# Figure 3.1. DATA PRESENTATION AND ANALYSIS: Looking at Data

## 1. Introduction

In data-based investigating, the Formulation, Design and Execution stages of the FDEAC cycle are followed by the Analysis stage – examinating, presentating and analysing the data to answer the question(s) of interest. The *first* step in assessing the information content of a data is to *look at the data*; "*look*" is to be taken literally, in the sense of examining pictorial and tabular presentations of the data in ways that are appropriate for the question(s) the data are intended to answer. It may take considerable work, involving a number of attempts, to develop forms of presentation which really *are* appropriate; care, honesty and imagination are essential to this process.

Data presentations have *two* functions:

- they are an essential component of the process whereby the investigator(s) achieve an understanding of the data; such understanding should *always* be a precondition of using the data to answer questions.
- pictorial displays and tables should be the primary means by which it is made clear to *others* that the data *do* yield the stated answers.

## 2. Principles of Graphical Excellence

In the book *The Visual Display of Quantitative Information* (Graphics Press, Cheshire, Connecticut, 1983), Edward R. Tufte gives (p. 51) the following five hallmarks of *graphical excellence*:

- it is the well-designed presentation of interesting information – a matter of *substance*, of *statistics* and of *design*;
- it consists of complex ideas communicated with clarity, precision, and efficiency;
- it is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space;
- it is nearly always multivariate;
- it requires telling the *truth* about the data.

More generally, these statements should be the goals of *all* data presentation and analysis; graphical excellence (or its absence) is the theme of the data displays in this Part 3.

## 3. Tabular Presentation of Data

The principles of graphical excellence listed above in Section 2 are equally applicable to the *tabular* presentation of data. As discussed in Figure 3.11a, constructing proper tables is as demanding a task as producing informative pictorial displays, and the two methods of presentation are complementary. The vertical and horizontal spacings of a table (*e.g.*, the distances between rows and between columns) should be chosen so that, even at a glance, there is appropriate separation between the various table components (*e.g.*, titles and headings, data, explanations and legends) and between data subcategories; this is proper *blocking* of the table contents. In addition, tables should be made as *self-contained* as feasible; for example, it is better to define abbreviations, to indicate special features or limitations of the data, and to describe method(s) of data collection, at least in summary form, in the *table* rather than only in accompanying text.

A characteristic of tables (but less commonly of graphical displays) is that they may become the primary repository of the original data; this can be of vital importance in journal publications of scientific papers, for example. The ability of a table to function as an historical record of data should be a significant consideration when the table is constructed. Tables also often provide a good way of allowing other people to check the *accuracy* and *appropriateness* of data manipulations involved in the answers which are presented.

A set of 64 tables, many of appreciable complexity, is given in the book (John Wiley & Sons, Inc., 1981) edited by J.O. Nriagu: *Cadmium in the Environment. Part 2.* Health Effects, in Chapter 4 entitled: *Distribution of Cadmium in Human Tissues.* The chapter illustrates the use of tabular displays as an integral part of a comprehensive presentation and assessment of a large multi-component data set.

## 4. A Classification of Pictorial Displays

In the book *Statistical mapping and the presentation of statistics* (Edward Arnold, London, Second edition, 1973), G.C. Dickinson (p. 17) gives a 'family tree' of pictorial displays. The primary division is into *statistical diagrams*, where spatial distribution (if any) is *not* an important aspect of the data, and *statistical maps*, where the spatial distribution of the data *is* important. In statistical *diagrams*, there is then a division on the basis of the chief *purpose* of the diagram. To show *relationships between quantities* or *trends*, the types of diagrams which can be used include *line graphs*, *bar graphs*, *circular graphs*, *log graphs* and *scatter diagrams*. Alternatively, to illustrate *division into components*, diagram types include *compound line*, *bar graphs*, *divided circles* ('pie charts') *and rectangles*, and *triangular graphs*.

Statistical *maps* are also divided into two categories: *non-quantitative* and *quantitative*. Non-quantitative maps convey their


(*continued overleaf*)

information using symbols, letters, shading, etc. Quantitative maps are divided into three subcategories on the basis of whether they show quantities distributed at specific *points*, within a given *area*, or along a *line.* The three subcategories then usually employ different methods of conveying information; for instance, proportional shapes (*e.g.*, circles, squares) can be used for *points*, shadings can cover *areas*, and bands of proportional width or graduated size may be useful along *lines.* There is much scope for achieving graphical excellence by the creative application of *combinations* of any of the techniques in this categorization.

*Tabulations of Mortality by County for Ontario Residents During 1964-68*, Faculty of Mathematics, University of Waterloo, April, 1983, illustrates the use of tables, shaded maps and histograms in presenting a large data set. The 14 graded shading patterns of the maps, originally in red, blue and black, were developed to remain visually distinct when reproduced in black and white.

## 5. Frequency Distributions

A useful summary of a data set is its *frequency distribution* – a set of mutually exclusive *intervals* (or *categories*) which span the full range of the observations, together with the *number* of observations that fall in each interval or category. Three data features (often conveying information about population attributes) which are seen from a properly-constructed frequency distribution are the *shape* of the distribution, where it is *located*, and how *spread out* it is. In addition, a frequency distribution should show if the data set contains *outliers* – observations (usually *few* in number) which are *well separated* from the main body of data. A useful (though not essential) first step in constructing a frequency distribution is to *order* the observations in the data set (from smallest to largest, say), although it is good statistical practice to retain a record of the *original* order of the data, because this order may sometimes be a source of useful information. The following Sections 6 and 7 describe two methods of presenting frequency distributions.

## 6. Stemplots (also called Stem and Leaf Plots)

Constructing a *stemplot* or *stem and leaf* plot is a useful first step in the examination of almost *any* data set; such displays can be produced by hand fairly quickly even for data sets containing several hundred observations, although some statisticians tend to confine their use to 'small' data sets. In addition to their ease of construction, stemplots have three other advantages over histograms:

- the actual observations can be recovered from the plot;
- the number of significant figures in the data (*i.e.*, the precision of measuring) is apparent;
- patterns in the observations (*e.g.*, only odd final digits) or in their spacing may become visible.

Stemplots have a number of variations which allow them to be applied to data sets with a wide range of features. Examples of the simplest variety of stemplot (with proper titles, labels, etc.) are given in Figures 3.6.

## 7. Histograms

A *histogram* can be thought of as a stemplot in which (vertical) *bars* are used in place of the (horizontal) rows of digits which form the leaves; however, histograms are more flexible because of the freedom which is available in the choice of the intervals which determine the widths of the bars. Proportions are represented in histograms by bar *area*; the vertical axis of a histogram thus has a *density* scale. Like *all* pictorial displays, histograms should have properly labelled axes and an informative title. In addition to exhibiting important features of a data set, histograms provide a useful visual link between empirical and theoretical probability distributions; *i.e.*, they facilitate *mathematical modelling* of data distributions. The method of constructing a histogram and features to look for in the finished display are described in Figure 3.7, which is followed by Figures 3.8 showing histograms (and bar graphs) from a variety of data sets. We have previously encountered histograms in these Course Materials in Figures 2.5c and 2.7b on pages 2.43 and 2.51.

A *bar graph* (or *bar chart*) is similar in appearance to a histogram, except that gaps are left between the bars to indicate lack of continuity from one category (*e.g.*, of nominal data) to the next and a density scale is seldom involved. Unfortunately, the distinction between the two displays is not always recognized and their names sometimes appear (wrongly) to be used interchangeably.

## 8. Misconceptions

Misconceptions about the processes of using data to answer questions are common; four of the more damaging ones are:

- it is not necessary to *look* at data; rather, it is only necessary to enter them into a computer, preferably one that has a variety of pre-programmed statistical software available, and answers will be forthcoming;
- informative data presentations (tabular *and* pictorial or graphical) are *easy* to construct;
- simple summaries of data are usually *un*informative;
- proper methods of data presentation and analysis are *complicated*; generally, the *more* complicated the method, the 'better' its answers.

These mistaken ideas can have serious consequences – inhibiting proper data presentation and analysis and thus contributing to the abuse of statistical methods and disseminating falsehoods.