1 Introduction to How to Display Data and the Scatter Plot

1.1 INTRODUCTION

The initial chapters of the book are related to data and how it should be portrayed. Often useful data is poorly served by poor data displays, which, while they might look attractive, are actually very difficult to interpret and mask trends in the data.

It has been said many times that 'a picture is worth a thousand words' and this 'original' thought has been attributed to at least two historical heavyweights (Mark Twain and Benjamin Disraeli). While tables of figures can be hard or difficult to interpret, some form of pictorial presentation of the data enables management to gain an immediate indication of the key issues highlighted within the data set. It enables senior management to identify some of the major trends within a complex data set without the requirement to undertake detailed mathematical work. It is important that the author of a pictorial presentation of data follows certain basic rules when plotting data to avoid introducing bias, either accidentally or deliberately, or producing inappropriate or misleading representations of the original data.

When asked to prepare a report for management which is either to analyse or present some data that has been accumulated, the first step is often to present it in a tabular format and then produce a simple presentation of the information, frequently referred to as a plot. It is claimed that a plot is interpreted with more ease than the actual data set out in some form of a table. Many businesses have standardised reporting packages, which enable data to be quickly transformed into a pictorial presentation, offering a variety of potential styles. While many of these software packages produce plots, they should be used with care. Just because a computer produces a graph does not mean it is an honest representation of the data. The key issue for the author of such a plot is to see if the key trends inherent in the data are better highlighted by the pictorial representation. If this is not the case then an alternative approach should be adopted.

Whenever you are seeking to portray data there are always a series of choices to be made:

1. What is the best way to show the data?
2. Can I amend the presentation so that key trends in the data are more easily seen?
3. Will the reader understand what the presentation means?

Often people just look at the options available on their systems and choose the version that looks the prettiest, without taking into consideration the best way in which the material should be portrayed.

Many people are put off by mathematics and statistics – perhaps rightly in many cases since the language and terminology are difficult to penetrate. The objective of good data presentation is not to master all the mathematical techniques, but rather to use those that are appropriate, given the nature of what you are trying to achieve.

In this chapter we consider some of the most commonly used graphical presentational approaches and try to assist you in establishing which is most appropriate for the particular

data set that is to be presented. We start with some of the simplest forms of data presentation, the scatter plot, the matrix plot and the histogram.

1.2 SCATTER PLOTS

Scatter plots are best used for data sets in which there is likely to be some form of relationship or association between two different elements included within the data. These different elements are generally referred to as *variables*. Scatter plots use horizontal and vertical axes to enable the author to input the information into the scatter plot, or, in mathematical jargon, to plot the various *data points*. This style of presentation effectively shows how one variable affects another. Such a relationship will reveal itself by highlighting any trend that will be apparent to the reader from a review of the chart.

1.3 DATA IDENTIFICATION

A scatter plot is a plot of the values of Y on the vertical axis, or *ordinate*, taken against the corresponding values of X on the horizontal axis, or *abscissa*. Here the letters X and Y are taken to replace the actual variables, which might be something like losses arising in a month (Y) against time (X).

- X is usually the independent variable.
- Y is usually the response or dependent variable that may be related to the independent variable.

We shall explain these terms further through consideration of a simple example.

1.3.1 An example of salary against age

Figure 1.1 presents the relationship between salary and age for 474 employees of a company. This type of data would be expected to show some form of trend since, as the staff gains experience, you would expect their value to the company to increase and therefore their salary to also increase.

The raw data were obtained from personnel records. The first individual sampled was 28.50 years old and had a salary of £16,080. To put this data onto a *scatter plot* we insert age onto the horizontal axis and salary onto the vertical axis. The different entries onto the plot are the 474 combinations of age and salary resulting from a selection of 474 employees, with each individual observation being a single point on the chart.

This figure shows that in fact for this company there is no obvious relation between salary and age. From the plot it can be seen that the age range of employees is from 23 to 65. It can also be seen that a lone individual earns a considerably higher salary than all the others and that starters and those nearing retirement are actually on similar salaries.

You will see that the length of the axis has been chosen to match the range of the available data. For instance, no employees were younger than 20 and none older than 70. It is not essential that the axis should terminate at the origin. The objective is to find the clearest way to show the data, so making best use of the full space available clearly makes sense. The process of starting from 20 for age and 6,000 for salaries is called *truncation* and enables the actual data to cover the whole of the area of the plot, rather than being stuck in one quarter.

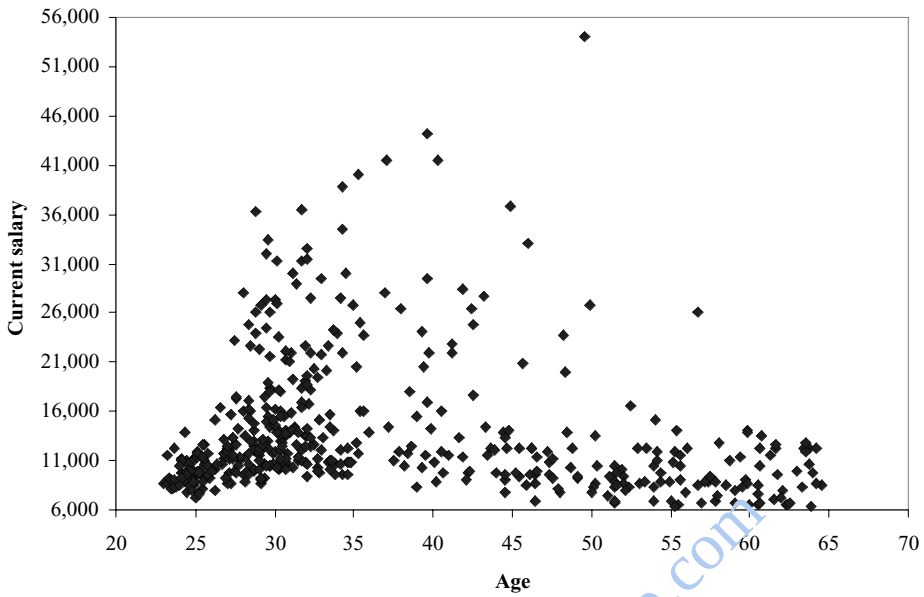

Figure 1.1 Scatter plot of current salary against age.

1.4 WHY DRAW A SCATTER PLOT?

Having drawn the plot it is necessary to interpret it. The author should do this before it is passed to any user. The most obvious relationship between the variables X and Y would be a straight line or a linear one. If such a relationship can be clearly demonstrated then it will be of assistance to the reader if this is shown explicitly on the scatter plot. This procedure is known as *linear regression* and is discussed in Chapter 13.

An example of data where a straight line would be appropriate would be as follows. Consider a company that always charges out staff at £1,000 per day, regardless of the size of the contract and never allows discounts. That would mean that a one-day contract would cost £1,000 whereas a 7-day contract would cost £7,000 (seven times the amount per day). If you were to plot 500 contracts of differing lengths by taking the value of the contract against the number of days, then this would represent a straight line scatter plot.

In looking at data sets, various questions may be posed. Scatter plots can provide answers to the following questions:

- Do two variables X and Y appear to be related? Given what the scatter plot portrays, could this be used to give some form of prediction of the potential value for Y that would correspond to a potential value of X?
- Are the two variables X and Y actually related in a straight line or linear relationship? Would a straight line fit through the data?
- Are the two variables X and Y instead related in some non-linear way? If the relationship is non-linear, will any other form of line be appropriate that might enable predictions of Y to be made? Might this be some form of distribution? If we are able to use a distribution this will enable us to use the underlying mathematics to make predictions about the variables. This is discussed in Chapter 7.

- Does the amount by which Y changes depend on the amount by which X changes? Does the coverage or spread in the Y values depend on the choice of X? This type of analysis always helps to gain an additional insight into the data being portrayed.
- Are there data points that sit away from the majority of the items on the chart, referred to as *outliers*? Some of these may highlight errors in the data set itself that may need to be rechecked.

1.5 MATRIX PLOTS

Scatter plots can also be combined into multiple plots on a single page if you have more than two variables to consider. This type of analysis is often seen in investment analysis, for example, where there could be a number of different things all impacting upon the same data set. Multiple plots enable the reader to gain a better understanding of more complex trends hidden within data sets that include more than two variables. If you wish to show more than two variables on a scatter plot grid, or matrix, then you still need to generate a series of pairs of data to input into the plots. Figure 1.2 shows a typical example.

In this example four variables (a, b, c, d) have been examined by producing all possible scatter plots. Clearly while you could technically include even more variables, this would make the plot almost impossible to interpret as the individual scatter plots become increasingly small.

Returning to the analysis we set out earlier of salary and age (Figure 1.1), let us now differentiate between male salaries and female salaries, by age. This plot is shown as Figure 1.3.

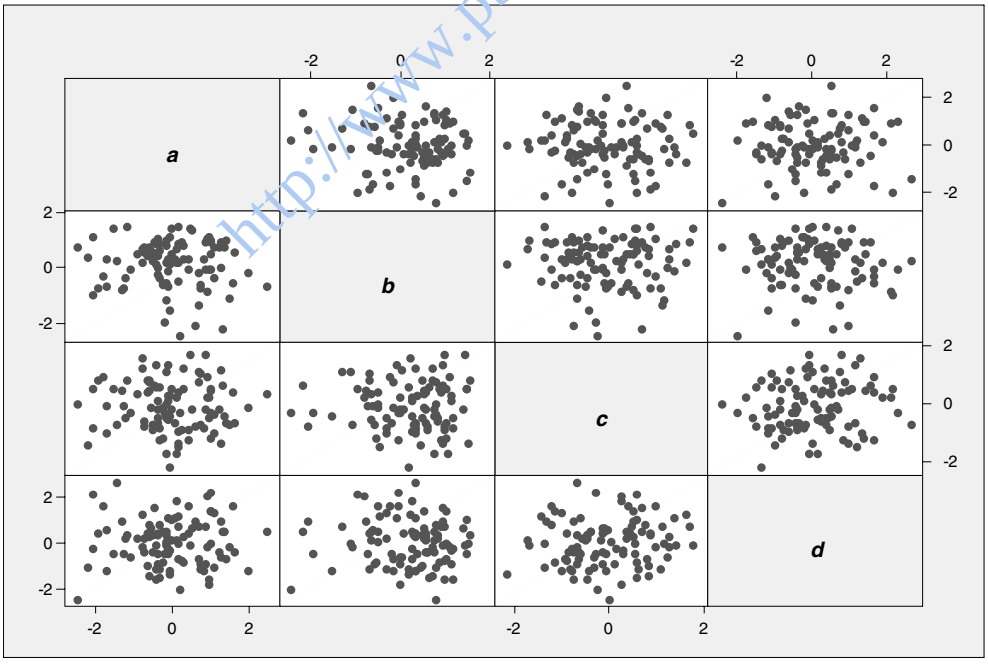

Figure 1.2 Example of a matrix plot.

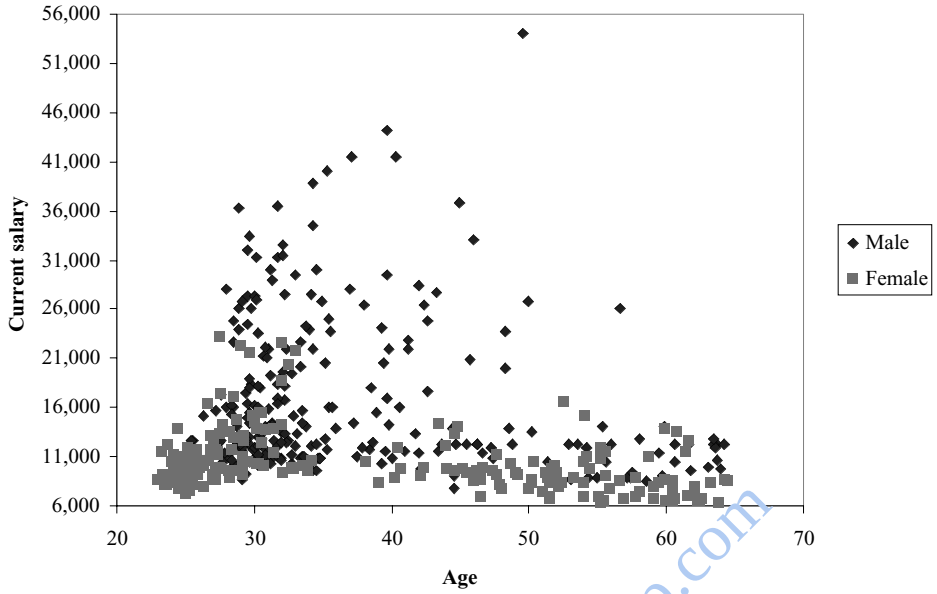

Figure 1.3 Scatter plot of current salary against age, including the comparison of male and female workers.

1.5.1 An example of salary against age: Revisited

It now becomes very clear that women have the majority of the lower paid jobs and that their salaries appear to be even less age dependent than those of men. This type of analysis would be of interest to the Human Resources function of the company to enable it to monitor compliance with legislation on sexual discrimination, for example. Of course there may be a range of other factors that need to be considered, including differentiating between full- and part-time employment by using either another colour or plotting symbol.

It is the role of the data presentation to facilitate the highlighting of trends that might be there. It is then up to the user to properly interpret the story that is being presented.

In summary the scatter plot attempts to uncover any relationship in the data. 'Relationship' means that there may be some structural association between two variables X and Y. Scatter plots are a useful diagnostic tool for highlighting whether there is any form of potential association, but they cannot in themselves suggest an underlying cause-and-effect mechanism. A scatter plot can never prove cause and effect; this needs to be achieved through further detailed investigation, which should use the scatter plot to set out the areas where the investigation into the underlying data should commence.