Teacher's Edition
Exploring Data

Exploring Data was prepared under the auspices of the American Statistical Association—National Council of Teachers of Mathematics Joint Committee on the Curriculum in Statistics and Probability.

This book is part of the Quantitative Literacy Project, which was funded in part by the National Science Foundation.

Teacher's Edition
Exploring Data

James M. Landwehr
AT&T Bell Laboratories
Murray Hill, New Jersey

Ann E. Watkins
Los Angeles Pierce College
Woodland Hills, California

DALE SEYMOUR PUBLICATIONS

Cover Design: John Edeen
Technical Art: Pat Rogondino
Colleen Donovan

This publication was prepared as part of the American Statistical Association Project—Quantitative Literacy—with partial support of the National Science Foundation Grant No. DPE-8317656. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily represent the views of the National Science Foundation. These materials shall be subject to a royalty-free, irrevocable, worldwide, nonexclusive license in the United States Government to reproduce, perform, translate, and otherwise use and to authorize others to use such materials for Government purposes.

ISBN 0-86651-322-1
Order Number DS01619

DALE SEYMOUR PUBLICATIONS
P.O. BOX 10888
PALO ALTO, CA 94303

b c d e f g h i j - M A - 8 9 3 2 1 0 9 8 7

ACKNOWLEDGMENTS

We are pleased to thank the following people for their assistance during the preparation of this book:

The many teachers who reviewed drafts of the manuscripts and/or participated in the field tests, including Carol Joyce Blumberg, Dorothy Brown, Jim Bruni, Gail Burrill, Pamela Coffield, Clare Conner, Sandra Crepps, Betty Dewey, Michael Dirks, Jim Eiden, Lyle Fisher, Anne Gavin, Reda Gill, Stephen Ginaitis, Landy Godbold, Jay Gowell, Jo Anne Hackworth, Beverly Hartter, Greta Healy, Daniel Hildebrandt, Richard Houde, Arlene Johnson, Randy Kaker, Robert Kolar, Roberta Koss, Ronald LaPorte, Mary Leland, Susan Levy, Bernice Loober, Charles Marion, Bill Medigovich, Crystal Mills, Vera Smith Page, Carl Rasmussen, Carol Rezba, Debra Rodenbaugh-Nelson, Peter Sciarrota, Al Shulte, Murray Siegel, Virginia Stimpson, Marilyn Tahl, Marj Thimmesch, Mary Ann Trovillion, Zal Usiskin, Dena Watson, and Diane Zmaczynski.

The many statisticians who offered their suggestions and reactions, including colleagues at AT&T Bell Laboratories, colleagues at Bell Communications Research, Larry Clevenson, and Gottfried Noether.

The many students who participated in the field tests.

Gail Burrill, for writing answers to a preliminary edition and for suggestions for student projects.

Jim Swift and Jim McBride, for earlier work on data analysis materials for high school students.

John Tukey, for thoughtfully and quickly reviewing our drafts and for his encouragement.

Dick Scheaffer, for his leadership of the Joint Committee and the Quantitative Literacy Project.

Annamaria Doney, for editing our manuscript and seeing it through to publication, and Ruth Cottrell for handling the teacher's edition.

Dorothy Perreca and Shirley Jones, for administering the field tests and teacher training workshops and for obtaining permissions.

Sue Tarczynski and Rose Burnet, for excellent word processing and secretarial services.

Jacque Landwehr, for helping to organize the preliminary editions and for searching for data.

Rebecca Landwehr and Mary Watkins, for their expertise in what kids find interesting and for working some applications for us, occasionally under duress.

Jacque and Bill, for their encouragement and endurance.

J.M.L.
A.E.W.
June 1986

CONTENTS

THE QUANTITATIVE LITERACY PROJECT

There is an excitement today about statistics. Its importance is underscored daily by its frequent use in the media. Statisticians are developing new and simpler techniques. Many states and districts have recently mandated the teaching of statistics. It is now considered to be a fundamental subject in elementary and secondary education.

This book is one of a series of four written by members of the Joint Committee on the Curriculum in Statistics and Probability of the American Statistical Association and the National Council of Teachers of Mathematics. In an effort to introduce the most important and up-to-date topics in statistics into the elementary and secondary curriculum, the Joint Committee initiated the Quantitative Literacy Project. The project, partially funded by the National Science Foundation, included the writing and field-testing of this book and others like it, holding regional conferences for teachers, and developing a videotape that serves as an introduction to the project.

These four books are a result of a collaboration between statisticians and teachers, who have agreed on both the statistical concepts that it is most important for the general public to know and the best ways to teach these concepts. The principles that have guided this collaboration include the following:

1. There is often more than one way to approach problems in statistics and probability. A probability problem can be solved either theoretically or by simulation. It is not unusual for two statisticians to make two different graphs to display the same data. This means that discussion and evaluation of different approaches can take up a large part of class time. It also means that the data may suggest more than one conclusion. Students must be encouraged to attack problems from different angles and to be prepared to support their conclusions.
2. Real data should be used whenever possible in statistics lessons. Real data give the study of statistics both its legitimacy and its excitement. In addition, real data are invariably messy. Values are often missing and are sometimes faulty. Students, who are accustomed to the neatness of the numbers in much of mathematics, need experience in dealing with numbers in the real world.
3. Traditional topics taught in introductory statistics—such as the standard deviation, the normal distribution, hypothesis testing, and Bayes' Theorem and other probability formulas—should be taught *after* the more basic ideas in these four books.
4. The emphasis in teaching statistics should be on good examples and on building intuition, not on showing how to lie with statistics or on probability paradoxes that destroy a student's confidence.
5. Finally, students enjoy and profit from project work, experiments, and other activities designed to give them practical experience in statistics.

ABOUT *EXPLORING DATA*

The student edition serves as an introduction to data analysis. From it students will learn how to make various kinds of graphs, including some that have been developed only recently and are fast becoming widely used. Traditional graphs such as bar graphs and pie charts are not included here because the newer techniques are simpler and quicker to use and the plots are easier to interpret. Students will also learn how to select the appropriate plots for a given set of data. Finally, and most important, they will learn how to examine the plots in order to describe the data, detect patterns in them, and make conjectures about them.

Students will be taught to look at data the way a good statistician does. Surprisingly, this is not at all complicated. A statistician's first step is to try to determine whether the data are reliable. Were they collected in a reasonable manner? Are values missing? Are values in error? Are they the right data for the question? The next step in this statistical analysis is to display the data in appropriate plots. The statistician is likely to use one or more of the plots taught in this book. Finally, the statistician examines the plots and tries to make some sense of the data. After learning the techniques given here, students should be able to analyze the data that they come across in the media and in their own work.

All of the sets of data in the student edition are real and have been selected because they are interesting to students. Students' interest in the data should make them want to explore the data, to argue about them, and to ask questions about them.

How to Use the Book

The major goal of the student edition is to help students learn how to interpret data by using various kinds of plots and graphs. A secondary goal is to teach students to make these plots and graphs. The fact that the interpretations, not the techniques, are the focus may seem strange to many students. In a mathematics class, students are used to getting full credit for a problem if they get the one right answer. That is not the approach in this book. Two students may write entirely different descriptions and yet each may get full credit. We have found that mathematically talented students have the most trouble adjusting to the fact that their grades will be based, not on whether they make plots without any mistakes, but on whether they write good descriptions of what the plots reveal. Those students who do not usually do well in their mathematics classes often accept this idea most readily. They may not be very good at computation or at manipulating algebraic expressions, but they can still feel a real sense of accomplishment in mathematics when they show that they are capable of thinking like statisticians.

Field-Testing

The book was field-tested, with careful selection of topics and different pacing, in grades six through thirteen. This wide range was possible because very few mathematics skills are prerequisites. The students themselves provide their own level of sophistication in the way they approach the data.

In the field tests, the book was used most successfully in four situations:

1. as a unit in a junior high school mathematics class.
2. as a supplement to a traditional text in a one-semester high school statistics course.

3. as a unit in a high school general math course.
4. when combined with the other three books in the Quantitative Literacy Series in a one-semester high school course.

As we have said, students who do not usually do well in their mathematics classes were the most enthusiastic about this book. As one wrote, "I feel I will probably be able to actually use some of this knowledge in real life situations, whereas most math that I am learning now seems to be fairly unnecessary for real life."

Some teachers met with resistance from honors students who wanted to push ahead with the "regular" math curriculum. As one such student wrote, "It was just busy work for us because they didn't want us to be ahead of the rest of our class." It is important to convince students like this of the central role of statistics in modern life, including its importance in many different professional careers.

Teaching Methods

Each section consists of introductory material that is followed by various applications. Typically, a class works through the introductory material together, learning the techniques and talking about the discussion questions. Then students are assigned to work on selected applications, either individually or in small groups. Some students are capable of understanding the introductory material on their own, but class discussions are generally more beneficial and more fun.

The most successful teachers are those who get the students so involved with the material that they ask questions about both the data and the techniques. The confirmation that you are doing a good job of teaching statistics comes when students ask questions that *you* cannot answer! Send them to others or to the library to find their own answers.

In the field tests, some students thought that the work was trivial because the plots were easy for them to construct. If your students have this reaction, challenge them to interpret the data *before* making their plots or reading the questions in the application. Then have them work through the problems, write perceptive interpretations, and defend or modify their initial opinions. Finally, have them do projects using their own data.

How long it takes you to cover the material in the book depends on which sections you select. It usually takes from three to nine weeks.

Helping Students Write Interpretations

Writing interpretations of statistical data is hard at first for almost all students. This fact should not be surprising because most students have never done anything like it before. Some teachers have had success in getting them started by putting a plot on the board or on an overhead projector and asking the class to make observations about it. These can be simple, "The smallest number is 17," or more insightful, "I'll bet there's a gap there because of the baseball strike a few years ago." You can write these comments on the board, and then you can help the class organize them into a paragraph or two. Emphasize the fact that there are no unique, correct answers. In fact, students should try to ask questions that they may not be able to answer about the data, such as "Why is the value for Missouri so big?"

With today's emphasis on writing across the curriculum, it is important for teachers from all fields to help students improve their writing skills. The English teachers at your school may be able to give you some pointers. Two books published by organizations of English teachers are:

Fulwiler, Toby, and Art Young, eds., *Language Connections: Writing and Reading across the Curriculum.* Urbana, IL: National Council of Teachers of English, 1982.

Walvoord, Barbara E. Fassler, *Helping Students Write Well: A Guide for Teachers in All Disciplines.* New York: Modern Language Association, 1982.

Using a Calculator

The student edition requires very little tedious computation. Nevertheless, we suggest that students be allowed to use a calculator whenever they wish. Their attention should not be distracted from exploring the data by a need to work out computations using pencil and paper.

Using a Computer

A computer is not required, but a computer is clearly useful for reducing some of the work involved in constructing the plots. Furthermore, once the data are entered into the computer, a good statistical computer package makes it easy to construct alternate plots, and this helps to improve the data analysis. None of the packages, however, does the really important job of *interpreting* the results, and that is what this book will help your students learn to do.

A number of statistical packages that are available for various personal computers will construct some of the plots presented here. One excellent package that is widely used in introductory college statistics courses is Minitab. If a computer and an appropriate statistical package are available, we suggest that you have students use it *after* they have first worked through a few applications by hand. This enables the students to learn the methods well before the computer takes over some of the dirty work.

As part of the Quantitative Literacy Project, a special computer package that will be keyed to all four books in the series is being developed. Contact Dale Seymour Publications for ordering information.

Using Graph Paper

Students should use graph paper to make their plots, including line plots, stem-and-leaf plots, and box plots. The graph paper helps students make accurate number lines and helps them keep the numbers in stem-and-leaf plots lined up vertically.

Using Outside References

Often the data bring up more questions than they answer. Students then find that they need to do some outside research in almanacs, encyclopedias, or other reference books to do a good job of examining the data. (When was that baseball strike?)

At first, they may resent this outside work, especially if they are in a mathematics class. However, this resentment often goes away quickly if you point out the detective aspect of the research.

The Technicalities of Making Plots

Stem-and-leaf plots, box plots, fitting a line, and smoothing are new techniques that were developed during the 1960s and 1970s. Consequently, they are still in a state of change. Such things as whether a fitted line should be called a *Tukey line,* a *median-fit line,* a *resistant line,* a *robust line,* or some other name has simply not yet been universally agreed upon. The techniques of making the plots themselves are not yet set in concrete either. For example, some people make box plots horizontally; some make them vertically. Some people put a dent in their box plot at the median; others draw a line across the box at the median. Students like to be told that the techniques of making these plots have not yet solidified. This means that they may invent a slightly better way of doing things. Encourage variations. Because not everyone agrees on the "correct" way to make these plots, there is no reason for a class to get caught up in the technicalities of making plots exactly the way this book does.

This Data or These Data?

Traditionally the word *data* is plural. Thus we should say, "These data are interesting," rather than, "This data is interesting." However, there is flexibility here as well and some people prefer the singular. We have tried to use the plural consistently, but you may correctly use the word either way.

Emphasis on the Median

Those of you who have taught statistics before may be surprised at the emphasis on the median. Many of the techniques involve the median, or middle value, in a set of numbers, rather than the mean, or average. There are two main reasons for this emphasis. First, the median is a simpler idea and requires less computation. Second, the median is not affected by a few extremely large or extremely small values as is the mean or average. For these reasons, statisticians often prefer the median in data analysis.

The mode is not used at all because it is generally not useful for interpreting and summarizing data. There are several reasons for this. First, in many sets of data, such as 2, 4, 5, 7, and 9, there is no mode. In contrast, the median and mean are always defined. Second, the mode is unstable. For example, the mode of 1, 1, 3, 5, 8, and 9 is 1, but if two values are changed slightly—say to 1, 2, 3, 5, 9, and 9—the mode becomes 9. Finally, as shown in the last example, the mode does not necessarily indicate the center of the data.

What Sections to Cover

With several exceptions, the sections of the student edition are independent of one another, and you may select the ones most likely to interest your students. The exceptions are that Section III, "Median, Mean, Quartiles, and Outliers," and Section VI, "Scatter Plots," must be completed before any later sections can be done. Also, in order to do the review sections, the students must have completed each of the previous sections.

It is important to go over all of the introductory material in each section. However, it is not necessary for a student to complete all applications. For example, depending on his or her interest, a student can complete either Application 1, "Rock Albums," or Application 2, "Causes of Death," or both, in Section I: Line Plots. Applications that can be omitted are identified in this Teacher's Edition.

BIBLIOGRAPHY

Denby, L., and J. M. Landwehr. "Examining Data: A General Strategy and One-Sample Methods" (Unit 629). In *UMAP Modules 1983.* Lexington, MA: Consortium for Mathematics and Its Applications, Inc., 1983.

Discusses overall strategy, questions, and methods to keep in mind when analyzing data.

Ehrenberg, A. S. C. *A Primer in Data Reduction: An Introductory Statistics Textbook.* New York: John Wiley, 1982.

For the teacher who would like to learn more about communicating data.

Freedman, David, Robert Pisani, and Roger Purves. *Statistics.* New York: W. W. Norton, 1978.

An excellent college-level introductory textbook that can be used with high school students with good verbal skills.

Hoaglin, David C., Frederick Mosteller, and John W. Tukey, eds. *Understanding Robust and Exploratory Data Analysis.* New York: John Wiley, 1983.

An advanced book giving rationale, historical and conceptual development, and mathematical support for the techniques of data analysis; also relates these techniques to classical statistical theory.

Hoffer, Alan. *Statistics and Information Organization.* Palo Alto, CA: Creative Publications, 1978.

The largest single source of reproducible classroom materials on descriptive statistics.

Huff, Darrell. *How to Lie with Statistics.* New York: W. W. Norton, 1954.

Entertaining reading for the student who would like to see how data can be misrepresented in charts and graphs.

Shulte, Albert P., and James R. Smart, eds. *Teaching Statistics and Probability.* 1981 Yearbook of the National Council of Teachers of Mathematics. Reston, VA: NCTM, 1981.

A collection of articles, most of which describe classroom activities.

Tufte, Edward R. *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press. 1983.

A beautiful book on excellence in graphing. Contains a selection of the best statistical graphics ever drawn.

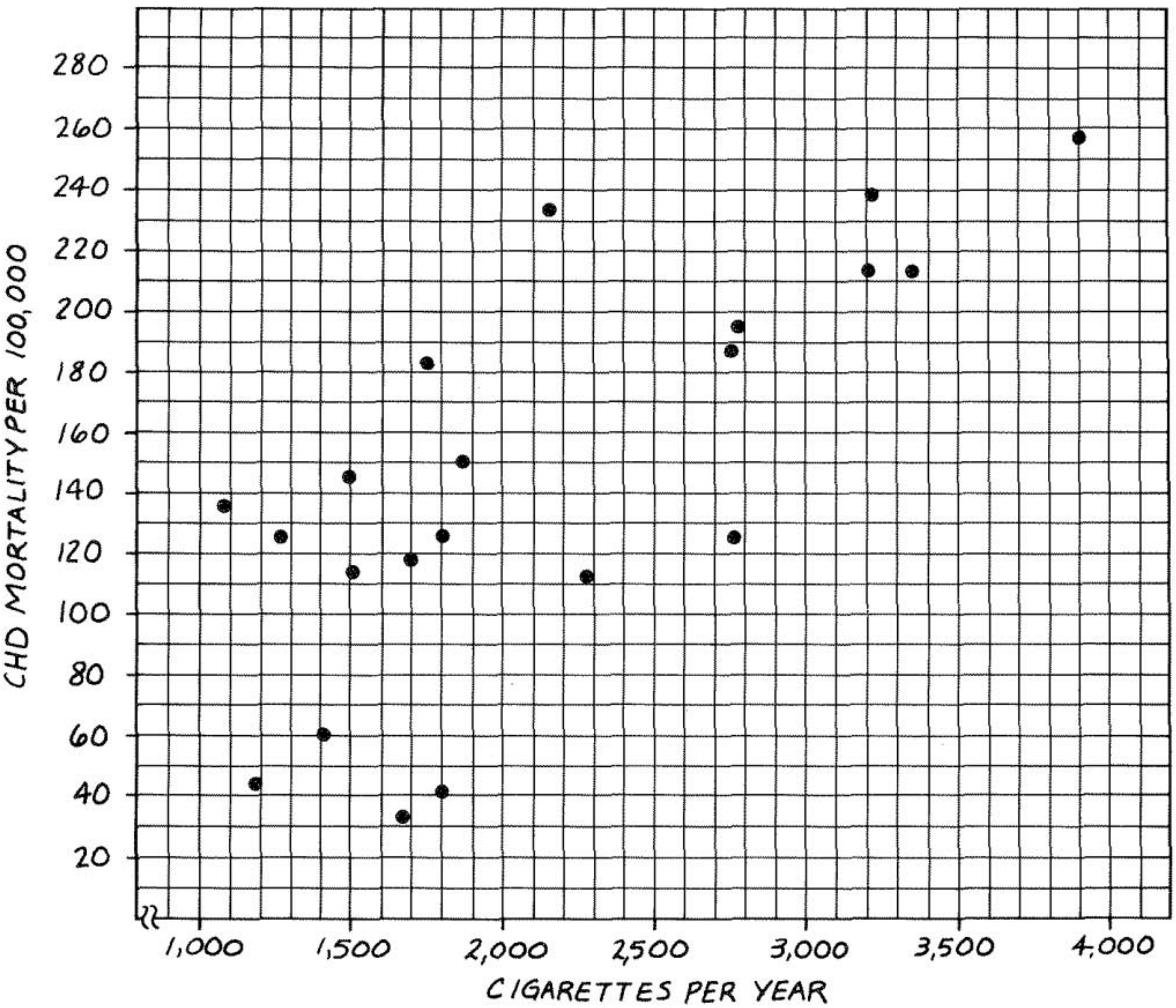
Velleman, Paul F., and David C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis.* Boston: Duxbury, 1981.

A more advanced explanation that includes the techniques presented in this book.

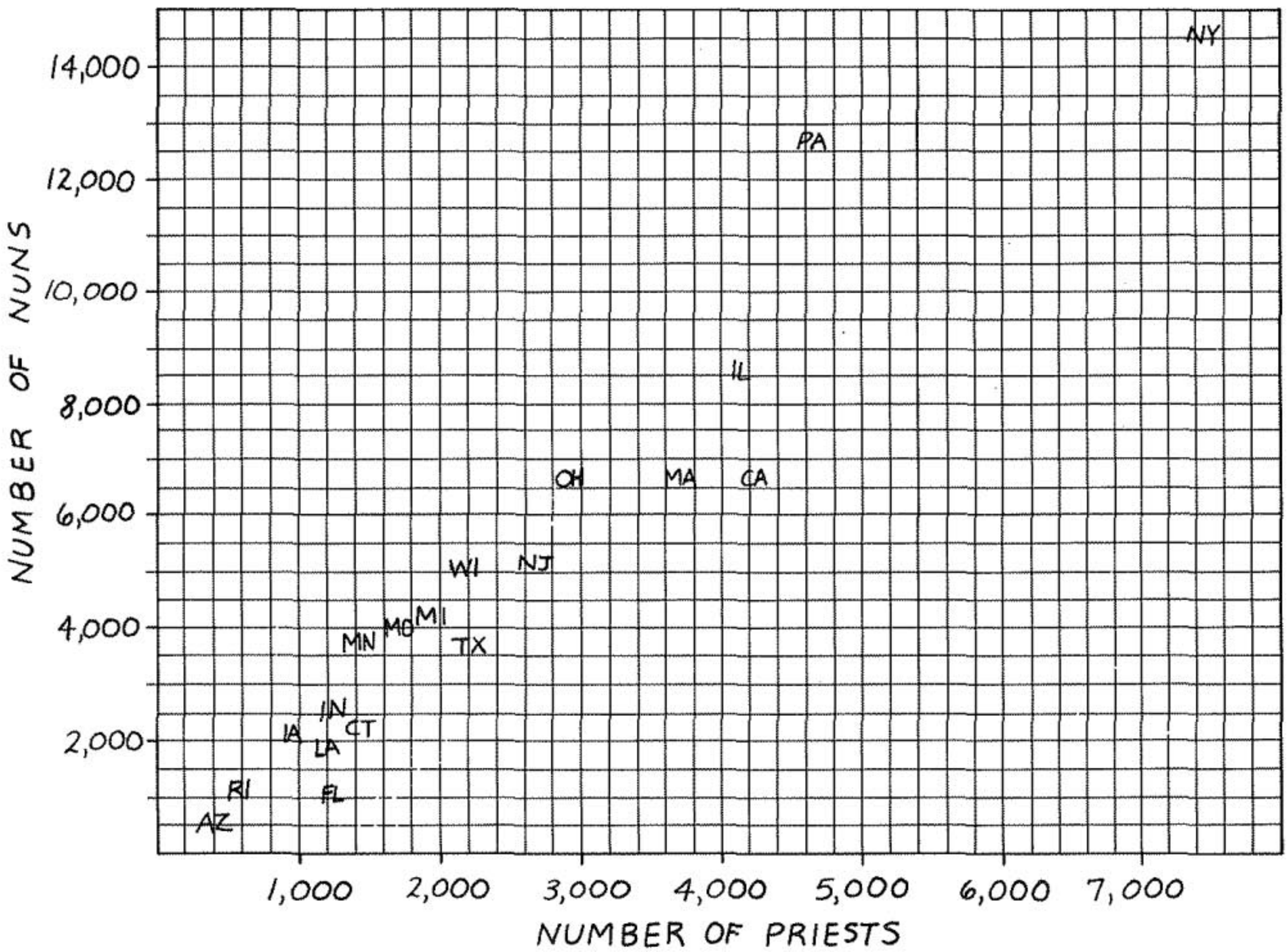
OPTIONAL GRAPHS

The graphs on the following pages are the same as the ones in the answers, except that the fitted line (or smoothed trend over time) is missing. We have included them here so that you can copy them and pass them out if you don't want the students to take the time to make the scatter plots themselves.

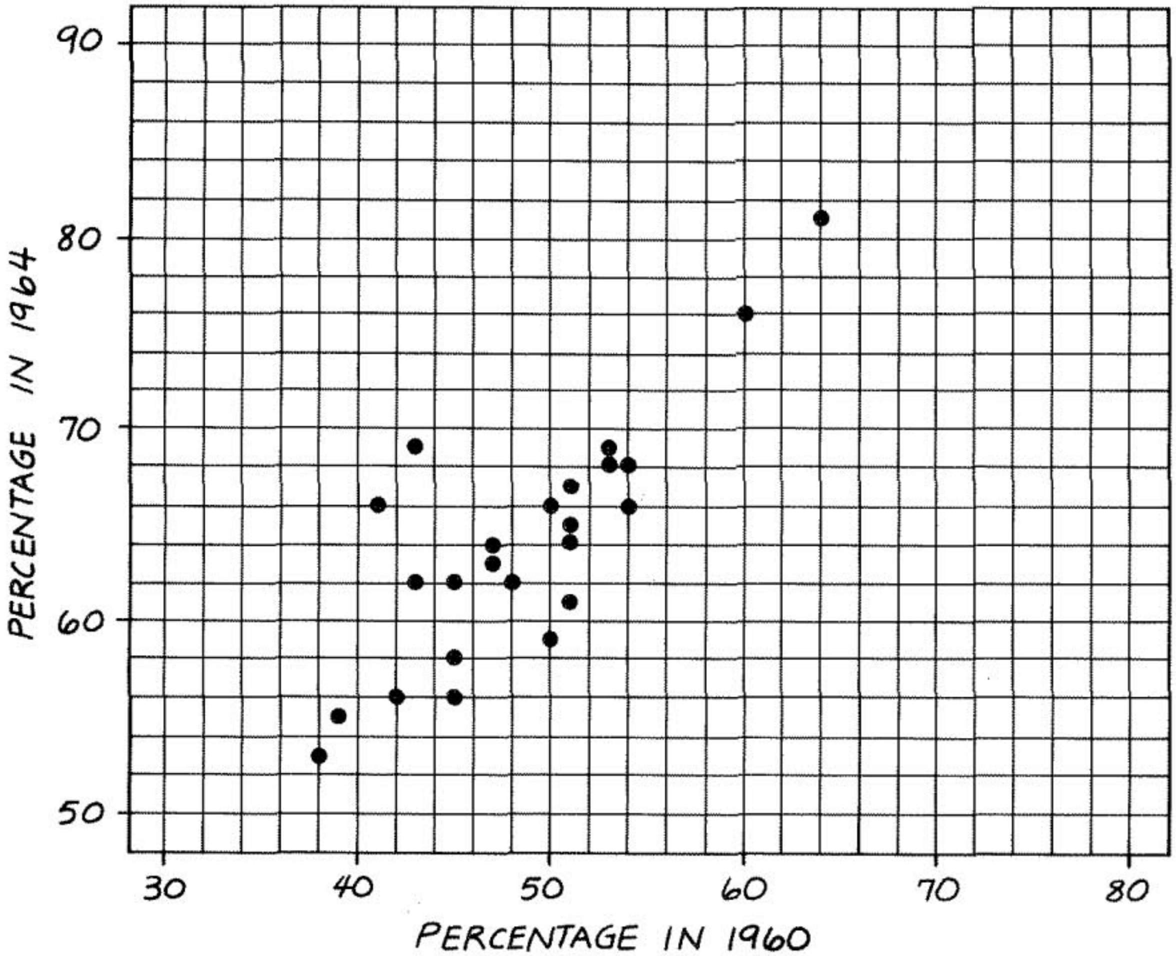
For use with Application 28
Pages 119–120



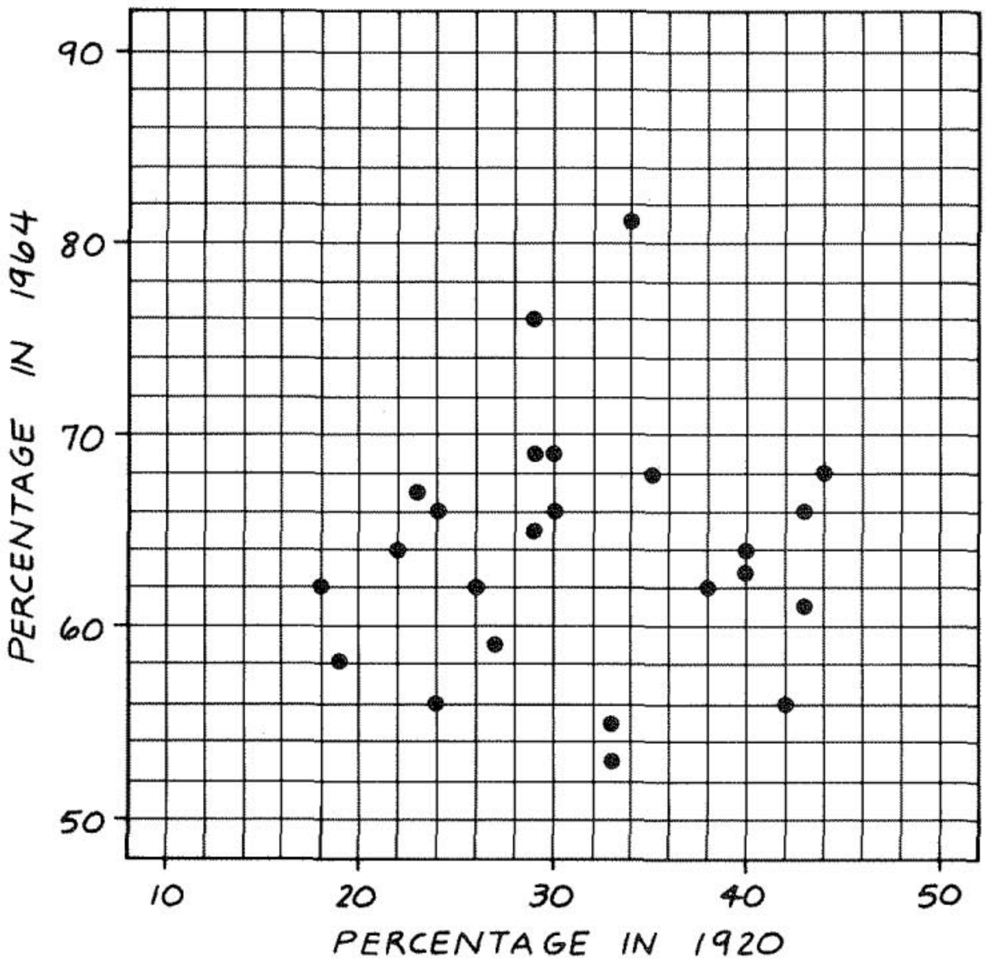
For use with Application 29
Pages 121–122



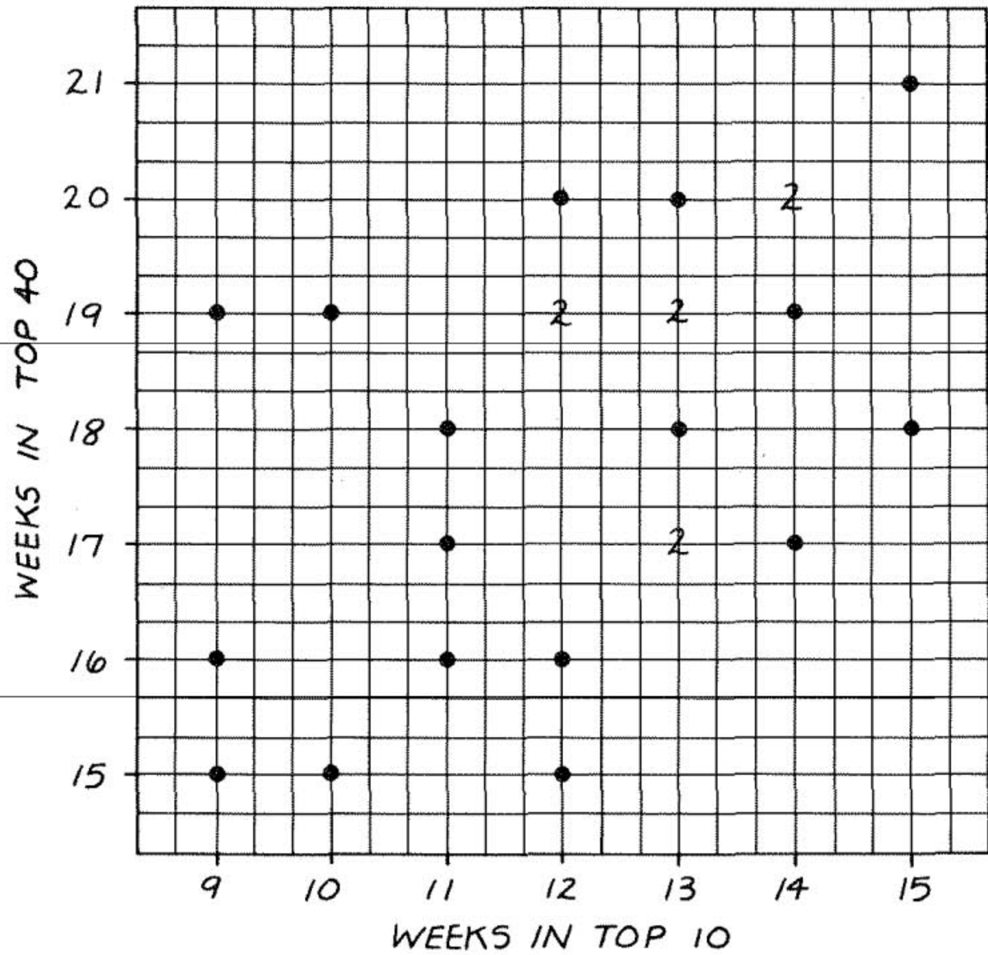
For use with Application 30
Pages 123–124



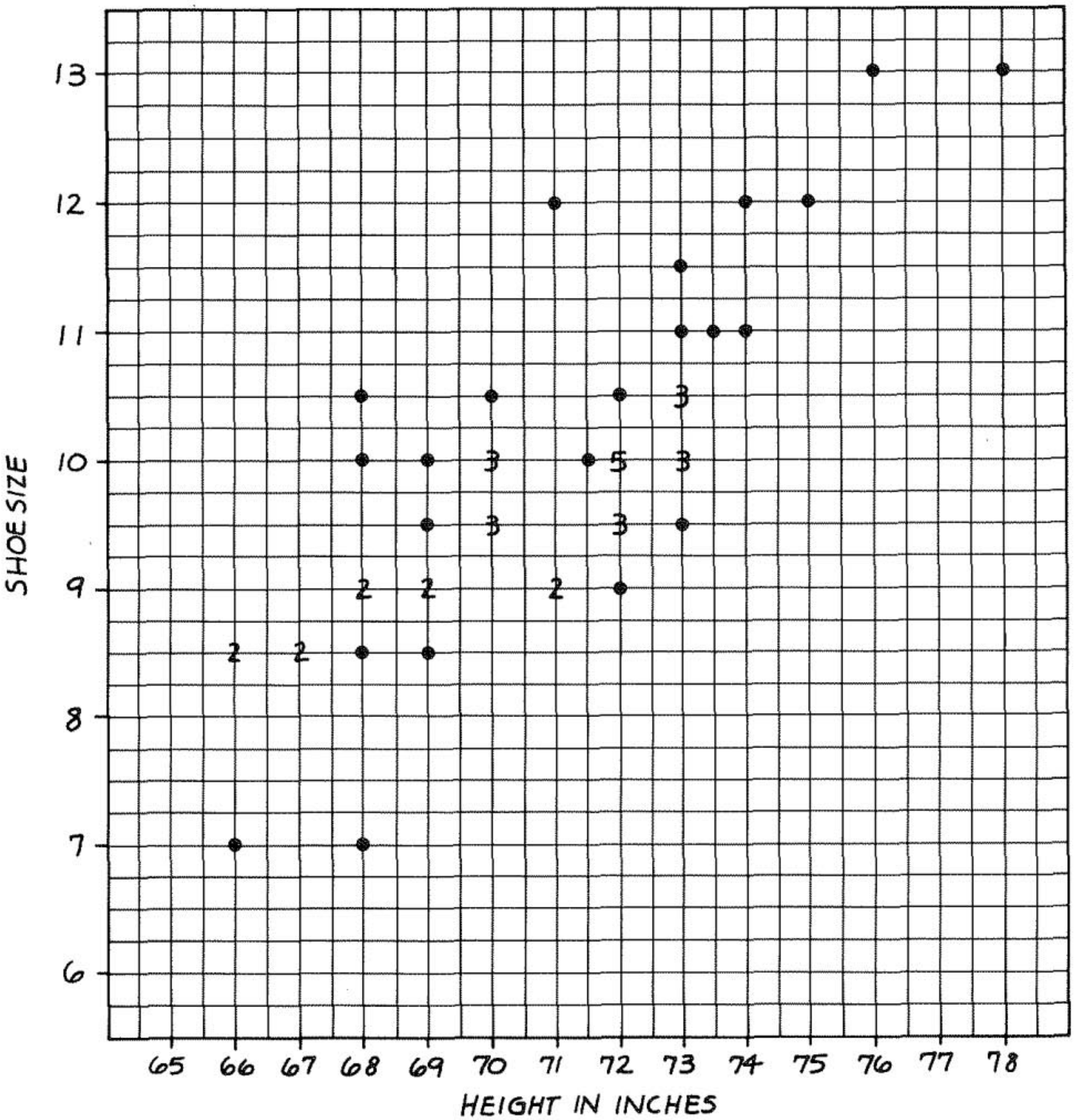
For use with Application 30
Pages 123–124



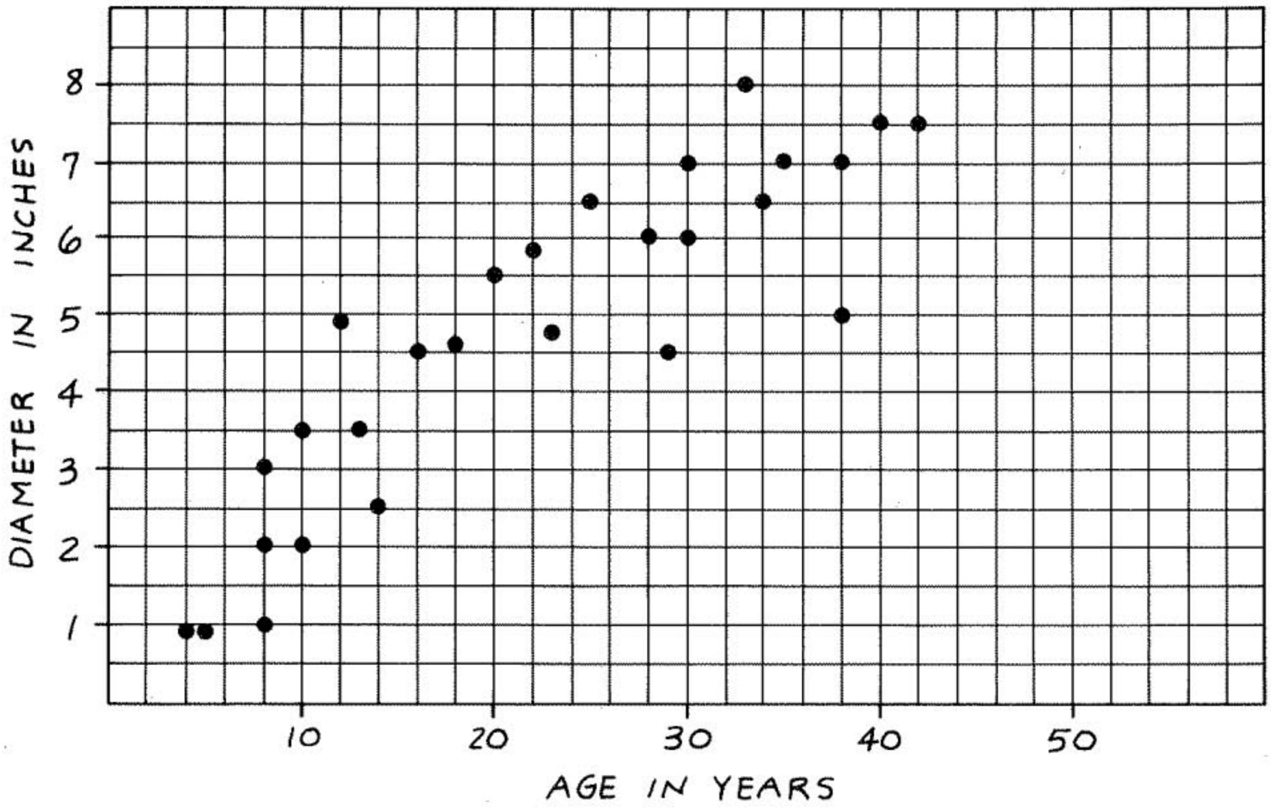
For use with Application 31
Pages 130-131



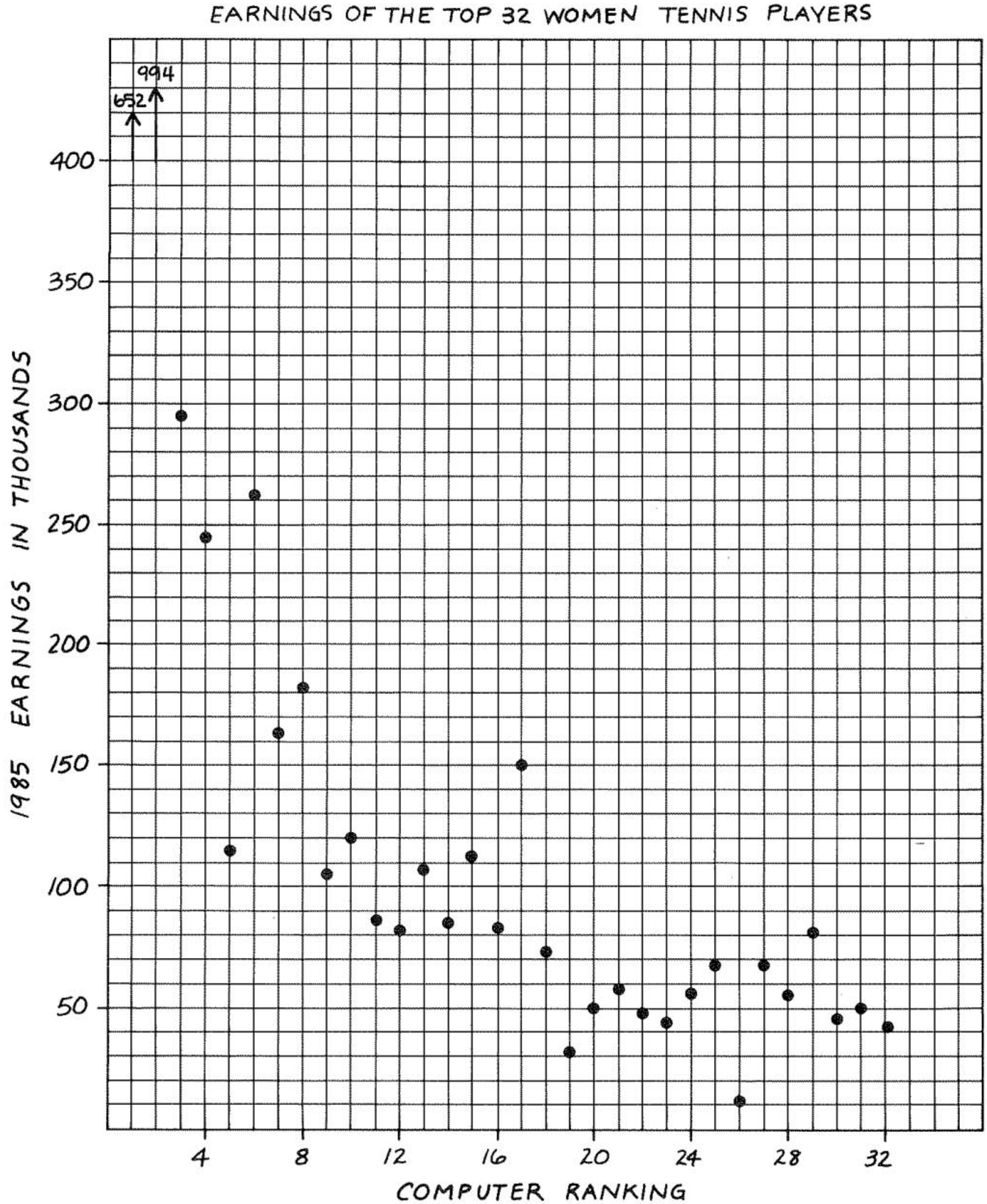
For use with Application 32
Pages 132–133



For use with Application 34
Pages 136–137

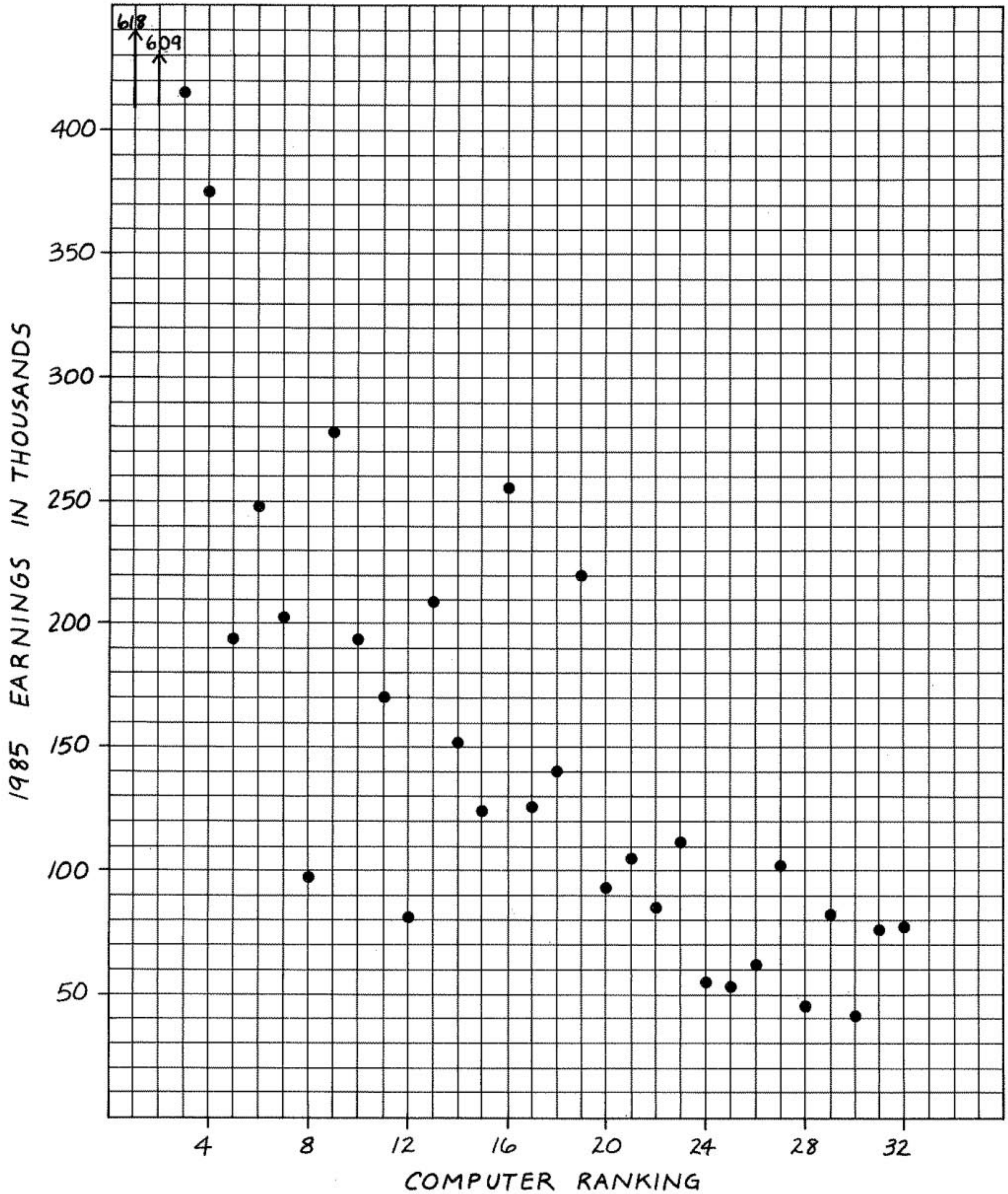


For use with Application 37
Pages 147-149



For use with Application 37
Pages 147–149

EARNINGS OF THE TOP 32 MEN TENNIS PLAYERS



QUIZZES

The following pages contain reproducible quizzes for the sections that introduce new material. There are no quizzes for the two review sections because you can use the applications themselves as end-of-unit projects or tests.

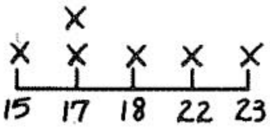
As in the applications, the quizzes contain problems such as, "Write a description of the information displayed in the plot." Remind students that the more complete and organized their descriptions are, the more points they will receive. Answers that are one sentence long will not receive full credit.

We suggest that you let students use calculators when they take all of these quizzes so that computation will not distract them from the statistics.

The answers to the quizzes appear immediately following the quizzes.

QUIZ ON LINE PLOTS

1. The following line plot is incorrect. Make the correct plot.



2. The following table gives the number of visits, *in millions*, to the most popular National Park Service Recreation Areas in 1980:

   a. How many people visited Olympic National Park in 1980?
   b. Which area was visited by the most people?
   c. Make a line plot of these data by rounding each number to the nearest million.
   d. Write a description of the information displayed in your line plot.

National Park Service Recreation Areas	Number of Visitors in Millions
Blue Ridge Parkway, Ga., N.C., Va.	16.7
Cape Hatteras National Seashore, N.C.	1.7
Chickamauga and Chattanooga National Military Park, Ga., Tenn.	14.2
Colonial National Historical Park, Va.	9.1
Death Valley National Monument, Calif.	0.6
Gateway National Recreation Area, N.J., N.Y.	9.4
Glacier National Park, Mont.	1.5
Glen Canyon National Recreation Area, Ariz., Utah	1.7
Golden Gate National Recreation Area, Calif.	18.4
Grand Canyon National Park, Ariz.	2.5
Grand Teton National Park, Wyo.	3.5
Great Smoky Mountains National Park, N.C., Tenn.	11.9
Hot Springs National Park, Ark.	5.3
Indiana Dunes National Lakeshore, Ind.	1.6
Kennesaw Mountain National Battlefield Park, Ga.	6.7
Kings Canyon National Park, Calif.	0.8
Lake Mead National Recreation Area, Ariz., Nev.	5.2
Natchez Trace Parkway, Miss., Tenn., Ala.	15.9
Olympic National Park, Wash.	2.5
Ozark National Scenic Riverways, Mo.	1.8
Rocky Mountain National Park, Colo.	2.6
Sequoia National Park, Calif.	0.9
Shenandoah National Park, Va.	1.8
Valley Forge National Historical Park, Pa.	11.5
Yellowstone National Park, Idaho, Mont., Wyo.	2.0
Yosemite National Park, Calif.	2.6

Source: *Statistical Abstract of the United States*, 1981.

QUIZ ON STEM-AND-LEAF PLOTS

1. The following table gives the number of years that a person born in 1981 could expect to live at the time of his or her birth for countries with 10 million or more population in the Americas and in Europe.

   a. In which European country is the life expectancy the longest?
   b. In which country in the Americas is the life expectancy the longest?
   c. Make a back-to-back stem-and-leaf plot of the countries in the Americas and in Europe. Decide how to spread it out. Be sure to include an explanation such as

   "7 | 3 represents. . ."

   d. Write a description of what you learned from your stem-and-leaf plot.

Americas		Europe	
Country	Life Expectancy at Birth (years)	Country	Life Expectancy at Birth (years)
Argentina	65	Czechoslovakia	71
Brazil	60	France	73
Canada	73	Germany, East	72
Chile	62	Germany, West	72
Colombia	59	Hungary	70
Mexico	60	Italy	73
Peru	55	Netherlands	75
United States	73	Poland	71
Venezuela	63	Romania	70
		Soviet Union	70
		Spain	73
		United Kingdom	73
		Yugoslavia	70

Source: *Statistical Abstract of the United States*, 1981, p. 871.

2. If an explanation on a stem-and-leaf plot is given as

   4 | 6 represents 460 to 469,

   the data have been truncated. Write the explanation that would be given if these same data had been rounded.

3. Suppose the data consist of 27 values between 42 and 99. To construct a stem-and-leaf plot, would you look through the data to find all the values in the 40s to fill in that row on the plot; then find all the values to fill in the 50s row, and so on? Why or why not?

(Continued from page 20)

4. Here is a stem-and-leaf plot of the amount of vitamin X in servings of fish (F), meats (M), vegetables (V), and starches (S). The leaves are in order.

0	S V F
1	V V V M
2	M S S M M S
3	F F M F
4	F

a. Which type of food is generally highest in vitamin X?
b. Which type of food is generally lowest in vitamin X?
c. True or false: Every meat item has more vitamin X per serving than any vegetable item.
d. True or false: Every fish item has more vitamin X per serving than any meat item.
e. Which type of food varies the most in the amount of vitamin X it contains?
f. What is the shape of the distribution?
g. Which two items appear to be the most different compared to the other items of their own type?

5. The following plot shows the lifetimes of several Brand A and Brand B batteries.

BRAND A		BRAND B
	1	
	·	7
	2	2
	·	6
2 1 1 0	3	
9 9 7 7 5	·	
3 2 2 1	4	2 2 3 3 3 4
	·	5 6 8 8 9
4	5	0
5	·	

|4| 2 REPRESENTS
420 - 429 HOURS

a. What is the longest that any battery lasted?
b. Are the data truncated, rounded, or is there not enough information to tell?
c. If you want to maximize your chances of getting a battery that will last at least 300 hours, which brand should you choose?
d. Which typically lasts longer, a Brand A battery or a Brand B battery?
e. If you want to maximize your chances of getting a battery that lasts more than 500 hours, which brand should you choose?
f. True or false: To show the comparisons more clearly, you should spread this plot out more.
g. Give a reason someone might prefer a Brand A battery.
h. Give a reason someone might prefer a Brand B battery.

QUIZ ON MEDIAN, MEAN, QUARTILES, AND OUTLIERS

1. The *Statistical Abstract of the United States* (1981, page 232) gives the median size of a home garden as 663 square feet.
   a. Explain the meaning of this statement.
   b. Explain why the median is used instead of the mean.

2. The following table gives the number of pounds of cotton produced per acre in the major cotton-producing states in 1980.
   a. Find the median number of pounds.
   b. Find the upper quartile.
   c. Find the lower quartile.
   d. Find the interquartile range.
   e. Use the 1.5 × IQR rule to find any outliers. Show your work.

State	Pounds per Acre
Alabama	411
Arizona	1,085
Arkansas	330
California	995
Georgia	258
Louisiana	390
Mississippi	488
Missouri	353
New Mexico	430
North Carolina	381
Oklahoma	174
South Carolina	309
Tennessee	349
Texas	234

Source: *Statistical Abstract of the United States*, 1981, p. 691.

3. Find the mean number of letters in the following words:

   MONDAY TUESDAY WEDNESDAY THURSDAY FRIDAY SATURDAY SUNDAY

   Do not round your answer. Leave it as a mixed number or decimal.

4. True or false: If there are only three data values, the median must equal the mean.

5. a. True or false: The upper quartile is always larger than or equal to the median.
   b. True or false: The upper quartile is always larger than or equal to the mean.

6. For summarizing a distribution of incomes by a single number, which is generally better to use, the median or the mean? Why?

(Continued from page 22)

7. A data set has lower extreme = 18, lower quartile = 30, median = 37, upper quartile = 40, mean = 42, and upper extreme = 70. Using the $1.5 \times$ IQR rule, tell whether each of the following observations is an outlier.
   a. 18
   b. 24
   c. 53
   d. 60
   e. 70
8. If a distribution is mound-shaped except for one outlier at the upper extreme, would you expect the mean to be larger, about the same, or smaller than the median? Explain.
9. A data set contains five observations. Four of them are 6, 12, 12, and 14. Find the fifth observation so that the median of all five equals the mean of all five. (*Hint:* Consider a line plot of the four given numbers, and then see how the median depends on the fifth number.)

QUIZ ON BOX PLOTS

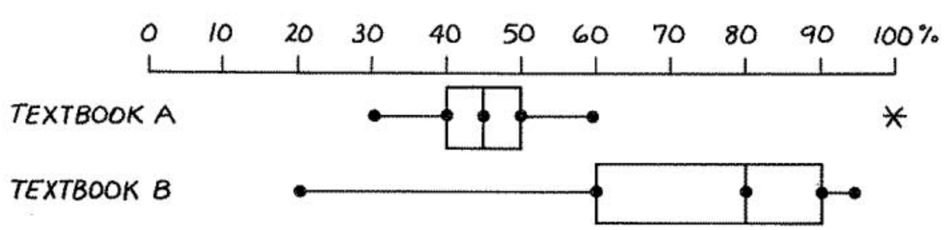
1. The following table shows the number of international passengers *in thousands* that departed from U.S. airports in 1980 for each country listed.

Country	Number of Departures in Thousands	Country	Number of Departures in Thousands
Bahamas, The	1,006	Italy	495
Bermuda	467	Jamaica	382
Brazil	291	Japan	1,602
Colombia	299	Mexico	2,886
Denmark	254	Netherlands	409
Dominican Republic	443	Netherlands Antilles	282
France	635	Spain	273
Germany, Fed. Rep. of	1,178	Switzerland	306
Greece	190	United Kingdom	2,840
Ireland	212	Venezuela	518

Source: *Statistical Abstract of the United States*, 1981, p. 240.

a. How many passengers departed for Denmark?
b. To which country did the largest number of passengers go?
c. Write the numbers of passengers from smallest to largest.
d. Find:
The lower extreme.
The upper extreme.
The median.
The lower quartile.
The upper quartile.
e. Determine whether there are any outliers. If so, which countries are outliers? Show your work.
f. Make a box plot, using *'s for any outliers.
g. Write a summary of the information displayed in your box plot.
h. (Bonus) One major foreign destination of U.S. travelers is not included in the table. Which country is this?

2. The following box plot shows the final exam scores in algebra for students using two different textbooks.



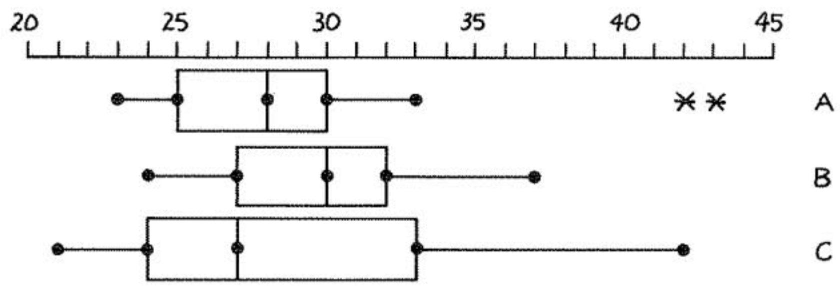
a. What was the lowest score for a student using Textbook A?
b. What proportion of the students using Textbook A got less than 50 percent?

(Continued from page 24)

c. Complete this sentence: Half of the students using Textbook B got ________ percent or more on the final exam.
d. Which textbook gave student scores that varied less? Explain your answer.
e. Which textbook do you think is better? Explain your answer.

3. For each of the following, decide whether a box plot or a stem-and-leaf plot would be more useful. Then write a sentence giving the reason for your choice.
   a. Showing clusters and gaps in the data.
   b. Comparing four groups of data.
   c. Comparing one data set with 150 values to another data set with 37 values.
   d. Presenting a plot to someone who wants to compute the mean.
   e. Judging whether the middle 50 percent of one data set is spread over a wider range than the middle 50 percent of a second data set.
   f. Emphasizing the median and the quartiles.
   g. Comparing a data set with 11 values to a second one with 9 values.

4. Here are box plots of the miles-per-gallon achieved by all the different car models made by three manufacturers, A, B, and C.



   a. If we compare manufacturers by looking at just the car with the very highest miles-per-gallon, which manufacturer does the best?
   b. If we compare manufacturers by looking at just the highest 25 percent of all their cars, which manufacturer does the best?
   c. If we compare manufacturers by looking at just the median miles-per-gallon, which manufacturer does the best?
   d. Which manufacturer makes cars whose miles-per-gallon varies least?
   e. Suppose you work for manufacturer C and you want to improve your miles-per-gallon compared to A and B. Should you put extra effort into improving your cars with the most miles-per-gallon, improving your cars with the fewest miles-per-gallon, or should you spread your extra effort over all the cars? Explain your answer.
   f. True or false: For manufacturer C, the median is not in the center of the box because there are more models above the median than below it.

5. From which of the following plots can you determine how many values are in the data set?
   a. line plot
   b. stem-and-leaf plot
   c. box plot

QUIZ ON SCATTER PLOTS

1. The following is a list of 22 Los Angeles high schools that reported the percentage of students in yearbook who were declared ineligible and the percentage of students in girls' track who were declared ineligible.

High School	Percent Ineligible Yearbook	Percent Ineligible Girls' Track
Banning	0	24
Belmont	7	7
Canoga Park	20	33
Chatsworth	33	31
El Camino Real	13	15
Franklin	21	44
Gardena	17	20
Hamilton	0	0
Kennedy	18	20
Lincoln	71	13
Locke	67	57
Los Angeles	39	17
Manual Arts	38	25
Monroe	6	24
San Fernando	24	33
South Gate	5	28
Taft	0	27
University	43	9
Van Nuys	0	7
Verdugo Hills	10	39
Washington	25	14
Westchester	0	0

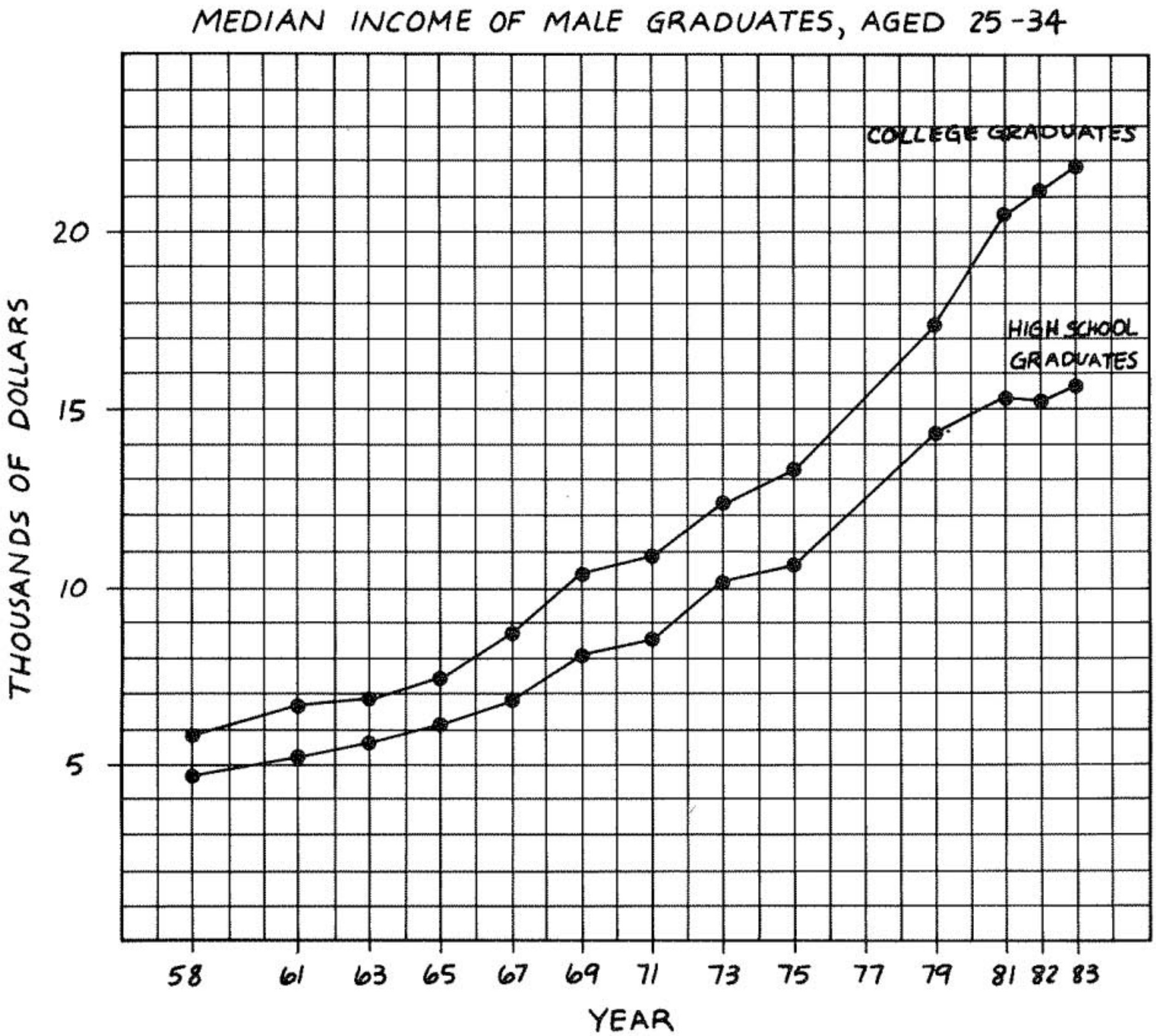
Source: *Los Angeles Times*, May 17, 1983.

a. Make a scatter plot of these data. Put percentage ineligible in yearbook on the horizontal scale. Instead of a dot, plot the first two letters in the high school's name. (Use LA for Los Angeles and use LO for Locke.)
b. Is there a positive association, a negative association, or no association between the percentage ineligible in yearbook and the percentage ineligible in girls' track?
c. Describe any clusters of schools that you find in the plot.
d. Which two high schools stand out on the scatter plot as most unusual? Explain how each is unusual.

2. Decide whether each pair of variables that follows would show a positive association, a negative association, or no association.
   a. A person's height and weight.
   b. An adult's intelligence and age.
   c. The amount of candy eaten and the number of cavities.

(Continued from page 26)

3. The following plot over time gives the median income of male college and high school graduates, 25 to 34 years old, for the years from 1958 to 1983 (in current dollars).



Source: U.S. Census Bureau.

a. Approximately how much did a typical male college graduate aged 25 to 34 earn in 1967?
b. In what year did the median income of high school graduates decrease from the year before?
c. Describe the information that you see in this plot.

QUIZ ON LINES ON SCATTER PLOTS

1. The scores on the first and second tests of the semester are given here for a small class.

Student	First Test	Second Test
Al	19	11
Ann	15	5
Barbara	11	3
Bill	24	14
Diana	14	14
Elizabeth	13	10
Gail	20	20
Jacque	15	9
Jim	24	17
Luis	17	7
Mary	18	14
Neil	6	6
Rebecca	5	1
Richard	17	10
Roberto	10	8
Shirley	14	7

   a. Make a scatter plot of the scores. Put the score for the first test on the horizontal axis.
   b. Fit a line to these points.
   c. Use your line to predict the score on the second test for a student who got a 22 on the first test.
   d. Which student is the farthest vertical distance from the line?
   e. What is this vertical distance?
   f. On your scatter plot, draw in the 45° line and label it as the 45° line.
   g. How many points are on this 45° line?
   h. What does it mean if a point is on this line?
   i. Are more students below the 45° line or above it?
   j. If a student got a higher score on the second test than on the first test, where would the point be?
   k. Write a description of the information given by the plot and its two lines.
   l. (For students who have studied algebra) Find the equation of the fitted line.

2. If a line is to be fitted to 23 points, how many points would ideally be in the center strip?

3. Why do we move the ruler and draw the line only one-third of the way from the two end *X*'s toward the center *X*?

4. True or false: The fitted line is not much affected by outliers.

5. Explain why one would want to fit a line to the data on a scatter plot.

QUIZ ON SMOOTHING

1. The following table gives the number of fine ounces of silver produced in the United States for various years. The numbers are in millions.

Year	Fine Ounces	Smoothed Values
1930	51	
1935	46	
1940	70	
1945	29	
1950	43	
1955	36	
1960	36	
1965	40	
1970	45	
1975	35	
1980	32	

Source: Bureau of Mines.

   a. Make a plot over time of the number of fine ounces produced.
   b. Explain why this plot is a good candidate for smoothing.
   c. Copy and complete the Smoothed Values column.
   d. Make a plot over time of the smoothed values.
   e. Describe the overall trend in silver production in the United States.

2. What happens to outliers after smoothing?

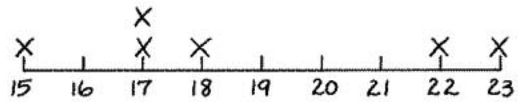
3. Construct an example to show why the rule for smoothing endpoints is often unsatisfactory.

ANSWERS TO QUIZZES

The following pages give the answers to the quizzes. When students are asked to write descriptions or give interpretations, the answers will vary. We have included sample answers that cover the points we expect the students to address.

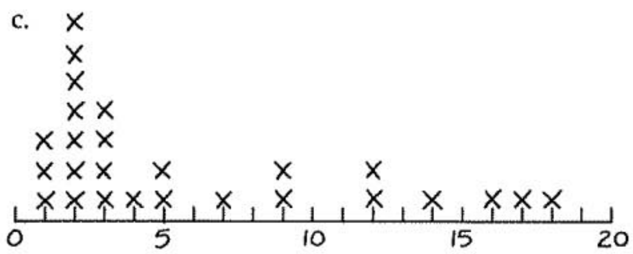
ANSWERS TO QUIZ ON LINE PLOTS

1.



2. a. 2,500,000 or 2.5 million
   b. Golden Gate National Recreation Area

   c.



   d. Answers will vary. Sample: Seventeen of these 26 National Park Service Recreation Areas had 6 million or fewer visitors. The remaining nine were spread out rather evenly between 7 million and 18 million. The most-visited area was Golden Gate National Recreation Area. It seems likely that this area is around the Golden Gate Bridge, which is in San Francisco. Looking down the table, it does not appear that any other of these areas is in a major city, so that could explain in part why its attendance is so high. The next two most-visited areas were Blue Ridge Parkway and Natchez Trace Parkway; these are the only two "parkways" listed. One wonders if these parkways consist mainly of a road, and if that is why their number of visitors is large. The next three most-visited areas, with from 12 to 14 million visitors, were in Georgia, Tennessee, Pennsylvania, and North Carolina. These are all in the east, where the population density is larger than in the west.

   The least-visited area was Death Valley, which may be appropriately named.

ANSWERS TO QUIZ ON STEM-AND-LEAF PLOTS

1. a. Netherlands
   b. United States and Canada
   c.

AMERICAS		EUROPE
	5	
	·	
5	·	
	·	
9	·	
0 0	6	
3 2	·	
5	·	
	·	
	·	
	7	0 0 0 0 1 1
3 3	·	2 2 3 3 3 3
	·	5
	·	
	·	
	8	

7\|3 REPRESENTS
73 YEARS

   d. Answers will vary. Sample: In general, life expectancy is about 10 years more in Europe than it is in the Americas. Life expectancy in Europe varies from 70 years in four countries to 75 years in the Netherlands. In contrast, in the Americas all but two countries, the United States and Canada with 73 years, have life expectancies of from 55 years to 65 years. Thus, even excluding the United States and Canada, life expectancy is more variable among the countries in the Americas than it is in the European countries. In terms of life expectancy, the United States and Canada can be thought of as typical of a European country.

2. 4 | 6 represents 455–464.

3. You should *not* look through the data to find all of the values in the 40s, then all of the values in the 50s, and so on. Instead, you should start at the top of the list of data and enter that value on the plot, then enter the second value in the list on the plot, and so on.

   Filling in all of the leaves for the 40 stem, then for the 50 stem, and so on is more likely to result in mistakes, and it takes much longer.

4. a. fish
   b. vegetables
   c. true
   d. false
   e. fish
   f. mound-shaped or normal
   g. The S on the first row is much smaller than the other three starches, and the F on the first row is much smaller than the other four fish.

5. a. 550 hours (Brand A)
   b. truncated
   c. Brand A
   d. Brand B
   e. Brand A
   f. false

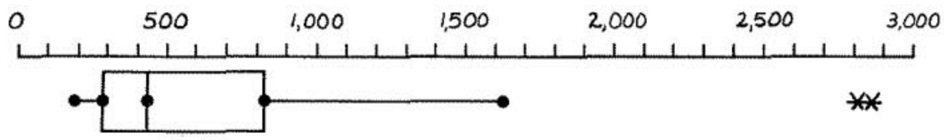
g. Answers will vary. Sample: There is less chance of a quick failure; that is, there is less chance of a failure in less than 300 hours. Also, there is a higher chance of getting a really long-lasting battery, say, more than 500 hours, than there is from Brand B. (Either reason is sufficient.)
h. Answers will vary. Sample: The typical (median) Brand B battery lasts longer than the typical Brand A battery by about 40 hours (430 versus 390). Also, if you don't get one of the few that fail early, all the Brand B batteries lasted at least 400 hours, and that is longer than the median life observed for Brand A batteries. (Either reason is sufficient.)

ANSWERS TO QUIZ ON MEDIAN, MEAN, QUARTILES, AND OUTLIERS

1. a. Half of the home gardens are larger than 663 square feet and half are smaller.
   b. Answers will vary. Sample: Some home gardens are very large. The median sized garden of 663 square feet could be 17 feet by 39 feet. Many gardens are much larger than this—thousands of square feet larger. Of course, there can't be any gardens that are thousands of square feet smaller. These large gardens will, like large incomes, make the mean larger than the size of most home gardens.
2. a. 367
   b. 430
   c. 309
   d. $430 - 309 = 121$
   e. $1.5 \times IQR = 1.5\,(121) = 181.5$
   $LQ - 1.5 \times IQR = 309 - 181.5 = 127.5$
   $UQ + 1.5 \times IQR = 430 + 181.5 = 611.5$
   The outliers are California and Arizona.
3. $\frac{6+7+9+8+6+8+6}{7} = \frac{50}{7} = 7\frac{1}{7}$
4. False. An example is 3, 5, 10. The median is 5. The mean is 6.
5. a. true
   b. false
6. The median. With incomes, there are likely to be a few very large values, and these can make the mean not at all representative of the distribution, whereas a few such values would have no effect on the median.
7. a. not an outlier
   b. not an outlier
   c. not an outlier
   d. outlier
   e. outlier
8. Larger. Without the outlier, we would expect the median and mean of a mound-shaped distribution to be about the same. The outlier would not change the median much, but it would increase the mean.
9. 16

ANSWERS TO QUIZ ON BOX PLOTS

1. a. 254,000
   b. Mexico
   c. 190, 212, 254, 273, 282, 291, 299
   306, 382, 409, 443, 467, 495, 518
   635, 1,006, 1,178, 1,602, 2,840, 2,886
   d. 190, the lower extreme; 2,886, the upper extreme; (409 + 443)/2 = 426, the median; (282 + 291)/2 = 286.5, the lower quartile; (635 + 1,006)/2 = 820.5, the upper quartile.
   e. The interquartile range is 820.5 – 286.5 = 534. Consequently, 1.5 interquartile ranges above the upper quartile is 1621.5 and 1.5 interquartile ranges below the lower quartile is less than 0. Thus the two outliers are Mexico and the United Kingdom.
   f.



   g. Answers will vary. Sample: The box plot shows the number of airline passengers that departed for twenty countries from U.S. airports in 1980. Two countries had a much larger number of passengers than the others. They were Mexico with 2,886,000 and the United Kingdom with 2,840,000. The remaining countries had between approximately 200,000 and 1,600,000 passengers. The median number of passengers departing was 426,000, so the bottom ten countries are close together in number of people who traveled there by air.
   These data include both U.S. citizens and citizens of other countries. It would be interesting to know how many of the passengers were U.S. citizens and how many weren't. It would also be interesting to know why the number for Germany is so much higher than the ones for France or Italy.
   h. Canada

2. a. 30 percent
   b. 3/4 or 75 percent
   c. 80
   d. Textbook A, because half of those students scored within a range of 10 points, at 40 to 50. In contrast, for Textbook B, half scored within the wider range of 28 points, from 60 to 88.
   e. Answers will vary. Sample: Textbook B was better, even though the highest score was earned by a student who used Textbook A and the lowest score was earned by a student who used Textbook B. Three-quarters of the students who used Textbook B got 60 percent or more, while all but one of the students who used Textbook A got 60 percent or less.

3. a. Stem-and-leaf plot. Clusters and gaps cannot be seen in a box plot.
   b. Box plot. They can be put next to each other easily for comparisons.
   c. Box plot. When the sizes of the two data sets are so different, stem-and-leaf plots are not so useful, but this difference in size does not cause a problem for box plots.

(Answers to quiz on box plots continued)

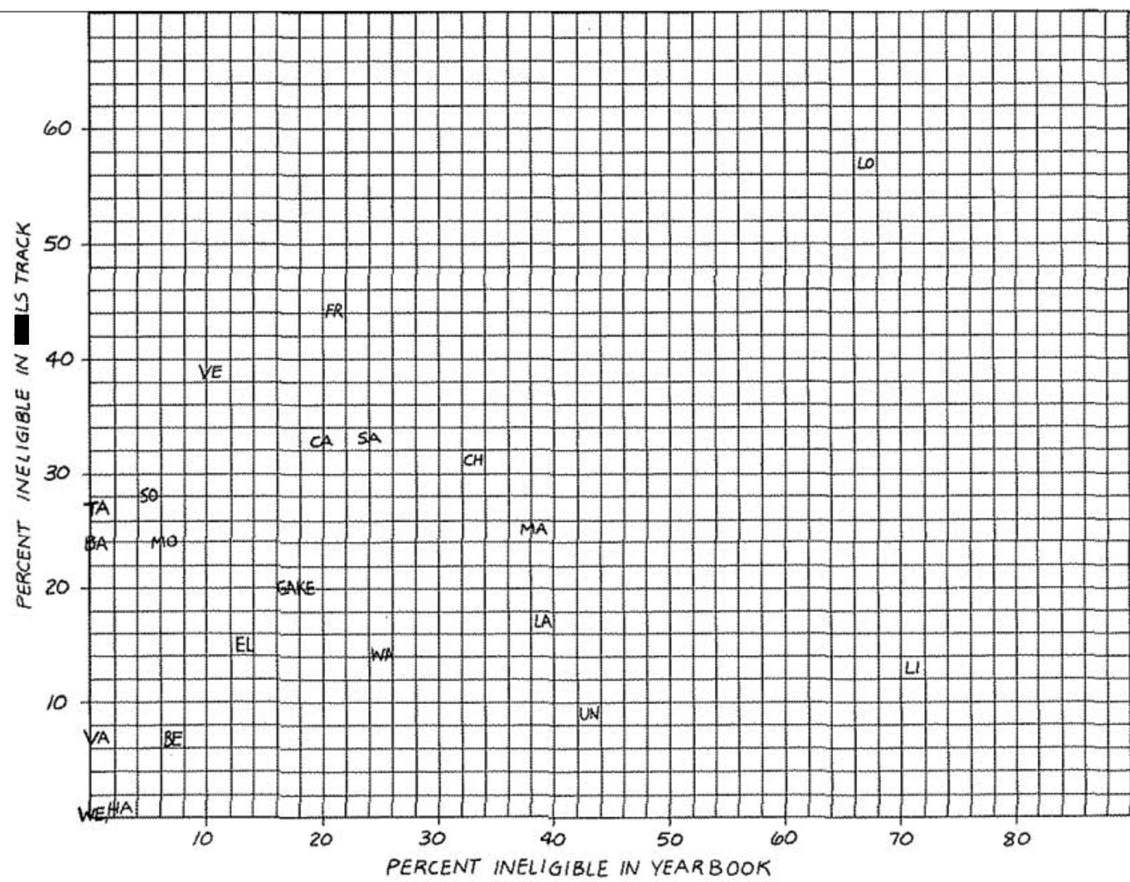
d. Stem-and-leaf plot. The box plot does not display individual values so the mean cannot be computed.
e. Box plot. The box plot shows the quartiles directly.
f. Box plot. These values can be calculated from a stem-and-leaf plot, but the box plot shows them directly.
g. Stem-and-leaf plot. Box plots are not useful with very small data sets because their appearance can change greatly with only small changes in the data.

4. a. manufacturer A
   b. manufacturer C
   c. manufacturer B
   d. manufacturer A or B
   e. You should put your extra effort into improving your cars with the fewest miles per gallon. The top 25 percent of manufacturer C's cars are already better than the top cars from A or B. But C's bottom 50 percent are worse than the bottom 50 percent from A or B.
   f. false

5. From line plots and stem-and-leaf plots, but not from box plots.

ANSWERS TO QUIZ ON SCATTER PLOTS

1. a.



b. no association

c. Answers will vary. Sample: Two schools, Hamilton and Westchester, have zero ineligibility in both activities. Two additional schools, Van Nuys and Belmont, have very low (less than 10 percent) ineligibility in both activities. We can think of this as a small cluster of two schools or as a larger cluster of four schools.

In addition, there are four schools—Banning, Taft, Monroe, and South Gate—that are low for yearbook (less than 10 percent) but moderately high (20–30 percent) for girls' track. It would be interesting to know if those schools have any other characteristics in common.

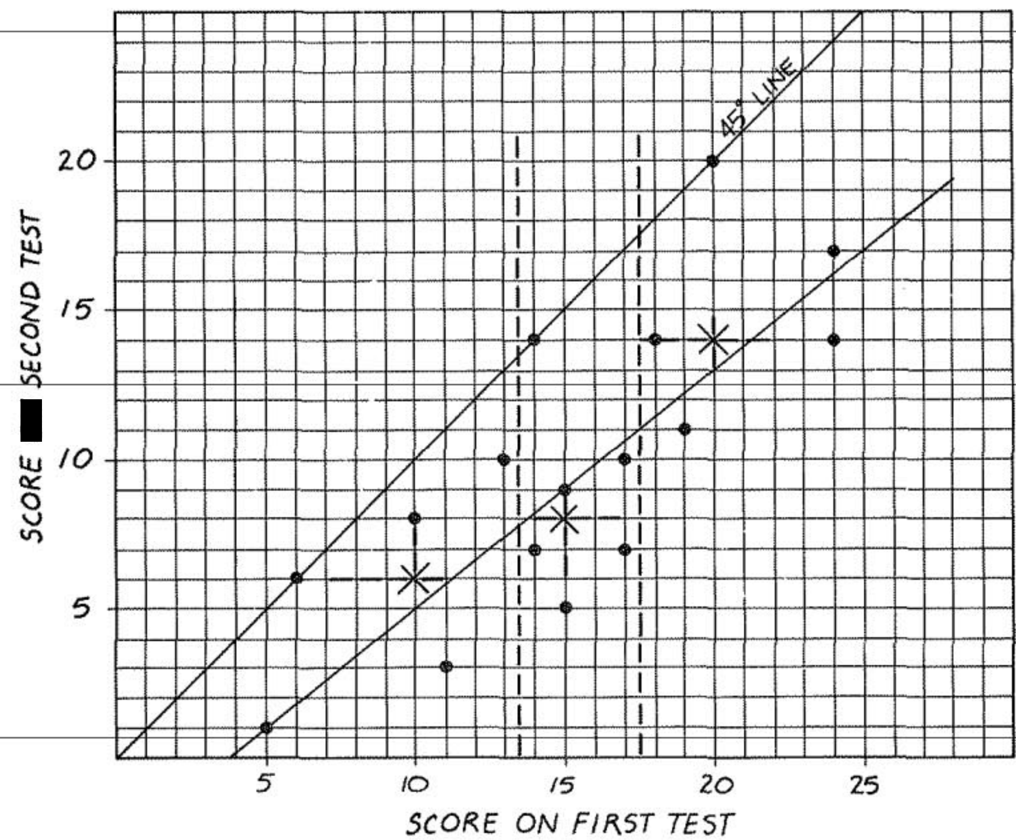
d. Locke stands out because it has very high ineligibility rates in both yearbook and girls' track. Lincoln stands out because it has an unusually high ineligibility rate in yearbook but not a high ineligibility rate in girls' track.

2. a. positive association
   b. no association
   c. positive association

3. a. approximately $8,800.
   b. 1982
   c. Answers will vary. Sample: The median income of high school graduates has risen steadily from about $4,700 in 1958 to about $15,800 in 1983. There is only one year that the median income decreased slightly. That was in 1982, a year of high unemployment. The income of college graduates was about $6,000 in 1958, only $1,300 more than that of high school graduates. The income of college graduates also rose steadily, staying within $3,000 of that for high school graduates, until 1981, when it went up steeply. By 1983, college graduates were earning about $22,000, or more than $6,000 more than high school graduates.

ANSWERS TO QUIZ ON LINES ON SCATTER PLOTS

1. a, b, and f



c. about 15
d. Gail
e. 7
g. three
h. The student's score was the same on both tests.
i. below
j. above the 45° line
k. Answers will vary. Sample: The scores on the first test varied from 5 to 24 and the scores on the second test from 1 to 20. The fact that all but three of the scores are below the 45° line shows that scores on the second test were lower than scores on the first test, which could have happened because the second test was harder, because students didn't study, or because the tests were graded differently. Three students got the same grade on both tests.

   The fitted line shows that, in general, students got about five points lower on the second test than they did on the first test. Gail was the student who was farthest from the predicted value on the second test. She got seven points more than the line predicts.
l. Because the line goes approximately through the two points (5, 1) and (20, 13), we can use the two-point form for the equation of a line to get $y = (4/5)x - 3$.

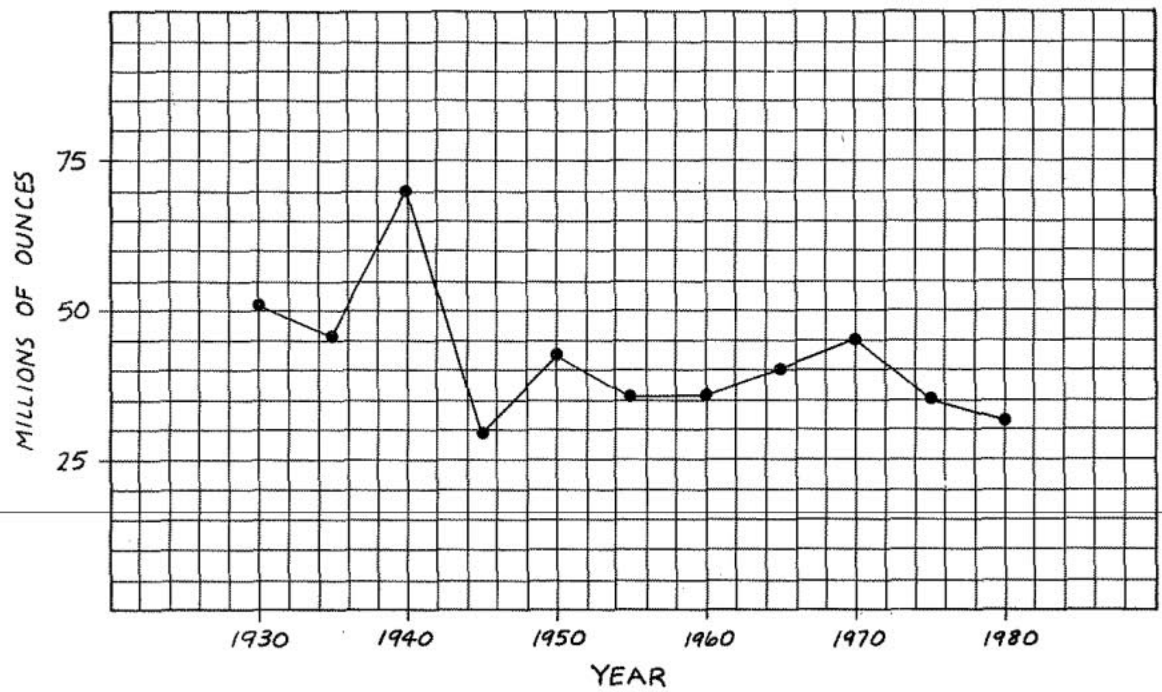
2. seven

3. If we move the ruler this distance, the two outside x's will have twice as much weight as does the center x in determining the position of the line, which makes sense because there are two of them and there is only one center point.
4. true
5. Answers will vary. Sample: A fitted line can be used to determine whether the data follow a straight line (linear) relationship or whether the data follow a curved relationship. In addition, a fitted line enables us to predict a value for the variable on the vertical axis if we are given a value for the variable on the horizontal axis.

ANSWERS TO QUIZ ON SMOOTHING

1. *NOTE TO TEACHERS:* A fine ounce of silver is an ounce by weight that is at least 99.9 percent silver. Sterling silver contains from 94 percent to 98 percent silver.

   a.



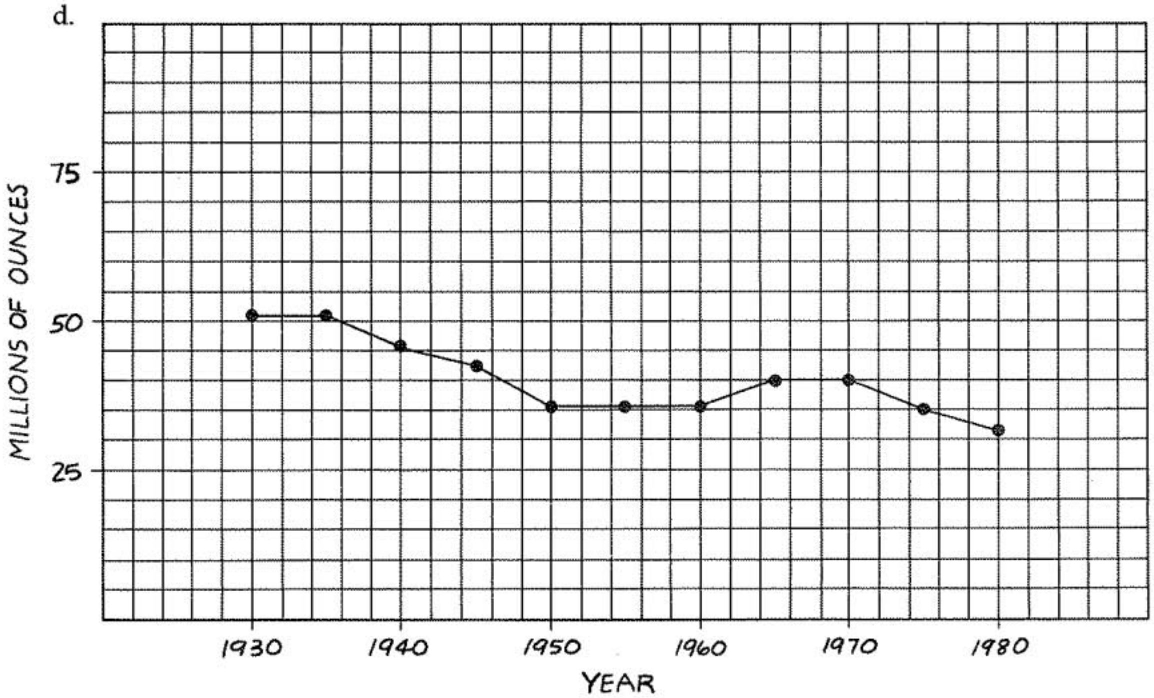
(Answers to quiz on smoothing continued)

b. Because of the sawtooth effect. For some years, such as 1940 and 1945, the number of ounces is unusually high or low.

c.

Year	Fine Ounces	Smoothed Values
1930	51	51
1935	46	51
1940	70	46
1945	29	43
1950	43	36
1955	36	36
1960	36	36
1965	40	40
1970	45	40
1975	35	35
1980	32	32

d.



e. Answers will vary. Sample: The plot of the smoothed values shows that U.S. silver production decreased gradually from 51 million ounces in 1930 to 32 million ounces in 1980. Several years were exceptions to this overall trend, which can be seen by holding the plots together in front of the light. In 1940 an unusually large amount was produced (70 million ounces), and in 1945 an unusually small amount was produced (29 million ounces).

2. They get averaged out (where the average is the median) and so they disappear from the plot of smoothed values.

(Answers to quiz on smoothing continued)

3. Answers will vary. Sample: With the following data, the allowance for 1980, like the one for 1985, is clearly an outlier that should be eliminated in smoothing. However, it is not eliminated because it is an endpoint.

Year	Allowance	Smoothed
1980	$0.25	$0.25
1981	3.00	2.75
1982	2.75	3.00
1983	3.25	3.25
1984	3.50	3.50
1985	15.75	3.75
1986	3.75	3.75

TEACHING NOTES AND ANSWERS

The following pages contain notes on teaching *Exploring Data,* with answers to the discussion questions and the applications. The notes indicate which applications can be omitted if time is a problem.

In many applications, we ask students to write interpretations. Of course the students' answers will all be different, but we give sample answers that cover the points we expect the students to address.

We have included reduced student pages along with the notes and answers so that you will have all the information you need at hand.

Page 1

NOTE TO TEACHERS: Students may ask if a single, isolated point, such as Finland, should be considered a cluster. It is probably better to refer to it as an isolated point and not as a cluster because the term *cluster* implies more than one point.

I. LINE PLOTS

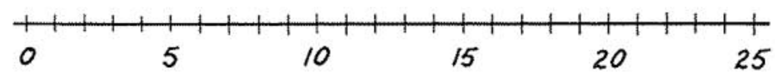
The 1984 Winter Olympics were held in Sarajevo, Yugoslavia. The table below lists the total number of gold, silver, and bronze medals won, by country.

Country	Total Medals	Country	Total Medals
Austria	1	Italy	2
Canada	4	Japan	1
Czechoslovakia	6	Liechtenstein	2
Finland	13	Norway	9
France	3	Sweden	8
Germany, East	24	Switzerland	5
Germany, West	4	USSR	25
Great Britain	1	United States	8
		Yugoslavia	1

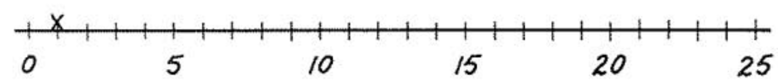
Source: *The World Almanac and Book of Facts,* 1985 edition.

Let's make a *line plot* of these data. First, make a horizontal line.

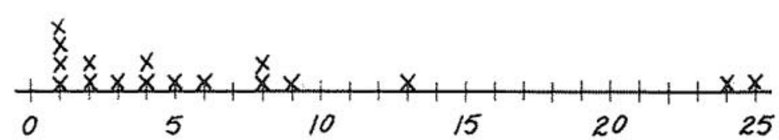
Then, put a scale of numbers on this line using a ruler. Since the smallest number of medals is 1 and the largest is 25, the scale might run from 0 to 25 as shown below.



The first country, Austria, won one medal. To represent Austria, put an *X* above the line at the number 1.



Continuing this way with the other countries, we can complete the line plot as shown below.



From a line plot, features of the data become apparent that were not as apparent from the list. These features include:

- *Outliers* — data values that are substantially larger or smaller than the other values
- *Clusters* — isolated groups of points
- *Gaps* — large spaces between points

It is also easy to spot the largest and smallest values from a line plot. If you see a cluster, try to decide if its members have anything special in common. For example, in the previous line plot the two largest values form a cluster. They are the USSR and East Germany — both eastern European countries. These two values are quite a bit larger than the rest, so we could also consider these points to be outliers.

Often, we would like to know the location of a particular point of interest. For these data, we might want to know how well the United States did compared to the other countries.

Discussion Questions

1. How many countries won only one medal?
2. How many countries won ten or more medals?
3. Do the countries seem to fall into clusters on the line plot?
4. Describe how the United States compares with the other countries.
5. In this book, you will often be asked to "describe what you learned from looking at the plot." Try to do this now with the plot of medal winners, then read the following sample.

Seventeen countries won medals in the 1984 Winter Olympics. Two countries, the USSR with 25 and East Germany with 24, won many more medals than the next country, Finland, with 13. The remaining countries were all clustered, with from 1 to 9 medals each. The United States won 8 medals, more than 11 countries but not many in comparison to the leaders. One noticeable feature about these 17 countries is that, with the exception of the United States, Canada, and Japan, they are all in Europe.

The list does not say how many countries did not win any medals. This might be interesting to find out.

Writing descriptions is probably new to you. When you look at the plot, jot down any observations you make and any questions that occur to you. Look specifically for outliers, clusters, and the other features we mentioned. Then organize and write your paragraphs as if you were composing them for your English teacher. The ability to organize, summarize, and communicate numerical information is a necessary skill in many occupations and is similar to your work with science projects and science laboratory reports.

Page 2: Discussion Questions

1. 4
2. 3
3. yes
4. The United States won eight medals. Only four countries won more medals than this, but the number the U.S. won is much less than the leading countries with 24 and 25.

Page 3:

NOTE TO TEACHERS: Either Application 1, "Rock Albums," or Application 2, "Causes of Death," may be omitted.

Application 1

1. 91
2. 6
3. 27
4. Answers will vary. It could be No. 6 for five weeks or No. 1 for two weeks and No. 6 for one week.
5. Answers will vary. A possible answer is No. 1 for three weeks, No. 2 for fifteen weeks, and No. 5 for three weeks.
6. "Born in the U.S.A." and "Like a Virgin"
7. yes
8. "Private Dancer," "Purple Rain," "Agent Provocateur," and "We Are the World"
9. "Centerfield," "Make It Big," "Beverly Hills Cop," and "No Jacket Required"
10. In the first five months of 1985, two albums, "Born in the U.S.A." with 183 points and "Like a Virgin" with 149 points, were much more popular than any other record. The remaining albums in the top 10 clustered into two groups. One cluster of four albums had from 93 to 108 points, and the other cluster of four had from 49 to 69 points.

Rock Albums

The following list of the top 10 record albums in the first five months of 1985 is based on *Billboard* magazine reports.

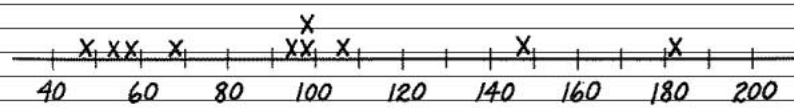
Artist	Title	Total Points
Bruce Springsteen	"Born in the U.S.A."	183
Madonna	"Like a Virgin"	149
Phil Collins	"No Jacket Required"	108
John Fogerty	"Centerfield"	97
Wham!	"Make It Big"	97
Soundtrack	"Beverly Hills Cop"	93
Tina Turner	"Private Dancer"	69
Prince	"Purple Rain"	59
Foreigner	"Agent Provocateur"	54
USA for Africa	"We Are the World"	49

Source: *Los Angeles Times*, May 25, 1985.

The total points were calculated by giving 10 points for each week an album was number 1 on the *Billboard* charts, 9 points for each week it was number 2, 8 points for each week it was number 3, and so forth.

1. If a record was number 1 for 3 weeks, number 2 for 5 weeks, and number 3 for 2 weeks, how many total points would it have?
2. How many points does a record earn by being number 5 for 1 week?
3. If a record was number 4 for 3 weeks and number 5 for 1 week, how many total points would it have?
4. Find two ways for a record to earn 25 points.
5. There were about 21 weeks in the first five months of 1985. Find a way for "Born in the U.S.A." to earn 183 points in these 21 weeks.

The following line plot was constructed from these data.



6. Which record(s) is an outlier?
7. Do the records seem to cluster into more than one group?
8. List the records in the lowest group.
9. List the records in the next lowest group.
10. Write a description of what you learned from studying this plot.

Application 2

Causes of Death

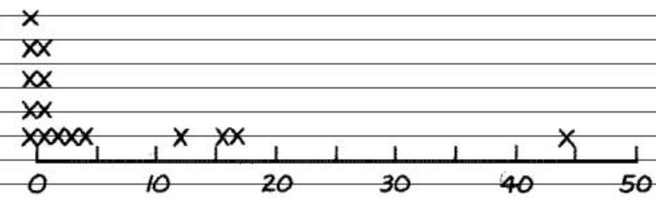
The United States Public Health Service issues tables giving death rates by cause of death. These are broken down by age group, and the table below is for people 15-24 years of age. It gives death rates per 100,000 population for 16 leading causes of death. As an example, a death rate of 1.7 for leukemia means that out of 100,000 people in the United States aged 15-24, we can expect 1.7 of them will die annually from leukemia.

Cause of Death	Death Rate (per 100,000 people aged 15-24 per year)
heart diseases	2.9
leukemia	1.7
cancers of lymph and blood other than leukemia	1.0
other cancers	3.6
strokes	1.0
motor vehicle accidents	44.8
other accidents	16.9
chronic lung diseases	0.3
pneumonia and influenza	0.8
diabetes	0.3
liver diseases	0.3
suicide	12.3
homicide	15.6
kidney diseases	0.3
birth defects	1.4
blood poisoning	0.2

Source: National Center for Health Statistics, Monthly Vital Statistics Report, August 1983.

1. Of 100,000 people aged 15-24, how many would you expect to die annually from pneumonia and influenza?
2. Of 1,000,000 people aged 15-24, how many would you expect to die annually from pneumonia and influenza?
3. Suppose there are 200,000 people, and 3 die from a certain cause. What is the death rate per 100,000 people?
4. Of 250,000 people aged 15-24, about how many would you expect to die annually from motor vehicle accidents?
5. Construct a line plot of these data. To avoid crowding when plotting the *X*'s, round each death rate to the nearest whole number.
6. Which cause of death is an outlier?

Page 4: Application 2

1. 0.8
2. 8
3. 1.5
4. 112
5. 

6. motor vehicle accidents

Page 5: Application 2 (*continued*)

7. other accidents, homicide, suicide
8. other cancers
9. Answers will vary. Sample: The leading cause of death for 15– to 24–year–olds is motor vehicle accidents. With 44.8 deaths per 100,000 people, this number is much greater than those for the next three causes of death: other accidents with 16.9, homicide with 15.6, and suicide with 12.3. It is interesting that these three causes, taken together, have the same death rate as the single leading cause. It is tragic that the four leading causes of death are all preventable.

   The next highest cause of death is other cancers with only 3.6. All of the remaining causes are due to medical problems and have very low rates compared to the leading four. It is interesting that *all* the medical causes, when taken together, have a total death rate of 13.8, which is about the same as each of the rates in the middle cluster of three causes.

   The table does not give some information that it would be interesting to know. For example, in which category are drug overdoses included? Do the death rates vary by gender? Do they vary by race? Are the motor vehicle accident victims primarily drivers, passengers, or pedestrians? When do the fatal accidents tend to occur?
10. Answers will vary.

7. Which three causes of death are in the cluster below the outlier?
8. Which medical problem has the largest death rate?
9. Write a summary of the information communicated by the line plot. Include a list of any questions you have about the data. (For example, in which category are drug overdoses included?)
10. (For class discussion) Suppose you want to reduce the total death rate for 15-24 year olds, and you have $10 million to spend. How would you spend it? On medical research, medical treatment, or in some other way?

Line Plots — Summary

Line plots are a quick, simple way to organize data. They work best when there are fewer than 25 numbers. With many more, the plot starts to look crowded.

From a line plot it is easy to spot the largest and smallest values, outliers, clusters, and gaps in the data. It is also possible to find the relative position of particular points of interest. Sometimes you can notice outliers, clusters, and gaps from the table of data. However, the line plot is easy to make and has several advantages. It makes it easy to spot these features, it gives a graphical picture of the relative sizes of the numbers, and it helps you to make sure that you aren't missing any important information.

When making line plots, be sure to place the X's for values that are approximately the same on top of each other rather than crowding them in. It is also usual to number the scale in multiples of 1, 5, 10, 100, or some other round number.

Suggestions for Student Projects

Collect data on one of the ideas listed below or on your own topic. Make a line plot of the data and write a summary of the information displayed by the plot.

1. heights of students in your class
2. grades on your math tests this year
3. grades on the last test for the members of your class
4. ages of the mothers of students in your class
5. number of hours of television you watch each day for two weeks
6. number of miles each student drives in a week
7. number of students in your class born in each of the 12 months (On the number line, 1 would represent January, 2 would represent February, and so forth.)

II. STEM-AND-LEAF PLOTS

The table below gives the amounts of calories, fat, carbohydrates (sugar and starch), and sodium (salt) in each serving of various fast food items. Fat and carbohydrates are measured in grams; sodium in milligrams.

Item	Calories	Fat (gm)	Carbohydrates (gm)	Sodium (mg)
HAMBURGERS				
Burger King Whopper	660	41	49	1083
Jack-in-the-Box Jumbo Jack	538	28	44	1007
McDonald's Big Mac	591	33	46	963
Wendy's Old Fashioned	413	22	29	708
SANDWICHES				
Roy Rogers Roast Beef	356	12	34	610
Burger King Chopped-Beef Steak	445	13	50	966
Hardee's Roast Beef	351	17	32	765
Arby's Roast Beef	370	15	36	869
FISH				
Long John Silver's	483	27	27	1333
Arthur Treacher's Original	439	27	27	421
McDonald's Filet-O-Fish	383	18	38	613
Burger King Whaler	584	34	50	968
CHICKEN				
Kentucky-Fried Chicken Snack Box	405	21	16	728
Arthur Treacher's Original Chicken	409	23	25	580
SPECIALTY ENTREES				
Wendy's Chili	266	9	29	1190
Pizza Hut Pizza Supreme	506	15	64	1281
Jack-in-the-Box Taco	429	26	34	926

Source: *Consumer Reports*, September 1979.

Suppose you decide to order a McDonald's Big Mac. It contains 33 grams of fat. How does this compare to the other items? By looking at the table, about all we can see is that it does not have the most fat nor the least. So that we can get a better picture of the grams of fat per serving, let's make a stem-and-leaf plot.

First, find the smallest value and the largest value.

The smallest value is 9 for Wendy's Chili and the largest is 41 for the Burger King Whopper.

The smallest value, 9, has a 0 in the ten's place and the largest value, 41, has a 4 in the ten's place. Therefore, we choose the *stems* to be the digits from 0 to 4.

Second, write these stems vertically with a line to their right.

```
0 |
1 |
2 |
3 |
4 |
```

Third, separate each value into a stem and a leaf and put the leaves on the plot to the right of the stem.

For example, the first value in the list is 41, for a Burger King Whopper. Its stem is 4 and its leaf is 1. It is placed on the plot as follows:

```
0 |
1 |
2 |
3 |
4 | 1
```

The second value in the list is 28. Its stem is 2 and its leaf is 8. Now the plot looks as shown below.

```
0 |
1 |
2 | 8
3 |
4 | 1
```

Continuing in this way gives the following plot:

```
0 | 9
1 | 2 3 7 5 8 5
2 | 8 2 7 7 1 3 6
3 | 3 4
4 | 1
```

Next, on a new plot arrange the leaves so they are ordered from smallest value to largest. (This final step is often omitted.)

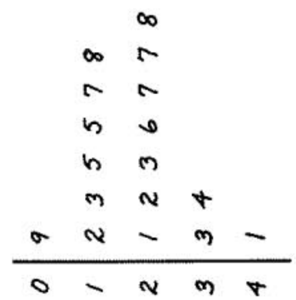
> Be sure to add an explanation like this to the left of your plot.

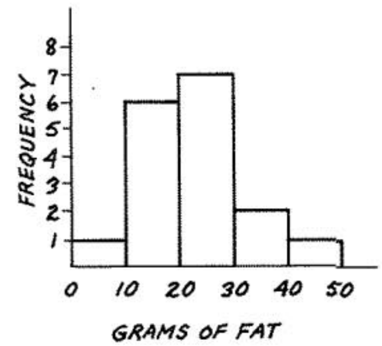
2|3 represents 23 grams of fat

```
0 | 9
1 | 2 3 5 5 7 8
2 | 1 2 3 6 7 7 8
3 | 3 4
4 | 1
```

The plot shows that most of the food items have grams of fat in the 10's and 20's and that there are a few large values. The McDonald's Big Mac with 33 grams has one of the larger amounts of fat.

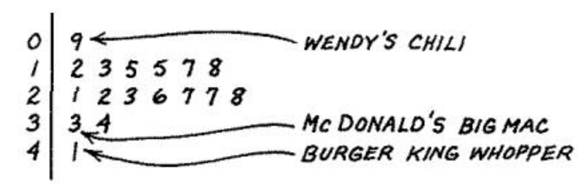
If we rotate the stem-and-leaf plot 90° counterclockwise, we get a plot that resembles a bar graph or histogram.





The stem-and-leaf plot is often better than the bar graph or histogram because it is easier to construct and all the original data values are displayed.

It is sometimes worthwhile to label specific items. For example, we might want to label the smallest value, the largest value, and a value of special interest such as McDonald's Big Mac. This is shown below.



Also, it is sometimes interesting to replace the leaves in the stem-and-leaf plot by symbols identifying the items. For example, replace each of the four hamburger leaves with an *H*, each of the four sandwich leaves with an *S*, each of the four fish leaves with an *F*, each of the two chicken leaves with a *C*, and each of the three special entree leaves with an *O* (for other). Replacing the leaves by symbols gives the following:

```
0 | O
1 | S S S O S F
2 | C H C O F F H
3 | H F
4 | H
```

When writing a description of a stem-and-leaf plot, look for the same features that you looked for with a line plot:

- largest and smallest values
- outliers
- clusters
- gaps
- the relative position of any item important to you

Our description of what we learned about fat in the fast food items from the stem-and-leaf plots follows:

There are no outliers separated far from the rest nor any large internal gaps among these values. Of these fast foods, the type that is generally highest in fat is the hamburger, which has three of the highest four values. One hamburger is lower in fat than the others and lies in about the middle of all these values; it is Wendy's Old Fashioned. Some possible reasons for its lower value are: it might be smaller than the others, it might be made from meat with a lower fat content, or it might be cooked differently.

From the data, the type of food that is second highest in fat is fish; the values are only slightly smaller than those for hamburgers. Again, one fish value, McDonald's Filet-O-Fish, is smaller than the other fish values. Although we generally think of fish as having a lot less fat than beef, perhaps these fish items are all fried and therefore high in fat.

The type of food lowest in fat is the roast beef sandwich, and chicken falls near the middle in these data. It is surprising that both the lowest and highest items are beef, but perhaps the sandwich is lowest because it is not fried. The other specialty items are spread throughout the data, but they include the single lowest item, Wendy's Chili. Is it just a coincidence that the hamburger that was lowest in fat was also from Wendy's?

When analyzing data throughout this book, you will need to examine the plots and to think about other information you may have from outside mathematics that can help to interpret the results. Sometimes, this process will lead to questions and possibilities about the problem that cannot be answered just from the data.

The stem-and-leaf plot shows the shape of the set of data more clearly than a line plot. The "shape" of a set of data is called its *distribution*. For example, some common types of distribution follow:

3	4
4	6 9 9
5	2 4 4 5 5 9
6	1 1 7
7	8

MOUND-SHAPED

2	5 5 8
3	2 3 4 4 5 9
4	6 7 7
5	4
6	1 1 3 8
7	0 1 3 4 4 5 6 8 8
8	2 3 5 5

U-SHAPED

3	4
4	7 8
5	2 2 3
6	1 1 2 4 4 5 7 8
7	0 1 2 2 2 3 6 8 8 9 9

J-SHAPED

3	2 2 3 8
4	1 5 7
5	0 4 4 9
6	1 1 5 7
7	3 6 8 8 9

RECTANGULAR-SHAPED

The mound-shaped distribution, sometimes called bell-shaped, is a shape that occurs often. The data values are fairly symmetrical, with lows balancing the highs. If the data follow a U-shaped distribution, it may be because there are really two underlying groups, each of which is mound-shaped, corresponding to the two peaks. Thus, when a U-shaped plot is observed, it is a good idea to see if there is any reason to treat the observations as two separate groups.

The J-shaped plot or the backward J-shaped plot does not occur as often as the first two types. Typically, it occurs because it is impossible to have observations above (or below) a particular limit. In the example above, this limit might be 80. In some problems, there is a lower limit of 0. If you observe a J-shaped plot, try to determine if there is a limit, what it is, and why it is there. For a rectangular-shaped distribution, sometimes called flat or uniform, there are often both lower and upper limits with the data values spread evenly between them. For the previous example, the limits might be 30 and 80. As with the J-shaped plot, you should try to understand if there are limits to the possible values of the data, and what the limits might mean.

Discussion Questions

1. Make a stem-and-leaf plot of the grams of carbohydrates in the fast food items. Label the smallest value, the largest value, and McDonald's Big Mac.
2. Make another stem-and-leaf plot of the grams of carbohydrates, but replace the leaves by the symbols:
   H for hamburger
   S for sandwich
   F for fish
   C for chicken
   O for other
3. Write a description of the information displayed in the stem-and-leaf plot of the grams of carbohydrates. Mention any interesting patterns. How does this plot compare to the one for fat?
4. All of the fast food information was given on a per item basis. However, the sizes of the items are different. Do you think this should be taken into account? How might you do this? Should price also be considered?
5. In judging fast food items, which is most important to you: calories, fat, carbohydrates, or sodium?
6. Give an example of data that are distributed a) U-shaped. b) mound-shaped. c) J-shaped. d) rectangular-shaped.

Page 12: Discussion Questions

1.

```
1 | 6 ← KENTUCKY FRIED CHICKEN SNACK BOX
2 | 5 7 7 9 9
3 | 2 4 4 6 8
4 | 4 6 9 ← MCDONALD'S BIG MAC
5 | 0 0
6 | 4 ← PIZZA HUT PIZZA SUPREME        1|6 REPRESENTS 16 GRAMS
                                        OF CARBOHYDRATES
```

2.

```
1 | C
2 | C F F H O
3 | S S O S F
4 | H H H
5 | S F
6 | O
```

3. Answers will vary. Sample: The lowest number of grams of carbohydrates is 16 and the highest is 64. The number of grams in Pizza Hut Pizza Supreme, 64, is quite a bit larger than the 50 grams in the next highest item. Otherwise, there are no especially large gaps or clusters in this distribution.

   The two items lowest in carbohydrates are both chicken. However, the other types of items are mixed up and show no strong patterns. It is interesting, though, that three of the four hamburgers are grouped together in the upper half of the distribution while the value of the fourth is substantially smaller. This hamburger, Wendy's Old Fashioned, is also the one that had less fat content.

   This plot does not look similar to the one for fat. The items highest in fat, hamburgers and fish, are not highest in carbohydrates.

4. Answers will vary.
5. Answers will vary.
6. Answers will vary. Samples: U-shaped: scores on an algebra test of all tenth graders in a school, some of whom have taken algebra and some of whom have not.

   Mound-shaped: heights of the boys or girls in your class.

   J-shaped: grades on an easy test.

   Rectangular-shaped: last digits of the phone numbers of students in your class.

Page 13

NOTE TO TEACHERS: Either Application 3, "Ages of U.S. Presidents at Their Death," or Application 4, "Thunderstorms," may be omitted.

Before students construct the stem-and-leaf plot of question 1, remind them to put the leaves on in the order that the presidents appear on the list. That is, students should not try to find all values that go on the first stem, then find the values that go on the second stem, and so on.

It is generally not important that the leaves be in order. If you or your students prefer to have them in order, you can make a second plot quickly from the first one.

Application 3

1\.

KENNEDY (4|6), GARFIELD (4|9), LINCOLN (5|6), MCKINLEY (5|8)

4	6 9
5	3 6 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 3 5 8
9	0 0

4|6 REPRESENTS 46 YEARS OLD

2. 7
3. Adams, Hoover
4. See the preceding plot.
5. mound-shaped
6. Answers will vary. Sample: The youngest president to die was Kennedy at age 46. In fact, of the seven who died before age 60, four were assassinated. They were Kennedy, Garfield, Lincoln, and McKinley.

   Only two presidents, Adams and Hoover, lived to be 90. Most presidents die in their sixties or seventies. There have been about as many deaths in the forties and fifties as in the eighties and nineties, giving a mound-shaped distribution.

Application 3

Ages of U.S. Presidents at Their Death

The table below lists the presidents of the United States and the ages at which they died.

Washington	67	Filmore	74	Roosevelt	60
Adams	90	Pierce	64	Taft	72
Jefferson	83	Buchanan	77	Wilson	67
Madison	85	Lincoln	56	Harding	57
Monroe	73	Johnson	66	Coolidge	60
Adams	80	Grant	63	Hoover	90
Jackson	78	Hayes	70	Roosevelt	63
Van Buren	79	Garfield	49	Truman	88
Harrison	68	Arthur	57	Eisenhower	78
Tyler	71	Cleveland	71	Kennedy	46
Polk	53	Harrison	67	Johnson	64
Taylor	65	McKinley	58		

1. Make a stem-and-leaf plot of the ages using these stems.

4	
5	
6	
7	
8	
9	

2. How many presidents died in their forties or fifties?
3. Who lived to be the oldest?
4. Label the four presidents who were assassinated.
5. What is the shape of this distribution?
6. Write a one-paragraph description of the information shown in the stem-and-leaf plot, including information about the presidents who were assassinated.

Application 4

Thunderstorms

The table below lists 81 U.S. cities with the number of days per year with thunderstorms.

Area	Number of Days	Area	Number of Days	Area	Number of Days
Akron, OH	39	Detroit, MI	33	Oklahoma City, OK	51
Albany, NY	28	El Paso, TX	36	Omaha, NE	51
Albuquerque, NM	43	Fargo, ND	30	Orlando, FL	85
Anchorage, AK	1	Fresno, CA	5	Philadelphia, PA	42
Atlanta, GA	50	Grand Rapids, MI	37	Phoenix, AZ	20
Austin, TX	40	Great Falls, MT	27	Pittsburgh, PA	35
Bakersfield, CA	3	Hartford, CT	28	Portland, ME	20
Baltimore, MD	24	Honolulu, HI	7	Portland, OR	7
Baton Rouge, LA	80	Houston, TX	59	Providence, RI	21
Beaumont, TX	63	Indianapolis, IN	47	Raleigh, NC	45
Biloxi, MS	80	Kansas City, MO	50	Richmond, VA	37
Birmingham, AL	65	Las Vegas, NV	13	Rochester, NY	29
Boise, ID	15	Little Rock, AR	56	Sacramento, CA	5
Boston, MA	19	Louisville, KY	52	Salt Lake City, UT	41
Buffalo, NY	30	Los Angeles, CA	6	San Antonio, TX	35
Burlington, VT	27	Manchester, NH	24	San Diego, CA	3
Charleston, SC	58	Memphis, TN	50	San Francisco, CA	2
Charleston, WV	45	Miami, FL	71	Seattle, WA	6
Chicago, IL	36	Milwaukee, WI	37	Shreveport, LA	58
Cincinnati, OH	52	Minneapolis, MN	36	Sioux Falls, SD	47
Cleveland, OH	38	Mobile, AL	86	St. Louis, MO	43
Columbia, SC	52	Nashville, TN	52	Tampa, FL	91
Columbus, OH	36	Nassau-Suffolk, NY	18	Tucson, AZ	28
Corpus Christi, TX	32	Newark, NJ	25	Tulsa, OK	53
Dallas, TX	41	New Orleans, LA	73	Washington, DC	28
Denver, CO	38	New York, NY	18	Wichita, KS	53
Des Moines, IA	55	Norfolk, VA	36	Wilmington, DE	30

Source: United States Weather Bureau.

Use this map with Application 4 to locate each city in the west, south, central, or northeast region.

Page 17: Application 4

1. Answers will vary.
2. Tampa, Mobile, Orlando, Biloxi, and Baton Rouge; they are near the Gulf Coast.
3.

0	W W W W W W W W W W
1	W W N N N
2	W N N N N N N W N N W N N
3	N C N S C N S C C S C S C C S C W C
4	S S W N W C N S C C
5	S C S S C C S S S S C C S S S S
6	S S
7	S S
8	S S S S
9	S

4. Answers will vary. Sample: The cities with 15 or fewer days of thunderstorms per year are all in the west. The cities with 56 or more are all in the south. In general, the west has the fewest, followed by the northeast, the central region, and the south.

   Three cities in the west—Albuquerque (43), Salt Lake City (41), and Denver (38)—all have many more thunderstorms than the next highest western city, Tucson (28). Thus, perhaps those three cities should really be classified as central.

NOTE TO TEACHERS: Here is a suggested way for students to make the regional stem-and-leaf plot. First copy the plot from page 17, making sure there is space between the rows. Then go through the list of cities and, for each one, write the label (W, S, C, or N) above the value in the plot. The result follows. You may want to have students work on this in groups.

Stem	Leaves
	W W W W W W W W W W
0	1 2 3 3 5 5 6 6 7 7
	W W N N N
1	3 5 8 8 9
	W N N N N N N W N N W N N
2	0 0 1 4 4 5 7 7 8 8 8 8 9
	N C N S C N S C C S C S C C S C W C
3	0 0 0 2 3 5 5 6 6 6 6 6 7 7 7 8 8 9
	S S W N W C N S C C
4	0 1 1 2 3 3 5 5 7 7
	S C S S C C S S S S C C S S S S
5	0 0 0 1 1 2 2 2 2 3 3 5 6 8 8 9
	S S
6	3 5
	S S
7	1 3
	S S S S
8	0 0 5 6
	S
9	1

6|3 REPRESENTS 63 THUNDERSTORMS PER YEAR

A stem-and-leaf plot of the number of days of thunderstorms is shown below. Notice that the stem for numbers less than 10 is 0.

Stem	Leaves
0	1 2 3 3 5 5 6 6 7 7
1	3 5 8 8 9
2	0 0 1 4 4 5 7 7 8 8 8 8 9
3	0 0 0 2 3 5 5 6 6 6 6 6 7 7 7 8 8 9
4	0 1 1 2 3 3 5 5 7 7
5	0 0 0 1 1 2 2 2 2 3 3 5 6 8 8 9
6	3 5
7	1 3
8	0 0 5 6
9	1

6|3 REPRESENTS 63 THUNDERSTORMS PER YEAR

1. How does your city, or the city nearest you, compare to the other cities?
2. Which five cities have the largest number of days with thunderstorms? What do these five cities have in common?
3. The map on page 15 shows the United States divided into four regions: west, south, central, and northeast. Make a stem-and-leaf plot, replacing each city with the label for its location:

   W for WEST
   S for SOUTH
   C for CENTRAL
   N for NORTHEAST
4. Write a summary of what you can see in this stem-and-leaf plot.

Application 5

Soft Drinks

The table below shows the number of *gallons* of soft drinks sold per person in 1977 for each state.

State	Gallons per Person	State	Gallons per Person
Alabama (AL)	36.8	Nebraska (NE)	32.9
Alaska (AK)	29.5	Nevada (NV)	34.5
Arizona (AZ)	29.1	New Hampshire (NH)	28.4
Arkansas (AR)	33.3	New Jersey (NJ)	28.7
California (CA)	32.2	New Mexico (NM)	28.7
Colorado (CO)	30.0	New York (NY)	31.7
Connecticut (CT)	31.3	North Carolina (NC)	39.9
Delaware (DE)	32.5	North Dakota (ND)	23.2
Florida (FL)	39.7	Ohio (OH)	34.1
Georgia (GA)	39.4	Oklahoma (OK)	31.0
Hawaii (HI)	31.3	Oregon (OR)	23.8
Idaho (ID)	20.7	Pennsylvania (PA)	26.5
Illinois (IL)	33.2	Rhode Island (RI)	28.5
Indiana (IN)	28.8	South Carolina (SC)	39.1
Iowa (IA)	29.0	South Dakota (SD)	25.5
Kansas (KS)	35.9	Tennessee (TN)	36.4
Kentucky (KY)	35.3	Texas (TX)	35.9
Louisiana (LA)	36.7	Utah (UT)	28.0
Maine (ME)	29.2	Vermont (VT)	26.6
Maryland (MD)	34.9	Virginia (VA)	38.3
Massachusetts (MA)	31.6	Washington (WA)	25.1
Michigan (MI)	33.4	Washington, D.C. (DC)	36.0
Minnesota (MN)	33.0	West Virginia (WV)	34.2
Mississippi (MS)	38.2	Wisconsin (WI)	28.8
Missouri (MO)	36.4	Wyoming (WY)	20.6
Montana (MT)	23.3		

Source: *Beverage World*, March 1978.

(After each state is its two-letter postal abbreviation. In some applications we will use these for identifying the states, so you may need to refer back to this list to check any that are unfamiliar.)

1. How many ounces are in a gallon?
2. In Alabama, 36.8 gallons were sold per person. How many ounces were sold per person? How many 12-ounce cans would 36.8 gallons fill?
3. For the number of gallons per person in your state, find the equivalent number of 12-ounce cans of soft drinks.

Page 18

NOTE TO TEACHERS: Before students construct the stem-and-leaf plot for question 4, remind them to put the leaves on in the order that the states appear on the list. That is, students should not try to find all values that go on the first stem, then all of the values that go on the second stem, and so forth. It is generally not important that the leaves be in order. If you or your students prefer to have them in order, you can make a second plot quickly from the first one.

Application 5

1. 128
2. 4710.4; about 392.5 cans
3. Answers will vary.

Page 19: Application 5 (*continued*)

4.

Stem	Leaves	
20	7 6	7 ← IDAHO; 6 ← WYOMING
21		
22		
23	3 2 8	
24		
25	5 1	
26	5 6	
27		
28	8 4 7 7 5 0 8	
29	5 1 0 2	
30	0	
31	3 3 6 7 0	
32	2 5 9	
33	3 2 4 0	
34	9 5 1 2	
35	9 3 9	
36	8 7 4 4 0	
37		
38	2 3	2 → MISSISSIPPI; 3 → VIRGINIA
39	7 4 9 1	1 → SOUTH CAROLINA; 9 → NORTH CAROLINA; 4 → GEORGIA; 7 → FLORIDA

20|6 REPRESENTS 20.6 GALLONS

5. Answers will vary.
6. See the preceding plot.
7. See the preceding plot.
8. the south; the temperature there (hot)
9. Answers will vary. Sample: These data might possibly have been obtained from an association of soft drink manufacturers, or from a survey taken by the U.S. Department of Commerce. They were undoubtedly obtained as gross sales for each state, then divided by population size to get the per person value.

4. These data are different from previous sets of data since the numbers contain decimals. The values go from 20.6 to 39.9, so we choose the stems to be 20, 21, 22, ..., 39. Copy and complete this stem-and-leaf plot of the gallons per person. The plot has been started with the values for Alabama and Alaska.

Stem	Leaves
20	
21	
22	
23	
24	
25	
26	
27	
28	
29	5
30	
31	
32	
33	
34	
35	
36	8
37	
38	
39	

20|6 REPRESENTS 20.6 GALLONS

5. Label your state.
6. Label the states that have the lowest soft drink consumption.
7. Label the states that have the highest soft drink consumption.
8. Which region of the country consumes the most soft drinks per person? What is your explanation for this?
9. (For class discussion) How could these data have been collected?

Back-to-Back Stem-and-Leaf Plots and Spreading Out Stem-and-Leaf Plots

Sometimes we want to compare two sets of data. For example, look at the following tables that contain the home run leaders for the National League and American League from 1921 to 1985.

Home Run Leaders

Year	National League	HR	American League	HR
1921	George Kelly, New York	23	Babe Ruth, New York	59
1922	Rogers Hornsby, St. Louis	42	Ken Williams, St. Louis	39
1923	Cy Williams, Philadelphia	41	Babe Ruth, New York	41
1924	Jacques Fournier, Brooklyn	27	Babe Ruth, New York	46
1925	Rogers Hornsby, St. Louis	39	Bob Meusel, New York	33
1926	Hack Wilson, Chicago	21	Babe Ruth, New York	47
1927	Hack Wilson, Chicago Cy Williams, Philadelphia	30	Babe Ruth, New York	60
1928	Hack Wilson, Chicago Jim Bottomley, St. Louis	31	Babe Ruth, New York	54
1929	Charles Klein, Philadelphia	43	Babe Ruth, New York	46
1930	Hack Wilson, Chicago	56	Babe Ruth, New York	49
1931	Charles Klein, Philadelphia	31	Babe Ruth, New York Lou Gehrig, New York	46
1932	Charles Klein, Philadelphia Mel Ott, New York	38	Jimmy Foxx, Philadelphia	58
1933	Charles Klein, Philadelphia	28	Jimmy Foxx, Philadelphia	48
1934	Rip Collins, St. Louis Mel Ott, New York	35	Lou Gehrig, New York	49
1935	Walter Berger, Boston	34	Jimmy Foxx, Philadelphia Hank Greenberg, Detroit	36
1936	Mel Ott, New York	33	Lou Gehrig, New York	49
1937	Mel Ott, New York Joe Medwick, St. Louis	31	Joe DiMaggio, New York	46
1938	Mel Ott, New York	36	Hank Greenberg, Detroit	58
1939	John Mize, St. Louis	28	Jimmy Foxx, Boston	35
1940	John Mize, St. Louis	43	Hank Greenberg, Detroit	41
1941	Dolph Camilli, Brooklyn	34	Ted Williams, Boston	37
1942	Mel Ott, New York	30	Ted Williams, Boston	36
1943	Bill Nicholson, Chicago	29	Rudy York, Detroit	34
1944	Bill Nicholson, Chicago	33	Nick Etten, New York	22
1945	Tommy Holmes, Boston	28	Vern Stephens, St. Louis	24
1946	Ralph Kiner, Pittsburgh	23	Hank Greenberg, Detroit	44
1947	Ralph Kiner, Pittsburgh John Mize, New York	51	Ted Williams, Boston	32
1948	Ralph Kiner, Pittsburgh John Mize, New York	40	Joe DiMaggio, New York	39
1949	Ralph Kiner, Pittsburgh	54	Ted Williams, Boston	43
1950	Ralph Kiner, Pittsburgh	47	Al Rosen, Cleveland	37
1951	Ralph Kiner, Pittsburgh	42	Gus Zernial, Chicago-Philadelphia	33
1952	Ralph Kiner, Pittsburgh Hank Sauer, Chicago	37	Larry Doby, Cleveland	32

Source: *The World Almanac and Book of Facts*, 1985 edition.

Home Run Leaders

Year	National League	HR	American League	HR
1953	Ed Mathews, Milwaukee	47	Al Rosen, Cleveland	43
1954	Ted Kluszewski, Cincinnati	49	Larry Doby, Cleveland	32
1955	Willie Mays, New York	51	Mickey Mantle, New York	37
1956	Duke Snider, Brooklyn	43	Mickey Mantle, New York	52
1957	Hank Aaron, Milwaukee	44	Roy Sievers, Washington	42
1958	Ernie Banks, Chicago	47	Mickey Mantle, New York	42
1959	Ed Mathews, Milwaukee	46	Rocky Colavito, Cleveland Harmon Killebrew, Washington	42
1960	Ernie Banks, Chicago	41	Mickey Mantle, New York	40
1961	Orlando Cepeda, San Francisco	46	Roger Maris, New York	61
1962	Willie Mays, San Francisco	49	Harmon Killebrew, Minnesota	48
1963	Hank Aaron, Milwaukee Willie McCovey, San Francisco	44	Harmon Killebrew, Minnesota	45
1964	Willie Mays, San Francisco	47	Harmon Killebrew, Minnesota	49
1965	Willie Mays, San Francisco	52	Tony Conigliaro, Boston	32
1966	Hank Aaron, Atlanta	44	Frank Robinson, Baltimore	49
1967	Hank Aaron, Atlanta	39	Carl Yastrzemski, Boston Harmon Killebrew, Minnesota	44
1968	Willie McCovey, San Francisco	36	Frank Howard, Washington	44
1969	Willie McCovey, San Francisco	45	Harmon Killebrew, Minnesota	49
1970	Johnny Bench, Cincinnati	45	Frank Howard, Washington	44
1971	Willie Stargell, Pittsburgh	48	Bill Melton, Chicago	33
1972	Johnny Bench, Cincinnati	40	Dick Allen, Chicago	37
1973	Willie Stargell, Pittsburgh	44	Reggie Jackson, Oakland	32
1974	Mike Schmidt, Philadelphia	36	Dick Allen, Chicago	32
1975	Mike Schmidt, Philadelphia	38	George Scott, Milwaukee Reggie Jackson, Oakland	36
1976	Mike Schmidt, Philadelphia	38	Graig Nettles, New York	32
1977	George Foster, Cincinnati	52	Jim Rice, Boston	39
1978	George Foster, Cincinnati	40	Jim Rice, Boston	46
1979	Dave Kingman, Chicago	48	Gorman Thomas, Milwaukee	45
1980	Mike Schmidt, Philadelphia	48	Reggie Jackson, New York Ben Oglivie, Milwaukee	41
1981	Mike Schmidt, Philadelphia	31	Bobby Grich, California Tony Armas, Oakland Dwight Evans, Boston Eddie Murray, Baltimore	22
1982	Dave Kingman, New York	37	Gorman Thomas, Milwaukee Reggie Jackson, California	39
1983	Mike Schmidt, Philadelphia	40	Jim Rice, Boston	39
1984	Mike Schmidt, Philadelphia Dale Murphy, Atlanta	36	Tony Armas, Boston	43
1985	Dale Murphy, Atlanta	37	Darrell Evans, Detroit	40

Source: *The World Almanac and Book of Facts*, 1985 edition.

In which league does the leader tend to hit more home runs? To find out, we make the following back-to-back stem-and-leaf plot of these data. Notice that the stems are in the center of the plot.

NATIONAL LEAGUE		AMERICAN LEAGUE
98887331	2	224
998887776666544331111100	3	22222223334566677777799999
9988877776655444433322110000	4	00111222333444455666666788999999
642211	5	24889
	6	01

|2|4 REPRESENTS 24 HOME RUNS

There are too many leaves per stem, so we will spread out the stem-and-leaf plot using the stems that follow.

NATIONAL LEAGUE		AMERICAN LEAGUE
	2	
	·	
	3	
	·	
	4	
	·	
	5	
	·	
	6	

We will put the leaves 0, 1, 2, 3, and 4 on the first line for each stem and the leaves 5, 6, 7, 8, and 9 on the second line. The reorganized plot is shown as follows:

NATIONAL LEAGUE		AMERICAN LEAGUE
3 3 1	2	2 2 4
9 8 8 8 7	·	
4 4 3 3 1 1 1 1 0 0	3	2 2 2 2 2 2 2 3 3 3 4
9 9 8 8 8 7 7 7 6 6 6 6 5	·	5 6 6 6 7 7 7 7 9 9 9 9 9
4 4 4 4 3 3 3 2 2 1 1 0 0 0 0	4	0 0 1 1 1 2 2 2 3 3 3 4 4 4 4
9 9 8 8 8 7 7 7 7 6 6 5 5	·	5 5 6 6 6 6 6 7 8 8 9 9 9 9 9 9
4 2 2 1 1	5	2 4
6	·	8 8 9
	6	0 1

|2|4 REPRESENTS 24 HOME RUNS

Page 23: Discussion Questions

1. American League
2. 1981, 1944, and 1945; strike in 1981 and World War II in 1944 and 1945.
3.

NATIONAL LEAGUE		AMERICAN LEAGUE
1	2	
3 3	·	2 2
	·	4
7	·	
9 8 8 8	·	
1 1 1 1 0 0	3	
3 3	·	2 2 2 2 2 2 2 3 3 3
5 4 4	·	4 5
7 7 7 6 6 6 6	·	6 6 6 7 7 7 7
9 9 8 8 8	·	9 9 9 9 9
1 1 0 0 0 0	4	0 0 1 1 1
3 3 3 2 2	·	2 2 2 3 3 3
5 5 4 4 4 4	·	4 4 4 4 5 5
7 7 7 7 6 6	·	6 6 6 6 6 7
9 9 8 8 8	·	8 8 9 9 9 9 9 9
1 1	5	
2 2	·	2
4	·	4
6	·	
	·	8 8 9
	6	0 1

9|3| REPRESENTS
39 HOME RUNS

4. We like the last plot best. There are neither too many nor too few leaves for each stem. The gaps show up in this plot while they weren't visible at all in the first plot, and they were less visible in the second plot than they are in the last plot.

Discussion Questions

1. Does the American League champion or the National League champion tend to hit the most home runs?
2. Which three years were unusually low in home runs hit in the American League? What happened in these three years?
3. Make a new back-to-back stem-and-leaf plot using the stems that follow. The home runs for the National League have been done for you. To construct this plot, you don't have to go back to the original list of data. Instead, take the values from one of the stem-and-leaf plots already constructed.

   For each stem, put the leaves:

   - 0 and 1 on the first line
   - 2 and 3 on the second line
   - 4 and 5 on the third line
   - 6 and 7 on the fourth line
   - 8 and 9 on the last line

NATIONAL LEAGUE		AMERICAN LEAGUE
1	2	
3 3	·	
	·	
7	·	
9 8 8 8	·	
1 1 1 1 0 0	3	
3 3	·	
5 4 4	·	
7 7 7 6 6 6 6	·	
9 9 8 8 8	·	
1 1 0 0 0 0	4	
3 3 3 2 2	·	
5 5 4 4 4 4	·	
7 7 7 7 6 6	·	
9 9 8 8 8	·	
1 1	5	
2 2	·	
4	·	
6	·	
	·	
	6	

9|3| REPRESENTS
39 HOME RUNS

4. Which of the three back-to-back stem-and-leaf plots for the home run data do you think displays the data best? Why?

From a back-to-back plot like this, we can see that there tends to be a slightly larger number of home runs in the American League. We reach this conclusion because the values at the high end, in the upper 50's and 60's, come more often from the American League. Also, the values at the low end, in the 20's, come more often from the National League. For the stems in the 30's and the 40's, the numbers of leaves for the two leagues are about equal. The lower 50's has more values in the National League, but the American League makes up for this by having more values in the upper 50's and 60's.

Back-to-back stem-and-leaf plots are useful for comparing two sets of data. Before making comparisons, however, check to see first that both sets have about the same total number of values. Also, make sure that the plot is drawn accurately with each leaf taking up the same amount of space. These checks are important because we make the comparisons mainly through comparing the numbers of leaves on both sides. If one side has more data values or each leaf takes more space on one side than on the other, it can be hard to make accurate comparisons. To get the sizes correct, it helps to construct the plot on graph paper.

To decide if one data set generally has larger values than the other, compare the number of leaves on the two sides for both the largest and smallest stems. Also, note if there are outliers or gaps in the data that are not the same on both sides, and whether or not the two sides have about the same shape.

Page 25: Application 6

1. 3.4
2. 29,375,000,000
3.

WEST OF THE MISSISSIPPI		EAST OF THE MISSISSIPPI
	1	
	·	6 7 7 8
2 2 2 1 1	2	1 1 1 1 1 2 3 3 3 4
7 7 6 6 6 6 5 5 5	·	5 5 6 8 8 9
2 2 2 0	3	0 1 2 3 4
9 8	·	
3 0	4	1 4
	·	

3|2 REPRESENTS 3.2 TRAFFIC DEATHS PER 100 MILLION MILES

Application 6

Traffic Deaths

The table below lists the 50 states and the District of Columbia with the number of deaths in 1983 per 100 million vehicle miles driven.

Motor Vehicle Traffic Deaths by State per 100 Million Vehicle Miles

Alabama	3.2	Montana	4.0
Alaska	3.9	Nebraska	2.1
Arizona	2.6	Nevada	3.8
Arkansas	3.2	New Hampshire	2.6
California	2.6	New Jersey	1.7
Colorado	2.6	New Mexico	4.3
Connecticut	2.1	New York	2.5
Delaware	2.3	North Carolina	2.8
District of Columbia	1.8	North Dakota	2.1
Florida	3.3	Ohio	2.1
Georgia	3.1	Oklahoma	2.7
Hawaii	2.2	Oregon	2.7
Idaho	3.2	Pennsylvania	2.4
Illinois	2.3	Rhode Island	1.6
Indiana	2.5	South Carolina	3.4
Iowa	2.5	South Dakota	2.6
Kansas	2.2	Tennessee	2.9
Kentucky	3.0	Texas	3.0
Louisiana	4.3	Utah	2.5
Maine	2.8	Vermont	2.3
Maryland	2.1	Virginia	2.1
Massachusetts	1.7	Washington	2.2
Michigan	2.1	West Virginia	4.4
Minnesota	1.8	Wisconsin	2.2
Mississippi	4.1	Wyoming	3.2
Missouri	2.5		

Source: National Safety Council.

1. If a state had 685 traffic deaths for 20,000,000,000 vehicle miles, what rate would be listed in the table above?
2. Alabama had a total of 940 auto deaths in 1983. How many miles were driven in Alabama that year?
3. How do the states east of the Mississippi River compare with the states west of it? To decide, construct a back-to-back stem-and-leaf plot with the stems spread out. You may want to use the map on page 15. Leave the values for Minnesota and Louisiana off the plot as the river goes through both states.

4. Which states east of the Mississippi River might be considered outliers?
5. Which state west of the Mississippi River has the highest traffic death rate? Would you call it an outlier?
6. Do states in the east or the west generally have larger traffic death rates?
7. Summarize what you learned from this back-to-back stem-and-leaf plot.
8. What factors do you think might help to explain the difference between the east and the west?
9. (For class discussion) How could these data have been collected?

Page 26: Application 6 (*continued*)

4. West Virginia and Mississippi
5. New Mexico; no
6. west
7. Answers will vary. Sample: States west of the Mississippi River tend to have a slightly higher number of deaths per 100 million miles driven than do states east of the Mississippi River, although there is a substantial overlap in the two distributions. However, two eastern states, West Virginia with 4.4 and Mississippi with 4.1, have death rates about the same as the two highest western states, New Mexico with 4.3 and Montana with 4.0. These two eastern states are outliers in the east. It is interesting that all four of these states are mainly rural.

   The lowest death rates are 1.6 in Rhode Island, 1.7 in New Jersey and Massachusetts, and 1.8 in the District of Columbia. It is interesting that these four regions are relatively urbanized.
8. Some possible answers are: higher speeds on long, straight western roads; smaller cars (Japanese) are driven more frequently in the west and these cars are more dangerous; lawless attitude in the west. (Encourage students to investigate their hypotheses.)

   One factor that does *not* explain the difference is that people drive more in the west, as these death rates are given per 100 million miles driven.

   An interesting question is whether or not this tendency for the rates to be a bit higher in the west really holds up, or if it is something that just happened by chance in 1983. To try to answer this question, students could obtain data for more recent years and make the comparisons.
9. To obtain the rates, you must know both the number of deaths and the total vehicle miles. It seems reasonable that states would keep records of the total traffic deaths, but how could anyone know the total vehicle miles driven? One possibility is through using the state gasoline tax receipts. These receipts would give the total number of gallons of gasoline sold; then by using some average miles-per-gallon figure for cars in the state, you could estimate the miles driven. Would you want to use the same miles-per-gallon figure for all states?

Stem-and-Leaf Plots Where the Data Should be Truncated

The following table lists the buildings in San Francisco that are over 360 feet tall.

Building	Height in Feet
Transamerica Pyramid	853
Bank of America	778
101 California Street	600
5 Fremont Center	600
Embarcadero Center, Number 4	570
Security Pacific Bank	569
One Market Plaza, Spear Street	565
Wells Fargo Building	561
Standard Oil	551
One Sansome-Citicorp	550
Shaklee Building	537
Aetna Life	529
First & Market Building	529
Metropolitan Life	524
Crocker National Bank	500
Hilton Hotel	493
Pacific Gas & Electric	492
Union Bank	487
Pacific Insurance	476
Bechtel Building	475
333 Market Building	474
Hartford Building	465
Mutual Benefit Life	438
Russ Building	435
Pacific Telephone Building	435
Pacific Gateway	416
Embarcadero Center, Number 3	412
Embarcadero Center, Number 2	412
595 Market Building	410
101 Montgomery Street	405
California State Automobile Association	399
Alcoa Building	398
St. Francis Hotel	395
Shell Building	386
Del Monte	378
Pacific 3-Apparel Mart	376
Meridien Hotel	374

Source: *The World Almanac and Book of Facts*, 1985 edition.

The shortest building, the Meridien Hotel, is 374 feet tall. The tallest, the Transamerica Pyramid, is 853 feet tall. Start the stem-and-leaf plot as follows:

```
3 |
. |
4 |
. |
5 |
. |
6 |
. |
7 |
. |
8 |
. |
```

To place the 778-feet tall Bank of America Building on the plot, truncate (cut off) the last digit. This leaves 77, which goes on the plot as follows:

```
3 |
. |
4 |
. |
5 |
. |
6 |
. |
7 |
. | 7
8 |
. |
```

The finished plot follows.

```
                    3 |
                    . | 7 7 7 8 9 9 9
                    4 | 0 1 1 1 1 3 3 3
                    . | 6 7 7 7 8 9 9
                    5 | 0 2 2 2 3
                    . | 5 5 6 6 6 7
                    6 | 0 0
                    . |
                    7 |
                    . | 7
3|7 REPRESENTS      8 |
370-379 FEET        . | 5
```

Page 29: Discussion Questions

1. 850 feet to 859 feet
2. city building code limitations on height; fear of earthquakes
3.

LOS ANGELES		SAN FRANCISCO
	3	
99966666	·	7778999
31	4	01111333
555	·	6777899
1	5	02223
7	·	556667
22	6	00
9	·	
3	7	
5	·	7
	8	
5	·	5

3|7 REPRESENTS 370-379 FEET

Discussion Questions

1. What heights can 8|5 represent?
2. The heights of all but two buildings stop abruptly at 600 feet. Can you think of a possible explanation for this?
3. The following table lists Los Angeles buildings taller than 360 feet.

Building	Height in Feet
First Interstate Bank	858
Crocker Center, North	750
Security Pacific National Bank	735
Atlantic Richfield Plaza (2 buildings)	699
Wells Fargo Bank	625
Crocker-Citizen Plaza	620
Century Plaza Towers (2 buildings)	571
Union Bank Square	516
City Hall	454
Equitable Life Building	454
Transamerica Center	452
Mutual Benefit Life Insurance Building	435
Broadway Plaza	414
1900 Avenue of Stars	398
1 Wilshire Building	395
The Evian	390
Bonaventure Hotel	367
400 South Hope Street	365
Beaudry Center	365
California Federal Savings & Loan Building	363
Century City Office Building	363

Source: *The World Almanac and Book of Facts*, 1985 edition.

Complete this back-to-back stem-and-leaf plot for the two cities.

LOS ANGELES		SAN FRANCISCO
	3	
	·	
	4	
	·	
	5	
	·	
	6	
	·	
	7	
	·	
	8	
	·	

Notice that San Francisco has 37 tall buildings, while Los Angeles has only 21. We don't need a stem-and-leaf plot to tell us that San Francisco has more tall buildings than Los Angeles. This plot can, however, help us answer the question of which city's buildings are relatively taller, apart from the total numbers of tall buildings. Unlike the last section, we cannot just look at the number of leaves, since San Francisco has more values and thus will generally have more leaves for each stem. Instead, we need to compare the two *shapes*, making a mental adjustment for the fact that San Francisco has about twice as many data values. Follow this procedure to answer the following question.

4. Considering only buildings over 360 feet tall, does Los Angeles or San Francisco tend to have relatively taller buildings?
5. In the previous stem-and-leaf plots, both the San Francisco and Los Angeles heights were truncated. Instead of truncating, we will now *round* each height to the nearest ten. Then we will see if the back-to-back stem-and-leaf plot gives the same impression as before. The San Francisco side of the plot below was made by rounding. Copy the plot and complete the Los Angeles side using rounding. The symbol 3|7 now represents 365-374 feet.

LOS ANGELES		SAN FRANCISCO
	3	
	·	7 8 8 9
	4	0 0 0 1 1 1 1 2 4 4 4
	·	7 7 8 8 9 9 9
	5	0 2 3 3 4
	·	5 5 6 7 7 7
	6	0 0
	·	
	7	
3|7 REPRESENTS	·	8
	8	
365 – 374 FEET	·	5

6. Is it faster to round or to truncate?
7. Does the back-to-back stem-and-leaf plot with rounded numbers give the same general impression as the one with truncated numbers? Are there any differences in what you learn from the two plots?
8. Do you think truncating is an appropriate procedure, or should the data be rounded?

Page 30: Discussion Questions (*continued*)

4. About the same, in general. More precisely, though, for the heights between 360 and 600 feet, the San Francisco heights are fairly rectangular-shaped, while the Los Angeles heights are more backwards J-shaped, with more at the low end. Thus, for heights in this range, the San Francisco ones tend to be a bit larger. For heights above 600 feet it is hard to say—there is not much difference.

5.

LOS ANGELES		SAN FRANCISCO
	3	
9 7 7 7 6 6	·	7 8 8 9
4 1 0 0	4	0 0 0 1 1 1 1 2 4 4 4
5 5 5	·	7 7 8 8 9 9 9
2	5	0 2 3 3 4
7	·	5 5 6 7 7 7
3 2	6	0 0
	·	
4 0	7	
3|7 REPRESENTS 5	·	8
	8	
365 – 374 FEET 6	·	5

6. truncate (Also, we make fewer mistakes.)
7. yes; no
8. Answers will vary. Students often object to truncating. We think that truncating is generally OK when making stem-and-leaf plots.

If you are like many students, you may feel that there is something wrong about truncating. It seems less accurate than rounding, and therefore worse. But is using 3|7 to represent 365-374 feet really more accurate for our purposes than using 3|7 to represent 370-379 feet?

Another point to consider is that the data we have may already be either rounded or truncated, and we don't know which. Are all the building heights exact multiples of one foot, with no inches or fractions of inches, as listed in the tables?

Finally, it is easy to make a mistake when rounding. In order to truncate, all we do is use a straightedge to cover the columns of digits not needed. To decide if truncating is appropriate for a specific problem, ask yourself if it is likely to make any difference in the interpretations you reach.

Application 7

Children's Books

The following table lists the children's books published in the U.S. since 1895 that have sold one million or more copies.

Green Eggs and Ham, by Dr. Seuss. 1960	5,940,776
One Fish, Two Fish, Red Fish, Blue Fish, by Dr. Seuss. 1960	5,842,024
Hop on Pop, by Dr. Seuss. 1963	5,814,101
Dr. Seuss' ABC, by Dr. Seuss. 1963	5,648,193
The Cat in the Hat, by Dr. Seuss. 1957	5,394,741
The Wonderful Wizard of Oz, by L. Frank Baum. 1900	(estimate) 5,000,000
Charlotte's Web, by E. B. White. 1952	4,670,516
The Cat in the Hat Comes Back, by Dr. Seuss. 1958	3,431,917
The Little Prince, by Antoine de Saint-Exupery. 1943	2,811,478
The Little House on the Prairie, by Laura Ingalls Wilder. 1953 edition	2,732,666
The Little House in the Big Woods, by Laura Ingalls Wilder. 1953 edition	2,527,203
My First Atlas. 1959	2,431,000
Love and the Facts of Life, by Evelyn Duvall and Sylvanus Duvall. 1950	2,360,000
Egermeier's Bible Story Book, by Elsie E. Egermeier. 1923	2,326,577
Go Ask Alice, Anonymous. 1971	2,245,605
Benji, by Leonore Fleischer. 1974	2,235,694
The Little Engine That Could, by Watty Piper. 1926	2,166,000
Stuart Little, by E. B. White. 1945	2,129,591
Freckles, by Gene Stratton Potter. 1904	2,089,523
The Girl of the Limberlost, by Gene Stratton Porter. 1909	2,053,892
Sounder, by William Armstrong. 1969	1,815,401
Harry, the Dirty Dog, by Gene Zion. 1956	1,690,339
Seventeen, by Booth Tarkington. 1916	(estimate) 1,682,891
Where the Wild Things Are, by Maurice Sendak. 1963	1,632,020
Laddie, by Gene Stratton Porter. 1913	1,586,529
The Big Book of Mother Goose. 1950	1,500,000
The Golden Dictionary, by Ellen Wales Walpole. 1944	1,450,000
A Friend is Someone Who Likes You, by Joan Walsh Anglund. 1958	1,423,432
Rebecca of Sunnybrook Farm, by Kate Douglas Wiggin. 1904	1,357,714
Love Is a Special Way of Feeling, by Joan Walsh Anglund. 1960	1,308,293
The Real Mother Goose. 1915	1,296,140
The Pigman, by Paul Zindel. 1968	1,265,876
Better Homes and Gardens Story Book. 1951	1,220,728
Trouble after School, by Jerrold Beim. 1957	1,145,570
Better Homes and Gardens Junior Cook Book. 1955	1,100,182
Pollyanna, by Eleanor H. Porter. 1913	1,059,000
Le Petit Prince, by Antoine de Saint-Exupery. 1943	1,018,373
Mary Poppins, by Pamela L. Travers. 1934	1,005,203
Winnie-the Pooh, by A. A. Milne. 1926	1,005,000
Pollyanna Grows Up, by Eleanor H. Porter. 1915	1,000,000
Little Black Sambo, by Helen Bannerman. 1899	(estimate) 1,000,000

Source: A. P. Hackett and J. H. Burke, *Eighty Years of Best Sellers.*

Page 33: Application 7

1.

```
1 | 0 0 0 0 0 0 1 1 2 2 2 3 3 4 4
. | 5 5 6 6 6 8
2 | 0 0 1 1 2 2 3 3 4
. | 5 7 8
3 | 4
. |
4 |
. | 6
5 | 0 3
. | 6 8 8 9
```

1|0 REPRESENTS 1,000,000 THROUGH 1,099,999 BOOKS SOLD

2. See the preceding plot.
3. Answers will vary; at the bottom; the books at the bottom have sold more copies than those at the top.
4. About twice as long as the top line; because that is the trend in the plot so far. There should be many more books in that category.
5. Answers will vary. Sample: Forty-one children's books published in the United States since 1895 had sold 1 million or more copies by 1977. There is a cluster of seven books at the higher end of sales, separated from the rest, that have all sold more than 4½ million copies. Of these seven, five are by Dr. Seuss, and they are the top five. (He must be very rich.) The other two are *The Wonderful Wizard of Oz* and *Charlotte's Web.* The remaining books have sold between 1 million and 3.5 million copies, with only four over 2.5 million.

   It is also interesting that only two of these 41 books were published since 1970. Is the reason just that fewer best sellers were written recently, or does it perhaps take a long time for a children's book to become a best seller, or are perhaps fewer books being sold now than earlier?

1. Make a stem-and-leaf plot of these data using these stems. *Green Eggs and Ham* has been placed on the plot to get you started. Truncate all digits except those in the millions and hundred-thousands places.

```
1 |
. |
2 |
. |
3 |
. |
4 |
. |
5 |
. | 9
```

1|0 REPRESENTS 1,000,000 THROUGH 1,099,999 BOOKS SOLD

2. Underline all digits representing books by Dr. Seuss.
3. Circle the digits representing the books you have read. Do these circles tend to be at the top or the bottom of the diagram? Why?
4. If another line were added to the top of the plot for books that sold 500,000-999,999 copies, how long do you think it would be? Why?
5. Write a summary of the information displayed in the plot.

Stem-and-Leaf Plots — Summary

Stem-and-leaf plots are a new way to quickly organize and display data. Unlike line plots, they are best used when there are more than 25 pieces of data. Statisticians use stem-and-leaf plots as a substitute for the less informative histograms and bar graphs.

Variations of stem-and-leaf plots that you should know how to construct are as follows:

- back-to-back
- truncated and rounded
- spread out

From a stem-and-leaf plot it is easy to identify the largest and smallest values, outliers, clusters, gaps, the relative position of any important value, and the shape of the distribution.

Suggestions for Student Projects

1. Collect data on a topic that interests you, make a stem-and-leaf plot, and then write a summary of the information displayed in the plot. Use one of the topics listed below or think of your own.

   a. Compare the ages in months of the boys and the girls in your class.

   b. Compare the heights of the boys and the girls in your class.

   c. Compare the heights of the buildings in two cities near you.

   d. Compare the gas mileage of foreign and domestic cars. (This information can be found in many almanacs.)

   e. Compare the scores of two different classes taking the same math test.

The next two projects involve comparing line plots with stem-and-leaf plots.

2. Devise a way to use symbols in a line plot to replace the individual data values, as we did for the stem-and-leaf plots in the fast foods and thunderstorm examples. Then, construct a line plot for one of these examples, using your method. Do the line and stem-and-leaf plots show any different information? Which is easier to interpret? Which do you prefer?

3. Devise a way of modifying a line plot to get a back-to-back line plot. Then, redo Application 6, or the building heights example, using your back-to-back line plot. Which is easier to construct, the back-to-back line plot or the stem-and-leaf plot? Do they show any different information? Which shows the information more clearly? Which do you prefer? Can you think of situations in which you might prefer the other plot?

4. In order to compare truncating and rounding, take any of the data in this section and make a back-to-back stem-and-leaf plot of the truncated against the rounded values. Do you see any difference, and if so what is it? Could you have predicted this?

5. In the fast foods example at the beginning of this section, we showed the type of food in the stem-and-leaf plot by replacing the leaves by letters. A way to show both the specific numerical values and labels is to keep the numerical leaf in the plot, and follow it by a label in parentheses. For instance, the next-to-bottom row in the fast foods example would be $3|3(H), 4(F)$. By keeping the number in the plot, we retain as much detailed numerical information as is generally needed. This idea is especially useful for displaying data where there is one number for each of the 50 states. The two-letter postal abbreviation can be used to identify each state. Find some interesting data where there is one value for each state. A good example would be each state's current population as found in an almanac. Make the plot just described, and write a summary of the information displayed.

Page 35: Discussion Questions

1. 78.3
2. C

III. MEDIAN, MEAN, QUARTILES, AND OUTLIERS

Median and Mean

You have probably learned how to compute the average of a set of numbers. For example, if Sally gets scores of 80, 96, 84, 95, and 90 on five math tests, then her average is:

$$\frac{80 + 96 + 84 + 95 + 90}{5}$$

$$= \frac{445}{5}$$

$$= 89.$$

Whenever we compute an average this way, we will call it the *mean*. Thus, the mean of Sally's test scores is 89. We need a new word for the average because there are other kinds of averages. Another type of average is the *median*. To find the median of Sally's test scores, first put them in order from smallest to largest.

80 84 [90] 95 96

The middle score, 90, is the median. Half of Sally's five test scores are lower than or equal to the median and half are higher than or equal to the median.

What do you do if there is an even number of scores? If Sally takes a sixth test and gets a 25, her scores are now:

25 80 [84 90] 95 96.

There are two scores in the middle, 84 and 90. The median is halfway between these two scores:

$$\frac{84 + 90}{2}$$

$$= \frac{174}{2}$$

$$= 87.$$

Half of her six test scores are lower than 87 and half are higher.

Discussion Questions

1. Compute the mean of Sally's six test scores. (Round to the nearest tenth.)
2. On the basis of this grading scale what grade would Sally receive if the mean of the six tests is used to determine her grade?

   A 90-100 B 80-89 C 70-79 D 60-69 E 0-59

3. What grade would she receive if the median of the six tests is used to determine her grade?
4. Does one extreme score cause a greater change in the median or in the mean?
5. Do you need to know all of the data values in order to find the median? For example, suppose that Sally has taken 6 tests and you only know 5 of her scores. Can you calculate the median?
6. Give a reason for choosing the median to summarize Sally's test scores.
7. Give a reason for choosing the mean to summarize Sally's test scores.
8. Which do you think is better to use, the mean or median?
9. Why do you think the median is generally used when discussing ages, average house prices, or average incomes, as in the following newspaper and magazine examples?
   a. "When only first-time marriages were considered, the agency [National Center for Health Statistics] placed the median age for brides at 21.8 years in 1980, up from 20.3 years in 1963. The median age for bridegrooms was 23.6 years, up from 22.5 years in 1963." *(Los Angeles Times 2/17/84)*
   b. According to the Census Bureau, "the counties with the highest median value of owner-occupied dwellings are: Pitkin, CO. - $200,000; Marin, CA - $151,000; Honolulu, HI - $130,400; San Mateo, CA - $124,400; Maui, HI - $113,600." *(USA Today 3/8/84)*
   c. According to the Census Bureau, "the median time spent on homework for students in American elementary and high schools was 5.4 hours a week . . . the sharpest difference was between types of schools, with students in private high schools doing 14.2 hours of homework weekly, as against 6.5 hours by their public school counterparts." *(The New York Times 11/29/84)*
   d. "The following drawing shows typical allowances (rounded to the nearest 25¢) for 8-to-13-year-olds, as reported by the 811 students in our survey who received allowances. The allowances of the 8-to-11-year-olds are all pretty much the same. They range from $2.00 to $2.75. But for the 12-year-olds, there's a jump of $1, and an even bigger jump for kids one year older.

      "The figures don't mean that all the three hundred thirty-eight 11-year-olds in our survey who receive an allowance are pocketing $2.75 every week. That $2.75 is the *median* allowance for that age. Median means right in the middle. Half the 11-year-olds are getting more than $2.75, and half are getting less. In fact, one-third report a weekly allowance of under $2, and about the same amount get more than $4 a week.

      169 get less $2.75   169 get more

      "The amount of your allowance seems to depend a lot on your age. But where you live and whether you are a boy or a girl do *not* seem to affect how much you get per week. Students all across

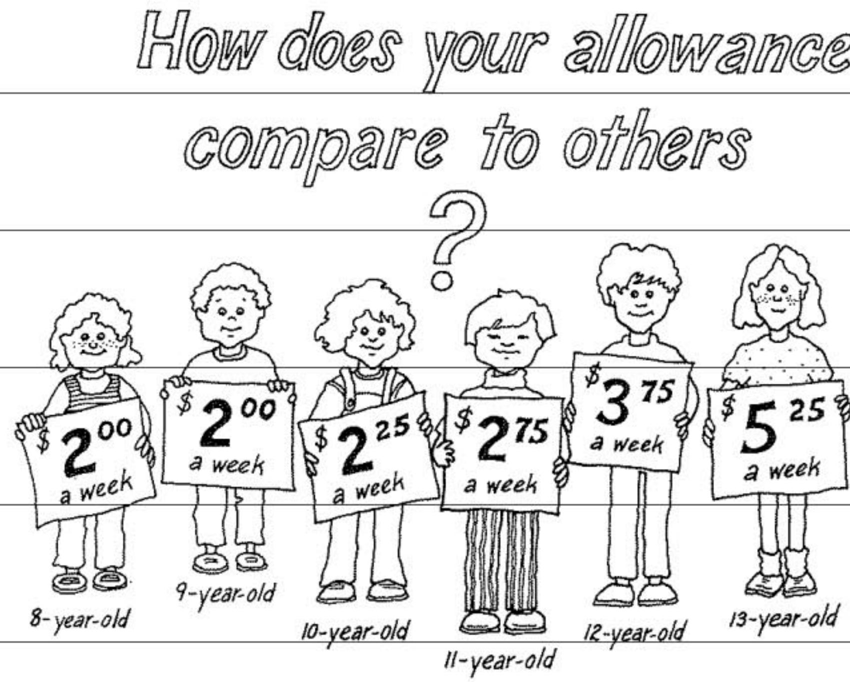
Page 36: Discussion Questions (*continued*)

3. B
4. in the mean
5. yes; no
6. Answers will vary. Sample: It does not penalize her so much for the one low score. She might have been ill.
7. Answers will vary. Sample: Using the mean will penalize Sally for the test score of 25. If she didn't study or didn't understand the material, this may be desirable.
8. Answers will vary. Sample: The teacher might prefer the mean so that the student will study for each test, while the student might prefer the median.
9. Answers will vary. Sample: With ages, incomes, and housing prices, the values tend to have a J-shaped distribution. That is, there will probably be some values much larger than the rest. These values tend to make the mean large. For example, most women first marry around ages 18 to 26, but there are first-time brides in their fifties and older. If we don't want a measure that would be affected greatly by a few extreme values, the median is more useful than the mean.

Page 37: Discussion Questions (*continued*)

10. mean; because the story says the average was increased by a few extraordinarily large awards.

the country, in cities and small towns, said they received pretty much the same amount. Boys and girls also reported similar allowances." (*Penny Power* 2/3/83)



10. In the following newspaper story, what do you think is the meaning of the word "average"? Give your reasons.

"[In a study of jury awards in civil trials, they] found that while the average award against corporate defendants was more than $120,000, the average against individuals was $18,500. The average against government defendants was $38,000, but it was $97,000 in cases that involved hospitals and other nonprofit entities.

'To some degree, the average awards against corporations and hospitals were so great because of a few extraordinarily large awards,' the report explained." (*Newark Star-Ledger* 8/20/85)

11. The following information seems to be incorrect.

"According to the latest enrollment analysis by age-categories, half of the [Los Angeles Community College] district's 128,000 students are over the age of 24. The average student is 29." (*Los Angeles Times* 9/20/81)

"In the region we are traveling west of Whitney, precipitation drops off and the average snow depth on April 1 for the southern Sierra is a modest 5 to 6 feet. And two winters out of three, the snow pack is below average." Ezra Bowen, *The High Sierra* (New York: Time-Life Books, 1972), p. 142.

a. Give an example of four students with a mean age of 29 and median age of 24.

b. Give an example of the snow depth for three winters that makes the quote from *The High Sierra* true.

Both the median and the mean summarize the data by giving a measure of the center of the data values. *For the kinds of data in this book,* the median generally gives a more reasonable summary since it is not affected by a few extreme values. When there are no outliers, there will generally not be much difference between the median and mean, and which we choose won't matter. Using a calculator, the mean is easy to compute. To find the median, however, the data must be ordered from smallest to largest. This can be tedious, but an easy method is to construct a stem-and-leaf plot.

Neither the median nor the mean can tell us as much about the data as a plot showing all the values, such as a line plot or a stem-and-leaf plot.

Page 38: Discussion Questions (*continued*)

11. a. Answers will vary. For example, 18, 24, 24, and 50.
    b. Answers will vary. For example, 2 feet, 2 feet, and 12 feet.

Page 39

NOTE TO TEACHERS: Either Application 8, "How Many Moons," or Application 9, "The Pop Meter," may be omitted.

Application 8

1. 6.67
2. 2
3. Jupiter, Saturn, and Uranus; they are large planets and it may be easier for them to hold on to moons.
4. Answers will vary. Sample: The median is better as the mean is large due to Jupiter, Saturn, and Uranus. Six of the planets are within two moons of the median, but none is within two moons of the mean. Neither the mean nor the median is really adequate. There are so few values, it would be best just to give her the table with an explanation about Jupiter, Saturn, and Uranus.

How Many Moons?

A visitor from the star Alpha Centauri has selected you to provide her with information about our solar system. She is filling out a form and asks how many moons are "average" for a planet in our solar system.

Study the table below.

Planet	Number of Moons
Mercury	0
Venus	0
Earth	1
Mars	2
Jupiter	16
Saturn	23
Uranus	15*
Neptune	2
Pluto	1

Source: The World Book, 1984.
*The published figure is 5 moons, but in January 1986, Voyager 2 discovered 10 additional moons around Uranus.

1. Compute the mean number of moons.
2. Compute the median number of moons.
3. Which three planets are the most different in number of moons compared to the others? Do you know any explanation for this?
4. Do you think the visitor from Alpha Centauri would get a more accurate impression about the typical number of moons from the median or the mean? Is either summary number adequate? Give your reasons.

Next, the visitor asks about the length of a typical day in our solar system. Study the following table.

Planet	Approximate Length of a Day in Earth Hours
Mercury	1416
Venus	5832
Earth	24
Mars	24.5
Jupiter	10
Saturn	11
Uranus	22
Neptune	16
Pluto	153

5. Compute the mean length of a day in our solar system in hours.
6. How many Earth days is this?
7. Find the median length of a day in our solar system.
8. Do you think it is better to give your visitor the mean length of a day or the median length of a day? Why? Are you happy about giving your visitor one single number? Why or why not?

Page 40: Application 8 (*continued*)

5. 834.28 hours
6. 34.76 days
7. 24 hours
8. Answers will vary. Sample: The median is better as only two planets, Mercury and Venus, have days longer than the mean of 834.3 hours. The very long days of these two planets make the mean large. Using the median, however, hides the fact that there are one long and two very long days. Thus neither the mean nor the median tells everything about these numbers.

Page 41: Application 9

1. Album	Mean	Median
"Little Creatures"	84.7	85
"Who's Zoomin' Who?"	80.0	82
"Youthquake"	49.7	50
"Boy in the Box"	48.6	51
"Invasion of Your Privacy"	44.7	27

2. See the preceding chart.
3. a. "Invasion of Your Privacy"
   b. Molly Ringwald, who gave this album a much higher rating than the regular reviewers
   c. median
4. a. Willman (42.8)
   b. Willman (27)
   c. Molly Ringwald, who gave the highest rating on four of the five albums

Application 9

The Pop Meter

Six of the pop music reviewers for the *Los Angeles Times* and a teenage actress and singer, Molly Ringwald, rated five new albums as follows:

Albums	Dennis Hunt	Lori E. Pike	Richard Cromelin	Connie Johnson	Chris Willman	Patrick Goldstein	Molly Ringwald
"Little Creatures" Talking Heads	75	84	85	75	88	91	95
"Who's Zoomin' Who?" Aretha Franklin	86	82	70	83	62	79	98
"Youthquake" Dead or Alive	78	72	50	30	12	36	70
"Boy in the Box" Corey Hart	60	60	20	49	25	51	75
"Invasion of Your Privacy" Ratt	65	20	20	25	27	66	90

The ratings system: 90-100, excellent; 70-89, good; 50-69, fair; 30-49, weak; 0-29, melt down.

Source: *Los Angeles Times*, September 1, 1985.

1. Compute the mean rating for each album.
2. Compute the median rating for each album.
3. a) For which album are the mean and median farthest apart?
   b) Which reviewer caused this?
   c) Is the mean or the median more representative of this album's overall rating?
4. a) If you judge by the mean rating, which reviewer is the hardest grader?
   b) If you judge by the median rating, which reviewer is the hardest grader?
   c) Which reviewer tends to be the most different from the others?

Range, Quartiles, and Interquartile Range

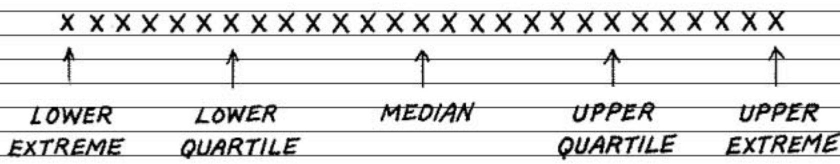
The number of grams of carbohydrates (starch and sugar) in a 1-ounce serving of thirteen breakfast cereals is shown below.

Cereal	Carbohydrates	Cereal	Carbohydrates
Life	19	Grape Nuts	23
Super Sugar Crisp	26	Special K	21
Rice Krispies	25	Raisin Bran	28
Product 19	24	Wheaties	23
Total	23	Puffed Rice	13
Sugar Corn Pops	26	Sugar Smacks	25
		Cheerios	20

To find the *range*, subtract the smallest number from the largest. The range for the carbohydrates is:

$$28 - 13 = 15 \text{ grams.}$$

We will also learn how to find the *lower quartile* and the *upper quartile*. If the numbers are arranged in order from smallest to largest, the lower quartile, the median, and the upper quartile divide them into four groups of roughly the same size.



To find the quartiles of the previous numbers, first arrange the numbers in order:

13 19 20 21 23 23 23 24 25 25 26 26 28

Second, find the median and draw a vertical line through it.

13 19 20 21 23 23 |23| 24 25 25 26 26 28

The median is 23. Six numbers are below this 23 and six are above it.

Third, consider only the data values to the left of the line.

13 19 20 | 21 23 23

The median of these six numbers is between 20 and 21. This is the lower quartile. Thus, the lower quartile is 20.5. We have drawn a vertical line at the median of these values in the same way as before.

Page 42

NOTE TO TEACHERS: In this section, students will learn to measure how spread out the data are. The measure they will learn, the interquartile range, is, like the median, resistant to outliers. In addition it is easier for students to compute, comprehend, and interpret than the more sophisticated standard deviation.

Page 43

NOTE TO TEACHERS: In some textbooks, the median is included when finding the quartiles. For example, when finding the upper quartile of the cereal data, these textbooks would find the median of 23, 24, 25, 25, 26, 26, and 28 and would get 25.

Discussion Questions

1. no
2. no
3. Median is 23.5 and quartiles are 22 and 25.5.
4. lower quartile by 1.5, median by 0.5, and upper quartile by 0
5. 9; 3.5

Finally, consider only the data values to the right of the line and find their median. This is the upper quartile. The upper quartile is 25.5.

24 25 25 | 26 26 28

We have divided the numbers into four groups:

13 19 20 | 21 23 23 | 23 24 25 25 | 26 26 28

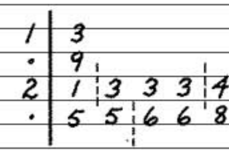
Notice that there are three numbers in each group.

The *interquartile range* is the difference between the upper quartile and the lower quartile. The interquartile range of the given numbers is:

$$25.5 - 20.5 = 5.$$

The *lower extreme* is the smallest value in the data. In this case, it is 13. Similarly, the *upper extreme* is the largest number in the data. In this case, it is 28.

The fastest way to order the numbers from smallest to largest is to make a stem-and-leaf plot of the data, with the leaves ordered. Then, count in from the top and bottom to mark the median and quartiles. As an example, suppose we did not have Cheerios in the list of cereals and we wanted the median and quartiles of the remaining 12 cereals. The median will then be between the sixth and seventh values. We draw the first line there and consider only the data values below and above this line, as before, to get the quartiles.



The vertical lines here are dotted. The median is 23.5, the lower quartile is 22, and the upper quartile is 25.5.

Discussion Questions

1. In these data, the median is the mean of the quartiles. Will the median always be the mean of the quartiles?
2. Is the interquartile range half of the range?
3. Cross the 13 grams from Puffed Rice off the list and find the new median and quartiles.
4. By how much did these values change?
5. Recompute the range and interquartile range.

6. By how much did these values change?
7. Find two different sets of seven numbers with:

   lower extreme - 3
   lower quartile - 5
   median - 10
   upper quartile - 12
   upper extreme - 13

8. The median is always between the two quartiles. Do you think the *mean* is always between the two quartiles?
9. Find a set of seven numbers where the mean is above the upper quartile.
10. Find a set of seven numbers where the mean is below the lower quartile.

Page 44: Discussion Questions (*continued*)

6. range by 6 and interquartile range by 1.5
7. Answers will vary. Examples are: 3, 5, 7, 10, 11, 12, 13 and 3, 5, 8, 10, 12, 12, 13.
8. no
9. Answers will vary. For example: 1, 2, 3, 4, 5, 6, 40
10. Answers will vary. For example: 1, 8, 8, 8, 9, 10, 11

Page 45: Application 10

1. BMX Freemag
2. BMX 34 Open Road

Application 10

Motocross Bike Ratings

The list below contains the ratings by *Penny Power* magazine of 22 motocross bikes.

Rating	Brand	Model	Price
Very Good	Raleigh	R-10 TUFF BMF	$190
Very Good	Raleigh	R-10 MK III	$150
Very Good	Schwinn	B43 Scrambler	$196
Very Good	Mongoose	BMX Wirewheel	$190
Very Good	Mongoose	BMX Freemag	$215
Good	Vista	GTX99	$125
Good	J.C.Penney	Eagle V	$190
Fair	Ross	142-25 THX	$165
Fair	Ross	Slinger	$125
Fair	Sears	Free Spirit BMX FS500	$150
Fair	Schwinn	B511 Thrasher	$143
Fair	Sears	BMX FS100	$100
Fair	Murray	X-20 Team Murray	$141
Fair	AMF	Hawk 4 BMX	$139
Fair	Huffy	Pro Thunder BMX	$160
Fair	Columbia	Pro Am 2236	$160
Poor	Murray	Team Murray BMX	$130
Poor	J.C.Penney	Dirt Tracker II	$110
Poor	Wards	BMX 34 Open Road	$80
Poor	AMF	Avenger Motocross	$100
Poor	Columbia	Formula 16 BMX	$110
Poor	Huffy	Thunder BMX	$100

Source: *Penny Power,* February 3, 1983.

1. What is the most expensive bike?
2. What is the least expensive bike?

3. Find the median price of the bikes rated:
   a. very good
   b. good
   c. fair
   d. poor
4. In general, do bikes with a higher price have a higher rating?
5. What is the range of the bike prices?
6. Find the lower quartile for all bikes.
7. Find the upper quartile.
8. What is the interquartile range of the bike prices?
9. Which of the bikes rated "very good" is priced below the upper quartile? Is this bike a good buy?
10. Which of the bikes rated "poor" is priced above the lower quartile? Is this bike a good buy?

Outliers

The following table lists all 15 records that reached number 1 for the first time in 1959, and the total number of weeks that each record held the number 1 spot.

Weeks	Record Title	Artist
3	"Smoke Gets in Your Eyes"	Platters
4	"Stagger Lee"	Lloyd Price
5	"Venus"	Frankie Avalon
4	"Come Softly to Me"	Fleetwoods
1	"The Happy Organ"	Dave 'Baby' Cortez
2	"Kansas City"	Wilbert Harrison
6	"The Battle of New Orleans"	Johnny Horton
4	"Lonely Boy"	Paul Anka
2	"A Big Hunk o' Love"	Elvis Presley
4	"The Three Bells"	Browns
2	"Sleep Walk"	Santo & Johnny
9	"Mack the Knife"	Bobby Darin
1	"Mr. Blue"	Fleetwoods
2	"Heartaches by the Number"	Guy Mitchell
1	"Why"	Frankie Avalon

Source: *The Billboard Book of Top 40 Hits*, 1985.

Page 46: Application 10 (*continued*)

3. a. $190
   b. $157.50
   c. $143
   d. $105
4. yes
5. $135
6. $110
7. $165
8. $55
9. R-10 MK III; yes
10. Team Murray MBX; no

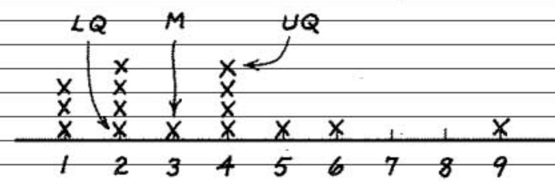
Page 47

NOTE TO TEACHERS: The 1.5 × IQR rule for finding outliers can be interpreted as follows. If the data were all drawn from a normal (bell-shaped) distribution, then about 1 of every 100 observations would be so large or small as to be called an outlier according to this rule. More precisely, this rule defines an outlier for a normal distribution as any value more than about 2.7 standard deviations from the mean. In real data we almost always observe more than 1 percent outliers; the corollary is that real data generally do not follow a normal distribution.

We have already used the word *outlier* several times to indicate values that are widely separated from the rest of the data. Would you say that any record in the list above is an outlier? If we think we have spotted an outlier, it is worth some special thought about why it is different from the rest. Trying to make sense out of the outliers can be an important part of interpreting data.

It is not reasonable, however, to automatically call the upper and lower extremes outliers. Any data set has extremes, and we don't want to put extra energy into trying to interpret them unless they are separated from the rest of the data. We could decide if an observation is an outlier by looking at a plot and making a decision, as we have done so far. However, it is helpful to have a rule to aid in making the decision, especially when there are a moderate to large number of observations (say 25 or more).

Thus, we say that an *outlier* is any number more than 1.5 interquartile ranges above the upper quartile, or more than 1.5 interquartile ranges below the lower quartile. A line plot of the hit record data, with the median (M) and quartiles (LQ and UQ) labeled, follows.



The interquartile range (IQR) is $4 - 2 = 2$, so $1.5 \times IQR = 3$. Thus, the upper cut-off is $4 + 3 = 7$. Since the data value 9 ("Mack the Knife") is greater than 7, we call it an outlier. For the lower end, the cut-off is $2 - 3 = -1$. Since no data value can be less than -1, there are no outliers at the lower end. An interpretation we can draw is that "Mack the Knife" was not only the most popular record in 1959, but that it really stands out as substantially more popular than the other 14 top hits. Before doing this calculation, did you feel that "Mack the Knife" was an outlier?

The rule just described is quick, easy, and straightforward to use. Multiplying the IQR by 1.5 rather than 1.0 or 2.0 generally produces results that are what we would like, if we were to decide which values should be labeled outliers. You might experiment using multipliers such as 1.0, 1.5, and 2.0 to decide which you prefer.

Ice Cream Cone Prices

In September 1985, the prices of a single-scoop ice cream cone at 17 Los Angeles stores are given in the table below.

Store (brand)	Price
Andi's (homemade)	$ .90
Baskin-Robbins	.75
Carvel	.95
Cecelia's (Dreyers)	.90
Cinema Sweet (homemade)	1.20
Clancy Muldoon	.95
Creamery (homemade)	1.05
Farrell's	.70
Foster's Freeze	.53
Haagen-Dazs	1.10
Humphrey Yogart	.95
Leatherby's (homemade)	.91
Magic Sundae (Buds)	.96
Robb's (homemade)	.95
Swensons	1.00
Thrifty Drug	.25
Will-Wright's (own recipe)	1.15

1. Make a stem-and-leaf plot of the prices.
2. Are there any gaps in the prices? Where?
3. Find the median price of an ice cream cone using the stem-and-leaf plot.
4. Find the mean price of an ice cream cone.
5. Thrifty Drug's cone is much cheaper than the others. If it is taken off the list, do you think the median or the mean will increase the most?
6. Cross Thrifty Drug's price off the list before determining the following:
   a. Find the median price of the remaining cones.
   b. Find the mean price of the remaining cones.
   c. Which increased more, the median or the mean?
7. Find the range in prices. (Include Thrifty Drug from exercise 7 through 13).
8. Find the lower quartile of the prices.

Page 48: Application 11

1.

2	5
3	
4	
5	3
6	
7	0 5
8	
9	0 0 1 5 5 5 5 6
10	0 5
11	0 5
12	0

2|5 REPRESENTS 25¢

2. between 25¢ and 53¢, 53¢ and 70¢, and 75¢ and 90¢
3. 95¢
4. 89.4¢
5. mean
6. a. 95¢
   b. 93.4¢
   c. mean
7. 95¢
8. 82.5¢

Page 49: Application 11 (*continued*)

9. \$1.025
10. median and lower quartile
11. 20¢
12. $UQ + 1.5 \times IQR = 102.5 + 30 = 132.5$
    $LQ - 1.5 \times IQR = 82.5 - 30 = 52.5$
    Thrifty Drug is the only outlier (but Foster's Freeze is surely close).
13. Answers will vary. Sample: It's much cheaper. It is sold at a drug store rather than a specialty ice cream store. Maybe the ice cream is not as good, or maybe the cone is a lot smaller, or maybe it is priced cheaply to encourage people to come into the drug store with the hope they will also buy other items (in other words, it is a "loss leader").

9. Find the upper quartile of the prices.
10. Is there a larger difference between the median and the lower quartile or between the median and the upper quartile?
11. Find the interquartile range.
12. Use the 1.5 × IQR rule to find any outliers.
13. How is the outlier different from the others? Can you think of any possible explanations for this?

Median, Mean, Quartiles, and Outliers — Summary

Both the median and the mean are single numbers that summarize the location of the data. Neither alone can tell the whole story about the data, but sometimes we do want a single, concise, summary value. Generally, the median is more valuable than the mean, especially if there is any possibility of having even a few unusually large or small values in the data.

The lower quartile, median, and upper quartile divide the data into four parts with approximately the same number of observations in each part. The interquartile range (IQR), the third quartile minus the first quartile, is a measure of how spread out the data are. If a number is more than 1.5 times the interquartile range above the upper quartile or below the lower quartile, we call it an outlier. If the data are grouped fairly tightly, there will be no outliers. When we do find an outlier, we should study it closely. It is worthwhile to try to find reasons for it, as they can be an important part of the overall interpretation of the data.

Suggestions for Student Projects

1. Choose 5 or 6 current popular records. Your teacher should select 5 or 6 reviewers from students in your class. These reviewers will fill in ratings as in Application 9, and the entire class will analyze the results.
2. Find examples of the use of the words "mean," "median," or "average" in a local newspaper. If you find "average," can you tell if they used the median, the mean, or some other method? If you find "mean" or "median," discuss whether or not the appropriate method was used.
3. The following data give the Number 1 hit records in each of 10 years. The class will work in groups. Each group takes the data from one year, makes a line plot, and identifies outliers using several different rules (for example, multipliers of 1.0, 1.5, and 2.0, or other appropriate multipliers). Then, each group decides which rule it likes the best for its data. Finally, discuss the results among the whole class. What is your choice?

1960

Weeks	Record Title	Artist
2	"El Paso"	Marty Robbins
3	"Running Bear"	Johnny Preston
2	"Teen Angel"	Mark Dinning
9	"The Theme from 'A Summer Place'"	Percy Faith
4	"Stuck on You"	Elvis Presley
5	"Cathy's Clown"	Everly Brothers
2	"Everybody's Somebody's Fool"	Connie Francis
1	"Alley-Oop"	Hollywood Argyles
3	"I'm Sorry"	Brenda Lee
1	"Itsy Bitsy Teenie Weenie Yellow Polkadot Bikini"	Brian Hyland
5	"It's Now or Never"	Elvis Presley
1	"The Twist"	Chubby Checker
2	"My Heart Has a Mind of Its Own"	Connie Francis
1	"Mr. Custer"	Larry Verne
3	"Save the Last Dance for Me"	Drifters
1	"I Want to Be Wanted"	Brenda Lee
1	"Georgia on My Mind"	Ray Charles
1	"Stay"	Maurice Williams & The Zodiacs
6	"Are You Lonesome To-Night?"	Elvis Presley

Source: *The Billboard Book of Top 40 Hits*, 1985.

1962

Weeks	Record Title	Artist
2	"The Twist"	Chubby Checker
3	"Peppermint Twist"	Joey Dee & The Starliters
3	"Duke of Earl"	Gene Chandler
3	"Hey! Baby"	Bruce Channel
1	"Don't Break the Heart That Loves You"	Connie Francis
2	"Johnny Angel"	Shelley Fabares
2	"Good Luck Charm"	Elvis Presley
3	"Soldier Boy"	Shirelles
1	"Stranger on the Shore"	Mr. Acker Bilk
5	"I Can't Stop Loving You"	Ray Charles
1	"The Stripper"	David Rose
4	"Roses Are Red"	Bobby Vinton
2	"Breaking Up Is Hard to Do"	Neil Sedaka
1	"The Loco-Motion"	Little Eva
2	"Sheila"	Tommy Roe
5	"Sherry"	4 Seasons
2	"Monster Mash"	Bobby "Boris" Pickett & The Crypt Kickers
2	"He's a Rebel"	Crystals
5	"Big Girls Don't Cry"	4 Seasons
3	"Telstar"	Tornadoes

Source: *The Billboard Book of Top 40 Hits*, 1985.

1964

Weeks	Record Title	Artist
4	"There! I've Said It Again"	Bobby Vinton
7	"I Want to Hold Your Hand"	Beatles
2	"She Loves You"	Beatles
5	"Can't Buy Me Love"	Beatles
1	"Hello, Dolly!"	Louis Armstrong
2	"My Guy"	Mary Wells
1	"Love Me Do"	Beatles
3	"Chapel of Love"	Dixie Cups
1	"A World Without Love"	Peter & Gordon
2	"I Get Around"	Beach Boys
2	"Rag Doll"	4 Seasons
2	"A Hard Day's Night"	Beatles
1	"Everybody Loves Somebody"	Dean Martin
2	"Where Did Our Love Go"	Supremes
3	"The House of the Rising Sun"	Animals
3	"Oh, Pretty Woman"	Roy Orbison
2	"Do Wah Diddy Diddy"	Manfred Mann
4	"Baby Love"	Supremes
1	"Leader of the Pack"	Shangri-Las
1	"Ringo"	Lorne Greene
1	"Mr. Lonely"	Bobby Vinton
2	"Come See about Me"	Supremes
3	"I Feel Fine"	Beatles

Source: *The Billboard Book of Top 40 Hits*, 1985.

1966

Weeks	Record Title	Artist
2	"The Sounds of Silence"	Simon & Garfunkel
3	"We Can Work It Out"	Beatles
2	"My Love"	Petula Clark
1	"Lightnin' Strikes"	Lou Christie
1	"These Boots Are Made for Walkin'"	Nancy Sinatra
5	"The Ballad of the Green Berets"	Sgt. Barry Sadler
3	"(You're My) Soul and Inspiration"	Righteous Brothers
1	"Good Lovin'"	Young Rascals
3	"Monday, Monday"	Mama's & Papa's
2	"When a Man Loves a Woman"	Percy Sledge
2	"Paint It, Black"	Rolling Stones
2	"Paperback Writer"	Beatles
1	"Strangers in the Night"	Frank Sinatra
2	"Hanky Panky"	Tommy James & The Shondells
2	"Wild Thing"	Troggs
3	"Summer in the City"	Lovin' Spoonful
1	"Sunshine Superman"	Donovan
2	"You Can't Hurry Love"	Supremes
3	"Cherish"	Association
2	"Reach Out I'll Be There"	Four Tops
1	"96 Tears"	?(Question Mark) & The Mysterians
1	"Last Train to Clarksville"	Monkees
1	"Poor Side of Town"	Johnny Rivers
2	"You Keep Me Hangin' On"	Supremes
3	"Winchester Cathedral"	New Vaudeville Band
1	"Good Vibrations"	Beach Boys
7	"I'm a Believer"	Monkees

Source: *The Billboard Book of Top 40 Hits*, 1985.

1968		
Weeks	**Record Title**	**Artist**
2	"Judy in Disguise (With Glasses)"	John Fred & His Playboy Band
1	"Green Tambourine"	Lemon Pipers
5	"Love Is Blue"	Paul Mauriat
4	"(Sittin' on) The Dock of the Bay"	Otis Redding
5	"Honey"	Bobby Goldsboro
2	"Tighten Up"	Archie Bell & The Drells
3	"Mrs. Robinson"	Simon & Garfunkel
4	"This Guy's in Love with You"	Herb Alpert
2	"Grazing in the Grass"	Hugh Masekela
2	"Hello, I Love You"	Doors
5	"People Got to Be Free"	Rascals
1	"Harper Valley P.T.A."	Jeannie C. Riley
9	"Hey Jude"	Beatles
2	"Love Child"	Diana Ross & The Supremes
7	"I Heard It through the Grapevine"	Marvin Gaye

Source: *The Billboard Book of Top 40 Hits*, 1985.

1980		
Weeks	**Record Title**	**Artist**
1	"Please Don't Go"	KC & The Sunshine Band
4	"Rock with You"	Michael Jackson
1	"Do That to Me One More Time"	Captain & Tennille
4	"Crazy Little Thing Called Love"	Queen
4	"Another Brick in the Wall (Part II)"	Pink Floyd
6	"Call Me"	Blondie
4	"Funkytown"	Lipps, Inc.
3	"Coming Up (Live at Glasgow)"	Paul McCartney & Wings
2	"It's Still Rock and Roll to Me"	Billy Joel
4	"Magic"	Olivia Newton-John
1	"Sailing"	Christopher Cross
4	"Upside Down"	Diana Ross
3	"Another One Bites the Dust"	Queen
3	"Woman In Love"	Barbra Streisand
6	"Lady"	Kenny Rogers
5	"(Just Like) Starting Over"	John Lennon

Source: *The Billboard Book of Top 40 Hits*, 1985.

1981

Weeks	Record Title	Artist
1	"The Tide Is High"	Blondie
2	"Celebration"	Kool & The Gang
2	"9 to 5"	Dolly Parton
2	"I Love a Rainy Night"	Eddie Rabbitt
1	"Keep on Loving You"	REO Speedwagon
2	"Rapture"	Blondie
3	"Kiss on My List"	Daryl Hall & John Oates
2	"Morning Train (Nine to Five)"	Sheena Easton
9	"Bette Davis Eyes"	Kim Carnes
1	"Stars on 45 Medley"	Stars on 45
1	"The One That You Love"	Air Supply
2	"Jessie's Girl"	Rick Springfield
9	"Endless Love"	Diana Ross & Lionel Richie
3	"Arthur's Theme (Best That You Can Do)"	Christopher Cross
2	"Private Eyes"	Daryl Hall & John Oates
10	"Physical"	Olivia Newton-John

Source: *The Billboard Book of Top 40 Hits*, 1985.

1982

Weeks	Record Title	Artist
1	"I Can't Go for That (No Can Do)"	Daryl Hall & John Oates
6	"Centerfold"	J. Geils Band
7	"I Love Rock 'n Roll"	Joan Jett & The Blackhearts
1	"Chariots of Fire"	Vangelis
7	"Ebony and Ivory"	Paul McCartney/ Stevie Wonder
3	"Don't You Want Me"	Human League
6	"Eye of the Tiger"	Survivor
2	"Abracadabra"	Steve Miller Band
2	"Hard to Say I'm Sorry"	Chicago
4	"Jack & Diane"	John Cougar
1	"Who Can It Be Now?"	Men At Work
3	"Up Where We Belong"	Joe Cocker & Jennifer Warnes
2	"Truly"	Lionel Richie
1	"Mickey"	Toni Basil
4	"Maneater"	Daryl Hall & John Oates

Source: *The Billboard Book of Top 40 Hits*, 1985.

1983

Weeks	Record Title	Artist
4	"Down Under"	Men At Work
1	"Africa"	Toto
2	"Baby, Come to Me"	Patti Austin & James Ingram
7	"Billie Jean"	Michael Jackson
1	"Come On Eileen"	Dexys Midnight Runners
3	"Beat It"	Michael Jackson
1	"Let's Dance"	David Bowie
6	"Flashdance...What a Feeling"	Irene Cara
8	"Every Breath You Take"	Police
1	"Sweet Dreams (Are Made of This)"	Eurythmics
2	"Maniac"	Michael Sembello
1	"Tell Her about It"	Billy Joel
4	"Total Eclipse of the Heart"	Bonnie Tyler
2	"Islands in the Stream"	Kenny Rogers with Dolly Parton
4	"All Night Long (All Night)"	Lionel Richie
6	"Say Say Say"	Paul McCartney & Michael Jackson

Source: *The Billboard Book of Top 40 Hits*, 1985.

1984

Weeks	Record Title	Artist
2	"Owner of a Lonely Heart"	Yes
3	"Karma Chameleon"	Culture Club
5	"Jump"	Van Halen
3	"Footloose"	Kenny Loggins
3	"Against All Odds (Take a Look at Me Now)"	Phil Collins
2	"Hello"	Lionel Richie
2	"Let's Hear It for the Boy"	Deniece Williams
2	"Time after Time"	Cyndi Lauper
2	"The Reflex"	Duran Duran
5	"When Doves Cry"	Prince
3	"Ghostbusters"	Ray Parker Jr.
3	"What's Love Got to Do with It"	Tina Turner
1	"Missing You"	John Waite
2	"Let's Go Crazy"	Prince
3	"I Just Called to Say I Love You"	Stevie Wonder
2	"Caribbean Queen (No More Love on the Run)"	Billy Ocean
3	"Wake Me Up Before You Go-Go"	WHAM!
2	"Out of Touch"	Daryl Hall & John Oates
6	"Like a Virgin"	Madonna

Source: *The Billboard Book of Top 40 Hits*, 1985.

Page 55

NOTE TO TEACHERS: In 1985 the Nielsen Company determined these ratings from electronic meters attached to the television sets in about 1,700 homes. Additionally, the people in these homes filled out diaries of the programs they watched.

Most of the programs listed are regular weekly shows; the titles in quotation marks are movies and other special features.

IV. BOX PLOTS

In the last section, we learned how to find the extremes, the quartiles and the median. These five numbers tell us a great deal about a set of data. In this section, we will describe a way of using them to make a plot.

The following tables give the ratings for national prime-time television for the week of April 29 through May 5, 1985, as compiled by the A. C. Nielsen Co. The 25.5 rating for *The Cosby Show* means that out of every 100 houses with televisions, 25.5 were watching *The Cosby Show* at the time it was on. Each ratings point represents 849,000 TV households.

TELEVISION RATINGS

	Program	Network	Rating
1.	The Cosby Show	NBC	25.5
2.	Family Ties	NBC	21.9
3.	Dallas	CBS	21.4
4.	Cheers	NBC	19.7
5.	Newhart	CBS	18.4
6.	Falcon Crest	CBS	18.3
7.	"Alfred Hitchcock Presents"	NBC	18.0
8.	60 Minutes	CBS	17.9
9.	Knots Landing	CBS	17.8
10.	A-Team	NBC	17.6
11.	Murder, She Wrote	CBS	17.6
12.	Night Court	NBC	17.6
13.	Highway to Heaven	NBC	17.0
14.	Facts of Life	NBC	16.8
15.	"Missing, Have You Seen This Person?"	NBC	16.5
16.	Kate & Allie	CBS	16.3
17.	Sara	NBC	16.3
18.	Who's the Boss?	ABC	15.9
19.	Trapper John, M.D.	CBS	15.7
20.	Love Boat	ABC	15.5
21.	Scarecrow & Mrs. King	CBS	15.4
22.	"Miss Hollywood '85"	ABC	15.4
23.	"Lace II," Part I	ABC	15.3
24.	Miami Vice	NBC	15.2
25.	Simon & Simon	CBS	15.2
26.	Riptide	NBC	15.2
27.	Cagney & Lacey	CBS	15.0
28.	"Adam"	NBC	14.9
29.	*Crazy Like a Fox*	CBS	14.6
30.	MacGruder and Loud	ABC	14.3
31.	20/20	ABC	14.3
32.	"Life's Embarrassing Moments"	ABC	14.2
33.	Hill Street Blues	NBC	14.0

Source: A.C. Nielsen Company.

TELEVISION RATINGS

	Program	Network	Rating
34.	St. Elsewhere	NBC	13.9
35.	Three's a Crowd	ABC	13.8
36.	Hail to the Chief	ABC	13.7
37.	"Joanna"	ABC	13.0
38.	Airwolf	CBS	12.7
39.	Remington Steele	NBC	12.6
40.	"Loving Couples"	CBS	12.4
41.	"Apocalypse Now"	ABC	12.4
42.	"Survival Anglia"	CBS	12.0
43.	Gimme a Break	NBC	12.0
44.	Knight Rider	NBC	11.8
45.	Hunter	NBC	11.6
46.	"Anything for a Laugh"	ABC	11.6
47.	T. J. Hooker	ABC	11.5
48.	Double Trouble	NBC	11.5
49.	Magnum, P. I.	CBS	11.4
50.	Diff'rent Strokes	NBC	10.7
51.	Benson	ABC	10.7
52.	"Ray Mancini Story"	CBS	10.6
53.	Mike Hammer	CBS	10.5
54.	Webster	ABC	10.4
55.	Under One Roof	NBC	10.4
56.	Half-Nelson	NBC	10.4
57.	Double Dare	CBS	9.6
58.	Best Times	NBC	9.5
59.	"Dr. No"	ABC	9.5
60.	Punky Brewster	NBC	9.0
61.	Ripley's Believe It or Not	ABC	8.5
62.	Cover Up	CBS	8.3
63.	Eye to Eye	ABC	8.3
64.	Street Hawk	ABC	7.9
65.	Silver Spoons	NBC	7.8
66.	Lucie Arnaz Show	CBS	7.5
67.	Jeffersons	CBS	7.1

Source: A.C. Nielsen Company.

Page 57: Discussion Questions

1. 50 percent
2. 25 percent
3. 75 percent
4. 50 percent
5. 25 percent

The following instructions will teach you how to make a box plot of the ratings of the 67 programs:

Step 1 Find the median rating.

There are 67 ratings, thus the median will be the 34th show. The 34th show, *St. Elsewhere*, has a rating of 13.9.

Step 2 Find the median of the upper half.

There are 33 ratings above the median. The median of these ratings is at the 17th show. This show is *Sara* with a rating of 16.3. This number 16.3 is the upper quartile.

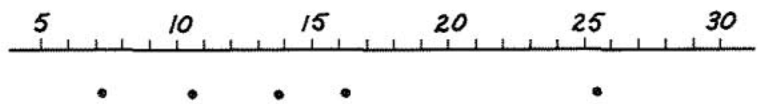
Step 3 Find the median of the lower half.

There are 33 ratings below the median. The median of these ratings is at the 51st show, which is *Benson* with a rating of 10.7. This number 10.7 is the lower quartile.

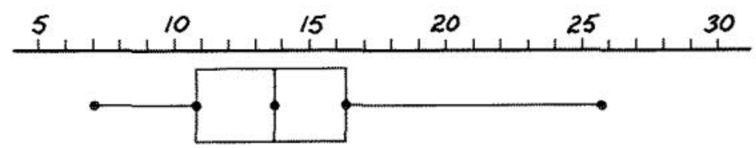
Step 4 Find the extremes.

The lowest rating is 7.1 and the highest is 25.5.

Step 5 Mark dots for the median, quartiles, and extremes below a number line.



Step 6 Draw a box between the two quartiles. Mark the median with a line across the box. Draw two "whiskers" from the quartiles to the extremes.



Discussion Questions

About what percent of the ratings are:

1. Below the median?
2. Below the lower quartile?
3. Above the lower quartile?
4. In the box?
5. In each whisker?

6. Is one whisker longer than the other? What does this mean?
7. Why isn't the median in the center of the box?
8. On May 8, 1985, CBS announced that it was cancelling *The Jeffersons*, *Cover Up*, *The Lucie Arnaz Show*, and *Double Dare*. The future of *Mike Hammer* was in doubt. Why do you think CBS is cancelling these shows? Are there any other programs CBS should consider cancelling?
9. Which shows do you think ABC cancelled?

The executives of the networks are interested in how the three compare in ratings. We learned that a back-to-back stem-and-leaf plot is good for such comparisons. Unfortunately, it has only two sides and there are three networks. Box plots are effective for comparing two or more sets of data. For example, let's plot the ratings for CBS, NBC, and ABC on separate box plots.

CBS has 22 shows listed. Their ratings are:

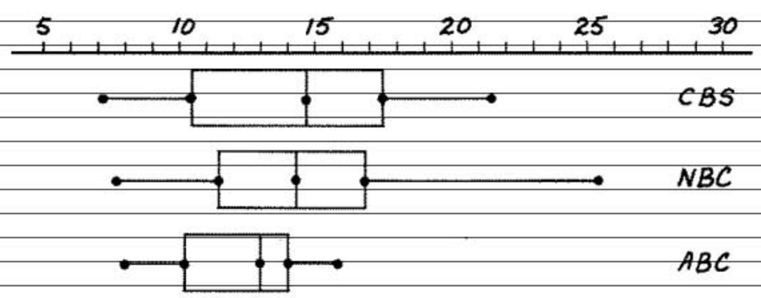
21.4	18.4	18.3	17.9	17.8	17.6	16.3	15.7
15.4	15.2	15.0	14.6	12.7	12.4	12.0	11.4
10.6	10.5	9.6	8.3	7.5	7.1		

The median is halfway between the 11th and 12th ratings, which are 15.0 and 14.6. Thus, the median is:

$$\frac{15.0 + 14.6}{2} = \frac{29.6}{2} = 14.8 .$$

The lower quartile is 10.6 and the upper quartile is 17.6. The extremes are 7.1 and 21.4.

The box plots for CBS, NBC, and ABC are shown below.



Page 58: Discussion Questions (*continued*)

6. Yes; it means the ratings of the top quarter of the shows are more spread out than those in the bottom quarter.
7. It's close to the center, but generally won't be exactly in the center because the values in the second and third quarters are unlikely to be equally spread out.
8. Low ratings mean advertisers will pay less to show their commercials; Magnum P.I.
9. On May 6, 1985, ABC announced that it was cancelling *Three's a Crowd, Eye to Eye, MacGruder and Loud, T. J. Hooker, Hail to the Chief,* and *Street Hawk.*

Page 59: Discussion Questions

1. They are actually 14.5, 11.5 and 17.0, 7.8 and 25.5.
2. They are actually 13.0, 10.4 and 14.3, 7.9 and 15.9.
3. CBS
4. NBC, CBS, ABC
5. CBS, NBC, ABC
6. CBS, NBC, ABC
7. Yes, the *Cosby Show* is just barely an outlier (25.5 versus 25.25).
8. no
9. Answers will vary. Sample: Most of the detail of individual values is omitted in box plots and we can then see relative standings better. In addition, we don't have any method of showing three different networks on a line plot or on a back-to-back stem-and-leaf plot.
10. In redrawing the box plot for ABC to reflect the hypothetical situation, we see the approximate values as follows: minimum 10.4, lower quartile 13.0, median 14.3, upper quartile 15.9, maximum 22.

Discussion Questions

1. Use the box plot to estimate the median, quartiles, and extremes for NBC.
2. Use the box plot to estimate the median, quartiles, and extremes for ABC.
3. Study the box plots to decide which network has the largest interquartile range.
4. If you say that the winning network is the one with the highest-rated show, which network is the winner? Which is second? Which is third?
5. If you say that the winning network is the one with the largest upper quartile, which network is the winner? Which is second? Which is third?
6. If you say that the winning network is the one with the largest median, which network is the winner? Which is second? Which is third?
7. Use the box plot to estimate if there are any outliers for NBC. (Hint: The length of the box is one interquartile range!)
8. Are any shows outliers for CBS or ABC?
9. Why are box plots a better way to compare the relative positions of the three networks than line plots or stem-and-leaf plots?
10. Write a description of the relative standings of the three networks. Then (don't peek) read the following example.

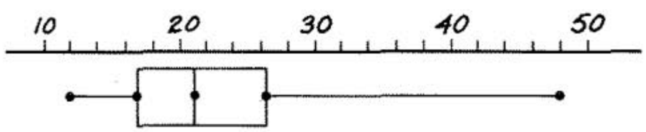
The median ratings of the three networks are very close — each around 14. The lower quartiles and lower extremes are also very close — around 11 and 7, respectively. This means that if you look at just the shows in the bottom half for each network, the three networks do about the same in the ratings. However, when looking at the top half of the ratings, NBC and CBS do much better than ABC. The ratings for ABC are all packed tightly between 13.0 and 15.9. In contrast, about 25% of the ratings for both CBS and NBC are larger than 17. It is clear that ABC is the losing network, but whether NBC or CBS is the winner is not so clear.

Even if ABC had cancelled the bottom quarter of their shows and replaced them all by shows that received a higher rating than their current top show — for example between 17 and 22 — they would still be a bit behind NBC and CBS in terms of the top shows. (As an exercise, redraw the boxplot for ABC to reflect this hypothetical situation.)

Application 12

Prices of Corn Poppers

The box plot below shows the dollar prices of twenty popcorn poppers as listed in *Consumer Reports Buying Guide*, 1981.



Source: *Consumer Reports Buying Guide*, 1981.

1. Approximately how much did the most expensive popcorn popper cost?
2. Approximately how much did the least expensive popcorn popper cost?
3. What was the median price for a popcorn popper?
4. What percentage of the poppers cost more than $26.50 (the upper quartile)?
5. What percentage of the poppers cost more than $17.00 (the lower quartile)?
6. If you had $21.00, how many of the twenty poppers could you afford?
7. If you had $26.50, how many of the twenty poppers could you afford?
8. Are any of the prices outliers? How can you tell?
9. Write a short description of the price of popcorn poppers.

Page 60

NOTE TO TEACHERS: Application 12, "Prices of Corn Poppers," may be omitted.

Application 12

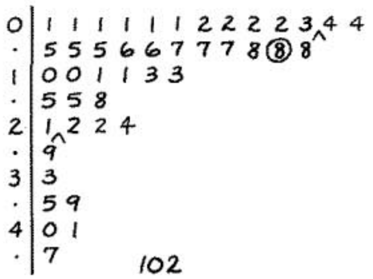
1. $48
2. $12
3. $21
4. 25 percent
5. 75 percent
6. 10
7. 15
8. Yes; because the upper extreme is more than 1.5 box lengths above the upper quartile.
9. Answers will vary. Sample: Of the twenty popcorn poppers listed in *Consumer Reports*, the most expensive was about $48 and the least expensive about $12. Half cost more than $21 and five more than about $26. Five cost under about $17. At least one popper was much more expensive than the others.

Page 61

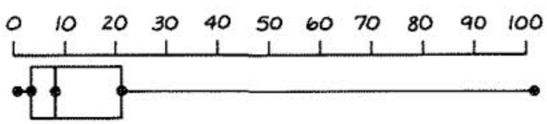
NOTE TO TEACHERS: Application 13 should be completed by *all* students because it introduces the method of showing an outlier on a box plot (question 5 on p. 62).

Application 13

1. The missing states probably have no roller skating clubs, so the values would be 0; Alaska, Idaho, Montana, North Dakota, South Dakota.
2. A stem-and-leaf plot like this helps in making the box plot:



The lower extreme is 1, the lower quartile is 3.5, the median is 8, the upper quartile is 21.5, and the upper extreme is 102.



3. An outlier will lie more than 1.5 interquartile ranges above the upper quartile; that is, above

$$21.5 + 1.5\,(21.5 - 3.5) = 48.5.$$

California, at 102, is larger than 48.5

4. California makes this whisker long. If California were omitted, the whisker would end at 47.

Application 13

Roller Skating Clubs

The following table gives the number of roller skating clubs by state for 45 states.

State	Number	State	Number
Alabama	11	Nebraska	8
Arizona	6	Nevada	1
Arkansas	5	New Hampshire	1
California	102	New Jersey	24
Colorado	11	New Mexico	1
Connecticut	7	New York	18
Delaware	2	*North Carolina*	15
Florida	39	Ohio	47
Georgia	8	Oklahoma	5
Hawaii	1	Oregon	13
Illinois	35	Pennsylvania	41
Indiana	21	Rhode Island	5
Iowa	7	South Carolina	2
Kansas	7	Tennessee	10
Kentucky	6	Texas	40
Louisiana	10	Utah	2
Maine	1	Vermont	1
Maryland	15	Virginia	33
Massachusetts	13	Washington	22
Michigan	29	West Virginia	4
Minnesota	4	Wisconsin	8
Mississippi	3	Wyoming	2
Missouri	22		

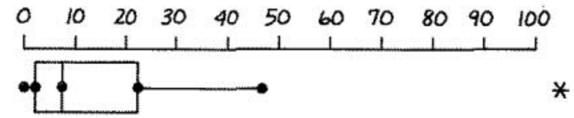
Source: Roller Skating Rink Operators Association.

1. Why do you think the data include only 45 and not 50 states? What values might the 5 remaining states have? Which states are missing?
2. Make a box plot of the 45 values. (Hint: The numbers must be put in order before you find the median and the quartiles. A quick way to do *this is to use a stem-and-leaf plot.)*
3. Show that California is an outlier.
4. Look at the upper whisker. Why is it so long? If you were to omit California from the list, how would the box plot change?

5. There is an alternate way to construct the box plot when there is an outlier, such as California. Copy your box plot, but stop the upper whisker at Ohio's 47. Then, put an asterisk at California's 102. Thus, there is a gap in the plot, corresponding to the gap between the largest and second-largest values.
6. Which of these plots do you think gives a more accurate picture of these data? Why?
7. Write a description of the information given in the box plot you constructed for question 5.

Page 62: Application 13 (*continued*)

5.



6. The plot in question 5. It shows that there is only one state with more than 47 roller skating clubs. From the plot in question 2, one might think there are many.
7. Answers will vary. Sample: This box plot shows that half of the forty-five states listed have 8 or fewer roller skating clubs. Another quarter of the states have between 8 and 21 clubs, and the top quarter between 22 and 47. One state, California, has 102 clubs, more than twice as many as the next state, Ohio, with 47.

   Five states—Alaska, Idaho, Montana, North Dakota, and South Dakota—are not listed. They probably have no roller skating clubs.

Page 63

NOTE TO TEACHERS: Application 14, "Sugar in Cereals," may be omitted.

Application 14

1. It could mean either percentage of weight or percentage of calories.

Application 14

Sugar in Cereals

Percentage of Sugar in Cereals

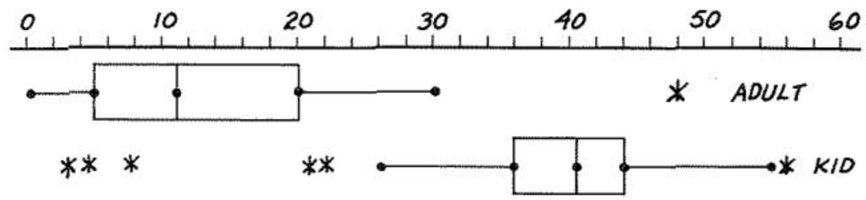
Product	% Sugar	Product	% Sugar
Sugar Smacks (K)	56.0	Kellogg Raisin Bran (A)	29.0
Apple Jacks (K)	54.6	C. W. Post, Raisin, (A)	29.0
Froot Loops (K)	48.0	C. W. Post (A)	28.7
General Foods Raisin Bran (A)	48.0	Frosted Mini Wheats (K)	26.0
Sugar Corn Pops (K)	46.0	Country Crisp (K)	22.0
Super Sugar Crisp (K)	46.0	Life, cinnamon (K)	21.0
Crazy Cow, chocolate (K)	45.6	100% Bran (A)	21.0
Corny Snaps (K)	45.5	All Bran (A)	19.0
Frosted Rice Krinkles (K)	44.0	Fortified Oat Flakes (A)	18.5
Frankenberry (K)	43.7	Life (A)	16.0
Cookie Crisp, vanilla (K)	43.5	Team (A)	14.1
Cap'n Crunch, crunch berries (K)	43.3	40% Bran (A)	13.0
Cocoa Krispies (K)	43.0	Grape Nuts Flakes (A)	13.3
Cocoa Pebbles (K)	42.6	Buckwheat (A)	12.2
Fruity Pebbles (K)	42.5	Product 19 (A)	9.9
Lucky Charms (K)	42.2	Concentrate (A)	9.3
Cookie Crisp, chocolate (K)	41.0	Total (A)	8.3
Sugar Frosted Flakes of Corn (K)	41.0	Wheaties (A)	8.2
Quisp (K)	40.7	Rice Krispies (K)	7.8
Crazy Cow, strawberry (K)	40.1	Grape Nuts (A)	7.0
Cookie Crisp, oatmeal (K)	40.1	Special K (A)	5.4
Cap'n Crunch (K)	40.0	Corn Flakes (A)	5.3
Count Chocula (K)	39.5	Post Toasties (A)	5.0
Alpha Bits (K)	38.0	Kix (K)	4.8
Honey Comb (K)	37.2	Rice Chex (A)	4.4
Frosted Rice (K)	37.0	Corn Chex (A)	4.0
Trix (K)	35.9	Wheat Chex (A)	3.5
Cocoa Puffs (K)	33.3	Cheerios (K)	3.0
Cap'n Crunch, peanut butter (K)	32.2	Shredded Wheat (A)	0.6
Golden Grahams (A)	30.0	Puffed Wheat (A)	0.5
Cracklin' Bran (A)	29.0	Puffed Rice (A)	0.1

Source: United States Department of Agriculture, 1979.

1. What do you think the table means when it says that "the percentage of sugar" in Sugar Smacks is 56.0?

We divided the list into "kid" and "adult" cereals as indicated by a (K) or an (A) following each name. (You may disagree and change some of these.)

The following box plots show the amount of sugar in "kid" and "adult" cereals.



2. For the "kid" cereals, estimate:
   a. the lower extreme
   b. the upper extreme
   c. the median
   d. the lower quartile
   e. the upper quartile

3. For the "adult" cereals, estimate
   a. the lower extreme
   b. the upper extreme
   c. the median
   d. the lower quartile
   e. the upper quartile

4. Write a paragraph comparing the percentage of sugar in "kid" and "adult" cereals.

Page 64: Application 14 (*continued*)

2. The actual values are as follows:
   a. 3.0
   b. 56.0
   c. 40.85
   d. 35.9
   e. 43.7

3. The actual values are as follows:
   a. 0.1
   b. 48.0
   c. 11.05
   d. 5.15
   e. 20.0

4. Answers will vary. Sample: These box plots show that there is a lot more sugar in "kid" cereal than in "adult" cereal. In fact, all but one of the "adult" cereals are 30 percent or less sugar while more than 75 percent of the "kid" cereals have more than this percentage of sugar. However, there are five outlier "kid" cereals that have far less sugar than the others. The three lowest—Cheerios, Kix, and Rice Krispies—even have a low amount of sugar relative to most "adult" cereals. The other two—Cinnamon Life and Country Crisp—are high relative to the "adult" cereals, but still quite low relative to the other "kid" cereals.

   One "adult" cereal, Raisin Bran, has more sugar than most of the "kid" cereals.

Page 65

NOTE TO TEACHERS: Application 15, "Automobile Safety," may be omitted.

Application 15

Automobile Safety

The Highway Loss Data Institute rated 181 models of 1982-84 cars based on the number of insurance claims filed for personal injury coverage. The cars are rated in relative terms; 100 represents the average for all cars. Lower numbers mean a better safety record. A rating of 122, for example, means 22% worse than average.

Station Wagons and Passenger Vans

Small Cars	Injury Rating	Midsize Cars	Injury Rating	Large Cars	Injury Rating
Volkswagen Vanagon	73	Volvo 240	56	Olds. Custom Cruiser	54
Mercury Lynx	83	American Eagle 30	69	Buick Electra	59
Toyota Tercel 4WD	91	Ford LTD	76	Dodge Caravan	63
Ford Escort	95	Oldsmobile Firenza	80	Plymouth Voyager	67
Subaru DL/GL 4WD	98	Chevrolet Celebrity	83	Chevrolet Caprice	69
Subaru DL/GL	100	Dodge Aries	91	Mercury Grand Marquis	69
Nissan Sentra	108	Plymouth Reliant	93	Ford Crown Victoria	70
		Pontiac 2000	94		
		Chevrolet Cavalier	94		
		Chrysler LeBaron	95		
		Nissan Maxima	100		

Source: Highway Loss Data Institute.

Sports and Specialty Models

Small Cars	Injury Rating	Midsize Cars	Injury Rating	Large Cars	Injury Rating
Mercedes 380SL Coupe	57	Lincoln Continental	72	Mercedes 300SD/380SE	60
Chevrolet Corvette	63	BMW 528e/533i	74	Jaguar X16	63
Porsche 944 Coupe	71	Audi 5000 4D	79	Mercedes-Benz 300D	64
Nissan 300ZX	100	BMW 318i/325e	81	Oldsmobile Toronado	65
VW Rabbit Convertible	102	Chrys. LeBaron Conv.	87	Cadillac De Ville 4D	67
Mazda RX-7	104	Ford Mustang Convertible	98	Cadillac Eldorado	71
Pontiac Fiero	119	Toyota Celica Supra	102	Lincoln Town Car	72
Ford EXP	124	Pontiac Firebird	107	Buick Riviera	73
		Mercury Capri	114	Cadillac Brougham 4D	75
		Chevrolet Camaro	116	Cadillac Seville	76
		Ford Mustang	127	Cadillac De Ville 2D	81

Source: Highway Loss Data Institute.

Four-Door Models

Small Cars	Injury Ratings	Midsize Cars	Injury Ratings	Large Cars	Injury Ratings
Saab 900	71	Chrysler E Class	75	Oldsmobile Delta 88	59
Honda Accord	89	Oldsmobile Cutlass	76	Buick LeSabre	62
Volkswagen Rabbit	92	Buick Regal	79	Oldsmobile Ninety Eight	62
Volkswagen Jetta	97	Pontiac Bonneville	80	Mercury Grand Marquis	65
Mazda 626	100	Mercury Topaz	81	Buick Electra	66
Nissan Stanza	107	Pontiac 6000	85	Chevrolet Caprice	68
Dodge Omni	114	Mercury Marquis	86	Ford LTD Crown Victoria	68
Renault Alliance	114	Dodge 600	86	Chrys. 5th Ave.	69
Ford Escort	117	Oldsmobile Ciera	86	Dodge Diplomat	72
Plymouth Horizon	118	Chrysler New Yorker	87	Chevrolet Impala	79
Mercury Lynx	120	Buick Century	87	Plymouth Grand Fury	101
Toyota Corolla	122	Chrysler LeBaron	88		
Subaru DL/GL Sedan	125	Volvo 240	89		
Toyota Tercel	127	Ford LTD	89		
Mazda GLC	130	Peugeot 505	91		
Pontiac 1000	139	Toyota Camry	91		
Isuzu T-Car/I-Mark	140	Toyota Cressida	92		
Chevrolet Chevette	143	Buick Skylark	92		
Dodge Colt	144	Cadillac Cimarron	93		
Nissan Sentra	145	Chevrolet Celebrity	94		
Mitsubishi Tredia	155	Chevrolet Citation	94		
Plymouth Colt	156	Audi 4000	96		
		Oldsmobile Omega	98		
		Ford Tempo	100		
		Pontiac Phoenix	101		
		Pontiac 2000	109		
		Dodge Aries	111		
		Plymouth Reliant	112		
		Chevrolet Cavalier	112		
		Oldsmobile Firenza	113		
		Buick Skyhawk	113		
		Nissan Maxima	121		

Source: Highway Loss Data Institute.

Page 67: Application 15

1. station wagons and passenger vans
2. two-door models

Two-Door Models

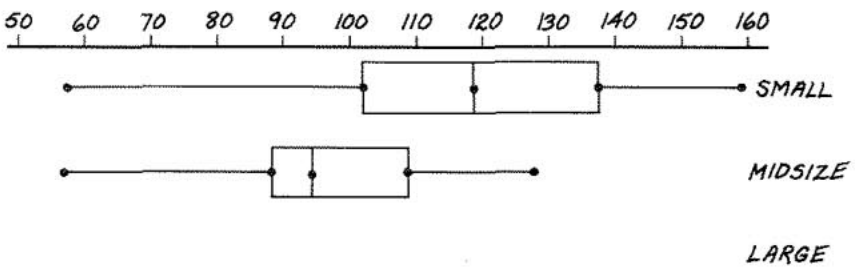
Small Cars	Injury Ratings	Midsize Cars	Injury Ratings	Large Cars	Injury Ratings
Saab 900	70	Oldsmobile Cutlass	88	Ford Crown Victoria	65
Honda Accord	102	Buick Regal	90	Buick LeSabre	70
Nissan Stanza	105	Oldsmobile Ciera	91	Oldsmobile Delta 88	70
Volkswagen Rabbit	106	Pontiac Grand Prix	92	Oldsmobile Ninety Eight	71
Mazda 626	106	Oldsmobile Omega	92	Mercury Grand Marquis	76
Volkswagen Scirocco	108	Pontiac 6000	94	Chevrolet Caprice	77
Mazda GLC	110	Buick Skylark	94	Buick Electra	81
Honda Prelude	114	Chevrolet Monte Carlo	98		
Honda Civic	115	Chrysler LeBaron	99		
Subaru Hardtop	117	Ford Thunderbird	100		
Renault Fuego	118	Buick Century	100		
Toyota Celica	120	Volvo 240	104		
Dodge Daytona	122	*Dodge 400/600*	105		
Subaru Hatchback	125	Chevrolet Celebrity	107		
Plymouth Horizon	128	Dodge Aries	109		
Chrysler Laser	128	Mercury Cougar	109		
Toyota Tercel	129	Chevrolet Citation	111		
Ford Escort	130	Pontiac Phoenix	112		
Renault Encore	130	Pontiac 2000	118		
Dodge Charger	132	Ford Tempo	118		
Mercury Lynx	137	Plymouth Reliant	119		
Nissan Sentra	137	Buick Skylark	123		
Renault Alliance	138	Oldsmobile Firenza	123		
Toyota Starlet	148	Chevrolet Cavalier	126		
Plymouth Colt	148				
Dodge Colt	149				
Mitsubishi Cordia	151				
Chevrolet Chevette	154				
Pontiac 1000	155				
Nissan Pulsar	158				

Source: Highway Loss Data Institute.

1. Which of the four groups of cars is the safest?

2. Which is the most dangerous group?

3. The box plot for all of the small cars and for midsize cars is shown below. (All four types of models were combined.) Make the box plot for large cars. Show any outliers as in Application 13, question 5.



4. Which would you say are closer in safety, small and midsize cars, or midsize and large cars? Why?

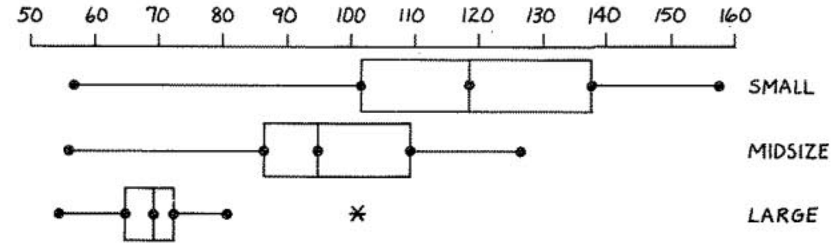
5. Write a paragraph giving an overall summary of the plots.

6. (Optional) Make box plots for American small cars and for Japanese small cars, or two other categories that interest you, and write a summary of the plots.

7. (For class discussion) Do you think that these injury ratings reflect just the inherent safety of these cars? Might they also relate to other factors such as different characteristics of the drivers, different mileages, or different types of driving that the cars receive? What other ways can you think of for comparing the safety of different automobiles?

Page 68: Application 15 (*continued*)

3.



For large cars, the lower extreme is 54, the lower quartile is 64.5, the median is 69, the upper quartile is 72.5, the upper whisker is 81, and the outlier is 101. (For small cars, the points plotted are 57, 102, 118, 137, and 158. For midsize cars, the points plotted are 56, 87, 94, 109, and 127.)

4. Small and midsize; their box plots overlap.
5. Answers will vary. Sample: The box plots show that the larger the car, the better safety record it tends to have. There are exceptions. Several small cars have very good safety records, and the safest small, midsize, and large cars have almost exactly the same safety ratings. One large car, the Plymouth Grand Fury, has a safety record much worse than the other large cars, and it is more typical of midsize cars.

   Another interesting thing is that the box plot for small cars is more spread out than that for midsize cars, and the plot for midsize cars is more spread out than that for large cars. This means that large cars tend to be alike in their safety records while there is more variation in midsize cars and still more in small cars.

   The distributions of the small and midsize cars overlap more than do the distributions of the midsize and large cars. More precisely, *any* large car (except for the Plymouth Grand Fury) is safer than three-fourths of the midsize cars and three-fourths of the small cars.
6. Answers will vary.
7. Answers will vary. Sample: The factors listed could also affect the ratings. It would be helpful to know more about how these numbers were compiled.

Page 69

NOTE TO TEACHERS: Application 16, "High School Eligibility," may be omitted.

Application 16

High School Eligibility

Data from the *Los Angeles Times* appear in the following table.

High School	*% Ineligible in Selected Activities*					
	Band	Drama	Yearbook	Baseball	Boys Track	Girls Track
Banning	27	19	0	9	38	24
Bell	37	19	—	0	22	13
Belmont	—	3	7	19	14	7
Birmingham	52	31	—	—	24	—
Canoga Park	30	19	20	17	25	33
Carson	—	0	—	13	21	—
Chatsworth	25	19	33	9	20	31
Cleveland	11	—	7	16	15	—
Crenshaw	68	36	—	20	19	17
Dorsey	8	28	—	15	31	31
Eagle Rock	0	—	0	—	—	—
El Camino Real	7	15	13	3	16	15
Fairfax	35	23	—	21	51	30
Francis Poly	4	28	—	0	22	31
Franklin	48	33	21	17	29	44
Fremont	—	43	—	32	32	38
Gardena	34	—	17	19	20	20
Garfield	21	—	—	7	16	23
Granada Hills	14	29	—	15	21	28
Grant	—	3	—	17	26	—
Hamilton	36	27	0	24	12	0
Hollywood	3	3	8	—	—	—
Huntington Park	40	33	44	15	22	—
Jefferson	61	58	—	8	62	—

High School	% Ineligible in Selected Activities					
	Band	Drama	Yearbook	Baseball	Boys Track	Girls Track
Jordan	49	70	50	38	32	—
Kennedy	20	14	18	0	18	20
Lincoln	32	28	71	6	17	13
Locke	30	—	67	45	30	57
Los Angeles	27	100	39	43	37	17
Manual Arts	45	35	38	21	18	25
Marshall	14	19	—	16	31	15
Monroe	23	30	6	3	21	24
Narbonne	21	35	0	—	—	—
North Hollywood	18	50	—	—	—	—
Palisades	4	14	—	14	30	30
Reseda	24	53	0	—	—	—
Roosevelt	—	39	—	12	12	35
San Fernando	19	64	24	24	44	33
San Pedro	11	10	—	8	18	0
South Gate	38	8	5	19	15	28
Sylmar	10	32	—	0	21	13
Taft	—	11	0	10	27	27
University	20	30	43	16	14	9
Van Nuys	22	21	0	17	30	7
Venice	11	21	—	5	16	13
Verdugo Hills	8	35	10	14	39	39
Washington	29	31	25	16	18	14
Westchester	19	11	0	17	3	0
Wilson	—	—	—	—	—	—

Source: *Los Angeles Times*, May 17, 1983.

Following a policy established by the Los Angeles Board of Education, students must maintain a C average and have no failing grades in order to participate in extracurricular activities. The table shows how the policy is affecting activities at high schools. Numbers represent the percentage of students in the activity who were declared ineligible. For example, 25% of band members at Chatsworth and 38% of the athletes in girls' track at Fremont were declared ineligible and could no longer participate.

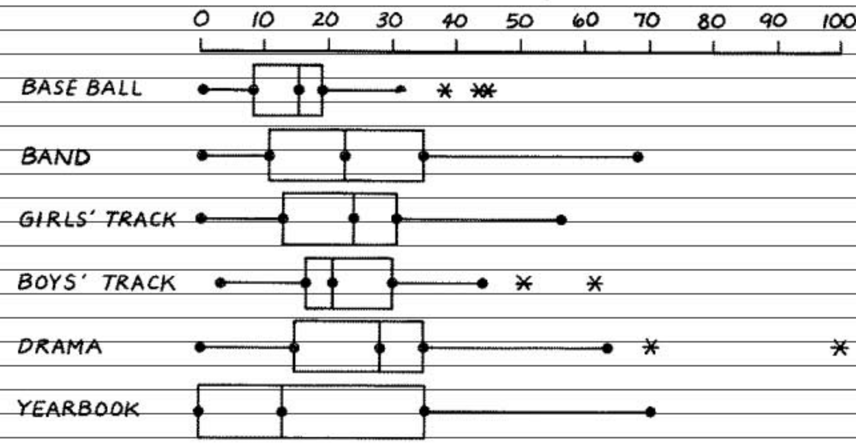
1. The newspaper does not say why the table contains blanks. How do you know that a blank does not mean that no students were ineligible? What do you think a blank means? For the rest of the worksheet, ignore the blanks.

2. The class should be divided into six groups. One group should construct a box plot of the percentage of students declared ineligible in band; one group should construct the plot for drama; another for yearbook, and so on. Use an asterisk for any outliers, as in Application 13, question 5.

Page 70: Application 16

1. Because a 0 in the table means no students were ineligible. Maybe it means that the activity is not offered at the school or that the school did not respond for that activity.

2. The box plots are based on the following points for the lower extreme, the lower quartile, the median, the upper quartile, the upper extreme or the point at which the upper whisker ends, and any outliers:
   Baseball: 0, 8, 15.5, 19, 32; 38, 43, 45
   Band: 0, 11, 22.5, 35, 68; —
   Girls' Track: 0, 13, 23.5, 31, 57; —
   Boys' Track: 3, 17, 21, 30, 44; 51, 62
   Drama: 0, 15, 28, 35, 64; 70, 100
   Yearbook: 0, 0, 13, 35.5, 71; —



Page 71: Application 16 (*continued*)

3. See the preceding box plots in question 2.
4. band members
5. Because more than 25 percent of the schools have an ineligibility rate of 0 percent
6. Answers will vary but could include the following information.

   The rates of ineligibility were generally lowest in baseball. Baseball also has the smallest range, from 0 percent to 45 percent. It is difficult to distinguish among the rest, as there is much overlap among the distributions.

   The ineligibility rates for yearbook appear to be different from the rates for other activities. Yearbook has the longest box by a substantial margin, indicating that there is a lot of variability among the schools. Further, at least one-fourth of the schools have ineligibility rates of 0 percent; this didn't happen for any other activity. Yearbook's median of 13 percent is the lowest of any activity. Thus about one-half of the schools have quite small rates of ineligibility for yearbook, but the others cover a very wide range.

   Drama generally had the largest rates of ineligibility. It has the largest median of 28 percent, its upper quartile of 35 percent is as large as that for any other activity, and its maximum is 100 percent. Nevertheless, about one-fourth of the schools had rates of 15 percent or smaller, and some had 0 percent ineligible for drama.

   The rates for band, girls' track, and boys' track are similar, being larger than those for baseball but smaller than drama.

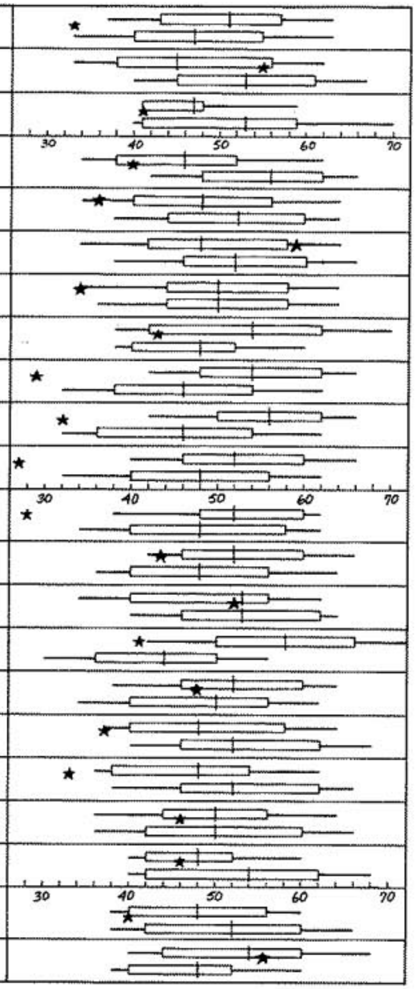
3. Make a number line on the blackboard or overhead projector. A representative from each group should draw its box plot under this number line.
4. Do band members or baseball players tend to have higher rates of ineligibility?
5. Why is there no lower whisker on the yearbook box plot?
6. Write a paragraph or two summarizing what you see in the six box plots.

The Use of Box Plots

It is becoming more and more common to use a box plot to tell people their results on a test. For example, students sometimes take tests to see how interested they are in various occupations. The results from one such test are reproduced below.

BASIC INTEREST SCALES

NATURE	33	V-LOW
ADVENTURE	55	MOD-H
MILITARY ACTIVITIES	41	V-LOW
MECHANICAL ACTIVITIES	40	AVER.
SCIENCE	36	LOW
MATHEMATICS	59	HIGH
MEDICAL SCIENCE	34	LOW
MEDICAL SERVICE	43	MOD-L
MUSIC/DRAMATICS	29	V-LOW
ART	32	V-LOW
WRITING	26	V-LOW
TEACHING	28	V-LOW
SOCIAL SERVICE	43	MOD-L
ATHLETICS	52	AVER.
DOMESTIC ARTS	41	LOW
RELIGIOUS ACTIVITIES	48	AVER.
PUBLIC SPEAKING	37	MOD-L
LAW/POLITICS	33	LOW
MERCHANDISING	45	AVER.
SALES	46	AVER.
BUSINESS MANAGEMENT	40	MOD-L
OFFICE PRACTICES	55	AVER.



Page 73: Discussion Questions

1. mathematics
2. This girl's ★ is below the box plot for girls in nature, medical science, music/dramatics, art, writing, teaching, domestic arts, law/politics.
3. The girls' median is higher than the boys' upper quartile in medical service, music/dramatics, art, domestic arts, office practices.
4. military activities, mechanical activities, sales
5. Answers will vary. Sample: The test you took rated your interest in various occupational categories. Compared to other girls, your interest in nature, medical science, music/dramatics, art, writing, teaching, domestic arts, and law/politics is very low.

   You have a high interest in mathematics and a moderately high interest in adventure. Your next highest interests were in office practices and athletics.

   Based on these interests, I suggest that you consider a career as a sports statistician or as a computer programmer who writes adventure games.

   There are more areas in which you have low interest than areas in which you have high interest. This might be due to a lack of knowledge about some of the low interest areas. Your interests may change as you learn more about these areas.

Let's examine the "Nature" result more carefully. There are two box plots for "Nature." The top one is for girls and the bottom one is for boys. The top box plot shows that the median interest score in nature for girls is about 51. (The scale is above "Mechanical Activities.") The score of the girl who took the test is marked on each scale by a ★. Thus, her interest in nature is very low compared to other girls who have taken the test previously.

Discussion Questions

1. For which subject(s) is this girl's interest score in the top 25% of all girls?
2. For which subjects is this girl's interest lowest?
3. Which subjects are girls much more interested in than are boys?
4. Which subjects are boys much more interested in than are girls?
5. Write a letter to this girl recommending possible career choices.

Box Plots — Summary

You may have found it difficult to see the advantages of using box plots. Some students are disturbed by the fact that most of the data disappears and only five summary numbers (the median, quartiles, and extremes) remain. It is true that we can no longer spot clusters and gaps, nor can we identify the shape of the distribution as clearly as with line plots or stem-and-leaf plots. However, we are able to focus on the relative positions of different sets of data and thereby compare them more easily.

Box plots are especially useful when the set of data contains hundreds or even thousands of numbers. A line plot or stem-and-leaf plot would be unwieldy with thousands of numbers on it!

To compare two (or more) sets of data using box plots, first look at the boxes to get an idea whether or not they are located in about the same place. Also, study their lengths, to determine whether or not the variabilities in the data sets are about the same. Then, you can focus on details. Check whether or not one data set has median, upper and lower quartiles, and extremes that are all larger than the corresponding values in the second data set. If it does, then the data in the first set tend to be larger than those in the second no matter which criterion we use for comparing them. If it does not, then there is more uncertainty about which data set is larger. In either case, the plot has helped us learn some details about the similarities and differences between the two data sets. Also, check to see if the pattern of outliers is the same in both data sets.

Notice that even if two (or more) sets of data have unequal numbers of values, this does not cause problems for making comparisons with box plots. This was not true for stem-and-leaf plots.

Suggestions for Student Projects

1. Collect some data on a topic that interests you, construct box plots, and interpret them. Topics that other students have used include:
   - number of hours students work per week
   - number of hours of TV watched per week by different types of students
   - allowances of girls and of boys in your class
   - scores of all the students in a school that take a certain test, separated so you can compare the different classes

2. One variation of box plots involves changing the width in proportion to the number of data values represented. For example, if a box representing 100 values is 1 cm wide, then a box representing 50 values would be 0.5 cm wide and a box representing 200 values would be 2 cm wide. Make box plots under the same number line for the small two-door models, midsize two-door models and large two-door models from Application 15. Make the width of the box proportional to the number of cars represented. Discuss the merit of this variation.

V. REVIEW OF ONE-VARIABLE TECHNIQUES

Which Method to Use?

This section is different from the previous four. Each of the previous four introduced some statistical method that can help to interpret data. Then, the method was used on several examples. Often more than one of these methods *could* be used to display and to help interpret a particular set of data. This section helps you to choose an appropriate method by giving some comparisons among them.

Before using any statistical method it is a good idea to ask yourself a few basic questions about the data. How were the numbers obtained? Are the values plausible? What would you like to learn from the data? Are there any specific questions that you know need answers? The purpose of statistical methods is to help us learn something useful or interesting from the data, so it is a good idea to keep questions such as these in mind throughout the analysis.

Suppose we have the starting weekly wage for 23 different jobs. We could display the values using a line plot (Section I), a stem-and-leaf plot (Section II), or a box plot (Section IV). We could calculate statistics such as the median, mean, range, and interquartile range (Section III). Which of these methods should we use, or, at least, which should we use first? There is no single, correct answer. However, there are some guidelines that can help you to make an appropriate choice of methods.

A reasonable general strategy is to use the simpler methods first. Then, if the interpretations of the data are very clear, there is no need to go on to more complicated displays and methods.

One Group and One Variable

Consider the above example of the starting wage for several jobs. In this example there is one *variable*, the wage. We can treat the various jobs as forming one *group* of jobs. Thus, we have measurements for *one group on one variable*. This is the simplest type of problem for which statistical methods and displays are needed. Most of the examples in Sections I, II, and III are this type of problem.

The line plot, the stem-and-leaf plot, and the box plot are three different displays that can be used for the one-group/one-variable situation. The following paragraphs describe their relative advantages and disadvantages.

> ***Line Plot***. The line plot is easy to construct and interpret. It gives a clear graphical picture, and a few values can be labeled easily. Constructing a line plot is also a useful first step for calculating the median, extremes, and quartiles. These statements are all true providing the number of values is not too large — fewer than about 25. As the number of values becomes larger, the line plot can become unwieldy and more difficult to interpret. When a specific value is repeated several times or when there are many

nearby values, the line plot can also become jumbled. Another disadvantage is that it is hard to read the exact numerical values from the line plot. In conclusion, the line plot is a useful first display for the one-group/one-variable situation, providing there are about 25 or fewer values in the data.

Stem-and-Leaf Plot. The stem-and-leaf plot shares many advantages of the line plot. It is easy to construct and interpret, values can be labeled, and it is a useful first step for calculating the median, extremes, and quartiles. In addition, exact numerical values can be read from the stem-and-leaf plot and repeated values and nearby values in the data cause no special problems. Stem-and-leaf plots do not get as unwieldy as line plots when the number of data values becomes large. On the other hand, a disadvantage is that to construct the stem-and-leaf plot you may have to decide whether or not to truncate or round. Further disadvantages are the need to decide which values to use for the stems, and how to spread out the plot. Thus, it may take more thought to construct the stem-and-leaf plot than the line plot. The stem-and-leaf plot can display more values than the line plot without becoming too confusing in appearance. However, it also has a limit to the number of values that is *reasonable* to display. With more than about 100 values, you will most likely spread out the stem-and-leaf plot. Then it can be useful for up to about 250 values. Above 250 it will be too large and jumbled to interpret easily. In conclusion, for the one-group/one-variable situation with about 25 or fewer values, either the stem-and-leaf plot or the line plot is a reasonable first display. The choice is partly a matter of personal preference. With about 25 to 250 data values, the stem-and-leaf plot is the most useful first display.

Box Plot. The box plot is more complicated to construct, since you must calculate the median, extremes, and quartiles first. Generally, the simplest way to do this is to construct the stem-and-leaf plot first and then count in from the ends to get the quartiles and median. Unlike the stem-and-leaf plot, once the box plot is constructed, specific data values cannot be read from it (except for outliers and the median, quartiles, and extremes). The main advantage of the box plot is that it is not cluttered by showing all the data values. It highlights only a few *important* features of the data. Thus, the box plot makes it easier to focus attention on the median, extremes, and quartiles and comparisons among them. Another advantage of the box plot is that it does not become more complicated with more data values. It is useful with any number of values. A disadvantage of the box plot occurs when there are only a few data values — less than about 15. Then, the plotted values might change greatly if only one or a few of the observations were changed.

The box plot is a *summary display* since it shows only certain statistics, not all the data. In conclusion, the box plot is not as useful as the line or stem-and-leaf plots for showing details, but it enables us to focus more easily on the median, extremes, and quartiles. Since the line and stem-and-leaf plots are useful for computing the statistics needed to construct the box plot, it is generally reasonable to make one of these two plots first even if you will eventually construct and use the box plot.

Several Groups and One Variable

Think again about the starting weekly wage example mentioned at the beginning of this section. Instead of considering the 23 jobs as *one group* of jobs, we could divide them into those jobs that require a high school diploma and those that require a college diploma. The jobs are divided into *two groups*. We want to compare the various salaries in these two groups. This is an example of the *two-group/one-variable* problem. Many of the examples in Sections II and IV are this type. The following paragraphs describe the relative advantages and disadvantages of the line, stem-and-leaf, and box plots for this situation.

Line plots can be placed next to each other to compare two groups, although we did not give any examples of this type. However, this becomes confusing if the two groups overlap a lot or if there are more than a total of about 25 data values.

Back-to-back stem-and-leaf plots are more useful for comparing two groups. They are easy to construct. Comparisons can be made by judging the number of leaves for various stems. However, if the number of data values in the two groups is not roughly equal, the comparisons get more difficult. The details shown in the stem-and-leaf plots can become an obstacle. Furthermore, as the number of values becomes large these plots become unwieldy. In summary, for comparing two groups of about equal size with around 100 or fewer data values in each group, back-to-back stem-and-leaf plots are easy to construct and generally adequate.

Box plots below the same number line can also be used to compare two groups. This gives the easiest and most direct comparisons of the two minimums, the two lower quartiles, the two medians, the two upper quartiles, and the two maximums. Of course, this does not show any other details, but these quantities are usually sufficient for comparing two groups. Moreover, there are no special problems caused by having a large number of data values, or by having a different number of values in the two groups.

Often, we need to compare more than two groups. For example, the jobs could be broken down into those not requiring a high school diploma, those requiring a high school diploma, those requiring a college degree, and those requiring a graduate degree. This gives four groups. It is an example of a *many-group/one-variable* problem.

There is no way to construct a stem-and-leaf plot for this situation. Several line plots placed next to each other can be useful, if there are not many data values. Box plots are the best choice. The reasons are the same as those given for comparing two groups.

A more concise way to compare two groups than any of these is simply to calculate a single number, such as the mean or median, for each group. But this number hides all the other information in the data. It also loses the advantage of graphical displays. Thus, for purposes of exploring and interpreting data, any of the graphical displays will be more valuable than calculating just means or medians. If it is necessary to give a single number to summarize the data, and if there is a possibility of even a few outliers, then the median is usually more valuable than the mean.

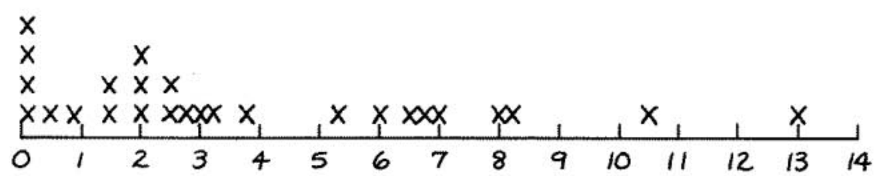
As a general conclusion, line plots, stem-and-leaf plots, and box plots each have a useful role for exploring various kinds of data sets. Often, it is worthwhile to make more than one plot. There are no hard and fast rules about which plot should be used, but the previous comparisons can help you make good choices.

The following applications will help you compare the different methods.

(Answers for p. 79 start here and continue on the facing page.)

Page 79: Application 17

1. E
2. Z
3. 10.5; 52.5
4. 38.2 percent
5.



6.

0	0 1 1 2 4 9
1	4 5
2	0 0 0 5 5 8
3	0 4 8
4	
5	3
6	0 5 8
7	0
8	0 2
9	
10	5
11	
12	
13	0

1/4 REPRESENTS 1.4%

(Answers for p. 79 continued from the facing page.)

7. median = 2.65; lower quartile = 1.4; upper quartile = 6.5
8.

9. E and T; from the line plot
10. Half are used about 2.5 percent of the time or less, so most are rarely used; from the box plot.
11.

VOWELS		CONSONANTS
	0	0 1 1 2 4 9
	1	4 5
5	2	0 0 0 5 8
	3	0 4 8
	4	
	5	3
5	6	0 8
	7	0
2 0	8	
	9	
	10	5
	11	
	12	
0	13	

1|4 REPRESENTS 1.4%.

12. There are only 5 vowels. There are not enough of them to make a box plot.
13. Answers will vary. Sample: Three of the vowels (E, A, and O) are used more than any consonant except T. The next vowel, I, is used more than all but three consonants, and the last vowel, U, is still used more than about half the consonants. Among the consonants, T is used far more than any other. Another group used frequently is N, R, S, and H.

Application 17

Letter Frequencies

The number of occurrences of each letter was counted in a very large amount of written material. The percentage that each letter occurred is given in the table below.

A	8.2	J	0.1	S	6.0
B	1.4	K	0.4	T	10.5
C	2.8	L	3.4	U	2.5
D	3.8	M	2.5	V	0.9
E	13.0	N	7.0	W	1.5
F	3.0	O	8.0	X	0.2
G	2.0	P	2.0	Y	2.0
H	5.3	Q	0.1	Z	0.07
I	6.5	R	6.8		

Source: National Council of Teachers of Mathematics.

1. What is the most-used letter?
2. What is the least-used letter?
3. How many *t*'s would you expect to find in a paragraph of 100 letters? In a paragraph of 500 letters?
4. As a group, vowels account for what percentage of letters used?
5. Make a line plot of the percentages.
6. Make a stem-and-leaf plot of the percentages.
7. Find the median percentage, the quartiles, and any outliers.
8. Make a box plot of the percentages.
9. Which two letters have the most unusual percentages? From which plot is it easiest to find this information?
10. Are most of the letters used rarely or used more frequently? From which plot is it easiest to find this information?
11. Make a back-to-back stem-and-leaf plot of vowels and consonants.
12. Why isn't it appropriate to make one box plot for vowels and another for consonants?
13. What conclusions can you make by looking at the stem-and-leaf plot you constructed for question 11?

Application 18

Salaries

The table below lists the median weekly salaries of workers employed full time. For example, the median salary for carpenters is $325 because half of the carpenters earn less than $325 and half earn more than $325.

Occupation	Median Weekly Earnings	Occupation	Median Weekly Earnings
Accountant	379	Machinist	356
Airplane Pilot	530	Mathematician	508
Architect	428	Newspaper Reporter	351
Auto Mechanic	285	Painter	271
Bank Teller	189	Pharmacist	463
Barber	327	Physician, Osteopath	501
Bookkeeper	227	Plumber	404
Carpenter	325	Police Officer	363
Cashier	168	Postal Clerk	400
Chemist	467	Printing Press Operator	320
Civil Engineer	505	Psychologist	394
College Teacher	444	Receptionist	200
Computer Programmer	422	Registered Nurse	332
Cooks and Chefs	171	Retail Sales Worker	178
Cosmetologist	179	School Counselor	396
Dental Assistant	183	Secondary Teacher	351
Dentist	352	Secretary	229
Drafter	343	Shoe Repairer	200
Electrician	419	Telephone Operator	240
Fire Fighter	362	Truck Driver (local)	314
Flight Attendant	365	Truck Driver (long distance)	517
Food Counter Worker	141	Typist	213
K-6 Teacher	322	Veterinarian	656
Lawyer	546	Waiter/Waitress	150
Librarian	320	Welder	334

Source: United States Bureau of Labor Statistics.

1. Which kind of worker earns the most?
2. Which kind of worker earns the least?
3. Which occupation listed would you most like to have someday?
4. Suppose you want to see how the salary of the occupation you chose compares to the other salaries. Which do you think is best for this use: a line plot, stem-and-leaf plot, or box plot?
5. Construct the plot you selected.
6. In one or two sentences, describe how the salary of the occupation you chose compares to the other salaries.

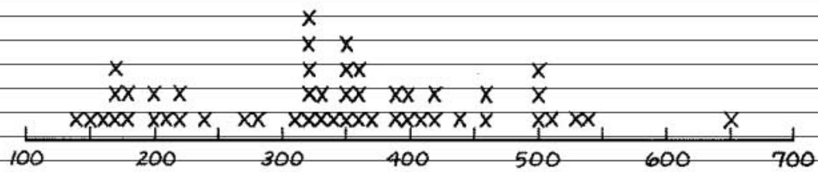
Page 80: Application 18

1. veterinarian
2. food counter worker
3. Answers will vary.
4. Answers will vary.
5. Stem-and-leaf plot:

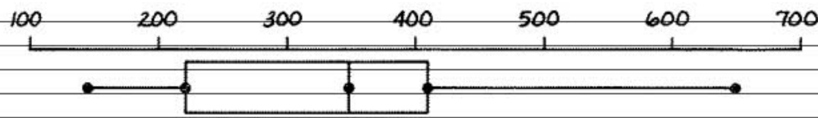
1 | 4
· | 5 6 7 7 7 8 8
2 | 0 0 1 2 2 4
· | 7 8
3 | 1 2 2 2 2 2 3 3 4
· | 5 5 5 5 6 6 6 7 9 9
4 | 0 0 1 2 2 4
· | 6 6
5 | 0 0 0 1 3 4
· |
6 |
· | 5

2|0 REPRESENTS $200 - $209
MEDIAN WEEKLY SALARY

Line plot:



Box plot:



(lower extreme = 141; lower quartile = 220; median = 345; upper quartile = 410; upper extreme = 656)

6. Answers will vary.

NOTE TO TEACHERS: Some students might question whether these numbers are accurate. It seems surprising that veterinarians have by

(Answers for p. 80 continue on the facing page.)

(Answers for p. 80 continued from the facing page.)

far the highest median salary, much larger than physicians, for example. These numbers were taken accurately from the source. Perhaps the answer is that many physicians are considered self-employed, not salaried, and therefore not included here. Perhaps salaried physicians are mainly the lower-paid interns and residents. Perhaps similar issues also apply for other occupations.

The point of this is that you cannot always just take published numbers at face value. You should ask, how were they collected and are they believable?

Page 81: Application 19

1. Answers will vary.
2. Answers will vary. Sample: Utah spends the least amount of money per student, but four other states also spend between $2,000 and $2,200. However, at the upper end of the distribution, Alaska with $6,378 spends far more than the second highest state, New York with $4,821. There are three other states that are relatively high—New Jersey, Washington, DC, and Wyoming. The median spent per student is about $2,900, and about half of the states are within $400 of that amount.

```
2 | 0 0 0 1 1
· | 2 2 3
· | 4 4 5
· | 6 6 6 7 7
· | 8 8 8 8 8 9 9 9 9 9
3 | 1 1 1 1
· | 2 3 3 3 3 3
· | 4
· | 6 6 6 7 7 7
· | 8 8
4 | 0
· |
· | 4 5
· | 6
· | 8
5 |
· |
· |
· |
· |       2|0 REPRESENTS $2000 - $2099
6 |       SPENT PER STUDENT
· | 3
```


Application 19

Money Spent Per Student

The values in the table below are the amount of money spent on education per student in 1983-84 for each of the 50 states and Washington, D.C.

State	Expense	State	Expense
Alabama	$2,082	Montana	$3,691
Alaska	$6,378	Nebraska	$2,913
Arizona	$2,685	Nevada	$2,882
Arkansas	$2,214	New Hampshire	$2,765
California	$2,981	New Jersey	$4,677
Colorado	$3,188	New Mexico	$2,866
Connecticut	$4,055	New York	$4,821
Delaware	$3,848	North Carolina	$2,455
D.C.	$4,574	North Dakota	$2,952
Florida	$3,169	Ohio	$2,996
Georgia	$2,317	Oklahoma	$3,146
Hawaii	$3,395	Oregon	$3,771
Idaho	$2,174	Pennsylvania	$3,707
Illinois	$3,384	Rhode Island	$3,811
Indiana	$2,583	South Carolina	$2,271
Iowa	$3,251	South Dakota	$2,639
Kansas	$3,392	Tennessee	$2,141
Kentucky	$2,646	Texas	$2,960
Louisiana	$2,707	Utah	$2,047
Maine	$2,839	Vermont	$3,491
Maryland	$3,771	Virginia	$2,853
Massachusetts	$3,692	Washington	$3,129
Michigan	$3,315	West Virginia	$2,488
Minnesota	$3,322	Wisconsin	$3,677
Mississippi	$2,090	Wyoming	$4,488
Missouri	$2,814		

Source: National Education Association.

1. Using the value for your state, and an estimate of the number of students in your school, give a rough estimate of the total cost of running your school in 1983-84.

2. Suppose you want to know how your state compares to the others. Construct a plot to help you make this comparison, and label your state.

Then, write a paragraph describing the overall distribution of expenses, and the relative position of your state.

3. Pick 3 to 5 nearby states that are similar to yours. Label them on the plot. Write another sentence or two describing how the expenses in your state compare to those of your neighbors.

4. Using the map of the United States on page 15, classify each state as being in the Northeast, Central, South, or West. Then, construct a plot to show how the expenses per student compare in the four regions of the country. Write a paragraph summarizing the comparisons.

Page 82: Application 19 (*continued*)

3. Answers will vary.
4. Based on the map on page 15 of *Exploring Data,* states will be classified as follows:

	State	Expense		State	Expense
S	Alabama	$2,082	W	Montana	$3,691
W	Alaska	$6,378	C	Nebraska	$2,913
W	Arizona	$2,685	W	Nevada	$2,882
S	Arkansas	$2,214	N	New Hampshire	$2,765
W	California	$2,981	N	New Jersey	$4,677
W	Colorado	$3,188	W	New Mexico	$2,866
N	Connecticut	$4,055	N	New York	$4,821
N	Delaware	$3,848	S	North Carolina	$2,455
N	D.C.	$4,574	C	North Dakota	$2,952
S	Florida	$3,169	C	Ohio	$2,996
S	Georgia	$2,317	S	Oklahoma	$3,146
W	Hawaii	$3,395	W	Oregon	$3,771
W	Idaho	$2,174	N	Pennsylvania	$3,707
C	Illinois	$3,384	N	Rhode Island	$3,811
C	Indiana	$2,583	S	South Carolina	$2,271
C	Iowa	$3,251	C	South Dakota	$2,639
C	Kansas	$3,392	S	Tennessee	$2,141
S	Kentucky	$2,646	S	Texas	$2,960
S	Louisiana	$2,707	W	Utah	$2,047
N	Maine	$2,839	N	Vermont	$3,491
N	Maryland	$3,771	S	Virginia	$2,853
N	Massachusetts	$3,692	W	Washington	$3,129
C	Michigan	$3,315	N	West Virginia	$2,488
C	Minnesota	$3,322	C	Wisconsin	$3,677
S	Mississippi	$2,090	W	Wyoming	$4,488
C	Missouri	$2,814			

(Answers for p. 82 continue on the facing page.)

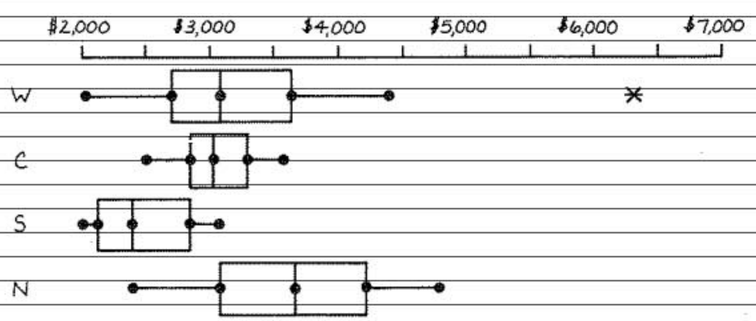
(Answers for p. 82 continued from the facing page.)

Stem-and-leaf plot:

2	W S S S W
.	S S S
.	S N C
.	C S W S N
.	C N S W W C C S W C
3	W S S W
.	C C C C C W
.	N
.	C W N N N W
.	N N
4	N
.	
.	W N
.	N
.	N
5	
.	
.	
.	
.	
6	
.	W

2| REPRESENTS $2000 - $2199

The values for lower extreme, lower quartile, median, upper quartile, and upper extreme, as calculated from a truncated stem-and-leaf plot, are as follows:
West: 2,000, 2,700, 3,100, 3,650, 4,400
Central: 2,500, 2,850, 3,050, 3,300, 3,600
South: 2,000, 2,150, 2,400, 2,850, 3,100
Northeast: 2,400, 3,100, 3,700, 4,250, 4,800



(Answers for p. 82 continue on the next page.)

VI. SCATTER PLOTS

The table below gives the box score for the first game of the 1985 National Basketball Association Championship series.

Los Angeles Lakers 114, Boston Celtics 148

LOS ANGELES

	Min	FG-A	FT-A	R	A	P	T
Worthy	37	8-19	4-6	8	5	1	20
Rambis	22	4-6	0-0	9	0	2	8
Jabbar	22	6-11	0-0	3	1	3	12
Magic Johnson	34	8-14	3-4	1	12	2	19
Scott	30	5-14	0-0	2	0	2	10
Cooper	24	1-5	2-2	2	2	3	4
McAdoo	21	6-13	0-0	3	0	5	12
McGee	15	4-7	4-5	2	2	1	14
Spriggs	15	4-7	0-2	3	4	1	8
Kupchak	16	3-3	1-2	2	1	3	7
Lester	4	0-1	0-0	0	1	0	0
Totals	240	49-100	14-21	35	28	23	114

Shooting field goals, 49.0%, free throws, 66.7%

BOSTON

	Min	FG-A	FT-A	R	A	P	T
Bird	31	8-14	2-2	6	9	1	19
McHale	32	10-16	6-9	9	0	1	26
Parish	28	6-11	6-7	8	1	1	18
Dennis Johnson	33	6-14	1-1	3	10	1	13
Ainge	29	9-15	0-0	5	6	1	19
Buckner	16	3-5	0-0	4	6	4	6
Williams	14	3-5	0-0	0	5	2	6
Wedman	23	11-11	0-2	5	2	4	26
Maxwell	16	1-1	1-2	3	1	0	3
Kite	10	3-5	1-2	3	0	1	7
Carr	4	1-3	0-0	1	0	1	3
Clark	4	1-2	0-0	1	3	0	2
Totals	240	62-102	17-25	48	43	17	148

Shooting field goals, 60.8%, free throws, 68.0%

Key for table

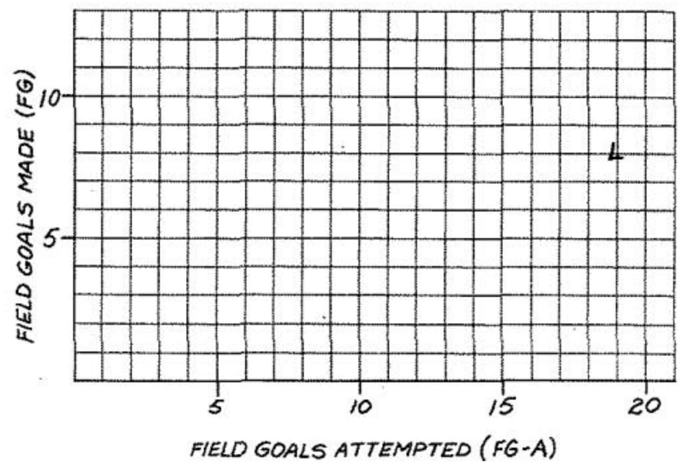
Min	Minutes played
FG-A	Field goals made - field goals attempted
FT-A	Free throws made - free throws attempted
R	Rebounds
A	Assists
P	Personal fouls
T	Total points scored

Source: *Los Angeles Times*, May 28, 1985.

Discussion Questions

1. How many rebounds did Kevin McHale make?
2. Which player played the most minutes?
3. Which player had the most assists?
4. How many field goals did James Worthy make? How many did he attempt? What percentage did he make?
5. Five players are on the court at one time for each team. Determine how many minutes are in a game.
6. Which team made a larger percentage of free throws?
7. How is the T (total points scored) column computed? Verify that this number is correct for Magic Johnson and for Kevin McHale. (Caution: Some of the field goals for other players were three point shots.)

Do you think that the players who *attempt* the most field goals are generally the players that *make* the most field goals? Of course! We can see this from the box score. To further investigate this question, we will make a *scatter plot* showing field goals made (FG) and field goals attempted (FG-A). First, set up a plot with field goals attempted on the horizontal axis and field goals made on the vertical axis.



Worthy, the first player, attempted 19 field goals and made 8 of them. The L on the preceding plot represents Worthy. The L is above 19 and across from 8. We used an L to show that he is a Los Angeles player.

(Answers for p. 82 continued from the previous page.)

Paragraph summaries will vary but may contain the following information:

In general, the northeast spends the most per student and the south spends the least. Among the northern states, West Virginia, New Hampshire, and Maine are separated at the low end, since each spends at least $600 less than any other northeastern state. In the south, only two states, Florida and Oklahoma, spend more than $3,000 per student, an amount that is exceeded by all northeastern states except the three first listed.

The central and western regions have about the same median, midway between the northeast and the south. The west, however, is more diverse. Two western states, Idaho and Utah, are among the lowest of all states, and Alaska spends far more than any other state. The central region is the least variable of the four regions.

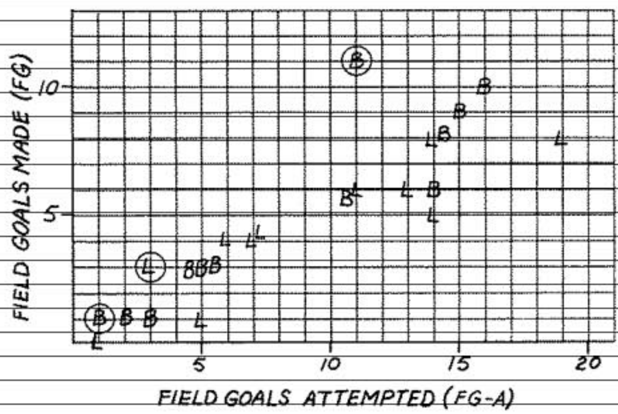
NOTE TO TEACHERS: These features can be seen using either a stem-and-leaf plot with four symbols or box plots. The stem-and-leaf plots make it easier to spot states that are different from the rest in their region. The box plots show the overall relationship more clearly, but we have to be a bit careful because there are not many observations in each group.

Page 84: Discussion Questions

1. 9
2. Worthy
3. Magic Johnson
4. 8; 19; 42 percent
5. 240/5 = 48 minutes
6. Boston with 68 percent
7. Multiply the number of field goals by two and add the number of free throws.
   Johnson: $8 \times 2 + 3 = 19$
   McHale: $10 \times 2 + 6 = 26$

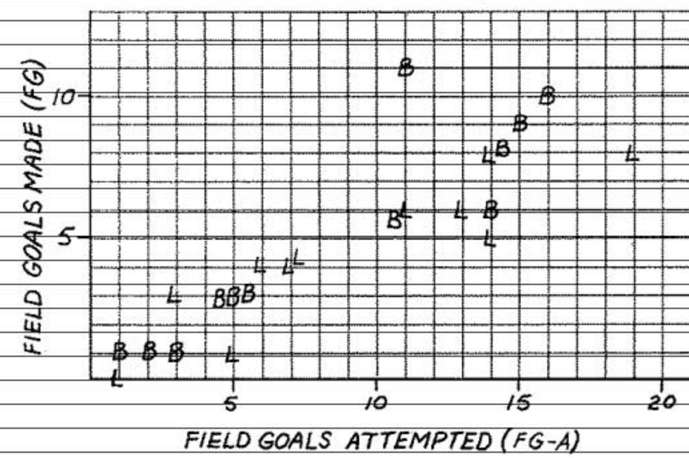
Page 85: Discussion Questions

1. They are circled below.



2. If there were a point in the upper left half, it would mean the player made more field goals than he attempted.
3. no
4. Cooper, Scott, and Worthy
5. 5
6. Answers will vary.

The completed scatter plot follows. Each B stands for a Boston player and each L for a Los Angeles player.



As we suspected, this plot shows that players who attempt more field goals generally make more field goals, and players who attempt few field goals make few field goals. Thus, there is a *positive* association between field goals attempted and field goals made.

However, we can see much more from this plot. First, a player who makes every basket will be represented by a point on the line through the points (0, 0), (1, 1), (2, 2), (3, 3), and so forth. Second, the players who are relatively far below this line were not shooting as well as the other players. Finally, we can observe the relative positions of the two teams in this plot.

Discussion Questions

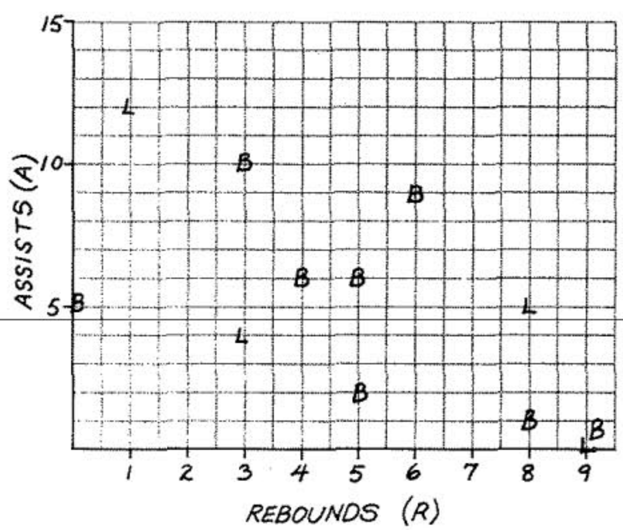
1. Using the scatter plot, find the points that represent the three perfect shooters.
2. Why are all the points below a diagonal line running from lower left to upper right?
3. Is there a different pattern for Los Angeles and Boston players?
4. Which three Laker players were not shooting very well that game?
5. Suppose a player attempts 9 field goals. About how many would you expect him to make?
6. Write a brief description of the information conveyed by this scatter plot. Then read the following sample discussion. Did you notice any information not listed in this sample discussion?

In this plot, we were not surprised to see a positive association between the number of field goals attempted and the number of field goals made. There were three players, two from Boston and one from Los Angeles, who made all the field goals they attempted. One of these Boston players was truly outstanding as he made eleven out of eleven attempts. The Laker players who attempted a great number of field goals generally did not make as many of them as did the Celtics who attempted a great number of field goals. This could have been the deciding factor in the game.

The points seem to cluster into two groups. The cluster on the upper right generally contains players who played over 20 minutes and the one on the lower left contains players who played less than 20 minutes.

An assist is a pass that leads directly to a basket. A player is credited with a rebound when he recovers the ball following a missed shot. Do you think that players who get a lot of rebounds also make a lot of assists? It is difficult to answer this question just by looking at the box score.

To answer this question, we will make a scatter plot showing rebounds (R) and assists (A). This plot includes all players who made at least four rebounds or four assists.



This plot shows that players who get *more* rebounds generally have *fewer* assists, and players who get *fewer* rebounds have *more* assists. Thus, there is a *negative* association between rebounds and assists.

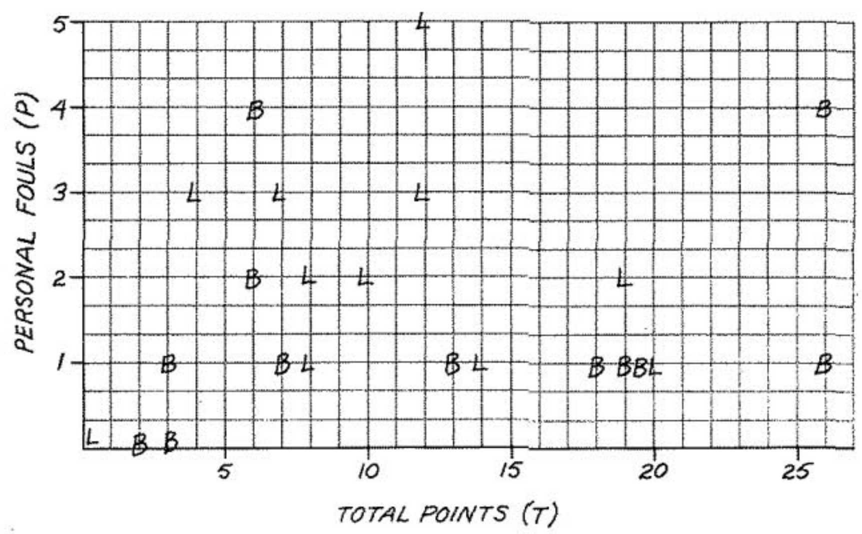
Page 87: Discussion Questions

1. no
2. 4
3. no
4. Answers will vary.
5. No; both might result from whether the player tends to play close to the basket or far away.
6. Answers will vary. Sample: They played too few minutes for any possible relationship between rebounds and assists to develop. Therefore, including them just clutters the plot.

Discussion Questions

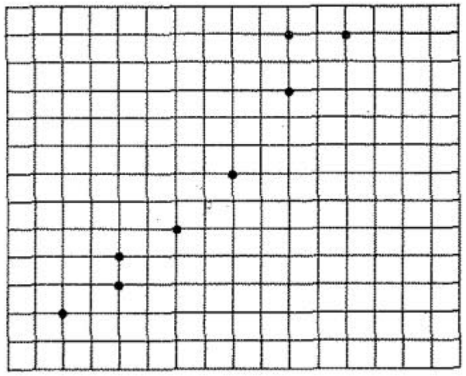
1. Do the players who get the most rebounds also make the most assists?
2. Suppose a player had 7 rebounds. About how many assists would you expect this player to have?
3. Is there a different pattern for Boston players than for Los Angeles players?
4. Why do you suppose players who get a lot of rebounds do not make a lot of assists?
5. If you were the coach and you wanted a player to make more assists, would you instruct him to make fewer rebounds?
6. Why didn't we include players who would have been in the lower left-hand corner of this plot?

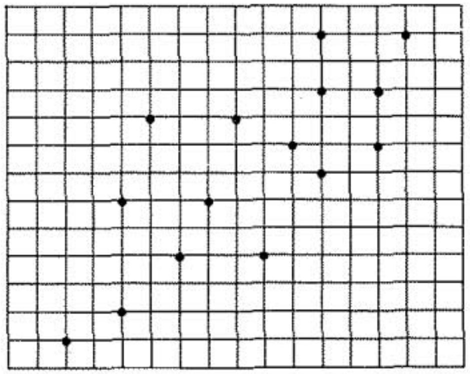
The following scatter plot shows total points and personal fouls for all players.



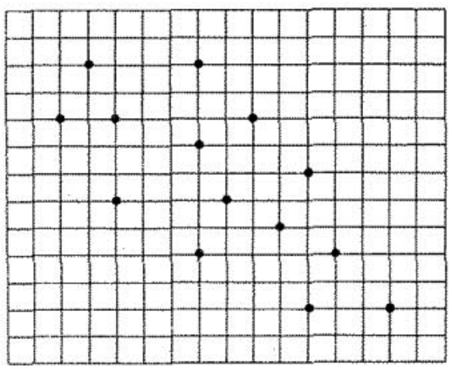
This plot shows *no association* between total points scored and the number of personal fouls committed.

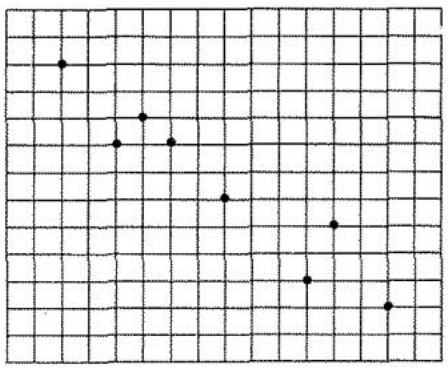
In summary, the following scatter plots show *positive association.*



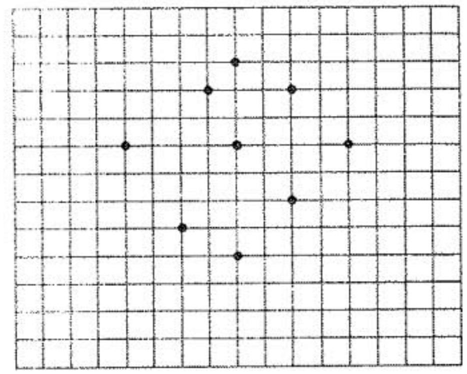


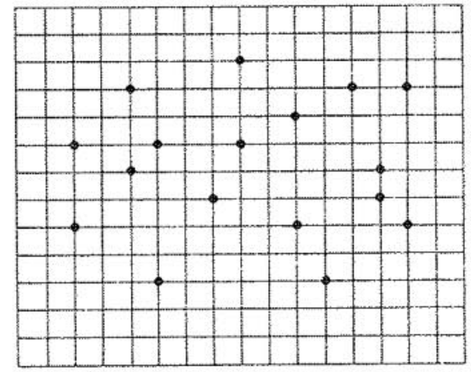
The following scatter plots show *negative association.*





The following scatter plots show *no association.*





Sometimes one or two points can make it appear that there is a positive or negative association when there is really no association. If you can cover up one or two points and make it look as if there is no association, there probably really is none.

When describing the information displayed on a scatter plot, you can discuss

1. whether there is positive, negative, or no association;

2. whether there are any clusters of points and whether the points in the clusters have anything in common; and

3. whether any points do not follow the general pattern.

It's not always safe to conclude that one variable *causes* another to happen (or not happen) just because there is an association.

Application 20

Box Office Hits

The table below shows production costs, promotion costs, and gross ticket sales for twelve of the most popular "dumb" movies. The box office grosses were obtained from studios and are estimates.

Dumbing for Dollars

	Year	Production Costs	Promotion Costs	Worldwide Ticket Sales
"Animal House"	1978	$2.9 million	$3 million	$150 million
"Meatballs"	1979	$1.4 million	$2 million	$70 million
"Caddyshack"	1980	$4.8 million	$4 million	$60 million
"Stripes"	1981	$10.5 million	$4.5 million	$85 million
"Spring Break"	1982	$4.5 million	$5 million	$24 million
"Porky's"	1982	$4.8 million	$9 million	$160 million
"Fast Times At Ridgemont High"	1982	$5 million	$4.9 million	$50 million
"Porky's II — The Next Day"	1983	$7 million	$7.5 million	$55 million
"Hot Dog — The Movie"	1984	$2 million	$4 million	$22 million
"Bachelor Party"	1984	$7 million	$7.5 million	$38 million
"Revenge of the Nerds"	1984	$7 million	$7.5 million	$42 million
"Police Academy"	1984	$4.5 million	$4 million	$150 million

Source: Peter H. Brown, "Dumbing for Dollars," *Los Angeles Times*, January 20, 1985.

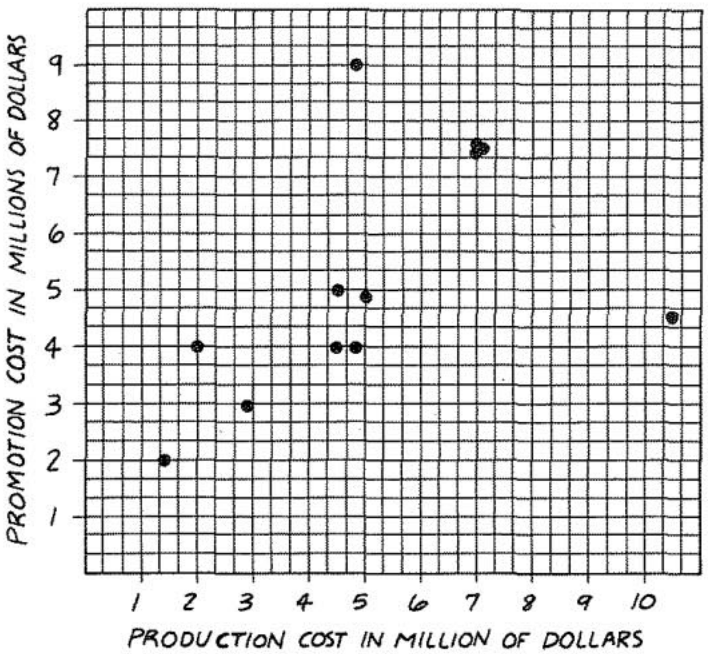
The scatter plot for total costs (production costs + promotion costs) and worldwide ticket sales follows.

Page 90

NOTE TO TEACHERS: All but one of Application 20, "Box Office Hits," Application 21, "Protein versus Fat," Application 22, "Walk-around Stereos," or Application 23, "SAT Scores," may be omitted.

Page 91: Application 20

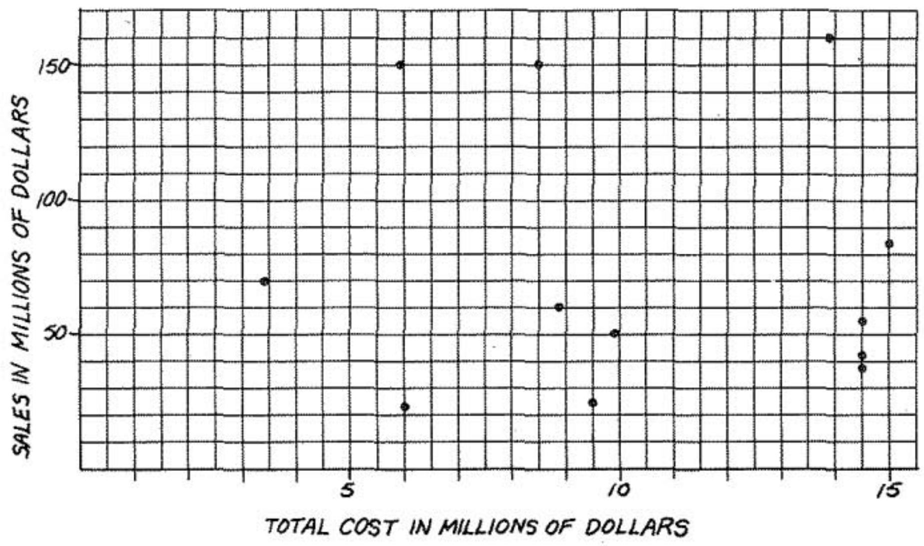
1. no association
2. *Animal House*
3. *Spring Break, Bachelor Party,* and *Revenge of the Nerds*
4.



5. positive association
6. about $4 to $5 million
7. *Stripes* and *Porky's*
8. Answers will vary but may contain the following information:

   There is no association between total costs and worldwide ticket sales for these movies. *Animal House* sold a lot of tickets at a relatively low total cost. The biggest grosser (so to speak) was *Porky's* with $160 million in sales; it was relatively expensive to make.

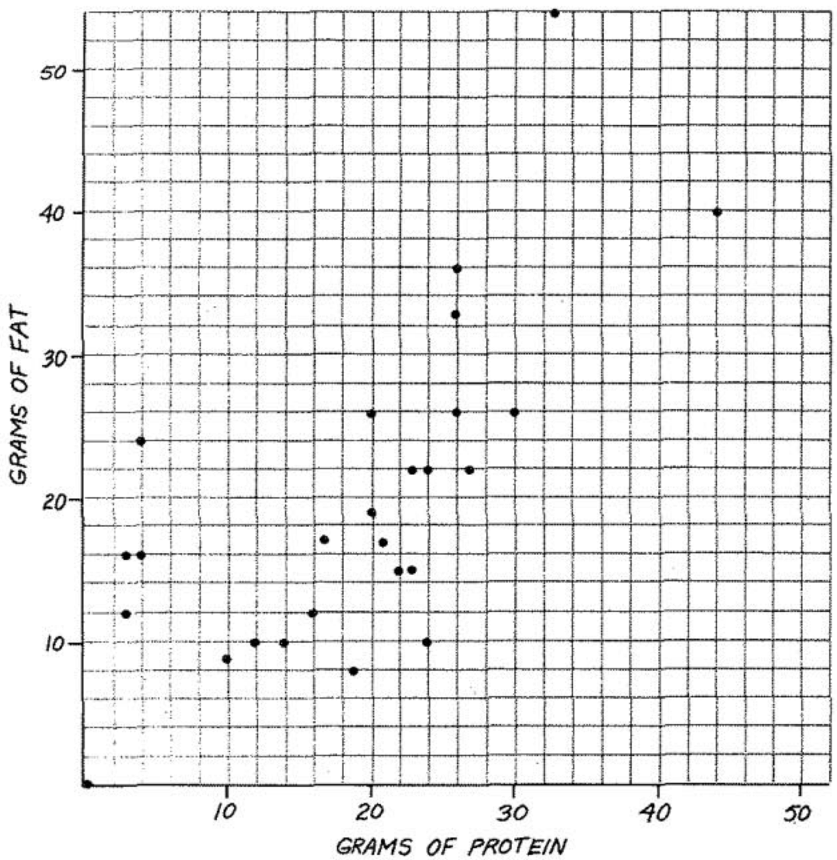
   There is a positive association between production costs and promotion costs. In general, the more spent producing a movie, the more spent promoting it. *Porky's* had the largest promotion costs, and these were also quite large compared to its production costs. *Stripes* had by far the largest production costs, but its promotion costs were relatively low.



1. Is there positive association, negative association, or no association between total costs and worldwide ticket sales?
2. Which movie(s) would you say did the best when costs are compared to ticket sales?
3. Which movie(s) would you say did the worst when costs are compared to ticket sales?
4. Make a scatter plot of promotion costs against production costs. Put production costs on the horizontal axis and promotion costs on the vertical axis.
5. Is there a positive, negative, or no association between production costs and promotion costs?
6. If a studio spends $4 million on production costs, about how much money would you expect the studio to spend promoting the movie?
7. Which two movies stand out on the scatter plot you made in question 4?
8. Write a description of the information displayed by the two scatter plots.

Protein versus Fat

The following scatter plot shows the grams of fat against the grams of protein in individual servings of lunch and dinner items sold at various fast food restaurants.



1. Suppose you want protein without much fat. Give the number of grams of protein and fat in the item that you would select.
2. What is the largest number of grams of protein in any item?

Page 92: Application 21

1. 24 grams of protein and 10 grams of fat, or 19 grams of protein and 8 grams of fat
2. 44

Page 93: Application 21 (*continued*)

3. 40
4. no
5. 0
6. 0
7. positive association
8. around 30
9. Yes; the four points at the left side are all low in protein and especially high in fat.

3. What is the number of grams of fat in the item in question 2?
4. Does the item in question 2 have an unusually large amount of fat considering how much protein it has?
5. What is the smallest number of grams of protein in any item?
6. How many grams of fat did the item in question 5 have?
7. Is there a positive, negative, or no association between grams of protein and grams of fat?
8. If a new item has 32 grams of protein, how many grams of fat would you expect it to have?
9. Do you see any clusters of points? Where?

The following table lists the items in the previous plot with their grams of protein and grams of fat.

	Protein grams	Fat grams
Big Mac — McDonald's	26	33
Cheeseburger — Hardee's	17	17
Double cheeseburger — Burger Chef	23	22
Cheeseburger w/Bacon Supreme — Jack-in-the-Box	33	54
Single — Wendy's	26	26
Double — Wendy's	44	40
Hamburger — McDonald's	12	10
Quarter Pounder — McDonald's	24	22
Whopper — Burger King	26	36
Roast beef — Arby's	22	15
Beef and cheese — Arby's	27	22
Roast beef — Hardee's	21	17
Big fish — Hardee's	20	26
Ham and cheese — Hardee's	23	15
Thick-crust cheese pizza — Pizza Hut	24	10
Super Supreme thin-crust pizza — Pizza Hut	30	26
Idiot's Delight pizza — Shakey's	14	10
Cheese pizza — Shakey's	16	12
Chicken McNuggets — McDonald's	20	19
Chili — Wendy's	19	8
French fries — McDonald's	3	12
Onion rings — Burger King	3	16
Chocolate shake — McDonald's	10	9
Apple turnover — Jack-in-the-Box	4	24
Chocolaty chip cookies — McDonald's	4	16
Carbonated beverages	0	0

Source: P. Hausman, *At-A-Glance Nutrition Counter*, 1984.

10. What is the item that you decided to order in question 1?
11. What kinds of items are in the cluster of question 9?
12. Do you see any single points in the scatter plot that could be outliers? That is, do you see points that don't follow the general relationship or that don't lie in a large cluster? If so, list the grams of protein and fat for those points. Which items are they? Can you give explanations for any of them?
13. With your fingers, cover up any points you identified for question 12 and the cluster from question 9, and look at the remaining points. Are they scattered fairly closely about a straight line?
14. Write a summary of the information displayed in the scatter plot.

Page 94: Application 21 (*continued*)

10. thick-crust cheese pizza, or chili
11. desserts and fried vegetables
12. Answers will vary. Sample: 0 grams of protein and 0 grams of fat—carbonated beverages. Also, 33 grams of protein and 54 grams of fat—the Cheeseburger w/Bacon Supreme. This item is the only one in the list that has bacon in it; maybe that is the reason it is relatively higher in fat than other hamburger-type sandwiches.
13. yes
14. Answers will vary. Sample: In general, the items with the most protein also have the most fat. If one is looking for high protein and low fat, the best bets are the thick-crust cheese pizza or the chili. A cluster of desserts—chocolate chip cookies and apple turnovers—and fries—potato and onion—are very low in protein and high in fat. One other item, the cheeseburger with bacon, is also high in fat relative to its amount of protein. Except for the items just mentioned, it is interesting that these items have about equal numbers of grams of fat and protein.

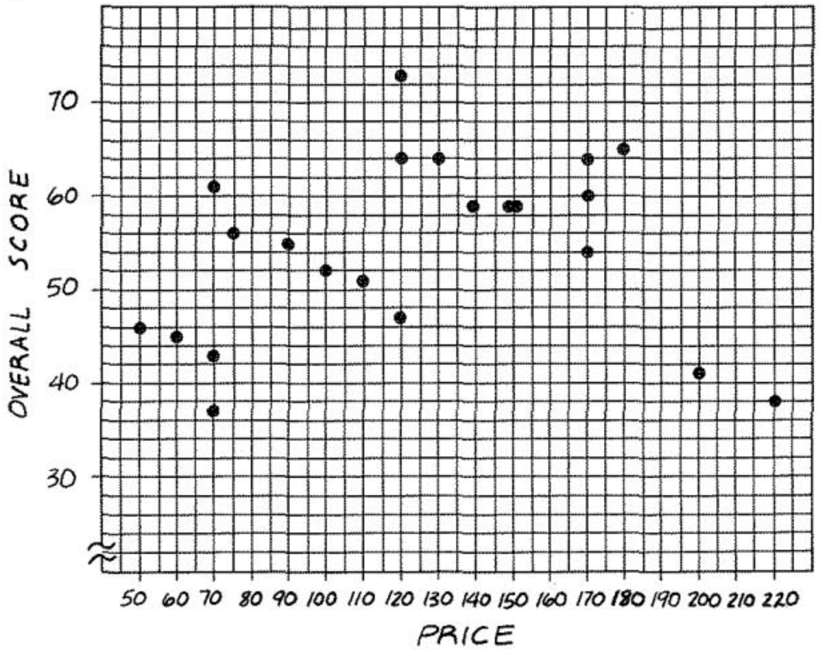
    It is interesting that three of the pizzas are in the main group of items although one, the thick-crust cheese, is relatively high in protein for the amount of fat. Is this because the crust of this pizza is thick, or because it perhaps has more or less cheese than the others, or because it perhaps does not have some high-fat topping the others have, or because of some other reason?

NOTE TO TEACHERS: In a scatter plot, a point can be an outlier in at least two different ways. One possibility is that the x (or y) value is itself an outlier compared to just the other x (or y) values, but the point follows the same relationship between x and y as do the rest of the data. An example here is Double-Wendys. Such an item might be similar in nature to others in the data but has values that are larger or smaller.

Another possibility is an item whose x and y values are not outliers when compared separately to the other x and y values; however, when taken together the (x, y) pair does not follow the same relationship as the other items. An example here is apple turnover—Jack-in-the-Box. Such an item is called a *bivariate outlier* because both variables together are required to show that the item is an outlier. Both kinds of outliers are important for interpreting scatter plots.

Page 95: Application 22

1. Answers will vary.
2.



Application 22

Walk-around Stereos

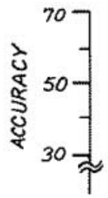
The following table lists 22 "walk-around stereos," each with its price and overall score. The overall score is based on "estimated overall quality as tape players, based on laboratory tests and judgments of features and convenience." A "perfect" walk-around stereo would have a score of 100. Consumers Union says that a difference of 7 points or less in overall score is not very significant.

Ratings of Walk-around Stereos

Brand and Model	Price	Overall Score
AIWA HSP02	$120	73
AIWA HSJ02	180	65
JVC CQ1K	130	64
Sanyo MG100	120	64
Sony Walkman WM7	170	64
Sanyo Sportster MG16D	70	61
Toshiba KTVS1	170	60
JVC CQF2	150	59
Panasonic RQJ20X	150	59
Sharp WF9BR	140	59
Sony Walkman WM4	75	56
General Electric Stereo Escape II 35275A	90	55
KLH Solo S200	170	54
Sanyo Sportster MG36D	100	52
Koss Music Box A2	110	51
Toshiba KTS3	120	47
Panasonic RQJ75	50	46
Sears Cat. No. 21162	60	45
General Electric Great Escape 35273A	70	43
Sony Walkman WMR2	200	41
Sony Walkman WMF2	220	38
Realistic SCP4	70	37

Source: *Consumer Reports Buying Guide*, 1985.

1. Which walk-around stereo do you think is the best buy?
2. A scatter plot will give a better picture of the relative price and overall score of the walk-around stereos. Make a scatter plot with price on the horizontal axis. You can make the vertical axis as follows:



The ≈ lines indicate that part of the vertical axis is not shown, so that the plot is not too tall.

3. Which stereo appears to be the best buy according to the scatter plot?
4. Is there a positive, negative, or no association between price and overall score?
5. Given their overall scores, which walk-around stereos are too expensive?

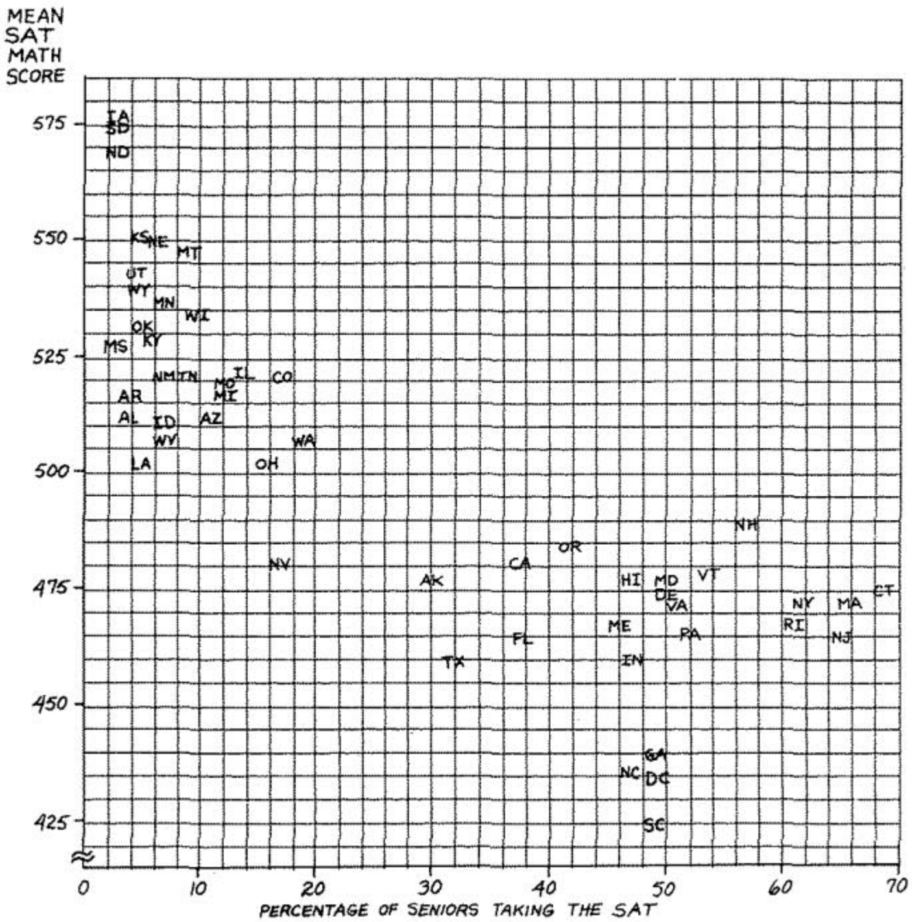
Page 96: Application 22 (*continued*)

3. Sanyo Sportster MG16D, or possibly the AIWA HSP02
4. With the exception of the two points at the far lower right, the remaining 20 show a positive association. These two look so different, though, that we could say that overall the 22 points show no association.
5. Sony Walkman WMR2 and Sony Walkman WMF2

Application 23

SAT Scores

The following plot shows the SAT math scores in each state in 1985 against the percentage of seniors in each state who took the test. Each state is identified by its postal code. For example, Mississippi is MS. The nationwide mean was 475.

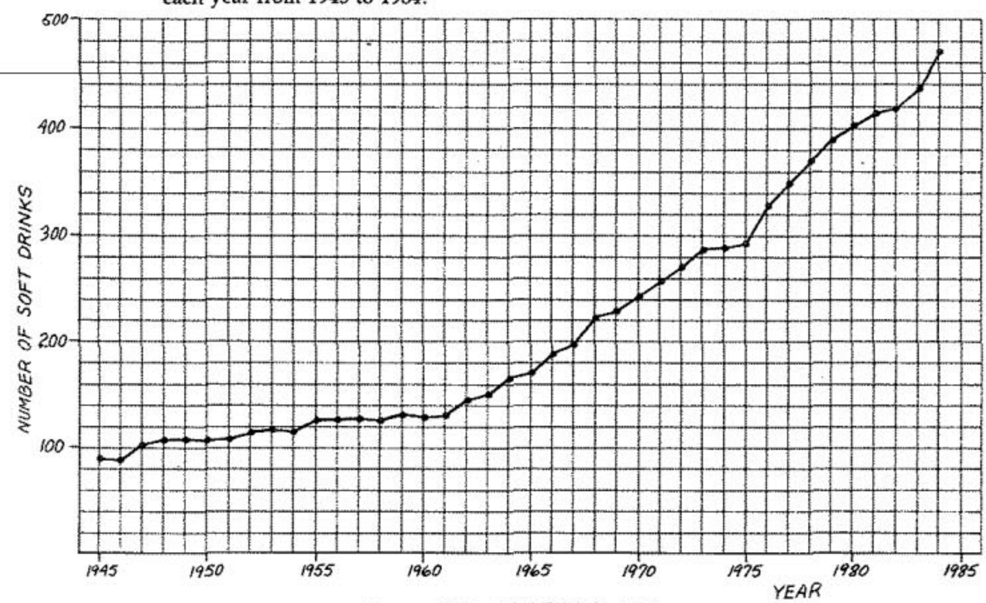


Source: The College Board.

1. In general, as a larger percentage of students take the test, what happens to the SAT math score?
2. Find the two clusters of states. Within the cluster on the left, is there a positive, negative, or no association between the percentage taking the test and the score?
3. Within the cluster on the right, is there a positive, negative, or no association?
4. Taking into account the percentage of students taking the test, which state(s) do you think have the best SAT math score? Which have the worst?
5. Using the facts you discovered in questions 1 through 4, write a summary of the information given in the scatter plot. Include an analysis of the position of your state.

Time Series Plots

Some scatter plots have year or some other time interval on the horizontal axis. Since there is only one value per year, we can connect the points in order to see the general trend. For example, the following *plot over time* shows how many 12-ounce soft drinks the average person in the U.S. drank each year from 1945 to 1984.



Source: National Soft Drink Association.

Page 98: Application 23

1. It goes down.
2. negative association
3. no association
4. For the cluster of states with 30 percent or more taking the SAT, New Hampshire and then Oregon have the highest scores. For the cluster with less than 20 percent taking the SAT, the states on the upper boundary are Iowa, South Dakota, Montana, and Colorado. These are the states that seem to be doing better, but it is hard to say one state is *the best.* Similarly, for the worst we look at the bottom of the two clusters, finding Louisiana, Nevada, North Carolina, Georgia, the District of Columbia, and South Carolina.
5. Answers will vary. Sample: The states fall into two clusters. Among states with fewer than 20 percent of the seniors taking the test, the mean SAT score ranges from about 575 in Iowa to 480 in Nevada. In general, the more students taking the test in this cluster, the lower the score.

   In the other cluster, with 30 percent or more of the students taking the test, there is no association between the percentage taking the test and the mean SAT score. All the states in the second cluster have a lower mean SAT score than every state in the first cluster except Nevada. Their mean scores range from about 490 in New Hampshire to about 425 in South Carolina.

   For a discussion of states that scored relatively high or low, see the answer to question 4. To analyze the position of a specific state, you will want to consider its position within its cluster. You may also want to consider some states that are geographic neighbors and see where they are in this plot.

Page 99: Discussion Questions

1. 105; 242
2. 410/6 or 68
3. 2; 8
4. about 600
5. 1960; introduction of diet drinks, introduction of aluminum cans
6. The amount the "average person" drank is computed by dividing the total number of 12-ounces consumed by the number of people in the United States.
7. Answers will vary but should resemble the summary in the text.

Discussion Questions

1. About how many soft drinks did the average person drink in 1950? In 1970?
2. About how many six-packs of soft drinks did the average person drink in 1980?
3. About how many soft drinks did the average person drink *per week* in 1950? In 1980?
4. If the trend continues, about how many 12-ounce soft drinks will the average person drink each year in 1990?
5. In what year did soft drink consumption start to "take off"? Can you think of any reason for this?
6. Who is the "average person"?
7. Write a summary of the trend in soft drink consumption shown by the plot. (Our summary of this plot follows.)

In the U.S. from 1945 until 1961, soft drink consumption rose gradually from about 90 twelve-ounce servings per year per person to about 130 twelve-ounce servings. In 1962, soft drink consumption started to rise rapidly until it was about 400 twelve-ounce servings in 1980. In other words, in these 18 years, soft drink consumption more than tripled in the United States.

What happened in 1962? Some ideas are as follows:

- Diet drinks might have been introduced.
- Soft drinks in aluminum cans might have become available.
- The economy might have improved so people started to spend more money on luxuries such as soft drinks.
- The post-war baby boom kids were reaching their teenage years.

There were very big increases in the late 70's. Then, the increase showed signs of leveling off. However, there were large increases again in 1983 and 1984.

Application 24

How Long Can You Expect to Live?

1. Study the table below. At your birth, how long could you expect to live?

Life Expectancy at Birth

Birth Year	White		Black and Other	
	Male	Female	Male	Female
1920	54.4	55.6	45.5	45.2
1930	59.7	63.5	47.3	49.2
1940	62.1	66.6	51.5	54.9
1950	66.5	72.2	59.1	62.9
1955	67.4	73.7	61.4	66.1
1960	67.4	74.1	61.1	66.3
1965	67.6	74.7	61.1	67.4
1970	68.0	75.6	61.3	69.4
1971	68.3	75.8	61.6	69.7
1972	68.3	75.9	61.5	69.9
1973	68.4	76.1	61.9	70.1
1974	68.9	76.6	62.9	71.3
1975	69.4	77.2	63.6	72.3
1976	69.7	77.3	64.1	72.6
1977	70.0	77.7	64.6	73.1
1978	70.2	77.8	65.0	73.6
1979, preliminary	70.6	78.3	65.5	74.5

Source: United States National Center for Health Statistics.

2. Can males or females expect to live longer?

3. Can whites or blacks and others expect to live longer?

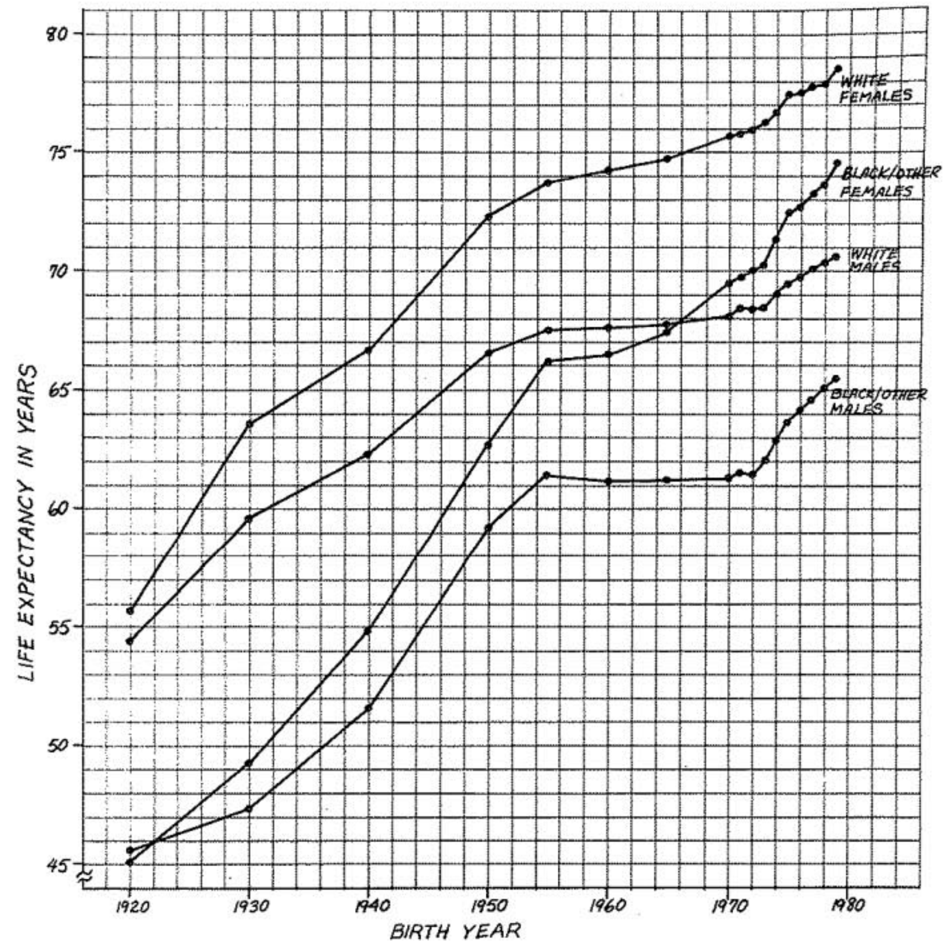
The life expectancies for each group have been placed on the following plot and the points have been connected by a line.

Page 100

NOTE TO TEACHERS: It is necessary to do only one of Application 24, "How Long Can You Expect to Live," Application 25, "Speeding," or Application 26, "Sex Ratio by Age."

Application 24

1. Answers will vary
2. females
3. whites


WHITE FEMALES
BLACK/OTHER FEMALES
WHITE MALES
BLACK/OTHER MALES
LIFE EXPECTANCY IN YEARS
80
75
70
65
60
55
50
45
1920
1930
1940
1950
1960
1970
1980
BIRTH YEAR

4. Which group born in 1979 could expect the longest life?
5. Which group made the greatest gain in life expectancy in the years from 1920 to 1979?
6. Which group has had the smallest increase in life expectancy since 1920?
7. During which decade did the largest increase in life expectancy occur for black and other females?
8. Within each race, males and females had about the same life expectancy in 1920. Was this still true in 1979?
9. Write a summary of the trends you see in the plot.

Page 102: Application 24 (*continued*)

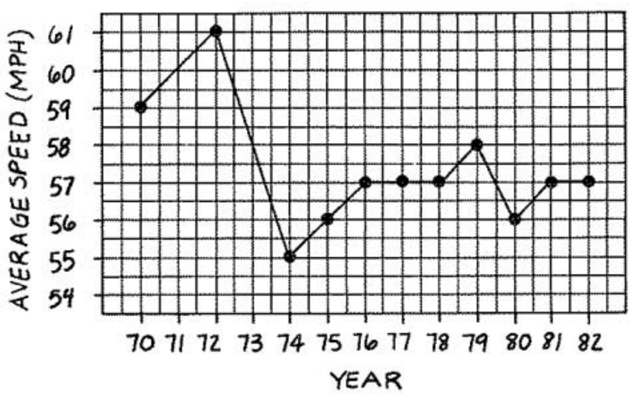
4. white females
5. black/other females
6. white males
7. 1940s
8. No; females live longer.
9. Answers will vary. Sample: All four groups have had large increases in life expectancy since 1920. The largest gain was made by black/other females and the smallest gain by white males.

   In 1920, black/others of both genders could expect to live to be about 45 years old and whites to be about 55 years old. This 10-year difference between white and black/others was reduced to about 5 years in 1980. Within each race, a separation between genders has occurred so that a female born in 1980 can expect to live 8 or 9 years longer than a male. In fact, black/other females can expect to live longer than white males.

   The greatest gains in life expectancy occurred in the years 1920–1955 and 1972–1980.

Page 103: Application 25

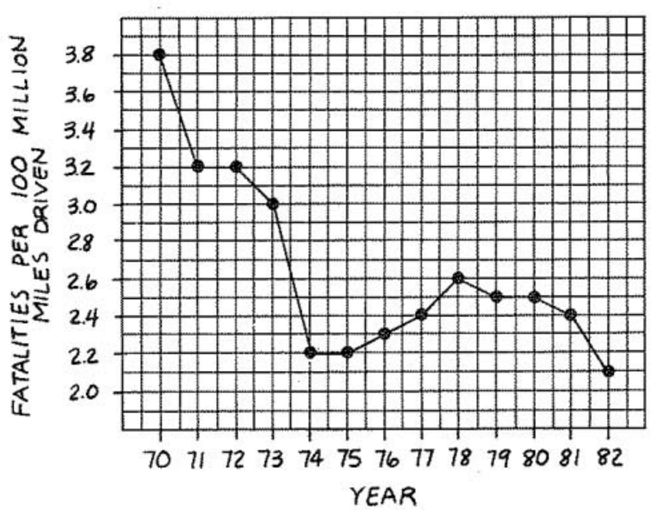
1.



2. 1973 or 1974

3. Although the average speed did drop to 55 miles per hour in 1974, it has since risen a bit above the speed limit. However, the average speed has been fairly constant from 1976 to 1982 (57 miles per hour plus or minus 1), and drivers have not been increasing their speed from one year to the next. Drivers are going slower than in the early 1970s.

4.



Application 25

Speeding

The following table shows average freeway speeds as recorded by highway monitoring devices in California. The newspaper gave no explanation why the average speed is missing for 1971 and 1973.

Year	Average Highway Speed in Miles per Hour
1970	59
1971	—
1972	61
1973	—
1974	55
1975	56
1976	57
1977	57
1978	57
1979	58
1980	56
1981	57
1982	57

Source: *Los Angeles Times*, May 22, 1983.

1. Construct a plot over time of the average speeds.
2. Can you guess what year the 55 miles per hour speed limit went into effect?
3. Some people think drivers are ignoring the 55 miles per hour speed limit. Do you think your plot shows that this is the case?
4. The fatalities in California per 100 million miles driven are shown in the following table. Construct a plot over time of these data.

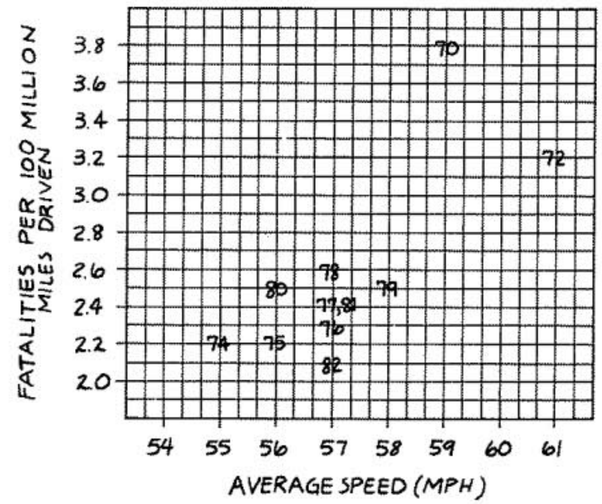
Year	Fatalities per 100 Million Miles
1970	3.8
1971	3.2
1972	3.2
1973	3.0
1974	2.2
1975	2.2
1976	2.3
1977	2.4
1978	2.6
1979	2.5
1980	2.5
1981	2.4
1982	2.1

Source: *Los Angeles Times*, May 22, 1983.

5. Was there a decrease in fatalities when the 55 miles per hour speed limit took effect?
6. Another way to display these data is with a scatter plot of fatalities against speed. Construct such a plot. Place the values for speed on the horizontal axis. Plot the last two digits of the year instead of a dot.
7. What do you learn from the plot in question 6?
8. Why is the plot in question 6 the best one?

Page 104: Application 25 (*continued*)

5. Yes. This plot suggests it took effect in 1974, not 1973.
6.



7. Answers will vary. Sample: Two years, 1970 and 1972, had a much higher number of fatalities per 100 million miles driven. The average freeway speed was also higher for these two years.
   The other years are in a cluster. Within the range for this cluster, 55 to 58 miles per hour, there does not appear to be a strong relationship between speed and fatalities, but there is more indication of a positive than a negative association.
8. It contains all of the information—the year, fatalities, and speed—and it shows clearly the relationship between speed and fatalities.

Page 105

NOTE TO TEACHERS: If your students have had algebra, ask them this question: If the sex ratio is 0.600, what is the percentage of males? To find the answer, in general, let P be the proportion of males and let r be the sex ratio. Then

$$\frac{P}{1-P} = r$$

$$P = r(1-P)$$

$$P = r - rP$$

$$P + rP = r$$

$$P(1+r) = r$$

$$P = \frac{r}{1+r}$$

So in this case, $P = \dfrac{0.600}{1+0.600}$

$= 0.375.$

The answer is 37.5 percent.

For a project, a student might want to compare this plot of sex ratio over time to one of percentage of males over time. Will the plots have the same shape?

Application 26

1. 1.500
2. 0.714
3. the same
4. more males than females
5. fewer males
6. a. whites
   b. "others"

Application 26

Sex Ratio by Age

The following table gives the ratio of males to females at different ages for whites, blacks, and other races in 1980. The sex ratio is computed by dividing the number of males by the number of females.

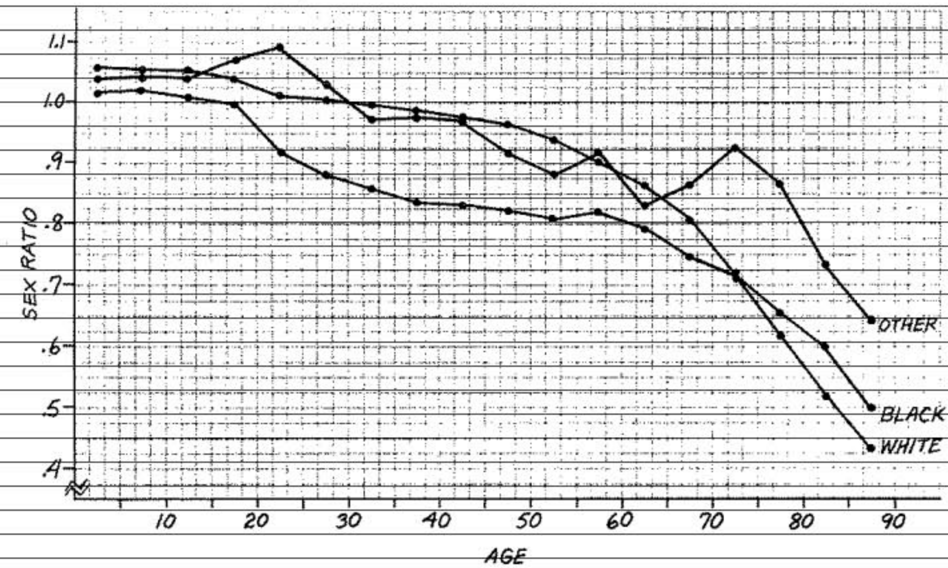
Sex Ratio by Age (total number male/total number female)

Age	White	Black	Other
0-4	1.054	1.016	1.035
5-9	1.053	1.016	1.036
10-14	1.050	1.011	1.035
15-19	1.037	.995	1.073
20-24	1.009	.913	1.087
25-29	1.003	.877	1.026
30-34	.994	.856	.971
35-39	.983	.832	.972
40-44	.974	.828	.973
45-49	.963	.821	.917
50-54	.939	.808	.878
55-59	.901	.818	.913
60-64	.869	.793	.864
65-69	.804	.745	.863
70-74	.720	.712	.925
75-79	.620	.651	.865
80-84	.524	.599	.730
85-	.429	.500	.642

Source: United States Census Bureau.

1. If there are 750 males and 500 females, what is the sex ratio?
2. If there are 500 males and 700 females, what is the sex ratio?
3. If the sex ratio is 1.000, are there more males than females, fewer males than females, or the same number of males as females?
4. If the sex ratio is 1.213, are there more males, fewer males, or the same number of males as females?
5. If the sex ratio is 0.736, are there more males, fewer males, or the same number of males as females?
6. Is there a higher percentage of males among
   a. 0-4 year old whites, 0-4 year old blacks, or 0-4 year old "others"?
   b. 80-84 year old whites, 80-84 year old blacks, or 80-84 year old "others"?

The following scatter plot shows the curves for whites, blacks, and "others."



7. Do the three curves look about the same overall?
8. What is one general characteristic of all three curves?
9. What does it mean when the curve is going downhill?
10. Where are the curves closest?
11. a) At what ages do the curves for whites and blacks first start separating?
    b) Can you think of any possible explanations for this?
12. a) How do the white and black curves compare at older ages?
    b) Again, can you think of any possible explanations for this?
13. Write a description of the information you see in this plot. Include any questions the plot suggests to you.

Page 106: Application 26 (*continued*)

7. yes
8. They go downhill.
9. The percentage of males is decreasing.
10. at ages 0–9
11. a. about 20–24
    b. For some reason, black males are dying at a rapid rate in these years.
12. a. very close again
    b. For some reason, older white males die at a more rapid rate compared to females than do older black males.
13. Answers will vary. Sample: At ages 0–4, there are slightly more males than females. This is true for whites, blacks, and others. However, there is a fairly steady drop in the sex ratio until by age 85+, there are 429 white males for every 1,000 white females, 500 black males for every 1,000 black females, and 642 other males for every 1,000 other females.

    It would be interesting to find out why the sex ratio *increases* for "others" during the teenage years and during the seventies. Why are females dying at a faster rate than males?

    We can also see that the black and white curves are close until about 20–24 and then separate until 70–74. For ages in the middle, whites have a higher percentage of males than do blacks. It would be interesting to learn why black males die at a more rapid rate compared to females than white males for these ages. However, for the oldest ages this effect reverses and, for some reason, blacks have a larger percentage of males than do whites.

Scatter Plots — Summary

Scatter plots are the best way to display data in which two numbers are given for each person or item. When you analyze a scatter plot, look for the following:

- positive, negative, or no association
- clusters of points
- points that do not follow the general pattern

If you find any of these features, ask yourself what could have caused them.

On time series plots, it is often helpful to connect the points in order to see the trend. Look for places where the general trend seems to change, and try to find possible explanations. If there is more than one time series on a plot, compare them to determine similarities and differences.

Suggestions for Student Projects

Think of a problem that interests you or select one of those below. Collect the data, make the appropriate plot(s), and write a summary of your results. Try to explain any trends or patterns.

1. Did the students who studied the most hours tend to get the higher grades on your last test?
2. Do students who get the most allowance tend to work more hours doing chores at home?
3. Can the students who do the most sit-ups in one minute also do the most push-ups?
4. Investigate whether there are relationships between certain physical characteristics by measuring a group of students. Some possibilities include the following:
   a. height and elbow-hand length
   b. circumference of closed fist and length of foot
   c. hand span and circumference of wrist
   d. weight and waist
   e. circumferences of head and neck
5. Construct a plot over time of the number of absences in your class on each day of the last six weeks. What trends do you see?

VII. LINES ON SCATTER PLOTS

The 45° Line

In the last section we interpreted scatter plots by looking for general relationships of positive, negative, and no association. We also looked for clusters of points that seemed special in some way. This section shows how interpretations of scatter plots are sometimes helped by adding a straight line to the plot. Two different straight lines are used. One is the 45° line going through the points (0, 0), (1, 1), (2, 2), and so forth. The second type is a straight line that is fitted to go through much of the data.

This table lists the number of black state legislators for each state in 1974 and 1984.

Number of Black State Legislators

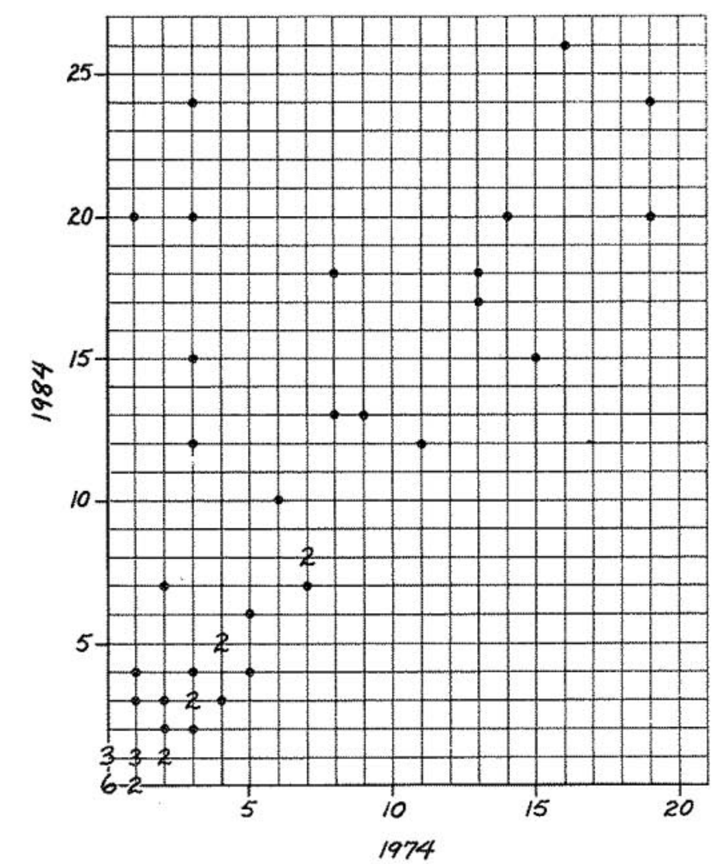
	1974	1984		1974	1984
Alabama	3	24	Montana	0	0
Alaska	2	1	Nebraska	1	1
Arizona	2	2	Nevada	3	3
Arkansas	4	5	New Hampshire	0	0
California	7	8	New Jersey	7	7
Colorado	4	3	New Mexico	1	0
Connecticut	6	10	New York	14	20
Delaware	3	3	North Carolina	3	15
District of Columbia	n/a	n/a	North Dakota	0	0
Florida	3	12	Ohio	11	12
Georgia	16	26	Oklahoma	4	5
Hawaii	0	0	Oregon	1	3
Idaho	0	0	Pennsylvania	13	18
Illinois	19	20	Rhode Island	1	4
Indiana	7	8	South Carolina	3	20
Iowa	1	1	South Dakota	0	0
Kansas	5	4	Tennessee	9	13
Kentucky	3	2	Texas	8	13
Louisiana	8	18	Utah	0	1
Maine	1	0	Vermont	0	1
Maryland	19	24	Virginia	2	7
Massachusetts	5	6	Washington	2	3
Michigan	13	17	West Virginia	1	1
Minnesota	2	1	Wisconsin	3	4
Mississippi	1	20	Wyoming	0	1
Missouri	15	15	Total	236	382

Source: Joint Center for Political Studies.

The scatter plot of the 1984 number against the 1974 number follows:

Page 109: Discussion Questions

1. 14



A striking feature of the plot is that the points all seem to lie above an (imaginary) diagonal line. Another feature is that there are many points in the lower left-hand corner. In fact, several states sometimes lie at exactly the same point. For example, Arkansas and Oklahoma both lie at (4, 5). To show this, we placed a 2 at (4, 5).

Discussion Questions

1. Place a ruler on the plot next to the line going through (0, 0), (10, 10), (20, 20), and so forth. For states on this line, the 1984 and 1974 numbers of black legislators are equal. How many points are exactly on this line?

2. If a point is above this line, the number of black legislators in that state in 1984 is larger than the number of black legislators that state had in 1974. Name three states for which this statement is true.

3. How many points fall below this line? What can we say about these states? What is the maximum (vertical) distance any of these is below the line? What does this mean in terms of the number of black legislators in 1974 and 1984?

4. Again, consider states above this line, those where the number of black legislators was larger in 1984 than in 1974. What are the names of the 7 or so states that lie farthest above the line? What do these states have in common?

5. The number of black legislators has generally increased from 1974 to 1984. Does this mean that the percentage of legislators who are black has necessarily increased? (Hint: Is the total number of legislators in a state necessarily the same in 1984 as in 1974?)

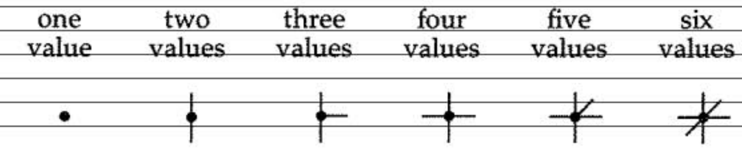
In summary, this 45° line (sometimes called the $y = x$ line) divides the plot into two regions. We should try to distinguish the characteristics of the points in the two regions. In this plot the top region contains states where the number of black legislators in 1984 is larger than it was in 1974. Most of the states lie in this region. The points in this region that are farthest from the line are those where the number has increased the most from 1974 to 1984. These states turn out to be states in the deep south. There are only a few points slightly below the 45° line, where the number of black legislators was greater in 1974 than in 1984. These are all states that had only 5 or fewer black legislators in 1974. Almost half the states are in the lower left-hand corner, with 5 or fewer in both years. Two states, Illinois and Maryland, had relatively large numbers in both years.

It would have been helpful to plot each state's abbreviation (such as NY for New York) instead of a dot. However, there wasn't room to do this for the states in the lower left corner.

Page 110: Discussion Questions (*continued*)

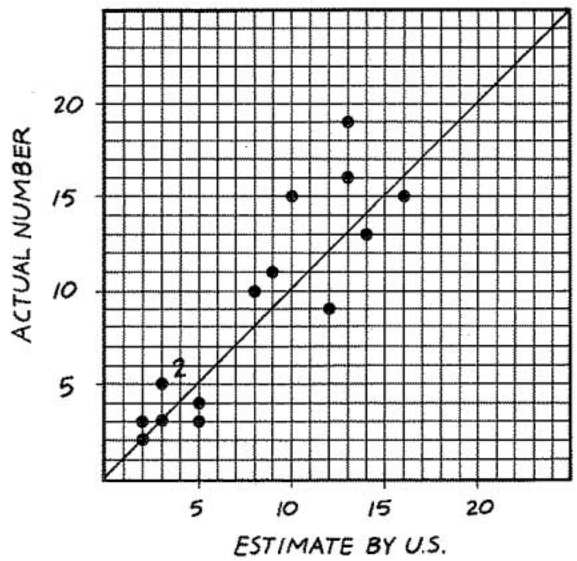
2. Answers will vary.
3. 7; there were fewer black legislators in 1984 than there were in 1974; 1; this means that there was, at most, one less black legislator in 1984 than in 1974 for any state.
4. Alabama, South Carolina, Mississippi, North Carolina, Florida, Louisiana, and Georgia; these southern states were the states that increased the number of black legislators by the largest amount.
5. Answers will vary. Sample: No; for example, suppose that a state had 15 black legislators out of 50 in 1974 and, as the population of the state had increased, there were 80 legislators in 1984 with 20 black. Then, in 1974 blacks were 15/50 or 30 percent of the legislators, but in 1984 they were only 20/80 or 25 percent.

NOTE TO TEACHERS: If there are, say, four values that lie at the same point on the graph, there is a way to show this other than writing the numeral 4 at that point. This method involves making a "sunflower" as we plot points. The first time a value comes up, we plot it as usual with a dot. The second time that value needs to be plotted, we add two "petals" and the dot becomes a small sunflower (see below), the two petals showing that there are two values at that point. For the third value, we add a third petal; for the fourth value, we add a fourth petal, then a fifth, a sixth, and so on, as shown in the following progression:



Page 111: Application 27

1.



2. See the preceding plot.
3. too low
4. above
5. underestimate
6. the point for month 14; 6 units
7. 4

Application 27

Submarine Sinkings

During World War II, the United States Navy tried to estimate how many German submarines were sunk each month. After the war, the Navy was able to get the actual numbers. The results follow:

Month	U.S. Estimate	Actual Number of Sinkings
1	3	3
2	2	2
3	4	6
4	2	3
5	5	4
6	5	3
7	9	11
8	12	9
9	8	10
10	13	16
11	14	13
12	3	5
13	4	6
14	13	19
15	10	15
16	16	15

Source: Mosteller, Fienberg, and Rourke, *Beginning Statistics with Data Analysis.*

1. Make a scatter plot of the data. Put the U.S. estimate on the horizontal axis.
2. Draw in the line that connects all the points where the number estimated by the U.S. Navy would be the same as the actual number of sinkings.
3. If a point is above the line, does it mean that the U.S. Navy's estimate was too high or too low?
4. Are more points above the line or below it?
5. Did the U.S. Navy tend to underestimate or overestimate the number of submarine sinkings?
6. Which point is farthest from the line? How many units away from the line is it? (Count the units vertically from the point to the line.)
7. How many points are three units or more from the line?

Fitting a Line

Not all ducks look alike, and it turns out that not all species of ducks behave alike, either. In an effort to study possible relationships between looks and behavior of ducks, two scales were created and an experiment performed. A plumage scale was devised to reflect the color and other characteristics of the duck's feathers. The scale ranged from 0 (looks just like a mallard with a green head and white neck-ring) to 20 (looks just like a pintail with a needle tail and neck stripe). Similarly, a behavior scale was devised ranging from 0 (generally congregate in pairs, just like mallards) to 15 (generally congregate in larger groups, just like pintails). The crucial scientific question is: After some interbreeding of mallards and pintails to produce ducks with a variety of looks and behavior, will we be able to predict how the ducks behave from their looks?

An experiment was performed. Mallards were mated with pintails and 11 second generation males were studied. For ease of identification, we have named the ducks. The results follow:

Duck	Plumage	Behavior
Rub	7	3
Stu	13	10
Ugly	14	11
Fred	6	5
Y.U.	14	15
Kold	15	15
Don	4	7
Ole	8	10
Van	7	4
Joe	9	9
Lou	14	11

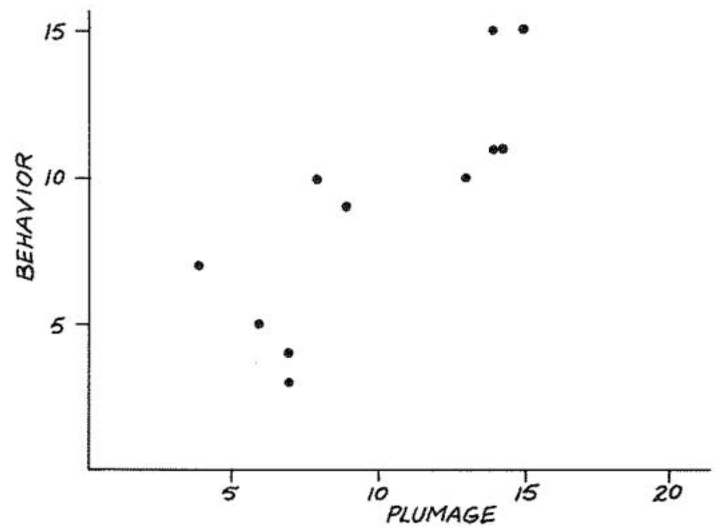
Source: Richard J. Larsen and Donna Fox Stroup, *Statistics in the Real World.*

Kold Duck looked the most like a pintail. Don Duck looked the most like a mallard. The scatter plot of these data follows:

Page 112

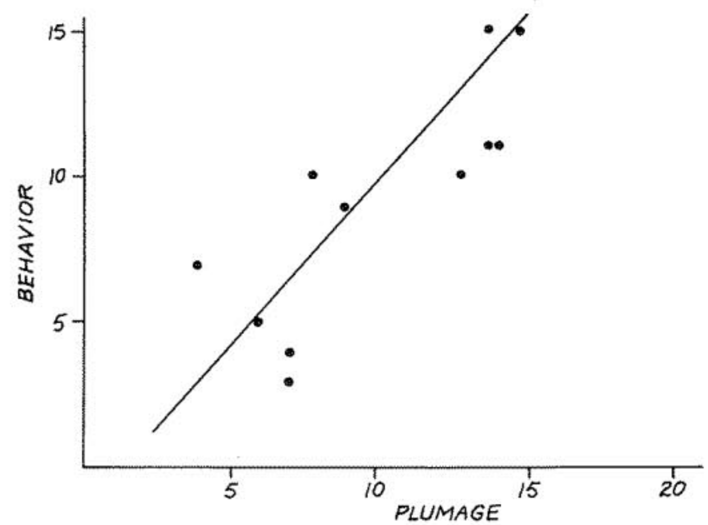
NOTE TO TEACHERS: The line introduced in this section is called a *robust* (or *resistant*) *regression* line. It is a simple alternative to the least squares regression line usually taught in introductory statistics classes. The robust line is not so affected by extreme values or outliers as is the least squares line. When there are no outliers, both lines will be about the same, but when there are outliers, the robust line fits the data better. Thus, for fitting real data, the robust line works about as well as or better than the least squares line. For a more advanced discussion and comparison of these methods, see Velleman and Hoaglin (Chapter 5) and Hoaglin et al. (Chapter 5).

Have students bring in pictures of a mallard and a pintail. Sometimes there will be a student in class who can describe the difference in behavior!

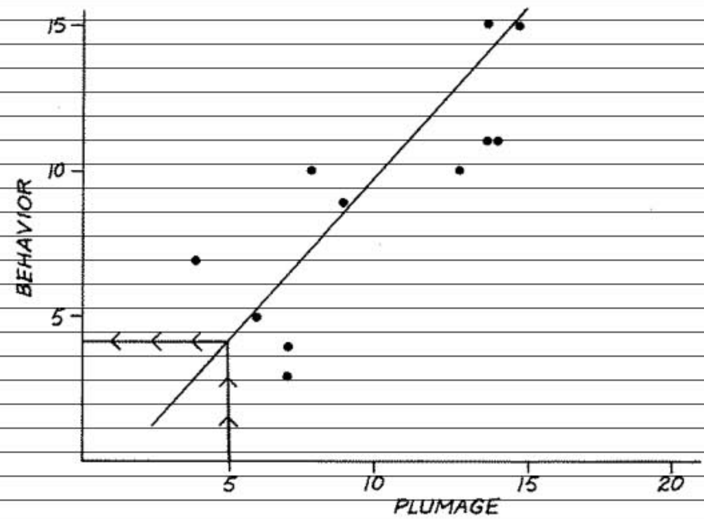


There is a positive association between a duck's plumage rating and his behavior rating. Ducks who look more like pintails tend to act more like pintails.

The same plot with a line through the points follows. This line is called the *fitted line.*



We can use this line to predict the behavior rating of a duck with a given plumage rating. For example, if a duck has a plumage rating of 5, what would you expect for his behavior rating?

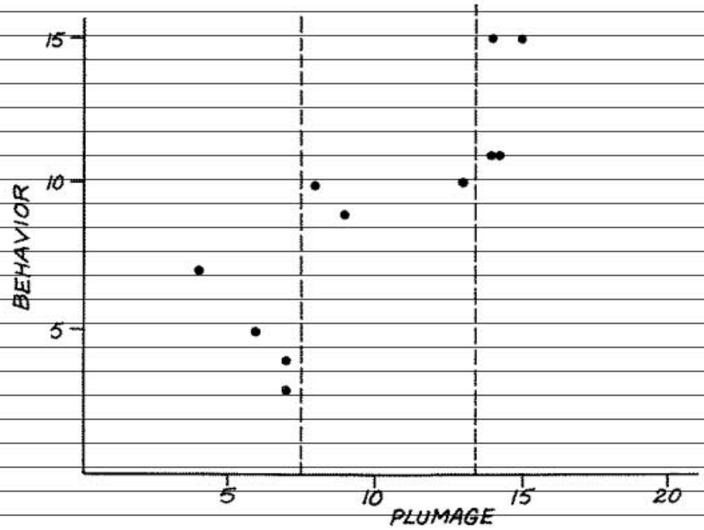


You should expect a behavior rating of 4.

Now we will describe a method for drawing a line through the data in order to predict a duck's behavior rating if we are given a plumage rating.

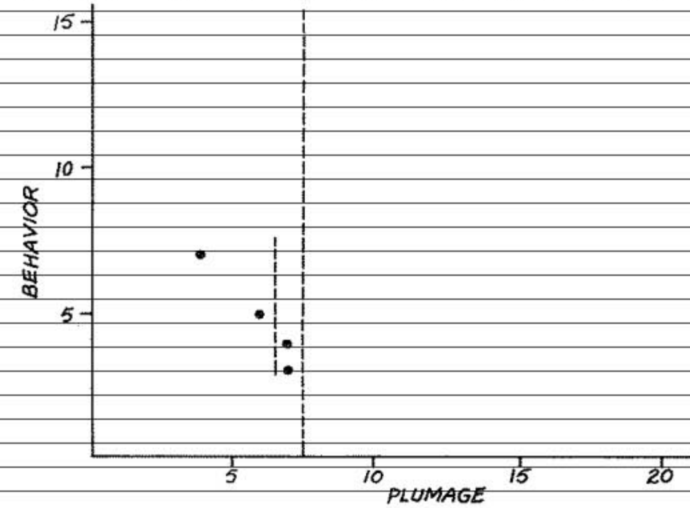
First, count the total number of points. Draw two vertical dashed lines so there are approximately the same number of points in each of the three strips. The two outer strips should have the same number of points, if possible.

In this case, we have 11 points. We will have four points in each outside strip and three points in the middle.

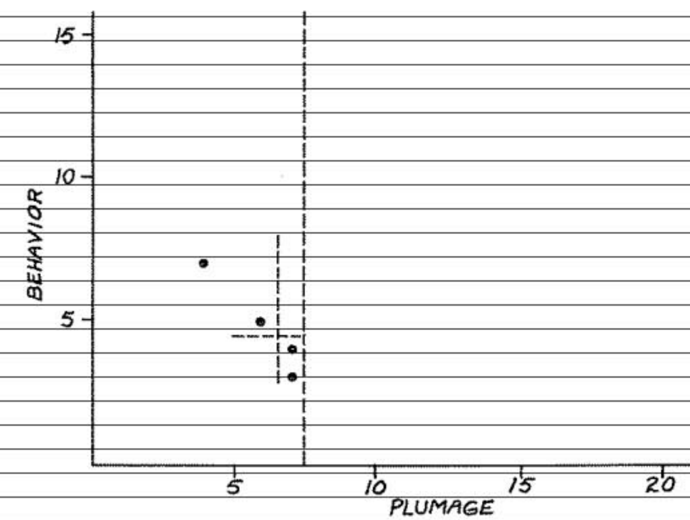


Second, place an X in each strip at the "center" of the points in that strip.

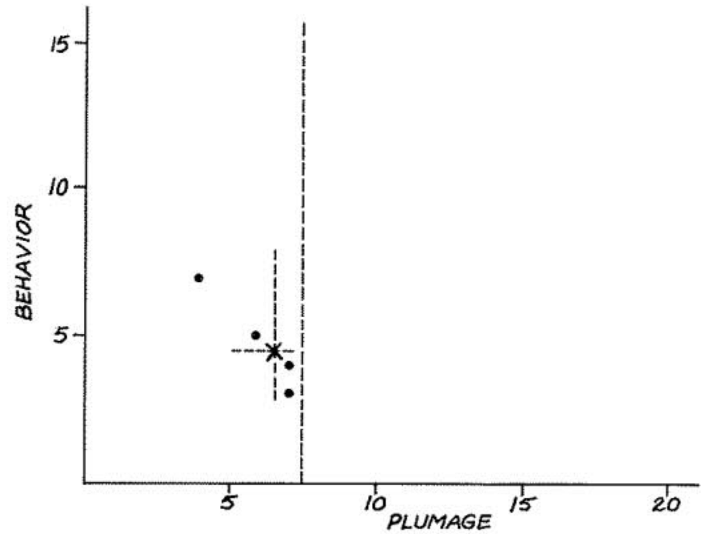
Study the left strip. It has four points. We want to find the median of the plumage ratings and the median of the behavior ratings. The median of the plumage ratings is halfway between the second and third points counting from the left. To find the median of the plumage ratings, place a ruler to the left of the points and move it toward the right until it is halfway between the second and third points. Draw a short vertical dashed line there.



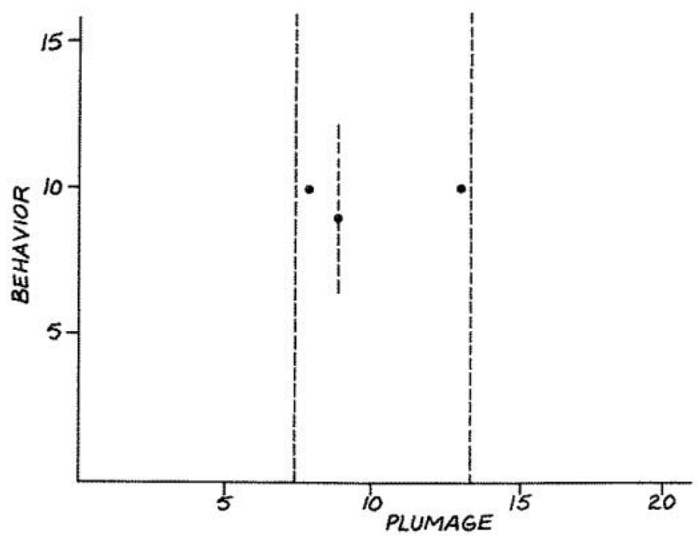
The median of the behavior ratings is halfway between the second and third points, counting from the bottom. Move the ruler up until it is halfway between these points and draw a horizontal dashed line there. The plot is shown as follows:



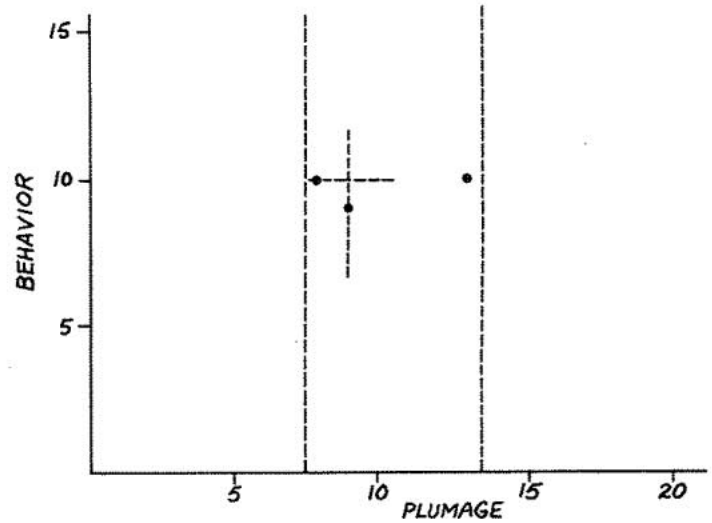
Mark an X where the dashed lines cross.



The center strip has three points. The median of the plumage ratings is at the second point, counting from the left.

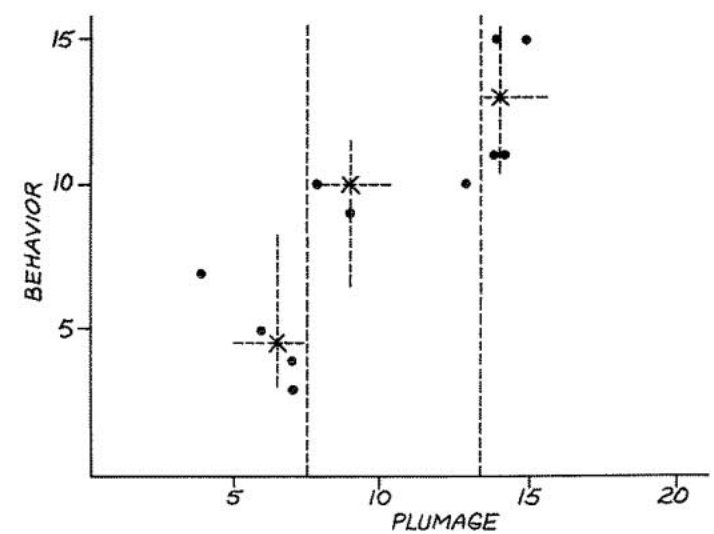


The median of the behavior ratings is at the second point, counting from the bottom.



Mark an X where the dashed lines cross.

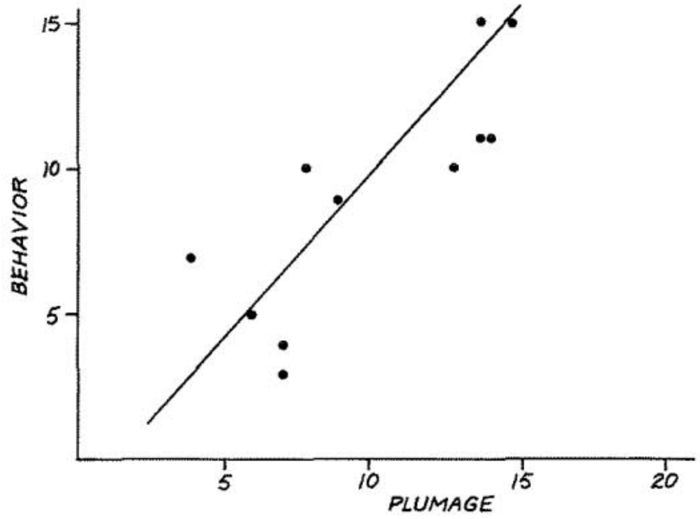
After the "center" of the right strip is also found, the plot looks as follows:



The third step is to decide whether or not the three X's lie close to a straight line. Use your ruler, balanced on its edge, to help decide. For this example, the X's lie approximately on a straight line.

Finally, place your ruler so that it connects the two X's in the outside strips. Now slide the ruler one-third of the way to the middle X and draw the line.

The finished plot including the fitted line is shown below. It is not necessary to include the dashed lines.



Discussion Questions

1. Which duck behaved the most like a pintail?
2. Which duck behaved the most like a mallard?
3. Why do we need a method for drawing a line? Why can't we just sketch one?
4. If a duck has a plumage rating of 10, what would you expect his behavior rating to be? Use the fitted line to get your answer.
5. If a duck has a plumage rating of 4, what would you expect his behavior rating to be?
6. To judge how much a duck's actual behavior differs from its predicted behavior, we measure the vertical distance from the point to the fitted line. Which duck is farthest from the line, and how many units is he from the line?
7. Which ducks are within two units of the line?
8. You might wonder why the fitted line has been constructed this way. Why have we used medians instead of means to form the X's? Why have we constructed three X's instead of two or four? Why have we constructed the slope of the line by using only the two end X's? After connecting the two end X's, why did we slide the ruler one-third of the way towards the middle X rather than some other fraction? Try to think of reasons for these choices or of alternate reasons for constructing a fitted line in a different way.

Page 118

NOTE TO TEACHERS: Question 6 introduces students to an important idea of statistics: residuals, or errors. The residual for a given duck is the difference between its actual behavior rating and the behavior rating predicted by the line. The sum of the squared residuals is often used in statistics as a measure of how well the line fits the data.

Discussion Questions

1. Y.U. and Kold
2. Rub
3. So we will all get the same line; sometimes it is hard to "eyeball" one accurately.
4. 9 or 10
5. 3 or 4
6. Rub, about 4
7. Y.U., Kold, Joe, and Fred
8. Answers will vary. Sample: Medians are not affected by a few outliers the way means are. With three X's, we can judge if the data follow a straight line at all by seeing if the X's approximately line up. If we used only two X's, they would automatically fall on a straight line, whether it makes sense to fit one or not. With four X's, it would be harder to use them sensibly to draw in the fitted line. To obtain the slope, we use the X's from the end strips so that we can get more stability in the estimate. To get the intercept, we slide the ruler one-third of the way because there are three X's and we want each to have equal importance for obtaining the intercept.

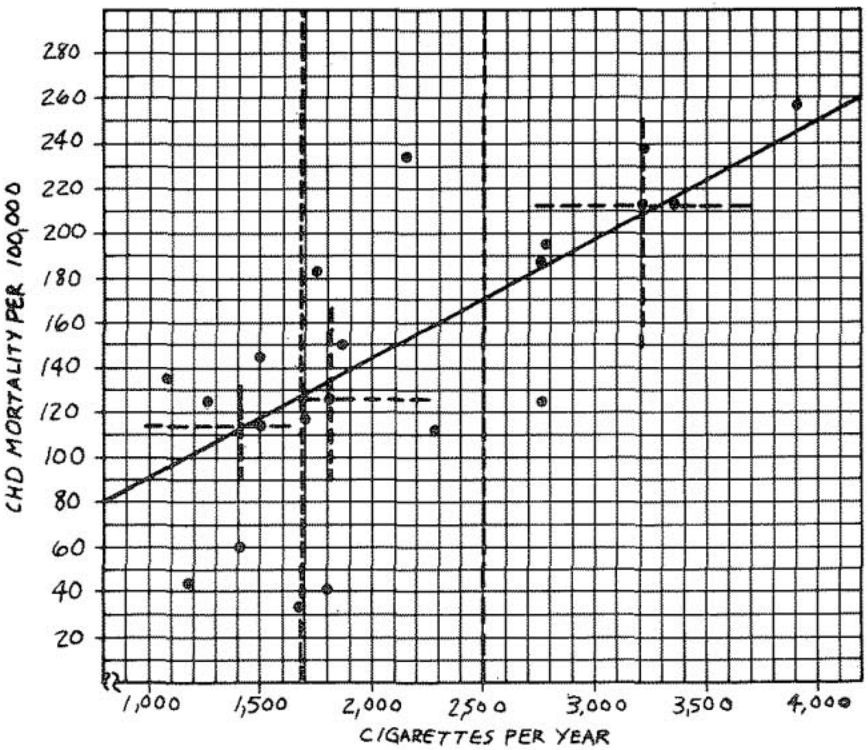
Page 119

NOTE TO TEACHERS: In this section, all but one of Application 28, "Smoking and Heart Disease," Application 29, "Catholic Clergy," or Application 30, "Voting for President," may be omitted.

If you feel that your students have had enough practice making scatter plots, a scatter plot that you can duplicate for students to use in answering question 5 appears on page 8 of this Teacher's Edition.

Application 28

1. United States
2. United States
3. Mexico
4. cigarette consumption
5. a. See the following plot.
   b. See the following plot.
   c. See the following plot.
   d. yes
   e. See the following plot.



Application 28

Smoking and Heart Disease

The following table lists 21 countries with the cigarette consumption per adult per year and the number of deaths per 100,000 people per year from coronary heart disease (CHD).

Country	Cigarette Consumption per Adult per Year	CHD Mortality per 100,000 (ages 35-64)
United States	3900	257
Canada	3350	212
Australia	3220	238
New Zealand	3220	212
United Kingdom	2790	194
Switzerland	2780	125
Ireland	2770	187
Iceland	2290	111
Finland	2160	233
West Germany	1890	150
Netherlands	1810	125
Greece	1800	41
Austria	1770	182
Belgium	1700	118
Mexico	1680	32
Italy	1510	114
Denmark	1500	145
France	1410	60
Sweden	1270	127
Spain	1200	44
Norway	1090	136

Source: *American Journal of Public Health.*

1. *In which country do adults smoke the largest number of cigarettes?*
2. Which country has the highest death rate from coronary heart disease?
3. Which country has the lowest death rate from coronary heart disease?
4. If we want to predict CHD mortality from cigarette consumption, which variable should be placed on the horizontal axis of a scatter plot?
5. a) Make a scatter plot of the data.
   b) Draw two vertical lines so there are seven points in each strip.
   c) Place an X in each strip at the median of the cigarette consumption and the median of the CHD mortality.
   d) Do the three X's lie close to a straight line?
   e) Draw in the fitted line.

6. a) Which three countries lie the farthest vertical distance from the line?
   b) How many units do they lie from the line?
   c) Considering the cigarette consumption, are these countries relatively high or low in CHD mortality?
7. If you were told that the adults in a country smoke an average of 2500 cigarettes a year, how many deaths from CHD would you expect?
8. If you were told that the adults in a country smoke an average of 1300 cigarettes a year, how many deaths from CHD would you expect?
9. (For class discussion) Sometimes strong association in a scatter plot is taken to mean that one of the variables *causes* the other one. Do you think that a high CHD death rate could cause cigarette consumption to be high? Could high cigarette consumption cause the CHD death rate to be high? Sometimes, though, there is not a causal relationship between the two variables. Instead, there is a hidden third variable. This variable could cause both of the variables to be large simultaneously. Do you think that this might be the situation for this example? Can you think of such a possible variable?
10. (For students who have studied algebra.) Choose two points on the fitted line, and from them find the equation of the line. Express it in the form $y = mx + b$, where y is mortality from coronary heart disease per 100,000 people (aged 35-64) per year, and x is cigarette consumption per adult per year. Using this equation, how many additional deaths per 100,000 people tend to result from an increase of 200 in cigarette consumption? What number of cigarettes per year is associated with one additional death from CHD per 100,000 people per year?

Page 120: Application 28 (*continued*)

6. a. Finland, Mexico, and Greece
   b. about 85
   c. Finland is high; Mexico and Greece are low.
7. about 170 per 100,000
8. about 105 per 100,000
9. Possible third variables are coffee drinking, stress, urbanization, or genetic differences among the nationalities.
10. Answers may vary slightly. The equation of the line obtained from the points (2,500, 170) and (1,500, 117) would be $y = 0.053x + 37.5$; 11; 1/0.053 or 19.

NOTE TO TEACHERS: A discussion of correlation versus causation should accompany this application. There is a positive association between cigarette consumption and CHD mortality. Does this positive association mean that cigarettes *necessarily* cause heart disease? It may provide some evidence that they do, but consider this: there is also a positive association between CHD mortality and cigarette consumption. (Think of CHD mortality on the x axis and cigarette consumption on the y axis.) Does this mean that heart disease causes cigarette smoking?

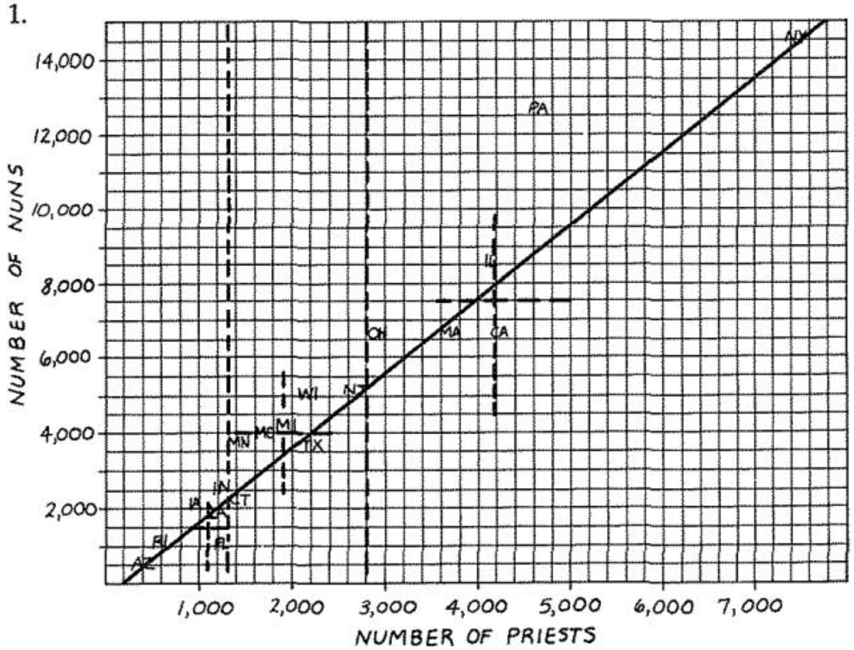
There are many examples of positive association between two variables when one does not cause the other. For example, in children there is a positive association between foot length and math achievement. But foot length doesn't *cause* math achievement. Instead, both variables increase with age.

Page 121

NOTE TO TEACHERS: A scatter plot that you can duplicate for students to use in answering question 1 appears on page 9 of this Teacher's Edition.

Application 29

1.



2. See the preceding plot.
3. yes
4. Yes; it's near the line.
5. Pennsylvania; California

Application 29

Catholic Clergy

Nineteen states have more than 500,000 residents who are Catholic. The following table lists these states, along with the number of priests and nuns in each state.

State	Number of Priests	Number of Nuns
Arizona	412	591
California	4242	6615
Connecticut	1298	2450
Florida	1224	1240
Illinois	4131	8564
Indiana	1229	2515
Iowa	982	2140
Louisiana	1236	1931
Massachusetts	3630	6715
Michigan	1892	4296
Minnesota	1403	3911
Missouri	1660	4049
New Jersey	2784	5102
New York	7334	14665
Ohio	2901	6685
Pennsylvania	4600	12785
Rhode Island	580	1105
Texas	2146	3832
Wisconsin	2167	5176

Source: *The Official Catholic Directory.*

Clearly, the number of priests and nuns varies greatly among these states. This application investigates whether there is any relationship between the number of priests and the number of nuns.

1. Make a scatter plot of the number of nuns on the vertical axis against the number of priests on the horizontal axis.
2. Fit a straight line to the scatter plot.
3. Do you feel that a straight line fits these data well, overall?
4. New York is the state with the largest number of Catholic clergy. Would you say that the two numbers for New York follow the same relationship as do the other states? Give your reasons.
5. Which state has a large number of nuns compared to its number of priests? Which state has a relatively small number of nuns compared to its number of priests?

6. (For students who have studied algebra.) Find the equation of the fitted line. Express it in the form $y = mx + b$, where y is the number of nuns and x is the number of priests. According to this equation, if one state had 100 more priests than a second state, how many more nuns would we expect the first state to have than the second? If there were 100 priests in a state, how many nuns would the equation predict? The moral is: One should be careful using fitted lines for values far to the left or right of the given points.

Page 122: Application 29 (*continued*)

6. Answers may vary slightly. The equation of the line obtained from the points (4,000, 7,600) and (2,000, 3,600) would be $y = 2x - 400$; 200; –200.

Page 123

NOTE TO TEACHERS: Scatter plots that you can duplicate for students to use in answering questions 2 and 9 appear on pages 10 and 11 of this Teacher's Edition.

Application 30

1. a. Republican
   b. hard to tell, but the Democrat won
   c. Democratic

Application 30

Voting for President

The following table gives the percentage of the vote received by the Democratic candidate in the presidential elections of 1920, 1960, and 1964. The percentages were calculated using only votes for the two major party candidates. The question we want to investigate here is whether the 1964 percentage can be predicted from either the 1920 or the 1960 percentage. Only states in the northeast and midwest are included.

Percentage Vote Received by Democrat

State	1920	1960	1964
Colorado	38	45	62
Connecticut	35	54	68
Delaware	43	51	61
Illinois	27	50	59
Indiana	42	45	56
Iowa	26	43	62
Kansas	33	39	55
Maine	30	43	69
Maryland	43	54	66
Massachusetts	29	60	76
Michigan	23	51	67
Minnesota	22	51	64
Nebraska	33	38	53
New Hampshire	40	47	64
New Jersey	30	50	66
New York	29	53	69
North Dakota	19	45	58
Ohio	40	47	63
Pennsylvania	29	51	65
Rhode Island	34	64	81
South Dakota	24	42	56
Vermont	24	41	66
West Virginia	44	53	68
Wisconsin	18	48	62

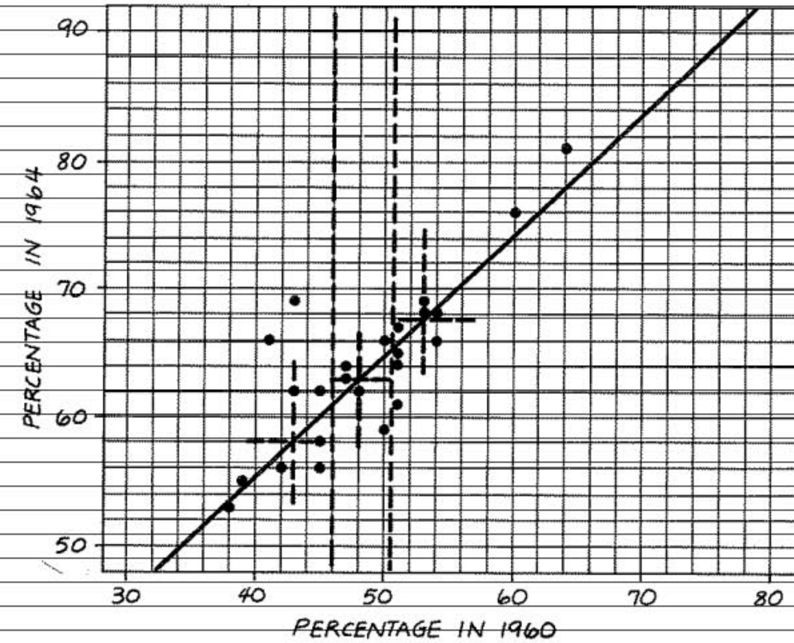
Source: United States Census Bureau.

1. By looking down the columns of percentages, do you think the Democratic or Republican candidate won the election in
   a. *1920?*
   b. 1960?
   c. 1964?

2. Make a scatter plot with the 1960 percentages on the horizontal axis and the 1964 percentages on the vertical axis.
3. Is there a positive, negative, or no association? Why?
4. Fit a straight line to the scatter plot. Due to the fact that three states have a 1960 percentage of 45 and four states have a 1960 percentage of 51, you will have to have 9 states in the left group, 5 in the middle group, and 10 in the right group.
5. Which two states lie the farthest vertical distance from the line?
6. Use your line to complete these sentences.
   a. A state with a 50% vote for the Democratic candidate in 1960 would give the Democratic candidate about a ________% vote in 1964.
   b. A state with a 60% vote for the Democratic candidate in 1960 would give the Democratic candidate about a ________% vote in 1964.
   c. As an approximation using the fitted line, the 1964 vote can be estimated by adding about ________% to the 1960 vote.
7. We call the vertical distance of each point from the fitted line the "error." With the exception of the two states in question 5, all the rest of the states give an "error" less than ________.
8. Putting together the information in questions 5, 6, and 7, we can say the following: The 1964 Democratic percentage equals the 1960 Democratic percentage plus ________%, with an error of less than ________% for all these states except for two, which are ________ and ________.
9. Now make a scatter plot with the 1920 percentages on the horizontal axis and the 1964 percentages on the vertical axis.
10. Is there positive, negative, or no association? Why?
11. Divide the plot into three vertical strips and mark the X in each strip. The three X's do not lie close to a straight line, so do not draw one in.
12. Is it possible to predict the 1964 vote if you are given the 1920 vote?
13. Summarize the information from these two scatter plots in a paragraph.
14. What two candidates ran in
   a. 1920?
   b. 1960?
   c. 1964?

Page 124: Application 30 (*continued*)

2.

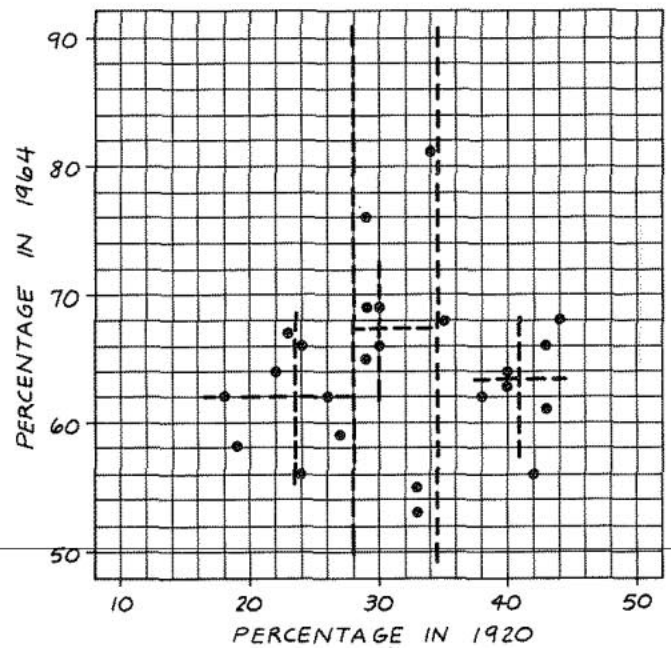


3. Positive; some states tend to vote Democratic and some tend to vote Republican.
4. See the preceding plot.
5. Maine and Vermont
6. a. 64 or 65
   b. 74
   c. 14
7. about 5 percent
8. 14; 5; Maine; Vermont

(Answers for p. 124 continue on the facing page.)

(Answers for p. 124 continued from the facing page.)

9.

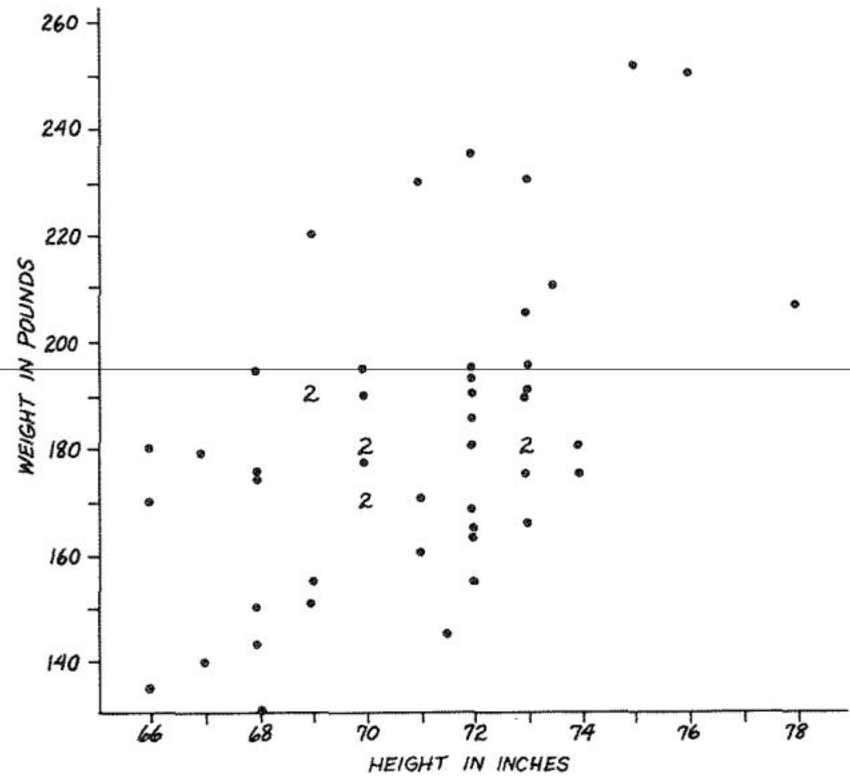


10. No association; apparently states that tended to vote Democratic in 1920 showed no particular tendency in 1964.
11. See the preceding plot.
12. no
13. Answers will vary. Sample: There is a positive association between the percentage of the vote given to the Democratic candidate in 1960 and in 1964 in these twenty-four states. With the exception of Maine and Vermont, the 1964 percentage can be estimated by adding 14 percent to the 1960 percentage. The error resulting is less than about 5 percent. However, it is impossible to predict the 1964 vote from the 1920 vote. States with a relatively high Democratic vote in 1920 did not tend to have a relatively high Democratic vote in 1964.
14. a. Warren Harding (R) and James Cox (D)
    b. John Kennedy (D) and Richard Nixon (R)
    c. Lyndon Johnson (D) and Barry Goldwater (R)

Fitting a Line with a More Complicated Example

When the scatter plot has more points on it than in the previous examples, we can still use the method that was described to fit a straight line. However, some parts of the construction and interpretation can be more complicated, so we will now work a larger example.

The following scatter plot shows the weights and heights of 52 men in an office. Notice that in several places there is a 2 in the plot. This means that two men had the same height and weight.

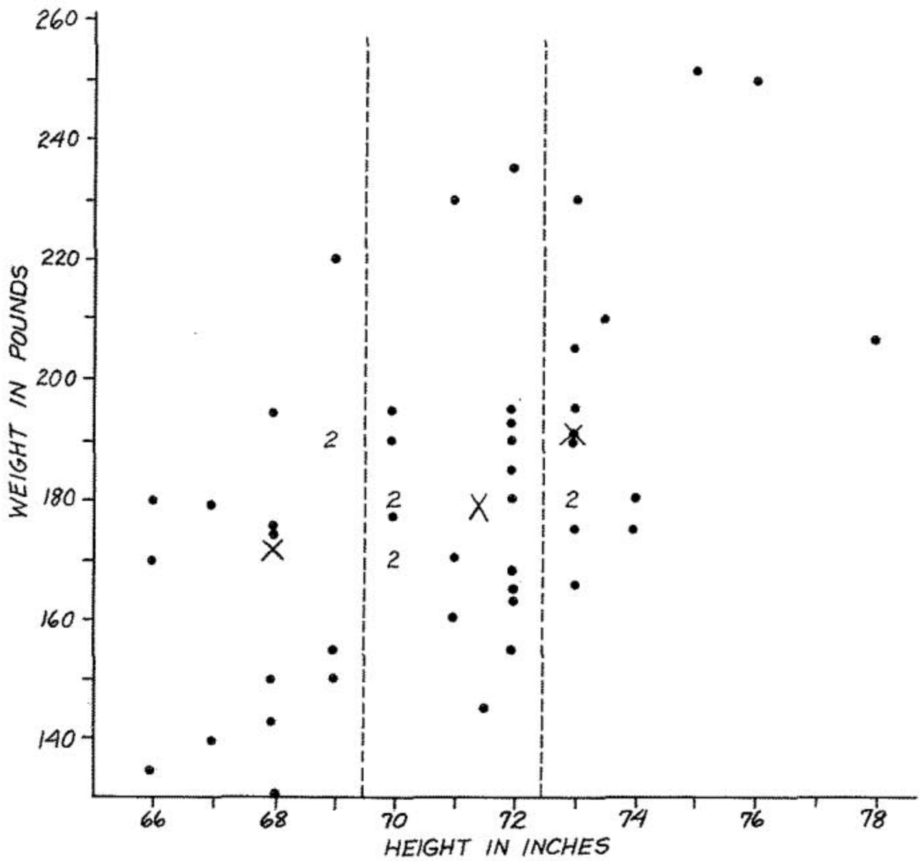


There are 52 points, so to construct the fitted line we would like to divide the points into groups of 17, 18, and 17 points. This division is not possible because different men have the same height. For example, for the left group there are 16 men with heights 69" or less, and 23 men with heights 70" or less. We cannot construct a group with exactly 17 men, so we choose the group with 16 by making the dividing line at 69.5". For the right group, counting in from the right side of the plot shows that 15 men have heights

73" or taller, and 25 men have heights 72" or taller. Similarly, we choose the dividing point to be 72.5", so the right group has 15 points. This choice leaves 21 points in the middle. The dividing lines are shown in the following scatter plot.

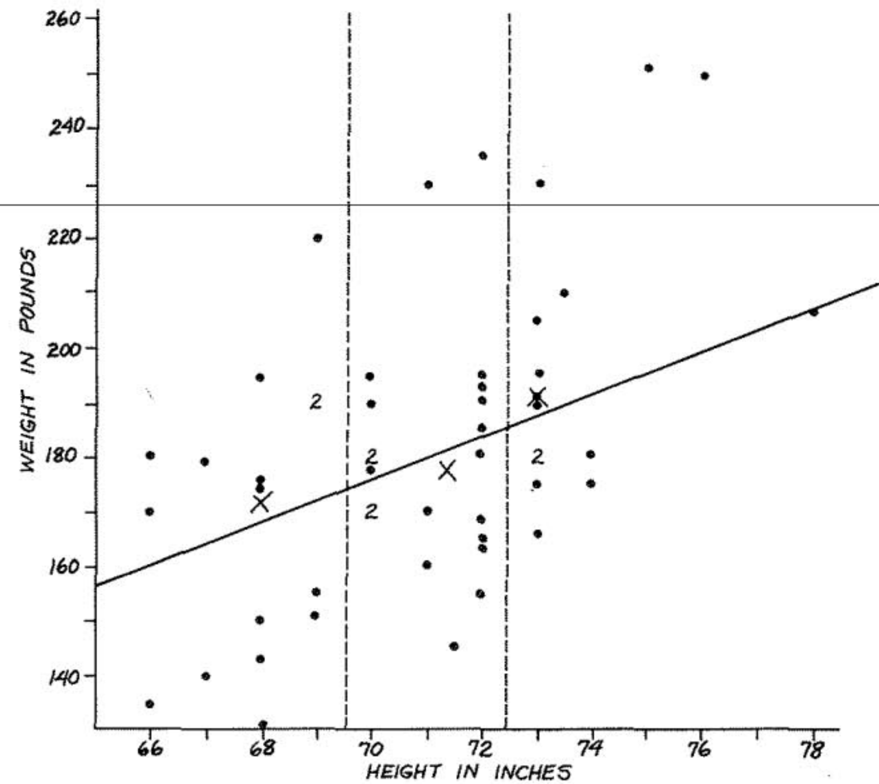
Next, we find the centers of the three groups, using the median method. For the left group of 16 points, both the eighth and ninth largest heights are 68", so the median height is 68". For the weights, the eighth largest is 170 and the ninth is 175, so the median weight is 172.5 pounds. These medians give the left *X* on the scatter plot. For the right group of 15 points, the eighth height is 73' and the eighth weight is 190 pounds. These medians give the right *X* on the plot. Similarly, the center *X* is obtained from the 21 points in the center group as before.

The scatter plot with the three *X*'s follows. It is important to stop now and see if the three *X*'s fall reasonably close to a straight line. If they do not, we would not continue to fit the straight line.



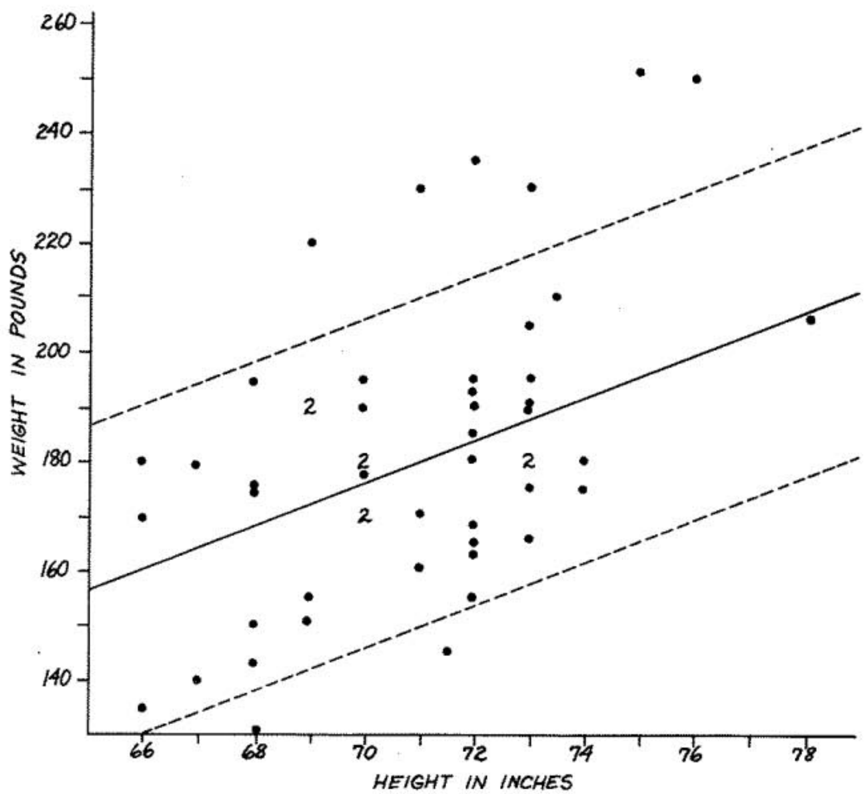
In this case the three X's are close to a straight line, so we continue. Draw the fitted line by first taking a straightedge and placing it along the two end X's. The middle X is below this line. We now slide the straightedge down one-third of the way towards the middle X and draw in the fitted line. This line is shown in the following scatter plot.

The fitted line does not go exactly through any of the three X's, but it goes close to each of them. From this straight line we can predict that a typical weight for a man 66" tall is 160 pounds, and a typical weight for a man 76" tall is 197 pounds. For a 10" increase in height there is a typical increase in weight of 37 pounds, so we could say that on the average for each one inch increase in height there is a 3.7 pound increase in weight. It would be difficult to draw a conclusion like this without fitting a line to the scatter plot.



It is also useful to examine the spread of the points about the fitted line. A good way to do this is to add two additional lines that are parallel to the fitted line. We want these new lines to be an equal distance above and below the fitted line. We also want them drawn far enough from the fitted line so that most, but not all, of the points lie between the two new lines. This lets us notice and focus our attention more easily on outlying points or on other unusual features of the data around the edges.

This has been done in the following plot, using lines giving weights that are 30 pounds more, and 30 pounds less, than the predicted weight for each height. The value 30 pounds was chosen by sliding a ruler parallel to the fitted line so that most, but not all, of the men would fall between these additional lines.



Page 129: Discussion Questions

1. 6; 2
2. 6/52 = 11.5 percent
3. 2/52 = 3.8 percent
4. More men who are very heavy; it is easier to be overweight.
5. unusually heavy
6. a. 7
   b. No; because he is on the line, which means his weight is typical of someone his height.

Discussion Questions

1. How many men fall above the top line? Below the bottom line?
2. What percentage of these 52 men would you say are unusually heavy for their height (above the top line)?
3. What percentage of these 52 men would you say are unusually light for their height?
4. Are there more men who are very heavy for their height, or are there more men who are very light for their height? Why do you think this is the case?
5. For those men whose weight is unusually heavy or unusually light for their height, which group has the more extreme values of weight?
6. Consider the man with height 78".
   a. How many men are heavier than he is?
   b. Do you think he is overweight? Why or why not?

Application 31

1980-84 Rock Hits

The following table lists the top 25 single records from 1980 through 1984 and the number of weeks each of these was in the Top 10 and the Top 40. Is there a close relationship between these two numbers? If we know how many weeks a hit record was in the Top 10, could we accurately predict the total length of time it would remain in the Top 40?

Top Record Hits, 1980-1984

Title — Artist	Number of weeks in Top 10	Top 40
"Physical" — Olivia Newton-John	15	21
"Endless Love" — Diana Ross & Lionel Richie	13	19
"Bette Davis Eyes" — Kim Carnes	14	20
"Every Breath You Take" — Police	13	20
"Billie Jean" — Michael Jackson	11	17
"I Love Rock 'n Roll" — Joan Jett & The Blackhearts	12	16
"Ebony and Ivory" — Paul McCartney & Stevie Wonder	12	15
"Flashdance ... What a Feeling" — Irene Cara	14	20
"Centerfold" — J. Geils Band	12	20
"Lady" — Kenny Rogers	13	19
"Call Me" — Blondie	12	19
"Eye of the Tiger" — Survivor	15	18
"Say Say Say" — Paul McCartney & Michael Jackson	13	18
"(Just Like) Starting Over" — John Lennon	14	19
"When Doves Cry" — Prince	11	16
"Jump" — Van Halen	10	15
"Total Eclipse of the Heart" — Bonnie Tyler	11	18
"Upside Down" — Diana Ross	14	17
"Another Brick in the Wall (part II)" — Pink Floyd	12	19
"Down Under" — Men At Work	10	19
"Rock with You" — Michael Jackson	9	19
"All Night Long (All Night)" — Lionel Richie	13	17
"Maneater" — Daryl Hall & John Oates	13	17
"Magic" — Olivia Newton-John	9	16
"Funkytown" — Lipps, Inc.	9	15

Source: *The Billboard Book of Top 40 Hits*, 1985.

1. Construct a scatter plot, putting weeks in the Top 40 on the vertical axis and weeks in the Top 10 on the horizontal axis.
2. Next divide the data into three groups. There are 25 points, so we would like to have three groups of 8, 9, and 8 points. However, notice that there are many records that are tied with the same Top 10 values.

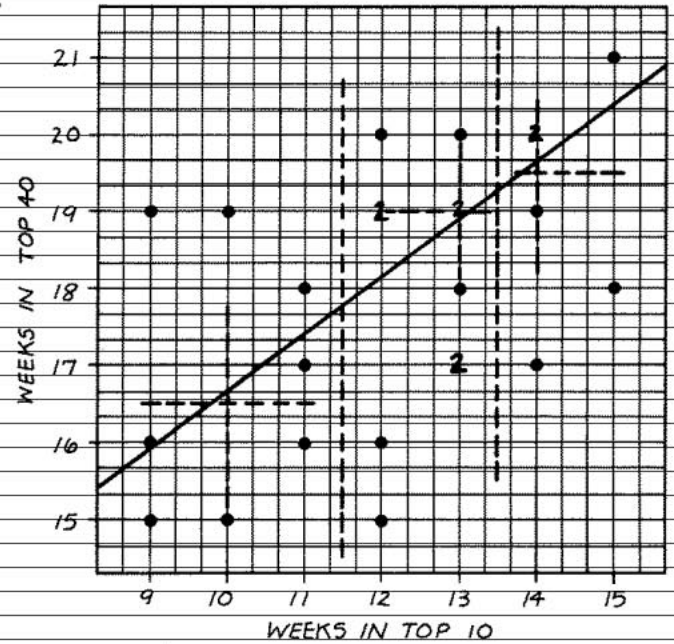
Page 130

NOTE TO TEACHERS: A scatter plot that you can duplicate for students to use in answering question 1 appears on page 12 of this Teacher's Edition.

Either Application 31, "1980–1984 Rock Hits," or Application 32, "52 Men in an Insurance Office," may be omitted.

Application 31

1.



2. 8 in left group and 11 in center group

Page 131: Application 31 (*continued*)

3. See the preceding plot.
4. about 16½ weeks
5. "Rock with You," relatively long; "Ebony and Ivory," relatively short
6. Answers will vary. Sample: For the top 25 single records of 1980 through 1984, in general the longer a record was in the Top 10, the longer the record was in the Top 40. This isn't surprising because more popular records tend to stay around longer. In addition, the number of weeks in the Top 40 includes the number of weeks in the Top 10. For these records, the general pattern was that they were in the Top 40 for a total of about 6 more weeks than they were in the Top 10. Their time in the Top 10 ranged from 9 to 15 weeks. Considering that they were in the Top 10 that long, an extra 6 weeks in positions 11 to 40 does not seem like a long time.

   The most popular record was "Physical" with 15 weeks in the Top 10 and 21 weeks in the Top 40. Two records that do not follow the general pattern are "Rock with You" with only 9 weeks in the Top 10 but 19 weeks in the Top 40, and "Ebony and Ivory" with 12 weeks in the Top 10 but only 15 weeks in the Top 40.

For the right group, if we include records 14 or more weeks in the Top 10, we would have 6 points. If we include records 13 or more weeks in the Top 10, we would have 12 points. In order to have enough points remaining to put into the other two groups, it seems reasonable to make the right group consist of the 6 records with 14 or more weeks in the Top 10. Decide how to form the left and center groups.

3. Using these three groups, fit the line to these data.
4. If a record stayed in the Top 10 for ten weeks, about how long would it stay in the Top 40?
5. Which records are farthest from the line? Did they spend a relatively long or short time in the Top 40 compared to their time in the Top 10? Can you think of any reasons?
6. Write a paragraph that summarizes these data.

Application 32

52 Men in an Insurance Office

The following table lists the heights, shoe sizes, and weights for 52 men in an office. These weights and heights were discussed earlier, on pages 125 to 129. Now we will consider shoe size against height to see if this relationship is similar to or different from the relationship with weight and height.

Height	Shoe Size	Weight	Height	Shoe Size	Weight
70	10.5	195	73	10	190
68	10.5	195	70	9.5	180
69	8.5	152	72	9	168
72	10.5	185	72	10	193
72	10	180	74	12	175
73	9.5	189	71	9	160
74	11	180	72	9.5	163
70	10	180	73	10.5	175
72	9.5	155	72	10	235
73	11	180	71	12	230
68	7	150	69	9.5	220
72	10	195	75	12	252
66	7	135	68	10	175
67	8.5	178	76	13	250
68	9	143	69	9	190
69	10	190	70	10	170
70	9.5	170	73	10	230
73	10.5	205	73	11.5	195
73	10	180	72	9.5	190
67	8.5	140	66	8.5	170
72	10	165	68	8.5	130
70	10	190	73	10.5	166
70	9.5	178	78	13	207
73.5	11	210	66	8.5	180
71.5	10	145	71	9	170
68	9	176	69	9	155

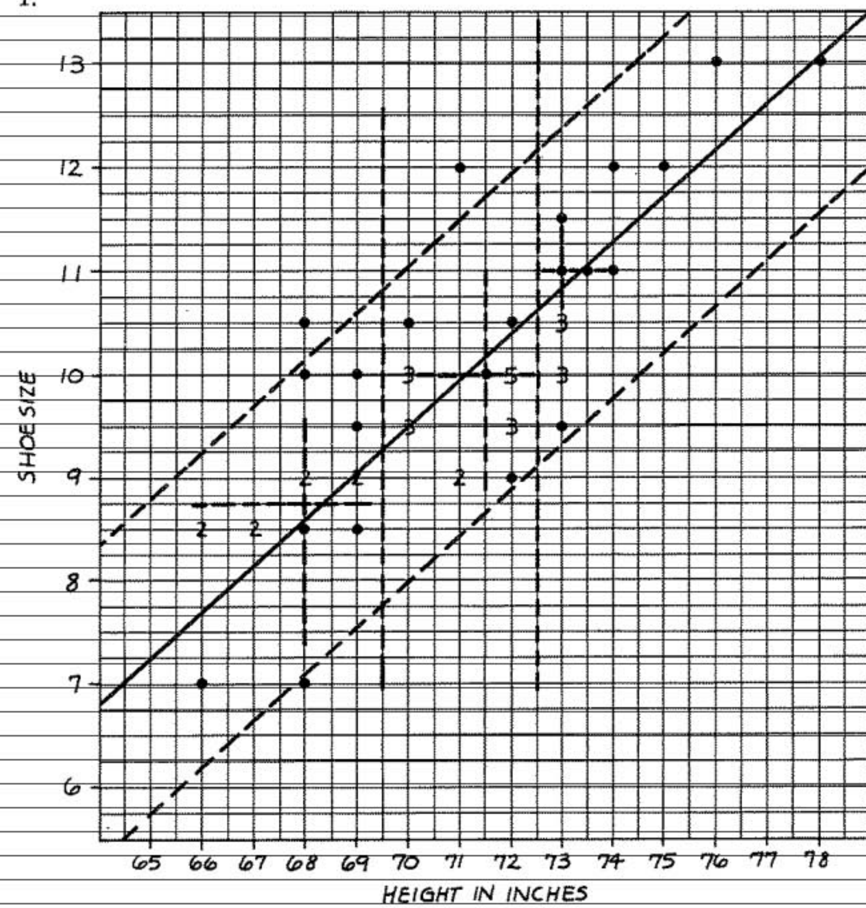
1. Construct a scatter plot of shoe size against height. Put height on the horizontal axis. There are several men with exactly the same height and shoe size. For example, 5 men have the same height of 72" and the same shoe size of 10, so there should be a 5 at that position on the plot. At first, you will want to make the scatter plot lightly with pencil so you can change the dots to numerals as necessary.
2. Use the method that was given to fit a line to these points. (Since there are many repeated heights on the horizontal axis, you will want the three groups to have 16, 21, and 15 points, from left to right.) Does the line fit well?

Page 132

NOTE TO TEACHERS: A scatter plot that you can duplicate for students to use in answering question 1 appears on page 13 of this Teacher's Edition.

Application 32

1.



2. yes

Page 133: Application 32 (*continued*)

3. 7½ or 8; 12; about 2 inches
4. See plot for question 1; yes, three points; two are above the top line and one is below the bottom line.
5. only the three points mentioned in question 4
6. Answers will vary. Sample: Shoe size against height; no; shoe size and height are both measuring the same thing—skeletal length. Further, you cannot control your shoe size or height the way you can control your weight. Recall that for weight against height, we discovered more people at the heavy end than at the light end. Apparently being extra heavy does not cause your feet to get extra long.

3. What shoe size would you predict for a man 66" tall? For a man 76" tall? About how many additional inches of height are needed for a man's predicted shoe size to increase by one whole size?
4. Draw lines 1-1/2 shoe sizes above and 1-1/2 shoe sizes below the fitted line. Are there many points falling outside this range? Are they primarily above the top line or below the bottom line?
5. Are there any outlying points in the plot that do not follow the relationship given by the fitted line?
6. Compare the plot of shoe size against height with the earlier plot of weight against height. Which plot indicates a closer, tighter relationship? Does this surprise you? Can you think of any explanation for this?

Fitted Straight Lines — Clustering and Curvature

In the previous section there were many scatter plots that can be appropriately fitted with straight lines. However, don't assume that it is always appropriate to fit a straight line to a scatter plot. Sometimes the points simply do not lie near a single straight line. Two possibilities are that the data could be *clustered* into two or more groups in the scatter plot or that the data might fall near a *curved* (not straight) line.

How can we tell if there is clustering or curvature, and what should we do about them? Look at the scatter plot as a whole, as you did in Section VI, to see if you observe clusters or a curved relationship. Sometimes clusters or curvature are more obvious after a straight line has been fitted. Always look at a plot again after fitting a line to see if something is apparent that wasn't before.

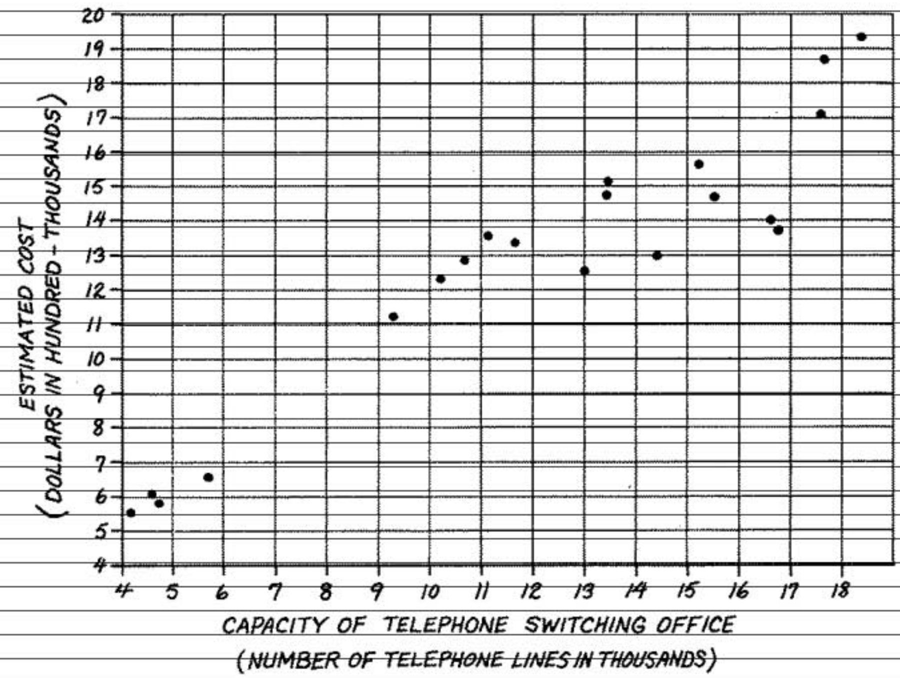
In some cases, a straight line fits well within one of the clusters but not to all the data. Then you can use this line for prediction or summary within the range of data corresponding to the cluster, but don't use a single line that is fitted to all the data. Sometimes you might fit two separate straight lines to different parts of the data. These lines can help you see that a single straight line does not fit well and that a curve might be better. Of course, you might decide instead that no straight or curved line fits well and none should be used for prediction or summary. This could be the best answer.

The following two applications have scatter plots containing clustering and curvature. For these plots it is best not to interpret the data in terms of a single straight line fit.

Application 33

Telephone Office Costs (Clustering)

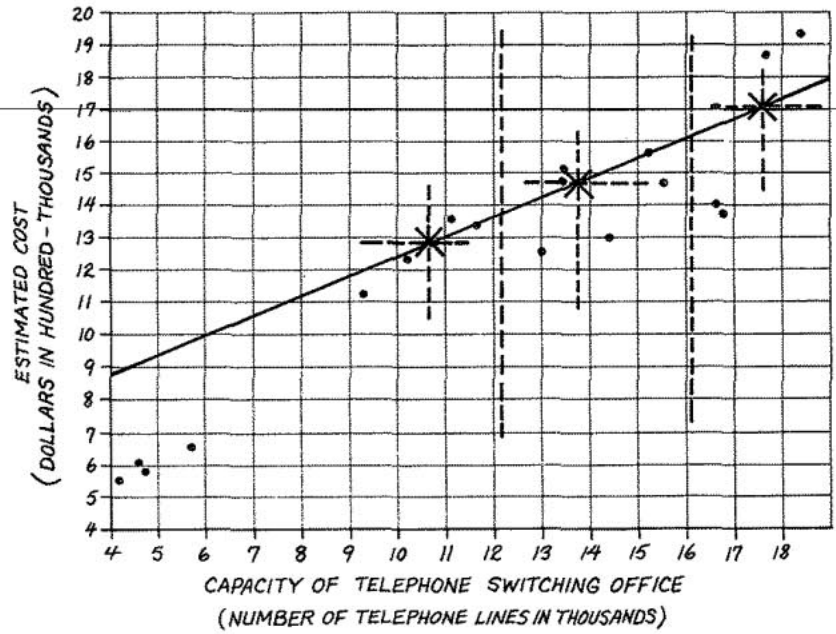
The following scatter plot involves some engineering data. The horizontal axis gives the number of telephone lines that can be handled by each of 20 telephone switching offices. (A telephone switching office is the place that local telephone calls pass through and one customer is connected to another.) The vertical axis gives an estimate of the total cost of constructing the office. The cost depends on more than just the number of telephone lines. Each point in the scatter plot represents one telephone switching office. The horizontal value is the number of telephone lines into the office and the vertical value is the total cost. We want to study the scatter plot to learn whether or not there is a close relationship between cost and capacity for these switching offices.



Page 135

NOTE TO TEACHERS: You can duplicate the scatter plot on page 134 in the student edition for students to use in answering question 2.

Application 33

1. about $600,000
2. 

3. about $1,750,000
4. about $950,000
5. not well at all; from about 9,000 to 18,000 telephone lines

The first general impression is that there is a large gap in the data, giving two separate groups of switching offices. The bottom four offices are all separated by over 3,000 lines from the smallest of the other 16. You might think that the topmost three points should also be treated as a separate cluster. Perhaps they should be, but the gap on the horizontal axis here is definitely smaller, only about 1,000 lines. Thus, as a first step, it seems sensible to treat the data as two clusters rather than one or three.

The data values for the 20 offices are listed in the following table. You will need to construct or trace a scatter plot such as the preceding one to answer the following questions.

Switching Office Capacity (lines)	Estimated Cost	Switching Office Capacity (lines)	Estimated Cost
4,200	$560,000	13,200	$1,470,000
4,600	610,000	13,300	1,510,000
4,700	580,000	14,400	1,300,000
5,700	660,000	15,200	1,580,000
9,300	1,120,000	15,500	1,480,000
10,200	1,230,000	16,700	1,400,000
10,700	1,270,000	16,800	1,370,000
11,100	1,360,000	17,600	1,710,000
11,600	1,340,000	17,700	1,870,000
13,000	1,250,000	18,400	1,930,000

1. For an office with 5,000 telephone lines, what cost would you estimate? Do not fit any straight line. Just scan the plot to get an estimate.
2. Fit a straight line to the cluster of 16 larger offices.
3. For offices of about 18,000 telephone lines, what cost does this line predict?
4. Extend the fitted line to the extreme left of the plot. What would it predict as the cost for an office of size 5,000?
5. How well does the line fit the four observations with small capacity? For what size offices does the fitted line give reasonable estimates of cost?

Application 34

Tree Age and Diameter (Curvature)

The table below lists 27 chestnut oak trees planted on a poor site with their ages and diameters at chest height. We would like to determine how their size increases with age.

Age in Years	Diameter at Chest Height in Inches
4	0.8
5	0.8
8	1.0
8	2.0
8	3.0
10	2.0
10	3.5
12	4.9
13	3.5
14	2.5
16	4.5
18	4.6
20	5.5
22	5.8
23	4.7
25	6.5
28	6.0
29	4.5
30	6.0
30	7.0
33	8.0
34	6.5
35	7.0
38	5.0
38	7.0
40	7.5
42	7.5

Source: Chapman and Demeritt, *Elements of Forest Mensuration*.

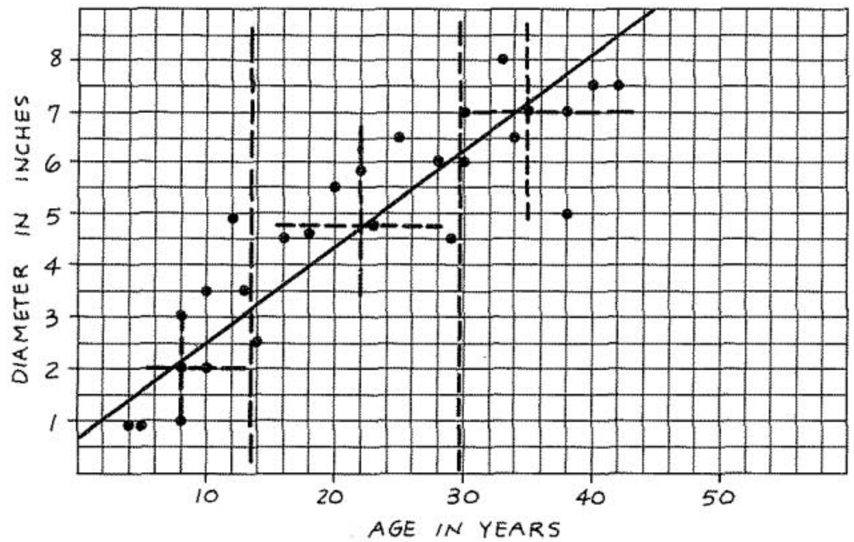
1. Make a scatter plot of these data. We want to predict diameter given age. Which variable will you put on the horizontal axis?
2. Divide the points into three strips. Mark the three X's and draw in the fitted line.
3. Do the three X's lie very close to a single straight line?
4. In the left strip, how many points are
   a. above the line?
   b. below the line?

Page 136

NOTE TO TEACHERS: A scatter plot that you can duplicate for students to use in answering question 1 appears on page 14 of this Teacher's Edition.

Application 34

1. See the following plot; age.



2. See the preceding plot.
3. yes
4. a. 4
   b. 5

Page 137: Application 34 (*continued*)

5. a. 6
   b. 3

6. a. 2
   b. 7

7. young

5. In the center strip, how many points are
   a. above the line?
   b. below the line?
6. In the right strip, how many points are
   a. above the line?
   b. below the line?

There are too many points above the line in the center strip and too many points below the line in both end strips. This means that a single straight line does not fit these data well. A curved line would summarize these data better. There are more complicated statistical methods for fitting a curve to data, but we won't investigate them. You could draw a free-hand curve through the middle of the data.

7. The fact that the points lie on a curved line tells us that trees do not grow at the same rate over their lifetime. Does the diameter increase at a faster rate when the tree is young or old?

Lines on Scatter Plots — Summary

The scatter plot is the basic method for learning about relationships between two variables. Sometimes interpretations are clear simply from studying the scatter plot. This section has dealt with problems where the interpretation becomes clearer by adding a straight line to the plot.

The method of adding the 45° line ($y = x$ line) through the points (0, 0), (1, 1), (2, 2), and so forth and then observing on which side of this line most points lie can assist us in learning whether the variable on the horizontal axis or the variable on the vertical axis is generally larger. This method does not require fitting a line to the data.

In some examples it is helpful to fit a straight line through the central part of the data. We have used a method based on medians. This method is not greatly affected by a few outlying points. If the data follow a straight-line relationship, the method described gives a line that fits the data closely. Moreover, looking at the data in terms of the three X's and the straight line can help us to recognize examples where the data do not fit a single straight line. These situations, such as clustering and curvature, need to be dealt with differently.

The critical feature about the 45° line and the fitted straight line is not just the method of constructing them. As with all the other methods in this book, their purpose is to assist you in the interpretation and analysis of the data. These straight lines can help identify interesting and important data points, find and summarize relationships between the variables, and predict the variable on the vertical axis from the variable on the horizontal axis.

Student Project

1. Take the scatter plots you made on your projects from Section VI and add straight lines when appropriate. Do the lines change any of your interpretations?

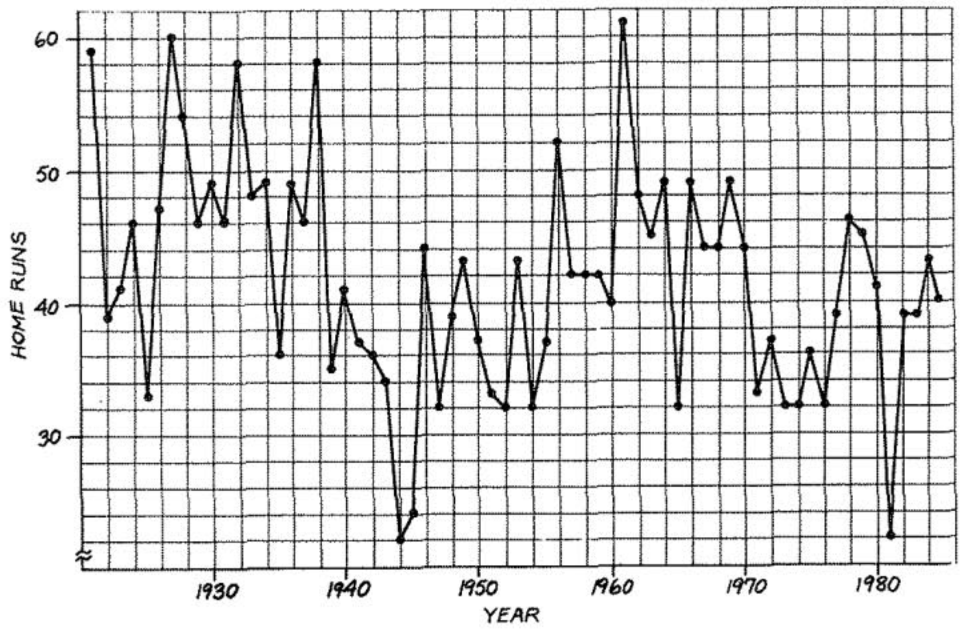
VIII. SMOOTHING PLOTS OVER TIME

The following table lists the American League home run champions from 1921 to 1985.

Year	American League	HR	Year	American League	HR
1921	Babe Ruth, New York	59	1957	Roy Sievers, Washington	42
1922	Ken Williams, St. Louis	39	1958	Mickey Mantle, New York	42
1923	Babe Ruth, New York	41	1959	Rocky Colavito, Cleveland	42
1924	Babe Ruth, New York	46		Harmon Killebrew, Washington	
1925	Bob Meusel, New York	33	1960	Mickey Mantle, New York	40
1926	Babe Ruth, New York	47	1961	Roger Maris, New York	61
1927	Babe Ruth, New York	60	1962	Harmon Killebrew, Minnesota	48
1928	Babe Ruth, New York	54	1963	Harmon Killebrew, Minnesota	45
1929	Babe Ruth, New York	46	1964	Harmon Killebrew, Minnesota	49
1930	Babe Ruth, New York	49	1965	Tony Conigliaro, Boston	32
1931	Babe Ruth, New York	46	1966	Frank Robinson, Baltimore	49
	Lou Gehrig, New York		1967	Carl Yastrzemski, Boston	44
1932	Jimmy Foxx, Philadelphia	58		Harmon Killebrew, Minnesota	
1933	Jimmy Foxx, Philadelphia	48	1968	Frank Howard, Washington	44
1934	Lou Gehrig, New York	49	1969	Harmon Killebrew, Minnesota	49
1935	Jimmy Foxx, Philadelphia	36	1970	Frank Howard, Washington	44
	Hank Greenberg, Detroit		1971	Bill Melton, Chicago	33
1936	Lou Gehrig, New York	49	1972	Dick Allen, Chicago	37
1937	Joe DiMaggio, New York	46	1973	Reggie Jackson, Oakland	32
1938	Hank Greenberg, Detroit	58	1974	Dick Allen, Chicago	32
1939	Jimmy Foxx, Boston	35	1975	George Scott, Milwaukee	36
1940	Hank Greenberg, Detroit	41		Reggie Jackson, Oakland	
1941	Ted Williams, Boston	37	1976	Graig Nettles, New York	32
1942	Ted Williams, Boston	36	1977	Jim Rice, Boston	39
1943	Rudy York, Detroit	34	1978	Jim Rice, Boston	46
1944	Nick Etten, New York	22	1979	Gorman Thomas, Milwaukee	45
1945	Vern Stephens, St. Louis	24	1980	Reggie Jackson, New York	41
1946	Hank Greenberg, Detroit	44		Ben Oglivie, Milwaukee	
1947	Ted Williams, Boston	32	1981	Bobby Grich, California	22
1948	Joe DiMaggio, New York	39		Tony Armas, Oakland	
1949	Ted Williams, Boston	43		Dwight Evans, Boston	
1950	Al Rosen, Cleveland	37		Eddie Murray, Baltimore	
1951	Gus Zernial, Chicago-Philadelphia	33	1982	Gorman Thomas, Milwaukee	39
1952	Larry Doby, Cleveland	32		Reggie Jackson, California	
1953	Al Rosen, Cleveland	43	1983	Jim Rice, Boston	39
1954	Larry Doby, Cleveland	32	1984	Tony Armas, Boston	43
1955	Mickey Mantle, New York	37	1985	Darrell Evans, Detroit	40
1956	Mickey Mantle, New York	52			

Source: *The World Almanac and Book of Facts,* 1985 edition.

From this list it is difficult to see any general trends in the number of home runs through the years. To try to determine the general trends, we will make a scatter plot over time of the number of home runs hit by the champions and connect these points.



This scatter plot looks all jumbled up! It is impossible to see general trends because of the large fluctuations in the number of home runs hit from year to year. For example, 58 home runs were hit in 1938 compared to only 35 the next year. This variation gives the plot a sawtooth effect. The highs and lows, not the overall pattern, capture our attention. To remove the large fluctuations from the data, we will use a method called *smoothing*.

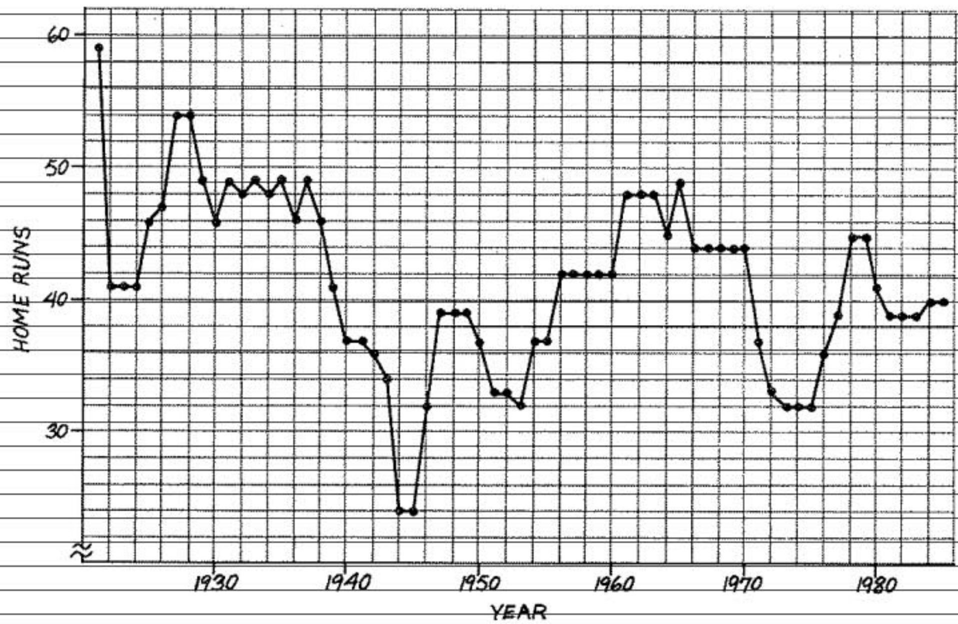
To illustrate, the smoothed version of the first ten years of the home run champions' data follows.

Year	Home Runs	Smoothed Values
1921	59	59
1922	39	41
1923	41	41
1924	46	41
1925	33	46
1926	47	47
1927	60	54
1928	54	54
1929	46	49
1930	49	46
1931	46	

To find the smoothed value for 1924, for example, the 46 home runs for that year are compared to the number of home runs for the year before, 41, and the number of home runs for the following year, 33. The median of the three numbers, 41, is entered into the smoothed values column.

For the first and last years, just copy the original data into the smoothed values column.

The plot of the connected smoothed values follows. Notice what has happened to the large fluctuation between 1938 and 1939. Since this plot is smoother than the previous one, we can see general trends better, such as the drop in the number of home runs in the 1940's.



Discussion Questions

1. Complete the smoothed value column through 1940 for the next ten American League home run champions.

2. Study the smoothed plot of the American League home run champions.

   a. What happened around 1940 that could have affected the number of home runs hit?

   b. Did the increase in the number of games from 154 to 162 in 1961 have an effect on the number of home runs hit?

Page 140

NOTE TO TEACHERS: Students may notice that there are no cases in which the smoothed value is bigger than the original number of home runs for two years in a row. Challenge them to explain why.

It is also true that it is impossible for the smoothed value to be smaller than the original number for two years in a row.

Discussion Questions

1.

Year	Home Runs	Smoothed Values
1931	46	49
1932	58	48
1933	48	49
1934	49	48
1935	36	49
1936	49	46
1937	46	49
1938	58	46
1939	35	41
1940	41	37

2. a. World War II
   b. probably

Page 141: Discussion Questions (*continued*)

3. 1926, no; 1931, no; 1959, no; 1969, no; 1969, no; 1971, possibly, as this might make pitchers less cautious about hitting batters, thereby giving an advantage to pitchers and decreasing home runs.
4. yes; no
5. Use the weighted average of 2/3 of the 1921 value plus 1/3 of the 1922 value.
6. about 40
7. Answers will vary.

3. Study the following rule changes. Do any of them seem to have affected the number of home runs hit by the champions?

   1926 — A ball hit over a fence that is less than 250 feet from home plate will not be counted as a home run.

   1931 — A fair ball that bounces over a fence will be counted as a double instead of a home run.

   1959 — New ballparks must have a minimum distance of 325 feet down the foul lines and 400 feet in center field.

   1969 — The strike zone is decreased in size to include only the area from the armpit to the top of the knee.

   1969 — The pitcher's mound is lowered, giving an advantage to the hitter.

   1971 — All batters must wear helmets.

4. In 1981 there was a strike that shortened the season. Can this be seen in the original data? In the smoothed values?

5. Since they were not smoothed, the endpoints may appear to be out of place. The number of home runs hit in 1921 seems too high. Can you determine a better rule for deciding what to write in the smoothed values column for the endpoints?

6. Imagine a curve through the smoothed values. Try to predict the number of home runs hit in 1986.

7. Some students feel that smoothing is not a legitimate method. For example, they do not like changing the original 33 home runs in 1925 to 46 home runs on the plot of smoothed values. Write a description of the trends that are visible in the smoothed plot that are not easily seen in the original plot. Try to convince a reluctant fellow student that smoothing is valuable. Then study the following answer. Did you mention features we omitted?

The original plot of the time series for home runs gives a very jagged appearance. There were values that were quite large for two years in the 1920's, two years in the 1930's, and also in 1961. Extremely low values occurred in the mid-1940's and in 1981. Using this plot, it is difficult to evaluate overall trends. However, the values in the 1940's and early 1950's seem lower than the values in the late 1920's and 1930's.

We get a stronger impression of trends from the smoothed plot of the home run data. In particular, for the years from 1927 to 1935, the values are generally higher than at any other time before or since. The only period that was nearly comparable was in the early 1960's. The original data show that the champions causing the earlier values to be large were Babe Ruth, Jimmy Foxx, and Lou Gehrig. In the 1960's, it was Roger Maris and Harmon Killebrew. These players clearly were outstanding home run hitters!

There was a steady decline in home runs from the late 1930's to a low period in the middle 1940's. There were also low periods in the early 1950's and in the early 1970's. It is interesting that these lows coincide roughly with World War II, the Korean War, and the Viet Nam War. These wars might be possible causes for the declines, although we have not proved this simply through observing this association. The values for the years since 1980 are near the middle compared to the whole 65-year series. The smoothed series has removed some of the individual highs (such as Maris' 61 in 1961) and lows (such as the 22 in the strike-shortened 1981 season). Therefore, the longer trends stand out more clearly.

Page 143

NOTE TO TEACHERS: In this section, all but one of Application 35, "Birth Months," Application 36, "Olympic Marathon," or Application 37, "Tennis Earnings," may be omitted.

Application 35

1. 296,000
2. July

Application 35

Birth Months

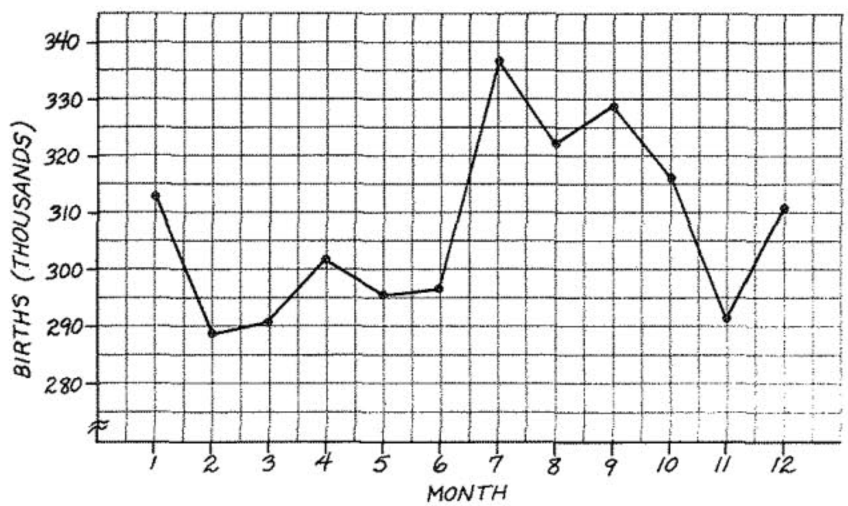
The following table gives the number of babies born in the United States for each month of 1984. The numbers are in thousands.

Month	Births (thousands)	Smoothed Values
January	314	
February	289	
March	291	
April	302	
May	296	
June	297	
July	336	
August	323	
September	329	
October	316	
November	292	
December	311	

Source: National Center for Health Statistics.

1. How many babies were born in May 1984?
2. In which month were the most babies born?

The time series plot for these data is given as follows. This plot is a good candidate for smoothing because of the sawtooth effect. This appearance is an indication that some points are unusually large or small.



3. Copy and complete the "Smoothed Values" column.

4. Make a scatter plot of the smoothed values.

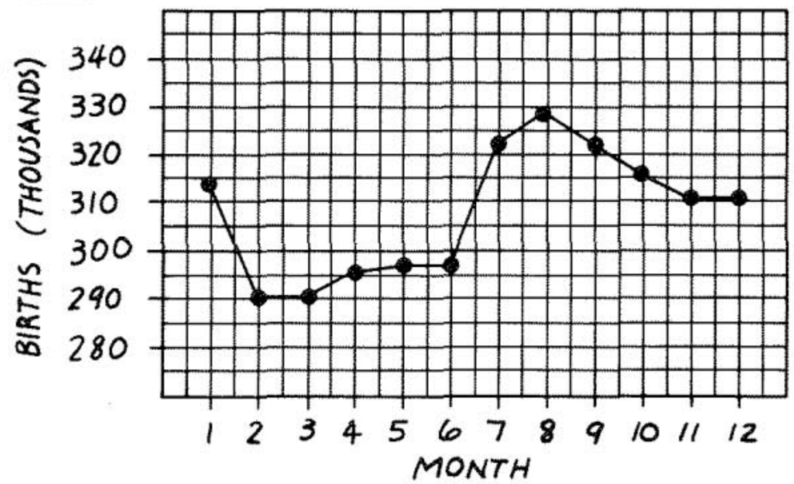
5. What is the general trend in the number of babies born throughout the year?

Page 144: Application 35 (*continued*)

3.

Month	Births (thousands)	Smoothed Values
January	314	314
February	289	291
March	291	291
April	302	296
May	296	297
June	297	297
July	336	323
August	323	329
September	329	323
October	316	316
November	292	311
December	311	311

4.



5. Answers will vary. Sample: The number of births is relatively low in February through June and relatively high in July through January. Further, from February to June, the number of births is fairly constant, in that the smoothed curve changes by only about 2 percent (6/291) over these five months. In July, however, there is a marked increase of 9 percent (26/297) over June. The values for August and September are also high, and then there is a gradual decline until January. The largest drop is from January to February, a decline of about 7 percent (23/314).

It is interesting that the smoothed values at the end of the year are close to the value for January, even though these are at the opposite ends of this 12-month series.

Page 145: Application 36

1. 1952
2. 1916, 1940, 1944; World Wars I and II
3. See the second-to-last column in the following table:

Year	Winner Name, Country	Time Hours	Time Minutes	Time in Minutes	Smoothed Values
1896	Loues, Greece	2	59	179	179
1900	Teato, France	3	0	180	180
1904	Hicks, U.S.A.	3	29	209	180
1908	Hayes, U.S.A.	2	55	175	175
1912	McArthur, South Africa	2	37	157	157
1920	Kolehmainen, Finland	2	33	153	157
1924	Stenroos, Finland	2	41	161	153
1928	El Ouafi, France	2	33	153	153
1932	Zabala, Argentina	2	32	152	152
1936	Son, Japan	2	29	149	152
1948	Cabrera, Argentina	2	35	155	149
1952	Zatopek, Czechoslovakia	2	23	143	145
1956	Mimoun, France	2	25	145	143
1960	Bikila, Ethiopia	2	15	135	135
1964	Bikila, Ethiopia	2	12	132	135
1968	Wolde, Ethiopia	2	20	140	132
1972	Shorter, U.S.A.	2	12	132	132
1976	Cierpinski, East Germany	2	10	130	131
1980	Cierpinski, East Germany	2	11	131	130
1984	Lopes, Portugal	2	9	129	129

Application 36

Olympic Marathon

The following table shows the winning times for the *marathon* run (slightly more than 26 miles) in the 1896-1984 Olympics. The times are rounded to the nearest minute.

Year	Winner Name, Country	Time		Time in Minutes	Smoothed Values
1896	Loues, Greece	2 hours	59 minutes	179	
1900	Teato, France	3	0	180	
1904	Hicks, U.S.A.	3	29	209	
1908	Hayes, U.S.A.	2	55	175	
1912	McArthur, South Africa	2	37	157	
1920	Kolehmainen, Finland	2	33	153	
1924	Stenroos, Finland	2	41	161	
1928	El Ouafi, France	2	33	153	
1932	Zabala, Argentina	2	32	152	
1936	Son, Japan	2	29	149	
1948	Cabrera, Argentina	2	35		
1952	Zatopek, Czechoslovakia	2	23		
1956	Mimoun, France	2	25		
1960	Bikila, Ethiopia	2	15		
1964	Bikila, Ethiopia	2	12		
1968	Wolde, Ethiopia	2	20		
1972	Shorter, U.S.A.	2	12		
1976	Cierpinski, East Germany	2	10		
1980	Cierpinski, East Germany	2	11		
1984	Lopes, Portugal	2	9		

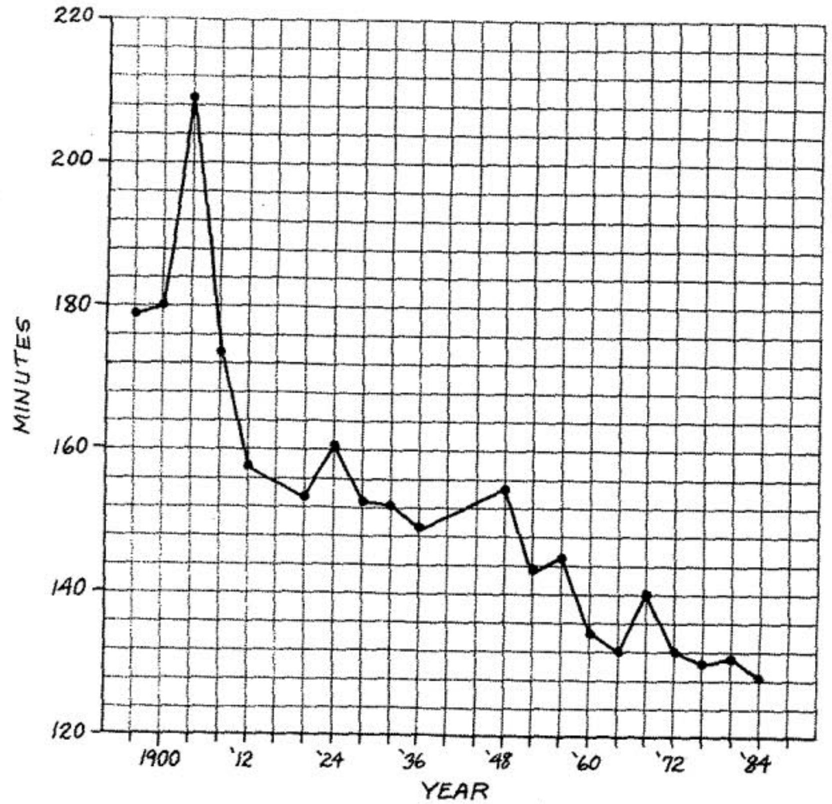
Source: *The World Almanac and Book of Facts*, 1985 edition.

1. The first Olympic women's marathon was not held until 1984. The winner was Joan Benoit of the United States with a time of 2 hours 25 minutes. What was the first year that a Olympic men's marathon winner was able to beat this time?

2. Find the three years when the Olympics were not held. Why were the Olympics not held in these years?

3. Complete the second to the last column of the previous table by converting each time to minutes. The first ten are done for you.

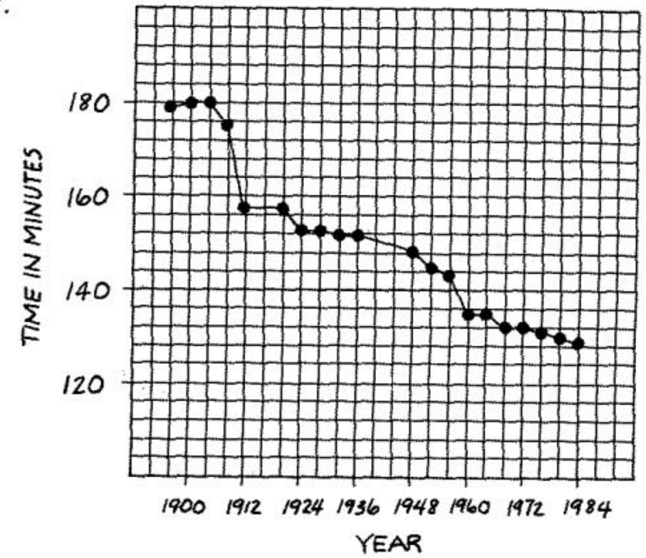
A plot over time with year on the horizontal axis and time in minutes on the vertical axis is shown as follows:



4. What trends do you see in this plot?
5. On the time series plot, which year is farthest from the general trend?
6. Complete the last column of the previous table by smoothing the "time in minutes" column.
7. Construct a plot over time for the smoothed values.
8. Study your plot over time for the smoothed values.
   a. When did the largest drop in time occur?
   b. What do you predict for the winning time in the 1988 Olympic marathon?
   c. Describe the patterns shown on your plot in a short paragraph.

Page 146: Application 36 (*continued*)

4. Times are decreasing at a fairly steady rate.
5. 1904
6. See the last column in the table for question 3.
7.



8. a. 1908–1912
   b. about 128 minutes
   c. Answers will vary. Sample: In the years from 1896 to 1908, the men's marathon was run in about 180 minutes. The time dropped greatly in 1912 to 157 minutes, and since then it has decreased steadily until it was 129 minutes in 1984.

      For the five Olympics held between World War I and World War II (1920 to 1936), there was a decrease of only about 5 minutes in the general trend, but for the first five Olympics after World War II (1948 to 1964), there was a decrease of 14 minutes, almost three times as large. The winning times still seem to be decreasing, but not by as much; for the last 20 years since 1964, the time has dropped only about 6 minutes.

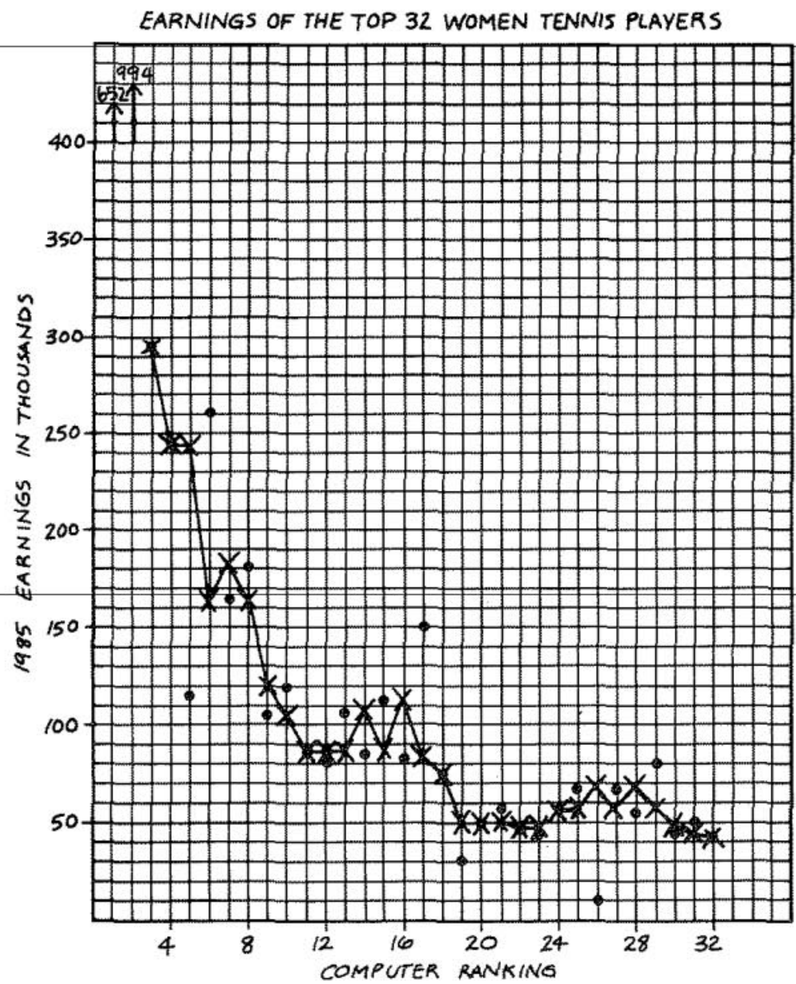
Page 147

NOTE TO TEACHERS: A scatter plot that you can duplicate for students to use in answering question 3 appears on page 15 of this Teacher's Edition.

There are extra data here that are not needed to answer the questions. The extra data are given in case a student wants to use them in a project on top tennis players.

Application 37

1. Plot of women tennis players' earnings.



Application 37

Tennis Earnings

The following two tables from *Tennis Championships Magazine* list the top tennis players of each sex and their earnings from tennis tournaments in the first part of 1985.

The Top 32 Women

Name	Birthplace	Height	Weight	Age	Computer Ranking	1985 Earnings
Chris Evert Lloyd	Ft. Lauderdale, FL	5'6"	118	30	1	$652,269
Martina Navratilova	Czechoslovakia	5'7"	145	28	2	994,579
Hana Mandlikova	Czechoslovakia	5'8"	130	23	3	294,872
Pam Shriver	Baltimore, MD	5'11"	130	23	4	244,653
Manuela Maleeva	Bulgaria	5'6"	114	18	5	115,113
Helena Sukova	Czechoslovakia	6'1"	139	20	6	261,512
Zina Garrison	Houston, TX	5'4"	128	21	7	162,732
Claudia Kohde-Kilsch	West Germany	6'0"	140	21	8	181,995
Wendy Turnbull	Australia	5'4"	120	32	9	104,795
Kathy Rinaldi	Stuart, FL	5'5"	110	18	10	120,315
Bonnie Gadusek	Pittsburgh, PA	5'6"	120	21	11	88,097
Steffi Graf	West Germany	5'5"	110	16	12	81,872
Catarina Lindqvist	Sweden	5'5"	125	22	13	107,805
Gabriela Sabatini	Argentina	5'7"	121	15	14	85,405
Carling Bassett	Canada	5'5"	118	17	15	113,173
Barbara Potter	Waterbury, CT	5'9"	135	23	16	82,949
Kathy Jordan	Bryn Mawr, PA	5'8"	130	25	17	149,763
Bettina Bunge	Switzerland	5'7"	120	22	18	72,090
Sylvia Hanika	West Germany	5'8"	128	25	19	32,310
Andrea Temesvari	Hungary	5'11"	125	19	20	49,810
Alycia Moulton	Sacramento, CA	5'11"	145	24	21	58,735
Peanut Louie	San Francisco, CA	5'5"	115	25	22	48,850
Pam Casale	Camden,NJ	5'8"	127	21	23	43,965
Gigi Fernandez	Puerto Rico	5'7"	140	21	24	56,850
Kathleen Horvath	Chicago, IL	5'7"	115	20	25	68,962
Michelle Torres	Chicago, IL	5'5"	107	18	26	10,950
Elise Burgin	Baltimore, MD	5'4"	115	23	27	68,806
Katerina Maleeva	Bulgaria	5'5"	110	16	28	54,897
Rosalyn Fairbank	South Africa	5'8"	140	24	29	81,301
Catherine Tanvier	France	5'8"	116	20	30	45,660
Virginia Ruzici	Romania	5'8"	128	30	31	49,757
Pascale Paradis	France	5'9"	135	19	32	42,017

Source: *Tennis Championships Magazine.*

The Top 32 Men

Name	Birthplace	Height	Weight	Age	Computer Ranking	1985 Earnings
John McEnroe	West Germany	5'11"	165	26	1	$618,852
Ivan Lendl	Czechoslovakia	6'2"	175	25	2	609,283
Mats Wilander	Sweden	6'1"	175	21	3	416,037
Jimmy Connors	Belleville, IL	5'10"	155	32	4	375,291
Kevin Curren	South Africa	6'1"	170	27	5	193,422
Anders Jarryd	Sweden	5'11"	155	24	6	248,133
Yannick Noah	France	6'4"	180	25	7	202,899
Andres Gomez	Ecuador	6'3"	190	25	8	99,794
Boris Becker	West Germany	6'2"	173	17	9	278,207
Joakim Nystrom	Sweden	6'2"	155	22	10	192,583
Stefan Edberg	Sweden	6'2"	158	19	11	169,920
Eliot Teltscher	Palos Verdes, CA	5'10"	150	26	12	81,092
Miloslav Mecir	Czechoslovakia	6'3"	180	21	13	209,172
Johan Kriek	South Africa	5'8"	155	27	14	151,991
Pat Cash	Australia	5'11"	170	20	15	123,244
Tim Mayotte	Springfield, MA	6'3"	180	25	16	255,174
Scott Davis	Santa Monica, CA	6'2"	170	22	17	126,324
Henrik Sundstrom	Sweden	6'2"	160	21	18	140,122
Tomas Smid	Czechoslovakia	6'3"	175	29	19	220,043
Brad Gilbert	Oakland, CA	6'1"	160	24	20	92,667
Martin Jaite	Argentina	5'11"	150	20	21	104,985
David Pate	Los Angeles, CA	6'0"	170	23	22	84,798
Aaron Krickstein	Ann Arbor, MI	5'10"	150	18	23	110,965
Greg Holmes	Covina, CA	5'10"	160	21	24	56,092
Vitas Gerulaitis	Brooklyn, NY	6'0"	155	31	25	54,329
Libor Pimek	Czechoslovakia	6'5"	172	22	26	61,542
Henri Leconte	France	6'1"	160	22	27	101,690
Jose Luis Clerc	Argentina	6'1"	176	27	28	46,356
Jan Gunnarsson	Sweden	6'0"	165	23	29	81,694
Ben Testerman	Knoxville, TN	6'3"	180	23	30	40,557
Sammy Giammalva	Houston, TX	5'10"	165	22	31	78,873
Jimmy Arias	Buffalo, NY	5'9"	145	21	32	79,941

Source: *Tennis Championships Magazine.*

Do this activity in pairs, with one of you taking the data for men and the other the data for women. After you each answer questions separately for your players, you will put your plots together to compare the women's earnings with the men's. You will need to coordinate with your partner so you both use the same size graph paper.

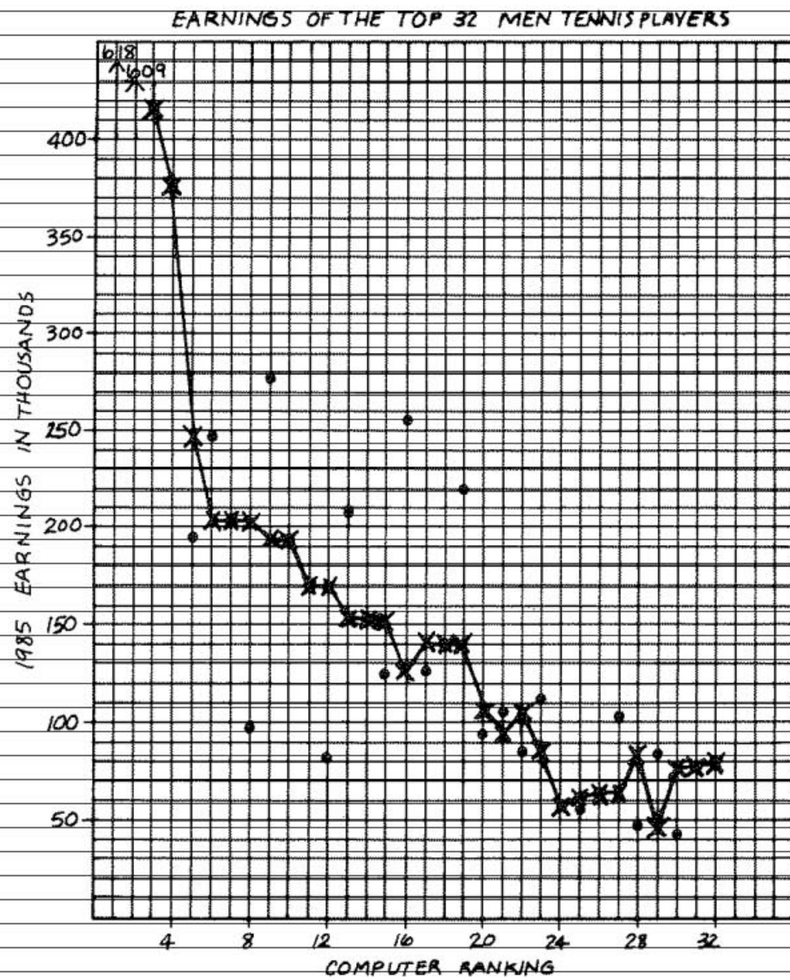
1. Construct a plot over time of the earnings against the computer ranking for your players. Begin by plotting the 32 values as dots; do not connect them with lines. Because the first few men and women earned so much more than the rest, a vertical axis that includes all the earnings would result in most of the earnings being too close together at the bottom. Instead, make the vertical axis from $0 to $400,000. For those

Page 148

NOTE TO TEACHERS: A scatter plot that you can duplicate for students to use in answering question 3 appears on page 16 of this Teacher's Edition.

Application 37 (*continued*)

1. See the plot on the previous page (women players' earnings) and the following plot (men players' earnings).



Page 149: Application 37 (*continued*)

2. See the two preceding plots.
3. See the two preceding plots.
4. Answers will vary. Sample: For women, those with unusually high earnings for their ranking are Sukova (6) and Jordan (17); those with low earnings are M. Maleeva (5) and Torres (26). For men, high earners for their ranking are Becker (9), Mayotte (16), and Smid (19); low earners are Gomez (8) and Teltscher (12).

   Here are some possible reasons why the earnings may not correspond exactly with the rankings. The time period for calculating the computer ranking, possibly a year, could be different from the time period for earnings (the first part of 1985). If a player wins a single lucrative tournament, this could increase the earnings more than the ranking (for example, Becker won Wimbledon in 1985). If a player enters a lot more tournaments than most, his or her earnings could be high. If a player just turned professional, his or her earnings could be low. The ranking probably reflects only singles play but the earnings include both singles and doubles, so a very good doubles player could have high earnings.
5. top ranked players
6. about $250,000
7. men's
8. Answers will vary. Sample: The very highest earnings is for a woman, but for about the top five rankings, men and women typically earn about the same, $200,000 and higher. Similarly, for positions from around 24 to 32, both men and women earned nearly the same, about $50,000.

   For the rankings between 5 to 24, however, the men typically earned more than the women. The men's earnings decreased steadily from rankings 5 to 25 and are, for example, about $150,000 at positions 13 to 15. The women's earnings decreased more quickly, and they are only about $90,000 to $110,000 at these ranks.

NOTE TO TEACHERS: The material on advanced smoothing on pages 149 through 155 is optional and may be omitted.

players who earned more than this, just write in their numbers at the top.

2. In the earlier examples, to get the smoothed earnings we constructed a column of smoothed values and then plotted them. This time we will save a step and do this directly on the plot. For each rank, plot an *X* at the median of the three earnings from that rank, the next lower rank, and the next higher rank. (You might also want to use a different color from the dots for the *X*'s to help distinguish the actual earnings from the smoothed earnings.)
3. Connect the *X*'s by lines. This gives a smooth curve relating the 1985 earnings to the computer rankings.
4. Name any players that have earned a relatively large amount, or a relatively small amount, considering their ranking. Can you think of any reasons for this to happen?
5. The earnings generally decrease as the computer ranking increases. Do the earnings decrease more quickly for the very top ranked players or for the lower ranked players?
6. Give an estimate of how much money you would expect the player who is fifth ranked in 1986 to earn in the corresponding part of 1986.

To answer the remaining questions, work with your partner so you have plots for both men and women.

7. Is smoothing more helpful for the men's data or the women's data to get a useful picture of how earnings relate to rank?
8. Which top tennis players earn more, men or women? To compare the earnings, it helps to place the two plots on top of each other and hold them up to a light. Write a paragraph summarizing how the women's and men's earnings compare.

Advanced Smoothing (Optional)

Often the smoothing method we have just used will give a smooth curve. Sometimes, however, it will still have fluctuations in it that can hide overall trends. In these cases, we will want to smooth the data a little bit more.

For example, in the plot of smoothed values for the American League home run leaders on page 140, the points for the years 1927, 1928, 1944, and 1945 are separated from the general trend. They still give that sawtooth appearance that obscures the overall pattern. A simple method for further smoothing is described in the following paragraphs.

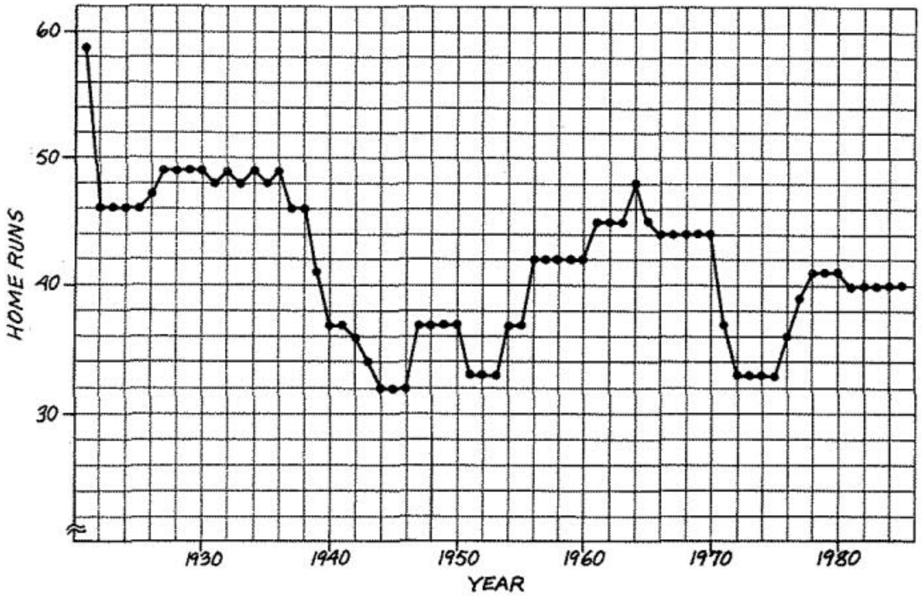
One result of what we did to the first ten years of American League home run data was to make some short strings where adjacent values are equal. For example, the smoothed values for 1922 to 1924 are all 41. One possibility is to treat such "horizontal ties" as single points, and then do the smoothing a second time.

To illustrate, the data for the first ten years, the first smoothed values, and the second smoothed values are listed in the following table.

Year	Home Runs	First Smoothed Values	Second Smoothed Values
1921	59	59	59
1922	39	41	46
1923	41	41	46
1924	46	41	46
1925	33	46	46
1926	47	47	47
1927	60	54	49
1928	54	54	49
1929	46	49	49
1930	49	46	

To find the second smoothed values, we use only the first smoothed values. For the first year, 1921, the value is simply retained. For 1922, we treat the three adjacent 41's as a single value and find the median of 59, 41, and 46, which is 46. For 1923 and 1924, we have the median of 59, 41, and 46 again. For 1925, use the median of 41, 46, and 47. For 1926, use the median of 46, 47, and 54. Use the median of 47, 54, and 49 for 1927 and 1928. For 1929, use the median of 54, 49, and 46.

The plot of the second smoothed values follows. Notice that these smoothed values show the overall trends somewhat more clearly than the earlier smoothed values. Almost all the points that lie far away from the others have been smoothed away. It is now easy to imagine a smooth curve that connects most of the points.



Page 151: Discussion Questions

1.

Year	Home Runs	Smoothed Values	Second Smoothed Values
1930	49	46	49
1931	46	49	48
1932	58	48	49
1933	48	49	48
1934	49	48	49
1935	36	49	48
1936	49	46	49
1937	46	49	46
1938	58	46	46
1939	35	41	41
1940	41	37	

2. late 1920s through early 1930s; Babe Ruth, Jimmy Foxx, and Lou Gehrig
3. 1940s and early 1970s; World War II and Vietnam War
4. Roger Maris in 1961, Babe Ruth in 1927, and Hank Greenberg in 1938 were all high. The champions in 1925, 1965, and 1981 were all much below the trend.

Discussion Questions

1. Complete the next ten values in the second column of smoothed values for the American League home run champions.
2. Which period had the most home runs? Who was responsible for this occurrence?
3. When were the periods of fewest home runs? What was happening during these years?
4. Compare the original home run champions' data to the smoother curve just shown. Which champions differed the most from the value of the overall trend when they played?

This same smoothing process can be repeated to get third smoothed values that are even smoother than the second ones. Using the second smoothed values and the same exact method that was used to calculate the second smoothed values from the first smoothed values, you can calculate the third smoothed values. The effect will be to remove even more of the "bumpiness." For these particular data, the third smoothed values will remove the small peak in 1947-1950 and lower the peak in 1964.

Application 38

National League Home Run Champions (Optional)

The following table lists the National League home run champions.

National League			First
Year		HR	Smoothed Values
1921	George Kelly, New York	23	23
1922	Rogers Hornsby, St. Louis	42	41
1923	Cy Williams, Philadelphia	41	41
1924	Jacques Foumier, Brooklyn	27	39
1925	Rogers Hornsby, St. Louis	39	27
1926	Hack Wilson, Chicago	21	30
1927	Hack Wilson, Chicago Cy Williams, Philadelphia	30	30
1928	Hack Wilson, Chicago Jim Bottomley, St. Louis	31	31
1929	Charles Klein, Philadelphia	43	43
1930	Hack Wilson, Chicago	56	43
1931	Charles Klein, Philadelphia	31	38
1932	Charles Klein, Philadelphia Mel Ott, New York	38	31
1933	Charles Klein, Philadelphia	28	35
1934	Rip Collins, St. Louis Mel Ott, New York	35	34
1935	Walter Berger, Boston	34	34
1936	Mel Ott, New York	33	33
1937	Mel Ott, New York Joe Medwick, St. Louis	31	33
1938	Mel Ott, New York	36	31
1939	John Mize, St. Louis	28	36
1940	John Mize, St. Louis	43	34
1941	Dolph Camilli, Brooklyn	34	34
1942	Mel Ott, New York	30	30
1943	Bill Nicholson, Chicago	29	30
1944	Bill Nicholson, Chicago	33	29
1945	Tommy Holmes, Boston	28	28
1946	Ralph Kiner, Pittsburgh	23	28
1947	Ralph Kiner, Pittsburgh John Mize, New York	51	40
1948	Ralph Kiner, Pittsburgh John Mize, New York	40	51
1949	Ralph Kiner, Pittsburgh	54	47
1950	Ralph Kiner, Pittsburgh	47	47
1951	Ralph Kiner, Pittsburgh	42	42

Source: *The World Almanac and Book of Facts,* 1985 edition.

Page 153: Application 38

1. Hack Wilson in 1930
2. Mike Schmidt and Ralph Kiner for seven seasons each

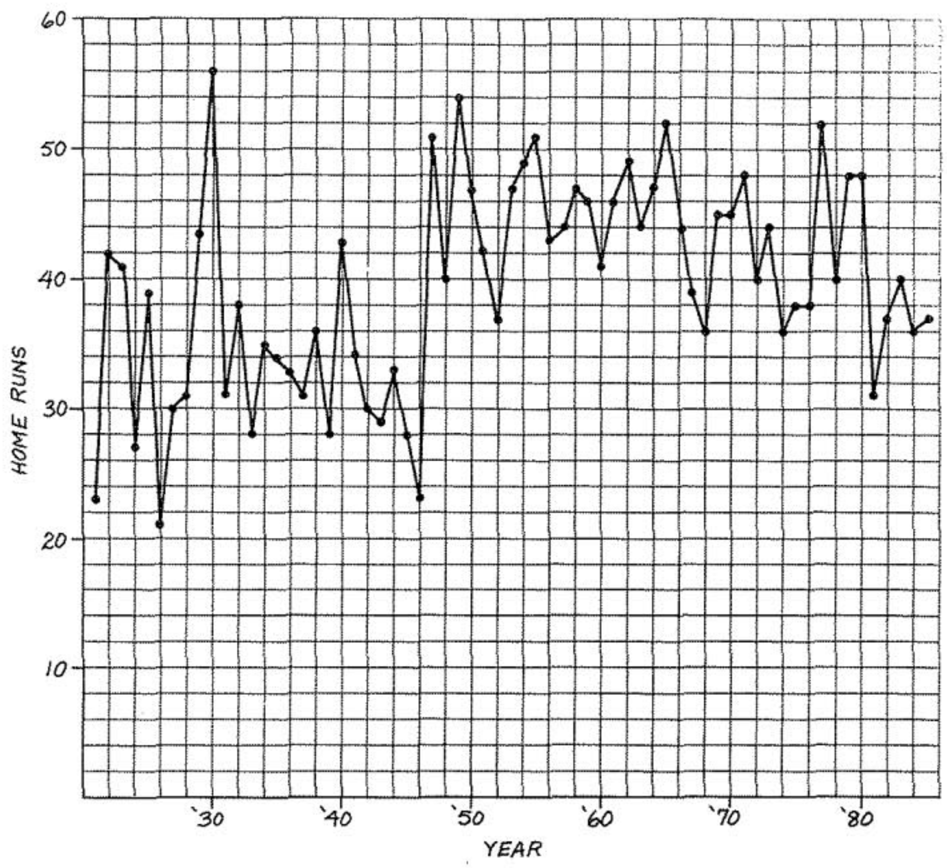
National League

Year		HR	First Smoothed Values
1952	Ralph Kiner, Pittsburgh Hank Sauer, Chicago	37	42
1953	Ed Mathews, Milwaukee	47	47
1954	Ted Kluszewski, Cincinnati	49	49
1955	Willie Mays, New York	51	49
1956	Duke Snider, Brooklyn	43	44
1957	Hank Aaron, Milwaukee	44	44
1958	Ernie Banks, Chicago	47	46
1959	Ed Mathews, Milwaukee	46	46
1960	Ernie Banks, Chicago	41	46
1961	Orlando Cepeda, San Francisco	46	46
1962	Willie Mays, San Francisco	49	46
1963	Hank Aaron, Milwaukee Willie McCovey, San Francisco	44	47
1964	Willie Mays, San Francisco	47	47
1965	Willie Mays, San Francisco	52	47
1966	Hank Aaron, Atlanta	44	44
1967	Hank Aaron, Atlanta	39	39
1968	Willie McCovey, San Francisco	36	39
1969	Willie McCovey, San Francisco	45	45
1970	Johnny Bench, Cincinnati	45	45
1971	Willie Stargell, Pittsburgh	48	45
1972	Johnny Bench, Cincinnati	40	44
1973	Willie Stargell, Pittsburgh	44	40
1974	Mike Schmidt, Philadelphia	36	38
1975	Mike Schmidt, Philadelphia	38	38
1976	Mike Schmidt, Philadelphia	38	38
1977	George Foster, Cincinnati	52	40
1978	George Foster, Cincinnati	40	48
1979	Dave Kingman, Chicago	48	48
1980	Mike Schmidt, Philadelphia	48	48
1981	Mike Schmidt, Philadelphia	31	37
1982	Dave Kingman, New York	37	37
1983	Mike Schmidt, Philadelphia	40	37
1984	Mike Schmidt, Philadelphia Dale Murphy, Atlanta	36	37
1985	Dale Murphy, Atlanta	37	37

Source: *The World Almanac and Book of Facts*, 1985 edition.

1. Which player hit the largest number of home runs in a season?
2. Which player was champion for the most seasons?

A plot over time of the number of home runs follows:



3. From this plot, it is easy to spot unusually high or low years. Which years stand out as the most unusual?

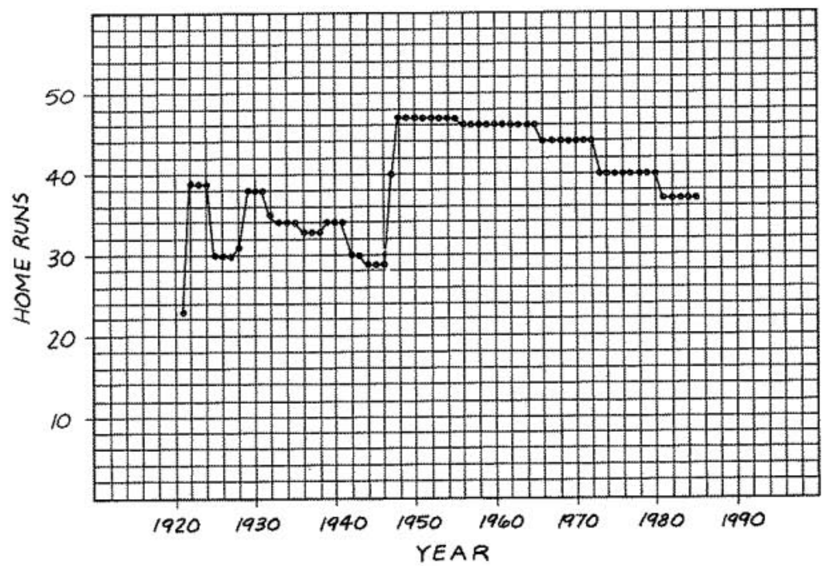
A plot over time of the smoothed values follows:

Page 154: Application 38 (*continued*)

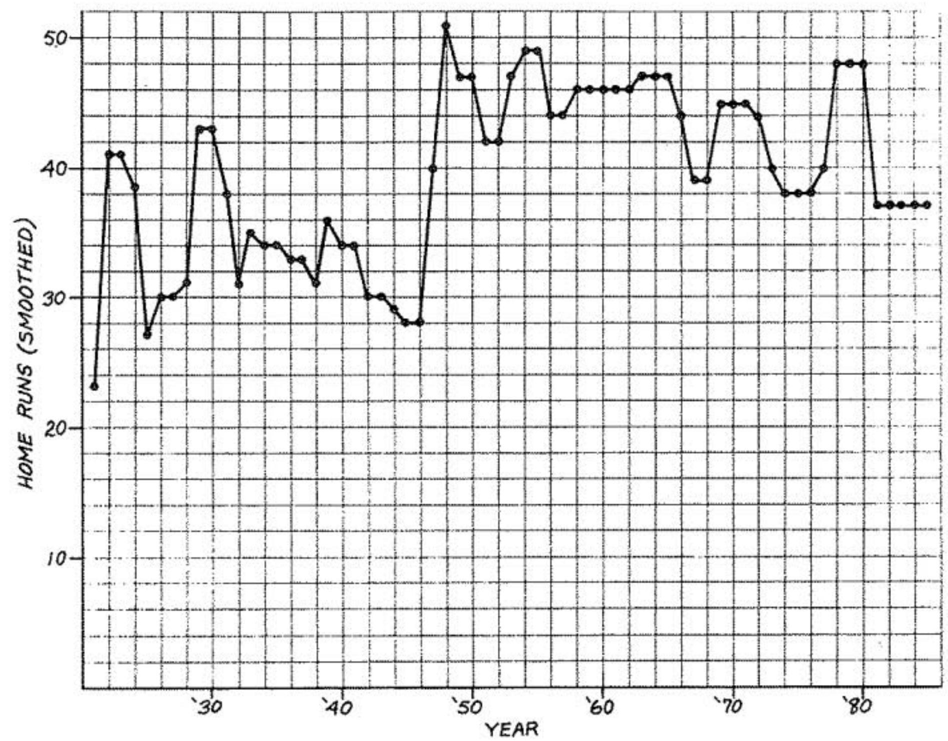
3. 1930 is the main one; maybe also 1940 and 1946.

Page 155: Application 38 (*continued*))

4. yes
5. the big increase in home runs following World War II and gradual decline since then
6. See the tables on p. 203 at the end of this section. The second smoothed values have been added to the tables on pp. 152 and 153.
7.



8. decreasing
9. generally lower
10. No; it's too low; see discussion question 5 on page 141 of the student edition.
11. late 1940s
12. about 38



4. Is there a dip in the early 1940's (during World War II) as there was for the American League?
5. Are there any other especially noticeable trends in this plot?
6. This is an example where a second smoothing might be useful for spotting overall trends. Using the method described just before this application, use the column of first smoothed values and add a column of second smoothed values.
7. Construct a plot over time using the second smoothed values.
8. What has been happening to the number of home runs since 1950?
9. How did the numbers of home runs in the 1920's and 1930's compare to the numbers in the 1960's and 1970's?
10. Do you think that the second smoothed value for 1921 is reasonable? Try to invent a method to smooth endpoints.
11. When did the largest increase in home runs occur?
12. What do you think was the winning number of home runs in 1986?

13. For which year is the actual data value the farthest above the second smoothed value? For which year is the data value the farthest below the second smoothed value?

14. Compare the second smoothed curve for the American League home runs with the second smoothed curve for the National League. What is one way that these curves are similar? What is one way that they are different?

15. Since 1960, are the trends in both leagues about the same?

Smoothing Plots Over Time — Summary

Smoothing is a technique that can be used with time series data where the horizontal axis is marked off in years, days, hours, ages, and so forth. We can use medians to obtain smoothed values, and these smoothed values can remove much of the sawtooth effect often seen in time series data. As a result, a clearer picture of where values are increasing and decreasing emerges.

Many students feel uncomfortable with smoothing. Try to think of it in the same way you think about computing, say, a mean. When you average your test scores in math, the original scores disappear and you are left with one number that summarizes how well you did overall. It is a similar idea with smoothing. Some of the original data disappear and you are left with a summary of overall trends.

Suggestions for Student Projects

1. If any of the scatter plots from your projects in Section VI were plots over time, smooth those plots. Does this show any of the trends more clearly than before?

2. Collect some time series data that interest you and analyze these data according to the methods of this section. Your topic might be one of the following:

   - the number of student absences in your class or school for each day of the last few months
   - daily sales in the school cafeteria during the last few months
   - the daily temperature maximums, minimums, or ranges as reported in the local newspaper
   - sports records for your school

3. A variation of the procedure for smoothing is to replace each value with the median of that value and the *two* values on either side. For example, in the American League home run data, the smoothed value for 1924 would be 41, which is the median of 39, 41, 46, 33, and 47. These are the number of home runs hit in 1922, 1923, 1924, 1925, and 1926. Use this method of "smoothing by medians of five values" on the American League home run data. Discuss the advantages and disadvantages over the usual method.

Page 156: Application 38 (*continued*)

13. 1930; 1924

14. relatively low in the early 1940s; American League higher in 1920s and 1930s

15. Answers will vary. Sample: No; both leagues increased following World War II, but the National League increased quickly to a peak around 1950, while the American League had a much slower increase, peaking in the early 1960s. At that time, both leagues were at about the same level. The American League had a sharp drop in the early 1970s that the National League did not have. There has been a steady decline in the National League since 1950, while the American League has fluctuated much more.

IX. REVIEW OF ALL TECHNIQUES

It might be helpful to reread the review of one variable techniques in Section V before reading this section.

Two Variable Techniques

Suppose that we have measured the cumulative grade point average and the SAT score for each senior in a school. We want to learn how grade point averages and SAT scores are related. This is called a *two variable* situation since we have two values, grade average and test score, for each person.

The basic display for this situation is the scatter plot (Section VI). From a scatter plot you can determine if there is positive, negative, or no association between the variables. You can also determine whether or not the data separate into several clusters of points and whether or not there are any outlying points that do not follow the general pattern. If you notice one of these features, try to find possible reasons for it as part of your interpretation. Often in scatter plots, one of the two variables is time. In these situations we have a plot over time (Section VI).

After constructing and studying a scatter plot, the relationship between the variables may be clear. If so, there is no need to supplement the scatter plot. However, important yet subtle interpretations, concerning both general relationships and specific data points, can often be brought out by adding an appropriate straight line to the scatter plot (Section VII). For plots over time, smoothing can help to show long-run underlying trends, as well as departures of specific points from these trends (Section VIII).

The following applications will help you to see the relative advantages and disadvantages of the statistical methods described in Sections I-VIII. No new techniques are given. These applications will take more time and thought than previous ones as you will have to decide which plot is the best.

There are no right or wrong answers to many of the questions. Your teacher will expect you to make plots that are appropriate and to write thoughtful and complete comments about the characteristics of the data shown in the plot.

Application 39

Presidential Autographs

The following table lists the U.S. presidents. With each is the lowest price you could expect to pay for his autograph (a plain signature).

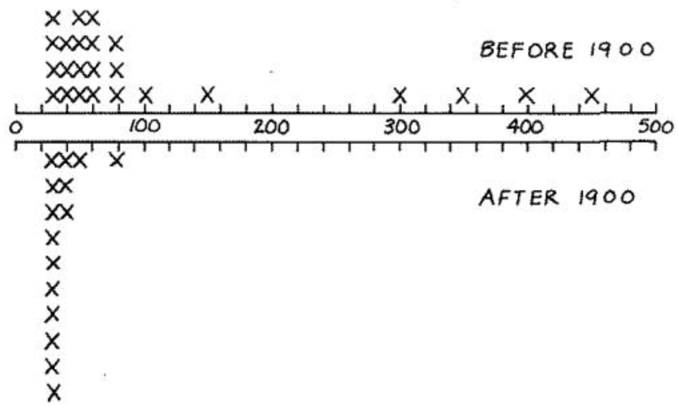
Washington, George	$450	Arthur, Chester A.	$30
Adams, John	300	Cleveland, Grover	26
Jefferson, Thomas	400	Harrison, Benjamin	28
Madison, James	100	McKinley, William	38
Monroe, James	75	Roosevelt, Theodore	32
Adams, John Q.	80	Taft, William H.	28
Jackson, Andrew	150	Wilson, Woodrow	38
Van Buren, Martin	65	Harding, Warren G.	28
Harrison, William H.	80	Coolidge, Calvin	28
Tyler, John	60	Hoover, Herbert	28
Polk, James K.	60	Roosevelt, Franklin	33
Taylor, Zachary	60	Truman, Harry	39
Fillmore, Millard	50	Eisenhower, Dwight D.	28
Pierce, Franklin	50	Kennedy, John F.	80
Buchanan, James	50	Johnson, Lyndon B.	35
Lincoln, Abraham	350	Nixon, Richard M.	50
Johnson, Andrew	50	Ford, Gerald	28
Grant, U. S.	40	Carter, James E.	25
Hayes, Rutherford B.	30	Reagan, Ronald W.	25
Garfield, James	38		

Source: *The Official Price Guide to Paper Collectibles, 1985.*

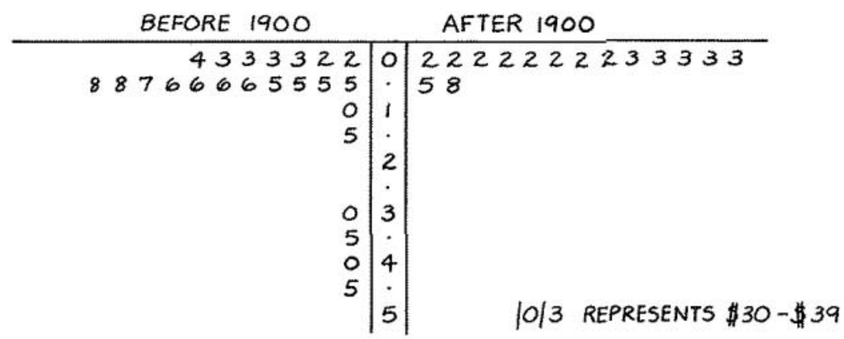
1. Which president's autograph costs the most?
2. Which president's autograph costs the least?
3. Theodore Roosevelt became president in 1901, and all those preceding him in this list were president before 1900. We want to compare the prices of autographs for those who were president before 1900 with the prices for those who were president since 1900. Use any two of the three types of plots — line, stem-and-leaf, or box — to make this comparison.
4. Which plot do you prefer? Why?
5. From this plot, estimate the median prices of autographs of presidents before 1900 and the median prices of autographs of presidents after 1900.
6. Do you think that presidents' autographs become more valuable as they get older? Construct the appropriate plot over time. If it seems to be helpful, make a plot of the smoothed values.
7. Write a summary of the information that you have learned about presidential autographs.

Page 158: Application 39

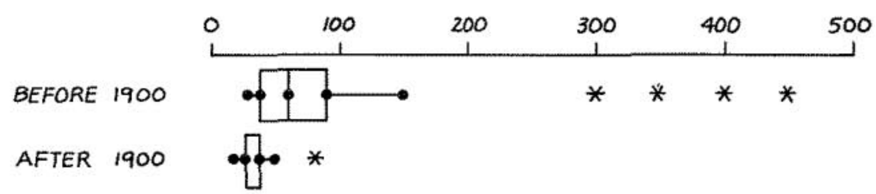
1. George Washington's
2. Jimmy Carter's and Ronald Reagan's
3. Line plot:



Stem-and-leaf plot:



Box plot:



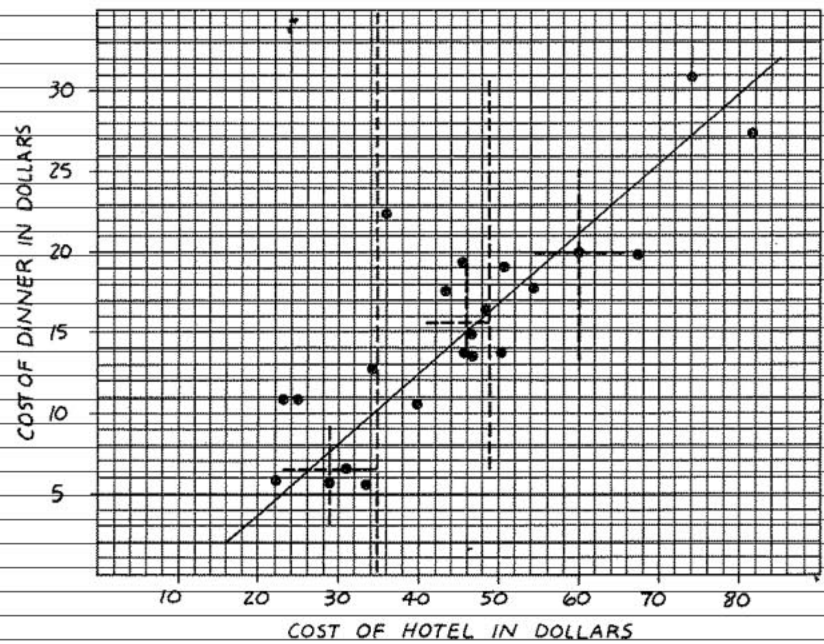
before 1900: lower quartile = 39, median = 60, upper quartile = 90
after 1900: lower quartile = 28, median = 28, upper quartile = 38

(Answers for p. 158 continue on p. 204 at the end of this section.)

Page 159: Application 40

1. Paris; Nairobi

Scatter plot students must construct to answer questions 2-4:



2. Johannesburg
3. about $17
4. Answers will vary. Sample: In general, cities with the most expensive hotels also have the most expensive dinners and cities with cheaper hotels have cheaper dinners. Johannesburg is the only city that is not close to the overall trend. For its $36 hotel price, we would expect a dinner for about $11, but it costs $22.52.

   An equation to predict the dinner price in a city from the hotel price is $y = 0.42x - 4.4$. This means that if a second city has a hotel price $10 higher than in the first city, then we would expect the dinner price to be about $4.20 higher in the second city.

Application 40

Least and Most Expensive Cities

The following table lists some major world cities. With each are the cost in dollars of one night for a single room in a good mid-range hotel and the cost of dinner for one including wine and a tip in a good restaurant.

CITY	HOTEL	DINNER
Athens	$23.73	$10.79
Caracas	24.82	10.95
New Delhi	34.18	12.70
Frankfurt	33.59	5.60
Hong Kong	45.86	19.11
Johannesburg	36.04	22.52
Lisbon	28.90	5.62
London	67.39	19.97
Madrid	30.81	6.56
Manila	81.80	27.27
Mexico City	46.82	13.38
Nairobi	22.22	5.93
New York	60.00	20.00
Paris	74.18	30.91
Rio de Janeiro	46.41	14.97
Rome	43.67	17.47
Stockholm	50.69	19.01
Sydney	54.11	17.75
Tokyo	48.24	16.35
Toronto	50.26	13.78
Vienna	39.77	10.60
Zurich	45.89	13.77

Source: Murray J. Brown, "Hotel and Dining Prices in Cities," *Los Angeles Times*, November 13, 1983.

1. Which city has the most expensive dinner? Which has the least expensive hotel?

To answer the following questions, you will have to decide which type of plot must be constructed and then construct it.

2. In which city is the cost of dinner relatively expensive compared to the cost of a hotel?
3. If the cost of a hotel room in a particular city is $50, what would you expect the cost of a dinner to be?
4. Write a description of the information displayed in your plot.

Who Was the Greatest Yankee Home Run Hitter?

The following table lists four of the greatest New York Yankees' home run hitters with the number of home runs each hit while a Yankee.

Babe Ruth		Lou Gehrig		Mickey Mantle		Roger Maris	
Year	Home Runs	Year	Home Runs	Year	Home Runs	Year	Home Runs
1920	54	1923	1	1951	13	1960	39
1921	59	1924	0	1952	23	1961	61
1922	35	1925	20	1953	21	1962	33
1923	41	1926	16	1954	27	1963	23
1924	46	1927	47	1955	37	1964	26
1925	25	1928	27	1956	52	1965	8
1926	47	1929	35	1957	34	1966	13
1927	60	1930	41	1958	42		
1928	54	1931	46	1959	31		
1929	46	1932	34	1960	40		
1930	49	1933	32	1961	54		
1931	46	1934	49	1962	30		
1932	41	1935	30	1963	15		
1933	34	1936	49	1964	35		
1934	22	1937	37	1965	19		
		1938	29	1966	23		
		1939	0	1967	22		
				1968	18		

Source: *Macmillan Baseball Encyclopedia*, 4th edition.

1. Study these records. Which player appears to be the greatest home run hitter? Why did you choose this player?
2. Your task now is to rank the four players. You may wish to compute means, medians, or quartiles, or make line plots, stem-and-leaf plots, box plots, plots over time, or smoothed plots over time.

   How did you rank the four players? Describe your reasons and include your plots.

Page 160

NOTE TO TEACHERS: This application or the next could be assigned as an end-of-unit project.

Application 41

1. Babe Ruth; answers will vary.
2. Answers will vary.

Page 161: Application 42

Answers will vary.

Application 42

Yankees Versus Mets

New York City has two baseball teams, the Yankees and the Mets. The following table gives the attendance and final standing for both teams each year since the Mets began play in 1962. There are no questions for this application. Your assignment is to make the plots you think are appropriate and interesting. Then write a report about your discoveries.

Here is a possible question to get you started: In a year when attendance for the Yankees is high does Mets attendance also tend to be high?

YANKEES			METS	
Finish	Attendance	Year	Attendance	Finish
Second	2,214,587	1985	2,751,437	Second
Third	1,821,815	1984	1,829,482	Second
Third	2,257,976	1983	1,103,808	Sixth
Fifth	2,041,219	1982	1,320,055	Sixth
First	1,614,533	1981	701,910	Fifth
First	2,627,417	1980	1,178,659	Fifth
Fourth	2,537,765	1979	788,905	Sixth
First	2,335,871	1978	1,007,328	Sixth
First	2,103,092	1977	1,066,825	Sixth
First	2,012,434	1976	1,468,754	Third
Third	1,288,048	1975	1,730,566	Third
Second	1,273,075	1974	1,722,209	Fifth
Fourth	1,262,077	1973	1,912,390	First
Fourth	966,328	1972	2,134,185	Third
Fourth	1,070,771	1971	2,266,680	Third
Second	1,136,879	1970	2,697,479	Third
Fifth	1,067,996	1969	2,175,373	First
Fifth	1,125,124	1968	1,781,657	Ninth
Ninth	1,141,714	1967	1,565,492	Tenth
Tenth	1,124,648	1966	1,932,693	Ninth
Sixth	1,213,552	1965	1,768,389	Tenth
First	1,305,636	1964	1,732,597	Tenth
First	1,308,920	1963	1,080,108	Tenth
First	1,493,574	1962	922,530	Tenth

Source: *Newark Star-Ledger*, April 7, 1986.

ACKNOWLEDGMENTS

Grateful acknowledgment is made to the following publishers, authors, and institutions for permission to use and adapt copyrighted materials.

Addison-Wesley Publishing Company for data on page 111 on World War II submarine sinkings, from Mosteller, Fienberg, and Rourke, *Beginning Statistics with Data Analysis*, © 1983, Addison-Wesley, Reading, Massachusetts. Pg. 79, Table 3-3. Reprinted with permission.

American Public Health Association for data on page 119 on coronary heart disease, from "Cigarette Smoking Related to Geographic Variations in Coronary Heart Disease Mortality and to Expectation of Life in the Two Sexes," Risteard Mulcahy, J.W. McGiluary, and Noel Hickey, in *American Journal of Public Health*, vol. 60, 1970.

Ballantine Books for data on page 158 on the value of presidential autographs, from *The Official Price Guide to Paper Collectibles*, edited by Thomas E. Hudgeons, © 1985, Ballantine Books, New York, NY 10022.

Beverage World for data on page 18 on U.S. soft drink consumption. Reprinted by permission from *Beverage World*, March 1978.

R. R. Bowker Company for data on page 32 on sales of children's books, from *Eighty Years of Best Sellers*, A. P. Hackett and J. H. Burke. Copyright © 1977, R. R. Bowker Company, New York, NY 10017.

Murray J. Brown for data on page 159 on least and most expensive cities, from "Hotel and Dining Prices in Cities" in the *Los Angeles Times*, November 13, 1983. Reprinted by permission of the author.

Peter H. Brown for the "Dumbing for Dollars" chart on page 90, adapted from an article in the *Los Angeles Times*, January 20, 1985. Reprinted by permission of the author.

Consumers Union for data on page 7 from "Nutritional Information for Fast Foods" in *Consumer Reports*; for data and excerpts on pages 36-37 from "How Does Your Allowance Compare to Others," and for data on page 45 from "Motocross Bike Ratings," both in *Penny Power*; also, for data on page 95 from "Walkaround Stereos" in the *1985 Buying Guide Issue*. Copyright 1979, 1983, 1985 by Consumers Union of the United States, Inc., Mount Vernon, NY 10553. Reprinted by permission from *Consumer Reports*, September 1979, *Penny Power*, February/March 1983, and the *1985 Buying Guide Issue*.

Highway Loss Data Institute for data on page 65 on automobile safety records. Copyright Highway Loss Data Institute, Washington, DC 20037.

Joint Center for Political Studies for data on page 108 on the number of black state legislators. Reprinted by permission.

P. J. Kenedy & Sons for data on page 121 on the Catholic clergy, from *The Official Catholic Directory 1985*, published by P. J. Kenedy & Sons, New York, NY 10022.

Los Angeles Times for data on page 3 on top 10 record albums, from "The King of Hearts vs. the Queen of Tarts," Robert Hilburn, copyright 1985, *Los Angeles Times*. Reprinted by permission. For data on page 41 on record album ratings, from "The Pop Meter," copyright 1985, *Los Angeles Times*. Reprinted by permission. For data on pages 69-70, from "Barring L.A. Students from Extracurricular Activities," copyright 1983, *Los Angeles Times*. Reprinted by permission. For data on page 83 on the Celtics-Lakers game, from "The Day in Sports," copyright 1985, *Los Angeles Times*. Reprinted by permission. For data on page 103 on highway speeds, from "Speeding on the Freeways," copyright 1983, *Los Angeles Times*. Reprinted by permission.

A. C. Nielsen Company for data on page 55 on television ratings, from "Prime Time Network Television Rankings, April 29-May 5, 1985." Copyright © A. C. Nielsen Company, New York, NY 10104.

Record Research for data on pages 46, 50-54, and 130 on number 1 hit records, from *The Billboard Book of Top 40 Hits*, Joel Whitburn. Copyright © 1985, Record Research Inc., P.O. Box 200, Menomonee Falls, WI 53051.

Roller Skating Rink Operators Association for data on page 61 on numbers of roller skating clubs. Copyright © 1978, Roller Skating Rink Operators Association, Lincoln, NE 68501.

The Star-Ledger for data on page 161 on the Yankees and the Mets, from April 7, 1985 edition. Copyright © 1985 *The Star-Ledger*, Newark, NJ 07101.

Williams Press for data on page 136 on tree age and diameter, from Chapman and Demeritt, *Elements of Forest Mensuration*, 2nd ed., copyright © 1936, Williams Press, Albany, NY 12204.

H. O. Zimman, Inc. for data on pages 147-48 on the top-ranked women and men tennis players, from *Tennis Championships Magazine*, Special U.S. Open edition, copyright 1985, H. O. Zimman, Inc., Lynn, MA 01901.

Macmillan Publishing Company for data on page 112 on duck plumage and behavior. Reprinted with permission of Macmillan Publishing Company from *Statistics in the Real World: A Book of Examples*, Richard J. Larsen and Donna Fox Stroup. Copyright © 1976 by Macmillan Publishing Company. For data on page 160 on Yankee home run hitters. Reprinted with permission of the publisher from *Macmillan Baseball Encyclopedia*, 4th ed., edited by Joseph L. Reichler. Copyright © 1969, 1974, 1976, 1979 Macmillan Publishing Company.

National Council of Teachers of Mathematics for data on page 79 on letter frequencies, from *Student Math Notes*, copyright © 1983, National Council of Teachers of Mathematics, Reston, VA 22011.

National Soft Drink Association for data on page 98 on U.S. soft drink consumption, from *Sales Survey of the Soft Drink Industry, NSDA 1984*, copyright National Soft Drink Association, Washington, DC 20036.

Newspaper Enterprise Association, Inc. for data on page 1 on 1984 Winter Olympic medal winners; for data on pages 20-21, 138, and 152-53 on National League and American League home run leaders; for data on pages 27 and 29 on heights of buildings in San Francisco and Los Angeles; and for data on page 145 on Olympic marathon times; from *The World Almanac and Book of Facts*, 1985 edition, copyright © Newspaper Enterprise Association, Inc., 1984, New York, NY 10166.

Tables for question 6 on p. 155 of *Exploring Data.* The second smoothed values have been added to the tables on pp. 152 and 153.

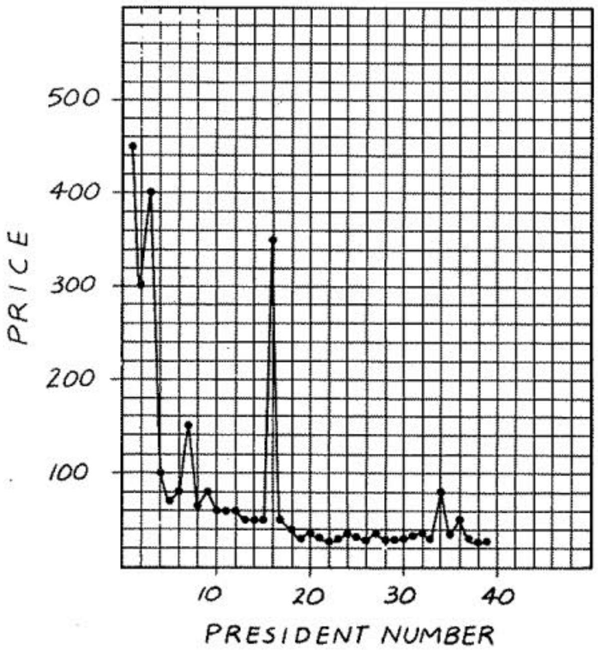
Year	Home Runs	Smoothed Values	Second Smoothed Values
1921	23	23	23
1922	42	41	39
1923	41	41	39
1924	27	39	39
1925	39	27	30
1926	21	30	30
1927	30	30	30
1928	31	31	31
1929	43	43	38
1930	56	43	38
1931	31	38	38
1932	38	31	35
1933	28	35	34
1934	35	34	34
1935	34	34	34
1936	33	33	33
1937	31	33	33
1938	36	31	33
1939	28	36	34
1940	43	34	34
1941	34	34	34
1942	30	30	30
1943	29	30	30
1944	33	29	29
1945	28	28	29
1946	23	28	29
1947	51	40	40
1948	40	51	47
1949	54	47	47
1950	47	47	47
1951	42	42	47
1952	37	42	47
1953	47	47	47

Year	Home Runs	Smoothed Values	Second Smoothed Values
1954	49	49	47
1955	51	49	47
1956	43	44	46
1957	44	44	46
1958	47	46	46
1959	46	46	46
1960	41	46	46
1961	46	46	46
1962	49	46	46
1963	44	47	46
1964	47	47	46
1965	52	47	46
1966	44	44	44
1967	39	39	44
1968	36	39	44
1969	45	45	44
1970	45	45	44
1971	48	45	44
1972	40	44	44
1973	44	40	40
1974	36	38	40
1975	38	38	40
1976	38	38	40
1977	52	40	40
1978	40	48	40
1979	48	48	40
1980	48	48	40
1981	31	37	37
1982	37	37	37
1983	40	37	37
1984	36	37	37
1985	37	37	37

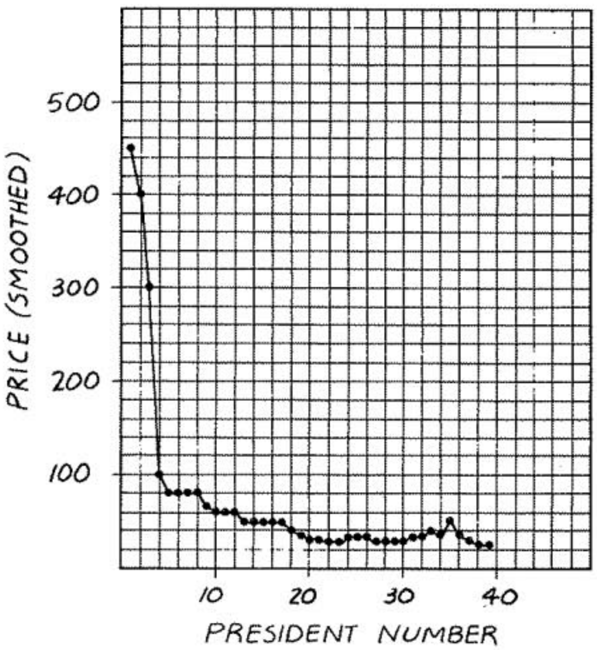
(Answers for p. 158 continued.)

4. Answers will vary.
5. \$60; \$28
6. Yes.

Time series plot for question 6.



Time series plot of smoothed values for question 6.



7. Answers will vary. Sample: The most expensive presidential autograph is George Washington's at \$450 and the least expensive are Jimmy Carter's and Ronald Reagan's at \$25 each. In general, men who were president before 1900 have more expensive autographs than men who were president after 1900. The median price of the earlier presidents is \$60 and that of the later presidents is only \$28.

    Further, about three-fourths of the presidents before 1900 have autographs costing \$40 or more, while only two of the 15 presidents since then have autographs costing that much. The three earliest presidents—Washington, Adams, and Jefferson—plus Lincoln—all have autographs substantially more expensive than the rest. Among the more recent presidents, only Kennedy's autograph stands out as unusually expensive.

    With the exception of Jackson and Lincoln, prices get gradually cheaper for the presidents from Washington up to about 1900. For the presidents since then, there has not been much change in the typical price, with the exception of Kennedy (and possibly also Nixon).