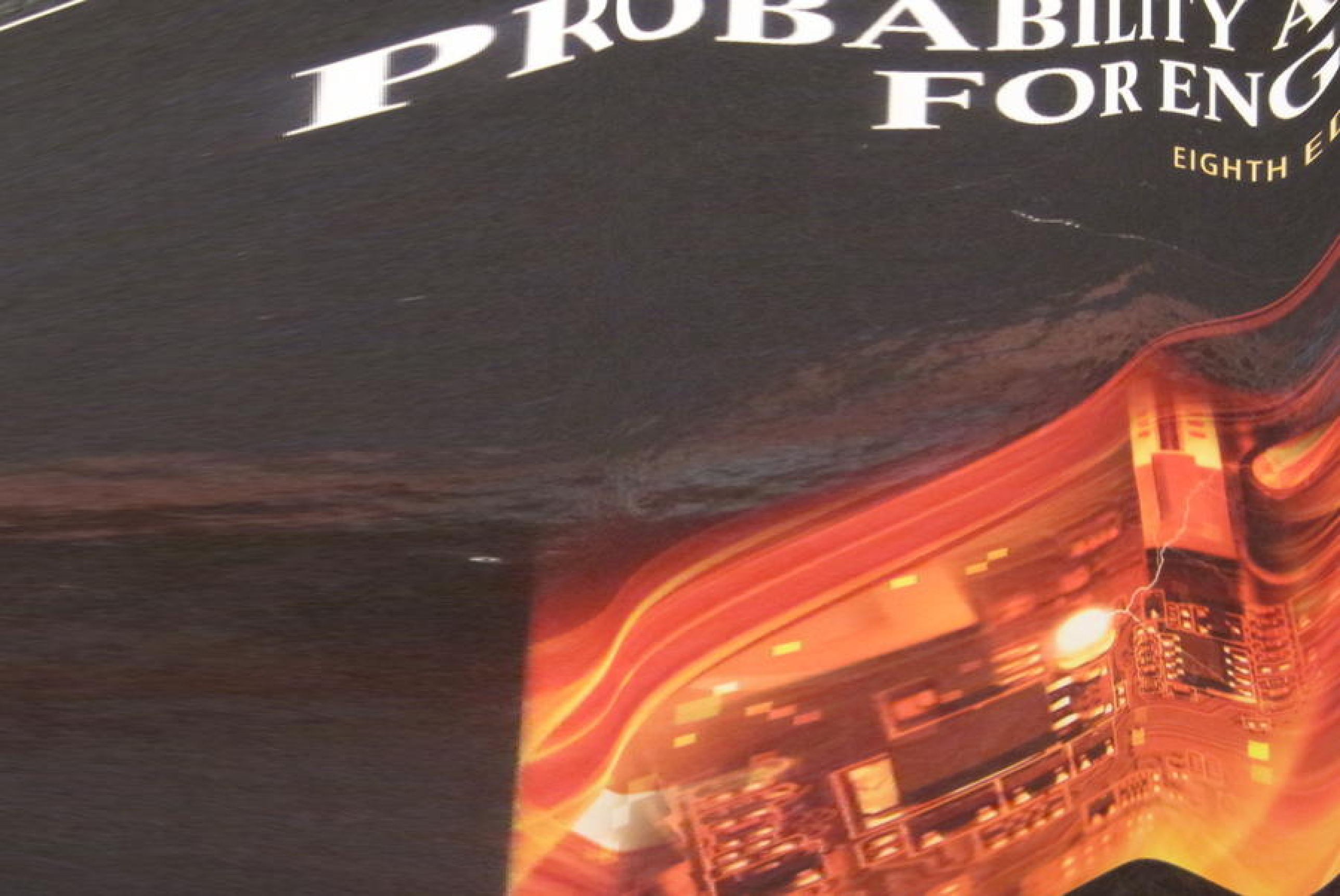
PROBABILITY A
FOR ENG

EIGHTH E

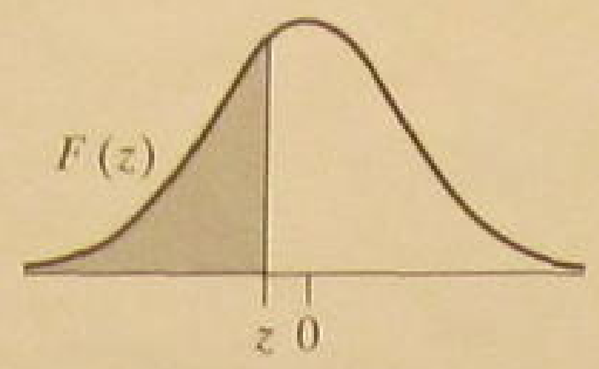


ND STATS
INEERS

DITION

Standard Normal Distribution Function

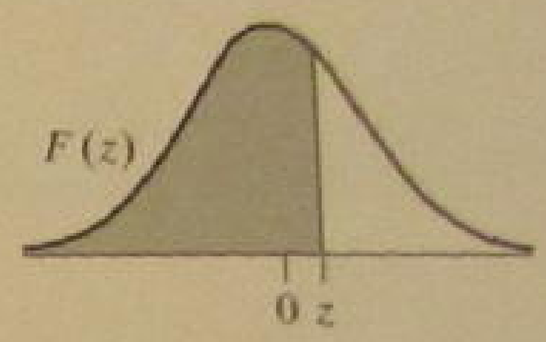
$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$



0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0000003									
0.00003									
0.0002									
0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0006	0.0003
0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Standard Normal Distribution Function

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5973	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998									
4.0	0.99997									
5.0	0.9999997									

n Chief: *Deirdre Lynch*
tions Editor: *Christopher Cummings*
t Editor: *Christina Lepre*
l Assistant: *Dana Jones*
Managing Editor: *Karen Wernholm*
e Managing Editor: *Bayani DeLeon*
Production Supervisor: *Tracy Patruno*
Manager: *Barbara Mack*
ctor: *Heather Scott*
Assets Manager: *Marianne Groth*
ng Manager: *Alex Gay*
ng Assistant: *Kathleen DeChavez*
nd Permissions Advisor: *Michael Joyce*
Manufacturing Buyer: *Carol Melville*
l Cover Design: *Tamara Newnam*
on Coordination, Composition, and Illustrations: *MPS Limited, A Macmillan Company*

hoto: *Shutterstock*

f the designations used by manufacturers and sellers to distinguish their products are claimed
marks. Where those designations appear in this book, and Pearson was aware of a trademark
he designations have been printed in initial caps or all caps.

y of Congress Cataloging-in-Publication Data

n, Richard Arnold.
er & Freund's probability and statistics for engineers / Richard A. Johnson,
Miller, John Freund. — 8th ed.
cm.
ides bibliographical references and index.
N 0-321-64077-2
ngineering—Statistical methods. 2. Probabilities. I. Miller, Irwin, 1928- II.
, John E. III. Title. IV. Title: Probability and statistics for engineers.
40.J66 2011
202'462—dc22 2009023072

ght © 2011, 2005, 2000 Pearson Education, Inc. All rights reserved. No part of this publication
e reproduced, stored in a retrieval system, or transmitted, in any form or by any means,
nic, mechanical, photocopying, recording, or otherwise, without the prior written permission
publisher. Printed in the United States of America. For information on obtaining permission
of material in this work, please submit a written request to Pearson Education, Inc.,
and Contracts Department, 501 Boylston Street, Suite 900, Boston, MA 02116, fax your
t to 617-671-3447, or e-mail at http://www.pearsoned.com/legal/permissions.htm.

4 5 6 7 8 9 10—EB—13 12 11 10 09

Prentice Hall
an imprint of

PEARSON

www.pearsonhighered.com

ISBN 10: 0-321-64077-2
ISBN 13: 978-0-321-64077-2

CONTENTS

Preface vii

Chapter 1 Introduction 1

1.1 Why Study Statistics? 1
1.2 Modern Statistics 2
1.3 Statistics and Engineering 2
1.4 The Role of the Scientist and Engineer in Quality Improvement 3
1.5 A Case Study: Visually Inspecting Data to Improve Product Quality 3
1.6 Two Basic Concepts—Population and Sample 5
Review Exercises 10
Key Terms 11

Chapter 2 Organization and Description of Data 12

2.1 Pareto Diagrams and Dot Diagrams 12
2.2 Frequency Distributions 14
2.3 Graphs of Frequency Distributions 17
2.4 Stem-and-Leaf Displays 21
2.5 Descriptive Measures 24
2.6 Quartiles and Percentiles 29
2.7 The Calculation of $\bar{x}$ and s 34
2.8 A Case Study: Problems with Aggregating Data 39
Review Exercises 42
Key Terms 44

Chapter 3 Probability 45

3.1 Sample Spaces and Events 45
3.2 Counting 49
3.3 Probability 55
3.4 The Axioms of Probability 57
3.5 Some Elementary Theorems 60
3.6 Conditional Probability 66
3.7 Bayes' Theorem 72
Review Exercises 78
Key Terms 80

Chapter 4 Probability Distributions 81

4.1 Random Variables 81
4.2 The Binomial Distribution 85
4.3 The Hypergeometric Distribution 89
4.4 The Mean and the Variance of a Probability Distribution 93
4.5 Chebyshev's Theorem 100
4.6 The Poisson Approximation to the Binomial Distribution 104
4.7 Poisson Processes 107
4.8 The Geometric and Negative Binomial Distribution 109
4.9 The Multinomial Distribution 111
4.10 Simulation 113
Review Exercises 116
Key Terms 118

ter 5 Probability Densities 119

Continuous Random Variables 119
The Normal Distribution 125
The Normal Approximation to the Binomial Distribution 132
Other Probability Densities 135
The Uniform Distribution 135
The Log-Normal Distribution 136
The Gamma Distribution 138
The Beta Distribution 141
The Weibull Distribution 142
5.10 Joint Distributions—Discrete and Continuous 145
5.11 Moment Generating Functions 157
5.12 Checking If the Data Are Normal 163
5.13 Transforming Observations to Near Normality 165
5.14 Simulation 167
Review Exercises 171
Key Terms 172

er 6 Sampling Distributions 175

opulations and Samples 175
he Sampling Distribution of the Mean r known) 179
he Sampling Distribution of the Mean r unknown) 187
he Sampling Distribution of the ariance 189
6.5 Representations of the Normal Theory Distributions 192
6.6 The Moment Generating Function Method to Obtain Distributions 194
6.7 Transformation Methods to Obtain Distributions 197
Review Exercises 202
Key Terms 203

er 7 Inferences Concerning a Mean 204

oint Estimation 204
terval Estimation 209
laximum Likelihood Estimation 215
ests of Hypotheses 222
ull Hypotheses and Tests of ypotheses 224
ypotheses Concerning One Mean 229
7.7 The Relation between Tests and Confidence Intervals 235
7.8 Power, Sample Size, and Operating Characteristic Curves 237
Review Exercises 243
Key Terms 244

er 8 Comparing Two Treatments 245

xperimental Designs for Comparing Two reatments 245
omparisons—Two Independent Large amples 246
omparisons—Two Independent Small amples 251
8.4 Matched Pairs Comparisons 259
8.5 Design Issues—Randomization and Pairing 263
Review Exercises 266
Key Terms 266

Chapter 9 Inferences Concerning Variances 268

9.1 The Estimation of Variances 268
9.2 Hypotheses Concerning One Variance 271
9.3 Hypotheses Concerning Two Variances 272
Review Exercises 276
Key Terms 277

Chapter 10 Inferences Concerning Proportions 278

10.1 Estimation of Proportions 278
10.2 Hypotheses Concerning One Proportion 283
10.3 Hypotheses Concerning Several Proportions 285
10.4 Analysis of $r \times c$ Tables 292
10.5 Goodness of Fit 295
Review Exercises 299
Key Terms 300

Chapter 11 Regression Analysis 301

11.1 The Method of Least Squares 301
11.2 Inferences Based on the Least Squares Estimators 309
11.3 Curvilinear Regression 321
11.4 Multiple Regression 326
11.5 Checking the Adequacy of the Model 329
11.6 Correlation 335
11.7 Multiple Linear Regression (Matrix Notation) 345
Review Exercises 350
Key Terms 353

Chapter 12 Analysis of Variance 354

12.1 Some General Principles 354
12.2 Completely Randomized Designs 357
12.3 Randomized-Block Designs 370
12.4 Multiple Comparisons 377
12.5 Analysis of Covariance 382
Review Exercises 388
Key Terms 390

Chapter 13 Factorial Experimentation 391

13.1 Two-Factor Experiments 391
13.2 Multifactor Experiments 398
13.3 2^n Factorial Experiments 409
13.4 The Graphic Presentation of 2^2 and 2^3 Experiments 416
13.5 Response Surface Analysis 430
13.6 Confounding in a 2^n Factorial Experiment 433
13.7 Fractional Replication 437
Review Exercises 442
Key Terms 446

apter 14 Nonparametric Tests 447

.1 Introduction 447
.2 The Sign Test 447
.3 Rank-Sum Tests 449
.4 Correlation Based on Ranks 453
.5 Tests of Randomness 455
14.6 The Kolmogorov-Smirnov and Anderson-Darling Tests 458
Review Exercises 461
Key Terms 462

apter 15 The Statistical Content of Quality-Improvement Programs 463

1 Quality-Improvement Programs 463
2 Starting a Quality-Improvement Program 465
3 Experimental Designs for Quality 467
4 Quality Control 469
5 Control Charts for Measurements 471
15.6 Control Charts for Attributes 476
15.7 Tolerance Limits 482
Review Exercises 484
Key Terms 486

apter 16 Application to Reliability and Life Testing 487

1 Reliability 487
2 Failure-Time Distribution 489
3 The Exponential Model in Life Testing 493
16.4 The Weibull Model in Life Testing 496
Review Exercises 501
Key Terms 502

endix A Bibliography 503
endix B Statistical Tables 504
endix C Using the R Software Program 528
Introduction to R 528
Entering Data 528
Arithmetic Operations 529
Descriptive Statistics 529
Probability Distributions 530
Normal Probability Calculations 530
Sampling Distributions 530
Confidence Intervals and Tests of Means 531
Inference about Proportions 531
Regression 531
One-Way Analysis of Variance (ANOVA) 532
pendix D Answers to Odd-Numbered Exercises 533
Index 540

PREFACE

This book is designed for an introductory course in probability and stat for students of engineering and the physical sciences. Each chapter b with an introductory statement and has a checklist of key terms with numbers. Each chapter concludes with a set of statistical guidelines for cor applying statistical procedures, and avoiding common pitfalls, under th *Do's and Don'ts*. Important formulas, theorems, and rules are set out from th in boxes. Confidence intervals are stressed as the major procedure for m inferences. Their properties are carefully described and their interpretation quently discussed. The steps for hypothesis testing are clearly and consi delineated in each application. The examples of hypotheses testing also str interpretation and calculation of a P value.

The text has been tested extensively in courses for university students as by in-plant training of engineers. The whole book can be covered in semester or three-quarter course consisting of three lectures a week. The bo makes an excellent basis for a one-semester course where the lecturer can topics to emphasize applications with varying amounts of theory. The auth ers most of the first eight chapters, straight-line regression, and the graphic tation of factorial designs in one semester.

The mathematical background expected of the reader is a year course i lus; actually, calculus is required mainly for Chapter 5, dealing with basic d tion theory in the continuous case, and in the new optional sections of Cha

In this eighth edition, we have continued to build on the strengths of th ous editions by adding more data sets and examples of the application of s in scientific investigations. The new data sets, like many of those alread text, arose in the author's consulting activities or in discussions with scient engineers about their statistical problems. Data from some companies ha disguised, but they still retain all of the features necessary to illustrate the s methods and the reasoning required to make generalizations from data c in an experiment. Altogether, there are a great number of new examp exercises.

The new Table 5.2 summarizes the common discrete and continuous tions. To improve both presentation and readability, we have split the Chapter 7 into a new Chapter 7 and Chapter 8, where the second chapter is to comparisons. Because of its popularity in engineering practice, maximu hood estimation is developed in a new Section 7.3. A new Section 13.5 added to introduce the idea of response surface methodology through an ex

A major new enhancement is the inclusion of four short new sections the essentials of a slightly more mathematical treatment of distributio Motivated by the suggestions from many reviewers and users of Section 5.11 covers moment generating functions, Section 6.6 the momen ating function method, Section 6.5 representations and relations among th theory distributions, and Section 6.7 transformation methods for obtain pling distributions. Any or all of the last four sections can be omitted witho ing the readability or comprehension, but an appreciation of the results can the understanding of statistical procedures.

Besides having the data sets available online at the Pearson Data S Site—www.pearsonhighered.com/datasets—we also have included e details of some of the new sections. This material covers convolution meth

Section 14.8 and its accompanying tables from Appendix B of the previous edition. Exercises are also included so instructors wishing to incorporate these topics into their courses can do so easily.

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 theory of statistics and, together with some of the material in Chapter 15, they are suitable for an introductory semester (or quarter) course on probability and statistics. Chapters 7, 8, 9, and 10 from 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 regression, analysis of variance, and experimental design. Chapter 14 concerns nonparametric rank tests and goodness-of-fit tests. 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.

We wish to thank MINITAB (State College, Pennsylvania) for permission to include commands and output from their *MINITAB* software package, and the SAS institute (Gary, North Carolina) for permission to include output from their SAS package. Appendix C presents an introduction to the software package R.

For their contributions to earlier additions, thanks go to Paul M. Berthouex, Michael Brajac, Smiley Cheng, Don Ermer, Jim Evans, Cherilyn Hatfield, David Steinberg, and Steve Verrill for contributing data sets. Thanks also to Kuo-Tsung Wu, Songyong Sim, Ruoji Li, and Erik Johnson for their help. I would like to especially thank Asit Bannerjee, Robert Blick, Song Jin, Chang Sik Lee, Shiyu Zhou, and others who contributed data sets to this edition. It is the resulting rich class of examples and exercises that convey the usefulness of statistics and make the subject interesting the students. The current edition also benefited greatly from the input of accuracy checkers Keith Friedman, University of Texas at Austin, and Kathryn Sharpe, Stony Brook University.

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

Richard A. Johnson

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 the downtimes of a computer, 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, in particular, 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? 1

1.2 Modern Statistics 2

1.3 Statistics and Engineering 2

1.4 The Role of the Scientist and Engineer in Quality Improvement 3

1.5 A Case Study: Visually Inspecting Data to Improve Product Quality 3

1.6 Two Basic Concepts—Population and Sample 5

Review Exercises 10

Key Terms 11

1.1 Why Study Statistics?

Answers provided by statistical analysis can provide the basis for making decisions or choosing 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, a civil engineer must determine the strength of supports for generators at a power plant. A number of those available must be loaded to failure, and their strengths will provide the basis for assessing the strength of other supports not subject to testing. The proportion of all supports available with strengths that lie below a design limit needs to be determined.

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 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, and the most important feature of its growth has been the shift in emphasis from descriptive statistics to **statistical inference**. Statistical inference concerns generalization 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 one makes a statistical inference, namely, an inference that goes beyond the information contained in a set of data, one must always proceed with caution. One must decide carefully how far to go in generalizing from a given set of data, whether such generalizations are at all reasonable or justifiable, whether it might be wise to wait until there are more data, and so forth. Indeed, some of the most important problems of statistical inference concern the appraisal of the risks and the consequences to which one might be exposed 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 shall approach the subject of statistics as a science, developing each statistical idea insofar as possible from its probabilistic foundation, and applying 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. However, we shall endeavor continually to make the reader aware that the subjective factors do exist, and to indicate whenever possible 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

There are few areas where the impact of the recent growth of statistics has been felt more strongly than 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 secretary 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

In the last half of the twentieth century and continuing in the present, 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

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

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

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

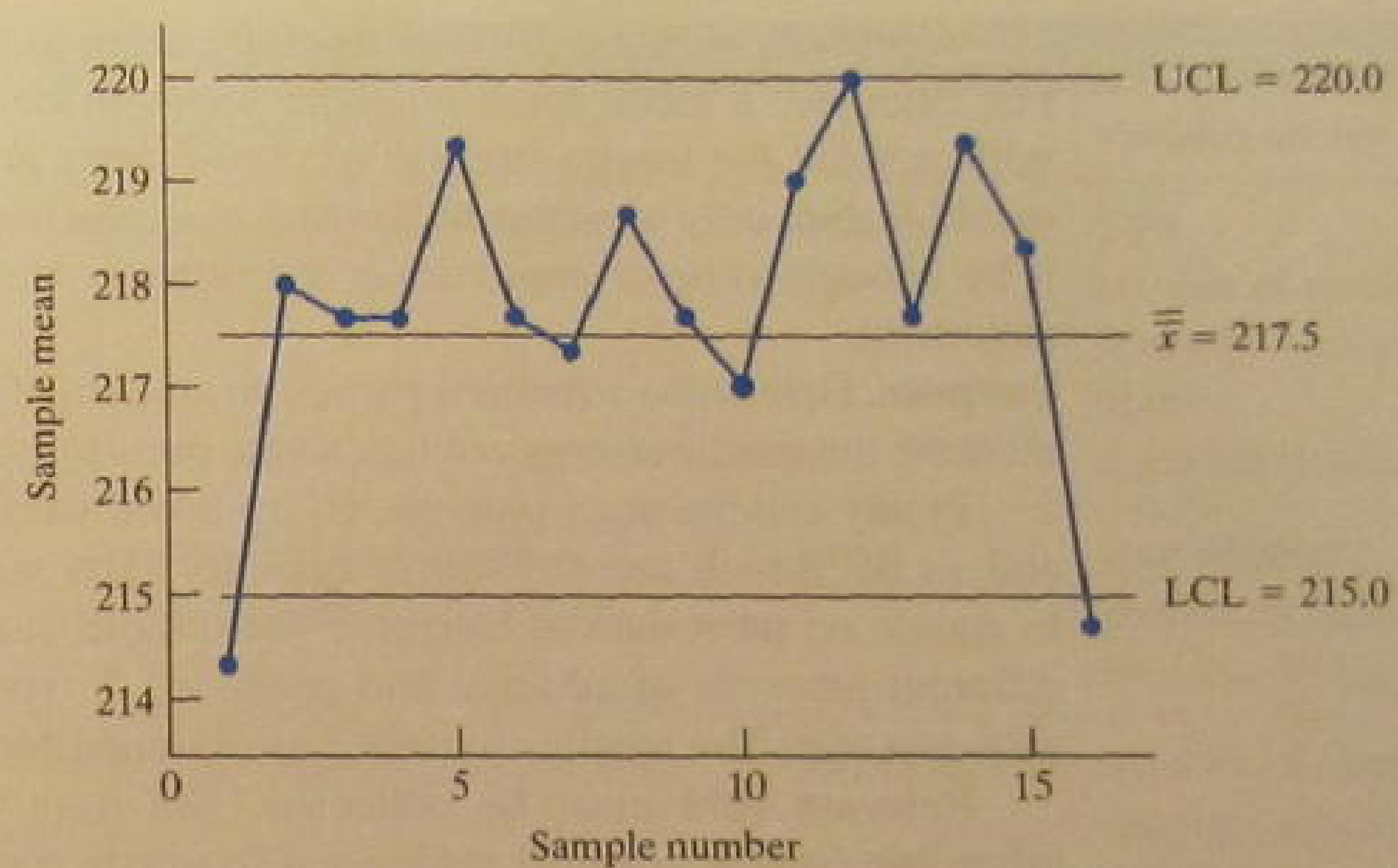


Figure 1.1
X-bar chart for depth

is large compared to the others. This evidence suggests that the pressing machi had not yet reached a steady state. The control chart suggests that it is necessary warm up the pressing machine before the first shift begins at 6 A.M. Managem and engineering implemented an early start-up and thereby improved the proce The operator and foreman did not have the authority to make this change. Dem claimed that 85% or more of our quality problems are in the system and that operator and others responsible for the day-to-day operation are responsible for 1 or less of our quality problems.

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

On Monday morning when the operator started up the pressing machine, of the dies broke. The cost of the die was over a thousand dollars. But this was the biggest cost. When a customer was called and told there would be a delay delivering the ceramic parts, he canceled the order. Certainly the loss of a custo is an expensive item. Deming referred to this type of cost as the unknown unknowable, but at the same time it is probably the most important cost of p quality.

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

1.6 Two Basic Concepts—Population and Sample

The preceding examples, in which the evaluation of actual information is essent for acquiring new knowledge, motivate the development of statistical reasoning a tools taught in this text. Most experiments and investigations conducted by engine in the course of investigating, be it a physical phenomenon, production process, 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 like *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 tried two different amounts of additive and produced 25 specimens of the plastic part with each mixture. The tensile strength was 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 has been 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 **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, 48 pass and 12 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 **Self-selected samples—a bad practice**

A magazine which features the latest computer hardware and software for home-office use enclosed a short questionnaire on a postcard. Readers were asked to indicate whether or not they owned specific new software packages or hardware products. In past issues, this magazine used similar information from cards that were returned 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 the magazine editors is not representative and is badly biased. Readers who always update and try most of the new software will be more likely to respond indicating their purchases. In contrast, those who did not purchase any of the software or hardware mentioned in the survey will be very likely not to return the postcard. That is, the proportion of purchasers of software package P in the sample of returned postcards 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 7 at the end of the book 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 **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 a

Table 1.3 Random digits (portion of Table 7)

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

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 **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 survey. Use the random digit Table 7 to select five phone numbers.

Solution We arbitrarily decide to start on the second page of Table 7 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, do select samples using a random device or random number table.

Don'ts

1. Don't unquestioningly accept self-selected samples.

Review Exercises

1.1 A consumer magazine article asks "How Safe Is the Air in Airplanes?" and goes on to say that the air quality was measured on 158 different flights for U.S.-based airlines. Let the variable of interest be a numerical measure of staleness. Identify the population and the sample.

1.2 A radio-show host announced that she wanted to know which rapper was the favorite among college students in your school. Listeners were asked to call and name their favorite rapper. 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 laptop computers owned by students at your university. You want to know the weight of the laptop.

(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) To learn about starting salaries for engineers graduating from a Midwestern university, twenty graduating seniors are asked to report their starting salary.

(b) Fifty computer memory chips were selected from the thousand manufactured that day. The computer memory chips were tested, and 5 were found to be defective.

(c) Tensile strength was measured on 20 specimens made of a new plastic material. The intent is to learn about the tensile strengths for all specimens that could conceivably be manufactured with the new plastic material.

1.5 For hard drives to operate properly, the distance between the read/write head and the disk must remain between tight limits. From each hour's production, 40 drives are selected and the distance 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 6 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
	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

Classical approach to statistics 2
Descriptive statistics 2
Population 6
Population of units 6
Quality improvement 3
Random number table 8
Reliability 3
Sample 6
Statement of purpose 6
Statistical inference 2
Statistical population 6
X-bar chart 4
Unit 6
Variable 6

CHAPTER

2 ORGANIZATION AND DESCRIPTION OF DATA

CHAPTER OUTLINE

2.1 Pareto Diagrams and Dot Diagrams 12
2.2 Frequency Distributions 14
2.3 Graphs of Frequency Distributions 17
2.4 Stem-and-Leaf Displays 21
2.5 Descriptive Measures 24
2.6 Quartiles and Percentiles 29
2.7 The Calculation of $\bar{x}$ and s 34
2.8 A Case Study: Problems with Aggregating Data 39
Review Exercises 42
Key Terms 44

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, for a computer-controlled lathe whose performance was below par, workers recorded the following causes 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 14.2.

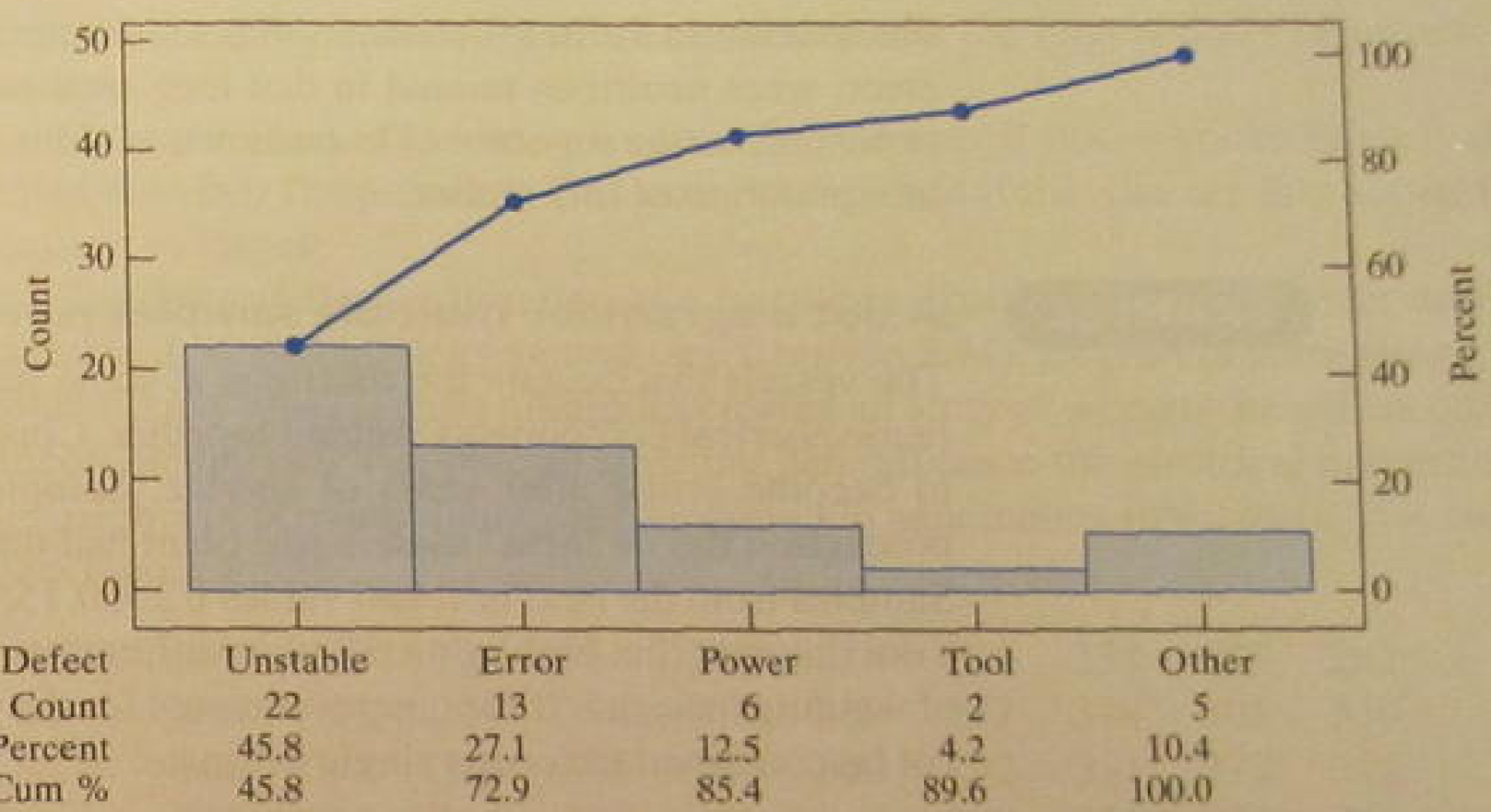


Figure 2.1
A Pareto diagram of failures

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

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

3 6 −2 4 7 4 3

are plotted as a **dot diagram** in Figure 2.2. The dot diagram visually summarizes information that the lathe is, generally, running fast. In Chapters 13 and 15 we develop efficient experimental designs and methods for identifying primary cau 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 pattern of variation. Still, it is a good idea to plot the data and look for unus features.

EXAMPLE **Dot diagrams expose outliers**

In 1987, for the first time, physicists observed neutrinos from a supernova occurred outside of our solar system. At a site in Kamiokande, Japan, the follow times (in seconds) between neutrinos were recorded:

0.107 0.196 0.021 0.283 0.179 0.854 0.58 0.19 7.3 1.18 2

Draw a dot diagram.

Solution We plot to the nearest 0.1 second to avoid crowding (see Figure 2.3). Note extremely long gap between 2.0 and 7.3 seconds. Statisticians call such an unu

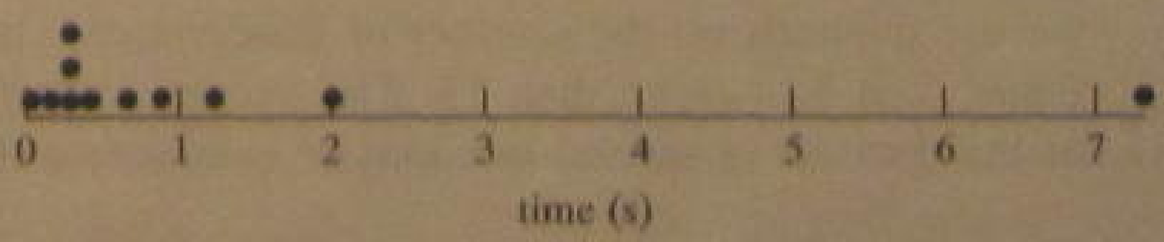


Figure 2.3
Dot diagram of time between neutrinos

observation an **outlier**. Usually, outliers merit further attention. Was there a recording error; were neutrinos missed in that long time interval; or were there two separate explosions in the supernova? Important questions in physics may hinge on the correct interpretation of this outlier. ■

EXAMPLE **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 Chapter 14 we 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. ■

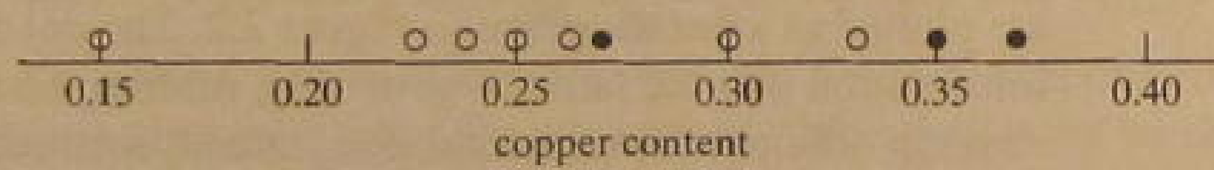


Figure 2.4
Dot diagram of copper content

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. Such a 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, this kind of grouping often brings out important features of the data, and the gain in "legibility" usually more than compensates for the loss of information. In what follows, 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 t smallest.

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

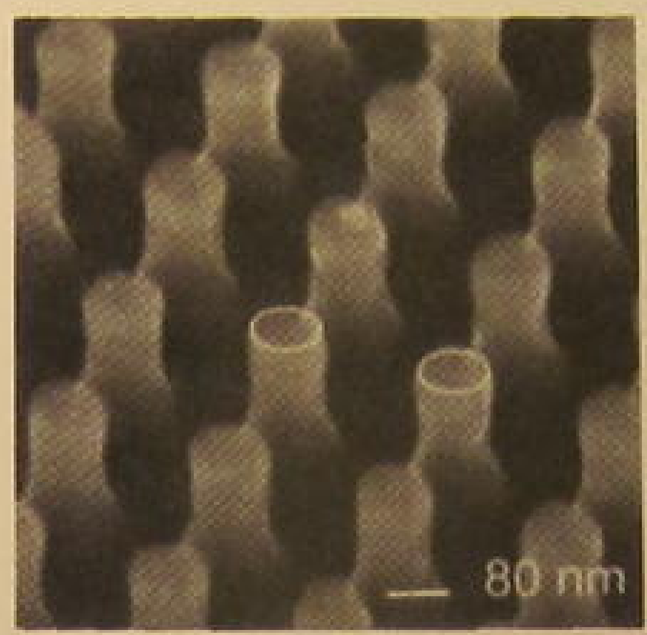


Figure 2.5
Nanopillars

To illustrate the construction of a frequency distribution, we consider data co lected in a nanotechnology setting (see Exercise 2.44). Engineers fabricating a n transmission-type electron multiplier created an array of silicon nanopillars on a f silicon membrane. The precise structure can influence the electrical properties, the heights of 50 nanopillars were measured in nanometers (nm), or $10^{-9} \times$ mete (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 rang $391 - 221 = 170$, we might choose five classes having the limits 206–245, 246– 286–325, 326–365, 366–405, or the six classes 216–245, 246–275, . . . , 366– Note that, in either case, **the classes do not overlap, they accommodate all data, and they are all of the same width**.

Initially, deciding on the first of these classifications, we count the numbe 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 ori data. Had the original data been given to one decimal place, we would have use class limits 205.9–245.0, 245.1–285.0, . . . , 365.1–405.0. If they had been round the nearest 10 nanometers, we would have used the class limits 210–240, 250– 290–320, 330–360, 370–400.

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

We make an **endpoint convention**. For the pillar height data, we can (205, 245] as the first class, (245, 285] as the second, and so on through (365,

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

That is, for this data set, we adopt the the convention that the right-hand endpoint is included but the left-hand endpoint is not. For other data sets we 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 limits or 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 **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 **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**.

If it is desirable to compare frequency distributions, it may be necessary (or at least advantageous) to convert them into **percentage distributions**. We simply divide each class frequency by the total frequency (the total number of observations in the distribution) and multiply by 100; in this way we indicate what percentage of the data falls into each class of the distribution. The same can also be done with cumulative distributions, thus converting them to **cumulative percentage 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; 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.

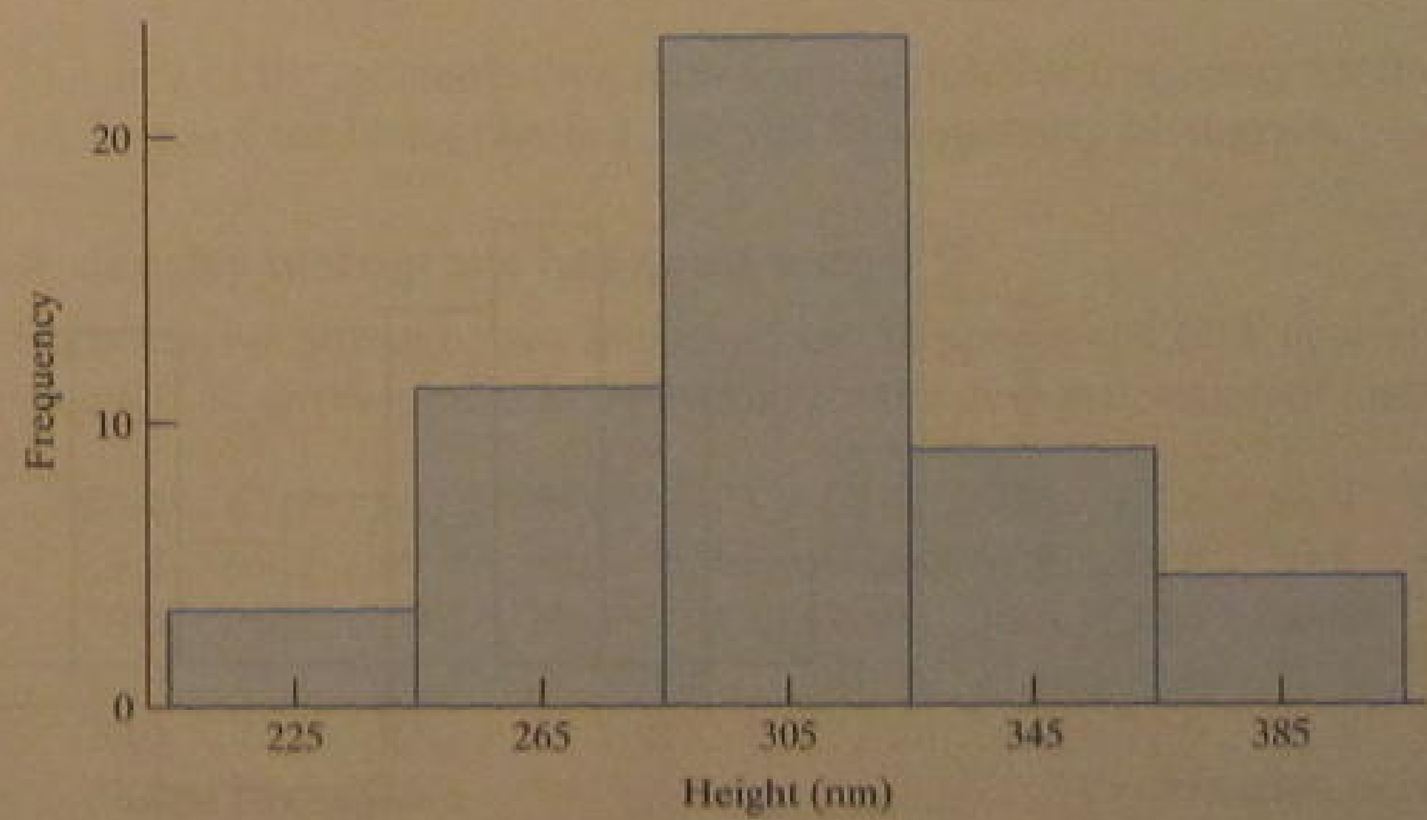


Figure 2.6
Histogram of pillar height

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 (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.

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. For instance, outlying observations which somehow do not fit the overall picture, that is, 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 **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. ■

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

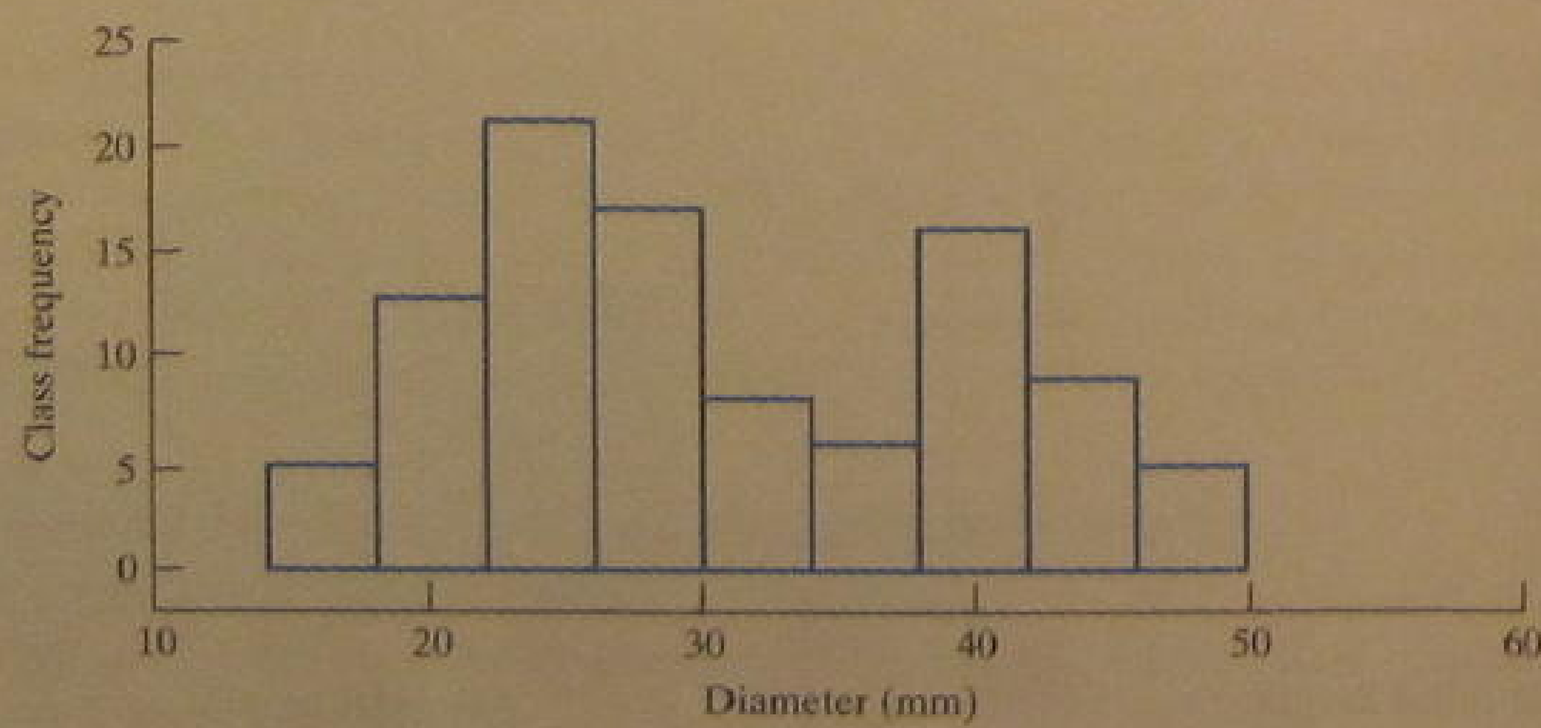


Figure 2.7
Histogram of pellet diameter

EXAMPLE **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).

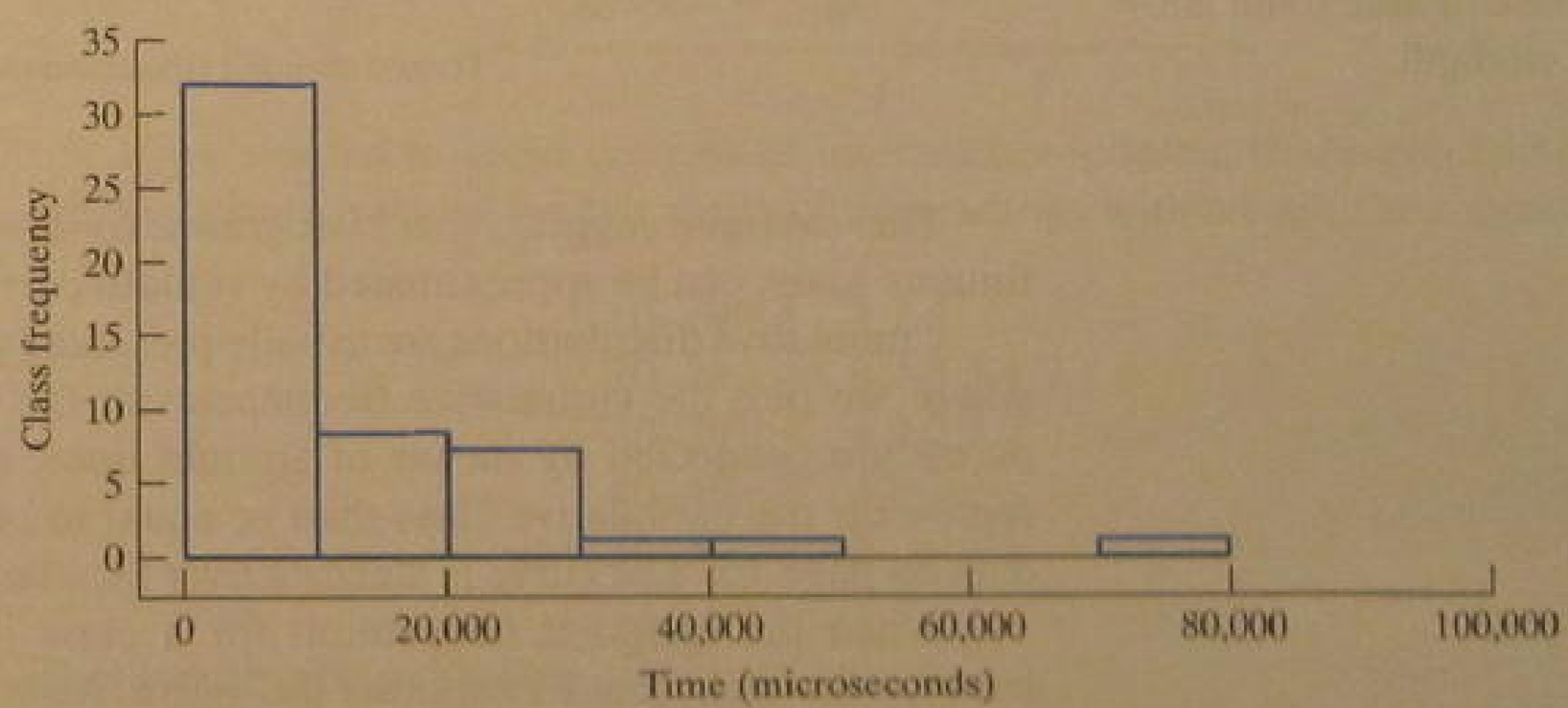


Figure 2.8
Histogram of interrequest time

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**.

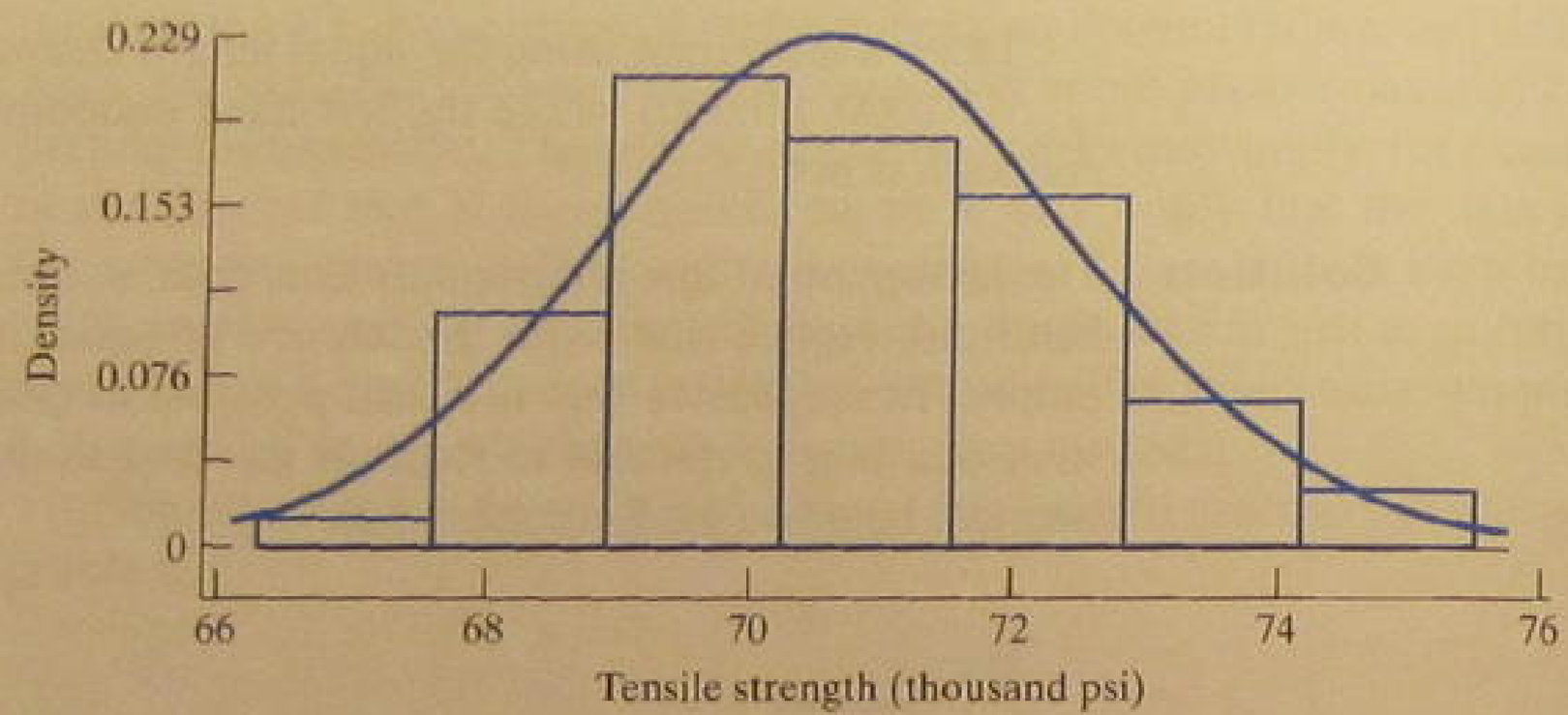
EXAMPLE **A density histogram has total area 1**

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 6, we will introduce this bell-shaped family of curves. ■

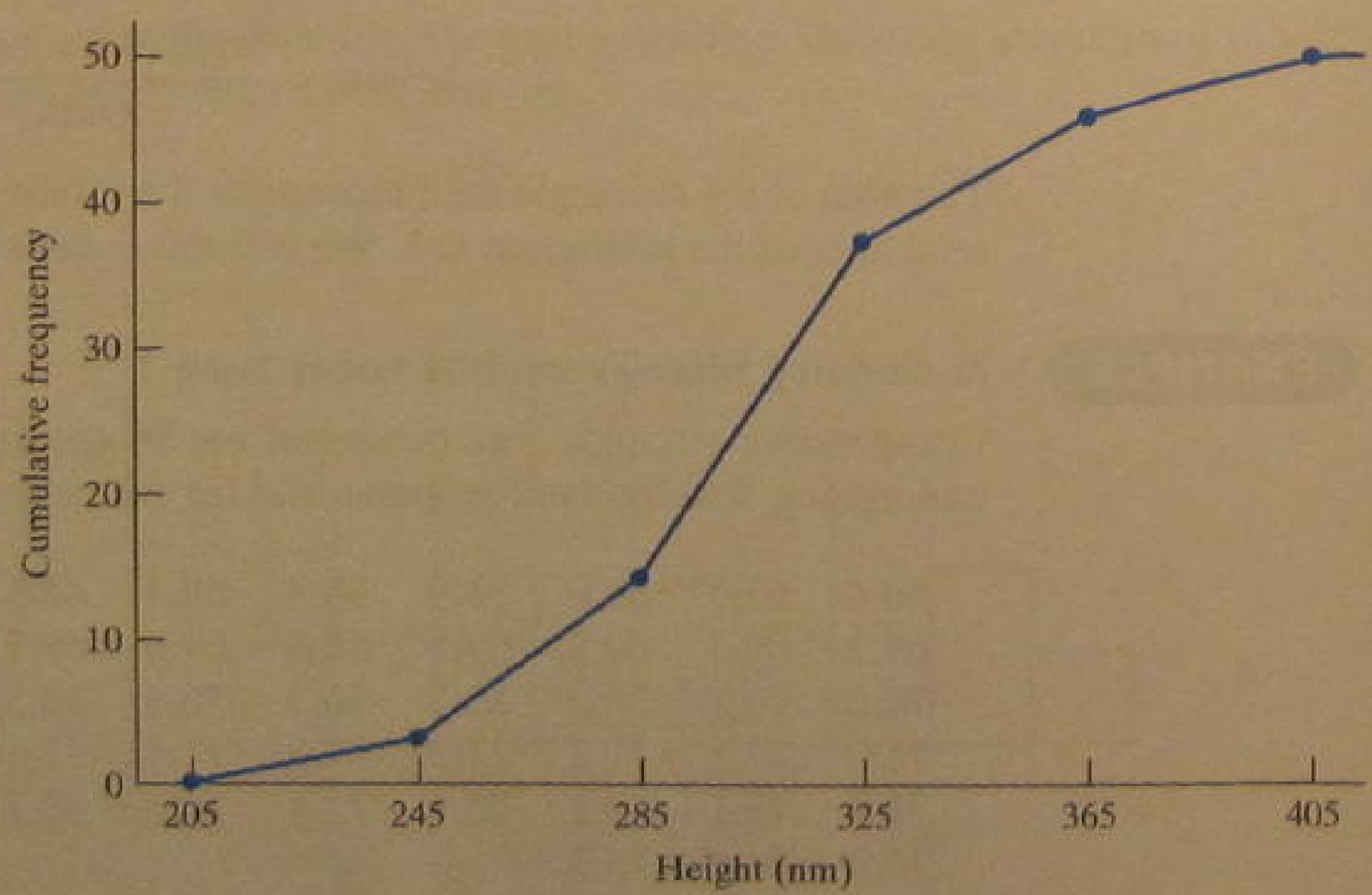


re 2.9
gram of aluminum alloy
e strength

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 15. 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.



e 2.10
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 Accidents at a potato-chip plant are categorized according to the area injured.

fingers	17
eyes	5
arm	2
leg	1

Draw a Pareto chart.

2.2 Damages at a paper mill (in thousands of dollars) due to breakage can be divided according to the product manufactured.

toilet paper	132
hand towels	85
napkins	43
12 other products	50

(a) Draw a Pareto chart.

(b) What percent of the loss occurs in making:

(1) toilet paper?

(2) toilet paper or hand towels?

2.3 Probes for noninvasive measurement of blood flow must not give off too much energy (milliwatts per square centimeter). A sample of 12 probes yielded the energy values:

9, 10, 6, 7, 9, 7, 9, 6, 8, 5, 10, and 12

Construct a dot diagram.

2.4 The following are 14 measurements on the strength (psi) of paper to be used in cardboard tubes:

121, 128, 129, 132, 135, 133, 127, 115, 131, 125, 118, 114, 120, 116

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 Jump River Electric serves part of Northern Wisconsin, and because much of the area is forested, it is prone to outages. One August there were 11 power outages. Their durations (in hours) are

2.5 2.0 1.5 3.0 1.0 1.5 2.0 1.5 1.0 10.0 1.0

Display the data in a dot diagram.

2.7 The breaking forces (lbf) of salt pellets, given to the nearest tenth, are grouped into a table having the classes [140.0, 160.0), [160.0, 180.0), [180.0, 200.0), [200.0, 220.0), [220.0, 240.0), and [240.0, 260.0), where the left-hand endpoint is included but the right-hand endpoint is not. Find

(a) the class marks

(b) the class interval

2.8 With reference to the preceding exercise, is it possible to determine from the grouped data how many pellets had breaking force

(a) less than 160.0?

(b) more than 160.0?

(c) at least 220.0?

(d) at most 240.0?

(e) from 220.0 to 260.0 inclusive?

2.9 The class marks of a distribution of temperature readings (given to the nearest degree Celsius) are 16, 25, 34, 43, 52, and 61. Find

(a) the class boundaries

(b) the class interval

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 ignition times of certain upholstery materials exposed to a flame (given to the nearest hundredth of a second):

2.58	2.51	4.04	6.43	1.58	4.32	2.20	4.19
4.79	6.20	1.52	1.38	3.87	4.54	5.12	5.15
5.50	5.92	4.56	2.46	6.90	1.47	2.11	2.32
6.75	5.84	8.80	7.40	4.72	3.62	2.46	8.75
2.65	7.86	4.71	6.25	9.45	12.80	1.42	1.92
7.60	8.79	5.92	9.65	5.09	4.11	6.37	5.40
11.25	3.90	5.33	8.64	7.41	7.95	10.60	3.81
3.78	3.75	3.10	6.43	1.70	6.40	3.24	1.79
4.90	3.49	6.77	5.62	9.70	5.11	4.50	2.50
5.21	1.76	9.20	1.20	6.85	2.80	7.35	11.75

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 In a two-week study of the productivity of workers, the following data were obtained on the total number of acceptable pieces which 100 workers produced:

65	36	49	84	79	56	28	43	67	36
43	78	37	40	68	72	55	62	22	82
88	50	60	56	57	46	39	57	73	65
59	48	76	74	70	51	40	75	56	45
35	62	52	63	32	80	64	53	74	34
76	60	48	55	51	54	45	44	35	51
21	35	61	45	33	61	77	60	85	68
45	53	34	67	42	69	52	68	52	47
63	65	55	61	73	50	53	59	41	54
41	74	82	58	26	35	47	50	38	70

Group these figures into a distribution having the classes 20–29, 30–39, 40–49, . . . , and 80–89, and plot a histogram using [20, 30), . . . , [80, 90), 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 automobile accidents that occurred at 60 major intersections in a certain city during a Fourth of July weekend:

0	2	5	0	1	4	1	0	2	1
5	0	1	3	0	0	2	1	3	1
1	4	0	2	4	1	2	4	0	4
3	5	0	1	3	6	4	2	0	2
0	2	3	0	4	2	5	1	1	2
2	1	6	5	0	3	3	0	0	4

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

2.17 Convert the distribution obtained in Exercise 2.16 into a cumulative "or more" distribution and draw its ogive.

2.18 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 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.

2.19 The pictogram of Figure 2.11 is intended to illustrate the fact that per capita income in the United States doubled from \$13,000 in 1988 to \$26,000 in 2006. Does this pictogram convey a fair impression of the actual change? If not, state how it might be modified.

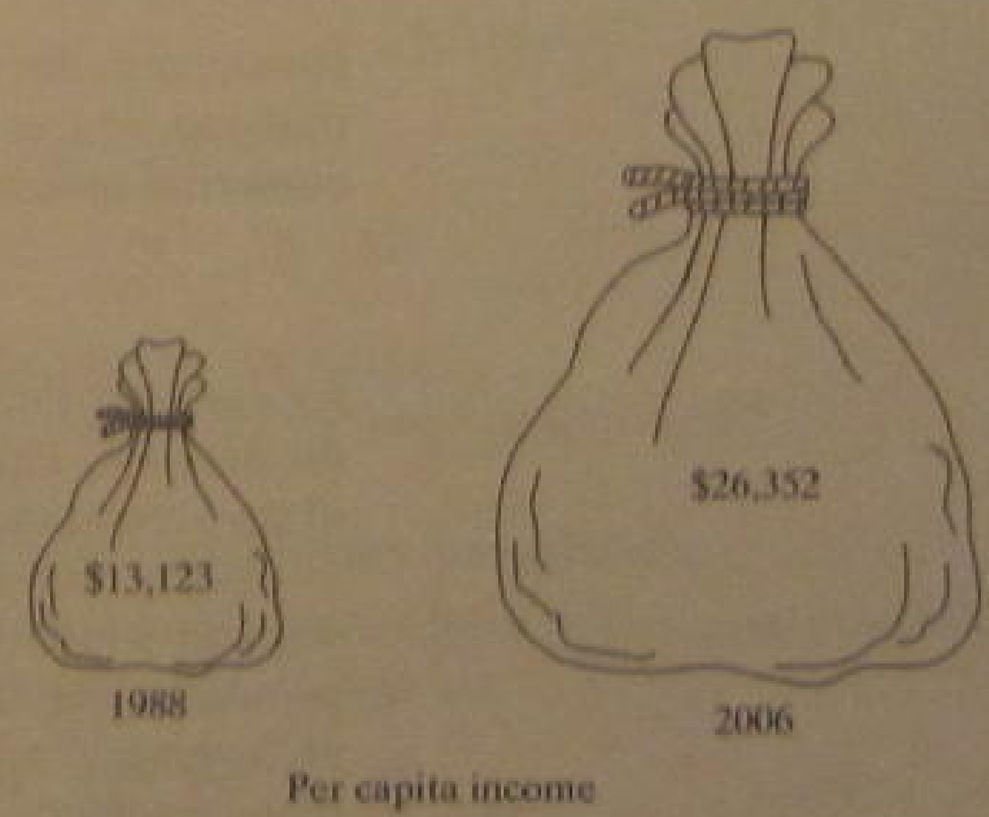


Figure 2.11 Pictogram for Exercise 2.19

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

2.20 Convert the distribution of nanopillar heights on page 15 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.

2.21 Given a set of observations $x_1, x_2, \ldots,$ and x_n, we define their **empirical cumulative distribution** as the function whose values $F(x)$ equal the proportion of the observations less than or equal to x. Graph the empirical cumulative distribution for the 12 measurements of Exercise 2.3.

2.22 The following are figures on an oil well's daily production in barrels: 214, 203, 226, 198, 243, 225, 207, 203, 208, 200, 217, 202, 208, 212, 205, and 220. Construct a stem-and-leaf display with the stem labels 19, 20, . . . , and 24.

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) 1 | 1 2 3 4 5 7 8. Leaf unit = 1.0.

(b) 23 | 0 0 1 4 6. Leaf unit = 1.0.

(c) 2 | 03 18 35 57. First leaf digit unit = 10.0

(d) 3.2 | 1 3 4 4 7. Leaf unit = 0.01

2.25 If we want 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 21, 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 might wish to have 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. We now develop numerical measures to describe 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

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

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

Sometimes it is preferable to use the **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 **Calculation of the sample mean and median**

In order to control costs, a company collects data on the weekly number of meals claimed on expense accounts. The numbers for five weeks are

15 14 2 27 13

Find the mean and the median.

Solution The mean is

$$\bar{x} = \frac{15 + 14 + 2 + 27 + 13}{5} = 14.2 \text{ meals}$$

and, ordering the data from smallest to largest

$$2 \quad 13 \quad \underbrace{14} \quad 15 \quad 27$$

the median is the third largest value, namely, 14 meals.

Both the mean and median give essentially the same central value. ■

EXAMPLE **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 9 17 19 4 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.

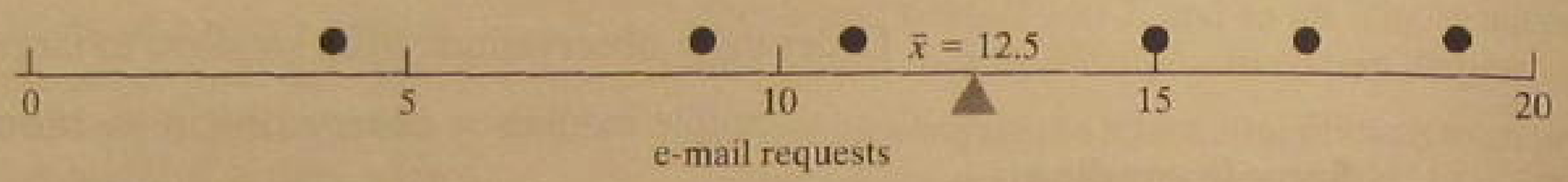


Figure 2.12
The interpretation of the sample mean as a balance point

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

A small company employs four young engineers, who each earn $70,000, and the owner (also an engineer), who gets $160,000. Comment on the claim that on the average the company pays $88,000 to its engineers and, hence, is a good place to work.

Solution The mean of the five salaries is $88,000, but it hardly describes the situation. The median, on the other hand, is $70,000, 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. Measures such as the mean and median describe one important aspect of a set of data—their "middle" or their "average"—but they tell us nothing about this other basic characteristic. 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 then 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, 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. Also, using divisor $n - 1$ produces an estimate that will not, on average, lead to consistent overestimation or consistent underestimation.

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 **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:

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

Then we set up the work required to find $\sum (x_i - \bar{x})^2$ in the following 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

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. In other 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 (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 **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

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

■

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 **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 mean, the median, the variance, and the standard deviation, but 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 **Calculation of percentiles from the nanopillar height data**

Obtain the quartiles and the 93rd percentile for heights of nanopillars data on page 15.

Solution The ordered heights of the nanopillars are

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

According to our calculation rule, $np = 50\left(\frac{1}{4}\right) = 12.5$, which we round up to 13. The first quartile is the 13th ordered observation

$$Q_1 = 278 \quad \text{nm}$$

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

$$np = 50\left(\frac{1}{2}\right) = 25$$

which is an integer. Therefore, we average the 25th and 26th ordered values

$$Q_2 = \frac{304 + 305}{2} = 304.5 \quad \text{nm}$$

The third quartile is the 38th observation, $Q_3 = 330$ nm. We could also have started at the largest value and counted down to the 13th position.

To obtain the 93rd percentile, we determine that $np = 50 \times 0.93 = 46.5$, which we round up to 47. Counting to the 47th position, we obtain

$$P_{0.93} = 366 \quad \text{nm}$$

The 93rd percentile provides a useful description regarding the tallest nanopillars. Only 7% of the pillars measured have height greater than 366 nm.

In the context of monitoring tall pillars, we also record that the maximum height measured was 391 nm. ■

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 **Calculation of range and interquartile range for nanopillar data**

Obtain the range and interquartile range for the heights of nanopillars data in the previous example.

Solution The minimum = 221. From the previous example, the maximum = 391, $Q_1 = 278$, and $Q_3 = 330$.

$$\text{range} = \text{maximum} - \text{minimum} = 391 - 221 = 170 \quad \text{nm}$$

$$\text{interquartile range} = Q_3 - Q_1 = 330 - 278 = 52 \quad \text{nm}$$ ■

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 nanopillar height data. The symmetry seen in the histogram is also evident in this boxplot.

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,

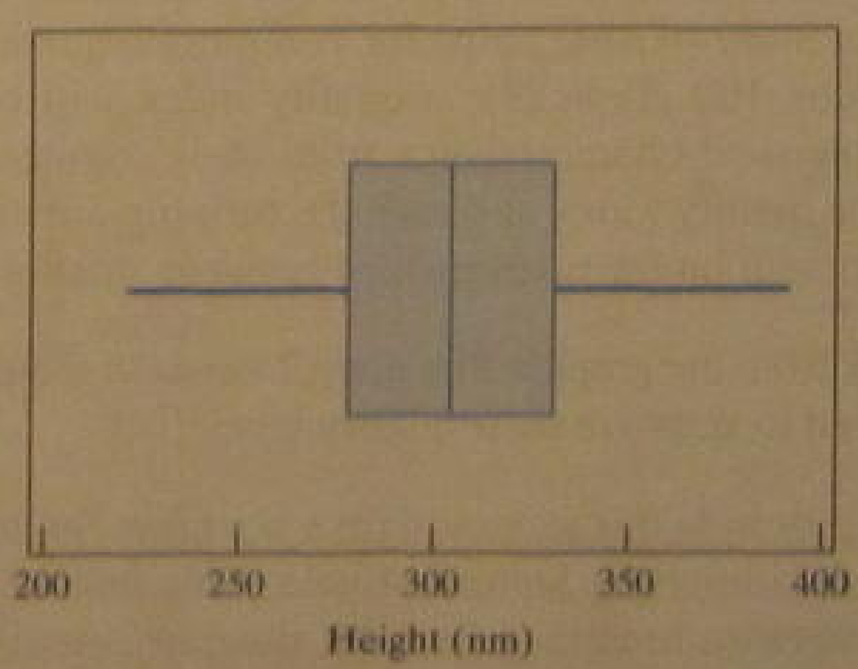


Figure 2.13
Boxplot of the nanopillar height data

it must be within $1.5 \times$ (interquartile range) units of Q_3. The line from 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 **A modified boxplot—possible outliers are detached**

Construct a modified boxplot for the neutrino interarrival time data (see page 13).

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. ■

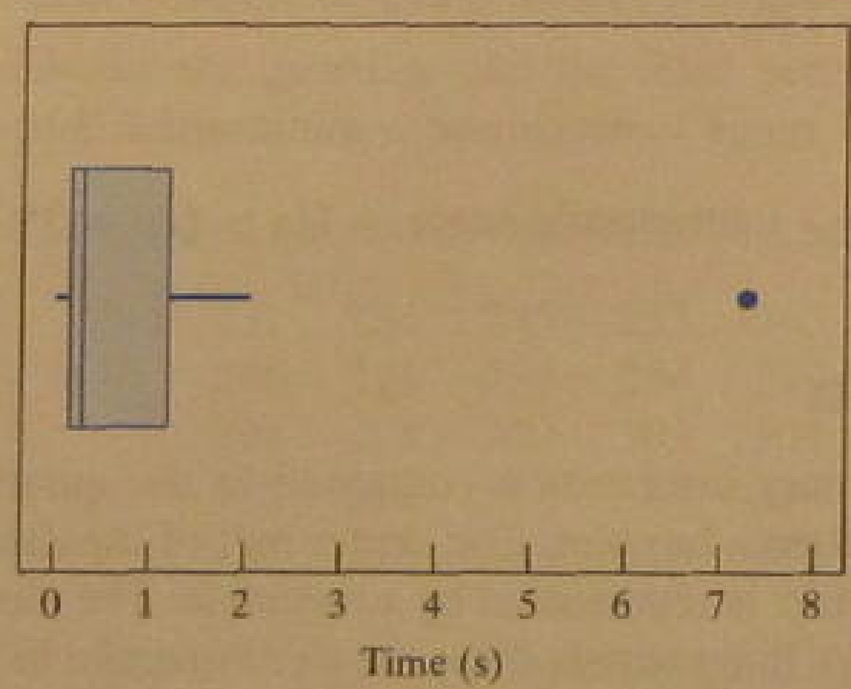


Figure 2.14
Modified boxplot for neutrino data

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

EXAMPLE **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 of data were summarized in a dot diagram, stem-and-leaf display, or boxplot. In

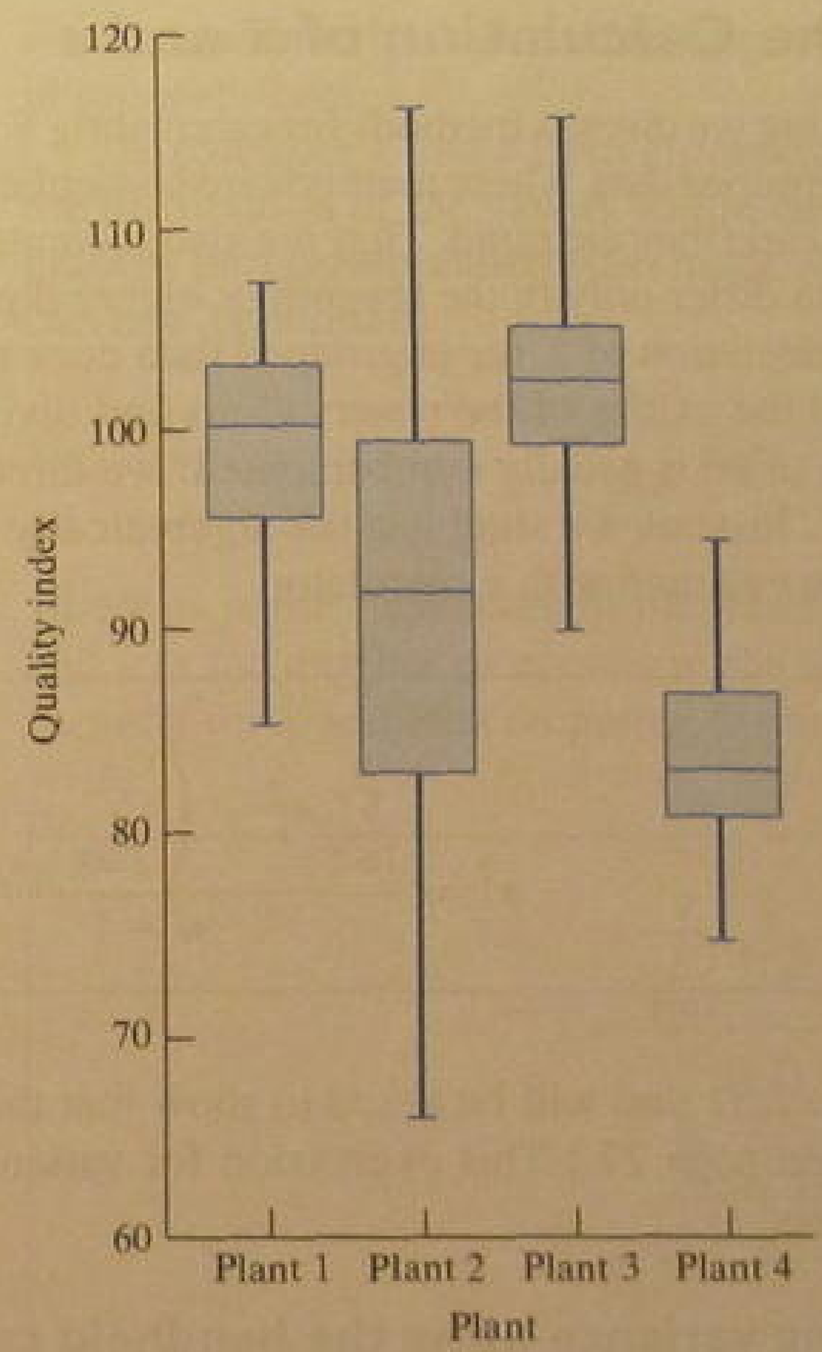


Figure 2.15
Boxplot of the quality index

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 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.

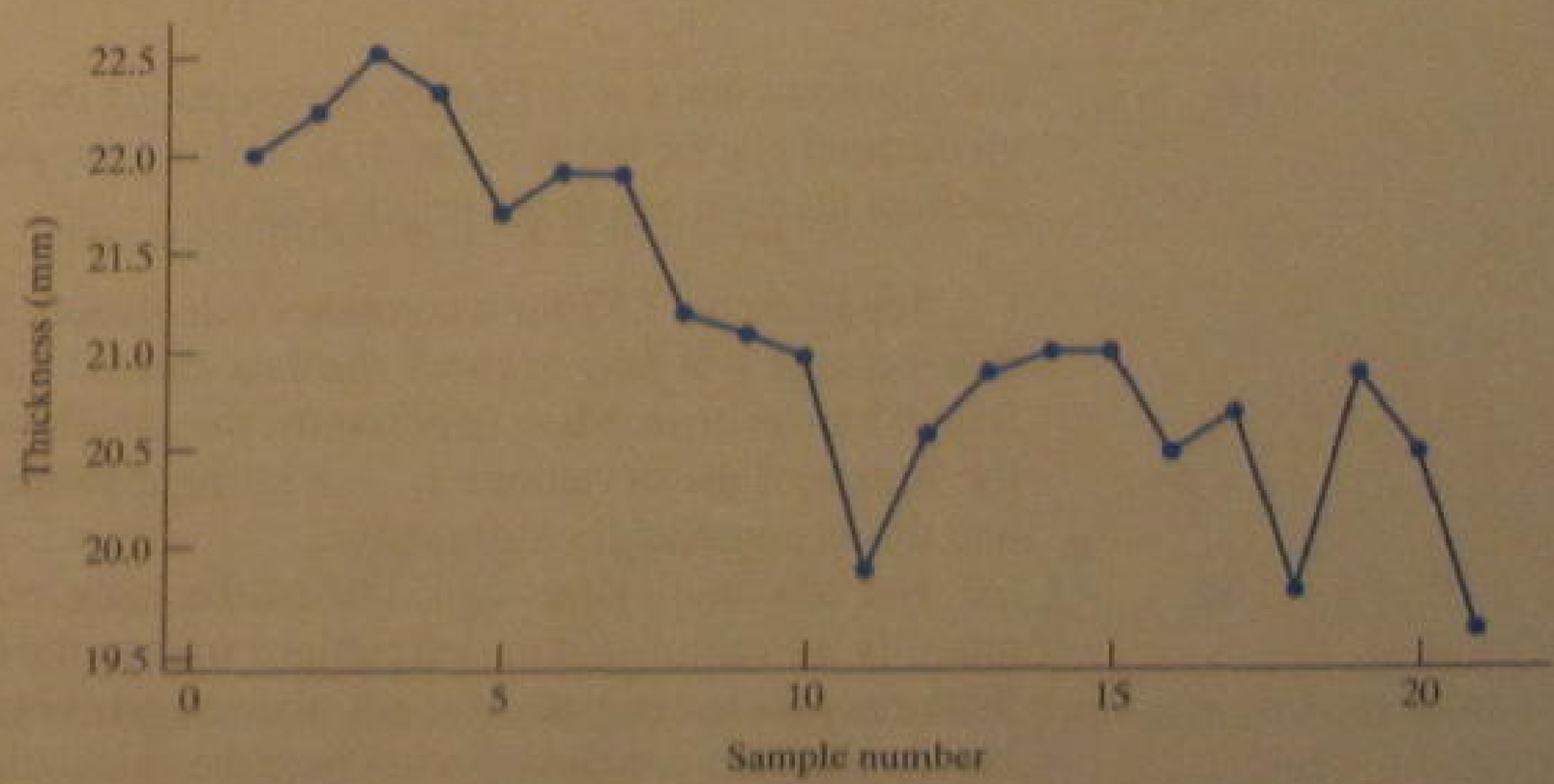


Figure 2.16
Machine measurement of thickness shows trend

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

In this section we discuss methods for calculating $\bar{x}$ and s for **raw data** (ungrouped) as well as grouped data. These methods are particularly well suited for small handheld calculators, and they are rapid. They are also accurate, except in extreme cases where, say, the data differ only in the seventh or higher digits.

The calculation of $\bar{x}$ for ungrouped data does not pose any problems; we have only to add the values of the observations and divide by n. On the other hand, the calculation of s^2 is usually cumbersome if we directly use the formula defining s^2 on page 27. Instead, we shall use the algebraically equivalent form which requires less labor to evaluate with a calculator.

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 27.) This expression for variance is without $\bar{x}$, which reduces roundoff error.

EXAMPLE **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.

Not too many years ago, one of the main reasons for grouping data was to simplify the calculation of descriptions such as the mean and the standard deviation. With easy access to statistical calculators and computers, this is no longer the case, but we shall nevertheless discuss here the calculation of $\bar{x}$ and s from grouped data

since some data (for instance, from government publications) may be available only in grouped form.

To calculate $\bar{x}$ and s from grouped data, we must make some assumption about the distribution of the values within each class. If we represent all values within a class by the corresponding class mark, the sum of the x's and the sum of their squares can now 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 **Calculating a mean and variance from grouped data**

Use the distribution obtained on page 16 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 Laser milling can result in a deposition of debris on a target surface, and this can cause difficulties when laser milling is employed for microstructuring. Mechanical engineers[3] have experimented with pickling followed by electrochemical polishing. For three test specimens, the surface roughness decreased 0.91, 0.99, and 0.98 (μm).

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

(b) Calculate sample standard deviation s.

2.33 Engineers[4] 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

$$35 \quad 37 \quad 38 \quad 34 \quad 30 \quad 24 \quad 13$$

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 national railway's high-horsepower locomotives was given to a large private company. After one year of experience with the maintenance program, those in charge of the program felt that major improvements could be made in the reliability of the locomotives. To document the current status, they collected data on the cost of materials for rebuilding traction motors. Use the data below to

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

(b) calculate the sample standard deviation s.

Materials costs for rebuilding traction motors (thousands of dollars):

1.41	1.70	1.03	0.99	1.68	1.09	1.68	1.94
1.53	2.25	1.60	3.07	1.78	0.67	1.76	1.17
1.54	0.99	0.99	1.17	1.54	1.68	1.62	0.67
0.67	1.78	2.12	1.52	1.01			

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 in Phoenix, Arizona, the normal daily maximum temperature for each month is 65, 69, 74, 84, 93, 102, 105, 102, 98, 88, 74, and 66 degrees Fahrenheit. Verify that the mean of these figures is 85 and comment on the claim that, in Phoenix, the average daily maximum temperature is a very comfortable 85 degrees.

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.[5] The agreement numbers for one pair of waveforms are

$$0.50 \quad 0.40 \quad 0.04 \quad 0.45 \quad 0.65 \quad 0.40 \quad 0.20 \quad 0.30 \quad 0.60 \quad 0.45$$

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

(b) Calculate sample standard deviation s.

[3] T. Dobrev, D. Pham, and S. Dimov, Techniques for improving surface quality after laser milling. *Proc. IMechE, 222 Part B: J. Engineering Manufacture* (2008), 55–65.

[4] 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.

[5] 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 The following are the numbers of minutes that a person had to wait for a bus to work on 15 working days:

10 1 13 9 5 9 2 10 3 8 6 17 2 10 15

(a) Find the mean.

(b) Find the median.

(c) Draw a boxplot.

2.40 With reference to the preceding exercise, find s^2 using

(a) the formula that defines s^2;

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

2.41 Material manufactured continuously before being cut and wound into large rolls must be monitored for thickness (caliper). A sample of 10 measurements on paper, in millimeters, yielded

32.2 32.0 30.4 31.0 31.2 31.2 30.3 29.6 30.5 30.7

Find the mean and quartiles for this sample.

2.42 For the four observations 9 7 15 5

(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 23, draw a boxplot.

2.44 A company was experiencing a chronic weld-defect problem with a water-outlet-tube assembly. Each assembly manufactured is leak tested in a water tank. Data were collected on a gap between the flange and the pipe for 6 bad assemblies that leaked and 6 good assemblies that passed the leak test.

Leaker 0.290 0.104 0.207 0.145 0.104 0.124

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

(b) Calculate the sample standard deviation s.

2.45 Refer to Exercise 2.44. The measurements for 6 assemblies that did not leak were

Good 0.207 0.124 0.062 0.301 0.186 0.124

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

(b) Calculate the sample standard deviation s.

(c) Does there appear to be a major difference in gap between assemblies that leaked and those that did not? The quality improvement group turned their focus to welding process variables.

2.46 Find the mean and the standard deviation of the 20 humidity readings on page 21 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 22 to find the mean and the variance of the nanopillar diameters.

2.48 Use the distribution obtained in Exercise 2.12 on page 23 to find the mean and the standard deviation of the ignition times. Also determine the coefficient of variation.

2.49 Use the distribution obtained in Exercise 2.14 on page 23 to find the coefficient of variation of the productivity 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 34 is equivalent to the one used to define s^2 on page 27.

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 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 $\frac{(n/2) - k}{j}$ 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 $\frac{n}{2}$ of the observations instead of $\frac{n+1}{2}$ as on page 25.

To illustrate, let us refer to the nanopillar height data on page 15 and the frequency distribution on page 16. 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 22 to find the median of the grouped nanopillar diameters.

(b) Use the distribution obtained in Exercise 2.12 on page 23 to find the median of the grouped ignition times.

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 15 and the frequency distribution on page 16, find Q_1, Q_3, and the interquartile range.

(b) Find Q_1 and Q_3 for the distribution of the ignition time 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) The average annual salaries paid to top-level management in three companies are \$164,000, \$172,000, and \$169,000. If the respective numbers of top-level executives in these companies are 4, 15, and 11, find the average salary paid to these 30 executives.

(b) In a nuclear engineering class there are 22 juniors, 18 seniors, and 10 graduate students. If the juniors averaged 71 in the midterm examination, the seniors averaged 78, and the graduate students averaged 89, what is the mean for the entire class?

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 2004 to 2010, 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

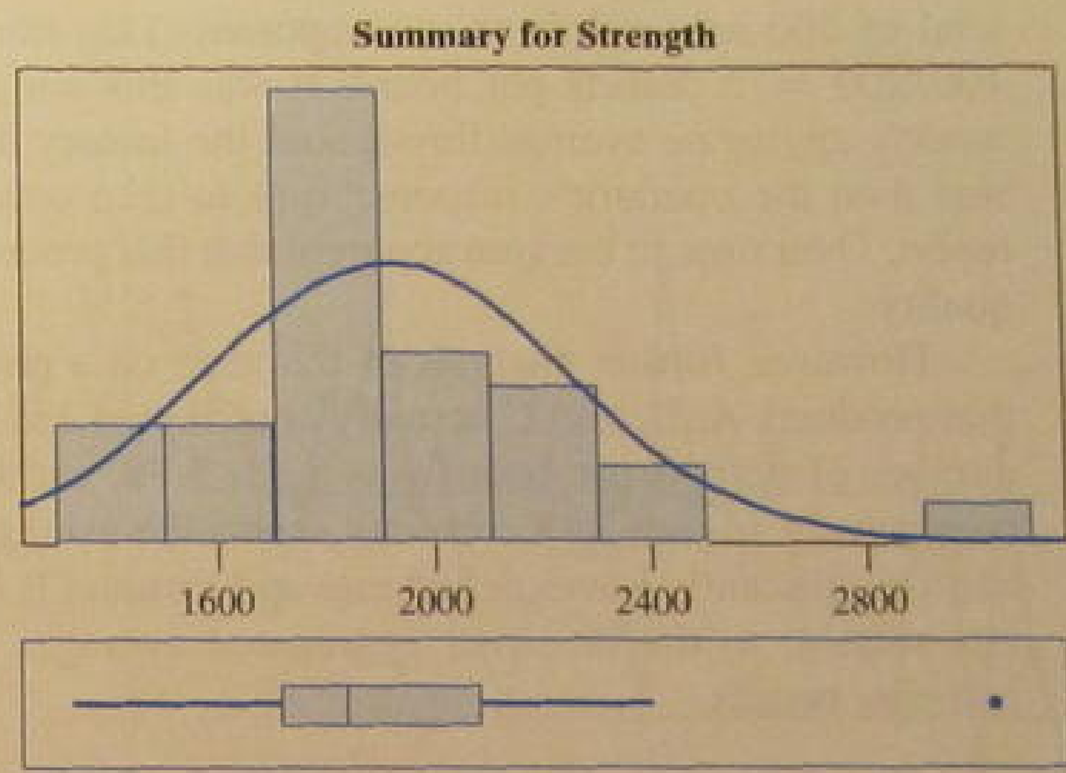
Type *strength* in **Variables**.

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 140

(b) the interrequest times on page 19

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 19.

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.[6] The data 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

[6]Courtesy of Don Ermer.

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.

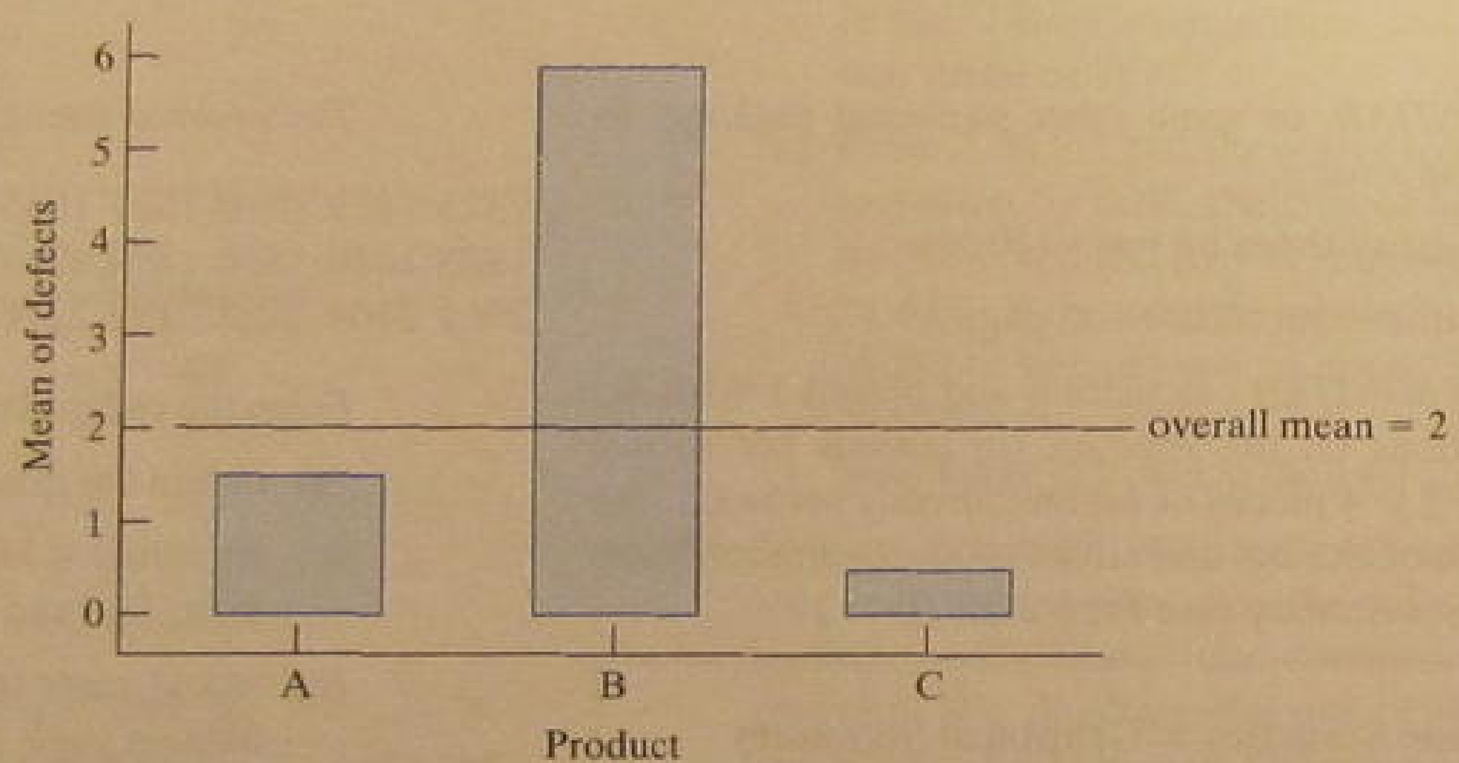


Figure 2.18
Average number of defects per product type

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.

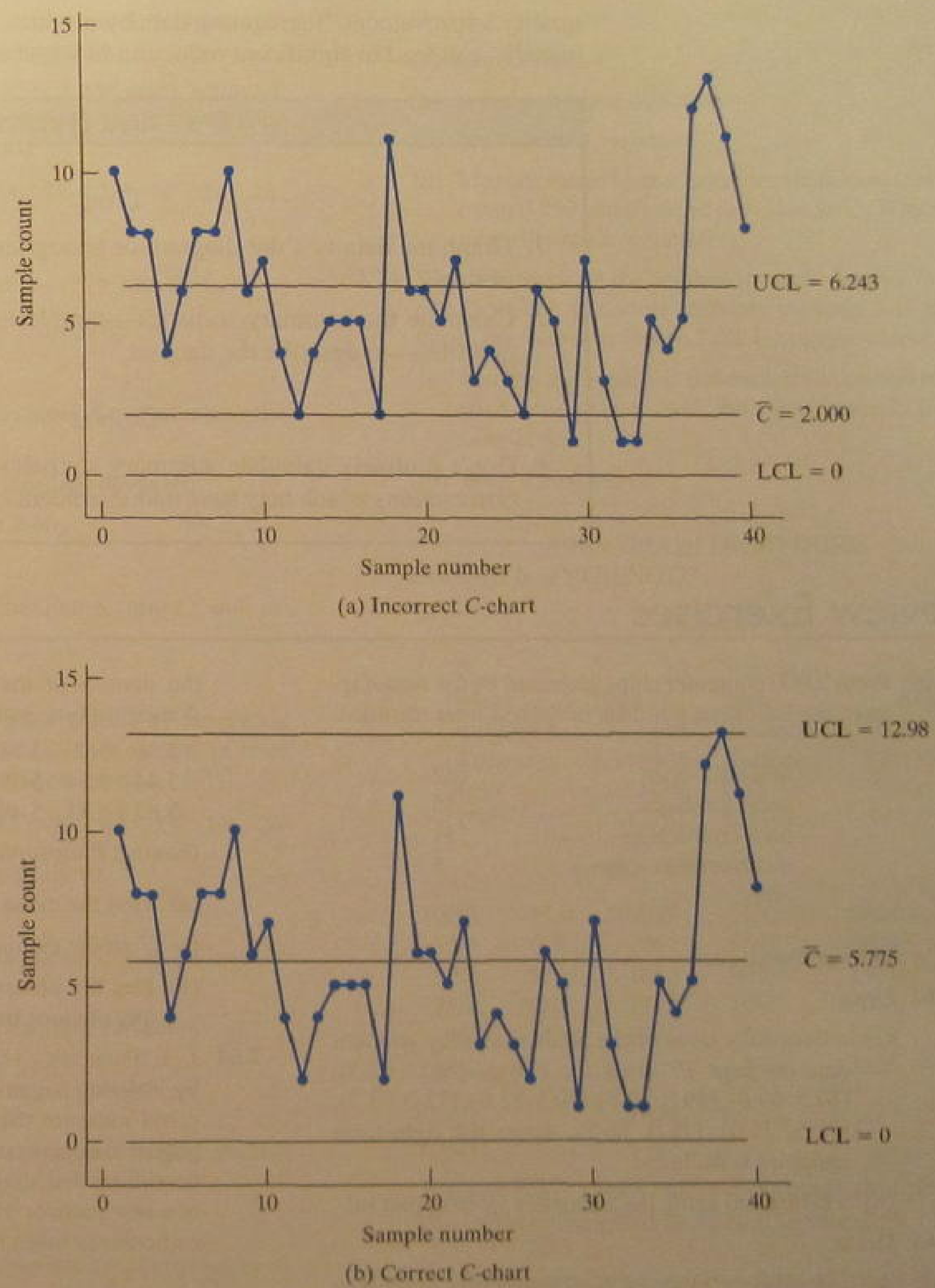


Figure 2.19
C-charts for defects

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.

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 2,000 computer chips inspected by the manufacturer, the following numbers of defects were recorded.

holes not open	182
holes too large	55
poor connections	31
incorrect size chip	5
other	7

Draw a Pareto chart.

2.61 Draw

(a) a frequency table of the aluminum alloy strength data on page 19 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 Draw

(a) a frequency table of the interrequest time data on page 19 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) Draw 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) draw a boxplot.

2.68 A civil engineer monitors water quality by measuring the amount of suspended solids in a sample of river water. Over 11 weekdays, she observed

14 12 21 28 30 63 29 63 55 19 20

suspended solids (parts per million).

(a) Draw a dot diagram.

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

(c) Find the variance and standard variation.

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 With reference to the aluminum-alloy strength data in the example on page 19,

(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, draw a boxplot.

2.72 With reference to the aluminum-alloy strength data in the example on page 19, make a stem-and-leaf display.

2.73 During the manufacture of hard disks, the height between the disk and the head, the headlift, must be controlled. One manufacturer recorded

0.239	0.246	0.245	0.243	0.239	0.241
0.248	0.246	0.243	0.242	0.251	0.246

Find the

(a) sample mean

(b) sample standard deviation

(c) coefficient of variation

(d) Measurements from a larger hard disk have sample mean 0.280 and standard deviation 0.05. Which 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) draw a boxplot.

UNIVARIATE PROCEDURE
VARIABLE = STRENGTH

Moments

N	30	Sum Wgts	30
Mean	1908.767	Sum	57263
Std Dev	327.115	Variance	107004.3
Skewness	1.118406	Kurtosis	2.883349
USS	1.124E8	CSS	3103123
CV	17.13751	Std Mean	59.72276

Quantiles (Def = 5)

100%	Max	2983	99%	2983
75%	Q3	2061	95%	2403
50%	Med	1863	90%	2301
25%	Q1	1712	10%	1561.5
0%	Min	1325	5%	1419
			1%	1325

Range	1658
Q3-Q1	349

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

2.75 Civil engineers must monitor flow on rivers wher power is generated. The following are the daily mea flow rates in millions of gallons per day (MGD) c the Namekagon River during the month of May fo 47 years.

602.0	517.5	572.5	392.4	505.8	547.5	389.1	497.
794.8	657.6	904.7	595.5	611.9	482.9	698.6	606.
986.4	567.7	400.1	634.9	448.4	479.1	1156.0	718.
575.6	743.3	1146.0	461.6	644.0	480.8	429.1	626.
833.9	889.0	752.6	516.5	817.2	895.8	572.2	563.
679.3	738.0	618.9	390.8	550.9	425.9	760.6	

(a) Obtain the quartiles.

(b) Obtain the 90th 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[7] 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.

[7]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

Absolute variation 28
Arithmetic mean 24
Bar chart 12
Boxplot 31
Categorical distribution 14
Class boundary 16
Class frequency 15
Class interval 16
Class limit 14
Class mark 16
Coefficient of variation 29
Cumulative distribution 16
Cumulative percentage distribution 17
Density histogram 19
Deviation from the mean 27
Dot diagram 13
Double-stem display 24
Empirical cumulative distribution 24
Endpoint convention 15
Exploratory data analysis 22
Five-stem display 24
Frequency distribution 14
Histogram 17
Interquartile range 31
Leaf 21
Maximum 31
Mean 24
Median 25
Minimum 31
Modified boxplot 31
Numerical distribution 14
Ogive 20
Outlier 14
Pareto diagram 12
Percentage distribution 17
Percentile 29
Pie chart 23
Quartile 30
Range 31
Raw data 34
Relative variation 28
Sample mean 25
Sample variance 27
Standard deviation 28
Stem 21
Stem-and-leaf display 21
Stem label 21
Variance 27
Weighted mean 38

CHAPTER 3

PROBABILITY

CHAPTER OUTLINE

3.1 Sample Spaces and Events 45
3.2 Counting 49
3.3 Probability 55
3.4 The Axioms of Probability 57
3.5 Some Elementary Theorems 60
3.6 Conditional Probability 66
3.7 Bayes' Theorem 72
Review Exercises 78
Key Terms 80

In the study of probability there are basically three kinds of questions: (1) What do we mean when we say that the probability of an event is, say, 0.50, 0.02, or 0.81? (2) How are the numbers we call probabilities determined, or measured in actual practice? (3) What are the mathematical rules that probabilities must obey?

After some mathematical preliminaries in Sections 3.1 and 3.2, we study the first two kinds of questions in Section 3.3 and the third kind of question in Sections 3.4 through 3.7.

3.1 Sample Spaces and Events

Probability allows us to quantify the variability in the outcome of any experiment whose exact outcome cannot be predicted with certainty. However, before we can introduce probability, it is necessary to specify the space of outcomes and the events on which it will be defined.

In statistics, a set of all possible outcomes of an experiment is called a **sample space**, because it usually consists of all the things that can happen when one takes a sample. Sample spaces are usually denoted by the letter S. To avoid misunderstandings about the words *experiment* and *outcome* as we have used them here, it should be understood that statisticians use these terms in a very wide sense. An **experiment** may consist of the simple process of noting whether a switch is turned on or off; it may consist of determining the time it takes a car to accelerate to 30 miles per hour; or it may consist of the complicated process of finding the mass of an electron. Thus, the **outcome** of an experiment may be a simple choice between two possibilities; it may be the result of a direct measurement or count, or it may be an answer obtained after extensive measurements and calculations.

When we study the outcomes of an experiment, we usually identify the various possibilities with numbers, points, or some other kinds of symbols. For instance, if four contractors bid on a highway construction job and we let a, b, c, and d denote that it is awarded to Mr. Adam, Mrs. Brown, Mr. Clark, or Ms. Dean, then the sample space for this experiment is the set $S = \{a, b, c, d\}$.

Also, if a government agency must decide where to locate two new computer research facilities and that (for a certain purpose) it is of interest to indicate how many of them will be located in Texas and how many in California, we can write the sample space as

$$S = \{(0, 0), (1, 0), (0, 1), (2, 0), (1, 1), (0, 2)\}$$

where the first coordinate is the number of research facilities that will be located in Texas and the second coordinate is the number that will be located in California. Geometrically, this sample space may be pictured as in Figure 3.1, from which it is apparent, for example, that in two of the six possibilities Texas and California will get an equal number of the new research facilities.

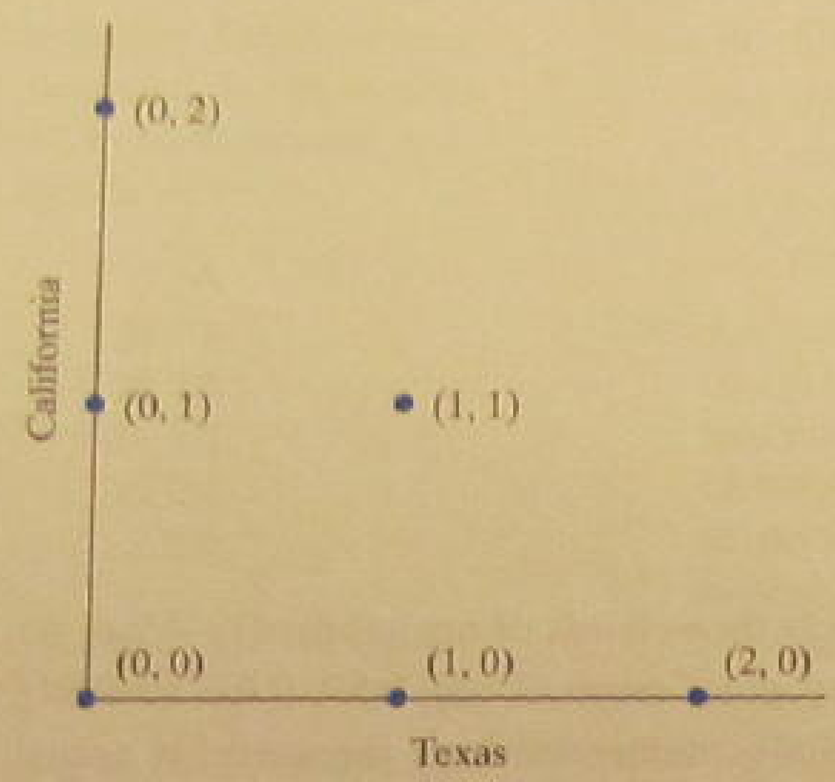


Figure 3.1
Sample space for the number of new computer research facilities to be located in Texas and in California

The use of points rather than letters or numbers has the advantage that it makes it easier to visualize the various possibilities, and perhaps discover some special features which several of the outcomes may have in common.

Generally, sample spaces are classified according to the number of elements (points) that they contain. In the two preceding examples, the sample spaces had four and six elements, and they are both referred to as **finite sample spaces**. Other examples of finite sample spaces are the one for the various ways in which a president and a vice president can be selected from among the 25 members of a union local and the one for the various ways in which a student can answer the 12 questions on a true-false test. As we see on page 50, the first of these sample spaces has 600 elements and the other has 4,096.

The following are examples of sample spaces that are not finite. If persons checking the nitrogen-oxide emission of cars are interested in the number of cars they have to inspect before they observe the first one that does not meet government regulations, it could be the first, the second, . . . , the fiftieth, . . . , and for all we know they may have to check thousands of cars before they find one that does not meet government regulations. Not knowing how far they may have to go, it is appropriate in an example like this to take as the sample space the whole set of natural numbers, of which there is a countable infinity. To go one step further, if they were interested in the nitrogen oxide emission of a given car in grams per mile, the sample space would have to consist of all the points on a continuous scale (a certain interval on the line of real numbers), of which there is a continuum.

In general, a sample space is said to be a **discrete sample space** if it has finitely many or a countable infinity of elements. If the elements (points) of a sample space constitute a continuum—for example, all the points on a line, all the points on a line segment, or all the points in a plane—the sample space is said to be a **continuous sample space**.

In the remainder of this chapter we shall consider only discrete and mainly finite sample spaces.

In statistics, any subset of a sample space is called an **event**. By subset we mean any part of a set, including the whole set and, trivially, a set called the **empty set** and denoted by ϕ, which has no elements at all. For instance, with reference to Figure 3.1,

$$C = \{(1, 0), (0, 1)\}$$

is the event that, between them, Texas and California will get one of the two research facilities,

$$D = \{(0, 0), (0, 1), (0, 2)\}$$

is the event that Texas will not get either of the two research facilities, and

$$E = \{(0, 0), (1, 1)\}$$

is the event that Texas and California will get an equal number of the facilities. Note that events C and E have no elements in common—they are **mutually exclusive events**.

In many probability problems we are interested in events which can be expressed in terms of two or more events by forming **unions, intersections**, and **complements**. Although the reader must surely be familiar with these terms, let us review briefly that if A and B are any two sets in a sample space S, their union $A \cup B$ is the subset of S that contains all elements that are either in A, in B, or in both; their intersection $A \cap B$ is the subset of S that contains all elements that are in both A and B; and the complement $\overline{A}$ of A is the subset of S that contains all the elements of S that are not in A.

EXAMPLE **Combining events by union, intersection, and complement**

With reference to the sample space of Figure 3.1 and the events C, D, and E just defined, list the outcomes comprising each of the following events and also express the events in words:

(a) $C \cup E$;

(b) $C \cap D$;

(c) $\overline{D}$.

Solution **(a)** Since $C \cup E$ contains all the elements that are in C, in E, or in both,

$$C \cup E = \{(1, 0), (0, 1), (0, 0), (1, 1)\}$$

is the event that neither Texas nor California will get both of the new research facilities.

(b) Since $C \cap D$ contains all the elements that are in both C and D,

$$C \cap D = \{(0, 1)\}$$

is the event that Texas will not get either of the two new facilities and California will get one.

(c) Since $\overline{D}$ contains all the elements of the sample space that are not in D,

$$\overline{D} = \{(1, 0), (1, 1), (2, 0)\}$$

is the event that Texas will get at least one of the new computer research facilities.

Sample spaces and events, particularly relationships among events, are often depicted by means of **Venn diagrams** like those of Figures 3.2–3.4. In each case the sample space is represented by a rectangle, whereas events are represented by regions within the rectangle, usually by circles or parts of circles. The shaded regions of the four Venn diagrams of Figure 3.2 represent event A, the complement of event A, the union of events A and B, and the intersection of events A and B.

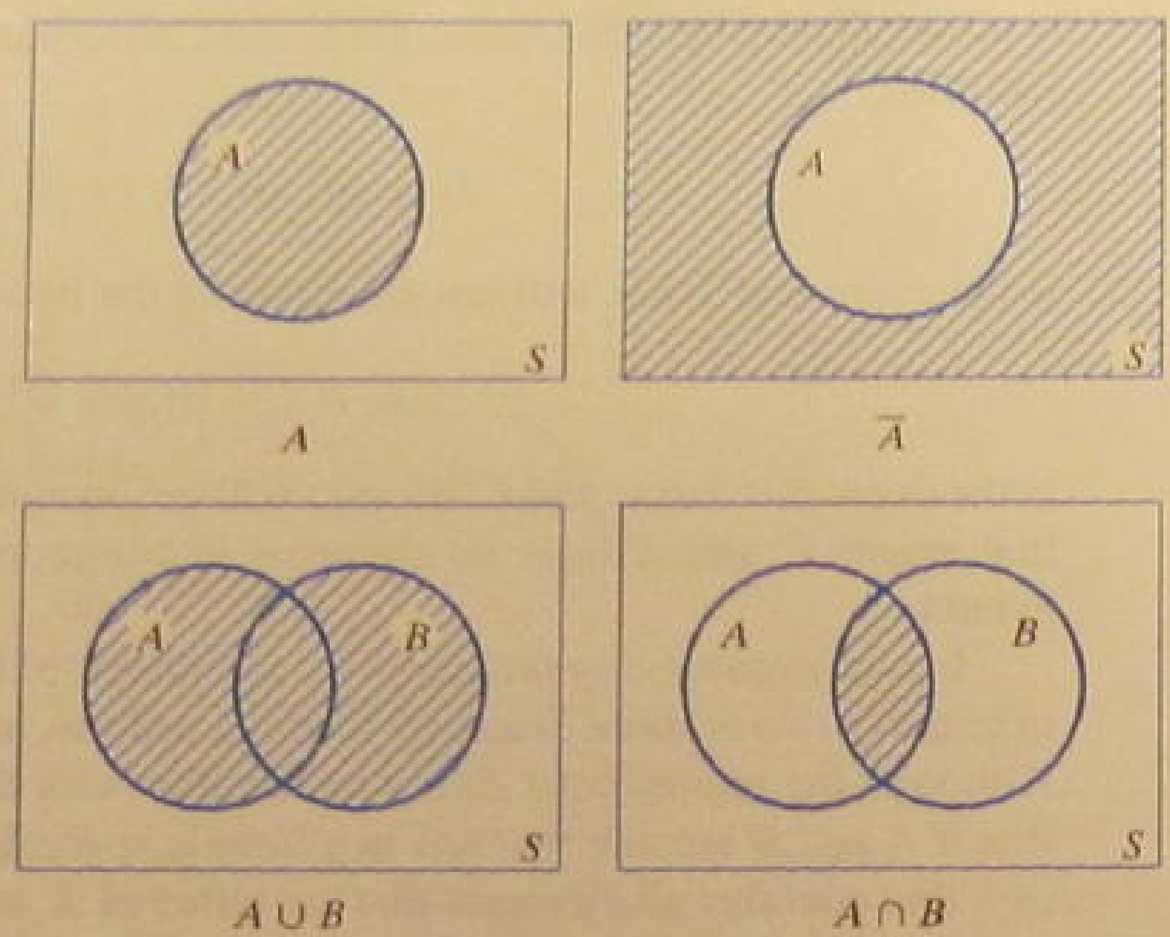


Figure 3.2
Venn diagrams showing complement, union, and intersection

EXAMPLE **Relation of regions in Venn diagrams to events**

If A is the event that a certain student is taking a course in calculus and B is the event that the student is taking a course in applied mechanics, what events are represented by the shaded regions of the four Venn diagrams of Figure 3.2?

Solution The shaded region of the first diagram represents the event that the student is taking a course in calculus; that of the second diagram represents the event that the student is not taking a course in calculus; that of the third diagram represents the event that the student is taking a course in calculus and/or a course in applied mechanics; and that of the fourth diagram represents the event that the student is taking a course in calculus as well as a course in applied mechanics. ■

Venn diagrams are often used to verify relationships among sets, thus making it unnecessary to give formal proofs based on the algebra of sets. To illustrate, let us show that $\overline{A \cup B} = \overline{A} \cap \overline{B}$, which expresses the fact that the complement of the union of two sets equals the intersection of their complements. To begin, note that the shaded region of the first Venn diagram of Figure 3.3 represents the set $\overline{A \cup B}$ (compare this diagram with the third diagram of Figure 3.2). The cross-hatched region of the second Venn diagram of Figure 3.3 was obtained by shading the region representing $\overline{A}$ with lines going in one direction and that representing $\overline{B}$ with lines going in another direction. Thus, the cross-hatched region represents the intersection of $\overline{A}$ and $\overline{B}$, and it can be seen that it is identical with the shaded region of the first Venn diagram of Figure 3.3.

When we deal with three events, we draw the circles as in Figure 3.4. In this diagram, the circles divide the sample space into eight regions, numbered 1 through 8, and it is easy to determine whether the corresponding events are parts of A or $\overline{A}$, B or $\overline{B}$, and C or $\overline{C}$.

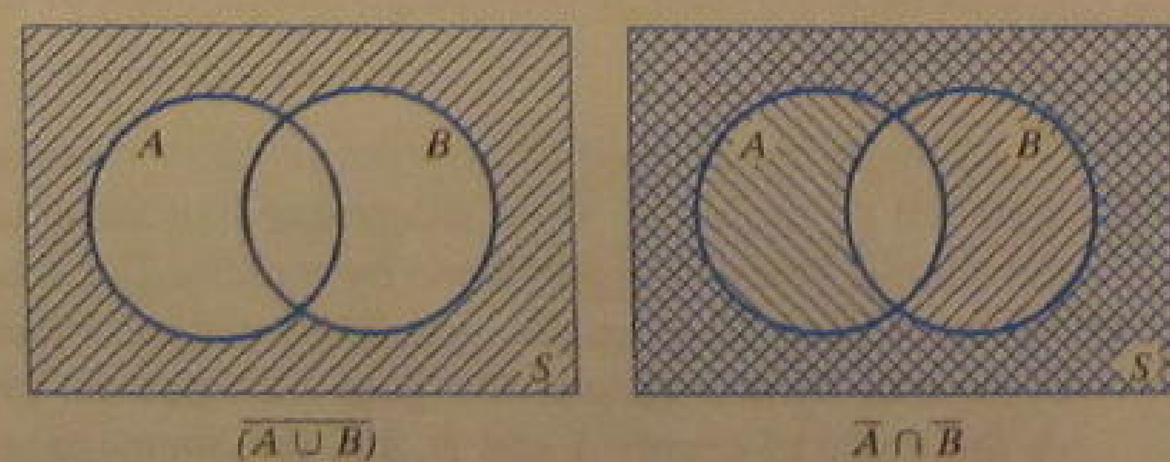


Figure 3.3
Use of Venn diagrams to show that $\overline{A \cup B} = \overline{A} \cap \overline{B}$

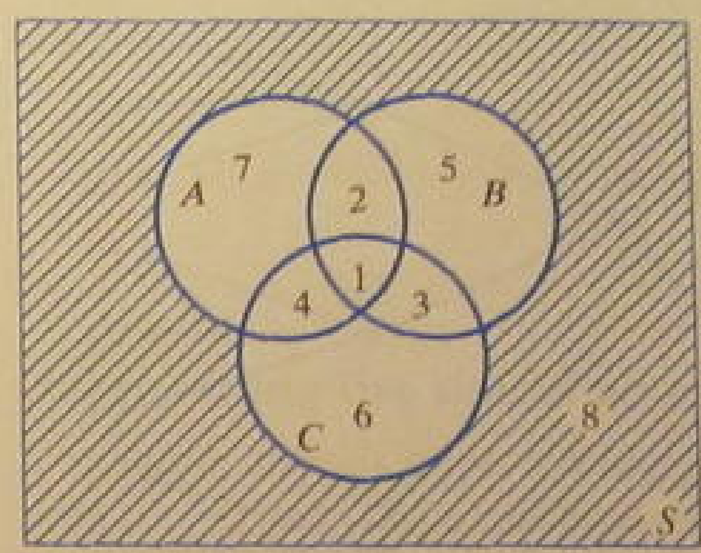


Figure 3.4
Venn diagram

EXAMPLE **Relating events to regions of the Venn diagram**

A manufacturer of small motors is concerned with three major types of defects. If A is the event that the shaft size is too large, B is the event that the windings are improper, and C is the event that the electrical connections are unsatisfactory, express in words what events are represented by the following regions of the Venn diagram of Figure 3.4:

(a) region 2
(b) regions 1 and 3 together
(c) regions 3, 5, 6, and 8 together

Solution

(a) Since this region is contained in A and B but not in C, it represents the event that the shaft is too large and the windings improper, but the electrical connections are satisfactory.

(b) Since this region is common to B and C, it represents the event that the windings are improper and the electrical connections are unsatisfactory.

(c) Since this is the entire region outside A, it represents the event that the shaft size is not too large. ■

3.2 Counting

At times it can be quite difficult, or at least tedious, to determine the number of elements in a finite sample space by direct enumeration. To illustrate, suppose all newer used cars in a large city can be classified as low, medium, or high current mileage; moderate or high priced; and be inexpensive, average, or expensive to operate. In how many ways can a used car be categorized?

Clearly, there are many possibilities; a used car can have low current mileage, be moderately priced, and be inexpensive to operate; have neither low or high mileage, be high priced, and be average cost to operate; and so on. Continuing in this way, we may be able to list all 18 possibilities, but the chances are that we will omit at least one or two.

To handle this kind of problem systematically, it helps to draw at **tree diagram** like that of Figure 3.5, where the three alternatives for current mileage are denoted by M_1, M_2, and M_3, where M_1 is low mileage. The price is either P_1 or P_2, where P_1 is moderate; and the three alternatives for operating costs are denoted by C_1, C_2, and C_3, where C_1 is inexpensive. Following a given path from left to right along the branches of the tree, we obtain a particular categorization, namely a particular element of the sample space. It can be seen that all together there are 18 possibilities.

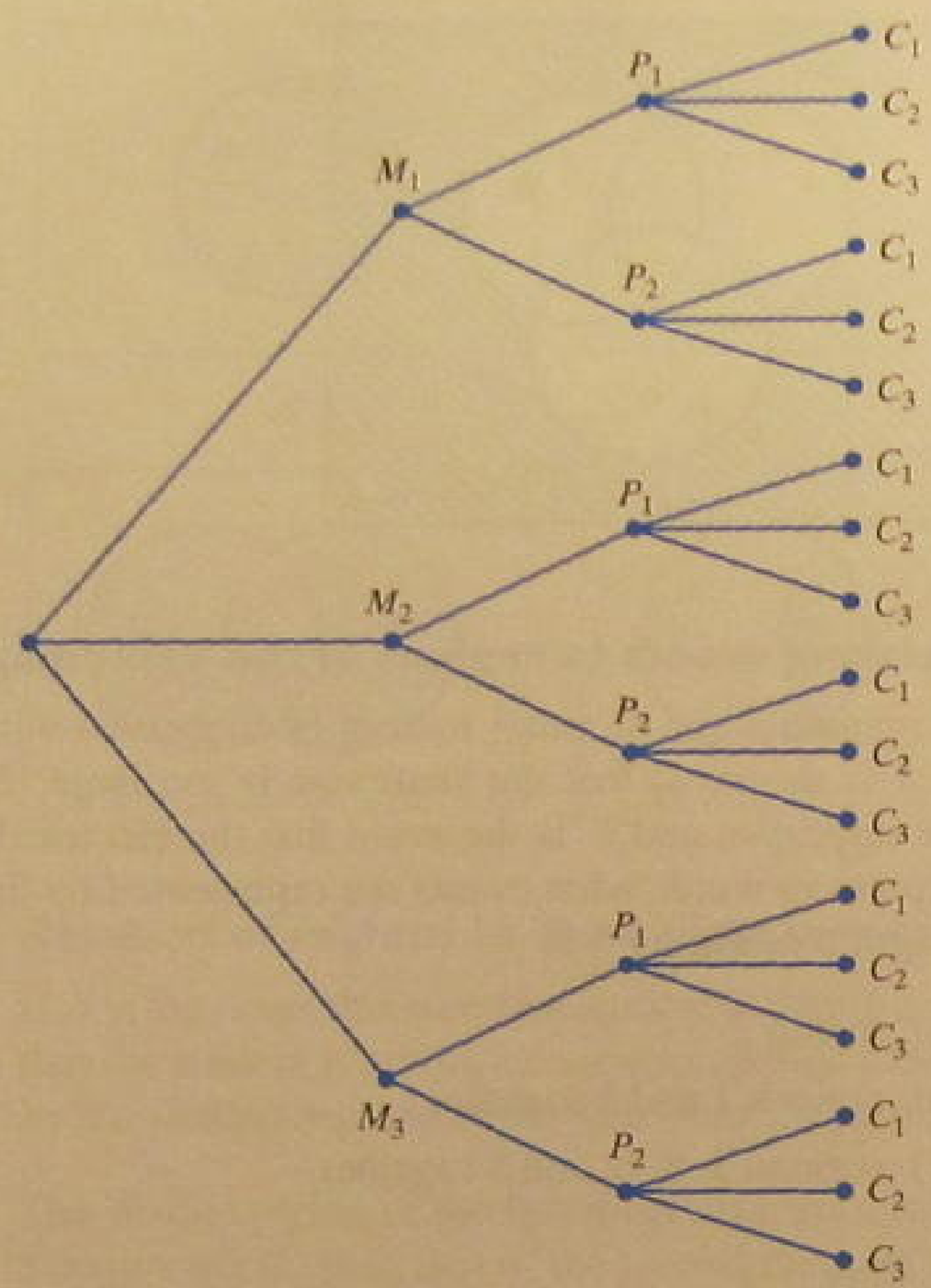


Figure 3.5
Tree diagram for used cars

This result could also have been obtained by observing that there are three M-branches, that each M-branch forks into two P-branches, and that each P-branch forks into three C-branches. Thus, there are $3 \cdot 2 \cdot 3 = 18$ combinations of branches, or paths. This result is a special case of the following theorem often called the fundamental theorem of counting.

Multiplication of choices

Theorem 3.1 If sets $A_1, A_2, \cdots, A_k$ contain, respectively, $n_1, n_2, \cdots, n_k$ elements, there are $n_1 \cdot n_2 \cdots n_k$ ways of choosing first an element of A_1, then an element of $A_2, \cdots$, and finally an element of A_k.

In our example we had $n_1 = 3$, $n_2 = 2$, and $n_3 = 3$, and hence, $3 \cdot 2 \cdot 3 = 18$ possibilities.

EXAMPLE **The multiplication rule for $k = 2$ stages of choices**

In how many different ways can a union local with a membership of 25 choose a vice president and a president?

Solution Since the vice president can be chosen in 25 ways and, subsequently, the president in 24 ways, there are altogether $25 \cdot 24 = 600$ ways in which the whole choice can be made. ■

EXAMPLE **The multiplication rule with $k = 12$ stages of choices**

If a test consists of 12 true-false questions, in how many different ways can a student mark the test paper with one answer to each question?

Solution Since each question can be answered in two ways, there are all together

$$2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^{12} = 4{,}096 \text{ possibilities}$$

EXAMPLE **Determining the size of an experiment**

A manufacturer is experiencing difficulty getting consistent readings of tens strength between three machines located on the production floor, research lab, a quality control lab, respectively. There are also four possible technicians—Tom, J Ken, and Carol—who operate at least one of the test machines regularly.

(a) How many operator-machine pairs must be included in a designed experime where every operator tries every machine?

(b) If each operator-machine pair is required to test eight specimens, how many test specimens are required for the entire procedure? Note: A specimen is destroyed when its tensile strength is measured.

Solution **(a)** There are $n_1 = 4$ operators and $n_2 = 3$ machines, so $4 \cdot 3 = 12$ pairs are required.

(b) There are $n_3 = 8$ test specimens required for each operator-machine pair, so $8 \cdot 12 = 96$ test specimens are required for the designed experiment.

As in the first of these three examples, the rule for the multiplication of choi is often used when several choices are made from one set and we are concer with the order in which they are made. In general, if r objects are chosen fro set of n distinct objects, any particular arrangement, or order, of these object called a **permutation**. For instance, 4 1 2 3 is a permutation of the first four posit integers, and Maine, Vermont, and Connecticut is a permutation, a particular orde arrangement, of three of the six New England states.

To find a formula for the total number of permutations of r objects selected fr a set of n distinct objects, we observe that the first selection is made from the wh set of n objects, the second selection is made from the $n - 1$ objects which rem after the first selection has been made, . . . , and the rth selection is made from $n - (r - 1) = n - r + 1$ objects which remain after the first $r - 1$ selections h been made. Therefore, by the rule for the multiplication of choices, the total num of permutations of r objects selected from a set of n distinct objects is

$${}_nP_r = n(n-1)(n-2)\cdots(n-r+1)$$

for $r = 1, 2, \ldots, n$.

Since products of consecutive integers arise in many problems relating to p mutations or other kinds of special selections, it will be convenient to introduce h the **factorial notation**, where $1! = 1$, $2! = 2 \cdot 1 = 2$, $3! = 3 \cdot 2 \cdot 1 = 6$, $4!$ $4 \cdot 3 \cdot 2 \cdot 1 = 24, \ldots$, and in general $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$. Also, to ma various formulas more generally applicable, we let $0! = 1$ by definition.

To express the formula for ${}_nP_r$ in terms of factorials, we multiply and divide $(n-r)!$, getting

$${}_nP_r = \frac{n(n-1)(n-2)\cdots(n-r+1)(n-r)!}{(n-r)!} = \frac{n!}{(n-r)!}$$

To summarize:

Number of permutations of n objects taken r at a time

Theorem 3.2 The number of permutations of r objects selected from a set of n distinct objects is

$$_nP_r = n(n-1)(n-2)\cdots(n-r+1)$$

or, in factorial notation,

$$_nP_r = \frac{n!}{(n-r)!}$$

Note that the second formula also holds for $r = 0$.

EXAMPLE **The number of ways to assemble chips in a controller**

An electronic controlling mechanism requires 5 distinct, but interchangeable, memory chips. In how many ways can this mechanism be assembled by placing the 5 chips in the 5 positions within the controller?

Solution For $n = 5$ and $r = 5$, the first formula yields

$$_5P_5 = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

and the second formula yields

$$_5P_5 = \frac{5!}{(5-5)!} = \frac{5!}{0!} = 5! = 120$$

The first formula for $_nP_r$ is generally easier to use unless we can use a calculator which directly yields factorials and/or ratios of factorials. ■

There are many problems in which we must find the number of ways in which r objects can be selected from a set of n objects, but we do not care about the order in which the selection is made. For instance, we may want to know in how many ways 3 of 20 laboratory assistants can be chosen to assist with an experiment. In general, there are $r!$ permutations of any r objects we select from a set of n distinct objects. So, the $_nP_r$ permutations of r objects, selected from a set of n objects, contains each set of r objects $r!$ times. Therefore, to find the number of ways in which r objects can be selected from a set of n distinct objects, also called the number of **combinations** of n objects taken r at a time and denoted by $_nC_r$ or $\binom{n}{r}$, we divide $_nP_r$ by $r!$ and get

Number of combinations of n objects taken r at a time

Theorem 3.3 The number of ways in which r objects can be selected from a set of n distinct objects is

$$\binom{n}{r} = \frac{n(n-1)(n-1)\cdots(n-r+1)}{r!}$$

or, in factorial notation,

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

EXAMPLE **Evaluating a combination**

In how many different ways can 3 of 20 laboratory assistants be chosen to assist with an experiment?

Solution For $n = 20$ and $r = 3$, the first formula for $\binom{n}{r}$ yields

$$\binom{20}{3} = \frac{20 \cdot 19 \cdot 18}{3!} = 1{,}140$$

■

EXAMPLE **Selection of machines for an experiment**

A calibration study needs to be conducted to see if the readings on 15 test machines are giving similar results. In how many ways can 3 of the 15 be selected for the initial investigation?

Solution

$$\binom{15}{3} = \frac{15 \cdot 14 \cdot 13}{3 \cdot 2 \cdot 1} = 455 \text{ ways}$$

Note that selecting which 3 machines to use is the same as selecting which 12 not to include. That is, according to the second formula,

$$\binom{15}{12} = \frac{15!}{12!\,3!} = \frac{15!}{3!\,12!} = \binom{15}{3}$$

■

EXAMPLE **The number of choices of new researchers**

In how many different ways can the director of a research laboratory choose 2 chemists from among 7 applicants and 3 physicists from among 9 applicants?

Solution The 2 chemists can be chosen in $\binom{7}{2} = 21$ ways and the 3 physicists can be chosen in $\binom{9}{3} = 84$ ways. By the multiplication rule, the whole selection can be made in $21 \cdot 84 = 1{,}764$ ways. ■

Exercises

3.1 An environmental engineer suspects mercury contamination in an area which contains three lakes and two streams. He will check all five for mercury contamination.

(a) Express each outcome using two coordinates, so (2, 1), for example, represents the event that two of the lakes and one of the streams will be contaminated. Draw a diagram similar to that of Figure 3.1 showing the 12 outcomes in the sample space.

(b) If R is the event that equally many lakes and streams are contaminated, T is the event that none of the streams is contaminated, and U is the event that fewer lakes than streams are contaminated, express each of these events symbolically by listing its elements.

3.2 With reference to Exercise 3.1, which of the three pairs of events, R and T, R and U, and T and U, are mutually exclusive?

3.3 With reference to Exercise 3.1, list the outcomes comprising each of the following events, and also express the events in words.

(a) $R \cup U$

(b) $R \cap T$

(c) $\overline{T}$

3.4 With reference to the sample space of Figure 3.1, express each of the following events in words.

(a) $F = \{(1, 0), (1, 1)\}$

(b) $G = \{(0, 2), (1, 1), (2, 0)\}$

(c) $F \cap G$

3.5 To construct sample spaces for experiments in which we deal with nonnumerical data, we often code the various alternatives by assigning them numbers. For instance, if a mechanic is asked whether work on a certain model car is very easy, easy, average, difficult, or very difficult, we might assign these alternatives the codes, 1, 2, 3, 4, and 5. If $A = \{3, 4\}$, $B = \{2, 3\}$, and $C = \{4, 5\}$, express each of the following symbolically by listing its elements and also in words.

(a) $A \cup B$

(b) $A \cap B$

(c) $A \cup \overline{B}$

(d) $\overline{C}$

3.6 With reference to Exercise 3.5, which of the three pairs of events, A and B, A and C, and B and C, are mutually exclusive?

3.7 Two professors and 3 graduate assistants are responsible for the supervision of a physics lab, and at least one professor and one graduate assistant have to be present at all times.

(a) Using two coordinates so that (1, 3), for example, represents the event that one professor and 3 graduate assistants are present, draw a diagram similar to that of Figure 3.1 showing the points of the corresponding sample space.

(b) Describe in words the events which are represented by $B = \{(1, 3), (2, 3)\}$, $C = \{(1, 1), (2, 2)\}$, and $D = \{(1, 2), (2, 1)\}$.

(c) With reference to part (b), express $C \cup D$ symbolically by listing its elements, and also express this event in words.

(d) With reference to part (b), are B and D mutually exclusive?

3.8 For each of the following experiments, decide whether it would be appropriate to use a sample space which is finite, countably infinite, or continuous.

(a) The amount of cosmic radiation to which passengers are exposed during a transcontinental jet flight is measured by means of a suitable counter.

(b) Five of the members of a professional society with 12,600 members are chosen to serve on a nominating committee.

(c) An experiment is conducted to measure the heat of vaporization of an unknown liquid.

(d) A study is made to determine in how many of 450 airplane accidents the main cause is pilot error.

(e) Measurements are made to determine the uranium content of a certain ore.

(f) In a torture test, a watch is dropped a number of times from a tall building until it stops running.

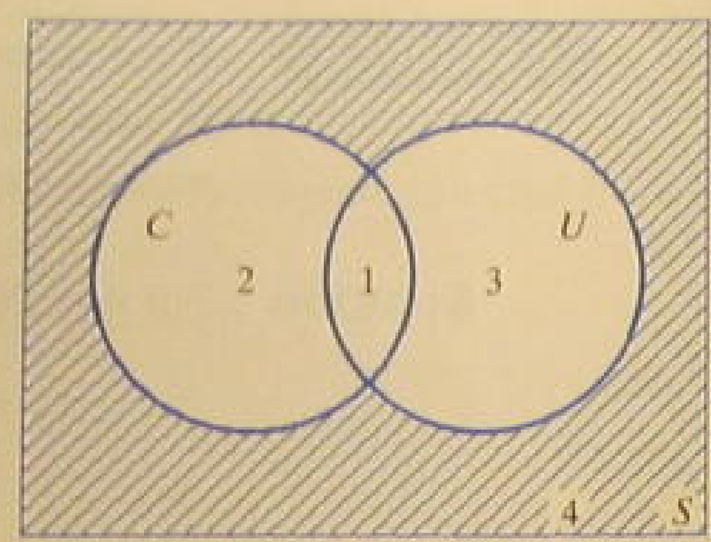


Figure 3.6 Venn diagram for Exercises 3.9 and 3.10

3.9 In Figure 3.6, C is the event that an ore contains copper and U is the event that it contains uranium. Explain in words what events are represented by regions 1, 2, 3, and 4.

3.10 With reference to Exercise 3.9, what events are represented by

(a) regions 1 and 3 together;

(b) regions 3 and 4 together;

(c) regions 1, 2, and 3 together?

3.11 With reference to Figure 3.4, what events are represented by

(a) region 5;

(b) regions 4 and 6 together;

(c) regions 7 and 8 together;

(d) regions 1, 2, 3 and 5 together?

3.12 With reference to Figure 3.4, what regions or combinations of regions represent the events that a motor will have

(a) none of the major defects;

(b) a shaft that is large and windings improper;

(c) a shaft that is large and/or windings improper but the electrical connections are satisfactory;

(d) a shaft that is large and the windings improper and/or the electrical connections are unsatisfactory?

3.13 Use Venn diagrams to verify that

(a) $\overline{A \cap B} = \overline{A} \cup \overline{B}$

(b) $A \cup (A \cap B) = A$

(c) $(A \cap B) \cup (A \cap \overline{B}) = A$

(d) $A \cup B = (A \cap B) \cup (A \cap \overline{B}) \cup (\overline{A} \cap B)$

(e) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

3.14 A building inspector has to check the wiring in a new apartment building either on Monday, Tuesday, Wednesday, or Thursday, and at 8 A.M., 1 P.M., or 2 P.M. Draw a tree diagram which shows the various ways in which the inspector can schedule the inspection of the wiring of the new apartment building.

3.15 If the five finalists in an international volleyball tournament are Spain, the United States, Uruguay, Portugal, and Japan, draw a tree diagram that shows the various possible first- and second-place finishers.

3.16 A biomechanical device for medical emergencies can operate for 0, 1, or 2 times a night. Construct a tree diagram to show that there are 10 different ways that it can be operated for a total of 6 times over 4 nights.

3.17 In an optics kit there are 6 concave lenses, 4 convex lenses, and 3 prisms. In how many ways can one choose one of the concave lenses, one of the convex lenses, and one of the prisms?

3.18 A special-purpose computer has 2 switches, each of which can be set in 3 different positions, and 1 switch that can be set in 2 positions. In how many ways can the computer's switches be set?

3.19 One engineering group consists of 6 men and 4 women.

(a) How many different project teams can be formed consisting of 2 men and 2 women?

(b) If 2 women have the same boyfriend and refuse to be on the same team together, how many different project teams can be formed consisting of 2 men and 2 women?

3.20 If there are 9 cars in a race, in how many different ways can they place first, second, and third?

3.21 In how many ordered ways can a television director schedule 6 different commercials during the 6 time slots allocated to commercials during the telecast of the first period of a hockey game?

3.22 If among n objects k are alike and the others are all distinct, the number of permutations of these n objects taken all together is $n!/k!$.

(a) How many permutations are there of the letters of the word *class*?

(b) In how many ways can the television director of Exercise 3.21 fill the 6 time slots allocated to commercials, if there are 4 different commercials, of which a given one is to be shown 3 times while each of the others is to be shown once?

3.23 Determine the number of ways in which a manufacturer can choose 2 of 15 locations for a new warehouse.

3.24 If the order does not matter, in how many different ways can 4 of 18 robotic arms be chosen for a special welding job?

3.25 A carton of 12 rechargeable batteries contains one that is defective. In how many ways can an inspector choose 3 of the batteries and

(a) get the one that is defective;

(b) not get the one that is defective?

3.26 With reference to Exercise 3.25, suppose that two of the batteries are defective. In how many ways can the inspector choose 3 of the batteries and get

(a) none of the defective batteries;

(b) one of the defective batteries;

(c) both of the defective batteries?

3.27 The supply department has 8 different electric motors and 5 different starting switches. In how many ways can 2 motors and 2 switches be selected for an experiment concerning a tracking antenna?

3.3 Probability

So far we have studied only what is possible in a given situation; now we shall go one step further and judge also what is probable and what is improbable. Historically, the oldest way of measuring uncertainties is the **classical probability concept**, which was developed originally in connection with games of chance. It applies when all possible outcomes are equally likely.

The classical probability concept

If there are m equally likely possibilities, of which one must occur and s are regarded as favorable, or as a "success," then the probability of a "success" is given by $\frac{s}{m}$.

In the application of this rule, the terms *favorable* and *success* are used rather loosely—favorable may mean that a television set does not work and success may mean that someone catches the flu.

EXAMPLE **Well-shuffled cards are equally likely to be selected**

What is the probability of drawing an ace from a well-shuffled deck of 52 playing cards?

Solution There are $s = 4$ aces among the $m = 52$ cards, so we get

$$\frac{s}{m} = \frac{4}{52} = \frac{1}{13}$$

■

Although equally likely possibilities are found mostly in games of chance, the classical probability concept applies also to a great variety of situations where gambling devices are used to make random selections; say, when offices are assigned to research assistants by lot, when laboratory animals are chosen for an experiment so that each one has the same chance of being selected, or when washing-machine parts are chosen for inspection so that each part produced has the same chance of being selected.

EXAMPLE **Random selection results in the equally likely case**

The next generation of miniaturized wireless capsules with active locomotion will require two miniture electric[1] motors to maneuver each capsule. Suppose 10 motors have been fabricated but that, in spite of tests performed on the individual motors, 2 will not operate satisfactorily when placed into a capsule.

To fabricate a new capsule, 2 motors will be randomly selected (that is, each pair of motors has the same chance of being selected). Find the probability that

(a) both motors will operate satisfactorily in the capsule

(b) one motor will operate satisfactorily and the other will not

Solution **(a)** There are $\binom{10}{2} = 45$ equally likely ways of choosing 2 of 10 motors, so $m = 45$.

The number of favorable outcomes is the number of ways in which two good motors can be selected from eight:

$$s = \binom{8}{2} = 28$$

so the probability that both motors will operate satisfactorily in the capsule is

$$\frac{s}{m} = \frac{28}{45}$$

or approximately 0.622.

(b) The number of favorable outcomes is the number of ways in which one satisfactory motor and one unsatisfactory motor can be selected, or

$$s = \binom{8}{1}\binom{2}{1} = 8 \cdot 2 = 16$$

It follows that the probability is

$$\frac{s}{m} = \frac{16}{45} = 0.356$$

■

[1] M. Quirini et al. 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.

A major shortcoming of the classical probability concept is its limited applicability, for there are many situations in which the various possibilities cannot all be regarded as equally likely. This would be the case, for example, if we are concerned with the question of whether it will rain the next day, whether a missile launching will be a success, whether a newly designed engine will function for at least 1,000 hours, or whether a certain candidate will win an election.

Among the various probability concepts, most widely held is the **frequency interpretation**.

The frequency interpretation of probability

The probability of an event (or outcome) is the proportion of times the event will occur in a long run of repeated experiments.

If we say that the probability is 0.78 that a jet from New York to Boston will arrive on time, we mean that such flights arrive on time 78% of the time. Also, if the Weather Service predicts that there is a 40% chance for rain (that the probability is 0.40), this means that under the same weather conditions it will rain 40% of the time. More generally, we say that an event has a probability of, say, 0.90, in the same sense in which we might say that in cold weather our car will start 90% of the time. We cannot guarantee what will happen on any particular occasion—the car may start and then it may not—but if we kept records over a long period of time, we should find that the proportion of "successes" is very close to 0.90.

In accordance with the frequency interpretation of probability, we estimate the probability of an event by observing what fraction of the time similar events have occurred in the past.

EXAMPLE **Long-run relative frequency approximation to probability**

If records show that 294 of 300 ceramic insulators tested were able to withstand a certain thermal shock, what is the probability that any one untested insulator will be able to withstand the thermal shock?

Solution Among the insulators tested, $\frac{294}{300} = 0.98$ were able to withstand the thermal shock, and we use this figure as an estimate of the probability. ■

An alternative point of view is to interpret probabilities as *personal* or *subjective evaluations*. Such **subjective probabilities** express the strength of one's belief with regard to the uncertainties that are involved, and they apply especially when there is little or no direct evidence, so that there is no choice but to consider collateral (indirect) evidence, educated guesses, and perhaps intuition and other subjective factors. Subjective probabilities are best determined by referring to risk taking, or betting situations, as will be explained in Exercise 3.53.

3.4 The Axioms of Probability

In this section we define probabilities mathematically as the values of **additive set functions**. Since the reader is probably most familiar with functions for which the elements of the domain and the range are all numbers, let us first give a very simple example where the elements of the domain are sets, while the elements of the range are nonnegative integers, namely, a **set function** that assigns to each subset A of a finite sample space S the number of elements in A, written $N(A)$. Suppose that 500 machine parts are inspected before they are shipped, that I denotes that a machine part is improperly assembled, D denotes that it contains one or more

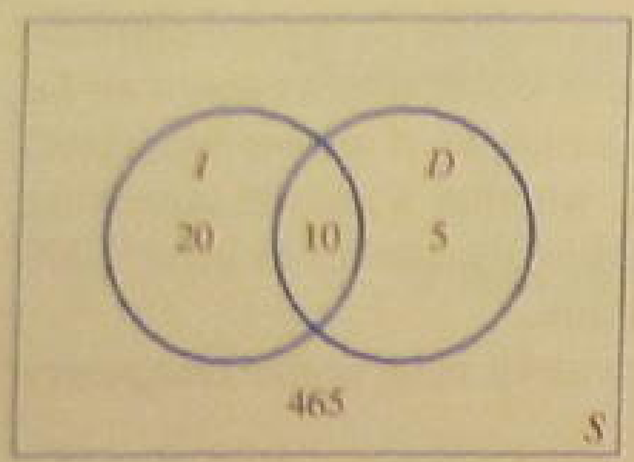


Figure 3.7
Classification of 500 machine parts

defective components, and the distribution of the 500 machine parts among the various categories is as shown in the Venn diagram of Figure 3.7.

The numbers in Figure 3.7 are $N(I \cap \overline{D}) = 20$, $N(I \cap D) = 10$, $N(\overline{I} \cap D) = 5$, and $N(\overline{I} \cap \overline{D}) = 465$. Using these values and the fact that the set function is **additive** (meaning that the number which it assigns to the union of two subsets which have no elements in common is the sum of the numbers assigned to the individual subsets), we can determine the value of $N(A)$ for any other subset A of S. For instance,

$$N(\overline{I}) = N(\overline{I} \cap D) + N(\overline{I} \cap \overline{D}) = 5 + 465 = 470$$

$$\begin{aligned} N(I \cup D) &= N(I \cap \overline{D}) + N(I \cap D) + N(\overline{I} \cap D) \\ &= 20 + 10 + 5 = 35 \end{aligned}$$

$$\begin{aligned} N(\overline{I} \cup \overline{D}) &= N(\overline{I} \cap D) + N(\overline{I} \cap \overline{D}) + N(\overline{I} \cap \overline{D}) \\ &= 10 + 5 + 465 = 480 \end{aligned}$$

and

$$N(D) = N(I \cap D) + N(\overline{I} \cap D) = 10 + 5 = 15$$

Using the concept of an additive set function, let us now explain what we mean by the probability of an event. Given a finite sample space S and an event A in S, we define $P(A)$, the probability of A, to be a value of an additive set function that satisfies the following three conditions.

The axioms of probability for a finite sample space

Axiom 1 $0 \leq P(A) \leq 1$ for each event A in S.
Axiom 2 $P(S) = 1$.
Axiom 3 If A and B are mutually exclusive events in S, then

$$P(A \cup B) = P(A) + P(B)$$

The first axiom states that probabilities are real numbers on the interval from 0 to 1, inclusive. The second axiom states that the sample space as a whole is assigned a probability of 1. Since S contains all possible outcomes, and one of these must always occur, S is certain to occur. The third axiom states that probability functions must be additive—the probability of the union is the sum of the two probabilities when the two events have no outcomes in common.

Axioms for a mathematical theory require no proof, but if such a theory is to be applied to the physical world, we must show somehow that the axioms are "realistic." Thus, let us show that the three postulates are consistent with the classical probability concept and the frequency interpretation.

So far as the first axiom is concerned, fractions of the form $\frac{s}{m}$, where $0 \leq s \leq m$ and m is a positive integer, cannot be negative or exceed 1, and the same is true also for the proportion of the time that an event will occur. To show that the second axiom is consistent with the classical probability concept and the frequency interpretation for a long series of repeated experiments, we have only to observe that for the whole sample space

$$P(S) = \frac{m}{m} = 1$$

and that some outcome must happen 100% of the time. So far as the third axiom is concerned, if

$$P(A) = \frac{s_1}{m}, \quad P(B) = \frac{s_2}{m}$$

and A and B are mutually exclusive, then

$$P(A \cup B) = \frac{s_1 + s_2}{m} = P(A) + P(B)$$

Also, if one event occurs in proportion 0.36 or 36% of the time, another event occurs 41% of the time, and the two events are mutually exclusive, then one or the other will occur in proportion $0.36 + 0.41 = 0.77$ or 77%.

Before we go any further, it is important to stress the point that the axioms of probability do not tell us how to assign probabilities to the various outcomes of an experiment; they merely restrict the ways in which it can be done. In actual practice, probabilities are assigned on the basis of past experience, on the basis of a careful analysis of conditions underlying the experiment, on the basis of subjective evaluations, or on the basis of assumptions—say, the common assumption that all the outcomes are equiprobable.

EXAMPLE **Checking possible assignments of probability**

If an experiment has the three possible and mutually exclusive outcomes A, B, and C, check in each case whether the assignment of probabilities is permissible:

(a) $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{3}$, and $P(C) = \frac{1}{3}$

(b) $P(A) = 0.64$, $P(B) = 0.38$, and $P(C) = -0.02$

(c) $P(A) = 0.35$, $P(B) = 0.52$, and $P(C) = 0.26$

(d) $P(A) = 0.57$, $P(B) = 0.24$, and $P(C) = 0.19$

Solution **(a)** The assignment of probabilities is permissible because the values are all on the interval from 0 to 1, and their sum is $\frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$.

(b) The assignment is not permissible because $P(C)$ is negative.

(c) The assignment is not permissible because $0.35 + 0.52 + 0.26 = 1.13$, which exceeds 1.

(d) The assignment is permissible because the values are all on the interval from 0 to 1 and their sum is $0.57 + 0.24 + 0.19 = 1$. ■

The approach in the last example extends to any experiment where the sample space S is discrete so the outcomes can be arranged in a sequence. An amount of probability p_i is assigned to the ith outcome, where

$$0 \le p_i \qquad \text{and} \qquad \sum_{\text{all outcomes in } S} p_i = 1$$

and then the probability of any event A is defined as

$$P(A) = \sum_{\text{all outcomes in } A} p_i$$

When probability is assigned in this manner, the axioms of probability are always satisfied.

Intuitively, we can think of the scientist as starting with a unit amount of clay (probability) and placing a proportion p_1 on the first outcome, p_2 on the second outcome, and so on. Some outcomes can be assigned a large amount and others lesser amounts. The total unit amount of clay (probability) is assigned to the outcomes in the sample space. Then, an event A is assigned the total of all the clay (probability) assigned to each outcome in A.

3.5 Some Elementary Theorems

With the use of mathematical induction, the third axiom of probability can be extended to include any number of mutually exclusive events; in other words, the following can be shown.

Generalization of the third axiom of probability

Theorem 3.4 If $A_1, A_2, \ldots, A_n$ are mutually exclusive events in a sample space S, then

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

In the next chapter we shall see how the third axiom of probability must be modified so that the axioms apply also to sample spaces which are not finite.

EXAMPLE **Probabilities add for mutually exclusive events**

The probability that a consumer testing service will rate a new antipollution device for cars very poor, poor, fair, good, very good, or excellent are 0.07, 0.12, 0.17, 0.32, 0.21, and 0.11. What are the probabilities that it will rate the device

(a) very poor, poor, fair, or good;

(b) good, very good, or excellent?

Solution Since the probabilities are all mutually exclusive, direct substitution into the formula of Theorem 3.4 yields

$$0.07 + 0.12 + 0.17 + 0.32 = 0.68$$

for part (a) and

$$0.32 + 0.21 + 0.11 = 0.64$$

for part (b). ■

As it can be shown that a sample space of n points (outcomes) has 2^n subsets, it would seem that the problem of specifying a probability function (namely, a probability for each subset or event) can easily become very tedious. Indeed, for $n = 20$ there are already more than 1 million possible events. Fortunately, this task can be simplified considerably by the use of the following theorem:

Rule for calculating probability of an event

Theorem 3.5 If A is an event in the finite sample space S, then $P(A)$ equals the sum of the probabilities of the individual outcomes comprising A.

To prove this theorem, let $E_1, E_2, \ldots, E_n$ be the n outcomes comprising event A, so that we can write $A = E_1 \cup E_2 \cup \cdots \cup E_n$. Since the E's are individual outcomes, they are mutually exclusive, and by Theorem 3.4 we have

$$\begin{aligned} P(A) &= P(E_1 \cup E_2 \cup \cdots \cup E_n) \\ &= P(E_1) + P(E_2) + \cdots + P(E_n) \end{aligned}$$

which completes the proof.

EXAMPLE **Using a Venn diagram to visualize probability calculations**

Refer to the used car classification example on page 49. Suppose that the probabilities of the 18 outcomes are as shown in Figure 3.8 (which, except for the the probabilities, is identical to Figure 3.5).

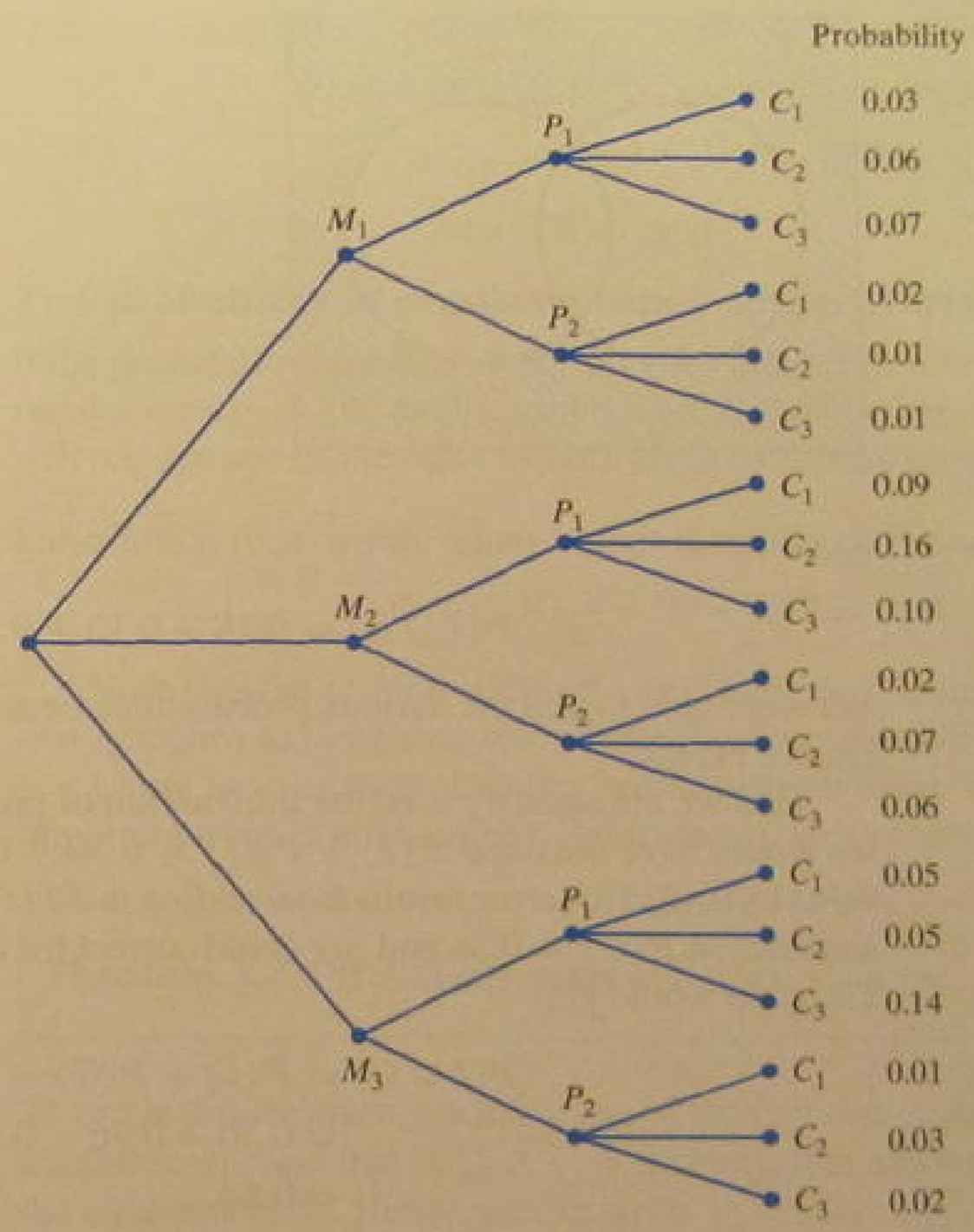


Figure 3.8
Used car classifications and their probabilities

Find $P(M_1)$, $P(P_1)$, $P(C_3)$, $P(M_1 \cap P_1)$, and $P(M_1 \cap C_3)$.

Solution Adding the probabilities of the outcomes comprising the respective events, we get

$$P(M_1) = 0.03 + 0.06 + 0.07 + 0.02 + 0.01 + 0.01 = 0.20$$
$$P(P_1) = 0.03 + 0.06 + 0.07 + 0.09 + 0.16 + 0.10 + 0.05$$
$$0.05 + 0.14 = 0.75$$
$$P(C_3) = 0.07 + 0.01 + 0.10 + 0.06 + 0.14 + 0.02 = 0.40$$
$$P(M_1 \cap P_1) = 0.03 + 0.06 + 0.07 = 0.16$$

and

$$P(M_1 \cap C_3) = 0.07 + 0.01 = 0.08$$

■

In Theorem 3.4 we saw that the third axiom of probability can be extended to include more than two mutually exclusive events. Another useful and important extension of this axiom allows us to find the probability of the union of any two events in S regardless of whether or not they are mutually exclusive. To motivate the theorem which follows, let us consider the Venn diagram of Figure 3.9, which concerns the job offers received by recent engineering-school graduates. The letters I and G stand for a job offer from industry and a job offer from the government, respectively.

It follows from the Venn diagram that

$$P(I) = 0.18 + 0.12 = 0.30$$
$$P(G) = 0.12 + 0.24 = 0.36$$

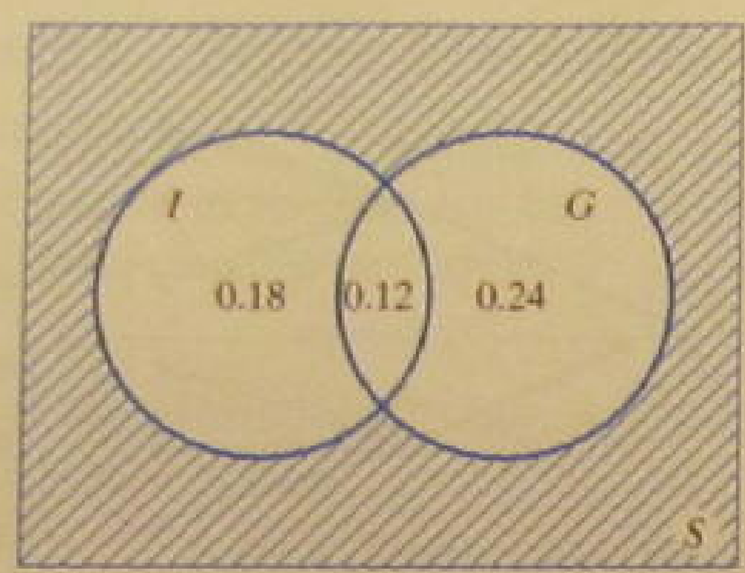


Figure 3.9
Venn diagram for job offers

and

$$P(I \cup G) = 0.18 + 0.12 + 0.24 = 0.54$$

We were able to add the various probabilities because they represent mutually exclusive events.

Had we erroneously used the third axiom of probability to calculate $P(I \cup G)$, we would have obtained $P(I) + P(G) = 0.30 + 0.36$, which exceeds the correct value by 0.12. This error results from adding in $P(I \cap G)$ twice, once in $P(I) = 0.30$ and once in $P(G) = 0.36$ and, we could correct for it by subtracting 0.12 from 0.66. Thus, we would get

$$\begin{aligned} P(I \cup G) &= P(I) + P(G) - P(I \cap G) \\ &= 0.30 + 0.36 - 0.12 \\ &= 0.54 \end{aligned}$$

and this agrees, as it should, with the result obtained before.

In line with this motivation, let us now state and prove the following theorem:

General addition rule for probability

Theorem 3.6 If A and B are any events in S, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

To prove this theorem, we make use of the identities of parts (c) and (d) of Exercise 3.13, getting

$$\begin{aligned} P(A \cup B) &= P(A \cap B) + P(A \cap \overline{B}) + P(\overline{A} \cap B) \\ &= [P(A \cap B) + P(A \cap \overline{B})] \\ &\quad + [P(A \cap B) + P(\overline{A} \cap B)] - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

where, in the third line, we add and subtract $P(A \cap B)$. Note that when A and B are mutually exclusive so that $P(A \cap B) = 0$, Theorem 3.6 reduces to the third axiom of probability. For this reason, we sometimes refer to the third axiom of probability as the **special addition rule**.

EXAMPLE **Using the general addition rule for probability**

With reference to the used car example of page 49, find the probability that a car will have high mileage or be expensive to operate, namely $P(M_1 \cup C_3)$.

Solution Making use of the results obtained on page 61, $P(M_1) = 0.20$, $P(C_3) = 0.40$, and $P(M_1 \cap C_3) = 0.08$, we substitute into the general addition rule of

Theorem 3.6 to get

$$\begin{aligned} P(M_1 \cup C_3) &= P(M_1) + P(C_3) - P(M_1 \cap C_3) \\ &= 0.20 + 0.40 - 0.08 \\ &= 0.52 \end{aligned}$$

■

EXAMPLE **The probability of requiring repair under warranty**

If the probabilities are 0.87, 0.36, and 0.29 that, while under warranty, a new car will require repairs on the engine, drive train, or both, what is the probability that a car will require one or the other or both kinds of repairs under the warranty?

Solution Substituting these given values into the formula of Theorem 3.6, we get

$$0.87 + 0.36 - 0.29 = 0.94$$

■

Note that the general addition rule, Theorem 3.6, can be generalized further so that it applies to more than two events (see Exercise 3.49).

Using axioms of probability, we can derive many other theorems which play important roles in applications. For instance, let us show the following:

Probability rule of the complement

Theorem 3.7 If A is any event in S, then $P(\overline{A}) = 1 - P(A)$.

To prove this theorem, we make use of the fact that A and $\overline{A}$ are mutually exclusive by definition, and that $A \cup \overline{A} = S$ (namely, that among them A and $\overline{A}$ contain all the elements in S). Hence we can write

$$\begin{aligned} P(A) + P(\overline{A}) &= P(A \cup \overline{A}) \\ &= P(S) \\ &= 1 \end{aligned}$$

so that $P(\overline{A}) = 1 - P(A)$. As a special case we find that $P(\phi) = 1 - P(S) = 0$ since the empty set ϕ is the complement of S.

EXAMPLE **Using the probability rule of the complement**

Referring to the used car example of page 49 and the results on page 61, find

(a) the probability that a used car will not have low mileage

(b) the probability that a used car will either not have low mileage or not be expensive to operate

Solution By the rule of the complement

(a) $P(\overline{M}_1) = 1 - P(M_1) = 1 - 0.20 = 0.80$

(b) Since $\overline{M}_1 \cup \overline{C}_3 = \overline{M_1 \cap C_3}$ according to the identity of part (a) of Exercise 3.13, by the rule of the complement we get

$$P(\overline{M}_1 \cup \overline{C}_3) = 1 - P(M_1 \cap C_3) = 1 - 0.08 = 0.92$$

■

Exercises

3.28 (a) A refrigerator manufacturer sold 2,756 units of a new model and 287 required repairs under the warranty. Estimate the probability that a new unit, which has just been sold, will require repairs under the warranty. Explain your reasoning.

(b) Last year 8,400 students applied for the 6,000 student season tickets available for football games. Next year you will apply and would like to estimate the probability of receiving a season ticket. Give your estimate and

comment on one factor that might influence your estimate.

3.29 When we roll a pair of balanced dice, what are the probabilities of getting

(a) 7;

(b) 11;

(c) 7 or 11;

(d) 3;

(e) 2 or 12;

(f) 2, 3, or 12?

3.30 A lottery sells tickets numbered from 00001 through 50000. What is the probability of drawing a number that is divisible by 200?

3.31 A car rental agency has 18 compact cars and 12 intermediate-size cars. If four of the cars are randomly selected for a safety check, what is the probability of getting two of each kind?

3.32 Among the first 842 convection ovens sold to consumers, 143 required some adjustment during the warranty period. Estimate the probability that a newly purchased convection oven will require some adjustment during the warranty period.

3.33 In a group of 160 graduate engineering students, 92 are enrolled in an advanced course in statistics, 63 are enrolled in a course in operations research, and 40 are enrolled in both. How many of these students are not enrolled in either course?

3.34 Among 150 persons interviewed as part of an urban mass transportation study, some live more than 3 miles from the center of the city (A), some now regularly drive their own car to work (B), and some would gladly switch to public mass transportation if it were available (C). Use the information given in Figure 3.10 to find

(a) $N(A)$;

(b) $N(B)$;

(c) $N(C)$;

(d) $N(A \cap B)$;

(e) $N(A \cap C)$;

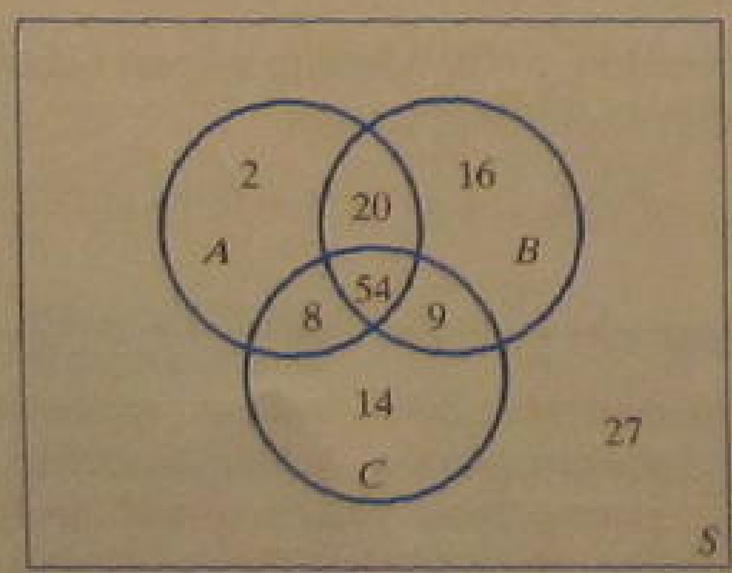


Figure 3.10 Diagram for Exercise 3.34

(f) $N(A \cap B \cap C)$;

(g) $N(A \cup B)$;

(h) $N(B \cup C)$;

(i) $N(\overline{A} \cup \overline{B} \cup C)$;

(j) $N[B \cap (A \cup C)]$.

3.35 An experiment has the four possible mutually exclusive outcomes A, B, C, and D. Check whether the following assignments of probability are permissible:

(a) $P(A) = 0.38$, $P(B) = 0.16$, $P(C) = 0.11$, $P(D) = 0.35$;

(b) $P(A) = 0.31$, $P(B) = 0.27$, $P(C) = 0.28$, $P(D) = 0.16$;

(c) $P(A) = 0.32$, $P(B) = 0.27$, $P(C) = -0.06$, $P(D) = 0.47$;

(d) $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{4}$, $P(C) = \frac{1}{8}$, $P(D) = \frac{1}{16}$;

(e) $P(A) = \frac{5}{18}$, $P(B) = \frac{1}{6}$, $P(C) = \frac{1}{3}$, $P(D) = \frac{2}{9}$.

3.36 With reference to Exercise 3.1, suppose that the points (0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2), (2, 0), (2, 1), (2, 2), (3, 0), (3, 1), and (3, 2) have the probabilities 0.060, 0.012, 0.006, 0.067, 0.014, 0.092, 0.260, 0.027, 0.080, 0.166, 0.110, and 0.106.

(a) Verify that this assignment of probabilities is permissible.

(b) Find the probabilities of events R, T, and U given in part (b) of that exercise.

(c) Calculate the probabilities that zero, one, or two streams are contaminated.

3.37 With reference to Exercise 3.7, suppose that each point (i, j) of the sample space is assigned the probability $\frac{15/28}{i+j}$.

(a) Verify that this assignment of probabilities is permissible.

(b) Find the probabilities of events B, C, and D described in part (b) of that exercise.

(c) Find the probabilities that one, two, or three of the graduate students will be supervising the physics lab.

3.38 Explain why there must be a mistake in each of the following statements:

(a) The probability that a mineral sample will contain silver is 0.38 and the probability that it will not contain silver is 0.52.

(b) The probability that a drilling operation will be a success is 0.34 and the probability that it will not be a success is −0.66.

(c) An air-conditioning repair person claims that the probability is 0.82 that the compressor is all right, 0.64 that the fan motor is all right, and 0.41 that they are both all right.

3.39 Refer to parts (c) and (d) of Exercise 3.13 to show that

(a) $P(A \cap B) \leq P(A)$;

(b) $P(A \cup B) \geq P(A)$.

3.40 Explain why there must be a mistake in each of the following statements:

(a) The probability that a student will get an A in a geology course is 0.3, and the probability that he or she will get either an A or a B is 0.27.

(b) A company is working on the construction of two shopping centers; the probability that the larger one will be completed on time is 0.35 and the probability that both will be completed on time is 0.42.

3.41 If A and B are mutually exclusive events, $P(A) = 0.26$, and $P(B) = 0.45$, find

(a) $P(\overline{A})$;

(b) $P(A \cup B)$;

(c) $P(A \cap \overline{B})$;

(d) $P(\overline{A} \cap \overline{B})$.

3.42 With reference to Exercise 3.34, suppose that the questionnaire filled in by one of the 150 persons is to be double-checked. If it is chosen in such a way that each questionnaire has a probability of $\frac{1}{150}$ of being selected, find the probabilities that the person

(a) lives more than 3 miles from the center of the city;

(b) regularly drives his or her car to work;

(c) does not live more than 3 miles from the center of the city and would not want to switch to public mass transportation if it were available;

(d) regularly drives his or her car to work but would gladly switch to public mass transportation if it were available.

3.43 A campus police department needs new bicycles for its patrol persons, and the probabilities are 0.17, 0.22, 0.03, 0.29, 0.21, and 0.08 that it will buy Bianshe, Cannonhill, Fishim, Giante, Trec, or HT. Find the probabilities that it will buy

(a) Cannonhill or Trec;

(b) Bianshe, Giante, or Trec;

(c) Fishim or Trec;

(d) Cannonhill, Giante, or HT.

3.44 The probabilities that a TV station will receive $0, 1, 2, 3, \ldots, 8$ or at least 9 complaints after showing a controversial program are, respectively, 0.01, 0.03, 0.07, 0.15, 0.19, 0.18, 0.14, 0.12, 0.09, and 0.02. What are the probabilities that after showing such a program the station will receive

(a) at most 4 complaints;

(b) at least 6 complaints;

(c) from 5 to 8 complaints?

3.45 If each point of the sample space of Figure 3.11 represents an outcome having the probability $\frac{1}{32}$, find

(a) $P(A)$;

(b) $P(B)$;

(c) $P(A \cap B)$;

(d) $P(A \cup B)$;

(e) $P(\overline{A} \cap B)$;

(f) $P(\overline{A} \cap \overline{B})$.

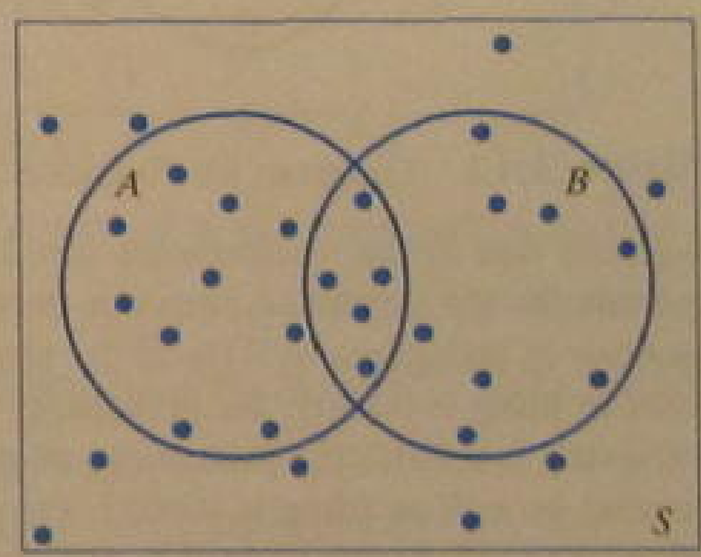


Figure 3.11 Diagram for Exercise 3.45

3.46 The probability that an integrated circuit chip will have defective etching is 0.06, the probability that it will have a crack defect is 0.03, and the probability that it has both defects is 0.02.

(a) What is the probability that a newly manufactured chip will have either an etching or a crack defect?

(b) What is the probability that a newly manufactured chip will have neither defect?

3.47 The probability that a new airport will get an award for its design is 0.16, the probability that it will get an award for the efficient use of materials is 0.24, and the probability that it will get both awards is 0.11.

(a) What is the probability that it will get at least one of the two awards?

(b) What is the probability that it will get only one of two awards?

3.48 Given $P(A) = 0.35$, $P(B) = 0.65$, and $P(A \cap B) = 0.12$, find

(a) $P(A \cup B)$;

(b) $P(\overline{A} \cap B)$;

(c) $P(A \cap \overline{B})$;

(d) $P(\overline{A} \cup \overline{B})$.

3.49 It can be shown that for any three events A, B, and C, the probability that at least one of them will occur is given by

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Verify that this formula holds for the probabilities of Figure 3.12.

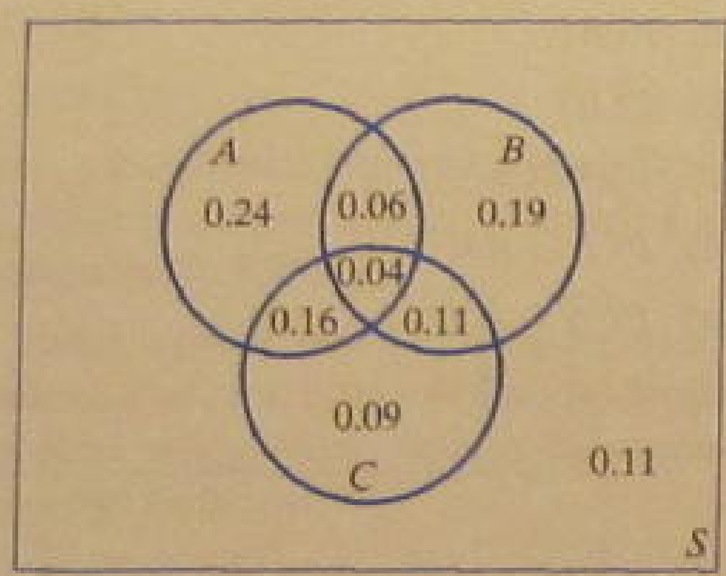


Figure 3.12 Diagram for Exercise 3.49

3.50 Suppose that in the maintenance of a large medical-records file for insurance purposes the probability of an error in processing is 0.0010, the probability of an error in filing is 0.0009, the probability of an error in retrieving is 0.0012, the probability of an error in processing as well as filing is 0.0002, the probability of an error in processing as well as retrieving is 0.0003, and the probability of an error in processing and filing as well as retrieving is 0.0001. What is the probability of making at least one of these errors?

3.51 If the probability of event A is p, then the **odds** that it will occur are given by the ratio of p to $1 - p$. Odds are usually given as a ratio of two positive integers having no common factor, and if an event is more likely not to occur than to occur, it is customary to give the odds that it will not occur rather than the odds that it will occur. What are the odds for or against the occurrence of an event if its probability is

(a) $\frac{4}{7}$; (b) 0.05; (c) 0.80?

3.52 Use the definition of Exercise 3.51 to show that if the odds for the occurrence of event A are a to b, where a and b are positive integers, then

$$p = \frac{a}{a+b}$$

3.53 The formula of Exercise 3.52 is often used to determine subjective probabilities. For instance, if an applicant for a job "feels" that the odds are 7 to 4 of getting the job, the subjective probability the applicant assigns to getting the job is

$$p = \frac{7}{7+4} = \frac{7}{11}$$

(a) If a businessperson feels that the odds are 3 to 2 that a new venture will succeed (say, by betting \$300 against \$200 that it will succeed), what subjective probability is he or she assigning to its success?

(b) If a student is willing to bet \$30 against \$10, but not \$40 against \$10, that he or she will get a passing grade in a certain course, what does this tell us about the subjective probability the student assigns to getting a passing grade in the course?

3.54 Subjective probabilities may or may not satisfy the third axiom of probability. When they do, we say that they are **consistent**; when they do not, they ought not to be taken too seriously.

(a) The supplier of delicate optical equipment feels that the odds are 7 to 5 against a shipment arriving late, and 11 to 1 against it not arriving at all. Furthermore, he feels that there is a 50/50 chance (the odds are 1 to 1) that such a shipment will either arrive late or not at all. Are the corresponding probabilities consistent?

(b) There are two Ferraris in a race, and an expert feels that the odds against their winning are, respectively, 2 to 1 and 3 to 1. Furthermore, she claims that there is a less-than-even chance that either of the two Ferraris will win. Discuss the consistency of these claims.

3.6 Conditional Probability

As we have defined probability, it is meaningful to ask for the probability of an event only if we refer to a given sample space S. To ask for the probability that an engineer earns at least \$90,000 a year is meaningless unless we specify whether we are referring to all engineers in the western hemisphere, all engineers in the United States, all those in a particular industry, all those affiliated with a university, and so forth. Thus, when we use the symbol $P(A)$ for the probability of A, we really mean the probability of A given some sample space S. Since the choice of S is by no means always evident, and since there are problems in which we are interested in the probabilities of A with respect to more sample spaces than one, the notation $P(A|S)$

is used to make it clear that we are referring to a particular sample space S. We read $P(A|S)$ as the conditional probability of A relative to S, and every probability is thus a conditional probability. Of course, we use the simplified notation $P(A)$ whenever the choice of S is clearly understood.

To illustrate some of the ideas connected with conditional probabilities, let us consider again the 500 machine parts of which some are improperly assembled and some contain one or more defective components as shown in Figure 3.7. Assuming equal probabilities in the selection of one of the machine parts for inspection, it can be seen that the probability of getting a part with one or more defective components is

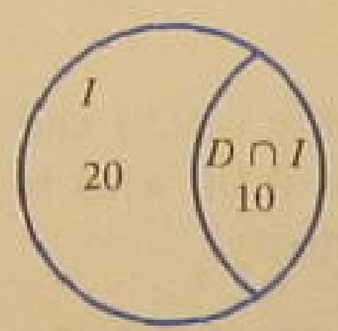


Figure 3.13
Reduced sample space

$$P(D) = \frac{10+5}{500} = \frac{3}{100}$$

To check whether the probability is the same if the choice is restricted to the machine parts that are improperly assembled, we have only to look at the reduced sample space of Figure 3.13 and assume that each of the 30 improperly assembled parts has the same chance of being selected. We thus get

$$P(D \mid I) = \frac{N(D \cap I)}{N(I)} = \frac{10}{30} = \frac{1}{3}$$

and it can be seen that the probability of getting a machine part with one or more defective components has increased from $\frac{3}{100}$ to $\frac{1}{3}$. Note that if we divide the numerator and denominator of the preceding formula for $P(D \mid I)$ by $N(S)$, we get

$$P(D \mid I) = \frac{\dfrac{N(D \cap I)}{N(S)}}{\dfrac{N(I)}{N(S)}} = \frac{P(D \cap I)}{P(I)}$$

where $P(D \mid I)$ is given by the ratio of $P(D \cap I)$ to $P(I)$.

Looking at this example in another way, note that with respect to the whole sample space S we have

$$P(D \cap I) = \frac{10}{500} = \frac{1}{50} \qquad \text{and} \qquad P(\overline{D} \cap I) = \frac{20}{500} = \frac{2}{50}$$

assuming, as before, that each of the 500 machine parts has the same chance of being selected. Thus, the probabilities that the machine part selected will or will not contain one or more defective components, given that it is improperly assembled, should be in the ratio 1 to 2. Since the probabilities of D and $\overline{D}$ in the reduced sample space must add up to 1, it follows that

$$P(D \mid I) = \frac{1}{3} \qquad \text{and} \qquad P(\overline{D} \mid I) = \frac{2}{3}$$

which agrees with the result obtained before. This explains why, in the last step, we had to divide by $P(I)$ to

$$P(D \mid I) = \frac{P(D \cap I)}{P(I)}$$

Division by $P(I)$, or multiplication by $1/P(I)$, takes care of the proportionality factor, which makes the sum of the probabilities over the reduced sample space equal to 1.

Following these observations, let us now make the following general definition:

Conditional probability

If A and B are any events in S and $P(B) \neq 0$, the conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

EXAMPLE **Calculating a conditional probability**

If the probability that a communication system will have high fidelity is 0.81 and the probability that it will have high fidelity and high selectivity is 0.18, what is probability that a system with high fidelity will also have high selectivity?

Solution If A is the event that a communication system has high selectivity and B is the event that it has high fidelity, we have $P(B) = 0.81$ and $P(A \cap B) = 0.18$, and substitution into the formula yields

$$P(A \mid B) = \frac{0.18}{0.81} = \frac{2}{9}$$

■

EXAMPLE **The conditional probability that a used car has low mileage given that it is expensive to operate**

Referring to the used car example, for which the probabilities of the individual outcomes are given in Figure 3.8, use the results on page 61 to find $P(M_1 \mid C_3)$.

Solution Since we had $P(M_1 \cap C_3) = 0.08$ and $P(C_1) = 0.40$, substitution into the formula for consitional probability yields

$$P(M_1 \mid C_3) = \frac{P(M_1 \cap C_3)}{P(C_3)} = \frac{0.08}{0.40} = 0.20$$

It is of interest to note that the value of the conditional probability obtained here, $P(M_1 \mid C_3) = 0.20$, equals the value for $P(M_1)$ obtained on page 61. This means that the probability a used car has low mileage is the same whether or not it is expensive to operate. We say that M_1 is **independent** of C_3. As the reader is asked to verify in Exercise 3.59, it also follows from the results on page 61 that M_1 is not independent of P_1, namely, that low mileage is related to the car's price. ■

In general, if A and B are any two events in a sample space S, we say that **A is independent of B if and only if $P(A \mid B) = P(A)$,** but as it can be shown that B is independent of A whenever A is independent of B, it is customary to say simply that **A and B are independent events**.

General multiplication rule of probability

Theorem 3.8 If A and B are any events in S, then

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B \mid A) \quad &\text{if} \quad P(A) \neq 0 \\ &= P(B) \cdot P(A \mid B) \quad &\text{if} \quad P(B) \neq 0 \end{aligned}$$

The second of these rules is obtained directly from the definition of conditional probability by multiplying both sides by $P(B)$; the first is obtained from the second by interchanging the letters A and B.

EXAMPLE **Using the general multiplication rule of probability**

The supervisor of a group of 20 construction workers wants to get the opinion of 2 of them (to be selected at random) about certain new safety regulations. If 12 workers favor the new regulations and the other 8 are against them, what is the probability that both of the workers chosen by the supervisor will be against the new safety regulations?

Solution Assuming equal probabilities for each selection (which is what we mean by the selections being random), the probability that the first worker selected will be against the new safety regulations is $\frac{8}{20}$, and the probability that the second worker selected will be against the new safety regulations given that the first one is against them is $\frac{7}{19}$. Thus, the desired probability is

$$\frac{8}{20} \cdot \frac{7}{19} = \frac{14}{95}$$

■

In the special case where A and B are independent so $P(A \mid B) = P(A)$, Theorem 3.8 leads to the following result:

Special product rule of probability

Theorem 3.9 Two events A and B are independent events if and only if

$$P(A \cap B) = P(A) \cdot P(B)$$

Thus, the probability that two independent events will both occur is simply the product of their probabilities. This rule is sometimes used as the definition of independence. It applies even when $P(A)$ or $P(B)$ or both equal 0. In any case, it may be used to determine whether two given events are independent.

EXAMPLE **The outcomes to unrelated parts of an experiment can be treated as independent**

What is the probability of getting two heads in two flips of a balanced coin?

Solution Since the probability of heads is $\frac{1}{2}$ for each flip and the two flips are independent, the probability is

$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

■

EXAMPLE **Independence and selection with and without replacement**

Two cards are drawn at random from an ordinary deck of 52 playing cards. What is the probability of getting two aces if

(a) the first card is replaced before the second card is drawn;

(b) the first card is not replaced before the second card is drawn?

Solution **(a)** Since there are four aces among the 52 cards, we get

$$\frac{4}{52} \cdot \frac{4}{52} = \frac{1}{169}$$

(b) Since there are only three aces among the 51 cards that remain after one ace has been removed from the deck, we get

$$\frac{4}{52} \cdot \frac{3}{51} = \frac{1}{221}$$

Note that

$$\frac{1}{221} \neq \frac{4}{52} \cdot \frac{4}{52}$$

so independence is violated when the sampling is without replacement. ■

EXAMPLE **Checking if two events are independent under an assigned probability**

If $P(C) = 0.65$, $P(D) = 0.40$, and $P(C \cap D) = 0.24$, are the events C and D independent?

Solution Since $P(C) \cdot P(D) = (0.65)(0.40) = 0.26$ and not 0.24, the two events are not independent. ■

In the preceding examples we have used the assigned probabilities to check if two events are independent. The concept of independence can be—and frequently is—employed when probabilities are assigned to events that concern unrelated parts of an experiment.

EXAMPLE **Assigning probability by the special product rule**

Let A be the event that raw material is available when needed and B be the event that the machining time is less than 1 hour. If $P(A) = 0.8$ and $P(B) = 0.7$, assign probability to the event $A \cap B$.

Solution Since the events A and B concern unrelated steps in the manufacturing process, we invoke independence and make the assignment

$$P(A \cap B) = P(A)P(B) = 0.8 \times 0.7 = 0.56$$ ■

The special product rule can easily be extended so that it applies to more than two independent events—again, we multiply together all the individual probabilities.

EXAMPLE **The extended special product rule of probability**

What is the probability of not rolling any 6's in four rolls of a balanced die?

Solution The probability is $\frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} = \frac{625}{1,296}$ ■

For three or more dependent events the multiplication rule becomes more complicated, as is illustrated in Exercise 3.70.

EXAMPLE **The probability of falsely signaling a pollution problem**

Many companies must monitor the effluent that is discharged from their plants into rivers and waterways. In some states, it is the law that some substances have water-quality limits that are below the limit of detection, L, for the current method of measurement. The effluent is judged to satisfy the quality limit if every test specimen is below the limit of detection L. Otherwise it will be declared to fail compliance with the quality limit. Suppose the water does not contain the contaminant of interest but

that the variability in the chemical analysis still gives a 1% chance that a measurement on a test specimen will exceed L.

(a) Find the probability that neither of two test specimens, both free of the contaminant, will fail to be in compliance.

(b) If one test specimen is taken each week for two years, and they are all free of the contaminant, find the probability that none of the test specimens will fail to be in compliance.

(c) Comment on the incorrect reasoning of having a fixed limit of detection no matter how many tests are conducted.

Solution **(a)** If the two samples are not taken too closely in time or space, we treat them as independent. We use the special product rule to obtain the probability that both are in compliance:

$$0.99 \times 0.99 = 0.9801$$

(b) Treating the results for different weeks as independent,

$$(0.99)^{104} = 0.35$$

so, even with excellent water quality, there is almost a two-thirds chance that at least once the water quality will be declared to fail to be in compliance with the law.

(c) With this type of law, no company would want to collect test specimens more than maybe once a year. This is in direct opposition to the scientific idea that more information is better than less information on water quality. Some effort should be made to allow for higher limits when the testing is more frequent. ■

EXAMPLE **Using probability to compare the accuracy of two schemes for sending messages**

Electrical engineers are considering two alternative systems for sending messages. A message consists of a word that is either a 0 or a 1. However, because of random noise in the channel, a 1 that is transmitted could be received as a 0 and vice versa. That is, there is a small probability, p, that

$$P[\text{A transmitted 1 is received as 0}] = p$$
$$P[\text{A transmitted 0 is received as 1}] = p$$

One scheme is to send a single digit. The message is short but may be unreliable. A second scheme is to repeat the selected digit three times in succession. At the receiving end, the majority rule will be used to decode. That is, when any of 101, 110, 011, or 111 are received, it is interpreted to mean a 1 was sent.

(a) Evaluate the probability that a transmitted 1 will be received as a 1 under the three-digit scheme when $p = 0.01, 0.02$, or 0.05. Compare this with the scheme where a single digit is transmitted as a word. Treat the results for different digits as independent.

(b) Suppose a message, consisting of the two words, a 1 followed by 0, is to be transmitted using the three-digit scheme. What is the probability that the total message will be correctly decoded under the majority rule with $p = 0.05$? Compare with the scheme where a single digit is transmitted as a word.

Solution **(a)** The three digits 111 are transmitted. By independence, the sequence 111 has probability $(1-p)(1-p)(1-p)$ of being received as 111. Also the

probability of receiving 011 is $p(1-p)(1-p)$, so the probability of exactly one 0 among the three received is $3p(1-p)^2$. Using the majority rule,

$$P[\text{Correct}] = P[\text{transmitted 1 received as 1}] = (1-p)^3 + 3p(1-p)^2$$

p	0.01	0.02	0.05
$P[\text{Correct}]$	0.9997	0.9988	0.9928

All three probabilities are considerably above the corresponding single-digit scheme probabilities 0.99, 0.98, and 0.95, respectively.

(b) Both of the words 1 and 0 must be received correctly. As in part (a), the probability that a 0 is received correctly is also 0.9928. Consequently, using independence, the probability that the total message is correctly received is $(0.9928)^2 = 0.986$. This improves over the scheme where single digits are sent for each word since that scheme has only the probability $(0.95)^2 = 0.903$ of correctly receiving the total message.

Redundancy helps improve accuracy, but more digits need to be transmitted, which results in a significantly lower throughput. ■

3.7 Bayes' Theorem

The general multiplication rules are useful in solving many problems in which the ultimate outcome of an experiment depends on the outcomes of various intermediate stages. Suppose, for instance, that an assembly plant receives its voltage regulators from three different suppliers, 60% from supplier B_1, 30% from supplier B_2, and 10% from supplier B_3. In other words, the probabilities that any one voltage regulator received by the plant comes from these three suppliers are 0.60, 0.30, and 0.10. Also suppose that 95% of the voltage regulators from B_1, 80% of those from B_2, and 65% of those from B_3 perform according to specifications. We would like to know the probability that any one voltage regulator received by the plant will perform according to specifications.

If A denotes the event that a voltage regulator received by the plant performs according to specifications, and B_1, B_2, and B_3 are the events that it comes from the respective suppliers, we can write

$$\begin{aligned} A &= A \cap [B_1 \cup B_2 \cup B_3] \\ &= (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \end{aligned}$$

where B_1, B_2, and B_3 are mutually exclusive events of which one must occur. It follows that $A \cap B_1$, $A \cap B_2$, and $A \cap B_3$ are also mutually exclusive. By the generalization of the third axiom of probability on page 60, we get

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

Then, if we apply the second of the general multiplication rules to $P(A \cap B_1)$, $P(A \cap B_2)$, and $P(A \cap B_3)$, we get

$$P(A) = P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2) + P(B_3) \cdot P(A \mid B_3)$$

and substitution of the given numerical values yields

$$\begin{aligned} P(A) &= (0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65) \\ &= 0.875 \end{aligned}$$

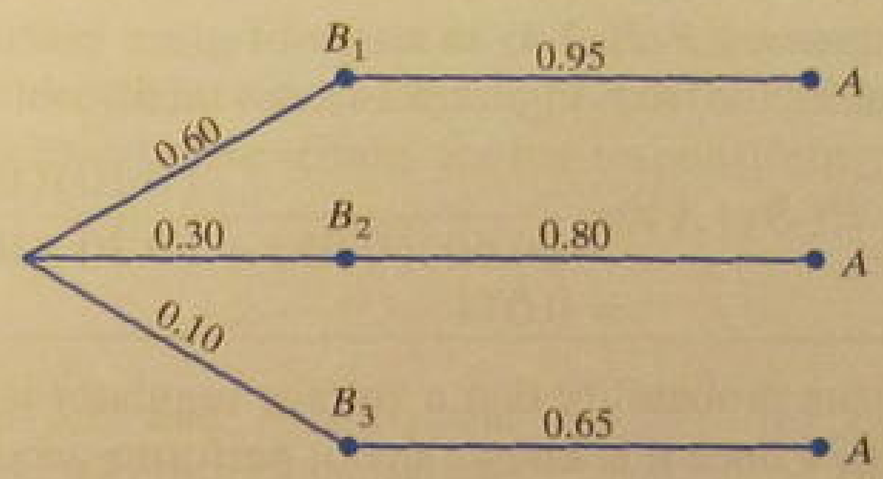


Figure 3.14
Tree diagram for example dealing with three suppliers of voltage regulators

for the probability that any one voltage regulator received by the given plant wil perform according to specifications.

To visualize this result, we have only to construct a tree diagram like that o Figure 3.14, where the probability of the final outcome is given by the sum of th products of the probabilities corresponding to each branch of the tree.

In the preceding example there were only 3 possibilities at the intermediate stag but if there are n mutually exclusive possibilities $B_1, B_2, \ldots, B_n$ at the intermediat stage, a similar argument will lead to the following result, sometimes called the **rul of elimination** or the **rule of total probability**:

Rule of total probability

Theorem 3.10 If $B_1, B_2, \ldots, B_n$ are mutually exclusive events of which one must occur, then

$$P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)$$

The tree diagram like that of Figure 3.15, where the probability of the fin outcome is again given by the sum of the products of the probabilities correspondir to each branch of the tree, graphically explains the calculation.

To consider a problem that is closely related to the one we have just discusse suppose we want to know the probability that a particular voltage regulator, which known to perform according to specifications, came from supplier B_3. Symbolicall we want to know the value of $P(B_3 \mid A)$, and to find a formula for this probabili we first write

$$P(B_3 \mid A) = \frac{P(A \cap B_3)}{P(A)}$$

Then, substituting $P(B_3) \cdot P(A \mid B_3)$ for $P(A \cap B_3)$ and $\sum_{i=1}^{3} P(B_i) \cdot P(A \mid B_i)$ f $P(A)$ in accordance with Theorems 3.8 and 3.10, we get

$$P(B_3 \mid A) = \frac{P(B_3) \cdot P(A \mid B_3)}{\sum_{i=1}^{3} P(B_i) \cdot P(A \mid B_i)}$$

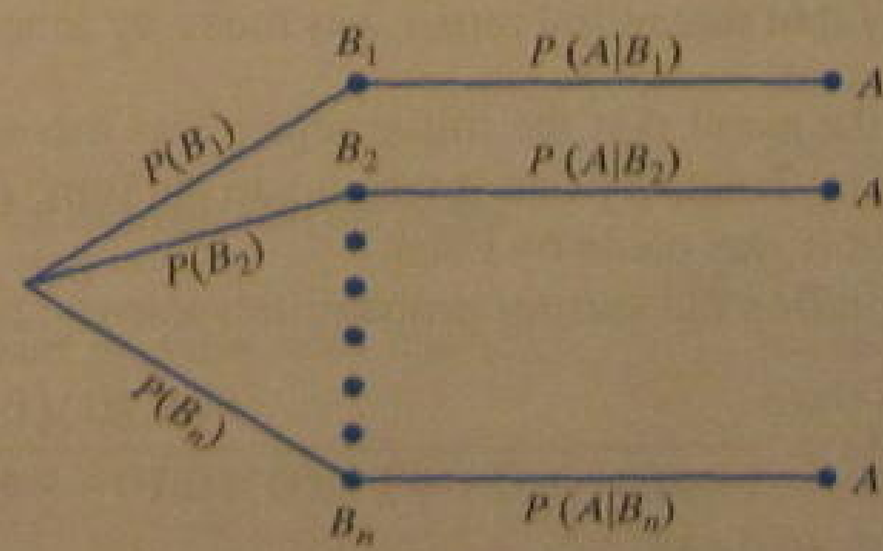


Figure 3.15
Tree diagram for rule of elimination

which expresses $P(B_3 \mid A)$ in terms of given probabilities. Substituting the values from page 72 (or from Figure 3.14), we finally obtain

$$P(B_3 \mid A) = \frac{(0.10)(0.65)}{(0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65)}$$
$$= 0.074$$

Note that the probability that a voltage regulator is supplied by B_3 decreases from 0.10 to 0.074 once it is known that it performs according to specifications.

The method used to solve the preceding example can easily be generalized to yield the following formula, called **Bayes' theorem**:

Bayes' theorem

Theorem 3.11 If $B_1, B_2, \ldots, B_n$ are mutually exclusive events of which one must occur, then

$$P(B_r \mid A) = \frac{P(B_r) \cdot P(A \mid B_r)}{\sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)}$$

for $r = 1, 2, \ldots, n$.

Note that the expression in the numerator is the probability of reaching A via the rth branch of the tree and that the expression in the denominator is the sum of the probabilities of reaching A via the n branches of the tree.

Bayes' theorem provides a formula for finding the probability that the "effect" A was "caused" by the event B_r. For instance, in our example we found the probability that an acceptable voltage regulator was made by supplier B_3. The probabilities $P(B_i)$ are called the *prior*, or *a priori*, probabilities of the "causes" B_i, and in practice it is often difficult to assign them numerical values. For many years Bayes' theorem was looked upon with suspicion because it was used with the often erroneous assumption that the prior probabilities are all equal. A good deal of the controversy once surrounding Bayes' theorem has been cleared up with the realization that the probabilities $P(B_i)$ must be determined separately in each case from the nature of the problem, preferably on the basis of specific knowledge or past experience.

EXAMPLE **Using Bayes' theorem**

Four technicians regularly make repairs when breakdowns occur on an automated production line. Janet, who services 20% of the breakdowns, makes an incomplete repair 1 time in 20; Tom, who services 60% of the breakdowns, makes an incomplete repair 1 time in 10; Georgia, who services 15% of the breakdowns, makes an incomplete repair 1 time in 10; and Peter, who services 5% of the breakdowns, makes an incomplete repair 1 time in 20. For the next problem with the production line diagnosed as being due to an initial repair that was incomplete, what is the probability that this initial repair was made by Janet?

Solution Let A be the event that the initial repair was incomplete, B_1 that the initial repair was made by Janet, B_2 that it was made by Tom, B_3 that it was made by Georgia, and B_4 that it was made by Peter.

Substituting the various probabilities into the formula of Theorem 3.11, we get

$$P(B_1 \mid A) = \frac{(0.20)(0.05)}{(0.20)(0.05) + (0.60)(0.10) + (0.15)(0.10) + (0.05)(0.05)}$$
$$= 0.114$$

and it is of interest to note that although Janet makes an incomplete repair only 1 out of 20 times, namely, 5% of the breakdowns she services, more than 11% of the incomplete repairs are her responsibility. ■

Exercises

3.55 With reference to Figure 3.7, find $P(I \mid D)$ and $P(I \mid \overline{D})$, assuming that originally each of the 500 machine parts has the same chance of being chosen for inspection.

3.56 (a) Would you expect the probability that a randomly selected car will need major repairs in the next year to be smaller, remain the same, or increase if you are told it already has high mileage? Explain.

(b) Would you expect the probability that a randomly selected senior would know the second law of thermodynamics, to be smaller, remain the same, or increase if the person selected is a mechanical engineering major? Explain.

(c) In Part (a), identify the two events with symbols A and B and the conditional probability of interest.

3.57 With reference to Exercise 3.34 and Figure 3.10, assume that each of 150 persons has the same chance of being selected, and find the probabilities that he or she

(a) lives more than 3 miles from the center of the city given that he or she would gladly switch to public mass transportation;

(b) regularly drives his or her car to work given that he or she lives more than 3 miles from the center of the city;

(c) would not want to switch to public mass transportation given that he or she does not regularly drive his or her car to work.

3.58 With reference to Figure 3.12, find

(a) $P(A \mid B)$;

(b) $P(B \mid \overline{C})$;

(c) $P(A \cap B \mid C)$;

(d) $P(B \cup C \mid \overline{A})$;

(e) $P(A \mid B \cup C)$;

(f) $P(A \mid B \cap C)$;

(g) $P(A \cap B \cap C \mid B \cap C)$;

(h) $P(A \cap B \cap C \mid B \cup C)$.

3.59 With reference to the used car example and the probabilities given in Figure 3.8, find

(a) $P(M_1 \mid P_1)$ and compare its value with that of $P(M_1)$;

(b) $P(C_3 \mid P_2)$ and compare its value with that of $P(C_3)$;

(c) $P(M_1 \mid P_1 \cap C_3)$ and compare its value with that of $P(M_1)$.

3.60 With reference to Exercise 3.47, find the probabilities that the airport will get the design award given that

(a) it got the award for the efficient use of materials;

(b) it did not get the award for the efficient use of materials.

3.61 Prove that $P(A \mid B) = P(A)$ implies that $P(B \mid A) = P(B)$ provided that $P(A) \neq 0$ and $P(B) \neq 0$.

3.62 If the probabilities are 0.98, 0.96, and 0.95 that a newly manufactured bottle of soap will be free of major defects, free of minor blemishes, and free of major defects and minor blemishes, find the probability that a bottle

(a) that is free of major defects will also be free of minor blemishes

(b) that is free of minor blemishes will also be free of major defects

3.63 Given that $P(A) = 0.60$, $P(B) = 0.40$, and $P(A \cap B) = 0.24$, verify that

(a) $P(A \mid B) = P(A)$;

(b) $P(A \mid \overline{B}) = P(A)$;

(c) $P(B \mid A) = P(B)$;

(d) $P(B \mid \overline{A}) = P(B)$.

3.64 Among the 24 invoices prepared by a billing department, 4 contain errors while the others do not. If we randomly check 2 of these invoices, what are the probabilities that

(a) both will contain errors;

(b) neither will contain an error?

3.65 Among 60 automobile repair parts loaded on a truck in San Francisco, 45 are destined for Seattle and 15 for Vancouver. If two of the parts are unloaded in Portland by mistake and the "selection" is random, what are the probabilities that

(a) both parts should have gone to Seattle;

(b) both parts should have gone to Vancouver;

(c) one should have gone to Seattle and one to Vancouver?

3.66 Two random digits will be selected using a random number generator (all 100 pairs of digits are equally likely). Find the probabilities of getting

(a) two 5's;

(b) first a 5 and then a digit less than 5.

3.67 If $P(A) = 0.60$, $P(B) = 0.45$, and $P(A \cap B) = 0.27$, are A and B independent?

3.68 If the odds are 5 to 3 that an event M will not occur, 2 to 1 that event N will occur, and 4 to 1 that they will not both occur, are the two events M and N independent?

3.69 Find the probabilities of getting

(a) eight heads in a row with a balanced coin;

(b) three 3's and then a 4 or a 5 in four rolls of a balanced die;

(c) five multiple-choice questions answered correctly, if for each question the probability of answering it correctly is $\frac{1}{3}$.

3.70 For three or more events which are not independent, the probability that they will all occur is obtained by multiplying the probability that one of the events will occur, times the probability that a second of the events will occur given that the first event has occurred, times the probability that a third of the events will occur given that the first two events have occurred, and so on. For instance, for three events we can write

$$P(A \cap B \cap C) = P(A) \cdot P(B \mid A) \cdot P(C \mid A \cap B)$$

and we find that the probability of drawing without replacement three aces in a row from an ordinary deck of 52 playing cards is

$$\frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} = \frac{1}{5{,}525}$$

(a) If six bullets, of which three are blanks, are randomly inserted into a gun, what is the probability that the first three bullets fired will all be blanks?

(b) In a certain city during the month of May, the probability that a rainy day will be followed by another rainy day is 0.80 and the probability that a sunny day will be followed by a rainy day is 0.60. Assuming that each day is classified as being either rainy or sunny and that the weather on any given day depends only on the weather the day before, find the probability that in the given city a rainy day in May is followed by two more rainy days, then a sunny day, and finally another rainy day.

(c) A department store which bills its charge-account customers once a month has found that if a customer pays promptly one month, the probability is 0.90 that he will also pay promptly the next month; however, if a customer does not pay promptly one month, the probability that he will pay promptly the next month is only 0.50. What is the probability that a customer who has paid promptly one month will not pay promptly the next three months?

(d) If 5 of a company's 12 delivery trucks do not meet emission standards and 4 of the 12 trucks are randomly picked for inspection, what is the probability that none of them meets emission standards?

3.71 At an electronics plant, it is known from past experience that the probability is 0.83 that a new worker who has attended the company's training program will meet the production quota and that the corresponding probability is 0.35 for a new worker who has not attended the company's training program. If 80% of all new workers attend the training program, what is the probability that a new worker will meet the production quota?

3.72 With reference to the preceding exercise, find the probability that a new worker who meets the production quota will have attended the company's training program.

3.73 A consulting firm rents cars from three agencies, 20% from agency D, 20% from agency E, and 60% from agency F. If 10% of the cars from D, 12% of the cars from E, and 4% of the cars from F have bad tires, what is the probability that the firm will get a car with bad tires?

3.74 With reference to Exercise 3.73, what is the probability that a car with bad tires rented by the firm came from agency F?

3.75 Use the information on the tree diagram of Figure 3.16 to determine the value of

(a) $P(A)$;

(b) $P(B \mid A)$;

(c) $P(B \mid \overline{A})$.

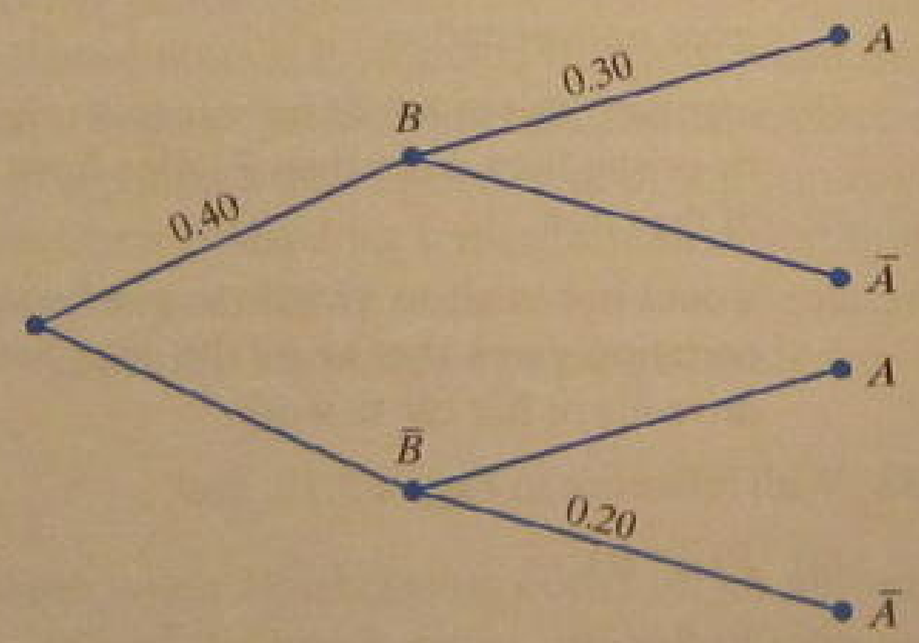


Figure 3.16 Diagram for Exercise 3.75

3.76 With reference to the example on page 72, find the probabilities that a voltage regulator that performs according to specifications came

(a) from supplier B_1

(b) from supplier B_2

3.77 With reference to the example on page 74, for a problem diagnosed as being due to an incomplete initial repair, find the probability that the initial repair was made by

(a) Tom;

(b) Georgia;

(c) Peter.

3.78 Two firms V and W consider bidding on a road-building job, which may or may not be awarded depending on the amounts of the bids. Firm V submits a bid and the probability is $\frac{3}{4}$ that it will get the job provided firm W does not bid. The probability is $\frac{3}{4}$ that W will bid, and if it does, the probability that V will get the job is only $\frac{1}{3}$.

(a) What is the probability that V will get the job?

(b) If V gets the job, what is the probability that W did not bid?

3.79 Engineers in charge of maintaining our nuclear fleet must continually check for corrosion inside the pipes that are part of the cooling systems. The inside condition of the pipes cannot be observed directly but a nondestructive test can give an indication of possible corrosion. This test is not infallible. The test has probability 0.7 of detecting corrosion when it is present but it also has probability 0.2 of falsely indicating internal corrosion. Suppose the probability that any section of pipe has internal corrosion is 0.1.

(a) Determine the probability that a section of pipe has internal corrosion, given that the test indicates its presence.

(b) Determine the probability that a section of pipe has internal corrosion, given that the test is negative.

3.80 An East Coast manufacturer of printed circuit boards exposes all finished boards to an online automated verification test. During one period, 900 boards were completed and 890 passed the test. The test is not infallible. Of 30 boards intentionally made to have noticeable defects, 25 were detected by the test. Use the relative frequencies to approximate the conditional probabilities needed below.

(a) Give an approximate value for $P[\text{Pass test} \mid \text{board has defects}]$.

(b) Explain why your answer in part a may be too small.

(c) Give an approximate value for the probability that a manufactured board will have defects. In order to answer the question, you need information about the conditional probability that a good board will fail the test. This is important to know but was not available at the time an answer was required. To proceed, you can assume that this probability is zero.

(d) Approximate the probability that a board has defects given that it passed the automated test.

Do's and Don'ts

Do's

1. Begin by creating a sample space S which specifies all possible outcomes.
2. Always assign probabilities to events that satisfy the axioms of probability. In the discrete case, the possible outcomes can be arranged in a sequence. The axioms are then automatically satisfied when probability p_i is assigned to the ith outcome, where

$$0 \le p_i \quad \text{and} \quad \sum_{\text{all outcomes in } S} p_i = 1$$

and the probability of any event A is defined as

$$P(A) = \sum_{\text{all outcomes in } A} p_i$$

3. Combine the probabilities of events according to rules of probability.

General Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Rule of the Complement: $P(\overline{A}) = 1 - P(A)$

General Multiplication Rule: $P(A \cap B) = P(A)P(B \mid A)$ if $P(A) \neq 0$
$= P(B)P(A \mid B)$ if $P(B) \neq 0$

Conditional Probability: $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ if $P(B) \neq 0$

Don'ts

1. Don't confuse independent events with mutually exclusive events. When A and B are mutually exclusive, only one of them can occur. Their intersection is empty and so has probability 0.
2. Don't assign probability to $A \cap B$ according to special product rule

$$P(A \cap B) = P(A)P(B)$$

unless the conditions for independence hold. Independence may be plausible when the events A and B pertain to physically unrelated parts of a large system and there are no common causes that jointly affect the occurrence of both events.

Review Exercises

3.81 (a) In a recent random check of 300 books housed in the Engineering Library, 27 were checked out. For a randomly selected book today, what is the probability that it is currently checked out?

(b) Last year, 380 students applied for 28 internships administered by the university. Next year you will apply for one of the 28 internships and would like to estimate the probability of receiving one. Give your estimate and comment on one factor that might influence your estimate.

3.82 A salesperson of industrial chemicals has four customers in Sacramento, whom he may or may not be able to visit on a 2-day trip to this city. He will not visit any of these customers more than once.

(a) Using two coordinates so that (2,1), for example, represents the event that he will visit two of his customers on the first day and one on the second day, draw a diagram similar to that of Figure 3.1 showing the points of the corresponding sample space.

(b) List the points of the sample space that constitute the events A, B, and C that he will visit all four of his customers, that he will visit more of this customers on the first day than on the second day, and that he will visit at least three of his customers on the second day.

(c) Which of the three pairs of events, A and B, B and C, or A and C, are mutually exclusive?

3.83 With reference to the preceding exercise, express each of the following events symbolically by listing its elements, and also express it in words:

(a) $\overline{A}$;

(b) $A \cup B$;

(c) $A \cap C$;

(d) $\overline{A} \cap B$.

3.84 Use Venn diagrams to verify that

(c) $\overline{\overline{A} \cup \overline{B}} = A \cap B$;

(b) $\overline{A \cap \overline{B}} = \overline{A} \cup B$.

3.85 The quality of surround sound from four digital movie systems is to be rated superior, average, or inferior, and we are interested only in how many of the systems get each of these ratings. Draw a tree diagram which shows the 12 different possibilities.

3.86 Alarm units are to be connected at four fixed positions along a pipeline. In how many ways can the four available alarm units be connected to the four positions along the line?

3.87 In how many ways can two out of seven chemical engineers be assigned to a new project?

3.88 Refer to the example of motors for miniaturized capsules on page 56, but instead suppose that 20 motors are available and that 4 will not operate satisfactorily, when placed in a capsule. If the scientist wishes to fabricate two capsules, with two motors each, find the probability that among the four randomly selected motors

(a) all four operate satisfactorily;

(b) three operate satisfactorily and one does not.

3.89 Given $P(A) = 0.35$, $P(B) = 0.40$, and $P(A \cap B) = 0.20$, find

(a) $P(A \cup B)$;

(b) $P(\overline{A} \cap B)$;

(c) $P(A \cap \overline{B})$;

(d) $P(\overline{A} \cup \overline{B})$.

(e) Are A and B independent?

3.90 In a sample of 446 cars stopped at a roadblock, only 67 of the drivers had their seatbelts fastened. Estimate the probability that a driver stopped on that road will have his or her seatbelt fastened.

3.91 The marketing manager reported to the head engineer regarding a survey concerning the company's portable cleaning tool. He claims that, among the 200 customers surveyed, 165 said the product is reliable, 117 said it is easy to use, 88 said it is both reliable and easy to use, and 33 said it is neither reliable nor easy to use. Explain why the head engineer should question this claim.

3.92 If the probabilities that a satellite launching rocket will explode during lift-off or have its guidance system fail in flight are 0.0002 and 0.0005, find the probabilities that such a rocket will

(a) not explode during lift-off;

(b) explode during lift-off or have its guidance system fail in flight;

(c) neither explode during lift-off nor have its guidance system fail in flight.

3.93 Given $P(A) = 0.20$, $P(B) = 0.45$, and $P(A \cap B) = 0.09$, verify that

(a) $P(A \mid B) = P(A)$;

(b) $P(A \mid \overline{B}) = P(A)$;

(c) $P(B \mid A) = P(B)$;

(d) $P(B \mid \overline{A}) = P(B)$.

3.94 If events A and B are independent and $P(A) = 0.45$ and $P(B) = 0.30$, find

(a) $P(A \cap B)$;

(b) $P(A \mid B)$;

(c) $P(A \cup B)$;

(d) $P(\overline{A} \cap \overline{B})$.

3.95 The following frequency table shows the classification of 58 landfills in a state according to their concentration of the three hazardous chemicals arsenic, barium, and mercury.

		Barium			
		High		Low	
		Mercury		Mercury	
		High	Low	High	Low
Arsenic	High	1	3	5	9
	Low	4	8	10	18

If a landfill is selected at random, find the probability that it has

(a) a high concentration of mercury;

(b) a high concentration of barium and low concentrations of arsenic and mercury;

(c) high concentrations of any two of the chemicals and low concentration of the third;

(d) a high concentration of any one of the chemicals and low concentrations of the other two.

3.96 Refer to Exercise 3.95. Given that a landfill, selected at random, is found to have a high concentration of barium, what is the probability that its concentration is

(a) high in mercury?

(b) low in both arsenic and mercury?

(c) high in either arsenic or mercury?

3.97 An explosion in an LNG storage tank in the process of being repaired could have occurred as the result of static electricity, malfunctioning electrical equipment, an open flame in contact with the liner, or purposeful action (industrial sabotage). Interviews with engineers who were analyzing the risks involved led to estimates that such a explosion would occur with probability 0.25 as a result of static electricity, 0.20 as a result of malfunctioning electric equipment, 0.40 as a result of an open flame, and 0.75 as a result of purposeful action. These interviews also yielded subjective estimates of the prior probabilities of these four causes of 0.30, 0.40, 0.15, and 0.15, respectively. What was the most likely cause of the explosion?

3.98 An engineer wants to move across the country from City A to City B to work. On different days, she mails a cover letter and resume to three different companies in City B. Suppose that a letter sent from City A to City B has a probability of 0.7 of reaching City B within 3 days.

(a) What is the probability that exactly 2 of the 3 letters will reach City B within 3 days?

(b) If the three letters are mailed together at the same time and location, how does your conclusion in part (a) change? Explain.

3.99 Amy commutes to work by two different routes A and B. If she comes home by route A, then she will be home no later than 6 P.M. with probability 0.8, but if she comes home by route B, then she will be home no later than 6 P.M. with probability 0.7. In the past, the proportion of times that Amy chose route A is 0.4.

(a) What proportion of times is Amy home no later than 6 P.M.?

(b) If Amy is home after 6 P.M. today, what is the probability that she took route B?

Key Terms

Additive set function 57
Axioms of probability 58
Bayes' theorem 74
Classical probability concept 55
Combination 52
Complement 47
Conditional probability 68
Consistency of probabilities 66
Continuous sample space 46
Discrete sample space 46
Empty set 46
Event 46
Experiment 45
Factorial notation 51
Finite sample space 46
Frequency interpretation 57
General addition rule 62
General multiplication rule 68
Independent events 68
Intersection 47
Multiplication of choices 50
Mutually exclusive events 47
Odds 66
Outcome 45
Permutation 51
Rule of complement 63
Rule of total probability 73
Sample space 45
Set function 57
Special addition rule 62
Special product rule 69
Subjective probability 57
Tree diagram 49
Union 47
Venn diagram 47

CHAPTER

PROBABILITY DISTRIBUTIONS

4

In most statistical problems we are concerned with one number or a few numbers that are associated with the outcomes of experiments. In the inspection of a manufactured product we may be interested only in the number of defectives; in the analysis of a road test we may be interested only in the average speed and the average fuel consumption; and in the study of the performance of a miniature rechargeable battery we may be interested only in its power and lifelength. All these numbers are associated with situations involving an element of chance—in other words, they are values of random variables.

In the study of random variables we are usually interested in their probability distributions, namely, in the probabilities with which they take on the various values in their range. The introduction to random variables and probability distributions in Section 4.1 is followed by a discussion of various special probability distributions in Sections 4.2, 4.3, 4.6, 4.7, 4.8, and 4.9, and descriptions of the most important features of probability distributions in Sections 4.4 and 4.5.

CHAPTER OUTLINE

4.1 Random Variables 81
4.2 The Binomial Distribution 85
4.3 The Hypergeometric Distribution 89
4.4 The Mean and the Variance of a Probability Distribution 93
4.5 Chebyshev's Theorem 100
4.6 The Poisson Approximation to the Binomial Distribution 104
4.7 Poisson Processes 107
4.8 The Geometric and Negative Binomial Distribution 109
4.9 The Multinomial Distribution 111
4.10 Simulation 113
Review Exercises 116
Key Terms 118

4.1 Random Variables

To be more explicit about the concept of a random variable, let us refer again to the used car example of page 49 and the corresponding probabilities shown in Figure 3.8. Now let us refer to M_1 (low current mileage), P_1 (moderate price), and C_1 (inexpensive to operate) as preferred attributes. Suppose we are interested only in the number of preferred attributes a used car will get. To find the probabilities that a used car will get 0, 1, 2, or 3 preferred attributes, let us refer to Figure 4.1, which is like Figure 3.8 in Chapter 3 except that we indicate for each outcome the number of preferred attributes. Adding the respective probabilities, we find that for 0 preferred attributes the probability is

$$0.07 + 0.06 + 0.03 + 0.02 = 0.18$$

and for one preferred attribute the probability is

$$0.01 + 0.01 + 0.16 + 0.10 + 0.02 + 0.05 + 0.14 + 0.01 = 0.50$$

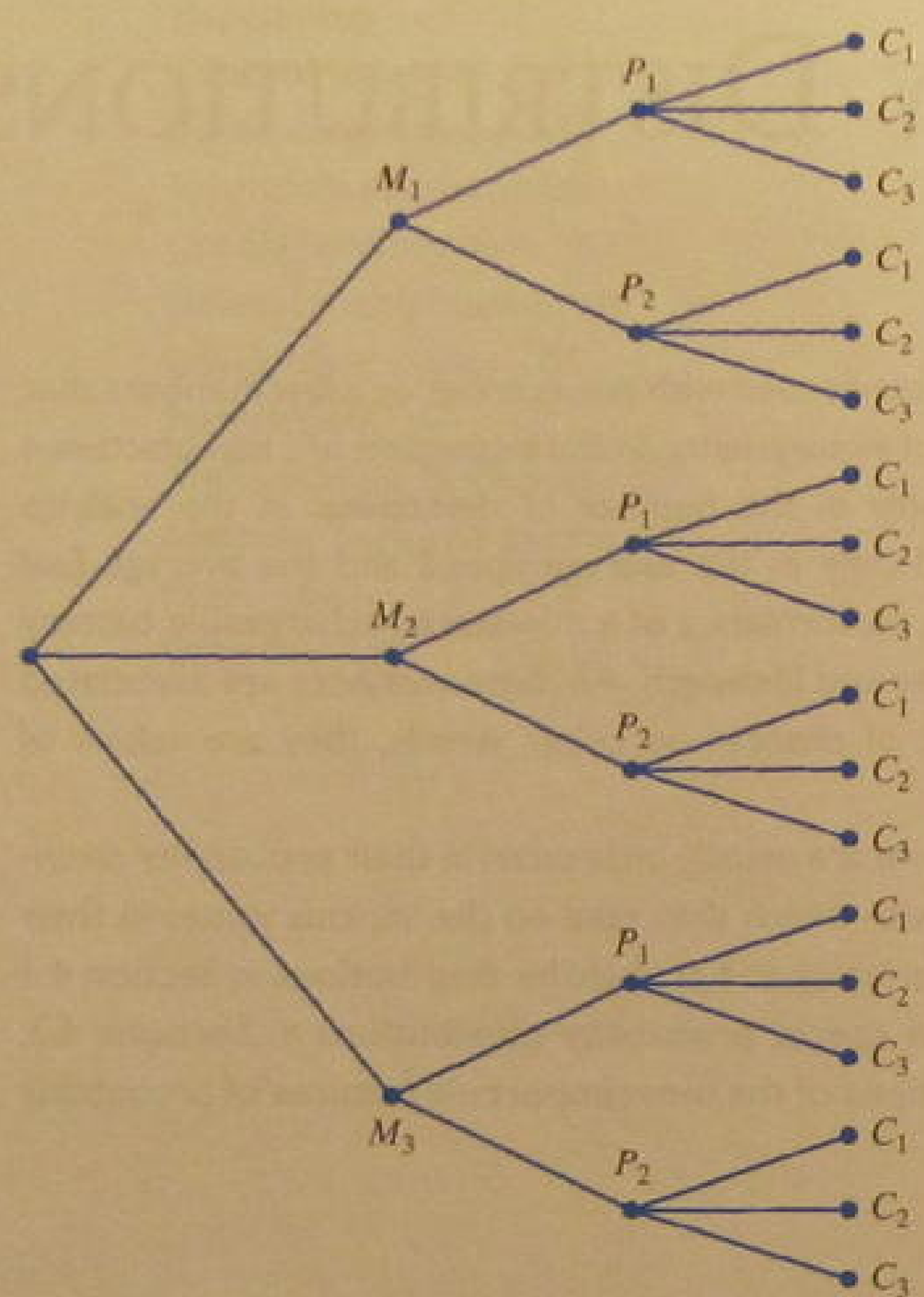
For two preferred attributes, the probability is

$$0.06 + 0.07 + 0.02 + 0.09 + 0.05 = 0.29$$

and for three preferred attributes the probability is 0.03.

These results may be summarized, as in the following table, where x denotes a possible number of preferred attributes

x	0	1	2	3
Probability	0.18	0.50	0.29	0.03



	Probability	Number of preferred ratings
$M_1P_1C_1$	0.03	3
$M_1P_1C_2$	0.06	2
$M_1P_1C_3$	0.07	2
$M_1P_2C_1$	0.02	2
$M_1P_2C_2$	0.01	1
$M_1P_2C_3$	0.01	1
$M_2P_1C_1$	0.09	2
$M_2P_1C_2$	0.16	1
$M_2P_1C_3$	0.10	1
$M_2P_2C_1$	0.02	1
$M_2P_2C_2$	0.07	0
$M_2P_2C_3$	0.06	0
$M_3P_1C_1$	0.05	2
$M_3P_1C_2$	0.05	1
$M_3P_1C_3$	0.14	1
$M_3P_2C_1$	0.01	1
$M_3P_2C_2$	0.03	0
$M_3P_2C_3$	0.02	0

Figure 4.1
Used cars and numbers of preferred attributes

The numbers 0, 1, 2, and 3 in this table are values of a **random variable**—the number of preferred attributes. Corresponding to each elementary outcome in the sample space there is exactly one value x for this random variable. That is, the random variable may be thought of as a function defined over the elements of the sample space. This is how we define random variables in general; they are functions defined over the elements of a sample space.

Random variables

A **random variable** is any function that assigns a numerical value to each possible outcome.

The numerical value should be chosen to quantify an important characteristic of the outcome.

Random variables are denoted by capital letters X, Y, and so on, to distinguish them from their possible values given in lowercase x, y.

To find the probability that a random variable will take on any one value within its range, we proceed as in the above example. Indeed, the table which we obtained displays another function, called the **probability distribution** of the random variable. To denote the values of a probability distribution, we shall use such symbols as $f(x)$, $g(x)$, $\varphi(y)$, $h(z)$, and so on. Strictly speaking, it is the function $f(x) = P(X = x)$ which assigns probability to each possible outcome x that is called the probability distribution. However, we will follow the common practice of also calling $f(x)$ the probability distribution, with the understanding that we are referring to the function and that the range of x values is part of the definition.

Random variables are usually classified according to the number of values they can assume. In this chapter we shall limit our discussion to **discrete random variables**, which can take on only a finite number, or a countable infinity of values; **continuous random variables** are taken up in Chapter 5.

Whenever possible, we try to express probability distributions by means of equations; otherwise, we must give a table that actually exhibits the correspondence between the values of the random variable and the associated probabilities. For instance,

$$f(x) = \frac{1}{6} \qquad \text{for } x = 1, 2, 3, 4, 5, 6$$

gives the probability distribution for the number of points we roll with a balanced die.

Of course, not every function defined for the values of a random variable can serve as a probability distribution. Since the values of probability distributions are probabilities and one value of a random variable must always occur, it follows that if $f(x)$ is a probability distribution, then

$$f(x) \geq 0 \qquad \text{for all } x$$

and

$$\sum_{\text{all } x} f(x) = 1$$

Probability distributions

The **probability distribution** of a discrete random variable X is a list of the possible values of X together with their probabilities

$$f(x) = P[X = x]$$

The probability distribution always satisfies the conditions

$$f(x) \geq 0 \quad \text{and} \quad \sum_{\text{all } x} f(x) = 1$$

EXAMPLE **Checking for nonnegativity and total probability equals one**

Check whether the following can serve as probability distributions:

(a) $f(x) = \dfrac{x-2}{2} \qquad \text{for } x = 1, 2, 3, 4$

(b) $h(x) = \dfrac{x^2}{25} \qquad \text{for } x = 0, 1, 2, 3, 4$

Solution

(a) This function cannot serve as a probability distribution because $f(1)$ is negative.

(b) The function cannot serve as a probability distribution because the sum of the five probabilities is $\dfrac{6}{5}$ and not 1. ■

It is often helpful to visualize probability distributions by means of graphs like those of Figure 4.2. The one on the left is called a **probability histogram;** the areas

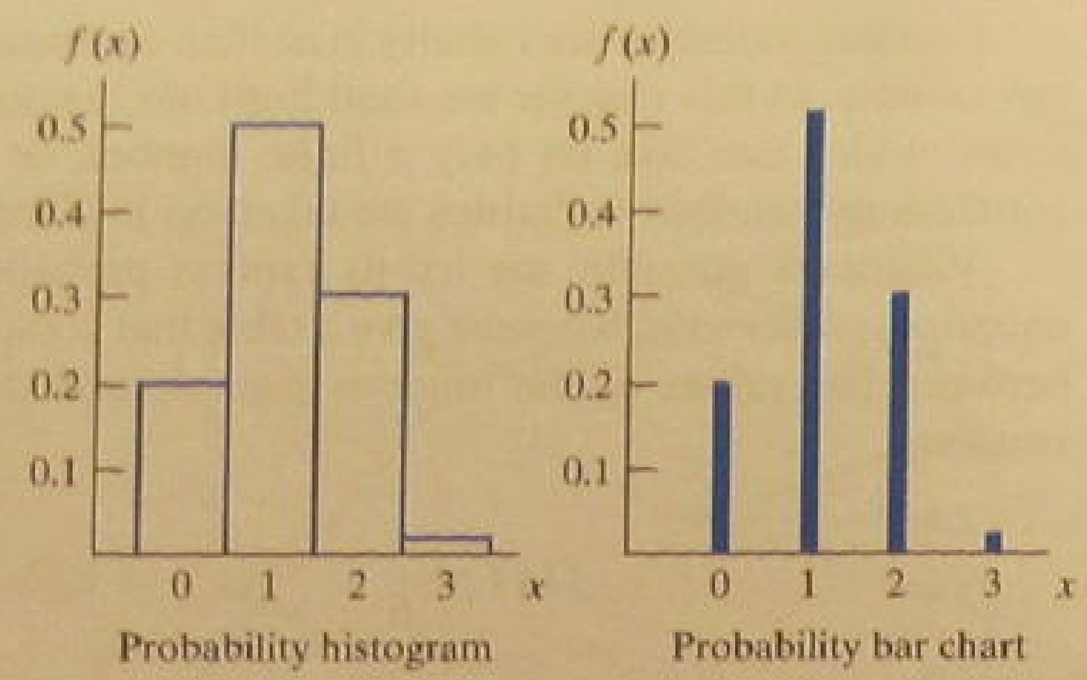


Figure 4.2
Graphs of the probability distribution of the number of preferred attributes

of the rectangles are equal to the corresponding probabilities so their heights are proportional to the probabilities. The bases touch so that there are no gaps between the rectangles representing the successive values of the random variable. The one on the right is called a **bar chart**; the heights of the rectangles are also proportional to the corresponding probabilities, but they are narrow and their width is of no significance.

As we see later in this chapter, there are many problems in which we are interested not only in the probability $f(x)$ that the value of a random variable is x, but also in the probability $F(x)$ that the value of a random variable is less than or equal to x. Specifically,

$$F(x) = P[X \leq x] \quad \text{for all } -\infty < x < \infty$$

and we refer to the function $F(x)$ as the **cumulative distribution function** or just the **distribution function** of the random variable. For any value x, it adds up, or accumulates, all the probability assigned to that value and smaller values.

Referring to the used car example and basing our calculations on the table on page 81, we get

x	0	1	2	3
$F(x)$	0.18	0.68	0.97	1.00

for the cumulative distribution function of the number of preferred attributes.

The cumulative distribution it jumps the amount at $f(x) = 0, 1, 2, 3$ and is constant between the values in the table as illustrated in Figure 4.3. The solid dots emphasize the fact that $F(x)$ takes the upper value at jumps and this makes it continuous from the right.

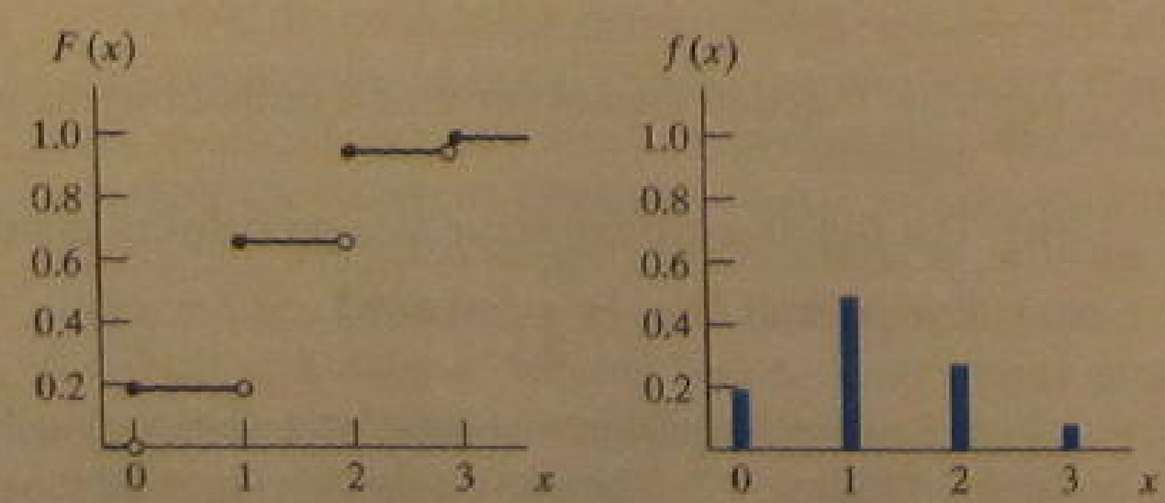


Figure 4.3
The cumulative distribution has jumps corresponding to $f(x) = P[X = x]$

4.2 The Binomial Distribution

Many statistical problems deal with the situations referred to as **repeated trials**. For example, we may want to know the probability that 1 of 5 rivets will rupture in a tensile test, the probability that 9 of 10 DVD players will run at least 1,000 hours, the probability that 45 of 300 drivers stopped at a roadblock will be wearing seatbelts, or the probability that 66 of 200 television viewers (interviewed by a rating service) will recall what products were advertised on a given program. To borrow from the language of games of chance, we might say that in each of these examples we are interested in the probability of getting x **successes** in n **trials**, or, in other words, x successes and $n - x$ failures in n attempts.

There are common features to each of the examples. They are all composed of a series of trials which we refer to as **Bernoulli trials** if the following assumptions hold.

1. **There are only two possible outcomes for each trial** (arbitrarily called "success" and "failure," without inferring that a success is necessarily desirable).
2. **The probability of success is the same for each trial.**
3. **The outcomes from different trials are independent.**

If the assumptions cannot be met, the theory we develop does not apply.

EXAMPLE **Checking the adequacy of the Bernoulli trials assumptions**

Can the following be treated as Bernoulli trials? Drivers stopped at a roadblock will be checked for failure to wear a seatbelt.

Solution There are only two outcomes, and we call not wearing a seatbelt a success. (Success in this context does not mean success in life.)

If all cars are treated alike, their drivers would all have the same probability of not wearing a seatbelt. If drivers are grouped by age, you may need different probabilities for persons under 20 than for those 30 to 40 years old. Then you would not have Bernoulli trials.

The results on seatbelt wear, for different drivers, should be independent. There is no obvious common cause. If someone caught without a seatbelt were to inform oncoming cars about the checkpoint, that would introduce dependence. ■

In the problems we study in this section, we add the additional assumption that the number of trials is fixed in advance.

4. **There are a fixed number n of Bernoulli trials conducted.**

EXAMPLE **Binomial probability distribution $n = 3$**

When a relay tower for wireless phone service breaks down, it quickly becomes an expensive proposition for the phone company, and the cost increases with the time it is inoperable. From company records, it is postulated that the probability is 0.90 that the breakdown can be repaired within one hour. For the next three breakdowns, on different days and different towers,

(a) List all possible outcomes in terms of success, S, repaired within one hour, and failure, F, not repaired within one hour.

(b) Find the probability distribution of the number of successes, X, among the 3 repairs.

Solution **(a)** We write FSS for the outcome where the first repair is not made within one hour and the second and third are successful. The $2 \times 2 \times 2 = 8$ possible outcomes can be systematically arranged as follows:

FFF	FFS	FSS	SSS
	FSF	SFS	
	SFF	SSF	
$X = 0$	$X = 1$	$X = 2$	$X = 3$

where the number of successes X is the same for each outcome in a column. This value is recorded at the bottom of the column.

(b) The results of repairs on different days and different towers should be independent. Also, the probability of success 0.90 is the same for each repair. Therefore, the probability that $X = 0$ is $0.1 \times 0.1 \times 0.1 = 0.001$. Next, the probability of FFS is $0.1 \times 0.1 \times 0.9 = 0.009$ and both SFS and FSS have the same probability. Consequently, the probability that $X = 1$ is $3 \times 0.009 = 0.027$. Note that the number of outcomes where $X = 1$ is just the number of ways to select 1 or 3 trials for an S and the others are F.
By similar reasoning, the probability that $X = 2$ is

$$3\,(0.1 \times 0.9 \times 0.9) = 0.243 = \binom{3}{2}\,(0.9)^2\,(0.1)^1$$

Finally, the probability that $X = 3$, no repair takes over one hour, is

$$0.9 \times 0.9 \times 0.9 = 0.729 = \binom{3}{3}\,(0.9)^3\,(0.1)^0$$

All of these probabilities can be expressed by the formula

$$f(x) = P(X = x) = \binom{3}{x}\,(0.9)^x\,(0.1)^{3-x} \qquad \text{for } x = 0, 1, 2, 3$$

This is the probability distribution for a binomial random variable when the success probability is $p = 0.9$ and there are $n = 3$ trials. ■

Let X be the random variable that equals the number of successes in n trials. To obtain probabilities concerning X, we proceed as follows: If p and $1 - p$ are the probabilities of success and failure on any one trial, then the probability of getting x successes and $n - x$ failures, in some specific order, is $p^x(1-p)^{n-x}$. Clearly, in this product of p's and $(1 - p)$'s there is one factor p for each success, one factor $1 - p$ for each failure. The x factors p and $n - x$ factors $1 - p$ are all multiplied together by virtue of the generalized multiplication rule for more than two independent events. Since this probability applies to any point of the sample space that represents x successes and $n - x$ failures (in any specific order), we have only to count how many points of this kind there are, and then multiply $p^x(1-p)^{n-x}$ by this. Clearly, the number of ways in which we can select the x trials on which there is to be a success is $\binom{n}{x}$, the number of combinations of x objects selected from a set of n objects, and we thus arrive at the following result:

Binomial distribution

$$b(x; n, p) = \binom{n}{x}\,p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

This probability distribution is called the **binomial distribution** because for $x = 0, 1, 2, \ldots,$ and n, the values of the probabilities are the successive terms of the binomial expansion of $[p + (1 - p)]^n$; for the same reason, the combinatorial quantities $\binom{n}{x}$ are referred to as **binomial coefficients**. Actually, the preceding equation defines a family of probability distributions, with each member characterized by a given value of the **parameter** p and the number of trials n.

EXAMPLE **Evaluating binomial probabilities**

It has been claimed that in 60% of all solar-heat installations the utility bill is reduced by at least one-third. Accordingly, what are the probabilities that the utility bill will be reduced by at least one-third in

(a) four of five installations;

(b) at least four of five installations?

Solution **(a)** Substituting $x = 4$, $n = 5$, and $p = 0.60$ into the formula for the binomial distribution, we get

$$b(4; 5, 0.60) = \binom{5}{4}(0.60)^4(1 - 0.60)^{5-4}$$
$$= 0.259$$

(b) Substituting $x = 5$, $n = 5$, and $p = 0.60$ into the formula for the binomial distribution, we get

$$b(5; 5, 0.60) = \binom{5}{5}(0.60)^5(1 - 0.60)^{5-5}$$
$$= 0.078$$

and the answer is $b(4; 5, 0.60) + b(5; 5, 0.60) = 0.259 + 0.078 = 0.337$. ■

If n is large, the calculation of binomial probabilities can become quite tedious. Many statistical software programs have binomial distribution commands (see Exercises 4.30 and 4.31), as do some statistical calculators. Otherwise it is convenient to refer to special tables. Table 1 at the end of the book gives the values of

$$B(x; n, p) = \sum_{k=0}^{x} b(k; n, p) \qquad \text{for } x = 0, 1, 2, \ldots, n$$

for $n = 2$ to $n = 20$ and $p = 0.05, 0.10, 0.15, \ldots, 0.90, 0.95$. We tabulated the **cumulative probabilities** rather than the values of $b(x; n, p)$, because the values of $B(x; n, p)$ are the ones needed more often in statistical applications. Note, however, that the values of $b(x; n, p)$ can be obtained by subtracting adjacent entries in Table 1. Because the two cumulative probabilities $B(x; n, p)$ and $B(x - 1; n, p)$ differ by the single term $b(x; n, p)$

$$b(x; n, p) = B(x; n, p) - B(x - 1; n, p)$$

where $B(-1) = 0$. The examples that follow illustrate the direct use of Table 1 and the use of this relationship.

EXAMPLE **Evaluating cumulative binomial probabilities**

If the probability is 0.05 that a certain wide-flange column will fail under a given axial load, what are the probabilities that among 16 such columns

(a) at most two will fail;

(b) at least four will fail?

Solution **(a)** Table 1 shows that $B(2; 16, 0.05) = 0.9571$.

(b) Since

$$\sum_{x=4}^{16} b(x; 16, 0.05) = 1 - B(3; 16, 0.05)$$

Table 1 yields $1 - 0.9930 = 0.0070$. ■

EXAMPLE **Finding a binomial probability using cumulative binomial probabilities**

The manufacture of large high-definition LCD panels is difficult, and a moderately high proportion have too many defective pixels to pass inspection. If the probability is 0.3 that an LCD panel will not pass inspection, what is the probability that 6 of 18 panels, randomly selected from production, will not pass inspection?

Solution Using the relationship to cumulative probabilities and then looking up these probabilities in Table 1, we get

$$\begin{aligned} b(6, 18, 0.30) &= B(6, 18, 0.30) - B(5, 18, 0.30) \\ &= 0.7217 - 0.5344 \\ &= 0.1873 \end{aligned}$$

■

The following example illustrates the use of the binomial distribution in a problem of decision making.

EXAMPLE **A binomial probability to guide decision making**

A manufacturer of external hard drives claims that only 10% of his drives require repairs within the warranty period of 12 months. If 5 of 20 of his drives required repairs within the first year, does this tend to support or refute the claim?

Solution Let us first find the probability that 5 or more of 20 of the drives will require repairs within a year when the probability that any one will require repairs within a year is 0.10. Using Table 1, we get

$$\begin{aligned} \sum_{x=5}^{20} b(x; 20, 0.10) &= 1 - B(4; 20, 0.10) \\ &= 1 - 0.9568 \\ &= 0.0432 \end{aligned}$$

Since this probability is very small, it would seem reasonable to reject the hard drive manufacturer's claim. ■

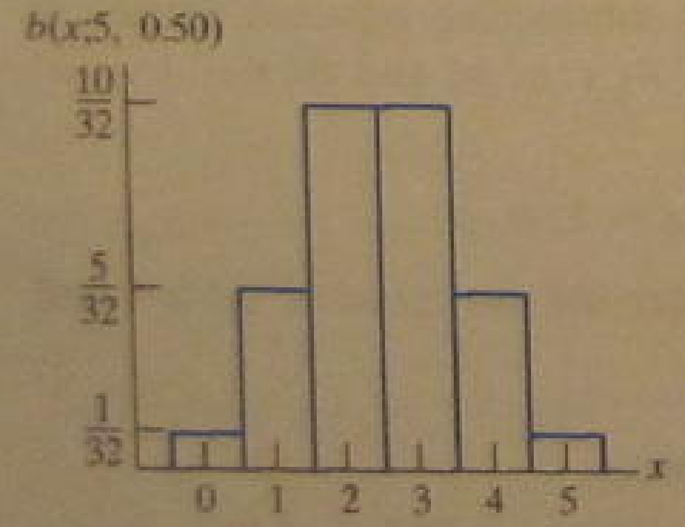


Figure 4.4
Symmetrical binomial distribution with $n = 5$ and $p = 0.50$

Important information about the shape of binomial distributions is shown in Figures 4.4 and 4.5. First, if $p = 0.50$, the equation for the binomial distribution is

$$b(x; n, 0.50) = \binom{n}{x} (0.5)^n$$

and since

$$\binom{n}{n-x} = \binom{n}{x}$$

it follows that $b(x; n, 0.50) = b(n - x; n, 0.50)$. For any n, the binomial distribution with $p = 0.5$ is a **symmetrical distribution**. This means that the probability histograms of such binomial distributions are symmetrical, as is illustrated in Figure 4.4. Note, however, that if p is less than 0.50, it is more likely that X will be

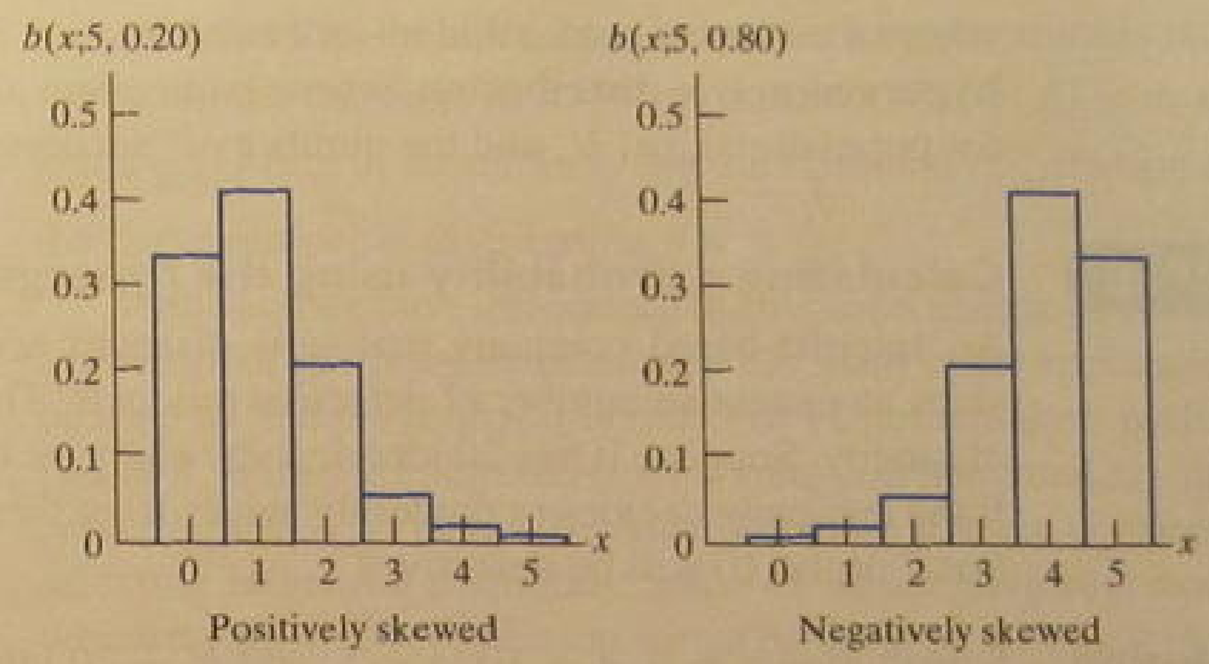


Figure 4.5
Skewed binomial distributions with $n = 5$ and $p = 0.20$ and $p = 0.80$

small rather than large compared to $n/2$ and that the opposite is true if p is greater than 0.50. This is illustrated in Figure 4.5, showing the histograms of binomial distributions with $n = 5$ and $p = 0.20$ and $p = 0.80$. These two are mirror images of each other as can be verified more generally in Exercise 4.7.

Finally, a probability distribution that has a probability histogram like either of those in Figure 4.5 is said to be a long-tailed or **skewed distribution**. It is said to be a **positively skewed distribution** if the tail is on the right, and it is said to be **negatively skewed** if the tail is on the left.

4.3 The Hypergeometric Distribution

Suppose that we are interested in the number of defectives in a sample of n units drawn without replacement from a lot containing N units, of which a are defective. Let the sample be drawn in such a way that at each successive drawing, whatever units are left in the lot have the same chance of being selected. The probability that the first drawing will yield a defective unit is $\frac{a}{N}$, but for the second drawing it is $\frac{a-1}{N-1}$ or $\frac{a}{N-1}$, depending on whether or not the first unit drawn was defective. Thus, the trials are not independent, the fourth assumption underlying the binomial distribution is not met, and the binomial distribution does not apply. Note that the binomial distribution would apply if we do **sampling with replacement**, namely, if each unit selected for the sample is replaced before the next one is drawn.

To solve the problem of **sampling without replacement** (that is, as we originally formulated the problem), let us proceed as follows: The x successes (defectives) can be chosen in $\binom{a}{x}$ ways, the $n - x$ failures (nondefectives) can be chosen in $\binom{N-a}{n-x}$ ways, and hence, x successes and $n - x$ failures can be chosen in $\binom{a}{x}\binom{N-a}{n-x}$ ways. Also, n objects can be chosen from a set of N objects in $\binom{N}{n}$ ways, and if we consider all the possibilities as equally likely, it follows that for sampling without replacement the probability of getting "x successes in n trials" is

Hypergeometric distribution

$$h(x; n, a, N) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}} \qquad \text{for } x = 0, 1, \ldots, n$$

where x cannot exceed a and $n - x$ cannot exceed $N - a$. This equation defines the **hypergeometric distribution**, whose parameters are the sample size n, the lot size (or population size) N, and the number of "successes" in the lot a.

EXAMPLE **Calculating a probability using the hypergeometric distribution**

An Internet-based company that sells discount accessories for cell phones often ships an excessive number of defective products. The company needs better control of quality. Suppose it has 20 identical car chargers on hand but that 5 are defective. If the company decides to randomly select 10 of these items, what is the probability that 2 of the 10 will be defective?

Solution Substituting $x = 2$, $n = 10$, $a = 5$, and $N = 20$ into the formula for the hypergeometric distribution, we get

$$h(2; 10, 5, 20) = \frac{\binom{5}{2}\binom{15}{8}}{\binom{20}{10}} = \frac{10 \times 6{,}435}{184{,}756} = 0.348$$

■

In the preceding example, n was not small compared to N, and if we had made the mistake of using the binomial distribution with $n = 10$ and $p = \frac{5}{20} = 0.25$ to calculate the probability of two defectives, the result would have been 0.282, which is much too small. However, when n is small compared to N, less than $\frac{N}{10}$, the composition of the lot is not seriously affected by drawing the sample without replacement, and the binomial distribution with the parameters n and $p = \frac{a}{N}$ will yield a good approximation.

EXAMPLE **A numerical comparison of the hypergeometric and binomial distributions**

Repeat the preceding example but with 100 car chargers, of which 25 are defective, by using

(a) the formula for the hypergeometric distribution;
(b) the formula for the binomial distribution as an approximation.

Solution (a) Substituting $x = 2$, $n = 10$, $a = 25$, and $N = 100$ into the formula for the hypergeometric distribution, we get

$$h(2; 10, 25, 100) = \frac{\binom{25}{2}\binom{75}{8}}{\binom{100}{10}} = 0.292$$

(b) Substituting $x = 2$, $n = 10$, and $p = \frac{25}{100} = 0.25$ into the formula for the binomial distribution, we get

$$b(2; 10, 0.25) = \binom{10}{2}(0.25)^2(1 - 0.25)^{10-2}$$
$$= 0.282$$

■

Observe that the difference between the two values is only 0.010. In general, it can be shown that $h(x;\, n, a, N)$ approaches $b(x;\, n, p)$ with $p = \frac{a}{N}$ when $N \to \infty$, and a good rule of thumb is to use the binomial distribution as an approximation to the hypergeometric distribution if $n \leq \frac{N}{10}$.

Although we have introduced the hypergeometric distribution in connection with a problem of sampling inspection, it has many other applications. For instance, it can be used to find the probability that 3 of 12 homemakers prefer Brand A detergent to Brand B, if they are selected from among 200 homemakers among whom 40 actually prefer Brand A to Brand B. Also, it can be used in connection with a problem of selecting industrial diamonds, some of which have superior qualities and some of which do not, or in connection with a problem of sampling income tax returns, where a among N returns filed contain questionable deductions, and so on.

Exercises

4.1 Suppose that a probability of $\frac{1}{12}$ is assigned to each point of the sample space of part (a) of Exercise 3.1 on page 53. Find the probability distribution of the total number of lakes and streams that are contaminated.

4.2 An experiment consists of four tosses of a coin. Denoting the outcomes *HHTH*, *THTT*, . . . and assuming that all 16 outcomes are equally likely, find the probability distribution for the total number of heads.

4.3 Determine whether the following can be probability distributions of a random variable which can take on only the values 1, 2, 3, and 4.

(a) $f(1) = 0.16$, $f(2) = 0.28$, $f(3) = 0.28$, and $f(4) = 0.28$;

(b) $f(1) = 0.24$, $f(2) = 0.24$, $f(3) = 0.24$, and $f(4) = 0.24$;

(c) $f(1) = 0.37$, $f(2) = 0.31$, $f(3) = 0.34$, and $f(4) = -0.02$.

4.4 Check whether the following can define probability distributions, and explain your answers.

(a) $f(x) = \frac{x}{14}$ for $x = 0, 1, 2, 3, 4$

(b) $f(x) = \frac{3 - x^2}{4}$ for $x = 0, 1, 2$

(c) $f(x) = \frac{1}{5}$ for $x = 5, 6, 7, 8, 9$

(d) $f(x) = \frac{2x + 1}{50}$ for $x = 1, 2, 3, 4, 5$

4.5 Given that $f(x) = \frac{k}{2^x}$ is a probability distribution for a random variable that can take on the values $x = 0, 1, 2, 3$, and 4, find k.

4.6 With reference to Exercise 4.5, find an expression for the distribution function $F(x)$ of the random variable.

4.7 Prove that $b(x;\, n, p) = b(n - x;\, n, 1 - p)$.

4.8 Prove that $B(x;\, n, p) = 1 - B(n - x - 1;\, n, 1 - p)$.

4.9 Do the assumptions for Bernoulli trials appear to hold? Explain. If the assumptions hold, identify success and the probability of interest.

(a) A TV ratings company will use their electronic equipment to check a sample of homes around the city to see which ones have a set tuned to the mayor's speech on the local channel.

(b) Among 6 nuclear power plants in a state, 2 have had serious violations in last five years. Two plants will be selected at random, one after the other, and the outcome of interest is a serious violation in the last five years.

4.10 What conditions for the binomial distribution, if any, fail to hold in the following situations?

(a) For each of a company's eight production facilities, record whether or not there was an accident in the past week. The largest facility has three times the number of production workers as the smallest facility.

(b) For each of three shifts, the number of units produced will be compared with the long-term average of 560 and it will be determined whether or not production exceeds 560 units. The second shift will know the result for the first shift before they start working, and the third shift will start with the knowledge of how the first two shifts performed.

4.11 Which conditions for the binomial distribution, if any, fail to hold in the following situations?

(a) The number of persons having a cold at a family reunion attended by 30 persons.

(b) Among 8 projectors in the department office, 2 do not work properly but are not marked defective. Two are selected and the number that do not work properly will be recorded.

4.12 Use Table 1 to find

(a) $B(8; 16, 0.40)$;

(b) $b(8; 16, 0.40)$;

(c) $B(9; 12, 0.60)$;

(d) $b(9; 12, 0.60)$;

(e) $\sum_{k=6}^{20} b(k; 20, 0.15)$;

(f) $\sum_{k=6}^{9} b(k; 9, 0.70)$.

4.13 Use Table 1 to find

(a) $B(7; 19, 0.45)$;

(b) $b(7; 19, 0.45)$;

(c) $B(8; 10, 0.95)$;

(d) $b(8; 10, 0.95)$;

(e) $\sum_{k=4}^{10} b(k; 10, 0.35)$;

(f) $\sum_{k=2}^{4} b(k; 9, 0.30)$.

4.14 Rework the decision problem on page 88, supposing that only 3 of the 20 hard drives required repairs within the first year.

4.15 Human error is given as the reason for 75% of all accidents in a plant. Use the formula for the binomial distribution to find the probability that human error will be given as the reason for two of the next four accidents.

4.16 If the probability is 0.40 that steam will condense in a thin-walled aluminum tube at 10 atm pressure, use the formula for the binomial distribution to find the probability that, under the stated conditions, steam will condense in 4 of 12 such tubes.

4.17 During one stage in the manufacture of integrated circuit chips, a coating must be applied. If 70% of chips receive a thick enough coating, use Table 1 to find the probabilities that, among 15 chips:

(a) at least 12 will have thick enough coatings;

(b) at most 6 will have thick enough coatings;

(c) exactly 10 will have thick enough coatings.

4.18 The probability that the noise level of a wide-band amplifier will exceed 2 dB is 0.05. Use Table 1 to find the probabilities that among 12 such amplifiers the noise level of

(a) one will exceed 2 dB;

(b) at most two will exceed 2 dB;

(c) two or more will exceed 2 dB.

4.19 An agricultural cooperative claims that 90% of the watermelons shipped out are ripe and ready to eat. Find the probabilities that among 18 watermelons shipped out

(a) all 18 are ripe and ready to eat;

(b) at least 16 are ripe and ready to eat;

(c) at most 14 are ripe and ready to eat.

4.20 A quality-control engineer wants to check whether (in accordance with specifications) 95% of the electronic components shipped by his company are in good working condition. To this end, he randomly selects 15 from each large lot ready to be shipped and passes the lot if the selected components are all in good working condition; otherwise, each of the components in the lot is checked. Find the probabilities that the quality-control engineer will commit the error of

(a) holding a lot for further inspection even though 95% of the components are in good working condition;

(b) letting a lot pass through without further inspection even though only 90% of the components are in good working condition;

(c) letting a lot pass through without further inspection even though only 80% of the components are in good condition.

4.21 A food processor claims that at most 10% of her jars of instant coffee contain less coffee than claimed on the label. To test this claim, 16 jars of her instant coffee are randomly selected and the contents are weighed; her claim is accepted if fewer than 3 of the jars contain less coffee than claimed on the label. Find the probabilities that the food processor's claim will be accepted when the actual percentage of her jars containing less coffee than claimed on the label is

(a) 5%; (b) 10%; (c) 15%; (d) 20%.

4.22 Refer to Exercise 4.2.

(a) Determine the cumulative probability distribution $F(x)$.

(b) Graph the probability distribution of $f(x)$ as a bar chart and below it graph $F(x)$.

4.23 Four emergency radios are available for rescue workers but one does not work properly. Two randomly selected radios are taken on a rescue mission. Let X be the number that work properly between the two.

(a) Determine the probability distribution $f(x)$ of X.

(b) Determine the cumulative probability distribution $F(x)$ of X.

(c) Graph $f(x)$ as a bar chart and below it graph $F(x)$.

4.24 Among the 12 solar collectors on display at a trade show, 9 are flat-plate collectors and the others are concentrating collectors. If a person visiting the show randomly selects 4 of the solar collectors to check out, what is the probability that 3 of them will be flat-plate collectors?

4.25 A quality-control engineer inspects a random sample of 3 batteries from each lot of 24 car batteries ready to be shipped. If such a lot contains 6 batteries with slight defects, what are the probabilities that the inspector's sample will contain

(a) none of the batteries with defects?

(b) only one of the batteries with defects?

(c) at least two of the batteries with defects?

4.26 If 6 of 18 new buildings in a city violate the building code, what is the probability that a building inspector, who randomly selects 4 of the new buildings for inspection, will catch

(a) none of the buildings that violate the building code?

(b) 1 of the new buildings that violate the building code?

(c) 2 of the new buildings that violate the building code?

(d) at least 3 of the new buildings that violate the building code?

4.27 Among the 16 cities that a professional society is considering for its next 3 annual conventions, 7 are in the western part of the United States. To avoid arguments, the selection is left to chance. If none of the cities can be chosen more than once, what are the probabilities that

(a) none of the conventions will be held in the western part of the United States?

(b) all of the conventions will be held in the western part of the United States?

4.28 A shipment of 120 burglar alarms contains 5 that are defective. If 3 of these alarms are randomly selected and shipped to a customer, find the probability that the customer will get one bad unit by using

(a) the formula for the hypergeometric distribution;

(b) the formula for the binomial distribution as an approximation.

4.29 Among the 300 employees of a company, 240 are union members, while the others are not. If 8 of the employees are chosen by lot to serve on the committee which administers the pension fund, find the probability that 5 of them will be union members while the others are not, by using

(a) the formula for hypergeometric distribution;

(b) the formula for binomial distribution as an approximation.

4.30 Binomial probabilities can be calculated using *MINITAB*.

> **Dialog box:**
> **Calc > Probability Distribution > Binomial**
> Choose **Probability.**
> Enter 7 in **Number of trials** and .33 in **Probability of success**.
> Choose **Input constant** and enter 2.
> Click **OK**.

Output:

Binomial with $n = 7$ and $p = 0.330000$

x	$P(X = x)$
2.00	0.3088

Find the binomial probabilities with $n = 27$ and $p = 0.47$.

4.31 Cumulative binomial probabilities can be calculated using *MINITAB*.

> **Dialog Box:**
> **Calc > Probability Distribution > Binomial**
> Choose **Cumulative Distribution**.
> Enter 7 in **Number of trials** and .33 in **Probability of success**.
> Choose **Input constant** and enter 2.
> Click **OK**.

Output:

Binomial with $n = 7$ and $p = 0.330000$

x	$P(X <= x)$
2.00	0.5783

Find the cumulative binomial probabilities with $n = 27$ and $p = 0.47$.

4.4 The Mean and the Variance of a Probability Distribution

Besides the binomial and hypergeometric distributions, there are many other probability distributions that have important engineering applications. However, before we go any further, let us discuss some general characteristics of probability distributions.

One such characteristic, that of the symmetry or **skewness** of a probability distribution, was illustrated in Figures 4.4 and 4.5; two other characteristics are

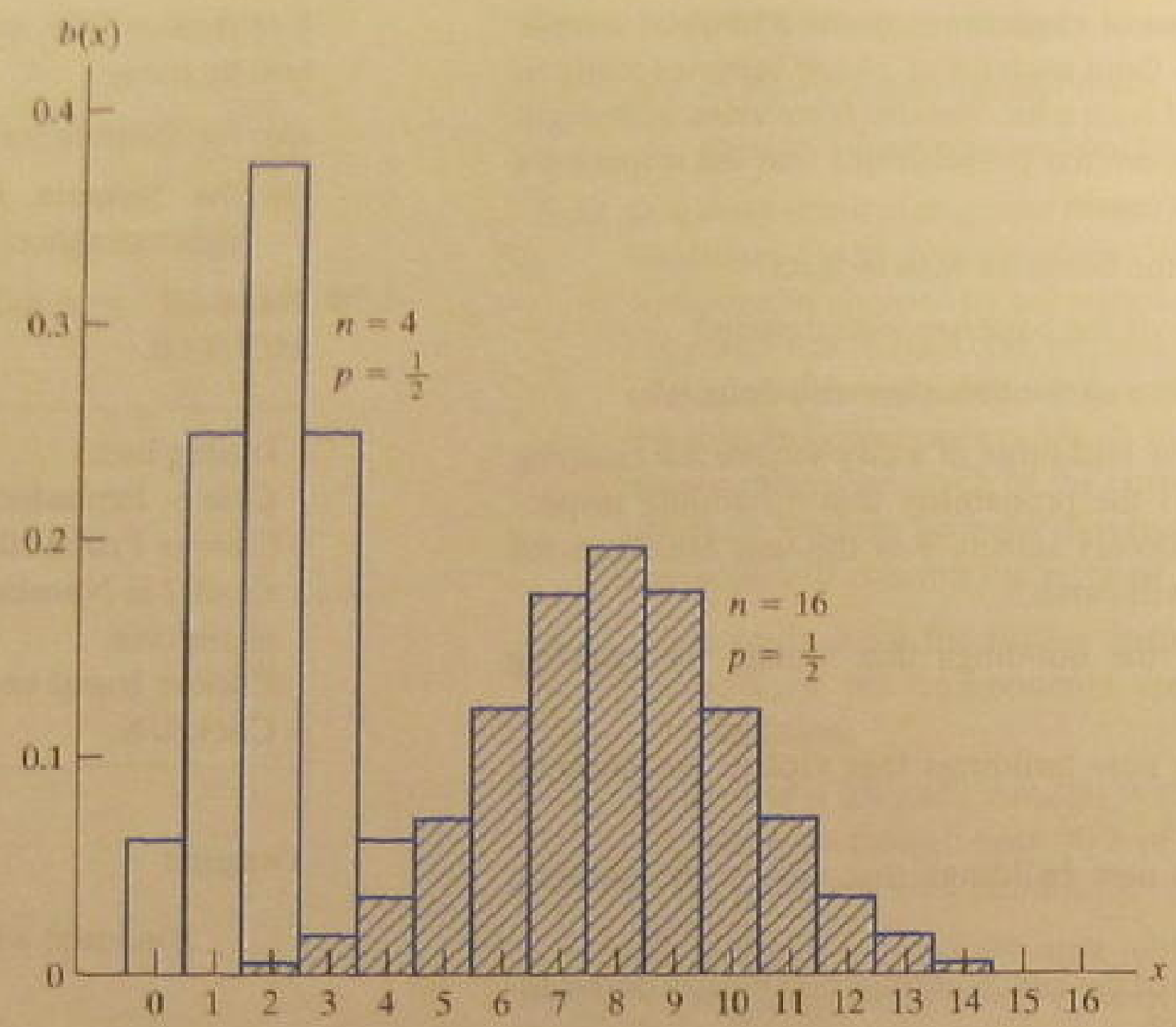


Figure 4.6
Probability histograms of two binomial distributions

apparent in Figure 4.6, which shows the probability histograms of two binomial distributions. One of these binomial distributions has the parameters $n = 4$ and $p = 1/2$, and the other has the parameters $n = 16$ and $p = 1/2$. Essentially, these two probability distributions differ in two respects. The first probability distribution is centered about $x = 2$, whereas the other (whose histogram is shaded) is centered about $x = 8$, and we say that the two distributions differ in their **location**. Another distinction is that the histogram of the second distribution is broader, and we say that the two distributions differ in **variation**. To make such comparisons more specific, we shall introduce in this section two of the most important statistical measures, describing the location and the variation of a probability distribution—the **mean** and the **variance**, respectively.

The mean of a probability distribution is simply the mathematical expectation of a random variable having that distribution. If a random variable X takes on the values $x_1, x_2, \ldots,$ or x_k, with the probabilities $f(x_1), f(x_2), \ldots,$ and $f(x_k)$, its mathematical expectation or expected value is

$$x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + \cdots + x_k \cdot f(x_k) = \sum \text{(value)} \times \text{(probability)}$$

using the $\sum$ notation. We can thus write

Mean of discrete probability distribution

$$\mu = \sum_{\text{all } x} x \cdot f(x)$$

where the mean is denoted by the Greek letter μ (*mu*). The mean of a probability distribution measures its center in the sense of an average, or by analogy to physics, in the sense of a center of gravity. Note that the above formula for μ is, in fact, that for the **first moment about the origin** of a discrete system of masses $f(x)$ arranged on a weightless straight line at distances x from the origin. We do not have to divide

here by

$$\sum_{\text{all } x} f(x)$$

as we do in the usual formula for the x-coordinate of the center of gravity, since the sum equals 1 by definition.

EXAMPLE **The mean number of heads in three tosses**

Find the mean of the probability distribution of the number of heads obtained in 3 flips of a balanced coin.

Solution The probabilities for 0, 1, 2, or 3 heads are $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}$, and $\frac{1}{8}$ as can easily be verified by counting equally likely possibilities or by using the formula for the binomial distribution with $n = 3$ and $p = \frac{1}{2}$. Thus,

$$\mu = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}$$

■

EXAMPLE **The mean number of preferred used car attributes**

With reference to the used car example and the probabilities given on page 82, find the mean of the probability distribution of the number of preferred attributes.

Solution Substituting $x = 0, 1, 2$, and 3 and the corresponding probabilities into the formula for μ, we get

$$\begin{aligned}\mu &= 0(0.18) + 1(0.50) + 2(0.29) + 3(0.03)\\ &= 1.17\end{aligned}$$

■

Returning to the second probability distribution of Figure 4.6, we could find its mean by calculating all the necessary probabilities (or by looking them up in Table 1) and substituting them into the formula for μ. However, if we reflect for a moment, we might argue that there is a 50-50 chance for a success on each trial, there are 16 trials, and it would seem reasonable to expect 8 heads and 8 tails (in the sense of a mathematical expectation). Similarly, we might argue that if a binomial distribution has the parameters $n = 200$ and $p = 0.20$, we can expect a success 20% of the time and, hence, on the average $200(0.20) = 40$ successes in 200 trials. These two values are, indeed, correct, and it can be shown in general that

Mean of binomial distribution

$$\boxed{\mu = n \cdot p}$$

for the mean of a binomial distribution. To prove this formula, we substitute the expression that defines $b(x; n, p)$ into the formula for μ, and we get

$$\mu = \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Then, making use of the fact that

$$\frac{x}{x!} = \frac{1}{(x-1)!}$$

and $n! = n(n-1)!$, and factoring out n and p, we obtain

$$\mu = np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1}(1-p)^{n-x}$$

where the summation starts with $x = 1$ since the original summand is zero for $x = 0$. If we now let $y = x - 1$ and $m = n - 1$, we obtain

$$\mu = np \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^{y}(1-p)^{m-y}$$

and this last sum can easily be recognized as that of all the terms of the binomial distribution with the parameters m and p. Hence, this sum equals 1 and it follows that $\mu = np$.

EXAMPLE **Using $\mu = np$ to find the mean number of heads in three tosses**

Find the mean of the probability distribution of the number of heads obtained in 3 flips of a balanced coin.

Solution For a binomial distribution with $n = 3$ and $p = \frac{1}{2}$, we get $\mu = 3 \cdot \frac{1}{2} = \frac{3}{2}$, and this agrees with the result obtained on page 95. ■

The formula $\mu = np$ applies, of course, only to binomial distributions, but for other special distributions, we can, similarly, express the mean in terms of their parameters. For instance, for the mean of the hypergeometric distribution with the parameters n, a, and N, we can write

Mean of hypergeometric distribution

$$\mu = n \cdot \frac{a}{N}$$

In Exercise 4.43, the reader will be asked to derive this formula by a method similar to the one we used to derive the formula for the mean of a binomial distribution.

EXAMPLE **Using the formula for the mean of a hypergeometric distribution**

With reference to the example on page 90 in which 5 of 20 car chargers were defective, find the mean of the probability distribution of the number of defectives in a sample of 10 randomly chosen for inspection.

Solution Substituting $n = 10$, $a = 5$, and $N = 20$ into the above formula for μ, we get

$$\mu = 10 \cdot \frac{5}{20} = 2.5$$

In other words, if we inspect 10 of the car chargers, we can expect 2.5 defectives, where *expect* is to be interpreted in the sense, it represents the long-run average number of defectives if 10 chargers are repeatedly selected from 20 chargers of which 5 are defective. ■

To study the second of the two properties of probability distributions mentioned on page 94, their variation, let us refer again to the two probability distributions of Figure 4.6. For the one where $n = 4$, there is a high probability of getting values close to the mean, but for the one where $n = 16$, there is a high probability of getting values scattered over considerable distances away from the mean. Using this property, it

may seem reasonable to measure the variation of a probability distribution with the quantity

$$\sum_{\text{all } x} (x - \mu) \cdot f(x)$$

namely, the average amount by which the values of the random variable deviate from the mean. Unfortunately,

$$\begin{aligned}\sum_{\text{all } x} (x - \mu) \cdot f(x) &= \sum_{\text{all } x} x \cdot f(x) - \sum_{\text{all } x} \mu \cdot f(x) \\ &= \mu - \mu \cdot \sum_{\text{all } x} f(x) = \mu - \mu = 0\end{aligned}$$

so that this expression is always equal to zero. However, since we are really interested in the magnitude of the deviations $x - \mu$ and not in their signs, it suggests itself that we average the absolute values of these deviations from the mean. This would, indeed, provide a measure of variation, but on purely theoretical grounds we prefer to work instead with the squares of the deviations from the mean. These quantities are also nonnegative, and their average is indicative of the spread or dispersion of a probability distribution. We thus define the **variance** of a probability distribution $f(x)$, or that of the random variable X which has that probability distribution, as

Variance of probability distribution

$$\sigma^2 = \sum_{\text{all } x} (x - \mu)^2 \cdot f(x)$$

where σ is the lowercase Greek letter for s. This measure is not in the same units (or dimension) as the values of the random variable, but we can adjust for this by taking the square root. This results in a measure of variation that is expressed in the same units in which the random variable is expressed. The **standard deviation** is defined as

Standard deviation of probability distribution

$$\sigma = \sqrt{\sum_{\text{all } x} (x - \mu)^2 \cdot f(x)}$$

EXAMPLE **Calculating the standard deviations of two probability distributions**

Compare the standard deviations of the two probability distributions of Figure 4.6, on page 94.

Solution Since $\mu = 4 \cdot \frac{1}{2} = 2$ for the binomial distribution with $n = 4$ and $p = \frac{1}{2}$, we find that the variance of this probability distribution is

$$\begin{aligned}\sigma^2 = (0-2)^2 \cdot \frac{1}{16} + (1-2)^2 \cdot \frac{4}{16} + (2-2)^2 \cdot \frac{6}{16} \\ + (3-2)^2 \cdot \frac{4}{16} + (4-2)^2 \cdot \frac{1}{16} = 1\end{aligned}$$

and, hence, that its standard deviation is $\sigma = 1$. Similarly, it can be shown that for the other distribution $\sigma = 2$, and we find that the second (shaded) distribution with the greater spread also has the greater standard deviation. ■

Solution **(a)** We write FSS for the outcome where the first repair is not made within one hour and the second and third are successful. The $2 \times 2 \times 2 = 8$ possible outcomes can be systematically arranged as follows:

FFF	FFS	FSS	SSS
	FSF	SFS	
	SFF	SSF	
$X = 0$	$X = 1$	$X = 2$	$X = 3$

where the number of successes X is the same for each outcome in a column. This value is recorded at the bottom of the column.

(b) The results of repairs on different days and different towers should be independent. Also, the probability of success 0.90 is the same for each repair. Therefore, the probability that $X = 0$ is $0.1 \times 0.1 \times 0.1 = 0.001$. Next, the probability of FFS is $0.1 \times 0.1 \times 0.9 = 0.009$ and both SFS and FSS have the same probability. Consequently, the probability that $X = 1$ is $3 \times 0.009 = 0.027$. Note that the number of outcomes where $X = 1$ is just the number of ways to select 1 or 3 trials for an S and the others are F.
By similar reasoning, the probability that $X = 2$ is

$$3(0.1 \times 0.9 \times 0.9) = 0.243 = \binom{3}{2} (0.9)^2 (0.1)^1$$

Finally, the probability that $X = 3$, no repair takes over one hour, is

$$0.9 \times 0.9 \times 0.9 = 0.729 = \binom{3}{3} (0.9)^3 (0.1)^0$$

All of these probabilities can be expressed by the formula

$$f(x) = P(X = x) = \binom{3}{x} (0.9)^x (0.1)^{3-x} \qquad \text{for } x = 0, 1, 2, 3$$

This is the probability distribution for a binomial random variable when the success probability is $p = 0.9$ and there are $n = 3$ trials. ■

Let X be the random variable that equals the number of successes in n trials. To obtain probabilities concerning X, we proceed as follows: If p and $1 - p$ are the probabilities of success and failure on any one trial, then the probability of getting x successes and $n - x$ failures, in some specific order, is $p^x (1 - p)^{n-x}$. Clearly, in this product of p's and $(1 - p)$'s there is one factor p for each success, one factor $1 - p$ for each failure. The x factors p and $n - x$ factors $1 - p$ are all multiplied together by virtue of the generalized multiplication rule for more than two independent events. Since this probability applies to any point of the sample space that represents x successes and $n - x$ failures (in any specific order), we have only to count how many points of this kind there are, and then multiply $p^x (1 - p)^{n-x}$ by this. Clearly, the number of ways in which we can select the x trials on which there is to be a success is $\binom{n}{x}$, the number of combinations of x objects selected from a set of n objects, and we thus arrive at the following result:

Binomial distribution

$$b(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

This probability distribution is called the **binomial distribution** because for $x = 0, 1, 2, \ldots,$ and n, the values of the probabilities are the successive terms of the binomial expansion of $[p + (1 - p)]^n$; for the same reason, the combinatorial quantities $\binom{n}{x}$ are referred to as **binomial coefficients**. Actually, the preceding equation defines a family of probability distributions, with each member characterized by a given value of the **parameter** p and the number of trials n.

EXAMPLE **Evaluating binomial probabilities**

It has been claimed that in 60% of all solar-heat installations the utility bill is reduced by at least one-third. Accordingly, what are the probabilities that the utility bill will be reduced by at least one-third in

(a) four of five installations;

(b) at least four of five installations?

Solution **(a)** Substituting $x = 4$, $n = 5$, and $p = 0.60$ into the formula for the binomial distribution, we get

$$b(4; 5, 0.60) = \binom{5}{4}(0.60)^4(1 - 0.60)^{5-4}$$
$$= 0.259$$

(b) Substituting $x = 5$, $n = 5$, and $p = 0.60$ into the formula for the binomial distribution, we get

$$b(5; 5, 0.60) = \binom{5}{5}(0.60)^5(1 - 0.60)^{5-5}$$
$$= 0.078$$

and the answer is $b(4; 5, 0.60) + b(5; 5, 0.60) = 0.259 + 0.078 = 0.337$. ■

If n is large, the calculation of binomial probabilities can become quite tedious. Many statistical software programs have binomial distribution commands (see Exercises 4.30 and 4.31), as do some statistical calculators. Otherwise it is convenient to refer to special tables. Table 1 at the end of the book gives the values of

$$B(x; n, p) = \sum_{k=0}^{x} b(k; n, p) \qquad \text{for } x = 0, 1, 2, \ldots, n$$

for $n = 2$ to $n = 20$ and $p = 0.05, 0.10, 0.15, \ldots, 0.90, 0.95$. We tabulated the **cumulative probabilities** rather than the values of $b(x; n, p)$, because the values of $B(x; n, p)$ are the ones needed more often in statistical applications. Note, however, that the values of $b(x; n, p)$ can be obtained by subtracting adjacent entries in Table 1. Because the two cumulative probabilities $B(x; n, p)$ and $B(x - 1; n, p)$ differ by the single term $b(x; n, p)$

$$b(x; n, p) = B(x; n, p) - B(x - 1; n, p)$$

where $B(-1) = 0$. The examples that follow illustrate the direct use of Table 1 and the use of this relationship.

EXAMPLE **Evaluating cumulative binomial probabilities**

If the probability is 0.05 that a certain wide-flange column will fail under a given axial load, what are the probabilities that among 16 such columns

(a) at most two will fail;

(b) at least four will fail?

Solution **(a)** Table 1 shows that $B(2; 16, 0.05) = 0.9571$.

(b) Since

$$\sum_{x=4}^{16} b(x; 16, 0.05) = 1 - B(3; 16, 0.05)$$

Table 1 yields $1 - 0.9930 = 0.0070$. ■

EXAMPLE **Finding a binomial probability using cumulative binomial probabilities**

The manufacture of large high-definition LCD panels is difficult, and a moderately high proportion have too many defective pixels to pass inspection. If the probability is 0.3 that an LCD panel will not pass inspection, what is the probability that 6 of 18 panels, randomly selected from production, will not pass inspection?

Solution Using the relationship to cumulative probabilities and then looking up these probabilities in Table 1, we get

$$\begin{aligned} b(6, 18, 0.30) &= B(6, 18, 0.30) - B(5, 18, 0.30) \\ &= 0.7217 - 0.5344 \\ &= 0.1873 \end{aligned}$$

■

The following example illustrates the use of the binomial distribution in a problem of decision making.

EXAMPLE **A binomial probability to guide decision making**

A manufacturer of external hard drives claims that only 10% of his drives require repairs within the warranty period of 12 months. If 5 of 20 of his drives required repairs within the first year, does this tend to support or refute the claim?

Solution Let us first find the probability that 5 or more of 20 of the drives will require repairs within a year when the probability that any one will require repairs within a year is 0.10. Using Table 1, we get

$$\begin{aligned} \sum_{x=5}^{20} b(x; 20, 0.10) &= 1 - B(4; 20, 0.10) \\ &= 1 - 0.9568 \\ &= 0.0432 \end{aligned}$$

Since this probability is very small, it would seem reasonable to reject the hard drive manufacturer's claim. ■

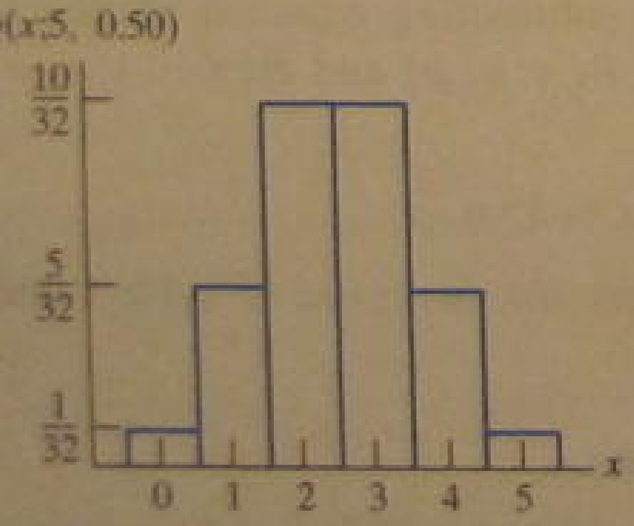


Figure 4.4
Symmetrical binomial distribution with $n = 5$ and $p = 0.50$

Important information about the shape of binomial distributions is shown in Figures 4.4 and 4.5. First, if $p = 0.50$, the equation for the binomial distribution is

$$b(x; n, 0.50) = \binom{n}{x} (0.5)^n$$

and since

$$\binom{n}{n-x} = \binom{n}{x}$$

it follows that $b(x; n, 0.50) = b(n - x; n, 0.50)$. For any n, the binomial distribution with $p = 0.5$ is a **symmetrical distribution**. This means that the probability histograms of such binomial distributions are symmetrical, as is illustrated in Figure 4.4. Note, however, that if p is less than 0.50, it is more likely that X will be

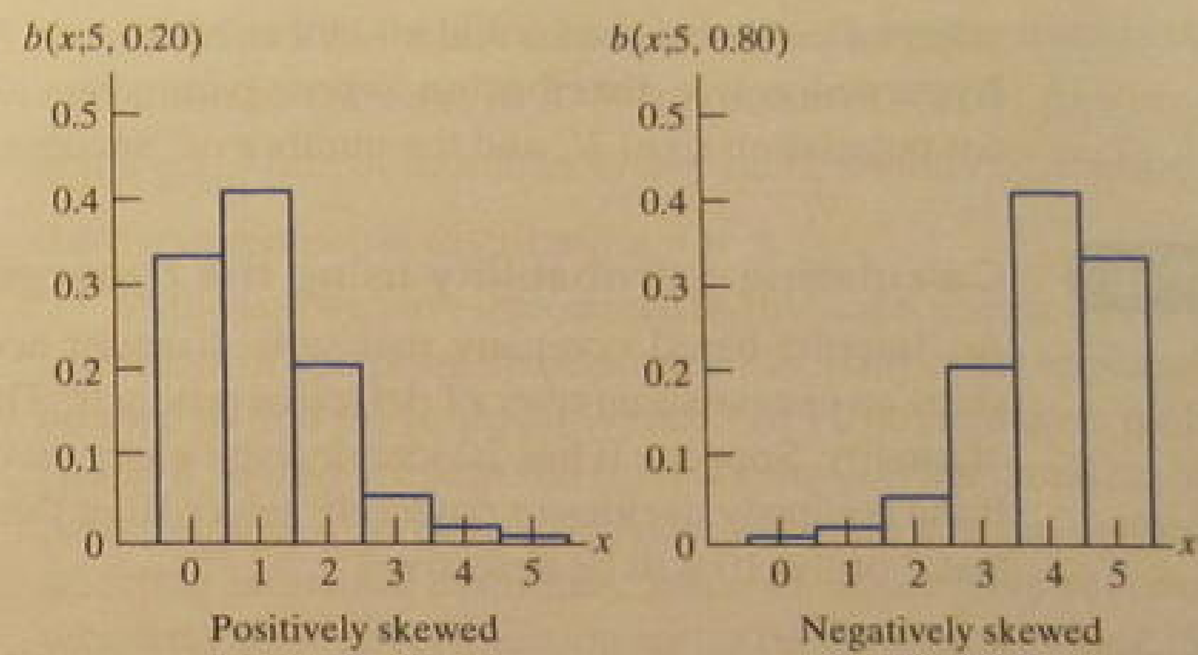


Figure 4.5
Skewed binomial distributions with $n = 5$ and $p = 0.20$ and $p = 0.80$

small rather than large compared to $n/2$ and that the opposite is true if p is greater than 0.50. This is illustrated in Figure 4.5, showing the histograms of binomial distributions with $n = 5$ and $p = 0.20$ and $p = 0.80$. These two are mirror images of each other as can be verified more generally in Exercise 4.7.

Finally, a probability distribution that has a probability histogram like either of those in Figure 4.5 is said to be a long-tailed or **skewed distribution**. It is said to be a **positively skewed distribution** if the tail is on the right, and it is said to be **negatively skewed** if the tail is on the left.

4.3 The Hypergeometric Distribution

Suppose that we are interested in the number of defectives in a sample of n units drawn without replacement from a lot containing N units, of which a are defective. Let the sample be drawn in such a way that at each successive drawing, whatever units are left in the lot have the same chance of being selected. The probability that the first drawing will yield a defective unit is $\frac{a}{N}$, but for the second drawing it is $\frac{a-1}{N-1}$ or $\frac{a}{N-1}$, depending on whether or not the first unit drawn was defective. Thus, the trials are not independent, the fourth assumption underlying the binomial distribution is not met, and the binomial distribution does not apply. Note that the binomial distribution would apply if we do **sampling with replacement**, namely, if each unit selected for the sample is replaced before the next one is drawn.

To solve the problem of **sampling without replacement** (that is, as we originally formulated the problem), let us proceed as follows: The x successes (defectives) can be chosen in $\binom{a}{x}$ ways, the $n - x$ failures (nondefectives) can be chosen in $\binom{N-a}{n-x}$ ways, and hence, x successes and $n - x$ failures can be chosen in $\binom{a}{x}\binom{N-a}{n-x}$ ways. Also, n objects can be chosen from a set of N objects in $\binom{N}{n}$ ways, and if we consider all the possibilities as equally likely, it follows that for sampling without replacement the probability of getting "x successes in n trials" is

Hypergeometric distribution

$$h(x; n, a, N) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}} \qquad \text{for } x = 0, 1, \ldots, n$$

where x cannot exceed a and $n - x$ cannot exceed $N - a$. This equation defines the **hypergeometric distribution**, whose parameters are the sample size n, the lot size (or population size) N, and the number of "successes" in the lot a.

EXAMPLE **Calculating a probability using the hypergeometric distribution**

An Internet-based company that sells discount accessories for cell phones often ships an excessive number of defective products. The company needs better control of quality. Suppose it has 20 identical car chargers on hand but that 5 are defective. If the company decides to randomly select 10 of these items, what is the probability that 2 of the 10 will be defective?

Solution Substituting $x = 2$, $n = 10$, $a = 5$, and $N = 20$ into the formula for the hypergeometric distribution, we get

$$h(2; 10, 5, 20) = \frac{\binom{5}{2}\binom{15}{8}}{\binom{20}{10}} = \frac{10 \times 6{,}435}{184{,}756} = 0.348$$

■

In the preceding example, n was not small compared to N, and if we had made the mistake of using the binomial distribution with $n = 10$ and $p = \frac{5}{20} = 0.25$ to calculate the probability of two defectives, the result would have been 0.282, which is much too small. However, when n is small compared to N, less than $\frac{N}{10}$, the composition of the lot is not seriously affected by drawing the sample without replacement, and the binomial distribution with the parameters n and $p = \frac{a}{N}$ will yield a good approximation.

EXAMPLE **A numerical comparison of the hypergeometric and binomial distributions**

Repeat the preceding example but with 100 car chargers, of which 25 are defective, by using

(a) the formula for the hypergeometric distribution;
(b) the formula for the binomial distribution as an approximation.

Solution **(a)** Substituting $x = 2$, $n = 10$, $a = 25$, and $N = 100$ into the formula for the hypergeometric distribution, we get

$$h(2; 10, 25, 100) = \frac{\binom{25}{2}\binom{75}{8}}{\binom{100}{10}} = 0.292$$

(b) Substituting $x = 2$, $n = 10$, and $p = \frac{25}{100} = 0.25$ into the formula for the binomial distribution, we get

$$b(2; 10, 0.25) = \binom{10}{2}(0.25)^2(1 - 0.25)^{10-2}$$
$$= 0.282$$

■

Observe that the difference between the two values is only 0.010. In general, it can be shown that $h(x;\, n, a, N)$ approaches $b(x;\, n, p)$ with $p = \frac{a}{N}$ when $N \to \infty$, and a good rule of thumb is to use the binomial distribution as an approximation to the hypergeometric distribution if $n \leq \frac{N}{10}$.

Although we have introduced the hypergeometric distribution in connection with a problem of sampling inspection, it has many other applications. For instance, it can be used to find the probability that 3 of 12 homemakers prefer Brand A detergent to Brand B, if they are selected from among 200 homemakers among whom 40 actually prefer Brand A to Brand B. Also, it can be used in connection with a problem of selecting industrial diamonds, some of which have superior qualities and some of which do not, or in connection with a problem of sampling income tax returns, where a among N returns filed contain questionable deductions, and so on.

Exercises

4.1 Suppose that a probability of $\frac{1}{12}$ is assigned to each point of the sample space of part (a) of Exercise 3.1 on page 53. Find the probability distribution of the total number of lakes and streams that are contaminated.

4.2 An experiment consists of four tosses of a coin. Denoting the outcomes *HHTH*, *THTT*, . . . and assuming that all 16 outcomes are equally likely, find the probability distribution for the total number of heads.

4.3 Determine whether the following can be probability distributions of a random variable which can take on only the values 1, 2, 3, and 4.

(a) $f(1) = 0.16$, $f(2) = 0.28$, $f(3) = 0.28$, and $f(4) = 0.28$;

(b) $f(1) = 0.24$, $f(2) = 0.24$, $f(3) = 0.24$, and $f(4) = 0.24$;

(c) $f(1) = 0.37$, $f(2) = 0.31$, $f(3) = 0.34$, and $f(4) = -0.02$.

4.4 Check whether the following can define probability distributions, and explain your answers.

(a) $f(x) = \frac{x}{14}$ for $x = 0, 1, 2, 3, 4$

(b) $f(x) = \frac{3 - x^2}{4}$ for $x = 0, 1, 2$

(c) $f(x) = \frac{1}{5}$ for $x = 5, 6, 7, 8, 9$

(d) $f(x) = \frac{2x + 1}{50}$ for $x = 1, 2, 3, 4, 5$

4.5 Given that $f(x) = \frac{k}{2^x}$ is a probability distribution for a random variable that can take on the values $x = 0, 1, 2, 3$, and 4, find k.

4.6 With reference to Exercise 4.5, find an expression for the distribution function $F(x)$ of the random variable.

4.7 Prove that $b(x;\, n, p) = b(n - x;\, n, 1 - p)$.

4.8 Prove that $B(x;\, n, p) = 1 - B(n - x - 1;\, n, 1 - p)$.

4.9 Do the assumptions for Bernoulli trials appear to hold? Explain. If the assumptions hold, identify success and the probability of interest.

(a) A TV ratings company will use their electronic equipment to check a sample of homes around the city to see which ones have a set tuned to the mayor's speech on the local channel.

(b) Among 6 nuclear power plants in a state, 2 have had serious violations in last five years. Two plants will be selected at random, one after the other, and the outcome of interest is a serious violation in the last five years.

4.10 What conditions for the binomial distribution, if any, fail to hold in the following situations?

(a) For each of a company's eight production facilities, record whether or not there was an accident in the past week. The largest facility has three times the number of production workers as the smallest facility.

(b) For each of three shifts, the number of units produced will be compared with the long-term average of 560 and it will be determined whether or not production exceeds 560 units. The second shift will know the result for the first shift before they start working, and the third shift will start with the knowledge of how the first two shifts performed.

4.11 Which conditions for the binomial distribution, if any, fail to hold in the following situations?

(a) The number of persons having a cold at a family reunion attended by 30 persons.

(b) Among 8 projectors in the department office, 2 do not work properly but are not marked defective. Two are selected and the number that do not work properly will be recorded.

4.12 Use Table 1 to find

(a) $B(8; 16, 0.40)$;

(b) $b(8; 16, 0.40)$;

(c) $B(9; 12, 0.60)$;

(d) $b(9; 12, 0.60)$;

(e) $\sum_{k=6}^{20} b(k; 20, 0.15)$;

(f) $\sum_{k=6}^{9} b(k; 9, 0.70)$.

4.13 Use Table 1 to find

(a) $B(7; 19, 0.45)$;

(b) $b(7; 19, 0.45)$;

(c) $B(8; 10, 0.95)$;

(d) $b(8; 10, 0.95)$;

(e) $\sum_{k=4}^{10} b(k; 10, 0.35)$;

(f) $\sum_{k=2}^{4} b(k; 9, 0.30)$.

4.14 Rework the decision problem on page 88, supposing that only 3 of the 20 hard drives required repairs within the first year.

4.15 Human error is given as the reason for 75% of all accidents in a plant. Use the formula for the binomial distribution to find the probability that human error will be given as the reason for two of the next four accidents.

4.16 If the probability is 0.40 that steam will condense in a thin-walled aluminum tube at 10 atm pressure, use the formula for the binomial distribution to find the probability that, under the stated conditions, steam will condense in 4 of 12 such tubes.

4.17 During one stage in the manufacture of integrated circuit chips, a coating must be applied. If 70% of chips receive a thick enough coating, use Table 1 to find the probabilities that, among 15 chips:

(a) at least 12 will have thick enough coatings;

(b) at most 6 will have thick enough coatings;

(c) exactly 10 will have thick enough coatings.

4.18 The probability that the noise level of a wide-band amplifier will exceed 2 dB is 0.05. Use Table 1 to find the probabilities that among 12 such amplifiers the noise level of

(a) one will exceed 2 dB;

(b) at most two will exceed 2 dB;

(c) two or more will exceed 2 dB.

4.19 An agricultural cooperative claims that 90% of the watermelons shipped out are ripe and ready to eat. Find the probabilities that among 18 watermelons shipped out

(a) all 18 are ripe and ready to eat;

(b) at least 16 are ripe and ready to eat;

(c) at most 14 are ripe and ready to eat.

4.20 A quality-control engineer wants to check whether (in accordance with specifications) 95% of the electronic components shipped by his company are in good working condition. To this end, he randomly selects 15 from each large lot ready to be shipped and passes the lot if the selected components are all in good working condition; otherwise, each of the components in the lot is checked. Find the probabilities that the quality-control engineer will commit the error of

(a) holding a lot for further inspection even though 95% of the components are in good working condition;

(b) letting a lot pass through without further inspection even though only 90% of the components are in good working condition;

(c) letting a lot pass through without further inspection even though only 80% of the components are in good condition.

4.21 A food processor claims that at most 10% of her jars of instant coffee contain less coffee than claimed on the label. To test this claim, 16 jars of her instant coffee are randomly selected and the contents are weighed; her claim is accepted if fewer than 3 of the jars contain less coffee than claimed on the label. Find the probabilities that the food processor's claim will be accepted when the actual percentage of her jars containing less coffee than claimed on the label is

(a) 5%; (b) 10%; (c) 15%; (d) 20%.

4.22 Refer to Exercise 4.2.

(a) Determine the cumulative probability distribution $F(x)$.

(b) Graph the probability distribution of $f(x)$ as a bar chart and below it graph $F(x)$.

4.23 Four emergency radios are available for rescue workers but one does not work properly. Two randomly selected radios are taken on a rescue mission. Let X be the number that work properly between the two.

(a) Determine the probability distribution $f(x)$ of X.

(b) Determine the cumulative probability distribution $F(x)$ of X.

(c) Graph $f(x)$ as a bar chart and below it graph $F(x)$.

4.24 Among the 12 solar collectors on display at a trade show, 9 are flat-plate collectors and the others are concentrating collectors. If a person visiting the show randomly selects 4 of the solar collectors to check out, what is the probability that 3 of them will be flat-plate collectors?

4.25 A quality-control engineer inspects a random sample of 3 batteries from each lot of 24 car batteries ready to be shipped. If such a lot contains 6 batteries with slight defects, what are the probabilities that the inspector's sample will contain

(a) none of the batteries with defects?

(b) only one of the batteries with defects?

(c) at least two of the batteries with defects?

4.26 If 6 of 18 new buildings in a city violate the building code, what is the probability that a building inspector, who randomly selects 4 of the new buildings for inspection, will catch

(a) none of the buildings that violate the building code?

(b) 1 of the new buildings that violate the building code?

(c) 2 of the new buildings that violate the building code?

(d) at least 3 of the new buildings that violate the building code?

4.27 Among the 16 cities that a professional society is considering for its next 3 annual conventions, 7 are in the western part of the United States. To avoid arguments, the selection is left to chance. If none of the cities can be chosen more than once, what are the probabilities that

(a) none of the conventions will be held in the western part of the United States?

(b) all of the conventions will be held in the western part of the United States?

4.28 A shipment of 120 burglar alarms contains 5 that are defective. If 3 of these alarms are randomly selected and shipped to a customer, find the probability that the customer will get one bad unit by using

(a) the formula for the hypergeometric distribution;

(b) the formula for the binomial distribution as an approximation.

4.29 Among the 300 employees of a company, 240 are union members, while the others are not. If 8 of the employees are chosen by lot to serve on the committee which administers the pension fund, find the probability that 5 of them will be union members while the others are not, by using

(a) the formula for hypergeometric distribution;

(b) the formula for binomial distribution as an approximation.

4.30 Binomial probabilities can be calculated using *MINITAB*.

> **Dialog box:**
> **Calc > Probability Distribution > Binomial**
> Choose **Probability.**
> Enter 7 in **Number of trials** and .33 in **Probability of success**.
> Choose **Input constant** and enter 2.
> Click **OK**.

Output:

Binomial with $n = 7$ and $p = 0.330000$

x	$P(X = x)$
2.00	0.3088

Find the binomial probabilities with $n = 27$ and $p = 0.47$.

4.31 Cumulative binomial probabilities can be calculated using *MINITAB*.

> **Dialog Box:**
> **Calc > Probability Distribution > Binomial**
> Choose **Cumulative Distribution**.
> Enter 7 in **Number of trials** and .33 in **Probability of success**.
> Choose **Input constant** and enter 2.
> Click **OK**.

Output:

Binomial with $n = 7$ and $p = 0.330000$

x	$P(X <= x)$
2.00	0.5783

Find the cumulative binomial probabilities with $n = 27$ and $p = 0.47$.

4.4 The Mean and the Variance of a Probability Distribution

Besides the binomial and hypergeometric distributions, there are many other probability distributions that have important engineering applications. However, before we go any further, let us discuss some general characteristics of probability distributions.

One such characteristic, that of the symmetry or **skewness** of a probability distribution, was illustrated in Figures 4.4 and 4.5; two other characteristics are

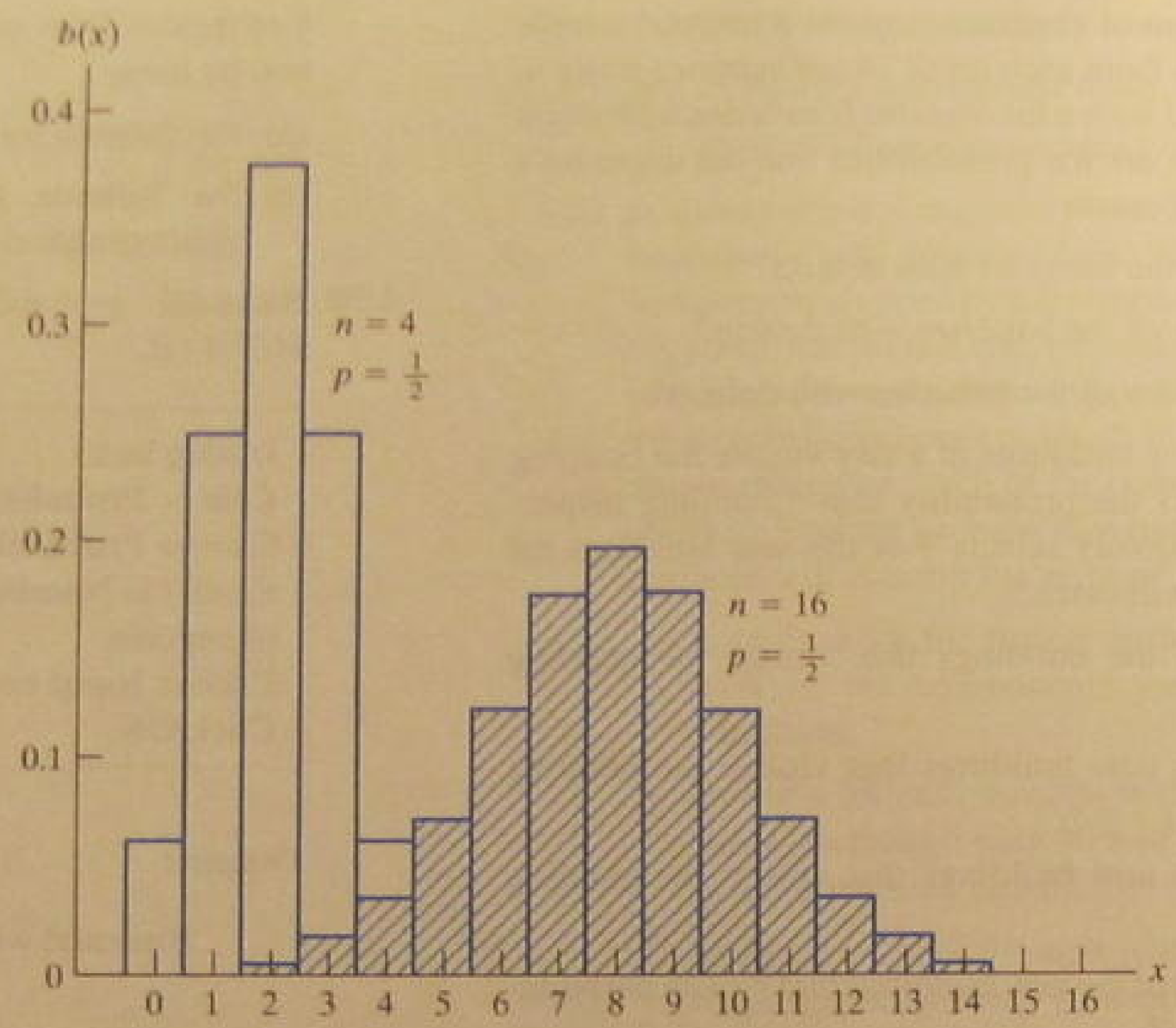


Figure 4.6
Probability histograms of two binomial distributions

apparent in Figure 4.6, which shows the probability histograms of two binomial distributions. One of these binomial distributions has the parameters $n = 4$ and $p = 1/2$, and the other has the parameters $n = 16$ and $p = 1/2$. Essentially, these two probability distributions differ in two respects. The first probability distribution is centered about $x = 2$, whereas the other (whose histogram is shaded) is centered about $x = 8$, and we say that the two distributions differ in their **location**. Another distinction is that the histogram of the second distribution is broader, and we say that the two distributions differ in **variation**. To make such comparisons more specific, we shall introduce in this section two of the most important statistical measures, describing the location and the variation of a probability distribution—the **mean** and the **variance**, respectively.

The mean of a probability distribution is simply the mathematical expectation of a random variable having that distribution. If a random variable X takes on the values $x_1, x_2, \ldots,$ or x_k, with the probabilities $f(x_1), f(x_2), \ldots,$ and $f(x_k)$, its mathematical expectation or expected value is

$$x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + \cdots + x_k \cdot f(x_k) = \sum \text{(value)} \times \text{(probability)}$$

using the $\sum$ notation. We can thus write

Mean of discrete probability distribution

$$\mu = \sum_{\text{all } x} x \cdot f(x)$$

where the mean is denoted by the Greek letter μ (*mu*). The mean of a probability distribution measures its center in the sense of an average, or by analogy to physics, in the sense of a center of gravity. Note that the above formula for μ is, in fact, that for the **first moment about the origin** of a discrete system of masses $f(x)$ arranged on a weightless straight line at distances x from the origin. We do not have to divide

here by

$$\sum_{\text{all } x} f(x)$$

as we do in the usual formula for the x-coordinate of the center of gravity, since the sum equals 1 by definition.

EXAMPLE **The mean number of heads in three tosses**

Find the mean of the probability distribution of the number of heads obtained in 3 flips of a balanced coin.

Solution The probabilities for 0, 1, 2, or 3 heads are $\frac{1}{8}$, $\frac{3}{8}$, $\frac{3}{8}$, and $\frac{1}{8}$ as can easily be verified by counting equally likely possibilities or by using the formula for the binomial distribution with $n = 3$ and $p = \frac{1}{2}$. Thus,

$$\mu = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}$$

■

EXAMPLE **The mean number of preferred used car attributes**

With reference to the used car example and the probabilities given on page 82, find the mean of the probability distribution of the number of preferred attributes.

Solution Substituting $x = 0, 1, 2$, and 3 and the corresponding probabilities into the formula for μ, we get

$$\begin{aligned}\mu &= 0(0.18) + 1(0.50) + 2(0.29) + 3(0.03)\\ &= 1.17\end{aligned}$$

■

Returning to the second probability distribution of Figure 4.6, we could find its mean by calculating all the necessary probabilities (or by looking them up in Table 1) and substituting them into the formula for μ. However, if we reflect for a moment, we might argue that there is a 50-50 chance for a success on each trial, there are 16 trials, and it would seem reasonable to expect 8 heads and 8 tails (in the sense of a mathematical expectation). Similarly, we might argue that if a binomial distribution has the parameters $n = 200$ and $p = 0.20$, we can expect a success 20% of the time and, hence, on the average $200(0.20) = 40$ successes in 200 trials. These two values are, indeed, correct, and it can be shown in general that

Mean of binomial distribution

$$\boxed{\mu = n \cdot p}$$

for the mean of a binomial distribution. To prove this formula, we substitute the expression that defines $b(x; n, p)$ into the formula for μ, and we get

$$\mu = \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Then, making use of the fact that

$$\frac{x}{x!} = \frac{1}{(x-1)!}$$

and $n! = n(n-1)!$, and factoring out n and p, we obtain

$$\mu = np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1}(1-p)^{n-x}$$

where the summation starts with $x = 1$ since the original summand is zero for $x = 0$. If we now let $y = x - 1$ and $m = n - 1$, we obtain

$$\mu = np \sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^{y}(1-p)^{m-y}$$

and this last sum can easily be recognized as that of all the terms of the binomial distribution with the parameters m and p. Hence, this sum equals 1 and it follows that $\mu = np$.

EXAMPLE **Using $\mu = np$ to find the mean number of heads in three tosses**

Find the mean of the probability distribution of the number of heads obtained in 3 flips of a balanced coin.

Solution For a binomial distribution with $n = 3$ and $p = \frac{1}{2}$, we get $\mu = 3 \cdot \frac{1}{2} = \frac{3}{2}$, and this agrees with the result obtained on page 95. ■

The formula $\mu = np$ applies, of course, only to binomial distributions, but for other special distributions, we can, similarly, express the mean in terms of their parameters. For instance, for the mean of the hypergeometric distribution with the parameters n, a, and N, we can write

Mean of hypergeometric distribution

$$\mu = n \cdot \frac{a}{N}$$

In Exercise 4.43, the reader will be asked to derive this formula by a method similar to the one we used to derive the formula for the mean of a binomial distribution.

EXAMPLE **Using the formula for the mean of a hypergeometric distribution**

With reference to the example on page 90 in which 5 of 20 car chargers were defective, find the mean of the probability distribution of the number of defectives in a sample of 10 randomly chosen for inspection.

Solution Substituting $n = 10$, $a = 5$, and $N = 20$ into the above formula for μ, we get

$$\mu = 10 \cdot \frac{5}{20} = 2.5$$

In other words, if we inspect 10 of the car chargers, we can expect 2.5 defectives, where *expect* is to be interpreted in the sense, it represents the long-run average number of defectives if 10 chargers are repeatedly selected from 20 chargers of which 5 are defective. ■

To study the second of the two properties of probability distributions mentioned on page 94, their variation, let us refer again to the two probability distributions of Figure 4.6. For the one where $n = 4$, there is a high probability of getting values close to the mean, but for the one where $n = 16$, there is a high probability of getting values scattered over considerable distances away from the mean. Using this property, it

may seem reasonable to measure the variation of a probability distribution with the quantity

$$\sum_{\text{all } x} (x - \mu) \cdot f(x)$$

namely, the average amount by which the values of the random variable deviate from the mean. Unfortunately,

$$\begin{aligned}\sum_{\text{all } x} (x - \mu) \cdot f(x) &= \sum_{\text{all } x} x \cdot f(x) - \sum_{\text{all } x} \mu \cdot f(x) \\ &= \mu - \mu \cdot \sum_{\text{all } x} f(x) = \mu - \mu = 0\end{aligned}$$

so that this expression is always equal to zero. However, since we are really interested in the magnitude of the deviations $x - \mu$ and not in their signs, it suggests itself that we average the absolute values of these deviations from the mean. This would, indeed, provide a measure of variation, but on purely theoretical grounds we prefer to work instead with the squares of the deviations from the mean. These quantities are also nonnegative, and their average is indicative of the spread or dispersion of a probability distribution. We thus define the **variance** of a probability distribution $f(x)$, or that of the random variable X which has that probability distribution, as

Variance of probability distribution

$$\sigma^2 = \sum_{\text{all } x} (x - \mu)^2 \cdot f(x)$$

where σ is the lowercase Greek letter for s. This measure is not in the same units (or dimension) as the values of the random variable, but we can adjust for this by taking the square root. This results in a measure of variation that is expressed in the same units in which the random variable is expressed. The **standard deviation** is defined as

Standard deviation of probability distribution

$$\sigma = \sqrt{\sum_{\text{all } x} (x - \mu)^2 \cdot f(x)}$$

EXAMPLE **Calculating the standard deviations of two probability distributions**

Compare the standard deviations of the two probability distributions of Figure 4.6, on page 94.

Solution Since $\mu = 4 \cdot \frac{1}{2} = 2$ for the binomial distribution with $n = 4$ and $p = \frac{1}{2}$, we find that the variance of this probability distribution is

$$\begin{aligned}\sigma^2 = (0-2)^2 \cdot \frac{1}{16} + (1-2)^2 \cdot \frac{4}{16} + (2-2)^2 \cdot \frac{6}{16} \\ + (3-2)^2 \cdot \frac{4}{16} + (4-2)^2 \cdot \frac{1}{16} = 1\end{aligned}$$

and, hence, that its standard deviation is $\sigma = 1$. Similarly, it can be shown that for the other distribution $\sigma = 2$, and we find that the second (shaded) distribution with the greater spread also has the greater standard deviation. ■

Given any probability distribution, we can always calculate σ^2 by substituting the corresponding probabilities $f(x)$ into the formula which defines the variance. As in the case of the mean, however, this work can be simplified to a considerable extent when we deal with special kinds of distributions. For instance, it can be shown that the variance of the binomial distribution with the parameters n and p is given by the formula

Variance of binomial distribution

$$\sigma^2 = n \cdot p \cdot (1 - p)$$

EXAMPLE **Using the formula for variance of the binomial distribution**

Verify the result stated in the preceding example, that $\sigma = 2$ for the binomial distribution with $n = 16$ and $p = \frac{1}{2}$.

Solution Substituting $n = 16$ and $p = \frac{1}{2}$ into the formula for the variance of a binomial distribution, we get

$$\sigma^2 = 16 \cdot \frac{1}{2} \cdot \frac{1}{2} = 4$$

and, hence, $\sigma = \sqrt{4} = 2$. ■

The variance of the hypergeometric distribution with the parameters n, a, and N is given by the formula

Variance of hypergeometric distribution

$$\sigma^2 = n\,\frac{a}{N}\left(1 - \frac{a}{N}\right)\left(\frac{N-n}{N-1}\right)$$

EXAMPLE **Using the formula for variance of the hypergeometric distribution**

With reference to the example on page 90 in which 5 of 20 car chargers were defective, find the standard deviation of the probability distribution of the number of defectives in a sample of 10 randomly chosen for inspection.

Solution Substituting $n = 10$, $a = 5$, and $N = 20$ into the formula for the variance of a hypergeometric distribution, we get

$$\sigma^2 = 10\,\frac{5}{20}\left(1 - \frac{5}{20}\right)\left(\frac{20-10}{20-1}\right) = \frac{75}{76}$$

and, hence, $\sigma = \sqrt{75/76} = 0.99$. ■

When we first defined the variance of a probability distribution, it may have occurred to the reader that the formula looked exactly like the one which we use in physics to define second moments, or moments of inertia. Indeed, it is customary in statistics to define the ***k*th moment about the origin** as

$$\mu'_k = \sum_{\text{all } x} x^k \cdot f(x)$$

and the ***k*th moment about the mean** as

$$\mu_k = \sum_{\text{all } x} (x - \mu)^k \cdot f(x)$$

Thus, the mean μ is the first moment about the origin, and the variance σ^2 is the second moment about the mean. Higher moments are often used in statistics to give further descriptions of probability distributions. For instance, the third moment about the mean (divided by σ^3 to make this measure independent of the scale of measurement) is used to describe the symmetry or skewness of a distribution; the fourth moment about the mean (divided by σ^4) is, similarly, used to describe its "peakedness," or **kurtosis**. To determine moments about the mean, it is usually easiest to express moments about the mean in terms of moments about the origin and then to calculate the necessary moments about the mean. For the second moment about the mean we thus have the important formula

Computing formula for the variance

$$\sigma^2 = \mu_2' - \mu^2$$

which the reader will be asked to prove in part (a) of Exercise 4.49; part (b) of that exercise pertains to a similar formula for expressing μ_3 in terms of moments about the origin.

EXAMPLE **Calculating variance using the alternative computing formula**

Use the preceding computing formula to determine the variance of the probability distribution of the number of points rolled with a balanced die.

Solution Since $f(x) = \frac{1}{6}$ for $x = 1, 2, 3, 4, 5$, and 6, we get

$$\begin{aligned}\mu &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{7}{2} \\ \mu_2' &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} \\ &= \frac{91}{6}\end{aligned}$$

and, hence,

$$\sigma^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

■

EXAMPLE **The mean and variance of the number of incorrect addresses**

As part of a quality-improvement project focused on the delivery of mail at a department office within a large company, data were gathered on the number of different addresses that had to be changed so the mail could be redirected to the correct mail stop. The distribution, given in the first two columns of the table below, describes the number of redirects per delivery. Compute the mean and variance.

Solution We determine the columns $xf(x)$ and $x^2 f(x)$

x	$f(x)$	$xf(x)$	$x^2 f(x)$
0	.05	.0	0.0
1	.20	.2	0.2
2	.45	.9	1.8
3	.20	.6	1.8
4	.10	.4	1.6
Total		2.1	5.4

so $\mu = 2.1$ and $\sigma^2 = 5.4 - (2.1)^2 = 0.990$. ■

4.5 Chebyshev's Theorem

Earlier in this chapter we used examples to show how the standard deviation measures the variation of a probability distribution, that is, how it reflects the concentration of probability in the neighborhood of the mean. If σ is large, there is a correspondingly higher probability of getting values farther away from the mean. Formally, the idea is expressed by the following theorem.

Chebyshev's theorem

Theorem 4.1 If a probability distribution has mean μ and standard deviation σ, the probability of getting a value which deviates from μ by at least $k\sigma$ is at most $\frac{1}{k^2}$.

Symbolically,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

where $P(|X - \mu| \geq k\sigma)$ is the probability associated with the set of outcomes for which x, the value of a random variable having the given probability distribution, is such that $|x - \mu| \geq k\sigma$.

Thus, the probability that a random variable will take on a value which deviates (differs) from the mean by at least 2 standard deviations is at most $\frac{1}{4}$, the probability that it will take on a value which deviates from the mean by at least 5 standard deviations is at most $\frac{1}{25}$, and the probability that it will take on a value which deviates from the mean by 10 standard deviations or more is less than or equal to $\frac{1}{100}$.

To prove this theorem, consider any probability distribution $f(x)$, having mean μ, and variance σ^2. Dividing the sum defining the variance into three parts as indicated in Figure 4.7, we have

$$\begin{aligned}\sigma^2 &= \sum_{\text{all } x} (x - \mu)^2 f(x) \\ &= \sum_{R_1} (x - \mu)^2 f(x) + \sum_{R_2} (x - \mu)^2 f(x) + \sum_{R_3} (x - \mu)^2 f(x)\end{aligned}$$

where R_1 is the region for which $x \leq \mu - k\sigma$, R_2 is the region for which $\mu - k\sigma < x < \mu + k\sigma$, and R_3 is the region for which $x \geq \mu + k\sigma$. Since $(x - \mu)^2 f(x)$

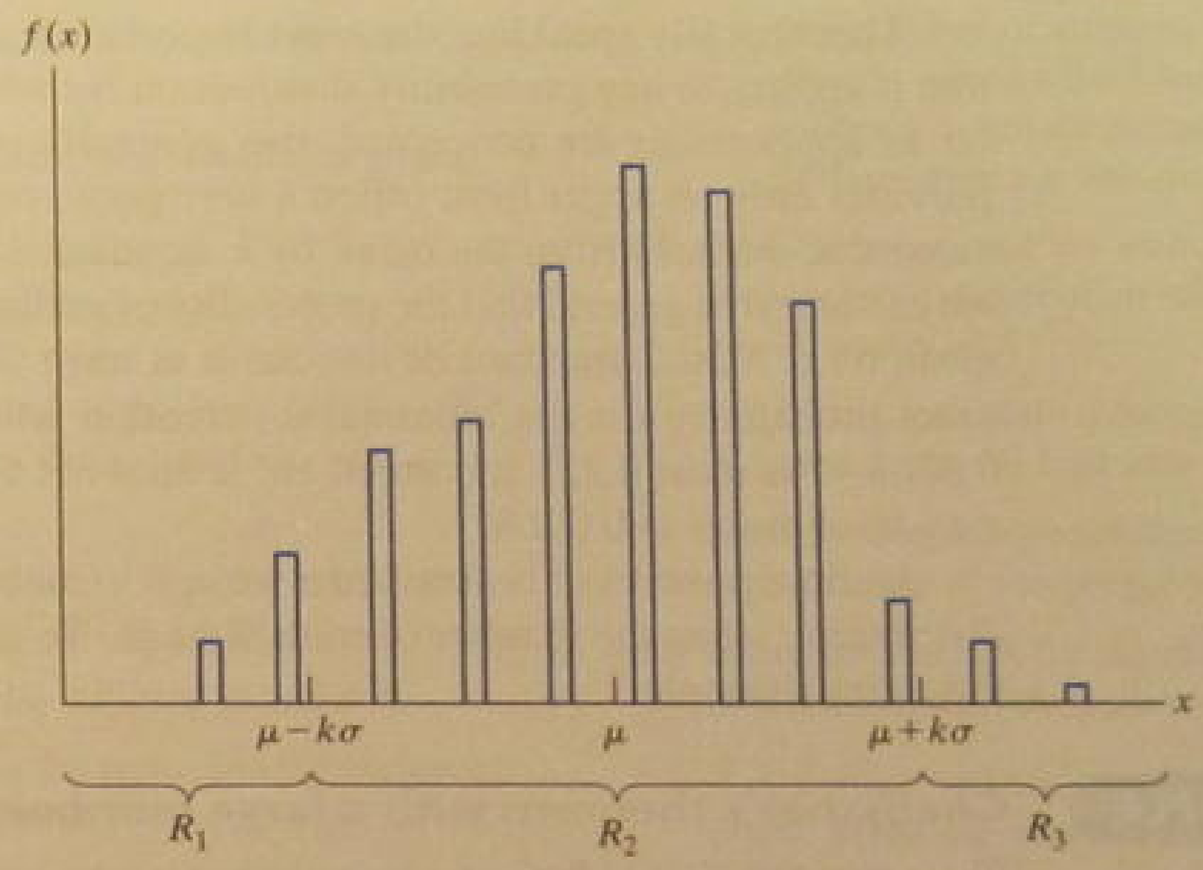


Figure 4.7
Diagram for proof of Chebyshev's theorem

cannot be negative, the above sum over R_2 is nonnegative, and without it the sum of the summations over R_1 and R_3 is less than or equal to σ^2; that is,

$$\sigma^2 \geq \sum_{R_1} (x-\mu)^2 f(x) + \sum_{R_3} (x-\mu)^2 f(x)$$

But $x - \mu \leq -k\sigma$ in the region R_1 and $x - \mu \geq k\sigma$ in the region R_3, so that in either case $|x - \mu| \geq k\sigma$; hence, in both regions $(x-\mu)^2 \geq k^2\sigma^2$. If we now replace $(x-\mu)^2$ in each sum by $k^2\sigma^2$, a number less than or equal to $(x-\mu)^2$, we obtain the inequality

$$\sigma^2 \geq \sum_{R_1} k^2\sigma^2 f(x) + \sum_{R_1} k^2\sigma^2 f(x)$$

or

$$\frac{1}{k^2} \geq \sum_{R_1} f(x) + \sum_{R_3} f(x)$$

Since

$$\sum_{R_1} f(x) + \sum_{R_3} f(x)$$

represents the probability assigned to the region $R_1 \cup R_3$, namely, $P(|X-\mu| \geq k\sigma)$, this completes the proof of Theorem 4.1.

To obtain an alternative form of Chebyshev's theorem, note that the event $|x - \mu| < k\sigma$ is the complement of the event $|x - \mu| \geq k\sigma$; hence, the probability of getting a value which deviates from μ by less than $k\sigma$ is at least $1 - \frac{1}{k^2}$.

EXAMPLE **A probability bound using Chebyshev's theorem**

The number of customers who visit a car dealer's showroom on a Saturday morning is a random variable with $\mu = 18$ and $\sigma = 2.5$. With what probability can we assert that there will be more than 8 but fewer than 28 customers?

Solution Let X be the number of customers. Since

$$k = \frac{28-18}{2.5} = \frac{18-8}{2.5} = 4$$

$$P(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2} \quad \text{and} \quad P(8 < X < 28) \geq 1 - \frac{1}{4^2} = \frac{15}{16}$$ ■

Theoretically speaking, the most important feature of Chebyshev's theorem is that it applies to any probability distribution for which μ and σ exist. However, so far as applications are concerned, this generally is also its greatest weakness—it provides only an upper limit (often a very poor one) to the probability of getting a value that deviates from the mean by k standard deviations or more. For instance, we can assert in general that the probability of getting a value which differs from the mean by at least 2 standard deviations is at most 0.25, whereas the corresponding exact probability for the binomial distribution with $n = 16$ and $p = \frac{1}{2}$ is only 0.0768—"at most 0.25" is correct, but it does not tell us that the actual probability may be as small as 0.0768.

An important result is obtained if we apply Chebyshev's theorem to the binomial distribution when the number of trials is large. To illustrate this result, consider the following example.

EXAMPLE **Chebyshev's theorem with a large number of Bernoulli trials**

Show that for 40,000 flips of a balanced coin, the probability is at least 0.99 that the proportion of heads will fall between 0.475 and 0.525.

Solution Since

$$\mu = 40{,}000 \cdot \frac{1}{2} = 20{,}000 \qquad \sigma = \sqrt{40{,}000 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 100$$

and

$$1 - \frac{1}{k^2} = 0.99$$

yields $k = 10$, the alternative form of Chebyshev's theorem tells us that the probability is at least 0.99 that we will get between $20{,}000 - 10(100) = 19{,}000$ and $20{,}000 + 10(100) = 21{,}000$ heads. Hence, the probability is at least 0.99 that the proportion of heads will fall between

$$\frac{19{,}000}{40{,}000} = 0.475 \qquad \text{and} \qquad \frac{21{,}000}{40{,}000} = 0.525$$

■

Correspondingly, the reader will be asked to show in Exercise 4.47 that for 1,000,000 flips of a balanced coin the probability is at least 0.99 that the proportion of heads will fall between 0.495 and 0.505, and these results suggest that when n is large, the chances are that the proportion of heads will be very close to $p = \frac{1}{2}$. When formulated for any binomial distribution with the parameters n and p, this result is referred to as the **law of large numbers**.

Exercises

4.32 Suppose that the probabilities are 0.4, 0.3, 0.2, and 0.1 that there will be 0, 1, 2, or 3 power failures in a certain city during the month of July. Use the formulas which define μ and σ^2 to find

(a) the mean of this probability distribution;

(b) the variance of this probability distribution.

4.33 Use the computing formula for σ^2 to rework part (b) of the preceding exercise.

4.34 The following table gives the probabilities that a certain computer will malfunction 0, 1, 2, 3, 4, 5, or 6 times on any one day:

Number of malfunctions: x	0	1	2	3	4	5	6
Probability: $f(x)$	0.17	0.29	0.27	0.16	0.07	0.03	0.01

Use the formulas which define μ and σ to find

(a) the mean of this probability distribution;

(b) the standard deviation of this probability distribution.

4.35 Use the computing formula for σ^2 to rework part (b) of the preceding exercise.

4.36 Find the mean and the variance of the uniform probability distribution given by

$$f(x) = \frac{1}{n} \qquad \text{for } x = 1, 2, 3, \ldots, n$$

[*Hint*: The sum of the first n positive integers is $n(n+1)/2$, and the sum of their squares is $n(n+1)(2n+1)/6$.]

4.37 Find the mean and the variance of the binomial distribution with $n = 4$ and $p = 0.70$ by using

(a) Table 1 and the formulas defining μ and σ^2;

(b) the special formulas for the mean and the variance of a binomial distribution.

4.38 As can easily be verified by means of the formula for the binomial distribution (or by listing all 32 possibilities), the probabilities of getting 0, 1, 2, 3, 4, or 5 heads in five flips of a balanced coin are

$$\frac{1}{32} \quad \frac{5}{32} \quad \frac{10}{32} \quad \frac{10}{32} \quad \frac{5}{32} \quad \frac{1}{32}$$

Find the mean of this probability distribution using

(a) the formula that defines μ;

(b) the special formula for the mean of a binomial distribution.

4.39 With reference to Exercise 4.38, find the variance of the probability distribution using

(a) the formula that defines σ^2;

(b) the computing formula for σ^2;

(c) the special formula for the variance of a binomial distribution.

4.40 If 95% of certain high-performance radial tires last at least 30,000 miles, find the mean and the standard deviation of the distribution of the number of these tires, among 20 selected at random, that last at least 30,000 miles, using

(a) Table 1, the formula which defines μ, and the computing formula for σ^2.

(b) the special formulas for the mean and the variance of a binomial distribution.

4.41 Find the mean and the standard deviation of the distribution of each of the following random variables (having binomial distributions):

(a) The number of heads obtained in 676 flips of a balanced coin.

(b) The number of 4's obtained in 720 rolls of a balanced die.

(c) The number of defectives in a sample of 600 parts made by a machine, when the probability is 0.04 that any one of the parts is defective.

(d) The number of students among 800 interviewed who do not like the food served at the university cafeteria, when the probability is 0.65 that any one of them does not like the food.

4.42 Find the mean and the standard deviation of the hypergeometric distribution with the parameters $n = 3$, $a = 4$, and $N = 8$

(a) by first calculating the necessary probabilities and then using the formulas which define μ and σ;

(b) by using the special formulas for the mean and the variance of a hypergeometric distribution.

4.43 Prove the formula for the mean of the hypergeometric distribution with the parameters n, a, and N, namely, $\mu = n \cdot \frac{a}{N}$.

[*Hint*: Make use of the identity

$$\sum_{r=0}^{k} \binom{m}{r}\binom{s}{k-r} = \binom{m+s}{k}$$

which can be obtained by equating the coefficients of x^k in $(1+x)^m(1+x)^s$ and in $(1+x)^{m+s}$.]

4.44 Construct a table showing the upper limits provided by Chebyshev's theorem for the probabilities of obtaining values differing from the mean by at least 1, 2, and 3 standard deviations and also the corresponding probabilities for the binomial distribution with $n = 16$ and $p = \frac{1}{2}$.

4.45 Over the range of cylindrical parts manufactured on a computer-controlled lathe, the standard deviation of the diameters is 0.002 millimeter.

(a) What does Chebyshev's theorem tell us about the probability that a new part will be within 0.006 unit of the mean μ for that run?

(b) If the 400 parts are made during the run, about what proportion do you expect will lie in the interval in part (a)?

4.46 In 1 out of 6 cases, material for bulletproof vests fails to meet puncture standards. If 405 specimens are tested, what does Chebyshev's theorem tell us about the probability of getting at most 30 or more than 105 cases that do not meet puncture standards?

4.47 Show that for 1 million flips of a balanced coin, the probability is at least 0.99 that the proportion of heads will fall between 0.495 and 0.505.

4.48 How many times do we have to flip a balanced coin to be able to assert with a probability of at most 0.01 that the difference between the proportion of tails and 0.50 will be at least 0.04?

4.49 Prove that

(a) $\sigma^2 = \mu_2' - \mu^2$;

(b) $\mu_3 = \mu_3' - 3\mu_2' \cdot \mu + 2\mu^3$.

4.6 The Poisson Approximation to the Binomial Distribution

The Poisson distribution often serves as a model for counts which do not have a natural upper bound. The **Poisson distribution**, with mean λ (lambda), has probabilities given by

Poisson distribution

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{for } x = 0, 1, 2, \ldots \qquad \lambda > 0$$

When n is large and p is small, binomial probabilities are often approximated by means of the Poisson distribution with λ equal to the product np. Before we justify this approximation, let us point out that $x = 0, 1, 2, \ldots$ means that there is a countable infinity of possibilities, and this requires that we modify the third axiom of probability given on page 58. In its place we substitute the following axiom.

Modification of third axiom of probability

Axiom 3′ If $A_1, A_2, A_3, \ldots$ is a finite or infinite sequence of mutually exclusive events in S, then

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

The other postulates remain unchanged. To verify that $P(S) = 1$ for this formula, we make use of Axiom 3′ and write

$$\sum_{x=0}^{\infty} f(x;\lambda) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$$

Since the infinite series in the expression on the right is the Maclaurin's series for e^{λ}, it follows that

$$\sum_{x=0}^{\infty} f(x;\lambda) = e^{-\lambda} \cdot e^{\lambda} = 1$$

Let us now show that when $n \to \infty$ and $p \to 0$, while $np = \lambda$ remains constant, the limiting form of the binomial distribution is the Poisson distribution given above. First let us substitute $\frac{\lambda}{n}$ for p into the formula for the binomial distribution and simplify the resulting expression; thus, we get

$$\begin{aligned} b(x;n,p) &= \frac{n!}{x!\,(n-x)!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!\,n^x}(\lambda)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)}{x!}(\lambda)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \end{aligned}$$

If we now let $n \to \infty$, we find that

$$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right) \to 1$$

and that

$$\left(1-\frac{\lambda}{n}\right)^{n-x} = \left[\left(1-\frac{\lambda}{n}\right)^{n/\lambda}\right]^{\lambda}\left(1-\frac{\lambda}{n}\right)^{-x} \to e^{-\lambda}$$

Hence, the binomial distribution $b(x; n, p)$ approaches

$$\frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{for } x = 0, 1, 2, \ldots$$

This completes our proof; the distribution at which we arrived is called the **Poisson distribution**, as we already indicated on page 104.

Since the Poisson distribution has many important applications besides approximating binomial probabilities (see Section 4.7), it has been extensively tabulated. Table 2 at the end of the book gives the values of the probabilities

$$F(x; \lambda) = \sum_{k=0}^{x} f(k; \lambda)$$

for values of λ in varying increments from 0.02 to 25, and its use is very similar to that of Table 1. Poisson probabilities are also calculated by many statistical software programs (see Exercises 4.70 and 4.71).

The following are some examples illustrating the Poisson approximation of binomial probabilities. An acceptable rule of thumb is to use this approximation if $n \geq 20$ and $p \leq 0.05$; if $n \geq 100$, the approximation is generally excellent so long as $np \leq 10$.

EXAMPLE **Comparing Poisson and binomial probabilities**

It is known that 5% of the books bound at a certain bindery have defective bindings. Find the probability that 2 of 100 books bound by this bindery will have defective bindings using

(a) the formula for the binomial distribution;

(b) the Poisson approximation to the binomial distribution.

Solution **(a)** Substituting $x = 2$, $n = 100$, and $p = 0.05$ into the formula for the binomial distribution, we get

$$b(2; 100, 0.05) = \binom{100}{2}(0.05)^2(0.95)^{98} = 0.081$$

(b) Substituting $x = 2$ and $\lambda = 100(0.05) = 5$ into the formula for the Poisson distribution, we get

$$f(2; 5) = \frac{5^2 \cdot e^{-5}}{2!} = 0.084$$

It is of interest to note that the difference between the two values we obtained (the error we would make by using the Poisson approximation) is only 0.003. [Had we used Table 2 instead of using a calculator to obtain e^{-5}, we would have obtained $f(2; 5) = F(2; 5) - F(1; 5) = 0.125 - 0.040 = 0.085$.] ■

EXAMPLE **A Poisson approximation to binomial probabilities**

A heavy machinery manufacturer has 3,840 large generators in the field that are under warranty. If the probability is 1/1,200 that any one will fail during the given year, find the probabilities that 0, 1, 2, 3, 4, . . . of the generators will fail during the given year.

Solution The binomial distribution could be used when appropriate computer software is available. However, the expected number is small and the number of generators is large so the Poisson approximation is valid. We take

$$\lambda = 3{,}840 \cdot \frac{1}{1{,}200} = 3.2$$

Consulting Table 2 with $\lambda = 3.2$, and using the identity $f(x;\lambda) = F(x;\lambda) - F(x-1;\lambda)$, we obtain the results shown in the probability histogram of Figure 4.8. ■

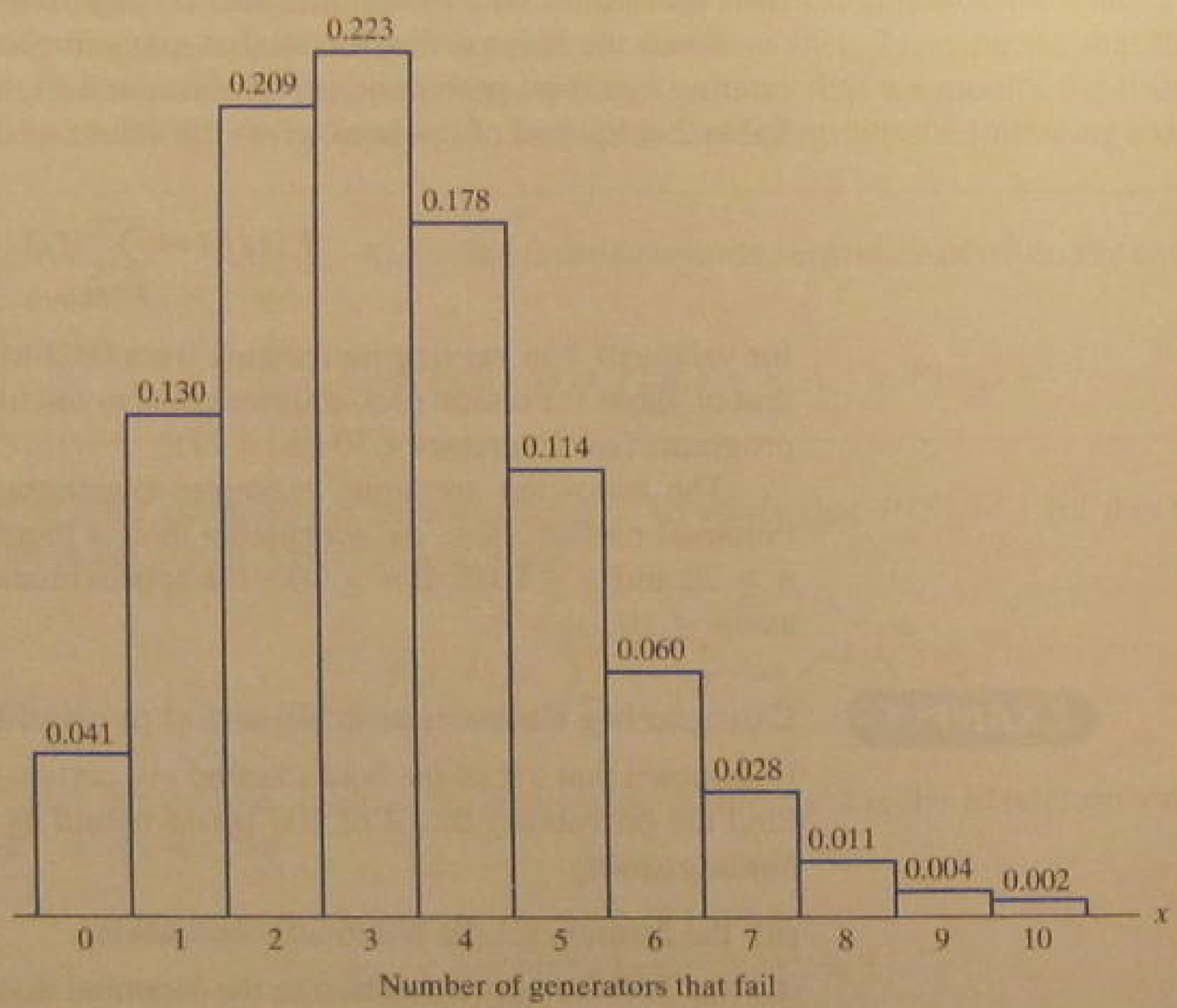


Figure 4.8
Probability histogram of Poisson distribution with $\lambda = 3.2$

Using a method similar to that employed on page 95 to derive the formula for the mean of the binomial distribution, we can show that the mean and the variance of the Poisson distribution with the parameter λ are given by

Mean and variance of Poisson distribution

$$\mu = \lambda \qquad \text{and} \qquad \sigma^2 = \lambda$$

The first of these results should really have been expected, for in our justification of the Poisson approximation to the binomial distribution we let $\lambda = np$. For the variance we can write $\sigma^2 = np(1-p) = \lambda(1-p)$, which approaches λ as $p \to 0$.

EXAMPLE **A Poisson distribution for counts of particles**

Over a 10-minute period, a counter records an average of 1.3 gamma particles per millisecond coming from a radioactive substance. To a good approximation, the

distribution of the count, X, of gamma particles during the next millisecond is Poisson. Determine (a) λ, (b) the probability of one or more gamma particles, and (c) the variance.

Solution **(a)** The mean of the Poisson distribution is λ, so we approximate that $\lambda = 1.3$ to agree with the long-run average.

(b) $P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-1.3} = 0.727$.

(c) The variance of the Poisson is also equal to λ, so $\sigma^2 = 1.3$. ■

4.7 Poisson Processes

In general, a **random process** is a physical process that is wholly or in part controlled by some sort of chance mechanism. It may be a sequence of repeated flips of a coin, measurements of the quality of manufactured products coming off an assembly line, the vibrations of airplane wings, the noise in a radio signal, or any one of numerous other phenomena. What characterizes such processes is their time dependence, namely, the fact that certain events do or do not take place (depending on chance) at regular intervals of time or throughout continuous intervals of time.

In this section we shall be concerned with processes taking place over continuous intervals of time or space, such as the occurrence of imperfections on a continuously produced bolt of cloth, the recording of radiation by means of a Geiger counter, the arrival of telephone calls at a virtual switchboard, or the passing by cars over an electronic counting device. We will now show that the mathematical model which we can use to describe many situations like these is that of the Poisson distribution. To find the probability of x successes during a time interval of length T, we divide the interval into n equal parts of length Δt, so that $T = n \cdot \Delta t$, and we assume that

1. The probability of a success during a very small interval of time Δt is given by $\alpha \cdot \Delta t$.
2. The probability of more than one success during such a small time interval Δt is negligible.
3. The probability of a success during such a time interval does not depend on what happened prior to that time.

This means that the assumptions underlying the binomial distribution are satisfied, and the probability of x successes in the time interval T is given by the binomial probability $b(x; n, p)$ with

$$n = \frac{T}{\Delta t} \qquad \text{and} \qquad p = \alpha \cdot \Delta t$$

Then, following the argument on page 104, we find that when $n \to \infty$ the probability of x successes during the time interval T is given by the corresponding Poisson probability with the parameter

$$\lambda = n \cdot p = \frac{T}{\Delta t} \cdot (\alpha \cdot \Delta t) = \alpha T$$

Since λ is the mean of this Poisson distribution, note that α is the average (mean) number of successes per unit time.

EXAMPLE **Calculating probabilities concerning bad checks**

If a bank receives on the average $\alpha = 6$ bad checks per day, what are the probabilities that it will receive

(a) 4 bad checks on any given day?

(b) 10 bad checks over any 2 consecutive days?

Solution **(a)** Substituting $x = 4$ and $\lambda = \alpha T = 6 \cdot 1 = 6$ into the formula for the Poisson distribution, we get

$$f(4;6) = \frac{6^4 \cdot e^{-6}}{4!} = \frac{1,296(0.00248)}{24} = 0.134$$

(b) Here $\lambda = \alpha \cdot 2 = 12$ so we want to find $f(10;12)$. We write

$$\begin{aligned} f(10;12) &= F(10;12) - F(9;12) \\ &= 0.347 - 0.242 \\ &= 0.105 \end{aligned}$$

where the values of $F(10;12)$ and $F(9;12)$ were obtained from Table 2. ■

EXAMPLE **Calculating the probabilities of imperfections**

In the inspection of tin plate produced by a continuous electrolytic process, 0.2 imperfection is spotted per minute, on average. Find the probabilities of spotting

(a) one imperfection in 3 minutes;
(b) at least two imperfections in 5 minutes;
(c) at most one imperfection in 15 minutes.

Solution Referring in each case to Table 2,

(a) With $\lambda = (0.2) \cdot 3 = 0.6$, we get

$$\begin{aligned} F(1;0.6) - F(0;0.6) &= 0.878 - 0.549 \\ &= 0.329 \end{aligned}$$

(b) With $\lambda = (0.2) \cdot 5 = 1.0$, we get

$$\begin{aligned} 1 - F(1;1.0) &= 1 - 0.736 \\ &= 0.264 \end{aligned}$$

(c) With $\lambda = (0.2) \cdot 15 = 3.0$ we get

$$F(1;3.0) = 0.199$$ ■

The Poisson distribution has many important applications in *queuing* problems, where we may be interested, for example, in the number of customers arriving for service at a cafeteria, the number of ships or trucks arriving to be unloaded at a receiving dock, the number of aircraft arriving at an airport, and so forth. Thus, if, on average, 0.3 customer arrives per minute at a cafeteria, then the probability that exactly 3 customers will arrive during a 5-minute span is

$$F(3;1.5) - F(2;1.5) = 0.934 - 0.809 = 0.125$$

and if, on the average, 3 trucks arrive per hour to be unloaded at a warehouse, then the probability that at most 20 will arrive during an 8-hour day is

$$F(20;24) = 0.243$$

Also, if, on the average, 14.5 planes arrive per day at a private airport, then the probability that the number of arrivals on any one day will be anywhere from 12 to 15, inclusive, is

$$F(15;14.5) - F(11;14.5) = 0.619 - 0.220 = 0.399$$

4.8 The Geometric and Negative Binomial Distribution

On page 46 we indicated that a countably infinite sample space would be needed if we are interested in the number of cars persons have to inspect until they find one whose nitrogen oxide emission does not meet government standards. To treat this kind of problem in general, suppose that in a sequence of trials we are interested in the number of the trial on which the first success occurs. The three assumptions for Bernoulli trials are satisfied but the extra assumption underlying the binomial distribution is not. In other words, n is not fixed.

Clearly, if the first success is to come on the xth trial, it has to be preceded by $x - 1$ failures, and if the probability of a successes is p, the probability of $x - 1$ failures in $x - 1$ trials is $(1 - p)^{x-1}$. Then, if we multiply this expression by the probability p of a success on the xth trial, we find that the probability of getting the first success on the xth trial is given by

Geometric distribution

$$g(x; p) = p(1 - p)^{x-1} \qquad \text{for } x = 1, 2, 3, 4, \ldots$$

This probability distribution is called the **geometric distribution**, and, as the reader will be asked to verify in Exercise 4.68 on page 111, its mean is

Mean of geometric distribution

$$\mu = \frac{1}{p}$$

EXAMPLE **Calculating a probability using the geometric distribution**

If the probability is 0.05 that a certain kind of measuring device will show excessive drift, what is the probability that the sixth measuring device tested will be the first to show excessive drift?

Solution Substituting $x = 6$ and $p = 0.05$ into the formula for the geometric distribution, we get

$$\begin{aligned} g(6; 0.05) &= (0.05)(1 - 0.05)^{6-1} \\ &= 0.039 \end{aligned}$$

■

The **negative binomial distribution** describes the total number of Bernoulli trials, X, to obtain a specified number r successes. When $r = 1$, the negative binomial reduces to the geometric distribution. Since, if the rth success occurs at trial number x, it must be that $r - 1$ successes occurred in the first $x - 1$ trials and the last trial is a success. The probability distribution is then the product of the binomial probability $b(r - 1; x - 1, p)$ and p.

Negative binomial distribution

$$f(x) = \binom{x-1}{r-1} p^r (1 - p)^{x-r} \qquad \text{for } x = r, r + 1, \ldots$$

Exercises

4.50 Prove that for the Poisson distribution

$$\frac{f(x+1;\lambda)}{f(x;\lambda)} = \frac{\lambda}{x+1}$$

for $x = 0, 1, 2, \ldots$.

4.51 Use the recursion formula of Exercise 4.50 to calculate the value of the Poisson distribution with $\lambda = 3$ for $x = 0, 1, 2, \ldots$, and 9, and draw the probability histogram of this distribution. Verify your results by referring to Table 2.

4.52 Use Table 2 to find

(a) $F(4;7)$; (b) $f(4;7)$; (c) $\sum_{k=6}^{19} f(k;8)$.

4.53 Use Table 2 to find

(a) $F(9;12)$; (b) $f(9;12)$; (c) $\sum_{k=3}^{12} f(k;7.5)$.

4.54 Use the Poisson distribution to approximate the binomial probability $b(3;100,0.03)$.

4.55 In a given city, 6% of all drivers get at least one parking ticket per year. Use the Poisson approximation to the binomial distribution to determine the probabilities that among 80 drivers (randomly chosen in this city):

(a) 4 will get at least one parking ticket in any given year;

(b) at least 3 will get at least one parking ticket in any given year;

(c) anywhere from 3 to 6, inclusive, will get at least one parking ticket in any given year.

4.56 If 0.8% of the fuses delivered to an arsenal are defective, use the Poisson approximation to determine the probability that 4 fuses will be defective in a random sample of 400.

4.57 The number of gamma rays emitted per second by a certain radioactive substance is a random variable having the Poisson distribution with $\lambda = 5.8$. If a recording instrument becomes inoperative when there are more than 12 rays per second, what is the probability that this instrument becomes inoperative during any given second?

4.58 A consulting engineer receives, on average, 0.7 requests per week. If the number of requests follows a Poisson process, find the probability that

(a) in a given week, there will be at least 1 request;

(b) in a given 4-week period there will be at least 3 requests.

4.59 At a checkout counter customers arrive at an average of 1.5 per minute. Find the probabilities that

(a) at most 4 will arrive in any given minute;

(b) at least 3 will arrive during an interval of 2 minutes;

(c) at most 15 will arrive during an interval of 6 minutes.

4.60 A company rents time on a computer for periods of t hours, for which it receives \$600 an hour. The number of times the computer breaks down during t hours is a random variable having the Poisson distribution with $\lambda = (0.8)t$, and if the computer breaks down x times during t hours, it costs $50x^2$ dollars to fix it. How should the company select t in order to maximize its expected profit?

4.61 In a "torture test," a light switch is turned on and off until it fails. If the probability that the switch will fail any time it is turned on or off is 0.001, what is the probability that the switch will fail *after* it has been turned on or off 1,200 times? Assume that the conditions underlying the geometric distribution are met. [*Hint*: Use the formula for the value of an infinite geometric progression.]

4.62 An automated weight monitor can detect underfilled cans of beverages with probability 0.98. What is the probability it fails to detect an underfilled can for the first time when it encounters the 10th underfilled can?

4.63 A company fabricates special-purpose robots, and records show that the probability is 0.10 that one of its new robots will require repairs during confirmation tests. What is the probability that the eighth robot it builds in a month is the first one to require repairs?

4.64 Referring to Exercise 4.63, find the probability that the 12th robot built in a month is the second to require repairs.

4.65 During an assembly process, parts arrive just as they are needed. However, at one station, the probability is 0.01 that a defective part will arrive in a one-hour period. Find the probability that

(a) exactly 1 defective part arrives in a 4-hour span;

(b) 1 or more defective parts arrive in a 4-hour span;

(c) exactly 1 defective part arrives in a 4-hour span and exactly 1 defective part arrives in the next 4-hour span.

4.66 The arrival of trucks at a receiving dock is a Poisson process with a mean arrival rate of 2 per hour.

(a) Find the probability that exactly 5 trucks arrive in a two-hour period.

(b) Find the probability that 8 or more trucks arrive in a two-hour period.

(c) Find the probability that exactly 2 trucks arrive in a one-hour period and exactly 3 trucks arrive in the next one-hour period.

4.67 The number of flaws in a fiber optic cable follows a Poisson process with an average of 0.6 per 100 feet.

(a) Find the probability of exactly 2 flaws in a 200-foot cable.

(b) Find the probability of exactly 1 flaw in the first 100 feet and exactly 1 flaw in the second 100 feet.

4.68 Differentiating with respect to p on both sides of the equation

$$\sum_{x=1}^{\infty} p(1-p)^{x-1} = 1$$

show that the geometric distribution

$$f(x) = p(1-p)^{x-1} \qquad \text{for } x = 1, 2, 3, \ldots$$

has the mean $1/p$.

4.69 Use the formulas defining μ and σ^2 to show that the mean and the variance of the Poisson distribution are both equal to λ.

4.70 Poisson probabilities can be calculated using *MINITAB*.

Dialog box:
Calc > Probability Distribution > Poisson
Choose **Probability.**
Choose **Input constant** and enter 2. Type 1.64 in **Mean.**
Click **OK**.

Output:

Poisson with	mu = 1.64000
x	$P(X = x)$
2.00	0.2609

Find the Poisson probabilities for $x = 2$ and $x = 3$ when

(a) $\lambda = 2.73$; (b) $\lambda = 4.33$.

4.71 Cumulative Poisson probabilities can be calculated using *MINITAB*.

Dialog box:
Calc > Probability Distribution > Poisson
Choose **Cumulative Distribution.**
Choose **Input constant** and enter 2. Type 1.64 in **Mean.**
Click **OK**.

Output:

Poisson with	mu = 1.64000
x	$P(X <= x)$
2.00	0.7730

Find the cumulative Poisson probabilities for $x = 2$ and $x = 3$ when

(a) $\lambda = 2.73$; (b) $\lambda = 4.33$.

4.9 The Multinomial Distribution

An immediate generalization of the binomial distribution arises when each trial can have more than two possible outcomes. This happens, for example, when a manufactured product is classified as superior, average, or poor, when a student's performance is graded as an A, B, C, D, or F, or when an experiment is judged successful, unsuccessful, or inconclusive. To treat this kind of problem in general, let us consider the case where there are n independent trials, with each trial permitting k mutually exclusive outcomes whose respective probabilities are

$$p_1, p_2, \ldots, p_k \qquad \text{with } \sum_{i=1}^{k} p_i = 1$$

Referring to the outcomes as being of the first kind, the second kind, ..., and the kth kind, we shall be interested in the probability $f(x_1, x_2, \ldots, x_k)$ of getting x_1 outcomes of the first kind, x_2 outcomes of the second kind, ..., and x_k outcomes of the kth kind, with

$$\sum_{i=1}^{k} x_i = n$$

Using arguments similar to those which we employed in deriving the equation for the binomial distribution in Section 4.2, it can be shown that the desired probability is given by

Multinomial distribution

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

for $x_i = 0, 1, \ldots, n$ for each i, but with the x_i subject to the restriction

$$\sum_{i=1}^{k} x_i = n$$

The **joint probability distribution** whose values are given by these probabilities is called the **multinomial distribution**; it owes its name to the fact that for the various values of the x_i the probabilities are given by the corresponding terms of the multinomial expansion of $(p_1 + p_2 + \cdots + p_k)^n$.

EXAMPLE **Calculating a probability using the multinomial distribution**

The probabilities that the light bulb of a certain kind of projector will last fewer than 40 hours of continuous use, anywhere from 40 to 80 hours of continuous use, or more than 80 hours of continuous use are 0.30, 0.50, and 0.20. Find the probability that among eight such bulbs 2 will last fewer than 40 hours, 5 will last anywhere from 40 to 80 hours, and 1 will last more than 80 hours.

Solution Substituting $n = 8$, $x_1 = 2$, $x_2 = 5$, $x_3 = 1$, $p_1 = 0.30$, $p_2 = 0.50$, and $p_3 = 0.20$ into the formula, we get

$$\begin{aligned} f(2, 5, 1) &= \frac{8!}{2!5!1!}(0.30)^2(0.50)^5(0.20)^1 \\ &= 0.0945 \end{aligned}$$

■

Exercises

4.72 Suppose that the probabilities are, respectively, 0.40, 0.40, and 0.20 that in city driving a certain kind of imported car will average less than 22 miles per gallon, anywhere from 22 to 25 miles per gallon, or more than 25 miles per gallon. Find the probability that among 12 such cars tested, 4 will average less than 22 miles per gallon, 6 will average anywhere from 22 to 25 miles per gallon, and 2 will average more than 25 miles per gallon.

4.73 As can easily be shown, the probabilities of getting 0, 1, or 2 heads with a pair of balanced coins are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$. What is the probability of getting 2 tails twice, 1 head and 1 tail 3 times, and 2 heads once in 6 tosses of a pair of balanced coins?

4.74 Suppose the probabilities are 0.89, 0.09, and 0.02 that the finish on a new car will be rated acceptable, easily repairable, or unacceptable. Find the probability that, among 20 cars painted one morning, 17 have acceptable finishes, 2 have repairable finishes, and 1 finish is unacceptable.

4.75 Using the same sort of reasoning as in the derivation of the formula for the hypergeometric distribution, we can derive a formula which is analogous to the multinomial distribution but applies to sampling without replacement. A set of N objects contains a_1 objects of the first kind, a_2 objects of the second kind, . . . , and a_k objects of the kth kind, so that $a_1 + a_2 + \cdots + a_k = N$. The number of ways in which we can select x_1 objects of the first kind, x_2 objects of the second kind, . . . , and x_k objects of the kth kind is given by the product of the number of ways in which we can select x_1 of the a_1 objects of the first kind, x_2 of the a_2 objects of the second kind, . . . , and x_k of the a_k objects of the kth kind. Thus, the probability of getting that many objects of each kind is simply this product divided by the total number of ways in which $x_1 + x_2 + \cdots + x_k = n$ objects can be selected from the whole set of N objects.

(a) Write a formula for the probability of obtaining x_1 objects of the first kind, x_2 objects of the second kind, . . . and x_k objects of the kth kind.

(b) If 20 defective glass bricks include 10 that have cracks but no discoloration, 7 that are discolored but have no cracks, and 3 that have cracks and discoloration, what is the probability that among 6 of the bricks chosen at random for further checks 3 will have cracks only, 2 will only be discolored, and 1 will have cracks as well as discoloration?

4.10 Simulation

In recent years, simulation techniques have been applied to many problems in the various sciences, and if the processes which are being simulated involve an element of chance, these techniques are referred to as **Monte Carlo methods**. Very often, the use of Monte Carlo simulation eliminates the cost of building and operating expensive equipment; it is used, for instance, in the study of collisions of photons with electrons, the scattering of neutrons, and similar complicated phenomena. Monte Carlo methods are also useful in situations where direct experimentation is impossible—say, in studies of the spread of cholera epidemics, which, of course, cannot be induced experimentally on human populations. In addition, Monte Carlo techniques are sometimes applied to the solution of mathematical problems which cannot be solved by direct means, or where a direct solution is too costly or requires too much time.

A classical example of the use of Monte Carlo methods in the solution of a problem of pure mathematics is the determination of π (the ratio of the circumference of a circle to its diameter) by probabilistic means. Early in the eighteenth century, George de Buffon, a French naturalist, proved that if a very fine needle of length a is thrown at random on a board ruled with equidistant parallel lines, the probability that the needle will intersect one of the lines is $2a/\pi b$, where b is the distance between the parallel lines. What is remarkable about this fact is that it involves the constant $\pi = 3.1415926\ldots$, which in elementary geometry is approximated by the circumferences of regular polygons enclosed in a circle of radius $\frac{1}{2}$. Buffon's result implies that if such a needle is actually tossed a great many times, the proportion of the time it crosses one of the lines gives an estimate of $2a/\pi b$ and, hence, an estimate of π since a and b are known. Early experiments of this kind yielded an estimate of 3.1596 (based on 5,000 trials) and an estimate of 3.155 (based on 3,204 trials) in the middle of the nineteenth century.

Although Monte Carlo methods are sometimes based on actual gambling devices (for example, the needle tossing in the estimation of π), it is usually expedient to use so-called **random digits** or **random numbers** generated by computer software. We will illustrate an application using a table of random numbers that consists of many pages on which the digits of 0, 1, 2, . . . , and 9 are set down in a "random" fashion, much as they would appear if they were generated one at a time by a gambling device giving each digit an equal probability of being selected. Actually, we could also construct such tables ourselves—say, by repeatedly drawing numbered slips out of a hat or by using a perfectly constructed spinner—but in practice such tables are usually generated by means of computers.

Although tables of random numbers are constructed so that the digits can be looked upon as values of a random variable having the discrete uniform distribution $f(x) = \frac{1}{10}$ for $x = 0, 1, 2, \ldots,$ or 9, they can be used to simulate values of any discrete random variable, and even continuous random variables.

To illustrate the use of a table of random numbers, let us simulate, say, tossing three balanced coins. The distribution for the number of heads is

Number of Heads	Probability
0	$1/8 = 0.125$
1	$3/8 = 0.375$
2	$3/8 = 0.375$
3	$1/8 = 0.125$

Since the probabilities in this distribution are given to three decimal places, we use three-digit random numbers. Our scheme is to allocate 125 (or one-eighth) of the 1,000 random numbers from 000 to 999 to 0 heads, 375 (or three-eighths) to 1 head, 375 (or three-eighths) to 2 heads, and 125 (or one-eighth) to 3 heads.

We use the following scheme:

Number of Heads	Probability	Cumulative Probability	Random Numbers
0	0.125	0.125	000–124
1	0.375	0.500	125–449
2	0.375	0.875	500–874
3	0.125	1.000	875–999

The column of **cumulative probabilities** was added to facilitate the assignment of the random numbers. Observe that in each case the last random digit is one less than the number formed by the three decimal digits of the corresponding cumulative probability.

With this scheme, if we arbitrarily use the twenty-second, twenty-third, and twenty-fourth columns of the first page of Table 7, starting with the sixth row and going down the page, we get 197, 365, 157, 520, 946, 951, 948, 568, 586, and 089, and we interpret this as 1, 1, 1, 2, 3, 3, 3, 2, 2, and 0 heads.

The method we have illustrated here with reference to a game of chance can be used to simulate observations of any random variable with a given probability distribution.

However, in practice it is much more efficient to use common computer software based on this scheme.

EXAMPLE **Simulation of arrival of cars at toll booth**

Suppose that the probabilities are 0.082, 0.205, 0.256, 0.214, 0.134, 0.067, 0.028, 0.010, 0.003, and 0.001 that 0, 1, 2, 3, . . . , or 9 cars will arrive at a toll booth of a turnpike during any one-minute interval in the early afternoon.

Use computer software to simulate the arrival of cars at the toll booth during 20 one-minute intervals in the early afternoon.

Solution We illustrate using *MINITAB* with the values set in C1 and the probabilities in C2.

Data:
C1: 0, 1, . . . , 9
C2: 0.082, 0.205, . . . , 0.001

Dialog box:
Calc > Random Data > Discrete
Type 20 after **Generate**. Type *C*3 below **Store**. Type *C*1 in **Values in:**. Type *C*2 in **Probabilities in** Click **OK**.

Output:

```
4 1 5 4 1 2 5 0 1 4
3 3 1 0 1 1 2 5 1 2
```

Suppose we are interested in a somewhat complex event, say, 11 or more cars arrive in at least one three-minute interval among the 20 one-minute intervals. It is a simple manner to repeat the simulation of 20 one-minute periods 100 times. The probability that 11 or more cars arrive in at least one three-minute interval is estimated by the proportion of times that event occurs. In the single sample of size 20 here, that event does not occur. ■

Exercises

4.76 Simulate tossing a coin.

(a) For a balanced coin, generate 100 flips.

(b) For a coin with probability of heads 0.8, generate 100 flips.

4.77 The probabilities that a computer software salesperson will make 0, 1, 2, 3, 4, or 5 sales on any one day are 0.14, 0.28, 0.27, 0.18, 0.09, and 0.04.

(a) Simulate the salesperson's sales on 25 days.

(b) Repeat the simulation of sales on 25 days a total of 100 times. Estimate the probability that the sales on three consecutive days are greater than 9.

4.78 Depending on the availability of parts, a company can manufacture 3, 4, 5, or 6 units of a certain item per week with corresponding probabilities of 0.10, 0.40, 0.30, and 0.20. The probabilities that there will be a weekly demand for 0, 1, 2, 3, . . . , or 8 units are, respectively, 0.05, 0.10, 0.30, 0.30, 0.10, 0.05, 0.05, 0.04, and 0.01. If a unit is sold during the week that it is made, it will yield a profit of \$100; this profit is reduced by \$20 for each week that a unit has to be stored. Simulate the operation of this company for 50 consecutive weeks and estimate its expected weekly profit.

Do's and Don'ts

Do's

1. Keep in mind that any scheme for assigning a numerical value to each possible outcome should quantify a feature of the outcome that is important to the scientist. That is, any random variable should convey pertinent information about the outcome.
2. Describe the chance behavior of a discrete random variable X by its probability distribution function

$$f(x) = P[X = x] \qquad \text{for each possible value } x$$

3. Summarize a probability distribution, or the random variable, by its

$$\text{mean: } \mu = \sum_{\text{all } x} x \cdot f(x)$$

$$\text{variance: } \sigma^2 = \sum_{\text{all } x} (x - \mu)^2 \cdot f(x)$$

4. Use a special family of distributions, for instance the binomial distribution

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } x = 0, 1, \ldots, n$$

having mean np and variance $np(1-p)$, if the underlying assumptions are reasonable. The hypergeometric distribution might be entertained when sampling without replacement from a finite collection of units each of which is one of two possible types. It will be well approximated by the binomial when the sample size n is a small fraction of the population size N.

5. For counts whose possible values do not have a specified upper limit, consider the Poisson distribution

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{for } x = 0, 1, 2, \ldots \qquad \lambda > 0$$

having mean λ and variance λ. You do need to check that the Poisson distribution is reasonable. The sample mean and variance should be about the same size.

Don'ts

1. Never apply the binomial distribution to counts without first checking that the conditions hold for Bernoulli trials: independent trials with the same probability of success for each trial. If the conditions are satisfied, then the binomial distribution is appropriate for the number of successes in a fixed number of trials.
2. Never use the formula $np(1-p)$ for the variance of a count of successes without checking that the trials are independent.

Review Exercises

4.79 A manufacturer of digital phones has the following probability distribution for the number of defects per phone:

x	$f(x)$
0	.89
1	.07
2	.03
3	.01

(a) Determine the probability of 2 or more defects.

(b) Is a randomly selected phone more likely to have 0 defects or 1 or more defects?

4.80 Upon reviewing recent use of conference rooms at an engineering consulting firm, an industrial engineer determined the following probability distribution for the number of requests for a conference room per half-day:

x	$f(x)$
0	.07
1	.15
2	.45
3	.25
4	.08

(a) Currently, the building has two conference rooms. What is the probability that the number of requests will exceed the number of rooms for a given half-day?

(b) What is the probability that the two conference rooms will not be fully utilized on a given half-day?

(c) How many additional conference rooms are required so that the probability of denying a request is not more than 0.10?

4.81 Refer to Exercise 4.80 and obtain the (a) mean; (b) variance; (c) standard deviation for the number of requests for conference rooms.

4.82 Determine whether the following can be probability distributions of a random variable that can take on only the values of 0, 1, and 2:

(a) $f(0) = 0.34 \quad f(1) = 0.34$ and $f(2) = 0.34$.

(b) $f(0) = 0.2 \quad f(1) = 0.6$ and $f(2) = 0.2$.

(c) $f(0) = 0.7 \quad f(1) = 0.4$ and $f(2) = -0.1$.

4.83 Check whether the following can define probability distributions, and explain your answers.

(a) $f(x) = \dfrac{x}{10}$, for $x = 0, 1, 2, 3, 4$.

(b) $f(x) = \frac{1}{3}$, for $x = -1, 0, 1$.

(c) $f(x) = \frac{(x-1)^2}{4}$, for $x = 0, 1, 2, 3$.

4.84 A basketball player makes 90% of her free throws. What is the probability she will miss for the first time on the seventh shot?

4.85 If the probability is 0.20 that a downtime of an automated production process will exceed 2 minutes, find the probability that 3 of 8 downtimes of the process will exceed 2 minutes using (a) the formula for the binomial distribution; (b) Table 1.

4.86 If the probability is 0.85 that a fully charged digital camera battery will take 150 or more pictures, find the probabilities that among 18 such batteries

(a) 16 will take 150 pictures or more;

(b) at least 14 will take 150 pictures or more;

(c) at most two will not take 150 pictures or more.

4.87 In 16 experiments studying the electrical behavior of single cells, 12 use micro-electrodes made of metal and the other 4 use micro-electrodes made from glass tubing. If 2 of the experiments are to be terminated for financial reasons, and they are selected at random, what are the probabilities that

(a) neither uses micro-electrodes made from glass tubing?

(b) only one uses micro-electrodes made from glass tubing?

(c) both use micro-electrodes made from glass tubing?

4.88 As can be easily verified by means of the formula for the binomial distribution, the probabilities of getting 0, 1, 2, or 3 heads in 3 flips of a coin whose probability of heads is 0.4 are 0.216, 0.432, 0.288, and 0.064. Find the mean of this probability distribution using

(a) the formula that defines μ;

(b) the special formula for the mean of a binomial distribution.

4.89 With reference to Exercise 4.88, find the variance of the probability distribution using

(a) the formula that defines σ^2;

(b) the special formula for the variance of a binomial distribution.

4.90 Find the mean and the standard deviation of the distribution of each of the following random variables (having binomial distributions):

(a) The number of heads in 440 flips of a balanced coin.

(b) The number of 6's in 300 rolls of a balanced die.

(c) The number of defectives in a sample of 700 parts made by a machine, when the probability is 0.03 that any one of the parts is defective.

4.91 Use the Poisson distribution to approximate the binomial probability $b(1; 100, 0.02)$.

4.92 With reference to Exercise 4.87, find the mean and the variance of the distribution of the number of micro-electrodes made from glass tubing using

(a) the probabilities obtained in that exercise;

(b) the special formulas for the mean and the variance of a hypergeometric distribution.

4.93 The daily number of orders filled by the parts department of a repair shop is a random variable with $\mu = 142$ and $\sigma = 12$. According to Chebyshev's theorem, with what probability can we assert that on any one day it will fill between 82 and 202 orders?

4.94 Records show that the probability is 0.00004 that a car will have a flat tire while driving through a certain tunnel. Use the formula for the Poisson distribution to approximate the probability that at least 2 of 10,000 cars passing through the tunnel will have a flat tire.

4.95 The number of weekly breakdowns of a computer is a random variable having a Poisson distribution with $\lambda = 0.2$. What is the probability that the computer will operate without a breakdown for 3 consecutive weeks?

4.96 A manufacturer determines that a big screen HDTV set had probabilities of 0.8, 0.15, 0.05, respectively, of being placed in the categories acceptable, minor defect, or major defect. If 3 HDTVs are inspected,

(a) find the probability that 2 are acceptable and 1 is a minor defect;

(b) find the marginal distribution of the number in minor defect;

(c) compare your answer in part (b) with the binomial probabilities $b(x; 3, 0.15)$. Comment.

4.97 Suppose that the probabilities are 0.2466, 0.3452, 0.2417, 0.1128, 0.0395, 0.0111, 0.0026, and 0.0005 that there will be 0, 1, 2, 3, 4, 5, 6, or 7 polluting spills in the Great Lakes on any one day. Simulate the numbers of polluting spills in the Great Lakes in 30 days.

4.98 A candidate invited for a visit has probability 0.6 of being hired. Let X be the number of candidates that visit before 2 are hired. Find

(a) $P(X \leq 4)$;

(b) $P(X \geq 5)$.

Key Terms

Bar chart 84
Bernoulli trials 85
Binomial coefficient 87
Binomial distribution 86
Chebyshev's theorem 100
Cumulative distribution function 84
Cumulative probability 87
Discrete random variable 83
Distribution function 84
Geometric distribution 109
Hypergeometric distribution 89
Joint probability distribution 112
kth moment about the mean 99
kth moment about the origin 98
Kurtosis 99
Law of large numbers 102
Mean 94
Monte Carlo methods 113
Multinomial distribution 111
Negative binomial distribution 109
Negatively skewed distribution 89
Parameter 87
Poisson distribution 104
Positively skewed distribution 89
Probability distribution 83
Probability histogram 83
Random numbers 113
Random process 107
Random variable 82
Repeated trials 85
Sampling with replacement 89
Sampling without replacement 89
Skewed distribution 89
Standard deviation 97
Symmetrical distribution 88
Variance 97

CHAPTER

PROBABILITY DENSITIES

5

Continuous sample spaces and continuous random variables arise when we deal with quantities that are measured on a continuous scale—for instance, when we measure the speed of a car, the amount of alcohol in a person's blood, the efficiency of a solar collector, or the tensile strength of a new alloy.

In this chapter we shall learn how to determine and work with probabilities relating to continuous sample spaces and continuous random variables. The introduction to probability densities in Section 5.1 is followed by a discussion of the normal distribution in Sections 5.2 and 5.3, and various other special probability densities in Sections 5.4 through 5.9. Problems involving more than one random variable are discussed in Section 5.10. Section 5.11 presents the moment generating function method, a tool for finding the distribution of the sum of independent random variables. A method for checking whether a data set appears to be generated by a normal distribution is introduced in Section 5.12.

CHAPTER OUTLINE

5.1 Continuous Random Variables 119
5.2 The Normal Distribution 125
5.3 The Normal Approximation to the Binomial Distribution 132
5.4 Other Probability Densities 135
5.5 The Uniform Distribution 135
5.6 The Log-Normal Distribution 136
5.7 The Gamma Distribution 138
5.8 The Beta Distribution 141
5.9 The Weibull Distribution 142
5.10 Joint Distributions—Discrete and Continuous 145
5.11 Moment Generating Functions 157
5.12 Checking If the Data Are Normal 163
5.13 Transforming Observations to Near Normality 165
5.14 Simulation 167
Review Exercises 171
Key Terms 172

5.1 Continuous Random Variables

When we first introduced the concept of a random variable in Chapter 4, we presented it as a real-valued function defined over the sample space of an experiment, and we illustrated this idea with the random variable giving the number of preferred attributes possessed by a used car, assigning the numbers 0, 1, 2, or 3 (whichever was appropriate) to the 18 possible outcomes of the experiment. In the continuous case, where random variables can assume values on a continuous scale, the procedure is very much the same. The outcomes of an experiment are represented by the points on a line segment or a line. Then, a random variable is created by appropriately assigning a number to each point by means of some rule or equation. When the value of a random variable is given directly by a measurement or observation, we usually do not bother to differentiate among the value of the random variable, the measurement which we obtain, and the outcome of the experiment, which is the corresponding point on the real axis. If an experiment consists of determining what force is required to break a given tensile-test specimen, the result itself, say, 138.4 pounds, is the value of the random variable, X, with which we are concerned. There is no real need in that case to add that the sample space of the experiment consists of all (or part of) the points on the positive real axis. In general, we shall write $P(a \leq X \leq b)$ for the probability associated with the points of the sample space for which the value of a random variable falls on the interval from a to b.

The problem of defining probabilities in connection with continuous sample spaces and continuous random variables involves some complications. To illustrate the nature of these complications, let us consider the following situation: Suppose we want to know the probability that if an accident occurs on a freeway whose length is 200 miles, it will happen at some given location or, perhaps, some particular stretch

of the road. The outcomes of this experiment can be looked upon as a continuum of points, namely, those on the continuous interval from 0 to 200. Suppose the probability that the accident occurs on any interval of length L is $L/200$, with L measured in miles. Note that this arbitrary assignment of probability is consistent with Axioms 1 and 2 on page 58, since the probabilities are all nonnegative and less than or equal to 1, and $P(S) = \frac{200}{200} = 1$. Of course, we are considering so far only events represented by intervals which form part of the line segment from 0 to 200. Using Axiom $3'$ on page 104, we can also obtain probabilities of events that are not intervals but which can be represented by the union of finitely many or countably many intervals. Thus, for two nonoverlapping intervals of length L_1 and L_2 we have a probability of

$$\frac{L_1 + L_2}{200}$$

and for an infinite sequence of nonoverlapping intervals of length $L_1, L_2, L_3, \ldots$, we have a probability of

$$\frac{L_1 + L_2 + L_3 + \cdots}{200}$$

Note that the probability that the accident occurs at any given point is equal to zero because we can look upon a point as an interval of zero length. However, the probability that the accident occurs in a very short interval is positive; for instance, for an interval of length 1 foot the probability is $(5{,}280 \times 200)^{-1} = 9.5 \times 10^{-7}$.

Thus, in extending the concept of probability to the continuous case, we again use Axioms 1, 2, and $3'$, but we shall have to restrict the meaning of the term *event*. So far as practical considerations are concerned, this restriction is of no consequence; we simply do not assign probabilities to some rather abstruse point sets, which cannot be expressed as the unions or intersections of finitely many or countably many intervals.

The way in which we assigned probabilities in the preceding example is, of course, very special; it is similar in nature to the way in which we assign equal probabilities to the six faces of a die, heads and tails, the 52 cards in a standard deck, and so forth. To treat the problem of associating probabilities with continuous random variables generally, suppose we are interested in the probability that a given random variable will take on a value on the interval from a to b, where a and b are constants with $a \leq b$. Suppose, furthermore, that we divide the interval from a to b into m equal subintervals of width Δx containing, respectively, the points $x_1, x_2, \ldots, x_m$, and that the probability that the random variable will take on a value in the subinterval containing x_i is given by $f(x_i) \cdot \Delta x$. Then the probability that the random variable with which we are concerned will take on a value in the interval from a to b is given by

$$P(a \leq X \leq b) = \sum_{i=1}^{m} f(x_i) \cdot \Delta x$$

Now, if f is an integrable function defined for all values of the random variable with which we are concerned, we shall define the probability that the value of the random variable falls between a and b by letting $\Delta x \to 0$, namely, as

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

As illustrated in Figure 5.1, this definition of probability in the continuous case presupposes the existence of an appropriate function f which, integrated from any constant a to any constant b (with $a \leq b$), gives the probability that the

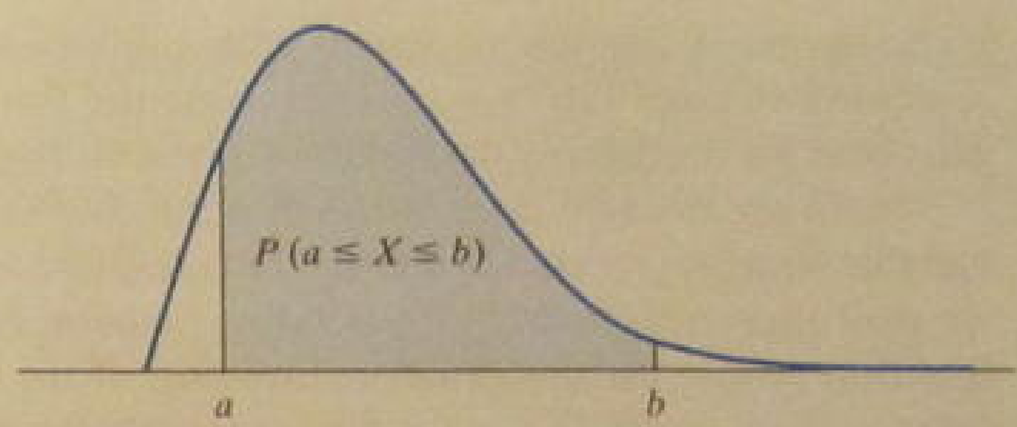


Figure 5.1
Probability as area under f

corresponding random variable takes on a value on the interval from a to b. Note that the value $f(x)$ does not give the probability that the corresponding random variable takes on the value x. In the continuous case, probabilities are given by integrals and not by the values $f(x)$.

To obtain the probability that a random variable will actually take on a given value x, we might first determine the probability that it will take on a value on the interval from $x - \Delta x$ to $x + \Delta x$, and then let $\Delta x \to 0$. However, if we did this it would become apparent that the result is always zero. The fact that the probability is always zero that a continuous random variable will take on any given value x should not be disturbing. Indeed, our definition of probability for the continuous case provides a remarkably good model for dealing with measurements or observations. Owing to the limits of our ability to measure, experimental data never seem to come from a continuous sample space. Thus, while temperatures are productively thought of as points on a continuous scale, if we report a temperature measurement of 74.8 degrees centigrade, we really mean that the temperature lies in the interval from 74.75 to 74.85 degrees centigrade, and not that it is exactly 74.800 . . . .

It is important to add that when we say that there is a zero probability that a random variable will take on any given value x, this does not mean that it is impossible that the random variable will take on the value x. In the continuous case, a zero probability does not imply logical impossibility, but the whole matter is largely academic since, owing to the limitations of our ability to measure and observe, we are always interested in probabilities connected with intervals and not with isolated points.

As an immediate consequence of the fact that in the continuous case probabilities associated with individual points are always zero, we find that if we speak of the probability associated with the interval from a to b, it does not matter whether either endpoint is included. Symbolically,

$$P(a \le X \le b) = P(a \le X < b) = P(a < X \le b) = P(a < X < b)$$

Drawing an analogy with the concept of a density function in physics, we call the functions f, whose existence we stipulated in extending our definition of probability to the continuous cases, **probability density functions**, or simply **probability densities**. Whereas density functions are integrated to obtain weights, probability density functions are integrated to obtain probabilities. We will follow the common practice of also calling $f(x)$ the probability density function with the understanding that we are referring to the function f which assigns the value $f(x)$ to x, for each x that is a possible value for the random variable X.

Since a probability density, integrated between any two constants a and b, gives the probability that a random variable assumes a value between these limits, f cannot be just any real-valued integrable function. However, if we impose the conditions that

$$f(x) \ge 0 \quad \text{for all } x$$

and

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

it can be shown that the axioms of probability (with the modification about events discussed on page 120) are satisfied. Note the similarity between these conditions and those for probability distributions given on page 83.

As in the discrete case, we shall write as $F(x)$ the probability that a random variable with the probability density $f(x)$ takes on a value less than or equal to x, and we shall again refer to the corresponding function F as the **cumulative distribution function** or just the **distribution function** of the random variable. Thus, for any value x, $F(x) = P(X \leq x)$ is the area under the probability density function over the interval $-\infty$ to x. In the usual calculus notation for the integral,

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

Consequently, the probability that the random variable will take on a value on the interval from a to b is $F(b) - F(a)$, and according to the fundamental theorem of integral calculus it follows that

$$\frac{dF(x)}{dx} = f(x)$$

wherever this derivative exists.

EXAMPLE **Calculating probabilities from the probability density function**

If a random variable has the probability density

$$f(x) = \begin{cases} 2e^{-2x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

find the probabilities that it will take on a value

(a) between 1 and 3;

(b) greater than 0.5.

Solution Evaluating the necessary integrals, we get

(a)
$$\int_{1}^{3} 2e^{-2x}\,dx = e^{-2} - e^{-6} = 0.133$$

(b)
$$\int_{0.5}^{\infty} 2e^{-2x}\,dx = e^{-1} = 0.368$$

■

Note that in the preceding example we made the domain of f include all the real numbers even though the probability is zero that x will be negative. This is a practice we shall follow throughout the book. It is also apparent from the graph of this function in Figure 5.2 that it has a discontinuity at $x = 0$; indeed, a probability density need not be everywhere continuous, as long as it is integrable between any two limits a and b (with $a \leq b$).

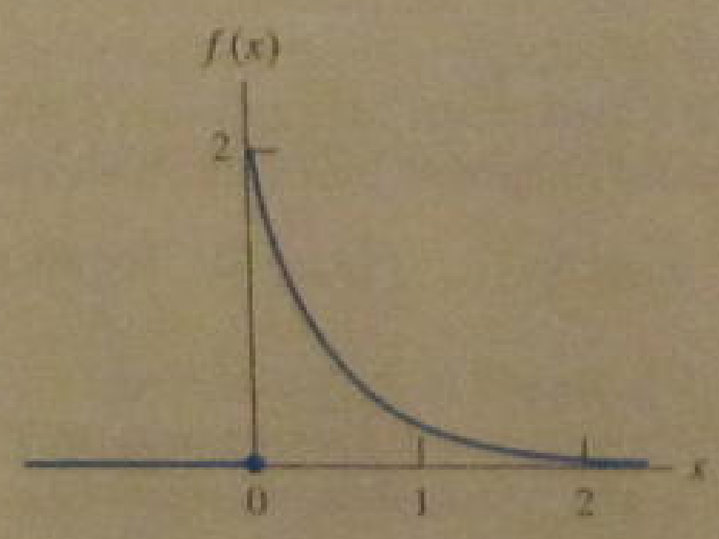


Figure 5.2
Graph of probability density $f(x) = 2e^{-2x}$, $x > 0$

EXAMPLE **Determining a distribution function from its density function**

With reference to the preceding example, find the distribution function and use it to determine the probability that the random variable will take on a value less than or equal to 1.

Solution Performing the necessary integrations, we get

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \int_0^x 2e^{-2t}\,dt = 1 - e^{-2x} & \text{for } x > 0 \end{cases}$$

and substitution of $x = 1$ yields

$$F(1) = 1 - e^{-2} = 0.865$$

■

Note that the distribution function of this example is nondecreasing and that $F(-\infty) = 0$ and $F(\infty) = 1$. Indeed, it follows by definition that these properties are shared by all distribution functions.

EXAMPLE **A probability density function assigns probability one to $(-\infty, \infty)$**

Find k so that the following can serve as the probability density of a random variable:

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ kxe^{-4x^2} & \text{for } x > 0 \end{cases}$$

Solution To satisfy the first of the two conditions on page 121, k must be nonnegative, and to satisfy the second condition we must have

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^{\infty} kxe^{-4x^2}\,dx = \int_0^{\infty} \frac{k}{8} \cdot e^{-u}\,du = \frac{k}{8} = 1$$

so that $k = 8$.

■

The statistical measures that are used to describe probability densities are very similar to the ones which we used to describe probability distributions. Replacing summations with integrals, we define the ***k*th moment about the origin** as

$$\mu'_k = \int_{-\infty}^{\infty} x^k \cdot f(x)\,dx$$

analogous to the definition we gave on page 98. The first moment about the origin is again referred to as the **mean**, and it is denoted by μ. It is also called the expected value of a random variable having the probability density $f(x)$.

Mean of a probability density

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx$$

We also define the ***k*th moment about the mean** as

$$\mu_k = \int_{-\infty}^{\infty} (x - \mu)^k \cdot f(x)\,dx$$

In particular, the second moment about the mean is again referred to as the **variance** and it is written as σ^2; as before, it measures the spread of a probability density in the sense that it gives the expected value of the squared deviation from the mean.

Variance of a probability density

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

Also, σ is again referred to as the **standard deviation**.

EXAMPLE **Determining the mean and variance using the probability density function**

With reference to the example on page 122, find the mean and the variance of the given probability density.

Solution Performing the necessary integrations, using integrations by parts, we get

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{0}^{\infty} x \cdot 2e^{-2x}\,dx = \frac{1}{2}$$

and

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{0}^{\infty} \left(x - \frac{1}{2}\right)^2 \cdot 2e^{-2x}\,dx = \frac{1}{4}$$ ■

Exercises

5.1 Verify that the function of the example on page 122 is, in fact, a probability density.

5.2 If the probability density of a random variable is given by

$$f(x) = \begin{cases} kx^2 & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

find the value k and the probability that the random variable takes on a value

(a) between $\frac{1}{4}$ and $\frac{3}{4}$; (b) greater than $\frac{2}{3}$.

5.3 With reference to the preceding exercise, find the corresponding distribution function and use it to determine the probabilities that a random variable having this distribution function will take on a value

(a) greater than 0.8; (b) between 0.2 and 0.4.

5.4 If the probability density of a random variable is given by

$$f(x) = \begin{cases} x & \text{for } 0 < x < 1 \\ 2 - x & \text{for } 1 \le x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

find the probabilities that a random variable having this probability density will take on a value

(a) between 0.2 and 0.8; (b) between 0.6 and 1.2.

5.5 With reference to the preceding exercise, find the corresponding distribution function, and use it to determine the probabilities that a random variable having the distribution function will take on a value

(a) greater than 1.8;

(b) between 0.4 and 1.6.

5.6 Given the probability density $f(x) = \dfrac{k}{1+x^2}$ for $-\infty < x < \infty$, find k.

5.7 If the distribution function of a random variable is given by

$$F(x) = \begin{cases} 1 - \dfrac{1}{x^2} & \text{for } x > 1 \\ 0 & \text{for } x \le 1 \end{cases}$$

find the probabilities that this random variable will take on a value

(a) less than 3; (b) between 4 and 5.

5.8 Find the probability density that corresponds to the distribution function of Exercise 5.7. Are there any points at which it is undefined? Also sketch the graphs of the distribution function and the probability density.

5.9 Let the phase error in a tracking device have probability density

$$f(x) = \begin{cases} \cos x & 0 < x < \pi/2 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability that the phase error is

(a) between 0 and $\pi/4$; (b) greater than $\pi/3$.

5.10 The length of satisfactory service (years) provided by a certain model of laptop computer is a random variable

having the probability density

$$f(x) = \begin{cases} \frac{1}{4.5} e^{-x/4.5} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

Find the probabilities that one of these laptops will provide satisfactory service for

(a) at most 2.5 years;

(b) anywhere from 4 to 6 years;

(c) at least 6.75 years.

5.11 In a certain city, the daily consumption of electric power (in millions of kilowatt-hours) is a random variable having the probability density

$$f(x) = \begin{cases} \frac{1}{9} xe^{-x/3} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

If the city's power plant has a daily capacity of 12 million kilowatt-hours, what is the probability that this power supply will be inadequate on any given day?

5.12 Prove that the identity $\sigma^2 = \mu_2' - \mu_2$ holds for any probability density for which these moments exist.

5.13 Find μ and σ^2 for the probability density of Exercise 5.2.

5.14 Find μ and σ^2 for the probability density of Exercise 5.4.

5.15 Find μ and σ for the probability density obtained in Exercise 5.8.

5.16 Find μ and σ for the distribution of the phase error of Exercise 5.9.

5.17 Find μ for the distribution of the satisfactory service of Exercise 5.10.

5.18 Show that μ_2' and, hence, σ^2 do not exist for the probability density of Exercise 5.6.

5.2 The Normal Distribution

Among the special probability densities we shall study in this chapter, the **normal probability density**, usually referred to simply as the **normal distribution**, is by far the most important.[1] It was studied first in the eighteenth century when scientists observed an astonishing degree of regularity in errors of measurement. They found that the patterns (distributions) they observed were closely approximated by a continuous distribution, which they referred to as the "normal curve of errors" and attributed to the laws of chance. The equation of the normal probability density, whose graph (shaped like the cross section of a bell) is shown in Figure 5.3, is

Normal distribution

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

and in Exercises 5.42 and 5.43, the reader will be asked to verify that its parameters μ and σ are indeed its mean and its standard deviation.

Since the normal probability density cannot be integrated in closed form between every pair of limits a and b, probabilities relating to normal distributions are usually

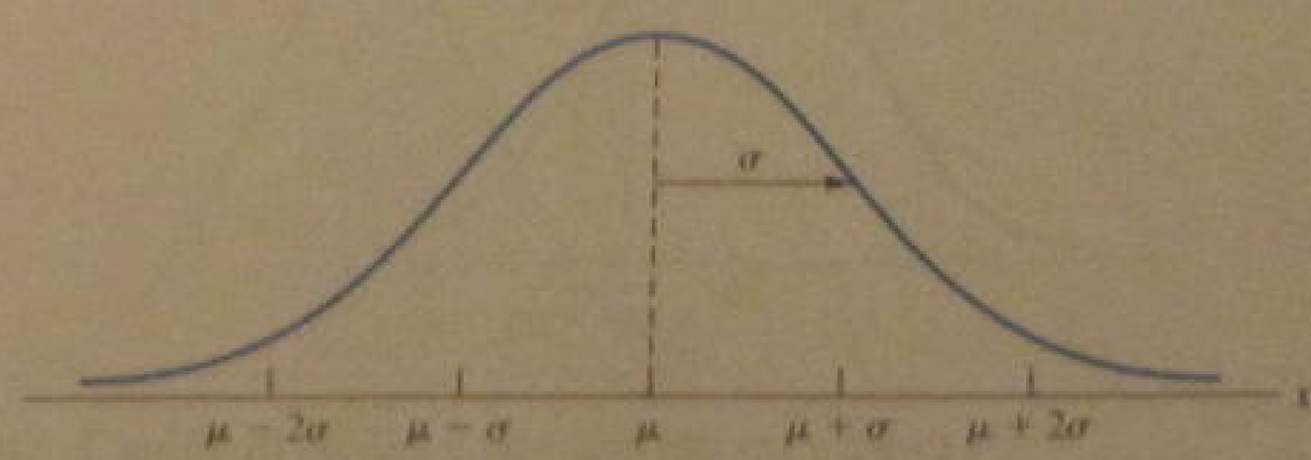


Figure 5.3
Graph of normal probability density

[1]The words *density* and *distribution* are often used interchangeably in the literature of applied statistics.

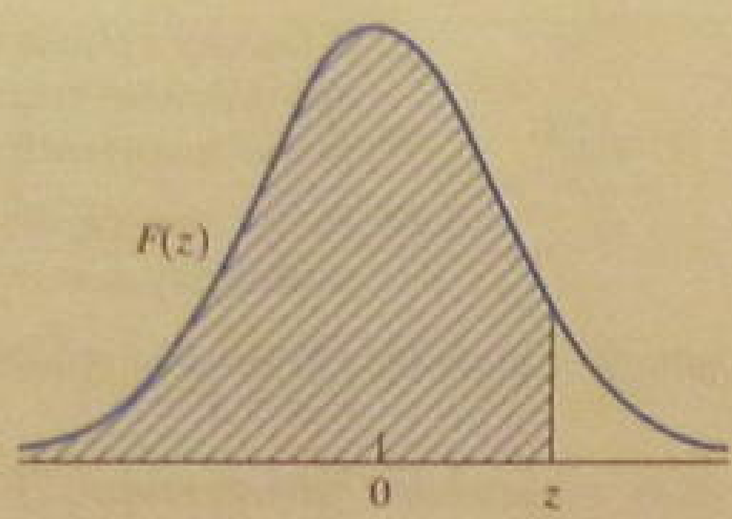


Figure 5.4
The standard normal probabilities
$F(z) = P(Z \leq z)$

obtained from special tables, such as Table 3 at the end of this book. This table pertains to the **standard normal distribution**, namely, the normal distribution with $\mu = 0$ and $\sigma = 1$, and its entries are the values of

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt = P(Z \leq z)$$

for positive or negative $z = 0.00, 0.01, 0.02, \ldots, 3.49$, and also $z = 3.50$, $z = 4.00$, and $z = 5.00$. The cumulative probabilities $F(z)$ correspond to the area under the standard normal density to the left of z, as shown by the shaded area in Figure 5.4.

To find the probability that a random variable having the standard normal distribution will take on a value between a and b, we use the equation

$$P(a < Z \leq b) = F(b) - F(a)$$

as shown by the shaded area in Figure 5.5. We also sometimes make use of the identity $F(-z) = 1 - F(z)$, which holds for all symmetric distributions centered around 0, which the reader will be asked to verify in Exercise 5.41.

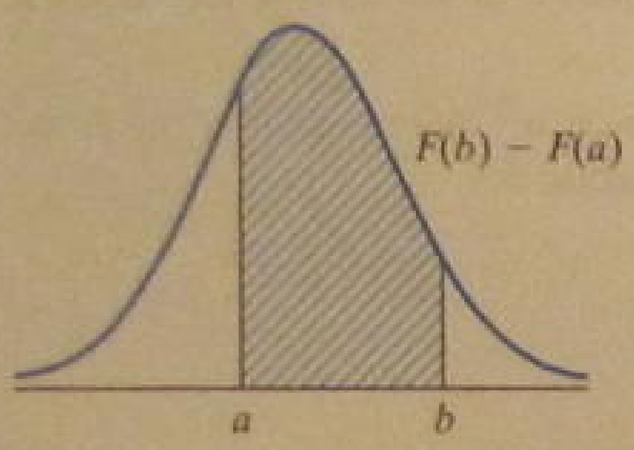


Figure 5.5
The standard normal probability $F(b) - F(a) = P(a < Z \leq b)$

EXAMPLE **Calculating some standard normal probabilities**

Find the probabilities that a random variable having the standard normal distribution will take on a value

(a) between 0.87 and 1.28;
(b) between −0.34 and 0.62;
(c) greater than 0.85;
(d) greater than −0.65.

Solution It is helpful to first indicate the area of interest in a graph as in Figure 5.6.

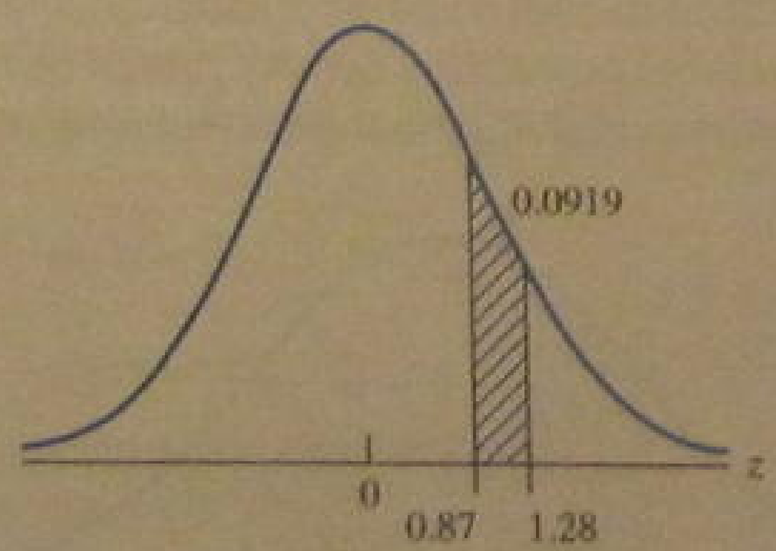


Figure 5.6
$P(0.87 < Z < 1.28)$

Looking up the necessary values in Table 3, for part (a) we get

$$\begin{aligned} F(1.28) - F(0.87) &= 0.8997 - 0.8078 \\ &= 0.0919 \end{aligned}$$

As indicated in Figure 5.7 for part (b),

$$F(0.62) - F(-0.34) = 0.7324 - 0.3669$$
$$= 0.3655$$

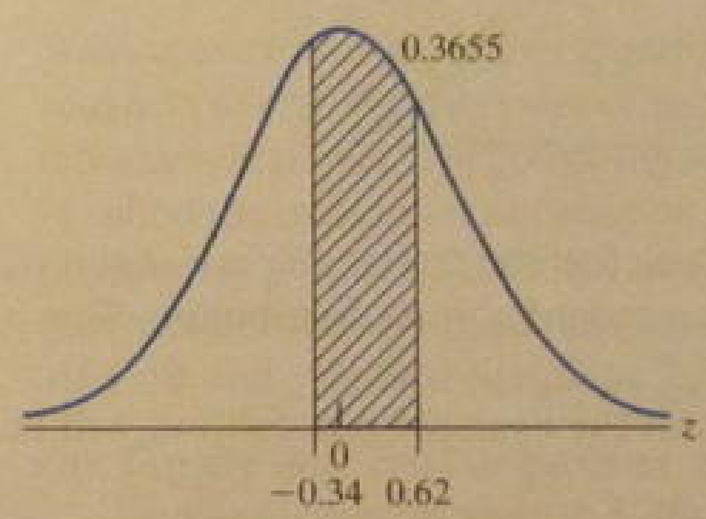


Figure 5.7
$P(-0.34 < Z < 0.62)$

As indicated in Figure 5.8 for part (c),

$$1 - F(0.85) = 1 - 0.8023$$
$$= 0.1977$$

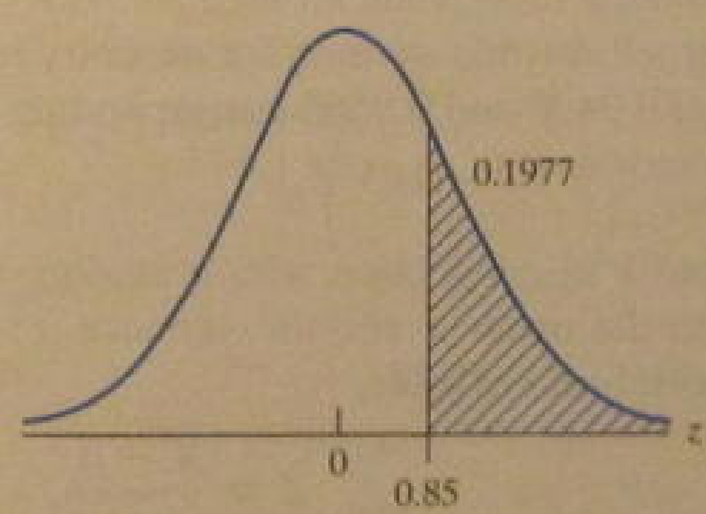


Figure 5.8
$P(Z > 0.85)$

As indicated in Figure 5.9 for part (d)

$$1 - F(-0.65) = 1 - 0.2578 = 0.7422$$

or, alternatively,

$$1 - F(-0.65) = 1 - [1 - F(0.65)]$$
$$= F(0.65)$$
$$= 0.7422$$

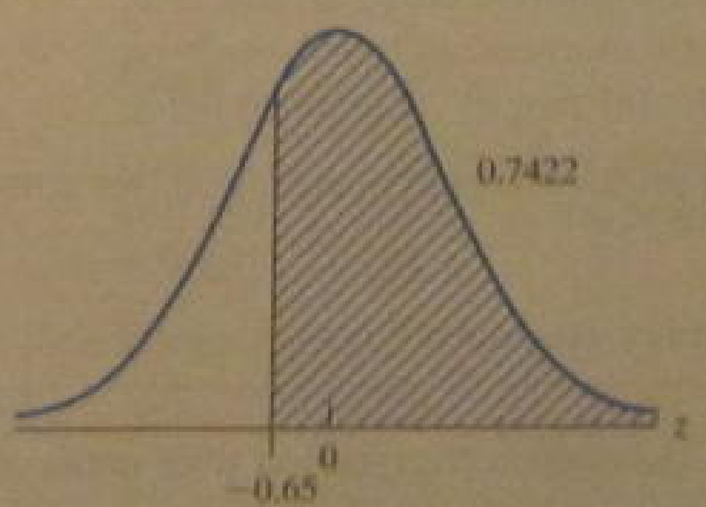


Figure 5.9
$P(Z > -0.65)$

■

There are also problems in which we are given probabilities relating to standard normal distributions and asked to find the corresponding values of z.

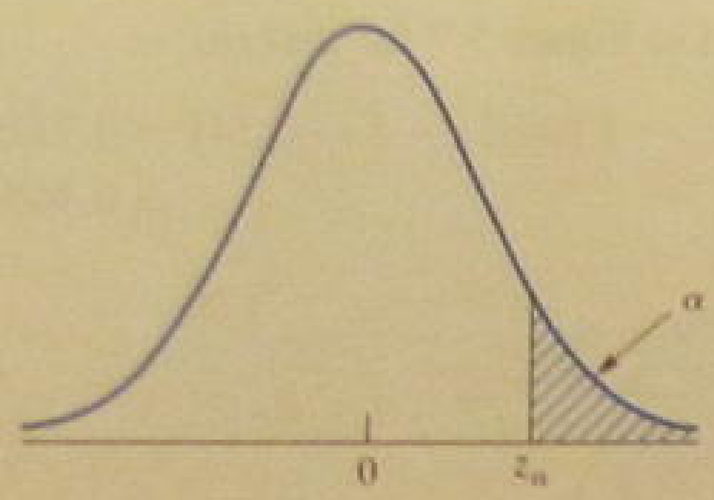


Figure 5.10
The z_α notation for a standard normal distribution

Let z_α be such that the probability is α that it will be exceeded by a random variable having the standard normal distribution. That is, $\alpha = P(Z > z_\alpha)$ as illustrated in Figure 5.10.

The results of the example that follows will be used extensively in subsequent chapters.

EXAMPLE **Two important values for z_α**

Find (a) $z_{0.01}$; (b) $z_{0.05}$.

Solution

(a) Since $F(z_{0.01}) = 0.99$, we look for the entry in Table 3 which is closest to 0.99 and get 0.9901 corresponding to $z = 2.33$. Thus $z_{0.01} = 2.33$.

(b) Since $F(z_{0.05}) = 0.95$, we look for the entry in Table 3 which is closest to 0.95 and get 0.9495 and 0.9505 corresponding to $z = 1.64$ and $z = 1.65$. Thus, by interpolation, $z_{0.05} = 1.645$. ■

To use Table 3 in connection with a random variable X which has a normal distribution with the mean μ and the variance σ^2, we refer to the corresponding **standardized random variable**,

$$Z = \frac{X - \mu}{\sigma}$$

which can be shown to have the standard normal distribution. Thus, to find the probability that the original random variable will take on a value less than or equal to a, in Table 3 we look up

$$F\left(\frac{a - \mu}{\sigma}\right)$$

Also, if we want to find the probability that a random variable having the normal distribution with the mean μ and the variance σ^2 will take on a value between a and b, we have only to calculate the probability that a random variable having the standard normal distribution will take on a value between

$$\frac{a - \mu}{\sigma} \quad \text{and} \quad \frac{b - \mu}{\sigma}$$

That is, to find probabilities concerning X, we convert its values to z **scores** using

$$Z = \frac{X - \mu}{\sigma}$$

Normal probabilities

When X has the normal distribution with mean μ and standard deviation σ,

$$P(a < X \leq b) = F\left(\frac{b - \mu}{\sigma}\right) - F\left(\frac{a - \mu}{\sigma}\right)$$

According to Figure 2.9 on page 20, the observations on the strength of an aluminum alloy appear to be normally distributed. The normal distribution is often used to model variation when the distribution is symmetric and has a single mode.

EXAMPLE **Calculation of probabilities using a normal distribution**

With an eye toward improving performance, industrial engineers studied the ability of scanners to read the bar codes of various food and household products. The maximum reduction in power, just before the scanner cannot read the bar code at a fixed distance, is called the maximum attenuation. This quantity, measured in decibels, varies from product to product. After collecting considerable data, the engineers decided to model the variation in maximum attenuation as a normal distribution with mean 10.1 dB and standard deviation 2.7 dB.

(a) For the next food or product, what is the probability that its maximum attenuation is between 8.5 dB and 13.0 dB?

(b) According to the normal model, what proportion of the products have maximum attenuation between 8.5 dB and 13.0 dB?

(c) What proportion of the products have maximum attenuation greater than 15.1 dB?

Solution **(a)** We treat the maximum attenuation of the next product, X, as a random selection for the normal distribution with $\mu = 10.1$ and $\sigma = 2.7$. Consequently, $Z = (X - 10.1)/2.7$ and, from Table 3, we get

$$\begin{aligned} F\left(\frac{13.0 - 10.1}{2.7}\right) - F\left(\frac{8.5 - 10.1}{2.7}\right) &= F(1.07) - F(-0.59) \\ &= 0.8577 - 0.2776 \\ &= 0.5801 \end{aligned}$$

corresponding to the shaded area in Figure 5.11.

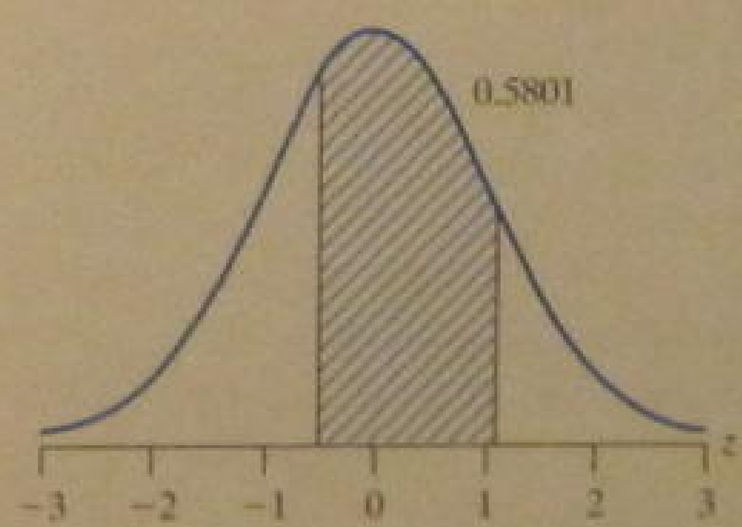


Figure 5.11
$P(8.5 < X < 13.0) =$
$P(-0.59 < Z < 1.07)$

(b) The variation in maximum attenuation for the vast, but finite, collection of all different products is modeled by a normal distribution. The proportion of products having maximum attenuation between 8.5 and 13.0 dB corresponds to the probability in part (a). When we consider the even larger infinite population of all existing products and those that could have been made, we still refer to 0.5801 as the proportion having maximum attenuation between 8.5 and 13.0 dB.

(c) Looking up the necessary value in Table 3,

$$\begin{aligned} 1 - F\left(\frac{15.1 - 10.1}{2.7}\right) &= 1 - F(1.85) \\ &= 1 - 0.9678 \\ &= 0.0322 \end{aligned}$$

corresponding to the shaded area in Figure 5.12, on page 130. ■

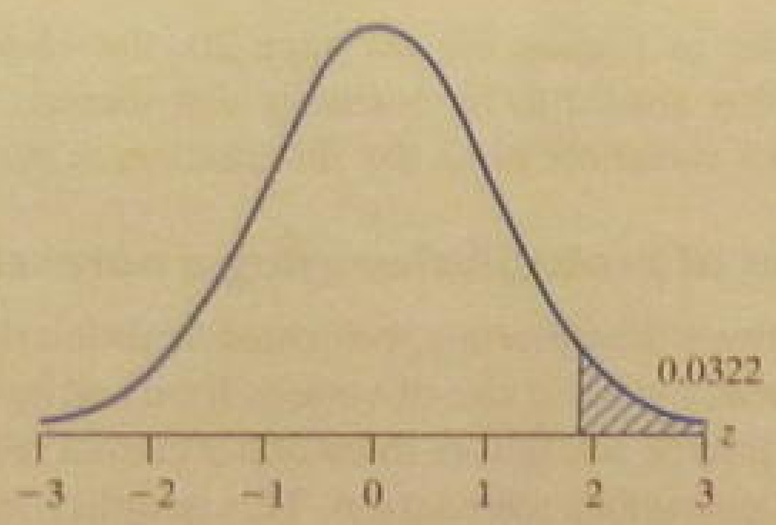


Figure 5.12
$P(X \geq 15.1)$

EXAMPLE **Using normal probabilities to determine the mean fill of jars**

The actual amount of instant coffee that a filling machine puts into "4-ounce" jars may be looked upon as a random variable having a normal distribution with $\sigma = 0.04$ ounce. If only 2% of the jars are to contain less than 4 ounces, what should be the mean fill of these jars?

Solution To find μ such that

$$F\left(\frac{4-\mu}{0.04}\right) = 0.02$$

we look for the entry in Table 3 closest to 0.02 and get 0.0202, corresponding to $z = -2.05$. As indicated in Figure 5.13,

$$\frac{4-\mu}{0.04} = -2.05$$

and, solving for μ, we find that $\mu = 4.082$ ounces. ■

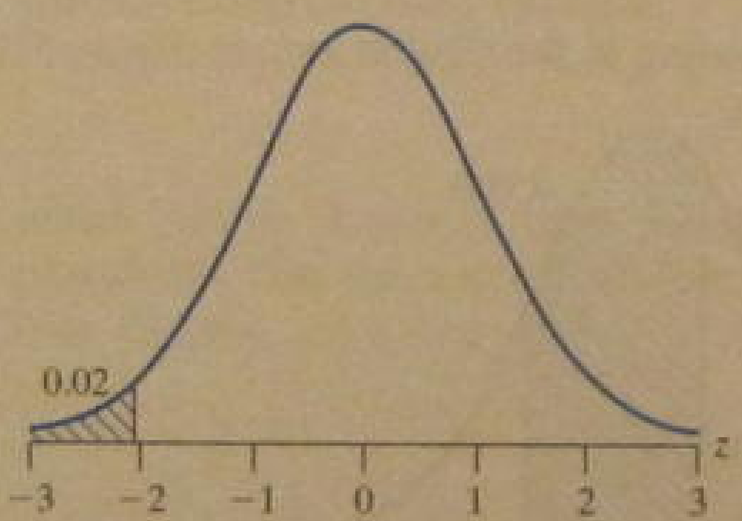


Figure 5.13
$P(Z < -2.05)$

EXAMPLE **Calculating probabilities when ln X has a normal distribution**

After a large number of assays of the gold content in rocks collected from an open pit mine, a mining engineer postulates that the natural log of the gold content (oz/st gold) follows a normal distribution with mean −4.6 and variance 1.21. Under this distribution, would it be unusual to get 0.0015 oz/st gold or less in an assay?

Solution Because it is ln X that has a normal distribution, the question concerns the standardized value

$$\frac{\ln(0.0015) - (-4.6)}{\sqrt{1.21}} = -1.729$$

The standard normal probability of obtaining this value or smaller (see Figure 5.14) is

$$F\left(\frac{\ln(0.0015) - (-4.6)}{\sqrt{1.21}}\right) = F(-1.73) = 0.0419$$

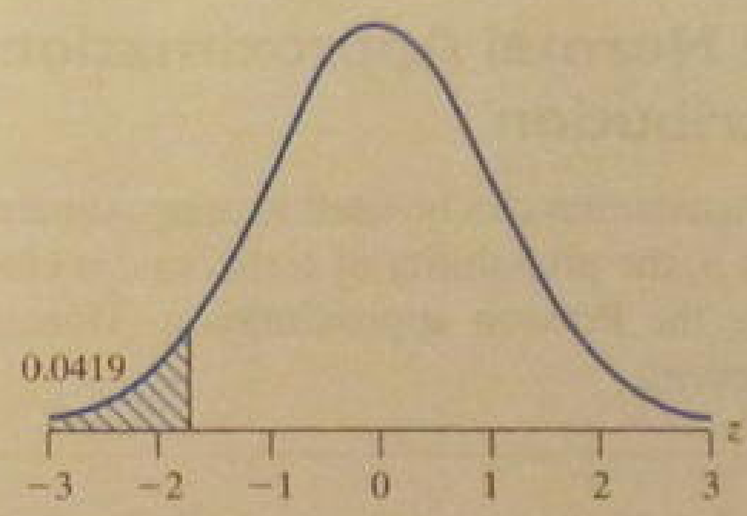


Figure 5.14
$P(Z \leq -1.73) = P(X \leq 0.0015)$

This probability is small, so we suspect that the postulated normal distribution with mean −4.6 does not apply. An assay with this small amount of gold content could suggest that the specimen was collected outside of the vein. ■

Although the normal distribution applies to continuous random variables, it is often used to approximate distributions of discrete random variables. Quite often, this yields satisfactory results, provided that we make the **continuity correction** illustrated in the following example.

EXAMPLE **A continuity correction to improve the normal approximation to a count variable**

In a certain city, the number of power outages per month is a random variable, having a distribution with $\mu = 11.6$ and $\sigma = 3.3$. If this distribution can be approximated closely with a normal distribution, what is the probability that there will be at least 8 outages in any one month?

Solution The answer is given by the area of the lined region of Figure 5.15—the area to the right of 7.5, not 8. The reason for this is that the number of outages is a discrete random variable, and if we want to approximate its distribution with a normal distribution, we must "spread" its values over a continuous scale. We do this by representing each integer k by the interval from $k - \frac{1}{2}$ to $k + \frac{1}{2}$. For instance, 3 is represented by the interval from 2.5 to 3.5, 10 is represented by the interval from 9.5 to 10.5, and "at least 8" is represented by the interval to the right of 7.5. Thus the desired probability is approximated by

$$\begin{aligned} 1 - F\left(\frac{7.5 - 11.6}{3.3}\right) &= 1 - F(-1.24) \\ &= F(1.24) \\ &= 0.8925 \end{aligned}$$

■

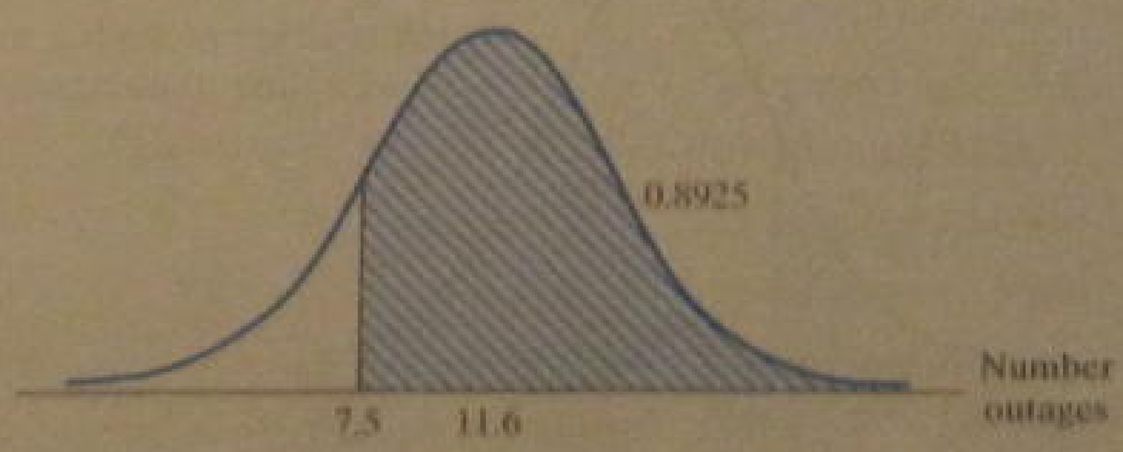


Figure 5.15
Diagram for example dealing with power outages

5.3 The Normal Approximation to the Binomial Distribution

The normal distribution can be used to approximate the binomial distribution when n is large and p, the probability of a success, is close to 0.50 and, hence, not small enough to use the Poisson approximation. Thus, let us state, without proof, the following theorem:

Normal approximation to binomial distribution

Theorem 5.1 If X is a random variable having the binomial distribution with the parameters n and p, the limiting form of the distribution function of the standardized random variable

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

as $n \to \infty$, is given by the standard normal distribution

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \qquad -\infty < z < \infty$$

Note that although X takes on only the values $0, 1, 2, \ldots, n$, in the limit as $n \to \infty$, the distribution of the corresponding standardized random variable is continuous, and the corresponding probability density is the standard normal density. Note that in the following example, which is an application of Theorem 5.1, we use again the continuity correction given on page 131.

EXAMPLE **A normal approximation to binomial probabilities**

If 20% of the memory chips made in a certain plant are defective, what are the probabilities that in a lot of 100 randomly chosen for inspection

(a) at most 15 will be defective; **(b)** exactly 15 will be defective?

Solution Since $\mu = 100\,(0.20) = 20$ and $\sigma = \sqrt{100\,(0.20)(0.80)} = 4$ for the binomial distribution with $n = 100$ and $p = 0.20$, we find that the normal approximation to the binomial distribution yields

(a)

$$F\left(\frac{15.5 - 20}{4}\right) = F(-1.13)$$
$$= 0.1292$$

as indicated in Figure 5.16.

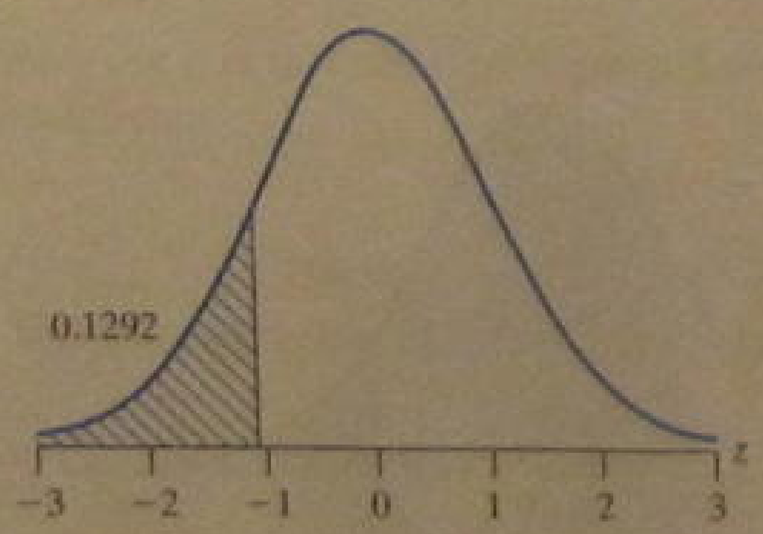


Figure 5.16 Normal approximation to $P(X \le 15)$

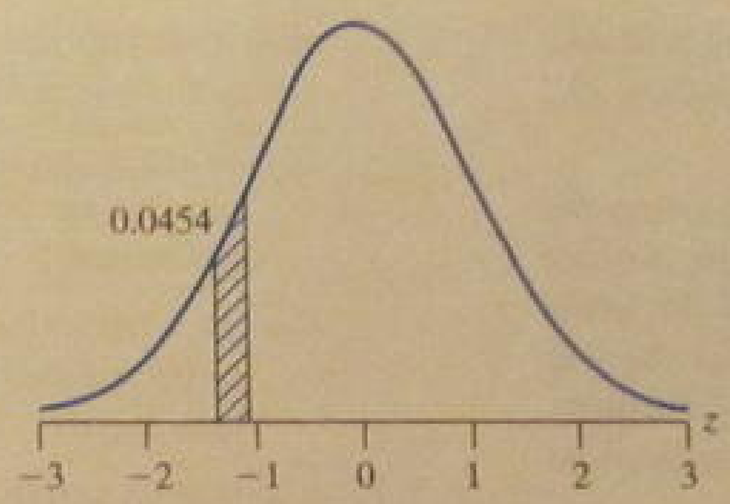


Figure 5.17
Normal approximation to $P(X = 15)$

(b)

$$F\left(\frac{15.5-20}{4}\right) - F\left(\frac{14.5-20}{4}\right) = F(-1.13) - F(-1.38)$$
$$= 0.1292 - 0.0838$$
$$= 0.0454$$

as indicated in Figure 5.17. ■

Had we done the exact binomial calculation on a computer instead of using normal approximation in the preceding example, we would have obtained 0.1285 instead of 0.1292 for part (a) and 0.0481 instead of 0.0454 for part (b). Thus, it can be seen that both approximations are very close. The exact calculation is always preferrable when p is given but the approximation is important for inference when it is not.

A good rule of thumb for the normal approximation

> Use the normal approximation to the binomial distribution only when np and $n(1-p)$ are both greater than 15.

Exercises

5.19 If a random variable has the standard normal distribution, find the probability that it will take on a value

(a) less than 1.75;

(b) less than −1.25;

(c) greater than 2.06;

(d) greater than −1.82.

5.20 If a random variable has the standard normal distribution, find the probability that it will take on a value

(a) between 0 and 2.3;

(b) between 1.22 and 2.43;

(c) between −1.45 and −0.45;

(d) between −1.70 and 1.35.

5.21 Find z if the probability that a random variable having the standard normal distribution will take on a value

(a) less than z is 0.9911;

(b) greater than z is 0.1093;

(c) greater than z is 0.6443;

(d) less than z is 0.0217;

(e) between $-z$ and z is 0.9298.

5.22 If a random variable has a normal distribution, what are the probabilities that it will take on a value within

(a) 1 standard deviation of the mean;

(b) 2 standard deviations of the mean;

(c) 3 standard deviations of the mean;

(d) 4 standard deviations of the mean?

5.23 Verify that

(a) $z_{0.005} = 2.575$;

(b) $z_{0.025} = 1.96$.

5.24 Given a random variable having the normal distribution with $\mu = 16.2$ and $\sigma^2 = 1.5625$, find the probabilities that it will take on a value

(a) greater than 16.8;

(b) less than 14.9;

(c) between 13.6 and 18.8;

(d) between 16.5 and 16.7.

5.25 The time for a super glue to set can be treated as a random variable having a normal distribution with mean

30 seconds. Find its standard deviation if the probability is 0.20 that it will take on a value greater than 39.2 seconds.

5.26 The time to microwave a bag of popcorn using the automatic setting can be treated as a random variable having a normal distribution with standard deviation 10 seconds. If the probability is 0.8212 that the bag will take less than 282.5 seconds to pop, find the probability that it will take longer than 258.3 seconds to pop.

5.27 The time required to assemble a piece of machinery is a random variable having approximately a normal distribution with $\mu = 12.9$ minutes and $\sigma = 2.0$ minutes. What are the probabilities that the assembly of a piece of machinery of this kind will take

(a) at least 11.5 minutes;

(b) anywhere from 11.0 to 14.8 minutes?

5.28 Find the *quartiles*

$$-z_{0.25} \quad z_{0.50} \quad z_{0.25}$$

of the standard normal distribution.

5.29 In a photographic process, the time to process 8×10 prints from a memory card may be looked upon as a random variable having the normal distribution with a mean of 10.28 seconds and a standard deviation of 0.12 second. Find the probability that it will take

(a) anywhere from 10.00 to 10.50 seconds to process one of the prints;

(b) at least 10.20 seconds to process one of the prints;

(c) at most 10.35 seconds to process one of the prints.

5.30 With reference to the preceding exercise, for which value is the probability 0.95 that it will be exceeded by the time it takes to process one of the prints?

5.31 Specifications for a certain job call for washers with an inside diameter of 0.300 ± 0.005 inch. If the inside diameters of the washers supplied by a given manufacturer may be looked upon as a random variable having the normal distribution with $\mu = 0.302$ inch and $\sigma = 0.003$ inch, what percentage of these washers will meet specifications?

5.32 With reference to the example on page 130, verify that if the variability of the filling machine is reduced to $\sigma = 0.025$ ounce, this will lower the required average amount of coffee to 4.05 ounces, yet keep 98% of the jars above 4 ounces.

5.33 A stamping machine produces can tops whose diameters are normally distributed with a standard deviation of 0.01 inch. At what "normal" (mean) diameter should the machine be set so that no more than 5% of the can tops produced have diameters exceeding 3 inches?

5.34 Extruded plastic rods are automatically cut into nominal lengths of 6 inches. Actual lengths are normally distributed about a mean of 6 inches and their standard deviation is 0.06 inch.

(a) What proportion of the rods have lengths that are outside the tolerance limits of 5.9 inches to 6.1 inches?

(b) To what value does the standard deviation need to be reduced if 99% of the rods must be within tolerance?

5.35 If a random variable has the binomial distribution with $n = 40$ and $p = 0.40$, use the normal approximation to determine the probabilities that it will take on

(a) the value 22;

(b) a value less than 8.

5.36 A manufacturer knows that, on average, 2% of the electric toasters that he makes will require repairs within 90 days after they are sold. Use the normal approximation to the binomial distribution to determine the probability that among 1,200 of these toasters at least 30 will require repairs within the first 90 days after they are sold.

5.37 The probability that an electronic component will fail in less than 1,000 hours of continuous use is 0.25. Use the normal approximation to find the probability that among 200 such components fewer than 45 will fail in less than 1,000 hours of continuous use.

5.38 A safety engineer feels that 30% of all industrial accidents in her plant are caused by failure of employees to follow instructions. If this figure is correct, find, approximately, the probability that among 84 industrialized accidents in this plant anywhere from 20 to 30 (inclusive) will be due to failure of employees to follow instructions.

5.39 If 62% of all clouds seeded with silver iodide show spectacular growth, what is the probability that among 40 clouds seeded with silver iodide at most 20 will show spectacular growth?

5.40 To illustrate the law of large numbers mentioned on page 102, find the probabilities that the proportion of heads will be anywhere from 0.49 to 0.51 when a balanced coin is flipped

(a) 1,000 times;

(b) 10,000 times.

5.41 Verify the identity $F(-z) = 1 - F(z)$ given on page 126.

5.42 Verify that the parameter μ in the expression for the normal density on page 125, is, in fact, its mean.

5.43 Verify that the parameter σ^2 in the expression for the normal density on page 125 is, in fact, its variance.

5.44 Normal probabilities can be calculated using *MINITAB*. Let X have a normal distribution with mean 11.3 and standard deviation 5.7. The following steps yield the cumulative probability of 9.31 or smaller, or $P(X \leq 9.31)$.

Dialog box:
Calc > **Probability Distribution** > **Normal**
Choose **Cumulative Distribution.** Choose **Input constant** and enter 9.31.
Type 11.3 in **Mean** and 5.7 in **standard deviation.**
Click **OK**.

Output: Normal with mean = 11.3000 and standard deviation = 5.70000

x	$P(X <= x)$
9.3100	0.3635

For this same normal distribution, find the probability

(a) of 8.493 or smaller;

(b) of 16.074 or smaller.

5.4 Other Probability Densities

In the application of statistics to problems in engineering and physical science, we shall encounter many probability densities other than the normal distribution. Among these are the t, F, and chi square distributions; the fundamental sampling distributions that we introduce in Chapter 6; and the exponential and Weibull distributions, which we apply to problems of reliability and life testing in Chapter 16.

In the remainder of this chapter we shall discuss five continuous distributions, the *uniform distribution*, the *log-normal distribution*, the *gamma distribution*, the *beta distribution*, and the *Weibull distribution*, for the twofold purpose of giving further examples of probability densities and of laying the foundation for future applications.

5.5 The Uniform Distribution

The **uniform distribution**, with the parameters α and β, has probability density function

Uniform distribution

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{for } \alpha < x < \beta \\ 0 & \text{elsewhere} \end{cases}$$

whose graph is shown in Figure 5.18. Note that all values of x from α to β are "equally likely" in the sense that the probability that x lies in an interval of width Δx entirely contained in the interval from α to β is equal to $\Delta x/(\beta - \alpha)$, regardless of the exact location of the interval.

To illustrate how a physical situation might give rise to a uniform distribution, suppose that a wheel of a locomotive has the radius r and that x is the location of a point on its circumference measured along the circumference from some reference point 0. When the brakes are applied, some point will make sliding contact with the

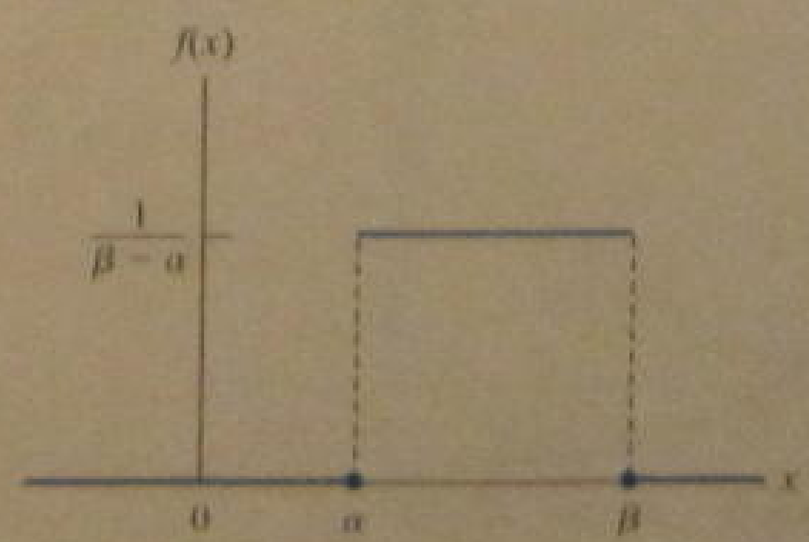


Figure 5.18
Graph of uniform probability density

rail, and heavy wear will occur at that point. For repeated application of the brakes, it would seem reasonable to assume that x is a value of a random variable having the uniform distribution with $\alpha = 0$ and $\beta = 2\pi r$. If this assumption were incorrect, that is, if some set of points on the wheel made contact more often than others, the wheel would eventually exhibit "flat spots" or wear out of round.

To determine the mean and the variance of the uniform distribution, we first evaluate the two integrals

$$\mu = \int_{\alpha}^{\beta} x \cdot \frac{1}{\beta - \alpha}\, dx = \frac{\alpha + \beta}{2}$$

and

$$\mu_2' = \int_{\alpha}^{\beta} x^2 \cdot \frac{1}{\beta - \alpha}\, dx = \frac{\alpha^2 + \alpha\beta + \beta^2}{3}$$

Thus,

Mean of uniform distribution

$$\mu = \frac{\alpha + \beta}{2}$$

and, making use of the formula $\sigma^2 = \mu_2' - \mu^2$, we find that

Variance of uniform distribution

$$\sigma^2 = \frac{1}{12}(\beta - \alpha)^2$$

5.6 The Log-Normal Distribution

The **log-normal distribution** occurs in practice whenever we encounter a random variable which is such that its logarithm has a normal distribution. Its probability density is given by

Log-normal distribution

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\,\beta} x^{-1} e^{-(\ln x - \alpha)^2/2\beta^2} & \text{for } x > 0,\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

where $\ln x$ is the natural logarithm of x. A graph of the log-normal distribution with $\alpha = 0$ and $\beta = 1$ is shown in Figure 5.19. It can be seen from the figure that this distribution is positively skewed, that is, it has a long right-hand tail.

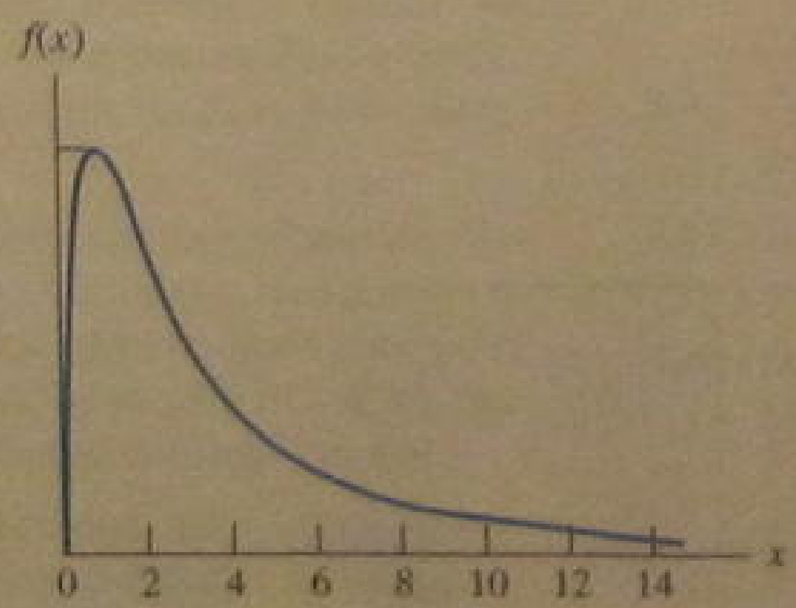


Figure 5.19 Graph of log-normal probability density, $\alpha = 0$, $\beta = 1$.

To find the probability that a random variable having the log-normal distribution will take on a value between a and b $(0 < a < b)$, we must evaluate the integral

$$\int_a^b \frac{1}{\sqrt{2\pi}\beta} x^{-1} e^{-(\ln x - \alpha)^2/2\beta^2}\, dx$$

Changing variable by letting $y = \ln x$ and identifying the integrand as the normal density with $\mu = \alpha$ and $\sigma = \beta$, we find that the desired probability is given by

$$\int_{\ln a}^{\ln b} \frac{1}{\sqrt{2\pi}\beta} e^{-(y-\alpha)^2/2\beta^2}\, dy = F\left(\frac{\ln b - \alpha}{\beta}\right) - F\left(\frac{\ln a - \alpha}{\beta}\right)$$

where F is the distribution function of the standard normal distribution.

EXAMPLE **Calculating a log-normal probability**

The current gain of certain transistors is measured in units which make it equal to the logarithm of I_o/I_i, the ratio of the output to the input current. If this logarithm is normally distributed with $\mu = 2$ and $\sigma^2 = 0.01$, find the probability that I_o/I_i will take on a value between 6.1 and 8.2.

Solution Since $\alpha = 2$ and $\beta = 0.1$, we get

$$F\left(\frac{\ln 8.2 - 2}{0.1}\right) - F\left(\frac{\ln 6.1 - 2}{0.1}\right) = F(1.0) - F(-1.92)$$
$$= 0.8139$$

■

EXAMPLE **Graphing a probability density function on top of a density histogram to help assess fit**

Make a density histogram of the interrequest times on page 19 and relate it to a log-normal distribution.

Solution Figure 5.20 gives the density histogram. To accurately portray the pattern, shorter intervals are used for the smaller times. We have also plotted the log-normal density with $\alpha = 8.85$ and $\beta = 1.03$. The log-normal fit is explored further in Section 5.12. (See also Exercise 5.103.) ■

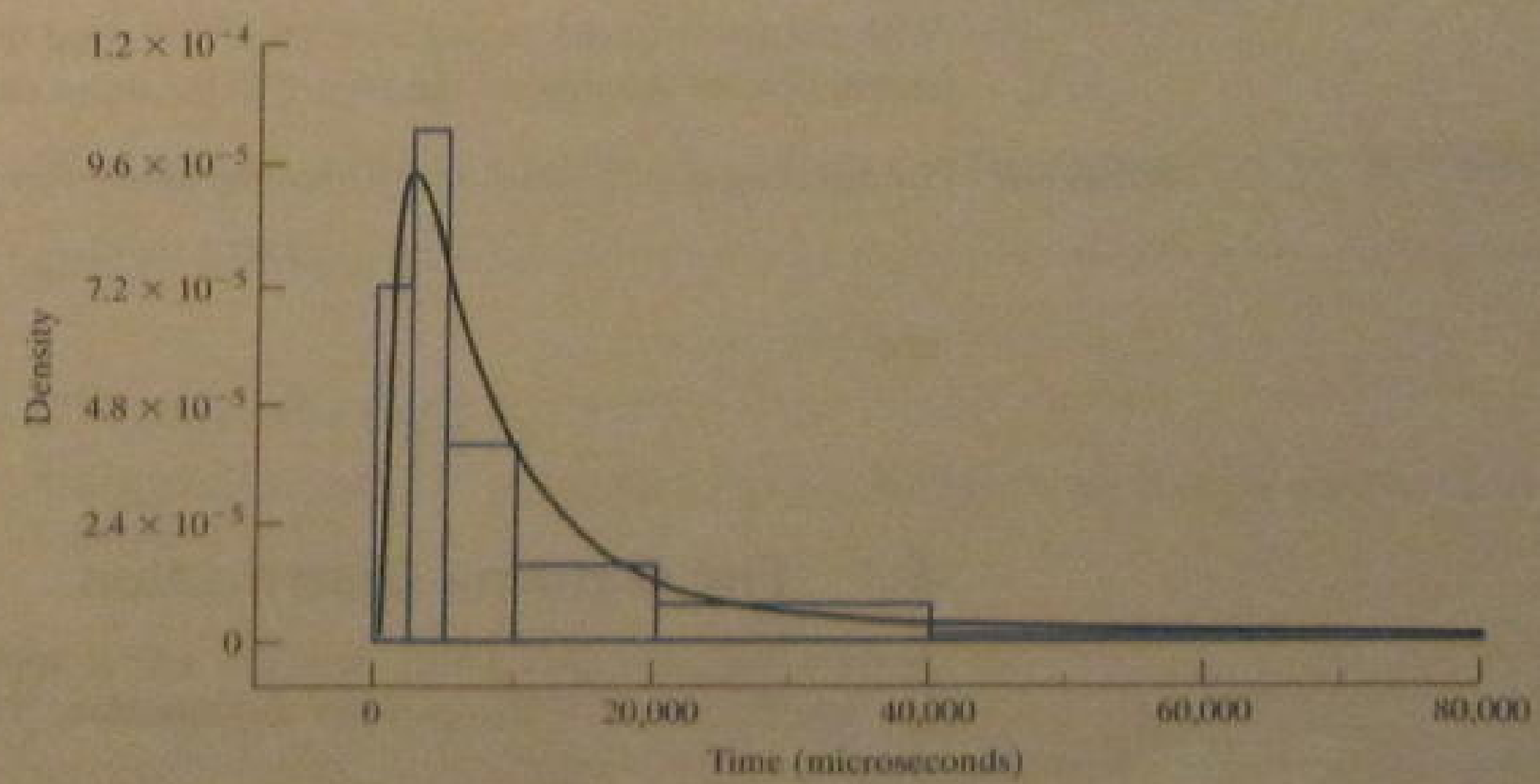


Figure 5.20
Density histogram of interrequest time

EXAMPLE **A log-normal probability calculation for a risk analysis**

As part of a risk analysis concerning a nuclear power plant, engineers must model the strength of steam generator supports in terms of their ability to withstand the peak acceleration caused by earthquakes. Expert opinion suggests that the ln (strength) is normally distributed with $\mu = 4.0$ and $\sigma^2 = 0.09$. Find the probability that the supports will survive a peak acceleration of 33.

Solution Since $\alpha = 4.0$ and $\beta = 0.30$, we find

$$1 - F\left(\frac{\ln(33) - 4.0}{0.30}\right) = 1 - F(-1.68) = 0.9535$$

■

To find a formula for the mean of the log-normal distribution, we write

$$\mu = \frac{1}{\sqrt{2\pi}\beta}\int_0^\infty x \cdot x^{-1} e^{-(\ln x - \alpha)^2/2\beta^2}\, dx$$

and upon letting $y = \ln x$, this becomes

$$\mu = \frac{1}{\sqrt{2\pi}\beta}\int_{-\infty}^\infty e^y e^{-(y-\alpha)^2/2\beta^2}\, dy$$

This integral can be evaluated by completing the square on the exponent $y - (y - \alpha)^2/2\beta^2$, thus obtaining an integrand which has the form of a normal density. The final result, which the reader will be asked to verify in Exercise 5.48, is

Mean of log-normal distribution

$$\mu = e^{\alpha + \beta^2/2}$$

Similar, but more lengthy, calculations yield

Variance of log-normal distribution

$$\sigma^2 = e^{2\alpha+\beta^2}(e^{\beta^2} - 1)$$

EXAMPLE **Calculating the mean and variance for a log-normal distribution**

With reference to the example on page 137, find the mean and the variance of the distribution of the ratio of the output to the input current.

Solution Substituting $\alpha = 2$ and $\beta = 0.1$ into the above formulas, we get

$$\mu = e^{2+(0.1)^2/2} = 7.43$$

and

$$\sigma^2 = e^{4+(0.1)^2}(e^{(0.1)^2} - 1) = 0.55$$

■

5.7 The Gamma Distribution

Several important probability densities whose applications will be discussed later are special cases of the **gamma distribution**. This distribution has probability density

Gamma distribution

$$f(x) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{for } x > 0,\ \alpha > 0,\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

where $\Gamma(\alpha)$ is a value of the **gamma function**, defined by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\, dx$$

Integration by parts shows that

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

for any $\alpha > 1$ and, hence, that $\Gamma(\alpha) = (\alpha - 1)!$ when α is a positive integer. Graphs of several gamma distributions are shown in Figure 5.21 and they exhibit the fact that these distributions are positively skewed. In fact, the skewness decreases as α increases for any fixed value of β.

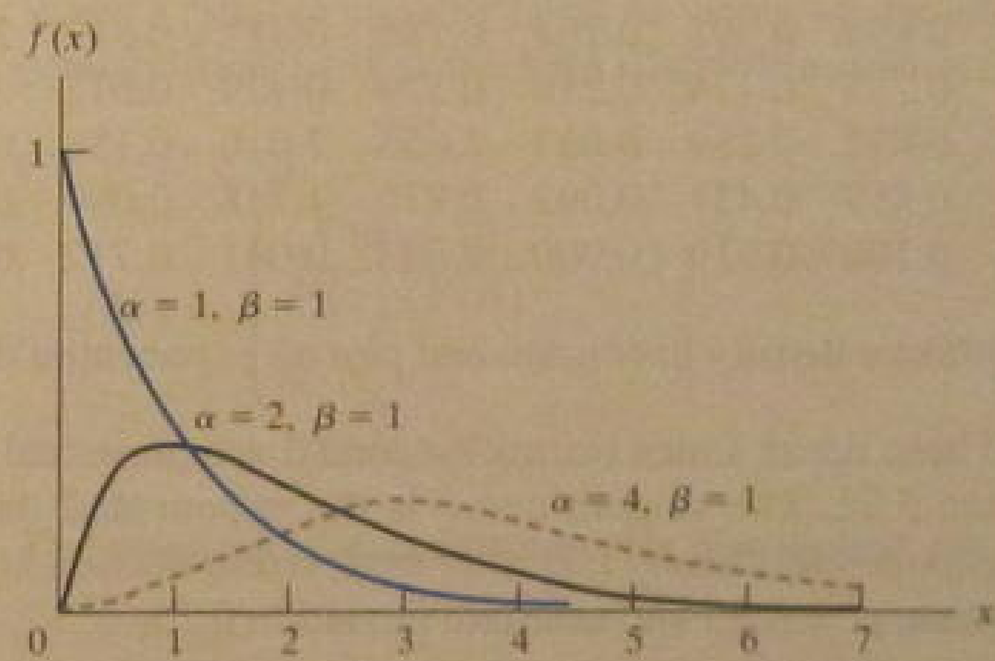


Figure 5.21
Graph of some gamma probability density functions

The mean and the variance of the gamma distribution may be obtained by making use of the gamma function and its special properties mentioned above. For the mean we have

$$\mu = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} x \cdot x^{\alpha-1} e^{-x/\beta}\, dx$$

and, after letting $y = x/\beta$, we get

$$\mu = \frac{\beta}{\Gamma(\alpha)} \int_0^{\infty} y^{\alpha} e^{-y}\, dy = \frac{\beta\Gamma(\alpha+1)}{\Gamma(\alpha)}$$

Then, making use of the identity $\Gamma(\alpha + 1) = \alpha \cdot \Gamma(\alpha)$, we arrive at the result that

Mean of gamma distribution

$$\mu = \alpha\beta$$

Using similar methods, it can also be shown that the variance of the gamma distribution is given by

Variance of gamma distribution

$$\sigma^2 = \alpha\beta^2$$

In the special case where $\alpha = 1$, we get the **exponential distribution**, whose probability density is thus

Exponential distribution

$$f(x) = \begin{cases} \dfrac{1}{\beta} e^{-x/\beta} & \text{for } x > 0,\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and whose mean and variance are $\mu = \beta$ and $\sigma^2 = \beta^2$. Note that the distribution of the example on page 122 is an exponential distribution with $\beta = \frac{1}{2}$.

EXAMPLE **An exponential density function on top of a density histogram**

An engineer observing a nuclear reaction measures the time intervals between the emissions of beta particles.

0.894	0.991	0.261	0.186	0.311	0.817	2.267	0.091	0.139	0.083
0.235	0.424	0.216	0.579	0.429	0.612	0.143	0.055	0.752	0.188
0.071	0.159	0.082	1.653	2.010	0.158	0.527	1.033	2.863	0.365
0.459	0.431	0.092	0.830	1.718	0.099	0.162	0.076	0.107	0.278
0.100	0.919	0.900	0.093	0.041	0.712	0.994	0.149	0.866	0.054

Make a density histogram and plot an exponential density as an approximation.

Solution These decay times (in milliseconds) are presented as a density histogram in Figure 5.22. The smooth curve is the exponential density with $\beta = 0.25$. Fit to an exponential density is further explored in Section 16.3. The exponential density has mean $\beta = 0.25$ and standard deviation 0.25. ■

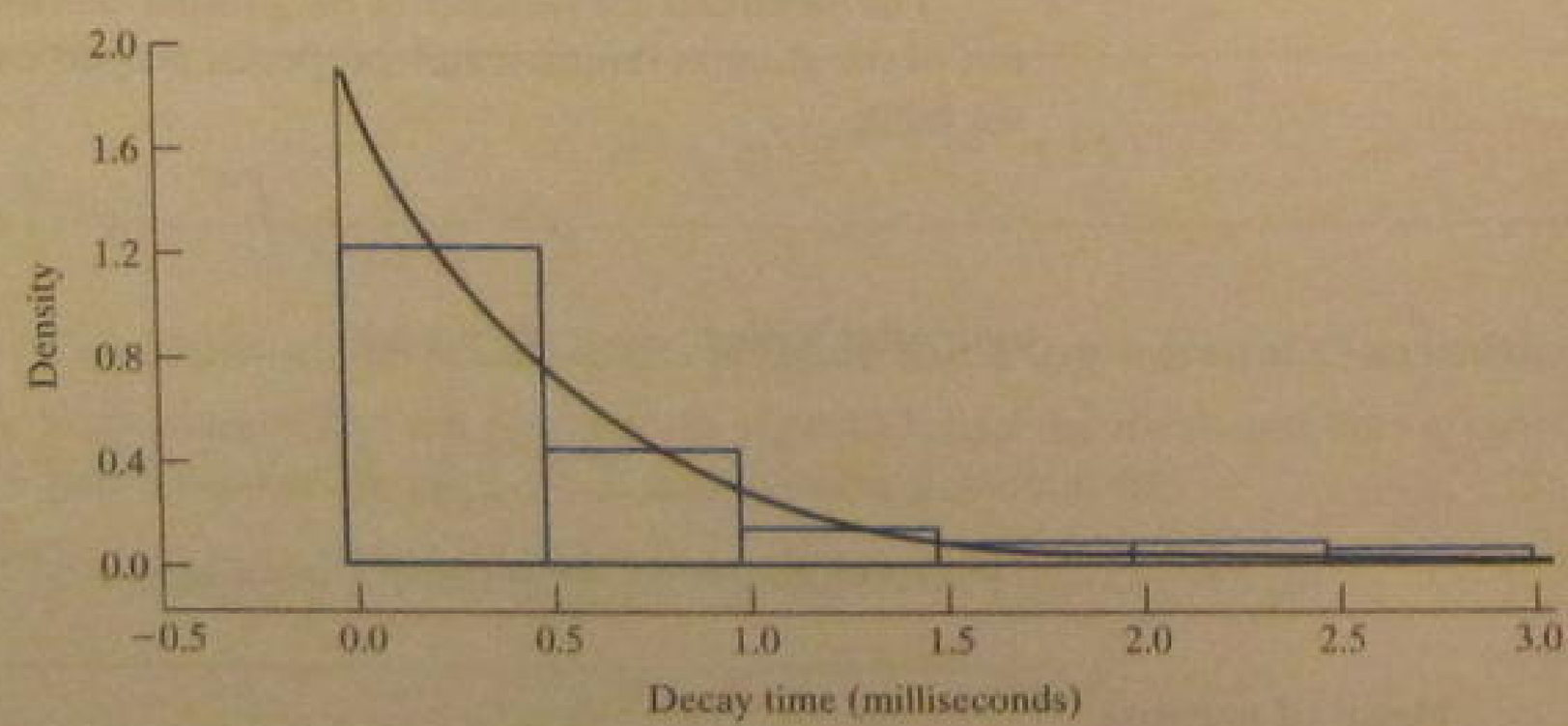


Figure 5.22
Density histogram of decay times

The exponential distribution has many important applications; for instance, it can be shown that in connection with Poisson processes (see Section 4.7) the **waiting time** between successive arrivals (successes) has an exponential distribution. More specifically, it can be shown that if in a Poisson process the mean arrival rate (average number of arrivals per unit time) is α, the time until the first arrival, or the waiting time between successive arrivals, has an exponential distribution with $\beta = \dfrac{1}{\alpha}$ (see Exercise 5.62).

EXAMPLE **Probability calculations using the exponential distribution**

With reference to the example on page 108, where on the average three trucks arrived per hour to be unloaded at a warehouse, what are the probabilities that the time between the arrival of successive trucks will be

(a) less than 5 minutes? (b) at least 45 minutes?

Solution Assuming the arrivals follow a Poisson process with $\alpha = 3$, then $\beta = \frac{1}{3}$ and we get

(a)
$$\int_0^{1/12} 3e^{-3x}\,dx = 1 - e^{-1/4} = 0.221$$

(b)
$$\int_{3/4}^{\infty} 3e^{-3x}\,dx = e^{-9/4} = 0.105$$

■

5.8 The Beta Distribution

When a random variable takes on values on the interval from 0 to 1, one choice of a probability density is the **beta distribution** whose probability density is

Beta distribution

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1, \alpha > 0, \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

The mean and the variance of this distribution are given by

Mean and variance of beta distribution

$$\mu = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Note that for $\alpha = 1$ and $\beta = 1$ we obtain as a special case the uniform distribution of Section 5.5 defined on the interval from 0 to 1. The following example, pertaining to a proportion, illustrates a typical application of the beta distribution.

EXAMPLE **Probability calculations using a beta distribution**

In a certain county, the proportion of highway sections requiring repairs in any given year is a random variable having the beta distribution with $\alpha = 3$ and $\beta = 2$ (shown in Figure 5.23, on page 142).

(a) On the average, what percentage of the highway sections require repairs in any given year?

(b) Find the probability that at most half of the highway sections will require repairs in any given year.

Solution (a) $\mu = \dfrac{3}{3+2} = 0.60,$

which means that on the average 60% of the highway sections require repairs in any given year.

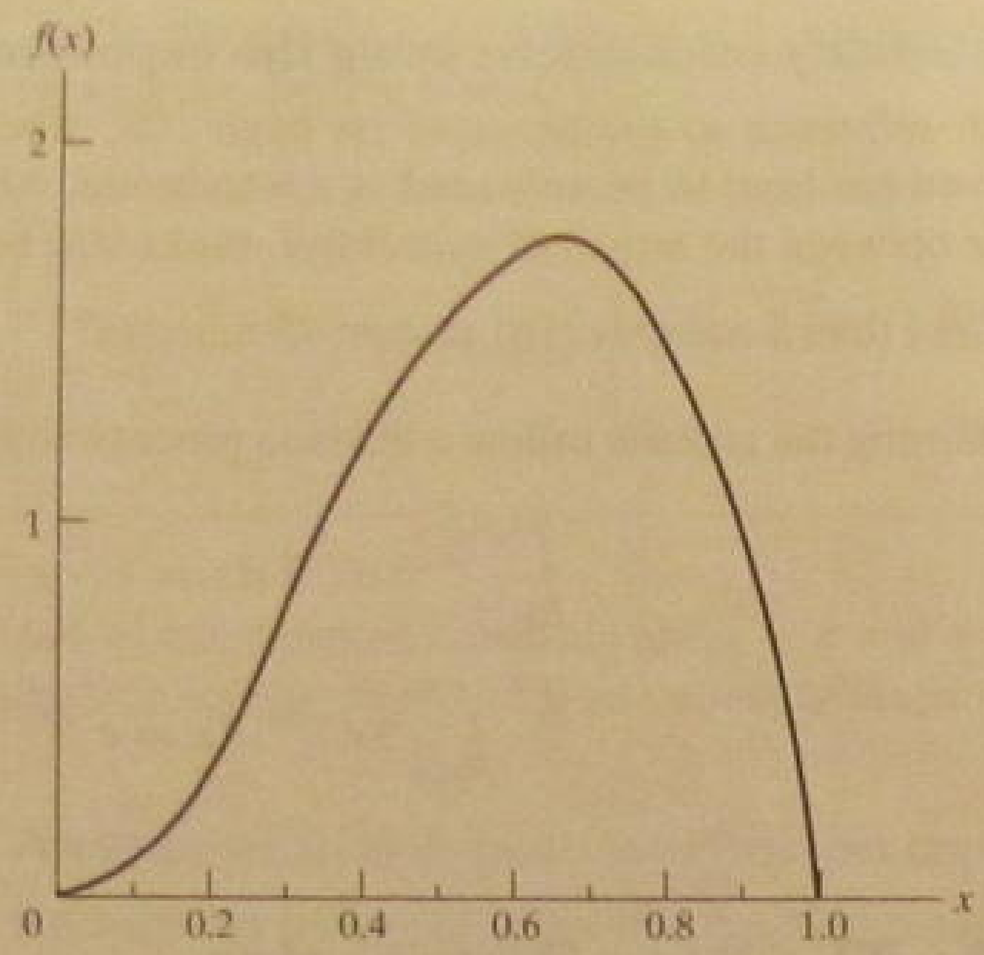


Figure 5.23
Graph of the beta density with $\alpha = 3$ and $\beta = 2$

(b) Substituting $\alpha = 3$ and $\beta = 2$ into the formula for the beta distribution and making use of the fact that $\Gamma(5) = 4! = 24$, $\Gamma(3) = 2! = 2$, and $\Gamma(2) = 1! = 1$, we get

$$f(x) = \begin{cases} 12x^2(1-x) & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Thus, the desired probability is given by

$$\int_0^{1/2} 12x^2(1-x)\,dx = \frac{5}{16}$$

■

In most realistically complex situations, probabilities relating to gamma and beta distributions are obtained from computer programs.

5.9 The Weibull Distribution

Closely related to the exponential distribution is the **Weibull distribution**, whose probability density is given by

Weibull distribution

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^\beta} & \text{for } x > 0,\ \alpha > 0,\ \beta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

To demonstrate this relationship we evaluate the probability that a random variable having the Weibull distribution will take on a value less than a, namely, the integral

$$\int_0^a \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}\,dx$$

Making the change of variable $y = x^\beta$, we get

$$\int_0^{a^\beta} \alpha e^{-\alpha y}\,dy = 1 - e^{-\alpha a^\beta}$$

and it can be seen that y is a value of a random variable having an exponential distribution. The graphs of several Weibull distributions with $\alpha = 1$ and $\beta = \frac{1}{2}$, 1, and 2 are shown in Figure 5.24.

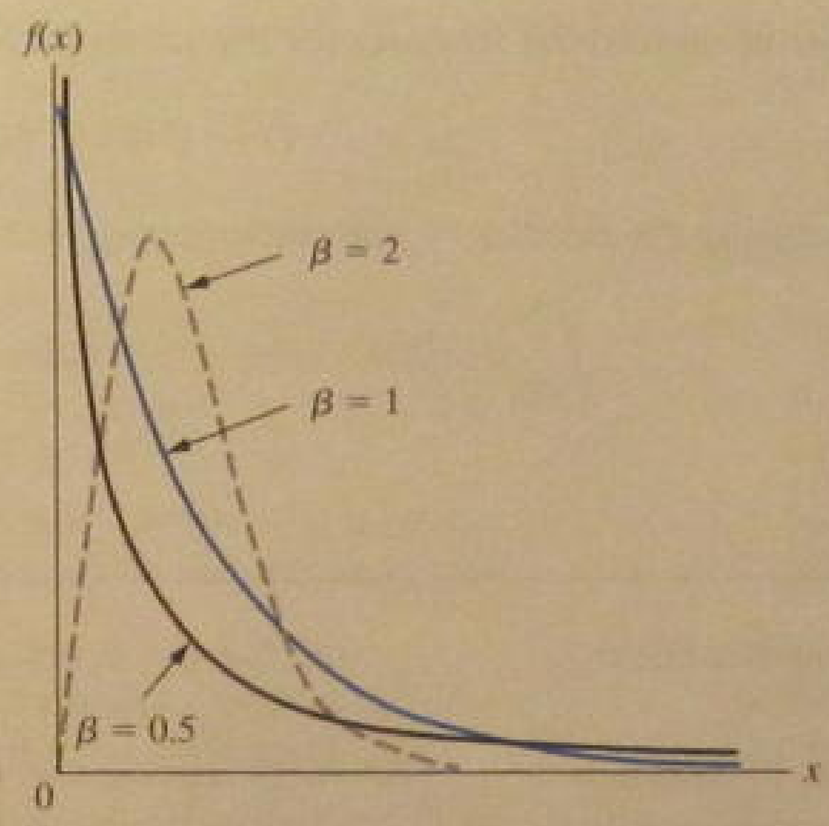


Figure 5.24
Graphs of Weibull densities with $\alpha = 1$ and $\beta = \frac{1}{2}$, 1, and 2

The mean of the Weibull distribution having the parameters α and β may be obtained by evaluating the integral

$$\mu = \int_0^\infty x \cdot \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}\, dx$$

Making the change of variable $u = \alpha x^\beta$, we get

$$\mu = \alpha^{-1/\beta} \int_0^\infty u^{1/\beta} e^{-u}\, du$$

and recognizing the integral as

$$\Gamma\left(1 + \frac{1}{\beta}\right)$$

namely, as a value of the gamma function which we defined on page 139, we find that the mean of the Weibull distribution is given by

Mean of Weibull distribution

$$\mu = \alpha^{-1/\beta}\, \Gamma\left(1 + \frac{1}{\beta}\right)$$

Using a similar method to determine first μ_2', the reader will be asked to show in Exercise 5.70 that the variance of this distribution is given by

Variance of Weibull distribution

$$\sigma^2 = \alpha^{-2/\beta} \left\{ \Gamma\left(1 + \frac{2}{\beta}\right) - \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2 \right\}$$

EXAMPLE **Probability calculations using a Weibull distribution**

Suppose that the lifetime of a certain kind of an emergency backup battery (in hours) is a random variable X having the Weibull distribution with $\alpha = 0.1$ and $\beta = 0.5$. Find

(a) the mean lifetime of these batteries;

(b) the probability that such a battery will last more than 300 hours.

Solution **(a)** Substitution into the formula for the mean yields

$$\mu = (0.1)^{-2}\,\Gamma(3) = 200 \text{ hours}$$

(b) Performing the necessary integration, we get

$$\int_{300}^{\infty} (0.05)x^{-0.5}\,e^{-0.1x^{0.5}}\,dx = e^{-0.1(300)^{0.5}}$$
$$= 0.177$$

■

Exercises

5.45 Find the distribution function of a random variable having a uniform distribution on (0, 1).

5.46 In certain experiments, the error made in determining the solubility of a substance is a random variable having the uniform density with $\alpha = -0.025$ and $\beta = 0.025$. What are the probabilities that such an error will be

(a) between 0.010 and 0.015;

(b) between −0.012 and 0.012?

5.47 From experience Mr. Harris has found that the low bid on a construction job can be regarded as a random variable having the uniform density

$$f(x) = \begin{cases} \dfrac{3}{4C} & \text{for } \dfrac{2C}{3} < x < 2C \\ 0 & \text{elsewhere} \end{cases}$$

where C is his own estimate of the cost of the job. What percentage should Mr. Harris add to his cost estimate when submitting bids to maximize his expected profit?

5.48 Verify the expression given on page 138 for the mean of the log-normal distribution.

5.49 With reference to the example on page 137, find the probability that I_o/I_i will take on a value between 7.0 and 7.5.

5.50 If a random variable has the log-normal distribution with $\alpha = -1$ and $\beta = 2$, find its mean and its standard deviation.

5.51 With reference to the preceding exercise, find the probabilities that the random variable will take on a value

(a) between 3.2 and 8.4;

(b) greater than 5.0.

5.52 If a random variable has the gamma distribution with $\alpha = 2$ and $\beta = 3$, find the mean and the standard deviation of this distribution.

5.53 With reference to Exercise 5.52, find the probability that the random variable will take on a value less than 5.

5.54 In a certain city, the daily consumption of electric power (in millions of kilowatt-hours) can be treated as a random variable having a gamma distribution with $\alpha = 3$ and $\beta = 2$. If the power plant of this city has a daily capacity of 12 million kilowatt-hours, what is the probability that this power supply will be inadequate on any given day?

5.55 With reference to the example on page 138, suppose the expert opinion is in error. Calculate the probability that the supports will survive if

(a) $\mu = 3.0$ and $\sigma^2 = 0.09$;

(b) $\mu = 4.0$ and $\sigma^2 = 0.25$.

5.56 Verify the expression for the variance of the gamma distribution given on page 139.

5.57 Show that when $\alpha > 1$, the graph of the gamma density has a relative maximum at $x = \beta(\alpha - 1)$. What happens when $0 < \alpha < 1$ and when $\alpha = 1$?

5.58 The amount of time that a surveillance camera will run without having to be reset is a random variable having the exponential distribution with $\beta = 50$ days. Find the probabilities that such a camera will

(a) have to be reset in less than 20 days;

(b) not have to be reset in at least 60 days.

5.59 With reference to Exercise 4.95, find the percent of the time that the interval between breakdowns of the computer will be

(a) less than 1 week;

(b) at least 5 weeks.

5.60 With reference to Exercise 4.58, find the probabilities that the time between successive requests for consulting will be

(a) less than 0.5 week;

(b) more than 3 weeks.

5.61 Given a Poisson process with on the average α arrivals per unit time, find the probability that there will be no arrivals during a time interval of length t, namely, the probability that the waiting times between successive arrivals will be at least of length t.

5.62 Use the result of Exercise 5.61 to find an expression for the probability density of the waiting time between successive arrivals.

5.63 Verify for $\alpha = 3$ and $\beta = 3$ that the integral of the beta density, from 0 to 1, is equal to 1.

5.64 If the annual proportion of erroneous income tax returns filed with the IRS can be looked upon as a random variable having a beta distribution with $\alpha = 2$ and $\beta = 9$, what is the probability that in any given year there will be fewer than 10% erroneous returns?

5.65 Suppose that the proportion of defectives shipped by a vendor, which varies somewhat from shipment to shipment, may be looked upon as a random variable having the beta distribution with $\alpha = 1$ and $\beta = 4$.

(a) Find the mean of this beta distribution, namely, the average proportion of defectives in a shipment from this vendor.

(b) Find the probability that a shipment from this vendor will contain 25% or more defectives.

5.66 Show that when $\alpha > 1$ and $\beta > 1$, the beta density has a relative maximum at

$$x = \frac{\alpha - 1}{\alpha + \beta - 2}$$

5.67 With reference to the example on page 143, find the probability that such a battery will not last 100 hours.

5.68 Suppose that the time to failure (in minutes) of certain electronic components subjected to continuous vibrations may be looked upon as a random variable having the Weibull distribution with $\alpha = \frac{1}{5}$ and $\beta = \frac{1}{3}$.

(a) How long can such a component be expected to last?

(b) What is the probability that such a component will fail in less than 5 hours?

5.69 Suppose that the service life (in hours) of a semiconductor device is a random variable having the Weibull distribution with $\alpha = 0.025$ and $\beta = 0.500$. What is the probability that such a device will still be in operating condition after 4,000 hours?

5.70 Verify the formula for the variance of the Weibull distribution given on page 143.

5.10 Joint Distributions—Discrete and Continuous

Discrete Variables

Often, experiments are conducted where two random variables are observed simultaneously in order to determine not only their individual behavior but also the degree of relationship between them.

For two discrete random variables X_1 and X_2, we write the probability that X_1 will take the value x_1 and X_2 will take the value x_2 as $P(X_1 = x_1, X_2 = x_2)$. Consequently, $P(X_1 = x_1, X_2 = x_2)$ is the probability of the intersection of the events $X_1 = x_1$ and $X_2 = x_2$. The distribution of probability is specified by listing the probabilities associated with all possible pairs of values x_1 and x_2, either by formula or in a table. We refer to the function $f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$ and the corresponding possible values (x_1, x_2) as the **joint probability distribution** of X_1 and X_2.

EXAMPLE **Calculating probabilities from a discrete joint probability distribution**

Let X_1 and X_2 have the joint probability distribution in the table below.

Joint Probability Distribution $f(x_1, x_2)$ of X_1 and X_2

		x_1 = 0	1	2
x_2	0	0.1	0.4	0.1
	1	0.2	0.2	0

(a) Find $P(X_1 + X_2 > 1)$.

(b) Find the probability distribution $f_1(x_1) = P(X_1 = x_1)$ of the individual random variable X_1.

Solution **(a)** The event $X_1 + X_2 > 1$ is composed of the pairs of values (1, 1), (2, 0), and (2,1). Adding their corresponding probabilities

$$P(X_1 + X_2 > 1) = f(1, 1) + f(2, 0) + f(2, 1) = 0.2 + 0.1 + 0 = 0.3$$

(b) Since the event $X_1 = 0$ is composed of the two pairs of values (0, 0) and (0, 1), we add their corresponding probabilities to obtain

$$P(X_1 = 0) = f(0, 0) + f(0, 1) = 0.1 + 0.2 = 0.3$$

Continuing, we obtain $P(X_1 = 1) = 0.6$ and $P(X_1 = 2) = 0.1$. In summary, $f_1(0) = 0.3$, $f_1(1) = 0.6$, and $f_1(2) = 0.1$ is the probability distribution of X_1.

Rewriting the frequency table but including the row and column totals,

Joint Probability Distribution $f(x_1, x_2)$ of X_1 and X_2 with Marginal Distributions

		x_1 0	1	2	Total $f_2(x_2)$
x_2	0	0.1	0.4	0.1	0.6
	1	0.2	0.2	0	0.4
Total	$f_1(x_1)$	0.3	0.6	0.1	1.0

Note that the probability distribution $f_1(x_1)$ of X_1 appears in the lower margin of this enlarged table. The probability distribution $f_2(x_2)$ of X_2 appears in the right-hand margin of the table. Consequently, the individual distributions are called **marginal probability distributions**. ■

From the example, we see that for each fixed value x_1 of X_1, the marginal probability distribution is obtained as

$$P(X_1 = x_1) = f_1(x_1) = \sum_{\text{values } x_2} f(x_1, x_2)$$

where the sum is over all possible values of the second variable with x_1 fixed.

Consistent with the definition of conditional probability of events when A is the event $X_1 = x_1$ and B is the event $X_2 = x_2$, the **conditional probability distribution** of X_1 given $X_2 = x_2$ is defined as

$$f_1(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \qquad \text{for all } x_1 \text{ provided } f_2(x_2) \neq 0$$

If $f_1(x_1 \mid x_2) = f_1(x_1)$ for all x_1 and x_2, so the conditional probability distribution is free of x_2, or, equivalently, if

$$f(x_1, x_2) = f_1(x_1) f_2(x_2) \qquad \text{for all } x_1, x_2$$

the two random variables are **independent**.

EXAMPLE **A conditional probability distribution**

With reference to the previous example, find the conditional probability distribution of X_1 given $X_2 = 1$. Are X_1 and X_2 independent?

Solution $f_1(0 \mid 1) = \dfrac{f(0,1)}{f_2(1)} = \dfrac{0.2}{0.4} = 0.5, \quad f_1(1 \mid 1) = \dfrac{f(1,1)}{f_2(1)} = \dfrac{0.2}{0.4} = 0.5$, and

$$f_1(2 \mid 1) = \frac{f(2,1)}{f_2(1)} = \frac{0}{0.4} = 0$$

Since $f_1(0 \mid 1) = 0.5 \neq 0.3 = f_1(0)$, the conditional probability distribution is not free of the value x_2. Equivalently, $f(0,1) = 0.2 \neq (0.3)(0.4) = f_1(0)f_2(1)$ so X_1 and X_2 are *dependent*. ■

Suppose that instead we are concerned with k random variables $X_1, X_2, \ldots, X_k$. Let x_1 be a possible value for the first random variable X_1, x_2 be a possible value for the second random variable X_2, and so on with x_k a possible value for the kth random variable. Then the probabilities $P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = f(x_1, x_2, \ldots, x_k)$ need to be specified. We refer to the function f and the corresponding k-tuples of possible values $(x_1, x_2, \ldots, x_k)$ as the **joint probability distribution** of these discrete random variables.

The probability distribution $f_i(x_i)$ of the individual variable X_i is called the **marginal probability distribution** of the ith random variable

$$f_i(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_k} f(x_1, x_2, \ldots, x_k)$$

where the summation is over all possible k-tuples where the ith component is held fixed at the specified value x_i.

Continuous Variables

There are many situations in which we describe an outcome by giving the values of several continuous random variables. For instance, we may measure the weight and the hardness of a rock; the volume, pressure, and temperature of a gas; or the thickness, color, compressive strength, and potassium content of a piece of glass. If $X_1, X_2, \ldots, X_k$ are k continuous random variables, we shall refer to $f(x_1, x_2, \ldots, x_k)$ as the **joint probability density** of these random variables, if the probability that $a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \ldots,$ and $a_k \leq X_k \leq b_k$ is given by the multiple integral

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_k}^{b_k} f(x_1, x_2, \ldots, x_k)\, dx_1\, dx_2 \ldots dx_k$$

Thus, not every function $f(x_1, x_2, \ldots, x_k)$ can serve as a joint probability density, but if

$$f(x_1, x_2, \ldots, x_k) \geq 0$$

for all values of $x_1, x_2, \ldots, x_k$, and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_k)\, dx_1\, dx_2 \ldots dx_k = 1$$

it can be shown that the axioms of probability (with the modification of the definition of "event" discussed in Section 5.1) are satisfied.

To extend the concept of a cumulative distribution function to the k-variable case, we write as $F(x_1, x_2, \ldots, x_k)$ the probability that the first random variable will take on a value less than or equal to x_1, the second random variable will take on a value less than or equal to $x_2, \ldots,$ and the kth random variable will take on a

value less than or equal to x_k, and we refer to the corresponding function F as the **joint cumulative distribution function** of the k random variables.

EXAMPLE **Calculating probabilities from a joint probability density function**

If the joint probability density of two random variables is given by

$$f(x_1, x_2) = \begin{cases} 6e^{-2x_1-3x_2} & \text{for } x_1 > 0,\ x_2 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

find the probabilities that

(a) the first random variable will take on a value between 1 and 2 and the second random variable will take on a value between 2 and 3;

(b) the first random variable will take on a value less than 2 and the second random variable will take on a value greater than 2.

Solution Performing the necessary integrations, we get

(a)

$$\int_1^2 \int_2^3 6e^{-2x_1-3x_2}\,dx_1\,dx_2 = (e^{-2} - e^{-4})(e^{-6} - e^{-9}) = 0.0003$$

(b)

$$\int_0^2 \int_2^\infty 6e^{-2x_1-3x_2}\,dx_1\,dx_2 = (1 - e^{-4})e^{-6} = 0.0025$$

■

EXAMPLE **Determining a joint cumulative distribution function**

Find the joint distribution function of the two random variables of the preceding exercise, and use it to find the probability that both random variables will take on values less than 1.

Solution By definition,

$$F(x_1, x_2) = \begin{cases} \displaystyle\int_0^{x_1} \int_0^{x_2} 6e^{-2u-3v}\,du\,dv & \text{for } x_1 > 0,\ x_2 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

so that

$$F(x_1, x_2) = \begin{cases} (1 - e^{-2x_1})(1 - e^{-3x_2}) & \text{for } x_1 > 0,\ x_2 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and, hence,

$$F(1, 1) = (1 - e^{-2})(1 - e^{-3}) = 0.8216$$

■

Given the joint probability density of k random variables, the probability density of the ith random variable can be obtained by integrating out the other variables; symbolically,

Marginal density

$$\boxed{f_i(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_k)\,dx_1 \ldots dx_{i-1}\,dx_{i+1} \ldots dx_k}$$

and, in this context, the function f_i is called the **marginal density** of the ith random variable. Integrating out only some of the k random variables, we can similarly define **joint marginal densities** of any two, three, or more of the k random variables.

EXAMPLE **Determining a marginal density from a joint density**

With reference to the example on page 148, find the marginal density of the first random variable.

Solution Integrating out x_2, we get

$$f_1(x_1) = \begin{cases} \int_0^\infty 6e^{-2x_1-3x_2}\,dx_2 & \text{for } x_1 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

or

$$f_1(x) = \begin{cases} 2e^{-2x_1} & \text{for } x_1 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

■

To explain what we mean by the **independence** of continuous random variables, we could proceed as with discrete random variables on page 146 and define **conditional densities** first; however, it will be easier to say that

Independent random variables

> k random variables $X_1, \ldots, X_k$ are **independent** if and only if
>
> $$F(x_1, x_2, \ldots, x_k) = F_1(x_1) \cdot F_2(x_2) \cdots F_k(x_k)$$
>
> for all values $x_1, x_2, \ldots, x_k$ of these random variables.

In this notation $F(x_1, x_2, \ldots, x_k)$ is, as before, the joint distribution function of the k random variables, while $F_i(x_i)$ for $i = 1, 2, \ldots, k$ are the corresponding individual distribution function of the respective random variables. The same condition applies for discrete random variables.

EXAMPLE **Checking independence via the joint cumulative distribution**

With reference to the example on page 148, check whether the two random variables are independent.

Solution As we already saw in the example on page 148, the joint distribution function of the two random variables is given by

$$F(x_1, x_2) = \begin{cases} (1-e^{-2x_1})(1-e^{-3x_2}) & \text{for } x_1 > 0 \text{ and } x_2 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Now, since $F_1(x_1) = F(x_1, \infty)$ and $F_2(x_2) = F(\infty, x_2)$, it follows that

$$F_1(x_1) = \begin{cases} 1-e^{-2x_1} & \text{for } x_1 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and

$$F_2(x_2) = \begin{cases} 1-e^{-3x_1} & \text{for } x_2 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Thus, $F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2)$ for all (x_1, x_2) and the two random variables are independent. ■

When k random variables have a joint probability density, the k **random variables are independent** if and only if their joint probability density equals the product of the corresponding values of the marginal densities of the k random variables; symbolically,

$$f(x_1, x_2, \ldots, x_k) = f_1(x_1) \cdot f_2(x_2) \cdots f_k(x_k) \qquad \text{for all } (x_1, \ldots, x_k).$$

EXAMPLE **Establishing independence by factoring the joint probability density**

With reference to the example on page 148, verify that

$$f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$$

Solution On page 149 we showed that

$$f_1(x_1) = \begin{cases} 2e^{-2x_1} & \text{for } x_1 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and in the same way it can be shown that

$$f_2(x_2) = \begin{cases} 3e^{-3x_2} & \text{for } x_2 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Thus,

$$f_1(x_1) \cdot f_2(x_2) = \begin{cases} 6e^{-2x_1 - 3x_2} & \text{for } x_1 > 0 \text{ and } x_2 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and it can be seen that $f_1(x_1) \cdot f_2(x_2) = f(x_1, x_2)$ for all (x_1, x_2). ■

Given two continuous random variables X_1 and X_2, we define the **conditional probability density** of the first given that the second takes on the value x_2 as

Conditional probability density

$$f_1(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \quad \text{provided } f_2(x_2) \neq 0$$

where $f(x_1, x_2)$ and $f_2(x_2)$ are, as before, the joint density of the two random variables and the marginal density of the second. Note that this definition parallels that of the conditional probability distribution on page 146. Also, the joint probability density is the product

$$f(x_1, x_2) = f_1(x_1 \mid x_2)\, f_2(x_2).$$

EXAMPLE **Determining a conditional probability density**

If two random variables have the joint probability density

$$f(x_1, x_2) = \begin{cases} \frac{2}{3}(x_1 + 2x_2) & \text{for } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{elsewhere} \end{cases}$$

find the conditional density of the first given that the second takes on the value x_2.

Solution First we find the marginal density of the second random variable by integrating out x_1, and we get

$$f_2(x_2) = \int_0^1 \frac{2}{3}(x_1 + 2x_2)\,dx_1 = \frac{1}{3}(1 + 4x_2) \qquad \text{for } 0 < x_2 < 1$$

and $f_2(x_2) = 0$ elsewhere. Hence, by definition, the conditional density of the first random variable given that the second takes on the value x_2 is given by

$$f_1(x_1 \mid x_2) = \frac{\frac{2}{3}(x_1 + 2x_2)}{\frac{1}{3}(1 + 4x_2)} = \frac{2x_1 + 4x_2}{1 + 4x_2} \qquad \text{for } 0 < x_1 < 1, 0 < x_2 < 1$$

and $f_1(x_1 \mid x_2) = 0$ for $x_1 \le 0$ or $x_1 \ge 1$ and $0 < x_2 < 1$. ■

Properties of Expectation

Consider a function $g(X)$ of a single random variable X. For instance, if X is an oven temperature in degrees centigrade, then

$$g(X) = \frac{9}{5}X + 32$$

is the same temperature in degrees Fahrenheit.

The **expectation** of the function $g(X)$ is again the *sum* of the products *value* × *probability*.

Expected value of $g(X)$

In the discrete case, where X has probability distribution $f(x)$

$$E[g(X)] = \sum_{x_i} g(x_i) f(x_i)$$

In the continuous case, where X has probability density function $f(x)$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,dx$$

If X has mean $\mu = E(X)$, then taking $g(x) = (x - \mu)^2$, we have $E[g(X)] = E(X - \mu)^2$, which is just the variance σ^2 of X.

For any random variable Y, let $E(Y)$ denote its expectation, which is also its mean μ_Y. Its variance is $Var(Y)$ which is also written as σ_Y^2.

When $g(x) = ax + b$, for given constants a and b, then random variable $g(X)$ has expectation

$$E(aX + b) = \int_{-\infty}^{\infty} (ax + b) f(x)\,dx = a\int_{-\infty}^{\infty} x f(x)\,dx + b\int_{-\infty}^{\infty} f(x)\,dx$$
$$= aE(X) + b$$

and variance

$$Var(aX + b) = \int_{-\infty}^{\infty} (ax + b - a\mu_X - b)^2 f(x)\,dx$$
$$= a^2 \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\,dx = a^2 Var(X)$$

To summarize,

For given constants a and b

$$E(aX + b) = aE(X) + b \quad \text{and} \quad Var(aX + b) = a^2 Var(X)$$

EXAMPLE **The mean and standard deviation of a standardized random variable**

Let X have mean μ and standard deviation σ. Use the properties of expectation to show that the standardized random variable

$$Z = \frac{X - \mu}{\sigma}$$

has mean 0 and standard deviation 1.

Solution Since Z is of the form

$$Z = \frac{X - \mu}{\sigma} = \frac{1}{\sigma} X - \frac{\mu}{\sigma} = aX + b$$

where $a = 1/\sigma$ and $b = -\mu/\sigma$,

$$E(Z) = \frac{1}{\sigma} E(X) - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$$

and the variance of Z is

$$\left(\frac{1}{\sigma}\right)^2 Var(X) = \frac{\sigma^2}{\sigma^2} = 1$$

because $E(X) = \mu$ and $Var(X) = \sigma^2$. ■

EXAMPLE **Determining the mean and variance of 20 X**

Suppose the daily amount of electricity X required for a plating process has mean 10 and standard deviation 3 kilowatt-hours. If the cost of electricity is 20 dollars per kilowatt hour, find the mean, variance, and standard deviation of the daily cost of electricity.

Solution The daily cost of electricity, $g(X) = 20X$, has mean $20E(X) = 20 \times 10 = 200$ dollars and variance $(20)^2 Var(X) = (20)^2 3^2 = 3{,}600$. Its standard deviation is $\sqrt{3{,}600} = 60$ dollars. ■

Given any collection of k random variables, the function $Y = g(X_1, X_2, \ldots, X_k)$ is also a random variable. Examples include $Y = X_1 - X_2$ when $g(x_1, x_2) = x_1 - x_2$ and $Y = 2X_1 + 3X_2$ when $g(x_1, x_2) = 2x_1 + 3x_2$. The random variable $g(X_1, X_2, \ldots, X_k)$ has expected value, or mean, which is the sum of the products *value × probability*.

Expected value of $g(X_1, X_2, \ldots, X_k)$

In the discrete case,

$$E[g(X_1, X_2, \ldots, X_k)] = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_k} g(x_1, x_2, \ldots, x_k) f(x_1, x_2, \ldots, x_k)$$

In the continuous case,

$$E[g(X_1, X_2, \ldots, X_k)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_k) f(x_1, x_2, \ldots, x_k)\, dx_1\, dx_2 \cdots dx_k$$

Several important properties of expectation can be deduced from this definition. Taking $g(x_1, x_2) = (x_1 - \mu_1)(x_2 - \mu_2)$, we see that the product $(x_1 - \mu_1)(x_2 - \mu_2)$ will be positive if both values x_1 and x_2 are above their respective means or both are below their respective means. Otherwise it will be negative. The expected value $E[(X_1 - \mu_1)(X_2 - \mu_2)]$ will tend to be positive when large X_1 and X_2 tend to occur together and small X_1 and X_2 tend to occur together, with high probability. This measure $E[(X_1 - \mu_1)(X_2 - \mu_2)]$ of joint variation is called the population **covariance** of X_1 and X_2.

If X_1 and X_2 are independent so $f(x_1, x_2) = f_1(x_1) f_2(x_2)$,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_1)(x_2-\mu_2)f(x_1,x_2)\,dx_1\,dx_2$$
$$=\int_{-\infty}^{\infty}(x_1-\mu_1)f_1(x_1)\,dx_1\cdot\int_{-\infty}^{\infty}(x_2-\mu_2)f_2(x_2)\,dx_2=0$$

This result concerning zero covariance can be stated as

Independence implies that the covariance is zero

When X_1 and X_2 are independent, their covariance

$$E[(X_1-\mu_1)(X_2-\mu_2)]=0$$

Further, the expectation of a linear combination of two independent random variables $Y = a_1X_1 + a_2X_2$ is

$$\begin{aligned}\mu_Y &= E(Y) = E(a_1X_1 + a_2X_2)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(a_1x_1+a_2x_2)f_1(x)f_2(x_2)\,dx_1\,dx_2\\ &= a_1\int_{-\infty}^{\infty}x_1f_1(x_1)\,dx_1\int_{-\infty}^{\infty}f_2(x_2)\,dx_2\\ &\quad + a_2\int_{-\infty}^{\infty}f_1(x_1)\,dx_1\int_{-\infty}^{\infty}x_2f_2(x_2)\,dx_2\\ &= a_1E(X_1)+a_2E(X_2)\end{aligned}$$

This result holds even if the two random variables are not independent. Also,

$$\begin{aligned}Var(Y) &= E(Y-\mu_Y)^2 = E[(a_1X_1+a_2X_2-a_1\mu_1-a_2\mu_2)^2]\\ &= E[(a_1(X_1-\mu_1)+a_2(X_2-\mu_2))^2]\\ &= E[a_1^2(X_1-\mu_1)^2+a_2^2(X_2-\mu_2)^2+2a_1a_2(X_1-\mu_1)(X_2-\mu_2)]\\ &= a_1^2E[(X_1-\mu_1)^2]+a_2^2E[(X_2-\mu_2)^2]+2a_1a_2E[(X_1-\mu_1)(X_2-\mu_2)]\\ &= a_1^2\,Var(X_1)+a_2^2\,Var(X_2)\end{aligned}$$

since the third term is zero because we assumed X_1 and X_2 are independent.

These properties hold for any number of random variables whether they are continuous or discrete.

The mean and variance of linear combinations

Let X_i have mean μ_i and variance σ_i^2 for $i = 1, 2, \ldots, k$. The linear combination $Y = a_1X_1 + a_2X_2 + \cdots + a_kX_k$ has

$$E(a_1X_1+a_2X_2+\cdots+a_kX_k)=a_1E(X_1)+a_2E(X_2)+\cdots+a_kE(X_k)$$

or

$$\mu_Y=\sum_{i=1}^{k}a_i\mu_i$$

When the random variables are independent,

$$Var(a_1X_1 + a_2X_2 + \cdots + a_kX_k) = a_1^2\, Var(X_1) + a_2^2\, Var(X_2) + \cdots + a_k^2\, Var(X_k)$$

or

$$\sigma_Y^2 = \sum_{i=1}^{k} a_i^2 \sigma_i^2$$

EXAMPLE **Variances of $X_1 - X_2$ and $X_1 + X_2$ when X_1 and X_2 are independent**

Let X_1 have mean μ_1 and variance σ_1^2 and let X_2 have mean μ_2 and variance σ_2^2. Find the mean and variance of

(a) $X_1 - X_2$ and **(b)** $X_1 + X_2$

if X_1 and X_2 are independent.

Solution **(a)** Note that $X_1 - X_2$ is of the form $a_1X_1 + a_2X_2$ with $a_1 = -a_2 = 1$ so it has mean

$$1 \cdot \mu_1 + (-1)\mu_2 = \mu_1 - \mu_2$$

and variance

$$(1)^2 \cdot \sigma_1^2 + (-1)^2\sigma_2^2 = \sigma_1^2 + \sigma_2^2$$

(b) Since $X_1 + X_2$ corresponds to the case with $a_1 = a_2 = 1$, it has mean

$$1 \cdot \mu_1 + 1 \cdot \mu_2 = \mu_1 + \mu_2$$

and variance

$$1^2 \cdot \sigma_1^2 + 1^2 \cdot \sigma_2^2 = \sigma_1^2 + \sigma_2^2$$

■

EXAMPLE **Finding the mean and variance of $2X_1 + X_2 - 5$**

If X_1 has mean 4 and variance 9 while X_2 has mean -2 and variance 5, and the two are independent, find

(a) $E(2X_1 + X_2 - 5)$

(b) $Var(2X_1 + X_2 - 5)$

Solution According to the properties of expectation, the constant -5 is added to the expectation of $2X_1 + X_2$ but the variance is unchanged.

(a)
$$\begin{aligned} E(2X_1 + X_2 - 5) &= E(2X_1 + X_2) - 5 \\ &= 2E(X_1) + E(X_2) - 5 = 2(4) + (-2) - 5 = 1 \end{aligned}$$

(b)
$$\begin{aligned} Var(2X_1 + X_2 - 5) &= Var(2X_1 + X_2) \\ &= 2^2\, Var(X_1) + Var(X_2) = 2^2(9) + 5 = 41 \end{aligned}$$

■

EXAMPLE **The mean and variance of total time to coat and rinse**

The time to complete a coating process, X_1, has mean 35 minutes and variance 11, while the time to rinse, X_2, has mean 8 minutes and variance 5. Find the mean and standard deviation of the total time to coat and rinse.

Solution According to properties of expectation, the total time $X_1 + X_2$ has mean $35 + 8 = 43$ minutes. Treating the coating and rinsing times as independent, the variance is $11 + 5 = 16$, so that the standard deviation is 4. ■

EXAMPLE **The mean and variance of the sample mean $\overline{X}$**

Let the n random variables $X_1, X_2, \ldots, X_n$ be independent and each have the same distribution with mean μ and variance σ^2. Use the properties of expectation to show that the sample mean $\overline{X}$ has

(a) mean: $\mu_{\overline{X}} = E(\overline{X}) = \mu$

(b) variance: $\sigma^2_{\overline{X}} = Var(\overline{X}) = \dfrac{\sigma^2}{n}$

Solution **(a)** The sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

is a linear combination with constants $a_i = 1/n$ for $i = 1, 2, \ldots, n$. Consequently,

$$E(\overline{X}) = \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n) = \sum_{i=1}^{n}\frac{1}{n}\mu = \frac{1}{n}n\mu = \mu$$

so the expected value or mean of $\overline{X}$ is the same as the mean of each observation.

(b) The variance of $\overline{X}$ is

$$Var(\overline{X}) = \left(\frac{1}{n}\right)^2 Var(X_1) + \left(\frac{1}{n}\right)^2 Var(X_2) + \cdots + \left(\frac{1}{n}\right)^2 Var(X_n)$$

$$= \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2\sigma^2 = \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$$

so the variance of $\overline{X}$ equals the variance of a single observation divided by n. ■

EXAMPLE **The expected value of the sample variance**

Let the n random variables $X_1, X_2, \ldots, X_n$ be independent and each have the same distribution with mean μ and variance σ^2. Use the properties of expectation to show that σ^2 is the mean, or expectation, of the sample variance

$$\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$$

Solution We write $(X_i - \overline{X})^2 = (X_i - \mu + \mu - \overline{X})^2 = (X_i - \mu)^2 + (\mu - \overline{X})^2 + 2(X_i - \mu)(\mu - \overline{X})$ so the numerator of the sample variance is

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i=1}^{n}(\mu - \overline{X})^2 + 2\sum_{i=1}^{n}(X_i - \mu)(\mu - \overline{X})$$

and the last term equals $-2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) = -2n(\overline{X} - \mu)^2$. Consequently,

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2$$

Since $E(X_i - \mu)^2 = Var(X_i) = \sigma^2$ and, by the previous example, $E(\overline{X}) = \mu$ so $E(\overline{X} - \mu)^2 = Var(\overline{X}) = \sigma^2/n$, taking expectation term by term and summing,

$$E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \sum_{i=1}^{n}\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2$$

Dividing both sides by $n - 1$, we conclude that σ^2 is the expected value of the sample variance. ■

Exercises

5.71 Two scanners are needed for an experiment. Of the five available, two have electronic defects, another one has a defect in the memory, and two are in good working order. Two units are selected at random.

(a) Find the joint probability distribution of X_1 = the number with electronic defects, and X_2 = the number with a defect in memory.

(b) Find the probability of 0 or 1 total defects among the two selected.

(c) Find the marginal probability distribution of X_1.

(d) Find the conditional probability distribution of X_1 given $X_2 = 0$.

5.72 Two random variables are independent and each has a binomial distribution with success probability 0.4 and 2 trials.

(a) Find the joint probability distribution.

(b) Find the probability that the second random variable is greater than the first.

5.73 If two random variables have the joint density

$$f(x_1, x_2) = \begin{cases} x_1 x_2 & \text{for } 0 < x_1 < 2,\ 0 < x_2 < 1 \\ 0 & \text{elsewhere} \end{cases}$$

find the probabilities that

(a) both random variables will take on values less than 1;

(b) the sum of the values taken on by the two random variables will be less than 1.

5.74 With reference to the preceding exercise, find the marginal densities of the two random variables.

5.75 With reference to Exercise 5.73, find the joint distribution function of the two random variables, the distribution functions of the individual random variables, and check whether the two random variables are independent.

5.76 If two random variables have the joint density

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & \text{for } 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

find the probability that $0.2 < X < 0.5$ and $0.4 < Y < 0.6$.

5.77 With reference to the preceding exercise, find the joint distribution function of the two random variables and use it to verify the value obtained for the probability.

5.78 With reference to Exercise 5.76, find both marginal densities and use them to find the probabilities that

(a) $X > 0.8$;

(b) $Y < 0.5$.

5.79 With reference to Exercise 5.76, find

(a) an expression for $f_1(x \mid y)$ for $0 < y < 1$;

(b) an expression for $f_1(x \mid 0.5)$;

(c) the mean of the conditional density of the first random variable when the second takes on the value 0.5.

5.80 With reference to the conditional probability example on page 150, find expressions for

(a) the conditional density of the first random variable when the second takes on the value $x_2 = 0.25$;

(b) the conditional density of the second random variable when the first takes on the value x_1.

5.81 If three random variables have the joint density

$$f(x, y, z) = \begin{cases} k(x + y)e^{-z} & \text{for } 0 < x < 1,\ 0 < y < 2,\ z > 0 \\ 0 & \text{elsewhere} \end{cases}$$

find

(a) the value of k;

(b) the probability that $X < Y$ and $Z > 1$.

5.82 With reference to the preceding exercise, check whether

(a) the three random variables are independent;

(b) any two of the three random variables are pairwise independent.

5.83 A pair of random variables has the **circular normal distribution** if their joint density is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma^2} e^{-[(x_1-\mu_1)^2 + (x_2-\mu_2)^2]/2\sigma^2}$$

for $-\infty < x_1 < \infty$ and $-\infty < x_2 < \infty$.

(a) If $\mu_1 = 2$ and $\mu_2 = -2$, and $\sigma = 10$, use Table 3 to find the probability that $-8 < X_1 < 14$ and $-9 < X_2 < 3$.

(b) If $\mu_1 = \mu_2 = 0$ and $\sigma = 3$, find the probability that (X_1, X_2) is contained in the region between the two circles $x_1^2 + x_2^2 = 9$ and $x_1^2 + x_2^2 = 36$.

5.84 A precision drill positioned over a target point will make an acceptable hole if it is within 5 microns of the target. Using the target as the origin of a rectangular system of coordinates, assume that the coordinates (x, y) of the point of contact are values of a pair of random variables having the circular normal distribution (see Exercise 5.83) with $\mu_1 = \mu_2 = 0$ and $\sigma = 2$. What is the probability that the hole will be acceptable?

5.85 With reference to Exercise 5.73, find the expected value of the random variable whose values are given by $g(x_1, x_2) = x_1 + x_2$.

5.86 With reference to Exercise 5.76, find the expected value of the random variable whose values are given by $g(x, y) = x^2 y$.

5.87 If measurements of the length and the width of a rectangle have the joint density

$$f(x, y) = \begin{cases} \dfrac{1}{ab} & \text{for } L - \dfrac{a}{2} < x < L + \dfrac{a}{2}, \\ & W - \dfrac{b}{2} < y < W + \dfrac{b}{2} \\ 0 & \text{elsewhere} \end{cases}$$

find the mean and the variance of the corresponding distribution of the area of the rectangle.

5.88 Establish a relationship between $f_1(x_1 \mid x_2)$, $f_2(x_2 \mid x_1)$, $f_1(x_1)$, and $f_2(x_2)$.

5.89 If X_1 has mean 1 and variance 5 while X_2 has mean -1 and variance 5, and the two are independent, find

(a) $E(X_1 + X_2)$;

(b) $Var(X_1 + X_2)$.

5.90 If X_1 has mean -3 and variance 2 while X_2 has mean 5 and variance 4, and the two are independent, find

(a) $E(X_1 - X_2)$;

(b) $Var(X_1 - X_2)$.

5.91 If X_1 has mean 1 and variance 3 while X_2 has mean -2 and variance 5, and the two are independent, find

(a) $E(X_1 + 2X_2 - 3)$;

(b) $Var(X_1 + 2X_2 - 3)$.

5.92 The time for an older machine to complete a check on a computer chip, X_1, has mean 65 milliseconds and variance 16, while the time for a newer model, X_2, has mean 45 milliseconds and variance 9. Find the expected time savings using the newer model when

(a) checking a single chip;

(b) checking 200 chips.

(c) Find the standard deviations in parts (a) and (b), assuming all of the checking times are independent.

5.93 Let $X_1, X_2, \ldots, X_{20}$ be independent and let each have the same marginal distribution with mean 10 and variance 3. Find

(a) $E(X_1 + X_2 + \cdots + X_{20})$;

(b) $Var(X_1 + X_2 + \cdots + X_{20})$.

5.11 Moment Generating Functions

An alternative to a probability distribution can sometimes greatly simplify the calculation of moments. The **moment generating function** (mgf) of a random variable X, or its probability distribution, is the function defined by

$$M(t) = E(e^{tX})$$

which is the expectation of the exponential function e^{tX}. In the discrete case,

$$M(t) = E(e^{tX}) = \sum_{\text{all } x_i} e^{tx_i} f(x_i)$$

and

$$M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

in the continuous case. For each fixed t, the integrand is a positive function of x so $M(t)$ is either finite or infinite. Note that $M(0) = E(e^{0X}) = E(1) = 1$ always exists, but we require $M(t)$ to exist for an interval of values of t.

The probability distribution, or random variable X, is said to **possess a moment generating function** $M(t)$ if this function is finite for t in some interval containing zero, say $|t| \leq T$ for some $T > 0$.

Under the condition that $M(t)$ is finite for $|t| \leq T$, for some $T > 0$, we can obtain successive derivatives by differentiating under the integral or summation sign. In the continuous case, we obtain

$$M'(t) = \frac{d}{dt}M(t) = \int_{-\infty}^{\infty} \frac{d}{dt}e^{tx} f(x)\,dx = \int_{-\infty}^{\infty} x\,e^{tx} f(x)\,dx$$

$$M''(t) = \frac{d^2}{dt^2}M(t) = \int_{-\infty}^{\infty} \frac{d}{dt}xe^{tx} f(x)\,dx = \int_{-\infty}^{\infty} x^2 e^{tx} f(x)\,dx$$

For either the continuous or discrete case,

$$M^{(k)}(t) = \frac{d^k}{dt^k}M(t) = E(X^k e^{tX}) \qquad \text{for } k = 1, 2, \ldots$$

Setting $t = 0$, we obtain the moments about the origin

$$M'(0) = E(X)$$
$$M''(0) = E(X^2)$$

Differentiating k times, the kth derivative is related to the kth moment

$$M^{(k)}(0) = E(X^k) \qquad \text{for } k = 1, 2, \ldots$$

Basic properties of moment generating functions

Theorem 5.2 If the moment generating function is finite for $|t| \leq T$, for some $T > 0$, it uniquely determines the the probability distribution.

Then, all moments exist and can be obtained from the relation

$$M^{(k)}(0) = E(X^k)$$

From the definition of expectation, the moments of any random variable must be obtained by integration or summation of a series. When the moment generating function is available, this process can be replaced by a straightforward differentiation. We find the moment generating functions and illustrate the calculation of mean and variance for several common distributions in the next examples.

EXAMPLE **Moment generating function for binomial distribution**

Let X have the binomial distribution with probability distribution

$$b(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } x = 0, 1, \ldots, n$$

Show that

(a) $M(t) = (1 - p + pe^t)^n$ for all t

(b) $E(X) = np$ and $Var(X) = np(1-p)$

Solution **(a)** By definition of the moment generating function

$$M(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x}$$
$$= \sum_{x=0}^{n} \binom{n}{x} (e^t p)^x (1-p)^{n-x}$$
$$= (pe^t + 1 - p)^n \qquad \text{for all } t$$

where we have used the binomial formula

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$$

(b) Differentiating $M(t)$, we find

$$M'(t) = n p e^t (p e^t + 1 - p)^{n-1}$$
$$M''(t) = (n-1) n p^2 e^{2t} (p e^t + 1 - p)^{n-2} + n p e^t (p e^t + 1 - p)^{n-1}$$

Evaluating these derivatives at $t = 0$, we obtain the moments

$$E(X) = np$$
$$E(X^2) = (n-1) n p^2 + np$$

Also, the variance is

$$Var(X) = E(X^2) - [E(X)]^2 = np(1-p)$$

■

EXAMPLE **Moment generating function for Poisson distribution**

Let X have the Poisson distribution with probability distribution

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad \text{for } x = 0, 1, \ldots, \infty$$

Show that

(a) $M(t) = e^{\lambda(e^t - 1)}$ for all t

(b) $E(X) = \lambda$ and $Var(X) = \lambda$

The mean and variance of the Poisson distribution are equal.

Solution **(a)** By definition of the moment generating function

$$M(t) = \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x}{x!} e^{-\lambda} = \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} e^{-\lambda}$$
$$= e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)} \qquad \text{for } x = 0, 1, \ldots, \infty$$

where we have used the series $e^y = \sum_{k=0}^{\infty} \frac{y^k}{k!}$

(b) Differentiating $M(t)$, we find

$$M'(t) = \lambda e^t e^{\lambda(e^t - 1)}$$
$$M''(t) = \lambda e^t e^{\lambda(e^t - 1)} + \lambda^2 e^{2t} e^{\lambda(e^t - 1)}$$

Evaluating these derivatives at $t = 0$, we obtain the moments

$$E(X) = \lambda$$
$$E(X^2) = \lambda + \lambda^2$$

Also, the variance is

$$Var(X) = E(X^2) - [E(X)]^2 = \lambda$$

■

EXAMPLE **Moment generating function for gamma distribution**

The gamma distribution has probability density function

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \qquad \text{for } x > 0$$

Show that its moment generating function is

$$M(t) = \frac{1}{(1 - \beta t)^\alpha}$$

and verify the mean and variance.

Solution

$$M(t) = \int_0^\infty e^{tx} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx$$
$$= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x(1-\beta t)/\beta} dx$$

This last integral is finite for all $t < 1/\beta$ and can be evaluated by multiplying and dividing by $(1 - \beta t)^\alpha$ to obtain a gamma density with parameters α and $\beta = \beta/(1 - \beta t)$. We conclude that

$$M(t) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \Gamma(\alpha) \frac{\beta^\alpha}{(1 - \beta t)^\alpha} = \frac{1}{(1 - \beta t)^\alpha}$$

To obtain the moments, we differentiate and find

$$M'(t) = \alpha \frac{1}{(1 - \beta t)^{\alpha+1}} \beta$$
$$M''(t) = (\alpha + 1)\alpha \frac{1}{(1 - \beta t)^{\alpha+2}} \beta^2$$

Setting $t = 0$,

$$E(X) = M'(0) = \alpha\beta \qquad \text{and} \qquad E(X^2) = M''(0) = \alpha(\alpha + 1)\beta^2$$

so $Var(X) = \alpha\beta^2$.

■

EXAMPLE **Moment generating function for chi square distribution**

Show that the moment generating function of the chi square distribution with ν degrees of freedom is

$$M(t) = \frac{1}{(1 - 2t)^{\nu/2}}$$

and that $E(X) = \nu$ and $Var(X) = 2\nu$.

Solution These results follow from the previous example because the chi square distribution with ν degrees of freedom is also a gamma distribution with $\alpha = \nu/2$ and $\beta = 2$. ■

EXAMPLE **Moment generating function for normal distribution**

Show that the normal distribution, whose probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad \text{has} \quad M(t) = e^{t\mu + \frac{1}{2}t^2\sigma^2}$$

which exists for all t. Also, verify the first two moments.

Solution To obtain the moment generating function, we use the identity

$$tx - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} = -\frac{1}{2}\frac{[x-(t\sigma^2+\mu)]^2}{\sigma^2} + t\mu + \frac{1}{2}t^2\sigma^2$$

obtained by completing the square. Then

$$M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}[x-(t\sigma^2+\mu)]^2/\sigma^2} dx \times e^{t\mu+\frac{1}{2}t^2\sigma^2} = e^{t\mu+\frac{1}{2}t^2\sigma^2}$$

To obtain the moments of the normal, we differentiate once to obtain

$$M'(t) = e^{t\mu + \frac{1}{2}t^2\sigma^2}(\mu + t\sigma^2)$$

and a second time to get

$$M''(t) = e^{t\mu + \frac{1}{2}t^2\sigma^2}[(\mu + t\sigma^2)^2 + \sigma^2].$$

Setting $t = 0$,

$$E[X] = M'(0) = \mu \quad \text{and} \quad E(X^2) = M''(0) = \sigma^2 + \mu^2$$

so $Var(X) = \sigma^2$ as the notation suggests. ■

A basic property relates the moment generating function of $a + bX$ to that of X.

Moment generating function of $a + bX$

Theorem 5.3 Let X have moment generating function $M(t)$ and let a and b be constants. Then

$$M_{a+bX}(t) = E(e^{(a+bX)t}) = e^{at} \cdot M(bt)$$

For instance, the moment generating function of $X - \mu$, corresponding to $b = 1$ and $a = -\mu$, is

$$M_{X-\mu}(t) = e^{-\mu t} \cdot M_X(t)$$

EXAMPLE **Converting to the standard normal distribution**

Let X be distributed as normal with mean μ and variance σ^2. Use moment generating functions to show that

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution.

Solution By the previous example, X has moment generating function

$$M(t) = e^{t\mu + \frac{1}{2}t^2\sigma^2}$$

Further

$$Z = \frac{X - \mu}{\sigma} = a + bX \qquad \text{with} \qquad a = \frac{-\mu}{\sigma} \qquad \text{and} \qquad b = \frac{1}{\sigma}$$

Therefore, the moment generating function of Z is

$$M_Z(t) = e^{at} \cdot M(bt) = e^{-(\mu/\sigma)t}\, e^{(t/\sigma)\mu + \frac{1}{2}(t/\sigma)^2\sigma^2} = e^{\frac{1}{2}t^2}$$

This last expression is the moment generating function of a normal distribution having mean 0 and variance 1. It exists for all t so the moment generating function uniquely determines the distribution. ■

One of most useful properties of moment generating functions is a multiplication property for independent random variables.

Moment generating function of sum under independence

Theorem 5.4 Let X and Y be independent random variables with moment generating functions M_X and M_Y. The sum $Z = X + Y$ has moment generating function

$$M_Z(t) = M_X(t)\, M_Y(t)$$

on the interval of t where $M_X(t)$ and $M_Y(t)$ exist.

Proof

$$M_Z(t) = E(e^{tZ}) = E(e^{t(X+Y)}) = E(e^{tX} e^{tY})$$

Then, from the assumption of independence,

$$= E(e^{tX} e^{tY}) = E(e^{tX})\, E(e^{tY}) = M_X(t)\, M_Y(t)$$

EXAMPLE **Sum of two independent normal random variables is normal**

Let X and Y be independent normal random variables. Let X have mean μ_X and variance σ_X^2 while Y has mean μ_Y and variance σ_Y^2. Use moment generating functions to show that

(a) $X + Y$ has a normal distribution with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$.

(b) $X - Y$ has a normal distribution with mean $\mu_X - \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$.

Solution **(a)** From a previous example, the two moment generating functions are

$$M_X(t) = e^{t\mu_X + \frac{1}{2}t^2\sigma_X^2}$$
$$M_Y(t) = e^{t\mu_Y + \frac{1}{2}t^2\sigma_Y^2}$$

Their product is

$$M_X(t)M_Y(t) = e^{t\mu_X + \frac{1}{2}t^2\sigma_X^2}\, e^{t\mu_Y + \frac{1}{2}t^2\sigma_Y^2}$$
$$= e^{t(\mu_X + \mu_Y) + \frac{1}{2}t^2(\sigma_X^2 + \sigma_Y^2)}$$

which we identify as the moment generating function of a normal random variable having mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$.

(b) Since X and $-Y$ are independent and $-Y$ has moment generating function $M_Y(-t)$, the result follows (see Exercise 5.98). ■

Exercises

5.94 Let $f(x) = 0.2$ for $x = 0, 1, 2, 3, 4$.

(a) Find the moment generating function.

(b) Obtain $E(X)$ and $E(X^2)$ by differentiating the moment generating function.

5.95 Let

$$f(x) = 0.25\binom{2}{x} \qquad \text{for } x = 0, 1, 2$$

(a) Find the moment generating function.

(b) Obtain $E(X)$ and $E(X^2)$ by differentiating the moment generating function.

5.96 Let Z have a normal distribution with mean 0 and variance 1.

(a) Find the moment generating function of Z^2.

(b) Identify the distribution of Z^2 by recognizing the form of the moment generating function.

5.97 Let X be a continuous random variable having probability density function

$$f(x) = \begin{cases} 2e^{-2x} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Find the moment generating function.

(b) Obtain $E(X)$ and $E(X^2)$ by differentiating the moment generating function.

5.98 Establish the result on page 162 concerning the difference of two independent normal random variables, X and Y.

5.99 Let X and Y be independent normal random variables with

$$E(X) = 2 \quad \text{and} \quad \sigma_X^2 = 4$$
$$E(Y) = 1 \quad \text{and} \quad \sigma_Y^2 = 9$$

(a) Use moment generating functions to show that $2X - 3Y + 5$ has a normal distribution.

(b) Find the mean and variance of the random variable in part (a).

5.100 Let X have the geometric distribution

$$f(x) = p(1-p)^{x-1} \qquad \text{for } x = 1, 2, \ldots$$

Obtain the moment generating function for

$$t < -\ln(1-p)$$

[*Hint:* Recall that $\sum_{k=0}^{\infty} r^k = \dfrac{1}{1-r}$ for $|r| < 1$.]

5.12 Checking If the Data Are Normal

In many instances, an experimenter needs to check whether a data set appears to be generated by a normally distributed random variable. As indicated in Figure 2.8, the normal distribution can serve to model variation in some quantities. Further, many commonly used statistical procedures, which we describe in later chapters, require that the probability distribution be nearly normal. Consequently, in a great number of applications it is prudent to check the assumption that the data are normal.

Although they involve an element of subjective judgment, graphical procedures are the most helpful for detecting serious departures from normality. Histograms can be checked for lack of symmetry. A single long tail certainly contradicts the assumption of a normal distribution. However, another special graph, called a **normal scores plot** or **normal quantile plot**, is even more effective in detecting departures from normality. To introduce such a plot, we consider a sample of size 4. In practice, we need a minimum of 15–20 observations in order to evaluate the agreement with normality.

The term **normal scores** refers to an idealized sample from the standard normal distribution. It consists of the values of z that divide the axes into equal probability intervals. For sample size $n = 4$, the normal scores are

$$\begin{aligned} m_1 &= -z_{0.20} = -0.84 \\ m_2 &= -z_{0.40} = -0.25 \\ m_3 &= z_{0.40} = 0.25 \\ m_4 &= z_{0.20} = 0.84 \end{aligned}$$

as illustrated in Figure 5.25 on page 164.

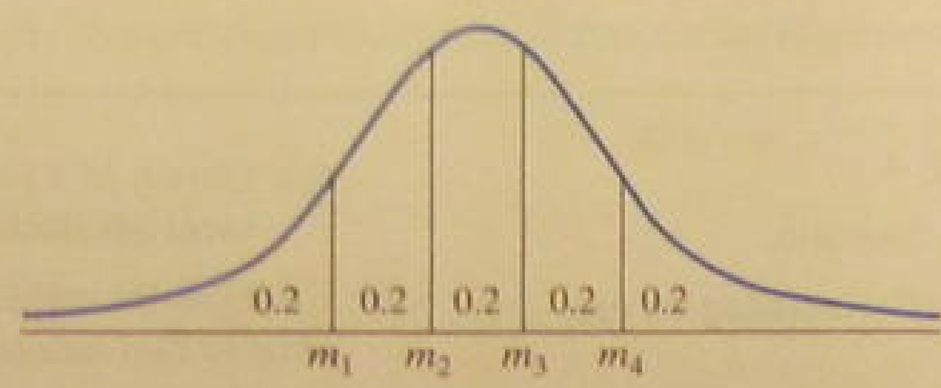


Figure 5.25
The standard normal distribution and the normal scores for $n = 4$

To construct a normal scores plot,

1. Order the data from smallest to largest;
2. Obtain the normal scores;
3. Plot the ith largest observation, versus the ith normal score m_i, for all i.

EXAMPLE **A simple normal scores plot**

Suppose the four observations are 67, 48, 76, 81. Construct a normal scores plot.

Solution The ordered observations are 48, 67, 76, 81. Earlier, we found that $m_1 = -z_{0.20} = -0.84$, so we plot the pair $(-0.84, 48)$. Continuing, we obtain Figure 5.26.

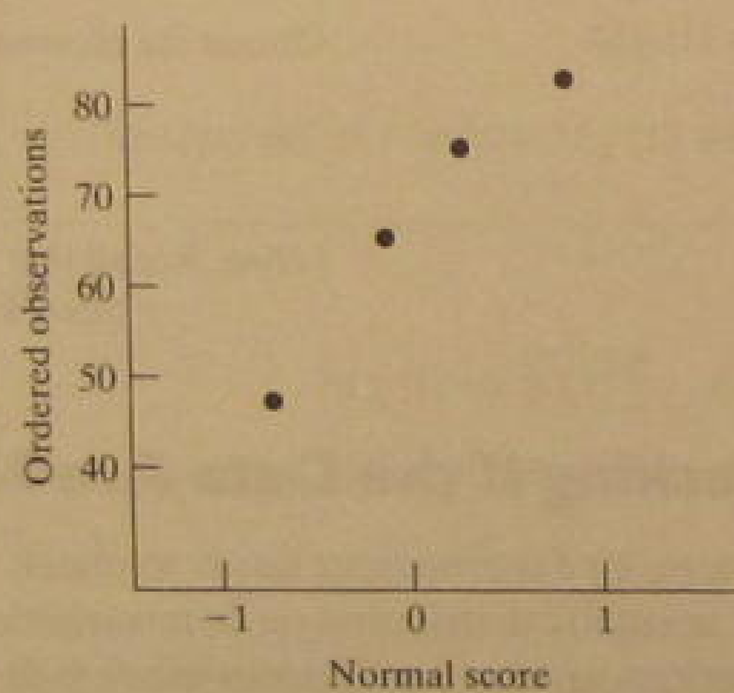


Figure 5.26
The normal scores plot

■

If the data were from a standard normal distribution, we would expect the ith largest observation to approximate the ith normal score so that the normal scores plot would resemble a 45° line through the origin. When the distribution is normal with an unspecified μ and σ,

$$z = \frac{x - \mu}{\sigma}$$

so the idealized z values can be converted to idealized x values through the relation $x = \mu + \sigma z$. Because the idealized values have this linear relation, it is sufficient to plot the ordered observations versus the normal scores obtained from the standard normal distribution. If the normal distribution prevails, the pattern should still be a straight line. But the line need not pass through the origin or have slope 1.

The construction of normal scores plots by hand is a difficult task at best. Fortunately, they can be treated easily with most statistical programs. (See Exercise 5.102.) Many slight variants are used in the calculation of the normal scores but the plots are very similar if more than 20 observations are plotted. Whichever computer program you use, if a normal distribution is plausible, the plot will have a straight-line appearance.

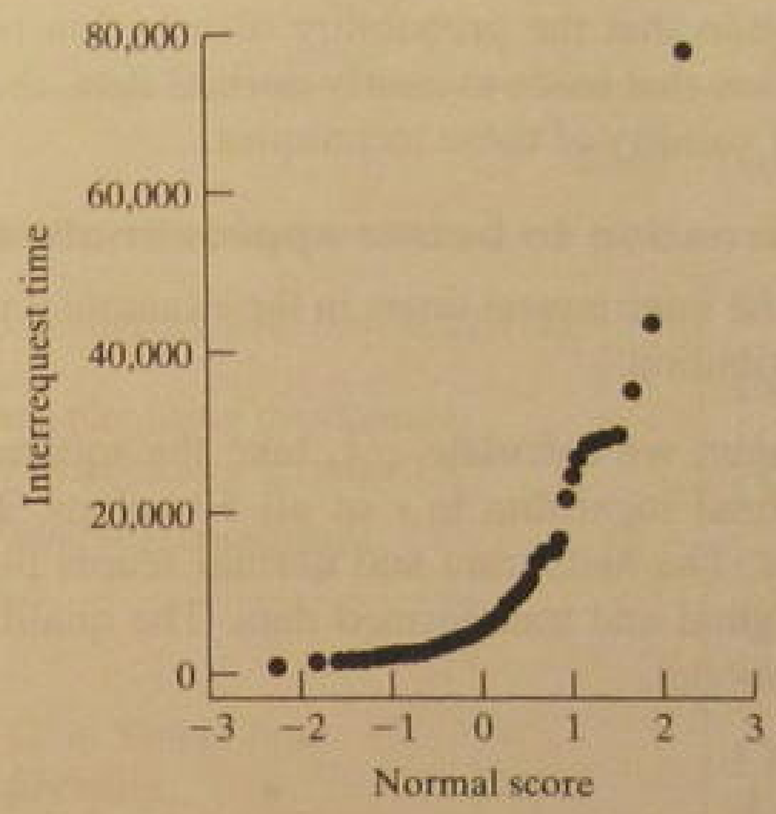


Figure 5.27
The normal scores plot of the interrequest times

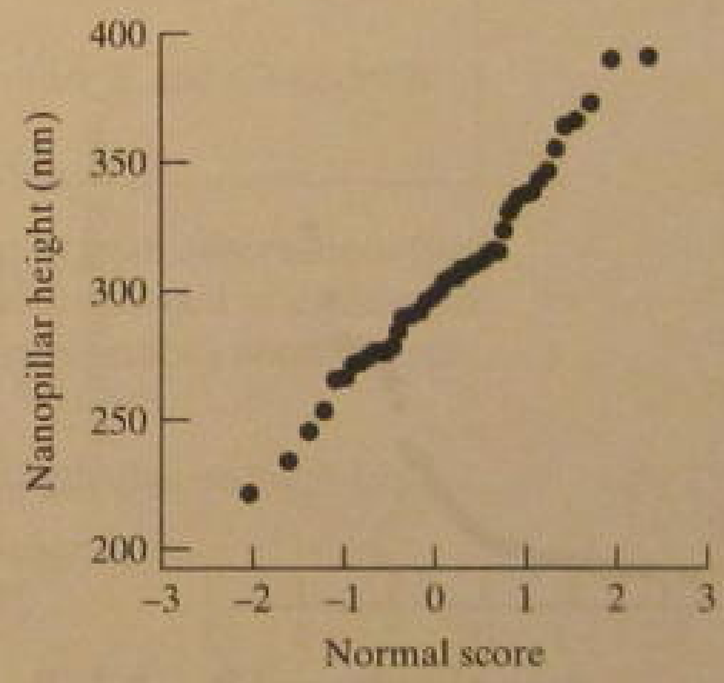


Figure 5.28
Normal scores plot of nanopillar heights

Figure 5.27 shows the normal scores plot for the interrequest times given on page 19. The bending shows that the largest values are larger than would be expected under a normal distribution. On the other hand, Figure 5.28 exhibits a normal scores plot of the nanopillar height data (see the example on page 15), and a normal distribution appears to be plausible.

5.13 Transforming Observations to Near Normality

When the histogram or normal scores plot indicate that the assumption of a normal distribution is invalid, transformations of the data can often improve the agreement with normality. Scientists regularly express their observations in natural logs. We consider a few other transformations, as indicated in Table 5.1.

Table 5.1 Some useful transformations

Make Large Values Smaller:				Make Large Values Larger:	
$-\frac{1}{x}$	$\ln x$	$x^{1/4}$	$\sqrt{x}$	x^2	x^3

If the observations are positive and the distribution has a long tail or a few stragglers on the right, then the $\ln x$ or $\sqrt{x}$ transformations will pull the large values down farther than they pull the central or small values. If the transformed observations have a nearly straight line normal scores plot, it is usually advantageous to use the normality of this new scale to perform any statistical analysis. Further, the validity of many of the powerful statistical methods described in later chapters rests on

the assumption that the probability distribution is nearly normal. By choosing a transformation that leads to nearly normal data, the investigator can greatly extend the range of validity of these techniques.

EXAMPLE **A transformation to better approximate a normal distribution**

Transform the interrequest times in the example on page 19 to better approximate a normal distribution.

Solution On a computer, we calculate $\sqrt{x}$, take the square root again to obtain $x^{1/4}$, and take the natural logarithm $\ln x$ of all 50 values. The transformation $\ln x$ appears to work best. The histogram and normal scores plot are shown in Figure 5.29 for both the original and transformed data. The quality of the fit further confirms the log-normal model. ■

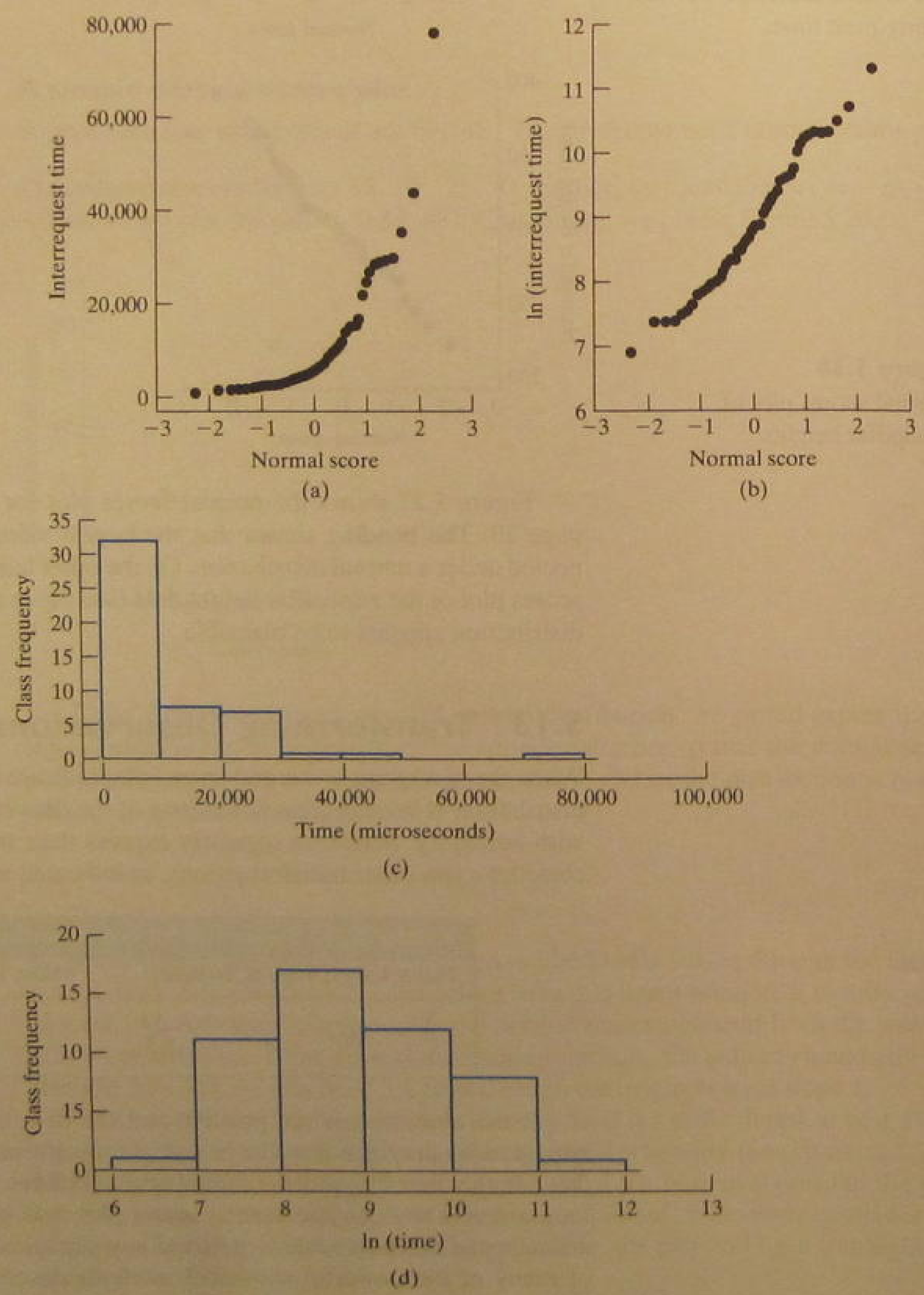


Figure 5.29
(a) The normal scores plot of interrequest time (b) The normal scores plot of ln (interrequest time) (c) Histogram of interrequest time (d) Histogram of ln (interrequest time)

Exercises

5.101 For any 11 observations,

(a) Use Table 3 to verify that the normal scores are

$-1.38 \quad -0.97 \quad -0.67 \quad -0.43 \quad -0.21 \quad 0 \quad 0.21 \quad 0.43 \quad 0.67 \quad 0.97 \quad 1.38$

(b) Construct a normal scores plot using the observations on the times between neutrinos on page 13.

5.102 (Normal scores plots) The *MINITAB* commands

> **Dialog box:**
>
> **Calc > Calculator** Type *C*2 in **Store**. Type *NORMAL SCORES* in **Expression**.
> Click **OK**.
>
> **Graph > Plot**. Type *C*1 under **Y** and *C*2 under **X**.
> Click **OK**.

will create a normal scores plot from observations that were set in C1. (*MINITAB* uses a variant of the normal scores, m_i, that we defined.) Construct a normal scores plot of

(a) the aluminum alloy strength data on page 19;

(b) the decay time data on page 140.

5.103 (Transformations) The *MINITAB* commands

> **Dialog box:**
>
> **Calc > Calculator** Type *C*2 in **Store**. Type *LOGE(C1)* in **Expression**.
> Click **OK**.
>
> **Calc > Calculator** Type *C*3 in **Store**. Type *SQRT(C1)* **Expression**.
> Click **OK**.
>
> **Calc > Calculator** Type *C*4 in **Store**. Type *SQRT(C3)* **Expression**.
> Click **OK**.

will place $\ln x$ in C2, $\sqrt{x}$ in C3, and $x^{1/4}$ in C4 for observations that are set in C1. Normal scores plots can then be constructed as in Exercise 5.102. Try these three transformations and construct the corresponding normal scores plots for

(a) the decay time data on page 140;

(b) the interrequest time data on page 19.

5.14 Simulation

Simulation techniques have grown up with computers. They are ideally suited for doing the repetitious calculations required. To **simulate** the observation of continuous random variables, we usually start with uniform random numbers and relate these to the distribution function of interest. We could use two- or three-digit random integers, perhaps selected from Table 7, but most software programs have a continuous uniform random number generator. That is, they produce approximations to random numbers from the uniform distribution

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Suppose we wish to simulate an observation from the exponential distribution

$$F(x) = 1 - e^{-0.3x}, \quad 0 < x < \infty$$

The computer would first produce the value u from the uniform distribution. Then we solve (see Exercise 5.104)

$$u = F(x) = 1 - e^{-0.3x}$$

so $x = [-\ln(1-u)]/0.3$ is the corresponding value of an exponential random variable. For instance, if $u = 0.45$, then $x = [-\ln(1-0.45)]/0.3 = 1.993$. This is illustrated graphically in Figure 5.30, where u is located on the vertical scale and the corresponding x value is read from the horizontal scale. (The theory on which this method is based involves the so-called *probability integral transformation*, which is presented in the example on page 197.) If we wish to simulate a sample from F, the preceding process is repeated with a different u for each new observation x.

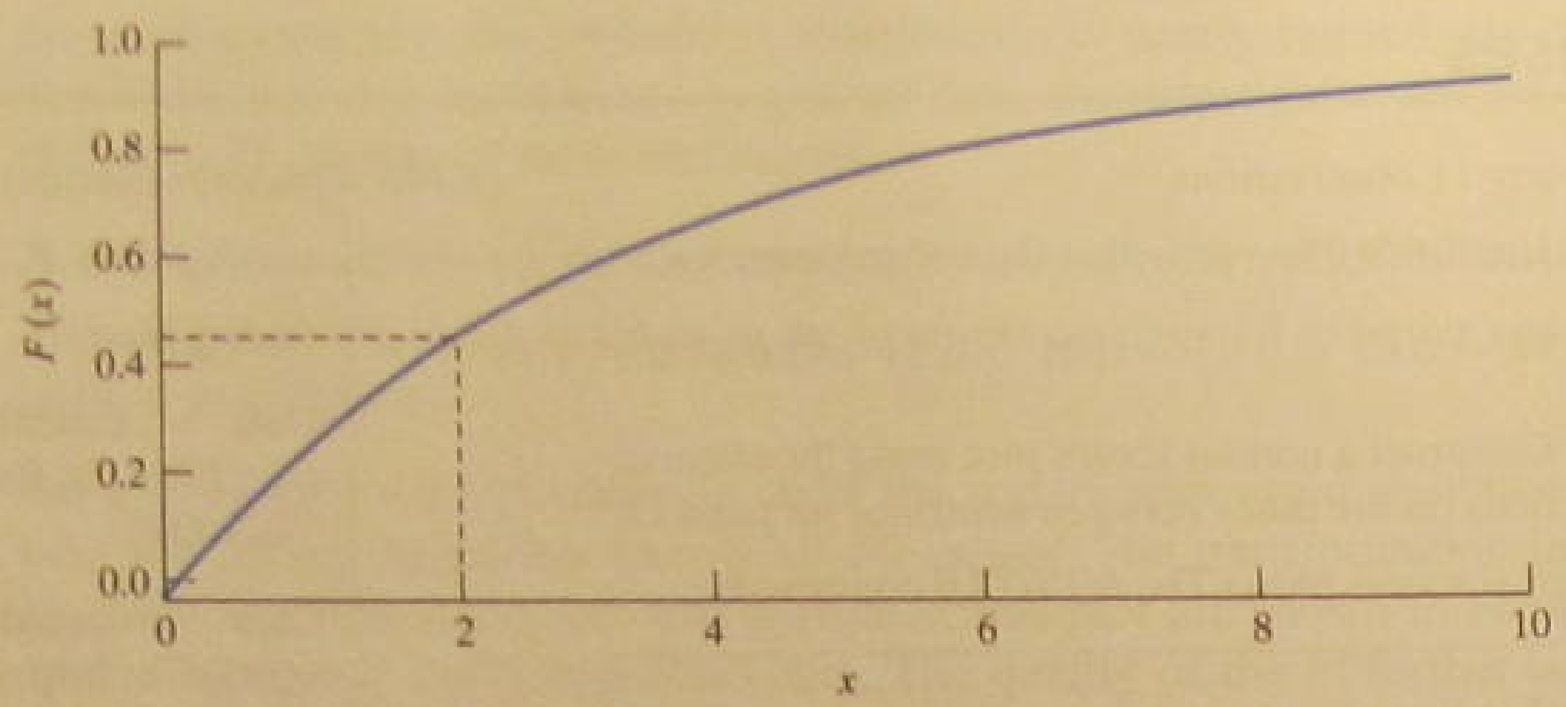


Figure 5.30
Exponential cumulative distribution with mean $\frac{10}{3}$

A similar procedure applies to the simulation of observations from a Weibull distribution. Starting with the value of a uniform variable u, we now solve (see Exercise 5.105)

$$u = F(x) = 1 - e^{-\alpha x^{\beta}}$$

for

$$x = \left[-\frac{1}{\alpha}\ln(1-u)\right]^{1/\beta}$$

which is the corresponding value of a Weibull random variable.

EXAMPLE **Simulating five values from a Weibull distribution**

Simulate five observations of a random variable having the Weibull distribution with $\alpha = 0.05$ and $\beta = 2.0$.

Solution A computer generates the five values 0.57, 0.74, 0.26, 0.77, 0.12. (Alternatively, they could have been obtained by reading two digits at a time from Table 7.) We calculate

$$x = [-20.0\ln(1-0.57)]^{1/2} = 4.108$$
$$x = [-20.0\ln(1-0.74)]^{1/2} = 5.191$$

The reader can show that the last three uniform numbers yield $x = 2.454, 5.422, 1.599$. ■

Suppose we need to simulate values from the normal distribution with a specified μ and σ^2. By the relation

$$z = \frac{x-\mu}{\sigma}$$

it follows that $x = \mu + \sigma z$, so a value x can be calculated from the value of a standard normal variable z. Although z can be obtained from the value for a uniform variable u by numerically solving $u = F(z)$, another approach called the Box-Muller-Marsaglia method is almost universally preferred. It starts with a pair of independent uniform variables (u_1, u_2) and produces two standard normal variables

$$z_1 = \sqrt{-2\ln(u_2)}\cos(2\pi u_1)$$
$$z_2 = \sqrt{-2\ln(u_2)}\sin(2\pi u_1)$$

where the angle is expressed in radians. Then $x_1 = \mu + \sigma z_1$ and $x_2 = \mu + \sigma z_2$ are treated as two independent observations of normal random variables (see Exercise 5.106). Most statistical packages include a normal random number generator (see Exercise 5.108).

EXAMPLE **Simulating two values from a normal distribution**

Simulate two observations of a random variable having the normal distribution with $\mu = 50$ and $\sigma = 5$.

Solution A computer generates the two values 0.253 and 0.531 from a uniform distribution. (Alternatively, they could have been obtained by reading three digits at a time from Table 7.) We first calculate the standard normal values

$$z_1 = \sqrt{-2\ln(0.531)}\cos(2\pi \cdot 0.253) = -0.021$$
$$z_2 = \sqrt{-2\ln(0.531)}\sin(2\pi \cdot 0.253) = 1.125$$

and then the normal values

$$x_1 = 50 + 5z_1 = 50 + 5(-0.021) = 49.895$$
$$x_2 = 50 + 5z_2 = 50 + 5(1.125) = 55.625$$

■

Exercises

5.104 Verify that

(a) the exponential density $0.3e^{-0.3x}$, $x > 0$ corresponds to the distribution function $F(x) = 1 - e^{-0.3x}$, $x > 0$;

(b) the solution of $u = F(x)$ is given by $x = [-\ln(1-u)]/0.3$.

5.105 Verify that

(a) the Weibull density $\alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}$, $x > 0$, corresponds to the distribution function $F(x) = 1 - e^{-\alpha x^{\beta}}$, $x > 0$;

(b) the solution of $u = F(x)$ is given by $x = \left[-\frac{1}{\alpha}\ln(1-u)\right]^{1/\beta}$.

5.106 Consider two independent standard normal variables whose joint probability density is

$$\frac{1}{2\pi}e^{-(z_1^2+z_2^2)/2}$$

Under a change to polar coordinates, $z_1 = r\cos(\theta)$, $z_2 = r\sin(\theta)$, we have $r^2 = z_1^2 + z_2^2$ and $dz_1\,dz_2 = r\,dr\,d\theta$, so the joint density of r and θ is

$$r e^{-r^2/2}\frac{1}{2\pi}, \quad 0 < \theta < 2\pi,\ r > 0$$

Show that

(a) r and θ are independent and that θ has a uniform distribution on the interval from 0 to 2π;

(b) $u_1 = \theta/2\pi$ and $u_2 = 1 - e^{-r^2/2}$ have independent uniform distributions;

(c) the relations between (u_1, u_2) and (z_1, z_2) on page 168 hold [note that $1 - u_2$ also has a uniform distribution, so $\ln(u_2)$ can be used in place of $\ln(1-u_2)$].

5.107 The statistical package *MINITAB* has a random number generator. To simulate 5 values from an exponential distribution having mean $\beta = 0.05$, choose

> **Dialog Box:**
>
> **Calc > Random Sample > Exponential**
> Type 5 after **Generate**, *C*1 in **Column** and 0.05 in **Mean**.
> Then click **OK**.

Output:
One call produced the output

0.031949 0.004643 0.030029 0.112834 0.064642

Generate 8 values from the exponential distribution with $\beta = 0.2$.

5.108 The statistical package *MINITAB* has a normal random number generator. To simulate 5 values from a normal distribution having mean 7 and standard deviation 4, and place them in C1, use the commands

> **Dialog Box:**
>
> **Calc > Random Sample > Normal**
> Type 5 after **Generate**, *C*1 in **Column**, 7 in **Mean** and 4 in **standard deviation**
> Click **OK**.

Output:
One call produced the output

5.42137 6.98061 9.41352 7.05932 5.87297

Generate 8 values for a normal variable with $\mu = 123$ and $\sigma^2 = 23.5$.

Do's and Don'ts

Do's

1. Describe the behavior of a continuous random variable X by specifying its probability density function which satisfies

$$f(x) \geq 0 \quad \text{for all } x \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1$$

2. Remember that it is only meaningful to talk about the probability that a continuous random variable X lies in an interval. It is always the case that $P(X = x) = 0$ for every possible value x.
3. Obtain the probability that the value of X will lie in an interval by finding the area under the curve f over the interval.

$$P(X \leq b) = \int_{-\infty}^{b} f(x)\,dx$$

$$= \text{area under the density function to the left of } x = b$$

$$P(a < X \leq b) = \int_{a}^{b} f(x)\,dx$$

$$= \text{area under the density function between } x = a \quad \text{and} \quad x = b$$

4. Summarize a probability density of the continuous random variable X by its

$$\text{mean:} \quad \mu = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$\text{variance: } \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx$$

5. When X has a normal distribution with mean μ and variance σ^2, obtain the probability of an interval $P(X \leq b)$ by converting the limit b to the standardized value $(b - \mu)/\sigma = z$ and obtaining the probability

$$P\left(Z \leq \frac{b - \mu}{\sigma}\right) = P(X \leq b)$$

from the standard normal table.

6. Use the properties of expectation and variance

$$E(aX + b) = aE(X) + b \quad \text{and} \quad Var(aX + b) = a^2\, Var(X)$$

More generally,

$$E(a_1X_1 + a_2X_2 + b) = a_1E(X_1) + a_2E(X_2) + b$$

and, if X_1 and X_2 are independent,

$$Var(a_1X_1 + a_2X_2 + b) = a_1^2\, Var(X_1) + a_2^2\, Var(X_2)$$

Don'ts

1. Never apply the normal approximation to the binomial

$$Z = \frac{X - np}{\sqrt{np(1 - p)}}$$

when the expected number of successes (or failures) is too small. That is, when either

$$np \quad \text{or} \quad n(1-p) \text{ is 15 or less}$$

2. Don't add variances according to

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2)$$

unless the two random variables are independent or have zero covariance.

3. Don't just assume that data come from a normal distribution. When there are at least 20 to 25 observations, it is good practice to construct a normal scores plot to check this assumption.

Review Exercises

5.109 If the probability density of a random variable is given by

$$f(x) = \begin{cases} k(1-x^2) & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

find the value of k and the probabilities that a random variable having this probability density will take on a value

(a) between 0.1 and 0.2;

(b) greater than 0.5.

5.110 With reference to the preceding exercise, find the corresponding distribution function and use it to determine the probabilities that a random variable having this distribution function will take on a value

(a) less than 0.3;

(b) between 0.4 and 0.6.

5.111 In certain experiments, the error made in determining the density of a silicon compound is a random variable having the probability density

$$f(x) = \begin{cases} 25 & \text{for } -0.02 < x < 0.02 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probabilities that such an error will be

(a) between −0.03 and 0.04;

(b) between −0.005 and 0.005.

5.112 Find μ and σ^2 for the probability density of Exercise 5.109.

5.113 If a random variable has the standard normal distribution, find the probability that it will take on a value

(a) between 0 and 2.50;

(b) between 1.22 and 2.35;

(c) between −1.33 and −0.33;

(d) between −1.60 and 1.80.

5.114 The burning time of an experimental rocket is a random variable having the normal distribution with $\mu = 4.76$ seconds and $\sigma = 0.04$ second. What is the probability that this kind of rocket will burn

(a) less than 4.66 seconds;

(b) more than 4.80 seconds;

(c) anywhere from 4.70 to 4.82 seconds?

5.115 Verify that

(a) $z_{0.10} = 1.28$;

(b) $z_{0.001} = 3.09$.

5.116 Referring to Exercise 5.28, find the *quartiles* of the normal distribution with $\mu = 102$ and $\sigma = 27$.

5.117 The probability density shown in Figure 5.20 is the log-normal distribution with $\alpha = 8.85$ and $\beta = 1.03$. Find the probability that

(a) the interrequest time is more than 200 microseconds;

(b) the interrequest time is less than 300 microseconds.

5.118 The probability density shown in Figure 5.22 is the exponential distribution

$$f(x) = \begin{cases} 0.25e^{-0.25x} & 0 < x \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability that

(a) the time to observe a particle is more than 200 microseconds;

(b) the time to observe a particle is less than 10 microseconds.

5.119 Referring to the normal scores in Exercise 5.101, construct a normal scores plot of the suspended solids data in Exercise 2.68.

5.120 Referring to the normal scores in Exercise 5.101, construct a normal scores plot of the velocity of light data in Exercise 2.66.

5.121 If n salespeople are employed in a door-to-door selling campaign, the gross sales volume in thousands of dollars may be regarded as a random variable having

the gamma distribution with $\alpha = 100\sqrt{n}$ and $\beta = \frac{1}{2}$. If the sales costs are \$5,000 per salesperson, how many salespeople should be employed to maximize the expected profit?

5.122 A mechanical engineer models the bending strength of a support beam in a transmission tower as a random variable having the Weibull distribution with $\alpha = 0.02$ and $\beta = 3.0$. What is the probability that the beam can support a load of 4.5?

5.123 Let the times to breakdown for the processors of a parallel processing machine have joint density

$$f(x, y) = \begin{cases} 0.04\, e^{-0.2x-0.2y} & \text{for } x > 0,\ y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

where x is the time for the first processor and y is the time for the second. Find

(a) the marginal distributions and their means;

(b) the expected value of the random variable whose values are given by $g(x, y) = x + y$.

(c) Verify in this example that the mean of a sum is the sum of the means.

5.124 Two random variables are independent and each has a binomial distribution with success probability 0.7 and 2 trials.

(a) Find the joint probability distribution.

(b) Find the probability that the second random variable is greater than the first.

5.125 If X_1 has mean -5 and variance 3 while X_2 has mean 1 and variance 4, and the two are independent, find

(a) $E(3X_1 + 5X_2 + 2)$;

(b) $Var(3X_1 + 5X_2 + 2)$.

5.126 Let $X_1, X_2, \ldots, X_{30}$ be independent and let each have the same marginal distribution with mean -5 and variance 2. Find

(a) $E(X_1 + X_2 + \cdots + X_{30})$;

(b) $Var(X_1 + X_2 + \cdots + X_{30})$.

5.127 Refer to the example of scanners on page 129. The maximum attenuation has a normal distribution with mean 10.1 dB and standard deviation 2.7 dB.

(a) What proportion of the products has maximum attenuation less than 6 dB?

(b) What proportion of the products has maximum attenuation between 6 dB and 14 dB?

5.128 Refer to the heights of pillars in the example on page 15. The variation in the population of heights of pillars can be modeled as a normal distribution with mean 306.6 nm and standard deviation 37.0 nm.

(a) For a pillar selected at random, what is the probability that its height is greater than 350 nm?

(b) According to the normal model, what proportion of all existing pillars has heights greater than 350 nm? Explain your answer.

(c) What proportion of the pillars has heights between 270 nm and 350 nm?

Summary of Distributions

The formulas for the discrete and continuous distributions, together with their means, variances, and moment generating functions, are given in Table 5.2(a) and (b).

Key Terms

Beta distribution 141
Circular normal distribution 157
Conditional probability density 150
Conditional probability distribution 146
Continuity correction 131
Covariance 152
Cumulative distribution function 122
Distribution function 122
Expectation 151
Exponential distribution 140
Gamma distribution 139
Gamma function 139
Independence 146
Joint cumulative distribution function 148
Joint marginal density 149
Joint probability density 147
Joint probability distribution 145
kth moment about the mean 123
kth moment about the origin 123
Log-normal distribution 136
Marginal density 149
Marginal probability distribution 146
Mean 123
Moment generating function 157
Normal distribution 125
Normal probability density 125
Normal scores 163
Normal scores plot 163
Probability density 121
Probability density function 121
Simulation 167
Standard deviation 124
Standard normal distribution 126
Standardized random variable 128
Uniform distribution 135
Variance 124
Waiting time 140
Weibull distribution 142
z-score 128

Table 5.2(a) Discrete distributions

Distribution	Probability Distribution $f(x)$	Mean	Variance	Moment Generating Function
Binomial $b(x;n,p)$	$\binom{n}{x} p^x (1-p)^{n-x}, \quad x=0,1,\ldots,n$	np	$np(1-p)$	$(pe^t+1-p)^n$
Geometric $g(x;p)$	$p(1-p)^{x-1}, \quad x=1,2,\ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
Hypergeometric $h(x;n,a,N)$	$\binom{a}{x}\binom{N-a}{n-x} \Big/ \binom{N}{n}$ $x=0,1,2,\ldots,\min(N-a,n)$	$n\frac{a}{N}$ $p=a/N$	$n\frac{a}{N}\left(1-\frac{a}{N}\right)\frac{(N-n)}{(N-1)}$	complicated
Poisson $f(x;\lambda)$	$\frac{\lambda^x e^{-\lambda}}{x!}, \quad x=0,1,\ldots$	λ	λ	$e^{\lambda(e^t-1)}$
Negative binomial	$\binom{x-1}{r-1} p^r q^{x-r}, \quad x=r,r+1,\ldots$	r/p	$r(1-p)/p^2$	$\left(\frac{pe^t}{1-(1-p)e^t}\right)^r$

Table 5.2(b) Continuous distributions

Distribution	Probability Density Function $f(x)$	Mean	Variance	Moment Generating Function
Uniform	$\frac{1}{\beta-\alpha}, \quad \alpha<x<\beta$	$\frac{\alpha+\beta}{2}$	$\frac{(\beta-\alpha)^2}{12}$	$\frac{e^{\beta t}-e^{\alpha t}}{(\beta-\alpha)t}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty<x<\infty$	μ	σ^2	$e^{\mu t+\frac{1}{2}\sigma^2 t^2}$
Exponential	$\frac{1}{\beta}e^{-x/\beta}, \quad 0\le x<\infty$	β	β^2	$\frac{1}{1-\beta t}, \quad t<\frac{1}{\beta}$
Lognormal	$\frac{1}{\sqrt{2\pi}\beta}x^{-1}e^{-(\ln x-\alpha)^2/2\beta^2}, \quad 0<x<\infty$	$e^{\alpha+\beta^2/2}$	$(e^{\beta^2}-1)e^{2\alpha+\beta^2}$	complicated
Gamma	$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}, \quad 0<x<\infty$	$\alpha\beta$	$\alpha\beta^2$	$\frac{1}{(1-\beta t)^\alpha}, \quad t<\frac{1}{\beta}$
Beta	$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \quad 0<x<1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	complicated
Weibull	$\alpha\beta x^{\beta-1}e^{-\alpha x^\beta}, \quad 0<x$	$\alpha^{-1/\beta}\Gamma(1+\frac{1}{\beta})$	$\alpha^{-2/\beta}\left\{\Gamma(1+\frac{2}{\beta})-\left[\Gamma(1+\frac{1}{\beta})\right]^2\right\}$	complicated

CHAPTER

SAMPLING DISTRIBUTIONS

6

In most of the inference methods we shall study in this book, it will be assumed that we are dealing with a particular kind of sample called a *random sample*. This attention to random samples, which we discuss in Section 6.1, is due to their permitting valid, or logical, generalizations from sample data. Then, in Sections 6.2 through 6.4, we see how certain statistics (that is, certain quantities determined from samples) can be expected to vary from sample to sample. Section 6.5 connects the sampling distributions arising from normal distributions. Techniques for deriving sampling distributions are described in Sections 6.6 and 6.7. The concept of a sampling distribution—the distribution of a statistic calculated on the basis of a random sample—is basic to all of statistical inference.

CHAPTER OUTLINE

6.1 Populations and Samples 175

6.2 The Sampling Distribution of the Mean (σ known) 179

6.3 The Sampling Distribution of the Mean (σ unknown) 187

6.4 The Sampling Distribution of the Variance 189

6.5 Representations of the Normal Theory Distributions 192

6.6 The Moment Generating Function Method to Obtain Distributions 194

6.7 Transformation Methods to Obtain Distributions 197

Review Exercises 202

Key Terms 203

6.1 Populations and Samples

Usage of the term **population** in statistics is a carryover from the days when statistics was applied mainly to sociological and economic phenomena. Recall from Chapter 1 that today it is applied to sets or collections of objects, actual or conceptual, and mainly to sets of numbers, measurements, or observations under investigation. For example, if we are interested in determining the average number of television sets per household in the United States, the totality of these numbers of sets, one for each household, constitutes the population for this study. Similarly, the population from which inspectors draw a sample to determine some quality characteristic of a manufactured product may be the corresponding measurements for all units in a given lot; depending on the objectives of the inspection, it may also consist of the corresponding measurements for all units that may conceivably be manufactured.

In some cases, such as the above example concerning the number of television sets per household, the population is **finite**. In other cases, such as the determination of some characteristic of all units past, present, and future that might conceivably be manufactured by a given process, it is convenient to think of the population as **infinite**. Similarly, we look upon the results obtained in an unending series of flips of a coin as a sample from the hypothetically infinite population consisting of all conceivably possible flips of the coin.

Populations are often described by the distribution of their values. It is common practice to refer to a population in terms of its corresponding probability distribution or frequency distribution. For example, we may refer to a fixed number of flips of a coin as a sample from a "binomial population" or to certain measurements as a sample from a "normal population." Hereafter, when referring to a "population $f(x)$" we shall mean a population described by a frequency distribution, a probability distribution, or a density $f(x)$.

If a population is infinite, it is impossible to observe all its values, and even if it is finite it may be impractical or uneconomical to observe it in its entirety. Thus, it is usually necessary to use a **sample**, a part of a population, and infer from it results pertaining to the entire population. Clearly, such results can be useful only if the sample is in some way "representative" of the population. It would be

unreasonable, for instance, to expect useful generalizations about the population of family incomes in the United States in the year 2010 on the basis of data pertaining to home owners only. Similarly, we can hardly expect reasonable generalizations about the performance of a tire if it is tested only on smooth roads. To assure that a sample is representative of the population from which it is obtained, and to provide a framework for the application of probability theory to problems of sampling, we shall limit our discussion to **random samples**. Before a random sample of size n is selected, the observations are modeled as the random variables $X_1, X_2, \ldots, X_n$. For sampling from finite populations, random samples are defined as follows:

Random sample (finite population)

A set of observations $X_1, X_2, \ldots, X_n$ constitutes a random sample of size n from a finite population of size N, if its values are chosen so that each subset of n of the N elements of the population has the same probability of being selected.

Note that this definition of randomness pertains essentially to the manner in which the sample values are selected. This holds also for the following definition of a random sample from an infinite population:

Random sample (infinite population)

A set of observations $X_1, X_2, \ldots, X_n$ constitutes a random sample of size n from the infinite population $f(x)$ if

1. Each X_i is a random variable whose distribution is given by $f(x)$.
2. These n random variables are independent.

We also apply the term **random sample** to the set of observed values $x_1, x_2, \ldots, x_n$ of the random variables. The lower case distinguishes the realization of a random sample from the upper case, which represents the random variables before they are observed.

There are several ways of assuring the selection of a sample that is at least approximately random. When dealing with a finite population, we can serially number the elements of the population and then select a sample with the aid of a table of random digits (see discussion on page 8). For instance, if a population has $N = 500$ elements and we wish to select a random sample of size $n = 10$, we can use three arbitrarily selected columns of Table 7 to obtain 10 different three-digit numbers less than or equal to 500, which will then serve as the serial numbers of the elements to be included in the sample.

When the population size is large, the use of random numbers can become very laborious and at times practically impossible. For instance, if a sample of five cartons of canned peaches is to be chosen for inspection from among the many thousands stored in a warehouse, one can hardly expect to number all the cartons, make a selection with the use of random numbers, and then pull out the ones that were chosen. In a situation like this, one really has very little choice but to make the selection relatively haphazard, hoping that this will not seriously violate the assumption of randomness which is basic to most statistical theory.

When dealing with infinite populations, the situation is somewhat different since we cannot physically number the elements of the population; but efforts should be made to approach conditions of randomness by the use of artificial devices. For example, in selecting a sample from a production line we may be able to approximate conditions of randomness by choosing one unit each half hour; when tossing a coin we can try to flip it in such a way that neither side is intentionally favored; and

so forth. The proper use of artificial or mechanical devices for selecting random samples is always preferable to human judgment, as it is extremely difficult to avoid unconscious biases when making almost any kind of selection.

Even with the careful choice of artificial devices, it is all too easy to commit gross errors in the selection of a random sample. To illustrate some of these pitfalls, suppose we have the task of selecting logs being fed into a sawmill by a constant-speed conveyor belt, for the purpose of obtaining a random sample of their lengths. One sampling device, which at first sight would seem to assure randomness, consists of measuring the logs which pass a given point at the end of a certain number of 10-minute intervals. However, further thought reveals that this method of selection favors the longer logs, since they require more time to pass the given point. Thus, the sample is not random since the longer logs have a better chance of being included. Another common mistake in selecting a sample is that of sampling from the wrong population or from a poorly specified population. As we have pointed out earlier, we would hardly get a sample from which we could generalize about family incomes in the United States if we limited our sample to home owners. Similarly, if we wanted to determine the effect of vibrations on a structural member, we should be careful to delineate the frequency band of vibrations that is of relevance, and to vibrate test specimens only at frequencies selected randomly from this band.

EXAMPLE **Selecting where to sample in an area that may be contaminated**

In many environmental cleanup studies, engineers are faced with the problem of evaluating the status of land areas or bodies of water. It is not always easy to collect a representative sample where the observations can be treated as independent and from the same distribution. To illustrate some of the key issues, consider sampling from a contaminated area in City C that covers a city block. Locations must be selected for taking the soil samples that will then be analyzed for the presence of heavy metals.

Solution One recommended approach for homogeneous land areas is to sample according to a rectangular grid as shown in Figure 6.1(a). This grid could even be randomly placed over the area.

However, more was known about this area. At the time the pollution occurred, a smelter was located in the position indicted by the shaded area in Figure 6.1(b). The soil contamination by heavy metals is definitely not homogeneous! Materials from the smelter flowed toward the north side of the smelter. This area is definitely a "hot spot." The average amount of heavy metals, obtained by averaging the measurements from all locations, is not representative of the contamination problem. Using the average from the whole area as a summary description downplays the seriousness of the contamination around the smelter.

The smelter runoff area should be treated as a separate population. Soil should be collected from at least two locations within the area of this suspected hot spot.

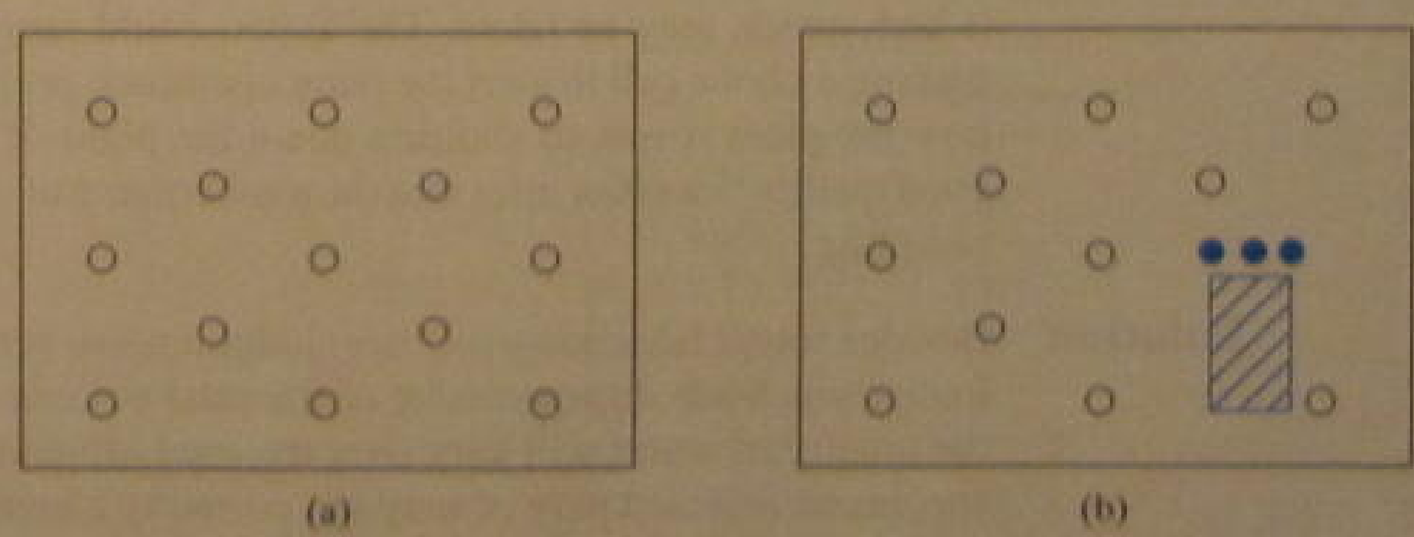


Figure 6.1
Choosing where to sample

We suggest three sites indicated by solid circles. The corresponding measurements of heavy metals should be used to describe only that area.

While sampling the "hot spot" separately might seem like the obvious approach, the owners of the site in the situation on which the example is based wanted to collect data from more locations throughout the whole block. This, they hoped, would lower the average amount of heavy metals enough so that only simple cleanup tactics would be required rather than resorting to the expensive solution of trucking out the soil. ■

EXAMPLE **Always replicate at least one key measurement**

Refer to the environmental cleanup study in the previous example. Explain why soil samples should be taken in at least two locations within the hot spot, rather than just doing two chemical analyses on essentially a single soil sample.

Solution One major aspect of understanding the contamination problem is to evaluate the condition of the known hot spot. Even within the hot spot, the amount of contamination could vary considerably because of, for instance, any particular location's position relative to the old stream of runoff from the smelter. We strongly recommend that soil be collected from at least two soil samples—not just that two chemical analyses be performed on soil from a single location. This approach will provide measurements that can also be used to estimate the total amount of variability in the measurements from the "hot spot."

By repeating the chemical analysis on soil from a single location, we could estimate the variability in the chemical analysis and possibly identify an outlier that might suggest a faulty chemical analysis. However, no matter how many times we repeated the chemical analysis, we would not know how much variation to expect if we sampled soil from the same hot spot but several feet from the first sample. That is, nearby locations are likely to be very much alike but those farther apart are less alike. We need to sample from at least two different locations to determine the degree of homogeneity within the hot spot. ■

The purpose of most statistical investigations is to generalize from information contained in random samples about the population from which the samples were obtained. In particular, we are usually concerned with the problem of making inferences about the **parameters** of populations, such as the mean μ or the standard deviation σ. In making such inferences, we use **statistics** such as $\bar{x}$ and s, namely quantities calculated on the basis of sample observations. In practice, the term *statistic* is also applied to the corresponding random variables.

EXAMPLE **Sample-to-sample variation must be understood to accurately assess total water quality**

The quality of water leaving a plant must be maintained. It is monitored by taking a tiny volume of water called a test specimen. If the quality of the water in the specimen is bad, action may be taken. The action could range from taking more specimens, making a phone call to alert the plant operators, preparing a written report, changing how the plant is run, to shutting down the plant. If the water in the specimen is of good quality, we often infer that the total volume of discharge is satisfactory. Discuss sampling.

Solution Besides sound laboratory practice, judgments depend crucially on getting test specimens which are representative of the total volume of effluent discharged. Because the quality of water will vary over the total effluent at any one time, the actual test specimens selected may or may not correctly convey the water quality.

The key idea in this discussion is that of sampling variability. Not all choices of actual test specimens will produce the same values for water quality or even a correct appraisal of quality. This variability can be overcome, to a greater or lesser extent, by taking a large enough number of test specimens. The results from examining enough test specimens should then quite accurately reflect water quality most of the time. ■

Since the selection of a random sample is controlled largely by chance, so are the values we obtain for statistics. The remainder of this chapter will be devoted to sampling distributions, namely, to distributions which describe the chance fluctuations of statistics calculated on the basis of random samples.

6.2 The Sampling Distribution of the Mean (σ known)

Suppose that a random sample of n observations has been taken from some population and that $\bar{x}$ has been computed, say, to estimate the mean of the population. It should be clear that if we took a second random sample of size n from the population, it would be quite unreasonable to expect the identical value for $\bar{x}$, and if we took several more samples, probably no two of the $\bar{x}$'s would be alike. The differences among such $\bar{x}$'s are generally attributed to chance, and this raises important questions concerning their distribution, specifically concerning the extent of their chance fluctuations.

To approach this question experimentally, suppose that 50 random samples of size $n = 10$ are to be taken from a population having the **discrete uniform distribution**

$$f(x) = \begin{cases} \dfrac{1}{10} & \text{for } x = 0, 1, 2, \ldots, 9 \\ 0 & \text{elsewhere} \end{cases}$$

Sampling is with replacement, so to speak, so that we are sampling from an infinite population. A convenient way of obtaining these samples is to use a table of random digits, like Table 7, letting each sample consist of 10 consecutive digits in arbitrarily chosen rows or columns. Actually proceeding in this way, we get 50 samples whose means are

4.4	3.2	5.0	3.5	4.1	4.4	3.6	6.5	5.3	4.4
3.1	5.3	3.8	4.3	3.3	5.0	4.9	4.8	3.1	5.3
3.0	3.0	4.6	5.8	4.6	4.0	3.7	5.2	3.7	3.8
5.3	5.5	4.8	6.4	4.9	6.5	3.5	4.5	4.9	5.3
3.6	2.7	4.0	5.0	2.6	4.2	4.4	5.6	4.7	4.3

Grouping these means into a distribution with the classes [2.0, 3.0), [3.0, 4.0), [4.0, 5.0), [5.0, 6.0), and [6.0, 7.0), where the left endpoint is included, we get

$\bar{x}$	Frequency
[2.0, 3.0)	2
[3.0, 4.0)	14
[4.0, 5.0)	19
[5.0, 6.0)	12
[6.0, 7.0)	3
	50

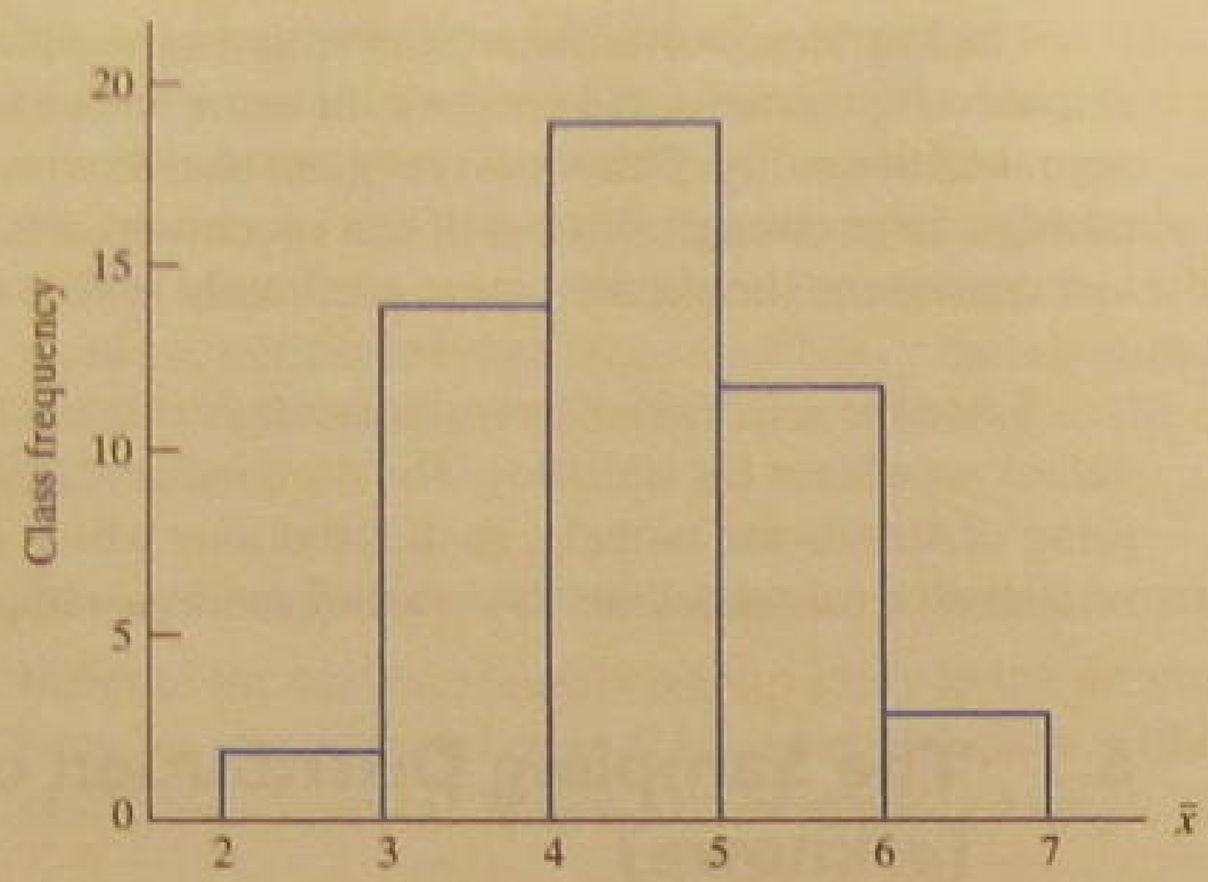


Figure 6.2
Experimental sampling distribution of the mean

and it is apparent from this distribution as well as its histogram shown in Figure 6.2 that the distribution of the means is fairly **bell-shaped**, even though the population itself has a uniform distribution. This raises the question whether our result is typical of what we might expect; that is, whether we would get similar distributions if we repeated the experiment again and again.

To answer this kind of question, we shall have to investigate the **theoretical sampling distribution** of the mean which, for the given example, provides us with the probabilities of getting means in the interval [2.0, 3.0) or in [3.0, 4.0), . . . , [6.0, 7.0) and perhaps values less than 2.0 or greater than or equal to 7.0. Although we could evaluate these probabilities for this particular example, it is usually sufficient to refer to some general theorems concerning sampling distributions. The first of these gives expressions for the mean $\mu_{\overline{X}}$ and the variance $\sigma^2_{\overline{X}}$ of sampling distributions of the mean $\overline{X}$.

Formulas for $\mu_{\overline{X}}$ and $\sigma^2_{\overline{X}}$

Theorem 6.1 If a random sample of size n is taken from a population having the mean μ and the variance σ^2, then $\overline{X}$ is a random variable whose distribution has the mean μ.

For samples from infinite populations the variance of this distribution is $\dfrac{\sigma^2}{n}$.

For samples from a finite population of size N the variance is $\dfrac{\sigma^2}{n} \cdot \dfrac{N-n}{N-1}$.

The result for infinite populations was established in the example on page 155 using the properties of expectation. Alternatively, we now prove that $\mu_{\overline{X}} = \mu$ for the continuous case directly starting from the definition on page 152.

$$\mu_{\overline{X}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \sum_{i=1}^{n} \frac{x_i}{n} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_i\, f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$

where $f(x_1, x_2, \ldots, x_n)$ is the joint density of the random variables which constitute the random sample. Using the assumption of a random sample, we can write

$$f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \ldots f(x_n)$$

and we now have

$$\mu_{\overline{X}} = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} f(x_1)\, dx_1 \ldots \int_{-\infty}^{\infty} x_i\, f(x_i)\, dx_i \ldots \int_{-\infty}^{\infty} f(x_n)\, dx_n$$

Since each integral except the one with the integrand $x_i\, f(x_i)$ equals 1, and the one with the integrand $x_i\, f(x_i)$ equals μ, we finally obtain

$$\mu_{\overline{X}} = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

and this completes the proof. (For the discrete case the proof follows the same steps, with integral signs replaced by $\sum$'s.)

To prove that $\sigma^2_{\overline{X}} = \sigma^2/n$ for the continuous case, we shall make the simplifying assumption that $\mu = 0$, which does not involve any loss of generality as the reader will be asked to show in Exercise 6.18. Using the definition on page 152, we thus have

$$\sigma^2_{\overline{X}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \overline{x}^2\, f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$

and making use of the fact that

$$\overline{x}^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i^2 + \sum\sum_{i\neq j} x_i\, x_j\right)$$

we obtain

$$\sigma^2_{\overline{X}} = \frac{1}{n^2}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_i^2\, f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$
$$+ \frac{1}{n^2}\sum\sum_{i\neq j}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_i\, x_j\, f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$

where $\sum\sum_{i\neq j}$ extends over all i and j from 1 to n, not including the terms where $i = j$. Again using the fact that

$$f(x_1, x_2, \ldots, x_n) = f(x_1)\, f(x_2) \ldots f(x_n)$$

we can write each of the preceding multiple integrals as a product of simple integrals, where each integral with integrand $f(x)$ equals 1. We thus obtain

$$\sigma^2_{\overline{X}} = \frac{1}{n^2}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x_i^2\, f(x_i)\, dx_i + \frac{1}{n^2}\sum\sum_{i\neq j}\int_{-\infty}^{\infty} x_i\, f(x_i)\, dx_i$$
$$\times \int_{-\infty}^{\infty} x_j\, f(x_j)\, dx_j$$

and since each integral in the first sum equals σ^2 while each integral in the second sum equals 0, we finally have

$$\sigma^2_{\overline{X}} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

This completes the proof of the second part of the theorem. We shall not prove the corresponding result for random samples from finite populations. But it should be noted that in the resulting formula for $\sigma^2_{\overline{X}}$ the factor

$$\frac{N-n}{N-1}$$

often called the **finite population correction factor**, is close to 1 (and can be omitted for most practical purposes) unless the sample constitutes a substantial portion of the population.

EXAMPLE **Calculating a finite population correction factor**

Find the value of the finite population correction factor for $n = 10$ and $N = 1{,}000$.

Solution

$$\frac{1{,}000 - 10}{1{,}000 - 1} = 0.991$$

■

Although it should not come as a surprise that $\mu_{\overline{X}} = \mu$, the fact that $\sigma^2_{\overline{X}} = \sigma^2/n$ for random samples from infinite populations is interesting and important. To point out its implications, let us apply Chebyshev's theorem to the sampling distribution of the mean, substituting $\overline{X}$ for X and $\sigma/\sqrt{n}$ for σ in the formula for the alternate form of the theorem (see page 101). We thus obtain

$$P\left(|\overline{X} - \mu| < \frac{k\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}$$

and, letting $k\sigma/\sqrt{n} = \varepsilon$, we get

$$P(|\overline{X} - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

Thus, for any given $\varepsilon > 0$, the probability that $\overline{X}$ differs from μ by less than ε can be made arbitrarily close to 1 by choosing n sufficiently large. In less rigorous language, the larger the sample size, the closer we can expect $\overline{X}$ to be to the mean of the population. In this sense we can say that the mean becomes more and more reliable as an estimate of μ as the sample size is increased. This result—that $\overline{X}$ becomes arbitrarily close to μ with arbitrarily high probability—is called the **law of large numbers**.

Law of large numbers

Theorem 6.2 Let $X_1, X_2, \ldots, X_n$ be independent random variables each having the same mean μ and variance σ^2. Then

$$P(|\overline{X} - \mu| > \varepsilon) \to 0 \quad \text{as } n \to \infty$$

As the sample size increases, unboundedly, the probability that the sample mean differs from the population mean μ, by more than an arbitrary amount ε, converges to zero.

EXAMPLE **Law of large numbers and long-run relative frequency**

Consider an experiment where a specified event A has probability p of occurring. Suppose that, when the experiment is repeated n times, outcomes from different trials are independent. Show that

$$\text{relative frequency of } A = \frac{\text{number of times } A \text{ occurs in } n \text{ trials}}{n}$$

becomes arbitrarily close to p, with arbitrarily high probability, as the number of times the experiment is repeated grows unboundedly.

Solution We can define n random variables $X_1, X_2, \ldots, X_n$ where $X_i = 1$ if A occurs on the ith trial and $X_i = 0$ otherwise. The X_i are independent and identically distributed with mean $\mu = p$ and variance $\sigma^2 = p(1-p)$ since $E(X_1^2) = 1^2 \cdot p + 0^2(1-p) = p$.

Then $X_1 + \cdots + X_n$ is the number of times that A occurs in n trials of the experiment and $\overline{X}$ is the relative frequency of A.

We apply the law of large numbers and conclude that, for an arbitrary positive amount ε,

$$P(|\text{relative frequency of } A - p| > \varepsilon) = P(|\overline{X} - p| > \varepsilon) \to 0 \quad \text{as } n \to \infty$$

Beginning with the axioms of probability, we are led to a theorem that determines the long-run relative frequency of an event. ■

The reliability of the mean as an estimate of μ is often measured by $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, also called the **standard error of the mean**. Note that this measure of the reliability of the mean decreases in proportion to the square root of n; for instance, it is necessary to quadruple the size of the sample in order to halve the standard deviation of the sampling distribution of the mean. This also indicates what might be called a "law of diminishing returns" so far as increasing the sample size is concerned. Usually it does not pay to take excessively large samples since the extra labor and expense is not accompanied by a proportional gain in reliability. For instance, if we increase the size of a sample from 25 to 2,500, the errors to which we are exposed are reduced only by a factor of 10.

Let us now return to the experimental sampling distribution on page 179, and let us check how closely its mean and variance correspond to the values we should expect in accordance with Theorem 6.1. Since the population from which the 50 samples of size $n = 10$ were obtained has the mean

$$\mu = \sum_{x=0}^{9} x \cdot \frac{1}{10} = 4.5$$

and the variance

$$\sigma^2 = \sum_{x=0}^{9} (x - 4.5)^2 \frac{1}{10} = 8.25$$

Theorem 6.1 leads us to expect a mean of $\mu_{\overline{X}} = 4.5$ and a variance of $\sigma^2_{\overline{X}} = 8.25/10 = 0.825$. Calculating the mean and the variance from the 50 sample means on page 179, we get $\overline{x}_{\overline{x}} = 4.43$ and $s^2_{\overline{x}} = 0.930$, which are reasonably close to the theoretical values.

Theorem 6.1 provides only partial information about the theoretical sampling distribution of the mean. In general, it is impossible to determine such a distribution exactly without knowledge of the actual form of the population. Even then, it can be quite difficult. But it is possible to find the limiting distribution as $n \to \infty$ of a random variable whose values are closely related to $\overline{X}$, assuming only that the population has a finite variance σ^2. The random variable we are referring to here is the **standardized sample mean**

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

whose values are given by the difference between $\overline{x}$ and μ divided by the standard error of the mean. With reference to this random variable, we can now state the following theorem, called the **central limit theorem:**

Central limit theorem

Theorem 6.3 If $\overline{X}$ is the mean of a random sample of size n taken from a population having the mean μ and the finite variance σ^2, then

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is a random variable whose distribution function approaches that of the standard normal distributions as $n \to \infty$.

The central limit theorem provides a normal distribution that allows us to assign probabilities to intervals of values for $\overline{X}$. Regardless of the form of the population distribution, the distribution of $\overline{X}$ is approximately normal with mean μ and variance σ^2/n whenever n is large. This tendency toward normality is illustrated in Figure 6.3 for a uniform population distribution and an exponential population distribution.

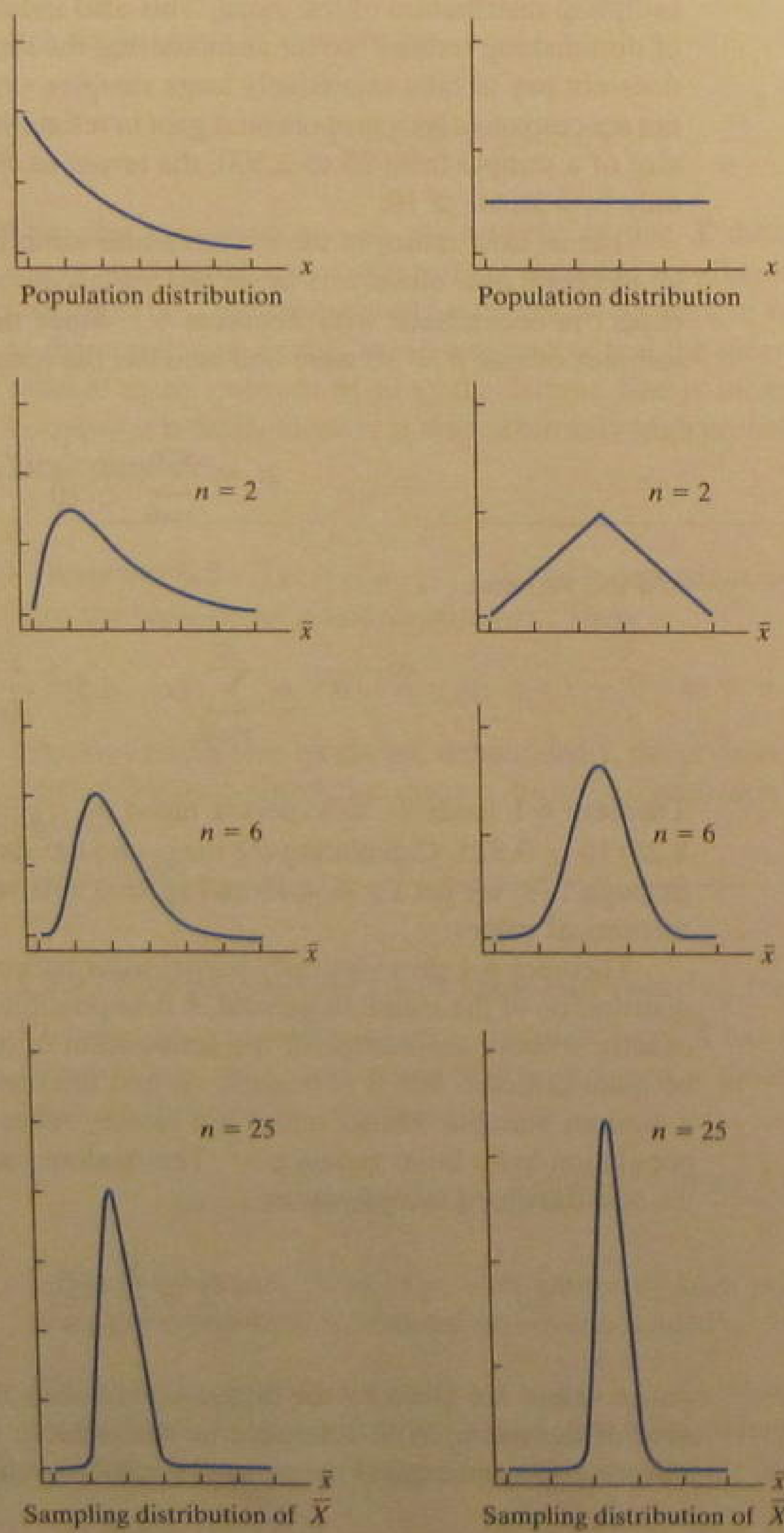


Figure 6.3
An illustration of the approach toward normality for the sampling distribution of $\overline{X}$ as sample size increases

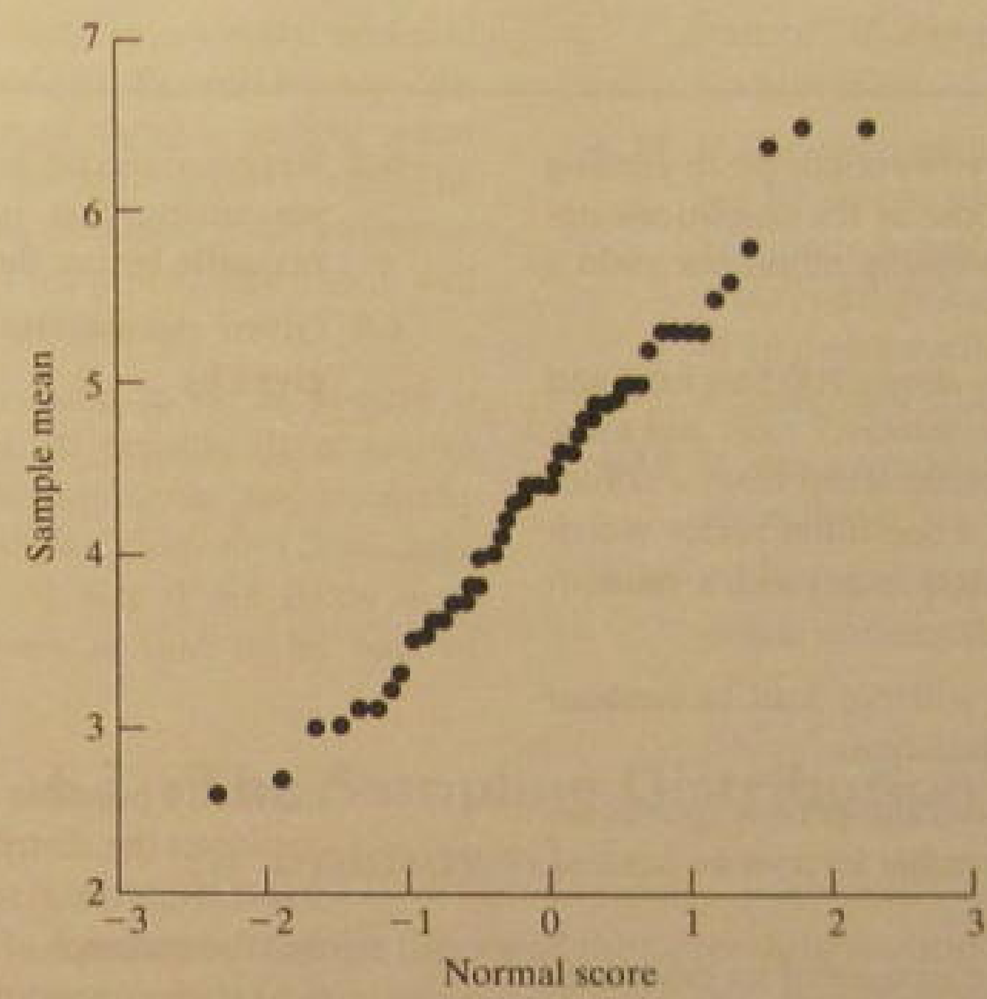


Figure 6.4
Experimental verification of the central limit theorem

Although proving the central limit theorem is beyond the scope of this text, we can obtain experimental verification by constructing a normal scores plot of the 50 sample means on page 179, which were obtained by sampling with replacement from a discrete uniform population. As can be seen in Figure 6.4, the points fall close to a straight line. It seems that, even for $n = 10$, the sampling distribution of the mean for this example has an overall pattern that generally resembles that of a normal distribution.

In practice, **the normal distribution provides an excellent approximation to the sampling distribution of the mean $\bar{X}$ for n as small as 25 or 30**, with hardly any restrictions on the shape of the population. As we saw in our example, the sampling distribution of the mean has the general shape of a normal distribution even for samples of size $n = 10$ from a discrete uniform distribution. **If the random samples come from a normal population, the sampling distribution of the mean is normal regardless of the size of the sample.**

EXAMPLE **Using the central limit theorem to calculate the probability of paint coverage**

If a 1-gallon can of paint covers on the average 513.3 square feet with a standard deviation of 31.5 square feet, what is the probability that the sample mean area covered by a sample of 40 of these 1-gallon cans will be anywhere from 510.0 to 520.0 square feet?

Solution By Theorem 6.3 we shall have to find the normal curve area between

$$z = \frac{510.0 - 513.3}{31.5/\sqrt{40}} = -0.066 \quad \text{and} \quad z = \frac{520.0 - 513.3}{31.5/\sqrt{40}} = 1.34$$

and checking these values in Table 3 we obtain a probability of 0.6553. Note that if $\bar{x}$ turned out to be much less than 513.3, say, less than 500.0, this might cause serious doubt whether the sample actually came from a population having $\mu = 513.3$ and $\sigma = 31.5$; the probability of obtaining such a small value (a z-value less than -2.67) is only 0.0038. ■

Exercises

6.1 An inspector examines every twentieth piece coming off an assembly line. List some of the conditions under which this method of sampling might not yield a random sample.

6.2 Clear plastic wrap is made in sheets 100 feet long and wound in rolls. An inspector selects 7 rolls and measures the clarity of a small piece taken from a corner at the end of each roll. List a condition under which this method of sampling might not yield a random sample.

6.3 Explain why the following will not lead to random samples from the desired populations.

(a) To determine what the average person spends on a vacation, a market researcher interviews passengers on a luxury cruise.

(b) To determine the average income of its graduates 10 years after graduation, the alumni office of a university sent questionnaires in 2010 to all the members of the class of 2000 and based its estimate on the questionnaires returned.

(c) To determine public sentiment about certain import restrictions, an interviewer asks voters: "Do you feel that this unfair practice should be stopped?"

6.4 A market research organization wants to try a new product in 8 of 50 states. Use Table 7 to make this selection.

6.5 How many different samples of size $n = 2$ can be chosen from a finite population of size

(a) $N = 7$;

(b) $N = 24$?

6.6 With reference to Exercise 6.5, what is the probability of each sample in part (a) and the probability of each sample in part (b), if the samples are to be random?

6.7 Take 30 slips of paper and label five each -4 and 4, four each -3 and 3, three each -2 and 2, and two each -1, 0 and 1.

(a) If each slip of paper has the same probability of being drawn, find the probability of getting $-4, -3, -2, -1, 0, 1, 2, 3, 4$ and find the mean and the variance of this distribution.

(b) Draw 50 samples of size 10 from this population, each sample being drawn without replacement, and calculate their means.

(c) Calculate the mean and the variance of the 50 means obtained in part (b).

(d) Compare the results obtained in part (c) with the corresponding values expected according to Theorem 6.1. [Note that μ and σ^2 were obtained in part (a).]

6.8 Repeat Exercise 6.7, but select each sample with replacement; that is, replace each slip of paper and reshuffle before the next one is drawn.

6.9 Given the infinite population whose distribution is given by

x	$f(x)$
1	0.25
2	0.25
3	0.25
4	0.25

list the 16 possible samples of size 2 and use this list to construct the distribution of $\overline{X}$ for random samples of size 2 from the given population. Verify that the mean and the variance of this sampling distribution are identical with corresponding values expected according to Theorem 6.1.

6.10 Suppose that we convert the 50 samples referred to on page 179 into 25 samples of size $n = 20$ by combining the first two, the next two, and so on. Find the means of these samples and calculate their mean and their standard deviation. Compare this mean and this standard deviation with the corresponding values expected in accordance with Theorem 6.1.

6.11 When we sample from an infinite population, what happens to the standard error of the mean if the sample size is

(a) increased from 50 to 200;

(b) increased from 400 to 900;

(c) decreased from 225 to 25;

(d) decreased from 640 to 40?

6.12 What is the value of the finite population correction factor in the formula for $\sigma^2_{\overline{X}}$ when

(a) $n = 5$ and $N = 250$;

(b) $n = 10$ and $N = 500$;

(c) $n = 100$ and $N = 5{,}000$?

6.13 For large sample size n, verify that there is a 50-50 chance that the mean of a random sample from an infinite population with the standard deviation σ will differ from μ by less than $0.6745 \cdot \sigma/\sqrt{n}$. It has been the custom to refer to this quantity as the **probable error of the mean**.

6.14 The mean of a random sample of size $n = 25$ is used to estimate the mean of an infinite population that has standard deviation $\sigma = 2.4$. What can we assert about the probability that the error will be less than 1.2, if we use

(a) Chebyshev's theorem;

(b) the central limit theorem?

6.15 Hard disks for computers must spin evenly, and one departure from level is called roll. The roll for any disk can be modeled as a random variable having mean 0.2250 mm and standard deviation 0.0042 mm. The sample mean roll $\overline{X}$ will be obtained from a random sample of 40 disks. What is the probability that $\overline{X}$ will lie between 0.2245 and 0.2260 mm?

6.16 A wire-bonding process is said to be in control if the mean pull strength is 10 pounds. It is known that the pull-strength measurements are normally distributed with a standard deviation of 1.5 pounds. Periodic random samples of size 4 are taken from this process and the process is said to be "out of control" if a sample mean is less than 7.75 pounds. Comment.

6.17 If the distribution of the weights of all men traveling by air between Dallas and El Paso has a mean of 163 pounds and a standard deviation of 18 pounds, what is the probability that the combined gross weight of 36 men traveling on a plane between these two cities is more than 6,000 pounds?

6.18 If X is a continuous random variable and $Y = X - \mu$, show that $\sigma_Y^2 = \sigma_X^2$.

6.19 Prove that $\mu_{\overline{X}} = \mu$ for random samples from discrete (finite or countably infinite) populations.

6.3 The Sampling Distribution of the Mean (σ unknown)

Application of the theory of the preceding section requires knowledge of the population standard deviation σ. If n is large, this does not pose any problems even when σ is unknown, as it is reasonable in that case to substitute for it the sample standard deviation s. However, when it comes to the random variable whose values are given by

$$\frac{\overline{x} - \mu}{s/\sqrt{n}}$$

very little is known about its exact sampling distribution for small values of n unless we make the assumption that the sample comes from a normal population. Under this assumption, one can prove the following:

A random variable having the t distribution

Theorem 6.4 If $\overline{X}$ is the mean of a random sample of size n taken from a normal population having the mean μ and the variance σ^2, and $S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$, then

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is a random variable having the t distribution with the parameter $v = n - 1$.

The lowercase t notation helps differentiate this important statistic from others. This theorem is more general than Theorem 6.3 in the sense that it does not require knowledge of σ; on the other hand, it is less general than Theorem 6.3 in the sense that it requires the assumption of a normal population.

As can be seen from Figure 6.5, the overall shape of a ***t* distribution** is similar to that of a normal distribution—both are bell-shaped and symmetrical about the

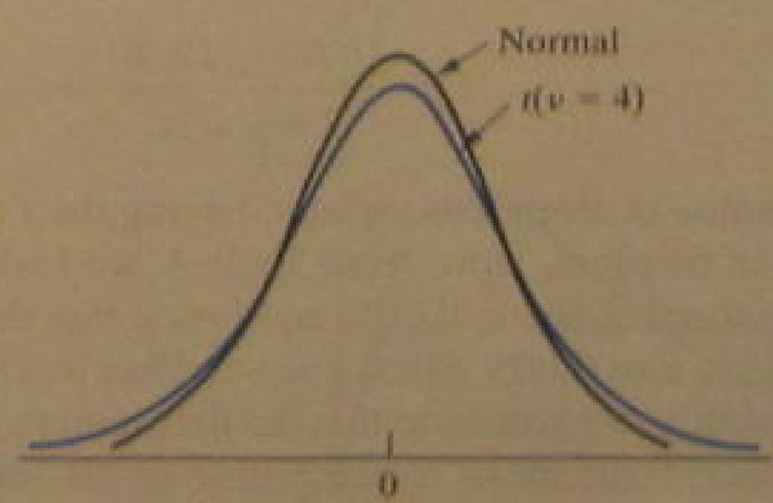


Figure 6.5
t distribution and standard normal distribution

mean. Like the standard normal distribution, the t distribution has the mean 0, but its variance depends on the parameter ν (nu), called the number of **degrees of freedom**. The variance of the t distribution exceeds 1, but it approaches 1 as $n \to \infty$. In fact, it can be shown that the t distribution with ν degrees of freedom approaches the standard normal distribution as $\nu \to \infty$.

Table 4 at the end of the book contains selected values of t_α for various values of ν, where t_α is such that the area under the t distribution to its right is equal to α. In this table the left-hand column contains values of ν, the column headings are areas α in the right-hand tail of the t distribution, and the entries are values of t_α. (See also Figure 6.6.) It is not necessary to tabulate values of t_α for $\alpha > 0.50$, as it follows from the symmetry of the t distribution that $t_{1-\alpha} = -t_\alpha$. Thus, the value of t that corresponds to a left-hand tail area of α is $-t_\alpha$.

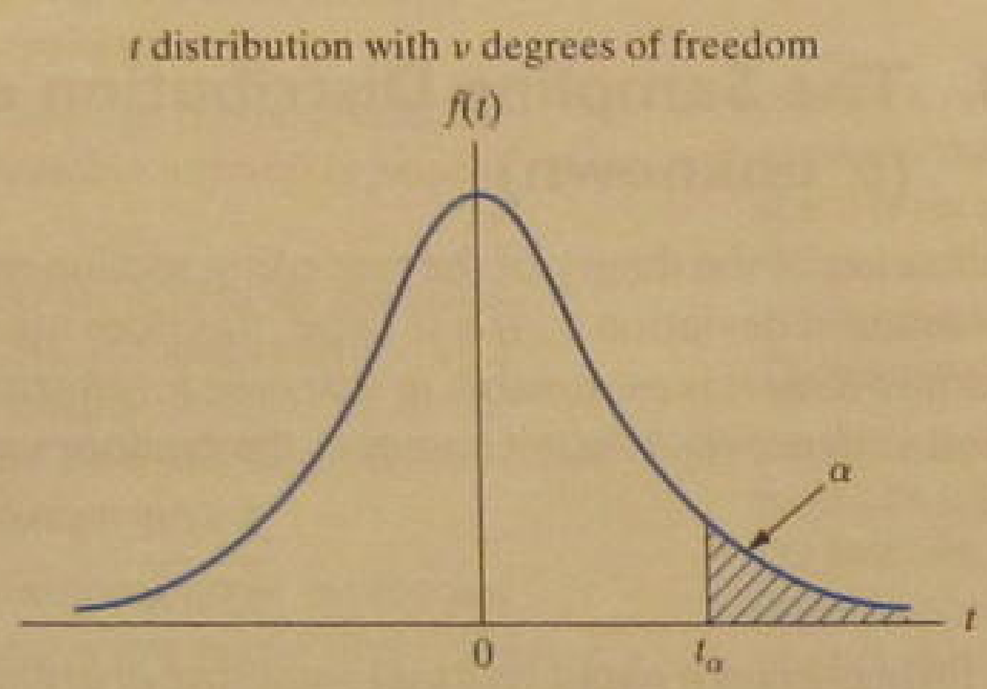


Figure 6.6
Tabulated value of t_α

Note that in the bottom row of Table 4 the entries correspond to the values of z that cut off right-hand tails of area α under the standard normal curve. Using the notation z_α for such a value of z, it can be seen, for example, that $z_{0.025} = 1.96 = t_{0.025}$ for $\nu = \infty$. In fact, observing that the values of t_α for 29 or more degrees of freedom are close to the corresponding values of z_α, we conclude that **the standard normal distribution provides a good approximation to the t distribution for samples of size 30 or more.**

EXAMPLE **Using a probability calculation from the t distribution to refute a claim**

A manufacturer of fuses claims that with a 20% overload, the fuses will blow in 12.40 minutes on the average. To test this claim, a sample of 20 of the fuses was subjected to a 20% overload, and the times it took them to blow had a mean of 10.63 minutes and a standard deviation of 2.48 minutes. If it can be assumed that the data constitute a random sample from a normal population, do they tend to support or refute the manufacturer's claim?

Solution First we calculate

$$t = \frac{10.63 - 12.40}{2.48/\sqrt{20}} = -3.19$$

which is a value of a random variable having the t distribution with $\nu = 20 - 1 = 19$ degrees of freedom. Now, from Table 4, we find that for $\nu = 19$ the probability that t will exceed 2.861 is 0.005, and hence that the probability that t will be less than -2.861 is also 0.005. Since $t = -3.19$ is less than -2.861 and 0.005 is a very small probability, we conclude that the data tend to refute the manufacturer's claim. In all likelihood, the mean blowing time of his fuses with a 20% overload is less than 12.40 minutes. ■

The assumption that the sample must come from a normal population is not so severe a restriction as it may seem. Studies have shown that the distribution of random variable

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is fairly close to a t distribution even for samples from certain nonnormal populations. In practice, it is necessary to make sure primarily that the population from which we are sampling is approximately bell-shaped and not too skewed. A practical way of checking this assumption is to construct a normal scores plot, as described on page 163. (If such a plot shows a distinct curve rather than a straight line, it may be possible to "straighten it out" by transforming the data—say, by taking their logarithms or their square roots, as discussed in Chapter 5, Section 5.13.)

6.4 The Sampling Distribution of the Variance

So far we have discussed only the sampling distribution of the mean, but if we had taken the medians or the standard deviations of the 50 samples mentioned in the example on page 179, we would similarly have obtained experimental sampling distributions of these statistics. In this section we shall be concerned with the theoretical sampling distribution of the sample variance for random samples from normal populations. Since S^2 cannot be negative, we should suspect that this sampling distribution is not a normal curve; in fact, it is related to the gamma distribution (see page 139) with $\alpha = \nu/2$ and $\beta = 2$, called the **chi square distribution**. Specifically, using the square of the Greek letter χ (chi), we have the following theorem.

A random variable having the chi square distribution

Theorem 6.5 If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

is a random variable having the chi square distribution with the parameter $\nu = n - 1$.

Table 5 at the end of the book contains selected values of χ^2_α for various values of ν, again called the number of **degrees of freedom**, where χ^2_α is such that the area under the chi square distribution to its right is equal to α. In this table the left-hand column contains values of ν, the column headings are areas α in the right-hand tail of the chi square distribution, and the entries are values of χ^2_α. (See also Figure 6.7.)

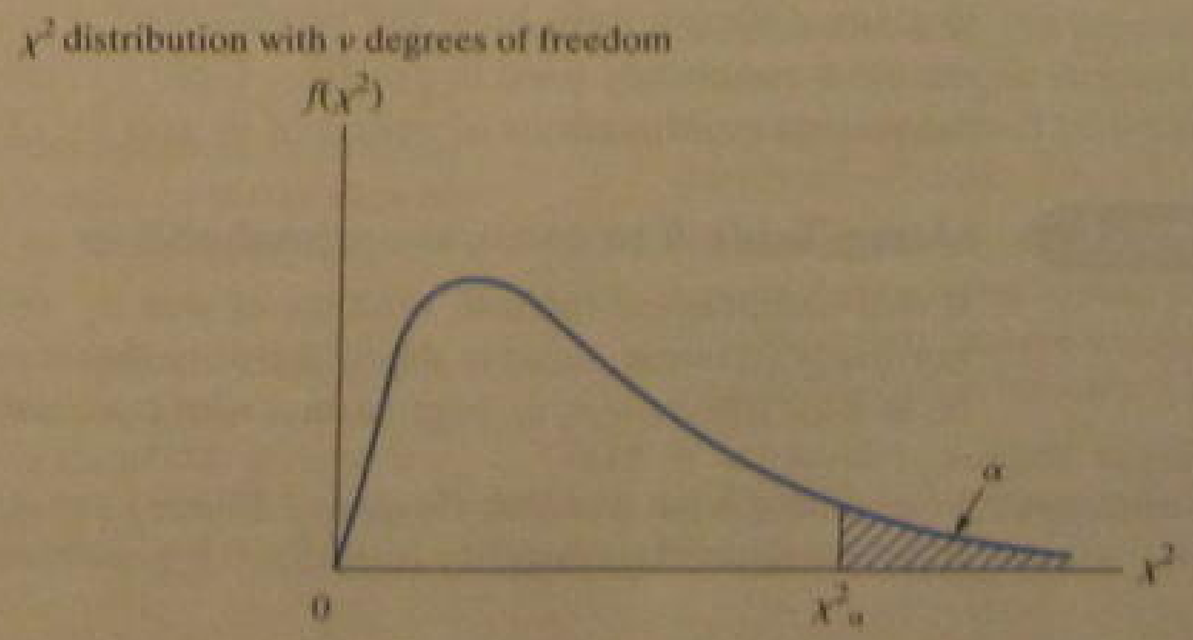


Figure 6.7
Tabulated values of chi square

Unlike the t distribution, it is necessary to tabulate values of χ^2_α for $\alpha > 0.50$, because the chi square distribution is not symmetrical.

EXAMPLE **A probability calculation based on the χ^2 distribution helps monitor variability**

An optical firm purchases glass to be ground into lenses, and it is known from past experience that the variance of the refractive index of this kind of glass is $1.26 \cdot 10^{-4}$. As it is important that the various pieces of glass have nearly the same index of refraction, the firm rejects such a shipment if the sample variance of 20 pieces selected at random exceeds $2.00 \cdot 10^{-4}$. Assuming that the sample values may be looked upon as a random sample from a normal population, what is the probability that a shipment will be rejected even though $\sigma^2 = 1.26 \cdot 10^{-4}$?

Solution Substituting into the formula for the chi square statistic, we get

$$\chi^2 = \frac{19(2.00 \cdot 10^{-4})}{1.26 \cdot 10^{-4}} = 30.2$$

and then we find from Table 5 that for 19 degrees of freedom, $\chi^2_{0.05} = 30.1$. Thus, the probability that a good shipment will erroneously be rejected is less than 0.05. ■

A problem closely related to that of finding the distribution of the sample variance is that a finding the distribution of the ratio of the variances of two independent random samples. This problem is important because it arises in tests in which we want to determine whether two samples come from populations having equal variances. If they do, the two sample variances should be nearly the same; that is, their ratio should be close to 1. To determine whether the ratio of two sample variances is too small or too large, we use the theory given in the following theorem:

A random variable having the F distribution

Theorem 6.6 If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2, respectively, taken from two normal populations having the same variance, then

$$F = \frac{S_1^2}{S_2^2}$$

is a random variable having the F distribution with the parameters $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

The F **distribution** is related to the beta distribution (page 141), and its two parameters, ν_1 and ν_2, are called the **numerator** and **denominator degrees of freedom**. As it would require too large a table to give values of F_α corresponding to many different right-hand tail probabilities α, and since $\alpha = 0.05$ and $\alpha = 0.01$ are most commonly used in practice, Table 6 contains only values $F_{0.05}$ and $F_{0.01}$ for various combinations of values of ν_1 and ν_2. (See also Figure 6.8.)

EXAMPLE **Using Table 6 to evaluate a probability**

If two independent random samples of size $n_1 = 7$ and $n_2 = 13$ are taken from a normal population, what is the probability that the variance of the first sample will be at least three times as large as that of the second sample?

Solution From Table 6 we find that $F_{0.05} = 3.00$ for $\nu_1 = 7 - 1 = 6$ and $\nu_2 = 13 - 1 = 12$; thus, the desired probability is 0.05. ■

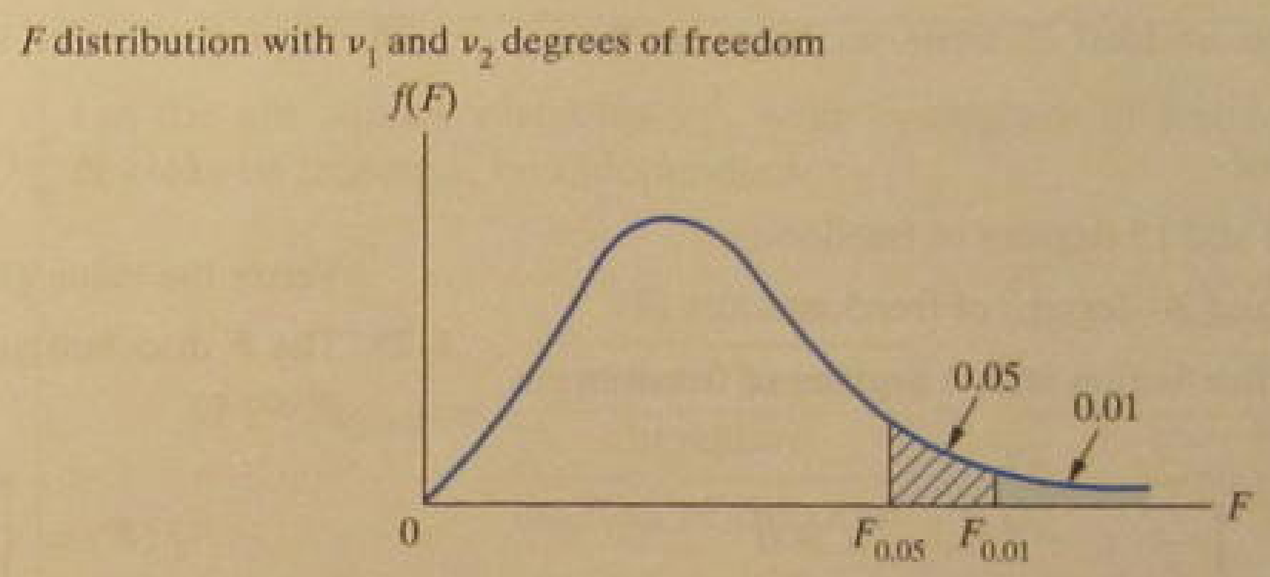


Figure 6.8
Tabulated values of F

It is possible to use Table 6 also to find values of F corresponding to left-hand tail probabilities of 0.05 or 0.01. Writing $F_\alpha(\nu_1, \nu_2)$ for F_α with ν_1 and ν_2 degrees of freedom, we simply use the identity

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)}$$

EXAMPLE **Using Table 6 to find a left-hand tail probability**

Find the value of $F_{0.95}$ (corresponding to a left-hand tail probability of 0.05) for $\nu_1 = 10$ and $\nu_2 = 20$ degrees of freedom.

Solution Making use of the identity and Table 6, we get

$$F_{0.95}(10, 20) = \frac{1}{F_{0.05}(20, 10)} = \frac{1}{2.77} = 0.36$$

■

Note that Theorems 6.4 and 6.5 require the assumption that we are sampling from normal populations. Unlike the situation with the t distribution, deviations from an underlying normal distribution, such as a long tail, may have a serious effect on these sampling distributions. Consequently, it is best to transform to near normality using the approach in Section 5.13 before invoking the sampling distributions in this section.

Exercises

6.20 The tensile strength (1,000 psi) of a new composite can be modeled as a normal distribution. A random sample of size 25 specimens has mean $\bar{x} = 45.3$ and standard deviation $s = 7.9$. Does this information tend to support or refute the claim that the mean of the population is 40.5?

6.21 The following are the times between 6 calls for an ambulance (in a certain city) and the patient's arrival at the hospital: 27, 15, 20, 32, 18, and 26 minutes. Use these figures to judge the reasonableness of the ambulance service's claim that it takes on the average of 20 minutes between the call for an ambulance and the patient's arrival at the hospital.

6.22 A process for making certain bearings is under control if the diameters of the bearings have a mean of 0.5000 cm. What can we say about this process if a sample of 10 of these bearings has a mean diameter of 0.5060 cm and a standard deviation of 0.0040 cm?

6.23 Hard disks for computers must spin evenly, and one departure from level is called pitch. Samples are regularly taken from production and each disk in the sample is placed in test equipment that yields a measurement of pitch. From many samples, it is concluded that the population is normal. The variance is $\sigma^2 = 0.065$ when the process is in control. A sample of size 10 is collected each week. The process will be declared out of control if the sample variance exceeds 0.122. What is the probability it will be declared out of control even though $\sigma^2 = 0.065$?

6.24 A random sample of 10 observations is taken from a normal population having the variance $\sigma^2 = 42.5$. Find the approximate probability of obtaining a sample standard deviation between 3.14 and 8.94.

6.25 If independent random samples of size $n_1 = n_2 = 8$ come from normal populations having the same variance, what is the probability that either sample

variance will be at least 7 times as large as the other?

6.26 Find the values of

(a) $F_{0.95}$ for 12 and 15 degrees of freedom;

(b) $F_{0.99}$ for 6 and 20 degrees of freedom.

6.27 The chi square distribution with 4 degrees of freedom is given by

$$f(x) = \begin{cases} \frac{1}{4} \cdot x \cdot e^{-x/2} & x > 0 \\ 0 & x \le 0 \end{cases}$$

Find the probability that the variance of a random sample of size 5 from a normal population with $\sigma = 12$ will exceed 180.

6.28 The t distribution with 1 degree of freedom is given by

$$f(t) = \frac{1}{\pi}(1 + t^2)^{-1} \qquad -\infty < t < \infty$$

Verify the value given for $t_{0.05}$ for $\nu = 1$ in Table 4.

6.29 The F distribution with 4 and 4 degrees of freedom is given by

$$f(F) = \begin{cases} 6F(1 + F)^{-4} & F > 0 \\ 0 & F \le 0 \end{cases}$$

If random samples of size 5 are taken from two normal populations having the same variance, find the probability that the ratio of the larger to the smaller sample variance will exceed 3.

6.5 Representations of the Normal Theory Distributions

The basic distributions of normal theory can all be defined in terms of independent standard normal random variables. The defining of a new random variable in terms of others is called a **representation**.

Let $Z, Z_1, Z_2, \ldots,$ be independent standard normal random variables with mean 0 and variance 1. First, we define a chi square variable.

Representation of chi square random variable

Let $Z_1, Z_2, \ldots, Z_\nu$ be independent standard normal random variables.

$$\chi^2_\nu = \text{sum of squares of } \nu \text{ independent standard normal variables}$$
$$= \sum_{i=1}^{\nu} Z_i^2 \text{ has a chi square distribution with } \nu \text{ degrees of freedom.}$$

Consider two chi square random variables which have the representations $\chi_1^2 = \sum_{i=1}^{\nu_1} Z_i^2$ and $\chi_2^2 = \sum_{i=\nu_1+1}^{\nu_1+\nu_2} Z_i^2$. Since they depend on different sets of Z_i's, they are independent. Adding these two representations we conclude, as in Exercise 6.33, that the **sum of two independent chi square variables, $\chi_1^2 + \chi_2^2$, has a chi square distribution with degrees of freedom $\nu_1 + \nu_2$.**

Next, since $\chi^2 = \sum_{i=1}^{\nu} Z_i^2$ depends only on $Z_1, Z_2, \ldots, Z_\nu$ and they are independent of Z this χ^2 and Z are independent. We define a t random variable in terms of two independent random variables Z and χ^2.

Representation of t random variable

Let the standard normal Z and chi square (χ^2), having ν degrees of freedom, be independent.

$$t = \frac{\text{standard normal}}{\sqrt{\dfrac{\text{chi square}}{\text{degrees of freedom}}}} = \frac{Z}{\sqrt{\dfrac{\chi^2}{\nu}}} = \frac{Z}{\sqrt{\dfrac{\sum_{i=1}^{\nu} Z_i^2}{\nu}}}$$

has a t distribution with ν degrees of freedom.

We define an F random variable in terms of two independent chi square variables χ_1^2 and χ_2^2 with ν_1 and ν_2 degrees of freedom, respectively.

Representation of F random variable

Let the chi square variables χ_1^2, with ν_1 degrees of freedom, and χ_2^2, with ν_2 degrees of freedom, be independent.

$$F_{\nu_1,\nu_2} = \frac{\dfrac{\text{chi square}}{\text{degrees of freedom}}}{\dfrac{\text{chi square}}{\text{degrees of freedom}}} = \frac{\dfrac{\chi_1^2}{\nu_1}}{\dfrac{\chi_2^2}{\nu_2}} = \frac{\dfrac{\sum_{i=1}^{\nu_1} Z_i^2}{\nu_1}}{\dfrac{\sum_{i=\nu_1+1}^{\nu_1+\nu_2} Z_i^2}{\nu_2}}$$

has an F distribution with (ν_1, ν_2) degrees of freedom.

The basic case arises starting with n independent normal random variables $X_1, X_2, \ldots, X_n$ all having the same mean μ and standard deviation σ. Then

$$Z_i = \frac{X_i - \mu}{\sigma}$$

has a standard normal distribution for each i. It then holds that

$$\sqrt{n}\,\overline{Z} = \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} Z_i = \sqrt{n}\left(\frac{\overline{X} - \mu}{\sigma}\right)$$

has a standard normal distribution.

Next,

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} (Z_i - \overline{Z} + \overline{Z})^2 = \sum_{i=1}^{n} (Z_i - \overline{Z})^2 + n\overline{Z}^2$$

The left-hand side has a chi square distribution with n degrees of freedom. The last term on the right is the square of the standard normal variable $\sqrt{n}\,\overline{Z}$ and so has a chi square distribution with 1 degree of freedom. It can be shown that the two terms on the right-hand side of the equation are independent and that

$$\sum_{i=1}^{n} (Z_i - \overline{Z})^2$$

has a chi square distribution with $n - 1$ degrees of freedom. Since

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} = \sum_{i=1}^{n} (Z_i - \overline{Z})^2$$

we conclude that $(n-1)S^2/\sigma^2$ has a chi square distribution with $n - 1$ degrees of freedom.

EXAMPLE **t^2 has an F distribution**

Let t be distributed as a t distribution with ν degrees of freedom.

(a) Use the representation of t to show that t^2 has an F distribution with $(1, \nu)$ degrees of freedom.

(b) Use part (a) to show that $t_{\alpha/2} = F_\alpha(1, \nu)$.

Solution (a) Using the representation of a t random variable,

$$t^2 = \left(\frac{Z}{\sqrt{\frac{\chi^2}{\nu}}} \right)^2 = \frac{Z^2}{\frac{\chi^2}{\nu}}$$

Since, by the first representation above, Z^2 has a chi square distribution with 1 degree of freedom and it is independent of the denominator, we confirm the representation of the F distribution with $(1, \nu)$ degrees of freedom.

(b)
$$1 - \alpha = P(-t_{\alpha/2} \le t \le t_{\alpha/2}) = P\left(t^2 \le t^2_{\alpha/2}\right)$$

By part (a), $t^2 = F$ so we have

$$1 - \alpha = P\left(F \le t^2_{\alpha/2}\right)$$

Because $t^2_{\alpha/2}$ satisfies the definition of $F_\alpha(1, \nu)$, the two must be equal. ■

Exercises

6.30 Let $Z_1, \ldots, Z_5$ be independent and let each have a standard normal distribution.

(a) Specify the distribution of $Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2$.

(b) Specify the distribution of

$$\frac{Z_1}{\sqrt{\frac{Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2}{4}}}$$

6.31 Let $Z_1, \ldots, Z_6$ be independent and let each have a standard normal distribution. Specify the distribution of

$$\frac{Z_1 - Z_2}{\sqrt{\frac{Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2}{8}}}$$

6.32 Let $Z_1, \ldots, Z_7$ be independent and let each have a standard normal distribution.

(a) Specify the distribution of $Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2$.

(b) Specify the distribution of $Z_5^2 + Z_6^2 + Z_7^2$.

(c) Specify the distribution of the sum of variables in part (a) and part (b).

6.33 Let the chi square variables χ_1^2, with ν_1 degrees of freedom, and χ_2^2, with ν_2 degrees of freedom, be independent. Establish the result on page 192, that their sum is a chi square variable with $\nu_1 + \nu_2$ degrees of freedom.

6.6 The Moment Generating Function Method to Obtain Distributions*

The mgf method is a very convenient tool for obtaining the distribution function of the sum of independent random variables. Let X_1 have mgf $M_1(t)$, X_2 have mgf $M_2(t)$, X_3 have mgf $M_3(t)$, and so on. Then, by independence, the mgf of the sum $X_1 + X_2 + X_3$ is

$$\begin{aligned} M_{X_1+X_2+X_3}(t) &= E(e^{t(X_1+X_2+X_3)}) = E(e^{tX_1}e^{tX_2}e^{tX_3}) \\ &= E(e^{tX_1}) \cdot E(e^{tX_2}) \cdot E(e^{tX_3}) \end{aligned}$$

or

$$M_{X_1+X_2+X_3}(t) = M_1(t) \cdot M_2(t) \cdot M_3(t)$$

For any number of independent random variables, we have the following result.

*This section may be skipped on first reading. Some key sampling distributions are verified.

Moment generating function for sum of n independent random variables

Theorem 6.7 Let $X_1, \ldots, X_n$ be independent random variables and let X_i have moment generating function $M_{X_i}(t)$ for $i = 1, \ldots, n$, where all moment generating functions exist for all $t \leq T$ some $T > 0$. Then the moment generating function of the sum exists for all $t \leq T$ and

$$M_{X_1+X_2+\cdots+X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdots M_{X_n}(t)$$

The mgf of the sum of random variables is the product of the component mgf, under independence. When the product can be identified, we know the distribution of the sum. This argument is called the **moment generating function method**.

EXAMPLE **Sum of n independent normal random variables is normal**

Let X_1 be $N(\mu_1, \sigma_1^2)$, X_2 be $N(\mu_2, \sigma_2^2)$, and X_3 be $N(\mu_3, \sigma_3^2)$, where the three random variables are independent.

(a) Find the distribution of $X_1 + X_2 + X_3$.

(b) Let X_i be $N(\mu_i, \sigma_i^2)$, for $i = 1, 2, \ldots, n$ and let the X_i be independent. Show that the distribution of their sum, $\sum_{i=1}^n X_i$, is normal with

$$\text{mean} = \sum_{i=1}^n \mu_i$$

$$\text{variance} = \sum_{i=1}^n \sigma_i^2$$

Solution **(a)** We know that X_1 has mgf

$$M_1(t) = e^{t\mu_1 + \frac{1}{2}t^2\sigma_1^2}$$

so

$$\begin{aligned} M_{X_1+X_2+X_3}(t) &= e^{t\mu_1+\frac{1}{2}t^2\sigma_1^2} \cdot e^{t\mu_2+\frac{1}{2}t^2\sigma_2^2} \cdot e^{t\mu_3+\frac{1}{2}t^2\sigma_3^2} \\ &= e^{t(\mu_1+\mu_2+\mu_3)+\frac{1}{2}t^2(\sigma_1^2+\sigma_2^2+\sigma_3^2)} \end{aligned}$$

This last form we identify as being an $N(\mu_1 + \mu_2 + \mu_3, \sigma_1^2 + \sigma_2^2 + \sigma_3^2)$. That is, the sum has a normal distribution where the mean is the sum of the component means and the variance is the sum of the variances.

(b)

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n e^{t\mu_i + \frac{1}{2}t^2\sigma_i^2} = e^{t\left(\sum_{i=1}^n \mu_i\right) + \frac{1}{2}t^2\left(\sum_{i=1}^n \sigma_i^2\right)}$$

so the sum has a normal distribution with mean equal the sum of the component means and variance equal to the sum of variances. ■

EXAMPLE **Sum of independent Poisson random variables is Poisson**

Let X_i have a Poisson distribution with parameter λ_i, for $i = 1, 2, \ldots, n$ and let the X_i be independent. Show that the distribution of the their sum, $\sum_{i=1}^n X_i$, is Poisson with parameter

$$\lambda = \sum_{i=1}^n \lambda_i$$

Solution We know that X_i has mgf

$$M_i(t) = e^{-\lambda_i + \lambda_i e^t}$$

Consequently,

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n e^{-\lambda_i + \lambda_i e^t} = e^{\left(\sum_{i=1}^n \lambda_i\right) + \left(\sum_{i=1}^n \lambda_i\right) e^t}$$

This is the mgf of a Poisson distribution with parameter $\sum_{i=1}^n \lambda_i$. ■

EXAMPLE **Sum of chi square random variables is chi square**

Let X_i have a chi square distribution with ν_i degrees of freedom, for $i = 1, 2, \ldots, n$ and let the X_i be independent. Use the moment generating function method to show that the distribution of their sum, $\sum_{i=1}^n X_i$ is chi square with degrees of freedom $\sum_{i=1}^n \nu_i$.

Solution **(a)** We know that X_i has mgf $(1-2t)^{-\nu_i/2}$ so that $\sum_{i=1}^n X_i$ has mgf

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n \frac{1}{(1-2t)^{\nu_i/2}} = \frac{1}{(1-2t)^{\sum_{i=1}^n \nu_i/2}}$$

which we identify as the mgf of a chi square distribution with $\sum_{i=1}^n \nu_i$ degrees of freedom. ■

Exercises

6.34 Let $X_1, X_2, \ldots, X_8$ be 8 independent random variables. Find the moment generating function

$$M_{\sum X_i}(t) = E(e^{t(X_1 + X_2 + \cdots + X_8)})$$

of the sum when X_i has a Poisson distribution with mean

(a) $\lambda_i = 0.5$

(b) $\lambda_i = 0.04$

6.35 Let $X_1, X_2, \ldots, X_5$ be 5 independent random variables. Find the moment generating function

$$M_{\sum X_i}(t) = E(e^{t(X_1 + X_2 + \cdots + X_5)})$$

of the sum when X_i has a gamma distribution with $\alpha_i = 2i$ and $\beta_i = 2$.

6.36 Let X_1, X_2, and X_3 be independent normal variables with

$$E(X_1) = 2 \quad \text{and} \quad \sigma_1^2 = 4$$
$$E(X_2) = 1 \quad \text{and} \quad \sigma_2^2 = 9$$
$$E(X_3) = -1 \quad \text{and} \quad \sigma_3^2 = 1$$

(a) Show that $X_1 + 2X_2 - 3X_3$ has a normal distribution.

(b) Find the mean and variance of the random variable in part (a).

6.37 Refer to Exercise 6.36.

(a) Show that $X_1 - 3X_2 + 2X_3 - 5$ has a normal distribution.

(b) Find the mean and variance of the random variable in part (b).

6.38 Let X_1, X_2, and X_3 be independent normal variables with

$$E(X_1) = -4 \quad \text{and} \quad \sigma_1^2 = 1$$
$$E(X_2) = 0 \quad \text{and} \quad \sigma_2^2 = 4$$
$$E(X_3) = 3 \quad \text{and} \quad \sigma_3^2 = 1$$

(a) Show that $2X_1 - X_2 + 5X_3$ has a normal distribution.

(b) Find the mean and variance of the random variable in part (a).

6.39 Refer to Exercise 6.38.

(a) Show that $7X_1 + X_2 - 2X_3 + 7$ has a normal distribution.

(b) Find the mean and variance of the random variable in part (a).

6.40 Let $X_1, X_2, \ldots, X_r$ be r independent random variables each having the same geometric distribution.

(a) Show that the moment generating function $M_{\sum X_i}(t) = E(e^{t(X_1+X_2+\cdots+X_r)})$ of the sum is

$$[pe^t/(1-(1-p)e^t)]^r$$

(b) Relate the sum to the total number of trials to obtain r successes. This distribution, is given by

$$\binom{x-1}{r-1} p^r q^{x-r}, x = r, r+1, \cdots$$

(see page 109)

(c) Obtain the first two moments of this negative binomial by differentiating the mgf.

6.41 Refer to Exercise 6.40. Let $X_1, X_2, \ldots, X_n$ be n independent random variables each having a negative binomial distribution with success probability p but where X_i has parameter r_i.

(a) Show that the mgf $M_{\sum X_i}(t) = E(e^{t(X_1+X_2+\cdots+X_r)})$ of the sum $\sum X_i$ is

$$[pe^t/(1-(1-p)e^t)]^{\sum_{i=0}^{n} r_i}$$

(b) Identify the form of this mgf and specify the distribution of $\sum X_i$.

6.7 Transformation Methods to Obtain Distributions*

We briefly[1] introduce two further techniques for obtaining the probability distribution, or density, of a random variable that is a function of a random variable whose distribution is known. These are the **distribution function method** and the **transformation method**.

Distribution Function Method

Under this method, we first obtain the distribution function $G(y)$ of $Y = h(X)$, where X has known distribution function $F(x)$. The density, if needed, can be obtained by differentiation.

$$G(y) = P(Y \le y) = P[h(X) \le y]$$

Two examples will illustrate this method.

EXAMPLE **The probability integral transformation**

Let X have distribution function $F(x)$ and density function $f(x)$ which is positive on an open interval and 0 elsewhere. Consider the **probability integral transformation** $Y = F(X)$ where the cumulative distribution distribution is evaluated at the random variable X. Show that $F(X)$ has a uniform distribution on (0, 1).

Solution Choose any value y between 0 and 1. Since $F(x)$ has a positive derivative, there is a unique value x such that $F(x) = y$. This correspondence can be written as a function $x = w(y)$ and $F(w(y)) = y$ for all $0 < y < 1$. Then,

$$\begin{aligned} G(y) = P(Y \le y) &= P(F(X) \le y) \\ &= P(X \le w(y)) \\ &= F(w(y) = y \end{aligned}$$

for any $0 < y < 1$. The cumulative distribution function $G(y) = y$ is that of the uniform distribution. ■

*This section may be skipped on first reading since the techniques are not used later in the book.

[1]An extended Section 6.7 is available at the Prentice Hall Web site.

EXAMPLE **Distribution function method applied to X^2**

Let X have distribution function $F(x)$ and density function $f(x)$.

(a) Show that its square, $Y = X^2$, has distribution function

$$G(y) = P(Y \le y) = F(\sqrt{y}) - F(-\sqrt{y})$$

(b) If X has a standard normal distribution, show that its square has

$$g(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2}$$

which is a chi square distribution with 1 degree of freedom

Solution **(a)** We have

$$\begin{aligned} G(y) = P(Y \le y) &= P(X^2 \le y) \\ &= P(-\sqrt{y} \le X \le \sqrt{y}) \\ &= F(\sqrt{y}) - F(-\sqrt{y}) \end{aligned}$$

(b) Upon differentiating,

$$g(y) = f(\sqrt{y})\frac{d\sqrt{y}}{dy} - f(-\sqrt{y})\frac{d-\sqrt{y}}{dy} = f(\sqrt{y})\frac{y^{-1/2}}{2} + f(-\sqrt{y})\frac{y^{-1/2}}{2}$$

The reader is asked, in Exercise 6.42, to verify the stated density. ■

For any differentiable strictly increasing function $h(x)$, with inverse function $w(y)$, we have $G(y) = F(w(y))$. By taking the derivative of both sides with respect to y, we obtain the expression for the density function of $h(X)$ presented next when discussing the transformation method.

Transformation Method

The probability density for a function of a random variable can be determined directly from the density of the original variable. Consider $Y = h(X)$ where X has density $f(x)$. Initially, we assume that $h(x)$ is differentiable and either strictly increasing or strictly decreasing. Then, $y = h(x)$ can be solved for x. That is, $h(x)$ has an inverse $w(y) = x$.

Theorem 6.8 Let $Y = h(X)$ where X has density $f(x)$, and let $h(x)$ be differentiable and either strictly increasing or strictly decreasing on the range where $f(x) \neq 0$. The inverse function $w(y)$ exists and the density of Y is given by

Density function of $h(X)$

$$g(y) = \begin{cases} f(w(y))\,|w'(y)| & \text{where } w'(y) \neq 0 \\ 0 & \text{elsewhere} \end{cases}$$

EXAMPLE **Transformation method: square root of chi square/ degrees of freedom**

Let X have a chi square distribution with ν degrees of freedom. Apply the transformation to show that the density of $Y = \sqrt{X/\nu}$ is

$$\frac{\nu^{\nu/2}}{\Gamma(\frac{\nu}{2})\,2^{(\nu-2)/2}} y^{\nu-1} e^{-\nu y^2/2}, \quad y > 0$$

Solution The density of the chi square distribution with ν degrees of freedom is given by

$$f(x) = \frac{1}{\Gamma(\frac{\nu}{2})\,2^{\nu/2}}\, x^{\frac{\nu}{2}-1} e^{-x/2}$$

and $y = \sqrt{x/\nu} = h(x)$ has inverse $x = \nu y^2 = w(y)$. Since $w'(y) = 2\nu y$ is continuous and greater than 0 for $y > 0$,

$$g(y) = \frac{1}{\Gamma(\frac{\nu}{2})\,2^{\nu/2}}\,(\nu y^2)^{\frac{\nu}{2}-1} e^{-\nu y^2/2}\, 2\nu y$$

$$= \frac{\nu^{\nu/2}}{\Gamma(\frac{\nu}{2})\,2^{(\nu-2)/2}}\, y^{\nu-1} e^{-\nu y^2/2}$$

■

We state two important transformations to obtain the sum or the ratio of two independent random variables.

Theorem 6.9 Let X and Y be independent and let X have density $f_X(x)$ and Y have density $f_Y(y)$. Then the density of $Z = X + Y$ is given by the **convolution formula**

Convolution formula

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx \quad \text{for all } z$$

The ratio of random variables $Z = Y/X$ has density

$$f_{Y/X}(z) = \int_{-\infty}^{\infty} |x|\, f_X(x) f_Y(xz)\,dx \quad \text{for all } z$$

EXAMPLE **Student's t distribution**

Let Y have a standard normal distribution and be independent of X which has a chi square distribution with ν degrees of freedom. Apply the transformation technique to show that the density of

$$\frac{\text{standard normal}}{\sqrt{\dfrac{\text{chi square}}{\text{degrees of freedom}}}} = \frac{Y}{\sqrt{\frac{X}{\nu}}}$$

is given by

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \text{for } -\infty < t < \infty$$

This distribution is called the **student's t distribution**, or the **t distribution**.

Solution Since Y has a standard normal distribution, using the conclusion from the previous example, but with Y replaced by X, the density of the ratio $T = Y/X$ is

$$\int_{-\infty}^{\infty} |x|\, f_X(x) f_Y(xt)\,dx$$

$$= \int_0^{\infty} |x| \frac{1}{\sqrt{2\pi}} e^{-(tx)^2/2} \frac{\nu^{\nu/2}}{\Gamma(\frac{\nu}{2})\,2^{(\nu-2)/2}}\, x^{\nu-1} e^{-\nu x^2/2}\,dx$$

$$= \int_0^{\infty} \frac{\nu^{\nu/2}}{\sqrt{\pi}\,\Gamma(\frac{\nu}{2})\,2^{(\nu-1)/2}}\, x^{\nu} e^{-\frac{1}{2}(\nu+t^2)x^2}\,dx$$

Making the change of variable $u = x^2(\nu + t^2)/2$, we obtain

$$(\nu + t^2)^{-\frac{\nu+1}{2}} \frac{\nu^{\nu/2}}{\sqrt{\pi}\,\Gamma(\frac{\nu}{2})} \int_0^\infty u^{\frac{\nu-1}{2}} e^{-u}\, du$$

and the result follows from the definition of $\Gamma(\frac{\nu+1}{2})$. ■

Convolution Formula for Discrete Random Variables

There is also a convolution formula for the sum, $Z = X + Y$, of two independent discrete random variables X and Y. Let $f_X(x)$ denote the probability distribution of X and $f_Y(y)$ denote the probability distribution of Y. We restrict attention to cases where X and Y take on nonnegative integer values.

To find $f_Z(z) = P(Z \le z)$, for each z, we recognize that the event $[Z = z]$ is the union of the disjoint events $[X = x \text{ and } Y = z - x]$ for $x = 0, 1, \ldots, z$. Consequently,

$$\begin{aligned} P(Z = z) = f_Z(z) &= \sum_{x=0}^{z} P(X = x \text{ and } Y = z - x) \\ &= \sum_{x=0}^{z} f_X(x) f_Y(z - x) \end{aligned}$$

where the last step follows by independence. This last result is called the **convolution formula** for discrete random variables.

Discrete convolution formula

Theorem 6.10 Let X and Y be non-negative and integer valued. The random variable $Z = X + Y$ has probability distribution $f_Z(z)$ given by

$$f_Z(z) = f_{X+Y}(z) = P(Z = z) = \sum_{x=0}^{z} f_X(x) f_Y(z - x) \text{ for } z = 0, 1, \ldots$$

EXAMPLE **Sum of two independent Poisson random variables**

Let X and Y be independent Poisson random variables where X has parameter (λ_1) and Y has parameter (λ_2). Show that the sum

$$X + Y \quad \text{has a Poisson distribution with parameter} \quad \lambda_1 + \lambda_2$$

Solution By the discrete convolution formula, $Z = X + Y$ has probability distribution

$$f_Z(z) = \sum_{x=0}^{z} f_X(x) f_Y(z - x)$$

so

$$\begin{aligned} f_Z(z) &= \sum_{x=0}^{z} \frac{\lambda_1^x}{x!} e^{-\lambda_1} \frac{\lambda_2^{z-x}}{(z-x)!} e^{-\lambda_2} \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{x=0}^{z} \frac{\lambda_1^x}{x!} \frac{\lambda_2^{z-x}}{(z-x)!} \end{aligned}$$

Use the binomial formula

$$(a+b)^m = \sum_{x=0}^{m} \binom{m}{x} a^x b^{m-x}$$

with $m = z$, $a = \lambda_1$, and $b = \lambda_2$, after multiplying and dividing by $z!$, to conclude that

$$\sum_{x=0}^{z} \frac{\lambda_1^x}{x!} \frac{\lambda_2^{z-x}}{(z-x)!} = \frac{(\lambda_1 + \lambda_2)^z}{z!}$$

and the result is established.
Remark: Note that the rate parameters λ_i add. ■

Exercises

6.42 Referring to the example on page 198, verify that

$$g(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2}$$

6.43 Use the distribution function method to obtain the density of Z^3 when Z has a standard normal distribution.

6.44 Use the distribution function method to obtain the density of $1 - e^{-X}$ when X has the exponential distribution with $\beta = 1$.

6.45 Use the distribution function method to obtain the density of $\ln(X)$ when X has the exponential distribution with $\beta = 1$.

6.46 Use the transformation method to obtain the density of of X^2 when X has density $f(x) = 0.5x$ for $0 < x < 2$.

6.47 Use the transformation method to obtain the distribution of $-\ln(X)$ when X has the uniform distribution on $(0, 1)$.

6.48 Use the convolution formula, Theorem 6.9, to obtain the density of $X + Y$ when X and Y are independent and each has the exponential distribution with $\beta = 1$.

6.49 Use the transformation method, Theorem 6.9, to obtain the distribution of the ratio Y/X when when X and Y are independent and each has the same gamma distribution.

6.50 Use the discrete convolution formula, Theorem 6.10, to obtain the probability distribution of $X + Y$ when X and Y are independent and each has the uniform distribution on $\{0, 1, 2\}$.

Do's and Don'ts

Do's

1. Understand the concept of a sampling distribution. Each observation is the value of a random variable so a sample of n observations varies from one possible sample to another. Consequently, a statistic such as a sample mean varies from one possible sample to another. The probability distribution or density function which describes the chance behavior of the sample mean is called its *sampling distribution*.
2. When the underlying distribution has mean μ and variance σ^2, remember that the sampling distribution of $\overline{X}$ has

$$\text{mean of } \overline{X} = \mu = \text{population mean}$$

$$\text{variance of } \overline{X} = \frac{\sigma^2}{n} = \frac{\text{population variance}}{n}$$

3. When the underlying distribution is normal with mean μ and variance σ^2, calculate exact probabilities for $\overline{X}$ using the normal distribution with mean μ and variance $\frac{\sigma^2}{n}$

$$P(\overline{X} \leq b) = P\left(Z \leq \frac{b-\mu}{\sigma/\sqrt{n}}\right)$$

4. Apply the central limit theorem, when the sample size is large, to approximate the sampling distribution of $\overline{X}$ by a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. The probability $P(\overline{X} \leq b)$ is approximately equal to the standard normal probability $P\left(Z \leq \frac{b-\mu}{\sigma/\sqrt{n}}\right)$.

Don'ts

1. Don't confuse the population distribution, which describes the variation for a single random variable, with the sampling distribution of a statistic.
2. When sampling from a finite population of size N, don't use $\sigma/\sqrt{n}$ as the standard deviation of $\overline{X}$ unless the finite population correction factor is nearly 1.
3. When the population distribution is noticeably nonnormal, don't try to conclude that the sampling distribution of $\overline{X}$ is normal unless the sample size is at least moderately large, 30 or more.

Review Exercises

6.51 The panel for a national science fair wishes to select 10 states from which a student representative will be chosen at random from the students participating in the state science fair.

(a) Use Table 7 to select the 10 states.

(b) Does the total selection process give each student who participates in some state science fair an equal chance of being selected to be a representative at the national science fair?

6.52 How many different samples of size $n = 2$ can be chosen from a finite population of size

(a) $N = 7$;

(b) $N = 18$?

6.53 With reference to Exercise 6.52, what is the probability of choosing each sample in part (a) and the probability of choosing each sample in part (b), if the samples are to be random?

6.54 Referring to Exercise 6.52, find the value of the finite population correction factor in the formula for $\sigma^2_{\overline{X}}$ for part (a) and part (b).

6.55 The time to check out and process payment information at an office supplies Web site can be modeled as a random variable with mean $\mu = 63$ seconds and variance $\sigma^2 = 81$. If the sample mean $\overline{X}$ will be based on a random sample of $n = 36$ times, what can we assert about the probability of getting a sample mean greater than 66.75, if we use

(a) Chebyshev's theorem;

(b) the central limit theorem?

6.56 The number of pieces of mail that a department receives each day can be modeled by a distribution having mean 44 and standard deviation 8. For a random sample of 35 days, what can be said about the probability that the sample mean will be less than 40 or greater than 48 using

(a) Chebyshev's theorem;

(b) the central limit theorem?

6.57 If measurements of the specific gravity of a metal can be looked upon as a sample from a normal population having a standard deviation of 0.04, what is the

probability that the mean of a random sample of size 25 will be "off" by at most 0.02?

6.58 Adding graphite to iron can improve its ductile qualities. If measurements of the diameter of graphite spheres within an iron matrix can be modeled as a normal distribution having standard deviation 0.16, what is the probability that the mean of a sample of size 36 will differ from the population mean by more than 0.06?

6.59 If 2 independent random samples of size $n_1 = 9$ and $n_2 = 16$ are taken from a normal population, what is the probability that the variance of the first sample will be at least 4 times as large as the variance of the second sample?

6.60 If 2 independent sample of sizes $n_1 = 26$ and $n_2 = 8$ are taken from a normal population, what is the probability that the variance of the second sample will be at least 2.4 times the variance of the first sample?

6.61 When we sample from an infinite population, what happens to the standard error of the mean if the sample size is

(a) increased from 100 to 200;

(b) increased from 200 to 300;

(c) decreased from 360 to 90?

6.62 A traffic engineer collects data on traffic flow at a busy intersection during the rush hour by recording the number of westbound cars that are waiting for a green light. The observations are made for each light change. Explain why this sampling technique will not lead to a random sample.

6.63 Explain why the following may not lead to random samples from the desired populations:

(a) To determine the smoothness of shafts, a manufacturer measures the roughness of the first piece made each morning.

(b) To determine the mix of cars, trucks, and buses in the rush hour, an engineer records the type of vehicle passing a fixed point at 1-minute intervals.

6.64 Several pickers are each asked to gather 30 ripe apples and put them in a bag.

(a) Would you expect all of the bags to weigh the same? For one bag, let X_1 be the weight of the first apple, X_2 the weight of the second apple, and so on. Relate the weight of this bag,

$$\sum_{i=1}^{30} X_i$$

to the approximate sampling distribution of $\overline{X}$.

(b) Explain how your answer to part (a) leads to the sampling distribution for the variation in bag weights.

(c) If the weight of an individual apple has mean $\mu = 0.2$ pound and standard deviation $\sigma = 0.03$ pound, find the probability that the total weight of the bag will exceed 6.2 pounds.

Key Terms

Central limit theorem 184
Chi square distribution 189
Convolution formula 199
Degrees of freedom 188
Discrete uniform distribution 179
Distribution function method 197
F distribution 190
Finite population 175
Finite population correction factor 181
Infinite population 175
Law of large numbers 182
Moment generating function method 195
Parameter 178
Population 175
Probability integral transformation 197
Probable error of the mean 186
Random sample 176
Representation 192
Sample 175
Sampling distribution 180
Standard error of the mean 183
Standardized sample mean 183
Statistic 178
t distribution 187
Transformation method 198

CHAPTER

7 Inferences Concerning a Mean

CHAPTER OUTLINE

7.1 Point Estimation 204
7.2 Interval Estimation 209
7.3 Maximum Likelihood Estimation 215
7.4 Tests of Hypotheses 222
7.5 Null Hypotheses and Tests of Hypotheses 224
7.6 Hypotheses Concerning One Mean 229
7.7 The Relation between Tests and Confidence Intervals 235
7.8 Power, Sample Size, and Operating Characteristic Curves 237
Review Exercises 243
Key Terms 244

In the beginning of Chapter 6, we stated that the purpose of most statistical investigations is to generalize from information contained in random samples about the populations from which the samples were obtained. In the classical approach the methods of statistical inference are divided into two major areas—estimation and tests of hypotheses. In Sections 7.1 and 7.2 we shall present some theory and some methods which pertain to the estimation of means. A general estimation procedure is introduced in Section 7.3. Sections 7.4 and 7.5 deal with the basic concepts of hypothesis testing, and Sections 7.6 and 7.7 deal with tests of hypotheses concerning a mean. Test performance, including power, is the subject of Section 7.8.

7.1 Point Estimation

Basically, **point estimation** concerns the choosing of a statistic, that is, a single number calculated from sample data (and perhaps other information) for which we have some expectation, or assurance, that it is reasonably close to the parameter it is supposed to estimate. To explain what we mean here by reasonably close is not an easy task; first, the value of the parameter is unknown, and second, the value of the statistic is unknown until after the sample has been obtained. Thus, we can only ask whether, upon repeated sampling, the distribution of the statistic has certain desirable properties akin to closeness.

EXAMPLE **Two components of closeness: variance and bias**

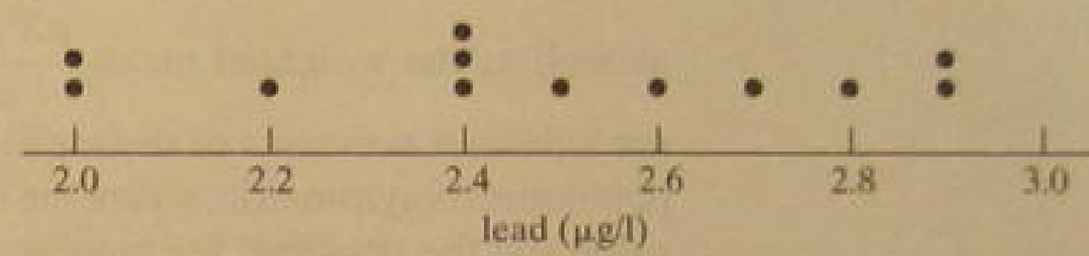
Scientists need to be able to detect small amounts of contaminants in the environment. As a check on the current capabilities, the following measurements were made on test specimens spiked with a known concentration 1.25 μg/l of lead. That is, the readings should average 1.25 if there is no background lead in the samples.

2.4 2.9 2.7 2.6 2.9 2.0 2.8 2.2 2.4 2.4 2.0 2.5

(a) Make a dot diagram;

(b) Compute the point estimator $\overline{X}$ and estimate its standard deviation. This latter quantity is called the **estimated standard error** of $\overline{X}$.

Solution (a)



(b) We calculate

$$\bar{x} = \frac{29.8}{12} = 2.483 \quad \text{and} \quad s^2 = \frac{75.08 - (29.8)^2/12}{12 - 1} = 0.09788$$

From Theorem 6.1 $\overline{X}$ has standard deviation $\sigma/\sqrt{n}$, so the estimated standard error is $s/\sqrt{n} = \sqrt{0.09788/12} = 0.090$.

There appears to be either a bias due to the laboratory procedure or some lead already in the samples before they were spiked. All 12 measurements are above the known concentration, 1.25 μg/l. ■

Closeness can be expressed as two components: (1) the bias, which is the expected value of the estimator minus the parameter; and (2) the standard deviation or variance of the estimator.

Point estimation of a mean

Parameter: Population mean μ
Data: A random sample $X_1, \ldots, X_n$
Estimator: $\overline{X}$
Estimate of standard error: $\dfrac{S}{\sqrt{n}}$

We know from Theorem 6.1 that the sampling distribution of the mean has the same mean as the population from which the sample is obtained; hence, we can expect that the means of repeated random samples from a given population will center on the mean of this population and not about some other value.

To formulate this property more generally, let θ be the parameter of interest and $\hat{\theta}$ be a statistic. The **hat notation** distinguishes the sample-based quantity from the parameter. We now make the following definition:

Unbiased estimator

A statistic $\hat{\theta}$ is said to be an **unbiased estimator**, or its value an unbiased estimate, if and only if the mean of the sampling distribution of the estimator equals θ, whatever the value of θ.

Thus, we call a statistic unbiased if "on the average" its values will equal the parameter it is supposed to estimate. Note that we have distinguished here between an estimator, a random variable, and an estimate, which is one of its values. Also, it is customary to apply the term *statistic* to both estimates and estimators.

It is a mathematical fact that $\overline{X}$ is an unbiased estimator of the population mean μ provided the observations are a random sample. However, the particular engineering application may contain unknown biases as in the spiked water sample example.

Generally speaking, the property of unbiasedness is one of the more desirable properties in point estimation, although it is by no means essential and it is sometimes out weighed by other factors. One shortcoming of the criterion of unbiasedness is that it will generally not provide a unique statistic for a given problem of estimation. For instance, it can be shown that for a random sample of size $n = 2$ the mean $\dfrac{X_1 + X_2}{2}$

as well as the weighted mean $\frac{aX_1 + bX_2}{a+b}$, where a and b are positive constants, are unbiased estimates of the mean of the population. If we further assume that the population is symmetric, so are the median and the midrange (the mean of the largest value and the smallest) for random samples of any size.

This suggests that we must seek a further criterion for deciding which of several unbiased estimators is best for estimating a given parameter. Such a criterion becomes evident when we compare the sampling distributions of the median and the mean for random samples of size n from the same normal population. The sampling distribution of the mean is normal and that of the median is nearly normal. Although these two sampling distributions have the same mean, the population mean μ, and although they are both symmetrical and bell-shaped, their variances differ. From Theorem 6.1, the variance of the sampling distribution of the mean for random samples from infinite populations is $\frac{\sigma^2}{n}$, and it can be shown that for random samples of the same size from normal populations the variance of the sampling distribution of the median is approximately $1.5708 \cdot \frac{\sigma^2}{n}$. Thus, it is more likely that the mean will be closer to μ than the median is to μ. Despite this long-run average property, given a particular sample we have no way of knowing which of the two is closest.

We formalize this important comparison of sampling distributions of statistics on the basis of their variances.

More efficient unbiased estimator

A statistic $\hat{\theta}_1$ is said to be a more efficient unbiased estimator of the parameter θ than the statistic $\hat{\theta}_2$ if

1. $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ;
2. the variance of the sampling distribution of the first estimator is no larger than that of the second and is smaller for at least one value of θ.

We have thus seen that for random samples from normal populations the mean $\overline{X}$ is more efficient than the median as an estimator of μ; in fact, it can be shown that in most practical situations where we estimate a population mean μ, the variance of the sampling distribution of no other statistic is less than that of the sampling distribution of the mean. In other words, in most practical situations the sample mean is an acceptable statistic for estimating a population mean μ. (There exist several other criteria for assigning the goodness of methods of point estimation, but we shall not discuss them in this book.)

When we use a sample mean to estimate the mean of a population, we know that although we are using a method of estimation which has certain desirable properties, the chances are slim, virtually nonexistent, that the estimate will actually equal μ. Hence, it would seem desirable to accompany such a point estimate of μ with some statement as to how close we might reasonably expect the estimate to be. The error, $\overline{X} - \mu$, is the difference between the estimator and the quantity it is supposed to estimate.

To examine this error, let us make use of the fact that for large n

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

is a random variable having approximately the standard normal distribution.

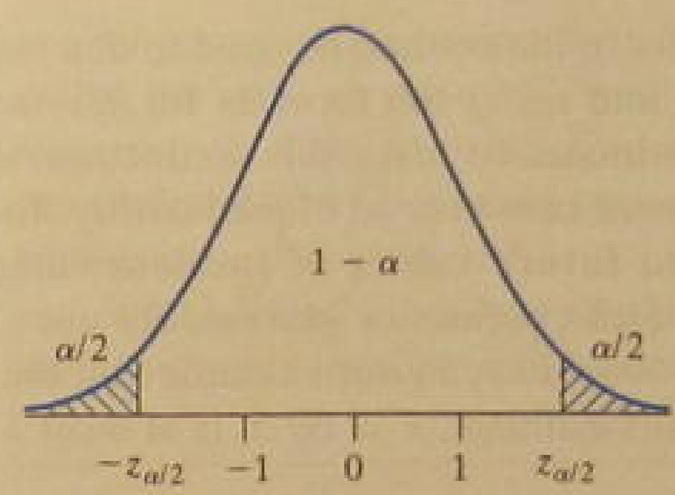


Figure 7.1
The sampling distribution of $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$

As shown in Figure 7.1, we can assert with probability $1-\alpha$ that the inequality

$$-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}$$

will be satisfied or that

$$\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \le z_{\alpha/2}$$

where $z_{\alpha/2}$ is such that the normal curve area to its right equals $\alpha/2$. If we now let E stand for the maximum of these values of $|\overline{X}-\mu|$, the maximum error of estimate, we have that the error $|\overline{X}-\mu|$, will be less than

Maximum error of estimate

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

with probability $1-\alpha$. In other words, if we intend to estimate μ with the mean of a large ($n \ge 30$) random sample, we can assert with probability $1-\alpha$ that the error, $|\overline{X}-\mu|$, will be at most $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$. The most widely used values for $1-\alpha$ are 0.95 and 0.99, and the corresponding values of $z_{\alpha/2}$ are $z_{0.025} = 1.96$ and $z_{0.005} = 2.575$. (See Exercise 5.23 on page 133.)

EXAMPLE **Specifying a high probability for the maximum error (σ known)**

An industrial engineer intends to use the mean of a random sample of size $n = 150$ to estimate the average mechanical aptitude (as measured by a certain test) of assembly line workers in a large industry. If, on the basis of experience, the engineer can assume that $\sigma = 6.2$ for such data, what can he assert with probability 0.99 about the maximum size of his error?

Solution Substituting $n = 150$, $\sigma = 6.2$, and $z_{0.005} = 2.575$ into the preceding formula for E, we get

$$E = 2.575 \cdot \frac{6.2}{\sqrt{150}} = 1.30.$$

Thus, the engineer can assert with probability 0.99 that his error will be at most 1.30. ■

Suppose now that the engineer of this example collects his data and gets $\overline{x} = 69.5$. Can he still assert with probability 0.99 that the error is at most 1.30? First of all, $\overline{x} = 69.5$ either differs from the true average by at most 1.30 or it does not, and he does not know which. Consequently, it must be understood that the 0.99

probability applies to the method he used to determine the maximum error (getting the sample data and using the formula for E) and not directly to the parameter he is trying to estimate. To make this distinction, it has become the custom to use the word **confidence** here instead of *probability*. **In general, we make probability statements about future values of random variables (say, the potential error of an estimate) and confidence statements once the data have been obtained.** Accordingly, we would say in our example that the engineer can be 99% confident that the error of his estimate, $\bar{x} = 69.5$, is at most 1.30.

Determination of Sample Size

The formula for E on page 207 can also be used to determine the sample size that is needed to attain a desired degree of precision. Suppose that we want to use the mean of a large random sample to estimate the mean of a population, and we want to be able to assert with probability $1 - \alpha$ that the error will be at most some prescribed quantity E [or assert later with $(1 - \alpha)100\%$ confidence that the error is at most E]. As before, we write

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

and upon solving this equation for n we get

Sample size determination

$$n = \left[\frac{z_{\alpha/2} \cdot \sigma}{E}\right]^2$$

To be able to use this formula we must know $1 - \alpha$, E, and σ, and for the latter we often substitute an estimate based on prior data of a similar kind (or, if necessary, a good guess).

EXAMPLE **Selecting the sample size**

A research worker wants to determine the average time it takes a mechanic to rotate the tires of a car, and she wants to be able to assert with 95% confidence that the mean of her sample is off by at most 0.50 minute. If she can presume from past experience that $\sigma = 1.6$ minutes, how large a sample will she have to take?

Solution Substituting $E = 0.50$, $\sigma = 1.6$, and $z_{0.025} = 1.96$ into the formula for n, we get

$$n = \left[\frac{1.96 \cdot 1.6}{0.50}\right]^2 = 39.3$$

or 40 rounded up to the nearest integer. Thus, the research worker will have to time 40 mechanics performing the task of rotating the tires of a car. ■

The methods discussed so far in this section require that σ be known or that it can be approximated with the sample standard deviation s, thus requiring that n be large. However, if it is reasonable to assume that we are sampling from a normal population, we can base our argument on Theorem 6.3 instead of Theorem 6.2, namely on the fact that

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is a random variable having the t distribution with $n-1$ degrees of freedom. Duplicating the steps on page 207, we thus arrive at the result that with probability $1-\alpha$ the error we make in using $\overline{X}$ to estimate μ will be at most $t_{\alpha/2}\, S/\sqrt{n}$. Here $t_{\alpha/2}$ has probability $\alpha/2$ of being exceeded by a t random variable having $n-1$ degrees of freedom. (See page 188.)

When $\overline{X}$ and S become available, we assert with $(1-\alpha)100\%$ confidence that the error made in using $\overline{x}$ to estimate μ is at most

Maximum error of estimate, normal population (σ unknown)

$$E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

EXAMPLE **A 98% confidence bound on the maximum error**

In six determinations of the melting point of tin, a chemist obtained a mean of 232.26 degrees Celsius with a standard deviation of 0.14 degree. If he uses this mean to estimate the actual melting point of tin, what can the chemist assert with 98% confidence about the maximum error?

Solution Substituting $n = 6$, $s = 0.14$, and $t_{0.01} = 3.365$ (for $n-1 = 5$ degrees of freedom) into the formula for E, we get

$$E = 3.365 \cdot \frac{0.14}{\sqrt{6}} = 0.19$$

Thus the chemist can assert with 98% confidence that his figure for the melting point of tin is off by at most 0.19 degree. ■

7.2 Interval Estimation

Since point estimates cannot really be expected to coincide with the quantities they are intended to estimate, it is sometimes preferable to replace them with **interval estimates**, that is, with intervals for which we can assert with a reasonable degree of certainty that they will contain the parameter under consideration. To illustrate the construction of such an interval, suppose that we have a large ($n \geq 30$) random sample from a population with the unknown mean μ and the known variance σ^2. Referring to the inequality

$$-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

which, as shown on page 207, will be satisfied with probability $1-\alpha$, we now apply simple algebra and rewrite it as $\overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \overline{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}$. When the observed value $\overline{x}$ becomes available, we obtain

Large sample confidence interval for μ (σ known)

$$\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Thus, when a sample has been obtained and the value of $\overline{x}$ has been calculated, we can claim with $(1-\alpha)100\%$ confidence that the interval from $\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ to

$\bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ contains μ. It is customary to refer to an interval of this kind as a **confidence interval** for μ having the **degree of confidence** $1 - \alpha$ or $(1 - \alpha)\,100\%$ and to its endpoints as the **confidence limits**.

EXAMPLE **Calculating and interpreting a large sample confidence interval**

A random sample of size $n = 100$ is taken from a population with $\sigma = 5.1$. Given that the sample mean is $\bar{x} = 21.6$, construct a 95% confidence interval for the population mean μ.

Solution Substituting the given values of $n, \bar{x}, \sigma$, and $z_{0.025} = 1.96$ into the confidence interval formula, we get

$$21.6 - 1.96 \cdot \frac{5.1}{\sqrt{100}} < \mu < 21.6 + 1.96 \cdot \frac{5.1}{\sqrt{100}}$$

or $20.6 < \mu < 22.6$. Of course, either the interval from 20.6 to 22.6 contains the population mean μ, or it does not, but we are 95% confident that it does. As was explained on page 208, this means that the method by which the interval was obtained "works" 95% of the time. In other words, in repeated applications of the confidence interval formula, 95% of the intervals can be expected to contain the means of the respective populations. ■

The preceding confidence interval formula is exact only for random samples from normal populations, but for large samples it will generally provide good approximations. Since σ is unknown in most applications, we may have to make the further approximation of substituting the sample standard deviation s for σ.

Large sample confidence interval for μ

$$\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

EXAMPLE **A 99% confidence interval for the mean nanopillar height**

With reference to the nanopillar height data on page 15, for which we have $n = 50$, $\bar{x} = 305.58$ nm, and $s^2 = 1{,}366.86$ (hence, $s = 36.97$ nm), construct a 99% confidence interval for the population mean of all nanopillars.

Solution Substituting into the confidence interval formula with $\bar{x} = 305.58$, $s = 36.97$, and $z_{0.005} = 2.575$, we get

$$305.58 - 2.575 \cdot \sqrt{\frac{36.97}{50}} < \mu < 305.58 + 2.575 \cdot \sqrt{\frac{36.97}{50}}$$

or $292.12 < \mu < 319.04$. We are 99% confident that the interval from 292.12 nm to 319.04 nm contains the true mean nanopillar height. ■

For small samples ($n < 30$), we proceed as on page 208, provided it is reasonable to assume that we are sampling from a normal population. Thus, with $t_{\alpha/2}$ defined as on page 209, we get the $(1 - \alpha)100\%$ confidence interval formula

Small sample confidence interval for μ of normal population

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

This formula applies to samples from normal populations, but in accordance with the discussion on page 189, it may be used as long as the sample does not exhibit any pronounced departures from normality.

EXAMPLE **A 99% confidence interval for the mean of a normal population**

The mean weight loss of $n = 16$ grinding balls after a certain length of time in mill slurry is 3.42 grams with a standard deviation of 0.68 gram. Construct a 99% confidence interval for the true mean weight loss of such grinding balls under the stated conditions.

Solution Substituting $n = 16$, $\bar{x} = 3.42$, $s = 0.68$, and $t_{0.05} = 2.947$ for $n - 1 = 15$ degrees of freedom into the small sample confidence interval formula for μ, we get

$$3.42 - 2.947 \cdot \frac{0.68}{\sqrt{16}} < \mu < 3.42 + 2.947 \cdot \frac{0.68}{\sqrt{16}}$$

or $2.92 < \mu < 3.92$. We are 99% confident that the interval from 2.92 grams to 3.92 grams contains the mean weight loss. ■

Because confidence intervals are an important way of making inferences, we review their interpretation in the context of 95% confidence intervals for μ.

Before the observations are made, $\bar{X}$ and S are random variables, so

1. The interval from $\bar{X} - t_{0.025}\frac{S}{\sqrt{n}}$ to $\bar{X} + t_{0.025}\frac{S}{\sqrt{n}}$ is a random interval. It is centered at $\bar{X}$ and its length is proportional to S.
2. The interval from $\bar{X} - t_{0.025}\frac{S}{\sqrt{n}}$ to $\bar{X} + t_{0.025}\frac{S}{\sqrt{n}}$ will cover the true (fixed) μ with probability 0.95.

Once the observations are made and we have the numerical value $\bar{x}$ and s,

3. The calculated interval from $\bar{x} - t_{0.025}\frac{s}{\sqrt{n}}$ to $\bar{x} + t_{0.025}\frac{s}{\sqrt{n}}$ is fixed. It is no longer possible to talk about the probability of covering μ. The interval either covers μ or it does not. Further, in any particular application, we have no way of knowing if μ is covered or not.

However, because 0.95 is the probability that we cover μ in each application, the long-run relative frequency interpretation of probability (or law of large numbers) promises that

$$\frac{\text{number of intervals that cover the true mean}}{\text{number of intervals calculated}} \to 0.95$$

when the intervals are calculated for a large number of different problems. This is what gives us 95% confidence! Over many different applications of the method, the proportion of intervals that cover μ should be nearly 0.95.

To emphasize these points, we simulated a sample of size $n = 10$ from a normal distribution with $\mu = 20$ and $\sigma = 5$. The 95% confidence interval was then calculated and graphed in Figure 7.2. This procedure was repeated 20 times. The

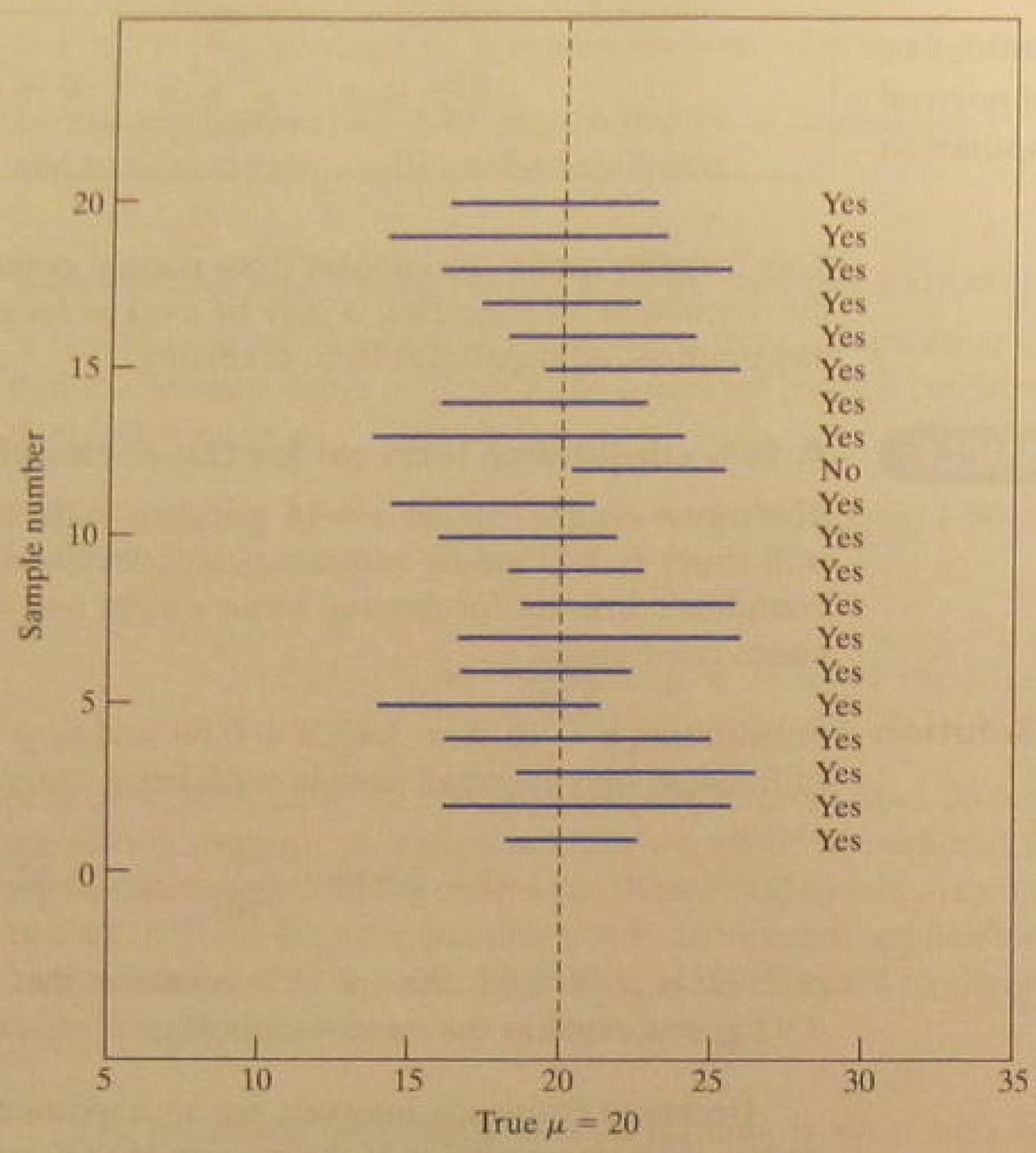


Figure 7.2
Interpretation of the confidence interval for population mean, true mean $\mu = 20$

different samples produce different values for $\bar{x}$ and, consequently, the intervals are centered at different points. The different values of the standard deviation s gave rise to intervals of different lengths. Unlike a real application, here we know that the true fixed mean is $\mu = 20$. The proportion of intervals that cover the true value of $\mu = 20$ should be near 0.95 and, in this instance, we happen to have exactly that proportion $19/20 = 0.95$.

Exercises

7.1 Civil engineers collected data from one area of Wisconsin on the amount of salt (tons) used to keep highways drivable during a snowstorm. The amount of salt for $n = 30$ storms

1111	2115	1573	2813	2815	2126	854	3965	1819	776
1484	2056	784	779	1373	1237	1701	1957	246	1730
2365	1902	2858	2236	1718	916	2830	2865	1574	1373

has $\bar{x} = 1798.4$ tons and $s^2 = 671{,}330.9$ so $s = 819.35$ tons. What can one assert with 95% confidence about the maximum error if $\bar{x} = 1798.4$ is used as a point estimate of the true population mean amount of salt required for a snowstorm?

7.2 With reference to the previous exercise, construct a 95% confidence interval for the true population mean amount of salt required for a snowstorm.

7.3 An industrial engineer collected data on the labor time required to produce an order of automobile mufflers using a heavy stamping machine. The data on times (hours) for $n = 52$ orders of different parts

2.15	2.27	0.99	0.63	2.45	1.30	2.63	2.20	0.99	1.00	1.05
3.44	0.49	0.93	2.52	1.05	1.39	1.22	3.17	0.85	1.18	2.27
1.52	0.48	1.33	4.20	1.37	2.70	0.63	1.13	3.81	0.20	1.08
2.92	2.87	2.62	1.03	2.76	0.97	0.78	4.68	5.20	1.90	0.55
1.00	2.95	0.45	0.70	2.43	3.65	4.55	0.33			

has $\bar{x} = 1.865$ hours and $s^2 = 1.5623$ so $s = 1.250$ hours. What can one assert with 95% confidence about the maximum error if $\bar{x} = 1.865$ hours is used as a point estimate of the true population mean labor time required to run the heavy stamping machine?

7.4 With reference to the previous exercise, construct a 95% confidence interval for the true population mean labor time.

7.5 The manufacture of large liquid crystal displays (LCD's) is difficult. Some defects are minor and can be removed; others are unremovable. The number of unremovable defects, for each of $n = 45$ displays (*Courtesy of Shiyu Zhou*)

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

has $\bar{x} = 2.667$ and $s = 3.057$ unremovable defects. What can one assert with 98% confidence about the maximum error if $\bar{x} = 1798.4$ is used as a point estimate of the true population mean number of unremovable defects?

7.6 With reference to the previous exercise, construct a 98% confidence interval for the true population mean number of unremovable defect per display.

7.7 With reference to the $n = 50$ interrequest time observations on page 19, which have mean 11,795 and standard deviation 14,054, what can one assert with 95% confidence about the maximum error if $\bar{x} = 11,795$ is used as a point estimate of the true population mean interrequest time?

7.8 With reference to the previous exercise, construct a 95% confidence interval for the true mean interrequest time.

7.9 In a study of automobile collision insurance costs, a random sample of 80 body repair costs for a particular kind of damage had a mean of \$472.36 and a standard deviation of \$62.35. If $\bar{x} = \$472.36$ is used as a point estimate of the true average repair cost of this kind of damage, with what confidence can one assert that the error does not exceed \$10?

7.10 If we want to determine the average mechanical aptitude of a large group of workers, how large a random sample will we need to be able to assert with probability 0.95 that the sample mean will not differ from the true mean by more than 3.0 points? Assume that it is known from past experience that $\sigma = 20.0$.

7.11 The dean of a college wants to use the mean of a random sample to estimate the average amount of time students take to get from one class to the next, and she wants to be able to assert with 99% confidence that the error is at most 0.25 minute. If it can be presumed from experience that $\sigma = 1.40$ minutes, how large a sample will she have to take?

7.12 One novel process of making green gasoline takes biomass in the form of sucrose and converts it into gasoline using catalytic reactions. At one step in a pilot plant process, a chemical engineer measures the output of carbon chains of length three. Nine runs with same catalyst produced the yields (gal)

0.63 2.64 1.85 1.68 1.09 1.67 0.73 1.04 0.68

What can the chemical engineer assert with 95% confidence about the maximum error if she uses the sample mean to estimate true mean yield?

7.13 With reference to the previous exercise, assume that yield has a normal distribution and obtain a 95% confidence interval for the true mean yield of the pilot plant process.

7.14 To monitor complex chemical processes, chemical engineers will consider key process indicators, which may be just yield but most often depend on several quantities. Before trying to improve a process, $n = 9$ measurements were made on a key performance indicator.

123 106 114 128 113 109 120 102 111

What can the engineer assert with 95% confidence about the maximum error if he uses the sample mean to estimate true mean value of the performance indicator?

7.15 With reference to the previous exercise, assume that the key performance indicator has a normal distribution and obtain a 95% confidence interval for the true value of the indicator.

7.16 Refer to Exercise 2.34, page 36, concerning material costs for rebuilding $n = 29$ traction motors. A computer calculation gives $\bar{x} = 1.4707$ and $s = 0.5235$ thousand dollars. Obtain a 90% confidence interval for the mean material costs to rebuild a motor.

7.17 Refer to the 2×4 lumber strength data in Exercise 2.58, page 38. According to the computer output, a sample of $n = 30$ specimens had $\bar{x} = 1908.8$ and $s = 327.1$. Find a 95% confidence interval for the population mean strength.

7.18 Refer to the data on page 40, on the number of defects per board for Product B. Obtain a 95% confidence interval for the population mean number of defects per board.

7.19 With reference to the thickness measurements in Exercise 2.41, page 37, obtain a 95% confidence interval for the mean thickness.

7.20 Ten bearings made by a certain process have a mean diameter of 0.5060 cm and a standard deviation of 0.0040 cm. Assuming that the data may be looked upon as a random sample from a normal population, construct a 95% confidence interval for the actual average diameter of bearings made by this process.

7.21 The freshness of produce at a mega-store is rated a scale of 1 to 5, with 5 being very fresh. From a random

sample of 36 customers, the average score was 3.5 with a standard deviation of 0.8.

(a) Obtain a 90% confidence interval for the population mean, μ, or the mean score for all customers.

(b) Does μ lie in your interval obtained in part (a)? Explain.

(c) In long series of repeated experiments, with new random samples collected for each experiment, what proportion of the resulting confidence intervals will contain the true population mean? Explain your reasoning.

7.22 A copy shop records that in $n = 64$ cases, the cartridge for the copy machine lasted an average of 18,300 copies with a standard deviation of 2,800 copies.

(a) Obtain a 95% confidence interval for μ, the population mean number of copies before a new cartridge is needed for the copy machine.

(b) Does μ lie in your interval obtained in part (a)? Explain.

(c) In long series of repeated experiments, what proportion of the respective confidence intervals contain the true mean? Explain your reasoning.

7.23 A sample of maximum pressure from 12 trials with a stamping machine yielded a mean of 7.2 thousand psi and a standard deviation of 1.2 thousand psi.

(a) Obtain a 95% confidence interval for the population mean of maximum pressure μ.

(b) Is the population mean contained in your interval in part (a)? Explain.

(c) What did you assume about the population in your answer to part (a)?

(d) Why are you 95% confident about the interval in part (a)?

7.24 In an air-pollution study performed at an experiment station, the following amount of suspended benzene-soluble organic matter (in micrograms per cubic meter) was obtained for eight different samples of air:

2.2 1.8 3.1 2.0 2.4 2.0 2.1 1.2

Assuming that the population sampled is normal, construct a 95% confidence interval for the corresponding true mean.

7.25 Modify the formula for E on page 207 so that it applies to large samples which constitute substantial portions of finite populations, and use the resulting formula for the following problems:

(a) A sample of 50 scores on the admission test for a school of engineering is drawn at random from the scores of the 420 persons who applied to the school in 2010. If the sample mean and the standard deviation are $\bar{x} = 546$ and $s = 85$, what can we assert with 95% confidence about the maximum error if $\bar{x} = 546$ is used as an estimate of the mean score of all the applicants?

(b) A random sample of 40 drums of a chemical, drawn from among 200 such drums whose weights can be expected to have the standard deviation $\sigma = 12.2$ pounds, has a mean weight of 240.8 pounds. If we estimate the mean weight of all 200 drums as 240.8 pounds, what can we assert with 99% confidence about the maximum error?

7.26 Instead of the large sample confidence interval formula for μ on page 209, we could have given the alternative formula

$$\bar{x} - z_{\alpha/3} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{2\alpha/3} \cdot \frac{\sigma}{\sqrt{n}}$$

Explain why the one on page 209 is narrower, and hence preferable, to the one given here.

7.27 Suppose that we observe a random variable having the binomial distribution. Let X be the number of successes in n trials.

(a) Show that $\frac{X}{n}$ is an unbiased estimate of the binomial parameter p.

(b) Show that $\frac{X+1}{n+2}$ is not an unbiased estimate of the binomial parameter p.

7.28 The statistical program *MINITAB* will calculate the small sample confidence interval for μ. With the nanopillar height data in C1,

Dialog box:
Calc > Stat > Basic Statistics > 1-Sample t
Type C1 in **variables**. Choose **confidence interval** and in **level**, enter 95.
Click **OK**.

produces the output

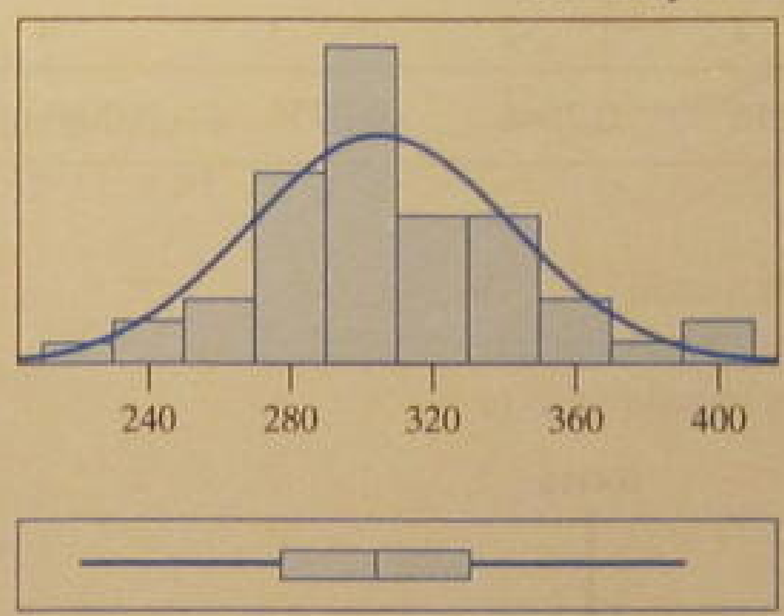
	N	Mean	StDev	SE Mean	95% CI
C1	50	305.580	36.971	5.229	(295.073, 316.087)

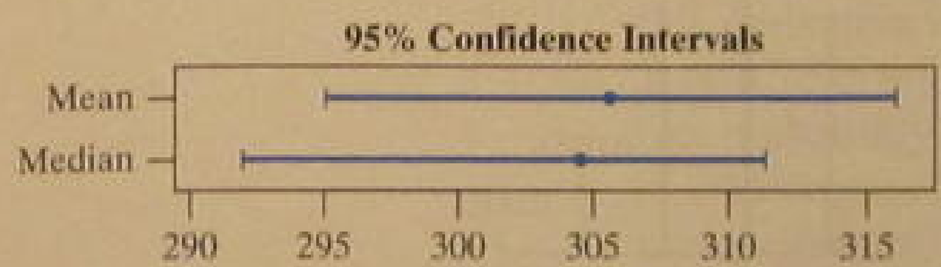
(a) Obtain a 90% confidence interval for μ.

(b) Obtain a 95% confidence interval for μ with the aluminum alloy data on page 19.

Alternatively, you can use the *MINITAB* commands
Stat > Basic statistics > Graphical summary
to produce the more complete output

Summary for height (nm)





Anderson-Darling Normality Test	
A-Squared	0.38
P-Value	0.398
Mean	305.58
StDev	36.97
Variance	1366.86
Skewness	0.260823
Kurtosis	0.202664
N	50
Minimum	221.00
1st Quartile	277.50
Median	304.50
3rd Quartile	330.75
Maximum	391.00
95% Confidence Interval for Mean	
295.07	316.09
95% Confidence Interval for Median	
292.00	311.33
95% Confidence Interval for StDev	
30.88	46.07

7.29 You can simulate the coverage of the small sample confidence intervals for μ by generating 20 samples of size 10 from a normal distribution with $\mu = 20$ and $\sigma = 5$ and computing the 95% confidence intervals according to the formula on page 211. Using *MINITAB*:

> **Calc > Random Data > Normal**
> Type 10 in **Generate**, $C1 - C20$ in **Store**, 20 in **Mean**, and 5 in **standard deviation**. Click **OK**.
>
> **Stat > Basic Statistics > 1-sample t**
> Type $C1 - C20$ in **Variables**. Click **OK**.

(a) From your output, determine the proportion of the 20 intervals that cover the true mean $\mu = 20$.

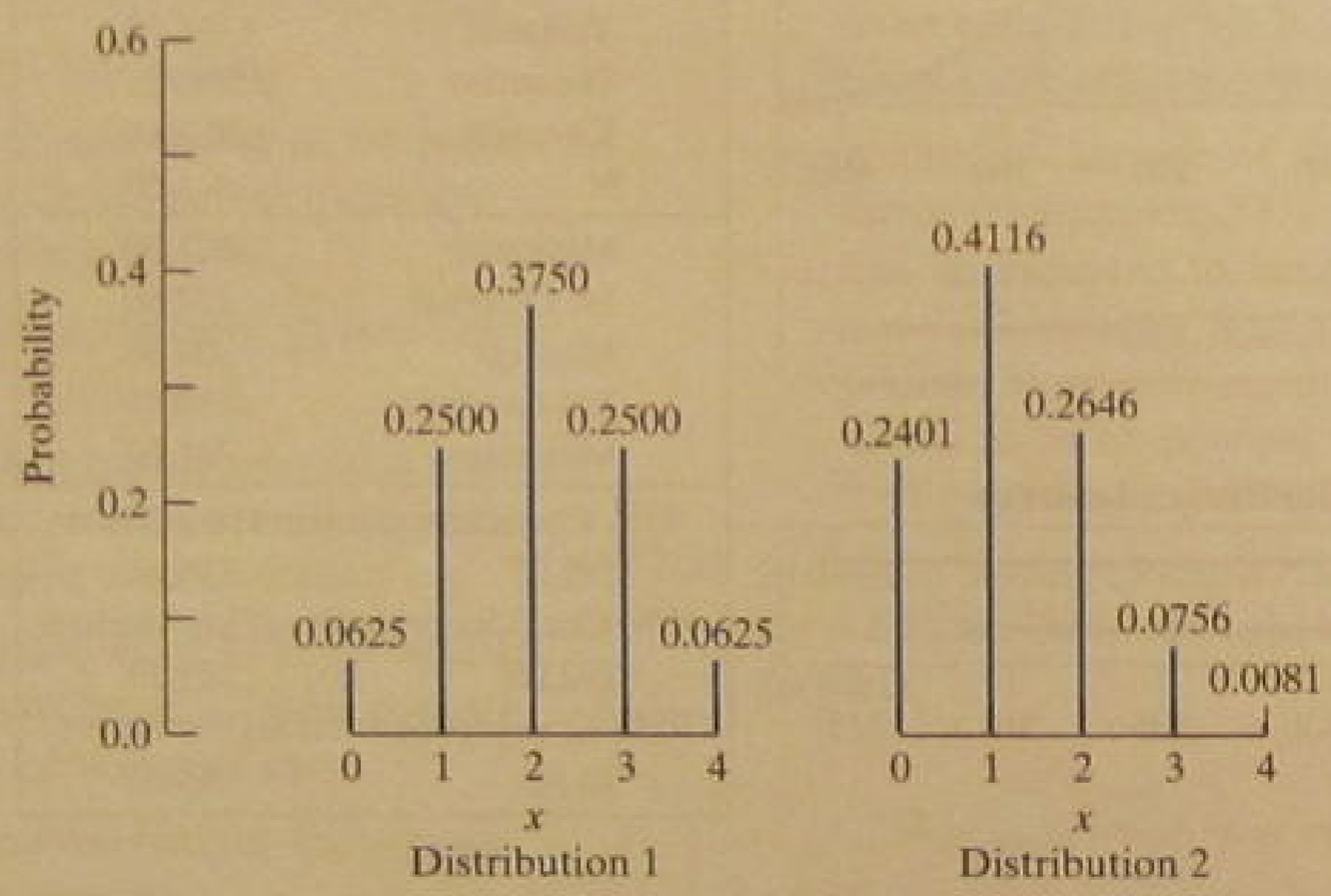
(b) Repeat with 20 samples of size 5.

7.3 Maximum Likelihood Estimation

Sometimes it is necessary to estimates parameter other than a mean or variance. A very general approach to estimation, proposed by R. A. Fisher, is called the method of maximum likelihood. To set the ideas, we begin with a special case. Suppose that one of just two distributions must prevail. For example, let X take the possible values 0, 1, 2, or 3 with probabilities specified by distribution 1 or with probabilities specified by distribution 2 (see Table 7.1 and Figure 7.3).

Table 7.1 Two Possible Distributions for X

Distribution 1					
x	0	1	2	3	4
$f(x)$	0.0625	0.2500	0.3750	0.2500	0.0625
Distribution 2					
x	0	1	2	3	4
$f(x)$	0.2401	0.4116	0.2646	0.0756	0.0081



gure 7.3
ne two possible distributions
r X

The first is the binomial distribution with $p = 0.5$ and the second the binomial with $p = 0.4$, but this fact is not important to the argument.

If we observe $X = 3$, should our estimate of the underlying distribution be distribution 1 or distribution 2? Suppose we take the attitude that we will select the distribution for which the observed value $x = 3$ has the highest probability of occurring. Because this calculation is done after the data are obtained, we use the terminology of maximizing *likelihood* rather than probability. For the first distribution, $P[X = 3] = 0.2500$ and for the second distribution $P[X = 3] = 0.0756$, so we estimate that the first distribution is the distribution that actually produced the observation 3.

If, instead, we observed $X = 1$, the estimate would be distribution 2 since 0.4116 is larger than 0.2500.

Let us take this example a step further and assume that X follows a binomial distribution with $n = 4$ but that $0 \le p \le 1$ is unknown. The count X then has the distribution

$$\binom{n}{x} p^x (1-p)^{4-x} \quad \text{for} \quad x = 0, 1, 2, 3, 4$$

If we again observe $X = 3$, we evaluate the binomial distribution at $x = 3$ and obtain

$$4p^3(1-p)^{4-3} \quad \text{for} \quad 0 \le p \le 1$$

which is a function of p. We now vary p to best explain the observed result. This curve, $L(p)$, is shown in Figure 7.4.

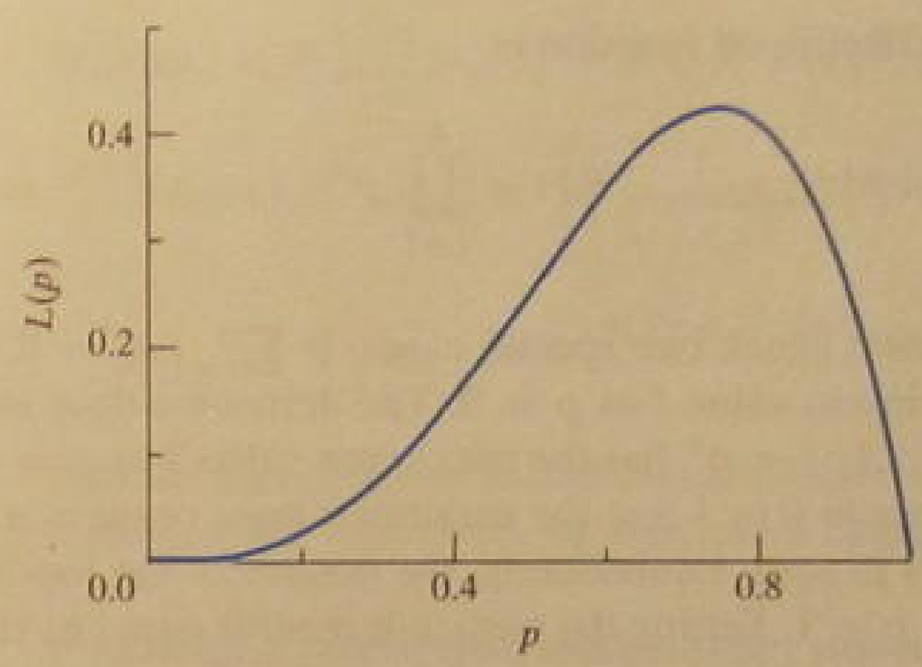


Figure 7.4
The likelihood curve
$L(p) = 4p^3(1-p)$

To obtain the best explanation for what we did observe, we choose a value for the unknown p at which the maximum occurs. Using calculus, the maximum occurs at the value of p for which the derivative is zero.

$$\frac{d}{dp} 4p^3(1-p) = 4(3p^2 - 4p^3) = 0$$

Since the solution $p = 0$ yields a minimum, our estimate is $\widehat{p} = 0.75$. Note that this derivative is positive for $p < 3/4$ and negative for $p > 3/4$, confirming that $\widehat{p} = 0.75$ gives the global maximum. To review, this value maximizes the after the fact probability, or likelihood, of observing the value 3.

More generally, a random sample of size n is taken from a probability distribution, or density, $f(x;\theta)$ that depends on a parameter θ. The random sample produces n values $x_1, x_2, \ldots, x_n$, which we substitute into the joint probability distribution, or probability density function, and then study the resulting function of θ.

The function of θ that is obtained by substituting the observed values of the random sample $X_1 = x_1, \ldots, X_n = x_n$ into the joint probability distribution or the density function for $X_1, X_2, \ldots, X_n$

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

is called the **likelihood function** for θ.

We often simplify the notation and write $L(\theta)$ with the understanding that the likelihood function does depend on the values $x_1, x_2, \ldots, x_n$ from the random sample.

Given the values $x_1, x_2, \ldots, x_n$ from a random sample, one distinctive feature of the likelihood function is the value or values of θ at which it attains its maximum.

> A statistic $\widehat{\theta}(X_1, \ldots, X_n)$ is a **maximum likelihood estimator** of θ if, for each sample $x_1, \ldots, x_n$, $\widehat{\theta}(x_1, \ldots, x_n)$ is a value for the parameter that maximizes the likelihood function $L(\theta \mid x_1, \ldots, x_n)$.

EXAMPLE **The maximum likelihood estimator with Bernoulli trials**

Consider a characteristic that occurs in proportion p of a population. Let $X_1, \ldots, X_n$ be a random sample of size n so

$$P[X_i = 0] = 1 - p \quad \text{and} \quad P[X_i = 1] = p \quad \text{for} \quad i = 1, \ldots, n$$

where $0 \le p \le 1$. Obtain the maximum likelihood estimator of p.

Solution The likelihood function is

$$L(p|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

We first check two special cases. If $\sum_{i=1}^{n} x_i = 0$, then $L(p) = (1-p)^n$ has the maximum value 1 at $p = 0$. The derivative does not vanish here. If $\sum_{i=1}^{n} x_i = n$, then $L(p) = p^n$ has the maximum value 1 at $p = 1$. Otherwise, $L(p)$ goes to 0 as p goes to 0 or 1 and the maximum must occur at a value of p where the derivative $L(p)$ is zero. Equivalently, we maximize the log-likelihood function $\ln L(p)$ over $0 < p < 1$. Setting the partial derivative equal to zero,

$$\frac{\partial}{\partial p} \ln L(p) = \frac{\partial}{\partial p}\left(\sum_{i=1}^{n} x_i \ln(p) + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)\right)$$

$$= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0$$

we obtain the maximum likelihood estimator $\widehat{p} = \sum_{i=1}^{n} x_i/n$. That is, $\widehat{p}$ is the fraction of persons in the sample that have the characteristic. Note that this definition of the estimator also includes the two special cases $\sum_{i=1}^{n} x_i = 0$ or $= 1$ even though the derivative does not vanish is these cases. ■

EXAMPLE **Maximum likelihood estimator: Poisson distribution**

Let $X_1, \ldots, X_n$ be a random sample of size n from the Poisson distribution

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where $0 \le \lambda < \infty$. Obtain the maximum likelihood estimator of λ.

Solution The likelihood function is

$$L(\lambda|x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} \times \frac{1}{\prod_{i=1}^{n} x_i!}$$

If $\sum_{i=1}^{n} x_i = 0$, then $L(\lambda) = e^{-n\lambda}$ has its maximum at $\widehat{\lambda} = 0$. Otherwise, we can maximize $\ln L(\lambda)$ by setting its derivative equal to zero.

$$\frac{\partial}{\partial \lambda} \ln L(\lambda) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0$$

so the maximum likelihood estimator is $\widehat{\lambda} = \sum_{i=1}^{n} x_i/n = \overline{x}$. This same formula works for the special case $\sum_{i=1}^{n} x_i = 0$. ■

The method of maximum likelihood also applies to continuous distributions.

EXAMPLE **Maximum likelihood estimator: normal distribution mean**

Let $X_1, \ldots, X_n$ be a random sample of size n from a normal distribution with known variance. Obtain the maximum likelihood estimator of μ.

Solution Writing $(x_i - \mu)^2 = (x_i - \overline{X} + \overline{X} - \mu)^2$, the likelihood function is

$$L(\mu|x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$
$$= e^{-n(\overline{x}-\mu)^2/2\sigma^2} \times \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(x_i-\overline{x})^2/2\sigma^2}$$

This likelihood is maximized over all values of μ when the exponent $n(\overline{x}-\mu)^2/2\sigma^2$ is minimized. Therefore, the maximum likelihood estimator $\widehat{\mu} = \overline{x}$. ■

EXAMPLE **Maximum likelihood estimator: normal distribution variance**

Let $X_1, \ldots, X_n$ be a random sample of size n from a normal distribution with known mean. Obtain the maximum likelihood estimator of σ^2.

Solution The likelihood function is

$$L(\sigma^2|x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2}$$

Recall that the function $z^b e^{-cz}$ has a maximum at $z = b/c$ for $z \geq 0$ when b and c are positive. Taking $z = 1/\sigma^2$, $b = n/2$, and $c = \sum_{i=1}^{n}(x_i - \mu)^2/2$, we obtain the maximum likelihood estimator $\widehat{\sigma^2} = \sum_{i=1}^{n}(x_i - \mu)^2/n$, when μ is known. ■

Suppose that, in the example above, we were interested in estimating σ rather than σ^2. The likelihood is still

$$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2}$$

but is now considered to be a function of σ. Taking logarithms and differentiating, you may verify that, for $d, c > 0$, $z^d e^{-cz^2}$ has a maximum at $z = \sqrt{d/2c}$ when $z > 0$. Consequently, taking $z = 1/\sigma$, $d = n$, and $c = \sum_{i=1}^{n}(x_i - \mu)^2/2$, we obtain the maximum likelihood estimator

$$\widehat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

This same argument can be extended to a more general **invariance** property for maximum likelihood estimators, which we state but do not prove. Specifically, if $g(\theta)$ is a continuous one-to-one function of θ and $\widehat{\theta}$ is the maximum likelihood estimator of θ, the maximum likelihood estimator of $g(\theta)$ is obtained by simple substitution.

$$\widehat{g(\theta)} = \text{maximum likelihood estimator of } g(\theta) = g(\widehat{\theta})$$

EXAMPLE **Maximum likelihood estimator of function of λ**

The number of defective hard drives produced daily by a production line can be modeled as a Poisson distribution. The counts for ten days are

7 3 1 2 4 1 2 3 1 2

Obtain the maximum likelihood estimate of the probability of 0 or 1 defectives on one day.

Solution From the example on page 218, the maximum likelihood estimate of λ is $\widehat{\lambda} = \overline{x} = 26/10 = 2.6$. Consequently, by the invariance property, the maximum likelihood

estimate of

$$P(X = 0 \quad \text{or} \quad 1) = e^{-\lambda} + \frac{\lambda e^{-\lambda}}{1!}$$

is

$$e^{-\hat{\lambda}} + \frac{\hat{\lambda} e^{-\hat{\lambda}}}{1!} = e^{-2.6} + \frac{2.6 \cdot e^{-2.6}}{1!} = 0.267$$

There will 1 or fewer defectives on just over one-quarter of the days. ■

The method of maximum likelihood applies to more than one parameter.

EXAMPLE **Maximum likelihood estimator: normal distributions**

Let $X_1, \ldots, X_n$ be a random sample of size n from a normal distribution. Obtain the maximum likelihood estimators of μ and σ^2. Also obtain the maximum likelihood estimator σ.

Solution Using the expression in the example for the joint probability density function (page 219), the likelihood function is

$$L(\mu, \sigma^2 \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$

$$= e^{-n(\bar{x}-\mu)^2/2\sigma^2} \times \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(x_i-\bar{x})^2/2\sigma^2}$$

Only the first term contains μ, and it is maximized at $\hat{\mu} = \bar{x}$ whatever the value of σ^2. The maximum of this first term is 1. Then, as in the example on page 219, the function $z^b e^{-cz}$ has a maximum at $z = b/c$. Taking $z = 1/\sigma^2, b = n/2$, and $c = \sum_{i=1}^{n}(x_i - \bar{x})^2/2$, we obtain the maximum likelihood estimator $\widehat{\sigma^2} = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$. In summary $\hat{\mu} = \bar{x}$ and $\widehat{\sigma^2} = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$ are the maximum likelihood estimators of μ and $\widehat{\sigma^2}$.

Here $\widehat{\sigma^2}$ contains the divisor n, not $n - 1$, so it is a biased estimate of σ^2.

Because the sample standard deviation σ is the square root of σ^2, the maximum likelihood estimator of σ is

$$\hat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}$$

Consequently, we commonly write $\hat{\sigma}^2$ for the maximum likelihood estimator of σ^2. ■

EXAMPLE **Estimation of yield for a green gas process**

One process of making green gasoline takes biomass in the form of sucrose and converts it into gasoline using catalytic reactions. At one step in a pilot plant process, the output includes carbon chains of length 3. Fifteen runs with same catalyst produced the yields (gal)

5.57	5.76	4.18	4.64	7.02	6.62	6.33	7.24
5.57	7.89	4.67	7.24	6.43	5.59	5.39	

Treating the yields as a random sample from a normal population,

(a) Obtain the maximum likelihood estimates of the mean yield and the variance.

(b) Obtain the maximum likelihood estimate of the coefficient of variation σ/μ.

Solution **(a)** We calculate

$$\widehat{\mu} = \overline{x} = \frac{5.57 + 5.76 + \cdots + 5.39}{15} = \frac{90.14}{15} = 6.009 \text{ gal}$$

Recall that the maximum likelihood estimate of variance uses divisor n, not $n - 1$.

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{15}(x_i - \overline{x})^2 = \frac{1}{15}(16.2631) = 1.084$$

(b) The coefficient of variation is a function of μ and σ^2, so its maximum likelihood estimate is that same function of $\widehat{\mu}$ and $\widehat{\sigma^2}$.

$$\widehat{\left(\frac{\sigma}{\mu}\right)} = \frac{\widehat{\sigma}}{\widehat{\mu}} = \frac{\sqrt{1.084}}{6.0109} = 0.173$$

■

Exercises

7.30 Refer to the example on page 57, where 294 out of 300 ceramic insulators were able to survive a thermal shock.

(a) Obtain the maximum likelihood estimate of the probability that a ceramic insulator will survive a thermal shock.

(b) Suppose a device contains 3 ceramic insulators and all must survive the shock in order for the device to work. Find the maximum likelihood estimate of the probability that all three will survive a thermal shock.

7.31 Refer to the example on page 284, where 48 of 60 transceivers passed inspection.

(a) Obtain the maximum likelihood estimate of the probability that a transceiver will pass inspection.

(b) Obtain the maximum likelihood estimate that the next two transceivers tested will pass inspection.

7.32 The daily number of accidental disconnects with a server follows a Poisson distribution. On five days

2 5 3 3 7

accidental disconnects are observed.

(a) Obtain the maximum likelihood estimate of λ.

(b) Find the maximum likelihood estimate of the probability that 3 or more accidental disconnects will occur.

7.33 In one area along the interstate, the number of dropped wireless phone connections per call follows a Poisson distribution. From four calls, the number of dropped connections is

2 0 3 1

(a) Find the maximum likelihood estimate of λ.

(b) Obtain the maximum likelihood estimate that the next two calls will be completed without any accidental drops.

7.34 Refer to Exercises 7.12 and 7.13.

(a) Obtain the maximum likelihood estimates of μ and σ and 7.13.

(b) Find the maximum likelihood of the probability that the next run will have a yield greater than 2 gallons.

7.35 Refer to Exercises 7.14 and 7.15.

(a) Obtain the maximum likelihood estimates of μ and σ.

(b) Find the maximum likelihood of the coefficient of variation σ/μ.

7.36 Find the maximum likelihood estimator of p when

$$f(x; p) = p^x(1-p)^{1-x} \quad \text{for } x = 0, 1$$

7.37 Let $x_1, \ldots, x_n$ be the observed values of a random sample of size n from the exponential distribution $f(x; \beta) = \beta^{-1}e^{-x/\beta}$ for $x > 0$.

(a) Find the maximum likelihood estimator of β.

(b) Obtain the maximum likelihood estimator of the probability that the next observation is greater than 1.

7.38 Let X have the negative binomial distribution

$$f(x) = \binom{x-1}{r-1} p^r(1-p)^{x-r} \quad \text{for } x = r, r+1, \ldots$$

(a) Obtain the maximum likelihood estimator of p.

(b) For one engineering application, it is best to use components with a superior finish. Suppose $X = 27$ identical components are inspected, one at a time, before the $r = 3$rd component with superior finish is found. Find the maximum likelihood estimate of the probability that a component will have a superior finish.

7.4 Tests of Hypotheses

There are many problems in which, rather than estimate the value of a parameter, we must decide whether a statement concerning a parameter is true or false; that is, we must test a hypothesis about a parameter. For example, in quality control work a random sample may serve to determine whether the "process mean" (for a given kind of measurement) has remained unchanged or whether it has changed to such an extent that the process has gone out of control and adjustments have to be made.

EXAMPLE **Not all samples will lead to a correct assessment of water quality**

Refer to the example on page 178 of monitoring the quality of water leaving a plant. Why does evaluating a sample of specimens not always lead to correct conclusions regarding water quality?

Solution The observed values of water quality will depend on the particular specimens in the sample. Because these values can vary from sample to sample, particular samples can produce misleading values and hence incorrect decisions. The possibility of making a mistake about water quality, on the basis of test specimens, cannot be completely eliminated unless the entire discharge could be accurately measured for the entire reporting period. This is, of course, technologically and economically infeasible. ■

To illustrate the general concepts involved in deciding whether or not a statement about the population is true or false, suppose that a consumer protection agency wants to test a paint manufacturer's claim that the average drying time of his new "fast-drying" paint is 20 minutes. It instructs a member of its research staff to paint each of 36 boards using a different 1-gallon can of the paint, with the intention of rejecting the claim if the mean of the drying times exceeds 20.75 minutes. Otherwise, it will accept the claim, and in either case it will take whatever action is called for in its plans.

This provides a clear-cut criterion for accepting or rejecting the claim, but unfortunately it is not infallible. Since the decision is based on a sample, there is the possibility that the sample mean may exceed 20.75 minutes even though the true mean drying time is $\mu = 20$ minutes, and there is also the possibility that the sample mean may be 20.75 minutes or less even though the true mean drying time is, say, $\mu = 21$ minutes. Thus, before adopting the criterion, it would seem wise to investigate the chances that the criterion may lead to a wrong decision.

Assuming that it is known from past experience that the standard deviation of such drying times can be expected to equal $\sigma = 2.4$ minutes, let us first investigate the possibility that the sample mean may exceed 20.75 minutes even though the true mean drying time is $\mu = 20$. The probability that this will happen purely due to chance is given by the area of the ruled region of Figure 7.5, and it can easily be determined by approximating the sampling distribution of the mean with a normal

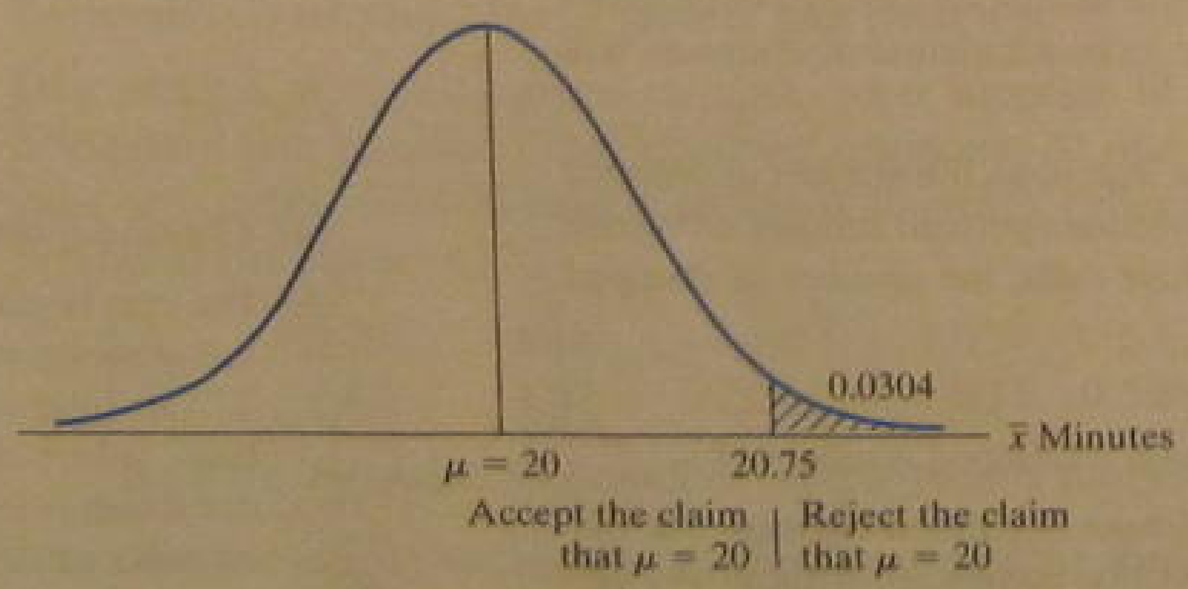


Figure 7.5
Probability of falsely rejecting claim

curve. Assuming that the population is large enough to be treated as an infinite, we have $\sigma_{\overline{X}} = \dfrac{2.4}{\sqrt{36}} = 0.4$, and the dividing line of the criterion in standard units is

$$z = \frac{20.75 - 20}{0.4} = 1.875$$

It follows from Table 3 that the area of the ruled region of Figure 7.5 is $1 - 0.9696 = 0.0304$ (by interpolation). Hence the probability of erroneously rejecting the hypothesis $\mu = 20$ minutes is approximately 0.03.

Let us now consider the other possibility, where the procedure fails to detect that $\mu > 20$ minutes. Suppose again, for the sake of argument, that the true mean drying time is $\mu = 21$ minutes, so that the probability of getting a sample mean less than or equal to 20.75 minutes (and, hence, erroneously accepting the claim that $\mu = 20$ minutes) is given by the area of the ruled region of Figure 7.6. As before, $\sigma_{\overline{X}} = 0.4$, so that the dividing line of the criterion, in standard units, is now

$$z = \frac{20.75 - 21}{0.4} = -0.625$$

It follows from Table 3 that the area of the ruled region of Figure 7.6 is 0.2660 (by interpolation), and hence that the probability of erroneously accepting, or failing to reject, the hypothesis $\mu = 20$ is approximately 0.27.

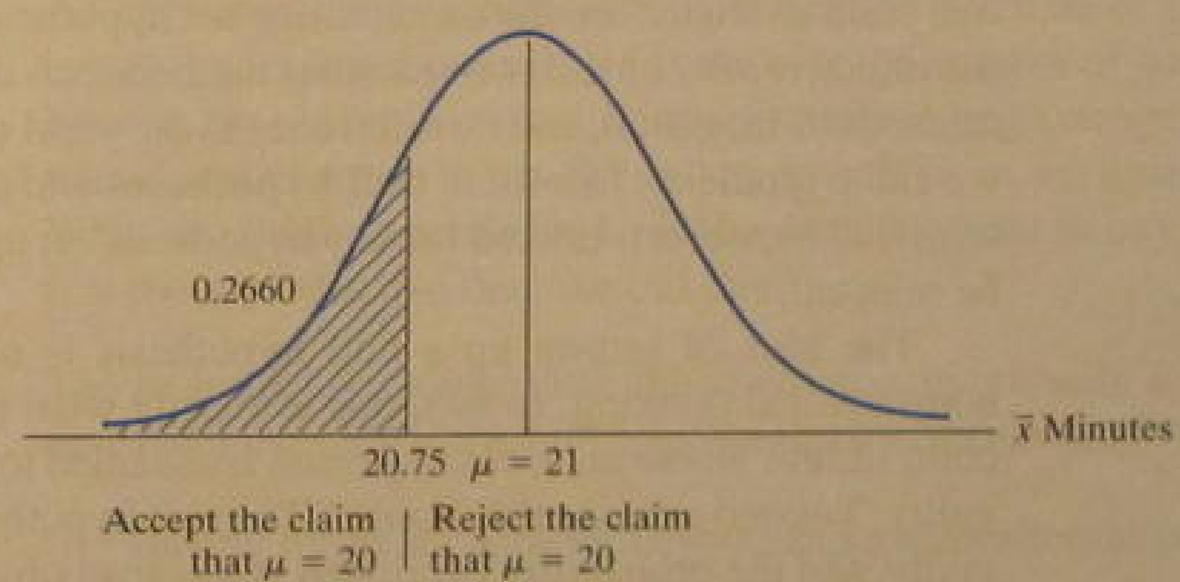


Figure 7.6
Probability of failing to reject claim

The situation described in this example is typical of testing a statistical hypothesis, and it may be summarized in the following table, where we refer to the hypothesis being tested as hypothesis H:

	Accept H	*Reject H*
H is true	Correct decision	Type I error
H is false	Type II error	Correct decision

If hypothesis H is true and not rejected or false and rejected, the decision is in either case correct. If hypothesis H is true but rejected, it is rejected in error, and if hypothesis H is false but not rejected, this is also an error. The first of these errors is called a **Type I error**. The probability of committing it, when the hypothesis is true, is designated by the Greek letter α (alpha). The second error is called a **Type II error** and the probability of committing it is designated by the Greek letter β (beta). Thus, in our example we showed that for the given test criterion, $\alpha = 0.03$ when $\mu = 20$ minutes, and $\beta = 0.27$ when $\mu = 21$ minutes.

In calculating the probability of a Type II error in our example, we arbitrarily choose the alternative value $\mu = 21$ minutes. However, in this problem, as in most

others, there are infinitely many other alternatives, and for each one of them there is a positive probability β of erroneously accepting the hypotheses H. What to do about this will be discussed further in Section 7.8.

7.5 Null Hypotheses and Tests of Hypotheses

In the drying time example of the preceding section, we were able to calculate the probability of a Type I error because we formulated the hypothesis H as a single value for the parameter μ; that is, we formulated the hypothesis H so that μ was completely specified. Had we formulated instead $\mu \leq 20$ minutes, where μ can take on more than one possible value, we would not have been able to calculate the probability of a Type I error without specifying by how much μ is less than 20 minutes.

We often formulate hypotheses to be tested as a single value for a parameter; at least, we do this whenever possible. This usually requires that we hypothesize the opposite of what we hope to prove. For instance, if we want to show that one method of teaching computer programming is more efficient than another, we hypothesize that the two methods are equally effective. Similarly, if we want to show that one method of irrigating the soil is more expensive than another, we hypothesize that the two methods are equally expensive; and if we want to show that a new copper-bearing steel has a higher yield strength than ordinary steel, we hypothesize that the two yield strengths are the same. Since we hypothesize that there is no difference in the effectiveness of the two teaching methods, no difference in the cost of the two methods of irrigation, and no difference in the yield strength of the two kinds of steel, we call hypotheses like these **null hypotheses** and denote them H_0. Nowadays, the term *null hypothesis* is used for any hypothesis set up primarily to see whether it can be rejected.

The idea of setting up a null hypothesis is not an uncommon one, even in nonstatistical thinking. In fact, this is exactly what is done in an American criminal court of law, where an accused person is assumed to be innocent unless he is proved guilty "beyond a reasonable doubt." The null hypothesis states that the accused is not guilty, and the probability expressed subjectively by the phrase "beyond reasonable doubt" reflects the probability α of risking a Type I error. Note that the "burden of proof" is always on the prosecution in the sense that the accused is found not guilty unless the null hypothesis of innocence is clearly disproved. This does not imply that the defendant has been proved innocent if found not guilty; it implies only that he has not been proved guilty. Of course, since we cannot legally "reserve judgment" if proof of guilt is not established, the accused is freed and we act as if the null hypothesis of innocence were accepted. Note that this is precisely what we may have to do in tests of statistical hypotheses, when we cannot afford the luxury of reserving judgment.

We will develop the procedures for hypothesis testing in the context of the drying-time example. When the goal is to refute the "fast-drying" claim, the alternative hypothesis should be $\mu > 20$ minutes.

To approach problems of hypothesis testing systematically, it will help to proceed as outlined in the following five steps.

1. **We formulate a null hypothesis and an appropriate alternative hypothesis which we accept when the null hypothesis must be rejected.**[1]

[1] See also the discussion on page 227.

In the drying time example, the null hypothesis is $\mu = 20$ minutes and the alternative hypothesis is $\mu > 20$ minutes. This kind of alternative, which is called a **one-sided alternative**, may also have the inequality going the other way. For instance, if we hope to be able to show that the average time required to do a certain job is less than 15 minutes, we would test the null hypothesis $\mu = 15$ against the alternative hypothesis $\mu < 15$.

The following is an example in which we would use the **two-sided alternative** $\mu \neq \mu_0$, where μ_0 is the value assumed under the null hypothesis: A food processor wants to check whether the average amount of coffee that goes into his 4-ounce jars is indeed 4 ounces. Since the food processor cannot afford to put much less than 4 ounces into each jar for fear of losing customer acceptance, nor can he afford to put much more than 4 ounces into each jar for fear of losing part of his profit, the appropriate alternative hypothesis is $\mu \neq 4$.

As in the examples of the two preceding paragraphs, **alternative hypotheses** usually specify that the population mean (or whatever other parameter may be of concern) is either not equal to, greater than, or less than the value assumed under the null hypothesis. For any given problem, the choice of an appropriate alternative depends mostly on what we hope to be able to show, or better, where we want to put the burden of proof.

EXAMPLE **Formulating the alternative hypothesis**

An appliance manufacturer is considering the purchase of a new machine for stamping out sheet metal parts. If μ_0 is the average number of good parts stamped out per hour by her old machine and μ is the corresponding average for the new machine, the manufacturer wants to test the null hypothesis $\mu = \mu_0$ against a suitable alternative. What should the alternative be if she does not want to buy the new machine unless it is more productive than the old one?

Solution The manufacturer should use the alternative hypothesis $\mu > \mu_0$ and purchase the new machine only if the null hypothesis can be rejected. ■

Having formulated the null hypothesis and an alternative hypothesis, we proceed with the following step:

2. **We specify the probability of a Type I error. If possible, desired, or necessary, we may also specify the probabilities of Type II errors for particular alternatives.**

The probability of a Type I error is also called the **level of significance**, and it is usually set at $\alpha = 0.05$ or $\alpha = 0.01$. Which value we choose in any given problem will have to depend on the risks, or consequences, of committing a Type I error. Observe, however, that we should not make the probability of a Type I error too small, because this will have the tendency to make the probabilities of serious Type II errors too large.

Step 2 can often be performed even when the null hypothesis specifies a range of values for the parameter. To illustrate, let us investigate briefly what might be done in the drying time example if we wanted to allow for the possibility that the paint may be better (faster drying) than claimed, and hence test the null hypothesis $\mu \leq 20$ minutes against the alternative hypothesis $\mu > 20$ minutes. Observe that if μ is less than 20 minutes, the normal curve of Figure 7.5 on page 222 is shifted to the left, and the area under the curve to the right of 20.75 becomes less than 0.0304. Thus, if the null hypothesis is $\mu \leq 20$ minutes, we can say that the probability of a Type I error is at most 0.0304, and we write $\alpha \leq 0.0304$. In general, if the null hypothesis is of the form $\mu \leq \mu_0$ or $\mu \geq \mu_0$, we can only specify the maximum

probability of a Type I error, and by performing the test as if the null hypothesis were $\mu = \mu_0$, we protect ourselves against the worst possibility. (See the example on page 232.)

After the null hypothesis, the alternative hypothesis, and the level of significance have been specified, the remaining steps are as follows:

3. **Based on the sampling distribution of an appropriate statistic, we construct a criterion for testing the null hypothesis against the given alternative.**
4. **We calculate from the data the value of the statistic on which the decision is to be based.**
5. **We decide whether to reject the null hypothesis or whether to fail to reject it.**

In the drying time example we studied the criterion using the normal-curve approximation to the sampling distribution of the mean. In general, step 3 depends not only on the statistic on which we want to base the decision and on its sampling distribution, but also on the alternative hypothesis we happen to choose. In the drying time example we used a **one-sided criterion (one-sided test** or **one-tailed test)** with the one-sided alternative $\mu > 20$ minutes, rejecting the null hypothesis only for large values of the sample mean. In the example dealing with the food processor and his 4-ounce jars of coffee, we would use a **two-sided criterion (two-sided test** or **two-tailed test)** to go with the two-sided alternative $\mu \neq 4$ ounces. In general, a test is said to be two-sided if the null hypothesis is rejected for values of the test statistic falling into either tail of its sampling distribution.

The purpose of the discussion in this and in the preceding section has been to introduce some of the basic problems connected with the testing of statistical hypotheses. Although the methods we have presented are objective—that is, two experimenters analyzing the same data under the same conditions would arrive at identical results—their use does entail some arbitrary, or subjective, considerations. For instance, in the example on page 222 it was partially a subjective decision to draw the line between satisfactory and unsatisfactory values of μ at 21 minutes. It is also partially a subjective decision to use a sample of 36 one-gallon cans of paint, and to reject the manufacturer's claim for values of $\overline{X}$ exceeding 20.75 minutes. Approaching the problem differently, the government agency investigating the paint manufacturer's claim could have specified values of α and β, thus controlling the risks to which they are willing to be exposed. The choice of α, the probability of a Type I error, could have been based on the consequences of making that kind of error, namely, the manufacturer's cost of having a good product condemned, the possible cost of subsequent litigation, the manufacturer's cost of unnecessarily adjusting his machinery, the cost to the public of not having the product available when needed, and so forth. The choice of β, the probability of a Type II error, could similarly have been based on the consequences of making that kind of error. Namely, the cost to the public of buying an inferior product, the manufacturer's savings in using inferior ingredients but loss in good will, again the cost of possible ligation, and so forth. It should be obvious that it would be extremely difficult to put cash values on all these eventualities, but they must nevertheless be considered, at least indirectly, in choosing suitable criteria for testing statistical hypotheses.

In this text we shall discuss mainly the **Neyman-Pearson theory**, also called the **classical theory of testing hypothesis**. This means that we shall consider cost factors and other considerations that are partly arbitrary and partly subjective only insofar as they will affect the choice of a sample size, the choice of an alternative hypothesis, the choice of α and β, and so forth.

In this approach, the maximum value of α is controlled over the null hypothesis. For instance, if the null hypothesis is $\mu \leq \mu_0$, then μ as well as σ is unspecified. This is a **composite hypothesis**. Otherwise, if all the parameters are completely specified, the hypothesis is **simple**. Even with a composite null hypothesis, we set the critical region so that the error probability is α on the boundary $\mu = \mu_0$. Then, the error probabilities will be even smaller under values of μ that are less than μ_0.

Because the probability of falsely rejecting the null hypothesis is controlled, the null hypothesis is retained unless the observations strongly contradict it. Consequently, if the goal of an experiment is to establish an assertion or hypothesis, that hypothesis must be taken as the alternative hypothesis.

Guideline for selecting the null hypothesis

When the goal of an experiment is to establish an assertion, the negation of the assertion should be taken as the null hypothesis. The assertion becomes the alternative hypothesis.

Similar reasoning suggests that even when costs are difficult to determine but the consequences of one error are much more serious than for the other, the hypotheses should be labeled so that the most serious error is the Type I error.

The alternative hypothesis is often denoted by H_1. In the paint drying example, H_1: $\mu > 20$. Recalling the notation H_0 for the null hypothesis, we summarize the concepts and notation for hypotheses, types of error, and the probability of error.

Notation for the hypotheses

H_1: The **alternative hypothesis** is the claim we wish to establish.
H_0: The **null hypothesis** is the negation of the claim.

The two kinds of error and their probabilities are

The errors and their probabilities

Type I error: Rejection of H_0 when H_0 is true.
Type II error: Nonrejection of H_0 when H_1 is true.
α = probability of making a Type I error (also called the **level of significance**)
β = probability of making a Type II error

It is important to understand that tests of hypotheses are structured to control the probability, α, of falsely rejecting the null hypothesis. To interpret this, suppose the null hypothesis prevails and the test is repeated many times with independent sets of data. When $\alpha = 0.05$, the law of large numbers tells us that

$$\frac{\text{number of times the null hypothesis is falsely rejected}}{\text{number of times a test is conducted}} \to \alpha$$

This is, of course, the long-run frequency interpretation of probability. In the long run only 1 in 20 times will the null hypothesis be falsely rejected.

Before we discuss various special tests about means in the remainder of this chapter, let us point out that the concepts we have introduced here apply equally well to hypotheses concerning proportions, standard deviations, the randomness of samples, and relationships among several variables.

Exercises

7.39 A computer manufacturer wants to establish that the average time to set up a new desktop computer is less than 2 hours.

(a) Formulate the null and alternative hypotheses.

(b) What error could be made if $\mu = 1.9$? Explain in the context of the problem.

(c) What error could be made if $\mu = 2.0$? Explain in the context of the problem.

7.40 A manufacturer of four-speed clutches for automobiles claims that the clutch will not fail until after 50,000 miles.

(a) Interpreting this as a statement about the mean, formulate a null and alternative hypothesis for verifying the claim.

(b) If the true mean is 55,000 miles, what error can be made? Explain your answer in the context of the problem.

(c) What error could be made if the true mean is 50,000 miles?

7.41 An airline claims that the typical flying time between two cities is 56 minutes.

(a) Formulate a test of hypotheses with the intent of establishing that the population mean flying time is different from the published time of 56 minutes.

(b) If the true mean is 50 minutes, what error can be made? Explain your answer in the context of the problem.

(c) What error could be made if the true mean is 56 minutes?

7.42 A company wants to establish that the mean life of its batteries, when used in a wireless mouse, is over 183 days. The data will consist of the life lengths of batteries in 64 different wireless mice.

(a) Formulate the null and alternative hypotheses.

(b) If the true mean is 190 days, what error can be made? Explain your answer in the context of the problem.

7.43 A statistical test of hypotheses includes the step of setting a maximum for the probability of falsely rejecting the null hypothesis. Engineers make many measurements on critical bridge components to decide if a bridge is safe or unsafe.

(a) Explain how you would formulate the null hypothesis.

(b) Would you prefer $\alpha = 0.05$ or $\alpha = 0.01$? Explain your reasoning.

7.44 Suppose you are scheduled to ride a space vehicle that will orbit the earth and return. A statistical test of hypotheses includes the step of setting a maximum for the probability of falsely rejecting the null hypothesis. Engineers need to make various measurements to decide if it is safe or unsafe to launch the vehicle.

(a) Explain how you would formulate the null hypothesis.

(b) Would you prefer $\alpha = 0.05$ or $\alpha = 0.01$? Explain your reasoning.

7.45 Suppose that an engineering firm is asked to check the safety of a dam. What type of error would it commit if it erroneously rejects the null hypothesis that the dam is safe? What type of error would it commit if it erroneously accepts the null hypothesis that the dam is safe? Would would the likely impact of these errors be?

7.46 Suppose that we want to test the null hypothesis that an antipollution device for cars is effective. Explain under what conditions we would commit a Type I error and under what conditions we would commit a Type II error.

7.47 If the criterion on page 222 is modified so that the paint manufacturer's claim is rejected for $\overline{X} > 20.50$ minutes, find

(a) the probability of a Type I error;

(b) the probability of a Type II error when $\mu = 21$ minutes.

7.48 Suppose that in the drying time example on page 222, n is changed from 36 to 50 while everything else remains the same. Find

(a) the probability of a Type I error;

(b) the probability of a Type II error when $\mu = 21$ minutes.

7.49 It is desired to test the null hypothesis $\mu = 100$ pounds against the alternative hypothesis $\mu < 100$ pounds on the basis of the weights of a random sample of size $n = 40$ packages shipped by truck. The population has $\sigma = 12$ pounds. For what values of $\overline{X}$ must the null hypothesis be rejected if the probability of a Type I error is to be $\alpha = 0.01$?

7.50 Several square inches of gold leaf are required in the manufacture of a high-end component. Suppose that, the population of the amount of gold leaf has $\sigma = 8.4$ square inches. We want to test the null hypothesis $\mu = 80.0$ square inches against the alternative hypothesis $\mu < 80.0$ square inches on the basis of a random sample of size $n = 100$.

(a) If the null hypothesis is rejected for $\overline{X} < 78.0$ square inches and otherwise it is accepted, what is the probability of Type I error?

(b) What is the answer to part (a) if the null hypothesis is $\mu \geq 80.0$ square inches instead of $\mu = 80.0$ square inches?

7.51 A producer of extruded plastic products finds that his mean daily inventory is 1,250 pieces. A new marketing policy has been put into effect and it is desired to test the null hypothesis that the mean daily inventory is still the same. What alternative hypothesis should be used if

(a) it is desired to know whether or not the new policy changes the mean daily inventory;

(b) it is desired to demonstrate that the new policy actually reduces the mean daily inventory;

(c) the new policy will be retained so long as it cannot be shown that it actually increases the mean daily inventory?

7.52 Specify the null hypothesis and the alternative hypothesis in each of the following cases.

(a) An automobile manufacturer wants to confirm a supplier's claim that the maximum resistance in a wiring harness is less than 50 ohms.

(b) An investigator wants to establish the research department's claim that a new ballast will increase mean bulb life to more than 5,000 hours.

7.6 Hypotheses Concerning One Mean

Having used tests concerning means to illustrate the basic principles of hypothesis testing, let us now see how we proceed in actual practice. Suppose, for instance, we want to establish that the thermal conductivity of a certain kind of cement brick differs from 0.340, the value claimed. We will test on the basis of $n = 35$ determinations and at the 0.05 level of significance. From information gathered in similar studies, we can expect that the variability of such determinations is given by $\sigma = 0.010$.

Following the outline of the preceding section, we begin with steps 1 and 2 by writing

1. *Null hypothesis*: $\mu = 0.340$
   Alternative hypothesis: $\mu \neq 0.340$
2. *Level of significance*: $\alpha = 0.05$

The alternative hypothesis is two-sided because we shall want to reject the null hypothesis if the mean of the determinations is significantly less than or significantly greater than 0.340.

Next, in step 3, we depart from the procedure used in the example of the preceding section and base the test on the standardized statistic

Statistic for test concerning mean (σ known)

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

instead of $\overline{X}$. The reason for working with standard units, or Z values, is that it enables us to formulate criteria which are applicable to a great variety of problems, not just one.

If z_α is, as before, such that the area under the standard normal curve to its right equals α, the **critical regions**, namely, the sets of values of Z for which we reject

the null hypothesis $\mu = \mu_0$, can be expressed as in the following table:

Critical Regions for Testing $\mu = \mu_0$ (normal population and σ known)	
Alternative hypothesis	**Reject null hypothesis if:**
$\mu < \mu_0$	$Z < -z_\alpha$
$\mu > \mu_0$	$Z > z_\alpha$
$\mu \neq \mu_0$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

If $\alpha = 0.05$, the dividing lines, or **critical values**, of the criteria are -1.645 and 1.645 for the one-sided alternatives, and -1.96 and 1.96 for the two-sided alternative. If $\alpha = 0.01$, the dividing lines of the criteria are -2.33 and 2.33 for the one-sided alternatives, and -2.575 and 2.575 for the two-sided alternative. These results come from the example on page 128 and Exercise 5.23.

Returning now to the example dealing with the thermal conductivity of the cement bricks, suppose that the mean of the 35 determinations is 0.343. So we continue by writing

3. *Criterion*: Reject the null hypothesis if $Z < -1.96$ or $Z > 1.96$, where

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

4. *Calculations*:

$$Z = \frac{0.343 - 0.340}{0.010/\sqrt{35}} = 1.77$$

5. *Decision*: Since $Z = 1.77$ falls on the interval from -1.96 to 1.96, the null hypothesis cannot be rejected; to put it another way, the difference between $\overline{x} = 0.343$ and $\mu = 0.340$ can be attributed to chance. We actually never establish that the null hypothesis holds. Instead, the conclusion is that we fail to reject the null hypothesis.

In problems like this, many research workers accompany the calculated value of Z with a corresponding **tail probability**, or ***P*-value**, which is the probability of getting a difference between $\overline{x}$ and μ_0 greater than or equal to that actually observed. For instance, in the above example, it is given by the total area under the standard normal curve to the left of -1.77 and to the right of 1.77, and it equals $2(1 - 0.9616) = 0.0768$. This P-value exceeds 0.05, which agrees with our earlier result. Figure 7.7 shows that the rejection region, the value of Z, and the two tail areas.

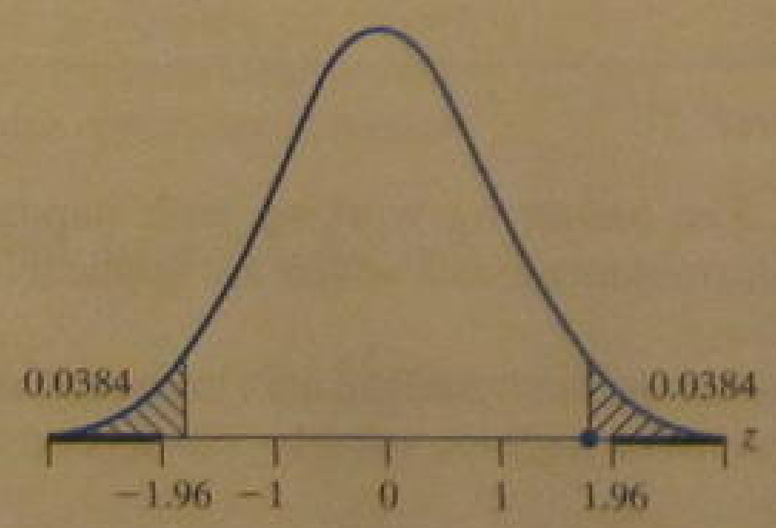


Figure 7.7
P-value $= 0.0384 + 0.0384 = 0.0768$ when $Z = 1.77$

EXAMPLE **Calculating the P-value**

A process for producing vinyl floor covering has been stable for a long period of time, and the surface hardness measurement of the flooring produced has a normal distribution with mean 4.5 and standard deviation $\sigma = 1.5$. A second shift has been hired and trained and their production needs to be monitored. Consider testing the hypothesis $H_0: \mu = 4.5$ versus $H_1: \mu \neq 4.5$. A random sample of hardness measurements is made of $n = 25$ vinyl specimens produced by the second shift. Calculate the P-value when using the test statistic

$$Z = \frac{\overline{X} - 4.5}{1.5/\sqrt{25}}$$

if $\overline{X} = 3.9$.

Solution The observed value of the test statistic is

$$z = \frac{3.9 - 4.5}{1.5/\sqrt{25}} = -2.00$$

Since the alternative hypothesis is two-sided, we must consider large negative values for Z as well as large positive values (see Figure 7.8).

From the normal table, $P(Z > 2.00) = 0.0228$. The probability that Z is smaller than -2.00 is also 0.0228. Consequently, the P-value is $0.0228 + 0.0228 = 0.0456$. This is the probability, under the null hypothesis, of getting the same or a more extreme value of the test statistic than was observed. The small P-value suggests that the mean of the second shift is not at the target value of 4.5. ■

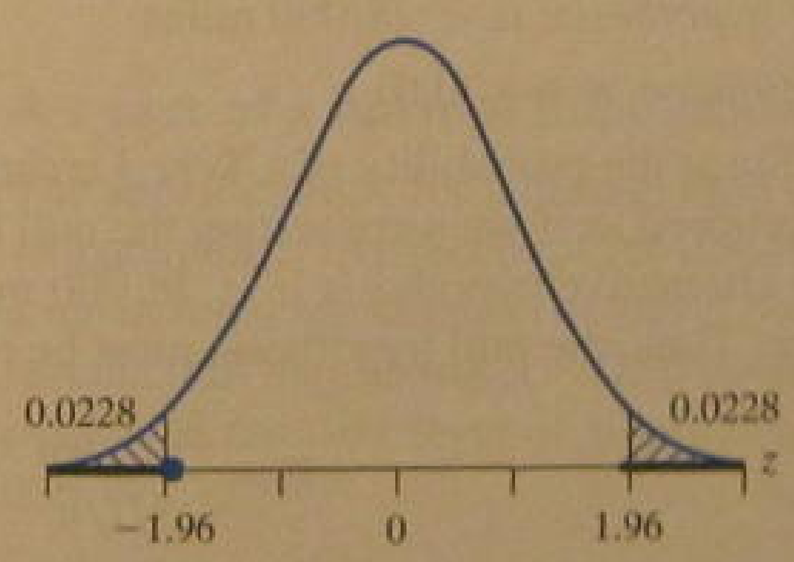


Figure 7.8
The P-value $= 0.0228 + 0.0228 = 0.0456$ for $Z = -2.00$

The P-value can be defined for any test statistic in any hypotheses testing problem.

P-value for a given test statistic and null hypothesis

The **P-value** is the probability of obtaining a value for the test statistic that is as extreme or more extreme than the value actually observed.
Probability is calculated under the null hypothesis.

If the alternative hypothesis is right-sided—for instance, $H_1: \mu > \mu_o$—then only values greater than the observed value are more extreme. If the alternative hypothesis is left-sided, only values less than the observed value are more extreme. For two-sided alternatives, values in both tails need to be considered as in the example above.

Observe that giving a tail probability does not relieve us of the responsibility of specifying the level of significance before the test is actually performed.

The test we have described in this section is essentially an approximate large sample test; it is exact only when the population we are sampling is normal and σ is known. In many practical situations where σ is unknown but the sample size is large,

we must make the further approximation of substituting for it the sample standard deviation S.

Statistic for large sample test concerning mean

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

Critical Regions for Testing $\mu = \mu_0$ (large sample)	
Alternative hypothesis	**Reject null hypothesis if:**
$\mu < \mu_0$	$Z < -z_\alpha$
$\mu > \mu_0$	$Z > z_\alpha$
$\mu \neq \mu_0$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

EXAMPLE **A large sample test of mean tire life**

A trucking firm is suspicious of the claim that the average lifetime of certain tires is at least 28,000 miles. To check the claim, the firm puts 40 of these tires on its trucks and gets a mean lifetime of 27,463 miles with a standard deviation of 1,348 miles. What can it conclude if the probability of a Type I error is to be at most 0.01?

Solution

1. *Null hypothesis*: $\mu \geq 28{,}000$ miles
   Alternative hypothesis: $\mu < 28{,}000$ miles
2. *Level of significance*: $\alpha \leq 0.01$
3. *Criterion*: Since the probability of a Type I error is greatest when $\mu = 28{,}000$ miles, we proceed as if we were testing the null hypothesis $\mu = 28{,}000$ miles against the alternative hypothesis $\mu < 28{,}000$ miles at the 0.01 level of significance. Thus, the null hypothesis must be rejected if $Z < -2.33$, where
$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$
4. *Calculations*:
$$Z = \frac{27{,}463 - 28{,}000}{1{,}348/\sqrt{40}} = -2.52$$
5. *Decision*: Since $Z = -2.52$ is less than -2.33, the null hypothesis must be rejected at level of significance 0.01. In other words, the trucking firm's suspicion that $\mu < 28{,}000$ miles is confirmed.

The small P-value 0.0059, as shown in Figure 7.9, strengthens the conclusion. ■

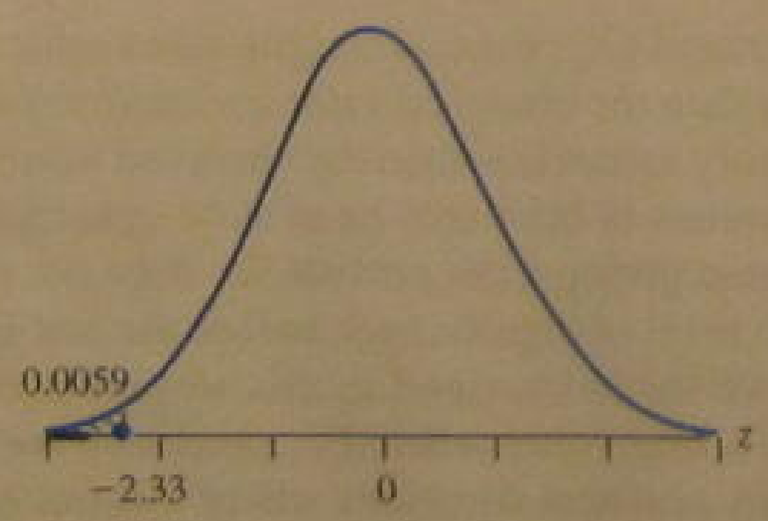


Figure 7.9
The P-value = 0.0059 for $Z = -2.52$

If the sample size is small and σ is unknown, the tests just described cannot be used. However, if the sample comes from a normal population (to within a reasonable degree of approximation), we can make use of the theory discussed in Section 6.3 and base the test of the null hypothesis $\mu = \mu_0$ on the statistic

Statistic for small sample test concerning mean (normal population)

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

which is a random variable having the t distribution with $n - 1$ degrees of freedom. The criteria for the **one sample t test** are like those for Z but are based on the t distribution.

Critical Regions for Testing $\mu = \mu_0$ (normal population and σ unknown) one sample t test

Alternative hypothesis	Reject null hypothesis if:
$\mu < \mu_0$	$t < -t_\alpha$
$\mu > \mu_0$	$t > t_\alpha$
$\mu \neq \mu_0$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

where t_α and $t_{\alpha/2}$ are based on $n - 1$ degrees of freedom.

EXAMPLE **A t test of a normal population mean**

The specifications for a certain kind of ribbon call for a mean breaking strength of 180 pounds. If five pieces of the ribbon (randomly selected from different rolls) have a mean breaking strength of 169.5 pounds with a standard deviation of 5.7 pounds, test the null hypothesis $\mu = 180$ pounds against the alternative hypothesis $\mu < 180$ pounds at the 0.01 level of significance. Assume that the population distribution is normal.

Solution

1. *Null hypothesis*: $\mu = 180$ pounds
   Alternative hypothesis: $\mu < 180$ pounds
2. *Level of significance*: $\alpha = 0.01$
3. *Criterion*: Reject the null hypothesis if $t < -3.747$, where 3.747 is the value of $t_{0.01}$ for $5 - 1 = 4$ degrees of freedom and

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

4. *Calculations*:

$$t = \frac{169.5 - 180}{5.7/\sqrt{5}} = -4.12$$

5. *Decision*: Since $t = -4.12$ is less than -3.747, the null hypothesis must be rejected at level $\alpha = 0.01$. In other words, the breaking strength is below specifications. The exact tail probability, or P-value, cannot be determined from Table 4, but it is 0.0073 (see Figure 7.10). The evidence against the mean breaking strength being 180 pounds is even stronger than 0.01. Only about 7 in

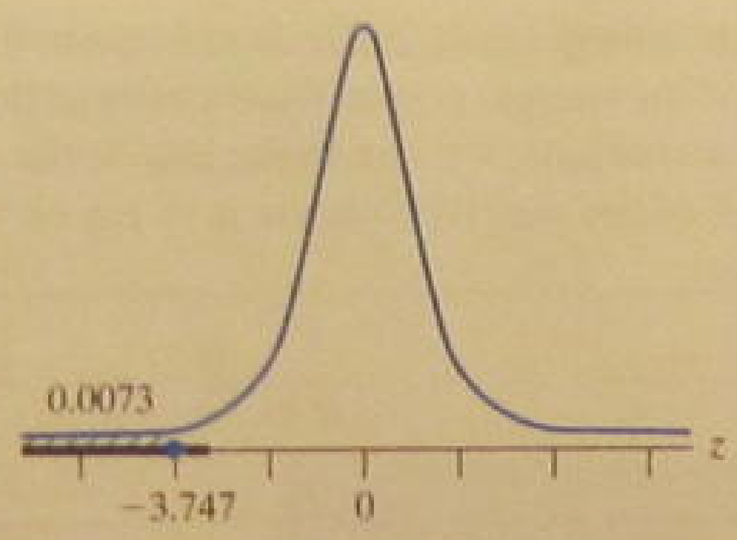


Figure 7.10
P-value = 0.0073 for $t = -4.12$

1,000 times would we observe a value of t that is -4.12 or smaller, if the mean really were 180 pounds. ■

Exercises

7.53 Refer to Exercise 7.1 where civil engineers recorded the amount of salt (tons) used to keep highways drivable during a snowstorm. The amount of salt for $n = 30$ storms has $\bar{x} = 1798.4$ tons and $s^2 = 671{,}330.9$, so $s = 819.35$ tons.

(a) Conduct a test of hypotheses with the intent of showing that the mean salt usage during a snowstorm is less than 2,000 tons. Take $\alpha = 0.05$.

(b) Based on your conclusion in part (a), what error could you have made? Explain in the context of the problem.

7.54 Refer to data in Exercise 7.3 on the labor time required to produce an order of automobile mufflers using a heavy stamping machine. The times (hours) for $n = 52$ orders of different parts has $\bar{x} = 1.865$ hours and $s^2 = 1.5623$, so $s = 1.250$ hours.

(a) Conduct a test of hypotheses with the intent of showing that the mean labor time is more than 1.5 hours. Take $\alpha = 0.05$.

(b) Based on your conclusion in part (a), what error could you have made? Explain in the context of the problem.

7.55 Refer to Exercise 7.5, where the number of unremovable defects, for each of $n = 45$ displays, has $\bar{x} = 2.667$ and $s = 3.057$ unremovable defects.

(a) Conduct a test of hypotheses with the intent of showing that the mean number of unremovable defects is less than 3.6. Take $\alpha = 0.025$.

(b) Based on your conclusion in part (a), what error could you have made? Explain in the context of the problem.

7.56 Refer to Exercise 7.12, where, at one step in a pilot plant process, the yield (gal) of carbon chains of length three was measured for $n = 9$ runs.

0.63 2.64 1.85 1.68 1.09 1.67 0.73 1.04 0.68

(a) Conduct a test of hypotheses with the intent of showing that the mean yield is less than 1.8. Take $\alpha = 0.05$ and assume a normal population.

(b) Based on your conclusion in part (a), what error could you have made? Explain in the context of the problem.

7.57 Refer to Exercise 7.14, where $n = 9$ measurements were made on a key performance indicator.

123 106 114 128 113 109 120 102 111

(a) Conduct a test of hypotheses with the intent of showing that the mean key performance indicator is different from 107. Take $\alpha = 0.05$ and assume a normal population.

(b) Based on your conclusion in part (a), what error could you have made? Explain in the context of the problem.

7.58 Refer to Exercise 7.22, where $n = 64$ cartridges for a copy machine produced a mean of 18,300 copies and a standard deviation of 2,800 copies.

(a) Conduct a test of hypotheses with the intent of showing that the mean number of copies is greater than 17,500 copies. Take $\alpha = 0.02$.

(b) Based on your conclusion in part (a), what error could you have made? Explain in the context of the problem.

7.59 Refer to Exercise 2.34, page 36, concerning material costs for rebuilding $n = 29$ traction motors. A computer calculation gives $\bar{x} = 1.4707$ and $s = 0.5235$ thousand dollars. At the 0.05 level of significance, conduct a test of hypotheses with the intent of showing the mean is greater than 1.3 thousand dollars.

7.60 In 64 randomly selected hours of production, the mean and the standard deviation of the number of acceptable pieces produced by a automatic stamping machine are $\bar{x} = 1{,}038$ and $s = 146$. At the 0.05 level of

significance, does this enable us to reject the null hypothesis $\mu = 1{,}000$ against the alternative hypothesis $\mu > 1{,}000$?

7.61 With reference to the thickness measurements in Exercise 2.41, test the null hypothesis that $\mu = 30.0$ versus a two-sided alternative. Take $\alpha = 0.05$.

7.62 A random sample of 6 steel beams has a mean compressive strength of 58,392 psi (pounds per square inch) with a standard deviation of 648 psi. Use this information and the level of significance $\alpha = 0.05$ to test whether the true average compressive strength of the steel from which this sample came is 58,000 psi. Assume normality.

7.63 A manufacturer claims that the average tar content of a certain kind of cigarette is $\mu = 14.0$. In an attempt to show that it differs from this value, five measurements are made of the tar content (mg per cigarette):

14.5 14.2 14.4 14.3 14.6

Show that the difference between the mean of this sample, $\bar{x} = 14.4$, and the average tar claimed by the manufacturer, $\mu = 14.0$, is significant at $\alpha = 0.05$. Assume normality.

7.64 Suppose that in the preceding exercise the first measurement is recorded incorrectly as 16.0 instead of 14.5. Show that, even though the mean of the sample increases to $\bar{x} = 14.7$, the null hypothesis H_0: $\mu = 14.0$ is not rejected at level $\alpha = 0.05$. Explain the apparent paradox that even though the difference between observed $\bar{x}$ and μ has increased, the null hypothesis is no longer rejected.

7.65 The statistical program *MINITAB* will calculate t tests. With the nanopillar height data in C1,

> **Dialog box:**
> **Calc > Stat > Basic Statistics > 1-Sample t**
> Type *C1* in **variables**. Choose **Test mean** and enter 300. Click **OK**.

produces the output

Test of mu = 300 vs not = 300

	N	Mean	StDev	SE Mean	T	P
C1	50	305.580	36.971	5.229	1.07	0.291

You must compare your preselected α with the printed P-value in order to obtain the conclusion of your test. To perform a Z test, you need to find the P-value corresponding to the value 1.07 for the test statistic.

Here, with the two-sided alternative, we cannot reject H_0: $\mu = 300$ nm unless we take the significance level almost as large as 0.30.

(a) Test H_0: $\mu = 295$ with $\alpha = 0.05$.

(b) Test H_0: $\mu = 16.0$ for the speed of light data in Exercise 2.66 with $\alpha = 0.05$.

7.7 The Relation between Tests and Confidence Intervals

We now describe an important connection between tests for two-sided alternatives and confidence intervals. This relation provides the reason that most statisticians prefer the information available in a confidence interval statement as opposed to the information that the null hypothesis $\mu = \mu_0$ was or was not rejected.

To develop the relation, we consider the $(1 - \alpha)100\%$ confidence interval for μ given on page 211:

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

This interval is closely connected to a level α test of H_0: $\mu = \mu_0$ versus the two-sided alternative H_1: $\mu \neq \mu_0$. In terms of the values of $\bar{x}$ and s, this test has critical region

$$\left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| = |t| \geq t_{\alpha/2}$$

The acceptance region of this test is obtained by reversing the inequality to obtain all the values of $\bar{x}$ and s that do not lead to the rejection of H_0: $\mu = \mu_0$.

$$\text{acceptance region: } \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| < t_{\alpha/2}$$

The acceptance region can also be expressed as

$$\text{acceptance region: } \bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu_0 < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where the limits of the interval are identical to the preceding confidence interval. That is, the null hypothesis will not be rejected at level α if μ_0 lies within the $(1-\alpha)100\%$ confidence interval for μ.

The $(1-\alpha)100\%$ confidence interval gives the interval of plausible values for μ, so if μ_0 is contained in this interval, then it cannot be ruled out (i.e., cannot be rejected). The set of plausible values for μ, as determined by the $(1-\alpha)100\%$ confidence interval, tells us at once about the outcome of all possible two-sided tests of hypothesis that specify a single value for μ.

EXAMPLE **Illustrating the relation between tests and confidence intervals**

Referring to the example on page 211, $n = 16$ measurements of weight loss yielded $\bar{x} = 3.42$ and $s = 0.68$. Since $t_{0.025} = 2.131$ with 15 degrees of freedom, the 95% confidence interval becomes

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}, \quad \text{or} \quad 3.06 < \mu < 3.78.$$

Use the relation between 95% confidence intervals and $\alpha = 0.05$ level tests to test the null hypothesis $\mu = 3.7$ versus the alternative hypothesis $\mu \neq 3.7$. Also test the null hypothesis $\mu = 3.0$ versus the alternative hypothesis $\mu \neq 3.0$.

Solution In view of the relation just established, a test of $\mu = 3.7$ versus $\mu \neq 3.7$ would not reject H_0: $\mu = 3.7$, at the 5% level, since $\mu = 3.7$ falls within the confidence interval.

On the other hand $\mu = 3.0$ does not fall within the 95% confidence interval and hence that null hypothesis would be rejected at level $\alpha = 0.05$. ■

The relation between tests and confidence intervals holds quite generally. Suppose, for any value θ_0 of a parameter θ, we have a level α test of the null hypothesis $\theta = \theta_0$ versus the alternative $\theta \neq \theta_0$. Collect all the values θ_0 that would not be rejected. These form a $(1-\alpha)100\%$ confidence interval for θ.

A confidence interval statement provides a more comprehensive inference than a statement concerning a two-sided test of a single null hypothesis. Consequently, we favor a confidence interval approach when one is available.

Exercises

7.66 Refer to the nonopillar height data on page 15. Using the 95% confidence interval, based on the t distribution, for the mean nanopillar height

N	Mean	StDev	SE Mean	95% CI
50	305.580	36.971	5.229	(295.073, 316.087)

(a) decide whether or not to reject $H_0: \mu = 320$ nm in favor of $H_1: \mu \neq 320$ at $\alpha = 0.05$;

(b) decide whether or not to reject $H_0: \mu = 310$ nm in favor of $H_1: \mu \neq 310$ at $\alpha = 0.05$.

(c) What is your decision in part (b) if $\alpha = 0.02$? Explain.

7.67 Repeat Exercise 7.66 but replace the t test with the large sample Z test.

7.68 Refer to the green gas data on page 220. Using the 95% confidence interval, based on the t distribution for the mean yield

N	Mean	StDev	SE Mean	95% CI
15	6.00933	1.07780	0.27829	(5.41247, 6.60620)

(a) decide whether or not to reject $H_0: \mu = 5.5$ gal in favor of $H_0: \mu \neq 5.5$ at $\alpha = 0.05$;

(b) decide whether or not to reject $H_0: \mu = 5.3$ gal in favor of $H_1: \mu \neq 5.3$ at $\alpha = 0.05$.

(c) Perform the t test for part (b) to verify your conclusion.

7.69 Refer to the labor time data in Exercise 7.3. Using the 90% confidence interval, based on the t distribution, for the mean labor time

N	Mean	StDev	SE Mean	90% CI
52	1.86462	1.24992	0.17333	(1.57423, 2.15500)

(a) decide whether or not to reject $H_0 : \mu = 1.6$ in favor of $H_1 : \mu \neq 1.6$ at $\alpha = 0.10$;

(b) decide whether or not to reject $H_0 : \mu = 2.2$ in favor of $H_1 : \mu \neq 2.2$ at $\alpha = 0.10$.

(c) What is your decision in part (a) if $\alpha = 0.05$? Explain.

7.70 Repeat Exercise 7.69 but replace the t test with the large sample Z test.

7.8 Power, Sample Size, and Operating Characteristic Curves*

So far we have not paid much attention to Type II errors. In the drying time example of Section 7.5 we calculated one probability of a Type II error. Since the choice of the alternative hypothesis $\mu = 21$ minutes in the drying time example was essentially arbitrary, it may be of interest to see how the testing procedure will perform when other values of μ prevail. The actual mean drying time depends on the chemical reactions governing the paint and the environmental conditions. Lacking a total scientific explanation, we do not know precisely what value to expect for this population mean. Consequently, we must investigate the probability of not rejecting (accepting) the null hypothesis under a range of possible values for μ. To this end, let

$$L(\mu) = \text{probability of accepting the null hypothesis when } \mu \text{ prevails}$$

Figure 7.11 presents the picture of a typical operating characteristic (OC) curve for the case where the alternative hypothesis is $\mu > \mu_0$. When the alternative hypothesis is $\mu < \mu_0$, the OC curve becomes the mirror image of that of Figure 7.11, reflected about the dashed vertical line through μ_0.

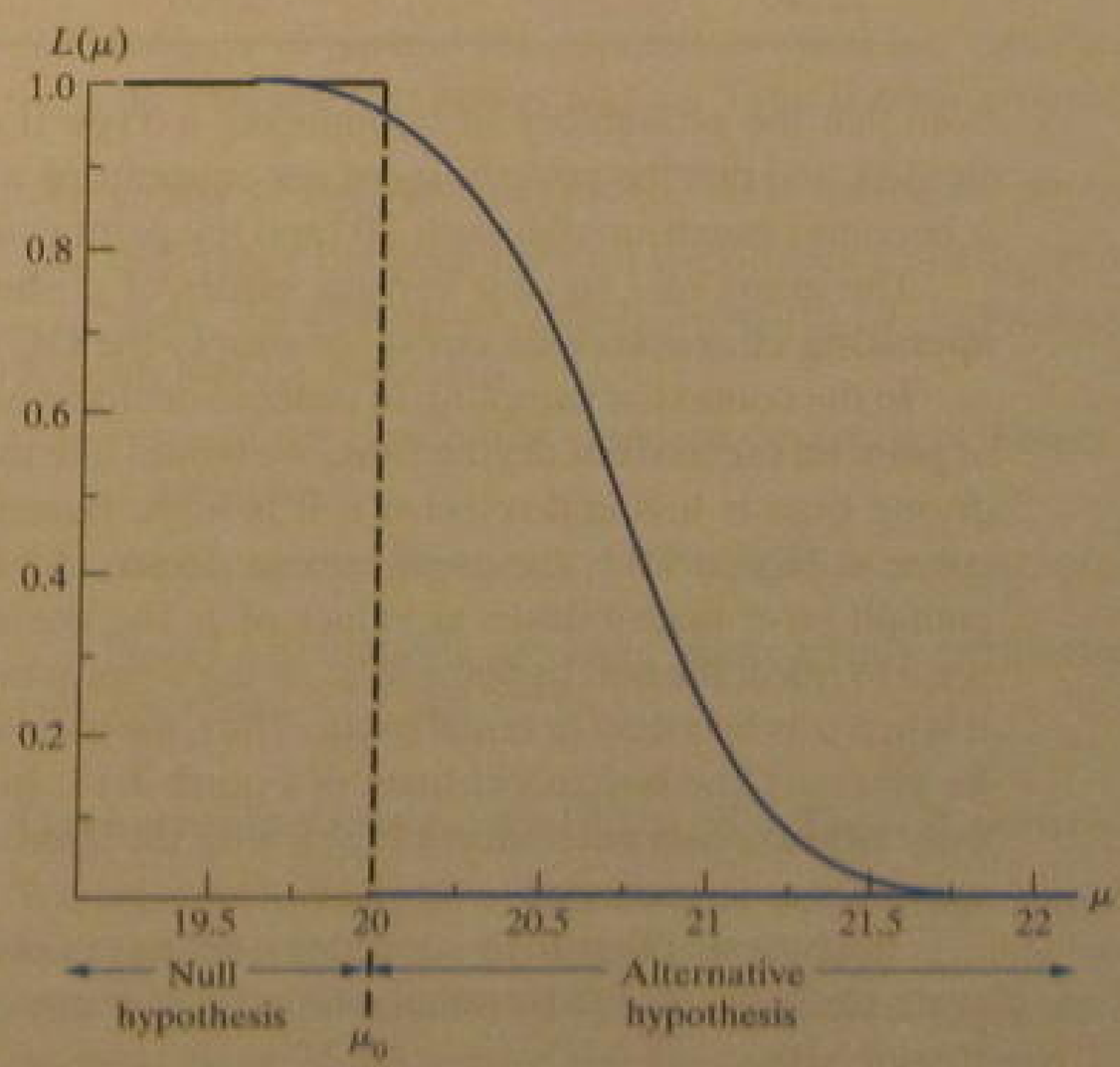


Figure 7.11
Operating characteristic curve

*This section can be omitted on first reading.

The function $L(\mu)$ completely characterizes the testing procedure whatever the value of population mean μ. If μ equals a value where the null hypothesis is true, then $1 - L(\mu)$ is the probability of the Type I error. When μ has a value where the alternative hypothesis is true, then $L(\mu)$ is the probability of a Type II error. That is, the function $L(\mu)$ carries complete information about the probabilities of both types of error.

To illustrate the calculation of $L(\mu)$, we continue with the drying time example on page 222, where we had $\mu_0 = 20, \sigma = 2.4, n = 36$, and the dividing line of the criterion was $\bar{x} = 20.75$. If the prevailing population mean is $\mu = 20.5$, then $Z = \sqrt{n}(\overline{X} - 20.5)/\sigma = 6(\overline{X} - 20.5)/2.4$ is a standard normal variable. We reason that $L(20.5)$ is the probability of observing

$$\overline{X} < 20.75, \quad \text{or} \quad \frac{6(\overline{X} - 20.5)}{2.4} < \frac{6(20.75 - 20.5)}{2.4} = 0.625$$

or $Z < 0.625$. Therefore, from Table 3, $L(20.5) = 0.73$. Continuing with other possible values for μ, we obtain the results shown in the following table and also in Figure 7.11 (see Exercise 7.75):

Values of μ	Probability of accepting null hypothesis
19.50	0.999
19.75	0.99
20.00	0.97
20.25	0.89
20.50	0.73
20.75	0.50
21.00	0.27
21.25	0.11
21.50	0.03
21.75	0.01
22.00	0.001

Note that the probability of committing a Type II error diminishes when μ is increased, and that the probability of not committing a Type I error approaches 1 when μ becomes much smaller than 20 (and the paint is even better than claimed).

The graph of $L(\mu)$ for various values of μ shown in Figure 7.11 is called the **operating characteristic curve**, or simply the **OC curve**, of the test.

In the context of sampling in order to decide whether or not to accept a shipment of paint on the basis of drying time, we would like to accept the shipment if the mean drying time is low and reject it if it is high. Based on the operating characteristic curve in Figure 7.11, the engineer can decide if the proposed procedure has small enough error probabilities at values of μ she deems important. Ideally we should want to reject the null hypothesis $\mu = \mu_0$ when actually μ exceeds μ_0, and to accept it when μ is less than or equal to μ_0. Thus, the ideal OC curve for our example would be given by the horizontal lines of Figure 7.11. In actual practice, OC curves can only approximate such ideal curves, with the approximation becoming better as the sample size is increased.

In contexts other than acceptance sampling plans, and for most statistical software packages, the performance of a test is expressed in terms of **power**:

$$\text{power} = \gamma(\mu) = P(\text{reject } H_0)$$

when μ is a value for the mean under the alternative hypothesis. Consequently, power and the operating characteristic are equivalent since $\gamma(\mu) = 1 - L(\mu)$.

For the one-sided alternative $H_0: \mu > \mu_0$, with σ known, the rejection region for a level α Z test is

$$\sqrt{n}\,\frac{(\overline{X}-\mu_0)}{\sigma} > z_\alpha \quad \text{or} \quad \overline{X} > \frac{\sigma}{\sqrt{n}}\,z_\alpha + \mu_0$$

When the mean has the particular value μ_1 greater than μ_0, $\sqrt{n}\,(\overline{X}-\mu_1)/\sigma$ has a standard normal distribution and H_0 is rejected when

$$\overline{X}-\mu_1 > \frac{\sigma}{\sqrt{n}}\,z_\alpha + \mu_0 - \mu_1$$

The power at μ_1 is

$$\begin{aligned}\gamma(\mu_1) &= P\left(\sqrt{n}\,\frac{(\overline{X}-\mu_1)}{\sigma} > z_\alpha + \sqrt{n}\,\frac{(\mu_0-\mu_1)}{\sigma}\right)\\ &= P\left(Z > z_\alpha + \sqrt{n}\,\frac{(\mu_0-\mu_1)}{\sigma}\right)\end{aligned}$$

when $H_1: \mu > \mu_0$ and $\mu_1 > \mu_0$.

Similarly,

$$\begin{aligned}\gamma(\mu_1) &= P\left(\sqrt{n}\,\frac{(\overline{X}-\mu_1)}{\sigma} < -z_\alpha + \sqrt{n}\,\frac{(\mu_0-\mu_1)}{\sigma}\right)\\ &= P\left(Z < -z_\alpha + \sqrt{n}\,\frac{(\mu_0-\mu_1)}{\sigma}\right)\end{aligned}$$

when $H_1: \mu < \mu_0$ and $\mu_1 < \mu_0$. Under a two-sided alternative,

$$\gamma(\mu_1) = P\left(Z < -z_{\alpha/2} + \sqrt{n}\,\frac{(\mu_0-\mu_1)}{\sigma}\right) + P\left(Z > z_{\alpha/2} + \sqrt{n}\,\frac{(\mu_0-\mu_1)}{\sigma}\right)$$

when $H_1: \mu \neq \mu_0$ and the population mean has value $\mu_1 \neq \mu_0$.

We observe that power and the Type II error probability β depend upon the

1. choice of significance level α. As α increases, power increases (β decreases).
2. difference between the null hypothesis value μ_0 and the particular value μ_1 under the alternative. Power increases (β decreases) as the difference between μ_0 and μ_1 increases.
3. value of the population standard deviation σ. Power decreases (β increases) as σ increases.
4. sample size n. Power increases (β decreases) as sample size increases.

Power calculations are most conveniently performed with statistical software (see Exercises 7.72 and 7.75).

EXAMPLE **Determining the probability of a Type II error—one-sided test**

We want to investigate a claim about the average sound intensity of certain vacuum cleaners. Suppose the sound is a random variable having a normal distribution with a standard deviation of 3.6 dB. Specifically, we shall want to test the null hypothesis $\mu = 75.20$ against the alternative hypothesis $\mu > 75.20$ on the basis of measurements of the sound intensity of $n = 15$ of these machines. If the probability of a Type I error is to be $\alpha = 0.05$, what is the probability of a Type II error for $\mu = 77.00$ dB?

Solution The test is one-sided, $\alpha = 0.05$, $z_{0.05} = 1.645$, and

$$z_\alpha + \sqrt{n}\,\frac{(\mu_0 - \mu_1)}{\sigma} = 1.645 + \sqrt{15}\,\frac{(75.20 - 77.0)}{3.6} = -0.291$$

so the power is $\gamma(77.0) = P(Z > -0.219) = 0.614$. The Type II error probability $\beta = 1 - 0.614 = 0.386$. ■

EXAMPLE **Determining the probability of a Type II error—two-sided test**

Suppose that the length of certain machine parts may be looked upon as a random variable having a normal distribution with a mean of 2.000 cm and a standard deviation of 0.050 cm. Specifically, we shall want to test the null hypothesis $\mu = 2.000$ against the alternative hypothesis $\mu \neq 2.000$ on the basis of the mean of a random sample of size $n = 30$. If the probability of a Type I error is to be $\alpha = 0.05$, what is the probability of a Type II error for $\mu = 2.010$?

Solution The test is two-tailed, $\alpha = 0.05$, $\alpha/2 = 0.025$, $z_{0.025} = 1.96$, and

$$-z_{\alpha/2} + \sqrt{n}\,\frac{(\mu_0 - \mu_1)}{\sigma} = -1.96 + \sqrt{30}\,\frac{(2.000 - 2.010)}{0.050} = -3.055$$

$$z_{\alpha/2} + \sqrt{n}\,\frac{(\mu_0 - \mu_1)}{\sigma} = 1.96 + \sqrt{30}\,\frac{(2.000 - 2.010)}{0.050} = 0.865$$

The power is

$$\gamma(2.010) = P(Z < -3.055) + P(Z > .865) = 0.001 + 0.194 = 0.196$$

The Type II error probability $\beta = 1 - 0.196 = 0.804$. ■

When the alternative is one-sided, we can obtain an equation for the sample size required to give a specified power $\gamma = 1 - \beta$ at some value μ_1. For, $H_1: \mu > \mu_0$ we require n so that

$$\gamma = 1 - \beta = P\left(Z > z_\alpha + \sqrt{n}\,\frac{(\mu_0 - \mu_1)}{\sigma}\right)$$

but then, by the definition z_β, we must have

$$-z_\beta = z_\alpha + \sqrt{n}\,\frac{(\mu_0 - \mu_1)}{\sigma}$$

Solving for n, we obtain the required sample size. The same expression for required sample size holds when $H_1: \mu < \mu_0$.

Required sample size

When conducting a one-sided large sample Z test, the required sample size n must be at least as large as

$$n = \left(\sigma\,\frac{(z_\beta + z_\alpha)}{(\mu_0 - \mu_1)}\right)^2$$

For two-sided tests, the calculation of sample size is best relegated to computer software (see Exercise 7.72).

EXAMPLE **Determining a sample size—one-sided test**

With reference to the drying time example on page 222, where we had $\mu_0 = 20$, $\sigma = 2.4$, and $\alpha = 0.05$, how large a sample would we need so that $\beta = 0.10$ for $\mu = 21$?

Solution The test is one-sided, $\mu_0 = 20$, $\mu_1 = 21$, and $\sigma = 2.4$. Also, $\alpha = 0.05$, $z_{0.05} = 1.645$, $\beta = 0.10$, $z_{0.10} = 1.28$. The required sample size is no smaller than

$$n = \left(\sigma \frac{(z_\beta + z_\alpha)}{(\mu_0 - \mu_1)}\right)^2 = \left(2.4 \frac{(1.28 + 1.645)}{(20 - 21)}\right)^2 = 49.2$$

so $n = 50$ is the required sample size. ■

Exercises

7.71 Refer to the example concerning average sound intensity on page 239. Calculate the the power at $\mu_1 = 77$ when

(a) the level is changed to $\alpha = 0.03$.

(b) $\alpha = 0.05$ but the alternative is changed to the two-sided $H_1 : \mu \neq 75.2$.

7.72 *MINITAB* **calculation of power**

These calculations pertain to normal populations with known variance and provide an accurate approximation in the large sample case where σ is unknown. To calculate the power of the Z test at μ_1, you need to enter the

$$\text{difference} = \mu_1 - \mu_0.$$

Although $n = 15$ is not large, we illustrate with reference to the example concerning average sound intensity on page 239, where $\alpha = 0.05$, $\sigma = 3.6$, and $H_1 : \mu > 75.2$. We are given $\mu_1 = 77$, so the difference $= \mu_1 - \mu_0 = 77 - 75.2 = 1.80$.

Dialog box:

Stat > **Power and Sample Size** > **1-Sample Z**. Type 15 in **Sample sizes,** 1.8 in **differences** and 3.6 in **standard deviation.**
Click **Options** and choose *greater than*. Type 0.05 in **significance level.**
Click **OK.** Click **OK.**

Notice that you have a choice of an alternative that is less than, not equal to or greater than.

Output: (partial)

Difference	Sample Size	Power
1.8	15	0. 0.614718

Referring to the example of machine parts on page 240.

(a) Calculate the power at $\mu_1 = 2.020$.

(b) Repeat part (a) but take $\alpha = 0.03$.

7.73 Use computer software to repeat Exercise 7.71.

7.74 *MINITAB* **calculation of sample size**

Refer to Exercise 7.72, but this time leave **Sample size** blank but

Type *0.90* in **power**

to obtain the partial output concerning sample size

Difference	Sample Size	Target Power	Actual Power
1.8	35	0.9	0.905440

Refer to the example concerning sound intensity on page 239. Find the required sample size if power must be at least 0.96 at $\mu_1 = 77$.

7.75 *MINITAB* **calculation of power or OC curve**
Refer to the steps in Exercise 7.72, but enter a range of values for the difference. Here 0:3/.1 goes in steps from 0 to 3 in steps of .1 for the drying time example.

> **Stat > Power and sample size > 1-Sample Z.** Type 36 in **Sample sizes,** 0 : 3/.1 in **differences** and 2.4 in **Standard deviation.** Click **Options** and choose **greater than.** Type 0.05 in **level of significance.** Type *C*1 in **Store differences in.** Type *C*2 in **Store power values in.**
> Click **OK**. Click **OK**.

Then calculate the values for μ = difference + μ_0 = C1 + 20 and store in C3. Plot C2 versus C3 to produce the power curve. The operating characteristic is obtained by calculating 1 − power = 1 − C2 and storing in C4. Plot C4 versus C2.

With reference to the drying time example on page 222, use computer software to obtain

(a) the power curve; (b) the operating characteristic curve.

Do's and Don'ts

Do's

1. Calculate the estimated standard error $s/\sqrt{n}$ to accompany the point estimate $\bar{x}$ of a population mean.
2. Whatever the population, when the sample size is large, calculate the $100(1-\alpha)\%$ confidence interval for the mean
$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$
3. When the population is normal, calculate the $100(1-\alpha)\%$ confidence interval for the mean
$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$
where $t_{\alpha/2}$ is obtained from the t distribution with $n-1$ degrees of freedom.
4. Understand the interpretation of a $100(1-\alpha)\%$ confidence interval. When the population is normal, before the data are collected,
$$\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$
is a random interval that will cover the fixed unknown mean with probability $1-\alpha$. In many repeated applications of this method, about proportion $1-\alpha$ of the times the interval will cover the respective population mean.
5. When conducting a test of hypothesis, formulate the assertion that the experiment seeks to confirm as the alternative hypothesis.
6. When the sample size is large, base a test of the null hypothesis H_0: $\mu = \mu_0$ on the test statistic
$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

which has, approximately, a standard normal distribution. When the population is normal, the same statistic has a t distribution with $n - 1$ degrees of freedom.

7. Understand the interpretation of a level α test. If the null hypothesis is true, before the data are collected, the probability is α that the experiment will produce observations that lead to the rejection of the null hypothesis. Consequently, after many independent experiments, the proportion that lead to rejection of the null hypothesis will be nearly α.

Don'ts

1. Don't routinely apply the statistical procedures above if the sample is not random but collected from convenient units or the data show a trend in time.

Review Exercises

7.76 Specify the null hypothesis and the alternative hypothesis in each of the following cases.

(a) An engineer hopes to establish that an additive will increase the viscosity of an oil.

(b) An electrical engineer hopes to establish that a modified circuit board will give a computer a higher average operating speed.

7.77 With reference to the example on page 19, find a 95% confidence interval for the mean strength of the aluminum alloy.

7.78 While performing a certain task under simulated weightlessness, the pulse rate of 32 astronaut trainees increased on the average by 26.4 beats per minute with a standard deviation of 4.28 beats per minute. What can one assert with 95% confidence about the maximum error if $\bar{x} = 26.4$ is used as a point estimate of the true average increase in the pulse rate of astronaut trainees performing the given task?

7.79 With reference to the preceding exercise, construct a 95% confidence interval for the true average increase in the pulse rate of astronaut trainees performing the given task.

7.80 It is desired to estimate the mean number of days of continuous use until a certain kind of computer will first require repairs. If it can be assumed that $\sigma = 48$ hours, how large a sample is needed so that one will be able to assert with 90% confidence that the sample mean is off by at most 10 hours?

7.81 A sample of 12 camshafts intended for use in gasoline engines has an average eccentricity of 1.02 and a standard deviation of 0.044 inch. Assuming the data may be treated as a random sample from a normal population, determine a 95% confidence interval for the actual mean eccentricity of the camshaft.

7.82 In order to test the durability of a new paint, a highway department had test strips painted across heavily traveled roads in 15 different locations. If on the average the test strips disappeared after they had been crossed by 146,692 cars with standard deviation of 14,380 cars, construct a 99% confidence interval for the true average number of cars it takes to wear off the paint. Assume a normal population.

7.83 Referring to Exercise 7.82 and using 14,380 as an estimate of σ, find the sample size that would have been needed to be able to assert with 95% confidence that the sample mean is off by at most 10,000. [*Hint*: First estimate n_1 by using $z = 1.96$, then use $t_{0.025}$ for $n_1 - 1$ degrees of freedom to obtain a second estimate n_2, and repeat this procedure until the last two values of n thus obtained are equal.]

7.84 A laboratory technician is timed 20 times in the performance of a task, getting $\bar{x} = 7.9$ and $s = 1.2$ minutes. If the probability of a Type I error is to be at most 0.05, does this constitute evidence against the null hypothesis that the average time is less than or equal to 7.5 minutes?

7.85 Suppose that in the drying time example on page 222, n is changed from 36 to 50 while the other quantities remain $\mu_0 = 20$, $\sigma = 2.4$, and $\alpha = 0.03$. Find

(a) the new dividing line of the test criterion;

(b) the probability of Type II errors for the same values of μ as shown in the table on page 238.

7.86 In an air-pollution study, ozone measurements were taken in a large California city at 5.00 P.M. The eight readings (in parts per million) were

7.9 11.3 6.9 12.7 13.2 8.8 9.3 10.6

Assuming the population sampled is normal, construct a 95% confidence interval for the corresponding true mean.

7.87 An industrial engineer concerned with service at a large medical clinic recorded the duration of time from the time a patient called until a doctor or nurse returned the call. A sample of size 180 calls had a mean of 1.65 hours and a standard deviation of 0.82.

(a) Obtain a 95% confidence interval for the population mean of time to return a call.

(b) Does μ lie in your interval obtained in part (a)? Explain.

(c) In a long series or repeated experiments, with new random samples collected for each experiment, what proportion of the resulting confidence intervals will contain the true population mean? Explain your reasoning.

7.88 Refer to Exercise 7.87.

(a) Perform a test with the intention of establishing that the mean time to return a call is greater than 1.5 hours. Use $\alpha = 0.05$.

(b) In light of your conclusion in part (a), what error could you have made? Explain in the context of this problem.

(c) In a long series of repeated experiments, with new random samples collected for each experiment, what proportion of the resulting tests would reject the null hypothesis if it prevailed? Explain your reasoning.

7.89 The compressive strength of parts made from a composite material are known to be nearly normally distributed. A scientist, using the testing device for the first time, obtains the tensile strength (psi) of 20 specimens

95	102	105	107	109	110	111	112	114	115
134	135	136	138	139	141	142	144	150	155

shown in Figure 7.12. Should the scientist report the 95% confidence interval based on the t-distribution? Explain your reasoning.

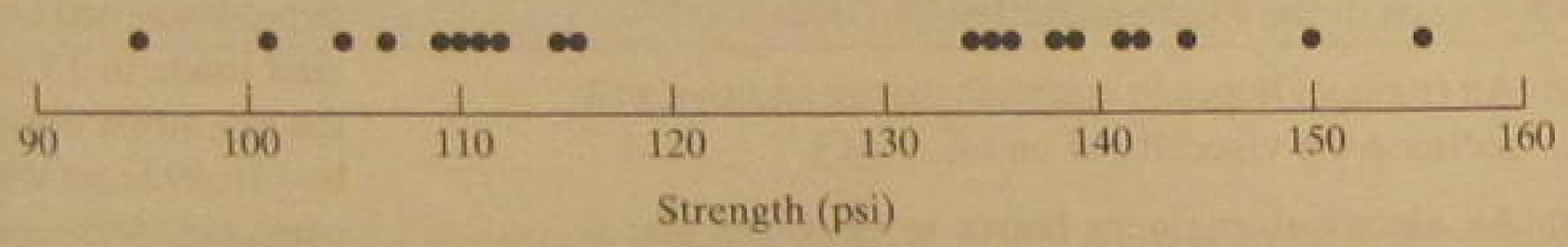


Figure 7.12
Dot diagram of tensile strength

Key Terms

Alternative hypothesis 225
Classical theory of testing hypotheses 226
Composite hypothesis 227
Confidence 208
Confidence interval 210
Confidence limits 210
Critical regions 229
Critical values 230
Degree of confidence 210
Estimated standard error 204
Hat notation 205
Interval estimate 209
Large sample Z test 232
Level of significance 225
Likelihood function 217
Maximum likelihood estimator 217
Neyman-Pearson theory 226
Null hypothesis 227
One sample t test 233
One-sided alternative 225
One-sided criterion 226
One-sided test 226
One-tailed test 226
Operating characteristic (OC) curve 238
Point estimation 204
Power 238
P value 231
Simple hypothesis 227
Tail probability 230
Two-sided alternative 225
Two-sided criterion 226
Two-sided test 226
Two-tailed test 226
Type I error 223
Type II error 223
Unbiased estimator 205

CHAPTER

8

Comparing Two Treatments

Advances occur in engineering when new ideas lead to better equipment, new materials, or revision of existing production processes. Any new procedure or device must be compared with the existing one and the amount of improvement assessed. As described in Section 8.1, statistical methods of comparison begin by deciding on one of two schemes for collecting data. Sections 8.2 and 8.3 concern the independent samples design and Section 8.4 the matched pairs sample design. These two designs are compared and contrasted in Section 8.5.

CHAPTER OUTLINE

8.1 Experimental Designs for Comparing Two Treatments 245

8.2 Comparisons—Two Independent Large Samples 246

8.3 Comparisons—Two Independent Small Samples 251

8.4 Matched Pairs Comparisons 259

8.5 Design Issues—Randomization and Pairing 263

Review Exercises 266

Key Terms 266

8.1 Experimental Designs for Comparing Two Treatments

Progress in science and engineering begins when new devices or materials are invented or when existing processes are revised. Advances occur whenever the new technique is shown to be better than the old. We perform experiments, collect data on performance, and then use statistical methods to make comparisons between the new and the old techniques.

It is common to use the statistical term **treatment** to refer to the procedures, machines, or processes that are being compared. The basic unit that is exposed to one treatment or the other is called the **experimental unit**, and the observed characteristic that forms the basis of the comparison is called the **response**.

EXAMPLE **Randomly assigning treatment to units**

It is an extremely expensive event for a cell phone company when one of its relay towers breaks down. Because of a shortage of experts, sometimes a novice must be sent to fix the problem. A cell phone company wants to conduct an experiment to compare the average time for an expert to fix a problem and the average time for a novice to fix the problem.

The next 15 breakdowns can be used in the experiment. Describe the scheme for conducting a comparative experiment.

Solution A breakdown is the unit, and expert and novice will be Treatment 1 and Treatment 2, respectively. The response is the time for the tower to be fixed. One person will be assigned to each breakdown. The times for the group of breakdowns where novices are assigned should be independent of the times for the group of breakdowns where experts are assigned. This is the *independent samples design*.

Before we know anything about the breakdown, its location, or its possible severity, seven of the breakdowns should be chosen to receive Treatment 1, expert. This selection should be made using random numbers. Numbering the order of breakdowns from 1 to 15, seven different random numbers should be selected within this range. Randomization, without restriction, gives rise to the alternative name *completely randomized design*. ■

EXAMPLE **Pairing to eliminate a known source of variation**

A civil engineer needs to compare the durability of two different paints for marking lanes on a divided highway. One year after applying the paint, she will return and assign a number to the current quality of the marking. Ten widely separated 1-mile sections of divided highway are available. From experience, she expects that the sections would carry substantially different traffic volumes. Describe an experiment for making the comparison.

Solution The two paints are the treatments which we call Treatment 1 and Treatment 2. The 1-mile sections, in either direction, will carry approximately the same traffic volume for the whole mile. A 1-mile section, with traffic going in one direction, is a unit. The response is quality of marking after one year. The civil engineer expects traffic volume to heavily influence the response. The traffic volume could be eliminated from the comparison if we pair the two 1-mile sections with traffic going in opposite directions. Then a comparison will only be made within the pair. This is called a *matched pairs design*.

Still the road for one side may be subject to more shade, higher temperature, or other conditions different from the other. For each 1-mile section, the engineer should flip a coin. If heads, the north or west direction receives Treatment 1. Randomization helps prevent these other uncontrolled variables from influencing the response in a systematic manner.

Note that by pairing like experimental units, we have eliminated traffic volume as an influencing variable. ■

The term **experimental design** refers to the manner in which units are chosen and assigned to receive treatments. As introduced in the last two examples, there are two basic designs for comparing two treatments:

1. **Independent samples (complete randomization)**
2. **Matched pairs sample (randomization within each matched pair)**

We investigate independent samples designs in the next section and matched pairs sample in the following section.

8.2 Comparisons—Two Independent Large Samples

In this section, we consider the independent samples design when both sample sizes are large. To state the assumptions, we use X and Y and the subscript 1 or 2 to distinguish the two populations.

Assumptions—Large Samples

1. $X_1, X_2, \ldots, X_{n_1}$ is a random sample of size n_1 from population 1 which has mean $= \mu_1$ and variance $= \sigma_1^2$.
2. $Y_1, Y_2, \ldots, Y_{n_2}$ is a random sample of size n_2 from population 2 which has mean $= \mu_2$ and variance $= \sigma_2^2$.
3. The two samples $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ are independent.

Inferences will be made about the difference in means $\mu_1 - \mu_2 = \delta$. Since, by Theorem 6.1,

$$E(\overline{X}) = \mu_1 \qquad Var(\overline{X}) = \frac{\sigma_1^2}{n_1} \qquad E(\overline{Y}) = \mu_2 \qquad Var(\overline{Y}) = \frac{\sigma_2^2}{n_2}$$

the mean of $\overline{X} - \overline{Y}$ is $E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2$ and, by independence (see the example on page 154),

$$Var(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

When the sample sizes n_1 and n_2 are large, the central limit theorem implies that both $\overline{X}$ and $\overline{Y}$ are approximately normal. Because they are independent, their difference is also approximately normal and the **two sample Z statistic**

$$Z = \frac{\overline{X} - \overline{Y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is approximately standard normal. Because the sample sizes n_1 and n_2 are large—namely, both are greater than or equal to 30—the normal approximation remains valid when the sample variances replace the population variances.

Statistic for large samples inference concerning difference between two means

When the sample sizes n_1 and n_2 are large—namely, $n_1, n_2 \geq 30$

$$Z = \frac{\overline{X} - \overline{Y} - \delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

is approximately standard normal.

Large Samples Confidence Intervals

Large samples confidence intervals, for the difference of means $\delta = \mu_1 - \mu_2$, are determined from the standard normal probability

$$1 - \alpha = P\left(-z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - \delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} < z_{\alpha/2}\right)$$

$$= P\left(-z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \overline{X} - \overline{Y} - \delta < z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$$

or

$$1 - \alpha = P\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \delta < \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$$

This last statement asserts that, before we obtain the data, the probability is $1 - \alpha$ that the random interval will cover the true unknown difference in the means $\delta = \mu_1 - \mu_2$.

Large samples confidence interval for $\mu_1 - \mu_2$

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This confidence interval can also be obtained from the acceptance regions for the two-sided test on page 249.

EXAMPLE **Large samples confidence interval for difference of means**

As a baseline for a study on the effects of changing electrical pricing for electricity during peak hours, July usage during peak hours was obtained for $n_1 = 45$ homes with air-conditioning and $n_2 = 55$ homes without.[1] The July on-peak usage (kWh) is summarized as

	Sample		
Population	Size	Mean	Variance
With	45	204.4	13,825.3
Without	55	130.0	8,632.0

Obtain a 95% confidence interval for $\delta = \mu_1 - \mu_2$.

Solution For a 95% confidence interval, $\alpha = 0.05$ and $z_{0.025} = 1.96$. We are given $n_1 = 45$, $n_2 = 55$, $\overline{x} = 204.4$, $s_1^2 = 13{,}825.3$, $\overline{y} = 130.0$, and $s_2^2 = 8{,}632.0$. Then the limits of the confidence interval are

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 204.4 - 130.0 \pm 1.96\sqrt{\frac{13{,}825.3}{45} + \frac{8{,}632.0}{55}}$$

$$= 74.4 \pm 1.96\sqrt{464.17}$$

so the 95% confidence interval is $(32.17, 116.63)$. The mean on-peak usage for homes with air-conditioning is higher than for homes without, from 32.17 to 116.63 kWh per month.

The confidence interval not only reveals that the two population means are statistically different, because the confidence interval does not cover 0, it also quantifies the amount of difference. ■

Large Samples Tests for Differences of Means

There are many statistical problems in which we are faced with decisions about the relative size of the means of two populations. For example, if two methods of welding are being considered, we may take samples and decide which is better by comparing their mean strengths.

Formulating the problem more generally, we shall consider two populations having the means μ_1 and μ_2 and the variances of σ_1^2 and σ_2^2. We want to test the null hypothesis

$$H_0\colon \mu_1 - \mu_2 = \delta_0$$

where δ_0 is a specified constant, on the basis of independent random samples of size n_1 and n_2. Analogous to the tests concerning one mean, we shall consider tests of this null hypothesis against each of the alternatives $\mu_1 - \mu_2 < \delta_0$, $\mu_1 - \mu_2 > \delta_0$, and $\mu_1 - \mu_2 \neq \delta_0$. The test itself will depend on the distance, measured in estimated standard deviation units, from the difference in sample means $\overline{X} - \overline{Y}$ to the

[1] Richard Johnson and Dean Wichern (2007), *Applied Multivariate Statistical Analysis*, 6th ed., page 289, Prentice Hall: Upper Saddle River, NJ.

hypothesized value, δ_0. When the sample sizes are large—namely, n_1 and n_2 are both greater than or equal to 30—we obtain this test statistic by specifying the null value δ_0 for the difference of means in the random Z defined above.

Test statistic for large samples concerning difference between two means

When $n_1, n_2 \geq 30$, test $H_0: \mu_1 - \mu_2 = \delta_0$ using

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

which has, approximately, a standard normal distribution.

Analogous to the table on page 232, the critical regions for testing the null hypothesis $\mu_1 - \mu_2 = \delta_0$, using the **two sample Z test** are;

Critical Regions for Testing $\mu_1 - \mu_2 = \delta_0$ (normal populations and σ_1 and σ_2 known, or large samples $n_1, n_2 \geq 30$)

Alternative hypothesis	Reject null hypothesis if:
$\mu_1 - \mu_2 < \delta_0$	$Z < -z_\alpha$
$\mu_1 - \mu_2 > \delta_0$	$Z > z_\alpha$
$\mu_1 - \mu_2 \neq \delta_0$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

Although δ_0 can be any constant, it is worth noting that in the great majority of problems its value is zero and we test the null hypothesis of no difference, namely, the null hypothesis $\mu_1 = \mu_2$.

EXAMPLE **Testing a difference in means with two large samples**

To test the claim that the resistance of electric wire can be reduced by more than 0.050 ohm by alloying, 32 values obtained for standard wire yielded $\overline{x} = 0.136$ ohm and $s_1 = 0.004$ ohm, and 32 values obtained for alloyed wire yielded $\overline{y} = 0.083$ ohm and $s_2 = 0.005$ ohm. At the 0.05 level of significance, does this support the claim?

Solution

1. *Null hypothesis*: $\mu_1 - \mu_2 = 0.050$
   Alternative hypothesis: $\mu_1 - \mu_2 > 0.050$
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $Z > 1.645$, where Z is given by the large samples formula above.
4. *Calculations*:

$$z = \frac{0.136 - 0.083 - 0.050}{\sqrt{\dfrac{(0.004)^2}{32} + \dfrac{(0.005)^2}{32}}} = 2.65$$

5. *Decision*: Since $z = 2.65$ exceeds 1.645, the null hypothesis must be rejected; that is, the data substantiate the claim. From Table 3, the P-value is 0.004 (see Figure 8.1), so the evidence for alloying is very strong. Only 4 in 1,000 times would Z be at least 2.65, if the mean difference was 0.05. ■

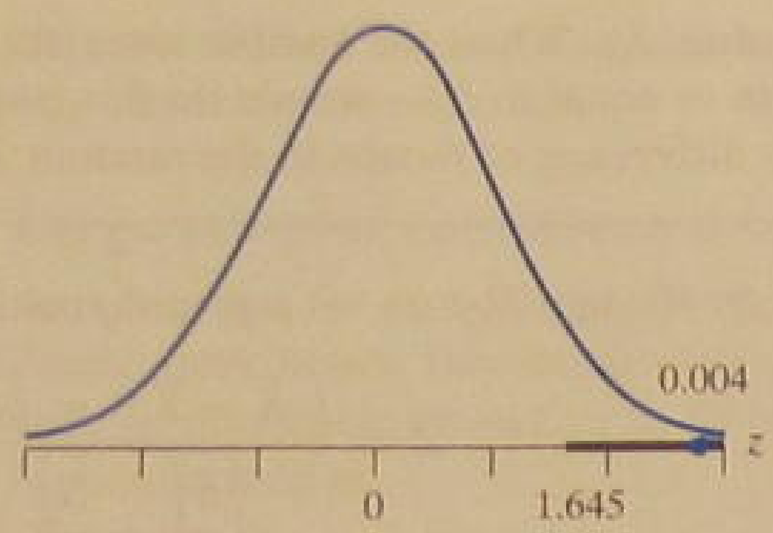


Figure 8.1
Large samples P-value = 0.004 for $Z = 2.65$

EXAMPLE **Large samples test of the null hypothesis of the equality of two means**

A company claims that its light bulbs are superior to those of its main competitor. If a study showed that a sample of $n_1 = 40$ of its bulbs has a mean lifetime of 1647 hours of continuous use with a standard deviation of 27 hours, while a sample of $n_2 = 40$ bulbs made by its main competitor had a mean lifetime of 1638 hours of continuous use with a standard deviation of 31 hours, does this substantiate the claim at the 0.05 level of significance?

Solution

1. *Null hypothesis*: $\mu_1 - \mu_2 = 0$
   Alternative hypothesis: $\mu_1 - \mu_2 > 0$
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $Z > 1.645$, where Z is given by the large samples formula on page 249.
4. *Calculations*:

$$z = \frac{1647 - 1638}{\sqrt{\frac{27^2}{40} + \frac{31^2}{40}}} = 1.38$$

5. *Decision*: Since $z = 1.38$ does not exceed 1.645, the null hypothesis cannot be rejected; that is, the observed difference between the two sample means is not significant.

   The P-value = 0.0838 (see Figure 8.2), so the evidence against equal means is not very strong. ■

To judge the strength of support for the null hypothesis when it is not rejected, we consider Type II errors, for which the probabilities depend on the actual alternative differences $\delta' = \mu_1 - \mu_2$. Fortunately, these can be determined using the single sample results (as long as we are sampling from normal populations with known standard deviations or both samples are large).

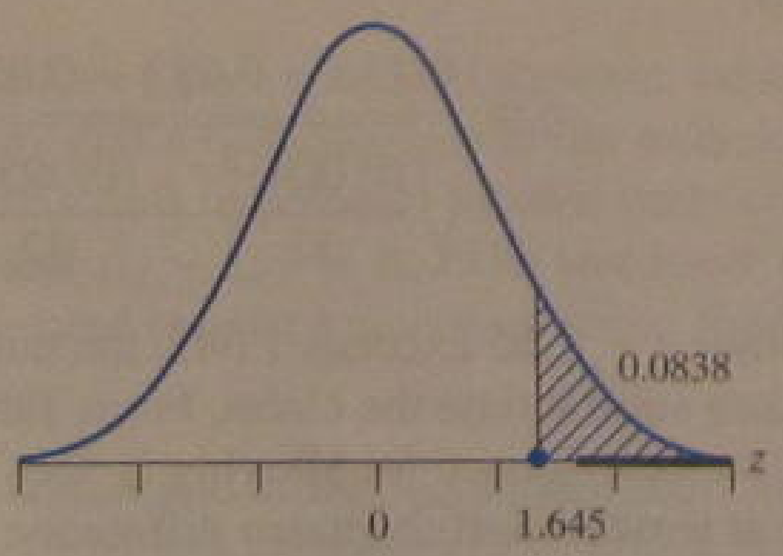


Figure 8.2
Large samples test: P-value = 0.0838 for $Z = 1.38$

The calculation of Type II error is based on the results for a single sample on page 239 and a fictitious single variance σ^2 and sample size n, where

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 \qquad n = \frac{\sigma_1^2 + \sigma_2^2}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Also, replace $\mu_0 - \mu_1$ by $\delta_0 - \delta'$.

The preferred method for calculation of Type II error is to use a computer software (see Exercise 7.72).

EXAMPLE **Approximating the probability of Type II error**

With reference to the preceding example, what is the probability of a Type II error for $\delta' = 16$ hours?

Solution Refer to the formula on page 239. Since $\delta_0 = 0$, $\sigma^2 = 27^2 + 31^2 = 1{,}690$, and $n = 40$

$$z_\alpha + \sqrt{n}\,\frac{(\delta_0 - \delta')}{\sigma} = 1.645 + \sqrt{40}\,\frac{(0-16)}{\sqrt{1{,}690}} = 1.645 - 2.642 = -0.817$$

so the power is $\gamma(16) = P(Z > -0.817) = 0.793$. The Type II error probability $\beta = 1 - 0.793 = 0.207$. ■

8.3 Comparisons—Two Independent Small Samples

When n_1 and n_2 are both small and the population variances are unknown, we must impose additional assumptions, which we label 4 or 5 to emphasize that the three original assumptions still prevail.

Additional Assumptions for Small Samples

4. Both populations are normal.
5. The two standard deviations have a common value $\sigma_1 = \sigma_2 = \sigma$.

Because the populations are normal, $\overline{X}$ and $\overline{Y}$ are normal and, because they are independent, their difference is also normal (see page 162). Recall, $E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 = \delta$. Under the assumption of common standard deviation σ, the expression above for the variance of $\overline{X} - \overline{Y}$ becomes

$$Var(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

and the standardized version of $\overline{X} - \overline{Y}$

$$Z = \frac{\overline{X} - \overline{Y} - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a normal distribution.

The unknown σ must be estimated. Reasoning that each squared deviation $(X_i - \overline{X})^2$ is an estimate of σ^2 and so is each $(Y_i - \overline{Y})^2$ from the second sample,

we estimate σ^2 by **pooling** the sums of squared deviations from the respective sample means. That is, we estimate σ^2 by the **pooled estimator**

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where $\sum(X_i - \overline{X})^2$ is the sum of the squared deviations from the mean for the first sample, while $\sum(Y_i - \overline{Y})^2$ is the sum of the squared deviations from the mean for the second sample. We divide by $n_1 + n_2 - 2$, since there are $n_1 - 1$ independent deviations from the mean in the first sample, $n_2 - 1$ in the second, and we thus have $n_1 + n_2 - 2$ independent deviations from their mean to estimate the population variance.

More specifically, from the single sample results we know that both S_1^2 and S_2^2 are estimates of σ^2 and that

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \quad \text{has a chi square distribution with } n_1 - 1 \text{ degrees of freedom}$$

$$\frac{(n_2 - 1)S_2^2}{\sigma^2} \quad \text{has a chi square distribution with } n_2 - 1 \text{ degrees of freedom}$$

and these two random quantities are independent since the samples on which they are based are independent. By either the result on page 192 or the example on page 196, the sum of the two chi square variables has a chi square distribution with degrees of freedom equal to the sum of the two degrees of freedom $n_1 + n_2 - 2$. Further,

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}$$

so

$$\frac{\dfrac{(n_1 - 1)S_1^2}{\sigma^2} + \dfrac{(n_2 - 1)S_2^2}{\sigma^2}}{n_1 + n_2 - 2} = \frac{\dfrac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}}{n_1 + n_2 - 2} = \frac{S_p^2}{\sigma^2}$$

and we conclude that

$$\frac{S_p}{\sigma} = \sqrt{\frac{\text{chi square variable}}{\text{degrees of freedom}}}$$

and this can be shown to be independent of the standard normal based on $\overline{X}_1 - \overline{X}_2$. Using the representation of t on page 192 as a standard normal over the square root of a chi square divided by its degrees of freedom, we obtain

Statistic for small sample test concerning difference between two means

$$t = \frac{\overline{X} - \overline{Y} - \delta}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \quad \text{where } S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

has a t distribution with $n_1 + n_2 - 2$ degrees of freedom (d.f.).

Note that by substituting S_p for σ in the expression for Z on page 251, we arrive at the same statistic. With small sample sizes, the distribution is not standard normal but a t.

The criteria for the **two sample t test** based on this statistic are like those for Z for testing the null hypothesis: H_0: $\mu_1 - \mu_2 = \delta_0$.

Critical Regions for Testing $\mu_1 - \mu_2 = \delta_0$ (normal populations with $\sigma_1 = \sigma_2$) two sample t test	
Alternative hypothesis	**Reject null hypothesis if:**
$\mu_1 - \mu_2 < \delta_0$	$t < -t_\alpha$
$\mu_1 - \mu_2 > \delta_0$	$t > t_\alpha$
$\mu_1 - \mu_2 \neq \delta_0$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

In the application of this test, n_1 and n_2 may be small, yet $n_1 + n_2 - 2$ may be 30 or more; in that case we use the normal critical value (also bottom line of Table 4.)

EXAMPLE **A two sample t test**

The following random samples are measurements of the heat-producing capacity (in millions of calories per ton) of specimens of coal from two mines:

Mine 1:	8,260	8,130	8,350	8,070	8,340	
Mine 2:	7,950	7,890	7,900	8,140	7,920	7,840

Use the 0.01 level of significance to test whether the difference between the means of these two samples is significant.

Solution

1. *Null hypothesis*: $\mu_1 - \mu_2 = 0$
   Alternative hypothesis: $\mu_1 - \mu_2 \neq 0$
2. *Level of significance*: $\alpha = 0.01$
3. *Criterion*: Reject the null hypothesis if $t < -3.250$ or $t > 3.250$, where 3.250 is the value of $t_{0.005}$ for $5 + 6 - 2 = 9$ degrees of freedom and t is given by the formula on page 252.
4. *Calculations*: The means and the variances of the two samples are

$$\bar{x} = 8{,}230, \quad \bar{y} = 7{,}940, \quad s_1^2 = \frac{63{,}000}{4} = 15{,}750, \quad \text{and}$$

$$s_2^2 = \frac{54{,}600}{5} = 10{,}920$$

so that $s_p^2 = (63{,}000 + 54{,}600)/(5 + 4) = 13{,}066.7$, $s_p = 114.31$, and

$$t = \frac{8{,}230 - 7{,}940}{114.31\sqrt{\frac{1}{5} + \frac{1}{6}}} = 4.19$$

5. *Decision*: Since $t = 4.19$ exceeds 3.250, the null hypothesis must be rejected at level $\alpha = 0.01$. We conclude that the average heat-producing capacity of the coal from the two mines is not the same. The very small P-value 0.0024 (see Figure 8.3) strengthens this conclusion. ■

In the preceding example we went ahead and performed the two sample t test, tacitly assuming that the population variances are equal. Fortunately, the test is not overly sensitive to small differences between the population variances, and the procedure used in this instance is justifiable. As a rule of thumb, if one variance is four times the other, we should be concerned. A transformation will often improve the situation. As another alternative there is the Smith-Satterthwaite test discussed below.

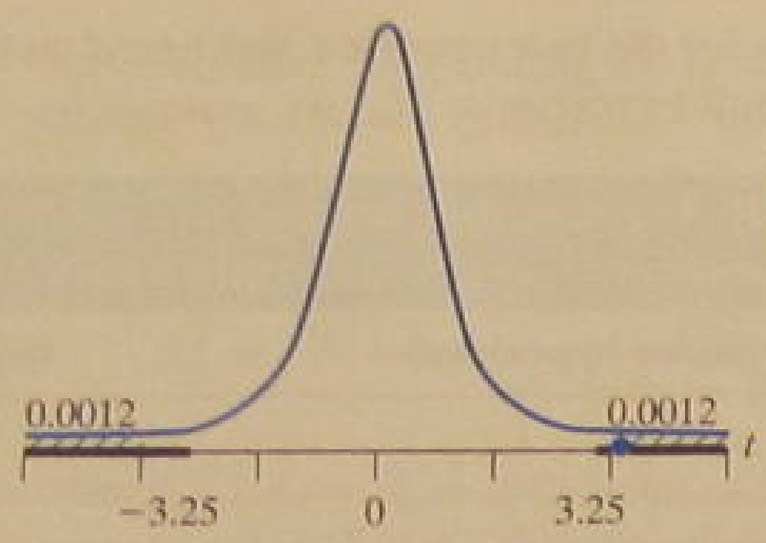


Figure 8.3
Two sample t test: P-value = 0.0024 for $t = 4.19$

Confidence intervals follow directly from the acceptance region for the tests. For two normal populations with equal variances,

Small sample confidence interval concerning difference between two means

The $(1-\alpha)100\%$ confidence interval for $\delta = \mu_1 - \mu_2$ has limits

$$\bar{x} - \bar{y} \pm t_{\alpha/2}\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

where $t_{\alpha/2}$ is based on $\nu = n_1 + n_2 - 2$ degrees of freedom.

EXAMPLE **Graphics to accompany a two sample t test**

Strength measurements on an aluminum alloy were introduced on page 19. A second alloy yielded measurements given in the following stem-and-leaf display. Find a 95% confidence interval for the difference in mean strength δ.

Solution We first place the observations on the two alloys in stem-and-leaf displays. Note that the observations from the first alloy appear normal, but those on the second alloy may deviate from a normal distribution. Since the sample sizes are relatively large, this will not cause any difficulty.

Alloy 1, $N = 58$
Leaf unit = 0.10

66	4
67	7
68	0 0 3 4 6 8 9
69	0 1 2 3 3 5 5 6 7 8 8 9
70	0 0 1 2 3 3 4 5 6 6 8 9
71	0 1 2 3 3 5 6 6 7 8 8 9
72	1 2 3 4 6 7 9
73	1 3 5
74	2 5
75	3
76	
77	
78	
79	

Alloy 2, $N = 27$
Leaf unit = 0.10

66	
67	
68	
69	
70	
71	2 8
72	6 8
73	4 7 9
74	4 9
75	5 9
76	3 5 7 9
77	1 3 6 7 8
78	1 2 4 6
79	0 3 8

A computer calculation gives the sample means and standard deviations.

	N	MEAN	STDEV
ALLOY 1	58	70.70	1.80
ALLOY 2	27	76.13	2.42

From another computer calculation (or by interpolation in Table 4), we find $t_{0.025} = 1.99$ for 83 degrees of freedom, so the 95% confidence interval is

$$\bar{x} - \bar{y} \pm t_{\alpha/2} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_1}}$$

$$= 70.70 - 76.13 \pm 1.99 \sqrt{\frac{57(1.80)^2 + 26(2.42)^2}{83}} \sqrt{\frac{1}{58} + \frac{1}{27}}$$

or

$$-6.4 < \mu_1 - \mu_2 < -4.5$$

We are 95% confident that the mean strength of alloy 2 is 4.5 to 6.4 thousand pounds per square inch higher than the mean strength of alloy 1. ■

It is good practice to show stem-and-leaf displays, boxplots, or histograms. Often they reveal more than a mean difference. For instance, in the last example, the first population is nearly symmetric but the second has a long tail to the left.

The large sample confidence interval is obtained from the acceptance regions for the test on page 249.

EXAMPLE **Comparing the two confidence intervals**

Referring to the previous example, find the 95% large sample confidence interval.

Solution $$\bar{x} - \bar{y} \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 70.70 - 76.13 \pm 1.96 \sqrt{\frac{(1.80)^2}{58} + \frac{(2.42)^2}{27}}$$

or

$$-6.5 < \mu_1 - \mu_2 < -4.4$$

There is not much difference between this 95% confidence interval and the one in the previous example, where the variances were pooled. ■

Small Sample Sizes but Unequal Standard Deviations—Normal Populations

When we deal with independent random samples from normal populations whose variances seem to be unequal, we should not pool. As long as the populations are normal, an approximate t distribution is available for making inferences. The statistic t' is the same as the large samples statistic but, because sample sizes are small, its distribution is approximated as a t distribution.

Statistic for small samples inference, $\sigma_1 \neq \sigma_2$, normal populations

For normal populations, when the sample sizes n_1 and n_2 are not large and $\sigma_1 \neq \sigma_2$,

$$t' = \frac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

is approximately distributed as a t with estimated degrees of freedom.

The estimated degrees of freedom for t' are calculated from the observed values of the sample variances s_1^2 and s_2^2.

$$\text{estimated degrees of freedom} = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$$

The estimated degrees of freedom are often rounded down to an integer so a t table can be consulted.

The test based on t' is called the **Smith-Satterthwaite test**.

EXAMPLE **Testing equality of mean product volume**

One process of making green gasoline takes sucrose, which can be derived from biomass, and converts it into gasoline using catalytic reactions. This is not a process for making a gasoline additive but fuel itself, so research is still at the pilot plant stage. At one step in a pilot plant process, the product consists of carbon chains of length 3. Nine runs were made with each of two catalysts and the product volumes (gal) are

catalyst 1	0.63	2.64	1.85	1.68	1.09	1.67	0.73	1.04	0.68
catalyst 2	3.71	4.09	4.11	3.75	3.49	3.27	3.72	3.49	4.26

The sample sizes $n_1 = n_2 = 9$ and the summary statistics are

$$\overline{x} = 1.334, \qquad s_1^2 = 0.4548 \qquad \overline{y} = 3.766 \qquad s_2^2 = 0.1089$$

A chemical engineer wants to show that the mean product volumes are different. Test with $\alpha = 0.05$.

Solution The test concerns $\delta = \mu_1 - \mu_2$ and the sample sizes $n_1 = n_2 = 9$ are small. We note that there are no outliers and no obvious departure from normality. However, $s_1^2 = 0.4548$ is more than four times $s_2^2 = 0.1089$. We should not pool.

1. *Null hypothesis*: $\delta = 0$
   Alternative hypothesis: $\delta \neq 0$
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: We choose the Smith-Satterthwaite test statistic with $\delta_0 = 0$,

$$t' = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

The null hypothesis will be rejected if $t' < -t_{0.025}$ or $t' > t_{0.025}$, but the value of $t_{0.025}$ depends on the estimated degrees of freedom.

4. *Calculations*: As a first step, we estimate the degrees of freedom

$$\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}=\frac{\left(\frac{0.4548}{9}+\frac{0.1089}{9}\right)^2}{\frac{(0.4548/9)^2}{9-1}+\frac{(0.1089/9)^2}{9-1}}=11.62$$

To use the t table, we round down to 11 and obtain $t_{0.025}=2.201$. The observed value of the test statistic is

$$t'=\frac{\overline{X}-\overline{Y}-\delta_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}=\frac{1.334-3.766-0}{\sqrt{\frac{0.4548}{9}+\frac{0.1089}{9}}}=-9.71$$

5. *Decision*: Since $t'=-9.71$ is less than -2.201, the null hypothesis must be rejected at level of significance 0.05. The value of t' is so small that the P-value is 0.0000 when rounded. In other words, there is extremely strong evidence that the mean product volumes are different for the two catalysts. ■

Confidence intervals can tell us what differences in means are plausible, not just that the means are different.

Confidence interval for $\delta=\mu_1-\mu_2$, normal populations $\sigma_1\neq\sigma_2$

A 100 $(1-\alpha)$% confidence interval for $\delta=\mu_1-\mu_2$

$$\left(\overline{x}-\overline{y}-t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},\ \overline{x}-\overline{y}+t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$$

where $t_{\alpha/2}$ has the degrees of freedom estimated for t'.

EXAMPLE **A confidence interval for the difference of mean yields when variances are unequal**

With reference to the previous example, obtain the 95% confidence interval for $\delta=\mu_1-\mu_2$

Solution From the previous example we have $\overline{x}=1.334$, $s_1^2=0.4548$, $\overline{y}=3.766$, $s_2^2=0.1089$, and $t_{0.025}=2.201$ for 11 degrees of freedom. We get

$$\left(1.334-3.776-2.201\sqrt{0.45489+\frac{0.1089}{9}},\ 1.334-3.776+2.201\sqrt{0.45489+\frac{0.1089}{9}}\right)$$

or $(-2.982, -1.880)$ gallons. The mean product volume for the second catalyst is greater than that of the first catalyst by 1.880 to 2.982 gallons. ■

Although the computations for unequal standard deviations seem tedious by hand, popular statistical software will allow this option in addition to pooling (see Exercise 8.35).

Exercises

8.1 Refer to Example 2.58, where $n_1=30$ specimens of 2×4 lumber have $\overline{x}=1{,}908.8$ and $s_1=327.1$ psi. A second sample of size $n_2=40$ specimens of larger dimension, 2×6, lumber yielded $\overline{y}=2{,}114.3$ and $s_2=472.3$. Test, with $\alpha=0.05$, the null hypothesis of equality of mean tensile strengths versus the one-sided alternative that the mean tensile strength for the second population is greater than that of the first.

8.2 Refer to Exercise 8.1 and obtain a 95% confidence interval for the difference in mean tensile strength.

8.3 The dynamic modulus of concrete is obtained for two different concrete mixes. For the first mix, $n_1 = 33$, $\bar{x} = 115.1$, and $s_1 = 0.47$ psi. For the second mix, $n_2 = 31$, $\bar{y} = 114.6$, and $s_2 = 0.38$. Test, with $\alpha = 0.05$, the null hypothesis of equality of mean dynamic modulus versus the two-sided alternative.

8.4 Refer to Exercise 8.3 and obtain a 95% confidence interval for the difference in mean dynamic modulus.

8.5 An investigation of two kinds of photocopying equipment showed that 75 failures of the first kind of equipment took on the average 83.2 minutes to repair with a standard deviation of 19.3 minutes, while 75 failures of the second kind of equipment took on the average 90.8 minutes to repair with a standard deviation of 21.4 minutes.

(a) Test the null hypothesis $\mu_1 - \mu_2 = 0$ (namely, the hypothesis that on the average it takes an equal amount of time to repair either kind of equipment) against the alternative hypothesis $\mu_1 - \mu_2 \neq 0$ at the level of significance $\alpha = 0.05$.

(b) Using 19.3 and 21.4 as estimates of σ_1 and σ_2, find the probability of failing to reject the null hypothesis $\mu_1 - \mu_2 = 0$ with the criterion of part (a) when actually $\mu_1 - \mu_2 = -12$.

8.6 Studying the flow of traffic at two busy intersections between 4 P.M. and 6 P.M. (to determine the possible need for turn signals), it was found that on 40 weekdays there were on the average 247.3 cars approaching the first intersection from the south that made left turns while on 30 weekdays there were on the average 254.1 cars approaching the second intersection from the south that made left turns. The corresponding sample standard deviations are $s_1 = 15.2$ and $s_2 = 18.7$.

(a) Test the null hypothesis $\mu_1 - \mu_2 = 0$ against the alternative hypothesis $\mu_1 - \mu_2 \neq 0$ at the level of significance $\alpha = 0.01$.

(b) Using 15.2 and 18.7 as estimates of σ_1 and σ_2, find the probability of failing to reject (accepting) the null hypothesis $\mu_1 - \mu_2 = 0$ when actually $|\mu_1 - \mu_2| = 15.6$.

8.7 Given the $n_1 = 3$ and $n_2 = 2$ observations from Population 1 and Population 2, respectively,

Population 1	6	2	7
Population 2	14	10	

(a) Calculate the three deviations $x - \bar{x}$ and two deviations $y - \bar{y}$.

(b) Use your results from part (a) to obtain the pooled variance.

8.8 Two procedures for etching integrated circuits are to be compared. Given 10 units, five are prepared using etching procedure A and five are prepared using etching procedure B.

(a) How would you assign etching procedures to the 10 units?

(b) The response is the percent of area on the integrated circuit where the etching was inadequate. Suppose the results are

Procedure A	Procedure B
5	1
2	3
9	4
6	0
3	2

Find a 95% confidence interval for the difference in means.

(c) What assumptions did you make for your answer to part (b)?

8.9 Measuring specimens of nylon yarn taken from two spinning machines, it was found that 8 specimens from the first machine had a mean denier of 9.67 with a standard deviation of 1.81, while 10 specimens from the second machine had a mean denier of 7.43 with a standard deviation of 1.48. Assuming that the populations sampled are normal and have the same variance, test the null hypothesis $\mu_1 - \mu_2 = 1.5$ against the alternative hypothesis $\mu_1 - \mu_2 > 1.5$ at the 0.05 level of significance.

8.10 As part of an industrial training program, some trainees are instructed by Method A, which is straight computer-based instruction, and some are instructed by Method B, which also involves the personal attention of an instructor. If random samples of size 10 are taken from large groups of trainees instructed by each of these two methods, and the scores which they obtained in an appropriate achievements test are

Method A: 71 75 65 69 73 66 68 71 74 68
Method B: 72 77 84 78 69 70 77 73 65 75

use the 0.05 level of significance to test the claim that Method B is more effective. Assume that the populations sampled can be approximated closely with normal distributions having the same variance.

8.11 The following are the number of sales which a sample of 9 salespeople of industrial chemicals in California and a sample of 6 salespeople of industrial chemicals in Oregon made over a certain fixed period of time:

California: 59 68 44 71 63 46 69 54 48
Oregon: 50 36 62 52 70 41

Assuming that the populations sampled can be approximated closely with normal distributions having the same variance, test the null hypothesis $\mu_1 - \mu_2 = 0$

against the alternative hypothesis $\mu_1 - \mu_2 \neq 0$ at the 0.01 level of significance.

8.12 With reference to the example on page 249, construct a 95% confidence interval for the true difference between the average resistance of the two kinds of wire.

8.13 In each of the parts below, first decide whether or not to use the pooled estimator of variance. Assume that the populations are normal.

(a) The following are the Brinell hardness values obtained for samples of two magnesium alloys before testing:

Alloy 1:	66.3	63.5	64.9	61.8	64.3	64.7	65.1	64.5	68.4	63.2
Alloy 2:	71.3	60.4	62.6	63.9	68.8	70.1	64.8	68.9	65.8	66.2

Use the 0.05 level of significance to test the null hypothesis $\mu_1 - \mu_2 = 0$ against the alternative hypothesis $\mu_1 - \mu_2 < 0$.

(b) To compare two kinds of bumper guards, 6 of each kind, were mounted on a certain kind, of compact car. Then each car was run into a concrete wall at 5 miles per hour, and the following are the costs of the repairs (in dollars):

Bumper guard 1:	407	448	423	465	402	419
Bumper guard 2:	434	415	412	451	433	429

Use the 0.01 level of significance to test whether the difference between the two sample means is significant.

8.4 Matched Pairs Comparisons

In the application of the two sample t test we need to watch that the samples are independent. For instance, the test cannot be used when we deal with "before and after" data, the IQs of husbands and wives, and numerous other kinds of situations where the data are naturally paired. Instead, comparisons are based on the **matched pairs**.

A manufacturer is concerned about the loss of weight of ceramic parts during a baking step. The readings before and after baking, on the same specimen, are naturally paired. It would make no sense to compare the before-baking weight of one specimen with the after-baking weight of another specimen. Let the pair of random variables (X_i, Y_i) denote the weight before and weight after baking for the ith specimen, for $i = 1, 2, \ldots, n$. A statistical analysis proceeds by considering the differences

$$D_i = X_i - Y_i \qquad \text{for } i = 1, 2, \ldots, n$$

This collection of (signed) differences is then treated as a random sample of size n from a population having mean μ_D. We interpret $\mu_D = 0$ as indicating that the means of the two responses are the same and $\mu_D > 0$ as indicating that the mean response of the first is higher than that of the second.

Tests of the null hypothesis H_0: $\mu_D = \mu_{D,0}$ are based on the ratio

$$\frac{\overline{D} - \mu_{D,0}}{S_D/\sqrt{n}} \quad \text{where } \overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} \quad S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}$$

If n is small, we treat this ratio as the one sample t statistic on page 233. Otherwise, we treat this ratio as the large sample statistic on page 232.

EXAMPLE **Conducting a paired t test**

The following are the average weekly losses of worker-hours due to accidents in 10 industrial plants before and after a certain safety program was put into operation:

Before:	45	73	46	124	33	57	83	34	26	17
After:	36	60	44	119	35	51	77	29	24	11

Use the 0.05 level of significance to test whether the safety program is effective.

Solution We cannot apply the independent samples test because the before and after weekly losses of worker-hours in the same industrial plant are correlated. Here there is the obvious pairing of these two observations.

1. *Null hypothesis*: $\mu_D = 0$
   Alternative hypothesis: $\mu_D > 0$
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $t > 1.833$, the value of $t_{0.005}$ for $10 - 1 = 9$ degrees of freedom, where

$$t = \frac{\overline{D} - 0}{S_D/\sqrt{n}}$$

   and $\overline{D}$ and S_D are the mean and the standard deviation of the differences.
4. *Calculations*: The differences are

$$9 \quad 13 \quad 2 \quad 5 \quad -2 \quad 6 \quad 6 \quad 5 \quad 2 \quad 6$$

   their mean is $\overline{d} = 5.2$, their standard deviation is $s_D = 4.08$, so that

$$t = \frac{5.2 - 0}{4.08/\sqrt{10}} = 4.03$$

5. *Decision*: Since $t = 4.03$ exceeds 1.833, the null hypothesis must be rejected at level $\alpha = 0.05$. We conclude that the industrial safety program is effective. The evidence is very strong, since a computer calculation gives the P-value 0.0015 (see Figure 8.4). If $\mu_D = 0$, only in 15 out of 10,000 times would we observe t greater than or equal to 4.03. ■

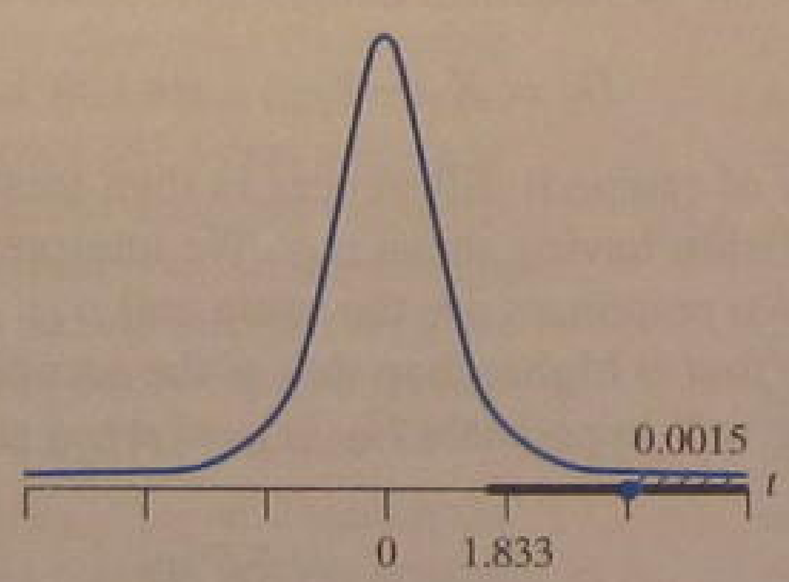


Figure 8.4
The paired t test: P-value = 0.0015 for $t = 4.03$ and degrees of freedom (d.f.) = 9

In connection with this kind of problem, the one sample t test is referred to as a **matched pairs t test** or just the **paired t test**.

EXAMPLE **A 90% confidence interval for the mean of a paired difference**

Referring to the previous example, find a 90% confidence interval for the mean improvement in lost worker-hours.

Solution The $n = 10$ differences have $\overline{d} = 5.2$ and $s_D = 4.08$. Since $t_{0.05} = 1.833$, the 90% confidence interval for μ_D, the mean improvement, is

$$5.2 - 1.833\,\frac{4.08}{\sqrt{10}} < \mu_D < 5.2 + 1.833\,\frac{4.08}{\sqrt{10}} \quad \text{or} \quad 2.8 < \mu_D < 7.7$$

worker-hours per week. We are 90% confident that between 2.8 and 7.7 fewer worker-hours per week are lost, on average, after the safety program. ■

The next example illustrates some practical points when conducting a matched pairs experiment, including randomization.

EXAMPLE **Comparing measurements made at two laboratories**

A state law requires municipal wastewater treatment plants to monitor their discharges into rivers and streams. A treatment plant could choose to send its samples to a commercial laboratory of its choosing. Concern over this self-monitoring led a civil engineer to design a matched pairs experiment.[2] Exactly the same bottle of effluent cannot be sent to two different laboratories. To match "identical" as closely as possible, she would take a sample of effluent in a large sample bottle and pour it back and forth over two open specimen bottles. When they were filled and capped, a coin was flipped to see if the one on the right was sent to Commercial Laboratory A or the Wisconsin State Laboratory of Hygiene. This process was repeated 11 times. The results, for the response suspended solids (SS) are

Sample	1	2	3	4	5	6	7	8	9	10	11
Commercial lab	27	23	64	44	30	75	26	124	54	30	14
State lab	15	13	22	29	31	64	30	64	56	20	21
Difference $x_i - y_i$	12	10	42	15	−1	11	−4	60	−2	10	−7

Obtain a 95% confidence interval and look for any unusual features in the data.

Solution The sample size is relatively small so we assume normality and base the confidence interval on the t distribution. We have $n = 11$ and calculate $\bar{d} = 13.27$ and $s_D^2 = 418.61$ Then, with $n - 1 = 11 - 1 = 10$ degrees of freedom, $t_{0.025} = 2.228$ and the 95% confidence interval is

$$\left(13.27 - 2.228\sqrt{\frac{418.61}{11}},\ 13.27 + 2.228\sqrt{\frac{418.61}{11}}\right) \quad \text{or} \quad (-0.47,\ 27.01)$$

This 95% confidence interval just covers 0, so no difference is indicated with this small sample size. But wait, look at the dot diagram of the differences in Figure 8.5. There are two very large differences that would be unusual if the sample were taken from a normal population. The validity of the confidence interval is, at least, under suspicion. In Exercise 8.17 you are asked to try the square root transformation to see if it improves the situation. ■

Figure 8.5
Dot diagram of differences in suspended solids; outliers present.

[2] R. Johnson and D. Wichern (2007), *Applied Multivariate Statistical Analysis*, Prentice Hall, page 276.

Exercises

8.14 A civil engineer wants to compare two instruments for measuring the amount of polychlorinated biphenyls (PCBs) in corn stalks. A sample of stalks is cut and crushed and two scoops of the material taken. One is measured with the first instrument and the other with the second instrument. This whole process is repeated five times. The results, in parts per billion, are as follows:

Sample No.	Instrument 1	Instrument 2
1	3	4
2	8	7
3	9	6
4	4	3
5	6	5

Find a 95% confidence interval for the mean difference in instrument readings assuming the differences have a normal distribution.

8.15 Refer to Exercise 8.14. Test, with $\alpha = 0.05$, that the mean difference is 0 versus a two-sided alternative.

8.16 The following data were obtained in an experiment designed to check whether there is a systematic difference in the weights obtained with two different scales:

	Weight in grams	
	Scale I	Scale II
Rock Specimen 1	11.23	11.27
Rock Specimen 2	14.36	14.41
Rock Specimen 3	8.33	8.35
Rock Specimen 4	10.50	10.52
Rock Specimen 5	23.42	23.41
Rock Specimen 6	9.15	9.17
Rock Specimen 7	13.47	13.52
Rock Specimen 8	6.47	6.46
Rock Specimen 9	12.40	12.45
Rock Specimen 10	19.38	19.35

Use the paired t test at the 0.05 level of significance to try to establish that the mean difference of the weights obtained with the two scales is nonzero.

8.17 Refer to the example concerning suspended solids in effluent from a treatment plant on page 261. Take the square root of each of the measurements and then take the difference.

(a) Construct a 95% confidence interval for μ_D.

(b) Conduct a level $\alpha = 0.05$ level test of H_0: $\mu_D = 0$ against a two-sided alternative. Verify that the conclusion is the same as that obtained from the confidence interval.

(c) Make a dot diagram of these differences and decide if the transformation has essentially removed the outliers.

8.18 Refer to the example concerning suspended solids in effluent from a treatment plant on page 261. Take the natural logarithm of each of the measurements and then take the difference.

(a) Construct a 95% confidence interval for μ_D.

(b) Conduct a level $\alpha = 0.05$ level test of H_0: $\mu_D = 0$ against a two-sided alternative. Verify that the conclusion is the same as that obtained from the confidence interval.

(c) Make a dot diagram of these differences and decide if the transformation has essentially removed the outliers.

8.19 A sunglass company wants potential customers to compare two types of lenses, the current lens A and one made of a new hi-tech material B. Glasses made with lens A and glasses made with lens B are available. Each person, in a sample of size 43, is asked to wear one of each type for a whole day. After a day in the sun, they are asked to score that day's pair on a scale of 1 to 7, with higher scores being best. The differences in scores (hi-tech lens B) − (current lens A) have mean 0.7 and variance 3.8. Construct a 95% confidence interval for the mean difference.

8.20 Referring to Exercise 8.19, conduct a test to show that the hi-tech lens is preferred by customers. Take $\alpha = 0.05$.

8.21 In a study of the effectiveness of physical exercise in weight reduction, a group of 16 persons engaged in a prescribed program of physical exercise for one month showed the following results:

Weight before (pounds)	Weight after (pounds)	Weight before (pounds)	Weight after (pounds)
209	196	170	164
178	171	153	152
169	170	183	179
212	207	165	162
180	177	201	199
192	190	179	173
158	159	243	231
180	180	144	140

Use the 0.01 level of significance to test whether the prescribed program of exercise is effective.

8.5 Design Issues—Randomization and Pairing

Often, the experimenter can make the choice of which of the two treatments is applied to an individual unit. Then it is possible to conduct either an independent samples comparison or a matched pairs comparison of means. The method of assignment of units to two groups or pairs can be crucial to the validity of the statistical procedures for comparing the means of two populations. We first emphasize the importance of randomization in the context of the independent samples design of a comparative experiment.

Independent Samples Design: Randomization

In many comparative studies the investigator applies one or the other treatment to an object we call an experimental unit. The method of assigning the treatments to the experimental units can be crucial to the validity of the statistical procedures. Suppose a chemist has a new formula for waterproofing that she applies to several pairs of shoes that are almost like new. She applies the old formula to several pairs of scuffed shoes. At the end of a month, she will measure the ability of each pair of shoes to withstand water. It doesn't take a statistician to see that this is not a good experimental design. The persons with scuffed shoes probably walk a lot more and do so in all kinds of weather. These sources of variation could very well lead to systematic biases that make the new formula seem better than the old even when this is not the case. The pairs of shoes need to be assigned to the treatments with old and new waterproofing formula in a random manner.

When possible, the $n = n_1 + n_2$ experimental units should be assigned at random to the two treatments. This means that all $\binom{n}{n_1}$ possible selections of n_1 units to receive the first treatment are equally likely. Practically, the assignment is accomplished by selecting n_1 random integers between 1 and n. The corresponding experimental units are assigned to the first treatment. Generally, a test will have more power if the two sample sizes are equal.

In summary, we must actively assign treatments at random to experimental units. This process is called **randomization**.

Purpose of randomization

> **Randomization** of treatments prevents uncontrolled sources of variation from exerting a systematic influence on the responses.

Matched Pairs Design: Pairing and Randomization

The object of **pairing** experimental units, according to a characteristic that is likely to influence the response, is to eliminate this source of variation from the comparison. In the context of waterproofing for shoes, each person could have the old formula on one shoe and the new formula on the other. Since the paired t analysis only uses differences from the same pair, this experimental strategy should eliminate most of the variation in response due to different terrain, distance covered, and weather conditions.

> **Pairing** according to some variable(s) thought to influence the response will remove the effect of that variable from analysis.

Even after units are paired, there is a need for randomization. For each pair, a fair coin should be flipped to assign the treatments. In the context of the waterproofing example, the old formula could be applied to the right shoe if heads and left shoe if tails. The new formula is applied to the other shoe. This randomization, restricted to be within pairs, would prevent systematic influences such as those caused by the fact that a majority of persons would tend to kick things with their right shoe.

> **Randomizing** the assignment of treatments within a pair helps prevent any other uncontrolled variables from influencing the responses in a systematic manner.

Notice that in the example on page 259, the experimenter had no control over the before and after. Many uncontrolled variables may also have changed over the course of the experiment: fewer working hours due to strikes, phasing out of an old type of equipment, etc. One of these could have been the cause for the improvement rather than the safety program.

We pursue the ideas of randomization and blocking in Chapter 12. Our purpose here was to show what practical steps can be taken to meet the idealistic assumptions of random samples when comparing two treatments.

Exercises

8.22 An investigator wants to compare two busy network protocols by recording the number of messages that are successfully passed by the network in a day. Describe how to select 5 of the next 10 working days for trying Protocol 1. Protocol 2 would be tried on the other 5 days.

8.23 An electrical engineer has developed a modified circuit board for elevators. Suppose 3 modified circuit boards and 6 elevators are available for a comparative test of the old versus the modified circuit boards.

(a) Describe how you would select the 3 elevators in which to install the modified circuit boards. The old circuit boards will be installed in the other 3 elevators.

(b) Alternatively, describe how you would conduct a paired comparison and then randomize within the pair.

8.24 It takes an average of 10 weeks to train a typical employee to run a computer-aided machine. The instructor has a new approach that she feels will lead to faster learning. She intends to teach 5 persons by the new method and then compare their learning times with those of 5 randomly selected persons trained by the old method. In order to obtain 5 students from the 25 available candidates, she asks for volunteers. Why is this a bad idea?

8.25 How would you randomize, for a two sample test, if 50 cars are available for an emissions study and you want to compare a modified air pollution device with the current production?

Do's and Don'ts

Do's

1. When sample sizes are large, determine the limits of a $100(1-\alpha)\%$ confidence interval for the difference of means $\mu_1 - \mu_2$ as

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

2. When each of the two samples are from normal populations having the same variance, determine the limits of a $100(1-\alpha)\%$ confidence interval for the difference of means $\mu_1 - \mu_2$ as

$$\bar{x} - \bar{y} \pm t_{\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the pooled estimate of variance

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

and $t_{\alpha/2}$ is based on $n_1 + n_2 - 2$ degrees of freedom.

3. When analyzing data from a matched pair design, use the results for one sample but applied to the differences from each matched pair. If the difference of paired measurements has a normal distribution, determine a $100(1-\alpha)\%$ confidence interval for the mean difference μ_D as

$$\left(\bar{d} - t_{\alpha/2}\frac{s_D}{\sqrt{n}},\ \bar{d} + t_{\alpha/2}\frac{s_D}{\sqrt{n}}\right)$$

where $t_{\alpha/2}$ is based on $n - 1$ degrees of freedom.

4. When comparing two treatments using the independent samples design, randomly assign the treatments to groups whenever possible. With the matched pair design, randomly assign the treatments within each pair.

Don'ts

1. Don't pool the two sample variances s_1^2 and s_2^2 if they are very different. We suggest a factor of 4 as being too different.

Confidence interval for σ^2

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

If we take the square root of each member of this inequality, we obtain a corresponding $(1-\alpha)100\%$ confidence interval for σ.

Note that confidence intervals for σ or σ^2 obtained by taking equal tails, as in the above formula, do not actually given the narrowest confidence intervals, because the chi square distribution is not symmetrical. (See Exercise 7.26.) Nevertheless, they are used in most applications in order to avoid fairly complicated calculations. For moderate degrees of freedom, the choice of equal tails is inconsequential.

EXAMPLE **Finding a 95% confidence interval for σ**

Returning to the example on page 190, suppose that the refractive indices of 20 pieces of glass (randomly selected from a large shipment purchased by the optical firm) have a variance of $1.20 \cdot 10^{-4}$. Construct a 95% confidence interval for σ, the standard deviation of the population sampled.

Solution For $20-1 = 19$ degrees of freedom, $\chi^2_{0.975} = 8.907$ and $\chi^2_{0.025} = 32.852$ according to Table 5, so that substitution into the formula yields

$$\frac{(19)(1.20 \cdot 10^{-4})}{32.852} < \sigma^2 < \frac{(19)(1.20 \cdot 10^{-4})}{8.907}$$

$$0.000069 < \sigma^2 < 0.000256$$

and, taking the square root,

$$0.0083 < \sigma < 0.0160$$

This means we are 95% confident that the interval from 0.0083 to 0.0160 contains σ, the true standard deviation of the refractive index. ■

The method which we have discussed applies only to random samples from normal populations (or at least to random samples from populations which can be approximated closely with normal distributions).

Exercises

9.1 Use the data of Exercise 7.57 to estimate σ for the key performance indicator in terms of

(a) the sample standard deviation;

(b) the sample range.

Compare the two estimates by expressing their difference as a percentage of the first.

9.2 With reference to the example on page 253, use the range of the second sample to estimate σ for the heat-producing capacity of coal from the second mine, and compare the result with the standard deviation of the second sample.

9.3 Use the data of part (a) of Exercise 8.13 to estimate σ for the Brinell hardness of Alloy 1 in terms of

(a) the sample standard deviation;

(b) the sample range.

Compare the two estimates by expressing their difference as a percentage of the first.

9.4 With reference to Exercise 7.56, construct a 99% confidence interval for the variance of the yield.

9.5 With reference to Exercise 7.63, construct a 99% confidence interval for the variance of the population sampled.

9.6 Use the value s obtained in Exercise 9.3 to construct a 98% confidence interval for σ, measuring the actual variability in the hardness of Alloy 1.

9.2 Hypotheses Concerning One Variance

In this section we shall consider the problem of testing the null hypothesis that a population variance equals a specified constant against a suitable one-sided or two-sided alternative; that is, we shall test the null hypothesis $\sigma^2 = \sigma_0^2$ against one of the alternatives $\sigma^2 < \sigma_0^2, \sigma^2 > \sigma_0^2$, or $\sigma^2 \neq \sigma_0^2$. Tests like these are important whenever it is desired to control the uniformity of a product or an operation. For example, suppose that a silicon disk or wafer is to be cut into small squares, or dice, to be used in the manufacture of a semiconductor device. Since certain electrical characteristics of the finished device may depend on the thickness of the die, it is important that all dice cut from a wafer have approximately the same thickness. Thus, not only must the mean thickness of a wafer be kept within specifications, but also the variation in thickness from location to location on the wafer.

Using the same sampling theory as on page 269, we base such tests on the fact that for a random sample from a normal population with the variance σ_0^2

Statistic for test concerning variance (normal population)

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

is a random variable having the chi square distribution with $n-1$ degrees of freedom. The critical regions for such tests are as shown in the following table:

Critical regions for testing $\sigma^2 = \sigma_0^2$ (normal population)	
Alternative hypothesis	**Reject null hypothesis if:**
$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha}$
$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{\alpha}$
$\sigma^2 \neq \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha/2}$ or $\chi^2 > \chi^2_{\alpha/2}$

In this table χ^2_α is as defined on page 189. Note that equal tails are used for the two-sided alternative, although this is actually not the best procedure since the chi square distribution is not symmetrical. For moderate degrees of freedom, the two tests are nearly the same.

EXAMPLE **Testing hypotheses concerning a standard deviation**

The lapping process which is used to grind certain silicon wafers to the proper thickness is acceptable only if σ, the population standard deviation of the thickness of dice cut from the wafers, is at most 0.50 mil. Use the 0.05 level of significance to test the null hypothesis $\sigma = 0.50$ against the alternative hypothesis $\sigma > 0.50$, if the thicknesses of 15 dice cut from such wafers have a standard deviation of 0.64 mil.

Solution

1. *Null hypothesis*: $\sigma = 0.50$
   Alternative hypothesis: $\sigma > 0.50$
2. *Level of significance*: $\alpha = 0.05$

3. *Criterion*: Reject the null hypothesis if $\chi^2 > 23.685$, the value of $\chi^2_{0.05}$ for 14 degrees of freedom, where

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

4. *Calculations*:

$$\chi^2 = \frac{(15-1)(0.64)^2}{(0.50)^2} = 22.94$$

5. *Decision*: Since $\chi^2 = 22.94$ does not exceed 23.685, the null hypothesis cannot be rejected; even though the sample standard deviation exceeds 0.50, there is not sufficient evidence to conclude that the lapping process is unsatisfactory. ■

Statistical software is readily available to obtain other values of χ^2_α. See Exercise 9.23 for the *MINITAB* and Appendix C for the *R* commands.

9.3 Hypotheses Concerning Two Variances

The two sample t test described in Section 8.2 requires that the variances of the two populations sampled are equal. In this section we describe a test of the null hypothesis $\sigma_1^2 = \sigma_2^2$, which applies to independent random samples from two normal populations. It must be used with some discretion as it is very sensitive to departures from this assumption.

If independent random samples of size n_1 and n_2 are taken from normal populations having the same variance, it follows from Theorem 6.6 that

Statistic for test of equality of two variances (normal populations)

$$F = \frac{S_1^2}{S_2^2}$$

is a random variable having the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Thus, if the null hypothesis $\sigma_1^2 = \sigma_2^2$ is true, the ratio of the sample variances S_1^2 and S_2^2 provides a statistic on which tests of the null hypothesis can be based.

The critical region for testing the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $\sigma_1^2 > \sigma_2^2$ is $F > F_\alpha$, where F_α is as defined on page 190. Similarly, the critical region for testing the null hypothesis against the alternative hypothesis $\sigma_1^2 < \sigma_2^2$ is $F < F_{1-\alpha}$, and this causes some difficulties since Table 6 only contains values corresponding to right-hand tails of $\alpha = 0.05$ and $\alpha = 0.01$. However, we can use the reciprocal of the original test statistic and make use of the relation

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)}$$

first given on page 191. Thus, we base the test on the statistic $F = S_2^2/S_1^2$ and the critical region for testing the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $\sigma_1^2 < \sigma_2^2$ becomes $F > F_\alpha$, where F_α is the appropriate critical value of F for $n_2 - 1$ and $n_1 - 1$ degrees of freedom.

For the two-sided alternative $\sigma_1^2 \neq \sigma_2^2$, the critical region is $F < F_{1-\alpha/2}$ or $F > F_{\alpha/2}$, where $F = S_1^2/S_2^2$ and the degrees of freedom are $n_1 - 1$ and $n_2 - 1$. In

practice, we modify this test as in the preceding paragraph, so that we can again use the table of F values corresponding to right-hand tails of $\alpha = 0.05$ and $\alpha = 0.01$. To this end we let S_M^2 represent the larger of the two sample variances, S_m^2 the smaller, and we write the corresponding sample sizes as n_M and n_m. Thus, the test statistic becomes $F = S_M^2/S_m^2$ and the critical region is as shown in the following table:

Critical regions for testing $\sigma_1^2 = \sigma_2^2$ (normal populations)

Alternative hypothesis	Test statistic	Reject null hypothesis if:
$\sigma_1^2 < \sigma_2^2$	$F = \dfrac{S_2^2}{S_1^2}$	$F > F_\alpha(n_2 - 1, n_1 - 1)$
$\sigma_1^2 > \sigma_2^2$	$F = \dfrac{S_1^2}{S_2^2}$	$F > F_\alpha(n_1 - 1, n_2 - 1)$
$\sigma_1^2 \neq \sigma_2^2$	$F = \dfrac{S_M^2}{S_m^2}$	$F > F_{\alpha/2}(n_M - 1, n_m - 1)$

The level of significance of these tests is α and the figures indicated in parentheses are the respective degrees of freedom. Note that, as in the chi square test, equal tails are used in the two-tailed test as a matter of mathematical convenience, even though the F distribution is not symmetrical.

EXAMPLE **A one-sided F test of the equality of two variances**

It is desired to determine whether there is less variability in the silver plating done by Company 1 than in that done by Company 2. If independent random samples of size 12 of the two companies' work yield $s_1 = 0.035$ mil and $s_2 = 0.062$ mil, test the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $\sigma_1^2 < \sigma_2^2$ at the 0.05 level of significance.

Solution

1. *Null hypothesis*: $\sigma_1^2 = \sigma_2^2$
   Alternative hypothesis: $\sigma_1^2 < \sigma_2^2$
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $F > 2.82$, the value of $F_{0.05}$ for 11 and 11 degrees of freedom, where

$$F = \frac{S_2^2}{S_1^2}$$

4. *Calculations*:

$$F = \frac{(0.062)^2}{(0.035)^2} = 3.14$$

5. *Decision*: Since $F = 3.14$ exceeds 2.82, the null hypothesis must be rejected; in other words, the data support the contention that the plating done by Company 1 is less variable than that done by Company 2. ■

EXAMPLE **A two-sided test for the equality of two variances**

With reference to the example dealing with the heat-producing capacity of coal from two mines on page 253, use the 0.02 level of significance to test whether it is reasonable to assume that the variances of the two populations sampled are equal.

Solution 1. *Null hypothesis*: $\sigma_1^2 = \sigma_2^2$
Alternative hypothesis: $\sigma_1^2 \neq \sigma_2^2$

2. *Level of significance*: $\alpha = 0.02$

3. *Criterion*: Reject the null hypothesis if $F > 11.4$, the value of $F_{0.01}$ for 4 and 5 degrees of freedom, where

$$F = \frac{S_1^2}{S_2^2}$$

since $s_1^2 = 15{,}750$ is greater than $s_2^2 = 10{,}920$.

4. *Calculations*:

$$F = \frac{15{,}750}{10{,}920} = 1.44$$

5. *Decision*: Since $F = 1.44$ does not exceed 11.4, the null hypothesis cannot be rejected; there is no real statistical reason to doubt the equality of the variances of the two populations. ■

To obtain confidence intervals for the ratio of variances, we need a slightly more general sampling distribution. From the single sample result,

$$\frac{(n_1 - 1)\,S_1^2}{\sigma_1^2} \quad \text{has a chi square distribution with} \quad n_1 - 1 \quad \text{degrees of freedom}$$

$$\frac{(n_2 - 1)\,S_2^2}{\sigma_2^2} \quad \text{has a chi square distribution with} \quad n_2 - 1 \quad \text{degrees of freedom}$$

and these two random quantities are independent since the samples on which they are based are independent. Consequently,

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\text{chi square/degrees of freedom}}{\text{chi square/degrees of freedom}} = F$$

where the right-hand side has an F distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom according to the representation on page 193.

Then, prior to sampling, we can assert with probability $1 - \alpha$ that

$$F_{1-\alpha/2}(n_1 - 1, n_2 - 1) < \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} < F_{\alpha/2}(n_1 - 1, n_2 - 1)$$

Solving this inequality for σ_2^2/σ_1^2, we obtain a confidence interval in terms of the observed s_1^2 and s_2^2.

Confidence interval for σ_2^2/σ_1^2 normal populations

$$F_{1-\alpha/2}(n_1 - 1, n_2 - 1)\frac{s_2^2}{s_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < F_{\alpha/2}(n_1 - 1, n_2 - 1)\frac{s_2^2}{s_1^2}$$

EXAMPLE **Confidence interval for ratio of variances of yield**

Refer to the example on page 256 of making green gasoline from sucrose. The equal sample sizes are $n_1 = n_2 = 9$, $s_1^2 = 0.4548$, and $s_2^2 = 0.1089$. Obtain a 98% confidence interval for σ_2^2/σ_1^2.

Solution Since the degrees of freedom for the F are $(n_1 - 1, n_2 - 1) = (8, 8)$ and $\alpha/2 = 0.01$, we find $F_{0.01} = 6.03$ and $F_{0.99} = 1/F_{0.01} = 1/6.03$. The 98% confidence interval for σ_2^2/σ_1^2 becomes

$$\left(\frac{1}{6.03} \frac{0.1089}{0.4548}, 6.03 \frac{0.1089}{0.4548} \right) \quad \text{or} \quad (0.04, 1.44)$$

The wideness of the interval illustrates the large amount of variability in variances when sample sizes are small. The second variance σ_2^2 could be as small as one-twenty-fifth of σ_1^2 or it could be larger than σ_1^2. ■

Statistical software is readily available to obtain other values of $F_\alpha(\nu_1, \nu_2)$. See Exercise 9.23 for the *MINITAB* and Appendix C for the *R* commands.

Caution

In marked contrast to the procedures for making inferences about μ, the validity of the procedures in this chapter depends rather strongly on the assumption that the underlying population is normal. The sampling variance of S^2 can change when the population departs from normality by having, for instance, a single long tail. It can be shown that, when the underlying population is normal, the sampling variance of S^2 is $2\sigma^4/(n-1)$. However, for nonnormal distributions, the sampling variance of S^2 depends not only on σ^2 but also on the population third and fourth moments, μ_3 and μ_4 (see page 94). Consequently, it could be much larger than $2\sigma^4/(n-1)$. This behavior completely invalidates any tests of hypothesis or confidence intervals for σ^2. We say that these procedures for making inferences about σ^2 are not **robust** with respect to deviations from normality.

Exercises

9.7 With reference to Exercise 7.62, test the null hypothesis $\sigma = 600$ psi for the compressive strength of the given kind of steel against the alternative hypothesis $\sigma > 600$ psi. Use the 0.05 level of significance.

9.8 If 12 determinations of the specific heat of iron have a standard deviation of 0.0086, test the null hypothesis that $\sigma = 0.010$ for such determinations. Use the alternative hypothesis $\sigma \neq 0.010$ and the level of significance $\alpha = 0.01$.

9.9 With reference to Exercise 8.5, test the null hypothesis that $\sigma = 15.0$ minutes for the time that is required for repairs of the first kind of photocopying equipment against the alternative hypothesis that $\sigma > 15.0$ minutes. Use the 0.05 level of significance and assume normality.

9.10 Use the 0.01 level of significance to test the null hypothesis that $\sigma = 0.015$ inch for the diameters of certain bolts against the alternative hypothesis that $\sigma \neq 0.015$ inch, given that a random sample of size 15 yielded $s^2 = 0.00011$.

9.11 Playing 10 rounds of golf on his home course, a golf professional averaged 71.3 with a standard deviation of 1.32. Test the null hypothesis that the consistency of his game on his home course is actually measured by $\sigma = 1.20$, against the alternative hypothesis that he is less consistent. Use the level of significance $\alpha = 0.05$.

9.12 The security department of a large office building wants to test the null hypothesis that $\sigma = 2.0$ minutes for the time it takes a guard to walk his round against the alternative hypothesis that $\sigma \neq 2.0$ minutes. What can it conclude at the 0.01 level of significance if a random sample of size $n = 31$ yields $s = 1.8$ minutes?

9.13 Help justify the use of the two sample t test in Exercise 8.9 by testing the null hypothesis that the two

populations have equal variances. Use the 0.02 level of significance.

9.14 With reference to Exercise 8.10, use the 0.02 level of significance to test the assumption that the two populations have equal variances.

9.15 Two different lighting techniques are compared by measuring the intensity of light at selected locations in areas lighted by the two methods. If 15 measurements in the first area had a standard deviation of 2.7 foot-candles and 21 measurements in the second area had a standard deviation of 4.2 foot-candles, can it be concluded that the lighting in the second area is less uniform? Use a 0.01 level of significance. What assumptions must be made as to how the two samples are obtained?

9.16 With reference to Exercise 8.6, where we had $n_1 = 40$, $n_2 = 30$, $s_1 = 15.2$, and $s_2 = 18.7$, use the 0.05 level of significance to test the claim that there is a greater variability in the number of cars which make left turns approaching from the south between 4 P.M. and 6 P.M. at the second intersection. Assume the distributions are normal.

Do's and Don'ts

Do's

1. Before applying the procedures in this chapter, always plot the data to look for outliers or presence of a long tail. Lack of normality can seriously affect tests of hypotheses and confidence intervals for variances.

Don'ts

1. Don't routinely calculate confidence intervals for variances or standard deviations using the formulas in this chapter. The confidence levels can deviate substantially from their specified level, say 95%, because of nonnormality.

Review Exercises

9.17 With reference to the example on page 233, construct a 95% confidence interval for the true standard deviation of the breaking strength of the given kind of ribbon.

9.18 If 31 measurements of the boiling point of sulfur have a standard deviation of 0.83 degree Celsius, construct a 98% confidence interval for the true standard deviation of such measurements. What assumption did you make about the population?

9.19 Past data indicate that the variance of measurements made on sheet metal stampings by experienced quality-control inspectors is 0.18 $(\text{inch})^2$. Such measurements made by an inexperienced inspector could have too large a variance (perhaps because of inability to read instruments properly) or too small a variance (perhaps because unusually high or low measurements are discarded). If a new inspector measures 101 stampings with a variance of 0.13 $(\text{inch})^2$, test at the 0.05 level of significance whether the inspector is making satisfactory measurements. Assume normality.

9.20 Pull-strength tests on 10 soldered leads for a semiconductor device yield the following results in pounds-force required to rupture the bond:

15.8 12.7 13.2 16.9 10.6 18.8 11.1 14.3 17.0 12.5

Another set of 8 leads was tested after encapsulation to determine whether the pull strength has been increased by encapsulation of the device, with the following results:

24.9 23.6 19.8 22.1 20.4 21.6 21.8 22.5

Use the 0.02 level of significance to test whether it is reasonable to assume that the two samples come from populations with equal variances.

9.21 With reference to the example on page 254, test the equality of the variances for the two aluminum alloys. Use the 0.02 level of significance.

9.22 With reference to the example on page 254, find a 98% confidence interval for the ratio of variances of the two aluminum alloys.

9.23 ***MINITAB* calculation of t_α, χ^2_ν, and F_α**

The software finds percentiles, so to obtain F_α, we first convert from α to $1-\alpha$. We illustrate with the calculation of $F_{0.025}(4, 7)$, where $1 - 0.025 = 0.975$.

> **Dialog box:**
>
> **Calc > Probability distributions > F**. Choose **Inverse cumulative probability**. *Type* 4 in **Numerator degrees of freedom**, 7 in **Denominator degrees of freedom**. Choose **Input constant** and *type* .975. Click **OK**.

Output:

F distribution with 4 DF in numerator and 7 DF in denominator

P(X <= x)	x
0.975	5.52259

In the first line, you may instead select **Chi square** or **t** and then there is only one **Degrees of Freedom** in the second line.

Obtain $F_{0.975}(7, 4)$ and check that it equals $1/F_{0.025}(4, 7) = 1/5.52259$.

9.24 A bioengineering company manufactures a device for externally measuring blood flow. Measurements of the electrical output (milliwatts) on a sample of 16 units yields the data

11	1	5	3	2	23	37	5
18	7	1	11	2	2	30	3

plotted in Figure 9.1.

(a) Should you report the 95% confidence interval for σ using the formula in this Chapter? Explain.

(b) What is your answer to part (a) if you first take natural logarithms and then calculate the confidence interval for the variance of ln (output)?

(c) Does your conclusion in part (b) readily imply anything about variance on the original scale? Explain.

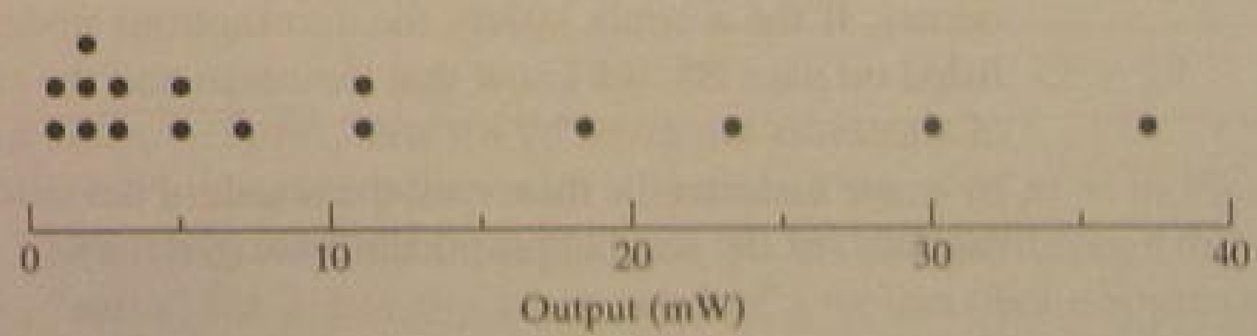


Figure 9.1
Output in milliwatts

Key Terms

Robust 275

Sample range 269

CHAPTER

10 Inferences Concerning Proportions

CHAPTER OUTLINE

10.1 Estimation of Proportions 278
10.2 Hypotheses Concerning One Proportion 283
10.3 Hypotheses Concerning Several Proportions 285
10.4 Analysis of $r \times c$ Tables 292
10.5 Goodness of Fit 295
Review Exercises 299
Key Terms 300

Many engineering problems deal with proportions, percentages, or probabilities. In acceptance sampling we are concerned with the proportion of defectives in a lot, and in life testing we are concerned with the percentage of certain components which will perform satisfactorily during a stated period of time, or the probability that a given component will last at least a given number of hours. It should be clear from these examples that problems concerning proportions, percentages, or probabilities are really equivalent; a percentage is merely a proportion multiplied by 100, and a probability may be interpreted as a proportion in a long series of trials.

Section 10.1 deals with the estimation of proportions; Section 10.2 deals with tests concerning proportions; Section 10.3 deals with tests concerning two or more proportions; in Section 10.4 we shall learn how to analyze data tallied into a two-way classification; and in Section 10.5 we shall learn how to judge whether differences between an observed frequency distribution and corresponding expectations can be attributed to chance.

10.1 Estimation of Proportions

The information that is usually available for the estimation of a proportion is the number of times, X, that an appropriate event occurs in n trials, occasions, or observations. The point estimator of the population proportion, itself, is usually the **sample proportion** $\frac{X}{n}$, namely, the proportion of the time that the event actually occurs. If the n trials satisfy the assumptions underlying the binomial distribution listed on page 85, we know that the mean and the standard deviation of the number of successes are given by np and $\sqrt{np(1-p)}$. If we divide both of these quantities by n, we find that the mean and the standard deviation of the proportion of successes (namely, of the sample proportion) are given by

$$\frac{np}{n} = p \quad \text{and} \quad \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$$

The first of these results shows that the sample proportion is an unbiased estimator of the binomial parameter p, namely, of the true proportion we are trying to estimate on the basis of a sample.

In the construction of confidence intervals for the binomial parameter p, we meet several obstacles. First, since x and $\frac{x}{n}$ are values of discrete random variables, it may be impossible to get an interval for which the degree of confidence is exactly $(1-\alpha)100\%$. Second, the standard deviation of the sampling distribution of the number of successes, as well as that of the proportion of successes, involves the parameter p that we are trying to estimate.

To construct a confidence interval for p having approximately the degree of confidence $(1-\alpha)100\%$, we first determine for a given set of values of p the corresponding quantities x_0 and x_1, where x_0 is the largest integer for which the binomial probabilities $b(k;n,p) = P[X=k]$ satisfy

$$\sum_{k=0}^{x_0} b(k;n,p) \leq \frac{\alpha}{2}$$

while x_1 is the smallest integer for which

$$\sum_{k=x_1}^{n} b(k;n,p) \leq \frac{\alpha}{2}$$

To emphasize the point that x_0 and x_1 depend on the value of p, we shall write these quantities as $x_0(p)$ and $x_1(p)$. Thus, we can assert with a probability of approximately $1-\alpha$, and at least $1-\alpha$, that the inequality

$$x_0(p) < x < x_1(p)$$

will be satisfied; here x is a value of a random variable and p is a fixed constant. To change inequalities like these into confidence intervals for p, we can use a simple graphical method illustrated by the following example: Suppose, for instance, that we want to find approximate 95% confidence intervals for p for samples of size $n = 20$. Using Table 1 at the end of the book, we first determine x_0 and x_1 for selected values of p such that x_0 is the largest integer for which

$$B(x_0; 20, p) \leq 0.025$$

while x_1 is the smallest integer for which

$$1 - B(x_1 - 1; 20, p) \leq 0.025$$

Letting p equal $0.1, 0.2, \ldots,$ and 0.9, we thus obtain the values shown in the following table:

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
x_0	–	0	1	3	5	7	9	11	14
x_1	6	9	11	13	15	17	19	20	–

Plotting the points with coordinates p and $x(p)$ as in Figure 10.1, and drawing smooth curves, one through the x_0 points and one through the x_1 points, we can now "solve" for p. For any given value of x we can obtain approximate 95% confidence limits for p by going horizontally to the two curves and marking off the corresponding values of p. (See Figure 10.1.) Thus, for $x = 4$ we obtain the approximate 95% confidence interval

$$0.06 < p < 0.45$$

Computer software can do this calculation quickly and more accurately. In Exercise 10.13, we obtain the confidence interval $(0.06, 0.44)$.

On page 133 we gave the general rule of thumb that the normal distribution provides a good approximation to the binomial distribution when np and $n(1-p)$ are both greater than 15. Thus, for $n = 50$ the normal curve approximation may be used if it can be assumed that p lies between 0.30 and 0.70; for $n = 200$ it may be used if it can be assumed that p lies between 0.075 and 0.925; and so forth. This is what we shall mean here, and later in this chapter, by "n being large."

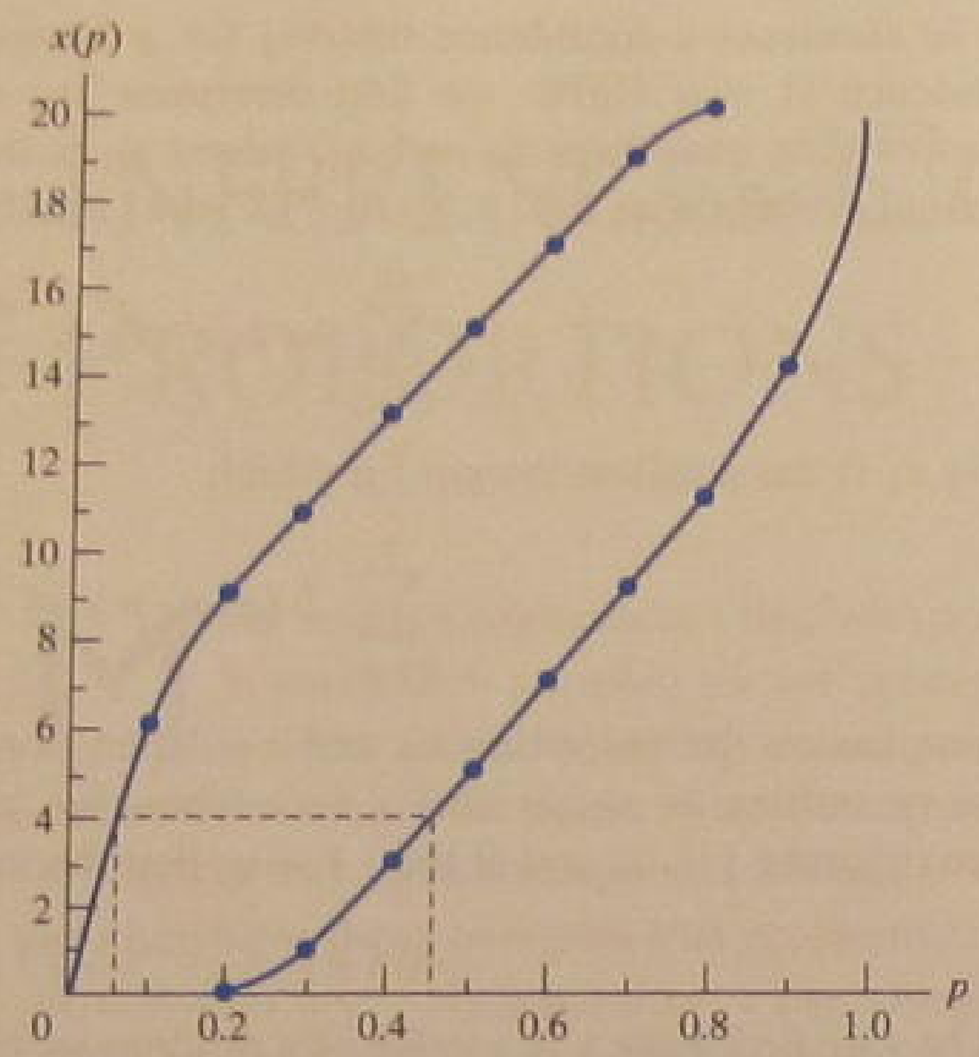


Figure 10.1
95% confidence intervals for proportions ($n = 20$)

When n is large, we can construct approximate confidence intervals for the binomial parameter p by using the normal approximation to the binomial distribution. Accordingly, we can assert with probability $1 - \alpha$ that the inequality

$$-z_{\alpha/2} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\alpha/2}$$

will be satisfied. Solving this quadratic inequality for p, we can obtain a corresponding set of approximate confidence limits for p in terms of the observed value x (see Exercise 10.15), but since the necessary calculations are complex, we shall make the further approximation of substituting $\frac{x}{n}$ for p in $\sqrt{np(1-p)}$. This yields

Large sample confidence interval for p

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

where the degree of confidence is $(1 - \alpha)100\%$.

EXAMPLE **A large sample 95% confidence interval for p**

If $x = 36$ of $n = 100$ persons interviewed are familiar with the tax incentives for installing certain energy-saving devices, construct a 95% confidence interval for the corresponding true proportion.

Solution Substituting

$$\frac{x}{n} = \frac{36}{100} = 0.36$$

and $z_{\alpha/2} = 1.96$ into the above formula, we get

$$0.36 - 1.96\sqrt{\frac{(0.36)(0.64)}{100}} < p < 0.36 + 1.96\sqrt{\frac{(0.36)(0.64)}{100}}$$

or

$$0.266 < p < 0.454$$

We are 95% confident that the population proportion of persons familiar with the tax incentives, p, is contained in the interval from 0.266 to 0.454. Note that if we had used the computer calculation in Exercise 10.13, we would have obtained

$$0.27 < p < 0.46$$ ■

The magnitude of the error we make when we use $\frac{X}{n}$ as an estimator of p is given by

$$\left| \frac{X}{n} - p \right|$$

Again using the normal approximation, we can thus assert with probability $1 - \alpha$ that the inequality

$$\left| \frac{X}{n} - p \right| \leq z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

will be satisfied, namely, that the error will be at most

$$z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

Maximum error of estimate

$$E = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

With the observed value $\frac{x}{n}$ substituted for p, we obtain an estimate of E.

EXAMPLE **An estimate of the maximum error**

In a sample survey conducted in a large city, 136 of 400 persons answered yes to the question of whether their city's public transportation is adequate. With 99% confidence, what can we say about the maximum error, if

$$\frac{x}{n} = \frac{136}{400} = 0.34$$

is used as an estimate of the corresponding true proportion?

Solution Substituting $\frac{x}{n} = 0.34$ and $z_{\alpha/2} = 2.575$ into the above formula, we find that the error is at most

$$E = 2.575 \sqrt{\frac{(0.34)(0.66)}{400}} = 0.061$$ ■

The preceding formula for E can also be used to determine the sample size that is needed to attain a desired degree of precision. Solving for n, we get

Sample size determination

$$n = p(1-p)\left[\frac{z_{\alpha/2}}{E}\right]^2$$

but this formula cannot be used as it stands unless we have some information about the possible size of p (on the basis of collateral data, say a pilot sample). If no such information is available, we can make use of the fact that $p(1-p)$ is at most $\frac{1}{4}$, corresponding to $p = \frac{1}{2}$, as can be shown by the methods of elementary calculus. If a range for p is known, the value closest to $\frac{1}{2}$ should be used.

Thus, if

Sample size (p unknown)

$$n = \frac{1}{4}\left[\frac{z_{\alpha/2}}{E}\right]^2$$

we can assert with a probability of at least $1-\alpha$ that the error in using $\frac{X}{n}$ as an estimate of p will not exceed E; once the data have been obtained, we will be able to assert with at least $(1-\alpha)100\%$ confidence that the error does not exceed E.

EXAMPLE **Selecting a sample size for estimating a proportion**

Suppose that we want to estimate the true proportion of defectives in a very large shipment of adobe bricks, and that we want to be at least 95% confident that the error is at most 0.04. How large a sample will we need if

(a) we have no idea what the true proportion might be;

(b) we know that the true proportion does not exceed 0.12?

Solution **(a)** Using the second of the two formulas for the sample size, we get

$$n = \frac{1}{4}\left[\frac{1.96}{0.04}\right]^2 = 600.25$$

or $n = 601$ rounded up to the nearest integer.

(b) Using the first of the two formulas for the sample size with $p = 0.12$ (the possible value closest to $p = \frac{1}{2}$), we get

$$n = (0.12)(0.88)\left[\frac{1.96}{0.04}\right]^2 = 253.55$$

or $n = 254$ rounded up to the nearest integer. This serves to illustrate how some collateral information about the possible size of p can substantially reduce the size of the required sample. ■

Exercises

10.1 In a random sample of 200 claims filed against an insurance company writing collision insurance on cars, 84 exceeded \$3,500. Construct a 95% confidence interval for the true proportion of claims filed against this insurance company that exceed \$3,500, using the large sample confidence interval formula.

10.2 With reference to Exercise 10.1, what can we say with 99% confidence about the maximum error if we use the sample proportion as an estimate of the true proportion of claims filed against this insurance company that exceed \$3,500?

10.3 In a random sample of 400 industrial accidents, it was found that 231 were due at least partially to unsafe working conditions. Construct a 99% confidence interval for the corresponding true proportion using the large sample confidence interval formula.

10.4 With reference to Exercise 10.3, what can we say with 95% confidence about the maximum error if we use the sample proportion to estimate the corresponding true proportion?

10.5 In a random sample of 90 sections of pipe in a chemical plant, 15 showed signs of serious corrosion. Construct a 95% confidence interval for the true proportion of pipe sections showing signs of serious corrosion, using the large sample confidence interval formula.

10.6 In a recent study, 69 of 120 meteorites were observed to enter the earth's atmosphere with a velocity of less than 26 miles per second. If we estimate the corresponding true proportion as $\frac{69}{120} = 0.575$, what can we say with 95% confidence about the maximum error?

10.7 Among 100 fish caught in a large lake, 18 were inedible due to the pollution of the environment. If we use $\frac{18}{100} = 0.18$ as an estimate of the corresponding true proportion, with what confidence can we assert that the error of this estimate is at most 0.065?

10.8 A random sample of 300 shoppers at a supermarket includes 204 who regularly use cents-off coupons. Construct a 98% confidence interval for the probability that any one shopper at the supermarket, selected at random, will regularly use cents-off coupons.

10.9 What is the size of the smallest sample required to estimate an unknown proportion of customers who would pay for an additional service, to within a maximum error of 0.06 with at least 95% confidence?

10.10 With reference to Exercise 10.9, how would the required sample size be affected if it is known that the proportion to be estimated is at least 0.75?

10.11 Suppose that we want to estimate what percentage of all drivers exceed the 55 mph speed limit on a certain stretch of road. How large a sample will we need to be at least 99% confident that the error of our estimate, the sample percentage, is at most 3.5%?

10.12 With reference to Exercise 10.11, how would the required sample size be affected if it is known that the percentage to be estimated is at most 40%?

10.13 ***MINITAB* determination of confidence interval for *p***
When the sample size is not large, the confidence interval for a proportion p can be obtained using the following commands. We illustrate the case $n = 20$ and $x = 4$.

> **Dialog box:**
>
> **Stat > Basic Statistics > 1-Proportion**. Choose **Summarized outcomes**.
> Type 20 in **Number of trials**, 4 **in Number of events**.
> Click **Options** and type 95.0 in **Confidence level**. Click **OK**. Click **OK**.

The partial output includes the 95% confidence interval (0.057334, 0.436614).

Obtain the 95% confidence interval when $n = 20$ and $x = 16$.

10.14 Use Exercise 10.13 or other software to obtain the interval requested in Exercise 10.3.

10.15 Show that the inequality on page 280 leads to the following $(1 - \alpha)100\%$ confidence limits:

$$\frac{x + \frac{1}{2} z_{\alpha/2}^2 \pm z_{\alpha/2} \sqrt{\frac{x(n-x)}{n} + \frac{1}{4} z_{\alpha/2}^2}}{n + z_{\alpha/2}^2}$$

10.16 Use the formula of Exercise 10.15 to rework Exercise 10.3.

10.17 A major automobile company was having trouble with the quality of paint on its gray cars. The first step was to collect data on the magnitude of the problem. Of 4,063 recently painted gray automobiles, 533 had paint problems that could easily be detected by visual inspection. Obtain a 95% confidence interval for the population proportion of defective gray paint jobs.

10.18 An international corporation needed several millions of words, from thousands of documents and manuals, translated. The work was contracted to a company that used computer-assisted translation, along with some human checks. The corporation conducted its own quality check by sampling the translation. Among the 2,037 mistakes found, 834 were an incorrect word. Obtain a 99% confidence interval for the population proportion of mistakes that are incorrect words.

10.2 Hypotheses Concerning One Proportion

Many of the methods used in sampling inspection, quality control, and reliability verification are based on tests of the null hypothesis that a proportion (percentage, or probability) equals some specified constant.

Although there are exact tests based on the binomial distribution that can be performed with the use of Table 1, we shall consider here only approximate large sample tests based on the normal approximation to the binomial distribution. In

other words, we shall test the null hypothesis $p = p_0$ against one of the alternatives $p < p_0$, $p > p_0$, or $p \neq p_0$ with the use of the statistic

Statistic for large sample test concerning p

$$Z = \frac{X - n\,p_0}{\sqrt{n\,p_0\,(1 - p_0)}}$$

which is a random variable having approximately the standard normal distribution.[1] The critical regions are like those shown in the table on page 232 with p and p_0 substituted for μ and μ_0.

Critical Regions for Testing $p = p_0$ (large sample)

Alternative hypothesis	Reject null hypothesis if:
$p < p_0$	$Z < -z_\alpha$
$p > p_0$	$Z > z_\alpha$
$p \neq p_0$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

EXAMPLE **A one-sided test of the proportion of transceivers**

Transceivers provide wireless communication among electronic components of consumer products. Responding to a need for a fast, low-cost test of Bluetooth-capable transceivers, engineers[2] developed a product test at the wafer level. In one set of trials with 60 devices selected from different wafer lots, 48 devices passed. Test the null hypothesis $p = 0.70$ against the alternative hypothesis $p > 0.70$ at the 0.95 level of significance.

Solution

1. *Null hypothesis*: $p = 0.70$
   Alternative hypothesis: $p > 0.70$
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $Z > 1.645$, where

$$Z = \frac{X - n\,p_0}{\sqrt{n\,p_0\,(1 - p_0)}}$$

4. *Calculations*: Substituting $x = 48$, $n = 60$, and $p_0 = 0.70$ into the formula above, we get

$$z = \frac{48 - 60\,(0.70)}{\sqrt{60\,(0.70)\,(0.30)}} = 1.69$$

5. *Decision*: Since $z = 1.69$ is greater than 1.645, we reject the null hypothesis at level 0.05. In other words, there is sufficient evidence to conclude that the proportion of good transceivers that would be produced is greater than 0.70. The P-value, $P(Z > 1.69) = .0455$, somewhat strengthens this conclusion. ■

[1] Some authors write the numerator of this formula for Z as $X \pm \frac{1}{2} - np_0$, whichever is numerically smaller, but there is generally no need for this continuity correction so long as n is large.

[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* (2008), 150–159.

10.3 Hypotheses Concerning Several Proportions

When we compare the consumer response (percentage favorable and percentage unfavorable) to two different products, when we decide whether the proportion of defectives of a given process remains constant from day to day, when we judge whether there is a difference in political persuasion among several nationality groups, and in many similar situations, we are interested in testing whether two or more binomial populations have the same parameter p. Referring to these parameters as $p_1, p_2, \ldots,$ and p_k, we are, in fact, interested in testing the null hypothesis

$$p_1 = p_2 = \cdots = p_k = p$$

against the alternative hypothesis that these population proportions are not all equal. To perform a suitable large sample test of this hypothesis, we require independent random samples of size $n_1, n_2, \ldots,$ and n_k from the k populations; then, if the corresponding numbers of "successes" are $X_1, X_2, \ldots,$ and X_k, the test we shall use is based on the fact that

1. for large samples the sampling distribution of

$$Z_i = \frac{X_i - n_i p_i}{\sqrt{n_i p_i (1 - p_i)}}$$

is approximately the standard normal distribution,

2. the square of random variable having the standard normal distribution is a random variable having the chi square distribution with 1 degree of freedom, and
3. the sum of k independent random variables having chi square distributions with 1 degree of freedom is a random variable having the chi square distribution with k degrees of freedom. (See Section 6.5, and the examples on pages 194 and 198 for proofs of these last two results.) Thus,

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i p_i)^2}{n_i p_i (1 - p_i)}$$

is a value of a random variable having approximately the chi square distribution with k degrees of freedom. In practice we substitute for the p_i, which under the null hypothesis are all equal, the pooled estimate

$$\hat{p} = \frac{x_1 + x_2 + \cdots + x_k}{n_1 + n_2 + \cdots + n_k}$$

Since the null hypothesis should be rejected if the differences between the x_i and the $n_i \hat{p}$ are large, the critical region is $\chi^2 > \chi^2_\alpha$, where χ^2_α is as defined on page 189 and the number of degrees of freedom is $k - 1$. The loss of one degree of freedom results from substituting for p the estimate $\hat{p}$.

In actual practice, when we compare two or more sample proportions, it is convenient to determine the value of the χ^2 statistic by looking at the data as arranged in the following way:

	Sample 1	*Sample 2*	$\cdots$	*Sample k*	*Total*
Successes	x_1	x_2	$\cdots$	x_k	x
Failures	$n_1 - x_1$	$n_2 - x_2$	$\cdots$	$n_k - x_k$	$n - x$
Total	n_1	n_2	$\cdots$	n_k	n

The notation is the same as before, except for x and n, which represent, respectively, the total number of successes and the total number of trials for all samples combined. With reference to this table, the entry in the cell belonging to the ith row and jth column is called the **observed cell frequency** o_{ij} with $i = 1, 2$ and $j = 1, 2, \ldots, k$.

Under the null hypothesis $p_1 = p_2 = \cdots = p_k = p$, we estimate p, as before, as the total number of successes divided by the total number of trials, which we now write as $\hat{p} = \frac{x}{n}$. Hence, the expected number of successes and failures for the jth sample are estimated by

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}$$

and

$$e_{2j} = n_j(1 - \hat{p}) = \frac{n_j(n - x)}{n}$$

The quantities e_{1j} and e_{2j} are called the **expected cell frequencies** for $j = 1, 2, \ldots, k$. Note that **the expected frequency for any given cell may be obtained by multiplying the totals of the column and the row to which it belongs and then dividing by the grand total** n.

A **chi square test** is based on the χ^2 statistic on page 285, with $\hat{p}$ substituted for the p_i. The χ^2 statistic can be written in the form

χ^2 statistic for test concerning difference among proportions

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

as the reader will be asked to verify in Exercise 10.36. This formula has the advantage that it can easily be extended to the more general case, to be treated in Section 10.4, where each trial permits more than two possible outcomes, and there are, thus, more than two rows in the tabular presentation of the various frequencies.

EXAMPLE **Testing the equality of three proportions using the χ^2 statistic**

Samples of three kinds of materials, subjected to extreme temperature changes, produced the results shown in the following table:

	Material A	*Material B*	*Material C*	*Total*
Crumbled	41	27	22	90
Remained intact	79	53	78	210
Total	120	80	100	300

Use the 0.05 level of significance to test whether, under the stated conditions, the probability of crumbling is the same for the three kinds of materials.

Solution

1. *Null hypothesis*: $p_1 = p_2 = p_3$
   Alternative hypothesis: p_1, p_2, and p_3 are not all equal.
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $\chi^2 > 5.991$, the value of $\chi^2_{0.05}$ for $3 - 1 = 2$ degrees of freedom, where χ^2 is given by the formula above.

4. *Calculations*: The expected frequencies for the first two cells of the first row are

$$e_{11} = \frac{90 \cdot 120}{300} = 36 \quad \text{and} \quad e_{12} = \frac{90 \cdot 80}{300} = 24$$

and, as it can be shown that **the sum of the expected frequencies for any row or column equals that of the corresponding observed frequencies** (see Exercise 10.37), we find by subtraction that $e_{13} = 90 - (36 + 24) = 30$, and that the expected frequencies for the second row are $e_{21} = 120 - 36 = 84$, $e_{22} = 80 - 24 = 56$, and $e_{23} = 100 - 30 = 70$. Then, substituting these values together with the observed frequencies into the formula for χ^2, we get

$$\begin{aligned}\chi^2 &= \frac{(41-36)^2}{36} + \frac{(27-24)^2}{24} + \frac{(22-30)^2}{30} \\ &\quad + \frac{(79-84)^2}{84} + \frac{(53-56)^2}{56} + \frac{(78-70)^2}{70} \\ &= 4.575\end{aligned}$$

5. *Decision*: Since $\chi^2 = 4.575$ does not exceed 5.991, the null hypothesis cannot be rejected; in other words, the data do not refute the hypothesis that, under the stated conditions, the probability of crumbling is the same for the three kinds of material. ■

Most of the entries of Table 5 are given to three decimal places, but, since rounding errors tend to average out, there is seldom a need to give more than two decimal places in the final value of the χ^2 statistic. The test we have been discussing here is only an approximate test since the sampling distribution of the χ^2 statistic is only approximately the chi square distribution, and it should not be used when one or more of the expected frequencies is less than 5. If this is the case, we can sometimes combine two or more of the samples in such a way that none of the e's is less than 5.

If the null hypothesis of equal proportions is rejected, it is a good practice to graph the confidence intervals (see page 280) for the individual proportions p_i. The graph helps illuminate differences between the proportions.

EXAMPLE **Graphical display of confidence intervals**

Four methods are under development for making disks of a superconducting material. Fifty disks are made by each method and they are checked for superconductivity when cooled with liquid nitrogen.

	Method 1	*Method 2*	*Method 3*	*Method 4*	*Total*
Superconductors	31	42	22	25	120
Failures	19	8	28	25	80
Total	50	50	50	50	200

Perform a chi square test with $\alpha = 0.05$. If there is a significant difference between the proportions of superconductors produced, plot the individual confidence intervals.

Solution

1. *Null hypothesis*: $p_1 = p_2 = p_3 = p_4$
   Alternative hypothesis: p_1, p_2, p_3, and p_4 are not all equal.
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $\chi^2 > 7.815$, the value of $\chi^2_{0.05}$ for $4 - 1 = 3$ degrees of freedom.

4. *Calculations*: Each cell in the first row has expected frequency

$$120 \cdot \frac{50}{200} = 30$$

and each cell in the second row has expected frequency

$$80 \cdot \frac{50}{200} = 20$$

The chi square statistic is

$$\begin{aligned}\chi^2 &= \frac{1}{30} + \frac{144}{30} + \frac{64}{30} + \frac{25}{30} \\ &\quad + \frac{1}{20} + \frac{144}{20} + \frac{64}{20} + \frac{25}{20} \\ &= 19.50\end{aligned}$$

5. *Decision*: Since 19.50 greatly exceeds 7.815, we reject the null hypothesis of equal proportions at the 5% level of significance.
The confidence intervals obtained from the large sample formula on page 280 are

$$0.62 \pm 0.13, \quad 0.84 \pm 0.10, \quad 0.44 \pm 0.14, \quad 0.50 \pm 0.14.$$

These are plotted in Figure 10.2. Note how Method 2 stands out as being better.

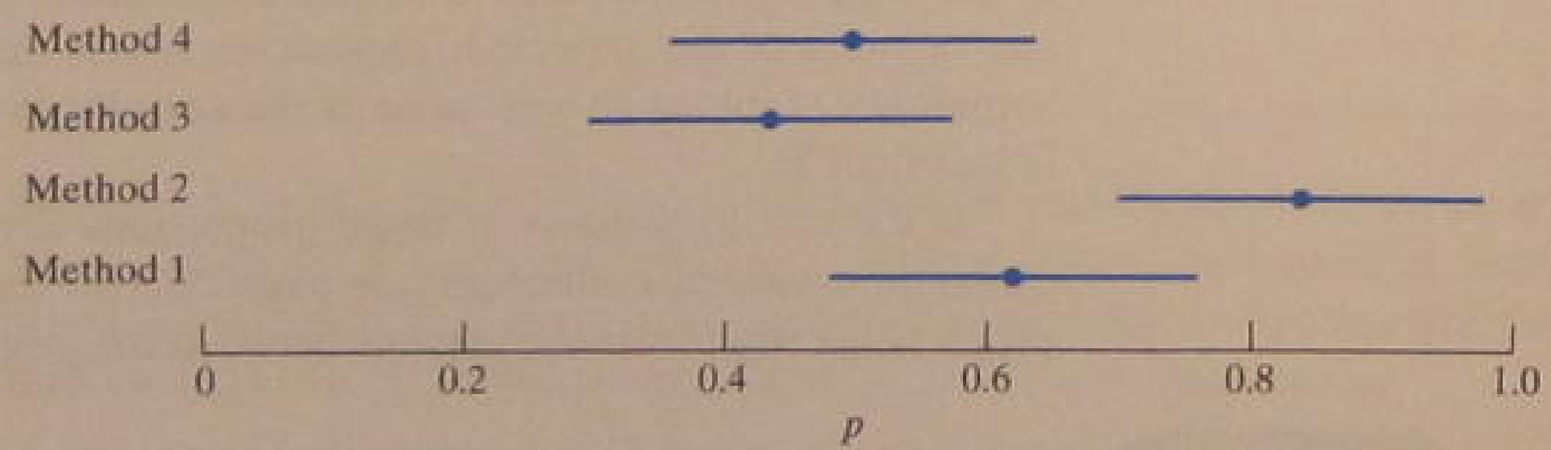


Figure 10.2
Confidence intervals for several proportions

Although there has been no mention of randomization in the development of the χ^2 statistic, wherever possible the experimental units should be randomly assigned to methods. In the example above, the disks could be numbered from 1 to 200 and random numbers selected from 1 to 200 without replacement. The disks corresponding to the first fifty numbers drawn would be assigned to method 1, and so on. This will prevent uncontrolled sources of variation from systematically influencing the test concerning the four methods.

So far, the alternative hypothesis has been that $p_1, p_2, \ldots,$ and p_k are not all equal, and for $k = 2$ this reduces to the alternative hypothesis $p_1 \neq p_2$. In problems where the alternative hypothesis may also be $p_1 < p_2$ or $p_1 > p_2$, we can base the test on the statistic

Statistic for test concerning difference between two proportions

$$Z = \frac{\dfrac{X_1}{n_1} - \dfrac{X_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \quad \text{with} \quad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

which, for large samples, is a random variable having approximately the standard normal distribution. The test based on this statistic is equivalent to the one based on the χ^2 statistic on page 286 with $k = 2$, in the sense that the square of this Z statistic actually equals the χ^2 statistic (see Exercise 10.38). The critical regions for this alternative test of the null hypothesis $p_1 = p_2$ are like those shown in the table on page 249 with p_1 and p_2 substituted for μ_1 and μ_2.

EXAMPLE **A large sample test concerning two proportions**

A study shows that 16 of 200 tractors produced on one assembly line required extensive adjustments before they could be shipped, while the same was true for 14 of 400 tractors produced on another assembly line. At the 0.01 level of significance, does this support the claim that the second production line does superior work?

Solution

1. *Null hypothesis*: $p_1 = p_2$
   Alternative hypothesis: $p_1 > p_2$
2. *Level of significance*: $\alpha = 0.01$
3. *Criterion*: Reject the null hypothesis if $Z > 2.33$, where Z is given by the formula on page 288.
4. *Calculations*: Substituting $x_1 = 16$, $n_1 = 200$, $x_2 = 14$, $n_2 = 400$, and

$$\hat{p} = \frac{16 + 14}{200 + 400} = 0.05$$

into the formula for Z, we get

$$Z = \frac{\dfrac{16}{200} - \dfrac{14}{400}}{\sqrt{(0.05)(0.95)\left(\dfrac{1}{200} + \dfrac{1}{400}\right)}} = 2.38$$

5. *Decision*: Since $Z = 2.38$ exceeds 2.33, the null hypothesis must be rejected; we conclude that the true proportion of tractors requiring extensive adjustments is greater for first assembly line than for the second.
   The P value is 0.0087 as shown in Figure 10.3. ■

The test we have described here applies to the null hypothesis $p_1 = p_2$, but it can easily be modified (see Exercise 10.34) so that it applies also to the null hypothesis $p_1 - p_2 = \delta_0$.

The statistic for testing $p_1 - p_2 = \delta_0$ leads to a confidence interval which provides the set of plausible values for $p_1 - p_2$.

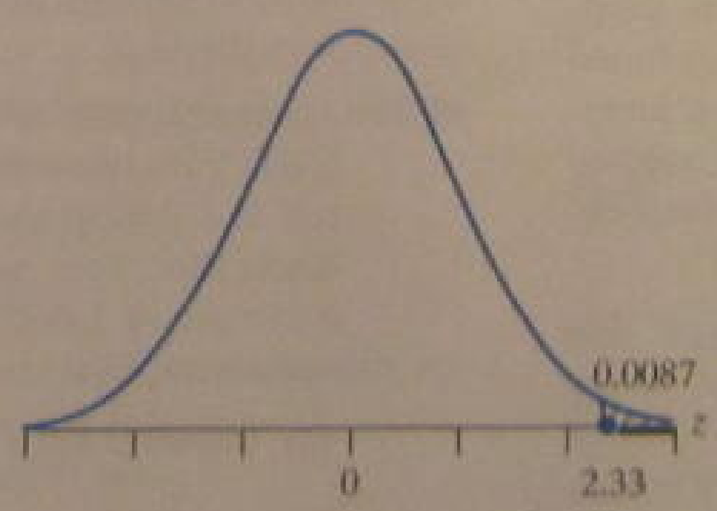


Figure 10.3
The P-value 0.0087

Large sample confidence interval for the difference of two proportions

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$$

EXAMPLE **A large sample confidence interval for the difference of two proportions**

With reference to the preceding example, find the large sample 95% confidence interval for $p_1 - p_2$.

Solution Since $\frac{x_1}{n_1} = \hat{p}_1 = \frac{16}{200} = 0.08$ and $\frac{x_2}{n_2} = \hat{p}_2 = \frac{14}{400} = 0.035$,

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$$

$$= 0.08 - 0.035 \pm 1.96\sqrt{\frac{(0.08)(0.92)}{200} + \frac{(0.035)(0.965)}{400}}$$

or $$0.003 < p_1 - p_2 < 0.087.$$

The first assembly line has a rate of extensive adjustment between 3 and 87 out of 1,000, higher than the rate for the second assembly line. ■

Exercises

10.19 A manufacturer of submersible pumps claims that at most 30% of the pumps require repairs within the first 5 years of operation. If a random sample of 120 of these pumps includes 47 which required repairs within the first 5 years, test the null hypothesis $p = 0.30$ against the alternative hypothesis $p > 0.30$ at the 0.05 level of significance.

10.20 The performance of a computer is observed over a period of 2 years to check the claim that the probability is 0.20 that its downtime will exceed 5 hours in any given week. Testing the null hypothesis $p = 0.20$ against the alternative hypothesis $p \neq 0.20$, what can we conclude at the level of significance $\alpha = 0.05$, if there were only 11 weeks in which the downtime of the computer exceeded 5 hours?

10.21 To check on an ambulance service's claim that at least 40% of its calls are life-threatening emergencies, a random sample was taken from its files, and it was found that only 49 of 150 calls were life-threatening emergencies. Can the null hypothesis $p = 0.40$ be rejected against the alternative hypothesis $p < 0.40$ if the probability of a Type I error is to be at most 0.01?

10.22 In a random sample of 600 cars making a right turn at a certain intersection, 157 pulled into the wrong lane. Test the null hypothesis that actually 30% of all drivers make this mistake at the given intersection, using the alternative hypothesis $p \neq 0.30$ and the level of significance

(a) $\alpha = 0.05$;

(b) $\alpha = 0.01$.

10.23 An airline claims that only 6% of all lost luggage is never found. If, in a random sample, 17 of 200 pieces of lost luggage are not found, test the null hypothesis $p = 0.06$ against the alternative hypothesis $p > 0.06$ at the 0.05 level of significance.

10.24 Suppose that 4 of 13 undergraduate engineering students state that they will go on to graduate school. Test the dean's claim that 60% of the undergraduate students will go on to graduate school, using the alternative hypothesis $p < 0.60$ and the level of significance $\alpha = 0.05$. [*Hint*: Use Table 1 to determine the probability of getting "at most 4 successes in 13 trials" when $p = 0.60$.]

10.25 A manufacturer of plasma high-definition televisions (HDTVs) claims that 95% of a high-end model will not fail during the one-year warrantee period. You doubt this claim and want to refute it on the basis of a sample of 1,000 sets where 937 did not fail under warrantee. First,

(a) Conduct a test of hypotheses using $\alpha = 0.05$.

(b) In light of your decision in part (a), what error could you have made? Explain in the context of this exercise.

10.26 Refer to Exercise 10.25. Suppose a sample of 450 sets-of an entry-level model plasma HDTV yielded 380 sets that did not fail under warrantee. Obtain a 98% confidence interval for the difference in proportions.

10.27 Tests are made on the proportion of defective castings produced by 5 different molds. If there were 14 defectives among 100 castings made with Mold I, 33 defectives among 200 castings made with Mold II, 21 defectives among 180 castings made with Mold III, 17 defectives among 120 castings made with Mold IV, and 25 defectives among 150 castings made with Mold V, use the 0.01 level of significance to test whether the true proportion of defectives is the same for each mold.

10.28 A study showed that 64 of 180 persons who saw a photocopying machine advertised during the telecast of a baseball game and 75 of 180 other persons who saw it advertised on a variety show remembered the brand name 2 hours later. Use the χ^2 statistic to test at the 0.05 level of significance whether the difference between the corresponding sample proportions is significant.

10.29 The following data come from a study in which random samples of the employees of three government agencies were asked questions about their pension plan:

	Agency 1	*Agency 2*	*Agency 3*
For the pension plan	67	84	109
Against the pension plan	33	66	41

Use the 0.01 level of significance to test the null hypothesis that the actual proportions of employees favoring the pension plan are the same.

10.30 The owner of a machine shop must decide which of two snack-vending machines to install in his shop. If each machine is tested 250 times and the first machine fails to work (neither delivers the snack nor returns the money) 13 times and the second machine fails to work 7 times, test at the 0.05 level of significance whether the difference between the corresponding sample proportions is significant, using

(a) the χ^2 statistic on page 286;

(b) the Z statistic on page 288.

10.31 With reference to the preceding exercise, verify that the square of the value obtained for Z in part (b) equals the value obtained for χ^2 in part (a).

10.32 Photolithography plays a central role in manufacturing integrated circuits made on thin disks of silicon. Prior to a quality-improvement program, too many rework operations were required. In a sample of 200 units, 26 required reworking of the photolithographic step. Following training in the use of Pareto charts and other approaches to identify significant problems, improvements were made. A new sample of size 200 had only 12 that needed rework.

Is this sufficient evidence to conclude at the 0.01 level of significance that the improvements have been effective in reducing the rework?

10.33 With reference to Exercise 10.32, find a large sample 99% confidence interval for the true difference of the proportions.

10.34 To test the null hypothesis that the difference between two population proportions equals some constant δ_0, not necessarily 0, we can use the statistic

$$Z = \frac{\dfrac{X_1}{n_1} - \dfrac{X_2}{n_2} - \delta_0}{\sqrt{\dfrac{\dfrac{X_1}{n_1}\left(1 - \dfrac{X_1}{n_1}\right)}{n_1} + \dfrac{\dfrac{X_2}{n_2}\left(1 - \dfrac{X_2}{n_2}\right)}{n_2}}}$$

which, for large samples, is a random variable having the standard normal distribution.

(a) With reference to Exercise 10.32, use this statistic to test at the 0.05 level of significance whether the true proportion of units requiring rework is now at least 4% less than before the improvements were made.

(b) In a true-false test, a test item is considered to be good if it discriminates between well-prepared students and poorly prepared students. If 205 of 250 well-prepared students and 137 of 250 poorly prepared students answer a certain item correctly, test at the 0.01 level of significance whether for the given item the proportion of correct answers can be expected to be at least 15% higher among well-prepared students than among poorly prepared students.

10.35 With reference to part (b) of Exercise 10.34, find a large sample 99% confidence interval for the true difference of the proportions.

10.36 Verify that the formulas for the χ^2 statistic on page 285 (with $\hat{p}$ substituted for the p_i) and on page 286 are equivalent.

10.37 Verify that if the expected frequencies are determined in accordance with the rule on page 286, the sum of the expected frequencies for each row and column equals the sum of the corresponding observed frequencies.

10.38 Verify that the square of the Z statistic on page 288 equals the χ^2 statistic on page 286 for $k = 2$.

10.4 Analysis of $r \times c$ Tables

As we suggested earlier, the method by which we analyzed the example on page 286 lends itself also to the analysis of $r \times c$ **tables**, or r-by-c tables; that is, tables in which data are tallied into a two-way classification having r rows and c columns. Such tables arise in essentially two kinds of problems. First, we might again have samples from several populations, with the distinction that now each trial permits more than two possible outcomes. This might happen, for example, if persons belonging to different income groups are asked whether they favor a certain political candidate, whether they are against him, or whether they are indifferent or undecided. The other situation giving rise to an $r \times c$ table is one in which we sample from one population but classify each item with respect to two (usually qualitative) categories. This might happen, for example, if a consumer testing service rates cars as excellent, superior, average, or poor with regard to performance and also with regard to appearance. Each car tested would then fall into one of the 16 cells of a 4×4 table, and it is mainly in connection with problems of this kind that $r \times c$ tables are referred to as **contingency tables**.

The essential difference between the two kinds of situations giving rise to $r \times c$ tables is that in the first case the column totals (the sample sizes) are fixed, while in the second case only the **grand total** (the total for the entire table) is fixed. As a result, there are also differences in the null hypotheses we shall want to test. In the first case we want to test whether the probability of obtaining an observation in the ith row is the same for each column; symbolically, we shall want to test the null hypothesis

$$p_{i1} = p_{i2} = \cdots = p_{ic} \qquad \text{for } i = 1, 2, \ldots, r$$

where p_{ij} is the probability of obtaining an observation belonging to the ith row and the jth column, and

$$\sum_{i=1}^{r} p_{ij} = 1$$

for each column, the alternative hypothesis is that the p's are not all equal for at least one row. In the second case we shall want to test the null hypothesis that the random variables represented by the two classifications are independent, so that p_{ij} is the product of the probability of getting a value belonging to the ith row and the probability of getting a value belonging to the jth column. The alternative hypothesis is that the two random variables are dependent.

In spite of the differences we have described, the analysis of an $r \times c$ table is the same for both cases. First we calculate the expected cell frequencies.

$$e_{ij} = \frac{(i\text{th row total}) \times (j\text{th column total})}{\text{grand total}}$$

When the column totals are fixed, the test is called a test of homogeneity.

Null hypothesis of homogeneity

$$H_0: p_{i1} = p_{i2} = \cdots = p_{ic} \qquad \text{for all rows } i = 1, 2, \ldots, r$$

EXAMPLE **Contingency table with column totals fixed**

Three different shops are used to repair electric motors. One hundred motors are sent to each shop. When a motor is returned, it is put in use and then the repair is classified

as complete, requiring an adjustment, or an incomplete repair. The column totals are fixed at 100 each and the grand total at 300. Shop 1 produced 78 complete repairs, 15 minor adjustments, and 7 incomplete repairs. Shop 2 produced 56, 30, and 14, respectively; while Shop 3 produced 54, 31, and 15 complete, minor adjustments, and incomplete repairs, respectively.

		Shop 1	*Shop 2*	*Shop 3*	*Total*
	Complete	78	56	54	188
Repair	*Adjustment*	15	30	31	76
	Incomplete	7	14	15	36
	Total	100	100	100	300

Calculate the expected frequencies.

Solution For the Complete–Shop 2 cell of the table,

$$e_{12} = \frac{\text{1st row total} \times \text{2nd column total}}{\text{grand total}} = \frac{188 \times 100}{300} = 62.67$$

Continuing, we obtain all of the expected frequencies, which are shown in bold below the frequencies.

		Shop 1	*Shop 2*	*Shop 3*	*Total*
	Complete	78	56	54	188
		62.7	**62.67**	**62.67**	
Repair	*Adjustment*	15	30	31	76
		25.33	**25.33**	**25.33**	
	Incomplete	7	14	15	36
		12.00	**12.00**	**12.00**	
	Total	100	100	100	300

Visually, Shop 1 has more than expected complete repairs and lower minor adjustments and incomplete repairs. A chi square test, described below, verifies that repair probabilities for the three shops are not homogeneous (see Exercise 10.39). ■

The observed frequencies and the expected frequencies total the same for each row and column, so that only $(r-1)(c-1)$ of the e_{ij} have to be calculated directly, while the others can be obtained by subtraction from appropriate row or column totals. To perform a **chi square test** we then substitute into the formula

χ^2 statistic for analysis of $r \times c$ table

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

and we reject the null hypothesis if the value of this statistic exceeds χ^2_α for $(r-1)(c-1)$ degrees of freedom. This expression for the number of degrees of freedom is justified by the above observation that after we determine $(r-1)(c-1)$ of the expected cell frequencies, the others are automatically determined; that is, they may be obtained by subtraction from appropriate row or column totals.

When only the grand total is fixed in advance, the null hypothesis specifies that each cell probability p_{ij} is the product of the marginal totals $p_{i\cdot} = \sum_{j=1}^{c} p_{ij}$ and $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$.

Null hypothesis of independence

$$H_0: p_{ij} = p_{i\cdot}\, p_{\cdot j} \qquad \text{for all } i,\ j$$

where $p_{i\cdot} = \sum_{j=1}^{c} p_{ij}$ and $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$

EXAMPLE **The chi square test of independence**

To determine whether there really is a relationship between an employee's performance in the company's training program and his or her ultimate success in the job, the company takes a sample of 400 cases from its very extensive files and obtains the results shown in the following table:

		Performance in training program			
		Below average	*Average*	*Above average*	*Total*
Success in job (employer's rating)	*Poor*	23	60	29	112
	Average	28	79	60	167
	Very good	9	49	63	121
	Total	60	188	152	400

Use the 0.01 level of significance to test the null hypothesis that performance in the training program and success in the job are independent.

Solution

1. *Null hypothesis*: Performance in training program and success in job are independent.
   Alternative hypothesis: Performance in training program and success in job are dependent.
2. *Level of significance*: $\alpha = 0.01$
3. *Criterion*: Reject the null hypothesis if $\chi^2 > 13.277$, the value of $\chi^2_{0.01}$ for $(3-1)(3-1) = 4$ degrees of freedom, where χ^2 is given by the formula above.
4. *Calculations*: Calculating first the expected cell frequencies for the first two cells of the first two rows, we get

$$e_{11} = \frac{112 \cdot 60}{400} = 16.80 \qquad e_{12} = \frac{112 \cdot 188}{400} = 52.64$$

$$e_{21} = \frac{167 \cdot 60}{400} = 25.05 \qquad e_{22} = \frac{167 \cdot 188}{400} = 78.49$$

Then, by subtraction, we find that the expected frequencies for the third cell of the first two rows are 42.56 and 63.46, and those for the third row are 18.15, 56.87, and 45.98. Thus,

$$\begin{aligned}\chi^2 &= \frac{(23-16.80)^2}{16.80} + \frac{(60-52.64)^2}{52.64} + \frac{(29-42.56)^2}{42.56} \\ &\quad + \frac{(28-25.05)^2}{25.05} + \frac{(79-78.49)^2}{78.49} + \frac{(60-63.46)^2}{63.46} \\ &\quad + \frac{(9-18.15)^2}{18.15} + \frac{(49-56.87)^2}{56.87} + \frac{(63-45.98)^2}{45.98} \\ &= 20.179\end{aligned}$$

5. *Decision*: Since $\chi^2 = 20.179$ exceeds 13.277, the null hypothesis must be rejected; we conclude that there is a dependence between an employee's performance in the training program and his or her success in the job. ■

We pursue this example further in order to determine the form of the dependence.

EXAMPLE **Exploring the form of dependence**

With reference to the preceding example, find the individual contributions to the chi square statistic.

Solution We display the contingency table, but this time we conclude the expected frequencies just below the observed frequencies.

		Performance in training program			
		Below average	*Average*	*Above average*	*Total*
Success in job (employer's rating)	*Poor*	23 **16.80**	60 **52.64**	29 **42.56**	112
	Average	28 **25.05**	79 **78.49**	60 **63.46**	167
	Very good	9 **18.15**	49 **56.87**	63 **45.98**	121
	Total	60	188	152	400

Also, we write the χ^2 statistic as the sum of the contributions.

$$\begin{aligned}\chi^2 &= 2.288 + 1.029 + 4.320 \\ &\quad + 0.347 + 0.003 + 0.189 \\ &\quad + 4.613 + 1.089 + 6.300 \\ &= 20.179\end{aligned}$$

From these two displays, it is clear that there is a positive dependence between performance in training and job success. For the three individual cells with the largest contributions to χ^2, the *above average–very good* cell frequency is high, whereas the *above average–poor* and *below average–very good* cell frequencies are low. ■

10.5 Goodness of Fit

We speak of **goodness of fit** when we try to compare an observed frequency distribution with the corresponding values of an expected, or theoretical, distribution. To illustrate, suppose that during 400 five-minute intervals the air-traffic control of an airport received 0, 1, 2, . . . , or 13 radio messages with respective frequencies of 3, 15, 47, 76, 68, 74, 39, 15, 9, 5, 2, 0, and 1. Suppose, furthermore, that we want to check whether these data substantiate the claim that the number of radio messages which they receive during a 5-minute interval may be looked upon as a random variable having the Poisson distribution with $\lambda = 4.6$. Looking up the corresponding Poisson probabilities in Table 2 and multiplying them by 400 to get the expected

frequencies, we arrive at the result shown in the following table, together with the original data:

Number of radio messages	Observed frequencies	Poisson probabilities	Expected frequencies
0	3 } 18	0.010	4.0 } 22.4
1	15	0.046	18.4
2	47	0.107	42.8
3	76	0.163	65.2
4	68	0.187	74.8
5	74	0.173	69.2
6	46	0.132	52.8
7	39	0.087	34.8
8	15	0.050	20.0
9	9	0.025	10.0
10	5 } 8	0.012	4.8 } 8.0
11	2	0.005	2.0
12	0	0.002	0.8
13	1	0.001	0.4
	400		400.0

Note that we combined some of the data so that none of the expected frequencies is less than 5.

To test whether the discrepancies between the observed and expected frequencies can be attributed to chance, we use the statistic

χ^2 statistic for test of goodness of fit

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where the o_i and e_i are the observed and expected frequencies. The sampling distribution of this statistic is approximately the chi square distribution with $k - m$ degrees of freedom, where k is the number of terms in the formula for χ^2 and m is the number of quantities, obtained from the observed data, that are needed to calculate the expected frequencies.

EXAMPLE **A chi square goodness of fit to the Poisson distribution**

With reference to the radio message data above, test at the 0.01 level of significance whether the data can be looked upon as values of a random variable having the Poisson distribution with $\lambda = 4.6$.

Solution

1. *Null hypothesis*: Random variable has a Poisson distribution with $\lambda = 4.6$.
   Alternative hypothesis: Random variable does not have the Poisson distribution with $\lambda = 4.6$.
2. *Level of significance*: $\alpha = 0.05$
3. *Criterion*: Reject the null hypothesis if $\chi^2 > 16.919$, the value of $\chi^2_{0.05}$ for $k - m = 10 - 1 = 9$ degrees of freedom, where χ^2 is given by the formula above. (The number of degrees of freedom is $10 - 1 = 9$, since only one quantity, the total frequency of 400, is needed from the observed data to calculate the expected frequencies.)

4. *Calculations*: Substitution into the formula for χ^2 yields

$$\chi^2 = \frac{(18-22.4)^2}{22.4} + \frac{(47-42.8)^2}{42.8} + \cdots + \frac{(9-10.0)^2}{10.0} + \frac{(8-8.0)^2}{8.0}$$
$$= 6.749$$

5. *Decision*: Since $\chi^2 = 6.749$ does not exceed 16.919, the null hypothesis cannot be rejected; we cannot reject that the Poisson distribution with $\lambda = 4.6$ provides a good fit at level $\alpha = 0.05$. ■

Exercises

10.39 Referring to the motor repair example on page 292, use the 0.05 level of significance to test whether there is homogeneity among the shops' repair distributions.

10.40 A large electronics firm that hires many workers with disabilities wants to determine whether their disabilities affect such workers' performance. Use the level of significance $\alpha = 0.05$ to decide on the basis of the sample data shown in the following table whether it is reasonable to maintain that the disabilities have no effect on the workers' performance:

	Performance		
	Above average	*Average*	*Below average*
Blind	21	64	17
Deaf	16	49	14
No disability	29	93	28

10.41 Tests of the fidelity and the selectivity of 190 digital radio receivers produced the results shown in the following table:

		Fidelity		
		Low	*Average*	*High*
Selectivity	*Low*	6	12	32
	Average	33	61	18
	High	13	15	0

Use the 0.01 level of significance to test whether there is a relationship (dependence) between fidelity and selectivity.

10.42 A quality-control engineer takes daily samples of $n = 4$ tractors coming off an assembly line and on 200 consecutive working days the data summarized in the following table are obtained:

Number requiring adjustments	Number of days
0	101
1	79
2	19
3	1

To test the claim that 10% of all the tractors coming off this assembly line require adjustments, look up the corresponding probabilities in Table 1, calculate the expected frequencies, and perform the chi square test at the 0.01 level of significance. (combine data so number of days ≥ 5.)

10.43 With reference to Exercise 10.42, verify that the mean of the observed distribution is 0.60, corresponding to 15% of the tractors requiring adjustments. Then look up the probabilities for $n = 4$ and $p = 0.15$ in Table 1, calculate the expected frequencies, and test at the 0.01 level of significance whether the binomial distribution with $n = 4$ and $p = 0.15$ provides a suitable model for this situation.

10.44 The following is the distribution of the hourly number of trucks arriving at a company's warehouse:

Trucks arriving per hour	Frequency
0	52
1	151
2	130
3	102
4	45
5	12
6	5
7	1
8	2

Find the mean of this distribution, and using it (rounded to one decimal place) as the parameter λ, fit a Poisson distribution. Test for goodness of fit at the 0.05 level of significance.

10.45 Among 100 purification filters used in an experiment, 46 had a service life of less than 20 hours, 19 had a service life of 20 or more but less than 40 hours, 17 had a service life of 40 or more but less than 60 hours, 12 had a service life of 60 or more but less than 80 hours, and 6 had a service life of 80 hours or more. Test at the 0.01 level of significance whether the lifetimes may be regarded as a sample from an exponential population with $\mu = 40$ hours.

Calculate $P[0 < X < 20]$ and multiply by 100 to obtain e_1, and so on.

10.46 A chi square test is easily implemented on a computer. With the counts

31	42	22	25
19	8	28	25

from the example on page 287 in columns 1–4, the *MINITAB* commands

> **Dialog box:**
> **Stat > Tables > Chisquare Test**
> Type $C1 - C4$ in **columns containing the table.**
> Click **OK**.

produce the output

Expected counts are printed below observed counts

	Method 1	Method 2	Method 3	Method 4	Total
1	31	42	22	25	120
	30.00	30.00	30.00	30.00	
2	19	8	28	25	80
	20.00	20.00	20.00	20.00	
Total	50	50	50	50	200

$$\text{Chi sq} = 0.033 + 4.800 + 2.133 + 0.050 + 7.200 + 3.200 + 1.250 = 19.550$$

$$\text{df} = 3$$

Repeat the analysis using only the data from the first three methods.

10.47 The procedure in Exercise 10.46 also calculates the chi square test for independence. Do Exercise 10.40 using the computer.

Do's and Don'ts

Do's

1. Remember that it usually takes a sample size of a few hundred to get precise estimates of a proportion. When the sample size is large, calculate the approximate $100(1-\alpha)\%$ confidence interval for the population proportion p

$$\frac{x}{n} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

Base a test of the null hypothesis H_0: $p = p_0$ on the approximately standard normal test statistic

$$Z = \frac{X - n\,p_0}{\sqrt{n\,p_0\,(1-p_0)}}$$

2. When both sample sizes are large, compare two proportions by calculating the approximate $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ with limits

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}$$

Base tests of the null hypothesis H_0: $p_1 = p_2$ on the approximately standard normal test statistic

$$Z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \text{with} \qquad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

3. Use the chi square test to analyze $r \times c$ contingency tables. The test statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

with estimated expected values

$$e_{ij} = \frac{(i\text{th row total}) \times (j\text{th column total})}{\text{grand total}}$$

and observed values o_{ij} is approximately chi square distributed with $(r - 1)(c - 1)$ degrees of freedom. This statistic applies to comparing several proportions when samples of size n_j are selected from the jth population. The same statistic applies to testing independence of the two sets of categories when a single sample of size n is cross-tabulated to create the $r \times c$ table.

Don'ts

1. Don't routinely apply the inference procedures for a proportion p without confirming that the outcomes from different trials are independent.
2. When sampling without replacement from a finite population of size N, don't forget to account for dependence when the finite population correction factor is substantially different from 1.

Review Exercises

10.48 In a sample of 100 ceramic pistons made for an experimental diesel engine, 18 were cracked. Construct a 95% confidence interval for the true proportion of cracked pistons using the large sample confidence interval formula.

10.49 With reference to Exercise 10.48, test the null hypothesis $p = 0.20$ versus the alternative hypothesis $p < 0.20$ at the 0.05 level.

10.50 In a random sample of 160 workers exposed to a certain amount of radiation, 24 experienced some ill effects. Construct a 99% confidence interval for the corresponding true percentage using the large sample confidence interval formula.

10.51 With reference to Exercise 10.50, test the null hypothesis $p = 0.18$ versus the alternative hypothesis $p \neq 0.18$ at the 0.01 level.

10.52 In a random sample of 100 packages shipped by air freight, 16 had some damage. Construct a 95% confidence interval for the true proportion of damaged packages using the large sample confidence interval formula.

10.53 With reference to Exercise 10.52, test the hypothesis $p = 0.10$ versus the alternative hypothesis $p > 0.10$ at the 0.01 level.

10.54 If 26 of 200 Brand A tires fail to last 20,000 miles, whereas the corresponding figures for 200 tires each of Brands B, C, and D are 23, 15, 32, use the 0.05 level of significance to test the null hypothesis that there is no difference in the quality of the four kinds of tires with regard to their durability.

10.55 One method of seeding clouds was successful in 57 of 150 attempts while another method was successful in 33 of 100 attempts. At the 0.05 level of significance, can we conclude that the first method is better than the second?

10.56 With reference to Exercise 10.55, find a large sample 95% confidence interval for the true difference of probabilities.

10.57 Two bonding agents, A and B, are available for making a laminated beam. Of 50 beams made with Agent A, 11 failed a stress test, whereas 19 of the 50 beams made with Agent B failed. At the 0.05 level, can we conclude that Agent A is better than Agent B?

10.58 With reference to Exercise 10.57, find a large sample 95% confidence interval for the true difference of the probabilities of failure.

10.59 Cooling pipes at three nuclear power plants are investigated for deposits that would inhibit the flow of

water. From 30 randomly selected spots at each plant, 13 from the first plant, 8 from the second plant, and 19 from the third were clogged.

(a) Use the 0.05 level to test the null hypothesis of equality.

(b) Plot the confidence intervals for the three probabilities of being clogged.

10.60 Suppose that in Exercise 10.54 we had been interested also in how many of the tires lasted more than 30,000 miles and obtained the results shown in the following table:

	Brand A	*Brand B*	*Brand C*	*Brand D*
Failed to last 20,000 *miles*	26	23	15	32
Lasted from 20,000 *to* 30,000	118	93	116	121
Lasted more than 30,000 *miles*	56	84	69	47

(a) Use the 0.01 level of significance to test the null hypothesis that there is no difference in the quality of the four kinds of tires with regard to their durability.

(b) Plot the four individual 99% confidence intervals for proportions that last more than 30,000 miles.

10.61 The following is the distribution of the daily number of power failures reported in a western city on 300 days:

Number of power failures	Number of days
0	9
1	43
2	64
3	62
4	42
5	36
6	22
7	14
8	6
9	2

Test at the 0.05 level of significance whether the daily number of power failures in this city is a random variable having the Poisson distribution with $\lambda = 3.2$.

10.62 With reference to the example on page 294, repeat the analysis after combining the categories *below average* and *average* in the training program and the categories *poor* and *average* in success. Comment on the form of the dependence.

10.63 Mechanical engineers, testing a new arc-welding technique, classified welds both with respect to appearance and an X-ray inspection.

		Appearance			
		Bad	*Normal*	*Good*	*Total*
X-ray	*Bad*	20	7	3	30
	Normal	13	51	16	80
	Good	7	12	21	40
	Total	40	70	40	150

Test for independence using $\alpha = 0.05$ and find the individual cell contributions to the χ^2 statistic.

Key Terms

Chi square test 286, 293
Contingency table 292
Expected cell frequency 286
Goodness of fit 295
Observed cell frequency 286
$r \times c$ table 292
Sample proportion 278

CHAPTER

11

Regression Analysis

The main objective of many statistical investigations is to make predictions, preferably on the basis of mathematical equations. For instance, an engineer may wish to predict the amount of oxide that will form on the surface of a metal baked in an oven for one hour at 200 degrees Celsius, or the amount of deformation of a ring subjected to a compressive force of 1,000 pounds, or the number of miles to wear out a tire as a function of tread thickness and composition. Usually, such predictions require that a formula be found which relates the dependent variable (whose value one wants to predict) to one or more independent variables.

Problems relating to predictions based on the known value of one variable are treated in Sections 11.1 through 11.3 and 11.6; the case where predictions are based on the known values of several variables is treated in Section 11.4. The importance of checking the assumptions concerning the prediction model is considered in Section 11.5.

CHAPTER OUTLINE

11.1 The Method of Least Squares 301

11.2 Inferences Based on the Least Squares Estimators 309

11.3 Curvilinear Regression 321

11.4 Multiple Regression 326

11.5 Checking the Adequacy of the Model 329

11.6 Correlation 335

11.7 Multiple Linear Regression (Matrix Notation) 345

Review Exercises 350

Key Terms 353

11.1 The Method of Least Squares

In this section we begin our study of the case where a dependent variable is to be predicted in terms of a single independent variable. In many problems of this kind the independent variable is observed without error, or with an error which is negligible when compared with the error (chance variation) in the dependent variable. For example, in measuring the amount of oxide on the surface of a metal specimen, the baking temperature can usually be controlled with good precision, but the oxide-thickness measurement may be subject to considerable chance variation. Thus, even though the independent variable may be fixed at x, repeated measurements of the dependent variable may lead to y values which differ considerably. Such differences among y values can be attributed to several causes, chiefly to errors of measurement and to the existence of other, uncontrolled variables which may influence the measured thickness, y, when x is fixed. Thus, measurements of the thickness of oxide layers may vary over several specimens baked for the same length of time at the same temperature because of the difficulty in measuring thickness as well as possible differences in the composition of the oven atmosphere, surface conditions of the specimens, and the like.

It should be apparent from this discussion that, in this context, the measured thickness of oxide layers Y is a random variable whose distribution depends on x. In most situations of this sort we are interested mainly in the relationship between x and the mean of the corresponding distribution of Y, and we refer to this relationship as the **regression curve of Y on x**. (For the time being we shall consider the case where x is fixed, that is, not random. In Section 11.6 we shall consider the case where x and y are both values of random variables.)

Let us first treat the case where the regression curve of Y on x is a **linear regression** curve; that is, where for any given x, the mean of the distribution of the Y's is given by $\alpha + \beta x$. In general, Y will differ from this mean, and we shall denote this difference by ε, writing

$$Y = \alpha + \beta x + \varepsilon$$

Thus, ε is a random variable and we can always choose α so that the mean of the distribution of this random variable is equal to zero. The value of ε for any given observation will depend on a possible error of measurement and on the values of variables other than x which might have an influence on Y.

To give an example where the regression curve of Y on x can reasonably be assumed to be linear, suppose that a tensile ring is to be calibrated by measuring the deflection at various loads. In the following table, which gives the results of 12 measurements, the x's are the load forces in thousands of pounds and the y values are the corresponding deflections in thousandths of an inch:

x	1	2	3	4	5	6	7	8	9	10	11	12
y	16	35	45	64	86	96	106	124	134	156	164	182

It is apparent from Figure 11.1, where these data have been plotted, that it is reasonable to assume that the relationship (regression curve) is linear. That is, a straight line gives a very good approximation over the range of the available data. This kind of diagram, which shows how the data points are scattered, is called a **scatter plot** or **scattergram**.

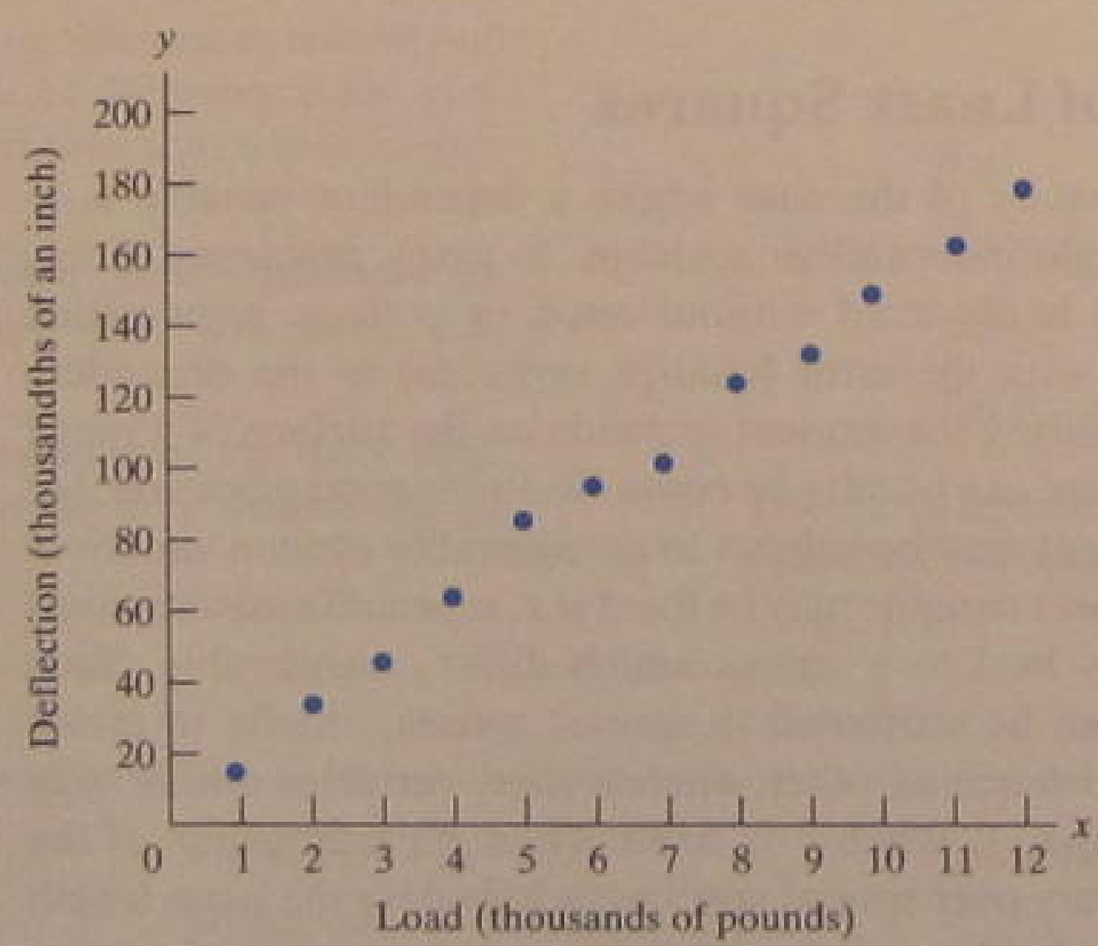


Figure 11.1
Scatter plot indicating linear regression

Now we face the problem of using the observed data to estimate the parameters α and β of the regression line that somehow provide the best fit to the data. In connection with Figure 11.1, it may well be satisfactory to do this by eye. If different experimenters were to fit a line in this way, they would probably all predict that for a load of 7,500 pounds the deflection should be close to 0.115 inch. However, if we have to deal with data such as those plotted in Figure 11.2, the problem of finding a best-fitting line is not so obvious. To handle problems of this kind, we must seek a nonsubjective method for fitting straight lines which reflects some desirable statistical properties.

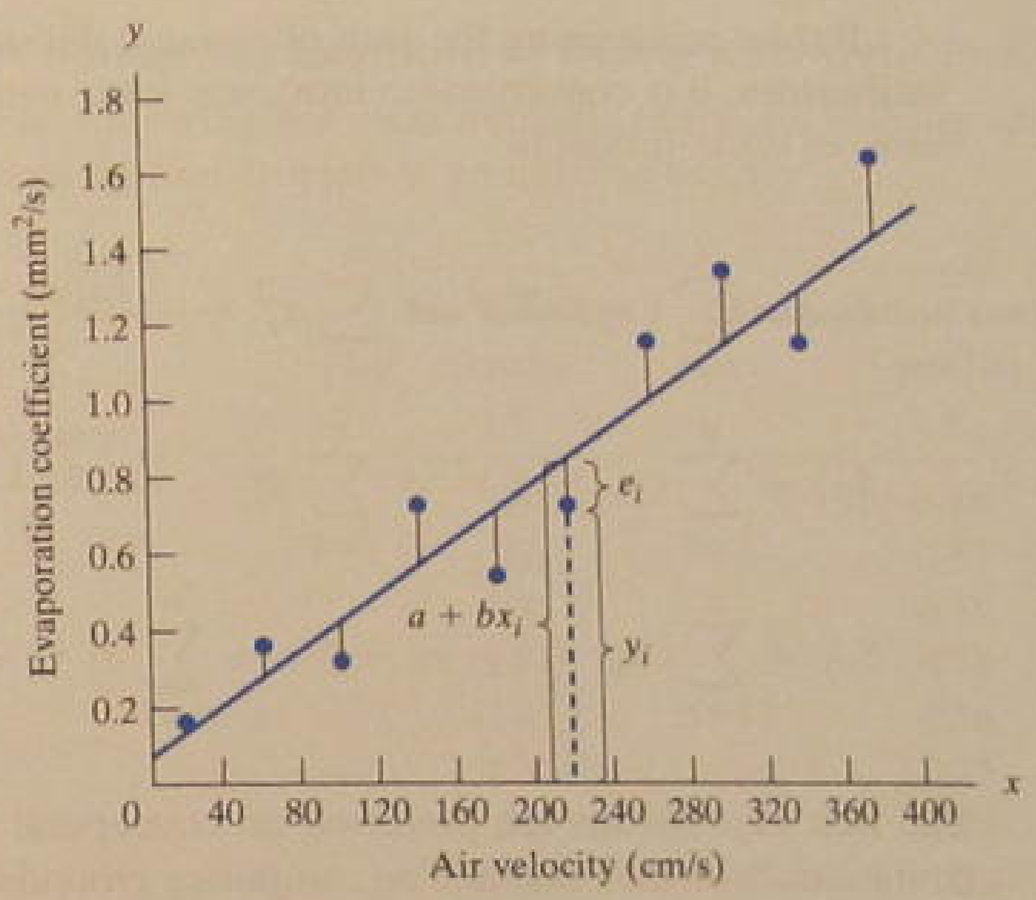


Figure 11.2
Diagram for least squares criterion showing the vertical deviations

To state the problem formally, we have n paired observations (x_i, y_i) for which it is reasonable to assume that the regression of Y on x is linear, and we want to determine the line (that is, the equation of the line) which in some sense provides the best fit. There are several ways in which we interpret the word "best," and the meaning we shall give it here may be explained as follows. If we predict y by means of the equation

$$\hat{y} = a + bx$$

where a and b are constants, then e_i, the error in predicting the value of y corresponding to the given x_i, is

$$e_i = y_i - \hat{y}_i$$

and we shall want to determine a and b so that these errors are in some sense as small as possible.

Since we cannot simultaneously minimize each of the e_i individually, we might try to make their sum $\sum_{i=1}^{n} e_i$ as close as possible to zero. However, since this sum can be made equal to zero by many choices of totally unsuitable lines for which the positive and negative errors cancel, we shall minimize the sum of the squares of the e_i (for the same reason we worked with the squares of the deviations from the mean in the definition of the standard deviation). In other words, we apply the **principle of least squares** and choose a and b so that

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [\, y_i - (a + b x_i)\,]^2$$

is a minimum. Note from Figure 11.2 that this is equivalent to minimizing the sum of the squares of the vertical distances from the points to the line.

The procedure of finding the equation of the line which best fits a given set of paired data, called the **method of least squares**, yields values for a and b (estimates of α and β) that have many desirable properties; some of these are mentioned on page 309.

Before minimizing the sum of squared deviations to obtain the **least squares estimators**, it is convenient to introduce some notation for the sums of squares and sums of cross-products.

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

The first expressions are preferred on conceptual grounds because they highlight deviations from the mean and on computing grounds because they are less susceptible to roundoff error. The second expressions are for handheld calculators.

Below, we show that the least squares estimates are

Least squares estimates

$$a = \overline{y} - b \cdot \overline{x} \quad \text{and} \quad b = \frac{S_{xy}}{S_{xx}}$$

where $\overline{x}$ and $\overline{y}$ are, respectively, the means of the values of x and y.

The least squares estimates determine the best-fitting line

Fitted (or estimated) regression line

$$\widehat{y} = a + b\,x$$

where the ^ on y indicates the estimated value.

The individual deviations of the observations y_i from their fitted values $\widehat{y}_i = a + b\,x_i$ are called the **residuals**.

Residuals

$$\text{observation} - \text{fitted value} = y_i - a - b\,x_i$$

The minimum value of the sum of squares is called the **residual sum of squares** or **error sum of squares**. Below we show that

$$\text{SSE} = \text{residual sum of squares} = \sum_{i=1}^{n} (y_i - a - b\,x_i)^2$$

$$= S_{yy} - S_{xy}^2 / S_{xx}$$

EXAMPLE **A numerical example of fitting a straight line by least squares**

The following are measurements of the air velocity and evaporation coefficient of burning fuel droplets in an impulse engine:

Air velocity (cm/s) x	Evaporation coefficient (mm^2/s) y
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17
380	1.65

Fit a straight line to these data by the method of least squares, and use it to estimate the evaporation coefficient of a droplet when the air velocity is 190 cm/s.

Solution For these $n = 10$ pairs (x_i, y_i) we first calculate

$$\sum_{i=1}^{n} x_i = 2{,}000 \qquad \sum_{i=1}^{n} x_i^2 = 532{,}000$$

$$\sum_{i=1}^{n} y_i = 8.35 \qquad \sum_{i=1}^{n} x_i y_i = 2{,}175.40$$

$$\sum_{i=1}^{n} y_i^2 = 9.1097$$

and then we obtain

$$S_{xx} = 532{,}000 - (2{,}000)^2/10 = 132{,}000$$
$$S_{xy} = 2{,}175.40 - (2{,}000)(8.35)/10 = 505.40$$
$$S_{yy} = 9.1097 - (8.35)^2/10 = 2.13745$$

Consequently,

$$b = \frac{S_{xy}}{S_{xx}} = \frac{505.40}{132{,}000} = 0.00383$$

and then

$$a = \bar{y} - b\bar{x} = \frac{8.35}{10} - 0.00383\,\frac{2{,}000}{10} = 0.069$$

Thus, the equation of the straight line that best fits the given data in the sense of least squares is

$$\hat{y} = 0.069 + 0.00383\,x$$

and for $x = 190$ we predict that the evaporation coefficient will be

$$\hat{y} = 0.069 + 0.00383(190) = 0.80 \text{ mm}^2/\text{s}$$

Finally, the residual sum of squares is

$$S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 2.13745 - \frac{(505.40)^2}{132{,}000} = 0.20238$$ ■

To avoid confusion, we make it clear that there are two possible regression lines.

EXAMPLE **One scatter plot but two different fitted lines**

Engineers fabricating a new transmission-type electron multiplier[1] created an array of silicon nanopillars (see Figure 2.5) on a flat silicon membrane. The precise structure can influence the electrical properties so, subsequently, the height and widths of 50 nanopillars (see Exercise 11.23) were measured in nanometers (nm) or $10^{-9} \times$ meters. The summary statistics, with $x =$ width and $y =$ height, are

$$n = 50 \qquad \bar{x} = 88.34 \qquad \bar{y} = 305.58$$
$$S_{xx} = 7{,}239.22 \quad S_{xy} = 17{,}840.1 \quad S_{yy} = 66{,}975.2$$

(a) Find the least squares line for predicting height from width.

(b) Find the least squares line for predicting width from height.

(c) Make a scatter plot and show both lines. Comment.

Solution **(a)** Here $y =$ height and the least squares estimates are

$$\text{slope} = b = \frac{S_{xy}}{S_{xx}} = \frac{17{,}840.1}{7{,}239.22} = 2.464 \quad \text{and}$$

$$a = \bar{y} - b\bar{x} = 305.58 - \frac{17{,}840.1}{7{,}239.22} \times 88.34 = 87.88$$

The fitted line is

$$\text{height} = 87.88 + 2.464 \text{ width}.$$

(b) Width is now the response variable and height the predictor, so x and y must be interchanged.

$$\text{slope} = b = \frac{17{,}840.1}{66{,}976.2} = 0.266 \quad \text{and}$$

$$a = 88.34 - 0.2664 \times 305.58 = 6.944$$

The fitted line is

$$\text{width} = 6.944 + 0.266 \text{ height}.$$

(c) Using the data in Exercise 11.23, we construct the scatter plot in Figure 11.3 and include the two lines. The line from part (b) is written as

$$\text{height} = -\frac{6.944}{0.266} + \frac{1}{0.266} \text{ width} = -26.11 + 3.759 \text{ width}$$

Notice that both pass through the mean point $(\bar{x}, \bar{y}) = (88.34, 305.58)$. The choice of fitted line depends on which variable you wish to predict. ■

[1] H. Qin, H. Kim, and R. Blick, Nanopillar arrays on semiconductor membranes as electron emission amplifiers, *Nanotechnology* 19 (2008), 095504 (5pp).

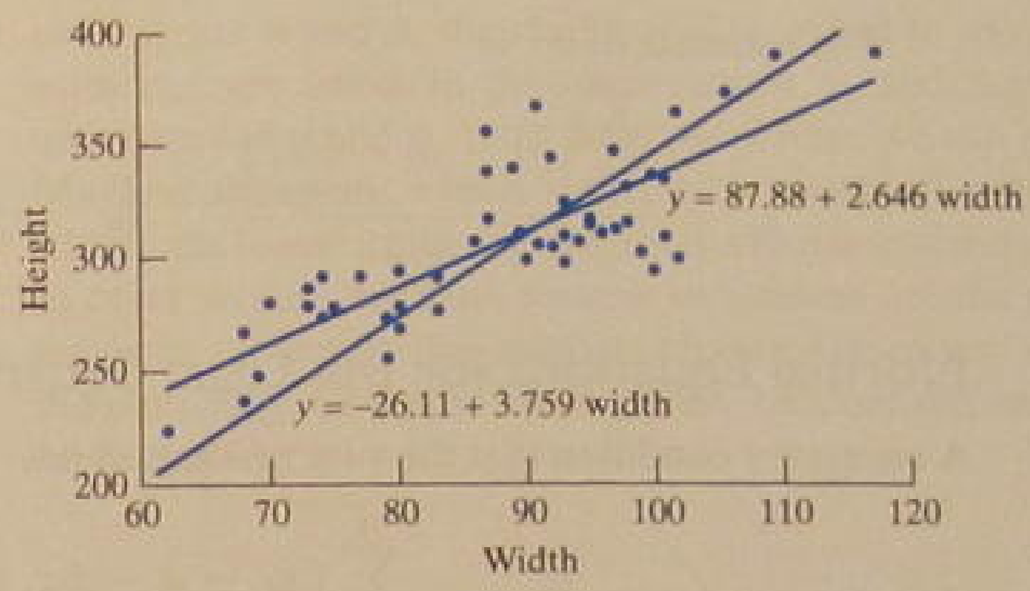


Figure 11.3
Scatter plot and two fitted lines

Determining the Least Squares Estimators

We now show that the choice of estimates

$$b = \frac{S_{xy}}{S_{xx}} \qquad a = \overline{y} - b\overline{x}$$

minimizes the sum of squares,

$$S(a, b) = \sum_{i=1}^{n} (y_i - a - b x_i)^2$$

over all choices of a and b. We first find an alternative expression for the sum of squares for any a and b. Adding and subtracting $b\overline{x} - \overline{y}$, we have

$$y_i - a - b x_i = (y_i - \overline{y}) - b(x_i - \overline{x}) + (\overline{y} - a - b\overline{x})$$

Squaring both sides, we obtain

$$\begin{aligned}(y_i - a - b x_i)^2 = {} & (y_i - \overline{y})^2 + b^2 (x_i - \overline{x})^2 + (\overline{y} - a - b\overline{x})^2 \\ & - 2b(y_i - \overline{y})(x_i - \overline{x}) \\ & - 2b(\overline{y} - a - b\overline{x})(x_i - \overline{x}) \\ & + 2(y_i - \overline{y})(\overline{y} - a - b\overline{x})\end{aligned}$$

Now, summing over both sides, the last two terms vanish so we obtain

$$\begin{aligned}S(a, b) &= S_{yy} + b^2 S_{xx} + n(\overline{y} - a - b\overline{x})^2 - 2b S_{xy} \\ &= n(\overline{y} - a - b\overline{x})^2 + \left(b^2 S_{xx} - 2b S_{xy} + \frac{S_{xy}^2}{S_{xx}}\right) + S_{yy} - \frac{S_{xy}^2}{S_{xx}} \\ &= n(\overline{y} - a - b\overline{x})^2 + \left(b\sqrt{S_{xx}} - \frac{S_{xy}}{\sqrt{S_{xx}}}\right)^2 + S_{yy} - \frac{S_{xy}^2}{S_{xx}}\end{aligned}$$

According to the principle of least squares, we must select values for a and b which will minimize this sum of squares. All three terms on the right-hand side are non-negative and the third term does not depend on the choice of a and b. The second term can be made to equal zero, its minimum value, by taking $b = S_{xy}/S_{xx}$. With this choice for b, the first term can be made equal to zero by taking $a = y - b\overline{x}$. This confirms the formula for the least squares estimators on page 304.

Further, we have shown that the minimized sum of squares equals the third term. That is, the sum of squares error

$$\text{SSE} = S_{yy} - S_{xy}^2/S_{xx}$$

as stated on page 304.

The estimate of σ^2 is

$$s_e^2 = \frac{\sum_{i=1}^{n} (y_i - a - bx_i)^2}{n-2} \quad \text{or} \quad s_e^2 = \frac{S_{yy} - S_{xy}^2/S_{xx}}{n-2}$$

Normal Equations for the Least Squares Estimators

A necessary condition that the sum of squared deviations,

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

be a minimum is the vanishing of the partial derivatives with respect to a and b. We thus have

$$2\sum_{i=1}^{n} [y_i - (a + bx_i)](-1) = 0$$

$$2\sum_{i=1}^{n} [y_i - (a + bx_i)](-x_i) = 0$$

and we can rewrite these two equations as

Normal equations

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

This set of two linear equations in the unknowns a and b, called the **normal equations**, gives the same values of a and b for the line which provides the best fit to a given set of paired data in accordance with the criterion of least squares.

EXAMPLE **The least squares estimates obtained from the normal equations**

Solve the normal equations for the data in the previous example and confirm the values for the least squares estimates.

Solution Using the calculations in the previous example, the normal equations are

$$8.35 = 10a + 2{,}000b$$
$$2{,}175.40 = 2{,}000a + 532{,}000b$$

Solving this system of equations by use of determinants or the method of elimination, we obtain $a = 0.069$ and $b = 0.00383$.

As they must be, these values are the same as those obtained in the previous example. ■

It is impossible to make any exact statement about the "goodness" of an estimate like this unless we make some assumptions about the underlying distributions of the random variables with which we are concerned and about the true nature of the regression. Looking upon the values of a and b as estimates of the actual regression

coefficients α and β, the reader will be asked to show in Exercise 11.20 that these estimates are linear in the observations y_i and that they are values of unbiased estimators of α and β. With these properties, we can refer to the remarkable **Gauss-Markov theorem**, which states that among all unbiased estimators for α and β which are linear in the Y_i, the least squares estimators have the smallest variance. In other words, the least squares estimators are the most reliable in the sense that they are subject to the smallest chance variations. A proof of the Gauss-Markov theorem may be found in the book by Johnson and Wichern referred to in the bibliography.

11.2 Inferences Based on the Least Squares Estimators

The method of the preceding section is used when the relationship between x and the mean of Y is linear or close enough to a straight line so that the least squares line yields reasonably good predictions. In what follows we shall assume that the regression *is* linear in x and, furthermore, that the n random variables Y_i are independently normally distributed with the means $\alpha + \beta x_i$ and the common variance σ^2 for $i = 1, 2, \ldots, n$. Equivalently, we write the model as

Statistical model for straight-line regression

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad \text{for } i = 1, 2, \ldots, n$$

where it is assumed that the ε_i are independent normally distributed random variables having zero means and the common variance σ^2.

The various assumptions we have made here are illustrated in Figure 11.4, showing the distributions of Y_i for several values of x_i. Note that these additional assumptions are required to discuss the goodness of predictions based on least squares equations, the properties of a and b as estimates of α and β, and so on. They were not required to obtain the original estimates based on the method of least squares.

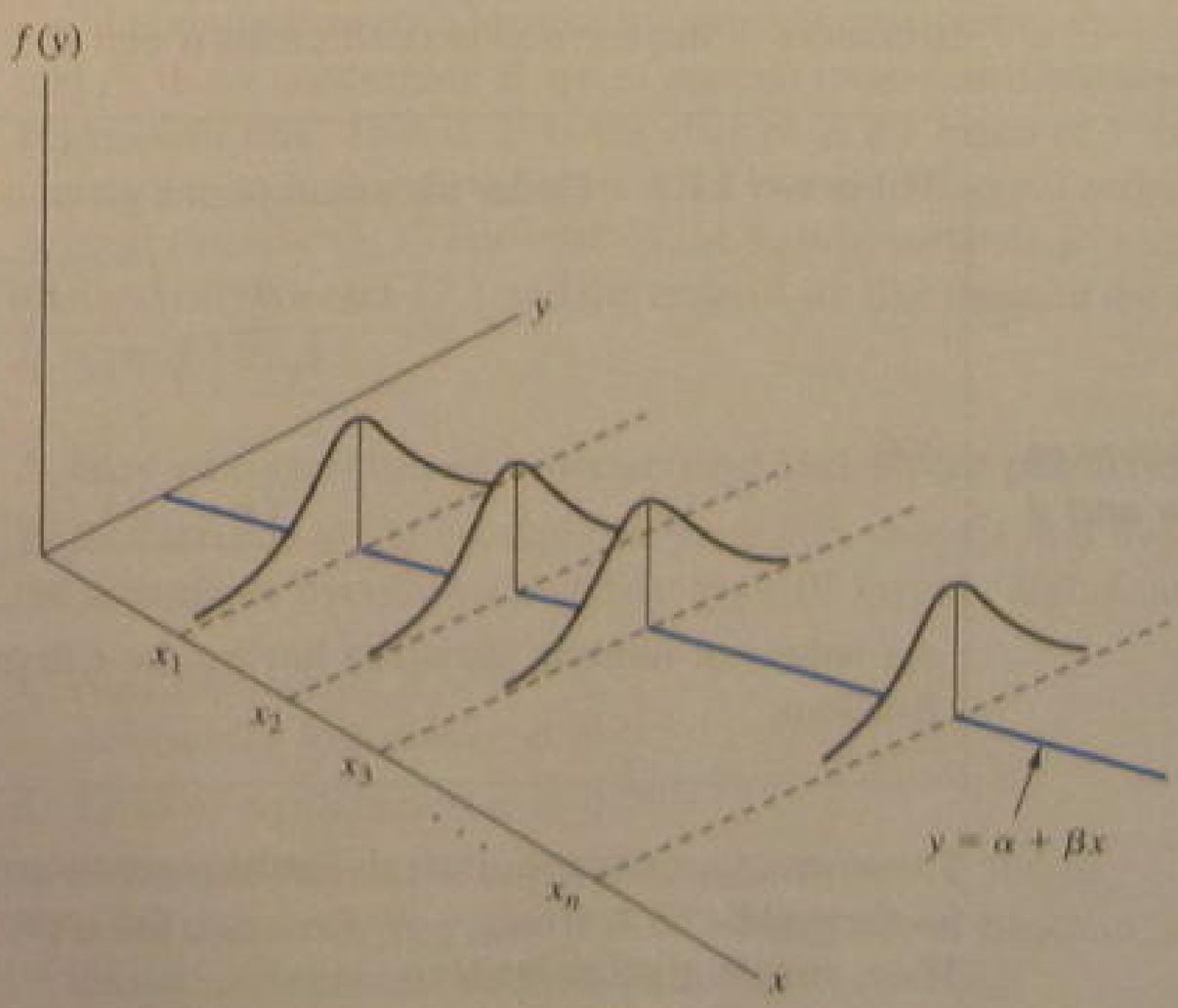


Figure 11.4
Diagram showing assumptions underlying Theorem 11.1