William S. Cleveland

Visualizing Data

AT&T Bell Laboratories, Murray Hill, New Jersey

Published by Hobart Press, Summit, New Jersey

Printed in the United States of America

ISBN 0-9634884-0-6 CLOTH

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 92-075077

PUBLISHER'S CATALOGING IN PUBLICATION

Cleveland, William S., 1943–
Visualizing data / by William S. Cleveland.
p. cm.
Includes bibliographical references and index.

1. Graphic methods. 2. Mathematical statistics–Graphic methods. I. Title.

QA90.C549 1993 511'.5
QBI93-693

Contents

Preface

Visualization is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones.

Tools

Tools matter. There are exceptionally powerful visualization tools, and there are others, some well known, that rarely outperform the best ones. The data analyst needs to be hard-boiled in evaluating the efficacy of a visualization tool. It is easy to be dazzled by a display of data, especially if it is rendered with color or depth. Our tendency is to be misled into thinking we are absorbing relevant information when we see a lot. But the success of a visualization tool should be based solely on the amount we learn about the phenomenon under study. Some tools in the book are new and some are old, but all have a proven record of success in the analysis of common types of statistical data that arise in science and technology.

Graphing and Fitting

There are two components to visualizing the structure of statistical data — graphing and fitting. Graphs are needed, of course, because visualization implies a process in which information is encoded on visual displays. Fitting mathematical functions to data is needed too. Just graphing raw data, without fitting them and without graphing the fits and residuals, often leaves important aspects of data undiscovered. The visualization tools in this book consist of methods for graphing and methods for fitting.

Applications

The book is organized around applications of the visualization tools to data sets from scientific studies. This shows the role each tool plays in data analysis, and the class of problems it solves. It also demonstrates the power of visualization; for many of the data sets, the tools reveal that effects were missed in the original analyses or incorrect assumptions were made about the behavior of the data. And the applications convey the excitement of discovery that visualization brings to data analysis.

The Legacy of the Past

The visualization of statistical data has always existed in one form or another in science and technology. For example, diagrams are the first methods presented in R. A. Fisher's *Statistical Methods for Research Workers*, the 1925 book that brought statistics to many in the scientific and technical community [38]. But with the appearance of John Tukey's pioneering 1977 book, *Exploratory Data Analysis*, visualization became far more concrete and effective [76]. Since 1977, changes in computer systems have changed how we carry out visualization, but not its goals.

Display Methods

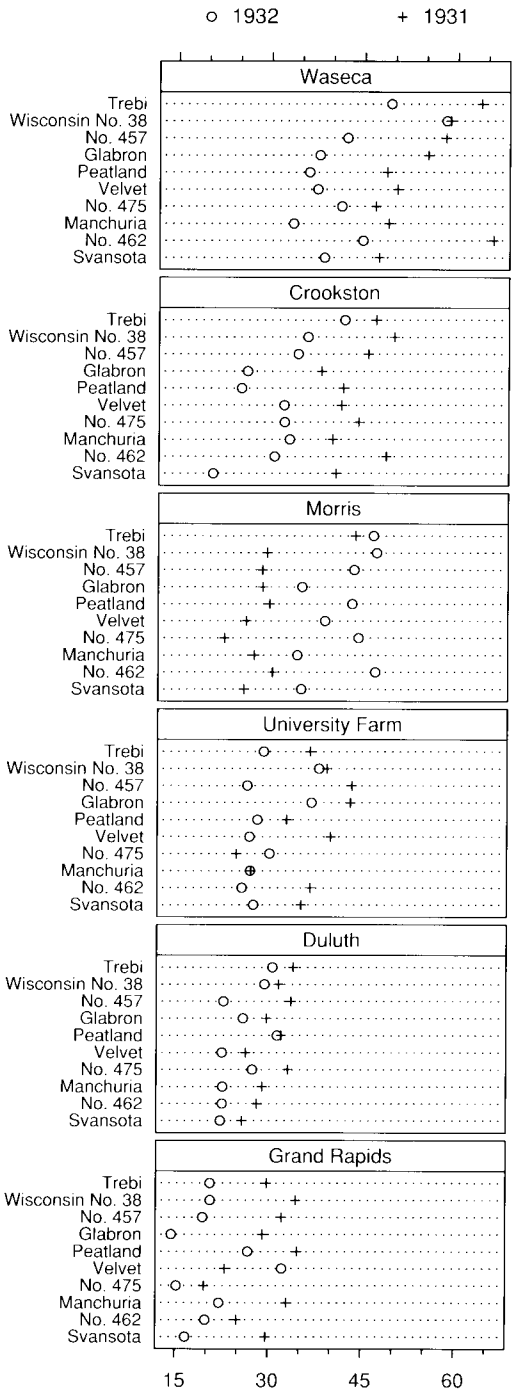
When a graph is made, quantitative and categorical information is encoded by a display method. Then the information is visually decoded. This visual perception is a vital link. No matter how clever the choice of the information, and no matter how technologically impressive the encoding, a visualization fails if the decoding fails. Some display methods lead to efficient, accurate decoding, and others lead to inefficient, inaccurate decoding. It is only through scientific study of visual perception that informed judgments can be made about display methods. Display methods are the main topic of *The Elements of Graphing Data* [20]. The visualization methods described here make heavy use of the results of *Elements* and other work in graphical perception.

Prerequisites

The reader should be familiar with basic statistics and the least-squares method of fitting equations to data. For example, an introductory course in statistics that included the fundamentals of regression analysis would be sufficient.

How to Read the Book

For most purposes, the chapters need to be read in order. Material in later chapters uses tools and ideas introduced in earlier chapters. There are two exceptions to this general rule. Chapter 6, which is about multiway data, does not use material beyond Section 4.6 in Chapter 4. Also, sections of the book labeled "For the Record" contain details that are not necessary for understanding and using the visualization tools. The details are meant for those who want to experiment with alterations of the methods, or want to implement the methods, or simply like to take in all of the detail.



1.1 A multiway dot plot graphs data from a barley experiment run in the 1930s. This visualization reveals an anomaly in the data that was missed by the experimenters and by others who subsequently analyzed the data.

1 Introduction

Visualization is an approach to data analysis that stresses a penetrating look at the structure of data. No other approach conveys as much information. As W. Edwards Deming puts it, visualization "retains the information in the data" [33]. Conclusions spring from data when this information is combined with the prior knowledge of the subject under investigation. An important discovery springs from Figure 1.1. It attests to the power of visualization.

In the early 1930s, agronomists in Minnesota ran a field trial to study the crop barley [55]. At six sites in Minnesota, ten varieties of barley were grown in each of two years. The data are the yields for all combinations of site, variety, and year, so there are $6 \times 10 \times 2 = 120$ observations. In Figure 1.1, each panel displays the 20 yields at a single site.

The barley data have been analyzed and re-analyzed for decades. Their first analysis appeared in a 1934 report published by the experimenters. The statistician and geneticist R. A. Fisher, who established the modern foundations of statistics, presented the data for five of the sites in his book, *The Design of Experiments* [39]. Francis J. Anscombe [3, 4] and Cuthbert Daniel [29], pioneers of diagnostic methods for determining when statistical models fit data, also analyzed them.

Now, the visualization of Figure 1.1 reveals an anomaly that was missed in these previous analyses. It occurs at Morris. For all other sites, 1931 produced a significantly higher overall yield than 1932. The reverse is true at Morris. But most importantly, the amount by which 1932 exceeds 1931 at Morris is similar to the amounts by which 1931 exceeds 1932 at the other sites. Thus we have a mystery. Either an extraordinary natural event, such as disease or a local weather anomaly, produced a strange coincidence, or the years for Morris were inadvertently reversed. The mystery is investigated at the end of the book; the conclusion will be revealed there, not here, to retain the suspense of the full story, which is complicated.

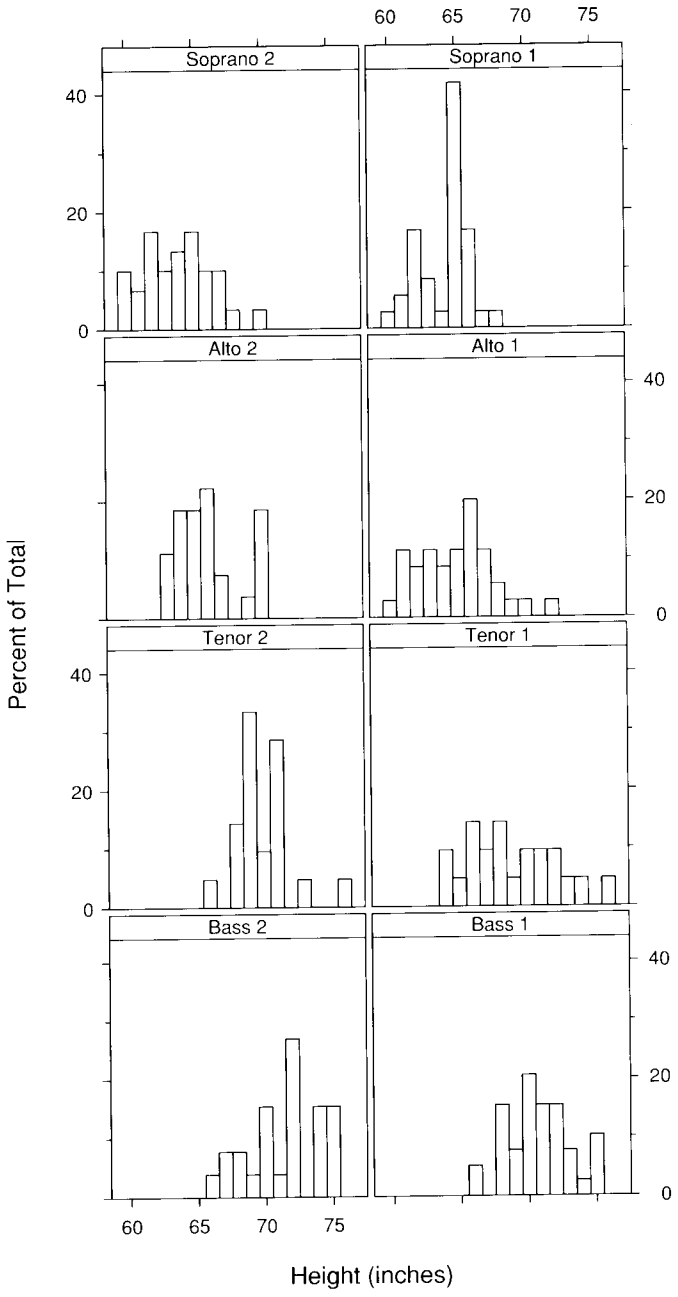
1.1 Tools and Data Types

Tools matter. The Morris anomaly is revealed in Figure 1.1 because the tool used to display the data, a *multiway dot plot*, is an effective one. In the analyses of the past, the methods were insufficient. Even the most adroit data analyst cannot divine effects in data. The critical revelation in Figure 1.1 is not simply that the year effects at Morris are reversed — that was noted in the past. The revelation is that at Morris, the 1932 yields minus the 1931 yields have about the same overall level as the 1931 yields minus the 1932 yields at the other sites. This observation triggers the thought that the years might have been reversed. As Chapter 6 will show, important aspects of the display method of the multiway dot plot contribute to this revelation and other revelations of importance in solving the mystery. Tools matter indeed.

The tools of this book are organized by type of data. Each chapter treats a different data type: univariate, bivariate, trivariate, hypervariate, and multiway.

Univariate Data

Figure 1.2 uses histograms to graph heights of singers in the New York Choral Society [16]. The singers are divided into eight voice parts, shown by the panel labels in the figure. Starting from the lower left panel of the display, and then going from left to right and from bottom to top, the pitch intervals of the voice parts increase. For example, the second basses sing lower pitches than the first basses, who in turn sing lower pitches than the second tenors. The goal of the analysis of the singer data is to determine if the heights tend to decrease as the pitch interval increases. The heights are univariate data: measurements of a single quantitative variable. In this case the variable is broken up into groups by the categorical variable, voice part. The visualization of univariate data is treated in Chapter 2.



1.2 Histograms graph the singer heights by voice part. The interval width is one inch.

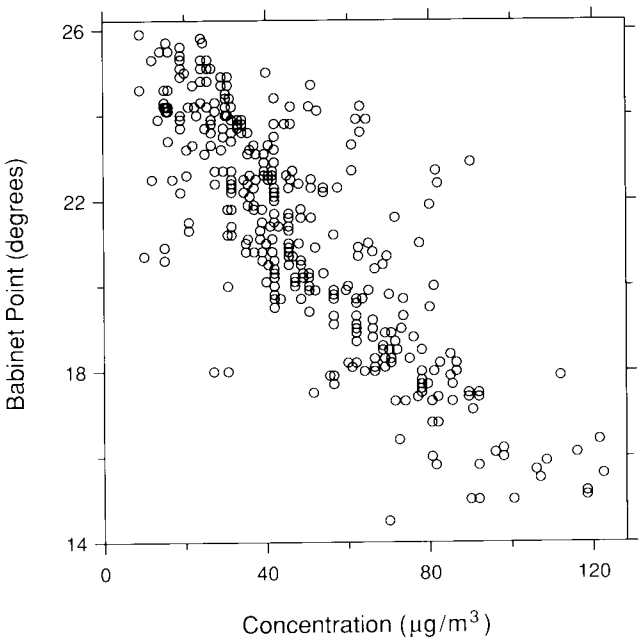
The histogram is a widely used graphical method that is at least a century old. But maturity and ubiquity do not guarantee the efficacy of a tool. The histogram is a poor method for comparing groups of univariate measurements. In Figure 1.2, it does not clearly reveal the relationship of height and voice range. True, we see that the heights in the bottom four panels tend to be greater than the heights in the top four panels. But the bottom heights are all men and the top are all women, so we have merely seen the obvious. The visualization tools in Chapter 2 show far more of the structure of the singer data. They include quantile plots, q-q plots, normal q-q plots, box plots, and fitting methods. The venerable histogram, an old favorite, but a weak competitor, will not be encountered again.

Bivariate Data

Figure 1.3 is a scatterplot of data from an experiment on the scattering of sunlight in the atmosphere [7]. The vertical scale is the Babinet point, the scattering angle at which the polarization of sunlight vanishes. The horizontal scale is the atmospheric concentration of solid particles in the air. The goal is to determine the dependence of the Babinet point on concentration, so the Babinet point is a response and the concentration is a factor. The polarization data are bivariate data: paired measurements of two quantitative variables. The visualization of bivariate data is treated in Chapter 3.

The polarization data have two components of variation. One component is a smooth underlying pattern — a decrease in the overall level of the Babinet point as concentration increases. Fitting such bivariate data means determining a smooth curve that describes the underlying pattern. The second component is residual variation about this underlying pattern — the vertical deviations of the points from the smooth curve.

The scatterplot is a useful exploratory method for providing a first look at bivariate data to see how they are distributed throughout the plane, for example, to see clusters of points, outliers, and so forth.



1.3 An exploratory scatterplot graphs the Babinet point against particulate concentration.

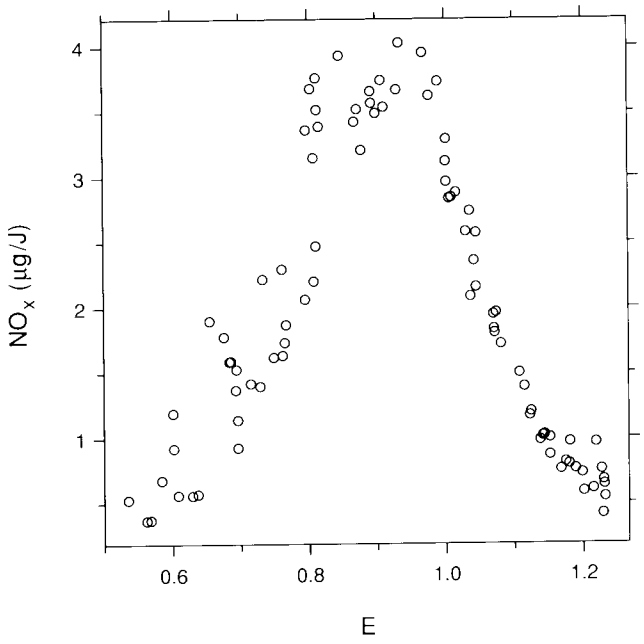
But for factor-response data such as the polarization data in Figure 1.3, we should be prepared to move almost immediately to fitting the data and visualizing the fit and residuals. It is part of the folklore of data display that a good method for putting a smooth curve through bivariate data is to stare at an unfitted scatterplot and fair a smooth curve through the data by the mind's eye. In fact, the residual variation often interferes with the visual smoothing. For example, Figure 1.3 suggests that the underlying pattern is linear in the middle with a hint of curvature at the ends, but it is not possible to assess this nonlinearity with precision, or to even determine if it exists. In Chapter 3, a curve is fitted to the polarization data using the fitting method *loess*, and a pattern not readily apparent from the scatterplot emerges. Conversely, when the underlying smooth pattern is a major component of the data, with steep slopes, the pattern interferes with our assessment of the residual variation. In Chapter 3, residual variation is visualized by many methods; for the polarization data, interesting patterns emerge.

Trivariate and Hypervariate Data

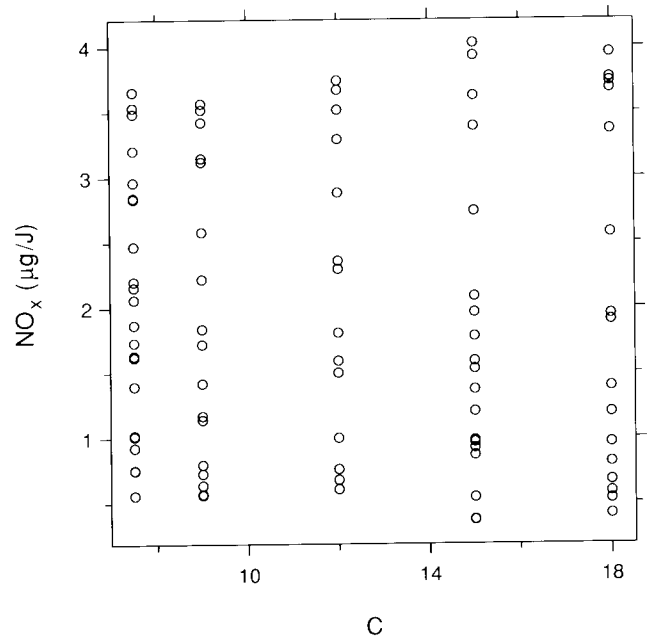
Oxides of nitrogen, NO_x, are one of the major pollutants in automobile exhaust. An experiment was run to study how the concentration of NO_x depends on two engine factors: E, the equivalence ratio, a measure of the richness of the air and fuel mixture, and C, the compression ratio of the engine [12]. The observations, which consist of 88 measurements of the three quantitative variables, are trivariate data. Measurements of four or more quantitative variables are hypervariate data. The visualization of trivariate data is discussed in Chapter 4, and the visualization of hypervariate data is discussed in Chapter 5.

Figures 1.4 and 1.5 graph NO_x against the factors. The scatterplot of NO_x against E reveals a strong nonlinear pattern. The scatterplot of NO_x against C shows little apparent relationship between the two variables. Should we conclude that concentration does not depend on C? There is a precedent for doing this [17]. Still, we will withhold judgment. The data live in three dimensions, but each scatterplot is a projection onto only two dimensions. It is possible for 2-D projections not to reveal 3-D structure. As we go from one point to the next on the scatterplot of NO_x against C, the value of E changes, so the graph is not providing a proper view of how NO_x depends on C for E held fixed. It would be imprudent to conclude at this point that NO_x does not depend on C. For example, a strong dependence of concentration on E could mask a subtler dependence on C.

We need a way of seeing the dependence of NO_X on C without the interference from E. Visualization tools discussed in Chapters 4 will do this for us. For example, the coplot is a particularly incisive method for studying conditional dependence. And, of course, we will fit the data.



1.4 Oxides of nitrogen are graphed against equivalence ratio.



1.5 Oxides of nitrogen are graphed against compression ratio.

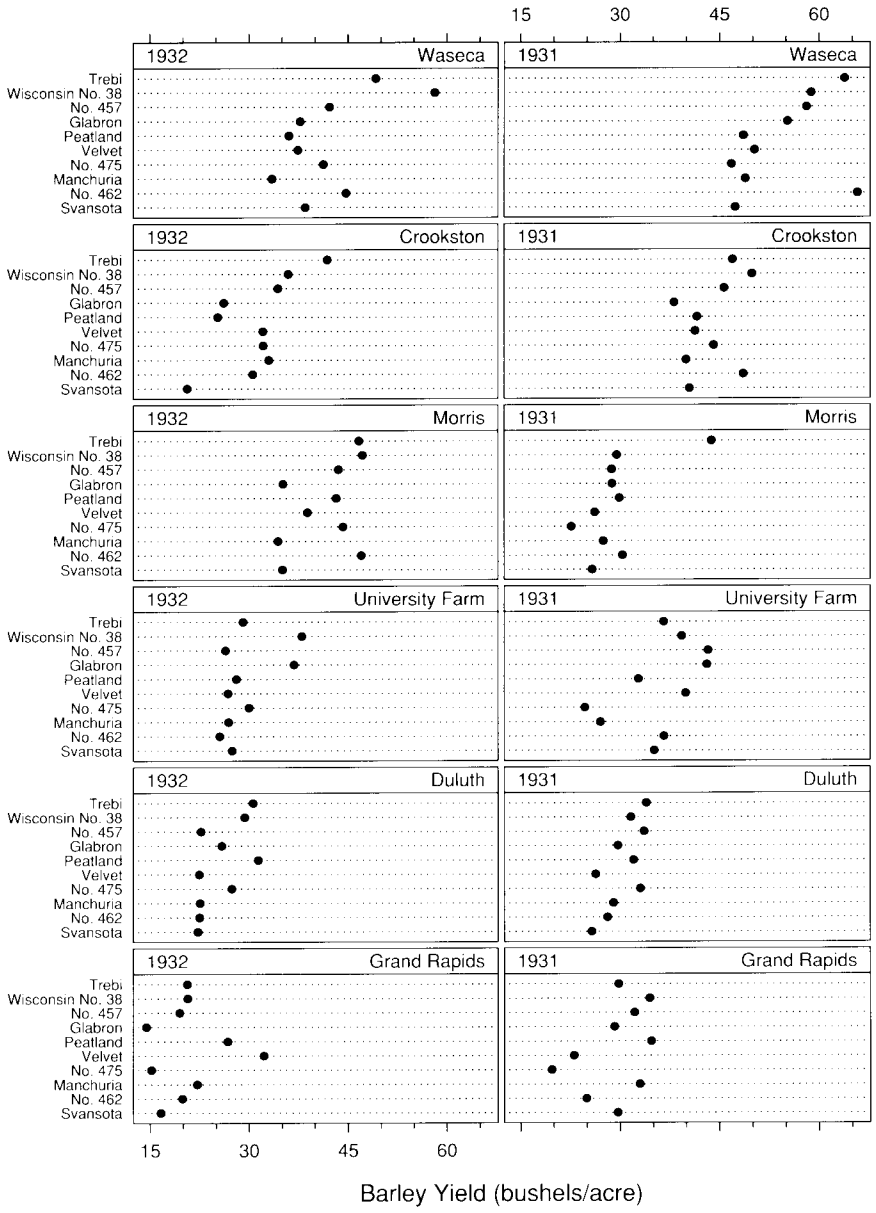
Multiway Data

The barley data discussed at the beginning of the chapter are graphed in Figure 1.6 by a multiway dot plot with a format different from that of the display of the data in Figure 1.1. The barley measurements are multiway data: a quantitative variable is measured for each combination of the levels of two or more categorical variables. The quantitative variable in this case is yield and the categorical variables are variety, site, and year. The visualization of multiway data is discussed in Chapter 6. The chief visualization tool is the multiway dot plot.

1.2 Visualization and Probabilistic Inference

Probabilistic inference is the classical paradigm for data analysis in science and technology. It rests on a foundation of randomness; variation in data is ascribed to a random process in which nature generates data according to a probability distribution. This leads to a codification of uncertainly by confidence intervals and hypothesis tests. Pascal, Fermat, and Huygens laid the foundations of probability theory in the second half of the 17th century, and by the beginning of the 18th century, the variation in scientific data was being described by probability models [47]. But the modern foundations of probabilistic inference as we practice it today were laid in the early part of the 20th century by R. A. Fisher [38, 39, 40].

Visualization — with its two components, graphing and fitting — is a different paradigm for learning from data. It stresses a penetrating look at the structure of data. What is learned from the look is guided by knowledge of the subject under study. Sometimes visualization can fully replace the need for probabilistic inference. We visualize the data effectively and suddenly, there is what Joseph Berkson called *interocular traumatic impact*: a conclusion that hits us between the eyes. In other cases, visualization is not enough and probabilistic inference is needed to help calibrate the uncertainty of a less certain issue. When this is so, visualization has yet another role to play — checking assumptions. The validity of methods of probabilistic inference rest on assumptions about the structure of the population from which the data came. But if assumptions are false, results are not valid. Despite its flippancy, the aphorism, "garbage in, garbage out", is an excellent characterization.



1.6 The barley data are graphed by a multiway dot plot with the data for each site on two panels, one for 1931 and one for 1932.

Without a careful checking of assumptions, validity is replaced by large leaps of faith; one can only hope that the probabilistic assertions of confidence intervals and hypothesis tests are valid. Visualization reduces leaps of faith by providing a framework for studying assumptions.

Carrying out probabilistic inference without checking assumptions deserves an unflattering name. *Rote data analysis* achieves this and describes its character well. Ample examples in the book attest to its dangers.

1.3 Direct Manipulation

Imagine us at the University of Leiden in 1637, trying to understand new data: measurements of the heights and weights of 25 adult males in Leiden. A colleague comes into the office and says, in Dutch, of course:

> I just met a Frenchman named René Descartes and he has an interesting idea for representing mathematical functions of a single variable. Associated with each point in the plane are two numbers: the distances of the point from two perpendicular lines. If the function is f, then the function value $f(x)$ is represented geometrically by showing the point associated with x and $f(x)$. This can be done for many values of x, and the result is a geometric representation of the function that gives you much insight into its behavior. I wonder if we could use this idea to study our height and weight data.

The colleague then reveals Descartes' *La Géométrie* [34].

Cartesian coordinates provide a visual medium within which data can be visually displayed. Most graphical methods use this medium. In 1637 in Leiden, had we been sufficiently creative, we might have exploited Cartesian coordinates and graphed the weights and heights by a scatterplot. This would have made us way ahead of our time. The scientific community only slowly exploited the medium, first on a limited basis in the 1600s and early to middle 1700s, and then with much more energy toward the end of the 1700s [28, 45].

In the 1960s, over three centuries after Descartes' *La Géométrie*, computer scientists created a new visual medium that would be as revolutionary for data display as Cartesian coordinates. *Direct manipulation graphics* appeared on computer screens. The user visually addresses a display using an input device such as a mouse, and causes the display to change in real time. Direct manipulation would become a standard medium not only for data display but also for user interfaces, affecting the basic way that people interface with computer software. But unlike three centuries earlier, scientists were quick off the mark to exploit this new medium for data display. For example, Edward Fowlkes, a statistician, saw the possibilities for graphing data and quickly invented several new methods [2, 43]. As the medium became widely available, the invention of direct manipulation graphical methods grew and intensified [6, 24, 41, 62, 73].

This book presents several direct manipulation methods. This is no small challenge to do on the static pages of a book. But the ideas manage to get through, if not the excitement. The excitement must await the video version of the book.