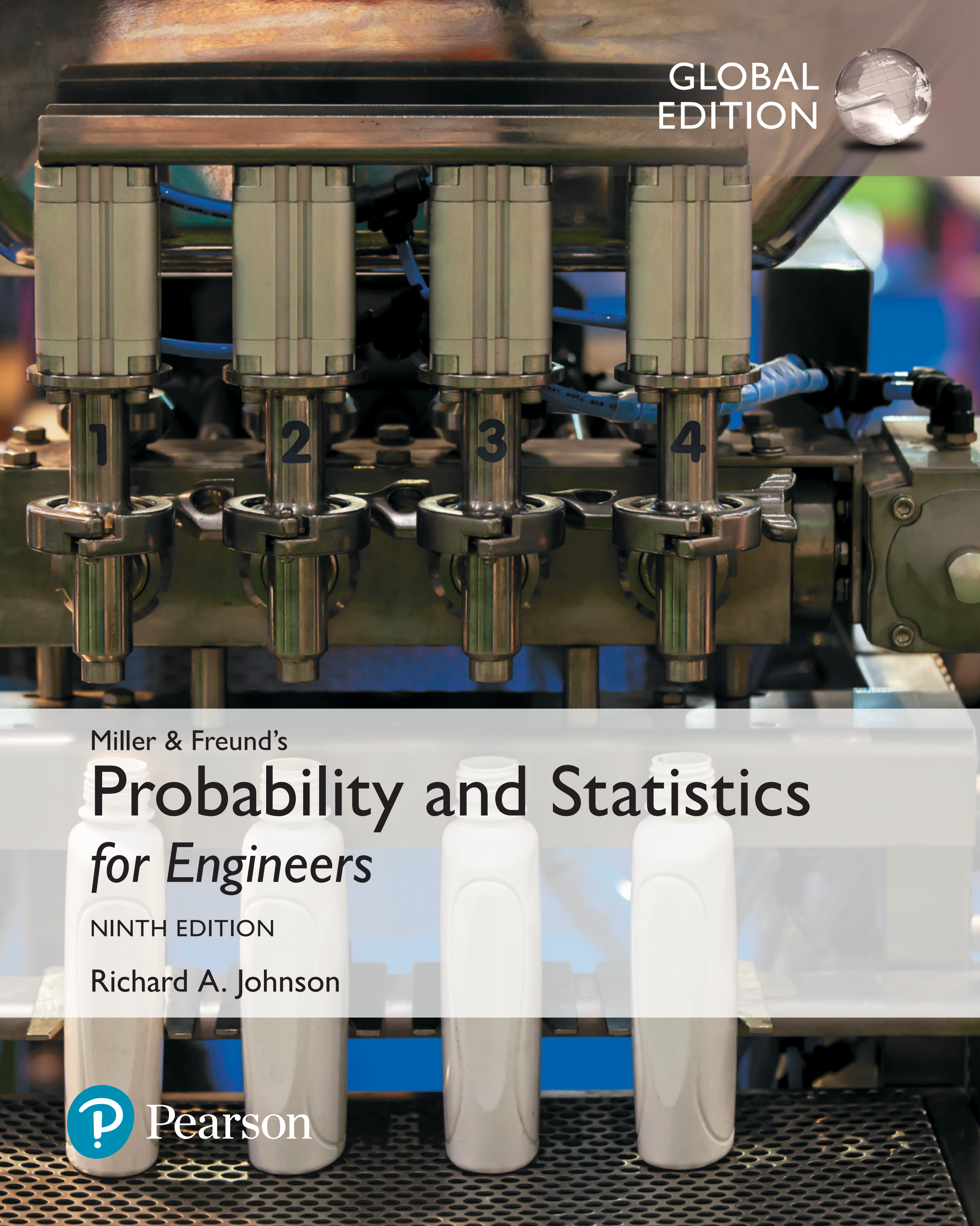


GLOBAL EDITION

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

For permission to use copyrighted material, grateful acknowledgement is made to these copyright holders: Screenshots from Minitab. Courtesy of Minitab Corporation. SAS Output Created with SAS® software. Copyright © 2013, SAS Institute Inc., Cary, NC, USA. All rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.

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

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

Printed and bound in Malaysia.

Contents

Preface 7

Chapter 1 Introduction 11

1.1 Why Study Statistics? 11
1.2 Modern Statistics 12
1.3 Statistics and Engineering 12
1.4 The Role of the Scientist and Engineer in Quality Improvement 13
1.5 A Case Study: Visually Inspecting Data to Improve Product Quality 13
1.6 Two Basic Concepts—Population and Sample 15
Review Exercises 20
Key Terms 21

Chapter 2 Organization and Description of Data 22

2.1 Pareto Diagrams and Dot Diagrams 22
2.2 Frequency Distributions 24
2.3 Graphs of Frequency Distributions 27
2.4 Stem-and-Leaf Displays 31
2.5 Descriptive Measures 34
2.6 Quartiles and Percentiles 39
2.7 The Calculation of $\overline{x}$ and s 44
2.8 A Case Study: Problems with Aggregating Data 49
Review Exercises 52
Key Terms 54

Chapter 3 Probability 56

3.1 Sample Spaces and Events 56
3.2 Counting 60
3.3 Probability 67
3.4 The Axioms of Probability 69
3.5 Some Elementary Theorems 72
3.6 Conditional Probability 78
3.7 Bayes' Theorem 84
Review Exercises 91
Key Terms 93

Chapter 4 Probability Distributions 94

4.1 Random Variables 94
4.2 The Binomial Distribution 98
4.3 The Hypergeometric Distribution 103
4.4 The Mean and the Variance of a Probability Distribution 107
4.5 Chebyshev's Theorem 114
4.6 The Poisson Distribution and Rare Events 118
4.7 Poisson Processes 122
4.8 The Geometric and Negative Binomial Distribution 124
4.9 The Multinomial Distribution 127
4.10 Simulation 128
Review Exercises 132
Key Terms 133

Chapter 5 Probability Densities 134

5.1 Continuous Random Variables 134
5.2 The Normal Distribution 140
5.3 The Normal Approximation to the Binomial Distribution 148
5.4 Other Probability Densities 151
5.5 The Uniform Distribution 151
5.6 The Log-Normal Distribution 152
5.7 The Gamma Distribution 155
5.8 The Beta Distribution 157
5.9 The Weibull Distribution 158
5.10 Joint Distributions—Discrete and Continuous 161
5.11 Moment Generating Functions 174
5.12 Checking If the Data Are Normal 180
5.13 Transforming Observations to Near Normality 182
5.14 Simulation 184
Review Exercises 188
Key Terms 190

Chapter 6 Sampling Distributions 193

6.1 Populations and Samples 193
6.2 The Sampling Distribution of the Mean (σ known) 197
6.3 The Sampling Distribution of the Mean (σ unknown) 205
6.4 The Sampling Distribution of the Variance 207
6.5 Representations of the Normal Theory Distributions 210
6.6 The Moment Generating Function Method to Obtain Distributions 213
6.7 Transformation Methods to Obtain Distributions 215
Review Exercises 221
Key Terms 222

Chapter 7 Inferences Concerning a Mean 223

7.1 Statistical Approaches to Making Generalizations 223
7.2 Point Estimation 224
7.3 Interval Estimation 229
7.4 Maximum Likelihood Estimation 236
7.5 Tests of Hypotheses 242
7.6 Null Hypotheses and Tests of Hypotheses 244
7.7 Hypotheses Concerning One Mean 249
7.8 The Relation between Tests and Confidence Intervals 256
7.9 Power, Sample Size, and Operating Characteristic Curves 257
Review Exercises 263
Key Terms 265

Chapter 8 Comparing Two Treatments 266

8.1 Experimental Designs for Comparing Two Treatments 266
8.2 Comparisons—Two Independent Large Samples 267
8.3 Comparisons—Two Independent Small Samples 272
8.4 Matched Pairs Comparisons 280
8.5 Design Issues—Randomization and Pairing 285
Review Exercises 287
Key Terms 288

Chapter 9 Inferences Concerning Variances 290

9.1 The Estimation of Variances 290
9.2 Hypotheses Concerning One Variance 293
9.3 Hypotheses Concerning Two Variances 295
Review Exercises 299
Key Terms 310

Chapter 10 Inferences Concerning Proportions 301

10.1 Estimation of Proportions 301
10.2 Hypotheses Concerning One Proportion 308
10.3 Hypotheses Concerning Several Proportions 310
10.4 Analysis of $r \times c$ Tables 318
10.5 Goodness of Fit 322
Review Exercises 325
Key Terms 326

Chapter 11 Regression Analysis 327

11.1 The Method of Least Squares 327
11.2 Inferences Based on the Least Squares Estimators 336
11.3 Curvilinear Regression 350
11.4 Multiple Regression 356
11.5 Checking the Adequacy of the Model 361
11.6 Correlation 366
11.7 Multiple Linear Regression (Matrix Notation) 377
Review Exercises 382
Key Terms 385

Chapter 12 Analysis of Variance 386

12.1 Some General Principles 386
12.2 Completely Randomized Designs 389
12.3 Randomized-Block Designs 402
12.4 Multiple Comparisons 410
12.5 Analysis of Covariance 415
Review Exercises 422
Key Terms 424

Chapter 13 Factorial Experimentation 425

13.1 Two-Factor Experiments 425
13.2 Multifactor Experiments 432
13.3 The Graphic Presentation of 2^2 and 2^3 Experiments 441
13.4 Response Surface Analysis 456
Review Exercises 459
Key Terms 463

Chapter 14 Nonparametric Tests 464

14.1 Introduction 464
14.2 The Sign Test 464
14.3 Rank-Sum Tests 466
14.4 Correlation Based on Ranks 469
14.5 Tests of Randomness 472
14.6 The Kolmogorov-Smirnov and Anderson-Darling Tests 475
Review Exercises 478
Key Terms 479

Chapter 15 The Statistical Content of Quality-Improvement Programs 480

15.1 Quality-Improvement Programs 480
15.2 Starting a Quality-Improvement Program 482
15.3 Experimental Designs for Quality 484
15.4 Quality Control 486
15.5 Control Charts for Measurements 488
15.6 Control Charts for Attributes 493
15.7 Tolerance Limits 499
Review Exercises 501
Key Terms 503

Chapter 16 Application to Reliability and Life Testing 504

16.1 Reliability 504
16.2 Failure-Time Distribution 506
16.3 The Exponential Model in Life Testing 510
16.4 The Weibull Model in Life Testing 513
Review Exercises 518
Key Terms 519

Appendix A Bibliography 521
Appendix B Statistical Tables 522
Appendix C Using the R Software Program 529
Introduction to R 529
Entering Data 529
Arithmetic Operations 530
Descriptive Statistics 530
Probability Distributions 531
Normal Probability Calculations 531
Sampling Distributions 531
Confidence Intervals and Tests of Means 532
Inference about Proportions 532
Regression 532
One-Way Analysis of Variance (ANOVA) 533
Appendix D Answers to Odd-Numbered Exercises 534
Index 541

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.		A first semester introduction that develops the tools of probability and some statistical inferences.	
Chapter 1	especially 1.6	Chapter 1	especially 1.6
Chapter 2		Chapter 2	
Chapter 3		Chapter 3	
Chapter 4	4.4–4.7	Chapter 4	4.4–4.7 4.8 (geometric, negative binomial)
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 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	Chapter 6	6.1–6.4 6.5–6.7 (Representations, mgf's, transformation)
Chapter 7	7.1–7.7	Chapter 7	7.1–7.7
Chapter 8		Chapter 8	
Chapter 9	(could skip)	Chapter 9	(could skip)
Chapter 10	10.1–10.4	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		

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

CHAPTER OUTLINE

1.1 Why Study Statistics? 11

1.2 Modern Statistics 12

1.3 Statistics and Engineering 12

1.4 The Role of the Scientist and Engineer in Quality Improvement 13

1.5 A Case Study: Visually Inspecting Data to Improve Product Quality 13

1.6 Two Basic Concepts—Population and Sample 15

Review Exercises 20

Key Terms 21

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.

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

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

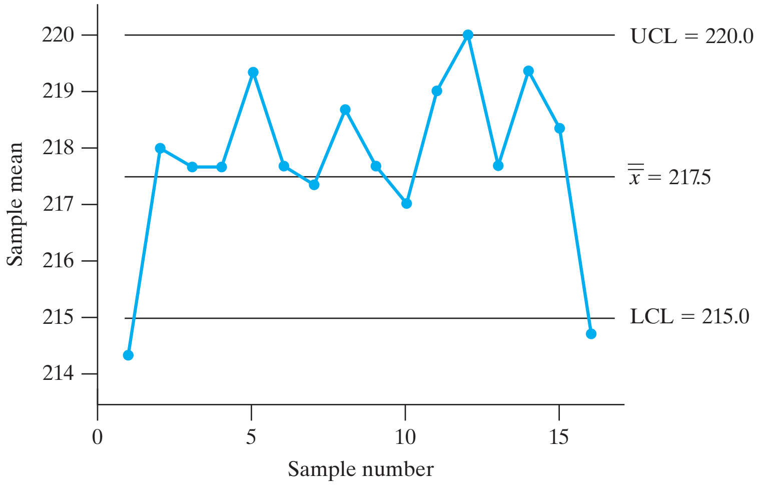


Figure 1.1
X-bar chart for depth

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 $\overline{x}$.

(b) Does the new $\overline{x}$ 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 $\overline{\overline{x}} = (4.25 - 3.00 - \cdots - 1.50 + 3.25)/24$.

(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	13	14	15	16	17	18	19	20	21	22	23	24
	10	−6	−1	−8	−14	−6	−1	8	−1	5	2	5	5	6	−5	−8	2	7	8	5	8	−5	−2	−1
	3	1	−3	−3	−5	−2	−6	−3	7	6	1	3	9	6	4	−5	8	7	13	4	1	7	−4	5
	6	−4	0	−7	−6	−1	−1	9	1	3	1	10	9	8	−5	1	−4	5	6	7	0	1	−7	9
	−2	−3	−7	−2	2	−6	7	11	7	2	4	4	7	10	−2	0	1	3	6	10	−6	2	7	0
$\bar{x}$	4.25	−3.00	−2.75	−5.00	−5.75	−3.75	−0.25	6.25	3.50	4.00	2.00	5.50	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

2.1 Pareto Diagrams and Dot Diagrams 22

2.2 Frequency Distributions 24

2.3 Graphs of Frequency Distributions 27

2.4 Stem-and-Leaf Displays 31

2.5 Descriptive Measures 34

2.6 Quartiles and Percentiles 39

2.7 The Calculation of $\overline{x}$ and s 44

2.8 A Case Study: Problems with Aggregating Data 49

Review Exercises 52

Key Terms 54

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

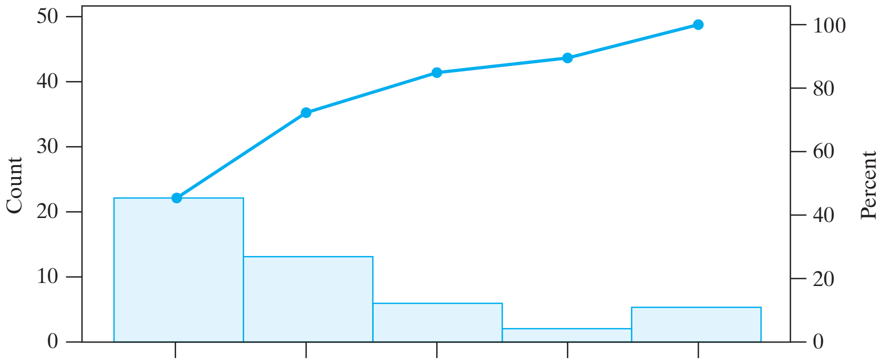
Data need to be collected to provide the vital information necessary to solve engineering problems. Once gathered, these data must be described and analyzed to produce summary information. Graphical presentations can often be the most effective way to communicate this information. To illustrate the power 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.

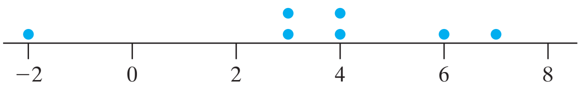


Figure 2.2
Dot diagram of cutting speed deviations

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

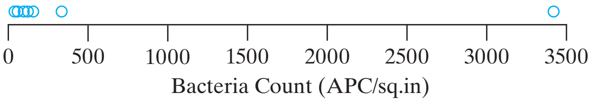


Figure 2.3
Maximum bacteria counts on seven days.

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. ■

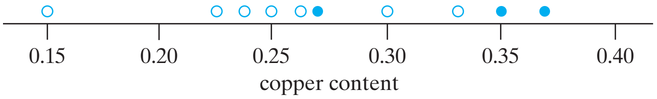


Figure 2.4
Dot diagram of copper content

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.

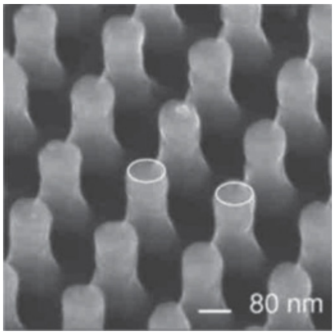


Figure 2.5
Nanopillars

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

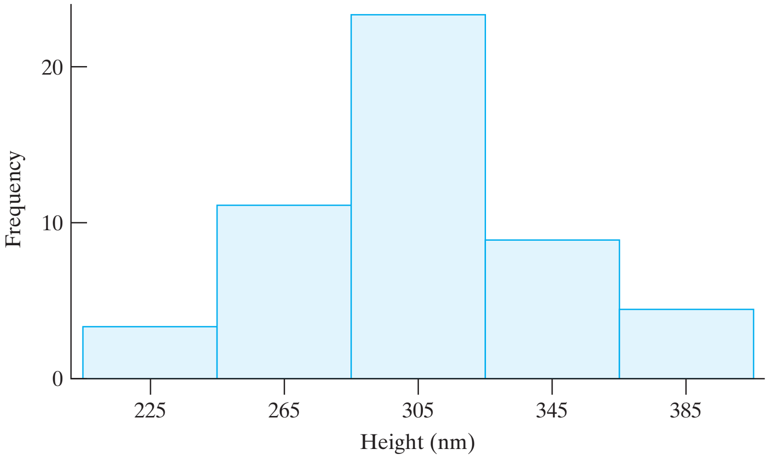


Figure 2.6
Histogram of pillar height

(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. ■

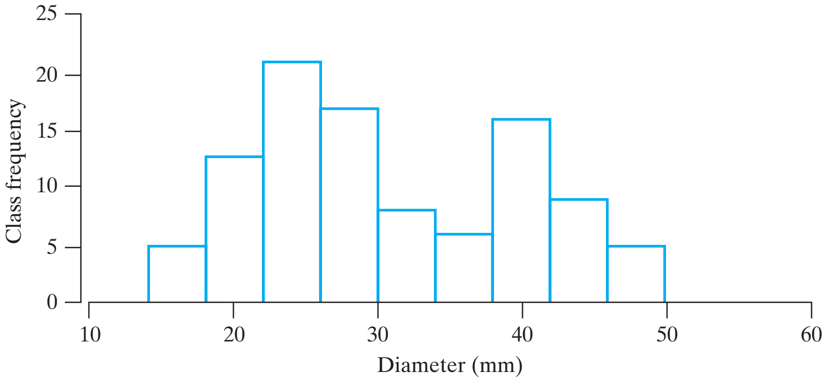


Figure 2.7
Histogram of pellet diameter

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). ■

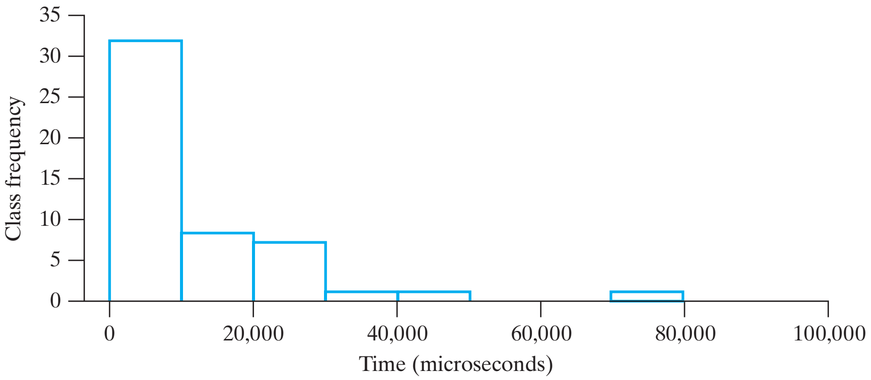


Figure 2.8
Histogram of interrequest time

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								