University of BRISTOL

IT SERVICES

Introduction to SPSS (version 23) for Windows

Practical workbook

Aims and Learning Objectives

By the end of this course you will be able to:

- prepare data for SPSS
- manipulate data using recode, compute, select and split
- create and run SPSS programs to do simple statistical data analysis.

This includes being able to:

- create frequency tables
- produce graphs
- use crosstabulation, correlation, t-test and non-parametric tests
- move data from other applications into SPSS
- export SPSS output for presentations and publications.

Document information

Course files

This document and any associated practice files (if needed) are available on the web. To find these, go to www.bristol.ac.uk/is/learning/resources and in the **Keyword** box, type the document code given in brackets at the top of this page.

Related documentation

Other related documents are available from the web at:

http://www.bristol.ac.uk/is/learning/resources

There is no introduction to STATA documentation produced by Information Services, but there are short courses on STATA available within the University of Bristol provided by the Department of Social Medicine.

Introduction

SPSS provides facilities for analysing and displaying information using a variety of techniques. This document uses version 23 of SPSS for Windows.

Prerequisites

Basic familiarity with Windows and at least an elementary knowledge of simple statistics such as t-tests, chi squared tests, p values and confidence intervals would be useful (statistical theory is not taught on this course).

Contents

Preparing your data for SPSS

1. **Entering data straight into SPSS - See section 5**

   This method is not a good way to enter a large amount of data, as you will find yourself typing ahead of yourself and making mistakes. Instead consider using Access or Excel.

2. **Data from an Access database - See help files on Access**

   Click on the relevant table of data and export it into an excel file.

3. **Data from an Excel workbook**

   SPSS will read data from an excel workbook, it allows you to specify which worksheet it should use and what range of row and columns it should use also.

Basic rules to allow Excel data to be read into SPSS

Structure

Subjects or samples go down the left side of the page, things that you have recorded about the subjects or sample go across the top

- Give all the subjects/samples a unique id number
- Only use one row of column headings
- Make the column headings unique
- If possible make the column headings unique within the first 8 digits
- Put all the rows of data onto the same spreadsheet. If you have several sets of data with the same column in each and which you intend to analyse together, for example pupils from different schools, include a column which says which data set they come from, rather than keep them as separate spreadsheets.
- Try not to leave blank rows and columns.

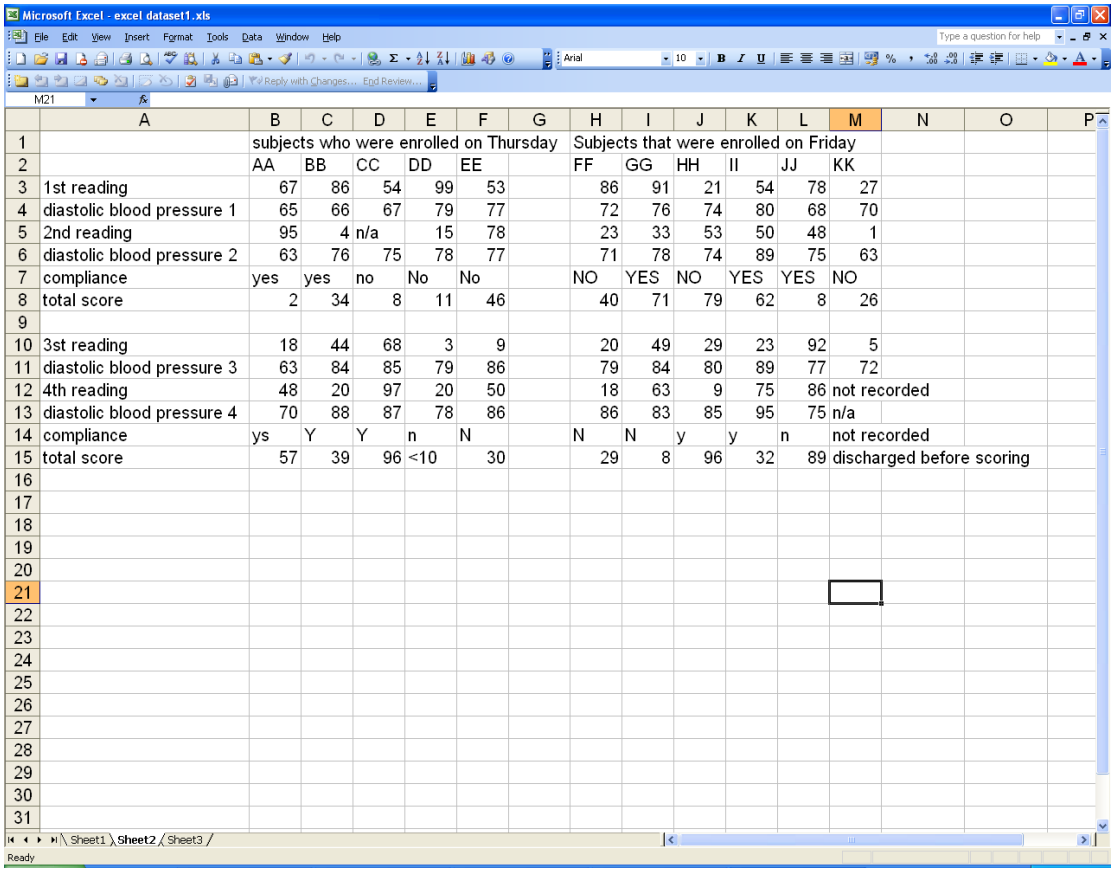
Content

- SPSS can read dates in Excel but only if Excel knows they are dates – check this is the case using the format cell feature in Excel.
- If you are entering numbers such as weight, make sure all the values are numbers. Remove ***?***, ***n/a***, ***<10***, etc.
- If the value is missing, either code with a missing value such as 9, or 99 etc, or leave with out entering anything. Do not enter a space or a fullstop.
- If you are entering text such as ***yes*** or ***no***, check that you have always spelt this in the same way. ie. ***Y*** will be seen by SPSS as different from ***y*** and different from "***Y*** " and "***YES***" and "***yes*** and ***Yes***.
- Do not merge cells.
- Do not sort data without taking steps to ensure that all the data is sorted together.
- Avoid putting totals, means, counts or graphs on the same spreadsheet

Task 1 Practical exercise in Data Preparation

Suppose this was a data set that your research assistant had prepared for you.

Can you convert this into a usable SPSS data set?

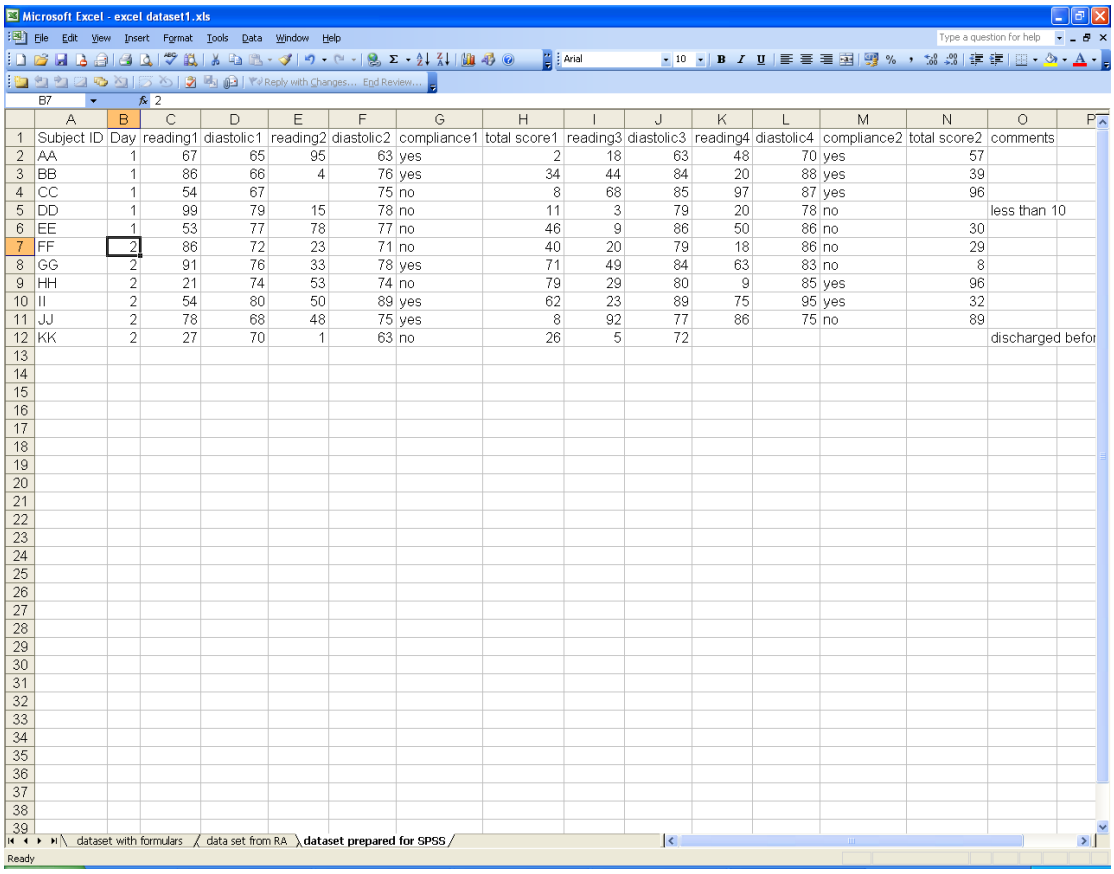


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		subjects who were enrolled on Thursday						Subjects that were enrolled on Friday								
2		AA	BB	CC	DD	EE		FF	GG	HH	II	JJ	KK			
3	1st reading	67	86	54	99	53		86	91	21	54	78	27			
4	diastolic blood pressure 1	65	66	67	79	77		72	76	74	80	68	70			
5	2nd reading	95	4	n/a	15	78		23	33	53	50	48	1			
6	diastolic blood pressure 2	63	76	75	78	77		71	78	74	89	75	63			
7	compliance	yes	yes	no	No	No		NO	YES	NO	YES	YES	NO			
8	total score	2	34	8	11	46		40	71	79	62	8	26			
9																
10	3st reading	18	44	68	3	9		20	49	29	23	92	5			
11	diastolic blood pressure 3	63	84	85	79	86		79	84	80	89	77	72			
12	4th reading	48	20	97	20	50		18	63	9	75	86	not recorded			
13	diastolic blood pressure 4	70	88	87	78	86		86	83	85	95	75	n/a			
14	compliance	ys	Y	Y	n	N		N	N	y	y	n	not recorded			
15	total score	57	39	96	<10	30		29	8	96	32	89	discharged before scoring			
16																
17																
18																
19																
20																
21																
22																
23																
24																
25																
26																
27																
28																
29																
30																
31																

Figure 1 - Example data set

Think about how you need to alter the structure and content of the data using the rules in section 1(3).

The data set that you produce should look something like this:



Microsoft Excel - excel dataset1.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Subject ID	Day	reading1	diastolic1	reading2	diastolic2	compliance1	total score1	reading3	diastolic3	reading4	diastolic4	compliance2	total score2	comments
2	AA	1	67	65	95	63	yes	2	18	63	48	70	yes	57	
3	BB	1	86	66	4	76	yes	34	44	84	20	88	yes	39	
4	CC	1	54	67		75	no	8	68	85	97	87	yes	96	
5	DD	1	99	79	15	78	no	11	3	79	20	78	no		less than 10
6	EE	1	53	77	78	77	no	46	9	86	50	86	no	30	
7	FF	2	86	72	23	71	no	40	20	79	18	86	no	29	
8	GG	2	91	76	33	78	yes	71	49	84	63	83	no	8	
9	HH	2	21	74	53	74	no	79	29	80	9	85	yes	96	
10	II	2	54	80	50	89	yes	62	23	89	75	95	yes	32	
11	JJ	2	78	68	48	75	yes	8	92	77	86	75	no	89	
12	KK	2	27	70	1	63	no	26	5	72					discharged befor

Figure 2 - Formatted data set

This has been achieved using copy and past special (transpose) and by removing text from the columns and giving the columns easy to read unique labels. The problem with two sections of data, one recorded on one day and the other the next day has been sorted out by using a column labelled day.

There is still one problem of a data point which was labelled less than 10. If you can find out what the threshold value of the assay is, you can sometimes substitute half the threshold value, or ask them to conduct the assay again so that the true value can be ascertained. Less than 10 is not a number SPSS will analyse.

Other data formats:

SPSS will read many types of data set including STATA data files (*.dta), Systat data files (*.sys), SPSS portable files (*.por), lotus files (*.w*), SAS data files, and even text files with delimited or fixed formatted data.

There is sometimes a problem with a new version of software, such that SPSS might not be able to recognise the new format of a data set. In this case it is usually possible to resave it in a previous version of the package so that SPSS can recognise it.

Task 2 Starting SPSS

2.1 To open SPSS:

- From the **Start** menu go to **All Programs** and select **IBM SPSS 23.0** followed by **IBM statistics SPSS 23.0**.

  Initially the **SPSS Data Editor** window opens up with a queries window superimposed (Figure 3).

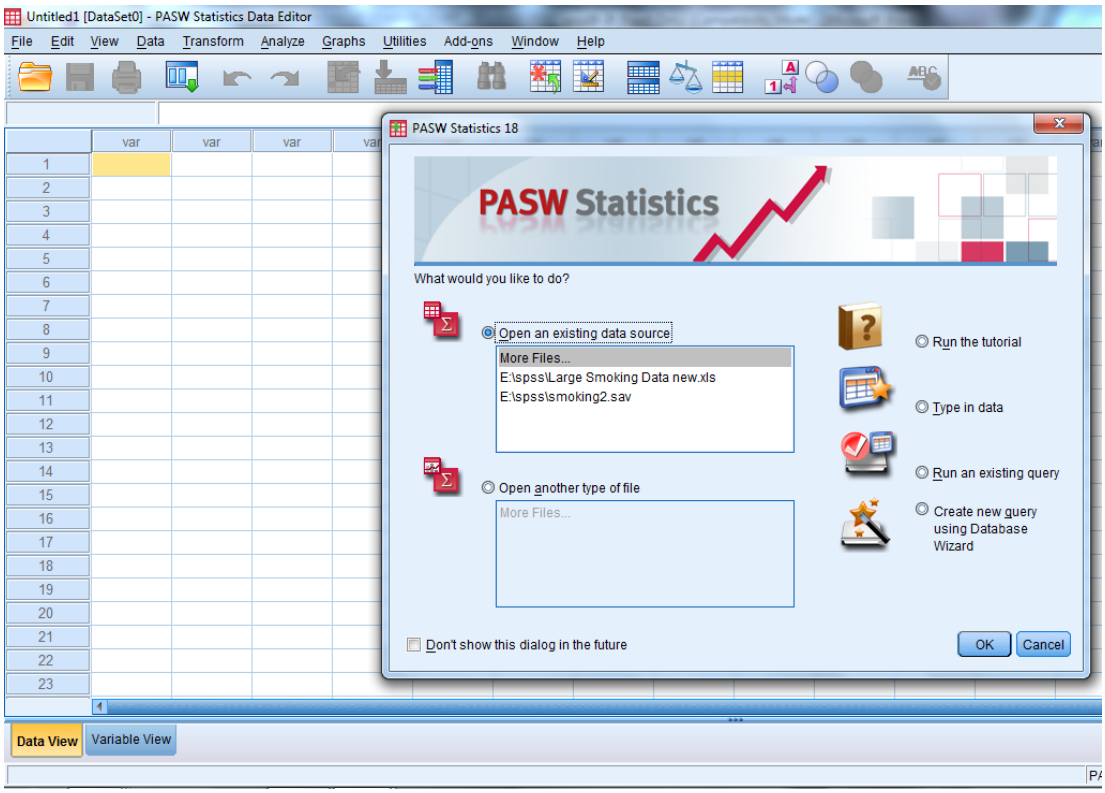


Figure 3 – the data view of the data editor window

- Select **Cancel** or **Type in data** to this query window. The query window then closes.

Tips

- When you have been using SPSS at your own PC in the recent past, it will give you the option of opening up the last few files that you accessed in SPSS just like the list available in most Microsoft packages

Task 3 Example dataset

Comments This document uses an example survey and considers some of the analyses that might be carried out on the data. An SPSS dataset is made up of a number of observations, each of which contains a value for each variable in the dataset. You will be instructed how to create the SPSS dataset for the example survey in due course but first you must understand the layout and meaning of the survey questionnaire.

3.1 Imagine that you interviewed some people on their smoking habits using the questions shown below:

Reference number ☐☐☐

University of BRISTOL

Smoking Questionnaire

1. How old were you on your last birthday? ☐☐
2. Indicate your sex — Male ☐ Female ☐
3. Do you smoke at all? — Yes ☐$_1$ No ☐$_2$
   If No, go to question 9
4. Do you smoke cigarettes? — Yes ☐$_1$ No ☐$_2$
   If No, go to question 6
5. On average how many cigarettes a day would you say that you smoked? ☐☐☐
6. Do you smoke a pipe? — Yes ☐$_1$ No ☐$_2$
7. Do you smoke cigars? — Yes ☐$_1$ No ☐$_2$
8. Have you ever tried to give up smoking? Yes ☐$_1$ No ☐$_2$
9. Tell me what you think on each of the following three statements:-

   Tax on tobacco is too high

   strongly disagree ☐$_1$ disagree ☐$_2$ agree ☐$_3$ strongly agree ☐$_4$

   Smoking is dangerous to your health

   strongly disagree ☐$_1$ disagree ☐$_2$ agree ☐$_3$ strongly agree ☐$_4$

   Smoking should not be allowed in cinemas

   strongly disagree ☐$_1$ disagree ☐$_2$ agree ☐$_3$ strongly agree ☐$_4$

10. Date questionnaire completed ☐☐/☐☐/☐☐☐☐
11. Date thesis submitted ☐☐/☐☐/☐☐☐☐

From 10 completed questionnaires it is possible to create a dataset like the one shown below.

This dataset has 10 observations and 12 variables. The data is in fixed column format; each measurement forms a column and the **values** in each column make up a **variable**. Note that blanks indicate missing values. Each of the items recorded - **Age**, **Sex** and so on - are data values. For **Sex**, **M** and **F** are used to denote male and female, rather than using numeric codes. All the information about a single person makes up one **observation**.

Ref Number	Age	Sex	Smoke	Smoke Cigs	How many	Pipe	Cigars	Give Up	Tax	Health	Cine ma	Date survey	Date submit
1	27	F	1	1	10	2	2	1	3	3	3	02/11/1995	25/09/1995
2	31	M	2						4	2	1	16/11/1995	04/09/1995
3	35	M	2						4	1	1	01/12/1995	21/07/1995
4	58	M	2						3	1	2	17/11/1995	05/09/1995
5	56	M	2						4	3	2	10/11/1995	30/09/1995
6	25	F	1	1	20	2	2	2	3	4	4	02/12/1995	10/08/1995
7	41	F	1	1	30	2	1	1	3	1	3	17/11/1995	24/08/1995
8	38	F	1	1	999	2	2	1	4	4	4	02/12/1995	30/08/1995
9	43	F	1	2		2	1	1	4	2	2	16/11/1995	01/08/1995
10	29	M	1	1	40	2	2	2	2	4	4	02/12/1995	26/09/1995

Table 1 - example data set

From this dataset you can see that the reference number of the first subject is 1, she is 27 and is female. She smokes, and smokes cigarettes, on average 10 per day. She doesn't smoke a pipe or cigars and has tried to give up. She agrees that tax on tobacco is too high and that smoking is dangerous to your health and that smoking should be banned from cinemas.

Later you will be asked to enter these values but first you are asked to consider how variables are defined and what attributes they can have.

Each variable needs to be given a variable name that is used in describing the variable to SPSS. Table 2 lists the names that are to be used in the example and specifies the order in which they are to be given to SPSS. It also suggests suitable labels that can later be associated with the variables to clarify output.

ref_no	Reference number
age	Age last birthday
sex	Sex of respondent
smoker	Do you smoke?
cigs	Do you smoke cigarettes?
num_cigs	How many cigarettes per day?
pipe	Do you smoke a pipe?
cigars	Do you smoke cigars?
give_up	Have you tried to give up smoking?
tax	Do you think tax on tobacco is too high?
danger	Do you think smoking is dangerous to your health?
cinemas	Do you think smoking should be allowed in cinemas?
Date_survey	Date the survey was completed
Date_submit	Date the subject submitted their thesis

Table 2 – labels given to each variable to clarify output

You should now have all the information necessary to start creating the example dataset in SPSS. First you will use the **Variable View** of the **Data Editor** window to specify the variable names and their attributes.

You should note that this data was collected in 1995 before any of the present day legislation on smoking in public had been introduced.

Task 4 The Data Editor

The **Data Editor** is a worksheet used for entering and editing data. It has two panes, the **Data View** and the **Variable View**. You switch between each by clicking on the tab at the bottom left of the **Data Editor** window.

In general, to enter data in a cell in the worksheet you can use the cursor to click on the desired cell, type in the entry and finish by using the **ENTER** key. If you make a mistake you can overtype all the contents of a cell by a similar process.

If you want to correct just part of a cell then double-click on the cell, then use the direction keys to locate the part you want to change. Correct the error using the **BACKSPACE** key and type the correction, and then use the **ENTER** key to complete the edit. Some cells in the **Variable View** window are limited to specific values. When you click on such a cell a square grey icon appears on the right of the cell. Clicking on this brings up a window that allows values to be entered and changed from a predefined list.

Task 5 Naming and defining variables

5.1 To select the **Variable View** of the **Data Editor** window:

- Click on the **VARIABLE VIEW** tab in the bottom left hand corner of the screen (Figure 4).

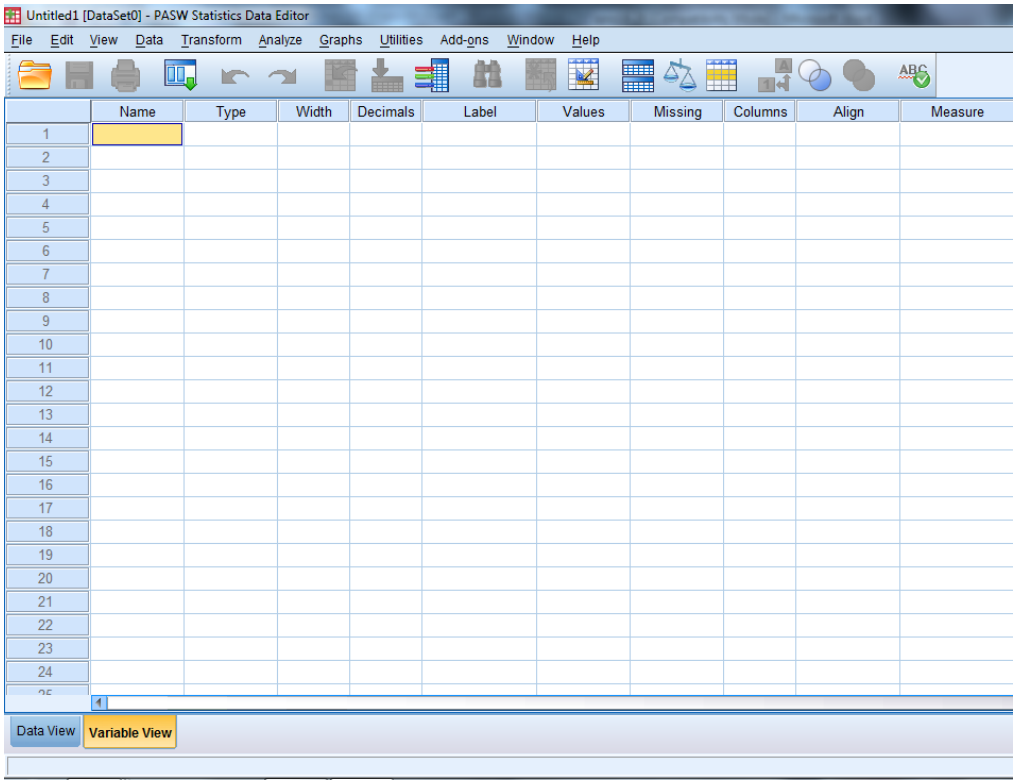


Figure 4 - the Variable View of the Data Editor window

Note This window already has a defined structure. There are eleven columns headed: - **Name**, **Type**, **Width**, **Decimals**, **Label**, **Values**, **Missing**, **Columns**, **Align, Measure** and **Role**.

Each of these headings is used to indicate some facet of the definition of each variable. Their use is described as we proceed to develop our sample dataset.

Variable names

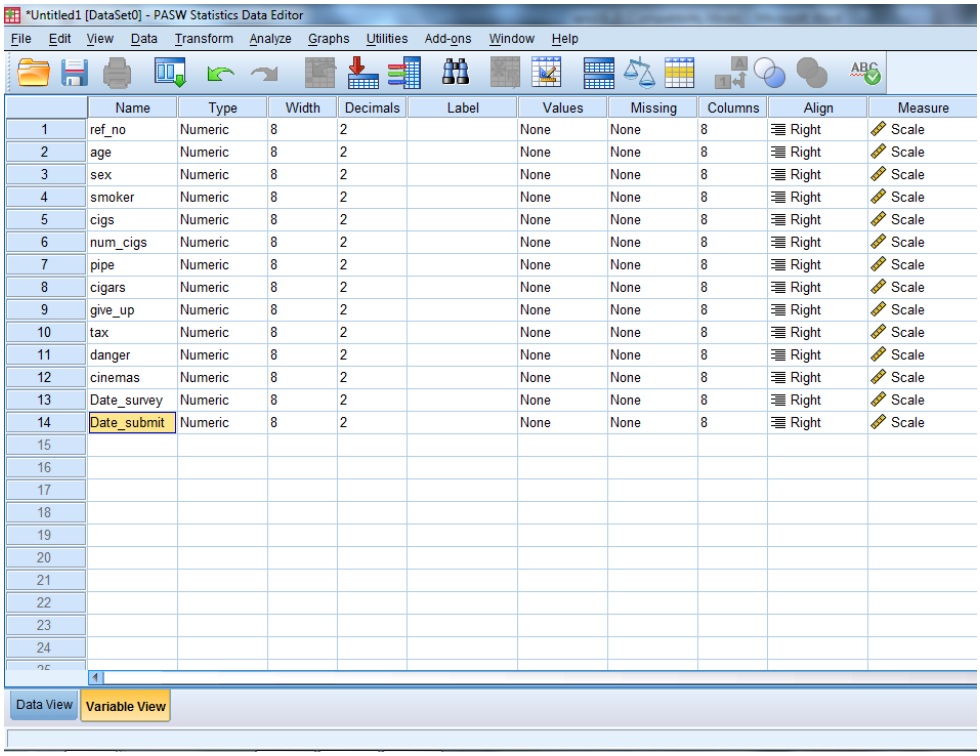
The rules for names are:

- the name must begin with a letter. The remaining characters can be any letter, any digit, a period, or the symbols @, #, _, or $;
- variable names cannot end with a period
- variable names that end with an underscore should be avoided
- the length of the name cannot exceed 64 bytes. Sixty-four bytes typically means 64 characters in single-byte languages (eg, English, French, German) and 32 characters in double-byte languages (eg, Japanese, Chinese, Korean).
- spaces and special characters (eg !, ?, ', and *) cannot be used;
- each variable name must be unique; duplication is not allowed;

- if SPSS comes across a duplication when reading a data set prepared in another package it will create a new variable name such as var0004;
- the underscore character is frequently used where a space is desired in names;
- SPSS may miss out spaces when reading in names from another package such as Excel, ie age group would become agegroup;
- Reserved keywords cannot be used as variable names – these are ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO and WITH;
- A mixture of uppercase and lowercase characters, and case is preserved for display.

5.2 Point and click on the cell in row **1** and column **1**. Type ***ref_no*** in this cell. Use the down arrow to move to row **2** column **1**.

- Type ***age*** in this cell. Use the down arrow to move to row **3** column **1**.
- Continue with this process until all 14 variable names from Table 2 are entered in this first column. Your screen should then look like Figure 5.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	ref_no	Numeric	8	2		None	None	8	Right	Scale
2	age	Numeric	8	2		None	None	8	Right	Scale
3	sex	Numeric	8	2		None	None	8	Right	Scale
4	smoker	Numeric	8	2		None	None	8	Right	Scale
5	cigs	Numeric	8	2		None	None	8	Right	Scale
6	num_cigs	Numeric	8	2		None	None	8	Right	Scale
7	pipe	Numeric	8	2		None	None	8	Right	Scale
8	cigars	Numeric	8	2		None	None	8	Right	Scale
9	give_up	Numeric	8	2		None	None	8	Right	Scale
10	tax	Numeric	8	2		None	None	8	Right	Scale
11	danger	Numeric	8	2		None	None	8	Right	Scale
12	cinemas	Numeric	8	2		None	None	8	Right	Scale
13	Date_survey	Numeric	8	2		None	None	8	Right	Scale
14	Date_submit	Numeric	8	2		None	None	8	Right	Scale
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										

Figure 5 - table with variable names in first column

Note When the down arrow is pressed, the 10 columns to the right of the first column fill with values (default values). We can examine these ten columns and if necessary change values in specific cells to reflect properties of our example dataset.

Variable types

The **Type** column is showing **Numeric** for all rows. This means that numeric (number) values will be expected in the dataset relating to these variables. This is correct for all the variables except sex where we have collected data in the form of F for female and M for male, and age where we would like SPSS to record this as a date.

5.3 To enter the data as M or F, change the variable type:

- Point and click on the grey box on the right of the cell giving **Type** for the variable **sex**.

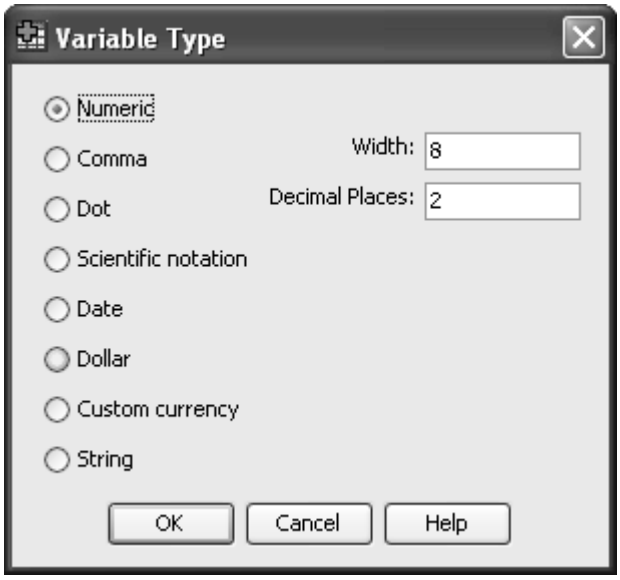


Figure 6 - Variable Type options box

A dropdown menu appears offering eight data types (Figure 6). The common items are **Numeric**, **Date**, **Custom currency** and **String**. For a full description of each of the variable types, click on the **Help** button.

5.4 To enter alpha characters into the variable **sex**, select data type **String** by clicking on its selection button. Notice that a string of 8 characters is the default. Whilst this would cause no problem, it is more efficient to reduce it to the actual number of characters you are going to input. Therefore change the default of **8** in the box to ***1***.

- Click **OK** to return to the **Variable View** window.
- Click on the variable type options box for the 2 **dates** at the bottom of the list.

  Choose **Date** for these variables.

  There are a large number of different ways in which the date can be written. The top option **dd-mmm-yyyy** is probably the safest option, as it is least ambiguous, but if you are using SPSS to enter the data you might consider choosing **dd.mm.yyyy** as the easiest one for entering data. You can switch back to **dd-mmm-yyyy** after the data has been entered.

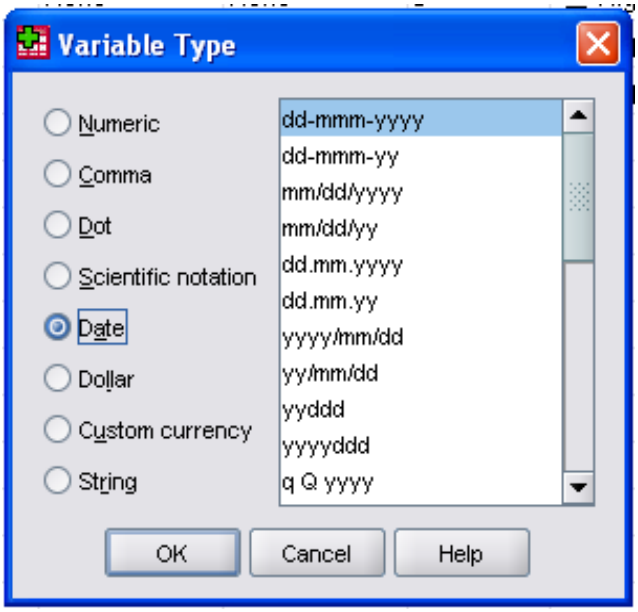


Figure 7 - Variable Type date options

Variable width and decimal places

The next column is headed **Width** and deals with the maximum number of characters that will be displayed for a particular variable in all output relating to this variable. It does not control the display in the **Data View** window, which is determined by **Columns** - see later. For a numeric variable it needs to be considered alongside the next column labelled **Decimals**. The value in this column indicates the number of decimal places that will be displayed in all output relating to this variable. By default the **Width** value is set to 8 and **Decimals** to 2. This may be felt to be appropriate and not need changing. For finer control of your output, you can alter values as necessary. For a string or date variable, the **Decimals** column has no meaning. For the **Date** variables, the **Width** and **Decimals** columns are fixed by the display version you have chosen.

For the example dataset, it would be better to choose:

Width *1*, **Decimals *0*** for variables smoker, cigs, pipe, cigars, give_up, tax, danger and cinemas.

Width *2*, **Decimals *0*** for variable age.

Width *3*, **Decimals *0*** for variable num_cigs.

Width *4*, **Decimals *0*** for variable ref_no.

Width *1* for string variable sex.

5.5 Starting in the first row, fourth column, click on the cell and either type the required decimals value (***0***) or use the small up/down arrows that appear to adjust the value. Move to the third column and change the value to two.

Note (If you attempt to reduce the **Width** to 2 first, it will fail since a number with 2 (by default) decimal places needs a minimum of 3 columns width).

➢ Repeat this process for all the other variables.

Variable labels

The next column is headed **Label** and is used to inform SPSS about the details associated with each variable name. The maximum length of any label is 255

characters and there are no restrictions on what may appear. Spaces are entered just as typed. If you want to specify where a new line appears in a label, type **\n** within the text and SPSS will wrap the label at this point.

5.6 Moving to the first row, fifth column, click on the cell and type in the words: ***Reference number***. The width of the column can be expanded to allow for the number of characters in the label. To do this, place the cursor on the divide between Label and Values in the table headings, where it will change to a two headed horizontal arrow, and then drag to the right as required.

- Move down to row 2, column 5 and type ***Age last birthday***.
- Continue entering the labels for all the other variables as given in Table 2.
- To correct any existing labels, double-click on the entry and edit as you would in a word processor.
- Alternatively use copy and paste to enter all the labels from a word table in one go.

Value labels

5.7 The next task is to enter **Value Labels** for each variable if appropriate. These will appear in the **Values** column. For our first two variables (**ref_no** and **age**) there are no **Value Labels**, so the default entry of **None** can remain. For **sex** you can indicate that **M** is male and **F** is female.

Figure 8 - the value labels options box

- Move to row 3 column 6 and click on the grey box on the right of the cell. A dropdown menu appears so you can provide Value Label information (Figure 8).
- In the box by the word **Value** type ***F***. In the box by the word **Label** type ***Female***.
- Click on **Add** and watch the value and its label move to the bottom box.
- In the box by the word **Value** now type ***M*** and the word ***Male*** in the **Label** box.

  Click on **Add**. Now that all the **Value Labels** for this variable are complete (your screen should look like Figure 9), click on **OK** to return to the **Variable View** window.

Figure 9 - the completed value labels options box

Note The first part of the Value Label that you entered appears in the appropriate cell.

5.8 The next variable needing a **Value Label** is smoker:

- Click on row **4** column **6** and its grey box**.** Enter the value ***1*** and label ***Yes*** in the dropdown menu box and click **Add**. Enter the Value ***2*** and Label ***No*** and click on **Add**. Click on **OK**.

5.9 The variables **cigs**, **pipe**, **cigars** and **give_up** all need the same **Value Labels** as **smoker**. Either you can repeat the above instructions for each variable in turn or take advantage of a useful shortcut:

- Click in the cell containing the values for **smoker**. Copy this cell by using either **Edit / Copy** from the main menu or by clicking the right mouse button and selecting **Copy**, or by pressing **CTRL + C**. Point at the cell for **Value** for variable **cigs** and paste the current clip board using either **Edit / Paste** or use the right mouse button and choose **Paste**, or use **CTRL + V**. Repeat this process for **pipe**, **cigars** and **give-up** values.

5.10 Finally for the variables **tax**, **danger** and **cinemas**, you will need to provide four value labels for each.

- Use the basic method to enter this information for the variable **tax**:
  - Value 1 has the label ***Strongly disagree***.
  - Value 2 has the label ***Disagree***.
  - Value 3 has the label ***Agree***.
  - Value 4 has the label ***Strongly agree***.
- Now copy and paste for the other two variables **danger** and **cinemas**.

Missing values

The next column of the **Variable View** sheet is **Missing Values**. In the statistical analysis of any dataset it is sometimes necessary to exclude cases where the information is not known or not appropriate. An example of this occurs in the variable **num_cigs** in this dataset. The information is missing in two situations; for

non-smokers it is not appropriate and in the data the appropriate cell has been left blank; for one interview, **ref_no 8**, the respondent failed to give an answer to this question (no he did not smoke 999 cigarettes per day!) A number that could not be expected as a genuine response is selected to represent this circumstance. However in any analysis, it should not be considered as it would seriously distort many statistical procedures.

Within SPSS there are two types of missing value - **system-missing** values and **user-defined** missing values. By default, for non-string variables, an empty cell is defined as a system missing value and does not need to be further declared. For user-defined missing values this column of the **Variable View** has to be used.

Figure 10 - the missing values options box

- Click on row 6, column 7 and then the small grey box at the right of the cell to produce the **Missing Values** dropdown menu (Figure 10).
- Select the **Discrete missing values** button and enter ***999*** in the first cell.
- Click on **OK** and return to the **Variable View** window. Notice that the value 999 now appears in the missing column for the variable **num_cigs**.

If entering missing values for string variables, the required discrete value should be entered as characters e.g. X to represent the letter X, a space to represent an empty cell.

Data display

The next two columns (**Columns** and **Align**) are concerned with the display of data in the **Data View** window. For the purposes of this example dataset, the default values of a column 8 characters wide and the values right aligned for numeric variables and left aligned for string variables are fine. When you have entered your data as instructed below, return to the **Variable View** window and change one or more of these values. Then flip to the **Data View** window and see the effect your choice has made.

Measurement scale of variables

The almost final column is concerned with the measurement scale properties of your variable. In statistics certain procedures are only appropriate for variables measured on specific scales of measurement. The measurement characteristics recognised by SPSS are as follows:

scale	to represent a numeric variable that can take discrete or continuous values along a range
ordinal	to represent values that, although numeric, only represent an ordered listing of such values
nominal	to represent values that are simply names

You should be able to recognise that in the example dataset, there are:

- 6 nominal measures - **sex**, **smoker**, **cigs**, **pipe**, **cigars** and **give up**.
- 3 ordinal measures - **tax**, **danger** and **cinemas**.
- 3 scale measures - **ref_no**, **age** and **num_cigs** (**ref_num** could be nominal, ordinal or scale!)
- Plus 2 date variables which are considered by SPSS to be scale variables and can be used as such.

5.11 Starting with **ref_no** in row 1, column 10, click on the cell and choose the appropriate measure. (You should not have to change this from the default).

➢ Move down the column making the appropriate choice in each case. The first change is for the variable **smoker**, which will need to be changed to **nominal**.

You have now defined all the information that SPSS needs to know about the characteristics of your specific dataset. Your variable view pane should look like Figure 11.

Role of variables

The final column is concerned with role your variable is going to take in the analysis. This is a relative new column (version 18 onwards). In statistics certain procedures are only appropriate for certain types of variable. It is not unusual for a research study to use one variable as the outcome for one set of analyses and as the confounder in another set of analyses, which means that using this column may mean rechecking these classifications before each analysis. The roles recognised by SPSS are as follows:

Input – this is variable can be used as an independent predictor.

Target – this is the outcome of the analysis

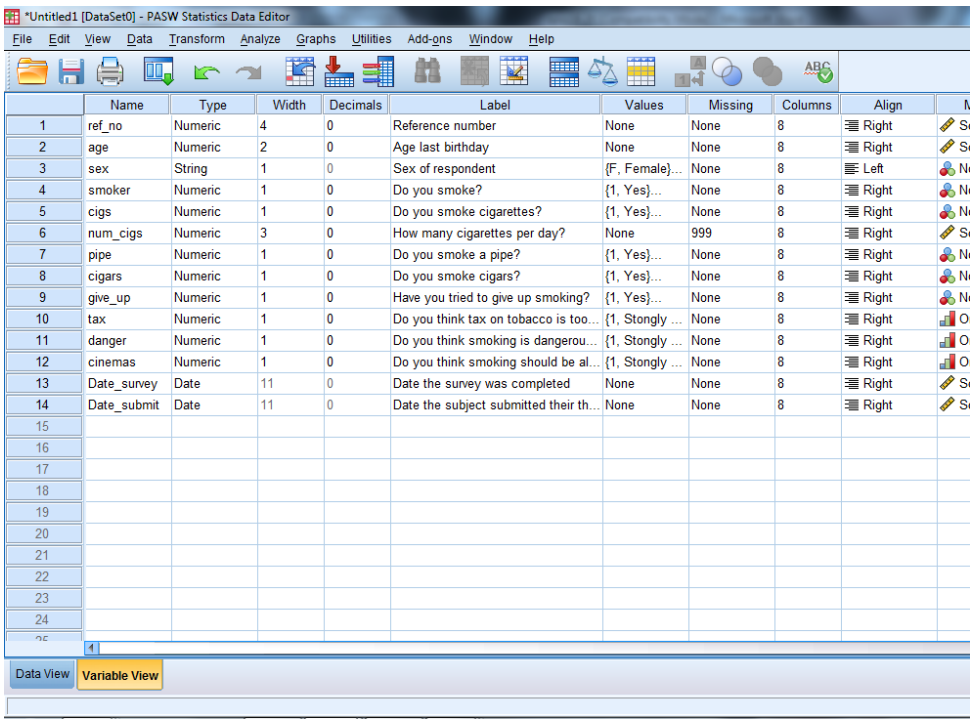
Both – this can be either target or input

None – no role assigned

Partition – this variable can be used to partition the data, such as a variable which defines a test or training data set.

Split – I believe this is included for compatibility with other PASW programmes

There are very few procedures which require the role to be defined. We will leave all variables with the default role.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	ref_no	Numeric	4	0	Reference number	None	None	8	Right
2	age	Numeric	2	0	Age last birthday	None	None	8	Right
3	sex	String	1	0	Sex of respondent	{F, Female}...	None	8	Left
4	smoker	Numeric	1	0	Do you smoke?	{1, Yes}...	None	8	Right
5	cigs	Numeric	1	0	Do you smoke cigarettes?	{1, Yes}...	None	8	Right
6	num_cigs	Numeric	3	0	How many cigarettes per day?	None	999	8	Right
7	pipe	Numeric	1	0	Do you smoke a pipe?	{1, Yes}...	None	8	Right
8	cigars	Numeric	1	0	Do you smoke cigars?	{1, Yes}...	None	8	Right
9	give_up	Numeric	1	0	Have you tried to give up smoking?	{1, Yes}...	None	8	Right
10	tax	Numeric	1	0	Do you think tax on tobacco is too...	{1, Stongly ...	None	8	Right
11	danger	Numeric	1	0	Do you think smoking is dangerou...	{1, Stongly ...	None	8	Right
12	cinemas	Numeric	1	0	Do you think smoking should be al...	{1, Stongly ...	None	8	Right
13	Date_survey	Date	11	0	Date the survey was completed	None	None	8	Right
14	Date_submit	Date	11	0	Date the subject submitted their th...	None	None	8	Right
15									
16									
17									
18									
19									
20									
21									
22									
23									
24									

Figure 11 - the variable view screen with defined information

You may return to the **Variable View** window at any time if further changes are needed.

Task 6 Entering data

The **Data View** pane of the **Data Editor** window is used to enter the data. Displayed initially is an empty spreadsheet with the variable names you have defined appearing as the column headings.

The data that you need to enter is given in Task 2, but for your convenience the data table is reproduced here.

Ref number	Age	Sex	Smoke	Smoke Cigs	How many	Pipe	Cigars	Give up	Tax	Health	Cine ma	Date survey	Date submit
1	27	F	1	1	10	2	2	1	3	3	3	02-Nov-1995	25-Sep-1995
2	31	M	2						4	2	1	16-Nov-1995	04-Sep-1995
3	35	M	2						4	1	1	01-Dec-1995	21-Jul-1995
4	58	M	2						3	1	2	17-Nov-1995	05-Sep-1995
5	56	M	2						4	3	2	10-Nov-1995	30-Sep-1995
6	25	F	1	1	20	2	2	2	3	4	4	02-Dec-1995	10-Aug-1995
7	41	F	1	1	30	2	1	1	3	1	3	17-Nov-1995	24-Aug-1995
8	38	F	1	1	999	2	2	1	4	4	4	02-Dec-1995	30-Aug-1995
9	43	F	1	2		2	1	1	4	2	2	16-Nov-1995	01-Aug-1995
10	29	M	1	1	40	2	2	2	2	4	4	02-Dec-1995	26-Sep-1995

6.1 Click on the **Data View** tab of the **Data Editor** Window

- To enter the first person's data, click the first cell of **ref_no**.
- Type ***1***.
- Press the **TAB** key or right arrow once and the heavy outline moves to the next column.
- Type in ***27*** and press the **TAB** key.
- Type in ***F*** and press the **TAB** key.
- Type in ***1*** and press the **TAB** key.
- Follow the same procedure along the first row until all twelve data values are entered.

6.2 Move back to row 2, column 1 and start to enter the values for interview 2. Press the **TAB** key twice to skip over a column. Notice that a dot appears in the cell. This is the system-missing value (see Figure 12).

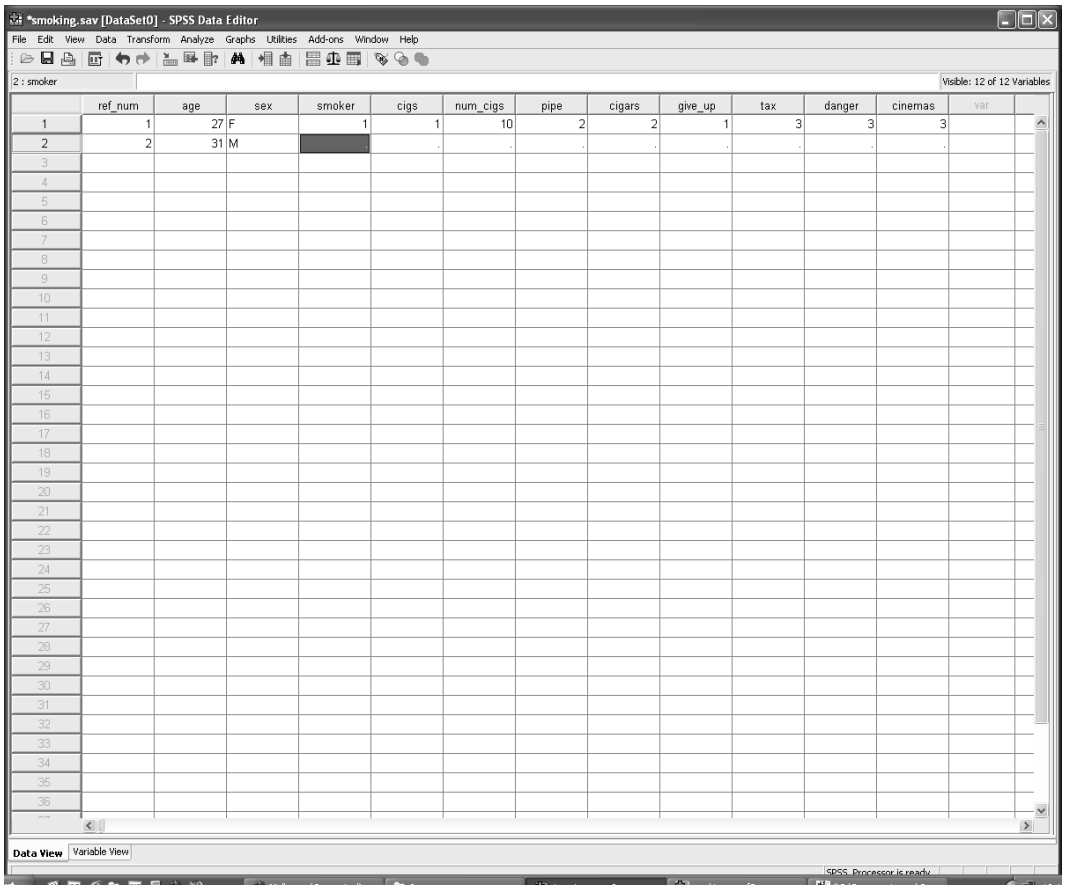
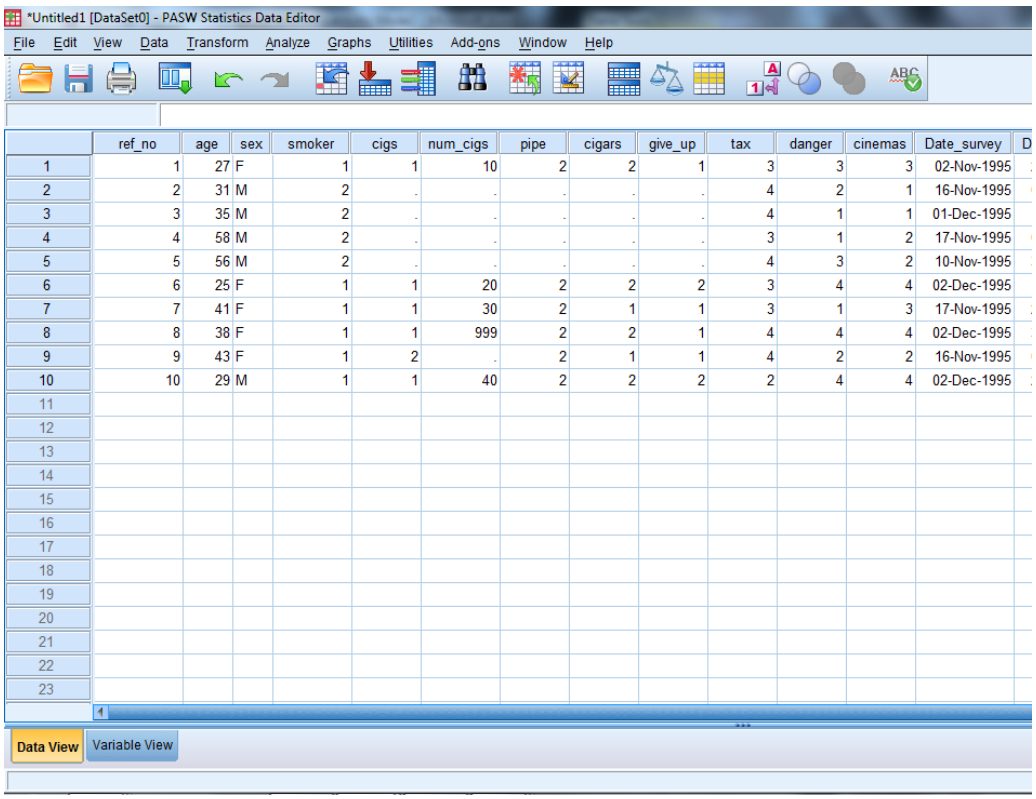


Figure 12 - Data Editor window with the first and part of the second interview entered

6.3 Continue the process until the data from all ten interviews are entered. Your **Data Editor** screen should now look like Figure below. Some people find it easier to enter data by column rather than by row. The method is similar except that you use the down arrow key instead of the **TAB** key. The **HOME** and **END** keys take the cursor to the first or last column of a particular case. **CTRL + HOME** will take you to row 1, column 1, and **CTRL + END** to the last used cell.



	ref_no	age	sex	smoker	cigs	num_cigs	pipe	cigars	give_up	tax	danger	cinemas	Date_survey
1	1	27	F	1	1	10	2	2	1	3	3	3	02-Nov-1995
2	2	31	M	2	.	.	.	.	.	4	2	1	16-Nov-1995
3	3	35	M	2	.	.	.	.	.	4	1	1	01-Dec-1995
4	4	58	M	2	.	.	.	.	.	3	1	2	17-Nov-1995
5	5	56	M	2	.	.	.	.	.	4	3	2	10-Nov-1995
6	6	25	F	1	1	20	2	2	2	3	4	4	02-Dec-1995
7	7	41	F	1	1	30	2	1	1	3	1	3	17-Nov-1995
8	8	38	F	1	1	999	2	2	1	4	4	4	02-Dec-1995
9	9	43	F	1	2	.	2	1	1	4	2	2	16-Nov-1995
10	10	29	M	1	1	40	2	2	2	2	4	4	02-Dec-1995
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													
21													
22													
23													

Figure 13 - the data editor window with all interview data entered

Task 7 Help system

More information about SPSS may be found in the on-line help facility. To access on-line help:

- Click **Help** from the main menu.

  A menu appears from which you can choose further topics.

- Click **Topics** and an output screen similar to the one shown in Figure 14 should open in a new window.

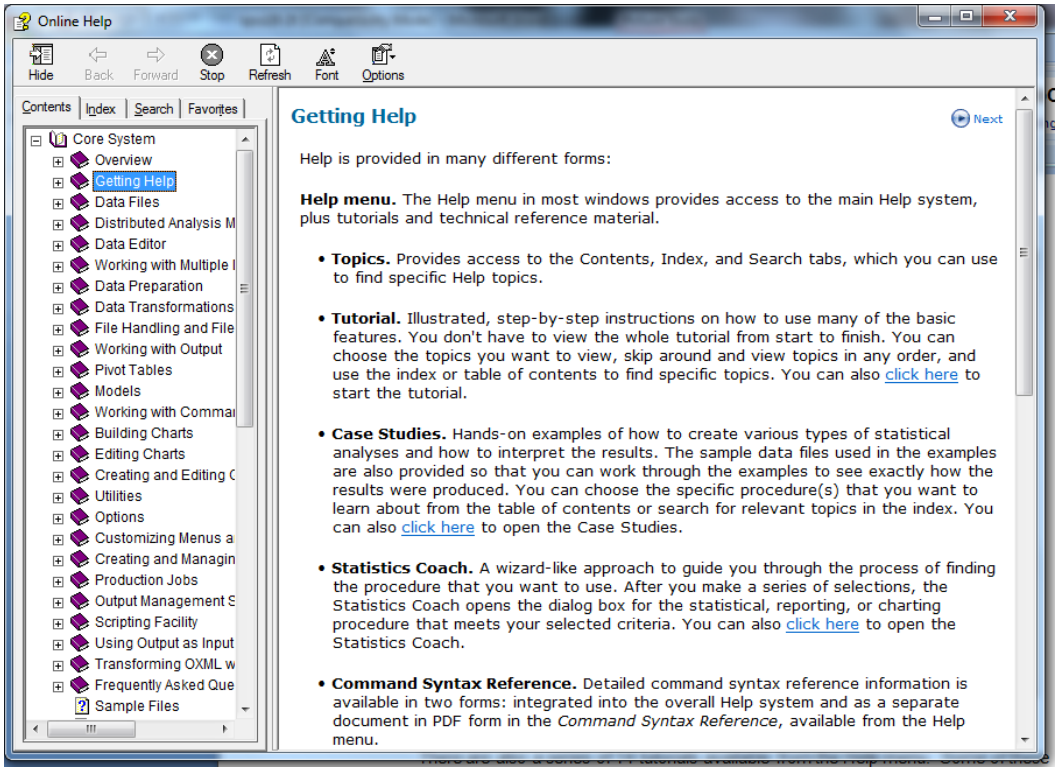


Figure 14 - a typical Help screen from SPSS

There are also a series of 14 tutorials available from the Help menu. Some of these tutorials are directed at a more commercial or financial perspective rather than an academic use, but they can be used to familiarise yourself with the way SPSS works. They are not interactive tutorials.

Task 8 Frequency tables - the frequencies procedure

8.1 If you want to know the number of males and females in this sample, you need to see the distribution of the variable **sex**, by using the **Frequencies** procedure:

- In the **Data Editor** window select **Analyze**.
- From the **Analyze** menu select **Descriptive Statistics**.

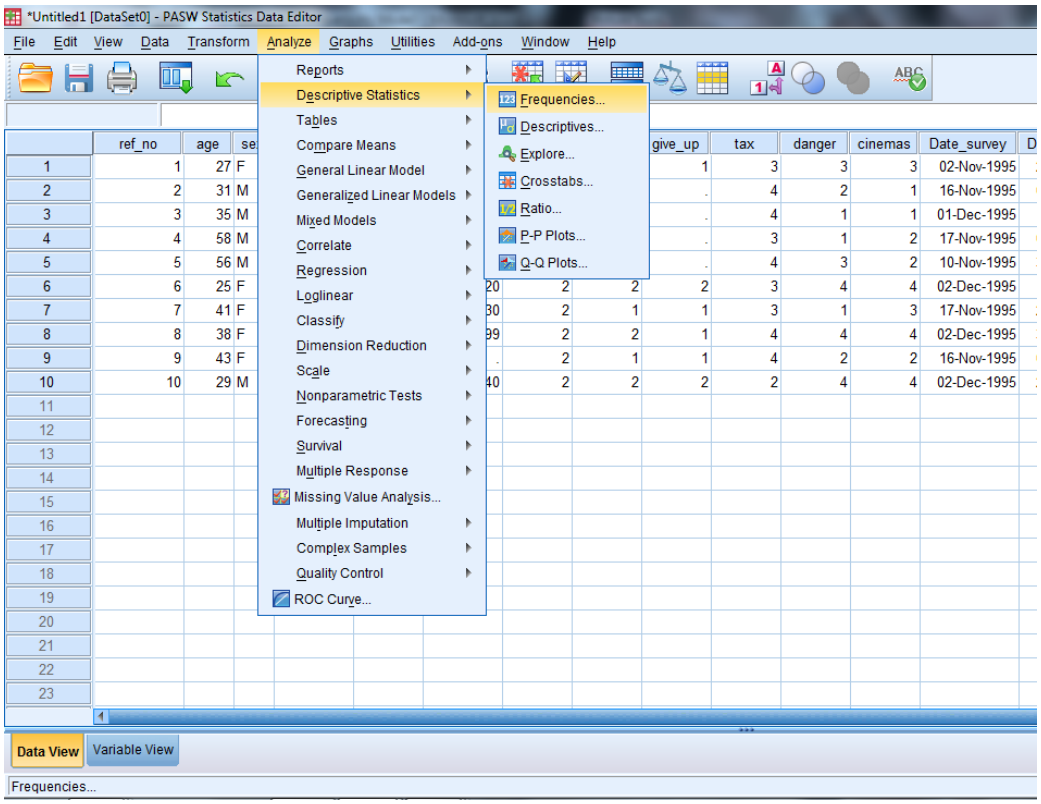


Figure 15 - the window with drop down menus from the Analyze command

- From the **Descriptive Statistics** submenu, select **Frequencies** (Figure 15).

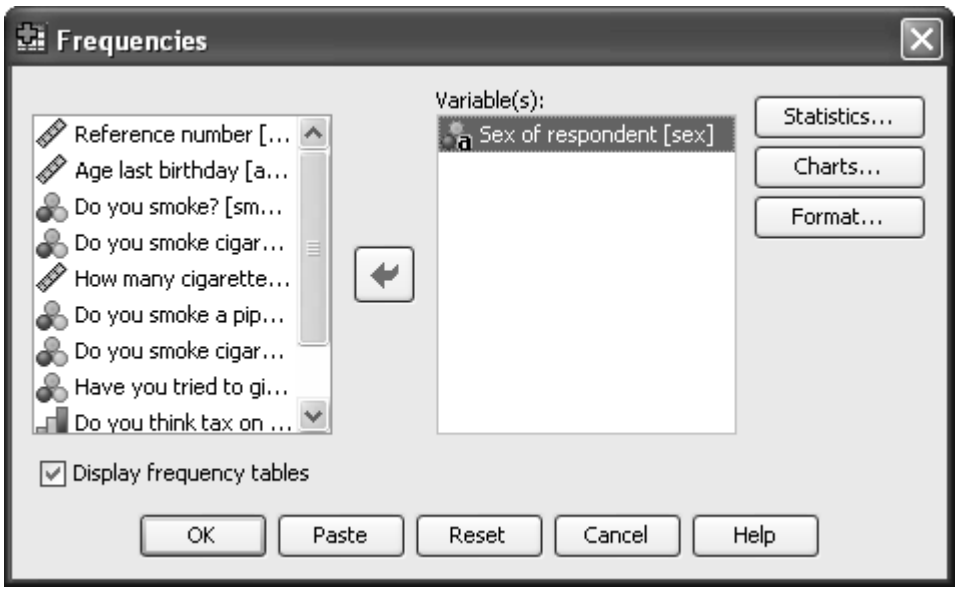


Figure 16 - the Frequencies options window

- Select **Sex of respondent**.
- Click the right pointing arrow head (➥) to move **sex** into the **Variables** box (see Figure 16).
- Click **OK**.

A frequency table is produced. Note that tables, statistics and charts are displayed in the **SPSS Viewer** window – a completely different window from the Data Editor (Figure 17).

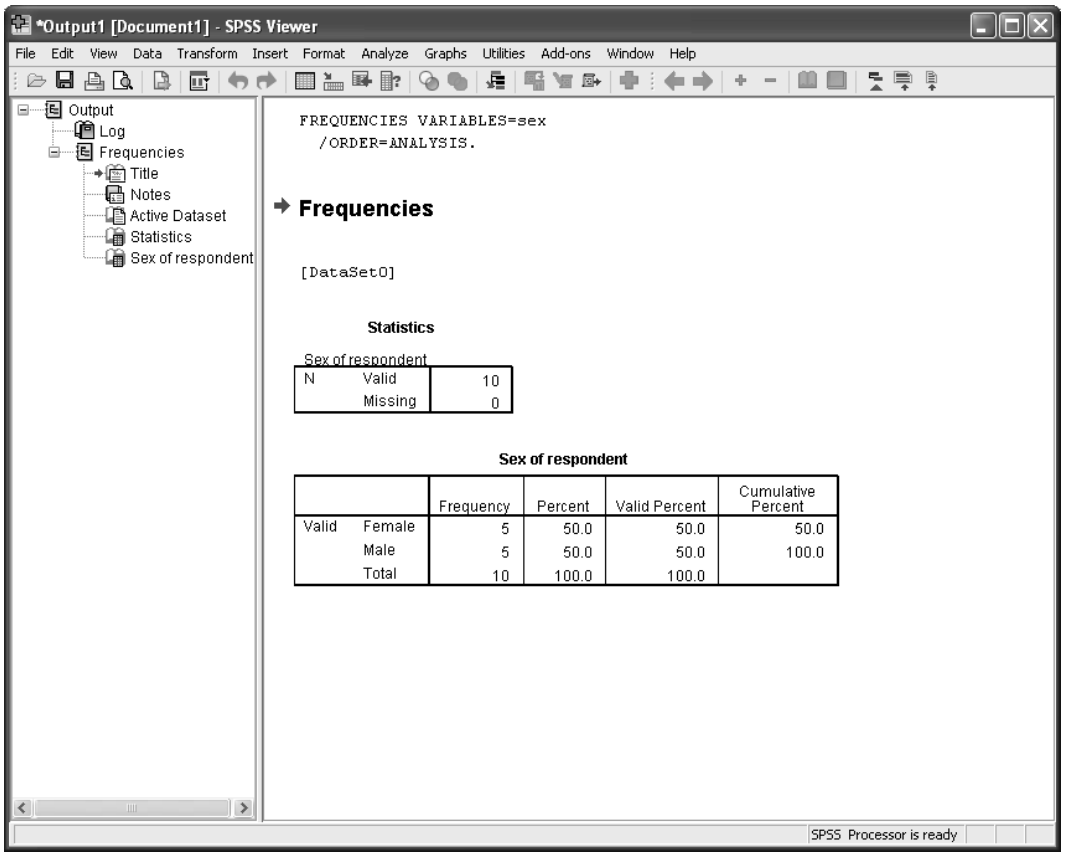


Figure 17 - the SPSS Viewer (Output) window with a frequency table

The **SPSS Viewer** window has two panes, each with its own bottom and right scroll bars. The left pane contains an outline (index) of the results so far. The right pane contains the detailed tables, graphs and text output. Clicking on the **+** and **-** symbols in the left pane controls what output is displayed in the right pane. Clicking on the other section names in the left pane moves the focus of the right pane display so it starts the display at the selected item.

You can use the **Window** command to select which window you want to be in at any particular time.

Check that you can see the command lines above the output:

```
FREQUENCIES VARIABLES=sex
  /ORDER=ANALYSIS.
```

If this is not visible go to **Edit** and choose **Options** from the **viewer** tab tick the bottom left button labelled **Display commands in the log**.
Click on **OK**. (this is dealt with in more depth in section 11)

8.2 Click **Window**.

Note that there are currently 2 types of window:

- Untitled [Dataset0] - SPSS Data Editor
- Output1 [Document 1] - SPSS Viewer

➢ Click **SPSS Data Editor**.

Task 9 Saving files in SPSS

Saving an SPSS data file

9.1 To save an SPSS **data** file:

- Select **File** in the **Data Editor** window (rather than **File** in the **Viewer** window!).
- Select **Save As** (see Figure 18).
- Ensure the desired directory is displayed in the **Look in** box. If you are using a file in one of the Computer Centre training rooms change the directory to: **C:\Training\Stats**.
- Type ***smoking*** in the **File name** box.
- Click **Save**.

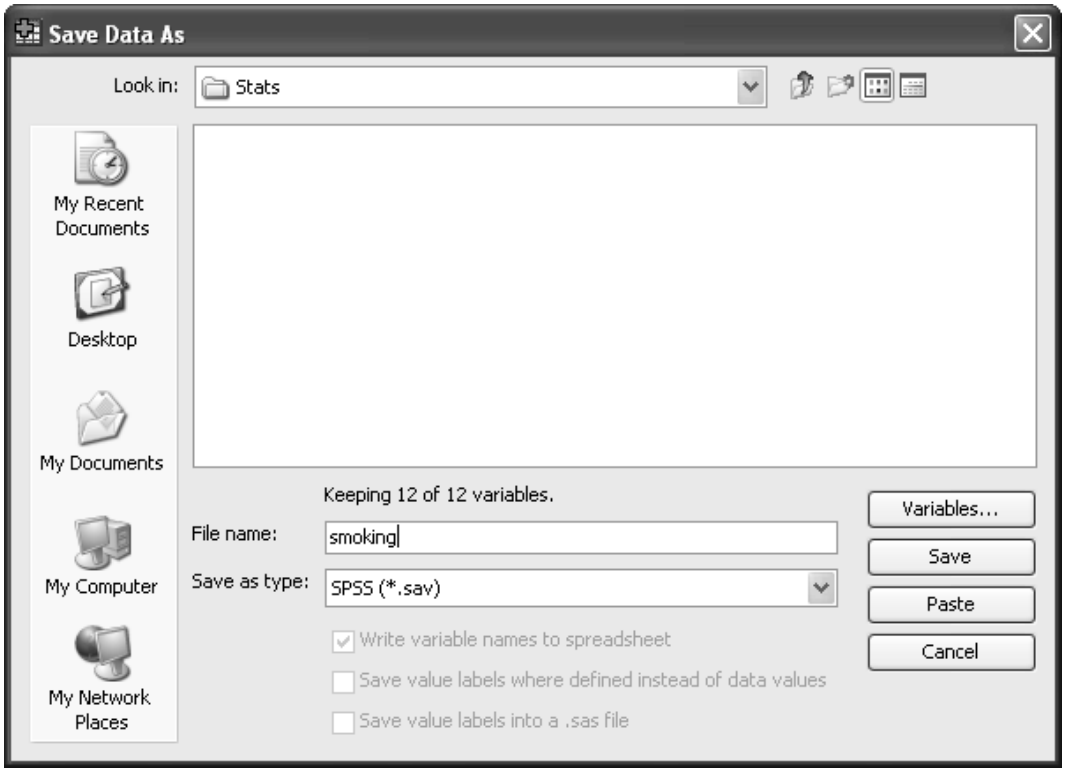


Figure 18 - the Save/Save As data options window

SPSS saves the file as **smoking.sav** in the specified directory. Normal Windows rules apply to file names.

Saving an SPSS output file

9.2 To save an SPSS **output** file:

- Select **File** in the **SPSS Viewer** window.
- Select **Save As** (see Figure 19).
- Ensure the desired directory is displayed in the **Save in** box. If you are using one of the Computer Centre training rooms change the directory to **C:\Training\Stats**.
- Type ***smoking*** in the **File name** box.

- Click **Save**.

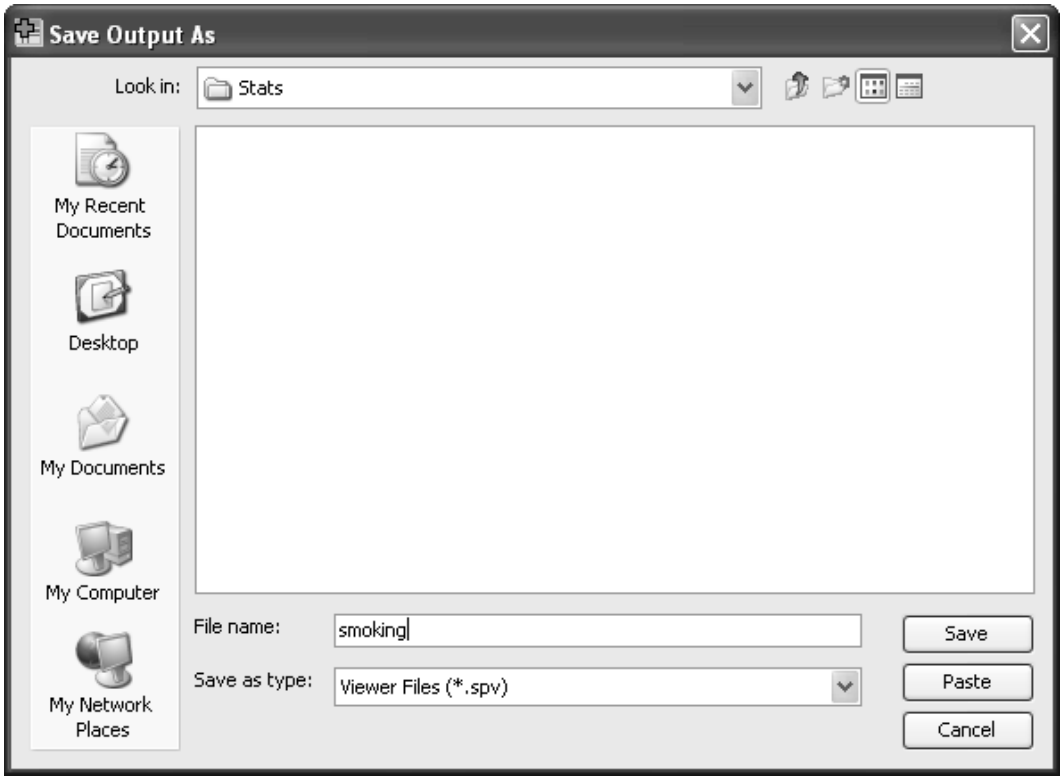


Figure 19 - the Save/Save As output options window

The saved **output** from SPSS appears in the file **smoking.spv**. This is the output produced by the various procedures, eg tables, means, plots. It does NOT contain all the values of the dataset variables nor all the label information so cannot be used like an SPSS data file for future analysis purposes. It is known as the spool file.

Note This file is not a text file - you must inspect such files by opening them via the **File** option in the **SPSS Viewer** window. To get other formats eg, text, you must select the desired items in the **Viewer outline** and use the **File Export** option (see Task 27 at the end of this document).

Task 10 Leaving SPSS

10.1 To leave SPSS:

- Click **File** in the **Viewer** or **Data Editor** window.
- Click **Exit**.

  SPSS shuts down all open windows and exits. It prompts you if data and/or output have not been saved before exiting.

Task 11 Opening a file

11.1 To open an existing file; in this case, **smoking.sav**:

- Go back into **SPSS 23.0**.
- Click on **smoking.sav** in the file list.
- Click **OK**.
- If **smoking.sav** is not listed, then click on **More files** and click on **OK**.
- Ensure the correct directory is specified in the **Look in** box (Figure 20)

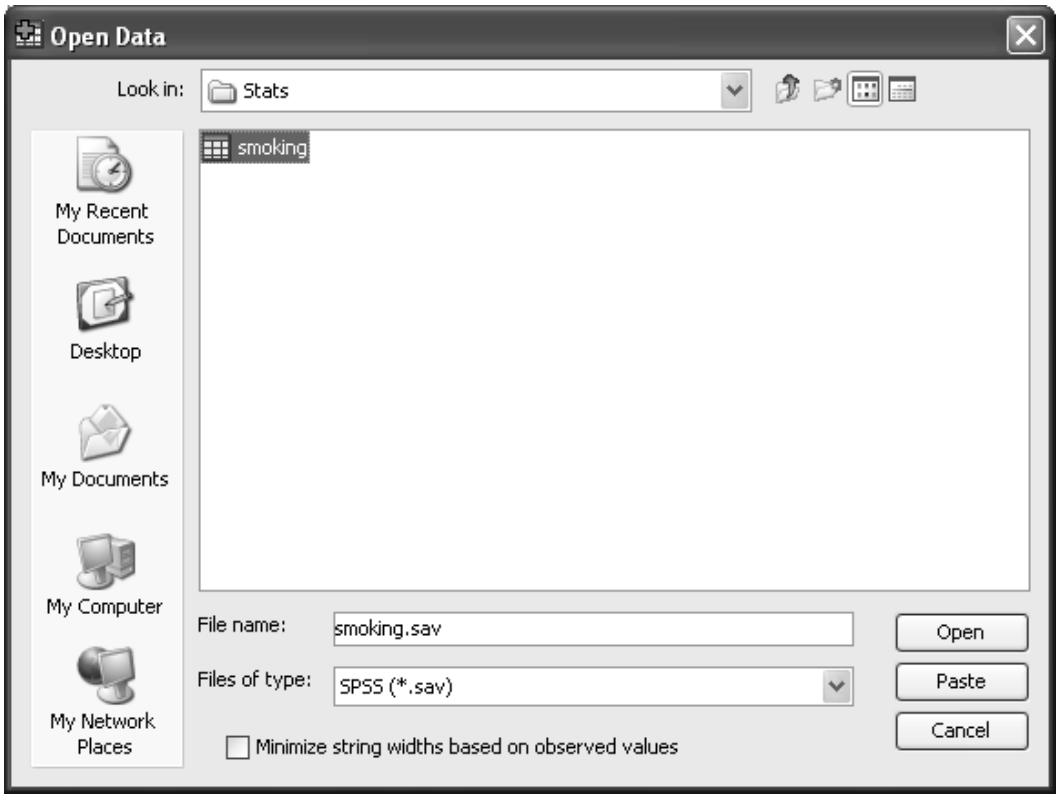


Figure 20 - the open data file options window

- Select **smoking**.
- Click **Open**.

Task 12 Controlling your output

SPSS provides a number of useful commands for directing your output in a specific fashion. You can experiment with most of them in your own time. However, there are a few minor changes that would help the remainder of this practical.

12.1 From the main menu click on **Edit**.

- Click on **Options** in the submenu.

  A dropdown menu of some complexity appears as in Figure 21.

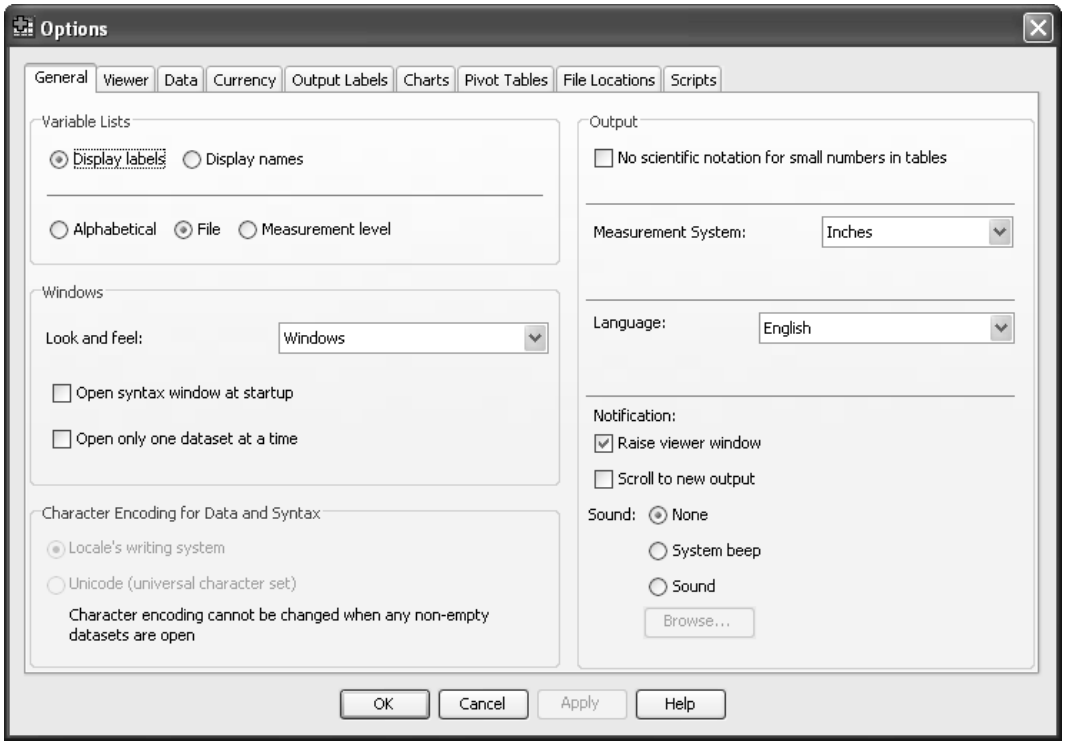


Figure 21 - the General window from the Edit Options submenu

- Select the **General** tab (it may be displayed automatically).

  At the top left is the **Variable Lists** section. These options specify how the list of variables offered in all SPSS procedures is composed. If you prefer your variables to be listed in alphabetical order, as opposed to entry (file) order, press the appropriate button. Also some users prefer to have the variable name and not its label displayed.

  In the **Notification panel** of the Output section, you will see a Scroll to new output option. You are advised to always have this ticked.

  If you make any changes to this **General** tab, click the **Apply** button before moving on. You may be advised that certain changes cannot be applied until you start SPSS again. Click on **OK** but remember when you have made all your changes you may need to leave SPSS and restart it from the SPSS 23.0 command.

12.2 Select the **Viewer** tab (Figure 22).

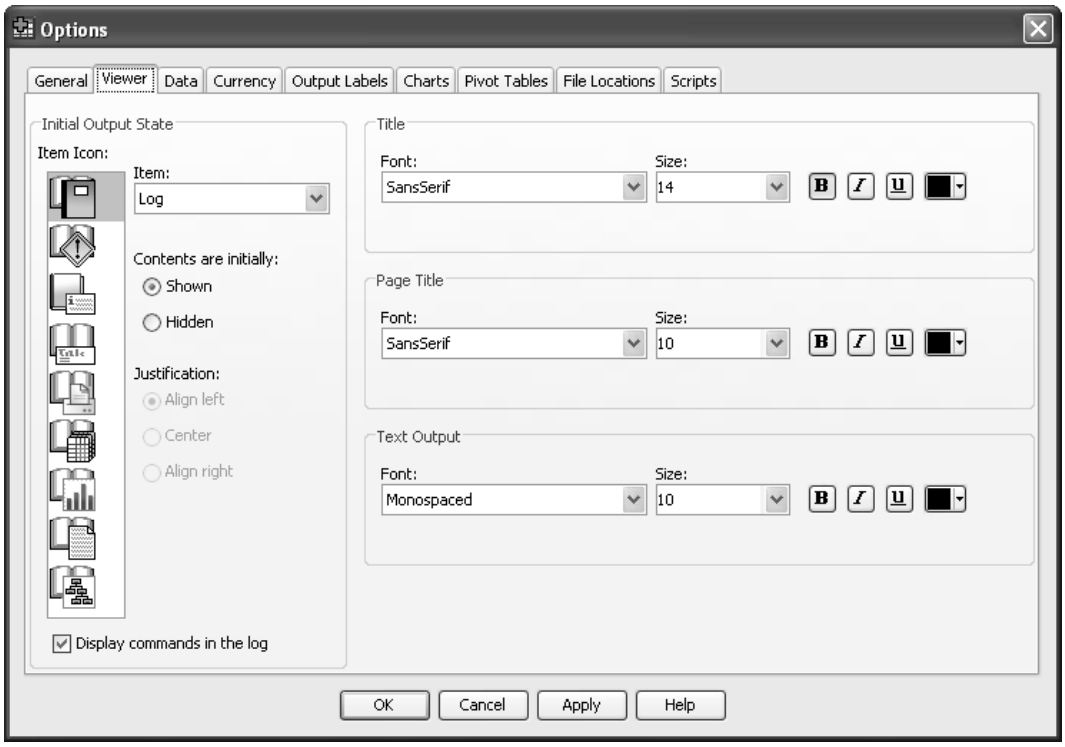


Figure 22 - the Viewer window from the Edit Options submenu

At the foot of the first column (**Initial Output State**) is a tick box labelled **Display commands in the log**. It is very helpful in any discussion of your output to have this command log available. If necessary tick the box by clicking on it. When you now run any SPSS command, your output file will include a representation of the choices you made from any drop-down menu. This takes the form of the commands generated in SPSS's own language. You can see this in Figure 23 on the following page where the first two lines in the right hand frame are the log generated by the particular request for a Frequencies count. This SPSS language is referred to as **syntax** and is a very powerful tool in the advanced use of the program.

The right hand frames of this screen allow you to change the default fonts used by SPSS in the production of output.

If you make any changes to this **Viewer** tab, click the **Apply** button before moving on. This practical does not intend to investigate any further tabs in the **Options** menu, so press **OK** to exit this section. (Remember to restart SPSS if you have been told it is necessary.)

Task 13 Procedure commands - Frequencies

Comments Having created an SPSS dataset, SPSS procedures may be used to analyse and process the data. There are many different procedures in SPSS. The following sections describe the procedures **Frequencies**, **Descriptives**, **Crosstabs**, **Correlations** and **Graphs**. Now you need to decide what you want SPSS to do.

The questions you might wish to ask of the data include:

What is the distribution of each variable?

How many men and women are there?

Is there a correlation between **tax**, **cinemas** and **danger**?

13.1 To generate a **Frequencies** table:

- In the **Viewer** or **Data Editor** window, select **Analyze**.
- From the **Analyze** menu, select **Descriptive Statistics**.
- From the **Descriptive Statistics** submenu, select **Frequencies**.

The **Frequencies** box contains a list of variables in the **Data Editor**. Generate two frequency tables one for **sex** and one for **smoker**.

- Click **Sex of respondent**.
- Click ➥.
- Click **Do you smoke? [smoker]**.
- Click ➥**,**
- Click **OK**.

  - Here, **Frequencies** is used to find the sex distribution of the sample and the distribution of the smokers. Examine the output. You should find that there are 5 women and 5 men, 6 smokers and 4 non-smokers and no missing values. The contents of the tables are the frequency and the percentage of the total number in the table.
  - You can alter the statistics included with this procedure. The default is number of valid cases and number of missing cases. The options available if you click the statistics button allow you to have other statistics such as mean, median, etc.

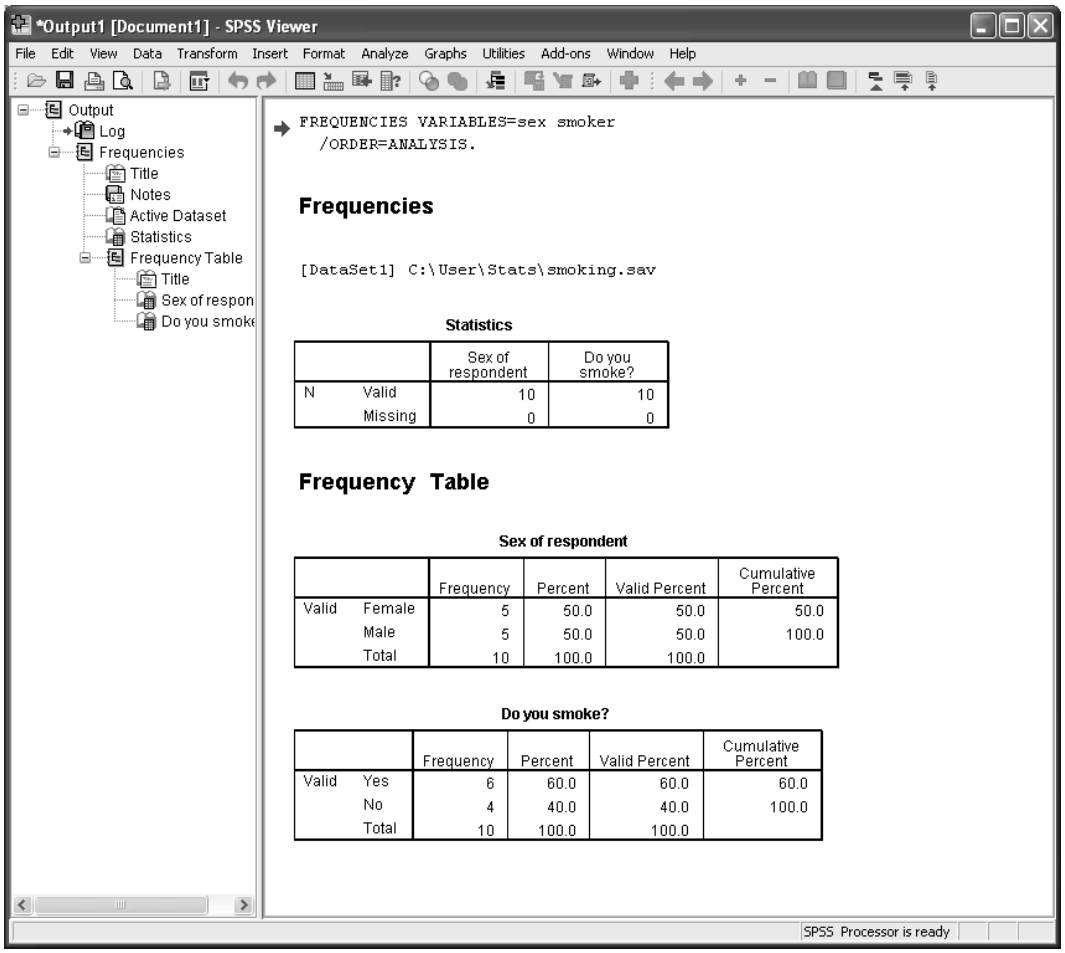


Figure 23 - the Output window from the Frequencies procedure

Notes for advanced SPSS use:

Objective : to be able to repeat your analysis on another dataset, or at another time quickly and accuratly. This is most useful If you are intending to publish the results of your analysis as it is wise to keep a copy of the syntax used to create them.

- The command (or syntax) used here is included by SPSS.
- Go back to the **analyze** menu and choose the **descriptive statistics** and **frequencies** menu options as before. SPSS will have remembered what you have just asked it to do, and will have retained the options you had chosen earlier.
- This time click on **Paste** instead of **OK**.
- You have created a record of your analysis request in a new **syntax** file.
- Save this file and give it the name **Analysis for Paper 1.sps**.
- When you want to redo the analysis all you will have to do is highlight the commands in the file and click on **run current** (run is the *green triangle*).

Task 14 Using summary statistics for continuous variables – the Descriptives procedure

With continuous variables, frequency tables are not always the best method of summarising. A better option would be to use a selection of summary statistics in place of a frequency table. Frequencies and crosstabulations are useful mainly for categorical variables, ie where the values represent categories such as male/female, nationality, class of university degree. However, variables like age and num_cigs have many values and for these continuous variables, statistics like the mean and standard deviation are sometimes useful.

14.1 In the **Viewer** or **Data Editor** window select **Analyze**.

- From the **Analyze** menu, select **Descriptive Statistics**.
- From the **Descriptive Statistics** submenu, select **Descriptives**.
- Select the variable **Age last birthday [age]** (see Figure 24).

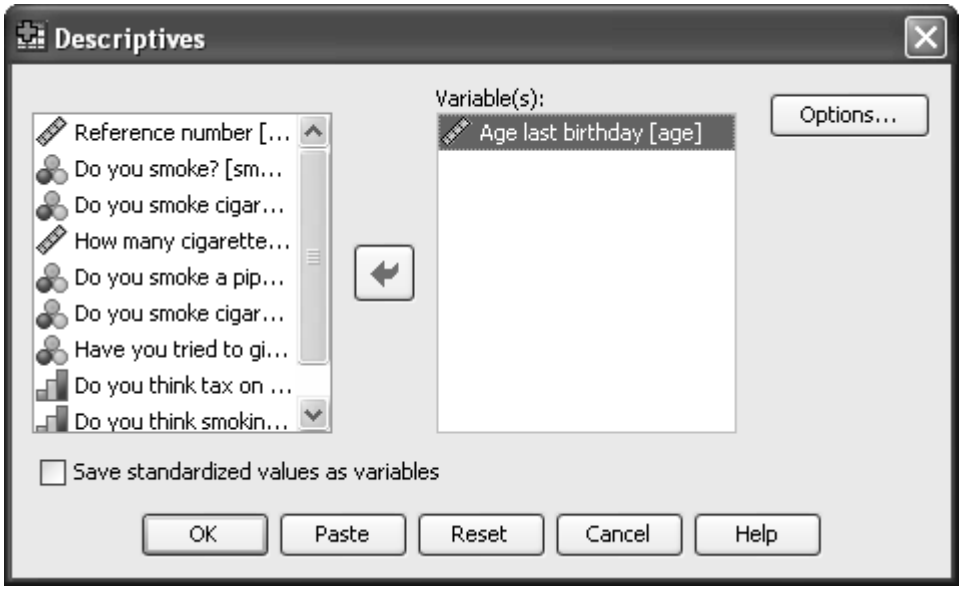


Figure 24 - the Descriptives option window

- Click the **Options** button in the **Descriptives** window.

  A number of statistics options which are appropriate for continuous variables are displayed (Figure 25).

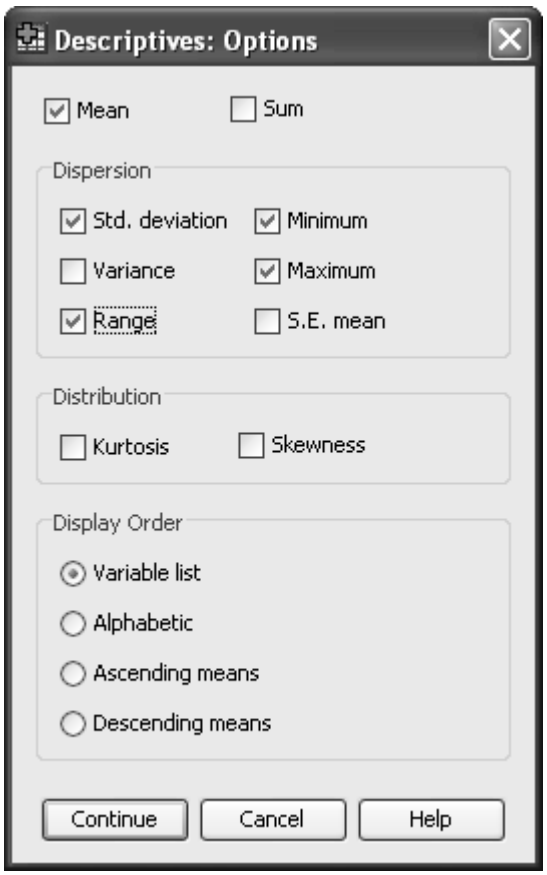


Figure 25 - the statistics options for continuous variables

- Select **Mean**, **Standard deviation**, **Minimum, Maximum** and **Range**.
- Click **Continue**.
- Click **OK**.

The required statistics are displayed in Figure 26. You may have to scroll down in the SPSS Viewer window to see this.

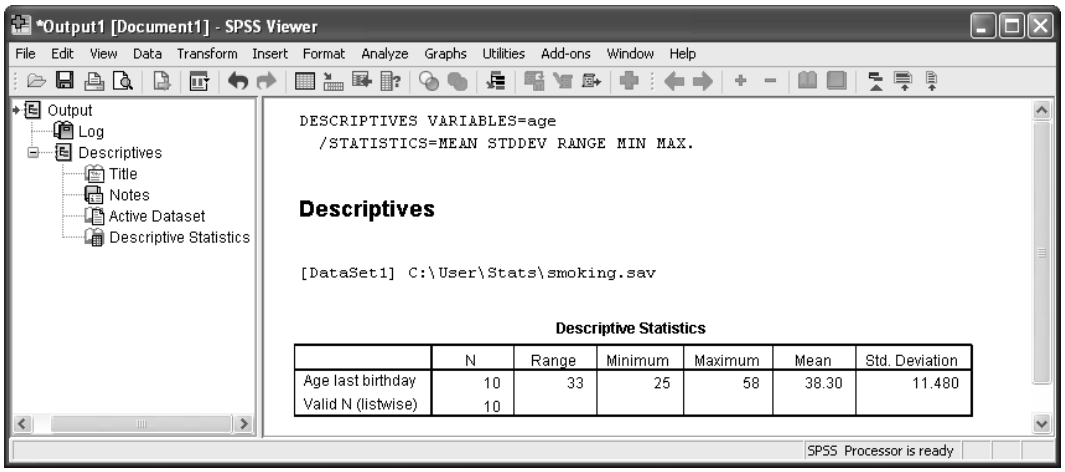


Figure 26 - SPSS viewer output from Descriptives procedure

Task 15 Producing a bar chart from frequencies

15.1 To create a bar chart for **sex** using the **Frequencies** box:

- In the **Viewer** or **Data Editor** window click **Analyze**.
- From the **Analyze** menu, click **Descriptive Statistics**.
- From the **Descriptive Statistics** submenu, click **Frequencies**.
- Click **Reset**.
- Select **Sex of respondent [Sex]** and then click ➥ (Figure 27).

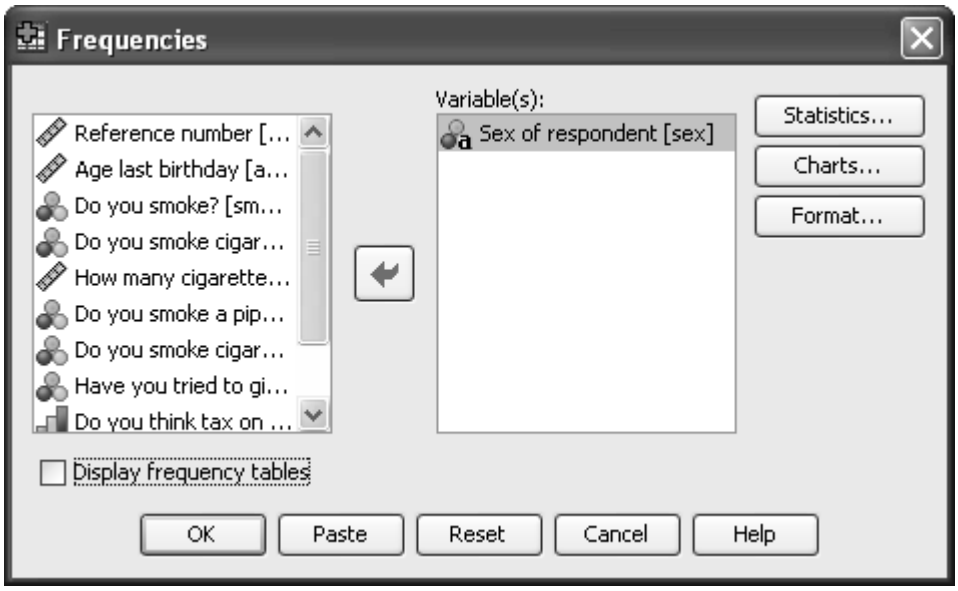


Figure 27 - the Frequencies option window

- Click **Charts** (see Figure 28).

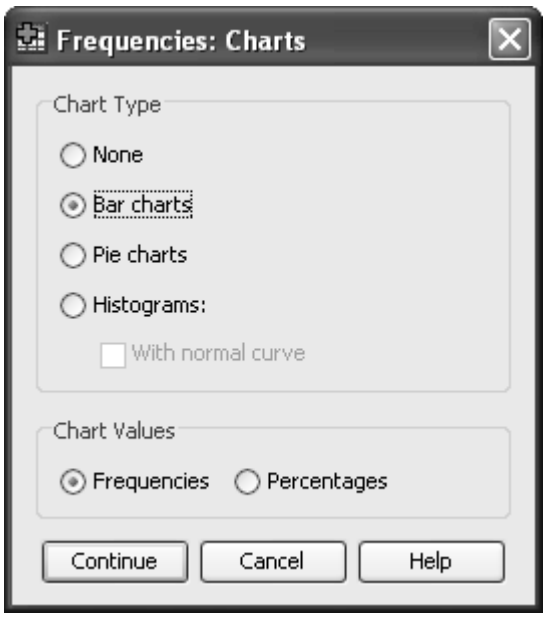


Figure 28 - the Charts submenu window from Descriptives

- Click the **Bar chart(s)** option and click on **Continue**.
- Click **Display Frequency Tables** to suppress the display of the frequency table.
- Click **OK**.

SPSS draws the chart and shows it in the **Output Viewer** window (Figure 29). (You may need to scroll down to see the complete chart.)

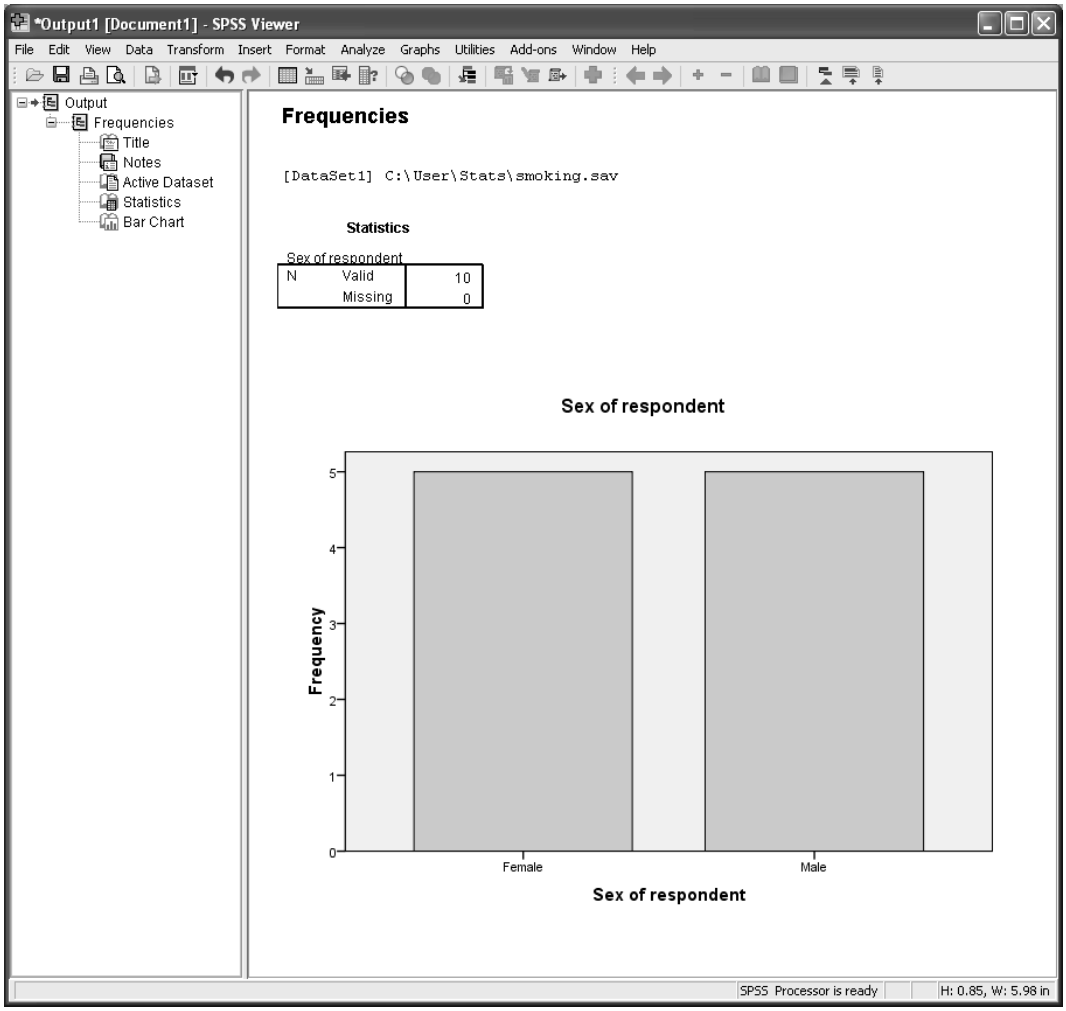


Figure 29 - SPSS viewer output produced using the Bar Chart option from the Frequencies procedure

To display the chart in a full screen window, double-click on the chart in the **Viewer** pane (this takes you into the SPSS **Chart Editor** window where the chart may be amended to suit personal preferences e.g. colour choice, labels, etc) (Figure 30).

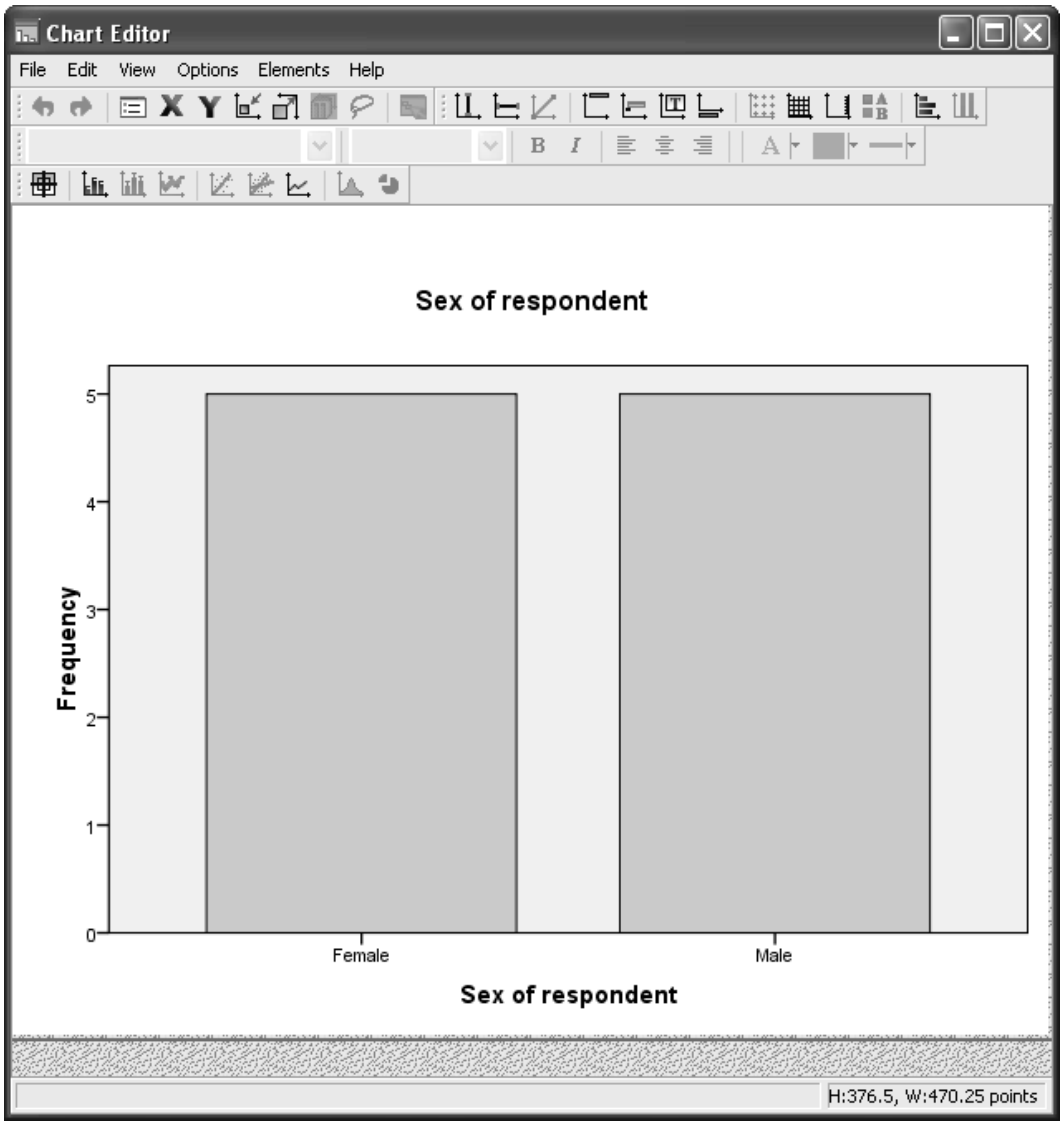


Figure 30 - SPSS Chart Editor window

- To leave the **Chart Editor** window click **File/Close**.

  If the chart editor is left open, then the image of the chart in the Viewer window is displayed as “crossed-out” and cannot be displayed properly until the chart editor is closed.

Task 16 Displaying histograms

16.1 Use the **Frequencies** dialog box to request a histogram:

- Select **Analyze**.
- Select **Descriptive Statistics**.
- Select **Frequencies**.
- Click **Reset**.
- Select **Age last birthday [age]**.
- Click **Charts**.
- Click **Histogram(s**).

16.2 To display a normal curve on the chart:

- Select **With normal curve**.
- Click **Continue** and then click **OK**.
- The histogram appears in the **Output Viewer** window (Figure 31).

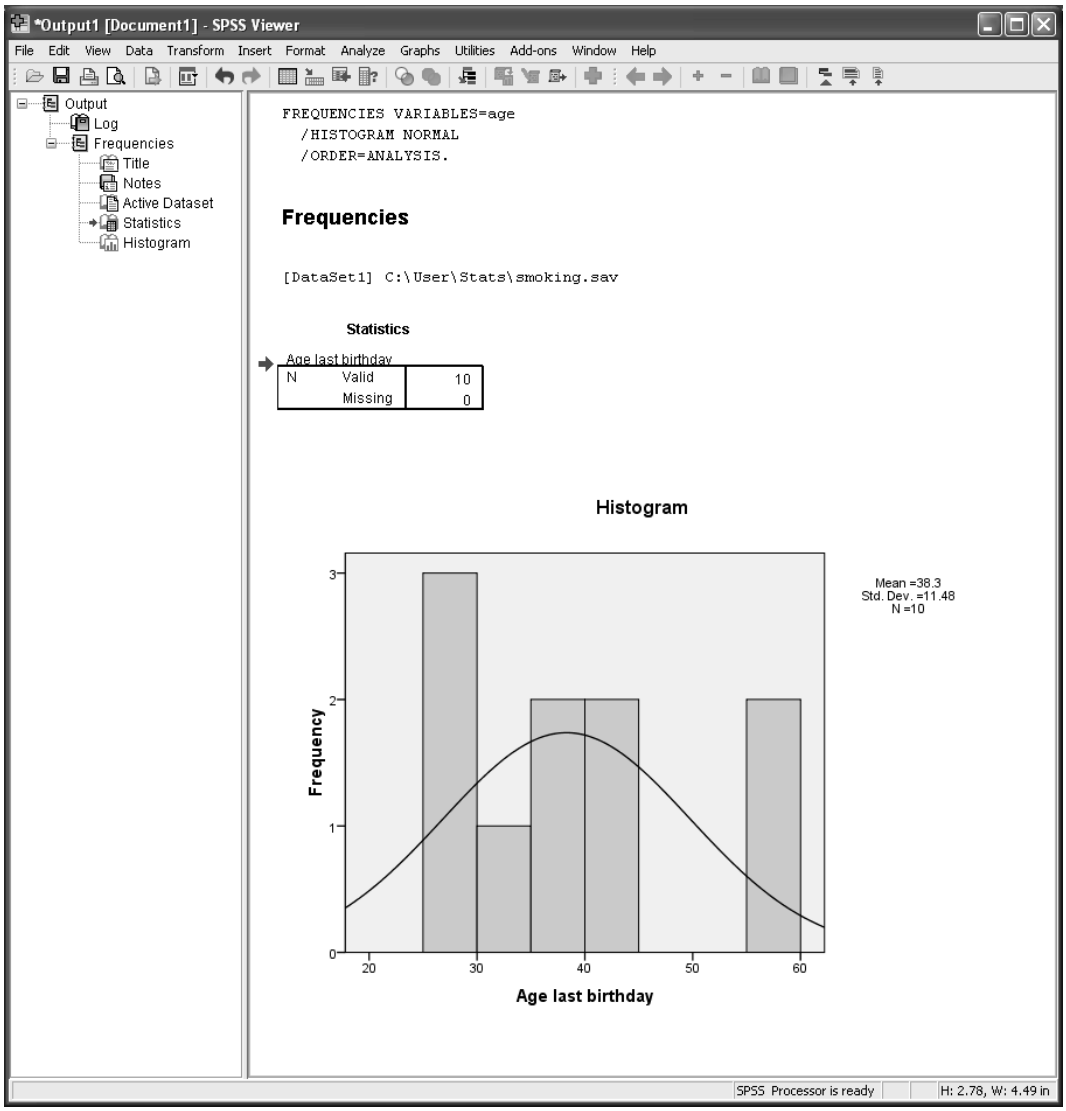


Figure 31 - SPSS viewer output produced using the histogram option from the Frequencies procedure

Task 17 Crosstabulation

The data may be broken down further with crosstabulation (multi-way) tables, which show the joint distribution of two variables' values. If you want to know how many women are smokers, the two numerical variables, **sex** and **smoker** are used in Crosstabs.

17.1 To get a crosstabulation:

- Select **Analyze**.
- Select **Descriptive Statistics**.
- Select **Crosstabs**.
- Select **Do you smoke? [smoker]** from the source variable list.
- Click ➥ adjacent to the **Row(s)** text box.
- From the source variable list select **Sex of respondent [sex]**.
- Click ➥ adjacent to the **Column(s)** text box.
- To see the crosstabulation click **OK**.

SPSS produces a crosstabulation of **smoker** by **sex**. The cells of the table show the **Counts**. You should find that all five women are smokers and only one man is a smoker.

Adding cell percents and the chi-square statistic

17.2 Select **Analyze**.

- Select **Descriptive Statistics**.
- Select **Crosstabs**.

The table contents can be changed by clicking on the **Cells** button and specifying options. Some useful options are:

Expected	prints expected values
Row	includes row percentages
Column	includes column percentages

17.3 Modify the **Crosstabs** table to request statistics and include the options **Row**, **Column**, **Total** and **Expected** as follows:

- Click the **Cells** button.
- Select the additional options **Expected**, **Row Percentages**, **Column Percentages** and **Total Percentages**.
- Click **Continue**.

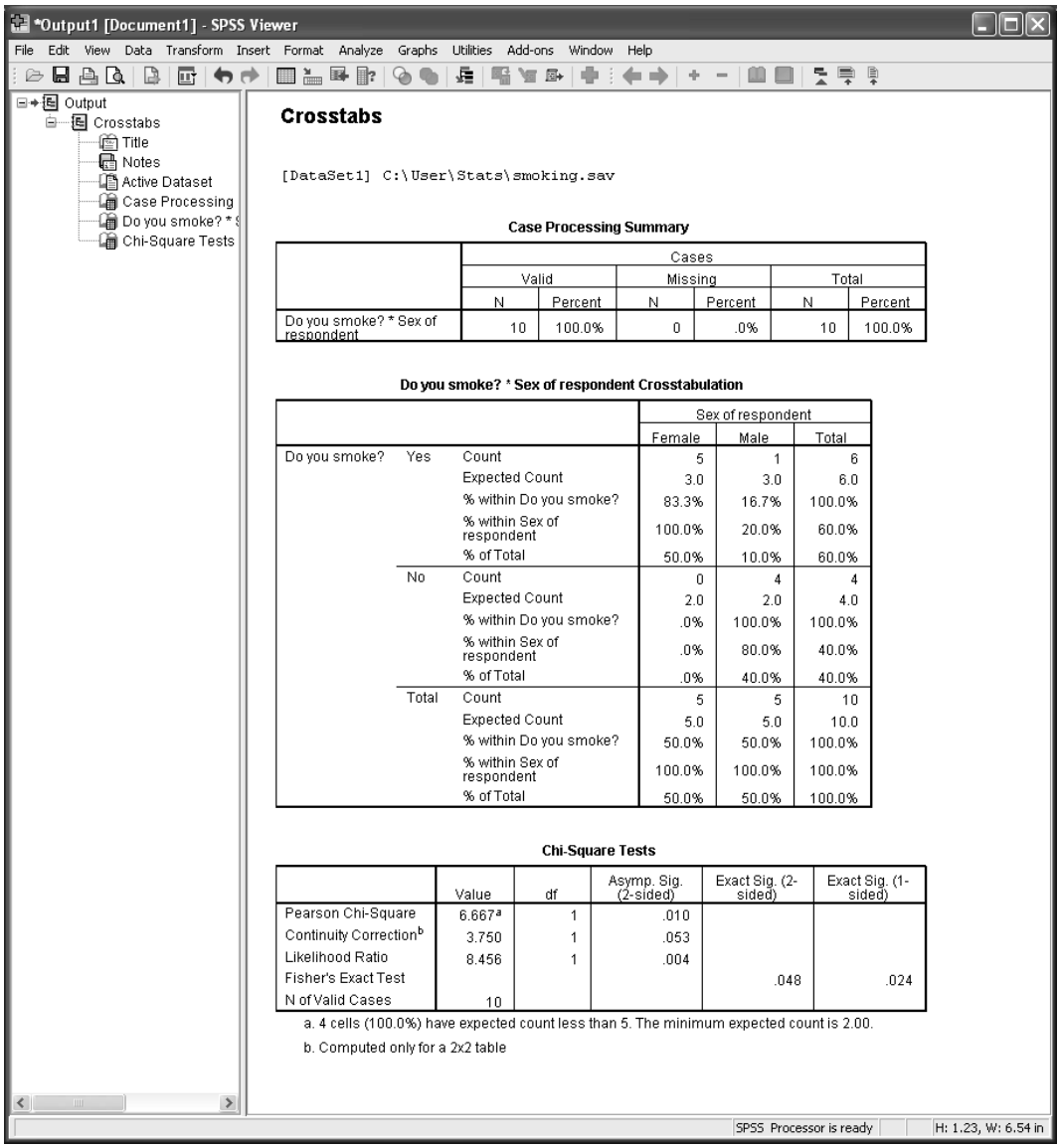
The **Statistics** button in the **Crosstabs** window requests statistics.

- Click the **Statistics** button.

Chi-square requests a Chi-Square (χ^2) test of independence and a Fisher's Exact test when there are fewer than 20 cases in a 2 x 2 table.

- Select the **Chi-square** option.
- Click **Continue**.

- Click **OK** (see the results in Figure 32).



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Do you smoke? * Sex of respondent	10	100.0%	0	.0%	10	100.0%

Do you smoke? * Sex of respondent Crosstabulation

			Sex of respondent		
			Female	Male	Total
Do you smoke?	Yes	Count	5	1	6
		Expected Count	3.0	3.0	6.0
		% within Do you smoke?	83.3%	16.7%	100.0%
		% within Sex of respondent	100.0%	20.0%	60.0%
		% of Total	50.0%	10.0%	60.0%
	No	Count	0	4	4
		Expected Count	2.0	2.0	4.0
		% within Do you smoke?	.0%	100.0%	100.0%
		% within Sex of respondent	.0%	80.0%	40.0%
		% of Total	.0%	40.0%	40.0%
	Total	Count	5	5	10
		Expected Count	5.0	5.0	10.0
		% within Do you smoke?	50.0%	50.0%	100.0%
		% within Sex of respondent	100.0%	100.0%	100.0%
		% of Total	50.0%	50.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.667[a]	1	.010		
Continuity Correction[b]	3.750	1	.053		
Likelihood Ratio	8.456	1	.004		
Fisher's Exact Test				.048	.024
N of Valid Cases	10				

a. 4 cells (100.0%) have expected count less than 5. The minimum expected count is 2.00.

b. Computed only for a 2x2 table

Figure 32 - SPSS viewer output from Crosstabs procedure (with Chi-squared tests)

Suppose that you wish to test the hypothesis: 'Men and women are not equally likely to smoke' against the alternative: 'Men and women are equally likely to smoke'. Since 100% of the cells have expected counts of less than 5, the Chi-Square test is not valid. However, the Fisher's Exact test (p=0.048 - statistically quoted as $p<0.05$) which is also given, lends support to the hypothesis that women and men in this sample are not equally likely to smoke.

Task 18 Clustered bar chart

18.1 A clustered bar chart can give a picture of the crosstabulation in Task 17:

- In the **Viewer** or **Data Editor** window select **Graphs**.
- Select Legacy Dialogs.
- Select **Bar** (for bar charts).
- To see the number of people who smoke or not, with separate bars for men and women click **Clustered** in the **Bar Charts** box.
- To see the counts of cases summarised in the bars click **Define** to specify the variables in the chart (see Figure 33).

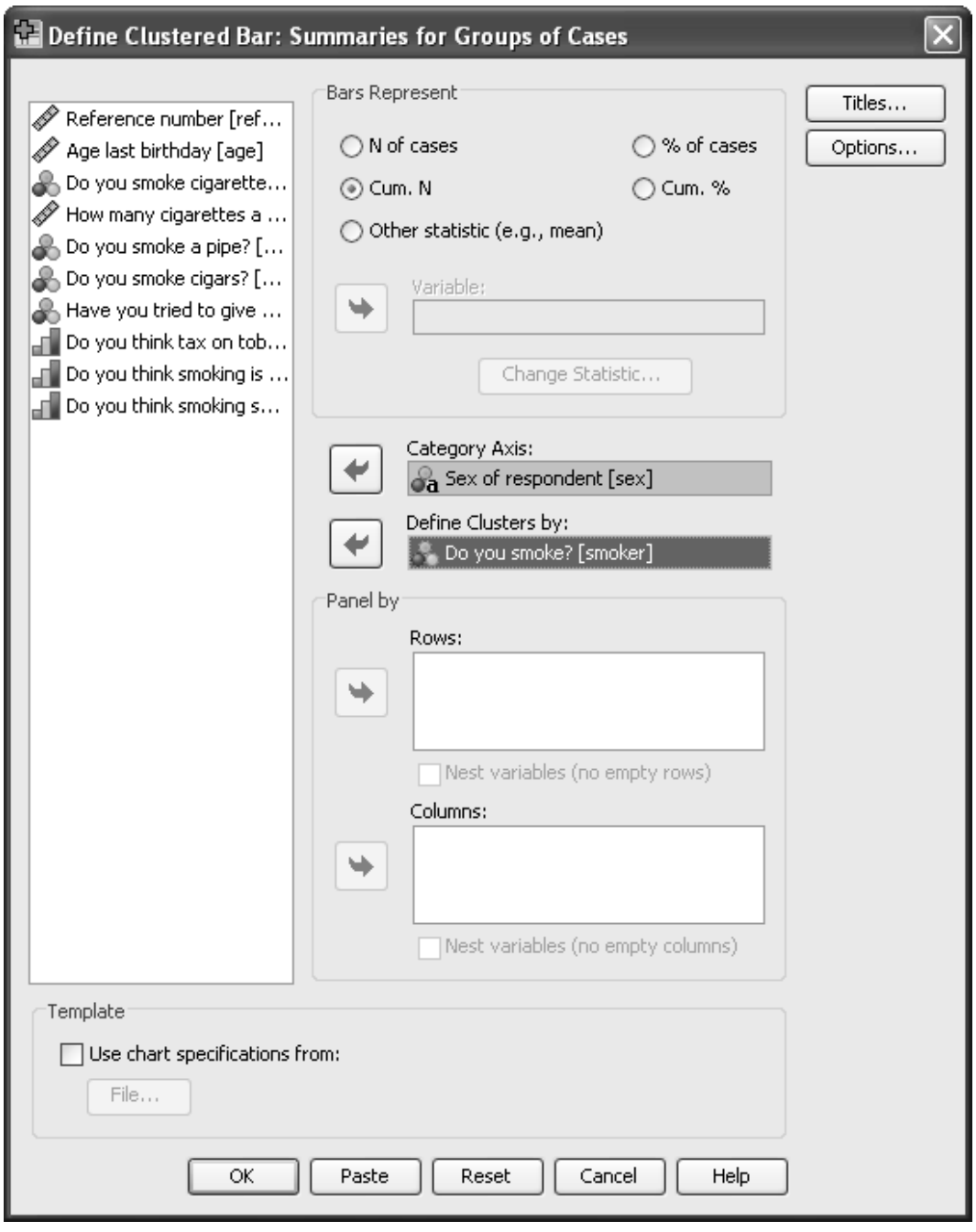


Figure 33 - the Define options window in the Clustered Bar Chart procedure

- Click **Cum. N** in the **Bars Represent** panel.
- From the source variable list select **Sex of respondent [sex]**.
- Click ➥ to place **sex** underneath **Category Axis**.
- From the source variable list select **Do you smoke? [smoker]**.
- Click ➥ adjacent to **Define Clusters by**.
- Click **OK**.

18.2 Is the graph what you expected? Look carefully at the right hand cluster of bars. (It should NOT look like the graph below in Figure 34)

Try re-running this procedure but this time in the **Define** section select **N of cases** in the **Bars Represent** panel. (You should now get a graph like Figure 34). See that you understand the difference between the two graphs you have produced using SPSS.

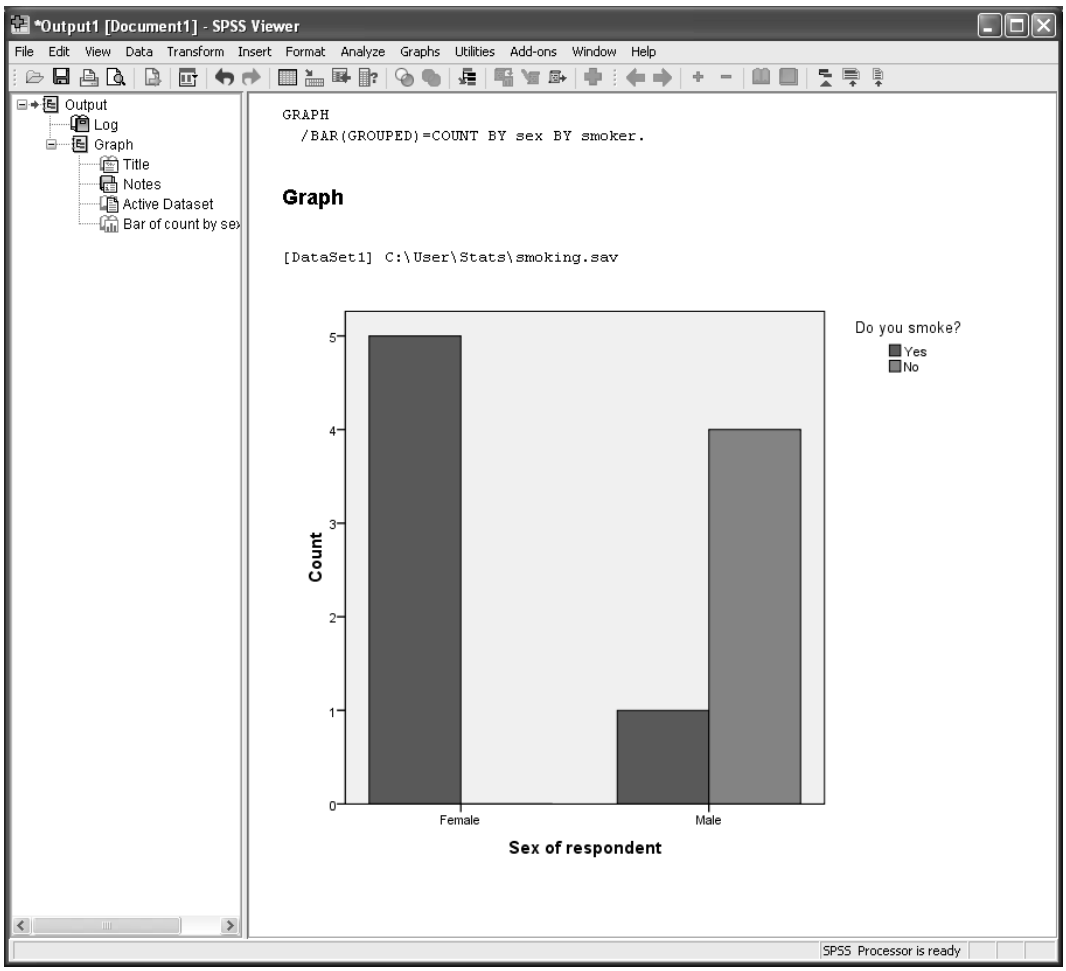


Figure 34 - SPSS viewer output from Clustered bar chart

18.3 Use options to include error bars on the bar chart

There are three different options for error bars – if you are not sure which to use please ask a statistician, all are valid options.

Task 19 Analysing data in subgroups - Split File

A procedure can produce analyses for separate subgroups rather than for all the data if the **Split File** option is selected before a procedure is carried out. The variable (or variables) that define the subset are given in the **Split File** command. For example, you may want to run descriptives first for the women, and then for the men and thus produce separate mean values of age for men and women.

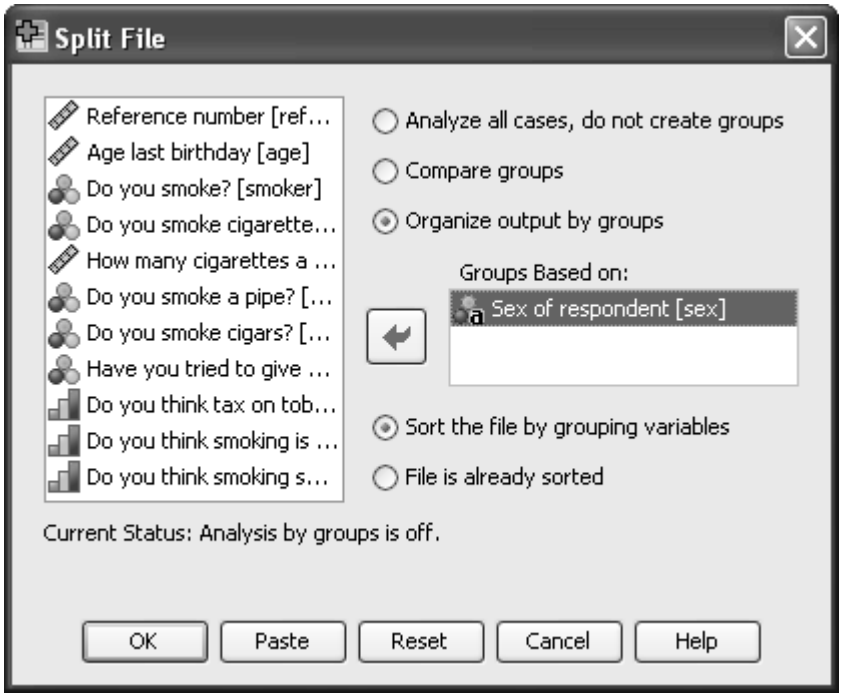


Figure 35 - the Split File option box

19.1 Select **Data**.

- Select **Split File** (see Figure 35).
- Select **Organise output by groups**.
- Move **Sex of respondent [sex]** to the **Groups Based on** box.
- Click **OK**.

Note **Split File** sorts your data. In the **Data Editor**, observations are no longer in **Ref_no** order.

19.2 To see how **Split File** processing works, from the **Analyze** menu, select the **Descriptive Statistics** option to run **Descriptives** for **Age last birthday [age]**.

- Click **OK**.

  In the **Viewer** window, you should find that the mean age for women is 34.8 years and the mean age for men is 41.8 years.

  All subsequent analyses are done separately for each group.

19.3 To turn off the **Split File** option:

- Select **Data**.
- From the **Data** menu, click **Split File**.
- Click **Analyze all cases**.

- Click **OK** to select analysis for the whole dataset.

Task 20 Excluding observations – Select cases

A procedure can produce analyses for a specific group rather than for all the data if the **select cases** option is selected before a procedure is carried out. The variable (or variables) that define the group retained for the analysis are given in the **Select Cases** command. For example, you may want to exclude all the data from people who do not smoke when looking at some of the smoking questions.

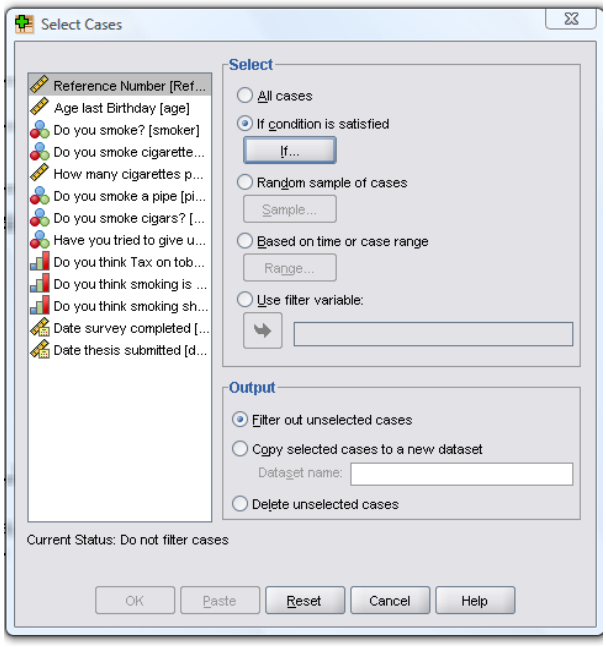


Figure 36 - the Select Cases option box

20.1 Select **Data**.

- Select **Select cases** (see Figure 36).
- Click on **If condition is satisfied**
- Choose **Filter out unselected cases** Choosing the option **delete unselected cases** will permanently delete the other cases and is **not recommended**!
- Click on the **If** under the words **If condition is satisfied**
- Choose **Do you smoke** from the list of variables
- Click the arrow to put the variable in the top right box
- The word smoker should appear in the window
- Type **=1** after the word smoker (see Figure 37)
- Click on **Continue**.
- Click on **OK**

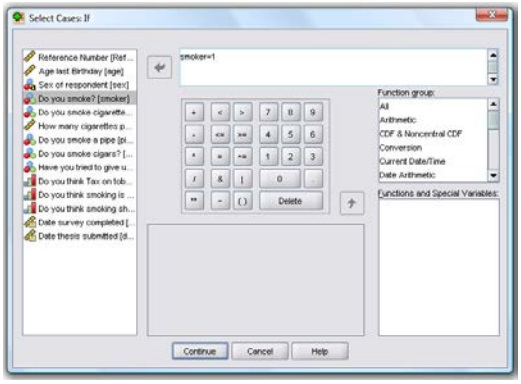


Figure 37 -the Select cases If option box

20.2 To see how **Select Cases** processing works, from the **Analyze** menu, select the **Descriptive Statistics** option to run **Frequencies** for the two smoking variables **Do you smoke cigarettes** and **have you tried to give up smoking**.

- Click **OK**.

  In the **Viewer** window, you should find that 83% of the smokers smoke cigarettes and two thirds of them have tried to give up smoking.

  All subsequent analyses are done separately for each group.

20.3 To turn off the **Select Cases** option:

- Select **Data**.
- From the **Data** menu, click **Select Cases.**
- Choose **All Cases**
- Click on **OK**.

Notes for advanced SPSS use

If you have identified an individual or group of individuals which should be excluded from the analysis for a paper you are intending to publish, save an account of those you are intending exclude. Use the paste function to create the syntax that you have used. Add text to the syntax file above the select statement to remind you why they have been excluded.

Task 21 Modifying variables

Recoding values into a new variable

Sometimes you might want to use data in a different form, such as looking at the age groups young and old (eg 30 and under, 31 and above) rather than exact age. For instance, the values of **age** could be recoded to 1 and 2, representing say, young (1) and old (2).

21.1 You can create a new variable to hold the new recoded values and preserve the original values:

- Select **Transform**.
- From the **Transform** menu click **Recode Into Different Variables**.

  A dialog box opens up where you can select the variable you want to recode (Figure 38).

21.2 To recode **age** to **agecat** (age category):

- Click **Age last birthday [age]** in the source variable list.
- Click ➥ to transfer **age**.

  Give an **Output Variable Name**, eg, ***agecat***.
- Click in the **Label** text box and type ***Age category***.
- Click **Change** to register the new output variable.

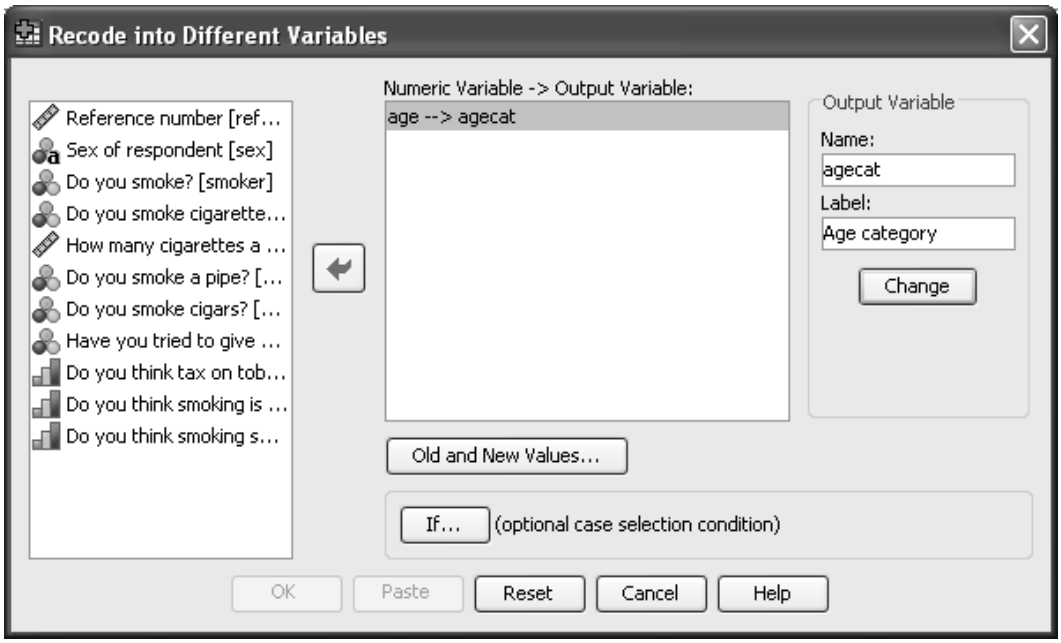


Figure 38 - the Recode submenu options window

Once you have decided which variables to recode you need to specify the old and new values.

21.3 Select the **Old and New Values** box (Figure 39).

To create a **young** (to be called **1**) and an **old** category (to be called **2**), recode the values 0 through to 30 into the new value 1:

- Click on the **Range, LOWEST through** button.
- Type ***30*** in the text box.

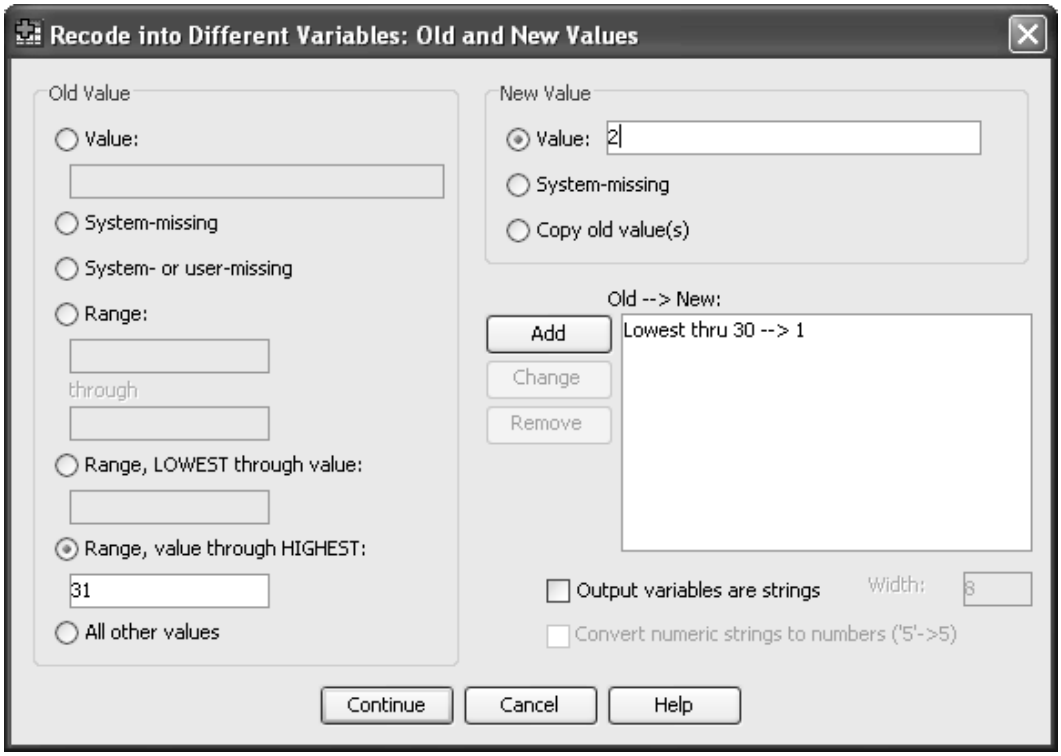
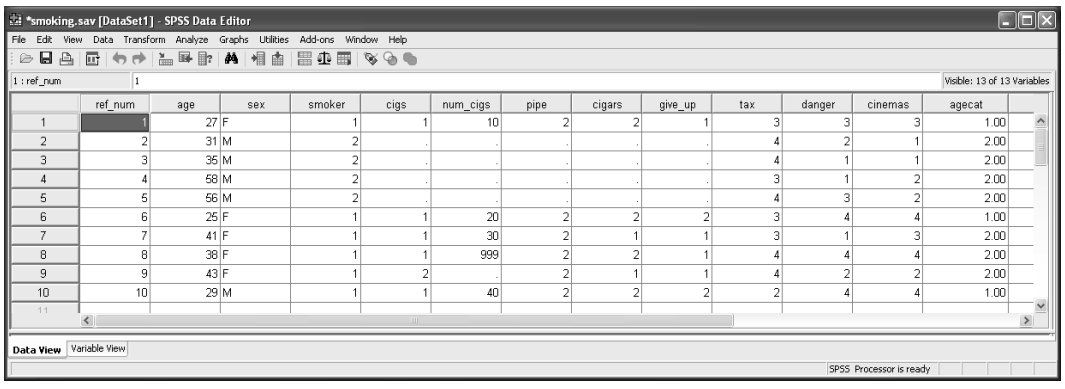


Figure 39 - the Old and New values options window in Recode

- Click **Value** in the **New Value** area.
- Type **1** in the **New Value** text box.
- Click **Add** to place this specification in the **Old-> New** text box.
- Repeat to recode values (ages) 31 and above to new value **2**, using **Range, value through HIGHEST**. Remember to click **Add**.
- Click **Continue** to return to the **Recode into Different Variables** window.
- Click **OK** to perform the recoding operation.

SPSS adds the new variable **agecat** to the next empty slot in the **Data Editor** window. Look at the right-most column in the **Data View** pane of the window (Figure 40). Also look at the **Variable View** pane to see that **agecat** has been added as variable number 13. Whilst in the variable view panel you might like to add the value labels for this new variable **agecat** (see Task 5 if you have forgotten how).



	ref_num	age	sex	smoker	cigs	num_cigs	pipe	cigars	give_up	tax	danger	cinemas	agecat
1	1	27	F	1	1	10	2	2	1	3	3	3	1.00
2	2	31	M	2	.	.	.	.	.	4	2	1	2.00
3	3	35	M	2	.	.	.	.	.	4	1	1	2.00
4	4	58	M	2	.	.	.	.	.	3	1	2	2.00
5	5	56	M	2	.	.	.	.	.	4	3	2	2.00
6	6	25	F	1	1	20	2	2	2	3	4	4	1.00
7	7	41	F	1	1	30	2	1	1	3	1	3	2.00
8	8	38	F	1	1	999	2	2	1	4	4	4	2.00
9	9	43	F	1	2	.	2	1	1	4	2	2	2.00
10	10	29	M	1	1	40	2	2	2	2	4	4	1.00

Figure 40 - SPSS Data Editor window with extra variable 'agecat' added

21.4 Use **Frequencies** to tabulate **agecat** and check the results. You should find that 3 subjects are aged 30 or less and 7 subjects are age 31 or more.

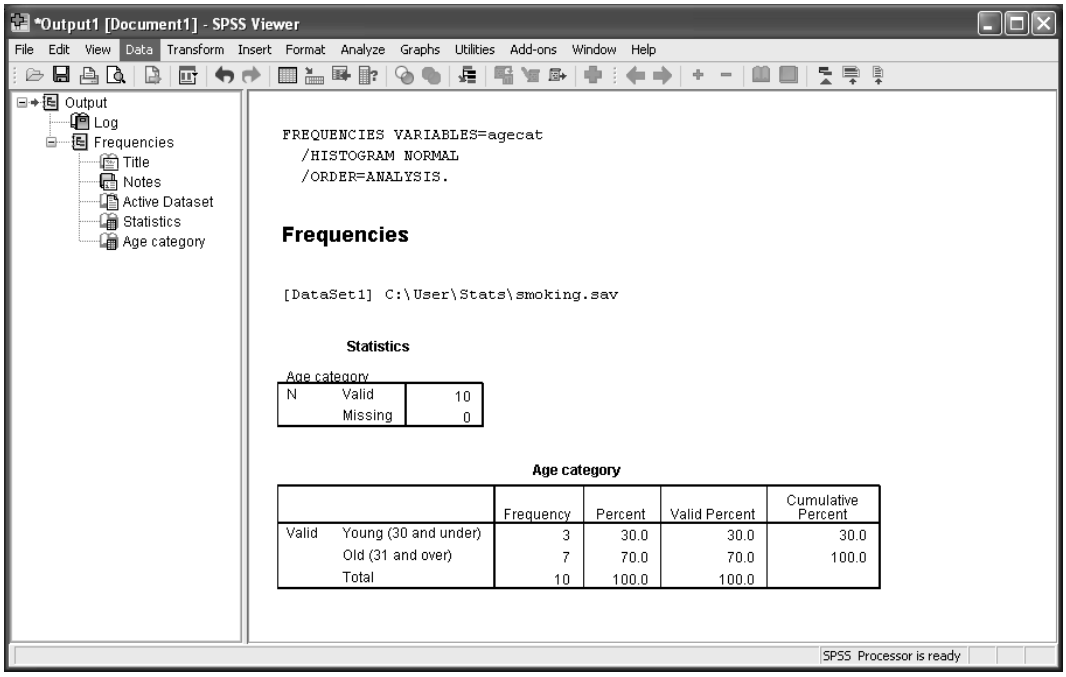


Figure 41 - SPSS viewer output from recoded variable 'agecat'

Notes for advanced SPSS use

If you are intending to publish this work at some point you will need a permanent record of the recoding statements that you have carried out. Use the Paste function to save a record to your file **Analysis for Paper 1.sps**

You may find this easier to do all your recoding using Syntax, once you learn how to alter the syntax successfully.

Computing new variables

Compute commands are so-called because they assign values to variables. They may be used, for example, when data are not in the form required. For example, the log of age may be required for analysis rather than actual age. The SPSS **Data Editor** window contains the **Transform** menu with the **Compute** option.

21.5 To create a new variable firstly choose a name for the variable, eg **logage**.

To calculate the log of **age**:

- Select **Transform**.
- Select **Compute**.
- Type ***logage*** in the **Target Variable** box (Figure 42).

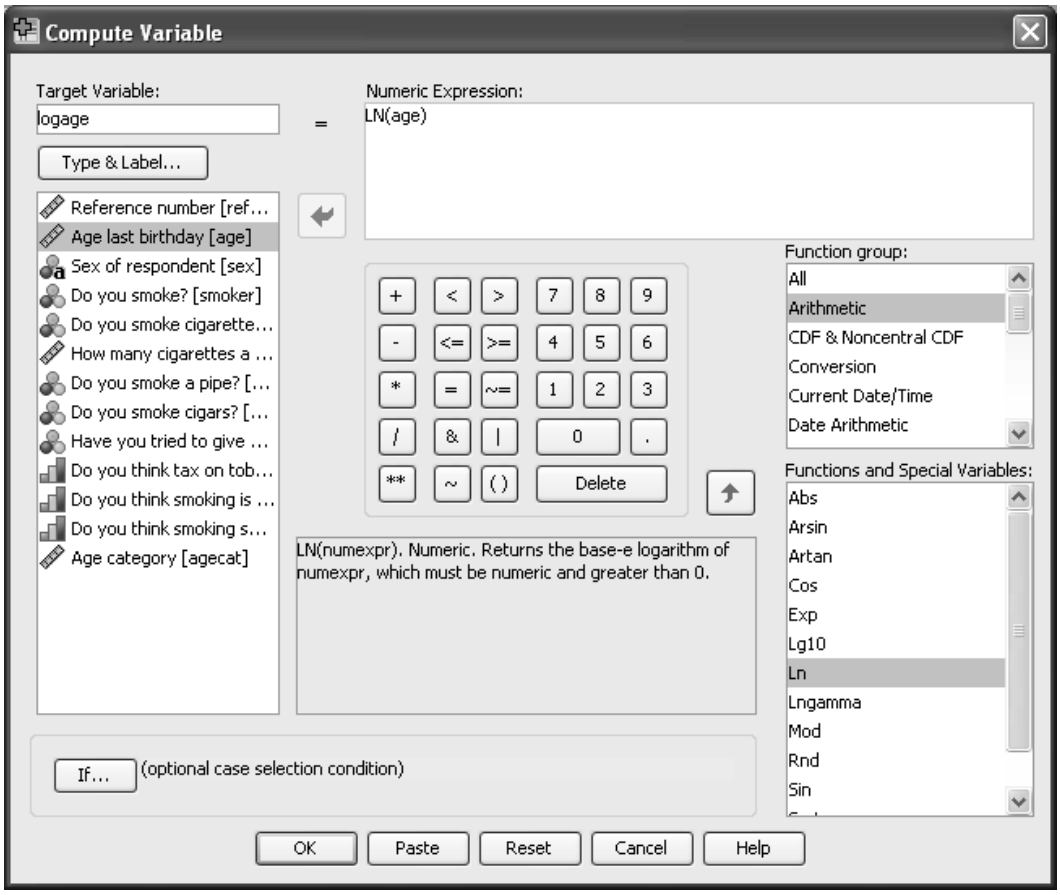


Figure 42 - the Compute Variable options window

Next, in the **Numeric Expression** text box you need to give the instructions for calculating the variable, in this case taking the log of **age**.

21.6 Initially you need to find the appropriate function. Log is an arithmetic function:

- Click on **Arithmetic** in the **Function group** window.
- Scroll down and click on **Ln** in the **Functions and Special Variables** window.
- Click the up arrow next to the **Function group** box.

Note A **Function group** must be highlighted as well as a **Function** otherwise the up arrow alongside the functions group does not operate.)

- Select **Age last birthday (age)**.
- Click ➥ to move **age** to replace the **?** in the function definition.

  Now you have a compute statement: **logage=LN(age)**.
- Click **OK**.

  SPSS creates the new variable and places it in the next free column in the **Data Editor** window. The newly created variable **logage** may now be used in SPSS procedures.

Using the IF Statement to Compute New Variables

The If statement can be used in conjuction with the Compute command you have just used to assign values to variables for a specific group of observations. They may be used, for example, when you are required to combine the results from

two or more variables. For example, you might want to identify the group of people who are smokers who strongly agree with the statement that smoking is dangerous to their health. The SPSS **Data Editor** window contains the **Transform** menu with the **Compute** option and the **If** box.

21.7 To create a new variable firstly choose a name for the variable, eg **dangersmok**.

We are going to give those who smoke and think it dangerous the value 1 for **dangersmok**:

- Select **Transform**.
- Select **Compute**.
- Type ***dangersmok*** in the **Target Variable** box (Figure 42).

Next, in the **Numeric Expression** text box you need to give the instructions for calculating the variable, in this case just type **1**.

21.8 Now you need to go to the **If** box so that you can say who it is who are going to have the value of **1**:

- Click on **IF**.
- Choose the lower option of **"Include if case satisfies condition:"**
- Put the variable smoke into the top window and type **=1 and** and then add the variable danger by clicking on "**Do you think smoking is dangerous?**". Then type **"=4**". The result should look like the figure bellow.

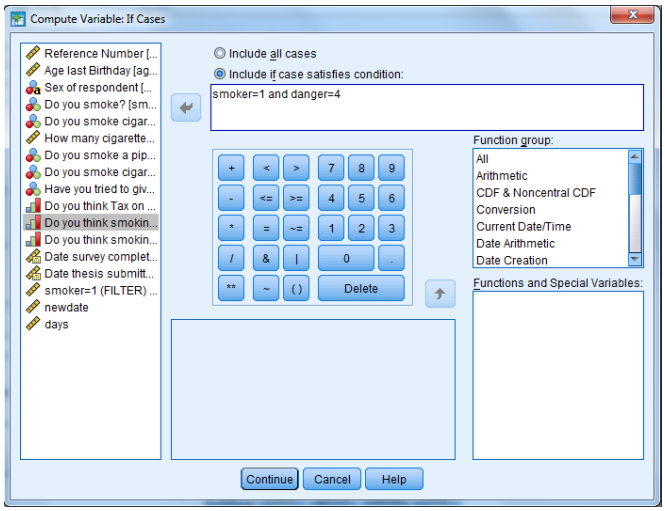


Figure 43 - The If box from the compute variable window

Note This will only compute 1 for those people who have both the features smoking =1 and danger=4. This means that all those who do not have both these functions will have a missing value for this new variable. If you would like to give the other people a value such as zero, you have to go through this procedure a second time, this time choosing **0** as the value you are calculating, and also using the statement "**smoke=2 or danger ne 4"** Here **ne** can be used by SPSS to mean not equal to, you can also use the button provided which has the symbol **~=.** Don't forget to use the **or** option to join the conditions together in the second statement instead of **and**.

- Click **OK**.

SPSS creates the new variable and places it in the next free column in the **Data Editor** window. To retain the syntax open the compute window back up and press paste.

Task 22 Working with Dates in SPSS

Dates can be manipulated in excel prior to importing the data, alternatively SPSS handles dates in a similar way to that used by excel. SPSS holds a date as the number of seconds since a particular date some time in 1582. Thus when two dates are subtracted from each other you need to divide by 60*60*24 to get back to the number of days.

We can use the two date functions here to ask how many days after the subject submitted their thesis they filled in the questionnaire.

- Select **Transform**.
- Select **Compute**.
- Type ***days*** in the **Target Variable** box (Figure 42).

Next, in the **Numeric Expression** text box you need to give the instructions for calculating the variable, in this case calculating the number of ***days***.

- Enter the expression:

(date_survey - date_submit)/(60*60*24)

The variables can be entered by typing their variable name **or** by selecting them from the list of variables.

- Click **OK**.

SPSS creates the new variable and places it in the next free column in the **Data Editor** window. The newly created variable **days** may now be used in SPSS procedures.

Important

Save this syntax using paste from the Figure 37 window you will need it for a future task 26

- Check the variable **days** by using the **frequency** procedure and asking for a **histogram** from the **charts** options.

Task 23 Correlations

The correlations procedure calculates the (Pearson parametric) correlation between variables and is used to measure the strength of linear association between 2 variables.

23.1 To obtain the Pearson correlation coefficients of **tax**, **danger** and **cinemas**:

- Select **Analyze**.
- Select **Correlate**.
- Select **Bivariate** (see Figure 44).

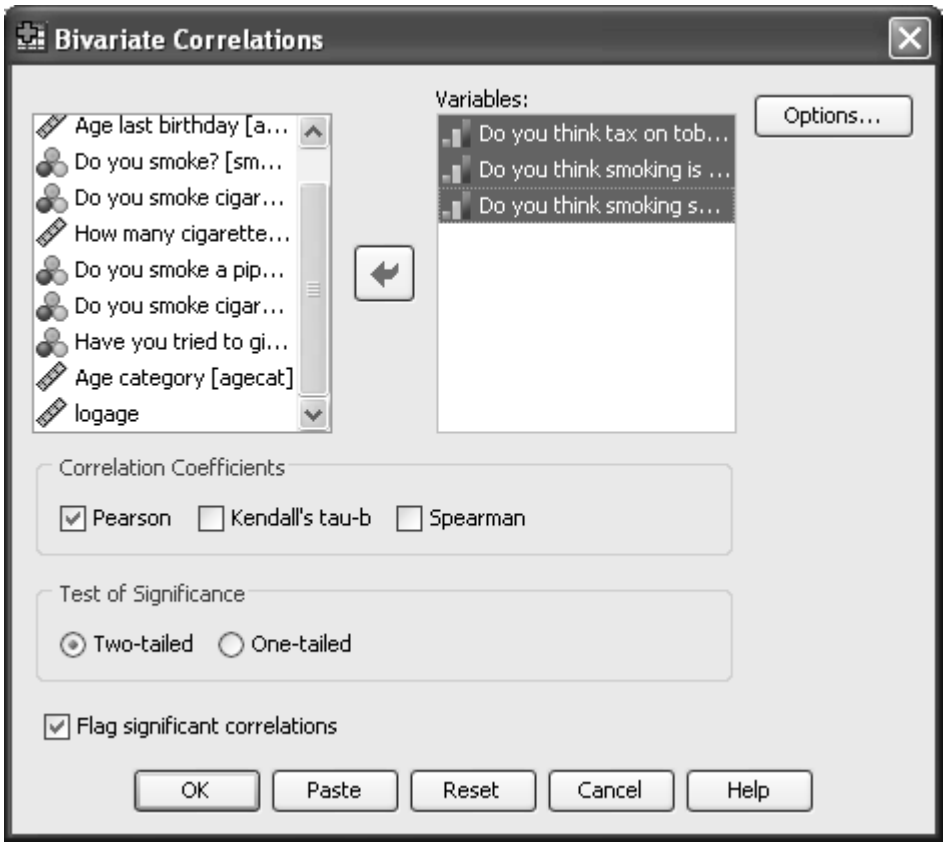


Figure 44 - the Bivariate Correlations options window

- Move the following to the **Variables** box:
  - **Do you think tax on tobacco too high? (tax)**
  - **Do you think smoking is dangerous to your health? (danger)**
  - **Do you think smoking should be allowed in cinemas? (cinemas)** to the **Variables** box.
- Click **Flag significant correlations** to put a tick in the box. (It may already be ticked)
- Click **OK**.

There is a significant correlation between **danger** and **cinemas** (0.746, $p<0.05$), which means that people who feel that smoking is dangerous also tend to think that smoking should not be allowed in cinemas.

Since the three variables in this correlation analysis are ordinal in measurement scale, it would have been statistically more proper to use Spearman's Rank correlation coefficients to measure the relationship.

23.2 Repeat the above analysis but this time select the **Spearman** correlation coefficient. Note that the same relationship remains significant, although its value is slightly lower than the Pearson correlation (Figure 45).

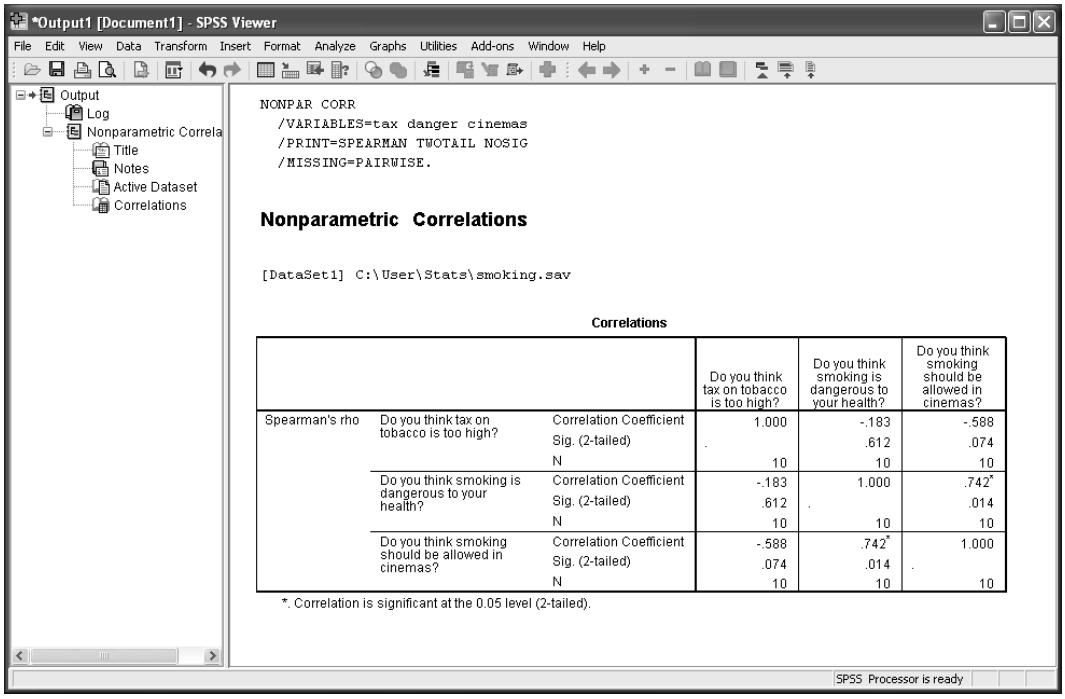


Figure 45 - SPSS viewer output from Spearman's Rank correlation

Task 24 Creating charts - drawing a scatter plot

A scatter plot graphs one variable against another, and often gives an idea of any associations in the data. In this example, we will plot **danger** along the horizontal (X) axis and **cinemas** along the vertical (Y) axis.

24.1 To use the **Graphs** menu to specify the scatter plot:

- Click **Graphs** and Legacy Dialogs.
- From the menu select **Scatter/Dot**.
- To specify the scatter plot click **Simple Scatter**.
- Click **Define** in the **Simple Scatterplot** dialog box (see Figure 46).

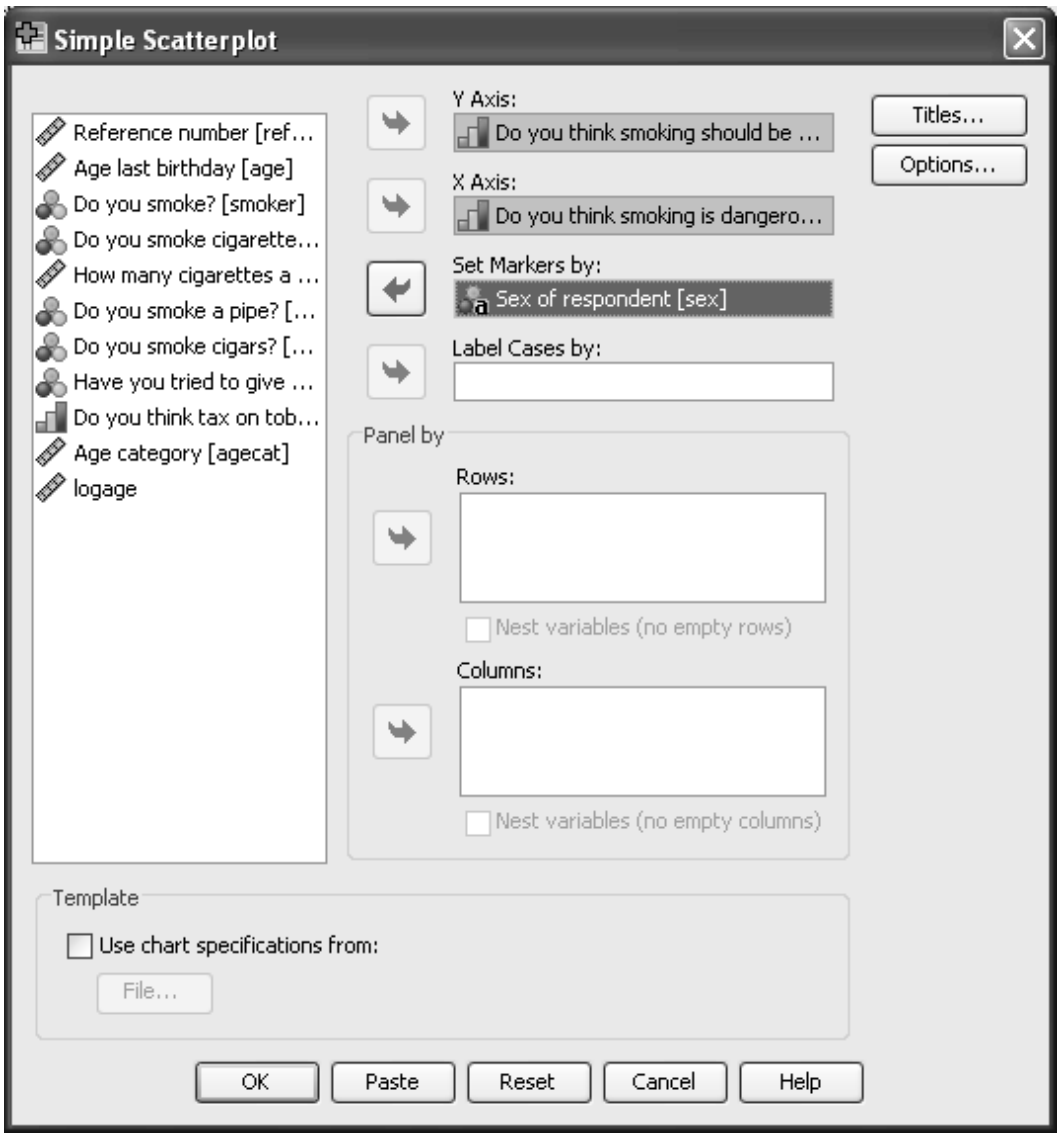


Figure 46 - the Scatterplot design window from Graph Scatter/Dot procedure

- Click on **Do you think smoking should be allowed in cinemas? [cinemas]** in the source variable list.
- Click ➥ to place **cinemas** in the **Y Axis** text box.

- Click on **Do you think smoking is dangerous to health? [danger]** in the source variable list.
- Click ➥ to place danger in the **X Axis** text box.

  Using a variable such as **sex**, we can also see if there is a difference between the men and women.
- Click **Sex of respondent [sex[** in the source variable list.
- Click ➥ to place **sex** in the **Set Markers by** box.
- Click **OK**.

  This produces a plot of **cinemas** versus **danger**; the variable **sex** is used to mark each observation on the plot as male and female. Where duplicate values occur, this may not be an accurate representation. Have a look at the data points for the data in Figure 47.

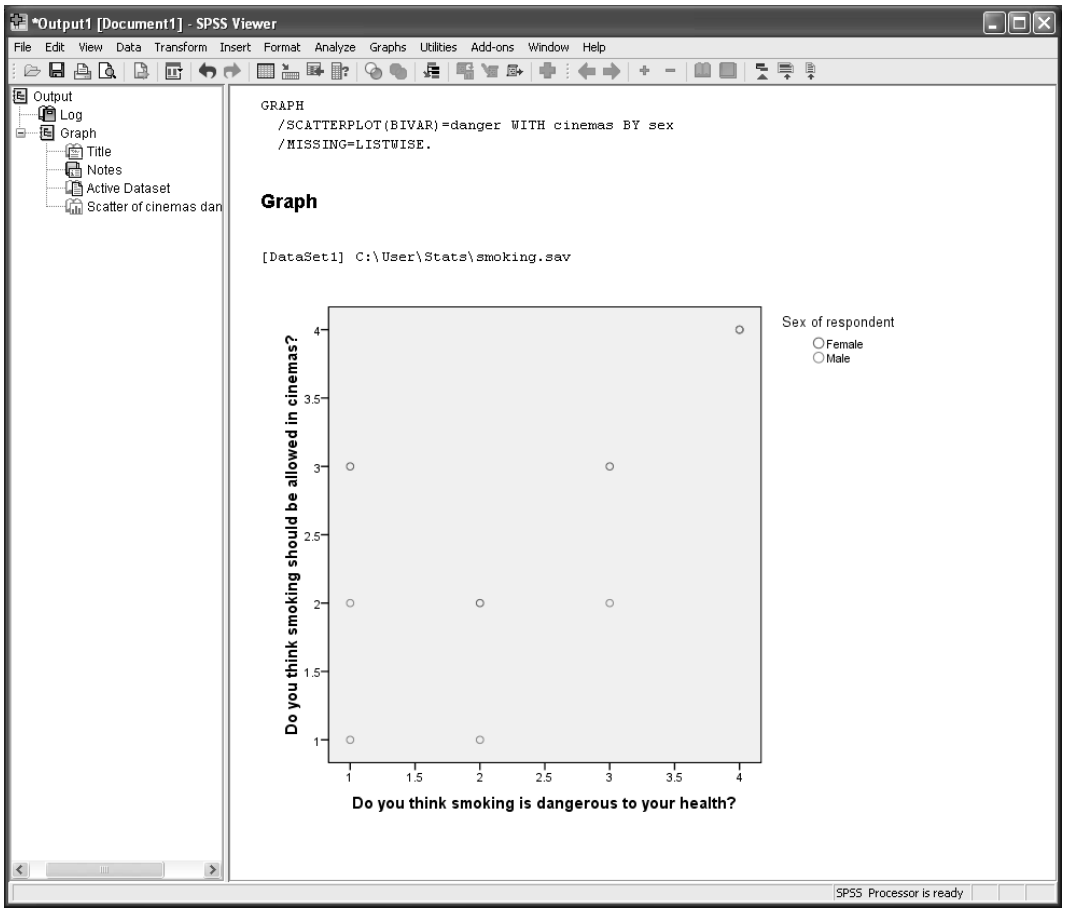


Figure 47 - the SPSS viewer output from Scatter/Dot in Graphs

Task 25 Saving an updated copy of the data

In order to save your recoded information and computed variables permanently, you must save the current copy of the data in the **Data Editor** window. For now, save the file under a new name, ***smoking2***. You can save the file under the same name or a different name.

25.1 To save the file:

- Click **File** in the **Data Editor** window.
- From the **File** menu click **Save As**.
- Ensure the directory in the **Save in** box is correct. If you are using one of the Computer Centre training rooms change the directory to **C:\Training\Stats**.
- Type ***smoking2.sav*** in the **File name** box.
- Click **Save**.

  An updated copy of the data is saved in **smoking2.sav**.

Task 26 Getting SPSS to read data from other spreadsheet formats e.g. Excel

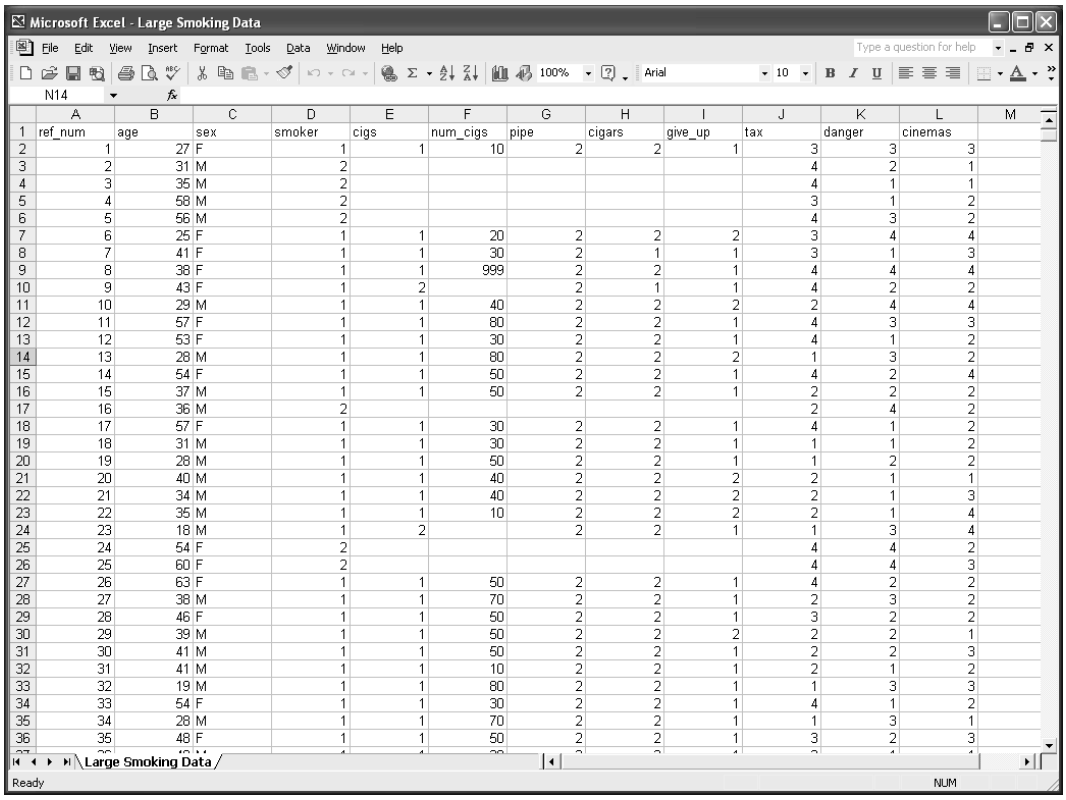
When working the Computer Centre training rooms you will notice that in the directory **C:\Training\Stats** there is a file called **Large Smoking Data.xls**. This is an **Excel** spreadsheet that has the same 12 variables as you have used so far in this workbook, but with many more cases than you have entered.

26.1 Browse to **C:\Training\Stats**

- Double click on the file name to enable Excel to open the file (Figure 48).

  You will notice that the first row of the Excel spreadsheet contains the names of the twelve variables exactly as you used them before. You will analyse this large file using SPSS.

- Close the Excel file at this point.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ref_num	age	sex	smoker	cigs	num_cigs	pipe	cigars	give_up	tax	danger	cinemas	
2	1	27	F	1	1	10	2	2	1	3	3	3	
3	2	31	M	2						4	2	1	
4	3	35	M	2						4	1	1	
5	4	58	M	2						3	1	2	
6	5	56	M	2						4	3	2	
7	6	25	F	1	1	20	2	2	2	3	4	4	
8	7	41	F	1	1	30	2	1	1	3	1	3	
9	8	38	F	1	1	999	2	2	1	4	4	4	
10	9	43	F	1	2		2	1	1	4	2	2	
11	10	29	M	1	1	40	2	2	2	2	4	4	
12	11	57	F	1	1	80	2	2	1	4	3	3	
13	12	53	F	1	1	30	2	2	1	4	1	2	
14	13	28	M	1	1	80	2	2	2	1	3	2	
15	14	54	F	1	1	50	2	2	1	4	2	4	
16	15	37	M	1	1	50	2	2	1	2	2	2	
17	16	36	M	2						2	4	2	
18	17	57	F	1	1	30	2	2	1	4	1	2	
19	18	31	M	1	1	30	2	2	1	1	1	2	
20	19	28	M	1	1	50	2	2	1	1	2	2	
21	20	40	M	1	1	40	2	2	2	2	1	1	
22	21	34	M	1	1	40	2	2	2	2	1	3	
23	22	35	M	1	1	10	2	2	2	2	1	4	
24	23	18	M	1	2		2	2	1	1	3	4	
25	24	54	F	2						4	4	2	
26	25	60	F	2						4	4	3	
27	26	63	F	1	1	50	2	2	1	4	2	2	
28	27	38	M	1	1	70	2	2	1	2	3	2	
29	28	46	F	1	1	50	2	2	1	3	2	2	
30	29	39	M	1	1	50	2	2	2	2	2	1	
31	30	41	M	1	1	50	2	2	1	2	2	3	
32	31	41	M	1	1	10	2	2	1	2	1	2	
33	32	19	M	1	1	80	2	2	1	1	3	3	
34	33	54	F	1	1	30	2	2	1	4	1	2	
35	34	28	M	1	1	70	2	2	1	1	3	1	
36	35	48	F	1	1	50	2	2	1	3	2	3	

Figure 48 - the first part of the Excel data file Large Smoking Data.xls

26.2 To input the Excel file:

- Go back into **SPSS 23.0 for Windows**.
- Select **Cancel** to the query window.
- From the **File** menu select **Open** and then select **Data**.
- Ensure the directory in the **Look in** box is correct. If you are in one of the Computer Centre training rooms change the directory to **C:\Training\Stats**.
- The **Files of type** window will be showing **SPSS**. This needs to be changed to **Excel** by clicking on the down arrow at the right-hand end of the **Files of Type** box

and selecting **Excel (*xls, *xlsx, *xlsm)**. The file name **Large Smoking Data** should now be visible (see Figure 49).

- Double click the file name.

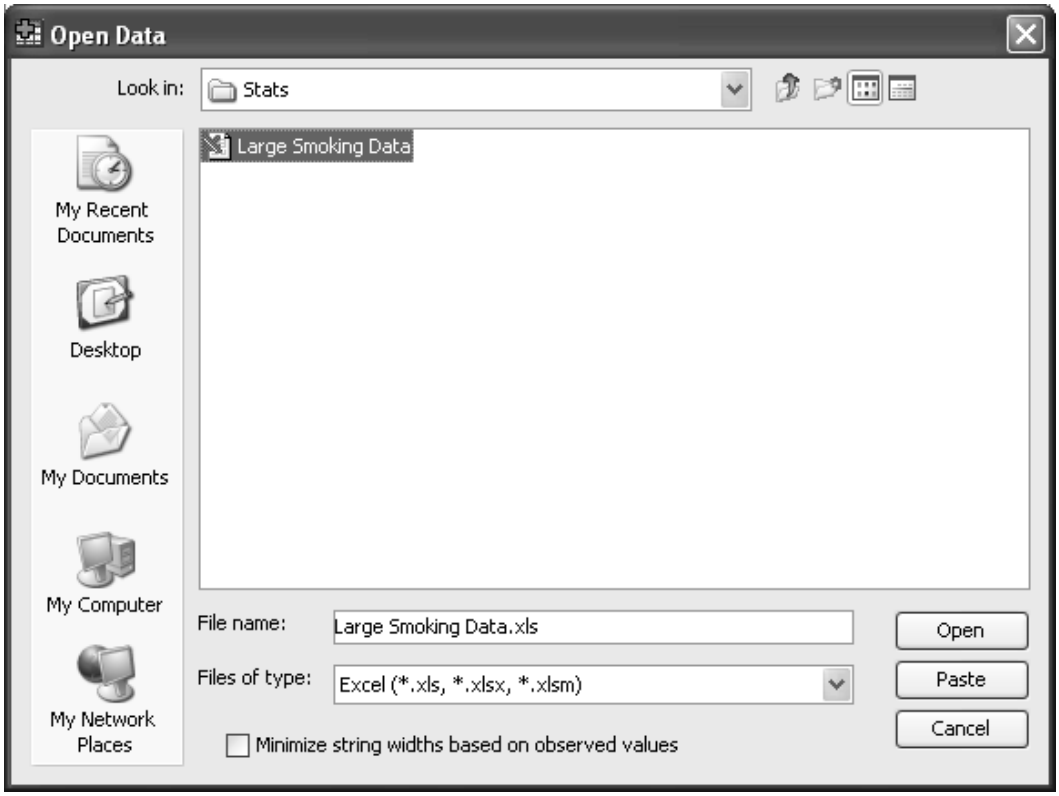


Figure 49 - the SPSS Open File menu window showing an Excel file for input

- The next menu confirms that the Excel file has been recognised. You should check that the box **Read variable names from first row of data** is ticked (Figure 50).
- Click **Continue**

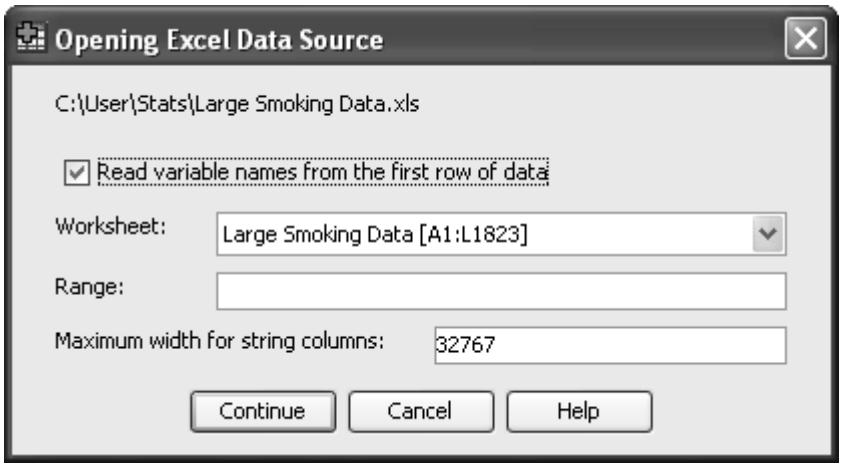


Figure 50 - the options window showing an Excel file for input

You now have a data file in SPSS with all the additional cases. Unfortunately all the information you entered into the **Variable View** part of your small original SPSS file is missing. You could re-type this information as you did earlier, but you can get SPSS to import the original.

26.3 In the **Data Editor** window click on **Data**.

- Click on **Copy Data Properties** (see Figure 51).

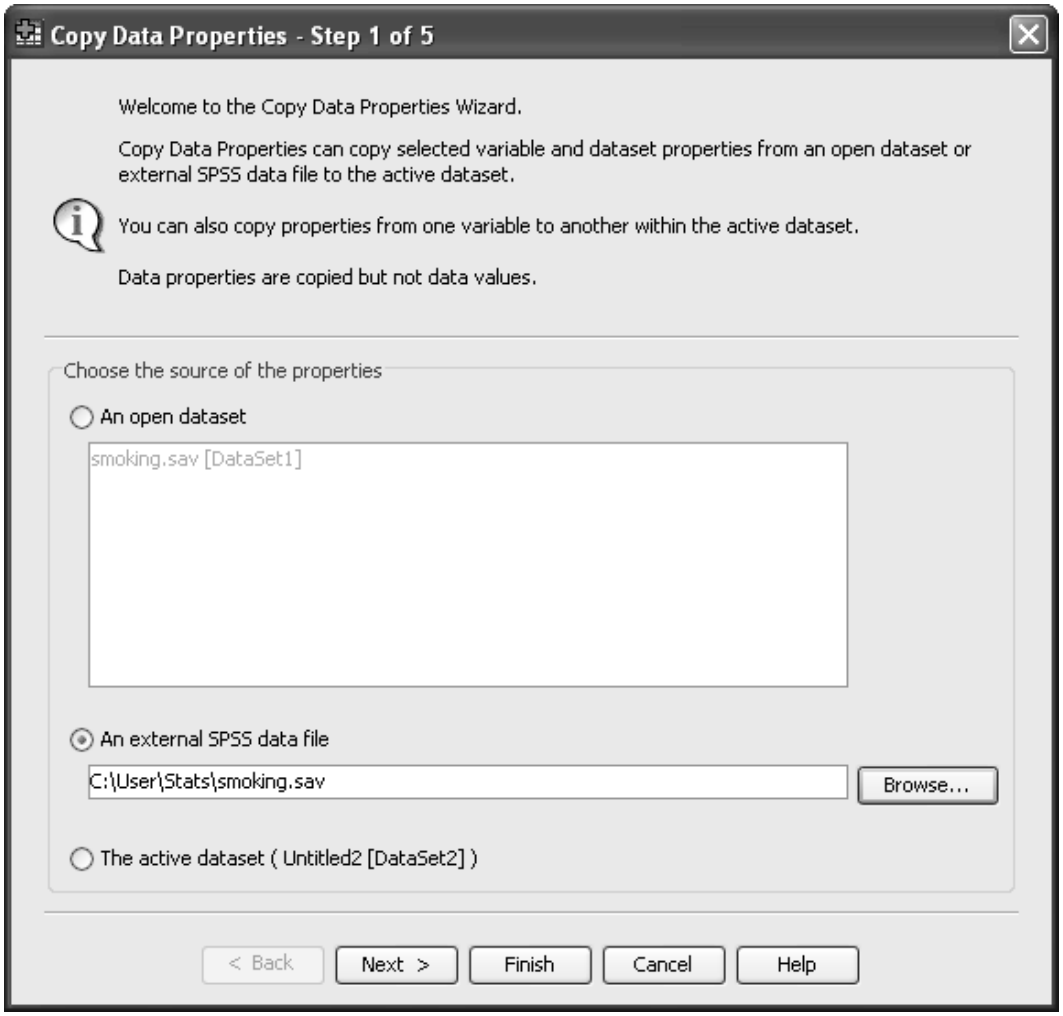


Figure 51 - the Copy Data Properties menu window

You wish to use the properties of the file **smoking.sav** which you saved earlier in C;\Training\Stats.

- Select **An external SPSS data file** and either by direct entry or browsing insert into the file name box ***C:\Training\Stats\smoking.sav***
- Click **Finish**. (If you were to click **Next**, you would have a number of alternatives offered that would allow you to control more finely what properties are copied)

26.4 Using procedures you mastered earlier, answer the following:- (keep your SPSS output for Task 24)

How many females answered the questionnaire? (Clue Task 12).

Show a histogram of respondent's ages? (Clue Task 15).

Are men and women equally likely to smoke? (Clue Task 16).

What is Spearman's rank correlation between **danger** and **cinemas**? (Clue Task 20).

Task 27 Saving output from SPSS into word processor documents e.g. Microsoft Word

The output file produced by SPSS can only be read by the SPSS program. This is complicated further by the fact that **versions from 15 onwards are unable to read the output from previous versions of SPSS**. Most users would like to take share their results and discuss them with other members of their research team, many of whom may not have access to SPSS, and if they do not quite the same version. Sharing results usually means transferring all or parts of their SPSS output into a Microsoft package. We shall learn how this can be achieved by moving output to a Microsoft Word document.

There are two ways of achieving this process. Let's first consider allowing SPSS to do most of the work.

27.1 Make sure you are in the **SPSS Viewer** window.

- From the **File** menu select **Export**.

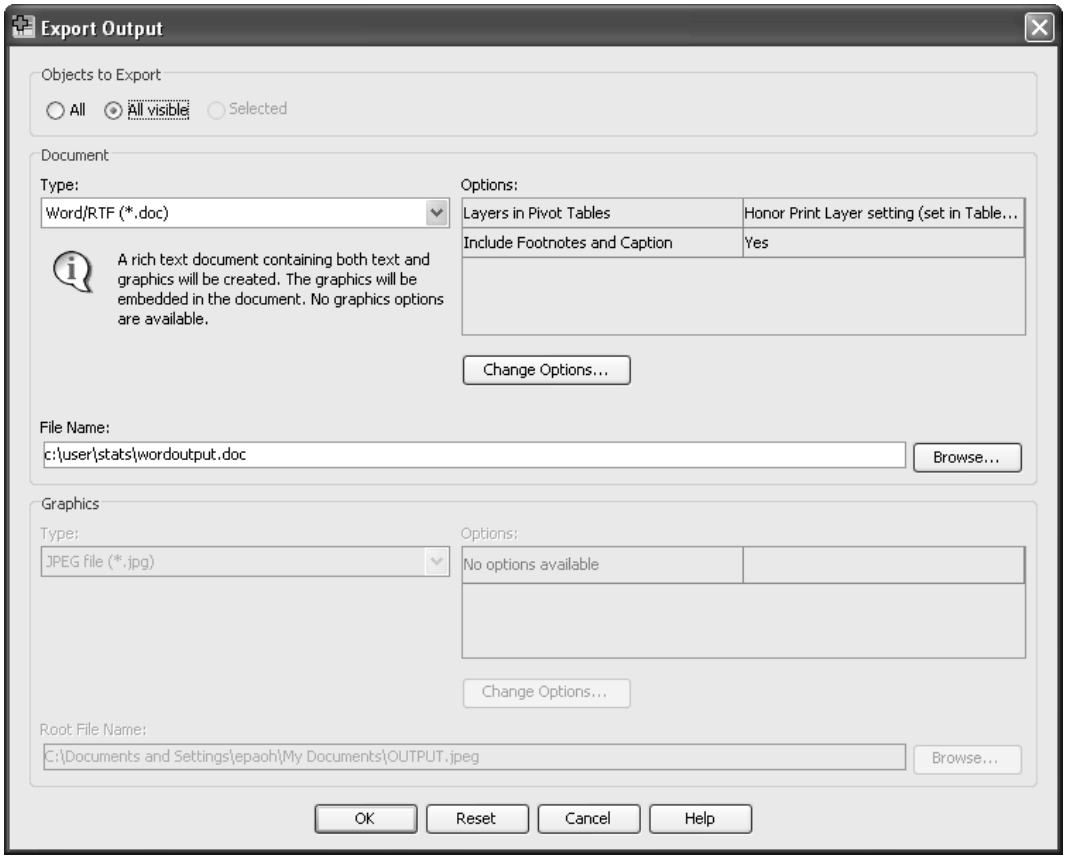


Figure 46 – the SPSS Export Output window

- In the **Objects to Export** box, ensure that **All Visible objects** is selected
- Under **Document Type** select **Word/RTF file (*.doc)** from the drop down menu.
- In the **File Name** box type ***C:\Training\Stats\wordoutput***.
- Click **OK**.

27.2 Minimise SPSS Viewer and Data Editor windows and then browse to and open the document **wordoutput.doc**. You will see that all your original output has been transferred.

27.3 There is an alternative way of achieving the movement of small amounts of output where you have much more control. Keep your **wordoutput** document open but position your cursor at the end of the present file. Minimise the Word document and maximise SPSS viewer file.

- Point and click on any single table or chart in your output. You will notice that a rectangular box appears around the table.
- Click on **Edit** and select **Copy.**
- Maximise your Word file and click on **Edit** and **Paste**.

  This process will work for transferring all text, tabular or chart output from SPSS.

Note **Never** copy and paste SPSS Graphs

Note **Always** export graphs before placing them in your Power point presentations. Failure to do this can make them **invisible** in the presentation if you are using a computer which does not have SPSS.

Task 28 Analysis – Box plot

Box plots are a good way to investigate continuous variables graphically. A continuous variable here is the number of days between handing in the thesis and completing the survey. Here we might want to see whether the smokers have submitted the survey in the same time as the non-smokers.

- Select Graphs
- Select Legacy dialogues
- Select Boxplot
- Select Simple and Summaries for groups of cases. See Figure 52

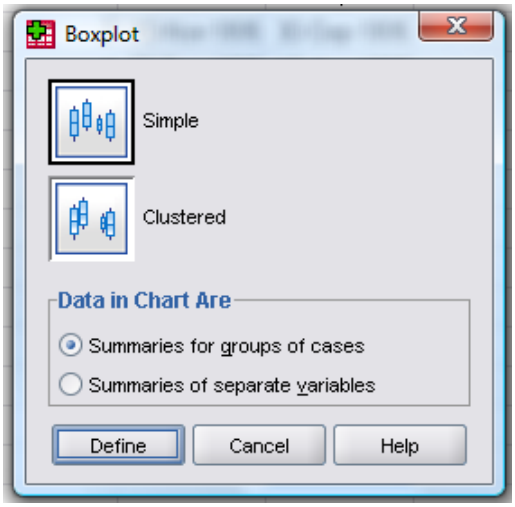


Figure 52 - box plot options

- Click on **Define**
- Select days and put it into the variable box
- Select Do you smoke (smoker) and put it into the Category Axis box (Figure 53).

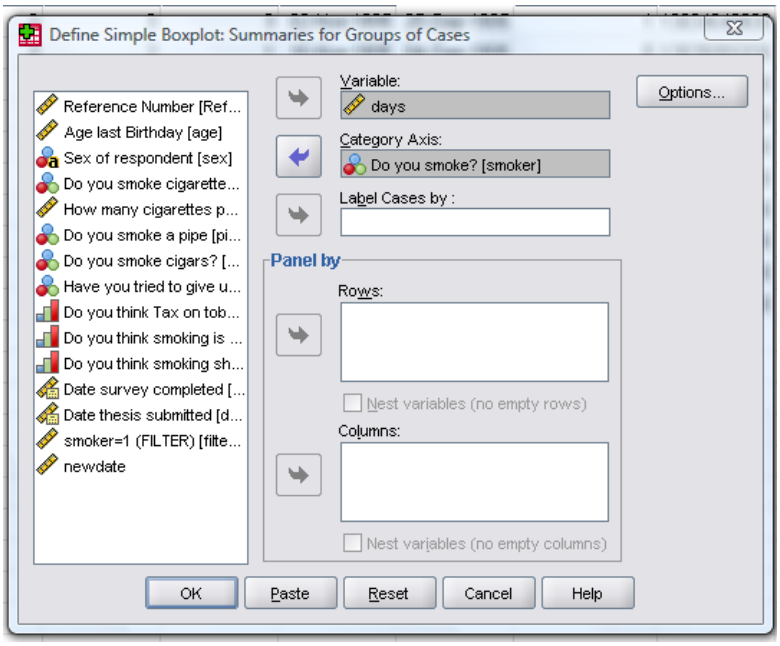


Figure 53 - define boxplot options

- Select **OK**

  You should have a box plot like the one below

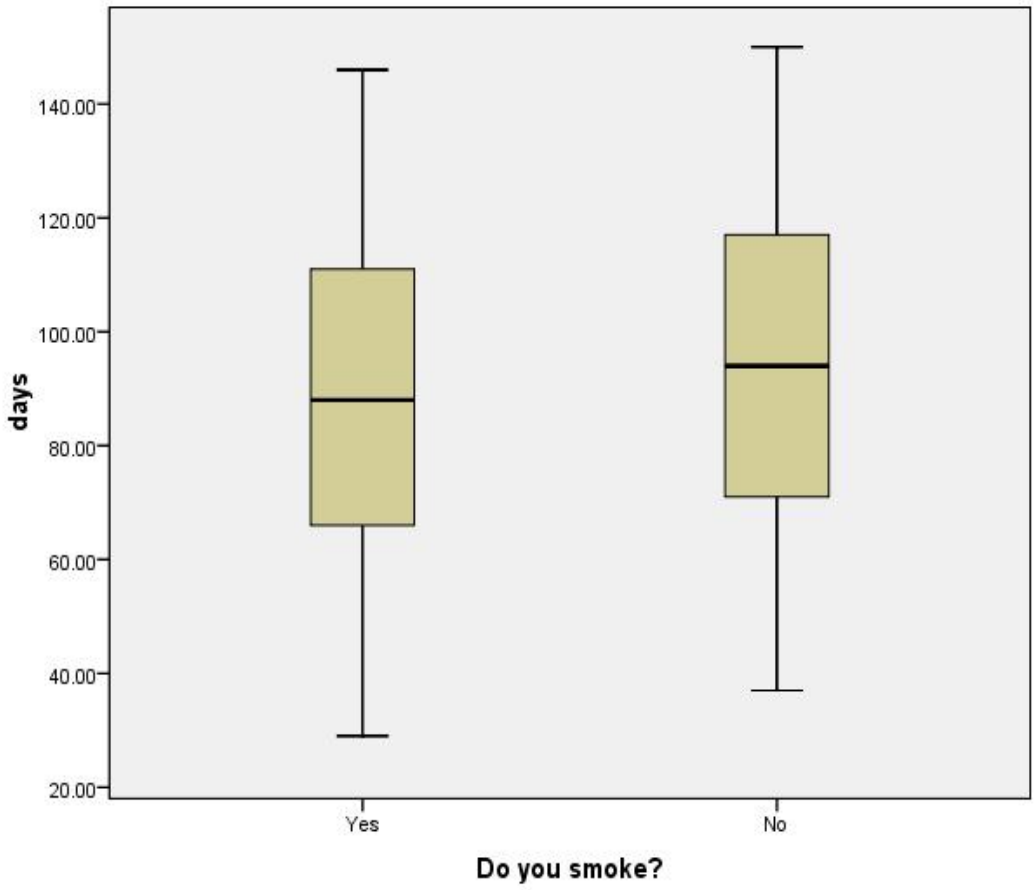


Figure 54 - finished box plot

- Click on this graph with the left mouse button to select it
- Click on the graph with the right mouse button to be given the option of editing it or exporting it.
- Go back to the define boxplot popup window to be given the chance to paste this syntax into your syntax window so that you can redo the analysis again at any time.

Task 29 Analysis - T-test

T-tests are used to demonstrate whether two groups are the same with respect to a variable which has a continuous normal distribution. For example a variable which has a continuous normal distribution in this data set is the number of days between submitting the thesis and filling in the survey, and two groups might be those people who smoke and those people who are not smokers as defined by the question 'Do you smoke?' We are testing whether the means demonstrated in the box plot above are statistically significantly different.

- Select **Analyze**
- Select **Compare Means**
- Select **Independent-Samples T Test**
- Put days in the Test Variable box
- Put Do you smoke in the Grouping variable box
- Select **Define Groups**
- Type **1** in the box labelled Group 1 (See Figure 55 - define groups)
- Type **2** in the box labelled Group 2
- Select **Continue**

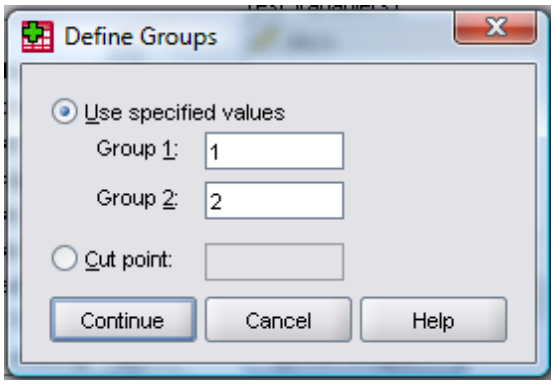


Figure 55 - define groups

The independent samples T test box should now look like this:

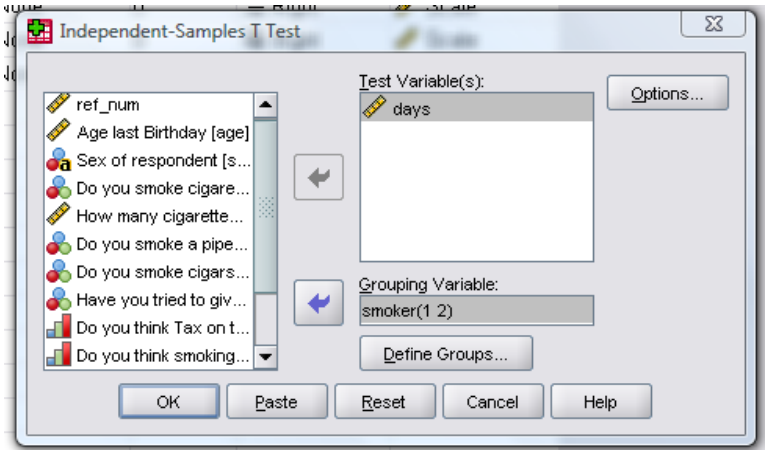


Figure 56 – Independent Samples T test

- Select **OK**

The output should look like this:

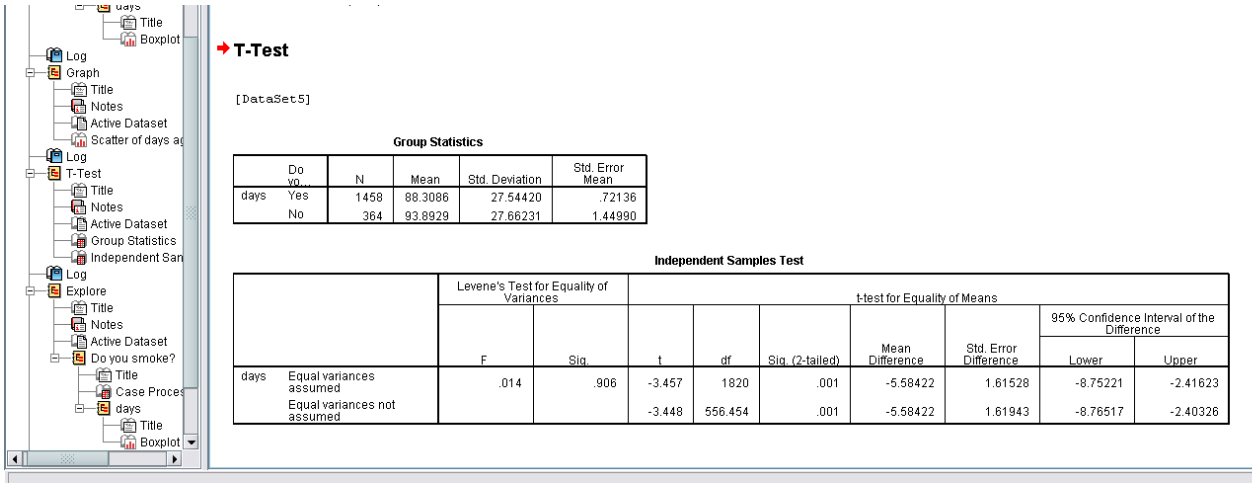


Figure 57 - T-test output

The value labelled Sig here is not the p value of the t-test. This is the value of the **Levene's Test for Equality of Variances.** The p-value for the t-test is labelled 'Sig (2-tailed).

- Can you find the mean difference and 95% confidence intervals?
  These are -5.58 with a 95% confidence interval of -8.72, -2.42
- Can you tell which group has the lower mean number of days?
- What is the mean number of days for each group?

Task 30 Analysis Non-parametric tests

The continuous variable of days analysed in Task 26 with a t-test could have been analysed using a Mann Whitney test instead:

- Select **Analyze**
- Select **Nonparametric tests**
- Select **2 Independent samples**
- Move **Days** to the **Test Variable** List
- Move **Smoker** to the **Grouping Variable** box
- Click on Define Groups
- Type **1** in the box labelled **Group 1 (**See Figure 55 - define groups)
- Type **2** in the box labelled **Group 2**
- Select continue

Test type should be Mann-Whitney U (See below)

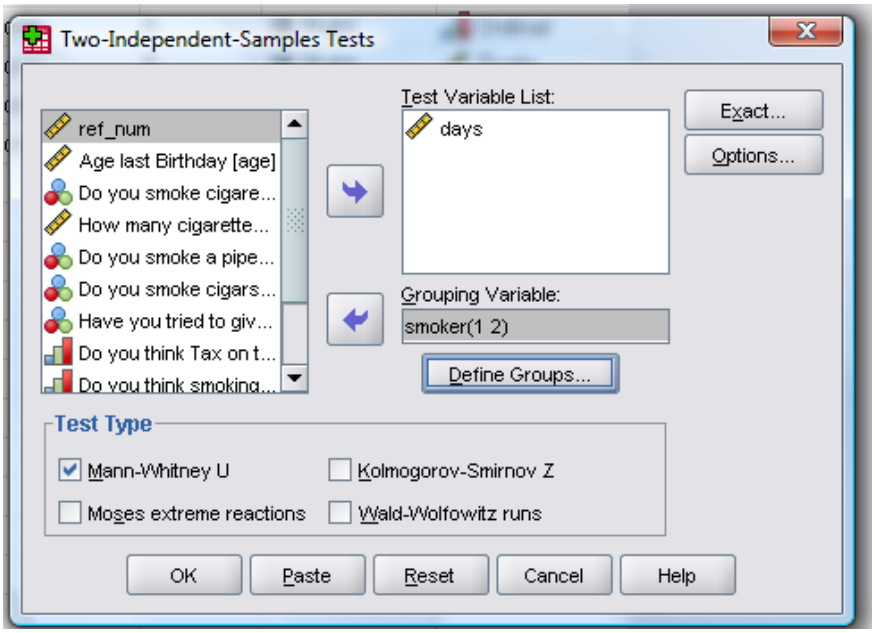


Figure 58 - Two–Independent–Samples Tests options

- Click on **OK**
- The output should appear as below in Figure 59.
- Note the **p value** is referred to as **Asymp Sig (2-tailed).**

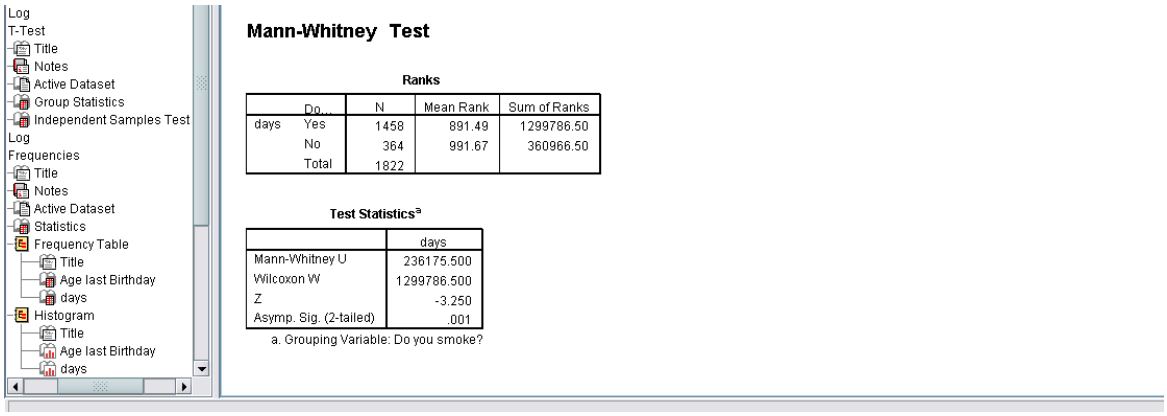


Figure 59 - Output for Mann-Whitney Test

Task 31 Analysis – Other

Objectives To identify where to find the statistics options that are relevant to the analysis you wish to do for your current work.

- One way **ANOVA** can be found under **Compare Means**.
- **Two way ANOVA, ANCOVA etc**, can be found under **General Linear Model - Univariate**.
- **Linear Regression** can be found under **Regression – Linear**.
- **Logistic Regression** can be found under either **Regression – Binary Logistic**, or **Loglinear**
- **Kaplin Meier Plots** can be found under **Survival – Kaplan Meier**
- **Cox proportional Hazard's modeling** can be found under **Survival - Cox Regression**
- **Factor analysis** can be found under **Data Reduction** as can **Principle components analysis**
- **Roc Curves** can be found under **Analyze** and not Graphs

Graphs

If you can't find the graph you are looking for under **Legacy dialogue**, you can create your own using **Chart Builder**. It is often quite time consuming to do this properly and outside the scope of this course.

Appendix A References

One of the reasons for using SPSS in the University is that it has a very comprehensive on-line documentation facility. When the software is installed you should take a copy of the 'Manuals' disc. On this you will find a series of 'pdf' files that give full documentation in each area of SPSS. Listed below are the titles of the important files.

Manuals relating to statistical issues:

- SPSS Brief Guide 23.0
- SPSS Base User's Guide 23.0
- SPSS Tables 23.0
- SPSS Data Preparation 23.0
- SPSS Advanced Models 23.0
- SPSS Regression Models 23.0
- SPSS Trends 23.0
- SPSS Categories 23.0
- SPSS Classification Trees 23.0
- SPSS Complex Samples 23.0
- SPSS Conjoint 23.0
- SPSS Exact Tests
- SPSS Missing Value Analysis 23.0

Manuals relating to computational issues:

- SPSS 23.0 Algorithms
- SPSS 23.0 Command Syntax Reference
- SPSS Programming and Data Management, 4th Edition