# 8 Checking Data

Before starting the statistical analysis it is important to check the data definition.

You can display and print the dictionary of the data file using **File | Display Data File Information | Working File**

SPSS will display all of the file information (dictionary) for each variable in the SPSS Viewer, as shown below.

*Note   The menus and toolbars in the SPSS viewer window have fewer options and buttons available. If these menus are not shown, select **View | Toolbars** and check the Viewer Standard and Viewer Outlining Toolbars.*
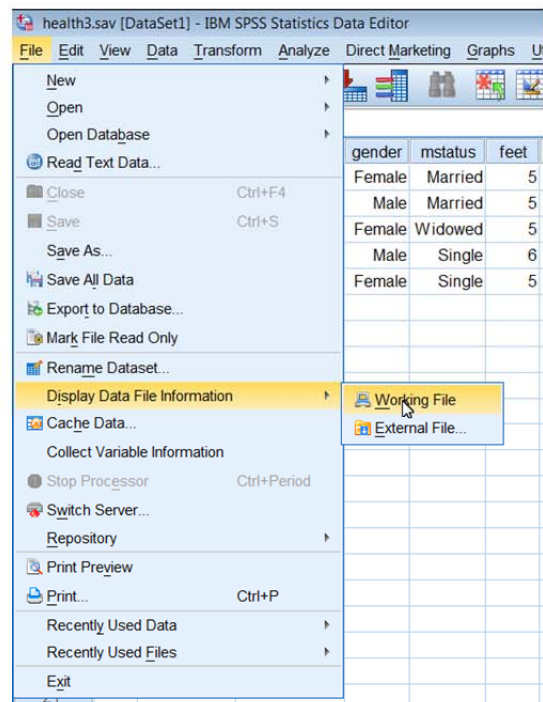
9 Viewing Output
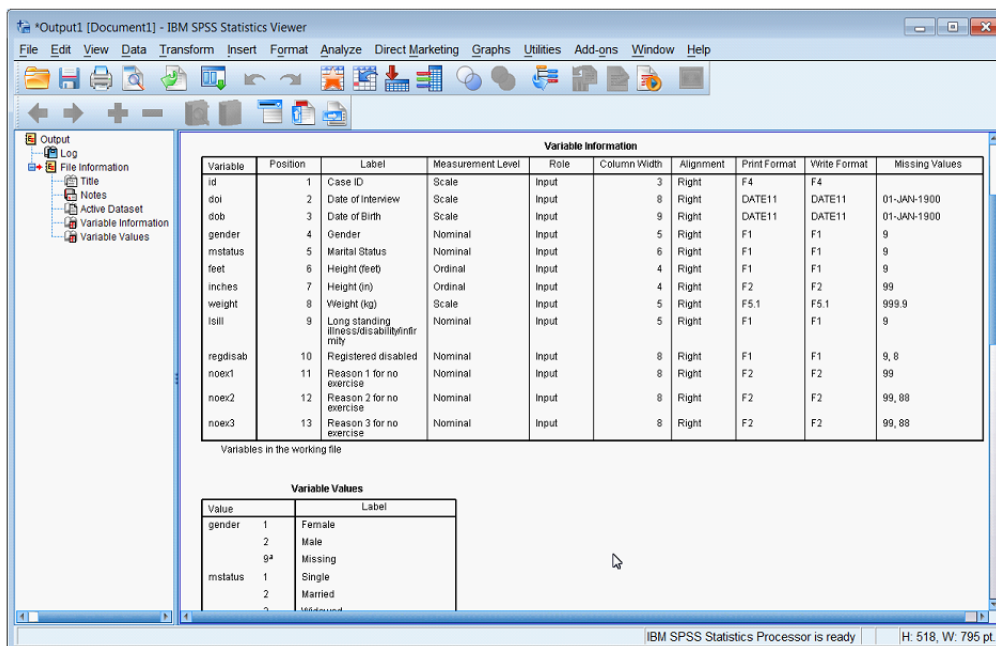
A useful tip is to remember to maximise the Viewer, if it is not already maximised. This makes it easier to view the output. Click on the maximise button ▢ at the top right of the Output window. Use the scroll bars to view the all of the output.

Checking Variable Definitions

Details of the definitions of the first four variables are shown below. The output in the viewer can be used to verify that the variables have been defined correctly.

Variable Information

Variable	Position	Label	Measurement Level	Role	Column Width	Alignment	Print Format	Write Format	Missing Values
id	1	Case ID	Scale	Input	3	Right	F4	F4	
doi	2	Date of Interview	Scale	Input	8	Right	DATE11	DATE11	01-JAN-1900
dob	3	Date of Birth	Scale	Input	9	Right	DATE11	DATE11	01-JAN-1900
gender	4	Gender	Nominal	Input	5	Right	F1	F1	9

Variable Values

	Value	Label
gender	1	Female
	2	Male
	9ª	Missing
mstatus	1	Single
	2	Married

The two tables Variable Information and Variable Values list the settings applied to these variables.

The next check you should do is to list the frequencies of all variables in the data set. This will allow you to check for errors in the data such as duplicate case numbers, or unexpected values. Although this can be done using menus, an easier way is to uses the Syntax Editor window and type the command directly. Open a Syntax Editor window by selecting:

File │ New │ Syntax

or if a Syntax Window is already open,

Click on the Syntax Editor button 📑 *Syntax1 - IB on the Windows task bar to make it the active window.

Type

freq all .

in the Syntax window.

*Note: There is a full stop after the word '**all**' which must be entered.
freq is short for Frequencies, the full word could be entered but it is not necessary.*

Click on the **Selection** button ▶ from the Syntax window toolbar to run this command.

Scroll through the output to check the frequency tables.
These frequency tables will help in the checking that:

- no duplicate case ids have been entered.
- values are within the expected range.
- missing values have been handled correctly.
- no unexpected values have been entered.

Further Checking

Further checking should be performed by cross tabulating categorical variables (for consistency checks and validation), producing histograms for continuous variables and scatter plots for pairs of continuous variables. These plots will give an idea of the average value, the variability, the shape of the distribution and whether there are any outlying or missing values. This screening will help in the choice of statistical techniques used to analyse the data.

Consistency of discrete variables (variables with a limited number of categories, e.g. gender) can be performed through the use of the Crosstabs command. For example, we may check the consistency of whether a case (i.e. respondent) with no long-standing illness has incorrectly selected "Yes" in response to the question, "If Yes, are you registered disabled'?

Select the Crosstabs dialogue box by clicking **Analyze | Descriptive Statistics | Crosstabs...**

Select *Long standing illness/disability/infirmity [lsill]* – by clicking on it, then click the ⮕ button to add it to the *Row(s)* panel.
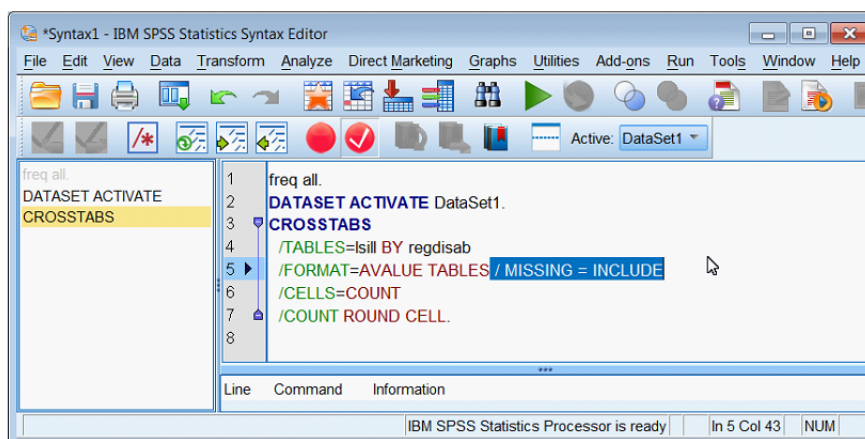Select *Registered disabled [regdisab]* by clicking on it, click the ⮕ button to add it to the *Column(s)* panel. Click **Paste**.

Note   An alternative method is to drag items from the list of variables into the Row and Column panels.
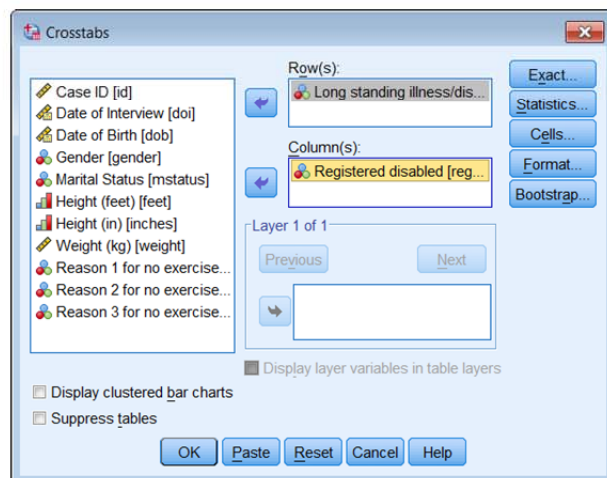
Switch to the *Syntax1 – SPSS Syntax Editor* window.
Add **/MISSING = INCLUDE** after the words AVALUE TABLES as shown below.

Run this Crosstabs command, by clicking the Selection command button ▶. Without the /MISSING = INCLUDE added you would not get any output in this instance because there is no option in the CROSSTABS dialogue to include data identified as Missing.

10 Printing Output

Before printing any output it is a good idea to use the Print Preview which is similar to the preview feature in Word. This shows you what will print on each page. Ensure that the

Output Viewer is the active window then click on the Print Preview button 🔍 or select **File │ Print Preview**.

If this output is what you want you can click on the **Print** button in Print Preview.

You can choose what you want to print from the Output window, for example, specific tables or charts. Select by **Clicking** on items listed in the Outline pane. Use the Windows techniques of **<Shift>** or **<Ctrl>** keys to select multiple items.

Click on the printer button 🖨 or, select **File | Print**.
The *Print range* panel lets you specify either *All visible Output*, or the current *Selection*.
Click on **OK** to print.

*Note   An alternative way to select items to print, is to delete items you don't want and then print all the output. To do this click on an item on the Outline pane and press **Delete** on the keyboard.*

11 Saving Output

You can save the contents of the Viewer to a file but note that that Output files can occupy a lot of disc space.

The saved file includes both the outline and contents panes. The contents of the file can be reloaded into the Statisitcs Viewer at any time. Ensure that the Viewer window is the active window, by clicking on its button on the Windows task bar, or by choosing it from the **Windows** option on the menu bar.

Click on the Save button 💾 or Select **File | Save As**.
Ensure that the folder you are saving to is MyDocuments and replace the file name with **Health1** in the *File name* box. **Click** the **Save** button. The file will be saved as Health1.spv.
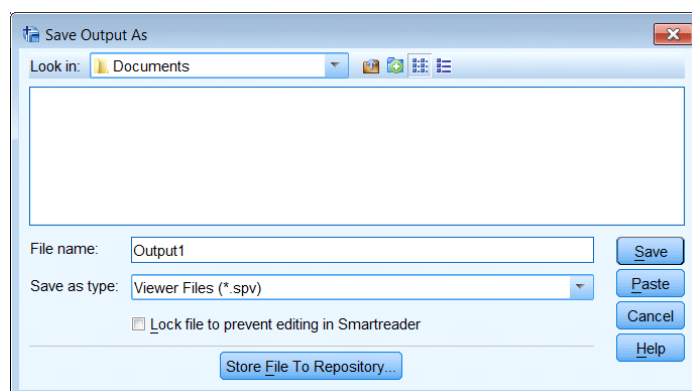
From the Statistics Viewer menu you can use **File │ Export** to export all or sections of the output. You can choose to do this with or without charts, or to export only charts (Graphics) in various formats including Excel, HTML, PDF, PowerPoint, Text and Word/RTF. Graphics can be exported as JPEG, BMP, PNG, EMF, TIFF or EPS formats.

12 Summarising Data

For the remainder of these notes we will use 150 cases from a data set similar to the one you set up earlier.

CLOSE SPSS before you start this section

- Go to web page www.som.soton.ac.uk/learn/resmethods/datasets/spss/
- Download the file **soton11.sav**, remember which folder it is saved to.
- The file will have a slightly different name, e.g. **soton11-sav-1261.sav.**
- You might want to copy the file to the Desktop and rename it to **soton11.sav**.so that the folders and filename used in these notes are identical.
- Start **SPSS** and cancel the opening dialogue.

Click on the Open file button 📁 .

Select the Desktop folder and select file *soton11.sav* using the Open Date File dialogue box.

Click on **Paste** (Not on OK).

This will enter two commands into the SPSS Syntax Window.

> **GET**
>   **FILE**='C:\Users\tnb\Desktop\soton11.SAV'
>   **DATASET NAME** DataSet1 WINDOW=FRONT.

Note   The folder name (C:Users\tnb\, will reflect your username.

Highlight these two commands, either by:

    Holding the left mouse button down and dragging over the text.

Or   by Pressing <**Ctrl A**>

Or   by Selecting **Edit | Select | All.**

Click on the **Selection** button ▶.

The next few pages will demonstrate the use of some of the **Analyze** commands in SPSS.

Other sources of information are,:

- **Which Statistical Test?**
  www.som.soton.ac.uk/learn/resmethods/statisticalnotes/which_test.htm

- **SPSS statistics coach** may assist in choosing the appropriate statistics, chart or statistical test for summarising your data. To use the statistics coach, from the menus in any SPSS window select: **Help | Statistics Coach**
  It is designed to provide general assistance for many of the basic, commonly used statistical techniques.

Categorical (nominal, ordinal) data

The **Frequencies** command is most often used for analysing categorical data. This command produces frequency tables that display the count of the number of times each code (category of the variable) occurs. Select:

Analyse | Descriptive Statistics | Frequencies

Highlight variables *lsill noex1 noex2 noex3 regdisab*

(Click on *lsill* then hold down the **Ctrl** key and **click** on each of the remaining variables).
Drag these items or click on the ➡ button to move these variables to the *Variable(s):* panel.
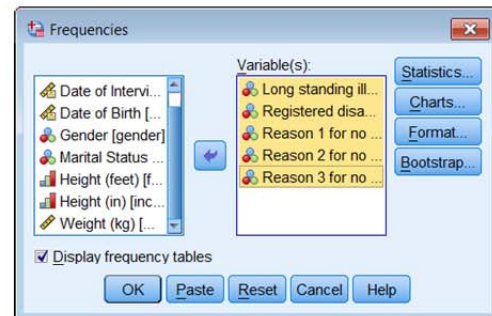
Note: *If you select the wrong variable and add it to this list. Highlight the wrong variable in the Variables list box and either drag it back to the variable list (on the left) or click the button ⬅ to remove it.*

Click on **Paste**.

Select the **SPSS Syntax Editor** button on the Windows taskbar ▣ *Syntax1 - IB* this will show the following command has been pasted into the Syntax window:

> **FREQUENCIES** VARIABLES=lsill regdisab noex1 noex2 noex3
> /ORDER ANALYSIS .

Click on ▶ to run this command. (As this is a single SPSS command it does not need to be highlighted before clicking on the Run Current syntax button).

The results appear in the Statistics Viewer window. They can be viewed as explained earlier.

SPSS can show the list of variables in various ways.
Right Click
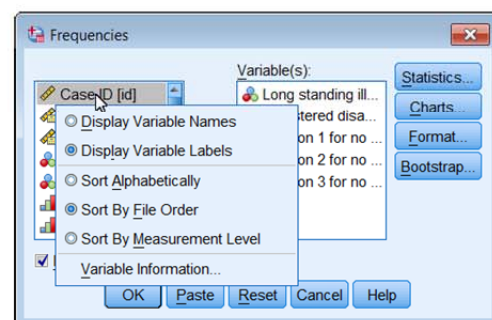
Continuous (scale, interval, ratio) data

The **Frequencies** command is just one of the commands that can be used to analyse variables of continuous data. An example is the weight of respondents, variable *weight*.
Select the Frequencies dialogue box again either by:

Analyse | Descriptive Statistics | Frequencies

Or **Click** on the Dialogue Recall button ▣. A drop down list with Frequencies on it will appear. **Click** on *Frequencies*.
Click on the **Reset** button.

Select *weight* and drag it to the Variable(s) panel.
Uncheck the *Display frequency tables* box.
An information message box about 'all output is turned off' will be displayed. **Click** on **OK**.
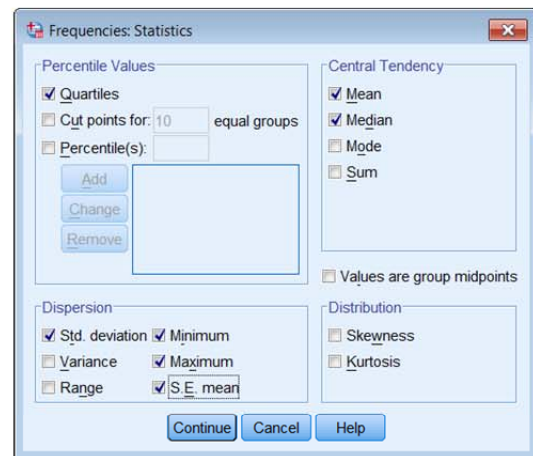
Click on the **Statistics** button.

A *Frequencies: Statistics* dialogue box will be displayed.

To select the required statistics:
Check the *Quartiles* boxes in the *Percentile Values* panel.
Check the *Mean* and *Median* boxes in the *Central Tendency* panel
Check the *Std. deviation, Minimum, Maximum* and *S.E. mean* boxes in the *Dispersion* panel.

Click on **Continue**.

Click on **Charts** in the *Frequencies* dialogue box.
A *Frequencies: Charts* dialogue box will be displayed.
Check the *Histograms* button and the *With normal curve* box in the *Chart Type* panel.
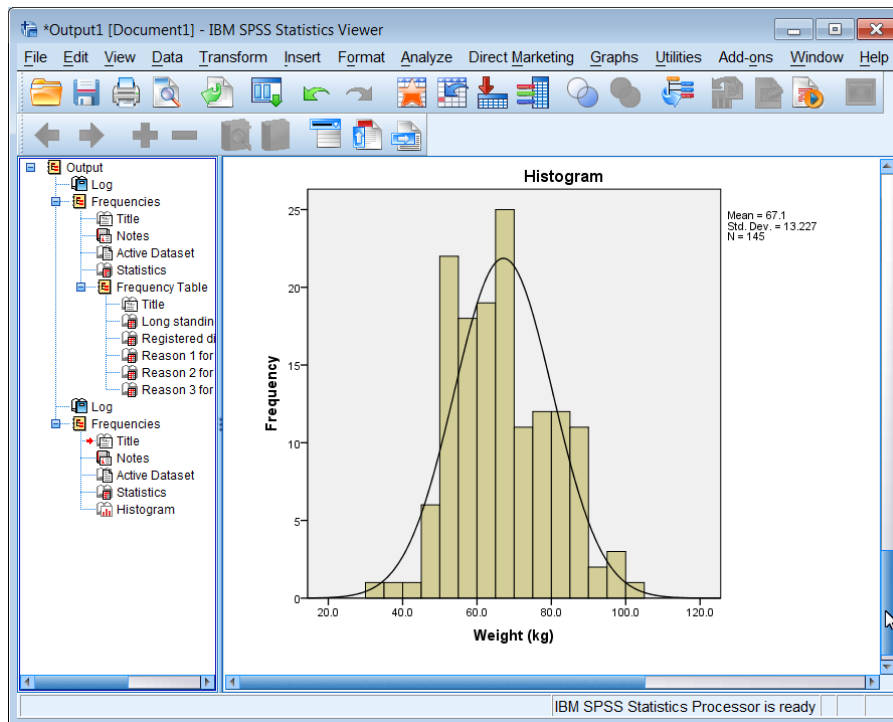**Click** on **Continue**.
Click on **Paste**.
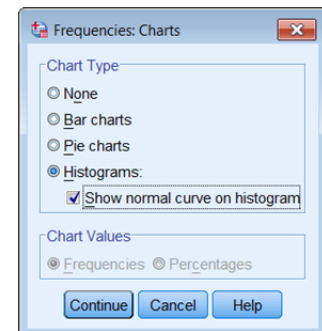
Move to the *Syntax1 - Syntax Editor* window to see that the following command has been pasted into the Syntax1:

```
FREQUENCIES VARIABLES=weight
 /FORMAT=NOTABLE
 /NTILES= 4
 /STATISTICS=STDDEV MINIMUM MAXIMUM SEMEAN MEAN MEDIAN
 /HISTOGRAM  NORMAL
 /ORDER  ANALYSIS .
```

Click on ▶ to run the command.

View the results in **Output**, window. The outline pane lists the output, including the histograms that have been created. To view the histogram of weight, **Click** on Histogram listed in the Outline (left) panel. You may need to move, by clicking and dragging the right border of the outline pane to read clearly the outline list.

It is also possible to use **Analyse | Descriptive Statistics | Descriptives** to produce similar statistics but this procedure does offer the option to create charts.

Another very useful command used with continuous variables is the **Explore** command. It is similar to the **Frequencies** command but it possesses a greater variety of plots and it can analyse subsets (specified as factors) of the data set. For example, suppose we want to analyse the weight of respondents by Gender.

Select **Analyse | Descriptive Statistics | Explore**
Highlight *weight*, add it to the *Dependent List* panel.
Highlight *gender*, add it to the *Factor List* panel.
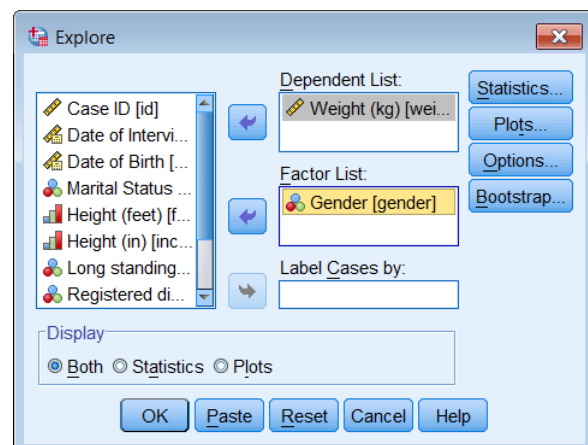
Note: If you do not specify a Factor List variable, Explore analyses all cases for each variable in the dependent list.
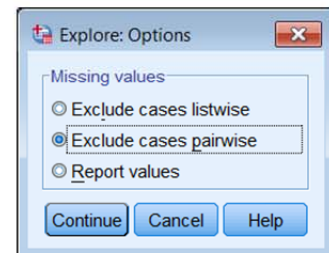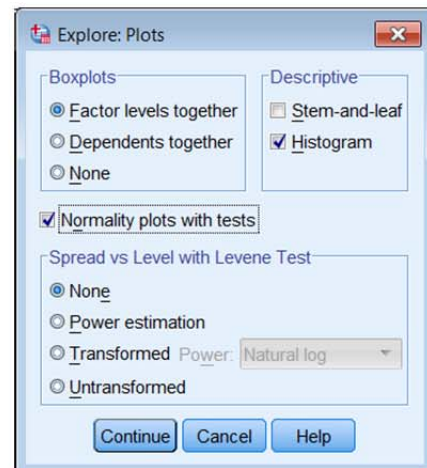
Click the **Plots** button.

UnCheck the *Stem-and-Leaf* box on the *Descriptive* panel.

Check the *Histogram* and *Normality plots with tests* check boxes ☑.

Click on **Continue**.

Click on **Options**.

Click the *Exclude cases pairwise* option button.

Note: *Although the Exclude cases pairwise option is not needed in this example, because we are only using two variables, this option is **important** when you are analysing several variables using the Explore procedure. The **Exclude cases pairwise** option uses the maximum information in a data set. Cases are included in the analysis even if they have missing data for some cases for some variables in this list. **Exclude cases listwise** excludes a case from analysis if data is missing for just one case for one variable.*
This option for the treatment of Missing Values is found in many SPSS procedures.

Click on **Continue**.
Click on **Paste**.

The Explore dialogue box creates an SPSS command called EXAMINE (not Explore!). It produces high resolution charts of Box plots and Histograms.

```
EXAMINE VARIABLES=weight BY gender
  /PLOT BOXPLOT HISTOGRAM NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING PAIRWISE
  /NOTOTAL.
```

Click on ▶ .

Note   Graphs can also create the same chart.
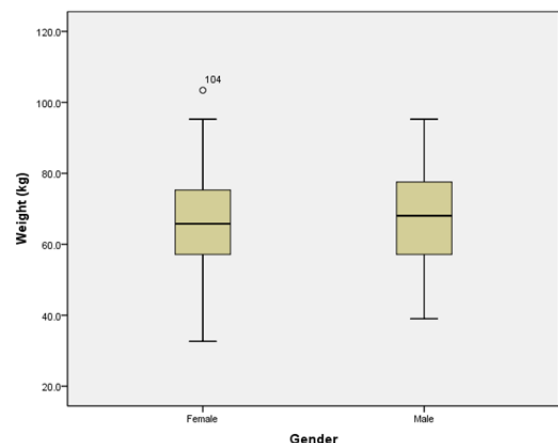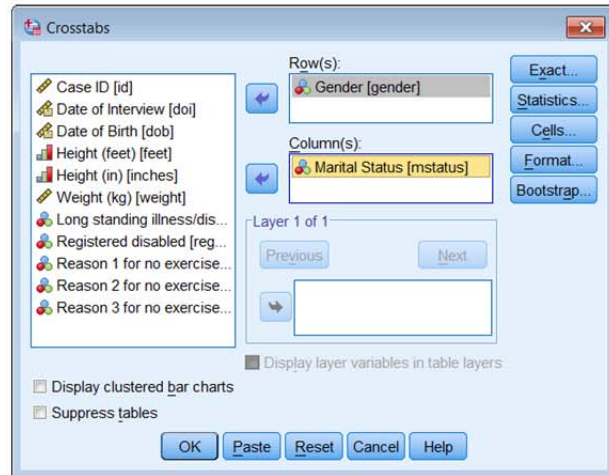
Crosstabulations

Tabulation of one categorical variable against one or more other categorical variables is done using the **Crosstabs** command. To cross tabulate *gender* by *marital status*, select:

Analyse | Descriptive Statistics | Crosstabs

Highlight *gender* and add it to the Row(s) list. (In the Output, row categories are printed down the page).
Highlight *mstatus* and add it to the Columns(s) list.
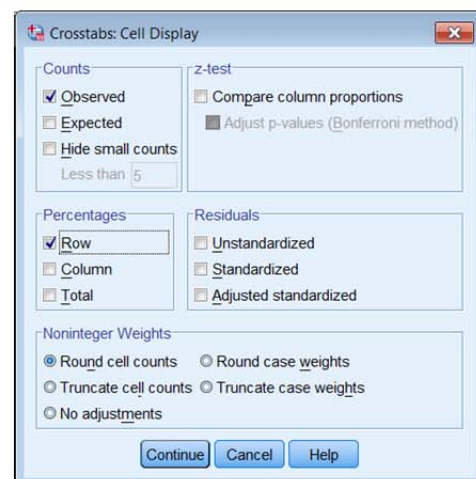In the Output, column categories will be printed across the page.

Click on **Cells**

Click on the *Row* check box in the *Percentages* panel.
Click on **Continue**.

Click on **Paste**.
Click on ▶ .

Other buttons on the Crosstabs dialogue allow the selection of Statistics, for example Chi-square. At this stage the cross tabulation table should be examined and percentages compared before any statistical hypothesis testing is performed.

13 Why all this pasting? - the SPSS Syntax file

You have already used the syntax window to execute SPSS commands and by typing **frcq all** into the window. Although SPSS uses menus and dialogue windows, it is basically a command driven program. The syntax window is where commands are pasted, or typed directly, to build a SPSS program. This program can be saved to a file. It can be recalled and executed again at a later date if the data changes. The importance of entering (pasting) commands into the Syntax window and saving them cannot be over emphasised.

Note: There are some SPSS commands and Options for some SPSS commands that can only be entered using the Syntax window because dialogue boxes for building them do not exist.

Saving and Printing the Syntax window

To save the contents of the syntax window ensure that it is the current window (The title of this window must be **SPSS Syntax Editor**. **Click** on the button on the windows task bar, or select it from the Window menu).
Either Select

File | Save As or **File | Save**

or **Click** on the Save button .

Chose an appropriate Documents folder and enter **<u>Health1</u>** (your chosen name in this example) into the *File name* text box and ensure that the directory and drive destinations within this dialogue box are correct.

Click on **Save**.

*Note   The name of the syntax window changes to **health1**. SPSS automatically adds .sps as the file extension. From now on **File | Save** or clicking on will save the contents of this window to health1.sps without asking if it should replace the existing contents.*

Retrieving and Editing the Syntax file

To retrieve syntax commands that were previously saved.

Either select **File | Open** or Click on on the Syntax window toolbar.

*Note: If the current window is not a Syntax Window, the **Files of type** box must be scrolled and select **Syntax (*.sps)**.*

The *Open* dialogue box will be displayed (it is identical to the Save dialogue box).
Select the appropriate drive, folder and filename details.
Double click on the filename or **Click** on **Open** and the file will appear in a second syntax window.

- Commands can be typed directly into the syntax window, or pasted from dialogue boxes.
- To execute one command at a time, ensure that the flashing insertion point is placed somewhere on the command line(s), and then click on .
- To run multiple commands, you must highlight all the lines containing the commands before clicking . To do this move to the top left or bottom right of the command lines, hold down the left mouse button and drag the mouse to the opposite corner

(*Note: Only part of the first and last commands must be highlighted*)
To execute all the lines within a syntax window.

Select **Edit | Select All**.
or   Press **<Ctrl A>**
All the lines in the syntax file will now be highlighted.

Click to execute all the commands.

When the Syntax window is active, clicking on the syntax button will launch a 'Syntax Help' window with details of the syntax for the current command. A general syntax help window will be presented if the cursor is not on a command.

Tip: It is possible to have more than one Syntax window (file) open concurrently. However, when more than one Syntax window is open, commands are pasted into the active or Designated Syntax window. To make an alternative Syntax window the Designated window by click the button. The active or designated syntax window shows a disabled designate button on the toolbar