6. An Introduction to Statistical Package for the Social Sciences

Nick Emtage and Stephen Duthy

This module provides an introduction to statistical analysis, particularly in regard to survey data. Some of the features of the Statistical Package for the Social Sciences (SPSS) are then explained, with reference to a farm forestry survey. Of necessity, this is a brief overview to the highly complex and powerful SPSS package.

1. INTRODUCTION

Computer based statistical packages are an important tool for researchers in the social sciences. The prospect of using statistics is sometimes either repugnant or simply frightening for people, yet most researchers recognise the potential utility of statistical analysis to aid them to describe, analyse, interpret and report their data. The mathematics behind statistical analysis can be daunting for those who have little formal training in either mathematics or the use of statistics. The development of specialist statistical analysis packages has greatly reduced the mathematical challenge of undertaking many analyses. It should be emphasised, however, that these packages have not reduced the need for researchers to understand the assumptions behind statistical analyses, and to be able to interpret their results. The packages have however reduced the need for researchers to be able to undertake many of the calculations that are required for statistical analyses. In this way they allow researchers to concentrate on understanding the assumptions behind the various methods, as well as the potential applications and limitations of various statistical tests.

Statistical software packages have, like other software packages, changed greatly since the advent of the personal computer a little over 20 years ago. Some of the authors still remember programming mainframe computers with paper cards. Holes were punched into the cards and these were then fed into the computer. Computers in those days were scarce, especially the big ones with four megabyte memory! Needless to say that statistical tests were difficult to perform unless the user had an advanced understanding of the mathematics required. More recent computer software packages are reasonably easy to use for people with some familiarity with computers. Most of the packages have features such as drop-down menus, ‘tree’ structure diagrams and on-line help systems. This said, it should be remembered that the packages discussed here are large and highly complex. While they are considerably easier to use today than they were even 10 years ago, like other large software packages, familiarity and ease of use are only developed through practice with the package. A user can become functionally proficient with a package such as Excel and Word after several weeks, use but development of a high level of expertise can take many months or even several years.

When choosing which package to use for statistical analyses a number of factors must be considered. These include the availability of a package, its cost, the functions it can perform, familiarity with the package and the availability of an expert statistician to assist with the analysis process. As discussed above, the packages take time and effort to learn, and many researchers prefer to continue using a particular package once they learned how to use it. Other factors may affect this however. Availability of a package is an important factor in deciding whether to use it or not. If an institution has already obtained the rights to use a particular package, it may be the only choice available. Buying copies of the latest versions of the specialist statistical packages is expensive, as is the cost of maintaining the license to use the package. If an institution already has a package that can provide the functions required then the researcher may be forced to use that package despite preferences for other software because of limited funds.

Where expert statisticians are available to assist with data analyses then the preferred package of the expert is likely to be the one used. As discussed in other modules it is important to discuss research projects with expert statisticians *during their design* to ensure that the data collected will be in a format that allows the use of the desired analysis techniques. It is also important at this stage to discuss the packages available to the researcher and the time available to access a computer for data entry, analysis and reporting. Where access to the computers with the statistical software is limited it may be possible for the researcher to enter the data into a spreadsheet program like Excel and then transfer the data set to the statistics package in order to carry out the analyses. In this case it is important to have some understanding of the formatting required by the statistical package to be used so as to avoid unnecessary reformatting of the data in the statistical package. Where possible data should be entered directly into the statistical package to avoid the potential need to reformat the data.

2. THE STATISTICAL PROGRAM FOR SOCIAL SCIENTISTS (SPSS)

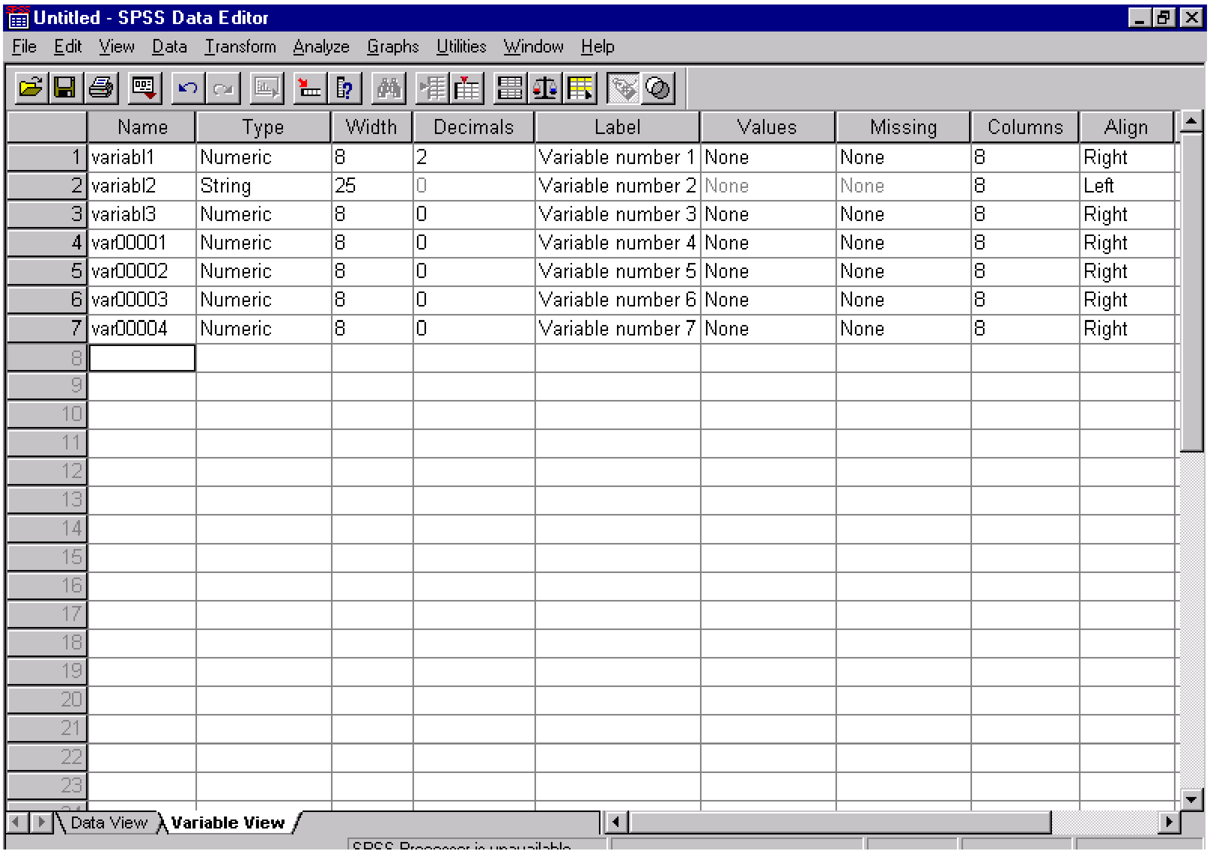
The SPSS Corporation first produced the SPSS software package in the early 1980's and has recently released version 11.0. It is presently one of the most commonly used statistical packages in Australian research institutions and is available at all Australian universities. The advantages of the package are its relative ease of use, its familiarity to many statistical experts and its functionality. One of SPSS's major disadvantages is its cost. The SPSS corporation appears to be progressively breaking up the program into different sections that can be purchased separately. For Australian students an individual users' license (one year) costs approximately $A100 for a 'base' student version and $A350 for a 'graduate pack' licensed for 5 years (as of March 2002). The different versions have varying analytical functions and different capacities in terms of the number of cases and variables that can be used. An institutional license costs even more, depending on the number of expected users. The different packages have licenses that also differ. In most cases licenses are set up to expire automatically after a limited period after which the package can no longer be used. The package is developed for a number of operating systems including Windows and Unix. Information about SPSS products is available on-line at www.SPSS.com.

Organisation of the SPSS package

The set-up of the version 10.0 package (used for illustration here) is organised into two main sections, for defining and entering data and for output. When defining and entering data, users can move between the 'variable' and 'data' 'views' by clicking on the tabs at the bottom of the screen. The third 'output' section opens in a separate window and displays the results of the statistical analyses. The 'output' data are saved as a separate file to the data set.

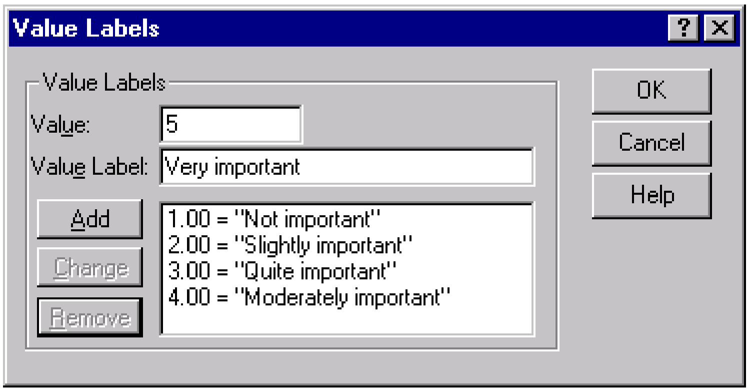
In the 'variable view' (Figure 1) the users sets up the data entry and analysis cells by naming and defining the variables included in the data set. Users are required to use names for the variables of eight or fewer characters. Names must begin with an alphabetic character. Longer descriptions of the variables can be added using the 'Labels' dialog box (Figure 2). A quick way to define the variable format (including the variable type, the number of characters used and labels) if a number of variables have a similar format is to copy the attributes of a variable then paste them into other variable fields.

Once the variables to be recorded have been named and defined the user can access the 'data view' to enter in the values for each variable. The SPSS data view looks similar to a spreadsheet program. The variables are organised as columns with each row as a single 'case' in the data set containing values for the variables relating to that case. It is common practice to use codes to enter data into the package and labels can be used to describe values where needed. For example, codes may be used to record the types of agriculture practiced on a landholding, or respondent's educational levels. The defined labels will appear, by clicking the drop-down list arrow on the right side of the cell, and the user can select the relevant value (Figure 3).


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is unavailable

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	variabl1	Numeric	8	2	Variable number 1	None	None	8	Right
2	variabl2	String	25	0	Variable number 2	None	None	8	Left
3	variabl3	Numeric	8	0	Variable number 3	None	None	8	Right
4	var00001	Numeric	8	0	Variable number 4	None	None	8	Right
5	var00002	Numeric	8	0	Variable number 5	None	None	8	Right
6	var00003	Numeric	8	0	Variable number 6	None	None	8	Right
7	var00004	Numeric	8	0	Variable number 7	None	None	8	Right
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									

Figure 1. Variable view in SPSS 10.0

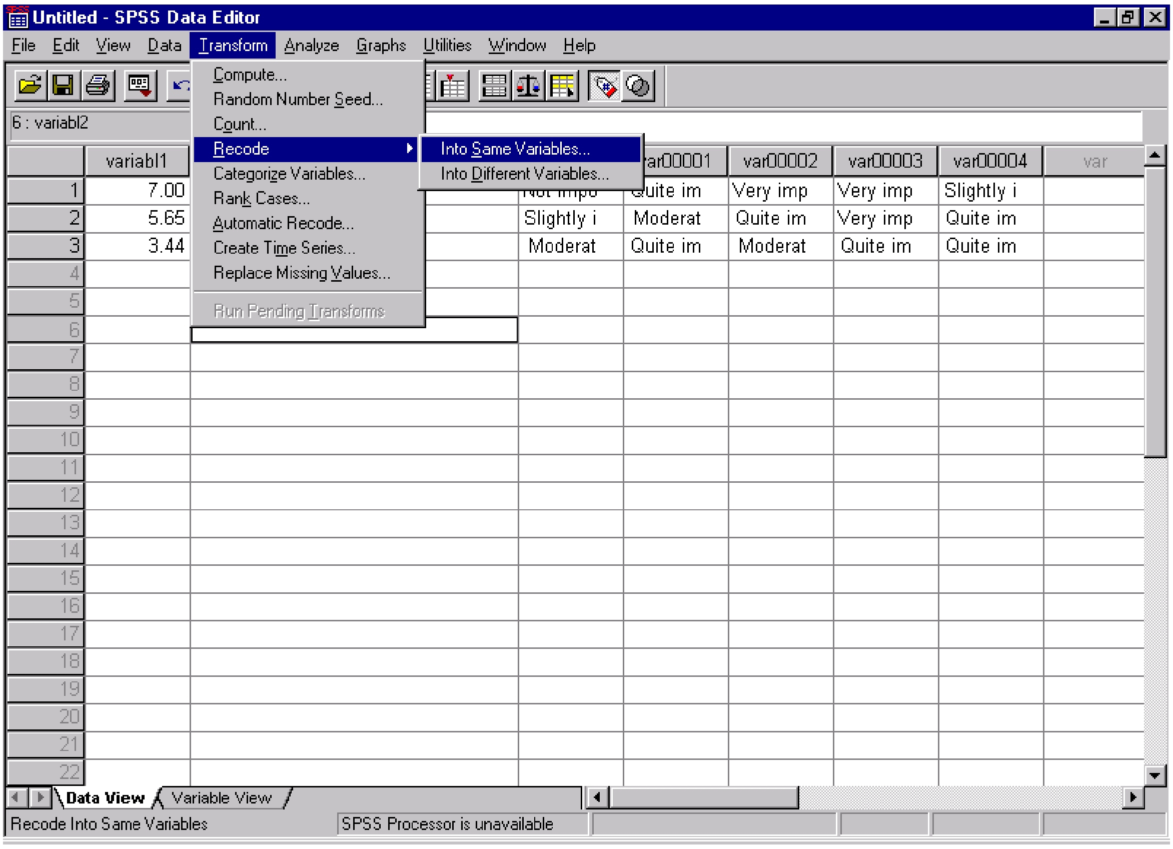


Value: 5

Value Label: Very important

1.00 = "Not important"
2.00 = "Slightly important"
3.00 = "Quite important"
4.00 = "Moderately important"

Figure 2. Defining variable labels using the Value labels dialog box



	variabl1			ar00001	var00002	var00003	var00004	var
1	7.00		Not impo	Quite im	Very imp	Very imp	Slightly i	
2	5.65		Slightly i	Moderat	Quite im	Very imp	Quite im	
3	3.44		Moderat	Quite im	Moderat	Quite im	Quite im	

Figure 3. Entering data into the SPSS

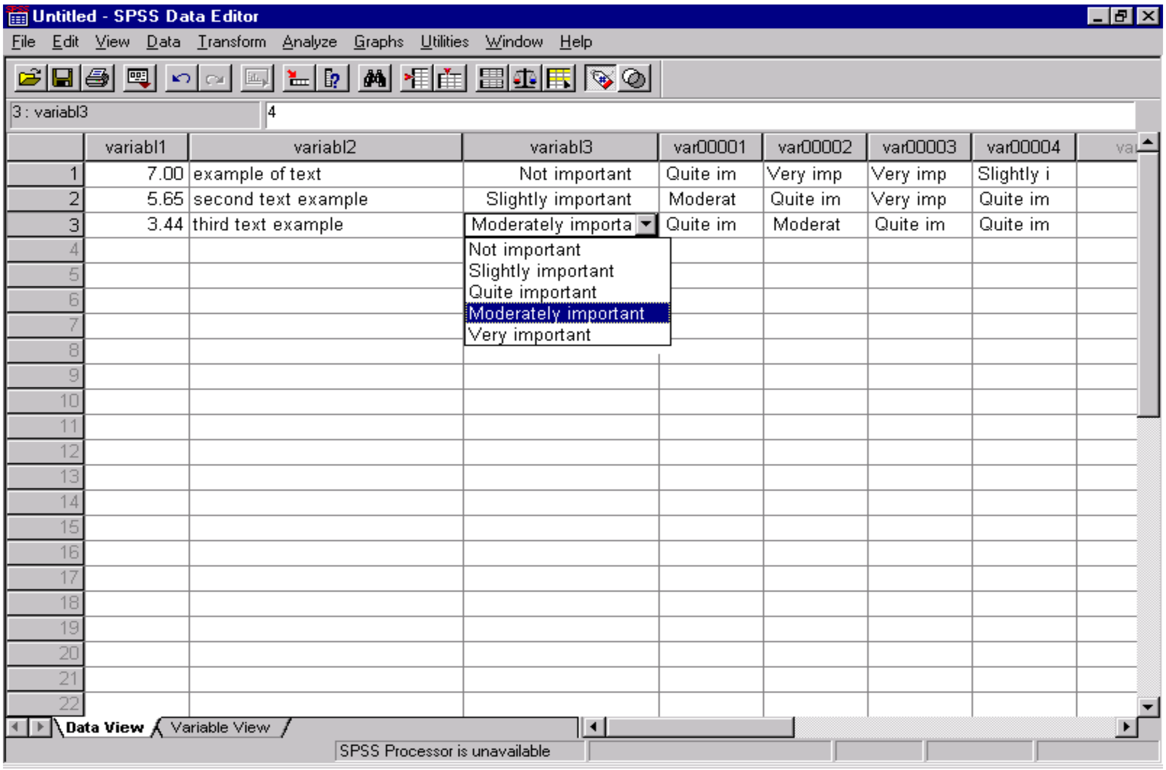
This is handy when there is a large number of possible responses, and thus codes, for a variable, and the user cannot remember all of them. The user can choose to have the codes or the labels displayed in the data view by selecting the 'Value labels' option under the 'View' menu.

Data analysis using SPSS

The SPSS 'student pack' has a wide range of analytical functions, from basic descriptive statistics to advanced general linear modeling capabilities. Specific functions are also included to allow the transformation of variables as preparation for different tests (e.g. for creating standardised or logarithmic values, or the calculation of scales from a number of variables) (Figure 4). The use of these functions allows researchers to calculate quickly new variables based on the values of other variables, test variations in category schemes used to classify responses to 'open ended' questions, and collapse categories where necessary.

Once the data are entered into the SPSS program it is important to check the database for typographic errors that may affect the results of statistical analyses. One means of achieving this is to examine the frequencies of categorical (nominal) data, and descriptive statistics of numeric (ordinal, scale or interval) data. All of the analytical functions available in SPSS can be accessed using the 'Analyse' menu (Figure 6). If the 'Descriptive statistics' then the 'Frequencies' options are selected, the dialog box illustrated in Figure 5 appears. This dialog box enables users to select the variables for which frequencies are computed as well as control the types and, to a limited extent, the formatting of displays of the analyses.

If calculation of descriptive statistics is required, users should select 'Descriptive statistics' and the 'Descriptives' options under the Analysis menu to reveal the 'Descriptives' dialog box (Figure 7).

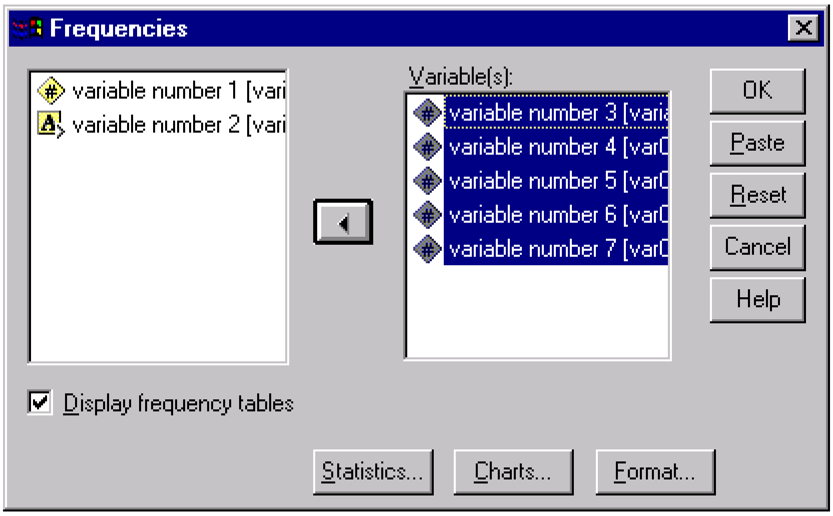


3 : variabl3 4

	variabl1	variabl2	variabl3	var00001	var00002	var00003	var00004	va
1	7.00	example of text	Not important	Quite im	Very imp	Very imp	Slightly i	
2	5.65	second text example	Slightly important	Moderat	Quite im	Very imp	Quite im	
3	3.44	third text example	Moderately importa	Quite im	Moderat	Quite im	Quite im	
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								

SPSS Processor is unavailable

Figure 4. Data transformation functions in SPSS



variable number 1 [vari
variable number 2 [vari

Variable(s):
variable number 3 [varia
variable number 4 [var0
variable number 5 [var0
variable number 6 [var0
variable number 7 [var0

☑ Display frequency tables

Figure 5. Frequencies dialog box

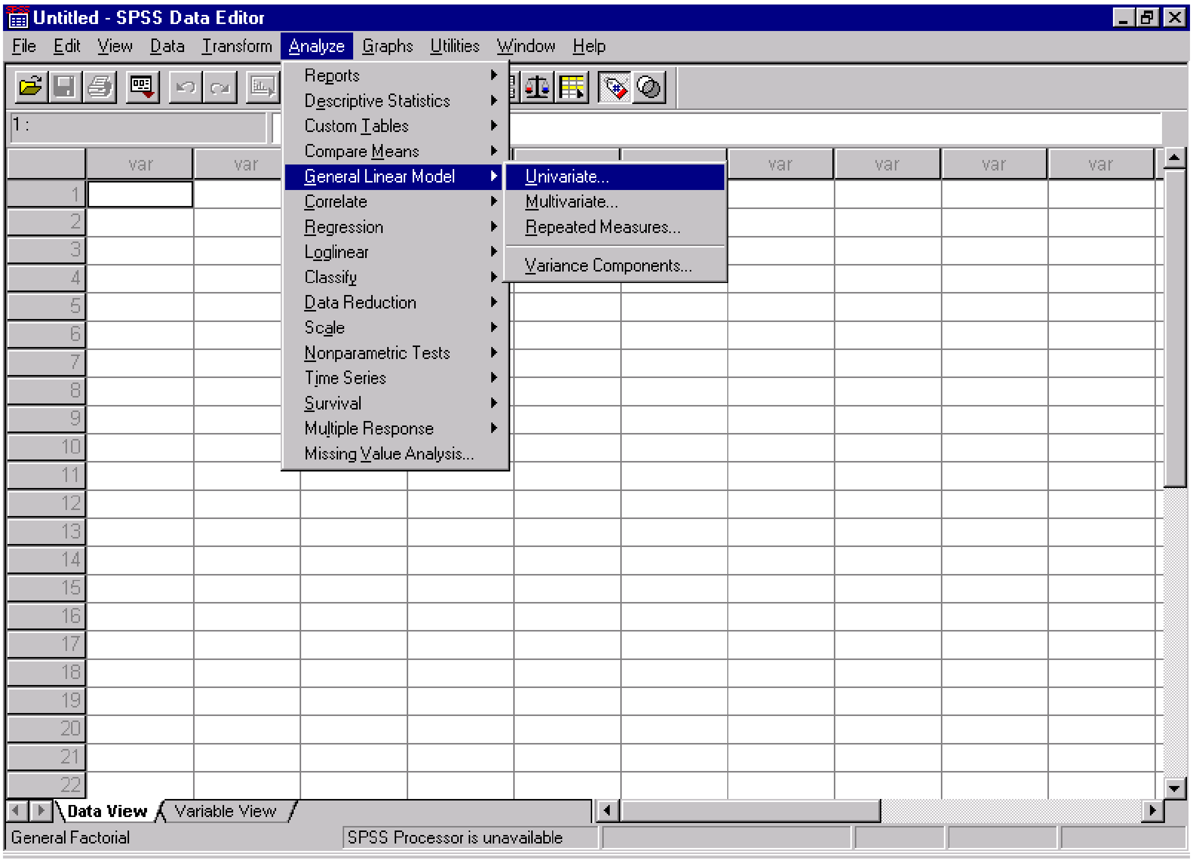


Figure 6. Analysis options available in SPSS

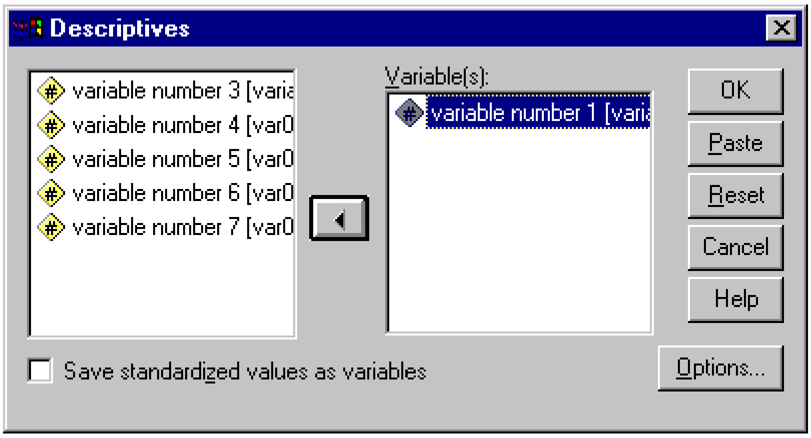


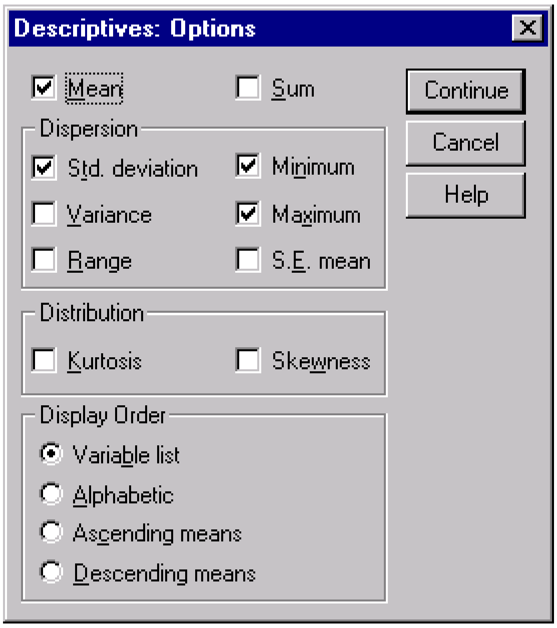
Figure 7. Descriptives function dialog box

Once the ‘Descriptives’ dialog box is shown, the variables to be included in the analyses are selected from the list on the left side of the box (Figure 7), and transferred to the list on the right side of the box (labeled ‘Variables’ in Figure 7) using the arrow in the centre of the box. The types of descriptive statistics that will be Calculated using this function can be selected by clicking on the ‘Options…’ button (Figure 7). This reveals the ‘Options’ dialog box for the Descriptives function (Figure 8).

Other analytical functions included in the SPSS student pack (Version 10) include chi-square tests, correlations, regressions, principal components analyses, ANOVA, cluster analyses, general linear modeling and more.

Whilst this paper does not attempt to provide the reader with statistical skills, the flowchart in Figure 9 may act as a guide for the reader to access quickly those functions in SPSS that will best serve their statistical analysis needs.

The analytical functions are adequate for all but the most advanced researchers or those requiring highly specific analyses. Most advanced or specific applications can be met as well, with SPSS open to manipulation via user compiled ‘Sax Basic’ computer code (also known as ‘scripts’ in SPSS). This is similar to the use of the Visual Basic programming language to develop and execute macros in Microsoft Excel. The Sax Basic language is compatible with Visual Basic for Applications.



Descriptives: Options

☑ Mean □ Sum

Dispersion

☑ Std. deviation ☑ Minimum

□ Variance ☑ Maximum

□ Range □ S.E. mean

Distribution

□ Kurtosis □ Skewness

Display Order

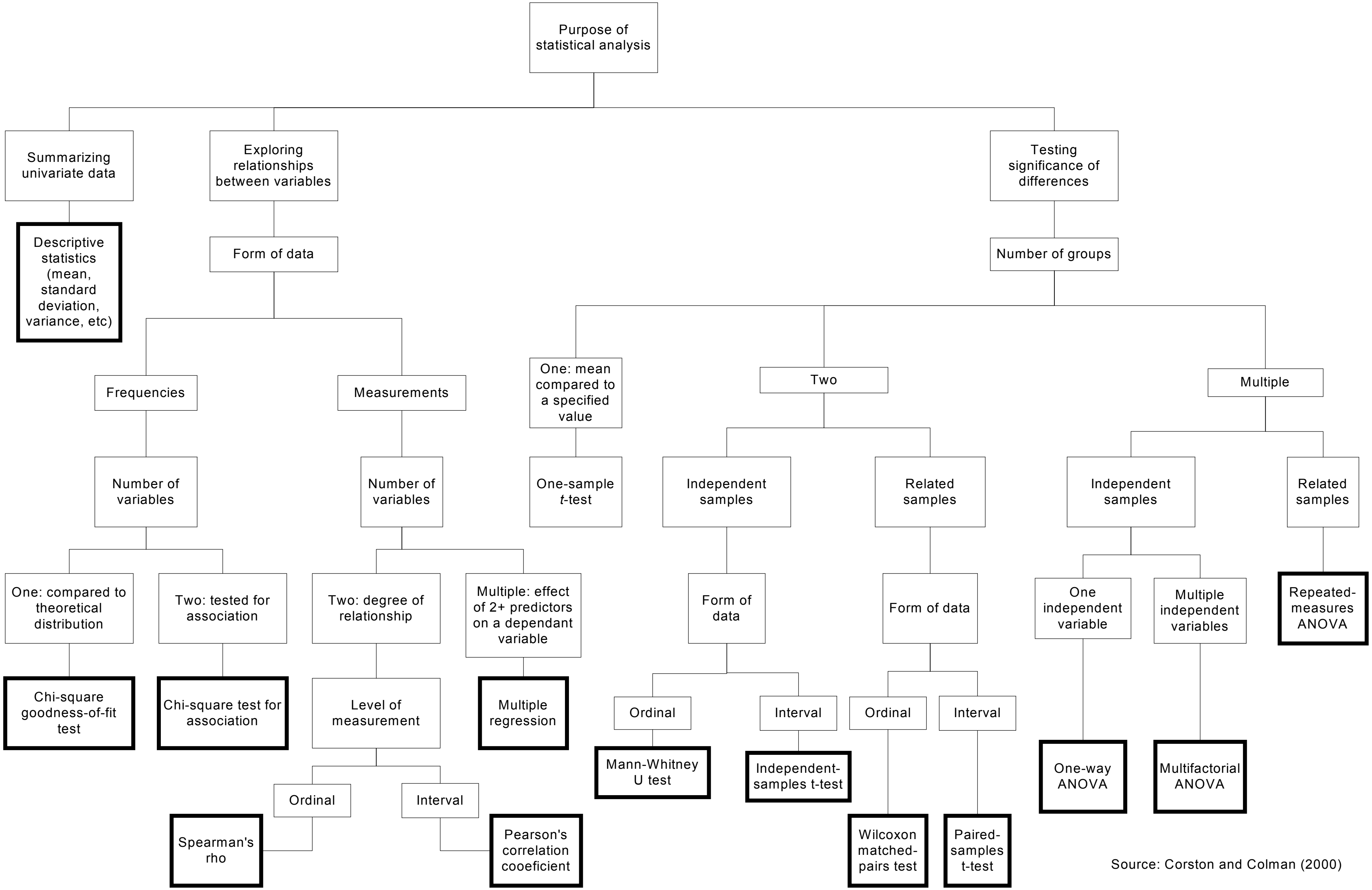
● Variable list

○ Alphabetic

○ Ascending means

○ Descending means

Figure 8. Options dialog box for the ‘Descriptives’ function dialog box


Purpose of statistical analysis
Summarizing univariate data
Descriptive statistics (mean, standard deviation, variance, etc)
Exploring relationships between variables
Form of data
Frequencies
Number of variables
One: compared to theoretical distribution
Chi-square goodness-of-fit test
Two: tested for association
Chi-square test for association
Measurements
Number of variables
Two: degree of relationship
Level of measurement
Ordinal
Spearman's rho
Interval
Pearson's correlation cooeficient
Multiple: effect of 2+ predictors on a dependant variable
Multiple regression
Testing significance of differences
Number of groups
One: mean compared to a specified value
One-sample *t*-test
Two
Independent samples
Form of data
Ordinal
Mann-Whitney U test
Interval
Independent-samples t-test
Related samples
Form of data
Ordinal
Wilcoxon matched-pairs test
Interval
Paired-samples t-test
Multiple
Independent samples
One independent variable
One-way ANOVA
Multiple independent variables
Multifactorial ANOVA
Related samples
Repeated-measures ANOVA

Source: Corston and Colman (2000)

Figure 9. Choosing an appropriate statistical procedure

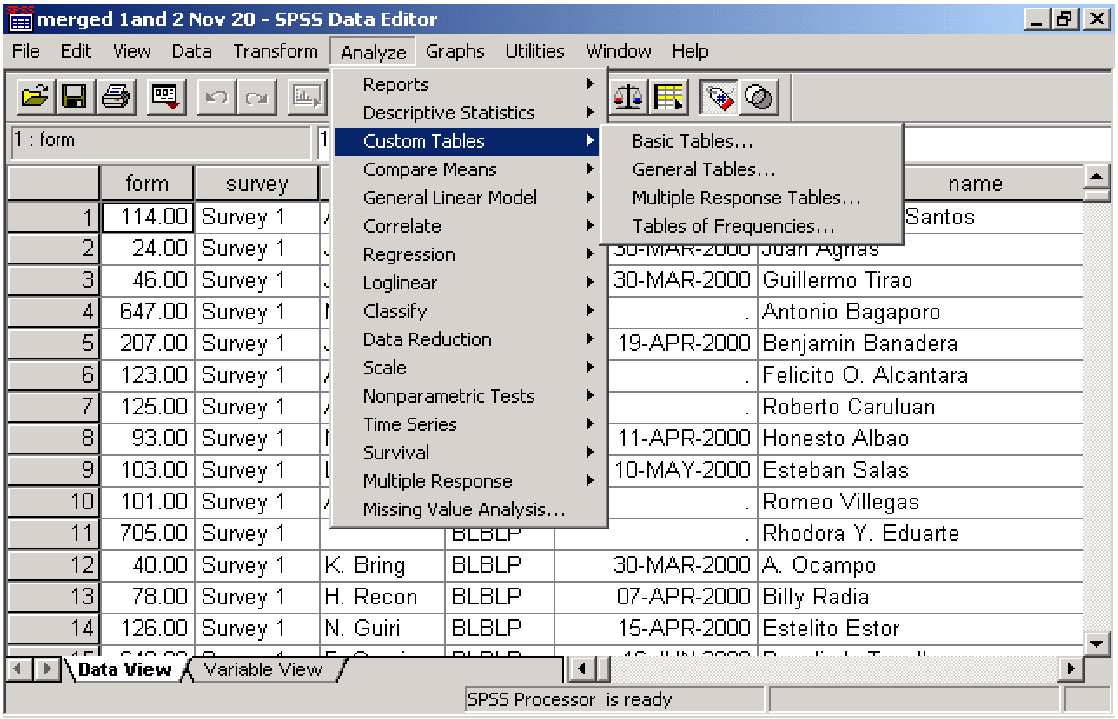
Using SPSS to Describe Data

Whilst computer-based statistical packages provide a high degree of functionality with regard to data analysis, they also provide a number of highly useful tools for the description and presentation of summaries of the dataset.

These functions include Descriptives and Frequencies as explained earlier and Crosstabs, also found under the Descriptive Statistics menu, and Basic Tables, General Tables, Multiple Response Tables and Tables of Frequencies all located under the Custom Tables menu item (Figure 10). It is often useful to undertake one or more of these processes before commencing data analysis to identify any weaknesses in the dataset such as poorly represented groups within the sample that may limit the statistical validity of some forms of analysis. Crosstabs are also an efficient way of presenting data summaries in research and project reports.

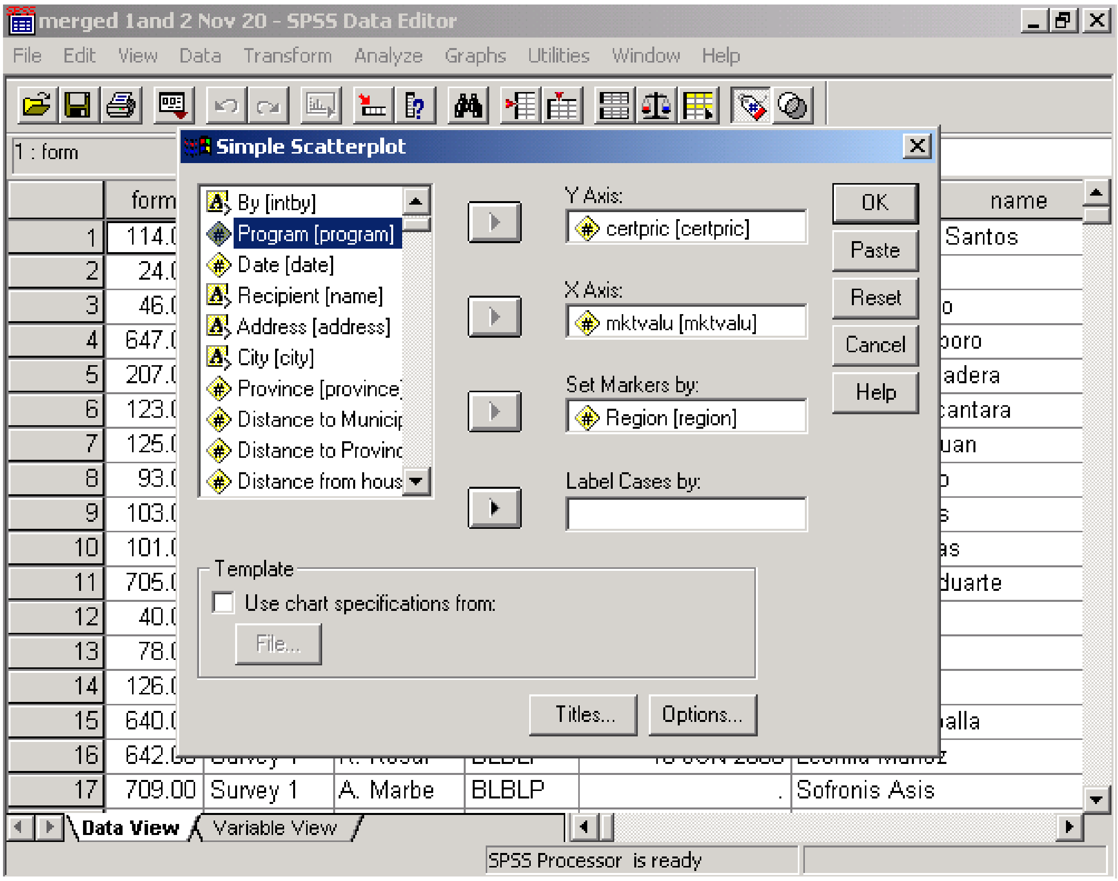
The charting functions available in SPSS also provide a number of techniques for the initial exploration and the presentation of data. Scatter Plots (Figures 11 and 12) can be used to identify quickly the presence and nature of any correlations between variables while Histograms (Figures 13 and 14) can be used to present a graphical representation of the shape of the distribution of the data for important variables.

There is a reasonable amount of literature available to assist users of the SPSS package produced by the SPSS Corporation and by independent authors. The tutorial and help facilities for the package are comprehensive, generally easy to understand and include the on-line Statistics Coach and Syntax Guide.



	form	survey				name
1	114.00	Survey 1				Santos
2	24.00	Survey 1			30-MAR-2000	Juan Agnas
3	46.00	Survey 1			30-MAR-2000	Guillermo Tirao
4	647.00	Survey 1			.	Antonio Bagaporo
5	207.00	Survey 1			19-APR-2000	Benjamin Banadera
6	123.00	Survey 1			.	Felicito O. Alcantara
7	125.00	Survey 1			.	Roberto Caruluan
8	93.00	Survey 1			11-APR-2000	Honesto Albao
9	103.00	Survey 1			10-MAY-2000	Esteban Salas
10	101.00	Survey 1			.	Romeo Villegas
11	705.00	Survey 1		BLBLP	.	Rhodora Y. Eduarte
12	40.00	Survey 1	K. Bring	BLBLP	30-MAR-2000	A. Ocampo
13	78.00	Survey 1	H. Recon	BLBLP	07-APR-2000	Billy Radia
14	126.00	Survey 1	N. Guiri	BLBLP	15-APR-2000	Estelito Estor

Figure 10. Custom Tables drop-down menu selections



Y Axis: certpric [certpric]

X Axis: mktvalu [mktvalu]

Set Markers by: Region [region]

Label Cases by:

Template

□ Use chart specifications from:

Figure 11. Design options for a Simple Scatterplot

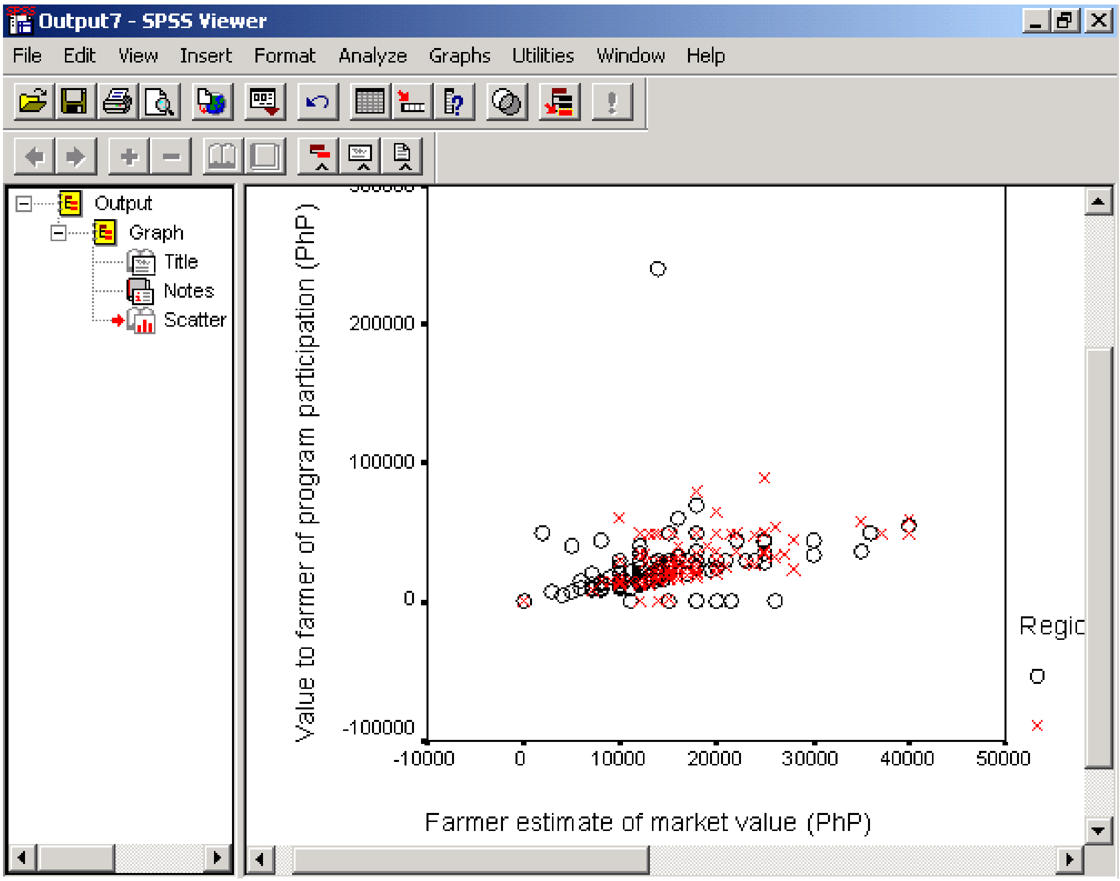
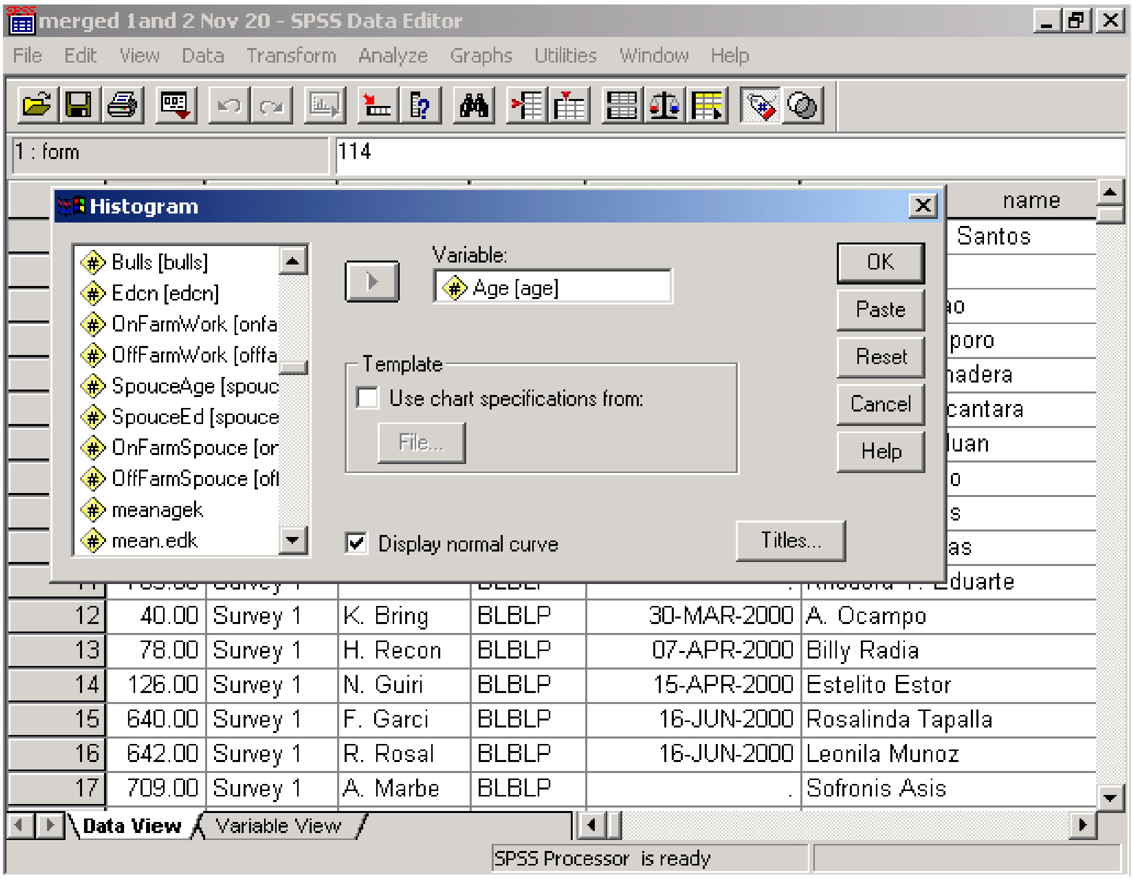

Output7 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Output
Graph
Title
Notes
Scatter
Value to farmer of program participation (PhP)
200000
100000
0
-100000
-10000
0
10000
20000
30000
40000
50000
Farmer estimate of market value (PhP)
Regic
○
×

Figure 12. Simple Scatterplot displayed in the Output Viewer



12	40.00	Survey 1	K. Bring	BLBLP	30-MAR-2000	A. Ocampo
13	78.00	Survey 1	H. Recon	BLBLP	07-APR-2000	Billy Radia
14	126.00	Survey 1	N. Guiri	BLBLP	15-APR-2000	Estelito Estor
15	640.00	Survey 1	F. Garci	BLBLP	16-JUN-2000	Rosalinda Tapalla
16	642.00	Survey 1	R. Rosal	BLBLP	16-JUN-2000	Leonila Munoz
17	709.00	Survey 1	A. Marbe	BLBLP	.	Sofronis Asis

Figure 13. Design options for a histogram

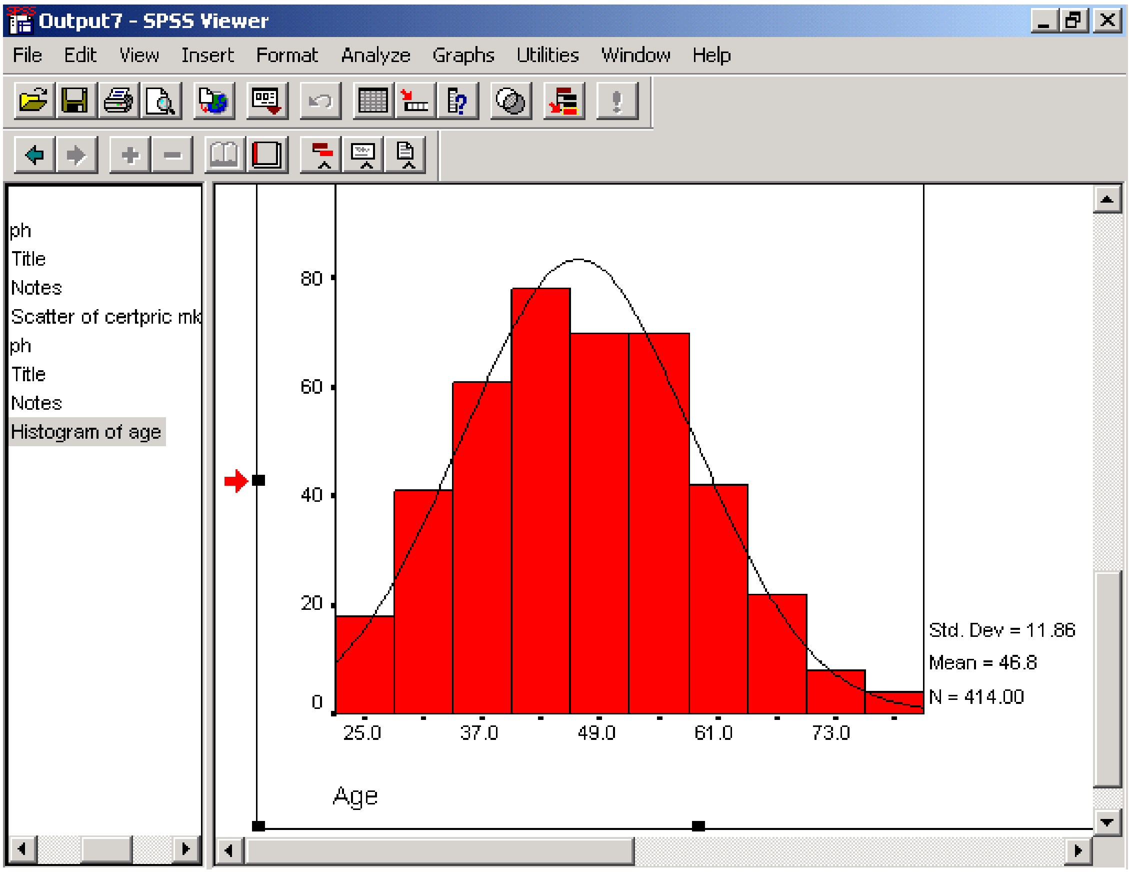


Figure 14. Histogram displayed in the Output Viewer

3. CONCLUDING COMMENTS

Researchers frequently collect large quantities of data, from surveys, experiments and other forms of observation. A statistical computing package provides a convenient means to store these data, and derive descriptive and inferential statistics. The Statistical Package for the Social Sciences (SPSS) is a widely used general-purpose survey analysis package, and hence a useful one to master. It is necessary to allow some learning time to become familiar with this package, and annual license fees can be a disincentive.

REFERENCES

Corston, R. and Colman, A. (2000), *A Crash Course in SPSS for Windows*, Blackwell, Oxford.