GLOBAL EDITION

Miller & Freund's

Probability and Statistics

for Engineers

NINTH EDITION

Richard A. Johnson

Pearson

MILLER & FREUND'S

PROBABILITY AND STATISTICS FOR ENGINEERS

NINTH EDITION
Global Edition

Richard A. Johnson

University of Wisconsin–Madison

Pearson

Boston Columbus Indianapolis New York San Francisco Amsterdam
Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editorial Director, Mathematics: *Christine Hoag*
Editor-in-Chief: *Deirdre Lynch*
Acquisitions Editor: *Patrick Barbera*
Project Team Lead: *Christina Lepre*
Project Manager: *Lauren Morse*
Editorial Assistant: *Justin Billing*
Acquisitions Editor: *Global Edition: Sourabh Maheshwari*
Program Team Lead: *Karen Wernholm*
Program Manager: *Tatiana Anacki*
Project Editor, Global Edition: *K.K. Neelakantan*
Illustration Design: *Studio Montage*
Cover Design: *Lumina Datamatics*
Program Design Lead: *Beth Paquin*
Marketing Manager: *Tiffany Bitzel*
Marketing Coordinator: *Brooke Smith*
Field Marketing Manager: *Evan St. Cyr*
Senior Author Support/Technology Specialist: *Joe Vetere*
Media Production Manager, Global Edition: *Vikram Kumar*
Senior Procurement Specialist: *Carol Melville*
Senior Manufacturing Controller, Global Editions: *Kay Holman*
Interior Design, Production Management, and Answer Art:
iEnergizer Aptara Limited/Falls Church
Cover Image: © MOLPIX/Shutterstock.com

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsonglobaleditions.com

Authorized adaptation from the United States edition, entitled Miller & Freund's Probability and Statistics for Engineers, 9th Edition, ISBN 978-0-321-98624-5, by Richard A. Johnson published by Pearson Education © 2017.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library
10 9 8 7 6 5 4 3 2 1

Typeset by iEnergizer Aptara Limited

Printed and bound in Malaysia.

ISBN 10: 1-292-17601-6
ISBN 13: 978-1-292-17601-7

Contents

PREFACE

This book introduces probability and statistics to students of engineering and the physical sciences. It is primarily applications focused but it contains optional enrichment material. Each chapter begins with an introductory statement and concludes with a set of statistical guidelines for correctly applying statistical procedures and avoiding common pitfalls. These *Do's and Don'ts* are then followed by a checklist of key terms. Important formulas, theorems, and rules are set out from the text in boxes.

The exposition of the concepts and statistical methods is especially clear. It includes a careful introduction to probability and some basic distributions. It continues by placing emphasis on understanding the meaning of confidence intervals and the logic of testing statistical hypotheses. Confidence intervals are stressed as the major procedure for making inferences. Their properties are carefully described and their interpretation is reviewed in the examples. The steps for hypothesis testing are clearly and consistently delineated in each application. The interpretation and calculation of the *P*-value is reinforced with many examples.

In this ninth edition, we have continued to build on the strengths of the previous editions by adding several more data sets and examples showing application of statistics in scientific investigations. The new data sets, like many of those already in the text, arose in the author's consulting activities or in discussions with scientists and engineers about their statistical problems. Data from some companies have been disguised, but they still retain all of the features necessary to illustrate the statistical methods and the reasoning required to make generalizations from data collected in an experiment.

The time has arrived when software computations have replaced table lookups for percentiles and probabilities as well as performing the calculations for a statistical analysis. Today's widespread availability of statistical software packages makes it imperative that students now become acquainted with at least one of them. We suggest using software for performing some analysis with larger samples and for performing regression analysis. Besides having several existing exercises describing the use of MINITAB, we now give the R commands within many of the examples. This new material augments the basics of the freeware R that are already in Appendix C.

NEW FEATURES OF THE NINTH EDITION INCLUDE:

Large number of new examples. Many new examples are included. Most are based on important current engineering or scientific data. The many contexts further strengthen the orientation towards an applications-based introduction to statistics.

More emphasis on *P*-values. New graphs illustrating *P*-values appear in several examples along with an interpretation.

More details about using R. Throughout the book, R commands are included in a number of examples. This makes it easy for students to check the calculations, on their own laptop or tablet, while reading an example.

Stress on key formulas and downplay of calculation formulas. Generally, computation formulas now appear only at the end of sections where they can easily be skipped. This is accomplished by setting key formulas in the context of an application which only requires all, or mostly all, integer arithmetic. The student can then check their results with their choice of software.

Visual presentation of 2^2 and 2^3 designs. Two-level factorial designs have a 50-year tradition in the teaching of engineering statistics at the University of Wisconsin. It is critical that engineering students become acquainted with the key ideas of (i) systematically varying several input variables at a time and (ii) how to interpret interactions. Major revisions have produced Section 13.3 that is now self-contained. Instructors can cover this material in two or three lectures at the end of course.

New data based exercises. A large number of exercises have been changed to feature real applications. These contexts help both stimulate interest and strengthen a student's appreciation of the role of statistics in engineering applications.

Examples and now numbered. All examples are now numbered within each chapter.

This text has been tested extensively in courses for university students as well as by in-plant training of engineers. The whole book can be covered in a two-semester or three-quarter course consisting of three lectures a week. The book also makes an excellent basis for a one-semester course where the lecturer can choose topics to emphasize theory or application. The author covers most of the first seven chapters, straight-line regression, and the graphic presentation of factorial designs in one semester (see the basic applications syllabus below for the details).

To give students an early preview of statistics, descriptive statistics are covered in Chapter 2. Chapters 3 through 6 provide a brief, though rigorous, introduction to the basics of probability, popular distributions for modeling population variation, and sampling distributions. Chapters 7, 8, and 9 form the core material on the key concepts and elementary methods of statistical inference. Chapters 11, 12, and 13 comprise an introduction to some of the standard, though more advanced, topics of experimental design and regression. Chapter 14 concerns nonparametric tests and goodness-of-fit test. Chapter 15 stresses the key underlying statistical ideas for quality improvement, and Chapter 16 treats the associated ideas of reliability and the fitting of life length models.

The mathematical background expected of the reader is a year course in calculus. Calculus is required mainly for Chapter 5 dealing with basic distribution theory in the continuous case and some sections of Chapter 6.

It is important, in a one-semester course, to make sure engineers and scientists become acquainted with the least squares method, at least in fitting a straight line. A short presentation of two predictor variables is desirable, if there is time. Also, not to be missed, is the exposure to 2-level factorial designs. Section 13.3 now stands alone and can be covered in two or three lectures.

For an audience requiring more exposure to mathematical statistics, or if this is the first of a two-semester course, we suggest a careful development of the properties of expectation (5.10), representations of normal theory distributions (6.5), and then moment generating functions (5.11) and their role in distribution theory (6.6).

For each of the two cases, we suggest a syllabus that the instructor can easily modify according to their own preferences.

One-semester introduction to probability and statistics emphasizing the understanding of basic applications of statistics.	
Chapter 1	especially 1.6
Chapter 2	
Chapter 3	
Chapter 4	4.4–4.7
Chapter 5	5.1–5.4, 5.6, 5.12 5.10 Select examples of joint distribution, independence, mean and variance of linear combinations.
Chapter 6	6.1–6.4
Chapter 7	7.1–7.7
Chapter 8	
Chapter 9	(could skip)
Chapter 10	10.1–10.4
Chapter 11	11.1–11.2 11.3 and 11.4 Examples
Chapter 13	13.3 2^2 and 2^3 designs also 13.1 if possible

A first semester introduction that develops the tools of probability and some statistical inferences.	
Chapter 1	especially 1.6
Chapter 2	
Chapter 3	
Chapter 4	4.4–4.7 4.8 (geometric, negative binomial)
Chapter 5	5.1–5.4, 5.6, 5.12 5.5, 5.7, 5.8 (gamma, beta) 5.10 Develop joint distributions, independence expectation and moments of linear combinations.
Chapter 6	6.1–6.4 6.5–6.7 (Representations, mgf's, transformation)
Chapter 7	7.1–7.7
Chapter 8	
Chapter 9	(could skip)
Chapter 10	10.1–10.4

Any table whose number ends in W can be downloaded from the book's section of the website

http://www.pearsonglobaleditions.com/Johnson

We wish to thank MINITAB (State College, Pennsylvania) for permission to include commands and output from their *MINITAB* software package, the SAS institute (Gary, North Carolina) for permission to include output from their SAS package and the software package R (R project http://CRAN.R-project.org), which we connect to many examples and discuss in Appendix C.

We wish to heartily thank all of those who contributed the data sets that appear in this edition. They have greatly enriched the presentation of statistical methods by setting each of them in the context of an important engineering problem.

The current edition benefited from the input of the reviewers.

Kamran Iqbal, University of Arakansas at Little Rock
Young Bal Moon, Syracuse University
Nabin Sapkota, University of Central Florida
Kiran Bhutani, Catholic University of America
Xianggui Qu, Oakland University
Christopher Chung, University of Houston.

All revisions in this edition were the responsibility of Richard. A. Johnson.

Richard A. Johnson

Pearson would like to thank and acknowledge the following for their contributions to the Global Edition.

Contributors
Vikas Arora

Reviewers
Antar Bandyopadhyay, Indian Statistical Institute
Somesh Kumar, Indian Institute of Technology Kanpur
Abhishek Kumar Umrawal, Delhi University

CHAPTER 1

INTRODUCTION

Everything dealing with the collection, processing, analysis, and interpretation of numerical data belongs to the domain of statistics. In engineering, this includes such diversified tasks as calculating the average length of computer downtimes, collecting and presenting data on the numbers of persons attending seminars on solar energy, evaluating the effectiveness of commercial products, predicting the reliability of a launch vehicle, and studying the vibrations of airplane wings.

In Sections 1.2, 1.3, 1.4, and 1.5 we discuss the recent growth of statistics and its applications to problems of engineering. Statistics plays a major role in the improvement of quality of any product or service. An engineer using the techniques described in this book can become much more effective in all phases of work relating to research, development, or production. In Section 1.6 we begin our introduction to statistical concepts by emphasizing the distinction between a population and a sample.

CHAPTER OUTLINE

1.1 Why Study Statistics?

Answers provided by statistical analysis can provide the basis for making better decisions and choices of actions. For example, city officials might want to know whether the level of lead in the water supply is within safety standards. Because not all of the water can be checked, answers must be based on the partial information from samples of water that are collected for this purpose. As another example, an engineer must determine the strength of supports for generators at a power plant. First, loading a few supports to failure, she obtains their strengths. These values provide a basis for assessing the strength of all the other supports that were not tested.

When information is sought, statistical ideas suggest a typical collection process with four crucial steps.

1. **Set clearly defined goals for the investigation.**
2. **Make a plan of what data to collect and how to collect it.**
3. **Apply appropriate statistical methods to efficiently extract information from the data.**
4. **Interpret the information and draw conclusions.**

These indispensable steps will provide a frame of reference throughout as we develop the key ideas of statistics. Statistical reasoning and methods can help you become efficient at obtaining information and making useful conclusions.

1.2 Modern Statistics

The origin of statistics can be traced to two areas of interest that, on the surface, have little in common: games of chance and what is now called political science. Mid-eighteenth-century studies in probability, motivated largely by interest in games of chance, led to the mathematical treatment of errors of measurement and the theory that now forms the foundation of statistics. In the same century, interest in the numerical description of political units (cities, provinces, countries, etc.) led to what is now called **descriptive statistics**. At first, descriptive statistics consisted merely of the presentation of data in tables and charts; nowadays, it includes the summarization of data by means of numerical descriptions and graphs.

In recent decades, the growth of statistics has made itself felt in almost every major phase of activity. The most important feature of its growth has been the shift in emphasis from descriptive statistics to **statistical inference**. Statistical inference concerns generalizations based on sample data. It applies to such problems as estimating an engine's average emission of pollutants from trial runs, testing a manufacturer's claim on the basis of measurements performed on samples of his product, and predicting the success of a launch vehicle in putting a communications satellite in orbit on the basis of sample data pertaining to the performance of the launch vehicle's components.

When making a statistical inference, namely, an inference that goes beyond the information contained in a set of data, always proceed with caution. One must decide carefully how far to go in generalizing from a given set of data. Careful consideration must be given to determining whether such generalizations are reasonable or justifiable and whether it might be wise to collect more data. Indeed, some of the most important problems of statistical inference concern the appraisal of the risks and the consequences that arise by making generalizations from sample data. This includes an appraisal of the probabilities of making wrong decisions, the chances of making incorrect predictions, and the possibility of obtaining estimates that do not adequately reflect the true situation.

We approach the subject of statistics as a science whenever possible, we develop each statistical idea from its probabilistic foundation, and immediately apply each idea to problems of physical or engineering science as soon as it has been developed. The great majority of the methods we shall use in stating and solving these problems belong to the **frequency** or **classical approach**, where statistical inferences concern fixed but unknown quantities. This approach does not formally take into account the various subjective factors mentioned above. When appropriate, we remind the reader that subjective factors do exist and also indicate what role they might play in making a final decision. This "bread-and-butter" approach to statistics presents the subject in the form in which it has successfully contributed to engineering science, as well as to the natural and social sciences, in the last half of the twentieth century, into the first part of the twenty-first century, and beyond.

1.3 Statistics and Engineering

The impact of the recent growth of statistics has been felt strongly in engineering and industrial management. Indeed, it would be difficult to overestimate the contributions statistics has made to solving production problems, to the effective use of materials and labor, to basic research, and to the development of new products. As in other sciences, statistics has become a vital tool to engineers. It enables them to understand phenomena subject to variation and to effectively predict or control them.

In this text, our attention will be directed largely toward engineering applications, but we shall not hesitate to refer also to other areas to impress upon the reader the great generality of most statistical techniques. The statistical method used to estimate the average coefficient of thermal expansion of a metal serves also to estimate the average time it takes a health care worker to perform a given task, the average thickness of a pelican eggshell, or the average IQ of first-year college students. Similarly, the statistical method used to compare the strength of two alloys serves also to compare the effectiveness of two teaching methods, or the merits of two insect sprays.

1.4 The Role of the Scientist and Engineer in Quality Improvement

During the last 3 decades, the United States has found itself in an increasingly competitive world market. This competition has fostered an international revolution in quality improvement. The teaching and ideas of W. Edwards Deming (1900–1993) were instrumental in the rejuvenation of Japanese industry. He stressed that American industry, in order to survive, must mobilize with a continuing commitment to quality improvement. From design to production, processes need to be continually improved. The engineer and scientist, with their technical knowledge and armed with basic statistical skills in data collection and graphical display, can be main participants in attaining this goal.

Quality improvement is based on the philosophy of "make it right the first time." Furthermore, one should not be content with any process or product but should continue to look for ways of improving it. We will emphasize the key statistical components of any modern quality-improvement program. In Chapter 15, we outline the basic issues of quality improvement and present some of the specialized statistical techniques for studying production processes. The experimental designs discussed in Chapter 13 are also basic to the process of quality improvement.

Closely related to quality-improvement techniques are the statistical techniques that have been developed to meet the **reliability** needs of the highly complex products of space-age technology. Chapter 16 provides an introduction to this area.

1.5 A Case Study: Visually Inspecting Data to Improve Product Quality

This study[1] dramatically illustrates the important advantages gained by appropriately plotting and then monitoring manufacturing data. It concerns a ceramic part used in popular coffee makers. This ceramic part is made by filling the cavity between two dies of a pressing machine with a mixture of clay, water, and oil. After pressing, but before the part is dried to a hardened state, critical dimensions are measured. The depth of the slot is of interest here.

Because of natural uncontrolled variation in the clay-water-oil mixture, the condition of the press, differences in operators, and so on, we cannot expect all of the slot measurements to be exactly the same. Some variation in the depth of slots is inevitable, but the depth needs to be controlled within certain limits for the part to fit when assembled.

[1]Courtesy of Don Ermer

Table 1.1 Slot depth (thousandths of an inch)

Time	6:30	7:00	7:30	8:00	8:30	9:00	9:30	10:00
1	214	218	218	216	217	218	218	219
2	211	217	218	218	220	219	217	219
3	218	219	217	219	221	216	217	218
Sum	643	654	653	653	658	653	652	656
$\bar{x}$	214.3	218.0	217.7	217.7	219.3	217.7	217.3	218.7
Time	**10:30**	**11:00**	**11:30**	**12:30**	**1:00**	**1:30**	**2:00**	**2:30**
1	216	216	218	219	217	219	217	215
2	219	218	219	220	220	219	220	215
3	218	217	220	221	216	220	218	214
Sum	653	651	657	660	653	658	655	644
$\bar{x}$	217.7	217.0	219.0	220.0	217.7	219.3	218.3	214.7

Slot depth was measured on three ceramic parts selected from production every half hour during the first shift from 6 A.M. to 3 P.M. The data in Table 1.1 were obtained on a Friday. The sample mean, or average, for the first sample of 214, 211, and 218 (thousandths of an inch) is

$$\frac{214 + 211 + 218}{3} = \frac{643}{3} = 214.3$$

This value is the first entry in row marked $\bar{x}$.

The graphical procedure, called an ***X*-bar** chart, consists of plotting the sample averages versus time order. This plot will indicate when changes have occurred and actions need to be taken to correct the process.

From a prior statistical study, it was known that the process was stable and that it varied about a value of 217.5 thousandths of an inch. This value will be taken as the central line of the X-bar chart in Figure 1.1.

$$\text{central line: } \bar{\bar{x}} = 217.5$$

It was further established that the process was capable of making mostly good ceramic parts if the average slot dimension for a sample remained between certain control limits.

$$\text{Lower control limit: LCL} = 215.0$$
$$\text{Upper control limit: UCL} = 220.0$$

What does the chart tell us? The mean of 214.3 for the first sample, taken at approximately 6:30 A.M., is outside the lower control limit. Further, a measure of the variation in this sample

$$\text{range} = \text{largest} - \text{smallest} = 218 - 211 = 7$$

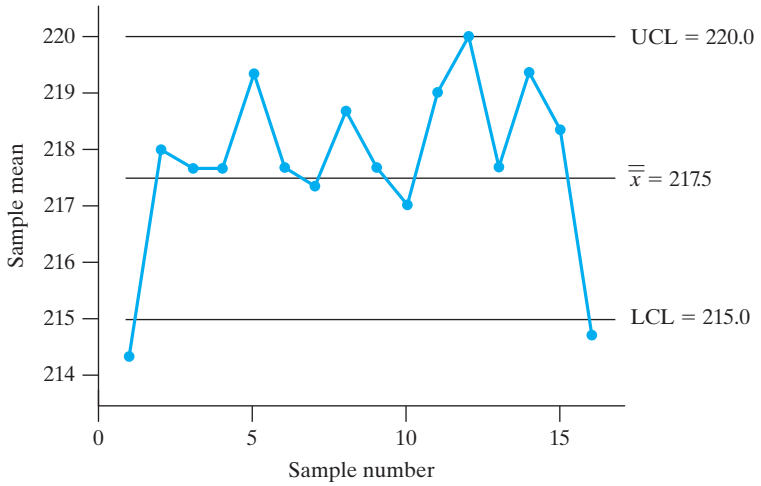


Figure 1.1
X-bar chart for depth

is large compared to the others. This evidence suggests that the pressing machine had not yet reached a steady state. The control chart suggests that it is necessary to warm up the pressing machine before the first shift begins at 6 A.M. Management and engineering implemented an early start-up and thereby improved the process. The operator and foreman did not have the authority to make this change. Deming claims that 85% or more of our quality problems are in the system and that the operator and others responsible for the day-to-day operation are responsible for 15% or less of our quality problems.

The *X*-bar chart further shows that, throughout the day, the process was stable but a little on the high side, although no points were out of control until the last sample of the day. Here an unfortunate oversight occurred. The operator did not report the out-of-control value to either the set-up person or the foreman because it was near the end of her shift and the start of her weekend. She also knew the set-up person was already cleaning up for the end of the shift and that the foreman was likely thinking about going across the street to the Legion Bar for some refreshments as soon as the shift ended. She did not want to ruin anyone's plans, so she kept quiet.

On Monday morning when the operator started up the pressing machine, one of the dies broke. The cost of the die was over a thousand dollars. But this was not the biggest cost. When a customer was called and told there would be a delay in delivering the ceramic parts, he canceled the order. Certainly the loss of a customer is an expensive item. Deming refers to this type of cost as the unknown and unknowable, but at the same time it is probably the most important cost of poor quality.

On Friday the chart had predicted a problem. Afterward it was determined that the most likely difficulty was that the clay had dried and stuck to the die, leading to the break. The chart indicated the problem, but someone had to act. For a statistical charting procedure to be truly effective, action must be taken.

1.6 Two Basic Concepts—Population and Sample

The preceding senarios which illustrate how the evaluation of actual information is essential for acquiring new knowledge, motivate the development of statistical reasoning and tools taught in this text. Most experiments and investigations conducted by engineers in the course of investigating, be it a physical phenomenon, production process, or manufactured unit, share some common characteristics.

A first step in any study is to develop a clear, well-defined **statement of purpose**. For example, a mechanical engineer wants to determine whether a new additive will increase the tensile strength of plastic parts produced on an injection molding machine. Not only must the additive increase the tensile strength, it needs to increase it by enough to be of engineering importance. He therefore created the following statement.

Purpose: Determine whether a particular amount of an additive can be found that will increase the tensile strength of the plastic parts by at least 10 pounds per square inch.

In any statement of purpose, try to avoid words such as *soft*, *hard*, *large enough*, and so on, which are difficult to quantify. The statement of purpose can help us to decide on what data to collect. For example, the mechanical engineer takes two different amounts of additive and produces 25 specimens of the plastic part with each mixture. The tensile strength is obtained for each of 50 specimens.

Relevant data must be collected. But it is often physically impossible or infeasible from a practical standpoint to obtain a complete set of data. When data are obtained from laboratory experiments, no matter how much experimentation is performed, more could always be done. To collect an exhaustive set of data related to the damage sustained by all cars of a particular model under collision at a specified speed, every car of that model coming off the production lines would have to be subjected to a collision!

In most situations, we must work with only partial information. The distinction between the data actually acquired and the vast collection of all potential observations is a key to understanding statistics.

The source of each measurement is called a **unit**. It is usually an object or a person. To emphasize the term *population* for the entire collection of units, we call the entire collection the **population of units**.

Units and population of units

unit: A single entity, usually an object or person, whose characteristics are of interest.
population of units: The complete collection of units about which information is sought.

Guided by the statement of purpose, we have a **characteristic of interest** for each unit in the population. The characteristic, which could be a qualitative trait, is called a **variable** if it can be expressed as a number.

There can be several characteristics of interest for a given population of units. Some examples are given in Table 1.2.

For any population there is the value, for each unit, of a characteristic or variable of interest. For a given variable or characteristic of interest, we call the collection of values, evaluated for every unit in the population, the **statistical population** or just the **population**. This collection of values is the population we will address in all later chapters. Here we refer to the collection of units as the **population of units** when there is a need to differentiate it from the collection of values.

Statistical population

A **statistical population** is the set of all measurements (or record of some quality trait) corresponding to each unit in the entire population of units about which information is sought.

Generally, any statistical approach to learning about the population begins by taking a sample.

Table 1.2 Examples of populations, units, and variables

Population	Unit	Variables/Characteristics
All students currently enrolled in school	student	GPA number of credits hours of work per week major right/left-handed
All printed circuit boards manufactured during a month	board	type of defects number of defects location of defects
All campus fast food restaurants	restaurant	number of employees seating capacity hiring/not hiring
All books in library	book	replacement cost frequency of checkout repairs needed

Samples from a population

A **sample** from a statistical population is the subset of measurements that are actually collected in the course of an investigation.

EXAMPLE 1 **Variable of interest, statistical population, and sample**

Transceivers provide wireless communication between electronic components of consumer products, especially transceivers of Bluetooth standards. Addressing a need for a fast, low-cost test of transceivers, engineers[2] developed a test at the wafer level. In one set of trials with 60 devices selected from different wafer lots, 49 devices passed.

Identify the population unit, variable of interest, statistical population, and sample.

Solution

The population unit is an individual wafer, and the population is all the wafers in lots currently on hand. There is some arbitrariness because we could use a larger population of all wafers that would arrive within some fixed period of time.

The variable of interest is pass or fail for each wafer.

The statistical population is the collection of pass/fail conditions, one for each population unit.

The sample is the collection of 60 pass/fail records, one for each unit in the sample. These can be summarized by their totals, 49 pass and 11 fail. ■

The sample needs both to be representative of the population and to be large enough to contain sufficient information to answer the questions about the population that are crucial to the investigation.

[2]G. Srinivasan, F. Taenzler, and A. Chatterjee, Loopback DFT for low-cost test of single-VCO-based wireless transceivers, *IEEE Design & Test of Computers* 25 (2008), 150–159.

EXAMPLE 2 **Self-selected samples—a bad practice**

A magazine which features the latest computer hardware and software for home-office use asks readers to go to their website and indicate whether or not they owned specific new software packages or hardware products. In past issues, this magazine used similar information to make such statements as "40% of readers have purchased software package *P*." Is this sample representative of the population of magazine readers?

Solution It is clearly impossible to contact all magazine readers since not all are subscribers. One must necessarily settle for taking a sample. Unfortunately, the method used by this magazine's editors is not representative and is badly biased. Readers who regularly upgrade their systems and try most of the new software will be more likely to respond positively indicating their purchases. In contrast, those who did not purchase any of the software or hardware mentioned in the survey will very likely not bother to report their status. That is, the proportion of purchasers of software package *P* in the sample will likely be much higher than it is for the whole population consisting of the *purchase/not purchase* record for each reader. ■

To avoid bias due to self-selected samples, we must take an active role in the selection process.

Using a random number table to select samples

The selection of a sample from a finite population must be done impartially and objectively. But writing the unit names on slips of paper, putting the slips in a box, and drawing them out may not only be cumbersome, but proper mixing may not be possible. However, the selection is easy to carry out using a chance mechanism called a **random number table**.

Random number table

Suppose ten balls numbered 0, 1, . . . , 9 are placed in an urn and shuffled. One is drawn and the digit recorded. It is then replaced, the balls shuffled, another one drawn, and the digit recorded. The digits in Table 7W[3] were actually generated by a computer that closely simulates this procedure. A portion of this table is shown as Table 1.3.

The chance mechanism that generated the random number table ensures that each of the single digits has the same chance of occurrence, that all pairs 00, 01, . . . , 99 have the same chance of occurrence, and so on. Further, any collection of digits is unrelated to any other digit in the table. Because of these properties, the digits are called *random*.

EXAMPLE 3 **Using the table of random digits**

Eighty specialty pumps were manufactured last week. Use Table 1.3 to select a sample of size $n = 5$ to carefully test and recheck for possible defects before they are sent to the purchaser. Select the sample without replacement so that the same pump does not appear twice in the sample.

Solution The first step is to number the pumps from 1 to 80, or to arrange them in some order so they can be identified. The digits must be selected two at a time because the population size $N = 80$ is a two-digit number. We begin by arbitrarily selecting

[3]The W indicates that the table is on the website for this book. See Appendix B for details.

Table 1.3 Random digits (portion of Table 7W)

1306	1189	5731	3968	5606	5084	8947	3897	1636	7810
0422	2431	0649	8085	5053	4722	6598	5044	9040	5121
6597	2022	6168	5060	8656	6733	6364	7649	1871	4328
7965	6541	5645	6243	7658	6903	9911	5740	7824	8520
7695	6937	0406	8894	0441	8135	9797	7285	5905	9539
5160	7851	8464	6789	3938	4197	6511	0407	9239	2232
2961	0551	0539	8288	7478	7565	5581	5771	5442	8761
1428	4183	4312	5445	4854	9157	9158	5218	1464	3634
3666	5642	4539	1561	7849	7520	2547	0756	1206	2033
6543	6799	7454	9052	6689	1946	2574	9386	0304	7945
9975	6080	7423	3175	9377	6951	6519	8287	8994	5532
4866	0956	7545	7723	8085	4948	2228	9583	4415	7065
8239	7068	6694	5168	3117	1568	0237	6160	9585	1133
8722	9191	3386	3443	0434	4586	4150	1224	6204	0937
1330	9120	8785	8382	2929	7089	3109	6742	2468	7025

a row and column. We select row 6 and column 21. Reading the digits in columns 21 and 22, and proceeding downward, we obtain

41 75 91 75 19 69 49

We ignore the number 91 because it is greater than the population size 80. We also ignore any number when it appears a second time, as 75 does here. That is, we continue reading until five different numbers in the appropriate range are selected. Here the five pumps numbered

41 75 19 69 49

will be carefully tested and rechecked for defects.

For situations involving large samples or frequent applications, it is more convenient to use computer software to choose the random numbers. ■

EXAMPLE 4 **Selecting a sample by random digit dialing**

Suppose there is a single three-digit exchange for the area in which you wish to conduct a phone survey. Use the random digit Table 7W to select five phone numbers.

Solution We arbitrarily decide to start on the second page of Table 7W at row 53 and column 13. Reading the digits in columns 13 through 16, and proceeding downward, we obtain

5619 0812 9167 3802 4449

These five numbers, together with the designated exchange, become the phone numbers to be called in the survey. Every phone number, listed or unlisted, has the same chance of being selected. The same holds for every pair, every triplet, and so on. Commercial phones may have to be discarded and another number drawn from the table. If there are two exchanges in the area, separate selections could be done for each exchange. ■

Do's and Don'ts

Do's

1. Create a clear statement of purpose before deciding upon which variables to observe.
2. Carefully define the population of interest.
3. Whenever possible, select samples using a random device or random number table.

Don'ts

1. Don't unquestioningly accept conclusions based on self-selected samples.

Review Exercises

1.1 An article in a civil engineering magazine asks "How Strong Are the Pillars of Our Overhead Bridges?" and goes on to say that samples were collected of materials being used in the construction of 294 overhead bridges across the country. Let the variable of interest be a numerical measure of quality. Identify the population and the sample.

1.2 A television channel announced a vote for their viewers' favorite television show. Viewers were asked to visit the channel's website and vote online for their favorite show. Identify the population in terms of preferences, and the sample. Is the sample likely to be representative? Comment. Also describe how to obtain a sample that is likely to be more representative.

1.3 Consider the population of all cars owned by women in your neighborhood. You want to know the model of the car.

(a) Specify the population unit.

(b) Specify the variable of interest.

(c) Specify the statistical population.

1.4 Identify the statistical population, sample, and variable of interest in each of the following situations:

(a) Tensile strength is measured on 20 specimens of super strength thread made of the same nanofibers. The intent is to learn about the strengths for all specimens that could conceivably be made by the same method.

(b) Fifteen calls to the computer help desk are selected from the hundreds received one day. Only 4 of these calls ended without a satisfactory resolution of the problem.

(c) Thirty flash memory cards are selected from the thousands manufactured one day. Tests reveal that 6 cards do not meet manufacturing specifications.

1.5 For ceiling fans to rotate effectively, the bending angle of the individual paddles of the fan must remain between tight limits. From each hour's production, 25 fans are selected and the angle is measured.

Identify the population unit, variable of interest, statistical population, and sample.

1.6 Ten seniors have applied to be on the team that will build a high-mileage car to compete against teams from other universities. Use Table 7 of random digits to select 5 of the 10 seniors to form the team.

1.7 Refer to the slot depth data in Table 1.1. After the machine was repaired, a sample of three new ceramic parts had slot depths 215, 216, and 213 (thousandths of an inch).

(a) Redraw the X-bar chart and include the additional mean $\bar{x}$.

(b) Does the new $\bar{x}$ fall within the control limits?

1.8 A Canadian manufacturer identified a critical diameter on a crank bore that needed to be maintained within a close tolerance for the product to be successful. Samples of size 4 were taken every hour. The values of the differences (measurement − specification), in ten-thousandths of an inch, are given in Table 1.4.

(a) Calculate the central line for an X-bar chart for the 24 hourly sample means. The centerline is $\bar{\bar{x}} = (4.25 - 3.00 - \cdots - 1.50 + 3.25)/24$.

(b) Is the average of all the numbers in the table, 4 for each hour, the same as the average of the 24 hourly averages? Should it be?

(c) A computer calculation gives the control limits

$$\text{LCL} = -4.48$$
$$\text{UCL} = \ \ 7.88$$

Construct the X-bar chart. Identify hours where the process was out of control.

Table 1.4 The differences (measurement – specification), in ten-thousandths of an inch

Hour	1	2	3	4	5	6	7	8	9	10	11	12
	10	−6	−1	−8	−14	−6	−1	8	−1	5	2	5
	3	1	−3	−3	−5	−2	−6	−3	7	6	1	3
	6	−4	0	−7	−6	−1	−1	9	1	3	1	10
	−2	−3	−7	−2	2	−6	7	11	7	2	4	4
$\bar{x}$	4.25	−3.00	−2.75	−5.00	−5.75	−3.75	−0.25	6.25	3.50	4.00	2.00	5.50

Hour	13	14	15	16	17	18	19	20	21	22	23	24
	5	6	−5	−8	2	7	8	5	8	−5	−2	−1
	9	6	4	−5	8	7	13	4	1	7	−4	5
	9	8	−5	1	−4	5	6	7	0	1	−7	9
	7	10	−2	0	1	3	6	10	−6	2	7	0
$\bar{x}$	7.50	7.50	−2.00	−3.00	1.75	5.50	8.25	6.50	0.75	1.25	−1.50	3.25

Key Terms

Characteristic of interest 16
Classical approach to statistics 12
Descriptive statistics 12
Population 16
Population of units 16
Quality improvement 13
Random number table 18
Reliability 13
Sample 17
Statement of purpose 16
Statistical inference 12
Statistical population 16
X-bar chart 14
Unit 16
Variable 16

CHAPTER

2

ORGANIZATION AND DESCRIPTION OF DATA

CHAPTER OUTLINE

Statistical data, obtained from surveys, experiments, or any series of measurements, are often so numerous that they are virtually useless unless they are condensed, or reduced into a more suitable form. We begin with the use of simple graphics in Section 2.1. Sections 2.2 and 2.3 deal with problems relating to the grouping of data and the presentation of such groupings in graphical form. In Section 2.4 we discuss a relatively new way of presenting data.

Sometimes it may be satisfactory to present data just as they are and let them speak for themselves; on other occasions it may be necessary only to group the data and present the result in tabular or graphical form. However, most of the time data have to be summarized further, and in Sections 2.5 through 2.7 we introduce some of the most widely used kinds of statistical descriptions.

2.1 Pareto Diagrams and Dot Diagrams

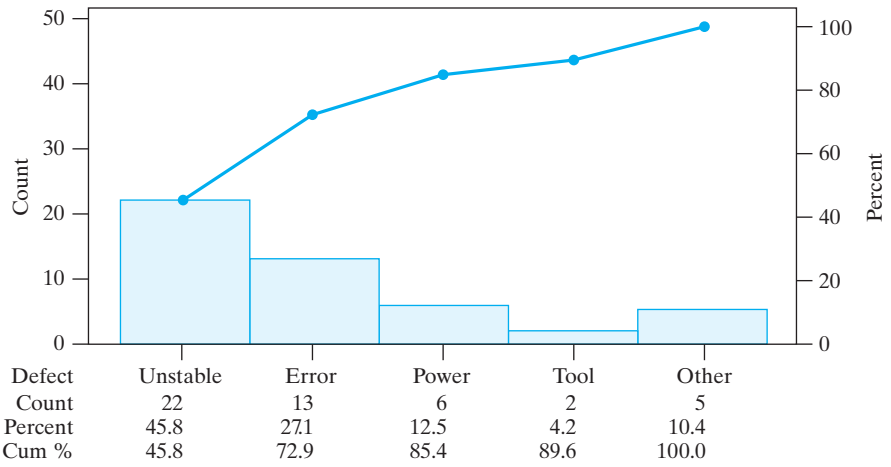
Data need to be collected to provide the vital information necessary to solve engineering problems. Once gathered, these data must be described and analyzed to produce summary information. Graphical presentations can often be the most effective way to communicate this information. To illustrate the power of graphical techniques, we first describe a **Pareto diagram**. This display, which orders each type of failure or defect according to its frequency, can help engineers identify important defects and their causes.

When a company identifies a process as a candidate for improvement, the first step is to collect data on the frequency of each type of failure. For example, the performance of a computer-controlled lathe is below par so workers record the following causes of malfunctions and their frequencies:

power fluctuations	6
controller not stable	22
operator error	13
worn tool not replaced	2
other	5

These data are presented as a special case of a **bar chart** called a **Pareto diagram** in Figure 2.1. This diagram graphically depicts Pareto's empirical law that any assortment of events consists of a few major and many minor elements. Typically, two or three elements will account for more than half of the total frequency.

Concerning the lathe, 22 or $100(22/48) = 46\%$ of the cases are due to an unstable controller and $22 + 13 = 35$ or $100(35/48) = 73\%$ are due to either unstable controller or operator error. These cumulative percentages are shown in Figure 2.1 as a line graph whose scale is on the right-hand side of the Pareto diagram, as appears again in Figure 15.2.



Defect	Unstable	Error	Power	Tool	Other
Count	22	13	6	2	5
Percent	45.8	27.1	12.5	4.2	10.4
Cum %	45.8	72.9	85.4	89.6	100.0

Figure 2.1
A Pareto diagram of failures

In the context of quality improvement, to make the most impact we want to select the few vital major opportunities for improvement. This graph visually emphasizes the importance of reducing the frequency of controller misbehavior. An initial goal may be to cut it in half.

As a second step toward improvement of the process, data were collected on the deviations of cutting speed from the target value set by the controller. The seven observed values of (cutting speed) − (target),

$$3 \quad 6 \quad -2 \quad 4 \quad 7 \quad 4 \quad 3$$

are plotted as a **dot diagram** in Figure 2.2. The dot diagram visually summarizes the information that the lathe is, generally, running fast. In Chapters 13 and 15 we will develop efficient experimental designs and methods for identifying primary causal factors that contribute to the variability in a response such as cutting speed.

Figure 2.2
Dot diagram of cutting speed deviations

When the number of observations is small, it is often difficult to identify any pattern of variation. Still, it is a good idea to plot the data and look for unusual features.

EXAMPLE 1 **Dot diagrams expose outliers**

A major food processor regularly monitors bacteria along production lines that include a stuffing process for meat products. An industrial engineer records the maximum amount of bacteria present along the production line, in the units Aerobic Plate Count per square inch (APC/in^2), for $n = 7$ days. (Courtesy of David Brauch)

$$96.3 \quad 155.6 \quad 3408.0 \quad 333.3 \quad 122.2 \quad 38.9 \quad 58.0$$

Create a dot diagram and comment.

Solution The ordered data

$$38.9 \quad 58.0 \quad 96.3 \quad 122.2 \quad 155.6 \quad 333.3 \quad 3408.0$$

are shown as the dot diagram in Figure 2.3. By using open circles, we help differentiate the crowded smaller values. The one very large bacteria count is the prominent

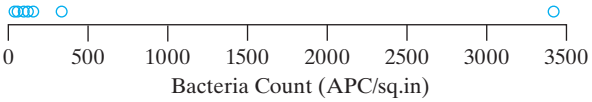


Figure 2.3
Maximum bacteria counts on seven days.

feature. It indicates a possible health concern. Statisticians call such an unusual observation an **outlier**. Usually, outliers merit further attention. ■

EXAMPLE 2 **A dot diagram for multiple samples reveals differences**

The vessels that contain the reactions at some nuclear power plants consist of two hemispherical components welded together. Copper in the welds could cause them to become brittle after years of service. Samples of welding material from one production run or "heat" used in one plant had the copper contents 0.27, 0.35, 0.37. Samples from the next heat had values 0.23, 0.15, 0.25, 0.24, 0.30, 0.33, 0.26. Draw a dot diagram that highlights possible differences in the two production runs (heats) of welding material. If the copper contents for the two runs are different, they should not be combined to form a single estimate.

Solution We plot the first group as solid circles and the second as open circles (see Figure 2.4). It seems unlikely that the two production runs are alike because the top two values are from the first run. (In Exercise 14.23, you are asked to confirm this fact.) The two runs should be treated separately.

The copper content of the welding material used at the power plant is directly related to the determination of safe operating life. Combining the sample would lead to an unrealistically low estimate of copper content and too long an estimate of safe life. ■

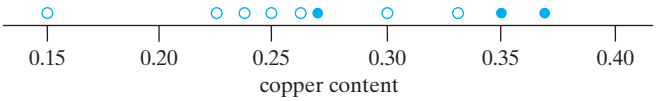


Figure 2.4
Dot diagram of copper content

When a set of data consists of a large number of observations, we take the approach described in the next section. The observations are first summarized in the form of a table.

2.2 Frequency Distributions

A **frequency distribution** is a table that divides a set of data into a suitable number of classes (categories), showing also the number of items belonging to each class. The table sacrifices some of the information contained in the data. Instead of knowing the exact value of each item, we only know that it belongs to a certain class. On the other hand, grouping often brings out important features of the data, and the gain in "legibility" usually more than compensates for the loss of information.

We shall consider mainly **numerical distributions**; that is, frequency distributions where the data are grouped according to size. If the data are grouped according to some quality, or attribute, we refer to such a distribution as a **categorical distribution**.

The first step in constructing a frequency distribution consists of deciding how many classes to use and choosing the **class limits** for each class. That is, deciding from where to where each class is to go. Generally speaking, the number of classes we use depends on the number of observations, but it is seldom profitable to use

fewer than 5 or more than 15. The exception to the upper limit is when the size of the data set is several hundred or even a few thousand. It also depends on the range of the data, namely, the difference between the largest observation and the smallest.

Once the classes are set, we count the number of observations in each class, called the **class frequencies**. This task is simplified if the data are first sorted from smallest to largest.

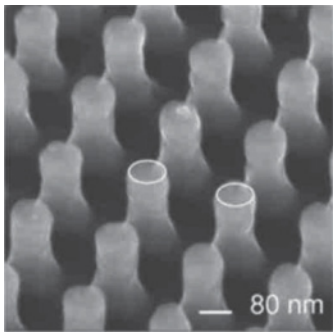


Figure 2.5
Nanopillars

To illustrate the construction of a frequency distribution, we consider data collected in a nanotechnology setting. Engineers fabricating a new transmission-type electron multiplier created an array of silicon nanopillars on a flat silicon membrane. The precise structure can influence the electrical properties, so the heights of 50 nanopillars were measured in nanometers (nm), or $10^{-9}\times$ meters. (See Figure 2.5.)[1]

245	333	296	304	276	336	289	234	253	292
366	323	309	284	310	338	297	314	305	330
266	391	315	305	290	300	292	311	272	312
315	355	346	337	303	265	278	276	373	271
308	276	364	390	298	290	308	221	274	343

Since the largest observation is 391 and the smallest is 221 and the range is $391 - 221 = 170$, we might choose five classes having the limits 206–245, 246–285, 286–325, 326–365, 366–405, or the six classes 216–245, 246–275, …, 366–395. Note that, in either case, **the classes do not overlap, they accommodate all the data, and they are all of the same width**.

Initially, deciding on the first of these classifications, we count the number of observations in each class to obtain the frequency distribution:

Limits of Classes	Frequency
206–245	3
246–285	11
286–325	23
326–365	9
366–405	4
Total	50

Note that the class limits are given to as many decimal places as the original data. Had the original data been given to one decimal place, we would have used the class limits 205.9–245.0, 245.1–285.0, …, 365.1–405.0. If they had been rounded to the nearest 10 nanometers, we would have used the class limits 210–240, 250–280, 290–320, 330–360, 370–400.

In the preceding example, the data on heights of nanopillars may be thought of as values of a continuous variable which, conceivably, can be any value in an interval. But if we use classes such as 205–245, 245–285, 285–325, 325–365, 365–405, there exists the possibility of ambiguities; 245 could go into the first class or the second, 285 could go into the second class or the third, and so on. To avoid this difficulty, we take an alternative approach.

We make an **endpoint convention**. For the pillar height data, we can take (205, 245] as the first class, (245, 285] as the second, and so on through (365, 405]. That is, for this data set, we adopt the convention that the right-hand endpoint is included

[1]Data and photo from H. Qin, H. Kim, and R. Blick, Nanopillar arrays on semiconductor membranes as electron emission amplifiers, *Nanotechnology* **19** (2008), used with permission from IOP Publishing Ltd.

but the left-hand endpoint is not. For other data sets we may prefer to reverse the endpoint convention so the left-hand endpoint is included but the right-hand endpoint is not. Whichever endpoint convention is adopted, it should appear in the description of the frequency distribution.

Under the convention that the right-hand endpoint is included, the frequency distribution of the nanopillar data is

Height (nm)	Frequency
(205, 245]	3
(245, 285]	11
(285, 325]	23
(325, 365]	9
(365, 405]	4
Total	50

The **class boundaries** are the endpoints of the intervals that specify each class. As we pointed out earlier, once data have been grouped, each observation has lost its identity in the sense that its exact value is no longer known. This may lead to difficulties when we want to give further descriptions of the data, but we can avoid them by representing each observation in a class by its midpoint, called the **class mark**. In general, the class marks of a frequency distribution are obtained by averaging successive class boundaries. If the classes of a distribution are all of equal length, as in our example, we refer to the common interval between any successive class marks as the **class interval** of the distribution. Note that the class interval may also be obtained from the difference between any successive class boundaries.

EXAMPLE 3 **Class marks and class interval for grouped data**

With reference to the distribution of the heights of nanopillars, find (a) the class marks and (b) the class interval.

Solution

(a) The class marks are

$$\frac{205 + 245}{2} = 225 \qquad \frac{245 + 285}{2} = 265, \quad 305, \quad 345, \quad 385$$

(b) The class interval is $245 - 205 = 40$. ■

There are several alternative forms of distributions into which data are sometimes grouped. Foremost among these are the "less than or equal to," "less than," "or more," and "equal or more" **cumulative distributions**. A cumulative "less than or equal to" distribution shows the total number of observations that are less than or equal to the given values. These values must be class boundaries, with an appropriate endpoint convention, when the data are grouped into a frequency distribution.

EXAMPLE 4 **Cumulative distribution of the nanopillar heights**

Convert the distribution of the heights of nanopillars into a distribution according to how many observations are less than or equal to 205, less than or equal to 245, …, less than or equal to 405.

Solution Since none of the values is less than 205, 3 are less than or equal to 245, $3 + 11 = 14$ are less than or equal to 285, $14 + 23 = 37$ are less than or equal to 325, $37+9 = 46$ are less than or equal to 365, and all 50 are less than or equal to 405, we have

Heights (mM)	Cumulative Frequency
(205, 245]	3
(245, 285]	14
(285, 325]	37
(325, 365]	46
(365, 405]	50

■

When the endpoint convention for a class includes the left-hand endpoint but not the right-hand endpoint, the cumulative distribution becomes a "less than" cumulative distribution.

Cumulative "more than" and "or more" distributions are constructed similarly by adding the frequencies, one by one, starting at the other end of the frequency distribution. In practice, "less than or equal to" cumulative distributions are used most widely, and it is not uncommon to refer to "less than or equal to" cumulative distributions simply as *cumulative distributions*.

2.3 Graphs of Frequency Distributions

Properties of frequency distributions relating to their shape are best exhibited through the use of graphs, and in this section we shall introduce some of the most widely used forms of graphical presentations of frequency distributions and cumulative distributions.

The most common form of graphical presentation of a frequency distribution is the **histogram**. The histogram of a frequency distribution is constructed of adjacent rectangles. Provided that the *class intervals are equal*, the heights of the rectangles represent the class frequencies and the bases of the rectangles extend between successive class boundaries. A histogram of the heights of nanopillars data is shown in Figure 2.6.

Using our endpoint convention, the interval (205, 245] that defines the first class has frequency 3, so the rectangle has height 3, the second rectangle, over the interval

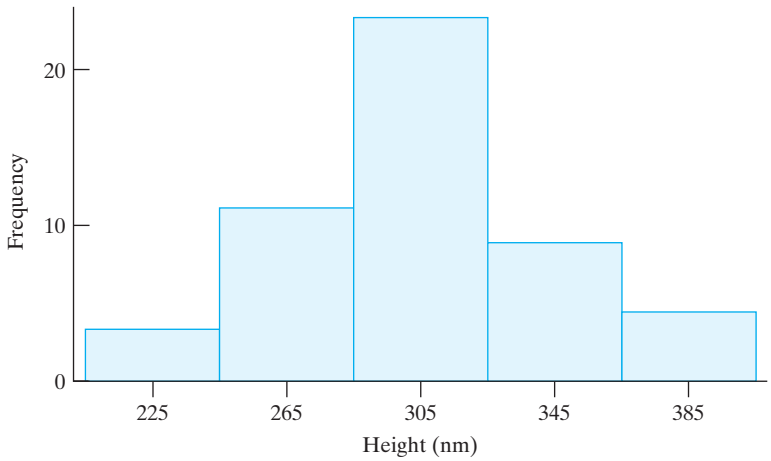


Figure 2.6
Histogram of pillar height

(245, 285], has height 9, and so on. The tallest rectangle is over the interval (285, 325] and has height 23. The histogram has a single peak and is reasonably symmetric. Almost half of the area, representing half of the observations, is over the interval 285 to 325 nanometers.

The choice of frequency, or relative frequency, for the vertical scale is only valid when all of the classes have the same width.

Inspection of the graph of a frequency distribution as a histogram often brings out features that are not immediately apparent from the data themselves. Aside from the fact that such a graph presents a good overall picture of the data, it can also emphasize irregularities and unusual features. It can reveal outlying observations which somehow do not fit the overall picture. Their distruption of the overall pattern of variation in the data may be due to errors of measurement, equipment failure, and similar causes. Also, the fact that a histogram exhibits two or more *peaks* (maxima) can provide pertinent information. The appearance of two peaks may imply, for example, a shift in the process that is being measured, or it may imply that the data come from two or more sources. With some experience one learns to spot such irregularities or anomalies, and an experienced engineer would find it just as surprising if the histogram of a distribution of integrated-circuit failure times were symmetrical as if a distribution of American men's hat sizes were bimodal.

Sometimes it can be enough to draw a histogram in order to solve an engineering problem.

EXAMPLE 5 **A histogram reveals the solution to a grinding operation problem**

A metallurgical engineer was experiencing trouble with a grinding operation. The grinding action was produced by pellets. After some thought he collected a sample of pellets used for grinding, took them home, spread them out on his kitchen table, and measured their diameters with a ruler. His histogram is displayed in Figure 2.7. What does the histogram reveal?

Solution The histogram exhibits two distinct peaks, one for a group of pellets whose diameters are centered near 25 and the other centered near 40.

By getting his supplier to do a better sort, so all the pellets would be essentially from the first group, the engineer completely solved his problem. Taking the action to obtain the data was the big step. The analysis was simple. ■

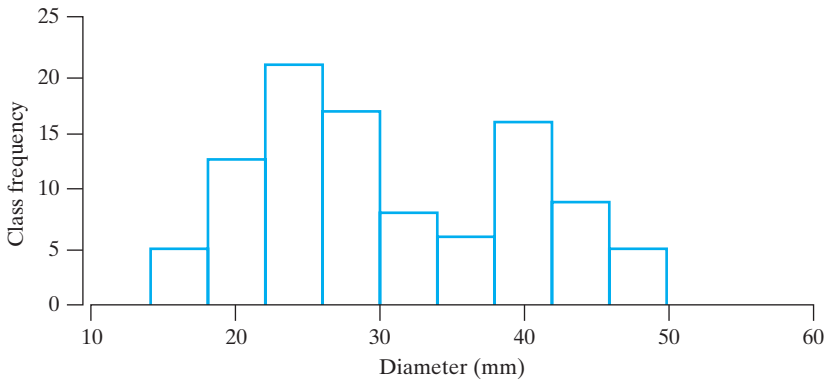


Figure 2.7
Histogram of pellet diameter

As illustrated by the next example concerning a system of supercomputers, not all histograms are symmetric.

EXAMPLE 6 **A histogram reveals the pattern of a supercomputer systems data**

A computer scientist, trying to optimize system performance, collected data on the time, in microseconds, between requests for a particular process service.

2,808	4,201	3,848	9,112	2,082	5,913	1,620	6,719	21,657
3,072	2,949	11,768	4,731	14,211	1,583	9,853	78,811	6,655
1,803	7,012	1,892	4,227	6,583	15,147	4,740	8,528	10,563
43,003	16,723	2,613	26,463	34,867	4,191	4,030	2,472	28,840
24,487	14,001	15,241	1,643	5,732	5,419	28,608	2,487	995
3,116	29,508	11,440	28,336	3,440				

Draw a histogram using the equal length classes [0, 10,000), [10,000, 20,000), . . . , [70,000, 80,000) where the left-hand endpoint is included but the right-hand endpoint is not.

Solution The histogram of this interrequest time data, shown in Figure 2.8, has a long right-hand tail. Notice that, with this choice of equal length intervals, two classes are empty. To emphasize that it is still possible to observe interrequest times in these intervals, it is preferable to regroup the data in the right-hand tail into classes of unequal lengths (see Exercise 2.62). ■

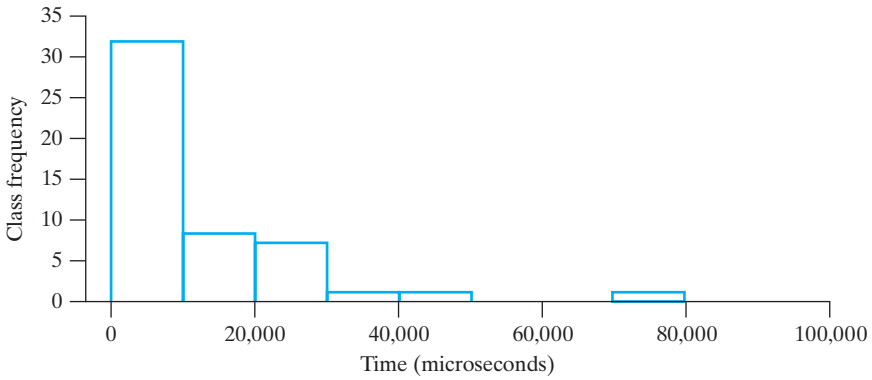


Figure 2.8
Histogram of interrequest time

When a histogram is constructed from a frequency table having classes of unequal lengths, the height of each rectangle must be changed to

$$\text{height} = \frac{\text{relative frequency}}{\text{width}}$$

The area of the rectangle then represents the relative frequency for the class and the total area of the histogram is 1. We call this a **density histogram**.

EXAMPLE 7 **A density histogram has total area 1**

Compressive strength was measured on 58 specimens of a new aluminum alloy undergoing development as a material for the next generation of aircraft.

66.4	67.7	68.0	68.0	68.3	68.4	68.6	68.8	68.9	69.0	69.1
69.2	69.3	69.3	69.5	69.5	69.6	69.7	69.8	69.8	69.9	70.0
70.0	70.1	70.2	70.3	70.3	70.4	70.5	70.6	70.6	70.8	70.9
71.0	71.1	71.2	71.3	71.3	71.5	71.6	71.6	71.7	71.8	71.8
71.9	72.1	72.2	72.3	72.4	72.6	72.7	72.9	73.1	73.3	73.5
74.2	74.5	75.3								

Draw a density histogram, that is, a histogram scaled to have a total area of 1 unit. For reasons to become apparent in Chapter 6, we call the vertical scale **density**.

Solution We make the height of each rectangle equal to *relative frequency/width*, so that its area equals the relative frequency. The resulting histogram, constructed by computer, has a nearly symmetric shape (see Figure 2.9). We have also graphed a continuous curve that approximates the overall shape. In Chapter 5, we will introduce this bell-shaped family of curves. ■

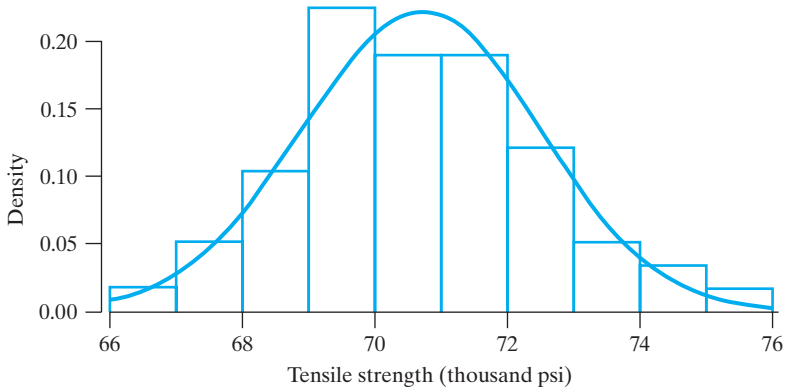


Figure 2.9
Histogram of aluminum alloy tensile strength

[Using **R**: **with (sample, hist (strength,prob=TRUE,las=1))** after **sample=read. table ("C2Ex.TXT",header=TRUE)**]

This example suggests that histograms, for observations that come from a continuous scale, can be approximated by smooth curves.

Cumulative distributions are usually presented graphically in the form of **ogives**, where we plot the cumulative frequencies at the class boundaries. The resulting points are connected by means of straight lines, as shown in Figure 2.10, which represents the cumulative "less than or equal to" distribution of nanopillar height data on page 25. The curve is steepest over the class with highest frequency.

When the endpoint convention for a class includes the left-hand endpoint but not the right-hand endpoint, the ogive represents a "less than" cumulative distribution.

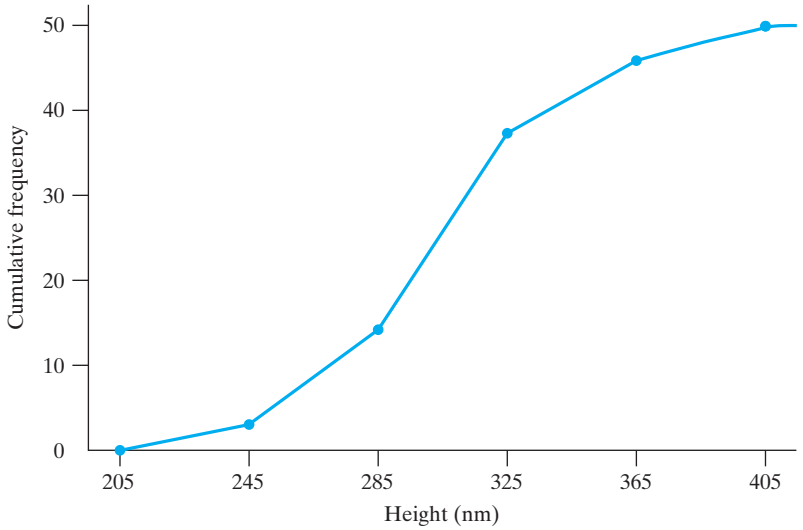


Figure 2.10
Ogive of heights of nanopillars

2.4 Stem-and-Leaf Displays

In the two preceding sections we directed our attention to the grouping of relatively large sets of data with the objective of putting such data into a manageable form. As we saw, this entailed some loss of information. Similar techniques have been proposed for the preliminary explorations of small sets of data, which yield a good overall picture of the data without any loss of information.

To illustrate, consider the following humidity readings rounded to the nearest percent:

29	44	12	53	21	34	39	25	48	23
17	24	27	32	34	15	42	21	28	37

Proceeding as in Section 2.2, we might group these data into the following distribution:

Humidity Readings	Frequency
10–19	3
20–29	8
30–39	5
40–49	3
50–59	1

If we wanted to avoid the loss of information inherent in the preceding table, we could keep track of the last digits of the readings within each class, getting

10–19	2 7 5
20–29	9 1 5 3 4 7 1 8
30–39	4 9 2 4 7
40–49	4 8 2
50–59	3

This can also be written as

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

where the left-hand column, the **stem**, gives the tens digits 10, 20, 30, 40, and 50. The numbers in a row, the leaves, have the unit 1.0. In the last step, the leaves are written in ascending order. The three numbers in the first row are 12, 15, and 17. This table is called a **stem-and-leaf display** or simply a **stem-leaf display**. The left-hand column forms the *stem*, and the numbers to the left of the vertical line are the **stem labels**, which in our example are 1, 2, . . . , 5. Each number to the right of the vertical line is a **leaf**. There should not be any gaps in the stem even if there are no leaves for that particular value.

Essentially, a stem-and-leaf display presents the same picture as the corresponding tally, yet it retains all the original information. For instance, if a stem-and-leaf display has the two-digit stem

1.2 | 0 2 3 5 8

where the leaf unit = 0.01, the corresponding data are 1.20, 1.22, 1.23, 1.25, and 1.28. If a stem-and-leaf display has the two digit leaves

0.3 | 03 17 55 89

where the first leaf digit unit = 0.01, the corresponding data are 0.303, 0.317, 0.355, and 0.389.

There are various ways in which stem-and-leaf displays can be modified to meet particular needs (see Exercises 2.25 and 2.26), but we shall not go into this here in any detail as it has been our objective to present only one of the relatively new techniques, which come under the general heading of **exploratory data analysis**.

Exercises

2.1 Damages at a factory manufacturing chairs are categorized according to the material wasted.

plastic	75
iron	31
cloth	22
spares	8

Draw a Pareto chart.

2.2 Losses at an oil refinery (in millions of dollars) due to excess heat can be divided according to the reason behind the generation of excessive heat.

oversupplying fuel	202
excess air	124
carelessness of operator	96
incomplete combustion	27

(a) Draw a Pareto chart.

(b) What percent of the loss occurs due to

(1) excess air?

(2) excess air and oversupplying fuel?

2.3 Tests were conducted to measure the running temperature for engines (in °F). A sample of 15 tests yielded the temperature values:

182	184	184	186	180	198	195	194
197	200	188	188	194	197	184	

Construct a dot diagram.

2.4 To determine the strengths of various detergents, the following are 20 measurements of the total dissolved salts (parts per million) in water:

168	170	148	160	168	164	175	178
165	168	152	170	172	192	182	164
152	160	170	172				

Construct a dot diagram.

2.5 Civil engineers help municipal wastewater treatment plants operate more efficiently by collecting data on the quality of the effluent. On seven occasions, the amounts of suspended solids (parts per million) at one plant were

14 12 21 28 30 65 26

Display the data in a dot diagram. Comment on your findings.

2.6 A dam on a river holds water in its reservoir to generate electricity. Because the dam is in a rainforest area, the flow of water is highly uncertain. In December last year, the overflow from the reservoir (in million cubic meters) on 14 different days was

26	24	25.5	23.5	25.5	23	23
24	25	24	26	23.5	25	20

Display the data in a dot diagram.

2.7 Physicists first observed neutrinos from a supernova that occurred outside of our solar system when the detector near Kamiokande, Japan, recorded twelve arrivals. The times(seconds) between the neutrinos are

0.107 0.196 0.021 0.281 0.179 0.854 0.58
0.19 7.30 1.18 2.00

(a) Draw a dot diagram.

(b) Identify any outliers.

2.8 The power generated (MW) by liquid hydrogen turbo pumps, given to the nearest tenth, is grouped into a table having the classes [40.0, 45.0), [45.0, 50.0), [50.0, 55.0), [55.0, 60.0) and [60.0, 65.0), where the left-hand endpoint is included but the right-hand endpoint is not. Find

(a) the class marks

(b) the class interval

2.9 With reference to the preceding exercise, is it possible to determine from the grouped data how many turbo pumps have a power generation of

(a) more than 50.0?

(b) less than 50.0?

(c) at most 60.0?

(d) at least 60.0?

(e) 50.0 to 60.0 inclusive?

2.10 To continually increase the speed of computers, electrical engineers are working on ever-decreasing scales.

The size of devices currently undergoing development is measured in nanometers (nm), or $10^{-9}\times$ meters. Engineers fabricating a new transmission-type electron multiplier[2] created an array of silicon nanopillars on a flat silicon membrane. Subsequently, they measured the diameters (nm) of 50 pillars.

62	68	69	80	68	79	83	70	74	73
74	75	80	77	80	83	73	79	100	93
92	101	87	96	99	94	102	95	90	98
86	93	91	90	95	97	87	89	100	93
92	98	101	97	102	91	87	110	106	118

Group these measurements into a frequency distribution and construct a histogram using (60,70], (70, 80], (80,90], (90,100], (100, 110], (110,120], where the right-hand endpoint is included but the left-hand endpoint is not.

2.11 Convert the distribution obtained in the preceding exercise into a cumulative "less than or equal to" distribution and graph its ogive.

2.12 The following are the sizes of particles of cement dust (given to the nearest hundredth of a micron) in a cement factory:

16.12	10.48	11.12	16.18	18.13	19.10	13.21	10.12
21.18	15.12	10.11	13.31	18.61	11.43	18.26	13.77
13.24	12.16	17.19	11.36	12.53	13.25	10.67	15.45
14.28	14.32	15.18	14.21	10.20	15.64	11.68	18.76
19.32	17.50	11.46	20.59	16.38	21.42	16.27	21.30
16.12	10.55	11.49	15.48	11.62	13.54	13.69	16.72
15.11	14.33	17.23	17.22	19.37	10.41	18.28	19.29
21.23	12.56	12.57	11.60	15.24	21.65	20.70	11.44
12.22	19.34	20.35	19.47	21.63	19.40	19.75	21.71
15.19	18.51	10.58	13.52	11.39	13.66	21.73	11.74

Group these figures into a table with a suitable number of equal classes and construct a histogram.

2.13 Convert the distribution obtained in Exercise 2.12 into a cumulative "less than" distribution and plot its ogive.

2.14 An engineer uses a thermocouple to monitor the temperature of a stable reaction. The ordered values of 50 observations (Courtesy of Scott Sanders), in tenths of °C, are

1.11	1.21	1.21	1.21	1.23	1.24	1.25	1.25	1.27	1.27	1.28
1.29	1.31	1.31	1.31	1.32	1.34	1.34	1.35	1.36	1.36	1.36
1.36	1.36	1.36	1.36	1.37	1.39	1.40	1.41	1.42	1.42	1.42
1.42	1.43	1.43	1.43	1.44	1.44	1.44	1.47	1.48	1.48	1.50
1.50	1.56	1.56	1.60	1.60	1.68					

Group these figures into a distribution having the classes 1.10–1.19, 1.20–1.29, 1.30–1.39, . . . , and 1.60–1.69, and plot a histogram using [1.10, 1.20), . . . , [1.60, 1.70), where the left-hand endpoint is included but the right-hand endpoint is not.

2.15 Convert the distribution obtained in Exercise 2.14 into a cumulative "less than" distribution and plot its ogive.

2.16 The following are the number of transistors failing a quality check per hour during 72 observed hours of production:

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

Group these data into a frequency distribution showing how often each of the values occurs and draw a bar chart.

2.17 Given a set of observations $x_1, x_2, \ldots, x_n$, we define their empirical cumulative distribution as the function whose values $F(x)$ equals the proportion of the observations less than or equal to x. Graph the empirical cumulative distribution for the 15 measurements of Exercise 2.3.

2.18 Referring to Exercise 2.17, graph the empirical cumulative distribution for the data in Exercise 2.16.

2.19 The pictogram of Figure 2.11 is intended to illustrate the fact that per capita income in the United States doubled from $21,385 in 1993 to $42,643 in 2012. Does this pictogram convey a fair impression of the actual change? If not, state how it might be modified.

Figure 2.11 Pictogram for Exercise 2.19

2.20 Categorical distributions are often presented graphically by means of **pie charts**, in which a circle is divided into sectors proportional in size to the frequencies (or percentages) with which the data are distributed among the categories. Draw a pie chart to represent the following data, obtained in a study in

[2]H. Qin, H. Kim, and R. Blick, *Nanotechnology* **19** (2008), 095504. (5pp)

which 40 drivers were asked to judge the maneuverability of a certain make of car:

Very good, good, good, fair, excellent, good, good, good, very good, poor, good, good, good, good, very good, good, fair, good, good, very poor, very good, fair, good, good, excellent, very good, good, good, good, fair, fair, very good, good, very good, excellent, very good, fair, good, good, and very good.

2.21 Convert the distribution of nanopillar heights on page 26 into a distribution having the classes (205, 245], (245, 325], (325, 365], (365, 405], where the right-hand endpoint is included. Draw two histograms of this distribution, one in which the class frequencies are given by the heights of the rectangles and one in which the class frequencies are given by the area of the rectangles. Explain why the first of these histograms gives a very misleading picture.

2.22 The following are figures on sacks of cement used daily at a construction site: 75, 77, 82, 45, 55, 90, 80, 81, 76, 47, 59, 52, 71, 83, 91, 76, 57, 59, 43 and 79. Construct a stem-and-leaf display with the stem labels 4, 5, . . . , and 9.

2.23 The following are determinations of a river's annual maximum flow in cubic meters per second: 405, 355, 419, 267, 370, 391, 612, 383, 434, 462, 288, 317, 540, 295, and 508. Construct a stem-and-leaf display with two-digit leaves.

2.24 List the data that correspond to the following stems of stem-and-leaf displays:

(a) 4 | 0 1 1 2 5 7 Leaf unit = 1.0

(b) 62 | 3 5 5 8 9 Leaf unit = 1.0

(c) 8 | 01 23 62 91 First leaf digit unit = 10.0

(d) 2.28 | 4 5 6 6 8 9 Leaf unit = 0.001

2.25 To construct a stem-and-leaf display with more stems than there would be otherwise, we might repeat each stem. The leaves 0, 1, 2, 3, and 4 would be attached to the first stem and leaves 5, 6, 7, 8, and 9 to the second. For the humidity readings on page 31, we would thus get the **double-stem display**:

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

where we doubled the number of stems by cutting the interval covered by each stem in half. Construct a double-stem display with one-digit leaves for the data in Exercise 2.14.

2.26 If the double-stem display has too few stems, we create 5 stems where the first holds leaves 0 and 1, the second holds 2 and 3, and so on. The resulting stem-and-leaf display is called a **five-stem display**.

(a) The following are the IQs of 20 applicants to an undergraduate engineering program: 109, 111, 106, 106, 125, 108, 115, 109, 107, 109, 108, 110, 112, 104, 110, 112, 128, 106, 111, and 108. Construct a five-stem display with one-digit leaves.

(b) The following is part of a five-stem display:

53	4 4 4 4 5 5	Leaf unit = 1.0
53	6 6 6 7	
53	8 9	
54	1	

List the corresponding measurements.

2.5 Descriptive Measures

Histograms, dot diagrams, and stem-and-leaf diagrams summarize a data set pictorially so we can visually discern the overall pattern of variation. Numerical measures can augment visual displays when describing a data set. To proceed, we introduce the notation

$$x_1, x_2, \ldots, x_i, \ldots, x_n$$

for a general sample consisting of n measurements. Here x_i is the ith observation in the list so x_1 represents the value of the first measurement, x_2 represents the value of the second measurement, and so on.

Given a set of n measurements or observations, $x_1, x_2, \ldots, x_n$, there are many ways in which we can describe their center (middle, or central location). Most popular among these are the **arithmetic mean** and the **median**, although other kinds

of "averages" are sometimes used for special purposes. The arithmetic mean—or, more succinctly, the **mean**—is defined as the sum of the observations divided by sample size.

Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The notation $\bar{x}$, read x bar, represents the mean of the x_i. To emphasize that it is based on the observations in a data set, we often refer to $\bar{x}$ as the **sample mean**.

Sometimes it is preferable to use the **sample median** as a descriptive measure of the center, or location, of a set of data. This is particularly true if it is desired to minimize the calculations or if it is desired to eliminate the effect of extreme (very large or very small) values. The median of n observations $x_1, x_2, \ldots, x_n$ can be defined loosely as the "middlemost" value once the data are arranged according to size. More precisely, if the observations are arranged according to size and n is an odd number, the median is the value of the observation numbered $\frac{n+1}{2}$; if n is an even number, the median is defined as the mean (average) of the observations numbered $\frac{n}{2}$ and $\frac{n+2}{2}$.

Sample median

Order the n observations from smallest to largest.

$$\begin{aligned} \text{sample median} &= \text{observation in position } \frac{n+1}{2}, && \text{if } n \text{ odd.} \\ &= \text{average of two observations in} \\ &\quad \text{positions } \frac{n}{2} \text{ and } \frac{n+2}{2}, && \text{if } n \text{ even.} \end{aligned}$$

EXAMPLE 8 **Calculation of the sample mean and median**

A sample of five university students responded to the question "How much time, in minutes, did you spend on the social network site yesterday?"

100 45 60 130 30

Find the mean and the median.

Solution The mean is

$$\bar{x} = \frac{100 + 45 + 60 + 130 + 30}{5} = 73 \text{ minutes}$$

and, ordering the data from smallest to largest

30 45 $\underbrace{60}$ 100 130

the median is the third largest value, namely, 60 minutes.

The two very large values cause the mean to be much larger than the median. ■

EXAMPLE 9 **Calculation of the sample median with even sample size**

An engineering group receives e-mail requests for technical information from sales and service. The daily numbers of e-mails for six days are

$$11 \quad 9 \quad 17 \quad 19 \quad 4 \quad 15$$

Find the mean and the median.

Solution The mean is

$$\bar{x} = \frac{11 + 9 + 17 + 19 + 4 + 15}{6} = 12.5 \text{ requests}$$

and, ordering the data from the smallest to largest

$$4 \quad 9 \quad \underbrace{11 \quad 15} \quad 17 \quad 19$$

the median, the mean of the third and fourth largest values, is 13 requests. ■

The sample mean has a physical interpretation as the balance point, or center of mass, of a data set. Figure 2.12 is the dot diagram for the data on the number of e-mail requests given in the previous example. In the dot diagram, each observation is represented by a ball placed at the appropriate distance along the horizontal axis. If the balls are considered as masses having equal weights and the horizontal axis is weightless, then the mean corresponds to the center of inertia or balance point of the data. This interpretation of the sample mean, as the balance point of the observations, holds for any data set.

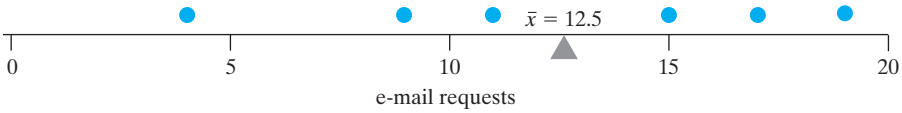


Figure 2.12
The interpretation of the sample mean as a balance point

Although the mean and the median each provide a single number to represent an entire set of data, the mean is usually preferred in problems of estimation and other problems of statistical inference. An intuitive reason for preferring the mean is that the median does not utilize all the information contained in the observations.

The following is an example where the median actually gives a more useful description of a set of data than the mean.

EXAMPLE 10 **The median is unaffected by a few outliers**

A small company employs four young engineers, who each earn $80,000, and the owner (also an engineer), who gets $200,000. Comment on the claim that on the average the company pays $104,000 to its engineers and, hence, is a good place to work.

Solution The mean of the five salaries is $104,000, but it hardly describes the situation. The median, on the other hand, is $80,000, and it is most representative of what a young engineer earns with the firm. Moneywise, the company is not such a good place for young engineers. ■

This example illustrates that there is always an inherent danger when summarizing a set of data in terms of a single number.

One of the most important characteristics of almost any set of data is that the values are not all alike; indeed, the extent to which they are unlike, or vary among themselves, is of basic importance in statistics. The mean and median describe one

important aspect of a set of data—their "middle" or their "average"—but they tell us nothing about the extent of variation.

We observe that the dispersion of a set of data is small if the values are closely bunched about their mean, and that it is large if the values are scattered widely about their mean. It would seem reasonable, therefore, to measure the variation of a set of data in terms of the amounts by which the values deviate from their mean.

If a set of numbers $x_1, x_2, \ldots, x_n$ has mean $\bar{x}$, the differences

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$$

are called the **deviations from the mean**. We might use the average of the deviations as a measure of variation in the data set. Unfortunately, this will not do. For instance, refer to the observations 11, 9, 17, 19, 4, 15, displayed above in Figure 2.12, where $\bar{x} = 12.5$ is the balance point. The six deviations are -1.5, -3.5, 4.5, 6.5, -8.5, and 2.5. The sum of positive deviations

$$4.5 + 6.5 + 2.5 = 13.5$$

exactly cancels the sum of the negative deviations

$$-1.5 \ - 3.5 \ - 8.5 = -13.5$$

so the sum of all the deviations is 0.

As you will be asked to show in Exercise 2.50, the sum of the deviations is always zero. That is,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

so the mean of the deviations is always zero. Because the deviations sum to zero, we need to remove their signs. Absolute value and square are two natural choices. If we take their absolute value, so each negative deviation is treated as positive, we would obtain a measure of variation. However, to obtain the most common measure of variation, we square each deviation. The **sample variance**, s^2, is essentially the average of the squared deviations from the mean, $\bar{x}$, and is defined by the following formula.

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Our reason for dividing by $n-1$ instead of n is that there are only $n-1$ independent deviations $x_i - \bar{x}$. Because their sum is always zero, the value of any particular one is always equal to the negative of the sum of the other $n - 1$ deviations.

If many of the deviations are large in magnitude, either positive or negative, their squares will be large and s^2 will be large. When all the deviations are small, s^2 will be small.

EXAMPLE 11 **Calculation of sample variance**

The delay times (handling, setting, and positioning the tools) for cutting 6 parts on an engine lathe are 0.6, 1.2, 0.9, 1.0, 0.6, and 0.8 minutes. Calculate s^2.

Solution First we calculate the mean:

$$\bar{x} = \frac{0.6 + 1.2 + 0.9 + 1.0 + 0.6 + 0.8}{6} = 0.85$$

To find $\sum (x_i - \bar{x})^2$, we set up the table:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.6	−0.25	0.0625
1.2	0.35	0.1225
0.9	0.05	0.0025
1.0	0.15	0.0225
0.6	−0.25	0.0625
0.8	−0.05	0.0025
5.1	0.00	0.2750

where the total of the third column $0.2750 = \sum(x_i - \bar{x})^2$.

We divide 0.2750 by $6 - 1 = 5$ to obtain

$$s^2 = \frac{0.2750}{5} = 0.055 \text{ (minute)}^2$$

By calculating the sum of deviations in the second column, we obtain a check on our work. For all data sets, this sum should be 0 up to rounding error. ■

Notice that the units of s^2 are not those of the original observations. The data are delay times in minutes, but s^2 has the unit $(\text{minute})^2$. Consequently, we define the **standard deviation** of n observations $x_1, x_2, \ldots, x_n$ as the square root of their variance, namely

Sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

The standard deviation is by far the most generally useful measure of variation. Its advantage over the variance is that it is expressed in the same units as the observations.

EXAMPLE 12 **Calculation of sample standard deviation**

With reference to the previous example, calculate s.

Solution From the previous example, $s^2 = 0.055$. Take the square root and get

$$s = \sqrt{0.055} = 0.23 \text{ minute}$$

[Using **R**: Enter data **x = c(.6, 1.2, .9, l, .6, .8)**. Then **mean(x)**, **var(x)**, and **sd(x)**] ■

The standard deviation s has a rough interpretation as the average distance from an observation to the sample mean.

The standard deviation and the variance are measures of **absolute variation**; that is, they measure the actual amount of variation in a set of data, and they depend on the scale of measurement. To compare the variation in several sets of data, it is generally desirable to use a measure of **relative variation**, for instance, the **coefficient of variation**, which gives the standard deviation as a percentage of the mean.

Coefficient of variation

$$V = \frac{s}{\bar{x}} \cdot 100\%$$

EXAMPLE 13 **The coefficient of variation for comparing relative preciseness**

Measurements made with one micrometer of the diameter of a ball bearing have a mean of 3.92 mm and a standard deviation of 0.0152 mm, whereas measurements made with another micrometer of the unstretched length of a spring have a mean of 1.54 inches and a standard deviation of 0.0086 inch. Which of these two measuring instruments is relatively more precise?

Solution For the first micrometer the coefficient of variation is

$$V = \frac{0.0152}{3.92} \cdot 100 = 0.39\%$$

and for the second micrometer the coefficient of variation is

$$V = \frac{0.0086}{1.54} \cdot 100 = 0.56\%$$

Thus, the measurements made with the first micrometer are relatively more precise. ■

In this section, we have limited the discussion to the sample mean, median, variance, and standard deviation. However, there are many other ways of describing sets of data.

2.6 Quartiles and Percentiles

In addition to the median, which divides a set of data into halves, we can consider other division points. When an ordered data set is divided into quarters, the resulting division points are called sample **quartiles**. The *first quartile*, Q_1, is a value that has one-fourth, or 25%, of the observations below its value. The first quartile is also the sample 25th **percentile** $P_{0.25}$. More generally, we define the sample 100 pth percentile as follows.

Sample percentiles

The sample 100 pth percentile is a value such that at least $100p\%$ of the observations are at or below this value, and at least $100(1 - p)\%$ are at or above this value.

As in the case of the median, which is the 50th percentile, this may not uniquely define a percentile. Our convention is to take an observed value for the sample percentile unless two adjacent values both satisfy the definition. In this latter case, take their mean. This coincides with the procedure for obtaining the median when the sample size is even. (Most computer programs linearly interpolate between the two adjacent values. For moderate or large sample sizes, the particular convention used to locate a sample percentile between the two observations is inconsequential.)

The following rule simplifies the calculation of sample percentiles.

Calculating the sample 100 pth percentile:

1. Order the n observations from smallest to largest.
2. Determine the product np.

   If np is not an integer, round it up to the next integer and find the corresponding ordered value.

   If np is an integer, say k, calculate the mean of the kth and $(k+1)$st ordered observations.

The quartiles are the 25th, 50th, and 75th percentiles.

Sample quartiles

first quartile	$Q_1 = 25$th percentile
second quartile	$Q_2 = 50$th percentile
third quartile	$Q_3 = 75$th percentile

EXAMPLE 14 **Calculation of percentiles for the strength of green materials**

Of all the waste materials entering landfills, a substantial proportion consists of construction and demolition materials. From the standpoint of green engineering, before incorporating these materials into the base for new or rehabilitated roadways, engineers must assess their strength. Generally, higher values imply a stiffer base which increases pavement life.

Measurements of the resiliency modulus (MPa) on $n = 18$ specimens of recycled concrete aggregate produce the ordered values (Courtesy of Tuncer Edil)

136	143	147	151	158	160
161	163	165	167	173	174
181	181	185	188	190	205

Obtain the quartiles and the 10th percentile.

Solution According to our calculation rule, $np = 18\left(\frac{1}{4}\right) = 4.5$, which we round up to 5. The first quartile is the 5th ordered observation

$$Q_1 = 158\,\text{MPa}$$

Since $p = \frac{1}{2}$ for the second quartile, or median,

$$np = 18\left(\frac{1}{2}\right) = 9$$

which is an integer. Therefore, we average the 9th and 10th ordered values

$$Q_2 = \frac{165 + 167}{2} = 166\,\text{MPa}$$

The third quartile is the 14th observation, $Q_3 = 181$ seconds. We could also have started at the largest value and counted down to the 5th position.

To obtain the 10th percentile, we determine that $np = 18 \times 0.10 = 1.8$, which we round up to 2. Counting to the 2nd position, we obtain

$$P_{0.10} = 143\,\text{MPa}$$

The 10th percentile provides a useful description regarding the resiliency modulus of the lowest 10% green pavement specimens.

In the context of monitoring green materials we also record that the maximum resiliency modulus measured was 205 MPa.

[Using **R**: **with(*x*, quantile(resiliency, c(.25,.5,.75,.10),type=2))** after ***x*=read.table("C2Ex14.TXT",header=TRUE)**] ■

The **minimum** and **maximum** observations also convey information concerning the amount of variability present in a set of data. Together, they describe the interval containing all of the observed values and whose length is the

$$\textbf{range} = \text{maximum} - \text{minimum}$$

Care must be taken when interpreting the range since a single large or small observation can greatly inflate its value.

The amount of variation in the middle half of the data is described by the

$$\textbf{interquartile range} = \text{third quartile} - \text{first quartile} = Q_3 - Q_1$$

EXAMPLE 15 **The range and interquartile range for the materials data**

Obtain the range and interquartile range for the resiliency modulus data in Example 14.

Solution The minimum = 136. From the previous example, the maximum = 205, $Q_1 = 158$, and $Q_3 = 181$.

$$\text{range} = \text{maximum} - \text{minimum} = 205 - 136 = 69\,\text{MPa}$$

$$\text{interquartile range} = Q_3 - Q_1 = 181 - 158 = 23\,\text{MPa}$$ ■

Boxplots

The summary information contained in the quartiles is highlighted in a graphic display called a **boxplot**. The center half of the data, extending from the first to the third quartile, is represented by a rectangle. The median is identified by a bar within this box. A line extends from the third quartile to the maximum, and another line extends from the first quartile to the minimum. (For large data sets the lines may only extend to the 95th and 5th percentiles.)

Figure 2.13 gives the boxplot for the green pavement data. The median is closer to Q_1 than Q_3.

A **modified boxplot** can both identify outliers and reduce their effect on the shape of the boxplot. The outer line extends to the largest observation only if it is not too far from the third quartile. That is, for the line to extend to the largest observation, it must be within 1.5 × (interquartile range) units of Q_3. The line from

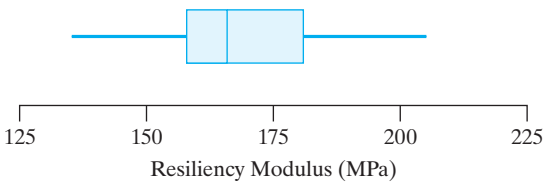


Figure 2.13
Boxplot of the resiliency modulus of green pavement.

Q_1 extends to the smallest observation if it is within that same limit. Otherwise the line extends to the next most extreme observations that fall within this interval.

EXAMPLE 16 **A modified boxplot—possible outliers are detached**

Physicists, trying to learn about neutrinos, detected twelve of them coming from a supernova outside of our solar system. The $n = 11$ times (seconds) between the arrivals are presented in their original order in Exercise 2.7, page 32.

The ordered interarrival times are

0.021 0.107 0.179 0.190 0.196 0.283 0.580 0.854 1.18 2.00 7.30

Construct a modified boxplot.

Solution Since $n/4 = 11/4 = 2.75$, the first quartile is the third ordered time 0.179 and $Q_3 = 1.18$, so the interquartile range is $1.18 - 0.179 = 1.001$. Further, $1.5 \times 1.001 = 1.502$ and the smallest observation is closer than this to $Q_1 = 0.179$, but

$$\text{maximum} - Q_3 = 7.30 - 1.18 = 6.12$$

exceeds $1.502 = 1.5 \times$ (interquartile range)

As shown in Figure 2.14, the line to the right extends to 2.00, the most extreme observation within 1.502 units, but not to the largest observation, which is shown as detached from the line.

[Using **R**: **with(x, boxplot(time,horizontal=TRUE)** after **x=read.table ("C2Exl4.TXT",header=TRUE)**] ■

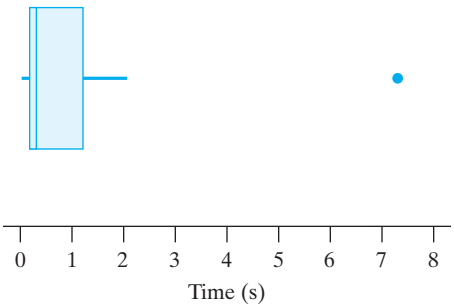


Figure 2.14
Modified boxplot for neutrino data

Boxplots are particularly effective for graphically portraying comparisons among sets of observations. They are easy to understand and have a high visual impact.

EXAMPLE 17 **Multiple boxplots can reveal differences and similarities**

Sometimes, with rather complicated components like hard-disk drives or random access memory (RAM) chips for computers, quality is quantified as an index with target value 100. Typically, a quality index will be based upon the deviations of several physical characteristics from their engineering specifications. Figure 2.15 shows the quality index at 4 manufacturing plants.

Comment on the relationships between quality at different plants.

Solution It is clear from the graphic that plant 2 needs to reduce its variability and that plants 2 and 4 need to improve their quality level. ■

We conclude this section with a warning. Sometimes it is a trend over time that is the most important feature of data. This feature would be lost entirely if the set

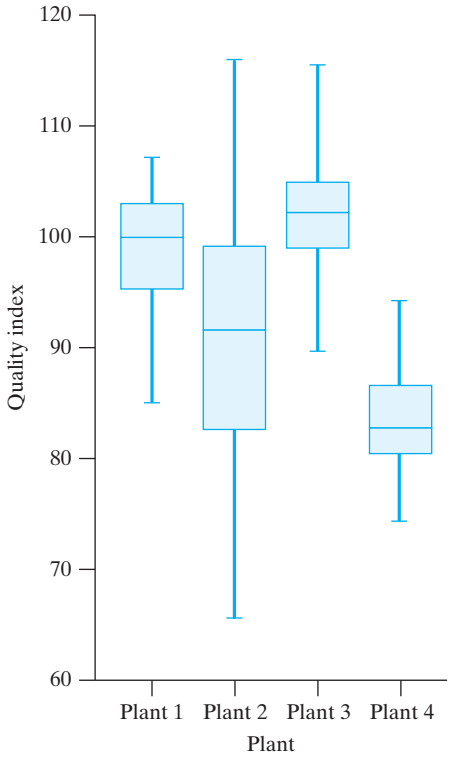


Figure 2.15
Boxplot of the quality index

of data were summarized in a dot diagram, stem-and-leaf display, or boxplot. In one instance, a major international company purchased two identical machines to rapidly measure the thickness of the material and test its strength. The machines were expensive but much faster than the current testing procedure. Before sending one across the United States and the other to Europe, engineers needed to confirm that the two machines were giving consistent results. Following one failed comparison, the problem machine was worked on for a couple of months by the engineers. In the second series of comparative trials, the average value from this machine was appropriate, but fortunately the individual values were plotted as in Figure 2.16. The

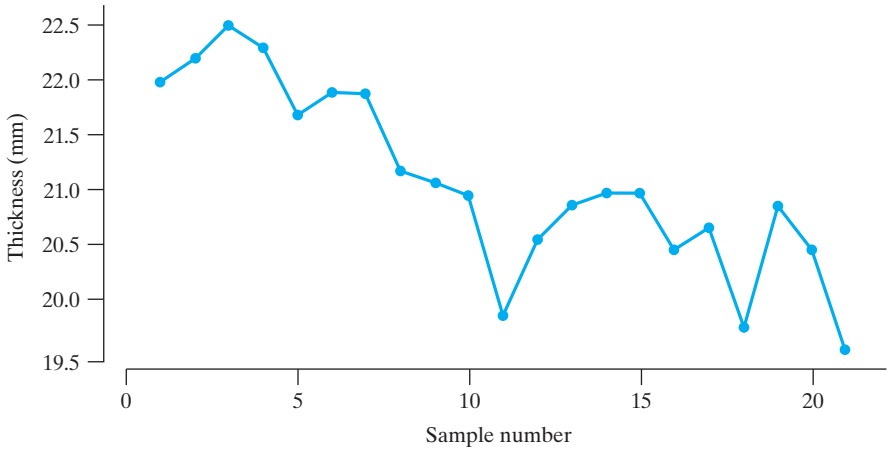


Figure 2.16
Machine measurement of thickness shows trend

time plot made it clear that the trend was the key feature, not the average, which was a poor summary. The testing machine required more work.

2.7 The Calculation of $\bar{x}$ and s

Here, we discuss methods for calculating $\bar{x}$ and s from data that are already grouped into intervals. These calculations are, in turn, based on the formulas for the mean and standard deviation for data consisting of all of the individual observations. In this latter case, we obtain $\bar{x}$ by summing all of the observations and dividing by the sample size n.

An alternative formula for s^2 forms the basis of the grouped data formula for variance. It was originally introduced to simplify hand calculations.

Variance (handheld calculator formula)

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}{n-1}$$

(In Exercise 2.51 you will be asked to show that this formula is, in fact, equivalent to the one on page 37.) This expression for variance is without $\bar{x}$, which reduces roundoff error when using a handheld calculator.

EXAMPLE 18 **Calculating variance using the handheld calculator formula**

Find the mean and the standard deviation of the following miles per gallon (mpg) obtained in 20 test runs performed on urban roads with an intermediate-size car:

19.7	21.5	22.5	22.2	22.6
21.9	20.5	19.3	19.9	21.7
22.8	23.2	21.4	20.8	19.4
22.0	23.0	21.1	20.9	21.3

Solution Using a calculator, we find that the sum of these figures is 427.7 and that the sum of their squares is 9,173.19. Consequently,

$$\bar{x} = \frac{427.7}{20} = 21.39 \text{ mpg}$$

and

$$s^2 = \frac{9{,}173.19 - (427.7)^2/20}{19} = 1.412$$

and it follows that $s = 1.19$ mpg. In computing the necessary sums we usually retain all decimal places, but at the end, as in this example, we usually round to one more decimal than we had in the original data. ■

See Exercise 2.58 for a computer calculation. This is the recommended procedure because it is easy to check the data entered for accuracy, and the calculation is free of human error. Most importantly, the calculation of variance can be done using the square of the deviations $x_i - \bar{x}$ rather than the squares of the observations x_i, and this is numerically more stable.

Historically, data were grouped to simplify the calculation of the mean and the standard deviation. Calculators and computers have eliminated the calculation

problem. Nevertheless, it is sometimes necessary to calculate $\bar{x}$ and s from grouped data since some data (for instance, from government publications) is available only in grouped form.

To calculate $\bar{x}$ and s from grouped data, we must assume something about the distribution of the values within each class. We represent each value within a class by the corresponding class mark. Then the sum of the x's and the sum of their squares can be written

$$\sum_{i=1}^{k} x_i f_i \quad \text{and} \quad \sum_{i=1}^{k} x_i^2 f_i$$

where x_i is the class mark of the ith class, f_i is the corresponding class frequency, and k is the number of classes in the distribution. Substituting these sums into the formula for $\bar{x}$ and the computing formula for s^2, we get

Mean and variance (grouped data)

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{k} x_i^2 f_i - \left(\sum_{i=1}^{k} x_i f_i \right)^2 / n}{n-1}$$

EXAMPLE 19 **Calculating a mean and variance from grouped data**

Use the distribution obtained on page 27 to calculate the mean, variance, and standard deviation of the nanopillar heights data.

Solution Recording the class marks and the class frequencies in the first two columns and the products $x_i f_i$ and $x_i^2 f_i$ in the third and fourth columns, we obtain

x_i	f_i	$x_i f_i$	$x_i^2 f_i$
225	3	675	151,875
265	11	2,915	772,475
305	23	7,015	2,139,575
345	9	3,105	1,071,225
385	4	1,540	592,900
Total	50	15,250	4,728,050

Then, substitution into the formula yields

$$\bar{x} = \frac{15{,}250}{50} = 305.0$$

and

$$s^2 = \frac{4{,}728{,}050 - 15{,}250^2/50}{49} = 1{,}567.35 \qquad \text{so} \qquad s = 39.6$$

For comparison, the original data have mean = 305.6 and standard deviation = 37.0. ■

Exercises

2.27 In each of the following situations, should your value be near the average or an outlier? If an outlier, should it be too large or too small?

(a) Income on your starting job

(b) Your score on the final exam in a physics class

(c) Your weight in 10 years

2.28 In each of the following situations, should your value be near the average or an outlier? If outlier, should it be too large or too small?

(a) The time you take to complete a lab assignment next week

(b) Your white blood cell count

2.29 Is the influence of a single outlier greater on the mean or the median? Explain.

2.30 Is the influence of a single outlier greater on the sample range or the interquartile range? Explain.

2.31 Referring to Exercise 1.8 in Chapter 1, we see that the sample of 4 deviations (observation − specification) during the second hour for a critical crank-bore diameter is

$$-6 \quad 1 \quad -4 \quad -3$$

ten-thousandths of an inch. For these 4 deviations

(a) calculate the sample mean $\bar{x}$

(b) calculate the sample standard deviation s

(c) On average, is the hole too large or too small?

2.32 At the end of 2012, nine skyscrapers in the world were over 300 meters tall. The ordered values of height are

366 381 442 452 484 492 508 601 828

The tallest is in Dubai.

(a) Calculate the sample mean

(b) Drop the largest value and re-calculate the mean.

(c) Comment on effect of dropping the single very large value.

2.33 Engineers[3] are developing a miniaturized robotic capsule for exploration of a human gastrointestinal tract. One novel solution uses motor-driven legs. The engineers' best design worked for a few trials, and then debris covered the tip of the leg and performance got worse. After cleaning, the next trial resulted in

35 37 38 34 30 24 13

distances covered (mm/min).

(a) Calculate the sample mean distance.

(b) Does the sample mean provide a good summary of these trials? If not, write a sentence or two to summarize more accurately.

2.34 A contract for the maintenance of a leading manufacturer's computers was given to a team of specialists. After six months, the supervisor of the team felt that computer performance could be improved by modifying the existing IC board. To document the current status, the team collected data on the number of IC board failures. Use the data below to:

(a) calculate the sample mean $\bar{x}$,

(b) calculate the sample standard deviation s.

Number of IC board failures:

12	3	8	6	19	1	2	5
1	11	14	3	13	2	9	8
2	1	4	13	3	11	9	15
14	5	12	7	6	16	10	0

2.35 If the mean annual compensation paid to the chief executives of three engineering firms is $175,000, can one of them receive $550,000?

2.36 Records show that the normal daily precipitation for each month in the Gobi desert, Asia is 1, 1, 2, 4, 7, 15, 29, 27, 10, 3, 2 and 1 mm. Verify that the mean of these figures is 8.5 and comment on the claim that the average daily precipitation is a very comfortable 8.5 mm.

2.37 The output of an instrument is often a waveform. With the goal of developing a numerical measure of closeness, scientists asked 11 experts to look at two waveforms on the same graph and give a number between 0 and 1 to quantify how well the two waveforms agree.[4] The agreement numbers for one pair of waveforms are

0.50 0.40 0.04 0.45 0.65 0.40 0.20 0.30 0.60 0.45

(a) Calculate the sample mean $\bar{x}$.

(b) Calculate sample standard deviation s.

[3]M. Quirini and S. Scapellato, Design and fabrication of a motor legged capsule for the active exploration of the gastrointestinal tract. *IEEE/ASME Transactions on Mechatronics* (2008) **13**, 169–179.

[4]L. Schwer, Validation metrics for response histories: Perspectives and case studies. *Engineering with Computers* **23** (2007), 295–309.

2.38 With reference to the preceding exercise, find s using

(a) the formula that defines s;

(b) the handheld calculator formula for s.

2.39 Meat products are regularly monitored for freshness. A trained inspector selects a sample of the product and assigns an offensive smell score between 1 and 7 where 1 is very fresh. The resulting offensive smell scores, for each of 16 samples, are (Courtesy of David Brauch)

3.2	3.9	1.7	5.0	1.9	2.6	2.4	5.3
1.0	2.7	3.8	5.2	1.0	6.3	3.3	4.3

(a) Find the mean.

(b) Find the median.

(c) Draw a boxplot.

2.40 With reference to Exercise 2.31, find s^2 using

(a) the formula that defines s^2;

(b) the handheld calculator formula for s^2.

2.41 The Aerokopter AK1-3 is an ultra-lightweight manned kit helicopter with a high rotor tip speed. A sample of 8 measurements of speed, in meters per second, yielded

204 208 205 211 207 201 201 203

Find the mean and quartiles for this sample.

2.42 For the five observations 8 2 10 6 9

(a) calculate the deviations $(x_i - \bar{x})$ and check that they add to 0.

(b) calculate the variance and the standard deviation.

2.43 With reference to Exercise 2.14 on page 34, draw a boxplot.

2.44 A factory experiencing a board-solder defect problem with an LED panel board product tested each board manufactured for LED failure. Data were collected on the area of the board on which LEDs were soldered for 8 bad panels and 8 good panels that passed the failure test.

Failure 32.5 34.5 33.5 36.5 34.0 32.25 33.75 35.25

(a) Calculate the sample mean $\bar{x}$.

(b) Calculate the sample standard deviation s.

2.45 Refer to Exercise 2.44. The measurements for the 8 panels that did not fail were

Good 33.5 32.25 34.75 34.25 35.5 33.0 36.25 35.75

(a) Calculate the sample mean $\bar{x}$.

(b) Calculate the sample standard deviation s.

(c) Does there appear to be a major difference in board area between panels in which LEDs failed and those in which LEDs did not?

2.46 Find the mean and the standard deviation of the 20 humidity readings on page 31 by using

(a) the raw (ungrouped) data

(b) the distribution obtained in that example

2.47 Use the distribution in Exercise 2.10 on page 32 to find the mean and the variance of the nanopillar diameters.

2.48 Use the distribution obtained in Exercise 2.12 on page 33 to find the mean and the standard deviation of the particle sizes. Also determine the coefficient of variation.

2.49 Use the distribution obtained in Exercise 2.14 on page 33 to find the coefficient of variation of the temperature data.

2.50 Show that

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

for any set of observations $x_1, x_2, \ldots, x_n$.

2.51 Show that the computing formula for s^2 on page 44 is equivalent to the one used to define s^2 on page 37.

2.52 If data are coded so that $x_i = c \cdot u_i + a$, show that $\bar{x} = c \cdot \bar{u} + a$ and $s_x = |c| \cdot s_u$.

2.53 **Median of grouped data** To find the *median* of a distribution obtained for n observations, we first determine the class into which the median must fall. Then, if there are j values in this class and k values below it, the median is located $\dfrac{(n/2) - k}{j}$ of the way into this class, and to obtain the median we multiply this fraction by the class interval and add the result to the lower boundary of the class into which the median must fall. This method is based on the assumption that the observations in each class are "spread uniformly" throughout the class interval, and this is why we count $\dfrac{n}{2}$ of the observations instead of $\dfrac{n+1}{2}$ as on page 35.

To illustrate, let us refer to the nanopillar height data on page 25 and the frequency distribution on page 26. Since $n = 50$, it can be seen that the median must fall in class (285, 325], which contains $j = 23$ observations. The class has width 40 and there are $k = 3 + 11 = 14$ values below it, so the median is

$$285 + \frac{25 - 14}{23} \times 40 = 264.13$$

(a) Use the distribution obtained in Exercise 2.10 on page 32 to find the median of the grouped nanopillar diameters.

(b) Use the distribution obtained in Exercise 2.12 on page 33 to find the median of the grouped particle sizes.

2.54 For each of the following distributions, decide whether it is possible to find the mean and whether it is possible to find the median. Explain your answers.

(a)

Grade	Frequency
40–49	5
50–59	18
60–69	27
70–79	15
80–89	6

(b)

IQ	Frequency
less than 90	3
90–99	14
100–109	22
110–119	19
more than 119	7

(c)

Weight	Frequency
110 or less	41
101–110	13
111–120	8
121–130	3
131–140	1

2.55 To find the first and third quartiles Q_1 and Q_3 for grouped data, we proceed as in Exercise 2.53, but count $\frac{n}{4}$ and $\frac{3n}{4}$ of the observations instead of $\frac{n}{2}$.

(a) With reference to the distribution of the nanopillar height data on page 25 and the frequency distribution on page 26, find Q_1, Q_3, and the interquartile range.

(b) Find Q_1 and Q_3 for the distribution of the particle size data obtained in Exercise 2.12.

2.56 If k sets of data consist, respectively, of $n_1, n_2, \ldots, n_k$ observations and have the means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, then the overall mean of all the data is given by the formula

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$$

(a) There are 15 students in semester I, 25 students in semester II and 16 students in semester III in an engineering program. If the average attendance of students is 82, 74, and 79 in semesters I, II and III respectively, what is the mean for the entire program?

(b) The average monthly expenses on repairs of machines in four factories are \$1,800, \$4,200, \$12,000 and \$800. If the number of machines in these factories is 12, 18, 42, and 8 respectively, find the average amount spent on repairs of these 80 machines.

2.57 The formula for the preceding exercise is a special case of the following formula for the **weighted mean**:

$$\bar{x}_w = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}$$

where w_i is a weight indicating the relative importance of the ith observation.

(a) If an instructor counts the final examination in a course four times as much as each 1-hour examination, what is the weighted average grade of a student who received grades of 69, 75, 56, and 72 in four 1-hour examinations and a final examination grade of 78?

(b) From 2010 to 2015, the cost of food in a certain city increased by 60%, the cost of housing increased by 30%, and the cost of transportation increased by 40%. If the average salaried worker spent 24% of his or her income on food, 33% on housing, and 15% on transportation, what is the combined percentage increase in the total cost of these items.

2.58 Modern computer software programs have come a long way toward removing the tedium of calculating statistics. *MINITAB* is one common and easy-to-use program. We illustrate the use of the computer using *MINITAB* commands. Other easy-to-use programs have a quite similar command structure.

The lumber used in the construction of buildings must be monitored for strength. Data for the strength of 2 × 4 pieces of lumber in pounds per square inch are in the file 2-58.TXT. We give the basic commands that calculate n, $\bar{x}$, and s as well as the quartiles.

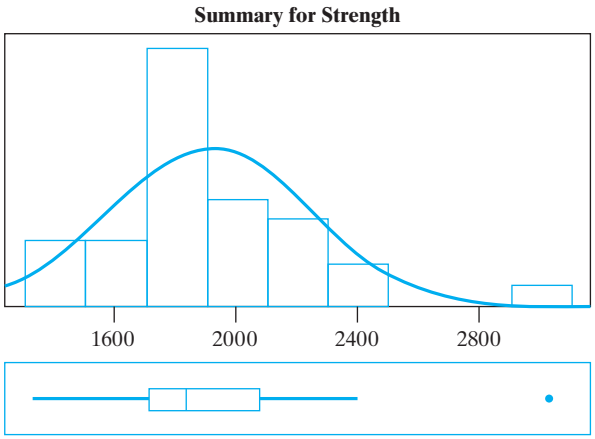
The session commands require the data to be set in the first column, C1, of the *MINITAB* work sheet. The command for creating a boxplot is also included.

Data in 2-58.TXT
strength
Dialog box:
Stat> **Basic Statistics** > **Descriptive Statistics**
Type *strength* in **Variables**.
Click **OK**.

Output (partial)

Variable	N	Mean	Median	StDev
Strength	30	1908.8	1863.0	327.1

Variable	Minimum	Maximum	Q1	Q3
Strength	1325.0	2983.0	1711.5	2071.8



Mean	1908.8
StDev	327.1
Variance	107004.3
Skewness	1.11841
Kurtosis	2.88335
N	30
Minimum	1325.0
1st Quartile	1711.5
Median	1863.0
3rd Quartile	2071.8
Maximum	2983.0

Figure 2.17
MINITAB 14 output

Use *MINITAB*, or some other statistical package, to find $\bar{x}$ and s for

(a) the decay times on page 156

(b) the interrequest times on page 29

2.59 (Further *MINITAB* calculation and graphs.) With the observations on the strength (in pounds per square inch) of 2×4 pieces of lumber already set in C1, the sequence of choices and clicks produces an even more complete summary (see Figure 2.17).

Stat> **Basic Statistics** > **Graphical Summary**
Type *strength* in **Variables**. Click **OK**.

The ordered strength data are

1325 1419 1490 1633 1645 1655 1710 1712 1725 1727 1745 1828 1840 1856 1859 1867 1889 1899 1943 1954 1976 2046 2061 2104 2168 2199 2276 2326 2403 2983

From the ordered data

(a) obtain the quartiles

(b) construct a histogram and locate the mean, median, Q_1, and Q_3 on the horizontal axes

(c) repeat parts (a) and (b) with the aluminum alloy data on page 29.

2.8 A Case Study: Problems with Aggregating Data

As circuit boards and other components move through a company's surface mount technology assembly line, a significant amount of data is collected for each assembly. The data (courtesy of Don Ermer) are recorded at several stages of manufacture in a serial tracking database by means of computer terminals located throughout the factory. The data include the board serial number, the type of defect, number of defects, and their location. The challenge here is to transform a large amount of data into manageable and useful information. When there is a variety of products and lots of data are collected on each, record management and the extraction of appropriate data for product improvement must be done well.

Originally, an attempt was made to understand this large database by *aggregating*, or grouping together, data from all products and performing an analysis of the data as if it were one product! This was a poor practice that decreased the resolution of the information obtained from the database. The products on the assembly line ranged in complexity, maturity, method of processing, and lot size.

To see the difficulties caused by aggregation, consider a typical week's production, where 100 printed circuit boards of Product A were produced, 40 boards of Product B, and 60 boards of Product C. Following a wave-soldering process, a total of 400 solder defects was reported. This translates to an overall average of $400/200 = 2$ defects per board. It was this company's practice to circulate the weekly aggregate average throughout the factory floor for review and comment.

It was then the operator's responsibility to take action according to the *misleading* report. Over time, it became apparent that this process was ineffective for improving quality.

However, further analysis of this data on a product-by-product basis revealed that products A, B, and C actually contributed 151, 231, and 18 defects. Thus, the number of defects per board was 1.51, 5.78, and 0.30 for products A, B, and C, respectively. Figure 2.18 correctly shows the average number of defects. Product C has a significantly lower defect rate and Product B has a significantly higher defect rate relative to the incorrect aggregated average. These latter are also the more complex boards.

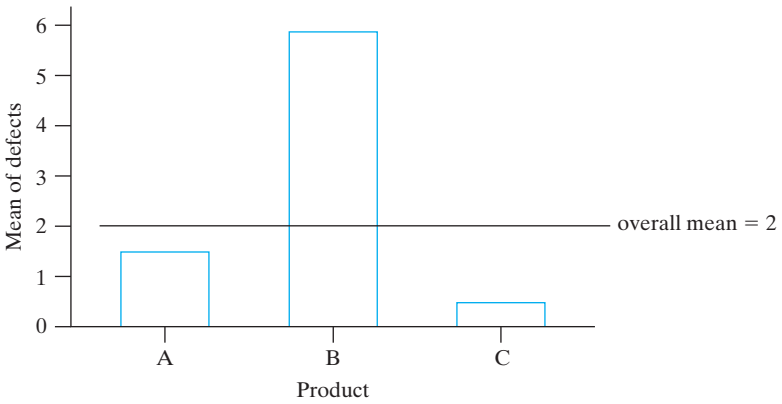


Figure 2.18
Average number of defects per product type

These data concern the number of defects that occurred when boards were wave-soldered after an assembly stage. The next step was to implement control charts for the number of defects for each of the three products. The numbers of defects for Product B were

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

The appropriate control chart is a time plot where the serial numbers of the product or sample are on the horizontal axis and the corresponding number of defects on the vertical axis. In this *C*-chart, the central line labeled $\overline{C}$ is the average number of defects over all cases in the plot. The dashed lines are the control limits set at three standard deviations about the central line. (For reasons explained in Section 15.6, we use $\sqrt{\overline{C}}$ rather than s when the data are numbers of defects.)

$$\text{LCL} = \overline{C} - 3\sqrt{\overline{C}}$$
$$\text{UCL} = \overline{C} + 3\sqrt{\overline{C}}$$

Figure 2.19(a) gives a *C*-chart constructed for Product B, but where the centerline is incorrectly calculated from the aggregated data is $\overline{C} = 2.0$. This is far too low and so is the upper control limit 6.24. The lower control limit is negative so we use 0. It looks like a great many of the observations are out of control because they exceed the upper control limit.

When the *C*-chart is correctly constructed on the basis of data from Product B alone, the centerline is $\overline{C} = 231/40 = 5.775$ and the upper control limit is 12.98. The lower control limit is again negative so we use 0. From Figure 2.19(b), the correct *C*-chart, the wave soldering process for Product B appears to be in control except for time 38 when 13 defects were observed.

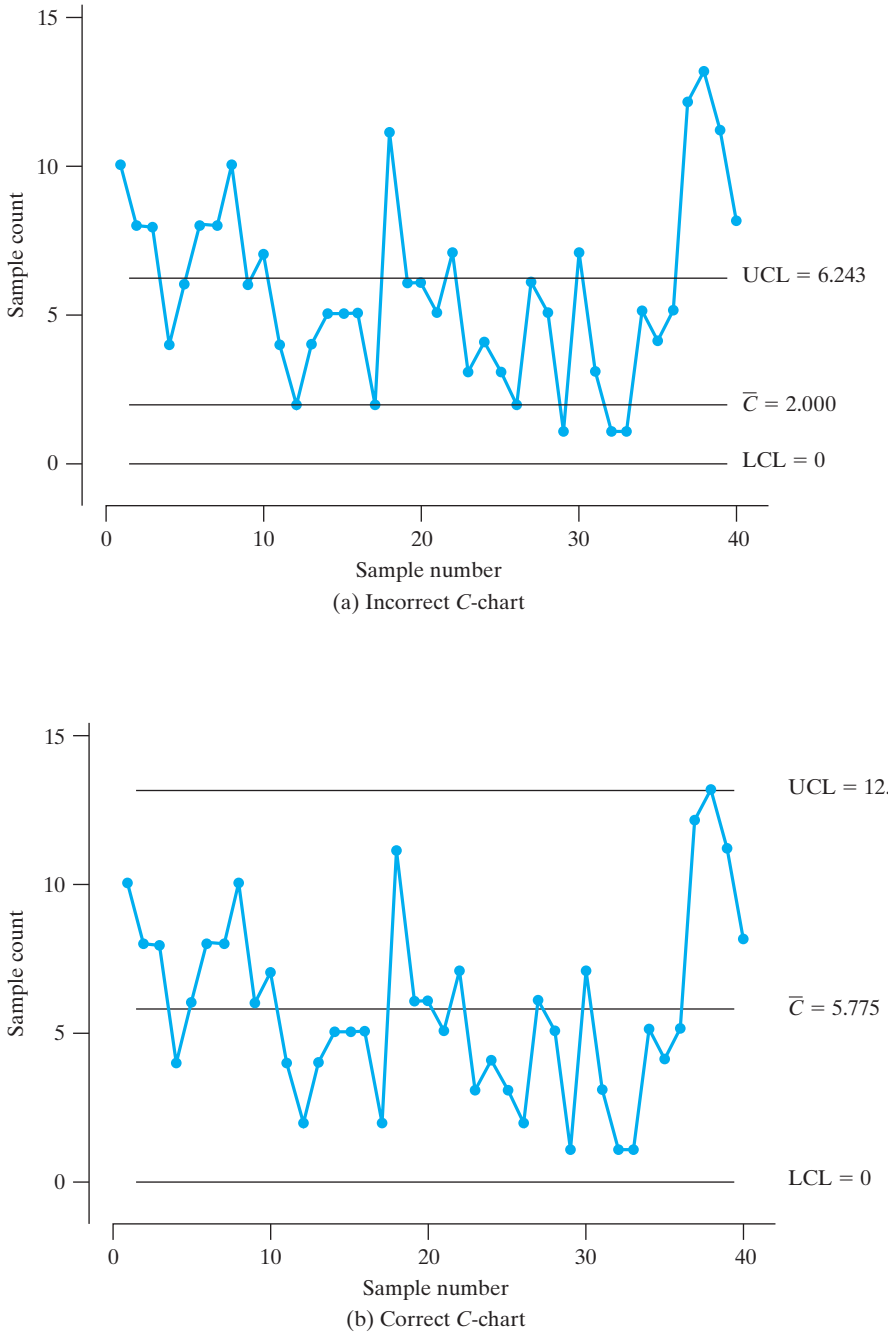


Figure 2.19
C-charts for defects

With the data segregated into products, separate charts were constructed for each of the three products. With this new outlook on data interpretation, a number of improvement opportunities surfaced that were previously disguised by aggregation. For example, by reducing the dimensions of an electrical pad, a significant reduction was achieved in the number of solder bridges between pins. This same design change was added to all of the board specifications and improvements were obtained on all products.

In summary, the aggregation of data from different products, or more generally from different sources, can lead to incorrect conclusions and mask opportunities for

quality improvement. Segregating data by product, although more time-consuming initially, can lead to significant reduction in waste and manufacturing costs.

Do's and Don'ts

Do's

1. Graph the data as a dot diagram or histogram to assess the overall pattern of data.
2. Calculate the summary statistics—sample mean, standard deviation, and quartiles—to describe the data set.

Don'ts

1. Don't routinely calculate summary statistics without identifying unusual observations which may have undue influence on the values of the statistics.

Review Exercises

2.60 From 1,500 wall clocks inspected by a manufacturer, the following defects were recorded.

hands touch each other	112
defective gears	16
faulty machinery	18
rotating pin	6
others	3

Create a Pareto chart.

2.61 Create

(a) a frequency table of the aluminum alloy strength data on page 29 using the classes [66.0, 67.5), [67.5, 69.0), [69.0, 70.5), [70.5, 72.0), [72.0, 73.5), [73.5, 75.0), [75.0, 76.5), where the right-hand endpoint is excluded

(b) a histogram using the frequency table in part (a)

2.62 Create

(a) a frequency table of the interrequest time data on page 29 using the intervals [0, 2,500), [2,500, 5,000), [5,000, 10,000), [10,000, 20,000), [20,000, 40,000), [40,000, 60,000), [60,000, 80,000), where the left-hand endpoint is included but the right-hand endpoint is not

(b) a histogram using the frequency table in part (a) (Note that the intervals are unequal, so make the height of the rectangle equal to relative frequency divided by width.)

2.63 Direct evidence of Newton's universal law of gravitation was provided from a renowned experiment by Henry Cavendish (1731–1810). In the experiment, masses of objects were determined by weighing, and the measured force of attraction was used to calculate the density of the earth. The values of the earth's density, in time order by row, are

5.36	5.29	5.58	5.65	5.57	5.53	5.62	5.29
5.44	5.34	5.79	5.10	5.27	5.39	5.42	5.47
5.63	5.34	5.46	5.30	5.75	5.68	5.85	

(*Source*: *Philosophical Transactions* 17 (1798); 469.)

(a) Find the mean and standard deviation.

(b) Find the median, Q_1, and Q_3.

(c) Plot the observations versus time order. Is there any obvious trend?

2.64 J. J. Thomson (1856–1940) discovered the electron by isolating negatively charged particles for which he could measure the mass/charge ratio. This ratio appeared to be constant over a wide range of experimental conditions and, consequently, could be a characteristic of a new particle. His observations, from two different cathode-ray tubes that used air as the gas, are

Tube 1	0.57	0.34	0.43	0.32	0.48	0.40	0.40
Tube 2	0.53	0.47	0.47	0.51	0.63	0.61	0.48

(*Source*: *Philosophical Magazine* 44; 5 (1897): 293.)

(a) Draw a dot diagram with solid dots for Tube 1 observations and circles for Tube 2 observations.

(b) Calculate the mean and standard deviation for the Tube 1 observations.

(c) Calculate the mean and standard deviation for the Tube 2 observations.

2.65 With reference to Exercise 2.64,

(a) calculate the median, maximum, minimum, and range for Tube 1 observations;

(b) calculate the median, maximum, minimum, and range for the Tube 2 observations.

2.66 A. A. Michelson (1852–1931) made many series of measurements of the speed of light. Using a revolving mirror technique, he obtained

12 30 30 27 30 39 18 27 48 24 18

for the differences

(velocity of light in air) − (229,700) km/s

(*Source: The Astrophysical Journal* 65 (1927): 11.)

(a) Create a dot diagram.

(b) Find the median and the mean. Locate both on the dot diagram.

(c) Find the variance and standard deviation.

2.67 With reference to Exercise 2.66,

(a) find the quartiles;

(b) find the minimum, maximum, range, and interquartile range;

(c) create a boxplot.

2.68 An electric engineer monitored the flow of current in a circuit by measuring the flow of electrons and the resistance of the medium. Over 11 hours, she observed a flow of

5 12 8 16 13 10 9 11 14 7 8

amperes.

(a) Create a dot diagram.

(b) Find the median and the mean. Locate both on the dot diagram.

(c) Find the variance and the standard deviation.

2.69 With reference to Exercise 2.68,

(a) find the quartiles;

(b) find the minimum, maximum, range, and interquartile range;

(c) construct a boxplot.

2.70 The weight (grams) of meat in a pizza product produced by a large meat processor is measured for a sample of $n = 20$ packages. The ordered values are (Courtesy of Dave Brauch)

16.12 16.77 16.87 16.91 16.96 16.99 17.02
17.19 17.20 17.26 17.36 17.39 17.39 17.62
17.63 17.76 17.85 17.86 17.91 19.00

(a) find the quartiles;

(b) find the minimum, maximum, range, and interquartile range;

(c) find the 10th percentile and 20th percentile.

2.71 With reference to Exercise 2.70, construct

(a) a boxplot.

(b) a modified boxplot.

2.72 With reference to the aluminum-alloy strength data in Example 7, make a stem-and-leaf display.

2.73 During the laying of gas pipelines, the depth of the pipeline (in mm) must be controlled. One service provider recorded depths of

418 428 431 420 412 425 423
433 417 420 410 431 429 425

(a) Find the sample mean.

(b) Find the sample standard deviation.

(c) Find the coefficient of variation.

(d) Measurements by another service provider have a sample mean of 425 and standard deviation of 6.36. Which provider's set of measurements is relatively more variable?

2.74 With reference to the lumber-strength data in Exercise 2.59, the statistical software package *SAS* produced the output in Figure 2.20. Using this output,

(a) identify the mean and standard deviation and compare these answers with the values given in Exercise 2.59.

(b) Create a boxplot.

The UNIVARIATE Procedure
Variable: Strength

Moments

N	30	Sum Weights	30
Mean	1908.76667	Sum Observations	57263
Std Deviation	327.115047	Variance	107004.254

Basic Statistical Measures

Location		Variability	
Mean	1908.767	Std Deviation	327.11505
Median	1863.000	Variance	107004
		Range	1658
		Interquartile Range	349.00000

Quantiles (Definition 5)

Level	Quantile
100% Max	2983.0
99%	2983.0
95%	2403.0
90%	2301.0
75% Q3	2061.0
50% Median	1863.0
25% Q1	1712.0
10%	1561.5
5%	1419.0
1%	1325.0
0% Min	1325.0

Figure 2.20 Selected SAS output to describe the lumber strength data from Exercise 2.59

2.75 An engineer was assigned the task of calculating the average time spent by vehicles waiting at traffic

signals. The signal timing (in seconds) would then be modified to reduce the pressure of traffic. The observations of average waiting time during the month of January are:

58	63	58	12	24	47
46	29	42	68	33	
43	37	39	52	35	
44	35	49	36	64	
53	28	55	27	53	
55	64	37	31	61	

(a) Obtain the quartiles.

(b) Obtain the 80th percentile.

(c) Construct a histogram.

2.76 The National Highway Traffic Safety Administration reported the relative speed (rounded to the nearest 5 mph) of automobiles involved in accidents one year. The percentages at different speeds were

20 mph or less	2.0%
25 or 30 mph	29.7%
35 or 40 mph	30.4%
45 or 50 mph	16.5%
55 mph	19.2%
60 or 65 mph	2.2%

(a) From these data, can we conclude that it is safe to drive at high speeds? Why or why not?

(b) Why do most accidents occur in the 35 or 40 mph and in the 25 or 30 mph ranges?

(c) Construct a density histogram using the endpoints 0, 22.5, 32.5, 42.5, 52.5, 57.5, 67.5 for the intervals.

2.77 Given a five-number summary,

minimum Q_1 Q_2 Q_3 maximum

is it possible to determine whether or not an outlier is present? Explain.

2.78 Given a stem-and-leaf display, is it possible to determine whether or not an outlier is present? Explain.

2.79 Traversing the same section of interstate highway on 11 different days, a driver recorded the number of cars pulled over by the highway patrol:

0 1 3 0 2 0 1 0 2 1 0

(a) Create a dot plot.

(b) There is a long tail to the right. You might expect the sample mean to be larger than the median. Calculate the sample mean and median and compare the two measures of center. Comment.

2.80 An experimental study of the atomization characteristics of biodiesel fuel[5] was aimed at reducing the pollution produced by diesel engines. Biodiesel fuel is recyclable and has low emission characteristics. One aspect of the study is the droplet size (μm) injected into the engine, at a fixed distance from the nozzle. From data provided by the authors on droplet size, we consider a sample of size 41 that has already been ordered.

2.1	2.2	2.2	2.3	2.3	2.4	2.5	2.5	2.5
2.8	2.9	2.9	2.9	3.0	3.1	3.1	3.2	3.3
3.3	3.3	3.4	3.5	3.6	3.6	3.6	3.7	3.7
4.0	4.2	4.5	4.9	5.1	5.2	5.3	5.7	6.0
6.1	7.1	7.8	7.9	8.9				

(a) Group these droplet sizes and obtain a frequency table using [2, 3), [3, 4), [4, 5) as the first three classes, but try larger classes for the other cases. Here the left-hand endpoint is included but the right-hand endpoint is not.

(b) Construct a density histogram.

(c) Obtain $\bar{x}$ and s^2.

(d) Obtain the quartiles.

[5]H. Kim, H. Suh, S. Park, and C. Lee, An experimental and numerical investigation of atomization characteristics of biodiesel, dimethyl ether, and biodiesel-ethanol blended fuel, *Energy and Fuels*, **22** (2008), 2091–2098.

Key Terms

Ogive 30
Outlier 24
Pareto diagram 22
Percentile 39
Pie chart 33
Quartile 40
Range 41
Relative variation 38
Sample mean 35
Sample median 35
Sample variance 37
Standard deviation 38
Stem 31
Stem-and-leaf display 31
Stem label 31
Variance 37
Weighted mean 48