Chapter 1 **Introduction to data display**

1.1 Introduction

This book has arisen from our extensive experience as researchers and teachers of medical statistics. We have frequently been appalled by the poor quality of data display even in major medical journals. While there is already a wealth of information about how to display data, it is scattered across many sources. Our purpose in writing this book is to bring together this information into a single volume and provide clear accessible advice for both researchers, and students alike.

Well-displayed data can clearly illuminate and enhance the interpretation of a study, while badly laid out data and results can obscure the message or at worst seriously mislead. Although the appropriate display of data in tables and graphs is an essential part of any report, paper or presentation, little space is devoted to it in the majority of textbooks. The purpose of this book is to address this deficit and give clear guidelines on appropriate methods for displaying quantitative information, using both graphs and tables.

There are many different types of graph and table available for displaying data; their purposes will be outlined in subsequent chapters. This chapter will outline the reasons why it is important to get display right, good principles to adhere to when displaying data and the types of data that will be covered in the rest of the book. The second chapter will cover some of the many ways in which the display of information can be badly done and the following chapters will then unpick these, and give clear guidance on how to do it well.

1.2 Types of data

To display data appropriately, one must first understand what types of data there are, as this determines the best method of displaying them. Figure 1.1 shows a basic hierarchy of data types, although there are others. Data are either *categorical* or *quantitative*. Data are described as categorical when they can

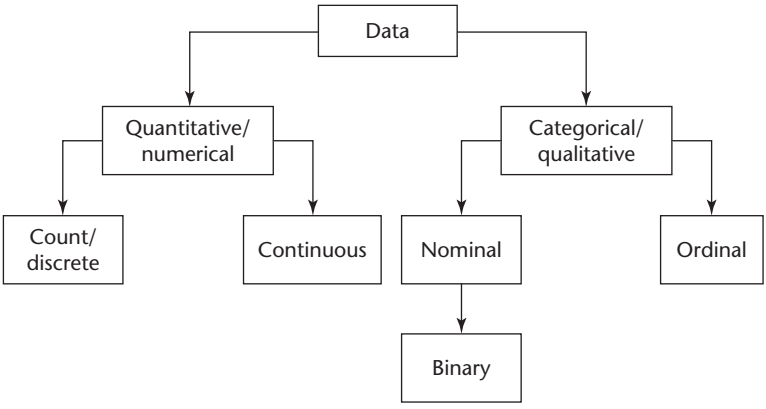

Figure 1.1 Types of data.

be categorised into distinct groups, such as ethnic group or disease severity. Although categorical data may be coded numerically, for example gender may be coded 1 for male and 2 for female, these codes have no intrinsic numerical value; it would be nonsense to calculate an average gender. Categorical data can be divided into either *nominal* or *ordinal*. Nominal data have no natural ordering and examples include eye colour, marital status and area of residence. *Binary* data is a special subcategory of nominal data, where there are only two possible values, for example male/female, yes/no, dead/alive. Ordinal data occurs when there can be said to be a natural ordering of the data values, such as better/same/worse, grades of breast cancer and social class.

Quantitative data can be either counted or continuous. *Count* data are also known as discrete data and, as the name implies, occur when the data can be counted, such as the number of children in a family or the number of visits to a GP in a year. Count data are similar to categorical data as they can only take discrete whole numbers. *Continuous* data are data that can be measured and they can take any value on the scale on which they are measured; they are limited only by the scale of measurement and examples include height, weight and blood pressure.

1.3 Where to start?

When displaying information visually, there are three questions one will find useful to ask as a starting point (Box 1.1). Firstly and most importantly, it is vital to have a clear idea about what is to be displayed; for example, is it important to demonstrate that two sets of data have different distributions or

Box 1.1 Useful questions to ask when considering how to display information

- What do you want to show?
- What methods are available for this?
- Is the method chosen the best? Would another have been better?

that they have different mean values? Having decided what the main message is, the next step is to examine the methods available and to select an appropriate one. Finally, once the chart or table has been constructed, it is worth reflecting upon whether what has been produced truly reflects the intended message. If not, then refine the display until satisfied; for example if a chart has been used would a table have been better or vice versa? This book will help you answer these questions and provide you with the means to best display your data.

1.4 Recommendations for the presentation of numbers

When summarising categorical data, both frequencies and percentages can be used. However, if percentages are reported, it is important that the denominator (i.e. total number of observations) is given. To summarise continuous numerical data, one should use the mean and standard deviation, or if the data have a skewed distribution use the median and range or interquartile range. However, for all of these calculated quantities it is important to state the total number of observations on which they are based.

In the majority of cases it is reasonable to treat count data, such as number of children in a family or number of visits to the GP in a year, as if they were continuous, at least as far as the statistical analysis goes. Ideally there should be a large number of different possible values, but in practice this is not always necessary. However, where ordered categories are numbered, such as stage of disease or social class, the temptation to treat these numbers as statistically meaningful must be resisted. For example, it is not sensible to calculate the average social class of a sample or stage of cancer for a group of patients, and in such cases the data should be treated in statistical analyses as if they are ordered categories.[1]

Numerical precision should be consistent throughout and summary statistics such as means and standard deviations should not have more than one extra decimal place (or significant digit) compared to the raw data. Spurious precision should be avoided although when certain measures are to be used for further calculations or when presenting the results of analyses, greater precision may sometimes be appropriate.[2]

1.5 Recommendations for presenting data and results in tables

There are a few basic rules of good presentation, both within the text of a document or presentation, and within tables, as outlined in Box 1.2. Tufte, in 1983, outlined a fundamental principle: always try to get as much information into a figure consistent with legibility. In other words, one should maximise the ratio of the amount of information given to the amount of ink used.[3] Tables, including column and row headings, should be clearly labelled and a brief summary of the contents of a table should always be given in words, either as part of the title or in the main body of the text.

Box 1.2 Recommendations when presenting data and results in tables

- The amount of information should be maximised for the minimum amount of ink.
- Numerical precision should be consistent throughout a paper or presentation, as far as possible.
- Avoid spurious accuracy. Numbers should be rounded to two effective digits.
- Quantitative data should be summarised using either the mean and standard deviation (for symmetrically distributed data) or the median and interquartile range or range (for skewed data). The number of observations on which these summary measures are based should be included.
- Categorical data should be summarised as frequencies and percentages. As with quantitative data, the number of observations should be included.
- Each table should have a title explaining what is being displayed and columns and rows should be clearly labelled.
- Solid lines in tables should be kept to a minimum.
- Where variables have no natural ordering, rows and columns should be ordered by size.

Solid lines should not be used in a table except to separate labels and summary measures from the main body of the data. However, their use should be kept to a minimum, particularly vertical gridlines, as they can interrupt eye movements, and thus the flow of information. White space can be used to separate data, such as different variables, from each other.[4]

The information in tables is easier to comprehend if the columns (rather than the rows) contain similar information, such as means or standard deviations, as it is easier to scan down a column than across a row.[4] However, it

is not always easy to do this, particularly when the information for several variables is contained in the same table and comparisons are to be made between different groups. This will be covered in more detail in Chapter 6. In addition, where there is no natural ordering of the rows (or indeed columns), they should be ordered by size (category with the highest frequency first, lowest frequency last) as this helps the reader to scan for patterns and exceptions in the data.[4] Table 1.1a shows the frequency distribution for marital status for 226 patients with leg ulcers who were recruited to a study to assess the effectiveness of specialist leg ulcers clinics compared to usual care.[5] The categories in this table are ordered alphabetically, whereas in Table 1.1b the marital status categories are ordered by frequency making it much easier to interpret than Table 1.1a.

Table 1.1 Marital status of 226 patients with leg ulcer recruited to a study to assess the effectiveness of specialist leg ulcer clinics using 4-layer compression bandaging compared to usual care[5]

	Frequency	Percent
(a) Unordered rows		
Divorced/separated	11	4.9
Married	104	46.0
Single	25	11.1
Widowed	86	38.1
Total	226	100.0
(b) Ordered rows		
Married	104	46.0
Widowed	86	38.1
Single	25	11.1
Divorced/separated	11	4.9
Total	226	100.0

1.6 Recommendations for construction of graphs

Box 1.3 outlines some basic recommendations for the construction and use of figures to display data. As with tables, a fundamental principle is that graphs should maximise the amount of information presented for the minimum amount of ink used.[3] Good graphs have the following four features in common: clarity of message, simplicity of design, clarity of text, and integrity of intention and action.[6] A graph should have a title explaining what is displayed and axes should be clearly labelled; if it is not immediately

Box 1.3 Guidelines for constructing graphs

- The amount of information should be maximised for the minimum amount of ink.
- Each graph should have a title explaining what is being displayed.
- Axes should be clearly labelled.
- Gridlines should be kept to a minimum.
- Avoid three-dimensional graphs as these can be difficult to read.
- The number of observations should be included.

obvious how many individuals the graph is based upon, this should also be stated. Gridlines should be kept to a minimum as they act as a distraction and can interrupt the flow of information. When using graphs for presentation purposes care must be taken to ensure that they are not misleading; an excellent exposition of the ways in which graphs can be used to mislead can be found in Huff.[7] Figure 1.2 shows a bar chart of the marital status data from Table 1.1 displayed using these principles. It includes a clear title (with the sample size), labelled axes, no gridlines and the marital status categories are ordered by their frequency.

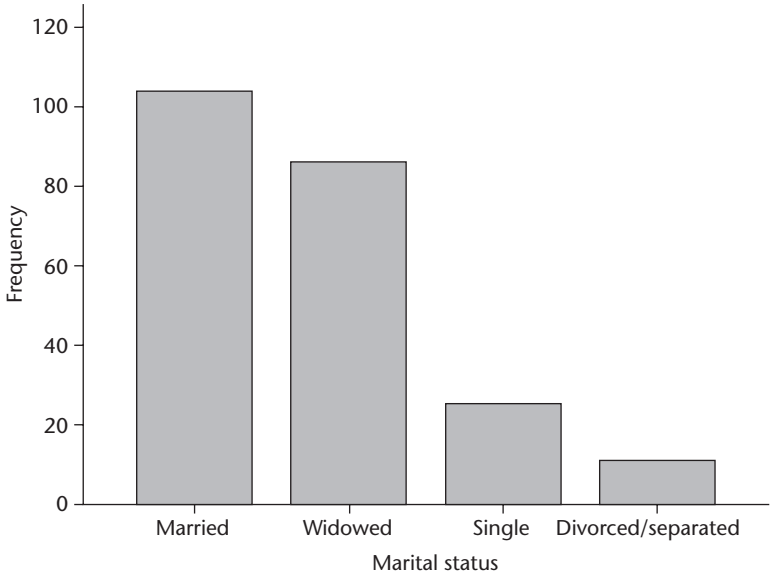

Figure 1.2 Bar chart of marital status for 226 patients recruited to the leg ulcer Study.[5]

1.7 Table or graph?

A fundamental point to consider is whether to use a table or graph (see Box 1.4). We define a table as a display of numbers in a rectangular grid, and a graph or chart as a picture in which the numbers are represented by points or lines. Plotting data is a useful first stage to any analysis and will show extreme observations together with any discernible patterns. In addition the relative sizes of categories are easier to see in a diagram (bar chart or pie chart) than in a table. Graphs are useful as they can be assimilated quickly, and are particularly helpful when presenting information to an audience. Tables can be useful for displaying information about many variables at once, while graphs can be useful for showing multiple observations on groups or individuals. Although there are no hard and fast rules about when to use a graph and when to use a table, in the context of a report or a paper it is often best to use tables so that the reader can scrutinise the numbers directly. Thus, for a talk or presentation, Figure 1.2 would be a good method of displaying the data. However, for a printed report or paper, Table 1.1b conveys the data more accurately and succinctly.

Box 1.4 Graph or table

Graph	Table
Usually better in presentations	Often better in papers
Can often show all the data	Usually can only show summaries
Usually show only a few variables	Better for multiple variables

1.8 Software

No single package can draw all the graphs necessary for displaying data. Simple graphs can be drawn in *Microsoft Excel*. However, you should be aware that some of the default settings are not ideal (see Chapter 2). For more complex graphs, any of the major statistical packages – *STATA*, *SPSS* or *SAS* – are useful. *S-Plus* is particularly good for superimposing several graphs into a single figure. In drawing the graphs for this book a variety of packages were used, although many were drawn in the specialist package Sigmaplot (Systat Software Inc 24, Vista Centre, 50, Salisbury Road, Hounslow, TW4 6JQ, London). Packages change regularly so we have not given explicit instructions on how to draw individual graphs in particular packages. The book simply outlines good practice for displaying data.

Summary

- The purpose of any attempt to present data and results, either in a presentation or on paper is to communicate with an audience.
- In the following chapters key methods using both graphs and tables will be outlined so that by the end of this book you should have the skills and knowledge to display your data appropriately.
- In addition, you will be able to distinguish between bad graphs and good graphs and know how to transform the former into the latter and you should be able to distinguish between a bad table and a good table and be able to transform the former into the latter.
- A variety of software packages is available for drawing graphs. In order to draw all of the graphs outlined in this book you will need to use several packages.

References

1 Freeman JV, Walters SJ. Examining relationships in quantitative data (inferential statistics). In: Gerrish K, Lacey A, editors. *The research process in nursing*, 5th ed. Oxford: Blackwell; 2006, pp. 454–74.
2 Altman DG, Bland JM. Presentation of numerical data. *British Medical Journal* 1996;**312**:572.
3 Tufte ER. *The visual display of quantitative information*. Cheshire, Connecticut: Graphics Press; 1983.
4 Ehrenberg ASC. *A primer in data reduction*. Chichester: John Wiley & Sons; 2000.
5 Morrell CJ, Walters SJ, Dixon S, Collins K, Brereton LML, Peters J, et al. Cost effectiveness of community leg ulcer clinic: randomised controlled trial. *British Medical Journal* 1998;**316**:1487–91.
6 Bigwood S, Spore M. *Presenting numbers, tables and charts*. Oxford: Oxford University Press; 2003.
7 Huff D. *How to lie with statistics*. London: Penguin Books; 1991.