A HANDBOOK OF STATISTICAL GRAPHICS USING SAS ODS

GEOFF DER

BRIAN S. EVERITT

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

A HANDBOOK OF STATISTICAL GRAPHICS USING SAS ODS

GEOFF DER
UNIVERSITY OF GLASGOW
UK

BRIAN S. EVERITT
PROFESSOR EMERITUS
KING'S COLLEGE
LONDON, UK

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C014174

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20140715

International Standard Book Number-13: 978-1-4665-9903-1 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

A HANDBOOK OF STATISTICAL GRAPHICS USING SAS ODS

Preface

Graphs, diagrams, plots etc. are essential components of almost all statistical analyses. They are needed in all stages of dealing with data, from an initial assessment of the data to suggesting what statistical models might be appropriate and for diagnosing the chosen models once they have been fitted to the data. And graphical material is often of great help to statisticians when discussing their results with clients such as psychologists, clinicians, psychiatrists and others as an aid to getting over the message the data they have collected has to tell. In this book we cover what might be termed the 'bread-and-butter' graphical methods needed in every statistician's toolkit and how to implement them using SAS Version 9.4.

SAS has two systems for producing graphs: the traditional SAS/GRAPH procedures and the newer ODS graphics. This book uses ODS graphics throughout as we believe this system offers a number of advantages. These include: their ease of use, the high quality of results, the consistency of appearance with tabular output and the convenience of semiautomatic graphs from the statistical procedures. For most users these and the new ODS graphical procedures will be all they need.

The SAS programs and the data used in this book are all available online at http://go.sas.com/hosgus/.

We hope the book will be useful for applied statisticians and others who use SAS in their work.

Contents

Chapter 1

An Introduction to Graphics: Good Graphics, Bad Graphics, Catastrophic Graphics and Statistical Graphics

1.1 The *Challenger* Disaster

January 28, 1986, was an unusually cold morning at the Kennedy Space Center (KSC) in Florida but after several days' delay the Space Shuttle *Challenger* was finally launched at 11:36 EST. There was more than the usual interest in this particular shuttle flight because of the presence on board of Christa McAuliffe, a schoolteacher who had been chosen to fly by the Teacher in Space project. A few miles from the KSC, the large crowd watching the launch that day included McAuliffe's parents and the President of the United States Ronald Reagan. Seventy-three seconds into the flight the *Challenger* broke apart leading to the deaths of its seven crew members.

The cause of the disaster was eventually traced to the O-rings that sealed the joints of the rocket motor. One of the O-ring seals had failed at liftoff allowing pressurized hot gas to escape from within the solid rocket motor leading to the eventual structural failure of the rocket; aerodynamic forces did the rest. The O-ring failure was ascribed to the low temperature at launch.

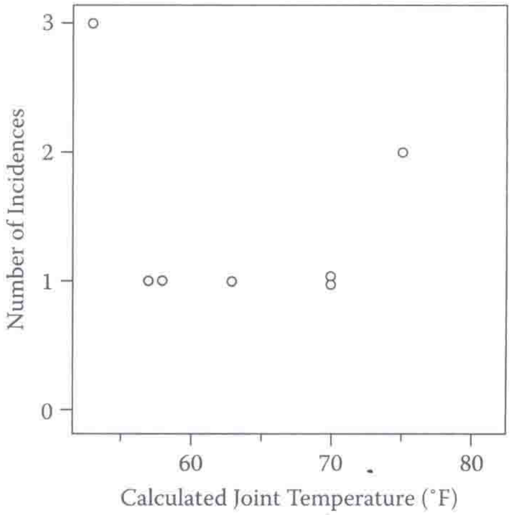

Figure 1.1 Data plotted by space shuttle engineers the evening before the *Challenger* accident to determine the dependence of O-ring failure on temperature.

Engineers had studied the possibility that low temperature might affect the performance of the O-rings the evening before the launch of *Challenger* by plotting data obtained from previous shuttle flights in which the O-rings had experienced thermal distress. The resulting graph (a simple *scatterplot*; see Chapter 6) is shown in Figure 1.1. The horizontal axis shows the O-ring temperature and the vertical axis shows the number of O-rings that had experienced thermal distress. The conclusion drawn by the engineers who examined this graph was that there was no relationship between temperature and incidences of thermal distress and so *Challenger* was allowed to take off when the temperature was 31°F with tragic consequences.

The data for no incidences of thermal distress were not included in the plot used by the shuttle engineers, as those involved believed that these data were irrelevant to the issue of dependence of thermal distress on temperature. They were mistaken, as shown by the plot in Figure 1.2, which includes *all* the data. Here a pattern does emerge and a dependence on temperature is revealed. Here choosing the wrong graph led to a disaster.

1.2 Graphical Displays

Just what is a graphical display? Edward Tufte gives a concise description in his now classic book, *The Visual Display of Quantitative Information*, first published in 1983:

> Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading and color.

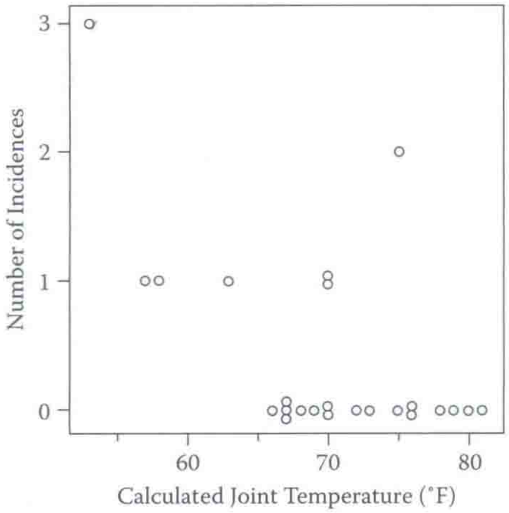

Figure 1.2 A plot of the complete O-ring data.

Graphical displays are very popular; it has been estimated (not sure by whom or how!) that between 900 billion (9 × 10^{11}) and 2 trillion (2 × 10^{12}) images of graphics are printed each year. Perhaps one of the main reasons for such popularity is that graphical presentation of data often provides the vehicle for discovering the unexpected (see Cleveland, 1993, for an example) because the human visual system is very powerful for detecting 'patterns', although it has to be remembered that whilst humans are undoubtedly good at discerning subtle patterns that are really there they are equally expert at imagining patters that are altogether absent.

Some of the advantages of graphical methods have been listed by Schmid (1954):

- In comparison with other types of presentations, well-designed charts are more effective in creating interest and in appealing to the attention of the reader.
- Visual relationships as portrayed by charts and graphs are more easily grasped and more easily remembered.
- The use of charts and graphs saves time, since the essential meaning of large measures of statistical data can be visualized at a glance.
- Charts and graphs provide a comprehensive picture of a problem that makes for a more complete and better balanced understanding than could be derived from tabular or textual forms of presentation.
- Charts and graphs can bring out hidden facts and relationships and can stimulate, as well as aid, analytical thinking and investigation.

Schmid's last point is reiterated by the legendary John Tukey in his observation that 'the greatest value of a picture is when it forces us to notice what we never expected to see'.

The prime objective of a graphical display is to communicate to others and ourselves. Graphic design must do everything it can to help people understand. In some cases a graphic is required to give an overview of the data and perhaps to tell a story about the data. In other cases a researcher may want a graphical display to suggest possible hypotheses for testing on new data and after some model has been fitted to the data a graphic that criticizes the model may be what is needed. Examples of using graphic displays in all these situations will be found throughout this text. But for now we will consider a little history and some early examples of graphical displays.

1.3 A Little History and Some Early Graphical Displays

Conveying complex information in words has always been considered difficult, and alternative methods of communication have been sought. Wainer (1997) points out that as far back as preclassical antiquity, paleolithic art provided an early and very striking example of graphic display with carvings of animals being intermixed with patterns of dots and strokes that archeologists have interpreted as a lunar notation system related to the animal's seasonal appearance. And Egyptian geographers turned spatial information into spatial diagrams and maps to keep track of land shifted by river floods.

But we have to make a leap into the 17th and 18th centuries to meet some early graphics that are vaguely familiar to us. William Playfair, for example, is often credited with inventing the *bar chart* (see Chapter 3) in the last part of the 18th century, although a Frenchman, Nicole Oresme, used a bar chart in a 14th century publication, *The Latitude of Forms*, to plot velocity of a constantly accelerating object against time. But it was Playfair who popularized the idea of graphic depiction of quantitative information. Figure 1.3 shows one of Playfair's earliest bar charts used to show imports and exports of Scotland.

Playfair's graphs are essentially one-dimensional and the next major graphical invention and the first truly two-dimensional graph is the *scatterplot* (or *scattergram*) an example of which introduced this chapter. According to Friendly and Denis (2006) the humble scatterplot may be considered the most versatile, polymorphic and generally useful invention in the entire history of statistical graphics. As we have already seen in Section 1.1, a scatterplot is quintessentially a plot of two variables x and y measured independently to produce bivariate pairs (x_i, y_i) and displayed as individual points of a coordinate grid typically defined by horizontal and vertical axes, where there is no necessary functional relation between x and y. It is worthwhile here quoting Tufte's comment about the scatterplot (1983):

> …the scatterplot and its variants is the greatest of all graphical designs. It links at least two variables encouraging and even imploring the viewer to assess the possible causal relationship between the plotted variables. It confronts causal theories that x causes y with empirical evidence as to the actual relationship between x and y.

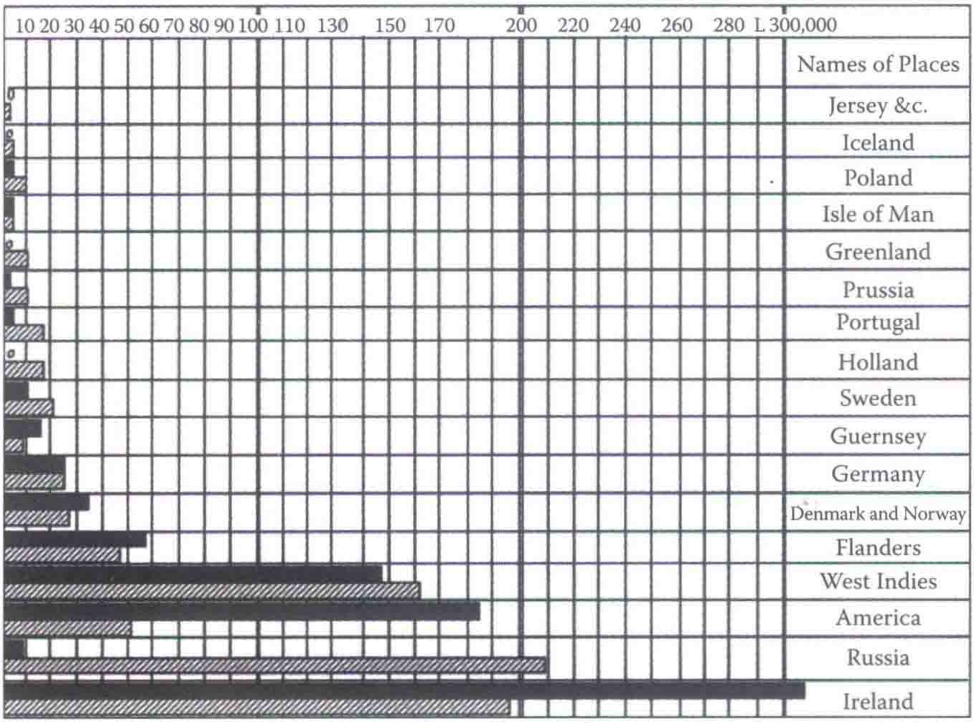

Figure 1.3 Playfair's bar chart for imports and exports of Scotland: imports are represented by crosshatch bars and exports by solid bars.

There are various claimants for the first genuine scatterplot, but Friendly and Denis (2006) come down on the side of Francis Galton's graphical displays constructed in his work on correlation, regression and heritability, although these are somewhat less than true scatterplots of data as used today being essentially bivariate frequency tables. The earliest known example of one of these charts is shown in Figure 1.4.

Friendly and Denis (2006) call diagrams such as Figure 1.4 'a poor-man's scatterplot' but note that for Galton such diagrams allowed him to 'smooth' the numbers by averaging the four adjacent cells to produce a simple *bivariate density estimator* (see Chapter 6), a procedure which eventually led to diagrams such as Figure 1.5, a graphic that led on to a host of important developments in statistics, for example, correlation, regression and partial correlation.

Another famous scatterplot which led to insights into how stars could be classified is the *Hertzsprung–Russell diagram* (H–R diagram), pioneered independently by Elnar Hertzsprung and Henry Norris Russell in the early 1900s. The H–R diagram is a plot of the luminosity of a star against its temperature; an example is given in Figure 1.6 and shows that stars preferentially fall into certain regions of the diagram with the majority falling along a curving diagonal line called the *main sequence*. The significance of the H-R diagram when

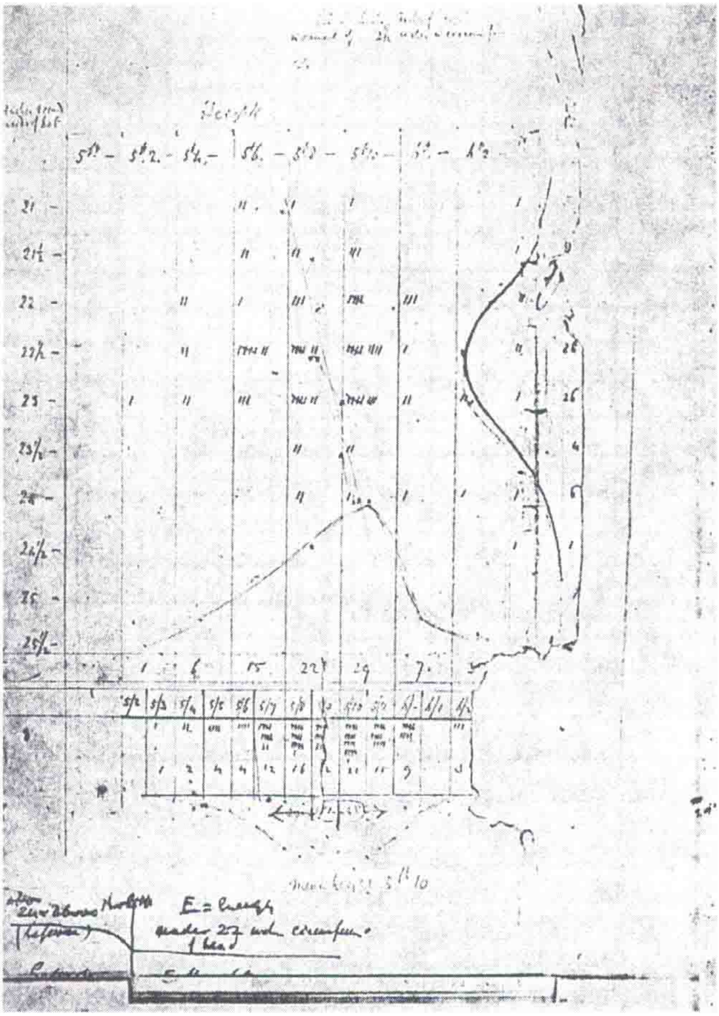

Figure 1.4 Galton's first correlation diagram, showing the relation between head circumference and height, from his undated notebook 'Special Peculiarities'. (From Hilts, V. L., 1975, *A Guide to Francis Galton's English Men of Science*, Philadelphia, American Philosophical Society, Figure 5, p. 26. With permission.)

it was first plotted is that stars were seen as concentrated in distinct regions rather than being distributed at random.

Lastly in this section we will consider a plot made in the 19th century by an early epidemiologist, John Snow (1813–1858), which was probably responsible for saving many lives. After an outbreak of cholera in central London in September 1854, Snow used data collected by the General Register Office and plotted the location of deaths on a map of the area and also showed the location of the area's eleven water pumps. The resulting map is shown in Figure 1.7 (deaths are marked

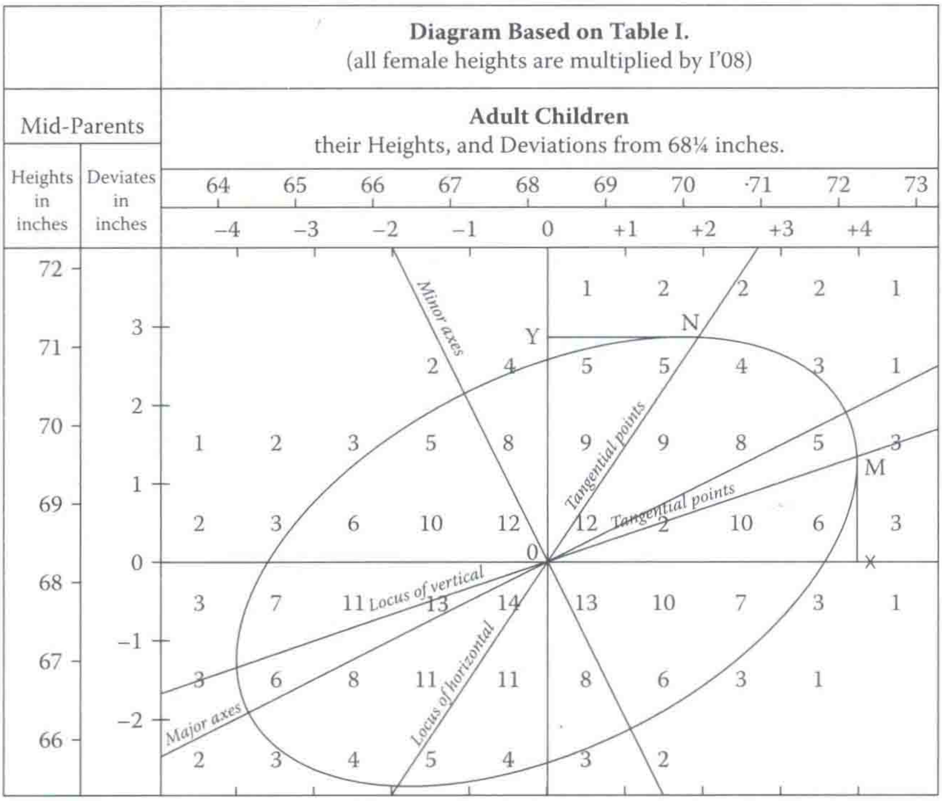

Figure 1.5 Galton's smoothed correlation diagram for the data on heights of parents and children, showing one ellipse of equal frequency. (From Galton, F., 1886, Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263, Plate X. With permission.)

by dots and water pumps by crosses). Examining the scatter over the surface of the map, Snow observed that nearly all the cholera deaths were among those who lived near the Broad Street pump. But before claiming that he had discovered a possible causal connection, Snow made a more detailed investigation of the deaths that had occurred near some other pump. He visited the families of ten of the deceased and found that five of these, because they preferred its taste, regularly sent for water from the Broad Street pump. Three others were children who attended a school near the Broad Street pump. One other finding that initially confused Snow was that there were no deaths amongst workers in a brewery close to the Broad Street pump, a confusion that was quickly resolved when it became apparent that the workers drank only beer, never water! Snow's findings were sufficiently compelling to persuade the authorities to remove the handle of the Broad Street pump and in days the neighbourhood epidemic that had taken more than 500 lives had ended.

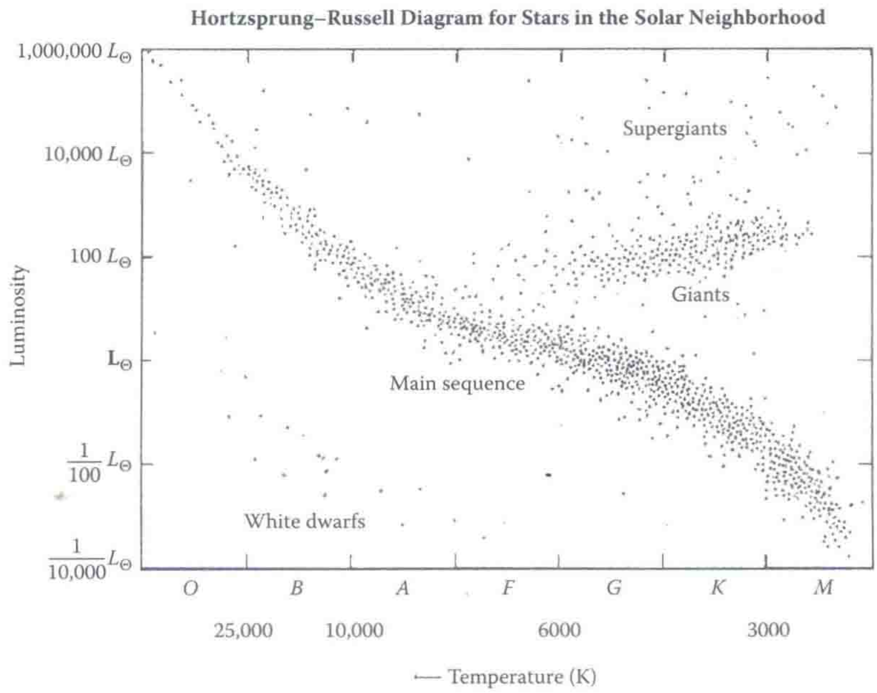

Figure 1.6 Herztsprung–Russell diagram.

1.4 Graphical Deception

Not all graphical displays are as honest as they should be and it is relatively easy to mislead the unwary with graphical material. For example, consider the plot of the death rate per million from cancer of the breast for several periods over the last three decades, shown in Figure 1.8. The rate appears to show a rather alarming increase. However, when the data are replotted with the vertical scale beginning at zero, as shown in Figure 1.9, the increase in the breast cancer death rate is altogether less startling. This example illustrates that undue exaggeration or compression of the scales is best avoided when drawing graphs (unless, of course, you are actually in the business of deceiving your audience).

A very common distortion introduced into the graphics most popular with newspapers, television and the media in general is when *both* dimensions of a *two-dimensional figure* or *icon* are varied simultaneously in response to changes in a single variable. The examples shown in Figure 1.10, both taken from Tufte (1983), illustrate this point. Tufte quantifies the distortion with what he calls the *lie factor* of a graphical display, which is defined as the size of the effect shown in the graph divided by the size of the effect in the data. Lie factor values close to unity show that the graphic is probably representing the underlying numbers reasonably accurately. The lie factor for the oil barrels is 9.4 since a 454% increase is depicted as 4280%. The lie factor for the shrinking doctors is 2.8.