SEEING AND BELIEVING:

A BEGINNER'S GUIDE

TO

STATISTICAL GRAPHICS

Edited by

Berton H. Gunter

Distributed by

Cleveland Chapter

American Statistical Association

SEEING AND BELIEVING:

A BEGINNER'S GUIDE TO STATISTICAL GRAPHICS

written under the auspices of
The Statistical Graphics Section,
The American Statistical Association

Edited by

Berton H. Gunter

Contributing Authors:

Carol Joyce Blumberg
Kazem Kazempour
Innis Sande
Paul Somerville

* * *

Additional copies of this guide may be purchased by sending a check for $5.00 made out to the CLEVELAND CHAPTER - AMERICAN STATISTICAL ASSOCIATION and mailing it to:

Jerry L. Moreno
Department of Mathematics
John Carroll University
University Heights, OH 44118.

TABLE OF CONTENTS

ACKNOWLEDGEMENT

The authors and editor wish to thank several people whose detailed criticisms of an earlier draft led to substantial improvements in the presentation. In particular, Dr. Howard Wainer of the Educational Testing Service in Princeton, NJ, who is well-known for his efforts to improve statistical graphics, reviewed our exposition to try to make sure we gave and followed good advice. Bob Peterson, mathematics coordinator of the Macomb Intermediate School District and a nationally recognized high school mathematics teacher, provided guidance on how to exposit to high school mathematics teachers and students and also provided the cartoons that appear in the introduction. Kathy Peterson and Rebecca Hickling read the manuscript from the viewpoint of practicing teachers and gave us valuable feedback on our efforts. We have attempted to heed the wise counsel of all -- especially their advice to simplify and clarify -- so that any shortcomings that remain do so despite their best efforts.

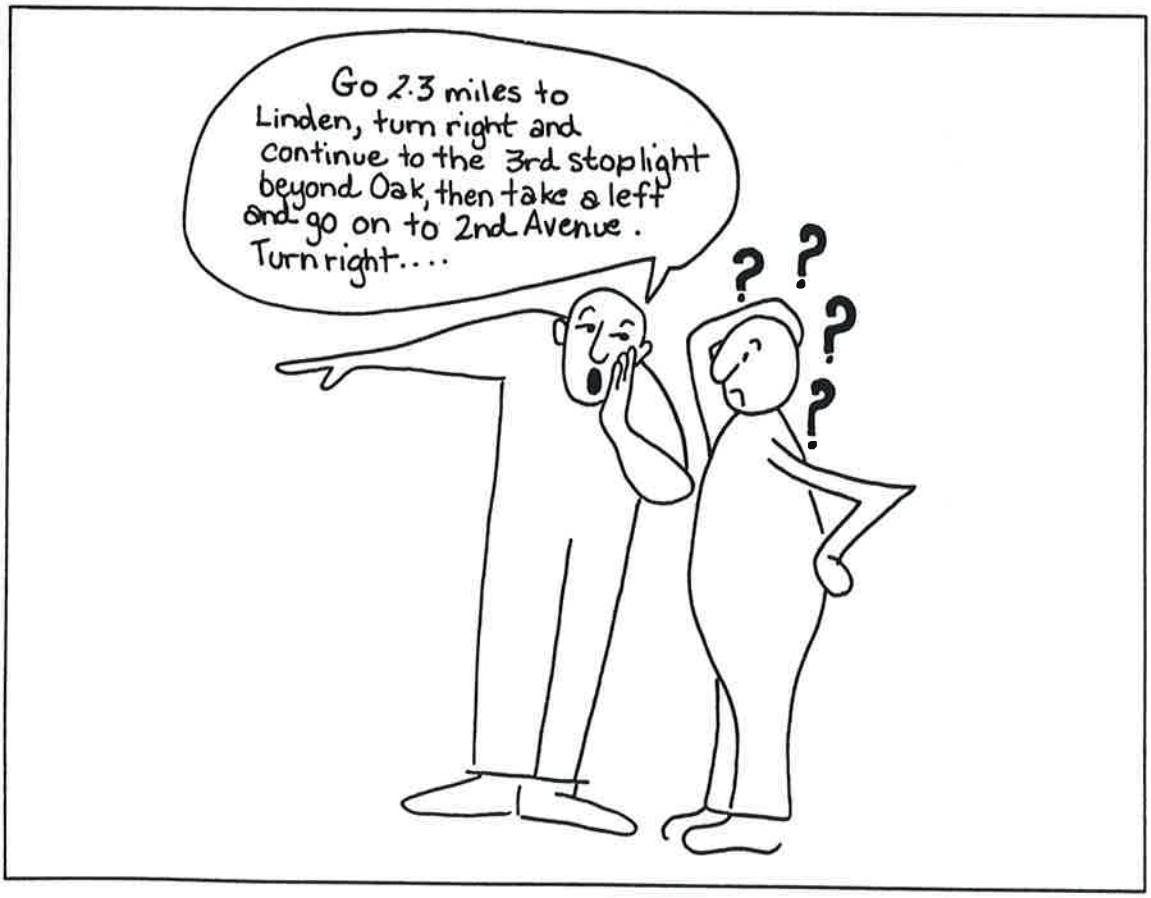
Finally, we would like to acknowledge the financial assistance of SAS, Institute Inc., whose generous contribution has helped defray the cost of printing and distributing the Guide.

INTRODUCTION

by Berton Gunter
Statistical Consultant

Do you remember the last time you saw someone whom you recognized but whose name you forgot? The face was so familiar but the name ... you just couldn't quite recall it.

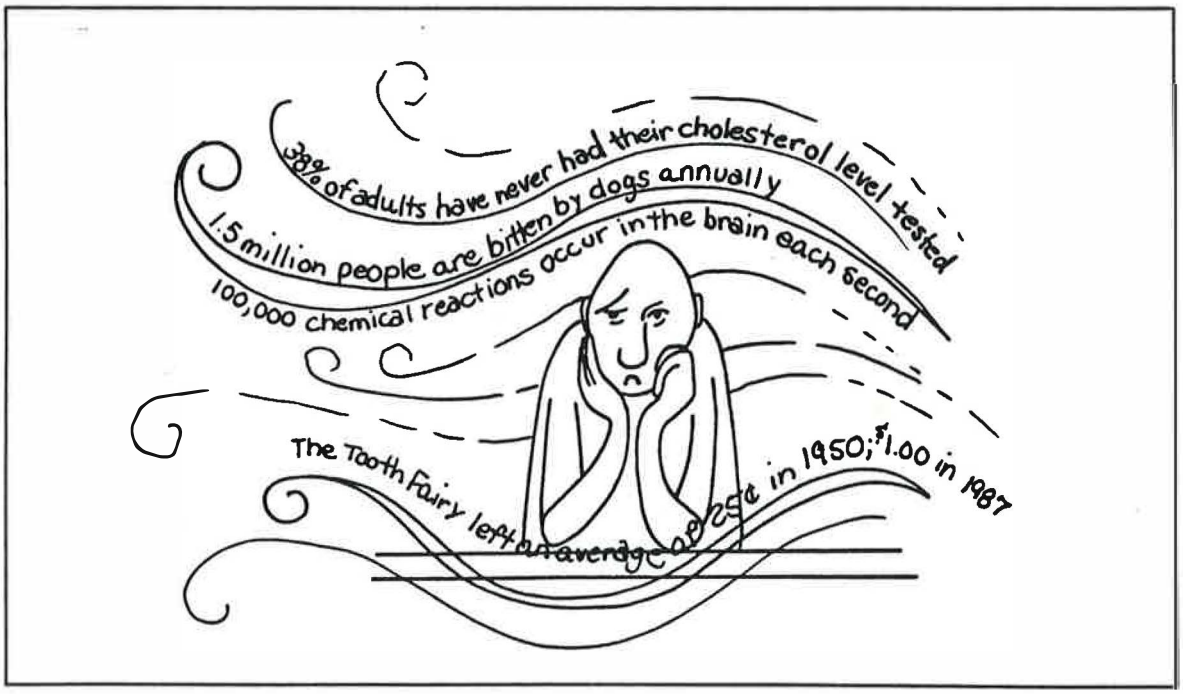
Or if you have lived somewhere else and haven't been back to your former home for a while, try to write down directions to tell someone how to get from your former house or apartment to the supermarket, drugstore, or some other familiar destination. Can you remember all the street names or the number of lights to pass before making turns? Probably not. Nonetheless, if you were suddenly transported to that home and were told to walk,bicycle, or drive the route, you'd probably have no problem finding your way.



Both of these examples illustrate a peculiar fact: even though it requires far less information to write down the name or the directions than to describe the face or the route, we remember the complex pictures and forget the simple words and numerals. This demonstrates a fundamental truth about the human animal -- our brains are wired for pattern recognition, not for symbolic processing. From the time we first greet the world and soon after recognize good old Mom, to the numerous images of people, places, and events that we store as we grow older, we tend to learn and process information as pictures (and sounds, and smells, and touches, and tastes). Numbers and words have their place -- as this introduction demonstrates -- but the way to really understand and communicate understanding is with pictures.

Of course, this is hardly a new idea, but only rather recently have people fully recognized its consequences for analyzing and communicating about data. Science and mathematics have long relied upon pictures and graphs, as a quick glance at practically any science or math textbook will show. But only within the past 20 to 30 years has statistics -- which is concerned with how best to acquire, understand, and communicate about data -- broadly recognized how important and effective graphics can be in data analysis.

This Guide is being published under the auspices of the Statistical Graphics Section of the American Statistical Association (ASA). It aims to introduce readers with no special knowledge of either statistics or graphics to some of the new and exciting ideas in this growing field. When most people think of statistics, visions of numbers and formulas come dancing into their heads. We hope that, after reading this Guide, graphs and pictures will frolic alongside.



We also hope that nearly everyone will find useful and interesting ideas in this booklet. We live in a world overflowing with data: economic and financial reports, political surveys, health and environmental studies, sports statistics, TV ratings, standardized test scores, and population projections are just a few of the "numbers games" that appear in the popular media every day. Graphics can help makes sense of these data.

Think about it. In 1800, most people only needed to know about farming to be good citizens. Only a few lawyers and politicians needed to know the "fancy" skills of reading and writing -- literacy. By 1900, literacy was essential for all, but only a few scientists or census takers needed to know anything about how to gather data and understand what they meant. How that situation has changed! As we approach the year 2000, we find ourselves swimming in a sea of numbers, fed and organized by one of mankind's greatest technological inventions, the digital computer. From choosing health foods and cosmetics, to purchasing a house, to choosing a career, to casting our vote, we confront data full of facts, fallacies, exaggerations, and confusion to help "guide" our choices. Clearly, in such an environment, we need a new kind of quantitative literacy to make sense of it all. The concepts and techniques of statistical graphics are an invaluable part of this quantitative literacy.

This Guide presents both underlying concepts and some specific graphical techniques. We have also included a short annotated bibliography for those who wish to learn more. The organization is as follows:

Chapter 1 gives an overview of some of the basic ideas underlying statistical graphics. The main theme is to answer the following question: what makes a good graph good and a bad graph bad? It turns out that this is not such a simple question to answer, and readers will discover that much -- maybe most-- of the graphics that one finds in popular media these days is pretty awful.

Chapter 2 provides some simple statistical graphical tools that students can use. Many of these ideas follow the examples of the National Council of Teachers of Mathematics/ASA joint Quantitative Literacy Series publications, so readers familar with these materials should find some old friends.

Chapter 3 introduces some more sophisticated -- but still very useful -- ideas and the essential role of the computer in implementing them. Even though statistical graphics changes our focus from numerics to patterns, we need the power of the computer in order to easily produce and manipulate these patterns.

We have tried to present the material clearly, relying on graphics and examples to demonstrate the ideas and avoiding numeric or algebraic manipulation. We hope that readers will find this fun to read as well as informative. But the proof of the pudding comes in the eating: we especially encourage readers to go out and try the ideas on data in which you are interested.

We can almost guarantee that you'll be amazed by what you see and by the effect it has on what you believe.

CHAPTER 1

PRINCIPLES OF GOOD GRAPHICAL DISPLAY

by Kazem Kazempour and Paul Somerville
University of Central Florida

Purposes and Message of a Statistical Graph

The purpose of statistical analysis is to draw conclusions from data. It converts the information in numbers into knowledge on which to base decisions. The decisions can range from which car to buy, to what interest policy is better for the national economy; from how many compact disks to produce for a new release, to predicting world population growth and agricultural needs; from determining the effectiveness and risks of a new drug, to deciding whether buildup of carbon dioxide in the atmosphere requires new world energy policies.

In today's world, the scientific method is used to investigate practically all questions of interest. What this means is that unsupported theories are not acceptable as a means to understand natural and/or social (political, economic, etc.) phenomena. Studies and experiments must be conducted to collect hard data. Theories and explanations that the data support are useful; those that the data contradict must be abandoned. Understanding such data always involves some sort of statistical analysis. For this reason, statistics is sometimes referred to as "the language of science."

Sometimes, statistical analysis requires nothing more than looking at a few numbers and drawing the obvious conclusions. But this is rare. More often than not, both the quantity of the data and the presence of variability make it difficult or impossible to determine much useful just by looking at the numbers (by variability, we mean measurement errors, systematic errors in sampling, experimental error, difficulties in getting reliable information, fuzziness of people's attitides and opinions, and so forth).

Statistical graphs are visual interpretations of data and information. As we noted in the introduction, the human brain is adept at identifying and interpreting patterns. Tables of data and numerical summaries present information, but because they are not in a form that the brain can readily deal with, understanding that information can be a problem. Moreover, a list of numbers is boring! A good graph avoids these problems and is both informative and appealing: the important information leaps out and grabs you.

The most important function of a graph is to summarize and meaningfully display the information in a table of numbers in the simplest and most useful way possible. Not only does this make comprehension easier, but it also helps both novice and expert viewer to remember the message longer.

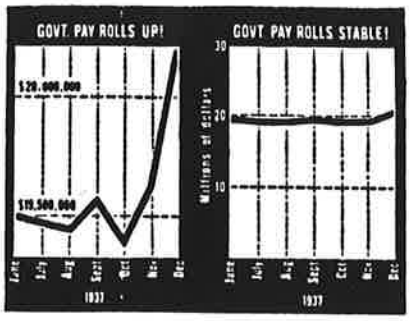
There are many ways to graph data. In this Guide, we shall show you a few simple techniques, but the variety of approaches is really endless. The guiding principle is: use whatever works. However, beware! -- there are no methods that work well in all applications. Methods that do well in one situation may be inappropriate for another. Before choosing any particular approach, you usually should try several in order to determine what the data have to say. Once you have decided what the main message(s) is, you can then choose the best method to convey it. Even so, several iterations of trying out a graph, examining it, editing it and trying out a modified version will probably be necessary before producing a final version.

Importance of Scaling

Scale is an essential part of any meaningful statistical graph. It is the ruler along which the data are displayed and compared. One of the most common errors in statistical graphs is to deliberately or unintentionally distort results by improper scaling. Following are two pairs of graphs, each pair portraying the same data set with the only difference being the scaling.

FIGURE 1.1

(a)



(b)

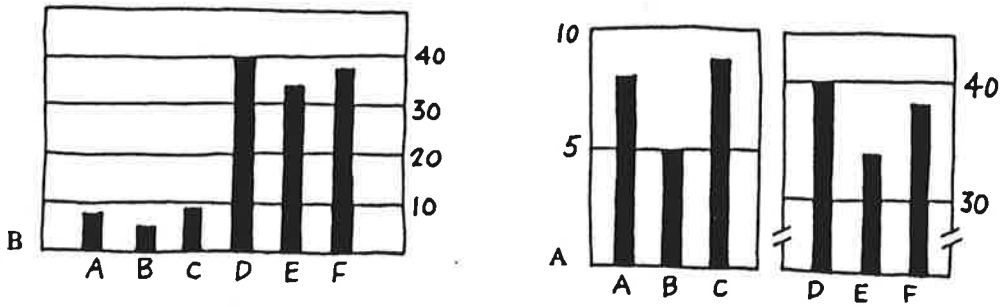


Figure 1.1(a) from How to Lie with Statistics by Darrell Huff, p.65. Reprinted by permission of W.W. Norton & Company. Copyright © 1954 and renewed 1982 by Darrell Huff and Irving Geis.
Figure 1.1(b) from The Designer's Guide to Creating Charts and Diagrams by Nigel Holmes, p.167. Reprinted by permission of Watson-Guptill Publications.

Notice how the scale affects your conclusions. In Figure 1.1a, the $800,000 increase (about 4%) can be hidden by extending the vertical scale to cover a range of $30,000,000 instead of about $1,000,000. In Figure 1.1b, by splitting the graph in half, putting a break in the vertical scale of the righthand part, and then using a different scale (each division equals 10 for D,E,F, and only 5 for A,B,C) it looks like A,B,and C are of the same size as D,E, and F. Such unscrupulous use of scales can give a bad name to statistics and graphics. However, such shenanigans make an important point: insofar as possible, the visual appearance of a graph should accurately reflect the numerical content. Clearly, the distortion in Figure 1.1b is caused by rescaling that forces the visual message (the heights of the bars from a common baseline) to contradict the data. For this reason, it is important always to carefully examine and understand the scaling when looking at a graph and to choose scaling carefully when making one.

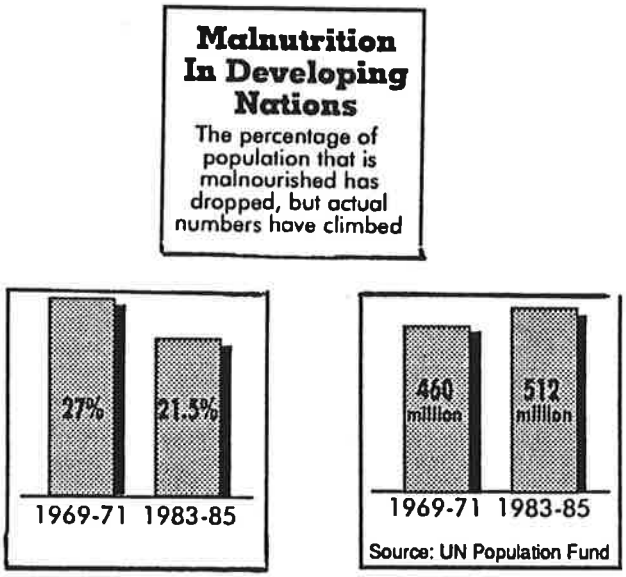
Figure 1.1a illustrates another problem that must be considered in choosing scales: the use of absolute value vs. percentage. Take a look at this pair of graphs again. The lefthand graph emphasizes that after several months of little absolute change, there has been a rapid absolute increase in payrolls. The righthand graph illustrates that on a relative (percentage) scale, the change hasn't been that great. Which is correct? -- It depends on the context, but the producer and consumers of this graph must be clear about what they wish to emphasize.

It is even possible for a particular measurement to go up in absolute value but down relative to the overall value. It may well be necessary to present graphs for <u>both</u> relative and absolute differences in this case in order not to mislead. For example, consider the following two graphs. One is the percentage change and the other the absolute change in malnutrition in developing nations. What the graphs show <u>together</u> is that while the relative amount of malnutrition declined between the two periods shown (27% to 21.5%), the actual number of malnourished people increased (from 460 to 512 million). Of course, the reason that this could happen is that the baseline, the total population, greatly increased between the two periods, so that there were many more people around. Portraying either graph alone would probably have been misleading; both are needed to tell the whole story.

In choosing scales for a graph of data, two principles should be generally followed. First, the range on the graph should include all of the data. Second, the data should fill in almost all the graph area. For example, if all data are in the range between 5 and 180, the scale on the graph should be such that 5 is near the left hand edge and 180 is near the right hand edge. The usual procedure is to find the two extremes of the data (say xmin and xmax) and to use a scale for the graph such that the largest and smallest numbers on the axis are "convenient" numbers which contain xmin and xmax. The space between the extremes is then marked off into equal intervals. For example, the axis might be partitioned into 8 intervals of 25 ranging from 0 to 200. The number of segments should not be too large (crowded) or small (sparse).

FIGURE 1.2

More Than One Graph May Be Necessary to Properly Convey the Information



"Malnutrition in Developing Nations" by David C. Walters, 5/15/90. Reprinted by permission from the Christian Science Monitor. Copyright © 1990 The Christian Science Publishing Society. All Rights Reserved.

In the case where one or a very few values is an outlier (i.e. is greatly separated from the rest of the data), it is often best to make the graph with that value excluded and separately indicated. This is in line with principle 2: including the outlier would leave too much white space in the graph.

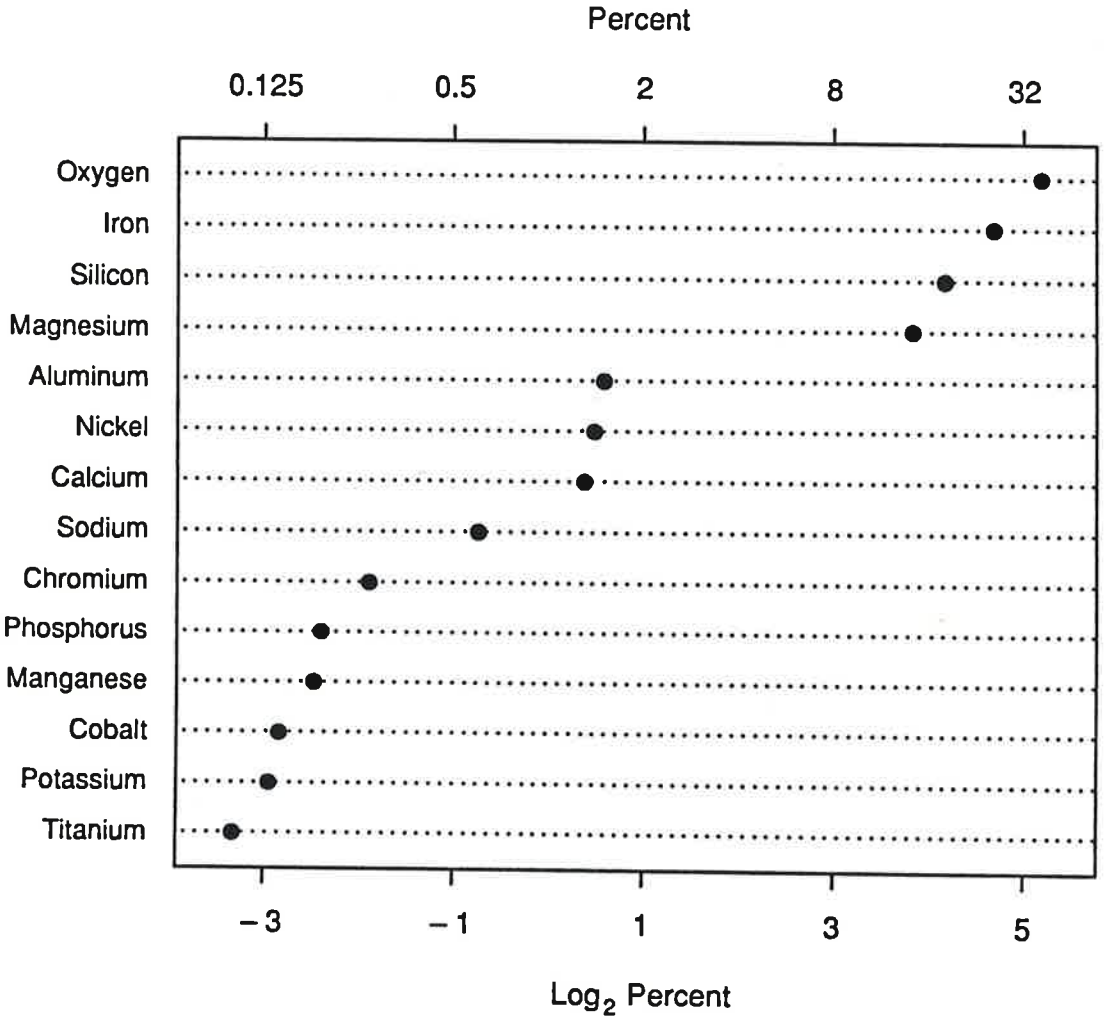
If the smallest value of a data set, xmin, is very far from zero, zero should be excluded from the scale in order to provide good resolution/detail (following principle 1). Some people think that excluding zero from the scale is a dishonest way of presenting data. However, it is almost always wiser not to sacrifice good resolution in order to include one specific value (zero) or some outliers. Of course, if points are excluded, this fact and their specific values should be clearly noted so that the viewer is made aware of this.

In practice, we often encounter data that are scattered unevenly over a very wide range of values but with a large proportion of the values near the lower end. Consequently, as x increases, the data become sparser and sparser. Here we have a definite pattern, not just a few unusual observations. Often, a way to improve such a graph is to "transform" the data. This is equivalent to using nonlinear scales.

If the x values are all positive, transforming the data by using the logarithm of the x value rather than the actual value can greatly improve the appearance of the graph. One should make sure that the viewer is aware that the graph represents transformed values, and of the type of transformation (e.g. logarithmic) used. Figure 1.3 is an example. It gives average percents of 14 common elements in stony meteorites. Note that each increase of 1 on the $\log_2$ scale represents a doubling of the percentage of iron.

FIGURE 1.3

Plotting the Data on a Log Scale Can Provide More Information



From The Elements of Graphing Data, by W.S. Cleveland, pp. 84-85. Reprinted by permission.

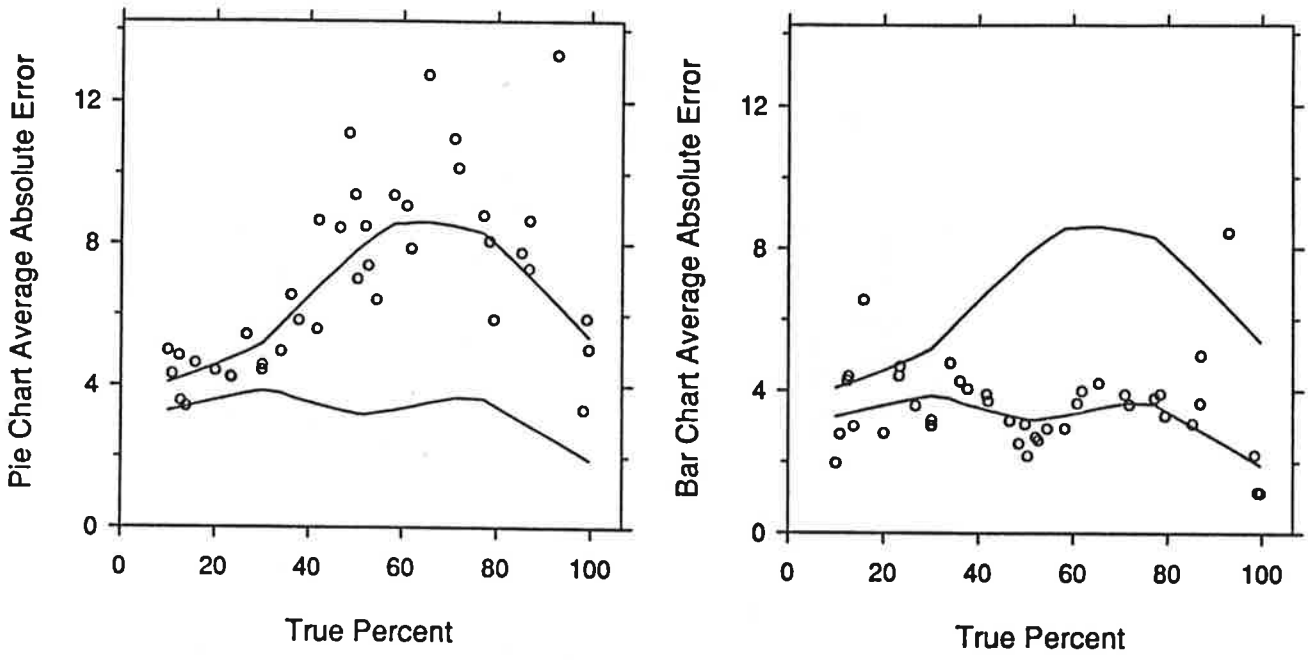
There are many other ways to transform scales in order to improve both appearance and comprehension of graphs. For example, even if the data are not all positive, one can first add a constant and then take logarithms. Other transformations (e.g., square root, reciprocal, "logit", rank, percentile) are all often useful, but beyond the scope of what we can discuss here. Readers may wish to consult some of the references for further ideas.

If it is desired to make comparisons between data on two separate graphs, then the scaling for the two graphs must be as alike as possible, even when this slightly violates principles (1) and (2). The graphs below are used to illustrate the results of a study to determine the difference in average absolute errors people made in interpreting pie vs. bar charts. At first glance, it looks like graph (b) did not make good use of the space: the largest value in the vertical axis is near 8 but the range for the y axis is 14. However, this is justified, because the purpose is comparing two graphs and the scales should be the same to facilitate the comparison.

First of all, a graph must be honest!

FIGURE 1.4

Extra White Space is OK When it is Needed to Maintain Scales for Comparisons



From The Elements of Graphing Data by W.S. Cleveland, p.205. Reprinted by permission

Simplicity of Graphs

Graphs should be simple and uncluttered. Elements that draw attention away from the main message of the graph should be eliminated. If no information or clarity is lost by removing some "ink", do not hesitate to remove it. Irrelevant decorative elements distract the viewer from the information the graph conveys and should be avoided.

Some people believe that statistical graphs must be "live", "communicatively dynamic", and heavily decorated and embellished. Perhaps this notion has stemmed from the fact that statistics --meaning tables of numbers -- are perceived to be dry and boring. Many corporate annual reports, government charts, and popular publications contain graphs that have been so profusely decorated that it is actually hard to see the information portrayed in the graph. In addition, sometimes the embellishments (intentionally or not) cause deception and/or distortion of the data. Here are a couple of examples.

FIGURE 1.5

Excessive Decoration can Distract from the Information

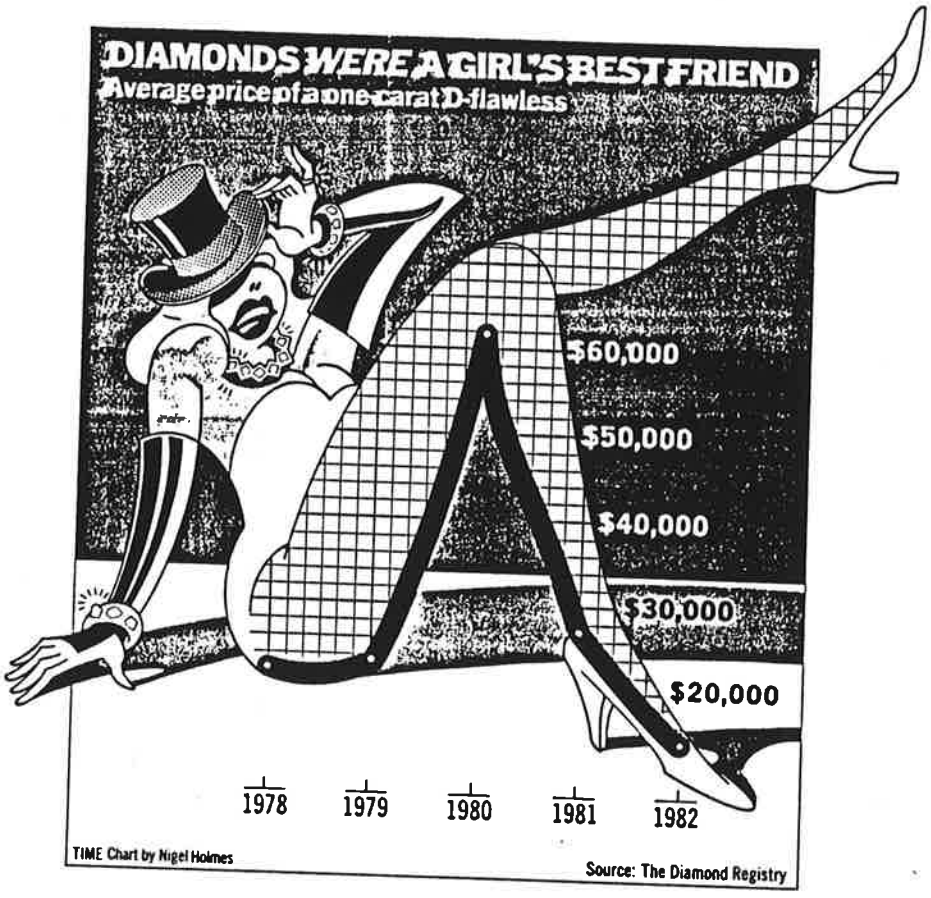
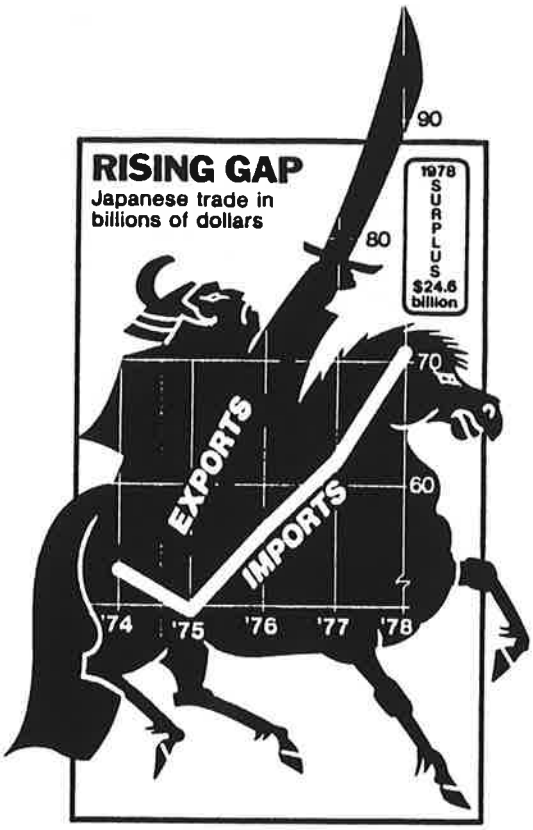


FIGURE 1.5 (cont.)

b)



From The Designer's Guide to Creating Charts and Diagrams by Nigel Holmes, pp.32 and 116. Reprinted by permission of Watson-Guptill Publications.

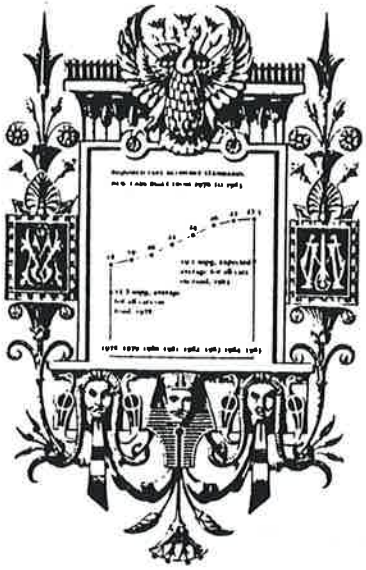
Good graphs demand the minimum possible effort of the viewer to interpret the information. The artistic effects in Figures 1.5a and b) add to the effort of understanding by forcing the reader to extract the graphic from the background visual clutter.

It is possible to decorate a graph without distorting the data, however. The graphs that follow are two examples. Note that in both cases the data part of the graphic is clearly separated from the decoration (in Figure 1.6a, by keeping the decoration outside the graph area; in 1.6b, by making the graph stand out in black from the gray envelope). This allows the viewer to focus on the information without difficulty.

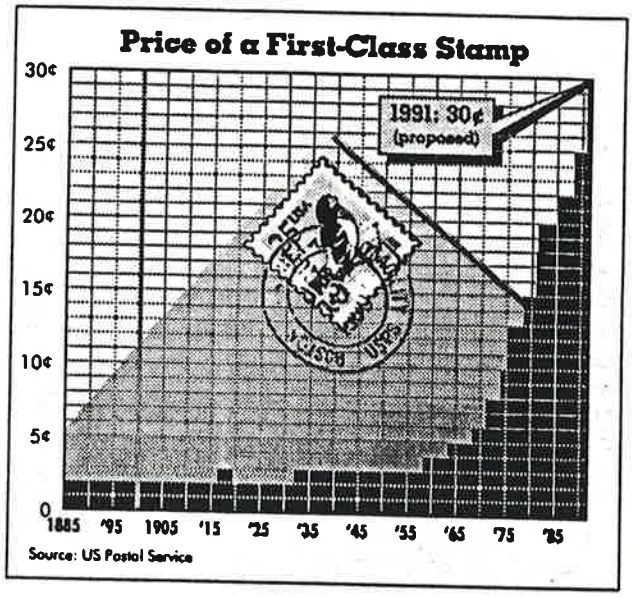
FIGURE 1.6

Decorating a Graph Without Obscuring the Information

a)



b)



1.6a from The Visual Display of Quantitative Information by E. Tufte. Reprinted by permission of the author.

1.6b by Elizabeth Ross from the 4/16/90 Christian Science Monitor. Reprinted by permission from The Christian Science Monitor Copyright © 1990 The Christian Science Publishing Society. All rights reserved.

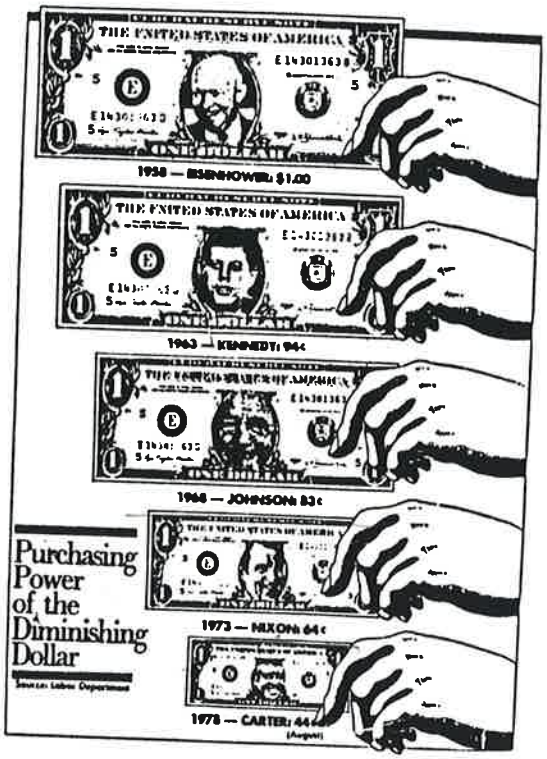
Adding unnecessary extra dimensions to a graph is one of the most common ways of distorting the information portrayed. Here are examples.

The graph in Figure 1.7a depicts the shrinking value of the dollar. However, the perception of the shrinkage far exceeds the actual shrinkage, because the dollar bill is shown in two dimensions, and each dimension is reduced. Howard Wainer, a keen and amusing chronicler of such graphical distortion (see his article, "How to Display Data Badly", listed in the references), has called this "squaring the eyeball to goose up the effect." A reduction by one half is perceived as a reduction by one quarter (½ in length x ½ in width). The distortion is even worse when a three dimensional picture is used to display a one-dimensional measurement, as is illustrated by the graph on 'The Shrinking Family Doctor' (Note: Both graphs appear in The Visual Display of Quantitative Information by E. Tufte).

FIGURE 1.7

Unnecessary Extra Dimensions Distort the Information

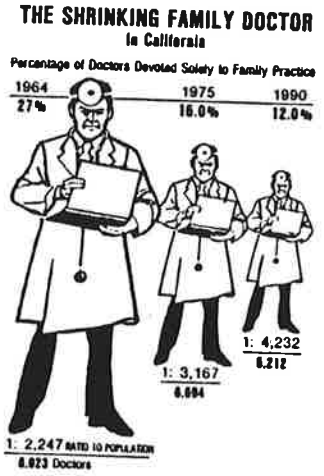
a)



From The Washington Post, 10/25/1978. Copyright © 1990 The Washington Post. Reprinted with Permission.

b)

FIGURE 1.7 (cont.)

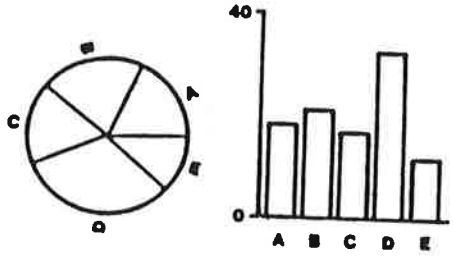


From the Los Angeles Times, 8/5/79. Copyright © , 1979, Los Angeles Times. Reprinted by permission.

In general, the simpler the visual processing task, the better the eye can grasp the message. One of the most interesting consequences of this observation is that the common pie chart should almost never be used. In figure 1.8, both the pie and bar charts give the same information, but most viewers can much more accurately make the comparisions in the bar chart (a linear comparison against a fixed baseline) than in the pie chart (an angular comparison with no fixed baseline).

FIGURE 1.8

Most of the Time, Don't Use Pie Charts



From "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods" by William S. Cleveland and Robert McGill, Journal of the American Statistical Association, 79, p.533.

The only advantage of the pie chart is that it makes clear that the individual pieces total 100%!

High Information Content

A good graphic can present an astounding amount of information in a very clear format. Population density maps, for example, are a kind of graphical display that visually represent tens of thousands -- or even hundreds of thousands -- of data points (population, latitude, and longitude at thousands of locations) on a single page. Indeed, there is probably no better way to deal with large amounts of information than with a good graphic.

High information content is therefore one of the hallmarks of good graphical display. Consequently, most of the ink in a graphic should be devoted to data-information, not to embellishment. One way to measure the information content of a graph is to measure what proportion of the consumed ink in a graph is used for data, data-ink. Data-ink is the portion of the ink that can't be erased without removing information. The ratio of data-ink to total ink used to print the graph is called the Data-ink ratio:

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{total ink to print the graph}}$$

It is clear that as this ratio tends toward 1 the graph has less embellishment. If the Data-ink ratio is 1, it means 100% of the ink is used to present data-information and nothing can be erased without losing some information. This ratio -- originally invented by Edward Tufte (see The Visual Display of Quantitative Information in the references) -- thus serves as a useful criterion for determining the information content of a graph.

Other criteria, such as Data Density Index ("the number of numbers plotted per square inch"), can also be useful in finding how well the allocated space in a graph is used. However, it is not so important to score high on any particular index as it is to remember that one of the great strengths of statistical graphics is the ability to convey large amounts of information that would be indigestible any other way. When there are only a few data items, a simple table -- or even some words -- may be sufficient. A graph would be overkill! But when large and complex sets of data must be understood and the information they contain communicated, there is no better way to do it than with a graph.

Some Good Graphs

We present here several famous examples of statistical graphs at their best. These examples are sometimes more complicated than the ideas presented in the later chapters of this Guide, but they illustrate what can be done with imagination and creativity. All of them present a lot of complex information clearly. The graphs are attractive and memorable (it is said that Frenchmen

wept when they saw Minard's graph of Napoleon's retreat), and they represent both the old and new, spanning over 100 years of excellence in graphical display.

1. The following graph gives a devastating history lesson: it tells the story of Napoleon's ill-fated invasion of Russia in 1812 - 1813. The width of the shaded gray path portrays the size of Napoleon's army at different places along his invasion route. He started with 422,000 men near the Niemen River (Polish - Russian border), but lost thousands as he progressed into Russia, reaching Moscow with only 100,000 men in September.

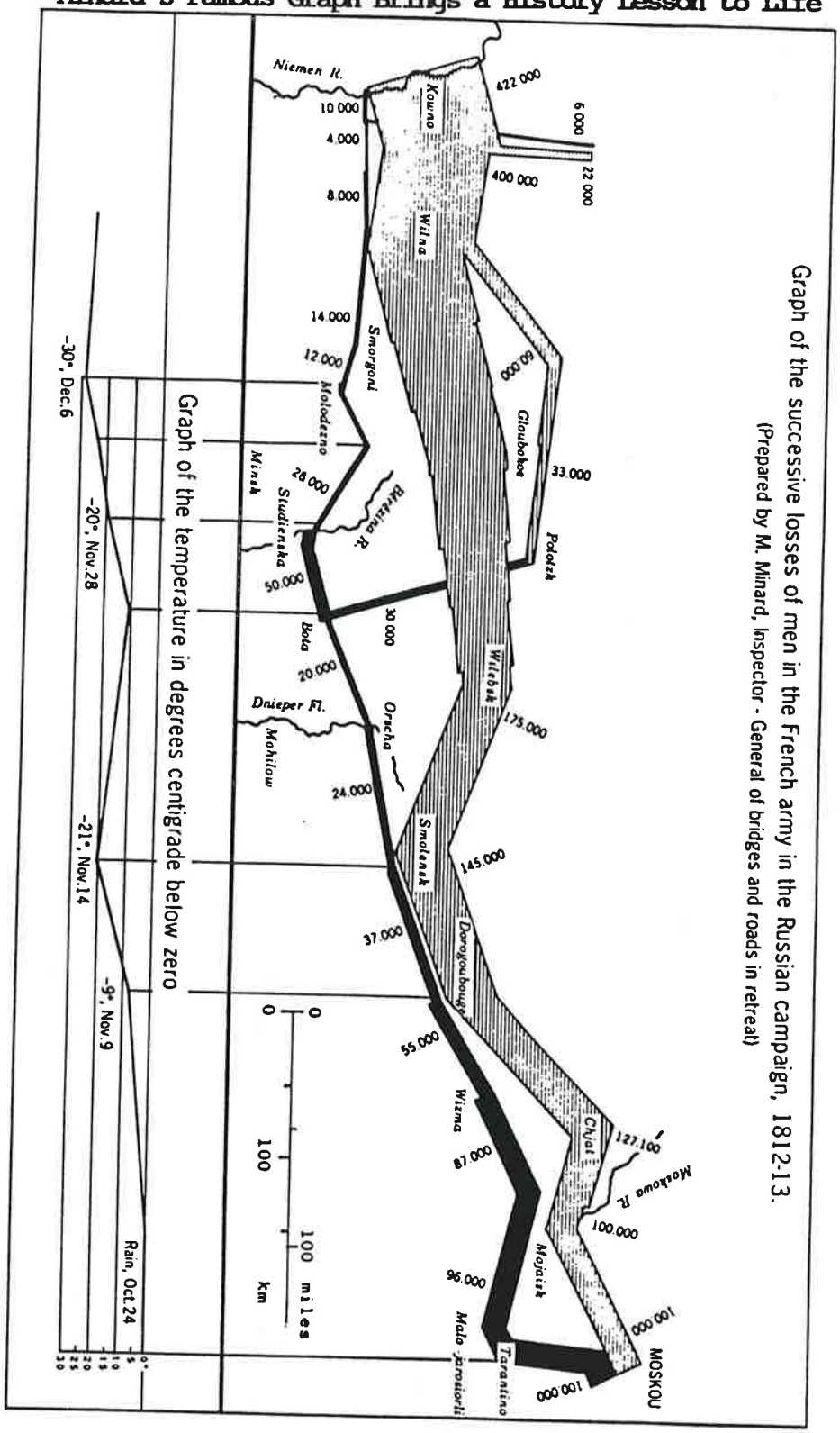
The dark lower band shows the army's return from Moscow and is linked to the temperature on the bottom of the graph. Because it was a bitterly cold winter, Napoleon lost many of his men due to the weather. As the graph shows, by the time the troops reach Orscha, the army has been reduced to 20,000. At Bota 30,000 men joined them. (60,000 men stayed behind in Gloubokoe, and the 30,000 troops that remained of these rejoined Napoleon when he reached Bota.) The army lost 22,000 men in a disastrous crossing of the Berezina River. By the time they reach the Polish - Russian border, only 10,000 men were left.

Note that data on 6 variables are presented with this graphic: size of the army, location on a map (2 variables), direction of movement, date, and temperature. The data-ink ratio is 1. Not a single thing can be erased without losing information. Few who have seen this graph could forget this history lesson!

Note: This Graph was originally drawn in 1861 by the French engineer, Charles Joseph Minard. It has been reproduced many times and in many places. Our version is taken from p.60 of Introductory Statistics for Business and Economics by Thomas H. and Ronald J. Wonnacott, 4th edition.

FIGURE 1.9

Minard's Famous Graph Brings a History Lesson to Life



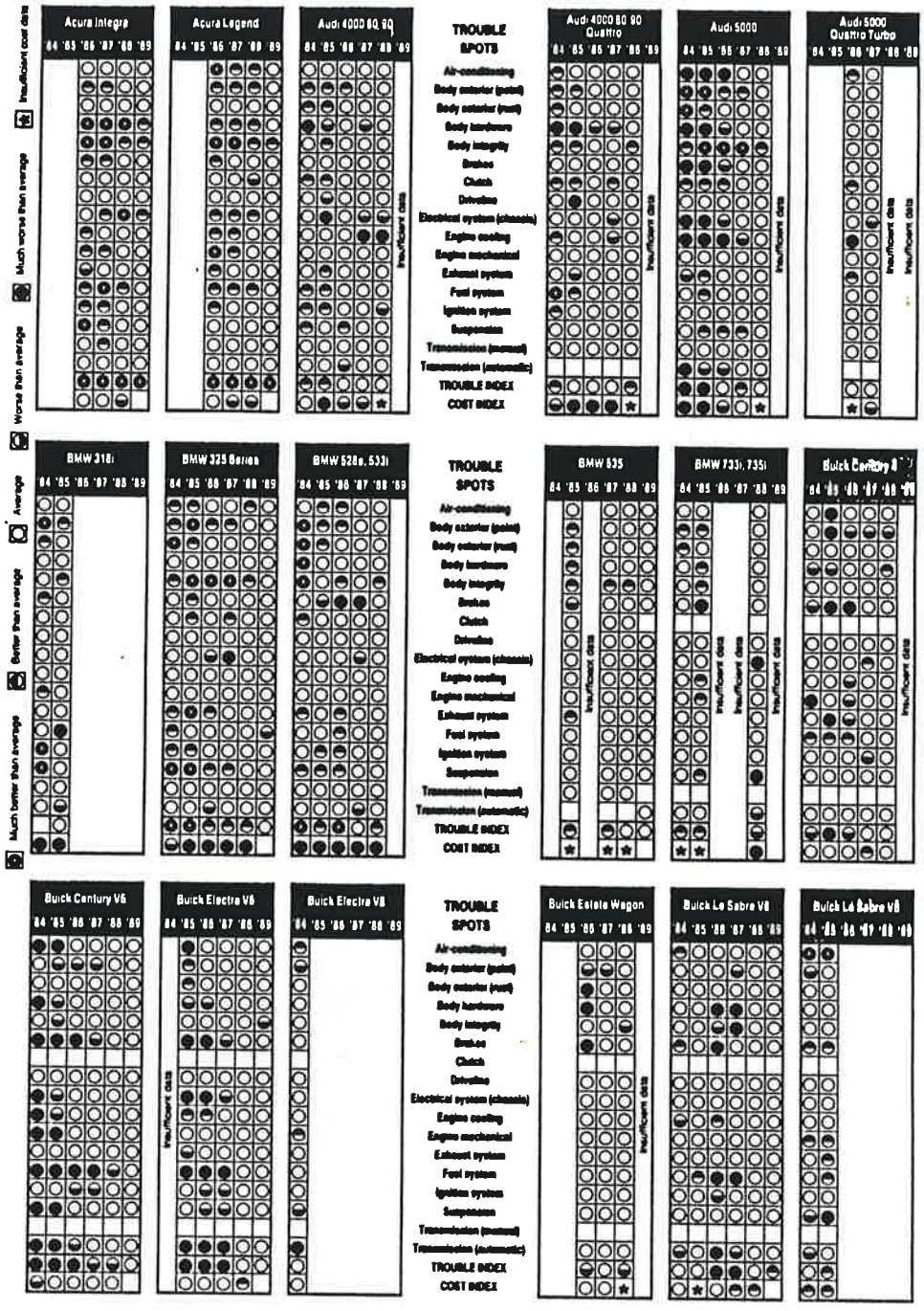
2. Consumer Reports magazine uses an excellent graphical approach to depict auto frequency-of-repair information on 19 variables (17 different systems and 2 indices) over 5-year periods on various models of cars and trucks. The graph on the next page is self-explanatory. In the magazine, color coding allows readers to quickly scan the page to pick out vehicles that are:

-- much better than average (red filled circle with small white center)
-- better than average (red on the upper half of the circle only)
-- average (open circle)
-- worse than average (black on the lower half of the circle only)
-- much worse than average (all black circle)

Note how the clever use of the white center in the red circle and filled upper and lower regions for "above" and "below" average,respectively, allows a black and white version to convey the same information, although not quite as easily as the color version.

FIGURE 1.10

Consumer Reports Uses Clever Choice of Symbols and a Clean Format to Help Readers Quickly Absorb Frequency-of-Repair Information



3. The two "Population Pyramids" that follow compare the United States and Brazilian population in 1985. They are clever examples of back-to-back histograms (to be discussed in the next chapter). They provide an enormous amount of information about each population. For example, we may observe the following:

1. Each small box represents one million persons. By adding them, one can obtain the male (or female) population of each country for various age groups.

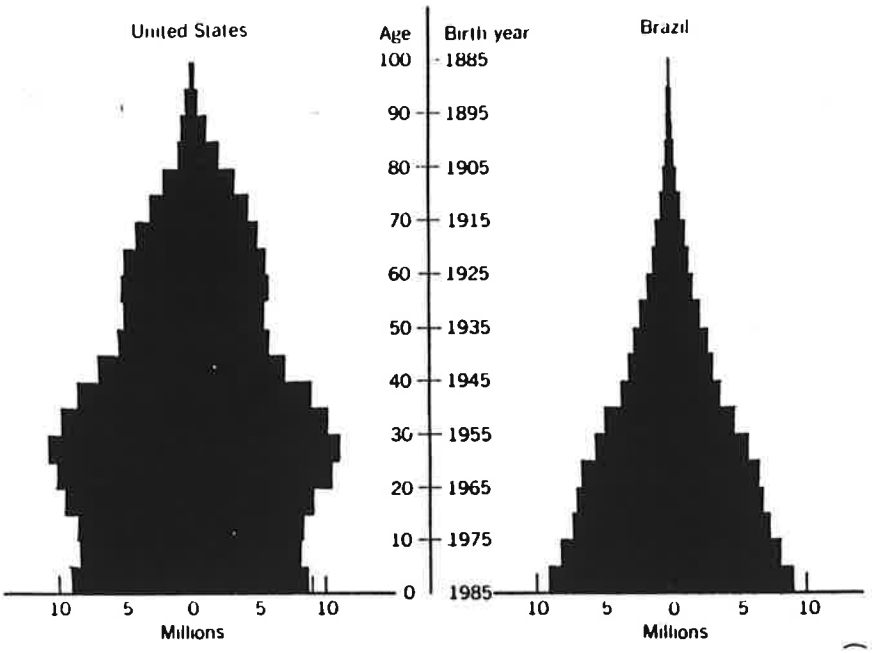
2. Life expectancy in the U.S. is longer then in Brazil.

3. Females live longer than males.

4. The U.S. shows a "developed nation" population profile: the present birthrate is lower than it has been in the past, infant and child mortality is low, and therefore the profile has a large "waist". Brazil, with its high birthrate and high mortality rate shows the typical undeveloped nation's pyramid.

FIGURE 1.11

Population Pyramids Show the "Big Picture"



From Introductory Statistics for Business and Economics by Thomas H. and Ronald J. Wonnacott, 4th edition. Copyright © 1986 John Wiley & Sons. Reprinted by permission.

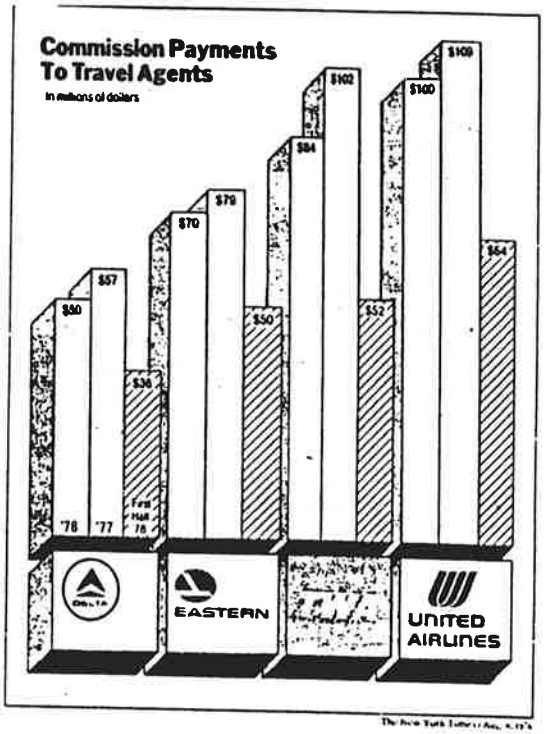
One other popular example of graphical excellence ought to be mentioned--the weather map in your daily newspaper (of which USA Today's is probably the best). Take a look at it. Depending on the details, it may contain information on moisture, temperature, wind speed, barometric pressure, and direction of movement and type of weather fronts for thousands of locations throughout the country. Though we take it for granted, this statistical graphic summarizes thousands of separate weather observations and is a superb example of the effectiveness of statistical graphics.

Some Bad Graphs

Unfortunately, the previous examples tend to be the exception rather than the rule these days. The development of computer graphics tools and the ease with which data can be gathered and processed has led to an explosion of graphical display. Sadly, many of these displays are designed more to demonstrate the cleverness of the artist than to present the information in the data. As we have stated, the purpose of a statistical graphic is first of all to inform. The following examples show how a failure to understand and follow this principle can lead to disastrous consequences.

FIGURE 1.12

Commission Payments Down in 1990? — Check the Fine Print



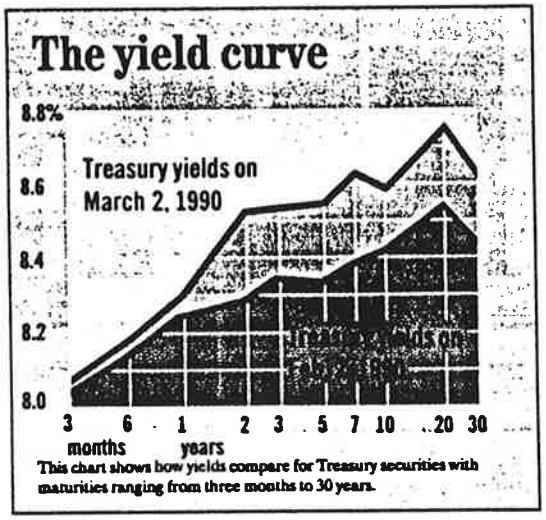
From the New York Times, 8/8/78. Copyright © 1978 by The New York Times Company. Reprinted by permission.

1. Figure 1.12 seems to portray a decline in the commission payments to travel agents in 1978 for four big airlines: Delta, Eastern, TWA and United. Only by examining the fine print in the '78 column for Delta, can the reader (barely) make out that the '78 results are only for a half year. Doubling them to provide a crude estimate of the whole year's results would show 1978 continuing the upward movement from '76 to '77. This is dishonest graphics at its crudest.

2. The following graph distorts the relationship between yield and time to maturity of U.S. Treasury securities. X-axis distance between 3 and 6 month bond maturities is actually larger than the distance between that of 6 month and 1 year maturities; the distortion is even worse for longer maturing securities. As a result, the yield curve shown gives a completely misleading impression of how the rate varies with time to maturity. In fact, the rate is almost constant, not rapidly increasing (the so-called "inversion" of the yield curve, an important phenomenon).

FIGURE 1.13

The Inconsistent Scale on the X-Axis Distorts the Information



From Money Magazine, April 1990. Reprinted by permission.

3. The last graphic appeared in an article syndicated by a national news service. It violates nearly every principle of good graphical display.

a. The artistic embellishment almost completely conceals the data . The purpose is clearly to show off the skills of the artist, not the information.

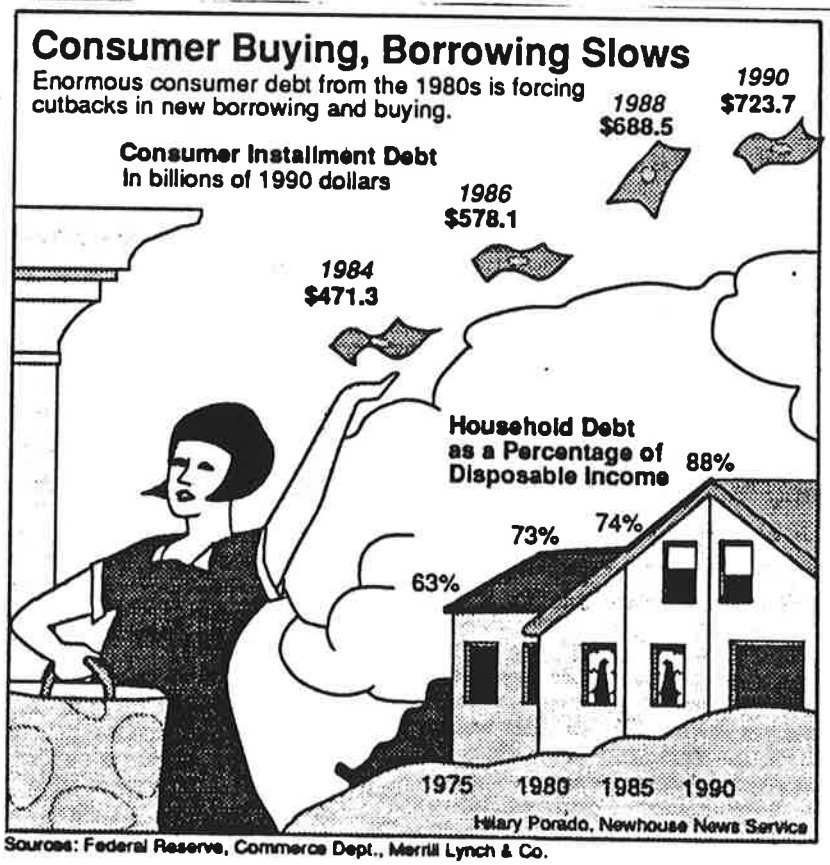
b. The data-ink ratio is almost zero -- almost all the ink is non-data ink.

c. So little information is presented that a simple table would have sufficed and been much clearer.

d. There are two different scales (with different scalings!) on the x-axis so that the relationship between consumer installment and household debt over time cannot be determined. Moreover, the consumer installment debt is reported in 1990 dollars, but the household debt is reported as a percentage of disposable income. Both should probably have been reported as percentages.

FIGURE 1.14

An Incoherent Graphic



From the Newhouse News Service. Reprinted by permission.

Summary:

What Makes a Good Graph Good and a Bad Graph Bad?

In this chapter, we have presented some of the principles of good statistical graphics and provided illustrations of both their use and abuse. Needless to say, there is a lot more to be said on these matters, and, indeed, research in these areas is ongoing. The references, particularly the works by Tufte and Cleveland, are wonderful resouces, and we have borrowed from them freely. The next two chapters in this Guide also provide further ideas and illustrations.

We have attempted to show here that graphical excellence is much more than a matter of applying a few principles and techniques. Rather, graphical excellence is as much an attitude as it is a set of techniques -- a commitment to presenting information in as clear, clean, and honest fashion as possible, to give the facts and permit the viewer to draw whatever conclusions can be drawn from them. The specific principles that we have discussed follow from this attitude. If, when producing a statistical display, you stop to ask yourself whether the information is clearly and honestly presented and make sure that it is, it will be hard to go wrong.

CHAPTER 2

SIMPLE GRAPHICAL DISPLAYS FOR LOOKING AT BUNCHES OF DATA

by Carol Joyce Blumberg
Winona State University

It is nearly impossible to look at a list or table of numbers -- raw data -- and make much sense of them. Consider Table 2.1, for example, which gives heights and weights of some college students (these data were taken from the 1985 Minitab Handbook by Ryan, Joiner, & Ryan, 1985). What is a typical "average" weight of the males? How spread out are the female weights compared to the male weights? Unless you are truly remarkable, it will be difficult just to stare at the data and answer these questions.

TABLE 2.1

Heights and Weights of 92 College Students Classified by Gender and Ordered by Height

Males		Females	
Height	Weight	Height	Weight
66.0	140	61.00	140
66.0	135	61.75	108
66.0	135	62.00	131
66.0	130	62.00	120
67.0	145	62.00	108
67.0	150	62.00	110
67.0	140	62.75	112
67.0	123	63.00	121
68.0	155	63.00	118
68.0	150	63.00	116
68.0	145	63.00	95
68.0	155	64.00	102
69.0	155	64.00	125
69.0	175	65.00	135
69.0	170	65.00	118
69.0	145	65.00	122
69.0	160	65.00	115
69.0	150	65.50	120
69.0	136	66.00	120
69.5	150	66.00	130
70.0	153	66.00	130
70.0	157	66.00	125
70.0	130	67.00	125
70.0	155	67.00	115

70.0	150	67.00	150
71.0	138	68.00	130
71.0	170	68.00	138
71.0	170	68.00	116
71.0	155	68.00	125
71.0	150	68.00	110
71.0	140	68.00	133
71.5	164	69.00	145
72.0	145	69.00	150
72.0	150	69.00	150
72.0	195	70.00	125
72.0	155		
72.0	175		
72.0	215		
72.0	180		
72.0	142		
73.0	190		
73.0	165		
73.0	170		
73.0	155		
73.0	155		
73.0	180		
73.0	155		
73.5	160		
73.5	155		
74.0	190		
74.0	160		
74.0	180		
74.0	190		
74.0	148		
75.0	185		
75.0	160		
75.0	190		

What we need are some simple ways to picture how the data are spread out--or distributed -- so we can easily see the answers to these and other similar questions. That is the subject of this chapter.

Dotplots

A dotplot (sometimes called a lineplot or point plot instead) is probably the simplest way to picture a bunch of numbers. It consists of the relevant portion of a number line on which each data value is indicated with a mark, typically a dot (hence the name dotplots). For example, here are the lengths in miles of the world's largest (by area) lakes, ordered by length:

TABLE 2.2

Lengths (in miles) of the World's Largest Lakes

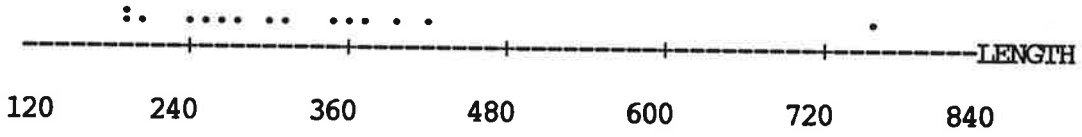
Name	Continent	Length in miles
Caspian Sea	Asia-Europe	760
Tanganyika	Africa	420
Baykal	Asia	395
Balkhash	Asia	376
Malawi	Africa	360
Superior	North America	350
Michigan	North America	307
Great Slave	North America	298
Aral Sea	Asia	280
Winnipeg	North America	266
Victoria	Africa	250
Erie	North America	241
Huron	North America	206
Ontario	North America	193
Great Bear	North America	192

(Source: The World Almanac and Book of Facts 1986)

To make a dotplot for these data, we let the scale on the number line range from a bit less than 192 (the minimum) to a bit more than 760 (the maximum) and for each length put a dot in the correct place above the number.

FIGURE 2.1a:

Dotplot of Lengths of the World's Largest Lakes



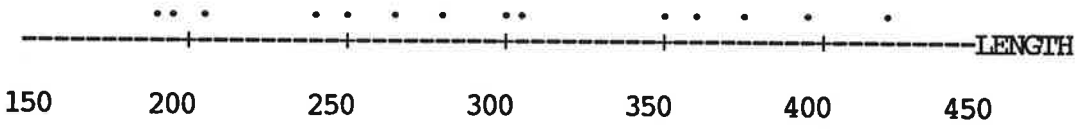
Ideally, a dotplot should have all the dots spaced out in one horizontal line, except, of course, when more than one data point has the same value. But, this is often not possible: the printing device may not be able to put the dots in exactly the right place (this was the situation above) or the scale may not have enough detail. In either case, some piling up of the dots may be necessary, as in the left hand side above.

An observation that is very big or very small compared to the rest of the data is called an outlier. For the above data, the length of the Caspian Sea is an outlier. Later in this chapter we will discuss outliers in more detail.

The dotplot below gives a dotplot of the lengths of the lakes with the Caspian Sea deleted from the data set. You can now see more clearly how the lengths of the rest of the lakes are related.

FIGURE 2.1b:

Dotplot of Lengths of the World's Largest Lakes with the Caspian Sea (length = 760 miles) Deleted



Looking carefully at the display, there appear to be three separate groups of lakes: three lakes around 200 miles in length, six lakes from 241 to 307 miles, and five lakes that are 350 miles or longer. No further patterns appear. It is often difficult to see overall patterns in dotplots because there is so much detail. Although this can sometimes be as advantage, often less detailed plots, such as histograms, stem-and-leaf displays, and boxplots allow us to see data patterns better. The remainder of this chapter will describe how to construct these other displays.

Histograms

A histogram is a bar graph where the frequencies (i.e., number of occurrences) of data values in certain intervals are represented by the lengths of the bars. In other words, a histogram puts the data into groups with the length of the bar for each group proportional to the amount of data in that group.

The first step in constructing a histogram is to decide the number and width of the intervals to be used. The intervals should be of equal widths whenever possible. For any particular data set there are many different ways of deciding the numbers and lengths of the intervals. No one way is "correct" -- in fact, looking at the data in more than one way is often desirable.

For example, consider the heights of the males and females combined in Table 2.1. For these data, either five intervals each with a length of three, or eight intervals each with a length of two could be used:

Length 3 Intervals		Length 2 Intervals	
Interval	Frequency	Interval	Frequency
		61.00-62.99	7
61.00-63.99	11	63.00-64.99	6
64.00-66.99	15	65.00-66.99	13
67.00-69.99	28	67.00-68.99	17
70.00-72.99	21	69.00-70.99	17
73.00-75.99	17	71.00-72.99	15
		73.00-74.99	14
		75.00-76.99	3

The second step is to count the number of data values (i.e., the frequency) for each interval.

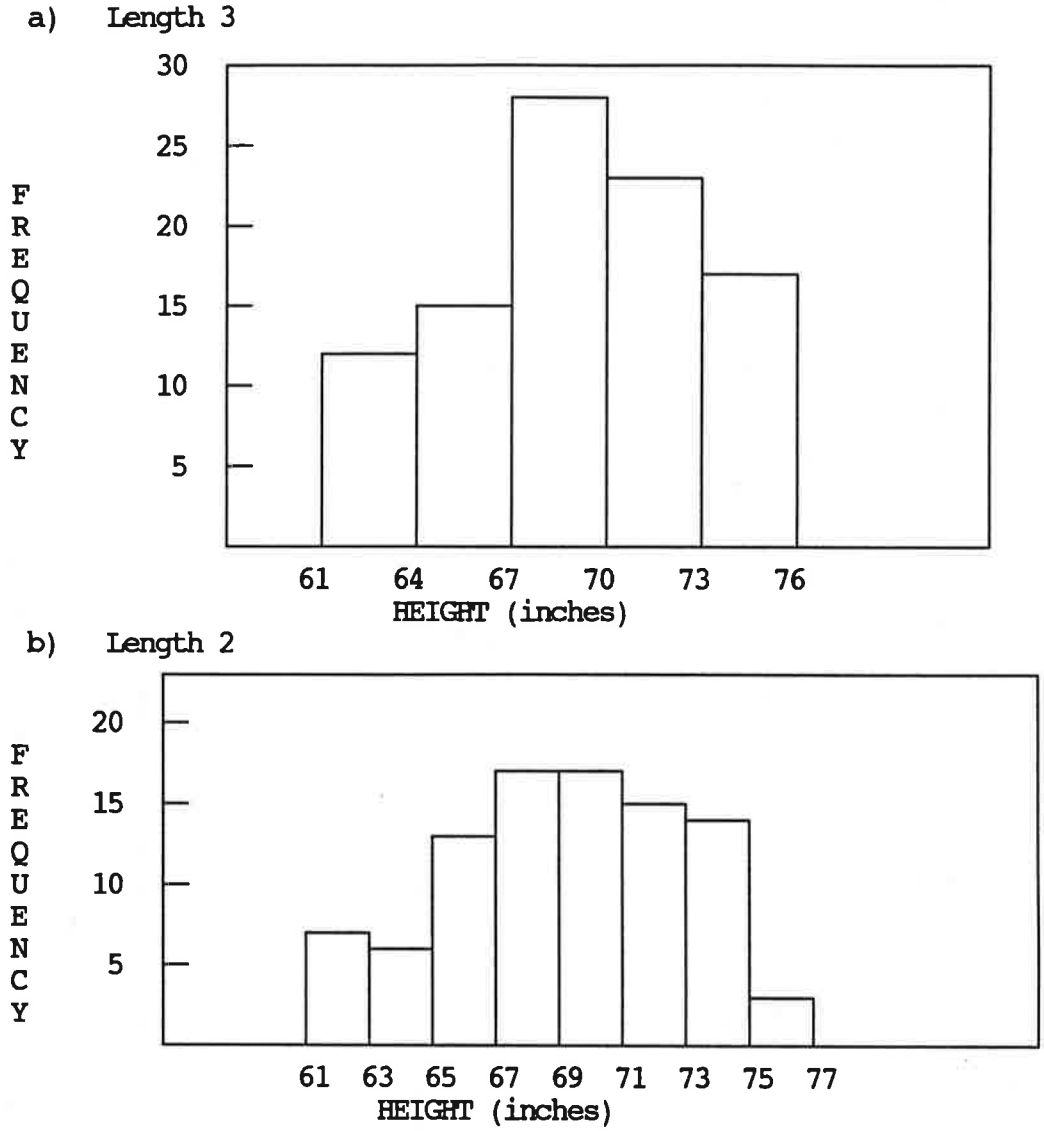
The next step is to draw the bars for each interval with length proportional to the frequency. If the bars are to be vertical then the vertical axis indicates the frequencies and the horizontal axis indicates the intervals. Start the horizontal axis at or near the lowest data value and choose the scaling so that the data fill up nearly the whole axis. Do not leave a lot of excess white space. Finally, it is important to clearly label the axes and give a descriptive title so that the viewer will know exactly what is plotted.

Some people, and many computer packages, prefer to make the bars horizontal (that is, sideways). The intervals are then along the vertical axis and the frequencies along the horizontal axis.

Figures 2.2a and 2.2b are the completed histograms (using vertical bars) for the two different interval widths given above. Which of these is the better display? Answer: Some people will say Figure 2.2a and some people will say Figure 2.2b. Neither is right nor wrong. Nevertheless, it is important to realize that the choice of intervals can make a difference in interpreting the data. Therefore, it is important to take care in making this choice, so that the histogram tells an accurate story.

FIGURE 2.2

Which Interval Length Does Better for the Student Height Data?



If the intervals chosen do not seem to portray the data well, (e.g., the histogram is too spread out or too compact) try a different set of intervals. The number of intervals is determined by the nature of the data and by the total number of observations in the data set. An often-used rule of thumb is to have about $\sqrt{n}$ intervals, where n is the number of data points. But, this is <u>not</u> a hard and fast rule. It is a good idea to experiment with different choices to see what seems to give the best picture of the data.

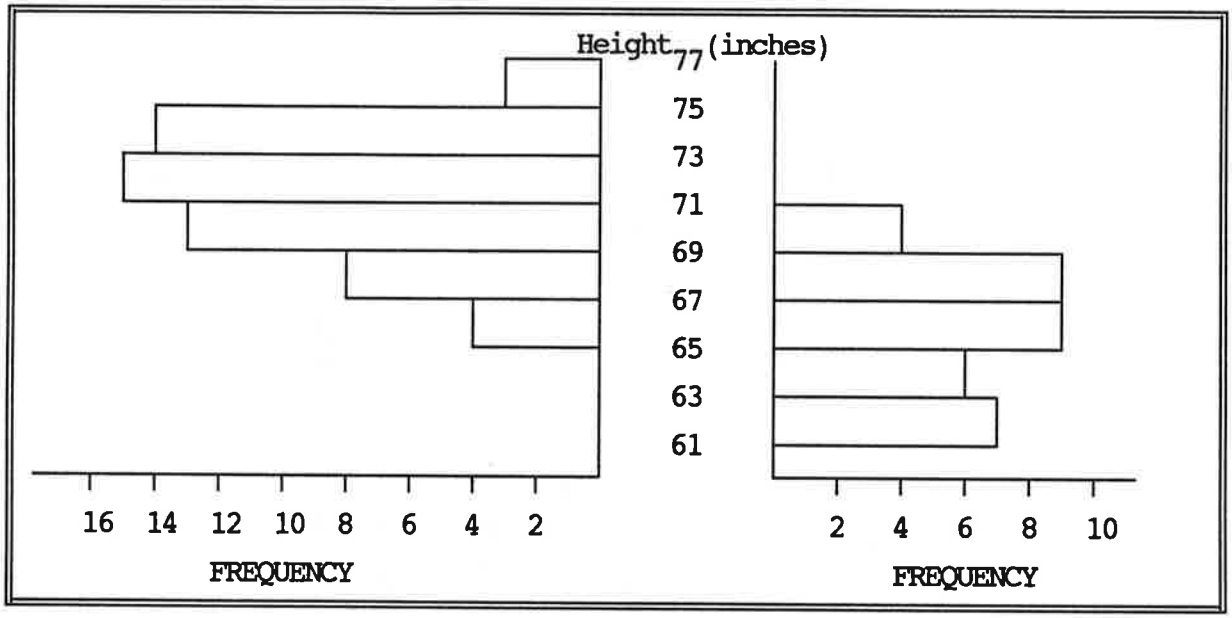
Sometimes, such as with the height and weight data, the individuals on which the data are being collected form two subgroups. You can then form a

histogram using horizontal bars for each group and put them back-to-back. This is called a back-to-back histogram. For the data in Table 2.1, the two obvious subgroups are the males and the females. Figure 2.3 shows a back-to-back histogram using horizontal bars for the heights for males and females. This gives a more complete picture of the data since the males and females have different patterns.

See also Figure 1.11 in Chapter 1 for an excellent example of this technique.

FIGURE 2.3

A Back-to-Back Histogram of the Male and Female Student Heights



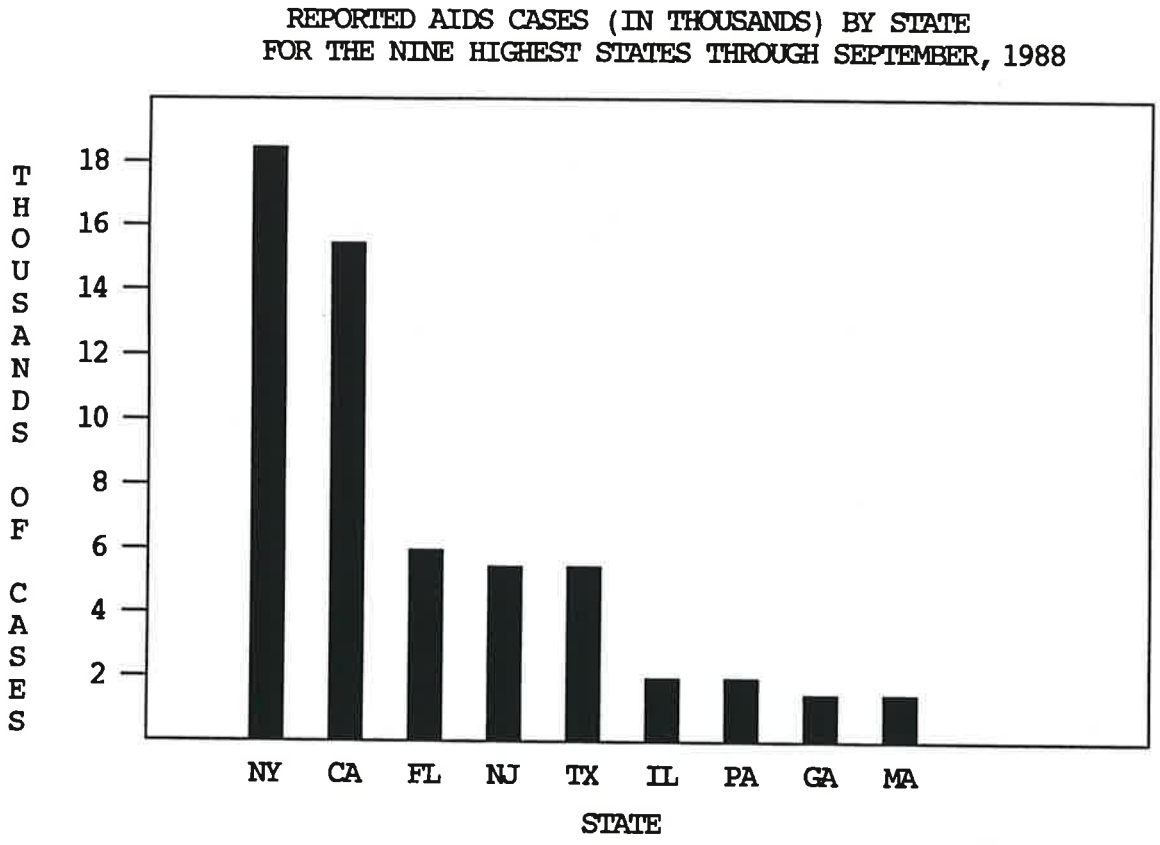
There are also occasions when the best way to describe the data is to have each interval represent only a single data value. In fact, this often occurs when year (e.g., 1987, 1988, 1989, etc.) is the grouping variable. For example, this would be the case when naking a histogram of the number of babies who were born each year in a city over a ten-year period.

Histograms can also be made for categorical data (sometimes referred to as attribute data or qualitative data) such as eye color, favorite music group, preferred brand of soda pop, or last digit of telephone numbers. Categorical data are data where numbers or labels are used to sort people or objects into groups or categories. If one wants vertical bars on the histogram then the vertical axis will indicate the frequencies and the horizontal axis will list the categories used. Of course, for horizontal histograms, the labelling is reversed. It is usually wise to put the categories in some logical order. Many times the best logical order is by decreasing frequency. In this case, the display is often called a Pareto chart. Figure 2.4 is a histogram of the number of AIDS cases reported by state for nine states through 1988. Note how

clearly ordering by frequency rather than, say, alphabetically, shows the three distinct groups of states.

FIGURE 2.4

A Pareto Chart is a Histogram Ordered by Decreasing Frequencies



An often used variation on the histogram is to use relative frequencies instead of actual frequencies. The relative frequency of an interval or category is defined as the number of observations in that interval or category divided by the total number of observations in the data set. These relative frequencies can be expressed as fractions, decimals, or percents, but percents seem to be the usual choice. For example, for the height data using length 2 intervals, the relative frequency for the first interval (61.00 to 61.99) would be $7/92 = .0761 = 7.61\%$. The histogram using the relative frequencies would be identical to Figure 2.2b in shape; the only difference would be that the vertical axis would be labelled by the relative frequencies instead of the frequencies.

Outliers

An outlier is any value in a data set that appears to be separated (either on the low side or high side) from the main body of the data. For example, for the weight data in Table 2.1, the male who weighs 215 pounds is an outlier. For the lake length data, the Caspian Sea is an outlier. There are many different methods that statisticians use to determine which data values, if any, are outliers. Discussion of these methods is beyond the scope of this Guide. However, there are several ideas concerning outliers that anyone analyzing data (graphically or otherwise) should consider.

First, complicated methods are often unnecessary to detect outliers: they often stick out like a sore thumb. For example, the 215 pound male is 20 pounds heavier than the next nearest person; the Caspian Sea is almost twice as long as the next longest lake. So, even if you don't use technical statistical procedures, you should always check for obvious outliers.

Second, outliers are often caused by errors made somewhere during the data collection, recording, or analysis stages. So always check data carefully for correctness at every stage. This goes even (some would say, especially) for data collected directly by computer.

Third, outliers can influence people's perceptions of graphical displays. We tend to focus too much on the outliers and not enough on the rest of the graphic. Hence, in constructing the graphical displays discussed in this and other chapters, try to make sure that the outliers do not dominate the graphic. As we saw in Chapter 1 and in Figure 2.1, one good way of avoiding this problem is to list the outliers separately and scale the display to the rest of the data. Comparing Figure 2.1a and 2.1b, it is easy to see how the outlier distorts the scale and therefore hides detail in the rest of the data.

Of course, sometimes outliers are not mistakes or nuisances, but potential great discoveries. This is another good reason for looking at them very carefully and not mixing them up with the rest of the data.

Stem-and-Leaf Displays

Stem-and-leaf displays are an improvement on the histogram. They were described in John Tukey's (1977) book and subsequently popularized (see, e.g., Velleman & Hoaglin, 1981). Stem-and-leaf displays are similar to histograms in that they provide a picture of the shape of the distribution of the data. But, in addition, they allow people to see more details about the data, because more digits can be retained in the stem-and-leaf display than in a histogram.

When constructing a stem-and-leaf display, each of the values for the variable to be displayed (e.g., the weights in Table 2.1) is divided into two parts called the stem (the main part) and the leaf (the secondary part). The best way to illustrate this division is by example. For the weight variable,

the hundreds' and the tens' digits become the stems and the ones' digit becomes the leaves.

The first step in constructing a stem-and-leaf display is to draw a vertical line. The stems (e.g., 11, 12, 13, etc. for the weight data) go on the left hand side and the leaves go on the right hand side of the line. It is very important to include each possible stem between the lowest and the highest values of the data, even if there are no data values (leaves) for that stem. This is because the stems act like the intervals on a histogram and omitting some of them would give a distorted picture of the data.

The next step is to write down each stem once (on the left) and each leaf the number of times it occurs in the data (on the right).

The last step is to put an appropriate legend near the display. A legend is a note that tells the viewer where the decimal point belongs and includes information on how many significant digits there are in the data. Make sure the legend is clearly visible but does not get in the way of the data.

As a first example, the stem-and-leaf display for the weights of the first 15 male students is:

```
12|3
13|055          LEGEND: 13|5 = 135
14|0055
15|00555
16|
17|05
```

Note that this is nothing more than a sideways histogram in which the length of the bars is given by the piling up of the data digits in the leaves. For this reason, it is important to make the leaves (the digits) uniform in width.

When making a stem-and-leaf, you should first write down the leaves in the order they appear in the data. DO NOT try to skip around and get them in order on the stems -- you'll make too many mistakes and it will take too long. After finishing this preliminary stem and leaf display, it is then almost always useful to copy the stem-and-leaf display over, this time putting the leaves in numerical order within each stem. This will be quick and accurate, and you'll then have neat ordered displays like those that appear here.

The reason for the blank space to the right of the 16 is that none of the first fifteen students had a weight between 160 and 169. The complete stem-and-leaf display for all 92 students and a back-to-back stem-and-leaf display with males on the left hand side and females on the right hand side are given in Figure 2.5.

FIGURE 2.5

Stem-and-Leaf Displays of Student Weights

a) All Students Combined

```
 9|5                          LEGEND:  12|3 = 123
10|288
11|002556688
12|00012355555
13|0000013555688
14|00002555558
15|000000000355555555557
16|000045
17|000055
18|0005
19|00005
20|
21|5
```

b) Males and Females Separately

```
          MALES          FEMALES
                      9|5              LEGEND:  12|3 = 123
                     10|288
                     11|002556688
                  3  12|0001255555
             865500  13|0001358
          855552000  14|05
7555555555530000000  15|000
             540000  16|
             550000  17|
               5000  18|
              50000  19|
                     20|
                  5  21|
```

Notice that the shapes of these displays are essentially the same as would be given by a histogram. However the stem-and-leaf displays offers several important advantages over histograms:

1. The original data can often be recovered from the stem-and-leaf display; with the histogram, they usually cannot.

2. Stem-and-leaf displays can show interesting patterns that histograms miss. For example, the lowest weight is exactly 95 lbs.; there are five people who weighed exactly 130 lbs., of whom two are male and three are female; most

interesting of all, almost all weights ended in a 0 or a 5 (only 8 out of 57 males and 13 out of 35 females did not). Can you think of a plausible explanation for this "strange" occurrence?

3. The stem-and-leaf display is a quick and easy way to order the data. This turns out to be useful for further analysis (e.g. with box-and-whisker plots--see the next section).

For the height data in Table 2.1, the construction of the stem-and-leaf display is a bit more complicated. First, it must be decided which digits are to form the stems. That is, should the stems be only the tens' digit or should the stems be both the tens' and ones' digits? The leaf is then the next digit to the right of the stem. When doing these displays by hand, all other digits to the right are ignored. The technical term for ignoring digits is truncation. We prefer to truncate, rather than round, because it is quicker and helps us avoid making the errors that often occur when rounding is used.

Let's look at several alternative ways of making a stem-and-leaf of the heights of the 35 female college students. Our first attempt, Figure 2.6a, uses the tens' digit as the stem and the ones' digit as the leaf. This display doesn't appear to be very helpful. To improve it, different intervals can be used. That is, instead of using ten leaf digits per stem, either two or five leaf digits per stem could be used. If five are used, each ten-digit stem is split up into two five-digit stems. This is done in Figure 2.6b. The "*" means that digits 0,1,2,3 and 4 appear as the leaves on that stem; the "." means that digits 5,6,7,8 and 9 are the leaf digits. This spreads out the display a bit, but still not enough.

FIGURE 2.6

These Stem-and-Leaf Displays are too Squeezed

a)

```
 6|1122222333344555556666777888888999    LEGEND:  6|1 = 61.00 to 69.99
 7|0
```

b)

```
6*|1122222333344                        LEGEND: 6*|1 = 61.00 to 64.99
6.|555556666777888888999
7*|0
```

For our next attempt, we split up each original ten-digit stem into five stems with two leaf digits per stem. This is shown in Figure 2.7a. At last we are getting somewhere! The labelling on the stems was invented by John Tukey and should be read as:

o = zeros (0's) and ones
t = twos and threes
f = fours and fives
s = sixes and sevens
. = eights and nines (our alphabetic luck ran out)

Finally, if we want still more detail, we can add another digit and truncate to three digits instead of two. When we do this, we get Figure 2.7b, in which we now have two-digit stems and the leaf digits are either 0,5 or 7 (corresponding to inches, ½ inches and 3/4 inches). Which of the last two displays you prefer is a matter of taste; they are both "correct". But you should compare them to the earlier histograms and note how much more information they give. Also note the use of a legend to help clarify the display and,in particular, to let the viewer know where the decimal point should go.

FIGURE 2.7

More Ways to Spread Out the Stem-and-Leaf Display of Student Heights

a)

6o	11
6t	222223333
6f	4455555
6s	6666777
6.	888888999
7o	0

LEGEND: 6o|1 = 60.00 to 61.99

b)

61	07
62	00007
63	0000
64	00
65	00005
66	0000
67	000
68	000000
69	000
70	0

LEGEND: 61|7 = 61.70 TO 61.79

The references listed in appendix 1 contain many examples of stem-and-leaf plots, as well as some further enhancements (the use of depths, for example). One very useful variation is a stem-and-leaf display where the leaf digits are replaced by symbols representing different categories. This is sometimes called an "inside-out plot." For example, for the lake data given in Table 2.2 the lakes are located on three different continents: Asia (for which we will use the symbol A), Africa (symbol F), and North America (symbol N). Figure 2.8a gives the usual stem-and-leaf display. Figure 2.8b gives the stem-and-leaf display when the leaves are replaced by the symbols representing the different continents (since the Caspian Sea is located in both Asia and Europe, we have arbitrarily labelled it as being in Asia) . Figure 2.8b shows that the five longest lakes are located outside of North America, while eight of the ten next longest lakes are located in North America. Of course this information is not available from 2.8a.

FIGURE 2.8

The Inside-out Plot

a) The Lengths of the Largest Lakes

```
1|99                    LEGEND: 1|9 = 190 TO 199
2|045689
3|05679
4|2
5|
6|
7|6
```

b) Replacing the Leaves by the Symbols of the Continents

```
1|NN                    LEGEND: 1|N = A LAKE IN NORTH AMERICA
2|NNFNAN                BETWEEN 100 AND 199 MILES IN LENGTH
3|NNFAA
4|F                     A = ASIA
5|                      F = AFRICA
6|                      N = NORTH AMERICA
7|A
```

In summary, the stem-and-leaf display is one of the most useful graphical techniques available for looking at bunches of data. Sometimes, however, you have several bunches of data that you want to compare, and too many stem-and-leaf displays would overwhelm the viewer with information. In such cases, what we need is a different way to graphically summarize the data that will neither overwhelm the viewer with detail, nor lose too much important information. This is the role of the box-and-whisker plot.

Box-and Whisker Plots

The box-and-whisker plot -- or "boxplot", for short --is a visual display of five pieces of information that together give some very useful facts about a set of data. These five pieces of information are the minimum value, the 25th percentile, the median (which is the 50th percentile), the 75th percentile, and the maximum.

We shall define these terms in a moment, but we must first warn you that the above definition is not universal. There are some slight variations that are often used. The biggest variation is with the use of the maximum and minimum values. If outliers are present, we usually don't want to plot them on the same scale as the rest of the data for the reasons discussed previously. To deal with this, John Tukey, the inventor of the boxplot, developed a slightly different boxplot procedure in which the maximum and minimum are replaced, if necessary, by essentially the largest and/or smallest nonoutlier values. Exactly what "if necessary" and "nonoutliers" mean are defined in his procedures. Readers interested in the (fairly simple) details should consult the references.

The other variation occur in the exact way in which the 25th, 50th, and 75th percentile are defined. However, the various definitions are only slightly different. These minor differences almost always have no (or only a negligible effect) on the boxplots. The definitions we have included here are the ones that most people find the easiest to use.

Now for the definitions. First, the median: the median is the middle value when a set of numbers is put in numerical order. For example, for the numbers 1, 1, 7, 12, and 13, the median is 7. For the numbers 1, 1, 7, 10, 12, and 13, the middle is halfway between the 7 and 10. For such cases, the median is defined as the average of the two middle values: (7+10)/2 = 8.5.

The general definition of a percentile is a bit awkward, and we will ignore it in this Guide. We can, however, easily define the 25th and 75th percentiles. Basically, the 25th percentile of a bunch of data (often called the first or lower quartile) can be thought of as that value for which 25% of the data values are below and 75% of the data values are above. Similarly, the 75th percentile (often called the third or upper quartile) is that value for which 75% of the data values are below and 25% are above. A simple working definition is to define the 25th percentile as the median of the data with values less than or equal to the overall median, and the 75th percentile as the median of the data with values greater than or equal to the overall median. Essentially, these are just the medians of the lower and upper halves of the ordered data.

For the height data for all 92 students, the minimum is 61, the 25th percentile is 66, the median is 69, the 75th percentile is 72 and the maximum is 75. For the 57 males the minimum is 66, the 25th percentile is 69, the median is 71, the 75th percentile is 73, and the maximum is 75. For the 35 females the minimum is 61, the 25th percentile is 63, the median is 65.5, the

75th percentile is 68, and the maximum is 70. The easiest way to see this is to construct either an ordered stem-and-leaf display (such as Figure 2.5) or a dotplot.

Now we are ready to draw a boxplot. We proceed as follows:

1. Draw a number line that contains the range of values of the data from the minimum to the maximum (or to the maximum and minimum nonoutliers, if severe outliers are present).

2. Put a small dot just slightly above the number line where the values of the minimum, 25th percentile, median, 75th percentile, and maximum are located.

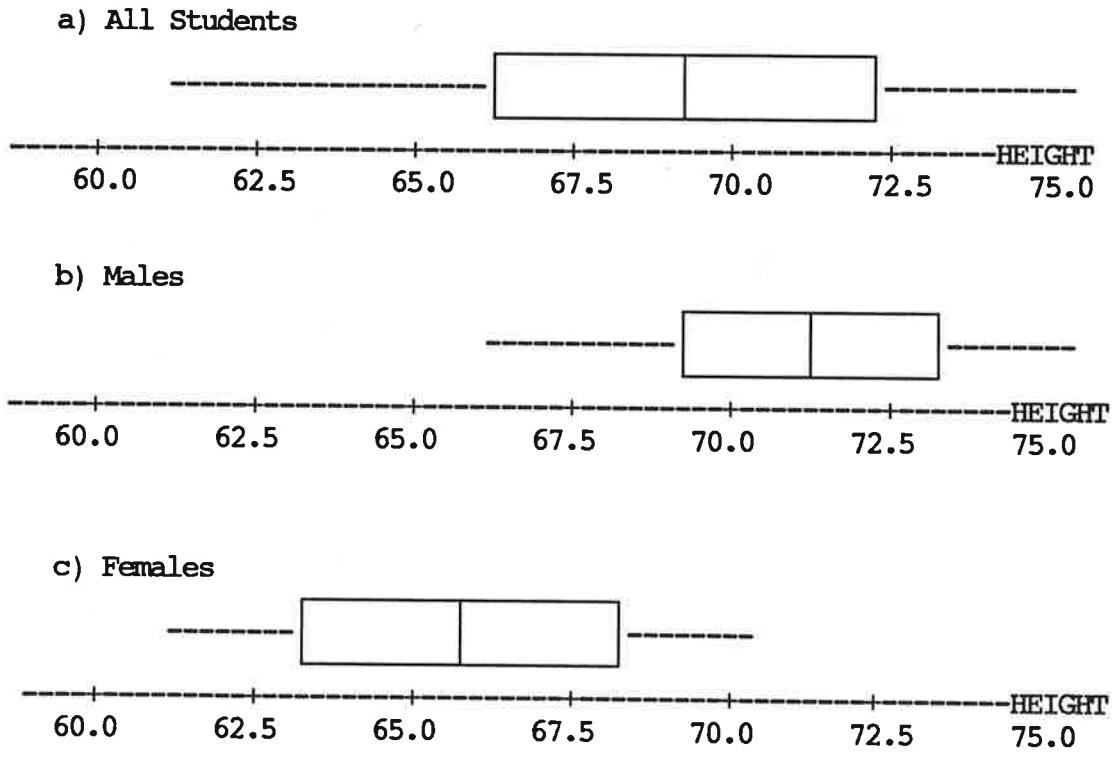
3. Draw a small vertical line through the dot for the median; connect the dots for the 25th and 75th percentiles with a box.

4. Finally, draw dashed lines from the 25th percentile to the minimum and from the 75th percentile to the maximum. These lines are often referred to as "whiskers". List any omitted outliers separately.

Figure 2.9a gives the boxplot for the heights of all the students. Figures 2.9b and 2.9c give boxplots for male students and female students separately.

FIGURE 2.9

Boxplots of Student Heights

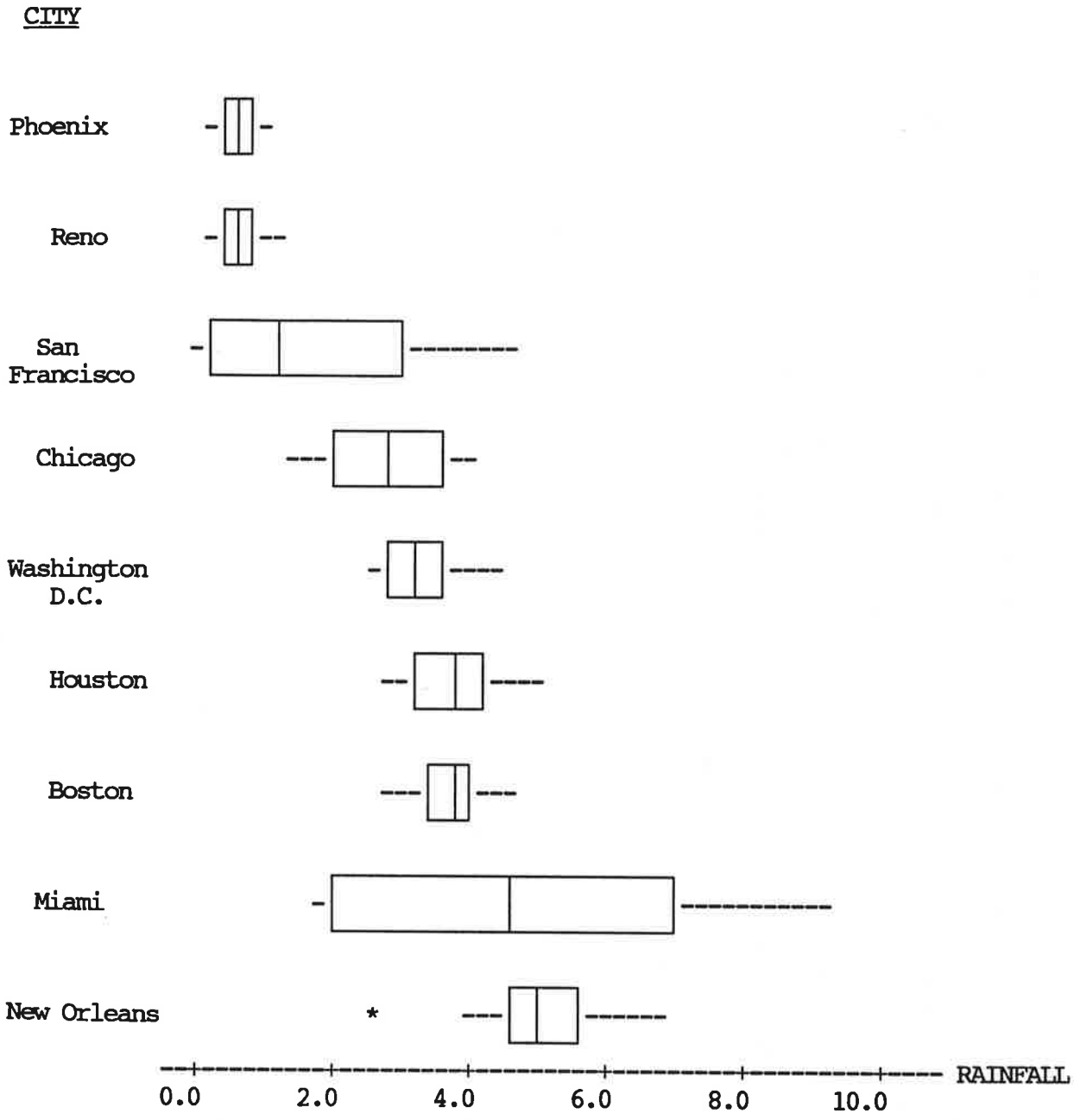


Even though it may seem at first that these boxplots give very little help in interpreting the data, this is not true. They allow you to see patterns that are not evident from the other displays. For example, from the boxplots in Figure 2.9, you can see that the middle 50% (i.e., those students whose heights fall within the box) for the females is more spread out than the middle 50% for the males. It is also immediately clear that 75% of the female heights are below the 25th percentile for the male heights and that all the female heights are below the median for the male heights.

Boxplots are particularly useful in comparing many different groups where the eye would be overwhelmed by the detail of too many stem-and-leaf displays or histograms. Figure 2.10 is an example. It gives boxplots of average monthly precipitations for nine U.S. cities (source: 1989 Statistical Abstract of the United States). By comparing the medians, we can quickly see which cities are the wettest and the driest. In addition, the lengths of the boxes indicate how variable (i.e., how spread out) monthly rainfall is and we thus easily can compare the cities in terms of variation in monthly rainfall. The whiskers show how variable the wetter and drier months are within each city. Note how variable Miami is in comparison to New Orleans, even though their median rainfalls are not that much different. The asterisk in the New Orleans boxplot indicates that one dry month (October) is very different from the rest of the year (i.e., it is an outlier). Finally, the asymmetry in the whiskers in the Miami and San Francisco boxplots shows that these locations tend to have rainy seasons -- a few months with a lot more rain than the rest of the year. For your information, in San Francisco, this occurs in the winter; in Miami, it is in the summer.

FIGURE 2.10

Boxplots of Monthly Average Rainfall in Inches for Nine Selected Cities



Although we could have made nine separate dotplots or stem-and-leaf displays, it is unlikely that we could see as much so easily as we can with the boxplots. Sometimes by removing information from a graph, more can be seen than when the information is left in.

This concludes our brief catalogue of simple displays for looking at bunches of data. Despite their simplicity, these procedures can produce attractive, information-rich graphics. They should always be tried when analyzing and comparing sets of data.

In the next chapter, we'll look at some more sophisticated graphs for answering another kind of question: how are various measured characteristics related to one another? For a particular example: how are the weights and heights of the 92 college students related? That is the realm of two (and more) dimensional scatterplots.

CHAPTER 3

USING GRAPHICS TO LOOK AT RELATIONSHIPS

by Innis Sande
Bell Communications Research

If you look at the weight and height data for the 92 college students of the last chapter, it appears that they are related: knowing someone's height may tell you something about his/her weight. So studying the height and weight of the college students as separate aspects doesn't tell the whole story. The relationship between them may also be imporant.

The study of statistical relationships reveals some of the structure of the world we live in and enables us to deal with it more intelligently. How do we determine what dose of a drug is effective? Does it make any difference whether we are dealing with an adult or a baby? How does a designer of ready-to-wear clothing patterns know how to change the pattern to accommodate different sizes? Is a person who weighs 140 lbs overweight?

Observed relationships may suggest that X causes Y, either directly or indirectly: as you increase the amount of fertilizer in a field, the plants grow taller, but if you add too much, they die. There may be a mutual dependence on other factors, such as that of height and weight on our genetics and diet. The maximum daily temperature varies in a predictable way depending on the time of year, as does household usage of electricity. This enables us to know what clothing to take on our vacation or make a good guess at what the electricity bill might be next month.

Of course, we can think of lots of good reasons why height and weight are probably related. But how about height and I.Q. scores? Does knowing someone's height tell you anything about their I.Q. scores -- or how well they will do on a math test? It's doubtful, but it would certainly be very interesting if it were so.

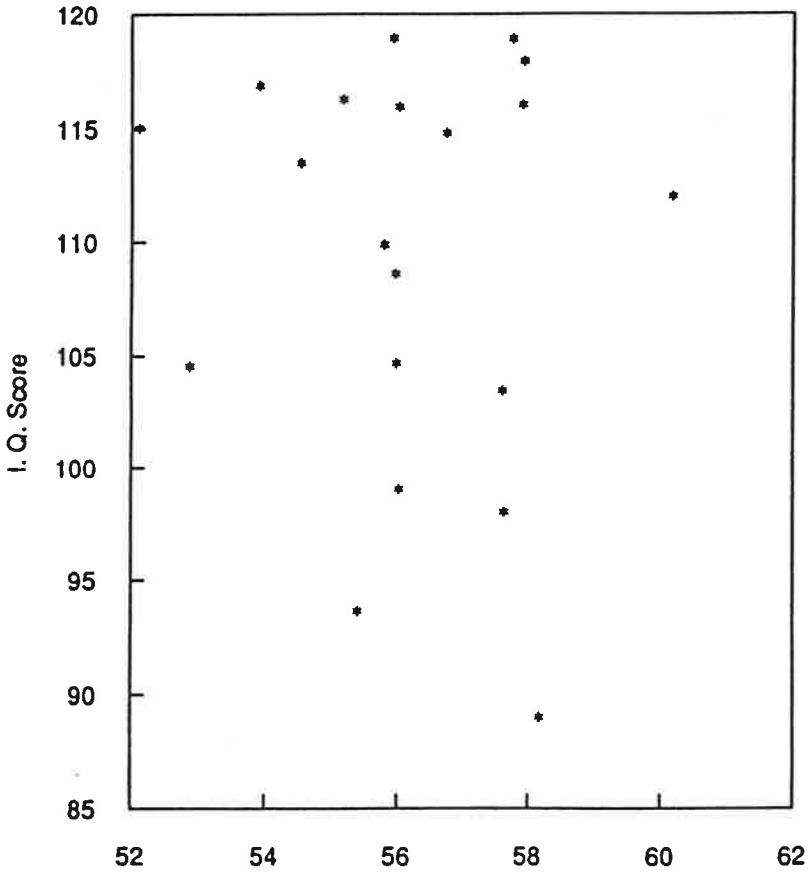
Rather than guess at the answers, we should turn to more scientific methods of finding out, and the easiest and most obvious method is to collect data and take a look at them.

Scatterplots

Once we know the I.Q. score and height for all the students in a class, we can plot them on some graph paper, using height as the x-value (horizontal distance) and I.Q. as the y-value (vertical distance). This type of graph is called a scatterplot. Each individual gives us a point on our plot (a height and I.Q. value), and the result might be something like Figure 3.1a. If we look at Figure 3.1a, we see that knowing that someone is relatively tall tells us nothing extra about his/her I.Q. Tall people and short people in this group have similar I.Q.'s. That is, there appears to be no relationship between height and I.Q.

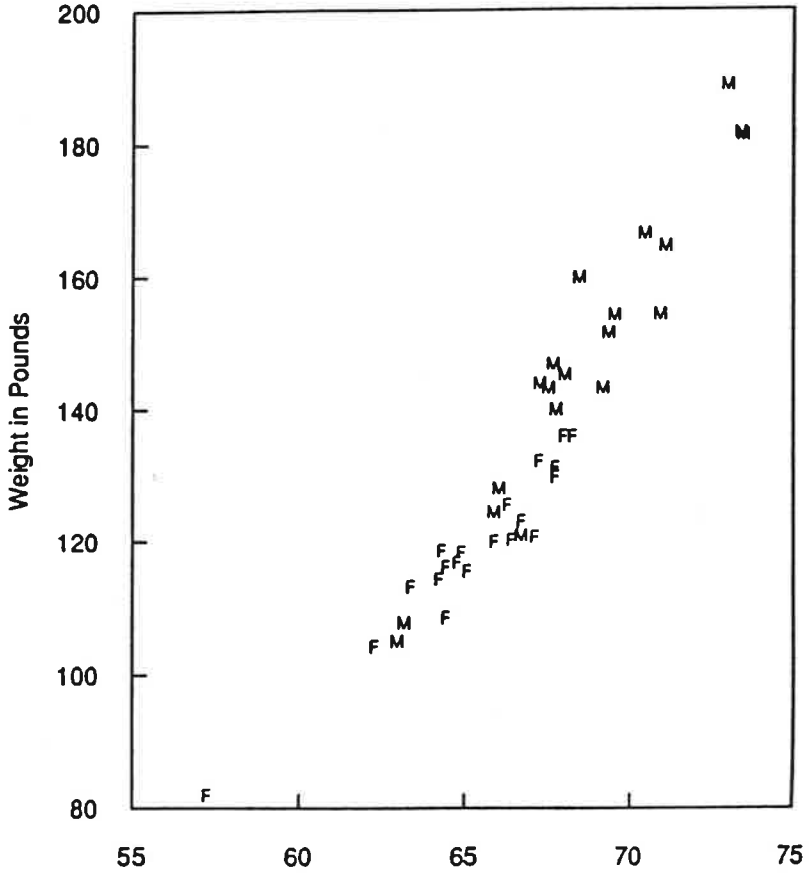
Figure 3.1: Showing Relationships through Plots

3.1a. No Relationship



I.Q. Score Plotted against Height

3.1b. Obvious Relationship



Weight Plotted against Height for a Mixed Class

This does not make this scatterplot useless and uninteresting. It is not hard to think of examples where the existence of no relationship between variables may be extremely important. To give just one, if the dosage of a proposed new drug was plotted on the x-axis and the percentage of people treated with it who suffered a bad side effect was plotted on the y-axis, seeing no relationship (at least over the range of dosages that would be used in practice) is exactly what the drug manufacturer would like to see. That would be evidence for the safety of the drug.

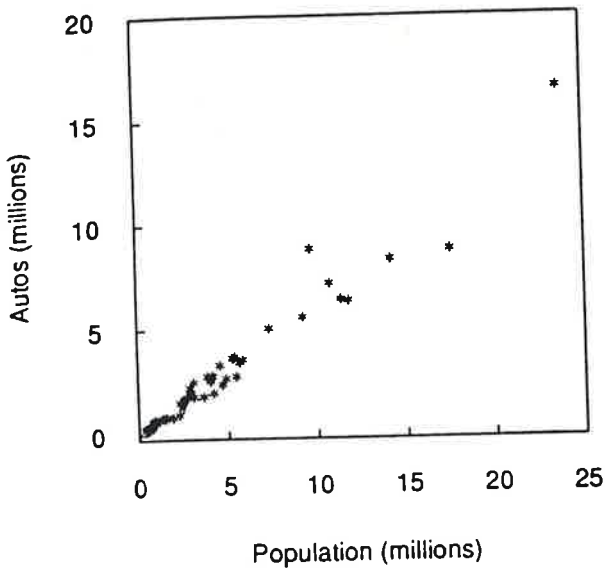
By contrast, if we collected the height and weight for a (different) group of people we might get a plot like Figure 3.1b. In this plot, measurements from males are plotted with M's instead of *'s, and measurements from females are plotted as F's. It is fairly obvious that not only does weight generally increase with height in this group, but males are generally heavier and taller than females. It is important to remember that the conclusions one might draw from a scatterplot apply only to the group represented by the individuals in the plot and that the same data plotted for different groups may result in different-looking plots. Try making the same kind of plot for the height/weight data in Table 2.1. Does it look the same (be careful with the scaling!)?

Now let's look at the relationship between the population and the number of automobiles for the 50 states and the District of Columbia in 1988. These data are represented in Figure 3.2a. Clearly there is quite a strong relationship between the two quantities. Which are the states with the largest populations and numbers of autos? In Figure 3.2b, instead of each state being represented by a star, a two-letter abbreviation has been used instead. Here it is easy to see that California is the most populous state in terms of both people and autos, and that New York,Texas, Florida, Ohio, Illinois, Pennsylvania, Michigan and New Jersey are next. Smaller than that, it is not possible to distinguish the individual identifiers; in fact, the plot looks like a mess and making the letters smaller is not going to help. The solution may be something like Figure 3.2c, where only the states with populations over 10 million or more than 7 million autos are represented by letters and the rest are plotted as circles.

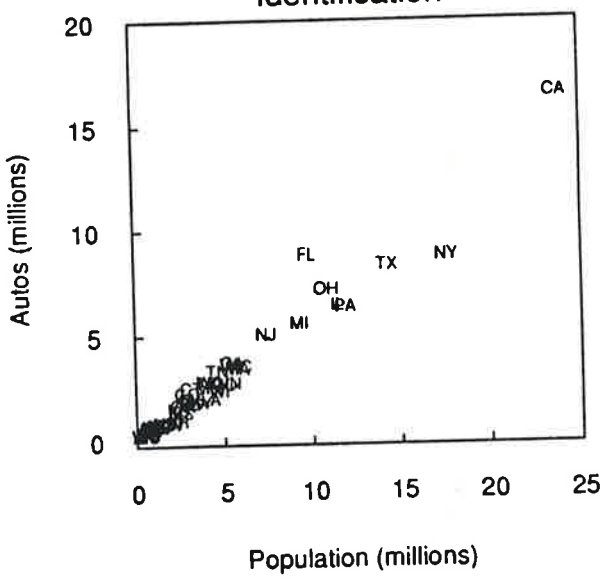
Although Figures 3.2a-c show us the large-scale relationship between numbers of autos and population, we essentially end up drawing conclusions based on the largest 9 states because they dominate the plots. If we want to look more closely at the 42 smallest states, we have to do something else. One option is to plot the data for the remaining 42 states, but before we do anything, let's take a look at the one-dimensional distributions of population and autos separately.

Figure 3.2: Relating Autos to Population

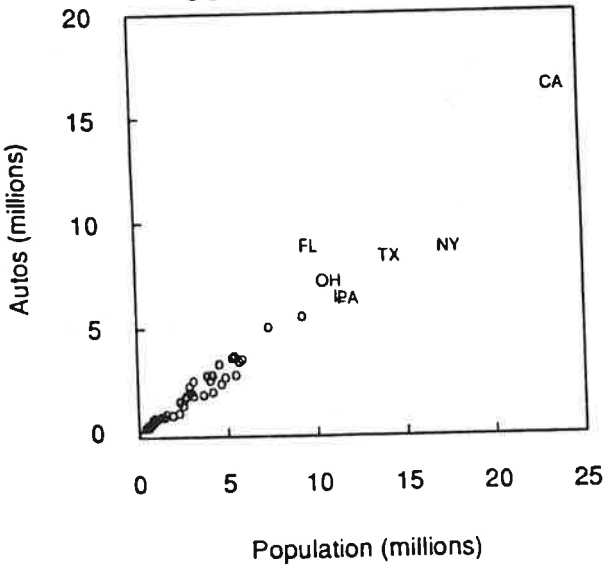
3.2a. Plain Scatter Plot



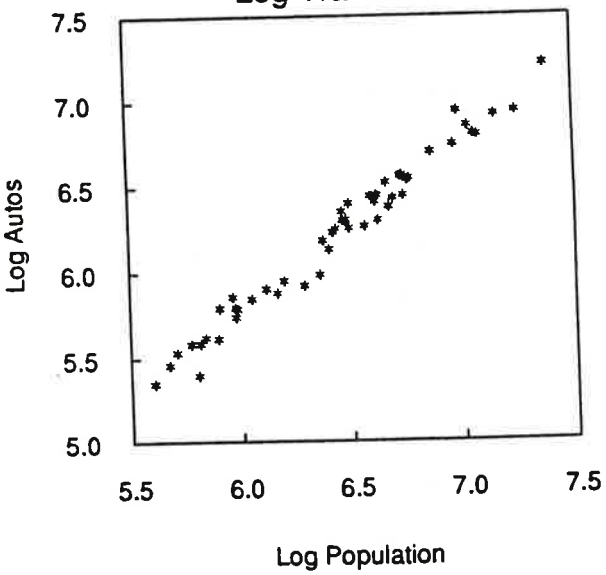
3.2b. Scatter Plot with Identification



3.2c. Scatter Plot with Selective Identification



3.2d. Same data, with Log Transform



These are stem-and-leaf plots of the populations of people and autos of the 50 states and DC:

People

Millions	Hundred-thousands
0	455667788999
1	013569
2	3456799
3	01179
4	1122679
5	35579
6	
7	4
8	
9	37
10	8
11	49
12	
13	
14	2

Too high to include: 17.558m
23.668m

Autos

Millions	Hundred-thousands
0	233344446666778889
1	0357888
2	000245578888
3	34567
4	
5	06
6	34
7	2
8	388

Too high: 16.473m

It is easy to see that both variables are quite skewed toward the smaller values. Simply deleting the largest values would not remove the problem - it would just change the scale. This kind of problem is usually solved by transforming the data. This means that, instead of plotting x, one plots f(x), where f is some convenient function that (in this case) spreads out the small values and squeezes the big ones. A function frequently used for this purpose is the log function. This is the same idea that was used in the plot of stony meteorite data in Chapter 1, Figure 1.3.

If we take logarithms to the base 10 of both populations and numbers of autos, the resulting stem-and-leaf plots (of the first two significant digits, with 2 different leaf values on each stem) look like this:

Log_{10}(Population)	
2	677
2	888899
3	000011
3	223
3	4444455555
3	666666777777
3	889
4	00011
4	22
4	4

Log_{10}(Autos)	
2	4455
2	66667
2	888899999
3	01
3	223333333
3	44444444555
3	6677
3	889999
4	
4	2

The transformed data are now more evenly spread out and more symmetrical than the original data.

If we now scatterplot the transformed data, we get Figure 3.2d. Here the points are quite well separated and California is not as far from the rest as it looked in Figures 3.2a-c. The relationship between population and number of autos by state is very clear.

Fitting a Prediction Line to the Plot

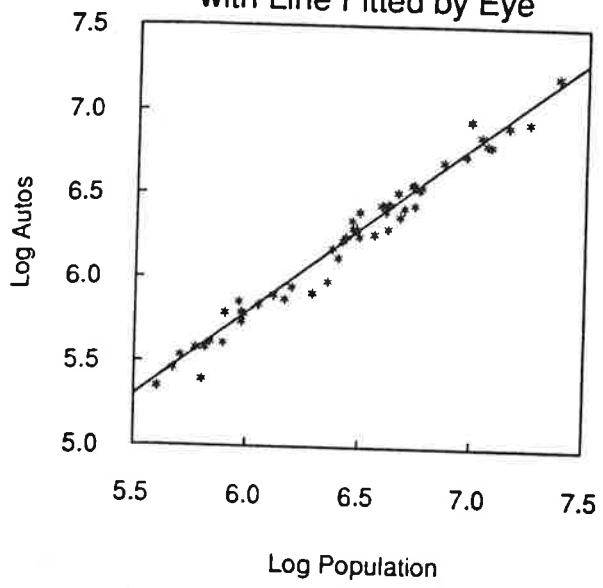
Suppose a new state were to be created -- for example, California or New York might be chopped in half. Could we then predict the number of autos the new state would have? Based on the plots in Figure 3.2, it looks as though it should be possible to predict quite well. The data look as though they lie very close to a straight line. If we can draw a line which represents or summarizes the data well, we simply use the line to "predict" the y- value (number of autos) corresponding to a given x-value (population).

In Figure 3.3a, a line that seems to fit the data well has been drawn "by eye" with a ruler. For data such as we have here, an informal procedure like this suffices. Often, however, the relationship obviously exists, but it is less obvious where to put the ruler. Then we need a more objective method of drawing lines. It turns out that there are lots of such methods from which to choose.

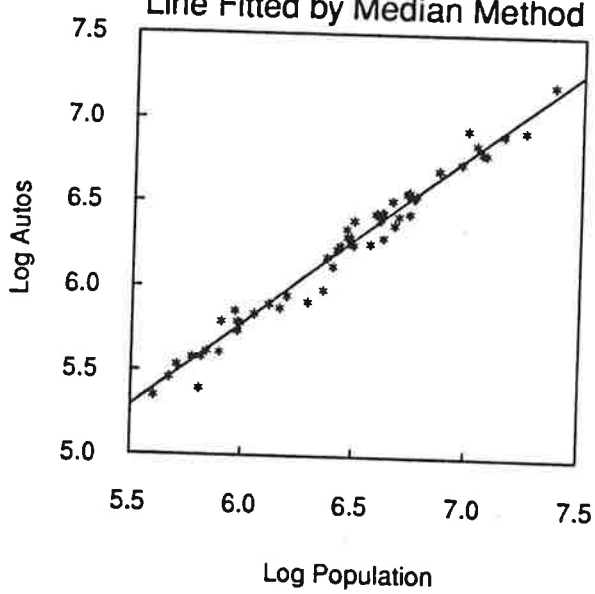
Figure 3.3: Relating Autos to Population (cont.)

Fitting Straight Lines and Plotting Deviations

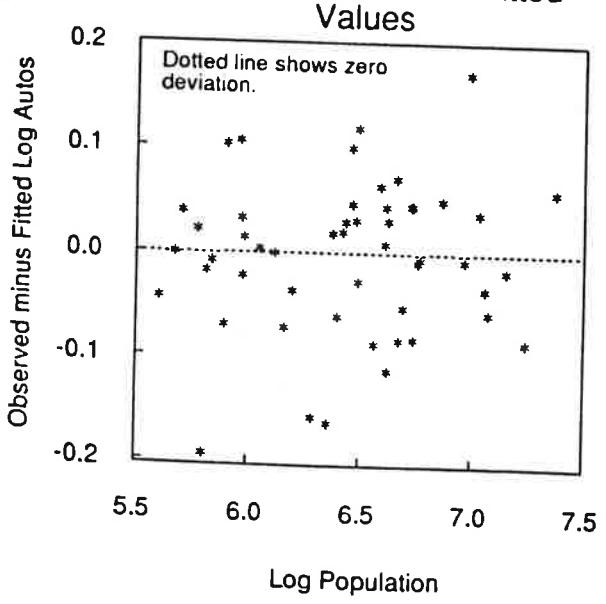
3.3a. Transformed Data, with Line Fitted by Eye



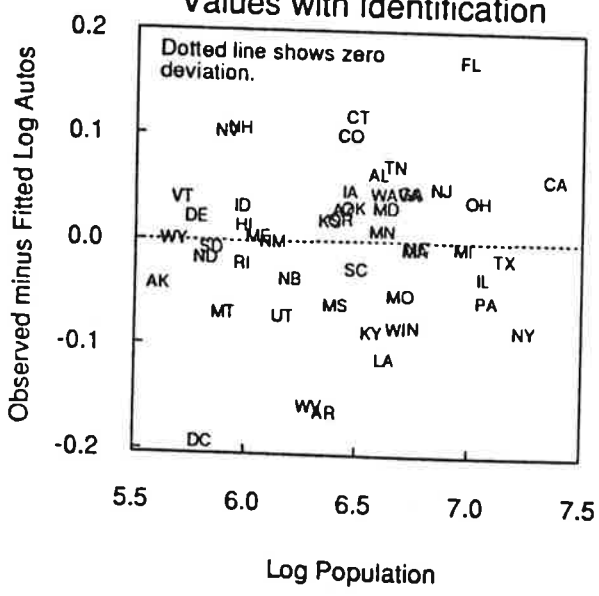
3.3b. Transformed Data with Line Fitted by Median Method



3.3c. Deviations from Fitted Values



3.3d. Deviations from Fitted Values with Identification



Methods of "fitting" lines (i.e., identifying "good" lines) to data vary in degree of sophistication. Here we will consider a very simple method, the median method, which will usually give a reasonable-looking line.

In the last chapter, the median of a group of numbers was defined as the middle value when they are arranged in order. The median is often used as a representative value for a set of values, in much the same way as an average. However, with a median, we know that it divides the data in half, whereas the average does not (except, sometimes, by accident).

To use the median method of fitting straight lines to (x,y) data like populations and autos proceed as in the following example:

Suppose, to keep the numbers simple, that the complete data are

x-values	1	2	3	4	5	6	7	8	9	10
y-values	3	1	4	2	5	8	6	4	7	5.

In other words, the observations are (x=1,y=3), (x=2,y=1), and so on.

1. Order the x-values.

2. Split the ordered x-values into three equal or nearly equal groups. For the above example (where the x-values are already ordered), 1 2 3 4 5 6 7 8 9 10 might be split into 1 2 3, 4 5 6, and 7 8 9 10.

3. Calculate the median of the x-values for the smallest and largest groups. For the example above they would be 2 and 8.5.

4. Now calculate the medians of the groups of y-values corresponding to the smallest and largest groups identified in step 3. In our example, the y-values corresponding to the smallest group of x-values are 3 1 4 and their median is 3, while the median y for the largest group is 5.5

5. Estimate the slope of the line as

   b = Difference between the median y's/Difference between the median x's.

   In the example we would get (5.5 - 3)/(8.5-2), or 2.5/6.5 = .38.

6. Estimate the intercept of the line as

   a = median of the values (y - bx)

   (The rationale for this is that, if we have a bunch of (x,y) pairs, each of which is supposed to lie approximately on the line y = a+ bx and if we know b approximately, then each pair gives us an estimate of the value of y - bx = a. A central value for these numbers gives us a better overall estimate of a.)

In the example, these values (corresponding to the ordered x-values above) would be 2.62, 0.23, 2.85, 0.46, 3.08, 5.69, 3.31, 0.92, 3.54, 1.15, and their median is 2.73.

7. Draw the line y = a + bx through the points on the scatterplot.

It is possible to refine this method somewhat, but this is enough to understand the principles without getting bogged down in details.

In Figure 3.3b, the line drawn by the median method does not appear to differ much from the by-eye line drawn in Figure 3.3a. The two lines are very close, in this case, but they are not the same. (In case you're wondering, we prefer to use medians rather than averages because they are less sensitive to outliers. If, in step 4 above, the y-value corresponding to x=9 were 19 instead of 7, the median y for that group would not change, but the average would change considerably. Try it.)

The fitted line can be used to predict a y-value for every x-value. (Statisticians often use the word "predict" to mean "make an intelligent guess at, based on related information.") If we invent a new state with population P and we want to predict A, the corresponding number of autos, we can put $x=\log_{10}(P)$ and read off the line (or from its formula) the corresponding value of $y=\log_{10}(A)$. The predicted value of A is then 10^y.

For example, from Figure 3.3b, if a new state had a population of P = 1.995262 million, $\log_{10}(P)$ is 6.3. The formula for the fitted line is $\log_{10}(A) = -.2 + \log_{10}(P)$. $\log_{10}(A)$ is then $-.2 + 6.3 = 6.1$ and $10^{6.1}$ is 1.258925 million. So we can predict that our new state will have about 1.26 million autos!

Although we actually know the number autos in each state, we can still use the fitted line to "predict" them from the state populations. We would like to compare the predicted number of autos with the actual number to see how good the fit is. Figure 3.3c is a plot of the difference between the actual value of $\log_{10}$(autos) and the value predicted from the line fitted by the median method against $\log_{10}$(population). That is, we have plotted

(actual - fitted)$\log_{10}$(autos) on the y axis

vs.

$\log_{10}$(population) on the x-axis.

(Note that we have changed the y-axis scale from the previous plots to show these results clearly.)

This is a plot of "error of prediction" vs the log population to see how well we predict over the entire population range and whether we have any systematic errors over that range. The plot also shows us how much variation in the 51 values of $\log_{10}$(autos) is left over after we have removed the

variation accounted for by $\log_{10}$(population). The dotted line indicates zero deviation from the predicted number of autos (perfect prediction).

Figure 3.3d is the same plot as Figure 3.3c, but with the individual states identified. We see that Florida has relatively more autos for its population size than any other state (actual - fitted is large), while DC has the smallest number of autos relative to size; if DC is excluded (it's not really a state), West Virginia and Arkansas have the next smallest number of autos relative to size.

We have seen that, not too surprisingly, population size does a pretty good job of explaining the number of autos in each state. Sometimes the addition of another variable can do more explaining. Let us try to explain the difference between the actual and predicted numbers of autos that we plotted in Figures 3.3c and 3.3d by introducing a new variable, average (i.e. per capita) income by state, for 1988.

Again, the distribution of average income across states is quite skew (can you explain why?), so we use $\log_{10}$(income) instead. Figure 3.4a plots the differences between the observed number of autos and the number predicted from population size alone against $\log_{10}$(income). Each state is identified. The plot shows some trend, although not as sharp as that in Figure 3.3a. In other words, it appears that as per capita state income increases, the actual number of autos tends to exceed that predicted by state population alone more and more. This sounds reasonable.

In Figure 3.4b, a median straight line is fitted to these data, and it exhibits a gradual upward slope. In Figure 3.4c, new deviations have been computed from the new fitted line and plotted against income and the individual states identified. DC stands out as having a relatively small number of autos for its population size and average income. The "real" states are quite comfortably clustered about the zero deviation line. Average income does seem to account for some of the variation remaining after taking out that due to population size.

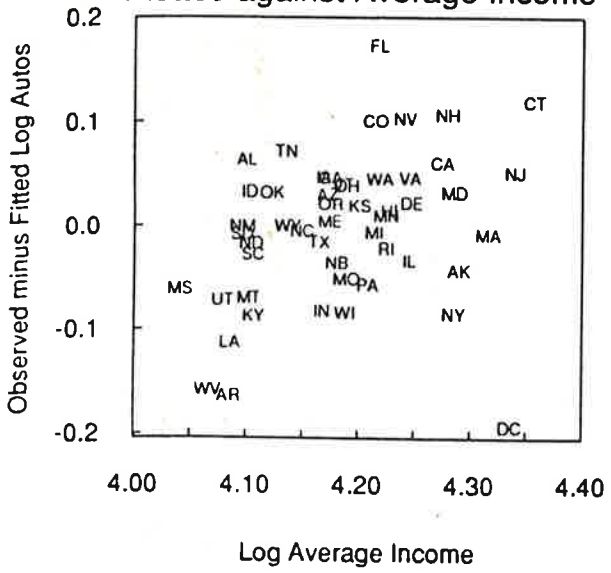
This technique of fitting a line and using the deviations as y-values for fitting a new line with new x-values only works if the new x-variable is more or less unrelated to to the original x-variable -- in this case, if income is unrelated to population size. Figure 3.4d is a plot of $\log_{10}$(income) against $\log_{10}$(population), and it does not appear that any relationship exists. In this case, we can actually translate the prediction lines from the graphs directly into an equation that expresses the relationship of (the logs of) number of autos to both population and per capita income simultaneously.

It turns out that this can be done with even more x-variables (maybe we should add size of the states) and even if the x-variables are related. However, the techniques for doing this are beyond what we can discuss here.

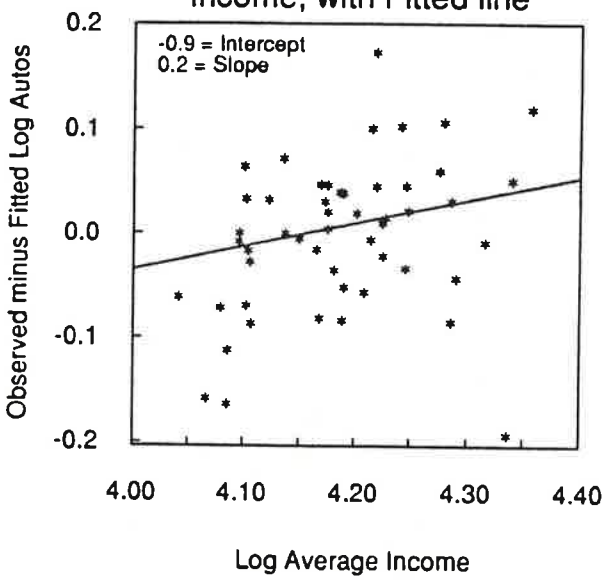
Figure 3.4: Relating Autos to Population (cont.)

Explaining the Deviations from the Fit

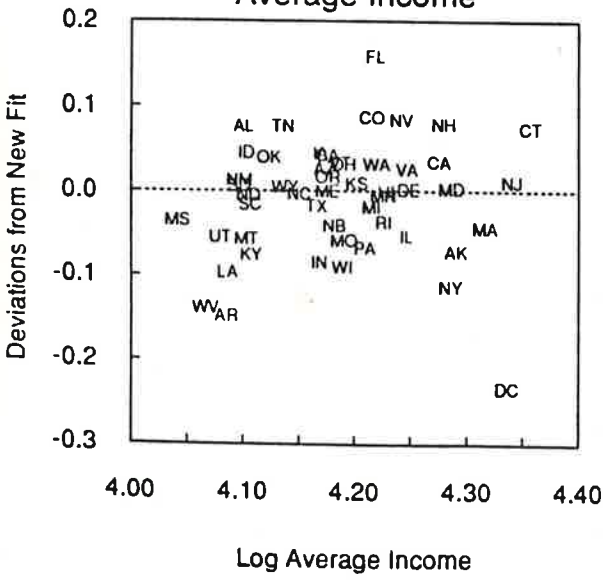
3.4a. Deviations from Fitted Values Plotted against Average Income



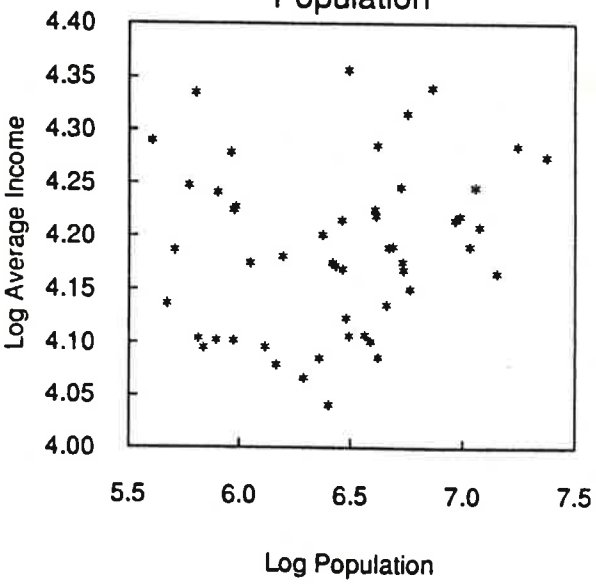
3.4b. Deviations vs. Average Income, with Fitted line



3.4c. New Deviations vs. Average Income



3.4d. Average Income vs. Population



Time Series Plots

Frequently data consist of a series of observations taken over time. For example, one might be interested in the daily maximum temperature or the monthly rainfall over several years. Data such as this is the basis for our knowledge of the weather in different parts of the world and how the climate is changing. If you plotted the monthly rainfall in your home town every month, what do you think the plot would look like over one year? What would it look like over 5 years? What would a very wet year look like compared to a very dry year? What would you expect the plot to look like if the rainfall were decreasing over many years?

Data observed over time are referred to as Time Series. It is often the kind of data we collect when we are monitoring something, such as rainfall, the price of gas or the unemployment rate. Data on weather, retail sales, prices, and so forth often exhibit interesting annual patterns, which we call cycles. Sunspots exhibit longer cycles. Our brains exhibit electrical activity with very short cycles. Time series data does not have to exhibit cycles to be interesting. The population of the United States increases over time, as does its consumption of energy. What will the population of the United States and its energy consumption be in the year 2050? While the proportion of deaths every year due to respiratory disease has been decreasing since the beginning of the 20th century, the proportion due to cancer has been increasing.

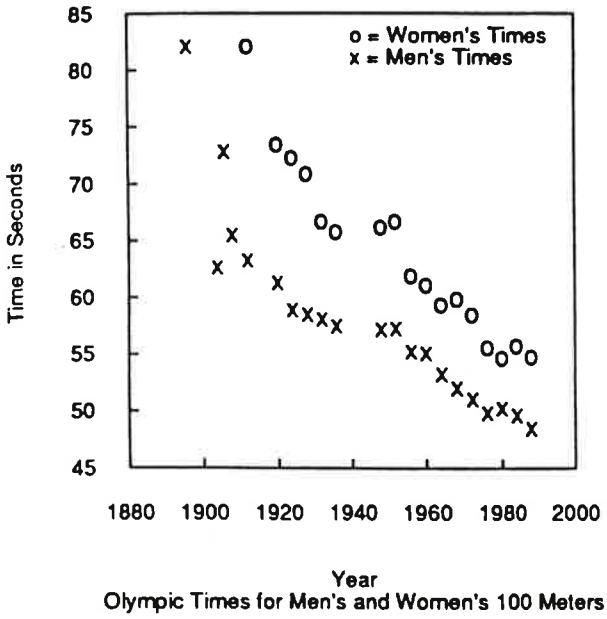
Unfortunately, we will not be able here to look at some of the "more interesting" series, simply because they tend to be very long and the techniques for looking at them are quite complex. However, to give you an idea of how we might think about a time series, we use a simple example.

Figure 3.5a shows plots of Men's and Women's Olympic times for the 100 Meter Freestyle swimming event. The Olympics occur roughly every 4 years (the first few since 1896 were more irregular and there were no Olympics in 1916, 1940 or 1944). The Men's Freestyle event has been held since 1896, while the Women's event was first held in 1912. The two plots are different. Men obviously swim the 100 Meters Freestyle faster than women do, but it is hard to say how much faster.

Are the women closing the gap? To examine this issue we might look at the difference between men's times and women's times, or the ratio of the two. Figure 3.5b shows the ratio of women's to men's times for the 17 years in which both men and women competed. The points are quite irregular, but they do suggest an increasing trend. However, it is impossible to guess from this plot whether the women are closing the gap or whether, in the long run, the ratio of men's times to women's times will flatten out at some limit which is less than 1.0.

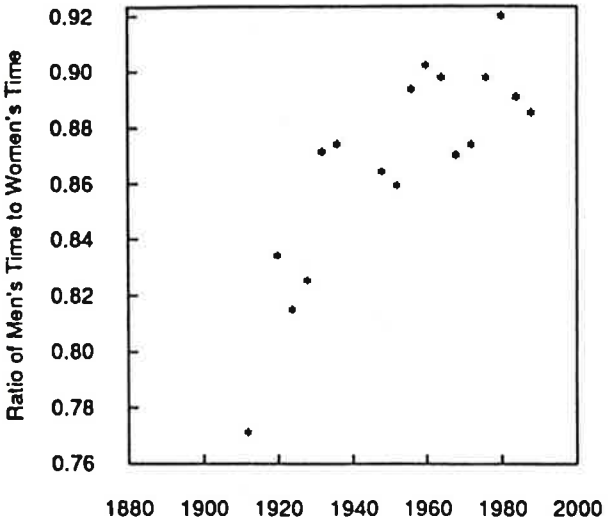
Figure 3.5: Exploring Time Series

3.5a. Two Time Series Plots



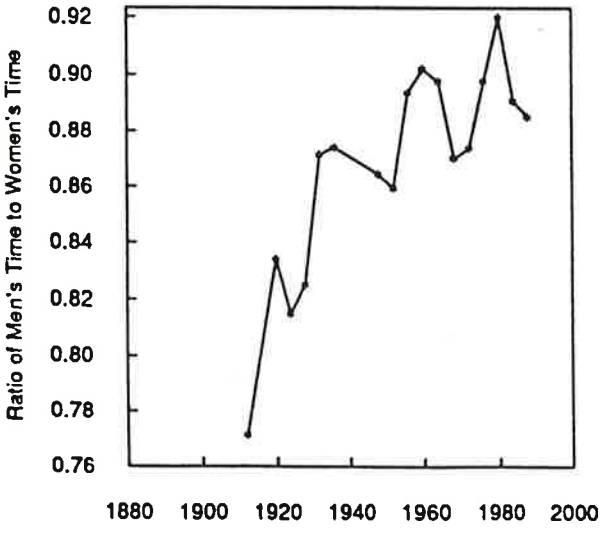
Olympic Times for Men's and Women's 100 Meters

3.5b. Deriving a New Plot



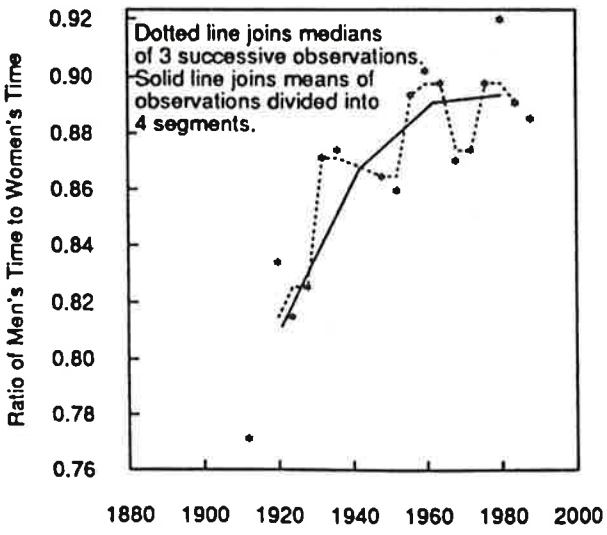
Comparison of Men's and Women's Times

3.5c. Looking for Trend on a Time Series Plot



Connecting Successive Points

3.5d. Smoothing a Time Series Plot



Smoothed Ratios of Times

In order get a better picture of the trend, we can try connecting the successive points of the plot. This is shown on Figure 3.5c. A jagged plot emerges which helps somewhat, but does not really solve our problem. We need to fit a curve to smooth out the plot. We prefer a curve because fitting a straight line seems inappropriate due to the definite curvature.

In order to produce a smoother version of Figure 3.5b, we shall again use a median procedure to fit smoothed values. This time the "x-values" are the time values: 1912, 1920, etc., and the "y-values" are the corresponding ratios: .77, .81, etc. Our smoothed y-values are produced as follows:

1. Group the ratio data into overlapping groups of three by sliding a "window" of width 3 along the data.

   In this example, the 1st group is (.77, .83, .81), corresponding to (1912,1920, 1924); the 2nd is (.83, .81, .83), corresponding to (1920, 1924, 1928); the 3rd is (.81, .83, .87), corresponding to (1924, 1928, 1932) and so on down to the last group of (.92, .89, .89), corresponding to (1980, 1984, 1988).

2. The smoothed value corresponds to the middle time period of each group and is just the median value of the y-values for that group. Note that this means that there is no predicted value corresponding to the very first and very last value in the time series.

   The medians of the above groups in the example are .81, .83, .83, down to .89. These are the smoothed (or fitted) values corresponding to .83, .81, .83, down to .89 (the second-to-last value).

The complete results are:

Year	Ratio Men/Women	Median (3-point)
1912	0.77	
1920	0.83	0.81
1924	0.81	0.83
1928	0.83	0.83
1932	0.87	0.87
1936	0.87	0.87
1948	0.86	0.86
1952	0.86	0.86
1956	0.89	0.89
1960	0.90	0.90
1964	0.90	0.90
1968	0.87	0.87
1972	0.87	0.87
1976	0.90	0.90
1980	0.92	0.90
1984	0.89	0.89
1988	0.89	

In short, the smoothed value for 1920 is the median of the 1912,1920 and 1924 observations, the predictione for 1924 is the median of the 1920, 1924 and 1928 observations, and so on.

These predicted values are plotted in 3.5d. If we took 5 points in each window instead of 3 points (the number of points is up to us), a slightly smoother curve would result, but there would be two fewer points (we'd have no predicted values corresponding to the first two or last two points) . This method is particularly appropriate for time series where we expect to see (say) annual or weekly cycles (or patterns) and we are interested in the cyclic structure of the series. Using a computer, it is very easy to try different window widths and explore the results.

Of course, just as with the median line fitting procedure, we can and should look at the errors of prediction to see what more might be learned. For example, we might want to plot these errors against some other "explanatory" variable to see if the remaining variation can be explained.

If we are just interested in trend and want to fit a curve to the data in order to get a better idea of where it is going, we can divide the data into (say) 4 adjacent segments of roughly equal numbers of points, and plot the mean (or median) y-value (ratio of men's time to women's time) vs. the mean (or median) x-value (time point). The 4 plotted points are then joined and the result is shown as the solid line in Figure 5d. It appears smoother than the dotted line not because it is better, but because fewer points are used, so that much of the variation between successive Olympics is ignored. From this trend line, it looks as if the ratio of men's to women's times is flattening out at about .9; i.e., the time made by the men will, in the long run, fall roughly 10% short of the time made by the women in the Freestyle.

Of course, we can never be certain about such a prediction. Future breakthroughs in women's athletic training might ennable them to jump to 95% or better, for all we know. This illustrates another problem: it is dangerous to predict outside the range of the data! Nevertheless, based on the information we have thus far, this seems to be a reasonable conclusion.

Summary

In this chapter, we have used variations on the scatterplot to study relationships among variables. Although these graphs can be drawn by hand for small amounts of data, when dealing with larger amounts, it is almost always necessary to have a computer and software so that different kinds of plots can be quickly and easily explored. This kind of interactive graphical exploration of data is rapidly growing as new computer hardware and software become available (and get cheaper!).

The plots in this chapter are somewhat different than those in Chapter 1. In Chapter 1, we were largely concerned with plotting data to communicate what we had learned in a clear and honest fashion. Here, and also in Chapter 2, we were more concerned with plotting the data to learn what we must communicate. Both modes of graphical use are important, and the techniques for each certainly overlap.

We hope that this Guide has shown you how statistical graphics can help you better understand and communicate about the world around you. As we said in the introduction, quantitative literacy is important in today's data-driven world, and graphics is a vital aspect of quantitative literacy.

The annotated reference list that follows will allow you to find out more about statistical graphics. We hope that you will continue to enjoy statistical graphics, and that this introduction has helped you to better understand this important and growing area of knowledge.

APPENDIX

A BRIEF ANNOTATED REFERENCE LIST FOR STATISTICAL GRAPHICS

The following list of references is designed to provide readers asking the question, "But where do I go to learn more," an answer. It is not meant to be exhaustive and, in fact, many useful topics are not covered here. Nevertheless, it should provide a reasonably comprehensive list to begin.

On the whole, no previous background is required to understand these works, though a bit of intellectual effort and perserverence may be required for some.

Boardman, T.J., (1985), "The Use of Simple Graphics to Study Hourly Data with Several Variables," in Experiments in Industry, Snee, R.D., Hare, L.B., and Trout, R.J. (Eds.), 127-142, available from the American Society for Quality Control's Quality Press.

-- This article may be hard to get, but it's worth trying. It shows how graphics can be used to analyze complex air pollution data and may provide good ideas for those interested in the statistical graphics poster competition. No special technical expertise is required to understand the article.

Cleveland, W.S. (1987), "Research in Statistical Graphics", special section on Statistical Graphics in (Journal of the American Statistical Association), 82, 419-423.

-- This article forms the introduction to a section on statistical graphics that includes several (mostly advanced) articles on statistical graphics. However, the article itself provides many references and is written at a basic level accessible to those without special knowledge. It's an excellent resource for those wishing to learn more about statistical graphics.

Cleveland, W.S. (1985), The Elements of Graphing Data, Monterey, CA: Wadsworth.

-- Gives a fairly elementary discussion of many important issues including scaling, choice of symbols, and how to do multiple graphs without confusion. The book also discusses results of psychological research on how people perceive graphics and uses these results to develop better ways to do some of the most commonly used graphs. Profusely illustrated with many examples, it is highly recommended reading.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983), Graphical Methods for Data Analysis, Boston: Duxbury Press.

-- A fine introduction to many modern topics in graphics, data smoothing, and computer approaches to data. This is a somewhat technical exposition, so much of it may be inappropriate for the beginner. Nevertheless, for the curious, it provides a nice exposition of some of the issues in modern statistical graphics. It also has lots of nice examples that can often be followed even without much technical knowledge. The "Further Reading" sections of each chapter are also useful for those seeking to learn more.

Department of Mathematics and Computer Science, North Carolina School of Science and Mathematics (1988), Data Analysis. Available from National Council of Teachers of Mathematics, 1906 Association Drive, Reston, VA 22091.

-- This book was written by high school teachers for high school students. It covers many of the topics in Chapters 2 and 3 of the guide, as well as others not included there. It contains excellent step-by-step descriptions for making the various displays and comes with a computer diskette.

Ehrenberg, A.S.C. (1982), A Primer in Data Reduction, Chichester: John Wiley and Sons.

-- Strictly speaking, this is not a book on statistical graphics, but an introductory statistics textbook that contains some good graphical examples. However, Part V of the book (titled "Communicating Data") consists of four short chapters on "Rounding", "Tables", Graphs", and "Words" that contain much good advice on how to communicate about data. The section on tables is especially valuable.

Huff, D. (1954), How to Lie With Statistics, New York: Norton.

-- A classic. It discusses the mistakes and misinterpretations that can be made (sometimes deliberately!) with graphical displays. It is easy to read, yet makes many important points. The entire book can be read in a couple of hours!

Landwehr, J.M. and Watkins, A.E. (1986), Exploring Data, Palo Alto, CA: Dale Seymour.

-- This is one of the books in the Quantitative Literacy series, written for junior high and high school students. It provides clear directions for making and interpreting stem-and-leafs,boxplots, and scatterplots. It also contains lots of good examples and datasets.

Tufte, E.R. (1983), The Visual Display of Quantitative Information, Cheshire, CT: Graphics Press.

-- A masterpiece! A beautifully illustrated, delightfully written exposition on graphical style. If you can only read ("look at" may be a much better phrase) one book, this is it. It is written for a general audience, so that no technical expertise is required. Yet it is full of wonderful advice and examples. Many of the ideas -- and some of the examples -- in Chapter 1 came from this book.

Tufte, E.R. (1990), Envisioning Information, Cheshire, CT: Graphics Press.

-- This book has a much broader scope than Visual Display. However, it is full of interesting pictures, and might suggest ideas for the poster competition.

Tukey, J.W.(1977), Exploratory Data Analysis, Reading, MA: Addison-Wesley.

-- This is a path-breaking work that expounds the philosophy of "exploratory" -- as opposed to "confirmatory" --data analysis. Many of the techniques introduced here are now firmly established standards in the modern data analytical toolkit. Tukey has an unusual writing style that makes the reading heavy-going at times, but it's worthwhile looking at some of the examples if you are interested in how some of the techniques came about. Also discussed here are more sophisticated versions of the median line fitting and medians of three time series smoothing techniques discussed in the Guide. Tukey does everything by hand, so the reader can follow along (if he has the stamina!). There are also loads of interesting datasets in the exercises.

Velleman, P.F. and Hoaglin, D.C. (1981), Applications, Basics, and Computing of Exploratory Data Analysis, Boston: Duxbury Press.

-- This book is essentially a translation of Tukey's book into a language and form that make it more accessible to the "ordinary" person. Although some technical matters requiring formal knowledge of statistics are discussed, there are also good discussions of boxplots, stem-and-leafs, smoothing and line fitting procedures, and so forth. The book also contains BASIC and FORTRAN listings that should ennable an ambitious student to get a computer to implement most of the techniques (however, commercial software is available that does everything, too).

Wainer, H. (1984), "How to Display Data Badly, The American Statistician, 38, (pp.).

-- Wainer is a colleague of Tufte's. This article is a funny discussion of many horrible examples of graphics in order to show what good graphics should be like. Many of Tufte's ideas and examples are in here, and the article itself is suitable for anyone. It's hard to believe that any article in a statistics journal could make you laugh out loud, but this one can!