CRC REVIVALS

Graphical Methods for Data Analysis

John M. Chambers, William S. Cleveland, Beat Kleiner, Paul A. Tukey

CRC Press
Taylor & Francis Group

GRAPHICAL METHODS FOR DATA ANALYSIS

GRAPHICAL METHODS FOR DATA ANALYSIS

John M. Chambers
William S. Cleveland
Beat Kleiner
Paul A. Tukey

Bell Laboratories

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

First published 1983 by CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Reissued 2018 by CRC Press

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.

Library of Congress Cataloging-in-Publication Data
Main entry under title:

Graphical methods for data analysis.

Bibliography: p.
Includes index.
ISBN 0-412-05271-7
1. Statistics—Graphic methods—Congresses.
2. Computer graphics—Congresses. I. Chambers, John M.
II. Series
QA276.3.G73 1983
001.4’22 83-3660

Publisher’s Note
The publisher has gone to great lengths to ensure the quality of this reprint but points out that some imperfections in the original copies may be apparent.

Disclaimer
The publisher has made every effort to trace copyright holders and welcomes correspondence from those they have been unable to contact.

ISBN 13: 978-1-315-89320-4 (hbk)
ISBN 13: 978-1-351-07230-4 (ebk)

Visit the Taylor & Francis Web site at http://www.taylorandfrancis.com and the
CRC Press Web site at http://www.crcpress.com

To our parents

Preface

WHAT IS IN THE BOOK?

This book presents graphical methods for analyzing data. Some methods are new and some are old, some methods require a computer and others only paper and pencil; but they are all powerful data analysis tools. In many situations a set of data — even a large set — can be adequately analyzed through graphical methods alone. In most other situations, a few well-chosen graphical displays can significantly enhance numerical statistical analyses.

There are several possible objectives for a graphical display. The purpose may be to record and store data compactly, it may be to communicate information to other people, or it may be to analyze a set of data to learn more about its structure. The methodology in this book is oriented toward the last of these objectives. Thus there is little discussion of communication graphics, such as pie charts and pictograms, which are seen frequently in the mass media, government publications, and business reports. However, it is often true that a graph designed for the analysis of data will also be useful to communicate the results of the analysis, at least to a technical audience.

The viewpoints in the book have been shaped by our own experiences in data analysis, and we have chosen methods that have proven useful in our work. These methods have been arranged according to data analysis tasks into six groups, and are presented in Chapters 2 to 7. More detail about the six groups is given in Chapter 1 which is an introduction. Chapter 8, the final one, discusses general principles and techniques that apply to all of the six groups. To see if the book is for you, finish reading the preface, table of contents, and Chapter 1, and glance at some of the plots in the rest of the book.

FOR WHOM IS THIS BOOK WRITTEN?

This book is written for anyone who either analyzes data or expects to do so in the future, including students, statisticians, scientists, engineers, managers, doctors, and teachers. We have attempted not to slant the techniques, writing, and examples to any one subject matter area. Thus the material is relevant for applications in physics, chemistry, business, economics, psychology, sociology, medicine, biology, quality control, engineering, education, or virtually any field where there are data to be analyzed. As with most of statistics, the methods have wide applicability largely because certain basic forms of data turn up in many different fields.

The book will accommodate the person who wants to study seriously the field of graphical data analysis and is willing to read from beginning to end; the book is wide in scope and will provide a good introduction to the field. It also can be used by the person who wants to learn about graphical methods for some specific task such as regression or comparing the distributions of two sets of data. Except for Chapters 2 and 3, which are closely related, and Chapter 8, which has many references to earlier material, the chapters can be read fairly independently of each other.

The book can be used in the classroom either as a supplement to a course in applied statistics, or as the text for a course devoted solely to graphical data analysis. Exercises are provided for classroom use. An elementary course can omit Chapters 7 and 8, starred sections in other chapters, and starred exercises; a more advanced course can include all of the material. Starred sections contain material that is either more difficult or more specialized than other sections, and starred exercises tend to be more difficult than others.

WHAT IS THE PREREQUISITE KNOWLEDGE NEEDED TO UNDERSTAND THE MATERIAL IN THIS BOOK?

Chapters 1 to 5, except for some of the exercises, assume a knowledge of elementary statistics, although no probability is needed. The material can be understood by almost anyone who wants to learn it and who has some experience with quantitative thinking. Chapter 6 is about probability plots (or quantile-quantile plots) and requires some knowledge of probability distributions; an elementary course in statistics should suffice. Chapter 7 requires more statistical background. It deals with graphical methods for regression and assumes that the reader is already familiar with the basics of regression methodology. Chapter 8 requires an understanding of some or most of the previous chapters.

ACKNOWLEDGMENTS

Our colleagues at Bell Labs contributed greatly to the book, both directly through patient reading and helpful comments, and indirectly through their contributions to many of the methods discussed here. In particular, we are grateful to those who encouraged us in early stages and who read all or major portions of draft versions. We also benefited from the supportive and challenging environment at Bell Labs during all phases of writing the book and during the research that underlies it. Special thanks go to Ram Gnanadesikan for his advice, encouragement and appropriate mixture of patience and impatience, throughout the planning and execution of the project.

Many thanks go to the automated text processing staff at Bell Labs — especially to Liz Quinzel — for accepting revision after revision without complaint and meeting all specifications, demands and deadlines, however outrageous, patiently learning along with us how to produce the book.

Marylyn McGill's contributions in the final stage of the project by way of organizing, preparing figures and text, compiling data sets, acquiring permissions, proofreading, verifying references, planning page lay-outs, and coordinating production activities at Bell Labs and at Wadsworth/Duxbury Press made it possible to bring all the pieces together and get the book out. The patience and cooperation of the staff at Wadsworth/Duxbury Press are also gratefully acknowledged.

Thanks to our families and friends for putting up with our periodic, seemingly antisocial behavior at critical points when we had to dig in to get things done.

A preliminary version of material in the book was presented at Stanford University. We benefited from interactions with students and faculty there.

Without the influence of John Tukey on statistics, this book would probably never have been written. His many contributions to graphical methods, his insights into the role good plots can play in statistics and

his general philosophy of data analysis have shaped much of the approach presented here. Directly and indirectly, he is responsible for much of the richness of graphical methods available today.

John M. Chambers

William S. Cleveland

Beat Kleiner

Paul A. Tukey

Contents

1

Introduction

1.1 WHY GRAPHICS?

There is no single statistical tool that is as powerful as a well-chosen graph. Our eye-brain system is the most sophisticated information processor ever developed, and through graphical displays we can put this system to good use to obtain deep insight into the structure of data. An enormous amount of quantitative information can be conveyed by graphs; our eye-brain system can summarize vast information quickly and extract salient features, but it is also capable of focusing on detail. Even for small sets of data, there are many patterns and relationships that are considerably easier to discern in graphical displays than by any other data analytic method. For example, the curvature in the pattern formed by the set of points in Figure 1.1 is readily appreciated in the plot, as are the two unusual points, but it is not nearly as easy to make such a judgment from an equivalent table of the data. (This figure is more fully discussed in Chapter 5.)

The graphical methods in this book enable the data analyst to explore data thoroughly, to look for patterns and relationships, to confirm or disprove the expected, and to discover new phenomena. The methods also can be used to enhance classical numerical statistical analyses. Most classical procedures are based, either implicitly or explicitly, on assumptions about the data, and the validity of the analyses depends upon the validity of the assumptions. Graphical methods provide powerful diagnostic tools for confirming assumptions, or, when the assumptions are not met, for suggesting corrective actions.

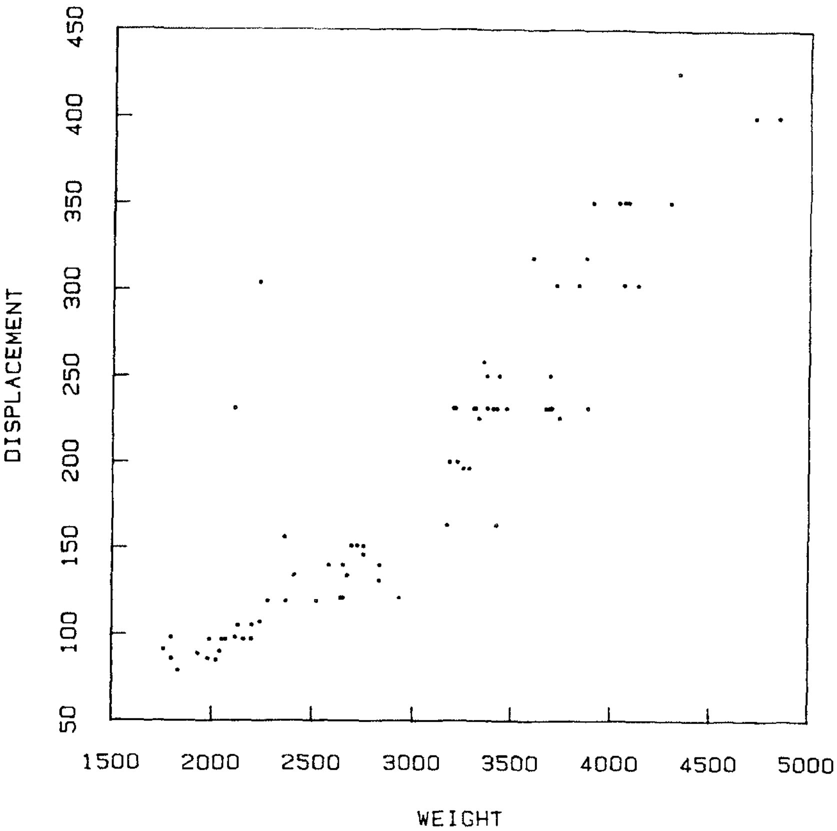

Figure 1.1 Scatter plot of displacement (in cubic inches) versus weight (in pounds) of 74 automobile models.

Without such tools, confirmation of assumptions can be replaced only by hope.

Until the mid-1970's, routine large-scale use of plots in data analysis was not feasible, since the hardware and software for computer graphics were not readily available to many people and making large numbers of plots by hand took too much time. We no longer have such an excuse. The field of computer graphics has matured. The recent rapid proliferation of graphics hardware — terminals, scopes, pen plotters, microfilm, color copiers, personal computers — has been accompanied by a steady development of software for graphical data analysis. Computer graphics facilities are now widely available at a reasonable cost, and this book has a relevance today that it would not have had prior to, say, 1970.

1.2 WHAT IS A GRAPHICAL METHOD FOR ANALYZING DATA?

The graphical displays in this book are visual portrayals of quantitative information. Most fall into one of two categories, displaying either the data themselves or quantities derived from the data. Usually, the first type of display is used when we are exploring the data and are not fitting models, and the second is used to enhance numerical statistical analyses that are based on assumptions about relationships in the data. For example, suppose the data are the heights x_i and weights y_i of a group of people. If we knew nothing about height and weight, we could still explore the association between them by plotting y_i against x_i; but if we have assumed the relationship to be linear and have fitted a linear function to the data using classical least squares, we will want to make a number of plots of derived quantities such as residuals from the fit to check the validity of the assumptions, including the assumptions implied by least squares.

If you have not already done so, you might want to stop reading for a moment, leaf through the book, and look at some of the figures. Many of them should look very familiar since they are standard Cartesian plots of points or curves. Figures 1.2 and 1.3, which reappear later in Chapters 3 and 7, are good examples. In these cases the main focus is not on the details of the vehicle, the Cartesian plot, but on what we choose to plot; although Figures 1.2 and 1.3 are superficially similar to each other, each being a simple plot of several dozen discrete points, they have very different meanings as data displays. While these displays are visually familiar, there are other displays that will probably seem unfamiliar. For example, Figure 1.4, which comes from Chapter 5, looks like a forest of misshapen trees. For such displays we discuss not only what to plot, but some of the steps involved in constructing the plot.

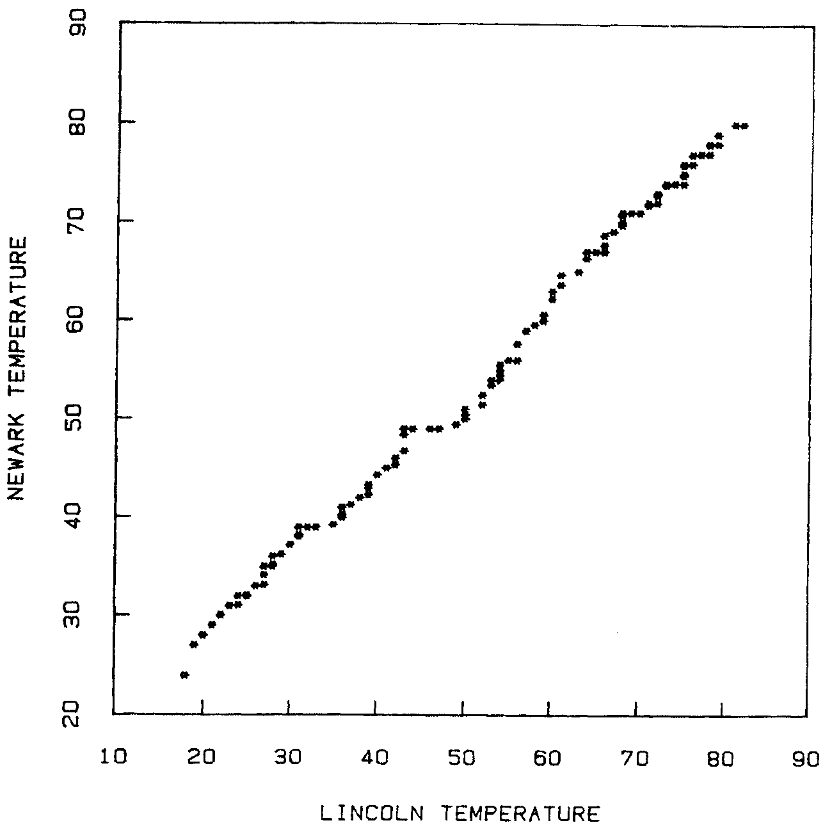

Figure 1.2 Empirical quantile-quantile plot of Newark and Lincoln monthly temperatures.

1.3 A SUMMARY OF THE CONTENTS

The book is organized according to the type of data to be analyzed and the complexity of the data analysis task. We progress from simple to complex situations. Chapters 2 to 5 contain mostly exploratory methods in which the raw data themselves are displayed. Chapter 2 describes methods for portraying the distribution of a single set of observations, for showing how the data spread out along the observation scale. Methods for comparing the distributions of several data sets are covered in Chapter 3. Chapter 4 deals with paired measurements, or two-

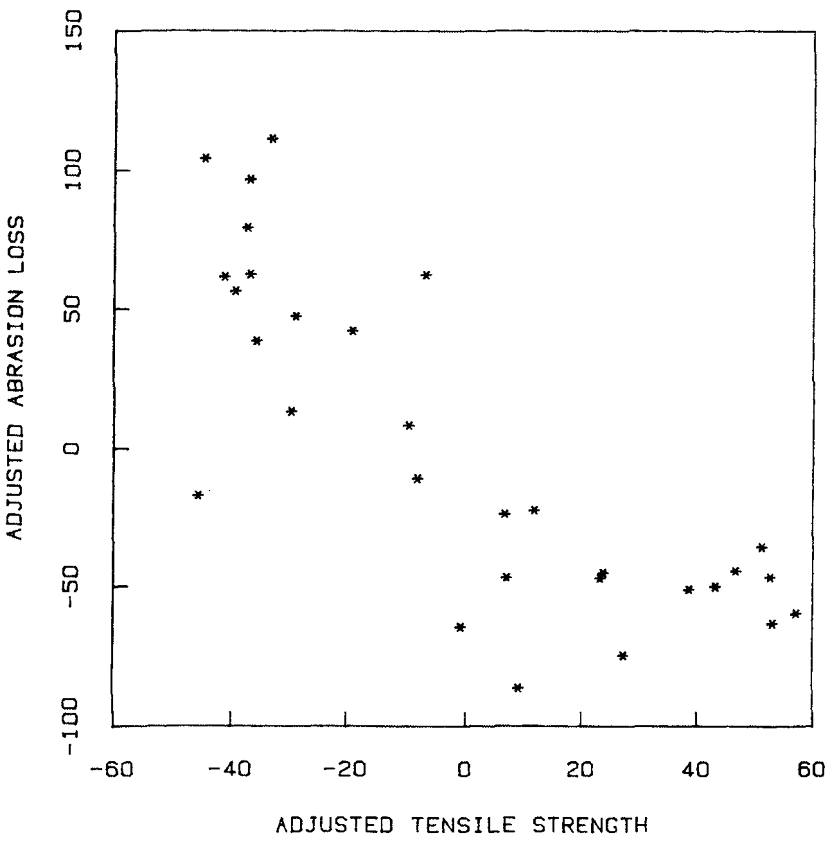

Figure 1.3 Adjusted variable plot of abrasion loss versus tensile strength, both variables adjusted for hardness.

dimensional data; the graphical methods there help us probe the relationship and association between the two variables. Chapter 5 does the same for measurements of more than two variables; an example of such multidimensional data is the heights, weights, blood pressures, pulse rates, and blood types of a group of people.

Chapters 6 and 7 present methods for studying data in the context of statistical models and for plotting quantities derived from the data. Here the displays are used to enhance standard numerical statistical analyses frequently carried out on data. The plots allow the investigator to probe the results of analyses and judge whether the data support the

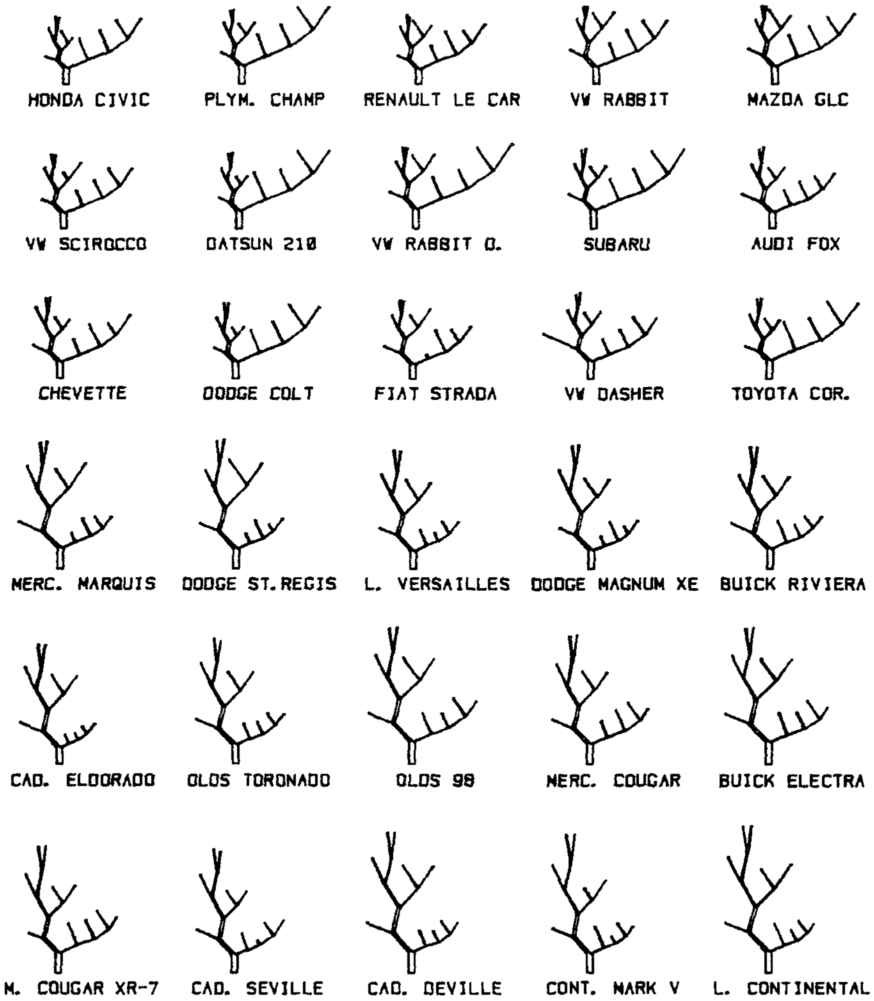

Figure 1.4 Kleiner-Hartigan trees.

underlying assumptions. Chapter 6 is about probability plots, which are designed for assessing formal distributional assumptions for the data. Chapter 7 covers graphical methods for regression, including methods for understanding the fit of the regression equation and methods for assessing the appropriateness of the regression model.

Chapter 8 is a general discussion of graphics including a number of principles that help us judge the strengths and weaknesses of graphical displays, and guide us in designing new ones.

The Appendix contains most of the data sets used in the examples of the book and other data sets referred to just in the exercises.

1.4 THE SELECTION AND PRESENTATION OF MATERIALS

We have selected a group of graphical methods to treat in detail. Our plan has been first to give all the information needed to construct a plot, then to illustrate the display by applying it to at least one set of data, and finally to describe the usefulness of the method and the role it plays in data analysis.

The process for selecting methods to feature was a parochial one: we chose methods that we use in our own work and that have proved successful. Such a selection process is necessary, for we cannot write intelligently about methods that we have not used. We have had to exclude many promising ones with which we are just beginning to have some experience and others that we are simply unfamiliar with. Some of these are briefly described and referenced in "Further Reading" sections at the ends of chapters.

1.5 DATA SETS

Almost all of the data sets used in this book to illustrate the methods are in the Appendix together with other data sets that are treated in the exercises. There are two reasons for this. One is to provide data for the reader to experiment with the graphical methods we describe. The second is to allow the reader to challenge more readily our methodology and devise still better graphical methods for data analysis. Naturally, we encourage readers to collect other data sets of suitable nature to experiment further.

1.6 QUALITY OF GRAPHICAL DISPLAYS

The plots shown in this book are generally in the form we would produce in the course of analyzing data. Most of them represent what you could expect to produce, routinely, from a good graphics package and a reasonably inexpensive graphics device, such as a pen plotter. A few plots have been done by hand. None were produced on special, expensive graphics devices. The point is that the value of graphs in data analysis comes when they show important patterns in the data, and plain, legible, well-designed plots can do this without the expense and delay involved with special presentation-quality graphics devices.

Naturally, when the plots are to be used for presentation or publication rather than for analysis, making the graphics elegant and aesthetically pleasing would be important. We have deliberately not made such changes here. These are working plots, part of the everyday business of data analysis.

1.7 HOW SHOULD THIS BOOK BE USED?

Readers who experiment with the graphical methods in this book by trying them in the exercises, on the data in the Appendix, and on their own data will learn far more from this book than passive readers.

It is usually easy to understand the details of making a particular plot. What is more difficult is to acquire the judgment necessary for successful application of the method: When should the method be used? For what types of data? For what types of problems? What patterns should be looked for? Which patterns are significant and which are spurious? What has been learned about the data in its application context by looking at the plots? The book can go just so far in dealing with these matters of judgment. Readers will need to take themselves the rest of the way.

2

Portraying the Distribution of a Set of Data

2.1 INTRODUCTION

A simple but common need arises in data analysis when we have a single set of numbers that are measurements, observations, or values of some variable, and we want to understand their basic characteristics as a collection. For example, if we consider the gross national product of all countries in the United Nations in 1980, we might ask: What is a "typical" or "average" or "central" value for the whole set? How spread out are the data around the center? How far are the most extreme values (both high and low) from the typical value? What fraction of the numbers are less than the value for one particular country (our own, say)?

In short, we need to understand the distribution of the set of data values: where they lie along the measurement axis, and what kind of pattern they form. This often means asking additional questions. What are the quartiles of the distribution (the 25 percent and 75 percent points along the observation scale)? Are any of the observations outliers, that is, values that seem to lie too far from the majority? Are there repeated values? What is the density or relative concentration of observations in various intervals along the measurement scale? Do the data accumulate at the middle of their range, or at one end, or at several places? Are the data symmetrically distributed?

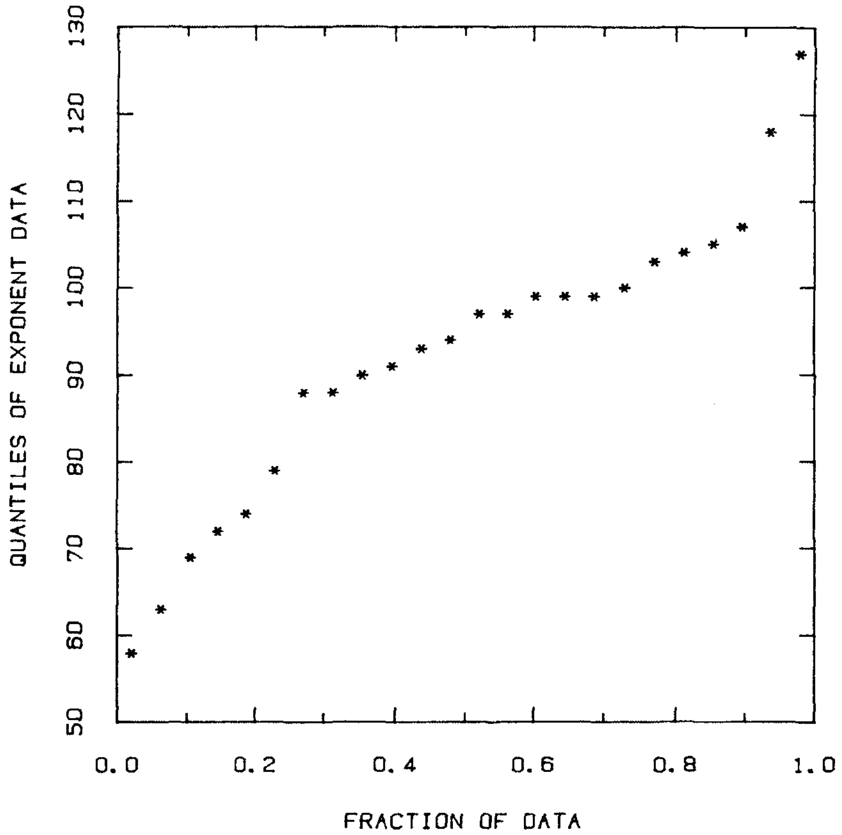

Figure 2.1 Quantile plot of the exponent data. The y coordinates of the plotted points are the ordered observations.

One way to present the distribution of a set of data is to present the data in a table. Many questions can be answered by carefully studying a table, especially if the data have first been ordered from smallest to largest (or the reverse). In a sense, a table contains all the answers, because apart from possible rounding, it presents all of the data.

However, many distributional questions are difficult to answer just from peering at a table. Plots of the data can be far more revealing, even though it may be harder to read exact data values from a plot. This chapter discusses a variety of plots designed for studying the distribution of a set of data.

Two sets of data will be used to illustrate the methodology. One is the daily maximum ozone concentrations at ground level recorded between May 1, 1974 and September 30, 1974 at a site in Stamford, Connecticut. (There are 17 missing days of data due to equipment malfunction.) The current federal standard for ozone states that the concentration should not exceed 120 parts per billion (ppb) more than one day per year at any particular location. A day with ozone concentration above 200 ppb is regarded as heavily polluted. The data are given in the Appendix.

The second set of data is from an experiment in perceptual psychology. A person asked to judge the relative areas of circles of varying sizes typically judges the areas on a perceptual scale that can be approximated by

$$\text{judged area} = \alpha(\text{true area})^{\beta}.$$

For most people the exponent β is between .6 and 1. Apart from random error, a person with an exponent of .7 who sees two circles, one twice the area of the other, would judge the larger one to be only $2^{.7} = 1.6$ times as large. Our second set of data is the set of measured exponents (multiplied by 100) for 24 people from one particular experiment (Cleveland, Harris, and McGill, 1982).

In this chapter we are concerned only with data values themselves, not with any particular ordering of them. (The ozone data have an ordering in time, for instance, and the exponent data could be ordered, say, by the ages of the people in the experiment.) We will usually refer to raw (unordered) data by "y_i for $i = 1$ to n", and to ordered data by "$y_{(i)}$ for $i = 1$ to n." The parentheses in the subscript simply mean that $y_{(1)}$ is the smallest value, $y_{(2)}$ is the second smallest, and so on.

2.2 QUANTILE PLOTS

A good preliminary look at a set of data is provided by the quantile plot which is shown for the exponent data in Figure 2.1. Before describing it, we must define "quantile".

The concept of quantile is closely connected with the familiar concept of percentile. When we say that a student's college board exam score is at the 85th percentile, we mean that 85 percent of all college board scores fall below that student's score, and that 15 percent of them fall above. Similarly, we will define the .85 quantile of a set of data to be a number on the scale of the data that divides the data into two groups, so that a fraction .85 of the observations fall below and a fraction .15 fall above. We will call this value $Q(.85)$. The only difference between percentile and quantile is that percentile refers to a percent of the set of data and quantile refers to a fraction of the set of data. Figure 2.2 depicts $Q(.85)$ for the ozone data plotted along a number line.

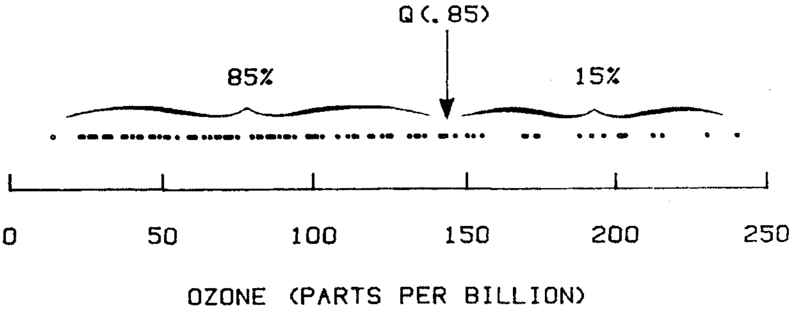

Figure 2.2 The Stamford ozone data, showing the .85 quantile.

Unfortunately, this definition runs into complications when we actually try to compute quantiles from a set of data. For instance, if we want to compute the .27 quantile from 10 data values, we find that each observation is 10 percent of the whole set, so we can split off a fraction of .2 or .3 of the data, but there is no value that will split off a fraction of exactly .27. Also, if we were to put the split point exactly at an observation, we would not know whether to count that observation in the lower or upper part.

To overcome these difficulties, we construct a convenient operational definition of quantile. Starting with a set of raw data y_i, for $i = 1$ to n, we order the data from smallest to largest, obtaining the sorted data $y_{(i)}$, for $i = 1$ to n. Letting p represent any fraction between 0 and 1, we begin by defining the quantile $Q(p)$ corresponding to the fraction p as follows: Take $Q(p)$ to be $y_{(i)}$ whenever p is one of the fractions $p_i = (i-.5)/n$, for $i = 1$ to n.

Thus, the quantiles $Q(p_i)$ of the data are just the ordered data values themselves, $y_{(i)}$. The quantile plot in Figure 2.1 is a plot of $Q(p_i)$ against p_i for the exponent data. The horizontal scale shows the fractions p_i and goes from 0 to 1. The vertical scale is the scale of the original data. Except for the way the horizontal axis is labeled, this plot would look identical to a plot of $y_{(i)}$ against i.

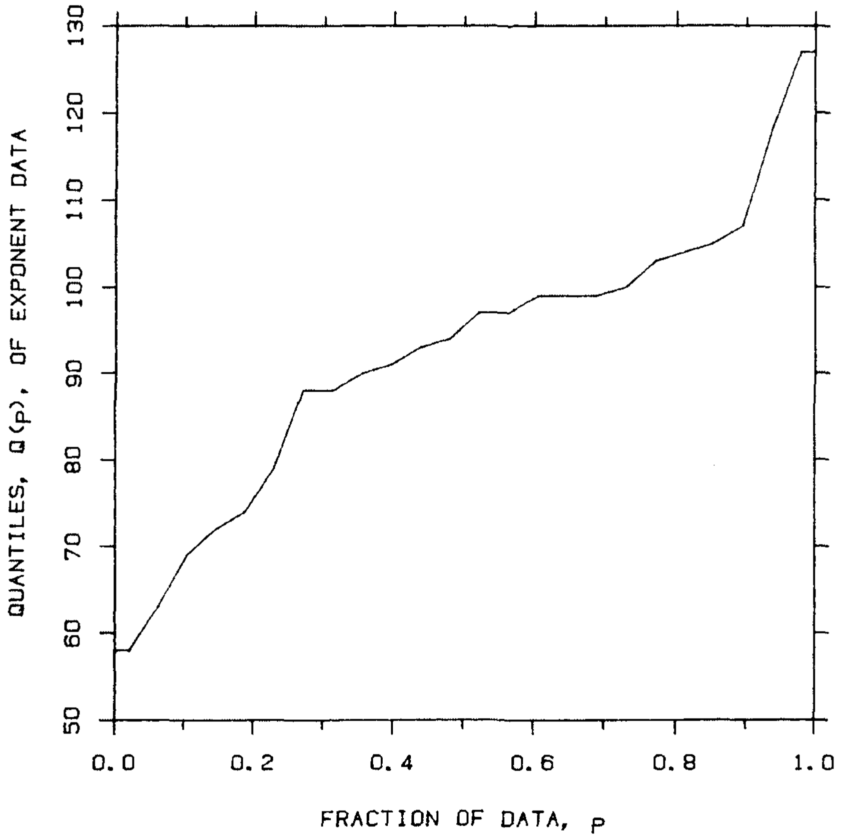

Figure 2.3 Interpolated quantiles for the exponent data.

So far, we have only defined the quantile function $Q(p)$ for certain discrete values of p, namely p_i. Often this is all we need; in other cases, we extend the definition of $Q(p)$ within the range of the data by simple interpolation. In Figure 2.1 this means connecting consecutive points with straight line segments, leading to Figure 2.3. In symbols, if p is a fraction f of the way from p_i to p_{i+1}, then $Q(p)$ is defined to be

$$Q(p) = (1-f)Q(p_i) + fQ(p_{i+1}).$$

We cannot use this formula to define $Q(p)$ outside the range of the data, where p is smaller than $.5/n$ or larger than $1\text{-}.5/n$. Extrapolation is a tricky business; if we must extrapolate we will play safe and define $Q(p) = y_{(1)}$ for $p < p_1$ and $Q(p) = y_{(n)}$ for $p > p_n$, which produces the short horizontal segments at the beginning and end of Figure 2.3.

Why do we take p_i to be $(i-.5)/n$ and not, say i/n? There are several reasons, most of which we will not go into here, since this is a minor technical issue. (Several other choices are reasonable, but we would be hard pressed to see a difference in any of our plots.) We will mention only that when we separate the ordered observations into two groups by splitting exactly on an observation, the use of $(i-.5)/n$ means that the observation is counted as being half in the lower group and half in the upper group.

The median, $Q(.5)$, is a very special quantile. It is the central value in a set of data, the value that divides the data into two groups of equal size. If n is odd, the median is $y_{((n+1)/2)}$; if n is even there are two values of $y_{(i)}$ equally close to the middle and our interpolation rule tells us to average them, giving $(y_{(n/2)} + y_{(n/2+1)})/2$. Two other important quantiles with special names are the lower and upper quartiles, defined as $Q(.25)$ and $Q(.75)$; they split off 25 percent and 75 percent of the data, respectively. The distance from the first to the third quartile, $Q(.75) - Q(.25)$, is called the interquartile range and can be used to judge the spread of the bulk of the data.

Many important properties of the distribution of a set of data are conveyed by the quantile plot. For example, the medians, quartiles, interquartile range, and other quantiles are quite easy to read from the plot. For the exponent data in Figure 2.1 we see that the median is about 95 and that a large fraction of points lie between 85 and 105. Thus, most of the subjects have a perceptual scale that does not deviate markedly from the area scale, which corresponds to the value 100. But a few subjects do have values quite different from 100. In fact, the total range (maximum minus minimum) is seen to be about 70. The subject with the smallest exponent, 58, comes close to judging some linear aspect of circles, such as diameter, rather than area. (A value of 50 corresponds to judging linear aspects exactly.)

Figure 2.4 is a quantile plot of the ozone data. It shows that the median ozone is about 80 ppb. The value 120 ppb is roughly the .75 quantile; thus the federal standard in Stamford was exceeded about 25% of the time. The highest concentration is somewhat less than 250 ppb and only 8 values are above 200 ppb (corresponding to days heavily polluted with ozone). The two smallest values of 14 ppb seem somewhat out of line with the pattern of points at the low end.

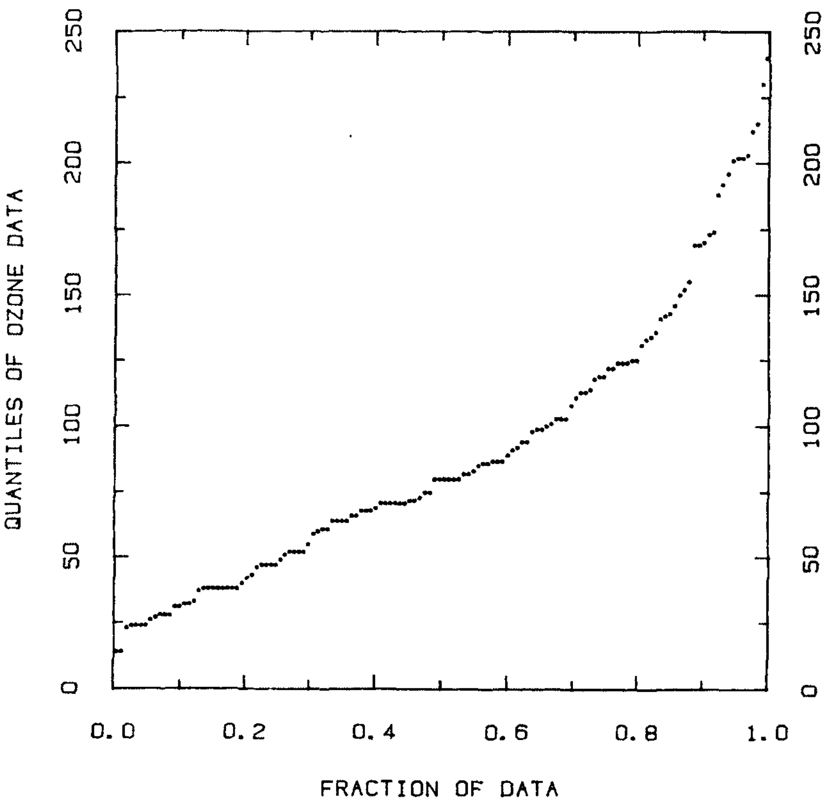

Figure 2.4 Quantile plot of the Stamford ozone data.

The local density or concentration of the data is conveyed by the local slope of the quantile plot; the flatter the slope, the greater the density of points. The rough overall density impression for the ozone data conveyed by Figure 2.4 is one in which the density decreases with larger ozone values. The highest local density of points occurs when there are many measurements with exactly the same value. This is revealed on the quantile plot by a string of horizontal points. For example, in Figure 2.4 there are two such strings of length 6 between 50 ppb and 100 ppb, and another of length 8 at about 35 ppb. A more detailed description of the ozone density will be given in Section 2.8 where a display specifically designed to convey density will be described.

The quantile plot is a good general purpose display since it is fairly easy to construct and does a good job of portraying many aspects of a distribution. Three convenient features of the plot are the following: First, in constructing it, we do not make any arbitrary choices of parameter values or cell boundaries (as we must for several of the displays to be described shortly), and no models for the data are fitted or assumed. Second, like a table, it is not a summary but a display of all the data. Third, on the quantile plot every point is plotted at a distinct location, even if there are exact duplicates in the data. The number of points that can be portrayed without overlap is limited only by the resolution of the plotting device. For a high resolution device several hundred points are easily distinguished.

2.3 SYMMETRY

We often use the idea of symmetry in data analysis. The essence of symmetry is that if you look at the reflection of a symmetric object in a mirror, its appearance remains the same. Since a mirror reverses left and right, this means that an object is symmetric if every detail that occurs on the left also occurs on the right, and at the same distance from an imaginary line down the center.

The distribution of a set of data is symmetric if a plot of the points along a simple number line is symmetric in the usual sense. The sketch in Figure 2.5 shows such a plot of six fictitious symmetric data values,

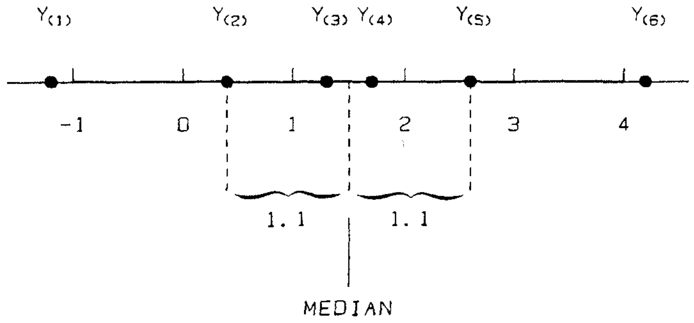

Figure 2.5 Six fictitious symmetric data values.

−1.2, 0.4, 1.3, 1.7, 2.6, and 4.2. The center of symmetry must be the median, and the sketch shows that $y_{(2)}$ and $y_{(5)}$ are equidistant from the center, that is,

$$\text{median} - y_{(2)} = y_{(5)} - \text{median} = 1.1.$$

The general requirement for symmetry is

$$\text{median} - y_{(i)} = y_{(n+1-i)} - \text{median}, \quad \text{for } i = 1 \text{ to } n/2.$$

(If n is odd we can use $(n+1)/2$ instead of $n/2$.) Of course, just as faces and others things that we regard as symmetric in real life are not exactly symmetric, so data will not be exactly symmetric. We will look for approximate symmetry.

We can also characterize symmetry in terms of the quantile function. Since the median is $Q(.5)$, we say that the data are symmetrically distributed if

$$Q(.5) - Q(p) = Q(1-p) - Q(.5) \quad \text{for all } p, \ 0 < p < .5.$$

When data are asymmetric in a way that makes the quantiles on the right progressively further from the median than the corresponding quantiles on the left, then we say that the data are skewed to the right, or toward large values.

The quantile plot can be used to examine data for symmetry. If the data are symmetric the plot itself will not be symmetric in the usual sense; rather, the points in the top half of the plot will stretch out toward the upper right in the same way that the points in the bottom half stretch out toward the lower left. This is shown for our artificial data in Figure 2.6. When the data are skewed toward large values, then the top of the quantile plot extends upward more sharply. Figure 2.4 shows that the ozone data are skewed, but in Figure 2.1 the exponent data appear to be nearly symmetric. Section 2.8 discusses a plot specifically designed for investigating symmetry in data.

There are several reasons why symmetry is an important concept in data analysis. First, the most important single summary of a set of data is the location of the center, and when data are symmetric the meaning of "center" is unambiguous. We can take center to mean any of the following three things, since they all coincide exactly for symmetric data, and they are close together for nearly symmetric data: (1) the center of symmetry, (2) the arithmetic average or center of gravity, (3) the median or 50% point. Furthermore, if the data have a single point of highest concentration instead of several (that is, they are unimodal), then we can add to the list (4) the point of highest concentration. When data are far from symmetric, we may have trouble

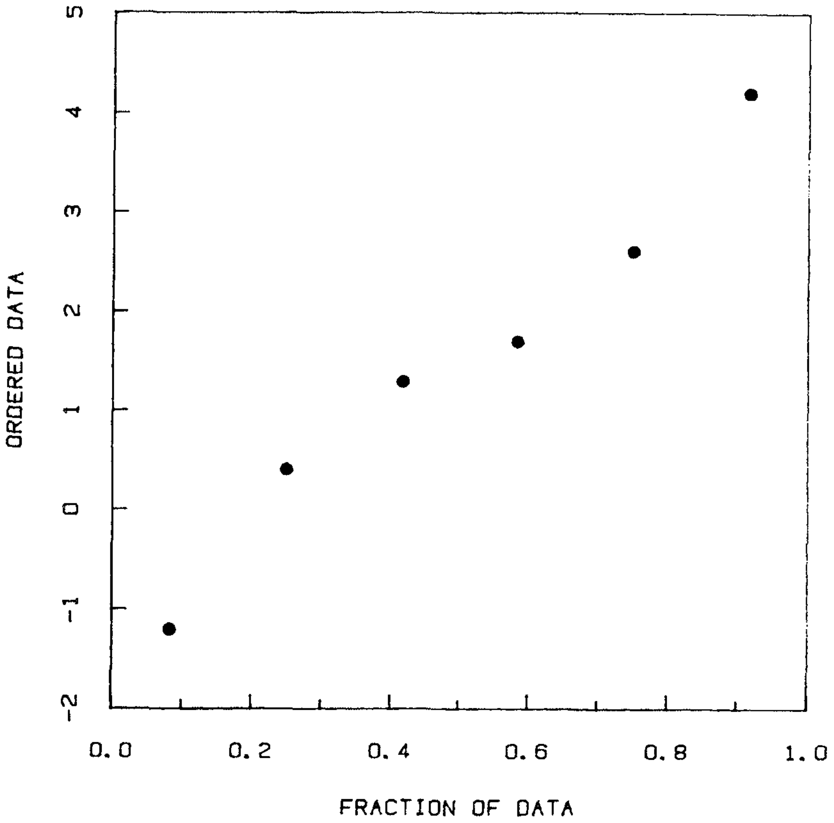

Figure 2.6 Quantile plot of the six fictitious symmetric data values of Figure 2.5.

even agreeing on what we mean by center; in fact, the center may become an inappropriate summary for the data.

Symmetry is also important because it can simplify our thinking about the distribution of a set of data. If we can establish that the data are (approximately) symmetric, then we no longer need to describe the shapes of both the right and left halves. (We might even combine the information from the two sides and have effectively twice as much data for viewing the distributional shape.)

Finally, symmetry is important because many statistical procedures are designed for, and work best on, symmetric data. For example, the simple and common practice of summarizing the spread of a set of data by quoting a single number such as the standard deviation or the interquartile range is only valid, in a sense, for symmetric data. For readers familiar with the normal or Gaussian distribution (which we do not discuss until Chapter 6), we mention that whereas the normal distribution is the foundation for many classical statistical procedures, symmetry alone underlies many modern robust statistical methods. The modern procedures have wider applicability because normality is often an unrealistic requirement for data, but approximate symmetry is often attainable. Interestingly, symmetry is a basic property of the normal distribution!

2.4 ONE-DIMENSIONAL SCATTER PLOTS

A simple way to portray the distribution of the data is to plot the data y_i along a number line or axis labeled according to the measurement scale. The resulting one-dimensional scatter diagram or scatter plot is shown in Figure 2.7 for the ozone data. Note that if we horizontally project the points on a quantile plot onto the vertical axis, the result is a vertical one-dimensional scatter plot. In this sense the quantile plot can be thought of as an expansion into two dimensions of the one-dimensional scatter plot.

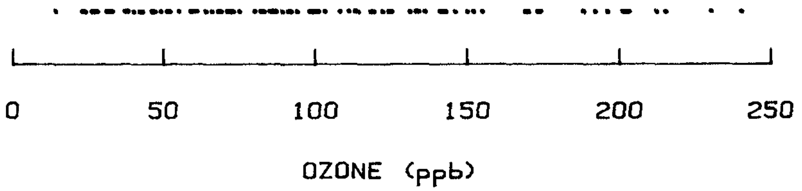

Figure 2.7 One-dimensional scatter plot of the ozone data.

The main virtue of the one-dimensional scatter plot is its compactness. This allows it to be used in the margins of other displays to add information. (An example will be shown later in the chapter.) In a one-dimensional scatter plot we can clearly see the maximum and minimum values of the data. Provided there is not too much overlap we can also get very rough impressions of the center of the data, the spread, local density, symmetry, and outliers. Furthermore the plot is easy to construct and to explain to others.

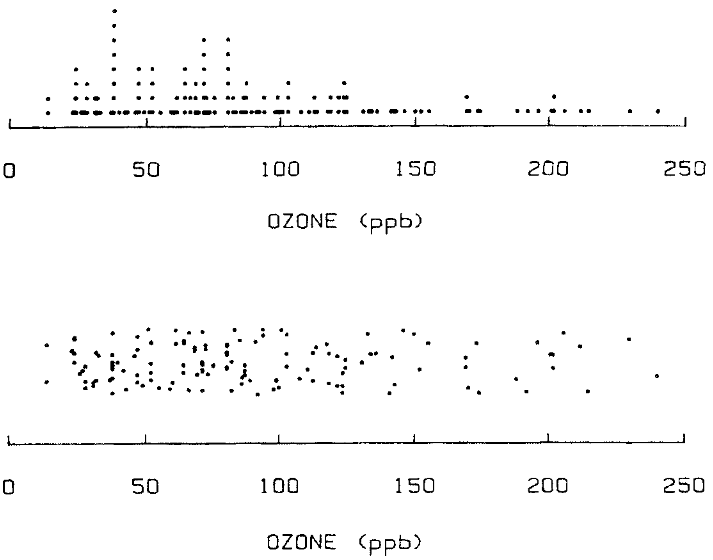

Figure 2.8 A one-dimensional scatter plot of the ozone data with stacking (top panel) and a jittered one-dimensional scatter plot (bottom panel).

However, a price is paid for collapsing the two-dimensional quantile plot to the one-dimensional scatter plot. Individual quantiles can no longer be found easily, and visual resolution of the points is more likely to be a problem even for moderately many points. We obtain maximum resolution by using a plotting character that is narrow such as a dot or a short vertical line instead of, say, an asterisk or an ×. But this does not solve the problem of exact duplicates. If $y_{(i)} = y_{(i+1)}$, then the plotting locations for $y_{(i)}$ and $y_{(i+1)}$ on the one-dimensional scatter plot are the same. (Note that this did not happen on the quantile plot.) For example, there are several repeated values in the ozone data which are not resolved in Figure 2.7. One way to alleviate this problem is to stack points, that is, to displace them vertically when they coincide with others. A one-dimensional scatter plot of the ozone data with stacking is shown in the top panel of Figure 2.8. This, however, is only a solution to the problem of exact overlap and does not help us when there are a lot of points that crowd one another. Another method that helps to alleviate both exact overlap and crowding is vertical jitter, which is illustrated in the bottom panel of Figure 2.8. Let u_i, $i = 1$ to n, be the integers 1 to n in random order. The vertical jitter is achieved by plotting u_i against y_i with u_i on the vertical axis and y_i on the horizontal axis. To keep the display nearly one-dimensional the range of the vertical axis — that is, the actual physical distance — is kept small compared to the range of the horizontal axis, and, of course, we do not need to indicate the vertical scale on the plot. The vertical jitter in Figure 2.8 appears to have done a good job of reducing the overlap in Figure 2.7.

2.5 BOX PLOTS

It is usually important to take an initial look at all of the data, perhaps with a quantile plot, to make sure that no unusual behavior goes undetected. But there are also situations and stages of analysis where it is useful to have summary displays of the distribution. One simple method of summarization, called a box plot (Tukey, 1977), is illustrated in Figure 2.9 for the ozone data and in Figure 2.10 for the exponent data.

In the box plot the upper and lower quartiles of the data are portrayed by the top and bottom of a rectangle, and the median is portrayed by a horizontal line segment within the rectangle. Dashed lines extend from the ends of the box to the *adjacent values* which are defined as follows. We first compute the interquartile range, $IQR = Q(.75) - Q(.25)$. In the case of the exponent data the quartiles are 83.5 and 101.5 so that $IQR = 18$. The upper adjacent value is defined to be the largest observation that is less than or equal to the upper quartile plus $1.5 \times IQR$. Since this latter value is 128.5 for the exponent data, the upper adjacent value is simply the largest observation, 127. The lower adjacent value is defined to be the smallest observation that is greater than or equal to the lower quartile minus $1.5 \times IQR$. For the exponent data, it is the smallest observation, 58. Thus for the exponent data, the adjacent values are the extreme values. If any y_i falls outside the range of the two adjacent values, it is called an *outside value* and is plotted as an individual point; for the exponent data there are no outside values and for the ozone data there are two.

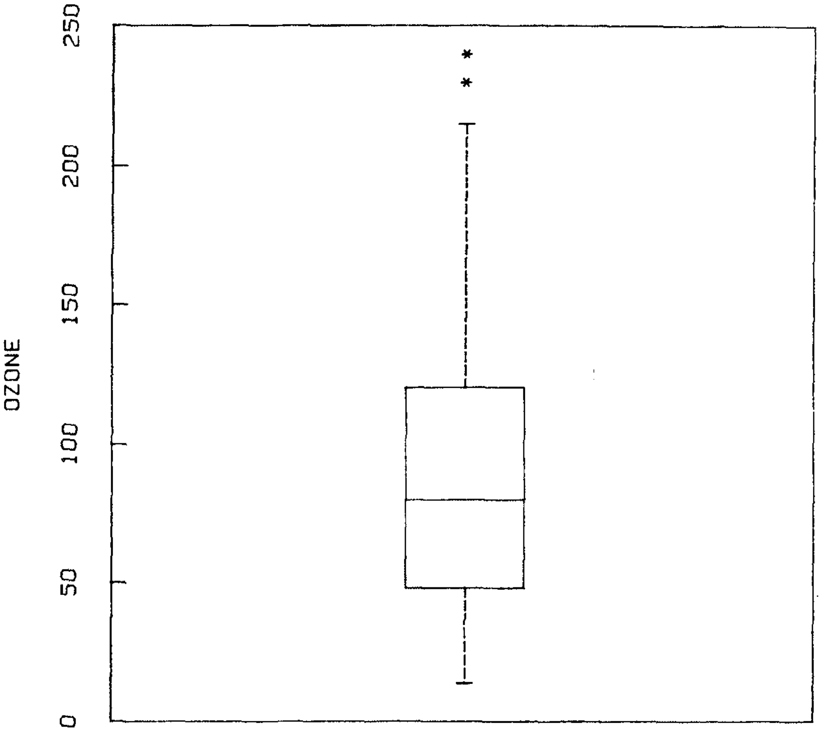

Figure 2.9 A box plot of the ozone data.

The box plot gives a quick impression of certain prominent features of the distribution. The median shows the center, or location, of the distribution. The spread of the bulk of the data (the central 50%) is seen as the length of the box. The lengths of the dashed lines relative to the box show how stretched the tails of the distribution are. The individual outside values give the viewer an opportunity to consider the question of outliers, that is, observations that seem unusually, or even implausibly, large or small. Outside values are not necessarily outliers (indeed, the ozone quantile plot suggests that the two ozone outside values are not), but any outliers will almost certainly appear as outside values.

The box plot allows a partial assessment of symmetry. If the distribution is symmetric then the box plot is symmetric about the median: the median cuts the box in half, the upper and lower dashed

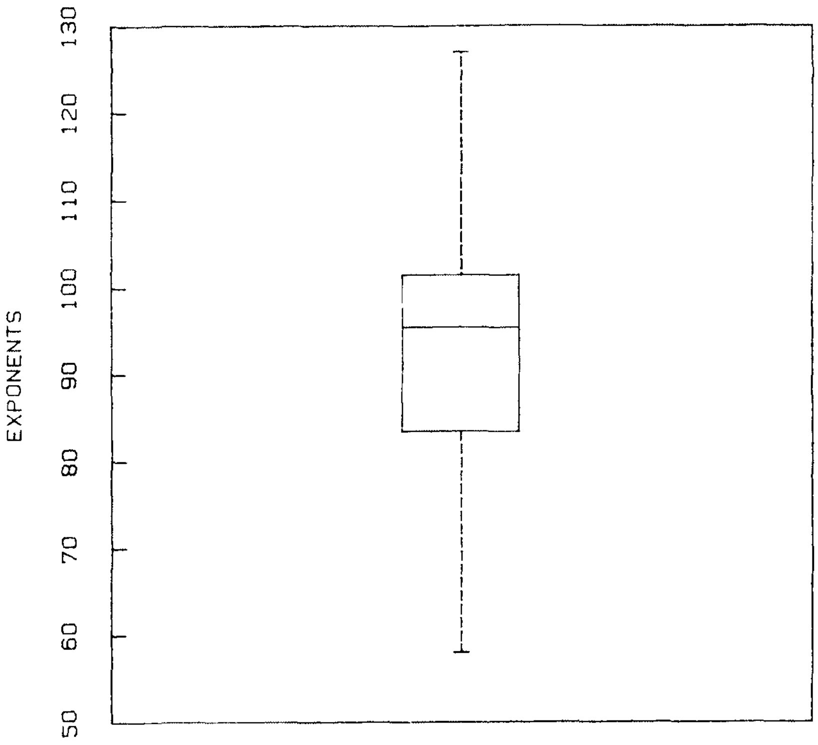

Figure 2.10 A box plot of the exponent data.

lines are about the same length, and the outside values at top and bottom, if any, are about equal in number and symmetrically placed. There can be asymmetry in the data not revealed by the box plot, but the plot usually gives a good rough indication. The box plot in Figure 2.9 shows that the ozone data are not symmetric. The upper components are stretched relative to their counterparts below the median, revealing that the distribution is skewed to the right. For the exponent data the box plot in Figure 2.10 suggests that the tails are symmetric, but that the median is high relative to the quartiles. Recall from Section 2.3 that the quantile plot of these data in Figure 2.1 suggests the data are approximately symmetric. To resolve this apparent contradiction, we can look more closely at Figure 2.1. Ignoring the two largest and two smallest values, the rest of the data appear slightly skewed toward small values, which explains the position of tHe median relative to the quartiles. But we should remember that the number of observations in this sample is small and that we would quite likely see different behavior in another sample.

Box plots are useful in situations where it is either not necessary or not feasible to portray all details of the distribution. For example, if many distributions are to be compared, it is difficult to try to compare all aspects of the distributions. In situations where the summary values of the box plot do a good job of conveying the prominent features of the distribution and the less prominent detailed features do not matter, it makes sense to use the box plot and eliminate the unneeded information.

The width of the box, as defined so far, has no particular meaning. The plot can be made quite narrow without affecting its visual impact so that it can be used in situations where compactness is important. This is useful in Chapter 3 when many distributions are being compared and in Chapter 4 when the box plot is added to the margin of another visual display.

2.6 HISTOGRAMS

Another way to summarize a data distribution, one that has a long history in statistics, is to partition the range of the data into several intervals of equal length, count the number of points in each interval, and plot the counts as bar lengths in a histogram. This has been done in Figure 2.11 for the ozone data. The relative heights of the bars represent the relative density of observations in the intervals.

The histogram is widely used and thus is familiar even to most nontechnical people and without extensive explanation. This makes it a convenient way to communicate distributional information to general audiences.

However, as a data analysis device it has some drawbacks. Figure 2.12 is a second histogram of the same ozone data. Below each histogram is a jittered one-dimensional scatter plot to show the relationship of the histogram to the original data. The two histograms give rather different visual impressions, and the differences depend on the fairly arbitrary choice of the number and placement of intervals. This choice determines whether we show more detail, as in Figure 2.12, or retain a smoothness or simplicity, as in Figure 2.11. But even Figure 2.11 is not genuinely smooth, because the bars have sharp corners. The

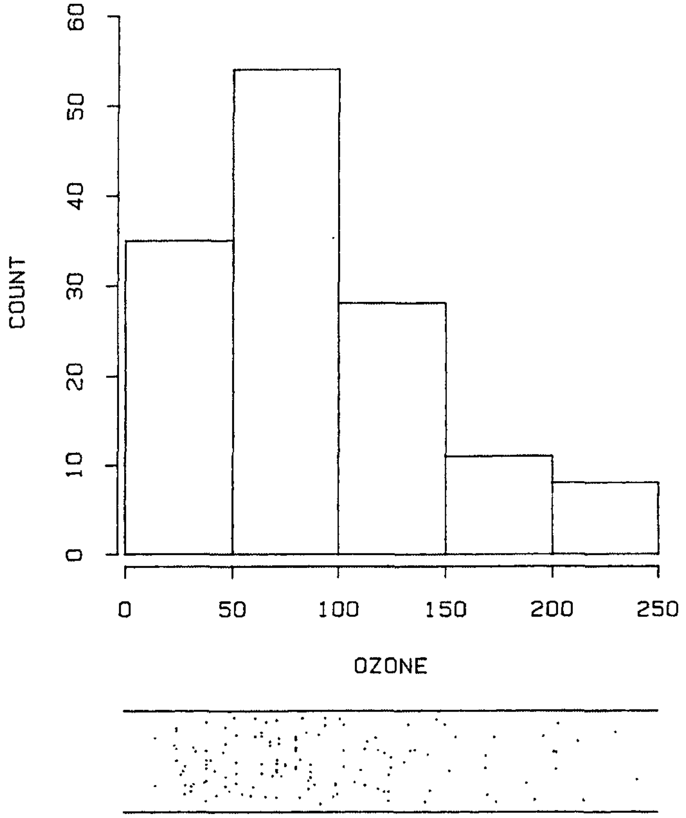

Figure 2.11 Histogram of the ozone data, with a jittered one-dimensional scatter plot.

positions of the corners have little to do with the data; they are an artifact of the histogram construction. (Smoother approaches are discussed in Section 2.9.) Figure 2.11 reveals an additional problem: following common practice, we put the ends of the intervals at convenient numbers (multiples of 50 ppb) so they can be easily read from the plot, but in doing so we have covered up the important nonzero lower bound for ozone and have created the erroneous impression that the density of the data points just below 50 ppb is much less than the density just above 50 ppb.

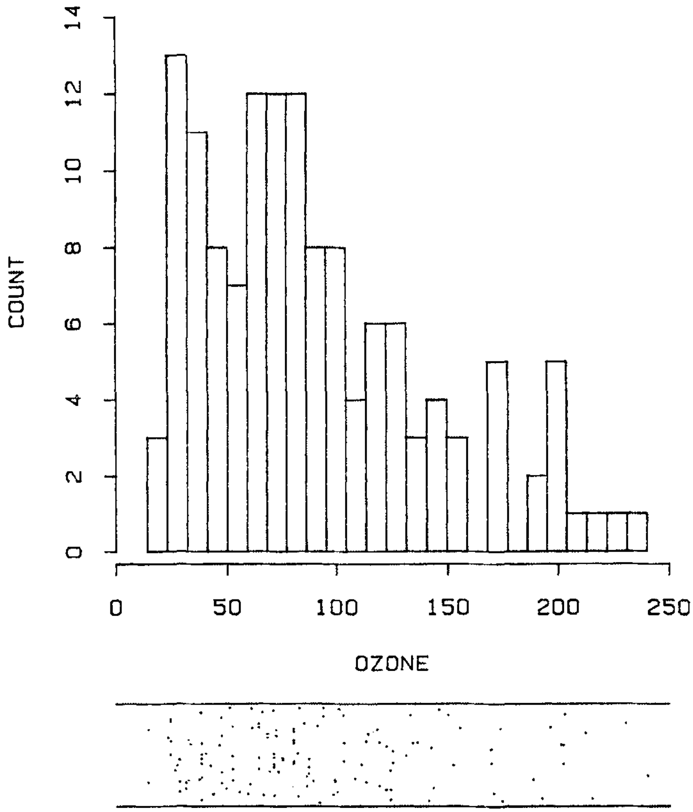

Figure 2.12 Histogram of the ozone data (with 25 intervals), with a jittered one-dimensional scatter plot.

2.7 STEM-AND-LEAF DIAGRAMS

Figure 2.13 shows a stem-and-leaf diagram of the ozone data (Tukey, 1977). It is a hybrid between a table and a graph since it shows numerical values as numerals but its profile is very much like a histogram.

To construct a stem-and-leaf diagram we first write down, to the left of a vertical line, all possible leading digits in the range of the data. Then we represent each data value by writing its trailing digit in the appropriate row to the right of the line. Thus the fifteenth row of the