



**Graduate School**

**Core Quantitative Methods**

*Also available as:*

***Core Quantitative Methods for DTC Students, option G***

*And alternative exit points:*

***Appreciating Quantitative Methods  
Selected Quantitative Methods***

Computer Exercises  
and  
Reference Information

2016-17

Mike Griffiths

# CORE QUANTITATIVE METHODS COMPUTER EXERCISES AND REFERENCE INFORMATION

## Contents

<b>1</b>	<b>Foreword .....</b>	<b>5</b>
1.1	How to use this booklet.....	5
1.2	Overview of inferential tests.....	5
<b>2</b>	<b>Introduction to SPSS.....</b>	<b>6</b>
2.1	Name, version and opening SPSS .....	6
2.2	Data entry – continuous variables.....	6
2.3	Descriptive statistics – continuous variables .....	9
2.4	Data entry – categorical variables.....	11
2.5	Descriptive statistics – categorical variables .....	12
<b>3</b>	<b>Introduction to Excel .....</b>	<b>13</b>
3.1	Introduction to Excel and its versions .....	13
3.2	Simple statistics in Excel.....	13
3.3	Graphs in Excel .....	15
3.3.1	Creating graphs.....	15
3.3.2	Editing and changing graphs.....	17
3.3.3	Bar charts with two independent variables/ data series.....	19
<b>4</b>	<b>Introduction to graphs in SPSS; Histograms; Chart Editor .....</b>	<b>20</b>
4.1	Introduction; Chart Builder .....	20
4.2	Histograms.....	20
4.3	Changing the appearance of a chart using the Chart Editor .....	21
4.4	Copying the chart into another application.....	23
<b>5</b>	<b>t-tests, Anovas and their non-parametric equivalents .....</b>	<b>23</b>
5.1	Introduction .....	23
5.2	Which test to use? .....	23
5.3	Entering Repeated Measures data .....	24
5.4	Paired samples t-test.....	25
5.5	Wilcoxon (Signed Ranks) test.....	29
5.6	Repeated Measures Anova .....	30
5.7	Friedman test.....	34
5.8	Independent-samples data - general .....	35
5.8.1	Entering independent-samples data.....	35
5.8.2	Descriptive statistics and histograms .....	37
5.9	Independent-samples t-test .....	38
5.10	Mann-Whitney U test.....	40
5.11	Independent-samples Anova.....	41
5.12	Kruskal-Wallis test .....	43
<b>6</b>	<b>Factorial Anovas.....</b>	<b>44</b>

6.1	Introduction .....	44
6.2	Outcomes .....	45
6.3	If the factorial Anova shows significant effects.....	46
6.4	Effect sizes .....	47
6.5	Two way independent-samples Anova .....	47
6.5.1	Following up a significant interaction.....	52
6.6	Two way repeated measures Anova.....	53
6.6.1	Following up a significant interaction.....	59
6.7	Two way mixed Anova .....	59
6.7.1	Following up a significant interaction.....	65
6.8	Anovas with more than two factors .....	66
<b>7</b>	<b>Chi-square tests of association.....</b>	<b>66</b>
7.1	Introduction; when they are used.....	66
7.2	The possible outcomes of a chi-square test.....	66
7.3	Example 1: entering individual cases into SPSS .....	67
7.4	Example 2: using the Weighted Cases procedure in SPSS.....	70
7.5	Showing percentages .....	73
7.6	Effect sizes .....	73
<b>8</b>	<b>Chi-square tests of a single categorical variable.....</b>	<b>74</b>
8.1	When they are used.....	74
8.2	Whether a categorical variable is evenly distributed .....	74
8.3	Whether a categorical variable is split in a given proportion .....	75
<b>9</b>	<b>Cochran's and McNemar's tests .....</b>	<b>78</b>
9.1	When to use Cochran's and McNemar's tests .....	78
9.2	Cochran's Q.....	78
9.3	McNemar's test.....	79
<b>10</b>	<b>Simple regression and correlation.....</b>	<b>81</b>
10.1	Scatterplots.....	81
10.2	Correlation .....	83
10.2.1	Parametric test of correlation (Pearson's $r$ ).....	83
10.2.2	Non-parametric test of correlation (Spearman's $\rho$ ) .....	83
10.3	Simple linear regression .....	84
10.3.1	Carrying out a regression.....	84
10.3.2	Regression output.....	84
10.3.3	Writing up regression .....	86
10.3.4	What it means .....	86
<b>11</b>	<b>Multiple regression and correlation.....</b>	<b>87</b>
<b>12</b>	<b>Introduction to statistics for questionnaires.....</b>	<b>91</b>
12.1	Overview.....	91
12.2	Entering the data .....	91
12.2.1	Introduction: example data.....	91
12.2.2	Variable view.....	92
12.2.3	Entering data.....	93
12.3	Calculating overall scores on a questionnaire .....	93
12.3.1	Introduction .....	94
12.3.2	Reverse-scored questions: what they are.....	94

12.3.3	Reverse-scored questions: How to deal with them .....	94
12.3.4	Adding up scores. ....	96
12.3.5	Mean scores .....	96
12.4	Your own scales: a very brief introduction .....	96
12.4.1	Checking for problematic questions .....	97
12.4.2	Cronbach's alpha: how to calculate it.....	99
12.4.3	How Cronbach's alpha is affected by individual questions.....	100
<b>13</b>	<b>Operations on the data file .....</b>	<b>101</b>
13.1	Overview: what SPSS can do.....	101
13.2	Calculating z scores .....	101
13.3	Calculations on one or more variables, using Compute Variable .....	102
13.4	Creating a case number (e.g. participant number) .....	103
13.5	Categorising data (e.g. pass/fail, high/low).....	103
13.5.1	Predefined split point(s) .....	103
13.5.2	Splitting into equal groups: e.g. median splits .....	104
13.6	Working with part of the data file .....	104
13.6.1	Selecting participants/ cases.....	104
13.6.2	Combining conditions, e.g. when you already have an <i>Include</i> variable. ....	106
13.6.3	Splitting the file with Data – Split File .....	106
13.6.4	Sorting the file .....	108
<b>14</b>	<b>Checking and screening data .....</b>	<b>108</b>
14.1	File control and excluding participants/ cases .....	108
14.1.1	Different versions of your file.....	108
14.1.2	Excluding participants/ cases; use of an <i>include</i> variable .....	108
14.2	Checking the data file; missing data .....	109
14.2.1	Check your data entry .....	109
14.2.2	Missing and illegal data – definitions.....	109
14.2.3	Detecting missing and illegal data – categorical variables. ....	109
14.2.4	Detecting missing and illegal data – continuous variables. ....	110
14.2.5	Finding the case(s) with missing or illegal values .....	111
14.2.6	Dealing with missing data and illegal values .....	111
14.3	Good practice: checking for assumptions .....	112
14.3.1	Overview; remember the checks we already do .....	112
14.3.2	Checking for outliers .....	113
14.3.3	Checking whether continuous variables are normally distributed .....	113
14.4	What to do if there are problems .....	114
14.4.1	A non-parametric test.....	114
14.4.2	Options relating to specific tests .....	114
14.4.3	Deleting outliers .....	114
14.4.4	A non-linear transformation.....	114
<b>Appendices .....</b>		<b>116</b>
<b>A. Reporting results .....</b>		<b>116</b>
What to include when reporting an inferential test.....		116
Reporting in APA Style .....		116

Formatting hints in Word .....	117
Rounding numbers .....	117
<b>B. Copying graphs and other objects into Word (or other applications)</b>	<b>118</b>
Converting charts and graphs to black and white .....	118
Straightforward copying .....	118
If there are problems with copying .....	118
Avoiding problems with objects moving around .....	119
<b>C. Help in SPSS .....</b>	<b>119</b>
<b>D. Boxplots and percentiles .....</b>	<b>120</b>
D1. Introduction .....	120
D2. Obtaining boxplots .....	121
D3 Interpreting boxplots .....	121
<b>E. Failing Levene's test .....</b>	<b>123</b>
<b>F. Some useful transformations and how to compute them .....</b>	<b>124</b>
Choosing a transformation .....	124
<b>G. Areas under the normal distribution .....</b>	<b>127</b>
<b>H. Overview of statistical tests .....</b>	<b>129</b>

# 1 Foreword

## 1.1 How to use this booklet

This booklet contains computer exercises for use in class and as examples of how to carry out various kinds of analysis. Generally speaking it can be read on its own (and kept for reference), but for the Core Quantitative Methods course it should be obviously read in conjunction with:

- the lectures; please print off the PowerPoint presentations. These will be posted on learn.gold.
- the Module Handbook, which gives details of recommended reading, content of classes, etc.

The order of material in this booklet has been chosen to make it read logically as a permanent reference book. For teaching reasons, the material will be covered in a different order in class. The booklet also includes a few sections which will not be covered in class at all, but which could be useful for future reference.

Information in boxes (like this) can be ignored on first reading, and may require an understanding of material which is covered later in the course. However, it may be useful when referring to the booklet later.

Unless otherwise specified, all the examples in this booklet use invented data.

## 1.2 Overview of inferential tests

**Table 1.1** is a quick reference to the inferential tests which take up much of this booklet. A more comprehensive overview of statistical tests is given as Appendix H.

**Table 1.1.** Overview of inferential tests in this booklet.

Variables	See chapter
Two: one categorical, one continuous	5
Two: both categorical	7
Two: both continuous	10
Several: More than one categorical; one continuous	6
Several: All continuous	11
One categorical (testing whether the categories are split in a given proportion)	8

Often, we conceptualise one (or more) of the variables as being an Independent Variable (IV) and the other as being a Dependent Variable (DV). However, all of these procedures simply test whether there is a statistically significant relationship between the variables. They do not demonstrate cause and effect; that depends on the validity of the study.

For example, consider a Mann-Whitney U test. This analyses whether there is any difference between groups on a continuous variable. Suppose we investigate whether there is any difference between students and non-students in how much alcohol they drink. We might have hypothesised that being a student makes people more likely to drink. However, the test would work just the same in any of the following circumstances:

- (a) if being a student makes people likely to drink more (IV = whether a student (yes/no); DV = amount of alcohol drunk (units per week)).
- (b) if people who drink more are more likely to become students (IV = amount of alcohol drunk; DV = becoming a student).
- (c) some other factor, such as parent's income, affects both, so neither variable is an IV.

## 2 Introduction to SPSS

### 2.1 Name, version and opening SPSS

SPSS is owned by IBM, who at one time changed its name to PASW; you may see various combinations of these names, especially in old books and manuals. This booklet was written using SPSS version 22. However, from version 15 onwards there have been very few changes.

To open SPSS:

- On your own computer, it will be in the usual places (e.g. in Windows, under **Programs** or **All Apps**, and you can create a desktop shortcut) named **SPSS** or **IBM SPSS**
- On college computers, arrangements may differ.

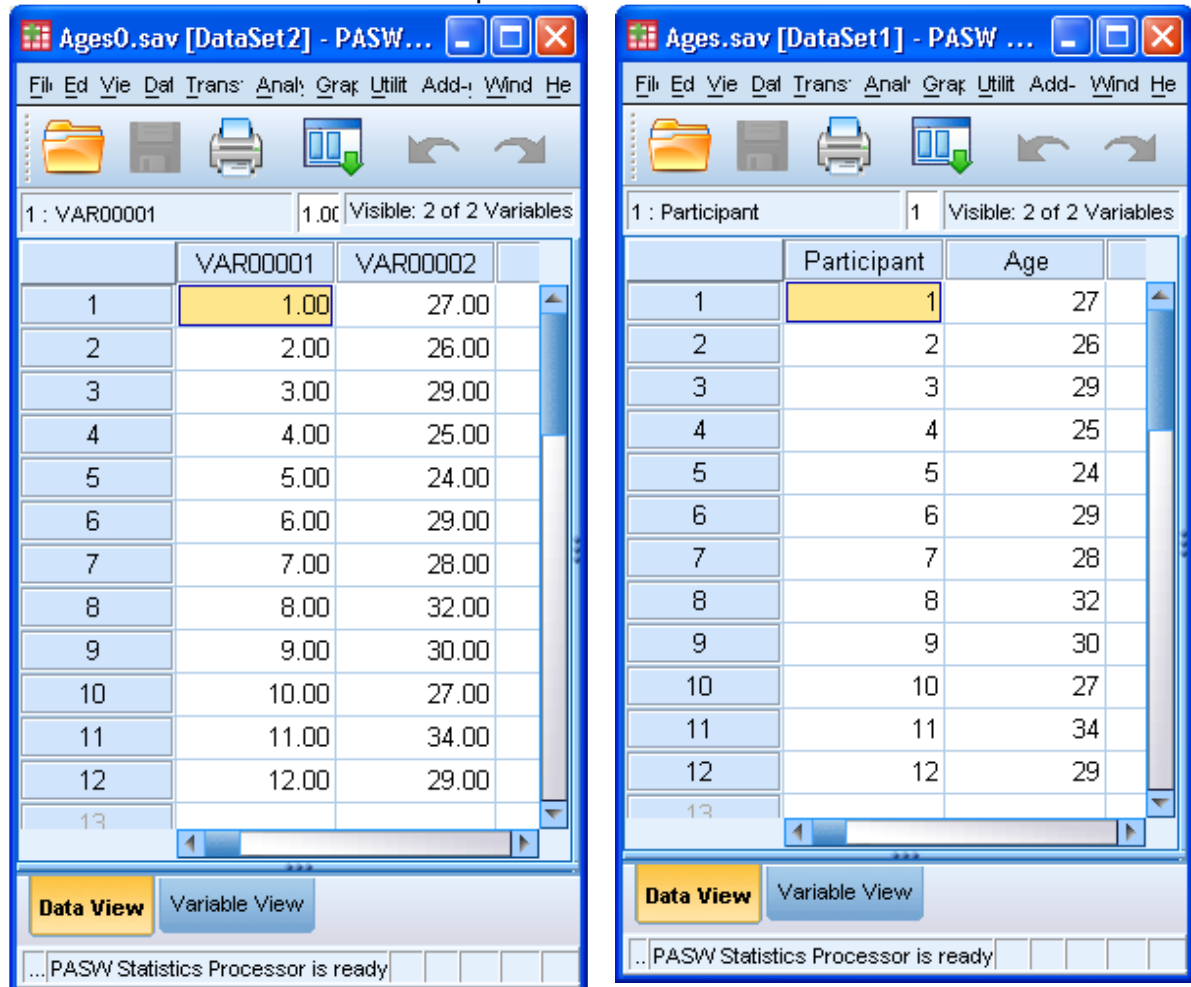
When you open SPSS, a dialogue box will appear. Cancel this for now. (If in future you decide that you never want it to appear, you can tick the "Don't show this dialog in the future" box.)

### 2.2 Data entry – continuous variables

When you open SPSS, the **Data Editor** opens. This uses the file extension .sav. Notice that it has two tabs:

- **Data View** where you enter the data.
- **Variable View** where you tell it about the variables, e.g. their names.

SPSS requires us to enter the data in a prescribed manner. All the information about one case (one person, etc.) goes in one row. For this demonstration, all we know about each person is their participant number and their age. If you enter those figures in **Data View**, the screen will look like 2.1(a). Notice that SPSS has entered some decimal points which we do not want.



(a) as first entered (b) after editing  
**Figure 2.1.** Data for first exercise in Data View.

The table will look more meaningful if we remove the unwanted decimal points and give our variables names, as in Figure 2.1(b). To do this, we need to go to **Variable View**. Click on the tab and your screen will look similar to Figure 2.2. Each row of **Variable View** specifies one column in **Data View**; for example the first row of Variable View specifies the first column of **Data View**.



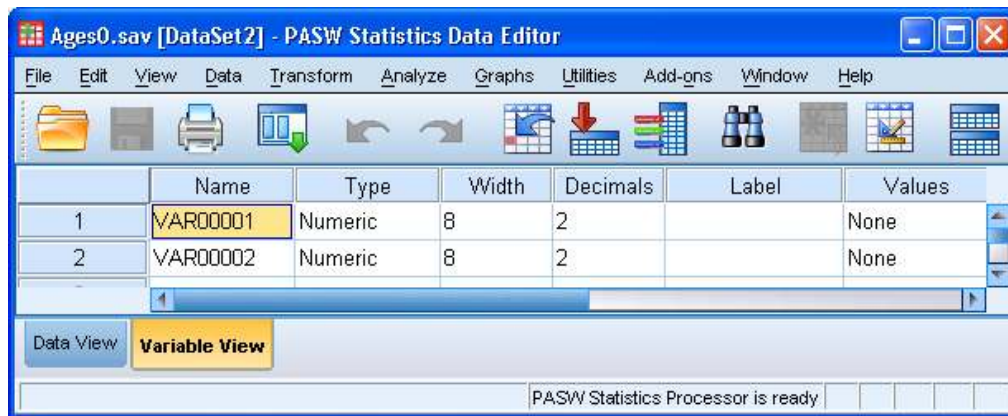
Firstly, let us give our variables names. Change *VAR00001* to *Participant*<sup>1</sup> and *VAR00002* to *Age*. Click back to **Data View** and notice that these names now appear at the top of the columns there.

Now change to the number of decimals appropriate to our data. Go back to **Variable View** and change **Decimals** from 2 to 0 for each of our two variables. Go back to **Data View** to see the effect. Your **Variable View** should now look like Figure 2.3 and your **Data View** should now look like Figure 2.1(b).

*A few points to note if you come back to this section for future reference:*

1. SPSS has some restrictions on the names you can give your variables; for example they cannot include spaces. You can add a more meaningful name in the 'labels' field (see section 5.8.1).
2. We changed the number of decimals to 0 because we were using whole numbers. If there are decimals in the data, keep the appropriate number of decimals in SPSS.
3. We entered the data first in **Data View**, and then set up the variables in **Variable View**. Once they are used to entering data, most people find it easier to set up the variables in **Variable View** first. You can swap between the two views as you please.

Save the file on your u: drive as *Ages.sav* so we can use it again next week.



**Figure 2.2.** Variable View associated with Figure 2.1(a).

<sup>1</sup> We will not be using the Participant variable for our calculations, but for various reasons it is good practice to make sure that each of your cases has a unique identifying number, even if you have to make one up for your SPSS file. Paragraph 13.4 shows how to create one automatically.

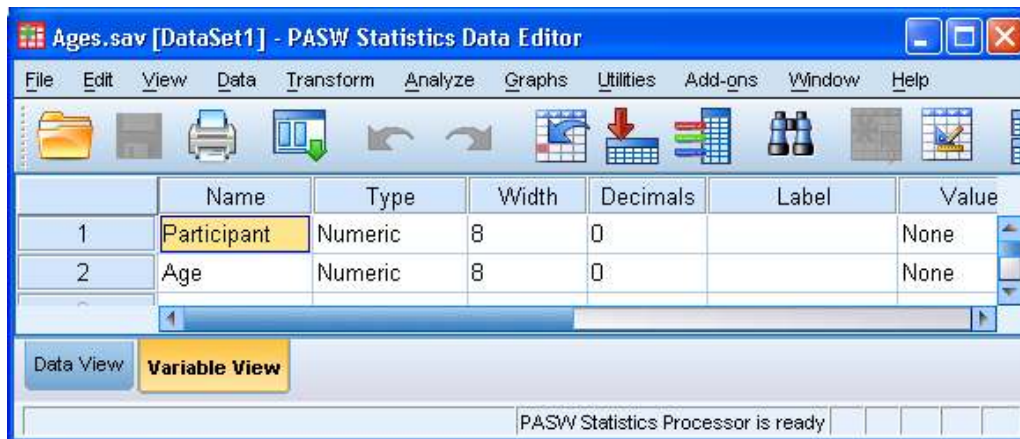


Figure 2.3. Variable View associated with Figure 2.1(b).

## 2.3 Descriptive statistics – continuous variables

Now we can calculate some statistics. On the drop-down menu click on **Analyze – Descriptive Statistics – Frequencies**<sup>2</sup>. A dialogue box comes up. When you have finished it will look as shown in Figure 2.4.

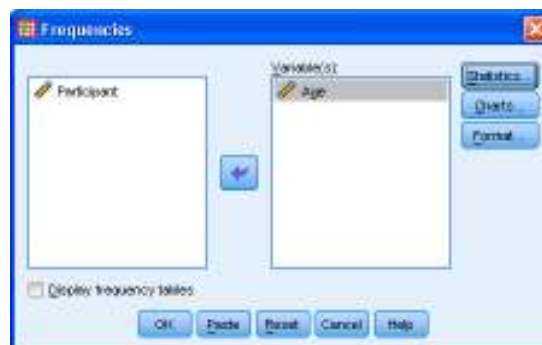


Figure 2.4. Frequencies dialogue box.

Click on *Age*, and then the arrow, to move *Age* into the box marked **Variable(s)**. Uncheck the box that says **Display frequency tables** (ignore the warning message that comes up). Click on **Statistics...** to choose what statistics you want to see. In this case we will ask for all the ones we have covered, as shown in Figure 2.5.

Click **Continue** and **OK**. The answers appear, as shown in Figure 2.6.

<sup>2</sup> Notice that there are several options under the same menu for getting descriptive statistics, but this is the easiest way of getting all the ones we want.

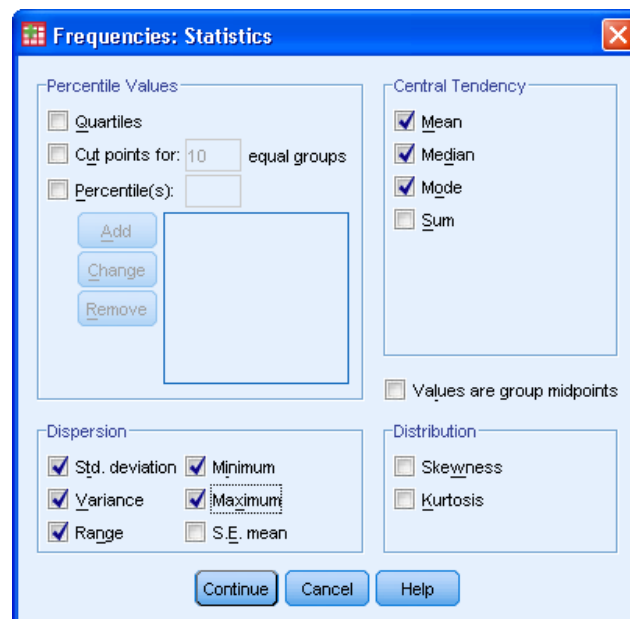


Figure 2.5. Frequencies: statistics dialogue box.

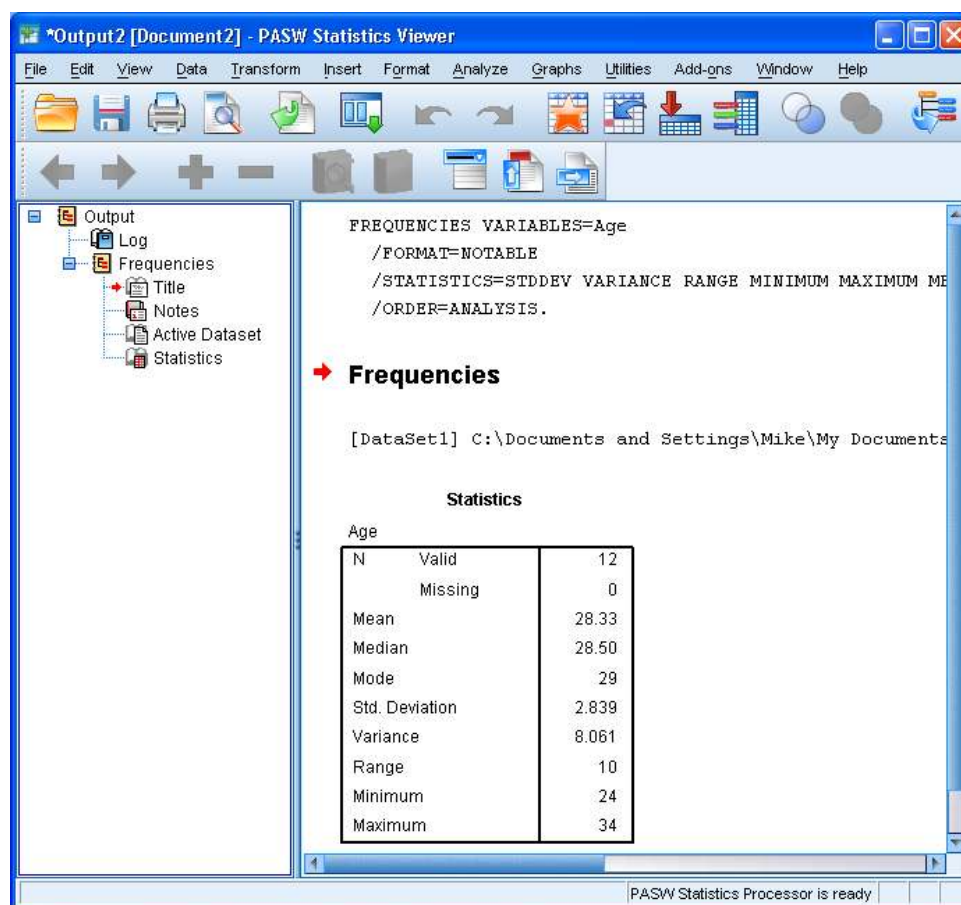


Figure 2.6. Descriptive statistics – numerical variables.


Notice that the Output window opened automatically. This is a separate file. If you save it, it will have the file extension .spv (.spo in SPSS 15 and earlier).


## 2.4 Data entry – categorical variables

When we enter categorical variables, we will represent each category by a number. This is just to make life easier for SPSS. It does not matter what numbers we use, but usually we use whole numbers starting with 1.

Suppose we wanted to extend the data in Figure 2.1 to include the participants' genders, as shown in Figure 2.7 (b). Enter the data with your chosen code, e.g. 1 = *male* and 2 = *female*.

Now we need to tell SPSS what these numbers mean. Go to **Variable View**. Firstly, name the variable as *Gender* and give it 0 decimal places. Then, on the same line, click on the box for **Values**. Three dots appear. Click on them, and a new dialogue box comes up. To show that 1 represents *male*, enter 1 in Value and *male* in **Label** (as shown in Figure 2.8). Click on **Add**. Repeat for 2 and *Female*.


Now if you go back to **Data View**, you can choose whether to show the genders by their numbers or their labels (Figure 2.7 (a) or (b) respectively). To change, click on **View – Value Labels**, or the  icon.



The screenshot shows the SPSS Data View window for a file named 'Agas.sav'. The window title bar includes 'IBM SPSS Statistics Data Editor'. The menu bar includes 'File', 'Edit', 'View', 'Data', 'Transform', 'Analyze', 'Direct Mode', 'Compute', 'Utilities', 'Window', and 'Help'. The toolbar contains icons for file operations, editing, and analysis. The status bar at the bottom indicates 'IBM SPSS Statistics Processor is ready.' The data grid has columns for 'Participant', 'Age', and 'Gender'. The 'Gender' column contains numerical values 1 and 2. The 'Data View' tab is selected at the bottom.

	Participant	Age	Gender
1	1	27	1
2	2	26	1
3	3	29	2
4	4	25	1
5	5	24	2
6	6	29	2
7	7	28	2
8	8	32	1
9	9	30	1
10	10	27	2
11	11	34	1
12	12	29	2

(a) Value labels off



The screenshot shows the SPSS Data View window for a file named 'Agas.sav'. The window title bar includes 'IBM SPSS Statistics Data Editor'. The menu bar includes 'File', 'Edit', 'View', 'Data', 'Transform', 'Analyze', 'Direct Mode', 'Compute', 'Utilities', 'Window', and 'Help'. The toolbar contains icons for file operations, editing, and analysis. The status bar at the bottom indicates 'IBM SPSS Statistics Processor is ready.' The data grid has columns for 'Participant', 'Age', and 'Gender'. The 'Gender' column contains text labels 'male' and 'female'. The 'Data View' tab is selected at the bottom.

	Participant	Age	Gender
1	1	27	male
2	2	26	male
3	3	29	female
4	4	25	male
5	5	24	female
6	6	29	female
7	7	28	female
8	8	32	male
9	9	30	male
10	10	27	female
11	11	34	male
12	12	29	female

(b) value labels on

**Figure 2.7.** Data view with a categorical variable.

## 2.5 Descriptive statistics – categorical variables

To get descriptive statistics for a categorical variable, click on **Analyse – Descriptive Statistics – Frequencies**. You get the same dialogue box as before (Figure 2.4). This time we simply want the default options (so if you have recently asked for statistics for a numerical variable, press the Reset button to undo all the options you asked for before). Now move the categorical variable(s) (*Gender* in this case) across to the **Variable(s)** box. Click **OK**. The output, shown in Figure 2.9, shows the number and the percentage of participants in each category.

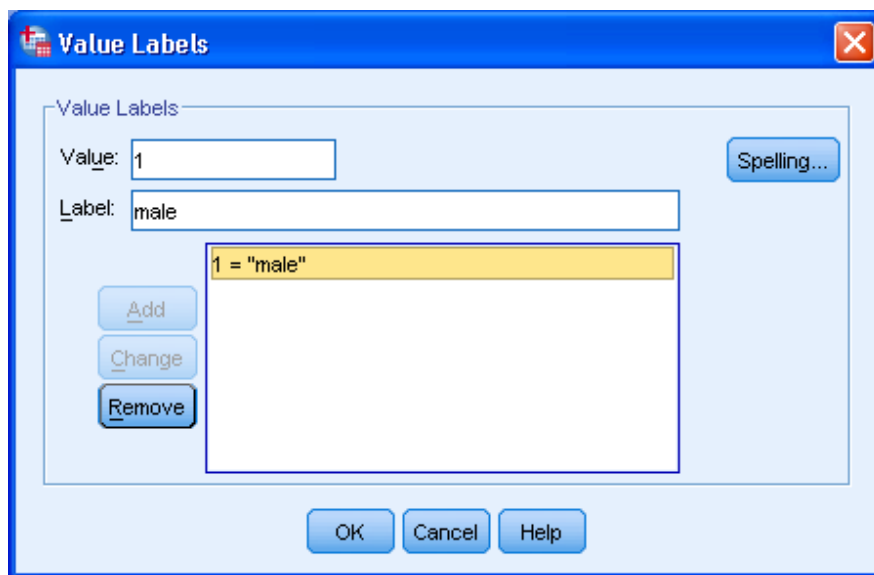


Figure 2.8. Value labels dialogue box.

Gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	6	50.0	50.0	50.0
	female	6	50.0	50.0	100.0
Total		12	100.0	100.0	

Figure 2.9. Output for a categorical variable.

## 3 Introduction to Excel

### 3.1 Introduction to Excel and its versions

Excel is a general purpose spreadsheet programme. It allows basic statistics calculations, but not to the depth that we need. It is however better than SPSS for at least two things:

- (a) Data manipulation. SPSS does allow some calculations on the data (see chapter 13) but Excel is specialised for this. We will only scratch the surface of what it can do.
- (b) Graphs and charts (these terms are almost interchangeable in Excel). We will cover various kinds of graphs and charts in SPSS during the course, but Excel is more flexible and is usually the best choice if you want to prepare a graph to a publishable standard.

This chapter is written for Excel 2013. The previous versions are quite similar, from 2007 onwards (for differences, see footnotes). Excel 2002 and earlier versions are a lot different (they do not have tabs along the top of the window); an appropriate version of these instructions is available on request.

Like SPSS, Excel has rows and columns. However, in Excel there is no fixed way to lay things out– it is like a blank sheet of paper.

You will quickly find that in Excel, there is often more than way to carry out any given action.

To open Excel 2013<sup>3</sup> in the RB, go to the **Start** button and select **All Programs – MS Office 2013 – Excel 2103**. Click to open a **Blank Workbook**.

### 3.2 Simple statistics in Excel

Enter the same data as in Figure 2.1. (If you still have the data in SPSS you can use a short cut. Highlight the data in SPSS and press **Control-C** to copy them. Select the top left cell in Excel and press **Control-V** to paste them.) Your screen should look similar to Figure 3.1.


---

<sup>3</sup> In earlier versions, you simply open Excel (via MS Office if Excel is not listed under **All Programs**) and it will display a blank workbook automatically.

	A	B	C
1	1	27	
2	2	26	
3	3	29	
4	4	25	
5	5	24	
6	6	29	
7	7	28	
8	8	32	
9	9	30	
10	10	27	
11	11	34	
12	12	29	
13			

**Figure 3.1.** Excel spreadsheet with sample data.

In Excel, if we want titles we must enter them ourselves. First, we must make some room. Click on cell A1 to select it. Ensure you are in the **Home** tab. In **Cells** (towards the right the ribbon at the top of the window), click on the small arrow next to **Insert** to bring up the drop-down menu. Click on **Insert Sheet Rows**. A new blank row appears on the sheet.

Enter *Student* and *Age* in the appropriate cells. To make the headings line up with the data, highlight the headings (click on one of them, and hold down the left hand mouse button whilst sweeping over the other one). In **Alignment**<sup>4</sup> (on the top ribbon) click on the right-alignment icon ()<sup>5</sup>.

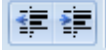
Now we can calculate some statistics. In cell B15 enter `=average(b2:b13)` and click **Enter**. Notice some things about the formula.

- The formula you entered appeared above the column headings as well as in the cell. The area where it appeared is called the *formula bar*. Once you pressed **Enter**, the formula in the formula bar turned into capitals. In the cell, the formula was replaced by the result of the calculation.

<sup>4</sup> This is under **Cells** in previous versions.

<sup>5</sup> As I have already mentioned, there is often more than one way to do things. If you are not familiar with Excel, you could experiment by clicking the little arrow at the bottom right of **Alignment**. This allows you to do the same thing, but also provides much more flexibility.



- All formulae in Excel start with '='.
- All formulae in Excel have a pair of brackets at the end. This is where you tell Excel about the data that you want it to work on.
- B2:B13 is the *range* of the cells for which you want to calculate the mean. It can be entered using the keyboard (as we did), or by highlighting the cells using the cursor.
- Although statisticians prefer the more precise term 'mean', Excel calls this the 'average'.
- The result is given to 5 decimal places – as many as will fit in the cell. Usually we only want to show one more decimal place than there was in the original data. You can increase or decrease the number of decimal places using the  icons in **Number**. Change it now to one decimal place<sup>6</sup>.

Excel is very flexible. That means it is also easy to make mistakes. For example, you need to be careful to enter the correct range of cells you want to calculate on.

As previously suggested, Excel is rather like a blank sheet of paper. If we want the reader to know that 28.3 is the mean, it is up to us to say so. Type the word *Mean* in cell A15.

Excel can also calculate the median [=median(b2:b13)], mode [=mode(b2:b13)] and standard deviation [=stdev(b2:b13)]. However, calculations like the interquartile range, or the inferential statistics we will cover in later weeks, are much harder to do.

### 3.3 Graphs in Excel

#### 3.3.1 Creating graphs

Excel is particularly good for graphs. Suppose that we have rainfall data over several months. Open a new Excel file: click **File – New**, and double-click **Blank Workbook**. Enter the data as shown in Figure 3.2.

To create the graph, click on the **Insert** tab. Highlight the data (including the headings). In **Charts**, select any of the main types and it will bring up a menu of subtypes. In Excel 2013<sup>7</sup> try **Column** (the top left icon), and the top left of the small icons. You should get a graph like Figure 3.3.

<sup>6</sup> Again, if you are not familiar with Excel you might want to click on the little arrow at the bottom right of **Number**, and experiment with some of the other options.

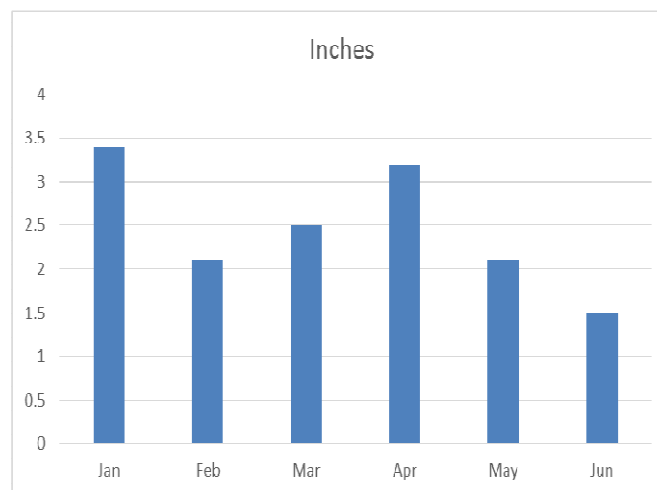
<sup>7</sup> In previous versions, select **Column** from the drop-down list, and the top left option.



	A	B
1	Month	Inches
2	Jan	3.4
3	Feb	2.1
4	Mar	2.5
5	Apr	3.2
6	May	2.1
7	Jun	1.5
8		

**Figure 3.2.** Data in Excel for sample graphs.

Notice that whenever you create a graph, a new pair of tabs (**Chart Tools: Design** and **Format**<sup>8</sup>) appear at the top. They also appear whenever you click on the chart to select it. I will have more to say about these tabs later, but for now notice that the **Design** tab came up. If you click on **Change Chart Type** (near the top right<sup>9</sup> of the ribbon) you can change your mind about what kind of graph you want. Try experimenting, for example with a line chart. Notice that Excel gives you a preview – you do not need to click **OK** to see what the chart would look like.



**Figure 3.3.** Default graph

<sup>8</sup> In earlier versions there are three; the additional tab is called **Layout** and (obviously) some items are to be found there.

<sup>9</sup> Left on earlier versions of Excel.

### 3.3.2 Editing and changing graphs

One of the strengths of creating graphs in Excel is its flexibility. If you can imagine a way you want a graph to look, there will probably be a way of doing it in Excel. Some of the ways of editing a graph are as follows; some of these are illustrated in Figure 3.4.

#### (a) Changing the data

If you make any changes to the data (including headings) they will automatically be carried through into the graph. Try changing *Inches* to *Centimetres*.

#### (b) Using the Chart Tools tabs

To use the **Chart Tools** tabs, select the graph. It is selected automatically when you first create it; otherwise you can left-click on it once with the mouse. As I mentioned above, there are two<sup>10</sup> special tabs under **Chart Tools**: **Design** and **Format**.

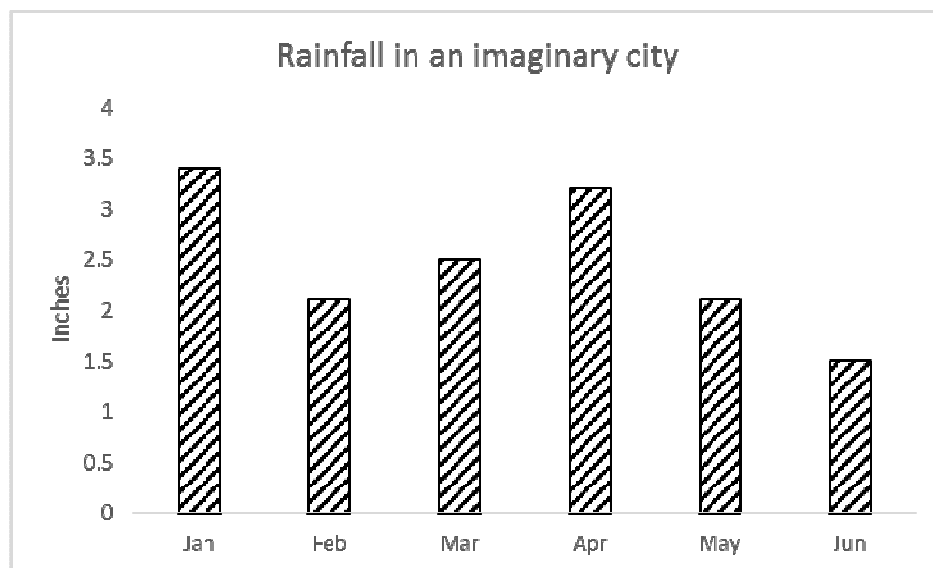


Figure 3.4. Amended bar chart.

For example, you can insert or remove a **Chart Title** (the overall title at the top, which says *Inches* in the default graph you have just created), **Axis Titles**, and/or

---

<sup>10</sup> In earlier versions there are three; the additional tab is called **Layout** and some items are to be found there.

a **Legend**<sup>11</sup>. To insert these (and/or other) elements in Excel 2013, go to the **Design** tab and click on **Add Chart Element**<sup>12</sup>. In Figure 3.4 I selected **Axis Titles, Primary Vertical**, then I edited the default wording (see next paragraph).

*(c) Typing on the graph itself*

Usually, wording can be changed by clicking on it and typing in your new wording<sup>13</sup>. In Figure 3.4 I have changed the wording on the **Primary Vertical Axis Title**, and the **Chart Title**.

*(d) The pop-up pane to the right of the window*

Another way of making changes (with the chart selected) is to hover with your mouse so that the names of different parts of the graph appear in the *screen tip* (the words in a box next to the insertion point). When you have found the item you want to change, there are several ways of accessing options. Double-clicking will bring up menus and/or icons in a right hand pane of the window (or if a right hand pane is already visible, you only have to single-click to change the features it refers to). Or if you right-click instead of left-clicking, you get a pop-up menu<sup>14</sup>; the last item on this (**Format**) brings up the same right-hand pane.

For example, in the chart area, in our default chart there is no fill<sup>15</sup>. If your chart is to appear in a research poster, you might want a coloured fill. Right-click on the plot area, and click on **Format Plot Area**. Click on the button next to **Fill**, then on the button for **Solid Fill**. Choose the **Color** you require. (If you are creating Figure 3.4, undo these actions.)

If on the other hand your chart is to go into a black and white document, you may want the bars to be picked out as patterns<sup>16</sup>. Click on one of the bars and make sure they are all selected (or if you have more than one series, as in section 3.3.3, select one series at a time). To do this, you may find you need to click away from the chart, and then back onto it. Then, in Excel 2013<sup>17</sup>, right-click on the bar, select **Format Data Series**, and click on the left hand icon near the top

---

<sup>11</sup> A Legend is useful when you have more than one data series. There is one on the right of the chart in Figure 3.6, for example. In some earlier versions of Excel a legend is created for this chart, even though it is not required.

<sup>12</sup> In earlier versions it is in **Layout**, the **Labels** tab.

<sup>13</sup> If this does not work, you may need to change the wording in the original data – see paragraph (a).

<sup>14</sup> In earlier versions, the effect of the right and left click buttons may be reversed, and where menus appear they may be overlaid on the window rather than to the side.

<sup>15</sup> Strictly speaking, there is a fill, but it is solid white. If there is no fill, the graph is transparent.

<sup>16</sup> This facility was disabled in Excel 2007, but is available in later as well as earlier versions of Excel.

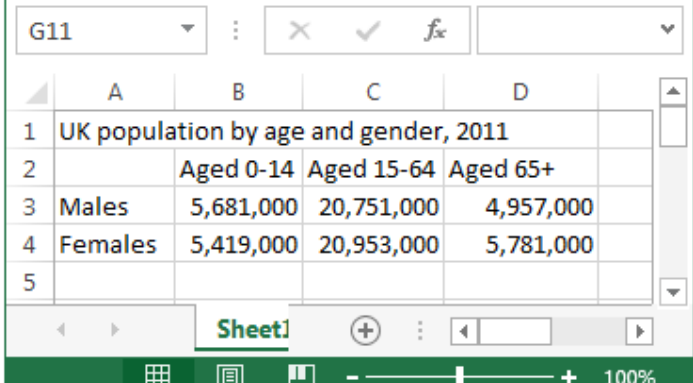
<sup>17</sup> In earlier versions: Right-click on the bar, select **Format Data Series**, and on the left hand side select **Fill**. Select **Pattern Fill**, and choose the fill you require. If necessary, change the **Foreground Color** and the **Background Color** to black and white. If the border is missing, select **Border Color** and change to **Solid Line**.

of the pane (**Fill and Line**). Under **Fill**, select **Pattern Fill** and choose the fill you require. If necessary, change the **Foreground Colour** and the **Background Colour** to black and white (further down the same pane). If the borders are missing from the bars, select **Border** (further again down the same pane) and change to **Solid Line**; change the colour to black if necessary.

One of the other options that comes up by right-clicking is to delete a feature altogether. I have done this for the Major Gridline (the lines across from the numbers on the vertical axis). The chart with all these changes is shown in Figure 3.4.

### 3.3.3 Bar charts with two independent variables/ data series

Bar charts are particularly useful when the data are split by two different variables, or (as Excel puts it) are in two data series. For example, consider the data in Figure 3.5<sup>18</sup>. Enter this into a new worksheet<sup>19</sup>. (By the way, I have enlarged the width of the columns so that all of the heading and numbers can be seen. This is only for our convenience; it does not affect the correct creation of the graph. One way to do this is to select the column headings – B, C and D in this case – and click on **Format – Autofit Column Width**. Also note that to avoid complications, I have put the word “Aged” at the beginning of each caption<sup>20</sup>.)



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D
1	UK population by age and gender, 2011			
2		Aged 0-14	Aged 15-64	Aged 65+
3	Males	5,681,000	20,751,000	4,957,000
4	Females	5,419,000	20,953,000	5,781,000
5				

Figure 3.5. Data with two independent variables

<sup>18</sup> Source: UK census, <http://www.ons.gov.uk/ons/rel/census/2011-census/population-and-household-estimates-for-the-united-kingdom/rft-table-1-census-2011.xls>

<sup>19</sup> To create a new worksheet within the same book, click on the + sign on the tabs at the bottom of the screen. I have of course suggested this only for convenience – Excel is totally flexible and does not mind how many different things you put onto one worksheet.

<sup>20</sup> In some older versions of Excel, if you enter simply “0-14”, Excel thinks you are trying to do a calculation. Should this problem ever arise, you can get round it by simply putting a single quote mark in front of the label you are trying to enter.

A graph such as the one in Figure 3.6 is produced just as before. Highlight the data and their titles (in this case, cells A2:D4) and insert a column graph just as you did in paragraph 3.3.1.

As always, you can edit the graph as required. An important additional feature is that you can change which way round the graph is displayed. For example, in this case you could put sex on the horizontal axis and the *age groups* as different colours. In the **Design** tab, click on **Switch Row/Column**.

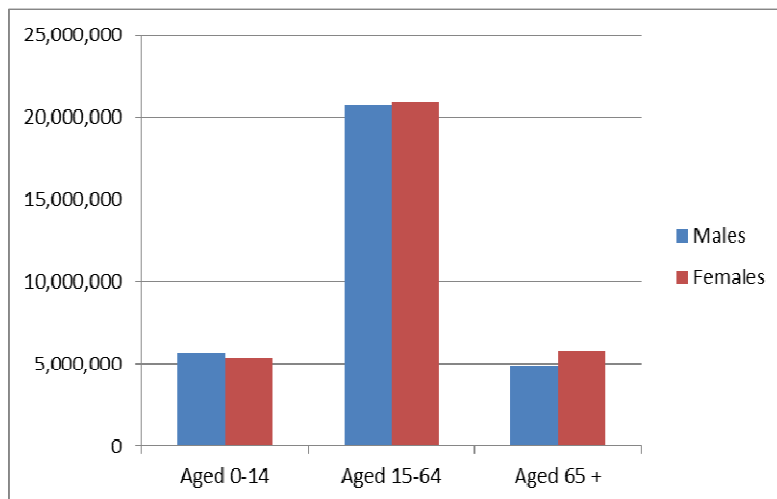


Figure 3.6. Chart with two independent variables.

## 4 Introduction to graphs in SPSS; Histograms; Chart Editor

### 4.1 Introduction; Chart Builder

In version 15 onwards of SPSS there is a versatile facility called Chart Builder. You may like to experiment with this. However, these versions also retain the previous methods of creating charts, under **Graphs – Legacy Dialogues**. Personally I think these are easier to teach, so I will use these ‘legacy’ dialogues.

### 4.2 Histograms

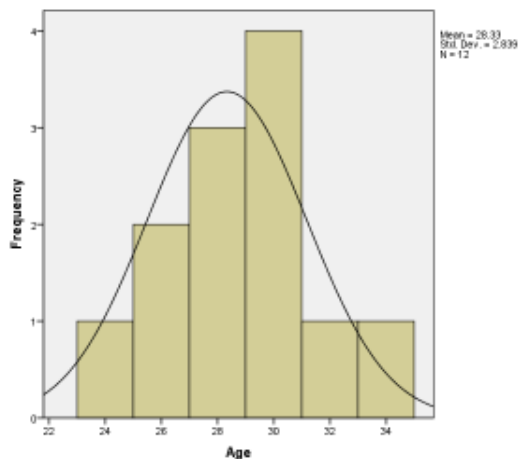
Histograms are a way of visualising a single variable. To produce a histogram in SPSS, for example for the data in Figure 2.1 (which you should have saved as Ages.sav):

- On the drop-down menu, click on **Graphs – Legacy Dialogs<sup>21</sup> – Histogram**
- Move the variable of interest (*Age*) into the **Variable** box

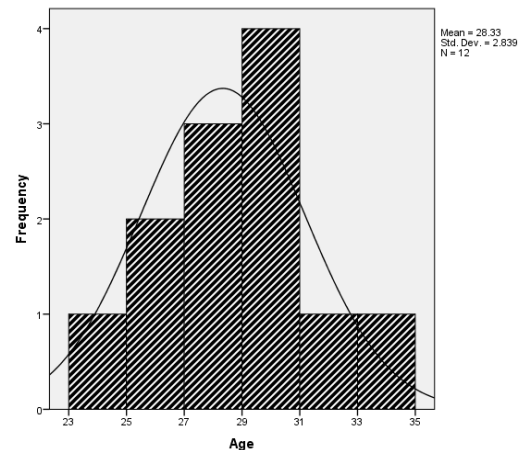
<sup>21</sup> If using SPSS 14 or earlier, omit this step

- Click the box to **Display normal curve** (if required. This helps to visualise whether the data are normal. If the sample was drawn from a normal population, you expect the histogram to lie close to the normal curve. As you will find out, this is quite hard to judge with small samples.)
- Click on **OK**.

The histogram appears in the **Output window**. As always, SPSS chooses default settings, shown here in Figure 4.1(a).



(a) before editing



(b) after editing

**Figure 4.1.** Histogram for the data in Figure 2.1.

### 4.3 Changing the appearance of a chart using the Chart Editor

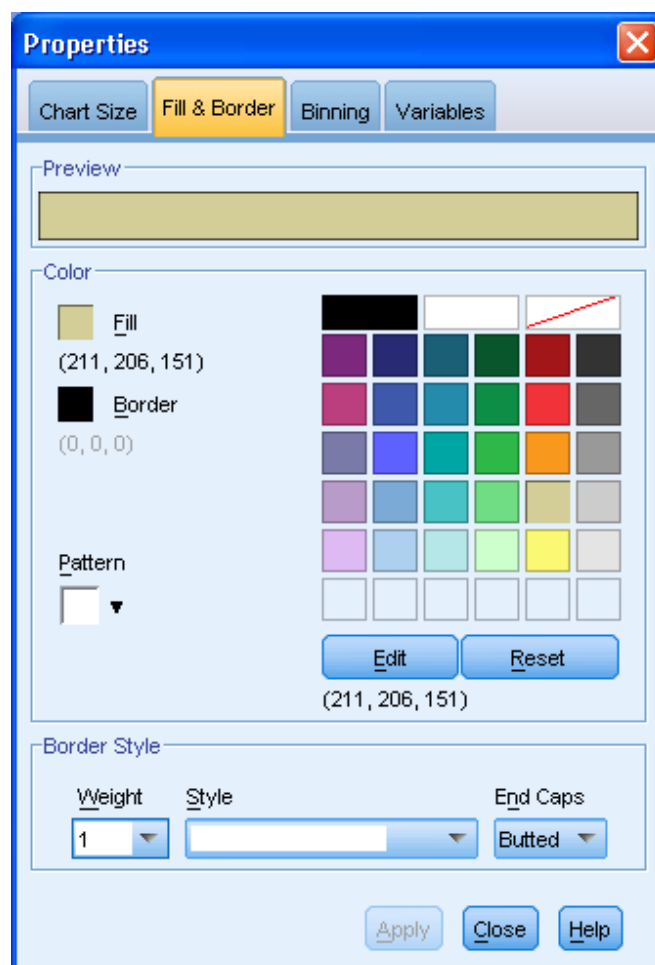
We can change the formatting and settings of any chart in SPSS using the **Chart Editor**. To open the Chart Editor, double-click on the graph. You can then bring up dialogue boxes to change various features, either from the drop-down menu at the top, or by double-clicking on the appropriate features. If double-clicking, you may need to single-click first, and/or try more than once in slightly different places to get the dialogue box you want.

For example, double-clicking on the numbers at the bottom brings up a Properties dialogue box with various things you can change relating to them. We don't need to change any of them on this histogram, but try some out to see how they work.

- I would prefer to show the ages at the edges of the bars rather than the middle. Click on the **Scale** tab, and change the **Minimum** to 23.
- Sometimes, SPSS shows more decimal places than you want. Click on the **Number Format** tab, and change **Decimal Places** to the number you require.

Here are some changes you can make by double-clicking on the bars themselves:

- In this case, SPSS has made a sensible choice on the *bins*, i.e. how many ages to collect together into one bar. However, you might not like SPSS's choice (for example, you might want a separate bar for each individual age). Choose the **Binning** tab<sup>22</sup> and (under X axis) click on **Custom**. You can choose either the number of bins, or the width of each bin.
- You can remove the colour of the bars, and/or add a pattern, much as we did in Excel. Choose the **Fill and Border** tab (Figure 4.2). Change the fill to white, by clicking on **Fill** and then the white patch. Click on the arrow to the right of **Pattern**, select the pattern you want, and click **Apply**. If you have finished, close the **Properties** box.



**Figure 4.2.** Dialogue box for changing colours and patterns.

When you have finished with the Chart Editor, click on the X at the top right to close it, and the edited chart will be saved back to the Output window.

<sup>22</sup> In SPSS 14, choose Histogram Options

## 4.4 Copying the chart into another application

You can copy and paste the chart into another application such as Word. . In the current version of SPSS, it is easiest to copy it from the Output window, using **Copy Special – Image**. For more tips on copying and pasting, see Appendix B.

# 5 t-tests, Anovas and their non-parametric equivalents

## 5.1 Introduction

This section deals with tests when one variable is categorical, and the other is continuous (ordinal, interval or ratio).

Often (but not always) the categorical variable is considered to be the Independent Variable (IV) and the continuous variable is considered to be the Dependent Variable (DV). See section 1.2 for further discussion of this. For example:

- the IV might be the amount of alcohol consumed (no alcohol, 1 unit, 2 units)
- the DV might be performance on a test (measured as a score)

Note: For this kind of test it is usually recommended that you have at least 20 cases (e.g. participants) for within-subjects designs, and 20 per condition for between-subject designs. This is just a rule of thumb – the best number of participants depends on various things including how big the effects are. In most of my illustration I use small numbers of cases, to make the data entry easier.

## 5.2 Which test to use?

Which test you use depends on the following factors, as shown in Table 5.1.

- the design (whether the categorical variable is repeated-measures or independent samples)
- the number of categories (or 'levels') of the categorical variables
- whether we presume that parametric assumptions are met for the continuous variable.

The table below covers the situations in this chapter (one categorical variable, one continuous). For a more comprehensive decision chart, see Appendix H.



**Table 5.1.** Choosing a test with one categorical variable and one continuous.

Categorical variable is: <sup>1</sup>	Categorical variable has how many categories ('levels')?	Parametric assumptions required for the continuous variable?	Test to use <sup>3</sup>
Repeated Measures (Within-subjects)	Two <sup>2</sup>	Yes	Paired-samples t-test
		No <sup>4</sup>	Wilcoxon (Signed Ranks) Test
	Two or more	Yes	Repeated measures Anova
		No <sup>4</sup>	Friedman test
Independent Samples (Between-Subjects)	Two <sup>2</sup>	Yes	Independent samples t-test
		No <sup>4</sup>	Mann-Whitney U test
	Two or more	Yes	Independent-samples Anova
		No <sup>4</sup>	Kruskal-Wallis test

#### Notes

- 1 It is almost always true to say that repeated-measures means the same as within-subjects, and that independent-samples means the same as between-subjects. Most people tend to use the terms interchangeably. There are a few exceptions, which are quite rare. For example, if there are control participants who are *individually* matched to each participant in the experimental condition, a repeated measures analysis is applicable.
- 2 You may wonder why we bother with tests for two levels. Surely you can use the tests that are for 'two or more' levels? Yes, you can. However, if there are only two levels, most people still use the separate tests (essentially for historical reasons) so you still need to understand them.
- 3 Unfortunately, most of these tests have more than one name. These are the names that SPSS uses.
- 4 These tests can be used whether parametric assumptions are met or not. However they are (slightly) less powerful than the parametric tests, so researchers prefer to use the parametric tests if possible.

### 5.3 Entering Repeated Measures data

Remember the rule of thumb: enter whatever we know about one case (e.g. one participant) on one line.

Enter the data in **Data View**. For our example, use the values in Figure 5.1.

	Participant	No_alc	Alc_1unit
1	1	9.6	9.5
2	2	9.0	8.4
3	3	11.9	11.7
4	4	8.6	8.2
5	5	9.4	9.3
6	6	11.6	11.8
7	7	9.5	9.1
8	8	11.6	11.0
9	9	11.2	11.0
10	10	10.3	9.3
11			

**Figure 5.1.** Data for paired samples t-test

In **Variable View**, give the variables:

- suitable names (for this example, *Participant*, *No\_alc*, *Alc*)
- fuller names (under **Label**): Participant number, No alcohol, Alcohol.
- suitable numbers of decimal places (0 for *Participant*, 1 for *No\_Alc* and *Alc*).

Save the file (e.g. on your home Goldsmiths drive) as *RMexample.sav* so we can use it again.

## 5.4 Paired samples t-test

(also known as related, matched pairs, within-subjects or repeated measures t-test)

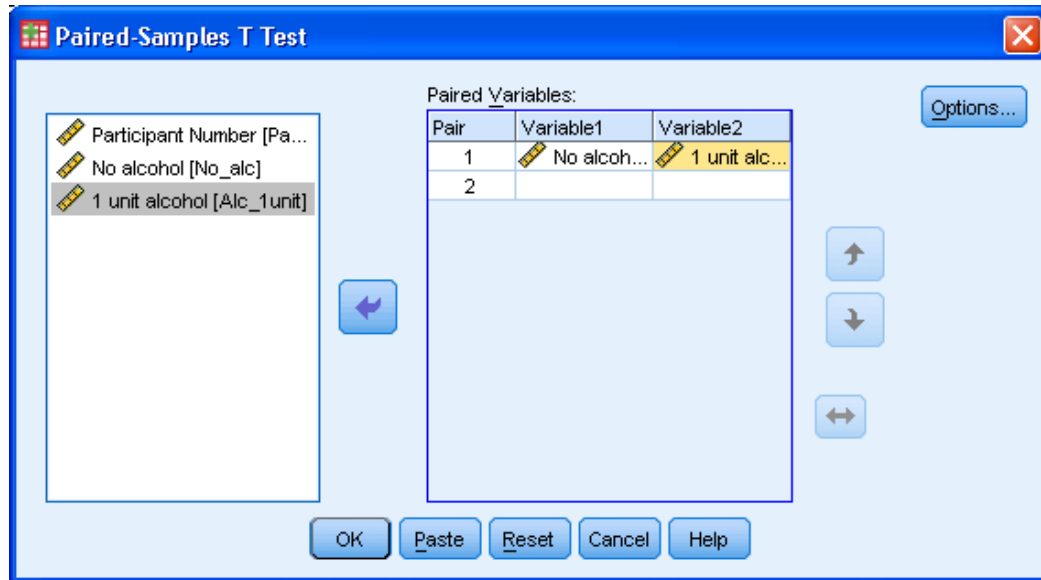
*Categorical variable: two categories, repeated-measures*

*Continuous variable: parametric assumptions made.*

On the drop-down menu, go to **Analyze – Compare Means – Paired-samples t test**. A dialogue box comes up. When you have completed the following, it will look like Figure 5.2. Click on the two categories you want to compare (in this case *No alcohol* and *1 unit alcohol*) and click them into the box marked *Paired Variables*. Click on **OK**.

*More advanced point for future reference:*

You can do more than one of these tests at the same time. Be careful that each pair of conditions you want to compare is on one line.



**Figure 5.2.** Dialogue box for paired samples t-test.

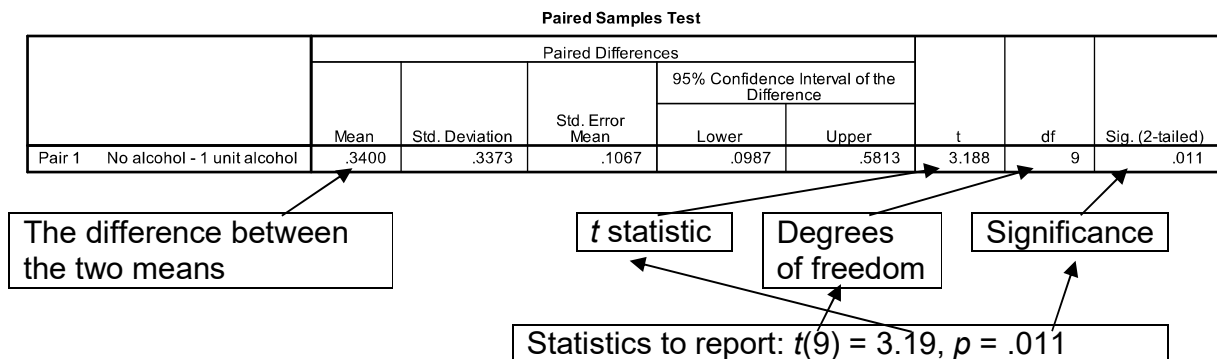
Examine the output. The first table (Figure 5.3) gives descriptive statistics.

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	No alcohol	10.270	10	1.2139	.3839
	1 unit alcohol	9.930	10	1.3300	.4206

**Figure 5.3.** Paired samples t-test descriptive statistics.

Ignore the second table. The third table (Figure 5.4) gives inferential statistics.

We consider a difference to be statistically significant if the significance level is less than 5% (i.e. less than .050). The significance level here is .011, so the difference is statistically significant.



**Figure 5.4.** Paired sample t-test results and how they are reported.

So, the results might be reported as follows: “With no alcohol, participants’ mean score was 10.27 ( $SD = 1.21$ ), and with alcohol the mean was 9.93 ( $SD = 1.33$ ). This difference was statistically significant,  $t(9) = 3.19, p = .011$ .” This way of reporting would, like the reporting of other tests in this booklet, accord with APA style. For more details of APA style, see Appendix A.

## Bar chart

You might want to illustrate the outcome with a bar chart. This is easily done. On the drop down menu, click on **Graphs – Legacy Dialogs – Bar**. A dialogue box appears (Figure 5.5). Click on **Simple; Summaries of separate variables**; and the **Define** button. In the next dialogue box, move the variables of interest (*No alcohol* and *1 unit alcohol*) into the **Bars Represent** box. (Note that the word ‘MEAN’ is shown, to confirm that SPSS will plot the means. We could have changed this, but the means are what we want.) Click **OK**.

The chart appears (Figure 5.6). As before, you can edit it by double-clicking to open the **Chart Editor**.

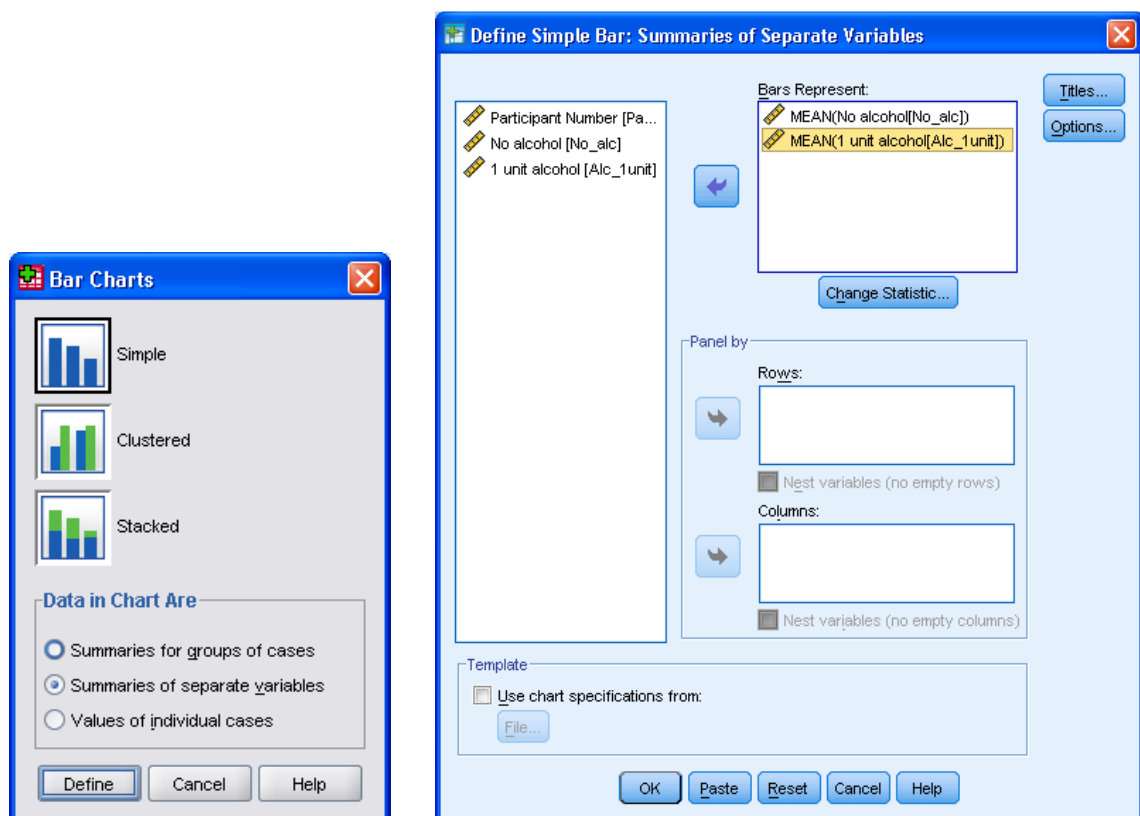


Figure 5.5. Bar charts dialogue boxes.

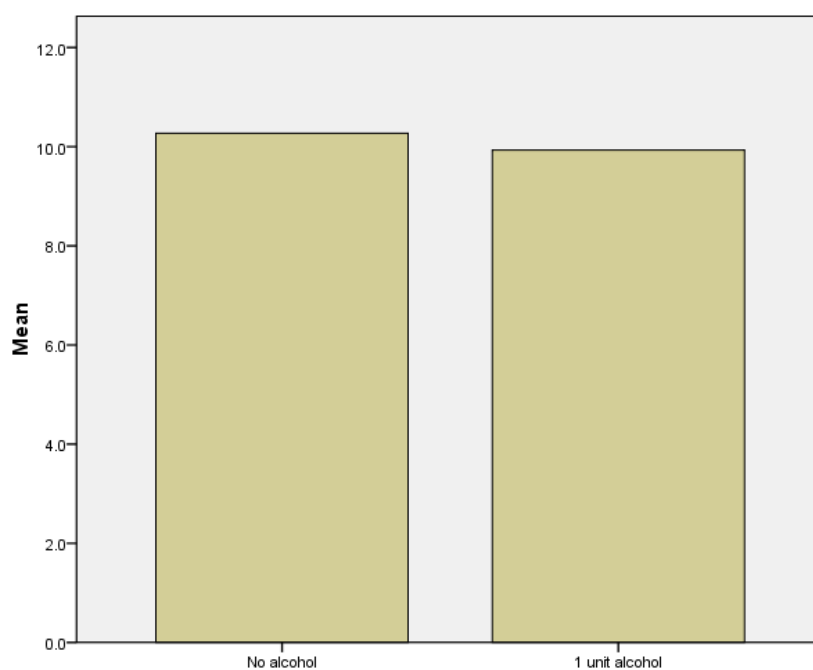


Figure 5.6. Default bar chart for paired sample data.

## 5.5 Wilcoxon (Signed Ranks) test

(also called the Wilcoxon matched pairs test)

*Categorical variable: two categories, repeated-measures*

*Continuous variable: parametric assumptions not required.*

For this example, we will use the same data as before (Figure 5.1), which you should have saved as RMexample.sav):

The method we will use is as follows<sup>23</sup>. On the drop-down menu go to **Analyze – Nonparametric tests – Legacy Dialogs – 2 Related Samples**. Move the two categories you want to compare (in this case *No alcohol* and *1 unit alcohol*) into the **Test pair(s)** box in a similar way to before (Figure 5.2). Click on **OK**. Examine the output. The figures we will report are from the last table: (Figure 5.7).

Test Statistics <sup>b</sup>	
	Alcohol - No alcohol
Z	-2.406 <sup>a</sup>
Asymp. Sig. (2-tailed)	.016

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

**Figure 5.7.** Wilcoxon test output.

You could report this as follows: “A Wilcoxon Signed Ranks test showed a significant difference in score between the groups,  $Z = 2.41$ ,  $p = .016$ .” (Note that when reporting  $Z$  we can ignore any negative sign.)

When reporting the results of non-parametric tests it is usual to report medians rather than means. We saw how to obtain these in section 2.3. As a reminder: Click on **Analyze – Descriptive Statistics – Frequencies**, and move the variables of interest into the box marked **Variable(s)**. Click on **Statistics**, and check the box that says **Median**. Click **Continue**, uncheck the box that says **Display frequency tables** and click **OK**. We get the output shown in Figure 5.8.

Statistics		No alcohol	1 unit alcohol
N	Valid	10	10
	Missing	0	0
Median		9.950	9.400

**Figure 5.8.** Descriptive statistics.

<sup>23</sup> There is a more direct menu option (Analyze – Nonparametric tests – Related Samples). A dialogue box appears; click on the Fields tab and put the conditions into the ‘Test Fields’ box. However, the output does not include the value of  $Z$ , which you may require for your report.

We would add this information to our report, e.g. “With no alcohol, the participants’ median score was 9.95, and with alcohol the median was 9.40.”

## 5.6 Repeated Measures Anova

*Categorical variable: two or more categories, repeated-measures  
Continuous variable: parametric assumptions made.*

Suppose that there are three categories (‘levels’) of the categorical variable. Extending our previous example, we might have tested participants with no alcohol, one unit of alcohol and two units of alcohol. In Data View, add in a further column with the results for two units, as shown in Figure 5.9. Give the new variable the Name *Alc\_2units* and the Label *2 units alcohol*.

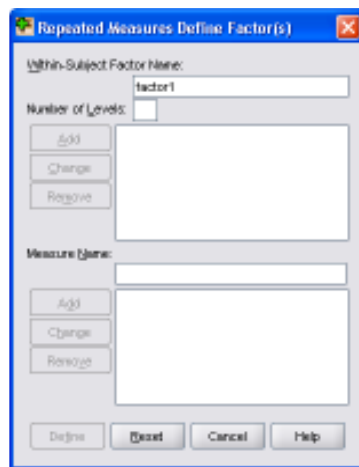
	Participant	No_alc	Alc_1unit	Alc_2units	va
1	1	9.6	9.5	8.9	
2	2	9.0	8.4	8.8	
3	3	11.9	11.7	11.7	
4	4	8.6	8.2	8.1	
5	5	9.4	9.3	9.0	
6	6	11.6	11.8	11.4	
7	7	9.5	9.1	8.6	
8	8	11.6	11.0	10.8	
9	9	11.2	11.0	11.0	
10	10	10.3	9.3	9.2	
11					

Figure 5.9. Data for Repeated Measures Anova.

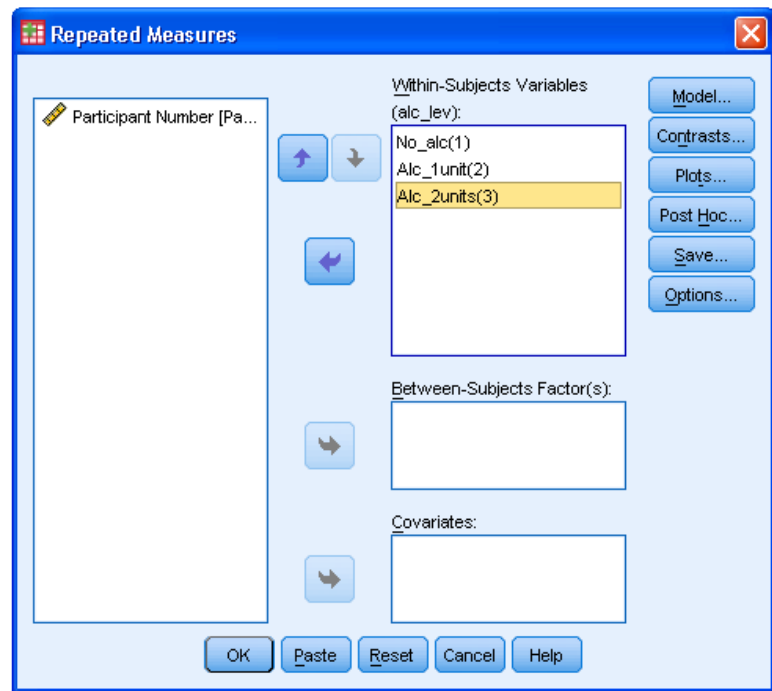
To do the test, click on **Analyze – General Linear Model – Repeated Measures**. A dialogue box appears (Figure 5.10).

Replace *factor 1* by the name you want to call the categorical variable. We will call it *alc\_lev*. In **number of levels** enter the number of levels (i.e. categories) of the categorical variable; 3 in this case (*no\_alc*, *alc\_1unit*, *alc\_2units*). Click on **Add**, then **Define**.

Another dialogue box appears. Click on the names representing our three categories ('levels') of the categorical variable and move them into the box headed **Within-Subjects Variables**, as in Figure 5.11.

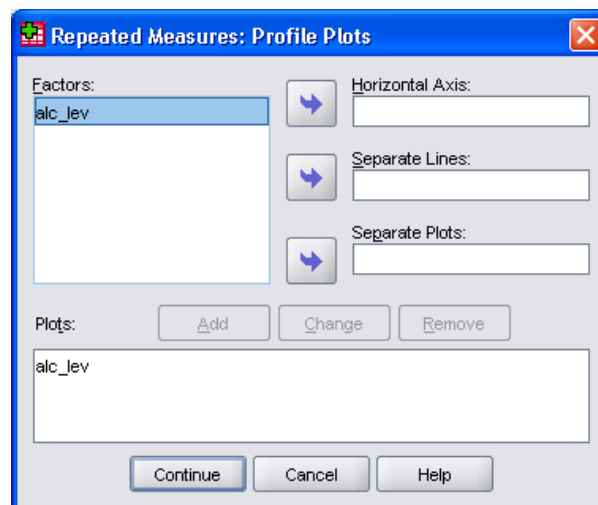


**Figure 5.10.** Define Factors dialogue box for Repeated Measures Anova.



**Figure 5.11.** Second dialogue box for Repeated Measures Anova.

Click on **Options** and check the box marked **Descriptive statistics**. Click **Continue**. Click on **Plots** and a new dialogue box appears (Figure 5.12). Move *alc\_lev* to the box headed **Horizontal Axis**. Click **Add**, then **Continue** and **OK**.



**Figure 5.12.** Plots dialogue box for Repeated Measures Anova.



### Effect size.

If you want a measure of effect size (see later lecture), when you click on **Options** also check the box that says **Estimates of effect size**. Click **Continue**. A measure of effect size, Partial Eta Squared, is shown in an extra column on the right. For this Anova, partial eta squared is roughly equivalent to the square of the correlation coefficient. Correlation coefficients and their squares will be discussed in the lecture on regression and correlation.

Examine the output. As usual, we do not need all of it.

Figure 5.13 shows the first two tables of the output. The first can be used to check we compared the conditions we wanted to. The second gives us the descriptive statistics. Ignore the one which says 'Multivariate tests.'

Mauchley's test of Sphericity (Figure 5.14) is a rather complex assumption associated with this Anova. But all we need to do with Mauchley's is to look at the significance level (under "Sig"). If Mauchley's is **not** significant (if  $p > .05$ ) we are happy. (If the categorical variable has only two categories, then Mauchley's is irrelevant, and a dot is printed in place of the significance level indicating that it cannot be calculated. We are happy in this case also.) If Mauchley's test **is** significant (if  $p < .05$ ), we need to report the results differently – see below.

#### Within-Subjects Factors

Measure: MEASURE\_1

alc_ lev	Dependent Variable
1	No_alc
2	Alc_1unit
3	Alc_2units

A reminder of the names we gave the levels of the IV.

#### Descriptive Statistics

	Mean	Std. Deviation	N
No alcohol	10.270	1.2139	10
1 unit alcohol	9.930	1.3300	10
2 units alcohol	9.750	1.3218	10

Descriptive statistics.

**Figure 5.13.** Relevant output from Repeated Measures Anova (part 1).

Mauchly's Test of Sphericity<sup>b</sup>

Mauchley's test – see text

Measure: MEASURE\_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon <sup>a</sup>		
					Greenhouse e-Geisser	Huynh-Feldt	Lower-bound
alc_lev	.964	.291	2	.865	.966	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

- May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.
- Design: Intercept

Within Subjects Design: alc\_lev

**Figure 5.14.** Relevant output from Repeated Measures Anova (part 2).

The actual Anova result is given in the table headed 'Tests of Within-subjects effects' (Figure 5.15).

Tests of Within-Subjects Effects					
Measure: MEASURE_1					
Source		Type III Sum of Squares	df	Mean Square	Sig.
alc_lev	Sphericity Assumed	1.395	2	.697	.000
	Greenhouse-Geisser	1.395	1.931	.722	.000
	Huynh-Feldt	1.395	2.000	.697	.000
	Lower-bound	1.395	1.000	1.395	.005
Error(alc_lev)	Sphericity Assumed	.932	18	.052	
	Greenhouse-Geisser	.932	17.380	.054	
	Huynh-Feldt	.932	18.000	.052	
	Lower-bound	.932	9.000	.104	

$F(2, 18) = 13.47, p < .001$

**Figure 5.15.** Anova result.

If Mauchley's test **is not** significant – as in this case – we take our figures from the lines marked 'Sphericity Assumed'. If Mauchley's **is** significant, we take our figures from the lines marked 'Greenhouse-Geisser'<sup>24</sup>. Remember, Mauchley's test is just to tell us which line to look at.

We need to report the following:

- Whether the result is statistically significant (yes in this example).
- the value of  $F$  (13.47 in this example).
- the degrees of freedom. There are now two to report:
  - one from the first line that says 'Sphericity Assumed' (against the name of our variable)

<sup>24</sup> In later lectures we will cover other possible options, such as transforming the data.

- and one from the second line that says 'Sphericity Assumed' (against the line that says 'Error' followed by the name of our variable).

In this example, they are 2 and 18.

- The significance level. SPSS has calculated this as .000. Remember we write this as '<.001').

Thus we can write: "A repeated-measures Anova showed a significant effect of alcohol on the score,  $F(2,18) = 13.47$ ,  $p < .001$ ."

*If Mauchley's test had been significant we could have written "A repeated-measures Anova with Greenhouse-Geisser correction showed a significant effect of alcohol on the score,  $F(1.9, 17.4) = 13.47$ ,  $p < .001$ ).*

In either case we must remember to report the descriptive statistics, e.g. "The mean scores (and standard deviations) with no alcohol, one unit and two units respectively were 10.27 (1.21), 9.93 (1.33) and 9.75 (1.32)."

You may find the chart (Figure 5.16) useful, but it will require editing if you want to publish it (e.g. remove the references to 'estimated marginal means'); you may also prefer to change it into a bar chart).

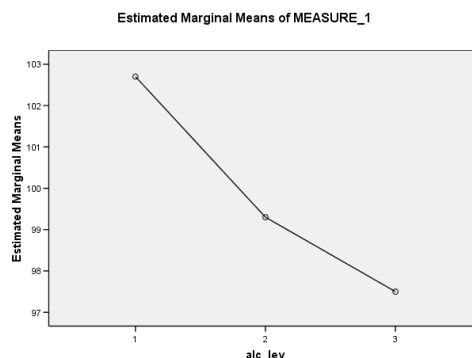


Figure 5.16. Default chart for repeated measures Anova.

## 5.7 Friedman test.

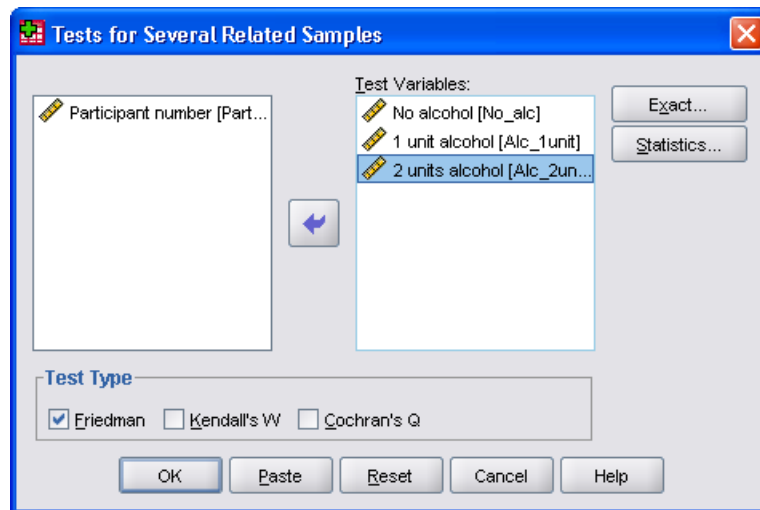
*Categorical variable: two or more categories, independent-samples*

*Continuous variable: parametric assumptions not required.*

Click on **Analyze – Nonparametric tests – Legacy Dialogs<sup>25</sup> – K Related Samples**. A dialogue box appears. As before, click on the names representing

<sup>25</sup> In versions of SPSS before 18, ignore the 'Legacy Dialogs' step. In SPSS 18 onwards there is a more direct menu option available (Analyse – Nonparametric tests – Related Samples). A dialogue box appears; click on the Fields tab, put the conditions into the 'Test Fields' box, and press 'Run'. However, the output does not include the value of chi-square or the degrees of freedom, which you may require for your report.

our three categories (levels) of the categorical variable and move them into the box headed **Test Variables**, so that the box looks like Figure 5.17.



**Figure 5.17.** Dialogue box for Friedman test.

Click on **OK**. The figures we report are from the 'Test Statistics' (Figure 5.18).

Test Statistics <sup>a</sup>	
N	10
Chi-Square	15.368
df	2
Asymp. Sig.	.000

Chi-square (2) =  
15.37,  
p < .001

a. Friedman Test

**Figure 5.18.** Output for Friedman test.

We could report the result as follows: “A Friedman test showed a significant effect of alcohol on the score, chi-square (2) = 15.37,  $p < .001$ .”<sup>26</sup>

As this is a non-parametric test, we would report the medians in each condition: “The median scores with no alcohol, one unit and two units respectively were 9.95, 9.40 and 9.10.” (See section 2.3 for a reminder of how to obtain these.)

## 5.8 Independent-samples data - general

### 5.8.1 Entering independent-samples data

Firstly, let us create an example data file. Remember the rule of thumb, that each line relates to one case (participant). So for each participant we will show their participant number, which condition they were in, and their score.

<sup>26</sup> To be more sophisticated, write chi-square in symbols (although it doesn't show up very well in this font): “A Friedman test showed that there was a significant effect of alcohol,  $\chi^2(2) = 15.37$ ,  $p < .001$ .” See Appendix A for how to do this.

Note carefully how this differs from a repeated-measures design. In an independent-samples design one of the columns represents a categorical variable (the condition).

Enter our sample data into **Data View**, as shown in Table 5.2. If you do this correctly, you will have 30 lines.

Table 5.2. **Sample data for independent-samples tests.**

Part	Group	Score
1	1	107
2	1	112
3	1	99
4	1	91
5	1	86
6	1	85
7	1	106
8	1	81
9	1	121
10	1	99
11	2	80
12	2	81
13	2	82
14	2	64
15	2	102

Part	Group	Score
16	2	88
17	2	97
18	2	80
19	2	90
20	2	71
21	3	98
22	3	69
23	3	85
24	3	98
25	3	81
26	3	83
27	3	99
28	3	84
29	3	83
30	3	95

In **Variable View**, give the first three variables the Names *Part*, *Group* and *Score*. Change the number of decimals to 0.


SPSS works with numbers, but to make the analysis clear to ourselves we will want to give each of the conditions a label (as in section 2.4). Participants in Group 1 had no alcohol, Group 2 had 1 unit, and Group 3 had 2 units. Go back into Variable View. In the line that says *Group*, click in the **Values** cell. Three dots come up, as shown in Figure 5.19. Click on them, and a dialogue box appears (Figure 5.20).

In **Value**, type 1. In **Value Label** type *No alcohol*, then press the **Add** button. Repeat the process with 2 for '1 unit' and 3 for '2 units'. When you have finished, click on **OK**.

You can choose whether Data View shows the numbers (1, 2 and 3) or the labels (*No alcohol*, *1 unit*, *2 units*). Click on **View**, and click against **Value Labels**.

Alternatively, click on the **Labels** icon (<sup>27</sup>). Either way, you will toggle between the two representations.

---

<sup>27</sup> In earlier versions of SPSS, the icon looks like this: .

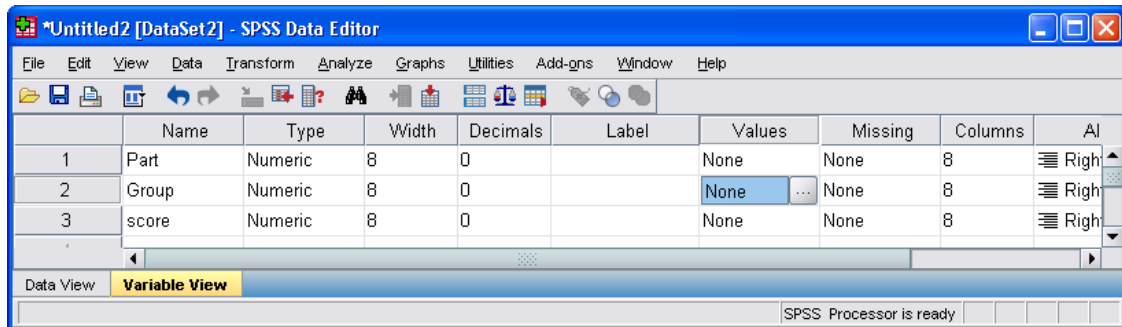


Figure 5.19. Variable View with Values cell selected.

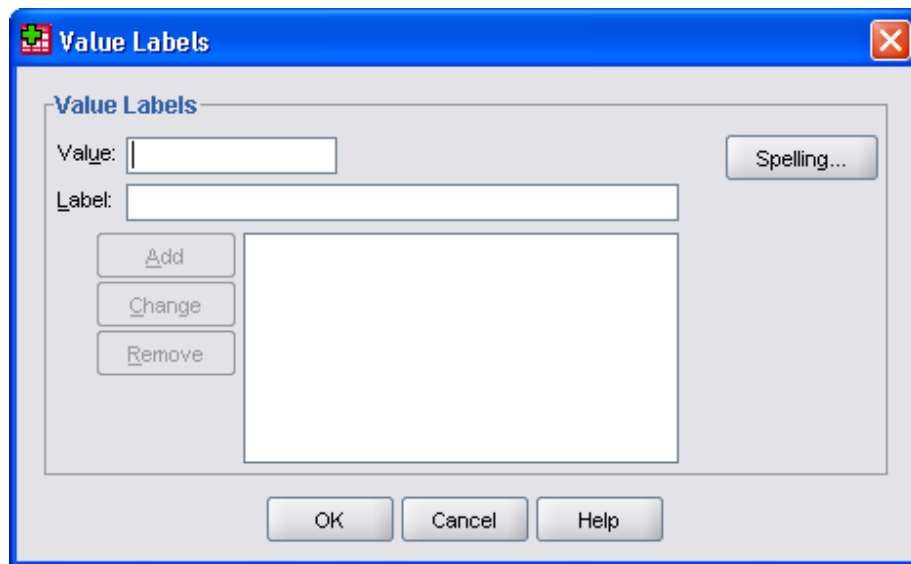


Figure 5.20. Value Labels dialogue box.

### 5.8.2 Descriptive statistics and histograms

Descriptive statistics are a bit harder to get with an independent-samples design. Often they are included with the test output. However, we might want the descriptive statistics without doing a test, or we might want the medians.

Click on **Data – Split file** and the Split File dialogue box appears (Figure 5.21).

Click on **Organise output by group**. Move the categorical variable we want to split by (*Group* in this case) into the **Groups based on** box. Click on **OK**. Now you can ask for your descriptive statistics in the normal way (**Analyze – Descriptive Statistics – Frequencies**; move *Score* into the **Variable(s)** box; uncheck **Display Frequency Tables** and click on **Statistics**; tell SPSS what statistics you want.) The output provides the statistics separately for each group.

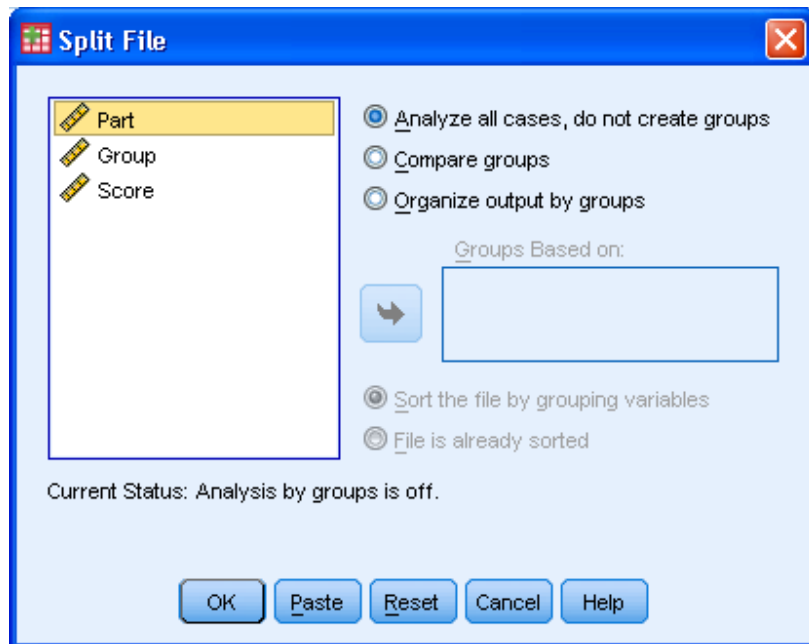


Figure 5.21. Split file dialogue box.

If you want histograms for each group, you can use the same procedure. Then you can go to **Graphs – Legacy Dialogs – Histogram** and move *score* into the **Variable** box in the normal manner.

Before you do any further analysis, go back to **Data – Split file** and click on **Analyze all cases**.

## 5.9 Independent-samples t-test

(also known as between-subjects t-test)

*Categorical variable: two categories, independent-samples*

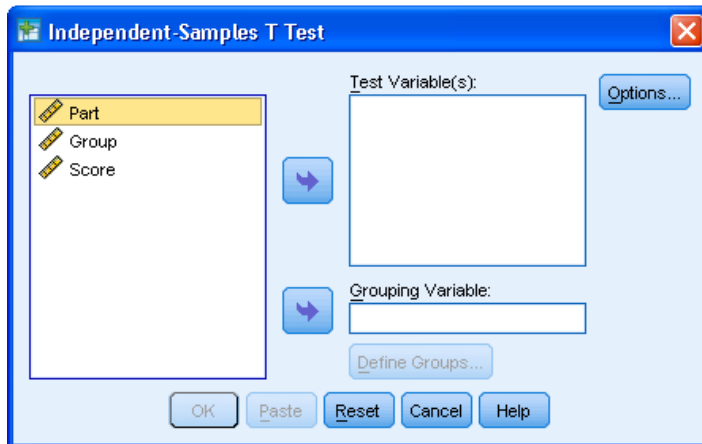
*Continuous variable: parametric assumptions made.*

Suppose for now we had only tested groups 1 and 2.

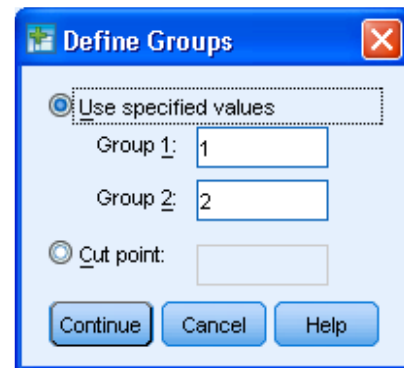
Click on **Analyze – Compare Means – Independent Samples T test**. A dialogue box appears (Figure 5.22).

Move the continuous variable (*Score* in this case) into **Test Variable(s)**. Move the categorical variable (*Group* in this case) into **Grouping Variable**, and the **Define Groups** button lights up. Press it.

A further dialogue box appears. Enter the numbers of the two groups we are comparing (1 and 2 in this case, as shown in Figure 5.23). Press **Continue** and **OK**.



**Figure 5.22.**  
Independent-Samples t-test dialogue box.



**Figure 5.23.**  
Define Groups dialogue box.

The output appears. The first item is our descriptive statistics (Figure 5.24).

Group Statistics					
Group		N	Mean	Std. Deviation	Std. Error Mean
Score	No alcohol	10	98.70	12.988	4.107
	1 unit	10	83.50	11.336	3.585

**Figure 5.24. Descriptive statistics.**

The test output for our independent samples t-test is shown in Figure 5.25. It is slightly more complicated than we saw for the paired-samples t-test.

The first thing we have to do is to see whether **Levene's test** is significant. If it is not significant, we are happy. If it is significant, one of the assumptions of the test has been violated, but this is not a serious problem as we can use a corrected result.

In this case the result of Levene's test is not significant, and we use the figures from the line that says 'Equal variances assumed'. If Levene's test is significant, we need to use the corrected result. This is not a problem, because the corrected result is printed on the second line: 'Equal variances not assumed'.

In this case we can report that "The mean score of the participants who did not drink alcohol was 98.7 (SD = 13.0) and that of the participants who drank alcohol was 83.5 (SD = 11.3). This difference was statistically significant,  $t(18) = 2.79$ ,  $p = .012$ ."



Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Score	Equal variances assumed	.324	.577	2.788	18	.012	15.200	5.451	3.747	26.653
	Equal variances not assumed			2.788	17.677	.012	15.200	5.451	3.732	26.668

Levene's test tells us which line to read (see text)

Report these statistics from the appropriate line:  
 $t$ , degrees of freedom,  $p$ .  
In this case we use the top line:  
 $t(18) = 2.8, p = .012$ .

Figure 5.25. Test output and interpretation.

## 5.10 Mann-Whitney U test

*Categorical variable: two categories, independent-samples*

*Continuous variable: parametric assumptions not required.*

Click on **Analyze – Nonparametric Tests – Legacy Dialogs<sup>28</sup> – 2 Independent Samples**. Move the continuous variable (*Score*) into **Test Variable list** and the categorical variable (*Group*) into **Grouping Variable**. Again, the **Define Groups** box lights up; say which are the groups we want to compare (1 and 2 in this case). Click on **Continue** and **OK**. We want the inferential statistics at the end of the output (Figure 5.26).

We can report this as “A Mann-Whitney U test revealed a significant difference in score between the groups,  $Z = 2.46, p = .014$ .” We would also report the median scores.

Test Statistics <sup>b</sup>	
	Score
Mann-Whitney U	17.500
Wilcoxon W	72.500
Z	-2.460
Asymp. Sig. (2-tailed)	.014
Exact Sig. [2*(1-tailed Sig.)]	.011 <sup>a</sup>

These are the figures we report (ignore any minus sign)

a. Not corrected for ties.  
b. Grouping Variable: Group

Figure 5.26. Mann-Whitney U test results.

<sup>28</sup> In versions of SPSS before 18, ignore the ‘Legacy Dialogs’ step. In SPSS 18 onwards there is an alternative route, but it has a number of complications and may not give you all the information you need to report.

## 5.11 Independent-samples Anova

Also known as between-subjects Anova.

*Categorical variable: two or more categories, independent-samples*

*Continuous variable: parametric assumptions made*

Now let us turn to an independent-samples test we can use with more than two conditions.

Click on **Analyze – General Linear Model – Univariate**. A dialogue box comes up. Move the continuous variable (*Score*) into the **Dependent Variable** box, and the categorical variable (*Group*) into the **Fixed Factor(s)** box, as shown in Figure 5.27.

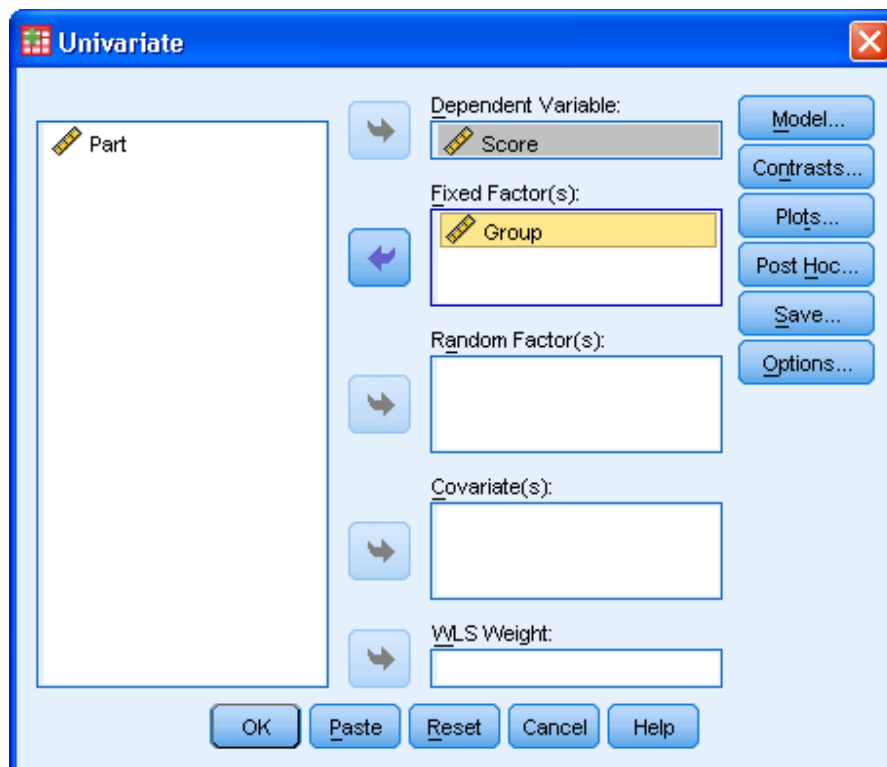


Figure 5.27. Dialogue box for independent-samples Anova.

Click on **Options**, and tick **Descriptive Statistics** and **Homogeneity Tests**, then click on **Continue**. Click on **Plots** and move the categorical variable (*Group*) into **Horizontal Axis**. Click **Add** and **Continue**. Finally back in the original dialogue box, click on **OK**.

### **Effect size.**

If you want a measure of effect size, click on **Options** and check the box that says **Estimates of effect size**. Click **Continue**. A measure of effect size, Partial Eta Squared, is shown in an extra column on the right. For this Anova, partial eta squared is roughly equivalent to the square of the correlation coefficient. Correlation coefficients and their squares will be discussed in the lecture on regression and correlation.

We are interested in the following output. Figure 5.28 provides the descriptive statistics for our report.

### **Descriptive Statistics**

Dependent Variable: Score

Group	Mean	Std. Deviation	N
No alcohol	98.70	12.988	10
1 unit	83.50	11.336	10
2 units	87.50	9.733	10
Total	89.90	12.823	30

**Figure 5.28.** Descriptive statistics.

Levene's test (Figure 5.29) tells us whether one of the assumptions of the Anova has been violated. We want it *non*-significant, as here. If it is significant, do the non-parametric test (Kruskal-Wallis – see below) instead<sup>29</sup>.

### **Levene's Test of Equality of Error Variances<sup>a</sup>**

Dependent Variable: Score

F	df1	df2	Sig.
.378	2	27	.689

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+Group

**Figure 5.29.** Results of Levene's test.

Figure 5.30 shows the figures for the Anova itself. So we can report this result as "There was a significant effect of alcohol on the score,  $F(2,27) = 4.75$ ,  $p = .017$ .

<sup>29</sup> Or see Appendix E for some more advanced options.

Tests of Between-Subjects Effects					
Dependent Variable: Score					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1241.600 <sup>a</sup>	2	620.800	4.752	.017
Intercept	242460.300	1	242460.300	1856.037	.000
Group	1241.600	2	620.800	4.752	.017
Error	3527.100	27	130.633		
Total	247229.000	30			
Corrected Total	4768.700	29			

a. R Squared = .260 (Adjusted R Squared = .206)

$$F(2,27) = 4.75, p = .017.$$

**Figure 5.30.** How to report test results for the independent-samples Anova.

The output also includes a graph, which you may wish to edit.

## 5.12 Kruskal-Wallis test

*Categorical variable: two or more categories, independent-samples*

*Continuous variable: parametric assumptions not required.*

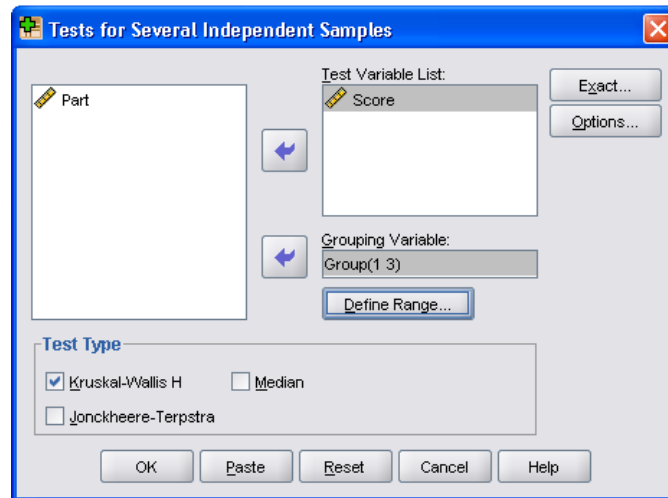
Click on **Analyze – Nonparametric tests – Legacy Dialogs<sup>30</sup> – K Independent Samples**. A dialogue box appears. Move the continuous variable (*Score*) into the **Test Variable** box and the categorical variable (*Group*) into the 'grouping variable' box. Click on **Define Range** and enter the highest and lowest numbers we used to define the groups: 1 as the **Minimum** and 3 as the **Maximum** in this case. Click **Continue** and return to the dialogue box, which should now look like Figure 5.31.

Click on **OK**. The important piece of output is shown in Figure 5.32.

We could report the result as follows: "A Kruskal-Wallis test showed a significant effect of alcohol on the score, chi-square (2) = 7.83,  $p = .020$ ."<sup>31</sup> As this is a non-parametric test, again we would report the medians in each condition. (See 5.8.2 for how to obtain these.)

<sup>30</sup> In versions before SPSS 18, ignore the 'Legacy Dialogs' step. In SPSS 18 there is an alternative route, but it has complications; see footnote 28.

<sup>31</sup> To be more sophisticated, write the  $\chi^2$  in symbols: "A Kruskal-Wallis test showed that there was a significant effect of alcohol on the score,  $\chi^2(2) = 7.83, p = .020$ ." See Appendix A for how to do this.



**Figure 5.31.**  
Kruskal-Wallis dialogue box.

**Test Statistics<sup>a,b</sup>**

	Score
Chi-Square	7.829
df	2
Asymp. Sig.	.020

a. Kruskal Wallis Test  
b. Grouping Variable: Group

Chi-square (2) = 7.83,  
p = .020

**Figure 5.32.**  
Kruskal-Wallis test result.

## 6 Factorial Anovas

### 6.1 Introduction

A 'factorial Anova' is an Anova with more than one categorical variable (but still one continuous variable). For example, a 'two way Anova' means that there are two categorical variables: e.g. the effect of gender and alcohol on performance. In Anova, an alternative name for the categorical variables is *factors*. (Do not get confused by the fact that this word also has other meanings.)

Each of the categorical variables could either be repeated-measures or independent-samples. For example, in a two way Anova, any of the following combinations can occur. Each requires a different procedure in SPSS.

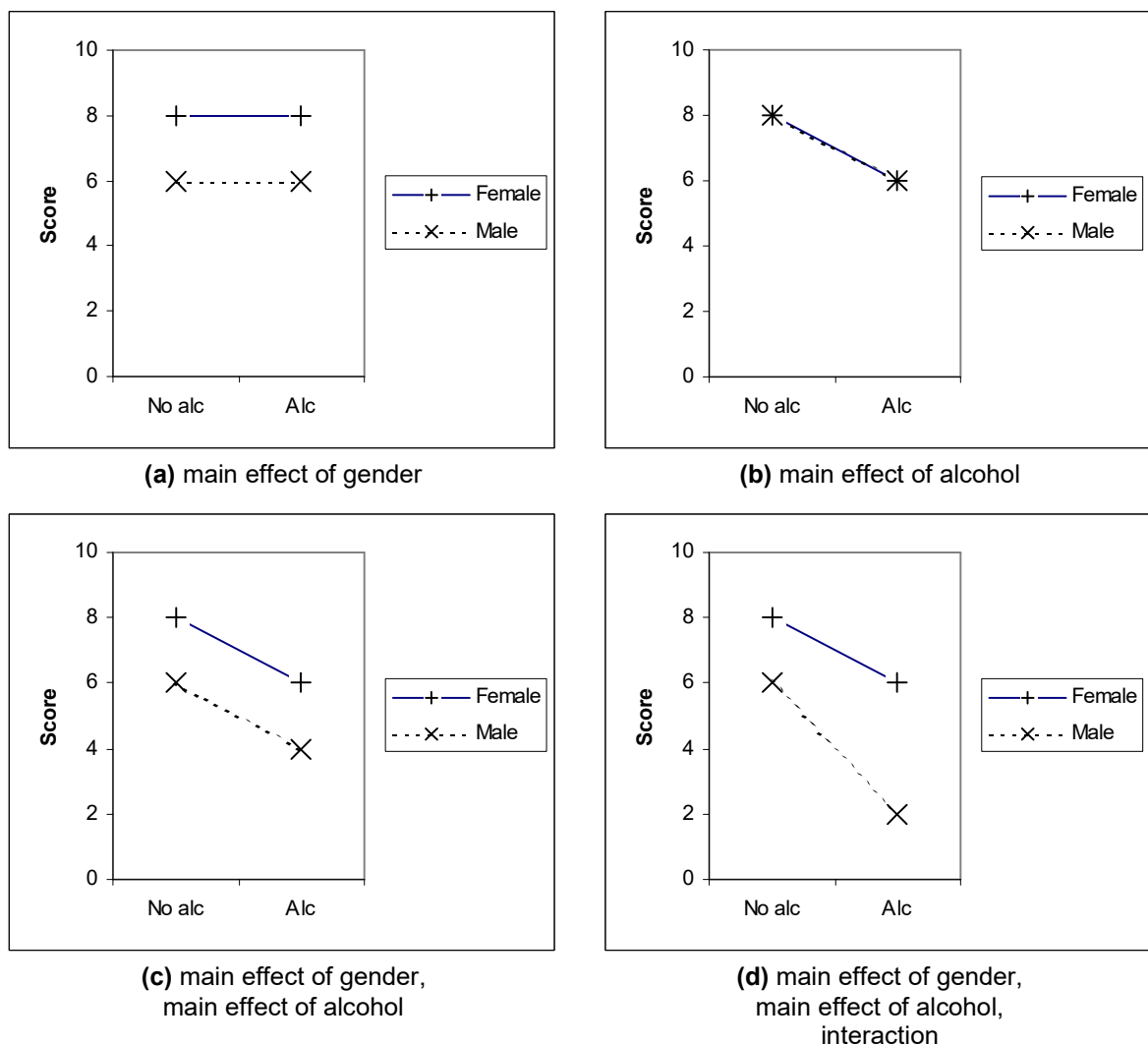
- (a) both categorical variables are independent-samples: requires a *two way independent-samples Anova* (section 6.5)
- (b) both categorical variables are repeated-measures: requires a *two way repeated-measures Anova* (section 6.6)

- (d) one categorical variable is independent-samples and the other is repeated-measures: requires a *two way mixed Anova* (section 6.7).

Each of the categorical variables can have two levels (categories), or more. In the following examples each will have two, but the principles are the same if they have more.

Almost invariably, a factorial Anova is used when the categorical variables are both considered to be Independent Variables (IVs) and the continuous variable is considered to be a Dependent Variable (DV). In fact, the results would be quite hard to interpret if this were not the case, and the explanations in this section presume this.

## 6.2 Outcomes



**Figure 6.1.** Some possible outcomes of a two way Anova.

Figure 6.1 shows some possible outcomes, illustrated using the kind of graph we can call up in SPSS.

The Anova calculates the effect of each categorical variable (the IVs) on the continuous variable (the DV). In the example illustrated, it examines the effect of alcohol on performance, and the effect of gender on performance. These are called *main effects*.

However, the main point of a two way Anova is that it enables us to see whether *the effect of one IV is different depending on the level of the other IV*. In our example, we might ask whether the effect of alcohol on performance is different for men and women? Such a difference is called an *interaction*. It is often said that the defining feature of an interaction on a graph is that the lines are not parallel. However, the lines will always be non-parallel, even if only because of sampling error. To be more precise, a significant interaction means that the lines differ significantly from being parallel.

The main effects and interactions are generically known as *effects*.

### 6.3 If the factorial Anova shows significant effects

If there is a significant interaction in a factorial Anova, you will probably want to break the results down further. Notice that the graphs (which we will call up in every instance) are a good way of visualising what we are likely to be interested in.

If there are (as here) only two levels (categories) of each Independent Variable, the logic would be as follows. You might want to ask

- “For the men, was there a significant difference in score between the alcohol and no-alcohol conditions?” (This known as the *simple main effect* of alcohol for men.)
- And “For the women, was there a significant difference in score between the alcohol and no-alcohol conditions?” (The simple main effect of alcohol for women.)

(In other words, is there an interaction because there was a difference only for one gender; or because there is a difference for both genders, but a bigger difference for one of the genders?)

Instead – or as well – you might ask “For the alcohol condition, was there a significant difference in score between the men and the women?” And “For the no-alcohol condition, was there a significant difference in score between the men and the women?”

You can examine these questions using exactly the same test(s) you would use if those were the only data in your file. For example, the question “For the men, was there a significant difference in score between alcohol and no alcohol?”

would require a t-test between the alcohol condition and the no-alcohol condition, just including the men in the analysis. Whether this is a paired-samples or an independent-samples t-test would, as always, depend on whether the same men or different men did the test in the two alcohol conditions. See paragraph 5.2, and detailed advice after each test in this section.

Remember to use a Bonferroni correction, since you are carrying out multiple comparisons.

If one or both of the Independent Variables has more than two levels, then to do all logically possible post hoc tests would be more complicated. Suppose the alcohol condition had three levels. Then to examine the simple main effect of alcohol for men, you could do an Anova of the effect of alcohol just for the men. If this was significant, you could follow up with three t-tests. You would do a Bonferroni correction at each stage; for 2 comparisons at the first stage and for 6 ( $2 \times 3$ ) at the second stage. To reduce the Bonferroni correction, it is preferable to carry out planned comparisons, rather than to follow this pedantic post hoc approach.

#### 6.4 Effect sizes

As with the Anovas in Chapter 5, you can ask for a measure of effect size. Under **Options**, select **Estimates of effect size**. As in Chapter 5, you will get a new column headed 'Partial eta squared'. For a factorial Anova, partial eta squared is roughly equivalent to the square of the partial correlation coefficient. The partial correlation coefficient, and its square, will be explained in the lecture on multiple regression.

#### 6.5 Two way independent-samples Anova

(also known as a two way between-subjects Anova)

Suppose we study the effect of sleep and alcohol on some kind of test. If we study participants with and without sleep, with and without alcohol, that makes four possible combinations:

- (a) without alcohol after normal sleep
- (b) without alcohol having missed a night's sleep
- (c) with alcohol after normal sleep
- (d) with alcohol having missed a night's sleep.

Suppose everybody provides data in only one of those combinations. That makes our design entirely between-subjects. The procedure is an extension of the one we used in section 5.11.

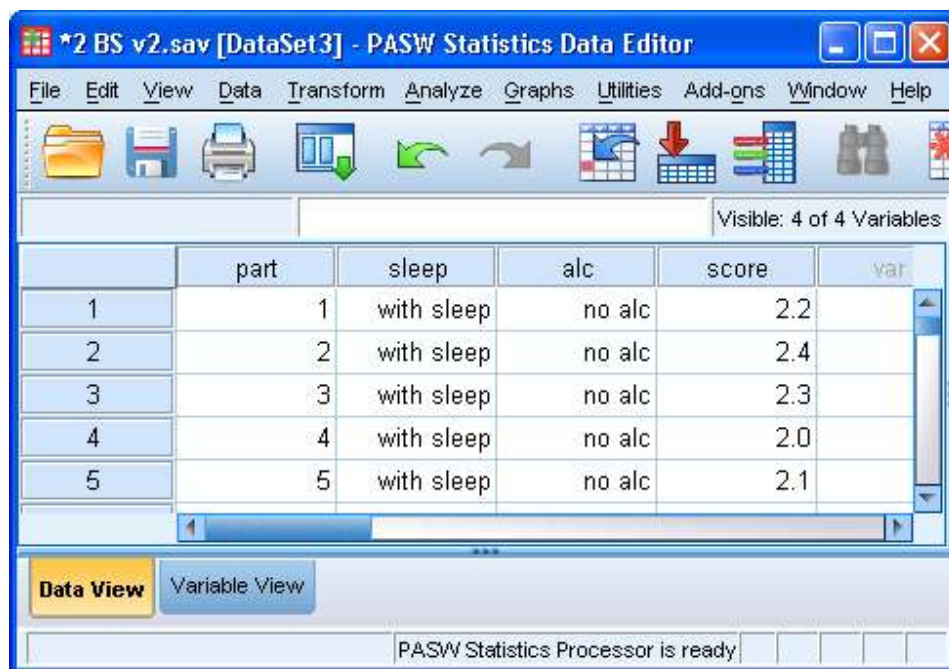
Our example data are in Table 6.1. Remembering to enter what we know about one person on one line, the data file needs to look like Figure 6.2. It will be helpful to set up the Variable View first.



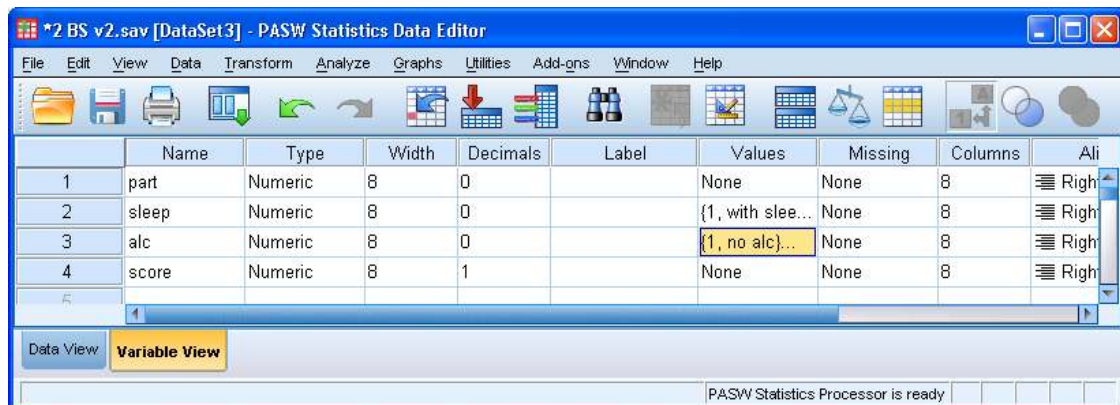
**Table 6.1.** Two way independent-samples example data.

Part	Sleep	Alc	Score	Part	Sleep	Alc	Score
1	with	no alc	2.2	17	without	no alc	1.8
2	with	no alc	2.4	18	without	no alc	1.9
3	with	no alc	2.3	19	without	no alc	1.4
4	with	no alc	2.0	20	without	no alc	1.5
5	with	no alc	2.1	21	without	no alc	1.5
6	with	no alc	1.7	22	without	no alc	1.8
7	with	no alc	2.0	23	without	no alc	1.2
8	with	no alc	2.8	24	without	no alc	1.4
9	with	with alc	1.8	25	without	with alc	0.5
10	with	with alc	1.8	26	without	with alc	0.5
11	with	with alc	1.5	27	without	with alc	0.1
12	with	with alc	1.4	28	without	with alc	0.9
13	with	with alc	2.1	29	without	with alc	0.9
14	with	with alc	1.8	30	without	with alc	0.7
15	with	with alc	2.3	31	without	with alc	0.4
16	with	with alc	1.6	32	without	with alc	0.3

Our Variable View is set up as shown in Figure 6.3. Remember that we use numbers to represent the between-subjects groups. We tell the computer what each number means, by using the **Values** cells. (Click on the cell and then on the three dots which appear. For more detail, refer back to section 5.8.1). For *Sleep*, we will set 1 = *with sleep* and 2 = *without sleep*. For *Alc*, we will set 1 = *no alc* and 2 = *with alc*. Notice also that there is one decimal place for the scores. There are no decimal places in the group numbers.



**Figure 6.2.** Data layout for two way independent-samples Anova.

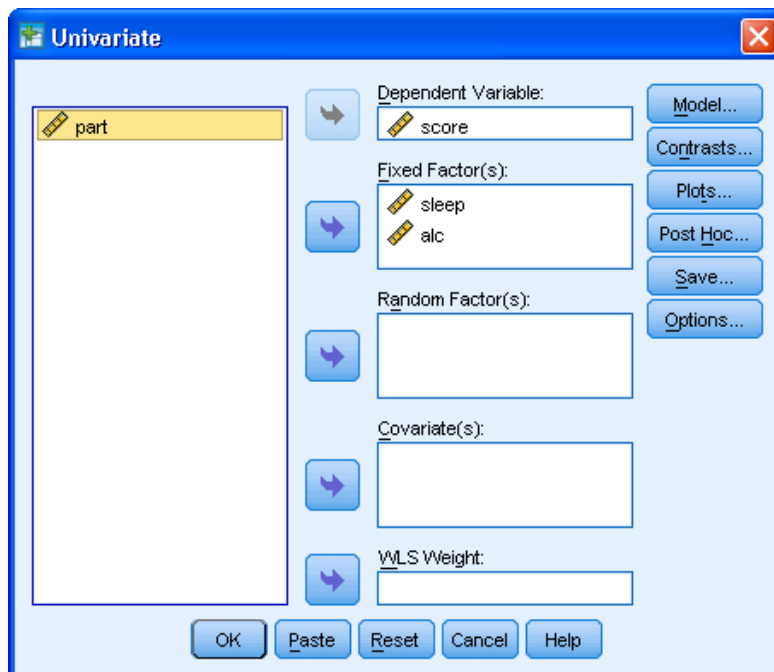


**Figure 6.3.** Variable view for independent-samples Anova.

Entering the groups is easiest using numbers, with value labels turned off. Remember you can do this using **View – Data Labels**, or the labels icon



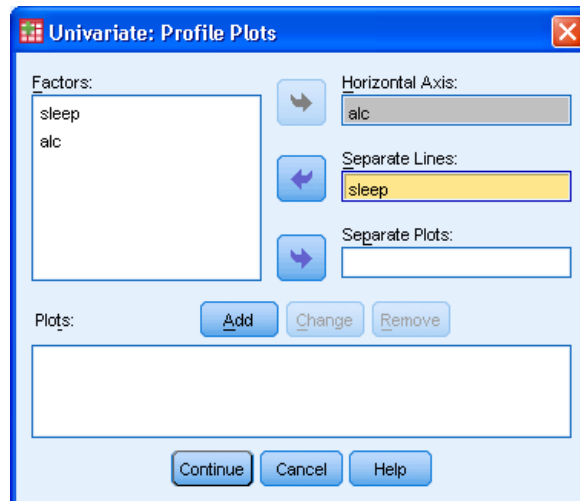
Once the data are entered, call up the test by going to **Analyze – General Linear Model – Univariate**. When the dialogue box appears, move the categorical variables into the box marked “Fixed Factors” and the continuous variable into the box marked “Dependent Variable”, as shown in Figure 6.4<sup>32</sup>.



<sup>32</sup> Remember that here, “factor” is another name for an independent variable (IV). As mentioned in 6.1 above, in factorial Anova it is usually presumed that the categorical variables are IVs.

**Figure 6.4.** Dialogue box for two way independent-samples Anova.

Click on **Options**, and tick the boxes marked **Descriptive statistics** and **Homogeneity tests**. Click **Continue**. Click on **Plots** and a new dialogue box appears (Figure 6.5). Click the factors into the **Horizontal Axis** and **Separate Lines** boxes<sup>33</sup>, and click on **Add**. Click **Continue**, then back in the main dialogue box click **OK**. Examine the output.



**Figure 6.5.** Plots dialogue box.

The descriptive statistics are in Figure 6.6. Notice they include an *N* column, which is a useful check that we have entered the correct number of cases for each combination of variables.

#### Descriptive Statistics

Dependent Variable: score

sleep	alc	Mean	Std. Deviation	N
with sleep	no alc	2.187	.3271	8
	with alc	1.788	.2997	8
	Total	1.987	.3667	16
without sleep	no alc	1.563	.2446	8
	with alc	.538	.2825	8
	Total	1.050	.5877	16
Total	no alc	1.875	.4266	16
	with alc	1.162	.7042	16
	Total	1.519	.6775	32

<sup>33</sup> If you are not sure which variable you want in which box, it is easiest to do it both ways round. Then look at the output and choose the most useful chart.

**Figure 6.6.** Descriptive statistics.

As with the one-way Anova, we should check that Levene's test is not significant<sup>34</sup>. In this case, it is not (Figure 6.7).

**Levene's Test of Equality of Error Variances<sup>a</sup>**

Dependent Variable: score

F	df1	df2	Sig.
.057	3	28	.982

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + sleep + alc + sleep \* alc

**Figure 6.7.** Levene's test result.

The Anova results are in Figure 6.8.

**Tests of Between-Subjects Effects**

Dependent Variable: score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	11.874 <sup>a</sup>	3	3.958	47.058	.000
Intercept	73.811	1	73.811	877.586	.000
sleep	7.031	1	7.031	83.599	.000
alc	4.061	1	4.061	48.287	.000
sleep * alc	.781	1	.781	9.289	.005
Error	2.355	28	.084		
Total	88.040	32			
Corrected Total	14.229	31			

a. R Squared = .834 (Adjusted R Squared = .817)

A significant effect of sleep,  $F(1,28) = 83.60$ ,  $p < .001$ .

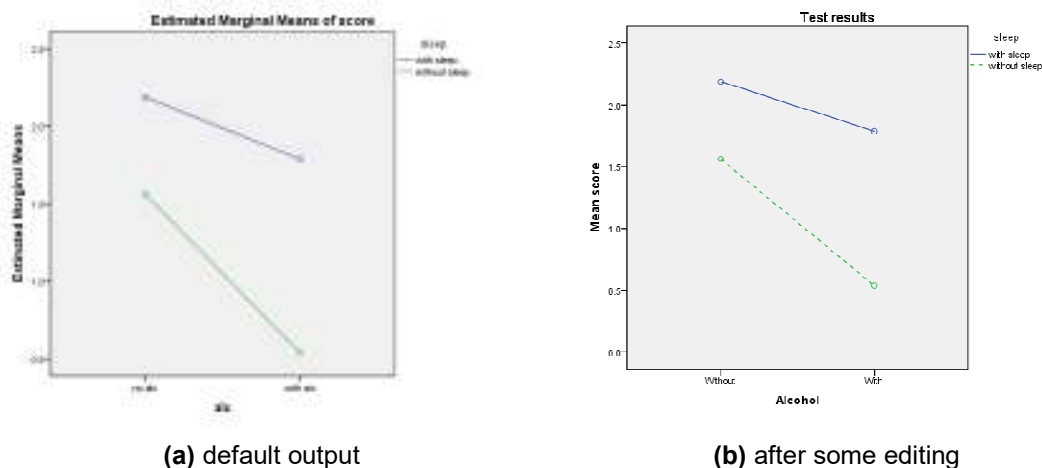
**Figure 6.8.** Anova results and partial interpretation.

Note that the information for each effect comes from its own line, except for the second figure in brackets (the error degrees of freedom): this comes from the same line (Error) for all effects. The line showing the interaction is always indicated by the two variable names with an asterisk between them.

<sup>34</sup> If it is significant, see Appendix E for some possible options.

Thus we can report: There was a significant effect of sleep on the score,  $F(1,28) = 83.60$ ,  $p < .001$ , a significant effect of alcohol,  $F(1,28) = 48.3$ ,  $p < .001$ , and a significant interaction between sleep and alcohol,  $F(1,28) = 9.3$ ,  $p = .005$ .

As always, your reader needs to know what these effects mean – did people do better with or without sleep, for example? Report the means and standard deviations. The graph may also help interpretation, but it will probably need to be edited (Figure 6.9(b) shows some of the changes that can be made) or re-created in Excel.



**Figure 6.9.** Graph of two factor independent-samples Anova.

### 6.5.1 Following up a significant interaction

If you have a significant interaction, you will probably want to follow up with post hoc tests, or planned comparisons (see paragraph 6.3). For a two way independent-samples Anova, this can be done by selecting cases, or splitting the file.

For example, looking at Figure 6.9(a), one question you could ask is “For the participants who had sleep, was there a significant difference between those who had no alcohol and those who had alcohol?” You could select just the participants who had had sleep (section 13.6.1) and do an independent-samples t-test (section 5.9) between *no alcohol* and *alcohol*.

But probably, you would want to ask the same question about participants who did not have sleep. You could kill two birds with one stone, as follows. Split the file by *sleep* (section 13.6.3) and call up an independent-samples t-test (section 5.9) between *no alcohol* and *alcohol*. SPSS will give you that result for participants who did have sleep (as in the previous paragraph), and also the result for participants who did not have sleep.

You might as well, or instead, want to know “For the participants who had alcohol, was there a significant difference between those who had no sleep and

those who had sleep?” (and the same question for participants who did not have alcohol). The procedure is the same as above, but swapping round the variables (i.e. split the file by *alcohol*, and call up an independent-samples t-test between *sleep* and no *sleep*).

## 6.6 Two way repeated measures Anova

(also known as a two-way within-subjects Anova).

Again suppose that we examine the effect of alcohol and sleep on a test. But now, every participant does the test in all four conditions

- (a) without alcohol after normal sleep
- (b) without alcohol having missed a night's sleep
- (c) with alcohol after normal sleep
- (d) with alcohol having missed a night's sleep.

So that makes the design entirely repeated-measures. (Note that if this an experimental design we would have to counterbalance both IVs; i.e. all four conditions. And again, as mentioned above, we would generally presume that the categorical variables here are IVs.)

Suppose that we test eight participants. Their scores are as shown in Table 6.2. Entering the data still follows our rule of thumb: what we know about one person goes on one line. Enter the data so that Data View looks like Figure 6.10 and Variable View looks like Figure 6.11. Notice it is a good idea to use names which are systematic and logical, so you know exactly what each combination means. Even so, you may want to put fuller names under **Labels**. (You can make more room for the Labels simply by pulling at the heading, using the mouse.)

**Table 6.2.** Data for two-way repeated-measures Anova.

Part	Score on the test:			
	No alcohol, with sleep	No alcohol, no sleep	With alcohol, with sleep	With alcohol, no sleep
1	17	18	14	10
2	17	11	24	4
3	20	18	23	14
4	28	21	18	0
5	20	17	16	12
6	15	18	18	16
7	21	16	17	16
8	21	20	17	12

1 : Part 1 Visible: 5 of 5 Variables

	Part	na_wsleep	na_nsleep	wa_wsleep	wa_nsleep	var
1	1	17	18	14	10	
2	2	17	11	24	4	
3	3	20	18	23	14	
4	4	28	21	18	0	
5	5	20	17	16	12	
6	6	15	18	18	16	
7	7	21	16	17	16	
8	8	21	20	17	12	

Data View Variable View

PASW Statistics Processor is ready

Figure 6.10. Data View for repeated measures Anova.

	Name	Type	Width	Decimals	Label	Values
1	Part	Numeric	8	0		None
2	na_wsleep	Numeric	8	0	No alc, with sleep	None
3	na_nsleep	Numeric	8	0	No alc, no sleep	None
4	wa_wsleep	Numeric	8	0	With alc, with sleep	None
5	wa_nsleep	Numeric	8	0	With alc, no sleep	None
6						

Data View Variable View

PASW Statistics Processor is ready

Figure 6.11. Variable view for Repeated Measures factorial Anova.

The analysis in SPSS is an extension of the one way repeated-measures Anova (section 5.6), but with some important differences.

Click on **Analyze – General Linear Model – Repeated Measures**. The **Define Factors** dialogue box appears (Figure 6.12). Replace the default name (*factor1*)



by the name of one of our factors<sup>35</sup>: it will make life easier if we start with the one which changes most slowly across our data; this is *alcohol* (since both of our first two columns of data are with no alcohol). The number of levels (i.e. categories) of this IV is 2; enter this. The dialogue box should now look like Figure 6.12(a). Click on **Add**, then repeat the process for the second factor (*sleep*). The dialogue box should now look like Figure 6.12(b).

Now click on **Define** and a dialogue box (similar to Figure 6.13) appears. Notice that underneath **Within-Subjects variables**, the two variables are named (*alc*, *sleep*). Carefully click names from the left box to the right box. 'Carefully' means that you need to ensure that the numbers are used consistently across all the variables. In our example, the first number, as shown at the top, represents *alcohol*. We use 1 to represent no alcohol (*na*) and 2 to represent with alcohol (*wa*). Similarly, the second number represents *sleep*. 1 represents with sleep (*wsleep*) and 2 represents no sleep (*nsleep*). (In fact, in this example we made sure that our first factor was the one which changed more slowly, so they were already in the correct order on the left hand side.) The dialogue box should now look exactly like Figure 6.13.

The dialog box is titled "Repeated Measures Defin...". It has two main sections. The top section is for the "Within-Subject Factor Name:" and "Number of Levels:". The "Within-Subject Factor Name:" field contains the text "alcohol". The "Number of Levels:" field contains the number "2". Below these fields are three buttons: "Add", "Change", and "Remove". The bottom section is for the "Measure Name:" and has three buttons: "Add", "Change", and "Remove". At the very bottom are four buttons: "Define", "Reset", "Cancel", and "Help".

(a) defining first factor

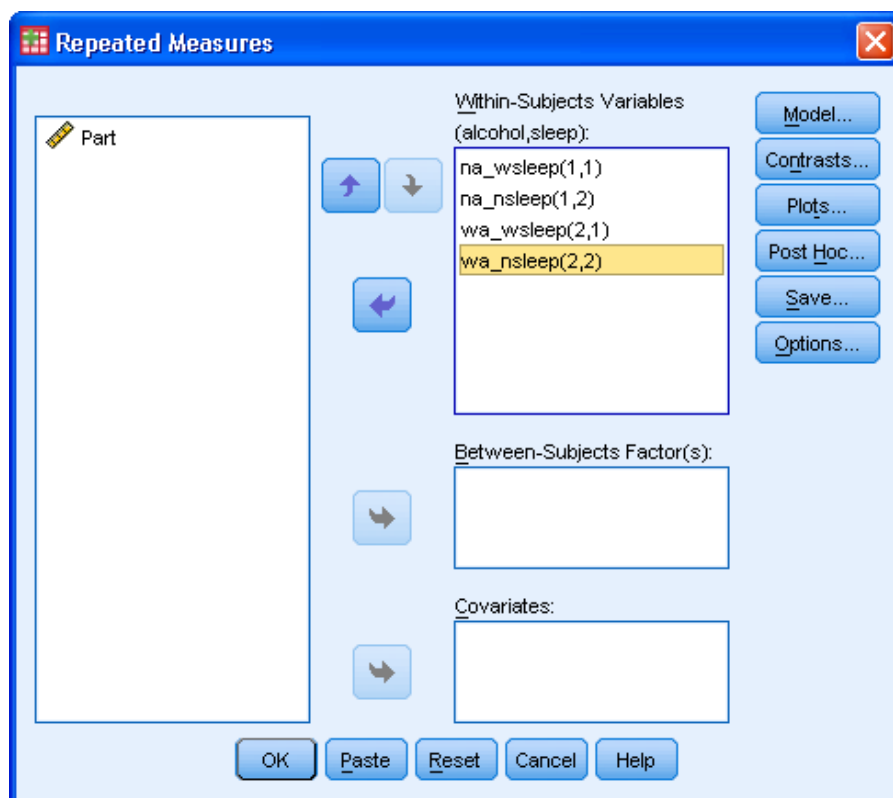
The dialog box is titled "Repeated Measures Defin...". It has two main sections. The top section is for the "Within-Subject Factor Name:" and "Number of Levels:". The "Within-Subject Factor Name:" field is empty. The "Number of Levels:" field is empty. Below these fields are three buttons: "Add", "Change", and "Remove". The bottom section is for the "Measure Name:" and has three buttons: "Add", "Change", and "Remove". At the very bottom are four buttons: "Define", "Reset", "Cancel", and "Help".

(b) on completion

**Figure 6.12.** Repeated Measures Define Factor(s) dialogue box at two stages.

<sup>35</sup> As mentioned in 6.1 above, 'factor' is used as another name here for a categorical independent variable.





**Figure 6.13.** Repeated Measures dialogue box with two factors.

Click on **Options** and check the box marked **Descriptive statistics**. Click **Continue**. Click on **Plots** and a new dialogue box appears, similar to Figure 6.5. Click the factors into the **Horizontal axis** and **Separate lines** boxes<sup>36</sup>, and click on **Add**. Click **Continue** and **OK**.

The first output of interest to us is the descriptive statistics (Figure 6.14). This shows the mean and standard deviation of the score in each condition, which we will need to report.

Descriptive Statistics			
	Mean	Std. Deviation	N
No alc, with sleep	19.88	3.944	8
No alc, no sleep	17.38	3.021	8
With alc, with sleep	18.38	3.420	8
With alc, no sleep	10.50	5.732	8

**Figure 6.14.** Descriptive statistics for two way repeated measures Anova.

<sup>36</sup> If you are not sure which variable you want in which box, it is easiest to do it both ways round. Then look at the output and choose the most useful chart.

Examine Mauchley's test of sphericity (Figure 6.15), although in this case the significance levels are blank because none of our factors has more than two levels. If one or more of the Mauchley's tests is significant, I recommend that you use the Greenhouse-Geisser correction (see section 5.6) for all of the effects in that test<sup>37</sup>.

Mauchly's Test of Sphericity <sup>b</sup>							
Measure: MEASURE_1							
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon <sup>a</sup>		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
alc	1.000	.000	0	.	1.000	1.000	1.000
sleep	1.000	.000	0	.	1.000	1.000	1.000
alc * sleep	1.000	.000	0	.	1.000	1.000	1.000

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.  
Design: Intercept  
Within Subjects Design: alc+sleep+alc\*sleep

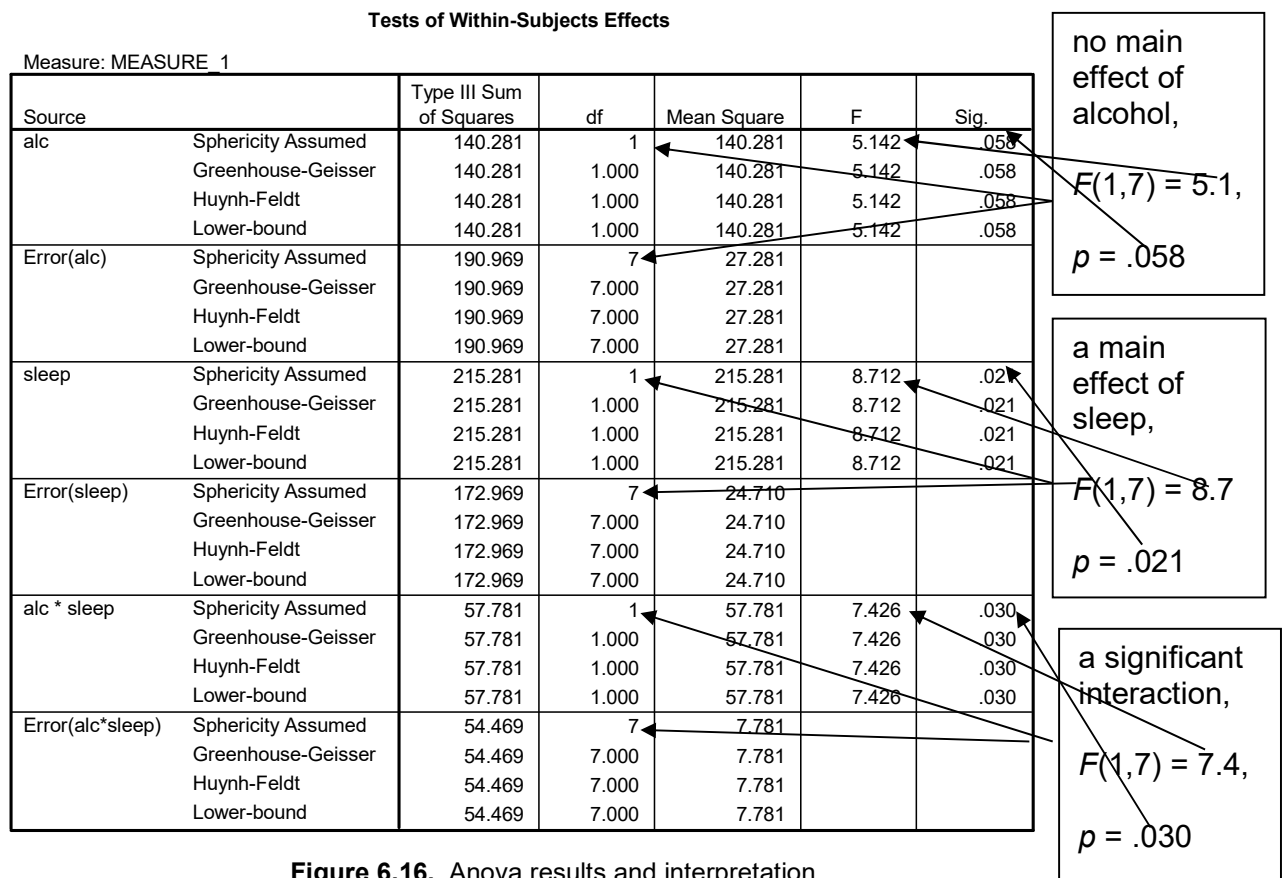
**Figure 6.15.** Mauchley's test of sphericity.

The Anova results are shown in Figure 6.16, which may seem daunting until you remember that you only have to read the lines marked 'Sphericity Assumed' (or 'Greenhouse-Geisser', as appropriate). In this case we may write that there was no main effect<sup>38</sup> of alcohol on the score,  $F(1,7) = 5.14$ ,  $p = .058$ ; there was a main effect of sleep,  $F(1,7) = 8.71$ ,  $p = .021$ ; there was a significant interaction between alcohol and sleep,  $F(1,7) = 7.43$ ,  $p = .030$ .

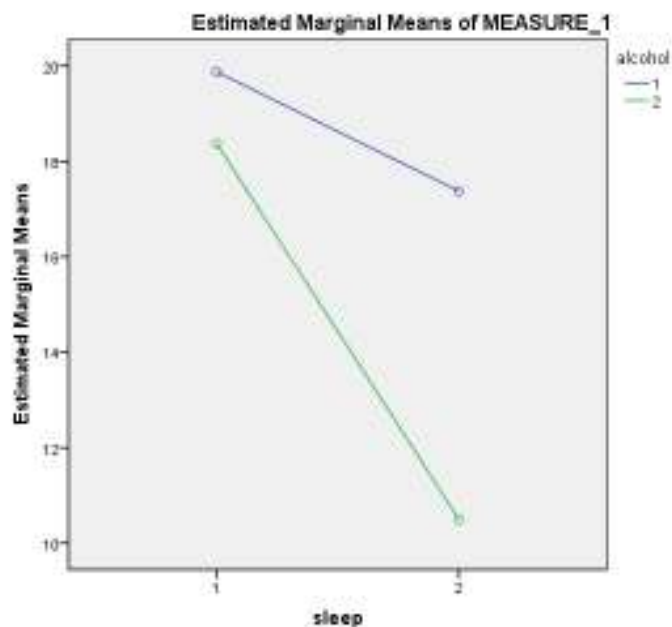
Remember to report the means and standard deviations (from Figure 6.14; you may find a table is the easiest way to do this). Finally, the graph (Figure 6.17) will help interpretation, although of course it will need editing or re-creating in Excel if you are to use it in your published work.

<sup>37</sup> Or you could try a transformation, see section 14.3.3; or more advanced texts cover other possibilities as well.

<sup>38</sup> Or we might report this as a trend; see Appendix A.



**Figure 6.16.** Anova results and interpretation.



**Figure 6.17.** Graph for two way repeated measures Anova.

### 6.6.1 Following up a significant interaction

If you have a significant interaction, you will probably want to follow up with post hoc tests, or planned comparisons (see paragraph 6.3). For a two way repeated-measures Anova, this is quite simple. Remember to look at the graph to see what comparisons you are interested in, then to do the appropriate tests. For example, after looking at Figure 6.17 you might like to know:

- For participants who have had alcohol, is there a significant difference in their score depending on whether they had sleep or not? Carry out a paired-samples t-test with the paired variables as *With alc, with sleep* and *With alc, no sleep*.
- For participants who have had no alcohol, is there a significant difference in their score depending on whether they had sleep or not? Carry out a paired-samples t-test with the paired variables as *No alc, with sleep* and *No alc, no sleep*.

If you need a reminder on how to carry out and report a paired-samples t-test, see paragraph 5.4. Remember to make an appropriate Bonferroni correction.

## 6.7 Two way mixed Anova

With two independent variables, it is possible that one might be repeated-measures and one might be independent-samples. For example, suppose that we carried out a study where

- one of the IVs was *gender* (which must be independent-samples: everyone can only provide data in one condition)
- and the other was *alcohol*, which we decided to make within-subjects, i.e. everyone contributed data with and without alcohol.

Our results might be as in Figure 6.18. Entering the data might seem hard at first, but just remember – use one line for everything you know about each participant. The Variable View for these results is shown in Figure 6.19. There is a between-subjects variable, so as usual we need to define it under **Values**; I have used 1 = *male*, 2 = *female*.

\*2 mixed.sav [DataSet1] - PASW Statistics Data Editor

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

Visible: 4 of 4 Variables

	Part	Gender	no_alc	with_alc
1	1	male	9	3
2	2	male	9	3
3	3	male	4	4
4	4	male	4	6
5	5	male	4	6
6	6	male	12	8
7	7	male	7	1
8	8	male	6	9
9	9	female	9	0
10	10	female	13	4
11	11	female	12	2
12	12	female	8	2
13	13	female	15	0
14	14	female	9	7
15	15	female	12	0
16	16	female	16	2
17				

Data View Variable View

PASW Statistics Processor is ready

Figure 6.18. Data for two way mixed Anova, entered into Data View.

\*2 mixed.sav [DataSet1] - PASW Statistics Data Editor

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

	Name	Type	Width	Decimals	Label	Values	Missing	C
1	Part	Numeric	8	0		None	None	8
2	Gender	Numeric	8	0		{1, male}...	None	8
3	no_alc	Numeric	8	0		None	None	8
4	with_alc	Numeric	8	0		None	None	8
5								

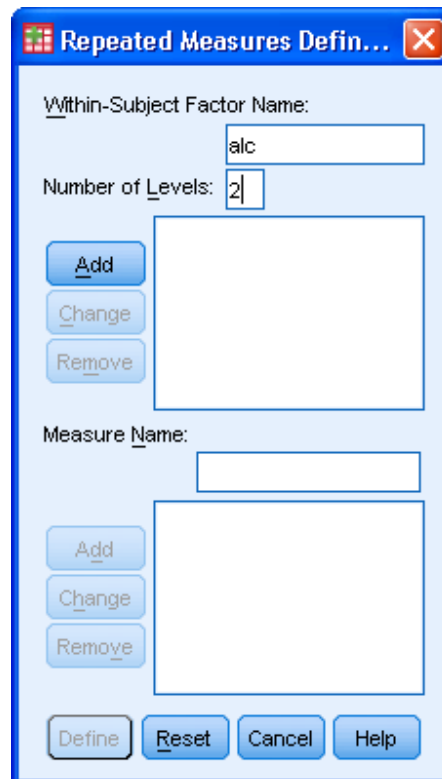
Data View Variable View

PASW Statistics Processor is ready

Figure 6.19. Variable View for mixed Anova.

When you have entered the data, click on **Analyze – General Linear Model – Repeated Measures**. In the first dialogue box, name our repeated measures IV

(i.e. within-subjects factor) and say how many levels there are (Figure 6.20). Click **Add** and **Define**.



**Figure 6.20.** Repeated measures dialogue box.

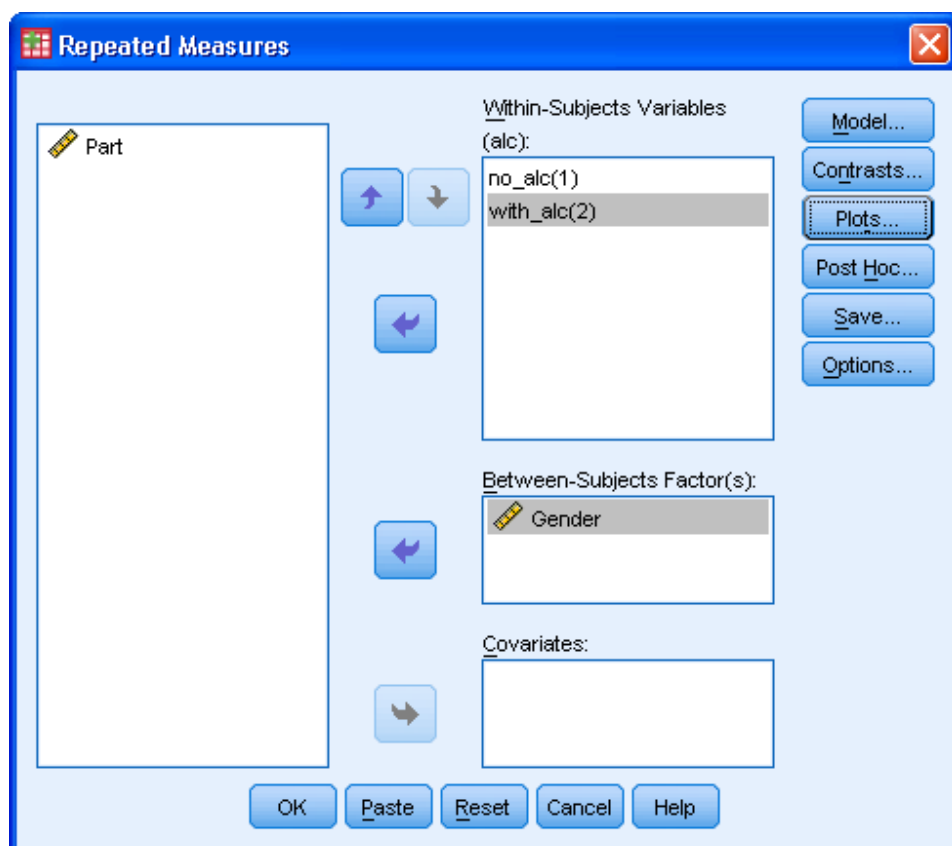
In the next dialogue box, put the two levels (categories) of the repeated-measures (within-subjects) variable into the **Within-Subjects Variables** box. Put the independent-samples IV (between-subjects factor) into the **Between-subjects factors** box. The dialogue box should then look like Figure 6.21.

As usual, under **Options**, request **Descriptive statistics**. Also under **Options**, ask for **homogeneity tests**, since we have a between-subjects factor. Under **Plots**, ask for a graph<sup>39</sup>. Our output is a cross between items we are used to. The first item of interest is the descriptive statistics (Figure 6.22).

We would normally need to check Mauchley's test (Figure 6.23), but in this case a dot is shown under 'Sig' – this means it is redundant, because we only have two levels of our within-subjects variable. As always, if it were significant we would need to use the Greenhouse-Geisser correction (see section 5.6), and I would recommend you do so for all within-subjects effects even if only one is significant.

---

<sup>39</sup> Also as usual, if you do not know which variable to put under Horizontal Axis and which under Separate Lines, you can try both and see which you like. Don't forget to click on Add after each combination.



**Figure 6.21.** Repeated measures dialogue box for two-way mixed Anova.

#### Descriptive Statistics

	Gender	Mean	Std. Deviation	N
no_alc	male	6.88	2.949	8
	female	11.75	2.915	8
	Total	9.31	3.790	16
with_alc	male	5.00	2.726	8
	female	2.13	2.416	8
	Total	3.56	2.898	16

**Figure 6.22.** Descriptive statistics for mixed Anova.

Mauchly's Test of Sphericity <sup>b</sup>							
Measure: MEASURE_1							
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon <sup>a</sup>		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
alc	1.000	.000	0	.	1.000	1.000	1.000

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.

Design: Intercept+Gender  
Within Subjects Design: alc

**Figure 6.23.** Mauchley's test for two-way mixed Anova example.

The within-subjects Anova result (and the interaction) is shown under 'Tests of within-subjects effects' (Figure 6.24).

Tests of Within-Subjects Effects						
Measure: MEASURE_1						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
alc	Sphericity Assumed	264.500	1	264.500	31.020	.000
	Greenhouse-Geisser	264.500	1.000	264.500	31.020	.000
	Huynh-Feldt	264.500	1.000	264.500	31.020	.000
	Lower-bound	264.500	1.000	264.500	31.020	.000
alc * Gender	Sphericity Assumed	120.125	1	120.125	14.088	.002
	Greenhouse-Geisser	120.125	1.000	120.125	14.088	.002
	Huynh-Feldt	120.125	1.000	120.125	14.088	.002
	Lower-bound	120.125	1.000	120.125	14.088	.002
Error(alc)	Sphericity Assumed	119.375	14	8.527		
	Greenhouse-Geisser	119.375	14.000	8.527		
	Huynh-Feldt	119.375	14.000	8.527		
	Lower-bound	119.375	14.000	8.527		

Significant effect of alcohol,  $F(1,14) = 31.02$ ,  $p < .001$

Significant interaction,  $F(1,14) = 14.09$ ,  $p = .002$

**Figure 6.24.** Tests of within-subjects effect and interpretation.

Before looking at the between-subjects results we check that Levene's test is not significant<sup>40</sup> (Figure 6.25; note that there is more than one to check).

<sup>40</sup> If either result is significant, see Appendix E for some options.



Levene's Test of Equality of Error Variances <sup>a</sup>				
	F	df1	df2	Sig.
no_alc	.007	1	14	.936
with_alc	.599	1	14	.452

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a.

Design: Intercept+Gender  
Within Subjects Design: alc

There are as many Levene's tests as there are within-subjects conditions. All should be non-significant.

**Figure 6.25.** Levene's test results.

The between-subjects Anova result is shown in Figure 6.26.

Tests of Between-Subjects Effects					
Measure: MEASURE_1					
Transformed Variable: Average					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1326.125	1	1326.125	197.771	.000
Gender	8.000	1	8.000	1.193	.293
Error	93.875	14	6.705		

No significant effect of gender,  $F(1,14) = 1.19$ ,  $p = .293$ .

**Figure 6.26.** Between-subjects Anova result.

Hence, there was a significant effect of alcohol on the score,  $F(1,14) = 31.02$ ,  $p < .001$ , no significant effect of gender,  $F(1,14) = 1.19$ ,  $p = .293$ , and a significant interaction,  $F(1,14) = 14.09$ ,  $p = .002$ . As always, report the means and standard deviations and consider using a graph (e.g. as Figure 6.27), especially if you have a significant interaction.

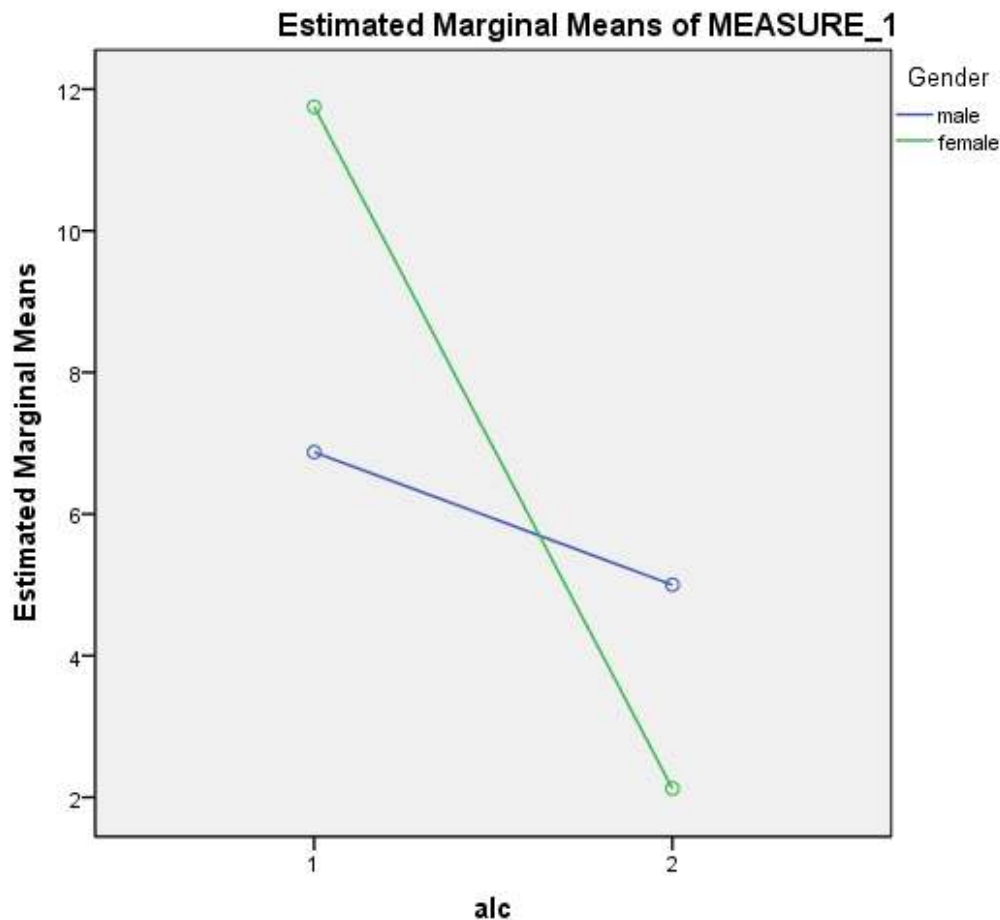


Figure 6.27. Graph for mixed Factorial Anova.

### 6.7.1 Following up a significant interaction

If you have a significant interaction, you will probably want to follow up with post hoc tests, or planned comparisons (see paragraph 6.3). With a mixed Anova, you need to think carefully about which factor(s) is (are) repeated-measures, and which is (are) independent-samples. As usual, the graph will help you.

Firstly, you might want to consider the two genders:

- For men, is there a significant difference between men who have alcohol and men who do not?
- For women, is there a significant difference between women who have had alcohol and women who have not?

In this case, *Gender* is a between-subjects variable. So you split the file by gender (paragraph 13.6.3). *Alcohol* is within-subjects variable. So you call up a

paired-samples t-test (paragraph 5.4) with *no\_alc* and *with\_alc* as the paired variables. **Remember to unsplit the file afterwards!**

Secondly, you might want to consider the two alcohol conditions:

- Considering people who had alcohol, is there a difference between men and women?
- Considering people who had not had alcohol, is there a difference between men and women?

The scores for people who have had alcohol are given in one column of SPSS, and gender is a between-subjects variable. So we can answer the first question with an independent-samples t-test (paragraph 5.9) with *with\_alc* as the **Test Variable**, and *Gender* as the **Grouping Variable**. Similarly, we can answer the second question with an independent-samples t-test with *no\_alc* as the **Test Variable**, and *Gender* as the **Grouping Variable**.

## 6.8 Anovas with more than two factors

Anovas with more than two factors can be analysed in SPSS using the same procedures as above. The statistical results are also interpreted in the same manner. However, the interpretation in words is more difficult. For example, a three way interaction between gender, sleep and alcohol would mean something like “the two way interaction between sleep and alcohol is significantly different for men and women” – or equivalently, one could swap the IVs round in any order!

## 7 Chi-square tests of association

### 7.1 Introduction; when they are used

Chi is a way of writing the Greek letter  $\chi$ , usually pronounced ‘kye’. To see how to write chi-square more neatly ( $\chi^2$ , although this works better in other fonts) see Appendix A.

A chi-square test is used when both of the variables are categorical. (It does not matter whether we have an independent and dependent variable.) Both variables should be independent-samples; if one of them is repeated-measures see chapter 9.

### 7.2 The possible outcomes of a chi-square test

If there is an *association* between the variables (the experimental or research hypothesis), the frequency of one variable will be different depending on the value of the other variable – for example, the people who took the drug were significantly less likely (or more likely!) still to have tumours.

If the variables are *independent* (the null hypothesis), there is no relationship between them – for example the people who took the drug were just as likely to still have tumours.

### 7.3 Example 1: entering individual cases into SPSS

A researcher thinks that employees in company A are more likely to be happy in their work than those in company B. He asks some sample workers “Are you happy in your work?: yes/no”. The responses are shown in Table 7.1.

Enter the data into SPSS. Remember to use a separate line for each case (i.e. 24 lines).

Notice that since this is a chi-square test, both variables are categorical and we will need to use the ‘values’ field to define them. Go to Variable View, and:

- Name the first variable *part* (short for participant).
- Name the second variable *firm*. Use the **Values** field to show that 1 means A and 2 means B.<sup>41</sup>
- Name the third variable *happy*. Use the **Values** field to show that 1 means yes and ‘2’ means no.

In Data View, enter figures as appropriate. Make sure that **Value Labels** (from the drop-down **View** menu) is ticked so you can check your entries.

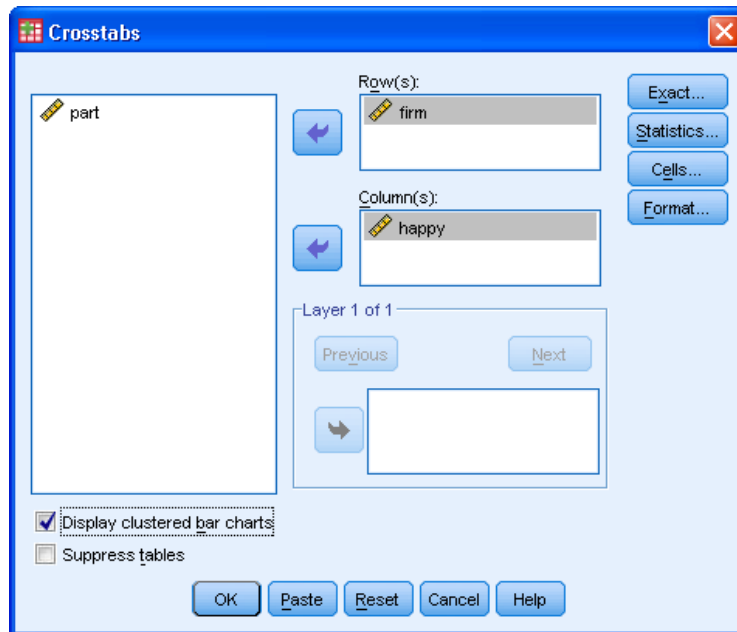
**Table 7.1.** Data for chi-square test example 1.

Part	Firm	Happy?	Part	Firm	Happy?
1	A	yes	12	B	yes
2	A	no	13	B	no
3	A	no	14	B	yes
4	A	yes	15	B	yes
5	A	no	16	B	yes
6	A	yes	17	B	yes
7	A	no	18	B	no
8	A	no	19	B	no
9	A	no	20	B	no
10	A	no	21	B	yes
11	A	no	22	B	yes
			23	B	yes
			24	B	yes

To begin the analysis, go to the **Analyze** drop-down menu. Click on **Analyze – Descriptive Statistics – Crosstabs**, and the **Crosstab** dialogue box appears

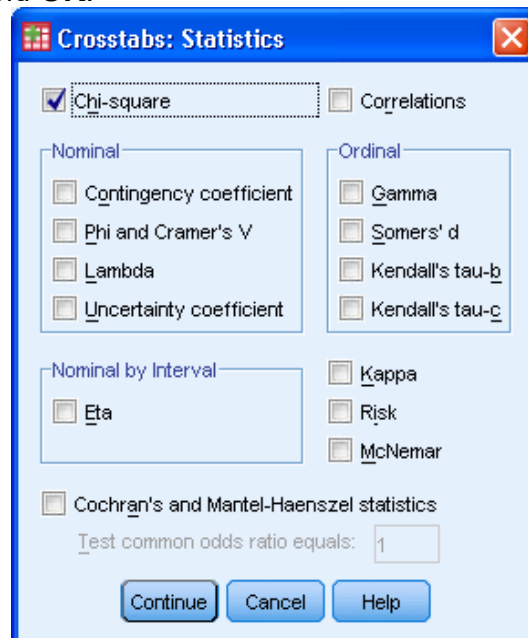
<sup>41</sup> If you need a reminder of how to do this, refer to section 2.4.

(Figure 7.1). Move one of the variables you are testing into **Row(s)** and the other into **Column(s)**. Which way round you do it will not affect the result of the test, only the layout of the output<sup>42</sup>. Tick **Display clustered bar charts**.



**Figure 7.1.** Crosstabs dialogue box.

Click on **Statistics** and in the next dialogue box (Figure 7.2) check **Chi-square**. Click on **Continue** and **OK**.



**Figure 7.2.** Crosstabs Statistics box.

<sup>42</sup> Personally, if there is an Independent and Dependent Variable I prefer to put the IV in Rows and the DV in columns.

Look first at the part of the output shown in Figure 7.3.

firm \* happy Crosstabulation

Count		happy		Total
		yes	no	
firm	A	3	8	11
	B	9	4	13
Total		12	12	24

**Figure 7.3.** Crosstabulation output.

Each combination of categories (e.g. firm A, responded yes) is known as a 'cell'. The table summarises our data. In fact, it represents the descriptive statistics for the study. For example, of the 11 interviewees in firm A, it shows that 3 responded yes and 8 no.

The result of the inferential (chi-squared) test is given in the first line of the test output (Figure 7.4). We also need to look at the footnotes to see if any of the 'expected values' are less than 5. (See lecture notes regarding constraints on the use of this test.)

Sample size, <i>N</i>	Value of the chi-square statistic	Degrees of freedom	Significance of the statistic		
Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.196 <sup>b</sup>	1	.041		
Continuity Correction <sup>a</sup>	2.685	1	.101		
Likelihood Ratio	4.332	1	.037		
Fisher's Exact Test				.100	.050
Linear-by-Linear Association	4.021	1	.045		
N of Valid Cases	24				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.50.

**Figure 7.4.** Chi-square test output.

We could write the result as "There was a significant difference in responses of interviewees in the two firms, chi-square(1, *N*=24) = 4.20, *p* = .041."<sup>43</sup> (Notice that we also include the sample size, *N*.) We would also need to give the

<sup>43</sup> or " $\chi^2(1) = 4.20, p = .041$ "; see Appendix A.

information from Figure 7.3, either in that format, in words, or in a bar chart such as the one which SPSS has given us.

## 7.4 Example 2: using the Weighted Cases procedure in SPSS

It would often be very tedious to enter this kind of data line by line into SPSS. There is an alternative, which breaks our usual rule about entering one line per case.

Consider an experiment by Cialdini, Reno and Kallgren (1990). People were handed a leaflet, as they were about to walk along a path which had a predetermined number of pieces of litter on it (placed there by the experimenters). They were observed to see whether they dropped the leaflet as litter. The results are shown in Table 7.2.

**Table 7.2.** Data for example 2.

Amount of litter on path	Behaviour of person	
	Dropped litter	Did not drop litter
0 or 1 piece ( <i>small</i> )	17	102
2 or 4 pieces ( <i>medium</i> )	28	91
8 or 16 pieces ( <i>large</i> )	49	71

In **Variable View** define three variables

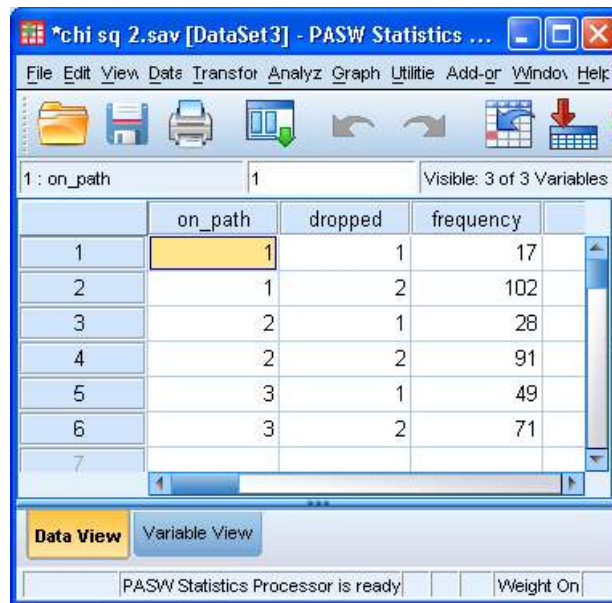
- Name: *on\_path*. Decimals: 0. Values: 1 = *small*, 2 = *med*, 3 = *large*.
- Name: *dropped*. Decimals: 0. Values: 1 = *yes*, 2 = *no*.
- Name: *frequency*. Decimals: 0.

Go back to **Data View** and enter the data as in Figure 7.5.

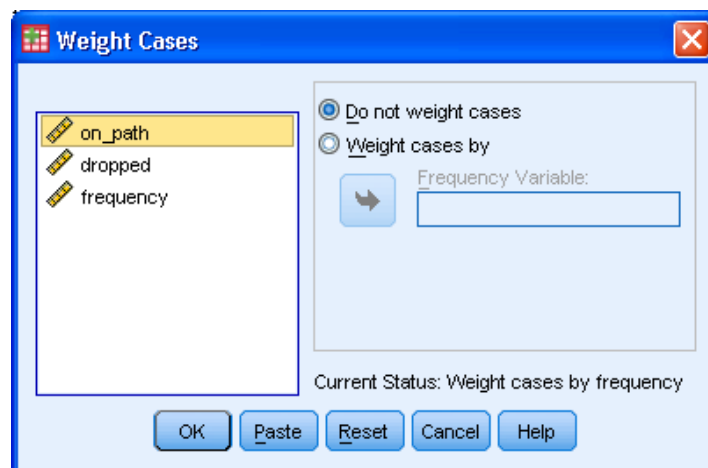
To use the special procedure, go to **Data – Weight Cases** on the drop-down menu. A new dialogue box appears (Figure 7.6).

Click on **Weight Cases by** and move *frequency* into the box marked **Frequency Variable**. Click **OK**. Now, the computer will think there are as many lines as there are in the *frequency* variable. For example, it will think that there are 17 lines in which *on\_path* is 'small' and *dropped* is 'yes'.

Now follow the same procedure as in paragraph 7.3 to do the chi-square test. (**Analyze – Descriptives – Crosstabs**. Ask for clustered bar charts. Under **Statistics**, tick **chi-square**. ) To get the table the same way round as the original one, put *on\_path* in **Rows** and *dropped* in **Columns**. (It does not make any difference to the chi-square test which we put in rows and which in columns. Notice that *Frequency* does *not* go in either rows or columns!)



**Figure 7.5.** Example 2 data entered into SPSS.



**Figure 7.6.** Weight Cases dialogue box.

Our output (Figure 7.7 and Figure 7.8) is similar to before; any differences in format are due to the extra columns in the table, not to the way we entered the data).

Check the footnote to see whether there are any problems with expected counts being less than 5 (see lecture notes about constraints on this test). In this case there are not. We report the result in just the same way as in the previous section. So we could say that whether people dropped litter was significantly affected by whether there was already litter on the path, chi-square (2,  $N = 358$ ) = 22.4,  $p < .001$ . Again, we would include the descriptive statistics in our report, and you may find the bar chart useful (Figure 7.9).



**on\_path \* dropped Crosstabulation**

Count		dropped		Total
		yes	no	
on_path	small	17	102	119
	med	28	91	119
	large	49	71	120
Total		94	264	358

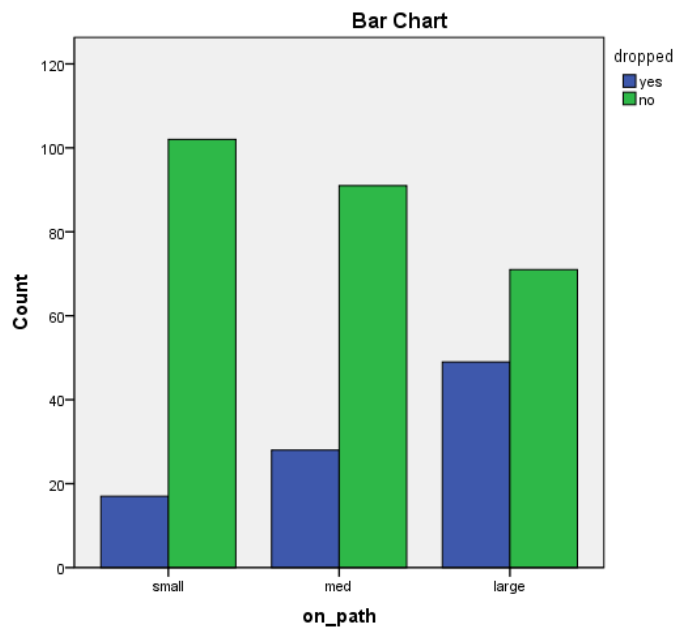
**Figure 7.7.** Example 2 output: table.

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22.433 <sup>a</sup>	2	.000
Likelihood Ratio	22.463	2	.000
Linear-by-Linear Association	21.706	1	.000
N of Valid Cases	358		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 31.25.

**Figure 7.8.** Example 2 output: test results.



**Figure 7.9.** Clustered bar graph.

## 7.5 Showing percentages

It is often useful to include percentages in tables. These can be obtained by choosing an extra option when following the procedure in paragraph 7.3 or 7.4. Click on **Cells** and you will see you get a choice of percentages by **Row**, **Column** or **Total**. If, for example, in example 2 you tick the box that says **Row**, you are shown what percentage of people dropped litter in each situation (Figure 7.10).

			dropped		Total
			yes	no	
on_path	small	Count	17	102	119
		% within on_path	14.3%	85.7%	100.0%
	med	Count	28	91	119
		% within on_path	23.5%	76.5%	100.0%
	large	Count	49	71	120
		% within on_path	40.8%	59.2%	100.0%
Total	Count	94	264	358	
	% within on_path	26.3%	73.7%	100.0%	

**Figure 7.10.** Example 2 output: table with percentages.

## 7.6 Effect sizes

You can report an effect size when doing a chi-square test. This can be obtained by choosing an extra option when following the procedure in paragraph 7.3 or 7.4. In the **Statistics** dialogue box, choose **Phi and Cramér's V** as well as **Chi-square**. You get the additional output shown in Figure 7.11. Report Cramér's V. (For a 2×2 table you can report Phi, but it will be the same anyway.)

**Symmetric Measures**

		Value	Approx. Sig.
Nominal by Nominal	Phi	.250	.000
	Cramer's V	.250	.000
N of Valid Cases		358	

- a. Not assuming the null hypothesis.
- b. Using the asymptotic standard error assuming the null hypothesis.

**Figure 7.11.** Output for Cramér's V.

## 8 Chi-square tests of a single categorical variable

### 8.1 When they are used

In addition to the uses in the previous chapter, we can use chi-squared tests to examine:

- (a) *whether a categorical variable is evenly distributed.* For example, if there are three computers in an office and we count how many people use each computer, is there a significant difference between the numbers using each computer?
- (b) *whether a categorical variable is distributed in a given proportion.* For example, if there are 13 boys and 17 girls in a class, is the teacher giving individual attention to the boys and girls in proportion to those numbers?

### 8.2 Whether a categorical variable is evenly distributed

Suppose that there are three computers in a room. A researcher finds out the number of times each computer was logged onto over the course of a week, as shown in Table 8.1.

**Table 8.1.** Number of times three computers were used.

Computer number	1	2	3
Times used	45	29	62

We could enter the data in 136 separate lines, with each showing a case number (1-136) and computer number (1, 2 or 3). However this would be pointless effort unless we already had a data file with this information on it. Here, we will use the 'weight cases' procedure, as we did in paragraph 6.4 above.

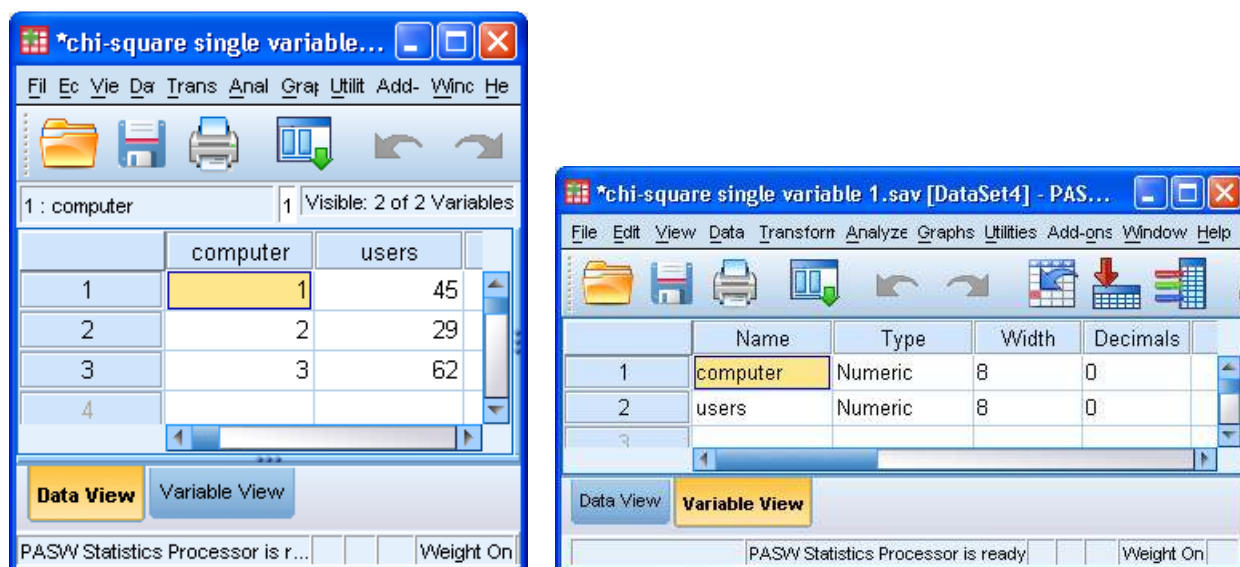
Enter the data into SPSS, as shown in Figure 8.1.

To weight the data, go to **Data – Weight cases**, click on **Weight cases by** and put the count (*users*) into the **Frequency Variable** box. Click **OK**.

To do the analysis, go to **Analyze – Nonparametric tests – Legacy Dialogs<sup>44</sup> – Chi-Square** and put our variable of interest (*computer*) into the **Test Variable** list. Check that under **Expected Values** it shows **All values equal** and click **OK**.

---

<sup>44</sup> In versions earlier than SPSS 18, ignore the 'Legacy Dialog' step. In SPSS 18, there is an alternative which does not use the 'Legacy Dialog' step but it seems unnecessarily complicated.



**Figure 8.1.** Data view and variable view for **Table 8.1**.

The first part of the output [Figure 8.2(a)] confirms the observed number of computer users in each condition, and that the 'Expected' numbers (the split we are testing against) are equal. The second part figure [Figure 8.2(b)] gives us the result of the chi-squared test. This is quite easy to read off as there is no extraneous information. Hence we can write: Computer 1 was used 45 times, computer 2 29 times, computer 3 62 times. This was significantly different from an even split,  $\chi^2(2) = 12.0, p = .002$ .

computer			
	Observed N	Expected N	Residual
1	45	45.3	-.3
2	29	45.3	-16.3
3	62	45.3	16.7
Total	136		

(a) counts

Test Statistics	
	computer
Chi-Square <sup>a</sup>	12.015
df	2
Asymp. Sig.	.002

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 45.3.

(b) test results

**Figure 8.2.** Chi-square test output.

### 8.3 Whether a categorical variable is split in a given proportion

Sometimes the proportion under the null hypothesis would not be evenly split. For example, suppose that there are 13 boys and 17 girls in a class. Is the teacher allocating her time fairly between the boys and the girls? If so we would not expect her to give equal time to boys and girls, but to allocate it in the proportion 13:17.

Perhaps a researcher finds that in a given period of time this teacher gives individual attention 50 times to boys and 40 times to girls. Enter the data into SPSS, as shown in Figure 8.3. (If you wish, use the **Data Labels** field in Variable View to show that *gender* 1 is male and *gender* 2 is female.)

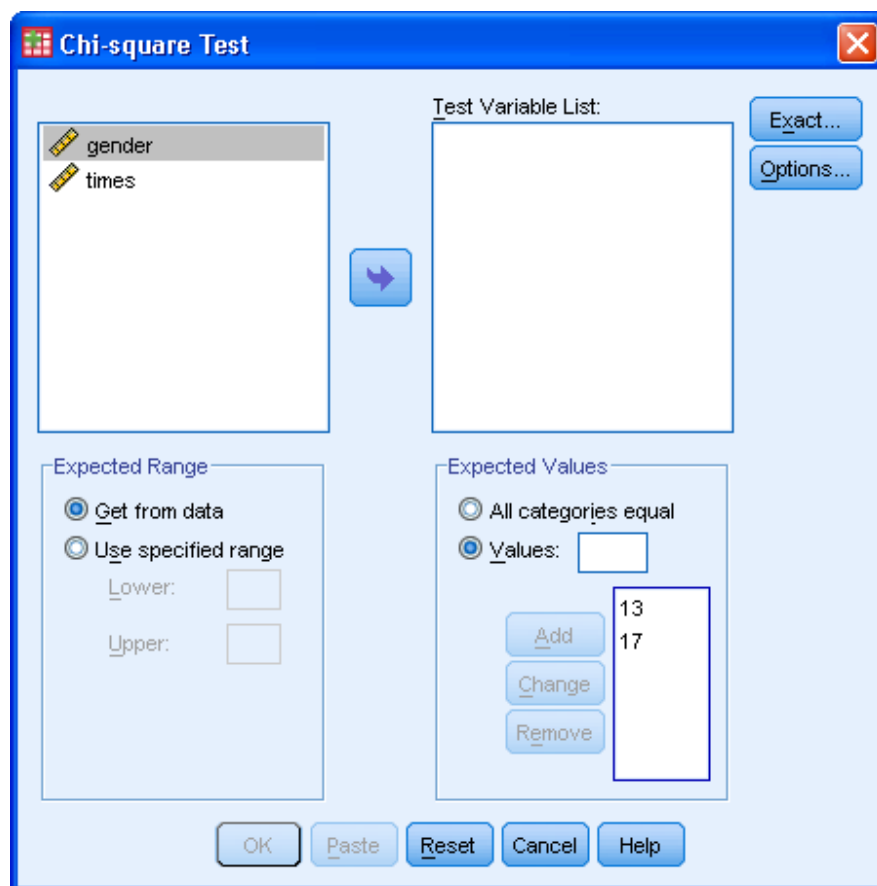
Use the weight cases procedure (**Data – Weight cases**) to weight the cases by *times*.

	gender	times
1	1	50
2	2	40

**Figure 8.3.** Data for gender example.

Go to **Analyze – Nonparametric Tests – Legacy Dialog<sup>45</sup> – Chi-Square** and a dialogue box will come up. Put the variable of interest (*gender*) into the **Test Variable** box. To tell SPSS what the expected proportions are (under the null hypothesis), go to **Expected Values** underneath the **Test Variable** list. Click on the second radio button, **Values**. It is very important that you put the values in the same order as the order of the categories in the test variable (i.e. in this case boys first, then girls). The expected proportions are the proportions of boys to girls, so put the number of boys (13) into the box next to **Values**. Click on **Add** and enter the number in the next category (17, for girls). Click **Add** again. Your dialogue box should now look like Figure 8.4. Click on **OK**.

<sup>45</sup> In versions earlier than SPSS 18, ignore the 'Legacy Dialog' step. In SPSS 18, there is an alternative which does not use the 'Legacy Dialog' step but it seems unnecessarily complicated.



**Figure 8.4.** Dialogue box for gender example.

The output (Figure 8.5) is quite similar to last time. The first part (a) tells you the number of times the teacher gave individual attention to each gender, and the number expected under the null hypothesis. The second part (b) gives the result of the chi-square test.

gender			
	Observed N	Expected N	Residual
boys	50	39.0	11.0
girls	40	51.0	-11.0
Total	90		

(a) counts

Test Statistics	
	gender
Chi-Square <sup>a</sup>	5.475
df	1
Asymp. Sig.	.019

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 39.0.

(b) test results

**Figure 8.5.** Chi-square test output.

Thus the teacher gave individual attention significantly more often to boys than to girls, chi-square (1) = 5.48,  $p = .019$ .

## 9 Cochran's and McNemar's tests

### 9.1 When to use Cochran's and McNemar's tests

These tests are the equivalent to chi-square tests, but when one of the variables (usually conceptualised as the Independent Variable) is repeated-measures (within-subjects).

### 9.2 Cochran's Q

Twenty drug addicts are asked whether they think that three different drugs (A, B and C) should be legalised. Their responses are shown in Table 9.1. Amongst our respondents, is there a statistically significant difference in attitude to legalisation of the three drugs?

**Table 9.1.** Data for Cochran's example.

Respondent	Drug A	Drug B	Drug C
1	yes	no	yes
2	yes	no	yes
3	no	yes	yes
4	no	yes	yes
5	no	yes	yes
6	no	yes	yes
7	no	yes	yes
8	no	yes	no
9	no	yes	no
10	no	yes	no
11	no	yes	no
12	no	yes	yes
13	no	yes	yes
14	no	yes	no
15	no	yes	no
16	no	no	no
17	no	no	no
18	no	no	no
19	no	no	no
20	no	no	no

Enter the data into SPSS. Remember in Variable View to set values for the variables, e.g. 1 for Yes and 2 for No.

Click on **Analyse – Nonparametric tests – Legacy Dialogs – K related samples**. In the dialogue box enter the three columns into the **Test Variables** box. De-select the **Friedman test** and select **Cochran's Q** (Figure 9.1).

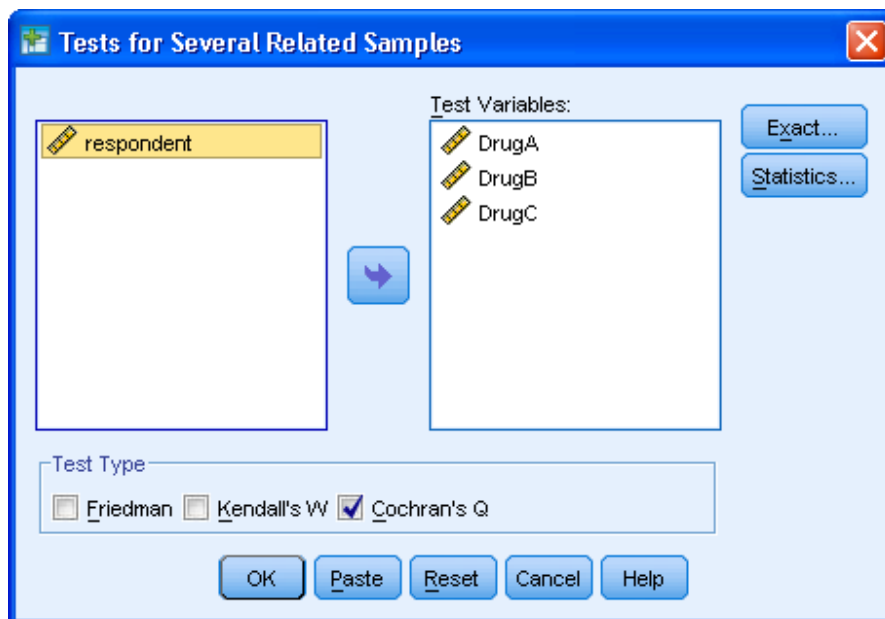


Figure 9.1. Dialogue box for Cochran's test.

Click on **OK**, and you get a cross-tabulation and the test result (Figure 9.2)

Frequencies			Test Statistics	
	Value			
	1	2		
DrugA	2	18	N	20
DrugB	13	7	Cochran's Q	12.400 <sup>a</sup>
DrugC	9	11	df	2
			Asymp. Sig.	.002

a. 1 is treated as a success.

Figure 9.2. Output for Cochran's test.

From the test statistics, we can say that there was a significant difference in attitudes to legalisation of the three drugs, Cochran's  $Q = 12.4$ ,  $p = .002$ . Don't forget to include the descriptive statistics (e.g. as shown in the table, or perhaps express them as percentages).

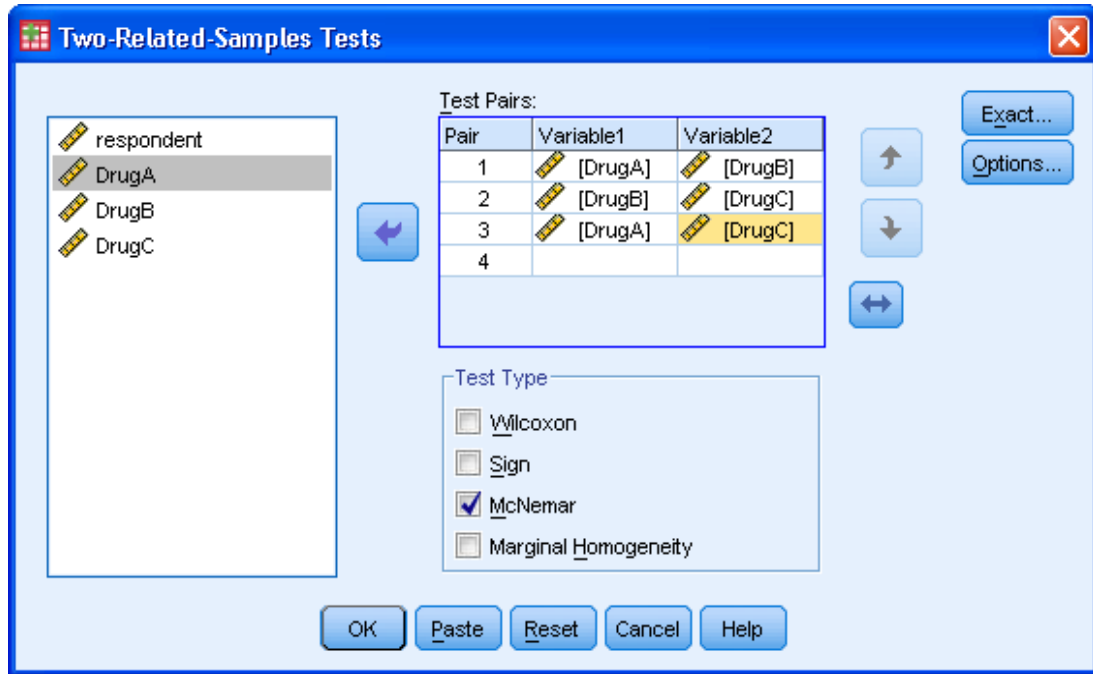
### 9.3 McNemar's test

Now let us examine whether there are significant differences in attitude between particular pairs of drugs.

Click on **Analyse – Nonparametric tests – Legacy Dialogs – 2 related samples**. In the dialogue box enter each pair of columns into the **Test Variables** box (we want to make three comparisons here, and we can do them all at the



same time). De-select the **Wilcoxon test** and select **McNemar** (Figure 9.3). Click **OK**.



**Figure 9.3.** Dialogue box for McNemar test.

You get a cross-tabulation for each pair of drugs (Figure 9.4), and the test results (Figure 9.5)

DrugA & DrugB			DrugB & DrugC			DrugA & DrugC		
DrugA	DrugB		DrugB	DrugC		DrugA	DrugC	
	1	2		1	2		1	2
1	0	2	1	7	6	1	2	0
2	13	5	2	2	5	2	7	11

**Figure 9.4.** Friedman's example: crosstabulations.

Test Statistics <sup>b</sup>			
	DrugA & DrugB	DrugB & DrugC	DrugA & DrugC
N	20	20	20
Exact Sig. (2-tailed)	.007 <sup>a</sup>	.289 <sup>a</sup>	.016 <sup>a</sup>

a. Binomial distribution used.

b. McNemar Test

**Figure 9.5.** Friedman's example: test results.

Remembering to use a Bonferroni correction, we could report: McNemar tests (Bonferroni-corrected for three comparisons) showed that there was a significant difference between attitudes to drugs A and B ( $p = .021$ ) and drugs A and C ( $p = .048$ ) but not between drugs B and C ( $p = .867$ ).

## 10 Simple regression and correlation

“Simple” in this context means that there are only two variables. (When we come to regression, we will have to choose one of these as an Independent Variable (IV) and the other as a Dependent Variable (DV). This is not necessary for correlation.)

We will work with the example data in Table 10.1: the length of time that eleven students studied for a test and their scores. Enter them into SPSS.

**Table 10.1.** Example data for correlation and regression.

Student	Hours studying	Test score
1	0	5
2	1	16
3	2	23
4	3	26
5	4	24
6	5	25
7	6	38
8	7	41
9	8	53
10	9	48
11	10	56

### 10.1 Scatterplots.

#### To create a scatterplot

- From the drop-down menu, click on **Graphs – Legacy Dialogs – Scatter/Dot**
- In the dialogue box, choose **Simple Scatter** and click **Define**
- Move *Hours studying* (our IV<sup>46</sup>) into **X Axis**
- Move *Test Score* (our DV) into **Y axis**
- Click on **OK**.

#### To add a trend line:

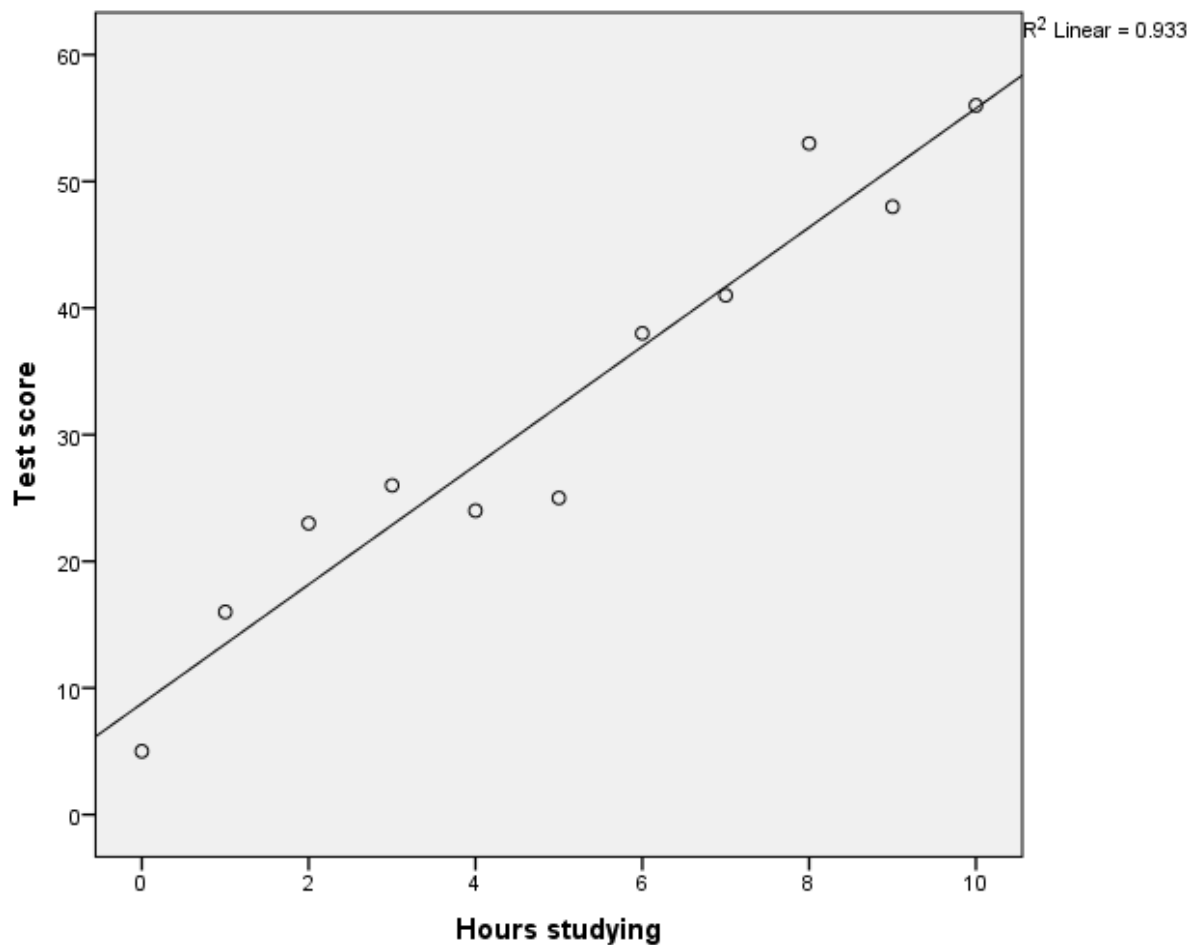
- Double click on the graph to open the **Chart Editor**

---

<sup>46</sup> If you have an IV and DV, it is conventional to put them this way round.

- On the drop down menu, click on **Elements – Fit Line at Total**
- Uncheck the box at bottom left marked **Attach label to line** (unless of course you want this label, which shows the regression equation on the line)
- Otherwise, keep the default options
- Click on **Close**.

Your graph should look like Figure 10.1.



**Figure 10.1.** Scatterplot for example data.

Notice that:

1. You can remove the box at the right ( $R^2 \text{ linear} = 0.933$ ) whilst in Chart Editor, by clicking on it and hitting **Delete** on your keyboard.
2. As usual with SPSS charts, if you want to copy it to another application, click on **Copy Special** and choose **Image**.

## 10.2 Correlation

For this example, we will create two sets of output (for Pearson's  $r$  and Spearman's rho). Normally we would only ask for one of these, depending on whether we wanted a parametric test (Pearson's  $r$ ) or non-parametric test (Spearman's rho).

From the drop-down menu click on **Analyze – Correlate – Bivariate**.

In the dialogue box:

- Move our two variables (*Hours studying* and *Test score*) into the box marked **Variables**
- Under **Correlation Coefficients** tick **Pearson** and **Spearman**
- Click **OK**.

### 10.2.1 Parametric test of correlation (Pearson's $r$ )

The result of the parametric test is shown in Figure 10.2. Yes, the information is all there twice! This layout would make more sense if you had asked for the correlations between several variables at the same time, but SPSS uses it even when there are only two variables.

Correlations				
		Hours studying	Test score	
Hours studying	Pearson Correlation	1	.966**	The correlation coefficient ( $r$ ) between 'hours' and 'score'
	Sig. (2-tailed)		.000	
	N	11	11	
Test score	Pearson Correlation	.966**	1	The significance level of the correlation
	Sig. (2-tailed)	.000		
	N	11	11	

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

The number of cases included

**Figure 10.2.** Pearson's test results.

We could write this up as 'There was a significant correlation between the hours of study and the test score,  $r = .966$ ,  $p < .001$ .' (Remember that if SPSS prints .000, we write  $< .001$ ).

Many people consider  $r^2$  to be more meaningful than  $r$ . It is the amount of *shared variance* between the variables, or "the extent to which one variable *explains* the other" (whether it really explains it depends on the validity of your study). You can calculate  $r^2$  by hand:  $r^2 = r \times r = .966 \times .966 = .933$ .

### 10.2.2 Non-parametric test of correlation (Spearman's $\rho$ )

The result of the non-parametric test has a similar layout (Figure 10.3).

We could write the result of this test as “There was a significant correlation between the hours of study and the test score, Spearman’s rho = .964,  $p < .001$ .”

Correlations				
Spearman's rho	Hours studying	Correlation Coefficient	1.000	.964**
		Sig. (2-tailed)	.	.000
		N	11	11
	Test score	Correlation Coefficient	.964**	1.000
		Sig. (2-tailed)	.000	.
		N	11	11

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

The correlation coefficient (rho) between 'hours' and 'score'

The significance level of the correlation

The number of cases included

Figure 10.3. Spearman’s test result.

## 10.3 Simple linear regression

### 10.3.1 Carrying out a regression

In regression, we have to choose one variable as the (hypothesised) Independent Variable (IV) and the other as the Dependent Variable (DV). The IV is often known as the ‘predictor’ and the DV as the ‘criterion’. However SPSS uses the familiar terms, IV and DV. ‘Simple’ regression means that there is only one IV.

The procedure assumes that any relationship between the IV and the DV is linear. It is good practice to do a scatterplot of the IV against the DV to check this, as above.

We will carry out a regression with the same data we used for our correlation example.

- From the drop-down menu, click on **Analyze – Regression – Linear**.
- In the dialogue box, move *Test score* into the **Dependent** box and *Hours studying* into the **Independent(s)** box.
- Click on **OK**.

### 10.3.2 Regression output

As usual, we are only interested in part of the output.

The Model Summary (Figure 10.4) provides information about correlations.  $R$  is the same as  $r$  from our correlation, and  $R^2$  is the same as  $r^2$ . Remember that  $R^2$  is the amount of shared variance. Whilst some people would use  $R^2$  as an estimate of the shared variance for the population, others prefer “Adjusted  $R$  square”, which is adjusted to allow for sample size.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.966 <sup>a</sup>	.933	.926	4.401

a. Predictors: (Constant), Hours studying

**Figure 10.4.** Model Summary output.

The Anova (Figure 10.5) tells us whether  $R$  is significantly different from zero – whether our equation is significantly better than just guessing which score relates to which number of hours studying. In this case it is significant,  $F(1,9) = 125.48$ ,  $p < .001$ .

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2429.900	1	2429.900	125.481	.000 <sup>a</sup>
	Residual	174.282	9	19.365		
	Total	2604.182	10			

a. Predictors: (Constant), Hours studying

b. Dependent Variable: Test score

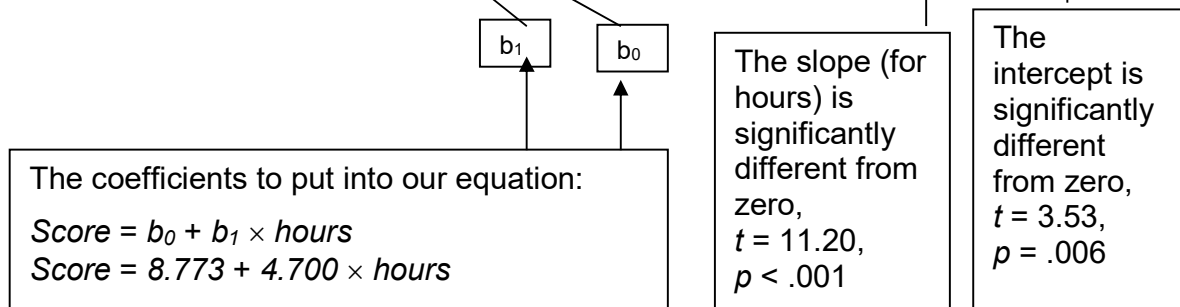
**Figure 10.5.** Anova table from regression.

The coefficients table gives a lot of information, as shown in Figure 10.6. The regression equation is the equation which describes the best fit line in Figure 10.1. 'Slope' is the slope of that line (how much *score* increases for an increase of 1 in *hours*) and 'intercept' is the intercept of that line (what the value is of *score* when *hours* = 0).

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.773	2.482		3.534	.006
	Hours studying	4.700	.420	.966	11.202	.000

a. Dependent Variable: Test score



**Figure 10.6.** Coefficients table and interpretation.

### 10.3.3 Writing up regression

You could say “A linear regression showed that the number of hours studying was a significant predictor of the score,  $R = .966$ ,  $R^2 = .933$ , adjusted  $R^2 = .926$ ,  $F(1,9) = 125.48$ ,  $p < .001$ . Coefficients are shown in table xxx”, where table xxx reproduces the SPSS coefficients table. Depending on what you were investigating, you might want to write out the regression equation, and/or explain it (e.g. “The equation estimates that each extra hour’s studying results in an increase of 4.7 in the score achieved in the test.”)

### 10.3.4 What it means

The regression equation was  $\text{Score} = 8.773 + 4.700 \times \text{Hours}$ . We can use this to predict the score for any given number of hours.

For example, if somebody had studied for 2 hours we predict that their score would be  $8.773 + 4.700 \times 2 = 8.773 + 9.400 = 18.173$ , which we can round to 18.2<sup>47</sup>.

Actually, the student who did study for 2 hours has a score of 23. The difference ( $23 - 18.173 = 4.827$ ) is known as a *residual*. This difference might arise because:

- Students’ performance may vary for reasons other than hours of study (e.g. ability of the student, mood, random fluctuations)
- The regression is not exact: for example, the relationship between hours of study and score is not exactly a straight line.

Similarly, you can estimate how much someone would score if they studied for 2.5 hours (20.55). Because this is in between two values of the IV that we have (2 and 3) it is known as *interpolation*.

The equation also allows you to estimate a score if someone had studied beyond the number of hours that were in the study (e.g. 11 hours). This is known as *extrapolation* and you need to beware of it. Can you rely on it? Would the score really go on increasing at the same rate for ever? For example, there is probably a maximum score on the test.

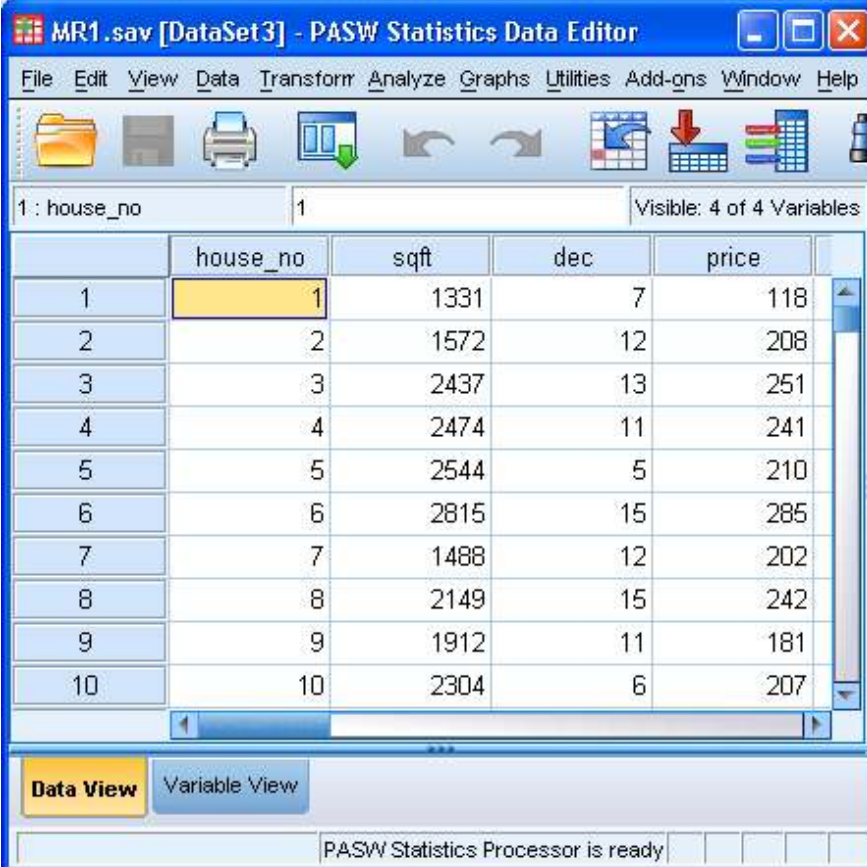
---

<sup>47</sup> There is a slightly difficult issue about rounding here. If you are going on to use a figure in subsequent calculations, it makes sense to keep as many decimal places as possible. If you are reporting a figure to somebody else, you do not want to suggest that it is more accurate than it really is by giving too many decimal places. You just need to apply common sense and decide what the figure is being used for.

## 11 Multiple regression and correlation

In Multiple Regression we still have one Dependent Variable (DV) but now we have more than one Independent Variable (IV). The basic procedure is the same as that for simple regression (section 10.3). However, due to some statistical problems that can occur with multiple regression, we will request some additional output.

Suppose that an estate agent thinks that the selling price of houses in her area (in thousands of pounds) is related to their size in square feet and to the state of decoration. Fortunately her data meet parametric assumptions. She looks up records for 100 houses and enters them into SPSS as shown in Figure 11.1.



MR1.sav [DataSet3] - PASW Statistics Data Editor

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

1 : house\_no 1 Visible: 4 of 4 Variables

	house_no	sqft	dec	price
1	1	1331	7	118
2	2	1572	12	208
3	3	2437	13	251
4	4	2474	11	241
5	5	2544	5	210
6	6	2815	15	285
7	7	1488	12	202
8	8	2149	15	242
9	9	1912	11	181
10	10	2304	6	207

Data View Variable View

PASW Statistics Processor is ready

**Figure 11.1.** Extract from Data View for multiple regression example.

Since this is a large data file, it will be provided on grad.gold as MR1.sav.

As for simple linear regression, the procedure assumes that any relationships between the IVs and the DV are linear. It is good practice to do a scatterplot of each IV against the DV to check this (see paragraph 10.1).



Then follow a similar procedure as for simple regression:

- From the drop-down menu, click on **Analyze – Regression – Linear**.
- In the dialogue box, move *price* into the **Dependent** box and *sqft* and *dec* into the **Independent(s)** box.
- Click on **Statistics** and ask for **Descriptives, Part and partial correlations, and Collinearity diagnostics** (in addition to **Model fit** and **Estimates**, which are selected by default). Press **Continue** and **OK**.

Examine the output. Some of it is the same as we got in simple linear regression.

The model summary (Figure 11.2) shows a correlation ( $R$ ). Now that we have more than one variable, this is a multiple correlation. This is best understood in terms of the squared multiple correlation ( $R^2$ , or  $R^2_{adj}$ ), which is the amount of variance in the DV that is shared or 'explained' by the IVs. Of course, whether the IVs really explain the DV depends on the validity of the study. Subject to that, our interpretation is that 60% of the variance in selling price of the houses is explained by their state of decoration and their size in square feet.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.777 <sup>a</sup>	.603	.595	21.871

a. Predictors: (Constant), dec, sqft

**Figure 11.2.** Model summary for multiple regression example.

The Anova (Figure 11.3) tells us whether  $R$  is significantly different from zero – whether our equation is significantly better than just guessing which price relates to which values of the IVs. In this case it is significant,  $F(2,97) = 73.77$ ,  $p < .001$ .

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70575.313	2	35287.656	73.773	.000 <sup>a</sup>
	Residual	46397.597	97	478.326		
	Total	116972.9	99			

a. Predictors: (Constant), dec, sqft

b. Dependent Variable: price

**Figure 11.3.** Anova for multiple regression example.

The Coefficients table (Figure 11.4) tells us more than ever. Firstly, it tells us the regression equation, as for a simple regression but a bit longer. If you publish

this result, you would include all the coefficients<sup>48</sup>, whether they were significant or not. Secondly, it tells us which coefficients were in fact significant. If you are carrying out an exploratory study you can use this information to say which IVs were in fact significant predictors of the DV (but read the rest of this section first!).

Some of the extra information we asked for is at the end of this table. If there are any big differences between the zero-order and the partial correlations, this shows that there are correlations between the Independent Variables. The *zero-order correlations* are the ordinary ones we have come across before, the correlation between each IV and the DV. The *partial correlations* are the unique correlation of that IV with the DV, that is to say how much of their relationship with the DV is not shared by any of the other IVs. If the two correlations are very different, you should tell your readers both.

Check the figures against each variable under VIF (Variance Inflation Factor). If any of them are too big (say, greater than 4<sup>49</sup>), that IV has too much shared variance with the other IVs – that is to say, it has a high correlation with one or more of them<sup>50</sup>. This messes up the maths and stops the regression from working properly. Think about whether you can re-run the analysis with one of the IVs removed – the high correlation may mean it is measuring something very similar to one of the other IVs anyway. If more than one VIF is too high, you can experiment with removing one IV at a time.

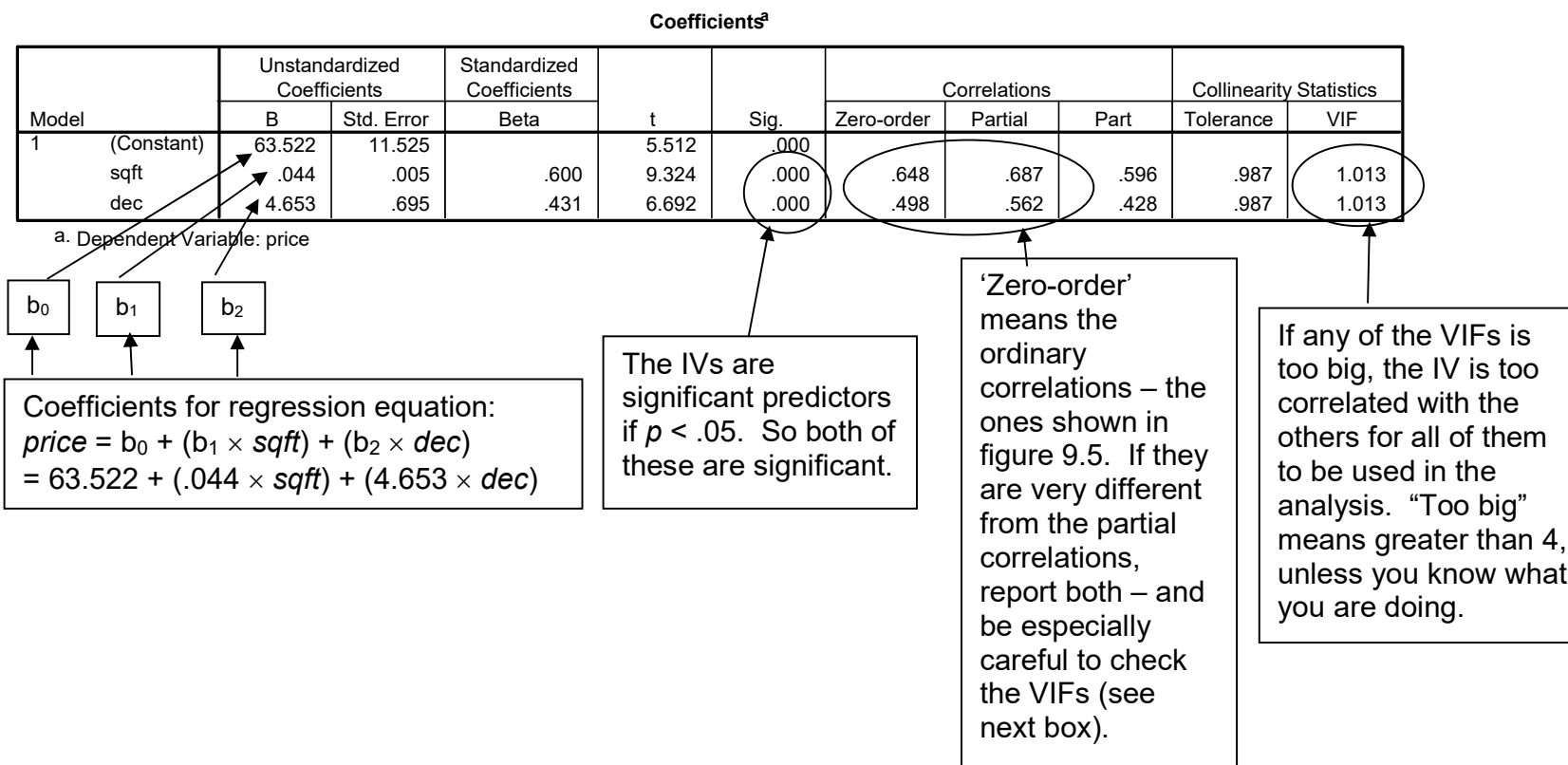
It is also a good idea to look at the correlations table (Figure 11.5), which was produced because we checked the ‘Descriptive statistics’ box. Check the significance of the correlations of each IV against the DV (notice that the significance levels are given as one-tailed; I recommend doubling them to give the two-tailed significance). If any of them are significant in this table, but those IVs are not significant in the Coefficients table, this is caused by correlations between the IVs. Again you should report both and be particularly careful to check the VIFs. Also, if the table showed a significant correlation between the IVs, we would report this and take the same precautions. However, in this case there is no significant correlation between the IVs *sqft* and *dec* ( $r = .112$ ,  $p = .134$ ).

---

<sup>48</sup> unless you repeated the analysis with the non-significant predictors excluded. However, some readers might find this controversial unless you had a prior hypothesis that these predictors would be nonsignificant.

<sup>49</sup> Some people will allow higher figures, such as 10, but that takes us into matters of opinion best left to people who are experienced at this kind of analysis.

<sup>50</sup> Strictly speaking, what is too high is the squared multiple correlation with the other IVs.



**Figure 11.4.** Coefficients table for multiple regression, and interpretation.

Correlations				
		price	sqft	dec
Pearson Correlation	price	1.000	.648	.498
	sqft	.648	1.000	.112
	dec	.498	.112	1.000
Sig. (1-tailed)	price	.	.000	.000
	sqft	.000	.	.134
	dec	.000	.134	.
N	price	100	100	100
	sqft	100	100	100
	dec	100	100	100

Figure 11.5. Correlations table for multiple regression example.

## 12 Introduction to statistics for questionnaires

### 12.1 Overview

Data files for questionnaires usually contain many variables (i.e. there are a lot of questions) and many lines (i.e. there are a lot of participants). We will use a sample questionnaire file to illustrate a number of general points about such files, as well as some that are specific to questionnaires. We will cover:

- how to set out and enter data for a questionnaire (section 12.2)
- some techniques to help us find errors in our data entry (sections 14.1 and 14.2)
- some calculations that might be needed, and how to do them (section 12.3)
- some statistical techniques particularly aimed at questionnaires (section 12.4).

### 12.2 Entering the data

#### 12.2.1 Introduction: example data

The data file (to be provided on grad.gold) contains the results of an imaginary questionnaire, given to 60 participants.

The data file contains the sort of information you might collect from a simple questionnaire:

- participant numbers (*Part*). It is good practice to give each participant a number, and to write the same number on each questionnaire. This allows you to check back to the questionnaire if you need to. Do not

- depend on the line numbers in SPSS, because SPSS sometimes changes the order of the lines.
- demographic information (*Gender* and *Age*)
  - participants' responses to six questions (*Q1* to *Q6*). These are related to job satisfaction (e.g. "I like the work I do"). Responses are on a Likert scale running from 1 (strongly disagree) to 7 (strongly agree).

The first few lines of the file are shown in Figure 12.1. (They may be shown as in Figure 12.2 depending on your settings; see paragraph 12.2.2).

	Part	Gender	Age	Q1	Q2	Q3	Q4	Q5	Q6
1	1	1	23	7	4	3	5	6	5
2	2	2	46	2	2	5	1	4	2
3	3	2	42	4	4	5	4	3	3
4	4	1	18	5	5	2	5	6	7
5	5	2	21	5	6	1	7	7	7

**Figure 12.1.** Example data file (with Data Labels turned off).

	Part	Gender	Age	Q1	Q2	Q3	Q4	Q5	Q6
1	1	Male	23	Strongly a...	Neutral Tend to dis...	Tend to agr...	Agree Tend to agr...	Disagree	
2	2	Female	46	Disagree	Disagree Tend to agr...	Strongly Di...	Neutral	Disagree	
3	3	Female	42	Neutral	Neutral Tend to agr...	Neutral Tend to dis...	Tend to dis...		
4	4	Male	18	Tend to agr...	Tend to agr...	Disagree Tend to agr...	Agree	Strongly a...	
5	5	Female	21	Tend to agr...	Agree Strongly Di...	Strongly a...	Strongly a...	Strongly a...	

**Figure 12.2.** Example data file (with Data labels turned on).

## 12.2.2 Variable view

The example file is already set up in Variable View. Notice that **Value Labels** have been set up for *Gender* and for the answers to questions 1 to 6 (*Q1* to *Q6*). (See section 2.4 if you need a reminder of how to set these up.) If you are entering your own data it is not essential to set up **Value Labels** for the answers, and if you decide you do want them later, you can always set them up then. As you know, in **Data View** you can show either the values or the **Value Labels** (as

in Figure 12.1 or Figure 12.2 respectively). You can swap between these views by clicking on **View – Value Labels**.

If several questions have the same set of responses, you can enter the **Value Labels** for one question, then (still in **Variable View**) copy and paste them to the other questions. But check the questionnaire carefully to make sure that all the questions really do have the same numbers for the same responses. (For example, if there are reverse-coded questions, 1 might mean *strongly disagree* for some questions and *strongly agree* for others.)

Since the responses will probably be whole numbers, it will help clarity if you set the variables to have 0 decimal places (as in the example file).

### 12.2.3 Entering data

Often, researchers collect data via the Internet, in which case your only problem is to make sure you know how to download it in a form you can read into SPSS.

If you have to enter the data by hand, you have two options. With the **Value Labels** off (Figure 12.1), you can simply enter the numbers. With the **Value Labels** on (Figure 12.2), you can click on each cell as you go along and choose the response from a drop-down list, as long as you have set up the values in Variable View (see paragraph 12.2.2).

If you have participants who have said ‘Don’t know’ to a question, or have failed to answer it, the easiest option is probably to leave the answer blank in SPSS. (Of course, this doesn’t apply to people who have simply marked the midpoint of a scale, who should be given the number appropriate to that midpoint.) If (instead) you use a number for ‘don’t know’, you will need to declare that number as a missing value in Variable View. (For how missing data are dealt with, see paragraph 14.2.5.)

Don’t forget to allocate each participant an identification number and enter it into SPSS. Write the same number on the questionnaire. You will find this procedure indispensable if you have any problems later. (E.g. see section 14.2.)

*When you have entered the data, you should check it. See sections 14.1 and 14.2 for some help with this. If you are following this exercise in class, we will go to those sections before coming back to section 12.3.*

## 12.3 Calculating overall scores on a questionnaire

*The sample file used in this section is one which has been checked and data entry errors have been corrected, as described in sections 14.1 and 14.2.*

### 12.3.1 Introduction

If the questions constitute a scale, you will need to add up (or average) the answers to different questions. For example, if the questions are all about job satisfaction (as in the example), we might have to add them up to get a total score for job satisfaction. There may be just one total, or sometimes different questions have to be added up to make sub-scales.

If using a published questionnaire, make sure you find the instructions. These may be in a manual or in a journal article. They will give important information on how to calculate the overall score (e.g. whether it is a total or an average, whether there are subscales, whether there are any reverse-scored questions, whether certain questions need to be ignored for any reason). The same source will also tell you who the questionnaire was tested on (the so-called *norm group*) and what their mean score was; you may wish to compare your own participants' mean against the norm group.

These calculations will give us a useful introduction to calculations using **Transform – Calculate Variable**.

### 12.3.2 Reverse-scored questions: what they are

We will see how to add up scores in SPSS in a minute, but first you may have to deal with reverse-scored questions.

If the questions are about how happy people are in their job, a typical question might ask people how much they agree with the statement “I enjoy coming to work.” Obviously the more people agree with this statement, the happier they are in their job. The people who most strongly agree with the statement get the highest score.

However there might be some questions such as “If I could give up work tomorrow, I would.” In this case, the more people agree with the statement, the *less* they are happy in their job. If the questionnaire has been devised and published by someone else, they should have made it clear if there are any such questions.

### 12.3.3 Reverse-scored questions: How to deal with them

If you entered the data with the data labels on (see section 12.2.3) you may have already taken account of the reverse-scoring, and given a score of 1 to the people who most agree with the statement.

If not, the score needs to be amended so that low scores are changed into high scores, and vice versa.

To keep an audit trail and prevent mistakes, it is advisable to keep the old variable as it is and to create a new variable with a new name, such as Q3rev for a reversed version of Q3.

Suppose the response scale runs from 1 to 7. We would want 1 (the lowest possible score) to change into 7 (the highest possible score). Similarly we want to change 2 into 6, 3 into 5, 4 to stay as 4, 5 into 3, 6 into 2, 7 into 1. One way to do this is to use **Transform – Recode into different variables** on the drop-down menu. (The detailed procedures for this method are not covered here.)

But there is an easier way in this situation (i.e. where there are scores going from 1 up to a maximum possible score). All we need to do is to add one to the highest number it is possible to score on the question, and then subtract each person's score from that number (see Table 12.1 for why this works). In the example data, the highest possible score is 7, so we need to subtract everyone's score from 8. Go to **Transform – Compute Variable**. Under **Target Variable** put Q3rev. Under **Numeric Expression** put 8-Q3. Click on **OK**. SPSS will do the calculation for each case in the file. You will probably also want to change the number of decimal places to 0. It is always a good idea to examine Data View and check that the calculation is correct for a few example cases. If you are following this document as a tutorial, do the calculation for Q3 and check that the first few scores have been correctly reversed<sup>51</sup>.

**Table 12.1.** Illustration of reverse-scoring on a scale from 1 to 7.

Participant's score	1	2	3	4	5	6	7
Reverse score	7	6	5	4	3	2	1
Total of score and reverse score	8	8	8	8	8	8	8

If the scores go from 0 up to the largest possible number, the procedure is only slightly different. Just subtract everyone's score from the highest possible number (see Table 12.2 for an explanation). So if the scale for Q3 had gone from 0 to 4, you would go to Transform – Compute Variable. Under Target Variable put the name you want for the reversed score (e.g. Q3rev). Under **Numeric Expression** put 4-Q3. Click on **OK**. Again, go to Data View and check the calculation for a few example cases. (But this does *not* apply to the example data!)

**Table 12.2.** Illustration of reverse-scoring on a scale from 0 to 4.

Participant's score	0	1	2	3	4
Reverse score	4	3	2	1	0
Total of score and reverse score	4	4	4	4	4

<sup>51</sup> The first few lines of Q3 are 3, 5, 5 and 2, so the correct values for Q3rev are 5, 3, 3, and 6.



### 12.3.4 Adding up scores.

Once we have checked all our data, and reversed any scores as necessary, we can add them up to get a total.

Once again, we do this using **Transform – Compute Variable**. Under **Target Variable** put the name you want for the total, e.g. *JobSatTotal*. Under **Numeric Expression**, put a formula to add up the scores, remembering to include only the ones you want, and using the reverse-scored ones as necessary. For the example file we would have

$$Q1 + Q2 + Q3rev + Q4 + Q5 + Q6$$

Enter this and click **OK**. SPSS adds the new variable at the end of the file. You may want to check that you have the correct answer for a few cases. For the example file, the first few totals should be 32, 14, and 21.

If a participant has missing data (has failed to answer any question), their score for the total will also be given as a missing value. This follows the general rule that in any analysis, SPSS only includes the cases that have values for all of the variables included in the analysis. (Section 14.2.6 refers.)

### 12.3.5 Mean scores

Sometimes, you want to calculate the mean score instead of the total. You can also do this using **Transform – Compute Variable**. Under **Target variable** put the name you want for the mean, e.g. *JobSatMean*. List all the questions as before, but put brackets round them. Then put a slash, followed by the number of questions, showing SPSS it should divide by that number. For the example file we would have

$$(Q1 + Q2 + Q3rev + Q4 + Q5 + Q6)/6$$

For the example file, the correct first few means are 5.33, 2.33 and 3.50.

Again, if a participant has missing data (has failed to answer any question), their score for the mean will also be given as a missing value<sup>52</sup>.

## 12.4 Your own scales: a very brief introduction

There is a vast literature on how to create scales, and if you plan to attempt this it would be wise to consult an appropriate textbook. But here is a very brief introduction to some of the procedures in SPSS. You may also want to run

---

<sup>52</sup> Instead, you could use a function in SPSS called Mean. However, if a participant has failed to answer some questions this procedure will give the mean score for the questions they did answer, which may not be valid. If you consider doing this, take further advice.

through some of these procedures if you want to check whether the scale works as consistently with your participants as it did with the original norm group.

Remember that these procedures are only appropriate for scales where you want to add up questions to get a total, because all the questions relate to the same thing.

Before starting, you should reverse-score any questions as necessary (see above).

#### 12.4.1 Checking for problematic questions

When you have given your questionnaire to a sample of people, you can check for questions which might cause a problem. There are three ways in which questions might stand out as being problematic. The first two are matters of opinion – their correlations and standard deviations, which are covered in this section. The third is to look at how they affect Cronbach's alpha (see paragraph 12.4.2).

Firstly, you can look at the correlations between questions. Go to **Analyse – Correlate – Bivariate**. Put all the relevant questions into the **Variables** box and click **OK**. Remember that if you have reverse-scored any questions, it is the reverse-scored version you want to use.

The output for the example file is shown in

		Correlations					
		Q1	Q2	Q3rev	Q4	Q5	Q6
Q1	Pearson Correlation	1	.592**	.537**	.541**	.489**	.666**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	58	58	58	58	58	58
Q2	Pearson Correlation	.592**	1	.644**	.578**	.496**	.505**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	58	58	58	58	58	58
Q3rev	Pearson Correlation	.537**	.644**	1	.613**	.569**	.477**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	58	58	58	58	58	58
Q4	Pearson Correlation	.541**	.578**	.613**	1	.482**	.458**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	58	58	58	58	58	58
Q5	Pearson Correlation	.489**	.496**	.569**	.482**	1	.417**
	Sig. (2-tailed)	.000	.000	.000	.000		.001
	N	58	58	58	58	58	58
Q6	Pearson Correlation	.666**	.505**	.477**	.458**	.417**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.001	
	N	58	58	58	58	58	58

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

Figure 12.3. For example, this shows that the correlation coefficient between questions 1 and 2 is .592, that this is highly significant ( $p < .001$ ) and that 58 people are included in that calculation. It may be appropriate to tidy up this table and put it in your Results section.

If one or more of the questions has a particularly low correlation with the others, it suggests that particular question is not getting at the same concept as the other questions. (You might think it is, but perhaps your participants are interpreting the question differently from you.) Read over the question and consider eliminating it from the analysis. Or, if some questions correlate with each other but not with the rest, it may indicate that your questions are tapping into two or more sub-scales, and you might want to consider a factor analysis (not covered on this course).

If one of the questions is negatively correlated with the others, it would appear that you forgot to reverse-score it; or that your respondents are interpreting the question very differently from the way you expected. If the latter, you may need to eliminate it.

Another thing you could consider looking at is the standard deviations (under **Analyse – Descriptive Statistics – Descriptives**). If one of the questions has a much smaller standard deviation than the others, it appears that there is very little

difference between participants as to how they answer that question, so perhaps it is not telling you anything. Consider whether it is worth keeping.

Of course, if you discard questions for any reason this is an important part of your findings and should be reported in your Results section.

		Correlations					
		Q1	Q2	Q3rev	Q4	Q5	Q6
Q1	Pearson Correlation	1	.592**	.537**	.541**	.489**	.666**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	58	58	58	58	58	58
Q2	Pearson Correlation	.592**	1	.644**	.578**	.496**	.505**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	58	58	58	58	58	58
Q3rev	Pearson Correlation	.537**	.644**	1	.613**	.569**	.477**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	58	58	58	58	58	58
Q4	Pearson Correlation	.541**	.578**	.613**	1	.482**	.458**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	58	58	58	58	58	58
Q5	Pearson Correlation	.489**	.496**	.569**	.482**	1	.417**
	Sig. (2-tailed)	.000	.000	.000	.000		.001
	N	58	58	58	58	58	58
Q6	Pearson Correlation	.666**	.505**	.477**	.458**	.417**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.001	
	N	58	58	58	58	58	58

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

**Figure 12.3.** Correlations for example file.

#### 12.4.2 Cronbach's alpha: how to calculate it

Cronbach's alpha is a measure of how much the questions measure the same thing (i.e. how much the questions as a whole correlate with each other). To call it up, go to **Analyse – Scale – Reliability Analysis**. Under Items, put all the questions concerned. (Again, if you have any reverse-scored questions the reverse-scored questions are the ones to use.) Click OK.

Cronbach's alpha is given in the output. In the example file, Cronbach's alpha for Q1, Q2, Q3rev, Q4, Q5 and Q6 is .874 (Figure 12.4).

Reliability Statistics	
Cronbach's Alpha	N of Items
.874	6

Figure 12.4. Cronbach's alpha output.

There is no hard-and-fast rule on what is an acceptable level for Cronbach's alpha. Most people would find a figure of above .8 good, and a figure above .7 acceptable. Some people (but not everyone) consider that it is possible for alpha to be too high, and say that if it is above .9 then the scale contains too many items that are just the same as each other, and is wasteful<sup>53</sup>.

If Cronbach's alpha is too low, you might consider excluding problematic questions (see sections 12.4.1 and 12.4.3). Or it might be appropriate to carry out a factor analysis to create two or more separate scales; each of these is likely to have a higher Cronbach's alpha than the overall scale.

### 12.4.3 How Cronbach's alpha is affected by individual questions

Usually, the more questions that make up a scale, the higher Cronbach's alpha becomes. It is possible to test whether this is true for your scale. When you carry out the procedure in paragraph 12.4.2, click on **Statistics** and tick the box (under **Descriptives**) for **Scale if Item Deleted**. This will bring up the output shown in Figure 12.5. See the final column of the table. In the case of our example data, Cronbach's alpha is indeed reduced if any of the items is deleted. If deleting any question increases Cronbach's alpha, this suggests that it is not measuring the same concept as the others. If there is one such item, perhaps it should be deleted (see paragraph 12.4.1); if there are several such items, it might be appropriate to consider a factor analysis.

Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Q1	21.0000	44.491	.718	.845
Q2	20.9828	45.807	.715	.847
Q3rev	21.0517	44.541	.721	.845
Q4	21.0345	45.472	.671	.853
Q5	20.9310	46.662	.609	.864
Q6	21.1207	45.968	.629	.861

Figure 12.5. Output from *Scale if Item Deleted*.

<sup>53</sup> For further discussion, see *Psychological Testing* (1982) by Kaplan and Saccuzzo, or other books on psychometrics.

## 13 Operations on the data file

### 13.1 Overview: what SPSS can do

Various kinds of calculations can be done on the SPSS data file. SPSS is not as versatile as Excel, but it can do most of the operations you are likely to want to do in preparation for statistical analysis. It is also possible to do calculations on selected cases, rather than the whole file. Some of the most useful facilities are included in this section.

Some other operations we will not cover, but can be found in other books, or in the SPSS Help file, include:

- Recoding variables. For example, suppose you had collected information on your participants' race, in 10 categories. Some of those categories may contain only a few participants. You could collapse these into fewer categories, e.g. *White British*, *White Irish* and *White Other* into *White*.
- Combining two data files into one to add extra information about the same participants. (N.B. both files must use the same numbering system to give each participant a unique number, and both must be sorted in order of that number.)
- Combining two data files into one to add new participants. (N.B. Each file must use different numbers for the participants!)

### 13.2 Calculating z scores

Reminder: a z score is how many standard deviations an individual score is from the mean. So if the mean IQ is 100, and the standard deviation is 15, someone with an IQ of 115 has a z-score of +1.

In section 12.3.4 we calculated a Job Satisfaction score (*JobSatTotal*) for each of our participants. The mean in the sample is 25.2, and the standard deviation is 8.0. However, we do not need to use these figures or a laborious calculation to find out each participant's z-score.

Open the file (the worked version, after screening and calculation of totals). Remember to exclude the cases we flagged up for exclusion, using **Data – Select Cases** (section 14.1.2).

There is a quick way of calculating z-scores in SPSS. Click on **Analyse – Descriptive Statistics – Descriptives**. Move the variable(s) of interest (*JobSatTotal*) into the **Variables** box. (You can call up any descriptive statistics you want at the same time as getting the z-scores. In fact, SPSS will not let you turn them all off. For this exercise, leave the default settings, including the means and standard deviations). However, the important thing for our present

purpose is to tick the box that says **Save Standardised values as variables** (Figure 13.1). Press **OK**.

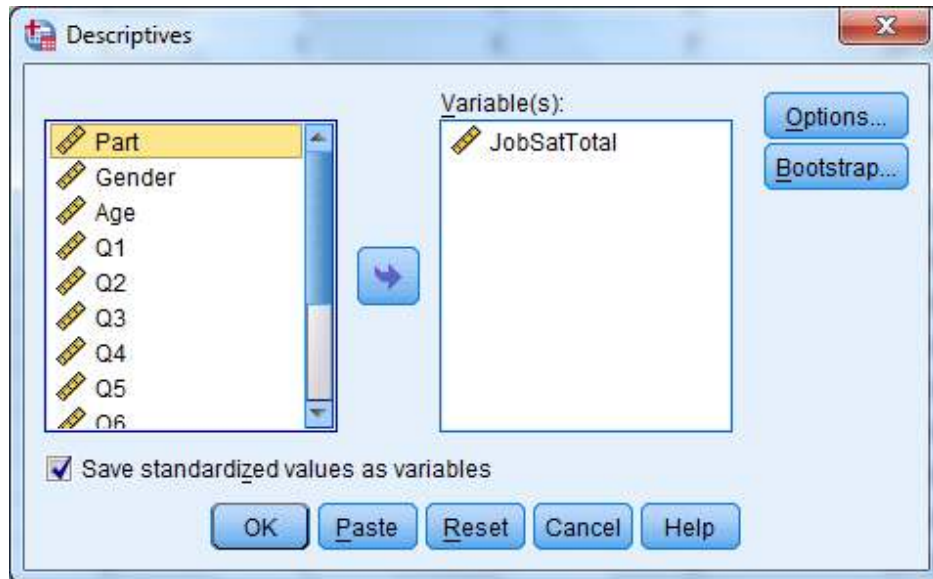


Figure 13.1. Descriptives dialogue box with 'Save standardised values' selected.

Whatever descriptive statistics we asked for are on the output file. More importantly, the z-scores have been saved back to the data file. They have the same name as the original variables, but with a z in front. You may like to check one of them to verify that SPSS has calculated it correctly.

### 13.3 Calculations on one or more variables, using Compute Variable

Using **Transform – Compute Variable**, SPSS can do all sorts of calculations, using one variable or several (or even none). It can change an existing variable, but as an audit trail it is better to create a new variable with a new name.

There are many possible reasons for doing calculations. We may simply want to add up scores on a questionnaire, or participant's total scores adding up two parts of a test. For some practical examples of this, see section 12.3. Section 12.3 also gives more detailed instructions on how to use this facility.

Another possible circumstance arises with data screening (see section 14.4). This is an instance where we may want to do a non-linear transformation, i.e. we create a new variable which is based on the original but does not correlate with it 100%.

## 13.4 Creating a case number (e.g. participant number)

It is good practice to give each participant or case a unique number, especially if we are going to be manipulating the file.

If you do not already have a variable that identifies each case, you can create one that simply corresponds to the line number in SPSS (at the time you create it). Go to **Transform – Compute Variable**. In **Target Variable** put the name you want your variable to have (e.g. *PartNo*), and in **Numeric Expression** simply type *\$casenum*.

## 13.5 Categorising data (e.g. pass/fail, high/low)

### 13.5.1 Predefined split point(s)

In section 12.3.4 we calculated a Job Satisfaction score (*JobSatTotal*) for each of our participants. Suppose that the questionnaire came with a manual that defined a score of over 30 as ‘highly motivated’. We want to classify the participants as highly motivated – yes or no. (Perhaps we want to use these categories in an Anova, or we just want to count how many ‘highly motivated’ respondents are in our sample.)

Remember that we have excluded some cases from analysis, and if you are re-opening the file you will need to tell SPSS to exclude them. (See section 14.1.2.)

Let us start by giving the highly motivated participants a value of 1. Click on **Transform – Compute Variable** and put a suitable name (e.g. in this case, *Highly\_Motivated*) in **Target Variable**. Type 1 in **Numeric Expression**. . Click on “If...” at the bottom left, and a new dialogue box comes up. Click on “**Include if case satisfies condition**” and enter a formula that defines our category. In this case enter *JobSatTotal* > 30, meaning *JobSatTotal* is greater than 30. Click **Continue** and **OK**.

Now let us give the other participants a value of 0. Click again on **Transform – Compute Variable** and leave the Target Variable as *Highly\_Motivated*. Type 0 in Numeric Expression. . Click on “If...” and change the rule to *JobSatTotal* <= 30, meaning *JobSatTotal* is less than or equal to 30. Click **Continue** and **OK**.

If you are doing this exercise and want to check your work, click on **Analyse – Descriptive Statistics – Frequencies** and put *Highly\_Motivated* into the **Variable(s)** box. You should find that there are 36 participants with a value of 0 and 22 with a value of 1.

You can of course go on to create even more categories if you want to. When you have finished, you will probably want to tidy up the variable in Variable View, for example giving names to the categories under Values.



**Warning.** SPSS will remember the **If...** condition for as long as the file is open. If you do other calculations you probably do not want to restrict them to these cases. Click on **If...** and re-check **Include all cases**.

### 13.5.2 Splitting into equal groups: e.g. median splits

Sometimes we want to split a variable up into two, but there is no specific pass mark. In particular, it is often useful to split a variable up into equal-sized groups of high and low scores. This is known as a median split.

To do this, go to **Transform – Rank Cases**. Put the variable of interest (e.g. *JobSatTotal*) into the **Variables** box. Click on **Rank Types** and uncheck **Rank**. Check **Ntiles** and change the figure to 2. Click **Continue** and **OK**. The new variable is automatically added to the file, with the name *NJobSatT* (i.e. *N* followed by the beginning of the name of the original variable). It also inserts a **Label** in Variable View. You may want to tidy all of this up.

You can use a similar procedure to split a variable into any other number of equally sized groups— just put the number of groups you want in place of 2.

## 13.6 Working with part of the data file

### 13.6.1 Selecting participants/ cases

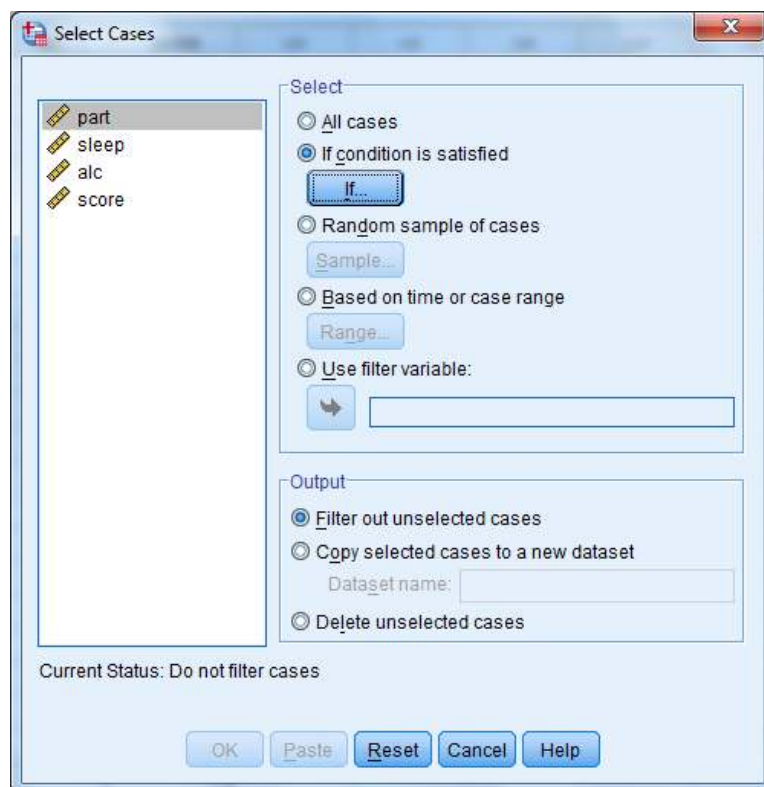
You might want to do an analysis<sup>54</sup> on only part of the file. For example, to do post hoc tests on the file used in section 6.6 of this booklet, you might want to carry out a t-test for just the participants who had sleep, to see whether alcohol made a difference to their score.

You can do this by clicking on **Data – Select Cases**. In the dialogue box that comes up (Figure 13.2). Select **If Condition is Satisfied** and click on the **If** button. A new dialogue box comes up. Enter the condition that describes what cases you want to describe – for example, in this instance, as shown in Figure 13.3. Now if you call up a statistical test, it will only be applied to the cases you selected.

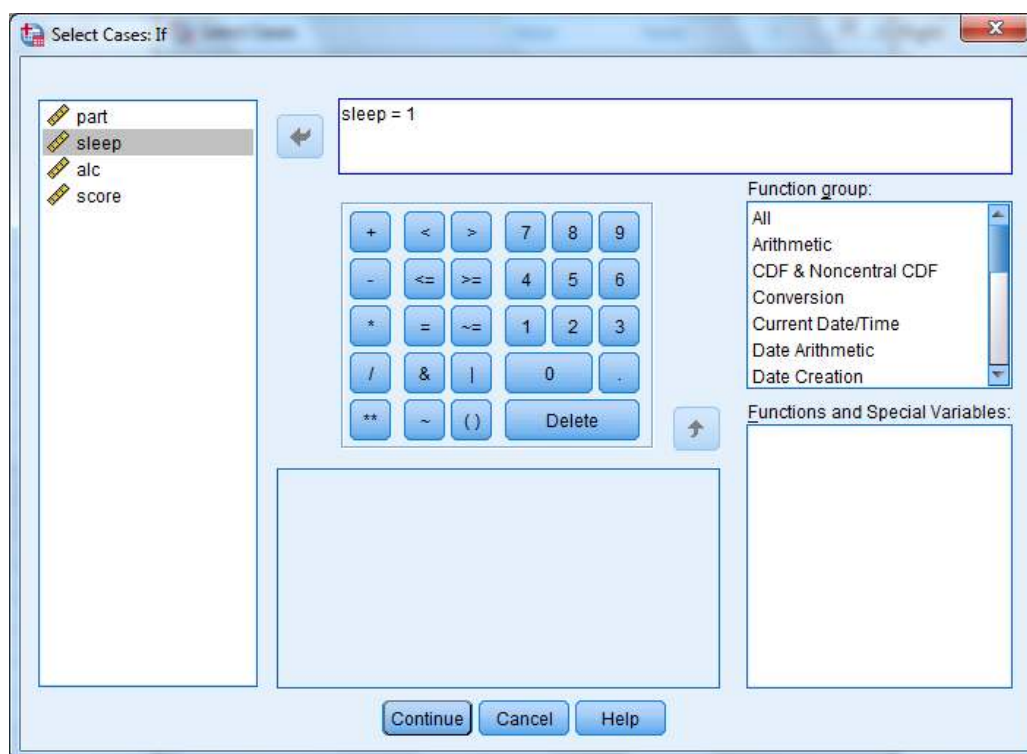
(For example, to do the t-test described above, we call it up: **Analyse – Compare Means – Independent Samples t-test**. The **Test Variable** is *Score* and the **Grouping Variable** is *alc*; click on **Define Groups** and the two groups are 1 and 2. You should find that the p-value is .023, which becomes .046 after a Bonferroni correction for 2 comparisons.

---

<sup>54</sup> Note that this procedure does not work for **Transform – Compute Variable** (paragraph 13.3). This has its own **If** button.



**Figure 13.2. Select Cases Dialogue Box.**



**Figure 13.3. Select Cases – If Dialogue box.**

(If you need a reminder of the full details on how to call up and report an independent-samples t-test, see section 5.9.)

Remember to turn off Select Cases when you have finished with it! Go back to **Data – Select Cases** and select the top radio button, **All Cases**.

### 13.6.2 Combining conditions, e.g. when you already have an *Include* variable.

You can specify a combination of conditions. For example, suppose you already had an *Include* variable (see paragraph 14.1.2), and that you want to include only cases where *Include* is 1, as well as the rule we have just specified (*sleep* = 1). You can join up the conditions using **&**, which is a 'logical and'. So our combined condition would be:

*sleep* = 1 & *include* = 1

### 13.6.3 Splitting the file with Data – Split File

Suppose the file includes a categorical variable (e.g. *sleep*) and you want to split the file by that variable<sup>55</sup>. For example, to do post hoc tests on the file used in section 6.6 of this booklet, you might want to carry out a t-test for just the participants who had sleep, to see whether alcohol made a difference to their score; and also for the participants who had not had sleep, to see if it made a difference to their score. You can do this by splitting the file.

Actually, we have already seen how to split a file, when we wanted to get descriptive statistics for individual levels (categories) of a categorical variable (see section 5.8.2). The procedure is repeated here for ease of reference.

Click on **Data – Split File**. A dialogue box comes up. Click on **Organise output by groups** and move the variable you want to split by (e.g. *sleep*) into the **Groups Based on** box. The dialogue box should look like Figure 13.4.

---

<sup>55</sup> Note that this procedure does not work for **Transform – Compute Variable** (paragraph 13.3). This has its own **If** button.

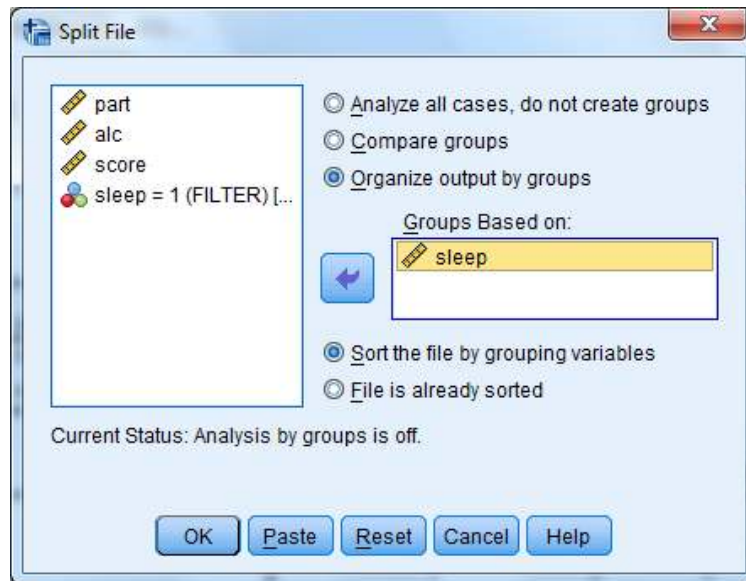


Figure 13.4. Split file dialogue box.

Now if you call up an analysis it will be done for each level of the categorical variable. For our current example it will be done twice: once for *sleep* = 1, and once for *sleep* = 2.

(So suppose we were doing the post hoc analysis for section 6.6. We split the file by *sleep*, then call up an independent-samples t-test: **Analyse – Compare Means – Independent Samples t-test**. The **Test Variable** is *Score* and the **Grouping Variable** is *alc*; click on **Define Groups** and the two groups are 1 and 2. You will get two results.

- With *sleep*, the p-value is .023 (.046 after a Bonferroni correction for two comparisons). So for participants who slept, there is a significant difference depending on whether they did or did not consume alcohol.
- Without *sleep*, the p-value is < .001 (<.002 after a Bonferroni correction for two comparisons). So for participants who did not sleep, there was also a significant difference in their score depending on whether they did or did not consume alcohol.

If you need a reminder of the full details on how to call up and report an independent-samples t-test, see section 5.9.)

Remember to turn off Split File when you have finished with it! Go back to **Data – Split File** and reinstate **Analyse all cases, do not create groups**.

### 13.6.4 Sorting the file

Sometimes it is helpful to sort the file into a different order. For example, with a large file the easiest way to find the lowest and highest values of a given variable is to sort the file in order of that variable.

First of all, make sure that you have a *Participants* variable<sup>56</sup> so you can re-sort back to the original order if you want to! Then go to **Data – Sort Cases** and put the variable concerned into the **Variable(s)** box. (Although this has space for more than one variable, only use one at a time.) Click **OK**. The file will be rearranged in order of that variable (with missing values at the top).

## 14 Checking and screening data

*In class, we will cover the first part of this chapter using the example questionnaire data (see section 12.2.1). We will cover the rest of this chapter in a later lecture, using two other examples.*

### 14.1 File control and excluding participants/ cases

#### 14.1.1 Different versions of your file

You may find that you end up with more than one version of your data file. For example, if you correct mistakes or add more participants you might still want a copy of the previous version of the file. The best way to keep control of different versions is to add a version number to the end of the file name, e.g. “Questionnaire data version 1.sav” etc.<sup>57</sup>.

#### 14.1.2 Excluding participants/ cases; use of an *include* variable

You may need to exclude the data from one or more participants (or cases in general – remember that statistics do not have to be about people). For example, when analysing questionnaires, you might want to exclude people who did not answer all your questions<sup>58</sup>. You could literally delete those lines and save the file as a new version. However, a more sophisticated way is to create an extra variable to show which participants (or cases, as SPSS calls them) are to be included and which are to be excluded. This makes it easier to keep track of changes, and to change your mind if you want to.

---

<sup>56</sup> If you need to create one specially, see paragraph 13.4

<sup>57</sup> This may seem obvious, but I have seen students who have files with names like ‘final final final’ or ‘real deal’. An increasing version number is much easier! Incidentally, if you prefer to use the date as part of your file name, if you use “international format” (Date, month, day, e.g. 2016-06-03 for 3 June 2016) the dates are in correct order by size.

<sup>58</sup> Remember that if you exclude participants from the analysis, you should report this in your write-up, especially if this is likely to affect how representative your sample is.

It is easy to set this up (although it does need careful management afterwards). On the drop-down menu, go to **Transform – Compute Variable**. Under **Target Variable** enter *Include* and under **Numeric Expression** enter 1. Click **OK** to create the variable. Go to Variable View and set up **Value Labels** to show that 1 means *Include* and 0 means *Exclude*. Also, change the number of decimals to 0. If you are following this as a tutorial, set this up on the example file.

So for now, of course, the variable shows all participants as being included. Later on, you can change it to 0 for any cases you want to exclude from analysis.

**If you use this method, beware! You will have to manually select these participants before doing any analysis, whenever you exclude any more cases, and again every time you open the file.** To do this, go to **Data – Select Cases**. Click on **If condition is satisfied** and the **If** button. In the dialogue box that comes up, type '*Include* = 1'. (Remember, here you are specifying which cases to *include*, not which ones to exclude.) Click on **Continue** and **OK**. In Data View, check that SPSS has put a cross through the line numbers of the cases which are to be excluded. (If you are trying this straight away, of course there are no such cases.)

## 14.2 Checking the data file; missing data

### 14.2.1 Check your data entry

Always check over your data to ensure that you have entered it correctly. However, if you have a lot of data, it is easy to miss mistakes. The following section gives methods for finding some of the worst errors.

### 14.2.2 Missing and illegal data – definitions

*Missing data:* If any of the data have no value against them at all, they will be shown in the data file as a dot. This may be because the participant really did not answer, but it may be that you made a mistake in data entry.

*Illegal data:* If you have made a mistake in data entry, some of the entries may be impossible. For example, if one of the variables is *age*, nobody can have an age of -1. Such an entry is known as 'illegal' data.

### 14.2.3 Detecting missing and illegal data – categorical variables.

One way to detect such problems is as follows. Click on **Analyse – Descriptive Statistics – Frequencies**. Move into the **Variable(s)** box your categorical variables. If you are using the example questionnaire data, the only categorical variable is *Gender*. In the main dialogue box, make sure that **Display Frequency Tables** is ticked. Click on **Statistics** and ask for **Minimum** and **Maximum**. Click on **Continue** and **OK**.

The output is shown in Figure 14.1 and Figure 14.2. Figure 14.1 shows us whether there are any missing values. (There are none in this case.) It also allows us to check for illegal values. In this case we can see that something is wrong, since the maximum is 22, when we know that the only values that are allowed for this variable are 1 and 2. Figure 14.2 clarifies this problem: there is one case for which the value has been entered as 22. We will correct this shortly (sections 14.2.5 and 14.2.6), but first we will see if there are any problems with the continuous variables.

Statistics		
Gender		
N	Valid	60
	Missing	0
Minimum		1
Maximum		22

Figure 14.1. Output for categorical variable(s) - part 1

Gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	29	48.3	48.3	48.3
	Female	30	50.0	50.0	98.3
	22	1	1.7	1.7	100.0
	Total	60	100.0	100.0	

Figure 14.2. Output for categorical variable(s) - part 2

#### 14.2.4 Detecting missing and illegal data – continuous variables.

One way to detect these problems is as follows. Click on **Analyse – Descriptive Statistics – Descriptives**. Put the continuous variables into the **Variable(s)** box. (In this case they are *Part*, *Age*, and *Q1* to *Q6*<sup>59</sup>.) Under **Options**, ensure that **Minimum** and **Maximum** are selected. Click **Continue** and **OK**.

The output is shown in Figure 14.3. You can spot missing values by looking at the column headed *N*, which shows the number of entries for that variable. If any are different from the others (or from the number of cases in your file), that indicates missing data. If you are using the example Questionnaire data, there are missing values for *Q1* and *Q5*.

<sup>59</sup> Actually we could have chosen to treat *Q1* to *Q6* as categorical variables, as they only had 7 possible values. For this purpose it doesn't really matter.



Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Part	60	1	60	30.50	17.464
Age	60	18	56	38.07	12.322
Q1	59	1	7	4.20	1.710
Q2	60	1	7	4.25	1.601
Q3	60	1	55	4.65	6.824
Q4	60	1	7	4.18	1.692
Q5	59	1	7	4.31	1.704
Q6	60	1	7	4.08	1.740
Valid N (listwise)	58				

Figure 14.3. Output for continuous variable(s).

You can also look at the Minimum and Maximum to see if any of those are implausible. Those for *Part* (participant no) and *Age* seem perfectly reasonable, as do those for most of the questions. However, for Q3 there is at least one case with a value of 55, which is not a possible value.

#### 14.2.5 Finding the case(s) with missing or illegal values

If it is too laborious to look through the file for the missing or illegal values, remember that you can change the order of the file so that any given variable is in order of size (with missing values at the top). First of all, make sure that you have a *Participants* variable<sup>60</sup> so you can re-sort back to the original order if you want to. Then go to **Data – Sort Cases** and put the variable concerned into the **Variable(s)** box. (Although this has space for more than one variable, only use one at a time.) Click **OK**.

#### 14.2.6 Dealing with missing data and illegal values

If possible, check the source of the original data to fill in any missing or wrong values. (If the source was questionnaires, I hope you followed my advice to write a participant number on each questionnaire and enter it into SPSS!)

If it is not possible to correct the data, the main options are:

- (a) *Out of range values*. You might decide to delete the case (either literally, or using the procedure in paragraph 14.1.2). An alternative, and the simplest option, is to delete the value (e.g. by selecting the value and back-spacing, so that SPSS shows a dot to indicate a missing value). This turns it into the next category:
- (b) *Missing values*. The simplest option is simply to carry on regardless. SPSS will exclude this case (this participant) from any analysis which

<sup>60</sup> If you need to create one specially, see paragraph 13.4



requires that value. Another option, if you want all your analyses to relate to a consistent set of participants, is to delete the case (either literally, or using the procedure in paragraph 14.1.2). If there is a variable for which you have a lot of missing data (e.g. a question that a lot of participants refused to answer) you might consider abandoning any analysis that would have used that variable.

Other options are beyond the scope of this course. If there are more than 5% of cases with missing values on variables you are using in your analysis, or if you are worried that the missing values will cause your sample to be seriously unrepresentative, you should consult a more advanced source<sup>61</sup>.

If you are following this document with the example Questionnaire file, we will take the following actions.

- Case 45, gender is 22, we will correct to 2.
- Case 53, Q3 is 55, we will correct it to 5.
- Cases 44 and 52 have missing values. We will change *Include* to 0 for these cases, and use the **Select Cases** procedure. See section 14.1.2 for how to do this.

*If you are following the Questionnaire exercise, return to section 12.3.*

## 14.3 Good practice: checking for assumptions

### 14.3.1 Overview; remember the checks we already do

There are certain checks we have already covered when doing tests, e.g.

- When doing t-tests and Anovas, you should carry out the tests of assumptions which are included in the write-ups in this booklet (e.g. Levene's test).
- When doing correlations and regressions you should look at scatterplots to check that the relationships between variables are linear (as best you can; in real life it is often hard to tell).

There are further things to check as good practice:

- If you are going to do a parametric test you should check for outliers in any continuous variables (section 14.3.2)
- Ideally, if doing a parametric test you should check whether continuous variables are normally distributed (section 14.3.3)
- There are various other checks you could do. These are beyond the scope of this course, but you may come across them in more advanced textbooks.

---

<sup>61</sup> For a definitive, albeit scary, treatment see Tabachnick and Fidell.

If you detect problems, make sure you have checked your data entry; double check if necessary. The other actions we will discuss are to delete outliers, or to carry out a transformation (paragraph 14.4). We will discuss this further in the lecture and the case studies.

Note that curing one problem will often cure others, so it is best to carry out all the checks you are going to do before you take any action. *Any action you take should be reported in your write-up.*

### 14.3.2 Checking for outliers

It is good practice to check whether your continuous variables contain any outliers. A conservative definition of an 'outlier' is any value of that variable which has a z-score below -3.29 or above +3.29<sup>62</sup>.

Before calculating these z-scores, you should split the file by any categorical variables that are to be used in your analysis (see paragraph 13.6.3 for how to split the file.) For example, if you were going to do a factorial Anova, you would split the file by the categorical variables in the Anova. Remember to un-split the file when you have finished!

Paragraph 13.2 tells you how to calculate z-scores. You can then sort the file by each z-score in turn (paragraph 13.6.4) to check for any high or low values.

### 14.3.3 Checking whether continuous variables are normally distributed

We will cover an introduction to this topic in the lecture and two case studies. We will use histograms. (We saw how to create these in section 4.2). Another common method is to look at boxplots<sup>63</sup>. Either involves some subjectivity.

When checking histograms for normality, just as when checking for outliers (paragraph 14.3.2) you should split the file by any categorical variables that are

---

<sup>62</sup> A less conservative approach (i.e. a more radical approach, likely to produce more outliers) would be to consider a z-score below -1.96 or above +1.96 to be an outlier. These figures (3.29 and 1.96) may sound arbitrary, but they correspond to a likelihood of .001, or .01, respectively of occurring, under the null hypothesis that the variable is normally distributed.

There are two other common ways of defining outliers. The first is to use a boxplot (Appendix D), which is also a fairly radical approach. The other is to examine a histogram (paragraph 4.2), but this is quite a subjective method.

As always, you should not take these different opinions as licence to play around with the data until you get the significant result you want. If in doubt, stick with my recommendation in the main text.

<sup>63</sup> This has the advantage that it can check for both normality and outliers at the same time. See footnote 62 for comments on outliers, and Appendix D for an explanation of boxplots.

to be used in your analysis<sup>64</sup>. Remember to un-split the file when you have finished.

## 14.4 What to do if there are problems

The options we will consider are:

### 14.4.1 A non-parametric test

Sometimes, you may be able to simply carry out a non-parametric test. Parametric tests do not require assumptions about outliers and/or non-normal distributions. (However, most statisticians prefer to make any necessary adjustments and to use parametric tests.)

### 14.4.2 Options relating to specific tests

If the only serious problem is a violation of the assumptions of a specific test, such as Levene's test, refer back to the advice given for the relevant test.

### 14.4.3 Deleting outliers

Sometimes it is appropriate to delete outliers (mentioning in your write-up what you have done, and how you defined outliers). However, outliers are often caused by non-normality, in which case it is usually preferable to do a transformation.

(To delete outliers, the easiest option is to literally delete the value from the data file. Another option is to remove the case (participant) from the analysis altogether. In either case I recommend that you keep the old version of the file. See paragraph 14.1 for further advice.)

If you delete outliers, you should check again with the reduced variable to ensure that the deletion has not created new outliers. If it has, this is likely to be a clue that a transformation would be more appropriate.

### 14.4.4 A non-linear transformation

Transforming a variable may help with one or more of the following problems. Often, the transformation will solve more than one problem at the same time.

- A skew in the variable. (This is often the underlying issue behind other problems)
- Outliers
- Non-linear relationships with other variables (thereby breaching assumptions of procedures such as correlation and regression)
- Breaches of other assumptions, such as Levene's test.

---

<sup>64</sup> Another method, which may be more convenient, is to use the options in the histogram dialogue box to "panel by" your categorical variable(s). Or if you use boxplots, these also provide options to separate them out.

Transformations are done using **Transform – Compute Variable** (see paragraph 13.3). The kind of transformation which could help here is called a *non-linear* transformation, i.e. it creates a new variable which does not correlate 100% with the old variable. Choosing an appropriate transformation is not an exact science; see Appendix F.

We will look in more detail at when a transformation might be appropriate, and how to choose one, in the lecture and our case studies.

After doing a transformation, you should check your new variable to ensure that the transformation was effective, i.e. that the histogram of the new variable is now an acceptable shape (remembering that it will never be perfect) and that you have eliminated outliers. (If the shape is otherwise good but there are still outliers, you could delete them at this stage.)

Now you can carry out your test, with the transformed variable in place of the original one. Remember to mention in your write-up what you did. When interpreting the results, remember that your findings are about the transformed variable(s), so you need to be careful in the conclusions you come to about the original variable(s).

## Appendices

### A. Reporting results

#### What to include when reporting an inferential test

You should always:

- Say whether the result is statistically significant or not. Usually we regard a result as statistically significant if  $p$  is less than .05. (You may find it easiest to think of this as .050, so it has the same number of decimal places as the SPSS output.) Some researchers/journals report  $p$  values between .05 and .10 as a 'trend', implying that they think there may have been an effect but that their study was not quite powerful enough to find it.
- Give the results of the statistical test that arrived at that judgement, e.g.  $t(9) = 3.19$ ,  $p = .011$ .
- Include appropriate descriptive statistics.

#### Reporting in APA Style

Styles of reporting statistics differ, even between journals in the same discipline. However, APA (American Psychological Association) style is used by many journals, and many others are similar. We use APA style on the course. The examples in this booklet are all reported in APA style.

Here are some of its most important features:

- To report the significance level,<sup>65</sup> write  $p =$  followed by the figure given by SPSS as the "Sig." level – see the descriptions of the individual tests for where this comes from in each case. The exception is if SPSS gives the "Sig." as .000. In that case we write  $p < .001$  ( $p$  is less than .001).
- if a Roman letter (e.g.  $t$ ,  $F$ ,  $p$ ,  $r$ ) is used for a statistic, it is printed in italics.
- $p$  is reported to three decimal places<sup>66</sup>.
- $F$  and  $t$  are reported to two decimal places.
- Practice differs on decimal places for  $r$  and rho; for this course please report them to three decimal places.
- degrees of freedom are put in brackets after the statistic. For example  $t(3) = 3.12$ . See write-ups of individual tests for more details.
- if a statistic cannot take a value higher than 1 (e.g.  $p$ ,  $r$ ), the 0 before the decimal point is omitted (e.g.  $p = .011$ )

---

<sup>65</sup> You may sometimes see that authors have reported simply  $p > .05$  ( $p$  is greater than .05) if the result is not significant, and  $p < .05$  ( $p$  is less than .05) if it is significant. Often, they may say  $p < .01$  and  $p < .001$  if appropriate. This is an old fashioned format which is no longer acceptable in APA style, except to save space in tables.

<sup>66</sup> Actually, the APA Publication Manual says that  $p$ -values may be given to 2 or 3 decimal places; this is an editorial decision for individual journals. Most journals use 3 figures, and 3 are expected on this course.

- $N$  or  $n$  represent the number of participants. In strict APA style,  $N$  means the total number of participants in the study and  $n$  means the number in a subgroup, but this often seems to be ignored.

### Formatting hints in Word

To insert a Greek letter (e.g.  $\chi$ , which looks a lot different from a Roman letter in some fonts, e.g. Times Roman) go to the Insert tab, and Symbol. Click on More Symbols (unless it is in the quick list because you have used it recently). In the dialogue box, choose '(normal text)' in the Font drop-down box (top left), then scroll down until you find the letter you want.

To insert a superscript (e.g. the <sup>2</sup> in  $\chi^2$ ), in the Home tab, highlight the text and click the appropriate icon under Font.

### Rounding numbers

It is often sensible to report figures to fewer decimal places than are given by SPSS or your computer. For example, when reporting a mean it is usually only meaningful to report one more decimal place than there was in the original data.

*How you do it.*

Take the number you wish to round (e.g. 2.361) and decide how many decimal places you wish to report (e.g. one). Cross out all the figures after that (in this case after the 3; 2.361). If the first crossed-out figure is 0, 1, 2, 3, or 4) leave the result unchanged. If the first crossed-out figure is 5, 6, 7, 8, or 9, add one to the last uncrossed-out figure. So here, the rounded result is 2.4.

*Why you do it.*

Readers are likely to think that the number of decimal places reflects how confident you are in your result. Suppose Anna and Bob give some children a test. Anna reports that her children's mean score is 5. Bob reports that his children's mean is 5.00.

Bob's score sounds more accurate than Anna's. This is because Anna's if children had a mean score anywhere between 4.5000 and 5.4999, she would still have reported it as 5 (as a round number). In other words, there could have been a range of 1 in the mean score and Anna would have reported it the same way. Bob's children would need an average between 4.9950 and 5.0050 for him to report it as 5.00 (to two decimal places). In other words there is only a range of 0.01 in the mean score for which Bob can have legitimately given the score he did.

So if the calculator (or SPSS) gives a mean score of 2.361, why not report it as 2.361? The reason is that we are not usually calculating numbers for the fun of it: we are using them to represent something. When we report the children's mean score, we are suggesting that this is our best estimate of something, for example what the likely score would be of other children who had the same learning experience. If we report it as 2.361 this suggests that we would expect other children to have a very similar score. This is known as *spurious accuracy*.

## B. Copying graphs and other objects into Word (or other applications)

### Converting charts and graphs to black and white

Most charts and graphs provided by SPSS and Excel are in colour, which may not be appropriate for published work. To change them to black and white, see section 3.3.2 or 4.3 as appropriate.

### Straightforward copying

Often, copying something from one programme to another is as simple as this:

- Click on it in the original program and ensure it is selected (sometimes this means it changes appearance in some way)
- Select **Edit – Copy** from the drop-down menu (or enter **Control-C**)
- Open the programme you want to move it to, and ensure that the cursor is at the point you want the object to appear
- Click on **Edit – Paste** (or enter **Control-V**).

### If there are problems with copying

If this does not work, or if the object does not behave itself in the new programme, read on.

To change the way that an object is copied, especially how much of its appearance it retains from the original, try one or more of the following.

- See if the application you are copying from has any other way of copying. For example,
  - to copy a graph from SPSS, you can open the Chart Editor first by double-clicking the chart, then select **Edit - Copy Chart**
  - Depending on the version of SPSS you are using, there may be an option such as **Edit – Copy Objects** or **Edit – Copy Special**.
- In the application you are copying to, try **Edit – Paste Special** instead of **Edit – Paste**. Experiment with all the options until you get the one you want. (For example, if you want a table in Word 2002 to look just like it did in SPSS, try Edit – Paste Special – Files.) Or in Word 2000, try Paste Special – Enhanced Metafile.

- If you are desperate, you can sometimes achieve your objective by copying into a different application, and then from there into the final destination. For example, you might copy from SPSS into Excel or PowerPoint, then copy again from there into Word.

Sometimes, it is appropriate to copy all of part of what is showing on the screen (a “screenshot”). Hit **PrtScrn** to copy the whole screen, or **Alt-PrtScrn** to copy the active window. The result is put onto the clipboard and can be pasted direct into an application such as Word, or edited using Paint (under **Programmes – Accessories**).

### Avoiding problems with objects moving around

Having copied any object into Word, it is advisable to right-click on it, click on **Format Objects** on the drop-down menu, click on the **Layout** tab, and under **Wrapping style** select **In line with text**. (This varies slightly depending on the version of Word you are using.) This ensures that it remains exactly where you placed it in relation to the text on the page.

## C. Help in SPSS

SPSS has the usual Help facilities. It also has (under Help) tutorials and case studies.

There are also quite substantial screen tips in many places. For example, on the output for a chi-square test, double-click on the **Chi-Square Tests** table so that it is surrounded by a shaded box. Click once on **Pearson Chi-square**, right-click and a drop-down menu appears. Click on **What’s this?** and an explanation appears.

Similarly, when using a dialogue box (e.g. **Crosstabs**), right-click on a part of it (e.g. **Row(s)**) and an explanation appears.

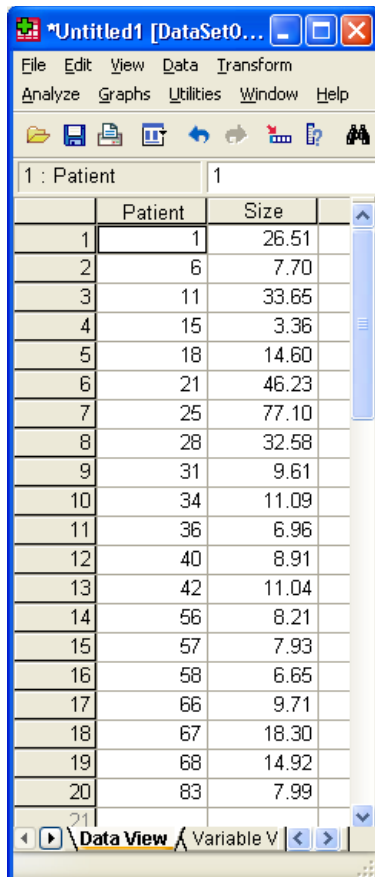


## D. Boxplots and percentiles

### D1. Introduction

Suppose we have a sample of 20 cancer patients and the sizes of their tumours (figure D1). We are going to look at a boxplot of these tumour sizes. To understand it, it will be helpful to rearrange the tumour sizes in order of size (see figure D2) and to look at the *percentiles*<sup>67</sup>.

Percentiles are also best understood when we have arranged the sizes into order. The percentile for a particular value shows what percentage of values in the sample are smaller than that value. In this case it shows, for each patient, what percentage of patients have a tumour size smaller than theirs. So if a patient is at the 75<sup>th</sup> percentile, 75% of the tumours are smaller than theirs<sup>68</sup>.



	Patient	Size
1	1	26.51
2	6	7.70
3	11	33.65
4	15	3.36
5	18	14.60
6	21	46.23
7	25	77.10
8	28	32.58
9	31	9.61
10	34	11.09
11	36	6.96
12	40	8.91
13	42	11.04
14	56	8.21
15	57	7.93
16	58	6.65
17	66	9.71
18	67	18.30
19	68	14.92
20	83	7.99

Fig D1. Data file in SPSS

Original line no	Patient	Size	Percentile
7	25	77.1	96
6	21	46.23	91
3	11	33.65	86
8	28	32.58	81
1	1	26.51	77
18	67	18.3	72
19	68	14.92	67
5	18	14.6	62
10	34	11.09	58
13	42	11.04	53
17	66	9.71	48
9	31	9.61	43
12	40	8.91	39
14	56	8.21	34
20	83	7.99	29
15	57	7.93	24
2	6	7.7	20
11	36	6.96	15
16	58	6.65	10
4	15	3.36	5

Fig D2. Data file re-ordered in order of tumour size.

<sup>67</sup> You can obtain percentiles in SPSS by going to **Transform – Rank Cases**. Click on **Rank Types**, select **Ntiles**, and enter the figure 100. To create a case number, see paragraph 13.4.

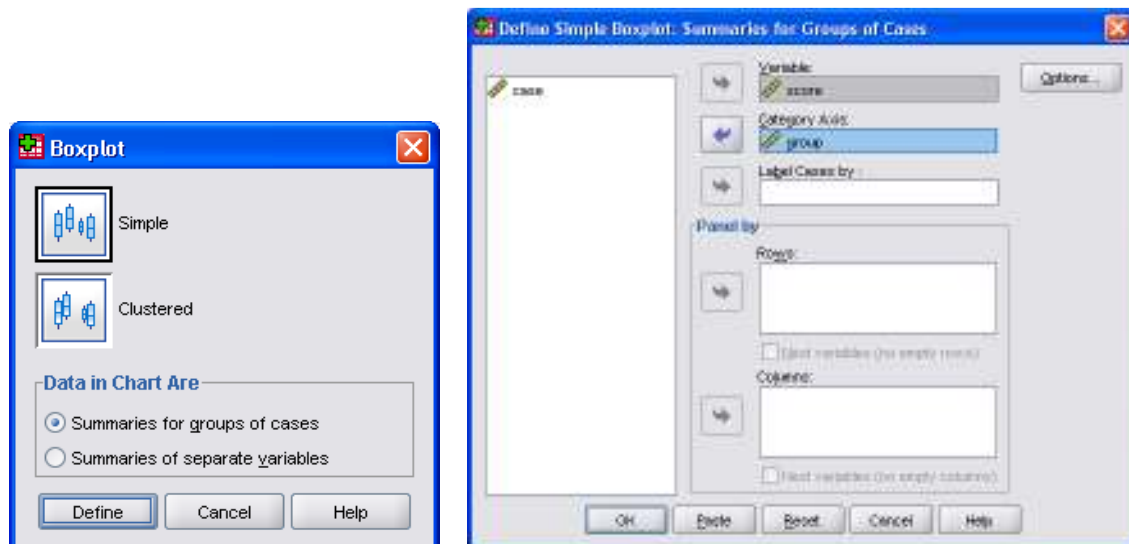
<sup>68</sup> If you remember, when the values are arranged in order of size, the value half way along is called the median. Therefore, the 50<sup>th</sup> percentile is the same thing as the median. Similarly, the 25<sup>th</sup> and 75<sup>th</sup> percentiles are also known as the lower and upper quartiles, respectively.

## D2. Obtaining boxplots

Go to **Graphs – Legacy Dialogs – Boxplot** on the drop-down menu. You obtain the dialogue box shown in figure D3(a).

If you have within-subjects data, click on **Summaries of separate variables** at the bottom. A second dialogue box appears; enter the variable(s) of interest under **Boxes represent**.

If you have between-subjects data, click on **Summaries for groups of cases**. Note that the file does not need to be split. A second dialogue box appears: Figure D3(b). Enter the variable for which you want the box plot (e.g. *score*) under **Variable** and your between-subjects variable (e.g. *group*) in the box that says **Category Axis**.



(a) first dialogue box

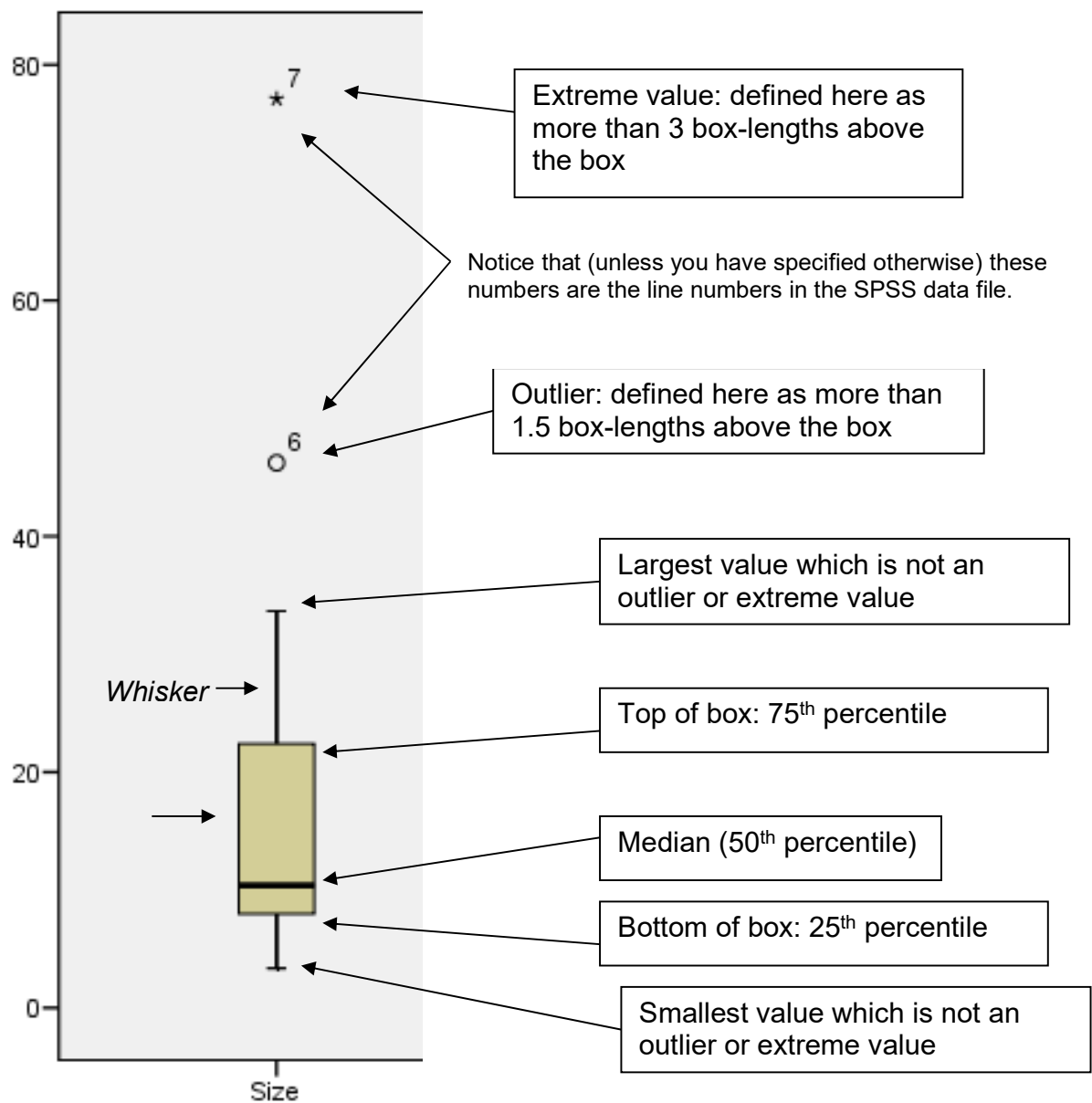
(b) second dialogue box for between-subjects data

Figure D3 Boxplot dialogue boxes.

## D3 Interpreting boxplots

A sample boxplot is shown in figure D4.

Note that outliers are defined by reference to the size of the box, which is in turn defined by the 25<sup>th</sup> and 75<sup>th</sup> percentiles. There are two categories of outlier depending on how extreme they are, but most researchers lump them together as simply 'outliers'.



**Figure D4. Boxplot<sup>69</sup>.**

Notice that in this boxplot, the median is well below the halfway point in the box, the top whisker is noticeably bigger than the bottom one, and there are two outliers or extreme values at the top. This indicates that the variable is skewed.

<sup>69</sup> Actually, some of my labels are slightly simplified. The top and bottom of the box ("Tukey's hinges") are defined slightly differently from the 25<sup>th</sup> and 75<sup>th</sup> percentile, so may not always match them. The whiskers are drawn 1.5 box lengths outside the box, or up to the data value which is furthest from the median but within that limit.

## E. Failing Levene's test

There are various options if you fail Levene's test.

1. You can do a non-parametric test, if there is one available. (There are even some in circumstances we have not covered, but most of these are too complicated for an introductory course.)
2. You may be able to correct the problem by eliminating outliers and/or transforming the DV (remembering to explain in your writeup exactly what you did). (See chapter 14, which is covered towards the end of the course.)
3. Presuming that there are no other serious problems such as outliers, the following advice is summarised from Tabachnick and Fidell (2007)<sup>70</sup>.
  - (a) If sample sizes are relatively equal (within a ratio of 4:1) you can look at the standard deviations (SDs) in each combination of conditions. You can accept a significant Anova result if the largest SD is not more than 3 times the smallest. (Tabachnick and Fidell actually define this in terms of variances – the largest variance should be no more than 10 times the smallest – but the variance is the square of the SD, and it is the SD which SPSS prints out for you in the Descriptives).
  - (b) Another option is to use a more stringent alpha level for the Anova. Instead of regarding the result as significant if  $p$  is less than .05, use .025 for "moderate violation" and .01 for "severe violation", although they don't define those expressions

What you must *not* do is to try out different methods until you get the result you want! Preferably, have a good reason for choosing one of them and stick to it, and/or be ready to show (perhaps with a footnote) that another method would have given a similar answer.

---

<sup>70</sup> Tabachnick, B. G. and Fidell, L. S. (2007). *Using Multivariate Statistics* (5<sup>th</sup> ed.). Boston: Pearson.

## F. Some useful transformations and how to compute them

Trying out different transformations is fine, as long as your criterion is to improve compliance with assumptions (normality, equality of variances, linear relationships etc.). Indeed, if a transformation improves compliance with one assumption, it is likely to improve compliance with others. What would be wrong would be if you were to choose a transformation on the basis that it produces the result you want in your eventual analysis. (For a fuller discussion see Howell, 2002, and/or Tabachnick and Fidell, 2007).

To carry out a transformation, use **Transform – Compute Variable** (paragraph 13.3). Under **Target Variable** enter the name you want for your new variable (represented by *new*) in the table.

### Choosing a transformation

The following tables show the most common transformations, with guidelines for when they are most likely to be useful and formulas to achieve them.

**Positive skew (i.e. the tail is longer on the right; any outliers<sup>71</sup> tend to have positive z-scores)**

Name	May be useful for:		Formula in SPSS (see notes)
	Amount of skew	If data is in groups	
Square root	Least	Variance is proportional to mean	$new = \text{SQRT}(old + a)$
Logarithm	Medium	SD is proportional to mean	$new = \text{LG10}(old + a)$
Inverse (or reciprocal)	Most; perhaps no left tail at all		$new = 1 / (old + a)$ [See also note b.]

/continued ...

---

<sup>71</sup> However, there may be outliers for other reasons than the skewed distribution. For example, with reaction times the distribution is usually positively skewed. But in addition there may be outliers to the left (because the participant made a random response, in less time than they could have reacted to the stimulus) and to the right (because they eventually made a random response, not in response to the stimulus). These may have to be removed independently of any decision about transforming the variable.

**Negative skew (i.e. the tail is longer on the left; any outliers<sup>72</sup> tend to have negative z-scores)**

Name	May be useful for:		Formula in SPSS (see notes)
	Amount of skew	If data is in groups	
Reflect and square root	Least	Variance is proportional to mean	$new = \text{SQRT}(c - old)$ [See also note d.]
Reflect and logarithm	Medium	SD is proportional to mean	$new = \text{LG10}(c - old)$ [See also note e.]
Reflect and inverse	Most; perhaps no right tail at all		$new = 1 / (c - old)$

#### Other

Name	Circumstance	Formula in SPSS
Arcsine	Is sometimes said to be useful when the old variable relates to proportions. It stretches out both tails.	$new = \text{ASIN}(old)$ [See note f.]

#### Notes

*old* = the name of the old variable (containing the scores you want to transform).

*new* = the new variable (the name you give to the transformed variable).

*a* If there are any negative scores in the old variable, *a* = the absolute value of the most negative score plus 1 (e.g. if the most negative value is -20, *a* = 21). If the old variable contains scores of 0 but no negative scores, *a* = 1. Otherwise, omit *a*.

*b* With the formula using *a*, an inverse transformation reverses the scores (i.e. the lowest old score becomes the highest new score). If you prefer to avoid this, you can use the transformation  $new = (b - old)$ . If there are no negative scores in the old variable, *b* = the highest score plus 1. If there are negative scores in the old variable, *b* = the highest score, plus 1, minus the lowest score (i.e. if the scores run from -3 to +4, *b* = 8.).

---

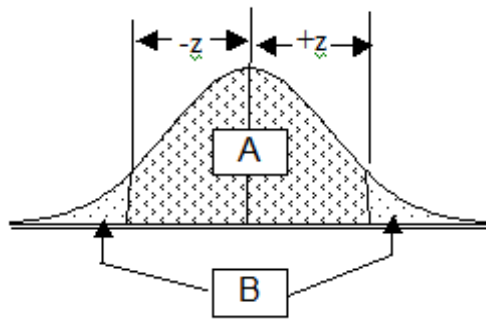
<sup>72</sup> See footnote 71.

- $c$  = a number such that when each score is subtracted from it, the smallest result is 1. If all the scores are all positive,  $c$  = the largest score + 1.
- $d$  = With the formula using  $c$ , the transformation reverses the scores (i.e. the lowest old score becomes the highest new score). If you prefer to avoid this, you can add a minus sign so that  $new = -\text{SQRT}(c - old)$ , or even  $new = d - \text{SQRT}(c - old)$ , where  $d$  is a number of your choice.
- $e$  = With the formula using  $c$ , the transformation reverses the scores (i.e. the lowest old score becomes the highest new score). If you prefer to avoid this, you can add a minus sign so that  $new = -\text{LG10}(c - old)$ , or even  $new = e - \text{LG10}(c - old)$ , where  $e$  is a number of your choice.
- $f$  Arcsine only works when all the values of the old score are within the range -1 to +1, which they will be if the variable represents a proportion. If the scores are not all in this range, you could try dividing the scores by the largest score in the dataset (ignoring the sign). In other words, use the formula  $new = \text{ASIN}(old / f)$  where  $f$  is the largest (absolute) value in the dataset.

And finally ... If the above advice does not give enough options, you could try experimenting with different values of the constants  $a$ ,  $b$ ,  $c$ ,  $d$ ,  $e$ ,  $f$  above.

But not all variables can be made normal by transforming them. Sometimes you may be prepared to use a parametric test with non-normal data. Other options include using a non-parametric test, or dividing the variable into two or more groups to create a new, categorical, variable.

## G.Areas under the normal distribution

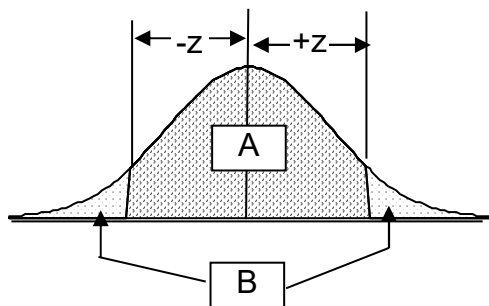


A = area within  $\pm z$  of mean

B = area outside  $\pm z$  of mean  
(2 tailed probability of a random score being at least as extreme as  $z$ )

z	A	B	z	A	B	z	A	B	z	A	B
0.00	0.000	1.000	0.50	0.383	0.617	1.00	0.683	0.317	1.50	0.866	0.134
0.01	0.008	0.992	0.51	0.390	0.610	1.01	0.688	0.312	1.51	0.869	0.131
0.02	0.016	0.984	0.52	0.397	0.603	1.02	0.692	0.308	1.52	0.871	0.129
0.03	0.024	0.976	0.53	0.404	0.596	1.03	0.697	0.303	1.53	0.874	0.126
0.04	0.032	0.968	0.54	0.411	0.589	1.04	0.702	0.298	1.54	0.876	0.124
0.05	0.040	0.960	0.55	0.418	0.582	1.05	0.706	0.294	1.55	0.879	0.121
0.06	0.048	0.952	0.56	0.425	0.575	1.06	0.711	0.289	1.56	0.881	0.119
0.07	0.056	0.944	0.57	0.431	0.569	1.07	0.715	0.285	1.57	0.884	0.116
0.08	0.064	0.936	0.58	0.438	0.562	1.08	0.720	0.280	1.58	0.886	0.114
0.09	0.072	0.928	0.59	0.445	0.555	1.09	0.724	0.276	1.59	0.888	0.112
0.10	0.080	0.920	0.60	0.451	0.549	1.10	0.729	0.271	1.60	0.890	0.110
0.11	0.088	0.912	0.61	0.458	0.542	1.11	0.733	0.267	1.61	0.893	0.107
0.12	0.096	0.904	0.62	0.465	0.535	1.12	0.737	0.263	1.62	0.895	0.105
0.13	0.103	0.897	0.63	0.471	0.529	1.13	0.742	0.258	1.63	0.897	0.103
0.14	0.111	0.889	0.64	0.478	0.522	1.14	0.746	0.254	1.64	0.899	0.101
0.15	0.119	0.881	0.65	0.484	0.516	1.15	0.750	0.250	1.65	0.901	0.099
0.16	0.127	0.873	0.66	0.491	0.509	1.16	0.754	0.246	1.66	0.903	0.097
0.17	0.135	0.865	0.67	0.497	0.503	1.17	0.758	0.242	1.67	0.905	0.095
0.18	0.143	0.857	0.68	0.503	0.497	1.18	0.762	0.238	1.68	0.907	0.093
0.19	0.151	0.849	0.69	0.510	0.490	1.19	0.766	0.234	1.69	0.909	0.091
0.20	0.159	0.841	0.70	0.516	0.484	1.20	0.770	0.230	1.70	0.911	0.089
0.21	0.166	0.834	0.71	0.522	0.478	1.21	0.774	0.226	1.71	0.913	0.087
0.22	0.174	0.826	0.72	0.528	0.472	1.22	0.778	0.222	1.72	0.915	0.085
0.23	0.182	0.818	0.73	0.535	0.465	1.23	0.781	0.219	1.73	0.916	0.084
0.24	0.190	0.810	0.74	0.541	0.459	1.24	0.785	0.215	1.74	0.918	0.082
0.25	0.197	0.803	0.75	0.547	0.453	1.25	0.789	0.211	1.75	0.920	0.080
0.26	0.205	0.795	0.76	0.553	0.447	1.26	0.792	0.208	1.76	0.922	0.078
0.27	0.213	0.787	0.77	0.559	0.441	1.27	0.796	0.204	1.77	0.923	0.077
0.28	0.221	0.779	0.78	0.565	0.435	1.28	0.799	0.201	1.78	0.925	0.075
0.29	0.228	0.772	0.79	0.570	0.430	1.29	0.803	0.197	1.79	0.927	0.073
0.30	0.236	0.764	0.80	0.576	0.424	1.30	0.806	0.194	1.80	0.928	0.072
0.31	0.243	0.757	0.81	0.582	0.418	1.31	0.810	0.190	1.81	0.930	0.070
0.32	0.251	0.749	0.82	0.588	0.412	1.32	0.813	0.187	1.82	0.931	0.069
0.33	0.259	0.741	0.83	0.593	0.407	1.33	0.816	0.184	1.83	0.933	0.067
0.34	0.266	0.734	0.84	0.599	0.401	1.34	0.820	0.180	1.84	0.934	0.066
0.35	0.274	0.726	0.85	0.605	0.395	1.35	0.823	0.177	1.85	0.936	0.064
0.36	0.281	0.719	0.86	0.610	0.390	1.36	0.826	0.174	1.86	0.937	0.063
0.37	0.289	0.711	0.87	0.616	0.384	1.37	0.829	0.171	1.87	0.939	0.061
0.38	0.296	0.704	0.88	0.621	0.379	1.38	0.832	0.168	1.88	0.940	0.060
0.39	0.303	0.697	0.89	0.627	0.373	1.39	0.835	0.165	1.89	0.941	0.059
0.40	0.311	0.689	0.90	0.632	0.368	1.40	0.838	0.162	1.90	0.943	0.057
0.41	0.318	0.682	0.91	0.637	0.363	1.41	0.841	0.159	1.91	0.944	0.056
0.42	0.326	0.674	0.92	0.642	0.358	1.42	0.844	0.156	1.92	0.945	0.055
0.43	0.333	0.667	0.93	0.648	0.352	1.43	0.847	0.153	1.93	0.946	0.054
0.44	0.340	0.660	0.94	0.653	0.347	1.44	0.850	0.150	1.94	0.948	0.052
0.45	0.347	0.653	0.95	0.658	0.342	1.45	0.853	0.147	1.95	0.949	0.051
0.46	0.354	0.646	0.96	0.663	0.337	1.46	0.856	0.144	1.96	0.950	0.050
0.47	0.362	0.638	0.97	0.668	0.332	1.47	0.858	0.142	1.97	0.951	0.049
0.48	0.369	0.631	0.98	0.673	0.327	1.48	0.861	0.139	1.98	0.952	0.048
0.49	0.376	0.624	0.99	0.678	0.322	1.49	0.864	0.136	1.99	0.953	0.047



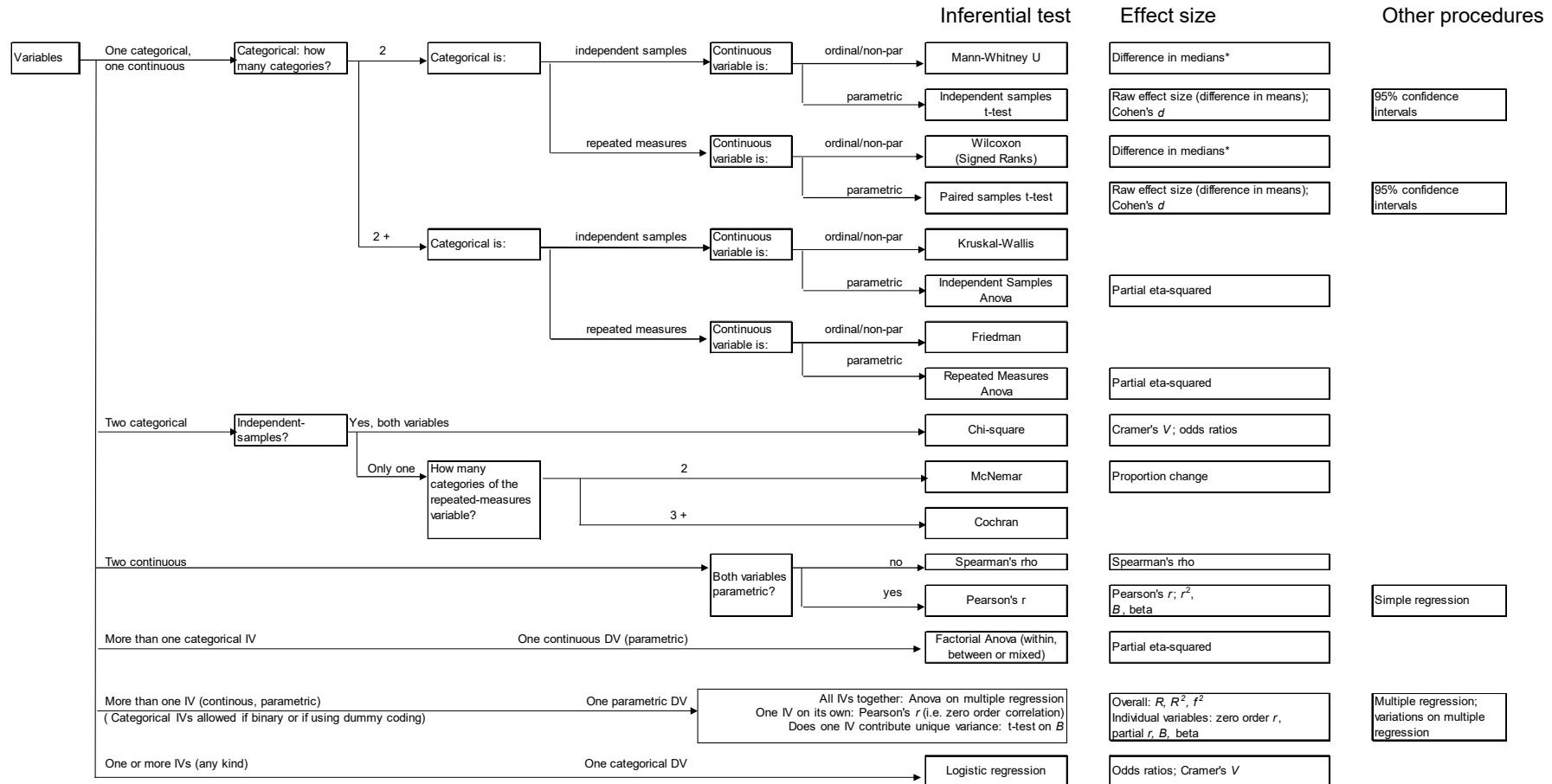


A = area within  $\pm z$  of mean

B = area outside  $\pm z$  of mean  
(2 tailed probability of a random score being at least as extreme as  $z$ )

z	A	B	z	A	B	z	A	B	z	A	B
2.00	0.954	0.046	2.50	0.988	0.012	3.00	0.997	0.003	3.29	0.999 000	0.001 000
2.01	0.956	0.044	2.51	0.988	0.012	3.01	0.997	0.003	3.89	0.999 900	0.000 100
2.02	0.957	0.043	2.52	0.988	0.012	3.02	0.997	0.003	4.41	0.999 990	0.000 010
2.03	0.958	0.042	2.53	0.989	0.011	3.03	0.998	0.002	5.07	0.999 999	0.000 001
2.04	0.959	0.041	2.54	0.989	0.011	3.04	0.998	0.002			
2.05	0.960	0.040	2.55	0.989	0.011	3.05	0.998	0.002	3.09	0.998 000	0.002 000
2.06	0.961	0.039	2.56	0.990	0.010	3.06	0.998	0.002	3.72	0.999 800	0.000 200
2.07	0.962	0.038	2.57	0.990	0.010	3.07	0.998	0.002	4.27	0.999 980	0.000 020
2.08	0.962	0.038	2.58	0.990	0.010	3.08	0.998	0.002	4.77	0.999 998	0.000 002
2.09	0.963	0.037	2.59	0.990	0.010	3.09	0.998	0.002			
2.10	0.964	0.036	2.60	0.991	0.009	3.10	0.998	0.002			
2.11	0.965	0.035	2.61	0.991	0.009	3.11	0.998	0.002			
2.12	0.966	0.034	2.62	0.991	0.009	3.12	0.998	0.002			
2.13	0.967	0.033	2.63	0.991	0.009	3.13	0.998	0.002			
2.14	0.968	0.032	2.64	0.992	0.008	3.14	0.998	0.002			
2.15	0.968	0.032	2.65	0.992	0.008	3.15	0.998	0.002			
2.16	0.969	0.031	2.66	0.992	0.008	3.16	0.998	0.002			
2.17	0.970	0.030	2.67	0.992	0.008	3.17	0.998	0.002			
2.18	0.971	0.029	2.68	0.993	0.007	3.18	0.999	0.001			
2.19	0.971	0.029	2.69	0.993	0.007	3.19	0.999	0.001			
2.20	0.972	0.028	2.70	0.993	0.007	3.20	0.999	0.001			
2.21	0.973	0.027	2.71	0.993	0.007	3.21	0.999	0.001			
2.22	0.974	0.026	2.72	0.993	0.007	3.22	0.999	0.001			
2.23	0.974	0.026	2.73	0.994	0.006	3.23	0.999	0.001			
2.24	0.975	0.025	2.74	0.994	0.006	3.24	0.999	0.001			
2.25	0.976	0.024	2.75	0.994	0.006	3.25	0.999	0.001			
2.26	0.976	0.024	2.76	0.994	0.006	3.26	0.999	0.001			
2.27	0.977	0.023	2.77	0.994	0.006	3.27	0.999	0.001			
2.28	0.977	0.023	2.78	0.995	0.005	3.28	0.999	0.001			
2.29	0.978	0.022	2.79	0.995	0.005	3.29	0.999	0.001			
2.30	0.979	0.021	2.80	0.995	0.005	3.30	0.999	0.001			
2.31	0.979	0.021	2.81	0.995	0.005	3.31	0.999	0.001			
2.32	0.980	0.020	2.82	0.995	0.005	3.32	0.999	0.001			
2.33	0.980	0.020	2.83	0.995	0.005	3.33	0.999	0.001			
2.34	0.981	0.019	2.84	0.995	0.005	3.34	0.999	0.001			
2.35	0.981	0.019	2.85	0.996	0.004	3.35	0.999	0.001			
2.36	0.982	0.018	2.86	0.996	0.004	3.36	0.999	0.001			
2.37	0.982	0.018	2.87	0.996	0.004	3.37	0.999	0.001			
2.38	0.983	0.017	2.88	0.996	0.004	3.38	0.999	0.001			
2.39	0.983	0.017	2.89	0.996	0.004	3.39	0.999	0.001			
2.40	0.984	0.016	2.90	0.996	0.004	3.40	0.999	0.001			
2.41	0.984	0.016	2.91	0.996	0.004	3.41	0.999	0.001			
2.42	0.984	0.016	2.92	0.996	0.004	3.42	0.999	0.001			
2.43	0.985	0.015	2.93	0.997	0.003	3.43	0.999	0.001			
2.44	0.985	0.015	2.94	0.997	0.003	3.44	0.999	0.001			
2.45	0.986	0.014	2.95	0.997	0.003	3.45	0.999	0.001			
2.46	0.986	0.014	2.96	0.997	0.003	3.46	0.999	0.001			
2.47	0.986	0.014	2.97	0.997	0.003	3.47	0.999	0.001			
2.48	0.987	0.013	2.98	0.997	0.003	3.48	0.999	0.001			
2.49	0.987	0.013	2.99	0.997	0.003	3.49	1.000	0.000			

## H. Overview of statistical tests



### Notes

Sometimes there is more than one possible test to use.  
 Not all procedures shown above are covered on this course.  
 Nonetheless, in circumstances not covered above, a more advanced test may be available

\*or possibly as for parametric data