1 Some Preliminaries

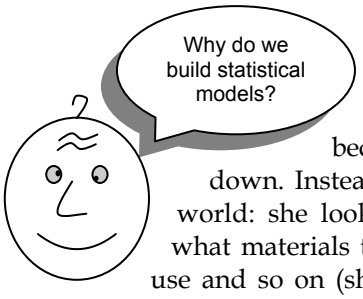
There are several things that I need to talk about before the main body of this book. Although it seems obvious that I would like you to read Chapter 1 first (otherwise I wouldn't have placed it at the beginning) I am aware that many students derive little pleasure from reading statistics books in their entirety and prefer to dip into relevant chapters. With this in mind, I urge you to read Chapter 1 before any other because the contents are important in understanding what follows. The two things I need to talk about are: (1) model building and linear models; and (2) the SPSS environment itself. Bear with me, it won't take long.

1.1. Statistical Models

1.1.1. Model Building

In the social sciences we are usually interested in discovering something about a phenomenon that we assume actually exists (something I refer to as a real-world phenomenon). These real-world phenomena can be anything from the behaviour of interest rates in the economic market to the behaviour of undergraduates at the end-of-exam party. Whatever the phenomenon we desire to explain, we seek to explain it by collecting data from the real world, and then using these data to draw conclusions about what is being studied. As statisticians our job is to take the available data and to use them in a meaningful way and this often involves building statistical models of the phenomenon of interest.

The reason for building statistical models of real-world data is best explained by analogy. Imagine an engineer wishes to build a bridge across a river. That engineer would be pretty daft if she just built any old bridge, because the chances are that it would fall down. Instead, an engineer collects data from the real world: she looks at bridges in the real world and sees what materials they are made from, what structures they use and so on (she might even collect data about whether these bridges are damaged!). She then uses this information to construct a model. She builds a scaled-down version of the real-world bridge because it is impractical, not to mention expensive, to build the actual



bridge itself. The model may differ from reality in a number of ways—it will be smaller for a start—but the engineer will try to build a model that best fits the situation of interest based on the data available. Once the model has been built, it can be used to predict things about the real world: for example, the engineer might test whether the bridge can withstand strong winds by placing the model in a wind tunnel. It seems obvious that it is important that the model is an accurate representation of the real world. Social scientists do much the same thing as engineers: we build models of real-world processes in an attempt to predict how these processes operate under certain conditions. We don't have direct access to the processes, so we collect data that represent the processes and then use these data to build statistical models (we reduce the process to a statistical model). We then use this statistical model to make predictions about the real-world phenomenon. Just like the engineer, we want our models to be as accurate as possible so that we can be confident that the predictions we make are also accurate. However, unlike engineers we don't have access to the real-world situation and so we can only ever *infer* things about psychological, societal or economic processes based upon the models we build. If we want our inferences to be accurate then the statistical model we build must represent the data collected (the *observed data*) as closely as possible. The degree to which a statistical model represents the data collected is known as the *fit* of the model and this is a term you will frequently come across.

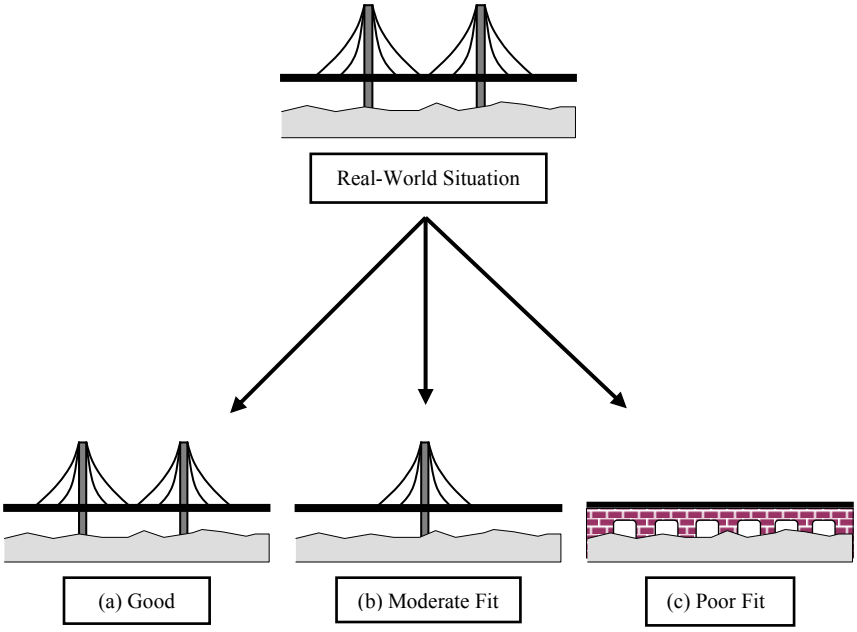


Figure 1.1: Fitting models to real-world data (see text for details)

Figure 1.1 illustrates the kinds of models that an engineer might build to represent the real-world bridge that she wants to create. The first model (a) is an excellent representation of the real-world situation and is said to be a *good fit* (i.e. there are a few small differences but the model is basically a very good replica of reality). If this model is used to make predictions about the real world, then the engineer can be confident that these predictions will be very accurate, because the model so closely resembles reality. So, if the model collapses in a strong wind, then there is a good chance that the real bridge would collapse also. The second model (b) has some similarities to the real world: the model includes some of the basic structural features, but there are some big differences from the real-world bridge (namely the absence of one of the supporting towers). This is what we might term a *moderate fit* (i.e. there are some differences between the model and the data but there are also some great similarities). If the engineer uses this model to make predictions about the real world then these predictions may be inaccurate and possibly catastrophic (for example, if the bridge collapses in strong winds this could be due to the absence of a second supporting tower). So, using this model results in predictions that we can have some confidence in but not complete confidence. The final model (c) is completely different to the real-world situation. This model bears no structural similarities to the real bridge and so could be termed a poor fit (in fact, it might more accurately be described as an abysmal fit!). As such, any predictions based on this model are likely to be completely inaccurate. Extending this analogy to the social sciences we can say that it is important when we fit a statistical model to a set of data that this model fits the data well. If our model is a poor fit of the observed data then the predictions we make from it will be equally poor.

1.1.2. Populations and Samples

As researchers, we are interested in finding results that apply to an entire population of people or things. For example, psychologists want to discover processes that occur in all humans, biologists might be interested in processes that occur in all cells, economists want to build models that apply to all salaries and so on. A population can be very general (all human beings) or very narrow (all male ginger cats called Bob), but in either case scientists rarely, if ever, have access to every member of a population. Psychologists cannot collect data from every human being and ecologists cannot observe every male ginger cat called Bob. Therefore, we collect data from a small subset of the population (known as a *sample*) and use these data to infer things about the population as a whole. The bridge-building engineer cannot make a full-size model of the bridge she wants to build and so she builds a small-

scale model and tests this model under various conditions. From the results obtained from the small-scale model the engineer infers things about how the full-sized bridge will respond. The small-scale model may respond differently to a full-sized version of the bridge, but the larger the model, the more likely it is to behave in the same way as the full-size bridge. This metaphor can be extended to social scientists. We never have access to the entire population (the real-size bridge) and so we collect smaller samples (the scaled-down bridge) and use the behaviour within the sample to infer things about the behaviour in the population. The bigger the sample, the more likely it is to reflect the whole population. If we take several random samples from the population, each of these samples will give us slightly different results. However, on average, large samples should be fairly similar.

1.1.3. Simple Statistical Models

1.1.3.1. The Mean, Sums of Squares, Variance and Standard Deviations

One of the simplest models used in statistics is the mean. Some of you may have trouble thinking of the mean as a model, but in fact it is because it represents a summary of data. The mean is a hypothetical value that can be calculated for any data set, it doesn't have to be a value that is actually observed in the data set. For example, if we took five statistics lecturers and measured the number of friends that they had, we might find the following data: 1, 2, 3, 3 and 4. If we take the mean number of friends, this can be calculated by adding the values we obtained, and dividing by the number of values measured: (1 + 2 + 3 + 3 + 4)/5 = 2.6. Now, we know that it is impossible to have 2.6 friends (unless you chop someone up with a chainsaw and befriend their arm) so the mean value is a *hypothetical* value. As such, the mean is a model created to summarize our data. Now, we can determine whether this is an accurate model by looking at how different our real data are from the model that we have created. One way to do this is to look at the difference between the data we observed and the model fitted. Figure 1.2 shows the number of friends that each statistics lecturer had, and also the mean number that we calculated earlier on. The line representing the mean can be thought of as our model, and the circles are the observed data. The diagram also has a series of vertical lines that connect each observed value to the mean value. These lines represent the differences between the observed data and our model and can be thought of as the error in the model. We can calculate the magnitude of these differences by simply subtracting the mean value ($\bar{x}$) from each of

the observed values (x_i).[1] For example, lecturer 1 had only 1 friend and so the difference is $x_1 - \bar{x} = 1 - 2.6 = -1.6$. You might notice that the difference is a minus number, and this represents the fact that our model *overestimates* this lecturer's popularity: it predicts that he will have 2.6 friends yet in reality he has only 1 friend (bless him!). Now, how can we use these differences to estimate the accuracy of the model? One possibility is to add up the differences (this would give us an estimate of the total error). If we were to do this we would find that:

$$\text{total error} = \text{sum of differences} = \sum(x_i - \bar{x}) = (-1.6) + (-0.6) + (0.4) + (0.4) + (1.4) = 0$$

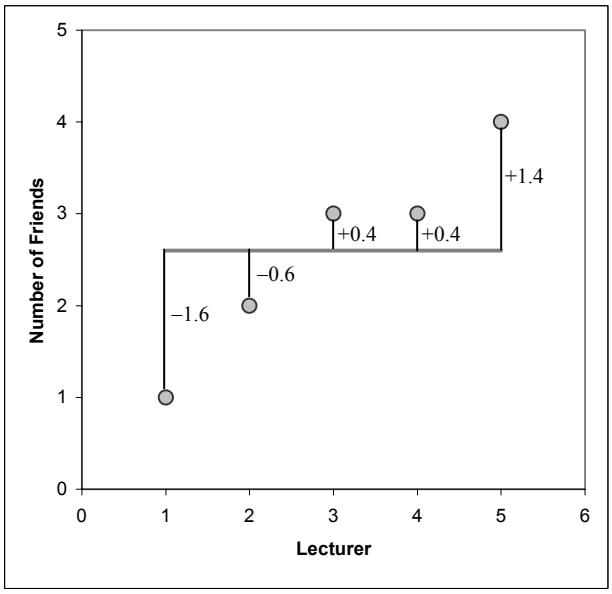


Figure 1.2: Graph showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends

So, in effect the result tells us that there is no total error between our model and the observed data, so, the mean is a perfect representation of the data. Now, this clearly isn't true: there were errors but some of them were positive, some were negative and they have simply cancelled each other out. It is clear that we need to avoid the problem of which direction the error is in and one mathematical way to do this is to square each error,[2] that is multiply each error by itself. So, rather than

[1] The x_i simply refers to the observed score for the ith person (so, the i can be replaced with a number that represents a particular individual). For these data: for lecturer 1, $x_i = x_1 = 1$; for lecturer 3, $x_i = x_3 = 3$; for lecturer 5, $x_i = x_5 = 4$.

[2] When you multiply a negative number by itself it becomes positive.

calculating the sum of errors, we calculate the sum of squared errors. In this example:

$$\begin{aligned}\text{sum of squared errors (SS)} &= \sum (x_i - \bar{x})(x_i - \bar{x}) \\ &= (-1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2 \\ &= 2.56 + 0.36 + 0.16 + 0.16 + 1.96 \\ &= 5.20\end{aligned}$$

The sum of squared errors (SS) is a good measure of the accuracy of our model. However, it is fairly obvious that the sum of squared errors is dependent upon the amount of data that has been collected—the more data points the higher the SS. To overcome this problem we calculate the average error by dividing the SS by the number of observations (N). If we are interested only in the average error for the sample, then we can divide by N alone. However, we are generally interested in using the error in the sample to estimate the error in the population and so we divide the SS by the number of observations minus 1 (the reason why is explained in section 7.1.3.2). This measure is known as the *variance* and is a measure that we will come across a great deal:

$$\text{variance } (s^2) = \frac{\text{SS}}{\text{N} - 1} = \frac{\sum (x_i - \bar{x})^2}{N - 1} = \frac{5.20}{4} = 1.3$$

The variance is, therefore, the average error between the mean and the observations made (and so is a measure of how well the model fits the actual data). There is one problem with the variance as a measure: it gives us a measure in units squared (because we squared each error in the calculation). In our example we would have to say that the average error in our data (the variance) was 1.3 friends squared. It makes little enough sense to talk about 1.3 friends, but it makes even less to talk about friends-squared! For this reason, we often take the square root of the variance (which ensures that the measure of average error is in the same units as the original measure). This measure is known as the standard deviation and is simply the square root of the variance. In this example the standard deviation is:

$$s = \sqrt{1.3} = 1.14$$

The standard deviation is, therefore, a measure of how well the mean represents the data. Small standard deviations (relative to the value of the mean itself) indicate that data points are close to the mean. A large standard deviation (relative to the mean) indicates that the data points are distant from the mean (i.e. the mean is not an accurate representation of the data). Figure 1.3 shows the overall ratings (on a five-point scale) of two lecturers after each of five different lectures. Both lecturers had an average rating of 2.6 out of 5 across the lectures.

However, the first lecturer had a standard deviation of 0.55 (relatively small compared to the mean). It should be clear from the graph that ratings for this lecturer were consistently close to the mean rating. There was a small fluctuation, but generally his lectures did not vary in popularity. As such, the mean is an accurate representation of his ratings. The mean is a good fit of the data. The second lecturer, however, had a standard deviation of 1.82 (relatively high compared to the mean). The ratings for this lecturer are clearly more spread from the mean, that is, for some lectures he received very high ratings, and for others his ratings were appalling. Therefore, the mean is not such an accurate representation of his performance because there was a lot of variability in the popularity of his lectures. The mean is a poor fit of the data. This illustration should hopefully make clear why the standard deviation is a measure of how well the mean represents the data.

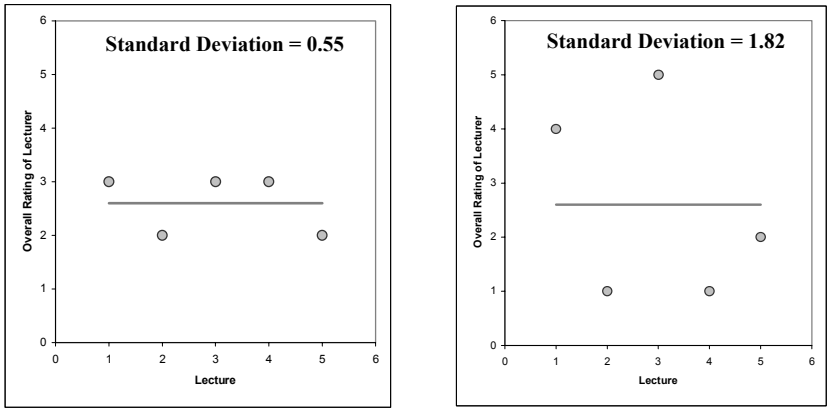


Figure 1.3: Graphs illustrating data that have the same mean but different standard deviations

The discussion of means, sums of squares and variance may seem a side-track from the initial point, but in fact the mean is probably one of the simplest statistical models that can be fitted to data. As such, the mean illustrates the concept of a statistical model and the variance and standard deviation illustrate how the goodness-of-fit of a model can be measured.

1.1.3.2. The Standard Error

Many students get confused about the difference between the standard deviation and the standard error (usually because the difference is never explained clearly). However, the standard error is an important concept to grasp, so I'll do my best to explain it to you. We have already learnt that social scientists use samples as a way of estimating the behaviour in

a population. I also mentioned that if you take several samples from a population, then these samples would differ slightly. Imagine that we were interested in the ratings of all lecturers (so, lecturers in general were the population). We could take five samples from this population with each sample containing five different lecturers.

Figure 1.4 illustrates this scenario. The ellipses represent the five samples, and contain the overall ratings of five lecturers (the grey dots). For each sample we can calculate the average, or *sample mean* (represented by a horizontal line). As you can see in the diagram, each sample has a slightly different mean: the lecturers in sample 1 had a mean rating of 1, whereas the lecturers in sample 4 had a mean rating of 3. If you calculated the average rating of all samples, then the value would be the mean of the sample means. The long dark horizontal line represents this overall mean (this value is the same as if you calculate the mean of all data points irrespective of the sample from which they come).

Figure 1.4 illustrates a situation in which we have taken only five samples, but imagine we took hundreds or even thousands of samples. For each sample we could calculate the average rating, and we could then calculate the average of all sample means. If we were to do this for hundreds of samples, then the average of all the sample means would be roughly equal to the mean of the whole population. Therefore, the long horizontal line in Figure 1.4 is going to be roughly equal to the population mean. If we take random samples, then the majority of these samples will have a mean that is equal, or very similar, to the population mean. However, the occasional sample will have a mean that is very different from the population (perhaps because by chance, that sample contained a lot of very good lecturers). This is also illustrated in Figure 1.4: the majority of samples have average values close to the population average (as shown by the short arrows between the population mean and the sample means), but the first and last samples have means that are distant from the population mean. If you think about this logically, if we want to infer things about a population, by using a sample, it is important that we know how well that sample represents the population. In this example, if we used samples 2, 3 or 4 we could be confident about any conclusions we make, because the sample means are representative of the population mean. However, if we happened to use either sample 1 or sample 5, then our conclusions would be inaccurate (because these two samples are not characteristic of the population as a whole). So, how can we gauge whether a sample is representative?

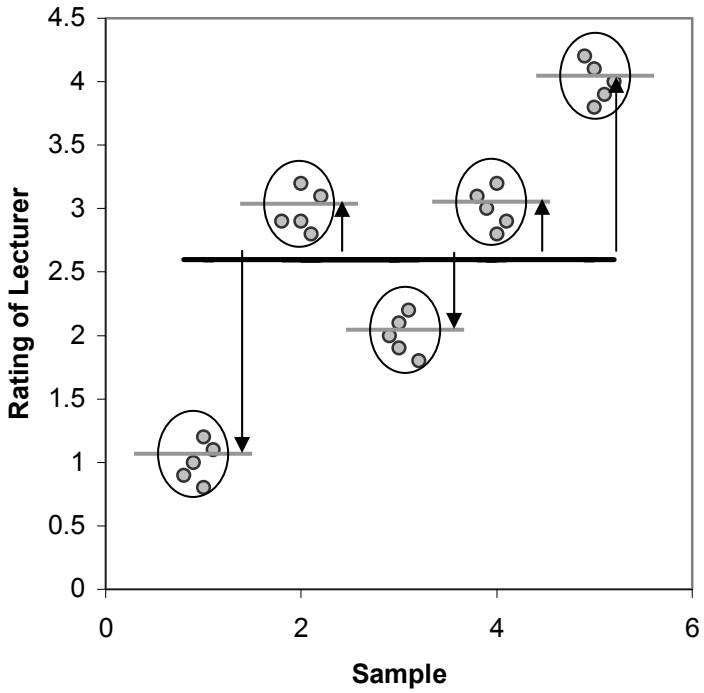


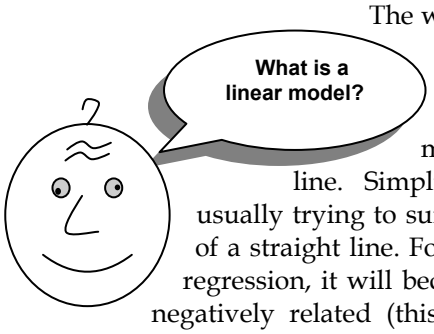
Figure 1.4: Graph illustrating the standard error (see text for details)

Think back to the discussion of the standard deviation. We used the standard deviation as a measure of how representative the mean was of the observed data. Small standard deviations represented a scenario in which most data points were close to the mean, a large standard deviation represented a situation in which data points were widely spread from the mean. If you were to calculate the standard deviation between *sample means* then this too would give you a measure of how much variability there was between the means of different samples. The standard deviation of sample means is known as the *standard error*. Therefore, the standard error could be calculated by taking the difference between each sample mean and the overall mean (the arrows in Figure 1.4), squaring these differences, adding them up, and then dividing by the number of samples. To clarify this point, look at the similarity between Figure 1.4 and Figure 1.2. Of course, in reality we cannot collect hundreds of samples and so we rely on approximations of the standard error (luckily for us lots of clever statisticians have calculated ways in which the standard error can be worked out from the sample standard deviation). So, in short, the standard error is the standard deviation of sample means. As such, it is a measure of how

representative a sample is likely to be of the population. A large standard error (relative to the sample mean) means that there is a lot of variability between the means of different samples and so the sample we have might not be representative of the population. A small standard error indicates that most sample means are similar to the population mean and so our sample is likely to be an accurate reflection of the population.

1.1.4. Linear Models

The mean is an example of what we call a statistical model, but you may well ask what other kinds of statistical models can be built. Well, if the truth is known there is only one model that is generally used, and this is known as the linear model. To some social scientists it may not be entirely obvious that my previous statement is correct, yet a statistician would acknowledge my sentiment much more readily. The reason for this is that there are a variety of different names given to statistical procedures that are based on the linear model. A classic example is that analysis of variance (ANOVA) and regression are identical systems (Cohen, 1968), yet they have different names and are used largely in different contexts (due to a divide in methodological philosophies—see Cronbach, 1957).



The word *linear* literally means 'relating to a line' but in statistical terms the line to which it refers is a straight one. A linear model is, therefore, a model that is based upon a straight line. Simplistically, this means that we are usually trying to summarize our observed data in terms of a straight line. For example, in the chapter describing regression, it will become clear that two variables can be negatively related (this just means that as values of one variable increase, values of the other variable decrease). In such circumstances, the relationship may be summarized by a straight line. Suppose we measured how many chapters of this book a person had read, and then measured their spiritual enrichment, we could represent these hypothetical data in the form of a scatterplot in which each dot represents an individual's score on both variables. Figure 1.5 shows such a graph, and also shows the same graph but with a line that summarizes the pattern of these data. A third version of the scatterplot is also included but has a curved line to summarize the general pattern of the data. As such, Figure 1.5 illustrates how we can fit different types of models to the same data. In this case we can use a straight line to represent our data and it shows that the more chapters a person reads,

the less their spiritual enrichment. However, we can also use a curved line to summarize the data and this shows that when most, or all, of the chapters have been read, spiritual enrichment seems to increase slightly (presumably because once the book is read everything suddenly makes sense—yeah, as if!). Neither of the two types of model is necessarily correct, but it will be the case that one model fits the data better than another and this is why when we use statistical models it is important for us to assess how well a given model fits the data.

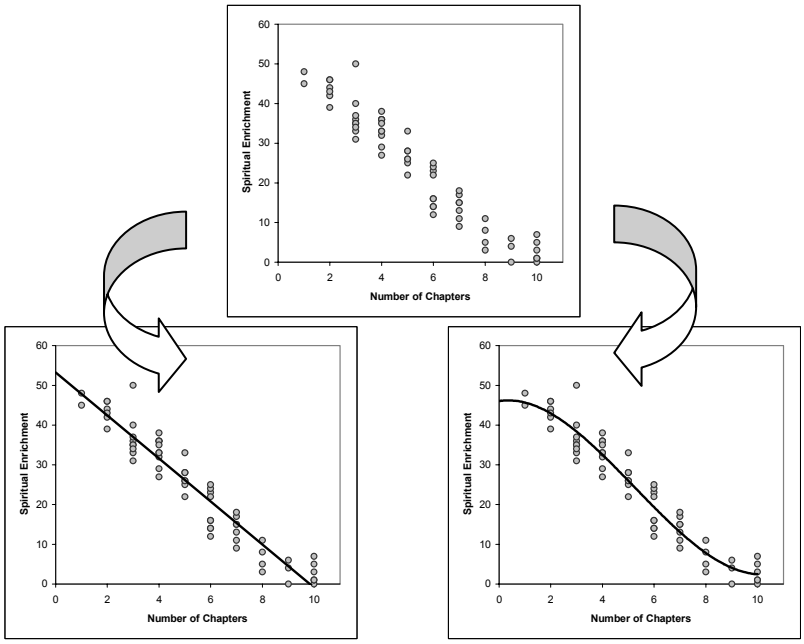


Figure 1.5: Shows a scatterplot of some data with no model fitted to the data, with a linear model fitted, and with a non-linear model fitted

Most of the statistics used in the social sciences are based on linear models, which means that we try to fit straight line models to the data collected. This is interesting because most published scientific studies are ones with statistically significant results. Given that most social scientists are only ever taught how to use techniques based on the linear model, published results will be those that have successfully used linear models. Data that fit a non-linear pattern are likely to be wrongly ignored (because the wrong model will have been applied to the data, leading to non-significant results). It is possible, therefore, that some areas of science are progressing in a biased way, so, if you collect data that look non-linear, why not try redressing the balance and investigating different statistical techniques!

1.2. The SPSS Environment

There are several excellent texts that give introductions to the general environment within which SPSS operates. The best ones include Kinnear and Gray (1997) and Foster (1998). These texts are well worth reading if you are unfamiliar with Windows and SPSS generally because I am assuming at least some knowledge of the system. However, I appreciate the limited funds of most students and so to make this text usable for those inexperienced with SPSS I will provide a brief guide to the SPSS environment—but for a more detailed account see the previously cited texts and the SPSS manuals. This book is based primarily on version 9.0 of SPSS (at least in terms of the diagrams); however, it also caters for versions 7.0, 7.5 and 8.0 (there are few differences between versions 7.0, 8.0 and 9.0 and any obvious differences are highlighted where relevant).

Once SPSS has been activated, the program will automatically load two windows: the data editor (this is where you input your data and carry out statistical functions) and the output window (this is where the results of any analysis will appear). There are a number of additional windows that can be activated. In versions of SPSS earlier than version 7.0, graphs appear in a separate window known as the *chart carousel*; however, versions 7.0 and after include graphs in the output window, which is called the *output navigator* (version 7.0) and the *output viewer* (version 8.0 and after). Another window that is useful is the syntax window, which allows you to enter SPSS commands manually (rather than using the window-based menus). At most levels of expertise, the syntax window is redundant because you can carry out most analyses by clicking merrily with your mouse. However, there are various additional functions that can be accessed using syntax and sick individuals who enjoy statistics can find numerous uses for it! I will pretty much ignore syntax windows because those of you who want to know about them will learn by playing around and the rest of you will be put off by their inclusion (interested readers should refer to Foster, 1998, Chapter 8).

1.2.1. The Data Editor

The main SPSS window includes a data editor for entering data. This window is where most of the action happens. At the top of this screen is a menu bar similar to the ones you might have seen in other programs (such as Microsoft Word). Figure 1.6 shows this menu bar and the data editor. There are several menus at the top of the screen (e.g. *File*, *Edit* etc.) that can be activated by using the computer mouse to move the on-screen arrow onto the desired menu and then pressing the left mouse button once (pressing this button is usually known as *clicking*). When

you have clicked on a menu, a menu box will appear that displays a list of options that can be activated by moving the on-screen arrow so that it is pointing at the desired option and then clicking with the mouse. Often, selecting an option from a menu makes a window appear; these windows are referred to as *dialog boxes*. When referring to selecting options in a menu I will notate the action using bold type with arrows indicating the path of the mouse (so, each arrow represents placing the on-screen arrow over a word and clicking the mouse's left button). So, for example, if I were to say that you should select the *Save As …* option in the *File* menu, I would write this as select **File⇒Save As …**.

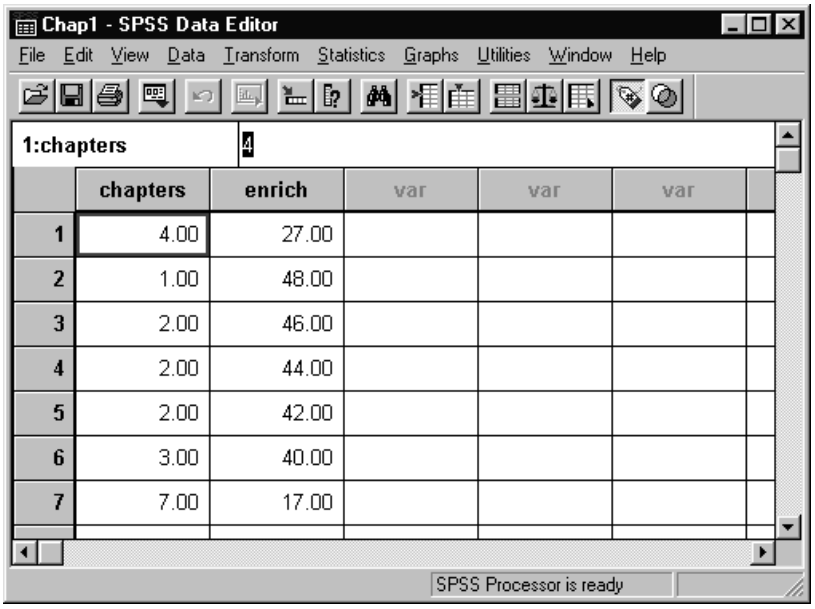
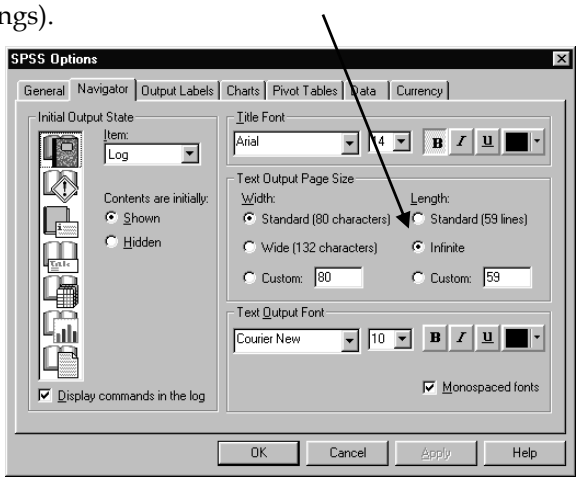


Figure 1.6: The SPSS data editor

Within these menus you will notice that some letters are underlined: these underlined letters represent the *keyboard shortcut* for accessing that function. It is possible to select many functions without using the mouse, and the experienced keyboard user may find these shortcuts faster than manoeuvring the mouse arrow to the appropriate place on the screen. The letters underlined in the menus indicate that the option can be obtained by simultaneously pressing ALT on the keyboard and the underlined letter. So, to access the *Save As…* option, using only the keyboard, you should press ALT and F on the keyboard simultaneously (which activates the *File* menu) then, keeping your finger on the ALT key, press A (which is the underlined letter).

Below is a brief reference guide to each of the menus and some of the options that they contain. This is merely a summary and we will discover the wonders of each menu as we progress through the book.

- **File**: This menu allows you to do general things such as saving data, graphs, or output. Likewise, you can open previously saved files and print graphs, data or output. In essence, it contains all of the options that are customarily found in *File* menus.
- **Edit**: This menu contains edit functions for the data editor. In SPSS for Windows it is possible to *cut* and *paste* blocks of numbers from one part of the data editor to another (which can be very handy when you realize that you've entered lots of numbers in the wrong place). You can also use the *Options* to select various preferences such as the font that is used for the output. The default preferences are fine for most purposes, the only thing you might want to change (for the sake of the environment) is to set the text output page size length of the viewer to infinite (this saves hundreds of trees when you come to print things).



- **Data**: This menu allows you to make changes to the data editor. The important features are *insert variable*, which is used to insert a new variable into the data editor (i.e. add a column); *insert case,* which is used to add a new row of data between two existing rows of data; *split file*, which is used to split the file by a grouping variable (see section 2.4.1); and *select cases*, which is used to run analyses on only a selected sample of cases.
- **Transform**: You should use this menu if you want to manipulate one of your variables in some way. For example, you can use *recode* to change the values of certain variables (e.g. if you wanted to adopt a slightly different coding scheme for some reason). The *compute* function is also useful for transforming data (e.g. you can create a

new variable that is the average of two existing variables). This function allows you to carry out any number of calculations on your variables (see section 6.2.2.1).

- **Analyze:** This menu is called **Statistics** in version 8.0 and earlier. The fun begins here, because the statistical procedures lurk in this menu. Below is a brief guide to the options in the statistics menu that will be used during the course of this book (this is only a small portion of what is available):
  (a) **Descriptive Statistics:** This menu is called **Summarize** in version 8.0 and earlier**.** This menu is for conducting descriptive statistics (mean, mode, median etc.), frequencies and general data exploration. There is also a command called *crosstabs* that is useful for exploring frequency data and performing tests such as chi-square, Fisher's exact test and Cohen's kappa.
  (b) **Compare Means**: This is where you can find *t*-tests (related and unrelated—Chapter 6) and one-way independent ANOVA (Chapter 7).
  (c) **General Linear Model**: This is called *ANOVA Models* in version 6 of SPSS. This menu is for complex ANOVA such as two-way (unrelated, related or mixed), one-way ANOVA with repeated measures and multivariate analysis of variance (MANOVA).
  (d) **Correlate**: It doesn't take a genius to work out that this is where the correlation techniques are kept! You can do bivariate correlations such as Pearson's *R*, Spearman's rho (ρ) and Kendall's tau (τ) as well as partial correlations (see Chapter 3).
  (e) **Regression**: There are a variety of regression techniques available in SPSS. You can do simple linear regression, multiple linear regression (Chapter 4) and more advanced techniques such as logistic regression (Chapter 5).
  (f) **Data Reduction**: You find factor analysis here (Chapter 11).
  (g) **Nonparametric**: There are a variety of non-parametric statistics available such the chi-square goodness-of-fit statistic, the binomial test, the Mann-Whitney test, the Kruskal-Wallis test, Wilcoxon's test and Friedman's ANOVA (Chapter 2).
- **Graphs**: SPSS comes with its own, fairly versatile, graphing package. The types of graphs you can do include: bar charts, histograms, scatterplots, box-whisker plots, pie charts and error bar graphs to name but a few. There is also the facility to edit any graphs to make them look snazzy—which is pretty smart if you ask me.
- **View**: This menu deals with system specifications such as whether you have grid lines on the data editor, or whether you display value labels (exactly what value labels are will become clear later).
- **Window**: This allows you to switch from window to window. So, if you're looking at the output and you wish to switch back to your

data sheet, you can do so using this menu. There are icons to shortcut most of the options in this menu so it isn't particularly useful.

- **Help**: This is an invaluable menu because it offers you on-line help on both the system itself and the statistical tests. Although the statistics help files are fairly useless at times (after all, the program is not supposed to teach you statistics) and certainly no substitute for acquiring a good knowledge of your own, they can sometimes get you out of a sticky situation.

As well as the menus there are also a set of *icons* at the top of the data editor window (see Figure 1.6) that are shortcuts to specific, frequently used, facilities. All of these facilities can be accessed via the menu system but using the icons will save you time. Below is a brief list of these icons and their function:

This icon gives you the option to open a previously saved file (if you are in the data editor SPSS assumes you want to open a data file, if you are in the output viewer, it will offer to open a viewer file).

This icon allows you to save files. It will save the file you are currently working on (be it data or output). If the file hasn't already been saved it will produce the *save data as* dialog box.

This icon activates a dialog box for printing whatever you are currently working on (either the data editor or the output). The exact print options will depend on the printer you use. One useful tip when printing from the output window is to highlight the text that you want to print (by holding the mouse button down and dragging the arrow over the text of interest). In version 7.0 onwards, you can also select parts of the output by clicking on branches in the viewer window (see section 1.2.4). When the *print* dialog box appears remember to click on the option to print only the selected text. Selecting parts of the output will save a lot of trees because by default SPSS will print everything in the output window.

Clicking this icon will activate a list of the last 12 dialog boxes that were used. From this list you can select any box from the list and it will appear on the screen. This icon makes it easy for you to repeat parts of an analysis.

This icon allows you to go directly to a case (i.e. a subject). This is useful if you are working on large data files. For example, if you were analysing a survey with 3000 respondents it would get pretty tedious scrolling down the data sheet to find a

particular subject's responses. This icon can be used to skip directly to a case (e.g. case 2407). Clicking on this icon activates a dialog box that requires you to type in the case number required.

Clicking on this icon will give you information about a specified variable in the data editor (a dialog box allows you to choose which variable you want summary information about).

This icon allows you to search for words or numbers in your data file and output window.

Clicking on this icon inserts a new case in the data editor (so, it creates a blank row at the point that is currently highlighted in the data editor). This function is very useful if you need to add new data or if you forget to put a particular subject's data in the data editor.

Clicking this icon creates a new variable to the left of the variable that is currently active (to activate a variable simply click once on the name at the top of the column).

Clicking on this icon is a shortcut to the **Data⇒Split File ...** function (see section 2.4.1). Social scientists often conduct experiments on different groups of people. In SPSS we differentiate groups of people by using a coding variable (see section 1.2.3.1), and this function lets us divide our output by such a variable. For example, we might test males and females on their statistical ability. We can code each subject with a number that represents their gender (e.g. 1 = female, 0 = male). If we then want to know the mean statistical ability of each gender we simply ask the computer to split the file by the variable **gender**. Any subsequent analyses will be performed on the men and women separately.

This icon shortcuts to the **Data⇒Weight Cases ...** function. This function is necessary when we come to input frequency data (see section 2.8.2) and is useful for some advanced issues in survey sampling.

This icon is a shortcut to the **Data⇒Select Cases ...** function. If you want to analyze only a portion of your data, this is the option for you! This function allows you to specify what cases you want to include in the analysis.

Clicking this icon will either display, or hide, the value labels of any coding variables. We often group people together and use a coding variable to let the computer know that a certain

subject belongs to a certain group. For example, if we coded gender as 1 = female, 0 = male then the computer knows that every time it comes across the value 1 in the **gender** column, that subject is a female. If you press this icon, the coding will appear on the data editor rather than the numerical values; so, you will see the words *male* and *female* in the **gender** column rather than a series of numbers. This idea will become clear in section 1.2.3.1.

1.2.2. Inputting Data

When you first load SPSS it will provide a blank data editor with the title *New Data*. When inputting a new set of data, you must input your data in a logical way. The SPSS data editor is arranged such that *each row represents data from one subject while each column represents a variable*. There is no discrimination between independent and dependent variables: both types should be placed in a separate column. The key point is that each row represents one participant's data. Therefore, any information about that case should be entered across the data editor. For example, imagine you were interested in sex differences in perceptions of pain created by hot and cold stimuli. You could place some people's hands in a bucket of very cold water for a minute and ask them to rate how painful they thought the experience was on a scale of 1 to 10. You could then ask them to hold a hot potato and again measure their perception of pain. Imagine I was a subject. You would have a single row representing my data, so there would be a different column for my name, my age, my gender, my pain perception for cold water, and my pain perception for a hot potato: Andy, 25, male, 7, 10. The column with the information about my gender is a grouping variable: I can belong to either the group of males or the group of females, but not both. As such, this variable is a between-group variable (different people belong to different groups). Therefore, between-group variables are represented by a single column in which the group to which the person belonged is defined using a number (see section 1.2.3.1). Variables that specify to which of several groups a person belongs can be used to split up data files (so, in the pain example you could run an analysis on the male and female subjects separately—see section 2.4.1). The two measures of pain are a repeated measure (all subjects were subjected to hot and cold stimuli). Therefore, levels of this variable can be entered in separate columns (one for pain to a hot stimulus and one for pain to a cold stimulus).

In summary, any variable measured with the same subjects (a repeated measure) should be represented by several columns (each column

representing one level of the repeated measures variable). However, when a between-group design was used (e.g. different subjects were assigned to each level of the independent variable) the data will be represented by two columns: one that has the values of the dependent variable and one that is a coding variable indicating to which group the subject belonged. This idea will become clearer as you learn about how to carry out specific procedures.

The data editor is made up of lots of *cells*, which are just boxes in which data values can be placed. When a cell is active it becomes highlighted with a black surrounding box (as in Figure 1.7). You can move around the data editor, from cell to cell, using the arrow keys ← ↑ ↓ → (found on the right of the keyboard) or by clicking the mouse on the cell that you wish to activate. To enter a number into the data editor simply move to the cell in which you want to place the data value, type the value, then press the appropriate arrow button for the direction in which you wish to move. So, to enter a row of data, move to the far left of the row, type the value and then press → (this process inputs the value and then moves you into the next cell on the left).

1.2.3. Creating a Variable

There are several steps to creating a variable in the SPSS data editor (see Figure 1.7):

- Move the on-screen arrow (using the mouse) to the grey area at the top of the first column (the area labelled *var*.
- Double-click (i.e. click two times in quick succession) with the left button of the mouse.
- A dialog box should appear that is labelled *define variable* (see Figure 1.7).
- In this dialog box there will be a default variable name (something like **var00001**) that you should delete. You can then give the variable a more descriptive name. There are some general rules about variable names, such as that they must be 8 characters or less and you cannot use a blank space. If you violate any of these rules the computer will tell you that the variable name is invalid when you click on OK. Finally, the SPSS data editor is not case sensitive, so if you use capital letters in this dialog box it ignores them. However, SPSS is case sensitive to labels typed into the *Variable Label* part of the *define labels* dialog box (see section 1.2.3.1); these labels are used in the output.
- If you click on OK at this stage then a variable will be created in the data editor for you. However, there are some additional options that you might find useful.

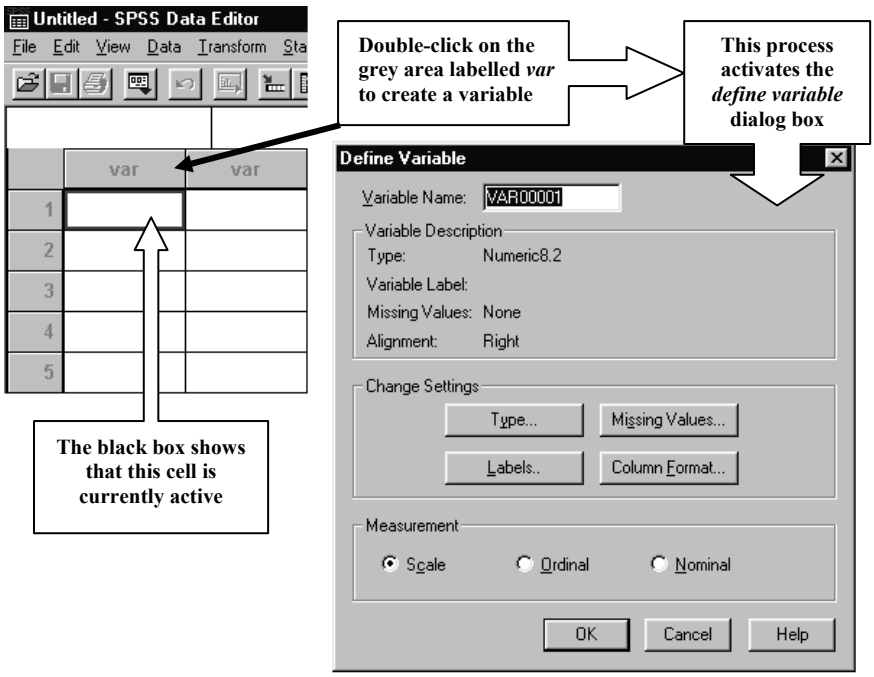


Figure 1.7: Creating a variable

In versions 8 and 9 of SPSS, the *define variable* dialog box contains three options for selecting the level of measurement at which the variable was measured (earlier versions do not have these options). If you are using the variable as a coding variable (next section) then the data are categorical (also called *nominal*) and so you should click on the *Nominal* option. For example, if we asked people whether reading this chapter bores them they will answer *yes* or *no*. Therefore, people fall into two categories: bored and not bored. There is no indication as to exactly how bored the bored people are and therefore the data are merely labels, or categories into which people can be placed. Interval data are scores that are measured on a scale along the whole of which intervals are equal. For example, rather than asking people if they are bored we could measure boredom along a 10-point scale (0 being very interested and 10 being very bored). For data to be interval it should be true that the increase in boredom represented by a change from 3 to 4 along the scale should be the same as the change in boredom represented by a change from 9 to 10. Ratio data have this property, but in addition we should be able to say that someone who had a score of 8 was twice as bored as someone who scored only 4. These two types of data are represented by the *Scale* option. It should be obvious that in some social sciences (notably psychology) it is extremely difficult to establish whether data are interval (can we really tell whether a change on the boredom scale

represents a genuine change in the experience of boredom?). A lower level of measurement is ordinal data, which does not quite have the property of interval data, but we can be confident that higher scores represent higher levels of a construct. We might not be sure that an increase in boredom of 1 on the scale represents the same change in experience between 1 and 2 as it does between 9 and 10. However, we can be confident that someone who scores 9 was, in reality, more bored than someone who scored only 8. These data would be ordinal and so you should select *Ordinal*. The *define variable* dialog box also has four buttons that you can click on to access other dialog boxes and these functions will be described in turn.

1.2.3.1. Creating Coding Variables

In the previous sections I have mentioned coding variables and this section is dedicated to a fuller description of this kind of variable (it is a type of variable that you will use a lot). A coding variable (also known as a grouping variable) is a variable consisting of a series of numbers that represent levels of a treatment variable. In experiments, coding variables are used to represent independent variables that have been measured between groups (i.e. different subjects were assigned to different groups). So, if you were to run an experiment with one group of subjects in an experimental condition and a different group of subjects in a control group, you might assign the experimental group a code of 1, and the control group a code of 0. When you come to put the data into the data editor, then you would create a variable (which you might call **group**) and type in the value 1 for any subjects in the experimental group, and 0 for any subject in the control group. These codes tell the computer that all of the cases that have been assigned the value 1 should be treated as belonging to the same group, and likewise for the subjects assigned the value 0.

There is a simple rule for how variables should be placed in the SPSS data editor: levels of the between-group variables go down the data editor whereas levels of within-subject (repeated measures) variables go across the data editor. We shall see exactly how we put this rule into operation in chapter 6.

To create a coding variable we create a variable in the usual way, but we have to tell the computer which numeric codes we are assigning to which groups. This can be done by using the 

button in the *define variable* dialog box (see Figure 1.7) to open the *define labels* dialog box (see Figure 1.8). In the *define labels* dialog box there is room to give your variable a more descriptive title. For the purposes of the data editor itself, I have already mentioned that variable labels have to be 8 characters or less and that they have to be lower case. However, for the

purposes of the output, it is possible to give our variable a more meaningful title (and this label can also have capital letters and space characters too—great!). If you want to give a variable a more descriptive title then simply click with the mouse in the white space next to where it says *Variable Label* in the dialog box. This will place the cursor in that space, and you can type a title: in Figure 1.8 I have chosen the title *Experimental Condition*. The more important use of this dialog box is to specify group codings. This can be done in three easy steps. First, click with the mouse in the white space next to where it says *Value* (or press ALT and U at the same time) and type in a code (e.g. 1). These codes are completely arbitrary: for the sake of convention people usually use 1, 2 and 3 etc., but in practice you could have a code of 495 if you were feeling particularly arbitrary. The second step is to click the mouse in the white space below, next to where it says *Value Label* (or press ALT and E at the same time) and type in an appropriate label for that group. In Figure 1.8 I have typed in 0 as my code and given this a label of *Control.* The third step is to add this coding to the list by clicking on Add. In Figure 1.8 I have already defined my code for the experimental group, to add the coding for the control group I must click on Add. When you have defined all of your coding values simply click on OK; if you click on OK and have forgotten to add your final coding to the list, SPSS will display a message warning you that any pending changes will be lost. In plain English this simply tells you to go back and click on Add.

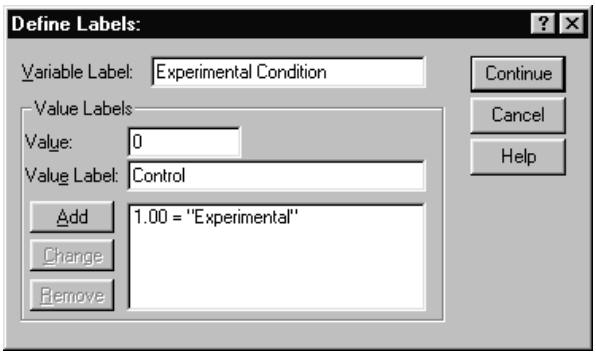


Figure 1.8: Defining coding values in SPSS

Having defined your codings, you can then go to the data editor and type these numerical values into the appropriate column. What is really groovy is that you can get the computer to display the codings themselves, or the value labels that you gave them by clicking on (see Figure 1.9). Figure 1.9 shows how the data should be arranged for a coding variable. Now remember that each row of the data editor represents one subject's data and so in this example it is clear that the first five subjects were in the experimental condition whereas subjects 6–

10 were in the control group. This example also demonstrates why grouping variables are used for variables that have been measured between subjects: because by using a coding variable it is impossible for a subject to belong to more than one group. This situation should occur in a between-group design (i.e. a subject should not be tested in both the experimental and the control group). However, in repeated measures designs (within subjects) each subject is tested in every condition and so we would not use this sort of coding variable (because each subject does take part in every experimental condition).

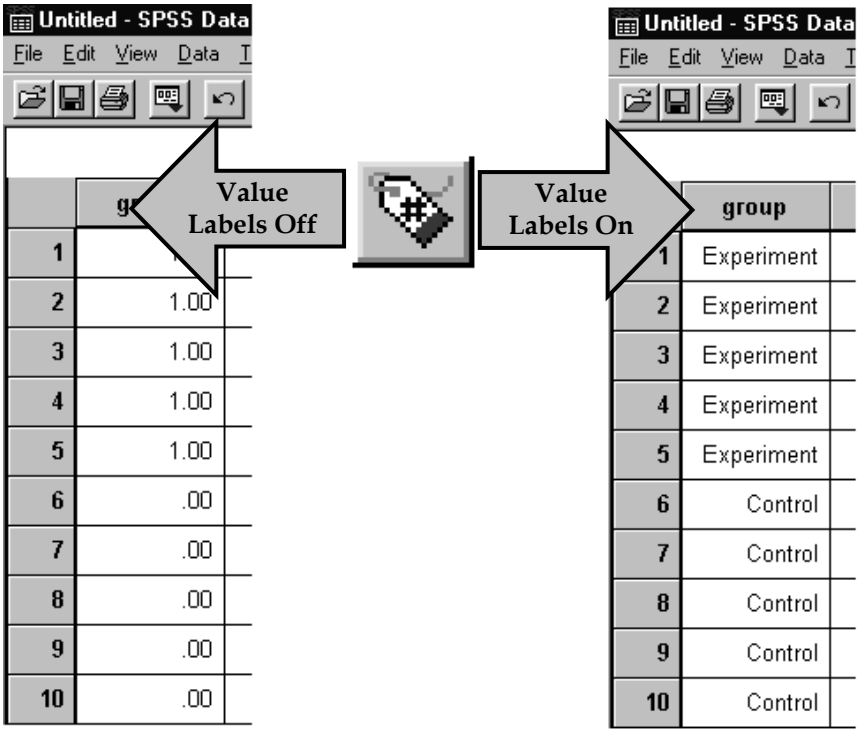


Figure 1.9: Coding values in the data editor with the value labels switched off and on

1.2.3.2. Types of Variables

There are different types of variables that can be used in SPSS. In the majority of cases you will find yourself using numeric variables. These variables are ones that contain numbers and include the type of coding variables that have just been described. However, one of the other options when you create a variable is to specify the type of variable and this is done by clicking on Type... in the *define variable* dialog box. Clicking this button will activate the dialog box in Figure 1.10, which

shows the default settings. By default, a variable is set up to store 8 digits, but you can change this value by typing a new number in the space labelled *Width* in the dialog box. Under normal circumstances you wouldn't require SPSS to retain any more than 8 characters unless you were doing calculations that need to be particularly precise. Another default setting is to have 2 decimal places displayed (in fact, you'll notice by default that when you type in whole numbers SPSS will add a decimal place with two zeros after it—this can be disconcerting initially!). It is easy enough to change the number of decimal places for a given variable by simply replacing the 2 with a new value depending on the level of precision you require.

The *define variable type* dialog box also allows you to specify a different type of variable. For the most part you will use numeric values. However, the other variable type of use is a string variable. A string variable is simply a line of text and could represent comments about a certain subject, or other information that you don't wish to analyze as a grouping variable (such as the subject's name). If you select the string variable option, SPSS lets you specify the width of the string variable (which by default is 8 characters) so that you can insert longer strings of text if necessary.

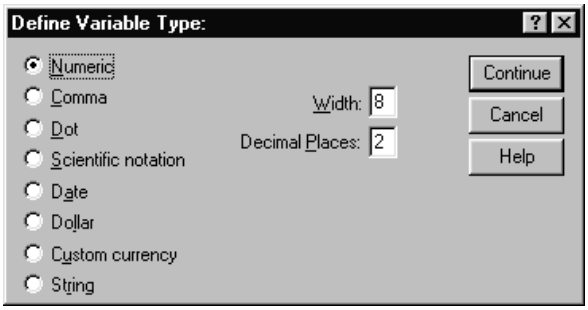


Figure 1.10: Defining the type of variable being used

1.2.3.3. Missing Values

Although as researchers we strive to collect complete sets of data, it is often the case that we have missing data. Missing data can occur for a variety of reasons: in long questionnaires participants accidentally miss out questions; in experimental procedures mechanical faults can lead to a datum not being recorded; and in research on delicate topics (e.g. sexual behaviour) subjects may exert their right not to answer a question. However, just because we have missed out on some data for a subject doesn't mean that we have to ignore the data we do have (although it sometimes creates statistical difficulties). However, we do

need to tell the computer that a value is missing for a particular subject. The principle behind missing values is quite similar to that of coding variables in that we choose a numeric value to represent the missing data point. This value simply tells the computer that there is no recorded value for a participant for a certain variable. The computer then ignores that cell of the data editor (it does not use the value you select in the analysis). You need to be careful that the chosen code doesn't correspond with any naturally occurring data value. For example, if we tell the computer to regard the value 9 as a missing value and several subjects genuinely scored 9, then the computer will treat their data as missing when, in reality, it is not.

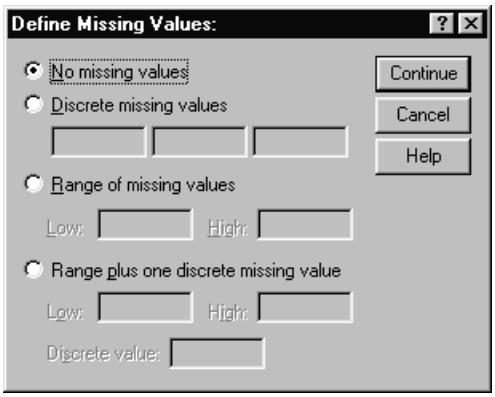
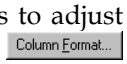


Figure 1.11: Defining missing values

To specify missing values you simply click on Missing Values... in the *define variable* dialog box to activate the *define missing values* dialog box (see Figure 1.11). By default SPSS assumes that no missing values exist but if you do have data with missing values you can choose to define them in one of three ways. The first is to select discrete values (by clicking on the circle next to where it says *Discrete missing values*) which are single values that represent missing data. SPSS allows you to specify up to three discrete values to represent missing data. The reason why you might choose to have several numbers to represent missing values is that you can assign a different meaning to each discrete value. For example, you could have the number 8 representing a response of 'not applicable', a code of 9 representing a 'don't know' response, and a code of 99 meaning that the subject failed to give any response. As far as the computer is concerned it will ignore any data cell containing these values; however, using different codes may be a useful way to remind you of why a particular score is missing. Usually, one discrete value is enough and in an experiment in which attitudes are measured on a 100-point scale (so scores vary from 1 to 100) you might choose 999 to represent missing values because this value cannot occur in the data that

have been collected. The second option is to select a range of values to represent missing data and this is useful in situations in which it is necessary to exclude data falling between two points. So, we could exclude all scores between 5 and 10. The final option is to have a range of values and one discrete value.

1.2.3.4. Changing the Column Format

The final option available to us when we define a variable is to adjust the formatting of the column within the data editor. Click on 

in the *define variable* dialog box and the dialog box in Figure 1.12 will appear. The default option is to have a column that is 8 characters wide with all numbers and text aligned to the right-hand side of the column. Both of these defaults can be changed: the column width by simply deleting the value of 8 and replacing it with a value suited to your needs, and the alignment by clicking on one of the deactivated circles (next to either *Left* or *Center*). It is very useful to adjust the column width when you have a coding variable with value labels that exceed 8 characters in length.

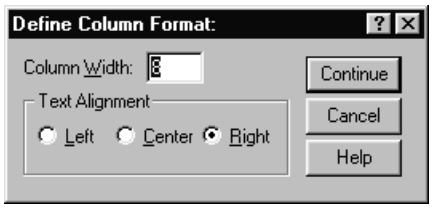


Figure 1.12: Defining the format of the column

1.2.4. The Output Viewer

Alongside the main SPSS window, there is a second window known as the output viewer (or *output navigator* in versions 7.0 and 7.5). In earlier versions of SPSS this is simply called the output window and its function is, in essence, the same. However, whereas the output window of old displayed only statistical results (in a very bland font I might add), the new, improved and generally amazing output viewer will happily display graphs, tables and statistical results and all in a much nicer font. Rumour has it that future versions of SPSS will even include a tea-making facility in the output viewer (I live in hope!).

Figure 1.13 shows the basic layout of the output viewer. On the right-hand side there is a large space in which the output is displayed. SPSS displays both graphs and the results of statistical analyses in this part of the viewer. It is also possible to edit graphs and to do this you simply

double-click on the graph you wish to edit (this creates a new window in which the graph can be edited). On the left-hand side of the output viewer there is a tree diagram illustrating the structure of the output. This tree diagram is useful when you have conducted several analyses because it provides an easy way of accessing specific parts of the output. The tree structure is fairly self-explanatory in that every time you conduct a procedure (such as drawing a graph or running a statistical procedure), SPSS lists this procedure as a main heading.

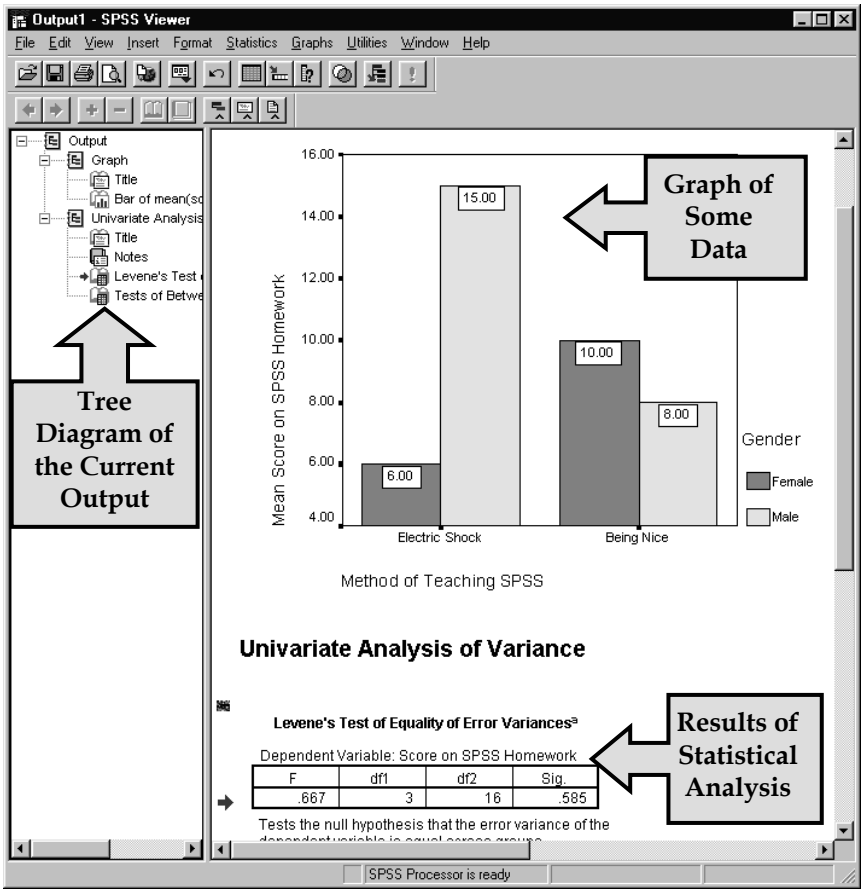


Figure 1.13 The output viewer

In Figure 1.13 I conducted a graphing procedure and then conducted a univariate analysis of variance (ANOVA) and so these names appear as main headings. For each procedure there are a series of sub-procedures, and these are listed as branches under the main headings. For example, in the ANOVA procedure there are a number of sections to the output such as a Levene's test (which tests the assumption of homogeneity of

variance) and the between-group effects (i.e. the *F*-test of whether the means are significantly different). You can skip to any one of these sub-components of the ANOVA output by clicking on the appropriate branch of the tree diagram. So, if you wanted to skip straight to the between-group effects you should move the on-screen arrow to the left-hand portion of the window and click where it says *Tests of Between-Subjects Effects*. This action will highlight this part of the output in the main part of the viewer. You can also use this tree diagram to select parts of the output (which is useful for printing). For example, if you decided that you wanted to print out a graph but you didn't want to print the whole output, you can click on the word *Graph* in the tree structure and that graph will become highlighted in the output. It is then possible through the print menu to select to print only the selected part of the output. In this context it is worth noting that if you click on a main heading (such as *Univariate Analysis of Variance*) then SPSS will highlight not only that main heading but all of the sub-components as well. This is extremely useful when you want to print the results of a single statistical procedure.

There are a number of icons in the output viewer window that help you to do things quickly without using the drop-down menus. Some of these icons are the same as those described for the data editor window so I will concentrate mainly on the icons that are unique to the viewer window.

As with the data editor window, this icon activates the print menu. However, when this icon is pressed in the viewer window it activates a menu for printing the output. When the print menu is activated you are given the default option of printing the whole output, or you can choose to select an option for printing the output currently visible on the screen, or most useful is an option to print a selection of the output. To choose this last option you must have already selected part of the output (see above).

This icon returns you to the data editor in a flash!

This icon takes you to the last output in the viewer (so, it returns you to the last procedure you conducted).

This icon *promotes* the currently active part of the tree structure to a higher branch of the tree. For example, in Figure 1.13 the *Tests of Between-Subjects Effects* are a sub-component under the heading of *Univariate Analysis of Variance*. If we wanted to promote this part of the output to a higher level (i.e. to make it a main heading) then this is done using this icon.

This icon is the opposite of the above in that it *demotes* parts of the tree structure. For example, in Figure 1.13 if we didn't want the *Univariate Analysis of Variance* to be a unique section we could select this heading and demote it so that it becomes part of the previous heading (the *Graph* heading). This button is useful for combining parts of the output relating to a specific research question.

This icon collapses parts of the tree structure, which simply means that it hides the sub-components under a particular heading. For example, in Figure 1.13 if we selected the heading *Univariate Analysis of Variance* and pressed this icon, all of the sub-headings would disappear. The sections that disappear from the tree structure don't disappear from the output itself; the tree structure is merely condensed. This can be useful when you have been conducting lots of analyses and the tree diagram is becoming very complex.

This icon expands any collapsed sections. By default all of the main headings are displayed in the tree diagram in their expanded form. If, however, you have opted to collapse part of the tree diagram (using the icon above) then you can use this icon to undo your dirty work.

This icon and the following one allow you to show and hide parts of the output itself. So, you can select part of the output in the tree diagram and click on this icon and that part of the output will disappear. It isn't erased, but it is hidden from view. So, this icon is similar to the collapse icon listed above except that it affects the output rather than the tree structure. This is useful for hiding less relevant parts of the output.

This icon undoes the previous one, so if you have hidden a selected part of the output from view and you click on this icon, that part of the output will reappear. By default, all parts of the output are shown and so this icon is not active: it will become active only once you have hidden part of the output.

Although this icon looks rather like a paint roller, it unfortunately does not paint the house for you. What it does do is to insert a new heading into the tree diagram. For example, if you had several statistical tests that related to one of many research questions you could insert a main heading and then demote the headings of the relevant analyses so that they all fall under this new heading.

Assuming you had done the above, you can use this icon to provide your new heading with a title. The title you type in will actually appear in your output. So, you might have a heading like 'Research Question number 1' which tells you that the analyses under this heading relate to your first research question.

This final icon is used to place a text box in the output window. You can type anything into this box. In the context of the previous two icons, you might use a text box to explain what your first research question is (e.g. 'My first research question is whether or not boredom has set in by the end of the first chapter of my book. The following analyses test the hypothesis that boredom levels will be significantly higher at the end of the first chapter than at the beginning').

1.2.5. Saving Files

Although most of you should be familiar with how to save files in Windows it is a vital thing to know and so I will briefly describe what to do. To save files simply use the  icon (or use the menus: **File⇒Save** or **File⇒Save As…**). If the file is a new file, then clicking this icon will activate the *Save As ...* dialog box (see Figure 1.14). If you are in the data editor when you select *Save As …* then SPSS will save the data file you are currently working on, but if you are in the viewer window then it will save the current output.

There are a number of features of the dialog box in Figure 1.14. First, you need to select a location at which to store the file. Typically, there are two types of locations where you can save data: the hard drive (or drives) and the floppy drive (and with the advent of rewritable CD-ROM drives, zip drives, jaz drives and the like you may have many other choices of location on your particular computer). The first thing to do is select either the floppy drive, by double clicking on  , or the hard drive, by double clicking on  . Once you have chosen a main location the dialog box will display all of the available folders on that particular device (you may not have any folders on your floppy disk in which case you can create a folder by clicking on ). Once you have selected a folder in which to save your file, you need to give your file a name. If you click in the space next to where it says *File name*, a cursor will appear and you can type a name of up to ten letters. By default, the file will be saved in an SPSS format, so if it is a data file it will have the file extension *.sav*, and if it is a viewer document it will have the file extension *.spo*. However, you can save data in different formats such as

Microsoft Excel files and tab-delimited text. To do this just click on where it says *Save as type* and a list of possible file formats will be displayed. Click on the file type you require. Once a file has previously been saved, it can be saved again (updated) by clicking . This icon appears in both the data editor and the viewer, and the file saved depends on the window that is currently active. The file will be saved in the location at which it is currently stored.

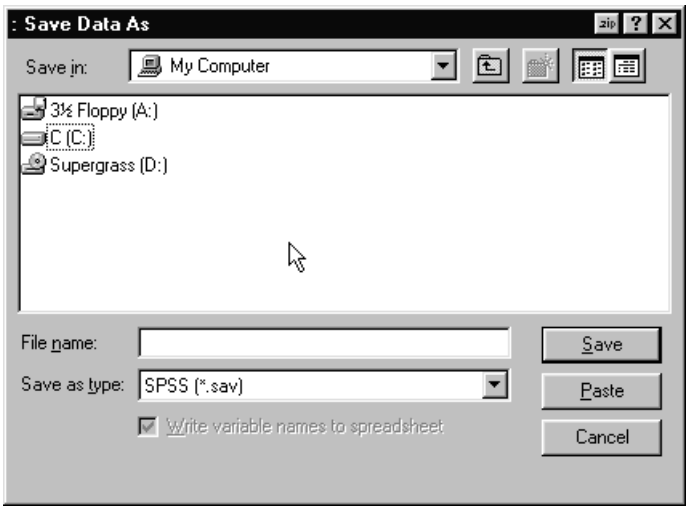


Figure 1.14: The s*ave data as* dialog box

1.2.6. Retrieving a File

Throughout this book you will work with data files that have been provided on a floppy disk. It is, therefore, important that you know how to load these data files into SPSS. The procedure is very simple. To open a file, simply use the icon (or use the menus: **File⇒Open**) to activate the dialog box in Figure 1.15. First, you need to find the location at which the file is stored. If you are loading a file from the floppy disk then access the floppy drive by clicking on where it says *Look in* and a list of possible location drives will be displayed. Once the floppy drive has been accessed you should see a list of files and folders that can be opened. As with saving a file, if you are currently in the data editor then SPSS will display only SPSS data files to be opened (if you are in the viewer window then only output files will be displayed). You can open a folder by double-clicking on the folder icon. Once you have tracked down the required file you can open it either by selecting it with the mouse and then clicking on Open, or by double-clicking on the icon next to the file you want (e.g. double-clicking on). The data/output

will then appear in the appropriate window. If you are in the data editor and you want to open a viewer file, then click on where it says *Files of type* and a list of alternative file formats will be displayed. Click on the appropriate file type (viewer document (**.spo*), Excel file (**.xls*), text file (**.dat, *.txt*)) and any files of that type will be displayed for you to open.

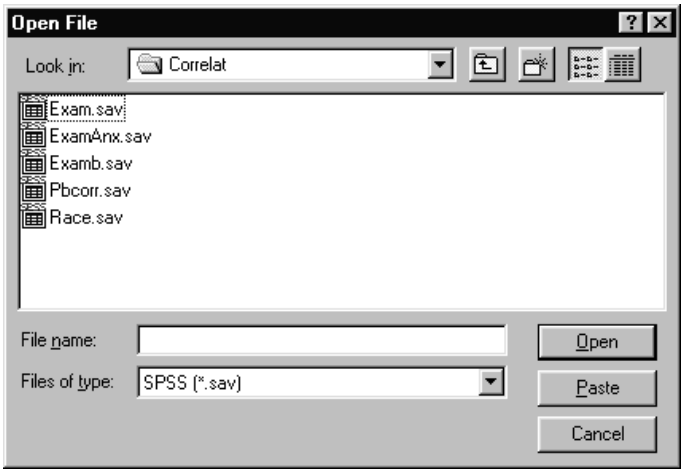


Figure 1.15: Dialog box to open a file

1.3. Further Reading

Einspruch, E. L. (1998). *An introductory guide to SPSS for Windows*. Thousand Oaks, CA: Sage.

Foster, J. J. (1998). *Data analysis using SPSS for Windows: a beginner's guide*. London: Sage.