** → **Open** → **Data** → **DATA06_School_Maths.sav**
(if not already done)
2. Select **Analyze** → **Regression** → **Linear**
 - ▶ If following straight on from T31.2, skip to step 8.
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **Alphabetical**, to match the list format in this tutorial.

3. Select **ENJOYMENT** and move it into the **Dependent** window.
4. Move **TEACHER_SUPPORT** to the **Independent(s)** window.
5. Move **CONFIDENCE** to the **Independent(s)** window.
6. Move **USEFUL_SELF** to the **Independent(s)** window.
7. Move **USEFUL_SOCIETY** to the **Independent(s)** window.
8. Open **Method** and select **Stepwise**.
 - ▶ The window appears like this →
9. Click **OK** to produce the output tables.



- ▶ As before, several tables are produced (we only show two of them here). This time they are bigger than in 31.2 as each table shows all the outputs for each of the successive models which are created as each variable is added (or removed). The models are numbered 1, 2, ... until the process is completed.
- ▶ Below we show the **Model Summary** table, which has three models containing 1, 2 and 3 variables (as listed in the footnotes). The R shows that the multiple correlation coefficient increases as more variables are added.

Model Summary^d

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.730 ^a	.533	.531	5.455
2	.815 ^b	.664	.662	4.633
3	.819 ^c	.671	.667	4.594

a. Predictors: (Constant), Confidence in maths (0-48) [Initial]

b. Predictors: (Constant), Confidence in maths (0-48) [Initial], Usefulness of maths to society (0-40) [Initial]

c. Predictors: (Constant), Confidence in maths (0-48) [Initial], Usefulness of maths to society (0-40) [Initial], Teacher supportiveness (0-48) [Initial]

d. Dependent Variable: Enjoyment of maths (0-40) [Initial]

- Below we show the Coefficients table. From this the actual models (regression equations) can be derived.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.958	1.352		.708	.480
	Confidence in maths (0-48) [Initial]	.678	.040	.730	17.123	.000
2	(Constant)	-8.227	1.470		-5.598	.000
	Confidence in maths (0-48) [Initial]	.502	.038	.541	13.237	.000
	Usefulness of maths to society (0-40) [Initial]	.551	.055	.409	10.015	.000
3	(Constant)	-9.642	1.582		-6.095	.000
	Confidence in maths (0-48) [Initial]	.469	.040	.505	11.660	.000
	Usefulness of maths to society (0-40) [Initial]	.507	.058	.376	8.758	.000
	Teacher supportiveness (0-48) [Initial]	.115	.050	.101	2.302	.022

a. Dependent Variable: Enjoyment of maths (0-40) [Initial]

The first model has just one variable (the most influential) – **Confidence** – leading to this regression equation:

$$\text{MODEL 1: } \text{Enjoyment} = 0.958 + 0.678 \times \text{Confidence in maths}$$

The second model has just another variable added – **Usefulness to society** – leading to this regression equation:

$$\text{MODEL 2: } \text{Enjoyment} = -8.227 + 0.502 \times \text{Confidence in maths} \\ + 0.551 \times \text{Usefulness to society}$$

The third model has just another variable added – **Teacher supportiveness** – leading to this regression equation:

$$\text{MODEL 3: } \text{Enjoyment} = -9.642 + 0.469 \times \text{Confidence in maths} \\ + 0.507 \times \text{Usefulness to society} \\ + 0.507 \times \text{Teacher supportiveness}$$

There is no fourth model as the variable **Usefulness to self** does not make a sufficient contribution and so is not entered (from 31.2 we know that it would produce a multiple correlation of 0.822).

- We do not show the other tables here. They are interpreted much as they were in T31.2.
- It may seem strange that in the four variable model derived in T31.2 the most influential variable is **Confidence** but in the three variable model it is **Usefulness to society**. The explanation is that when **Usefulness to self** is excluded its contribution is mostly taken up by **Usefulness to society**. This is because **Usefulness to self** correlates much more highly with **Usefulness to self** (+0.739) than it does with the other two independent variables (+0.476 and +0.487).

TUTORIAL T32: Logistic Regression

Logistic Regression is a variant of multiple regression when the dependent variable (to be predicted) is dichotomous (i.e. taking only two values) – e.g. Pass/Fail, Include/Exclude, High grade / Low grade.

T32.1 Logistic Regression – using Entry method ‘Forward LR’

Here we develop a regression model to predict the value of a variable which can take one of just two values. In this example the dataset comes from a research project (2008-9) investigating the factors which determine whether or not a student will continue studying mathematics into Y12.

For details of the project and the derivation of the variables see the Appendix.

Briefly, there are five variables we will use – derived from a Final Questionnaire taken in May 2009 by Y11 students, before public examination results were known – to be used to predict whether or not an individual student will continue to study mathematics in Y12 (starting Autumn 2009). (The dataset has several other variables which might be used.)

The ‘F’ at the end of the variable names indicates that this is the result from the Final questionnaire rather than the Initial questionnaire which was completed at the beginning of the academic year (September 2008).

Dependent variable (binary): **Y12_STUDYING_MATHS** (this records what actually happened)

Independent variables:	STUDENT_SEX	Categorical:	Female (1) or Male (2)
	CONFIDENCE_F	Scale:	0 to 48
	ENJOYMENT_F	Scale:	0 to 40
	TEACHER_SUPPORT_F	Scale:	0 to 48
	USEFUL_SELF_F	Scale:	0 to 48
	USEFUL_SOCIETY_F	Scale:	0 to 40

1. Load data file: **File** → **Open** → **Data** → **DATA06_School_Maths.sav**

► Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **Alphabetical**, to match the list format in this tutorial.

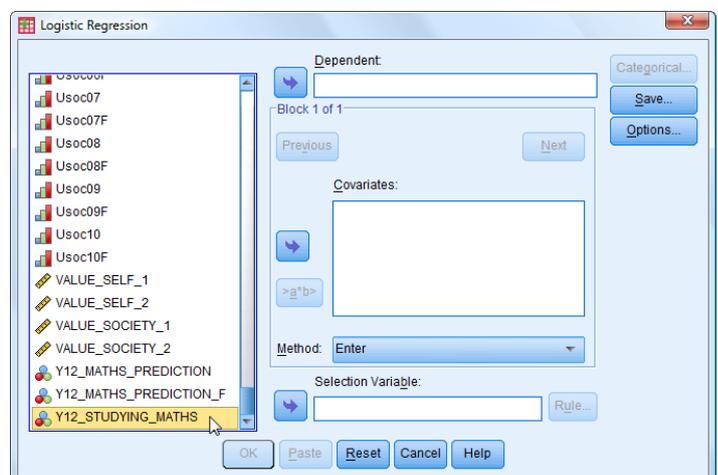
2. Select **Analyze** → **Regression** → **Binary Logistic**

► The **Linear Regression** window opens →

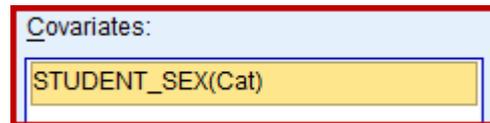
3. Scroll down to the bottom of the variable list and select **Y12_STUDYING_MATHS** and move it to the **Dependent** window.

4. Move **STUDENT_SEX** to the **Covariates** window.

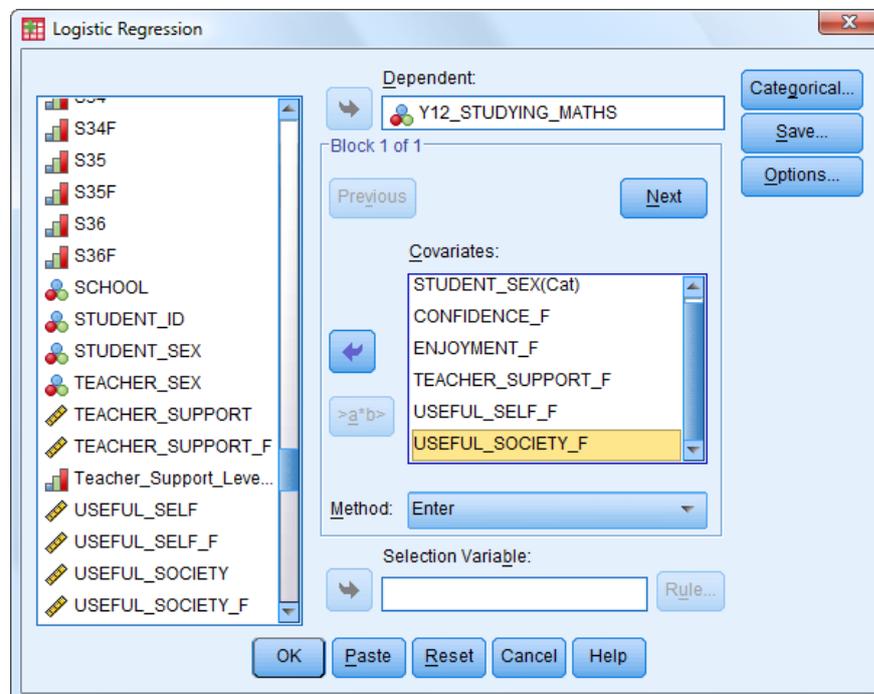
► Unlike all the other independent variables, this is nominal (categorical) and this fact needs to be declared.



5. Click on the **Categorical** button and move the **STUDENT_SEX** variable into the **Categorical Covariates** box which appears.
6. Click **Continue**, which will display this →
7. Move **CONFIDENCE_F** to the **Covariates** window.
8. Move **ENJOYMENT_F** to the **Covariates** window.
9. Move **TEACHER_SUPPORT_F** to the **Covariates** window.
10. Move **USEFUL_SELF_F** to the **Covariates** window.
11. Move **USEFUL_SOCIETY_F** to the **Covariates** window.



► The window appears like this:



12. Change the **Method** from the default **Enter** to **Forward: LR**

- This method chooses the most influential independent variable and adds it to the model, and continues adding a variable until no significant improvement is achieved.
- In contrast, the default method adds in all the selected variables *en bloc* (shown in T32.2).

13. **Click OK** which generates a large number of output tables, as explained below:

- With all *SPSS* analyses there is the option to select just a subset of cases, and therefore exclude others – e.g. select only females. (This process is explained in Part 1 of this Guide in Section 8.2.) However, in this example there are no unselected cases (the default), so all possible cases are included.
- This procedure produces a plethora of tables, most of which can be skimmed over by the beginner. They are all introduced here, but by far the most important are Tables 8 and 9.

- ▶ Table 1 (**Case Processing Summary**) reports that all 282 cases have been selected and that of these 200 are included in the analysis and 82 are missing (due to missing data), and that there are no unselected cases.

Case Processing Summary

Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	200	70.9
	Missing Cases	82	29.1
	Total	282	100.0
Unselected Cases		0	.0
Total		282	100.0

a. If weight is in effect, see classification table for the total number of cases.

- ▶ Table 2 (**Dependent Variable Encoding**) reports that the dependent variable – **Y12_STUDYING_MATHS** – has two values with names 'No' and 'Yes' – with codes '0' and '1'. If different codes are used in the dataset, *SPSS* automatically reassigns them to '0' and '1' for the analysis.

Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

- ▶ Table 3 (**Categorical Variables Coding**) reports the one categorical variable (Student's Gender) and the frequencies of the associated values.

Categorical Variables Codings

		Frequency	Parameter coding
			(1)
Student's Gender	Female	99	1.000
	Male	101	.000

- Table 4 (**Classification Table** – for Block 0: Beginning Block) reports the first stage of the modeling process. The model’s purpose to predict the ‘0’ and ‘1’ correctly for as many cases as possible. The first stage (‘Block 0’ here) just uses a constant predictor. As there are 118 ‘No’ and 82 ‘Yes’ cases, taking the cut value as half (the default), SPSS calculates that ‘No’ has more cases than ‘Yes’ and so assigns ‘No’ as the constant prediction for all 200 cases. This ensures that the percentage correct prediction is over 50%.

Block 0: Beginning Block

Classification Table^{a,b}

Observed			Predicted		
			Studying Maths in Y12		Percentage Correct
			No	Yes	
Step 0	Studying Maths in Y12	No	118	0	100.0
		Yes	82	0	.0
	Overall Percentage				59.0

- a. Constant is included in the model.
- b. The cut value is .500

- Table 4 (above) reports that the overall percentage is 59% because all 118 ‘No’ cases are predicted correctly and none of the 82 ‘Yes’ cases are predicted correctly. Not a very sophisticated predictive model, of course.
- Tables 5 and 6 (**Variables ...**) report the initial stage of the modeling process, when there is just a constant in the model and none of the proposed variables in the model.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	-.364	.144	6.409	1	.011	.695

Variables not in the Equation

	Score	df	Sig.
Step 0 Variables			
STUDENT_SEX	1.066	1	.302
CONFIDENCE_F	55.364	1	.000
TEACHER_SUPPORT_F	49.896	1	.000
USEFUL_SELF_F	45.305	1	.000
ENJOYMENT_F	74.915	1	.000
USEFUL_SOCIETY_F	26.213	1	.000
Overall Statistics	84.463	6	.000

- Table 7 (**Omnibus Test of Model Coefficients**) is somewhat complicated. The first point to note is that the Model and the Block rows are the same (the default) so 'Block' can be ignored here. The second point to note is that the word 'Step' is used in two different ways! Steps 1, 2, ... record the process of successively adding in another predictor (one new variable each time). Within each Step, 'Step' shows the effect on Chi-square for that Step, and 'Model' shows the overall Chi-square for that Step.

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	91.255	1	.000
	Block	91.255	1	.000
	Model	91.255	1	.000
Step 2	Step	6.817	1	.009
	Block	98.072	2	.000
	Model	98.072	2	.000
Step 3	Step	4.252	1	.039
	Block	102.324	3	.000
	Model	102.324	3	.000

- Note: The larger Chi-square is, the more significant (predictive) the model is. The aim, therefore, is to find a model with as large a Chi-square as possible, by choice of variables.
- Step 1 is the constant model and this has a highly significant effect (**Sig.** = 0.000 means $p < 0.0005$). So the constant is assessed to be a useful (significant) predictor.
- Step 2 is the constant plus one variable (note it does not report which variable at this point). The Step 2 Step Chi-square entry of 6.817 is to be added to the Step 1 Model Chi-square value (91.255) to give the corresponding Step 2 Model Chi-square value which is 98.072. This Step too has a highly significant effect (**Sig.** = 0.009). So the constant plus one variable is assessed to be a useful (significant) predictor.
- Step 3 is the constant plus two variables. The Step Chi-square entry of 4.252 is to be added to the previous value (98.072) to give the corresponding Model Chi-square value which is 102.324. This Step too has a significant effect (**Sig.** = 0.039). So the constant plus two variables is assessed to be a useful (significant) predictor.
- Step 4 is NOT there because adding any further variable does not significantly improve the prediction (i.e. Chi-square does not increase much). So the modeling process is terminated.

- ▶ Table 8 (**Model Summary**) shows a measure (- 2 Log likelihood) of how well the model fits the data – a perfect fit would be zero (this column can be ignored here).
- ▶ Table 8 also provides two different estimates of the R² value which indicates what percentage of the dependent variable can be ‘explained’ by the model. Note that this increases with each step, for both ‘conservative’ and ‘optimistic’ estimators. It appears here that the model explains about half the variability.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	179.488 ^a	.366	.494
2	172.671 ^b	.388	.523
3	168.419 ^b	.400	.540

- ▶ Table 9 (**Classification Table** – for Block 1) is important because it shows the predictive ability of the model at each stage.
- ▶ Step three (the final model in this case) reports that the model correctly predicts 80.0% of cases. Of the 118 who actually do not go on to study maths in Y12, it correctly predicts 99 will not (which is 83.9%) and of the 82 who do who do it correctly predicts that 61 will (which is 74.4%), giving overall 80.0% accuracy.

Classification Table^a

Observed			Predicted		
			Studying Maths in Y12		Percentage Correct
			No	Yes	
Step 1	Studying Maths in Y12	No	97	21	82.2
		Yes	22	60	73.2
Overall Percentage					78.5
Step 2	Studying Maths in Y12	No	99	19	83.9
		Yes	22	60	73.2
Overall Percentage					79.5
Step 3	Studying Maths in Y12	No	99	19	83.9
		Yes	21	61	74.4
Overall Percentage					80.0

- ▶ Three further tables appear – Tables 10 to 12 – but only Table 10 (**Variables in the Equation**) is important and reproduced here. It shows which variables are included in the model at each stage. It also gives indicators of the variables' importance (contributions) although it is not always as easy to interpret as for simple linear regression.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	ENJOYMENT_F	.213	.030	49.877	1	.000	1.237
	Constant	-5.486	.771	50.306	1	.000	.004
Step 2 ^b	TEACHER_SUPPORT_F	.084	.034	6.166	1	.013	1.088
	ENJOYMENT_F	.169	.033	26.451	1	.000	1.184
	Constant	-7.156	1.104	41.987	1	.000	.001
Step 3 ^c	CONFIDENCE_F	.071	.036	3.945	1	.047	1.074
	TEACHER_SUPPORT_F	.070	.035	4.078	1	.043	1.072
	ENJOYMENT_F	.135	.036	13.877	1	.000	1.144
	Constant	-8.251	1.321	39.039	1	.000	.000

a. Variable(s) entered on step 1: ENJOYMENT_F.

b. Variable(s) entered on step 2: TEACHER_SUPPORT_F.

c. Variable(s) entered on step 3: CONFIDENCE_F.

- ▶ The **B** column indicator indicates the weight used in the model for the given variable. The bigger the weight the more important the contribution. However, the weight depends upon the scale used for the variable – the bigger the range the smaller will be the weight.
- ▶ In this example, except for **STUDENT_SEX**, the scales are 0 to 48 and 0 to 40 so they are more-or-less comparable and **B** itself is quite a good indicator for comparing the contributions of the variables. Here the Step 3 **B** column entries show that **Enjoyment** (0.135) is by far the most important variable, with **Confidence** (0.71) and **Teacher Support** (0.70) much less important and about equal.
- ▶ The **Wald** column is a measure of the significance of the **B** value for the variable: higher values indicate greater contribution, and take into account the df (degrees of freedom). Here df is '1' for all variables so one can directly look at the Wald numbers. It confirms that **Enjoyment** is by far the most important variable (**Sig.** = 0.000), with **Confidence** and **Teacher Support** much less so and about equal. (For these last two the p values are almost 0.05 – the default entry criterion, so they only just qualify for inclusion.)
- ▶ For the sake of completeness, and not expecting full understanding from all readers, the regression equation for this model is given below. Unlike the simple linear regression model, this produces a probability. If the probability is above the cut value (usually chosen as 0.5) the outcome is considered '1' i.e. 'Yes' and otherwise is considered '0' i.e. 'No'. The equation involves logarithms which can be re-expressed in terms of exponentials like this:

Prob ('1') =

$$1 / \{1 + \exp(-B_0) \times \exp(-B_1 \times \text{Variable 1}) \times \exp(B_2 \times \text{Variable 2}) \times \exp(-B_3 \times \text{Variable 3}) \times \dots\}$$

- ▶ Here $B_0 = -8.251$ Constant weight
- $B_1 = +0.135$ Variable weight for **ENJOYMENT_F**
- $B_2 = +0.070$ Variable weight for **TEACHER_SUPPORT_F**
- $B_3 = +0.071$ Variable weight for **CONFIDENCE_F**

So Prob ('Study maths in Y12') =

$$1 / \{1 + \exp(+8.251) \times \exp(-0.135 \times \text{ENJOYMENT_F}) \times \exp(-0.070 \times \text{TEACHER_SUPPORT_F}) \times \exp(-0.071 \times \text{CONFIDENCE_F})\}$$

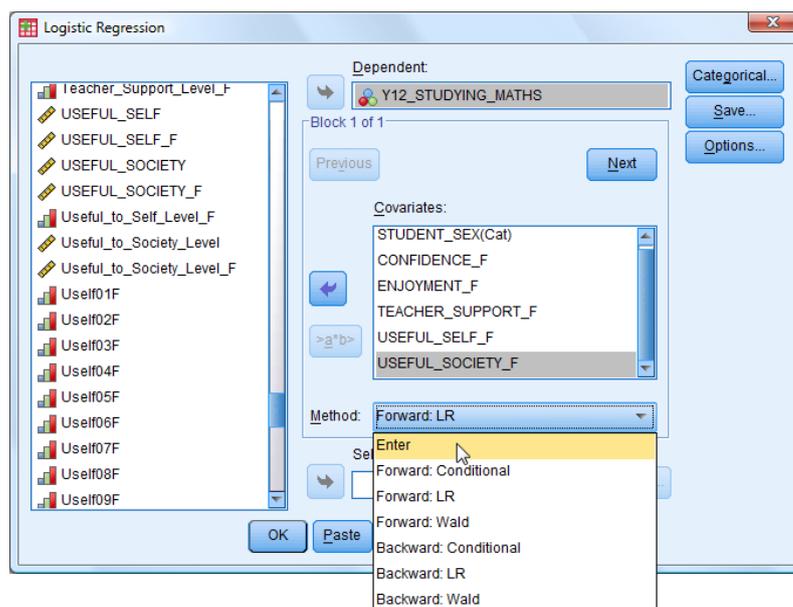
T32.2 Logistic Regression – using Entry Method ‘Enter’

This follows directly on from T32.1 but uses the ‘Enter’ entry method which forces all the selected variables to be used all at once in the model. This is clearly quicker, but does not allow SPSS to decide which variables are not worth including.

If continuing directly from T32.1, you need only repeat step 2 and proceed directly to step 12:

1. Load data file: **File** → **Open** → **Data** → DATA06_School_Maths.sav
2. Select **Analyze** → **Regression** → **Binary Logistic**
3. Select **Y12_STUDYING_MATHS** and move it to the **Dependent** window.
4. Move **STUDENT_SEX** to the **Covariates** window.
5. Click on the **Categorical** button and move the **STUDENT_SEX** variable into the **Categorical Covariates** box which appears.
6. Click **Continue**.
7. Move **CONFIDENCE_F** to the **Covariates** window.
8. Move **ENJOYMENT_F** to the **Covariates** window.
9. Move **TEACHER_SUPPORT_F** to the **Covariates** window.
10. Move **USEFUL_SELF_F** to the **Covariates** window.
11. Move **USEFUL_SOCIETY_F** to the **Covariates** window.
12. Ensure the **Method** for entry is **Enter**.

► It will be **Forward: LR** if continuing directly from T32.1 and must be changed.



13. Click **OK** which generates fewer output tables than before; only the most important two are shown below.

- ▶ The **Classification Table** shows that including all six variables gives an accuracy rate of 82.5% (compared to 80.0% with three variables).

Classification Table^a

Observed			Predicted		
			Studying Maths in Y12 - confirmed		Percentage Correct
			No	Yes	
Step 1	Studying Maths in Y12 - confirmed	No	101	17	85.6
		Yes	18	64	78.0
Overall Percentage					82.5

a. The cut value is .500

- ▶ The **Variables in the Equation Table** shows that the only significant contributors (from the **Wald** and **Sig.** columns) are:

ENJOYMENT_F main contributor with Sig. = 0.002
TEACHER_SUPPORT_F second contributor with Sig. = 0.017
USEFUL_SOCIETY_F marginal contributor with Sig. = 0.047

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a						
STUDENT_SEX(1)	-.184	.412	.200	1	.654	.832
CONFIDENCE_F	.064	.039	2.781	1	.095	1.067
ENJOYMENT_F	.133	.043	9.488	1	.002	1.142
TEACHER_SUPPORT_F	.096	.040	5.667	1	.017	1.101
USEFUL_SELF_F	.055	.033	2.794	1	.095	1.057
USEFUL_SOCIETY_F	-.094	.048	3.928	1	.047	.910
Constant	-7.921	1.495	28.057	1	.000	.000

a. Variable(s) entered on step 1: STUDENT_SEX, CONFIDENCE_F, ENJOYMENT_F, TEACHER_SUPPORT_F, USEFUL_SELF_F, USEFUL_SOCIETY_F.

- ▶ The reason this model has a different third variable is that there are correlations between all the pairs of variables and including the extra three diminishes the influence of **CONFIDENCE_F**.
- ▶ Although **STUDENT_SEX** has a relatively large **B** value, it is in fact of negligible importance (Sig. = 0.654). Its weight **B** is large because it only takes small values 1 and 2 so its range is 1 and its mean about 1.5. The other variables have actual ranges of about close to 40 and 48 and means around 20 to 30 (see the Descriptive Statistics table below to confirm this).

Descriptive Statistics

	N	Minimum	Maximum	Mean
Student's Gender	282	1	2	1.55
Confidence in maths (0-48) [Final]	242	2	48	32.02
Teacher supportiveness (0-48) [Final]	242	4	48	31.95
Usefulness of maths to self (0-48) [Final]	242	1	48	31.72
Enjoyment of maths (0-40) [Final]	232	1	40	22.77
Usefulness of maths to society (0-40) [Final]	230	3	40	27.42
Valid N (listwise)	230			

TUTORIAL T33: Reliability Analysis

T33.1 Reliability Analysis – Introduction

Reliability is the ability of a test to be consistent in its outcome. This differs from Validity which is the ability of a test to measure accurately what it is designed to measure. Both are important in a questionnaire or test. One aspect to achieving a reliable test instrument is to have a series of similar questions about the topic, attitude or opinion under investigation – some worded positively, others negatively – and to assess how consistent the answers are. There are various ways to achieve this – here we introduce Cronbach's Alpha method which is the most popular method. Reliability is measured on a scale of 0 to 1 with 1 indicating perfect reliability and 0 no reliability. A value above 0.75 is generally considered good, and a value above 0.9 is something to aspire to.

Cronbach's Alpha is derived from the mean of the correlations between all pairs of items (r) and the number of items (n). For those interested, the formula is: $\alpha = n \times r / (1 + (n - 1) \times r)$. This means that as n gets larger, alpha will get closer to 1 (however small r is), and, as r gets closer to +1, alpha will get closer to 1 (however small n is).

Cronbach's Alpha method can be used in two ways: to develop a valid coherent set of questions – a 'scale' (by weeding out 'poor' questions) and to provide evidence that a questionnaire used in research is indeed reliable, so conclusions derived have a solid foundation.

Factor Analysis (the subject of T34) can be useful in determining which items go together to form a coherent scale (to produce a scale that is uni-dimensional, or to identify sub-scales within it). The set of identified items forming the scale, or sets forming subscales, can then be tested for reliability.

Here we look at a set of 12 questions about the perceived usefulness of mathematics to school students, which was part of the research project introduced in T31, about which further details can be found in the Appendix.

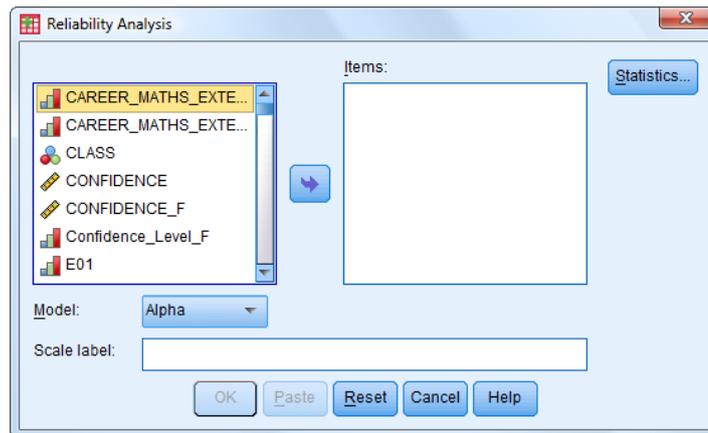
Name	Type	Width	Decimals	Label
Uself01F	Numeric	2	0	USEFUL to SELF 01+ [Final] from S03: Knowing Maths will help me earn a living [pos]
Uself02F	Numeric	2	0	USEFUL to SELF 02+ [Final] from S05: Maths will not be important to me in my life's work [neg]
Uself03F	Numeric	2	0	USEFUL to SELF 03+ [Final] from S08: I'll need Maths for my future work [pos]
Uself04F	Numeric	2	0	USEFUL to SELF 04+ [Final] from S10: I don't expect to use much Maths when I leave school [neg]
Uself05F	Numeric	2	0	USEFUL to SELF 05+ [Final] from S13: Maths is a worthwhile, necessary subject [pos]
Uself06F	Numeric	2	0	USEFUL to SELF 06+ [Final] from S16: Taking Maths is a waste of time [neg]
Uself07F	Numeric	2	0	USEFUL to SELF 07+ [Final] from S21: I will use Maths in many ways as an adult [pos]
Uself08F	Numeric	2	0	USEFUL to SELF 08+ [Final] from S22: I see Maths as something I won't use very often when I leave school [neg]
Uself09F	Numeric	2	0	USEFUL to SELF 09+ [Final] from S26: I'll need a good understanding of Maths for my future work [pos]
Uself10F	Numeric	2	0	USEFUL to SELF 10+ [Final] from S29: Doing well in Maths is not important for my future [neg]
Uself11F	Numeric	2	0	USEFUL to SELF 11+ [Final] from S32: Maths is not important for my life [neg]
Uself12F	Numeric	2	0	USEFUL to SELF 12+ [Final] from S34: I study maths because I know how useful it is [pos]

A typical 'scale' will consist of a set of multiple choice questions on a 5-point scale (or 7-point scale). Before the responses to the questionnaire can be analysed for reliability, the results for all negatively worded questions must be 'turned round', so that for every question on a 5-point scale a '5' means a very positive attitude to the attribute being assessed, and '1' means a very negative attitude.

For a multiple choice question on a 5-point scale this is easily achieved by computing a new variable for a negatively worded question whose values are given by 'new code = 6 – original code'. In this example this has already been done and the 12 'positive' variables produced are shown above, which also shows the actual questions used.

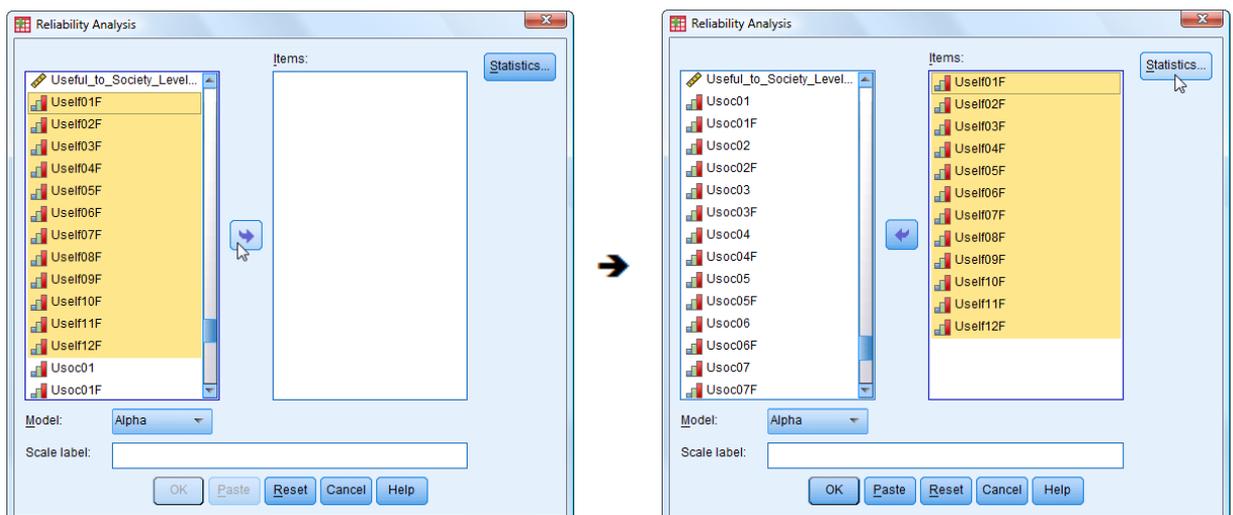
T33.2 Reliability Analysis – Cronbach’s Alpha method – Example 1

1. Load data file: **File** → **Open** → **Data** → DATA06_School_Maths.sav
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **Alphabetical**, to match the list format in this tutorial.
2. Select **Analyze** → **Scale** → **Reliability Analysis**
 - ▶ The **Reliability Analysis** window opens:



- ▶ Note that the default **Model** is **Alpha**, as required.

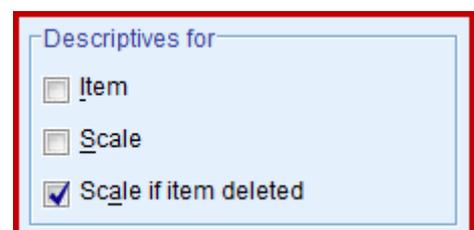
3. Scroll down through the long list of variables to locate and select **Usel01F** to **Usel12F** (this is most easily done by enlarging the window so all the 12 variables are visible, then click the first variable and shift-click the last variable) and move them all together into the **Items** box, as illustrated:



4. Click on the **Statistics** button and choose the **Descriptives for** option **Scale if item deleted**.

- ▶ You could select the other two **Descriptives for** options as well to get further output.

5. Click **Continue**.



6. Click **OK**.

- ▶ This produces the output as a series of tables.
- ▶ Table 1 (**Case Processing Summary**) reports that of the 282 cases available for the analysis, 40 were excluded (due to missing data).

Case Processing Summary

		N	%
Cases	Valid	242	85.8
	Excluded ^a	40	14.2
	Total	282	100.0

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

- ▶ Table 2 (**Reliability Statistics**) reports that for the scale of 12 items Cronbach's Alpha is 0.919. This is a very high value, indicating excellent reliability.

Reliability Statistics

Cronbach's Alpha	N of Items
.919	12

- ▶ Table 3 (**Item-Total Statistics**) lists the 12 items in the scale (and shows their labels which include the questions). The last column is very important as it indicates whether the scale could be improved by excluding any questions.
- ▶ In this case the overall Alpha value is 0.919 but if the last item were removed the Alpha would rise a little to 0.921.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
USEFUL to SELF 01+ [Final] from S03: Knowing Maths will help me earn a living [pos]	39.67	70.984	.622	.914
USEFUL to SELF 02+ [Final] from S05: Maths will not be important to me in my life's work [neg]	40.10	67.313	.693	.911
USEFUL to SELF 03+ [Final] from S08: I'll need Maths for my future work [pos]	40.27	69.037	.674	.912
USEFUL to SELF 04+ [Final] from S10 I don't expect to use much Maths when I leave school [neg]	40.25	66.353	.747	.908
USEFUL to SELF 05+ [Final] from S13: Maths is a worthwhile, necessary subject [pos]	39.69	70.429	.636	.913
USEFUL to SELF 06+ [Final] from S16: Taking Maths is a waste of time [neg]	39.73	67.957	.689	.911
USEFUL to SELF 07+ [Final] from S21: I will use Maths in many ways as an adult [pos]	40.13	67.767	.743	.909
USEFUL to SELF 08+ [Final] from S22: I see Maths as something I won't use very often when I leave school [neg]	40.24	64.683	.801	.906
USEFUL to SELF 09+ [Final] from S26: I'll need a good understanding of Maths for my future work [pos]	40.30	68.311	.728	.910
USEFUL to SELF 10+ [Final] from S29: Doing well in Maths is not important for my future [neg]	40.32	69.788	.520	.919
USEFUL to SELF 11+ [Final] from S32: Maths is not important for my life [neg]	39.98	67.207	.712	.910
USEFUL to SELF 12+ [Final] from S34: I study maths because I know how useful it is [pos]	40.21	71.011	.464	.921

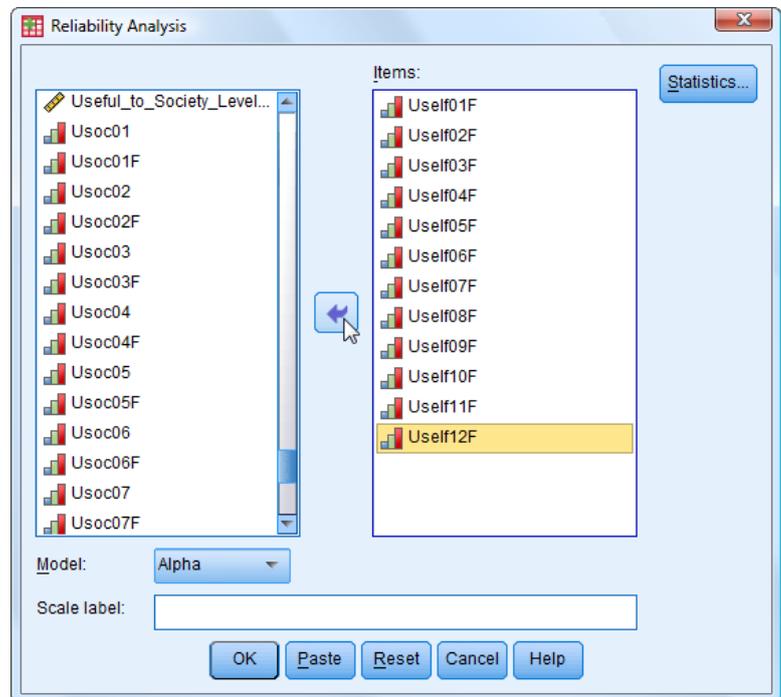
The next step, then, is to repeat the analysis excluding that item, as follows:

7. Select **Analyze** → **Scale** → **Reliability Analysis**

8. Select **Uself12F** and use the blue arrow to remove it from the Items box.

9. Click **OK**.

► This produces the same tables as before.



► This time the Cronbach's Alpha value is 0.921, with N = 11, as predicted in the **Item-Total Statistics** table above.

Reliability Statistics

Cronbach's Alpha	N of Items
.921	11

► What is a little surprising is that new **Item-Total Statistics** Table now says that the Cronbach's Alpha value can be further improved by removing item 10 (see below).

USEFUL to SELF 09+ [Final] from S26: I'll need a good understanding of Maths for my future work [pos]	36.79	59.312	.727	.912
USEFUL to SELF 10+ [Final] from S29: Doing well in Maths is not important for my future [neg]	36.81	60.594	.524	.922
USEFUL to SELF 11+ [Final] from S32: Maths is not important for my life [neg]	36.47	58.208	.716	.912

► The Cronbach's Alpha is now 0.922 using a 10-item scale, and cannot be further improved.

Reliability Statistics

Cronbach's Alpha	N of Items
.922	10

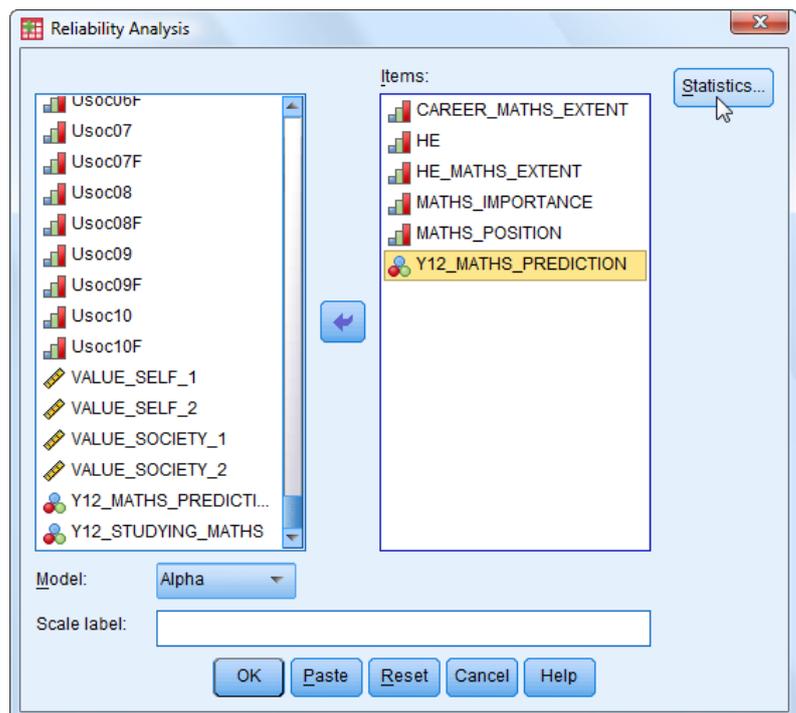
T33.3 Reliability Analysis – Cronbach’s Alpha method – Example 2

Here we attempt to create a scale to measure the commitment of a student to mathematics, using responses to six questions.

1. Load data file: **File** → **Open** → **Data** → DATA06_School_Maths.sav (if not loaded)
 - ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **Alphabetical**, to match the list format in this tutorial.
2. Select **Analyze** → **Scale** → **Reliability Analysis**
3. Click **Reset**.
4. Widen the **Reliability Analysis** window (to be able to see the longest variable names) and locate and move the following seven variables into the **Items** box:

CAREER_MATHS_EXTENT
HE
HE_MATHS_EXTENT
MATHS_IMPORTANCE
MATHS_POSITION
MATHS_Y12
Y12_MATHS_PREDICTION

5. Click the **Statistics** button and choose **Scale** if item deleted.



6. Click **Continue**.
7. Click **OK** to produce, among others, the following tables:

Reliability Statistics

Cronbach's Alpha	N of Items
.544	7

- ▶ Clearly Cronbach's Alpha at 0.544 is low, and bordering on unacceptable (0.5).

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
To what extent do you think Mathematics will feature in your future career? [Initial]	12.86	11.867	.549	.437
Do you expect to go into Higher Education (University or College)? [Initial]	13.54	14.086	.157	.542
To what extent do you expect Mathematics to be part of your study (in HE)? [Initial]	12.86	11.567	.610	.418
How important do you consider Mathematics to be for you? [Initial]	13.25	12.199	.565	.446
Where do you place Maths in your list of favourite academic subjects? [Initial]	10.28	5.128	.470	.521
Do you intend to continue studying Mathematics in Y12? [Initial]	12.69	11.459	.463	.441
Predicted Doing Maths in Y12 [Initial]	14.36	17.494	-.634	.662

- ▶ Studying the **Corrected Item-Total Correlation** column in the table above shows that the second item (**HE**) has a very low correlation (0.157) with the other items combined, but its exclusion at this point would not increase the Alpha value (although it should be removed anyway).
- ▶ The last item in that column (showing the label for variable **Y12_MATHS_PREDICTION**) is negatively correlated with the remaining combined variables. This is not acceptable in a scale. It could be deleted to improve the scale (Alpha would then rise to 0.662). However, as its correlation is actually quite high (–0.634), it is better to keep it but make the correlation positive by reversing the direction of the coding. This we will now do, using the procedure explained in Reference Section 9.1.
- ▶ By looking in **Variable View** at the variable **Y12_MATHS_PREDICTION** (variable 45) we see that currently 'Yes' is coded '1' and 'No' is coded '0'.
- ▶ We will change this so that 'Yes' is coded '1' and 'No' is coded '2'. This can be done by:
 - either **Transform** → **Recode into Different Variables** using '0' → '2' and '1' → '1'
 - or **Transform** → **Compute Variable** using 'new variable' = 2 – 'old variable'

8. Select **Transform** → **Compute Variable**
9. In the **Target Variable** box type in **Y12_MP_NEW**.
10. In the **Numeric Expression** box type '2 – Y12_MATHS_PREDICTION'.
11. Click **OK**.

12. Select **Variable View** and scroll to the bottom of the list of variables to find **Y12_MP_NEW**.
13. Change the **Decimals** to '0'.
14. Type in **Label** 'Y12_PREDICTION'
15. Click on the **Values** cell to open the **Value Labels** window.
16. In the **Value** box enter '1' and in the **Label** box enter 'Yes' and click **Add**.
17. In the **Value** box enter '2' and in the **Label** box enter 'No' and click **Add**.
18. Click **OK**.
19. Make sure the **Measure** is set to 'Nominal'.

► We can now replace the variable **Y12_MATHS_PREDICTION** with **Y12_MP_NEW** and repeat the analysis.

20. Select **Analyze** → **Scale** → **Reliability Analysis**

21. Select **Y12_MATHS_PREDICTION** in the **Items** box and remove it using the blue arrow.

22. Locate and select **Y12_MP_NEW** and move it into the **Items** box.

23. Click **OK**.

Reliability Statistics

Cronbach's Alpha	N of Items
.703	7

► Now Cronbach's Alpha has risen considerably, to 0.703

► Note that in the **Item-Total Statistics** table the correlation of the last item (**Y12_MP_NEW**) is now positive (0.634).

► The last column of the **Item-Total Statistics** table shows that removing the second item would raise Alpha to 0.713 (which has an unacceptably low correlation anyway), and removing the fifth item would raise Alpha to 0.793. Removing both should increase Alpha further. These removals can be done one at a time, or both together.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
To what extent do you think Mathematics will feature in your future career? [Initial]	13.63	16.471	.563	.650
Do you expect to go into Higher Education (University or College)? [Initial]	14.31	19.041	.177	.713
To what extent do you expect Mathematics to be part of your study (in HE)? [Initial]	13.63	16.116	.623	.639
How important do you consider Mathematics to be for you? [Initial]	14.02	16.886	.573	.655
Where do you place Maths in your list of favourite academic subjects? [Initial]	11.05	8.562	.494	.793
Do you intend to continue studying Mathematics in Y12? [Initial]	13.46	15.104	.619	.624
Y12_MP_NEW	14.36	17.494	.634	.662

24. Select **Analyze** → **Scale** → **Reliability Analysis**
25. Select **MATHS_POSITION** in the **Items** box and remove it using the blue arrow.
26. Select **HE** in the **Items** box and remove it using the blue arrow.
27. Click **OK**.
 - ▶ Now Cronbach's Alpha has risen further, to 0.828, which represents high reliability.

Reliability Statistics

Cronbach's Alpha	N of Items
.828	5

- ▶ The last column of the **Item-Total Statistics** table shows that the Item-Total correlations are all reasonably good, so no more variables need to be removed, and removing any would reduce the reliability.
- ▶ The final scale has been determined, containing five items, with reliability 0.828.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
To what extent do you think Mathematics will feature in your future career? [Initial]	7.52	4.975	.646	.788
To what extent do you expect Mathematics to be part of your study (in HE)? [Initial]	7.52	4.866	.678	.779
How important do you consider Mathematics to be for you? [Initial]	7.91	5.512	.550	.814
Do you intend to continue studying Mathematics in Y12? [Initial]	7.35	4.273	.657	.796
Y12_MP_NEW	8.25	5.735	.695	.793

TUTORIAL T34: Factor Analysis

Factor analysis is used to refine a large number of variables (often question responses) into a much smaller number of related groups called components (factors). Here we will begin with a quite small set – ten questions designed to measure a student's level of enjoyment of studying mathematics. This uses the same data set as that in T31 to T33.

T34.1 Factor Analysis – Example 1: 10 variables

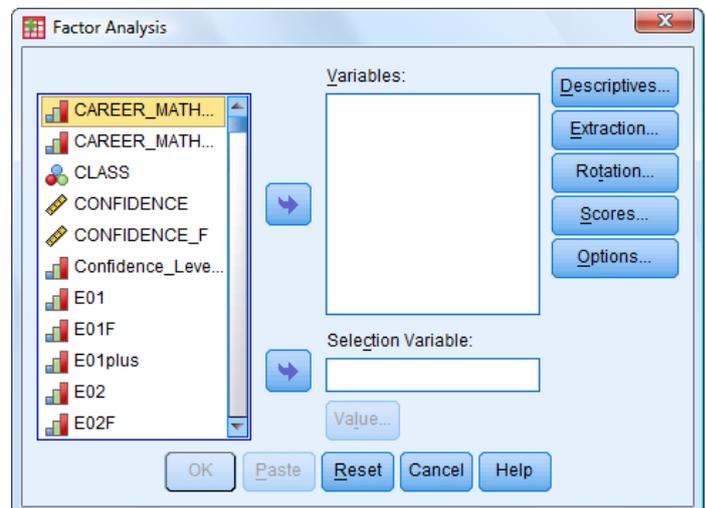
We investigate how many separate components or factors are evident in a set of ten questions purporting to measure 'enjoyment'.

1. Load data file: **File** → **Open** → **Data** → **DATA06_School_Maths.sav** (if not already loaded).

- ▶ Use **Edit** → **Options** to check that in the **General** window the **Variable Lists** choices are **Display names** and **Alphabetical**, to match the list format in this tutorial.

2. Select **Analyze** → **Dimension Reduction** → **Factor**

- ▶ The **Factor Analysis** window opens →
- ▶ The left window lists all scale and ordinal variables which could be included in the analysis.
- ▶ On the right are five buttons which give access to the many options available.



3. Enlarge the window vertically so you can see all ten variables **E01F** to **E10F**, then click on the first and CTRL-click each of the rest to select all ten, then move them into the **Variables** box.

4. Open the **Descriptives** window and select **KMO** and **Bartlett's test of sphericity**.

- ▶ One other option will already be selected – **Initial solution**.

5. Click **Continue**.

6. Open the **Extraction** window.

- ▶ The required **Method** option will already be selected – **Principal components**.
- ▶ The required **Display** option will already be selected – **Unrotated factor solution**.
- ▶ The required **Extract** option will already be selected – **Eigenvalues greater than: 1**.
- ▶ The required **Maximum Iterations** will already be selected: **25**.

7. Click **Continue**.

8. Open the **Rotation window**.
 - ▶ The required **Display** option will already be selected – **Rotated solution**.
9. Select **Varimax**.
10. Click **Continue**.
11. Open the **Options** window.
 - ▶ The required **Missing Values** option will already be selected – **Exclude cases listwise**.
12. Select the **Coefficient Display Format** option **Sorted by size**.
13. Click **Continue**.
14. Click **OK**.
 - ▶ Table 1 (**KMO and Bartlett's Test**) provides reports on two important tests that the data is suitable for Factor Analysis.
 - ▶ **KMO** provides a measure of whether the distributions of values in the variables is suitable. The scale is 0 to 1 with 0.5 the minimum acceptable. The following descriptions for the values obtained have been provided:

0.9+ = marvelous, 0.8+ = meritorious, 0.7+ = middling, 0.6+ = mediocre, 0.5+ = miserable.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.897
Bartlett's Test of Sphericity	Approx. Chi-Square	1602.714
	df	45
	Sig.	.000

- ▶ The value for our set of variables is 0.897 which is bordering on 'marvelous'.
- ▶ **Bartlett's Test of Sphericity** is a measure of the normality of the distributions of values in the variables. A significance value < 0.05 is required.
- ▶ The value for our set of variables is 0.000 (so $p < 0.0005$) which is excellent.
- ▶ Given these two positive results, it is valid to continue with the analysis.

- Table 2 (**Communalities**) lists the 10 variables.

Communalities

	Initial	Extraction
EF01: I enjoy going beyond the work set and trying to solve new problems in Maths [+]	1.000	.575
EF02: Maths is enjoyable and stimulating to me [+]	1.000	.758
EF03: Maths makes me feel uneasy and confused [-]	1.000	.714
EF04: I have never liked Maths, and it is my most dreaded subject [-]	1.000	.720
EF05: I have always enjoyed studying Maths in school [+]	1.000	.689
EF06: I would like to develop my Mathematical skills and study this subject more [+]	1.000	.775
EF07: Maths makes me feel uncomfortable and nervous [-]	1.000	.743
EF08: I am interested and willing to acquire further knowledge of Maths [+]	1.000	.736
EF09: Maths is dull and boring because it leaves no room for personal opinion [-]	1.000	.633
EF10: Maths is very interesting, and I have usually enjoyed courses in this subject [+]	1.000	.752

Extraction Method: Principal Component Analysis.

- It shows that to start the iteration process the 10 variables are all given an initial communality of 1. **Communality** is the amount of variance in the variable explained by all the factors (yet to be found). The value will reduce as the analysis continues, and must lie between 0 and 1 (similar to a multiple correlation).
- It also shows the **Extraction** values for each variable – that is the amount of variance of the variable attributable to the set of components which have been found ('extracted') by the Factor Analysis process – in this case there are two as shown below in the **Total Variance Explained** table. High values are therefore good.
- Table 3 (**Total Variance Explained**) lists the 10 initial eigenvalues for the initially assumed 10 components (or factors) – the same number as there are variables. An eigenvalue of at least 1 is the normal criterion for the existence of a component.
- From the **Initial Eigenvalues Total** column we see that there are just two components identified. From the **Cumulative %** column we see that these account for 71% of the total variance.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.004	60.038	60.038	6.004	60.038	60.038	4.456	44.561	44.561
2	1.091	10.908	70.946	1.091	10.908	70.946	2.639	26.385	70.946
3	.696	6.962	77.908						
4	.470	4.695	82.603						
5	.436	4.362	86.965						
6	.420	4.199	91.164						
7	.290	2.905	94.069						
8	.238	2.378	96.446						
9	.183	1.833	98.279						
10	.172	1.721	100.000						

Extraction Method: Principal Component Analysis.

- Table 4 (**Component Matrix**) lists the 10 variables and shows for each how much of its variance is attributable to each (of the two) components. This is of only limited interest. The rotated component matrix, which follows, is what really matters.

Component Matrix^a

	Component	
	1	2
EF10: Maths is very interesting, and I have usually enjoyed courses in this subject [+]	.863	.085
EF06: I would like to develop my Mathematical skills and study this subject more [+]	.846	.245
EF02: Maths is enjoyable and stimulating to me [+]	.840	.228
EF05: I have always enjoyed studying Maths in school [+]	.823	.108
EF09: Maths is dull and boring because it leaves no room for personal opinion [-]	-.796	-.001
EF08: I am interested and willing to acquire further knowledge of Maths [+]	.792	.329
EF04: I have never liked Maths, and it is my most dreaded subject [-]	-.754	.390
EF01: I enjoy going beyond the work set and trying to solve new problems in Maths [+]	.710	.267
EF03: Maths makes me feel uneasy and confused [-]	-.667	.519
EF07: Maths makes me feel uncomfortable and nervous [-]	-.620	.599

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

- Table 5 (**Rotated Component Matrix**) shows the result of a mathematical axis rotation designed to maximize the effect of one component on a variable and minimize the effect of all other components. The details are not so important. The result is what matters.
- What the table shows is that the first 7 variables listed (sorted by size of eigenvalue as requested in step 12) correspond to one component and the last three correspond to the other component. Note that in the Component 1 column the variances start at 0.840 and go down to -0.659 and then suddenly jump lower at which point Component 2's variances become large ('take over').

Rotated Component Matrix^a

	Component	
	1	2
EF08: I am interested and willing to acquire further knowledge of Maths [+]	.840	-.172
EF06: I would like to develop my Mathematical skills and study this subject more [+]	.837	-.272
EF02: Maths is enjoyable and stimulating to me [+]	.824	-.283
EF10: Maths is very interesting, and I have usually enjoyed courses in this subject [+]	.762	-.414
EF05: I have always enjoyed studying Maths in school [+]	.742	-.373
EF01: I enjoy going beyond the work set and trying to solve new problems in Maths [+]	.737	-.177
EF09: Maths is dull and boring because it leaves no room for personal opinion [-]	-.659	.446
EF07: Maths makes me feel uncomfortable and nervous [-]	-.177	.843
EF03: Maths makes me feel uneasy and confused [-]	-.260	.804
EF04: I have never liked Maths, and it is my most dreaded subject [-]	-.405	.746

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

- ▶ Table 6 (**Component Transformation Matrix**) contains the multiplying matrix used to convert the initial **Transformation Matrix** into the **Rotated Transformation Matrix**. It is likely to be of interest only to advanced users.

Component Transformation Matrix

Component	1	2
1	.828	-.561
2	.561	.828

Extraction Method: Principal
Component Analysis.
Rotation Method: Varimax with
Kaiser Normalization.

- ▶ What we have done so far is – perhaps – the easy bit. What remains is the interpretation. This requires a careful look at what components constitute each component. This requires knowledge of the research field and the source of the variables. Even though this is not your own research, here it is not too difficult. The source is a set of 10 questions about ‘enjoyment’ included in a much bigger questionnaire given to Y11 students.
- ▶ If you read the 7 questions in Component 1 you will observe that they are all about ‘interest’ in the subject (one is negatively worded).
- ▶ If you read the 3 questions in Component 2 you will observe that they are all about ‘feeling uncomfortable’ about the subject.
- ▶ It does seem that the Factor Analysis has picked out two distinct components in what the researchers originally thought was a single ‘construct’ (to use a psychological term).
- ▶ This was a very small set – sets of hundreds of variables (questions) are quite normal and, in such cases, obviously the computer is a vital tool in finding components. But the interpretation has to be done by a person!

T34.2 Factor Analysis – Example 2: 36 variables

1. Load data file: **File** → **Open** → **Data** → DATA06_School_Maths.sav (if not already loaded).
2. Select **Analyze** → **Dimension Reduction** → **Factor**
 - ▶ The **Factor Analysis** window opens.
3. If variables are already selected click **Reset**.
4. Enlarge the Factor Analysis window vertically as much as possible to reveal as many variables in the list as you can.
5. We need to select **S01** to **S36** and move them all into the **V**ariables box. The variables will probably be listed in 'Display Variable Names' order which is not very helpful here as the ones we want will appear mixed in with others we don't want. It is better if they are listed in 'Display Variable Labels' order, then all 36 variables S01 to S36 will be listed next to each other.

Do this by right-clicking on the variable list and selecting 'Display Variable Labels', then click on the first (**S01**) and SHIFT-click on the last (**S36**), then move them into the **V**ariables box. (Do it in two halves if necessary.)
6. Open the **D**escriptives window and select **KMO** and **Bartlett's test of sphericity**.
7. Click **Continue**.
8. Open the **R**otation window.
9. Select **Varimax**.
10. Click **Continue**.
11. Open the **O**ptions window.
12. Select **S**orted by size.
13. Click **Continue**.
14. Click **OK**.

- ▶ In Table 1 **KMO** provides a measure of whether the distributions of values in the variables is suitable.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.919
Bartlett's Test of Sphericity	Approx. Chi-Square	5027.596
	df	630
	Sig.	.000

- ▶ The value for our set of 36 variables is 0.919 which is in the 'marvellous' category.
- ▶ **Bartlett's Test of Sphericity** is a measure of the normality of the distributions of values in the variables. A significance value < 0.05 is required.

- ▶ The value for our set of variables is 0.000 (so $p < 0.0005$) which is excellent.
- ▶ Given these two positive results, it is valid to continue with the analysis.
- ▶ Table 2 (**Communalities**) lists the 36 variables in name order (**S01 to S36**).

Communalities

	Initial	Extraction
S01: I am sure that I can learn Maths [C+]	1.000	.466
S02: My teachers have been interested in my progress in Maths [T+]	1.000	.727
S03: Knowing Maths will help me earn a living [U+]	1.000	.648
S04: I don't think I could do advanced Maths [C-]	1.000	.540
S05: Maths will not be important to me in my life's work [U-]	1.000	.548
S06: Getting a teacher to take me seriously in Maths is a problem [T-]	1.000	.569
S07: Maths is hard for me [C-]	1.000	.621
S08: I'll need Maths for my future work [U+]	1.000	.752
S09: I am sure of myself when I do Maths [C+]	1.000	.542
S10: I don't expect to use much Maths when I leave school [U-]	1.000	.622
S11: I would talk to my Maths teachers about a career that uses Maths [T+]	1.000	.495
S12: It's hard to get Maths teachers to respect me [T-]	1.000	.678
S13: Maths is a worthwhile, necessary subject [U+]	1.000	.654
S14: I'm not the type to do well in Maths [C-]	1.000	.669
S15: My teachers have encouraged me to study more Maths [T+]	1.000	.637
S16: Taking Maths is a waste of time [U-]	1.000	.688
S17: I have a hard time getting teachers to talk seriously with me about Maths [T-]	1.000	.703
S18: Maths has been my worst subject [C-]	1.000	.586
S19: I think I could handle more difficult Maths [C+]	1.000	.612
S20: My teachers think advanced Maths will be a waste of time for me [T-]	1.000	.644
S21: I will use Maths in many ways as an adult [U+]	1.000	.665
S22: I see Maths as something I won't use very often when I leave school [U-]	1.000	.681
S23: I feel that Maths teachers ignore me when I try to talk about something serious [T-]	1.000	.661
S24: Most subjects I can handle OK, but I just can't do a good job with Maths [C-]	1.000	.673
S25: I can get good grades at Maths [C+]	1.000	.593
S26: I'll need a good understanding of Maths for my future work [U+]	1.000	.708
S27: My teachers want me to take all the Maths I can [T+]	1.000	.659
S28: I know I can do well at Maths [C+]	1.000	.678
S29: Doing well in Maths is not important for my future [U-]	1.000	.587
S30: My teachers would not take me seriously if I told them I was interested in a career in Science and/or Maths [T-]	1.000	.449
S31: I am sure I could do advanced work in Maths [C+]	1.000	.659
S32: Maths is not important for my life [U-]	1.000	.580
S33: I'm no good at Maths [C-]	1.000	.638
S34: I study maths because I know how useful it is [U+]	1.000	.555
S35: Maths teachers have made me feel I have the ability to go on in Maths [T+]	1.000	.669
S36: My teachers think I'm the kind of person who could do well in Maths [T+]	1.000	.575

Extraction Method: Principal Component Analysis.

- ▶ Table 3 (**Total Variance Explained** – partly shown) lists the initial 36 eigenvalues. An eigenvalue of at least 1 is the normal criterion for the existence of a component.
- ▶ From the **Initial Eigenvalues Total** column we see that there are 7 components identified. From the **Cumulative %** column we see that these account for 62% of the total variance.

Component	Total Variance Explained				
	Initial Eigenvalues			Extraction Sums of Squares	
	Total	% of Variance	Cumulative %	Total	% of Variance
1	11.336	31.488	31.488	11.336	31.488
2	3.463	9.620	41.109	3.463	9.620
3	2.793	7.758	48.867	2.793	7.758
4	1.670	4.639	53.506	1.670	4.639
5	1.117	3.103	56.609	1.117	3.103
6	1.039	2.886	59.495	1.039	2.886
7	1.014	2.818	62.312	1.014	2.818
8	.950	2.638	64.950		
9	.877	2.436	67.386		

- ▶ Table 4 (**Component Matrix** – not shown) lists the 36 variables and shows for each how much of its variance is attributable to each of the (two) components.
- ▶ Table 5 (**Rotated Component Matrix**) shows that:
 - the first 12 variables listed correspond to Component 1,
 - the next 12 variables listed correspond to Component 2,
 - the next 6 variables listed correspond to Component 3.

These three components account for 30 of the 36 variables (i.e. questions).
 The last four components account for the last 6 variables.

Rotated Component Matrix^a

	Component						
	1	2	3	4	5	6	7
S07: Maths is hard for me [C-]	.780	.023	.063	.007	.008	-.001	.091
S14: I'm not the type to do well in Maths [C-]	.748	-.172	.202	.053	-.160	.039	-.096
S19: I think I could handle more difficult Maths [C+]	.740	.198	.039	.084	.119	.020	.042
S24: Most subjects I can handle OK, but I just can't do a good job with Maths [C-]	.737	-.164	.287	-.076	.108	-.050	-.024
S09: I am sure of myself when I do Maths [C+]	.707	.085	-.038	.097	-.004	-.105	.116
S31: I am sure I could do advanced work in Maths [C+]	.703	.320	-.020	.124	.211	-.046	.012
S04: I don't think I could do advanced Maths [C-]	.690	-.187	.003	-.062	-.098	-.069	.102
S33: I'm no good at Maths [C-]	.690	-.235	.226	-.123	.074	-.026	-.184
S18: Maths has been my worst subject [C-]	.674	-.110	.167	-.145	.240	-.091	.068
S28: I know I can do well at Maths [C+]	-.609	.216	-.219	.029	.177	-.049	.422
S01: I am sure that I can learn Maths [C+]	-.523	.270	-.167	-.173	.209	.129	.031
S36: My teachers think I'm the kind of person who could do well in Maths [T+]	-.465	.135	-.152	.352	.274	.346	.098
S08: I'll need Maths for my future work [U+]	-.114	.761	-.011	-.047	.285	.251	-.153
S21: I will use Maths in many ways as an adult [U+]	-.113	.757	-.096	.082	.046	.011	.248
S26: I'll need a good understanding of Maths for my future work [U+]	-.175	.723	.044	-.065	.332	.172	-.101
S22: I see Maths as something I won't use very often when I leave school [U-]	.254	-.714	.194	-.154	-.159	.108	-.092
S10: I don't expect to use much Maths when I leave school [U-]	.304	-.704	.121	-.053	-.096	-.083	.012
S13: Maths is a worthwhile, necessary subject [U+]	-.173	.675	-.249	.024	-.049	.100	.305
S03: Knowing Maths will help me earn a living [U+]	-.161	.659	-.093	-.079	-.013	.371	.189
S29: Doing well in Maths is not important for my future [U-]	.266	-.641	.151	-.101	.057	.129	.229
S32: Maths is not important for my life [U-]	.165	-.635	.285	-.210	.021	.154	.000
S05: Maths will not be important to me in my life's work [U-]	.067	-.628	.076	.121	.073	-.070	.344
S34: I study maths because I know how useful it is [U+]	.020	.610	-.056	.304	.053	-.261	.126
S16: Taking Maths is a waste of time [U-]	.259	-.588	.407	-.091	.216	-.115	-.210
S17: I have a hard time getting teachers to talk seriously with me about Maths [T-]	.098	-.093	.818	-.061	-.105	.009	-.019
S12: It's hard to get Maths teachers to respect me [T-]	.070	-.091	.790	-.023	.017	-.131	-.152
S23: I feel that Maths teachers ignore me when I try to talk about something serious [T-]	.163	-.195	.765	-.040	-.085	-.026	-.040
S06: Getting a teacher to take me seriously in Maths is a problem [T-]	.067	-.127	.723	-.102	.076	-.090	.042
S30: My teachers would not take me seriously if I told them I was interested in a career in Science and/or Maths [T-]	.265	-.219	.530	-.154	-.097	-.046	.123
S20: My teachers think advanced Maths will be a waste of time for me [T-]	.466	.003	.464	.023	-.414	-.100	.200
S15: My teachers have encouraged me to study more Maths [T+]	-.047	-.001	-.095	.779	.107	.076	.036
S35: Maths teachers have made me feel I have the ability to go on in Maths [T+]	-.240	.240	-.301	.639	.076	.197	-.081
S27: My teachers want me to take all the Maths I can [T+]	-.205	.203	.036	.434	.076	.003	.001
S11: I would talk to my Maths teachers about a career that uses Maths [T+]	.005	.150	-.350	.155	.490	-.001	.294
S02: My teachers have been interested in my progress in Maths [T+]	.025	.108	-.291	.277	.010	.741	.055
S25: I can get good grades at Maths [C+]	-.519	.051	.016	-.057	.074	.197	.522

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 11 iterations.

- ▶ The Factor Analysis reported in Table 5 clearly identifies three components and, less convincingly, four more. We now look at the first three in detail.
- ▶ Component 1:

	1
S07: Maths is hard for me [C-]	.780
S14: I'm not the type to do well in Maths [C-]	.748
S19: I think I could handle more difficult Maths [C+]	-.740
S24: Most subjects I can handle OK, but I just can't do a good job with Maths [C-]	.737
S09: I am sure of myself when I do Maths [C+]	-.707
S31: I am sure I could do advanced work in Maths [C+]	-.703
S04: I don't think I could do advanced Maths [C-]	.690
S33: I'm no good at Maths [C-]	.690
S18: Maths has been my worst subject [C-]	.674
S28: I know I can do well at Maths [C+]	-.609
S01: I am sure that I can learn Maths [C+]	-.523
S36: My teachers think I'm the kind of person who could do well in Maths [T+]	-.455

- ▶ It can be seen that 11 of the 12 variables were designated as 'Confidence' questions. The twelfth (S36), although designated as a 'Teacher support' question can be viewed as relating to 'Confidence'. This is a very convincing component (or factor). It is a near-perfect match with the researcher's intended factor.

- ▶ Component 2:

S08: I'll need Maths for my future work [U+]	-.114	.761
S21: I will use Maths in many ways as an adult [U+]	-.113	.757
S26: I'll need a good understanding of Maths for my future work [U+]	-.175	.723
S22: I see Maths as something I won't use very often when I leave school [U-]	.254	-.714
S10: I don't expect to use much Maths when I leave school [U-]	.304	-.704
S13: Maths is a worthwhile, necessary subject [U+]	-.173	.675
S03: Knowing Maths will help me earn a living [U+]	-.161	.659
S29: Doing well in Maths is not important for my future [U-]	.266	-.641
S32: Maths is not important for my life [U-]	.165	-.635
S05: Maths will not be important to me in my life's work [U-]	.067	-.628
S34: I study maths because I know how useful it is [U+]	.020	.610
S16: Taking Maths is a waste of time [U-]	.259	-.586

- ▶ It can be seen that all 12 of these variables were designated as 'Useful to self' questions. This is a very convincing component (or factor). It is a perfect match with the researcher's intended factor.

► Component 3:

S17: I have a hard time getting teachers to talk seriously with me about Maths [T-]	.098	-.093	.818
S12: It's hard to get Maths teachers to respect me [T-]	.070	-.091	.790
S23: I feel that Maths teachers ignore me when I try to talk about something serious [T-]	.163	-.195	.765
S06: Getting a teacher to take me seriously in Maths is a problem [T-]	.067	-.127	.723
S30: My teachers would not take me seriously if I told them I was interested in a career in Science and/or Maths [T-]	.265	-.219	.530
S20: My teachers think advanced Maths will be a waste of time for me [T-]	.456	.003	.464

- It can be seen that all 6 of these variables were designated as 'Teacher support' questions. Although there are 6 other 'Teacher support' questions which have 'got lost elsewhere' this is a convincing component (or factor). It is in one sense a perfect match with the researcher's intended factor.
- It is not worth spending time trying to make sense of the remaining four components – especially as they have so few questions. In a much larger study the researcher would wish to investigate further.
- As a final point, there are links between Reliability Analysis (TUTORIAL T33) and Factor Analysis (TUTORIAL T34). Having developed and trialled a large set of questions, one can test the reliability of subsets (potential components) and use Factor Analysis to confirm the validity of these and also discover other subsets for further investigation and trialling, leading to the discovery of new 'constructs'. Thus the two methods can work well together.

DATA SET 1: 100 Top-Selling Books

SPSS Data File name: DATA01_100Books.sav

This data file contains details of the top 100 best selling books in the period 1989 to 2010.

Cases: 100

Variables: 15: position in the top 100, title, author(s), publisher (imprint), publisher group, number of books sold (volume), sales value (£), recommended retail price (RRP), average selling price (ASP), type of binding (paperback or hardback), month of publication, year of publication, product class code (detailed coding used by publishers to categorise books), genre (crime, fiction, etc), type (adult fiction, children's fiction, non-fiction).

Data source: The Guardian, Nielsen.

QUESTIONS

1. What are the mean, median and standard deviation of the average selling price (ASP) of books in the top 100 best sellers list?
2. What are the mean, median and standard deviation of the number of books sold?
3. What genre of books sold the most?
4. What is the number of hardback books and paperback books in the 100 best sellers list?
5. Which publisher has sold the most books?
6. Which author has sold the most books?
7. How many authors have more than two books in the list?
8. How many of the best-selling books were published in the five years 2006 to 2010? How does this compare to the previous five years 2001 to 2005?
9. Which month appears to be the best month for launching (a) a paperback, (b) a hardback? Would you offer advice to publishers on this data? If not, why not?
10. Is there a significant difference between the recommended retail price (RRP) of the books compared to the average selling price (ASP)?
11. Is there a significant difference between the average selling price (ASP) of fiction and non-fiction books?
12. Is the distribution of average selling price (ASP) of books normal?

DATA SET 2: VLE Questionnaire

SPSS Data File name: DATA03_LSquestionnaire.sav

These questions relate to responses from 150 Information Science students to a questionnaire concerning a university's VLE. The questionnaire itself is presented on the next page.

Cases: 150

Variables: 24

Data source: Confidential

QUESTIONS

1. What are the differences between undergraduate and postgraduate use of the VLE? Are they important?
 2. Are there any differences in use of the VLE between different programmes?
 3. Are there any differences in use of the VLE between full-time and part-time students?
 4. Are there any differences in use of the VLE between male and female students?
 5. How many students did not access the VLE?
 6. How many modules do students typically access on the VLE? Are their differences for different groups of students?
 7. What information do students access on the VLE? Are their differences for different groups of students?
 8. What do students think about the VLE? Are their differences for different groups of students?
 9. Which individual module correlates most highly with the final programme mark?
-

RCUK Open Access Survey – Introduction

INTRODUCTION

SQWconsulting and LISU (Loughborough University) were commissioned by Research Councils UK (RCUK) to identify the effects and assess the impact of Open Access to research outputs on pay-to-publish and self-archiving publishing models.

Open Access models provide free online access to research literature either by publishing in an Open Access journal which does not charge ('Gold' OA) or by archiving peer-reviewed articles published in subscription journals ('Green' OA).

INSTITUTIONAL DATA

A data file supplied with this Guide (DATA04_OpenAccess_HEIs.sav) includes responses from all 39 replying institutions (of 168 contacted); just under half of the questions in the survey are included.

The questions for which response data are provided in DATA04_OpenAccess_HEIs.sav are included on the following pages.

RESEARCHER DATA

A data file supplied with this Guide (DATA05_OpenAccess_Researchers.sav) includes responses from 418 replying individuals (of 2122 total responses); about half of the questions in the survey are included. The institutional variables are included in the Researchers' dataset.

The questions for which response data are provided in DATA05_OpenAccess_Researchers.sav are included on the following pages. See the end of this Guide for links to the Open Access Report.

QUESTION SETS IN THIS GUIDE FOR USE WITH THE SUPPLIED DATASETS

There are three sets of questions:

Set 3: General questions covering both datasets, not specifically linked to SPSS.

Set 4: Questions covering the Institutional dataset, specifically linked to SPSS.

Set 5: Questions covering the Researcher dataset, specifically linked to SPSS.

QUESTIONNAIRES

The actual questionnaires used by SQWconsulting/LISU are presented, in abbreviated form, after the three sets of questions which now follow.

DATA SET 3: RCUK Survey on Open Access – General**SPSS Data File**

DATA04_OpenAccess_HEIs.sav

Cases: 39**Variables:** 27**Data source:** LISU**SPSS Data File**

DATA05_OpenAccess_Researchers.sav

Cases: 418**Variables:** 77**Data source:** LISU**QUESTIONS**

1. According to researchers, what percentage of institutions have their own repository? How does this differ from what the institutions say? Why do you think there is a difference?
2. According to researchers, is self-archiving of research outputs in (a) institutional repositories, (b) subject-based repositories or (c) on project websites *mandated, encouraged, tolerated* or *discouraged*? Does this differ from what the institutions say on the matter?
3. Where have researchers published their research in the previous five years? Is this influenced by the category of researcher?
4. What dates were given for the researchers' most recent open access publication? Was one category of researcher more prolific than the others?
5. What were the researchers' main reasons for publishing in an open access journal or repository?
6. What were the researchers' general views regarding open access?
7. Who did the researchers think should bear the cost of publication of research outputs? How does this differ from how the institutions say open access is funded at their institutions?
8. When do researchers anticipate open access becoming the normal route for publication of research outputs in their discipline?
9. According to institutions, does their library include open access publications in their catalogues?
10. How is material deposited in the university repositories where they exist?
11. Have measures been put in place to encourage authors to deposit material where it is not mandated/required?
12. What was the mean number of items deposited or downloaded in the 2006-07 academic year? Did this vary between different types of institutions?
13. What were the mean number of total items held and the number individual depositors? Did this vary between different types of institutions?
14. Describe the running cost of the institutional repositories in 2006-07.

DATA SET 4: RCUK Survey on Open Access - Institutions

SPSS Data File name: DATA04_OpenAccess_HEIs.sav

QUESTIONS

1. (a) Use **Frequencies** to check the accuracy of the 'No of responses' column in Table B-1 below
[Source: *LISU/SQWconsulting* Open Access to Research Outputs: Annexes, page 11].

(b) Use **Crosstabs** to check the accuracy of the 'FTE students' column.

Table B-1 Institution survey response rate

Institution type	No of responses	FTE students *
RLUK ¹ member	12	206,575
Other pre-1992 university	12	117,020
Post-1992 university	13	203,669
HE college	2	14,455
Total	39	541,719
Response rate	23%	31%

Source: HESA, 2006-07

2. Use **Crosstabs** to check the accuracy of Table B-3 below [Source: Annexes page 13].

Table B-3 Does your institution have a written policy on open access to research outputs?

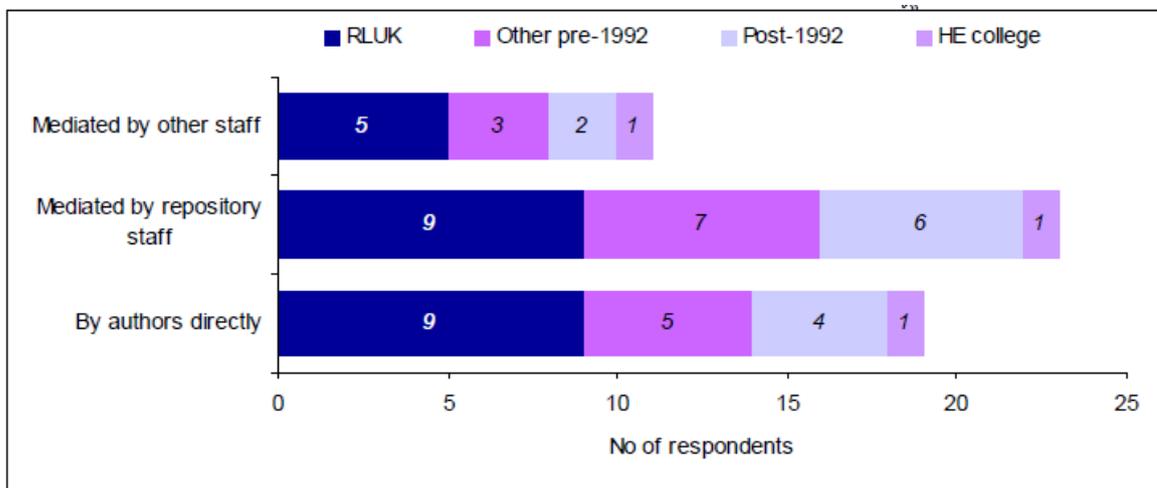
	RLUK	Other pre-1992	Post-1992	HE college	Total	
Yes, a policy is in place	4	3	2	-	9	23%
No, but a policy is planned	5	3	5	-	13	33%
No, such a policy has been rejected	-	2	-	-	2	5%
No, such a policy has not been considered to my knowledge	3	4	6	2	15	38%
Total	12	12	13	2	39	100%

3. Use **Frequencies** to check the veracity of the statement below [Source: Annexes page 18].

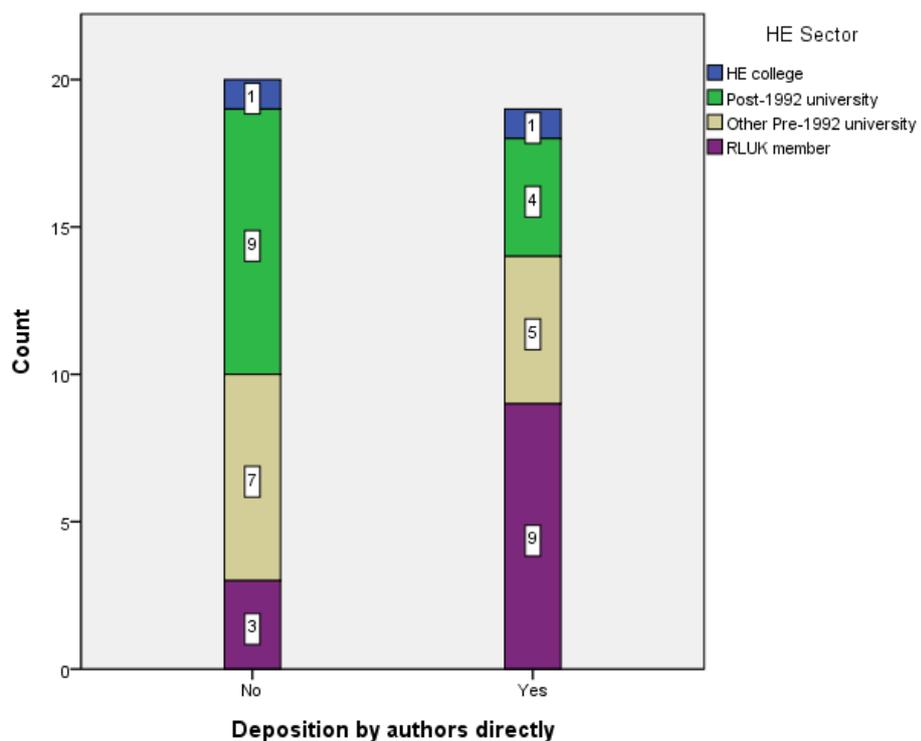
Two-thirds of responding institutions had their own repositories, and a further quarter had plans to introduce one.

- Use an appropriate procedure to check the accuracy of the data in Figure B-2 below [Source: Annexes page 21]

Figure B-2 How is material deposited in the repository?



- The chart below is essentially the same as that above for 'Mediated by repository staff'. Use **Chart Builder** to produce a similar chart for 'By authors directly', The chart above (Figure B-2) can be used as a check you have the correct data.



6. Use an appropriate procedure to check the accuracy of Table B-13 below [Source: Annexes page 23]

You will find some discrepancies. Look at the 'No. included' column which may help – at least partly – to explain how these discrepancies may have arisen.

Table B-13 Summary statistics for institutional repositories (2006-07 academic year)

	Minimum	Median	Maximum	No. included	No. "not known"
No of items deposited in year	0	185	9,000	18	1
No of downloads in year	0	34,000	350,000	11	5
Total items held currently	0	1,089	9,800	21	-
Number of individual depositors	0	202	1,700	10	4
Staff costs	£0	£15,000	£77,000	10	1
Hardware	£0	£0	£9,000	13	1
Software	£0	£0	£26,000	13	-
Other costs	£0	£0	£1,000	6	3
Total costs	£0	£9,000	£77,000	7	1

DATA SET 5: RCUK Survey on Open Access – Researchers

SPSS Data File name: DATA05_OpenAccess_Researchers.sav

QUESTIONS

1. Use **Frequencies** to check the whether the statement below [Source: Annexes page 18] is in accord with the responses provided by Researchers in the sample data set.

Two-thirds of responding institutions had their own repositories, and a further quarter had plans to introduce one.

2. Table B-15 below shows on the right the number of researchers in each category based on the valid responses made by 2116 of the 2122 cases.
[Source: LISU/SQWconsulting Open Access to Research Outputs: Annexes, page 24].

The smaller data set supplied for use with this Guide has 418 cases, which is a 20% sample.

Use the **One-sample Chi-square Test** to check whether the sample of 418 cases is representative of the whole survey dataset of 2116 valid cases for 'Researcher category' (i.e. check that it has appropriate numbers in each category of researcher to accurately preserve proportionality).

Table B-15 Demographics of the respondents

a) Institution type	Count	%	b) Researcher category	Count	%
RLUK member	1,243	59%	Professor	661	31%
Other pre-1992 university	652	31%	Senior lecturer/ reader	466	22%
New university	125	6%	Lecturer	272	13%
HE college	4	0%	Research associate / fellow	238	11%
Other research institute	98	5%	Postgraduate student	408	19%
Total	2,122	100%	Other	71	3%
			Total	2,116*	100%

* Six respondents did not indicate a research category

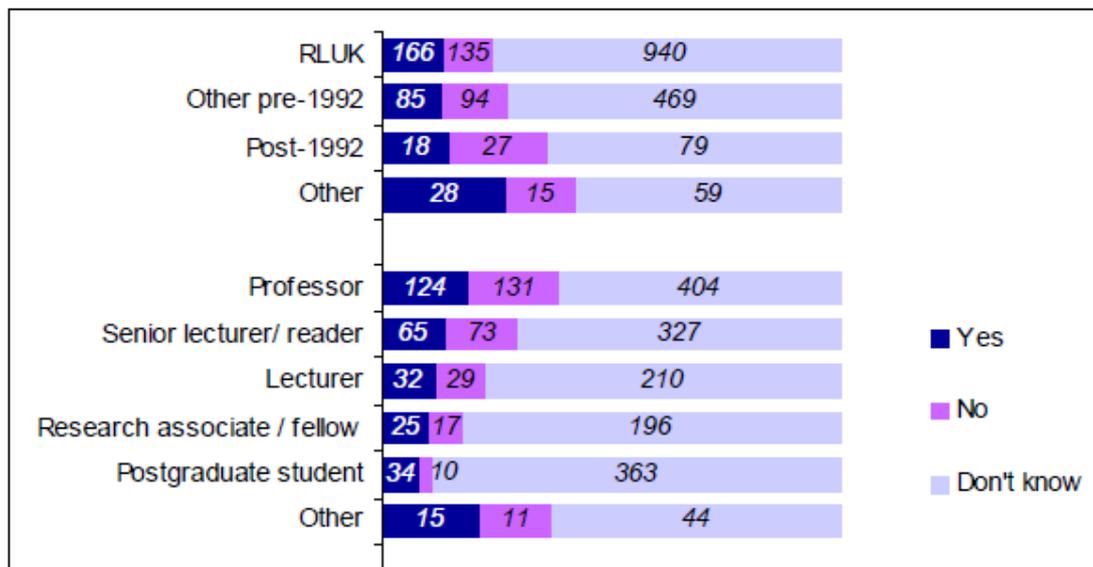
Notes:

- (a) The **One-sample Chi-square Test** is found by selecting
Analyze → **Nonparametric Tests** → **Legacy Dialogs** → **Chi-square**
- (b) You will need to analyse Q02, entering into the **Expected Values** box each number for the six **Researcher categories** in the **Count** column above. [See TUTORIAL T26 for an example.]

- Use **Chart Builder** to replicate the lower chart in Figure B-3 below, for Research staff

[Source: Annexes page 27].

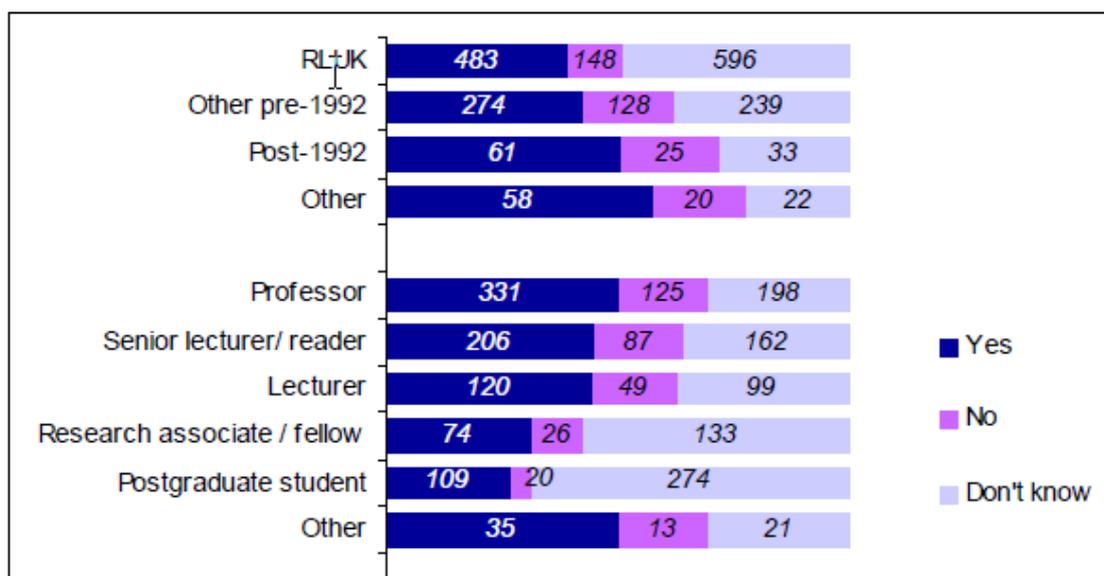
Figure B-3 Does your institution have a written policy on open access to research outputs?



Note: The detailed procedures for obtaining the upper chart demonstrated given in TUTORIAL T15.2.

- Figure B-6 below [Source: Annexes page 29] reports on whether the researcher said his institution does or does not have its own repository. Carry out an analysis to see how many of those who claimed to know, were correct.

Figure B-6 Does your institution have its own repository?



Notes:

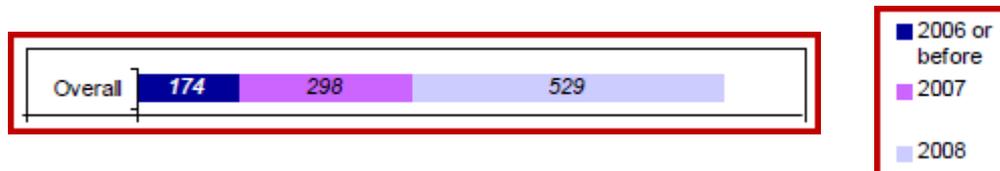
This will entail using the Researchers' data file (**DATA05_OpenAccess_Researchers.sav**), which includes all the variables from the Institutions data file, which are named **Inst.Q1**, **Inst.Q2a**, etc.

5. (a) Use **Crosstabs** and the associated **Chi-square Test** to check the accuracy of the statement below [Source: Annexes page 38], by comparing with results from the analysis of the sample data set.

Authors were asked to rate a series of seven possible reasons for choosing to publish open access on this occasion on a scale from 1 (very important) to 5 (not at all important). The most important factor, with 74% of respondents rating this at 1 or 2, was speed of dissemination. There were differences in response pattern between the different categories of researcher, but not by institution type.

6. Use the **One-sample Chi-square Test** to check whether the sample of 418 cases is representative of the whole survey dataset of 2122 cases for the variable 'Year of most recent Open Access publication'. Use the data in Figure B-9 below [Source: Annexes page 36] to provide the 'Expected Values'.

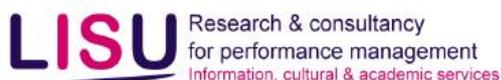
Figure B-9 In what year was your most recent open access publication published?



Notes:

- (a) This is similar to Question 2 above. [See T16.3 for an example.]
- (b) You will need to recode the variable to have just three values for 'Year'.

RCUK Survey – RESEARCHER QUESTIONNAIRE



Survey 1: Open access - researchers' view and practices

The aim of this survey is to investigate researchers' views and practices with regard to open access to research outputs, whether this may be publishing in open access journals, self-archiving material published in subscription based journals, or any other form.

1. What is the name of your institution?

.....

2. Which category of researcher are you?

- Professor
- Senior lecturer/reader
- Lecturer
- Research associate/fellow
- Postgraduate student
- Other Please specify :

3. Does your institution have a written policy on open access to research outputs?

- Yes
- No
- Don't know

4. Does your institution have its own repository?

- Yes
- No
- Don't know

5. At your institution, are the following ways of self-archiving research outputs generally:

	Mandated/ Required?	Encouraged?	Tolerated?	Discouraged?	Don't know	Not applicable
a. Institutional repository	<input type="checkbox"/>					
b. Subject-based repository	<input type="checkbox"/>					
c. Author/project website	<input type="checkbox"/>					

6. In the last 5 years, have you published research outputs in: (select all that apply)

- a. Subscription-based journals?
- b. Research monographs?
- c. Open access or hybrid journals: Free to authors?
- d. Open access or hybrid journals: Paid to publish?
- e. Institutional repository?
- f. Subject-based repository?
- g. Own/project website?
- h. Other? Please specify:

7. In what year was your most recent open access publication published?

.....

8. What was the type of output for this publication? (select all that apply)

- a. Article in open access journal
- b. Open access article in hybrid journal
- c. Deposit in own institutional repository
- d. Deposit in other repository
- e. Self publication on own/project website
- f. Other? Please specify:

9. How important were the following reasons in your decision to publish open access on this occasion?

	Very Important	—————>			Not important at all
	1	2	3	4	5
a. Principle of free access for all	<input type="checkbox"/>				
b. Prestige/quality impact of journal	<input type="checkbox"/>				
c. Speed of dissemination	<input type="checkbox"/>				
d. Availability to researchers with limited access to subscribed journals	<input type="checkbox"/>				
e. Possibility of increased citations to the output	<input type="checkbox"/>				
f. Mandate by institution	<input type="checkbox"/>				
g. Mandate by funder	<input type="checkbox"/>				

10. How much do you agree with the following statements?

	Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
a. Open access materials are an important source of material for	<input type="checkbox"/>				
b. Open access journals publish material more quickly than other	<input type="checkbox"/>				
c. Publishing in open access journals means a work is read	<input type="checkbox"/>				

	<i>Strongly Agree</i>	<i>Agree</i>	<i>Neutral</i>	<i>Disagree</i>	<i>Strongly Disagree</i>
d. I am wary of material published only on the author's own website	<input type="checkbox"/>				
e. The proliferation of versions associated with open access is	<input type="checkbox"/>				
f. Open access outputs are likely to be of lower quality than non open	<input type="checkbox"/>				
g. It is valuable to see the datasets underlying research papers	<input type="checkbox"/>				
h. I only cite peer-reviewed material in my own research	<input type="checkbox"/>				
i. Publishers are entitled to impose embargo periods before something they have published can be made available by open access	<input type="checkbox"/>				
j. The embargo periods specified by most publishers are fair	<input type="checkbox"/>				
k. Open access has changed the way I work with researchers elsewhere	<input type="checkbox"/>				
l. I am happy to cite pre-prints in my published research	<input type="checkbox"/>				
m. There are open access journals with high status in my field of	<input type="checkbox"/>				
n. Publishing my work in open access journals may adversely	<input type="checkbox"/>				

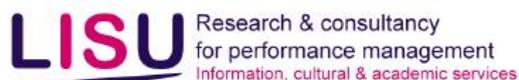
11. How much responsibility do you think the following should bear for the cost of publication of research outputs?

	<i>Most</i>	—————→			<i>Least</i>
	1	2	3	4	5
a. Author	<input type="checkbox"/>				
b. Reader/user	<input type="checkbox"/>				
c. Project funder	<input type="checkbox"/>				
d. Department	<input type="checkbox"/>				
e. Library	<input type="checkbox"/>				
f. Institution	<input type="checkbox"/>				
g. Government	<input type="checkbox"/>				
h. Publisher	<input type="checkbox"/>				

12. When do you anticipate open access becoming the normal route for publication of research outputs in your discipline?

- It is already
- Within 5 years
- Within 10 years
- Longer than that
- Never
- Don't know

RCUK Survey – INSTITUTIONAL QUESTIONNAIRE



Survey 2: Open access – institutional policies

The aim of this survey is to investigate academic institutions' policies with regard to open access to research outputs, and to survey the current position in respect of institutional repositories of research material.

1. Does your institution have a written policy on open access to research outputs?

- Yes, a policy is in place
- No, but a policy is planned
- No, such a policy has been rejected
- No, such a policy has not been considered to my knowledge
- Don't know

2. How is pay-to-publish open access funded at your institution? (select all that apply)

- Where specifically included in research funding
- From indirect costs administered at faculty/department level
- From indirect costs administered centrally
- From author's own resources e.g. discretionary funds
- We do not have any mechanism to support author pays
- Don't know Other Please specify:

3. Are the following ways of self-archiving research outputs generally:

	Mandated/ Required?	Encouraged?	Tolerated/	Discouraged?	Don't know	Not applicable
a. Institutional repository	<input type="checkbox"/>					
b. Subject-based repository	<input type="checkbox"/>					
c. Author/project website	<input type="checkbox"/>					

4. Does the library include open access publications in its catalogue?

- Yes
- No
- Don't know

If yes: please state (approximately) how many titles are included:

- a. Journals
- b. E-books

5. Does your institution have its own repository? (Select one box only)

- No, but a repository is planned
- No, a repository has been rejected
- No, having a repository has not been considered to my knowledge
- No, but we participate in a repository consortium
- Yes

6. How is material deposited in the repository? (select all that apply)

- a. By authors directly
- b. Mediated by repository staff
- c. Mediated by other staff

7. Are any measures in place to encourage authors to deposit material where it is not mandated/required?

- Yes
- No
- Don't know

If yes please describe

.....

.....

.....

8. Does the library include material from its own repository in its catalogue?

- Yes
- No
- Don't know

9. Please give approximate figures for:

- a. Number of items deposited (2006-07 academic year)
- b. Number of downloads (2006-07 academic year)
- c. Total items held (currently)
- d. Number of individual depositors

10. If available, what were the approximate running costs of the repository in the 2006-07 academic year (to nearest £1,000)?

- a. Staff costs
- b. Hardware
- c. Software
- d. Other
- e. Total

DATA SET 6: School Maths Research Project NCETM

SPSS Data File name: DATA07_School_Maths.sav

Cases: 282

Variables: 180

Data source: Confidential

This data file contains numerous statistics for 282 Y11 school students (aged 15-16 years) participating in a National Centre for Excellence in Teaching Mathematics (NCETM) funded research project investigating, among other things, factors which determine whether a student will continue studying mathematics into Y12. Each Y11 student participant was given an 'Initial Questionnaire' at the start of the project, in September 2008, and essentially the same questionnaire was administered again near the end of the project – the 'Final Questionnaire' – in May 2009, before public examinations were sat.

A slightly reduced version of the Initial Questionnaire is provided on the following pages. The reader is encouraged to look at this questionnaire, as it will make it much easier to understand the description of variables which now follows.

Names and labels for variables in the dataset which derived from the Initial Questionnaire either have no special identifier or are indicated by '[Initial]'.

Names and labels for variables in the dataset derived from the Final Questionnaire have the special identifier 'F' or '[Final]'.

Apart from seeking demographic information, there were individual questions about favourite subjects, future aspirations and study intentions, and then large sets of related questions aimed at investigating the students' views on five aspects of mathematics:

- Confidence (12 questions)
- Teacher-Support (12 questions)
- Usefulness-to-oneself (12 questions)
- Enjoyment (10 questions)
- Usefulness-to-society (10 questions)

All these questions were in multiple choice format, using a five-point scale with '5' indicating 'Strongly Agree' and '1' indicating 'Strongly Disagree'.

The first three sets of questions were mixed in together (S01 to S36).

The last two sets stood alone (E01 to E10 and Usoc01 to Usoc10).

The Final questionnaire's equivalent variables are S01F to S36F, E01F to E10F, Usoc01F to Usoc10F.

Some questions were positively worded and some negatively. For creating composite indices and in regression analysis the negative responses had to be 'turned round' and made 'positive'.

The Usefulness-to-self questions have been copied from the S01–S30 set and transformed to all be positive and stored as Uself01–Uself12. Those from the Final Questionnaire are: Uself01F–Uself12F.

The Enjoyment questions have been copied and transformed to all be positive and stored as E01plus–E10plus.

A small number of other variables have been derived from the above. All variables have explanatory labels.

SCHOOL MATHS RESEARCH PROJECT (NCETM) QUESTIONS

Note: Most of these questions require the more advanced statistical tests and SPSS procedures.

T TEST

1. Use the Paired-Samples t Test to see if each of these three pairs of variables can be considered to have come from populations with the same mean:
 - (a) **CONFIDENCE** and **CONFIDENCE_F**
 - (b) **TEACHER_SUPPORT** and **TEACHER_SUPPORT_F**
 - (c) **USEFUL_SELF** and **USEFUL_SELF_F**
2. Use the Independent-Samples t Test to see if the scores for female students (code = 1) and male students (code = 2) can be considered to have the same mean:
 - (a) **CONFIDENCE**
 - (b) **ENJOYMENT**
 - (c) **USEFUL_SOCIETY**

REGRESSION

3. Create a simple regression equation to predict **USEFUL_SELF** from **USEFUL_SOCIETY**. Report on its validity and accuracy.
4. Create a multiple regression equation using the 'Enter' entry method to predict **USEFUL_SELF** from **MATHS_IMPORTANCE**, **MATHS_POSITION**, **MATHS_Y12**. Report on its validity and accuracy.
5. Repeat 4 using the 'Forward' entry method. Report on its validity and accuracy, and compare to 4.
6. Use logistic regression with the 'Enter' entry method to predict the binary variable **GCSE_AstarA** from **CONFIDENCE_F**, **ENJOYMENT_F**, **STUDENT_SEX**, **TEACHER_SEX**, **TEACHER_SUPPORT_F**. Report its validity and accuracy.
7. Repeat 6 using the 'Forward: LR' entry method. Report its validity and accuracy, and compare to that obtained in 6.
8. There is a flaw in applying the analysis in 6 and 7 to this particular dataset because quite a lot of the students included in the project had already taken and passed their GCSE mathematics the year before (usually with high grades), and were studying for AS mathematics or other mathematical qualifications. It would be more sensible to exclude them from the analysis.

This is done using **Data** → **Select Cases** as described in Section 8.2. The variable to use is **GCSE_YEAR** which should be '2009'. If you do this you will find somewhat different results. The excluded cases will have a horizontal line through the case number. Also the message 'Filter on' will appear at the bottom right of the **Data Editor** window in the Status bar (it can be easily overlooked; it is not shown at the bottom of the **Viewer** Output window).

Important Note: It is very important afterwards to turn off this selection by revisiting **Select Cases** again using **Data** → **Select Cases** and clicking the **All cases** button. Then check that the horizontal lines through the case numbers have all disappeared and that the 'Filter On' message has gone from the **Status** bar.

ANOVA

9. Perform a One-way ANOVA to test to determine whether the variable **CONFIDENCE** is significantly different for male students and female students.
10. Perform a One-way ANOVA to test to determine whether the variable **ENJOYMENT** is significantly different for the three levels of teacher support recorded in the variable **Teacher_Support_Level**, which has 1 as lowest level and 3 as highest level.
11. Perform a One-way ANOVA to test to determine whether the variable **CONFIDENCE_F** is significantly different for the four levels recorded in the variable **Useful_to_Society_Level_F**, which has 1 as lowest level and 4 as highest level.
12. Perform a One-way ANOVA to determine whether the mean of the variable **TEACHER_SUPPORT** is significantly different from that for **TEACHER_SUPPORT_F**.
13. Perform a One-way ANOVA to test to determine whether the means of the four variables **USEFUL_SELF**, **USEFUL_SELF_F**, **USEFUL_SOCIETY** and **USEFUL_SOCIETY_F** have significantly different means. (When you finished this task, but not before, be sure to look at 14.)
14. There is a flaw in applying the ANOVA to the variables in 13 because the variables are of two different scales (0-48) and (0-40). Use **Transform** → **Compute Variable** to overcome this problem, and repeat the analysis, and compare your findings.
15. Perform a Two-way ANOVA to test to determine whether the variable **CONFIDENCE** is significantly different for male students and female students, for classes taught by male teachers and those taught by female teachers, and whether there is an interaction effect.
16. Perform a Two-way ANOVA to test to determine whether the variable **ENJOYMENT** is significantly different for male students and female students, for the four levels recorded in the variable **Useful_to_Society_Level**, which has 1 as lowest level and 4 as highest level, and whether there is an interaction effect.

KOLMOGOROV-SMIRNOV TEST

17. Use the Kolmogorov-Smirnov test to determine which of the following can be considered to have normal distributions: **CONFIDENCE**, **TEACHER_SUPPORT**, **USEFUL_SELF**.
18. Use the Kolmogorov-Smirnov test to determine which of the following can be considered to have normal distributions: **CONFIDENCE_F**, **TEACHER_SUPPORT_F**, **USEFUL_SELF_F** and compare with the results in 13.

RELIABILITY

19. Use Cronbach's Alpha method to determine the best possible scale to measure the 'commitment' of a student to mathematics, initially using the following six variables. (It doesn't work out as one would expect, despite being very similar to the example in T33.3.)

CAREER_MATHS_EXTENT
HE
HE_MATHS_EXTENT
MATHS_IMPORTANCE
MATHS_POSITION
MATHS_Y12

20. Use Cronbach's Alpha method to determine the best possible scale to measure the 'joy-of-maths' of a student to mathematics, initially using these variables:

E01plus, E02plus, E03plus, E04plus, E05plus, E07plus, E09plus, E10plus
MATHS_IMPORTANCE
MATHS_Y12_REV

21. (a) Starting from the 12 relevant variables within the 36 variables **S01** to **S36**, create a single variable 'CONSCALE' as a scale for Confidence-with-mathematics, with '0' as its lowest possible value. You will need to see which 12 of the 36 variables relate to Confidence. Do this by looking at their labels.

Method 1

This can be done in 3 stages:

Step 1: Make all 12 variables positive either by recoding or by computing to produce 12 new variables ('CON01' to 'CON12').

[If using **T**ransform → **R**ecode ... the recoding steps are 1→5, 2→4, 3→3, 4→2, 5→1.]

[If using **T**ransform → **C**ompute Variable the equation is 'new variable' = 6 – 'old variable'.]

Step 2: Use **C**ompute Variable to produce a new variable 'CONTOTAL' by adding together the 12 variables resulting from Step 1.

Step 3: Note that with individual scores ranging from 1 to 5, the range of 12 variables added will not have '0' as its minimum. So a constant must be subtracted from 'CONTOTAL' to produce 'CONSCALE'.

Method 2

The above can be done *all in one step* using the **C**ompute Variable procedure – but it is not easy to get right first time. It is wise to write down the equation on paper and test it 'by hand' first. Make sure '0' is the minimum achievable.

- (b) Use **F**requencies or some other procedure to compare the distribution for 'CONSCALE' with that of 'CONFIDENCE' which is a variable already in the dataset. They should match exactly.
- (c) Having created the 12 variables 'CON01' to 'CON12', perform a Reliability Analysis to calculate Cronbach's Alpha reliability score, and determine whether removal of any of the 12 variables would improve the score.

FACTOR ANALYSIS

22. Perform a Factor Analysis on the ten variables **Usoc01** to **Usoc10** in the Initial Questionnaire. Discuss your findings. (You should find three components.) (See 23 for related analysis.)
23. Perform a Factor Analysis on the ten variables **Usoc01F** to **Usoc10F** in the Final Questionnaire (the questions were identical to those for **Usoc01** to **Usoc10** analysed in 22). Discuss your findings and compare with those from 10. (You should find two components.)
24. Perform a Factor Analysis on the 22 variables **Usef01F** to **Usef12F** and **Usoc01F** to **Usoc10F** which purport to cover the two aspects of usefulness – for oneself and for society. Discuss your findings. (You should find three components.)

SCHOOL MATHS RESEARCH PROJECT (NCETM) QUESTIONNAIRE

NCETM QUESTIONNAIRE PART 1			
GROUP:	<input style="width: 90%;" type="text"/>		
SCHOOL:	<input style="width: 90%;" type="text"/>		
DATE:	<input style="width: 50%;" type="text"/>	DATE OF BIRTH:	<input style="width: 50%;" type="text"/>
GENDER:	MALE:	<input type="checkbox"/>	FEMALE: <input type="checkbox"/>
1 In which year are you taking GCSE Mathematics?			
	2008	<input type="checkbox"/>	2009 <input type="checkbox"/>
2 What are your 3 favourite academic subjects at school (in order)?			
	1st favourite	<input style="width: 90%;" type="text"/>	
	2nd favourite	<input style="width: 90%;" type="text"/>	
	3rd favourite	<input style="width: 90%;" type="text"/>	
If Mathematics is not included above, in what position would you put it?			
			<input type="checkbox"/>
3 How important do you consider Mathematics to be <u>for you</u>?			
	Extremely important	<input type="checkbox"/>	
	Quite important	<input type="checkbox"/>	
	Not very important	<input type="checkbox"/>	
	Not at all important	<input type="checkbox"/>	
4 Do you intend to continue studying Mathematics in Y12?			
	I will definitely study Mathematics in Y12	<input type="checkbox"/>	
	I am likely to study Mathematics in Y12	<input type="checkbox"/>	
	I am unlikely to study Mathematics in Y12	<input type="checkbox"/>	
	I am not going to study Mathematics in Y12	<input type="checkbox"/>	
5 Do you intend to continue studying Mathematics in Y13?			
	I will definitely study Mathematics in Y13	<input type="checkbox"/>	
	I am likely to study Mathematics in Y13	<input type="checkbox"/>	
	I am unlikely to study Mathematics in Y13	<input type="checkbox"/>	
	I am not going to study Mathematics in Y13	<input type="checkbox"/>	

NCETM QUESTIONNAIRE PART 1 continued

6 Do you intend to study Further Mathematics in Y12?	
I will definitely study Further Mathematics in Y12	<input type="checkbox"/>
I am likely to study Further Mathematics in Y12	<input type="checkbox"/>
I am unlikely to study Further Mathematics in Y12	<input type="checkbox"/>
I am not going to study Further Mathematics in Y12	<input type="checkbox"/>
7 Do you expect to go into Higher Education (University or College)?	
Definitely	<input type="checkbox"/>
Likely	<input type="checkbox"/>
Unlikely	<input type="checkbox"/>
Definitely not	<input type="checkbox"/>
8 If you go into Higher Education what subject area or areas are you likely to study?	
<input style="width: 100%;" type="text"/>	
<input style="width: 100%;" type="text"/>	
9 If you go into Higher Education, to what extent do you expect Mathematics to be part of your study?	
Mathematics will be very important	<input type="checkbox"/>
Mathematics will be quite important	<input type="checkbox"/>
Mathematics will be of little importance	<input type="checkbox"/>
Mathematics will not feature	<input type="checkbox"/>
It is not possible for me to say	<input type="checkbox"/>
10 What ideas do you have about your future career?	
<input style="width: 100%;" type="text"/>	
<input style="width: 100%;" type="text"/>	
11 To what extent do you think Mathematics will feature in your future career?	
Mathematics will be very important	<input type="checkbox"/>
Mathematics will be quite important	<input type="checkbox"/>
Mathematics will be of little importance	<input type="checkbox"/>
Mathematics will not feature	<input type="checkbox"/>
It is not possible for me to say	<input type="checkbox"/>

NCETM QUESTIONNAIRE PART 2

This questionnaire is about how you feel about yourself and Mathematics.	
Below are some statements. You are to mark your sheet to show how you feel about them	
For example, suppose a statement says: " Algebra is boring "	
As you read the statement, you will know whether you agree or disagree.	
If you strongly agree , circle 5 , next to the statement.	
If you agree , but not so strongly, or you only "sort of agree", circle 4 .	
If you strongly disagree circle 1 .	
If you disagree , but not so strongly, circle 2 .	
If you are not sure and are undecided , circle 3 .	
NOTE: There are no "right" or "wrong" answers.	STRONGLY DISAGREE DISAGREE Undecided AGREE STRONGLY AGREE
The only correct responses are those that are true FOR YOU .	
Remember that '5' means you strongly agree with the statement.	
STATEMENTS	
1 I am sure that I can learn Maths.	5 4 3 2 1
2 My teachers have been interested in my progress in Maths.	5 4 3 2 1
3 Knowing Mathematics will help me earn a living.	5 4 3 2 1
4 I don't think I could do advanced Maths.	5 4 3 2 1
5 Maths will not be important to me in my life's work.	5 4 3 2 1
6 Getting a teacher to take me seriously in Maths is a problem.	5 4 3 2 1
7 Maths is hard for me.	5 4 3 2 1
8 I'll need Mathematics for my future work.	5 4 3 2 1
9 I am sure of myself when I do Maths.	5 4 3 2 1
10 I don't expect to use much Maths when I leave school.	5 4 3 2 1
11 I would talk to my Maths teachers about a career that uses Maths.	5 4 3 2 1
12 It's hard to get Maths teachers to respect me.	5 4 3 2 1
13 Maths is a worthwhile, necessary subject.	5 4 3 2 1
14 I'm not the type to do well in Maths.	5 4 3 2 1
15 My teachers have encouraged me to study more Maths.	5 4 3 2 1
16 Taking Maths is a waste of time.	5 4 3 2 1
17 I have a hard time getting teachers to talk seriously with me about Maths.	5 4 3 2 1
18 Maths has been my worst subject.	5 4 3 2 1
19 I think I could handle more difficult Maths.	5 4 3 2 1
20 My teachers think advanced Maths will be a waste of time for me.	5 4 3 2 1
21 I will use Mathematics in many ways as an adult.	5 4 3 2 1
22 I see Mathematics as something I won't use very often when I leave school.	5 4 3 2 1
23 I feel that Maths teachers ignore me when I try to talk about something seri	5 4 3 2 1
24 Most subjects I can handle OK, but I just can't do a good job with Maths.	5 4 3 2 1

25	I can get good grades in Maths.	5	4	3	2	1
26	I'll need a good understanding of Maths for my future work.	5	4	3	2	1
27	My teachers want me to take all the Maths I can.	5	4	3	2	1
28	I know I can do well in Maths.	5	4	3	2	1
29	Doing well in Maths is not important for my future.	5	4	3	2	1
30	My teachers would not take me seriously if I told them I was interested in a career in Science and/or Mathematics.	5	4	3	2	1
31	I am sure I could do advanced work in Maths.	5	4	3	2	1
32	Maths is not important for my life.	5	4	3	2	1
33	I'm no good at Maths.	5	4	3	2	1
34	I study Maths because I know how useful it is.	5	4	3	2	1
35	Maths teachers have made me feel I have the ability to go on in Mathematics.	5	4	3	2	1
36	My teachers think I'm the kind of person who could do well in Maths.	5	4	3	2	1

NCETM QUESTIONNAIRE PART 3

Please indicate your level of agreement or disagreement with each of these statements.		STRONGLY DISAGREE	DISAGREE	Undecided	AGREE	STRONGLY AGREE
<u>ENJOYMENT</u>						
1	I enjoy going beyond the work set and trying to solve new problems in Maths	5	4	3	2	1
2	Mathematics is enjoyable and stimulating to me	5	4	3	2	1
3	Mathematics makes me feel uneasy and confused	5	4	3	2	1
4	I have never liked Mathematics, and it is my most dreaded subject	5	4	3	2	1
5	I have always enjoyed studying Mathematics in school	5	4	3	2	1
6	I would like to develop my Mathematical skills and study this subject more	5	4	3	2	1
7	Mathematics makes me feel uncomfortable and nervous	5	4	3	2	1
8	I am interested and willing to acquire further knowledge of Mathematics	5	4	3	2	1
9	Mathematics is dull and boring because it leaves no room for personal opinion	5	4	3	2	1
10	Mathematics is very interesting, and I have usually enjoyed courses in this subject	5	4	3	2	1
<u>USEFULNESS</u>						
1	Mathematics has contributed greatly to science and other fields of knowledge	5	4	3	2	1
2	Mathematics is less important to people than art or literature	5	4	3	2	1
3	Mathematics is not important for the advance of civilization and society	5	4	3	2	1
4	Mathematics is a very worthwhile and necessary subject	5	4	3	2	1
5	An understanding of Mathematics is needed by artists and writers as well as scientists	5	4	3	2	1
6	Mathematics helps develop people's minds and teaches them to think	5	4	3	2	1
7	Mathematics is not important in every day life	5	4	3	2	1
8	Mathematics is needed in designing practically everything	5	4	3	2	1
9	Mathematics is needed in order to keep the world running	5	4	3	2	1
10	There is nothing creative about Mathematics; it's just memorizing formulas and things	5	4	3	2	1

DATA SET 7: IT Piracy Worldwide

SPSS Data File name: DATA07_IT_Piracy.sav

This data file contains data on IT piracy for 109 countries, published by Business Software Alliance, who define the IT piracy rate as the percentage of all software in use which is pirated. See the end of this Guide for details of their annual reports which explain the methodology and provide the raw data.

Cases: 109

Variables: 21: Country, Region of the world, Region of the world (as used by BSA), Population (2008, 2009, 2010), GDP (2008, 2009, 2010), IT piracy rates for 2005 to 2010, IT piracy values (US\$ millions) for 2005 to 2010.

Data sources: Business Software Alliance, The World Bank, The IMF.

QUESTIONS

1. Which region had the highest (a) piracy rates and (b) piracy values in each of the years 2005 to 2010?
2. Which 10 countries have the highest (a) piracy rates and (b) highest piracy values in 2010?
3. For **Q2** are the 10 countries the same for (a) and (b)? If not why do you think there is a difference?
4. Which 10 countries have the least (a) piracy rates and (b) least piracy values in 2010?
5. For **Q4** are the 10 countries the same for (a) and (b)? If not why do you think there is a difference?
6. (a) How strongly is piracy rate correlated to population size and with GDP?
(b) What is the correlation with population when the effect of GDP is removed? (i.e. find the partial correlation).
(c) What is the correlation with GDP when the effect of population is removed? (i.e. find the partial correlation).
7. (a) How strongly is piracy value correlated to population size and with GDP?
(b) What is its partial correlation with population when the effect of GDP is removed?
(c) What is its partial correlation with GDP when the effect of population is removed?
8. Do your answers to **Q6** and **Q7** show the same trends? What is your explanation?
9. Did the **piracy rates** increase or decrease over the period 2005-2010? (Hint: draw a graph of the countries with the highest piracy rates for 2010 over the period 2005-2010, and another graph for countries with the lowest piracy rates).
10. Do the same as in **Q9** for **piracy values**. Do these show the same trends? What is your explanation?
11. Use boxplots to compare the distributions of piracy rates by region.
12. Use the Chi-square test to compare the mean piracy rates by region.
13. Calculate the piracy value per capita for each country, and compare the mean by region.

DATA SET 8: Facebook Users Worldwide

SPSS Data File name: DATA08_Facebook_Users.sav

This data file contains statistics for 157 of the world's largest countries (population at least 1 million in 2011) for whom the information is available, provided by Internet World Stats (IWS). See the end of this Guide for links to the IWS reports and datasets.

Cases: 157

Variables: 6: Country, region of world, population mid-2011 (estimate March 2011), GDP in purchasing power parity (PPP) per capita per annum in international dollars (estimates at April 2011), number of Facebook users (June 2011).

Data sources: Internet World Statistics, IMF.

QUESTIONS

1. Calculate the Facebook penetration rate for each country, defined as:

$$\text{Number of Facebook Users} / \text{Population size} \times 100$$

[Preferably, round these numbers to the nearest integer, using RND.]

2. What is the correlation between population size and penetration? What do you conclude?
3. What is the effect of GDP per capita on Facebook membership?
4. Does the region affect the Facebook penetration rate? (Hint: Perform a suitable test to compare mean penetration rates for the regions.)
5. Does the size of a country's population affect the Facebook penetration rate? (Hint: Assign the countries to different size categories, and perform a suitable test to compare mean penetration rates for the categories.)

DATA SET 9: Internet Users in Europe

SPSS Data File name: DATA09_Internet_Users_Europe.sav

This data file contains statistics on internet usage for 31 European countries, provided by Europa. See the end of this Guide for links to the Europa report and datasets.

Cases: 31

Variables: 9: Country, region within Europe, GNI (Gross National Income) in purchasing power parity (PPP) per capita per annum in US\$ 2009, Internet usage rate by 16-24 year-olds in 2009, Internet usage rate by 16-74 year-olds in 2009, Internet access rate in 2007 and in 2009, Internet buyer rate for males and for females 2009.

Data sources: Europa, The World Bank.

QUESTIONS

1. Calculate for each country the Internet penetration rate for 2007 and for 2009, defined as:

$$\text{Number of Internet Users} / \text{Population size} \times 100$$

[Preferably, round these numbers to the nearest integer, using RND.]

2. Are there significant regional differences in the Internet penetration rates found in **Q1**? If so, is there an explanation?

3. Calculate for each country the internet user growth rate from 2007 to 2009, defined as:

$$\{ (\text{Internet users 2009} - \text{Internet users 2007}) / \text{Population 2007} \} \times 50$$

[Preferably, round these numbers to the nearest integer, using RND.]

4. Are there significant regional differences in the Internet user growth rates found in **Q3**? If so, is there an explanation?
5. Is there a significant correlation between population size and number of internet users? Investigate separately for each year. What is the explanation of your findings?
6. Is there a significant correlation between population size and internet user growth rate? What is the explanation of your findings?
7. Is there a significant correlation between GNI and number of internet users? Investigate separately for each year. What is the explanation of your findings?
8. Is there a significant correlation between GNI and Internet penetration rate? Investigate separately for each year. What is the explanation of your findings?
9. Is there a significant correlation between GNI and internet user growth rate? What is the explanation of your findings?

10. Is there a significant correlation between GNI and Internet penetration rate? What is the explanation of your findings?
11. Are there differences in usage rates for younger people and the whole population? What is the explanation of your findings?

[Note: the data for 16-24 year-old is included in the data for 16-74 year-olds, and unfortunately cannot be separated out as only rates and not numbers are provided. It could be done approximately if the percentage of 16-24 year olds in the population of 16-74 year olds were found.]

12. Are there regional differences in **Q11**? What is the explanation of your findings?
 13. Are there differences in usage rates for males and females? What is the explanation of your findings?
 14. Are there regional differences in **Q13**? What is the explanation of your findings?
-

DATA SET 10: Demographics Worldwide

SPSS Data File name: DATA10_Demographics.sav

This data file contains demographic statistics for 155 of the world's largest countries (population at least 1 million in 2010) for whom at least some of the information is available, provided by UNESCO Institute for Statistics, The World Bank and Internet World Stats (IWS). See the end of this Guide for links to their reports and datasets.

Cases: 155

Variables: 15: Country, region of the world, subregion, land area, adult literacy rate (latest data), population (2008, 2009, 2010), GNI per capita (2008, 2009, 2010) in US\$, GDP total (2008, 2009, 2010) in US\$ millions.

Data sources: UNESCO Institute for Statistics, World Bank, Internet World Stats.

QUESTIONS

LITERACY

1. Is there a significant correlation between Literacy rate and population?
2. Is there a significant correlation between Literacy rate and GNI per capita? If so, what is the explanation?
3. Is there a significant correlation between Literacy rate and GDP total? What is the explanation?
4. What is the average person's Literacy rate? (NB This is not the average of the rates for countries.)
5. Using the variables Region, GNI per capita and GDP to find the regression equation to predict a country's literacy rate.

GNI and GDP

6. Calculate for each country the GDP per capita for each year (2008, 2009, 2010) and, using correlation, compare GDP per capita with GNI per capita. What do you expect to find? What is the explanation of what you actually find?
7. Does GNI per capita vary significantly between the regions of the world?
8. What is the average person's GNI per capita? (NB This is not the average of the rates for countries.)
9. Comparing the average GNI per capita for each year, is there a trend? What is the explanation?
10. Comparing the total GDP for each year, is there a trend? What is the explanation? Compare with **Q9**.

POPULATION

11. Calculate for each country the average annual population growth rate from 2008 to 2010, defined as:

$$\{ (\text{Population 2010} - \text{Population 2008}) / \text{Population 2008} \} \times 50$$

12. Are there significant regional differences in their population growth rates? If so, is there an explanation?
13. Is there a significant correlation between population size (based on 2008) and population growth rate? What is the explanation of your findings?
14. Are there significant differences in the population growth rates depending on the size of the country? If so, is there an explanation?
15. Estimate for the world the annual population growth rate from 2008 to 2010. What factors may limit the accuracy (validity) of this calculation?

DATA SET 11: Internet Users Worldwide

SPSS Data File name: DATA11_Internet_Users_Worldwide.sav

This data file contains internet user statistics for the seven regions of the world for the Years 2000 and 2011.

Cases: 7

Variables: 3: Region of the World, Population of the Region in Year 2011 (millions), Number of Internet Users in the Region in Year 2011 (millions),

Data source: Internet World Stats.

Data sources

We gratefully acknowledge permission to use data from the following sources.

Population of Countries

The World Bank. *Data > Indicators > Population, total.*

<<http://data.worldbank.org/indicator/SP.POP.TOTL>>, [2011], [accessed 04.08.11].

International Monetary Fund. *Data and Statistics > Data > World Economic Outlook Databases (WEO) > By Countries > 1 Select Country Group > 2 Select Country > 3 Select Subjects > People > Population > 4 Select Date Range > Prepare Report.* <<http://www.imf.org/external/index.htm>>, [25 July 2011], [accessed 04.08.11].

Land area of Countries

Internet World Stats. *ALPHABETICAL LIST OF COUNTRIES.*

<<http://www.internetworldstats.com/list2.htm>>, [2011], [accessed 04.08.11]

World Countries and Territories by Region

United Nations Statistics Division. *Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings.*

<unstats.un.org/unsd/methods/m49/m49regin.htm>, [2011], [accessed 04.08.11].

Social Class Population Profiles for UK

businessballs.com. *demographics classifications - free social grade definitions and demographics classifications and geodemographic classes.*

<<http://www.businessballs.com/demographicsclassifications.htm>>, [2011], [accessed 04.08.11]

businessballs.com *CACI ACORN Profile – Population.*

<<http://www.businessballs.com/freespecialresources/acorn-uk-demographics-figures-2010.pdf>>, [27 September 2010], [accessed 04.08.11].

Gross National Income (GNI) and Gross Domestic Product (GDP) for Countries

The World Bank. *Data > Indicators > GDP (current US\$).*

<<http://data.worldbank.org/indicator/NY.GDP.MKTP>>, [2011], [accessed 04.08.11].

The World Bank. *Data > Indicators > GNI per capita, Atlas method (current US\$).*

<<http://data.worldbank.org/indicator/NY.GNP.PCAP.CD>>, [2011], [accessed 04.08.11].

International Monetary Fund. *Data and Statistics > Data > World Economic Outlook Databases (WEO) > By Countries > 1 Select Country Group > 2 Select Country > 3 Select Subjects > National Accounts > Gross domestic product per capita, current prices U.S. dollars > 4 Select Date Range > Prepare Report.* <<http://www.imf.org/external/index.htm>>, [25 July 2011], [accessed 04.08.11].

International Monetary Fund. *Data and Statistics > Data > World Economic Outlook Databases (WEO) > By Countries > 1 Select Country Group > 2 Select Country > 3 Select Subjects > National Accounts > Gross domestic product, current prices U.S. dollars > 4 Select Date Range > Prepare Report.*

<<http://www.imf.org/external/index.htm>>, [25 July 2011], [accessed 04.08.11].

100 top-selling books 'of all time' (1998-2010)

Guardian.co.uk. *DATABLOG 100 top-selling books of all time (Nielsen 1998-2010)*.

<<http://www.guardian.co.uk/news/datablog/2011/jan/01/top-100-books-of-all-time>>, [January 2011], [accessed 04.08.11].

For the actual dataset: < <http://www.guardian.co.uk/news/datablog/2011/jan/01/top-100-books-of-all-time#data>>

Research Councils UK National Survey on Open Access

LISU and SQWconsulting. *Open Access to Research Outputs Annexes Final Report to RCUK September 2008*. <<http://www.rcuk.ac.uk/media/news/2009news/Pages/090422.aspx>>, [23 April 2009], [accessed 04.08.11].

Internet Users and Usage in Europe

Europa Press releases RAPID. *Internet access and use in 2009 - STAT/09/176*.

<<http://europa.eu/rapid/pressReleasesAction.do?reference=STAT/09/176&format=HTML&aged=0&language=EN&guiLanguage=en>>, [8 December 2009], [accessed 04.08.11].

Internet Users and Usage Worldwide

Internet World Stats. *INTERNET USAGE STATISTICS*. <<http://www.internetworldstats.com/stats.htm>>, [31 March 2011, accessed 04.08.11] and [31 Dec 2011, accessed 29.02.12]

Regional and individual country data are available. Updates occur regularly:

Africa:	< http://www.internetworldstats.com/stats1.htm > [31.03.11 and 31.12.11]
Asia:	< http://www.internetworldstats.com/stats3.htm > [30.06.11 and 31.12.11]
Europe:	< http://www.internetworldstats.com/stats4.htm > [31.03.11 and 30.06.11]
Middle East:	< http://www.internetworldstats.com/stats5.htm > [30.06.11 and 31.12.11]
Oceania:	< http://www.internetworldstats.com/stats6.htm > [31.03.11 and 31.12.11]
Latin America/Carib:	< http://www.internetworldstats.com/stats10.htm > [30.06.11 and 31.12.11]
North America:	< http://www.internetworldstats.com/stats14.htm > [31.03.11 and 31.12.11]

Also available other groupings, including:

Americas:	< http://www.internetworldstats.com/stats2.htm >
Language used:	< http://www.internetworldstats.com/stats7.htm >
European Union:	< http://www.internetworldstats.com/stats9.htm >
Caribbean:	< http://www.internetworldstats.com/stats11.htm >
Central America:	< http://www.internetworldstats.com/stats12.htm >
Spanish Speaking:	< http://www.internetworldstats.com/stats13.htm >
South America:	< http://www.internetworldstats.com/stats15.htm >

IT Piracy Rates and Values Worldwide

Business Software Alliance. *Eighth Annual BSA/IDC Global Software Piracy Study May 2011 – Study in Brief*. <http://portal.bsa.org/globalpiracy2010/downloads/study_pdf/2010_BSA_Piracy_Study-InBrief.pdf>, [May 2011], [accessed 04.08.11].

This has annual data for 2006-2010.

Business Software Alliance. *Seventh Annual BSA/IDC Global Software Piracy Study May 2010 – Study in Brief*. <http://portal.bsa.org/globalpiracy2009/studies/09%20Piracy_In%20Brief_A4_111010.pdf>, [May 2010], [accessed 04.08.11].

This has annual data for 2005-2009.

Facebook Users Worldwide

Internet World Stats. *FACEBOOK USERS IN THE WORLD*.

<<http://www.internetworldstats.com/facebook.htm>>, [30 June 2011], [accessed 01.08.11]

On the date accessed (04.08.11) the above led to the following (updates occur regularly):

Africa: <<http://www.internetworldstats.com/africa.htm>> [30 June 2011] - individual countries
 Asia: Middle East: <<http://www.internetworldstats.com/stats3.htm>> [30 June 2011] - list
 Europe: <<http://www.internetworldstats.com/europa2.htm>> [30 June 2011] - individual countries
 Middle East: <<http://www.internetworldstats.com/stats5.htm>> [30 June 2011] - list
 Oceania/Australia: <<http://www.internetworldstats.com/pacific.htm>> [30 June 2011] - individual countries
 The Caribbean: <<http://www.internetworldstats.com/stats11.htm>> [30 June 2011] - list
 Latin America: <<http://www.internetworldstats.com/stats10.htm>> [30 June 2011] - list
 North America: <<http://www.internetworldstats.com/america.htm>> [30 June 2011] - individual countries

Adult Literacy Rates Worldwide

UNESCO Institute of Statistics. *National adult literacy rates (15+), youth literacy rates (15-24), and elderly literacy rates (65+) – Release April 2011*.

<<http://www.uis.unesco.org/Literacy/Pages/default.aspx>>, [2011], [accessed 01.08.11].

Note: Many developed countries no longer collect or publish literacy rate data. For these a rate of 99.0% is assumed.

Additional sources

In a few cases GDP, population and literacy data could not be found from the above sources, for which recourse was made elsewhere:

CIA. *The World Factbook*.

<<https://www.cia.gov/library/publications/the-world-factbook/geography/tw.html>>, [2011], [accessed 04.08.11].

U.S. Census Bureau. *International Programs*.

<<http://www.census.gov/population/international/data/idb/country.php>>, [2011], [accessed 04/08.11].

HOW MANY COUNTRIES ARE THERE?

There is no easy answer to this, it seems. In creating several of the datasets for this guide the issue arose as to what 'countries' to include. On 1st August 2011 there were 193 'countries' - or the equivalent - recognised by the UN. The latest was South Sudan (July 2011). Not on this list are Vatican City, Kosovo and Taiwan, making 196.

There are many disputed territories and very small 'nations'. These have generally been omitted from our datasets. For political reasons the UN is not allowed to refer to Taiwan (despite being autonomous with a population of nearing 24 million!). Data about Taiwan can be hard to come by.

Interestingly, Hong Kong and the much smaller Macao are usually reported separately from China (being SARs – Special Administrative Regions), whereas Taiwan is reported by The International Monetary Fund as 'Taiwan, Province of China' and simply omitted from tables by The United Nations and The World Bank.

For an interesting discussion of what constitutes a country, territory, colony, dependency or other nation group see:

<http://geography.about.co./countries/a/numbercountries.htm>