

BRASE • BRASE

TENTH EDITION

Understandable Statistics

Concepts and Methods



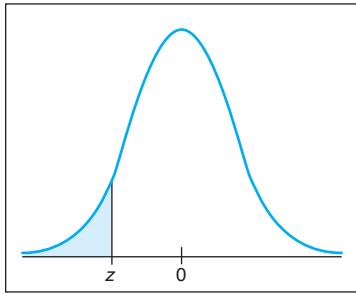


Table entry for z is the area to the left of z .

Areas of a Standard Normal Distribution

(a) Table of Areas to the Left of z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

For values of z less than -3.49, use 0.000 to approximate the area.

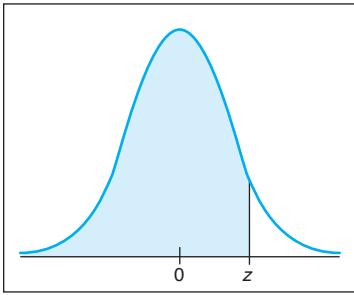


Table entry for z is the area to the left of z .

Areas of a Standard Normal Distribution *continued*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

For z values greater than 3.49, use 1.000 to approximate the area.

Areas of a Standard Normal Distribution *continued*

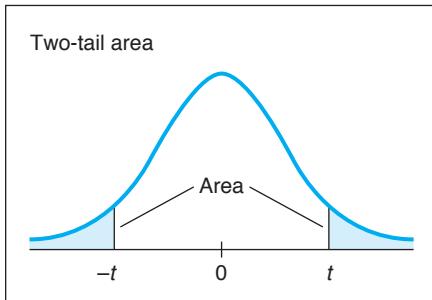
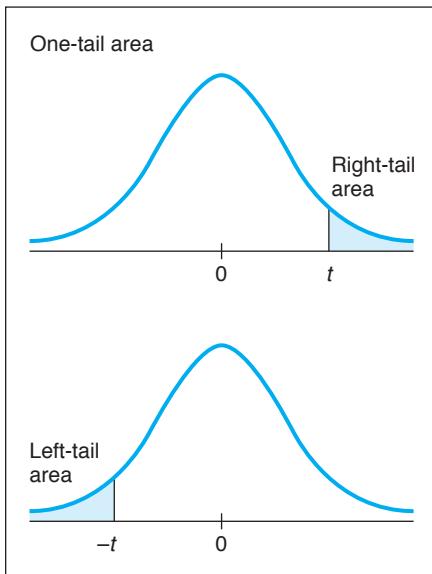
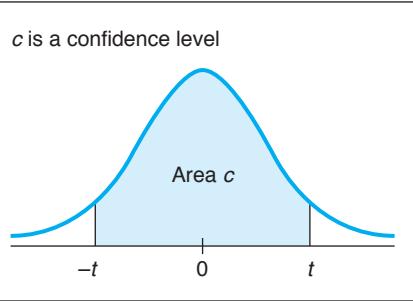
(b) Confidence Interval
Critical Values z_c

Level of Confidence c	Critical Value z_c
0.70, or 70%	1.04
0.75, or 75%	1.15
0.80, or 80%	1.28
0.85, or 85%	1.44
0.90, or 90%	1.645
0.95, or 95%	1.96
0.98, or 98%	2.33
0.99, or 99%	2.58

Areas of a Standard Normal Distribution *continued*

(c) Hypothesis Testing, Critical Values z_0

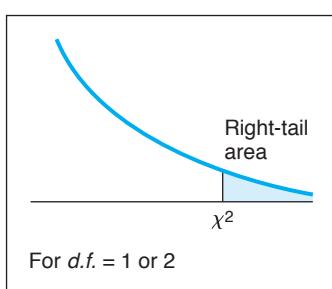
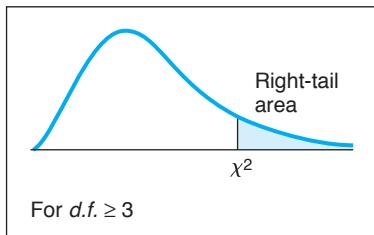
Level of Significance	$\alpha = 0.05$	$\alpha = 0.01$
Critical value z_0 for a left-tailed test	-1.645	-2.33
Critical value z_0 for a right-tailed test	1.645	2.33
Critical values $\pm z_0$ for a two-tailed test	± 1.96	± 2.58



Critical Values for Student's t Distribution

one-tail area	0.250	0.125	0.100	0.075	0.050	0.025	0.010	0.005	0.0005
two-tail area	0.500	0.250	0.200	0.150	0.100	0.050	0.020	0.010	0.0010
d.f. \ c	0.500	0.750	0.800	0.850	0.900	0.950	0.980	0.990	0.999
1	1.000	2.414	3.078	4.165	6.314	12.706	31.821	63.657	636.619
2	0.816	1.604	1.886	2.282	2.920	4.303	6.965	9.925	31.599
3	0.765	1.423	1.638	1.924	2.353	3.182	4.541	5.841	12.924
4	0.741	1.344	1.533	1.778	2.132	2.776	3.747	4.604	8.610
5	0.727	1.301	1.476	1.699	2.015	2.571	3.365	4.032	6.869
6	0.718	1.273	1.440	1.650	1.943	2.447	3.143	3.707	5.959
7	0.711	1.254	1.415	1.617	1.895	2.365	2.998	3.499	5.408
8	0.706	1.240	1.397	1.592	1.860	2.306	2.896	3.355	5.041
9	0.703	1.230	1.383	1.574	1.833	2.262	2.821	3.250	4.781
10	0.700	1.221	1.372	1.559	1.812	2.228	2.764	3.169	4.587
11	0.697	1.214	1.363	1.548	1.796	2.201	2.718	3.106	4.437
12	0.695	1.209	1.356	1.538	1.782	2.179	2.681	3.055	4.318
13	0.694	1.204	1.350	1.530	1.771	2.160	2.650	3.012	4.221
14	0.692	1.200	1.345	1.523	1.761	2.145	2.624	2.977	4.140
15	0.691	1.197	1.341	1.517	1.753	2.131	2.602	2.947	4.073
16	0.690	1.194	1.337	1.512	1.746	2.120	2.583	2.921	4.015
17	0.689	1.191	1.333	1.508	1.740	2.110	2.567	2.898	3.965
18	0.688	1.189	1.330	1.504	1.734	2.101	2.552	2.878	3.922
19	0.688	1.187	1.328	1.500	1.729	2.093	2.539	2.861	3.883
20	0.687	1.185	1.325	1.497	1.725	2.086	2.528	2.845	3.850
21	0.686	1.183	1.323	1.494	1.721	2.080	2.518	2.831	3.819
22	0.686	1.182	1.321	1.492	1.717	2.074	2.508	2.819	3.792
23	0.685	1.180	1.319	1.489	1.714	2.069	2.500	2.807	3.768
24	0.685	1.179	1.318	1.487	1.711	2.064	2.492	2.797	3.745
25	0.684	1.198	1.316	1.485	1.708	2.060	2.485	2.787	3.725
26	0.684	1.177	1.315	1.483	1.706	2.056	2.479	2.779	3.707
27	0.684	1.176	1.314	1.482	1.703	2.052	2.473	2.771	3.690
28	0.683	1.175	1.313	1.480	1.701	2.048	2.467	2.763	3.674
29	0.683	1.174	1.311	1.479	1.699	2.045	2.462	2.756	3.659
30	0.683	1.173	1.310	1.477	1.697	2.042	2.457	2.750	3.646
35	0.682	1.170	1.306	1.472	1.690	2.030	2.438	2.724	3.591
40	0.681	1.167	1.303	1.468	1.684	2.021	2.423	2.704	3.551
45	0.680	1.165	1.301	1.465	1.679	2.014	2.412	2.690	3.520
50	0.679	1.164	1.299	1.462	1.676	2.009	2.403	2.678	3.496
60	0.679	1.162	1.296	1.458	1.671	2.000	2.390	2.660	3.460
70	0.678	1.160	1.294	1.456	1.667	1.994	2.381	2.648	3.435
80	0.678	1.159	1.292	1.453	1.664	1.990	2.374	2.639	3.416
100	0.677	1.157	1.290	1.451	1.660	1.984	2.364	2.626	3.390
500	0.675	1.152	1.283	1.442	1.648	1.965	2.334	2.586	3.310
1000	0.675	1.151	1.282	1.441	1.646	1.962	2.330	2.581	3.300
∞	0.674	1.150	1.282	1.440	1.645	1.960	2.326	2.576	3.291

For degrees of freedom $d.f.$ not in the table, use the closest $d.f.$ that is *smaller*.



The χ^2 Distribution

d.f.	Right-tail Area									
	.995	.990	.975	.950	.900	.100	.050	.025	.010	.005
1	0.0 ⁴ 393	0.0 ³ 157	0.0 ³ 982	0.0 ² 393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	8.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.21	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.80	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

Source: Biometrika, June 1964, The χ^2 Distribution, H. L. Herter (Table 7). Used by permission of Oxford University Press.



INSTRUCTOR'S ANNOTATED EDITION

TENTH EDITION

Understandable Statistics

Concepts and Methods

Charles Henry Brase
Regis University

Corrinne Pellillo Brase
Arapahoe Community College



Australia • Brazil • Japan • Korea • Mexico • Singapore •
Spain • United Kingdom • United States

This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN#, author, title, or keyword for materials in your areas of interest.

*This book is dedicated to the memory of
a great teacher, mathematician, and friend*

Burton W. Jones

Professor Emeritus, University of Colorado



Understandable Statistics: Concepts and Methods, Tenth Edition

Charles Henry Brase, Corrinne Pellillo Brase
Editor in Chief: Michelle Julet
Publisher: Richard Stratton
Senior Sponsoring Editor: Molly Taylor
Senior Editorial Assistant: Shaylin Walsh
Media Editor: Andrew Coppola
Marketing Manager: Ashley Pickering
Marketing Communications Manager:
Mary Anne Payumo
Content Project Manager: Jill Clark
Art Director: Linda Helcher
Senior Manufacturing Buyer: Diane Gibbons
Senior Rights Acquisition Specialist, Text:
Katie Huha
Rights Acquisition Specialist, Images:
Mandy Groszko
Text Permissions Editor: Sue Howard
Production Service: Elm Street
Publishing Services
Cover Designer: RHDG
Cover Image: © Anup Shah
Compositor: Integra Software Services,
Ltd. Pvt.

© 2012, 2009, 2006 Brooks/Cole, Cengage Learning

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including but not limited to photocopying, recording, scanning, digitizing, taping, Web distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the publisher.

For product information and technology assistance, contact us at
Cengage Learning Customer & Sales Support, 1-800-354-9706

For permission to use material from this text or product,
submit all requests online at www.cengage.com/permissions.
Further permissions questions can be emailed to
permissionrequest@cengage.com.

Library of Congress Control Number: 2009942998

Student Edition:
ISBN-13: 978-0-8400-4838-7
ISBN-10: 0-8400-4838-6

Annotated Instructor's Edition:
ISBN-13: 978-0-8400-5456-2
ISBN-10: 0-8400-5456-4

Brooks/Cole
20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil and Japan. Locate your local office at international.cengage.com/region

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit www.cengage.com. Purchase any of our products at your local college store or at our preferred online store www.cengagbrain.com.

Printed in the United States of America
1 2 3 4 5 6 7 14 13 12 11 10



CONTENTS

1	PREFACE	xv
	Table of Prerequisite Material	1
2	GETTING STARTED	2
	FOCUS PROBLEM: Where Have All the Fireflies Gone?	3
	1.1 What Is Statistics? 4	
	1.2 Random Samples 12	
	1.3 Introduction to Experimental Design 21	
	Summary 30	
	Important Words & Symbols 30	
	Chapter Review Problems 31	
	Data Highlights: Group Projects 34	
	Linking Concepts: Writing Projects 34	
	USING TECHNOLOGY	35
2	ORGANIZING DATA	38
	FOCUS PROBLEM: Say It with Pictures	39
	2.1 Frequency Distributions, Histograms, and Related Topics 40	
	2.2 Bar Graphs, Circle Graphs, and Time-Series Graphs 54	
	2.3 Stem-and-Leaf Displays 63	
	Summary 71	
	Important Words & Symbols 71	
	Chapter Review Problems 72	
	Data Highlights: Group Projects 75	
	Linking Concepts: Writing Projects 77	
	USING TECHNOLOGY	78
3	AVERAGES AND VARIATION	80
	FOCUS PROBLEM: The Educational Advantage	81
	3.1 Measures of Central Tendency: Mode, Median, and Mean 82	
	3.2 Measures of Variation 93	
	3.3 Percentiles and Box-and-Whisker Plots 110	
	Summary 120	
	Important Words & Symbols 120	
	Chapter Review Problems 121	
	Data Highlights: Group Projects 123	
	Linking Concepts: Writing Projects 125	
	USING TECHNOLOGY	126
	CUMULATIVE REVIEW PROBLEMS: <i>Chapters 1–3</i> 127	

4

ELEMENTARY PROBABILITY THEORY

130

FOCUS PROBLEM: How Often Do Lie Detectors Lie? 131

- 4.1 What Is Probability? 132
 4.2 Some Probability Rules—Compound Events 142
 4.3 Trees and Counting Techniques 162
 Summary 172
 Important Words & Symbols 173
 Chapter Review Problems 174
 Data Highlights: Group Projects 176
 Linking Concepts: Writing Projects 178

USING TECHNOLOGY 179

5

THE BINOMIAL PROBABILITY DISTRIBUTION AND RELATED TOPICS

180

FOCUS PROBLEM: Personality Preference Types: Introvert or Extrovert? 181

- 5.1 Introduction to Random Variables and Probability Distributions 182
 5.2 Binomial Probabilities 195
 5.3 Additional Properties of the Binomial Distribution 210
 5.4 The Geometric and Poisson Probability Distributions 222
 Summary 239
 Important Words & Symbols 240
 Chapter Review Problems 241
 Data Highlights: Group Projects 244
 Linking Concepts: Writing Projects 245

USING TECHNOLOGY 247

6

NORMAL CURVES AND SAMPLING DISTRIBUTIONS

248

FOCUS PROBLEM: Impulse Buying 249

- 6.1 Graphs of Normal Probability Distributions 250
 6.2 Standard Units and Areas Under the Standard Normal Distribution 266
 6.3 Areas Under Any Normal Curve 276
 6.4 Sampling Distributions 291
 6.5 The Central Limit Theorem 296
 6.6 Normal Approximation to Binomial Distribution and to \hat{p} Distribution 308
 Summary 318
 Important Words & Symbols 319
 Chapter Review Problems 319
 Data Highlights: Group Projects 322
 Linking Concepts: Writing Projects 323

USING TECHNOLOGY 325CUMULATIVE REVIEW PROBLEMS: **Chapters 4–6** 329

7

ESTIMATION

332

FOCUS PROBLEM: The Trouble with Wood Ducks 333

- 7.1 Estimating μ When σ Is Known 334
7.2 Estimating μ When σ Is Unknown 347
7.3 Estimating p in the Binomial Distribution 360
7.4 Estimating $\mu_1 - \mu_2$ and $p_1 - p_2$ 372
 Summary 395
 Important Words & Symbols 395
 Chapter Review Problems 396
 Data Highlights: Group Projects 400
 Linking Concepts: Writing Projects 402

USING TECHNOLOGY 404

8

HYPOTHESIS TESTING

408

FOCUS PROBLEM: Benford's Law: The Importance of Being Number 1 409

- 8.1 Introduction to Statistical Tests 410
8.2 Testing the Mean μ 425
8.3 Testing a Proportion p 442
8.4 Tests Involving Paired Differences (Dependent Samples) 452
8.5 Testing $\mu_1 - \mu_2$ and $p_1 - p_2$ (Independent Samples) 466
 Summary 490
 Finding the P-Value Corresponding to a Sample Test Statistic 491
 Important Words & Symbols 491
 Chapter Review Problems 492
 Data Highlights: Group Projects 495
 Linking Concepts: Writing Projects 496

USING TECHNOLOGY 497

9

CORRELATION AND REGRESSION

500

FOCUS PROBLEM: Changing Populations and Crime Rate 501

- 9.1 Scatter Diagrams and Linear Correlation 502
9.2 Linear Regression and the Coefficient of Determination 520
9.3 Inferences for Correlation and Regression 541
9.4 Multiple Regression 559
 Summary 575
 Important Words & Symbols 575
 Chapter Review Problems 576
 Data Highlights: Group Projects 579
 Linking Concepts: Writing Projects 580

USING TECHNOLOGY 581CUMULATIVE REVIEW PROBLEMS: **Chapters 7-9** 586

10**CHI-SQUARE AND *F* DISTRIBUTIONS****590****FOCUS PROBLEM: Archaeology in Bandelier National Monument 591****Part I: Inferences Using the Chi-Square Distribution 592**

Overview of the Chi-Square Distribution 592

10.1 Chi-Square: Tests of Independence and of Homogeneity 593

10.2 Chi-Square: Goodness of Fit 608

10.3 Testing and Estimating a Single Variance or Standard Deviation 618

Part II: Inferences Using the *F* Distribution 630Overview of the *F* Distribution 630

10.4 Testing Two Variances 631

10.5 One-Way ANOVA: Comparing Several Sample Means 640

10.6 Introduction to Two-Way ANOVA 656

Summary 668

Important Words & Symbols 668

Chapter Review Problems 669

Data Highlights: Group Projects 672

Linking Concepts: Writing Projects 673

USING TECHNOLOGY 674**11****NONPARAMETRIC STATISTICS****676****FOCUS PROBLEM: How Cold? Compared to What? 677**

11.1 The Sign Test for Matched Pairs 678

11.2 The Rank-Sum Test 686

11.3 Spearman Rank Correlation 694

11.4 Runs Test for Randomness 705

Summary 714

Important Words & Symbols 714

Chapter Review Problems 714

Data Highlights: Group Projects 716

Linking Concepts: Writing Projects 717

CUMULATIVE REVIEW PROBLEMS: Chapters 10-11 718**APPENDIX I: ADDITIONAL TOPICS****A1**

Part I: Bayes's Theorem A1

Part II: The Hypergeometric Probability Distribution A5

APPENDIX II: TABLES**A9**

Table 1: Random Numbers A9; **Table 2:** Binomial Coefficients $C_{n,r}$ A10; **Table 3:** Binomial Probability Distribution $C_{n,r} p^r q^{n-r}$ A11; **Table 4:** Poisson Probability Distribution A16; **Table 5:** Areas of a Standard Normal Distribution A22; **Table 6:** Critical Values for Student's *t* Distribution A24; **Table 7:** The χ^2 Distribution A25; **Table 8:** Critical Values for *F* Distribution A26; **Table 9:** Critical Values for Spearman Rank Correlation, r_s A36; **Table 10:** Critical Values for Number of Runs *R* A37

Answers and Key Steps to Odd-Numbered Problems A39

Answers to Selected Even-Numbered Problems A74

Index I1

Critical Thinking

Students need to develop critical thinking skills in order to understand and evaluate the limitations of statistical methods. *Understandable Statistics: Concepts and Methods* makes students aware of method appropriateness, assumptions, biases, and justifiable conclusions.

CRITICAL THINKING

Bias and Variability

Whenever we use a sample statistic as an estimate of a population parameter, we need to consider both *bias* and *variability* of the statistic.

A sample statistic is **unbiased** if the mean of its sampling distribution equals the value of the parameter being estimated.

The spread of the sampling distribution indicates the **variability of the statistic**. The spread is affected by the sampling method and the sample size. Statistics from larger random samples have spreads that are smaller.

We see from the central limit theorem that the sample mean \bar{x} is an unbiased estimator of the mean μ when $n \geq 30$. The variability of \bar{x} decreases as the sample size increases.

In Section 6.6, we will see that the sample proportion \hat{p} is an unbiased estimator of the population proportion of successes p in binomial experiments with sufficiently large numbers of trials n . Again, we will see that the variability of \hat{p} decreases with increasing numbers of trials.

The sample variance s^2 is an unbiased estimator for the population variance σ^2 .

◀ Critical Thinking

Critical thinking is an important skill for students to develop in order to avoid reaching misleading conclusions. The Critical Thinking feature provides additional clarification on specific concepts as a safeguard against incorrect evaluation of information.

Interpretation ►

Increasingly, calculators and computers are used to generate the numeric results of a statistical process. However, the student still needs to correctly interpret those results in the context of a particular application. The Interpretation feature calls attention to this important step. Interpretation is stressed in examples, guided exercises, and in the problem sets.

1, 22, and 23, we'll apply the theorem to solve a sum of random

The probability is 0.9988 that \bar{x} is between 2350 and 2650.

(c) **Interpretation** At the end of each day, the inspector must decide to accept or reject the accumulated milk that has been held in cold storage awaiting shipment. Suppose the 42 samples taken by the inspector have a mean bacteria count \bar{x} that is *not* between 2350 and 2650. If you were the inspector, what would your comment on this situation?

SOLUTION: The probability that \bar{x} is between 2350 and 2650 for milk that is not contaminated is very high. If the inspector finds that the average bacteria count for the 42 samples is not between 2350 and 2650, then it is reasonable to conclude that there is something wrong with the milk. If \bar{x} is less than 2350, you might suspect someone added chemicals to the milk to artificially reduce the bacteria count. If \bar{x} is above 2650, you might suspect some other kind of biologic contamination.

11. **Interpretation** A job-performance evaluation form has these categories:
1 = excellent; 2 = good; 3 = satisfactory; 4 = poor; 5 = unacceptable
Based on 15 client reviews, one employee had
median rating of 4; mode rating of 1
The employee was pleased that most clients had rated her as excellent. The supervisor said improvement was needed because at least half the clients had rated the employee at the poor or unacceptable level. Comment on the different perspectives.
12. **Critical Thinking: Data Transformation** In this problem, we explore the effect on the mean, median, and mode of adding the same number to each data value. Consider the data set 2, 2, 3, 6, 10.
- Compute the mode, median, and mean.
 - Add 5 to each of the data values. Compute the mode, median, and mean.
 - Compare the results of parts (a) and (b). In general, how do you think the mode, median, and mean are affected when the same constant is added to each data value in a set?

◀ NEW! Critical Thinking and Interpretation Exercises

In every section and chapter problem set, Critical Thinking problems provide students with the opportunity to test their understanding of the application of statistical methods and their interpretation of their results. Interpretation problems ask students to apply statistical results to the particular application.

Statistical Literacy

No language can be spoken without learning the vocabulary, including statistics. *Understandable Statistics: Concepts and Methods* introduces statistical terms with deliberate care.

SECTION 6.1 PROBLEMS

1. **Statistical Literacy** Which, if any, of the curves in Figure 6-10 look(s) like a normal curve? If a curve is not a normal curve, tell why.
2. **Statistical Literacy** Look at the normal curve in Figure 6-11, and find μ , $\mu + \sigma$, and σ .

FIGURE 6-10

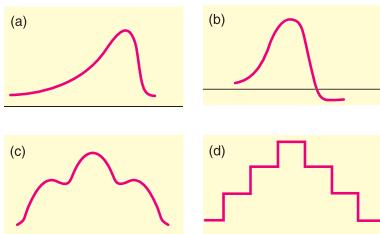
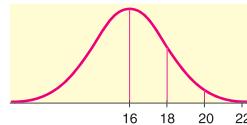


FIGURE 6-11



◀ Statistical Literacy Problems

In every section and chapter problem set, Statistical Literacy problems test student understanding of terminology, statistical methods, and the appropriate conditions for use of the different processes.

Definition Boxes ▶

Whenever important terms are introduced in text, yellow definition boxes appear within the discussions. These boxes make it easy to reference or review terms as they are used further.

Five-number summary

Box-and-Whisker Plots

The quartiles together with the low and high data values give us a very useful five-number summary of the data and their spread.

Five-number summary

Lowest value, Q_1 , median, Q_3 , highest value

We will use these five numbers to create a graphic sketch of the data called a *box-and-whisker plot*. Box-and-whisker plots provide another useful technique from exploratory data analysis (EDA) for describing data.

IMPORTANT WORDS & SYMBOLS

Section 4.1

Probability of an event A , $P(A)$ 132
Intuition 132
Relative frequency 132
Equally likely outcomes 132
Law of large numbers 134
Statistical experiment 134
Event 134
Simple event 134
Sample space 134
Complement of event A^c 136

$P(A \mid B)$ 143

More than two independent events 147
Probability of A or B 147
Event A and B 147
Event A or B 147
Mutually exclusive events 149
Addition rules (for mutually exclusive and general events) 149
More than two mutually exclusive events 151
Basic probability rules 154

Section 4.2

Independent events 143
Dependent events 143
Probability of A and B 143
Event $A \mid B$ 143
Conditional probability 143
Multiplication rules of probability (for independent and dependent events) 143

Section 4.3

Multiplication rule of counting 163
Tree diagram 163
Factorial notation 166
Permutations rule 167
Combinations rule 168

▲ REVISED! Important Words & Symbols

The Important Words & Symbols within the Chapter Review feature at the end of each chapter summarizes the terms introduced in the Definition Boxes for student review at a glance. Page numbers for first occurrence of term are given for easy reference.

Statistical Literacy

Linking Concepts: Writing Projects ►

Much of statistical literacy is the ability to communicate concepts effectively. The Linking Concepts: Writing Projects feature at the end of each chapter tests both statistical literacy and critical thinking by asking the student to express their understanding in words.

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. An average is an attempt to summarize a collection of data into just *one* number. Discuss how the mean, median, and mode all represent averages in this context. Also discuss the differences among these averages. Why is the mean a balance point? Why is the median a midway point? Why is the mode the most common data point? List three areas of daily life in which you think one of the mean, median, or mode would be the best choice to describe an “average.”
2. Why do we need to study the variation of a collection of data? Why isn’t the average by itself adequate? We have studied three ways to measure variation. The range, the standard deviation, and, to a large extent, a box-and-whisker plot all indicate the variation within a data collection. Discuss similarities and differences among these ways to measure data variation. Why would it seem reasonable to pair the median with a box-and-whisker plot and to pair the mean with the standard deviation? What are the advantages and disadvantages of each method of describing data spread? Comment on statements such as the following: (a) The range is easy to compute, but it doesn’t give much information; (b) although the standard deviation is more complicated to compute, it has some significant applications; (c) the box-and-whisker plot is fairly easy to construct, and it gives a lot of information at a glance.

9. **Basic Computation: Testing μ , σ Unknown** A random sample of 25 values is drawn from a mound-shaped and symmetric distribution. The sample mean is 10 and the sample standard deviation is 2. Use a level of significance of 0.05 to conduct a two-tailed test of the claim that the population mean is 9.5.
(a) **Check Requirements** Is it appropriate to use a Student’s t distribution?
Explain. How many degrees of freedom do we use?
(b) What are the hypotheses?
(c) Compute the sample test statistic t .
(d) Estimate the P -value for the test.
(e) Do we reject or fail to reject H_0 ?
(f) **Interpret** the results.

◀ NEW! Basic Computation Problems

These problems focus student attention on relevant formulas, requirements, and computational procedures. After practicing these skills, students are more confident as they approach real-world applications.

Expand Your Knowledge Problems ►

Expand Your Knowledge problems present optional enrichment topics that go beyond the material introduced in a section. Vocabulary and concepts needed to solve the problems are included at point-of-use, expanding students’ statistical literacy.



17. (b) Suppose the EPA has established an average chlorine compound concentration target of no more than 58 mg/l. Comment on whether this wetlands system meets the target standard for chlorine compound concentration.



18. (b) **Expand Your Knowledge: Harmonic Mean** When data consist of rates of change, such as speeds, the *harmonic mean* is an appropriate measure of central tendency, for n data values,

$$\text{Harmonic mean} = \frac{n}{\sum_x^1}, \text{ assuming no data value is 0}$$

Suppose you drive 60 miles per hour for 100 miles, then 75 miles per hour for 100 miles. Use the harmonic mean to find your average speed.

18. (c) **Expand Your Knowledge: Geometric Mean** When data consist of percentages, ratios, growth rates, or other rates of change, the *geometric mean* is a useful measure of central tendency. For n data values,

$$\text{Geometric mean} = \sqrt[n]{\text{product of the } n \text{ data values}}, \text{ assuming all data values are positive}$$

To find the *average growth factor* over 5 years of an investment in a mutual fund with growth rates of 10% the first year, 12% the second year, 14.8% the third year, 3.8% the fourth year, and 6% the fifth year, take the geometric mean of 1.10, 1.12, 1.148, 1.038, and 1.16. Find the average growth factor of this investment.

Note that for the same data, the relationships among the harmonic, geometric, and arithmetic means are harmonic mean \leq geometric mean \leq arithmetic mean (Source: *Oxford Dictionary of Statistics*).

Direction and Purpose

Real knowledge is delivered through direction, not just facts. *Understandable Statistics: Concepts and Methods* ensures the student knows what is being covered and why at every step along the way to statistical literacy.

NORMAL CURVES AND SAMPLING DISTRIBUTIONS

Chapter Preview ► Questions

Preview Questions at the beginning of each chapter give the student a taste of what types of questions can be answered with an understanding of the knowledge to come.

PREVIEW QUESTIONS

What are some characteristics of a normal distribution? What does the empirical rule tell you about data spread around the mean? How can this information be used in quality control? (SECTION 6.1)

Can you compare apples and oranges, or maybe elephants and butterflies? In most cases, the answer is no—unless you first standardize your measurements. What are a standard normal distribution and a standard z score? (SECTION 6.2)

How do you convert any normal distribution to a standard normal distribution? How do you find probabilities of “standardized events”? (SECTION 6.3)

As humans, our experiences are finite and limited. Consequently, most of the important decisions in our lives are based on sample (incomplete) information. What is a probability sampling distribution? How will sampling distributions help us make good decisions based on incomplete information? (SECTION 6.4)

There is an old saying: All roads lead to Rome. In statistics, we could recast this saying: All probability distributions average out to be normal distributions (as the sample size increases). How can we take advantage of this in our study of sampling distributions? (SECTION 6.5)

The binomial and normal distributions are two of the most important probability distributions in statistics. Under certain limiting conditions, the binomial can be thought to evolve (or envelope) a concept in the real world? (SECTION 6.6)



Dana White/PhotoEdit

FOCUS PROBLEM

Benford's Law: The Importance of Being Number 1

Benford's Law states that in a wide variety of circumstances, numbers have "1" as their first nonzero digit disproportionately often. Benford's Law applies to such diverse topics as the drainage areas of rivers; properties of chemicals; populations of towns; figures in newspapers, magazines, and government reports; and the half-lives of radioactive atoms!

Specifically, such diverse measurements begin with "1" about 30% of the time, with "2" about 18% of time, and with "3" about 12.5% of the time. Larger digits occur less often. For example, less than 5% of the numbers in circumstances such as these begin with the digit 9. This is in dramatic contrast to a random sampling situation, in which each of the digits 1 through 9 has an equal chance of appearing.

The first nonzero digits of numbers taken from large bodies of numerical records such as tax returns, population studies, government records, and so forth, show the probabilities of occurrence as displayed in the table on the next page.



Corbis

8. **Focus Problem: Benford's Law** Again, suppose you are the auditor for a very large corporation. The revenue file contains millions of numbers in a large computer data bank (see Problem 7). You draw a random sample of $n = 228$ numbers from this file and $r = 92$ have a first nonzero digit of 1. Let p represent the population proportion of all numbers in the computer file that have a leading digit of 1.
 - i. Test the claim that p is more than 0.301. Use $\alpha = 0.01$.
 - ii. If p is in fact larger than 0.301, it would seem there are too many numbers in the file with leading 1's. Could this indicate that the books have been "cooked" by artificially lowering numbers in the file? Comment from the point of view of the Internal Revenue Service. Comment from the perspective of the Federal Bureau of Investigation as it looks for "profit skimming" by unscrupulous employees.
 - iii. Comment on the following statement: "If we reject the null hypothesis at level of significance α , we have not *proved* H_0 to be false. We can say that the probability is α that we made a mistake in rejecting H_0 ." Based on the outcome of the test, would you recommend further investigation before accusing the company of fraud?

▲ Chapter Focus Problems

The Preview Questions in each chapter are followed by Focus Problems, which serve as more specific examples of what questions the student will soon be able to answer. The Focus Problems are set within appropriate applications and are incorporated into the end-of-section exercises, giving students the opportunity to test their understanding.

Direction and Purpose

Focus Points ►

Each section opens with bulleted Focus Points describing the primary learning objectives of the section.

SECTION 3.1

Measures of Central Tendency: Mode, Median, and Mean

FOCUS POINTS

- Compute mean, median, and mode from raw data.
- Interpret what mean, median, and mode tell you.
- Explain how mean, median, and mode can be affected by extreme data values.
- What is a trimmed mean? How do you compute it?
- Compute a weighted average.

The average price of an ounce of gold is \$1200. The Zippy car averages 39 miles per gallon on the highway. A survey showed the average shoe size for women is size 9.

In each of the preceding statements, *one* number is used to describe the entire sample or population. Such a number is called an *average*. There are many ways to compute averages, but we will study only three of the major ones.

The easiest average to compute is the *mode*.

Average

Mode

The mode of a data set is the value that occurs most frequently.



EXAMPLE 1

MODE

Count the letters in each word of this sentence and give the mode. The numbers of letters in the words of the sentence are

5 3 7 2 4 4 2 4 8 3 4 3 4

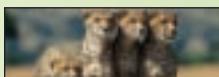
Scanning the data, we see that 4 is the mode because more words have 4 letters than any other number. For larger data sets, it is useful to order—or sort—the data before scanning them for the mode.

LOOKING FORWARD

In our future work with inferential statistics, we will use the mean \bar{x} from a random sample to estimate the population parameter μ (Chapter 7) or to make decisions regarding the value of μ (Chapter 8).

▲ NEW! Looking Forward

This feature shows students where the presented material will be used later. It helps motivate students to pay a little extra attention to key topics.



Chapter Review

SUMMARY

Organizing and presenting data are the main purposes of the branch of statistics called descriptive statistics. Graphs provide an important way to show how the data are distributed.

- Frequency tables show how the data are distributed within set classes. The classes are chosen so that they cover all data values and so that each data value falls within only one class. The number of classes and the class width determine the class limits and class boundaries. The number of data values falling within a class is the class frequency.
- A histogram is a graphical display of the information in a frequency table. Classes are shown on the horizontal axis, with corresponding frequencies on the vertical axis. Relative-frequency histograms show relative

frequencies on the vertical axis. Ogives show cumulative frequencies on the vertical axis. Dotplots are like histograms, except that the classes are individual data values.

- Bar graphs, Pareto charts, and pie charts are useful to show how quantitative or qualitative data are distributed over chosen categories.
- Time-series graphs show how data change over set intervals of time.
- Stem-and-leaf displays are effective means of ordering data and showing important features of the distribution.

Graphs aren't just pretty pictures. They help reveal important properties of the data distribution, including the shape and whether or not there are any outliers.

◀ Chapter Summaries

The Summary within each Chapter Review feature now also appears in bulleted form, so students can see what they need to know at a glance.

Real-World Skills

Statistics is not done in a vacuum. *Understandable Statistics: Concepts and Methods* gives students valuable skills for the real world with technology instruction, genuine applications, actual data, and group projects.

REVISED! Tech Notes ►

Tech Notes appearing throughout the text give students helpful hints on using TI-84 Plus and TI-nspire (with 84 Plus keypad) and TI-83 calculators, Microsoft Excel 2007, and Minitab to solve a problem. They include display screens to help students visualize and better understand the solution.



Stem-and-leaf display

TI-84Plus/TI-83Plus/TI-nspire Does not support stem-and-leaf displays. You can sort the data by using keys **Stat ▶ Edit ▶ 2:SortA**.

Excel 2007 Enter your data and select the data you want to sort. On the Home ribbon, click the Sort and Filter button in the Editing group of the ribbon and select the desired sorting option.

Minitab Use the menu selections **Graph ▶ Stem-and-Leaf** and fill in the dialogue box.

Minitab Stem-and-Leaf Display (for Data in Guided Exercise 4)

Stem-and-Leaf of Scores N=35		
Leaf Unit=1.0	1	8 3
	5	9 2789
	14	10 123455669
(11)		11 01222267789
	10	12 045568
	4	13 125
	1	14 3

The values shown in the left column represent depth. Numbers above the value in parentheses show the cumulative number of values from the top to the stem of the middle value. Numbers below the value in parentheses show the cumulative number of values from the bottom to the stem of the middle value. The number in parentheses shows how many values are on the same line as the middle value.

USING TECHNOLOGY

Binomial Distributions

Although tables of binomial probabilities can be found in most libraries, such tables are often inadequate. Either the value of p (the probability of success on a trial) you are looking for is not in the table, or the value of n (the number of trials) you are looking for is too large for the table. In Chapter 6, we will study the normal approximation to the binomial. This approximation is a great help in many practical applications. Even so, we sometimes use the formula for the binomial probability distribution on a computer or graphing calculator to compute the probability we want.

Applications

The following percentages were obtained over many years of observation by the U.S. Weather Bureau. All data listed are for the month of December.

Location	Long-Term Mean % of Clear Days in Dec.
Juneau, Alaska	18%
Seattle, Washington	24%
Hilo, Hawaii	36%
Honolulu, Hawaii	60%
Las Vegas, Nevada	75%
Phoenix, Arizona	77%

Adapted from *Local Climatological Data*, U.S. Weather Bureau publication, "Normals, Means, and Extremes" Table.

In the locations listed, the month of December is a relatively stable month with respect to weather. Since weather patterns from one day to the next are more or less the same, it is reasonable to use a binomial probability model.

- Let r be the number of clear days in December. Since December has 31 days, $0 \leq r \leq 31$. Using appropriate computer software or calculators available to you, find the probability $P(r)$ for each of the listed locations when $r = 0, 1, 2, \dots, 31$.
- For each location, what is the expected value of the probability distribution? What is the standard deviation?

You may find that using cumulative probabilities and appropriate subtraction of probabilities, rather than addition of probabilities, will make finding the solutions to Applications 3 to 7 easier.

- Estimate the probability that Juneau will have at most 7 clear days in December.
- Estimate the probability that Seattle will have from 5 to 10 (including 5 and 10) clear days in December.
- Estimate the probability that Hilo will have at least 12 clear days in December.
- Estimate the probability that Phoenix will have 20 or more clear days in December.
- Estimate the probability that Las Vegas will have from 20 to 25 (including 20 and 25) clear days in December.

Technology Hints

TI-84Plus/TI-83Plus/TI-nspire (with TI-84 Plus keypad), Excel 2007, Minitab

The Tech Note in Section 5.2 gives specific instructions for binomial distribution functions on the TI-84plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) calculators, Excel 2007, and Minitab.

SPSS

In SPSS, the function `PDF.BINOM(q,n,p)` gives the probability of q successes out of n trials, where p is the probability of success on a single trial. In the data editor, name a variable r and enter values 0 through n . Name another variable $Prob_r$. Then use the menu choices **Transform ▶ Compute**. In the dialogue box, use $Prob_r$ for the target variable. In the function group, select **PDF and Noncentral PDF**. In the function box, select `PDF.BINOM(q,n,p)`. Use the variable r for q and appropriate values for n and p . Note that the function `CDF.BINOM(q,n,p)`, from the CDF and Noncentral CDF group, gives the cumulative probability of 0 through q successes.

◀ REVISED! Using Technology

Further technology instruction is available at the end of each chapter in the Using Technology section. Problems are presented with real-world data from a variety of disciplines that can be solved by using TI-84 Plus, TI-nspire (with 84 Plus keypad) and TI-83 Plus calculators, Microsoft Excel 2007, and Minitab.

Real-World Skills

EXAMPLE 13

CENTRAL LIMIT THEOREM

A certain strain of bacteria occurs in all raw milk. Let x be the bacteria count per milliliter of milk. The health department has found that if the milk is not contaminated, then x has a distribution that is more or less mound-shaped and symmetrical. The mean of the x distribution is $\mu = 2500$, and the standard deviation is $\sigma = 300$. In a large commercial dairy, the health inspector takes 42 random samples of the milk produced each day. At the end of the day, the bacteria count in each of the 42 samples is averaged to obtain the sample mean bacteria count \bar{x} .

(a) Assuming the milk is not contaminated, what is the distribution of \bar{x} ?

SOLUTION: The sample size is $n = 42$. Since this value exceeds 30, the central limit theorem applies, and we know that \bar{x} will be approximately normal, with mean and standard deviation

Most exercises in each section ►
are applications problems.

DATA HIGHLIGHTS: GROUP PROJECTS



Old Faithful Geyser, Yellowstone National Park

FIGURE 3-11

Typical Behavior of Old Faithful Geyser Before 1959 Quake

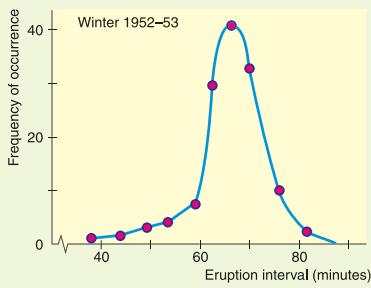
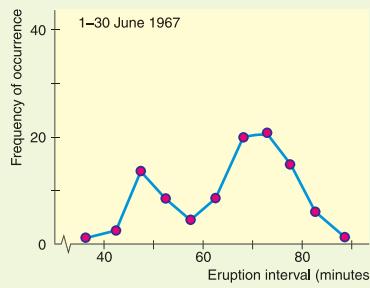


FIGURE 3-12

Typical Behavior of Old Faithful Geyser After 1959 Quake



◀ UPDATED! Applications

Real-world applications are used from the beginning to introduce each statistical process. Rather than just crunching numbers, students come to appreciate the value of statistics through relevant examples.

entered about the mean in which almost all the data points are found.

10. **Vending Machine: Soft Drinks** A vending machine automatically pours soft drinks into cups. The amount of soft drink dispensed into a cup is normally distributed with a mean of 7.6 ounces and standard deviation of 0.4 ounce. Examine Figure 6-3 and answer the following questions.
(a) Estimate the probability that the machine will overflow an 8-ounce cup.
(b) Estimate the probability that the machine will not overflow an 8-ounce cup.
(c) The machine has just been loaded with 850 cups. How many of these do you expect will overflow when served?
11. **Pain Management: Laser Therapy** "Effect of Helium-Neon Laser Auriculotherapy on Experimental Pain Threshold" is the title of an article in the journal *Physical Therapy* (Vol. 70, No. 1, pp. 24–30). In this article, laser therapy was discussed as a useful alternative to drugs in pain management of chronically ill patients. To measure pain threshold, a machine was used that delivered low-voltage direct current to different parts of the body (wrist, neck, and back). The machine measured current in milliamperes (mA). The pretreatment experimental group in the study had an average threshold of pain (pain was first detectable) at $\mu = 3.15$ mA with

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

1. *The Story of Old Faithful* is a short book written by George Marler and published by the Yellowstone Association. Chapter 7 of this interesting book talks about the effect of the 1959 earthquake on eruption intervals for Old Faithful Geyser. Dr. John Rinehart (a senior research scientist with the National Oceanic and Atmospheric Administration) has done extensive studies of the eruption intervals before and after the 1959 earthquake. Examine Figure 3-11. Notice the general shape. Is the graph more or less symmetrical? Does it have a single mode or frequency? The mean interval between eruptions has remained steady at about 65 minutes for the past 100 years. Therefore, the 1959 earthquake did not significantly change the mean, but it did change the distribution of eruption intervals. Examine Figure 3-12. Would you say there are really two frequency modes, one shorter and the other longer? Explain. The overall mean is about the same for both graphs, but one graph has a much larger standard deviation (for eruption intervals) than the other. Do no calculations, just look at both graphs, and then explain which graph has the smaller and which has the larger standard deviation. Which distribution will have the larger coefficient of variation? In everyday terms, what would this mean if you were actually at Yellowstone waiting to see the next eruption of Old Faithful? Explain your answer.

◀ Data Highlights: Group Projects

Using Group Projects, students gain experience working with others by discussing a topic, analyzing data, and collaborating to formulate their response to the questions posed in the exercise.

Making the Jump

Get to the "Aha!" moment faster. *Understandable Statistics: Concepts and Methods* provides the push students need to get there through guidance and example.

PROCEDURE

HOW TO TEST μ WHEN σ IS UNKNOWN

Requirements

Let x be a random variable appropriate to your application. Obtain a simple random sample (of size n) of x values from which you compute the sample mean \bar{x} and the sample standard deviation s . If you can assume that x has a normal distribution or simply a mound-shaped and symmetric distribution, then any sample size n will work. If you cannot assume this, use a sample size $n \geq 30$.

Procedure

1. In the context of the application, state the *null and alternate hypotheses* and set the *level of significance α* .
2. Use \bar{x} , s , and n from the sample, with μ from H_0 , to compute the *sample test statistic*.
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad \text{with degrees of freedom } d.f. = n - 1$$
3. Use the Student's t distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If $P\text{-value} \leq \alpha$, then reject H_0 . If $P\text{-value} > \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.

d.f. for testing μ when σ unknown

◀ REVISED! Procedures and Requirements

Procedure display boxes summarize simple step-by-step strategies for carrying out statistical procedures and methods as they are introduced. Requirements for using the procedures are also stated. Students can refer back to these boxes as they practice using the procedures.

Guided Exercises ▶

Students gain experience with new procedures and methods through Guided Exercises. Beside each problem in a Guided Exercise, a completely worked-out solution appears for immediate reinforcement.

GUIDED EXERCISE 11

Probability regarding \bar{x}

In mountain country, major highways sometimes use tunnels instead of long, winding roads over high passes. However, too many vehicles in a tunnel at the same time can cause a hazardous situation. Traffic engineers are studying a long tunnel in Colorado. If x represents the time for a vehicle to go through the tunnel, it is known that the x distribution has mean $\mu = 12.1$ minutes and standard deviation $\sigma = 3.8$ minutes under ordinary traffic conditions. From a histogram of x values, it was found that the x distribution is mound-shaped with some symmetry about the mean.

Engineers have calculated that, *on average*, vehicles should spend from 11 to 13 minutes in the tunnel. If the time is less than 11 minutes, traffic is moving too fast for safe travel in the tunnel. If the time is more than 13 minutes, there is a problem of bad air quality (too much carbon monoxide and other pollutants).

Under ordinary conditions, there are about 50 vehicles in the tunnel at one time. What is the probability that the mean time for 50 vehicles in the tunnel will be from 11 to 13 minutes? We will answer this question in steps.

- (a) Let \bar{x} represent the sample mean based on samples of size 50. Describe the \bar{x} distribution.

From the central limit theorem, we expect the \bar{x} distribution to be approximately normal, with mean and standard deviation

$$\mu_{\bar{x}} = \mu = 12.1 \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.8}{\sqrt{50}} \approx 0.54$$

- (b) Find $P(11 < \bar{x} < 13)$.

We convert the interval

$$11 < \bar{x} < 13$$

to a standard z interval and use the standard normal probability table to find our answer. Since

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 12.1}{0.54}$$

$$\bar{x} = 11 \text{ converts to } z \approx \frac{11 - 12.1}{0.54} = -2.04$$

$$\text{and } \bar{x} = 13 \text{ converts to } z \approx \frac{13 - 12.1}{0.54} = 1.67$$

Therefore,

$$\begin{aligned} P(11 < \bar{x} < 13) &= P(-2.04 < z < 1.67) \\ &= 0.9525 - 0.0207 \\ &= 0.9318 \end{aligned}$$

- (c) *Interpret* your answer to part (b).

It seems that about 93% of the time, there should be no safety hazard for average traffic flow.



Jupiter Images



PREFACE

Welcome to the exciting world of statistics! We have written this text to make statistics accessible to everyone, including those with a limited mathematics background. Statistics affects all aspects of our lives. Whether we are testing new medical devices or determining what will entertain us, applications of statistics are so numerous that, in a sense, we are limited only by our own imagination in discovering new uses for statistics.

Overview

The tenth edition of *Understandable Statistics: Concepts and Methods* continues to emphasize concepts of statistics. Statistical methods are carefully presented with a focus on understanding both the *suitability of the method* and the *meaning of the result*. Statistical methods and measurements are developed in the context of applications.

Critical thinking and interpretation are essential in understanding and evaluating information. Statistical literacy is fundamental for applying and comprehending statistical results. In this edition we have expanded and highlighted the treatment of statistical literacy, critical thinking, and interpretation.

We have retained and expanded features that made the first nine editions of the text very readable. Definition boxes highlight important terms. Procedure displays summarize steps for analyzing data. Examples, exercises, and problems touch on applications appropriate to a broad range of interests.

New with the tenth edition is CourseMate, encompassing all interactive online products and services with this text. Online homework powered by a choice of Enhanced WebAssign or Aplia is now available through CengageBrain.com. Also available through CourseMate are over 100 data sets (in Microsoft Excel, Minitab, SPSS, and TI-84Plus/TI-83Plus/TI-nspire with 84plus keypad ASCII file formats), lecture aids, a glossary, statistical tables, instructional video (also available on DVDs), an Online Multimedia eBook, and interactive tutorials.

Major Changes in the Tenth Edition

With each new edition, the authors reevaluate the scope, appropriateness, and effectiveness of the text's presentation and reflect on extensive user feedback. Revisions have been made throughout the text to clarify explanations of important concepts and to update problems.

Critical Thinking, Interpretation, and Statistical Literacy

The tenth edition of this text continues and expands the emphasis on critical thinking, interpretation, and statistical literacy. Calculators and computers are very good at providing numerical results of statistical processes. However, numbers from a computer or calcultor display are meaningless unless the user knows how to interpret the results and if the statistical process is appropriate. This text helps students determine whether or not a statistical method or process is appropriate. It helps students understand what a statistic measures. It helps students interpret the results of a confidence interval, hypothesis test, or liner regression model.

New Problems Featuring Basic Computation

Calculators and computer software automatically calculate designated statistical measurements. However, students gain an appreciation and understanding of what the measurements mean by studying and using the basic formulas. Basic computation problems focus attention on using formulas with small data sets. Students see not only how the formulas work, but also how the resulting measurements relate to the displayed data set.

There are more than 200 new and revised problems that feature basic computations, interpretation, and statistical literacy. Students are asked to check that the use of specific probability distributions and inferential methods are appropriate.

Normal Distributions and Sampling Distributions in Same Chapter

Chapter 6 of the tenth edition includes both an introduction to normal distributions as well as an introduction to sampling distributions. Putting both topics in the same chapter streamlines the course, and gives an immediate, important application of normal distributions. The chapter also includes the normal approximation to the binomial distribution.

New Content

The uniform probability distribution and the exponential probability distribution are introduced in Expand Your Knowledge problems in Chapter 6. Polynomial regression (also known as curvilinear regression) is discussed in Expand Your Knowledge problems of Section 9.4 Multiple Regression.

Excel 2007 and TI-nspire Calculator (with 84plus keypad)

Excel 2007 instructions are included in the *Tech Notes* and *Using Technology* features.

The TI-nspire calculator with the 84Plus keypad is also addressed. Instructions for using the TI-nspire calculator with the nspire keypad are included in the separate Technology Guide for the TI-84Plus graphing calculator.

Looking Forward

A new *Looking Forward* feature briefly points out how current subject matter will be used in later chapters.

Other Changes

Examples in probability (Chapter 4) have been revised to incorporate sample spaces that are small and easy for students to visualize.

Throughout the tenth edition, examples, guided exercises, and problem sets feature interpretation. They also address the suitability of using a specified statistical method or process for analysis.

Chapter 6 Normal Curves and Sampling Distribution of the tenth edition combines material from Chapters 6 and 7 of the ninth edition. Chapter 7 Estimation, Chapter 8 Hypothesis Testing, Chapter 9 Correlation and Regression, Chapter 10 Chi-Square and F Distribution, and Chapter 11 Nonparametric Statistics of the tenth edition correspond respectively to Chapters 8–12 of the ninth edition.

Continuing Content

Introduction of Hypothesis Testing Using *P*-Values

In keeping with the use of computer technology and standard practice in research, hypothesis testing is introduced using *P*-values. The critical region method is still supported, but not given primary emphasis.

Use of Student's *t* Distribution in Confidence Intervals and Testing of Means

If the normal distribution is used in confidence intervals and testing of means, then the *population standard deviation must be known*. If the population standard deviation is not known, then under conditions described in the text, the Student's *t* distribution is used. This is the most commonly used procedure in statistical research. It is also used in statistical software packages such as Microsoft Excel, Minitab, SPSS, and TI-84Plus/TI-83Plus/TI-nspire calculators.

Confidence Intervals and Hypothesis Tests of Difference of Means

If the normal distribution is used, then both population standard deviations must be known. When this is not the case, the Student's *t* distribution incorporates an approximation for *t*, with a commonly used conservative choice for the degrees of freedom. Satterthwaite's approximation for the degrees of freedom as used in computer software is also discussed. The pooled standard deviation is presented for appropriate applications ($\sigma_1 \approx \sigma_2$).

Features in the Tenth Edition

Chapter and Section Lead-ins

- *Preview Questions* at the beginning of each chapter are keyed to the sections.
- *Focus Problems* at the beginning of each chapter demonstrate types of questions students can answer once they master the concepts and skills presented in the chapter.
- *Focus Points* at the beginning of each section describe the primary learning objectives of the section.

Carefully Developed Pedagogy

- *Examples* show students how to select and use appropriate procedures.
- *Guided Exercises* within the sections give students an opportunity to work with a new concept. Completely worked-out solutions appear beside each exercise to give immediate reinforcement.
- *Definition boxes* highlight important definitions throughout the text.
- *Procedure displays* summarize key strategies for carrying out statistical procedures and methods. Conditions required for using the procedure are also stated.
- NEW! *Looking Forward* features give a brief preview of how a current topic is used later.
- *Labels* for each example or guided exercise highlight the technique, concept, or process illustrated by the example or guided exercise. In addition, labels for

section and chapter problems describe the field of application and show the wide variety of subjects in which statistics is used.

- *Section and chapter problems* require the student to use all the new concepts mastered in the section or chapter. Problem sets include a variety of real-world applications with data or settings from identifiable sources. Key steps and solutions to odd-numbered problems appear at the end of the book.
- NEW! *Basic Computation problems* ask students to practice using formulas and statistical methods on very small data sets. Such practice helps students understand what a statistic measures.
- *Statistical Literacy problems* ask students to focus on correct terminology and processes of appropriate statistical methods. Such problems occur in every section and chapter problem set.
- NEW! *Interpretation problems* ask students to explain the meaning of the statistical results in the context of the application.
- *Critical Thinking problems* ask students to analyze and comment on various issues that arise in the application of statistical methods and in the interpretation of results. These problems occur in every section and chapter problem set.
- *Expand Your Knowledge problems* present enrichment topics such as negative binomial distribution; conditional probability utilizing binomial, Poisson, and normal distributions; estimation of standard deviation from a range of data values; and more.
- *Cumulative review problem sets* occur after every third chapter and include key topics from previous chapters. Answers to *all* cumulative review problems are given at the end of the book.
- *Data Highlights and Linking Concepts* provide group projects and writing projects.
- *Viewpoints* are brief essays presenting diverse situations in which statistics is used.
- *Design and photos* are appealing and enhance readability.

Technology within the Text

- *Tech Notes* within sections provide brief point-of-use instructions for the TI-84Plus, TI-83Plus and TI-nspire (with 84plus keypad) calculators, Microsoft Excel 2007 and Minitab.
- *Using Technology* sections have been revised to show the use of SPSS as well as the TI-84Plus, TI-83Plus and TI-nspire (with 84plus keypad) calculators, Microsoft Excel, and Minitab.

Alternate Routes Through the Text

Understandable Statistics: Concepts and Methods, Tenth Edition, is designed to be flexible. It offers the professor a choice of teaching possibilities. In most one-semester courses, it is not practical to cover all the material in depth. However, depending on the emphasis of the course, the professor may choose to cover various topics. For help in topic selection, refer to the Table of Prerequisite Material on page 1.

- *Introducing linear regression early.* For courses requiring an early presentation of linear regression, the descriptive components of linear regression (Sections 9.1 and 9.2) can be presented any time after Chapter 3. However, inference topics involving predictions, the correlation coefficient ρ , and the slope of the least-squares line β require an introduction to confidence intervals (Sections 7.1 and 7.2) and hypothesis testing (Sections 8.1 and 8.2).
- *Probability.* For courses requiring minimal probability, Section 4.1 (What Is Probability?) and the first part of Section 4.2 (Some Probability Rules—Compound Events) will be sufficient.

Acknowledgments

It is our pleasure to acknowledge the prepublication reviewers of this text. All of their insights and comments have been very valuable to us. Reviewers of this text include:

Reza Abbasian, Texas Lutheran University
Paul Ache, Kutztown University
Kathleen Almy, Rock Valley College
Polly Amstutz, University of Nebraska at Kearney
Delores Anderson, Truett-McConnell College
Robert J. Astalos, Feather River College
Lynda L. Ballou, Kansas State University
Mary Benson, Pensacola Junior College
Larry Bennett, Benedictine University
Kiran Bhutani, The Catholic University of America
Kristy E. Bland, Valdosta State University
John Bray, Broward Community College
Bill Burgin, Gaston College
Toni Carroll, Siena Heights University
Pinyuen Chen, Syracuse University
Emmanuel des-Bordes, James A. Rhodes State College
Jennifer M. Dollar, Grand Rapids Community College
Larry E. Dunham, Wor-Wic Community College
Andrew Ellett, Indiana University
Ruby Evans, Keiser University
Mary Fine, Moberly Area Community College
Rebecca Fouquette, Santa Rosa Junior College
Rene Garcia, Miami-Dade Community College
Larry Green, Lake Tahoe Community College
Shari Harris, John Wood Community College
Janice Hector, DeAnza College
Jane Keller, Metropolitan Community College
Raja Khoury, Collin County Community College
Diane Koenig, Rock Valley College
Charles G. Laws, Cleveland State Community College
Michael R. Lloyd, Henderson State University
Beth Long, Pellissippi State Technical and Community College
Lewis Lum, University of Portland
Darcy P. Mays, Virginia Commonwealth University
Charles C. Okeke, College of Southern Nevada, Las Vegas
Peg Pankowski, Community College of Allegheny County
Ram Polepeddi, Westwood college, Denver North Campus
Ron Spicer, Colorado Technical University
Azar Raiszadeh, Chattanooga State Technical Community College
Traei Reed, St.Johns River Community College
Michael L. Russo, Suffolk County Community College
Janel Schultz, Saint Mary's University of Minnesota
Sankara Sethuraman, Augusta State University
Stephen Soltys, West Chester university of Pennsylvania
Winson Taam, Oakland University
Jennifer L. Taggart, Rockford College
William Truman, University of North Carolina at Pembroke
Bill White, University of South Carolina Upstate
Jim Wienckowski, State University of New York at Buffalo

Stephen M. Wilkerson, Susquehanna University
Hongkai Zhang, East Central University
Shunpu Zhang, University of Alaska, Fairbanks
Cathy Zucco-Teveloff, Trinity College

We would especially like to thank George Pasles for his careful accuracy review of this text. We are especially appreciative of the excellent work by the editorial and production professionals at Brooks/Cole Cengage. In particular we thank Molly Taylor, Shaylin Walsh, Jill Clark, and Heather Johnson.

Without their creative insight and attention to detail, a project of this quality and magnitude would not be possible. Finally, we acknowledge the cooperation of Minitab, Inc., SPSS, Texas Instruments, and Microsoft.

Charles Henry Brase

Corrinne Pellillo Brase



ADDITIONAL RESOURCES—GET MORE FROM YOUR TEXTBOOK!

Instructor Resources

Annotated Instructors's Edition (AIE) Answers to all exercises, teaching comments, and pedagogical suggestions appear in the margin, or at the end of the text in the case of large graphs.

Solution Builder Contains complete solutions to all exercises in the text, including those in the Chapter Review and Cumulative Review Problems in online format. Solution Builder allows instructors to create customized, secure PDF printouts of solutions matched exactly to the exercises assigned for class. Available to adoptions by signing up at www.cengage.com/solutionbuilder

ExamView® Allows instructors to create, deliver, and customize tests for class in print and online formats and features automatic grading. This electronic test bank features more than 450 questions based on the text. All test items are also provided in PDF and Microsoft® Word formats for instructors who opt not to use the software component

PowerLecture Provides the instructor with dynamic media tools for teaching. Create, deliver, and customize tests (both print and online) in minutes with ExamView® Computerized Testing Featuring Algorithmic Equations. Easily build solution sets for homework or exams using Solution Builder's online solutions manual. Microsoft® PowerPoint® lecture slides and figures from the book are also included on this CD-ROM.

Student Resources

Student Solutions Manual Provides solutions to the odd-numbered section and chapter exercises and to all the Cumulative Review exercises in the student textbook.



Instructional DVDs Hosted by Dana Moseley, these text-specific DVDs cover all sections of the text and provide explanations of key concepts, examples, exercises, and applications in a lecture-based format. DVDs are close-captioned for the hearing-impaired.

Aplia Is an online interactive learning solution that helps students improve comprehension—and their grade—by integrating a variety of

mediums and tools such as video, tutorials, practice tests, and an interactive eBook. Created by a professor to enhance his own courses, Aplia provides automatically graded assignments with detailed, immediate feedback on every question, and innovative teaching materials. More than 1,000,000 students have used Aplia at over 1,800 institutions.

MINITAB® and IBM SPSS Statistics CD-ROMs These statistical software packages manipulate and interpret data to produce textual, graphical, and tabular results. MINITAB and/or SPSS may be packaged with the textbook. Student versions are available.

CourseMate Brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. Watch student comprehension soar as your class works with the printed textbook and the textbook-specific website. Statistics CourseMate goes beyond the book to deliver what you need. Find the following and more at www.cengage.com/statistics/brase.

- **Engagement Tracker** a first-of-its-kind tool that monitors student engagement in the course. Online student quizzes
- **Technology Guides** Separate guides exist with information and examples for each of four technology tools. Guides are available for the TI-84Plus, TI-S3Plus, and TI-*nspire* graphing calculators, Minitab software (version 14) Microsoft Excel (2008/2007), and SPSS Statistics software.
- **Interactive Teaching and Learning Tools** include glossary flashcards, online datasets (in Microsoft Excel, Minitab, SPSS, and TI-84Plus/TI-83Plus/TI-*nspire* with 84Plus keypad ASCII file formats), statistical tables and formulae, and more.
- **Multimedia eBook** Integrates numerous assets such as video explanations and tutorials to expand upon and reinforce concepts as they appear in the text.

Enhanced WebAssign Offers an extensive online program for Statistics to encourage the practice that's so critical for concept mastery. The meticulously crafted pedagogy and exercises in Brase and Brase's text become even more effective in Enhanced WebAssign.

CengageBrain.com Provides the freedom to purchase online homework and other materials a-la-carte exactly what you need, when you need it.

For more information, visit <http://www.cengage.com/statistics/brase> or contact your local Cengage Learning sales representative.



TABLE OF PREREQUISITE MATERIAL

Chapter	Prerequisite Sections
1 Getting Started	None
2 Organizing Data	1.1, 1.2
3 Averages and Variation	1.1, 1.2, 2.1
4 Elementary Probability Theory	1.1, 1.2, 2.1, 3.1, 3.2
5 The Binomial Probability Distribution and Related Topics	1.1, 1.2, 2.1, 3.1, 3.2, 4.1, 4.2 4.3 useful but not essential
6 Normal Curves and Sampling Distributions (omit 6.6) (include 6.6)	1.1, 1.2, 2.1, 3.1, 3.2, 4.1, 4.2, 5.1 also 5.2, 5.3
7 Estimation (omit 7.3 and parts of 7.4) (include 7.3 and all of 7.4)	1.1, 1.2, 2.1, 3.1, 3.2, 4.1, 4.2, 5.1, 6.1, 6.2, 6.3, 6.4, 6.5 also 5.2, 5.3, 6.6
8 Hypothesis Testing (omit 8.3 and part of 8.5) (include 8.3 and all of 8.5)	1.1, 1.2, 2.1, 3.1, 3.2, 4.1, 4.2, 5.1, 6.1, 6.2, 6.3, 6.4, 6.5 also 5.2, 5.3, 6.6
9 Correlation and Regression (9.1 and 9.2) (9.3 and 9.4)	1.1, 1.2, 3.1, 3.2 also 4.1, 4.2, 5.1, 6.1, 6.2, 6.3, 6.4, 6.5, 7.1, 7.2, 8.1, 8.2
10 Chi-Square and F Distributions (omit 10.3) (include 10.3)	1.1, 1.2, 2.1, 3.1, 3.2, 4.1, 4.2, 5.1, 6.1, 6.2, 6.3, 6.4, 6.5, 8.1 also 7.1
11 Nonparametric Statistics	1.1, 1.2, 2.1, 3.1, 3.2, 4.1, 4.2, 5.1, 6.1, 6.2, 6.3, 6.4, 6.5, 8.1, 8.3



1

- 1.1 What Is Statistics?
- 1.2 Random Samples
- 1.3 Introduction to Experimental Design



Bettmann/Corbis

Chance favors the prepared mind.

—LOUIS PASTEUR

Statistical techniques are tools of

thought . . . not substitutes for thought.

—ABRAHAM KAPLAN



MLB Photos/Getty Images Sport/Getty Images

Louis Pasteur (1822–1895) is the founder of modern bacteriology. At age 57, Pasteur was studying cholera. He accidentally left some bacillus culture unattended in his laboratory during the summer. In the fall, he injected laboratory animals with this bacilli. To his surprise, the animals did not die—in fact, they thrived and were resistant to cholera.

When the final results were examined, it is said that Pasteur remained silent for a minute and then exclaimed, as if he had seen a vision, “Don’t you see they have been vaccinated!” Pasteur’s work ultimately saved many human lives.

Most of the important decisions in life involve incomplete information. Such decisions often involve so many complicated factors that a complete analysis is not practical or even possible. We are often forced into the position of making a guess based on limited information.

As the first quote reminds us, our chances of success are greatly improved if we have a “prepared mind.” The statistical methods you will learn in this book will help you achieve a prepared mind for the study of many different fields. The second quote reminds us that statistics is an important tool, but it is not a replacement for an in-depth knowledge of the field to which it is being applied.

The authors of this book want you to understand and enjoy statistics. The reading material will *tell you* about the subject. The examples will *show you* how it works. To understand, however, you must *get involved*. Guided exercises, calculator and computer applications, section and chapter problems, and writing exercises are all designed to get you involved in the subject. As you grow in your understanding of statistics, we believe you will enjoy learning a subject that has a world full of interesting applications.

For online student resources, visit the Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>

GETTING STARTED

PREVIEW QUESTIONS

Why is statistics important? (SECTION 1.1)

What is the nature of data? (SECTION 1.1)

How can you draw a random sample? (SECTION 1.2)

What are other sampling techniques? (SECTION 1.2)

How can you design ways to collect data? (SECTION 1.3)

FOCUS PROBLEM

Where Have All the Fireflies Gone?

A feature article in *The Wall Street Journal* discusses the disappearance of fireflies. In the article, Professor Sara Lewis of Tufts University and other scholars express concern about the decline in the worldwide population of fireflies.

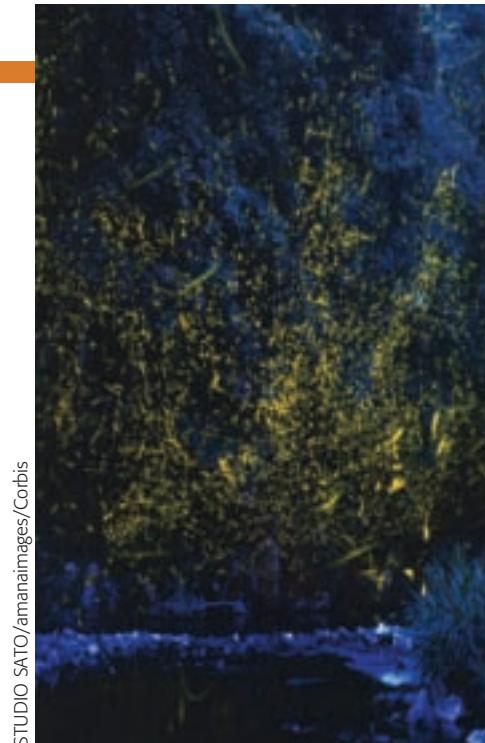
There are a number of possible explanations for the decline, including habitat reduction of woodlands, wetlands, and open fields; pesticides; and pollution. Artificial nighttime lighting might interfere with the Morse-code-like mating ritual of the fireflies. Some chemical companies pay a bounty for fireflies because the insects contain two rare chemicals used in medical research and electronic detection systems used in spacecraft.

What does any of this have to do with statistics?

The truth, at this time, is that no one really knows (a) how much the world firefly population has declined or (b) how to explain the decline. The population of all fireflies is simply too large to study in its entirety.

In any study of fireflies, we must rely on incomplete information from samples. Furthermore, from these samples we must draw realistic conclusions that have statistical integrity. This is the kind of work that makes use of statistical methods to determine ways to collect, analyze, and investigate data.

Suppose you are conducting a study to compare firefly populations exposed to normal daylight/darkness conditions with firefly populations exposed to continuous light (24 hours a day). You set up two firefly colonies in a laboratory environment. The two colonies are identical except that one colony is exposed to normal daylight/darkness



STUDIO SATO/amanaimages/Corbis



Courtesy of Corinne and Charles Brase

Adapted from Ohio State University Firefly Files logo

conditions and the other is exposed to continuous light. Each colony is populated with the same number of mature fireflies. After 72 hours, you count the number of living fireflies in each colony.

After completing this chapter, you will be able to answer the following questions.

- (a) Is this an experiment or an observation study? Explain.
- (b) Is there a control group? Is there a treatment group?
- (c) What is the variable in this study?
- (d) What is the level of measurement (nominal, interval, ordinal, or ratio) of the variable?

(See Problem 11 of the Chapter 1 Review Problems.)

SECTION 1.1

What Is Statistics?

FOCUS POINTS

- Identify variables in a statistical study.
- Distinguish between quantitative and qualitative variables.
- Identify populations and samples.
- Distinguish between parameters and statistics.
- Determine the level of measurement.
- Compare descriptive and inferential statistics.

Introduction

Decision making is an important aspect of our lives. We make decisions based on the information we have, our attitudes, and our values. Statistical methods help us examine information. Moreover, statistics can be used for making decisions when we are faced with uncertainties. For instance, if we wish to estimate the proportion of people who will have a severe reaction to a flu shot without giving the shot to everyone who wants it, statistics provides appropriate methods. Statistical methods enable us to look at information from a small collection of people or items and make inferences about a larger collection of people or items.

Procedures for analyzing data, together with rules of inference, are central topics in the study of statistics.

Statistics

Statistics is the study of how to collect, organize, analyze, and interpret numerical information from data.

The statistical procedures you will learn in this book should supplement your built-in system of inference—that is, the results from statistical procedures and good sense should dovetail. Of course, statistical methods themselves have no power to work miracles. These methods can help us make some decisions, but not all conceivable decisions. Remember, even a properly applied statistical procedure is no more accurate than the data, or facts, on which it is based. Finally, statistical results should be interpreted by one who understands not only the methods, but also the subject matter to which they have been applied.

The general prerequisite for statistical decision making is the gathering of data. First, we need to identify the individuals or objects to be included in the study and the characteristics or features of the individuals that are of interest.

Individuals
Variable

Individuals are the people or objects included in the study.
A variable is a characteristic of the individual to be measured or observed.

For instance, if we want to do a study about the people who have climbed Mt. Everest, then the individuals in the study are all people who have actually made it to the summit. One variable might be the height of such individuals. Other variables might be age, weight, gender, nationality, income, and so on. Regardless of the variables we use, we would not include measurements or observations from people who have not climbed the mountain.

The variables in a study may be *quantitative* or *qualitative* in nature.

Quantitative variable
Qualitative variable

A quantitative variable has a value or numerical measurement for which operations such as addition or averaging make sense. A **qualitative variable** describes an individual by placing the individual into a category or group, such as male or female.

For the Mt. Everest climbers, variables such as height, weight, age, or income are *quantitative* variables. *Qualitative variables* involve nonnumerical observations such as gender or nationality. Sometimes qualitative variables are referred to as *categorical variables*.

Another important issue regarding data is their source. Do the data comprise information from *all* individuals of interest, or from just *some* of the individuals?

Population data
Sample data

In **population data**, the data are from *every* individual of interest.
In **sample data**, the data are from *only some* of the individuals of interest.

It is important to know whether the data are population data or sample data. Data from a specific population are fixed and complete. Data from a sample may vary from sample to sample and are *not* complete.

Population parameter
Sample statistic

A population parameter is a numerical measure that describes an aspect of a population.
A sample statistic is a numerical measure that describes an aspect of a sample.

For instance, if we have data from *all* the individuals who have climbed Mt. Everest, then we have population data. The proportion of males in the *population* of all climbers who have conquered Mt. Everest is an example of a *parameter*.

On the other hand, if our data come from just some of the climbers, we have sample data. The proportion of male climbers in the *sample* is an example of a *statistic*. Note that different samples may have different values for the proportion of male climbers. One of the important features of sample statistics is that they can vary from sample to sample, whereas population parameters are fixed for a given population.

LOOKING FORWARD

In later chapters we will use information based on a sample and sample statistics to estimate population parameters (Chapter 7) or make decisions about the value of population parameters (Chapter 8).

EXAMPLE 1**USING BASIC TERMINOLOGY**

Joe Sohm/Riser/Getty Images

The Hawaii Department of Tropical Agriculture is conducting a study of ready-to-harvest pineapples in an experimental field.

- (a) The pineapples are the *objects* (individuals) of the study. If the researchers are interested in the individual weights of pineapples in the field, then the *variable* consists of weights. At this point, it is important to specify units of measurement and degrees of accuracy of measurement. The weights could be measured to the nearest ounce or gram. Weight is a *quantitative* variable because it is a numerical measure. If weights of *all* the ready-to-harvest pineapples in the field are included in the data, then we have a *population*. The average weight of all ready-to-harvest pineapples in the field is a *parameter*.
- (b) Suppose the researchers also want data on taste. A panel of tasters rates the pineapples according to the categories “poor,” “acceptable,” and “good.” Only some of the pineapples are included in the taste test. In this case, the *variable* is taste. This is a *qualitative* or *categorical* variable. Because only some of the pineapples in the field are included in the study, we have a *sample*. The proportion of pineapples in the sample with a taste rating of “good” is a *statistic*.

Throughout this text, you will encounter *guided exercises* embedded in the reading material. These exercises are included to give you an opportunity to work immediately with new ideas. The questions guide you through appropriate analysis. Cover the answers on the right side (an index card will fit this purpose). After you have thought about or written down *your own response*, check the answers. If there are several parts to an exercise, check each part before you continue. You should be able to answer most of these exercise questions, but don’t skip them—they are important.

GUIDED EXERCISE 1**Using basic terminology**

Television station QUE wants to know the proportion of TV owners in Virginia who watch the station’s new program at least once a week. The station asks a group of 1000 TV owners in Virginia if they watch the program at least once a week.

- (a) Identify the individuals of the study and the variable. ➡ The individuals are the 1000 TV owners surveyed. The variable is the response does, or does not, watch the new program at least once a week.
- (b) Do the data comprise a sample? If so, what is the underlying population? ➡ The data comprise a sample of the population of responses from all TV owners in Virginia.
- (c) Is the variable qualitative or quantitative? ➡ Qualitative—the categories are the two possible responses, does or does not watch the program.
- (d) Identify a quantitative variable that might be of interest. ➡ Age or income might be of interest.
- (e) Is the proportion of viewers in the sample who watch the new program at least once a week a statistic or a parameter? ➡ Statistic—the proportion is computed from sample data.

Levels of Measurement: Nominal, Ordinal, Interval, Ratio

We have categorized data as either qualitative or quantitative. Another way to classify data is according to one of the four *levels of measurement*. These levels indicate the type of arithmetic that is appropriate for the data, such as ordering, taking differences, or taking ratios.

Levels of Measurement

Nominal level

Ordinal level

Interval level

Ratio level

Levels of Measurement

The **nominal level of measurement** applies to data that consist of names, labels, or categories. There are no implied criteria by which the data can be ordered from smallest to largest.

The **ordinal level of measurement** applies to data that can be arranged in order. However, differences between data values either cannot be determined or are meaningless.

The **interval level of measurement** applies to data that can be arranged in order. In addition, differences between data values are meaningful.

The **ratio level of measurement** applies to data that can be arranged in order. In addition, both differences between data values and ratios of data values are meaningful. Data at the ratio level have a true zero.

EXAMPLE 2

Michelle Dulieu, 2009/Used under license from Shutterstock.com



LEVELS OF MEASUREMENT

Identify the type of data.

- (a) Taos, Acoma, Zuni, and Cochiti are the names of four Native American pueblos from the population of names of all Native American pueblos in Arizona and New Mexico.

SOLUTION: These data are at the *nominal* level. Notice that these data values are simply names. By looking at the name alone, we cannot determine if one name is “greater than or less than” another. Any ordering of the names would be numerically meaningless.

- (b) In a high school graduating class of 319 students, Jim ranked 25th, June ranked 19th, Walter ranked 10th, and Julia ranked 4th, where 1 is the highest rank.

SOLUTION: These data are at the *ordinal* level. Ordering the data clearly makes sense. Walter ranked higher than June. Jim had the lowest rank, and Julia the highest. However, numerical differences in ranks do not have meaning. The difference between June’s and Jim’s ranks is 6, and this is the same difference that exists between Walter’s and Julia’s ranks. However, this difference doesn’t really mean anything significant. For instance, if you looked at grade point average, Walter and Julia may have had a large gap between their grade point averages, whereas June and Jim may have had closer grade point averages. In any ranking system, it is only the relative standing that matters. Differences between ranks are meaningless.

- (c) Body temperatures (in degrees Celsius) of trout in the Yellowstone River.

SOLUTION: These data are at the *interval* level. We can certainly order the data, and we can compute meaningful differences. However, for Celsius-scale temperatures, there is not an inherent starting point. The value 0°C may seem to be a starting point, but this value does not indicate the state of “no heat.” Furthermore, it is not correct to say that 20°C is twice as hot as 10°C .

Korban Schwab/Stockphoto.com



- (d) Length of trout swimming in the Yellowstone River.

SOLUTION: These data are at the *ratio* level. An 18-inch trout is three times as long as a 6-inch trout. Observe that we can divide 6 into 18 to determine a meaningful *ratio* of trout lengths.



In summary, there are four levels of measurement. The nominal level is considered the lowest, and in ascending order we have the ordinal, interval, and ratio levels. In general, calculations based on a particular level of measurement may not be appropriate for a lower level.

PROCEDURE

HOW TO DETERMINE THE LEVEL OF MEASUREMENT

The levels of measurement, listed from lowest to highest, are nominal, ordinal, interval, and ratio. To determine the level of measurement of data, state the *highest level* that can be justified for the entire collection of data. Consider which calculations are suitable for the data.

Level of Measurement	Suitable Calculation
Nominal	We can put the data into categories.
Ordinal	We can order the data from smallest to largest or "worst" to "best." Each data value can be <i>compared</i> with another data value.
Interval	We can order the data and also take the differences between data values. At this level, it makes sense to compare the differences between data values. For instance, we can say that one data value is 5 more than or 12 less than another data value.
Ratio	We can order the data, take differences, and also find the ratio between data values. For instance, it makes sense to say that one data value is twice as large as another.

GUIDED EXERCISE 2

Levels of measurement

The following describe different data associated with a state senator. For each data entry, indicate the corresponding *level of measurement*.

- (a) The senator's name is Sam Wilson.



Nominal level

- (b) The senator is 58 years old.

Ratio level. Notice that age has a meaningful zero. It makes sense to give age ratios. For instance, Sam is twice as old as someone who is 29.

- (c) The years in which the senator was elected to the Senate are 1998, 2004, and 2010.



Interval level. Dates can be ordered, and the difference between dates has meaning. For instance, 2004 is six years later than 1998. However, ratios do not make sense. The year 2000 is not twice as large as the year 1000. In addition, the year 0 does not mean "no time."

Continued

GUIDED EXERCISE 2 *continued*

- (d) The senator's total taxable income last year was \$878,314.  Ratio level. It makes sense to say that the senator's income is 10 times that of someone earning \$87,831.40.
- (e) The senator surveyed his constituents regarding his proposed water protection bill. The choices for response were strong support, support, neutral, against, or strongly against.  Ordinal level. The choices can be ordered, but there is no meaningful numerical difference between two choices.
- (f) The senator's marital status is "married."  Nominal level
- (g) A leading news magazine claims the senator is ranked seventh for his voting record on bills regarding public education.  Ordinal level. Ranks can be ordered, but differences between ranks may vary in meaning.



CRITICAL THINKING

"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay." Sherlock Holmes said these words in *The Adventure of the Copper Beeches* by Sir Arthur Conan Doyle.

Reliable statistical conclusions require reliable data. This section has provided some of the vocabulary used in discussing data. As you read a statistical study or conduct one, pay attention to the nature of the data and the ways they were collected.

When you select a variable to measure, be sure to specify the process and requirements for measurement. For example, if the variable is the weight of ready-to-harvest pineapples, specify the unit of weight, the accuracy of measurement, and maybe even the particular scale to be used. If some weights are in ounces and others in grams, the data are fairly useless.

Another concern is whether or not your measurement instrument truly measures the variable. Just asking people if they know the geographic location of the island nation of Fiji may not provide accurate results. The answers may reflect the fact that the respondents want you to think they are knowledgeable. Asking people to locate Fiji on a map may give more reliable results.

The level of measurement is also an issue. You can put numbers into a calculator or computer and do all kinds of arithmetic. However, you need to judge whether the operations are meaningful. For ordinal data such as restaurant rankings, you can't conclude that a 4-star restaurant is "twice as good" as a 2-star restaurant, even though the number 4 is twice 2.

Are the data from a sample, or do they comprise the entire population? Sample data can vary from one sample to another! This means that if you are studying the same statistic from two different samples of the same size, the data values may be different. In fact, the ways in which sample statistics vary among different samples of the same size will be the focus of our study from Section 6.4 on.

Interpretation When you work with sample data, carefully consider the population from which they are drawn. Observations and analysis of the sample are applicable to only the population from which the sample is drawn.



LOOKING FORWARD

The purpose of collecting and analyzing data is to obtain information. Statistical methods provide us tools to obtain information from data. These methods break into two branches.

Descriptive statistics

Inferential statistics

Descriptive statistics involves methods of organizing, picturing, and summarizing information from samples or populations.

Inferential statistics involves methods of using information from a sample to draw conclusions regarding the population.

We will look at methods of descriptive statistics in Chapters 2, 3, and 9. These methods may be applied to data from samples or populations.

Sometimes we do not have access to an entire population. At other times, the difficulties or expense of working with the entire population is prohibitive. In such cases, we will use inferential statistics together with probability. These are the topics of Chapters 4 through 11.

VIEWPOINT

The First Measured Century

The 20th century saw measurements of aspects of American life that had never been systematically studied before. Social conditions involving crime, sex, food, fun, religion, and work were numerically investigated. The measurements and survey responses taken over the entire century reveal unsuspected statistical trends. The First Measured Century is a book by Caplow, Hicks, and Wattenberg. It is also a PBS documentary available on video. For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the PBS First Measured Century documentary.

SECTION 1.1 PROBLEMS

1. | **Statistical Literacy** What is the difference between an individual and a variable?
2. | **Statistical Literacy** Are data at the nominal level of measurement quantitative or qualitative?
3. | **Statistical Literacy** What is the difference between a parameter and a statistic?
4. | **Statistical Literacy** For a set population, does a parameter ever change? If there are three different samples of the same size from a set population, is it possible to get three different values for the same statistic?

5. | **Critical Thinking** Numbers are often assigned to data that are categorical in nature.
 - (a) Consider these number assignments for category items describing electronic ways of expressing personal opinions:

1 = Twitter; 2 = e-mail; 3 = text message; 4 = Facebook; 5 = blog

Are these numerical assignments at the ordinal data level or higher? Explain.

- (b) Consider these number assignments for category items describing usefulness of customer service:

1 = not helpful; 2 = somewhat helpful; 3 = very helpful;
4 = extremely helpful

Are these numerical assignments at the ordinal data level? Explain. What about at the interval level or higher? Explain.

6. **Interpretation** Lucy conducted a survey asking some of her friends to specify their favorite type of TV entertainment from the following list of choices:
- sitcom; reality; documentary; drama; cartoon; other
- Do Lucy's observations apply to *all* adults? Explain. From the description of the survey group, can we draw any conclusions regarding age of participants, gender of participants, or education level of participants?
7. **Marketing: Fast Food** A national survey asked 1261 U.S. adult fast-food customers which meal (breakfast, lunch, dinner, snack) they ordered.
- Identify the variable.
 - Is the variable quantitative or qualitative?
 - What is the implied population?
8. **Advertising: Auto Mileage** What is the average miles per gallon (mpg) for all new cars? Using *Consumer Reports*, a random sample of 35 new cars gave an average of 21.1 mpg.
- Identify the variable.
 - Is the variable quantitative or qualitative?
 - What is the implied population?
9. **Ecology: Wetlands** Government agencies carefully monitor water quality and its effect on wetlands (Reference: *Environmental Protection Agency Wetland Report EPA 832-R-93-005*). Of particular concern is the concentration of nitrogen in water draining from fertilized lands. Too much nitrogen can kill fish and wildlife. Twenty-eight samples of water were taken at random from a lake. The nitrogen concentration (milligrams of nitrogen per liter of water) was determined for each sample.
- Identify the variable.
 - Is the variable quantitative or qualitative?
 - What is the implied population?
10. **Archaeology: Ireland** The archaeological site of Tara is more than 4000 years old. Tradition states that Tara was the seat of the high kings of Ireland. Because of its archaeological importance, Tara has received extensive study (Reference: *Tara: An Archaeological Survey* by Conor Newman, Royal Irish Academy, Dublin). Suppose an archaeologist wants to estimate the density of ferromagnetic artifacts in the Tara region. For this purpose, a random sample of 55 plots, each of size 100 square meters, is used. The number of ferromagnetic artifacts for each plot is determined.
- Identify the variable.
 - Is the variable quantitative or qualitative?
 - What is the implied population?
11. **Student Life: Levels of Measurement** Categorize these measurements associated with student life according to level: nominal, ordinal, interval, or ratio.
- Length of time to complete an exam
 - Time of first class
 - Major field of study
 - Course evaluation scale: poor, acceptable, good
 - Score on last exam (based on 100 possible points)
 - Age of student
12. **Business: Levels of Measurement** Categorize these measurements associated with a robotics company according to level: nominal, ordinal, interval, or ratio.
- Salesperson's performance: below average, average, above average
 - Price of company's stock
 - Names of new products
 - Temperature ($^{\circ}\text{F}$) in CEO's private office
 - Gross income for each of the past 5 years
 - Color of product packaging

13. **Fishing: Levels of Measurement** Categorize these measurements associated with fishing according to level: nominal, ordinal, interval, or ratio.

- Species of fish caught: perch, bass, pike, trout
- Cost of rod and reel
- Time of return home
- Guidebook rating of fishing area: poor, fair, good
- Number of fish caught
- Temperature of water

14. **Education: Teacher Evaluation** If you were going to apply *statistical methods* to analyze teacher evaluations, which question form, A or B, would be better?

Form A: In your own words, tell how this teacher compares with other teachers you have had.

Form B: Use the following scale to rank your teacher as compared with other teachers you have had.

1 worst	2 below average	3 average	4 above average	5 best
------------	-----------------------	--------------	-----------------------	-----------

15. **Critical Thinking** You are interested in the weights of backpacks students carry to class and decide to conduct a study using the backpacks carried by 30 students.

- Give some instructions for weighing the backpacks. Include unit of measure, accuracy of measure, and type of scale.
- Do you think each student asked will allow you to weigh his or her backpack?
- Do you think telling students ahead of time that you are going to weigh their backpacks will make a difference in the weights?

SECTION 1.2

Random Samples

FOCUS POINTS

- Explain the importance of random samples.
- Construct a simple random sample using random numbers.
- Simulate a random process.
- Describe stratified sampling, cluster sampling, systematic sampling, multistage sampling, and convenience sampling.

Simple Random Samples

Eat lamb—20,000 coyotes can't be wrong!

This slogan is sometimes found on bumper stickers in the western United States. The slogan indicates the trouble that ranchers have experienced in protecting their flocks from predators. Based on their experience with this sample of the coyote population, the ranchers concluded that *all* coyotes are dangerous to their flocks and should be eliminated! The ranchers used a special poison bait to get rid of the coyotes. Not only was this poison distributed on ranch land, but with government cooperation, it also was distributed widely on public lands.

The ranchers found that the results of the widespread poisoning were not very beneficial. The sheep-eating coyotes continued to thrive while the general population of coyotes and other predators declined. What was the problem? The sheep-eating coyotes that the ranchers had observed were not a representative sample of all coyotes. Modern methods of predator control, however, target the sheep-eating

coyotes. To a certain extent, the new methods have come about through a closer examination of the sampling techniques used.

In this section, we will examine several widely used sampling techniques. One of the most important sampling techniques is a *simple random sample*.

Simple random sample

A **simple random sample** of n measurements from a population is a subset of the population selected in such a manner that every sample of size n from the population has an equal chance of being selected.

In a simple random sample, not only does every sample of the specified size have an equal chance of being selected, but every individual of the population also has an equal chance of being selected. However, the fact that each individual has an equal chance of being selected does not necessarily imply a simple random sample. Remember, for a simple random sample, every sample of the given size must also have an equal chance of being selected.

GUIDED EXERCISE 3

Simple random sample

Is open space around metropolitan areas important? Players of the Colorado Lottery might think so, since some of the proceeds of the game go to fund open space and outdoor recreational space. To play the game, you pay \$1 and choose any six different numbers from the group of numbers 1 through 42. If your group of six numbers matches the winning group of six numbers selected by simple random sampling, then you are a winner of a grand prize of at least \$1.5 million.

- (a) Is the number 25 as likely to be selected in the winning group of six numbers as the number 5?  Yes. Because the winning numbers constitute a simple random sample, each number from 1 through 42 has an equal chance of being selected.
- (b) Could all the winning numbers be even?  Yes, since six even numbers is one of the possible groups of six numbers.
- (c) Your friend always plays the numbers
1 2 3 4 5 6
Could she ever win?  Yes. In a simple random sample, the listed group of six numbers is *as likely as any* of the 5,245,786 possible groups of six numbers to be selected as the winner. (See Section 4.3 to learn how to compute the number of possible groups of six numbers that can be selected from 42 numbers.)

Random-number table

How do we get random samples? Suppose you need to know if the emission systems of the latest shipment of Toyotas satisfy pollution-control standards. You want to pick a random sample of 30 cars from this shipment of 500 cars and test them. One way to pick a random sample is to number the cars 1 through 500. Write these numbers on cards, mix up the cards, and then draw 30 numbers. The sample will consist of the cars with the chosen numbers. If you mix the cards sufficiently, this procedure produces a random sample.

An easier way to select the numbers is to use a *random-number table*. You can make one yourself by writing the digits 0 through 9 on separate cards and mixing up these cards in a hat. Then draw a card, record the digit, return the card, and mix up the cards again. Draw another card, record the digit, and so on. Table 1 in the Appendix is a ready-made random-number table (adapted from Rand

Corporation, *A Million Random Digits with 100,000 Normal Deviates*). Let's see how to pick our random sample of 30 Toyotas by using this random-number table.



EXAMPLE 3

RANDOM-NUMBER TABLE

Use a random-number table to pick a random sample of 30 cars from a population of 500 cars.

SOLUTION: Again, we assign each car a different number between 1 and 500, inclusive. Then we use the random-number table to choose the sample. Table 1 in the Appendix has 50 rows and 10 blocks of five digits each; it can be thought of as a solid mass of digits that has been broken up into rows and blocks for user convenience.

You read the digits by beginning anywhere in the table. We dropped a pin on the table, and the head of the pin landed in row 15, block 5. We'll begin there and list all the digits in that row. If we need more digits, we'll move on to row 16, and so on. The digits we begin with are

99281 59640 15221 96079 09961 05371

Since the highest number assigned to a car is 500, and this number has three digits, we regroup our digits into blocks of 3:

992 815 964 015 221 960 790 996 105 371

To construct our random sample, we use the first 30 car numbers we encounter in the random-number table when we start at row 15, block 5. We skip the first three groups—992, 815, and 964—because these numbers are all too large. The next group of three digits is 015, which corresponds to 15. Car number 15 is the first car included in our sample, and the next is car number 221. We skip the next three groups and then include car numbers 105 and 371. To get the rest of the cars in the sample, we continue to the next line and use the random-number table in the same fashion. If we encounter a number we've used before, we skip it.

COMMENT When we use the term *(simple) random sample*, we have very specific criteria in mind for selecting the sample. One proper method for selecting a simple random sample is to use a computer- or calculator-based random-number generator or a table of random numbers as we have done in the example. The term *random* should not be confused with *haphazard*!

LOOKING FORWARD

The runs test for randomness discussed in Section 11.4 shows how to determine if two symbols are randomly mixed in an ordered list of symbols.

PROCEDURE

HOW TO DRAW A RANDOM SAMPLE

1. Number all members of the population sequentially.
2. Use a table, calculator, or computer to select random numbers from the numbers assigned to the population members.
3. Create the sample by using population members with numbers corresponding to those randomly selected.

LOOKING FORWARD

Simple random samples are key components in methods of inferential statistics that we will study in Chapters 7–11. In fact, in order to draw conclusions about a population, the methods we will study *require* that we have simple random samples from the populations of interest.

Another important use of random-number tables is in *simulation*. We use the word *simulation* to refer to the process of providing numerical imitations of “real” phenomena. Simulation methods have been productive in studying a diverse array of subjects such as nuclear reactors, cloud formation, cardiology (and medical science in general), highway design, production control, shipbuilding, airplane design, war games, economics, and electronics. A complete list would probably include something from every aspect of modern life. In Guided Exercise 4 we’ll perform a brief simulation.

Simulation

A **simulation** is a numerical facsimile or representation of a real-world phenomenon.

GUIDED EXERCISE 4**Simulation**

Use a random-number table to simulate the outcomes of tossing a balanced (that is, fair) penny 10 times.

- (a) How many outcomes are possible when you toss a coin once?  Two—heads or tails
- (b) There are several ways to assign numbers to the two outcomes. Because we assume a fair coin, we can assign an even digit to the outcome “heads” and an odd digit to the outcome “tails.” Then, starting at block 3 of row 2 of Table 1 in the Appendix, list the first 10 single digits.  7 1 5 4 9 4 4 8 4 3
- (c) What are the outcomes associated with the 10 digits?  T T T H T H H H H T
- (d) If you start in a different block and row of Table 1 in the Appendix, will you get the same sequence of outcomes?  It is possible, but not very likely. (In Section 4.3 you will learn how to determine that there are 1024 possible sequences of outcomes for 10 tosses of a coin.)

TECH NOTES**Sampling with replacement**

Most statistical software packages, spreadsheet programs, and statistical calculators generate random numbers. In general, these devices sample with replacement. **Sampling with replacement** means that although a number is selected for the sample, it is *not removed* from the population. Therefore, the same number may be selected for the sample more than once. If you need to sample without replacement, generate more items than you need for the sample. Then sort the sample and remove duplicate values. Specific procedures for generating random samples using the TI-84Plus/TI-83Plus/TI-nspire (with TI-84 Plus keypad) calculator, Excel 2007, Minitab, and SPSS are shown in Using Technology at the end of this chapter. More details are given in the separate *Technology Guides* for each of these technologies.

Other Sampling Techniques

Although we will assume throughout this text that (simple) random samples are used, other methods of sampling are also widely used. Appropriate statistical techniques exist for these sampling methods, but they are beyond the scope of this text.

Stratified sampling

One of these sampling methods is called *stratified sampling*. Groups or classes inside a population that share a common characteristic are called *strata* (plural of *stratum*). For example, in the population of all undergraduate college students, some strata might be freshmen, sophomores, juniors, or seniors. Other strata might be men or women, in-state students or out-of-state students, and so on. In the method of stratified sampling, the population is divided into at least two distinct strata. Then a (simple) random sample of a certain size is drawn from each stratum, and the information obtained is carefully adjusted or weighted in all resulting calculations.

The groups or strata are often sampled in proportion to their actual percentages of occurrence in the overall population. However, other (more sophisticated) ways to determine the optimal sample size in each stratum may give the best results. In general, statistical analysis and tests based on data obtained from stratified samples are somewhat different from techniques discussed in an introductory course in statistics. Such methods for stratified sampling will not be discussed in this text.

Systematic sampling

Another popular method of sampling is called *systematic sampling*. In this method, it is assumed that the elements of the population are arranged in some natural sequential order. Then we select a (random) starting point and select every k th element for our sample. For example, people lining up to buy rock concert tickets are “in order.” To generate a systematic sample of these people (and ask questions regarding topics such as age, smoking habits, income level, etc.), we could include every fifth person in line. The “starting” person is selected at random from the first five.

The advantage of a systematic sample is that it is easy to get. However, there are dangers in using systematic sampling. When the population is repetitive or cyclic in nature, systematic sampling should not be used. For example, consider a fabric mill that produces dress material. Suppose the loom that produces the material makes a mistake every 17th yard, but we check only every 16th yard with an automated electronic scanner. In this case, a random starting point may or may not result in detection of fabric flaws before a large amount of fabric is produced.

Cluster sampling

Cluster sampling is a method used extensively by government agencies and certain private research organizations. In cluster sampling, we begin by dividing the demographic area into sections. Then we randomly select sections or clusters. Every member of the cluster is included in the sample. For example, in conducting a survey of school children in a large city, we could first randomly select five schools and then include all the children from each selected school.

Multistage samples

Often a population is very large or geographically spread out. In such cases, samples are constructed through a *multistage sample design* of several stages, with the final stage consisting of clusters. For instance, the government Current Population Survey interviews about 60,000 households across the United States each month by means of a multistage sample design.

Convenience sampling

For the Current Population Survey, the first stage consists of selecting samples of large geographic areas that do not cross state lines. These areas are further broken down into smaller blocks, which are stratified according to ethnic and other factors. Stratified samples of the blocks are then taken. Finally, housing units in each chosen block are broken into clusters of nearby housing units. A random sample of these clusters of housing units is selected, and each household in the final cluster is interviewed.

Convenience sampling simply uses results or data that are conveniently and readily obtained. In some cases, this may be all that is available, and in many cases, it is better than no information at all. However, convenience sampling does run the risk of being severely biased. For instance, consider a newscaster who wishes to get

the “opinions of the people” about a proposed seat tax to be imposed on tickets to all sporting events. The revenues from the seat tax will then be used to support the local symphony. The newscaster stands in front of a concert hall and surveys the first five people exiting after a symphony performance who will cooperate. This method of choosing a sample will produce some opinions, and perhaps some human interest stories, but it certainly has bias. It is hoped that the city council will not use these opinions as the sole basis for a decision about the proposed tax. It is good advice to be very cautious indeed when the data come from the method of convenience sampling.

Sampling Techniques

Random sampling: Use a simple random sample from the entire population.

Stratified sampling: Divide the entire population into distinct subgroups called strata. The strata are based on a specific characteristic such as age, income, education level, and so on. All members of a stratum share the specific characteristic. Draw random samples from each stratum.

Systematic sampling: Number all members of the population sequentially. Then, from a starting point selected at random, include every k th member of the population in the sample.

Cluster sampling: Divide the entire population into pre-existing segments or clusters. The clusters are often geographic. Make a random selection of clusters. Include every member of each selected cluster in the sample.

Multistage sampling: Use a variety of sampling methods to create successively smaller groups at each stage. The final sample consists of clusters.

Convenience sampling: Create a sample by using data from population members that are readily available.



Sampling frame

We call the list of individuals from which a sample is actually selected the *sampling frame*. Ideally, the sampling frame is the entire population. However, from a practical perspective, not all members of a population may be accessible. For instance, using a telephone directory as the sample frame for residential telephone contacts would not include unlisted numbers.

Undercoverage

When the sample frame does not match the population, we have what is called *undercoverage*. In demographic studies, undercoverage could result if the homeless, fugitives from the law, and so forth, are not included in the study.

A **sampling frame** is a list of individuals from which a sample is actually selected.

Undercoverage results from omitting population members from the sample frame.

Sampling error

In general, even when the sampling frame and the population match, a sample is not a perfect representation of a population. Therefore, information drawn from a sample may not exactly match corresponding information from the population. To the extent that sample information does not match the corresponding population information, we have an error, called a *sampling error*.

Nonsampling error

A **sampling error** is the difference between measurements from a sample and corresponding measurements from the respective population. It is caused by the fact that the sample does not perfectly represent the population.

A **nonsampling error** is the result of poor sample design, sloppy data collection, faulty measuring instruments, bias in questionnaires, and so on.

Sampling errors do not represent mistakes! They are simply the consequences of using samples instead of populations. However, be alert to nonsampling errors, which may sometimes occur inadvertently.

**VIEWPOINT****Extraterrestrial Life?**

Do you believe intelligent life exists on other planets? Using methods of random sampling, a Fox News opinion poll found that about 54% of all U.S. men do believe in intelligent life on other planets, whereas only 47% of U.S. women believe there is such life. How could you conduct a random survey of students on your campus regarding belief in extraterrestrial life?

**SECTION 1.2
PROBLEMS**

1. **Statistical Literacy** Explain the difference between a stratified sample and a cluster sample.
2. **Statistical Literacy** Explain the difference between a simple random sample and a systematic sample.
3. **Statistical Literacy** Marcie conducted a study of the cost of breakfast cereal. She recorded the costs of several boxes of cereal. However, she neglected to take into account the number of servings in each box. Someone told her not to worry because she just had some sampling error. Comment on that advice.
4. **Critical Thinking** Consider the students in your statistics class as the population and suppose they are seated in four rows of 10 students each. To select a sample, you toss a coin. If it comes up heads, you use the 20 students sitting in the first two rows as your sample. If it comes up tails, you use the 20 students sitting in the last two rows as your sample.
 - (a) Does every student have an equal chance of being selected for the sample? Explain.
 - (b) Is it possible to include students sitting in row 3 with students sitting in row 2 in your sample? Is your sample a simple random sample? Explain.
 - (c) Describe a process you could use to get a simple random sample of size 20 from a class of size 40.
5. **Critical Thinking** Suppose you are assigned the number 1, and the other students in your statistics class call out consecutive numbers until each person in the class has his or her own number. Explain how you could get a random sample of four students from your statistics class.
 - (a) Explain why the first four students walking into the classroom would not necessarily form a random sample.
 - (b) Explain why four students coming in late would not necessarily form a random sample.
 - (c) Explain why four students sitting in the back row would not necessarily form a random sample.
 - (d) Explain why the four tallest students would not necessarily form a random sample.

6. **Critical Thinking** In each of the following situations, the sampling frame does not match the population, resulting in undercoverage. Give examples of population members that might have been omitted.
- The population consists of all 250 students in your large statistics class. You plan to obtain a simple random sample of 30 students by using the sampling frame of students present next Monday.
 - The population consists of all 15-year-olds living in the attendance district of a local high school. You plan to obtain a simple random sample of 200 such residents by using the student roster of the high school as the sampling frame.
7. **Sampling: Random** Use a random-number table to generate a list of 10 random numbers between 1 and 99. Explain your work.
8. **Sampling: Random** Use a random-number table to generate a list of eight random numbers from 1 to 976. Explain your work.
9. **Sampling: Random** Use a random-number table to generate a list of six random numbers from 1 to 8615. Explain your work.
10. **Simulation: Coin Toss** Use a random-number table to simulate the outcomes of tossing a quarter 25 times. Assume that the quarter is balanced (i.e., fair).
11. **Computer Simulation: Roll of a Die** A die is a cube with dots on each face. The faces have 1, 2, 3, 4, 5, or 6 dots. The table below is a computer simulation (from the software package Minitab) of the results of rolling a fair die 20 times.

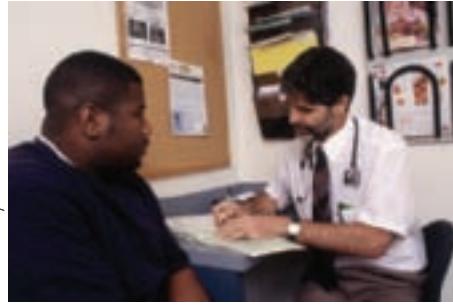
DATA DISPLAY

ROW	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	5	2	2	2	5	3	2	3	1	4
2	3	2	4	5	4	5	3	5	3	4

- (a) Assume that each number in the table corresponds to the number of dots on the upward face of the die. Is it appropriate that the same number appears more than once? Why? What is the outcome of the fourth roll?
- (b) If we simulate more rolls of the die, do you expect to get the same sequence of outcomes? Why or why not?
12. **Simulation: Birthday Problem** Suppose there are 30 people at a party. Do you think any two share the same birthday? Let's use the random-number table to simulate the birthdays of the 30 people at the party. Ignoring leap year, let's assume that the year has 365 days. Number the days, with 1 representing January 1, 2 representing January 2, and so forth, with 365 representing December 31. Draw a random sample of 30 days (with replacement). These days represent the birthdays of the people at the party. Were any two of the birthdays the same? Compare your results with those obtained by other students in the class. Would you expect the results to be the same or different?
13. **Education: Test Construction** Professor Gill is designing a multiple-choice test. There are to be 10 questions. Each question is to have five choices for answers. The choices are to be designated by the letters *a*, *b*, *c*, *d*, and *e*. Professor Gill wishes to use a random-number table to determine which letter choice should correspond to the correct answer for a question. Using the number correspondence 1 for *a*, 2 for *b*, 3 for *c*, 4 for *d*, and 5 for *e*, use a random-number table to determine the letter choice for the correct answer for each of the 10 questions.
14. **Education: Test Construction** Professor Gill uses true-false questions. She wishes to place 20 such questions on the next test. To decide whether to place a true statement or a false statement in each of the 20 questions, she uses a random-number table. She selects 20 digits from the table. An even digit tells her to use a true statement. An odd digit tells her to use a false statement. Use a random-number table to pick a sequence of 20 digits, and describe the corresponding sequence of 20 true-false questions. What would the test key for your sequence look like?

15. **Sampling Methods: Benefits Package** An important part of employee compensation is a benefits package, which might include health insurance, life insurance, child care, vacation days, retirement plan, parental leave, bonuses, etc. Suppose you want to conduct a survey of benefits packages available in private businesses in Hawaii. You want a sample size of 100. Some sampling techniques are described below. Categorize each technique as *simple random sample*, *stratified sample*, *systematic sample*, *cluster sample*, or *convenience sample*.
- (a) Assign each business in the Island Business Directory a number, and then use a random-number table to select the businesses to be included in the sample.
 - (b) Use postal ZIP Codes to divide the state into regions. Pick a random sample of 10 ZIP Code areas and then include all the businesses in each selected ZIP Code area.
 - (c) Send a team of five research assistants to Bishop Street in downtown Honolulu. Let each assistant select a block or building and interview an employee from each business found. Each researcher can have the rest of the day off after getting responses from 20 different businesses.
 - (d) Use the Island Business Directory. Number all the businesses. Select a starting place at random, and then use every 50th business listed until you have 100 businesses.
 - (e) Group the businesses according to type: medical, shipping, retail, manufacturing, financial, construction, restaurant, hotel, tourism, other. Then select a random sample of 10 businesses from each business type.
16. **Sampling Methods: Health Care** Modern Managed Hospitals (MMH) is a national for-profit chain of hospitals. Management wants to survey patients discharged this past year to obtain patient satisfaction profiles. They wish to use a sample of such patients. Several sampling techniques are described below. Categorize each technique as *simple random sample*, *stratified sample*, *systematic sample*, *cluster sample*, or *convenience sample*.
- (a) Obtain a list of patients discharged from all MMH facilities. Divide the patients according to length of hospital stay (2 days or less, 3–7 days, 8–14 days, more than 14 days). Draw simple random samples from each group.
 - (b) Obtain lists of patients discharged from all MMH facilities. Number these patients, and then use a random-number table to obtain the sample.
 - (c) Randomly select some MMH facilities from each of five geographic regions, and then include all the patients on the discharge lists of the selected hospitals.
 - (d) At the beginning of the year, instruct each MMH facility to survey every 500th patient discharged.
 - (e) Instruct each MMH facility to survey 10 discharged patients this week and send in the results.

Frank Siteman/Index Stock Imagery/
Photolibrary



SECTION 1.3

Introduction to Experimental Design

Focus points

- Discuss what it means to take a census.
- Describe simulations, observational studies, and experiments.
- Identify control groups, placebo effects, completely randomized experiments, and randomized block experiments.
- Discuss potential pitfalls that might make your data unreliable.

Planning a Statistical Study

Planning a statistical study and gathering data are essential components of obtaining reliable information. Depending on the nature of the statistical study, a great deal of expertise and resources may be required during the planning stage. In this section, we look at some of the basics of planning a statistical study.

PROCEDURE**BASIC GUIDELINES FOR PLANNING A STATISTICAL STUDY**

1. First, identify the individuals or objects of interest.
2. Specify the variables as well as the protocols for taking measurements or making observations.
3. Determine if you will use an entire population or a representative sample. If using a sample, decide on a viable sampling method.
4. In your data collection plan, address issues of ethics, subject confidentiality, and privacy. If you are collecting data at a business, store, college, or other institution, be sure to be courteous and to obtain permission as necessary.
5. Collect the data.
6. Use appropriate descriptive statistics methods (Chapters 2, 3, and 9) and make decisions using appropriate inferential statistics methods (Chapters 7–11).
7. Finally, note any concerns you might have about your data collection methods and list any recommendations for future studies.

Census

One issue to consider is whether to use the entire population in a study or a representative sample. If we use data from the entire population, we have a *census*.

In a **census**, measurements or observations from the *entire* population are used.

When the population is small and easily accessible, a census is very useful because it gives complete information about the population. However, obtaining a census can be both expensive and difficult. Every 10 years, the U.S. Department of Commerce Census Bureau is required to conduct a census of the United States. However, contacting some members of the population—such as the homeless—is almost impossible. Sometimes members of the population will not respond. In such cases, statistical estimates for the missing responses are often supplied.

Overcounting, that is, counting the same person more than once, is also a problem the Census Bureau is addressing. In fact, in 2000, slightly more people were counted twice than the estimated number of people missed. For instance,

Sample

a college student living on campus might be counted on a parent's census form as well as on his or her own census form.

If we use data from only part of the population of interest, we have a *sample*.

In a **sample**, measurements or observations from *part* of the population are used.

Simulation

In the previous section, we examined several sampling strategies: simple random, stratified, cluster, systematic, multistage, and convenience. In this text, we will study methods of inferential statistics based on simple random samples.

As discussed in Section 1.2, *simulation* is a numerical facsimile of real-world phenomena. Sometimes simulation is called a "dry lab" approach, in the sense that it is a mathematical imitation of a real situation. Advantages of simulation are that numerical and statistical simulations can fit real-world problems extremely well. The researcher can also explore procedures through simulation that might be very dangerous in real life.

Experiments and Observation

When gathering data for a statistical study, we want to distinguish between observational studies and experiments.

Observational study

In an **observational study**, observations and measurements of individuals are conducted in a way that doesn't change the response or the variable being measured.

Experiment

In an **experiment**, a *treatment* is deliberately imposed on the individuals in order to observe a possible change in the response or variable being measured.



EXAMPLE 4

EXPERIMENT

In 1778, Captain James Cook landed in what we now call the Hawaiian Islands. He gave the islanders a present of several goats, and over the years these animals multiplied into wild herds totaling several thousand. They eat almost anything, including the famous silver sword plant, which was once unique to Hawaii. At one time, the silver sword grew abundantly on the island of Maui (in Haleakala, a national park on that island, the silver sword can still be found), but each year there seemed to be fewer and fewer plants. Biologists suspected that the goats were partially responsible for the decline in the number of plants and conducted a statistical study that verified their theory.

- To test the theory, park biologists set up stations in remote areas of Haleakala. At each station two plots of land similar in soil conditions, climate, and plant count were selected. One plot was fenced to keep out the goats, while the other was not. At regular intervals a plant count was made in each plot. This study involved an *experiment* because a *treatment* (the fence) was imposed on one plot.
- The experiment involved two plots at each station. The plot that was not fenced represented the *control* plot. This was the plot on which a treatment was specifically not imposed, although the plot was similar to the fenced plot in every other way.



Silver sword plant, Haleakala National Park

Placebo effect

Statistical experiments are commonly used to determine the effect of a treatment. However, the design of the experiment needs to *control* for other possible causes of the effect. For instance, in medical experiments, the *placebo effect* is the improvement or change that is the result of patients just believing in the treatment, whether or not the treatment itself is effective.

The **placebo effect** occurs when a subject receives no treatment but (incorrectly) believes he or she is in fact receiving treatment and responds favorably.

Treatment group**Completely randomized experiment**

To account for the placebo effect, patients are divided into two groups. One group receives the prescribed treatment. The other group, called the *control group*, receives a dummy or placebo treatment that is disguised to look like the real treatment. Finally, after the treatment cycle, the medical condition of the patients in the *treatment group* is compared to that of the patients in the control group.

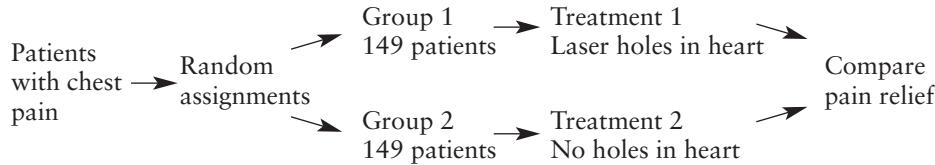
A common way to assign patients to treatment and control groups is by using a random process. This is the essence of a *completely randomized experiment*.

A **completely randomized experiment** is one in which a random process is used to assign each individual to one of the treatments.

EXAMPLE 5**COMPLETELY RANDOMIZED EXPERIMENT**

Can chest pain be relieved by drilling holes in the heart? For more than a decade, surgeons have been using a laser procedure to drill holes in the heart. Many patients report a lasting and dramatic decrease in angina (chest pain) symptoms. Is the relief due to the procedure, or is it a placebo effect? A recent research project at Lenox Hill Hospital in New York City provided some information about this issue by using a completely randomized experiment. The laser treatment was applied through a less invasive (catheter laser) process. A group of 298 volunteers with severe, untreatable chest pain were randomly assigned to get the laser or not. The patients were sedated but awake. They could hear the doctors discuss the laser process. Each patient thought he or she was receiving the treatment.

The experimental design can be pictured as



The laser patients did well. But shockingly, the placebo group showed more improvement in pain relief. The medical impacts of this study are still being investigated.

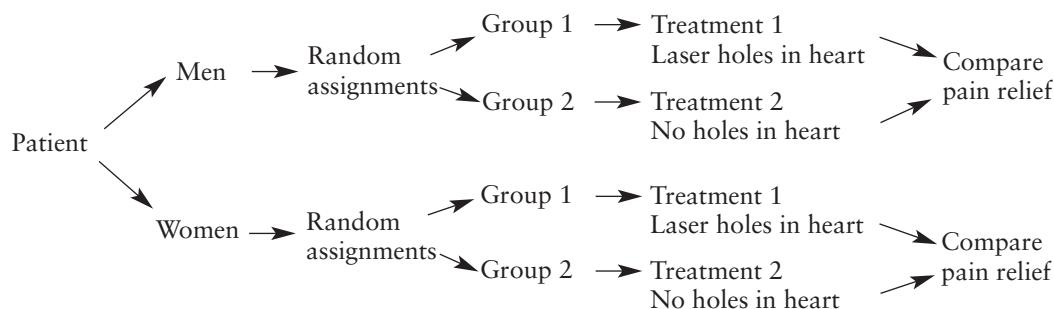
It is difficult to control all the variables that might influence the response to a treatment. One way to control some of the variables is through *blocking*.

Block**Randomized block experiment**

A **block** is a group of individuals sharing some common features that might affect the treatment.

In a **randomized block experiment**, individuals are first sorted into blocks, and then a random process is used to assign each individual in the block to one of the treatments.

A randomized block design utilizing gender for blocks in the experiment involving laser holes in the heart would be



The study cited in Example 5 has many features of good experimental design.

Control group

There is a **control group**. This group receives a dummy treatment, enabling the researchers to control for the placebo effect. In general, a control group is used to account for the influence of other known or unknown variables that might be an underlying cause of a change in response in the experimental group. Such variables are called **lurking** or **confounding variables**.

Randomization

Randomization is used to assign individuals to the two treatment groups. This helps prevent bias in selecting members for each group.

Replication

Replication of the experiment on many patients reduces the possibility that the differences in pain relief for the two groups occurred by chance alone.

Double-blind experiment

Many experiments are also *double-blind*. This means that neither the individuals in the study nor the observers know which subjects are receiving the treatment. Double-blind experiments help control for subtle biases that a doctor might pass on to a patient.

LOOKING FORWARD

One-way and two-way ANOVA (Sections 10.5 and 10.6) are analysis techniques used to study results from completely randomized experiments with several treatments or several blocks with multiple treatments.

GUIDED EXERCISE 5**Collecting data**

Which technique for gathering data (sampling, experiment, simulation, or census) do you think might be the most appropriate for the following studies?

- (a) Study of the effect of stopping the cooling process of a nuclear reactor.



Simulation, since you probably do not want to risk a nuclear meltdown.

Continued

GUIDED EXERCISE 5 *continued*

(b) Study of the amount of time college students taking a full course load spend watching television.

Sampling and using an observational study would work well. Notice that obtaining the information from a student will probably not change the amount of time the student spends watching television.

(c) Study of the effect on bone mass of a calcium supplement given to young girls.

Experimentation. A study by Tom Lloyd reported in the *Journal of the American Medical Association* utilized 94 young girls. Half were randomly selected and given a placebo. The other half were given calcium supplements to bring their daily calcium intake up to about 1400 milligrams per day. The group getting the experimental treatment of calcium gained 1.3% more bone mass in a year than the girls getting the placebo.

(d) Study of the credit hour load of *each* student enrolled at your college at the end of the drop/add period this semester.

Census. The registrar can obtain records for *every* student.



Spencer Grant/PhotoEdit

Likert scale**Survey pitfalls**
Nonresponse**Hidden bias****Voluntary response****Surveys**

Once you decide whether you are going to use sampling, census, observation, or experiments, a common means to gather data about people is to ask them questions. This process is the essence of *surveying*. Sometimes the possible responses are simply yes or no. Other times the respondents choose a number on a scale that represents their feelings from, say, strongly disagree to strongly agree. Such a scale is called a *Likert scale*. In the case of an open-ended, discussion-type response, the researcher must determine a way to convert the response to a category or number.

A number of issues can arise when using a survey.

Some Potential Pitfalls of a Survey

Nonresponse: Individuals either cannot be contacted or refuse to participate. Nonresponse can result in significant undercoverage of a population.

Truthfulness of response: Respondents may lie intentionally or inadvertently.

Faulty recall: Respondents may not accurately remember when or whether an event took place.

Hidden bias: The question may be worded in such a way as to elicit a specific response. The order of questions might lead to biased responses. Also, the number of responses on a Likert scale may force responses that do not reflect the respondent's feelings or experience.

Vague wording: Words such as "often," "seldom," and "occasionally" mean different things to different people.

Interviewer influence: Factors such as tone of voice, body language, dress, gender, authority, and ethnicity of the interviewer might influence responses.

Voluntary response: Individuals with strong feelings about a subject are more likely than others to respond. Such a study is interesting but not reflective of the population.

Lurking and confounding variables

Sometimes our goal is to understand the cause-and-effect relationships between two or more variables. Such studies can be complicated by *lurking variables* or *confounding variables*.

A **lurking variable** is one for which no data have been collected but that nevertheless has influence on other variables in the study.

Two variables are **confounded** when the effects of one cannot be distinguished from the effects of the other. Confounding variables may be part of the study, or they may be outside lurking variables.

For instance, consider a study involving just two variables, amount of gasoline used to commute to work and time to commute to work. Level of traffic congestion is a likely lurking variable that increases both of the study variables. In a study involving several variables such as grade point average, difficulty of courses, IQ, and available study time, some of the variables might be confounded. For instance, students with less study time might opt for easier courses.

Some researchers want to generalize their findings to a situation of wider scope than that of the actual data setting. The true scope of a new discovery must be determined by repeated studies in various real-world settings. Statistical experiments showing that a drug had a certain effect on a collection of laboratory rats do not guarantee that the drug will have a similar effect on a herd of wild horses in Montana.

The sponsorship of a study is another area of concern. Subtle bias may be introduced. For instance, if a pharmaceutical company is paying freelance researchers to work on a study, the researchers may dismiss rare negative findings about a drug or treatment.

Generalizing results**Study sponsor****GUIDED EXERCISE 6****Cautions about data**

Comment on the usefulness of the data collected as described.

- (a) A uniformed police officer interviews a group of 20 college freshmen. She asks each one his or her name and then if he or she has used an illegal drug in the last month.  Respondents may not answer truthfully. Some may refuse to participate.
- (b) Jessica saw some data that show that cities with more low-income housing have more homeless people. Does building low-income housing cause homelessness?  There may be some other confounding or lurking variables, such as the size of the city. Larger cities may have more low-income housing and more homeless.
- (c) A survey about food in the student cafeteria was conducted by having forms available for customers to pick up at the cash register. A drop box for completed forms was available outside the cafeteria.  The voluntary response likely produced more negative comments.
- (d) Extensive studies on coronary problems were conducted using men over age 50 as the subjects.  Conclusions for men over age 50 may or may not generalize to other age and gender groups. These results may be useful for women or younger people, but studies specifically involving these groups may need to be performed.

Choosing Data Collection Techniques

We've briefly discussed three common techniques for gathering data: observational studies, experiments, and surveys. Which technique is best? The answer depends on the number of variables of interest and the level of confidence needed regarding statements of relationships among the variables.

- Surveys may be the best choice for gathering information across a wide range of many variables. Many questions can be included in a survey. However, great care must be taken in the construction of the survey instrument and in the administration of the survey. Nonresponse and other issues discussed earlier can introduce bias.
- Observational studies are the next most convenient technique for gathering information on many variables. Protocols for taking measurements or recording observations need to be specified carefully.
- Experiments are the most stringent and restrictive data-gathering technique. They can be time-consuming, expensive, and difficult to administer. In experiments, the goal is often to study the effects of changing only one variable at a time. Because of the requirements, the number of variables may be more limited. Experiments must be designed carefully to ensure that the resulting data are relevant to the research questions.

COMMENT *An experiment is the best technique for reaching valid conclusions.* By carefully controlling for other variables, the effect of changing one variable in a treatment group and comparing it to a control group yields results carrying high confidence.

The next most effective technique for obtaining results that have high confidence is the use of observational studies. Care must be taken that the act of observation does not change the behavior being measured or observed.

The least effective technique for drawing conclusions is the survey. Surveys have many pitfalls and by their nature cannot give exceedingly precise results. A medical study utilizing a survey asking patients if they feel better after taking a specific drug gives some information, but not precise information about the drug's effects. However, surveys are widely used to gauge attitudes, gather demographic information, study social and political trends, and so on.

VIEWPOINT

Is the Placebo Effect a Myth?

Henry Beecher, former Chief of Anesthesiology at Massachusetts General Hospital, published a paper in the Journal of the American Medical Association (1955) in which he claimed that the placebo effect is so powerful that about 35% of patients would improve simply if they believed a dummy treatment (placebo) was real. However, two Danish medical researchers refute this widely accepted claim in the New England Journal of Medicine. They say the placebo effect is nothing more than a "regression effect," referring to a well-known statistical observation that patients who feel especially bad one day will almost always feel better the next day, no matter what is done for them. However, other respected statisticians question the findings of the Danish researchers. Regardless of the new controversy surrounding the placebo effect, medical researchers agree that placebos are still needed in clinical research. Double-blind research using placebos prevents researchers from inadvertently biasing results.

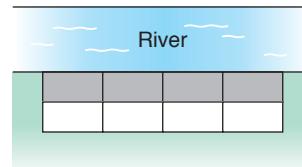
**SECTION 1.3
PROBLEMS**

1. **Statistical Literacy** A study involves three variables: income level, hours spent watching TV per week, and hours spent at home on the Internet per week. List some ways the variables might be confounded.
2. **Statistical Literacy** Consider a completely randomized experiment in which a control group is given a placebo for congestion relief and a treatment group is given a new drug for congestion relief. Describe a double-blind procedure for this experiment and discuss some benefits of such a procedure.
3. **Interpretation** Zane is examining two studies involving how different generations classify specified items as either luxuries or necessities. In the first study, the Echo generation is defined to be people ages 18–29. The second study defined the Echo generation to be people ages 20–31. Zane notices that the first study was conducted in 2006 while the second one was conducted in 2008.
 - (a) Are the two studies inconsistent in their description of the Echo generation?
 - (b) What are the birth years of the Echo generation?
4. **Interpretation** Suppose you are looking at the 2006 results of how the Echo generation classified specified items as either luxuries or necessities. Do you expect the results to reflect how the Echo generation would classify items in 2016? Explain.
5. **Ecology: Gathering Data** Which technique for gathering data (observational study or experiment) do you think was used in the following studies?
 - (a) The Colorado Division of Wildlife netted and released 774 fish at Quincy Reservoir. There were 219 perch, 315 blue gill, 83 pike, and 157 rainbow trout.
 - (b) The Colorado Division of Wildlife caught 41 bighorn sheep on Mt. Evans and gave each one an injection to prevent heartworm. A year later, 38 of these sheep did not have heartworm, while the other three did.
 - (c) The Colorado Division of Wildlife imposed special fishing regulations on the Deckers section of the South Platte River. All trout under 15 inches had to be released. A study of trout before and after the regulation went into effect showed that the average length of a trout increased by 4.2 inches after the new regulation.
 - (d) An ecology class used binoculars to watch 23 turtles at Lowell Ponds. It was found that 18 were box turtles and 5 were snapping turtles.
6. **General: Gathering Data** Which technique for gathering data (sampling, experiment, simulation, or census) do you think was used in the following studies?
 - (a) An analysis of a sample of 31,000 patients from New York hospitals suggests that the poor and the elderly sue for malpractice at one-fifth the rate of wealthier patients (*Journal of the American Medical Association*).
 - (b) The effects of wind shear on airplanes during both landing and takeoff were studied by using complex computer programs that mimic actual flight.
 - (c) A study of all league football scores attained through touchdowns and field goals was conducted by the National Football League to determine whether field goals account for more scoring events than touchdowns (*USA Today*).
 - (d) An Australian study included 588 men and women who already had some precancerous skin lesions. Half got a skin cream containing a sunscreen with a sun protection factor of 17; half got an inactive cream. After 7 months, those using the sunscreen with the sun protection had fewer new precancerous skin lesions (*New England Journal of Medicine*).

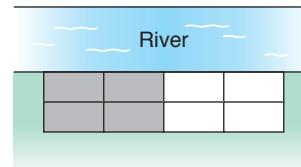
7. **General: Completely Randomized Experiment** How would you use a completely randomized experiment in each of the following settings? Is a placebo being used or not? Be specific and give details.
- A veterinarian wants to test a strain of antibiotic on calves to determine their resistance to common infection. In a pasture are 22 newborn calves. There is enough vaccine for 10 calves. However, blood tests to determine resistance to infection can be done on all calves.
 - The Denver Police Department wants to improve its image with teenagers. A uniformed officer is sent to a school one day a week for 10 weeks. Each day the officer visits with students, eats lunch with students, attends pep rallies, and so on. There are 18 schools, but the police department can visit only half of these schools this semester. A survey regarding how teenagers view police is sent to all 18 schools at the end of the semester.
 - A skin patch contains a new drug to help people quit smoking. A group of 75 cigarette smokers have volunteered as subjects to test the new skin patch. For one month, 40 of the volunteers receive skin patches with the new drug. The other volunteers receive skin patches with no drugs. At the end of two months, each subject is surveyed regarding his or her current smoking habits.
8. **Surveys: Manipulation** The *New York Times* did a special report on polling that was carried in papers across the nation. The article pointed out how readily the results of a survey can be manipulated. Some features that can influence the results of a poll include the following: the number of possible responses, the phrasing of the questions, the sampling techniques used (voluntary response or sample designed to be representative), the fact that words may mean different things to different people, the questions that precede the question of interest, and finally, the fact that respondents can offer opinions on issues they know nothing about.
- Consider the expression “over the last few years.” Do you think that this expression means the same time span to everyone? What would be a more precise phrase?
 - Consider this question: “Do you think fines for running stop signs should be doubled?” Do you think the response would be different if the question “Have you ever run a stop sign?” preceded the question about fines?
 - Consider this question: “Do you watch too much television?” What do you think the responses would be if the only responses possible were yes or no? What do you think the responses would be if the possible responses were “rarely,” “sometimes,” or “frequently”?
9. **Critical Thinking** An agricultural study is comparing the harvest volume of two types of barley. The site for the experiment is bordered by a river. The field is divided into eight plots of approximately the same size. The experiment calls for the plots to be blocked into four plots per block. Then, two plots of each block will be randomly assigned to one of the two barley types.

Two blocking schemes are shown below, with one block indicated by the white region and the other by the grey region. Which blocking scheme, A or B, would be better? Explain.

Scheme A



Scheme B





SUMMARY

Chapter Review

In this chapter, you've seen that statistics is the study of how to collect, organize, analyze, and interpret numerical information from populations or samples. This chapter discussed some of the features of data and ways to collect data. In particular, the chapter discussed

- Individuals or subjects of a study and the variables associated with those individuals
- Data classification as qualitative or quantitative, and levels of measurement of data
- Sample and population data. Summary measurements from sample data are called statistics, and those from populations are called parameters.

- Sampling strategies, including simple random, stratified, systematic, multistage, and convenience. Inferential techniques presented in this text are based on simple random samples.
- Methods of obtaining data: Use of a census, simulation, observational studies, experiments, and surveys
- Concerns: Undercoverage of a population, nonresponse, bias in data from surveys and other factors, effects of confounding or lurking variables on other variables, generalization of study results beyond the population of the study, and study sponsorship

IMPORTANT WORDS & SYMBOLS

Section 1.1*

- Statistics 4
 Individuals 5
 Variable 5
 Quantitative variable 5
 Qualitative variable 5
 Population data 5
 Sample data 5
 Population parameter 5
 Sample statistic 5
 Levels of measurement 7
 Nominal 7
 Ordinal 7
 Interval 7
 Ratio 7
 Descriptive statistics 10
 Inferential statistics 10

Section 1.2

- Simple random sample 13
 Random-number table 13
 Simulation 15
 Sampling with replacement 15
 Stratified sampling 16
 Systematic sampling 16
 Cluster sampling 16
 Multistage sample 16
 Convenience sampling 16

- Sampling frame 17
 Undercoverage 17
 Sampling error 17
 Nonsampling error 18

Section 1.3

- Census 21
 Observational study 22
 Experiment 22
 Placebo effect 23
 Treatment group 23
 Completely randomized experiment 23
 Block 24
 Randomized block experiment 24
 Control group 24
 Randomization 24
 Replication 24
 Double-blind experiment 24
 Survey 25
 Likert scale 25
 Nonresponse 25
 Hidden bias 25
 Voluntary response 25
 Lurking variable 26
 Confounding variable 26
 Generalizing results 26
 Study sponsor 26

*Indicates section of first appearance.

VIEWPOINT**Is Chocolate Good for Your Heart?**

A study of 7841 Harvard alumni showed that the death rate was 30% lower in those who ate candy compared with those who abstained. It turns out that candy, especially chocolate, contains antioxidants that help slow the aging process. Also, chocolate, like aspirin, reduces the activity of blood platelets that contribute to plaque and blood clotting. Furthermore, chocolate seems to raise levels of high-density lipoprotein (HDL), the good cholesterol. However, these results are all preliminary. The investigation is far from complete. A wealth of information on this topic was published in the August 2000 issue of the Journal of Nutrition. Statistical studies and reliable experimental design are indispensable in this type of research.

CHAPTER REVIEW PROBLEMS

Joe McDaniel/Stockphoto.com



1. **Critical Thinking** Sudoku is a puzzle consisting of squares arranged in 9 rows and 9 columns. The 81 squares are further divided into nine 3×3 square boxes. The object is to fill in the squares with numerals 1 through 9 so that each column, row, and box contains all nine numbers. However, there is a requirement that each number appear only once in any row, column, or box. Each puzzle already has numbers in some of the squares. Would it be appropriate to use a random-number table to select a digit for each blank square? Explain.
2. **Critical Thinking** Alisha wants to do a statistical study to determine how long it takes people to complete a Sudoku puzzle (see Problem 1 for a description of the puzzle). Her plan is as follows:
 - Download 10 different puzzles from the Internet.
 - Find 10 friends willing to participate.
 - Ask each friend to complete one of the puzzles and time him- or herself.
 - Gather the completion times from each friend.
 Describe some of the problems with Alisha's plan for the study. (Note: Puzzles differ in difficulty, ranging from beginner to very difficult.) Are the results from Alisha's study anecdotal, or do they apply to the general population?
3. **Statistical Literacy** You are conducting a study of students doing work-study jobs on your campus. Among the questions on the survey instrument are:
 - How many hours are you scheduled to work each week? Answer to the nearest hour.
 - How applicable is this work experience to your future employment goals?
 Respond using the following scale: 1 = not at all, 2 = somewhat, 3 = very
 - Suppose you take random samples from the following groups: freshmen, sophomores, juniors, and seniors. What kind of sampling technique are you using (simple random, stratified, systematic, cluster, multistage, convenience)?
 - Describe the individuals of this study.
 - What is the variable for question A? Classify the variable as qualitative or quantitative. What is the level of the measurement?
 - What is the variable for question B? Classify the variable as qualitative or quantitative. What is the level of the measurement?
 - Is the proportion of responses "3 = very" to question B a statistic or a parameter?
 - Suppose only 40% of the students you selected for the sample respond. What is the nonresponse rate? Do you think the nonresponse rate might introduce bias into the study? Explain.
 - Would it be appropriate to generalize the results of your study to all work-study students in the nation? Explain.

4. **Radio Talk Show: Sample Bias** A radio talk show host asked listeners to respond either yes or no to the question, “Is the candidate who spends the most on a campaign the most likely to win?” Fifteen people called in and nine said yes. What is the implied population? What is the variable? Can you detect any bias in the selection of the sample?
5. **Simulation: TV Habits** One cable station knows that approximately 30% of its viewers have TIVO and can easily skip over advertising breaks. You are to design a simulation of how a random sample of seven station viewers would respond to the question, “Do you have TIVO?” How would you assign the random digits 0 through 9 to the responses “Yes” or “No” to the TIVO question? Use your random-digit assignment and the random-number table to generate the responses from a random sample of seven station viewers.
6. **General: Type of Sampling** Categorize the type of sampling (simple random, stratified, systematic, cluster, or convenience) used in each of the following situations.
- To conduct a preelection opinion poll on a proposed amendment to the state constitution, a random sample of 10 telephone prefixes (first three digits of the phone number) was selected, and all households from the phone prefixes selected were called.
 - To conduct a study on depression among the elderly, a sample of 30 patients in one nursing home was used.
 - To maintain quality control in a brewery, every 20th bottle of beer coming off the production line was opened and tested.
 - Subscribers to the magazine *Sound Alive* were assigned numbers. Then a sample of 30 subscribers was selected by using a random-number table. The subscribers in the sample were invited to rate new compact disc players for a “What the Subscribers Think” column.
 - To judge the appeal of a proposed television sitcom, a random sample of 10 people from each of three different age categories was selected and those chosen were asked to rate a pilot show.
7. **General: Gathering Data** Which technique for gathering data (observational study or experiment) do you think was used in the following studies? Explain.
- The U.S. Census Bureau tracks population age. In 1900, the percentage of the population that was 19 years old or younger was 44.4%. In 1930, the percentage was 38.8%; in 1970, the percentage was 37.9%; and in 2000, the percentage in that age group was down to 28.5% (Reference: *The First Measured Century*, T. Caplow, L. Hicks, and B. J. Wattenberg).
 - After receiving the same lessons, a class of 100 students was randomly divided into two groups of 50 each. One group was given a multiple-choice exam covering the material in the lessons. The other group was given an essay exam. The average test scores for the two groups were then compared.
8. **General: Experiment** How would you use a completely randomized experiment in each of the following settings? Is a placebo being used or not? Be specific and give details.
- A charitable nonprofit organization wants to test two methods of fundraising. From a list of 1000 past donors, half will be sent literature about the successful activities of the charity and asked to make another donation. The other 500 donors will be contacted by phone and asked to make another donation. The percentage of people from each group who make a new donation will be compared.
 - A tooth-whitening gel is to be tested for effectiveness. A group of 85 adults have volunteered to participate in the study. Of these, 43 are to be given a gel that contains the tooth-whitening chemicals. The remaining 42 are to be given a similar-looking package of gel that does not contain the tooth-whitening chemicals. A standard method will be used to evaluate the whiteness of teeth for all participants. Then the results for the two groups will be compared. How could this experiment be designed to be double-blind?

- (c) Consider the experiment described in part (a). Describe how you would use a randomized block experiment with blocks based on age. Use three blocks: donors under 30 years old, donors 30 to 59 years old, donors 60 and over.
9. **Student Life: Data Collection Project** Make a statistical profile of your own statistics class. Items of interest might be
- Height, age, gender, pulse, number of siblings, marital status
 - Number of college credit hours completed (as of beginning of term); grade point average
 - Major; number of credit hours enrolled in this term
 - Number of scheduled work hours per week
 - Distance from residence to first class; time it takes to travel from residence to first class
 - Year, make, and color of car usually driven
- What directions would you give to people answering these questions? For instance, how accurate should the measurements be? Should age be recorded as of last birthday?
10. **Census: Web Site** *Census and You*, a publication of the Census Bureau, indicates that “Wherever your Web journey ends up, it should start at the Census Bureau’s site.” Visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find a link to the Census Bureau’s site, as well as to Fedstats, another extensive site offering links to federal data. The Census Bureau site touts itself as the source of “official statistics.” But it is willing to share the spotlight. The web site now has links to other “official” sources: other federal agencies, foreign statistical agencies, and state data centers. If you have access to the Internet, try the Census Bureau’s site.
11. **Focus Problem: Fireflies** Suppose you are conducting a study to compare firefly populations exposed to normal daylight/darkness conditions with firefly populations exposed to continuous light (24 hours a day). You set up two firefly colonies in a laboratory environment. The two colonies are identical except that one colony is exposed to normal daylight/darkness conditions and the other is exposed to continuous light. Each colony is populated with the same number of mature fireflies. After 72 hours, you count the number of living fireflies in each colony.
- Is this an experiment or an observation study? Explain.
 - Is there a control group? Is there a treatment group?
 - What is the variable in this study?
 - What is the level of measurement (nominal, interval, ordinal, or ratio) of the variable?

**DATA HIGHLIGHTS:
GROUP PROJECTS**

1. Use a random-number table or random-number generator to simulate tossing a fair coin 10 times. Generate 20 such simulations of 10 coin tosses. Compare the simulations. Are there any strings of 10 heads? of 4 heads? Does it seem that in most of the simulations, half the outcomes are heads? half are tails? In Chapter 5, we will study the probabilities of getting from 0 to 10 heads in such a simulation.
2. Use a random-number table or random-number generator to generate a random sample of 30 distinct values from the set of integers 1 to 100. Instructions for doing this using the TI-84Plus/TI-83Plus/TI-nspire (with TI-84 Plus keypad), Excel 2007, Minitab, or SPSS are given in Using Technology at the end of this chapter. Generate five such samples. How many of the samples include the number 1? the number 100? Comment about the differences among the samples. How well do the samples seem to represent the numbers between 1 and 100?

**LINKING CONCEPTS:
WRITING PROJECTS**

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. What does it mean to say that we are going to use a sample to draw an inference about a population? Why is a random sample so important for this process? If we wanted a random sample of students in the cafeteria, why couldn't we just choose the students who order Diet Pepsi with their lunch? Comment on the statement, "A random sample is like a miniature population, whereas samples that are not random are likely to be biased." Why would the students who order Diet Pepsi with lunch not be a random sample of students in the cafeteria?
2. In your own words, explain the differences among the following sampling techniques: simple random sample, stratified sample, systematic sample, cluster sample, multistage sample, and convenience sample. Describe situations in which each type might be useful.

USING TECHNOLOGY

General spreadsheet programs such as Microsoft's Excel, specific statistical software packages such as Minitab or SPSS, and graphing calculators such as the TI-84Plus/TI-83Plus/TI-nspire all offer computing support for statistical methods. Applications in this section may be completed using software or calculators with statistical functions. Select keystroke or menu choices are shown for the TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) calculators, Minitab, Excel 2007, and SPSS in the Technology Hints portion of this section. More details can be found in the software-specific *Technology Guide* that accompanies this text.

Applications

Most software packages sample *with replacement*. That is, the same number may be used more than once in the sample. If your applications require sampling without replacement, draw more items than you need. Then use sort commands in the software to put the data in order, and delete repeated data.

1. Simulate the results of tossing a fair die 18 times. Repeat the simulation. Are the results the same? Did you expect them to be the same? Why or why not? Do there appear to be equal numbers of outcomes 1 through 6 in each simulation? In Chapter 4, we will encounter the law of large numbers, which tells us that we would expect equal numbers of outcomes only when the simulation is very large.
2. A college has 5000 students, and the registrar wishes to use a random sample of 50 students to examine credit hour enrollment for this semester. Write a brief description of how a random sample can be drawn. Draw a random sample of 50 students. Are you sampling with or without replacement?

Technology Hints: Random Numbers

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad)

The TI-nspire calculator with the TI-84Plus keypad installed works exactly like other TI-84Plus calculators. Instructions for the TI-84Plus, TI-83Plus, and the TI-nspire (with the TI-84 Plus keypad installed) calculators are included directly in this text as well as in a separate *Technology Guide*. When the nspire keypad is installed, the required keystrokes and screen displays are different from those with the TI-84Plus keypad.

A separate *Technology Guide* for this text provides instructions for using the nspire keypad to perform statistical operations, create statistical graphs, and apply statistical tests.

The instructions that follow apply to the TI-84Plus, TI-83Plus, and TI-nspire (with the TI-84 keypad installed) calculators.

To select a random set of integers between two specified values, press the **MATH** key and highlight **PRB** with **5:randInt** (low value, high value, sample size). Press Enter and fill in the low value, high value, and sample size. To store the sample in list L1, press the **STO→** key and then L1. The screen display shows two random samples of size 5 drawn from the integers between 1 and 100.

```
randInt(1,100,5)
{63 89 13 46 47}
randInt(1,100,5)
{29 82 99 50 41}
```

Excel 2007

Many statistical processes that we will use throughout this text are included in the **Analysis ToolPak** add-in of Excel. This is an add-in that is included with the standard versions of Excel 2007. To see if you have the add-in

installed, click the Office Button  in the upper-left corner of the spreadsheet, press the **Excel Options** button, and select **Add-ins**. Check that **Analysis ToolPak** is in the Active Application Add-ins list. If it is not, select it for an active application.

To select a random number between two specified integer values, first select a cell in the active worksheet. Then type the command =RANDBETWEEN(bottom number, top number) in the formula bar, where the bottom number is the lower specified value and the top number is the higher specified value.

Alternatively, access a dialogue box for the command by clicking the **insert function** button  on the ribbon. You will see an insert function dialogue box. Select the category **All**, and then scroll down until you reach **RANDBETWEEN** and click **OK**. Fill in the bottom and top numbers. In the display shown, the bottom number is 1 and the top number is 100.

RANDBETWEEN					
					=RANDBETWEEN(1,100)
1	46				
2	47				
3	49				
4	73				
5	11				

Minitab

To generate random integers between specified values, use the menu selection **Calc > Random Data > Integer**. Fill in the dialogue box to get five random numbers between 1 and 100.

Worksheet 2 ***	
	C1
↓	
1	8
2	35
3	33
4	9
5	15

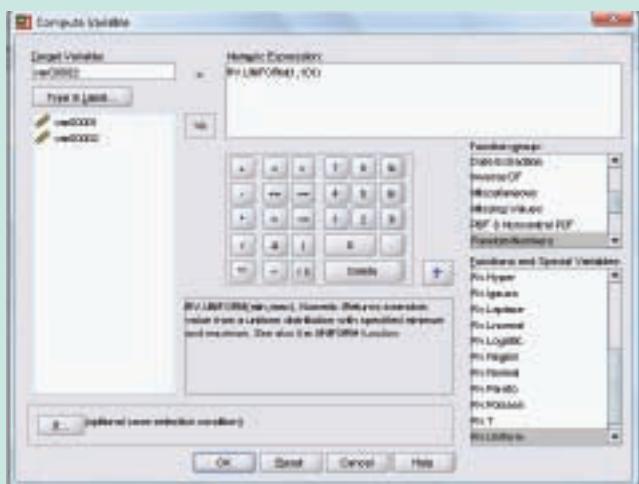
SPSS

SPSS is a research statistical package for the social sciences. Data are entered in the data editor, which has a spreadsheet format. In the data editor window, you have a choice of data view (default) or variable view. In the variable view, you name variables, declare type (numeric for measurements, string for category), determine format, and declare measurement type. The choices for measurement type are scale (for ratio or interval data), ordinal, or nominal. Once you have entered data, you can use the menu bar at the top of the screen to select activities, graphs, or analysis appropriate to the data.

SPSS supports several random sample activities. In particular, you can select a random sample from an existing data set or from a variety of probability distributions.

Selecting a random integer between two specified values involves several steps. First, in the data editor, enter the sample numbers in the first column. For instance, to generate five random numbers, list the values 1 through 5 in the first column. Notice that the label for the first column is now var00001. SPSS does not have a direct function for selecting a random sample of integers. However, there is a function for sampling values from the uniform distribution of all real numbers between two specified values. We will use that function and then truncate the values to obtain a random sample of integers between two specified values.

Use the menu options **Transform > Compute**. In the dialogue box, type in var00002 as the target variable. In the Function group select **Random Number**. Then under Functions and Special Variables, select **Rv.Uniform**. In the Numeric Expression box, replace the two question marks by the minimum and maximum. Use 1 as the minimum and 101 as the maximum. The maximum is 101 because numbers between 100 and 101 truncate to 100.



The random numbers from the uniform distribution now appear in the second column under var00002. You can visually truncate the values to obtain random integers. However, if you want SPSS to truncate the values for you, you can again use the menu choices **Transform > Compute**. In the dialogue box, enter var00003 for the target variable. In the Function Group select **Arithmetic**. Then in the Functions and Special Variables box select **Trunc(1)**. In the Numeric Expression box use var00002 in place of the question mark representing numexpr. The random integers between 1 and 100 appear in the third column under var00003.

	var00001	var00002	var00003
1	1.00	14.96	14.00
2	2.00	44.13	44.00
3	3.00	62.22	62.00
4	4.00	30.00	30.00
5	5.00	16.57	16.00



2

- 2.1 Frequency Distributions, Histograms, and Related Topics
- 2.2 Bar Graphs, Circle Graphs, and Time-Series Graphs
- 2.3 Stem-and-Leaf Displays



Fyodor Savintsev/ITAR-TASS/Corbis News/Corbis



World History Archive/Alamy

In dwelling upon the vital importance of sound observation, it must never be lost sight of what observation is for. It is not for the sake of piling up miscellaneous information or curious facts, but for the sake of saving life and increasing health and comfort.

—FLORENCE NIGHTINGALE,
Notes on Nursing

Florence Nightingale (1820–1910) has been described as a “passionate statistician” and a “relevant statistician.” She viewed statistics as a science that allowed one to transcend his or her narrow individual experience and aspire to the broader service of humanity. She was one of the first nurses to use graphic representation of statistics, illustrating with charts and diagrams how improved sanitation decreased the rate of mortality. Her statistical reports about the appalling sanitary conditions at Scutari (the main British hospital during the Crimean War) were taken very seriously by the English Secretary at War, Sidney Herbert. When sanitary reforms recommended by Nightingale were instituted in military hospitals, the mortality rate dropped from an incredible 42.7% to only 2.2%.

For online student resources, visit the Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>

ORGANIZING DATA

PREVIEW QUESTIONS

What are histograms? When are they used? (SECTION 2.1)

What are common distribution shapes? (SECTION 2.1)

How can you select graphs appropriate for given data sets? (SECTION 2.2)

How can you quickly order data and, at the same time, reveal the distribution shape? (SECTION 2.3)



Sportstock/iStockphoto

FOCUS PROBLEM

Say It with Pictures

Edward R. Tufte, in his book *The Visual Display of Quantitative Information*, presents a number of guidelines for producing good graphics. According to the criteria, a graphical display should

- show the data;
- induce the viewer to think about the substance of the graphic rather than about the methodology, the design, the technology, or other production devices;
- avoid distorting what the data have to say.

As an example of a graph that violates some of the criteria, Tufte includes a graphic that appeared in a well-known newspaper. Figure 2-1(a), on the next page, shows a facsimile of the problem graphic, whereas part (b) of the figure shows a better rendition of the data display.

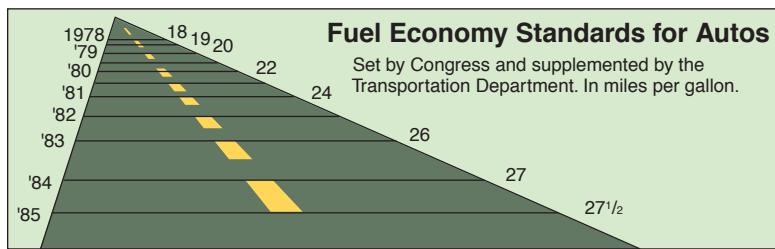
After completing this chapter, you will be able to answer the following questions.

- (a) Look at the graph in Figure 2-1(a). Is it essentially a bar graph? Explain. What are some of the flaws of Figure 2-1(a) as a bar graph?
- (b) Examine Figure 2-1(b), which shows the same information. Is it essentially a time-series graph? Explain. In what ways does the second graph seem to display the information in a clearer manner?

(See Problem 5 of the Chapter 2 Review Problems.)

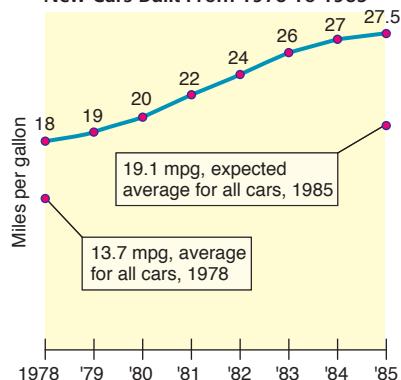


Vandykstadt/Photo Researchers Inc

FIGURE 2-1**(a) Fuel Economy Standards for Autos**

Source: Copyright © 1978 by The New York Times Company. Reprinted by permission.

Note: Fuel-economy standards for passenger cars remained at 27.5 miles per gallon through 2009. It is anticipated that the mileage standard will increase soon.

(b) Required Fuel Economy Standards: New Cars Built From 1978 To 1985

Source: Redrawn from *The Visual Display of Quantitative Information* by Edward R. Tufte, p. 57. Copyright © 1983. Reprinted by permission of Graphics Press.

SECTION 2.1**Frequency Distributions, Histograms, and Related Topics****FOCUS POINTS**

- Organize raw data using a frequency table.
- Construct histograms, relative-frequency histograms, and ogives.
- Recognize basic distribution shapes: uniform, symmetric, skewed, and bimodal.
- Interpret graphs in the context of the data setting.

Frequency Tables

When we have a large set of quantitative data, it's useful to organize it into smaller intervals or *classes* and count how many data values fall into each class. A frequency table does just that.

Frequency table

A frequency table partitions data into classes or intervals and shows how many data values are in each class. The classes or intervals are constructed so that each data value falls into exactly one class.

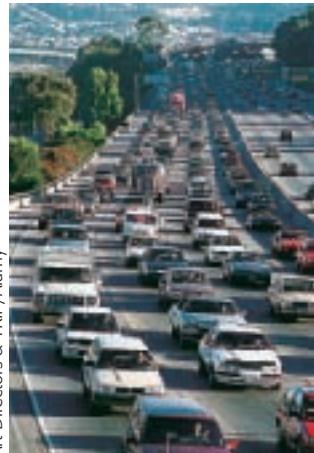
Constructing a frequency table involves a number of steps. Example 1 demonstrates the steps.

**EXAMPLE 1****FREQUENCY TABLE**

A task force to encourage car pooling did a study of one-way commuting distances of workers in the downtown Dallas area. A random sample of 60 of these workers was taken. The commuting distances of the workers in the sample are given in Table 2-1. Make a frequency table for these data.

SOLUTION:

- (a) First decide how many classes you want. Five to 15 classes are usually used. If you use fewer than five classes, you risk losing too much information. If you use more than 15 classes, the data may not be sufficiently summarized. Let the



Art Directors & TRIP/Alamy

Class width**TABLE 2-1 One-Way Commuting Distances (in Miles) for 60 Workers in Downtown Dallas**

13	47	10	3	16	20	17	40	4	2
7	25	8	21	19	15	3	17	14	6
12	45	1	8	4	16	11	18	23	12
6	2	14	13	7	15	46	12	9	18
34	13	41	28	36	17	24	27	29	9
14	26	10	24	37	31	8	16	12	16

spread of the data and the purpose of the frequency table be your guides when selecting the number of classes. In the case of the commuting data, let's use *six* classes.

- (b) Next, find the *class width* for the six classes.

PROCEDURE**HOW TO FIND THE CLASS WIDTH (INTEGER DATA)**

1. Compute $\frac{\text{Largest data value} - \text{smallest data value}}{\text{Desired number of classes}}$
2. Increase the computed value to the next highest whole number.

Note: To ensure that all the classes taken together cover the data, we need to increase the result of Step 1 to the *next whole number*, even if Step 1 produced a whole number. For instance, if the calculation in Step 1 produces the value 4, we make the class width 5.

To find the class width for the commuting data, we observe that the largest distance commuted is 47 miles and the smallest is 1 mile. Using six classes, the class width is 8, since

$$\text{Class width} = \frac{47 - 1}{6} \approx 7.7 \quad (\text{increase to } 8)$$

- (c) Now we determine the data range for each class.

Class limits

The **lower class limit** is the lowest data value that can fit in a class. The **upper class limit** is the highest data value that can fit in a class. The **class width** is the difference between the *lower class limit* of one class and the *lower class limit* of the next class.

The smallest commuting distance in our sample is 1 mile. We use this *smallest* data value as the lower class limit of the *first* class. Since the class width is 8, we add 8 to 1 to find that the *lower class limit* for the *second* class is 9. Following this pattern, we establish *all* the *lower class limits*. Then we fill in the *upper class limits* so that the classes span the entire range of data. Table 2-2, on the next page, shows the upper and lower class limits for the commuting distance data.

- (d) Now we are ready to tally the commuting distance data into the six classes and find the frequency for each class.

TABLE 2-2

Frequency Table of One-Way Commuting Distances for 60 Downtown Dallas Workers (Data in Miles)

Class Limits Lower–Upper	Class Boundaries Lower–Upper	Tally	Frequency	Class Midpoint
1–8	0.5–8.5		14	4.5
9–16	8.5–16.5		21	12.5
17–24	16.5–24.5		11	20.5
25–32	24.5–32.5		6	28.5
33–40	32.5–40.5		4	36.5
41–48	40.5–48.5		4	44.5

PROCEDURE**HOW TO TALLY DATA**

Tallying data is a method of counting data values that fall into a particular class or category.

To tally data into classes of a frequency table, examine each data value. Determine which class contains the data value and make a tally mark or vertical stroke (|) beside that class. For ease of counting, each fifth tally mark of a class is placed diagonally across the prior four marks (NN).

Class frequency

The *class frequency* for a class is the number of tally marks corresponding to that class.

Table 2-2 shows the tally and frequency of each class.

Class midpoint or class mark

- (e) The center of each class is called the *midpoint* (or *class mark*). The midpoint is often used as a representative value of the entire class. The midpoint is found by adding the lower and upper class limits of one class and dividing by 2.

$$\text{Midpoint} = \frac{\text{Lower class limit} + \text{upper class limit}}{2}$$

Table 2-2 shows the class midpoints.

Class boundaries

- (f) There is a space between the upper limit of one class and the lower limit of the next class. The halfway points of these intervals are called *class boundaries*. These are shown in Table 2-2.

PROCEDURE

Problems 17 and 18 show the procedure for finding the class width, class limits, and class boundaries for noninteger decimal data.

HOW TO FIND CLASS BOUNDARIES (INTEGER DATA)

To find **upper class boundaries**, add 0.5 unit to the upper class limits.

To find **lower class boundaries**, subtract 0.5 unit from the lower class limits.



Relative frequency

Basic frequency tables show how many data values fall into each class. It's also useful to know the *relative frequency* of a class. The relative frequency of a class is the proportion of all data values that fall into that class. To find the

TABLE 2-3 Relative Frequencies of One-Way Commuting Distances

Class	Frequency f	Relative Frequency f/n
1–8	14	14/60 ≈ 0.23
9–16	21	21/60 ≈ 0.35
17–24	11	11/60 ≈ 0.18
25–32	6	6/60 ≈ 0.10
33–40	4	4/60 ≈ 0.07
41–48	4	4/60 ≈ 0.07

relative frequency of a particular class, divide the class frequency f by the total of all frequencies n (sample size).

$$\text{Relative frequency} = \frac{f}{n} = \frac{\text{Class frequency}}{\text{Total of all frequencies}}$$

Table 2-3 shows the relative frequencies for the commuter data of Table 2-1. Since we already have the frequency table (Table 2-2), the relative-frequency table is obtained easily. The sample size is $n = 60$. Notice that the sample size is the total of all the frequencies. Therefore, the relative frequency for the first class (the class from 1 to 8) is

$$\text{Relative frequency} = \frac{f}{n} = \frac{14}{60} \approx 0.23$$

The symbol \approx means “approximately equal to.” We use the symbol because we rounded the relative frequency. Relative frequencies for the other classes are computed in a similar way.

The total of the relative frequencies should be 1. However, rounded results may make the total slightly higher or lower than 1.

LOOKING FORWARD

Sorting data into classes can be somewhat tedious. It is an easier task if the data are first ordered from smallest to largest. Stem-and-leaf diagrams, presented in Section 2.3, provide a convenient way to order data by hand.

Let’s summarize the procedure for making a frequency table that includes relative frequencies.

PROCEDURE

HOW TO MAKE A FREQUENCY TABLE

1. Determine the number of classes and the corresponding class width.
2. Create the distinct classes. We use the convention that the *lower class limit* of the first class is the smallest data value. Add the class width to this number to get the *lower class limit* of the next class.
3. Fill in *upper class limits* to create distinct classes that accommodate all possible data values from the data set.
4. Tally the data into classes. Each data value should fall into exactly one class. Total the tallies to obtain each *class frequency*.
5. Compute the *midpoint* (class mark) for each class.
6. Determine the *class boundaries*.

PROCEDURE

Relative-frequency table

HOW TO MAKE A RELATIVE-FREQUENCY TABLE

First make a frequency table. Then, for each class, compute the *relative frequency*, f/n , where f is the class frequency and n is the total sample size.

Histogram

Relative-frequency histogram

Histograms and Relative-Frequency Histograms

Histograms and *relative-frequency histograms* provide effective visual displays of data organized into frequency tables. In these graphs, we use bars to represent each class, where the width of the bar is the class width. For histograms, the height of the bar is the class frequency, whereas for relative-frequency histograms, the height of the bar is the relative frequency of that class.

PROCEDURE**HOW TO MAKE A HISTOGRAM OR A RELATIVE-FREQUENCY HISTOGRAM**

1. Make a frequency table (including relative frequencies) with the designated number of classes.
2. Place class boundaries on the horizontal axis and frequencies or relative frequencies on the vertical axis.
3. For each class of the frequency table, draw a bar whose width extends between corresponding class boundaries. For histograms, the height of each bar is the corresponding class frequency. For relative-frequency histograms, the height of each bar is the corresponding class relative frequency.

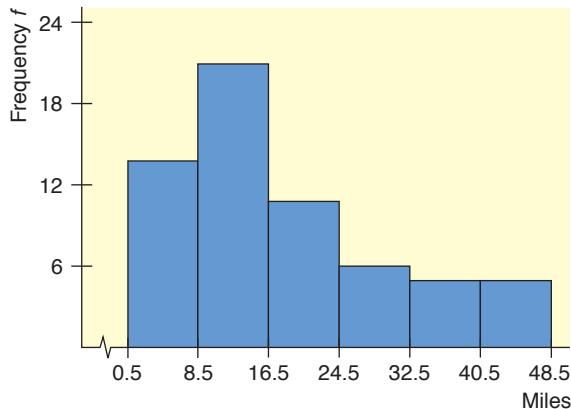
EXAMPLE 2**HISTOGRAM AND RELATIVE-FREQUENCY HISTOGRAM**

Make a histogram and a relative-frequency histogram with six bars for the data in Table 2-1 showing one-way commuting distances.

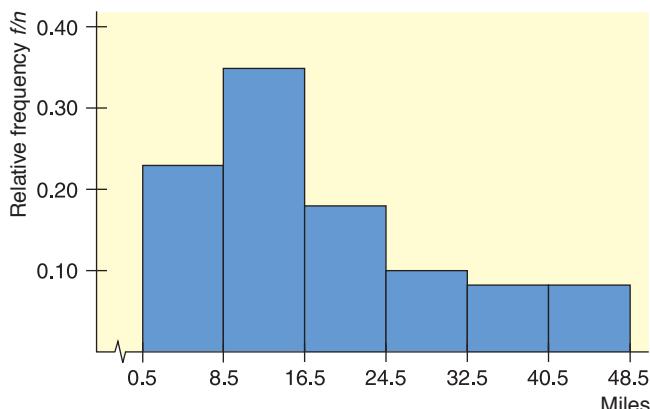
SOLUTION: The first step is to make a frequency table and a relative-frequency table with six classes. We'll use Table 2-2 and Table 2-3. Figures 2-2 and 2-3

FIGURE 2-2

Histogram for Dallas Commuters:
One-Way Commuting Distances

**FIGURE 2-3**

Relative-Frequency Histogram for Dallas Commuters:
One-Way Commuting Distances



show the histogram and relative-frequency histogram. In both graphs, class boundaries are marked on the horizontal axis. For each class of the frequency table, make a corresponding bar with horizontal width extending from the lower boundary to the upper boundary of the respective class. For a histogram, the height of each bar is the corresponding class frequency. For a relative-frequency histogram, the height of each bar is the corresponding relative frequency. Notice that the basic shapes of the graphs are the same. The only difference involves the vertical axis. The vertical axis of the histogram shows frequencies, whereas that of the relative-frequency histogram shows relative frequencies.

COMMENT The use of class boundaries in histograms assures us that the bars of the histogram touch and that no data fall on the boundaries. Both of these features are important. But a histogram displaying class boundaries may look awkward. For instance, the mileage range of 8.5 to 16.5 miles shown in Figure 2-2 isn't as natural a choice as a mileage range of 8 to 16 miles. For this reason, many magazines and newspapers do not use class boundaries as labels on a histogram. Instead, some use lower class limits as labels, with the convention that *a data value falling on the class limit is included in the next higher class (class to the right of the limit)*. Another convention is to label midpoints instead of class boundaries. Determine the default convention being used before creating frequency tables and histograms on a computer.

GUIDED EXERCISE 1

Histogram and relative-frequency histogram

An irate customer called Dollar Day Mail Order Company 40 times during the last two weeks to see why his order had not arrived. Each time he called, he recorded the length of time he was put “on hold” before being allowed to talk to a customer service representative. See Table 2-4.

TABLE 2-4 Length of Time on Hold, in Minutes

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

- (a) What are the largest and smallest values in Table 2-4? If we want five classes in a frequency table, what should the class width be?

The largest value is 13; the smallest value is 1. The class width is

$$\frac{13 - 1}{5} = 2.4 \approx 3 \quad \text{Note: Increase the value to 3.}$$

- (b) Complete the following frequency table.

TABLE 2-5 Time on Hold

Class Limits		Frequency	Midpoint
Lower-Upper	Tally		
1–3	____	_____	_____
4–6	____	_____	_____
7–9	____	_____	_____
10–12	____	_____	_____
13–15	____	_____	_____

TABLE 2-6 Completion of Table 2-5

Class Limits		Frequency	Midpoint
Lower-Upper	Tally		
1–3		3	2
4–6		5	5
7–9		5	8
10–12		4	11
13–15		1	14

Continued

GUIDED EXERCISE 1 *continued*

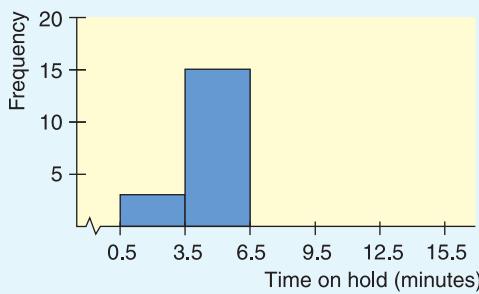
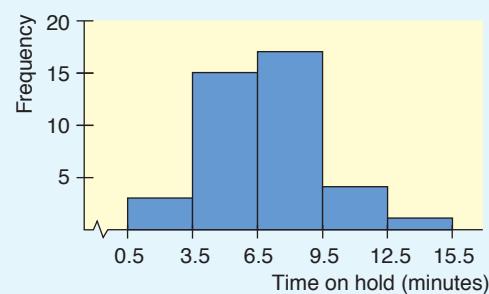
- (c) Recall that the class boundary is halfway between the upper limit of one class and the lower limit of the next. Use this fact to find the class boundaries in Table 2-7 and to complete the partial histogram in Figure 2-4.

TABLE 2-7 Class Boundaries

Class Limits	Class Boundaries
1–3	0.5–3.5
4–6	3.5–6.5
7–9	6.5–_____
10–12	_____–_____
13–15	_____–_____

TABLE 2-8 Completion of Table 2-7

Class Limits	Class Boundaries
1–3	0.5–3.5
4–6	3.5–6.5
7–9	6.5–9.5
10–12	9.5–12.5
13–15	12.5–15.5

FIGURE 2-4**FIGURE 2-5 Completion of Figure 2-4**

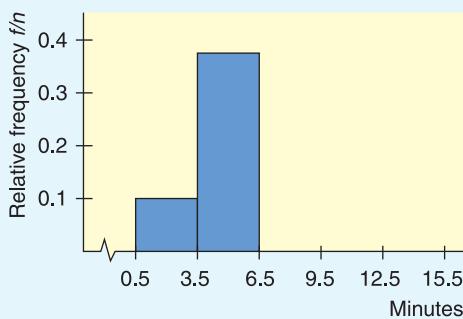
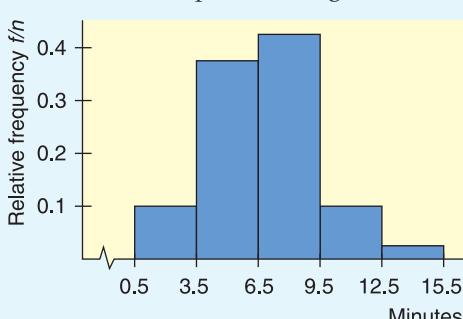
- (d) Compute the relative class frequency f/n for each class in Table 2-9 and complete the partial relative-frequency histogram in Figure 2-6.

TABLE 2-9 Relative Class Frequency

Class	f/n
1–3	$3/40 = 0.075$
4–6	$15/40 = 0.375$
7–9	_____
10–12	_____
13–15	_____

TABLE 2-10 Completion of Table 2-9

Class	f/n
1–3	0.075
4–6	0.375
7–9	0.425
10–12	0.100
13–15	0.025

FIGURE 2-6**FIGURE 2-7 Completion of Figure 2-6**

LOOKING FORWARD

We will see relative-frequency distributions again when we study probability in Chapter 4. There we will see that if a random sample is large enough, then we can estimate the probability of an event by the relative frequency of the event. The relative-frequency distribution then can be interpreted as a *probability distribution*. Such distributions will form the basis of our work in inferential statistics.

Distribution Shapes

Histograms are valuable and useful tools. If the raw data came from a random sample of population values, the histogram constructed from the sample values should have a distribution shape that is reasonably similar to that of the population.

Several terms are commonly used to describe histograms and their associated population distributions.

Mound-shaped symmetrical distribution

Uniform distribution

Skewed left
Skewed right

Bimodal distribution

(a) **Mound-shaped symmetrical:** This term refers to a histogram in which both sides are (more or less) the same when the graph is folded vertically down the middle. Figure 2-8(a) shows a typical mound-shaped symmetrical histogram.

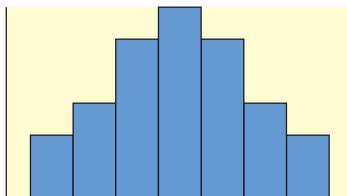
(b) **Uniform or rectangular:** These terms refer to a histogram in which every class has equal frequency. From one point of view, a uniform distribution is symmetrical with the added property that the bars are of the same height. Figure 2-8(b) illustrates a typical histogram with a uniform shape.

(c) **Skewed left or skewed right:** These terms refer to a histogram in which one tail is stretched out longer than the other. The direction of skewness is on the side of the *longer* tail. So, if the longer tail is on the left, we say the histogram is skewed to the left. Figure 2-8(c) shows a typical histogram skewed to the left and another skewed to the right.

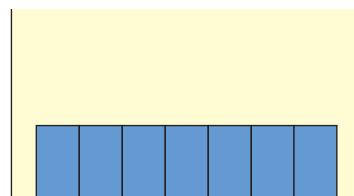
(d) **Bimodal:** This term refers to a histogram in which the two classes with the largest frequencies are separated by at least one class. The top two frequencies of these classes may have slightly different values. This type of situation sometimes indicates that we are sampling from two different populations. Figure 2-8(d) illustrates a typical histogram with a bimodal shape.

FIGURE 2-8
Types of Histograms

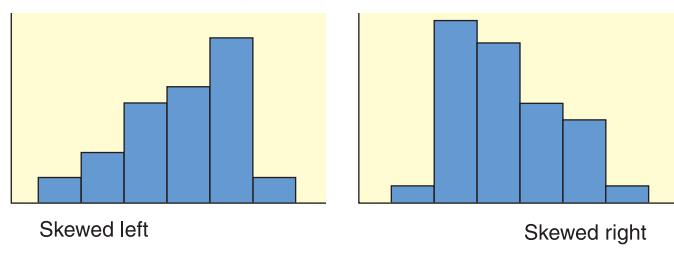
(a) Typical mound-shaped symmetrical histogram



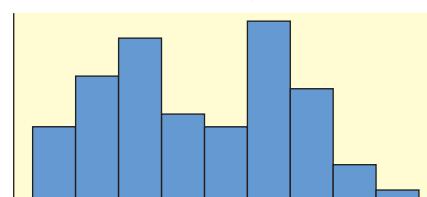
(b) Typical uniform or rectangular histogram



(c) Typical skewed histogram



(d) Typical bimodal histogram



LOOKING FORWARD

Normal and Student's t distributions are two important probability distributions used extensively in inferential statistics (Chapters 6, 7, 8). Both of these distributions are mound-shaped symmetrical. Two other useful probability distributions are the Chi-square and F distributions (Chapter 10), which are skewed right.

**CRITICAL THINKING**

A bimodal distribution shape might indicate that the data are from two different populations. For instance, a histogram showing the heights of a random sample of adults is likely to be bimodal because two populations, male and female, were combined.

If there are gaps in the histogram between bars at either end of the graph, the data set might include *outliers*.

Outliers

Outliers in a data set are the data values that are very different from other measurements in the data set.

Outliers may indicate data-recording errors. Valid outliers may be so unusual that they should be examined separately from the rest of the data. For instance, in a study of salaries of employees at one company, the chief CEO's salary may be so high and unique for the company that it should be considered separately from the other salaries. Decisions about outliers that are not recording errors need to be made by people familiar with both the field and the purpose of the study.

**TECH NOTES**

The TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab all create histograms. However, each technology automatically selects the number of classes to use. In Using Technology at the end of this chapter, you will see instructions for specifying the number of classes yourself and for generating histograms such as those we create "by hand."



Cumulative-Frequency Tables and Ogives

Sometimes we want to study cumulative totals instead of frequencies. Cumulative frequencies tell us how many data values are smaller than an upper class boundary. Once we have a frequency table, it is a fairly straightforward matter to add a column of cumulative frequencies.

Cumulative frequency

The **cumulative frequency** for a class is the sum of the frequencies for *that class and all previous classes*.

Ogive

An *ogive* (pronounced "oh-jī've") is a graph that displays cumulative frequencies.

PROCEDURE**HOW TO MAKE AN OGIVE**

1. Make a frequency table showing class boundaries and cumulative frequencies.
2. For each class, make a dot over the *upper class boundary* at the height of the cumulative class frequency. The coordinates of the dots are (upper class boundary, cumulative class frequency). Connect these dots with line segments.
3. By convention, an ogive begins on the horizontal axis at the lower class boundary of the first class.

EXAMPLE 3**CUMULATIVE-FREQUENCY TABLE AND OGIVE**

Aspen, Colorado, is a world-famous ski area. If the daily high temperature is above 40°F, the surface of the snow tends to melt. It then freezes again at night. This can result in a snow crust that is icy. It also can increase avalanche danger.

Table 2-11 gives a summary of daily high temperatures (°F) in Aspen during the 151-day ski season.

TABLE 2-11 High Temperatures During the Aspen Ski Season (°F)

Class Boundaries		Frequency	Cumulative Frequency
Lower	Upper		
10.5	20.5	23	23
20.5	30.5	43	66 (sum 23 + 43)
30.5	40.5	51	117 (sum 66 + 51)
40.5	50.5	27	144 (sum 117 + 27)
50.5	60.5	7	151 (sum 144 + 7)

Ph. Royer/Photo Researchers Inc.



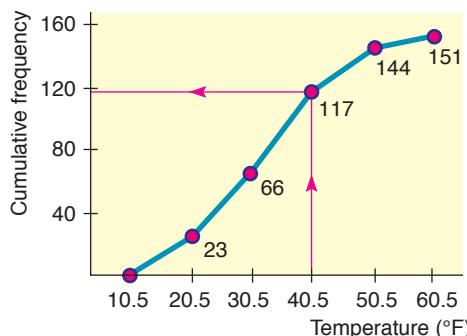
(a) The cumulative frequency for a class is computed by adding the frequency of that class to the frequencies of previous classes. Table 2-11 shows the cumulative frequencies.

(b) To draw the corresponding ogive, we place a dot at cumulative frequency 0 on the lower class boundary of the first class. Then we place dots over the *upper class boundaries* at the height of the cumulative class frequency for the corresponding class. Finally, we connect the dots. Figure 2-9 shows the corresponding ogive.

(c) Looking at the ogive, estimate the total number of days with a high temperature lower than or equal to 40°F.

FIGURE 2-9

Ogive for Daily High Temperatures (°F) During Aspen Ski Season



SOLUTION: Following the red lines on the ogive in Figure 2-9, we see that 117 days have had high temperatures of no more than 40°F.



VIEWPOINT

Mush, You Huskies!

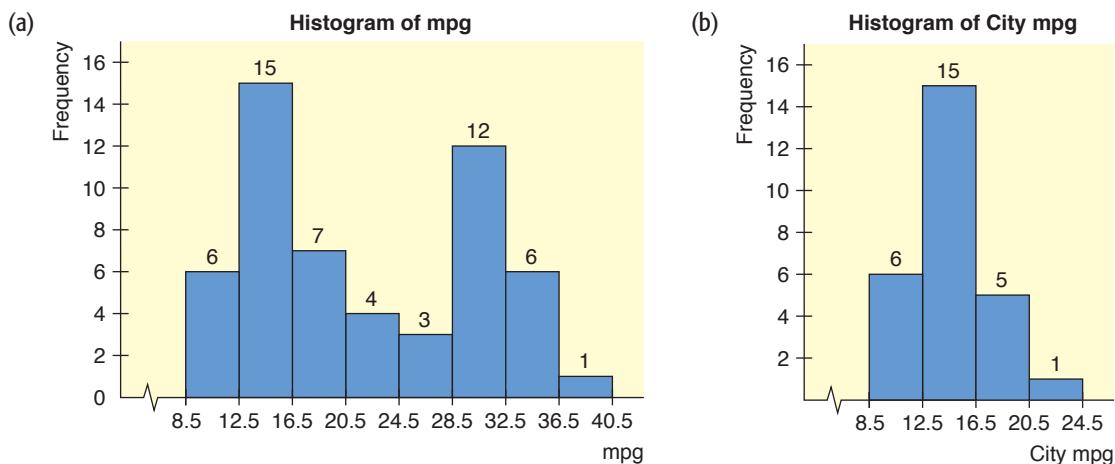
In 1925, the village of Nome, Alaska, had a terrible diphtheria epidemic. Serum was available in Anchorage but had to be brought to Nome by dogsled over the 1161-mile Iditarod Trail. Since 1973, the Iditarod Dog Sled Race from Anchorage to Nome has been an annual sporting event, with a current purse of more than \$600,000. Winning times range from more than 20 days to a little over 9 days.

To collect data on winning times, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Iditarod. Make a frequency distribution for these times.

SECTION 2.1 PROBLEMS

1. **Statistical Literacy** What is the difference between a class boundary and a class limit?
2. **Statistical Literacy** A data set has values ranging from a low of 10 to a high of 52. What's wrong with using the class limits 10–19, 20–29, 30–39, 40–49 for a frequency table?
3. **Statistical Literacy** A data set has values ranging from a low of 10 to a high of 50. What's wrong with using the class limits 10–20, 20–30, 30–40, 40–50 for a frequency table?
4. **Statistical Literacy** A data set has values ranging from a low of 10 to a high of 50. The class width is to be 10. What's wrong with using the class limits 10–20, 21–31, 32–42, 43–53 for a frequency table with a class width of 10?
5. **Basic Computation: Class Limits** A data set with whole numbers has a low value of 20 and a high value of 82. Find the class width and class limits for a frequency table with 7 classes.
6. **Basic Computation: Class Limits** A data set with whole numbers has a low value of 10 and a high value of 120. Find the class width and class limits for a frequency table with 5 classes.
7. **Interpretation** You are manager of a specialty coffee shop and collect data throughout a full day regarding waiting time for customers from the time they enter the shop until the time they pick up their order.
 - (a) What type of distribution do you think would be most desirable for the waiting times: skewed right, skewed left, mound-shaped symmetrical? Explain.
 - (b) What if the distribution for waiting times were bimodal? What might be some explanations?
8. **Critical Thinking** A web site rated 100 colleges and ranked the colleges from 1 to 100, with a rank of 1 being the best. Each college was ranked, and there were no ties. If the ranks were displayed in a histogram, what would be the shape of the histogram: skewed, uniform, mound-shaped?
9. **Critical Thinking** Look at the histogram in Figure 2–10(a), which shows mileage, in miles per gallon (mpg), for a random selection of passenger cars (Reference: *Consumer Reports*).
 - (a) Is the shape of the histogram essentially bimodal?

FIGURE 2-10



- (b) Jose looked at the raw data and discovered that the 54 data values included both the city and the highway mileages for 27 cars. He used the city mileages for the 27 cars to make the histogram in Figure 2–10(b). Using this information and Figure 2–10, parts (a) and (b), construct a histogram for the highway mileages of the same cars. Use class boundaries 16.5, 20.5, 24.5, 28.5, 32.5, 36.5, and 40.5.

10. **Critical Thinking** The following data represent salaries, in thousands of dollars, for employees of a small company. Notice that the data have been sorted in increasing order.

10. **Critical Thinking** The following data represent salaries, in thousands of dollars, for employees of a small company. Notice that the data have been sorted in increasing order.

24	25	25	27	27	29	30	35	35	35	36	38	38
39	39	40	40	40	45	45	45	45	47	52	52	52
58	59	59	61	61	67	68	68	68	250			

- (a) Make a histogram using the class boundaries 23.5, 69.5, 115.5, 161.5, 207.5, 253.5.
 - (b) Look at the last data value. Does it appear to be an outlier? Could this be the owner's salary?
 - (c) Eliminate the high salary of 250 thousand dollars. Make a new histogram using the class boundaries 23.5, 32.5, 41.5, 50.5, 59.5, 68.5. Does this histogram reflect the salary distribution of most of the employees better than the histogram in part (a)?

For Problems 11–16, use the specified number of classes to do the following.

- (a) Find the class width.
 - (b) Make a frequency table showing class limits, class boundaries, midpoints, frequencies, relative frequencies, and cumulative frequencies.
 - (c) Draw a histogram.
 - (d) Draw a relative-frequency histogram.
 - (e) Categorize the basic distribution shape as uniform, mound-shaped symmetrical, bimodal, skewed left, or skewed right.
 - (f) Draw an ogive.

11. **Sports: Dog Sled Racing** How long does it take to finish the 1161-mile Iditarod Dog Sled Race from Anchorage to Nome, Alaska (see Viewpoint)? Finish times (to the nearest hour) for 57 dogsled teams are shown below.

261	271	236	244	279	296	284	299	288	288	247	256
338	360	341	333	261	266	287	296	313	311	307	307
299	303	277	283	304	305	288	290	288	289	297	299
332	330	309	328	307	328	285	291	295	298	306	315
310	318	318	320	333	321	323	324	327			



Benoit Rousseaum/istockphoto

Use five classes.



12.

Medical: Glucose Testing The following data represent glucose blood levels (mg/100 ml) after a 12-hour fast for a random sample of 70 women (Reference: *American Journal of Clinical Nutrition*, Vol. 19, pp. 345–351). Note: These data are also available for download at the Online Study Center.

45	66	83	71	76	64	59	59
76	82	80	81	85	77	82	90
87	72	79	69	83	71	87	69
81	76	96	83	67	94	101	94
89	94	73	99	93	85	83	80
78	80	85	83	84	74	81	70
65	89	70	80	84	77	65	46
80	70	75	45	101	71	109	73
73	80	72	81	63	74		

Use six classes.



13.

Medical: Tumor Recurrence Certain kinds of tumors tend to recur. The following data represent the lengths of time, in months, for a tumor to recur after chemotherapy (Reference: D. P. Byar, *Journal of Urology*, Vol. 10, pp. 556–561). Note: These data are also available for download at the Online Study Center.

19	18	17	1	21	22	54	46	25	49
50	1	59	39	43	39	5	9	38	18
14	45	54	59	46	50	29	12	19	36
38	40	43	41	10	50	41	25	19	39
27	20								

Use five classes.



14.

Archaeology: New Mexico The Wind Mountain excavation site in New Mexico is an important archaeological location of the ancient Native American Anasazi culture. The following data represent depths (in cm) below surface grade at which significant artifacts were discovered at this site (Reference: A. I. Woosley and A. J. McIntyre, *Mimbres Mogollon Archaeology*, University of New Mexico Press). Note: These data are also available for download at the Online Study Center.

85	45	75	60	90	90	115	30	55	58
78	120	80	65	65	140	65	50	30	125
75	137	80	120	15	45	70	65	50	45
95	70	70	28	40	125	105	75	80	70
90	68	73	75	55	70	95	65	200	75
15	90	46	33	100	65	60	55	85	50
10	68	99	145	45	75	45	95	85	65
65	52	82							

Use seven classes.



15. **Environment: Gasoline Consumption** The following data represent highway fuel consumption in miles per gallon (mpg) for a random sample of 55 models of passenger cars (Source: Environmental Protection Agency). Note: These data are also available for download at the Online Study Center.

30	27	22	25	24	25	24	15
35	35	33	52	49	10	27	18
20	23	24	25	30	24	24	24
18	20	25	27	24	32	29	27
24	27	26	25	24	28	33	30
13	13	21	28	37	35	32	33
29	31	28	28	25	29	31	

Use five classes.

16. **Advertising: Readability** “Readability Levels of Magazine Ads,” by F. K. Shuptrine and D. D. McVicker, is an article in the *Journal of Advertising Research*. (For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to DASL, the Carnegie Mellon University Data and Story Library. Look in Data Subjects under Consumer and then Magazine Ads Readability file.) The following is a list of the number of three-syllable (or longer) words in advertising copy of randomly selected magazine advertisements.

34	21	37	31	10	24	39	10	17	18	32
17	3	10	6	5	6	6	13	22	25	3
5	2	9	3	0	4	29	26	5	5	24
15	3	8	16	9	10	3	12	10	10	10
11	12	13	1	9	43	13	14	32	24	15

Use eight classes.



17. **Expand Your Knowledge: Decimal Data** The following data represent tonnes of wheat harvested each year (1894–1925) from Plot 19 at the Rothamsted Agricultural Experiment Stations, England.

2.71	1.62	2.60	1.64	2.20	2.02	1.67	1.99	2.34	1.26	1.31
1.80	2.82	2.15	2.07	1.62	1.47	2.19	0.59	1.48	0.77	2.04
1.32	0.89	1.35	0.95	0.94	1.39	1.19	1.18	0.46	0.70	

- (a) Multiply each data value by 100 to “clear” the decimals.
- (b) Use the standard procedures of this section to make a frequency table and histogram with your whole-number data. Use six classes.
- (c) Divide class limits, class boundaries, and class midpoints by 100 to get back to your original data values.



18. **Decimal Data: Batting Averages** The following data represent baseball batting averages for a random sample of National League players near the end of the baseball season. The data are from the baseball statistics section of the *Denver Post*.

0.194	0.258	0.190	0.291	0.158	0.295	0.261	0.250	0.181
0.125	0.107	0.260	0.309	0.309	0.276	0.287	0.317	0.252
0.215	0.250	0.246	0.260	0.265	0.182	0.113	0.200	

- (a) Multiply each data value by 1000 to “clear” the decimals.
- (b) Use the standard procedures of this section to make a frequency table and histogram with your whole-number data. Use five classes.
- (c) Divide class limits, class boundaries, and class midpoints by 1000 to get back to your original data.



19. **Expand Your Knowledge: Dotplot** Another display technique that is somewhat similar to a histogram is a *dotplot*. In a dotplot, the data values are displayed along the horizontal axis. A dot is then plotted over each data value in the data set.

PROCEDURE

HOW TO MAKE A DOTPLOT

Display the data along a horizontal axis. Then plot each data value with a dot or point above the corresponding value on the horizontal axis. For repeated data values, stack the dots.

Dotplot

The next display shows a *dotplot* generated by Minitab (**Graph** **Dotplot**) for the number of licensed drivers per 1000 residents by state, including the District of Columbia (Source: U.S. Department of Transportation).

Dotplot for Licensed Drivers per 1000 Residents



- (a) From the dotplot, how many states have 600 or fewer licensed drivers per 1000 residents?
- (b) About what percentage of the states (out of 51) seem to have close to 800 licensed drivers per 1000 residents?
- (c) Consider the intervals 550 to 650, 650 to 750, and 750 to 850 licensed drivers per 1000 residents. In which interval do most of the states fall?
20. **Dotplot: Dog Sled Racing** Make a dotplot for the data in Problem 11 regarding the finish time (number of hours) for the Iditarod Dog Sled Race. Compare the dotplot to the histogram of Problem 11.
21. **Dotplot: Tumor Recurrence** Make a dotplot for the data in Problem 13 regarding the recurrence of tumors after chemotherapy. Compare the dotplot to the histogram of Problem 13.

SECTION 2.2

Bar Graphs, Circle Graphs, and Time-Series Graphs

FOCUS POINTS

- Determine types of graphs appropriate for specific data.
- Construct bar graphs, Pareto charts, circle graphs, and time-series graphs.
- Interpret information displayed in graphs.

Histograms provide a useful visual display of the distribution of data. However, the data must be quantitative. In this section, we examine other types of graphs, some of which are suitable for qualitative or category data as well.

Let's start with *bar graphs*. These are graphs that can be used to display quantitative or qualitative data.

Bar graph

Features of a Bar Graph

1. Bars can be vertical or horizontal.
2. Bars are of uniform width and uniformly spaced.
3. The lengths of the bars represent values of the variable being displayed, the frequency of occurrence, or the percentage of occurrence. The same measurement scale is used for the length of each bar.
4. The graph is well annotated with title, labels for each bar, and vertical scale or actual value for the length of each bar.

EXAMPLE 4**BAR GRAPH**

Figure 2-11 shows two bar graphs depicting the life expectancies for men and women born in the designated year. Let's analyze the features of these graphs.

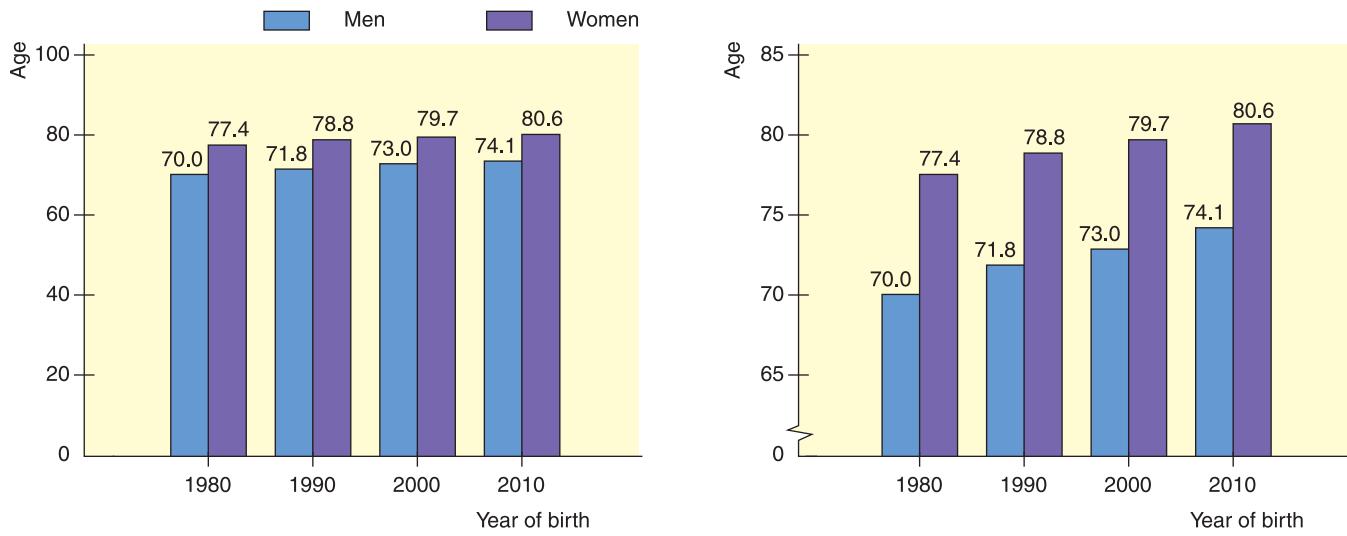
SOLUTION: The graphs are called *clustered bar graphs* because there are two bars for each year of birth. One bar represents the life expectancy for men, and the other represents the life expectancy for women. The height of each bar represents the life expectancy (in years).

Changing Scale

An important feature illustrated in Figure 2-11(b) is that of a *changing scale*. Notice that the scale between 0 and 65 is compressed. The changing scale amplifies the apparent difference between life spans for men and women, as well as the increase in life spans from those born in 1980 to the projected span of those born in 2010.

Changing Scale

Whenever you use a change in scale in a graphic, warn the viewer by using a squiggle \curvearrowleft on the changed axis. Sometimes, if a single bar is unusually long, the bar length is compressed with a squiggle in the bar itself.

FIGURE 2-11**Life Expectancy**

Source: U.S.Census Bureau

LOOKING FORWARD

Cluster bar graphs displaying percentages are particularly useful visuals for showing how different populations fit different categories. For instance, in Example 2 in Chapter 10, we see a cluster graph showing preferences of males and females for dogs, cats, other pets, or no pets. This graph displays the sample data for an inferential test called a *test of homogeneity*.

Pareto charts

Quality control is an important aspect of today's production and service industries. Dr. W. Edwards Deming was one of the developers of total quality management (TQM). In his book *Out of Crisis*, he outlines many strategies for monitoring and improving service and production industries. In particular, Dr. Deming recommends the use of some statistical methods to organize and analyze data from industries so that sources of problems can be identified and then corrected. *Pareto* (pronounced "Pah-rāy-tō") charts are among the many techniques used in quality-control programs.

A Pareto chart is a bar graph in which the bar height represents frequency of an event. In addition, the bars are arranged from left to right according to decreasing height.

GUIDED EXERCISE 2

Pareto charts

This exercise is adapted from *The Deming Management Method* by Mary Walton. Suppose you want to arrive at college 15 minutes before your first class so that you can feel relaxed when you walk into class. An early arrival time also allows room for unexpected delays. However, you always find yourself arriving "just in time" or slightly late. What causes you to be late? Charlotte made a list of possible causes and then kept a checklist for 2 months (Table 2-12). On some days more than one item was checked because several events occurred that caused her to be late.

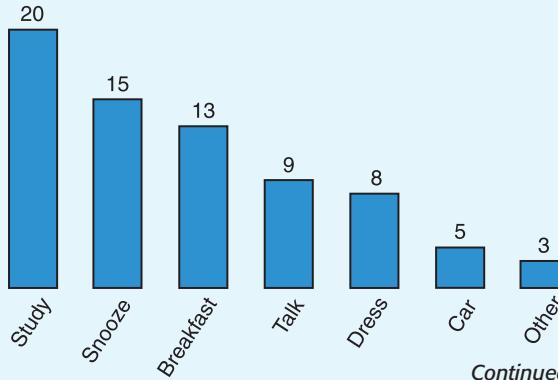
TABLE 2-12 Causes for Lateness (September–October)

Cause	Frequency
Snoozing after alarm goes off	15
Car trouble	5
Too long over breakfast	13
Last-minute studying	20
Finding something to wear	8
Talking too long with roommate	9
Other	3

- (a) Make a Pareto chart showing the causes for lateness. Be sure to label the causes, and draw the bars using the same vertical scale.



Figure 2-12 Pareto Chart: Conditions That Might Cause Lateness



Continued

GUIDED EXERCISE 2 *continued*

- (b) Looking at the Pareto chart, what recommendations do you have for Charlotte?



According to the chart, rearranging study time, or getting up earlier to allow for studying, would cure her most frequent cause for lateness. Repairing the car might be important, but for getting to campus early, it would not be as effective as adjusting study time.

Circle graphs or pie charts

Another popular pictorial representation of data is the *circle graph* or *pie chart*. It is relatively safe from misinterpretation and is especially useful for showing the division of a total quantity into its component parts. The total quantity, or 100%, is represented by the entire circle. Each wedge of the circle represents a component part of the total. These proportional segments are usually labeled with corresponding percentages of the total. Guided Exercise 3 shows how to make a circle graph.

In a circle graph or pie chart, wedges of a circle visually display proportional parts of the total population that share a common characteristic.

GUIDED EXERCISE 3

Circle graph

How long do we spend talking on the telephone after hours (at home after 5 P.M.)? The results from a recent survey of 500 people (as reported in *USA Today*) are shown in Table 2-13. We'll make a circle graph to display these data.

TABLE 2-13 Time Spent on Home Telephone After 5 P.M.

Time	Number	Fractional Part	Percentage	Number of Degrees
Less than $\frac{1}{2}$ hour	296	296/500	59.2	$59.2\% \times 360^\circ \approx 213^\circ$
$\frac{1}{2}$ hour to 1 hour	83	83/500	16.6	$16.6\% \times 360^\circ \approx 60^\circ$
More than 1 hour	121	—	—	—
Total	—	—	—	—

- (a) Fill in the missing parts in Table 2-13 for “More than 1 hour.” Remember that the central angle of a circle is 360° . Round to the nearest degree.



For “More than 1 hour,” Fractional Part = $121/500$; Percentage = 24.2%; Number of Degrees = $24.2\% \times 360^\circ \approx 87^\circ$. The symbol \approx means “approximately equal.”

- (b) Fill in the totals. What is the total number of responses? Do the percentages total 100% (within rounding error)? Do the numbers of degrees total 360° (within rounding error)?

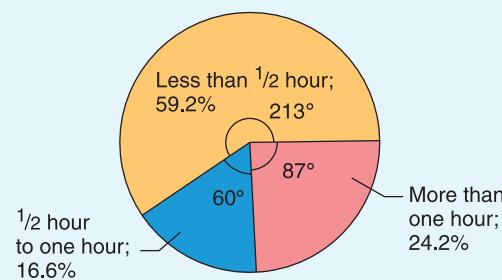


The total number of responses is 500. The percentages total 100%. You must have such a total in order to create a circle graph. The numbers of degrees total 360° .

- (c) Draw a circle graph. Divide the circle into pieces with the designated numbers of degrees. Label each piece, and show the percentage corresponding to each piece. The numbers of degrees are usually omitted from pie charts shown in newspapers, magazines, journals, and reports.



Figure 2-13 Hours on Home Telephone After 5 P.M.



Suppose you begin an exercise program that involves walking or jogging for 30 minutes. You exercise several times a week but monitor yourself by logging the distance you cover in 30 minutes each Saturday. How do you display these data in a meaningful way? Making a bar chart showing the frequency of distances you cover might be interesting, but it does not really show how the distance you cover in 30 minutes has changed over time. A graph showing the distance covered on each date will let you track your performance over time.

Time-series graph

We will use a *time-series graph*. A time-series graph is a graph showing data measurements in chronological order. To make a time-series graph, we put time on the horizontal scale and the variable being measured on the vertical scale. In a basic time-series graph, we connect the data points by line segments.

In a **time-series graph**, data are plotted in order of occurrence at regular intervals over a period of time.

EXAMPLE 5

TIME-SERIES GRAPH

Suppose you have been in the walking/jogging exercise program for 20 weeks, and for each week you have recorded the distance you covered in 30 minutes. Your data log is shown in Table 2-14.



TABLE 2-14 Distance (in Miles) Walked/Jogged in 30 Minutes

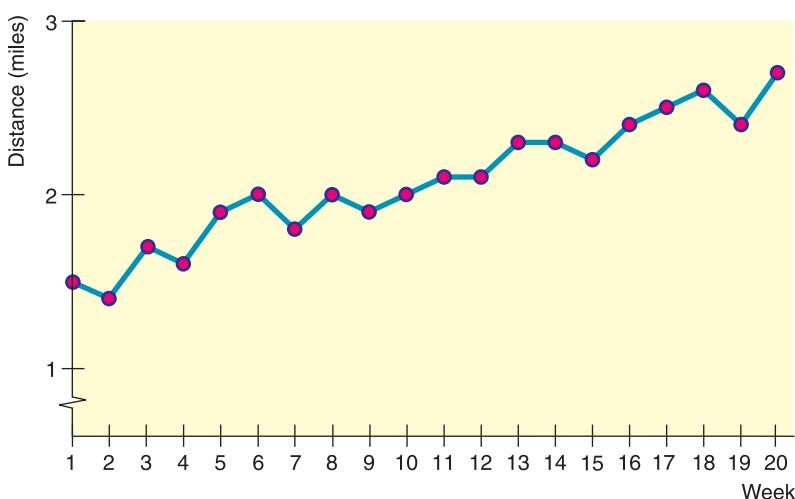
Week	1	2	3	4	5	6	7	8	9	10
Distance	1.5	1.4	1.7	1.6	1.9	2.0	1.8	2.0	1.9	2.0
Week	11	12	13	14	15	16	17	18	19	20
Distance	2.1	2.1	2.3	2.3	2.2	2.4	2.5	2.6	2.4	2.7

(a) Make a time-series graph.

SOLUTION: The data are appropriate for a time-series graph because they represent the same measurement (distance covered in a 30-minute period) taken at different times. The measurements are also recorded at equal time intervals (every week). To make our time-series graph, we list the weeks in order on the horizontal scale. Above each week, plot the distance covered that week on the vertical scale. Then connect the dots. Figure 2-14 shows the time-series graph. Be sure the scales are labeled.

FIGURE 2-14

Time-Series Graph of Distance (in miles) Jogged in 30 Minutes



(b) From looking at Figure 2-14, can you detect any patterns?

SOLUTION: There seems to be an upward trend in distance covered. The distances covered in the last few weeks are about a mile farther than those for the first few weeks. However, we cannot conclude that this trend will continue. Perhaps you have reached your goal for this training activity and now wish to maintain a distance of about 2.5 miles in 30 minutes.

Time series

Data sets composed of similar measurements taken at regular intervals over time are called *time series*. Time series are often used in economics, finance, sociology, medicine, and any other situation in which we want to study or monitor a similar measure over a period of time. A time-series graph can reveal some of the main features of a time series.

Time-series data consist of measurements of the same variable for the same subject taken at regular intervals over a period of time.



We've seen several styles of graphs. Which kinds are suitable for a specific data collection?

PROCEDURE

HOW TO DECIDE WHICH TYPE OF GRAPH TO USE

Bar graphs are useful for quantitative or qualitative data. With qualitative data, the frequency or percentage of occurrence can be displayed. With quantitative data, the measurement itself can be displayed, as was done in the bar graph showing life expectancy. Watch that the measurement scale is consistent or that a jump scale squiggle is used.

Pareto charts identify the frequency of events or categories in decreasing order of frequency of occurrence.

Circle graphs display how a *total* is dispersed into several categories.

The circle graph is very appropriate for qualitative data, or any data for which percentage of occurrence makes sense. Circle graphs are most effective when the number of categories or wedges is 10 or fewer.

Time-series graphs display how data change over time. It is best if the units of time are consistent in a given graph. For instance, measurements taken every day should not be mixed on the same graph with data taken every week.

For any graph: Provide a title, label the axes, and identify units of measure.

As Edward Tufte suggests in his book *The Visual Display of Quantitative Information*, don't let artwork or skewed perspective cloud the clarity of the information displayed.

TECH NOTES

Bar graphs, circle graphs, and time-series graphs

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Only time-series. Place consecutive values 1 through the number of time segments in list L1 and corresponding data in L2. Press Stat Plot and highlight an *xy* line plot.

Excel 2007 First enter the data into the spreadsheet. Then click the Insert Tab and select the type of chart you want to create. A variety of bar graphs, pie charts, and line graphs that can be used as time-series graphs are available. Use the Design Tab and the Layout Tab on the Chart Tools menu to access options such as title, axis labels, etc., for your chart.

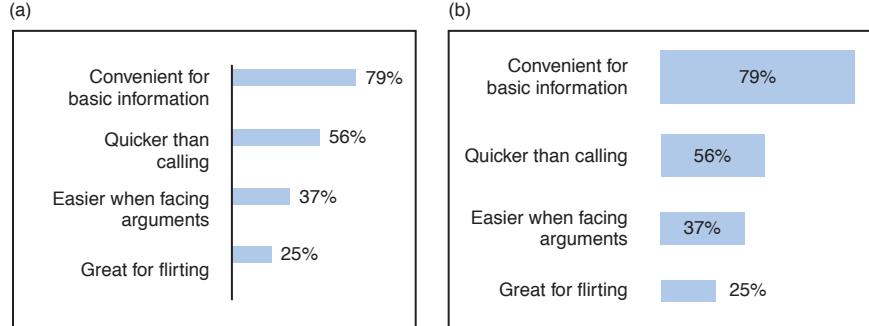
Minitab Use the menu selection **Graph**. Select the desired option and follow the instructions in the dialogue boxes.

VIEWPOINT**Do Ethical Standards Vary by the Situation?**

The Lutheran Brotherhood conducted a national survey and found that nearly 60% of all U.S. adults claim that ethics vary by the situation; 33% claim that there is only one ethical standard; and 7% are not sure. How could you draw a circle graph to make a visual impression of Americans' views on ethical standards?

SECTION 2.2 PROBLEMS

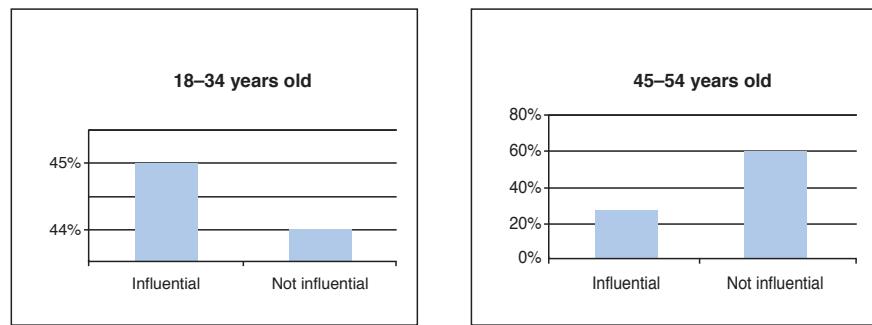
1. **Interpretation** Consider graph (a) of Reasons People Like Texting on Cell Phones, based on a GfK Roper survey of 1000 adults.

Reasons People Like Texting on Cell Phones

- (a) Do you think respondents could select more than one response? Explain.
 (b) Could the same information be displayed in a circle graph? Explain.
 (c) Is graph (a) a Pareto chart?

2. **Interpretation** Look at graph (b) of Reasons People Like Texting on Cell Phones. Is this a proper bar graph? Explain.
3. **Critical Thinking** A personnel office is gathering data regarding working conditions. Employees are given a list of five conditions that they might want to see improved. They are asked to select the one item that is most critical to them. Which type of graph, circle graph or Pareto chart, would be most useful for displaying the results of the survey? Why?
4. **Critical Thinking** Your friend is thinking about buying shares of stock in a company. You have been tracking the closing prices of the stock shares for the past 90 trading days. Which type of graph for the data, histogram or time-series, would be best to show your friend? Why?

5. ***Education: Does College Pay Off?*** It is costly in both time and money to go to college. Does it pay off? According to the Bureau of the Census, the answer is yes. The average annual income (in thousands of dollars) of a *household* headed by a person with the stated education level is as follows: 16.1 if ninth grade is the highest level achieved, 34.3 for high school graduates, 48.6 for those holding associate degrees, 62.1 for those with bachelor's degrees, 71.0 for those with master's degrees, and 84.1 for those with doctoral degrees. Make a bar graph showing household income for each education level.
6. ***Interpretation*** Consider the two graphs depicting the influence of advertisements on making large purchases for two different age groups, those 18–34 years old and those 45–54 years old (based on a Harris Poll of about 2500 adults aged 18 or older). *Note:* Other responses such as “not sure” and “not applicable” were also possible.

Influence of Advertising on Most Recent Large Purchase

- (a) Taking a quick glance at the graphs, Jenna thought that there was very little difference (maybe less than 1%) in the percentage of the two age groups who said that ads were influential. How would you change the graphs so that Jenna would not be misled so easily? *Hint:* Look at the vertical scales of the two graphs.
- (b) Take the information from the two graphs and make a cluster bar graph showing the percentage by age group reporting to be influenced by ads and those reporting they were not influenced by ads.
7. ***Commercial Fishing: Gulf of Alaska*** It's not an easy life, but it's a good life! Suppose you decide to take the summer off and sign on as a deck hand for a commercial fishing boat in Alaska that specializes in deep-water fishing for groundfish. What kind of fish can you expect to catch? One way to answer this question is to examine government reports on groundfish caught in the Gulf of Alaska. The following list indicates the types of fish caught annually in thousands of metric tons (Source: *Report on the Status of U.S. Living Marine Resources*, National Oceanic and Atmospheric Administration): flatfish, 36.3; Pacific cod, 68.6; sablefish, 16.0; Walleye pollock, 71.2; rockfish, 18.9. Make a Pareto chart showing the annual harvest for commercial fishing in the Gulf of Alaska.
8. ***Archaeology: Ireland*** Commercial dredging operations in ancient rivers occasionally uncover archaeological artifacts of great importance. One such artifact is Bronze Age spearheads recovered from ancient rivers in Ireland. A recent study gave the following information regarding discoveries of ancient bronze spearheads in Irish rivers.

River	Bann	Blackwater	Erne	Shannon	Barrow
No. of spearheads	19	8	15	33	14

(Based on information from *Crossing the Rubicon, Bronze Age Studies 5*, Lorraine Bourke, Department of Archaeology, National University of Ireland, Galway.)

- (a) Make a Pareto chart for these data.
 (b) Make a circle graph for these data.



9. **Lifestyle: Hide the Mess!** A survey of 1000 adults (reported in *USA Today*) uncovered some interesting housekeeping secrets. When unexpected company comes, where do we hide the mess? The survey showed that 68% of the respondents toss their mess into the closet, 23% shove things under the bed, 6% put things into the bathtub, and 3% put the mess into the freezer. Make a circle graph to display this information.
10. **Education: College Professors' Time** How do college professors spend their time? *The National Education Association Almanac of Higher Education* gives the following average distribution of professional time allocation: teaching, 51%; research, 16%; professional growth, 5%; community service, 11%; service to the college, 11%; and consulting outside the college, 6%. Make a pie chart showing the allocation of professional time for college professors.
11. **FBI Report: Hawaii** In the Aloha state, you are very unlikely to be murdered! However, it is considerably more likely that your house might be burgled, your car might be stolen, or you might be punched in the nose. That said, Hawaii is still a great place to vacation or, if you are very lucky, to live. The following numbers represent the crime rates per 100,000 population in Hawaii: murder, 2.6; rape, 33.4; robbery, 93.3; house burglary, 911.6; motor vehicle theft, 550.7; assault, 125.3 (Source: *Crime in the United States*, U.S. Department of Justice, Federal Bureau of Investigation).
- Display this information in a Pareto chart, showing the crime rate for each category.
 - Could the information as reported be displayed as a circle graph? Explain. Hint: Other forms of crime, such as arson, are not included in the information. In addition, some crimes might occur together.
12. **Driving: Bad Habits** Driving would be more pleasant if we didn't have to put up with the bad habits of other drivers. *USA Today* reported the results of a Valvoline Oil Company survey of 500 drivers, in which the drivers marked their complaints about other drivers. The top complaints turned out to be tailgating, marked by 22% of the respondents; not using turn signals, marked by 19%; being cut off, marked by 16%; other drivers driving too slowly, marked by 11%; and other drivers being inconsiderate, marked by 8%. Make a Pareto chart showing percentage of drivers listing each stated complaint. Could this information as reported be put in a circle graph? Why or why not?
13. **Ecology: Lakes** Pyramid Lake, Nevada, is described as the pride of the Paiute Indian Nation. It is a beautiful desert lake famous for very large trout. The elevation of the lake surface (feet above sea level) varies according to the annual flow of the Truckee River from Lake Tahoe. The U.S. Geological Survey provided the following data:

Year	Elevation	Year	Elevation	Year	Elevation
1986	3817	1992	3798	1998	3811
1987	3815	1993	3797	1999	3816
1988	3810	1994	3795	2000	3817
1989	3812	1995	3797		
1990	3808	1996	3802		
1991	3803	1997	3807		

Make a time-series graph displaying the data. For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Pyramid Lake Fisheries.

Ryan McVay/Photodisc/Jupiter Images



14. **Vital Statistics: Height** How does average height for boys change as boys get older? According to *Physician's Handbook*, the average heights at different ages are as follows:

Age (years)	0.5	1	2	3	4	5	6	7
Height (inches)	26	29	33	36	39	42	45	47
Age (years)	8	9	10	11	12	13	14	
Height (inches)	50	52	54	56	58	60	62	

Make a time-series graph for average height for ages 0.5 through 14 years.

SECTION 2.3

Stem-and-Leaf Displays

FOCUS POINTS

- Construct a stem-and-leaf display from raw data.
- Use a stem-and-leaf display to visualize data distribution.
- Compare a stem-and-leaf display to a histogram.

EDA

Exploratory Data Analysis

Together with histograms and other graphics techniques, the stem-and-leaf display is one of many useful ways of studying data in a field called *exploratory data analysis* (often abbreviated as *EDA*). John W. Tukey wrote one of the definitive books on the subject, *Exploratory Data Analysis* (Addison-Wesley). Another very useful reference for EDA techniques is the book *Applications, Basics, and Computing of Exploratory Data Analysis* by Paul F. Velleman and David C. Hoaglin (Duxbury Press). Exploratory data analysis techniques are particularly useful for detecting patterns and extreme data values. They are designed to help us explore a data set, to ask questions we had not thought of before, or to pursue leads in many directions.

EDA techniques are similar to those of an explorer. An explorer has a general idea of destination but is always alert for the unexpected. An explorer needs to assess situations quickly and often simplify and clarify them. An explorer makes pictures—that is, maps showing the relationships of landscape features. The aspects of rapid implementation, visual displays such as graphs and charts, data simplification, and robustness (that is, analysis that is not influenced much by extreme data values) are key ingredients of EDA techniques. In addition, these techniques are good for exploration because they require very few prior assumptions about the data.

EDA methods are especially useful when our data have been gathered for general interest and observation of subjects. For example, we may have data regarding the ages of applicants to graduate programs. We don't have a specific question in mind. We want to see what the data reveal. Are the ages fairly uniform or spread out? Are there exceptionally young or old applicants? If there are, we might look at other characteristics of these applicants, such as field of study. EDA methods help us quickly absorb some aspects of the data and then may lead us to ask specific questions to which we might apply methods of traditional statistics.

In contrast, when we design an experiment to produce data to answer a specific question, we focus on particular aspects of the data that are useful to us. If we want to determine the average highway gas mileage of a specific sports car, we use that model car in well-designed tests. We don't need to worry about unexpected

road conditions, poorly trained drivers, different fuel grades, sudden stops and starts, etc. Our experiment is designed to control outside factors. Consequently, we do not need to “explore” our data as much. We can often make valid assumptions about the data. Methods of traditional statistics will be very useful to analyze such data and answer our specific questions.

Stem-and-Leaf Display

In this text, we will introduce two EDA techniques: stem-and-leaf displays and, in Section 3.3, box-and-whisker plots. Let’s first look at a stem-and-leaf display.

Stem-and-leaf display

A stem-and-leaf display is a method of exploratory data analysis that is used to rank-order and arrange data into groups.

We know that frequency distributions and histograms provide a useful organization and summary of data. However, in a histogram, we lose most of the specific data values. A stem-and-leaf display is a device that organizes and groups data but allows us to recover the original data if desired. In the next example, we will make a stem-and-leaf display.



EXAMPLE 6

STEM-AND-LEAF DISPLAY

Many airline passengers seem weighted down by their carry-on luggage. Just how much weight are they carrying? The carry-on luggage weights in pounds for a random sample of 40 passengers returning from a vacation to Hawaii were recorded (see Table 2-15).

TABLE 2-15 **Weights of Carry-On Luggage in Pounds**

30	27	12	42	35	47	38	36	27	35
22	17	29	3	21	0	38	32	41	33
26	45	18	43	18	32	31	32	19	21
33	31	28	29	51	12	32	18	21	26

Stem and Leaf



Monika Wisniewska, 2009/Used under license from Shutterstock.com

To make a stem-and-leaf display, we break the digits of each data value into *two* parts. The left group of digits is called a *stem*, and the remaining group of digits on the right is called a *leaf*. We are free to choose the number of digits to be included in the stem.

The weights in our example consist of two-digit numbers. For a two-digit number, the stem selection is obviously the left digit. In our case, the tens digits will form the stems, and the units digits will form the leaves. For example, for the weight 12, the stem is 1 and the leaf is 2. For the weight 18, the stem is again 1, but the leaf is 8. In the stem-and-leaf display, we list each possible stem once on the left and all its leaves in the same row on the right, as in Figure 2-15(a). Finally, we order the leaves as shown in Figure 2-15(b).

Figure 2-15 shows a stem-and-leaf display for the weights of carry-on luggage. From the stem-and-leaf display in Figure 2-15, we see that two bags weighed 27 lb, one weighed 3 lb, one weighed 51 lb, and so on. We see that most of the weights were in the 30-lb range, only two were less than 10 lb, and six were over 40 lb. Note that the lengths of the lines containing the leaves give the visual impression that a sideways histogram would present.

As a final step, we need to indicate the scale. This is usually done by indicating the value represented by the stem and one leaf.

FIGURE 2-15

Stem-and-Leaf Displays of Airline Carry-On Luggage Weights

(a) Leaves Not Ordered

3 2	represents 32 lb
Stem	Leaves
0	3 0
1	2 7 8 8 9 2 8
2	7 7 2 9 1 6 1 8 9 1 6
3	0 5 8 6 5 8 2 3 2 1 2 3 1 2
4	2 7 1 5 3
5	1

(b) Final Display with Leaves Ordered

3 2	represents 32 lb
Stem	Leaves
0	0 3
1	2 2 7 8 8 8 9
2	1 1 1 2 6 6 7 7 8 9 9
3	0 1 1 2 2 2 2 3 3 5 5 6 8 8
4	1 2 3 5 7
5	1

There are no firm rules for selecting the group of digits for the stem. But whichever group you select, you must list all the possible stems from smallest to largest in the data collection.



PROCEDURE

HOW TO MAKE A STEM-AND-LEAF DISPLAY

1. Divide the digits of each data value into two parts. The leftmost part is called the *stem* and the rightmost part is called the *leaf*.
2. Align all the stems in a vertical column from smallest to largest. Draw a vertical line to the right of all the stems.
3. Place all the leaves with the same stem in the same row as the stem, and arrange the leaves in increasing order.
4. Use a label to indicate the magnitude of the numbers in the display. We include the decimal position in the label rather than with the stems or leaves.

LOOKING FORWARD

Ordering or sorting data is an essential first step to finding the median or center value of the data distribution. We will discuss the median in Section 3.1. The data also need to be ordered before we can construct a box-and-whisker plot (Section 3.3). After a little practice, you will find that stem-and-leaf diagrams provide a very efficient way of ordering data by hand.

GUIDED EXERCISE 4

Stem-and-leaf display

What does it take to win at sports? If you're talking about basketball, one sportswriter gave the answer. He listed the winning scores of the conference championship games over the last 35 years. The scores for those games follow below.

132	118	124	109	104	101	125	83	99
131	98	125	97	106	112	92	120	103
111	117	135	143	112	112	116	106	117
119	110	105	128	112	126	105	102	

Michael J. Doolittle/The Image Works



Continued

GUIDED EXERCISE 4 *continued*

To make a stem-and-leaf display, we'll use the first *two* digits as the stems (see Figure 2-16). Notice that the distribution of scores is fairly symmetrical.

- (a) Use the first *two* digits as the stem. Then order the leaves. Provide a label that shows the meaning and units of the first stem and first leaf.



Figure 2-16 Winning Scores

08 3	represents 083 or 83 points
08 3	
09 2 7 8 9	
10 1 2 3 4 5 5 6 6 9	
11 0 1 2 2 2 2 6 7 7 8 9	
12 0 4 5 5 6 8	
13 1 2 5	
14 3	

- (b) Looking at the distribution, would you say that it is fairly symmetrical?



Yes. Notice that stem 11 has the most data.

COMMENT Stem-and-leaf displays organize the data, let the data analyst spot extreme values, and are easy to create. In fact, they can be used to organize data so that frequency tables are easier to make. However, at this time, histograms are used more often in formal data presentations, whereas stem-and-leaf displays are used by data analysts to gain initial insights about the data.



TECH NOTES

Stem-and-leaf display

TI-84Plus/TI-83Plus/TI-nspire Does not support stem-and-leaf displays. You can sort the data by using Stat \blacktriangleright Edit \blacktriangleright 2:SortA.

Excel 2007 Enter your data and select the data you want to sort. On the Home ribbon, click the Sort and Filter button in the Editing group of the ribbon and select the desired sorting option.

Minitab Use the menu selections Graph \blacktriangleright Stem-and-Leaf and fill in the dialogue box.

Minitab Stem-and-Leaf Display (for Data in Guided Exercise 4)

Stem-and-Leaf of Scores		N=35
	Leaf Unit=1.0	
1		8 3
5		9 2789
14		10 123455669
(11)		11 01222267789
10		12 045568
4		13 125
1		14 3

The values shown in the left column represent depth. Numbers above the value in parentheses show the cumulative number of values from the top to the stem of the

middle value. Numbers below the value in parentheses show the cumulative number of values from the bottom to the stem of the middle value. The number in parentheses shows how many values are on the same line as the middle value.



CRITICAL THINKING

Problems 5 and 6 show how to split a stem.



Problem 10 discusses back-to-back stem-and-leaf displays.

Stem-and-leaf displays show each of the original or truncated data values. By looking at the display “sideways,” you can see the distribution shape of the data. If there are large gaps between stems containing leaves, especially at the top or bottom of the display, the data values in the first or last lines may be outliers. Outliers should be examined carefully to see if they are data errors or simply unusual data values. Someone very familiar with the field of study as well as the purpose of the study should decide how to treat outliers.

VIEWPOINT

What Does It Take to Win?

Scores for NFL Super Bowl games can be found at the NFL web site. Visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the NFL. Once at the NFL web site, follow the links to the Super Bowl. Of special interest in football statistics is the spread, or difference, between scores of the winning and losing teams. If the spread is too large, the game can appear to be lopsided, and TV viewers become less interested in the game (and accompanying commercial ads). Make a stem-and-leaf display of the spread for the NFL Super Bowl games and analyze the results.

SECTION 2.3 PROBLEMS

1. **Cowboys: Longevity** How long did *real* cowboys live? One answer may be found in the book *The Last Cowboys* by Connie Brooks (University of New Mexico Press). This delightful book presents a thoughtful sociological study of cowboys in west Texas and southeastern New Mexico around the year 1890. A sample of 32 cowboys gave the following years of longevity:

58	52	68	86	72	66	97	89	84	91	91
92	66	68	87	86	73	61	70	75	72	73
85	84	90	57	77	76	84	93	58	47	

(a) Make a stem-and-leaf display for these data.
(b) **Interpretation** Consider the following quote from Baron von Richthofen in his *Cattle Raising on the Plains of North America*: “Cowboys are to be found among the sons of the best families. The truth is probably that most were not a drunken, gambling lot, quick to draw and fire their pistols.” Does the data distribution of longevity lend credence to this quote?
2. **Ecology: Habitat** Wetlands offer a diversity of benefits. They provide a habitat for wildlife, spawning grounds for U.S. commercial fish, and renewable timber resources. In the last 200 years, the United States has lost more than half its wetlands. *Environmental Almanac* gives the percentage of wetlands lost in each

state in the last 200 years. For the lower 48 states, the percentage loss of wetlands per state is as follows:

46	37	36	42	81	20	73	59	35	50
87	52	24	27	38	56	39	74	56	31
27	91	46	9	54	52	30	33	28	35
35	23	90	72	85	42	59	50	49	
48	38	60	46	87	50	89	49	67	

Make a stem-and-leaf display of these data. Be sure to indicate the scale. How are the percentages distributed? Is the distribution skewed? Are there any gaps?

3. **Health Care: Hospitals** The American Medical Association Center for Health Policy Research, in its publication *State Health Care Data: Utilization, Spending, and Characteristics*, included data, by state, on the number of community hospitals and the average patient stay (in days). The data are shown in the table. Make a stem-and-leaf display of the data for the average length of stay in days. Comment about the general shape of the distribution.

State	No. of Hospitals	Average Length of Stay	State	No. of Hospitals	Average Length of Stay	State	No. of Hospitals	Average Length of Stay
Alabama	119	7.0	Kentucky	107	6.9	New Mexico	37	5.5
Alaska	16	5.7	Louisiana	136	6.7	Oregon	66	5.3
Arizona	61	5.5	Maine	38	7.2	Ohio	193	6.6
Arkansas	88	7.0	Maryland	51	6.8	Oklahoma	113	6.7
California	440	6.0	Massachusetts	101	7.0	Pennsylvania	236	7.5
Colorado	71	6.8	Michigan	175	7.3	Rhode Island	12	6.9
Connecticut	35	7.4	Minnesota	148	8.7	S. Carolina	68	7.1
Delaware	8	6.8	Mississippi	102	7.2	S. Dakota	52	10.3
Dist. of Columbia	11	7.5	Missouri	133	7.4	Tennessee	122	6.8
Florida	227	7.0	Montana	53	10.0	Texas	421	6.2
Georgia	162	7.2	N. Carolina	117	7.3	Utah	42	5.2
Hawaii	19	9.4	N. Dakota	47	11.1	Vermont	15	7.6
Idaho	41	7.1	Nebraska	90	9.6	Virginia	98	7.0
Indiana	113	6.6	Nevada	21	6.4	Washington	92	5.6
Illinois	209	7.3	New Hampshire	27	7.0	W. Virginia	59	7.1
Iowa	123	8.4	New York	231	9.9	Wisconsin	129	7.3
Kansas	133	7.8	New Jersey	96	7.6	Wyoming	27	8.5

4. **Health Care: Hospitals** Using the number of hospitals per state listed in the table in Problem 3, make a stem-and-leaf display for the number of community hospitals per state. Which states have an unusually high number of hospitals?



5. **Expand Your Knowledge: Split Stem** The Boston Marathon is the oldest and best-known U.S. marathon. It covers a route from Hopkinton, Massachusetts, to downtown Boston. The distance is approximately 26 miles. Visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Boston Marathon. Search the marathon site to find a wealth of information about the history of the race. In particular, the site gives the winning times for the Boston Marathon. They are all over 2 hours. The following data are the minutes over 2 hours for the winning male runners:

1961–1980										
23	23	18	19	16	17	15	22	13	10	
18	15	16	13	9	20	14	10	9	12	
1981–2000										
9	8	9	10	14	7	11	8	9	8	
11	8	9	7	9	9	10	7	9	9	

- (a) Make a stem-and-leaf display for the minutes over 2 hours of the winning times for the years 1961 to 1980. Use two lines per stem.

PROCEDURE

HOW TO SPLIT A STEM

When a stem has many leaves, it is useful to split the stem into two lines or more. For two lines per stem, place leaves 0 to 4 on the first line and leaves 5 to 9 on the next line.

- (b) Make a stem-and-leaf display for the minutes over 2 hours of the winning times for the years 1981 to 2000. Use two lines per stem.
 (c) *Interpretation* Compare the two distributions. How many times under 15 minutes are in each distribution?



6. **Split Stem: Golf** The U.S. Open Golf Tournament was played at Congressional Country Club, Bethesda, Maryland, with prizes ranging from \$465,000 for first place to \$5000. Par for the course was 70. The tournament consisted of four rounds played on different days. The scores for each round of the 32 players who placed in the money (more than \$17,000) were given on a web site. For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to golf. The scores for the first round were as follows:

71	65	67	73	74	73	71	71	74	73	71
70	75	71	72	71	75	75	71	71	74	75
66	75	75	75	71	72	72	73	71	67	

The scores for the fourth round for these players were as follows:

69	69	73	74	72	72	70	71	71	70	72
73	73	72	71	71	71	69	70	71	72	73
74	72	71	68	69	70	69	71	73	74	

- (a) Make a stem-and-leaf display for the first-round scores. Use two lines per stem. (See Problem 5.)
 (b) Make a stem-and-leaf display for the fourth-round scores. Use two lines per stem.
 (c) *Interpretation* Compare the two distributions. How do the highest scores compare? How do the lowest scores compare?

Are cigarettes bad for people? Cigarette smoking involves tar, carbon monoxide, and nicotine. The first two are definitely not good for a person's health, and the last ingredient can cause addiction. Problems 7, 8, and 9 refer to Table 2-16, which was taken from the web site maintained by the *Journal of Statistics Education*. For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the *Journal of Statistics Education*. Follow the links to the cigarette data.

TABLE 2-16 Milligrams of Tar, Nicotine, and Carbon Monoxide (CO) per One Cigarette

Brand	Tar	Nicotine	CO	Brand	Tar	Nicotine	CO
Alpine	14.1	0.86	13.6	MultiFilter	11.4	0.78	10.2
Benson & Hedges	16.0	1.06	16.6	Newport Lights	9.0	0.74	9.5
Bull Durham	29.8	2.03	23.5	Now	1.0	0.13	1.5
Camel Lights	8.0	0.67	10.2	Old Gold	17.0	1.26	18.5
Carlton	4.1	0.40	5.4	Pall Mall Lights	12.8	1.08	12.6
Chesterfield	15.0	1.04	15.0	Raleigh	15.8	0.96	17.5
Golden Lights	8.8	0.76	9.0	Salem Ultra	4.5	0.42	4.9
Kent	12.4	0.95	12.3	Tareyton	14.5	1.01	15.9
Kool	16.6	1.12	16.3	True	7.3	0.61	8.5
L&M	14.9	1.02	15.4	Viceroy Rich Light	8.6	0.69	10.6
Lark Lights	13.7	1.01	13.0	Virginia Slims	15.2	1.02	13.9
Marlboro	15.1	0.90	14.4	Winston Lights	12.0	0.82	14.9
Merit	7.8	0.57	10.0				

Source: Federal Trade Commission, USA (public domain).

7. **Health: Cigarette Smoke** Use the data in Table 2-16 to make a stem-and-leaf display for milligrams of tar per cigarette smoked. Are there any outliers?

8. **Health: Cigarette Smoke** Use the data in Table 2-16 to make a stem-and-leaf display for milligrams of carbon monoxide per cigarette smoked. Are there any outliers?

9. **Health: Cigarette Smoke** Use the data in Table 2-16 to make a stem-and-leaf display for milligrams of nicotine per cigarette smoked. In this case, truncate the measurements at the tenths position and use two lines per stem (see Problem 5, part a).

10. **Expand Your Knowledge: Back-to-Back Stem Plot** In archaeology, the depth (below surface grade) at which artifacts are found is very important. Greater depths sometimes indicate older artifacts, perhaps from a different archaeological period. Figure 2-17 is a *back-to-back stem plot* showing the depths of artifact locations at two different archaeological sites. These sites are from similar geographic locations. Notice that the stems are in the center of the diagram. The leaves for Site I artifact depths are shown to the left of the stem, while the leaves for Site II are to the right of the stem (see *Mimbres Mogollon Archaeology* by A. I. Woosley and A. J. McIntyre, University of New Mexico Press).

(a) What are the least and greatest depths of artifact finds at Site I? at Site II?

(b) Describe the data distribution of depths of artifact finds at Site I and at Site II.

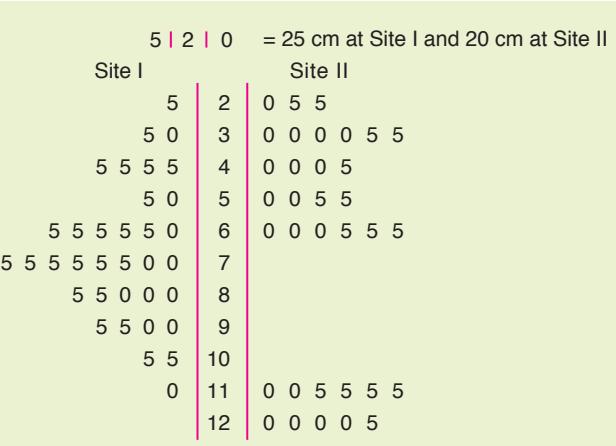
(c) **Interpretation** At Site II, there is a gap in the depths at which artifacts were found. Does the Site II data distribution suggest that there might have been a period of no occupation?



Back-to-back stem plot

FIGURE 2-17

Depth (in cm) of Artifact Location





SUMMARY

Chapter Review

Organizing and presenting data are the main purposes of the branch of statistics called descriptive statistics. Graphs provide an important way to show how the data are distributed.

- Frequency tables show how the data are distributed within set classes. The classes are chosen so that they cover all data values and so that each data value falls within only one class. The number of classes and the class width determine the class limits and class boundaries. The number of data values falling within a class is the class frequency.
- A histogram is a graphical display of the information in a frequency table. Classes are shown on the horizontal axis, with corresponding frequencies on the vertical axis. Relative-frequency histograms show relative

frequencies on the vertical axis. Ogives show cumulative frequencies on the vertical axis. Dotplots are like histograms, except that the classes are individual data values.

- Bar graphs, Pareto charts, and pie charts are useful to show how quantitative or qualitative data are distributed over chosen categories.
- Time-series graphs show how data change over set intervals of time.
- Stem-and-leaf displays are effective means of ordering data and showing important features of the distribution.

Graphs aren't just pretty pictures. They help reveal important properties of the data distribution, including the shape and whether or not there are any outliers.

IMPORTANT WORDS & SYMBOLS

Section 2.1

- Frequency table 40
 Frequency distribution 40
 Class width 41
 Class, lower limit, upper limit 41
 Class frequency 42
 Class midpoint 42
 Class mark 42
 Class boundaries 42
 Relative frequency 42
 Relative-frequency table 44
 Histogram 44
 Relative frequency histogram 44
 Mound-shaped symmetrical distribution 47
 Uniform distribution 47
 Skewed left 47
 Skewed right 47
 Bimodal distribution 47
 Outlier 48

- Cumulative frequency 48
 Ogive 48
 Dotplot 54

Section 2.2

- Bar graph 54
 Changing scale 55
 Pareto chart 56
 Circle graphs or pie charts 57
 Time-series graph 58
 Time series 59

Section 2.3

- EDA 63
 Stem-and-leaf display 64
 Stem 64
 Leaf 64
 Split stem 68
 Back-to-back stem plot 70

VIEWPOINT

Personality Clash!

Karl Pearson and Sir Ronald Fisher are two very famous mathematicians who contributed a great deal to the understanding and practice of modern statistics. Each was an outstanding person in his own right; however, the men had a terrific personality clash in the world of statistics. In Chapter 9 we will study the Pearson product moment correlation coefficient. Fisher's F distribution is the

central probability distribution in the field known as analysis of variance. In Chapter 8 we will examine two interpretations of statistical testing: the classical method of critical regions, favored by Karl Pearson, and the P-value method, favored by Sir Ronald Fisher. For a fixed level of significance, the methods can be shown to be equivalent. However, the P-value method is favored strongly by research and technology and is therefore emphasized in this text. The personality battles of Pearson and Fisher are chronicled in the popular book *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century* by David Salsburg.

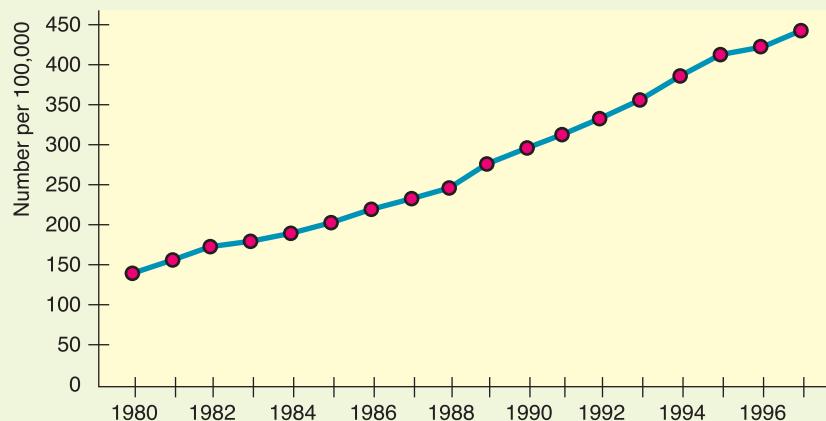
CHAPTER REVIEW PROBLEMS



1. **Critical Thinking** Consider these types of graphs: histogram, bar graph, Pareto chart, pie chart, stem-and-leaf display.
 - (a) Which are suitable for qualitative data?
 - (b) Which are suitable for quantitative data?
2. **Critical Thinking** A consumer interest group is tracking the percentage of household income spent on gasoline over the past 30 years. Which graphical display would be more useful, a histogram or a time-series graph? Why?
3. **Critical Thinking** Describe how data outliers might be revealed in histograms and stem-and-leaf plots.
4. **Expand Your Knowledge** How are dotplots and stem-and-leaf displays similar? How are they different?
5. **Focus Problem: Fuel Economy** Solve the focus problem at the beginning of this chapter.
6. **Criminal Justice: Prisoners** The time plot in Figure 2-18 gives the number of state and federal prisoners per 100,000 population (Source: *Statistical Abstract of the United States*, 120th Edition).
 - (a) Estimate the number of prisoners per 100,000 people for 1980 and for 1997.
 - (b) **Interpretation** During the time period shown, there was increased prosecution of drug offenses, longer sentences for common crimes, and reduced access to parole. What does the time-series graph say about the prison population change per 100,000 people?
 - (c) In 1997, the U.S. population was approximately 266,574,000 people. At the rate of 444 prisoners per 100,000 population, about how many prisoners were in the system? The projected U.S. population for the year 2020 is 323,724,000. If the rate of prisoners per 100,000 stays the same as in 1997, about how many prisoners do you expect will be in the system in 2020? To obtain the most recent information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Census Bureau.

FIGURE 2-18

Number of State and Federal Prisoners per 100,000 Population



7. **IRS: Tax Returns** Almost everyone files (or will someday file) a federal income tax return. A research poll for TurboTax (a computer software package to aid in tax-return preparation) asked what aspect of filing a return people thought to be the most difficult. The results showed that 43% of the respondents said understanding the IRS jargon, 28% said knowing deductions, 10% said getting the right form, 8% said calculating the numbers, and 10% didn't know. Make a circle graph to display this information. *Note:* Percentages will not total 100% because of rounding.
8. **Law Enforcement: DUI** Driving under the influence of alcohol (DUI) is a serious offense. The following data give the ages of a random sample of 50 drivers arrested while driving under the influence of alcohol. This distribution is based on the age distribution of DUI arrests given in the *Statistical Abstract of the United States* (112th Edition).

46	16	41	26	22	33	30	22	36	34
63	21	26	18	27	24	31	38	26	55
31	47	27	43	35	22	64	40	58	20
49	37	53	25	29	32	23	49	39	40
24	56	30	51	21	45	27	34	47	35

- (a) Make a stem-and-leaf display of the age distribution.
- (b) Make a frequency table using seven classes.
- (c) Make a histogram showing class boundaries.
- (d) Identify the shape of the distribution.

9. **Agriculture: Apple Trees** The following data represent trunk circumferences (in mm) for a random sample of 60 four-year-old apple trees at East Malling Agriculture Research Station in England (Reference: S. C. Pearce, University of Kent at Canterbury). *Note:* These data are also available for download at the Online Study Center.

108	99	106	102	115	120	120	117	122	142
106	111	119	109	125	108	116	105	117	123
103	114	101	99	112	120	108	91	115	109
114	105	99	122	106	113	114	75	96	124
91	102	108	110	83	90	69	117	84	142
122	113	105	112	117	122	129	100	138	117

- (a) Make a frequency table with seven classes showing class limits, class boundaries, midpoints, frequencies, and relative frequencies.
- (b) Draw a histogram.
- (c) Draw a relative-frequency histogram.
- (d) Identify the shape of the distribution.
- (e) Draw an ogive.

10. **Law: Corporation Lawsuits** Many people say the civil justice system is overburdened. Many cases center on suits involving businesses. The following data are based on a *Wall Street Journal* report. Researchers conducted a study of lawsuits involving 1908 businesses ranked in the Fortune 1000 over a 20-year period.



They found the following distribution of civil justice caseloads brought before the federal courts involving the businesses:

Case Type	Number of Filings (in thousands)
Contracts	107
General torts (personal injury)	191
Asbestos liability	49
Other product liability	38
All other	21

Note: Contracts cases involve disputes over contracts between businesses.

- (a) Make a Pareto chart of the caseloads. Which type of cases occur most frequently?
 (b) Make a pie chart showing the percentage of cases of each type.

11. ***Archaeology: Tree-Ring Data*** The Sand Canyon Archaeological Project, edited by W. D. Lipe and published by Crow Canyon Archaeological Center, contains the stem-and-leaf diagram shown in Figure 2-19. The study uses tree rings to accurately determine the year in which a tree was cut. The figure gives the tree-ring-cutting dates for samples of timbers found in the architectural units at Sand Canyon Pueblo. The text referring to the figure says, “The three-digit numbers in the left column represent centuries and decades A.D. The numbers to the right represent individual years, with each number derived from an individual sample. Thus, 124|2 2 represents three samples dated to A.D. 1242.” Use Figure 2-19 and the verbal description to answer the following questions.

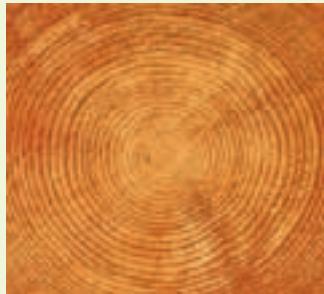
(a) Which decade contained the most samples?

(b) How many samples had a tree-ring-cutting date between 1200 A.D. and 1239 A.D., inclusive?

(c) What are the dates of the longest interval during which no tree-cutting samples occurred? What might this indicate about new construction or renovation of the pueblo structures during this period?

FIGURE 2-19

Tree-Ring-Cutting Dates from Architectural Units of Sand Canyon Pueblo: *The Sand Canyon Archaeological Project*

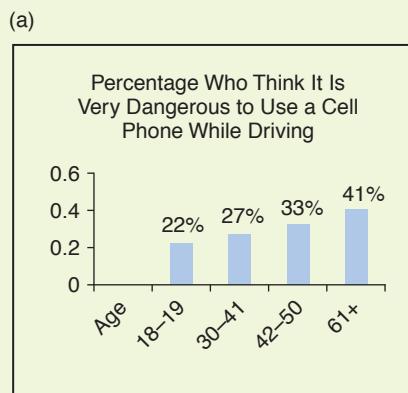


Yulia Podlesnova, 2009/Used under license from Shutterstock.com

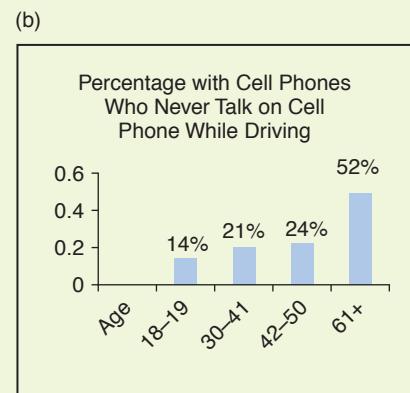
12. **Interpretation** A Harris Poll surveyed 2085 U.S. adults regarding use of cell phones while driving. All the adults were asked their opinion regarding how dangerous it is for a driver to use a cell phone while driving. Graph (a) shows the percentage responding “very dangerous” by age group. Only the adults who drive and who have a cell phone were asked how often they talk on the cell phone while driving. Graph (b) shows the percentage responding “never” by age group.

Cell Phone Use and Driving

(a)



(b)

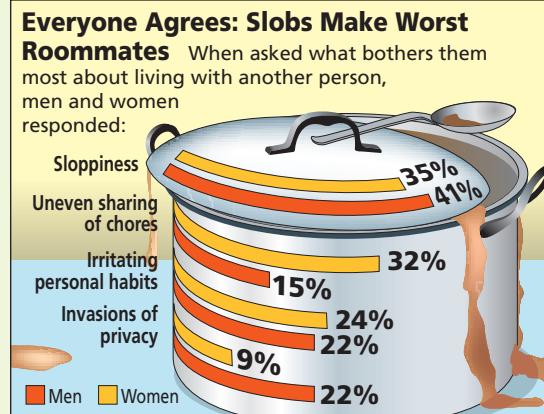


- (a) What trend does this survey portray regarding age group and the opinion that using a cell phone while driving is very dangerous?
- (b) How does the behavior of never using a cell phone while driving compare to the opinion that using a cell phone while driving is dangerous? Do you think that some of the differences in the behavior (never use a cell phone while driving) and the opinion (using a cell phone while driving is very dangerous) can be attributed to the differences in the survey population? Explain.

**DATA HIGHLIGHTS:
GROUP PROJECTS**

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

- Examine Figure 2-20, “Everyone Agrees: Slobs Make Worst Roommates.” This is a clustered bar graph because two percentages are given for each response category: responses from men and responses from women. Comment about how the artistic rendition has slightly changed the format of a bar graph. Do the bars seem to have lengths that accurately reflect the relative percentages of the responses? In your own opinion, does the artistic rendition enhance or confuse the information? Explain. Which characteristic of “worst roommates” does the graphic seem to illustrate? Can this graph be considered a Pareto chart for men? for women? Why or why not? From the information given in the figure, do you think the survey just listed the four given annoying characteristics? Do you think a respondent could choose more than one characteristic? Explain your answer in

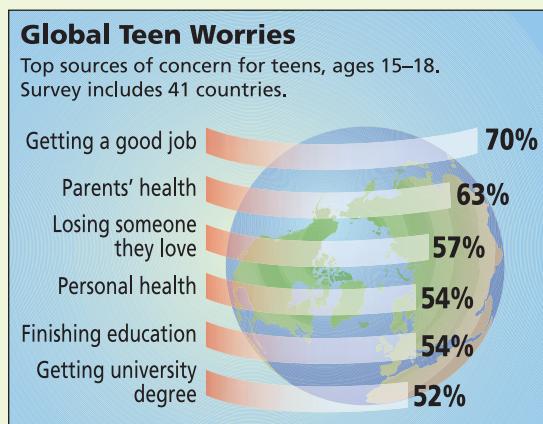
FIGURE 2-20

Source: Advantage Business Research for Mattel Compatibility

terms of the percentages given and in terms of the explanation given in the graphic. Could this information also be displayed in one circle graph for men and another for women? Explain.

2. Examine Figure 2-21, “Global Teen Worries.” How many countries were contained in the sample? The graph contains bars and a circle. Which bar is the longest? Which bar represents the greatest percentage? Is this a bar graph or not? If not, what changes would need to be made to put the information in a bar graph? Could the graph be made into a Pareto chart? Could it be made into a circle graph? Explain.

FIGURE 2-21



Source: BrainWaves Group's New World Teen Study

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. In your own words, explain the differences among histograms, relative-frequency histograms, bar graphs, circle graphs, time-series graphs, Pareto charts, and stem-and-leaf displays. If you have nominal data, which graphic displays might be useful? What if you have ordinal, interval, or ratio data?
2. What do we mean when we say a histogram is skewed to the left? to the right? What is a bimodal histogram? Discuss the following statement: “A bimodal histogram usually results if we draw a sample from two populations at once.” Suppose you took a sample of weights of college football players and with this sample you included weights of cheerleaders. Do you think a histogram made from the combined weights would be bimodal? Explain.
3. Discuss the statement that stem-and-leaf displays are quick and easy to construct. How can we use a stem-and-leaf display to make the construction of a frequency table easier? How does a stem-and-leaf display help you spot extreme values quickly?
4. Go to the library and pick up a current issue of *The Wall Street Journal*, *Newsweek*, *Time*, *USA Today*, or other news medium. Examine each newspaper or magazine for graphs of the types discussed in this chapter. List the variables used, method of data collection, and general types of conclusions drawn from the graphs. Another source for information is the Internet. Explore several web sites, and categorize the graphs you find as you did for the print media. For interesting web sites, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find links to the Social Statistics Briefing Room, to law enforcement, and to golf.

USING TECHNOLOGY

Applications

The following tables show the first-round winning scores of the NCAA men's and women's basketball teams.

TABLE 2-17 Men's Winning First-Round NCAA Tournament Scores

95	70	79	99	83	72	79	101
69	82	86	70	79	69	69	70
95	70	77	61	69	68	69	72
89	66	84	77	50	83	63	58

TABLE 2-18 Women's Winning First-Round NCAA Tournament Scores

80	68	51	80	83	75	77	100
96	68	89	80	67	84	76	70
98	81	79	89	98	83	72	100
101	83	66	76	77	84	71	77

1. Use the software or method of your choice to construct separate histograms for the men's and women's winning scores. Try 5, 7, and 10 classes for each. Which number of classes seems to be the best choice? Why?
2. Use the same class boundaries for histograms of men's and of women's scores. How do the scores for the two groups compare? What general shape do the histograms follow?
3. Use the software or method of your choice to make stem-and-leaf displays for each set of scores. If your software does not make stem-and-leaf displays, sort the data first and then make a back-to-back display by hand. Do there seem to be any extreme values in either set? How do the data sets compare?

Technology Hints: Creating Histograms

The default histograms produced by the TI-84Plus/TI-83Plus/TI-nspire calculators, Minitab, and Excel all determine automatically the number of classes to use. To control the number of classes the technology uses, follow the key steps as indicated. The display screens are generated for data found in Table 2-1—One-Way Commuting Distances (in Miles) for 60 Workers in Downtown Dallas.

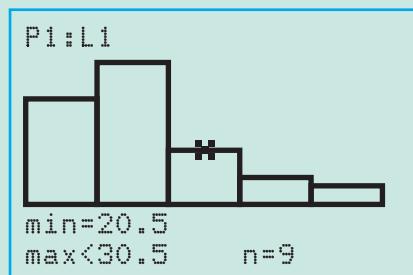
TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad)

Determine the class width for the number of classes you want and the lower class boundary for the first class. Enter the data in list L1.

Press STATPLOT and highlight On and the histogram plot.

Press WINDOW and set X min = lowest class boundary, Xscl = class width. Use appropriate values for the other settings.

Press GRAPH. TRACE gives boundaries and frequency.



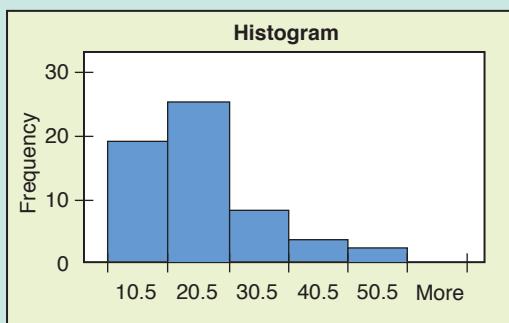
Excel 2007

Determine the upper class boundaries for the five classes. Enter the data in one column. In a separate column, enter the upper class boundaries. Click the Data tab, and then click the Data Analysis button on the ribbon in the Analysis group. Select Histogram and click OK.

In the Histogram dialogue box, put the data range in the Input Range. Put the upper class boundaries range in the Bin Range. Then check New Workbook and Chart Output and click OK.

To make the bars touch, right click on a bar and select Format Data Series... Then under Series Options move the Gap Width slider to 0% (No Gap) and click Close.

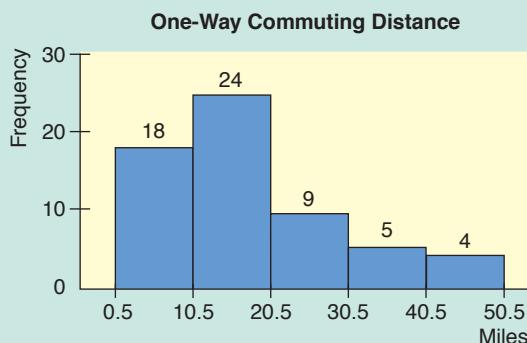
Further adjustments to the histogram can be made by clicking the Layout tab and selecting options in the labels or axes section of the ribbon.



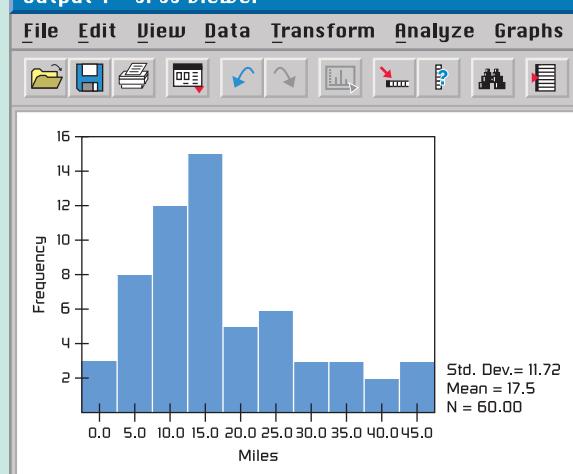
Minitab

Determine the class boundaries. Enter the data. Use the menu selection **Graph > Histogram**.

Choose **Simple** and click **OK**. Select the graph variable and click **OK** to obtain a histogram with automatically selected classes. To set your own class boundaries, double click the displayed histogram. In the dialogue box, select **Binning**. Then choose **Cutpoint** and enter the class boundaries as cutpoint positions.



Output 1 - SPSS Viewer



Note that SPSS requires that the measurement level of the data be set before using the **Chart Builder**. In SPSS, the **scale level** includes both interval and ratio levels of measurement. The **nominal** and **ordinal** levels are as described in this text.

Specific instructions for setting class boundaries (cutpoints) of a histogram are provided in the *Technology Guide* that accompanies this text.

SPSS

The SPSS screen shot shows the default histogram created by the menu choices **Analyze > Descriptive Statistics > Frequencies**. In the dialogue box, move the variable containing the data into the variables window. Click **Charts** and select **Histograms**. Click the Continue button and then the OK button.

In SPSS Student Version 17 there are a number of additional ways to create histograms and other graphs. These options utilize the **Chart Builder** or **Graphboard Template Chooser** items in the **Graphs** menu. Under these options you can set the number of classes and the boundaries of the histogram. Specific instructions for using these options are provided in the *Technology Guide* for SPSS that accompanies this text.



3

- 3.1 Measures of Central Tendency: Mode, Median, and Mean
- 3.2 Measures of Variation
- 3.3 Percentiles and Box-and-Whisker Plots



While the individual man is an insolvable puzzle, in the aggregate he becomes a mathematical certainty.

Imago/Getty

You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to.

—ARTHUR CONAN DOYLE,
The Sign of Four



PARAMOUNT TELEVISION/THE KOBAL COLLECTION/Picture Desk

Sherlock Holmes spoke the words quoted here to his colleague, Dr. Watson, as the two were unraveling a mystery. The detective was implying that if a single member is drawn at random from a population, we cannot predict exactly what that member will look like. However, there are some “average” features of the entire population that an individual is likely to possess. The degree of certainty with which we would expect to observe such average features in any individual depends on our knowledge of the variations among individuals in the population. Sherlock Holmes has led us to two of the most important statistical concepts: average and variation.

For online student resources, visit the Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>

AVERAGES AND VARIATION

PREVIEW QUESTIONS

What are commonly used measures of central tendency? What do they tell you? (SECTION 3.1)

How do variance and standard deviation measure data spread? Why is this important? (SECTION 3.2)

How do you make a box-and-whisker plot, and what does it tell about the spread of the data? (SECTION 3.3)



Nancy Sheehan/PhotoEdit

FOCUS PROBLEM

The Educational Advantage

Is it really worth all the effort to get a college degree? From a philosophical point of view, the love of learning is sufficient reason to get a college degree. However, the U.S. Census Bureau also makes another relevant point. Annually, college graduates (bachelor's degree) earn on average \$23,291 more than high school graduates. This means college graduates earn about 83.4% more than high school graduates, and according to "Education Pays" on the next page, the gap in earnings is increasing. Furthermore, as the College Board indicates, for most Americans college remains relatively affordable.

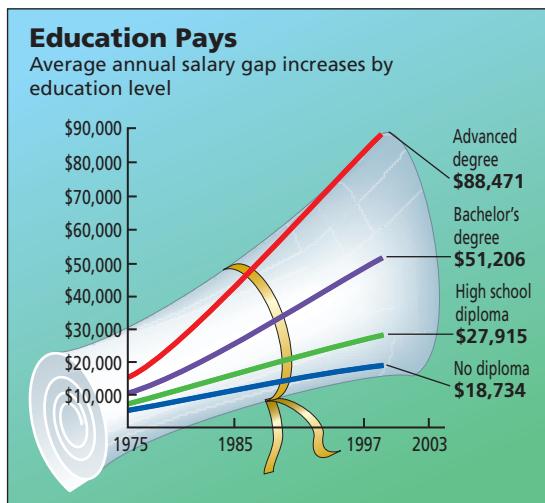
After completing this chapter, you will be able to answer the following questions.

- Does a college degree *guarantee* someone an 83.4% increase in earnings over a high school degree? Remember, we are using only *averages* from census data.
- Using census data (not shown in "Education Pays"), it is estimated that the standard deviation of college-graduate earnings is about \$8500. Compute a 75% Chebyshev confidence interval centered on the mean (\$51,206) for bachelor's degree earnings.
- How much does college tuition cost? That depends, of course, on where you go to college. Construct a weighted average. Using the data from "College Affordable for Most," estimate midpoints for the cost intervals. Say 46% of tuitions cost about

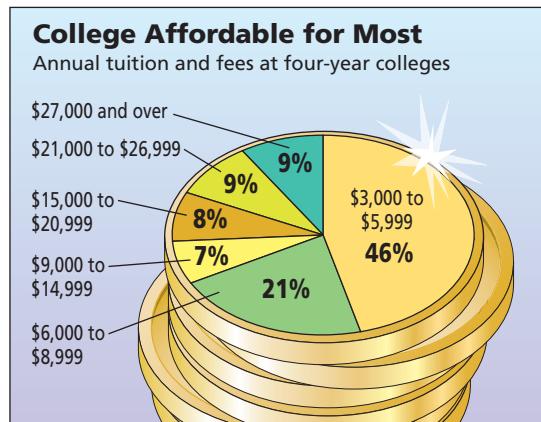


Myriam Ferguson Cate/PhotoEdit

\$4500; 21% cost about \$7500; 7% cost about \$12,000; 8% cost about \$18,000; 9% cost about \$24,000; and 9% cost about \$31,000. Compute the weighted average of college tuition charged at all colleges. (See Problem 9 in the Chapter Review Problems.)



Source: Census Bureau



Source: The College Board

SECTION 3.1

Measures of Central Tendency: Mode, Median, and Mean

FOCUS POINTS

- Compute mean, median, and mode from raw data.
- Interpret what mean, median, and mode tell you.
- Explain how mean, median, and mode can be affected by extreme data values.
- What is a trimmed mean? How do you compute it?
- Compute a weighted average.

The average price of an ounce of gold is \$1200. The Zippy car averages 39 miles per gallon on the highway. A survey showed the average shoe size for women is size 9.

In each of the preceding statements, *one* number is used to describe the entire sample or population. Such a number is called an *average*. There are many ways to compute averages, but we will study only three of the major ones.

The easiest average to compute is the *mode*.

Average

Mode

The **mode** of a data set is the value that occurs most frequently.



EXAMPLE 1

MODE

Count the letters in each word of this sentence and give the mode. The numbers of letters in the words of the sentence are

5 3 7 2 4 4 2 4 8 3 4 3 4

Scanning the data, we see that 4 is the mode because more words have 4 letters than any other number. For larger data sets, it is useful to order—or sort—the data before scanning them for the mode.



Not every data set has a mode. For example, if Professor Fair gives equal numbers of A's, B's, C's, D's, and F's, then there is no modal grade. In addition, the mode is not very stable. Changing just one number in a data set can change the mode dramatically. However, the mode is a useful average when we want to know the most frequently occurring data value, such as the most frequently requested shoe size.

Median

Another average that is useful is the *median*, or central value, of an ordered distribution. When you are given the median, you know there are an equal number of data values in the ordered distribution that are above it and below it.

PROCEDURE

HOW TO FIND THE MEDIAN

The **median** is the central value of an ordered distribution. To find it,

1. Order the data from smallest to largest.
2. For an *odd* number of data values in the distribution,

$$\text{Median} = \text{Middle data value}$$

3. For an *even* number of data values in the distribution,

$$\text{Median} = \frac{\text{Sum of middle two values}}{2}$$

EXAMPLE 2

MEDIAN

What do barbecue-flavored potato chips cost? According to *Consumer Reports*, Vol. 66, No. 5, the prices per ounce in cents of the rated chips are

19 19 27 28 18 35

- (a) To find the median, we first order the data, and then note that there are an even number of entries. So the median is constructed using the two middle values.

18 19 19 27 28 35
 \ /
 middle values

$$\text{Median} = \frac{19 + 27}{2} = 23 \text{ cents}$$

- (b) According to *Consumer Reports*, the brand with the lowest overall taste rating costs 35 cents per ounce. Eliminate that brand, and find the median price per ounce for the remaining barbecue-flavored chips. Again order the data. Note that there are an odd number of entries, so the median is simply the middle value.

18 19 19 27 28
 ↑
 middle values

$$\text{Median} = \text{middle value} = 19 \text{ cents}$$

- (c) One ounce of potato chips is considered a small serving. Is it reasonable to budget about \$10.45 to serve the barbecue-flavored chips to 55 people?

Yes, since the median price of the chips is 19 cents per small serving. This budget for chips assumes that there is plenty of other food!



The median uses the *position* rather than the specific value of each data entry. If the extreme values of a data set change, the median usually does not change. This is why the median is often used as the average for house prices. If one mansion costing several million dollars sells in a community of much-lower-priced homes, the median selling price for houses in the community would be affected very little, if at all.

GUIDED EXERCISE 1

Median and mode

Belleview College must make a report to the budget committee about the average credit hour load a full-time student carries. (A 12-credit-hour load is the minimum requirement for full-time status. For the same tuition, students may take up to 20 credit hours.) A random sample of 40 students yielded the following information (in credit hours):

17	12	14	17	13	16	18	20	13	12
12	17	16	15	14	12	12	13	17	14
15	12	15	16	12	18	20	19	12	15
18	14	16	17	15	19	12	13	12	15

- (a) Organize the data from smallest to largest number of credit hours.

→ 12 12 12 12 12 12 12 12 12 12
13 13 13 13 14 14 14 14 15 15
15 15 15 15 16 16 16 16 17 17
17 17 17 18 18 18 18 19 19 20

- (b) Since there are an _____ (odd, even) number of values, we add the two middle values and divide by 2 to get the median. What is the median credit hour load?

→ There are an even number of entries. The two middle values are circled in part (a).

$$\text{Median} = \frac{15 + 15}{2} = 15$$

- (c) What is the mode of this distribution? Is it different from the median? *Interpretation* If the budget committee is going to fund the college according to the average student credit hour load (more money for higher loads), which of these two averages do you think the college will report?

→ The mode is 12. It is different from the median. Since the median is higher, the college will probably report it and indicate that the average being used is the median.

Note: For small ordered data sets, we can easily scan the set to find the *location* of the median. However, for large ordered data sets of size n , it is convenient to have a formula to find the middle of the data set.

For an ordered data set of size n ,

$$\text{Position of the middle value} = \frac{n + 1}{2}$$

For instance, if $n = 99$ then the middle value is the $(99 + 1)/2$ or 50th data value in the ordered data. If $n = 100$, then $(100 + 1)/2 = 50.5$ tells us that the two middle values are in the 50th and 51st positions.

Mean

An average that uses the exact value of each entry is the *mean* (sometimes called the *arithmetic mean*). To compute the mean, we add the values of all the entries and then divide by the number of entries.

$$\text{Mean} = \frac{\text{Sum of all entries}}{\text{Number of entries}}$$

The mean is the average usually used to compute a test average.

EXAMPLE 3**MEAN**

To graduate, Linda needs at least a B in biology. She did not do very well on her first three tests; however, she did well on the last four. Here are her scores:

58 67 60 84 93 98 100

Compute the mean and determine if Linda's grade will be a B (80 to 89 average) or a C (70 to 79 average).

SOLUTION:

$$\begin{aligned}\text{Mean} &= \frac{\text{Sum of scores}}{\text{Number of scores}} = \frac{58 + 67 + 60 + 84 + 93 + 98 + 100}{7} \\ &= \frac{560}{7} = 80\end{aligned}$$

Since the average is 80, Linda will get the needed B.

Jose Carillo/PhotoEdit

**Summation symbol, Σ** **Formulas for the mean**

COMMENT When we compute the mean, we sum the given data. There is a convenient notation to indicate the sum. Let x represent any value in the data set. Then the notation

$$\Sigma x \text{ (read "the sum of all given } x \text{ values")}$$

means that we are to sum all the data values. In other words, we are to sum all the entries in the distribution. The *summation symbol* Σ means *sum the following* and is capital sigma, the S of the Greek alphabet.

The symbol for the mean of a *sample* distribution of x values is denoted by \bar{x} (read “ x bar”). If your data comprise the entire *population*, we use the symbol μ (lowercase Greek letter mu, pronounced “mew”) to represent the mean.

PROCEDURE

When data consist of rates of change, the *harmonic mean* is the appropriate average to use. Problem 25 discusses the harmonic mean.



Problem 26 discusses the *geometric mean*, which is a useful average for percentages, ratios, or growth rates.

HOW TO FIND THE MEAN

1. Compute Σx ; that is, find the sum of all the data values.
2. Divide the sum total by the number of data values.

Sample statistic \bar{x}

$$\bar{x} = \frac{\Sigma x}{n}$$

Population parameter μ

$$\mu = \frac{\Sigma x}{N}$$

where n = number of data values in the sample

N = number of data values in the population

LOOKING FORWARD

In our future work with inferential statistics, we will use the mean \bar{x} from a random sample to estimate the population parameter μ (Chapter 7) or to make decisions regarding the value of μ (Chapter 8).

CALCULATOR NOTE It is very easy to compute the mean on *any* calculator: Simply add the data values and divide the total by the number of data. However, on calculators with a statistics mode, you place the calculator in that mode, *enter* the data, and then press the key for the mean. The key is usually designated \bar{x} . Because the formula for the population mean is the same as that for the sample mean, the same key gives the value for μ .

We have seen three averages: the mode, the median, and the mean. For later work, the mean is the most important. A disadvantage of the mean, however, is that it can be affected by exceptional values.

Resistant measure

A *resistant measure* is one that is not influenced by extremely high or low data values. The mean is not a resistant measure of center because we can make the mean as large as we want by changing the size of only one data value. The median, on the other hand, is more resistant. However, a disadvantage of the median is that it is not sensitive to the specific size of a data value.

Trimmed mean

A measure of center that is more resistant than the mean but still sensitive to specific data values is the *trimmed mean*. A trimmed mean is the mean of the data values left after “trimming” a specified percentage of the smallest and largest data values from the data set. Usually a 5% trimmed mean is used. This implies that we trim the lowest 5% of the data as well as the highest 5% of the data. A similar procedure is used for a 10% trimmed mean.

PROCEDURE

HOW TO COMPUTE A 5% TRIMMED MEAN

1. Order the data from smallest to largest.
2. Delete the bottom 5% of the data and the top 5% of the data. *Note:* If the calculation of 5% of the number of data values does not produce a whole number, *round* to the nearest integer.
3. Compute the mean of the remaining 90% of the data.

GUIDED EXERCISE 2

Mean and trimmed mean

Barron's *Profiles of American Colleges*, 19th Edition, lists average class size for introductory lecture courses at each of the profiled institutions. A sample of 20 colleges and universities in California showed class sizes for introductory lecture courses to be

(14)	20	20	20	20	23	25	30	30	30
35	35	35	40	40	42	50	50	80	(80)

(a) Compute the mean for the entire sample.



Add all the values and divide by 20:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{719}{20} \approx 36.0$$

Continued

GUIDED EXERCISE 2 *continued*

(b) Compute a 5% trimmed mean for the sample.



The data are already ordered. Since 5% of 20 is 1, we eliminate one data value from the bottom of the list and one from the top. These values are circled in the data set. Then we take the mean of the remaining 18 entries.

$$5\% \text{ trimmed mean} = \frac{\Sigma x}{n} = \frac{625}{18} \approx 34.7$$

(c) Find the median of the original data set.



Note that the data are already ordered.

$$\text{Median} = \frac{30 + 35}{2} = 32.5$$

(d) Find the median of the 5% trimmed data set.

Does the median change when you trim the data?



The median is still 32.5. Notice that trimming the same number of entries from both ends leaves the middle position of the data set unchanged.

(e) Is the trimmed mean or the original mean closer to the median?



The trimmed mean is closer to the median.

 TECH NOTES

Minitab, Excel 2007, and TI-84Plus/TI-83Plus/TI-nspire calculators all provide the mean and median of a data set. Minitab and Excel 2007 also provide the mode. The TI-84Plus/TI-83Plus/TI-nspire calculators sort data, so you can easily scan the sorted data for the mode. Minitab provides the 5% trimmed mean, as does Excel 2007.

All this technology is a wonderful aid for analyzing data. However, *a measurement is worthless if you do not know what it represents or how a change in data values might affect the measurement*. The defining formulas and procedures for computing the measures tell you a great deal about the measures. Even if you use a calculator to evaluate all the statistical measures, pay attention to the information the formulas and procedures give you about the components or features of the measurement.


 CRITICAL THINKING

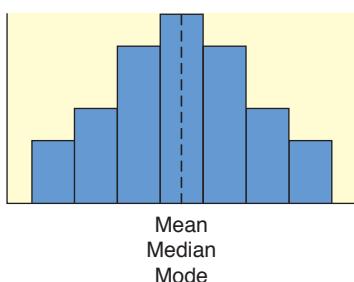
In Chapter 1, we examined four levels of data: nominal, ordinal, interval, and ratio. The mode (if it exists) can be used with all four levels, including nominal. For instance, the modal color of all passenger cars sold last year might be blue. The median may be used with data at the ordinal level or above. If we ranked the passenger cars in order of customer satisfaction level, we could identify the median satisfaction level. For the mean, our data need to be at the interval or ratio level (although there are exceptions in which the mean of ordinal-level data is computed). We can certainly find the mean model year of used passenger cars sold or the mean price of new passenger cars.

Another issue of concern is that of taking the average of averages. For instance, if the values \$520, \$640, \$730, \$890, and \$920 represent the mean monthly rents for five different apartment complexes, we can't say that \$740 (the mean of the five numbers) is the mean monthly rent of all the apartments. We need to know the number of apartments in each complex before we can determine an average based on the number of apartments renting at each designated amount.

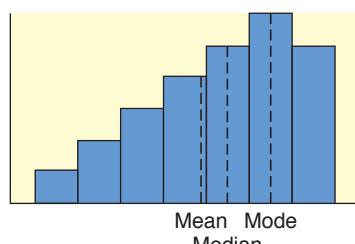
Data types and averages

FIGURE 3-1**Distribution Types and Averages**

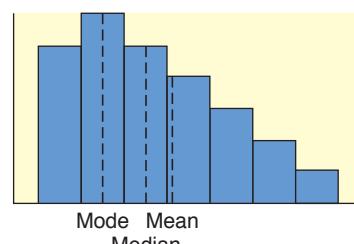
(a) Mound-shaped symmetrical



(b) Skewed left



(c) Skewed right

**Distribution shapes and averages**

In general, when a data distribution is mound-shaped symmetrical, the values for the mean, median, and mode are the same or almost the same. For skewed-left distributions, the mean is less than the median and the median is less than the mode. For skewed-right distributions, the mode is the smallest value, the median is the next largest, and the mean is the largest. Figure 3-1 shows the general relationships among the mean, median, and mode for different types of distributions.

**Weighted Average****Weighted average**

Sometimes we wish to average numbers, but we want to assign more importance, or weight, to some of the numbers. For instance, suppose your professor tells you that your grade will be based on a midterm and a final exam, each of which is based on 100 possible points. However, the final exam will be worth 60% of the grade and the midterm only 40%. How could you determine an average score that would reflect these different weights? The average you need is the *weighted average*.

$$\text{Weighted average} = \frac{\sum xw}{\sum w}$$

where x is a data value and w is the weight assigned to that data value. The sum is taken over all data values.

EXAMPLE 4**WEIGHTED AVERAGE**

Suppose your midterm test score is 83 and your final exam score is 95. Using weights of 40% for the midterm and 60% for the final exam, compute the weighted average of your scores. If the minimum average for an A is 90, will you earn an A?

SOLUTION: By the formula, we multiply each score by its weight and add the results together. Then we divide by the sum of all the weights. Converting the percentages to decimal notation, we get

$$\begin{aligned}\text{Weighted average} &= \frac{83(0.40) + 95(0.60)}{0.40 + 0.60} \\ &= \frac{33.2 + 57}{1} = 90.2\end{aligned}$$

Your average is high enough to earn an A.

**TECH NOTES**

The TI-84Plus/TI-83Plus/TI-nspire calculators directly support weighted averages. Both Excel 2007 and Minitab can be programmed to provide the averages.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84 Plus keypad) Enter the data into one list, such as L1, and the corresponding weights into another list, such as L2. Then press **Stat ▶ Calc ▶ 1: 1-Var Stats**. Enter the list containing the data, followed by a comma and the list containing the weights.

**VIEWPOINT****What's Wrong with Pitching Today?**

One way to answer this question is to look at averages. Batting averages and average hits per game are shown for selected years from 1901 to 2000 (Source: Wall Street Journal).

Year	1901	1920	1930	1941	1951	1961	1968	1976	1986	2000
B.A.	0.277	0.284	0.288	0.267	0.263	0.256	0.231	0.256	0.262	0.276
Hits	19.2	19.2	20.0	18.4	17.9	17.3	15.2	17.3	17.8	19.1

A quick scan of the averages shows that batting averages and average hits per game are virtually the same as they were almost 100 years ago. It seems there is nothing wrong with today's pitching! So what's changed? For one thing, the rules have changed! The strike zone is considerably smaller than it once was, and the pitching mound is lower. Both give the hitter an advantage over the pitcher. Even so, pitchers don't give up hits with any greater frequency than they did a century ago (look at the averages). However, modern hits go much farther, which is something a pitcher can't control.

**SECTION 3.1
PROBLEMS**

1. **Statistical Literacy** Consider the mode, median, and mean. Which average represents the middle value of a data distribution? Which average represents the most frequent value of a distribution? Which average takes all the specific values into account?
2. **Statistical Literacy** What symbol is used for the arithmetic mean when it is a sample statistic? What symbol is used when the arithmetic mean is a population parameter?
3. **Basic Computation: Mean, Median, Mode** Find the mean, median, and mode of the data set
8 2 7 2 6
4. **Basic Computation: Mean, Median, Mode** Find the mean, median, and mode of the data set
10 12 20 15 20
5. **Basic Computation: Mean, Median, Mode** Find the mean, median, and mode of the data set
8 2 7 2 6 5
6. **Critical Thinking** If a data set has an even number of data, is it true or false that the median is never equal to a value in the data set? Explain.
7. **Critical Thinking** When a distribution is mound-shaped symmetrical, what is the general relationship among the values of the mean, median, and mode?

8. **Critical Thinking** Consider the following types of data that were obtained from a random sample of 49 credit card accounts. Identify all the averages (mean, median, or mode) that can be used to summarize the data.
- Outstanding balance on each account
 - Name of credit card (e.g., MasterCard, Visa, American Express, etc.)
 - Dollar amount due on next payment
9. **Critical Thinking** Consider the numbers
 2 3 4 5 5
- Compute the mode, median, and mean.
 - If the numbers represent codes for the colors of T-shirts ordered from a catalog, which average(s) would make sense?
 - If the numbers represent one-way mileages for trails to different lakes, which average(s) would make sense?
 - Suppose the numbers represent survey responses from 1 to 5, with 1 = disagree strongly, 2 = disagree, 3 = agree, 4 = agree strongly, and 5 = agree very strongly. Which averages make sense?
10. **Critical Thinking** Consider two data sets.
 Set A: $n = 5; \bar{x} = 10$ Set B: $n = 50; \bar{x} = 10$
- Suppose the number 20 is included as an additional data value in Set A. Compute \bar{x} for the new data set. Hint: $\Sigma x = n\bar{x}$. To compute \bar{x} for the new data set, add 20 to Σx of the original data set and divide by 6.
 - Suppose the number 20 is included as an additional data value in Set B. Compute \bar{x} for the new data set.
 - Why does the addition of the number 20 to each data set change the mean for Set A more than it does for Set B?
11. **Interpretation** A job-performance evaluation form has these categories:
 1 = excellent; 2 = good; 3 = satisfactory; 4 = poor; 5 = unacceptable
 Based on 15 client reviews, one employee had
 median rating of 4; mode rating of 1
- The employee was pleased that most clients had rated her as excellent. The supervisor said improvement was needed because at least half the clients had rated the employee at the poor or unacceptable level. Comment on the different perspectives.
12. **Critical Thinking: Data Transformation** In this problem, we explore the effect on the mean, median, and mode of adding the same number to each data value. Consider the data set 2, 2, 3, 6, 10.
- Compute the mode, median, and mean.
 - Add 5 to each of the data values. Compute the mode, median, and mean.
 - Compare the results of parts (a) and (b). In general, how do you think the mode, median, and mean are affected when the same constant is added to each data value in a set?
13. **Critical Thinking: Data Transformation** In this problem, we explore the effect on the mean, median, and mode of multiplying each data value by the same number. Consider the data set 2, 2, 3, 6, 10.
- Compute the mode, median, and mean.
 - Multiply each data value by 5. Compute the mode, median, and mean.
 - Compare the results of parts (a) and (b). In general, how do you think the mode, median, and mean are affected when each data value in a set is multiplied by the same constant?

- (d) Suppose you have information about average heights of a random sample of airplane passengers. The mode is 70 inches, the median is 68 inches, and the mean is 71 inches. To convert the data into centimeters, multiply each data value by 2.54. What are the values of the mode, median, and mean in centimeters?
14. **Critical Thinking** Consider a data set of 15 distinct measurements with mean A and median B .
- If the highest number were increased, what would be the effect on the median and mean? Explain.
 - If the highest number were decreased to a value still larger than B , what would be the effect on the median and mean?
 - If the highest number were decreased to a value smaller than B , what would be the effect on the median and mean?
15. **Environmental Studies: Death Valley** How hot does it get in Death Valley? The following data are taken from a study conducted by the National Park System, of which Death Valley is a unit. The ground temperatures ($^{\circ}\text{F}$) were taken from May to November in the vicinity of Furnace Creek.

146	152	168	174	180	178	179
180	178	178	168	165	152	144

Compute the mean, median, and mode for these ground temperatures.

16. **Ecology: Wolf Packs** How large is a wolf pack? The following information is from a random sample of winter wolf packs in regions of Alaska, Minnesota, Michigan, Wisconsin, Canada, and Finland (Source: *The Wolf*, by L. D. Mech, University of Minnesota Press). Winter pack size:

13	10	7	5	7	7	2	4	3
2	3	15	4	4	2	8	7	8

Compute the mean, median, and mode for the size of winter wolf packs.

17. **Medical: Injuries** The Grand Canyon and the Colorado River are beautiful, rugged, and sometimes dangerous. Thomas Myers is a physician at the park clinic in Grand Canyon Village. Dr. Myers has recorded (for a 5-year period) the number of visitor injuries at different landing points for commercial boat trips down the Colorado River in both the Upper and Lower Grand Canyon (Source: *Fateful Journey* by Myers, Becker, and Stevens).

Upper Canyon: Number of Injuries per Landing Point Between North Canyon and Phantom Ranch

2	3	1	1	3	4	6	9	3	1	3
---	---	---	---	---	---	---	---	---	---	---

Lower Canyon: Number of Injuries per Landing Point Between Bright Angel and Lava Falls

8	1	1	0	6	7	2	14	3	0	1	13	2	1
---	---	---	---	---	---	---	----	---	---	---	----	---	---

- Compute the mean, median, and mode for injuries per landing point in the Upper Canyon.
- Compute the mean, median, and mode for injuries per landing point in the Lower Canyon.
- Compare the results of parts (a) and (b).
- The Lower Canyon stretch had some extreme data values. Compute a 5% trimmed mean for this region, and compare this result to the mean for the Upper Canyon computed in part (a).





Dale Walsh/istockphoto.com

18. ***Football: Age of Professional Players*** How old are professional football players? The 11th Edition of *The Pro Football Encyclopedia* gave the following information. Random sample of pro football player ages in years:
- | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 24 | 23 | 25 | 23 | 30 | 29 | 28 | 26 | 33 | 29 |
| 24 | 37 | 25 | 23 | 22 | 27 | 28 | 25 | 31 | 29 |
| 25 | 22 | 31 | 29 | 22 | 28 | 27 | 26 | 23 | 21 |
| 25 | 21 | 25 | 24 | 22 | 26 | 25 | 32 | 26 | 29 |
- (a) Compute the mean, median, and mode of the ages.
 (b) ***Interpretation*** Compare the averages. Does one seem to represent the age of the pro football players most accurately? Explain.
19. ***Leisure: Maui Vacation*** How expensive is Maui? If you want a vacation rental condominium (up to four people), visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase>, find the link to Maui, and then search for accommodations. The *Maui News* gave the following costs in dollars per day for a random sample of condominiums located throughout the island of Maui.
- | | | | | | | | | | |
|----|----|-----|----|-----|-----|-----|----|----|-----|
| 89 | 50 | 68 | 60 | 375 | 55 | 500 | 71 | 40 | 350 |
| 60 | 50 | 250 | 45 | 45 | 125 | 235 | 65 | 60 | 130 |
- (a) Compute the mean, median, and mode for the data.
 (b) Compute a 5% trimmed mean for the data, and compare it with the mean computed in part (a). Does the trimmed mean more accurately reflect the general level of the daily rental costs?
 (c) ***Interpretation*** If you were a travel agent and a client asked about the daily cost of renting a condominium on Maui, what average would you use? Explain. Is there any other information about the costs that you think might be useful, such as the spread of the costs?
20. ***Basic Computation: Weighted Average*** Find the weighted average of a data set where
 10 has a weight of 5; 20 has a weight of 3; 30 has a weight of 2
21. ***Basic Computation: Weighted Average*** Find the weighted average of a data set where
 10 has a weight of 2; 20 has a weight of 3; 30 has a weight of 5
22. ***Grades: Weighted Average*** In your biology class, your final grade is based on several things: a lab score, scores on two major tests, and your score on the final exam. There are 100 points available for each score. However, the lab score is worth 25% of your total grade, each major test is worth 22.5%, and the final exam is worth 30%. Compute the weighted average for the following scores: 92 on the lab, 81 on the first major test, 93 on the second major test, and 85 on the final exam.
23. ***Merit Pay Scale: Weighted Average*** At General Hospital, nurses are given performance evaluations to determine eligibility for merit pay raises. The supervisor rates the nurses on a scale of 1 to 10 (10 being the highest rating) for several activities: promptness, record keeping, appearance, and bedside manner with patients. Then an average is determined by giving a weight of 2 for promptness, 3 for record keeping, 1 for appearance, and 4 for bedside manner with patients. What is the average rating for a nurse with ratings of 9 for promptness, 7 for record keeping, 6 for appearance, and 10 for bedside manner?
24. ***EPA: Wetlands*** Where does all the water go? According to the Environmental Protection Agency (EPA), in a typical wetland environment, 38% of the water is outflow; 47% is seepage; 7% evaporates; and 8% remains as water volume

in the ecosystem (Reference: U.S. Environmental Protection Agency Case Studies Report 832-R-93-005). Chloride compounds as residuals from residential areas are a problem for wetlands. Suppose that in a particular wetland environment the following concentrations (mg/l) of chloride compounds were found: outflow, 64.1; seepage, 75.8; remaining due to evaporation, 23.9; in the water volume, 68.2.

- Compute the weighted average of chlorine compound concentration (mg/l) for this ecological system.
- Suppose the EPA has established an average chlorine compound concentration target of no more than 58 mg/l. Comment on whether this wetlands system meets the target standard for chlorine compound concentration.



25. **Expand Your Knowledge: Harmonic Mean** When data consist of rates of change, such as speeds, the *harmonic mean* is an appropriate measure of central tendency. For n data values,

$$\text{Harmonic mean} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}, \quad \text{assuming no data value is 0}$$

Suppose you drive 60 miles per hour for 100 miles, then 75 miles per hour for 100 miles. Use the harmonic mean to find your average speed.



26. **Expand Your Knowledge: Geometric Mean** When data consist of percentages, ratios, growth rates, or other rates of change, the *geometric mean* is a useful measure of central tendency. For n data values,

$$\text{Geometric mean} = \sqrt[n]{\text{product of the } n \text{ data values}}, \quad \text{assuming all data values are positive}$$

To find the *average growth factor* over 5 years of an investment in a mutual fund with growth rates of 10% the first year, 12% the second year, 14.8% the third year, 3.8% the fourth year, and 6% the fifth year, take the geometric mean of 1.10, 1.12, 1.148, 1.038, and 1.16. Find the average growth factor of this investment.

Note that for the same data, the relationships among the harmonic, geometric, and arithmetic means are harmonic mean \leq geometric mean \leq arithmetic mean (Source: *Oxford Dictionary of Statistics*).

SECTION 3.2

Measures of Variation

FOCUS POINTS

- Find the range, variance, and standard deviation.
- Compute the coefficient of variation from raw data. Why is the coefficient of variation important?
- Apply Chebyshev's theorem to raw data. What does a Chebyshev interval tell us?

An average is an attempt to summarize a set of data using just one number. As some of our examples have shown, an average taken by itself may not always be very meaningful. We need a statistical cross-reference that measures the spread of the data.

The *range* is one such measure of variation.

Range

The **range** is the difference between the largest and smallest values of a data distribution.

EXAMPLE 5**RANGE**

A large bakery regularly orders cartons of Maine blueberries. The average weight of the cartons is supposed to be 22 ounces. Random samples of cartons from two suppliers were weighed. The weights in ounces of the cartons were

Supplier I:	17	22	22	22	27
Supplier II:	17	19	20	27	27

- (a) Compute the range of carton weights from each supplier.

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

$$\text{Supplier I range} = 27 - 17 = 10 \text{ ounces}$$

$$\text{Supplier II range} = 27 - 17 = 10 \text{ ounces}$$

- (b) Compute the mean weight of cartons from each supplier. In both cases the mean is 22 ounces.

- (c) Look at the two samples again. The samples have the same range and mean. How do they differ? The bakery uses one carton of blueberries in each blueberry muffin recipe. It is important that the cartons be of consistent weight so that the muffins turn out right.

Supplier I provides more cartons that have weights closer to the mean. Or, put another way, the weights of cartons from Supplier I are more clustered around the mean. The bakery might find Supplier I more satisfactory.

David Young-Wolff/PhotoEdit



Blueberry patch

Variance
Standard deviation

Variance and Standard Deviation

We need a measure of the distribution or spread of data around an expected value (either \bar{x} or μ). *Variance* and *standard deviation* provide such measures. Formulas and rationale for these measures are described in the next Procedure display. Then, examples and guided exercises show how to compute and interpret these measures.

As we will see later, the formulas for variance and standard deviation differ slightly, depending on whether we are using a sample or the entire population.

PROCEDURE

HOW TO COMPUTE THE SAMPLE VARIANCE AND SAMPLE STANDARD DEVIATION

Quantity	Description
x	The variable x represents a data value or outcome.
Mean $\bar{x} = \frac{\Sigma x}{n}$	This is the average of the data values , or what you "expect" to happen the next time you conduct the statistical experiment. Note that n is the sample size.
$x - \bar{x}$	This is the difference between what happened and what you expected to happen. This represents a "deviation" away from what you "expect" and is a measure of risk.

Continued

Sum of squares

$$\Sigma(x - \bar{x})^2$$

The expression $\Sigma(x - \bar{x})^2$ is called the **sum of squares**. The $(x - \bar{x})$ quantity is squared to make it nonnegative. The sum is over all the data. If you don't square $(x - \bar{x})$, then the sum $\Sigma(x - \bar{x})$ is equal to 0 because the negative values cancel the positive values. This occurs even if some $(x - \bar{x})$ values are large, indicating a large deviation or risk.

Sum of squares

$$\begin{aligned} &\Sigma(x - \bar{x})^2 \\ &\text{or} \\ &\Sigma x^2 - \frac{(\Sigma x)^2}{n} \end{aligned}$$

This is an **algebraic simplification of the sum of squares** that is easier to compute.

The **defining formula** for the sum of squares is the upper one.

The **computation formula** for the sum of squares is the lower one. Both formulas give the same result.

Sample variance, s^2 **Sample variance**

$$\begin{aligned} s^2 &= \frac{\Sigma(x - \bar{x})^2}{n - 1} \\ &\text{or} \\ s^2 &= \frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1} \end{aligned}$$

The **sample variance is s^2** . The variance can be thought of as a kind of average of the $(x - \bar{x})^2$ values. However, for technical reasons, we divide the sum by the quantity $n - 1$ rather than n . This gives us the best mathematical estimate for the sample variance.

The **defining formula** for the variance is the upper one. The **computation formula** for the variance is the lower one. Both formulas give the same result.

Sample standard deviation, s **Sample standard deviation**

$$\begin{aligned} s &= \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} \\ &\text{or} \\ s &= \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}} \end{aligned}$$

This is the **sample standard deviation, s** . Why do we take the square root? Well, if the original x units were, say, days or dollars, then the s^2 units would be days squared or dollars squared (wow, what's that?). We take the square root to return to the original units of the data measurements. The standard deviation can be thought of as a measure of variability or risk. Larger values of s imply greater variability in the data.

The **defining formula** for the standard deviation is the upper one. The **computation formula** for the standard deviation is the lower one. Both formulas give the same result.

COMMENT Why is s called a *sample standard deviation*? First, it is computed from sample data. Then why do we use the word *standard* in the name? We know s is a measure of deviation or risk. You should be aware that there are other statistical measures of risk that we have not yet mentioned. However, s is the one that everyone uses, so it is called the “standard” (like standard time).

In statistics, the sample standard deviation and sample variance are used to describe the spread of data about the mean \bar{x} . The next example shows how to find these quantities by using the defining formulas. Guided Exercise 3 shows how to use the computation formulas.

As you will discover, for “hand” calculations, the computation formulas for s^2 and s are much easier to use. However, the defining formulas for s^2 and s emphasize the fact that the variance and standard deviation are based on the differences between each data value and the mean.

Defining Formulas (Sample Statistic)

$$\text{Sample variance} = s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} \quad (1)$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \quad (2)$$

where x is a member of the data set, \bar{x} is the mean, and n is the number of data values. The sum is taken over all data values.

Computation Formulas (Sample Statistic)

$$\text{Sample variance} = s^2 = \frac{\sum x^2 - (\sum x)^2/n}{n - 1} \quad (3)$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n - 1}} \quad (4)$$

where x is a member of the data set, \bar{x} is the mean, and n is the number of data values. The sum is taken over all data values.

**EXAMPLE 6****SAMPLE STANDARD DEVIATION (DEFINING FORMULA)**

Big Blossom Greenhouse was commissioned to develop an extra large rose for the Rose Bowl Parade. A random sample of blossoms from Hybrid A bushes yielded the following diameters (in inches) for mature peak blooms.

2 3 3 8 10 10

Use the defining formula to find the sample variance and standard deviation.

SOLUTION: Several steps are involved in computing the variance and standard deviation. A table will be helpful (see Table 3-1). Since $n = 6$, we take the sum of the entries in the first column of Table 3-1 and divide by 6 to find the mean \bar{x} .

$$\bar{x} = \frac{\sum x}{n} = \frac{36}{6} = 6.0 \text{ inches}$$

TABLE 3-1 **Diameters of Rose Blossoms (in inches)**

Column I x	Column II $x - \bar{x}$	Column III $(x - \bar{x})^2$
2	$2 - 6 = -4$	$(-4)^2 = 16$
3	$3 - 6 = -3$	$(-3)^2 = 9$
3	$3 - 6 = -3$	$(-3)^2 = 9$
8	$8 - 6 = 2$	$(2)^2 = 4$
10	$10 - 6 = 4$	$(4)^2 = 16$
10	$10 - 6 = 4$	$(4)^2 = 16$
$\Sigma x = 36$		$\Sigma(x - \bar{x})^2 = 70$

Using this value for \bar{x} , we obtain Column II. Square each value in the second column to obtain Column III, and then add the values in Column III. To get the sample variance, divide the sum of Column III by $n - 1$. Since $n = 6$, $n - 1 = 5$.

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{70}{5} = 14$$



Now obtain the sample standard deviation by taking the square root of the variance.

$$s = \sqrt{s^2} = \sqrt{14} \approx 3.74$$

(Use a calculator to compute the square root. Because of rounding, we use the approximately equal symbol, \approx .)

GUIDED EXERCISE 3

Sample standard deviation (computation formula)

Big Blossom Greenhouse gathered another random sample of mature peak blooms from Hybrid B. The six blossoms had the following widths (in inches):

5 5 5 6 7 8

- (a) Again, we will construct a table so that we can find the mean, variance, and standard deviation more easily. In this case, what is the value of n ? Find the sum of Column I in Table 3-2, and compute the mean.

TABLE 3-2 Complete Columns I and II

I x	II x^2
5	_____
5	_____
5	_____
6	_____
7	_____
8	_____
$\Sigma x =$ _____	$\Sigma x^2 =$ _____

➡ $n = 6$. The sum of Column I is $\Sigma x = 36$, so the mean is

$$\bar{x} = \frac{36}{6} = 6$$

TABLE 3-3 Completion of Table 3-2

I x	II x^2
5	25
5	25
5	25
6	36
7	49
8	64
$\Sigma x = 36$	$\Sigma x^2 = 224$

- (b) What is the value of n ? of $n - 1$? Use the computation formula to find the sample variance s^2 . *Note:* Be sure to distinguish between Σx^2 and $(\Sigma x)^2$. For Σx^2 , you square the x values first and then sum them. For $(\Sigma x)^2$, you sum the x values first and then square the result.
- (c) Use a calculator to find the square root of the variance. Is this the standard deviation?

➡ $n = 6; n - 1 = 5$.

$$s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}$$

$$= \frac{224 - (36)^2/6}{5} = \frac{8}{5} = 1.6$$

$$s = \sqrt{s^2} = \sqrt{1.6} \approx 1.26$$

Yes.

Let's summarize and compare the results of Guided Exercise 3 and Example 6. The greenhouse found the following blossom diameters for Hybrid A and Hybrid B:

Hybrid A: Mean, 6.0 inches; standard deviation, 3.74 inches

Hybrid B: Mean, 6.0 inches; standard deviation, 1.26 inches

Interpretation In both cases, the means are the same: 6 inches. But the first hybrid has a larger standard deviation. This means that the blossoms of Hybrid A are less consistent than those of Hybrid B. If you want a rosebush that occasionally

has 10-inch blooms and occasionally has 2-inch blooms, use the first hybrid. But if you want a bush that consistently produces roses close to 6 inches across, use Hybrid B.

ROUNDING NOTE Rounding errors cannot be completely eliminated, even if a computer or calculator does all the computations. However, software and calculator routines are designed to minimize the errors. If the mean is rounded, the value of the standard deviation will change slightly, depending on how much the mean is rounded. If you do your calculations “by hand” or reenter intermediate values into a calculator, try to carry one or two more digits than occur in the original data. If your resulting answers vary slightly from those in this text, do not be overly concerned. The text answers are computer- or calculator-generated.

In most applications of statistics, we work with a random sample of data rather than the entire population of *all* possible data values. However, if we have data for the entire population, we can compute the *population mean* μ , *population variance* σ^2 , and *population standard deviation* σ (lowercase Greek letter sigma) using the following formulas:

Population mean, μ
 Population variance, σ^2
 Population standard deviation, σ



What if you do not have access to raw data, but you do have a histogram or frequency table of the data and you want to estimate the mean and standard deviation? Problems 22–25 address grouped data, with the formulas shown before Problem 22.



Problem 26 discusses moving averages, and explores the difference in the standard deviations of raw data compared to the standard deviation of the moving average of the raw data.

Population size, N

Population Parameters

$$\text{Population mean} = \mu = \frac{\Sigma x}{N}$$

$$\text{Population variance} = \sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$

where N is the number of data values in the population and x represents the individual data values of the population.

We note that the formula for μ is the same as the formula for \bar{x} (the sample mean) and that the formulas for σ^2 and σ are the same as those for s^2 and s (sample variance and sample standard deviation), except that the population size N is used instead of $n - 1$. Also, μ is used instead of \bar{x} in the formulas for σ^2 and σ .

In the formulas for s and σ , we use $n - 1$ to compute s and N to compute σ . Why? The reason is that N (capital letter) represents the *population size*, whereas n (lowercase letter) represents the sample size. Since a random sample usually will not contain extreme data values (large or small), we divide by $n - 1$ in the formula for s to make s a little larger than it would have been had we divided by n . Courses in advanced theoretical statistics show that this procedure will give us the best possible estimate for the standard deviation σ . In fact, s is called the *unbiased estimate* for σ . If we have the population of all data values, then extreme data values are, of course, present, so we divide by N instead of $N - 1$.

COMMENT The computation formula for the population standard deviation is

$$\sigma = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/N}{N}}$$

We've seen that the standard deviation (sample or population) is a measure of data spread. We will use the standard deviation extensively in later chapters.

**TECH NOTES**

Most scientific or business calculators have a statistics mode and provide the mean and sample standard deviation directly. The TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab provide the median and several other measures as well.

Many technologies display only the sample standard deviation s . You can quickly compute σ if you know s by using the formula

$$\sigma = s\sqrt{\frac{n-1}{n}}$$

The mean given in displays can be interpreted as the sample mean \bar{x} or the population mean μ as appropriate.

The three displays, show output for the hybrid rose data of Guided Exercise 3.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus Keypad) Display Press STAT \blacktriangleright CALC \blacktriangleright 1:1-Var Stats. S_x is the sample standard deviation. σ_x is the population standard deviation.

1-Var Stats	
\bar{x}	=6
Σx	=36
Σx^2	=224
S_x	=1.264911064
σ_x	=1.154700538
n	=6

Excel 2007 Display On the Home ribbon, click the Data Tab. Then in the Analysis Group click Data Analysis. Select Descriptive Statistics. In the dialogue box, fill in the input range with the column and row numbers containing your data. Then check the Summary Statistics box.

Column 1	
Mean	6
Standard Error	0.516398
Median	5.5
Mode	5
Standard Deviation	1.264911
Sample Variance	1.6
Kurtosis	-0.78125
Skewness	0.889391
Range	3
Minimum	5
Maximum	8
Sum	36
Count	6

Minitab Display Menu choices: Stat \blacktriangleright Basic Statistics \blacktriangleright Display Descriptive Statistics. StDev is the sample standard deviation. TrMean is a 5% trimmed mean.

N	Mean	Median	TrMean	StDev	SEMean
6	6.000	5.500	6.000	1.265	0.516
Minimum	5.000	5.000	5.000	7.250	
Maximum	8.000	8.000	8.000		
Q1					
Q3					



Now let's look at two immediate applications of the standard deviation. The first is the coefficient of variation, and the second is Chebyshev's theorem.

Coefficient of variation, CV

A disadvantage of the standard deviation as a comparative measure of variation is that it depends on the units of measurement. This means that it is difficult to use the standard deviation to compare measurements from different populations. For this reason, statisticians have defined the *coefficient of variation*, which expresses the standard deviation as a percentage of the sample or population mean.

If \bar{x} and s represent the sample mean and sample standard deviation, respectively, then the sample **coefficient of variation** CV is defined to be

$$CV = \frac{s}{\bar{x}} \cdot 100$$

If μ and σ represent the population mean and population standard deviation, respectively, then the population coefficient of variation CV is defined to be

$$CV = \frac{\sigma}{\mu} \cdot 100$$

Notice that the numerator and denominator in the definition of CV have the same units, so CV itself has no units of measurement. This gives us the advantage of being able to directly compare the variability of two different populations using the coefficient of variation.

In the next example and guided exercise, we will compute the CV of a population and of a sample and then compare the results.

EXAMPLE 7

COEFFICIENT OF VARIATION

The Trading Post on Grand Mesa is a small, family-run store in a remote part of Colorado. The Grand Mesa region contains many good fishing lakes, so the Trading Post sells spinners (a type of fishing lure). The store has a very limited selection of spinners. In fact, the Trading Post has only eight different types of spinners for sale. The prices (in dollars) are

2.10 1.95 2.60 2.00 1.85 2.25 2.15 2.25

Since the Trading Post has only eight different kinds of spinners for sale, we consider the eight data values to be the *population*.

- (a) Use a calculator with appropriate statistics keys to verify that for the Trading Post data, $\mu \approx \$2.14$ and $\sigma \approx \$0.22$.

SOLUTION: Since the computation formulas for \bar{x} and μ are identical, most calculators provide the value of \bar{x} only. Use the output of this key for μ . The computation formulas for the sample standard deviation s and the population standard deviation σ are slightly different. Be sure that you use the key for σ (sometimes designated as σ_n or σ_x).



Ron Watts/CORBIS

- (b) Compute the CV of prices for the Trading Post and comment on the meaning of the result.

SOLUTION:

$$CV = \frac{\sigma}{\mu} \times 100 = \frac{0.22}{2.14} \times 100 = 10.28\%$$

Interpretation The coefficient of variation can be thought of as a measure of the spread of the data relative to the average of the data. Since the Trading Post is very small, it carries a small selection of spinners that are all priced similarly. The CV tells us that the standard deviation of the spinner prices is only 10.28% of the mean.

GUIDED EXERCISE 4

Coefficient of variation

Cabela's in Sidney, Nebraska, is a very large outfitter that carries a broad selection of fishing tackle. It markets its products nationwide through a catalog service. A random sample of 10 spinners from Cabela's extensive spring catalog gave the following prices (in dollars):

1.69 1.49 3.09 1.79 1.39 2.89 1.49 1.39 1.49 1.99

- (a) Use a calculator with sample mean and sample standard deviation keys to compute \bar{x} and s .  $\bar{x} = \$1.87$ and $s \approx \$0.62$.
- (b) Compute the CV for the spinner prices at Cabela's.  $CV = \frac{s}{\bar{x}} \times 100 = \frac{0.62}{1.87} \times 100 = 33.16\%$
- (c) **Interpretation** Compare the mean, standard deviation, and CV for the spinner prices at the Grand Mesa Trading Post (Example 7) and Cabela's. Comment on the differences.  The CV for Cabela's is more than three times the CV for the Trading Post. Why? First, because of the remote location, the Trading Post tends to have somewhat higher prices (larger μ). Second, the Trading Post is very small, so it has a rather limited selection of spinners with a smaller variation in price.

Chebyshev's Theorem

From our earlier discussion about standard deviation, we recall that the spread or dispersion of a set of data about the mean will be small if the standard deviation is small, and it will be large if the standard deviation is large. If we are dealing with a symmetrical, bell-shaped distribution, then we can make very definite statements about the proportion of the data that must lie within a certain number of standard deviations on either side of the mean. This will be discussed in detail in Chapter 6 when we talk about normal distributions.

However, the concept of data spread about the mean can be expressed quite generally for *all data distributions* (skewed, symmetric, or other shapes) by using the remarkable theorem of Chebyshev.

Chebyshev's theorem

Chebyshev's Theorem

For *any* set of data (either population or sample) and for any constant k greater than 1, the proportion of the data that must lie within k standard deviations on either side of the mean is *at least*

$$1 - \frac{1}{k^2}$$

Results of Chebyshev's Theorem

For *any* set of data:

- at least 75% of the data fall in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.
- at least 88.9% of the data fall in the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.
- at least 93.8% of the data fall in the interval from $\mu - 4\sigma$ to $\mu + 4\sigma$.

The results of Chebyshev's theorem can be derived by using the theorem and a little arithmetic. For instance, if we create an interval $k = 2$ standard deviations on either side of the mean, Chebyshev's theorem tells us that

$$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} \text{ or } 75\%$$

is the minimum percentage of data in the $\mu - 2\sigma$ to $\mu + 2\mu$ interval.

Notice that Chebyshev's theorem refers to the *minimum* percentage of data that must fall within the specified number of standard deviations of the mean. If the distribution is mound-shaped, an even *greater* percentage of data will fall into the specified intervals (see the Empirical Rule in Section 6.1).

EXAMPLE 8

CHEBYSHEV'S THEOREM

Students Who Care is a student volunteer program in which college students donate work time to various community projects such as planting trees. Professor Gill is the faculty sponsor for this student volunteer program. For several years, Dr. Gill has kept a careful record of x = total number of work hours volunteered by a student in the program each semester. For a random sample of students in the program, the mean number of hours was $\bar{x} = 29.1$ hours each semester, with a standard deviation of $s = 1.7$ hours each semester. Find an interval A to B for the number of hours volunteered into which at least 75% of the students in this program would fit.

SOLUTION: According to results of Chebyshev's theorem, at least 75% of the data must fall within 2 standard deviations of the mean. Because the mean is $\bar{x} = 29.1$ and the standard deviation is $s = 1.7$, the interval is

$$\begin{aligned} & \bar{x} - 2s \text{ to } \bar{x} + 2s \\ & 29.1 - 2(1.7) \text{ to } 29.1 + 2(1.7) \\ & 25.7 \text{ to } 32.5 \end{aligned}$$

At least 75% of the students would fit into the group that volunteered from 25.7 to 32.5 hours each semester.

Bob Daemmrich/The Image Works



GUIDED EXERCISE 5

Chebyshev interval

The *East Coast Independent News* periodically runs ads in its classified section offering a month's-free subscription to those who respond. In this way, management can get a sense about the number of subscribers who read the classified section each day. Over a period of 2 years, careful records have been kept. The mean number of responses per ad is $\bar{x} = 525$ with standard deviation $s = 30$.

Continued

GUIDED EXERCISE 5 *continued*

Determine a Chebyshev interval about the mean in which at least 88.9% of the data fall.



By Chebyshev's theorem, at least 88.9% of the data fall into the interval

$$\bar{x} - 3s \text{ to } \bar{x} + 3s$$

Because $\bar{x} = 525$ and $s = 30$, the interval is

$$525 - 3(30) \text{ to } 525 + 3(30)$$

or from 435 to 615 responses per ad.

LOOKING FORWARD

In Chapter 6 we will study normal distributions. For normal distributions we can make much more precise statements about the percentage of data falling within 1, 2, or 3 standard deviations of the mean. In fact, the empirical rule (see Section 6.1) tells us that for normal distributions, about 68% of the data fall within 1 standard deviation of the mean, about 95% fall within 2 standard deviations of the mean, and almost all (99.7%) fall within 3 standard deviations of the mean.

CRITICAL THINKING

Averages such as the mean are often referred to in the media. However, an average by itself does not tell much about the way data are distributed about the mean. Knowledge about the standard deviation or variance, along with the mean, gives a much better picture of the data distribution.

Chebyshev's theorem tells us that no matter what the data distribution looks like, at least 75% of the data will fall within 2 standard deviations of the mean. As we will see in Chapter 6, when the distribution is mound-shaped and symmetrical, about 95% of the data are within 2 standard deviations of the mean. Data values beyond 2 standard deviations from the mean are less common than those closer to the mean.

In fact, one indicator that a data value might be an *outlier* is that it is more than 2.5 standard deviations from the mean (Source: *Oxford Dictionary of Statistics*, Oxford University Press).

Outlier

VIEWPOINT

Socially Responsible Investing

Make a difference and make money! Socially responsible mutual funds tend to screen out corporations that sell tobacco, weapons, and alcohol, as well as companies that are environmentally unfriendly. In addition, these funds screen out companies that use child labor in sweatshops. There are 68 socially responsible funds tracked by the Social Investment Forum. For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to social investing.

How do these funds rate compared to other funds? One way to answer this question is to study the annual percent returns of the funds using both the mean and standard deviation. (See Problem 20 of this section.)

SECTION 3.2 PROBLEMS

1. **Statistical Literacy** Which average—mean, median, or mode—is associated with the standard deviation?
2. **Statistical Literacy** What is the relationship between the variance and the standard deviation for a sample data set?
3. **Statistical Literacy** When computing the standard deviation, does it matter whether the data are sample data or data comprising the entire population? Explain.
4. **Statistical Literacy** What symbol is used for the standard deviation when it is a sample statistic? What symbol is used for the standard deviation when it is a population parameter?
5. **Basic Computation: Range, Standard Deviation** Consider the data set

$$\begin{array}{ccccc} 2 & 3 & 4 & 5 & 6 \end{array}$$
 - (a) Find the range.
 - (b) Use the defining formula to compute the sample standard deviation s .
 - (c) Use the defining formula to compute the population standard deviation σ .
6. **Basic Computation: Range, Standard Deviation** Consider the data set

$$\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array}$$
 - (a) Find the range.
 - (b) Use the defining formula to compute the sample standard deviation s .
 - (c) Use the defining formula to compute the population standard deviation σ .
7. **Critical Thinking** For a given data set in which not all data values are equal, which value is smaller, s or σ ? Explain.
8. **Critical Thinking** Consider two data sets with equal sample standard deviations. The first data set has 20 data values that are not all equal, and the second has 50 data values that are not all equal. For which data set is the difference between s and σ greater? Explain. Hint: Consider the relationship $\sigma = s\sqrt{(n - 1)/n}$.
9. **Critical Thinking** Each of the following data sets has a mean of $\bar{x} = 10$.

(i) 8 9 10 11 12	(ii) 7 9 10 11 13	(iii) 7 8 10 12 13
------------------	-------------------	--------------------

 - (a) Without doing any computations, order the data sets according to increasing value of standard deviations.
 - (b) Why do you expect the difference in standard deviations between data sets (i) and (ii) to be greater than the difference in standard deviations between data sets (ii) and (iii)? Hint: Consider how much the data in the respective sets differ from the mean.
10. **Critical Thinking: Data Transformation** In this problem, we explore the effect on the standard deviation of adding the same constant to each data value in a data set. Consider the data set 5, 9, 10, 11, 15.
 - (a) Use the defining formula, the computation formula, or a calculator to compute s .
 - (b) Add 5 to each data value to get the new data set 10, 14, 15, 16, 20. Compute s .
 - (c) Compare the results of parts (a) and (b). In general, how do you think the standard deviation of a data set changes if the same constant is added to each data value?
11. **Critical Thinking: Data Transformation** In this problem, we explore the effect on the standard deviation of multiplying each data value in a data set by the same constant. Consider the data set 5, 9, 10, 11, 15.

- (a) Use the defining formula, the computation formula, or a calculator to compute s .
- (b) Multiply each data value by 5 to obtain the new data set 25, 45, 50, 55, 75. Compute s .
- (c) Compare the results of parts (a) and (b). In general, how does the standard deviation change if each data value is multiplied by a constant c ?
- (d) You recorded the weekly distances you bicycled in miles and computed the standard deviation to be $s = 3.1$ miles. Your friend wants to know the standard deviation in kilometers. Do you need to redo all the calculations? Given 1 mile ≈ 1.6 kilometers, what is the standard deviation in kilometers?
12. **Critical Thinking: Outliers** One indicator of an outlier is that an observation is more than 2.5 standard deviations from the mean. Consider the data value 80.
- (a) If a data set has mean 70 and standard deviation 5, is 80 a suspect outlier?
- (b) If a data set has mean 70 and standard deviation 3, is 80 a suspect outlier?
13. **Basic Computation: Variance, Standard Deviation** Given the sample data
- | | | | | | |
|------|----|----|----|----|----|
| $x:$ | 23 | 17 | 15 | 30 | 25 |
|------|----|----|----|----|----|
- (a) Find the range.
- (b) Verify that $\Sigma x = 110$ and $\Sigma x^2 = 2568$.
- (c) Use the results of part (b) and appropriate computation formulas to compute the sample variance s^2 and sample standard deviation s .
- (d) Use the defining formulas to compute the sample variance s^2 and sample standard deviation s .
- (e) Suppose the given data comprise the entire population of all x values. Compute the population variance σ^2 and population standard deviation σ .
14. **Basic Computation: Coefficient of Variation, Chebyshev Interval** Consider sample data with $\bar{x} = 15$ and $s = 3$.
- (a) Compute the coefficient of variation.
- (b) Compute a 75% Chebyshev interval around the sample mean.
15. **Basic Computation: Coefficient of Variation, Chebyshev Interval** Consider population data with $\mu = 20$ and $\sigma = 2$.
- (a) Compute the coefficient of variation.
- (b) Compute an 88.9% Chebyshev interval around the population mean.
16. **Investing: Stocks and Bonds** Do bonds reduce the overall risk of an investment portfolio? Let x be a random variable representing annual percent return for Vanguard Total Stock Index (all stocks). Let y be a random variable representing annual return for Vanguard Balanced Index (60% stock and 40% bond). For the past several years, we have the following data (Reference: Morningstar Research Group, Chicago).
- | | | | | | | | | | | |
|------|----|----|----|----|----|----|----|-----|-----|-----|
| $x:$ | 11 | 0 | 36 | 21 | 31 | 23 | 24 | -11 | -11 | -21 |
| $y:$ | 10 | -2 | 29 | 14 | 22 | 18 | 14 | -2 | -3 | -10 |
- (a) Compute Σx , Σx^2 , Σy , and Σy^2 .
- (b) Use the results of part (a) to compute the sample mean, variance, and standard deviation for x and for y .
- (c) Compute a 75% Chebyshev interval around the mean for x values and also for y values. Use the intervals to compare the two funds.
- (d) **Interpretation:** Compute the coefficient of variation for each fund. Use the coefficients of variation to compare the two funds. If s represents risks and \bar{x} represents expected return, then s/\bar{x} can be thought of as a measure of risk per unit of expected return. In this case, why is a smaller CV better? Explain.



- 17. Space Shuttle: Epoxy** Kevlar epoxy is a material used on the NASA space shuttles. Strands of this epoxy were tested at the 90% breaking strength. The following data represent time to failure (in hours) for a random sample of 50 epoxy strands (Reference: R. E. Barlow, University of California, Berkeley). Let x be a random variable representing time to failure (in hours) at 90% breaking strength. **Note:** These data are also available for download at the Online Study Center.

0.54	1.80	1.52	2.05	1.03	1.18	0.80	1.33	1.29	1.11
3.34	1.54	0.08	0.12	0.60	0.72	0.92	1.05	1.43	3.03
1.81	2.17	0.63	0.56	0.03	0.09	0.18	0.34	1.51	1.45
1.52	0.19	1.55	0.02	0.07	0.65	0.40	0.24	1.51	1.45
1.60	1.80	4.69	0.08	7.89	1.58	1.64	0.03	0.23	0.72

- (a) Find the range.
- (b) Use a calculator to verify that $\Sigma x = 62.11$ and $\Sigma x^2 \approx 164.23$.
- (c) Use the results of part (b) to compute the sample mean, variance, and standard deviation for the time to failure.
- (d) **Interpretation** Use the results of part (c) to compute the coefficient of variation. What does this number say about time to failure? Why does a small CV indicate more consistent data, whereas a larger CV indicates less consistent data? Explain.

- 18. Archaeology: Ireland** The Hill of Tara in Ireland is a place of great archaeological importance. This region has been occupied by people for more than 4000 years. Geomagnetic surveys detect subsurface anomalies in the earth's magnetic field. These surveys have led to many significant archaeological discoveries. After collecting data, the next step is to begin a statistical study. The following data measure magnetic susceptibility (centimeter-gram-second $\times 10^{-6}$) on two of the main grids of the Hill of Tara (Reference: *Tara: An Archaeological Survey* by Conor Newman, Royal Irish Academy, Dublin).

Grid E: x variable

13.20	5.60	19.80	15.05	21.40	17.25	27.45
16.95	23.90	32.40	40.75	5.10	17.75	28.35

Grid H: y variable

11.85	15.25	21.30	17.30	27.50	10.35	14.90
48.70	25.40	25.95	57.60	34.35	38.80	41.00
31.25						

- (a) Compute Σx , Σx^2 , Σy , and Σy^2 .
- (b) Use the results of part (a) to compute the sample mean, variance, and standard deviation for x and for y .
- (c) Compute a 75% Chebyshev interval around the mean for x values and also for y values. Use the intervals to compare the magnetic susceptibility on the two grids. Higher numbers indicate higher magnetic susceptibility. However, extreme values, high or low, could mean an anomaly and possible archaeological treasure.
- (d) **Interpretation** Compute the sample coefficient of variation for each grid. Use the CV's to compare the two grids. If s represents variability in the signal (magnetic susceptibility) and \bar{x} represents the expected level of the signal, then s/\bar{x} can be thought of as a measure of the variability per unit of expected signal. Remember, a considerable variability in the signal (above or below average) might indicate buried artifacts. Why, in this case, would a large CV be better, or at least more exciting? Explain.

19. **Wildlife: Mallard Ducks and Canada Geese** For mallard ducks and Canada geese, what percentage of nests are successful (at least one offspring survives)? Studies in Montana, Illinois, Wyoming, Utah, and California gave the following percentages of successful nests (Reference: *The Wildlife Society Press*, Washington, D.C.).
- x:* Percentage success for mallard duck nests
- | | | | | |
|----|----|----|----|----|
| 56 | 85 | 52 | 13 | 39 |
|----|----|----|----|----|
- y:* Percentage success for Canada goose nests
- | | | | | |
|----|----|----|----|----|
| 24 | 53 | 60 | 69 | 18 |
|----|----|----|----|----|
- (a) Use a calculator to verify that $\Sigma x = 245$; $\Sigma x^2 = 14,755$; $\Sigma y = 224$; and $\Sigma y^2 = 12,070$.
- (b) Use the results of part (a) to compute the sample mean, variance, and standard deviation for *x*, the percent of successful mallard nests.
- (c) Use the results of part (a) to compute the sample mean, variance, and standard deviation for *y*, the percent of successful Canada goose nests.
- (d) **Interpretation** Use the results of parts (b) and (c) to compute the coefficient of variation for successful mallard nests and Canada goose nests. Write a brief explanation of the meaning of these numbers. What do these results say about the nesting success rates for mallards compared to those of Canada geese? Would you say one group of data is more or less consistent than the other? Explain.
20. **Investing: Socially Responsible Mutual Funds** Pax World Balanced is a highly respected, socially responsible mutual fund of stocks and bonds (see Viewpoint). Vanguard Balanced Index is another highly regarded fund that represents the entire U.S. stock and bond market (an index fund). The mean and standard deviation of annualized percent returns are shown below. The annualized mean and standard deviation are based on the years 1993 through 2002 (Source: Morningstar).
- Pax World Balanced: $\bar{x} = 9.58\%$; $s = 14.05\%$
- Vanguard Balanced Index: $\bar{x} = 9.02\%$; $s = 12.50\%$
- (a) **Interpretation** Compute the coefficient of variation for each fund. If \bar{x} represents return and s represents risk, then explain why the coefficient of variation can be taken to represent risk per unit of return. From this point of view, which fund appears to be better? Explain.
- (b) **Interpretation** Compute a 75% Chebyshev interval around the mean for each fund. Use the intervals to compare the two funds. As usual, past performance does not guarantee future performance.
21. **Medical: Physician Visits** In some reports, the mean and coefficient of variation are given. For instance, in *Statistical Abstract of the United States*, 116th Edition, one report gives the average number of physician visits by males per year. The average reported is 2.2, and the reported coefficient of variation is 1.5%. Use this information to determine the standard deviation of the annual number of visits to physicians made by males.

Grouped data



Expand Your Knowledge: Grouped Data

When data are grouped, such as in a frequency table or histogram, we can estimate the mean and standard deviation by using the following formulas. Notice that all data values in a given class are treated as though each of them equals the midpoint *x* of the class.

Continued

Approximating \bar{x} and s from grouped data

Sample Mean for a Frequency Distribution

$$\bar{x} = \frac{\sum xf}{n} \quad (5)$$

Sample Standard Deviation for a Frequency Distribution

$$s = \sqrt{\frac{\sum(x - \bar{x})^2 f}{n - 1}} \quad (6)$$

Computation Formula for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum x^2 f - (\sum xf)^2 / n}{n - 1}} \quad (7)$$

where

x is the midpoint of a class,

f is the number of entries in that class,

n is the total number of entries in the distribution, and $n = \sum f$.

The summation Σ is over all classes in the distribution.

Use formulas (5) and (6) or (5) and (7) to solve Problems 22–25. To use formulas (5) and (6) to evaluate the sample mean and standard deviation, use the following column heads:

Midpoint x	Frequency f	xf	$(x - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
--------------	---------------	------	-----------------	-------------------	---------------------

For formulas (5) and (7), use these column heads:

Midpoint x	Frequency f	xf	x^2	$x^2 f$
--------------	---------------	------	-------	---------

Note: On the TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) calculators, enter the midpoints in column L₁ and the frequencies in column L₂. Then use 1-VarStats L₁, L₂.



22. **Grouped Data: Anthropology** What was the age distribution of prehistoric Native Americans? Extensive anthropologic studies in the southwestern United States gave the following information about a prehistoric extended family group of 80 members on what is now the Navajo Reservation in northwestern New Mexico (Source: Based on information taken from *Prehistory in the Navajo Reservation District*, by F. W. Eddy, Museum of New Mexico Press).

Age range (years)	1–10*	11–20	21–30	31 and over
Number of individuals	34	18	17	11

*Includes infants.

For this community, estimate the mean age expressed in years, the sample variance, and the sample standard deviation. For the class 31 and over, use 35.5 as the class midpoint.



23. **Grouped Data: Shoplifting** What is the age distribution of adult shoplifters (21 years of age or older) in supermarkets? The following is based on information taken from the National Retail Federation. A random sample of 895 incidents of shoplifting gave the following age distribution:

Age range (years)	21–30	31–40	41 and over
Number of shoplifters	260	348	287

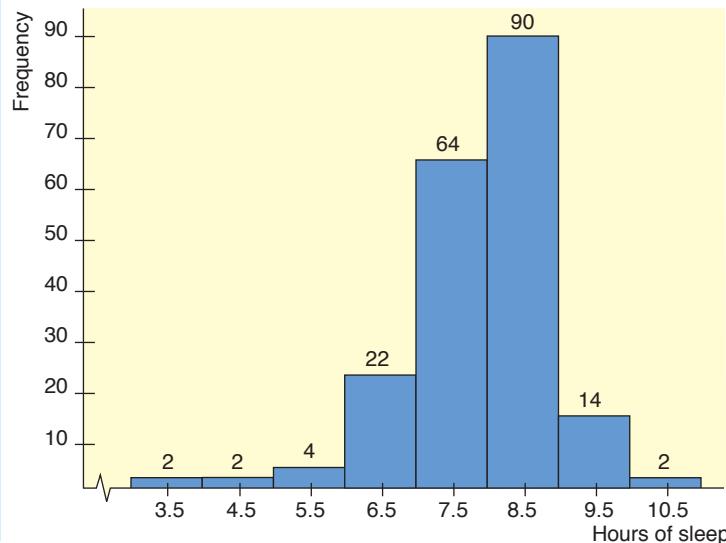
Estimate the mean age, sample variance, and sample standard deviation for the shoplifters. For the class 41 and over, use 45.5 as the class midpoint.



24. **Grouped Data: Hours of Sleep per Day** Alexander Borbely is a professor at the University of Zurich Medical School, where he is director of the sleep laboratory. The histogram in Figure 3-2 is based on information from his book *Secrets of Sleep*. The histogram displays hours of sleep per day for a random sample of 200 subjects. Estimate the mean hours of sleep, standard deviation of hours of sleep, and coefficient of variation.

FIGURE 3-2

Hours of Sleep Each Day (24-hour period)



25. **Grouped Data: Business Administration** What are the big corporations doing with their wealth? One way to answer this question is to examine profits as percentage of assets. A random sample of 50 Fortune 500 companies gave the following information (Source: Based on information from *Fortune 500*, Vol. 135, No. 8).

Profit as percentage of assets	8.6–12.5	12.6–16.5	16.6–20.5	20.6–24.5	24.6–28.5
Number of companies	15	20	5	7	3



- Estimate the sample mean, sample variance, and sample standard deviation for profit as percentage of assets.
26. **Expand Your Knowledge: Moving Averages** You do not need a lot of money to invest in a mutual fund. However, if you decide to put some money into an investment, you are usually advised to leave it in for (at least) several years. Why? Because good years tend to cancel out bad years, giving you a better overall return with less risk. To see what we mean, let's use a 3-year *moving average* on the Calvert Social Balanced Fund (a socially responsible fund).

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
% Return	1.78	17.79	7.46	5.95	-4.74	25.85	9.03	18.92	17.49	6.80	-2.38

Source: Morningstar

- (a) Use a calculator with mean and standard deviation keys to verify that the mean annual return for all 11 years is approximately 9.45%, with standard deviation 9.57%.

- (b) To compute a 3-year moving average for 1992, we take the data values for 1992 and the prior two years and average them. To compute a 3-year moving average for 1993, we take the data values for 1993 and the prior two years and average them. Verify that the following 3-year moving averages are correct.

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000
3-year moving average	9.01	10.40	2.89	9.02	10.05	17.93	15.15	14.40	7.30

- (c) Use a calculator with mean and standard deviation keys to verify that for the 3-year moving average, the mean is 10.68% with sample standard deviation 4.53%.
- (d) *Interpretation* Compare the results of parts (a) and (c). Suppose we take the point of view that risk is measured by standard deviation. Is the risk (standard deviation) of the 3-year moving average considerably smaller? This is an example of a general phenomenon that will be studied in more detail in Chapter 6.

27. **Brain Teaser: Sum of Squares** If you like mathematical puzzles or love algebra, try this! Otherwise, just trust that the computational formula for the sum of squares is correct. We have a sample of x values. The sample size is n . Fill in the details for the following steps.

$$\begin{aligned}\Sigma(x - \bar{x})^2 &= \Sigma x^2 - 2\bar{x}\Sigma x + n\bar{x}^2 \\ &= \Sigma x^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \Sigma x^2 - \frac{(\Sigma x)^2}{n}\end{aligned}$$

SECTION 3.3

Percentiles and Box-and-Whisker Plots

FOCUS POINTS

- Interpret the meaning of percentile scores.
- Compute the median, quartiles, and five-number summary from raw data.
- Make a box-and-whisker plot. Interpret the results.
- Describe how a box-and-whisker plot indicates spread of data about the median.

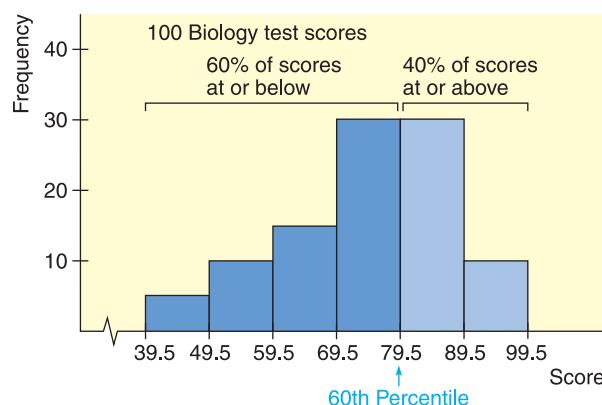
We've seen measures of central tendency and spread for a set of data. The arithmetic mean \bar{x} and the standard deviation s will be very useful in later work. However, because they each utilize every data value, they can be heavily influenced by one or two extreme data values. In cases where our data distributions are heavily skewed or even bimodal, we often get a better summary of the distribution by utilizing relative position of data rather than exact values.

Recall that the median is an average computed by using relative position of the data. If we are told that 81 is the median score on a biology test, we know that after the data have been ordered, 50% of the data fall at or below the median value of 81. The median is an example of a *percentile*; in fact, it is the 50th percentile. The general definition of the P th percentile follows.

For whole numbers P (where $1 \leq P \leq 99$), the P th percentile of a distribution is a value such that $P\%$ of the data fall at or below it and $(100 - P)\%$ of the data fall at or above it.

Percentile

FIGURE 3-3
A Histogram with the 60th Percentile Shown



In Figure 3-3, we see the 60th percentile marked on a histogram. We see that 60% of the data lie below the mark and 40% lie above it.

GUIDED EXERCISE 6

Percentiles

You took the English achievement test to obtain college credit in freshman English by examination.

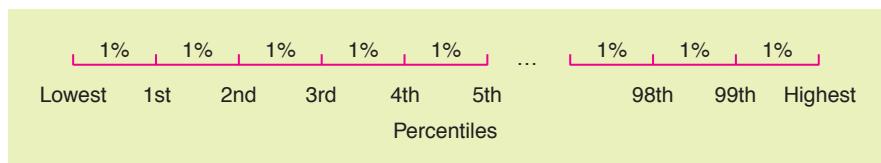
- (a) If your score is at the 89th percentile, what percentage of scores are at or below yours?
- (b) If the scores ranged from 1 to 100 and your raw score is 95, does this necessarily mean that your score is at the 95th percentile?

- The percentile means that 89% of the scores are at or below yours.
- No, the percentile gives an indication of relative position of the scores. The determination of your percentile has to do with the number of scores at or below yours. If everyone did very well and only 80% of the scores fell at or below yours, you would be at the 80th percentile even though you got 95 out of 100 points on the exam.

There are 99 percentiles, and in an ideal situation, the 99 percentiles divide the data set into 100 equal parts. (See Figure 3-4.) However, if the number of data elements is not exactly divisible by 100, the percentiles will not divide the data into equal parts.

FIGURE 3-4

Percentiles



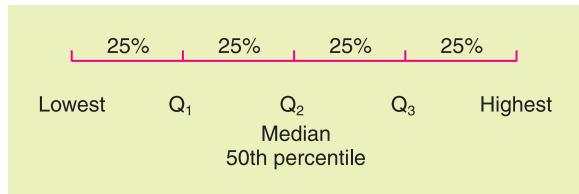
There are several widely used conventions for finding percentiles. They lead to slightly different values for different situations, but these values are close together. For all conventions, the data are first *ranked* or ordered from smallest to

Quartiles

largest. A natural way to find the P th percentile is to then find a value such that $P\%$ of the data fall at or below it. This will not always be possible, so we take the nearest value satisfying the criterion. It is at this point that there is a variety of processes to determine the exact value of the percentile.

We will not be very concerned about exact procedures for evaluating percentiles in general. However, *quartiles* are special percentiles used so frequently that we want to adopt a specific procedure for their computation.

Quartiles are those percentiles that divide the data into fourths. The *first quartile* Q_1 is the 25th percentile, the *second quartile* Q_2 is the median, and the *third quartile* Q_3 is the 75th percentile. (See Figure 3-5.)

FIGURE 3-5**Quartiles**

Again, several conventions are used for computing quartiles, but the following convention utilizes the median and is widely adopted.

PROCEDURE**HOW TO COMPUTE QUARTILES**

1. Order the data from smallest to largest.
2. Find the median. This is the second quartile.
3. The first quartile Q_1 is then the median of the lower half of the data; that is, it is the median of the data falling *below* the Q_2 position (and not including Q_2).
4. The third quartile Q_3 is the median of the upper half of the data; that is, it is the median of the data falling *above* the Q_2 position (and not including Q_2).

Interquartile range

In short, all we do to find the quartiles is find three medians.

The median, or second quartile, is a popular measure of the center utilizing relative position. A useful measure of data spread utilizing relative position is the *interquartile range* (*IQR*). It is simply the difference between the third and first quartiles.

$$\text{Interquartile range} = Q_3 - Q_1$$

The interquartile range tells us the spread of the middle half of the data. Now let's look at an example to see how to compute all of these quantities.

**EXAMPLE 9****QUARTILES**

In a hurry? On the run? Hungry as well? How about an ice cream bar as a snack? Ice cream bars are popular among all age groups. *Consumer Reports* did a study of ice cream bars. Twenty-seven bars with taste ratings of at least "fair" were listed, and cost per bar was included in the report. Just how much does an ice cream bar cost? The data, expressed in dollars, appear in Table 3-4. As you can see, the cost varies quite a bit, partly because the bars are not of uniform size.

TABLE 3-4 Cost of Ice Cream Bars (in dollars)

0.99	1.07	1.00	0.50	0.37	1.03	1.07	1.07
0.97	0.63	0.33	0.50	0.97	1.08	0.47	0.84
1.23	0.25	0.50	0.40	0.33	0.35	0.17	0.38
0.20	0.18	0.16					

TABLE 3-5 Ordered Cost of Ice Cream Bars (in dollars)

0.16	0.17	0.18	0.20	0.25	0.33	0.33	0.35
0.37	0.38	0.40	0.47	0.50	0.50	0.50	0.63
0.84	0.97	0.97	0.99	1.00	1.03	1.07	1.07
1.07	1.08	1.23					

(a) Find the quartiles.

SOLUTION: We first order the data from smallest to largest. Table 3-5 shows the data in order. Next, we find the median. Since the number of data values is 27, there are an odd number of data, and the median is simply the center or 14th value. The value is shown boxed in Table 3-5.

$$\text{Median} = Q_2 = 0.50$$

There are 13 values below the median position, and Q_1 is the median of these values. It is the middle or seventh value and is shaded in Table 3-5.

$$\text{First quartile} = Q_1 = 0.33$$

There are also 13 values above the median position. The median of these is the seventh value from the right end. This value is also shaded in Table 3-5.

$$\text{Third quartile} = Q_3 = 1.00$$

(b) Find the interquartile range.

SOLUTION:

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 1.00 - 0.33 \\ &= 0.67 \end{aligned}$$

This means that the middle half of the data has a cost spread of 67¢.

Gary Conner/PhotoEdit

GUIDED EXERCISE 7

Quartiles

Many people consider the number of calories in an ice cream bar as important as, if not more important than, the cost. The *Consumer Reports* article also included the calorie count of the rated ice cream bars (Table 3-6). There were 22 vanilla-flavored bars rated. Again, the bars varied in size, and some of the smaller bars had fewer calories. The calorie counts for the vanilla bars follow.

(a) Our first step is to order the data. Do so.

Table 3-6 Calories in Vanilla-Flavored Ice Cream Bars

342	377	319	353	295
234	294	286	377	182
310	439	111	201	182
197	209	147	190	151
131	151			

Table 3-7 Ordered Data

111	131	147	151	151	182
182	190	197	201	209	234
286	294	295	310	319	342
353	377	377	439		

Continued

GUIDED EXERCISE 1 *continued*

(b) There are 22 data values. Find the median.

→ Average the 11th and 12th data values boxed together in Table 3-7.

$$\begin{aligned}\text{Median} &= \frac{209 + 234}{2} \\ &= 221.5\end{aligned}$$

(c) How many values are below the median position? Find Q_1 .

→ Since the median lies halfway between the 11th and 12th values, there are 11 values below the median position. Q_1 is the median of these values.

$$Q_1 = 182$$

(d) There are the same number of data above as below the median. Use this fact to find Q_3 .

→ Q_3 is the median of the upper half of the data. There are 11 values in the upper portion.

$$Q_3 = 319$$

(e) Find the interquartile range and comment on its meaning.

$$\begin{aligned}IQR &= Q_3 - Q_1 \\ &= 319 - 182 \\ &= 137\end{aligned}$$

The middle portion of the data has a spread of 137 calories.

Box-and-Whisker Plots

Five-number summary

The quartiles together with the low and high data values give us a very useful *five-number summary* of the data and their spread.

Five-Number Summary

Lowest value, Q_1 , median, Q_3 , highest value

Box-and-whisker plot

We will use these five numbers to create a graphic sketch of the data called a *box-and-whisker plot*. Box-and-whisker plots provide another useful technique from exploratory data analysis (EDA) for describing data.

PROCEDURE

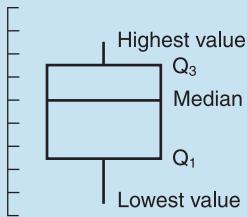
HOW TO MAKE A BOX-AND-WHISKER PLOT

1. Draw a vertical scale to include the lowest and highest data values.
2. To the right of the scale, draw a box from Q_1 to Q_3 .
3. Include a solid line through the box at the median level.
4. Draw vertical lines, called *whiskers*, from Q_1 to the lowest value and from Q_3 to the highest value.

Whiskers

FIGURE 3-6

Box-and-Whisker Plot



The next example demonstrates the process of making a box-and-whisker plot.

EXAMPLE 10 BOX-AND-WHISKER PLOT

Using the data from Guided Exercise 7, make a box-and-whisker plot showing the calories in vanilla-flavored ice cream bars. Use the plot to make observations about the distribution of calories.

- (a) In Guided Exercise 7, we ordered the data (see Table 3-7) and found the values of the median, Q_1 , and Q_3 . From this previous work we have the following five-number summary:

low value = 111; Q_1 = 182; median = 221.5; Q_3 = 319; high value = 439

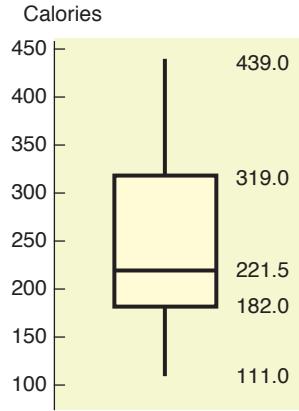
- (b) We select an appropriate vertical scale and make the plot (Figure 3-7).

- (c) **Interpretation** A quick glance at the box-and-whisker plot reveals the following:

- (i) The box tells us where the middle half of the data lies, so we see that half of the ice cream bars have between 182 and 319 calories, with an interquartile range of 137 calories.
- (ii) The median is slightly closer to the lower part of the box. This means that the lower calorie counts are more concentrated. The calorie counts above the median are more spread out, indicating that the distribution is slightly skewed toward the higher values.
- (iii) The upper whisker is longer than the lower, which again emphasizes skewness toward the higher values.

FIGURE 3-7

Box-and-Whisker Plot for Calories in Vanilla-Flavored Ice Cream Bars



COMMENT In exploratory data analysis, *hinges* rather than quartiles are used to create the box. Hinges are computed in a manner similar to the method used to compute quartiles. However, in the case of an odd number of data values, include the median itself in both the lower and upper halves of the data (see *Applications, Basics, and Computing of Exploratory Data Analysis* by Paul Velleman and David Hoaglin, Duxbury Press). This has the effect of shrinking the box and moving the ends of the box slightly toward the median. For an even number of data, the quartiles as we computed them equal the hinges.

GUIDED EXERCISE 8**Box-and-whisker plot**

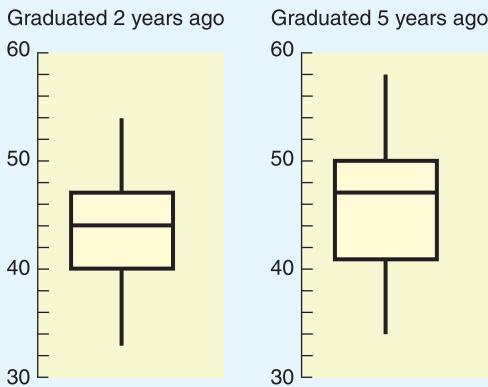
The Renata College Development Office sent salary surveys to alumni who graduated 2 and 5 years ago. The voluntary responses received are summarized in the box-and-whisker plots shown in Figure 3-8 on the next page.

- (a) From Figure 3-8, estimate the median and extreme values of salaries of alumni graduating 2 years ago. In what range are the middle half of the salaries?



The median seems to be about \$44,000. The extremes are about \$33,000 and \$54,000. The middle half of the salaries fall between \$40,000 and \$47,000.

Continued

GUIDED EXERCISE 1 *continued***FIGURE 3-8****Box-and-Whisker Plots for Alumni Salaries
(in thousands of dollars)**

- (b) From Figure 3-8, estimate the median and the extreme values of salaries of alumni graduating 5 years ago. What is the location of the middle half of the salaries?
- (c) **Interpretation** Compare the two box-and-whisker plots and make comments about the salaries of alumni graduating 2 and 5 years ago.

→ The median seems to be \$47,000. The extremes are \$34,000 and \$58,000. The middle half of the data is enclosed by the box with low side at \$41,000 and high side at \$50,000.

→ The salaries of the alumni graduating 5 years ago have a larger range. They begin slightly higher than and extend to levels about \$4000 above the salaries of those graduating 2 years ago. The middle half of the data is also more spread out, with higher boundaries and a higher median.

CRITICAL THINKING

 Problem 12 discusses how to identify possible outliers on a box-and-whisker plot.

Box-and-whisker plots provide a graphic display of the spread of data about the median. The box shows the location of the middle half of the data. One-quarter of the data is located along each whisker.

To the extent that the median is centered in the box and the whiskers are about the same length, the data distribution is symmetric around the median. If the median line is near one end of the box, the data are skewed toward the other end of the box.

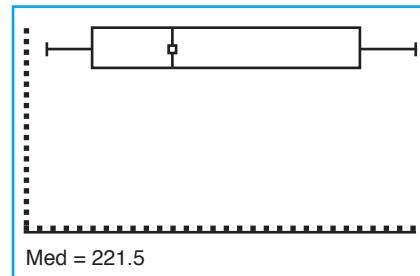
We have developed the skeletal box-and-whisker display. Other variations include *fences*, which are marks placed on either side of the box to represent various portions of data. Values that lie beyond the fences are *outliers*. Problem 12 of this section discusses some criteria for locating fences and identifying outliers.



TECH NOTES**Box-and-Whisker Plot**

Both Minitab and the TI-84Plus/TI-83Plus/TI-nspire calculators support box-and-whisker plots. On the TI-84Plus/TI-83Plus/TI-nspire, the quartiles Q_1 and Q_3 are calculated as we calculate them in this text. In Minitab and Excel 2007, they are calculated using a slightly different process.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus Keypad) Press STATPLOT > On. Highlight box plot. Use Trace and the arrow keys to display the values of the five-number summary. The display shows the plot for calories in ice cream bars.



Excel 2007 Does not produce box-and-whisker plot. However, each value of the five-number summary can be found. On the **Home** ribbon, click the **Insert Function f_x**. In the dialogue box, select **Statistical** as the category and scroll to **Quartile**. In the dialogue box, enter the data location and then enter the number of the value you want. For instance, enter 1 in the quartile box for the first quartile.

Minitab Press **Graph ▶ Boxplot**. In the dialogue box, set **Display to IQRange Box**.

VIEWPOINT

Is Shorter Higher?

Can you estimate a person's height from the pitch of his or her voice? Is a soprano shorter than an alto? Is a bass taller than a tenor? A statistical study of singers in the New York Choral Society provided information. For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to DASL, the Carnegie Mellon University Data and Story Library. From Data Subjects, select Music and then Singers. Methods of this chapter can be used with new methods we will learn in Chapters 8 and 9 to examine such questions from a statistical point of view.

SECTION 3.3 PROBLEMS

1. **Statistical Literacy** Angela took a general aptitude test and scored in the 82nd percentile for aptitude in accounting. What percentage of the scores were at or below her score? What percentage were above?
2. **Statistical Literacy** One standard for admission to Redfield College is that the student rank in the upper quartile of his or her graduating high school class. What is the minimal percentile rank of a successful applicant?
3. **Critical Thinking** The town of Butler, Nebraska, decided to give a teacher-competency exam and defined the passing scores to be those in the 70th percentile or higher. The raw test scores ranged from 0 to 100. Was a raw score of 82 necessarily a passing score? Explain.
4. **Critical Thinking** Clayton and Timothy took different sections of Introduction to Economics. Each section had a different final exam. Timothy scored 83 out of 100 and had a percentile rank in his class of 72. Clayton scored 85 out of 100 but his percentile rank in his class was 70. Who performed better with respect to the rest of the students in the class, Clayton or Timothy? Explain your answer.
5. **Basic Computation: Five-Number Summary, Interquartile Range** Consider the following ordered data:
2 5 5 6 7 7 8 9 10
(a) Find the low, Q_1 , median, Q_3 , high.
(b) Find the interquartile range.
(c) Make a box-and-whisker plot.

6. **Basic Computation: Five-Number Summary, Interquartile Range** Consider the following ordered data:

2 5 5 6 7 8 8 9 10 12

- Find the low, Q_1 , median, Q_3 high.
- Find the interquartile range.
- Make a box-and-whisker plot.

7. **Health Care: Nurses** At Center Hospital there is some concern about the high turnover of nurses. A survey was done to determine how long (in months) nurses had been in their current positions. The responses (in months) of 20 nurses were

23 2 5 14 25 36 27 42 12 8
7 23 29 26 28 11 20 31 8 36

Make a box-and-whisker plot of the data. Find the interquartile range.

8. **Health Care: Staff** Another survey was done at Center Hospital to determine how long (in months) clerical staff had been in their current positions. The responses (in months) of 20 clerical staff members were

25 22 7 24 26 31 18 14 17 20
31 42 6 25 22 3 29 32 15 72

- Make a box-and-whisker plot. Find the interquartile range.
- Compare this plot with the one in Problem 7. Discuss the locations of the medians, the locations of the middle halves of the data banks, and the distances from Q_1 and Q_3 to the extreme values.

9. **Sociology: College Graduates** What percentage of the general U.S. population have bachelor's degrees? The *Statistical Abstract of the United States*, 120th Edition, gives the percentage of bachelor's degrees by state. For convenience, the data are sorted in increasing order.

17 18 18 18 19 20 20 20 21 21
21 21 22 22 22 22 22 22 23 23
24 24 24 24 24 24 24 24 25 26
26 26 26 26 26 27 27 27 27 27
28 28 29 31 31 32 32 34 35 38

- Make a box-and-whisker plot and find the interquartile range.
- Illinois has a bachelor's degree percentage rate of about 26%. Into what quartile does this rate fall?

10. **Sociology: High School Dropouts** What percentage of the general U.S. population are high school dropouts? The *Statistical Abstract of the United States*, 120th Edition, gives the percentage of high school dropouts by state. For convenience, the data are sorted in increasing order.

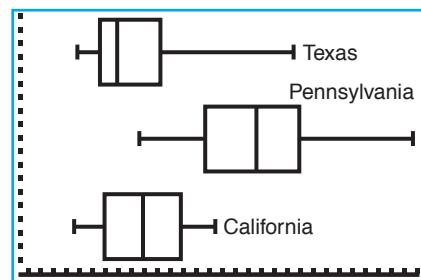
5 6 7 7 7 7 8 8 8 8
8 9 9 9 9 9 9 9 10 10
10 10 10 10 10 10 11 11 11 11
11 11 11 11 12 12 12 12 13 13
13 13 13 13 14 14 14 14 14 15

- Make a box-and-whisker plot and find the interquartile range.
- Wyoming has a dropout rate of about 7%. Into what quartile does this rate fall?

11. **Auto Insurance: Interpret Graphs** *Consumer Reports* rated automobile insurance companies and listed annual premiums for top-rated companies in several states. Figure 3-9 shows box-and-whisker plots for annual premiums for urban customers (married couple with one 17-year-old son) in three states.

The box-and-whisker plots in Figure 3-9 were all drawn using the same scale on a TI-84Plus/TI-83Plus/TI-nspire calculator.

FIGURE 3-9
Insurance Premium (annual, urban)



- Which state has the lowest premium? the highest?
- Which state has the highest median premium?
- Which state has the smallest range of premiums? the smallest interquartile range?
- Figure 3-10 gives the five-number summaries generated on the TI-84Plus/TI-83Plus/TI-nspire calculators for the box-and-whisker plots of Figure 3-9. Match the five-number summaries to the appropriate box-and-whisker plots.

FIGURE 3-10

Five-Number Summaries for Insurance Premiums

(a)

1-Var Stats
 $n=10$
 $\min X=2382$
 $Q_1=2758$
 $Med=2991$
 $Q_3=3652$
 $\max X=5715$

(b)

1-Var Stats
 $n=10$
 $\min X=3314$
 $Q_1=4326$
 $Med=5116.5$
 $Q_3=5801$
 $\max X=7527$

(c)

1-Var Stats
 $n=10$
 $\min X=2323$
 $Q_1=2801$
 $Med=3377.5$
 $Q_3=3966$
 $\max X=4482$



12. **Expand Your Knowledge: Outliers** Some data sets include values so high or so low that they seem to stand apart from the rest of the data. These data are called *outliers*. Outliers may represent data collection errors, data entry errors, or simply valid but unusual data values. It is important to identify outliers in the data set and examine the outliers carefully to determine if they are in error. One way to detect outliers is to use a box-and-whisker plot. Data values that fall beyond the limits

$$\text{Lower limit: } Q_1 - 1.5 \times (IQR)$$

$$\text{Upper limit: } Q_3 + 1.5 \times (IQR)$$

where IQR is the interquartile range, are suspected outliers. In the computer software package Minitab, values beyond these limits are plotted with asterisks (*).

Students from a statistics class were asked to record their heights in inches. The heights (as recorded) were

65	72	68	64	60	55	73	71	52	63	61	74
69	67	74	50	4	75	67	62	66	80	64	65

- Make a box-and-whisker plot of the data.
- Find the value of the interquartile range (IQR).
- Multiply the IQR by 1.5 and find the lower and upper limits.
- Are there any data values below the lower limit? above the upper limit? List any suspected outliers. What might be some explanations for the outliers?



SUMMARY

To characterize numerical data, we use both measures of center and of spread.

- Commonly used measures of center are the arithmetic mean, the median, and the mode. The weighted average and trimmed mean are also used as appropriate.
- Commonly used measures of spread are the variance, the standard deviation, and the range. The variance and standard deviation are measures of spread about the mean.
- Chebyshev's theorem enables us to estimate the data spread about the mean.

- The coefficient of variation lets us compare the relative spreads of different data sets.
- Other measures of data spread include percentiles, which indicate the percentage of data falling at or below the specified percentile value.
- Box-and-whisker plots show how the data are distributed about the median and the location of the middle half of the data distribution.

In later work, the average we will use most often is the mean; the measure of variation we will use most often is the standard deviation.

IMPORTANT WORDS & SYMBOLS

Section 3.1

- Average 82
Mode 82
Median 83
Mean 85
Summation symbol, Σ 85
Sample mean, \bar{x} 85
Population mean, μ 85
Resistant measure 86
Trimmed mean 86
Weighted average 88
Harmonic mean 93
Geometric mean 93

Section 3.2

- Range 93
Variance 94
Standard deviation 94
Sum of squares, $\sum(x - \bar{x})^2$ 95

- Sample variance, s^2 95
Sample standard deviation, s 95
Population mean, μ 98
Population variance, σ^2 98
Population standard deviation, σ 98
Population size, N page 98
Coefficient of variation, CV 100
Chebyshev's theorem 101
Outlier page 103
Mean of grouped data 107
Standard deviation of grouped data 108

Section 3.3

- Percentile 110
Quartiles 112
Interquartile range, IQR 112
Five-number summary 114
Box-and-whisker plot 114
Whisker 114

VIEWPOINT

The Fujita Scale

How do you measure a tornado? Professor Fujita and Allen Pearson (Director of the National Severe Storm Forecast Center) developed a measure based on wind speed and type of damage done by a tornado. The result is an excellent example of both descriptive and inferential statistical methods. For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the tornado project. Then look up Fujita scale. If we group the data a little, the scale becomes

FS	WS	%
F0 & F1	40–112	67
F2 & F3	113–206	29
F4 & F5	207–318	4

where FS represents Fujita scale; WS, wind speed in miles per hour; and %, percentage of all tornadoes.
Out of 100 tornadoes, what would you estimate for the mean and standard deviation of wind speed?

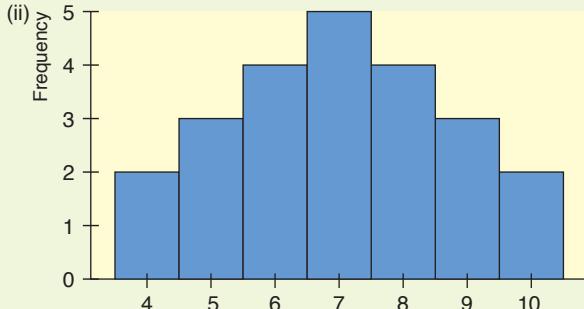
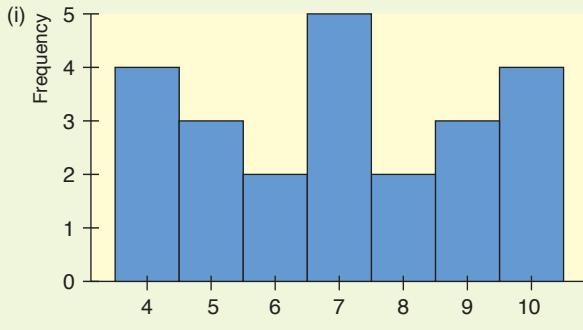
CHAPTER REVIEW PROBLEMS

1. Statistical Literacy

- (a) What measures of variation indicate spread about the mean?
- (b) Which graphic display shows the median and data spread about the median?

2. Critical Thinking

Look at the two histograms below. Each involves the same number of data. The data are all whole numbers, so the height of each bar represents the number of values equal to the corresponding midpoint shown on the horizontal axis. Notice that both distributions are symmetric.



- (a) Estimate the mode, median, and mean for each histogram.
 - (b) Which distribution has the larger standard deviation? Why?
3. | **Critical Thinking** Consider the following Minitab display of two data sets.

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
C1	20	20.00	1.62	7.26	7.00	15.00	20.00	25.00	31.00
C2	20	20.00	1.30	5.79	7.00	20.00	22.00	22.00	31.00

- (a) What are the respective means? the respective ranges?
 - (b) Which data set seems more symmetric? Why?
 - (c) Compare the interquartile ranges of the two sets. How do the middle halves of the data sets compare?
4. | **Consumer: Radon Gas** “Radon: The Problem No One Wants to Face” is the title of an article appearing in *Consumer Reports*. Radon is a gas emitted from the ground that can collect in houses and buildings. At certain levels it can cause lung cancer. Radon concentrations are measured in picocuries per liter (pCi/L). A radon level of 4 pCi/L is considered “acceptable.” Radon levels in a house vary from week to week. In one house, a sample of 8 weeks had the following readings for radon level (in pCi/L):

1.9 2.8 5.7 4.2 1.9 8.6 3.9 7.2

- (a) Find the mean, median, and mode.
 - (b) Find the sample standard deviation, coefficient of variation, and range.
5. | **Political Science: Georgia Democrats** How Democratic is Georgia? County-by-county results are shown for a recent election. For your convenience, the data

have been sorted in increasing order (Source: *County and City Data Book*, 12th Edition, U.S. Census Bureau).

Percentage of Democratic Vote by Counties in Georgia

31	33	34	34	35	35	35	36	38	38	38	39	40	40	40	40
41	41	41	41	41	41	41	42	42	43	44	44	44	45	45	46
46	46	46	47	48	49	49	49	49	50	51	52	52	53	53	53
53	53	55	56	56	57	57	59	62	66	66	68				

- (a) Make a box-and-whisker plot of the data. Find the interquartile range.
- (b) **Grouped Data** Make a frequency table using five classes. Then estimate the mean and sample standard deviation using the frequency table. Compute a 75% Chebyshev interval centered about the mean.
- (c) If you have a statistical calculator or computer, use it to find the actual sample mean and sample standard deviation. Otherwise, use the values $\Sigma x = 2769$ and $\Sigma x^2 = 132,179$ to compute the sample mean and sample standard deviation.
6. **Grades: Weighted Average** Professor Cramer determines a final grade based on attendance, two papers, three major tests, and a final exam. Each of these activities has a total of 100 possible points. However, the activities carry different weights. Attendance is worth 5%, each paper is worth 8%, each test is worth 15%, and the final is worth 34%.
- (a) What is the average for a student with 92 on attendance, 73 on the first paper, 81 on the second paper, 85 on test 1, 87 on test 2, 83 on test 3, and 90 on the final exam?
- (b) Compute the average for a student with the above scores on the papers, tests, and final exam, but with a score of only 20 on attendance.
7. **General: Average Weight** An elevator is loaded with 16 people and is at its load limit of 2500 pounds. What is the mean weight of these people?
8.  **Agriculture: Harvest Weight of Maize** The following data represent weights in kilograms of maize harvest from a random sample of 72 experimental plots on St. Vincent, an island in the Caribbean (Reference: B. G. F. Springer, *Proceedings, Caribbean Food Corps. Soc.*, Vol. 10, pp. 147–152). Note: These data are also available for download at the Online Study Center. For convenience, the data are presented in increasing order.
- | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| 7.8 | 9.1 | 9.5 | 10.0 | 10.2 | 10.5 | 11.1 | 11.5 | 11.7 | 11.8 |
| 12.2 | 12.2 | 12.5 | 13.1 | 13.5 | 13.7 | 13.7 | 14.0 | 14.4 | 14.5 |
| 14.6 | 15.2 | 15.5 | 16.0 | 16.0 | 16.1 | 16.5 | 17.2 | 17.8 | 18.2 |
| 19.0 | 19.1 | 19.3 | 19.8 | 20.0 | 20.2 | 20.3 | 20.5 | 20.9 | 21.1 |
| 21.4 | 21.8 | 22.0 | 22.0 | 22.4 | 22.5 | 22.5 | 22.8 | 22.8 | 23.1 |
| 23.1 | 23.2 | 23.7 | 23.8 | 23.8 | 23.8 | 23.8 | 24.0 | 24.1 | 24.1 |
| 24.5 | 24.5 | 24.9 | 25.1 | 25.2 | 25.5 | 26.1 | 26.4 | 26.5 | 26.7 |
| 27.1 | 29.5 | | | | | | | | |
- (a) Compute the five-number summary.
- (b) Compute the interquartile range.
- (c) Make a box-and-whisker plot.
- (d) **Interpretation** Discuss the distribution. Does the lower half of the distribution show more data spread than the upper half?
9. **Focus Problem: The Educational Advantage** Solve the focus problem at the beginning of this chapter.
10. **Agriculture: Bell Peppers** The pathogen *Phytophthora capsici* causes bell pepper plants to wilt and die. A research project was designed to study the effect of soil water content and the spread of the disease in fields of bell pep-

pers (Source: *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 2, No. 2). It is thought that too much water helps spread the disease. The fields were divided into rows and quadrants. The soil water content (percent of water by volume of soil) was determined for each plot. An important first step in such a research project is to give a statistical description of the data.

Soil Water Content for Bell Pepper Study

15	14	14	14	13	12	11	11	11	11	10	11	13	16	10
9	15	12	9	10	7	14	13	14	8	9	8	11	13	13
15	12	9	10	9	9	16	16	12	10	11	11	12	15	6
10	10	10	11	9										

- (a) Make a box-and-whisker plot of the data. Find the interquartile range.
- (b) **Grouped Data** Make a frequency table using four classes. Then estimate the mean and sample standard deviation using the frequency table. Compute a 75% Chebyshev interval centered about the mean.
- (c) If you have a statistical calculator or computer, use it to find the actual sample mean and sample standard deviation.

11. **Performance Rating: Weighted Average** A performance evaluation for new sales representatives at Office Automation Incorporated involves several ratings done on a scale of 1 to 10, with 10 the highest rating. The activities rated include new contacts, successful contacts, total contacts, dollar volume of sales, and reports. Then an overall rating is determined by using a weighted average. The weights are 2 for new contacts, 3 for successful contacts, 3 for total contacts, 5 for dollar value of sales, and 3 for reports. What would the overall rating be for a sales representative with ratings of 5 for new contacts, 8 for successful contacts, 7 for total contacts, 9 for dollar volume of sales, and 7 for reports?

DATA HIGHLIGHTS: GROUP PROJECTS

Steve Maehl 2009/Used under license from Shutterstock.com



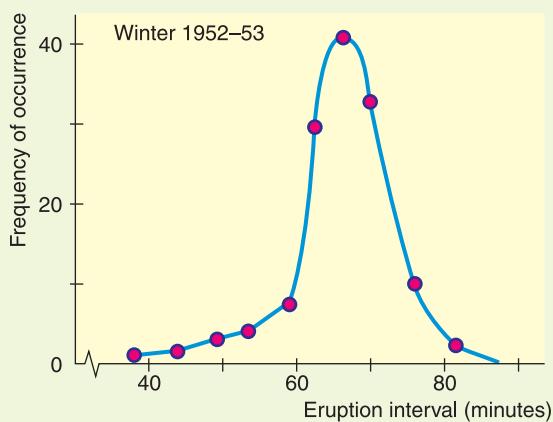
Old Faithful Geyser, Yellowstone National Park

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

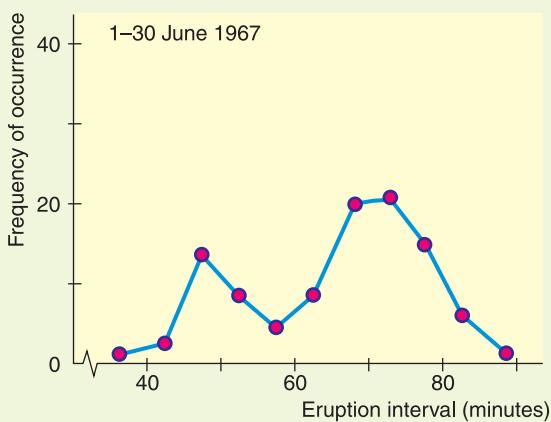
1. *The Story of Old Faithful* is a short book written by George Marler and published by the Yellowstone Association. Chapter 7 of this interesting book talks about the effect of the 1959 earthquake on eruption intervals for Old Faithful Geyser. Dr. John Rinehart (a senior research scientist with the National Oceanic and Atmospheric Administration) has done extensive studies of the eruption intervals before and after the 1959 earthquake. Examine Figure 3-11. Notice the general shape. Is the graph more or less symmetrical? Does it have a single mode frequency? The mean interval between eruptions has remained steady at about 65 minutes for the past 100 years. Therefore, the 1959 earthquake did not significantly change the mean, but it did change the distribution of eruption intervals. Examine Figure 3-12. Would you say there are really two frequency modes, one shorter and the other longer? Explain. The overall mean is about the same for both graphs, but one graph has a much larger standard deviation (for eruption intervals) than the other. Do no calculations, just look at both graphs, and then explain which graph has the smaller and which has the larger standard deviation. Which distribution will have the larger coefficient of variation? In everyday terms, what would this mean if you were actually at Yellowstone waiting to see the next eruption of Old Faithful? Explain your answer.
2. Most academic advisors tell students to major in a field they really love. After all, it is true that money cannot buy happiness! Nevertheless, it is interesting to at least look at some of the higher-paying fields of study. After all, a field like

FIGURE 3-11

Typical Behavior of Old Faithful Geyser Before 1959 Quake

**FIGURE 3-12**

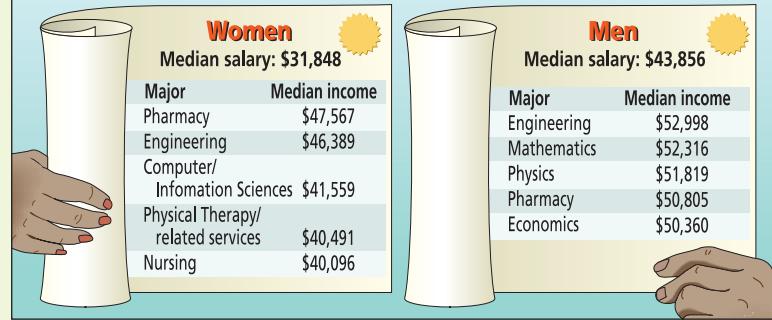
Typical Behavior of Old Faithful Geyser After 1959 Quake



mathematics can be a lot of fun, once you get into it. We see that women's salaries tend to be less than men's salaries. However, women's salaries are rapidly catching up, and this benefits the entire workforce in different ways. Figure 3-13 shows the median incomes for college graduates with different majors. The employees in the sample are all at least 30 years old. Does it seem reasonable to assume that many of the employees are in jobs beyond the entry level? Explain. Compare the median incomes shown for all women aged 30 or older holding bachelor's degrees with the median incomes for men of similar age holding bachelor's degrees. Look at the particular majors listed. What percentage of men holding bachelor's degrees in mathematics make \$52,316 or more? What percentage of women holding computer/information science degrees make \$41,559 or more? How do median incomes for men and women holding engineering degrees compare? What about pharmacy degrees?

FIGURE 3-13**Lucrative Majors for Bachelor's Degrees**

The U.S. median income for women and men (age 30 or older) with bachelor's degrees



Source: Bureau of Labor Statistics

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. An average is an attempt to summarize a collection of data into just *one* number. Discuss how the mean, median, and mode all represent averages in this context. Also discuss the differences among these averages. Why is the mean a balance point? Why is the median a midway point? Why is the mode the most common data point? List three areas of daily life in which you think the mean, median, or mode would be the best choice to describe an “average.”
2. Why do we need to study the variation of a collection of data? Why isn’t the average by itself adequate? We have studied three ways to measure variation. The range, the standard deviation, and, to a large extent, a box-and-whisker plot all indicate the variation within a data collection. Discuss similarities and differences among these ways to measure data variation. Why would it seem reasonable to pair the median with a box-and-whisker plot and to pair the mean with the standard deviation? What are the advantages and disadvantages of each method of describing data spread? Comment on statements such as the following: (a) The range is easy to compute, but it doesn’t give much information; (b) although the standard deviation is more complicated to compute, it has some significant applications; (c) the box-and-whisker plot is fairly easy to construct, and it gives a lot of information at a glance.
3. Why is the coefficient of variation important? What do we mean when we say that the coefficient of variation has no units? What advantage can there be in having no units? Why is *relative size* important?

Consider robin eggs; the mean weight of a collection of robin eggs is 0.72 ounce and the standard deviation is 0.12 ounce. Now consider elephants; the mean weight of elephants in the zoo is 6.42 tons, with a standard deviation 1.07 tons. The units of measurement are different and there is a great deal of difference between the weight of an elephant and that of a robin’s egg. Yet the coefficient of variation is about the same for both. Comment on this from the viewpoint of the size of the standard deviation relative to that of the mean.
4. What is Chebyshev’s theorem? Suppose you have a friend who knows very little about statistics. Write a paragraph or two in which you describe Chebyshev’s theorem for your friend. Keep the discussion as simple as possible, but be sure to get the main ideas across to your friend. Suppose he or she asks, “What is this stuff good for?” and suppose you respond (a little sarcastically) that Chebyshev’s theorem applies to everything from butterflies to the orbits of the planets! Would you be correct? Explain.

USING TECHNOLOGY

Raw Data

Application

Using the software or calculator available to you, do the following.

1. Trade winds are one of the beautiful features of island life in Hawaii. The following data represent total air movement in miles per day over a weather station in Hawaii as determined by a continuous anemometer recorder. The period of observation is January 1 to February 15, 1971.

26	14	18	14
113	50	13	22
27	57	28	50
72	52	105	138
16	33	18	16
32	26	11	16
17	14	57	100
35	20	21	34
18	13	18	28
21	13	25	19
11	19	22	19
15	20		

Source: United States Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Data Service. *Climatological Data, Annual Summary, Hawaii*, Vol. 67, No. 13. Asheville: National Climatic Center, 1971, pp. 11, 24.

- (a) Use the computer to find the sample mean, median, and (if it exists) mode. Also, find the range, sample variance, and sample standard deviation.
- (b) Use the five-number summary provided by the computer to make a box-and-whisker plot of total air movement over the weather station.
- (c) Four data values are exceptionally high: 113, 105, 138, and 100. The strong winds of January 5 (113 reading) brought in a cold front that dropped snow on Haleakala National Park (at the 8000 ft elevation). The residents were so excited that they drove up to see the snow and caused such a massive traffic jam that the Park Service had to close the road. The winds of January 15, 16, and 28 (readings 105, 138, and 100) accompanied a storm with funnel clouds that did much damage. Eliminate these values (i.e., 100, 105,

113, and 138) from the data bank and redo parts (a) and (b). Compare your results with those previously obtained. Which average is most affected? What happens to the standard deviation? How do the two box-and-whisker plots compare?

Technology Hints: Raw Data

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad), Excel 2007, Minitab

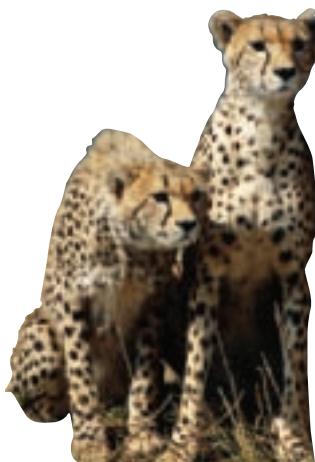
The Tech Note of Section 3.2 gives brief instructions for finding summary statistics for raw data using the TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab. The Tech Note of Section 3.3 gives brief instructions for constructing box-and-whisker plots using the TI-84Plus/TI-83Plus/TI-nspire calculators and Minitab.

SPSS

Many commands in SPSS provide an option to display various summary statistics. A direct way to display summary statistics is to use the menu choices **Analyze > Descriptive Statistics > Descriptives**. In the dialogue box, move the variable containing your data into the variables box. Click **Options...** and then check the summary statistics you wish to display. Click **Continue** and then **OK**. Notice that the median is not available. A more complete list of summary statistics is available with the menu choices **Analyze > Descriptive Statistics > Frequencies**. Click the **Statistics** button and check the summary statistics you wish to display.

For box-and-whisker plots, use the menu options **Graphs > Legacy Dialogues > Interactive > Boxplot**. In the dialogue box, place the variable containing your data in the box along the vertical axis. After selecting the options you want, click **OK**.

SPSS Student Version 17 has other options for creating box-and-whisker plots. The Tech Guide for SPSS that accompanies this text has instructions for using the **Chart Builder** and **Graphboard Template Chooser** options found in the **Graphs** menu. Note that SPSS requires the appropriate level of measurement for processing data. In SPSS, the scale level corresponds to both interval and ratio level. The nominal and ordinal levels are as described in this text.



Cumulative Review Problems

CHAPTERS 1–3

Critical Thinking and Literacy

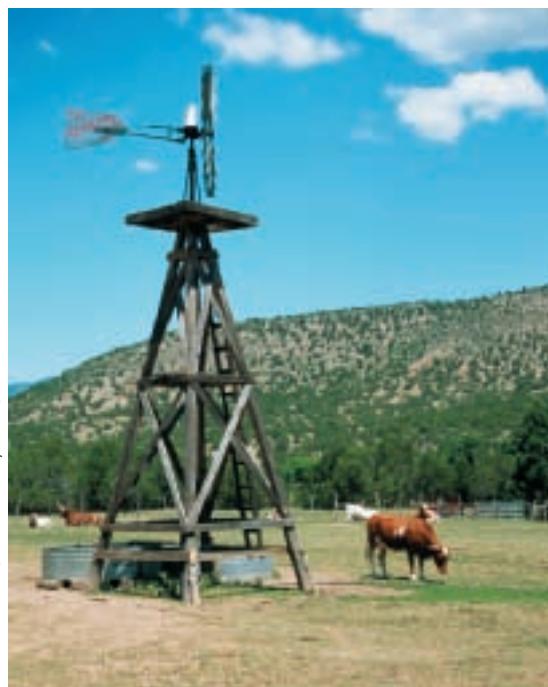
1. Consider the following measures: mean, median, variance, standard deviation, percentile.
 - (a) Which measures utilize relative position of the data values?
 - (b) Which measures utilize actual data values regardless of relative position?
2. Describe how the presence of possible outliers might be identified on
 - (a) histograms.
 - (b) dotplots.
 - (c) stem-and-leaf displays.
 - (d) box-and-whisker plots.
3. Consider two data sets, A and B. The sets are identical except that the high value of data set B is three times greater than the high value of data set A.
 - (a) How do the medians of the two data sets compare?
 - (b) How do the means of the two data sets compare?
 - (c) How do the standard deviations of the two data sets compare?
 - (d) How do the box-and-whisker plots of the two data sets compare?
4. You are examining two data sets involving test scores, set A and set B. The score 86 appears in both data sets. In which of the following data sets does 86 represent a higher score? Explain.
 - (a) The percentile rank of 86 is higher in set A than in set B.
 - (b) The mean is 80 in both data sets, but set A has a higher standard deviation.

Applications



In west Texas, water is extremely important. The following data represent pH levels in ground water for a random sample of 102 west Texas wells. A pH less

than 7 is acidic and a pH above 7 is alkaline. Scanning the data, you can see that water in this region tends to be hard (alkaline). Too high a pH means the water is unusable or needs expensive treatment to make it usable (Reference: C. E. Nichols and V. E. Kane, Union Carbide Technical Report K/UR-1).



Dave G. Houser/Documentary Value/Corbis

These data are also available for download at the Online Study Center. For convenience, the data are presented in increasing order.

x: pH of Ground Water in 102 West Texas Wells

7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.1 7.1 7.1 7.1
7.1 7.1 7.1 7.1 7.1 7.1 7.2 7.2 7.2 7.2 7.2 7.2
7.2 7.2 7.2 7.2 7.3 7.3 7.3 7.3 7.3 7.3 7.3 7.3
7.3 7.3 7.3 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4
7.5 7.5 7.5 7.5 7.5 7.5 7.6 7.6 7.6 7.6 7.6 7.6
7.6 7.6 7.6 7.6 7.6 7.7 7.7 7.7 7.7 7.7 7.7 7.8
7.8 7.8 7.8 7.9 7.9 7.9 7.9 7.9 8.0 8.1 8.1
8.1 8.1 8.1 8.1 8.2 8.2 8.2 8.2 8.2 8.2 8.2
8.4 8.5 8.6 8.7 8.8 8.8

5. Write a brief description in which you outline how you would obtain a random sample of 102 west Texas water wells. Explain how random numbers would be used in the selection process.
6. Is the given data nominal, ordinal, interval, or ratio? Explain.
7. Make a stem-and-leaf display. Use five lines per stem so that leaf values 0 and 1 are on one line, 2 and 3 are on the next line, 4 and 5 are on the next, 6 and 7 are on the next, and 8 and 9 are on the last line of the stem.
8. Make a frequency table, histogram, and relative-frequency histogram using five classes. Recall that for decimal data, we “clear the decimal” to determine classes for whole-number data and then reinsert the decimal to obtain the classes for the frequency table of the original data.
9. Make an ogive using five classes.

10. Compute the range, mean, median, and mode for the given data.
11. (a) Verify that $\Sigma x = 772.9$ and $\Sigma x^2 = 5876.6$.
(b) Compute the sample variance, sample standard deviation, and coefficient of variation for the given data. Is the sample standard deviation small relative to the mean pH?
12. Compute a 75% Chebyshev interval centered on the mean.
13. Make a box-and-whisker plot. Find the interquartile range.

Interpretation

Wow! In Problems 5–13 you constructed a lot of information regarding the pH of west Texas ground water based on sample data. Let’s continue the investigation.

14. Look at the histogram. Is the pH distribution for these wells symmetric or skewed? Are lower or higher values more common?
15. Look at the ogive. What percent of the wells have a pH less than 8.15? Suppose a certain crop can tolerate irrigation water with a pH between 7.35 and 8.55. What percent of the wells could be used for such a crop?
16. Look at the stem-and-leaf plot. Are there any unusually high or low pH levels in this sample of wells? How many wells are neutral (pH of 7)?

17. Use the box-and-whisker plot to describe how the data are spread about the median. Are the pH values above the median more spread out than those below? Is this observation consistent with the skew of the histogram?
18. Suppose you are working for the regional water commissioner. You have been asked to submit a brief report about the pH level in ground water in the west Texas region. Write such a report and include appropriate graphs.



4

4.1 What Is Probability?

4.2 Some Probability Rules—Compound Events

4.3 Trees and Counting Techniques



North Wind/North Wind Picture Archive

We see that the theory of probabilities is at bottom only common sense reduced to calculation; it makes us appreciate with exactitude what reasonable minds feel by a sort of instinct, often without being able to account for it.

—PIERRE-SIMON LAPLACE

For online student resources, visit the Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>.



Nathan Griffith/Comet/Corbis

This quote explains how the great mathematician Pierre-Simon Laplace (1749–1827) described the theory of mathematical probability. The discovery of the mathematical theory of probability was shared by two Frenchmen: Blaise Pascal and Pierre Fermat. These 17th-century scholars were attracted to the subject by the inquiries of the Chevalier de Méré, a gentleman gambler.

Although the first applications of probability were to games of chance and gambling, today the subject seems to pervade almost every aspect of modern life. Everything from the orbits of spacecraft to the social behavior of woodchucks is described in terms of probabilities.

ELEMENTARY PROBABILITY THEORY

PREVIEW QUESTIONS

Why would anyone study probability? Hint: Most big issues in life involve uncertainty. (SECTION 4.1)

What are the basic definitions and rules of probability? (SECTION 4.2)

What are counting techniques, trees, permutations, and combinations? (SECTION 4.3)

FOCUS PROBLEM

How Often Do Lie Detectors Lie?

James Burke is an educator who is known for his interesting science-related radio and television shows aired by the British Broadcasting Corporation. His book *Chances: Risk and Odds in Everyday Life* (Virginia Books, London) contains a great wealth of fascinating information about probabilities. The following quote is from Professor Burke's book:

If I take a polygraph test and lie, what is the risk I will be detected? According to some studies, there's about a 72 percent chance you will be caught by the machine.

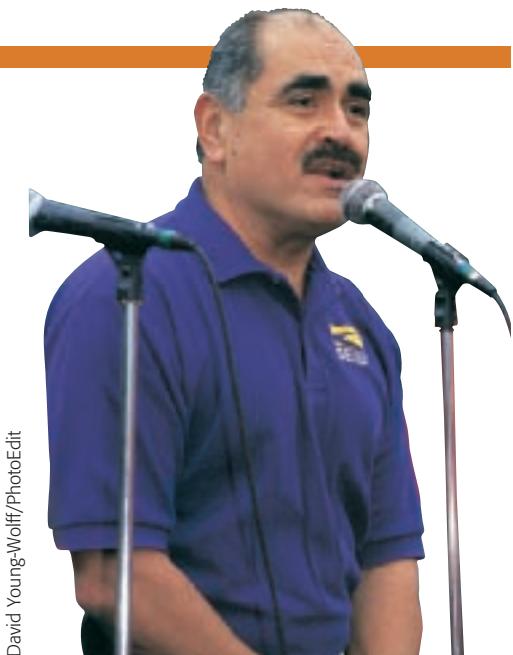
What is the risk that if I take a polygraph test it will incorrectly say that I lied? At least 1 in 15 will be thus falsely accused.

Both of these statements contain conditional probabilities, which we will study in Section 4.2. Information from that section will enable us to answer the following:

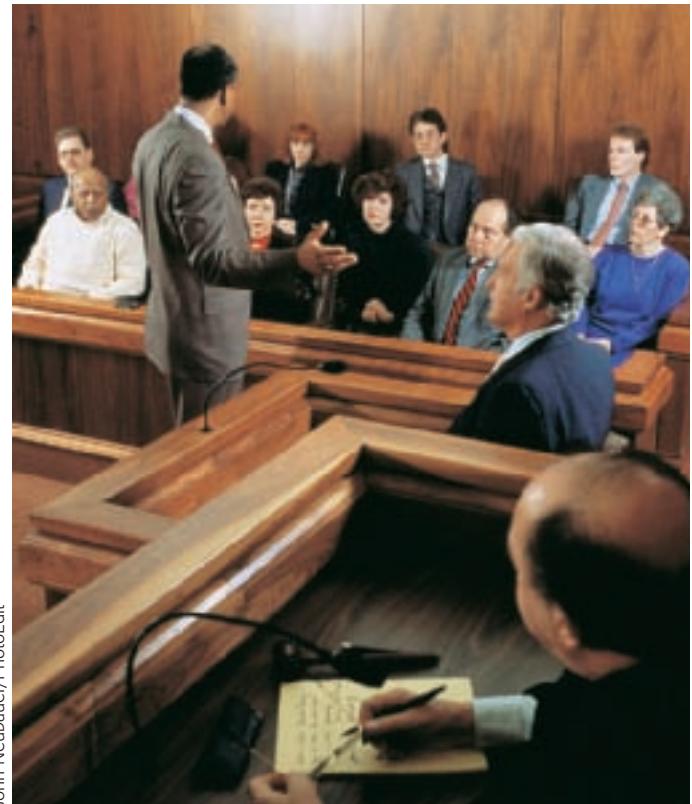
Suppose a person answers 10% of a long battery of questions with lies. Assume that the remaining 90% of the questions are answered truthfully.

1. Estimate the percentage of answers the polygraph will *wrongly* indicate as lies.
2. Estimate the percentage of answers the polygraph will *correctly* indicate as lies.

If the polygraph indicated that 30% of the questions were answered as lies, what would you estimate for the *actual* percentage of questions the person answered as lies? (See Problems 25 and 26 in Section 4.2.)



David Young-Wolff/PhotoEdit



John Neubauer/PhotoEdit

SECTION 4.1

What Is Probability?**FOCUS POINTS**

- Assign probabilities to events.
- Explain how the law of large numbers relates to relative frequencies.
- Apply basic rules of probability in everyday life.
- Explain the relationship between statistics and probability.

We encounter statements given in terms of probability all the time. An excited sports announcer claims that Sheila has a 90% chance of breaking the world record in the upcoming 100-yard dash. Henry figures that if he guesses on a true–false question, the probability of getting it right is $1/2$. The Right to Health lobby claims that the probability is 0.40 of getting an erroneous report from a medical laboratory at a low-cost health center. It is consequently lobbying for a federal agency to license and monitor all medical laboratories.

When we use probability in a statement, we're using a *number between 0 and 1* to indicate the likelihood of an event.

Probability is a numerical measure between 0 and 1 that describes the likelihood that an event will occur. Probabilities closer to 1 indicate that the event is more likely to occur. Probabilities closer to 0 indicate that the event is less likely to occur.

Probability of an event A, $P(A)$

$P(A)$, read “P of A,” denotes the **probability of event A**.

If $P(A) = 1$, event A is certain to occur.

If $P(A) = 0$, event A is certain not to occur.

It is important to know what probability statements mean and how to determine probabilities of events, because probability is the language of inferential statistics.

Intuition**Probability assignments**

1. A probability assignment based on **intuition** incorporates past experience, judgment, or opinion to estimate the likelihood of an event.
2. A probability assignment based on **relative frequency** uses the formula

$$\text{Probability of event} = \text{relative frequency} = \frac{f}{n} \quad (1)$$

where f is the frequency of the event occurrence in a sample of n observations.

Relative frequency

3. A probability assignment based on **equally likely outcomes** uses the formula

$$\text{Probability of event} = \frac{\text{Number of outcomes favorable to event}}{\text{Total number of outcomes}} \quad (2)$$

EXAMPLE 1**PROBABILITY ASSIGNMENT**

Consider each of the following events, and determine how the probability is assigned.

- (a) A sports announcer claims that Sheila has a 90% chance of breaking the world record in the 100-yard dash.

SOLUTION: It is likely the sports announcer used intuition based on Sheila's past performance.

- (b) Henry figures that if he guesses on a true-false question, the probability of getting it right is 0.50.

SOLUTION: In this case there are two possible outcomes: Henry's answer is either correct or incorrect. Since Henry is guessing, we assume the outcomes are equally likely. There are $n = 2$ equally likely outcomes, and only one is correct. By formula (2),

$$P(\text{correct answer}) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{1}{2} = 0.50$$

- (c) The Right to Health lobby claims that the probability of getting an erroneous medical laboratory report is 0.40, based on a random sample of 200 laboratory reports, of which 80 were erroneous.

SOLUTION: Formula (1) for relative frequency gives the probability, with sample size $n = 200$ and number of errors $f = 80$.

$$P(\text{error}) = \text{relative frequency} = \frac{f}{n} = \frac{80}{200} = 0.40$$

We've seen three ways to assign probabilities: intuition, relative frequency, and—when outcomes are equally likely—a formula. Which do we use? Most of the time it depends on the information that is at hand or that can be feasibly obtained. Our choice of methods also depends on the particular problem. In Guided Exercise 1, you will see three different situations, and you will decide the best way to assign the probabilities. *Remember, probabilities are numbers between 0 and 1, so don't assign probabilities outside this range.*

GUIDED EXERCISE 1**Determine a probability**

Assign a probability to the indicated event on the basis of the information provided. Indicate the technique you used: intuition, relative frequency, or the formula for equally likely outcomes.

- (a) A random sample of 500 students at Hudson College were surveyed and it was determined that 375 wear glasses or contact lenses. Estimate the probability that a Hudson College student selected at random wears corrective lenses.



In this case we are given a sample size of 500, and we are told that 375 of these students wear corrective lenses. It is appropriate to use a relative frequency for the desired probability:

$$P(\text{student needs corrective lenses}) = \frac{f}{n} = \frac{375}{500} = 0.75$$

Continued

GUIDED EXERCISE 1 *continued*

(b) The Friends of the Library hosts a fundraising barbecue. George is on the cleanup committee. There are four members on this committee, and they draw lots to see who will clean the grills. Assuming that each member is equally likely to be drawn, what is the probability that George will be assigned the grill-cleaning job?



There are four people on the committee, and each is equally likely to be drawn. It is appropriate to use the formula for equally likely events. George can be drawn in only one way, so there is only one outcome favorable to the event.

(c) Joanna photographs whales for Sea Life Adventure Films. On her next expedition, she is to film blue whales feeding. Based on her knowledge of the habits of blue whales, she is almost certain she will be successful. What specific number do you suppose she estimates for the probability of success?



Since Joanna is almost certain of success, she should make the probability close to 1. We could say $P(\text{success})$ is above 0.90 but less than 1. This probability assignment is based on intuition.

Law of large numbers

The technique of using the relative frequency of an event as the probability of that event is a common way of assigning probabilities and will be used a great deal in later chapters. The underlying assumption we make is that if events occurred a certain percentage of times in the past, they will occur about the same percentage of times in the future. In fact, this assumption can be strengthened to a very general statement called the *law of large numbers*.

Law of large numbers

In the long run, as the sample size increases and increases, the relative frequencies of outcomes get closer and closer to the theoretical (or actual) probability value.

The law of large numbers is the reason businesses such as health insurance, automobile insurance, and gambling casinos can exist and make a profit.

No matter how we compute probabilities, it is useful to know what outcomes are possible in a given setting. For instance, if you are going to decide the probability that Hardscrabble will win the Kentucky Derby, you need to know which other horses will be running.

To determine the possible outcomes for a given setting, we need to define a *statistical experiment*.

Statistical experiment

Event

Simple event

Sample space

A **statistical experiment** or **statistical observation** can be thought of as any random activity that results in a definite outcome.

An **event** is a collection of one or more outcomes of a statistical experiment or observation.

A **simple event** is one particular outcome of a statistical experiment.

The set of all simple events constitutes the **sample space** of an experiment.

**EXAMPLE 2****USING A SAMPLE SPACE**

Human eye color is controlled by a single pair of genes (one from the father and one from the mother) called a *genotype*. Brown eye color, B, is dominant over blue eye color, ℓ . Therefore, in the genotype $B\ell$, consisting of one brown gene B

and one blue gene ℓ , the brown gene dominates. A person with a $B\ell$ genotype has brown eyes.

If both parents have brown eyes and have genotype $B\ell$, what is the probability that their child will have blue eyes? What is the probability the child will have brown eyes?

SOLUTION: To answer these questions, we need to look at the sample space of all possible eye-color genotypes for the child. They are given in Table 4-1.

TABLE 4-1 Eye Color Genotypes for Child

Father	Mother	
	B	ℓ
B	BB	$B\ell$
ℓ	ℓB	$\ell\ell$

According to genetics theory, the four possible genotypes for the child are equally likely. Therefore, we can use Formula (2) to compute probabilities. Blue eyes can occur only with the $\ell\ell$ genotype, so there is only one outcome favorable to blue eyes. By formula (2),

$$P(\text{blue eyes}) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{1}{4}$$

Brown eyes occur with the three remaining genotypes: BB, $B\ell$, and ℓB . By formula (2),

$$P(\text{brown eyes}) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{3}{4}$$

Bonnie Kamin/PhotoEdit



GUIDED EXERCISE 2

Using a sample space

Professor Gutierrez is making up a final exam for a course in literature of the southwest. He wants the last three questions to be of the true-false type. To guarantee that the answers do not follow his favorite pattern, he lists all possible true-false combinations for three questions on slips of paper and then picks one at random from a hat.

- (a) Finish listing the outcomes in the given sample space.  The missing outcomes are FFT and FFF.

TTT	FTT	TFT	_____
TTF	FTF	TFF	_____

- (b) What is the probability that all three items will be false? Use the formula

$$P(\text{all } F) = \frac{\text{No. of favorable outcomes}}{\text{Total no. of outcomes}}$$

$$P(\text{all } F) = \frac{1}{8}$$

 There is only one outcome, FFF, favorable to all false, so

- (c) What is the probability that exactly two items will be true?

 There are three outcomes that have exactly two true items: TTF, TFT, and FTT. Thus,

$$P(\text{two } T) = \frac{\text{No. of favorable outcomes}}{\text{Total no. of outcomes}} = \frac{3}{8}$$

There is another important point about probability assignments of simple events.

The **sum** of the probabilities of all simple events in a sample space must equal 1.

We can use this fact to determine the probability that an event will not occur. For instance, if you think the probability is 0.65 that you will win a tennis match, you assume the probability is 0.35 that your opponent will win.

The *complement* of an event A is the event that A *does not occur*. We use the notation A^c to designate the complement of event A . Figure 4-1 shows the event A and its complement A^c .

Notice that the two distinct events A and A^c make up the entire sample space. Therefore, the sum of their probabilities is 1.

Complement of an event A^c

The **complement of event A** is the event that A *does not occur*. A^c designates the complement of event A . Furthermore,

1. $P(A) + P(A^c) = 1$
 2. $P(\text{event } A \text{ does not occur}) = P(A^c) = 1 - P(A)$
- (3)



EXAMPLE 3 COMPLEMENT OF AN EVENT

The probability that a college student who has not received a flu shot will get the flu is 0.45. What is the probability that a college student will *not* get the flu if the student has not had the flu shot?

SOLUTION: In this case, we have

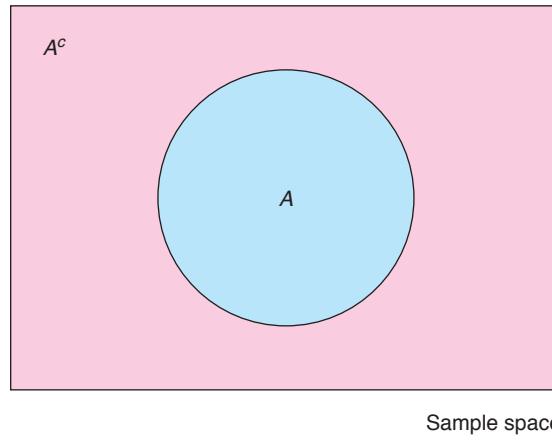
$$P(\text{will get flu}) = 0.45$$

$$P(\text{will not get flu}) = 1 - P(\text{will get flu}) = 1 - 0.45 = 0.55$$



FIGURE 4-1

The Event A and Its Complement A^c



GUIDED EXERCISE 3**Complement of an event**

A veterinarian tells you that if you breed two cream-colored guinea pigs, the probability that an offspring will be pure white is 0.25. What is the probability that an offspring will not be pure white?

$$(a) P(\text{pure white}) + P(\text{not pure white}) = \underline{\hspace{2cm}} \rightarrow 1$$

$$(b) P(\text{not pure white}) = \underline{\hspace{2cm}} \rightarrow 1 - 0.25, \text{ or } 0.75$$

LOOKING FORWARD

The complement rule has special application in Chapter 5. There we study probabilities associated with binomial experiments. These are experiments that have only two outcomes, “success” and “failure.” One feature of these experiments is that the outcomes can be categorized according to the number of “successes” out of n independent trials.

Simple Events of Sample Space of a Binomial Experiment
0 successes, 1 success, 2 successes, 3 successes, ..., n successes

A convenient way to compute the probability of “at least one” success is to observe that the complement of “at least one success” is the event “0 successes.” Then, by the complement rule

$$P(\text{at least one success}) = 1 - P(0 \text{ successes})$$

Problems 8, 9, and 10 of this section show a similar use of the complement rule.



Sometimes, instead of being told the probability of an event, we are given the odds for the event. Problems 19 and 20 discuss odds in favor and odds against an event.

Summary: Some important facts about probability

1. A statistical experiment or statistical observation is any random activity that results in a definite outcome. A **simple event** consists of one and only one outcome of the experiment. The **sample space** is the set of all simple events. An **event A** is any subset of the sample space.
2. The probability of an event A is denoted by $P(A)$.
3. The probability of an event is a number between 0 and 1. The closer to 1 the probability is, the more likely it is the event will occur. The closer to 0 the probability is, the less likely it is the event will occur.
4. The sum of the probabilities of all simple events in a sample space is 1.
5. Probabilities can be assigned by using intuition, relative frequencies, or the formula for equally likely outcomes. Additional ways to assign probabilities will be introduced in later chapters.
6. The **complement** of an event A is denoted by A^c . So, A^c is the event that A does not occur.
7. $P(A) + P(A^c) = 1$

Interpreting Probabilities

As stressed in item 3 in the summary about probabilities, the closer the probability is to 1, the more likely the event is to occur. The closer the probability is to 0, the less likely the event is to occur. However, just because the probability of an event is very high, it is not a certainty that the event will occur. Likewise, even though the probability of an event is very low, the event might still occur.

Events with low probability, but big consequences, are of special concern. Such events are sometimes referred to as “black swan events,” as discussed in the very popular book *The Black Swan* by Nassim Nicholas Taleb. In his varied careers Taleb has been a stock trader, author, and professor at the University of Massachusetts. His special interests are mathematical problems of luck, uncertainty, probability, and general knowledge. In his book *The Black Swan*, Taleb discusses “the impact of the highly improbable.” If an event has (a) little impact, bearing, or meaning on a person’s life and (b) very low probability, then (c) it seems reasonable to ignore the event.

However, some of the really big mistakes in a person’s life can result from misjudging either (a) the size of an event’s impact or (b) the likelihood the event will occur. An event of great importance cannot be ignored even if it has a low probability of occurrence.

So what should a person do?

Taleb describes himself as a hyperskeptic when others are gullible, and gullible when others are skeptical. In particular, he recommends being skeptical about (mentally) confirming an event when the errors of confirmation are costly. Lots and lots of data should not be thought of as confirming an event when a simple instance can disconfirm an important and costly event. Taleb recommends that one be skeptical in the face of wild randomness (and big consequences) and gullible when the randomness (and consequences) are mild. One way to estimate the “randomness” of a situation is to estimate its probability. Estimating the “consequences” of an event in your own life must be a personal decision.

Probability Related to Statistics

We conclude this section with a few comments on the nature of statistics versus probability. Although statistics and probability are closely related fields of mathematics, they are nevertheless separate fields. It can be said that probability is the medium through which statistical work is done. In fact, if it were not for probability theory, inferential statistics would not be possible.

Put very briefly, probability is the field of study that makes statements about what will occur when samples are drawn from a *known population*. Statistics is the field of study that describes how samples are to be obtained and how inferences are to be made about *unknown populations*.

A simple but effective illustration of the difference between these two subjects can be made by considering how we treat the following examples.

Example of a probability application

Condition: We *know* the exact makeup of the *entire* population.

Example: Given 3 green marbles, 5 red marbles, and 4 white marbles in a bag, draw 6 marbles at random from the bag. What is the probability that none of the marbles is red?

Example of a statistical application

Condition: We have only *samples* from an otherwise *unknown* population.

Example: Draw a random sample of 6 marbles from the (unknown) population of all marbles in a bag and observe the colors. Based on the sample results, make a conjecture about the colors and numbers of marbles in the entire population of all marbles in the bag.

In another sense, probability and statistics are like flip sides of the same coin. On the probability side, you know the overall description of the population. The central problem is to compute the likelihood that a specific outcome will happen.

On the statistics side, you know only the results of a sample drawn from the population. The central problem is to describe the sample (descriptive statistic) and to draw conclusions about the population based on the sample results (inferential statistics).

In statistical work, the inferences we draw about an unknown population are not claimed to be absolutely correct. Since the population remains unknown (in a theoretical sense), we must accept a “best guess” for our conclusions and act using the most probable answer rather than absolute certainty.

Probability is the topic of this chapter. However, we will not study probability just for its own sake. Probability is a wonderful field of mathematics, but we will study mainly the ideas from probability that are needed for a proper understanding of statistics.

VIEWPOINT

What Makes a Good Teacher?

A survey of 735 students at nine colleges in the United States was taken to determine instructor behaviors that help students succeed. Data from this survey can be found by visiting the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and finding the link to DASL, the Carnegie Mellon University Data and Story Library. Once at the DASL site, select Data Subjects, then Psychology, and then Instructor Behavior. You can estimate the probability of how a student would respond (very positive, neutral, very negative) to different instructor behaviors. For example, more than 90% of the students responded “very positive” to the instructor’s use of real-world examples in the classroom.

SECTION 4.1 PROBLEMS

1. | **Statistical Literacy** List three methods of assigning probabilities.
2. | **Statistical Literacy** Suppose the newspaper states that the probability of rain today is 30%. What is the complement of the event “rain today”? What is the probability of the complement?
3. | **Statistical Literacy** What is the probability of
 - (a) an event A that is certain to occur?
 - (b) an event B that is impossible?
4. | **Statistical Literacy** What is the law of large numbers? If you were using the relative frequency of an event to estimate the probability of the event, would it be better to use 100 trials or 500 trials? Explain.
5. | **Basic Computation: Probability as Relative Frequency** A recent Harris Poll survey of 1010 U.S. adults selected at random showed that 627 consider the occupation of firefighter to have very great prestige. Estimate the probability (to the nearest hundredth) that a U.S. adult selected at random thinks the occupation of firefighter has very great prestige.
6. | **Basic Computation: Probability of Equally Likely Events** What is the probability that a day of the week selected at random will be a Wednesday?
7. | **Interpretation** An investment opportunity boasts that the chance of doubling your money in 3 years is 95%. However, when you research the details of the investment, you estimate that there is a 3% chance that you could lose the entire investment. Based on this information, are you certain to make money on this investment? Are there risks in this investment opportunity?
8. | **Interpretation** A sample space consists of 4 simple events: A , B , C , D . Which events comprise the complement of A ? Can the sample space be viewed as having two events, A and A^c ? Explain.

9. **Critical Thinking** Consider a family with 3 children. Assume the probability that one child is a boy is 0.5 and the probability that one child is a girl is also 0.5, and that the events “boy” and “girl” are independent.
- List the equally likely events for the gender of the 3 children, from oldest to youngest.
 - What is the probability that all 3 children are male? Notice that the complement of the event “all three children are male” is “at least one of the children is female.” Use this information to compute the probability that at least one child is female.
10. **Critical Thinking** Consider the experiment of tossing a fair coin 3 times. For each coin, the possible outcomes are heads or tails.
- List the equally likely events of the sample space for the three tosses.
 - What is the probability that all three coins come up heads? Notice that the complement of the event “3 heads” is “at least one tail.” Use this information to compute the probability that there will be at least one tail.
11. **Critical Thinking** On a single toss of a fair coin, the probability of heads is 0.5 and the probability of tails is 0.5. If you toss a coin twice and get heads on the first toss, are you guaranteed to get tails on the second toss? Explain.
12. **Critical Thinking**
- Explain why -0.41 cannot be the probability of some event.
 - Explain why 1.21 cannot be the probability of some event.
 - Explain why 120% cannot be the probability of some event.
 - Can the number 0.56 be the probability of an event? Explain.
13. **Probability Estimate: Wiggle Your Ears** Can you wiggle your ears? Use the students in your statistics class (or a group of friends) to estimate the percentage of people who can wiggle their ears. How can your result be thought of as an estimate for the probability that a person chosen at random can wiggle his or her ears? *Comment:* National statistics indicate that about 13% of Americans can wiggle their ears (Source: Bernice Kanner, *Are You Normal?*, St. Martin’s Press, New York).
14. **Probability Estimate: Raise One Eyebrow** Can you raise one eyebrow at a time? Use the students in your statistics class (or a group of friends) to estimate the percentage of people who can raise one eyebrow at a time. How can your result be thought of as an estimate for the probability that a person chosen at random can raise one eyebrow at a time? *Comment:* National statistics indicate that about 30% of Americans can raise one eyebrow at a time (see source in Problem 13).
15. **Myers–Briggs: Personality Types** Isabel Briggs Myers was a pioneer in the study of personality types. The personality types are broadly defined according to four main preferences. Do married couples choose similar or different personality types in their mates? The following data give an indication (Source: I. B. Myers and M. H. McCaulley, *A Guide to the Development and Use of the Myers–Briggs Type Indicators*).

Similarities and Differences in a Random Sample of 375 Married Couples

Number of Similar Preferences	Number of Married Couples
All four	34
Three	131
Two	124
One	71
None	15

Suppose that a married couple is selected at random.

- Use the data to estimate the probability that they will have 0, 1, 2, 3, or 4 personality preferences in common.



- (b) Do the probabilities add up to 1? Why should they? What is the sample space in this problem?
16. **General: Roll a Die**
- If you roll a single die and count the number of dots on top, what is the sample space of all possible outcomes? Are the outcomes equally likely?
 - Assign probabilities to the outcomes of the sample space of part (a). Do the probabilities add up to 1? Should they add up to 1? Explain.
 - What is the probability of getting a number less than 5 on a single throw?
 - What is the probability of getting 5 or 6 on a single throw?
17. **Psychology: Creativity** When do creative people get their *best* ideas? *USA Today* did a survey of 966 inventors (who hold U.S. patents) and obtained the following information:

Time of Day When Best Ideas Occur

Time	Number of Inventors
6 A.M.–12 noon	290
12 noon–6 P.M.	135
6 P.M.–12 midnight	319
12 midnight–6 A.M.	222

- (a) Assuming that the time interval includes the left limit and all the times up to but not including the right limit, estimate the probability that an inventor has a best idea during each time interval: from 6 A.M. to 12 noon, from 12 noon to 6 P.M., from 6 P.M. to 12 midnight, from 12 midnight to 6 A.M.
- (b) Do the probabilities of part (a) add up to 1? Why should they? What is the sample space in this problem?
18. **Agriculture: Cotton** A botanist has developed a new hybrid cotton plant that can withstand insects better than other cotton plants. However, there is some concern about the germination of seeds from the new plant. To estimate the probability that a seed from the new plant will germinate, a random sample of 3000 seeds was planted in warm, moist soil. Of these seeds, 2430 germinated.
- Use relative frequencies to estimate the probability that a seed will germinate. What is your estimate?
 - Use relative frequencies to estimate the probability that a seed will *not* germinate. What is your estimate?
 - Either a seed germinates or it does not. What is the sample space in this problem? Do the probabilities assigned to the sample space add up to 1? Should they add up to 1? Explain.
 - Are the outcomes in the sample space of part (c) equally likely?



19. **Expand Your Knowledge: Odds in Favor** Sometimes probability statements are expressed in terms of odds.

The *odds in favour* of an event A are the ratio $\frac{P(A)}{P(\text{not } A)} = \frac{P(A)}{P(A^c)}$.

For instance, if $P(A) = 0.60$, then $P(A^c) = 0.40$ and the odds in favor of A are

$$\frac{0.60}{0.40} = \frac{6}{4} = \frac{3}{2}, \text{ written as 3 to 2 or 3:2}$$

- (a) Show that if we are given the *odds in favor* of event A as $n:m$, the probability of event A is given by $P(A) = \frac{n}{n+m}$. Hint: Solve the equation $\frac{n}{m} = \frac{P(A)}{1 - P(A)}$ for $P(A)$.

- (b) A telemarketing supervisor tells a new worker that the odds of making a sale on a single call are 2 to 15. What is the probability of a successful call?
 (c) A sports announcer says that the odds a basketball player will make a free throw shot are 3 to 5. What is the probability the player will make the shot?



- 20.** *Expand Your Knowledge: Odds Against* Betting odds are usually stated against the event happening (against winning).

The *odds against* event W are the ratio $\frac{P(\text{not } W)}{P(W)} = \frac{P(W^c)}{P(W)}$.

In horse racing, the betting odds are based on the probability that the horse does *not* win.

- (a) Show that if we are given the *odds against* an event W as $a:b$, the probability of *not* W is $P(W^c) = \frac{a}{a+b}$. Hint: Solve the equation $\frac{a}{b} = \frac{P(W^c)}{1 - P(W^c)}$ for $P(W^c)$.
 (b) In a recent Kentucky Derby, the betting odds for the favorite horse, Point Given, were 9 to 5. Use these odds to compute the probability that Point Given would lose the race. What is the probability that Point Given would win the race?
 (c) In the same race, the betting odds for the horse Monarchs were 6 to 1. Use these odds to estimate the probability that Monarchs would lose the race. What is the probability that Monarchs would win the race?
 (d) Invisible Ink was a long shot, with betting odds of 30 to 1. Use these odds to estimate the probability that Invisible Ink would lose the race. What is the probability the horse would win the race? For further information on the Kentucky Derby, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Kentucky Derby.

- 21.** *Business: Customers* John runs a computer software store. Yesterday he counted 127 people who walked by his store, 58 of whom came into the store. Of the 58, only 25 bought something in the store.

- (a) Estimate the probability that a person who walks by the store will enter the store.
 (b) Estimate the probability that a person who walks into the store will buy something.
 (c) Estimate the probability that a person who walks by the store will come in *and* buy something.
 (d) Estimate the probability that a person who comes into the store will buy nothing.



SECTION 4.2

Some Probability Rules—Compound Events

FOCUS POINTS

- Compute probabilities of general compound events.
- Compute probabilities involving independent events or mutually exclusive events.
- Use survey results to compute conditional probabilities.

Conditional Probability and Multiplication Rules

You roll two dice. What is the probability that you will get a 5 on each die? Consider a collection of 6 balls identical in every way except in color. There are 3 green balls, 2 blue balls, and 1 red ball. You draw two balls at random from the collection *without replacing* the first ball before drawing the second. What is the probability that both balls will be green?

It seems that these two problems are nearly alike. They are alike in the sense that in each case, you are to find the probability of two events occurring *together*. In the first problem, you are to find

$$P(5 \text{ on 1st die and } 5 \text{ on 2nd die})$$

In the second, you want

$$P(\text{green ball on 1st draw and green ball on 2nd draw})$$

The two problems differ in one important aspect, however. In the dice problem, the outcome of a 5 on the first die does not have any effect on the probability of getting a 5 on the second die. Because of this, the events are *independent*.

Independent events

Two events are **independent** if the occurrence or nonoccurrence of one event does *not* change the probability that the other event will occur.

Dependent events

Probability of A and B

In the problem concerning a collection of colored balls, the probability that the first ball is green is $3/6$, since there are 6 balls in the collection and 3 of them are green. If you get a green ball on the first draw, the probability of getting a green ball on the second draw is changed to $2/5$, because one green ball has already been drawn and only 5 balls remain. Therefore, the two events in the ball-drawing problem are *not* independent. They are, in fact, *dependent*, since the outcome of the first draw changes the probability of getting a green ball on the second draw.

Why does the *independence* or *dependence* of two events matter? The type of events determines the way we compute the probability of the two events happening together. If two events A and B are *independent*, then we use formula (4) to compute the probability of the event A and B :

Multiplication rule for independent events

$$P(A \text{ and } B) = P(A) \cdot P(B) \quad (4)$$

Conditional probability
 $P(A | B)$

If the events are *dependent*, then we must take into account the changes in the probability of one event caused by the occurrence of the other event. The notation $P(A, \text{given } B)$ denotes the probability that event A will occur *given* that event B has occurred. This is called a *conditional probability*. We read $P(A, \text{given } B)$ as “probability of A given B .” If A and B are dependent events, then $P(A) \neq P(A, \text{given } B)$ because the occurrence of event B has changed the probability that event A will occur. A standard notation for $P(A, \text{given } B)$ is $P(A | B)$.

General multiplication rule for any events

$$P(A \text{ and } B) = P(A) \cdot P(B | A) \quad (5)$$

$$P(A \text{ and } B) = P(B) \cdot P(A | B) \quad (6)$$

Multiplication rules of probability

We will use either formula (5) or formula (6) according to the information available.

Formulas (4), (5), and (6) constitute the *multiplication rules of probability*. They help us compute the probability of events happening together when the sample space is too large for convenient reference or when it is not completely known.

Note: For conditional probability, observe that the multiplication rule

$$P(A \text{ and } B) = P(B) \cdot P(A | B)$$

can be solved for $P(A | B)$, leading to

Conditional probability (when $P(B) \neq 0$)

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

We will see some applications of this formula in later chapters.

Let's use the multiplication rules to complete the dice and ball-drawing problems just discussed. We'll compare the results with those obtained by using the sample space directly.

EXAMPLE 4**MULTIPLICATION RULE, INDEPENDENT EVENTS**

Suppose you are going to throw two fair dice. What is the probability of getting a 5 on each die?

SOLUTION USING THE MULTIPLICATION RULE: The two events are independent, so we should use formula (4). $P(5 \text{ on 1st die and } 5 \text{ on 2nd die}) = P(5 \text{ on 1st}) \cdot P(5 \text{ on 2nd})$. To finish the problem, we need to compute the probability of getting a 5 when we throw one die.

There are six faces on a die, and on a fair die each is equally likely to come up when you throw the die. Only one face has five dots, so by formula (2) for equally likely outcomes,

$$P(5 \text{ on die}) = \frac{1}{6}$$

Now we can complete the calculation.

$$\begin{aligned} P(5 \text{ on 1st die and } 5 \text{ on 2nd die}) &= P(5 \text{ on 1st}) \cdot P(5 \text{ on 2nd}) \\ &= \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \end{aligned}$$

SOLUTION USING SAMPLE SPACE: The first task is to write down the sample space. Each die has six equally likely outcomes, and each outcome of the second die can be paired with each of the first. The sample space is shown in Figure 4-2. The total number of outcomes is 36, and only one is favorable to a 5 on the first die and a 5 on the second. The 36 outcomes are equally likely, so by formula (2) for equally likely outcomes,

Philip Harvey/Spirit/Corbis

FIGURE 4-2

Sample Space for Two Dice

1st	2nd										
•	•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•	•

$$P(5 \text{ on 1st and } 5 \text{ on 2nd}) = \frac{1}{36}$$

The two methods yield the same result. The multiplication rule was easier to use because we did not need to look at all 36 outcomes in the sample space for tossing two dice.

EXAMPLE 5

MULTIPLICATION RULE, DEPENDENT EVENTS

Consider a collection of 6 balls that are identical except in color. There are 3 green balls, 2 blue balls, and 1 red ball. Compute the probability of drawing 2 green balls from the collection if the first ball is *not replaced* before the second ball is drawn.

MULTIPLICATION RULE METHOD: These events are *dependent*. The probability of a green ball on the first draw is $3/6$, but on the second draw the probability of a green ball is only $2/5$ if a green ball was removed on the first draw. By the multiplication rule for dependent events,

$$\begin{aligned} P(\text{green ball on 1st draw and green ball on 2nd draw}) &= P(\text{green on 1st}) \cdot P(\text{green on 2nd} | \text{green on 1st}) \\ &= \frac{3}{6} \cdot \frac{2}{5} = \frac{1}{5} = 0.2 \end{aligned}$$

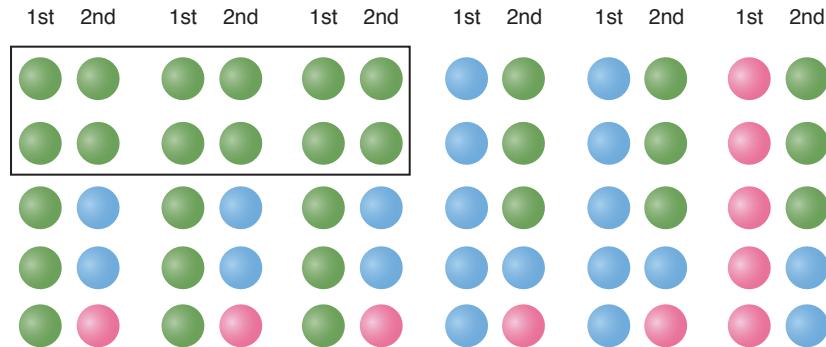
SAMPLE SPACE METHOD: Each of the 6 possible outcomes for the 1st draw must be paired with each of the 5 possible outcomes for the second draw. This means that there are a total of 30 possible pairs of balls. Figure 4-3 shows all the possible pairs of balls. In 6 of the pairs, both balls are green.

$$P(\text{green ball on 1st draw and green ball on 2nd draw}) = \frac{6}{30} = 0.2$$

Again, the two methods agree.

FIGURE 4-3

Sample Space of Drawing Two Balls Without Replacement from a Collection of 3 Green Balls, 2 Blue Balls, and 1 Red Ball



The multiplication rules apply whenever we wish to determine the probability of two events happening *together*. To indicate “together,” we use *and* between the events. But before you use a multiplication rule to compute the probability of *A and B*, you must determine if *A* and *B* are independent or dependent events.

PROCEDURE**HOW TO USE THE MULTIPLICATION RULES**

1. First determine whether A and B are independent events.
If $P(A) = P(A | B)$, then the events are independent.
2. If A and B are independent events,

$$P(A \text{ and } B) = P(A) \cdot P(B) \quad (4)$$

3. If A and B are any events,

$$P(A \text{ and } B) = P(A) \cdot P(B | A) \quad (5)$$

or $P(A \text{ and } B) = P(B) \cdot P(A | B)$ (6)

Let's practice using the multiplication rule.

GUIDED EXERCISE 4**Multiplication rule**

Andrew is 55, and the probability that he will be alive in 10 years is 0.72. Ellen is 35, and the probability that she will be alive in 10 years is 0.92. Assuming that the life span of one will have no effect on the life span of the other, what is the probability they will both be alive in 10 years?

- (a) Are these events dependent or independent?  Since the life span of one does not affect the life span of the other, the events are independent.
- (b) Use the appropriate multiplication rule to find $P(\text{Andrew alive in 10 years and Ellen alive in 10 years})$.  We use the rule for independent events:
- $$\begin{aligned} P(A \text{ and } B) &= P(A) \cdot P(B) \\ P(\text{Andrew alive and Ellen alive}) &= P(\text{Andrew alive}) \cdot P(\text{Ellen alive}) \\ &= (0.72)(0.92) \approx 0.66 \end{aligned}$$

GUIDED EXERCISE 5**Dependent events**

A quality-control procedure for testing Ready-Flash disposable cameras consists of drawing two cameras at random from each lot of 100 without replacing the first camera before drawing the second. If both are defective, the entire lot is rejected. Find the probability that both cameras are defective if the lot contains 10 defective cameras. Since we are drawing the cameras at random, assume that each camera in the lot has an equal chance of being drawn.

- (a) What is the probability of getting a defective camera on the first draw?  The sample space consists of all 100 cameras. Since each is equally likely to be drawn and there are 10 defective ones,
- $$P(\text{defective camera}) = \frac{10}{100} = \frac{1}{10}$$

Continued

GUIDED EXERCISE 5 *continued*

- (b) The first camera drawn is not replaced, so there are only 99 cameras for the second draw. What is the probability of getting a defective camera on the second draw if the first camera was defective?

→ If the first camera was defective, then there are only 9 defective cameras left among the 99 remaining cameras in the lot.

$$P(\text{def. cam. on 2nd draw} \mid \text{def. cam. on 1st})$$

$$= \frac{9}{99} = \frac{1}{11}$$

- (c) Are the probabilities computed in parts (a) and (b) different? Does drawing a defective camera on the first draw change the probability of getting a defective camera on the second draw? Are the events dependent?

→ The answer to all these questions is yes.

- (d) Use the formula for dependent events,

$$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$$

to compute $P(\text{1st camera defective and 2nd camera defective})$.

→ $P(\text{1st defective and 2nd defective}) = \frac{1}{10} \cdot \frac{1}{11}$

$$= \frac{1}{110}$$

$$\approx 0.009$$

More than two independent events

The multiplication rule for independent events extends to *more than two independent events*. For instance, if the probability that a single seed will germinate is 0.85, and you plant 3 seeds, the probability that they will all germinate (assuming seed germinations are independent) is

$$P(\text{1st germinates and 2nd germinates and 3rd germinates}) = (0.85)(0.85)(0.85)$$

$$\approx 0.614$$

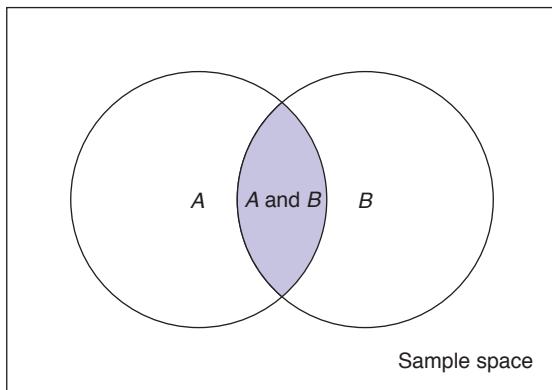
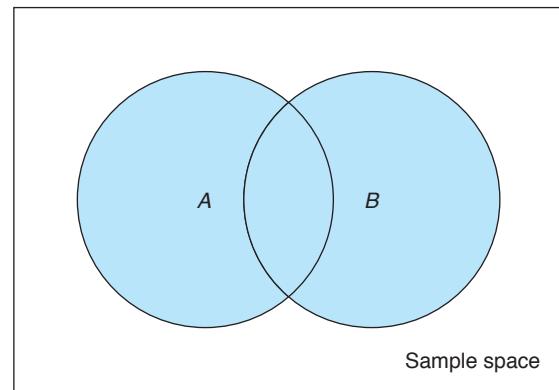
Addition Rules

One of the multiplication rules can be used any time we are trying to find the probability of two events happening *together*. Pictorially, we are looking for the probability of the shaded region in Figure 4-4(a).

Another way to combine events is to consider the possibility of one event *or* another occurring. For instance, if a sports car saleswoman gets an extra bonus if she sells a convertible or a car with leather upholstery, she is interested in the

Probability of *A or B*

FIGURE 4-4

(a) The Event *A and B*(b) The Event *A or B*

probability that you will buy a car that is a convertible *or* that has leather upholstery. Of course, if you bought a convertible with leather upholstery, that would be fine, too. Pictorially, the shaded portion of Figure 4-4(b) represents the outcomes satisfying the *or* condition. Notice that the condition *A or B* is satisfied by any one of the following conditions:

1. Any outcome in *A* occurs.
2. Any outcome in *B* occurs.
3. Any outcome in both *A* and *B* occurs.

It is important to distinguish between the *or* combinations and the *and* combinations because we apply different rules to compute their probabilities.

GUIDED EXERCISE 6

Combining events

Indicate how each of the following pairs of events are combined. Use either the *and* combination or the *or* combination.

- (a) Satisfying the humanities requirement by taking a course in the history of Japan or by taking a course in classical literature  Use the *or* combination.
- (b) Buying new tires and aligning the tires  Use the *and* combination.
- (c) Getting an A not only in psychology but also in biology  Use the *and* combination.
- (d) Having at least one of these pets: cat, dog, bird, rabbit  Use the *or* combination.

Once you decide that you are to find the probability of an *or* combination rather than an *and* combination, what formula do you use? It depends on the situation. In particular, it depends on whether or not the events being combined share any outcomes. Example 6 illustrates two situations.

EXAMPLE 6

PROBABILITY OF EVENTS COMBINED WITH *OR*

Consider an introductory statistics class with 31 students. The students range from freshmen through seniors. Some students are male and some are female. Figure 4-5 shows the sample space of the class.

FIGURE 4-5

Sample Space for Statistics Class

		F	M		
		Designates Female	Designates Male		
		Freshman	Sophomore	Junior	Senior
		FFFFF	FFF	FFF	F
		FFFF			
		MM	MM	MM	M
		MM	MM		
		15 students	8 students	6 students	2 students

- (a) Suppose we select one student at random from the class. Find the probability that the student is either a freshman or a sophomore.

Since there are 15 freshmen out of 31 students, $P(\text{freshmen}) = \frac{15}{31}$.

Since there are 8 sophomores out of 31 students, $P(\text{sophomore}) = \frac{8}{31}$.

$$P(\text{freshman or sophomore}) = \frac{15}{31} + \frac{8}{31} = \frac{23}{31} \approx 0.742$$

Notice that we can simply add the probability of freshman to the probability of sophomore to find the probability that a student selected at random will be either a freshman or a sophomore. No student can be both a freshman and a sophomore at the same time.

- (b) Select one student at random from the class. What is the probability that the student is either a male or a sophomore? Here we note that

$$P(\text{sophomore}) = \frac{8}{31} \quad P(\text{male}) = \frac{14}{31} \quad P(\text{sophomore and male}) = \frac{5}{31}$$

If we simply add $P(\text{sophomore})$ and $P(\text{male})$, we're including $P(\text{sophomore and male})$ twice in the sum. To compensate for this double summing, we simply subtract $P(\text{sophomore and male})$ from the sum. Therefore,

$$\begin{aligned} P(\text{sophomore or male}) &= P(\text{sophomore}) + P(\text{male}) - P(\text{sophomore and male}) \\ &= \frac{8}{31} + \frac{14}{31} - \frac{5}{31} = \frac{17}{31} \approx 0.548 \end{aligned}$$



Mutually exclusive events

We say the events A and B are *mutually exclusive* or *disjoint* if they cannot occur together. This means that A and B have no outcomes in common or, put another way, that $P(A \text{ and } B) = 0$.

Two events are **mutually exclusive** or **disjoint** if they cannot occur together. In particular, events A and B are mutually exclusive if $P(A \text{ and } B) = 0$.

Formula (7) is the *addition rule for mutually exclusive events A and B* .

Addition rules

Addition rule for mutually exclusive events A and B

$$P(A \text{ or } B) = P(A) + P(B) \tag{7}$$

If the events are not mutually exclusive, we must use the more general formula (8), which is the *general addition rule for any events A and B* .

General addition rule for any events A and B

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{8}$$

You may ask: Which formula should I use? The answer is: Use formula (7) only if you know that A and B are mutually exclusive (i.e., cannot occur together); if you do not know whether A and B are mutually exclusive, then use

formula (8). Formula (8) is valid either way. Notice that when A and B are mutually exclusive, then $P(A \text{ and } B) = 0$, so formula (8) reduces to formula (7).

PROCEDURE

HOW TO USE THE ADDITION RULES

1. First determine whether A and B are mutually exclusive events.

If $P(A \text{ and } B) = 0$, then the events are mutually exclusive.

2. If A and B are mutually exclusive events,

$$P(A \text{ or } B) = P(A) + P(B) \quad (7)$$

3. If A and B are any events,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad (8)$$

GUIDED EXERCISE 7

Mutually exclusive events

The Cost Less Clothing Store carries remainder pairs of slacks. If you buy a pair of slacks in your regular waist size without trying them on, the probability that the waist will be too tight is 0.30 and the probability that it will be too loose is 0.10.

- (a) Are the events “too tight” and “too loose” mutually exclusive?  The waist cannot be both too tight and too loose at the same time, so the events are mutually exclusive.
- (b) If you choose a pair of slacks at random in your regular waist size, what is the probability that the waist will be too tight or too loose?  Since the events are mutually exclusive,

$$\begin{aligned} P(\text{too tight or too loose}) &= P(\text{too tight}) + P(\text{too loose}) \\ &= 0.30 + 0.10 \\ &= 0.40 \end{aligned}$$

GUIDED EXERCISE 8

General addition rule

Professor Jackson is in charge of a program to prepare people for a high school equivalency exam. Records show that 80% of the students need work in math, 70% need work in English, and 55% need work in both areas.

- (a) Are the events “needs math” and “needs English” mutually exclusive?  These events are not mutually exclusive, since some students need both. In fact,

$$P(\text{needs math and needs English}) = 0.55$$
- (b) Use the appropriate formula to compute the probability that a student selected at random needs math or needs English.  Since the events are not mutually exclusive, we use formula (8):

$$\begin{aligned} P(\text{needs math or needs English}) &= P(\text{needs math}) + P(\text{needs English}) \\ &\quad - P(\text{needs math and English}) \\ &= 0.80 + 0.70 - 0.55 \\ &= 0.95 \end{aligned}$$

More than two mutually exclusive events

The addition rule for mutually exclusive events can be extended to apply to the situation in which we have more than two events, all of which are mutually exclusive to all the other events.

EXAMPLE 7

MUTUALLY EXCLUSIVE EVENTS

Laura is playing Monopoly. On her next move she needs to throw a sum bigger than 8 on the two dice in order to land on her own property and pass Go. What is the probability that Laura will roll a sum bigger than 8?

SOLUTION: When two dice are thrown, the largest sum that can come up is 12. Consequently, the only sums larger than 8 are 9, 10, 11, and 12. These outcomes are mutually exclusive, since only one of these sums can possibly occur on one throw of the dice. The probability of throwing more than 8 is the same as

$$P(9 \text{ or } 10 \text{ or } 11 \text{ or } 12)$$

Since the events are mutually exclusive,

$$\begin{aligned} P(9 \text{ or } 10 \text{ or } 11 \text{ or } 12) &= P(9) + P(10) + P(11) + P(12) \\ &= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} \\ &= \frac{10}{36} = \frac{5}{18} \end{aligned}$$

To get the specific values of $P(9)$, $P(10)$, $P(11)$, and $P(12)$, we used the sample space for throwing two dice (see Figure 4-2 on page 144). There are 36 equally likely outcomes—for example, those favorable to 9 are 6, 3; 3, 6; 5, 4; and 4, 5. So $P(9) = 4/36$. The other values can be computed in a similar way.

LOOKING FORWARD

Chapters 5 and 6 involve several probability distributions. In these chapters, most of the events of interest are mutually exclusive as well as independent. This means we can compute probabilities by using the extended addition rule for mutually exclusive events and the extended multiplication rule for independent events.

Further Examples

Most of us have been asked to participate in a survey. Schools, retail stores, news media, and government offices all conduct surveys. There are many types of surveys, and it is not our intention to give a general discussion of this topic. Let us study a very popular method called the *simple tally survey*. Such a survey consists of questions to which the responses can be recorded in the rows and columns of a table called a *contingency table*. These questions are appropriate to the information you want and are designed to cover the *entire* population of interest. In addition, the questions should be designed so that you can partition the sample space of responses into distinct (that is, mutually exclusive) sectors.

If the survey includes responses from a reasonably large random sample, then the results should be representative of your population. In this case, you can estimate simple probabilities, conditional probabilities, and the probabilities of some combinations of events directly from the results of the survey.

EXAMPLE 8**SURVEY**

At Hopewell Electronics, all 140 employees were asked about their political affiliations. The employees were grouped by type of work, as executives or production workers. The results with row and column totals are shown in Table 4-2.

TABLE 4-2 Employee Type and Political Affiliation

Employee Type	Political Affiliation			Row Total
	Democrat (D)	Republican (R)	Independent (I)	
Executive (E)	5	34	9	48
Production worker (PW)	63	21	8	92
Column Total	68	55	17	140
Total				Grand Total



Mark Richards / PhotoEdit

Suppose an employee is selected at random from the 140 Hopewell employees. Let us use the following notation to represent different events of choosing: E = executive; PW = production worker; D = Democrat; R = Republican; I = Independent.

(a) Compute $P(D)$ and $P(E)$.

SOLUTION: To find these probabilities, we look at the *entire* sample space.

$$P(D) = \frac{\text{Number of Democrats}}{\text{Number of employees}} = \frac{68}{140} \approx 0.486$$

$$P(E) = \frac{\text{Number of executives}}{\text{Number of employees}} = \frac{48}{140} \approx 0.343$$

(b) Compute $P(D | E)$.

SOLUTION: For the conditional probability, we restrict our attention to the portion of the sample space satisfying the condition of being an executive.

$$P(D | E) = \frac{\text{Number of executives who are Democrats}}{\text{Number of executives}} = \frac{5}{48} \approx 0.104$$

(c) Are the events D and E independent?

SOLUTION: One way to determine if the events D and E are independent is to see if $P(D) = P(D | E)$ [or equivalently, if $P(E) = P(E | D)$]. Since $P(D) \approx 0.486$ and $P(D | E) \approx 0.104$, we see that $P(D) \neq P(D | E)$. This means that the events D and E are *not* independent. The probability of event D “depends on” whether or not event E has occurred.

(d) Compute $P(D \text{ and } E)$.

SOLUTION: This probability is not conditional, so we must look at the entire sample space.

$$P(D \text{ and } E) = \frac{\text{Number of executives who are Democrats}}{\text{Total number of employees}} = \frac{5}{140} \approx 0.036$$

Let's recompute this probability using the rules of probability for dependent events.

$$P(D \text{ and } E) = P(E) \cdot P(D | E) = \frac{48}{140} \cdot \frac{5}{48} = \frac{5}{140} \approx 0.036$$

The results using the rules are consistent with those using the sample space.

- (e) Compute $P(D \text{ or } E)$.

SOLUTION: From part (d), we know that the events “Democrat” and “executive” are not mutually exclusive, because $P(D \text{ and } E) \neq 0$. Therefore,

$$\begin{aligned} P(D \text{ or } E) &= P(D) + P(E) - P(D \text{ and } E) \\ &= \frac{68}{140} + \frac{48}{140} - \frac{5}{140} = \frac{111}{140} \approx 0.793 \end{aligned}$$

GUIDED EXERCISE 9

Survey

Using Table 4-2 on page 152, let’s consider other probabilities regarding the types of employees at Hopewell and their political affiliations. This time let’s consider the production worker and the affiliation of Independent. Suppose an employee is selected at random from the group of 140.

- (a) Compute $P(I)$ and $P(PW)$.

$$\Rightarrow P(I) = \frac{\text{No. of Independents}}{\text{Total no. of employees}}$$

$$= \frac{17}{140} \approx 0.121$$

$$\begin{aligned} P(PW) &= \frac{\text{No. of production workers}}{\text{Total no. of employees}} \\ &= \frac{92}{140} \approx 0.657 \end{aligned}$$

- (b) Compute $P(I | PW)$. This is a conditional probability. Be sure to restrict your attention to production workers, since that is the condition given.

$$\Rightarrow P(I | PW) = \frac{\text{No. of Independent production workers}}{\text{No. of production workers}}$$

$$= \frac{8}{92} \approx 0.087$$

- (c) Compute $P(I \text{ and } PW)$. In this case, look at the entire sample space and the number of employees who are both Independent and in production.

$$\Rightarrow P(I \text{ and } PW) = \frac{\text{No. of Independent production workers}}{\text{Total no. employees}}$$

$$= \frac{8}{140} \approx 0.057$$

- (d) Use the multiplication rule for dependent events to calculate $P(I \text{ and } PW)$. Is the result the same as that of part (c)?

$$\Rightarrow \text{By the multiplication rule,}$$

$$P(I \text{ and } PW) = P(PW) \cdot P(I | PW)$$

$$= \frac{92}{140} \cdot \frac{8}{92} = \frac{8}{140} \approx 0.057$$

The results are the same.

- (e) Compute $P(I \text{ or } PW)$. Are the events mutually exclusive?

$$\Rightarrow \text{Since the events are not mutually exclusive,}$$

$$P(I \text{ or } PW) = P(I) + P(PW) - P(I \text{ and } PW)$$

$$= \frac{17}{140} + \frac{92}{140} - \frac{8}{140}$$

$$= \frac{101}{140} \approx 0.721$$

As you apply probability to various settings, keep the following rules in mind.

Basic probability rules

Summary of basic probability rules

A statistical experiment or statistical observation is any random activity that results in a recordable outcome. The sample space is the set of all simple events that are the outcomes of the statistical experiment and cannot be broken into other “simpler” events. A general event is any subset of the sample space. The notation $P(A)$ designates the probability of event A .

1. $P(\text{entire sample space}) = 1$
2. For any event A : $0 \leq P(A) \leq 1$
3. A^c designates the complement of A : $P(A^c) = 1 - P(A)$
4. Events A and B are independent events if $P(A) = P(A | B)$.
5. Multiplication Rules

General: $P(A \text{ and } B) = P(A) \cdot P(B | A)$

$P(A \text{ and } B) = P(B) \cdot P(A | B)$

Independent events: $P(A \text{ and } B) = P(A) \cdot P(B)$

6. Conditional Probability: $P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$

7. Events A and B are mutually exclusive if $P(A \text{ and } B) = 0$.

8. Addition Rules

General: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Mutually exclusive events: $P(A \text{ or } B) = P(A) + P(B)$

CRITICAL THINKING

Translating events described by common English phrases into events described using *and*, *or*, *complement*, or *given* takes a bit of care. Table 4-3 shows some common phrases and their corresponding translations into symbols.

TABLE 4-3 English Phrases and Corresponding Symbolic Translations

Consider the following events for a person selected at random from the general population:

A = person is taking college classes

B = person is under 30 years old

Phrase	Symbolic Expression
1. The probability that a person is under 30 years old and is taking college classes is 40%.	$P(B \text{ and } A) = 0.40$ or $P(A \text{ and } B) = 0.40$
2. The probability that a person under 30 years old is taking college classes is 45%.	$P(A B) = 0.45$
3. The probability is 45% that a person is taking college classes if the person is under 30.	$P(A B) = 0.45$
4. The probability that a person taking college classes is under 30 is 0.60.	$P(B A) = 0.60$
5. The probability that a person is not taking college classes or is under 30 years old is 0.75.	$P(A^c \text{ or } B) = 0.75$

In this section, we have studied some important rules that are valid in all probability spaces. The rules and definitions of probability are not only interesting, but also have extensive *applications* in our everyday lives. If you are inclined to continue your study of probability a little further, we recommend reading about *Bayes's theorem* in Appendix I. The Reverend Thomas Bayes (1702–1761) was an English mathematician who discovered an important relation for conditional probabilities.

VIEWPOINT

The Psychology of Odors

The Smell and Taste Treatment Research Foundation of Chicago collected data on the time required to complete a maze while subjects were smelling different scents. Data for this survey can be found by visiting the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and finding the link to DASL, the Carnegie Mellon University Data and Story Library. Once at the DASL site, select Data Subjects, then Psychology, and then Scents. You can estimate conditional probabilities regarding response times for smokers, nonsmokers, and types of scents.

SECTION 4.2 PROBLEMS

1. | **Statistical Literacy** If two events are mutually exclusive, can they occur concurrently? Explain.
2. | **Statistical Literacy** If two events A and B are independent and you know that $P(A) = 0.3$, what is the value of $P(A \mid B)$?
3. | **Basic Computation: Addition Rule** Given $P(A) = 0.3$ and $P(B) = 0.4$:
 - (a) If A and B are mutually exclusive events, compute $P(A \text{ or } B)$.
 - (b) If $P(A \text{ and } B) = 0.1$, compute $P(A \text{ or } B)$.
4. | **Basic Computation: Addition Rule** Given $P(A) = 0.7$ and $P(B) = 0.4$:
 - (a) Can events A and B be mutually exclusive? Explain.
 - (b) If $P(A \text{ and } B) = 0.2$, compute $P(A \text{ or } B)$.
5. | **Basic Computation: Multiplication Rule** Given $P(A) = 0.2$ and $P(B) = 0.4$:
 - (a) If A and B are independent events, compute $P(A \text{ and } B)$.
 - (b) If $P(A \mid B) = 0.1$, compute $P(A \text{ and } B)$.
6. | **Basic Computation: Multiplication Rule** Given $P(A) = 0.7$ and $P(B) = 0.8$:
 - (a) If A and B are independent events, compute $P(A \text{ and } B)$.
 - (b) If $P(B \mid A) = 0.3$, compute $P(A \text{ and } B)$.
7. | **Basic Computations: Rules of Probability** Given $P(A) = 0.2$; $P(B) = 0.5$, $P(A \mid B) = 0.3$:
 - (a) Compute $P(A \text{ and } B)$.
 - (b) Compute $P(A \text{ or } B)$.
8. | **Basic Computation: Rules of Probability** Given $P(A^c) = 0.8$, $P(B) = 0.3$, $P(B \mid A) = 0.2$:
 - (a) Compute $P(A \text{ and } B)$.
 - (b) Compute $P(A \text{ or } B)$.
9. | **Critical Thinking** Suppose two events A and B are mutually exclusive, with $P(A) \neq 0$ and $P(B) \neq 0$. By working through the following steps, you'll see why two mutually exclusive events are not independent.
 - (a) For mutually exclusive events, can event A occur if event B has occurred? What is the value of $P(A \mid B)$?
 - (b) Using the information from part (a), can you conclude that events A and B are *not* independent if they are mutually exclusive? Explain.

10. **Critical Thinking** Suppose two events A and B are independent, with $P(A) \neq 0$ and $P(B) \neq 0$. By working through the following steps, you'll see why two independent events are not mutually exclusive.

- (a) What formula is used to compute $P(A \text{ and } B)$? Is $P(A \text{ and } B) \neq 0$? Explain.
 (b) Using the information from part (a), can you conclude that events A and B are *not* mutually exclusive?

11. **Critical Thinking** Consider the following events for a driver selected at random from the general population:

$$\begin{aligned} A &= \text{driver is under 25 years old} \\ B &= \text{driver has received a speeding ticket} \end{aligned}$$

Translate each of the following phrases into symbols.

- (a) The probability the driver has received a speeding ticket and is under 25 years old
 (b) The probability a driver who is under 25 years old has received a speeding ticket
 (c) The probability a driver who has received a speeding ticket is 25 years old or older
 (d) The probability the driver is under 25 years old or has received a speeding ticket
 (e) The probability the driver has not received a speeding ticket or is under 25 years old

12. **Critical Thinking** Consider the following events for a college student selected at random:

$$\begin{aligned} A &= \text{student is female} \\ B &= \text{student is majoring in business} \end{aligned}$$

Translate each of the following phrases into symbols.

- (a) The probability the student is male or is majoring in business
 (b) The probability a female student is majoring in business
 (c) The probability a business major is female
 (d) The probability the student is female and is not majoring in business
 (e) The probability the student is female and is majoring in business

13. **General: Candy Colors** M&M plain candies come in various colors. According to the M&M/Mars Department of Consumer Affairs (link to the Mars company web site from the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase>), the distribution of colors for plain M&M candies is

Color	Purple	Yellow	Red	Orange	Green	Blue	Brown
Percentage	20%	20%	20%	10%	10%	10%	10%

Suppose you have a large bag of plain M&M candies and you choose one candy at random. Find

- (a) $P(\text{green candy or blue candy})$. Are these outcomes mutually exclusive? Why?
 (b) $P(\text{yellow candy or red candy})$. Are these outcomes mutually exclusive? Why?
 (c) $P(\text{not purple candy})$

14. **Environmental: Land Formations** Arches National Park is located in southern Utah. The park is famous for its beautiful desert landscape and its many natural sandstone arches. Park Ranger Edward McCarrick started an inventory (not yet complete) of natural arches within the park that have an opening of at least 3 feet. The following table is based on information taken from the book *Canyon Country Arches and Bridges* by F. A. Barnes. The height of the arch opening is rounded to the nearest foot.

Height of arch, feet	3–9	10–29	30–49	50–74	75 and higher
Number of arches in park	111	96	30	33	18

For an arch chosen at random in Arches National Park, use the preceding information to estimate the probability that the height of the arch opening is

- (a) 3 to 9 feet tall
 - (b) 30 feet or taller
 - (c) 3 to 49 feet tall
 - (d) 10 to 74 feet tall
 - (e) 75 feet or taller
15. **General: Roll Two Dice** You roll two fair dice, a green one and a red one.
- (a) Are the outcomes on the dice independent?
 - (b) Find $P(5 \text{ on green die and } 3 \text{ on red die})$.
 - (c) Find $P(3 \text{ on green die and } 5 \text{ on red die})$.
 - (d) Find $P((5 \text{ on green die and } 3 \text{ on red die}) \text{ or } (3 \text{ on green die and } 5 \text{ on red die}))$.
16. **General: Roll Two Dice** You roll two fair dice, a green one and a red one.
- (a) Are the outcomes on the dice independent?
 - (b) Find $P(1 \text{ on green die and } 2 \text{ on red die})$.
 - (c) Find $P(2 \text{ on green die and } 1 \text{ on red die})$.
 - (d) Find $P((1 \text{ on green die and } 2 \text{ on red die}) \text{ or } (2 \text{ on green die and } 1 \text{ on red die}))$.
17. **General: Roll Two Dice** You roll two fair dice, a green one and a red one.
- (a) What is the probability of getting a sum of 6?
 - (b) What is the probability of getting a sum of 4?
 - (c) What is the probability of getting a sum of 6 or 4? Are these outcomes mutually exclusive?
18. **General: Roll Two Dice** You roll two fair dice, a green one and a red one.
- (a) What is the probability of getting a sum of 7?
 - (b) What is the probability of getting a sum of 11?
 - (c) What is the probability of getting a sum of 7 or 11? Are these outcomes mutually exclusive?
- Problems 19–22 involve a standard deck of 52 playing cards. In such a deck of cards there are four suits of 13 cards each. The four suits are: hearts, diamonds, clubs, and spades. The 26 cards included in hearts and diamonds are red in color. The 26 cards included in clubs and spades are black in color. The 13 cards in each suit are: 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, and Ace. This means there are four Aces, four Kings, four Queens, four 10's, etc., down to four 2's in each deck.
19. **General: Deck of Cards** You draw two cards from a standard deck of 52 cards without replacing the first one before drawing the second.
- (a) Are the outcomes on the two cards independent? Why?
 - (b) Find $P(\text{Ace on 1st card and King on 2nd})$.
 - (c) Find $P(\text{King on 1st card and Ace on 2nd})$.
 - (d) Find the probability of drawing an Ace and a King in either order.
20. **General: Deck of Cards** You draw two cards from a standard deck of 52 cards without replacing the first one before drawing the second.
- (a) Are the outcomes on the two cards independent? Why?
 - (b) Find $P(3 \text{ on 1st card and } 10 \text{ on 2nd})$.
 - (c) Find $P(10 \text{ on 1st card and } 3 \text{ on 2nd})$.
 - (d) Find the probability of drawing a 10 and a 3 in either order.
21. **General: Deck of Cards** You draw two cards from a standard deck of 52 cards, but before you draw the second card, you put the first one back and reshuffle the deck.
- (a) Are the outcomes on the two cards independent? Why?
 - (b) Find $P(\text{Ace on 1st card and King on 2nd})$.
 - (c) Find $P(\text{King on 1st card and Ace on 2nd})$.
 - (d) Find the probability of drawing an Ace and a King in either order.

22. **General: Deck of Cards** You draw two cards from a standard deck of 52 cards, but before you draw the second card, you put the first one back and reshuffle the deck.
- Are the outcomes on the two cards independent? Why?
 - Find $P(3 \text{ on 1st card and } 10 \text{ on 2nd})$.
 - Find $P(10 \text{ on 1st card and } 3 \text{ on 2nd})$.
 - Find the probability of drawing a 10 and a 3 in either order.
23. **Marketing: Toys USA Today** gave the information shown in the table about ages of children receiving toys. The percentages represent all toys sold.
- | What is the probability that a toy is purchased for someone | Age (years) | Percentage of Toys |
|--|-------------|--------------------|
| (a) 6 years old or older? | 2 and under | 15% |
| (b) 12 years old or younger? | 3–5 | 22% |
| (c) between 6 and 12 years old? | 6–9 | 27% |
| (d) between 3 and 9 years old? | 10–12 | 14% |
| <i>Interpretation</i> A child between 10 and 12 years old looks at this probability distribution and asks, “Why are people more likely to buy toys for kids older than I am [13 and over] than for kids in my age group [10–12]?” How would you respond? | 13 and over | 22% |
24. **Health Care: Flu** Based on data from the *Statistical Abstract of the United States*, 112th Edition, only about 14% of senior citizens (65 years old or older) get the flu each year. However, about 24% of the people under 65 years old get the flu each year. In the general population, there are 12.5% senior citizens (65 years old or older).
- What is the probability that a person selected at random from the general population is a senior citizen who will get the flu this year?
 - What is the probability that a person selected at random from the general population is a person under age 65 who will get the flu this year?
 - Answer parts (a) and (b) for a community that is 95% senior citizens.
 - Answer parts (a) and (b) for a community that is 50% senior citizens.
25. **Focus Problem: Lie Detector Test** In this problem, you are asked to solve part of the Focus Problem at the beginning of this chapter. In his book *Chances: Risk and Odds in Everyday Life*, James Burke says that there is a 72% chance a polygraph test (lie detector test) will catch a person who is, in fact, lying. Furthermore, there is approximately a 7% chance that the polygraph will falsely accuse someone of lying.
- Suppose a person answers 90% of a long battery of questions truthfully. What percentage of the answers will the polygraph *wrongly* indicate are lies?
 - Suppose a person answers 10% of a long battery of questions with lies. What percentage of the answers will the polygraph *correctly* indicate are lies?
 - Repeat parts (a) and (b) if 50% of the questions are answered truthfully and 50% are answered with lies.
 - Repeat parts (a) and (b) if 15% of the questions are answered truthfully and the rest are answered with lies.
26. **Focus Problem: Expand Your Knowledge** This problem continues the Focus Problem. The solution involves applying several basic probability rules and a little algebra to solve an equation.
- If the polygraph of Problem 25 indicated that 30% of the questions were answered with lies, what would you estimate for the actual percentage of lies in the answers? Hint: Let $B =$ event detector indicates a lie. We are given $P(B) = 0.30$. Let $A =$ event person is lying, so $A^c =$ event person is not lying. Then
- $$P(B) = P(A \text{ and } B) + P(A^c \text{ and } B)$$
- $$P(B) = P(A)P(B | A) + P(A^c)P(B | A^c)$$





Replacing $P(A^c)$ by $1 - P(A)$ gives

$$P(B) = P(A) \cdot P(B | A) + [1 - P(A)] \cdot P(B | A^c)$$

Substitute known values for $P(B)$, $P(B | A)$, and $P(B | A^c)$ into the last equation and solve for $P(A)$.

- (b) If the polygraph indicated that 70% of the questions were answered with lies, what would you estimate for the actual percentage of lies?
27. **Survey: Sales Approach** In a sales effectiveness seminar, a group of sales representatives tried two approaches to selling a customer a new automobile: the aggressive approach and the passive approach. For 1160 customers, the following record was kept:

	Sale	No Sale	Row Total
Aggressive	270	310	580
Passive	416	164	580
Column Total	686	474	1160

Suppose a customer is selected at random from the 1160 participating customers. Let us use the following notation for events: A = aggressive approach, Pa = passive approach, S = sale, N = no sale. So, $P(A)$ is the probability that an aggressive approach was used, and so on.

- (a) Compute $P(S)$, $P(S | A)$, and $P(S | Pa)$.
 (b) Are the events S = sale and Pa = passive approach independent? Explain.
 (c) Compute $P(A \text{ and } S)$ and $P(Pa \text{ and } S)$.
 (d) Compute $P(N)$ and $P(N | A)$.
 (e) Are the events N = no sale and A = aggressive approach independent? Explain.
 (f) Compute $P(A \text{ or } S)$.

28. **Survey: Medical Tests** Diagnostic tests of medical conditions can have several types of results. The test result can be positive or negative, whether or not a patient has the condition. A positive test (+) indicates that the patient has the condition. A negative test (−) indicates that the patient does not have the condition. Remember, a positive test does not prove that the patient has the condition. Additional medical work may be required. Consider a random sample of 200 patients, some of whom have a medical condition and some of whom do not. Results of a new diagnostic test for the condition are shown.

	Condition Present	Condition Absent	Row Total
Test Result +	110	20	130
Test Result −	20	50	70
Column Total	130	70	200

Assume the sample is representative of the entire population. For a person selected at random, compute the following probabilities:

- (a) $P(+ | \text{condition present})$; this is known as the *sensitivity* of a test.
 (b) $P(- | \text{condition present})$; this is known as the *false-negative rate*.
 (c) $P(- | \text{condition absent})$; this is known as the *specificity* of a test.
 (d) $P(+ | \text{condition absent})$; this is known as the *false-positive rate*.
 (e) $P(\text{condition present and } +)$; this is the *predictive value* of the test.
 (f) $P(\text{condition present and } -)$.

29. **Survey: Lung/Heart** In an article entitled “Diagnostic accuracy of fever as a measure of postoperative pulmonary complications” (*Heart Lung*, Vol. 10, No. 1, p. 61), J. Roberts and colleagues discuss using a fever of 38°C or higher as a diagnostic

indicator of postoperative atelectasis (collapse of the lung) as evidenced by x-ray observation. For fever $\geq 38^\circ\text{C}$ as the diagnostic test, the results for postoperative patients are

	Condition Present	Condition Absent	Row Total
Test Result +	72	37	109
Test Result -	82	79	161
Column Total	154	116	270

For the meaning of + and -, see Problem 28. Complete parts (a) through (f) from Problem 28.

30. **Survey: Customer Loyalty** Are customers more loyal in the east or in the west? The following table is based on information from *Trends in the United States*, published by the Food Marketing Institute, Washington, D.C. The columns represent length of customer loyalty (in years) at a primary supermarket. The rows represent regions of the United States.

	Less Than 1 Year	1–2 Years	3–4 Years	5–9 Years	10–14 Years	15 or More Years	Row Total
East	32	54	59	112	77	118	452
Midwest	31	68	68	120	63	173	523
South	53	92	93	158	106	158	660
West	41	56	67	78	45	86	373
Column Total	157	270	287	468	291	535	2008

What is the probability that a customer chosen at random

- (a) has been loyal 10 to 14 years?
- (b) has been loyal 10 to 14 years, given that he or she is from the east?
- (c) has been loyal *at least* 10 years?
- (d) has been loyal *at least* 10 years, given that he or she is from the west?
- (e) is from the west, given that he or she has been loyal less than 1 year?
- (f) is from the south, given that he or she has been loyal less than 1 year?
- (g) has been loyal *1 or more* years, given that he or she is from the east?
- (h) has been loyal *1 or more* years, given that he or she is from the west?
- (i) Are the events “from the east” and “loyal 15 or more years” independent? Explain.

31. **Franchise Stores: Profits** Wing Foot is a shoe franchise commonly found in shopping centers across the United States. Wing Foot knows that its stores will not show a profit unless they gross over \$940,000 per year. Let A be the event that a new Wing Foot store grosses over \$940,000 its first year. Let B be the event that a store grosses over \$940,000 its second year. Wing Foot has an administrative policy of closing a new store if it does not show a profit in *either* of the first 2 years. The accounting office at Wing Foot provided the following information: 65% of *all* Wing Foot stores show a profit the first year; 71% of *all* Wing Foot stores show a profit the second year (this includes stores that did not show a profit the first year); however, 87% of Wing Foot stores that showed a profit the first year also showed a profit the second year. Compute the following:

- (a) $P(A)$
- (b) $P(B)$
- (c) $P(B | A)$
- (d) $P(A \text{ and } B)$
- (e) $P(A \text{ or } B)$

- (f) What is the probability that a new Wing Foot store will not be closed after 2 years? What is the probability that a new Wing Foot store will be closed after 2 years?
32. **Education: College of Nursing** At Litchfield College of Nursing, 85% of incoming freshmen nursing students are female and 15% are male. Recent records indicate that 70% of the entering female students will graduate with a BSN degree, while 90% of the male students will obtain a BSN degree. If an incoming freshman nursing student is selected at random, find
- $P(\text{student will graduate} \mid \text{student is female})$.
 - $P(\text{student will graduate and student is female})$.
 - $P(\text{student will graduate} \mid \text{student is male})$.
 - $P(\text{student will graduate and student is male})$.
 - $P(\text{student will graduate})$. Note that those who will graduate are either males who will graduate or females who will graduate.
 - The events described by the phrases “will graduate and is female” and “will graduate, given female” seem to be describing the same students. Why are the probabilities $P(\text{will graduate and is female})$ and $P(\text{will graduate} \mid \text{female})$ different?
33. **Medical: Tuberculosis** The state medical school has discovered a new test for tuberculosis. (If the test indicates a person has tuberculosis, the test is positive.) Experimentation has shown that the probability of a positive test is 0.82, given that a person has tuberculosis. The probability is 0.09 that the test registers positive, given that the person does not have tuberculosis. Assume that in the general population, the probability that a person has tuberculosis is 0.04. What is the probability that a person chosen at random will
- have tuberculosis and have a positive test?
 - not have tuberculosis?
 - not have tuberculosis and have a positive test?
34. **Therapy: Alcohol Recovery** The Eastmore Program is a special program to help alcoholics. In the Eastmore Program, an alcoholic lives at home but undergoes a two-phase treatment plan. Phase I is an intensive group-therapy program lasting 10 weeks. Phase II is a long-term counseling program lasting 1 year. Eastmore Programs are located in most major cities, and past data gave the following information based on percentages of success and failure collected over a long period of time: The probability that a client will have a relapse in phase I is 0.27; the probability that a client will have a relapse in phase II is 0.23. However, if a client did not have a relapse in phase I, then the probability that this client will not have a relapse in phase II is 0.95. If a client did have a relapse in phase I, then the probability that this client will have a relapse in phase II is 0.70. Let A be the event that a client has a relapse in phase I and B be the event that a client has a relapse in phase II. Let C be the event that a client has no relapse in phase I and D be the event that a client has no relapse in phase II. Compute the following:
- $P(A)$, $P(B)$, $P(C)$, and $P(D)$
 - $P(B \mid A)$ and $P(D \mid C)$
 - $P(A \text{ and } B)$ and $P(C \text{ and } D)$
 - $P(A \text{ or } B)$
 - What is the probability that a client will go through both phase I and phase II without a relapse?
 - What is the probability that a client will have a relapse in both phase I and phase II?
 - What is the probability that a client will have a relapse in either phase I or phase II?

Brain Teasers Assume A and B are events such that $0 < P(A) < 1$ and $0 < P(B) < 1$. Answer questions 35–49 true or false and give a brief explanation for each answer. Hint: Review the summary of basic probability rules.

35. | $P(A \text{ and } A^c) = 0$
36. | $P(A \text{ or } A^c) = 1$
37. | $P(A \mid A^c) = 1$
38. | $P(A \text{ or } B) = P(A) + P(B)$
39. | $P(A \mid B) \geq P(A \text{ and } B)$
40. | $P(A \text{ or } B) \geq P(A)$ if A and B are independent events
41. | $P(A \text{ and } B) \leq P(A)$
42. | $P(A \mid B) > P(A)$ if A and B are independent events
43. | $P(A^c \text{ and } B^c) \leq 1 - P(A)$
44. | $P(A^c \text{ or } B^c) \leq 2 - P(A) - P(B)$
45. | If A and B are independent events, they must also be mutually exclusive events.
46. | If A and B are mutually exclusive, they must also be independent.
47. | If A and B are both mutually exclusive and independent, then at least one of $P(A)$ or $P(B)$ must be zero.
48. | If A and B are mutually exclusive, then $P(A \mid B) = 0$.
49. | $P(A \mid B) + P(A^c \mid B) = 1$
50. | **Brain Teaser** The Reverend Thomas Bayes (1702–1761) was an English mathematician who discovered an important rule of probability (see Bayes's theorem, Appendix I part I). A key feature of Bayes's theorem is the formula

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid A^c) \cdot P(A^c)$$
Explain why this formula is valid. Hint: See Figure A1-1 in Appendix I.

SECTION 4.3

Trees and Counting Techniques

FOCUS POINTS

- Organize outcomes in a sample space using tree diagrams.
- Compute number of ordered arrangements of outcomes using permutations.
- Compute number of (nonordered) groupings of outcomes using combinations.
- Explain how counting techniques relate to probability in everyday life.

When outcomes are equally likely, we compute the probability of an event by using the formula

$$P(A) = \frac{\text{Number of outcomes favorable to the event } A}{\text{Number of outcomes in the sample space}}$$

The probability formula requires that we be able to determine the number of outcomes in the sample space. In the problems we have done in previous sections, this task has not been difficult because the number of outcomes was small or the sample space consisted of fairly straightforward events. The tools we present in this section will help you count the number of possible outcomes in larger sample spaces or those formed by more complicated events.

When an outcome of an experiment is composed of a series of events, the multiplication rule gives us the *total number* of outcomes.

Multiplication rule of counting**Multiplication rule of counting**

Consider the series of events E_1 through E_m , where n_1 is the number of possible outcomes for event E_1 , n_2 is the number of possible outcomes for event E_2 , and n_m designates the number of possible outcomes for event E_m . Then the product

$$n_1 \times n_2 \times \cdots \times n_m$$

gives the total number of possible outcomes for the series of events E_1 , followed by E_2 , up through event E_m .

EXAMPLE 9**MULTIPLICATION RULE**

Jacqueline is in a nursing program and is required to take a course in psychology and one in physiology (A and P) next semester. She also wants to take Spanish II. If there are two sections of psychology, two of A and P, and three of Spanish II, how many different class schedules can Jacqueline choose from? (Assume that the times of the sections do not conflict.)

SOLUTION: Creating a class schedule can be considered an experiment with a series of three events. There are two possible outcomes for the psychology section, two for the A and P section, and three for the Spanish II section. By the multiplication rule, the total number of class schedules possible is

$$2 \times 2 \times 3 = 12$$

Tree diagram

A *tree diagram* gives a visual display of the total number of outcomes of an experiment consisting of a series of events. From a tree diagram, we can determine not only the total number of outcomes, but also the individual outcomes.

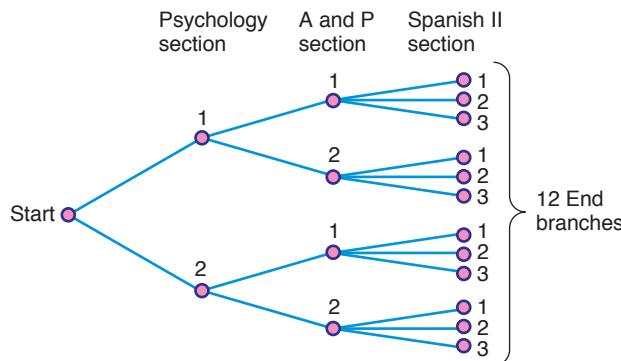
EXAMPLE 10**TREE DIAGRAM**

Using the information from Example 9, let's make a tree diagram that shows all the possible course schedules for Jacqueline.

SOLUTION: Figure 4-6 shows the tree diagram. Let's study the diagram. There are two branches from Start. These branches indicate the two possible choices for psychology sections. No matter which section of psychology Jacqueline chooses, she can choose from the two available A and P sections. Therefore, we have two

FIGURE 4-6

Tree Diagram for Selecting Class Schedules



branches leading from *each* psychology branch. Finally, after the psychology and A and P sections are selected, there are three choices for Spanish II. That is why there are three branches from *each* A and P section.

The tree ends with a total of 12 branches. The number of end branches tells us the number of possible schedules. The outcomes themselves can be listed from the tree by following each series of branches from Start to End. For instance, the top branch from Start generates the schedules shown in Table 4-4. The other six schedules can be listed in a similar manner, except they begin with the second section of psychology.

TABLE 4-4 Schedules Utilizing Section 1 of Psychology

Psychology Section	A and P Section	Spanish II Section
1	1	1
1	1	2
1	1	3
1	2	1
1	2	2
1	2	3

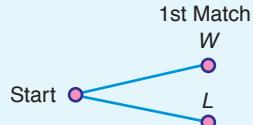
GUIDED EXERCISE 10

Tree diagram and multiplication rule

Louis plays three tennis matches. Use a tree diagram to list the possible win and loss sequences Louis can experience for the set of three matches.

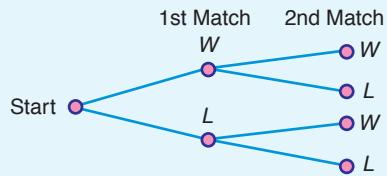
- (a) On the first match Louis can win or lose. From Start, indicate these two branches.

FIGURE 4-7 $W = \text{Win}$, $L = \text{Lose}$



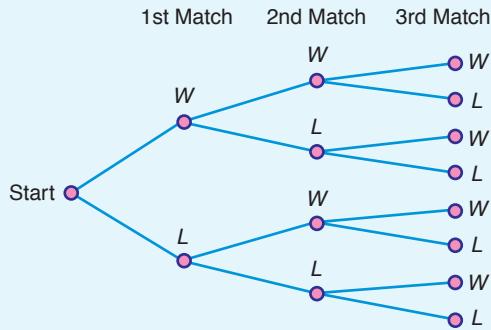
- (b) Regardless of whether Louis wins or loses the first match, he plays the second and can again win or lose. Attach branches representing these two outcomes to *each* of the first match results.

FIGURE 4-8



- (c) Louis may win or lose the third match. Attach branches representing these two outcomes to *each* of the second match results.

FIGURE 4-9



Continued

GUIDED EXERCISE 10 *continued*

- (d) How many possible win–lose sequences are there for the three matches?  Since there are eight branches at the end, there are eight sequences.
- (e) Complete this list of win–lose sequences.  The last four sequences all involve a loss on Match 1.
- | 1st | 2nd | 3rd | 1st | 2nd | 3rd |
|------|------|------|------|------|------|
| W | W | W | L | W | W |
| W | W | L | L | W | L |
| W | L | W | L | L | W |
| W | L | L | L | L | L |
| ____ | ____ | ____ | ____ | ____ | ____ |
| ____ | ____ | ____ | ____ | ____ | ____ |
| ____ | ____ | ____ | ____ | ____ | ____ |
| ____ | ____ | ____ | ____ | ____ | ____ |
- (f) Use the multiplication rule to compute the total number of outcomes for the three matches.  The number of outcomes for a series of three events, each with two outcomes, is

$$2 \times 2 \times 2 = 8$$

Tree diagrams help us display the outcomes of an experiment involving several stages. If we label each branch of the tree with an appropriate probability, we can use the tree diagram to help us compute the probability of an outcome displayed on the tree. One of the easiest ways to illustrate this feature of tree diagrams is to use an experiment of drawing balls out of an urn. We do this in the next example.

 EXAMPLE 11

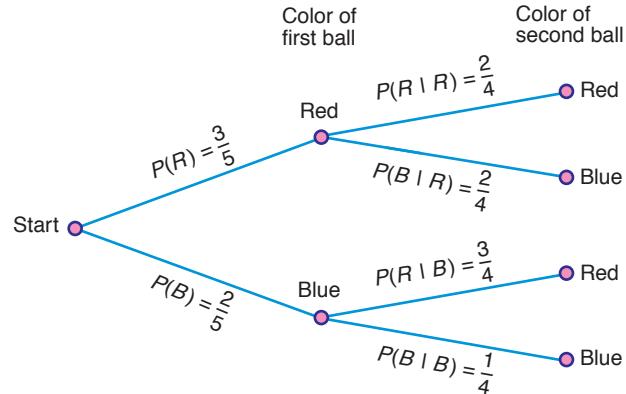
TREE DIAGRAM AND PROBABILITY

Suppose there are five balls in an urn. They are identical except for color. Three of the balls are red and two are blue. You are instructed to draw out one ball, note its color, and set it aside. Then you are to draw out another ball and note its color. What are the outcomes of the experiment? What is the probability of each outcome?

SOLUTION: The tree diagram in Figure 4-10 will help us answer these questions. Notice that since you did not replace the first ball before drawing the second one, the two stages of the experiment are dependent. The probability associated with

FIGURE 4-10

Tree Diagram for Urn Experiment



the color of the second ball depends on the color of the first ball. For instance, on the top branches, the color of the first ball drawn is red, so we compute the probabilities of the colors on the second ball accordingly. The tree diagram helps us organize the probabilities.

From the diagram, we see that there are four possible outcomes to the experiment. They are

$$RR = \text{red on 1st and red on 2nd}$$

$$RB = \text{red on 1st and blue on 2nd}$$

$$BR = \text{blue on 1st and red on 2nd}$$

$$BB = \text{blue on 1st and blue on 2nd}$$

To compute the probability of each outcome, we will use the multiplication rule for dependent events. As we follow the branches for each outcome, we will find the necessary probabilities.

$$P(\text{R on 1st and R on 2nd}) = P(R) \cdot P(R | R) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$$

$$P(\text{R on 1st and B on 2nd}) = P(R) \cdot P(B | R) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$$

$$P(\text{B on 1st and R on 2nd}) = P(B) \cdot P(R | B) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}$$

$$P(\text{B on 1st and B on 2nd}) = P(B) \cdot P(B | B) = \frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$$

Notice that the probabilities of the outcomes in the sample space add to 1, as they should.



Sometimes when we consider n items, we need to know the number of different *ordered arrangements* of the n items that are possible. The multiplication rules can help us find the number of possible ordered arrangements. Let's consider the classic example of determining the number of different ways in which eight people can be seated at a dinner table. For the first chair at the head of the table, there are eight choices. For the second chair, there are seven choices, since one person is already seated. For the third chair, there are six choices, since two people are already seated. By the time we get to the last chair, there is only one person left for that seat. We can view each arrangement as an outcome of a series of eight events. Event 1 is *fill the first chair*, event 2 is *fill the second chair*, and so forth. The multiplication rule will tell us the number of different outcomes.

Choices for	1st	2nd	3rd	4th	5th	6th	7th	8th	Chair position
	↓	↓	↓	↓	↓	↓	↓	↓	(8) (7) (6) (5) (4) (3) (2) (1) = 40,320

In all, there are 40,320 different seating arrangements for eight people. It is no wonder that it takes a little time to seat guests at a dinner table!

The multiplication pattern shown above is not unusual. In fact, it is an example of the multiplication indicated by the *factorial notation* $8!$.

$!$ is read “factorial”

$8!$ is read “8 factorial”

$$8! = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

In general, $n!$ indicates the product of n with each of the positive counting numbers less than n . By *special definition*, $0! = 1$.

Factorial notation

For a counting number n ,

$$0! = 1$$

$$1! = 1$$

$$n! = n(n - 1)(n - 2) \cdots 1$$

GUIDED EXERCISE 11**Factorial**

(a) Evaluate $3!$.



$$3! = 3 \cdot 2 \cdot 1 = 6$$

(b) In how many different ways can three objects be arranged in order? How many choices do you have for the first position? for the second position? for the third position?



You have three choices for the first position, two for the second position, and one for the third position. By the multiplication rule, you have

$$(3)(2)(1) = 3! = 6 \text{ arrangements}$$

Permutations rule

We have considered the number of ordered arrangements of n objects taken as an entire group. But what if we don't arrange the entire group? Specifically, we considered a dinner party for eight and found the number of ordered seating arrangements for all eight people. However, suppose you have an open house and have only five chairs. How many ways can five of the eight people seat themselves in the chairs? The formula we use to compute this number is called the *permutation formula*. As we see in the next example, the *permutations rule* is really another version of the multiplication rule.

Counting rule for permutations

The number of ways to *arrange in order* n distinct objects, taking them r at a time, is

$$P_{n,r} = \frac{n!}{(n - r)!} \quad (9)$$

where n and r are whole numbers and $n \geq r$. Another commonly used notation for permutations is nPr .

EXAMPLE 12**PERMUTATIONS RULE**

Let's compute the number of possible ordered seating arrangements for eight people in five chairs.

SOLUTION: In this case, we are considering a total of $n = 8$ different people, and we wish to arrange $r = 5$ of these people. Substituting into formula (9), we have

$$\begin{aligned} P_{n,r} &= \frac{n!}{(n - r)!} \\ P_{8,5} &= \frac{8!}{(8 - 5)!} = \frac{8!}{3!} = \frac{40,320}{6} = 6720 \end{aligned}$$

Using the multiplication rule, we get the same results

$$\begin{array}{ccccccccc} \text{Chair} & 1 & 2 & 3 & 4 & 5 \\ \text{Choices for} & 8 & \times & 7 & \times & 6 & \times & 5 & \times & 4 = 6720 \end{array}$$

The permutations rule has the advantage of using factorials. Most scientific calculators have a factorial key (!) as well as a permutations key (nPr) (see Tech Notes).



TECH NOTES

Most scientific calculators have a factorial key, often designated $x!$ or $n!$. Many of these same calculators have the permutation function built in, often labeled nPr. They also have the combination function, which is discussed next. The combination function is often labeled nCr.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) The factorial, permutation, and combination functions are all under MATH, then PRB.

Excel 2007 Click on the insert function $\boxed{f_x}$, then select all for the category. Fact gives factorials, Permut gives permutations, and Combin gives combinations.



In each of our previous counting formulas, we have taken the *order* of the objects or people into account. However, suppose that in your political science class you are given a list of 10 books. You are to select 4 to read during the semester. The order in which you read the books is not important. We are interested in the *different groupings* or *combinations* of 4 books from among the 10 on the list. The next formula tells us how to compute the number of different combinations.

Counting rule for combinations

The number of *combinations* of n objects taken r at a time is

$$C_{n,r} = \frac{n!}{r!(n-r)!} \quad (10)$$

where n and r are whole numbers and $n \geq r$. Other commonly used notations for combinations include nCr and $\binom{n}{r}$.

Notice the difference between the concepts of permutations and combinations. When we consider permutations, we are considering groupings *and order*. When we consider combinations, we are considering only the number of different groupings. For combinations, order within the groupings is not considered. As a result, the number of combinations of n objects taken r at a time is generally smaller than the number of permutations of the same n objects taken r at a time. In fact, the combinations formula is simply the permutations formula with the number of permutations of each distinct group divided out. In the formula for combinations, notice the factor of $r!$ in the denominator.

Combinations rule

Now let's look at an example in which we use the *combinations rule* to compute the number of *combinations* of 10 books taken 4 at a time.



EXAMPLE 13

COMBINATIONS

In your political science class, you are assigned to read any 4 books from a list of 10 books. How many different groups of 4 are available from the list of 10?

SOLUTION: In this case, we are interested in *combinations*, rather than permutations, of 10 books taken 4 at a time. Using $n = 10$ and $r = 4$, we have

$$C_{n,r} = \frac{n!}{r!(n-r)!} = \frac{10!}{4!(10-4)!} = 210$$

There are 210 different groups of 4 books that can be selected from the list of 10. An alternate solution method is to use the combinations key (often nCr or $C_{n,r}$) on a calculator.

LOOKING FORWARD

We will see the combinations rule again in Section 5.2 when we discuss the formula for the binomial probability distribution.

PROCEDURE

HOW TO DETERMINE THE NUMBER OF OUTCOMES OF AN EXPERIMENT

1. If the experiment consists of a series of stages with various outcomes, use the multiplication rule or a tree diagram.
2. If the outcomes consist of ordered subgroups of r items taken from a group of n items, use the permutations rule, $P_{n,r}$.

$$P_{n,r} = \frac{n!}{(n-r)!} \quad (9)$$

3. If the outcomes consist of nonordered subgroups of r items taken from a group of n items, use the combinations rule, $C_{n,r}$.

$$C_{n,r} = \frac{n!}{r!(n-r)!} \quad (10)$$

GUIDED EXERCISE 12

Permutations and combinations

The board of directors at Belford Community Hospital has 12 members.

- (i) Three officers—president, vice president, and treasurer—must be elected from among the members. How many different slates of officers are possible? We will view a slate of officers as a list of three people, with the president listed first, the vice president listed second, and the treasurer listed third. For instance, if Mr. Acosta, Ms. Hill, and Mr. Smith wish to be on a slate together, there are several different slates possible, depending on the person listed for each office. Not only are we asking for the number of different groups of three names for a slate, we are also concerned about order.

- (a) Do we use the permutations rule or the combinations rule? What is the value of n ? What is the value of r ?



We use the permutations rule, since order is important. The size of the group from which the slates of officers are to be selected is n . The size of each slate is r .

$$n = 12 \text{ and } r = 3$$

- (b) Use the permutations rule with $n = 12$ and $r = 3$ to compute $P_{12,3}$.



$$P_{n,r} = \frac{n!}{(n-r)!} = \frac{12!}{(12-3)!} = 1320$$

An alternative is to use the permutations key on a calculator.

Continued

GUIDED EXERCISE 12 *continued*

(ii) Three members from the group of 12 on the board of directors at Belford Community Hospital will be selected to go to a convention (all expenses paid) in Hawaii. How many different groups of 3 are there?

(c) Do we use the permutations rule or the combinations rule? What is the value of n ? What is the value of r ?

(d) Use the combinations rule with $n = 12$ and $r = 3$ to compute $C_{12,3}$.



We use the combinations rule, because order is not important. The size of the board is $n = 12$ and the size of each group going to the convention is $r = 3$.



$$C_{n,r} = \frac{n!}{r!(n-r)!} = \frac{12!}{3!(12-3)!} = 220$$

An alternative is to use the combinations key on a calculator.

VIEWPOINT**Powerball**

Powerball is a multistate lottery game that consists of drawing five distinct whole numbers from the numbers 1 through 59 in any order. Then one more number from the numbers 1 through 39 is selected as the Powerball number (this number could be one of the original five). Powerball numbers are drawn every Wednesday and Saturday. If you match all six numbers, you win the jackpot, which is worth at least \$10 million. Use methods of this section to show that there are 195,249,054 possible Powerball plays. For more information about the game of Powerball and the probability of winning different prizes, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Multi-State Lottery Association. Then select Powerball.

**SECTION 4.3
PROBLEMS**

1. **Statistical Literacy** What is the main difference between a situation in which the use of the permutations rule is appropriate and one in which the use of the combinations rule is appropriate?
2. **Statistical Literacy** Consider a series of events. How does a tree diagram help you list all the possible outcomes of a series of events? How can you use a tree diagram to determine the total number of outcomes of a series of events?
3. **Critical Thinking** For each of the following situations, explain why the combinations rule or the permutations rule should be used.
 - (a) Determine the number of different groups of 5 items that can be selected from 12 distinct items.
 - (b) Determine the number of different arrangements of 5 items that can be selected from 12 distinct items.
4. **Critical Thinking** You need to know the number of different arrangements possible for five distinct letters. You decide to use the permutations rule, but your friend tells you to use $5!$. Who is correct? Explain.
5. **Tree Diagram**
 - (a) Draw a tree diagram to display all the possible head-tail sequences that can occur when you flip a coin three times.
 - (b) How many sequences contain exactly two heads?
 - (c) **Probability Extension** Assuming the sequences are all equally likely, what is the probability that you will get exactly two heads when you toss a coin three times?

6. | **Tree Diagram**

- (a) Draw a tree diagram to display all the possible outcomes that can occur when you flip a coin and then toss a die.
- (b) How many outcomes contain a head and a number greater than 4?
- (c) *Probability Extension* Assuming the outcomes displayed in the tree diagram are all equally likely, what is the probability that you will get a head *and* a number greater than 4 when you flip a coin and toss a die?

7. | **Tree Diagram** There are six balls in an urn. They are identical except for color. Two are red, three are blue, and one is yellow. You are to draw a ball from the urn, note its color, and set it aside. Then you are to draw another ball from the urn and note its color.

- (a) Make a tree diagram to show all possible outcomes of the experiment. Label the probability associated with each stage of the experiment on the appropriate branch.
- (b) *Probability Extension* Compute the probability for each outcome of the experiment.

8. | **Tree Diagram**

- (a) Make a tree diagram to show all the possible sequences of answers for three multiple-choice questions, each with four possible responses.
- (b) *Probability Extension* Assuming that you are guessing the answers so that all outcomes listed in the tree are equally likely, what is the probability that you will guess the one sequence that contains all three correct answers?

9. | **Multiplication Rule** Four wires (red, green, blue, and yellow) need to be attached to a circuit board. A robotic device will attach the wires. The wires can be attached in any order, and the production manager wishes to determine which order would be fastest for the robot to use. Use the multiplication rule of counting to determine the number of possible sequences of assembly that must be tested. *Hint:* There are four choices for the first wire, three for the second, two for the third, and only one for the fourth.10. | **Multiplication Rule** A sales representative must visit four cities: Omaha, Dallas, Wichita, and Oklahoma City. There are direct air connections between each of the cities. Use the multiplication rule of counting to determine the number of different choices the sales representative has for the order in which to visit the cities. How is this problem similar to Problem 9?11. | **Counting: Agriculture** Barbara is a research biologist for Green Carpet Lawns. She is studying the effects of fertilizer type, temperature at time of application, and water treatment after application. She has four fertilizer types, three temperature zones, and three water treatments to test. Determine the number of different lawn plots she needs in order to test each fertilizer type, temperature range, and water treatment configuration.12. | **Counting: Outcomes** You toss a pair of dice.

- (a) Determine the number of possible pairs of outcomes. (Recall that there are six possible outcomes for each die.)
- (b) There are three even numbers on each die. How many outcomes are possible with even numbers appearing on each die?
- (c) *Probability extension:* What is the probability that both dice will show an even number?

13. | Compute $P_{5,2}$.17. | Compute $C_{5,2}$.14. | Compute $P_{8,3}$.18. | Compute $C_{8,3}$.15. | Compute $P_{7,7}$.19. | Compute $C_{7,7}$.16. | Compute $P_{9,9}$.20. | Compute $C_{8,8}$.21. | **Counting: Hiring** There are three nursing positions to be filled at Lilly Hospital. Position 1 is the day nursing supervisor; position 2 is the night nursing supervisor;

and position 3 is the nursing coordinator position. There are 15 candidates qualified for all three of the positions. Determine the number of different ways the positions can be filled by these applicants.

22. **Counting: Lottery** In the Cash Now lottery game there are 10 finalists who submitted entry tickets on time. From these 10 tickets, three grand prize winners will be drawn. The first prize is \$1 million, the second prize is \$100,000, and the third prize is \$10,000. Determine the total number of different ways in which the winners can be drawn. (Assume that the tickets are not replaced after they are drawn.)
23. **Counting: Sports** The University of Montana ski team has five entrants in a men's downhill ski event. The coach would like the first, second, and third places to go to the team members. In how many ways can the five team entrants achieve first, second, and third places?
24. **Counting: Sales** During the Computer Daze special promotion, a customer purchasing a computer and printer is given a choice of 3 free software packages. There are 10 different software packages from which to select. How many different groups of software packages can be selected?
25. **Counting: Hiring** There are 15 qualified applicants for 5 trainee positions in a fast-food management program. How many different groups of trainees can be selected?
26. **Counting: Grading** One professor grades homework by randomly choosing 5 out of 12 homework problems to grade.
 - (a) How many different groups of 5 problems can be chosen from the 12 problems?
 - (b) *Probability Extension* Jerry did only 5 problems of one assignment. What is the probability that the problems he did comprised the group that was selected to be graded?
 - (c) Silvia did 7 problems. How many different groups of 5 did she complete? What is the probability that one of the groups of 5 she completed comprised the group selected to be graded?
27. **Counting: Hiring** The qualified applicant pool for six management trainee positions consists of seven women and five men.
 - (a) How many different groups of applicants can be selected for the positions?
 - (b) How many different groups of trainees would consist entirely of women?
 - (c) *Probability Extension* If the applicants are equally qualified and the trainee positions are selected by drawing the names at random so that all groups of six are equally likely, what is the probability that the trainee class will consist entirely of women?



Chapter Review

SUMMARY

In this chapter we explored basic features of probability.

- The probability of an event A is a number between 0 and 1, inclusive. The more likely the event, the closer the probability of the event is to 1.
- Three main ways to determine the probability of an event are: the method of relative frequency, the method of equally likely outcomes, and intuition. Other important ways will be discussed later.

- The law of large numbers indicates that as the number of trials of a statistical experiment or observation increases, the relative frequency of a designated event becomes closer to the theoretical probability of that event.
- Events are mutually exclusive if they cannot occur together. Events are independent if the occurrence of one event does not change the probability of the occurrence of the other.

- Conditional probability is the probability that one event will occur, given that another event has occurred.
- The complement rule gives the probability that an event will not occur. The addition rule gives the probability that at least one of two specified events will occur. The multiplication rule gives the probability that two events will occur together.
- To determine the probability of equally likely events, we need to know how many outcomes

are possible. Devices such as tree diagrams and counting rules—such as the multiplication rule of counting, the permutations rule, and the combinations rule—help us determine the total number of outcomes of a statistical experiment or observation.

In most of the statistical applications of later chapters, we will use the addition rule for mutually exclusive events and the multiplication rule for independent events.

IMPORTANT WORDS & SYMBOLS

Section 4.1

Probability of an event A , $P(A)$ 132
 Intuition 132
 Relative frequency 132
 Equally likely outcomes 132
 Law of large numbers 134
 Statistical experiment 134
 Event 134
 Simple event 134
 Sample space 134
 Complement of event A^c 136

Section 4.2

Independent events 143
 Dependent events 143
 Probability of A and B 143
 Event $A \mid B$ 143
 Conditional probability 143
 Multiplication rules of probability (for independent and dependent events) 143

$P(A \mid B)$ 143

More than two independent events 147

Probability of A or B 147

Event A and B 147

Event A or B 147

Mutually exclusive events 149

Addition rules (for mutually exclusive and general events) 149

More than two mutually exclusive events 151

Basic probability rules 154

Section 4.3

Multiplication rule of counting 163
 Tree diagram 163
 Factorial notation 166
 Permutations rule 167
 Combinations rule 168

VIEWPOINT

Deathday and Birthday

Can people really postpone death? If so, how much can the timing of death be influenced by psychological, social, or other influential factors? One special event is a birthday. Do famous people try to postpone their deaths until an important birthday? Both Thomas Jefferson and John Adams died on July 4, 1826, when the United States was celebrating its 50th birthday. Is this only a strange coincidence, or is there an unexpected connection between birthdays and deathdays? The probability associated with a death rate decline of famous people just before important birthdays has been studied by Professor D. P. Phillips of the State University of New York and is presented in the book Statistics, A Guide to the Unknown, edited by J. M. Tanur.

CHAPTER REVIEW PROBLEMS

- for a raise received the raise. If a man is selected at random from the survey population of men, find the following probabilities: $P(\text{man asked for a raise})$; $P(\text{man received raise, given he asked for one})$; $P(\text{man asked for raise and received raise})$.
11. **General: Thumbtack** Drop a thumbtack and observe how it lands.
- Describe how you could use a relative frequency to estimate the probability that a thumbtack will land with its flat side down.
 - What is the sample space of outcomes for the thumbtack?
 - How would you make a probability assignment to this sample space if, when you drop 500 tacks, 340 land flat side down?
12. **Survey: Reaction to Poison Ivy** Allergic reactions to poison ivy can be miserable. Plant oils cause the reaction. Researchers at Allergy Institute did a study to determine the effects of washing the oil off within 5 minutes of exposure. A random sample of 1000 people with known allergies to poison ivy participated in the study. Oil from the poison ivy plant was rubbed on a patch of skin. For 500 of the subjects, it was washed off *within* 5 minutes. For the other 500 subjects, the oil was washed off *after* 5 minutes. The results are summarized in Table 4-5.
- | Reaction | Within 5 Minutes | After 5 Minutes | Row Total |
|--------------|------------------|-----------------|-----------|
| None | 420 | 50 | 470 |
| Mild | 60 | 330 | 390 |
| Strong | 20 | 120 | 140 |
| Column Total | 500 | 500 | 1000 |
- Let's use the following notation for the various events: W = washing oil off within 5 minute, A = washing oil off after 5 minutes, N = no reaction, M = mild reaction, S = strong reaction. Find the following probabilities for a person selected at random from this sample of 1000 subjects.
- $P(N)$, $P(M)$, $P(S)$
 - $P(N \mid W)$, $P(S \mid W)$
 - $P(N \mid A)$, $P(S \mid A)$
 - $P(N \text{ and } W)$, $P(M \text{ and } W)$
 - $P(N \text{ or } M)$. Are the events N = no reaction and M = mild reaction mutually exclusive? Explain.
 - Are the events N = no reaction and W = washing oil off within 5 minutes independent? Explain.
13. **General: Two Dice** In a game of craps, you roll two fair dice. Whether you win or lose depends on the sum of the numbers appearing on the tops of the dice. Let x be the random variable that represents the sum of the numbers on the tops of the dice.
- What values can x take on?
 - What is the probability distribution of these x values (that is, what is the probability that $x = 2, 3$, etc.)?
14. **Academic: Passing French** Class records at Rockwood College indicate that a student selected at random has probability 0.77 of passing French 101. For the student who passes French 101, the probability is 0.90 that he or she will pass French 102. What is the probability that a student selected at random will pass both French 101 and French 102?
15. **Combination: City Council** There is money to send two of eight city council members to a conference in Honolulu. All want to go, so they decide to choose the members to go to the conference by a random process. How many different combinations of two council members can be selected from the eight who want to go to the conference?

16. | **Basic Computation** Compute. (a) $P_{7,2}$ (b) $C_{7,2}$ (c) $P_{3,3}$ (d) $C_{4,4}$
17. | **Counting: Exam Answers** There are five multiple-choice questions on an exam, each with four possible answers. Determine the number of possible answer sequences for the five questions. Only one of the sets can contain all five correct answers. If you are guessing, so that you are as likely to choose one sequence of answers as another, what is the probability of getting all five answers correct?
18. | **Scheduling: College Courses** A student must satisfy the literature, social science, and philosophy requirements this semester. There are four literature courses to select from, three social science courses, and two philosophy courses. Make a tree diagram showing all the possible sequences of literature, social science, and philosophy courses.
19. | **General: Combination Lock** To open a combination lock, you turn the dial to the right and stop at a number; then you turn it to the left and stop at a second number. Finally, you turn the dial back to the right and stop at a third number. If you used the correct sequence of numbers, the lock opens. If the dial of the lock contains 10 numbers, 0 through 9, determine the number of different combinations possible for the lock. Note: The same number can be reused.
20. | **General: Combination Lock** You have a combination lock. Again, to open it you turn the dial to the right and stop at a first number; then you turn it to the left and stop at a second number. Finally, you turn the dial to the right and stop at a third number. Suppose you remember that the three numbers for your lock are 2, 9, and 5, but you don't remember the order in which the numbers occur. How many sequences of these three numbers are possible?

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

1. Look at Figure 4-11, “Who’s Cracking the Books?”
 - (a) Does the figure show the probability distribution of grade records for male students? for female students? Describe all the grade-record probability distributions shown in Figure 4-11. Find the probability that a male student selected at random has a grade record showing mostly A’s.
 - (b) Is the probability distribution shown for all students making mostly A’s? Explain your answer. *Hint:* Do the percentages shown for mostly A’s add up

FIGURE 4-11

Who’s Cracking the Books?

Undergraduate Grade Record by Student Characteristic

Student characteristic	C's and D's or lower	B's and C's	Mostly B's	A's and B's	Mostly A's
Gender:	Men	38.8 %	16.6 %	22.6 %	9.6 %
	Women	29.4	16.2	26.2	12.0
Class level:	Graduating senior	15.8 %	21.9 %	34.5 %	14.7 %
	All other class levels	35.4	15.8	23.6	10.5
Age:	18 or younger	42.6 %	14.7 %	23.4 %	9.4 %
	19 to 23	38.1	19.0	25.1	9.4
	24 to 29	33.3	16.9	24.7	10.3
	30 to 39	23.1	13.2	25.9	14.8
	40 and older	20.1	10.1	22.0	14.8
					33.0

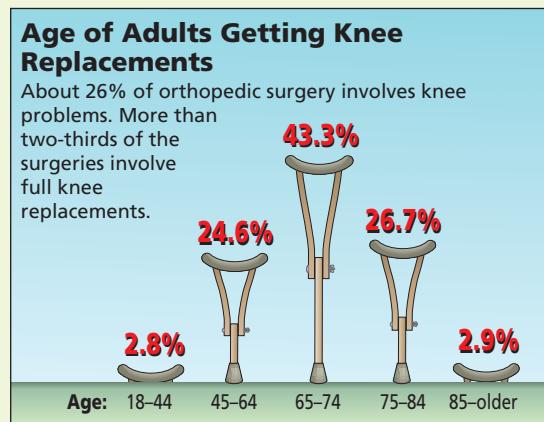
Source: U.S. Department of Education

to 1? Can Figure 4-11 be used to determine the probability that a student selected at random has mostly A's? Can it be used to determine the probability that a female student selected at random has mostly A's? What is the probability?

- (c) Can we use the information shown in the figure to determine the probability that a graduating senior has grades consisting of mostly B's or higher? What is the probability?
 - (d) Does Figure 4-11 give sufficient information to determine the probability that a student selected at random is in the age range 19 to 23 *and* has grades that are mostly B's? What is the probability that a student selected at random has grades that are mostly B's, *given* that he or she is in the age range 19 to 23?
 - (e) Suppose that 65% of the students at State University are between 19 and 23 years of age. What is the probability that a student selected at random is in this age range *and* has grades that are mostly B's?
2. Consider the information given in Figure 4-12, "Vulnerable Knees." What is the probability that an orthopedic case selected at random involves knee problems? Of those cases, estimate the probability that the case requires full knee replacement. Compute the probability that an orthopedic case selected at random involves a knee problem *and* requires a full knee replacement. Next, look at the probability distribution for ages of patients requiring full knee replacements. Medicare insurance coverage begins when a person reaches age 65. What is the probability that the age of a person receiving a knee replacement is 65 or older?

FIGURE 4-12

Vulnerable Knees



Source: American Academy of Orthopedic Surgeons

**LINKING CONCEPTS:
WRITING PROJECTS**

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas as appropriate.

1. Discuss the following concepts and give examples from everyday life in which you might encounter each concept. *Hint:* For instance, consider the “experiment” of arriving for class. Some possible outcomes are not arriving (that is, missing class), arriving on time, and arriving late.
 - (a) Sample space.
 - (b) Probability assignment to a sample space. In your discussion, be sure to include answers to the following questions.
 - (i) Is there more than one valid way to assign probabilities to a sample space? Explain and give an example.
 - (ii) How can probabilities be estimated by relative frequencies? How can probabilities be computed if events are equally likely?
2. Discuss the concepts of mutually exclusive events and independent events. List several examples of each type of event from everyday life.
 - (a) If A and B are mutually exclusive events, does it follow that A and B cannot be independent events? Give an example to demonstrate your answer. *Hint:* Discuss an election where only one person can win the election. Let A be the event that party A’s candidate wins, and let B be the event that party B’s candidate wins. Does the outcome of one event determine the outcome of the other event? Are A and B mutually exclusive events?
 - (b) Discuss the conditions under which $P(A \text{ and } B) = P(A) \cdot P(B)$ is true. Under what conditions is this not true?
 - (c) Discuss the conditions under which $P(A \text{ or } B) = P(A) + P(B)$ is true. Under what conditions is this not true?
3. Although we learn a good deal about probability in this course, the main emphasis is on statistics. Write a few paragraphs in which you talk about the distinction between probability and statistics. In what types of problems would probability be the main tool? In what types of problems would statistics be the main tool? Give some examples of both types of problems. What kinds of outcomes or conclusions do we expect from each type of problem?

USING TECHNOLOGY

Demonstration of the Law of Large Numbers

Computers can be used to simulate experiments. With packages such as Excel 2007, Minitab, and SPSS, programs using random-number generators can be designed (see the *Technology Guide*) to simulate activities such as tossing a die.

The following printouts show the results of the simulations for tossing a die 6, 500, 50,000, 500,000, and 1,000,000 times. Notice how the relative frequencies of the outcomes approach the theoretical probabilities of $1/6$ or 0.16667 for each outcome. Do you expect the same results every time the simulation is done? Why or why not?

Results of tossing one die 6 times

Outcome	Number of Occurrences	Relative Frequency
•	0	0.00000
••	1	0.16667
•••	2	0.33333
••••	0	0.00000
•••••	1	0.16667
••••••	2	0.33333

Results of tossing one die 500 times

Outcome	Number of Occurrences	Relative Frequency
•	87	0.17400
••	83	0.16600
•••	91	0.18200
••••	69	0.13800
•••••	87	0.17400
••••••	83	0.16600

Results of tossing one die 50,000 times

Outcome	Number of Occurrences	Relative Frequency
•	8528	0.17056
••	8354	0.16708
•••	8246	0.16492
••••	8414	0.16828
•••••	8178	0.16356
••••••	8280	0.16560

Results of tossing one die 500,000 times

Outcome	Number of Occurrences	Relative Frequency
•	83644	0.16729
••	83368	0.16674
•••	83398	0.16680
••••	83095	0.16619
•••••	83268	0.16654
••••••	83227	0.16645

Results of tossing one die 1,000,000 times

Outcome	Number of Occurrences	Relative Frequency
•	166643	0.16664
••	166168	0.16617
•••	167391	0.16739
••••	165790	0.16579
•••••	167243	0.16724
••••••	166765	0.16677



5

5.1 Introduction to Random Variables and Probability Distributions

5.2 Binomial Probabilities

5.3 Additional Properties of the Binomial Distribution

5.4 The Geometric and Poisson Probability Distributions



National Portrait Gallery,
Smithsonian Institution/Art
Resource, NY

*Education is the
key to unlock the
golden door of
freedom.*

—GEORGE
WASHINGTON CARVER



AP Photo/Matt Rourke

George Washington Carver (1859–1943) won international fame for agricultural research. After graduating from Iowa State College, he was appointed a faculty member in the Iowa State Botany Department. Carver took charge of the greenhouse and started a fungus collection that later included more than 20,000 species. This collection brought him professional acclaim in the field of botany.

At the invitation of his friend Booker T. Washington, Carver joined the faculty of the Tuskegee Institute, where he spent the rest of his long and distinguished career. Carver's creative genius accounted for more than 300 inventions from peanuts, 118 inventions from sweet potatoes, and 75 inventions from pecans.

Gathering and analyzing data were important components of Carver's work. Methods you will learn in this course are widely used in research in every field, including agriculture.

For online student resources, visit The Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>

THE BINOMIAL PROBABILITY DISTRIBUTION AND RELATED TOPICS

PREVIEW QUESTIONS

What is a random variable? How do you compute μ and σ for a discrete random variable? How do you compute μ and σ for linear combinations of independent random variables? (SECTION 5.1)

Many of life's experiences consist of some successes together with some failures. Suppose you make n attempts to succeed at a certain project. How can you use the binomial probability distribution to compute the probability of r successes? (SECTION 5.2)

How do you compute μ and σ for the binomial distribution? (SECTION 5.3)

How is the binomial distribution related to other probability distributions, such as the geometric and Poisson? (SECTION 5.4)



Andresr/istockphoto.com

FOCUS PROBLEM

Personality Preference Types: Introvert or Extrovert?

Isabel Briggs Myers was a pioneer in the study of personality types. Her work has been used successfully in counseling, educational, and industrial settings. In the book *A Guide to the Development and Use of the Myers–Briggs Type Indicators*, by Myers and McCaulley, it was reported that based on a very large sample (2282 professors), approximately 45% of all university professors are extroverted.

After completing this chapter, you will be able to answer the following questions. Suppose you have classes with six different professors.

- What is the probability that all six are extroverts?
- What is the probability that none of your professors is an extrovert?
- What is the probability that at least two of your professors are extroverts?
- In a group of six professors selected at random, what is the *expected number* of extroverts? What is the *standard deviation* of the distribution?



Mitch Wojnarowicz/The Image Works

- (e) Suppose you were assigned to write an article for the student newspaper and you were given a quota (by the editor) of interviewing at least three extroverted professors. How many professors selected at random would you need to interview to be at least 90% sure of filling the quota?

(See Problem 24 of Section 5.3.)

COMMENT Both extroverted and introverted professors can be excellent teachers.

SECTION 5.1

Introduction to Random Variables and Probability Distributions

FOCUS POINTS

- Distinguish between discrete and continuous random variables.
- Graph discrete probability distributions.
- Compute μ and σ for a discrete probability distribution.
- Compute μ and σ for a linear function of a random variable x .
- Compute μ and σ for a linear combination of two independent random variables.

Random Variables

For our purposes, we say that a *statistical experiment* or *observation* is any process by which measurements are obtained. For instance, you might count the number of eggs in a robin's nest or measure daily rainfall in inches. It is common practice to use the letter x to represent the quantitative result of an experiment or observation. As such, we call x a variable.

Random Variable

A quantitative variable x is a **random variable** if the value that x takes on in a given experiment or observation is a chance or random outcome.

Discrete random variable

A **discrete random variable** can take on only a finite number of values or a countable number of values.

Continuous random variable

A **continuous random variable** can take on any of the countless number of values in a line interval.

The distinction between discrete and continuous random variables is important because of the different mathematical techniques associated with the two kinds of random variables.

In most of the cases we will consider, a *discrete random variable* will be the result of a count. The number of students in a statistics class is a discrete random variable. Values such as 15, 25, 50, and 250 are all possible. However, 25.5 students is not a possible value for the number of students.

Most of the *continuous random variables* we will see will occur as the result of a measurement on a continuous scale. For example, the air pressure in an automobile tire represents a continuous random variable. The air pressure could, in theory, take on any value from 0 lb/in² (psi) to the bursting pressure of the tire. Values such as 20.126 psi, 20.12678 psi, and so forth are possible.

GUIDED EXERCISE 1

Discrete or continuous random variables

Which of the following random variables are discrete and which are continuous?

- (a) Measure the time it takes a student selected at random to register for the fall term.  Time can take on any value, so this is a continuous random variable.

Continued

GUIDED EXERCISE 1 *continued*

- (b) *Count* the number of bad checks drawn on Upright Bank on a day selected at random.
- (c) *Measure* the amount of gasoline needed to drive your car 200 miles.
- (d) Pick a random sample of 50 registered voters in a district and find the number who voted in the last county election.
- The number of bad checks can be only a whole number such as 0, 1, 2, 3, etc. This is a discrete variable.
- We are measuring volume, which can assume any value, so this is a continuous random variable.
- This is a count, so the variable is discrete.

Probability Distribution of a Discrete Random Variable

A random variable has a probability distribution whether it is discrete or continuous.

Probability distribution

A **probability distribution** is an assignment of probabilities to each distinct value of a discrete random variable or to each interval of values of a continuous random variable.

Features of the probability distribution of a discrete random variable

1. The probability distribution has a probability assigned to *each* distinct value of the random variable.
2. The sum of all the assigned probabilities must be 1.

EXAMPLE 1

DISCRETE PROBABILITY DISTRIBUTION

Dr. Mendoza developed a test to measure boredom tolerance. He administered it to a group of 20,000 adults between the ages of 25 and 35. The possible scores were 0, 1, 2, 3, 4, 5, and 6, with 6 indicating the highest tolerance for boredom. The test results for this group are shown in Table 5-1.

- (a) If a subject is chosen at random from this group, the probability that he or she will have a score of 3 is 6000/20,000, or 0.30. In a similar way, we can use relative frequencies to compute the probabilities for the other scores (Table 5-2). These probability assignments make up the probability distribution. Notice that the scores are mutually exclusive: No one subject has two scores. The sum of the probabilities of all the scores is 1.

TABLE 5-1
Boredom Tolerance Test Scores for 20,000 Subjects

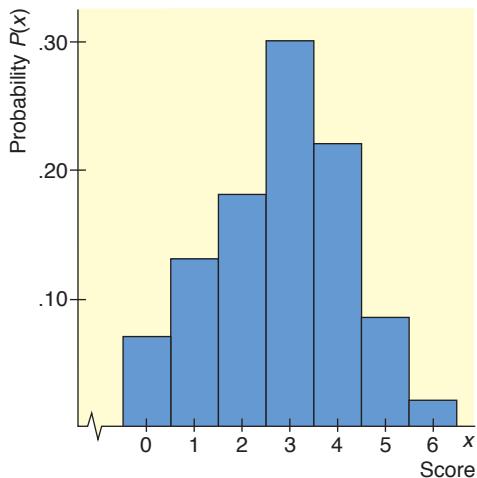
Score	Number of Subjects
0	1400
1	2600
2	3600
3	6000
4	4400
5	1600
6	400

TABLE 5-2
Probability Distribution of Scores on Boredom Tolerance Test

Score x	Probability $P(x)$
0	0.07
1	0.13
2	0.18
3	0.30
4	0.22
5	0.08
6	0.02
$\Sigma P(x) = 1$	

FIGURE 5-1

Graph of the Probability Distribution of Test Scores



- (b) The graph of this distribution is simply a relative-frequency histogram (see Figure 5-1) in which the height of the bar over a score represents the probability of that score. Since each bar is one unit wide, the area of the bar over a score equals the height and thus represents the probability of that score. Since the sum of the probabilities is 1, the area under the graph is also 1.
- (c) The Topnotch Clothing Company needs to hire someone with a score on the boredom tolerance test of 5 or 6 to operate the fabric press machine. Since the scores 5 and 6 are mutually exclusive, the probability that someone in the group who took the boredom tolerance test made either a 5 or a 6 is the sum

$$\begin{aligned}P(5 \text{ or } 6) &= P(5) + P(6) \\&= 0.08 + 0.02 = 0.10\end{aligned}$$

Notice that to find $P(5 \text{ or } 6)$, we could have simply added the *areas* of the bars over 5 and over 6. One out of 10 of the group who took the boredom tolerance test would qualify for the position at Topnotch Clothing.

**GUIDED EXERCISE 2****Discrete probability distribution**

One of the elementary tools of cryptanalysis (the science of code breaking) is to use relative frequencies of occurrence of different letters in the alphabet to break standard English alphabet codes. Large samples of plain text such as newspaper stories generally yield about the same relative frequencies for letters. A sample 1000 letters long yielded the information in Table 5-3.

- (a) Use the relative frequencies to compute the omitted probabilities in Table 5-3.  Table 5-4 shows the completion of Table 5-3.

Continued

GUIDED EXERCISE 2 *continued***TABLE 5-3 Frequencies of Letters in a 1000-Letter Sample**

Letter	Freq.	Prob.	Letter	Freq.	Prob.
A	73	—	N	78	0.078
B	9	0.009	O	74	—
C	30	0.030	P	27	0.027
D	44	0.044	Q	3	0.003
E	130	—	R	77	0.077
F	28	0.028	S	63	0.063
G	16	0.016	T	93	0.093
H	35	0.035	U	27	—
I	74	—	V	13	0.013
J	2	0.002	W	16	0.016
K	3	0.003	X	5	0.005
L	35	0.035	Y	19	0.019
M	25	0.025	Z	1	0.001

Source: From *Elementary Cryptanalysis: A Mathematical Approach*, by Abraham Sinkov. Copyright © 1968 and renewed 1996 by Yale University. Used by permission of Random House, Inc. Copyright The Mathematical Association of America 2010. All rights reserved

- (b) Do the probabilities of all the individual letters add up to 1?
- (c) If a letter is selected at random from a newspaper story, what is the probability that the letter will be a vowel?

TABLE 5-4 Entries for Table 5-3

Letter	Relative Frequency	Probability
A	$\frac{73}{1,000}$	0.073
E	$\frac{130}{1,000}$	0.130
I	$\frac{74}{1,000}$	0.074
O	$\frac{74}{1,000}$	0.074
U	$\frac{27}{1,000}$	0.027

Yes.

If a letter is selected at random,

$$\begin{aligned}
 P(a, e, i, o, \text{ or } u) &= P(a) + P(e) + P(i) + \\
 &\quad P(o) + P(u) \\
 &= 0.073 + 0.130 + 0.074 + \\
 &\quad 0.074 + 0.027 \\
 &= 0.378
 \end{aligned}$$

Mean and standard deviation of a discrete probability distribution

A probability distribution can be thought of as a relative-frequency distribution based on a very large n . As such, it has a mean and standard deviation. If we are referring to the probability distribution of a *population*, then we use the Greek letters μ for the mean and σ for the standard deviation. When we see the Greek letters used, we know the information given is from the *entire population* rather than just a sample. If we have a sample probability distribution, we use \bar{x} (x bar) and s , respectively, for the mean and standard deviation.

The mean and the standard deviation of a discrete population probability distribution are found by using these formulas:

$$\mu = \sum xP(x); \mu \text{ is called the expected value of } x$$

$$\sigma = \sqrt{\sum (x - \mu)^2 P(x)}; \sigma \text{ is called the standard deviation of } x$$

where x is the value of a random variable,

$P(x)$ is the probability of that variable, and

the sum Σ is taken for all the values of the random variable.

Note: μ is the *population mean* and σ is the underlying *population standard deviation* because the sum Σ is taken over *all* values of the random variable (i.e., the entire sample space).

Expected value

The mean of a probability distribution is often called the *expected value* of the distribution. This terminology reflects the idea that the mean represents a “central point” or “cluster point” for the entire distribution. Of course, the mean or expected value is an average value, and as such, it *need not be a point of the sample space*.

The standard deviation is often represented as a measure of *risk*. A larger standard deviation implies a greater likelihood that the random variable x is different from the expected value μ .

EXAMPLE 2**EXPECTED VALUE, STANDARD DEVIATION**

Are we influenced to buy a product by an ad we saw on TV? National Infomercial Marketing Association determined the number of times *buyers* of a product had watched a TV infomercial *before* purchasing the product. The results are shown here:



Kelly-Mooney Photography/Encyclopedia/Corbis

Number of Times Buyers Saw Infomercial	1	2	3	4	5*
Percentage of Buyers	27%	31%	18%	9%	15%

*This category was 5 or more, but will be treated as 5 in this example.

We can treat the information shown as an estimate of the probability distribution because the events are mutually exclusive and the sum of the percentages is 100%. Compute the mean and standard deviation of the distribution.

SOLUTION: We put the data in the first two columns of a computation table and then fill in the other entries (see Table 5-5). The average number of times a buyer views the infomercial before purchase is

$$\mu = \sum xP(x) = 2.54 \text{ (sum of column 3)}$$

To find the standard deviation, we take the square root of the sum of column 6:

$$\sigma = \sqrt{\sum (x - \mu)^2 P(x)} \approx \sqrt{1.869} \approx 1.37$$

TABLE 5-5 Number of Times Buyers View Infomercial Before Making Purchase

x (number of viewings)	$P(x)$	$xP(x)$	$x - \mu$	$(x - \mu)^2$	$(x - \mu)^2 P(x)$
1	0.27	0.27	-1.54	2.372	0.640
2	0.31	0.62	-0.54	0.292	0.091
3	0.18	0.54	0.46	0.212	0.038
4	0.09	0.36	1.46	2.132	0.192
5	0.15	0.75	2.46	6.052	0.908
$\mu = \sum xP(x) = 2.54$				$\Sigma (x - \mu)^2 P(x) = 1.869$	

CALCULATOR NOTE Some calculators, including the TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) models, accept fractional frequencies. If yours does, you can get μ and σ directly by entering the outcomes in list L₁ with corresponding probabilities in list L₂. Then use **1-Var Stats L₁, L₂**.

GUIDED EXERCISE 3**Expected value**

At a carnival, you pay \$2.00 to play a coin-flipping game with three fair coins. On each coin one side has the number 0 and the other side has the number 1. You flip the three coins at one time and you win \$1.00 for every 1 that appears on top. Are your expected earnings equal to the cost to play? We'll answer this question in several steps.

- (a) In this game, the random variable of interest counts the number of 1's that show. What is the sample space for the values of this random variable?

→ The sample space is {0, 1, 2, 3}, since any of these numbers of 1's can appear.

- (b) There are eight equally likely outcomes for throwing three coins. They are 000, 001, 010, 011, 100, 101, _____, and _____.

→ 110 and 111.

- (c) Complete Table 5-6.

TABLE 5-6

Number of 1's,			
<i>x</i>	Frequency	<i>P(x)</i>	<i>xP(x)</i>
0	1	0.125	0
1	3	0.375	_____
2	3	_____	_____
3	_____	_____	_____

- (d) The expected value is the sum

$$\mu = \Sigma xP(x)$$

Sum the appropriate column of Table 5-6 to find this value. Are your expected earnings less than, equal to, or more than the cost of the game?

TABLE 5-7 Completion of Table 5-6

Number of 1's,			
<i>x</i>	Frequency	<i>P(x)</i>	<i>xP(x)</i>
0	1	0.125	0
1	3	0.375	0.375
2	3	0.375	0.750
3	1	0.125	0.375

→ The expected value can be found by summing the last column of Table 5-7. The expected value is \$1.50. It cost \$2.00 to play the game; the expected value is less than the cost. The carnival is making money. In the long run, the carnival can expect to make an average of about 50 cents per player.

We have seen probability distributions of discrete variables and the formulas to compute the mean and standard deviation of a discrete population probability distribution. Probability distributions of continuous random variables are similar except that the probability assignments are made to intervals of values rather than to specific values of the random variable. We will see an important example of a discrete probability distribution, the binomial distribution, in the next section, and one of a continuous probability distribution in Chapter 6 when we study the normal distribution.

We conclude this section with some useful information about combining random variables.

Linear Functions of a Random Variable

Let a and b be any constants, and let x be a random variable. Then the new random variable $L = a + bx$ is called a *linear function of x*. Using some more advanced mathematics, the following can be proved.

Linear Function of a random variable

Let x be a random variable with mean μ and standard deviation σ . Then the **linear function** $L = a + bx$ has mean, variance, and standard deviation as follows:

$$\begin{aligned}\mu_L &= a + b\mu \\ \sigma_L^2 &= b^2\sigma^2 \\ \sigma_L &= \sqrt{b^2\sigma^2} = |b|\sigma\end{aligned}$$

Linear combination of two independent random variables

Linear Combinations of Independent Random Variables

Suppose we have two random variables x_1 and x_2 . These variables are *independent* if any event involving x_1 by itself is *independent* of any event involving x_2 by itself. Sometimes, we want to combine independent random variables and examine the mean and standard deviation of the resulting combination.

Let x_1 and x_2 be independent random variables, and let a and b be any constants. Then the new random variable $W = ax_1 + bx_2$ is called a *linear combination of x_1 and x_2* . Using some more advanced mathematics, the following can be proved.

Let x_1 and x_2 be independent random variables with respective means μ_1 and μ_2 , and variances σ_1^2 and σ_2^2 . For the **linear combination** $W = ax_1 + bx_2$, the mean, variance, and standard deviation are as follows:

$$\begin{aligned}\mu_W &= a\mu_1 + b\mu_2 \\ \sigma_W^2 &= a^2\sigma_1^2 + b^2\sigma_2^2 \\ \sigma_W &= \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\end{aligned}$$

Note: The formula for the mean of a linear combination of random variables is valid regardless of whether the variables are independent. However, *the formulas for the variance and standard deviation are valid* only if x_1 and x_2 are *independent* random variables. In later work (Chapter 6 on), we will use independent random samples to ensure that the resulting variables (usually means, proportions, etc.) are statistically independent.



EXAMPLE 3

LINEAR COMBINATIONS OF INDEPENDENT RANDOM VARIABLES

Let x_1 and x_2 be independent random variables with respective means $\mu_1 = 75$ and $\mu_2 = 50$, and standard deviations $\sigma_1 = 16$ and $\sigma_2 = 9$.

(a) Let $L = 3 + 2x_1$. Compute the mean, variance, and standard deviation of L .

SOLUTION: L is a **linear function** of the random variable x_1 . Using the formulas with $a = 3$ and $b = 2$, we have

$$\begin{aligned}\mu_L &= 3 + 2\mu_1 = 3 + 2(75) = 153 \\ \sigma_L^2 &= 2^2\sigma_1^2 = 4(16)^2 = 1024 \\ \sigma_L &= |2|\sigma_1 = 2(16) = 32\end{aligned}$$

Notice that the variance and standard deviation of the linear function are influenced only by the coefficient of x_1 in the linear function.

(b) Let $W = x_1 + x_2$. Find the mean, variance, and standard deviation of W .

SOLUTION: W is a linear combination of the independent random variables x_1 and x_2 . Using the formulas with both a and b equal to 1, we have

$$\begin{aligned}\mu_W &= \mu_1 + \mu_2 = 75 + 50 = 125 \\ \sigma_W^2 &= \sigma_1^2 + \sigma_2^2 = 16^2 + 9^2 = 337 \\ \sigma_W &= \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{337} \approx 18.36\end{aligned}$$

(c) Let $W = x_1 - x_2$. Find the mean, variance, and standard deviation of W .

SOLUTION: W is a linear combination of the independent random variables x_1 and x_2 . Using the formulas with $a = 1$ and $b = -1$, we have

$$\begin{aligned}\mu_W &= \mu_1 - \mu_2 = 75 - 50 = 25 \\ \sigma_W^2 &= 1^2\sigma_1^2 + (-1)^2\sigma_2^2 = 16^2 + 9^2 = 337 \\ \sigma_W &= \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{337} \approx 18.36\end{aligned}$$

(d) Let $W = 3x_1 - 2x_2$. Find the mean, variance, and standard deviation of W .

SOLUTION: W is a linear combination of the independent random variables x_1 and x_2 . Using the formulas with $a = 3$ and $b = -2$, we have

$$\begin{aligned}\mu_W &= 3\mu_1 - 2\mu_2 = 3(75) - 2(50) = 125 \\ \sigma_W^2 &= 3^2\sigma_1^2 + (-2)^2\sigma_2^2 = 9(16^2) + 4(9^2) = 2628 \\ \sigma_W &= \sqrt{2628} \approx 51.26\end{aligned}$$

LOOKING FORWARD

Problem 24 of Section 9.1 shows how to find the mean, variance, and standard deviation of a linear combination of two *linearly dependent* random variables.

VIEWPOINT

The Rosetta Project

Around 196 B.C., Egyptian priests inscribed a decree on a granite slab affirming the rule of 13-year-old Ptolemy V. The proclamation was in Egyptian hieroglyphics with a translation in a form of ancient Greek. By 1799, the meaning of Egyptian hieroglyphics had been lost for many centuries. However, Napoleon's troops discovered the granite slab (Rosetta Stone). Linguists then used the Rosetta Stone and their knowledge of ancient Greek to unlock the meaning of the Egyptian hieroglyphics.

Linguistic experts say that because of industrialization and globalization, by the year 2100 as many as 90% of the world's languages may be extinct. To help preserve some of these languages for future generations, 1000 translations of the first three chapters of Genesis have been inscribed in tiny text onto 3-inch nickel disks and encased in hardened glass balls that are expected to last at least 1000 years. Why Genesis? Because it is the most translated text in the world. The Rosetta Project is sending the disks to libraries and universities all over the world. It is very difficult to send information into the future. However, if in the year 2500 linguists are using the "Rosetta Disks" to unlock the meaning of a lost language, you may be sure they will use statistical methods of cryptanalysis (see Guided Exercise 2). To find out more about the Rosetta Project, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Rosetta Project site.

SECTION 5.1 PROBLEMS

1. **Statistical Literacy** Which of the following are continuous variables, and which are discrete?
 - (a) Number of traffic fatalities per year in the state of Florida
 - (b) Distance a golf ball travels after being hit with a driver
 - (c) Time required to drive from home to college on any given day
 - (d) Number of ships in Pearl Harbor on any given day
 - (e) Your weight before breakfast each morning
2. **Statistical Literacy** Which of the following are continuous variables, and which are discrete?
 - (a) Speed of an airplane
 - (b) Age of a college professor chosen at random
 - (c) Number of books in the college bookstore
 - (d) Weight of a football player chosen at random
 - (e) Number of lightning strikes in Rocky Mountain National Park on a given day
3. **Statistical Literacy** Consider each distribution. Determine if it is a valid probability distribution or not, and explain your answer.

(a)	x	0	1	2
	$P(x)$	0.25	0.60	0.15

(b)	x	0	1	2
	$P(x)$	0.25	0.60	0.20
4. **Statistical Literacy** Consider the probability distribution of a random variable x . Is the expected value of the distribution necessarily one of the possible values of x ? Explain or give an example.
5. **Basic Computation: Expected Value and Standard Deviation** Consider the probability distribution shown in Problem 3(a). Compute the expected value and the standard deviation of the distribution.
6. **Basic Computation: Expected Value** For a fundraiser, 1000 raffle tickets are sold and the winner is chosen at random. There is only one prize, \$500 in cash. You buy one ticket.
 - (a) What is the probability you will win the prize of \$500?
 - (b) Your expected earnings can be found by multiplying the value of the prize by the probability you will win the prize. What are your expected earnings?
 - (c) **Interpretation** If a ticket costs \$2, what is the difference between your “costs” and “expected earnings”? How much are you effectively contributing to the fundraiser?
7. **Critical Thinking: Simulation** We can use the random-number table to simulate outcomes from a given discrete probability distribution. Jose plays basketball and has probability 0.7 of making a free-throw shot. Let x be the random variable that counts the number of successful shots out of 10 attempts. Consider the digits 0 through 9 of the random-number table. Since Jose has a 70% chance of making a shot, assign the digits 0 through 6 to “making a basket from the free throw line” and the digits 7 through 9 to “missing the shot.”
 - (a) Do 70% of the possible digits 0 through 9 represent “making a basket”?
 - (b) Start at line 2, column 1 of the random-number table. Going across the row, determine the results of 10 “trials.” How many free throw shots are successful in this simulation?
 - (c) Your friend decides to assign the digits 0 through 2 to “missing the shot” and the digits 3 through 9 to “making the basket.” Is this assignment valid? Explain. Using this assignment, repeat part (b).
8. **Marketing: Age** What is the age distribution of promotion-sensitive shoppers? A *supermarket super shopper* is defined as a shopper for whom at least 70% of the items purchased were on sale or purchased with a coupon. The following table is based on information taken from *Trends in the United States* (Food Marketing Institute, Washington, D.C.).

Age range, years	18–28	29–39	40–50	51–61	62 and over
Midpoint x	23	34	45	56	67
Percent of super shoppers	7%	44%	24%	14%	11%

For the 62-and-over group, use the midpoint 67 years.

- (a) Using the age midpoints x and the percentage of super shoppers, do we have a valid probability distribution? Explain.
- (b) Use a histogram to graph the probability distribution of part (a).
- (c) Compute the expected age μ of a super shopper.
- (d) Compute the standard deviation σ for ages of super shoppers.
9. **Marketing: Income** What is the income distribution of super shoppers (see Problem 8). In the following table, income units are in thousands of dollars, and each interval goes up to but does not include the given high value. The midpoints are given to the nearest thousand dollars.

Income range	5–15	15–25	25–35	35–45	45–55	55 or more
Midpoint x	10	20	30	40	50	60
Percent of super shoppers	21%	14%	22%	15%	20%	8%

- (a) Using the income midpoints x and the percent of super shoppers, do we have a valid probability distribution? Explain.
- (b) Use a histogram to graph the probability distribution of part (a).
- (c) Compute the expected income μ of a super shopper.
- (d) Compute the standard deviation σ for the income of super shoppers.
10. **History: Florence Nightingale** What was the age distribution of nurses in Great Britain at the time of Florence Nightingale? Thanks to Florence Nightingale and the British census of 1851, we have the following information (based on data from the classic text *Notes on Nursing*, by Florence Nightingale). Note: In 1851 there were 25,466 nurses in Great Britain. Furthermore, Nightingale made a strict distinction between nurses and domestic servants.

Age range (yr)	20–29	30–39	40–49	50–59	60–69	70–79	80+
Midpoint x	24.5	34.5	44.5	54.5	64.5	74.5	84.5
Percent of nurses	5.7%	9.7%	19.5%	29.2%	25.0%	9.1%	1.8%

- (a) Using the age midpoints x and the percent of nurses, do we have a valid probability distribution? Explain.
- (b) Use a histogram to graph the probability distribution of part (a).
- (c) Find the probability that a British nurse selected at random in 1851 was 60 years of age or older.
- (d) Compute the expected age μ of a British nurse contemporary to Florence Nightingale.
- (e) Compute the standard deviation σ for ages of nurses shown in the distribution.
11. **Fishing: Trout** The following data are based on information taken from *Daily Creel Summary*, published by the Paiute Indian Nation, Pyramid Lake, Nevada. Movie stars and U.S. presidents have fished Pyramid Lake. It is one of the best places in the lower 48 states to catch trophy cutthroat trout. In this



Sean Locke/istockphoto.com



table, x = number of fish caught in a 6-hour period. The percentage data are the percentages of fishermen who catch x fish in a 6-hour period while fishing from shore.

x	0	1	2	3	4 or more
%	44%	36%	15%	4%	1%

- (a) Convert the percentages to probabilities and make a histogram of the probability distribution.
 - (b) Find the probability that a fisherman selected at random fishing from shore catches one or more fish in a 6-hour period.
 - (c) Find the probability that a fisherman selected at random fishing from shore catches two or more fish in a 6-hour period.
 - (d) Compute μ , the expected value of the number of fish caught per fisherman in a 6-hour period (round 4 or more to 4).
 - (e) Compute σ , the standard deviation of the number of fish caught per fisherman in a 6-hour period (round 4 or more to 4).
12. **Criminal Justice: Parole** USA Today reported that approximately 25% of all state prison inmates released on parole become repeat offenders while on parole. Suppose the parole board is examining five prisoners up for parole. Let x = number of prisoners out of five on parole who become repeat offenders. The methods of Section 5.2 can be used to compute the probability assignments for the x distribution.
- | x | 0 | 1 | 2 | 3 | 4 | 5 |
|--------|-------|-------|-------|-------|-------|-------|
| $P(x)$ | 0.237 | 0.396 | 0.264 | 0.088 | 0.015 | 0.001 |
- (a) Find the probability that one or more of the five parolees will be repeat offenders. How does this number relate to the probability that none of the parolees will be repeat offenders?
 - (b) Find the probability that two or more of the five parolees will be repeat offenders.
 - (c) Find the probability that four or more of the five parolees will be repeat offenders.
 - (d) Compute μ , the expected number of repeat offenders out of five.
 - (e) Compute σ , the standard deviation of the number of repeat offenders out of five.
13. **Fundraiser: Hiking Club** The college hiking club is having a fundraiser to buy new equipment for fall and winter outings. The club is selling Chinese fortune cookies at a price of \$1 per cookie. Each cookie contains a piece of paper with a different number written on it. A random drawing will determine which number is the winner of a dinner for two at a local Chinese restaurant. The dinner is valued at \$35. Since the fortune cookies were donated to the club, we can ignore the cost of the cookies. The club sold 719 cookies before the drawing.
- (a) Lisa bought 15 cookies. What is the probability she will win the dinner for two? What is the probability she will not win?
 - (b) **Interpretation** Lisa's expected earnings can be found by multiplying the value of the dinner by the probability that she will win. What are Lisa's expected earnings? How much did she effectively contribute to the hiking club?
14. **Spring Break: Caribbean Cruise** The college student senate is sponsoring a spring break Caribbean cruise raffle. The proceeds are to be donated to the Samaritan Center for the Homeless. A local travel agency donated the cruise, valued at \$2000. The students sold 2852 raffle tickets at \$5 per ticket.
- (a) Kevin bought six tickets. What is the probability that Kevin will win the spring break cruise to the Caribbean? What is the probability that Kevin will not win the cruise?

- (b) **Interpretation** Expected earnings can be found by multiplying the value of the cruise by the probability that Kevin will win. What are Kevin's expected earnings? Is this more or less than the amount Kevin paid for the six tickets? How much did Kevin effectively contribute to the Samaritan Center for the Homeless?
15. **Expected Value: Life Insurance** Jim is a 60-year-old Anglo male in reasonably good health. He wants to take out a \$50,000 term (that is, straight death benefit) life insurance policy until he is 65. The policy will expire on his 65th birthday. The probability of death in a given year is provided by the Vital Statistics Section of the *Statistical Abstract of the United States* (116th Edition).

$x = \text{age}$	60	61	62	63	64
$P(\text{death at this age})$	0.01191	0.01292	0.01396	0.01503	0.01613

- Jim is applying to Big Rock Insurance Company for his term insurance policy.
- (a) What is the probability that Jim will die in his 60th year? Using this probability and the \$50,000 death benefit, what is the expected cost to Big Rock Insurance?
- (b) Repeat part (a) for years 61, 62, 63, and 64. What would be the total expected cost to Big Rock Insurance over the years 60 through 64?
- (c) **Interpretation** If Big Rock Insurance wants to make a profit of \$700 above the expected total cost paid out for Jim's death, how much should it charge for the policy?
- (d) **Interpretation** If Big Rock Insurance Company charges \$5000 for the policy, how much profit does the company expect to make?
16. **Expected Value: Life Insurance** Sara is a 60-year-old Anglo female in reasonably good health. She wants to take out a \$50,000 term (that is, straight death benefit) life insurance policy until she is 65. The policy will expire on her 65th birthday. The probability of death in a given year is provided by the Vital Statistics Section of the *Statistical Abstract of the United States* (116th Edition).

$x = \text{age}$	60	61	62	63	64
$P(\text{death at this age})$	0.00756	0.00825	0.00896	0.00965	0.01035

- Sara is applying to Big Rock Insurance Company for her term insurance policy.
- (a) What is the probability that Sara will die in her 60th year? Using this probability and the \$50,000 death benefit, what is the expected cost to Big Rock Insurance?
- (b) Repeat part (a) for years 61, 62, 63, and 64. What would be the total expected cost to Big Rock Insurance over the years 60 through 64?
- (c) **Interpretation** If Big Rock Insurance wants to make a profit of \$700 above the expected total cost paid out for Sara's death, how much should it charge for the policy?
- (d) **Interpretation** If Big Rock Insurance Company charges \$5000 for the policy, how much profit does the company expect to make?
17. **Combination of Random Variables: Golf** Norb and Gary are entered in a local golf tournament. Both have played the local course many times. Their scores are random variables with the following means and standard deviations.

$$\text{Norb, } x_1: \mu_1 = 115; \sigma_1 = 12 \quad \text{Gary, } x_2: \mu_2 = 100; \sigma_2 = 8$$

In the tournament, Norb and Gary are not playing together, and we will assume their scores vary independently of each other.

- (a) The difference between their scores is $W = x_1 - x_2$. Compute the mean, variance, and standard deviation for the random variable W .



- (b) The average of their scores is $W = 0.5x_1 + 0.5x_2$. Compute the mean, variance, and standard deviation for the random variable W .
- (c) The tournament rules have a special handicap system for each player. For Norb, the handicap formula is $L = 0.8x_1 - 2$. Compute the mean, variance, and standard deviation for the random variable L .
- (d) For Gary, the handicap formula is $L = 0.95x_2 - 5$. Compute the mean, variance, and standard deviation for the random variable L .
18. **Combination of Random Variables: Repair Service** A computer repair shop has two work centers. The first center examines the computer to see what is wrong, and the second center repairs the computer. Let x_1 and x_2 be random variables representing the lengths of time in minutes to examine a computer (x_1) and to repair a computer (x_2). Assume x_1 and x_2 are independent random variables. Long-term history has shown the following times:
- Examine computer, $x_1: \mu_1 = 28.1$ minutes; $\sigma_1 = 8.2$ minutes
- Repair computer, $x_2: \mu_2 = 90.5$ minutes; $\sigma_2 = 15.2$ minutes
- (a) Let $W = x_1 + x_2$ be a random variable representing the total time to examine and repair the computer. Compute the mean, variance, and standard deviation of W .
- (b) Suppose it costs \$1.50 per minute to examine the computer and \$2.75 per minute to repair the computer. Then $W = 1.50x_1 + 2.75x_2$ is a random variable representing the service charges (without parts). Compute the mean, variance, and standard deviation of W .
- (c) The shop charges a flat rate of \$1.50 per minute to examine the computer, and if no repairs are ordered, there is also an additional \$50 service charge. Let $L = 1.5x_1 + 50$. Compute the mean, variance, and standard deviation of L .

19. **Combination of Random Variables: Insurance Risk** Insurance companies know the risk of insurance is greatly reduced if the company insures not just one person, but many people. How does this work? Let x be a random variable representing the expectation of life in years for a 25-year-old male (i.e., number of years until death). Then the mean and standard deviation of x are $\mu = 50.2$ years and $\sigma = 11.5$ years (Vital Statistics Section of the *Statistical Abstract of the United States*, 116th Edition).

Suppose Big Rock Insurance Company has sold life insurance policies to Joel and David. Both are 25 years old, unrelated, live in different states, and have about the same health record. Let x_1 and x_2 be random variables representing Joel's and David's life expectancies. It is reasonable to assume x_1 and x_2 are independent.

$$\text{Joel, } x_1: \mu_1 = 50.2; \sigma_1 = 11.5$$

$$\text{David, } x_2: \mu_2 = 50.2; \sigma_2 = 11.5$$

If life expectancy can be predicted with more accuracy, Big Rock will have less risk in its insurance business. Risk in this case is measured by σ (larger σ means more risk).

- (a) The average life expectancy for Joel and David is $W = 0.5x_1 + 0.5x_2$. Compute the mean, variance, and standard deviation of W .
- (b) Compare the mean life expectancy for a single policy (x_1) with that for two policies (W).
- (c) Compare the standard deviation of the life expectancy for a single policy (x_1) with that for two policies (W).
- (d) The mean life expectancy is the same for a single policy (x_1) as it is for two policies (W), but the standard deviation is smaller for two policies. What happens to the mean life expectancy and the standard deviation when we include more policies issued to people whose life expectancies have the same mean



and standard deviation (i.e., 25-year-old males)? For instance, for three policies, $W = (\mu + \mu + \mu)/3 = \mu$ and $\sigma_W^2 = (1/3)^2\sigma^2 + (1/3)^2\sigma^2 + (1/3)^2\sigma^2 = (1/3)^2(3\sigma^2) = (1/3)\sigma^2$ and $\sigma_W = \frac{1}{\sqrt{3}}\sigma$. Likewise, for n such policies, $W = \mu$ and $\sigma_W^2 = (1/n)\sigma^2$ and $\sigma_W = \frac{1}{\sqrt{n}}\sigma$. Looking at the general result, is it appropriate to say that when we increase the number of policies to n , the risk decreases by a factor of $\sigma_W = \frac{1}{\sqrt{n}}$?

SECTION 5.2

Binomial Probabilities

FOCUS POINTS

- List the defining features of a binomial experiment.
- Compute binomial probabilities using the formula $P(r) = C_{n,r} p^r q^{n-r}$.
- Use the binomial table to find $P(r)$.
- Use the binomial probability distribution to solve real-world applications.

Binomial Experiment

On a TV game show, each contestant has a try at the wheel of fortune. The wheel of fortune is a roulette wheel with 36 slots, one of which is gold. If the ball lands in the gold slot, the contestant wins \$50,000. No other slot pays. What is the probability that the game show will have to pay the fortune to three contestants out of 100?

In this problem, the contestant and the game show sponsors are concerned about only two outcomes from the wheel of fortune: The ball lands on the gold, or the ball does not land on the gold. This problem is typical of an entire class of problems that are characterized by the feature that there are exactly two possible outcomes (for each trial) of interest. These problems are called *binomial experiments*, or *Bernoulli experiments*, after the Swiss mathematician Jacob Bernoulli, who studied them extensively in the late 1600s.

Binomial experiments

Features of a binomial experiment

Number of trials, n

Independent trials

Success, S

Failure, F

$$P(S) = p$$

$$P(F) = q = 1 - p$$

Number of successes, r

Features of a binomial experiment

1. There is a *fixed number of trials*. We denote this number by the letter n .
2. The n trials are *independent* and repeated under identical conditions.
3. Each trial has only *two outcomes*: success, denoted by S , and failure, denoted by F .
4. For each individual trial, the *probability of success is the same*. We denote the probability of success by p and that of failure by q . Since each trial results in either success or failure, $p + q = 1$ and $q = 1 - p$.
5. The central problem of a binomial experiment is to find the *probability of r successes out of n trials*.

EXAMPLE 4

BINOMIAL EXPERIMENT

Let's see how the wheel of fortune problem meets the criteria of a binomial experiment. We'll take the criteria one at a time.

SOLUTION:

1. Each of the 100 contestants has a trial at the wheel, so there are $n = 100$ trials in this problem.

2. Assuming that the wheel is fair, the *trials are independent*, since the result of one spin of the wheel has no effect on the results of other spins.
3. We are interested in only two outcomes on each spin of the wheel: The ball either lands on the gold, or it does not. Let's call landing on the gold *success (S)* and not landing on the gold *failure (F)*. In general, the assignment of the terms *success* and *failure* to outcomes does not imply good or bad results. These terms are assigned simply for the user's convenience.
4. On each trial the probability p of success (landing on the gold) is $1/36$, since there are 36 slots and only one of them is gold. Consequently, the probability of failure is

$$q = 1 - p = 1 - \frac{1}{36} = \frac{35}{36}$$

on each trial.

5. We want to know the probability of 3 successes out of 100 trials, so $r = 3$ in this example. It turns out that the probability the quiz show will have to pay the fortune to 3 contestants out of 100 is about 0.23. Later in this section we'll see how this probability was computed.



Anytime we make selections from a population *without replacement*, we do not have *independent trials*. However, replacement is often not practical. If the number of trials is quite small with respect to the population, we *almost* have independent trials, and we can say the situation is *closely approximated* by a binomial experiment. For instance, suppose we select 20 tuition bills at random from a collection of 10,000 bills issued at one college and observe if each bill is in error or not. If 600 of the 10,000 bills are in error, then the probability that the first one selected is in error is $600/10,000$, or 0.0600. If the first is in error, then the probability that the second is in error is $599/9999$, or 0.0599. Even if the first 19 bills selected are in error, the probability that the 20th is also in error is $581/9981$, or 0.0582. All these probabilities round to 0.06, and we can say that the independence condition is approximately satisfied.

GUIDED EXERCISE 4

Binomial experiment

Let's analyze the following binomial experiment to determine p , q , n , and r :

According to the *Textbook of Medical Physiology*, 5th Edition, by Arthur Guyton, 9% of the population has blood type B. Suppose we choose 18 people at random from the population and test the blood type of each. What is the probability that three of these people have blood type B? Note: Independence is approximated because 18 people is an extremely small sample with respect to the entire population.

- (a) In this experiment, we are observing whether or not a person has type B blood. We will say we have a success if the person has type B blood. What is failure?  Failure occurs if a person does not have type B blood.
- (b) The probability of success is 0.09, since 9% of the population has type B blood. What is the probability of failure, q ?  The probability of failure is

$$q = 1 - p$$

$$= 1 - 0.09 = 0.91$$
- (c) In this experiment, there are $n =$ _____ trials.  In this experiment, $n = 18$.

Continued

GUIDED EXERCISE 4 *continued*

- (d) We wish to compute the probability of 3 successes out of 18 trials. In this case,
 $r = \underline{\hspace{2cm}}$.

 In this case, $r = 3$.

Next, we will see how to compute the probability of r successes out of n trials when we have a binomial experiment.

Computing Probabilities for a Binomial Experiment Using the Binomial Distribution Formula

The central problem of a binomial experiment is finding the probability of r successes out of n trials. Now we'll see how to find these probabilities.

A model with three trials

Suppose you are taking a timed final exam. You have three multiple-choice questions left to do. Each question has four suggested answers, and only one of the answers is correct. You have only 5 seconds left to do these three questions, so you decide to mark answers on the answer sheet without even reading the questions. Assuming that your answers are randomly selected, what is the probability that you get zero, one, two, or all three questions correct?

This is a binomial experiment. Each question can be thought of as a trial, so there are $n = 3$ trials. The possible outcomes on each trial are success S , indicating a correct response, or failure F , meaning a wrong answer. The trials are independent—the outcome of any one trial does not affect the outcome of the others.

What is the *probability of success* on anyone question? Since you are guessing and there are four answers from which to select, the probability of a correct answer is 0.25. The probability q of a wrong answer is then 0.75. In short, we have a binomial experiment with $n = 3$, $p = 0.25$, and $q = 0.75$.

Now, what are the possible outcomes in terms of success or failure for these three trials? Let's use the notation SSF to mean success on the first question, success on the second, and failure on the third. There are eight possible combinations of S 's and F 's. They are

$SSS \quad SSF \quad SFS \quad FSS \quad SFF \quad FSF \quad FFS \quad FFF$

To compute the probability of each outcome, we can use the multiplication law because the trials are independent. For instance, the probability of success on the first two questions and failure on the last is

$$P(SSF) = P(S) \cdot P(S) \cdot P(F) = p \cdot p \cdot q = p^2q = (0.25)^2(0.75) \approx 0.047$$

In a similar fashion, we can compute the probability of each of the eight outcomes. These are shown in Table 5-8, along with the number of successes r associated with each outcome.

TABLE 5-8 Outcomes for a Binomial Experiment with $n = 3$ Trials

Outcome	Probability of Outcome	r (number of successes)
SSS	$P(SSS) = P(S)P(S)P(S) = p^3 = (0.25)^3 \approx 0.016$	3
SSF	$P(SSF) = P(S)P(S)P(F) = p^2q = (0.25)^2(0.75) \approx 0.047$	2
SFS	$P(SFS) = P(S)P(F)P(S) = p^2q = (0.25)^2(0.75) \approx 0.047$	2
FSS	$P(FSS) = P(F)P(S)P(S) = p^2q = (0.25)^2(0.75) \approx 0.047$	2
SFF	$P(SFF) = P(S)P(F)P(F) = pq^2 = (0.25)(0.75)^2 \approx 0.141$	1
FSF	$P(FSF) = P(F)P(S)P(F) = pq^2 = (0.25)(0.75)^2 \approx 0.141$	1
FFS	$P(FFS) = P(F)P(F)P(S) = pq^2 = (0.25)(0.75)^2 \approx 0.141$	1
FFF	$P(FFF) = P(F)P(F)P(F) = q^3 = (0.75)^3 \approx 0.422$	0

Now we can compute the probability of r successes out of three trials for $r = 0, 1, 2$, or 3 . Let's compute $P(1)$. The notation $P(1)$ stands for the probability of one success. For three trials, there are three different outcomes that show exactly one success. They are the outcomes SFF , FSF , and FFS . Since the outcomes are mutually exclusive, we can add the probabilities. So,

$$\begin{aligned} P(1) &= P(SFF \text{ or } FSF \text{ or } FFS) = P(SFF) + P(FSF) + P(FFS) \\ &= pq^2 + pq^2 + pq^2 \\ &= 3pq^2 \\ &= 3(0.25)(0.75)^2 \\ &= 0.422 \end{aligned}$$

In the same way, we can find $P(0)$, $P(2)$, and $P(3)$. These values are shown in Table 5-9.

TABLE 5-9 $P(r)$ for $n = 3$ Trials, $p = 0.25$

r (number of successes)	$P(r)$ (probability of r successes in 3 trials)	$P(r)$ for $p = 0.25$
0	$P(0) = P(FFF) = q^3$	0.422
1	$P(1) = P(SFF) + P(FSF) + P(FFS) = 3pq^2$	0.422
2	$P(2) = P(SSF) + P(SFS) + P(FSS) = 3p^2q$	0.141
3	$P(3) = P(SSS) = p^3$	0.016

We have done quite a bit of work to determine your chances of $r = 0, 1, 2$, or 3 successes on three multiple-choice questions if you are just guessing. Now we see that there is only a small chance (about 0.016) that you will get them all correct.

Table 5-9 can be used as a model for computing the probability of r successes out of only *three* trials. How can we compute the probability of 7 successes out of 10 trials? We can develop a table for $n = 10$, but this would be a tremendous task because there are 1024 possible combinations of successes and failures on 10 trials. Fortunately, mathematicians have given us a direct formula to compute the probability of r successes for any number of trials.

General formula for binomial probability distribution

Formula for the binomial probability distribution

$$P(r) = \frac{n!}{r!(n-r)!} p^r q^{n-r} = C_{n,r} p^r q^{n-r}$$

where n = number of trials

p = probability of success on each trial

$q = 1 - p$ = probability of failure on each trial

r = random variable representing the number of successes out of n trials ($0 \leq r \leq n$)

$!$ = factorial notation. Recall from Section 4.3 that the factorial symbol $n!$ designates the product of all the integers between 1 and n . For instance, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. Special cases are $1! = 1$ and $0! = 1$.

Table for $C_{n,r}$

$c_{n,r} = \frac{n!}{r!(n-r)!}$ is the binomial coefficient. Table 2 of Appendix II gives values of $C_{n,r}$ for select n and r . Many calculators have a key designated nCr that gives the value of $C_{n,r}$ directly.

Binomial coefficient $C_{n,r}$

Note: The binomial coefficient $C_{n,r}$ represents the number of combinations of n distinct objects (n = number of trials in this case) taken r at a time (r = number of successes). For more information about $C_{n,r}$, see Section 4.3.

Let's look more carefully at the formula for $P(r)$. There are two main parts. The expression $p^r q^{n-r}$ is the probability of getting one outcome with r successes and $n - r$ failures. The binomial coefficient $C_{n,r}$ counts the number of outcomes that have r successes and $n - r$ failures. For instance, in the case of $n = 3$ trials, we saw in Table 5-8 that the probability of getting an outcome with one success and two failures was pq^2 . This is the value of $p^r q^{n-r}$ when $r = 1$ and $n = 3$. We also observed that there were three outcomes with one success and two failures, so $C_{3,1}$ is 3.

Now let's take a look at an application of the binomial distribution formula in Example 5.

 **EXAMPLE 5****COMPUTE $P(r)$ USING THE BINOMIAL DISTRIBUTION FORMULA**

Privacy is a concern for many users of the Internet. One survey showed that 59% of Internet users are somewhat concerned about the confidentiality of their e-mail. Based on this information, what is the probability that for a random sample of 10 Internet users, 6 are concerned about the privacy of their e-mail?

SOLUTION:

- (a) This is a binomial experiment with 10 trials. If we assign success to an Internet user being concerned about the privacy of e-mail, the probability of success is 59%. We are interested in the probability of 6 successes. We have

$$n = 10 \quad p = 0.59 \quad q = 0.41 \quad r = 6$$

By the formula,

$$\begin{aligned} P(6) &= C_{10,6}(0.59)^6(0.41)^{10-6} \\ &= 210(0.59)^6(0.41)^4 \\ &\approx 210(0.0422)(0.0283) \\ &\approx 0.25 \end{aligned}$$

Use Table 2 of Appendix II or a calculator.

Use a calculator.

FIGURE 5-2

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Display

```
10 nCr 6*.59^6*.
41^(10-6)
.2503034245
```

There is a 25% chance that *exactly* 6 of the 10 Internet users are concerned about the privacy of e-mail.

- (b) Many calculators have a built-in combinations function. On the TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) calculators, press the MATH key and select PRB. The combinations function is designated nCr. Figure 5-2 displays the process for computing $P(6)$ directly on these calculators.

 **Using a Binomial Distribution Table**

In many cases we will be interested in the probability of a range of successes. In such cases, we need to use the addition rule for mutually exclusive events. For instance, for $n = 6$ and $p = 0.50$,

$$\begin{aligned} P(4 \text{ or fewer successes}) &= P(r \leq 4) \\ &= P(r = 4 \text{ or } 3 \text{ or } 2 \text{ or } 1 \text{ or } 0) \\ &= P(4) + P(3) + P(2) + P(1) + P(0) \end{aligned}$$

It would be a bit of a chore to use the binomial distribution formula to compute all the required probabilities. Table 3 of Appendix II gives values of $P(r)$ for

selected p values and values of n through 20. To use the table, find the appropriate section for n , and then use the entries in the columns headed by the p values and the rows headed by the r values.

Table 5-10 is an excerpt from Table 3 of Appendix II showing the section for $n = 6$. Notice that all possible r values between 0 and 6 are given as row headers. The value $p = 0.50$ is one of the column headers. For $n = 6$ and $p = 0.50$, you can find the value of $P(4)$ by looking at the entry in the row headed by 4 and the column headed by 0.50. Notice that $P(4) = 0.234$.

TABLE 5-10 Excerpt from Table 3 of Appendix II for $n = 6$

n	r	P												
		.01	.05	.1030507085	.90	.95
:														
6	0	.941	.735	.531118016001000	.000	.000
	1	.057	.232	.354303094010000	.000	.000
	2	.001	.031	.098324234060006	.001	.000
	3	.000	.002	.015185312185042	.015	.002
	4	.000	.000	.001060234324176	.098	.031
	5	.000	.000	.000010094303399	.354	.232
	6	.000	.000	.000001016118377	.531	.735

Likewise, you can find other values of $P(r)$ from the table. In fact, for $n = 6$ and $p = 0.50$,

$$\begin{aligned} P(r \leq 4) &= P(4) + P(3) + P(2) + P(1) + P(0) \\ &= 0.234 + 0.312 + 0.234 + 0.094 + 0.016 = 0.890 \end{aligned}$$

Alternatively, to compute $P(r \leq 4)$ for $n = 6$, you can use the fact that the total of all $P(r)$ values for r between 0 and 6 is 1 and the complement rule. Since the complement of the event $r \leq 4$ is the event $r \geq 5$, we have

$$\begin{aligned} P(r \leq 4) &= 1 - P(5) - P(6) \\ &= 1 - 0.094 - 0.016 = 0.890 \end{aligned}$$

Note: In Table 3 of Appendix II, probability entries of 0.000 do not mean the probability is exactly zero. Rather, to three digits after the decimal, the probability rounds to 0.000.

CRITICAL THINKING

In Chapter 4, we saw the complement rule of probability. As we saw in the previous discussion, this rule provides a useful strategy to simplify binomial probability computations for a range of successes. For example, in a binomial experiment with $n = 7$ trials, the sample space for the number of successes r is

0 1 2 3 4 5 6 7

Notice that $r = 0$ successes is the complement of $r \geq 1$ successes. By the complement rule,

$$P(r \geq 1) = 1 - P(r = 0)$$

It is faster to compute or look up $P(r = 0)$ and subtract than it is to compute or look up all the probabilities $P(r = 1)$ through $P(r = 7)$.

Likewise, the outcome $r \leq 2$ is the complement of the outcome $r \geq 3$, so,

$$P(r \geq 3) = 1 - P(r \leq 2) = 1 - P(r = 2) - P(r = 1) - P(r = 0)$$

Before you use the complement rule for computing probabilities, be sure the outcomes used comprise complementary events. Complements can be expressed in several ways. For instance, the complement of the event $r \leq 2$ can be expressed as event $r > 2$ or event $r \geq 3$.



Because of rounding in the binomial probability table, probabilities computed by using the addition rule directly might differ slightly from corresponding probabilities computed by using the complement rule.

EXAMPLE 6

USING THE BINOMIAL DISTRIBUTION TABLE TO FIND $P(r)$

A biologist is studying a new hybrid tomato. It is known that the seeds of this hybrid tomato have probability 0.70 of germinating. The biologist plants six seeds.

- (a) What is the probability that *exactly* four seeds will germinate?

SOLUTION: This is a binomial experiment with $n = 6$ trials. Each seed planted represents an independent trial. We'll say germination is success, so the probability for success on each trial is 0.70.

$$n = 6 \quad p = 0.70 \quad q = 0.30 \quad r = 4$$

We wish to find $P(4)$, the probability of exactly four successes.

In Table 3, Appendix II, find the section with $n = 6$ (excerpt is given in Table 5-10). Then find the entry in the column headed by $p = 0.70$ and the row headed by $r = 4$. This entry is 0.324.

$$P(4) = 0.324$$

- (b) What is the probability that *at least* four seeds will germinate?

SOLUTION: In this case, we are interested in the probability of four or more seeds germinating. This means we are to compute $P(r \geq 4)$. Since the events are mutually exclusive, we can use the addition rule

$$P(r \geq 4) = P(r = 4) \quad \text{or} \quad r = 5 \quad \text{or} \quad r = 6) = P(4) + P(5) + P(6)$$

We already know the value of $P(4)$. We need to find $P(5)$ and $P(6)$.

Use the same part of the table but find the entries in the row headed by the r value 5 and then the r value 6. Be sure to use the column headed by the value of p , 0.70.

$$P(5) = 0.303 \quad \text{and} \quad P(6) = 0.118$$

Now we have all the parts necessary to compute $P(r \geq 4)$.

$$\begin{aligned} P(r \geq 4) &= P(4) + P(5) + P(6) \\ &= 0.324 + 0.303 + 0.118 \\ &= 0.745 \end{aligned}$$



In Guided Exercise 5 you'll practice using the formula for $P(r)$ in one part, and then you'll use Table 3, Appendix II, for $P(r)$ values in the second part.



GUIDED EXERCISE 5**Find $P(r)$**

A rarely performed and somewhat risky eye operation is known to be successful in restoring the eyesight of 30% of the patients who undergo the operation. A team of surgeons has developed a new technique for this operation that has been successful in four of six operations. Does it seem likely that the new technique is much better than the old? We'll use the binomial probability distribution to answer this question. We'll compute the probability of at least four successes in six trials for the old technique.

(a) Each operation is a binomial trial. In this case, $\rightarrow n = 6, p = 0.30, q = 1 - 0.30 = 0.70, r = 4$
 $n = \underline{\quad}, p = \underline{\quad}, q = \underline{\quad}, r = \underline{\quad}$.

(b) Use your values of n , p , and q , as well as Table 2 of Appendix II (or your calculator), to compute $P(4)$ from the formula:

$$P(r) = C_{n,r} p^r q^{n-r}$$

(c) Compute the probability of *at least* four successes out of the six trials.

$$\begin{aligned} P(r \geq 4) &= P(r = 4 \text{ or } r = 5 \text{ or } r = 6) \\ &= P(4) + P(5) + P(6) \end{aligned}$$

Use Table 3 of Appendix II to find values of $P(4)$, $P(5)$, and $P(6)$. Then use these values to compute $P(r \geq 4)$.

(d) **Interpretation** Under the older operation technique, the probability that at least four patients out of six regain their eyesight is $\underline{\quad}$. Does it seem that the new technique is better than the old? Would you encourage the surgeon team to do more work on the new technique?

$$n = 6, p = 0.30, q = 1 - 0.30 = 0.70, r = 4$$

$$\begin{aligned} P(4) &= C_{6,4}(0.30)^4(0.70)^2 \\ &= 15(0.0081)(0.490) \\ &\approx 0.060 \end{aligned}$$

To find $P(4)$, $P(5)$, and $P(6)$ in Table 3, we look in the section labeled $n = 6$. Then we find the column headed by $p = 0.30$. To find $P(4)$, we use the row labeled $r = 4$. For the values of $P(5)$ and $P(6)$, we look in the same column but in the rows headed by $r = 5$ and $r = 6$, respectively.

$$\begin{aligned} P(r \geq 4) &= P(4) + P(5) + P(6) \\ &= 0.060 + 0.010 + 0.001 = 0.071 \end{aligned}$$

It seems the new technique is better than the old since, by pure chance, the probability of four or more successes out of six trials is only 0.071 for the old technique. This means one of the following two things may be happening:

- (i) The new method is no better than the old method, and our surgeons have encountered a rare event (probability 0.071), or
- (ii) The new method is in fact better. We think it is worth encouraging the surgeons to do more work on the new technique.

Using Technology to Compute Binomial Probabilities

Some calculators and computer software packages support the binomial distribution. In general, these technologies will provide both the probability $P(r)$ for an exact number of successes r and the cumulative probability $P(r \leq k)$, where k is a specified value less than or equal to the number of trials n . Note that most of the technologies use the letter x instead of r for the random variable denoting the number of successes out of n trials.

**TECH NOTES**

The software packages Minitab and Excel 2007, as well as the TI-84Plus/TI-83Plus/TI-nspire calculators, include built-in binomial probability distribution options. These options give the probability $P(r)$ of a specific number of successes r , as well as the cumulative total probability for r or fewer successes.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Press the DISTR key and scroll to binompdf (n, p, r). Enter the number of trials n , the probability of success on a single trial p , and the number of successes r . This gives $P(r)$. For the cumulative probability that there are r or fewer successes, use binomcdf (n, p, r).

```
binompdf(6,.3,4)
```

$P(r) = 4$.059535

```
binomcdf(6,.3,4)
```

$P(r \leq 4)$.989065

Excel 2007 Click the **Insert Function** (f_x). In the dialogue box, select **Statistical** for the category, and then select **Binomdist**. In the next dialogue box, fill in the values r , n , and p . For $P(r)$, use false; for $P(\text{at least } r \text{ successes})$, use true.

Minitab First, enter the r values $0, 1, 2, \dots, n$ in a column. Then use menu choice **Calc** \blacktriangleright **Probability Distribution** \blacktriangleright **Binomial**. In the dialogue box, select **Probability** for $P(r)$ or **Cumulative** for $P(\text{at least } r \text{ successes})$. Enter the number of trials n , the probability of success p , and the column containing the r values. A sample printout is shown in Problem 25 at the end of this section.

LOOKING FORWARD

There are several ways to find the probability of r successes out of n binomial trials. In particular we have used the general formula for the binomial probability distribution, a table of binomial probabilities such as Table 3 of Appendix II, or results from calculator or computer software that is based on the formula. However, depending on the number of trials and the size of the probability of success, the calculations can become tedious, or rounding errors can become an issue (see *Linking Concepts*, Problem 2). In Section 6.6 we will see how to use the *normal* probability distribution to approximate the binomial distribution in the case of sufficiently large numbers of trials n . Section 5.4 shows how to use a *Poisson* probability distribution to approximate the binomial distribution when the probability of success on a single trial p is very small and the number of trials is 100 or more.

Common expressions and corresponding inequalities



Problems 29 and 30 show how to compute conditional probabilities for binomial experiments. In these problems we see how to compute the probability of r successes out of n binomial trials, given that a certain number of successes will occur.

Many times we are asked to compute the probability of a range of successes. For instance, in a binomial experiment with n trials, we may be asked to compute the probability of four or more successes. Table 5-11 shows how common English expressions such as “four or more successes” translate to inequalities involving r .

**TABLE 5-11 Common English Expressions and Corresponding Inequalities
(consider a binomial experiment with n trials and r successes)**

Expression	Inequality
Four or more successes	$r \geq 4$
At least four successes	That is, $r = 4, 5, 6, \dots, n$
No fewer than four successes	
Not less than four successes	
Four or fewer successes	$r \leq 4$
At most four successes	That is, $r = 0, 1, 2, 3, \text{ or } 4$
No more than four successes	
The number of successes does not exceed four	
More than four successes	$r > 4$
The number of successes exceeds four	That is, $r = 5, 6, 7, \dots, n$
Fewer than four successes	$r < 4$
The number of successes is not as large as four	That is, $r = 0, 1, 2, 3$

Interpretation Often we are not interested in the probability of a *specific number* of successes out of n binomial trials. Rather, we are interested in a minimum number of successes, a maximum number of successes, or a range of a number of successes. In other words, we are interested in the probability of *at least* a certain number of successes or *no more than* a certain number of successes, or the probability that the number of successes is between two given values.

For instance, suppose engineers have determined that at least 3 of 5 rivets on a bridge connector need to hold. If the probability that a single rivet holds is 0.80, and the performances of the rivets are independent, then the engineers are interested in the probability that *at least* 3 of the rivets hold. Notice that $P(r \geq 3 \text{ rivets hold}) = 0.943$, while $P(r = 3 \text{ rivets hold}) = 0.205$. The probability of *at least* 3 successes is much higher than the probability of *exactly* 3 successes. Safety concerns require that 3 of the rivets hold. However, there is a greater margin of safety if more than 3 rivets hold.

As you consider binomial experiments, determine whether you are interested in a *specific number* of successes or a *range* of successes.

Sampling Without Replacement: Use of the Hypergeometric Probability Distribution

If the population is relatively small and we draw samples without replacement, the assumption of independent trials is not valid and we should not use the binomial distribution.

The *hypergeometric distribution* is a probability distribution of a random variable that has two outcomes when sampling is done without replacement. This is the distribution that is appropriate when the sample size is so small that sampling without replacement results in trials that are not even approximately independent. A discussion of the hypergeometric distribution can be found in Appendix I.

Hypergeometric probability distribution

VIEWPOINT

Lies! Lies!! Lies!!! The Psychology of Deceit

This is the title of an intriguing book by C. V. Ford, professor of psychiatry. The book recounts the true story of Floyd "Buzz" Fay, who was falsely convicted of murder on the basis of a failed polygraph examination. During his 2 ½ years of wrongful imprisonment, Buzz became a polygraph expert. He taught inmates, who freely confessed guilt, how to pass a polygraph examination.

(For more information on this topic, see Problem 21.)

SECTION 5.2 PROBLEMS

1. | **Statistical Literacy** What does the random variable for a binomial experiment of n trials measure?
2. | **Statistical Literacy** What does it mean to say that the trials of an experiment are independent?
3. | **Statistical Literacy** For a binomial experiment, how many outcomes are possible for each trial? What are the possible outcomes?
4. | **Statistical Literacy** In a binomial experiment, is it possible for the probability of success to change from one trial to the next? Explain.
5. | **Interpretation** Suppose you are a hospital manager and have been told that there is no need to worry that respirator monitoring equipment might fail because the probability any one monitor will fail is only 0.01. The hospital has 20 such moni-

- tors and they work independently. Should you be more concerned about the probability that *exactly one* of the 20 monitors fails, or that *at least one* fails? Explain.
6. **Interpretation** From long experience a landlord knows that the probability an apartment in a complex will not be rented is 0.10. There are 20 apartments in the complex, and the rental status of each apartment is independent of the status of the others. When a minimum of 16 apartment units are rented, the landlord can meet all monthly expenses. Which probability is more relevant to the landlord in terms of being able to meet expenses: the probability that there are *exactly four* unrented units or the probability that there are *four or fewer* unrented units? Explain.
7. **Critical Thinking** In an experiment, there are n independent trials. For each trial, there are three outcomes, A, B, and C. For each trial, the probability of outcome A is 0.40; the probability of outcome B is 0.50; and the probability of outcome C is 0.10. Suppose there are 10 trials.
- Can we use the binomial experiment model to determine the probability of four outcomes of type A, five of type B, and one of type C? Explain.
 - Can we use the binomial experiment model to determine the probability of four outcomes of type A and six outcomes that are not of type A? Explain. What is the probability of success on each trial?
8. **Critical Thinking** In a carnival game, there are six identical boxes, one of which contains a prize. A contestant wins the prize by selecting the box containing it. Before each game, the old prize is removed and another prize is placed at random in one of the six boxes. Is it appropriate to use the binomial probability distribution to find the probability that a contestant who plays the game five times wins exactly twice? Check each of the requirements of a binomial experiment and give the values of n , r , and p .
9. **Critical Thinking** According to the college registrar's office, 40% of students enrolled in an introductory statistics class this semester are freshmen, 25% are sophomores, 15% are juniors, and 20% are seniors. You want to determine the probability that in a random sample of five students enrolled in introductory statistics this semester, exactly two are freshmen.
- Describe a trial. Can we model a trial as having only two outcomes? If so, what is success? What is failure? What is the probability of success?
 - We are sampling without replacement. If only 30 students are enrolled in introductory statistics this semester, is it appropriate to model 5 trials as independent, with the same probability of success on each trial? Explain. What other probability distribution would be more appropriate in this setting?
10. **Critical Thinking: Simulation** Central Eye Clinic advertises that 90% of its patients approved for LASIK surgery to correct vision problems have successful surgeries.
- In the random-number table, assign the digits 0 through 8 to the event "successful surgery" and the digit 9 to the event "unsuccessful surgery." Does this assignment of digits simulate 90% successful outcomes?
 - Use the random-digit assignment model of part (a) to simulate the outcomes of 15 trials. Begin at column 1, line 2.
 - Your friend assigned the digits 1 through 9 to the event "successful surgery" and the digit 0 to the event "unsuccessful surgery." Does this assignment of digits simulate 90% successful outcomes? Using this digit assignment, repeat part (b).

In each of the following problems, the binomial distribution will be used. Answers may vary slightly depending on whether the binomial distribution formula, the binomial distribution table, or distribution results from a calculator or computer are used. Please answer the following questions and then complete the problem.

What makes up a trial? What is a success? What is a failure?

What are the values of n , p , and q ?

11. **Basic Computation: Binomial Distribution** Consider a binomial experiment with $n = 7$ trials where the probability of success on a single trial is $p = 0.30$.
- Find $P(r = 0)$.
 - Find $P(r \geq 1)$ by using the complement rule.
12. **Basic Computation: Binomial Distribution** Consider a binomial experiment with $n = 7$ trials where the probability of success on a single trial is $p = 0.60$.
- Find $P(r = 7)$.
 - Find $P(r \leq 6)$ by using the complement rule.
13. **Basic Computation: Binomial Distribution** Consider a binomial experiment with $n = 6$ trials where the probability of success on a single trial is $p = 0.85$.
- Find $P(r \leq 1)$.
 - Interpretation** If you conducted the experiment and got fewer than 2 successes, would you be surprised? Why?
14. **Basic Computation: Binomial Distribution** Consider a binomial experiment with $n = 6$ trials where the probability of success on a single trial is $p = 0.20$.
- Find $P(0 < r \leq 2)$.
 - Interpretation** If you conducted the experiment and got 1 or 2 successes, would you be surprised? Why?
15. **Binomial Probabilities: Coin Flip** A fair quarter is flipped three times. For each of the following probabilities, use the formula for the binomial distribution and a calculator to compute the requested probability. Next, look up the probability in Table 3 of Appendix II and compare the table result with the computed result.
- Find the probability of getting exactly three heads.
 - Find the probability of getting exactly two heads.
 - Find the probability of getting two or more heads.
 - Find the probability of getting exactly three tails.
16. **Binomial Probabilities: Multiple-Choice Quiz** Richard has just been given a 10-question multiple-choice quiz in his history class. Each question has five answers, of which only one is correct. Since Richard has not attended class recently, he doesn't know any of the answers. Assuming that Richard guesses on all 10 questions, find the indicated probabilities.
- What is the probability that he will answer all questions correctly?
 - What is the probability that he will answer all questions incorrectly?
 - What is the probability that he will answer at least one of the questions correctly? Compute this probability two ways. First, use the rule for mutually exclusive events and the probabilities shown in Table 3 of Appendix II. Then use the fact that $P(r \geq 1) = 1 - P(r = 0)$. Compare the two results. Should they be equal? Are they equal? If not, how do you account for the difference?
 - What is the probability that Richard will answer at least half the questions correctly?
17. **Ecology: Wolves** The following is based on information taken from *The Wolf in the Southwest: The Making of an Endangered Species*, edited by David Brown (University of Arizona Press). Before 1918, approximately 55% of the wolves in the New Mexico and Arizona region were male, and 45% were female. However, cattle ranchers in this area have made a determined effort to exterminate wolves. From 1918 to the present, approximately 70% of wolves in the region are male, and 30% are female. Biologists suspect that male wolves are more likely than females to return to an area where the population has been greatly reduced.
- Before 1918, in a random sample of 12 wolves spotted in the region, what was the probability that 6 or more were male? What was the probability that 6 or more were female? What was the probability that fewer than 4 were female?
 - Answer part (a) for the period from 1918 to the present.

18. **Sociology: Ethics** The one-time fling! Have you ever purchased an article of clothing (dress, sports jacket, etc.), worn the item *once* to a party, and then returned the purchase? This is called a *one-time fling*. About 10% of all adults deliberately do a one-time fling and feel no guilt about it (Source: *Are You Normal?*, by Bernice Kanner, St. Martin's Press). In a group of seven adult friends, what is the probability that
(a) no one has done a one-time fling?
(b) at least one person has done a one-time fling?
(c) no more than two people have done a one-time fling?
19. **Sociology: Mother-in-Law** Sociologists say that 90% of married women claim that their husband's mother is the biggest bone of contention in their marriages (sex and money are lower-rated areas of contention). (See the source in Problem 18.) Suppose that six married women are having coffee together one morning. What is the probability that
(a) all of them dislike their mother-in-law?
(b) none of them dislike their mother-in-law?
(c) at least four of them dislike their mother-in-law?
(d) no more than three of them dislike their mother-in-law?
20. **Sociology: Dress Habits** A research team at Cornell University conducted a study showing that approximately 10% of all businessmen who wear ties wear them so tightly that they actually reduce blood flow to the brain, diminishing cerebral functions (Source: *Chances: Risk and Odds in Everyday Life*, by James Burke). At a board meeting of 20 businessmen, all of whom wear ties, what is the probability that
(a) at least one tie is too tight?
(b) more than two ties are too tight?
(c) no tie is too tight?
(d) at least 18 ties are *not* too tight?
21. **Psychology: Deceit** Aldrich Ames is a convicted traitor who leaked American secrets to a foreign power. Yet Ames took routine lie detector tests and each time passed them. How can this be done? Recognizing control questions, employing unusual breathing patterns, biting one's tongue at the right time, pressing one's toes hard to the floor, and counting backward by 7 are countermeasures that are difficult to detect but can change the results of a polygraph examination (Source: *Lies! Lies!! Lies!!! The Psychology of Deceit*, by C. V. Ford, professor of psychiatry, University of Alabama). In fact, it is reported in Professor Ford's book that after only 20 minutes of instruction by "Buzz" Fay (a prison inmate), 85% of those trained were able to pass the polygraph examination even when guilty of a crime. Suppose that a random sample of nine students (in a psychology laboratory) are told a "secret" and then given instructions on how to pass the polygraph examination without revealing their knowledge of the secret. What is the probability that
(a) all the students are able to pass the polygraph examination?
(b) more than half the students are able to pass the polygraph examination?
(c) no more than four of the students are able to pass the polygraph examination?
(d) all the students fail the polygraph examination?
22. **Hardware Store: Income** Trevor is interested in purchasing the local hardware/sporting goods store in the small town of Dove Creek, Montana. After examining accounting records for the past several years, he found that the store has been grossing over \$850 per day about 60% of the business days it is open. Estimate the probability that the store will gross over \$850
(a) at least 3 out of 5 business days.
(b) at least 6 out of 10 business days.
(c) fewer than 5 out of 10 business days.

- (d) fewer than 6 out of the next 20 business days. **Interpretation** If this actually happened, might it shake your confidence in the statement $p = 0.60$? Might it make you suspect that p is less than 0.60? Explain.
- (e) more than 17 out of the next 20 business days. **Interpretation** If this actually happened, might you suspect that p is greater than 0.60? Explain.
23. **Psychology: Myers–Briggs** Approximately 75% of all marketing personnel are extroverts, whereas about 60% of all computer programmers are introverts (Source: *A Guide to the Development and Use of the Myers–Briggs Type Indicator*, by Myers and McCaulley).
- (a) At a meeting of 15 marketing personnel, what is the probability that 10 or more are extroverts? What is the probability that 5 or more are extroverts? What is the probability that all are extroverts?
- (b) In a group of 5 computer programmers, what is the probability that none are introverts? What is the probability that 3 or more are introverts? What is the probability that all are introverts?
24. **Business Ethics: Privacy** Are your finances, buying habits, medical records, and phone calls really private? A real concern for many adults is that computers and the Internet are reducing privacy. A survey conducted by Peter D. Hart Research Associates for the Shell Poll was reported in *USA Today*. According to the survey, 37% of adults are concerned that employers are monitoring phone calls. Use the binomial distribution formula to calculate the probability that
- (a) out of five adults, none is concerned that employers are monitoring phone calls.
- (b) out of five adults, all are concerned that employers are monitoring phone calls.
- (c) out of five adults, exactly three are concerned that employers are monitoring phone calls.
25. **Business Ethics: Privacy** According to the same poll quoted in Problem 24, 53% of adults are concerned that Social Security numbers are used for general identification. For a group of eight adults selected at random, we used Minitab to generate the binomial probability distribution and the cumulative binomial probability distribution (menu selections \blacktriangleright Calc \blacktriangleright Probability Distributions \blacktriangleright Binomial).

Jeff Titcomb/Photographer's
Choice/Getty Images



Number	r	$P(r)$	$P(\leq r)$
0	0	0.002381	0.00238
1	1	0.021481	0.02386
2	2	0.084781	0.10864
3	3	0.191208	0.29985
4	4	0.269521	0.56937
5	5	0.243143	0.81251
6	6	0.137091	0.94960
7	7	0.044169	0.99377
8	8	0.006226	1.00000

Find the probability that out of eight adults selected at random,

- (a) at most five are concerned about Social Security numbers being used for identification. Do the problem by adding the probabilities $P(r = 0)$ through $P(r = 5)$. Is this the same as the cumulative probability $P(r \leq 5)$?
- (b) more than five are concerned about Social Security numbers being used for identification. First, do the problem by adding the probabilities $P(r = 6)$ through $P(r = 8)$. Then do the problem by subtracting the cumulative probability $P(r \leq 5)$ from 1. Do you get the same results?

26. **Health Care: Office Visits** What is the age distribution of patients who make office visits to a doctor or nurse? The following table is based on information taken from the Medical Practice Characteristics section of the *Statistical Abstract of the United States* (116th Edition).

Age group, years	Under 15	15–24	25–44	45–64	65 and older
Percent of office visitors	20%	10%	25%	20%	25%

Suppose you are a district manager of a health management organization (HMO) that is monitoring the office of a local doctor or nurse in general family practice. This morning the office you are monitoring has eight office visits on the schedule. What is the probability that

- (a) at least half the patients are under 15 years old? First, explain how this can be modeled as a binomial distribution with 8 trials, where success is visitor age is under 15 years old and the probability of success is 20%.
- (b) from 2 to 5 patients are 65 years old or older (include 2 and 5)?
- (c) from 2 to 5 patients are 45 years old or older (include 2 and 5)? Hint: Success is 45 or older. Use the table to compute the probability of success on a single trial.
- (d) all the patients are under 25 years of age?
- (e) all the patients are 15 years old or older?

27. **Binomial Distribution Table: Symmetry** Study the binomial distribution table (Table 3, Appendix II). Notice that the probability of success on a single trial p ranges from 0.01 to 0.95. Some binomial distribution tables stop at 0.50 because of the symmetry in the table. Let's look for that symmetry. Consider the section of the table for which $n = 5$. Look at the numbers in the columns headed by $p = 0.30$ and $p = 0.70$. Do you detect any similarities? Consider the following probabilities for a binomial experiment with five trials.
- (a) Compare $P(3 \text{ successes})$, where $p = 0.30$, with $P(2 \text{ successes})$, where $p = 0.70$.
 - (b) Compare $P(3 \text{ or more successes})$, where $p = 0.30$, with $P(2 \text{ or fewer successes})$, where $p = 0.70$.
 - (c) Find the value of $P(4 \text{ successes})$, where $p = 0.30$. For what value of r is $P(r \text{ successes})$ the same using $p = 0.70$?
 - (d) What column is symmetrical with the one headed by $p = 0.20$?

28. **Binomial Distribution: Control Charts** This problem will be referred to in the study of control charts (Section 6.1). In the binomial probability distribution, let the number of trials be $n = 3$, and let the probability of success be $p = 0.0228$. Use a calculator to compute
- (a) the probability of two successes.
 - (b) the probability of three successes.
 - (c) the probability of two or three successes.



29. **Expand Your Knowledge: Conditional Probability** In the western United States, there are many dry land wheat farms that depend on winter snow and spring rain to produce good crops. About 65% of the years, there is enough moisture to produce a good wheat crop, depending on the region (Reference: *Agricultural Statistics*, U.S. Department of Agriculture).

- (a) Let r be a random variable that represents the number of good wheat crops in $n = 8$ years. Suppose the Zimmer farm has reason to believe that at least 4 out of 8 years will be good. However, they need at least 6 good years out of 8 to survive financially. Compute the probability that the Zimmers will get at least 6 good years out of 8, given what they believe is true; that is, compute $P(6 \leq r \mid 4 \leq r)$. See part (d) for a hint.
- (b) Let r be a random variable that represents the number of good wheat crops in $n = 10$ years. Suppose the Montoya farm has reason to believe that at least 6 out of 10 years will be good. However, they need at least 8 good years



out of 10 to survive financially. Compute the probability that the Montoyas will get at least 8 good years out of 10, given what they believe is true; that is, compute $P(8 \leq r | 6 \leq r)$.

- List at least three other areas besides agriculture to which you think conditional binomial probabilities can be applied.
- Hint for solution:* Review item 6, conditional probability, in the summary of basic probability rules at the end of Section 4.2. Note that

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

and show that in part (a),

$$P(6 \leq r | 4 \leq r) = \frac{P((6 \leq r) \text{ and } (4 \leq r))}{P(4 \leq r)} = \frac{P(6 \leq r)}{P(4 \leq r)}$$



30.

Conditional Probability: Blood Supply Only about 70% of all donated human blood can be used in hospitals. The remaining 30% cannot be used because of various infections in the blood. Suppose a blood bank has 10 newly donated pints of blood. Let r be a binomial random variable that represents the number of “good” pints that can be used.

- Based on questionnaires completed by the donors, it is believed that at least 6 of the 10 pints are usable. What is the probability that at least 8 of the pints are usable, given this belief is true? Compute $P(8 \leq r | 6 \leq r)$.
- Assuming the belief that at least 6 of the pints are usable is true, what is the probability that all 10 pints can be used? Compute $P(r = 10 | 6 \leq r)$.

Hint: See Problem 29.

SECTION 5.3

Additional Properties of the Binomial Distribution

FOCUS POINTS

- Make histograms for binomial distributions.
- Compute μ and σ for a binomial distribution.
- Compute the minimum number of trials n needed to achieve a given probability of success $P(r)$.

Graphing a Binomial Distribution

Any probability distribution may be represented in graphic form. How should we graph the binomial distribution? Remember, the binomial distribution tells us the probability of r successes out of n trials. Therefore, we'll place values of r along the horizontal axis and values of $P(r)$ on the vertical axis. The binomial distribution is a *discrete* probability distribution because r can assume only whole-number values such as 0, 1, 2, 3, . . . Therefore, a histogram is an appropriate graph of a binomial distribution.

PROCEDURE

HOW TO GRAPH A BINOMIAL DISTRIBUTION

- Place r values on the horizontal axis.
- Place $P(r)$ values on the vertical axis.
- Construct a bar over each r value extending from $r - 0.5$ to $r + 0.5$. The height of the corresponding bar is $P(r)$.

Let's look at an example to see exactly how we'll make these histograms.

EXAMPLE 7**GRAPH OF A BINOMIAL DISTRIBUTION**

A waiter at the Green Spot Restaurant has learned from long experience that the probability that a lone diner will leave a tip is only 0.7. During one lunch hour, the waiter serves six people who are dining by themselves. Make a graph of the binomial probability distribution that shows the probabilities that 0, 1, 2, 3, 4, 5, or all 6 lone diners leave tips.

SOLUTION: This is a binomial experiment with $n = 6$ trials. Success is achieved when the lone diner leaves a tip, so the probability of success is 0.7 and that of failure is 0.3:

$$n = 6 \quad p = 0.7 \quad q = 0.3$$

We want to make a histogram showing the probability of r successes when $r = 0, 1, 2, 3, 4, 5$, or 6. It is easier to make the histogram if we first make a table of r values and the corresponding $P(r)$ values (Table 5-12). We'll use Table 3 of Appendix II to find the $P(r)$ values for $n = 6$ and $p = 0.70$.

FIGURE 5-3
Graph of the Binomial Distribution for $n = 6$ and $p = 0.7$

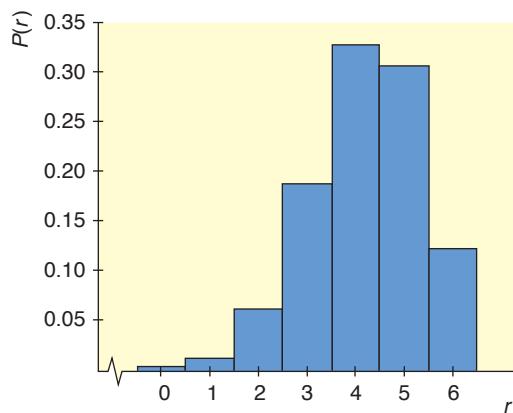


TABLE 5-12 Binomial Distribution for $n = 6$ and $p = 0.70$

r	P(r)
0	0.001
1	0.010
2	0.060
3	0.185
4	0.324
5	0.303
6	0.118

Mark Richards/PhotoEdit



To construct the histogram, we'll put r values on the horizontal axis and $P(r)$ values on the vertical axis. Our bars will be one unit wide and will be centered over the appropriate r value. The height of the bar over a particular r value tells the probability of that r (see Figure 5-3).

The probability of a particular value of r is given not only by the height of the bar over that r value but also by the *area* of the bar. Each bar is only one unit wide, so its area (area = height times width) equals its height. Since the area of each bar represents the probability of the r value under it, the sum of the areas of the bars must be 1. In this example, the sum turns out to be 1.001. It is not exactly equal to 1 because of rounding error.

GUIDED EXERCISE 6**Graph of a binomial distribution**

Jim enjoys playing basketball. He figures that he makes about 50% of the field goals he attempts during a game. Make a histogram showing the probability that Jim will make 0, 1, 2, 3, 4, 5, or 6 shots out of six attempted field goals.

Continued

GUIDED EXERCISE 6 *continued*

(a) This is a binomial experiment with $n = \underline{\hspace{2cm}}$ trials. In this situation, we'll say success occurs when Jim makes an attempted field goal. What is the value of p ?

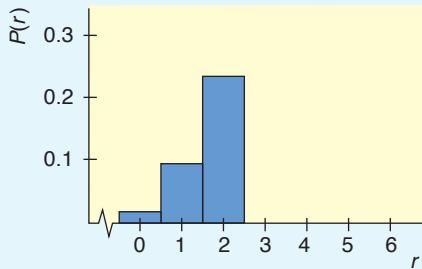
(b) Use Table 3 of Appendix II to complete Table 5-13 of $P(r)$ values for $n = 6$ and $p = 0.5$.

TABLE 5-13

r	$P(r)$
0	0.016
1	0.094
2	0.234
3	_____
4	_____
5	_____
6	_____

In this example, $n = 6$ and $p = 0.5$.

(c) Use the values of $P(r)$ given in Table 5-14 to complete the histogram in Figure 5-4.

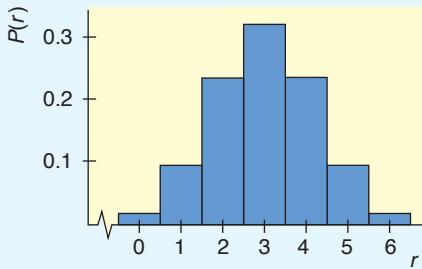
FIGURE 5-4 Beginning of Graph of Binomial Distribution for $n = 6$ and $p = 0.5$ 

(d) The area of the bar over $r = 2$ is 0.234. What is the area of the bar over $r = 4$? How does the probability that Jim makes exactly two field goals out of six compare with the probability that he makes exactly four field goals out of six?

TABLE 5-14 Completion of Table 5-13

r	$P(r)$
3	0.312
4	0.234
5	0.094
6	0.016

FIGURE 5-5 Completion of Figure 5-4



The area of the bar over $r = 4$ is also 0.234. Jim is as likely to make two out of six field goals attempted as he is to make four out of six.

In Example 7 and Guided Exercise 6, we see the graphs of two binomial distributions associated with $n = 6$ trials. The two graphs are different because the probability of success p is different in the two cases. In Example 7, $p = 0.7$ and the graph is skewed to the left—that is, the left tail is longer. In Guided Exercise 6, p is equal to 0.5 and the graph is symmetrical—that is, if we fold it in half, the two halves coincide exactly. Whenever p equals 0.5, the graph of the binomial distribution will be symmetrical no matter how many trials we have. In Chapter 6, we will see that if the number of trials n is quite large, the binomial distribution is almost symmetrical over the bars containing most of the area even when p is not close to 0.5.

Mean and Standard Deviation of a Binomial Distribution

Two other features that help describe the graph of any distribution are the balance point of the distribution and the spread of the distribution about that balance point. The *balance point* is the mean μ of the distribution, and the *measure of spread* that is most commonly used is the standard deviation σ . The mean μ is the *expected value* of the number of successes.

For the binomial distribution, we can use two special formulas to compute the mean μ and the standard deviation σ . These are easier to use than the general formulas in Section 5.1 for μ and σ of any discrete probability distribution.

PROCEDURE

Mean of a binomial distribution

Standard deviation of a binomial distributions

HOW TO COMPUTE μ AND σ FOR A BINOMIAL DISTRIBUTION

$\mu = np$ is the **expected number of successes** for the random variable r

$\sigma = \sqrt{npq}$ is the **standard deviation** for the random variable r

where

r is a random variable representing the number of successes in a binomial distribution,

n is the number of trials,

p is the probability of success on a single trial, and

$q = 1 - p$ is the probability of failure on a single trial.

EXAMPLE 8

COMPUTE μ AND σ

Let's compute the mean and standard deviation for the distribution of Example 7 that describes that probabilities of lone diners leaving tips at the Green Spot Restaurant.

SOLUTION: In Example 7,

$$n = 6 \quad p = 0.7 \quad q = 0.3$$

For the binomial distribution,

$$\mu = np = 6(0.7) = 4.2$$

The balance point of the distribution is at $\mu = 4.2$. The standard deviation is given by

$$\sigma = \sqrt{npq} = \sqrt{6(0.7)(0.3)} = \sqrt{1.26} \approx 1.12$$

The mean μ is not only the balance point of the distribution; it is also the *expected value* of r . Specifically, in Example 7, the waiter can expect 4.2 lone diners out of 6 to leave a tip. (The waiter would probably round the expected value to 4 tippers out of 6.)

GUIDED EXERCISE 7

Expected value and standard deviation

When Jim (of Guided Exercise 6) shoots field goals in basketball games, the probability that he makes a shot is only 0.5.

Continued

GUIDED EXERCISE 7 *continued*

- (a) The mean of the binomial distribution is the expected value of r successes out of n trials. Out of six throws, what is the expected number of goals Jim will make?
- (b) For six trials, what is the standard deviation of the binomial distribution of the number of successful field goals Jim makes?



The expected value is the mean μ :

$$\mu = np = 6(0.5) = 3$$

Jim can expect to make three goals out of six tries.



$$\sigma = \sqrt{npq} = \sqrt{6(0.5)(0.5)} = \sqrt{1.5} \approx 1.22$$


CRITICAL THINKING

Unusual Values

Chebyshev's theorem tells us that no matter what the data distribution looks like, at least 75% of the data will fall within 2 standard deviations of the mean. As we will see in Chapter 6, when the distribution is mound-shaped and symmetrical, about 95% of the data are within 2 standard deviations of the mean. Data values beyond 2 standard deviations from the mean are less common than those closer to the mean.

In fact, one indicator that a data value might be an outlier is that it is more than 2.5 standard deviations from the mean (Source: *Statistics*, by G. Upton and I. Cook, Oxford University Press).

Unusual values

For a binomial distribution, it is unusual for the number of successes r to be higher than $\mu + 2.5\sigma$ or lower than $\mu - 2.5\sigma$.

We can use this indicator to determine whether a specified number of successes out of n trials in a binomial experiment are unusual.

For instance, consider a binomial experiment with 20 trials for which probability of success on a single trial is $p = 0.70$. The expected number of successes is $\mu = 14$, with a standard deviation of $\sigma \approx 2$. A number of successes above 19 or below 9 would be considered unusual. However, such numbers of successes are possible.



Quota Problems: Minimum Number of Trials for a Given Probability

In applications, you do not want to confuse the expected value of r with certain probabilities associated with r . Guided Exercise 8 illustrates this point.

GUIDED EXERCISE 8

Find the minimum value of n for a given $P(r)$

A satellite is powered by three solar cells. The probability that any one of these cells will fail is 0.15, and the cells operate or fail independently.

Part I: In this part, we want to find the least number of cells the satellite should have so that the *expected value* of the number of working cells is no smaller than 3. In this situation, n represents the number of cells, r is the number of successful or working cells, p is the probability that a cell will work, q is the probability that a cell will fail, and μ is the expected value, which should be no smaller than 3.

Continued

GUIDED EXERCISE 8 *continued*(a) What is the value of q ? of p ?

→ $q = 0.15$, as given in the problem. p must be 0.85, since $p = 1 - q$.

(b) The expected value μ for the number of working cells is given by $\mu = np$. The expected value of the number of working cells should be no smaller than 3, so

$$3 \leq \mu = np$$

From part (a), we know the value of p . Solve the inequality $3 \leq np$ for n .

(c) Since n is between 3 and 4, should we round it to 3 or to 4 to make sure that μ is at least 3?

→ $3 \leq np$
 $3 \leq n(0.85)$
 $\frac{3}{0.85} \leq n$ Divide both sides by 0.85.
 $3.53 \leq n$

→ n should be at least 3.53. Since we can't have a fraction of a cell, we had best make $n = 4$. For $n = 4$, $\mu = 4(0.85) = 3.4$. This value satisfies the condition that μ be at least 3.

Part II: In this part, we want to find the smallest number of cells the satellite should have to be 97% sure that there will be adequate power—that is, that at least three cells work.

(a) The letter r has been used to denote the number of successes. In this case, r represents the number of working cells. We are trying to find the number n of cells necessary to ensure that (choose the correct statement)

→ $P(r \geq 3) = 0.97$

(i) $P(r \geq 3) = 0.97$ or

(ii) $P(r \leq 3) = 0.97$

(b) We need to find a value for n such that

$$P(r \geq 3) = 0.97$$

Try $n = 4$. Then, $r \geq 3$ means $r = 3$ or 4, so,

$$P(r \geq 3) = P(3) + P(4)$$

Use Table 3 (Appendix II) with $n = 4$ and $p = 0.85$ to find values of $P(3)$ and $P(4)$. Then, compute $P(r \geq 3)$ for $n = 4$. Will $n = 4$ guarantee that $P(r \geq 3)$ is at least 0.97?

(c) Now try $n = 5$ cells. For $n = 5$,

$$P(r \geq 3) = P(3) + P(4) + P(5)$$

since r can be 3, 4, or 5. Are $n = 5$ cells adequate? [Be sure to find new values of $P(3)$ and $P(4)$, since we now have $n = 5$.]

→ $P(3) = 0.368$

$P(4) = 0.522$

$P(r \geq 3) = 0.368 + 0.522 = 0.890$

Thus, $n = 4$ is *not* sufficient to be 97% sure that at least three cells will work. For $n = 4$, the probability that at least three cells will work is only 0.890.

→ $P(r \geq 3) = P(3) + P(4) + P(5)$
 $= 0.138 + 0.392 + 0.444$
 $= 0.974$

Thus, $n = 5$ cells are required if we want to be 97% sure that at least three cells will work.

In Part I and Part II, we got different values for n . Why? In Part I, we had $n = 4$ and $\mu = 3.4$. This means that if we put up lots of satellites with four cells, we can expect that an *average* of 3.4 cells will work per satellite. But for $n = 4$ cells, there is a probability of only 0.89 that at least three cells will work in any one satellite. In Part II, we are trying to find the number of cells necessary so that the probability is 0.97 that at least three cells will work in any *one* satellite. If we use $n = 5$ cells, then we can satisfy this requirement.

Quota problems

Quotas occur in many aspects of everyday life. The manager of a sales team gives every member of the team a weekly sales quota. In some districts, police have a monthly quota for the number of traffic tickets issued. Nonprofit organizations have recruitment quotas for donations or new volunteers. The basic ideas

used to compute quotas also can be used in medical science (how frequently checkups should occur), quality control (how many production flaws should be expected), or risk management (how many bad loans a bank should expect in a certain investment group). In fact, Part II of Guided Exercise 8 is a *quota problem*. To have adequate power, a satellite must have a quota of three working solar cells. Such problems come from many different sources, but they all have one thing in common: They are solved using the binomial probability distribution.

To solve quota problems, it is often helpful to use equivalent formulas for expressing binomial probabilities. These formulas involve the complement rule and the fact that binomial events are independent. Equivalent probabilities will be used in Example 9.

PROCEDURE

HOW TO EXPRESS BINOMIAL PROBABILITIES USING EQUIVALENT FORMULAS

$$P(\text{at least one success}) = P(r \geq 1) = 1 - P(0)$$

$$P(\text{at least two successes}) = P(r \geq 2) = 1 - P(0) - P(1)$$

$$P(\text{at least three successes}) = P(r \geq 3) = 1 - P(0) - P(1) - P(2)$$

$$P(\text{at least } m \text{ successes}) = P(r \geq m) = 1 - P(0) - P(1) - \cdots - P(m-1), \\ \text{where } 1 \leq m \leq \text{number of trials}$$

For a discussion of the mathematics behind these formulas, see Problem 26 at the end of this section.

Example 9 is a quota problem. Junk bonds are sometimes controversial. In some cases, junk bonds have been the salvation of a basically good company that has had a run of bad luck. From another point of view, junk bonds are not much more than a gambler's effort to make money by shady ethics.

The book *Liar's Poker*, by Michael Lewis, is an exciting and sometimes humorous description of his career as a Wall Street bond broker. Most bond brokers, including Mr. Lewis, are ethical people. However, the book does contain an interesting discussion of Michael Milken and shady ethics. In the book, Mr. Lewis says, "If it was a good deal, the brokers kept it for themselves; if it was a bad deal, they'd try to sell it to their customers." In Example 9, we use some binomial probabilities for a brief explanation of what Mr. Lewis's book is talking about.

EXAMPLE 9

QUOTA

Junk bonds can be profitable as well as risky. Why are investors willing to consider junk bonds? Suppose you can buy junk bonds at a tremendous discount. You try to choose "good" companies with a "good" product. The company should have done well but for some reason did not. Suppose you consider only companies with a 35% estimated risk of default, and your financial investment goal requires four bonds to be "good" bonds in the sense that they will not default before a certain date. Remember, junk bonds that do not default are usually very profitable because they carry a very high rate of return. The other bonds in your investment group can default (or not) without harming your investment plan. Suppose you want to be 95% certain of meeting your goal (quota) of at least four good bonds. How many junk bond issues should you buy to meet this goal?

SOLUTION: Since the probability of default is 35%, the probability of a "good" bond is 65%. Let success S be represented by a good bond. Let n be the number of bonds purchased, and let r be the number of good bonds in this group. We want

$$P(r \geq 4) \geq 0.95$$

This is equivalent to

$$1 - P(0) - P(1) - P(2) - P(3) \geq 0.95$$

Since the probability of success is $p = P(S) = 0.65$, we need to look in the binomial table under $p = 0.65$ and different values of n to find the *smallest value of n* that will satisfy the preceding relation. Table 3 of Appendix II shows that if $n = 10$ when $p = 0.65$, then,

$$1 - P(0) - P(1) - P(2) - P(3) = 1 - 0 - 0 - 0.004 - 0.021 = 0.975$$

The probability 0.975 satisfies the condition of being greater than or equal to 0.95. We see that 10 is the smallest value of n for which the condition

$$P(r \geq 4) \geq 0.95$$

is satisfied. Under the given conditions (a good discount on price, no more than 35% chance of default, and a fundamentally good company), you can be 95% sure of meeting your investment goal with $n = 10$ (carefully selected) junk bond issues.

In this example, we see that by carefully selecting junk bonds, there is a high probability of getting some good bonds that will produce a real profit. What do you do with the other bonds that aren't so good? Perhaps the quote from *Liar's Poker* will suggest what is sometimes attempted.



VIEWPOINT

Kodiak Island, Alaska

Kodiak Island is famous for its giant brown bears. The sea surrounding the island is also famous for its king crab. The state of Alaska's, Department of Fish and Game has collected a huge amount of data regarding ocean latitude, ocean longitude, and size of king crab. Of special interest to commercial fishing skippers is the size of crab. Those too small must be returned to the sea. To find locations and sizes of king crab catches near Kodiak Island, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the StatLib site hosted by the Department of Statistics at Carnegie Mellon University. Once at StatLib, go to crab data. From this information, it is possible to use methods of this chapter and Chapter 7 to estimate the proportion of legal crab in a sea skipper's catch.

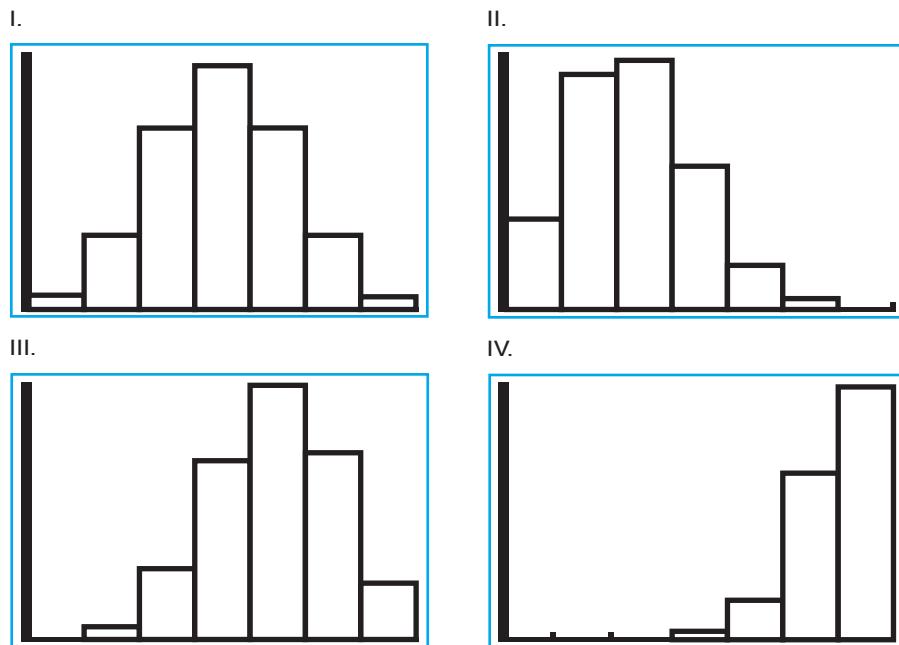
SECTION 5.3 PROBLEMS

1. **Statistical Literacy** What does the expected value of a binomial distribution with n trials tell you?
2. **Statistical Literacy** Consider two binomial distributions, with n trials each. The first distribution has a higher probability of success on each trial than the second. How does the expected value of the first distribution compare to that of the second?
3. **Basic Computation: Expected Value and Standard Deviation** Consider a binomial experiment with $n = 8$ trials and $p = 0.20$.
 - (a) Find the expected value and the standard deviation of the distribution.
 - (b) **Interpretation** Would it be unusual to obtain 5 or more successes? Explain. Confirm your answer by looking at the binomial probability distribution table.
4. **Basic Computation: Expected Value and Standard Deviation** Consider a binomial experiment with $n = 20$ trials and $p = 0.40$.
 - (a) Find the expected value and the standard deviation of the distribution.
 - (b) **Interpretation** Would it be unusual to obtain fewer than 3 successes? Explain. Confirm your answer by looking at the binomial probability distribution table.

5. **Critical Thinking** Consider a binomial distribution of 200 trials with expected value 80 and standard deviation of about 6.9. Use the criterion that it is unusual to have data values more than 2.5 standard deviations above the mean or 2.5 standard deviations below the mean to answer the following questions.
- Would it be unusual to have more than 120 successes out of 200 trials? Explain.
 - Would it be unusual to have fewer than 40 successes out of 200 trials? Explain.
 - Would it be unusual to have from 70 to 90 successes out of 200 trials? Explain.
6. **Critical Thinking** Consider a binomial distribution with 10 trials. Look at Table 3 (Appendix II) showing binomial probabilities for various values of p , the probability of success on a single trial.
- For what value of p is the distribution symmetric? What is the expected value of this distribution? Is the distribution centered over this value?
 - For small values of p , is the distribution skewed right or left?
 - For large values of p , is the distribution skewed right or left?
7. **Binomial Distribution: Histograms** Consider a binomial distribution with $n = 5$ trials. Use the probabilities given in Table 3 of Appendix II to make histograms showing the probabilities of $r = 0, 1, 2, 3, 4$, and 5 successes for each of the following. Comment on the skewness of each distribution.
- The probability of success is $p = 0.50$.
 - The probability of success is $p = 0.25$.
 - The probability of success is $p = 0.75$.
 - What is the relationship between the distributions shown in parts (b) and (c)?
 - If the probability of success is $p = 0.73$, do you expect the distribution to be skewed to the right or to the left? Why?
8. **Binomial Distributions: Histograms** Figure 5-6 shows histograms of several binomial distributions with $n = 6$ trials. Match the given probability of success with the best graph.
- $p = 0.30$ goes with graph _____.
 - $p = 0.50$ goes with graph _____.
 - $p = 0.65$ goes with graph _____.
 - $p = 0.90$ goes with graph _____.
 - In general, when the probability of success p is close to 0.5, would you say that the graph is more symmetrical or more skewed? In general, when the probability of success p is close to 1, would you say that the graph is skewed to the right or to the left? What about when p is close to 0?

FIGURE 5-6

Binomial Probability Distributions with $n = 6$ (generated on the TI-84Plus calculator)



9. **Marketing: Photography** Does the *kid factor* make a difference? If you are talking photography, the answer may be yes! The following table is based on information from *American Demographics* (Vol. 19, No. 7).

Ages of children in household, years	Under 2	None under 21
Percent of U.S. households that buy photo gear	80%	50%

Let us say you are a market research person who interviews a random sample of 10 households.

- (a) Suppose you interview 10 households with children under the age of 2 years. Let r represent the number of such households that buy photo gear. Make a histogram showing the probability distribution of r for $r = 0$ through $r = 10$. Find the mean and standard deviation of this probability distribution.
 - (b) Suppose that the 10 households are chosen to have no children under 21 years old. Let r represent the number of such households that buy photo gear. Make a histogram showing the probability distribution of r for $r = 0$ through $r = 10$. Find the mean and standard deviation of this probability distribution.
 - (c) **Interpretation** Compare the distributions in parts (a) and (b). You are designing TV ads to sell photo gear. Could you justify featuring ads of parents taking pictures of toddlers? Explain your answer.
10. **Quality Control: Syringes** The quality-control inspector of a production plant will reject a batch of syringes if two or more defective syringes are found in a random sample of eight syringes taken from the batch. Suppose the batch contains 1% defective syringes.
- (a) Make a histogram showing the probabilities of $r = 0, 1, 2, 3, 4, 5, 6, 7$, and 8 defective syringes in a random sample of eight syringes.
 - (b) Find μ . What is the expected number of defective syringes the inspector will find?
 - (c) What is the probability that the batch will be accepted?
 - (d) Find σ .
11. **Private Investigation: Locating People** Old Friends Information Service is a California company that is in the business of finding addresses of long-lost friends. Old Friends claims to have a 70% success rate (Source: *Wall Street Journal*). Suppose that you have the names of six friends for whom you have no addresses and decide to use Old Friends to track them.
- (a) Make a histogram showing the probability of $r = 0$ to 6 friends for whom an address will be found.
 - (b) Find the mean and standard deviation of this probability distribution. What is the expected number of friends for whom addresses will be found?
 - (c) **Quota Problem** How many names would you have to submit to be 97% sure that at least two addresses will be found?
12. **Insurance: Auto** The Mountain States Office of State Farm Insurance Company reports that approximately 85% of all automobile damage liability claims are made by people under 25 years of age. A random sample of five automobile insurance liability claims is under study.
- (a) Make a histogram showing the probability that $r = 0$ to 5 claims are made by people under 25 years of age.
 - (b) Find the mean and standard deviation of this probability distribution. For samples of size 5, what is the expected number of claims made by people under 25 years of age?
13. **Education: Illiteracy** USA Today reported that about 20% of all people in the United States are illiterate. Suppose you interview seven people at random off a city street.

- (a) Make a histogram showing the probability distribution of the number of illiterate people out of the seven people in the sample.
- (b) Find the mean and standard deviation of this probability distribution. Find the expected number of people in this sample who are illiterate.
- (c) **Quota Problem** How many people would you need to interview to be 98% sure that at least seven of these people can read and write (are not illiterate)?
14. **Rude Drivers: Tailgating** Do you tailgate the car in front of you? About 35% of all drivers will tailgate before passing, thinking they can make the car in front of them go faster (Source: Bernice Kanner, *Are You Normal?*, St. Martin's Press). Suppose that you are driving a considerable distance on a two-lane highway and are passed by 12 vehicles.
- (a) Let r be the number of vehicles that tailgate before passing. Make a histogram showing the probability distribution of r for $r = 0$ through $r = 12$.
- (b) Compute the expected number of vehicles out of 12 that will tailgate.
- (c) Compute the standard deviation of this distribution.
15. **Hype: Improved Products** The *Wall Street Journal* reported that approximately 25% of the people who are told a product is *improved* will believe that it is, in fact, improved. The remaining 75% believe that this is just hype (the same old thing with no real improvement). Suppose a marketing study consists of a random sample of eight people who are given a sales talk about a new, *improved* product.
- (a) Make a histogram showing the probability that $r = 0$ to 8 people believe the product is, in fact, improved.
- (b) Compute the mean and standard deviation of this probability distribution.
- (c) **Quota Problem** How many people are needed in the marketing study to be 99% sure that at least one person believes the product to be improved? Hint: Note that $P(r \geq 1) = 0.99$ is equivalent to $1 - P(0) = 0.99$, or $P(0) = 0.01$.
16. **Quota Problem: Archaeology** An archaeological excavation at Burnt Mesa Pueblo showed that about 10% of the flaked stone objects were finished arrow points (Source: *Bandelier Archaeological Excavation Project: Summer 1990 Excavations at Burnt Mesa Pueblo*, edited by Kohler, Washington State University). How many flaked stone objects need to be found to be 90% sure that at least one is a finished arrow point? Hint: Use a calculator and note that $P(r \geq 1) \geq 0.90$ is equivalent to $1 - P(0) \geq 0.90$, or $P(0) \leq 0.10$.
17. **Criminal Justice: Parole** USA Today reports that about 25% of all prison parolees become repeat offenders. Alice is a social worker whose job is to counsel people on parole. Let us say success means a person does not become a repeat offender. Alice has been given a group of four parolees.
- (a) Find the probability $P(r)$ of r successes ranging from 0 to 4.
- (b) Make a histogram for the probability distribution of part (a).
- (c) What is the expected number of parolees in Alice's group who will not be repeat offenders? What is the standard deviation?
- (d) **Quota Problem** How large a group should Alice counsel to be about 98% sure that three or more parolees will not become repeat offenders?
18. **Defense: Radar Stations** The probability that a single radar station will detect an enemy plane is 0.65.
- (a) **Quota Problem** How many such stations are required for 98% certainty that an enemy plane flying over will be detected by at least one station?
- (b) If four stations are in use, what is the expected number of stations that will detect an enemy plane?
19. **Criminal Justice: Jury Duty** Have you ever tried to get out of jury duty? About 25% of those called will find an excuse (work, poor health, travel out of town, etc.) to avoid jury duty (Source: Bernice Kanner, *Are You Normal?*, St. Martin's Press, New York). If 12 people are called for jury duty,

- (a) what is the probability that all 12 will be available to serve on the jury?
(b) what is the probability that 6 or more will *not* be available to serve on the jury?
(c) Find the expected number of those available to serve on the jury. What is the standard deviation?
(d) **Quota Problem** How many people n must the jury commissioner contact to be 95.9% sure of finding at least 12 people who are available to serve?
20. **Public Safety: 911 Calls** *The Denver Post* reported that a recent audit of Los Angeles 911 calls showed that 85% were not emergencies. Suppose the 911 operators in Los Angeles have just received four calls.
(a) What is the probability that all four calls are, in fact, emergencies?
(b) What is the probability that three or more calls are not emergencies?
(c) **Quota Problem** How many calls n would the 911 operators need to answer to be 96% (or more) sure that at least one call is, in fact, an emergency?
21. **Law Enforcement: Property Crime** Does crime pay? The *FBI Standard Survey of Crimes* shows that for about 80% of all property crimes (burglary, larceny, car theft, etc.), the criminals are never found and the case is never solved (Source: *True Odds*, by James Walsh, Merrit Publishing). Suppose a neighborhood district in a large city suffers repeated property crimes, not always perpetuated by the same criminals. The police are investigating six property crime cases in this district.
(a) What is the probability that none of the crimes will ever be solved?
(b) What is the probability that at least one crime will be solved?
(c) What is the expected number of crimes that will be solved? What is the standard deviation?
(d) **Quota Problem** How many property crimes n must the police investigate before they can be at least 90% sure of solving one or more cases?
22. **Security: Burglar Alarms** A large bank vault has several automatic burglar alarms. The probability is 0.55 that a single alarm will detect a burglar.
(a) **Quota Problem** How many such alarms should be used for 99% certainty that a burglar trying to enter will be detected by at least one alarm?
(b) Suppose the bank installs nine alarms. What is the expected number of alarms that will detect a burglar?
23. **Criminal Justice: Convictions** Innocent until proven guilty? In Japanese criminal trials, about 95% of the defendants are found guilty. In the United States, about 60% of the defendants are found guilty in criminal trials (Source: *The Book of Risks*, by Larry Laudan, John Wiley and Sons). Suppose you are a news reporter following seven criminal trials.
(a) If the trials were in Japan, what is the probability that all the defendants would be found guilty? What is this probability if the trials were in the United States?
(b) Of the seven trials, what is the expected number of guilty verdicts in Japan? What is the expected number in the United States? What is the standard deviation in each case?
(c) **Quota Problem** As a U.S. news reporter, how many trials n would you need to cover to be at least 99% sure of two or more convictions? How many trials n would you need if you covered trials in Japan?
24. **Focus Problem: Personality Types** We now have the tools to solve the Chapter Focus Problem. In the book *A Guide to the Development and Use of the Myers–Briggs Type Indicators* by Myers and McCaully, it was reported that approximately 45% of all university professors are extroverted. Suppose you have classes with six different professors.
(a) What is the probability that all six are extroverts?
(b) What is the probability that none of your professors is an extrovert?
(c) What is the probability that at least two of your professors are extroverts?

- (d) In a group of six professors selected at random, what is the *expected number* of extroverts? What is the *standard deviation* of the distribution?
- (e) **Quota Problem** Suppose you were assigned to write an article for the student newspaper and you were given a quota (by the editor) of interviewing at least three extroverted professors. How many professors selected at random would you need to interview to be at least 90% sure of filling the quota?
25. **Quota Problem: Motel Rooms** The owners of a motel in Florida have noticed that in the long run, about 40% of the people who stop and inquire about a room for the night actually rent a room.
- Quota Problem** How many inquiries must the owner answer to be 99% sure of renting at least one room?
 - If 25 separate inquiries are made about rooms, what is the expected number of inquiries that will result in room rentals?
26. **Critical Thinking** Let r be a binomial random variable representing the number of successes out of n trials.
- Explain why the sample space for r consists of the set $\{0, 1, 2, \dots, n\}$ and why the sum of the probabilities of all the entries in the entire sample space must be 1.
 - Explain why $P(r \geq 1) = 1 - P(0)$.
 - Explain why $P(r \geq 2) = 1 - P(0) - P(1)$.
 - Explain why $P(r \geq m) = 1 - P(0) - P(1) - \dots - P(m-1)$ for $1 \leq m \leq n$.

SECTION 5.4

The Geometric and Poisson Probability Distributions

FOCUS POINTS

- In many activities, the *first* to succeed wins everything! Use the geometric distribution to compute the probability that the n th trial is the first success.
- Use the Poisson distribution to compute the probability of the occurrence of events spread out over time or space.
- Use the Poisson distribution to approximate the binomial distribution when the number of trials is large and the probability of success is small.

In this chapter, we have studied binomial probabilities for the discrete random variable r , the number of successes in n binomial trials. Before we continue in Chapter 6 with continuous random variables, let us examine two other discrete probability distributions, the *geometric* and the *Poisson* probability distributions. These are both related to the binomial distribution.

Geometric Distribution

Suppose we have an experiment in which we repeat binomial trials until we get our *first success*, and then we stop. Let n be the number of the trial on which we get our *first success*. In this context, n is not a fixed number. In fact, n could be any of the numbers 1, 2, 3, and so on. What is the probability that our first success comes on the n th trial? The answer is given by the *geometric probability distribution*.

Geometric probability distribution

Geometric probability distribution

$$P(n) = p(1 - p)^{n-1}$$

where n is the number of the binomial trial on which the *first success* occurs ($n = 1, 2, 3, \dots$) and p is the probability of success on each trial. Note: p must be the same for each trial.

Using some mathematics involving infinite series, it can be shown that the **population mean** and **standard deviation** of the geometric distribution are

$$\mu = \frac{1}{p} \quad \text{and} \quad \sigma = \frac{\sqrt{1-p}}{p}$$

In many real-life situations, we keep on trying until we achieve success. This is true in areas as diverse as diplomacy, military science, real estate sales, general marketing strategies, medical science, engineering, and technology.

To Engineer Is Human: The Role of Failure in Successful Design is a fascinating book by Henry Petroski (a professor of engineering at Duke University). Reviewers for the *Los Angeles Times* describe this book as serious, amusing, probing, and sometimes frightening. The book examines topics such as the collapse of the Tacoma Narrows suspension bridge, the collapse of the Kansas City Hyatt Regency walkway, and the explosion of the space shuttle *Challenger*. Professor Petroski discusses such topics as “success in foreseeing failure” and the “limits of design.” What is meant by expressions such as “foreseeing failure” and the “limits of design”? In the next example, we will see how the geometric probability distribution might help us “forecast” failure.

EXAMPLE 10

FIRST SUCCESS

An automobile assembly plant produces sheet-metal door panels. Each panel moves on an assembly line. As the panel passes a robot, a mechanical arm will perform spot welding at different locations. Each location has a magnetic dot painted where the weld is to be made. The robot is programmed to locate the magnetic dot and perform the weld. However, experience shows that on each trial the robot is only 85% successful at locating the dot. If it cannot locate the magnetic dot, it is programmed to *try again*. The robot will keep trying until it finds the dot (and does the weld) or the door panel passes out of the robot’s reach.



ricardo azoury/stockphoto

- (a) What is the probability that the robot’s first success will be on attempts $n = 1, 2$, or 3 ?

SOLUTION: Since the robot will keep trying until it is successful, the geometric distribution is appropriate. In this case, success S means that the robot finds the correct location. The probability of success is $p = P(S) = 0.85$. The probabilities are

- n $P(n) = p(1 - p)^{n-1} = 0.85(0.15)^{n-1}$
- 1 $0.85(0.15)^0 = 0.85$
- 2 $0.85(0.15)^1 = 0.1275$
- 3 $0.85(0.15)^2 \approx 0.0191$

- (b) The assembly line moves so fast that the robot has a maximum of only three chances before the door panel is out of reach. What is the probability that the robot will be successful before the door panel is out of reach?

SOLUTION: Since $n = 1$ or 2 or 3 is mutually exclusive, then

$$\begin{aligned} P(n = 1 \text{ or } 2 \text{ or } 3) &= P(1) + P(2) + P(3) \\ &\approx 0.85 + 0.1275 + 0.0191 \\ &= 0.9966 \end{aligned}$$

This means that the weld should be correctly located about 99.7% of the time.

- (c) **Interpretation** What is the probability that the robot will not be able to locate the correct spot within three tries? If 10,000 panels are made, what is the expected number of defectives? Comment on the meaning of this answer in the context of “forecasting failure” and the “limits of design.”

SOLUTION: The probability that the robot will correctly locate the weld is 0.9966, from part (b). Therefore, the probability that it will not do so (after three unsuccessful tries) is $1 - 0.9966 = 0.0034$. If we made 10,000 panels, we would expect (forecast) $(10,000)(0.0034) = 34$ defectives. We could reduce this by inspecting every door, but such a solution is most likely too costly. If a defective weld of this type is not considered too dangerous, we can accept an expected 34 failures out of 10,000 panels due to the limits of our production design—that is, the speed of the assembly line and the accuracy of the robot. If this is not acceptable, a new (perhaps more costly) design is needed.



Problem 30 introduces the negative binomial distribution, with additional applications in Problems 31 and 32. Problem 33 explores the proof of the negative binomial distribution formula.

Poisson probability distribution

COMMENT The geometric distribution deals with binomial trials that are repeated until we have our *first success* on the n th trial. Suppose we repeat a binomial trial n times until we have k *successes* (not just one). In the literature, the probability distribution for the random variable n is called the *negative binomial distribution*. For more information on this topic, see Problems 30, 31, 32, and 33 at the end of this section.

Poisson Probability Distribution

If we examine the binomial distribution as the number of trials n gets larger and larger while the probability of success p gets smaller and smaller, we obtain the *Poisson distribution*. Siméon Denis Poisson (1781–1840) was a French mathematician who studied probabilities of rare events that occur infrequently in space, time, volume, and so forth. The Poisson distribution applies to accident rates, arrival times, defect rates, the occurrence of bacteria in the air, and many other areas of everyday life.

As with the binomial distribution, we assume only two outcomes: A particular event occurs (success) or does not occur (failure) during the specified time period or space. The events need to be independent so that one success does not change the probability of another success during the specified interval. We are interested in computing the probability of r occurrences in the given time period, space, volume, or specified interval.

Poisson distribution

Let λ (Greek letter lambda) be the mean number of successes over time, volume, area, and so forth. Let r be the number of successes ($r = 0, 1, 2, 3, \dots$) in a corresponding interval of time, volume, area, and so forth. Then the probability of r successes in the interval is

$$P(r) = \frac{e^{-\lambda} \lambda^r}{r!}$$

Continued

where e is approximately equal to 2.7183.

Using some mathematics involving infinite series, it can be shown that the **population mean** and **standard deviation** of the Poisson distribution are

$$\mu = \lambda \quad \text{and} \quad \sigma = \sqrt{\lambda}$$

Note: e^x is a key found on most calculators. Simply use 1 as the exponent, and the calculator will display a decimal approximation for e .

There are many applications of the Poisson distribution. For example, if we take the point of view that waiting time can be subdivided into many small intervals, then the actual arrival (of whatever we are waiting for) during any one of the very short intervals could be thought of as an infrequent (or rare) event. This means that the Poisson distribution can be used as a mathematical model to describe the probabilities of arrivals such as cars to a gas station, planes to an airport, calls to a fire station, births of babies, and even fish on a fisherman's line.

EXAMPLE 11

POISSON DISTRIBUTION

Pyramid Lake is located in Nevada on the Paiute Indian Reservation. The lake is described as a lovely jewel in a beautiful desert setting. In addition to its

natural beauty, the lake contains some of the world's largest cutthroat trout. Eight- to ten-pound trout are not uncommon, and 12- to 15-pound trophies are taken each season. The Paiute Nation uses highly trained fish biologists to study and maintain this famous fishery. In one of its publications, *Creel Chronicle* (Vol. 3, No. 2), the following information was given about the November catch for boat fishermen.

$$\text{Total fish per hour} = 0.667$$

Suppose you decide to fish Pyramid Lake for 7 hours during the month of November.

- (a) Use the information provided by the fishery biologist in *Creel Chronicle* to find a probability distribution for r , the number of fish (of all sizes) you catch in a period of 7 hours.

SOLUTION: For fish of all sizes, the mean success rate per hour is 0.667.

$$\lambda = 0.667/1\text{ hour}$$

Since we want to study a *7-hour interval*, we use a little arithmetic to adjust λ to 7 hours. That is, we adjust λ so that it represents the average number of fish expected in a 7-hour period.

$$\lambda = \frac{0.667}{1\text{ hour}} \cdot \left(\frac{7}{7}\right) = \frac{4.669}{7\text{ hour}}$$

For convenience, let us use the rounded value $\lambda = 4.7$ for a 7-hour period. Since r is the number of successes (fish caught) in the corresponding 7-hour period and $\lambda = 4.7$ for this period, we use the Poisson distribution to get

$$P(r) = \frac{e^{-\lambda} \lambda^r}{r!} = \frac{e^{-4.7} (4.7)^r}{r!}$$

Recall that $e \approx 2.7183$ Most calculators have e^x , y^x , and $n!$ keys (see your calculator manual), so the Poisson distribution is not hard to compute.

- (b) What is the probability that in 7 hours you will get 0, 1, 2, or 3 fish of any size?



Pyramid Lake, Nevada

SOLUTION: Using the result of part (a), we get

$$\begin{aligned}P(0) &= \frac{e^{-4.7}(4.7)^0}{0!} \approx 0.0091 \approx 0.01 \\P(1) &= \frac{e^{-4.7}(4.7)^1}{1!} \approx 0.0427 \approx 0.04 \\P(2) &= \frac{e^{-4.7}(4.7)^2}{2!} \approx 0.1005 \approx 0.10 \\P(3) &= \frac{e^{-4.7}(4.7)^3}{3!} \approx 0.1574 \approx 0.16\end{aligned}$$

The probabilities of getting 0, 1, 2, or 3 fish are about 1%, 4%, 10%, and 16%, respectively.

- (c) What is the probability that you will get four or more fish in the 7-hour fishing period?

SOLUTION: The sample space of all r values is $r = 0, 1, 2, 3, 4, 5, \dots$. The probability of the entire sample space is 1, and these events are mutually exclusive. Therefore,

$$1 = P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + \dots$$

So,

$$\begin{aligned}P(r \geq 4) &= P(4) + P(5) + \dots = 1 - P(0) - P(1) - P(2) - P(3) \\&\approx 1 - 0.01 - 0.04 - 0.10 - 0.16 \\&= 0.69\end{aligned}$$

There is about a 69% chance that you will catch four or more fish in a 7-hour period.

Use of tables

Table 4 of Appendix II is a table of the Poisson probability distribution for selected values of λ and the number of successes r . Table 5-15 is an excerpt from that table.

To find the value of $P(r = 2)$ when $\lambda = 0.3$, look in the column headed by 0.3 and the row headed by 2. From the table, we see that $P(2) = 0.0333$.

TABLE 5-15 Excerpt from Appendix II, Table 4, "Poisson Probability Distribution"

r	λ				
	0.1	0.2	0.3	0.4	0.5
0	.9048	.8187	.7408	.6703	.6065
1	.0905	.1637	.2222	.2681	.3033
2	.0045	.0164	.0333	.0536	.0758
3	.0002	.0011	.0033	.0072	.0126
4	.0000	.0001	.0003	.0007	.0016

LOOKING FORWARD

In Example 11 we studied the probability distribution of r , the *number* of fish caught. Now, what about the probability distribution of x , the *waiting time* between catching one fish and the next? It turns out that the random variable x has a continuous distribution called an *exponential distribution*. Continuous distributions will be studied in Chapter 6. To learn more about the exponential distribution and waiting time between catching fish, please see Problem 20 at the end of Section 6.1.



The TI-84Plus/TI-83Plus/TI-nspire calculators have commands for the geometric and Poisson distributions. Excel 2007 and Minitab support the Poisson distribution. All the technologies have both the probability distribution and the cumulative distribution.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Use the DISTR key and scroll to `geometpdf(p, n)` for the probability of first success on trial number n ; scroll to `geomecdf(p, n)` for the probability of first success on trial number $\leq n$. Use `poissonpdf(λ, r)` for the probability of r successes. Use `poissoncdf(λ, r)` for the probability of at least r successes. For example, when $\lambda = 0.25$ and $r = 2$, we get the following results.

```
poissonpdf(.25, 2)
.0243375245
poissoncdf(.25, 2)
.9978385033
```

Excel 2007 Click on the Insert Function f_x . In the dialogue box, select Statistical for the category, and then select Poisson. Enter the trial number r , and use λ for the mean. False gives the probability $P(r)$ and True gives the cumulative probability $P(\text{at least } r)$.

Minitab Put the r values in a column. Then use the menu choice Calc \blacktriangleright Probability Distribution \blacktriangleright Poisson. In the dialogue box, select probability or cumulative, enter the column number containing the r values, and use λ for the mean.

Poisson Approximation to the Binomial Probability Distribution

In the preceding examples, we have seen how the Poisson distribution can be used over intervals of time, space, area, and so on. However, the Poisson distribution also can be used as a probability distribution for “rare” events. In the binomial distribution, if the number of trials n is large while the probability p of success is quite small, we call the event (success) a “rare” event. Put another way, it can be shown that for most practical purposes, the Poisson distribution will be a very good *approximation to the binomial distribution*, provided the number of trials n is larger than or equal to 100 and $\lambda = np$ is less than 10. As n gets larger and p gets smaller, the approximation becomes better and better.

Poisson approximation to the binomial

PROCEDURE



Problems 28 and 29 show how to compute conditional probabilities for binomial experiments using the Poisson approximation for the binomial distribution. In these problems we see how to compute the probability of r successes out of n binomial trials, given that a certain number of successes will occur.

HOW TO APPROXIMATE BINOMIAL PROBABILITIES USING POISSON PROBABILITIES

Suppose you have a binomial distribution with

n = number of trials

r = number of successes

p = probability of success on each trial

Continued

If $n \geq 100$ and $np < 10$, then r has a binomial distribution that is approximated by a Poisson distribution with $\lambda = np$.

$$P(r) \approx \frac{e^{-\lambda} \lambda^r}{r!}$$

Note: $\lambda = np$ is the expected value of the binomial distribution.

EXAMPLE 12

POISSON APPROXIMATION TO THE BINOMIAL

Isabel Briggs Myers was a pioneer in the study of personality types. Today the Myers–Briggs Type Indicator is used in many career counseling programs as well as in many industrial settings where people must work closely together as a team. The 16 personality types are discussed in detail in the book *A Guide to the Development and Use of the Myers–Briggs Type Indicators*, by Myers and McCaulley. Each personality type has its own special contribution in any group activity. One of the more “rare” types is INFJ (introverted, intuitive, feeling, judgmental), which occurs in only about 2.1% of the population. Suppose a high school graduating class has 167 students, and suppose we call “success” the event that a student is of personality type INFJ.

- (a) Let r be the number of successes (INFJ students) in the $n = 167$ trials (graduating class). If $p = P(S) = 0.021$, will the Poisson distribution be a good approximation to the binomial?

SOLUTION: Since $n = 167$ is greater than 100 and $\lambda = np = 167(0.021) \approx 3.5$ is less than 10, the Poisson distribution should be a good approximation to the binomial.

- (b) Estimate the probability that this graduating class has 0, 1, 2, 3, or 4 people who have the INFJ personality type.

SOLUTION: Since Table 4 (Appendix II) for the Poisson distribution includes the values $\lambda = 3.5$ and $r = 0, 1, 2, 3$, or 4, we may simply look up the values for $P(r)$, $r = 0, 1, 2, 3, 4$:

$$\begin{aligned} P(r = 0) &= 0.0302 & P(r = 1) &= 0.1057 \\ P(r = 2) &= 0.1850 & P(r = 3) &= 0.2158 \\ P(r = 4) &= 0.1888 \end{aligned}$$

Since the outcomes $r = 0, 1, 2, 3$, or 4 successes are mutually exclusive, we can compute the probability of 4 or fewer INFJ types by using the addition rule for mutually exclusive events:

$$\begin{aligned} P(r \leq 4) &= P(0) + P(1) + P(2) + P(3) + P(4) \\ &= 0.0302 + 0.1057 + 0.1850 + 0.2158 + 0.1888 \\ &= 0.7255 \end{aligned}$$

The probability that the graduating class will have four or fewer INFJ personality types is about 0.73.

- (c) Estimate the probability that this class has five or more INFJ personality types.

SOLUTION: Because the outcomes of a binomial experiment are all mutually exclusive, we have

$$P(r \leq 4) + P(r \geq 5) = 1$$

$$\text{or } P(r \geq 5) = 1 - P(r \leq 4) = 1 - 0.7255 = 0.2745$$

The probability is approximately 0.27 that there will be five or more INFJ personality types in the graduating class.



EXAMPLE 12

POISSON APPROXIMATION TO THE BINOMIAL

Isabel Briggs Myers was a pioneer in the study of personality types. Today the Myers–Briggs Type Indicator is used in many career counseling programs as well as in many industrial settings where people must work closely together as a team. The 16 personality types are discussed in detail in the book *A Guide to the Development and Use of the Myers–Briggs Type Indicators*, by Myers and McCaulley. Each personality type has its own special contribution in any group activity. One of the more “rare” types is INFJ (introverted, intuitive, feeling, judgmental), which occurs in only about 2.1% of the population. Suppose a high school graduating class has 167 students, and suppose we call “success” the event that a student is of personality type INFJ.

- (a) Let r be the number of successes (INFJ students) in the $n = 167$ trials (graduating class). If $p = P(S) = 0.021$, will the Poisson distribution be a good approximation to the binomial?

SOLUTION: Since $n = 167$ is greater than 100 and $\lambda = np = 167(0.021) \approx 3.5$ is less than 10, the Poisson distribution should be a good approximation to the binomial.

- (b) Estimate the probability that this graduating class has 0, 1, 2, 3, or 4 people who have the INFJ personality type.

SOLUTION: Since Table 4 (Appendix II) for the Poisson distribution includes the values $\lambda = 3.5$ and $r = 0, 1, 2, 3$, or 4, we may simply look up the values for $P(r)$, $r = 0, 1, 2, 3, 4$:

$$\begin{aligned} P(r = 0) &= 0.0302 & P(r = 1) &= 0.1057 \\ P(r = 2) &= 0.1850 & P(r = 3) &= 0.2158 \\ P(r = 4) &= 0.1888 \end{aligned}$$

Since the outcomes $r = 0, 1, 2, 3$, or 4 successes are mutually exclusive, we can compute the probability of 4 or fewer INFJ types by using the addition rule for mutually exclusive events:

$$\begin{aligned} P(r \leq 4) &= P(0) + P(1) + P(2) + P(3) + P(4) \\ &= 0.0302 + 0.1057 + 0.1850 + 0.2158 + 0.1888 \\ &= 0.7255 \end{aligned}$$

The probability that the graduating class will have four or fewer INFJ personality types is about 0.73.

- (c) Estimate the probability that this class has five or more INFJ personality types.

SOLUTION: Because the outcomes of a binomial experiment are all mutually exclusive, we have

$$P(r \leq 4) + P(r \geq 5) = 1$$

$$\text{or } P(r \geq 5) = 1 - P(r \leq 4) = 1 - 0.7255 = 0.2745$$

The probability is approximately 0.27 that there will be five or more INFJ personality types in the graduating class.



Summary

In this section, we have studied two discrete probability distributions. The Poisson distribution gives us the probability of r successes in an interval of time or space. The Poisson distribution also can be used to approximate the binomial distribution when $n \geq 100$ and $np < 10$. The geometric distribution gives us the probability that our first success will occur on the n th trial. In the next guided exercise, we will see situations in which each of these distributions applies.

PROCEDURE

HOW TO IDENTIFY DISCRETE PROBABILITY DISTRIBUTIONS

<i>Distribution</i>	<i>Conditions and Setting</i>	<i>Formulas</i>
Binomial distribution	<ol style="list-style-type: none"> 1. There are n independent trials, each repeated under identical conditions. 2. Each trial has two outcomes, S = success and F = failure. 3. $P(S) = p$ is the same for each trial, as is $P(F) = q = 1 - p$ 4. The random variable r represents the number of successes out of n trials. $0 \leq r \leq n$ 	<p>The probability of exactly r successes out of n trials is</p> $P(r) = \frac{n!}{r!(n-r)!} p^r q^{n-r}$ $= C_{n,r} p^r q^{n-r}$ <p>For r,</p> $\mu = np \text{ and } \sigma = \sqrt{npq}$ <p>Table 3 of Appendix II has $P(r)$ values for selected n and p.</p>
Geometric distribution	<ol style="list-style-type: none"> 1. There are n independent trials, each repeated under identical conditions. 2. Each trial has two outcomes, S = success and F = failure. 3. $P(S) = p$ is the same for each trial, as is $P(F) = q = 1 - p$ 4. The random variable n represents the number of the trial on which the <i>first success</i> occurs. $n = 1, 2, 3, \dots$ 	<p>The probability that the first success occurs on the nth trial is</p> $P(n) = pq^{n-1}$ <p>For n,</p> $\mu = \frac{1}{p} \text{ and } \sigma = \frac{\sqrt{q}}{p}$
Poisson distribution	<ol style="list-style-type: none"> 1. Consider a random process that occurs over time, volume, area, or any other quantity that can (in theory) be subdivided into smaller and smaller intervals. 2. Identify success in the context of the interval (time, volume, area, ...) you are studying. 3. Based on long-term experience, compute the mean or average number of successes that occur over the interval (time, volume, area, ...) you are studying. λ = mean number of successes over designated interval 4. The random variable r represents the number of successes that occur over the interval on which you perform the random process. $r = 0, 1, 2, 3, \dots$ 	<p>The probability of r successes in the interval is</p> $P(r) = \frac{e^{-\lambda} \lambda^r}{r!}$ <p>For r,</p> $\mu = \lambda \text{ and } \sigma = \sqrt{\lambda}$ <p>Table 4 of Appendix II gives $P(r)$ for selected values of λ and r.</p>

Continued

Distribution	Conditions and Setting	Formulas
Poisson approximation to the binomial distribution	<ol style="list-style-type: none"> There are n independent trials, each repeated under identical conditions. Each trial has two outcomes, S = success and F = failure. $P(S) = p$ is the same for each trial. In addition, $n \geq 100$ and $np < 10$. The random variable r represents the number of successes out of n trials in a binomial distribution. 	$\lambda = np$ the expected value of r . The probability of r successes on n trials is $P(r) \approx \frac{e^{-\lambda} \lambda^r}{r!}$ Table 4 of Appendix II gives values of $P(r)$ for selected λ and r .

GUIDED EXERCISE 9**Select appropriate distribution**

For each problem, first identify the type of probability distribution needed to solve the problem: binomial, geometric, Poisson, or Poisson approximation to the binomial. Then solve the problem.

- (1) Denver, Colorado, is prone to severe hailstorms. Insurance agents claim that a homeowner in Denver can expect to replace his or her roof (due to hail damage) once every 10 years. What is the probability that in 12 years, a homeowner in Denver will need to replace the roof twice because of hail?

- (a) Consider the problem stated in part (I). What is success in this case? We are interested in the probability of two successes over a specified time interval of 12 years. Which distribution should we use?

- (b) In part (I), we are told that the average roof replacement is once every 10 years. What is the average number of times the roof needs to be replaced in 12 years? What is λ for the 12-year period?

- (c) To finish part (I), use the Poisson distribution to find the probability of two successes in the 12-year period.



Here we can say success is needing to replace a roof because of hail. Because we are interested in the probability of two successes over a time interval, we should use the Poisson distribution.



We are given a value of $\lambda = 1$ for 10 years. To compute λ for 12 years, we convert the denominator to 12 years.

$$\lambda = \frac{0.1}{1 \text{ year}} \cdot \frac{12}{12} = \frac{1.2}{12 \text{ year}}$$

For 12 years, we have $\lambda = 1.2$.



We may use Table 4 of Appendix II because $\lambda = 1.2$ and $r = 2$ are values in the table. The table gives $P(r = 2) = 0.2169$. Using the formula or a calculator gives the same result.

$$P(r) = \frac{e^{-\lambda} \lambda^r}{r!}$$

$$P(r = 2) = \frac{e^{-1.2} (1.2)^2}{2!} \approx 0.2169$$

There is about a 21.7% chance the roof will be damaged twice by hail during a 12-year period.

- (II) A telephone network substation will keep trying to connect a long-distance call to a trunk line until the fourth attempt has been made. After the fourth unsuccessful attempt, the call number goes into a buffer memory bank, and the caller gets a recorded message to be patient. During

Continued

GUIDED EXERCISE 9 *continued*

peak calling periods, the probability of a call connecting into a trunk line is 65% on each try. What percentage of all calls made during peak times will wind up in the buffer memory bank?

- (d) Consider part (II). What is success? We are interested in the probability that a call will wind up in the buffer memory. This will occur if success does not occur before which trial? Since we are looking at the probability of a first success by a specified trial number, which probability distribution do we use?

- (e) What is the probability of success on a single trial? Use this information and the formula for the geometric distribution to compute the probability that the first success occurs on trial 1, 2, 3, or 4. Then use this information to compute $P(n \geq 5)$, where n is the trial number of the first success. What percentage of the calls go to the buffer?



Success is connecting a long-distance call to a trunk line. The call will go into the buffer memory if success is not achieved during the first four attempts. In symbols, the call will go into the buffer if the trial number n of the first success is such that $n \geq 5$. We use the geometric distribution.



Success means the call connects to the trunk line. According to the description in the problem,

$$P(S) = 0.65 = p$$

By the formula for the geometric probability distribution, where n represents the trial number of the first success,

$$P(n) = p(1 - p)^{n-1}$$

Therefore,

$$P(1) = (0.65)(0.35)^0 = 0.65$$

$$P(2) = (0.65)(0.35)^1 = 0.2275$$

$$P(3) = (0.65)(0.35)^2 \approx 0.0796$$

$$P(4) = (0.65)(0.35)^3 \approx 0.0279$$

$$P(n \geq 5) = 1 - P(1) - P(2) - P(3) - P(4)$$

$$\approx 1 - 0.65 - 0.2275 - 0.0796$$

$$- 0.0279$$

$$= 0.015$$

About 1.5% of the calls go to the buffer.

- (III) The murder rate is 3.6 murders per 100,000 inhabitants (Reference: U.S. Department of Justice, Federal Bureau of Investigation). In a community of 1254 people, what is the probability that at least one person will be murdered?

- (f) Consider part (III). What is success in this case? What is the value of n ? Find p , the probability of success on a single trial, to six places after the decimal.



Success is a person being murdered.

$$n = 1254$$

$$p = \frac{3.6}{100,000} \\ = 0.000036$$

- (g) Compute np to three decimal places. Is it appropriate to use the Poisson approximation to the binomial? What is the value of λ to three decimal places?



$$np \approx 0.045$$

Yes.

$$\lambda = np \approx 0.045$$

- (h) Estimate $P(r = 0)$ to three decimal places.



$$P(r) \approx \frac{e^{-\lambda} \lambda^r}{r!} \approx \frac{e^{-0.045}(0.045)^0}{0!} \approx 0.956$$

Recall that $0! = 1$.

- (i) Use the relation $P(r \geq 1) = 1 - P(0)$ to estimate the probability that at least one person will be murdered.



$$P(r \geq 1) \approx 1 - 0.956 \\ \approx 0.044$$

VIEWPOINT**When Do Cracks Become Breakthroughs?**

No one wants to learn by mistakes! However, learning by our successes will not take us beyond the state of the art! Each new idea, technology, social plan, or engineering structure can be considered a new trial. In the meantime, we the laypeople, whose spokesperson is often a poet or writer, will be threatened by both **failures** and **successes**. This is the nature not only of science, technology, and engineering but also of all human endeavors. [For more discussion on this topic, see Problem 18, as well as To Engineer Is Human: The Role of Failure in Successful Design by Professor Petroski (Duke University Press).]

**SECTION 5.4
PROBLEMS**

1. **Statistical Literacy** For a binomial experiment, what probability distribution is used to find the probability that the *first* success will occur on a specified trial?
2. **Statistical Literacy** When using the Poisson distribution, which parameter of the distribution is used in probability computations? What is the symbol used for this parameter?
3. **Critical Thinking** Suppose we have a binomial experiment with 50 trials, and the probability of success on a single trial is 0.02. Is it appropriate to use the Poisson distribution to approximate the probability of two successes? Explain.
4. **Critical Thinking** Suppose we have a binomial experiment, and the probability of success on a single trial is 0.02. If there are 150 trials, is it appropriate to use the Poisson distribution to approximate the probability of three successes? Explain.
5. **Basic Computation: Geometric Distribution** Given a binomial experiment with probability of success on a single trial $p = 0.40$, find the probability that the first success occurs on trial number $n = 3$.
6. **Basic Computation: Geometric Distribution** Given a binomial experiment with probability of success on a single trial $p = 0.30$, find the probability that the first success occurs on trial number $n = 2$.
7. **Basic Computation: Poisson Distribution** Given a binomial experiment with $n = 200$ trials and probability of success on a single trial $p = 0.04$, find the value of λ and then use the Poisson distribution to estimate the probability of $r = 8$ successes.
8. **Basic Computation: Poisson Distribution** Given a binomial experiment with $n = 150$ trials and probability of success on a single trial $p = 0.06$, find the value of λ and then use the Poisson distribution to estimate the probability of $r \leq 2$ successes.
9. **College: Core Requirement** Susan is taking western civilization this semester on a pass/fail basis. The department teaching the course has a history of passing 77% of the students in western civilization each term. Let $n = 1, 2, 3, \dots$ represent the number of times a student takes western civilization until the *first* passing grade is received. (Assume the trials are independent.)
 - (a) Write out a formula for the probability distribution of the random variable n .
 - (b) What is the probability that Susan passes on the first try ($n = 1$)?
 - (c) What is the probability that Susan first passes on the second try ($n = 2$)?
 - (d) What is the probability that Susan needs three or more tries to pass western civilization?
 - (e) What is the expected number of attempts at western civilization Susan must make to have her (first) pass? Hint: Use μ for the geometric distribution and round.

10. **Law: Bar Exam** Bob is a recent law school graduate who intends to take the state bar exam. According to the National Conference on Bar Examiners, about 57% of all people who take the state bar exam pass (Source: *The Book of Odds* by Shook and Shook, Signet). Let $n = 1, 2, 3, \dots$ represent the number of times a person takes the bar exam until the *first* pass.
- Write out a formula for the probability distribution of the random variable n .
 - What is the probability that Bob first passes the bar exam on the second try ($n = 2$)?
 - What is the probability that Bob needs three attempts to pass the bar exam?
 - What is the probability that Bob needs more than three attempts to pass the bar exam?
 - What is the expected number of attempts at the state bar exam Bob must make for his (first) pass? Hint: Use μ for the geometric distribution and round.
11. **Sociology: Hawaiians** On the leeward side of the island of Oahu, in the small village of Nanakuli, about 80% of the residents are of Hawaiian ancestry (Source: *The Honolulu Advertiser*). Let $n = 1, 2, 3, \dots$ represent the number of people you must meet until you encounter the *first* person of Hawaiian ancestry in the village of Nanakuli.
- Write out a formula for the probability distribution of the random variable n .
 - Compute the probabilities that $n = 1$, $n = 2$, and $n = 3$.
 - Compute the probability that $n \geq 4$.
 - In Waikiki, it is estimated that about 4% of the residents are of Hawaiian ancestry. Repeat parts (a), (b), and (c) for Waikiki.
12. **Agriculture: Apples** Approximately 3.6% of all (untreated) Jonathan apples had bitter pit in a study conducted by the botanists Ratkowsky and Martin (Source: *Australian Journal of Agricultural Research*, Vol. 25, pp. 783–790). (Bitter pit is a disease of apples resulting in a soggy core, which can be caused either by overwatering the apple tree or by a calcium deficiency in the soil.) Let n be a random variable that represents the first Jonathan apple chosen at random that has bitter pit.
- Write out a formula for the probability distribution of the random variable n .
 - Find the probabilities that $n = 3$, $n = 5$, and $n = 12$.
 - Find the probability that $n \geq 5$.
 - What is the expected number of apples that must be examined to find the first one with bitter pit? Hint: Use μ for the geometric distribution and round.
13. **Fishing: Lake Trout** At Fontaine Lake Camp on Lake Athabasca in northern Canada, history shows that about 30% of the guests catch lake trout over 20 pounds on a 4-day fishing trip (Source: Athabasca Fishing Lodges, Saskatoon, Canada). Let n be a random variable that represents the *first* trip to Fontaine Lake Camp on which a guest catches a lake trout over 20 pounds.
- Write out a formula for the probability distribution of the random variable n .
 - Find the probability that a guest catches a lake trout weighing at least 20 pounds for the *first* time on trip number 3.
 - Find the probability that it takes more than three trips for a guest to catch a lake trout weighing at least 20 pounds.
 - What is the expected number of fishing trips that must be taken to catch the first lake trout over 20 pounds? Hint: Use μ for the geometric distribution and round.
14. **Archaeology: Artifacts** At Burnt Mesa Pueblo, in one of the archaeological excavation sites, the artifact density (number of prehistoric artifacts per 10 liters of sediment) was 1.5 (Source: *Bandelier Archaeological Excavation Project: Summer 1990 Excavations at Burnt Mesa Pueblo and Casa del Rito*, edited by Kohler, Washington State University Department of Anthropology). Suppose you are going to dig up and examine 50 liters of sediment at this site. Let $r = 0, 1, 2, 3, \dots$ be a random variable that represents the number of prehistoric artifacts found in your 50 liters of sediment.

(a) Explain why the Poisson distribution would be a good choice for the probability distribution of r . What is λ ? Write out the formula for the probability distribution of the random variable r .

(b) Compute the probabilities that in your 50 liters of sediment you will find two prehistoric artifacts, three prehistoric artifacts, and four prehistoric artifacts.

(c) Find the probability that you will find three or more prehistoric artifacts in the 50 liters of sediment.

(d) Find the probability that you will find fewer than three prehistoric artifacts in the 50 liters of sediment.

15. ***Ecology: River Otters*** In his doctoral thesis, L. A. Beckel (University of Minnesota, 1982) studied the social behavior of river otters during the mating season. An important role in the bonding process of river otters is very short periods of social grooming. After extensive observations, Dr. Beckel found that one group of river otters under study had a frequency of initiating grooming of approximately 1.7 for every 10 minutes. Suppose that you are observing river otters for 30 minutes. Let $r = 0, 1, 2, \dots$ be a random variable that represents the number of times (in a 30-minute interval) one otter initiates social grooming of another.

(a) Explain why the Poisson distribution would be a good choice for the probability distribution of r . What is λ ? Write out the formula for the probability distribution of the random variable r .

(b) Find the probabilities that in your 30 minutes of observation, one otter will initiate social grooming four times, five times, and six times.

(c) Find the probability that one otter will initiate social grooming four or more times during the 30-minute observation period.

(d) Find the probability that one otter will initiate social grooming fewer than four times during the 30-minute observation period.

16. ***Law Enforcement: Shoplifting*** *The Denver Post* reported that, on average, a large shopping center has had an incident of shoplifting caught by security once every three hours. The shopping center is open from 10 A.M. to 9 P.M. (11 hours). Let r be the number of shoplifting incidents caught by security in the 11-hour period during which the center is open.

(a) Explain why the Poisson probability distribution would be a good choice for the random variable r . What is λ ?

(b) What is the probability that from 10 A.M. to 9 P.M. there will be at least one shoplifting incident caught by security?

(c) What is the probability that from 10 A.M. to 9 P.M. there will be at least three shoplifting incidents caught by security?

(d) What is the probability that from 10 A.M. to 9 P.M. there will be no shoplifting incidents caught by security?

17. ***Vital Statistics: Birthrate*** *USA Today* reported that the U.S. (annual) birthrate is about 16 per 1000 people, and the death rate is about 8 per 1000 people.

(a) Explain why the Poisson probability distribution would be a good choice for the random variable r = number of births (or deaths) for a community of a given population size.

(b) In a community of 1000 people, what is the (annual) probability of 10 births? What is the probability of 10 deaths? What is the probability of 16 births? 16 deaths?

(c) Repeat part (b) for a community of 1500 people. You will need to use a calculator to compute $P(10 \text{ births})$ and $P(16 \text{ births})$.

(d) Repeat part (b) for a community of 750 people.

18. ***Engineering: Cracks*** Henry Petroski is a professor of civil engineering at Duke University. In his book *To Engineer Is Human: The Role of Failure in Successful Design*, Professor Petroski says that up to 95% of all structural failures, including those of bridges, airplanes, and other commonplace products





Yuan Zhang/Stockphoto.com

of technology, are believed to be the result of crack growth. In most cases, the cracks grow slowly. It is only when the cracks reach intolerable proportions and still go undetected that catastrophe can occur. In a cement retaining wall, occasional hairline cracks are normal and nothing to worry about. If these cracks are spread out and not too close together, the wall is considered safe. However, if a number of cracks group together in a small region, there may be real trouble. Suppose a given cement retaining wall is considered safe if hairline cracks are evenly spread out and occur on the average of 4.2 cracks per 30-foot section of wall.

- (a) Explain why a Poisson probability distribution would be a good choice for the random variable r = number of hairline cracks for a given length of retaining wall.
- (b) In a 50-foot section of safe wall, what is the probability of three (evenly spread-out) hairline cracks? What is the probability of three or more (evenly spread-out) hairline cracks?
- (c) Answer part (b) for a 20-foot section of wall.
- (d) Answer part (b) for a 2-foot section of wall. Round λ to the nearest tenth.
- (e) Consider your answers to parts (b), (c), and (d). If you had three hairline cracks evenly spread out over a 50-foot section of wall, should this be cause for concern? The probability is low. Could this mean that you are lucky to have so few cracks? On a 20-foot section of wall [part (c)], the probability of three cracks is higher. Does this mean that this distribution of cracks is closer to what we should expect? For part (d), the probability is very small. Could this mean you are not so lucky and have something to worry about? Explain your answers.
19. **Meteorology: Winter Conditions** Much of Trail Ridge Road in Rocky Mountain National Park is over 12,000 feet high. Although it is a beautiful drive in summer months, in winter the road is closed because of severe weather conditions. *Winter Wind Studies in Rocky Mountain National Park* by Glidden (published by Rocky Mountain Nature Association) states that sustained gale-force winds (over 32 miles per hour and often over 90 miles per hour) occur on the average of once every 60 hours at a Trail Ridge Road weather station.
- (a) Let r = frequency with which gale-force winds occur in a given time interval. Explain why the Poisson probability distribution would be a good choice for the random variable r .
- (b) For an interval of 108 hours, what are the probabilities that $r = 2, 3$, and 4 ? What is the probability that $r < 2$?
- (c) For an interval of 180 hours, what are the probabilities that $r = 3, 4$, and 5 ? What is the probability that $r < 3$?
20. **Earthquakes: San Andreas Fault** USA Today reported that Parkfield, California, is dubbed the world's earthquake capital because it sits on top of the notorious San Andreas fault. Since 1857, Parkfield has had a major earthquake on the average of once every 22 years.
- (a) Explain why a Poisson probability distribution would be a good choice for r = number of earthquakes in a given time interval.
- (b) Compute the probability of at least one major earthquake in the next 22 years. Round λ to the nearest hundredth, and use a calculator.
- (c) Compute the probability that there will be no major earthquake in the next 22 years. Round λ to the nearest hundredth, and use a calculator.
- (d) Compute the probability of at least one major earthquake in the next 50 years. Round λ to the nearest hundredth, and use a calculator.
- (e) Compute the probability of no major earthquakes in the next 50 years. Round λ to the nearest hundredth, and use a calculator.

21. **Real Estate: Sales** Jim is a real estate agent who sells large commercial buildings. Because his commission is so large on a single sale, he does not need to sell many buildings to make a good living. History shows that Jim has a record of selling an average of eight large commercial buildings every 275 days.
- Explain why a Poisson probability distribution would be a good choice for r = number of buildings sold in a given time interval.
 - In a 60-day period, what is the probability that Jim will make no sales? one sale? two or more sales?
 - In a 90-day period, what is the probability that Jim will make no sales? two sales? three or more sales?
22. **Law Enforcement: Burglaries** *The Honolulu Advertiser* stated that in Honolulu there was an average of 661 burglaries per 100,000 households in a given year. In the Kohola Drive neighborhood there are 316 homes. Let r = number of these homes that will be burglarized in a year.
- Explain why the Poisson approximation to the binomial would be a good choice for the random variable r . What is n ? What is p ? What is λ to the nearest tenth?
 - What is the probability that there will be no burglaries this year in the Kohola Drive neighborhood?
 - What is the probability that there will be no more than one burglary in the Kohola Drive neighborhood?
 - What is the probability that there will be two or more burglaries in the Kohola Drive neighborhood?
23. **Criminal Justice: Drunk Drivers** *Harper's Index* reported that the number of (Orange County, California) convicted drunk drivers whose sentence included a tour of the morgue was 569, of which only 1 became a repeat offender.
- Suppose that of 1000 newly convicted drunk drivers, all were required to take a tour of the morgue. Let us assume that the probability of a repeat offender is still $p = 1/569$. Explain why the Poisson approximation to the binomial would be a good choice for r = number of repeat offenders out of 1000 convicted drunk drivers who toured the morgue. What is λ to the nearest tenth?
 - What is the probability that $r = 0$?
 - What is the probability that $r > 1$?
 - What is the probability that $r > 2$?
 - What is the probability that $r > 3$?
24. **Airlines: Lost Bags** *USA Today* reported that for all airlines, the number of lost bags was
- | | |
|-------------------------------|-------------------------------------|
| May: 6.02 per 1000 passengers | December: 12.78 per 1000 passengers |
|-------------------------------|-------------------------------------|
- Note:* A passenger could lose more than one bag.
- Let r = number of bags lost per 1000 passengers in May. Explain why the Poisson distribution would be a good choice for the random variable r . What is λ to the nearest tenth?
 - In the month of May, what is the probability that out of 1000 passengers, no bags are lost? that 3 or more bags are lost? that 6 or more bags are lost?
 - In the month of December, what is the probability that out of 1000 passengers, no bags are lost? that 6 or more bags are lost? that 12 or more bags are lost? (Round λ to the nearest whole number.)
25. **Law Enforcement: Officers Killed** *Chances: Risk and Odds in Everyday Life*, by James Burke, reports that the probability a police officer will be killed in the line of duty is 0.5% (or less).
- In a police precinct with 175 officers, let r = number of police officers killed in the line of duty. Explain why the Poisson approximation to the binomial would be a good choice for the random variable r . What is n ? What is p ? What is λ to the nearest tenth?

- (b) What is the probability that no officer in this precinct will be killed in the line of duty?
 (c) What is the probability that one or more officers in this precinct will be killed in the line of duty?
 (d) What is the probability that two or more officers in this precinct will be killed in the line of duty?
26. ***Business Franchise: Shopping Center Chances: Risk and Odds in Everyday Life***, by James Burke, reports that only 2% of all local franchises are business failures. A Colorado Springs shopping complex has 137 franchises (restaurants, print shops, convenience stores, hair salons, etc.).
 (a) Let r be the number of these franchises that are business failures. Explain why a Poisson approximation to the binomial would be appropriate for the random variable r . What is n ? What is p ? What is λ (rounded to the nearest tenth)?
 (b) What is the probability that none of the franchises will be a business failure?
 (c) What is the probability that two or more franchises will be business failures?
 (d) What is the probability that four or more franchises will be business failures?
27. ***Poisson Approximation to the Binomial: Comparisons***
 (a) For $n = 100$, $p = 0.02$, and $r = 2$, compute $P(r)$ using the formula for the binomial distribution and your calculator:
- $$P(r) = C_{n,r} p^r (1 - p)^{n-r}$$
- (b) For $n = 100$, $p = 0.02$, and $r = 2$, estimate $P(r)$ using the Poisson approximation to the binomial.
 (c) Compare the results of parts (a) and (b). Does it appear that the Poisson distribution with $\lambda = np$ provides a good approximation for $P(r = 2)$?
 (d) Repeat parts (a) to (c) for $r = 3$.
28.  ***Expand Your Knowledge: Conditional Probability*** Pyramid Lake is located in Nevada on the Paiute Indian Reservation. This lake is famous for large cutthroat trout. The mean number of trout (large and small) caught from a boat is 0.667 fish per hour (Reference: *Creel Chronicle*, Vol. 3, No. 2). Suppose you rent a boat and go fishing for 8 hours. Let r be a random variable that represents the number of fish you catch in the 8-hour period.
 (a) Explain why a Poisson probability distribution is appropriate for r . What is λ for the 8-hour fishing trip? Round λ to the nearest tenth so that you can use Table 4 of Appendix II for Poisson probabilities.
 (b) If you have already caught three trout, what is the probability you will catch a total of seven or more trout? Compute $P(r \geq 7 | r \geq 3)$. See Hint below.
 (c) If you have already caught four trout, what is the probability you will catch a total of fewer than nine trout? Compute $P(r < 9 | r \geq 4)$. See Hint below.
 (d) List at least three other areas besides fishing to which you think conditional Poisson probabilities can be applied.

Hint for solution: Review item 6, conditional probability, in the summary of basic probability rules at the end of Section 4.2. Note that

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

and show that in part (b),

$$P(r \geq 7 | r \geq 3) = \frac{P((r \geq 7) \text{ and } (r \geq 3))}{P(r \geq 3)} = \frac{P(r \geq 7)}{P(r \geq 3)}$$



29. ***Conditional Probability: Hail Damage*** In western Kansas, the summer density of hailstorms is estimated at about 2.1 storms per 5 square miles. In most cases, a hailstorm damages only a relatively small area in a square mile (Reference: *Agricultural Statistics*, U.S. Department of Agriculture). A crop insurance



Negative binomial distribution



30.

Expand Your Knowledge: Negative Binomial Distribution Suppose you have binomial trials for which the probability of success on each trial is p and the probability of failure is $q = 1 - p$. Let k be a fixed whole number greater than or equal to 1. Let n be the number of the trial on which the k th success occurs. This means that the first $k - 1$ successes occur within the first $n - 1$ trials, while the k th success actually occurs on the n th trial. Now, if we are going to have k successes, we must have at least k trials. So, $n = k, k + 1, k + 2, \dots$ and n is a random variable. In the literature of mathematical statistics, the probability distribution for n is called the *negative binomial distribution*. The formula for the probability distribution of n is shown in the next display (see Problem 33 for a derivation).

Negative binomial distribution

Let $k \geq 1$ be a fixed whole number. The probability that the k th success occurs on trial number n is

$$P(n) = C_{n-1, k-1} p^k q^{n-k}$$

where $C_{n-1, k-1} = \frac{(n-1)!}{(k-1)!(n-k)!}$
 $n = k, k+1, k+2, \dots$

The expected value and standard deviation of this probability distribution are

$$\mu = \frac{k}{p} \text{ and } \sigma = \sqrt{\frac{kp}{p^2}} = \sqrt{\frac{k}{p}}$$

Note: If $k = 1$, the negative binomial distribution is called the *geometric distribution*.

In eastern Colorado, there are many dry land wheat farms. The success of a spring wheat crop is dependent on sufficient moisture in March and April. Assume that the probability of a successful wheat crop in this region is about 65%. So, the probability of success in a single year is $p = 0.65$, and the probability of failure is $q = 0.35$. The Wagner farm has taken out a loan and needs $k = 4$ successful crops to repay it. Let n be a random variable representing the year in which the fourth successful crop occurs (after the loan was made).

- Write out the formula for $P(n)$ in the context of this application.
- Compute $P(n = 4), P(n = 5), P(n = 6)$, and $P(n = 7)$.
- What is the probability that the Wagners can repay the loan within 4 to 7 years? *Hint:* Compute $P(4 \leq n \leq 7)$.
- What is the probability that the Wagners will need to farm for 8 or more years before they can repay the loan? *Hint:* Compute $P(n \geq 8)$.
- What are the expected value μ and standard deviation σ of the random variable n ? Interpret these values in the context of this application.



31.

Negative Binomial Distribution: Marketing Susan is a sales representative who has a history of making a successful sale from about 80% of her sales contacts. If

she makes 12 successful sales this week, Susan will get a bonus. Let n be a random variable representing the number of contacts needed for Susan to get the 12th sale.

- (a) Explain why a negative binomial distribution is appropriate for the random variable n . Write out the formula for $P(n)$ in the context of this application.

Hint: See Problem 30.

- (b) Compute $P(n = 12)$, $P(n = 13)$, and $P(n = 14)$.

- (c) What is the probability that Susan will need from 12 to 14 contacts to get the bonus?

- (d) What is the probability that Susan will need more than 14 contacts to get the bonus?

- (e) What are the expected value μ and standard deviation σ of the random variable n ? Interpret these values in the context of this application.



32.

Negative Binomial Distribution: Type A Blood Donors Blood type A occurs in about 41% of the population (Reference: *Laboratory and Diagnostic Tests* by F. Fischbach). A clinic needs 3 pints of type A blood. A donor usually gives a pint of blood. Let n be a random variable representing the number of donors needed to provide 3 pints of type A blood.

- (a) Explain why a negative binomial distribution is appropriate for the random variable n . Write out the formula for $P(n)$ in the context of this application.

Hint: See Problem 30.

- (b) Compute $P(n = 3)$, $P(n = 4)$, $P(n = 5)$, and $P(n = 6)$.

- (c) What is the probability that the clinic will need from three to six donors to obtain the needed 3 pints of type A blood?

- (d) What is the probability that the clinic will need more than six donors to obtain 3 pints of type A blood?

- (e) What are the expected value μ and standard deviation σ of the random variable n ? Interpret these values in the context of this application.



33.

Expand Your Knowledge: Brain Teaser If you enjoy a little abstract thinking, you may want to derive the formula for the negative binomial probability distribution. Use the notation of Problem 30. Consider two events, A and B .

$$A = \{\text{event that the first } n = 1 \text{ trials contain } k - 1 \text{ successes}\}$$

$$B = \{\text{event that the } n \text{th trial is a success}\}$$

- (a) Use the binomial probability distribution to show that the probability of A is $P(A) = C_{n-1, k-1} p^{k-1} q^{(n-1)-(k-1)}$.

- (b) Show that the probability of B is that of a single trial in a binomial experiment, $P(B) = p$.

- (c) Why is $P(A \text{ and } B) = P(A) \cdot P(B)$? *Hint:* Binomial trials are independent.

- (d) Use parts (a), (b), and (c) to compute and simplify $P(A \text{ and } B)$.

- (e) Compare $P(A \text{ and } B)$ with the negative binomial formula and comment on the meaning of your results.



Chapter Review

SUMMARY

This chapter discusses random variables and important probability distributions associated with discrete random variables.

- The value of a *random variable* is determined by chance. Formulas for the mean, variance, and standard deviation of linear functions and linear combinations of independent random variables are given.

- Random variables are either *discrete* or *continuous*.

- A probability distribution of a discrete random variable x consists of all distinct values of x and the corresponding probabilities $P(x)$. For each x , $0 \leq P(x) \leq 1$ and $\sum P(x) = 1$.

- A discrete probability distribution can be displayed visually by a *probability histogram* in which the values of the random variable x are displayed on the horizontal axis, the height of each bar is $P(x)$, and each bar is 1 unit wide.
 - For discrete probability distributions,
- $$\mu = \sum xP(x) \quad \text{and} \quad \sigma = \sqrt{\sum (x - \mu)^2 P(x)}$$
- The mean μ is called the *expected value* of the probability distribution.
 - A *binomial experiment* consists of a fixed number n of independent trials repeated under identical conditions. There are two outcomes for each trial, called *success* and *failure*. The probability p of success on each trial is the same.
 - The number of successes r in a binomial experiment is the random variable for the binomial probability distribution. Probabilities can be computed using a formula or probability distribution outputs from a computer or calculator. Some probabilities can be found in Table 3 of Appendix II.

- For a binomial distribution,
- $$\mu = np \quad \text{and} \quad \sigma = \sqrt{npq},$$
- where $q = 1 - p$.
- For a binomial experiment, the number of successes is usually within the interval from $\mu - 2.5\sigma$ to $\mu + 2.5\sigma$. A number of successes outside this range of values is unusual but can occur.
 - The *geometric probability distribution* is used to find the probability that the first success of a binomial experiment occurs on trial number n .
 - The *Poisson distribution* is used to compute the probability of r successes in an interval of time, volume, area, and so forth.
 - The *Poisson distribution* can be used to approximate the binomial distribution when $n \geq 100$ and $np < 10$.
 - The *hypergeometric distribution* (discussed in Appendix I) is a probability distribution of a random variable that has two outcomes when sampling is done without replacement.

It is important to check the conditions required for the use of each probability distribution.

IMPORTANT WORDS & SYMBOLS

Section 5.1

- Random variable 182
 Discrete 182
 Continuous 182
 Probability distribution 183
 Mean μ of a probability distribution 185
 Standard deviation σ of a probability distribution 185
 Expected value μ 186
 Linear function of a random variable 188
 Linear combination of two independent random variables 188

Section 5.2

- Binomial experiment 195
 Number of trials, n 195
 Independent trials 195
 Successes and failures in a binomial experiment 195
 Probability of success $P(S) = p$ 195

Probability of failure $P(F) = q = 1 - p$ 195

Number of successes, r 195

Binomial coefficient $C_{n,r}$ 198

Binomial probability distribution

$$P(r) = C_{n,r} p^r q^{n-r} 198$$

Hypergeometric probability distribution (see Appendix I) 204

Section 5.3

Mean for the binomial distribution

$$\mu = np 213$$

Standard deviation for the binomial distribution $\sigma = \sqrt{npq}$ 213

Quota problem 215

Section 5.4

- Geometric probability distribution 222
 Poisson probability distribution 224
 Poisson approximation to the binomial 227
 Negative binomial distribution 238

VIEWPOINT

What's Your Type?

Are students and professors really compatible? One way of answering this question is to look at the Myers–Briggs Type Indicators for personality preferences. What is the probability that your professor is introverted and judgmental? What is the probability that you are

extroverted and perceptive? Are most of the leaders in student government extroverted and judgmental? Is it true that members of Phi Beta Kappa have personality types more like those of the professors? We will consider questions such as these in more detail in Chapter 7 (estimation) and Chapter 8 (hypothesis testing), where we will continue our work with binomial probabilities. In the meantime, you can find many answers regarding careers, probability, and personality types in *Applications of the Myers–Briggs Type Indicator in Higher Education*, edited by J. Provost and S. Anchors.

CHAPTER REVIEW PROBLEMS

1. | **Statistical Literacy** What are the requirements for a probability distribution?
2. | **Statistical Literacy** List the criteria for a binomial experiment. What does the random variable of a binomial experiment measure?
3. | **Critical Thinking** For a binomial probability distribution, it is unusual for the number of successes to be less than $\mu - 2.5\sigma$ or greater than $\mu + 2.5\sigma$.
 - (a) For a binomial experiment with 10 trials for which the probability of success on a single trial is 0.2, is it unusual to have more than five successes? Explain.
 - (b) If you were simply guessing on a multiple-choice exam consisting of 10 questions with 5 possible responses for each question, would you be likely to get more than half of the questions correct? Explain.
4. | **Critical Thinking** Consider a binomial experiment. If the number of trials is increased, what happens to the expected value? to the standard deviation? Explain.
5. | **Probability Distribution: Auto Leases** Consumer Banker Association released a report showing the lengths of automobile leases for new automobiles. The results are as follows.

Lease Length in Months	Percent of Leases
13–24	12.7%
25–36	37.1%
37–48	28.5%
49–60	21.5%
More than 60	0.2%

- (a) Use the midpoint of each class, and call the midpoint of the last class 66.5 months, for purposes of computing the expected lease term. Also find the standard deviation of the distribution.
 (b) Sketch a graph of the probability distribution for the duration of new auto leases.
6. | **Ecology: Predator and Prey** Isle Royale, an island in Lake Superior, has provided an important study site of wolves and their prey. In the National Park Service Scientific Monograph Series 11, *Wolf Ecology and Prey Relationships on Isle Royale*, Peterson gives results of many wolf–moose studies. Of special interest is the study of the number of moose killed by wolves. In the period from 1958 to 1974, there were 296 moose deaths identified as wolf kills. The age distribution of the kills is as follows.

Age of Moose in Years	Number Killed by Wolves
Calf (0.5 yr)	112
1–5	53
6–10	73
11–15	56
16–20	2



georgesarkar.com/Alamy

- (a) For each age group, compute the probability that a moose in that age group is killed by a wolf.
- (b) Consider all ages in a class equal to the class midpoint. Find the expected age of a moose killed by a wolf and the standard deviation of the ages.
7. **Insurance: Auto** State Farm Insurance studies show that in Colorado, 55% of the auto insurance claims submitted for property damage are submitted by males under 25 years of age. Suppose 10 property damage claims involving automobiles are selected at random.
- (a) Let r be the number of claims made by males under age 25. Make a histogram for the r -distribution probabilities.
- (b) What is the probability that six or more claims are made by males under age 25?
- (c) What is the expected number of claims made by males under age 25? What is the standard deviation of the r -probability distribution?
8. **Quality Control: Pens** A stationery store has decided to accept a large shipment of ball-point pens if an inspection of 20 randomly selected pens yields no more than two defective pens.
- (a) Find the probability that this shipment is accepted if 5% of the total shipment is defective.
- (b) Find the probability that this shipment is not accepted if 15% of the total shipment is defective.
9. **Criminal Justice: Inmates** According to *Harper's Index*, 50% of all federal inmates are serving time for drug dealing. A random sample of 16 federal inmates is selected.
- (a) What is the probability that 12 or more are serving time for drug dealing?
- (b) What is the probability that 7 or fewer are serving time for drug dealing?
- (c) What is the expected number of inmates serving time for drug dealing?
10. **Airlines: On-Time Arrivals** *Consumer Reports* rated airlines and found that 80% of the flights involved in the study arrived on time (that is, within 15 minutes of scheduled arrival time). Assuming that the on-time arrival rate is representative of the entire commercial airline industry, consider a random sample of 200 flights. What is the expected number that will arrive on time? What is the standard deviation of this distribution?
11. **Agriculture: Grapefruit** It is estimated that 75% of a grapefruit crop is good; the other 25% have rotten centers that cannot be detected until the grapefruit are cut open. The grapefruit are sold in sacks of 10. Let r be the number of good grapefruit in a sack.
- (a) Make a histogram of the probability distribution of r .
- (b) What is the probability of getting no more than one bad grapefruit in a sack? What is the probability of getting at least one good grapefruit in a sack?
- (c) What is the expected number of good grapefruit in a sack?
- (d) What is the standard deviation of the r -probability distribution?
12. **Restaurants: Reservations** The Orchard Café has found that about 5% of the diners who make reservations don't show up. If 82 reservations have been made, how many diners can be expected to show up? Find the standard deviation of this distribution.
13. **College Life: Student Government** The student government claims that 85% of all students favor an increase in student fees to buy indoor potted plants for the classrooms. A random sample of 12 students produced 2 in favor of the project. What is the probability that 2 or fewer in the sample will favor the project, assuming the student government's claim is correct? **Interpretation** Do the data support the student government's claim, or does it seem that the percentage favoring the increase in fees is less than 85%?
14. **Quota Problem: Financial** Suppose you are a (junk) bond broker who buys only bonds that have a 50% chance of default. You want a portfolio with at least five bonds that do not default. You can dispose of the other bonds in the

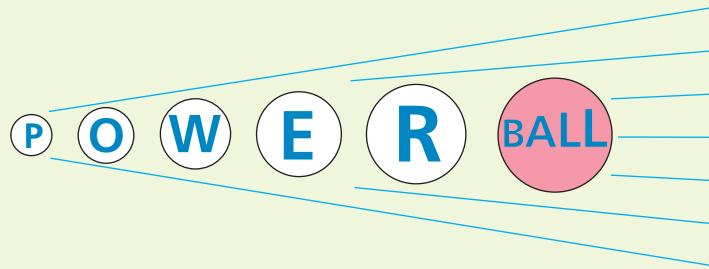
- portfolio with no great loss. How many such bonds should you buy if you want to be 94.1% sure that five or more will not default?
15. **Theater: Coughs** A person with a cough is a *persona non grata* on airplanes, elevators, or at the theater. In theaters especially, the irritation level rises with each muffled explosion. According to Dr. Brian Carlin, a Pittsburgh pulmonologist, in any large audience you'll hear about 11 coughs per minute (Source: *USA Today*).
(a) Let r = number of coughs in a given time interval. Explain why the Poisson distribution would be a good choice for the probability distribution of r .
(b) Find the probability of three or fewer coughs (in a large auditorium) in a 1-minute period.
(c) Find the probability of at least three coughs (in a large auditorium) in a 30-second period.
16. **Accident Rate: Small Planes** Flying over the western states with mountainous terrain in a small aircraft is 40% riskier than flying over similar distances in flatter portions of the nation, according to a General Accounting Office study completed in response to a congressional request. The accident rate for small aircraft in the 11 mountainous western states is 2.4 accidents per 100,000 flight operations (Source: *The Denver Post*).
(a) Let r = number of accidents for a given number of operations. Explain why the Poisson distribution would be a good choice for the probability distribution of r .
(b) Find the probability of no accidents in 100,000 flight operations.
(c) Find the probability of at least 4 accidents in 200,000 flight operations.
17. **Banking: Loan Defaults** Records over the past year show that 1 out of 350 loans made by Mammon Bank have defaulted. Find the probability that 2 or more out of 300 loans will default. Hint: Is it appropriate to use the Poisson approximation to the binomial distribution?
18. **Car Theft: Hawaii** In Hawaii, the rate of motor vehicle theft is 551 thefts per 100,000 vehicles (Reference: U.S. Department of Justice, Federal Bureau of Investigation). A large parking structure in Honolulu has issued 482 parking permits.
(a) What is the probability that none of the vehicles with a permit will eventually be stolen?
(b) What is the probability that at least one of the vehicles with a permit will eventually be stolen?
(c) What is the probability that two or more of the vehicles with a permit will eventually be stolen?
Note: The vehicles may or may not be stolen from the parking structure. Hint: Is it appropriate to use the Poisson approximation to the binomial? Explain.
19. **General: Coin Flip** An experiment consists of tossing a coin a specified number of times and recording the outcomes.
(a) What is the probability that the *first* head will occur on the second trial? Does this probability change if we toss the coin three times? What if we toss the coin four times? What probability distribution model do we use to compute these probabilities?
(b) What is the probability that the *first* head will occur on the fourth trial? after the fourth trial?
20. **Testing: CPA Exam** Cathy is planning to take the Certified Public Accountant Examination (CPA exam). Records kept by the college of business from which she graduated indicate that 83% of the students who graduated pass the CPA exam. Assume that the exam is changed each time it is given. Let $n = 1, 2, 3, \dots$ represent the number of times a person takes the CPA exam until the *first* pass. (Assume the trials are independent.)
(a) What is the probability that Cathy passes the CPA exam on the first try?
(b) What is the probability that Cathy passes the CPA exam on the second or third try?

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

1. Powerball! Imagine, you could win a jackpot worth at least \$10 million. Some jackpots have been worth more than \$250 million! Powerball is a multistate lottery. To play Powerball, you purchase a \$1 ticket. On the ticket you select five distinct white balls (numbered 1 through 59) and then one red Powerball (numbered 1 through 39). The red Powerball number may be any of the numbers 1 through 39, including any such numbers you selected for the white balls. Every Wednesday and Saturday there is a drawing. If your chosen numbers match those drawn, you win! Figure 5-7 shows all the prizes and the probability of winning each prize and specifies how many numbers on your ticket must match those drawn to win the prize. The Multi-State Lottery Association maintains a web site that displays the results of each drawing, as well as a history of the results of previous drawings. For updated Powerball data, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Multi-State Lottery Association.

FIGURE 5-7



Match	Approximate Probability	Prize
5 white balls + Powerball	0.0000000512	Jackpot*
5 white balls	0.000000195	\$200,000
4 white balls + Powerball	0.000000138	\$10,000
4 white balls	0.00000525	\$100
3 white balls + Powerball	0.00000733	\$100
3 white balls	0.00279	\$7
2 white balls + Powerball	0.00127	\$7
1 white ball + Powerball	0.00810	\$4
0 white balls + Powerball	0.0162	\$3
Overall chance of winning	0.0285	

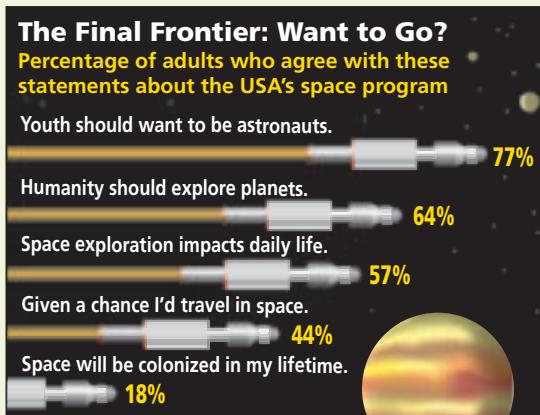
*The Jackpot will be divided equally (if necessary) among multiple winners and is paid in 30 annual installments or in a reduced lump sum.

- (a) Assume the jackpot is \$10 million and there will be only one jackpot winner. Figure 5-7 lists the prizes and the probability of winning each prize. What is the probability of *not winning* any prize? Consider all the prizes and their respective probabilities, and the prize of \$0 (no win) and its probability. Use all these values to estimate your expected winnings μ if you play one ticket. How much do you effectively contribute to the state in which you purchased the ticket (ignoring the overhead cost of operating Powerball)?
- (b) Suppose the jackpot increased to \$25 million (and there was to be only one winner). Compute your expected winnings if you buy one ticket. Does the probability of winning the jackpot change because the jackpot is higher?
- (c) Pretend that you are going to buy 10 Powerball tickets when the jackpot is \$10 million. Use the random-number table to select your numbers. Check

the Multi-State Lottery Association web site (or any other Powerball site) for the most recent drawing results to see if you would have won a prize.

- (d) The probability of winning *any* prize is about 0.0285. Suppose you decide to buy five tickets. Use the binomial distribution to compute the probability of winning (any prize) at least once. *Note:* You will need to use the binomial formula. Carry at least three digits after the decimal.
 - (e) The probability of winning *any* prize is about 0.0285. Suppose you play Powerball 100 times. Explain why it is appropriate to use the Poisson approximation to the binomial to compute the probability of winning at least one prize. Compute $\lambda = np$. Use the Poisson table to estimate the probability of winning at least one prize.
2. Would you like to travel in space, if given a chance? According to Opinion Research for Space Day Partners, if your answer is yes, you are not alone. Forty-four percent of adults surveyed agreed that they would travel in space if given a chance. Look at Figure 5-8, and use the information presented to answer the following questions.

FIGURE 5-8



Source: Opinion Research for Space Day Partners

- (a) According to Figure 5-8, the probability that an adult selected at random agrees with the statement that humanity should explore planets is 64%. Round this probability to 65%, and use this estimate with the binomial distribution table to determine the probability that of 10 adults selected at random, at least half would agree that humanity should explore planets.
- (b) Does space exploration have an impact on daily life? Find the probability that of 10 adults selected at random, at least 9 would agree that space exploration does have an impact on daily life. *Hint:* Use the formula for the binomial distribution.
- (c) In a room of 35 adults, what is the expected number who would travel in space, given a chance? What is the standard deviation?
- (d) What is the probability that the first adult (selected at random) you asked would agree with the statement that space will be colonized in the person's lifetime? *Hint:* Use the geometric distribution.

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. Discuss what we mean by a binomial experiment. As you can see, a binomial process or binomial experiment involves a lot of assumptions! For example, all

the trials are supposed to be independent and repeated under identical conditions. Is this always true? Can we always be completely certain that the probability of success does not change from one trial to the next? In the real world, there is almost nothing we can be absolutely sure about, so the *theoretical* assumptions of the binomial probability distribution often will not be completely satisfied. Does that mean we cannot use the binomial distribution to solve practical problems? Looking at this chapter, the answer seems to be that we can indeed use the binomial distribution even if not all the assumptions are *exactly* met. We find in practice that the conclusions are sufficiently accurate for our intended application. List three applications of the binomial distribution for which you think, although some of the assumptions are not exactly met, there is adequate reason to apply the binomial distribution anyhow.

2. Why do we need to learn the formula for the binomial probability distribution? Using the formula repeatedly can be very tedious. To cut down on tedious calculations, most people will use a binomial table such as the one found in Appendix II of this book.
 - (a) However, there are many applications for which a table in the back of *any* book is not adequate. For instance, compute

$$P(r = 3) \quad \text{where } n = 5 \text{ and } p = 0.735$$

Can you find the result in the table? Do the calculation by using the formula. List some other situations in which a table might not be adequate to solve a particular binomial distribution problem.

- (b) The formula itself also has limitations. For instance, consider the difficulty of computing

$$P(r \geq 285) \quad \text{where } n = 500 \text{ and } p = 0.6$$

What are some of the difficulties you run into? Consider the calculation of $P(r = 285)$. You will be raising 0.6 and 0.4 to very high powers; this will give you very, very small numbers. Then you need to compute $C_{500,285}$, which is a very, very large number. When combining extremely large and extremely small numbers in the same calculation, most accuracy is lost unless you carry a huge number of significant digits. If this isn't tedious enough, consider the steps you need to compute

$$P(r \geq 285) = P(r = 285) + P(r = 286) + \cdots + P(r = 500)$$

Does it seem clear that we need a better way to estimate $P(r \geq 285)$? In Chapter 6, you will learn a much better way to estimate binomial probabilities when the number of trials is large.

3. In Chapter 3, we learned about means and standard deviations. In Section 5.1, we learned that probability distributions also can have a mean and standard deviation. Discuss what is meant by the expected value and standard deviation of a binomial distribution. How does this relate back to the material we learned in Chapter 3 and Section 5.1?
4. In Chapter 2, we looked at the shapes of distributions. Review the concepts of skewness and symmetry; then categorize the following distributions as to skewness or symmetry:
 - (a) A binomial distribution with $n = 11$ trials and $p = 0.50$
 - (b) A binomial distribution with $n = 11$ trials and $p = 0.10$
 - (c) A binomial distribution with $n = 11$ trials and $p = 0.90$

In general, does it seem true that binomial probability distributions in which the probability of success is close to 0 are skewed right, whereas those with probability of success close to 1 are skewed left?

USING TECHNOLOGY

Binomial Distributions

Although tables of binomial probabilities can be found in most libraries, such tables are often inadequate. Either the value of p (the probability of success on a trial) you are looking for is not in the table, or the value of n (the number of trials) you are looking for is too large for the table. In Chapter 6, we will study the normal approximation to the binomial. This approximation is a great help in many practical applications. Even so, we sometimes use the formula for the binomial probability distribution on a computer or graphing calculator to compute the probability we want.

Applications

The following percentages were obtained over many years of observation by the U.S. Weather Bureau. All data listed are for the month of December.

Location	Long-Term Mean % of Clear Days in Dec.
Juneau, Alaska	18%
Seattle, Washington	24%
Hilo, Hawaii	36%
Honolulu, Hawaii	60%
Las Vegas, Nevada	75%
Phoenix, Arizona	77%

Adapted from *Local Climatological Data*, U.S. Weather Bureau publication, "Normals, Means, and Extremes" Table.

In the locations listed, the month of December is a relatively stable month with respect to weather. Since weather patterns from one day to the next are more or less the same, it is reasonable to use a binomial probability model.

- Let r be the number of clear days in December. Since December has 31 days, $0 \leq r \leq 31$. Using appropriate computer software or calculators available to you, find the probability $P(r)$ for each of the listed locations when $r = 0, 1, 2, \dots, 31$.
- For each location, what is the expected value of the probability distribution? What is the standard deviation?

You may find that using cumulative probabilities and appropriate subtraction of probabilities, rather than addition of probabilities, will make finding the solutions to Applications 3 to 7 easier.

- Estimate the probability that Juneau will have at most 7 clear days in December.
- Estimate the probability that Seattle will have from 5 to 10 (including 5 and 10) clear days in December.
- Estimate the probability that Hilo will have at least 12 clear days in December.
- Estimate the probability that Phoenix will have 20 or more clear days in December.
- Estimate the probability that Las Vegas will have from 20 to 25 (including 20 and 25) clear days in December.

Technology Hints

TI-84Plus/TI-83Plus/TI-nspire (with TI-84 Plus keypad), Excel 2007, Minitab

The Tech Note in Section 5.2 gives specific instructions for binomial distribution functions on the TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) calculators, Excel 2007, and Minitab.

SPSS

In SPSS, the function **PDF.BINOM(q,n,p)** gives the probability of q successes out of n trials, where p is the probability of success on a single trial. In the data editor, name a variable r and enter values 0 through n . Name another variable Prob_r . Then use the menu choices **Transform > Compute**. In the dialogue box, use Prob_r for the target variable. In the function group, select **PDF and Noncentral PDF**. In the function box, select **PDF.BINOM(q,n,p)**. Use the variable r for q and appropriate values for n and p . Note that the function **CDF.BINOM(q,n,p)**, from the **CDF and Noncentral CDF** group, gives the cumulative probability of 0 through q successes.



6



- 6.1 Graphs of Normal Probability Distributions
- 6.2 Standard Units and Areas Under the Standard Normal Distribution
- 6.3 Areas Under Any Normal Curve
- 6.4 Sampling Distributions
- 6.5 The Central Limit Theorem
- 6.6 Normal Approximation to Binomial Distribution and to \hat{p} Distribution



Historical Pictures/Stock Montage

*One cannot escape
the feeling that these
mathematical
formulas have an
independent existence*

*and an intelligence of their own, that they
are wiser than we are, wiser even than
their discoverers, that we get more out of
them than was originally put into them.*

—HEINRICH HERTZ

*How can it be that mathematics, a
product of human thought independent of
experience, is so admirably adapted to the
objects of reality?*

—ALBERT EINSTEIN

For online student resources, visit The Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>.

NORMAL CURVES AND SAMPLING DISTRIBUTIONS

PREVIEW QUESTIONS

What are some characteristics of a normal distribution? What does the empirical rule tell you about data spread around the mean? How can this information be used in quality control? (SECTION 6.1)

Can you compare apples and oranges, or maybe elephants and butterflies? In most cases, the answer is no—unless you first standardize your measurements. What are a standard normal distribution and a standard z score? (SECTION 6.2)

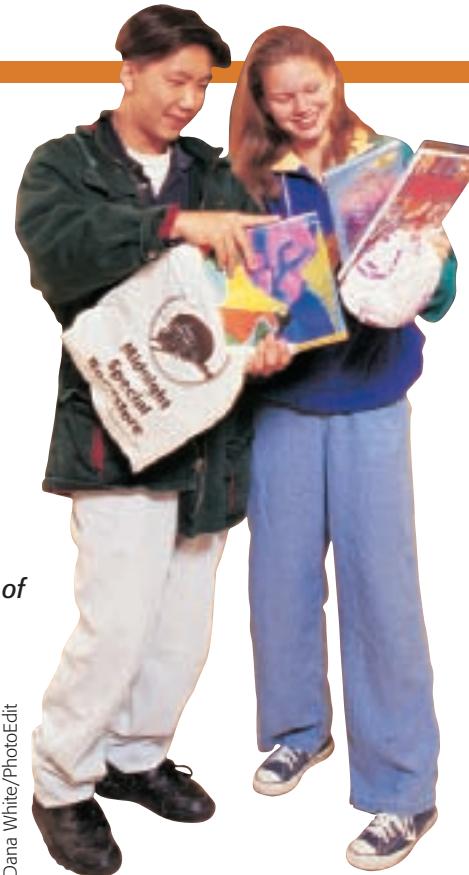
How do you convert any normal distribution to a standard normal distribution? How do you find probabilities of “standardized events”? (SECTION 6.3)

As humans, our experiences are finite and limited. Consequently, most of the important decisions in our lives are based on sample (incomplete) information. What is a probability sampling distribution? How will sampling distributions help us make good decisions based on incomplete information? (SECTION 6.4)

There is an old saying: All roads lead to Rome. In statistics, we could recast this saying: All probability distributions average out to be normal distributions (as the sample size increases). How can we take advantage of this in our study of sampling distributions? (SECTION 6.5)

The binomial and normal distributions are two of the most important probability distributions in statistics. Under certain limiting conditions, the binomial can be thought to evolve (or envelope) into the normal distribution. How can you apply this concept in the real world? (SECTION 6.6)

Many issues in life come down to success or failure. In most cases, we will not be successful all the time, so proportions of successes are very important. What is the probability sampling distribution for proportions? (SECTION 6.6)



Dana White/PhotoEdit

FOCUS PROBLEM

Impulse Buying

The Food Marketing Institute, Progressive Grocer, New Products News, and Point of Purchase Advertising Institute are organizations that analyze supermarket sales. One of the interesting discoveries was that the average amount of impulse buying in a grocery store is very time-dependent. As reported in *The Denver Post*, “When you dilly dally in a store for



Syracuse Newspapers/David Lassman/The Image Works Image

10 unplanned minutes, you can kiss nearly \$20 good-bye.” For this reason, it is in the best interest of the supermarket to keep you in the store longer. In the *Post* article, it was pointed out that long checkout lines (near end-aisle displays), “samplefest” events of free tasting, video kiosks, magazine and book sections, and so on, help keep customers in the store longer. On average, a single customer who strays from his or her grocery list can plan on impulse spending of \$20 for every 10 minutes spent wandering about in the supermarket.

Let x represent the dollar amount spent on supermarket impulse buying in a 10-minute (unplanned) shopping interval. Based on the *Post* article, the mean of the x distribution is about \$20 and the (estimated) standard deviation is about \$7.

- Consider a random sample of $n = 100$ customers, each of whom has 10 minutes of unplanned shopping time in a supermarket. From the central limit theorem, what can you say about the probability distribution of \bar{x} , the *average* amount spent by these customers due to impulse buying? Is the \bar{x} distribution approximately normal? What are the mean and standard deviation of the \bar{x} distribution? Is it necessary to make any assumption about the x distribution? Explain.
- What is the probability that \bar{x} is between \$18 and \$22?
- Let us assume that x has a distribution that is approximately normal. What is the probability that x is between \$18 and \$22?
- In part (b), we used \bar{x} , the *average* amount spent, computed for 100 customers. In part (c), we used x , the amount spent by only *one* individual customer. The answers to parts (b) and (c) are very different. Why would this happen? In this example, \bar{x} is a much more predictable or reliable statistic than x . Consider that almost all marketing strategies and sales pitches are designed for the *average* customer and *not* the *individual* customer. How does the central limit theorem tell us that the average customer is much more predictable than the individual customer? (See Problem 18 of Section 6.5.)

SECTION 6.1

Graphs of Normal Probability Distributions

FOCUS POINTS

- Graph a normal curve and summarize its important properties.
- Apply the empirical rule to solve real-world problems.
- Use control limits to construct control charts. Examine the chart for three possible out-of-control signals.

Normal distribution

One of the most important examples of a continuous probability distribution is the *normal distribution*. This distribution was studied by the French mathematician Abraham de Moivre (1667–1754) and later by the German mathematician Carl Friedrich Gauss (1777–1855), whose work is so important that the normal distribution is sometimes called *Gaussian*. The work of these mathematicians provided a foundation on which much of the theory of statistical inference is based.

Applications of a normal probability distribution are so numerous that some mathematicians refer to it as “a veritable Boy Scout knife of statistics.” However, before we can apply it, we must examine some of the properties of a normal distribution.

A rather complicated formula, presented later in this section, defines a normal distribution in terms of μ and σ , the mean and standard deviation of the population distribution. It is only through this formula that we can verify if a distribution is normal. However, we can look at the graph of a normal

FIGURE 6-1
A Normal Curve

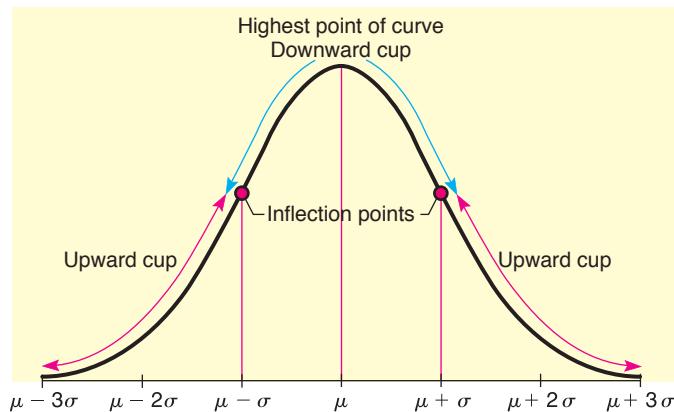


Two other continuous probability distributions are discussed in the section problems. Problems 16 and 17 discuss the *uniform probability distribution*. Problems 18 through 20 discuss the *exponential probability distribution*.

Normal curve

Downward cup
Upward cup

Symmetry of normal curves



distribution and get a good pictorial idea of some of the essential features of any normal distribution.

The graph of a normal distribution is called a *normal curve*. It possesses a shape very much like the cross section of a pile of dry sand. Because of its shape, blacksmiths would sometimes use a pile of dry sand in the construction of a mold for a bell. Thus the normal curve is also called a *bell-shaped curve* (see Figure 6-1).

We see that a general normal curve is smooth and symmetrical about the vertical line extending upward from the mean μ . Notice that the highest point of the curve occurs over μ . If the distribution were graphed on a piece of sheet metal, cut out, and placed on a knife edge, the balance point would be at μ . We also see that the curve tends to level out and approach the horizontal (x axis) like a glider making a landing. However, in mathematical theory, such a glider would never quite finish its landing because a normal curve never touches the horizontal axis.

The parameter σ controls the spread of the curve. The curve is quite close to the horizontal axis at $\mu + 3\sigma$ and $\mu - 3\sigma$. Thus, if the standard deviation σ is large, the curve will be more spread out; if it is small, the curve will be more peaked. Figure 6-1 shows the normal curve cupped downward for an interval on either side of the mean μ . Then it begins to cup upward as we go to the lower part of the bell. The exact places where the *transition* between the upward and downward cupping occur are above the points $\mu + \sigma$ and $\mu - \sigma$. In the terminology of calculus, transition points such as these are called *inflection points*.

Important properties of a normal curve

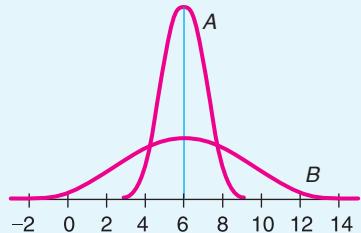
1. The curve is bell-shaped, with the highest point over the mean μ .
2. The curve is symmetrical about a vertical line through μ .
3. The curve approaches the horizontal axis but never touches or crosses it.
4. The inflection (transition) points between cupping upward and downward occur above $\mu + \sigma$ and $\mu - \sigma$.
5. The area under the entire curve is 1.

The parameters that control the shape of a normal curve are the mean μ and the standard deviation σ . When both μ and σ are specified, a specific normal curve is determined. In brief, μ locates the balance point and σ determines the extent of the spread.

GUIDED EXERCISE 1**Identify μ and σ on a normal curve**

Look at the normal curves in Figure 6-2.

FIGURE 6-2



- (a) Do these distributions have the same mean? If so, what is it?
 (b) One of the curves corresponds to a normal distribution with $\sigma = 3$ and the other to one with $\sigma = 1$. Which curve has which σ ?

- The means are the same, since both graphs have the high point over 6. $\mu = 6$.
 → Curve A has $\sigma = 1$ and curve B has $\sigma = 3$. (Since curve B is more spread out, it has the larger σ value.)

COMMENT The normal distribution curve is always above the horizontal axis. The area beneath the curve and above the axis is exactly 1. As such, the normal distribution curve is an example of a *density curve*. The formula used to generate the shape of the normal distribution curve is called the *normal density function*. If x is a normal random variable with mean μ and standard deviation σ , the formula for the normal density function is

$$f(x) = \frac{e^{(-1/2)((x-\mu)/\sigma)^2}}{\sigma\sqrt{2\pi}}$$

In this text, we will not use this formula explicitly. However, we will use tables of areas based on the normal density function.

The total area under any normal curve studied in this book will *always* be 1. The graph of the normal distribution is important because the portion of the *area* under the curve above a given interval represents the *probability* that a measurement will lie in that interval.

In Section 3.2, we studied Chebyshev's theorem. This theorem gives us information about the *smallest* proportion of data that lies within 2, 3, or k standard deviations of the mean. This result applies to *any* distribution. However, for normal distributions, we can get a much more precise result, which is given by the *empirical rule*.

Empirical rule

Empirical rule

For a distribution that is symmetrical and bell-shaped (in particular, for a normal distribution):

Approximately 68% of the data values will lie within 1 standard deviation on each side of the mean.

Approximately 95% of the data values will lie within 2 standard deviations on each side of the mean.

Approximately 99.7% (or almost all) of the data values will lie within 3 standard deviations on each side of the mean.

The preceding statement is called the *empirical rule* because, for symmetrical, bell-shaped distributions, the given percentages are observed in practice. Furthermore, for the normal distribution, the empirical rule is a direct consequence of the very nature of the distribution (see Figure 6-3). Notice that the empirical rule is a stronger statement than Chebyshev's theorem in that it gives *definite percentages*, not just lower limits. Of course, the empirical rule applies only to normal or symmetrical, bell-shaped distributions, whereas Chebyshev's theorem applies to all distributions.

FIGURE 6-3

Area Under a Normal Curve

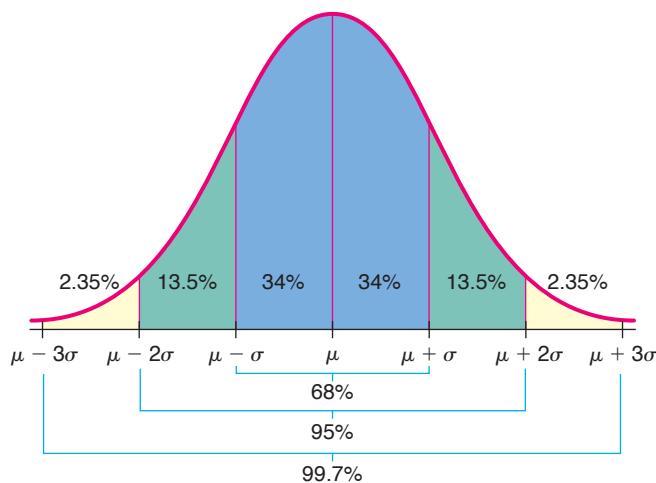
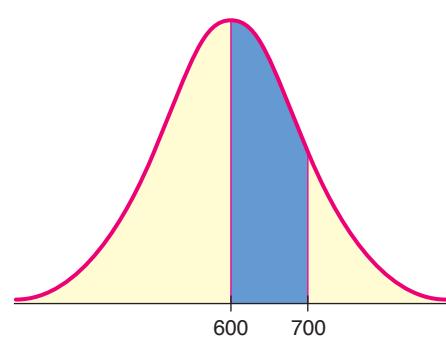


FIGURE 6-4

Distribution of Playing Times

**EXAMPLE 1****EMPIRICAL RULE**

The playing life of a Sunshine radio is normally distributed with mean $\mu = 600$ hours and standard deviation $\sigma = 100$ hours. What is the probability that a radio selected at random will last from 600 to 700 hours?

SOLUTION: The probability that the playing life will be between 600 and 700 hours is equal to the percentage of the total area under the curve that is shaded in Figure 6-4. Since $\mu = 600$ and $\mu + \sigma = 600 + 100 = 700$, we see that the shaded area is simply the area between μ and $\mu + \sigma$. The area from μ to $\mu + \sigma$ is 34% of the total area. This tells us that the probability a Sunshine radio will last between 600 and 700 playing hours is about 0.34.

GUIDED EXERCISE 2**Empirical rule**

The yearly wheat yield per acre on a particular farm is normally distributed with mean $\mu = 35$ bushels and standard deviation $\sigma = 8$ bushels.

- (a) Shade the area under the curve in Figure 6-5 (next page) that represents the probability that an acre will yield between 19 and 35 bushels.

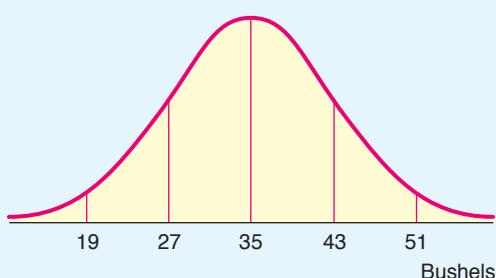
See Figure 6-6.

- (b) Is the area the same as the area between $\mu - 2\sigma$ and μ ? Yes, since $\mu = 35$ and $\mu - 2\sigma = 35 - 2(8) = 19$.

Continued

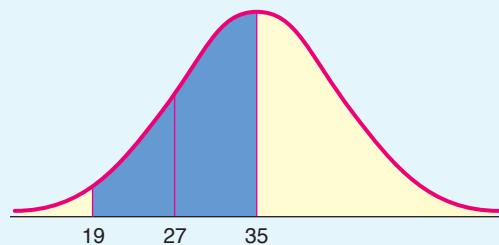
GUIDED EXERCISE 2 *continued*

FIGURE 6-5



- (c) Use Figure 6-3 to find the percentage of area over the interval between 19 and 35.
- (d) What is the probability that the yield will be between 19 and 35 bushels per acre?

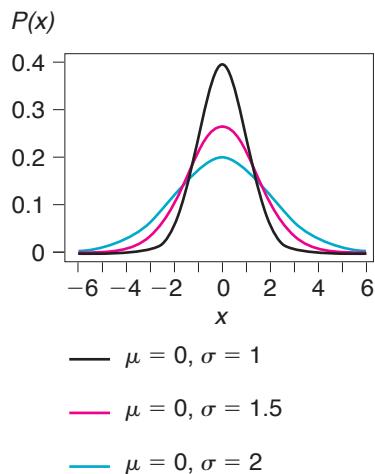
FIGURE 6-6 Completion of Figure 6-5



- The area between the values $\mu - 2\sigma$ and $\mu + 2\sigma$ is 47.5% of the total area.
- It is 47.5% of the total area, which is 1. Therefore, the probability is 0.475 that the yield will be between 19 and 35 bushels.

**TECH NOTES**

We can graph normal distributions using the TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab. In each technology, set the range of x values between $\mu - 3.5\sigma$ and $\mu + 3.5\sigma$. Then use the built-in normal density functions to generate the corresponding y values.



TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Press the **Y =** key. Then, under **DISTR**, select **1:normalpdf (x,μ,σ)** and fill in desired μ and σ values. Press the **WINDOW** key. Set **Xmin** to $\mu - 3\sigma$ and **Xmax** to $\mu + 3\sigma$. Finally, press the **ZOOM** key and select option **0:ZoomFit**.

Excel 2007 In one column, enter x values from $\mu - 3.5\sigma$ to $\mu + 3.5\sigma$ in increments of 0.2σ . In the next column, generate y values by using the ribbon choices **Insert function** f_x

In the dialogue box, select **Statistical** for the Category and then for the Function, select **NORMDIST** ($x,\mu,\sigma,\text{false}$). Next click the **Insert** tab and in the **Charts** group, click **Scatter**. Select the scatter diagram with smooth lines (third grouping).

Minitab In one column, enter x values from -3.5σ to 3.5σ in increments of 0.2σ . In the next column, enter y values by using the menu choices **Calc** \blacktriangleright **Probability Distribution** \blacktriangleright **Normal**. Fill in the dialogue box. Next, use menu choices **Graph** \blacktriangleright **Plot**. Fill in the dialogue box. Under **Display**, select **connect**.

**LOOKING FORWARD**

Normal probability distributions will be used extensively in our later work. For instance, when we repeatedly take samples of the same size from a distribution and compute the same mean for each sample, we'll find that the sample means follow a distribution that is normal or approximately normal (Section 6.5). Also, when the number of trials is sufficiently large, the binomial distribution can be approximated by a normal distribution (Section 6.6). The distribution of the sample proportion of successes in a fixed number of binomial trials also can be approximated by a normal distribution (Section 6.6).

Control charts

Control Charts

If we are examining data over a period of equally spaced time intervals or in some sequential order, then *control charts* are especially useful. Business managers and people in charge of production processes are aware that there exists an inherent amount of variability in any sequential set of data. The sugar content of bottled drinks taken sequentially off a production line, the extent of clerical errors in a bank from day to day, advertising expenses from month to month, or even the number of new customers from year to year are examples of sequential data. There is a certain amount of variability in each.

A random variable x is said to be in *statistical control* if it can be described by the *same* probability distribution when it is observed at successive points in time. Control charts combine graphic and numerical descriptions of data with probability distributions.

Control charts were invented in the 1920s by Walter Shewhart at Bell Telephone Laboratories. Since a control chart is a *warning device*, it is not absolutely necessary that our assumptions and probability calculations be precisely correct. For example, the x distributions need not follow a normal distribution exactly. Any mound-shaped and more or less symmetrical distribution will be good enough.

PROCEDURE

HOW TO MAKE A CONTROL CHART FOR THE RANDOM VARIABLE x

A control chart for a random variable x is a plot of observed x values in time sequence order.

1. Find the mean μ and standard deviation σ of the x distribution by
 - (a) using past data from a period during which the process was “in control” or
 - (b) using specified “target” values for μ and σ .
2. Create a graph in which the vertical axis represents x values and the horizontal axis represents time.
3. Draw a horizontal line at height μ and horizontal, dashed control-limit lines at $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$.
4. Plot the variable x on the graph in time sequence order. Use line segments to connect the points in time sequence order.

How do we pick values for μ and σ ? In most practical cases, values for μ (population mean) and σ (population standard deviation) are computed from past data for which the process we are studying was known to be *in control*. Methods for choosing the sample size to fit given error tolerances can be found in Chapter 7.

Sometimes values for μ and σ are chosen as *target values*. That is, μ and σ values are chosen as set goals or targets that reflect the production level or service level at which a company hopes to perform. To be realistic, such target assignments for μ and σ should be reasonably close to actual data taken when the process was operating at a satisfactory production level. In Example 2, we will make a control chart; then we will discuss ways to analyze it to see if a process or service is “in control.”



EXAMPLE 2

CONTROL CHART

Susan Tamara is director of personnel at the Antlers Lodge in Denali National Park, Alaska. Every summer Ms. Tamara hires many part-time employees from all over the United States. Most are college students seeking summer employment. One of the biggest activities for the lodge staff is that of “making up” the



Mt. McKinley, Denali National Park

FIGURE 6-7

Number of Rooms Not Made Up by 3:30 P.M.



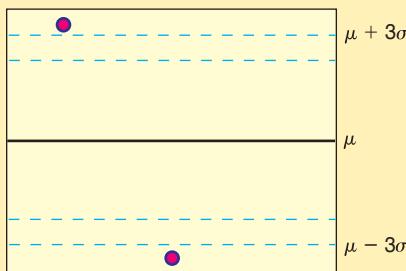
Once we have made a control chart, the main question is the following: As time goes on, is the x variable continuing in this same distribution, or is the distribution of x values changing? If the x distribution is continuing in more or less the same manner, we say it is *in statistical control*. If it is not, we say it is *out of control*.

Many popular methods can set off a warning signal that a process is out of control. Remember, a random variable x is said to be *out of control* if successive time measurements of x indicate that it is no longer following the target probability distribution. We will assume that the target distribution is (approximately) normal and has (user-set) target values for μ and σ .

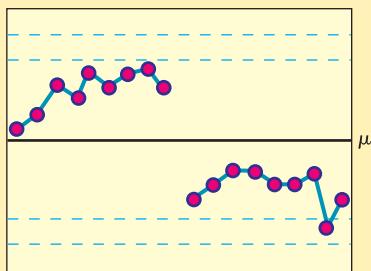
Three of the most popular warning signals are described next.

Out-of-control warning signals

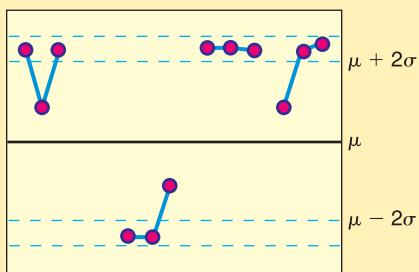
Out-of-Control Signal I



Out-of-Control Signal II



Out-of-Control Signal III

**Out-of-control signals****1. Out-of-Control Signal I: One point falls beyond the 3σ level**

What is the probability that signal I will be a false alarm? By the empirical rule, the probability that a point lies within 3σ of the mean is 0.997. The probability that signal I will be a false alarm is $1 - 0.997 = 0.003$. Remember, a false alarm means that the x distribution is really on the target distribution, and we simply have a very rare (probability of 0.003) event.

2. Out-of-Control Signal II: A run of nine consecutive points on one side of the center line (the line at target value μ)

To find the probability that signal II is a false alarm, we observe that if the x distribution and the target distribution are the same, then there is a 50% chance that the x values will lie above or below the center line at μ . Because the samples are (time) independent, the probability of a run of nine points on one side of the center line is $(0.5)^9 = 0.002$. If we consider both sides, this probability becomes 0.004. Therefore, the probability that signal II is a false alarm is approximately 0.004.

3. Out-of-Control Signal III: At least two of three consecutive points lie beyond the 2σ level on the same side of the center line

To determine the probability that signal III will produce a false alarm, we use the empirical rule. By this rule, the probability that an x value will be above the 2σ level is about 0.023. Using the binomial probability distribution (with success being the point is above 2σ), the probability of two or more successes out of three trials is

$$\frac{3!}{2!1!} (0.023)^2(0.997)^1 + \frac{3!}{3!0!} (0.023)^3(0.997)^0 \approx 0.002$$

Taking into account *both* above and below the center line, it follows that the probability that signal III is a false alarm is about 0.004.

Remember, a control chart is only a warning device, and it is possible to get a false alarm. A false alarm happens when one (or more) of the out-of-control signals occurs, but the x distribution is really on the target or assigned distribution. In this case, we simply have a rare event (probability of 0.003 or 0.004). In practice, whenever a control chart indicates that a process is out of control, it is usually a good precaution to examine what is going on. If the process is out of control, corrective steps can be taken before things get a lot worse. The rare false alarm is a small price to pay if we can avert what might become real trouble.

Type of Warning Signal	Probability of a False Alarm
Type I: Point beyond 3σ	0.003
Type II: Run of nine consecutive points, all below center line μ or all above center line μ	0.004
Type III: At least two out of three consecutive points beyond 2σ	0.004

From an intuitive point of view, signal I can be thought of as a blowup, something dramatically out of control. Signal II can be thought of as a slow drift out of control. Signal III is somewhere between a blowup and a slow drift.

**EXAMPLE 3****INTERPRETING A CONTROL CHART**

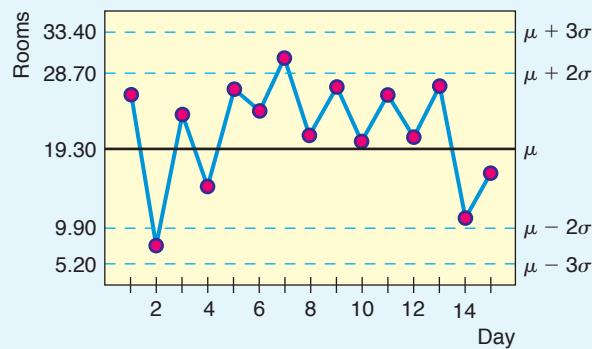
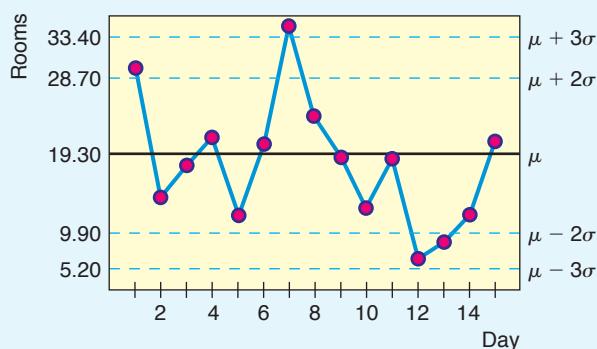
Ms. Tamara of the Antlers Lodge examines the control chart for housekeeping. During the staff meetings, she makes recommendations about improving service or, if all is going well, she gives her staff a well-deserved “pat on the back.” Look at the control chart created in Example 2 (Figure 6-7 on page 256) to determine if the housekeeping process is out of control.

SOLUTION: The x values are more or less evenly distributed about the mean $\mu = 19.3$. None of the points are outside the $\mu \pm 3\sigma$ limit (i.e., above 33.40 or below 5.20 rooms). There is no run of nine consecutive points above or below μ . No two of three consecutive points are beyond the $\mu \pm 2\sigma$ limit (i.e., above 28.7 or below 9.90 rooms).

It appears that the x distribution is “in control.” At the staff meeting, Ms. Tamara should tell her employees that they are doing a reasonably good job and they should keep up the fine work!

**GUIDED EXERCISE 3****Interpreting a control chart**

Figures 6-8 and 6-9 show control charts of housekeeping reports for two other 15-day periods.

FIGURE 6-8 Report II**FIGURE 6-9 Report III**

(a) **Interpret** the control chart in Figure 6-8.



Days 5 to 13 are above $\mu = 19.3$. We have nine consecutive days on one side of the mean. This is a warning signal! It would appear that the mean μ is slowly drifting up beyond the target value of 19.3. The chart indicates that housekeeping is “out of control.” Ms. Tamara should take corrective measures at her staff meeting.

(b) **Interpret** the control chart in Figure 6-9.



On day 7, we have a data value beyond $\mu + 3\sigma$ (i.e., above 33.40). On days 11, 12, and 13, we have two of three data values beyond $\mu - 2\sigma$ (i.e., below 9.90). The occurrences during both of these periods are out-of-control warning signals. Ms. Tamara might ask her staff about both of these periods. There may be a lesson to be learned from day 7, when housekeeping apparently had a lot of trouble. Also, days 11, 12, and 13 were very good days. Perhaps a lesson could be learned about why things went so well.

COMMENT Uniform Distribution and Exponential Distribution

Normal distributions are the central topic of this chapter. Normal distributions are very important in general probability and statistics. However, there are other, more specialized distributions that also have many applications. Two such distributions are the *uniform distribution* and the *exponential distribution*. To learn more about these distributions, please see Problems 16 to 20 at the end of this section.

VIEWPOINT

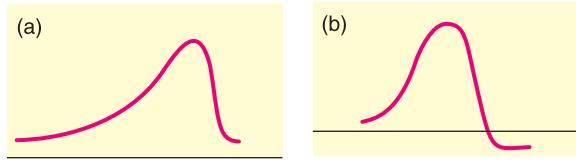
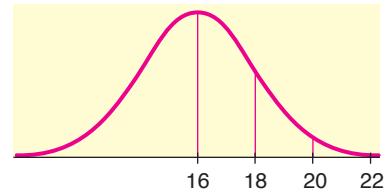
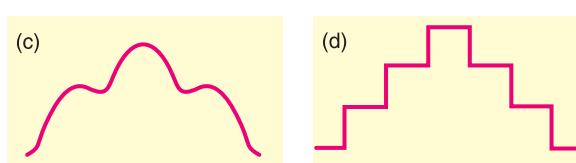
In Control? Out of Control?

If you care about quality, you also must care about control! Dr. Walter Shewhart invented control charts when he was working for Bell Laboratories. The great contribution of control charts is that they separate variation into two sources: (1) random or chance causes (in control) and (2) special or assignable causes (out of control). A process is said to be in statistical control when it is no longer afflicted with special or assignable causes. The performance of a process that is in statistical control is predictable. Predictability and quality control tend to be closely associated.

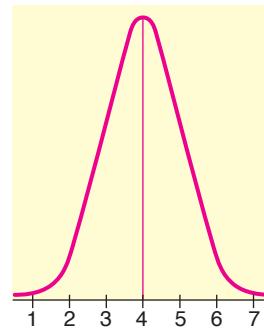
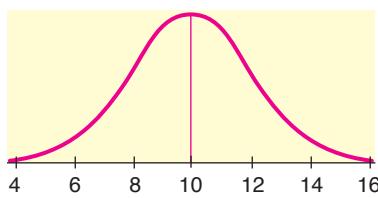
(Source: Adapted from the classic text *Statistical Methods from the Viewpoint of Quality Control*, by W. A. Shewhart, with foreword by W. E. Deming, Dover Publications.)

**SECTION 6.1
PROBLEMS**

- | **Statistical Literacy** Which, if any, of the curves in Figure 6-10 look(s) like a normal curve? If a curve is not a normal curve, tell why.
- | **Statistical Literacy** Look at the normal curve in Figure 6-11, and find μ , $\mu + \sigma$, and σ .

FIGURE 6-10**FIGURE 6-11**

- | **Critical Thinking** Look at the two normal curves in Figures 6-12 and 6-13. Which has the larger standard deviation? What is the mean of the curve in Figure 6-12? What is the mean of the curve in Figure 6-13?

FIGURE 6-13**FIGURE 6-12**

4. **Critical Thinking** Sketch a normal curve
- with mean 15 and standard deviation 2.
 - with mean 15 and standard deviation 3.
 - with mean 12 and standard deviation 2.
 - with mean 12 and standard deviation 3.
 - Consider two normal curves. If the first one has a larger mean than the second one, must it have a larger standard deviation as well? Explain your answer.
5. **Basic Computation: Empirical Rule** What percentage of the area under the normal curve lies
- to the left of μ ?
 - between $\mu - \sigma$ and $\mu + \sigma$?
 - between $\mu - 3\sigma$ and $\mu + 3\sigma$?
6. **Basic Computation: Empirical Rule** What percentage of the area under the normal curve lies
- to the right of μ ?
 - between $\mu - 2\sigma$ and $\mu + 2\sigma$?
 - to the right of $\mu + 3\sigma$?
7. **Distribution: Heights of Coeds** Assuming that the heights of college women are normally distributed with mean 65 inches and standard deviation 2.5 inches (based on information from *Statistical Abstract of the United States*, 112th Edition), answer the following questions. Hint: Use Problems 5 and 6 and Figure 6-3.
- What percentage of women are taller than 65 inches?
 - What percentage of women are shorter than 65 inches?
 - What percentage of women are between 62.5 inches and 67.5 inches?
 - What percentage of women are between 60 inches and 70 inches?
8. **Distribution: Rhode Island Red Chicks** The incubation time for Rhode Island Red chicks is normally distributed with a mean of 21 days and standard deviation of approximately 1 day (based on information from *World Book Encyclopedia*). Look at Figure 6-3 and answer the following questions. If 1000 eggs are being incubated, how many chicks do we expect will hatch
- in 19 to 23 days?
 - in 20 to 22 days?
 - in 21 days or fewer?
 - in 18 to 24 days? (Assume all eggs eventually hatch.)
- Note:* In this problem, let us agree to think of a single day or a succession of days as a continuous interval of time.
9. **Archaeology: Tree Rings** At Burnt Mesa Pueblo, archaeological studies have used the method of tree-ring dating in an effort to determine when prehistoric people lived in the pueblo. Wood from several excavations gave a mean of (year) 1243 with a standard deviation of 36 years (*Bandelier Archaeological Excavation Project: Summer 1989 Excavations at Burnt Mesa Pueblo*, edited by Kohler, Washington State University Department of Anthropology). The distribution of dates was more or less mound-shaped and symmetrical about the mean. Use the empirical rule to
- estimate a range of years centered about the mean in which about 68% of the data (tree-ring dates) will be found.
 - estimate a range of years centered about the mean in which about 95% of the data (tree-ring dates) will be found.
 - estimate a range of years centered about the mean in which almost all the data (tree-ring dates) will be found.
10. **Vending Machine: Soft Drinks** A vending machine automatically pours soft drinks into cups. The amount of soft drink dispensed into a cup is normally



- distributed with a mean of 7.6 ounces and standard deviation of 0.4 ounce. Examine Figure 6-3 and answer the following questions.
- Estimate the probability that the machine will overflow an 8-ounce cup.
 - Estimate the probability that the machine will not overflow an 8-ounce cup.
 - The machine has just been loaded with 850 cups. How many of these do you expect will overflow when served?
11. **Pain Management: Laser Therapy** “Effect of Helium-Neon Laser Auriculotherapy on Experimental Pain Threshold” is the title of an article in the journal *Physical Therapy* (Vol. 70, No. 1, pp. 24–30). In this article, laser therapy was discussed as a useful alternative to drugs in pain management of chronically ill patients. To measure pain threshold, a machine was used that delivered low-voltage direct current to different parts of the body (wrist, neck, and back). The machine measured current in milliamperes (mA). The pretreatment experimental group in the study had an average threshold of pain (pain was first detectable) at $\mu = 3.15$ mA with standard deviation $\sigma = 1.45$ mA. Assume that the distribution of threshold pain, measured in milliamperes, is symmetrical and more or less mound-shaped. Use the empirical rule to
 - estimate a range of milliamperes centered about the mean in which about 68% of the experimental group had a threshold of pain.
 - estimate a range of milliamperes centered about the mean in which about 95% of the experimental group had a threshold of pain.
12. **Control Charts: Yellowstone National Park** Yellowstone Park Medical Services (YPMS) provides emergency health care for park visitors. Such health care includes treatment for everything from indigestion and sunburn to more serious injuries. A recent issue of *Yellowstone Today* (National Park Service Publication) indicated that the average number of visitors treated each day by YPMS is 21.7. The estimated standard deviation is 4.2 (summer data). The distribution of numbers treated is approximately mound-shaped and symmetrical.
 - For a 10-day summer period, the following data show the number of visitors treated each day by YPMS:

Day	1	2	3	4	5	6	7	8	9	10
Number treated	25	19	17	15	20	24	30	19	16	23

 Make a control chart for the daily number of visitors treated by YPMS, and plot the data on the control chart. Do the data indicate that the number of visitors treated by YPMS is “in control”? Explain your answer.
 - For another 10-day summer period, the following data were obtained:

Day	1	2	3	4	5	6	7	8	9	10
Number treated	20	15	12	21	24	28	32	36	35	37

 Make a control chart, and plot the data on the chart. **Interpretation** Do the data indicate that the number of visitors treated by YPMS is “in control” or “out of control”? Explain your answer. Identify all out-of-control signals by type (I, II, or III). If you were the park superintendent, do you think YPMS might need some (temporary) extra help? Explain.
13. **Control Charts: Bank Loans** Tri-County Bank is a small independent bank in central Wyoming. This is a rural bank that makes loans on items as small as horses and pickup trucks to items as large as ranch land. Total monthly loan requests are used by bank officials as an indicator of economic business conditions in this rural community. The mean monthly loan request for the past several years has been 615.1 (in thousands of dollars) with a standard deviation of 11.2 (in thousands of dollars). The distribution of loan requests is approximately mound-shaped and symmetrical.



- (a) For 12 months, the following monthly loan requests (in thousands of dollars) were made to Tri-County Bank:

Month	1	2	3	4	5	6
Loan request	619.3	625.1	610.2	614.2	630.4	615.9
Month	7	8	9	10	11	12

Month	7	8	9	10	11	12
Loan request	617.2	610.1	592.7	596.4	585.1	588.2

Make a control chart for the total monthly loan requests, and plot the preceding data on the control chart. *Interpretation* From the control chart, would you say the local business economy is heating up or cooling down? Explain your answer by referring to any trend you may see on the control chart. Identify all out-of-control signals by type (I, II, or III).

- (b) For another 12-month period, the following monthly loan requests (in thousands of dollars) were made to Tri-County Bank:

Month	1	2	3	4	5	6
Loan request	608.3	610.4	615.1	617.2	619.3	622.1
Month	7	8	9	10	11	12

Month	7	8	9	10	11	12
Loan request	625.7	633.1	635.4	625.0	628.2	619.8

Make a control chart for the total monthly loan requests, and plot the preceding data on the control chart. *Interpretation* From the control chart, would you say the local business economy is heating up, cooling down, or about normal? Explain your answer by referring to the control chart. Identify all out-of-control signals by type (I, II, or III).

14. **Control Charts: Motel Rooms** The manager of Motel 11 has 316 rooms in Palo Alto, California. From observation over a long period of time, she knows that on an average night, 268 rooms will be rented. The long-term standard deviation is 12 rooms. This distribution is approximately mound-shaped and symmetrical.

- (a) For 10 consecutive nights, the following numbers of rooms were rented each night:

Night	1	2	3	4	5	6
Number of rooms	234	258	265	271	283	267
Night	7	8	9	10		

Night	7	8	9	10		
Number of rooms	290	286	263	240		

Make a control chart for the number of rooms rented each night, and plot the preceding data on the control chart. *Interpretation* Looking at the control chart, would you say the number of rooms rented during this 10-night period has been unusually low? unusually high? about what you expected? Explain your answer. Identify all out-of-control signals by type (I, II, or III).

- (b) For another 10 consecutive nights, the following numbers of rooms were rented each night:

Night	1	2	3	4	5	6
Number of rooms	238	245	261	269	273	250
Night	7	8	9	10		

Night	7	8	9	10
Number of rooms	241	230	215	217

Make a control chart for the number of rooms rented each night, and plot the preceding data on the control chart. **Interpretation** Would you say the room occupancy has been high? low? about what you expected? Explain your answer. Identify all out-of-control signals by type (I, II, or III).

15. **Control Chart: Air Pollution** The visibility standard index (VSI) is a measure of Denver air pollution that is reported each day in the *Rocky Mountain News*. The index ranges from 0 (excellent air quality) to 200 (very bad air quality). During winter months, when air pollution is higher, the index has a mean of about 90 (rated as fair) with a standard deviation of approximately 30. Suppose that for 15 days, the following VSI measures were reported each day:

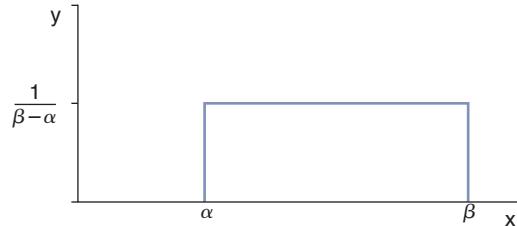
Day	1	2	3	4	5	6	7	8	9
VSI	80	115	100	90	15	10	53	75	80

Day	10	11	12	13	14	15
VSI	110	165	160	120	140	195

Make a control chart for the VSI, and plot the preceding data on the control chart. Identify all out-of-control signals (high or low) that you find in the control chart by type (I, II, or III).

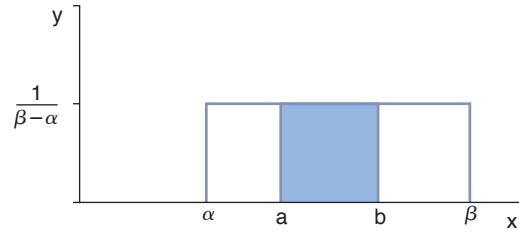


16. **Expand Your Knowledge: Continuous Uniform Probability Distribution** Let α and β be any two constants such that $\alpha < \beta$. Suppose we choose a point x at random in the interval from α to β . In this context the phrase *at random* is taken to mean that the point x is as likely to be chosen from one particular part of the interval as any other part. Consider the rectangle.



The base of the rectangle has length $\beta - \alpha$ and the height of the rectangle is $1/(\beta - \alpha)$, so the area of the rectangle is 1. As such, this rectangle's top can be thought of as part of a probability density curve. Since we specify that x must lie between α and β , the probability of a point occurring outside the interval $[\alpha, \beta]$ is, by definition, 0. From a geometric point of view, x chosen at random from α to β means we are equally likely to land anywhere in the interval from α to β . For this reason, the top of the (rectangle's) density curve is flat or uniform.

Now suppose that a and b are numbers such that $\alpha \leq a < b \leq \beta$. What is the probability that a number x chosen at random from α to β will fall in the interval $[a, b]$? Consider the graph



Because x is chosen at random from $[\alpha, \beta]$, the area of the rectangle that lies above $[a, b]$ is the probability that x lies in $[a, b]$. This area is

$$P(a < x < b) = \frac{b - a}{\beta - \alpha}$$

In this way we can assign a probability to any interval inside $[\alpha, \beta]$. This probability distribution is called the *continuous uniform distribution* (also called the rectangular distribution). Using some extra mathematics, it can be shown that if x is a random variable with this distribution, then the mean and standard deviation of x are

$$\mu = \frac{\alpha + \beta}{2} \quad \text{and} \quad \sigma = \frac{\beta - \alpha}{\sqrt{12}}$$

Sedimentation experiments are very important in the study of biology, medicine, hydrodynamics, petroleum engineering, civil engineering, and so on. The size (diameter) of approximately spherical particles is important since larger particles hinder and sometimes block the movement of smaller particles. Usually the size of sediment particles follows a uniform distribution (Reference: Y. Zimmels, "Theory of Kindred Sedimentation of Polydisperse Mixtures," *AIChE Journal*, Vol. 29, No. 4, pp. 669–676).

Suppose a veterinary science experiment injects very small, spherical pellets of low-level radiation directly into an animal's bloodstream. The purpose is to attempt to cure a form of recurring cancer. The pellets eventually dissolve and pass through the animal's system. Diameters of the pellets are uniformly distributed from 0.015 mm to 0.065 mm. If a pellet enters an artery, what is the probability that it will be the following sizes?

- (a) 0.050 mm or larger. *Hint:* All particles are between 0.015 mm and 0.065 mm, so larger than 0.050 means $0.050 \leq x \leq 0.065$.
- (b) 0.040 mm or smaller
- (c) between 0.035 mm and 0.055 mm
- (d) Compute the mean size of the particles.
- (e) Compute the standard deviation of particle size.



17.

Uniform Distribution: Measurement Errors Measurement errors from instruments are often modeled using the uniform distribution (see Problem 16). To determine the range of a large public address system, acoustical engineers use a method of triangulation to measure the shock waves sent out by the speakers. The time at which the waves arrive at the sensors must be measured accurately. In this context, a negative error means the signal arrived too early. A positive error means the signal arrived too late. Measurement errors in reading these times have a uniform distribution from -0.05 to $+0.05$ microseconds. (Reference: J. Perruzzi, and E. Hilliard, "Modeling Time Delay Measurement Errors," *Journal of the Acoustical Society of America*, Vol. 75, No. 1, pp. 197–201.) What is the probability that such measurements will be in error by

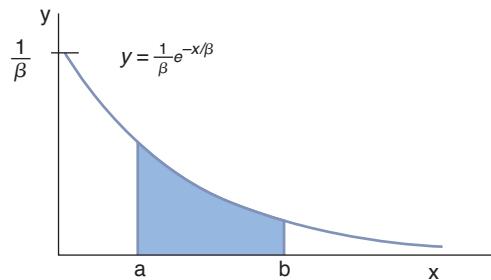
- (a) less than $+0.03$ microsecond (i.e., $-0.05 \leq x < 0.03$)?
- (b) more than -0.02 microsecond?
- (c) between -0.04 and $+0.01$ microsecond?
- (d) Find the mean and standard deviation of measurement errors. Measurements from an instrument are called *unbiased* if the mean of the measurement errors is zero. Would you say the measurements for these acoustical sensors are unbiased? Explain.



18.

Expand Your Knowledge: Exponential Distribution The Poisson distribution (Section 5.4) gives the probability for the *number of occurrences* for a "rare" event. Now, let x be a random variable that represents the *waiting time* between rare events. Using some mathematics, it can be shown that x has an *exponential distribution*. Let $x > 0$ be a random variable and let $\beta > 0$ be a constant. Then

$y = \frac{1}{\beta} e^{-x/\beta}$ is a curve representing the exponential distribution. Areas under this curve give us exponential probabilities.



If a and b are any numbers such that $0 < a < b$, then using some extra mathematics, it can be shown that the area under the curve above the interval $[a, b]$ is

$$P(a < x < b) = e^{-a/\beta} - e^{-b/\beta}$$

Notice that by definition, x cannot be negative, so, $P(x < 0) = 0$. The random variable x is called an *exponential random variable*. Using some more mathematics, it can be shown that the mean and standard deviation of x are

$$\mu = \beta \quad \text{and} \quad \sigma = \beta$$

Note: The number $e = 2.71828 \dots$ is used throughout probability, statistics, and mathematics. The key e^x is conveniently located on most calculators.

Comment: The Poisson and exponential distributions have a special relationship. Specifically, it can be shown that the *waiting time* between successive Poisson arrivals (i.e., successes or rare events) has an exponential distribution with $\beta = 1/\lambda$, where λ is the average number of Poisson successes (rare events) per unit of time. For more on this topic, please see Problem 20.)

Fatal accidents on scheduled domestic passenger flights are rare events. In fact, airlines do all they possibly can to prevent such accidents. However, around the world such fatal accidents do occur. Let x be a random variable representing the waiting time between fatal airline accidents. Research has shown that x has an exponential distribution with a mean of approximately 44 days (Reference: R. Pyke, "Spacings," *Journal of the Royal Statistical Society B*, Vol. 27, No. 3, p. 426.)

We take the point of view that x (measured in days as units) is a continuous random variable. Suppose a fatal airline accident has just been reported on the news. What is the probability that the waiting time to the next reported fatal airline accident is

- (a) less than 30 days (i.e., $0 \leq x < 30$)?
- (b) more than 50 days (i.e., $50 < x < \infty$)? Hint: $e^{-\infty} = 0$.
- (c) between 20 and 60 days?
- (d) What are the mean and the standard deviation of the waiting times x ?



19.

Exponential Distribution: Supply and Demand Another application for exponential distributions (see Problem 18) is supply/demand problems. The operator of a pumping station in a small Wyoming town has observed that demand for water on a typical summer afternoon is exponentially distributed with a mean of 75 cfs (cubic feet per second). Let x be a random variable that represents the town's demand for water (in cfs). What is the probability that on a typical summer afternoon, this town will have a water demand x

- (a) more than 60 cfs (i.e., $60 < x < \infty$)? Hint: $e^{-\infty} = 0$.
- (b) less than 140 cfs (i.e., $0 < x < 140$)?
- (c) between 60 and 100 cfs?
- (d) **Brain teaser** How much water c (in cfs) should the station pump to be 80% sure that the town demand x (in cfs) will not exceed the supply c ? Hint: First explain why the equation $P(0 < x < c) = 0.80$ represents the problem as stated. Then solve for c .



blueflames/Stockphoto.com



20.

- Exponential Distribution: Waiting Time Between Poisson Events** A common application of exponential distributions (see Problem 18) is waiting time between Poisson events (e.g., successes in the Poisson distribution; see Section 5.4). In our study of the Poisson distribution (Example 11, Section 5.4), we saw that the mean success rate per hour of catching a fish at Pyramid Lake is $\lambda = 0.667$ fish/hour. From this we see that the mean waiting time between fish can be thought of as $\beta = 1/\lambda = 1/0.667 \approx 1.5$ hours/fish. Remember, the fish at Pyramid Lake tend to be large. Suppose you have just caught a fish. Let x be a random variable representing the waiting time (in hours) to catch the next fish. Use the exponential distribution to determine the probability that the waiting time is
- less than half an hour (i.e., $0 < x < 0.5$).
 - more than 3 hours (i.e., $3 < x < \infty$). Hint: $e^{-\infty} = 0$.
 - between 1 and 3 hours (i.e., $1 < x < 3$).
 - What are the mean and the standard deviation of the x distribution?

SECTION 6.2

Standard Units and Areas Under the Standard Normal Distribution

FOCUS POINTS

- Given μ and σ , convert raw data to z scores.
- Given μ and σ , convert z scores to raw data.
- Graph the standard normal distribution, and find areas under the standard normal curve.

z Scores and Raw Scores

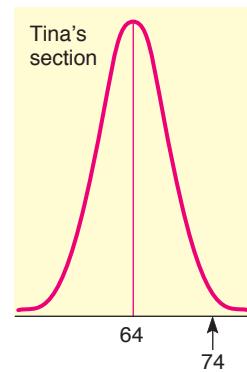
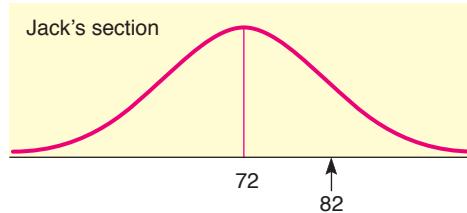
Normal distributions vary from one another in two ways: The mean μ may be located anywhere on the x axis, and the bell shape may be more or less spread according to the size of the standard deviation σ . The differences among the normal distributions cause difficulties when we try to compute the area under the curve in a specified interval of x values and, hence, the probability that a measurement will fall into that interval.

It would be a futile task to try to set up a table of areas under the normal curve for each different μ and σ combination. We need a way to standardize the distributions so that we can use *one* table of areas for *all* normal distributions. We achieve this standardization by considering how many standard deviations a measurement lies from the mean. In this way, we can compare a value in one normal distribution with a value in another, different normal distribution. The next situation shows how this is done.

Suppose Tina and Jack are in two different sections of the same course. Each section is quite large, and the scores on the midterm exams of each section follow a normal distribution. In Tina's section, the average (mean) was 64 and her score was 74. In Jack's section, the mean was 72 and his score was 82. Both Tina and Jack were pleased that their scores were each 10 points above the average of each respective section. However, the fact that each was 10 points above average does not really tell us how each did *with respect to the other students in the section*. In Figure 6-14, we see the normal distribution of grades for each section.

Tina's 74 was higher than most of the other scores in her section, while Jack's 82 is only an upper-middle score in his section. Tina's score is far better with respect to her class than Jack's score is with respect to his class.

FIGURE 6-14
Distributions of Midterm Scores



The preceding situation demonstrates that it is not sufficient to know the difference between a measurement (x value) and the mean of a distribution. We need also to consider the spread of the curve, or the standard deviation. What we really want to know is the number of standard deviations between a measurement and the mean. This “distance” takes both μ and σ into account.

We can use a simple formula to compute the number z of standard deviations between a measurement x and the mean μ of a normal distribution with standard deviation σ :

$$\left(\frac{\text{Number of standard deviations}}{\text{between the measurement and the mean}} \right) = \left(\frac{\text{Difference between the measurement and the mean}}{\text{Standard deviation}} \right)$$

z score or standard score

The ***z value*** or ***z score*** gives the number of standard deviations between the original measurement x and the mean μ of the x distribution.

$$z = \frac{x - \mu}{\sigma}$$

TABLE 6-2

x Values and Corresponding z Values

<i>x</i> Value in Original Distribution	Corresponding <i>z</i> Value or Standard Unit
$x = \mu$	$z = 0$
$x > \mu$	$z > 0$
$x < \mu$	$z < 0$

Standard units

The mean is a special value of a distribution. Let’s see what happens when we convert $x = \mu$ to a z value:

$$z = \frac{x - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$$

The mean of the original distribution is always zero, in *standard units*. This makes sense because the mean is zero standard variations from itself.

An x value in the original distribution that is *above* the mean μ has a corresponding z value that is *positive*. Again, this makes sense because a measurement above the mean would be a positive number of standard deviations from the mean. Likewise, an x value *below* the mean has a *negative* z value. (See Table 6-2.)

Note

Unless otherwise stated, in the remainder of the book we will take the word *average* to be either the sample arithmetic mean \bar{x} or the population mean μ .

EXAMPLE 4**STANDARD SCORE**

A pizza parlor franchise specifies that the average (mean) amount of cheese on a large pizza should be 8 ounces and the standard deviation only 0.5 ounce. An inspector picks out a large pizza at random in one of the pizza parlors and finds that it is made with 6.9 ounces of cheese. Assume that the amount of cheese on a pizza follows a normal distribution. If the amount of cheese is below the mean by more than 3 standard deviations, the parlor will be in danger of losing its franchise.

How many standard deviations from the mean is 6.9? Is the pizza parlor in danger of losing its franchise?

SOLUTION: Since we want to know the number of standard deviations from the mean, we want to convert 6.9 to standard z units.

$$z = \frac{x - \mu}{\sigma} = \frac{6.9 - 8}{0.5} = -2.20$$

Interpretation The amount of cheese on the selected pizza is only 2.20 standard deviations below the mean. The fact that z is negative indicates that the amount of cheese is 2.20 standard deviations *below* the mean. The parlor will not lose its franchise based on this sample.

Lois Ellen Frank/CORBIS

Raw score, x

We have seen how to convert from x measurements to standard units z . We can easily reverse the process to find the original *raw score* x if we know the mean and standard deviation of the original x distribution. Simply solve the z score formula for x .

Given an x distribution with mean μ and standard deviation σ , the **raw score** x corresponding to a z score is

$$x = z\sigma + \mu$$

GUIDED EXERCISE 4**Standard score and raw score**

Rod figures that it takes an average (mean) of 17 minutes with a standard deviation of 3 minutes to drive from home, park the car, and walk to an early morning class.

- (a) One day it took Rod 21 minutes to get to class.
How many standard deviations from the average is that? Is the z value positive or negative?
Explain why it should be either positive or negative.



The number of standard deviations from the mean is given by the z value:

$$z = \frac{x - \mu}{\sigma} = \frac{21 - 17}{3} \approx 1.33$$

The z value is positive. We should expect a positive z value, since 21 minutes is *more* than the mean of 17.

- (b) What commuting time corresponds to a standard score of $z = -2.5$? **Interpretation** Could Rod count on making it to class in this amount of time or less?



$$\begin{aligned} x &= z\sigma + \mu \\ &= (-2.5)(3) + 17 \\ &= 9.5 \text{ minutes} \end{aligned}$$

No, commute times at or less than 2.5 standard deviations below the mean are rare.

LOOKING FORWARD

The basic structure of the formula for the standard score of a distribution is very general. When we verbalize the formula, we see it is

$$z = \frac{\text{measurement} - \text{mean of the distribution}}{\text{standard deviation of the distribution}}$$

We will see this general formula used again and again. In particular, when we look at sampling distributions for the mean (Section 6.4) and when we use the normal approximation of the binomial distribution (Section 6.6), we'll see this formula. We'll also see it when we discuss the sampling distribution for proportions (Section 6.6). Further uses occur in computations for confidence intervals (Chapter 7) and hypothesis testing (Chapter 8).

Standard Normal Distribution

If the original distribution of x values is normal, then the corresponding z values have a normal distribution as well. The z distribution has a mean of 0 and a standard deviation of 1. The normal curve with these properties has a special name.

Standard normal distribution

The standard normal distribution is a normal distribution with $\mu = 0$ and standard deviation $\sigma = 1$ (Figure 6-15).

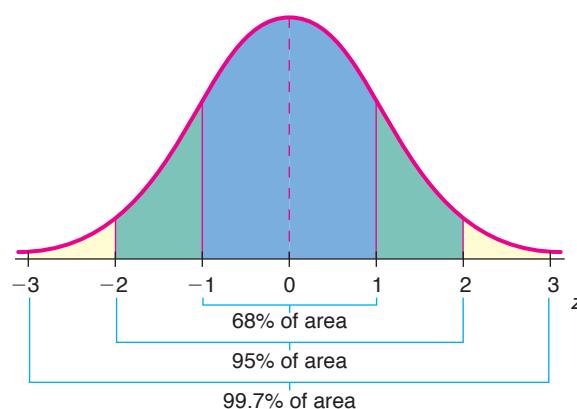
Any normal distribution of x values can be converted to the standard normal distribution by converting all x values to their corresponding z values. The resulting standard distribution will always have mean $\mu = 0$ and standard deviation $\sigma = 1$.

Areas Under the Standard Normal Curve

We have seen how to convert any normal distribution to the standard normal distribution. We can change any x value to a z value and back again. But what is the advantage of all this work? The advantage is that there are extensive tables that show the area under the standard normal curve for almost any interval along the z axis. The areas are important because each area is equal to the probability that the measurement of an item selected at random falls in this interval. Thus, the standard normal distribution can be a tremendously helpful tool.

FIGURE 6-15

The Standard Normal Distribution
($\mu = 0, \sigma = 1$)



Using a Standard Normal Distribution Table

Using a table to find areas and probabilities associated with the standard normal distribution is a fairly straightforward activity. However, it is important to first observe the range of z values for which areas are given. This range is usually depicted in a picture that accompanies the table.

Left-tail style table

In this text, we will use the left-tail style table. This style table gives cumulative areas to the left of a specified z . Determining other areas under the curve utilizes the fact that the area under the entire curve is 1. Taking advantage of the symmetry of the normal distribution is also useful. The procedures you learn for using the left-tail style normal distribution table apply directly to cumulative normal distribution areas found on calculators and in computer software packages such as Excel 2007 and Minitab.



EXAMPLE 5

STANDARD NORMAL DISTRIBUTION TABLE

Use Table 5 of Appendix II to find the described areas under the standard normal curve.

- (a) Find the area under the standard normal curve to the left of $z = -1.00$.

SOLUTION: First, shade the area to be found on the standard normal distribution curve, as shown in Figure 6-16. Notice that the z value we are using is negative. This means that we will look at the portion of Table 5 of Appendix II for which the z values are negative. In the upper-left corner of the table we see the letter z . The column under z gives us the units value and tenths value for z . The other column headings indicate the hundredths value of z . Table entries give areas under the standard normal curve to the left of the listed z values. To find the area to the left of $z = -1.00$, we use the row headed by -1.0 and then move to the column headed by the hundredths position, .00. This entry is shaded in Table 6-3. We see that the area is 0.1587.

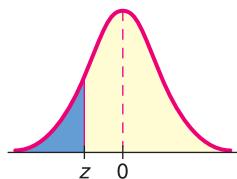


TABLE 6-3 Excerpt from Table 5 of Appendix II Showing Negative z Values

z	.00	.0107	.08	.09
-3.4	.0003	.00030003	.0003	.0002
:						
-1.1	.1357	.13351210	.1190	.1170
-1.0	.1587	.15621423	.1401	.1379
-0.9	.1841	.18141660	.1635	.1611
:						
-0.0	.5000	.49604721	.4681	.4641

- (b) Find the area to the left of $z = 1.18$, as illustrated in Figure 6-17.

FIGURE 6-16

Area to the Left of $z = -1.00$

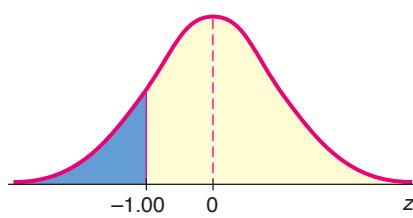
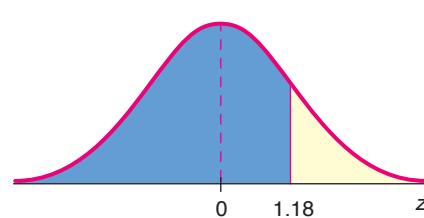


FIGURE 6-17

Area to the Left of $z = 1.18$



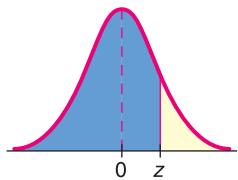


TABLE 6-4

Excerpt from Table 5 of Appendix II Showing Positive z Values

<i>z</i>	.00	.01	.0208	.09
0.0	.5000	.5040	.50805319	.5359
:						
0.9	.8159	.8186	.82128365	.8359
1.0	.8413	.8438	.84618599	.8621
1.1	.8643	.8665	.86868810	.8830
:						
3.4	.9997	.9997	.99979997	.9998

SOLUTION: In this case, we are looking for an area to the left of a positive z value, so we look in the portion of Table 5 that shows positive z values. Again, we first sketch the area to be found on a standard normal curve, as shown in Figure 6-17. Look in the row headed by 1.1 and move to the column headed by .08. The desired area is shaded (see Table 6-4). We see that the area to the left of 1.18 is 0.8810.

GUIDED EXERCISE 5**Using the standard normal distribution table**

Table 5, Areas of a Standard Normal Distribution, is located in Appendix II as well as in the end-papers of the text. Spend a little time studying the table, and then answer these questions.

- (a) As z values increase, do the areas to the left of z increase? Yes. As z values increase, we move to the right on the normal curve, and the areas increase.
- (b) If a z value is negative, is the area to the left of z less than 0.5000? Yes. Remember that a negative z value is on the left side of the standard normal distribution. The entire left half of the normal distribution has area 0.5, so any area to the left of $z = 0$ will be less than 0.5.
- (c) If a z value is positive, is the area to the left of z greater than 0.5000? Yes. Positive z values are on the right side of the standard normal distribution, and any area to the left of a positive z value includes the entire left half of the normal distribution.

Using Table 5 to find other areas

Table 5 gives areas under the standard normal distribution that are to the *left of a z value*. How do we find other areas under the standard normal curve?

PROCEDURE**HOW TO USE A LEFT-TAIL STYLE STANDARD NORMAL DISTRIBUTION TABLE**

- For areas to the left of a specified z value, use the table entry directly.
- For areas to the right of a specified z value, look up the table entry for z and subtract the area from 1.
Note: Another way to find the same area is to use the symmetry of the normal curve and look up the table entry for $-z$.
- For areas between two z values, z_1 and z_2 (where $z_2 > z_1$), subtract the table area for z_1 from the table area for z_2 .

FIGURE 6-18

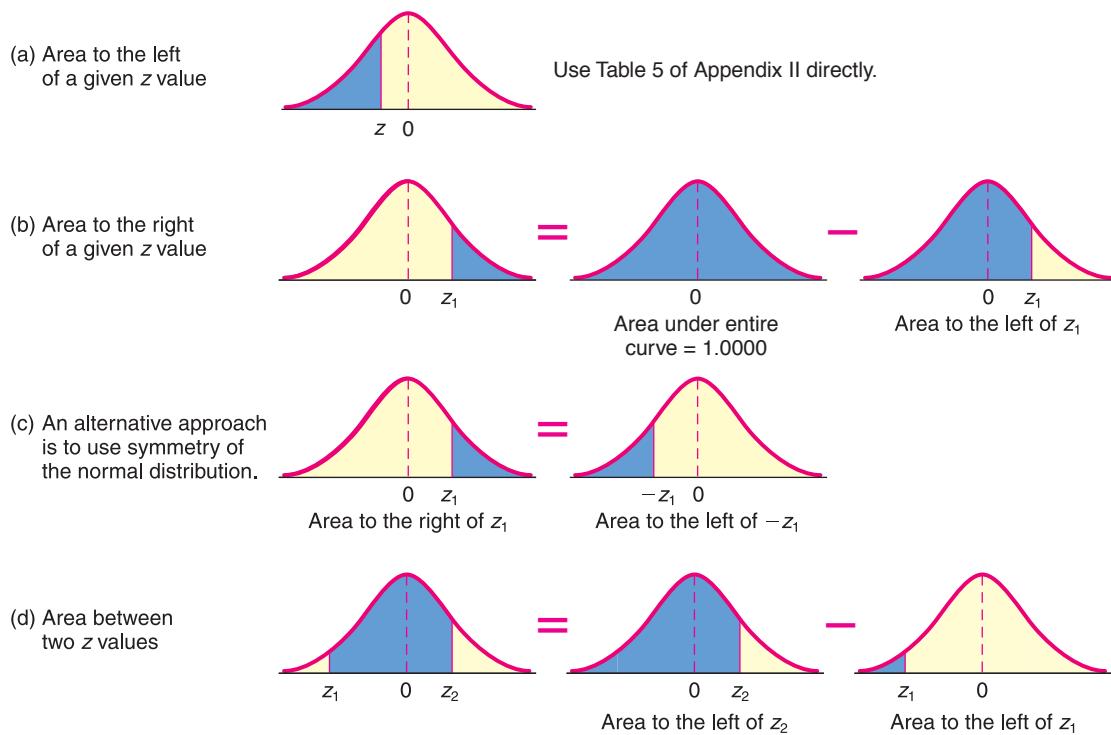


Figure 6-18 illustrates the procedure for using Table 5, Areas of a Standard Normal Distribution, to find any specified area under the standard normal distribution. Again, it is useful to sketch the area in question before you use Table 5.

COMMENT Notice that the z values shown in Table 5 of Appendix II are formatted to the hundredths position. It is convenient to *round or format z values to the hundredths position* before using the table. The areas are all given to four places after the decimal, so give your answers to four places after the decimal.

COMMENT The smallest z value shown in Table 5 is -3.49 , while the largest value is 3.49 . These values are, respectively, far to the left and far to the right on the standard normal distribution, with very little area beyond either value. We will follow the common convention of treating any area to the left of a z value smaller than -3.49 as 0.000 . Similarly, we will consider any area to the right of a z value greater than 3.49 as 0.000 . We understand that there is some area in these extreme tails. However, these areas are each less than 0.0002 . Now let's get real about this! Some very specialized applications, beyond the scope of this book, do need to measure areas and corresponding probabilities in these extreme tails. But in most practical applications, *we follow the convention of treating the areas in the extreme tails as zero*.

Convention for using Table 5 of Appendix II

1. Treat any area to the left of a z value smaller than -3.49 as 0.000 .
2. Treat any area to the left of a z value greater than 3.49 as 1.000 .



EXAMPLE 6

USING TABLE TO FIND AREAS

Use Table 5 of Appendix II to find the specified areas.

- (a) Find the area between $z = 1.00$ and $z = 2.70$.

SOLUTION: First, sketch a diagram showing the area (see Figure 6-19). Because we are finding the area between two z values, we subtract corresponding table entries.

$$\begin{aligned}(\text{Area between } 1.00 \text{ and } 2.70) &= (\text{Area left of } 2.70) - (\text{Area left of } 1.00) \\&= 0.9965 - 0.8413 \\&= 0.1552\end{aligned}$$

(b) Find the area to the right of $z = 0.94$.

SOLUTION: First, sketch the area to be found (see Figure 6-20).

$$\begin{aligned}(\text{Area to right of } 0.94) &= (\text{Area under entire curve}) - (\text{Area to left of } 0.94) \\&= 1.000 - 0.8264 \\&= 0.1736\end{aligned}$$

Alternatively,

$$\begin{aligned}(\text{Area to right of } 0.94) &= (\text{Area to left of } -0.94) \\&= 0.1736\end{aligned}$$

FIGURE 6-19

Area from $z = 1.00$ to $z = 2.70$

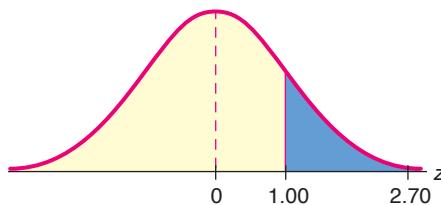
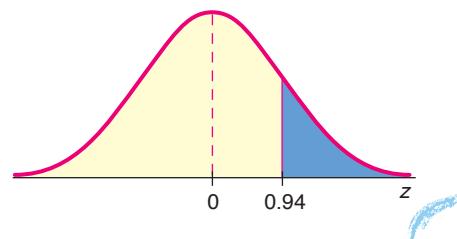


FIGURE 6-20

Area to the Right of $z = 0.94$.



Probabilities associated with the standard normal distribution

We have practiced the skill of finding areas under the standard normal curve for various intervals along the z axis. This skill is important because *the probability* that z lies in an interval *is given by the area* under the standard normal curve above that interval.

Because the normal distribution is continuous, there is no area under the curve exactly over a specific z . Therefore, probabilities such as $P(z \geq z_1)$ are the same as $P(z > z_1)$. When dealing with probabilities or areas under a normal curve that are specified with inequalities, *strict inequality* symbols can be used *interchangeably* with *inequality or equal* symbols.

GUIDED EXERCISE 6

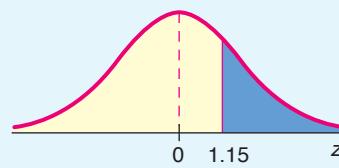
Probabilities associated with the standard normal distribution

Let z be a random variable with a standard normal distribution.

- (a) $P(z \geq 1.15)$ refers to the probability that z values lie to the right of 1.15. Shade the corresponding area under the standard normal curve (Figure 6-21) and find $P(z \geq 1.15)$.



FIGURE 6-21 Area to Be Found



$$P(z \geq 1.15) = 1.000 - P(z \leq 1.15) = 1.000 - 0.8749 = 0.1251$$

Alternatively,

$$P(z \geq 1.15) = P(z \leq -1.15) = 0.1251$$

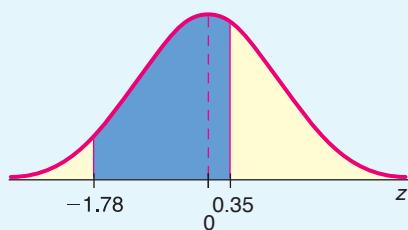
Continued

GUIDED EXERCISE 6 *continued*

- (b) Find $P(-1.78 \leq z \leq 0.35)$.

First, sketch the area under the standard normal curve corresponding to the area (Figure 6-22).

FIGURE 6-22 Area to Be Found



$$\begin{aligned} P(-1.78 \leq z \leq 0.35) &= P(z \leq 0.35) - P(z \leq -1.78) \\ &= 0.6368 - 0.0375 = 0.5993 \end{aligned}$$

TECH NOTES

The TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab all provide cumulative areas under any normal distribution, including the standard normal. The Tech Note of Section 6.3 shows examples.

VIEWPOINT

Mighty Oaks from Little Acorns Grow!

Just how big is that acorn? What if we compare it with other acorns? Is that oak tree taller than an average oak tree? How does it compare with other oak trees? What do you mean, this oak tree has a larger geographic range? Compared with what? Answers to questions such as these can be given only if we resort to standardized statistical units. Can you compare a single oak tree with an entire forest of oak trees? The answer is yes, if you use standardized z scores. For more information about sizes of acorns, oak trees, and geographic locations, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to DASL, the Carnegie Mellon University Data and Story Library. From the DASL site, find Biology under Data Subjects, and select Acorns.

SECTION 6.2
PROBLEMS

In these problems, assume that all distributions are *normal*. In all problems in Chapter 6, *average* is always taken to be the arithmetic mean \bar{x} or μ .

1. | **Statistical Literacy** What does a standard score measure?
2. | **Statistical Literacy** Does a raw score less than the mean correspond to a positive or negative standard score? What about a raw score greater than the mean?
3. | **Statistical Literacy** What is the value of the standard score for the mean of a distribution?
4. | **Statistical Literacy** What are the values of the mean and standard deviation of a standard normal distribution?
5. | **Basic Computation: z Score and Raw Score** A normal distribution has $\mu = 30$ and $\sigma = 5$.
 - (a) Find the z score corresponding to $x = 25$.
 - (b) Find the z score corresponding to $x = 42$.
 - (c) Find the raw score corresponding to $z = -2$.
 - (d) Find the raw score corresponding to $z = 1.3$.

Comstock/Jupiter Images



6. **Basic Computation: z Score and Raw Score** A normal distribution has $\mu = 10$ and $\sigma = 2$.
- Find the z score corresponding to $x = 12$.
 - Find the z score corresponding to $x = 4$.
 - Find the raw score corresponding to $z = 1.5$.
 - Find the raw score corresponding to $z = -1.2$.
7. **Critical Thinking** Consider the following scores:
- Score of 40 from a distribution with mean 50 and standard deviation 10
 - Score of 45 from a distribution with mean 50 and standard deviation 5
- How do the two scores compare relative to their respective distributions?
8. **Critical Thinking** Raul received a score of 80 on a history test for which the class mean was 70 with standard deviation 10. He received a score of 75 on a biology test for which the class mean was 70 with standard deviation 2.5. On which test did he do better relative to the rest of the class?
9. **z Scores: First Aid Course** The college physical education department offered an advanced first aid course last semester. The scores on the comprehensive final exam were normally distributed, and the z scores for some of the students are shown below:
- | | | |
|--------------|------------|--------------|
| Robert, 1.10 | Juan, 1.70 | Susan, -2.00 |
| Joel, 0.00 | Jan, -0.80 | Linda, 1.60. |
- Which of these students scored above the mean?
 - Which of these students scored on the mean?
 - Which of these students scored below the mean?
 - If the mean score was $\mu = 150$ with standard deviation $\sigma = 20$, what was the final exam score for each student?
10. **z Scores: Fawns** Fawns between 1 and 5 months old in Mesa Verde National Park have a body weight that is approximately normally distributed with mean $\mu = 27.2$ kilograms and standard deviation $\sigma = 4.3$ kilograms (based on information from *The Mule Deer of Mesa Verde National Park*, by G. W. Mierau and J. L. Schmidt, Mesa Verde Museum Association). Let x be the weight of a fawn in kilograms. Convert each of the following x intervals to z intervals.
- $x < 30$
 - $19 < x$
 - $32 < x < 35$
- Convert each of the following z intervals to x intervals.
- $-2.17 < z$
 - $z < 1.28$
 - $-1.99 < z < 1.44$
- Interpretation** If a fawn weighs 14 kilograms, would you say it is an unusually small animal? Explain using z values and Figure 6-15.
 - Interpretation** If a fawn is unusually large, would you say that the z value for the weight of the fawn will be close to 0, -2, or 3? Explain.
11. **z Scores: Red Blood Cell Count** Let x = red blood cell (RBC) count in millions per cubic millimeter of whole blood. For healthy females, x has an approximately normal distribution with mean $\mu = 4.8$ and standard deviation $\sigma = 0.3$ (based on information from *Diagnostic Tests with Nursing Implications*, edited by S. Loeb, Springhouse Press). Convert each of the following x intervals to z intervals.
- $4.5 < x$
 - $x < 4.2$
 - $4.0 < x < 5.5$
- Convert each of the following z intervals to x intervals.
- $z < -1.44$
 - $1.28 < z$
 - $-2.25 < z < -1.00$
- Interpretation** If a female had an RBC count of 5.9 or higher, would that be considered unusually high? Explain using z values and Figure 6-15.
12. **Normal Curve: Tree Rings** Tree-ring dates were used extensively in archaeological studies at Burnt Mesa Pueblo (*Bandelier Archaeological Excavation Project: Summer 1989 Excavations at Burnt Mesa Pueblo*, edited by Kohler, Washington State University Department of Anthropology). At one site on the mesa, tree-ring

dates (for many samples) gave a mean date of $\mu_1 = \text{year } 1272$ with standard deviation $\sigma_1 = 35$ years. At a second, removed site, the tree-ring dates gave a mean of $\mu_2 = \text{year } 1122$ with standard deviation $\sigma_2 = 40$ years. Assume that both sites had dates that were approximately normally distributed. In the first area, an object was found and dated as $x_1 = \text{year } 1250$. In the second area, another object was found and dated as $x_2 = \text{year } 1234$.

- Convert both x_1 and x_2 to z values, and locate both of these values under the standard normal curve of Figure 6-15.
- Interpretation** Which of these two items is the more unusual as an archaeological find in its location?

Basic Computation: Finding Areas Under the Standard Normal Curve In Problems 13–30, sketch the areas under the standard normal curve over the indicated intervals, and find the specified areas.

- | | |
|---|---|
| 13. To the right of $z = 0$ | 14. To the left of $z = 0$ |
| 15. To the left of $z = -1.32$ | 16. To the left of $z = -0.47$ |
| 17. To the left of $z = 0.45$ | 18. To the left of $z = 0.72$ |
| 19. To the right of $z = 1.52$ | 20. To the right of $z = 0.15$ |
| 21. To the right of $z = -1.22$ | 22. To the right of $z = -2.17$ |
| 23. Between $z = 0$ and $z = 3.18$ | 24. Between $z = 0$ and $z = -1.93$ |
| 25. Between $z = -2.18$ and $z = 1.34$ | 26. Between $z = -1.40$ and $z = 2.03$ |
| 27. Between $z = 0.32$ and $z = 1.92$ | 28. Between $z = 1.42$ and $z = 2.17$ |
| 29. Between $z = -2.42$ and $z = -1.77$ | 30. Between $z = -1.98$ and $z = -0.03$ |

Basic Computation: Finding Probabilities In Problems 31–50, let z be a random variable with a standard normal distribution. Find the indicated probability, and shade the corresponding area under the standard normal curve.

- | | |
|----------------------------------|----------------------------------|
| 31. $P(z \leq 0)$ | 32. $P(z \geq 0)$ |
| 33. $P(z \leq -0.13)$ | 34. $P(z \leq -2.15)$ |
| 35. $P(z \leq 1.20)$ | 36. $P(z \leq 3.20)$ |
| 37. $P(z \geq 1.35)$ | 38. $P(z \geq 2.17)$ |
| 39. $P(z \geq -1.20)$ | 40. $P(z \geq -1.50)$ |
| 41. $P(-1.20 \leq z \leq 2.64)$ | 42. $P(-2.20 \leq z \leq 1.40)$ |
| 43. $P(-2.18 \leq z \leq -0.42)$ | 44. $P(-1.78 \leq z \leq -1.23)$ |
| 45. $P(0 \leq z \leq 1.62)$ | 46. $P(0 \leq z \leq 0.54)$ |
| 47. $P(-0.82 \leq z \leq 0)$ | 48. $P(-2.37 \leq z \leq 0)$ |
| 49. $P(-0.45 \leq z \leq 2.73)$ | 50. $P(-0.73 \leq z \leq 3.12)$ |

SECTION 6.3

Areas Under Any Normal Curve

FOCUS POINTS

- Compute the probability of “standardized events.”
- Find a z score from a given normal probability (inverse normal).
- Use the inverse normal to solve guarantee problems.

Normal Distribution Areas

In many applied situations, the original normal curve is not the standard normal curve. Generally, there will not be a table of areas available for the original normal curve. This does not mean that we cannot find the probability that a

measurement x will fall into an interval from a to b . What we must do is *convert* the original measurements x , a , and b to z values.

PROCEDURE



To compute a z score, we need to know the standard deviation of the distribution. But what if we don't know σ ? If we have a range of data values, we can estimate σ . Problems 31 to 34 show how this is done.

HOW TO WORK WITH NORMAL DISTRIBUTIONS

To find areas and probabilities for a random variable x that follows a normal distribution with mean μ and standard deviation σ , convert x values to z values using the formula

$$z = \frac{x - \mu}{\sigma}$$

Then use Table 5 of Appendix II to find corresponding areas and probabilities.

EXAMPLE 7

NORMAL DISTRIBUTION PROBABILITY

Let x have a normal distribution with $\mu = 10$ and $\sigma = 2$. Find the probability that an x value selected at random from this distribution is between 11 and 14. In symbols, find $P(11 \leq x \leq 14)$.

SOLUTION: Since probabilities correspond to areas under the distribution curve, we want to find the area under the x curve above the interval from $x = 11$ to $x = 14$. To do so, we will convert the x values to standard z values and then use Table 5 of Appendix II to find the corresponding area under the standard curve.

We use the formula

$$z = \frac{x - \mu}{\sigma}$$

to convert the given x interval to a z interval.

$$z_1 = \frac{11 - 10}{2} = 0.50 \quad (\text{Use } x = 11, \mu = 10, \sigma = 2.)$$

$$z_2 = \frac{14 - 10}{2} = 2.00 \quad (\text{Use } x = 14, \mu = 10, \sigma = 2.)$$

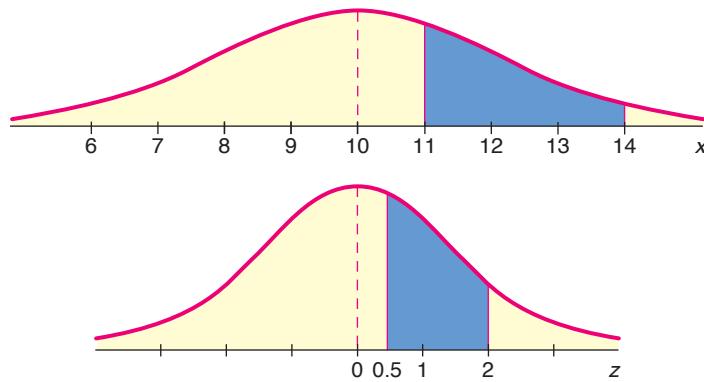
The corresponding areas under the x and z curves are shown in Figure 6-23. From Figure 6-23 we see that

$$\begin{aligned} P(11 \leq x \leq 14) &= P(0.50 \leq z \leq 2.00) \\ &= P(z \leq 2.00) - P(z \leq 0.50) \\ &= 0.9772 - 0.6915 \quad (\text{From Table 5, Appendix II}) \\ &= 0.2857 \end{aligned}$$

Interpretation The probability is 0.2857 that an x value selected at random from a normal distribution with mean 10 and standard deviation 2 lies between 11 and 14.

FIGURE 6-23

Corresponding Areas Under the x Curve and z Curve

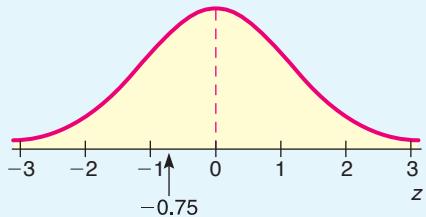


GUIDED EXERCISE 7**Normal distribution probability**

Sunshine Stereo cassette decks have a deck life that is normally distributed with a mean of 2.3 years and a standard deviation of 0.4 year. What is the probability that a cassette deck will break down during the guarantee period of 2 years?

- (a) Let x represent the life of a cassette deck. The statement that the cassette deck breaks during the 2-year guarantee period means the life is less than 2 years, or $x \leq 2$. Convert this to a statement about z .

- (b) Indicate the area to be found in Figure 6-24. Does this area correspond to the probability that $z \leq -0.75$?

FIGURE 6-24

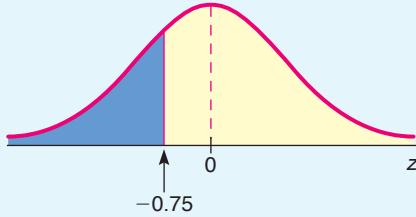
- (c) Use Table 5 of Appendix II to find $P(z \leq -0.75)$.

$$\Rightarrow z = \frac{x - \mu}{\sigma} = \frac{2 - 2.3}{0.4} = -0.75$$

So, $x \leq 2$ means $z \leq -0.75$.

See Figure 6-25.

Yes, the shaded area does correspond to the probability that $z \leq -0.75$.

FIGURE 6-25 $z \leq -0.75$ 

- (d) **Interpretation** What is the probability that the cassette deck will break before the end of the guarantee period?
[Hint: $P(x \leq 2) = P(z \leq -0.75)$.]

The probability is

$$\begin{aligned} P(x \leq 2) &= P(z \leq -0.75) \\ &= 0.2266 \end{aligned}$$

This means that the company will repair or replace about 23% of the cassette decks.

**TECH NOTES**

The TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab all provide areas under any normal distribution. Excel 2007 and Minitab give the left-tail area to the left of a specified x value. The TI-84Plus/TI-83Plus/TI-nspire has you specify an interval from a lower bound to an upper bound and provides the area under the normal curve for that interval. For example, to solve Guided Exercise 7 regarding the probability a cassette deck will break during the guarantee period, we find $P(x \leq 2)$ for a normal distribution with $\mu = 2.3$ and $\sigma = 0.4$.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Press the DISTR key, select 2: normalcdf (lower bound, upper bound, μ , σ) and press Enter. Type in the specified values. For a left-tail area, use a lower bound setting at about 4 standard deviations below the mean. Likewise, for a right-tail area, use an upper bound setting about 4 standard deviations above the mean. For our example, use a lower bound of $\mu - 4\sigma = 2.3 - 4(0.4) = 0.7$.

```
normalcdf(.7, 2, 2.3,
          .4)
.2265955934
```

Excel 2007 Select **Insert Function** f_x In the dialogue box, select **Statistical** for the Category and then for the Function, select **NORMDIST**. Fill in the dialogue box using True for cumulative.

f_x	=NORMDIST(2.2,3,0.4, TRUE)
C	D
0.226627	



What if we need to compute conditional probabilities based on the normal distribution? Problems 39 and 40 of this section show how we can do this.

Minitab Use the menu selection **Calc** \blacktriangleright **Probability Distribution** \blacktriangleright **Normal**. Fill in the dialogue box, marking cumulative.

Cumulative Distribution Function

Normal with mean = 2.3 and standard deviation = 0.4

x	$P(X \leq x)$
2.0	0.2266



Inverse Normal Distribution

Finding z or x, given a probability

Inverse normal probability distribution

EXAMPLE 8

FIND x, GIVEN PROBABILITY

Magic Video Games, Inc., sells an expensive video games package. Because the package is so expensive, the company wants to advertise an impressive guarantee for the life expectancy of its computer control system. The guarantee policy will refund the full purchase price if the computer fails during the guarantee period. The research department has done tests that show that the mean life for the computer is 30 months, with standard deviation of 4 months. The computer life is normally distributed. How long can the guarantee period be if management does not want to refund the purchase price on more than 7% of the Magic Video packages?

SOLUTION: Let us look at the distribution of lifetimes for the computer control system, and shade the portion of the distribution in which the computer lasts fewer months than the guarantee period. (See Figure 6-26.)

FIGURE 6-26

7% of the Computers Have a Lifetime Less Than the Guarantee Period

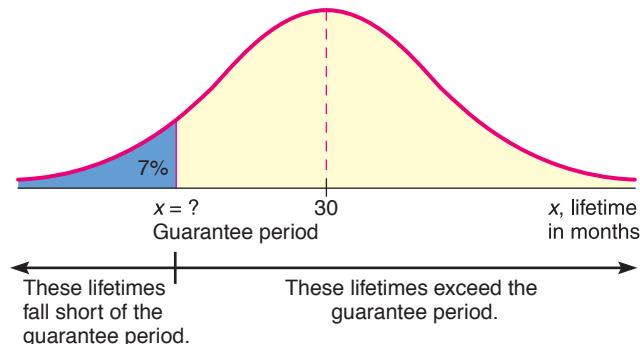


TABLE 6-5 Excerpt from Table 5 of Appendix II

z	.0007	.08	.09
:					
-1.4	.0808		.0708	.0694	.0681
				↑ 0.0700	

If a computer system lasts fewer months than the guarantee period, a full-price refund will have to be made. The lifetimes requiring a refund are in the shaded region in Figure 6-26. This region represents 7% of the total area under the curve.

We can use Table 5 of Appendix II to find the z value such that 7% of the total area under the *standard* normal curve lies to the left of the z value. Then we convert the z value to its corresponding x value to find the guarantee period.

We want to find the z value with 7% of the area under the standard normal curve to the left of z . Since we are given the area in a left tail, we can use Table 5 of Appendix II directly to find z . The area value is 0.0700. However, this area is not in our table, so we use the closest area, which is 0.0694, and the corresponding z value of $z = -1.48$ (see Table 6-5).

To translate this value back to an x value (in months), we use the formula

$$\begin{aligned}x &= z\sigma + \mu \\&= -1.48(4) + 30 \quad (\text{Use } \sigma = 4 \text{ months and } \mu = 30 \text{ months.}) \\&= 24.08 \text{ months}\end{aligned}$$

Interpretation The company can guarantee the Magic Video Games package for $x = 24$ months. For this guarantee period, they expect to refund the purchase price of no more than 7% of the video games packages.

Pierre Arsenault/Alamy

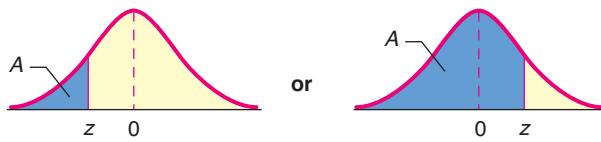


FIGURE 6-27

Inverse Normal: Use Table 5 of Appendix II to Find z Corresponding to a Given Area A ($0 < A < 1$)

(a) **Left-tail case:**

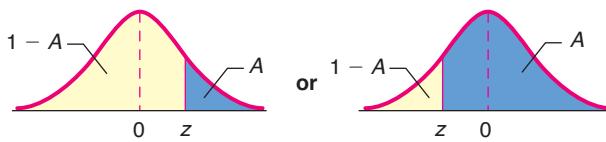
The given area A is to the left of z .



For the left-tail case, look up the number A in the body of the table and use the corresponding z value.

(b) **Right-tail case:**

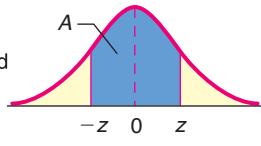
The given area A is to the right of z .



For the right-tail case, look up the number $1 - A$ in the body of the table and use the corresponding z value.

(c) **Center case:**

The given area A is symmetric and centered above $z = 0$. Half of A lies to the left and half lies to the right of $z = 0$.



For the center case, look up the number $\frac{1 - A}{2}$ in the body of the table and use the corresponding $\pm z$ value.

COMMENT When we use Table 5 of Appendix II to find a z value corresponding to a given area, we usually use the nearest area value rather than interpolating between values. However, when the area value given is exactly halfway between two area values of the table, we use the z value halfway between the z values of the corresponding table areas. Example 9 demonstrates this procedure. However, this interpolation convention is not always used, especially if the area is changing slowly, as it does in the tail ends of the distribution. *When the z value corresponding to an area is smaller than -2 , the standard convention is to use the z value corresponding to the smaller area. Likewise, when the z value is larger than 2 , the standard convention is to use the z value corresponding to the larger area.* We will see an example of this special case in Guided Exercise 9.

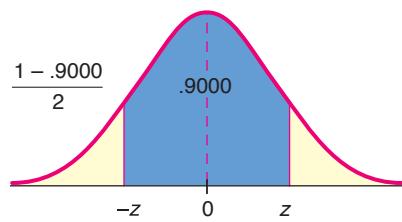
EXAMPLE 9

FIND z

Find the z value such that 90% of the area under the standard normal curve lies between $-z$ and z .

SOLUTION: Sketch a picture showing the described area (see Figure 6-28).

FIGURE 6-28
Area Between $-z$ and z Is 90%



We find the corresponding area in the left tail.

$$\begin{aligned} (\text{Area left of } -z) &= \frac{1 - 0.9000}{2} \\ &= 0.0500 \end{aligned}$$

Looking in Table 6-6, we see that 0.0500 lies exactly between areas 0.0495 and 0.0505. The halfway value between $z = -1.65$ and $z = -1.64$ is $z = -1.645$. Therefore, we conclude that 90% of the area under the standard normal curve lies between the z values -1.645 and 1.645 .

TABLE 6-6 Excerpt from Table 5 of Appendix II

z04	.05
:			
-1.6		.0505	.0495
		↑ 0.0500	

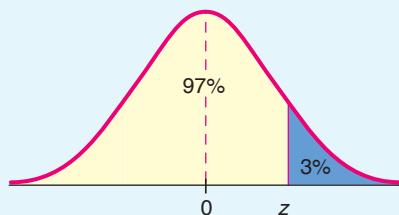
GUIDED EXERCISE 8**Find z**

Find the z value such that 3% of the area under the standard normal curve lies to the right of z .

- (a) Draw a sketch of the standard normal distribution showing the described area (Figure 6-29).



FIGURE 6-29 3% of Total Area Lies to the Right of z



- (b) Find the area to the left of z .



$$\text{Area to the left of } z = 1 - 0.0300 = 0.9700.$$

- (c) Look up the area in Table 6-7 and find the corresponding z .



The closest area is 0.9699. This area is to the left of $z = 1.88$.

TABLE 6-7 Excerpt from Table 5 of Appendix II

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

- (d) Suppose the time to complete a test is normally distributed with $\mu = 40$ minutes and $\sigma = 5$ minutes. After how many minutes can we expect all but about 3% of the tests to be completed?



We are looking for an x value such that 3% of the normal distribution lies to the right of x . In part (c), we found that 3% of the standard normal curve lies to the right of $z = 1.88$. We convert $z = 1.88$ to an x value.

$$\begin{aligned}x &= z\sigma + \mu \\&= 1.88(5) + 40 = 49.4 \text{ minutes}\end{aligned}$$

All but about 3% of the tests will be complete after 50 minutes.

- (e) Use Table 6-8 to find a z value such that 3% of the area under the standard normal curve lies to the left of z .



The closest area is 0.0301. This is the area to the left of $z = -1.88$.

TABLE 6-8 Excerpt from Table 5 of Appendix II

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294

- (f) Compare the z value of part (c) with the z value of part (e). Is there any relationship between the z values?



One z value is the negative of the other. This result is expected because of the symmetry of the normal distribution.

**TECH NOTES**

When we are given a z value and we find an area to the left of z , we are using a normal distribution function. When we are given an area to the left of z and we find the corresponding z , we are using an inverse normal distribution function. The

TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab all have inverse normal distribution functions for any normal distribution. For instance, to find an x value from a normal distribution with mean 40 and standard deviation 5 such that 97% of the area lies to the left of x , use the described instructions.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Press the DISTR key and select 3:invNorm(area, μ , σ).

invNorm(.97,40,5)
49.40396805

Excel 2007 Select Insert Function f_x . In the dialogue box, select Statistical for the Category and then for the Function, select NORMINV. Fill in the dialogue box.

f_x	=NORMINV(0.97,40,5)
C	D
49.40395	

Minitab Use the menu selection Calc \blacktriangleright Probability Distribution \blacktriangleright Normal. Fill in the dialogue box, marking Inverse Cumulative.

Inverse Cumulative Distribution Function
 Normal with mean = 40.000 and
 standard deviation = 5.00000
 $P(X \leq x) \quad x$
 0.9700 49.4040

LOOKING FORWARD

In our work with confidence intervals (Chapter 7), we will use inverse probability distributions for the normal distribution and for the Student's t distribution (a similar distribution introduced in Chapter 7). Just as in Example 9, we'll use inverse probability distributions to identify values such that 90%, 95%, or 99% of the area under the distribution graph centered over the mean falls between the values.



Checking for Normality

How can we tell if data follow a normal distribution? There are several checks we can make. The following procedure lists some guidelines.

PROCEDURE

HOW TO DETERMINE WHETHER DATA HAVE A NORMAL DISTRIBUTION

The following guidelines represent some useful devices for determining whether or not data follow a normal distribution.

1. **Histogram:** Make a histogram. For a normal distribution, the histogram should be roughly bell-shaped.

Continued

2. **Outliers:** For a normal distribution, there should not be more than one outlier. One way to check for outliers is to use a box-and-whisker plot. Recall that outliers are those data values that are

above Q_3 by an amount greater than $1.5 \times$ interquartile range

below Q_1 by an amount greater than $1.5 \times$ interquartile range

3. **Skewness:** Normal distributions are symmetric. One measure of skewness for sample data is given by Pearson's index:

$$\text{Pearson's index} = \frac{3(\bar{x} - \text{median})}{s}$$

An index value greater than 1 or less than -1 indicates skewness.

Skewed distributions are not normal.

4. **Normal quantile plot (or normal probability plot):** This plot is provided through statistical software on a computer or graphing calculator. The Using Technology feature, Application 1, at the end of this chapter gives a brief description of how such plots are constructed. The section also gives commands for producing such plots on the TI-84Plus/TI-83Plus/TI-nspire calculators, Minitab, or SPSS.

Examine a normal quantile plot of the data.

If the points lie close to a straight line, the data come from a distribution that is approximately normal.

If the points do not lie close to a straight line or if they show a pattern that is not a straight line, the data are likely to come from a distribution that is not normal.

EXAMPLE 10

ASSESSING NORMALITY

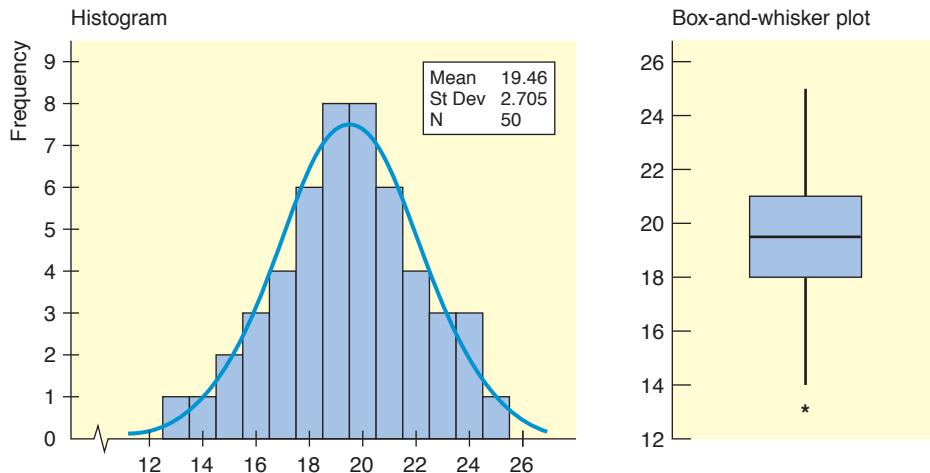
Consider the following data, which are rounded to the nearest integer.

19	19	19	16	21	14	23	17	19	20	18	24	20	13	16
17	19	18	19	17	21	24	18	23	19	21	22	20	20	20
24	17	20	22	19	22	21	18	20	22	16	15	21	23	21
18	18	20	15	25										

- (a) Look at the histogram and box-and-whisker plot generated by Minitab in Figure 6-30 and comment about normality of the data from these indicators.

FIGURE 6-30

Histogram and Box-and-Whisker Plot



SOLUTION: Note that the histogram is approximately normal. The box-and-whisker plot shows just one outlier. Both of these graphs indicate normality.

(b) Use Pearson's index to check for skewness.

SOLUTION: Summary statistics from Minitab:

Variable	N	N*	Mean	Se Mean	StDev	Minimum	Q1	Median	Q3
C2	50	0	19.460	0.382	2.705	13.000	18.000	19.500	21.000
Variable									
C2			Maximum			25.000			

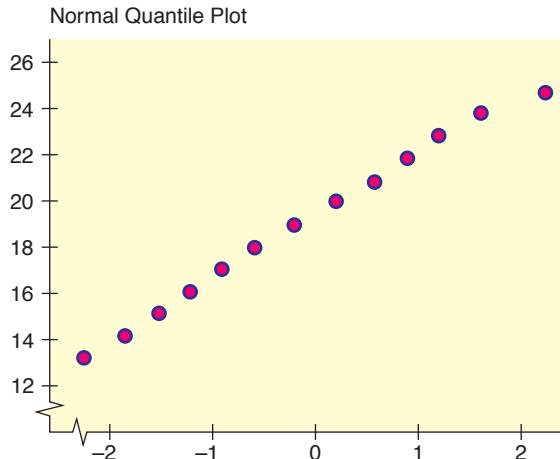
We see that $\bar{x} = 19.46$, median = 19.5, and $s = 2.705$.

$$\text{Pearson's index} = \frac{3(19.46 - 19.5)}{2.705} \approx -0.04$$

Since the index is between -1 and 1 , we detect no skewness. The data appear to be symmetric.

(c) Look at the normal quantile plot in Figure 6-31 and comment on normality.

FIGURE 6-31
Normal Quantile Plot



SOLUTION: The data fall close to a straight line, so the data appear to come from a normal distribution.

(d) *Interpretation* Interpret the results.

SOLUTION: The histogram is roughly bell-shaped, there is only one outlier, Pearson's index does not indicate skewness, and the points on the normal quantile plot lie fairly close to a straight line. It appears that the data are likely from a distribution that is approximately normal.

VIEWPOINT

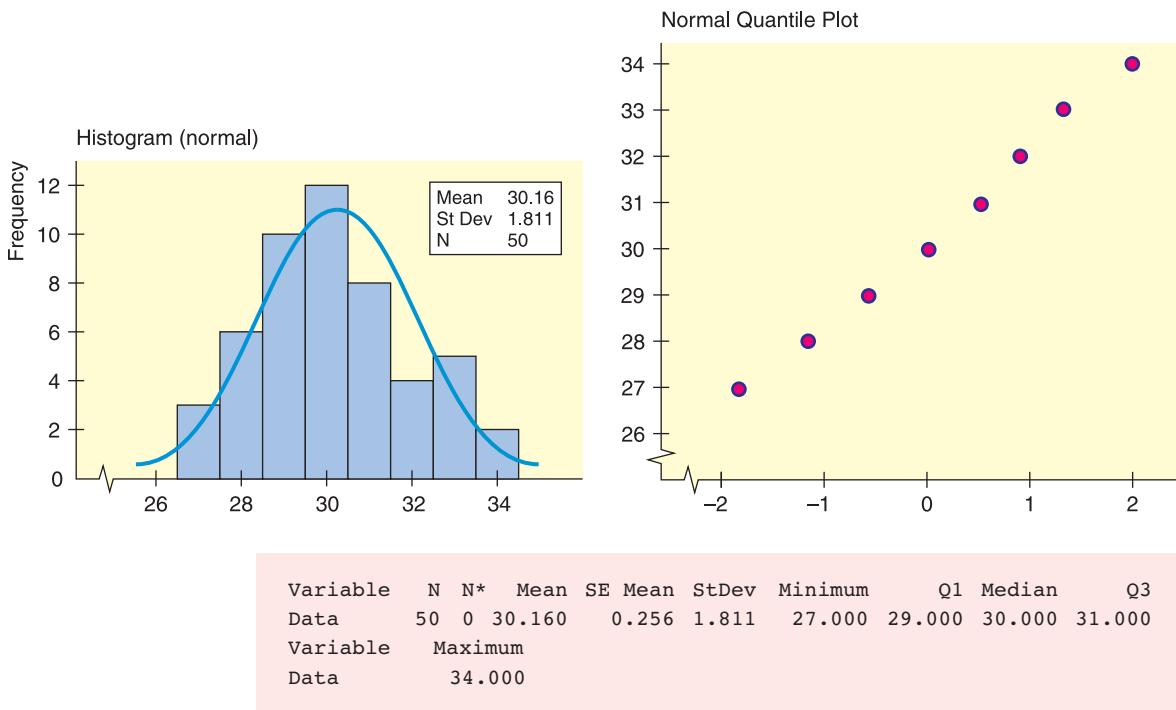
Want to Be an Archaeologist?

Each year about 4500 students work with professional archaeologists in scientific research at the Crow Canyon Archaeological Center, Cortez, Colorado. In fact, Crow Canyon was included in The Princeton Review Guide to America's Top 100 Internships. The nonprofit, multidisciplinary program at Crow Canyon enables students and laypeople with little or no archaeology background to get started in archaeological research. The only requirement is that you be interested in Native American culture and history. By the way, a knowledge of introductory statistics could come in handy for this internship. For more information about the program, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to Crow Canyon.

SECTION 6.3 PROBLEMS

- Statistical Literacy** Consider a normal distribution with mean 30 and standard deviation 2. What is the probability a value selected at random from this distribution is greater than 30?
- Statistical Literacy** Suppose 5% of the area under the standard normal curve lies to the right of z . Is z positive or negative?
- Statistical Literacy** Suppose 5% of the area under the standard normal curve lies to the left of z . Is z positive or negative?
- Critical Thinking: Normality** Consider the following data. The summary statistics, histogram, and normal quantile plot were generated by Minitab.

27 27 27 28 28 28 28 28 29 29 29 29 29 29
 29 29 29 29 30 30 30 30 30 30 30 30 30 30
 30 31 31 31 31 31 31 31 31 32 32 32 32 33 33
 33 33 33 34 34



- Does the histogram indicate normality for the data distribution? Explain.
- Does the normal quantile plot indicate normality for the data distribution? Explain.
- Compute the interquartile range and check for outliers.
- Compute Pearson's index. Does the index value indicate skewness?
- Using parts (a) through (d), would you say the data are from a normal distribution?

Basic Computation: Find Probabilities In Problems 5–14, assume that x has a normal distribution with the specified mean and standard deviation. Find the indicated probabilities.

- $P(3 \leq x \leq 6); \mu = 4; \sigma = 2$
- $P(10 \leq x \leq 26); \mu = 15; \sigma = 4$
- $P(50 \leq x \leq 70); \mu = 40; \sigma = 15$
- $P(7 \leq x \leq 9); \mu = 5; \sigma = 1.2$
- $P(8 \leq x \leq 12); \mu = 15; \sigma = 3.2$
- $P(40 \leq x \leq 47); \mu = 50; \sigma = 15$
- $P(x \geq 30); \mu = 20; \sigma = 3.4$
- $P(x \geq 120); \mu = 100; \sigma = 15$
- $P(x \geq 90); \mu = 100; \sigma = 15$
- $P(x \geq 2); \mu = 3; \sigma = 0.25$

Basic Computation: Find z Values In Problems 15–24, find the z value described and sketch the area described.

15. Find z such that 6% of the standard normal curve lies to the left of z .
16. Find z such that 5.2% of the standard normal curve lies to the left of z .
17. Find z such that 55% of the standard normal curve lies to the left of z .
18. Find z such that 97.5% of the standard normal curve lies to the left of z .
19. Find z such that 8% of the standard normal curve lies to the right of z .
20. Find z such that 5% of the standard normal curve lies to the right of z .
21. Find z such that 82% of the standard normal curve lies to the right of z .
22. Find z such that 95% of the standard normal curve lies to the right of z .
23. Find the z value such that 98% of the standard normal curve lies between $-z$ and z .
24. Find the z value such that 95% of the standard normal curve lies between $-z$ and z .
25. **Medical: Blood Glucose** A person's blood glucose level and diabetes are closely related. Let x be a random variable measured in milligrams of glucose per deciliter (1/10 of a liter) of blood. After a 12-hour fast, the random variable x will have a distribution that is approximately normal with mean $\mu = 85$ and standard deviation $\sigma = 25$ (Source: *Diagnostic Tests with Nursing Implications*, edited by S. Loeb, Springhouse Press). Note: After 50 years of age, both the mean and standard deviation tend to increase. What is the probability that, for an adult (under 50 years old) after a 12-hour fast,
 - (a) x is more than 60?
 - (b) x is less than 110?
 - (c) x is between 60 and 110?
 - (d) x is greater than 140 (borderline diabetes starts at 140)?
26. **Medical: Blood Protoplasm** Porphyrin is a pigment in blood protoplasm and other body fluids that is significant in body energy and storage. Let x be a random variable that represents the number of milligrams of porphyrin per deciliter of blood. In healthy adults, x is approximately normally distributed with mean $\mu = 38$ and standard deviation $\sigma = 12$ (see reference in Problem 25). What is the probability that
 - (a) x is less than 60?
 - (b) x is greater than 16?
 - (c) x is between 16 and 60?
 - (d) x is more than 60? (This may indicate an infection, anemia, or another type of illness.)
27. **Archaeology: Hopi Village** Thickness measurements of ancient prehistoric Native American pot shards discovered in a Hopi village are approximately normally distributed, with a mean of 5.1 millimeters (mm) and a standard deviation of 0.9 mm (Source: *Homol'ovi II: Archaeology of an Ancestral Hopi Village, Arizona*, edited by E. C. Adams and K. A. Hays, University of Arizona Press). For a randomly found shard, what is the probability that the thickness is
 - (a) less than 3.0 mm?
 - (b) more than 7.0 mm?
 - (c) between 3.0 mm and 7.0 mm?
28. **Law Enforcement: Police Response Time** Police response time to an emergency call is the difference between the time the call is first received by the dispatcher and the time a patrol car radios that it has arrived at the scene (based on information from *The Denver Post*). Over a long period of time, it has been determined that the police response time has a normal distribution with a mean of 8.4 minutes and a standard deviation of 1.7 minutes. For a

randomly received emergency call, what is the probability that the response time will be

- (a) between 5 and 10 minutes?
- (b) less than 5 minutes?
- (c) more than 10 minutes?

29. **Guarantee: Batteries** Quick Start Company makes 12-volt car batteries. After many years of product testing, the company knows that the average life of a Quick Start battery is normally distributed, with a mean of 45 months and a standard deviation of 8 months.

- (a) If Quick Start guarantees a full refund on any battery that fails within the 36-month period after purchase, what percentage of its batteries will the company expect to replace?
- (b) **Inverse Normal Distribution** If Quick Start does not want to make refunds for more than 10% of its batteries under the full-refund guarantee policy, for how long should the company guarantee the batteries (to the nearest month)?

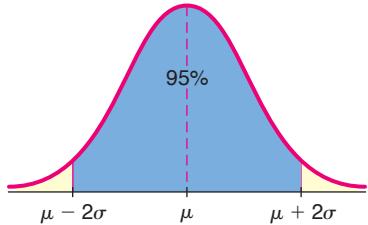
30. **Guarantee: Watches** Accrotime is a manufacturer of quartz crystal watches. Accrotime researchers have shown that the watches have an average life of 28 months before certain electronic components deteriorate, causing the watch to become unreliable. The standard deviation of watch lifetimes is 5 months, and the distribution of lifetimes is normal.

- (a) If Accrotime guarantees a full refund on any defective watch for 2 years after purchase, what percentage of total production should the company expect to replace?
- (b) **Inverse Normal Distribution** If Accrotime does not want to make refunds on more than 12% of the watches it makes, how long should the guarantee period be (to the nearest month)?



- 31. Expand Your Knowledge: Estimating the Standard Deviation** Consumer Reports gave information about the ages at which various household products are replaced. For example, color TVs are replaced at an average age of $\mu = 8$ years after purchase, and the (95% of data) range was from 5 to 11 years. Thus, the range was $11 - 5 = 6$ years. Let x be the age (in years) at which a color TV is replaced. Assume that x has a distribution that is approximately normal.

- (a) The empirical rule (Section 6.1) indicates that for a symmetrical and bell-shaped distribution, approximately 95% of the data lies within two standard deviations of the mean. Therefore, a 95% range of data values extending from $\mu - 2\sigma$ to $\mu + 2\sigma$ is often used for “commonly occurring” data values. Note that the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$ is 4σ in length. This leads to a “rule of thumb” for estimating the standard deviation from a 95% range of data values.



Estimating the standard deviation

For a symmetric, bell-shaped distribution,

$$\text{standard deviation} \approx \frac{\text{range}}{4} \approx \frac{\text{high value} - \text{low value}}{4}$$

where it is estimated that about 95% of the commonly occurring data values fall into this range.

Use this “rule of thumb” to approximate the standard deviation of x values, where x is the age (in years) at which a color TV is replaced.

- (b) What is the probability that someone will keep a color TV more than 5 years before replacement?

- (c) What is the probability that someone will keep a color TV fewer than 10 years before replacement?

(d) **Inverse Normal Distribution** Assume that the average life of a color TV is 8 years with a standard deviation of 1.5 years before it breaks. Suppose that a company guarantees color TVs and will replace a TV that breaks while under guarantee with a new one. However, the company does not want to replace more than 10% of the TVs under guarantee. For how long should the guarantee be made (rounded to the nearest tenth of a year)?



32. **Estimating the Standard Deviation: Refrigerator Replacement** *Consumer Reports* indicated that the average life of a refrigerator before replacement is $\mu = 14$ years with a (95% of data) range from 9 to 19 years. Let x = age at which a refrigerator is replaced. Assume that x has a distribution that is approximately normal.

(a) Find a good approximation for the standard deviation of x values. Hint: See Problem 31.

(b) What is the probability that someone will keep a refrigerator fewer than 11 years before replacement?

(c) What is the probability that someone will keep a refrigerator more than 18 years before replacement?

(d) **Inverse Normal Distribution** Assume that the average life of a refrigerator is 14 years, with the standard deviation given in part (a) before it breaks. Suppose that a company guarantees refrigerators and will replace a refrigerator that breaks while under guarantee with a new one. However, the company does not want to replace more than 5% of the refrigerators under guarantee. For how long should the guarantee be made (rounded to the nearest tenth of a year)?



33. **Estimating the Standard Deviation: Veterinary Science** The resting heart rate for an adult horse should average about $\mu = 46$ beats per minute with a (95% of data) range from 22 to 70 beats per minute, based on information from the *Merck Veterinary Manual* (a classic reference used in most veterinary colleges). Let x be a random variable that represents the resting heart rate for an adult horse. Assume that x has a distribution that is approximately normal.

(a) Estimate the standard deviation of the x distribution. Hint: See Problem 31.

(b) What is the probability that the heart rate is fewer than 25 beats per minute?

(c) What is the probability that the heart rate is greater than 60 beats per minute?

(d) What is the probability that the heart rate is between 25 and 60 beats per minute?

(e) **Inverse Normal Distribution** A horse whose resting heart rate is in the upper 10% of the probability distribution of heart rates may have a secondary infection or illness that needs to be treated. What is the heart rate corresponding to the upper 10% cutoff point of the probability distribution?



34. **Estimating the Standard Deviation: Veterinary Science** How much should a healthy kitten weigh? A healthy 10-week-old (domestic) kitten should weigh an average of $\mu = 24.5$ ounces with a (95% of data) range from 14 to 35 ounces. (See reference in Problem 33.) Let x be a random variable that represents the weight (in ounces) of a healthy 10-week-old kitten. Assume that x has a distribution that is approximately normal.

(a) Estimate the standard deviation of the x distribution. Hint: See Problem 31.

(b) What is the probability that a healthy 10-week-old kitten will weigh less than 14 ounces?



- (c) What is the probability that a healthy 10-week-old kitten will weigh more than 33 ounces?
- (d) What is the probability that a healthy 10-week-old kitten will weigh between 14 and 33 ounces?
- (e) **Inverse Normal Distribution** A kitten whose weight is in the bottom 10% of the probability distribution of weights is called *undernourished*. What is the cutoff point for the weight of an undernourished kitten?
35. **Insurance: Satellites** A relay microchip in a telecommunications satellite has a life expectancy that follows a normal distribution with a mean of 90 months and a standard deviation of 3.7 months. When this computer-relay microchip malfunctions, the entire satellite is useless. A large London insurance company is going to insure the satellite for \$50 million. Assume that the only part of the satellite in question is the microchip. All other components will work indefinitely.
- (a) **Inverse Normal Distribution** For how many months should the satellite be insured to be 99% confident that it will last beyond the insurance date?
- (b) If the satellite is insured for 84 months, what is the probability that it will malfunction before the insurance coverage ends?
- (c) If the satellite is insured for 84 months, what is the expected loss to the insurance company?
- (d) If the insurance company charges \$3 million for 84 months of insurance, how much profit does the company expect to make?
36. **Conversion Center: Exhibition Show Attendance** Attendance at large exhibition shows in Denver averages about 8000 people per day, with standard deviation of about 500. Assume that the daily attendance figures follow a normal distribution.
- (a) What is the probability that the daily attendance will be fewer than 7200 people?
- (b) What is the probability that the daily attendance will be more than 8900 people?
- (c) What is the probability that the daily attendance will be between 7200 and 8900 people?
37. **Exhibition Shows: Inverse Normal Distribution** Most exhibition shows open in the morning and close in the late evening. A study of Saturday arrival times showed that the average arrival time was 3 hours and 48 minutes after the doors opened, and the standard deviation was estimated at about 52 minutes. Assume that the arrival times follow a normal distribution.
- (a) At what time after the doors open will 90% of the people who are coming to the Saturday show have arrived?
- (b) At what time after the doors open will only 15% of the people who are coming to the Saturday show have arrived?
- (c) Do you think the probability distribution of arrival times for Friday might be different from the distribution of arrival times for Saturday? Explain.
38. **Budget: Maintenance** The amount of money spent weekly on cleaning, maintenance, and repairs at a large restaurant was observed over a long period of time to be approximately normally distributed, with mean $\mu = \$615$ and standard deviation $\sigma = \$42$.
- (a) If \$646 is budgeted for next week, what is the probability that the actual costs will exceed the budgeted amount?
- (b) **Inverse Normal Distribution** How much should be budgeted for weekly repairs, cleaning, and maintenance so that the probability that the budgeted amount will be exceeded in a given week is only 0.10?



39. **Expand Your Knowledge: Conditional Probability** Suppose you want to eat lunch at a popular restaurant. The restaurant does not take reservations, so there is usually a waiting time before you can be seated. Let x represent the length of time waiting to be seated. From past experience, you know that the mean waiting time is $\mu = 18$ minutes with $\sigma = 4$ minutes. You assume that the x distribution is approximately normal.

- What is the probability that the waiting time will exceed 20 minutes, given that it has exceeded 15 minutes? Hint: Compute $P(x > 20 | x > 15)$.
- What is the probability that the waiting time will exceed 25 minutes, given that it has exceeded 18 minutes? Hint: Compute $P(x > 25 | x > 18)$.
- Hint for solution:* Review item 6, conditional probability, in the summary of basic probability rules at the end of Section 4.2. Note that

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

and show that in part (a),

$$P(x > 20 | x > 15) = \frac{P((x > 20) \text{ and } (x > 15))}{P(x > 15)} = \frac{P(x > 20)}{P(x > 15)}$$



40. **Conditional Probability: Cycle Time** A cement truck delivers mixed cement to a large construction site. Let x represent the cycle time in minutes for the truck to leave the construction site, go back to the cement plant, fill up, and return to the construction site with another load of cement. From past experience, it is known that the mean cycle time is $\mu = 45$ minutes with $\sigma = 12$ minutes. The x distribution is approximately normal.

- What is the probability that the cycle time will exceed 60 minutes, given that it has exceeded 50 minutes? Hint: See Problem 39, part (c).
- What is the probability that the cycle time will exceed 55 minutes, given that it has exceeded 40 minutes?

AP Photo/Judi Bottoni



SECTION 6.4

Sampling Distributions

FOCUS POINTS

- Review such commonly used terms as *random sample*, *relative frequency*, *parameter*, *statistic*, and *sampling distribution*.
- From raw data, construct a relative frequency distribution for \bar{x} values and compare the result to a theoretical sampling distribution.

Let us begin with some common statistical terms. Most of these have been discussed before, but this is a good time to review them.

From a statistical point of view, a *population* can be thought of as a complete set of measurements (or counts), either existing or conceptual. We discussed populations at some length in Chapter 1. A *sample* is a subset of measurements from the population. For our purposes, the most important samples are *random samples*, which were discussed in Section 1.2.

When we compute a descriptive measure such as an average, it makes a difference whether it was computed from a population or from a sample.

Statistic
Parameter

A **statistic** is a numerical descriptive measure of a *sample*.
A **parameter** is a numerical descriptive measure of a *population*.

It is important to notice that for a given population, a specified parameter is a fixed quantity. On the other hand, the value of a statistic might vary depending on which sample has been selected.

Population parameter**Some commonly used statistics and corresponding parameters**

Measure	Statistic	Parameter
Mean	\bar{x} (x bar)	μ (mu)
Variance	s^2	σ^2 (sigma squared)
Standard deviation	s	σ (sigma)
Proportion	\hat{p} (p hat)	p

Often we do not have access to all the measurements of an entire population because of constraints on time, money, or effort. So, we must use measurements from a sample instead. In such cases, we will use a statistic (such as \bar{x} , s , or \hat{p}) to make *inferences* about a corresponding *population parameter* (e.g., μ , σ , or p). The principal types of inferences we will make are the following.

Types of inferences

1. **Estimation:** In this type of inference, we estimate the *value* of a population parameter.
2. **Testing:** In this type of inference, we formulate a *decision* about the value of a population parameter.
3. **Regression:** In this type of inference, we make *predictions* or *forecasts* about the value of a statistical variable.

Sampling distribution

To evaluate the reliability of our inferences, we will need to know the probability distribution for the statistic we are using. Such a probability distribution is called a *sampling distribution*. Perhaps Example 11 below will help clarify this discussion.

A **sampling distribution** is a probability distribution of a sample statistic based on all possible simple random samples of the *same* size from the same population.

EXAMPLE 11**SAMPLING DISTRIBUTION FOR \bar{x}**

Pinedale, Wisconsin, is a rural community with a children's fishing pond. Posted rules state that all fish under 6 inches must be returned to the pond, only children under 12 years old may fish, and a limit of five fish may be kept per day. Susan is a college student who was hired by the community last summer to make sure the rules were obeyed and to see that the children were safe from accidents. The pond contains only rainbow trout and has been well stocked for many years. Each child has no difficulty catching his or her limit of five trout.

As a project for her biometrics class, Susan kept a record of the lengths (to the nearest inch) of all trout caught last summer. Hundreds of children visited the pond and caught their limit of five trout, so Susan has a lot of data. To make Table 6-9, Susan selected 100 children at random and listed the lengths of each of the five trout caught by a child in the sample. Then, for each child, she listed the mean length of the five trout that child caught.

Now let us turn our attention to the following question: What is the average (mean) length of a trout taken from the Pinedale children's pond last summer?

SOLUTION: We can get an idea of the average length by looking at the far-right column of Table 6-9. But just looking at 100 of the \bar{x} values doesn't tell us much.



TABLE 6-9 Length Measurements of Trout Caught by a Random Sample of 100 Children at the Pinedale Children's Pond

Sample	\bar{x} = Sample Mean						Sample	\bar{x} = Sample Mean					
1	11	10	10	12	11	10.8	51	9	10	12	10	9	10.0
2	11	11	9	9	9	9.8	52	7	11	10	11	10	9.8
3	12	9	10	11	10	10.4	53	9	11	9	11	12	10.4
4	11	10	13	11	8	10.6	54	12	9	8	10	11	10.0
5	10	10	13	11	12	11.2	55	8	11	10	9	10	9.6
6	12	7	10	9	11	9.8	56	10	10	9	9	13	10.2
7	7	10	13	10	10	10.0	57	9	8	10	10	12	9.8
8	10	9	9	9	10	9.4	58	10	11	9	8	9	9.4
9	10	10	11	12	8	10.2	59	10	8	9	10	12	9.8
10	10	11	10	7	9	9.4	60	11	9	9	11	11	10.2
11	12	11	11	11	13	11.6	61	11	10	11	10	11	10.6
12	10	11	10	12	13	11.2	62	12	10	10	9	11	10.4
13	11	10	10	9	11	10.2	63	10	10	9	11	7	9.4
14	10	10	13	8	11	10.4	64	11	11	12	10	11	11.0
15	9	11	9	10	10	9.8	65	10	10	11	10	9	10.0
16	13	9	11	12	10	11.0	66	8	9	10	11	11	9.8
17	8	9	7	10	11	9.0	67	9	11	11	9	8	9.6
18	12	12	8	12	12	11.2	68	10	9	10	9	11	9.8
19	10	8	9	10	10	9.4	69	9	9	11	11	11	10.2
20	10	11	10	10	10	10.2	70	13	11	11	9	11	11.0
21	11	10	11	9	12	10.6	71	12	10	8	8	9	9.4
22	9	12	9	10	9	9.8	72	13	7	12	9	10	10.2
23	8	11	10	11	10	10.0	73	9	10	9	8	9	9.0
24	9	12	10	9	11	10.2	74	11	11	10	9	10	10.2
25	9	9	8	9	10	9.0	75	9	11	14	9	11	10.8
26	11	11	12	11	11	11.2	76	14	10	11	12	12	11.8
27	10	10	10	11	13	10.8	77	8	12	10	10	9	9.8
28	8	7	9	10	8	8.4	78	8	10	13	9	8	9.6
29	11	11	8	10	11	10.2	79	11	11	11	13	10	11.2
30	8	11	11	9	12	10.2	80	12	10	11	12	9	10.8
31	11	9	12	10	10	10.4	81	10	9	10	10	13	10.4
32	10	11	10	11	12	10.8	82	11	10	9	9	12	10.2
33	12	11	8	8	11	10.0	83	11	11	10	10	10	10.4
34	8	10	10	9	10	9.4	84	11	10	11	9	9	10.0
35	10	10	10	10	11	10.2	85	10	11	10	9	7	9.4
36	10	8	10	11	13	10.4	86	7	11	10	9	11	9.6
37	11	10	11	11	10	10.6	87	10	11	10	10	10	10.2
38	7	13	9	12	11	10.4	88	9	8	11	10	12	10.0
39	11	11	8	11	11	10.4	89	14	9	12	10	9	10.8
40	11	10	11	12	9	10.6	90	9	12	9	10	10	10.0
41	11	10	9	11	12	10.6	91	10	10	8	6	11	9.0
42	11	13	10	12	9	11.0	92	8	9	11	9	10	9.4
43	10	9	11	10	11	10.2	93	8	10	9	9	11	9.4
44	10	9	11	10	9	9.8	94	12	11	12	13	10	11.6
45	12	11	9	11	12	11.0	95	11	11	9	9	9	9.8
46	13	9	11	8	8	9.8	96	8	12	8	11	10	9.8
47	10	11	11	11	10	10.6	97	13	11	11	12	8	11.0
48	9	9	10	11	11	10.0	98	10	11	8	10	11	10.0
49	10	9	9	10	10	9.6	99	13	10	7	11	9	10.0
50	10	10	6	9	10	9.0	100	9	9	10	12	12	10.4

TABLE 6-10 Frequency Table for 100 Values of \bar{x}

Class	Class Limits		$f = \text{Frequency}$	$f/100 = \text{Relative Frequency}$
	Lower	Upper		
1	8.39	8.76	1	0.01
2	8.77	9.14	5	0.05
3	9.15	9.52	10	0.10
4	9.53	9.90	19	0.19
5	9.91	10.28	27	0.27
6	10.29	10.66	18	0.18
7	10.67	11.04	12	0.12
8	11.05	11.42	5	0.05
9	11.43	11.80	3	0.03

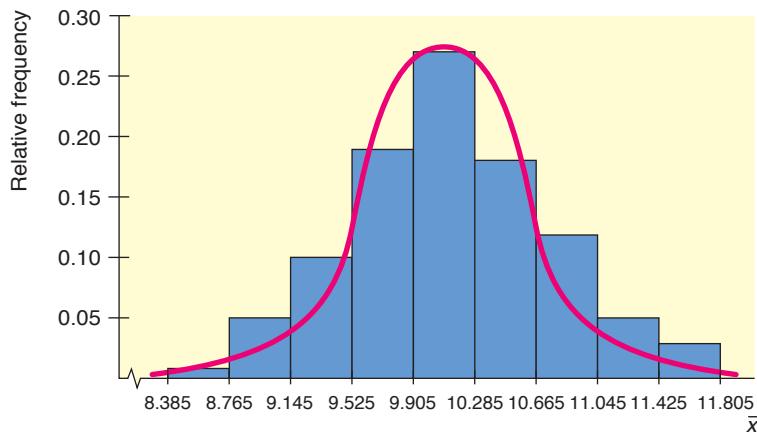
Let's organize our \bar{x} values into a frequency table. We used a class width of 0.38 to make Table 6-10.

Note: Techniques of Section 2.1 dictate a class width of 0.4. However, this choice results in the tenth class being beyond the data. Consequently, we shortened the class width slightly and also started the first class with a value slightly smaller than the smallest data value.

The far-right column of Table 6-10 contains relative frequencies $f/100$. Recall that relative frequencies may be thought of as probabilities, so we effectively have a probability distribution. Because \bar{x} represents the mean length of a trout (based on samples of five trout caught by each child), we estimate the probability of \bar{x} falling into each class by using the relative frequencies. Figure 6-32 is a relative-frequency or probability distribution of the \bar{x} values.

FIGURE 6-32

Estimates of Probabilities of \bar{x} Values



The bars of Figure 6-32 represent our estimated probabilities of \bar{x} values based on the data of Table 6-9. The bell-shaped curve represents the theoretical probability distribution that would be obtained if the number of children (i.e., number of \bar{x} values) were much larger.

Figure 6-32 represents a *probability sampling distribution* for the sample mean \bar{x} of trout lengths based on random samples of size 5. We see that the distribution is mound-shaped and even somewhat bell-shaped. Irregularities are due to the small number of samples used (only 100 sample means) and the rather small sample size (five trout per child). These irregularities would become less obvious and even disappear if the sample of children became much larger, if we used a larger number of classes in Figure 6-32, and if the number of trout in each sample became larger. In fact, the curve would eventually become a perfect

bell-shaped curve. We will discuss this property at some length in the next section, which introduces the *central limit theorem*.



There are other sampling distributions besides the \bar{x} distribution. Section 6.6 shows the sampling distribution for \hat{p} . In the chapters ahead, we will see that other statistics have different sampling distributions. However, the \bar{x} sampling distribution is very important. It will serve us well in our inferential work in Chapters 7 and 8 on estimation and testing.

Let us summarize the information about sampling distributions in the following exercise.

GUIDED EXERCISE 9**Terminology**

- (a) What is a population parameter? Give an example.
- (b) What is a sample statistic? Give an example.
- (c) What is a sampling distribution?
- (d) In Table 6-9, what makes up the members of the sample? What is the sample statistic corresponding to each sample? What is the sampling distribution? To which population parameter does this sampling distribution correspond?
- (e) Where will sampling distributions be used in our study of statistics?
- A population parameter is a numerical descriptive measure of a population. Examples are μ , σ , and p . (There are many others.)
- A sample statistic or statistic is a numerical descriptive measure of a sample. Examples are \bar{x} , s , and \hat{p} .
- A sampling distribution is a probability distribution for the sample statistic we are using based on all possible samples of the same size.
- There are 100 samples, each of which comprises five trout lengths. In the first sample, the five trout have lengths 11, 10, 10, 12, and 11. The sample statistic is the sample mean $\bar{x} = 10.8$. The sampling distribution is shown in Figure 6-32. This sampling distribution relates to the population mean μ of all lengths of trout taken from the Pinedale children's pond (i.e., trout over 6 inches long).
- Sampling distributions will be used for statistical inference. (Chapter 7 will concentrate on a method of inference called *estimation*. Chapter 8 will concentrate on a method of inference called *testing*.)

VIEWPOINT**"Chance Favors the Prepared Mind"**

—Louis Pasteur

It also has been said that a discovery is nothing more than an accident that meets a prepared mind. Sampling can be one of the best forms of preparation. In fact, sampling may be the primary way we humans venture into the unknown. Probability sampling distributions can provide new information for the sociologist, scientist, or economist. In addition, ordinary human sampling of life can help writers and artists develop preferences, styles, and insights. Ansel Adams became famous for photographing lyrical, unforgettable land scapes such as "Moonrise, Hernandez, New Mexico." Adams claimed that he was a strong believer in the quote by Pasteur. In fact, he claimed that the Hernandez photograph was just such a favored chance happening that his prepared mind readily grasped. During his lifetime, Adams made over \$25 million from sales and royalties on the Hernandez photograph.

SECTION 6.4 PROBLEMS

This is a good time to review several important concepts, some of which we have studied earlier. Please write out a careful but brief answer to each of the following questions.

1. | **Statistical Literacy** What is a population? Give three examples.
2. | **Statistical Literacy** What is a random sample from a population? *Hint:* See Section 1.2.
3. | **Statistical Literacy** What is a population parameter? Give three examples.
4. | **Statistical Literacy** What is a sample statistic? Give three examples.
5. | **Statistical Literacy** What is the meaning of the term *statistical inference*? What types of inferences will we make about population parameters?
6. | **Statistical Literacy** What is a sampling distribution?
7. | **Critical Thinking** How do frequency tables, relative frequencies, and histograms showing relative frequencies help us understand sampling distributions?
8. | **Critical Thinking** How can relative frequencies be used to help us estimate probabilities occurring in sampling distributions?
9. | **Critical Thinking** Give an example of a specific sampling distribution we studied in this section. Outline other possible examples of sampling distributions from areas such as business administration, economics, finance, psychology, political science, sociology, biology, medical science, sports, engineering, chemistry, linguistics, and so on.

SECTION 6.5

The Central Limit Theorem

FOCUS POINTS

- For a normal distribution, use μ and σ to construct the theoretical sampling distribution for the statistic \bar{x} .
- For large samples, use sample estimates to construct a good approximate sampling distribution for the statistic \bar{x} .
- Learn the statement and underlying meaning of the central limit theorem well enough to explain it to a friend who is intelligent, but (unfortunately) doesn't know much about statistics.

The \bar{x} Distribution, Given x Is Normal

In Section 6.4, we began a study of the distribution of \bar{x} values, where \bar{x} was the (sample) mean length of five trout caught by children at the Pinedale children's fishing pond. Let's consider this example again in the light of a very important theorem of mathematical statistics.

THEOREM 6.1 For a Normal Probability Distribution Let x be a random variable with a *normal distribution* whose mean is μ and whose standard deviation is σ . Let \bar{x} be the sample mean corresponding to random samples of size n taken from the x distribution. Then the following are true:

- (a) The \bar{x} distribution is a *normal distribution*.
- (b) The mean of the \bar{x} distribution is μ .
- (c) The standard deviation of the \bar{x} distribution is σ/\sqrt{n} .

We conclude from Theorem 6.1 that when x has a normal distribution, the \bar{x} distribution will be normal for any sample size n . Furthermore, we can convert the \bar{x} distribution to the standard normal z distribution using the following formulas.

$$\begin{aligned}\mu_{\bar{x}} &= \mu \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ z &= \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\end{aligned}$$

where n is the sample size,

μ is the mean of the x distribution, and

σ is the standard deviation of the x distribution.

Theorem 6.1 is a wonderful theorem! It states that the \bar{x} distribution will be normal provided the x distribution is normal. The sample size n could be 2, 3, 4, or any other (fixed) sample size we wish. Furthermore, the mean of the \bar{x} distribution is μ (same as for the x distribution), but the standard deviation is σ/\sqrt{n} (which is, of course, smaller than σ). The next example illustrates Theorem 6.1.

EXAMPLE 12

PROBABILITY REGARDING x AND \bar{x}

Suppose a team of biologists has been studying the Pinedale children's fishing pond. Let x represent the length of a single trout taken at random from the pond. This group of biologists has determined that x has a normal distribution with mean $\mu = 10.2$ inches and standard deviation $\sigma = 1.4$ inches.

- (a) What is the probability that a *single* trout taken at random from the pond is between 8 and 12 inches long?

SOLUTION: We use the methods of Section 6.3, with $\mu = 10.2$ and $\sigma = 1.4$, to get

$$z = \frac{x - \mu}{\sigma} = \frac{x - 10.2}{1.4}$$

Therefore,

$$\begin{aligned}P(8 < x < 12) &= P\left(\frac{8 - 10.2}{1.4} < z < \frac{12 - 10.2}{1.4}\right) \\ &= P(-1.57 < z < 1.29) \\ &= 0.9015 - 0.0582 = 0.8433\end{aligned}$$

Therefore, the probability is about 0.8433 that a *single* trout taken at random is between 8 and 12 inches long.

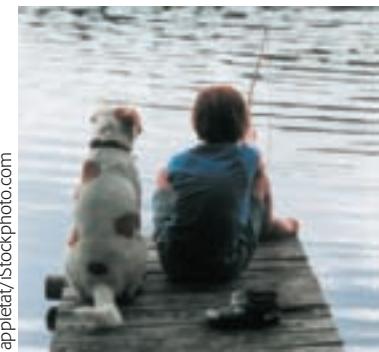
- (b) What is the probability that the *mean length* \bar{x} of five trout taken at random is between 8 and 12 inches?

SOLUTION: If we let $\mu_{\bar{x}}$ represent the mean of the distribution, then Theorem 6.1, part (b), tells us that

$$\mu_{\bar{x}} = \mu = 10.2$$

If $\sigma_{\bar{x}}$ represents the standard deviation of the $\sigma_{\bar{x}}$ distribution, then Theorem 6.1, part (c), tells us that

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 1.4/\sqrt{5} \approx 0.63$$



appletat/Stockphoto.com

To create a standard z variable from \bar{x} , we subtract $\mu_{\bar{x}}$ and divide by $\sigma_{\bar{x}}$:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 10.2}{0.63}$$

To standardize the interval $8 < \bar{x} < 12$, we use 8 and then 12 in place of \bar{x} in the preceding formula for z .

$$\begin{aligned} 8 &< \bar{x} < 12 \\ \frac{8 - 10.2}{0.63} &< z < \frac{12 - 10.2}{0.63} \\ -3.49 &< z < 2.86 \end{aligned}$$

Theorem 6.1, part (a), tells us that \bar{x} has a normal distribution. Therefore,

$$P(8 < \bar{x} < 12) = P(-3.49 < z < 2.86) = 0.9979 - 0.0002 = 0.9977$$

The probability is about 0.9977 that the mean length based on a sample size of 5 is between 8 and 12 inches.

- (c) Looking at the results of parts (a) and (b), we see that the probabilities (0.8433 and 0.9977) are quite different. Why is this the case?

SOLUTION: According to Theorem 6.1, both x and \bar{x} have a normal distribution, and both have the same mean of 10.2 inches. The difference is in the standard deviations for x and \bar{x} . The standard deviation of the x distribution is $\sigma = 1.4$. The standard deviation of the \bar{x} distribution is

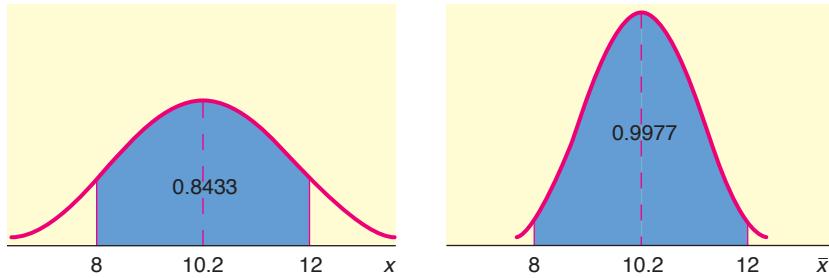
$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 1.4/\sqrt{5} \approx 0.63$$

The standard deviation of the \bar{x} distribution is less than half the standard deviation of the x distribution. Figure 6-33 shows the distributions of x and \bar{x} .

FIGURE 6-33

General Shapes of the x and \bar{x} Distributions

- (a) The x distribution with $\mu = 10.2$ and $\sigma = 1.4$
- (b) The \bar{x} distribution with $\mu_{\bar{x}} = 10.2$ and $\sigma_{\bar{x}} = 0.63$ for samples of size $n = 5$



Looking at Figure 6-33(a) and (b), we see that both curves use the same scale on the horizontal axis. The means are the same, and the shaded area is above the interval from 8 to 12 on each graph. It becomes clear that the smaller standard deviation of the \bar{x} distribution has the effect of gathering together much more of the total probability into the region over its mean. Therefore, the region from 8 to 12 has a much higher probability for the \bar{x} distribution.

Standard error of the mean

Theorem 6.1 describes the distribution of a particular statistic: namely, the distribution of sample mean \bar{x} . The standard deviation of a statistic is referred to as the *standard error* of that statistic.

The **standard error** is the standard deviation of a sampling distribution. For the \bar{x} sampling distribution,

$$\text{standard error} = \sigma_{\bar{x}} = \sigma/\sqrt{n}$$

Statistical software

The expression *standard error* appears commonly on printouts and refers to the standard deviation of the sampling distribution being used. (In Minitab, the expression SE MEAN refers to the standard error of the mean.)

Central limit theorem**The \bar{x} Distribution, Given x Follows Any Distribution**

Theorem 6.1 gives complete information about the \bar{x} distribution, provided the original x distribution is known to be normal. What happens if we don't have information about the shape of the original x distribution? The *central limit theorem* tells us what to expect.

THEOREM 6.2 **The Central Limit Theorem for Any Probability Distribution** If x possesses *any* distribution with mean μ and standard deviation σ , then the sample mean \bar{x} based on a random sample of size n will have a distribution that approaches the distribution of a normal random variable with mean μ and standard deviation σ/\sqrt{n} as n increases without limit.

The central limit theorem is indeed surprising! It says that x can have *any* distribution whatsoever, but that as the sample size gets larger and larger, the distribution of \bar{x} will approach a *normal* distribution. From this relation, we begin to appreciate the scope and significance of the normal distribution.

In the central limit theorem, the degree to which the distribution of \bar{x} values fits a normal distribution depends on both the selected value of n and the original distribution of x values. A natural question is: How large should the sample size be if we want to apply the central limit theorem? After a great deal of theoretical as well as empirical study, statisticians agree that if n is 30 or larger, the \bar{x} distribution will appear to be normal and the central limit theorem will apply. However, this rule should not be applied blindly. If the x distribution is definitely not symmetrical about its mean, then the \bar{x} distribution also will display a lack of symmetry. In such a case, a sample size larger than 30 may be required to get a reasonable approximation to the normal.

In practice, it is a good idea, when possible, to make a histogram of sample x values. If the histogram is approximately mound-shaped, and if it is more or less symmetrical, then we may be assured that, for all practical purposes, the \bar{x} distribution will be well approximated by a normal distribution and the central limit theorem will apply when the sample size is 30 or larger. The main thing to remember is that in almost all practical applications, a sample size of 30 or more is adequate for the central limit theorem to hold. However, in a few rare applications, you may need a sample size larger than 30 to get reliable results.

Let's summarize this information for convenient reference: For almost all x distributions, if we use a random sample of size 30 or larger, the \bar{x} distribution will be approximately normal. The larger the sample size becomes, the closer the \bar{x} distribution gets to the normal. Furthermore, we may convert the \bar{x} distribution to a standard normal distribution using the following formulas.

Using the central limit theorem to convert the \bar{x} distribution to the standard normal distribution

$$\begin{aligned}\mu_{\bar{x}} &= \mu \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ z &= \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\end{aligned}$$

Continued

where n is the sample size ($n \geq 30$),
 μ is the mean of the x distribution, and
 σ is the standard deviation of the x distribution.

Guided Exercise 10 shows how to standardize when appropriate. Then, Example 13 demonstrates the use of the central limit theorem in a decision-making process.

GUIDED EXERCISE 10

Central limit theorem

- (a) Suppose x has a *normal* distribution with mean $\mu = 18$ and standard deviation $\sigma = 13$. If you draw random samples of size 5 from the x distribution and \bar{x} represents the sample mean, what can you say about the \bar{x} distribution? How could you standardize the \bar{x} distribution?
- (b) Suppose you know that the x distribution has mean $\mu = 75$ and standard deviation $\sigma = 12$, but you have no information as to whether or not the x distribution is normal. If you draw samples of size 30 from the x distribution and \bar{x} represents the sample mean, what can you say about the \bar{x} distribution? How could you standardize the \bar{x} distribution?
- (c) Suppose you did not know that x had a normal distribution. Would you be justified in saying that the \bar{x} distribution is approximately normal if the sample size were $n = 8$?



Since the x distribution is given to be *normal*, the \bar{x} distribution also will be normal even though the sample size is much less than 30. The mean is $\mu_{\bar{x}} = \mu = 18$. The standard deviation is

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 13/\sqrt{5} \approx 1.3$$

We could standardize \bar{x} as follows:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 18}{1.3}$$



Since the sample size is large enough, the \bar{x} distribution will be an approximately normal distribution. The mean of the \bar{x} distribution is $\mu_{\bar{x}} = \mu = 75$

The standard deviation of the \bar{x} distribution is $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 12/\sqrt{30} \approx 2.2$

We could standardize \bar{x} as follows:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 75}{2.2}$$



No, the sample size should be 30 or larger if we don't know that x has a normal distribution.



EXAMPLE 13

CENTRAL LIMIT THEOREM

A certain strain of bacteria occurs in all raw milk. Let x be the bacteria count per milliliter of milk. The health department has found that if the milk is not contaminated, then x has a distribution that is more or less mound-shaped and symmetrical. The mean of the x distribution is $\mu = 2500$, and the standard deviation is $\sigma = 300$. In a large commercial dairy, the health inspector takes 42 random samples of the milk produced each day. At the end of the day, the bacteria count in each of the 42 samples is averaged to obtain the sample mean bacteria count \bar{x} .

- (a) Assuming the milk is not contaminated, what is the distribution of \bar{x} ?

SOLUTION: The sample size is $n = 42$. Since this value exceeds 30, the central limit theorem applies, and we know that \bar{x} will be approximately normal, with mean and standard deviation

$$\begin{aligned}\mu_{\bar{x}} &= \mu = 2500 \\ \sigma_{\bar{x}} &= \sigma/\sqrt{n} = 300/\sqrt{42} \approx 46.3\end{aligned}$$

- (b) Assuming the milk is not contaminated, what is the probability that the average bacteria count \bar{x} for one day is between 2350 and 2650 bacteria per milliliter?

SOLUTION: We convert the interval

$$2350 \leq \bar{x} \leq 2650$$

to a corresponding interval on the standard z axis.

$$\begin{aligned}z &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 2500}{46.3} \\ \bar{x} = 2350 &\text{ converts to } z = \frac{2350 - 2500}{46.3} \approx -3.24 \\ \bar{x} = 2650 &\text{ converts to } z = \frac{2650 - 2500}{46.3} \approx 3.24\end{aligned}$$

Therefore,

$$\begin{aligned}P(2350 \leq \bar{x} \leq 2650) &= P(-3.24 \leq z \leq 3.24) \\ &= 0.9994 - 0.0006 \\ &= 0.9988\end{aligned}$$

The probability is 0.9988 that \bar{x} is between 2350 and 2650.

- (c) **Interpretation** At the end of each day, the inspector must decide to accept or reject the accumulated milk that has been held in cold storage awaiting shipment. Suppose the 42 samples taken by the inspector have a mean bacteria count \bar{x} that is *not* between 2350 and 2650. If you were the inspector, what would be your comment on this situation?

SOLUTION: The probability that \bar{x} is between 2350 and 2650 for milk that is not contaminated is very high. If the inspector finds that the average bacteria count for the 42 samples is not between 2350 and 2650, then it is reasonable to conclude that there is something wrong with the milk. If \bar{x} is less than 2350, you might suspect someone added chemicals to the milk to artificially reduce the bacteria count. If \bar{x} is above 2650, you might suspect some other kind of biologic contamination.

Aleksas Kvedorcas/Stockphoto.com



In Problems 21, 22, and 23, we'll apply the central limit theorem to solve problems involving a *sum* of random variables.



PROCEDURE

HOW TO FIND PROBABILITIES REGARDING \bar{x}

Given a probability distribution of x values where

n = sample size

μ = mean of the x distribution

σ = standard deviation of the x distribution

1. If the x distribution is *normal*, then the \bar{x} distribution is *normal*.
2. Even if the x distribution is *not* normal, if the *sample size* $n \geq 30$, then, by the central limit theorem, the \bar{x} distribution is *approximately normal*.
3. Convert \bar{x} to z using the formula

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

4. Use the standard normal distribution to find the corresponding probabilities of events regarding \bar{x} .

GUIDED EXERCISE 11**Probability regarding \bar{x}**

In mountain country, major highways sometimes use tunnels instead of long, winding roads over high passes. However, too many vehicles in a tunnel at the same time can cause a hazardous situation. Traffic engineers are studying a long tunnel in Colorado. If x represents the time for a vehicle to go through the tunnel, it is known that the x distribution has mean $\mu = 12.1$ minutes and standard deviation $\sigma = 3.8$ minutes under ordinary traffic conditions. From a histogram of x values, it was found that the x distribution is mound-shaped with some symmetry about the mean.

Engineers have calculated that, *on average*, vehicles should spend from 11 to 13 minutes in the tunnel. If the time is less than 11 minutes, traffic is moving too fast for safe travel in the tunnel. If the time is more than 13 minutes, there is a problem of bad air quality (too much carbon monoxide and other pollutants).

Under ordinary conditions, there are about 50 vehicles in the tunnel at one time. What is the probability that the mean time for 50 vehicles in the tunnel will be from 11 to 13 minutes? We will answer this question in steps.

- (a) Let \bar{x} represent the sample mean based on samples of size 50. Describe the \bar{x} distribution.



From the central limit theorem, we expect the \bar{x} distribution to be approximately normal, with mean and standard deviation

$$\mu_{\bar{x}} = \mu = 12.1 \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.8}{\sqrt{50}} \approx 0.54$$

- (b) Find $P(11 < \bar{x} < 13)$.



We convert the interval

$$11 < \bar{x} < 13$$

to a standard z interval and use the standard normal probability table to find our answer. Since

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 12.1}{0.54}$$

$$\bar{x} = 11 \text{ converts to } z \approx \frac{11 - 12.1}{0.54} = -2.04$$

$$\text{and } \bar{x} = 13 \text{ converts to } z \approx \frac{13 - 12.1}{0.54} = 1.67$$

Therefore,

$$\begin{aligned} P(11 < \bar{x} < 13) &= P(-2.04 < z < 1.67) \\ &= 0.9525 - 0.0207 \\ &= 0.9318 \end{aligned}$$

- (c) *Interpret* your answer to part (b).



It seems that about 93% of the time, there should be no safety hazard for average traffic flow.

Jupiter Images


CRITICAL THINKING
Bias and Variability

Whenever we use a sample statistic as an estimate of a population parameter, we need to consider both *bias* and *variability* of the statistic.

A sample statistic is **unbiased** if the mean of its sampling distribution equals the value of the parameter being estimated.

The spread of the sampling distribution indicates the **variability of the statistic**. The spread is affected by the sampling method and the sample size. Statistics from larger random samples have spreads that are smaller.

We see from the central limit theorem that the sample mean \bar{x} is an unbiased estimator of the mean μ when $n \geq 30$. The variability of \bar{x} decreases as the sample size increases.

In Section 6.6, we will see that the sample proportion \hat{p} is an unbiased estimator of the population proportion of successes p in binomial experiments with sufficiently large numbers of trials n . Again, we will see that the variability of \hat{p} decreases with increasing numbers of trials.

The sample variance s^2 is an unbiased estimator for the population variance σ^2 .

LOOKING FORWARD

Sampling distributions for the mean \bar{x} and for proportions \hat{p} (Section 6.6) will form the basis of our work with estimation (Chapter 7) and hypothesis testing (Chapter 8). With these inferential statistics methods, we will be able to use information from a sample to make statements regarding the population. These statements will be made in terms of probabilities derived from the underlying sampling distributions.

VIEWPOINT

Chaos!

Is there a different side to random sampling? Can sampling be used as a weapon? According to the Wall Street Journal, the answer could be yes! The acronym for Create Havoc Around Our System is CHAOS. The Association of Flight Attendants (AFA) is a union that successfully used CHAOS against Alaska Airlines in 1994 as a negotiation tool. CHAOS involves a small sample of random strikes—a few flights at a time—instead of a mass walkout. The president of the AFA claims that by striking randomly, “We take control of the schedule.” The entire schedule becomes unreliable, and that is something management cannot tolerate. In 1986, TWA flight attendants struck in a mass walkout, and all were permanently replaced! Using CHAOS, only a few jobs are put at risk, and these are usually not lost. It appears that random sampling can be used as a weapon.

SECTION 6.5 PROBLEMS

In these problems, the word *average* refers to the arithmetic mean \bar{x} or μ , as appropriate.

1. | **Statistical Literacy** What is the standard error of a sampling distribution?
2. | **Statistical Literacy** What is the standard deviation of a sampling distribution called?
3. | **Statistical Literacy** List two unbiased estimators and their corresponding parameters.
4. | **Statistical Literacy** Describe how the variability of the \bar{x} distribution changes as the sample size increases.
5. | **Basic Computation: Central Limit Theorem** Suppose x has a distribution with a mean of 8 and a standard deviation of 16. Random samples of size $n = 64$ are drawn.

- (a) Describe the \bar{x} distribution and compute the mean and standard deviation of the distribution.
- (b) Find the z value corresponding to $\bar{x} = 9$.
- (c) Find $P(\bar{x} > 9)$.
- (d) **Interpretation** Would it be unusual for a random sample of size 64 from the x distribution to have a sample mean greater than 9? Explain.
6. **Basic Computation: Central Limit Theorem** Suppose x has a distribution with a mean of 20 and a standard deviation of 3. Random samples of size $n = 36$ are drawn.
- (a) Describe the \bar{x} distribution and compute the mean and standard deviation of the distribution.
- (b) Find the z value corresponding to $\bar{x} = 19$.
- (c) Find $P(\bar{x} < 19)$.
- (d) **Interpretation** Would it be unusual for a random sample of size 36 from the x distribution to have a sample mean less than 19? Explain.
7. **Statistical Literacy**
- (a) If we have a distribution of x values that is more or less mound-shaped and somewhat symmetrical, what is the sample size needed to claim that the distribution of sample means \bar{x} from random samples of that size is approximately normal?
- (b) If the original distribution of x values is known to be normal, do we need to make any restriction about sample size in order to claim that the distribution of sample means \bar{x} taken from random samples of a given size is normal?
8. **Critical Thinking** Suppose x has a distribution with $\mu = 72$ and $\sigma = 8$.
- (a) If random samples of size $n = 16$ are selected, can we say anything about the \bar{x} distribution of sample means?
- (b) If the original x distribution is *normal*, can we say anything about the \bar{x} distribution of random samples of size 16? Find $P(68 \leq \bar{x} \leq 73)$.
9. **Critical Thinking** Consider two \bar{x} distributions corresponding to the same x distribution. The first \bar{x} distribution is based on samples of size $n = 100$ and the second is based on samples of size $n = 225$. Which \bar{x} distribution has the smaller standard error? Explain.
10. **Critical Thinking** Consider an x distribution with standard deviation $\sigma = 12$.
- (a) If specifications for a research project require the standard error of the corresponding \bar{x} distribution to be 2, how large does the sample size need to be?
- (b) If specifications for a research project require the standard error of the corresponding \bar{x} distribution to be 1, how large does the sample size need to be?
11. **Critical Thinking** Suppose x has a distribution with $\mu = 15$ and $\sigma = 14$.
- (a) If a random sample of size $n = 49$ is drawn, find $\mu_{\bar{x}}$, $\sigma_{\bar{x}}$, and $P(15 \leq \bar{x} \leq 17)$.
- (b) If a random sample of size $n = 64$ is drawn, find $\mu_{\bar{x}}$, $\sigma_{\bar{x}}$, and $P(15 \leq \bar{x} \leq 17)$.
- (c) Why should you expect the probability of part (b) to be higher than that of part (a)? *Hint:* Consider the standard deviations in parts (a) and (b).
12. **Critical Thinking** Suppose an x distribution has mean $\mu = 5$. Consider two corresponding \bar{x} distributions, the first based on samples of size $n = 49$ and the second based on samples of size $n = 81$.
- (a) What is the value of the mean of each of the two \bar{x} distributions?
- (b) For which \bar{x} distribution is $P(\bar{x} > 6)$ smaller? Explain.
- (c) For which \bar{x} distribution is $P(4 < \bar{x} < 6)$ greater? Explain.
13. **Coal: Automatic Loader** Coal is carried from a mine in West Virginia to a power plant in New York in hopper cars on a long train. The automatic hopper car loader is set to put 75 tons of coal into each car. The actual weights of coal loaded into each car are *normally distributed*, with mean $\mu = 75$ tons and standard deviation $\sigma = 0.8$ ton.

- (a) What is the probability that one car chosen at random will have less than 74.5 tons of coal?
- (b) What is the probability that 20 cars chosen at random will have a mean load weight \bar{x} of less than 74.5 tons of coal?
- (c) **Interpretation** Suppose the weight of coal in one car was less than 74.5 tons. Would that fact make you suspect that the loader had slipped out of adjustment? Suppose the weight of coal in 20 cars selected at random had an average \bar{x} of less than 74.5 tons. Would that fact make you suspect that the loader had slipped out of adjustment? Why?
14. **Vital Statistics: Heights of Men** The heights of 18-year-old men are approximately *normally distributed*, with mean 68 inches and standard deviation 3 inches (based on information from *Statistical Abstract of the United States*, 112th Edition).
- (a) What is the probability that an 18-year-old man selected at random is between 67 and 69 inches tall?
- (b) If a random sample of nine 18-year-old men is selected, what is the probability that the mean height \bar{x} is between 67 and 69 inches?
- (c) **Interpretation** Compare your answers to parts (a) and (b). Is the probability in part (b) much higher? Why would you expect this?
15. **Medical: Blood Glucose** Let x be a random variable that represents the level of glucose in the blood (milligrams per deciliter of blood) after a 12-hour fast. Assume that for people under 50 years old, x has a distribution that is approximately normal, with mean $\mu = 85$ and estimated standard deviation $\sigma = 25$ (based on information from *Diagnostic Tests with Nursing Applications*, edited by S. Loeb, Springhouse). A test result $x < 40$ is an indication of severe excess insulin, and medication is usually prescribed.
- (a) What is the probability that, on a single test, $x < 40$?
- (b) Suppose a doctor uses the average \bar{x} for two tests taken about a week apart. What can we say about the probability distribution of \bar{x} ? Hint: See Theorem 6.1. What is the probability that $\bar{x} < 40$?
- (c) Repeat part (b) for $n = 3$ tests taken a week apart.
- (d) Repeat part (b) for $n = 5$ tests taken a week apart.
- (e) **Interpretation** Compare your answers to parts (a), (b), (c), and (d). Did the probabilities decrease as n increased? Explain what this might imply if you were a doctor or a nurse. If a patient had a test result of $\bar{x} < 40$ based on five tests, explain why either you are looking at an extremely rare event or (more likely) the person has a case of excess insulin.
16. **Medical: White Blood Cells** Let x be a random variable that represents white blood cell count per cubic milliliter of whole blood. Assume that x has a distribution that is approximately normal, with mean $\mu = 7500$ and estimated standard deviation $\sigma = 1750$ (see reference in Problem 15). A test result of $x < 3500$ is an indication of leukopenia. This indicates bone marrow depression that may be the result of a viral infection.
- (a) What is the probability that, on a single test, x is less than 3500?
- (b) Suppose a doctor uses the average \bar{x} for two tests taken about a week apart. What can we say about the probability distribution of \bar{x} ? What is the probability of $\bar{x} < 3500$?
- (c) Repeat part (b) for $n = 3$ tests taken a week apart.
- (d) **Interpretation** Compare your answers to parts (a), (b), and (c). How did the probabilities change as n increased? If a person had $\bar{x} < 3500$ based on three tests, what conclusion would you draw as a doctor or a nurse?
17. **Wildlife: Deer** Let x be a random variable that represents the weights in kilograms (kg) of healthy adult female deer (does) in December in Mesa Verde National Park. Then x has a distribution that is approximately normal, with mean $\mu = 63.0$ kg and standard deviation $\sigma = 7.1$ kg (Source: *The Mule Deer*



of Mesa Verde National Park, by G. W. Mierau and J. L. Schmidt, Mesa Verde Museum Association). Suppose a doe that weighs less than 54 kg is considered undernourished.

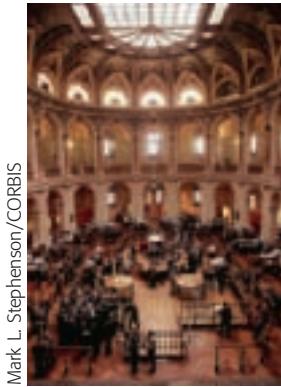
- What is the probability that a single doe captured (weighed and released) at random in December is undernourished?
- If the park has about 2200 does, what number do you expect to be undernourished in December?
- Interpretation** To estimate the health of the December doe population, park rangers use the rule that the average weight of $n = 50$ does should be more than 60 kg. If the average weight is less than 60 kg, it is thought that the entire population of does might be undernourished. What is the probability that the average weight \bar{x} for a random sample of 50 does is less than 60 kg (assume a healthy population)?
- Interpretation** Compute the probability that $\bar{x} < 64.2$ kg for 50 does (assume a healthy population). Suppose park rangers captured, weighed, and released 50 does in December, and the average weight was $\bar{x} = 64.2$ kg. Do you think the doe population is undernourished or not? Explain.

18. **Focus Problem: Impulse Buying** Let x represent the dollar amount spent on supermarket impulse buying in a 10-minute (unplanned) shopping interval. Based on a *Denver Post* article, the mean of the x distribution is about \$20 and the estimated standard deviation is about \$7.

- Consider a random sample of $n = 100$ customers, each of whom has 10 minutes of unplanned shopping time in a supermarket. From the central limit theorem, what can you say about the probability distribution of \bar{x} , the average amount spent by these customers due to impulse buying? What are the mean and standard deviation of the \bar{x} distribution? Is it necessary to make any assumption about the x distribution? Explain.
- What is the probability that \bar{x} is between \$18 and \$22?
- Let us assume that x has a distribution that is approximately normal. What is the probability that x is between \$18 and \$22?
- Interpretation:** In part (b), we used \bar{x} , the *average* amount spent, computed for 100 customers. In part (c), we used x , the amount spent by only *one* customer. The answers to parts (b) and (c) are very different. Why would this happen? In this example, \bar{x} is a much more predictable or reliable statistic than x . Consider that almost all marketing strategies and sales pitches are designed for the *average* customer and *not the individual* customer. How does the central limit theorem tell us that the average customer is much more predictable than the individual customer?

19. **Finance: Templeton Funds** Templeton World is a mutual fund that invests in both U.S. and foreign markets. Let x be a random variable that represents the monthly percentage return for the Templeton World fund. Based on information from the *Morningstar Guide to Mutual Funds* (available in most libraries), x has mean $\mu = 1.6\%$ and standard deviation $\sigma = 0.9\%$.

- Templeton World fund has over 250 stocks that combine together to give the overall monthly percentage return x . We can consider the monthly return of the stocks in the fund to be a sample from the population of monthly returns of all world stocks. Then we see that the overall monthly return x for Templeton World fund is itself an average return computed using all 250 stocks in the fund. Why would this indicate that x has an approximately normal distribution? Explain. Hint: See the discussion after Theorem 7.2.
- After 6 months, what is the probability that the *average* monthly percentage return \bar{x} will be between 1% and 2%? Hint: See Theorem 7.1, and assume that x has a normal distribution as based on part (a).
- After 2 years, what is the probability that \bar{x} will be between 1% and 2%?
- Compare your answers to parts (b) and (c). Did the probability increase as n (number of months) increased? Why would this happen?



Mark L. Stephenson/CORBIS

- (e) **Interpretation:** If after 2 years the average monthly percentage return \bar{x} was less than 1%, would that tend to shake your confidence in the statement that $\mu = 1.6\%$? Might you suspect that μ has slipped below 1.6%? Explain.
20. **Finance: European Growth Fund** A european growth mutual fund specializes in stocks from the British Isles, continental Europe, and Scandinavia. The fund has over 100 stocks. Let x be a random variable that represents the monthly percentage return for this fund. Based on information from *Morningstar* (see Problem 19), x has mean $\mu = 1.4\%$ and standard deviation $\sigma = 0.8\%$.
- Let's consider the monthly return of the stocks in the European growth fund to be a sample from the population of monthly returns of all European stocks. Is it reasonable to assume that x (the average monthly return on the 100 stocks in the European growth fund) has a distribution that is approximately normal? Explain. Hint: See Problem 19, part (a).
 - After 9 months, what is the probability that the *average* monthly percentage return \bar{x} will be between 1% and 2%? Hint: See Theorem 6.1 and the results of part (a).
 - After 18 months, what is the probability that the *average* monthly percentage return \bar{x} will be between 1% and 2%?
 - Compare your answers to parts (b) and (c). Did the probability increase as n (number of months) increased? Why would this happen?
 - Interpretation:** If after 18 months the average monthly percentage return \bar{x} is more than 2%, would that tend to shake your confidence in the statement that $\mu = 1.4\%$? If this happened, do you think the European stock market might be heating up? Explain.
21. **Expand Your Knowledge: Totals Instead of Averages** Let x be a random variable that represents checkout time (time spent in the actual checkout process) in minutes in the express lane of a large grocery. Based on a consumer survey, the mean of the x distribution is about $\mu = 2.7$ minutes, with standard deviation $\sigma = 0.6$ minute. Assume that the express lane always has customers waiting to be checked out and that the distribution of x values is more or less symmetrical and mound-shaped. What is the probability that the *total* checkout time for the next 30 customers is less than 90 minutes? Let us solve this problem in steps.
- Let x_i (for $i = 1, 2, 3, \dots, 30$) represent the checkout time for each customer. For example, x_1 is the checkout time for the first customer, x_2 is the checkout time for the second customer, and so forth. Each x_i has mean $\mu = 2.7$ minutes and standard deviation $\sigma = 0.6$ minute. Let $w = x_1 + x_2 + \dots + x_{30}$. Explain why the problem is asking us to compute the probability that w is less than 90.
 - Use a little algebra and explain why $w < 90$ is mathematically equivalent to $w/30 < 3$. Since w is the total of the 30 x values, then $w/30 = \bar{x}$. Therefore, the statement $\bar{x} < 3$ is equivalent to the statement $w < 90$. From this we conclude that the probabilities $P(\bar{x} < 3)$ and $P(w < 90)$ are equal.
 - What does the central limit theorem say about the probability distribution of \bar{x} ? Is it approximately normal? What are the mean and standard deviation of the \bar{x} distribution?
 - Use the result of part (c) to compute $P(\bar{x} < 3)$. What does this result tell you about $P(w < 90)$?
22. **Totals Instead of Averages: Airplane Takeoff Time** The taxi and takeoff time for commercial jets is a random variable x with a mean of 8.5 minutes and a standard deviation of 2.5 minutes. Assume that the distribution of taxi and takeoff times is approximately normal. You may assume that the jets are lined up on a runway so that one taxies and takes off immediately after another, and that they take off one at a time on a given runway. What is the probability that for 36 jets on a given runway, total taxi and takeoff time will be

- (a) less than 320 minutes?
 (b) more than 275 minutes?
 (c) between 275 and 320 minutes?

Hint: See Problem 21.



23.

Totals Instead of Averages: Escape Dunes It's true—sand dunes in Colorado rival sand dunes of the Great Sahara Desert! The highest dunes at Great Sand Dunes National Monument can exceed the highest dunes in the Great Sahara, extending over 700 feet in height. However, like all sand dunes, they tend to move around in the wind. This can cause a bit of trouble for temporary structures located near the “escaping” dunes. Roads, parking lots, campgrounds, small buildings, trees, and other vegetation are destroyed when a sand dune moves in and takes over. Such dunes are called “escape dunes” in the sense that they move out of the main body of sand dunes and, by the force of nature (prevailing winds), take over whatever space they choose to occupy. In most cases, dune movement does not occur quickly. An escape dune can take years to relocate itself. Just how fast does an escape dune move? Let x be a random variable representing movement (in feet per year) of such sand dunes (measured from the crest of the dune). Let us assume that x has a normal distribution with $\mu = 17$ feet per year and $\sigma = 3.3$ feet per year. (For more information, see *Hydrologic, Geologic, and Biologic Research at Great Sand Dunes National Monument and Vicinity, Colorado*, proceedings of the National Park Service Research Symposium.)

Under the influence of prevailing wind patterns, what is the probability that

- (a) an escape dune will move a total distance of more than 90 feet in 5 years?
 (b) an escape dune will move a total distance of less than 80 feet in 5 years?
 (c) an escape dune will move a total distance of between 80 and 90 feet in 5 years?

Hint: See Problem 21 and Theorem 6.1.



SECTION 6.6

Normal Approximation to Binomial Distribution and to \hat{p} Distribution

FOCUS POINTS

- State the assumptions needed to use the normal approximation to the binomial distribution.
- Compute μ and σ for the normal approximation.
- Use the continuity correction to convert a range of r values to a corresponding range of normal x values.
- Convert the x values to a range of standardized z scores and find desired probabilities.
- Describe the sampling distribution for proportions \hat{p} .

The probability that a new vaccine will protect adults from cholera is known to be 0.85. The vaccine is administered to 300 adults who must enter an area where the disease is prevalent. What is the probability that more than 280 of these adults will be protected from cholera by the vaccine?

This question falls into the category of a binomial experiment, with the number of trials n equal to 300, the probability of success p equal to 0.85, and the number of successes r greater than 280. It is possible to use the formula for the binomial distribution to compute the probability that r is greater than 280. However, this approach would involve a number of tedious and long calculations. There is an easier way to do this problem, for under the conditions stated below, the normal distribution can be used to approximate the binomial distribution.

Normal approximation to the binomial distribution

Consider a binomial distribution where

n = number of trials

r = number of successes

p = probability of success on a single trial

$q = 1 - p$ = probability of failure on a single trial

Criteria $np > 5$ and $nq > 5$

If $np > 5$ and $nq > 5$, then r has a binomial distribution that is approximated by a normal distribution with

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{npq}$$

Note: As n increases, the approximation becomes better.

Example 14 demonstrates that as n increases, the normal approximation to the binomial distribution improves.



EXAMPLE 14

BINOMIAL DISTRIBUTION GRAPHS

Graph the binomial distributions for which $p = 0.25$, $q = 0.75$, and the number of trials is first $n = 3$, then $n = 10$, then $n = 25$, and finally $n = 50$.

SOLUTION: The authors used a computer program to obtain the binomial distributions for the given values of p , q , and n . The results have been organized and graphed in Figures 6-34, 6-35, 6-36, and 6-37.

FIGURE 6-34

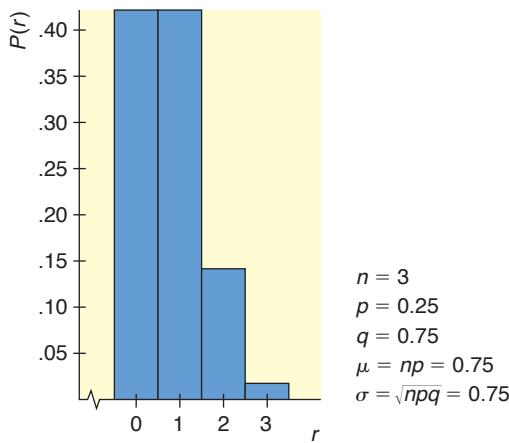


FIGURE 6-35

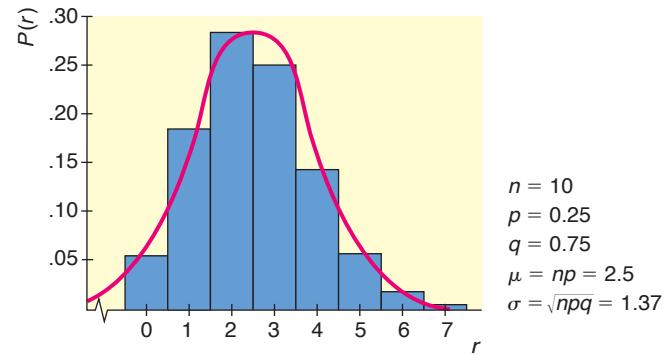


FIGURE 6-36

Good Normal Approximation; $np > 5$ and $nq > 5$

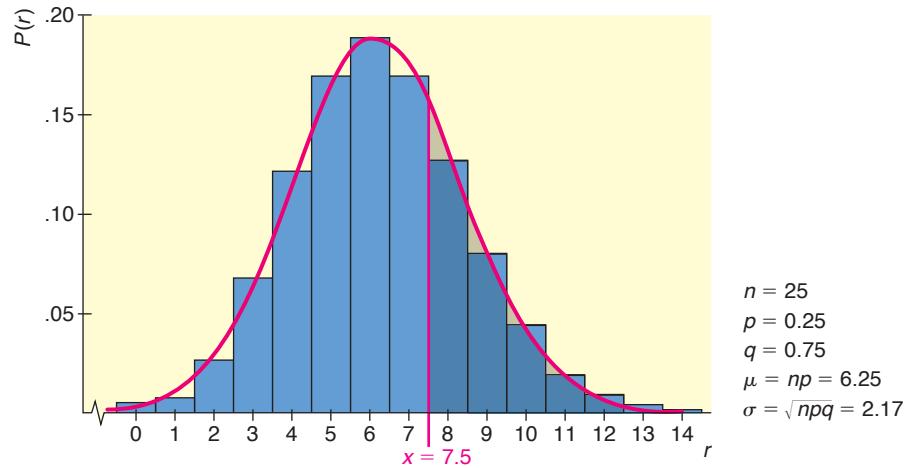
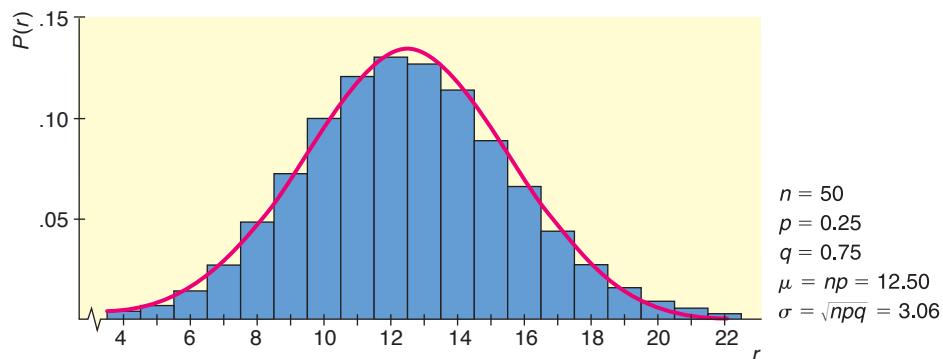


FIGURE 6-37

Good Normal Approximation; $np > 5$
and $nq > 5$



When $n = 3$, the outline of the histogram does not even begin to take the shape of a normal curve. But when $n = 10, 25$, or 50 , it does begin to take a normal shape, indicated by the red curve. From a theoretical point of view, the histograms in Figures 6-35, 6-36, and 6-37 would have bars for all values of r from $r = 0$ to $r = n$. However, in the construction of these histograms, the bars of height less than 0.001 unit have been omitted—that is, in this example, probabilities less than 0.001 have been rounded to 0 .

EXAMPLE 15

NORMAL APPROXIMATION

The owner of a new apartment building must install 25 water heaters. From past experience in other apartment buildings, she knows that Quick Hot is a good brand. A Quick Hot heater is guaranteed for 5 years only, but from the owner's past experience, she knows that the probability it will last 10 years is 0.25 .

- (a) What is the probability that 8 or more of the 25 water heaters will last at least 10 years? Define success to mean a water heater that lasts at least 10 years.

SOLUTION: In this example, $n = 25$ and $p = 0.25$, so Figure 6-36 (on the preceding page) represents the probability distribution we will use. Let r be the binomial random variable corresponding to the number of successes out of $n = 25$ trials. We want to find $P(r \geq 8)$ by using the normal approximation. This probability is represented graphically (Figure 6-36) by the area of the bar over 8 plus the areas of all bars to the right of the bar over 8 .

Let x be a normal random variable corresponding to a normal distribution, with $\mu = np = 25(0.25) = 6.25$ and $\sigma = \sqrt{npq} = \sqrt{25(0.25)(0.75)} \approx 2.17$. This normal curve is represented by the red line in Figure 6-36. The area under the normal curve from $x = 7.5$ to the right is approximately the same as the areas of the bars from the bar over $r = 8$ to the right. It is important to notice that we start with $x = 7.5$ because the bar over $r = 8$ really starts at $x = 7.5$.

The areas of the bars and the area under the corresponding red (normal) curve are approximately equal, so we conclude that $P(r \geq 8)$ is approximately equal to $P(x \geq 7.5)$.

When we convert $x = 7.5$ to standard units, we get

$$z = \frac{x - \mu}{\sigma} = \frac{7.5 - 6.25}{2.17} \quad (\text{Use } \mu = 6.25 \text{ and } \sigma = 2.17.) \\ \approx 0.58$$

The probability we want is

$$P(x \geq 7.5) = P(z \geq 0.58) = 1 - P(z \leq 0.58) = 1 - 0.7190 = 0.2810$$

- (b) How does this result compare with the result we can obtain by using the formula for the binomial probability distribution with $n = 25$ and $p = 0.25$?

SOLUTION: Using the binomial distribution function on the TI-84Plus/TI-83Plus/TI-nspire model calculators, the authors computed that



$P(r \geq 8) \approx 0.2735$. This means that the probability is approximately 0.27 that 8 or more water heaters will last at least 10 years.

(c) How do the results of parts (a) and (b) compare?

SOLUTION: The error of approximation is the difference between the approximate normal value (0.2810) and the binomial value (0.2735). The error is only $0.2810 - 0.2735 = 0.0075$, which is negligible for most practical purposes.

We knew in advance that the normal approximation to the binomial probability would be good, since $np = 25(0.25) = 6.25$ and $nq = 25(0.75) = 18.75$ are both greater than 5. These are the conditions that assure us that the normal approximation will be sufficiently close to the binomial probability for most practical purposes.

Continuity correction: Converting r values to x values

PROCEDURE

HOW TO MAKE THE CONTINUITY CORRECTION

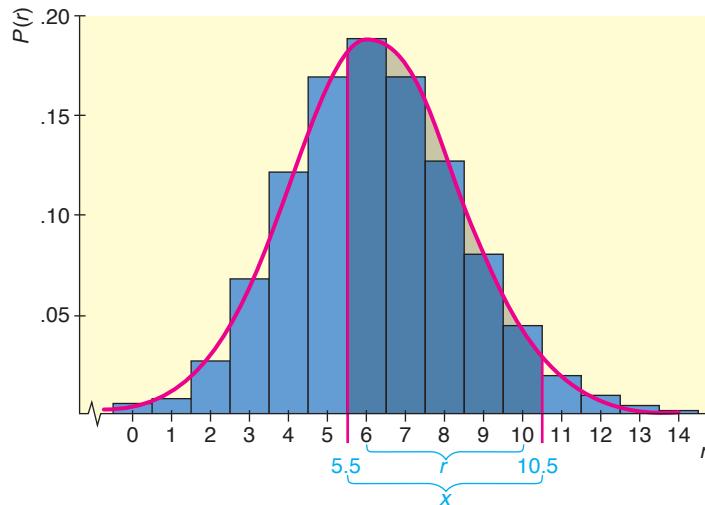
Convert the discrete random variable r (number of successes) to the continuous normal random variable x by doing the following:

- If r is a left point of an interval, subtract 0.5 to obtain the corresponding normal variable x ; that is, $x = r - 0.5$.
- If r is a right point of an interval, add 0.5 to obtain the corresponding normal variable x ; that is, $x = r + 0.5$.

For instance, $P(6 \leq r \leq 10)$, where r is a binomial random variable, is approximated by $P(5.5 \leq x \leq 10.5)$, where x is the corresponding normal random variable (see Figure 6-38).

COMMENT Both the binomial and Poisson distributions are for *discrete* random variables. Therefore, adding or subtracting 0.5 to r was not necessary when we approximated the binomial distribution by the Poisson distribution (Section 5.4). However, the normal distribution is for a *continuous* random variable. In this case, adding or subtracting 0.5 to or from (as appropriate) r will improve the approximation of the normal to the binomial distribution.

FIGURE 6-38
 $P(6 \leq r \leq 10)$ Is Approximately Equal to $P(5.5 \leq x \leq 10.5)$



GUIDED EXERCISE 12**Continuity correction**

From many years of observation, a biologist knows that the probability is only 0.65 that any given Arctic tern will survive the migration from its summer nesting area to its winter feeding grounds. A random sample of 500 Arctic terns were banded at their summer nesting area. Use the normal approximation to the binomial and the following steps to find the probability that between 310 and 340 of the banded Arctic terns will survive the migration. Let r be the number of surviving terns.



William Mullins/Photo Researchers

Arctic tern

- (a) To approximate $P(310 \leq r \leq 340)$, we use the normal curve with $\mu = \underline{\hspace{2cm}}$ and $\sigma = \underline{\hspace{2cm}}$.



We use the normal curve with

$$\mu = np = 500(0.65) = 325 \text{ and}$$

$$\sigma = \sqrt{npq} = \sqrt{500(0.65)(0.35)} \approx 10.67$$

- (b) $P(310 \leq r \leq 340)$ is approximately equal to $P(\underline{\hspace{2cm}} \leq x \leq \underline{\hspace{2cm}})$, where x is a variable from the normal distribution described in part (a).



Since 310 is the left endpoint, we subtract 0.5, and since 340 is the right endpoint, we add 0.5. Consequently,

$$P(310 \leq r \leq 340) \approx P(309.5 \leq x \leq 340.5)$$

- (c) Convert the condition $309.5 \leq x \leq 340.5$ to a condition in standard units.



Since $\mu = 325$ and $\sigma \approx 10.67$, the condition $309.5 \leq x \leq 340.5$ becomes

$$\frac{309.5 - 325}{10.67} \leq z \leq \frac{340.5 - 325}{10.67}$$

or

$$-1.45 \leq z \leq 1.45$$

- (d) $P(310 \leq r \leq 340) = P(309.5 \leq x \leq 340.5)$
 $= P(-1.45 \leq z \leq 1.45)$
 $= \underline{\hspace{2cm}}$



$$\begin{aligned} P(-1.45 \leq z \leq 1.45) &= P(z \leq 1.45) - P(z \leq -1.45) \\ &= 0.9265 - 0.0735 \\ &= 0.8530 \end{aligned}$$

- (e) Will the normal distribution make a good approximation to the binomial for this problem? Explain your answer.



Since

$$np = 500(0.65) = 35$$

and

$$nq = 500(0.35) = 175$$

are both greater than 5, the normal distribution will be a good approximation to the binomial.

Sampling Distributions for Proportions

In Sections 6.4 and 6.5, we studied the sampling distribution for the mean. Now we have the tools to look at sampling distributions for proportions. Suppose we repeat a binomial experiment with n trials again and again and, for each n trials,

record the sample proportion of successes $\hat{p} = r/n$. The \hat{p} values form a sampling distribution for proportions.

Sampling distribution for \hat{p}

Sampling distribution for the proportion $\hat{p} = \frac{r}{n}$

Given n = number of binomial trials (fixed constant)

r = number of successes

p = probability of success on each trial

$q = 1 - p$ = probability of failure on each trial

If $np > 5$ and $nq > 5$, then the random variable $\hat{p} = r/n$ can be approximated by a normal random variable (x) with mean and standard deviation

$$\mu_{\hat{p}} = p \quad \text{and} \quad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

Standard error of a proportion

TERMINOLOGY The *standard error* for the \hat{p} distribution is the standard deviation $\sigma_{\hat{p}}$ of the \hat{p} sampling distribution.

COMMENT To obtain the information regarding the sampling distribution for the proportion $\hat{p} = r/n$, we consider the sampling distribution for r , the number of successes out of n binomial trials. Earlier we saw that when $np > 5$ and $nq > 5$, the r distribution is approximately normal, with mean $\mu_r = np$ and standard deviation $\sigma_r = \sqrt{npq}$. Notice that $\hat{p} = r/n$ is a linear function of r . This means that the \hat{p} distribution is also approximately normal when np and nq are both greater than 5. In addition, from our work in Section 5.1 with linear functions of random variables, we know that $\mu_{\hat{p}} = \mu_r/n = np/n = p$ and $\sigma_{\hat{p}} = \sigma_r/n = \sqrt{npq}/n = \sqrt{pq}/n$. Although the values r/n are discrete for a fixed n , we do not use a continuity correction for the \hat{p} distribution. This is the accepted standard practice for applications in inferential statistics, especially considering the requirements that $np > 5$ and $nq >$ are met.

We see from the sampling distribution for proportions that the mean of the \hat{p} distribution is p , the population proportion of successes. This means that \hat{p} is an *unbiased estimator* for p .

LOOKING FORWARD

We will use the sampling distribution for proportions in our work with estimation (Chapter 7) and hypothesis testing (Chapter 8).



EXAMPLE 16

SAMPLING DISTRIBUTION OF \hat{p}

The annual crime rate in the Capital Hill neighborhood of Denver is 111 victims per 1000 residents. This means that 111 out of 1000 residents have been the victim of at least one crime (Source: *Neighborhood Facts*, Piton Foundation). For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Piton Foundation. These crimes range from relatively minor crimes (stolen hubcaps or purse snatching) to major crimes (murder). The Arms is an apartment building in this neighborhood that has 50 year-round residents. Suppose we view each of the $n = 50$ residents as a binomial trial.



The random variable r (which takes on values $0, 1, 2, \dots, 50$) represents the number of victims of at least one crime in the next year.

- (a) What is the population probability p that a resident in the Capital Hill neighborhood will be the victim of a crime next year? What is the probability q that a resident will not be a victim?

SOLUTION: Using the Piton Foundation report, we take

$$p = 111/1000 = 0.111 \quad \text{and} \quad q = 1 - p = 0.889$$

- (b) Consider the random variable

$$\hat{p} = \frac{r}{n} = \frac{r}{50}$$

Can we approximate the \hat{p} distribution with a normal distribution? Explain.

SOLUTION: $np = 50(0.111) = 5.55$

$$nq = 50(0.889) = 44.45$$

Since both np and nq are greater than 5, we can approximate the \hat{p} distribution with a normal distribution.

- (c) What are the mean and standard deviation for the \hat{p} distribution?

SOLUTION: $\mu_{\hat{p}} = p = 0.111$

$$\begin{aligned}\sigma_{\hat{p}} &= \sqrt{\frac{pq}{n}} \\ &= \sqrt{\frac{(0.111)(0.889)}{50}} \approx 0.044\end{aligned}$$



VIEWPOINT

Sunspots, Tree Rings, and Statistics

Ancient Chinese astronomers recorded extreme sunspot activity, with a peak around 1200 A.D. Mesa Verde tree rings in the period between 1276 and 1299 were unusually narrow, indicating a drought and/or a severe cold spell in the region at that time. A cooling trend could have narrowed the window of frost-free days below the approximately 80 days needed for cultivation of aboriginal corn and beans. Is this the reason the ancient Anasazi dwellings in Mesa Verde were abandoned? Is there a connection to the extreme sunspot activity? Much research and statistical work continue to be done on this topic.

Reference: *Prehistoric Astronomy in the Southwest*, by J. McKim Malville and C. Putnam, Department of Astronomy, University of Colorado.

SECTION 6.6 PROBLEMS

Note: When we say *between a and b*, we mean every value from *a* to *b*, *including a and b*. Due to rounding, your answers might vary slightly from answers given in the text.

1. **Statistical Literacy** Binomial probability distributions depend on the number of trials n of a binomial experiment and the probability of success p on each trial. Under what conditions is it appropriate to use a normal approximation to the binomial?
2. **Statistical Literacy** When we use a normal distribution to approximate a binomial distribution, why do we make a continuity correction?
3. **Basic Computation: Normal Approximation to a Binomial Distribution** Suppose we have a binomial experiment with $n = 40$ trials and a probability of success $p = 0.50$.

- (a) Is it appropriate to use a normal approximation to this binomial distribution? Why?
- (b) Compute μ and σ of the approximating normal distribution.
- (c) Use a continuity correction factor to convert the statement $r \geq 23$ successes to a statement about the corresponding normal variable x .
- (d) Estimate $P(r \geq 23)$.
- (e) **Interpretation** Is it unusual for a binomial experiment with 40 trials and probability of success 0.50 to have 23 or more successes? Explain.
4. **Basic Computation: Normal Approximation to a Binomial Distribution** Suppose we have a binomial experiment with $n = 40$ trials and probability of success $p = 0.85$.
- (a) Is it appropriate to use a normal approximation to this binomial distribution? Why?
- (b) Compute μ and σ of the approximating normal distribution.
- (c) Use a continuity correction factor to convert the statement $r < 30$ successes to a statement about the corresponding normal variable x .
- (d) Estimate $P(r < 30)$.
- (e) **Interpretation** Is it unusual for a binomial experiment with 40 trials and probability of success 0.85 to have fewer than 30 successes? Explain.
5. **Critical Thinking** You need to compute the probability of 5 or fewer successes for a binomial experiment with 10 trials. The probability of success on a single trial is 0.43. Since this probability of success is not in the table, you decide to use the normal approximation to the binomial. Is this an appropriate strategy? Explain.
6. **Critical Thinking** Consider a binomial experiment with 20 trials and probability 0.45 of success on a single trial.
- (a) Use the binomial distribution to find the probability of exactly 10 successes.
- (b) Use the normal distribution to approximate the probability of exactly 10 successes.
- (c) Compare the results of parts (a) and (b).
- In the following problems, check that it is appropriate to use the normal approximation to the binomial. Then use the normal distribution to estimate the requested probabilities.
7. **Health: Lead Contamination** More than a decade ago, high levels of lead in the blood put 88% of children at risk. A concerted effort was made to remove lead from the environment. Now, according to the *Third National Health and Nutrition Examination Survey (NHANES III)* conducted by the Centers for Disease Control, only 9% of children in the United States are at risk of high blood-lead levels.
- (a) In a random sample of 200 children taken more than a decade ago, what is the probability that 50 or more had high blood-lead levels?
- (b) In a random sample of 200 children taken now, what is the probability that 50 or more have high blood-lead levels?
8. **Insurance: Claims** Do you try to *pad* an insurance claim to cover your deductible? About 40% of all U.S. adults will try to pad their insurance claims! (Source: *Are You Normal?*, by Bernice Kanner, St. Martin's Press.) Suppose that you are the director of an insurance adjustment office. Your office has just received 128 insurance claims to be processed in the next few days. What is the probability that
- (a) half or more of the claims have been padded?
- (b) fewer than 45 of the claims have been padded?
- (c) from 40 to 64 of the claims have been padded?
- (d) more than 80 of the claims have *not* been padded?
9. **Longevity: 90th Birthday** It is estimated that 3.5% of the general population will live past their 90th birthday (*Statistical Abstract of the United States*, 112th Edition). In a graduating class of 753 high school seniors, what is the probability that

- (a) 15 or more will live beyond their 90th birthday?
(b) 30 or more will live beyond their 90th birthday?
(c) between 25 and 35 will live beyond their 90th birthday?
(d) more than 40 will live beyond their 90th birthday?
10. **Fishing: Billfish** Ocean fishing for billfish is very popular in the Cozumel region of Mexico. In *World Record Game Fishes* (published by the International Game Fish Association), it was stated that in the Cozumel region, about 44% of strikes (while trolling) result in a catch. Suppose that on a given day a fleet of fishing boats got a total of 24 strikes. What is the probability that the number of fish caught was
(a) 12 or fewer?
(b) 5 or more?
(c) between 5 and 12?
11. **Grocery Stores: New Products** *The Denver Post* stated that 80% of all new products introduced in grocery stores fail (are taken off the market) within 2 years. If a grocery store chain introduces 66 new products, what is the probability that within 2 years
(a) 47 or more fail?
(b) 58 or fewer fail?
(c) 15 or more succeed?
(d) fewer than 10 succeed?
12. **Crime: Murder** What are the chances that a person who is murdered actually knew the murderer? The answer to this question explains why a lot of police detective work begins with relatives and friends of the victim! About 64% of people who are murdered actually knew the person who committed the murder (*Chances: Risk and Odds in Everyday Life*, by James Burke). Suppose that a detective file in New Orleans has 63 current unsolved murders. What is the probability that
(a) at least 35 of the victims knew their murderers?
(b) at most 48 of the victims knew their murderers?
(c) fewer than 30 victims did *not* know their murderers?
(d) more than 20 victims did *not* know their murderers?
13. **Supermarkets: Free Samples** Do you take the free samples offered in supermarkets? About 60% of all customers will take free samples. Furthermore, of those who take the free samples, about 37% will buy what they have sampled. (See reference in Problem 8.) Suppose you set up a counter in a supermarket offering free samples of a new product. The day you are offering free samples, 317 customers pass by your counter.
(a) What is the probability that more than 180 take your free sample?
(b) What is the probability that fewer than 200 take your free sample?
(c) What is the probability that a customer takes a free sample *and* buys the product? Hint: Use the multiplication rule for *dependent* events. Notice that we are given the conditional probability $P(\text{buy} \mid \text{sample}) = 0.37$, while $P(\text{sample}) = 0.60$.
(d) What is the probability that between 60 and 80 customers will take the free sample *and* buy the product? Hint: Use the probability of success calculated in part (c).
14. **Ice Cream: Flavors** What's your favorite ice cream flavor? For people who buy ice cream, the all-time favorite is still vanilla. About 25% of ice cream sales are vanilla. Chocolate accounts for only 9% of ice cream sales. (See reference in Problem 8.) Suppose that 175 customers go to a grocery store in Cheyenne, Wyoming, today to buy ice cream.
(a) What is the probability that 50 or more will buy vanilla?
(b) What is the probability that 12 or more will buy chocolate?

Darren Baker, 2009. Used under permission from Shutterstock.com



- (c) A customer who buys ice cream is not limited to one container or one flavor. What is the probability that someone who is buying ice cream will buy chocolate or vanilla? *Hint:* Chocolate flavor and vanilla flavor are not mutually exclusive events. Assume that the choice to buy one flavor is independent of the choice to buy another flavor. Then use the multiplication rule for independent events, together with the addition rule for events that are not mutually exclusive, to compute the requested probability. (See Section 4.2.)
- (d) What is the probability that between 50 and 60 customers will buy chocolate or vanilla ice cream? *Hint:* Use the probability of success computed in part (c).
15. **Airline Flights: No-Shows** Based on long experience, an airline has found that about 6% of the people making reservations on a flight from Miami to Denver do not show up for the flight. Suppose the airline overbooks this flight by selling 267 ticket reservations for an airplane with only 255 seats.
- What is the probability that a person holding a reservation will show up for the flight?
 - Let $n = 267$ represent the number of ticket reservations. Let r represent the number of people with reservations who show up for the flight. Which expression represents the probability that a seat will be available for everyone who shows up holding a reservation?
- $$P(255 \leq r); \quad P(r \leq 255); \quad P(r \leq 267); \quad P(r = 255)$$
- (c) Use the normal approximation to the binomial distribution and part (b) to answer the following question: What is the probability that a seat will be available for every person who shows up holding a reservation?
16. **General: Approximations** We have studied *two* approximations to the binomial, the normal approximation and the Poisson approximation (Section 5.4). Write a brief but complete essay in which you discuss and summarize the *conditions* under which each approximation would be used, the *formulas* involved, and the *assumptions* made for each approximation. Give details and examples in your essay. How could you apply these statistical methods in your everyday life?
17. **Statistical Literacy** Under what conditions is it appropriate to use a normal distribution to approximate the \hat{p} distribution?
18. **Statistical Literacy** What is the formula for the standard error of the normal approximation to the \hat{p} distribution? What is the mean of the \hat{p} distribution?
19. **Statistical Literacy** Is \hat{p} an unbiased estimator for p when $np > 5$ and $nq > 5$? Recall that a statistic is an unbiased estimator of the corresponding parameter if the mean of the sampling distribution equals the parameter in question.
20. **Basic Computation: \hat{p} Distribution** Suppose we have a binomial experiment in which success is defined to be a particular quality or attribute that interests us.
- Suppose $n = 33$ and $p = 0.21$. Can we approximate the \hat{p} distribution by a normal distribution? Why? What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?
 - Suppose $n = 25$ and $p = 0.15$. Can we safely approximate the \hat{p} distribution by a normal distribution? Why or why not?
 - Suppose $n = 48$ and $p = 0.15$. Can we approximate the \hat{p} distribution by a normal distribution? Why? What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?
21. **Basic Computation: \hat{p} Distribution** Suppose we have a binomial experiment in which success is defined to be a particular quality or attribute that interests us.
- Suppose $n = 100$ and $p = 0.23$. Can we safely approximate the \hat{p} distribution by a normal distribution? Why? Compute $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$.
 - Suppose $n = 20$ and $p = 0.23$. Can we safely approximate the \hat{p} distribution by a normal distribution? Why or why not?



SUMMARY

Chapter Review

In this chapter, we examined properties and applications of the normal probability distribution.

- A normal probability distribution is a distribution of a continuous random variable. Normal distributions are bell-shaped and symmetric around the mean. The high point occurs over the mean, and most of the area occurs within 3 standard deviations of the mean. The mean and median are equal.

- The empirical rule for normal distributions gives areas within 1, 2, and 3 standard deviations of the mean. Approximately

68% of the data lie within the interval

$$\mu \pm \sigma$$

95% of the data lie within the interval

$$\mu \pm 2\sigma$$

99.7% of the data lie within the interval

$$\mu \pm 3\sigma$$

- For symmetric, bell-shaped distributions,

$$\text{standard deviation} \approx \frac{\text{range of data}}{4}$$

- A z score measures the number of standard deviations a raw score x lies from the mean.

$$z = \frac{x - \mu}{\sigma} \quad \text{and} \quad x = z\sigma + \mu$$

- For the standard normal distribution, $\mu = 0$ and $\sigma = 1$.

- Table 5 of Appendix II gives areas under a standard normal distribution that are to the left of a specified value of z .

- After raw scores x have been converted to z scores, the standard normal distribution table can be used to find probabilities associated with intervals of x values from any normal distribution.

- The inverse normal distribution is used to find z values associated with areas to the left of z . Table 5 of Appendix II can be used to find approximate z values associated with specific probabilities.

- Tools for assessing the normality of a data distribution include:

Histogram of the data. A roughly bell-shaped histogram indicates normality.

Presence of outliers. A limited number indicates normality.

Skewness. For normality, Pearson's index is between -1 and 1 .

Normal quantile plot. For normality, points lie close to a straight line.

- Control charts are an important application of normal distributions.
- Sampling distributions give us the basis for inferential statistics. By studying the distribution of a sample statistic, we can learn about the corresponding population parameter.
- For random samples of size n , the \bar{x} distribution is the sampling distribution for the sample mean of an x distribution with population mean μ and population standard deviation σ . If the x distribution is normal, then the corresponding \bar{x} distribution is normal.

By the central limit theorem, when n is sufficiently large ($n \geq 30$), the \bar{x} distribution is approximately normal even if the original x distribution is not normal.

In both cases,

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- For n binomial trials with probability of success p on each trial, the \hat{p} distribution is the sampling distribution of the sample proportion of successes. When $np > 5$ and $nq > 5$, the \hat{p} distribution is approximately normal with

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

- The binomial distribution can be approximated by a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$ provided

$$np > 5 \text{ and } nq > 5, \text{ with } q = 1 - p$$

and a continuity correction is made.

Data from many applications follow distributions that are approximately normal. We will see normal distributions used extensively in later chapters.

IMPORTANT WORDS & SYMBOLS

Section 6.1

- Normal distributions 250
- Normal curves 251
- Downward cup and upward cup on normal curves 251
- Symmetry of normal curves 251
- Normal density function 252
- Empirical rule 252
- Control chart 255
- Out-of-control signals 256
- Uniform distribution 263
- Exponential distribution 264

Section 6.2

- z value, z score or standard score 267
- Standard units 267
- Raw score, x 268
- Standard normal distribution ($\mu = 0$ and $\sigma = 1$) 269
- Area under the standard normal curve 269

Section 6.3

- Areas under any normal curve 277
- Inverse normal distribution 279

Normality indicators 283

Section 6.4

- Statistic 291
- Population parameter 291
- Sampling distribution 292

Section 6.5

- $\mu_{\bar{x}}$ 297
- $\sigma_{\bar{x}}$ 297
- Standard error of the mean 298
- Central limit theorem 299
- Large sample 299

Section 6.6

- Normal approximation to the binomial distribution 309
- Continuity correction 311
- Sampling distribution for \hat{p} 313
- $\mu_{\hat{p}}$ 313
- $\sigma_{\hat{p}}$ 313
- Standard error of a proportion 313

VIEWPOINT

Nenana Ice Classic

The Nenana Ice Classic is a betting pool offering a large cash prize to the lucky winner who can guess the time, to the nearest minute, of the ice breakup on the Tanana River in the town of Nenana, Alaska. Official breakup time is defined as the time when the surging river dislodges a tripod on the ice. This breaks an attached line and stops a clock set to Yukon Standard Time. The event is so popular that the first state legislature of Alaska (1959) made the Nenana Ice Classic an official statewide lottery. Since 1918, the earliest breakup has been April 20, 1940, at 3:27 P.M., and the latest recorded breakup was May 20, 1964, at 11:41 A.M. Want to make a statistical guess predicting when the ice will break up? Breakup times from the years 1918 to 1996 are recorded in *The Alaska Almanac*, published by Alaska Northwest Books, Anchorage.

CHAPTER REVIEW PROBLEMS

1. | **Statistical Literacy** Describe a normal probability distribution.
2. | **Statistical Literacy** According to the empirical rule, approximately what percentage of the area under a normal distribution lies within 1 standard deviation of the mean? within 2 standard deviations? within 3 standard deviations?
3. | **Statistical Literacy** Is a process in control if the corresponding control chart for data having a normal distribution shows a value beyond 3 standard deviations of the mean?
4. | **Statistical Literacy** Can a normal distribution always be used to approximate a binomial distribution? Explain.
5. | **Statistical Literacy** What characteristic of a normal quantile plot indicates that the data follow a distribution that is approximately normal?

6. **Statistical Literacy** For a normal distribution, is it likely that a data value selected at random is more than 2 standard deviations above the mean?
7. **Statistical Literacy** Give the formula for the *standard error* of the sample mean \bar{x} distribution, based on samples of size n from a distribution with standard deviation σ .
8. **Statistical Literacy** Give the formula for the *standard error* of the sample proportion \hat{p} distribution, based on n binomial trials with probability of success p on each trial.
9. **Critical Thinking** Let x be a random variable representing the amount of sleep each adult in New York City got last night. Consider a sampling distribution of sample means \bar{x} .
 - (a) As the sample size becomes increasingly large, what distribution does the \bar{x} distribution approach?
 - (b) As the sample size becomes increasingly large, what value will the mean $\mu_{\bar{x}}$ of the \bar{x} distribution approach?
 - (c) What value will the standard deviation $\sigma_{\bar{x}}$ of the sampling distribution approach?
 - (d) How do the two \bar{x} distributions for sample size $n = 50$ and $n = 100$ compare?
10. **Critical Thinking** If x has a normal distribution with mean $\mu = 15$ and standard deviation $\sigma = 3$, describe the distribution of \bar{x} values for sample size n , where $n = 4$, $n = 16$ and $n = 100$. How do the \bar{x} distributions compare for the various sample sizes?
11. **Basic Computation: Probability** Given that x is a normal variable with mean $\mu = 47$ and standard deviation $\sigma = 6.2$, find
 - (a) $P(x \leq 60)$
 - (b) $P(x \geq 50)$
 - (c) $P(50 \leq x \leq 60)$
12. **Basic Computation: Probability** Given that x is a normal variable with mean $\mu = 110$ and standard deviation $\sigma = 12$, find
 - (a) $P(x \leq 120)$
 - (b) $P(x \geq 80)$
 - (c) $P(108 \leq x \leq 117)$
13. **Basic Computation: Inverse Normal** Find z such that 5% of the area under the standard normal curve lies to the right of z .
14. **Basic Computation: Inverse Normal** Find z such that 99% of the area under the standard normal curve lies between $-z$ and z .
15. **Medical: Blood Type** Blood type AB is found in only 3% of the population (*Textbook of Medical Physiology*, by A. Guyton, M.D.). If 250 people are chosen at random, what is the probability that
 - (a) 5 or more will have this blood type?
 - (b) between 5 and 10 will have this blood type?
16. **Customer Complaints: Time** The Customer Service Center in a large New York department store has determined that the amount of time spent with a customer about a complaint is normally distributed, with a mean of 9.3 minutes and a standard deviation of 2.5 minutes. What is the probability that for a randomly chosen customer with a complaint, the amount of time spent resolving the complaint will be
 - (a) less than 10 minutes?
 - (b) longer than 5 minutes?
 - (c) between 8 and 15 minutes?
17. **Recycling: Aluminum Cans** One environmental group did a study of recycling habits in a California community. It found that 70% of the aluminum cans sold in the area were recycled.
 - (a) If 400 cans are sold today, what is the probability that 300 or more will be recycled?

- (b) Of the 400 cans sold, what is the probability that between 260 and 300 will be recycled?
18. **Guarantee: Disc Players** Future Electronics makes compact disc players. Its research department found that the life of the laser beam device is normally distributed, with mean 5000 hours and standard deviation 450 hours.
- Find the probability that the laser beam device will wear out in 5000 hours or less.
 - Inverse Normal Distribution** Future Electronics wants to place a guarantee on the players so that no more than 5% fail during the guarantee period. Because the laser pickup is the part most likely to wear out first, the guarantee period will be based on the life of the laser beam device. How many playing hours should the guarantee cover? (Round to the next playing hour.)
19. **Guarantee: Package Delivery** Express Courier Service has found that the delivery time for packages is normally distributed, with mean 14 hours and standard deviation 2 hours.
- For a package selected at random, what is the probability that it will be delivered in 18 hours or less?
 - Inverse Normal Distribution** What should be the guaranteed delivery time on all packages in order to be 95% sure that the package will be delivered before this time? *Hint:* Note that 5% of the packages will be delivered at a time beyond the guaranteed time period.
20. **Control Chart: Landing Gear** Hydraulic pressure in the main cylinder of the landing gear of a commercial jet is very important for a safe landing. If the pressure is not high enough, the landing gear may not lower properly. If it is too high, the connectors in the hydraulic line may spring a leak.
- In-flight landing tests show that the actual pressure in the main cylinders is a variable with mean 819 pounds per square inch and standard deviation 23 pounds per square inch. Assume that these values for the mean and standard deviation are considered safe values by engineers.
- For nine consecutive test landings, the pressure in the main cylinder is recorded as follows:
- | Landing number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Pressure | 870 | 855 | 830 | 815 | 847 | 836 | 825 | 810 | 792 |
- Make a control chart for the pressure in the main cylinder of the hydraulic landing gear, and plot the data on the control chart. Looking at the control chart, would you say the pressure is “in control” or “out of control”? Explain your answer. Identify any out-of-control signals by type (I, II, or III).
- For 10 consecutive test landings, the pressure was recorded on another plane as follows:
- | Landing number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Pressure | 865 | 850 | 841 | 820 | 815 | 789 | 801 | 765 | 730 | 725 |
- Make a control chart and plot the data on the chart. Would you say the pressure is “in control” or not? Explain your answer. Identify any out-of-control signals by type (I, II, or III).
21. **Job Interview: Length** The personnel office at a large electronics firm regularly schedules job interviews and maintains records of the interviews. From the past records, they have found that the length of a first interview is normally distributed, with mean $\mu = 35$ minutes and standard deviation $\sigma = 7$ minutes.
- What is the probability that a first interview will last 40 minutes or longer?
 - Nine first interviews are usually scheduled per day. What is the probability that the average length of time for the nine interviews will be 40 minutes or longer?

Jeff Greenberg/PhotoEdit



22. **Drugs: Effects** A new muscle relaxant is available. Researchers from the firm developing the relaxant have done studies that indicate that the time lapse between administration of the drug and beginning effects of the drug is normally distributed, with mean $\mu = 38$ minutes and standard deviation $\sigma = 5$ minutes.
- The drug is administered to one patient selected at random. What is the probability that the time it takes to go into effect is 35 minutes or less?
 - The drug is administered to a random sample of 10 patients. What is the probability that the average time before it is effective for all 10 patients is 35 minutes or less?
 - Comment on the differences of the results in parts (a) and (b).
23. **Psychology: IQ Scores** Assume that IQ scores are normally distributed, with a standard deviation of 15 points and a mean of 100 points. If 100 people are chosen at random, what is the probability that the sample mean of IQ scores will not differ from the population mean by more than 2 points?
24. **Hatchery Fish: Length** A large tank of fish from a hatchery is being delivered to a lake. The hatchery claims that the mean length of fish in the tank is 15 inches, and the standard deviation is 2 inches. A random sample of 36 fish is taken from the tank. Let \bar{x} be the mean sample length of these fish. What is the probability that \bar{x} is within 0.5 inch of the claimed population mean?
25. **Basic Computation: \hat{p} Distribution** Suppose we have a binomial distribution with $n = 24$ trials and probability of success $p = 0.4$ on each trial. The sample proportion of successes is $\hat{p} = r/n$.
- Is it appropriate to approximate the \hat{p} distribution with a normal distribution? Explain.
 - What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?
26. **Green Behavior: Purchasing Habits** A recent Harris Poll on green behavior showed that 25% of adults often purchase used items instead of new ones. Consider a random sample of 75 adults. Let \hat{p} be the sample proportion of adults who often purchase used instead of new items.
- Is it appropriate to approximate the \hat{p} distribution with a normal distribution? Explain.
 - What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?

DATA HIGHLIGHTS: GROUP PROJECTS



Mike Mazzaschi/Stock Boston

Wild iris

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

Iris setosa is a beautiful wildflower that is found in such diverse places as Alaska, the Gulf of St. Lawrence, much of North America, and even in English meadows and parks. R. A. Fisher, with his colleague Dr. Edgar Anderson, studied these flowers extensively. Dr. Anderson described how he collected information on irises:

I have studied such irises as I could get to see, in as great detail as possible, measuring iris standard after iris standard and iris fall after iris fall, sitting squat-legged with record book and ruler in mountain meadows, in cypress swamps, on lake beaches, and in English parks. [E. Anderson, "The Irises of the Gaspé Peninsula," *Bulletin, American Iris Society*, Vol. 59 pp. 2–5, 1935.]

The data in Table 6-11 were collected by Dr. Anderson and were published by his friend and colleague R. A. Fisher in a paper entitled "The Use of Multiple Measurements in Taxonomic Problems" (*Annals of Eugenics*, part II, pp. 179–188, 1936). To find these data, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to DASL, the Carnegie Mellon University Data and Story Library. From the DASL site, look under famous data sets.

Let x be a random variable representing petal length. Using a TI-84Plus/TI-83Plus/TI-nspire calculator, it was found that the sample mean is $\bar{x} = 1.46$ centimeters (cm) and the sample standard deviation is $s = 0.17$ cm. Figure 6-39 shows a histogram for the given data generated on a TI-84Plus/TI-83Plus/TI-nspire calculator.

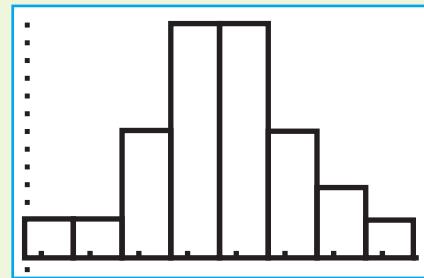
- Examine the histogram for petal lengths. Would you say that the distribution is approximately mound-shaped and symmetrical? Our sample has only 50 irises; if many thousands of irises had been used, do you think the distribution would look even more like a normal curve? Let x be the petal length of *Iris setosa*. Research has shown that x has an approximately normal distribution, with mean $\mu = 1.5$ cm and standard deviation $\sigma = 0.2$ cm.
- Use the empirical rule with $\mu = 1.5$ and $\sigma = 0.2$ to get an interval into which approximately 68% of the petal lengths will fall. Repeat this for 95% and 99.7%. Examine the raw data and compute the percentage of the raw data that actually fall into each of these intervals (the 68% interval, the 95% interval, and the 99.7% interval). Compare your computed percentages with those given by the empirical rule.
- Compute the probability that a petal length is between 1.3 and 1.6 cm. Compute the probability that a petal length is greater than 1.6 cm.
- Suppose that a random sample of 30 irises is obtained. Compute the probability that the average petal length for this sample is between 1.3 and 1.6 cm. Compute the probability that the average petal length is greater than 1.6 cm.
- Compare your answers to parts (c) and (d). Do you notice any differences? Why would these differences occur?

TABLE 6-11

Petal Length in Centimeters for *Iris setosa*

1.4	1.4	1.3	1.5	1.4
1.7	1.4	1.5	1.4	1.5
1.5	1.6	1.4	1.1	1.2
1.5	1.3	1.4	1.7	1.5
1.7	1.5	1	1.7	1.9
1.6	1.6	1.5	1.4	1.6
1.6	1.5	1.5	1.4	1.5
1.2	1.3	1.4	1.3	1.5
1.3	1.3	1.3	1.6	1.9
1.4	1.6	1.4	1.5	1.4

FIGURE 6-39

Petal Length (cm) for *Iris setosa* (TI-84Plus/TI-83Plus/TI-nspire)

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

- If you look up the word *empirical* in a dictionary, you will find that it means “relying on experiment and observation rather than on theory.” Discuss the empirical rule in this context. The empirical rule certainly applies to the normal distribution, but does it also apply to a wide variety of other distributions that are not *exactly* (theoretically) normal? Discuss the terms *mound-shaped* and *symmetrical*. Draw several sketches of distributions that are mound-shaped and symmetrical. Draw sketches of distributions that are not mound-shaped or symmetrical. To which distributions will the empirical rule apply?

2. Why are standard z values so important? Is it true that z values have no units of measurement? Why would this be desirable for comparing data sets with *different* units of measurement? How can we assess differences in quality or performance by simply comparing z values under a standard normal curve? Examine the formula for computing standard z values. Notice that it involves *both* the mean and the standard deviation. Recall that in Chapter 3 we commented that the mean of a data collection is not entirely adequate to describe the data; you need the standard deviation as well. Discuss this topic again in light of what you now know about normal distributions and standard z values.
3. Most companies that manufacture a product have a division responsible for quality control or quality assurance. The purpose of the quality-control division is to make reasonably certain that the products manufactured are up to company standards. Write a brief essay in which you describe how the statistics you have learned so far could be applied to an industrial application (such as control charts and the Antlers Lodge example).
4. Most people would agree that increased information should give better predictions. Discuss how sampling distributions actually enable better predictions by providing more information. Examine Theorem 6.1 again. Suppose that x is a random variable with a *normal* distribution. Then \bar{x} , the sample mean based on random samples of size n , also will have a normal distribution for *any* value of $n = 1, 2, 3, \dots$

What happens to the standard deviation of the \bar{x} distribution as n (the sample size) increases? Consider the following table for different values of n .

n	1	2	3	4	10	50	100
σ/\sqrt{n}	1σ	0.71σ	0.58σ	0.50σ	0.32σ	0.14σ	0.10σ

In this case, “increased information” means a larger sample size n . Give a brief explanation as to why a *large* standard deviation will usually result in poor statistical predictions, whereas a *small* standard deviation usually results in much better predictions. Since the standard deviation of the sampling distribution \bar{x} is σ/\sqrt{n} , we can decrease the standard deviation by increasing n . In fact, if we look at the preceding table, we see that if we use a sample size of only $n = 4$, we cut the standard deviation of \bar{x} by 50% of the standard deviation σ of x . If we were to use a sample size of $n = 100$, we would cut the standard deviation of \bar{x} to 10% of the standard deviation σ of x .

Give the preceding discussion some thought and explain why you should get much better predictions for μ by using \bar{x} from a sample of size n rather than by just using x . Write a brief essay in which you explain why sampling distributions are an important tool in statistics.

5. In a way, the central limit theorem can be thought of as a kind of “grand central station.” It is a connecting hub or center for a great deal of statistical work. We will use it extensively in Chapters 7, 8, and 9. Put in a very elementary way, the central limit theorem states that as the sample size n increases, the distribution of the sample mean \bar{x} will always approach a normal distribution, no matter where the original x variable came from. For most people, it is the complete generality of the central limit theorem that is so awe inspiring: It applies to practically everything. List and discuss at least three variables from everyday life for which you expect the variable x itself *not* to follow a normal or bell-shaped distribution. Then discuss what would happen to the sampling distribution \bar{x} if the sample size were increased. Sketch diagrams of the \bar{x} distributions as the sample size n increases.

USING TECHNOLOGY

Application 1

How can we determine if data originated from a normal distribution? We can look at a stem-and-leaf plot or histogram of the data to check for general symmetry, skewness, clusters of data, or outliers. However, a more sensitive way to check that a distribution is normal is to look at a special graph called a *normal quantile plot* (or a variation of this plot called a *normal probability plot* in some software packages). It really is not feasible to make a normal quantile plot by hand, but statistical software packages provide such plots. A simple version of the basic idea behind normal quantile plots involves the following process:

- Arrange the observed data values in order from smallest to largest, and determine the percentile occupied by each value. For instance, if there are 20 data values, the smallest datum is at the 5% point, the next smallest is at the 10% point, and so on.
- Find the z values that correspond to the percentile points. For instance, the z value that corresponds to the percentile 5% (i.e., percent in the left tail of the distribution) is $z = -1.645$.
- Plot each data value x against the corresponding percentile z score. If the data are close to a normal distribution, the plotted points will lie close to a straight line. (If the data are close to a standard normal distribution, the points will lie close to the line $x = z$.)

The actual process that statistical software packages use to produce the z scores for the data is more complicated.

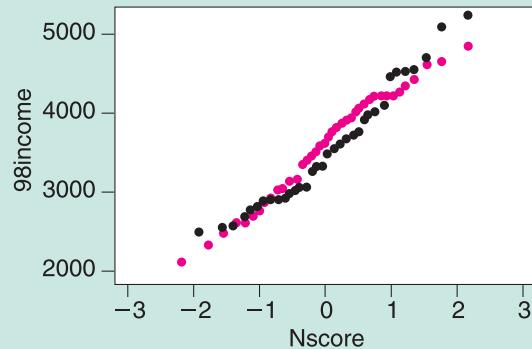
Interpreting normal quantile plots

If the points of a normal quantile plot lie close to a straight line, the plot indicates that the data follow a normal distribution. Systematic deviations from a straight line or bulges in the plot indicate that the data distribution is not normal. Individual points off the line may be outliers.

Consider Figure 6-40. This figure shows Minitab-generated quantile plots for two data sets. The black dots show the normal quantile plot for the salary data of the first application. The red dots show the normal quantile plot for a random sample of 42 data values drawn from a theoretical normal distribution with the same mean and standard deviation as the salary data ($\mu \approx 3421$, $\sigma \approx 709$).

FIGURE 6-40 Normal Quantile Plots

- Salary data for (city) government employees
- A random sample of 42 values from a theoretical normal distribution with the same mean and standard deviation as the salary data



- Do the black dots lie close to a straight line? Do the salaries appear to follow a normal distribution? Are there any outliers on the low or high side? Would you say that any of the salaries are “out of line” for a normal distribution?
- Do the red dots lie close to a straight line? We know the red dots represent a sample drawn from a normal distribution. Is the normal quantile plot for the red dots consistent with this fact? Are there any outliers shown?

Technology Hints

TI-84Plus/TI-83Plus/TI-nspire

Enter the data. Press STATPLOT and select one of the plots. Highlight ON. Then highlight the sixth plot option. To get a plot similar to that of Figure 6-40, choose Y as the data axis.

Minitab

Minitab has several types of normal quantile plots that use different types of scales. To create a normal quantile plot similar to that of Figure 6-40, enter the data in column C1. Then use the menu choices Calc ► Calculator. In the dialogue box listing the functions, scroll to **Normal Scores**. Use **NSCOR(C1)** and store the results in column C2. Finally, use the menu choices Graph ► Plot. In the dialogue box, use C1 for variable y and C2 for variable x .

SPSS

Enter the data. Use the menu choices **Analyze ▶ Descriptive Statistics ▶ Explore**. In the dialogue box, move your data variable to the dependent list. Click **Plots . . .**. Check “Normality plots with tests.” The graph appears in the output window.

Application 2

As we have seen in this chapter, the value of a sample statistic such as \bar{x} varies from one sample to another. The central limit theorem describes the distribution of the sample statistic \bar{x} when samples are sufficiently large.

We can use technology tools to generate samples of the same size from the same population. Then we can look at the statistic \bar{x} for each sample, and the resulting \bar{x} distribution.

Project Illustrating the Central Limit Theorem

Step 1: Generate random samples of specified size n from a population.

The random-number table enables us to sample from the uniform distribution of digits 0 through 9. Use either the random-number table or a random-number generator to generate 30 samples of size 10.

Step 2: Compute the sample mean \bar{x} of the digits in each sample.

Step 3: Compute the sample mean of the means (i.e., $\bar{x}_{\bar{x}}$) as well as the standard deviation $s_{\bar{x}}$ of the sample means.

The population mean of the uniform distribution of digits from 0 through 9 is 4.5. How does $\bar{x}_{\bar{x}}$ compare to this value?

Step 4: Compare the sample distribution of \bar{x} values to a normal distribution having the mean and standard deviation computed in Step 3.

- Use the values of $\bar{x}_{\bar{x}}$ and $s_{\bar{x}}$ computed in Step 3 to create the intervals shown in column 1 of Table 6-12.
- Tally the sample means computed in Step 2 to determine how many fall into each interval of column 2. Then compute the percent of data in each interval and record the results in column 3.
- The percentages listed in column 4 are those from a normal distribution (see Figure 6-3 showing the empirical rule). Compare the percentages in column 3 to those in column 4. How do the sample percentages compare with the hypothetical normal distribution?

Step 5: Create a histogram showing the sample means computed in Step 2.

Look at the histogram and compare it to a normal distribution with the mean and standard deviation of the \bar{x} s (as computed in Step 3).

Step 6: Compare the results of this project to the central limit theorem.

Increase the sample size of Step 1 to 20, 30, and 40 and repeat Steps 1 to 5.

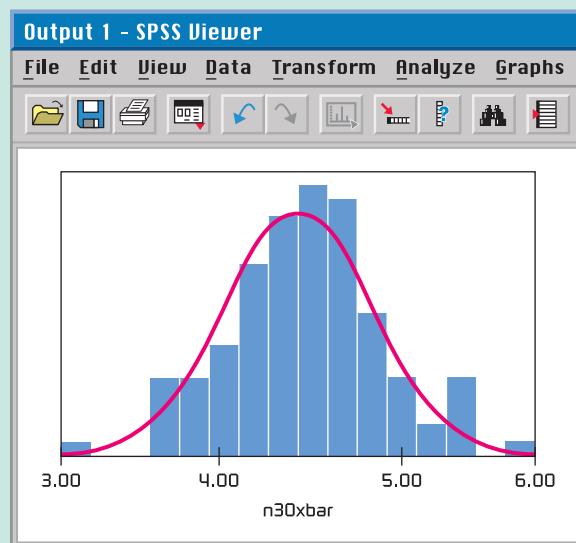
TABLE 6-12 Frequency Table of Sample Means

1. Interval	2. Frequency	3. Percent	4. Hypothetical Normal Distribution
$\bar{x} - 3s$ to $\bar{x} - 2s$	Tally the sample	Compute	2 or 3%
$\bar{x} - 2s$ to $\bar{x} - s$	means computed	percents	13 or 14%
$\bar{x} - s$ to \bar{x}	in step 2 and	from	About 34%
\bar{x} to $\bar{x} + s$	place here.	column 2	About 34%
$\bar{x} + s$ to $\bar{x} + 2s$		and place	13 or 14%
$\bar{x} + 2s$ to $\bar{x} + 3s$		here.	2 or 3%

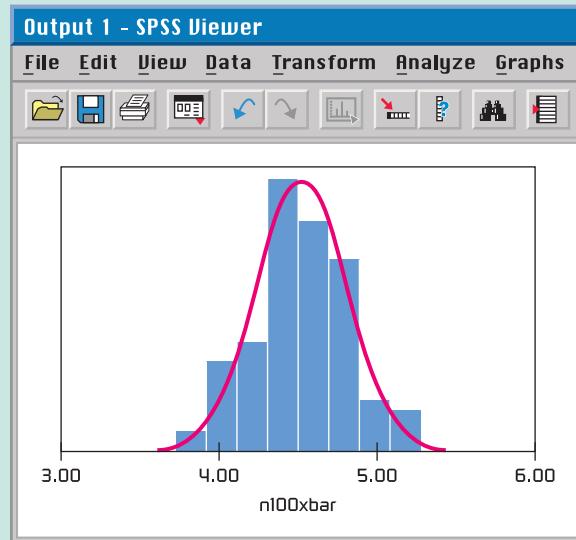
FIGURE 6-41

SPSS-Generated Histograms for Samples of Size 30 and Size 100

(a) $n = 30$



(b) $n = 100$



Technology Hints

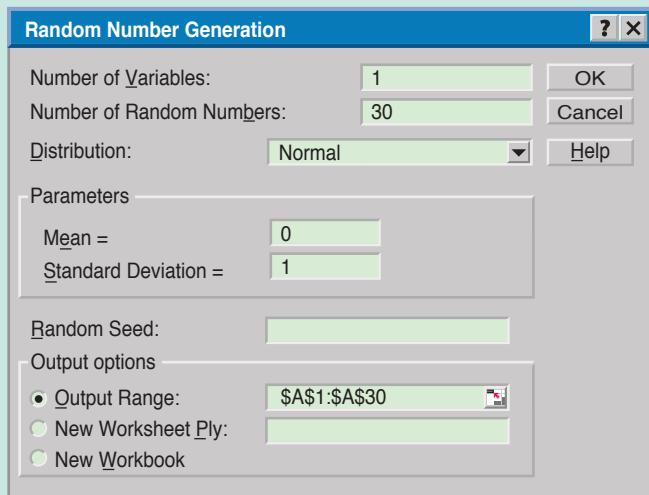
The TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, Minitab, and SPSS all support the process of drawing random samples from a variety of distributions. Macros can be written in Excel 2007, Minitab, and the professional version of SPSS to repeat the six steps of the project. Figure 6-41 shows histograms generated by SPSS for random samples of size 30 and size 100. The samples are taken from a uniform probability distribution.

TI-84Plus/TI-83Plus/TI-nspire

You can generate random samples from uniform, normal, and binomial distributions. Press **MATH** and select **PRB**. Selection 5:**randInt(lower, upper, sample size m)** generates m random integers from the specified interval. Selection 6:**randNorm(μ , σ , sample size m)** generates m random numbers from a normal distribution with mean μ and standard deviation σ . Selection 7:**randBin(number of trials n, p, sample size m)** generates m random values (number of successes out of n trials) for a binomial distribution with probability of success p on each trial. You can put these values in lists by using **Edit** under **Stat**. Highlight the list header, press **Enter**, and then select one of the options discussed.

Excel 2007

On the Home screen, click on the **Data** tab. In the **Analysis** group, select **Data Analysis**. In the dialogue box, select **Random Number Generator**. The next dialogue box provides choices for the population distribution, including uniform, binomial, and normal distributions. Fill in the required parameters and designate the location for the output.



Minitab

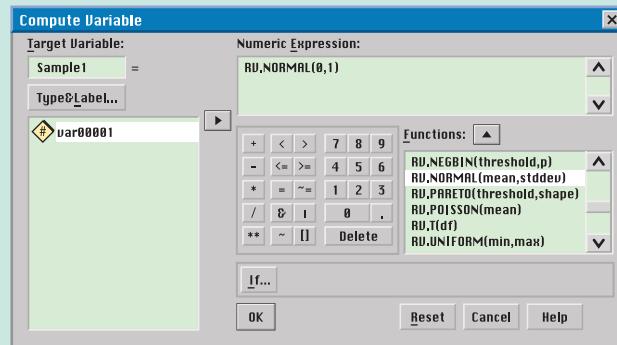
Use the menu selections **Calc > Random Data**. Then select the population distribution. The choices include uniform, binomial, and normal distributions. Fill in the dialogue box, where the number of rows indicates the number of data in the sample.

SPSS

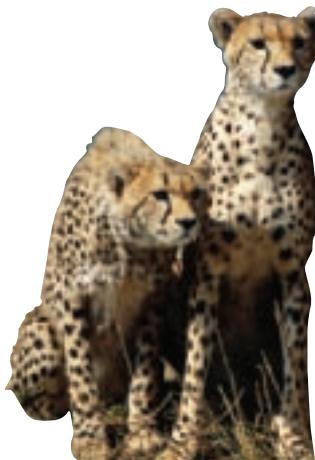
SPSS supports random samples from a variety of distributions, including binomial, normal, and uniform. In data view, generate a column of consecutive integers from 1 to n , where n is the sample size. In variable view, name the variables sample1, sample2, and so on, through sample30. These variables head the columns containing each of the 30 samples of size n . Then use the menu choices **Transform > Compute**. In the dialogue box, use sample1 as the target variable for the first sample, and so forth.

In the function group, select **Random Numbers** and in the functions and special variables group, select **Rv.Uniform**

for samples from a uniform distribution. Functions **Rv.Normal** and **Rv.Binom** provide random samples from normal and binomial distributions, respectively. For each function, the necessary parameters are described.



Hubert Stadler/Corbis



Cumulative Review Problems

CHAPTERS 4–6

The Hill of Tara is located in south-central Meath, not far from Dublin, Ireland. Tara is of great cultural and archaeological importance, since it is by legend the seat of the ancient high kings of Ireland. For more information, see *Tara: An Archaeological Survey*, by Conor Newman, Royal Irish Academy, Dublin.



Hubert Städler/Corbis

Magnetic surveying is one technique used by archaeologists to determine anomalies arising from variations in magnetic susceptibility. Unusual changes in magnetic susceptibility might (or might not) indicate an important archaeological discovery. Let x be a random variable that represents a magnetic susceptibility (MS) reading for a randomly chosen site on the Hill of Tara. A random sample of 120 sites gave the readings shown in Table A below.

TABLE A Magnetic Susceptibility Readings,
centimeter-gram-second $\times 10^{-6}$ (cmg $\times 10^{-6}$)

Comment	Magnetic Susceptibility	Number of Readings	Estimated Probability
"cool"	$0 \leq x < 10$	30	$30/120 = 0.25$
"neutral"	$10 \leq x < 20$	54	$54/120 = 0.45$
"warm"	$20 \leq x < 30$	18	$18/120 = 0.15$
"very interesting"	$30 \leq x < 40$	12	$12/120 = 0.10$
"hot spot"	$40 \leq x$	6	$6/120 = 0.05$

Answers may vary slightly due to rounding.

1. **Statistical Literacy: Sample Space** What is a statistical experiment? How could the magnetic susceptibility intervals $0 \leq x < 10$, $10 \leq x < 20$, and so on, be considered events in the sample space of all possible readings?
2. **Statistical Literacy: Probability** What is probability? What do we mean by relative frequency as a probability estimate for events? What is the law of large numbers? How would the law of large numbers apply in this context?
3. **Statistical Literacy: Probability Distribution** Do the probabilities shown in Table A add up to 1? Why should they total to 1?
4. **Probability Rules** For a site chosen at random, estimate the following probabilities.
(a) $P(0 \leq x < 30)$ (b) $P(10 \leq x < 40)$
(c) $P(x < 20)$ (d) $P(x \geq 20)$
(e) $P(30 \leq x)$ (f) $P(x \text{ not less than } 10)$
(g) $P(0 \leq x < 10 \text{ or } 40 \leq x)$ (h) $P(40 \leq x \text{ and } 20 \leq x)$



5. **Conditional Probability** Suppose you are working in a "warm" region in which all MS readings are 20 or higher. In this same region, what is the probability that you will find a "hot spot" in which the readings are 40 or higher? Use conditional probability to estimate $P(40 \leq x | 20 \leq x)$. Hint: See Problem 39 of Section 6.3.

6. **Discrete Probability Distribution** Consider the midpoint of each interval. Assign the value 45 as the midpoint for the interval $40 \leq x$. The midpoints constitute the sample space for a discrete random variable. Using Table A, compute the expected value μ and the standard deviation σ .

Midpoint x	5	15	25	35	45
$P(x)$					

7. **Binomial Distribution** Suppose a reading between 30 and 40 is called “very interesting” from an archaeological point of view. Let us say you take readings at $n = 12$ sites chosen at random. Let r be a binomial random variable that represents the number of “very interesting” readings from these 12 sites.
- Let us call “very interesting” a binomial success. Use Table A to find p , the probability of success on a single trial, where $p = P(\text{success}) = P(30 \leq x < 40)$.
 - What is the expected value μ and standard deviation σ for the random variable r ?
 - What is the probability that you will find *at least* one “very interesting” reading in the 12 sites?
 - What is the probability that you will find *fewer than* three “very interesting” readings in the 12 sites?
8. **Geometric Distribution** Suppose a “hot spot” is a site with a reading of 40 or higher.
- In a binomial setting, let us call success a “hot spot.” Use Table A to find $p = P(\text{success}) = P(40 \leq x)$ for a single trial.
 - Suppose you decide to take readings at random until you get your *first* “hot spot.” Let n be a random variable representing the trial on which you get your first “hot spot.” Use the geometric probability distribution to write out a formula for $P(n)$.
 - What is the probability that you will need more than four readings to find the first “hot spot”? Compute $P(n > 4)$.
9. **Poisson Approximation to the Binomial** Suppose an archaeologist is looking for geomagnetic “hot spots” in an unexplored region of Tara. As in Problem 8, we have a binomial setting where success is a “hot spot.” In this case, the probability of success is $p = P(40 \leq x)$. The archaeologist takes $n = 100$ magnetic susceptibility readings in the new, unexplored area. Let r be a binomial random variable representing the number of “hot spots” in the 100 readings.
- We want to approximate the binomial random variable r by a Poisson distribution. Is this appropriate? What requirements must be satisfied before we can do this? Do you think these requirements are satisfied in this case? Explain. What is the value of λ ?
 - What is the probability that the archaeologists will find six or fewer “hot spots?” Hint: Use Table 4 of Appendix II.
 - What is the probability that the archaeologists will find more than eight “hot spots”?
10. **Normal Approximation to the Binomial** Consider a binomial setting in which “neutral” is defined to be a success. So, $p = P(\text{success}) = P(10 \leq x < 20)$. Suppose $n = 65$ geomagnetic readings are taken. Let r be a binomial random variable that represents the number of “neutral” geomagnetic readings.
- We want to approximate the binomial random variable r by a normal variable x . Is this appropriate? What requirements must be satisfied before we can do this? Do you think these requirements are satisfied in this case? Explain.
 - What is the probability that there will be at least 20 “neutral” readings out of these 65 trials?
 - Why would the Poisson approximation to the binomial *not* be appropriate in this case? Explain.
11. **Normal Distribution** Oxygen demand is a term biologists use to describe the oxygen needed by fish and other aquatic organisms for survival. The Environmental Protection Agency conducted a study of a wetland area in Marin County, California. In this wetland environment, the mean oxygen demand was $\mu = 9.9$ mg/L with 95% of the data ranging from 6.5 mg/L to 13.3 mg/L (Reference: EPA Report 832-R-93-005). Let x be a random variable that represents oxygen demand in this wetland environment. Assume x has a probability distribution that is approximately normal.
- Use the 95% data range to estimate the standard deviation for oxygen demand. Hint: See Problem 31 of Section 6.3.
 - An oxygen demand below 8 indicates that some organisms in the wetland environment may be dying. What is the probability that the oxygen demand will fall below 8 mg/L?
 - A high oxygen demand can also indicate trouble. An oxygen demand above 12 may indicate an overabundance of organisms that endanger some types of plant life. What is the probability that the oxygen demand will exceed 12 mg/L?
12. **Statistical Literacy** Please give a careful but brief answer to each of the following questions.
- What is a population? How do you get a simple random sample? Give examples.
 - What is a sample statistic? What is a sampling distribution? Give examples.
 - Give a careful and complete statement of the central limit theorem.
 - List at least three areas of everyday life to which the above concepts can be applied. Be specific.



13. ***Sampling Distribution*** Workers at a large toxic cleanup project are concerned that their white blood cell counts may have been reduced. Let x be a random variable that represents white blood cell count per cubic millimeter of whole blood in a healthy adult. Then $\mu = 7500$ and $\sigma \approx 1750$ (Reference: *Diagnostic Tests with Nursing Applications*, S. Loeb). A random sample of $n = 50$ workers from the toxic cleanup site were given a blood test that showed $\bar{x} = 6820$. What is the probability that, for healthy adults, \bar{x} will be this low or lower?
- (a) How does the central limit theorem apply? Explain.
- (b) Compute $P(\bar{x} \leq 6820)$.
- (c) ***Interpretation*** Based on your answer to part (b), would you recommend that additional facts be obtained, or would you recommend that the workers' concerns be dismissed? Explain.
14. ***Sampling Distribution*** Do you have a great deal of confidence in the advice given to you by your medical doctor? About 45% of all adult Americans claim they do have a great deal of confidence in their M.D.s (Reference: *National Opinion Research Center*, University of Chicago). Suppose a random sample of $n = 32$ adults in a health insurance program are asked about their confidence in the medical advice their doctors give.
- (a) Is the normal approximation to the proportion $\hat{p} = r/n$ valid?
- (b) Find the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$.
15. ***Summary*** Write a brief but complete essay in which you describe the probability distributions you have studied so far. Which apply to discrete random variables? Which apply to continuous random variables? Under what conditions can the binomial distribution be approximated by the normal? by the Poisson?



7

- 7.1 Estimating μ When σ Is Known
- 7.2 Estimating μ When σ Is Unknown
- 7.3 Estimating p in the Binomial Distribution
- 7.4 Estimating $\mu_1 - \mu_2$ and $p_1 - p_2$



Library of Congress

*We dance round in a
ring and suppose, But
the Secret sits in the
middle and knows.*

—ROBERT FROST,
“THE SECRET SITS”*



Rusty Jarrett/Getty Images for NASCAR/Getty Images

In Chapter 1, we said that statistics is the study of how to collect, organize, analyze, and interpret numerical data. That part of statistics concerned with analysis, interpretation, and forming conclusions about the source of the data is called *statistical inference*. Problems of statistical inference require us to draw a *random sample* of observations from a larger *population*. A sample usually contains incomplete information, so in a sense we must “dance round in a ring and suppose,” to quote the words of the celebrated American poet Robert Lee Frost (1874–1963). Nevertheless, conclusions about the population can be obtained from sample data by the use of statistical estimates. This chapter introduces you to several widely used methods of estimation.

* “The Secret Sits,” from *The Poetry of Robert Frost*, edited by Edward Connery Lathem. Copyright 1942 by Robert Frost, © 1970 by Lesley Frost Ballantine, © 1969 by Henry Holt and Company, Inc. Reprinted by permission of Henry Holt and Company, Inc.

For online student resources, visit The Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>.

ESTIMATION

PREVIEW QUESTIONS

How do you estimate the expected value of a random variable? What assumptions are needed?

How much confidence should be placed in such estimates? (SECTION 7.1)

At the beginning design stage of a statistical project, how large a sample size should you plan to get? (SECTION 7.1)

What famous statistician worked for Guinness brewing company in Ireland? What has this got to do with constructing estimates from sample data? (SECTION 7.2)

How do you estimate the proportion p of successes in a binomial experiment? How does the normal approximation fit into this process? (SECTION 7.3)

Sometimes even small differences can be extremely important. How do you estimate differences? (SECTION 7.4)

Bobby Deal / RealDealPhoto,
2010/Used under license from
Shutterstock.com



FOCUS PROBLEM

The Trouble with Wood Ducks

The National Wildlife Federation published an article entitled "The Trouble with Wood Ducks" (*National Wildlife*, Vol. 31, No. 5). In this article, wood ducks are described as beautiful birds living in forested areas such as the Pacific Northwest and southeast United States. Because of overhunting and habitat destruction, these birds were in danger of extinction. A federal ban on hunting wood ducks in 1918 helped save the species from extinction. Wood ducks like to nest in tree cavities. However, many such trees were disappearing due to heavy timber cutting. For a period of time it seemed that nesting boxes were the solution to disappearing trees. At first, the wood duck population grew, but after a few seasons, the population declined sharply. Good biology research combined with good statistics provided an answer to this disturbing phenomenon.

Cornell University professors of ecology Paul Sherman and Brad Semel found that the nesting boxes were placed too close to each other. Female wood ducks prefer a secluded nest that is a considerable distance from the next wood duck nest. In fact, female wood duck behavior changed when the nests were too close to each other. Some females would lay their



eggs in another female's nest. The result was too many eggs in one nest. The biologists found that if there were too many eggs in a nest, the proportion of eggs that hatched was considerably reduced. In the long run, this meant a decline in the population of wood ducks.

In their study, Sherman and Semel used two placements of nesting boxes. Group I boxes were well separated from each other and well hidden by available brush. Group II boxes were highly visible and grouped closely together.

In group I boxes, there were a total of 474 eggs, of which a field count showed that about 270 hatched. In group II boxes, there were a total of 805 eggs, of which a field count showed that, again, about 270 hatched.

The material in Chapter 7 will enable us to answer many questions about the hatch ratios of eggs from nests in the two groups.

- (a) Find a point estimate \hat{p}_1 for p_1 , the proportion of eggs that hatch in group I nest box placements. Find a 95% confidence interval for p_1 .
- (b) Find a point estimate \hat{p}_2 for p_2 , the proportion of eggs that hatch in group II nest box placements. Find a 95% confidence interval for p_2 .
- (c) Find a 95% confidence interval for $p_1 - p_2$. Does the interval indicate that the proportion of eggs hatched from group I nest box placements is higher than, lower than, or equal to the proportion of eggs hatched from group II nest boxes?
- (d) What conclusions about placement of nest boxes can be drawn? In the article, additional concerns are raised about the higher cost of placing and maintaining group I nest boxes. Also at issue is the cost efficiency per successful wood duck hatch. Data in the article do not include information that would help us answer questions of cost efficiency. However, the data presented do help us answer questions about the proportions of successful hatches in the two nest box configurations. (See Problem 26 of Section 7.4.)

SECTION 7.1

Estimating μ When σ Is Known

FOCUS POINTS

- Explain the meanings of confidence level, error of estimate, and critical value.
- Find the critical value corresponding to a given confidence level.
- Compute confidence intervals for μ when σ is known. Interpret the results.
- Compute the sample size to be used for estimating a mean μ .

Because of time and money constraints, difficulty in finding population members, and so forth, we usually do not have access to *all* measurements of an *entire* population. Instead we rely on information from a sample.

In this section, we develop techniques for estimating the population mean μ using sample data. We assume the population standard deviation σ is known.

Let's begin by listing some basic assumptions used in the development of our formulas for estimating μ when σ is known.

Assumptions about the random variable x

1. We have a *simple random sample* of size n drawn from a population of x values.
2. The value of σ , the population standard deviation of x , is known.
3. If the x distribution is normal, then our methods work for *any sample size n*.

4. If x has an unknown distribution, then we require a *sample size* $n \geq 30$. However, if the x distribution is distinctly skewed and definitely not mound-shaped, a sample of size 50 or even 100 or higher may be necessary.

Point estimate

An estimate of a population parameter given by a single number is called a *point estimate* for that parameter. It will come as no great surprise that we use \bar{x} (the sample mean) as the point estimate for μ (the population mean).

A point estimate of a population parameter is an estimate of the parameter using a single number.

\bar{x} is the **point estimate** for μ .

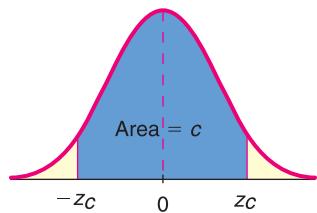
Margin of error

Even with a large random sample, the value of \bar{x} usually is not *exactly* equal to the population mean μ . The *margin of error* is the magnitude of the difference between the sample point estimate and the true population parameter value.

When using \bar{x} as a point estimate for μ , the **margin of error** is the magnitude of $|\bar{x} - \mu|$ or $|\bar{x} - \mu|$.

Confidence level, c

FIGURE 7-1
Confidence Level c and Corresponding Critical Value z_c Shown on the Standard Normal Curve

**Finding the critical value**

We cannot say exactly how close \bar{x} is to μ when μ is unknown. Therefore, the exact margin of error is unknown when the population parameter is unknown. Of course, μ is usually not known, or there would be no need to estimate it. In this section, we will use the language of probability to give us an idea of the size of the margin of error when we use \bar{x} as a point estimate for μ .

First, we need to learn about *confidence levels*. The reliability of an estimate will be measured by the confidence level.

Suppose we want a confidence level of c (see Figure 7-1). Theoretically, we can choose c to be any value between 0 and 1, but usually c is equal to a number such as 0.90, 0.95, or 0.99. In each case, the value z_c is the number such that the area under the standard normal curve falling between $-z_c$ and z_c is equal to c . The value z_c is called the *critical value* for a confidence level of c .

For a confidence level c , the **critical value** z_c is the number such that the area under the standard normal curve between $-z_c$ and z_c equals c .

The area under the normal curve from $-z_c$ to z_c is the probability that the standardized normal variable z lies in that interval. This means that

$$P(-z_c < z < z_c) = c$$

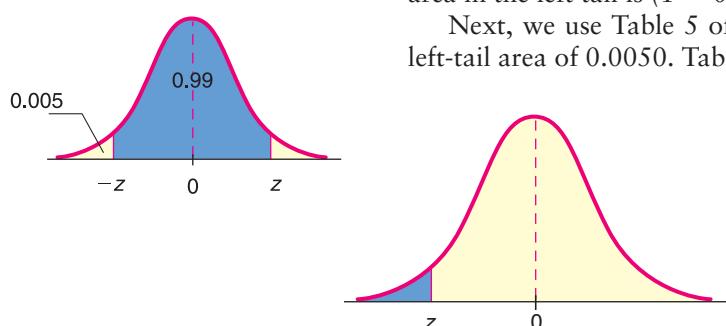
EXAMPLE 1**FIND A CRITICAL VALUE**

Let us use Table 5 of Appendix II to find a number $z_{0.99}$ such that 99% of the area under the standard normal curve lies between $-z_{0.99}$ and $z_{0.99}$. That is, we will find $z_{0.99}$ such that

$$P(-z_{0.99} < z < z_{0.99}) = 0.99$$

SOLUTION: In Section 6.3, we saw how to find the z value when we were given an area between $-z$ and z . The first thing we did was to find the corresponding area

FIGURE 7-2
Area Between $-z$ and z Is 0.99



to the left of $-z$. If A is the area between $-z$ and z , then $(1 - A)/2$ is the area to the left of z . In our case, the area between $-z$ and z is 0.99. The corresponding area in the left tail is $(1 - 0.99)/2 = 0.005$ (see Figure 7-2).

Next, we use Table 5 of Appendix II to find the z value corresponding to a left-tail area of 0.0050. Table 7-1 shows an excerpt from Table 5 of Appendix II.

TABLE 7-1 Excerpt from Table 5 of Appendix II

z	.0007	.08	.09
-3.4	.0003		.0003	.0003	.0002
:					
-2.5	.0062		.0051	.0049	.0048

↑ .0050

From Table 7-1, we see that the desired area, 0.0050, is exactly halfway between the areas corresponding to $z = -2.58$ and $z = -2.57$. Because the two area values are so close together, we use the more conservative z value -2.58 rather than interpolate. In fact, $z_{0.99} \approx 2.576$. However, to two decimal places, we use $z_{0.99} = 2.58$ as the critical value for a confidence level of $c = 0.99$. We have

$$P(-2.58 < z < 2.58) \approx 0.99$$

The results of Example 1 will be used a great deal in our later work. For convenience, Table 7-2 gives some levels of confidence and corresponding critical values z_c . The same information is provided in Table 5(b) of Appendix II.

An estimate is not very valuable unless we have some kind of measure of how “good” it is. The language of probability can give us an idea of the size of the margin of error caused by using the sample mean \bar{x} as an estimate for the population mean.

Remember that \bar{x} is a random variable. Each time we draw a sample of size n from a population, we can get a different value for \bar{x} . According to the central limit theorem, if the sample size is large, then \bar{x} has a distribution that is approximately normal with mean $\mu_{\bar{x}} = \mu$, the population mean we are trying to estimate. The standard deviation is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. If x has a normal distribution, these results are true for any sample size. (See Theorem 6.1.)

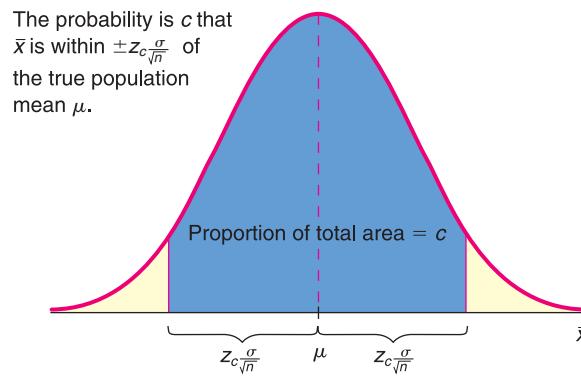
This information, together with our work on confidence levels, leads us (as shown in the optional derivation that follows) to the probability statement

$$P\left(-z_c \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_c \frac{\sigma}{\sqrt{n}}\right) = c \quad (1)$$

TABLE 7-2 Some Levels of Confidence and Their Corresponding Critical Values

Level of Confidence c	Critical Value z_c
0.70, or 70%	1.04
0.75, or 75%	1.15
0.80, or 80%	1.28
0.85, or 85%	1.44
0.90, or 90%	1.645
0.95, or 95%	1.96
0.98, or 98%	2.33
0.99, or 99%	2.58

FIGURE 7-3
Distribution of Sample Means \bar{x}



Equation (1) uses the language of probability to give us an idea of the size of the margin of error for the corresponding confidence level c . In words, Equation (1) states that the probability is c that our point estimate \bar{x} is within a distance $\pm z_c(\sigma/\sqrt{n})$ of the population mean μ . This relationship is shown in Figure 7-3.

In the following optional discussion, we derive Equation (1). If you prefer, you may jump ahead to the discussion about the margin of error.

Optional derivation of Equation (1)

For a c confidence level, we know that

$$P(-z_c < z < z_c) = c \quad (2)$$

This statement gives us information about the size of z , but we want information about the size of $\bar{x} - \mu$. Is there a relationship between z and $\bar{x} - \mu$? The answer is yes since, by the central limit theorem, \bar{x} has a distribution that is approximately normal, with mean μ and standard deviation σ/\sqrt{n} . We can convert \bar{x} to a standard z score by using the formula

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (3)$$

Substituting this expression for z into Equation (2) gives

$$P\left(-z_c < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_c\right) = c \quad (4)$$

Multiplying all parts of the inequality in (4) by σ/\sqrt{n} gives us

$$P\left(-z_c \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_c \frac{\sigma}{\sqrt{n}}\right) = c \quad (1)$$

Equation (1) is precisely the equation we set out to derive.

Maximal margin of error, E

The *margin of error* (or absolute error) using \bar{x} as a point estimate for μ is $|\bar{x} - \mu|$. In most practical problems, μ is unknown, so the margin of error is also unknown. However, Equation (1) allows us to compute an *error tolerance* E that serves as a bound on the margin of error. Using a $c\%$ level of confidence, we can say that the point estimate \bar{x} differs from the population mean μ by a *maximal margin of error*

$$E = z_c \frac{\sigma}{\sqrt{n}} \quad (5)$$

Note: Formula (5) for E is based on the fact that the sampling distribution for \bar{x} is exactly normal, with mean μ and standard deviation σ/\sqrt{n} . This occurs whenever the x distribution is normal with mean μ and standard deviation σ . If the x

distribution is not normal, then according to the central limit theorem, large samples ($n \geq 30$) produce an \bar{x} distribution that is approximately normal, with mean μ and standard deviation σ/\sqrt{n} .

Using Equations (1) and (5), we conclude that

$$P(-E < \bar{x} - \mu < E) = c \quad (6)$$

Equation (6) states that the probability is c that the difference between \bar{x} and μ is no more than the maximal error tolerance E . If we use a little algebra on the inequality

$$-E < \bar{x} - \mu < E \quad (7)$$

for μ , we can rewrite it in the following mathematically equivalent way:

$$\bar{x} - E < \mu < \bar{x} + E \quad (8)$$

Since formulas (7) and (8) are mathematically equivalent, their probabilities are the same. Therefore, from (6), (7), and (8), we obtain

$$P(\bar{x} - E < \mu < \bar{x} + E) = c \quad (9)$$

Equation (9) states that there is a chance c that the interval from $\bar{x} - E$ to $\bar{x} + E$ contains the population mean μ . We call this interval a c confidence interval for μ .

A c confidence interval for μ is an interval computed from sample data in such a way that c is the probability of generating an interval containing the actual value of μ . In other words, c is the proportion of confidence intervals, based on random samples of size n , that actually contain μ .

We may get a different confidence interval for each different sample that is taken. Some intervals will contain the population mean μ and others will not. However, in the long run, the proportion of confidence intervals that contain μ is c .

PROCEDURE

HOW TO FIND A CONFIDENCE INTERVAL FOR μ WHEN σ IS KNOWN

Requirements

Let x be a random variable appropriate to your application. Obtain a simple random sample (of size n) of x values from which you compute the sample mean \bar{x} . The value of σ is already known (perhaps from a previous study).

If you can assume that x has a normal distribution, then any sample size n will work. If you cannot assume this, then use a sample size of $n \geq 30$.

Confidence interval for μ when σ is known

$$\bar{x} - E < \mu < \bar{x} + E \quad (10)$$

where \bar{x} = sample mean of a simple random sample

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

c = confidence level ($0 < c < 1$)

z_c = critical value for confidence level c based on the standard normal distribution (see Table 5(b) of Appendix II for frequently used values).

EXAMPLE 2**CONFIDENCE INTERVAL FOR μ WITH σ KNOWN**

Julia enjoys jogging. She has been jogging over a period of several years, during which time her physical condition has remained constantly good. Usually, she jogs 2 miles per day. The standard deviation of her times is $\sigma = 1.80$ minutes. During the past year, Julia has recorded her times to run 2 miles. She has a random sample of 90 of these times. For these 90 times, the mean was $\bar{x} = 15.60$ minutes. Let μ be the mean jogging time for the entire distribution of Julia's 2-mile running times (taken over the past year). Find a 0.95 confidence interval for μ .

SOLUTION: *Check Requirements* We have a simple random sample of running times, and the sample size $n = 90$ is large enough for the \bar{x} distribution to be approximately normal. We also know σ . It is appropriate to use the normal distribution to compute a confidence interval for μ .

To compute E for the 95% confidence interval $\bar{x} - E$ to $\bar{x} + E$, we use the fact that $z_c = 1.96$ (see Table 7-2), together with the values $n = 90$ and $\sigma = 1.80$. Therefore,

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

$$E = 1.96 \left(\frac{1.80}{\sqrt{90}} \right)$$

$$E \approx 0.37$$

Using Equation (10), the given value of \bar{x} , and our computed value for E , we get the 95% confidence interval for μ .

$$\bar{x} - E < \mu < \bar{x} + E$$

$$15.60 - 0.37 < \mu < 15.60 + 0.37$$

$$15.23 < \mu < 15.97$$

Interpretation We conclude with 95% confidence that the interval from 15.23 minutes to 15.97 minutes is one that contains the population mean μ of jogging times for Julia.

Myrlene Ferguson Cate/PhotoEdit

**CRITICAL THINKING****Interpreting Confidence Intervals**

A few comments are in order about the general meaning of the term *confidence interval*.

- Since \bar{x} is a random variable, the endpoints $\bar{x} \pm E$ are also random variables. Equation (9) states that we have a chance c of obtaining a sample such that the interval, once it is computed, will contain the parameter μ .
- After the confidence interval is numerically fixed for a specific sample, it either does or does not contain μ . So, the probability is 1 or 0 that the interval, when it is fixed, will contain μ .

A nontrivial probability statement can be made only about variables, not constants.

- Equation (9), $P(\bar{x} - E < \mu < \bar{x} + E) = c$, really states that if we draw many random samples of size n and get lots of confidence intervals, then the proportion of all intervals that will turn out to contain the mean μ is c .

For example, in Figure 7-4, on the next page, the horizontal lines represent 0.90 confidence intervals for various samples of the same size from an x distribution. Some of these intervals contain μ and others do not. Since the

intervals are 0.90 confidence intervals, about 90% of all such intervals should contain μ . For each sample, the interval goes from $\bar{x} - E$ to $\bar{x} + E$.

- Once we have a *specific* confidence interval for μ , such as $3 < \mu < 5$, all we can say is that we are $c\%$ confident that we have one of the intervals that actually contains μ . Another appropriate statement is that at the c confidence level, our interval contains μ .

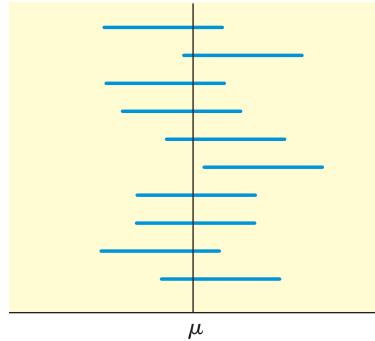


COMMENT Please see Using Technology at the end of this chapter for a computer demonstration of this discussion about confidence intervals.

FIGURE 7-4

0.90 Confidence Intervals for Samples of the Same Size

For each sample, the interval goes from $\bar{x} - E$ to $\bar{x} + E$



GUIDED EXERCISE 1

Confidence interval for μ with σ known

Walter usually meets Julia at the track. He prefers to jog 3 miles. From long experience, he knows that $\sigma = 2.40$ minutes for his jogging times. For a random sample of 90 jogging sessions, the mean time was $\bar{x} = 22.50$ minutes. Let μ be the mean jogging time for the entire distribution of Walter's 3-mile running times over the past several years. Find a 0.99 confidence interval for μ .

- | | |
|---|--|
| (a) Check Requirements Is the \bar{x} distribution approximately normal? Do we know σ ? | → Yes; we know this from the central limit theorem.
Yes, $\sigma = 2.40$ minutes. |
| (b) What is the value of $z_{0.99}$? (See Table 7-2.) | → $z_{0.99} = 2.58$ |
| (c) What is the value of E ? | → $E = z_c \frac{\sigma}{\sqrt{n}} = 2.58 \left(\frac{2.40}{\sqrt{90}} \right) \approx 0.65$ |
| (d) What are the endpoints for a 0.99 confidence interval for μ ? | → The endpoints are given by
$\bar{x} - E \approx 22.50 - 0.65 = 21.85$
$\bar{x} + E \approx 22.50 + 0.65 = 23.15$ |
| (e) Interpretation Explain what the confidence interval tells us. | → We are 99% certain that the interval from 21.85 to 23.15 is an interval that contains the population mean time μ . |

When we use samples to estimate the mean of a population, we generate a small error. However, samples are useful even when it is possible to survey the entire population, because the use of a sample may yield savings of time or effort in collecting data.

LOOKING FORWARD

The basic structure for most confidence intervals for population parameters is

$$\text{sample statistic} - E < \text{population parameter} < \text{sample statistic} + E$$

where E is the maximal margin of error based on the sample statistic distribution and the level of confidence c . We will see this same format used for confidence intervals of the mean when σ is unknown (Section 7.2), for proportions (Section 7.3), for differences of means from independent samples (Section 7.4), for differences of proportions (Section 7.4), and for parameters of linear regression (Chapter 9). This structure for confidence intervals is so basic that some software packages, such as Excel simply give the value of E for a confidence interval and expect the user to finish the computation.

**TECH NOTES**

The TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab all support confidence intervals for μ from large samples. The level of support varies according to the technology. When a confidence interval is given, the standard mathematical notation (lower value, upper value) is used. For instance, the notation (15.23, 15.97) means the interval from 15.23 to 15.97.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) These calculators give the most extensive support. The user can opt to enter raw data or just summary statistics. In each case, the value of σ must be specified. Press the STAT key, then select TESTS, and use 7:ZInterval. The TI-84Plus/TI-83Plus/TI-nspire output shows the results for Example 2.

ZInterval	
Inpt:Data	Stats
σ :	1.8
\bar{x} :	15.6
n:	90
C-Level:	95
Calculate	

ZInterval
(15.228, 15.972)
\bar{x} =15.6
n=90

Excel 2007 Excel gives only the value of the maximal error of estimate E . On the Home screen click the Insert Function f_x . In the dialogue box, select Statistical for the category, and then select Confidence. In the resulting dialogue box, the value of alpha is 1 - confidence level. For example, alpha is 0.05 for a 95% confidence interval. The values of σ and n are also required. The Excel output shows the value of E for Example 2.

f_x	=CONFIDENCE(0.05,1.8,90)
C	D
0.371876	

An alternate approach incorporating raw data (using the Student's t distribution presented in the next section) uses a selection from the Data Analysis package. Click the Data tab on the home ribbon. From the Analysis group, select Data Analysis. In the dialogue box, select Descriptive Statistics. Check the box by Confidence Level for Mean, and enter the confidence level. Again, the value of E for the interval is given.

Minitab Raw data are required. Use the menu choices Stat \blacktriangleright Basic Statistics \blacktriangleright 1-SampleZ.

Sample Size for Estimating the Mean μ

In the design stages of statistical research projects, it is a good idea to decide in advance on the confidence level you wish to use and to select the *maximal* margin of error E you want for your project. How you choose to make these decisions depends on the requirements of the project and the practical nature of the problem.

Whatever specifications you make, the next step is to determine the sample size. Solving the formula that gives the maximal margin of error E for n enables us to determine the minimal sample size.

PROCEDURE

HOW TO FIND THE SAMPLE SIZE n FOR ESTIMATING μ WHEN σ IS KNOWN

Requirements

The distribution of sample means \bar{x} is approximately normal.

Formula for sample size

$$n = \left(\frac{z_c \sigma}{E} \right)^2 \quad (11)$$

where E = specified maximal margin of error

σ = population standard deviation

z_c = critical value from the normal distribution for the desired confidence level c . Commonly used values of z_c can be found in Table 5(b) of Appendix II.

If n is not a whole number, increase n to the next higher whole number. Note that n is the minimal sample size for a specified confidence level and maximal error of estimate E .

COMMENT If you have a *preliminary study* involving a sample size of 30 or larger, then for most practical purposes it is safe to approximate σ with the sample standard deviation s in the formula for sample size.

EXAMPLE 3

SAMPLE SIZE FOR ESTIMATING μ

A wildlife study is designed to find the mean weight of salmon caught by an Alaskan fishing company. A preliminary study of a random sample of 50 salmon showed $s \approx 2.15$ pounds. How large a sample should be taken to be 99% confident that the sample mean \bar{x} is within 0.20 pound of the true mean weight μ ?

SOLUTION: In this problem, $z_{0.99} = 2.58$ (see Table 7-2) and $E = 0.20$. The preliminary study of 50 fish is large enough to permit a good approximation of σ by $s = 2.15$. Therefore, Equation (6) becomes

$$n = \left(\frac{z_c \sigma}{E} \right)^2 \approx \left(\frac{(2.58)(2.15)}{0.20} \right)^2 = 769.2$$

Note: In determining sample size, any fractional value of n is always rounded to the *next higher whole number*. We conclude that a sample size of 770 will be large enough to satisfy the specifications. Of course, a sample size larger than 770 also works.



Salmon moving upstream

VIEWPOINT**Music and Techno Theft**

Performing rights organizations ASCAP (American Society of Composers, Authors, and Publishers) and BMI (Broadcast Music, Inc.) collect royalties for songwriters and music publishers. Radio, television, cable, nightclubs, restaurants, elevators, and even beauty salons play music that is copyrighted by a composer or publisher. The royalty payment for this music turns out to be more than a billion dollars a year (Source: Wall Street Journal). How do ASCAP and BMI know who is playing what music? The answer is, they don't! Instead of tracking exactly what gets played, they use random sampling and confidence intervals. For example, each radio station (there are more than 10,000 in the United States) has randomly chosen days of programming analyzed every year. The results are used to assess royalty fees. In fact, Deloitte & Touche (a financial services company) administers the sampling process.

Although the system is not perfect, it helps bring order into an otherwise chaotic accounting system. Such methods of "copyright policing" help prevent techno theft, ensuring that many songwriters and recording artists get a reasonable return for their creative work.

**SECTION 7.1
PROBLEMS**

In Problems 1–8, answer true or false. Explain your answer.

1. **Statistical Literacy** The value z_c is a value from the standard normal distribution such that $P(-z_c < z < z_c) = c$.
2. **Statistical Literacy** The point estimate for the population mean μ of an x distribution is \bar{x} , computed from a random sample of the x distribution.
3. **Statistical Literacy** Consider a random sample of size n from an x distribution. For such a sample, the margin of error for estimating μ is the magnitude of the difference between \bar{x} and μ .
4. **Statistical Literacy** Every random sample of the same size from a given population will produce exactly the same confidence interval for μ .
5. **Statistical Literacy** A larger sample size produces a longer confidence interval for μ .
6. **Statistical Literacy** If the original x distribution has a relatively small standard deviation, the confidence interval for μ will be relatively short.
7. **Statistical Literacy** If the sample mean \bar{x} of a random sample from an x distribution is relatively small, then the confidence interval for μ will be relatively short.
8. **Statistical Literacy** For the same random sample, when the confidence level c is reduced, the confidence interval for μ becomes shorter.
9. **Critical Thinking** Sam computed a 95% confidence interval for μ from a specific random sample. His confidence interval was $10.1 < \mu < 12.2$. He claims that the probability that μ is in this interval is 0.95. What is wrong with his claim?
10. **Critical Thinking** Sam computed a 90% confidence interval for μ from a specific random sample of size n . He claims that at the 90% confidence level, his confidence interval contains μ . Is his claim correct? Explain.

Answers may vary slightly due to rounding.

11. **Basic Computation: Confidence Interval** Suppose x has a normal distribution with $\sigma = 6$. A random sample of size 16 has sample mean 50.
 - (a) **Check Requirements** Is it appropriate to use a normal distribution to compute a confidence interval for the population mean μ ? Explain.

- (b) Find a 90% confidence interval for μ .
(c) **Interpretation** Explain the meaning of the confidence interval you computed.
12. **Basic Computation: Confidence Interval** Suppose x has a mound-shaped distribution with $\sigma = 9$. A random sample of size 36 has sample mean 20.
(a) **Check Requirements** Is it appropriate to use a normal distribution to compute a confidence interval for the population mean μ ? Explain.
(b) Find a 95% confidence interval for μ .
(c) **Interpretation** Explain the meaning of the confidence interval you computed.
13. **Basic Computation: Sample Size** Suppose x has a mound-shaped distribution with $\sigma = 3$.
(a) Find the minimal sample size required so that for a 95% confidence interval, the maximal margin of error is $E = 0.4$.
(b) **Check Requirements** Based on this sample size, can we assume that the \bar{x} distribution is approximately normal? Explain.
14. **Basic Computation: Sample Size** Suppose x has a normal distribution with $\sigma = 1.2$.
(a) Find the minimal sample size required so that for a 90% confidence interval, the maximal margin of error is $E = 0.5$.
(b) **Check Requirements** Based on this sample size and the x distribution, can we assume that the \bar{x} distribution is approximately normal? Explain.
15. **Zoology: Hummingbirds** Allen's hummingbird (*Selasphorus sasin*) has been studied by zoologist Bill Alther (Reference: *Hummingbirds* by K. Long and W. Alther). A small group of 15 Allen's hummingbirds has been under study in Arizona. The average weight for these birds is $\bar{x} = 3.15$ grams. Based on previous studies, we can assume that the weights of Allen's hummingbirds have a normal distribution, with $\sigma = 0.33$ gram.
(a) Find an 80% confidence interval for the average weights of Allen's hummingbirds in the study region. What is the margin of error?
(b) What conditions are necessary for your calculations?
(c) **Interpret** Compare your results in the context of this problem.
(d) **Sample Size** Find the sample size necessary for an 80% confidence level with a maximal margin of error $E = 0.08$ for the mean weights of the hummingbirds.
16. **Diagnostic Tests: Uric Acid** Overproduction of uric acid in the body can be an indication of cell breakdown. This may be an advance indication of illness such as gout, leukemia, or lymphoma (Reference: *Manual of Laboratory and Diagnostic Tests* by F. Fischbach). Over a period of months, an adult male patient has taken eight blood tests for uric acid. The mean concentration was $\bar{x} = 5.35$ mg/dl. The distribution of uric acid in healthy adult males can be assumed to be normal, with $\sigma = 1.85$ mg/dl.
(a) Find a 95% confidence interval for the population mean concentration of uric acid in this patient's blood. What is the margin of error?
(b) What conditions are necessary for your calculations?
(c) **Interpret** Compare your results in the context of this problem.
(d) **Sample Size** Find the sample size necessary for a 95% confidence level with maximal margin of error $E = 1.10$ for the mean concentration of uric acid in this patient's blood.
17. **Diagnostic Tests: Plasma Volume** Total plasma volume is important in determining the required plasma component in blood replacement therapy for a person undergoing surgery. Plasma volume is influenced by the overall health and physical activity of an individual. (Reference: See Problem 16.) Suppose that a random sample of 45 male firefighters are tested and that they have a plasma

- volume sample mean of $\bar{x} = 37.5$ ml/kg (milliliters plasma per kilogram body weight). Assume that $\sigma = 7.50$ ml/kg for the distribution of blood plasma.
- Find a 99% confidence interval for the population mean blood plasma volume in male firefighters. What is the margin of error?
 - What conditions are necessary for your calculations?
 - Interpret** Compare your results in the context of this problem.
 - Sample Size** Find the sample size necessary for a 99% confidence level with maximal margin of error $E = 2.50$ for the mean plasma volume in male firefighters.
18. **Agriculture: Watermelon** What price do farmers get for their watermelon crops? In the third week of July, a random sample of 40 farming regions gave a sample mean of $\bar{x} = \$6.88$ per 100 pounds of watermelon. Assume that σ is known to be \$1.92 per 100 pounds (Reference: *Agricultural Statistics*, U.S. Department of Agriculture).
- Find a 90% confidence interval for the population mean price (per 100 pounds) that farmers in this region get for their watermelon crop. What is the margin of error?
 - Sample Size** Find the sample size necessary for a 90% confidence level with maximal margin of error $E = 0.3$ for the mean price per 100 pounds of watermelon.
 - A farm brings 15 tons of watermelon to market. Find a 90% confidence interval for the population mean cash value of this crop. What is the margin of error? Hint: 1 ton is 2000 pounds.
19. **FBI Report: Larceny** Thirty small communities in Connecticut (population near 10,000 each) gave an average of $\bar{x} = 138.5$ reported cases of larceny per year. Assume that σ is known to be 42.6 cases per year (Reference: *Crime in the United States*, Federal Bureau of Investigation).
- Find a 90% confidence interval for the population mean annual number of reported larceny cases in such communities. What is the margin of error?
 - Find a 95% confidence interval for the population mean annual number of reported larceny cases in such communities. What is the margin of error?
 - Find a 99% confidence interval for the population mean annual number of reported larceny cases in such communities. What is the margin of error?
 - Compare the margins of error for parts (a) through (c). As the confidence levels increase, do the margins of error increase?
 - Critical Thinking:** Compare the lengths of the confidence intervals for parts (a) through (c). As the confidence levels increase, do the confidence intervals increase in length?
20. **Confidence Intervals: Values of σ** A random sample of size 36 is drawn from an x distribution. The sample mean is 100.
- Suppose the x distribution has $\sigma = 30$. Compute a 90% confidence interval for μ . What is the value of the margin of error?
 - Suppose the x distribution has $\sigma = 20$. Compute a 90% confidence interval for μ . What is the value of the margin of error?
 - Suppose the x distribution has $\sigma = 10$. Compute a 90% confidence interval for μ . What is the value of the margin of error?
 - Compare the margins of error for parts (a) through (c). As the standard deviation decreases, does the margin of error decrease?
 - Critical Thinking** Compare the lengths of the confidence intervals for parts (a) through (c). As the standard deviation decreases, does the length of a 90% confidence interval decrease?
21. **Confidence Intervals: Sample Size** A random sample is drawn from a population with $\sigma = 12$. The sample mean is 30.
- Compute a 95% confidence interval for μ based on a sample of size 49. What is the value of the margin of error?

- (b) Compute a 95% confidence interval for μ based on a sample of size 100. What is the value of the margin of error?
- (c) Compute a 95% confidence interval for μ based on a sample of size 225. What is the value of the margin of error?
- (d) Compare the margins of error for parts (a) through (c). As the sample size increases, does the margin of error decrease?
- (e) **Critical Thinking** Compare the lengths of the confidence intervals for parts (a) through (c). As the sample size increases, does the length of a 90% confidence interval decrease?
22. **Ecology: Sand Dunes** At wind speeds above 1000 centimeters per second (cm/sec), significant sand-moving events begin to occur. Wind speeds below 1000 cm/sec deposit sand, and wind speeds above 1000 cm/sec move sand to new locations. The cyclic nature of wind and moving sand determines the shape and location of large dunes (Reference: *Hydraulic, Geologic, and Biologic Research at Great Sand Dunes National Monument and Vicinity, Colorado*, Proceedings of the National Park Service Research Symposium). At a test site, the prevailing direction of the wind did not change noticeably. However, the velocity did change. Sixty wind speed readings gave an average velocity of $\bar{x} = 1075$ cm/sec. Based on long-term experience, σ can be assumed to be 265 cm/sec.
- (a) Find a 95% confidence interval for the population mean wind speed at this site.
- (b) **Interpretation** Does the confidence interval indicate that the population mean wind speed is such that the sand is always moving at this site? Explain.
23. **Profits: Banks** Jobs and productivity! How do banks rate? One way to answer this question is to examine annual profits per employee. *Forbes Top Companies*, edited by J. T. Davis (John Wiley & Sons), gave the following data about annual profits per employee (in units of one thousand dollars per employee) for representative companies in financial services. Companies such as Wells Fargo, First Bank System, and Key Banks were included. Assume $\sigma \approx 10.2$ thousand dollars.
- | | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|------|
| 42.9 | 43.8 | 48.2 | 60.6 | 54.9 | 55.1 | 52.9 | 54.9 | 42.5 | 33.0 | 33.6 |
| 36.9 | 27.0 | 47.1 | 33.8 | 28.1 | 28.5 | 29.1 | 36.5 | 36.1 | 26.9 | 27.8 |
| 28.8 | 29.3 | 31.5 | 31.7 | 31.1 | 38.0 | 32.0 | 31.7 | 32.9 | 23.1 | 54.9 |
| 43.8 | 36.9 | 31.9 | 25.5 | 23.2 | 29.8 | 22.3 | 26.5 | 26.7 | | |
- (a) Use a calculator or appropriate computer software to verify that, for the preceding data, $\bar{x} \approx 36.0$.
- (b) Let us say that the preceding data are representative of the entire sector of (successful) financial services corporations. Find a 75% confidence interval for μ , the average annual profit per employee for all successful banks.
- (c) **Interpretation** Let us say that you are the manager of a local bank with a large number of employees. Suppose the annual profits per employee are less than 30 thousand dollars per employee. Do you think this might be somewhat low compared with other successful financial institutions? Explain by referring to the confidence interval you computed in part (b).
- (d) **Interpretation** Suppose the annual profits are more than 40 thousand dollars per employee. As manager of the bank, would you feel somewhat better? Explain by referring to the confidence interval you computed in part (b).
- (e) Repeat parts (b), (c), and (d) for a 90% confidence level.
24. **Profits: Retail** Jobs and productivity! How do retail stores rate? One way to answer this question is to examine annual profits per employee. The following data give annual profits per employee (in units of one thousand dollars per employee) for companies in retail sales. (See reference in Problem 23.) Companies such as Gap, Nordstrom, Dillards, JCPenney, Sears, Wal-Mart, Office Depot, and Toys "Я" Us are included. Assume $\sigma \approx 3.8$ thousand dollars.

4.4	6.5	4.2	8.9	8.7	8.1	6.1	6.0	2.6	2.9	8.1	-1.9
11.9	8.2	6.4	4.7	5.5	4.8	3.0	4.3	-6.0	1.5	2.9	4.8
-1.7	9.4	5.5	5.8	4.7	6.2	15.0	4.1	3.7	5.1	4.2	

- (a) Use a calculator or appropriate computer software to verify that, for the preceding data, $\bar{x} \approx 5.1$.
- (b) Let us say that the preceding data are representative of the entire sector of retail sales companies. Find an 80% confidence interval for μ , the average annual profit per employee for retail sales.
- (c) **Interpretation** Let us say that you are the manager of a retail store with a large number of employees. Suppose the annual profits per employee are less than 3 thousand dollars per employee. Do you think this might be low compared with other retail stores? Explain by referring to the confidence interval you computed in part (b).
- (d) **Interpretation** Suppose the annual profits are more than 6.5 thousand dollars per employee. As store manager, would you feel somewhat better? Explain by referring to the confidence interval you computed in part (b).
- (e) Repeat parts (b), (c), and (d) for a 95% confidence interval.
25. **Ballooning: Air Temperature** How hot is the air in the top (crown) of a hot air balloon? Information from *Ballooning: The Complete Guide to Riding the Winds* by Wirth and Young (Random House) claims that the air in the crown should be an average of 100°C for a balloon to be in a state of equilibrium. However, the temperature does not need to be exactly 100°C. What is a reasonable and safe range of temperatures? This range may vary with the size and (decorative) shape of the balloon. All balloons have a temperature gauge in the crown. Suppose that 56 readings (for a balloon in equilibrium) gave a mean temperature of $\bar{x} = 97^\circ\text{C}$. For this balloon, $\sigma \approx 17^\circ\text{C}$.
- (a) Compute a 95% confidence interval for the average temperature at which this balloon will be in a steady-state equilibrium.
- (b) **Interpretation** If the average temperature in the crown of the balloon goes above the high end of your confidence interval, do you expect that the balloon will go up or down? Explain.

Images Etc Ltd /Getty Images



SECTION 7.2

Estimating μ When σ Is Unknown

FOCUS POINTS

- Learn about degrees of freedom and Student's t distributions.
- Find critical values using degrees of freedom and confidence levels.
- Compute confidence intervals for μ when σ is unknown. What does this information tell you?

In order to use the normal distribution to find confidence intervals for a population mean μ , we need to know the value of σ , the population standard deviation. However, much of the time, when μ is unknown, σ is unknown as well. In such cases, we use the sample standard deviation s to approximate σ . When we use s to approximate σ , the sampling distribution for \bar{x} follows a new distribution called a *Student's t distribution*.

Student's t distribution

Student's t Distributions

Student's t distributions were discovered in 1908 by W. S. Gosset. He was employed as a statistician by Guinness brewing company, a company that discouraged publication of research by its employees. As a result, Gosset published

his research under the pseudonym *Student*. Gosset was the first to recognize the importance of developing statistical methods for obtaining reliable information from samples of populations with unknown σ . Gosset used the variable t when he introduced the distribution in 1908. To this day and in his honor, it is still called a Student's t distribution. It might be more fitting to call this distribution *Gosset's t distribution*; however, in the literature of mathematical statistics, it is known as a *Student's t distribution*.

The variable t is defined as follows. A Student's t distribution depends on sample size n .

Assume that x has a normal distribution with mean μ . For samples of size n with sample mean \bar{x} and sample standard deviation s , the t variable

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad (12)$$

Degrees of freedom, $d.f.$

has a Student's t distribution with degrees of freedom $d.f. = n - 1$.

If many random samples of size n are drawn, then we get many t values from Equation (12). These t values can be organized into a frequency table, and a histogram can be drawn, thereby giving us an idea of the shape of the t distribution (for a given n).

Fortunately, all this work is not necessary because mathematical theorems can be used to obtain a formula for the t distribution. However, it is important to observe that these theorems say that the shape of the t distribution depends only on n , provided the basic variable x has a normal distribution. So, *when we use a t distribution, we will assume that the x distribution is normal*.

Table 6 of Appendix II gives values of the variable t corresponding to what we call the number of *degrees of freedom*, abbreviated $d.f.$ For the methods used in this section, the number of degrees of freedom is given by the formula

$$d.f. = n - 1 \quad (13)$$

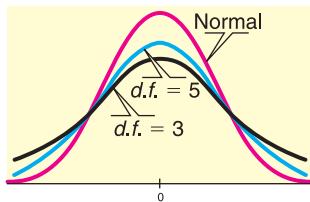
where $d.f.$ stands for the degrees of freedom and n is the sample size. Each choice for $d.f.$ gives a different t distribution.

The graph of a t distribution is always symmetrical about its mean, which (as for the z distribution) is 0. The main observable difference between a t distribution and the standard normal z distribution is that a t distribution has somewhat thicker tails.

Figure 7-5 shows a standard normal z distribution and Student's t distribution with $d.f. = 3$ and $d.f. = 5$.

FIGURE 7-5

A Standard Normal Distribution and Student's t Distribution with $d.f. = 3$ and $d.f. = 5$



Properties of a Student's t distribution

1. The distribution is *symmetric* about the mean 0.
2. The distribution depends on the *degrees of freedom*, $d.f.$ ($d.f. = n - 1$ for μ confidence intervals).
3. The distribution is *bell-shaped*, but has thicker tails than the standard normal distribution.
4. As the degrees of freedom increase, the t distribution *approaches* the standard normal distribution.
5. The area under the entire curve is 1.

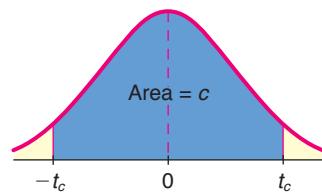
Critical values t_c **FIGURE 7-6**Area Under the t Curve Between $-t_c$ and t_c **Using Table 6 to Find Critical Values for Confidence Intervals**

Table 6 of Appendix II gives various t values for different degrees of freedom $d.f.$. We will use this table to find *critical values* t_c for a c confidence level. In other words, we want to find t_c such that an area equal to c under the t distribution for a given number of degrees of freedom falls between $-t_c$ and t_c . In the language of probability, we want to find t_c such that

$$P(-t_c < t < t_c) = c$$

This probability corresponds to the shaded area in Figure 7-6.

Table 6 of Appendix II has been arranged so that c is one of the column headings, and the degrees of freedom $d.f.$ are the row headings. To find t_c for any specific c , we find the column headed by that c value and read down until we reach the row headed by the appropriate number of degrees of freedom $d.f.$ (You will notice two other column headings: one-tail area and two-tail area. We will use these later, but for the time being, just ignore them.)

Convention for using a Student's t distribution table

If the degrees of freedom $d.f.$ you need are not in the table, use the closest $d.f.$ in the table that is *smaller*. This procedure results in a critical value t_c that is more conservative, in the sense that it is larger. The resulting confidence interval will be longer and have a probability that is slightly higher than c .

EXAMPLE 4**STUDENT'S t DISTRIBUTION**

Use Table 7-3 (an excerpt from Table 6 of Appendix II) to find the critical value t_c for a 0.99 confidence level for a t distribution with sample size $n = 5$.

SOLUTION:

- First, we find the column with c heading 0.990.
- Next, we compute the number of degrees of freedom:
 $d.f. = n - 1 = 5 - 1 = 4$.
- We read down the column under the heading $c = 0.99$ until we reach the row headed by 4 (under $d.f.$). The entry is 4.604. Therefore, $t_{0.99} = 4.604$.

**TABLE 7-3 Student's t Distribution Critical Values
(Excerpt from Table 6, Appendix II)**

one-tail area	—	—	—	—	
two-tail area	—	—	—	—	
<i>d.f.</i>	<i>c</i>	... 0.900	0.950	0.980	0.990 ...
:					
3		... 2.353	3.182	4.541	5.841 ...
4		... 2.132	2.776	3.747	4.604 ...
:					
7		... 1.895	2.365	2.998	3.449 ...
8		... 1.860	2.306	2.896	3.355 ...

GUIDED EXERCISE 2**Student's t distribution table**

Use Table 6 of Appendix II (or Table 7-3, showing an excerpt from the table) to find t_c for a 0.90 confidence level for a t distribution with sample size $n = 9$.

Continued

GUIDED EXERCISE 2 *continued*

- (a) We find the column headed by $c = \underline{\hspace{2cm}}$.  $c = 0.900.$
- (b) The degrees of freedom are given by  $d.f. = n - 1 = 9 - 1 = 8.$
 $d.f. = n - 1 = \underline{\hspace{2cm}}.$
- (c) Read down the column found in part (a) until you reach the entry in the row headed by  $d.f. = 8$. The value of $t_{0.90}$ is $\underline{\hspace{2cm}}$ for a sample of size $n = 9$.
- (d) Find t_c for a 0.95 confidence level for a t distribution with sample size $n = 9$.  $t_{0.95} = 2.306$ for a sample of size $n = 9$.

LOOKING FORWARD

Student's t distributions will be used again in Chapter 8 when testing μ and when testing differences of means. The distributions are also used for confidence intervals and testing of parameters of linear regression (Sections 9.3 and 9.4).

Confidence Intervals for μ When σ Is Unknown

Maximal margin of error, E

In Section 7.1, we found bounds $\pm E$ on the margin of error for a c confidence level. Using the same basic approach, we arrive at the conclusion that

$$E = t_c \frac{s}{\sqrt{n}}$$

is the maximal margin of error for a c confidence level when σ is unknown (i.e., $|\bar{x} - \mu| < E$ with probability c). The analogue of Equation (1) in Section 7.1 is

$$P\left(-t_c \frac{s}{\sqrt{n}} < \bar{x} - \mu < t_c \frac{s}{\sqrt{n}}\right) = c \quad (14)$$

COMMENT Comparing Equation (14) with Equation (1) in Section 7.1, it becomes evident that we are using the same basic method on the t distribution that we used on the z distribution.

Likewise, for samples from normal populations with unknown σ , Equation (9) of Section 7.1 becomes

$$P(\bar{x} - E < \mu < \bar{x} + E) = c \quad (15)$$

where $E = t_c(s/\sqrt{n})$. Let us organize what we have been doing in a convenient summary.

PROCEDURE

Confidence interval for μ with σ unknown

HOW TO FIND A CONFIDENCE INTERVAL FOR μ WHEN σ IS UNKNOWN

Requirements

Let x be a random variable appropriate to your application. Obtain a simple random sample (of size n) of x values from which you compute the sample mean \bar{x} and the sample standard deviation s .

Continued



Problem 23 presents an alternate method for computing confidence intervals when σ is unknown but $n \geq 30$. The alternate method approximates σ by the sample standard deviation s and utilizes the normal distribution.

If you can assume that x has a normal distribution or simply a mound-shaped, symmetric distribution, then any sample size n will work. If you cannot assume this, then use a sample size of $n \geq 30$.

Confidence interval for μ when σ is unknown

$$\bar{x} - E < \mu < \bar{x} + E \quad (16)$$

where \bar{x} = sample mean of a simple random sample

$$E = t_c \frac{s}{\sqrt{n}}$$

c = confidence level ($0 < c < 1$)

t_c = critical value for confidence level c and degrees of freedom

$$d.f. = n - 1$$

(See Table 6 of Appendix II.)

COMMENT In our applications of Student's t distributions, we have made the basic assumption that x has a normal distribution. However, the same methods apply even if x is only approximately normal. In fact, the main requirement for using a Student's t distribution is that the distribution of x values be reasonably symmetrical and mound-shaped. If this is the case, then the methods we employ with the t distribution can be considered valid for most practical applications.

EXAMPLE 5

CONFIDENCE INTERVAL FOR μ , σ UNKNOWN

Suppose an archaeologist discovers seven fossil skeletons from a previously unknown species of miniature horse. Reconstructions of the skeletons of these seven miniature horses show the shoulder heights (in centimeters) to be

45.3 47.1 44.2 46.8 46.5 45.5 47.6

For these sample data, the mean is $\bar{x} \approx 46.14$ and the sample standard deviation is $s \approx 1.19$. Let μ be the mean shoulder height (in centimeters) for this entire species of miniature horse, and assume that the population of shoulder heights is approximately normal.

Find a 99% confidence interval for μ , the mean shoulder height of the entire population of such horses.

SOLUTION: *Check Requirements* We assume that the shoulder heights of the reconstructed skeletons form a random sample of shoulder heights for all the miniature horses of the unknown species. The x distribution is assumed to be approximately normal. Since σ is unknown, it is appropriate to use a Student's t distribution and sample information to compute a confidence interval for μ .

In this case, $n = 7$, so $d.f. = n - 1 = 7 - 1 = 6$. For $c = 0.990$, Table 6 of Appendix II gives $t_{0.99} = 3.707$ (for $d.f. = 6$). The sample standard deviation is $s = 1.19$.

$$E = t_c \frac{s}{\sqrt{n}} = (3.707) \frac{1.19}{\sqrt{7}} \approx 1.67$$

The 99% confidence interval is

$$\bar{x} - E < \mu < \bar{x} + E$$

$$46.14 - 1.67 < \mu < 46.14 + 1.67$$

$$44.5 < \mu < 47.8$$

Interpretation The archaeologist can be 99% confident that the interval from 44.5 cm to 47.8 cm is an interval that contains the population mean μ for shoulder height of this species of miniature horse.



GUIDED EXERCISE 3

Confidence interval for μ , σ unknown

A company has a new process for manufacturing large artificial sapphires. In a trial run, 37 sapphires are produced. The distribution of weights is mound-shaped and symmetric. The mean weight for these 37 gems is $\bar{x} = 6.75$ carats, and the sample standard deviation is $s = 0.33$ carat. Let μ be the mean weight for the distribution of all sapphires produced by the new process.

- (a) **Check Requirements** Is it appropriate to use a Student's t distribution to compute a confidence interval for μ ? → Yes, we assume that the 37 sapphires constitute a simple random sample of all sapphires produced under the new process. The requirement that the x distribution be approximately normal can be dropped since the sample size is large enough to ensure that the \bar{x} distribution is approximately normal.
- (b) What is $d.f.$ for this setting? → $d.f. = n - 1$, where n is the sample size. Since $n = 37$, $d.f. = 37 - 1 = 36$.
- (c) Use Table 6 of Appendix II to find $t_{0.95}$. Note that $d.f. = 36$ is not in the table. Use the $d.f.$ closest to 36 that is *smaller* than 36. → $d.f. = 35$ is the closest $d.f.$ in the table that is *smaller* than 36. Using $d.f. = 35$ and $c = 0.95$, we find $t_{0.95} = 2.030$.
- (d) Find E . →
$$E = t_{0.95} \frac{s}{\sqrt{n}}$$
$$\approx 2.030 \frac{0.33}{\sqrt{37}} \approx 0.11 \text{ carat}$$
- (e) Find a 95% confidence interval for μ . →
$$\bar{x} - E < \mu < \bar{x} + E$$
$$6.75 - 0.11 < \mu < 6.75 + 0.11$$
$$6.64 \text{ carats} < \mu < 6.86 \text{ carats}$$
- (f) **Interpretation** What does the confidence interval tell us in the context of the problem? → The company can be 95% confident that the interval from 6.64 to 6.86 is an interval that contains the population mean weight of sapphires produced by the new process.

We have several formulas for confidence intervals for the population mean μ . How do we choose an appropriate one? We need to look at the sample size, the distribution of the original population, and whether or not the population standard deviation σ is known.

Summary: Confidence intervals for the mean

Assume that you have a random sample of size n from an x distribution and that you have computed \bar{x} and s . A confidence interval for μ is

$$\bar{x} - E < \mu < \bar{x} + E$$

where E is the margin of error. How do you find E ? It depends on how much you know about the x distribution.

Continued

Situation I (most common)

You don't know the population standard deviation σ . In this situation, you use the t distribution with margin of error

$$E = t_c \frac{s}{\sqrt{n}}$$

where degrees of freedom

$$d.f. = n - 1$$

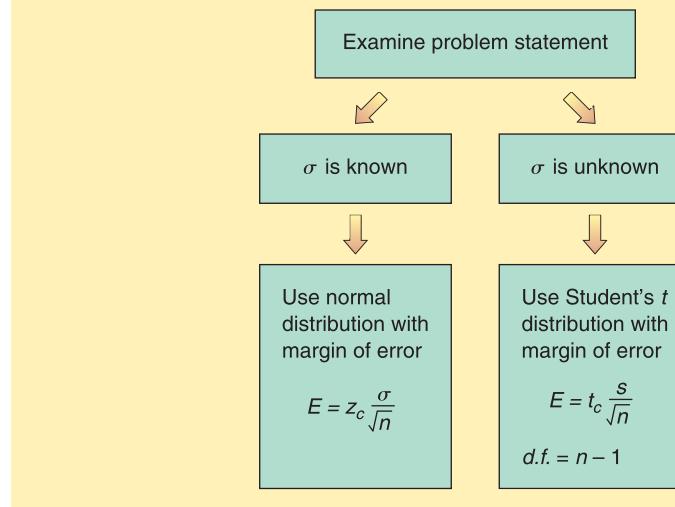
Although a t distribution can be used in many situations, you need to observe some guidelines. If n is less than 30, x should have a distribution that is mound-shaped and approximately symmetric. It's even better if the x distribution is normal. If n is 30 or more, the central limit theorem (Chapter 6) implies that these restrictions can be relaxed.

Situation II (almost never happens!)

You actually know the population value of σ . In addition, you know that x has a normal distribution. If you don't know that the x distribution is normal, then your sample size n must be 30 or larger. In this situation, you use the standard normal z distribution with margin of error

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

Which distribution should you use for \bar{x} ?



COMMENT To find confidence intervals for μ based on small samples, we need to know that the population distribution is approximately normal. What if this is not the case? A procedure called *bootstrap* utilizes computer power to generate an approximation for the \bar{x} sampling distribution. Essentially, the bootstrap method treats the sample as if it were the population. Then, using repetition, it takes many samples (often thousands) from the original sample. This process is called *resampling*. The sample mean \bar{x} is computed for each resample and a distribution of sample means is created. For example, a 95% confidence interval reflects the range for the middle 95% of the bootstrap \bar{x} distribution. If you read Using Technology at the end of this chapter, you will find one (of many) bootstrap methods (Reference: *An Introduction to the Bootstrap* by B. Efron and R. Tibshirani).

**TECH NOTES**

The TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab support confidence intervals using the Student's t distribution.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Press the STAT key, select TESTS, and choose the option 8:TInterval. You may use either raw data in a list or summary statistics.

Excel 2007 Excel gives only the value of the maximal margin of error E . You can easily construct the confidence interval by computing $\bar{x} - E$ and $\bar{x} + E$. On the Home screen, click the Data tab. In the Analysis group, click Data Analysis. In the dialogue box, select Descriptive Statistics. In the dialogue box, check summary statistics and check confidence level for mean. Then set the desired confidence level. Under these choices, Excel uses the Student's t distribution.

Minitab Use the menu choices Stat > Basic Statistics > 1-Sample t. In the dialogue box, indicate the column that contains the raw data. The Minitab output shows the confidence interval for Example 5.

T Confidence Intervals						99.0 % CI
Variable	N	Mean	StDev	SE Mean		
C1	7	46.143	1.190	0.450	(44.475,	47.810)

**VIEWPOINT****Earthquakes!**

California, Washington, Nevada, and even Yellowstone National Park all have earthquakes. Some earthquakes are severe! Earthquakes often bring fear and anxiety to people living near the quake. Is San Francisco due for a really big quake like the 1906 major earthquake? How big are the sizes of recent earthquakes compared with really big earthquakes? What is the duration of an earthquake? How long is the time span between major earthquakes? One way to answer questions such as these is to use existing data to estimate confidence intervals for the average size, duration, and time interval between quakes. Recent data sets for computing such confidence intervals can be found at the National Earthquake Information Service of the U.S. Geological Survey web site. To access the site, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to National Earthquake Information Service.

**SECTION 7.2
PROBLEMS**

1. Use Table 6 of Appendix II to find t_c for a 0.95 confidence level when the sample size is 18.
2. Use Table 6 of Appendix II to find t_c for a 0.99 confidence level when the sample size is 4.
3. Use Table 6 of Appendix II to find t_c for a 0.90 confidence level when the sample size is 22.
4. Use Table 6 of Appendix II to find t_c for a 0.95 confidence level when the sample size is 12.
5. **Statistical Literacy** Student's t distributions are symmetric about a value of t . What is that t value?
6. **Statistical Literacy** As the degrees of freedom increase, what distribution does the Student's t distribution become more like?

7. | **Critical Thinking** Consider a 90% confidence interval for μ . Assume σ is not known. For which sample size, $n = 10$ or $n = 20$, is the critical value t_c larger?
8. | **Critical Thinking** Consider a 90% confidence interval for μ . Assume σ is not known. For which sample size, $n = 10$ or $n = 20$, is the confidence interval longer?
9. | **Critical Thinking** Lorraine computed a confidence interval for μ based on a sample of size 41. Since she did not know σ , she used s in her calculations. Lorraine used the normal distribution for the confidence interval instead of a Student's t distribution. Was her interval longer or shorter than one obtained by using an appropriate Student's t distribution? Explain.
10. | **Critical Thinking** Lorraine was in a hurry when she computed a confidence interval for μ . Because σ was not known, she used a Student's t distribution. However, she accidentally used degrees of freedom n instead of $n - 1$. Was her confidence interval longer or shorter than one found using the correct degrees of freedom $n - 1$? Explain.

Answers may vary slightly due to rounding.

11. | **Basic Computation: Confidence Interval** Suppose x has a mound-shaped distribution. A random sample of size 16 has sample mean 10 and sample standard deviation 2.
 - (a) **Check Requirements** Is it appropriate to use a Student's t distribution to compute a confidence interval for the population mean μ ? Explain.
 - (b) Find a 90% confidence interval for μ .
 - (c) **Interpretation** Explain the meaning of the confidence interval you computed.
12. | **Basic Computation: Confidence Interval** A random sample of size 81 has sample mean 20 and sample standard deviation 3.
 - (a) **Check Requirements** Is it appropriate to use a Student's t distribution to compute a confidence interval for the population mean μ ? Explain.
 - (b) Find a 95% confidence interval for μ .
 - (c) **Interpretation** Explain the meaning of the confidence interval you computed.

In Problems 13–19, assume that the population of x values has an approximately normal distribution.

13. | **Archaeology: Tree Rings** At Burnt Mesa Pueblo, the method of tree-ring dating gave the following years A.D. for an archaeological excavation site (*Bandelier Archaeological Excavation Project: Summer 1990 Excavations at Burnt Mesa Pueblo*, edited by Kohler, Washington State University):

1189 1271 1267 1272 1268 1316 1275 1317 1275

- (a) Use a calculator with mean and standard deviation keys to verify that the sample mean year is $\bar{x} \approx 1272$, with sample standard deviation $s \approx 37$ years.
- (b) Find a 90% confidence interval for the mean of all tree-ring dates from this archaeological site.
- (c) **Interpretation** What does the confidence interval mean in the context of this problem?

14. | **Camping: Cost of a Sleeping Bag** How much does a sleeping bag cost? Let's say you want a sleeping bag that should keep you warm in temperatures from 20°F to 45°F. A random sample of prices (\$) for sleeping bags in this temperature range was taken from *Backpacker Magazine: Gear Guide* (Vol. 25, Issue 157, No. 2). Brand names include American Camper, Cabela's, Camp 7, Caribou, Cascade, and Coleman.

80	90	100	120	75	37	30	23	100	110
105	95	105	60	110	120	95	90	60	70

- (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx \$83.75$ and $s \approx \$28.97$.

- (b) Using the given data as representative of the population of prices of all summer sleeping bags, find a 90% confidence interval for the mean price μ of all summer sleeping bags.
- (c) **Interpretation** What does the confidence interval mean in the context of this problem?
15. **Wildlife: Mountain Lions** How much do wild mountain lions weigh? *The 77th Annual Report of the New Mexico Department of Game and Fish*, edited by Bill Montoya, gave the following information. Adult wild mountain lions (18 months or older) captured and released for the first time in the San Andres Mountains gave the following weights (pounds):
- | | | | | | |
|----|-----|-----|-----|----|----|
| 68 | 104 | 128 | 122 | 60 | 64 |
|----|-----|-----|-----|----|----|
- (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} = 91.0$ pounds and $s \approx 30.7$ pounds.
- (b) Find a 75% confidence interval for the population average weight μ of all adult mountain lions in the specified region.
- (c) **Interpretation** What does the confidence interval mean in the context of this problem?
16. **Franchise: Candy Store** Do you want to own your own candy store? With some interest in running your own business and a decent credit rating, you can probably get a bank loan on startup costs for franchises such as Candy Express, The Fudge Company, Karmel Corn, and Rocky Mountain Chocolate Factory. Startup costs (in thousands of dollars) for a random sample of candy stores are given below (Source: *Entrepreneur Magazine*, Vol. 23, No. 10).
- | | | | | | | | | |
|----|-----|-----|----|----|----|-----|-----|----|
| 95 | 173 | 129 | 95 | 75 | 94 | 116 | 100 | 85 |
|----|-----|-----|----|----|----|-----|-----|----|
- (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 106.9$ thousand dollars and $s \approx 29.4$ thousand dollars.
- (b) Find a 90% confidence interval for the population average startup costs μ for candy store franchises.
- (c) **Interpretation** What does the confidence interval mean in the context of this problem?
17. **Diagnostic Tests: Total Calcium** Over the past several months, an adult patient has been treated for tetany (severe muscle spasms). This condition is associated with an average total calcium level below 6 mg/dl (Reference: *Manual of Laboratory and Diagnostic Tests* by F. Fischbach). Recently, the patient's total calcium tests gave the following readings (in mg/dl).
- | | | | | | | |
|-----|------|------|-----|-----|-----|------|
| 9.3 | 8.8 | 10.1 | 8.9 | 9.4 | 9.8 | 10.0 |
| 9.9 | 11.2 | 12.1 | | | | |
- (a) Use a calculator to verify that $\bar{x} = 9.95$ and $s \approx 1.02$.
- (b) Find a 99.9% confidence interval for the population mean of total calcium in this patient's blood.
- (c) **Interpretation** Based on your results in part (b), does it seem that this patient still has a calcium deficiency? Explain.
18. **Hospitals: Charity Care** What percentage of hospitals provide at least some charity care? The following problem is based on information taken from *State Health Care Data: Utilization, Spending, and Characteristics* (American Medical Association). Based on a random sample of hospital reports from eastern states, the following information was obtained (units in percentage of hospitals providing at least some charity care):
- | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| 57.1 | 56.2 | 53.0 | 66.1 | 59.0 | 64.7 | 70.1 | 64.7 | 53.5 | 78.2 |
|------|------|------|------|------|------|------|------|------|------|
- (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 62.3\%$ and $s \approx 8.0\%$.

- (b) Find a 90% confidence interval for the population average μ of the percentage of hospitals providing at least some charity care.
- (c) **Interpretation** What does the confidence interval mean in the context of this problem?
19. **Critical Thinking: Boxplots and Confidence Intervals** The distribution of heights of 18-year-old men in the United States is approximately normal, with mean 68 inches and standard deviation 3 inches (U.S. Census Bureau). In Minitab, we can simulate the drawing of random samples of size 20 from this population (\triangleright Calc \triangleright Random Data \triangleright Normal, with 20 rows from a distribution with mean 68 and standard deviation 3). Then we can have Minitab compute a 95% confidence interval and draw a boxplot of the data (\triangleright Stat \triangleright Basic Statistics \triangleright 1—Sample t, with boxplot selected in the graphs). The boxplots and confidence intervals for four different samples are shown in the accompanying figures. The four confidence intervals are
- | VARIABLE | N | MEAN | STDEV | SEMEAN | 95.0 % CI |
|----------|----|--------|-------|--------|-------------------|
| Sample 1 | 20 | 68.050 | 2.901 | 0.649 | (66.692 , 69.407) |
| Sample 2 | 20 | 67.958 | 3.137 | 0.702 | (66.490 , 69.426) |
| Sample 3 | 20 | 67.976 | 2.639 | 0.590 | (66.741 , 69.211) |
| Sample 4 | 20 | 66.908 | 2.440 | 0.546 | (65.766 , 68.050) |
- (a) Examine the figure [parts (a) to (d)]. How do the boxplots for the four samples differ? Why should you expect the boxplots to differ?
- 95% Confidence Intervals for Mean Height of 18-Year-Old Men (Sample size 20)**
- (a) Boxplot of Sample 1
(with 95% t-confidence interval for the mean)
-
- (b) Boxplot of Sample 2
(with 95% t-confidence interval for the mean)
-
- (c) Boxplot of Sample 3
(with 95% t-confidence interval for the mean)
-
- (d) Boxplot of Sample 4
(with 95% t-confidence interval for the mean)
-
- (b) Examine the 95% confidence intervals for the four samples shown in the printout. Do the intervals differ in length? Do the intervals all contain the expected population mean of 68 inches? If we draw more samples, do you expect all of the resulting 95% confidence intervals to contain $\mu = 68$? Why or why not?
20. **Crime Rate: Denver** The following data represent crime rates per 1000 population for a random sample of 46 Denver neighborhoods (Reference: *The Piton Foundation*, Denver, Colorado).



63.2	36.3	26.2	53.2	65.3	32.0	65.0
66.3	68.9	35.2	25.1	32.5	54.0	42.4
77.5	123.2	66.3	92.7	56.9	77.1	27.5
69.2	73.8	71.5	58.5	67.2	78.6	33.2
74.9	45.1	132.1	104.7	63.2	59.6	75.7
39.2	69.9	87.5	56.0	154.2	85.5	77.5
84.7	24.2	37.5	41.1			

- (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 64.2$ and $s \approx 27.9$ crimes per 1000 population.
- (b) Let us say that the preceding data are representative of the population crime rates in Denver neighborhoods. Compute an 80% confidence interval for μ , the population mean crime rate for all Denver neighborhoods.
- (c) **Interpretation** Suppose you are advising the police department about police patrol assignments. One neighborhood has a crime rate of 57 crimes per 1000 population. Do you think that this rate is below the average population crime rate and that fewer patrols could safely be assigned to this neighborhood? Use the confidence interval to justify your answer.
- (d) **Interpretation** Another neighborhood has a crime rate of 75 crimes per 1000 population. Does this crime rate seem to be higher than the population average? Would you recommend assigning more patrols to this neighborhood? Use the confidence interval to justify your answer.
- (e) Repeat parts (b), (c), and (d) for a 95% confidence interval.
- (f) **Check Requirement** In previous problems, we assumed the x distribution was normal or approximately normal. Do we need to make such an assumption in this problem? Why or why not? *Hint:* See the central limit theorem in Section 6.5.

21. **Finance: P/E Ratio** The price of a share of stock divided by a company's estimated future earnings per share is called the P/E ratio. High P/E ratios usually indicate "growth" stocks, or maybe stocks that are simply overpriced. Low P/E ratios indicate "value" stocks or bargain stocks. A random sample of 51 of the largest companies in the United States gave the following P/E ratios (Reference: *Forbes*).

11	35	19	13	15	21	40	18	60	72	9	20
29	53	16	26	21	14	21	27	10	12	47	14
33	14	18	17	20	19	13	25	23	27	5	16
8	49	44	20	27	8	19	12	31	67	51	26
19	18	32									

- (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 25.2$ and $s \approx 15.5$.
- (b) Find a 90% confidence interval for the P/E population mean μ of all large U.S. companies.
- (c) Find a 99% confidence interval for the P/E population mean μ of all large U.S. companies.
- (d) **Interpretation** Bank One (now merged with J.P. Morgan) had a P/E of 12, AT&T Wireless had a P/E of 72, and Disney had a P/E of 24. Examine the confidence intervals in parts (b) and (c). How would you describe these stocks at the time the sample was taken?
- (e) **Check Requirements** In previous problems, we assumed the x distribution was normal or approximately normal. Do we need to make such an assumption in this problem? Why or why not? *Hint:* See the central limit theorem in Section 6.5.

22. **Baseball: Home Run Percentage** The home run percentage is the number of home runs per 100 times at bat. A random sample of 43 professional baseball players gave the following data for home run percentages (Reference: *The Baseball Encyclopedia*, Macmillan).

1.6	2.4	1.2	6.6	2.3	0.0	1.8	2.5	6.5	1.8
2.7	2.0	1.9	1.3	2.7	1.7	1.3	2.1	2.8	1.4
3.8	2.1	3.4	1.3	1.5	2.9	2.6	0.0	4.1	2.9
1.9	2.4	0.0	1.8	3.1	3.8	3.2	1.6	4.2	0.0
1.2	1.8	2.4							

- (a) Use a calculator with mean and standard deviation keys to verify that $\bar{x} \approx 2.29$ and $s \approx 1.40$.
- (b) Compute a 90% confidence interval for the population mean μ of home run percentages for all professional baseball players. *Hint:* If you use Table 6 of Appendix II, be sure to use the closest *d.f.* that is *smaller*.
- (c) Compute a 99% confidence interval for the population mean μ of home run percentages for all professional baseball players.
- (d) **Interpretation** The home run percentages for three professional players are

Tim Huelett, 2.5 Herb Hunter, 2.0 Jackie Jensen, 3.8

Examine your confidence intervals and describe how the home run percentages for these players compare to the population average.

- (e) **Check Requirements** In previous problems, we assumed the x distribution was normal or approximately normal. Do we need to make such an assumption in this problem? Why or why not? *Hint:* See the central limit theorem in Section 6.5.



23.

- Expand Your Knowledge: Alternate Method for Confidence Intervals** When σ is unknown and the sample is of size $n \geq 30$, there are two methods for computing confidence intervals for μ .

Method 1: Use the Student's t distribution with $d.f. = n - 1$.

This is the method used in the text. It is widely employed in statistical studies. Also, most statistical software packages use this method.

Method 2: When $n \geq 30$, use the sample standard deviation s as an estimate for σ , and then use the standard normal distribution.

This method is based on the fact that for large samples, s is a fairly good approximation for σ . Also, for large n , the critical values for the Student's t distribution approach those of the standard normal distribution.

Consider a random sample of size $n = 31$, with sample mean $\bar{x} = 45.2$ and sample standard deviation $s = 5.3$.

- (a) Compute 90%, 95%, and 99% confidence intervals for μ using Method 1 with a Student's t distribution. Round endpoints to two digits after the decimal.
- (b) Compute 90%, 95%, and 99% confidence intervals for μ using Method 2 with the standard normal distribution. Use s as an estimate for σ . Round endpoints to two digits after the decimal.
- (c) Compare intervals for the two methods. Would you say that confidence intervals using a Student's t distribution are more conservative in the sense that they tend to be longer than intervals based on the standard normal distribution?
- (d) Repeat parts (a) through (c) for a sample of size $n = 81$. With increased sample size, do the two methods give respective confidence intervals that are more similar?

SECTION 7.3

Estimating p in the Binomial Distribution

FOCUS POINTS

- Compute the maximal margin of error for proportions using a given level of confidence.
- Compute confidence intervals for p and interpret the results.
- Interpret poll results.
- Compute the sample size to be used for estimating a proportion p when we have an estimate for p .
- Compute the sample size to be used for estimating a proportion p when we have no estimate for p .

Basic criteria

The binomial distribution is completely determined by the number of trials n and the probability p of success on a single trial. For most experiments, the number of trials is chosen in advance. Then the distribution is completely determined by p . In this section, we will consider the problem of estimating p under the assumption that n has already been selected.

We are employing what are called *large-sample methods*. We will assume that the normal curve is a good approximation to the binomial distribution, and when necessary, we will use sample estimates for the standard deviation. Empirical studies have shown that these methods are quite good, provided *both*

Basic criteria

$$np > 5 \quad \text{and} \quad nq > 5, \quad \text{where } q = 1 - p$$

Point estimate for p

Let r be the number of successes out of n trials in a binomial experiment. We will take the sample proportion of successes \hat{p} (read “ p hat”) = r/n as our *point estimate for p* , the population proportion of successes.

The point estimates for p and q are

$$\begin{aligned}\hat{p} &= \frac{r}{n} \\ \hat{q} &= 1 - \hat{p}\end{aligned}$$

where n = number of trials and r = number of successes.

Margin of error

For example, suppose that 800 students are selected at random from a student body of 20,000 and that they are each given a shot to prevent a certain type of flu. These 800 students are then exposed to the flu, and 600 of them do not get the flu. What is the probability p that the shot will be successful for any single student selected at random from the entire population of 20,000 students? We estimate p for the entire student body by computing r/n from the sample of 800 students. The value $\hat{p} = r/n$ is 600/800, or 0.75. The value $\hat{p} = 0.75$ is then the *point estimate for p* .

The difference between the actual value of p and the estimate \hat{p} is the size of our error caused by using \hat{p} as a point estimate for p . The magnitude of $\hat{p} - p$ is called the *margin of error* for using $\hat{p} = r/n$ as a point estimate for p . In absolute value notation, the margin of error is $|\hat{p} - p|$.

To compute the bounds for the margin of error, we need some information about the distribution of $\hat{p} = r/n$ values for different samples of the same size n . It turns out that, for large samples, the distribution of \hat{p} values is well approximated by a *normal curve* with

$$\text{mean } \mu = p \quad \text{and} \quad \text{standard error } \sigma = \sqrt{pq/n}$$

Since the distribution of $\hat{p} = r/n$ is approximately normal, we use features of the standard normal distribution to find the bounds for the difference $\hat{p} - p$. Recall that z_c is the number such that an area equal to c under the standard normal curve falls between $-z_c$ and z_c . Then, in terms of the language of probability,

$$P\left(-z_c \sqrt{\frac{pq}{n}} < \hat{p} - p < z_c \sqrt{\frac{pq}{n}}\right) = c \quad (17)$$

Equation (17) says that the chance is c that the numerical difference between \hat{p} and p is between $-z_c \sqrt{pq/n}$ and $z_c \sqrt{pq/n}$. With the c confidence level, our estimate \hat{p} differs from p by no more than

Maximal margin of error, E

$$E = z_c \sqrt{pq/n}$$

As in Section 7.1, we call E the *maximal margin of error*.

Optional derivation of Equation (17)

First, we need to show that $\hat{p} = r/n$ has a distribution that is approximately normal, with $\mu = p$ and $\sigma = \sqrt{pq/n}$. From Section 6.6, we know that for sufficiently large n , the binomial distribution can be approximated by a normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. If r is the number of successes out of n trials of a binomial experiment, then r is a binomial random variable with a binomial distribution. When we convert r to standard z units, we obtain

$$z = \frac{r - \mu}{\sigma} = \frac{r - np}{\sqrt{npq}}$$

For sufficiently large n , r will be approximately normally distributed, so z will be too.

If we divide both numerator and denominator of the last expression by n , the value of z will not change.

$$z = \frac{\frac{r - np}{n}}{\frac{\sqrt{npq}}{n}} \quad \text{Simplified, we find } z = \frac{\frac{r}{n} - p}{\sqrt{\frac{pq}{n}}} \quad (18)$$

Equation (18) tells us that the $\hat{p} = r/n$ distribution is approximated by a normal curve with $\mu = p$ and $\sigma = \sqrt{pq/n}$.

The probability is c that z lies in the interval between $-z_c$ and z_c because an area equal to c under the standard normal curve lies between $-z_c$ and z_c . Using the language of probability, we write

$$P(-z_c < z < z_c) = c$$

From Equation (18), we know that

$$z = \frac{\frac{\hat{p} - p}{\sqrt{pq}}}{\sqrt{\frac{pq}{n}}}$$

If we put this expression for z into the preceding equation, we obtain

$$P\left(-z_c < \frac{\frac{\hat{p} - p}{\sqrt{pq}}}{\sqrt{\frac{pq}{n}}} < z_c\right) = c$$

If we multiply all parts of the inequality by $\sqrt{pq/n}$, we obtain the equivalent statement

$$P\left(-z_c \sqrt{\frac{pq}{n}} < \hat{p} - p < z_c \sqrt{\frac{pq}{n}}\right) = c \quad (17)$$

Confidence interval for p

To find a c confidence interval for p , we will use E in place of the expression $z_c \sqrt{pq/n}$ in Equation (17). Then we get

$$P(-E < \hat{p} - p < E) = c \quad (19)$$

Some algebraic manipulation produces the mathematically equivalent statement

$$P(\hat{p} - E < p < \hat{p} + E) = c \quad (20)$$

Equation (20) states that the probability is c that p lies in the interval from $\hat{p} - E$ to $\hat{p} + E$. Therefore, the interval from $\hat{p} - E$ to $\hat{p} + E$ is the c confidence interval for p that we wanted to find.

There is one technical difficulty in computing the c confidence interval for p . The expression $E = z_c \sqrt{pq/n}$ requires that we know the values of p and q . In most situations, we will not know the actual values of p or q , so we will use our point estimates

$$p \approx \hat{p} \quad \text{and} \quad q = 1 - p \approx 1 - \hat{p}$$

to estimate E . These estimates are reliable for most practical purposes, since we are dealing with large-sample theory ($np > 5$ and $nq > 5$).

For convenient reference, we'll summarize the information about c confidence intervals for p , the probability of success in a binomial distribution.

PROCEDURE**HOW TO FIND A CONFIDENCE INTERVAL FOR A PROPORTION p** **Requirements**

Consider a binomial experiment with n trials, where p represents the population probability of success on a single trial and $q = 1 - p$ represents the population probability of failure. Let r be a random variable that represents the number of successes out of the n binomial trials.

The point estimates for p and q are

$$\hat{p} = \frac{r}{n} \quad \text{and} \quad \hat{q} = 1 - \hat{p}$$

The number of trials n should be sufficiently large so that both $n\hat{p} > 5$ and $n\hat{q} > 5$.

Confidence interval for p

$$\hat{p} - E < p < \hat{p} + E$$

$$\text{where } E \approx z_c \sqrt{\frac{\hat{p}\hat{q}}{n}} = z_c \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

c = confidence level ($0 < c < 1$)

z_c = critical value for confidence level c based on the standard normal distribution (See Table 5(b) of Appendix II for frequently used values.)



Problem 28 of this section shows the method for computing a "plus four confidence interval for p ." This is an alternate method that generally results in a slightly smaller confidence interval than the ones computed with standard methods presented in this section and used in most statistical technology packages.

COMMENT Problem 6 asks you to show that the two conditions $n\hat{p} > 5$ and $n\hat{q} > 5$ are equivalent to the two conditions that the number of successes $r > 5$ and the number of failures $n - r > 5$.

EXAMPLE 6**CONFIDENCE INTERVAL FOR p**

Let's return to our flu shot experiment described at the beginning of this section. Suppose that 800 students were selected at random from a student body of 20,000 and given shots to prevent a certain type of flu. All 800 students were exposed to the flu, and 600 of them did not get the flu. Let p represent the probability that the shot will be successful for any single student selected at random from the entire population of 20,000. Let q be the probability that the shot is not successful.



Stock Boston

- (a) What is the number of trials n ? What is the value of r ?

SOLUTION: Since each of the 800 students receiving the shot may be thought of as a trial, then $n = 800$, and $r = 600$ is the number of successful trials.

- (b) What are the point estimates for p and q ?

SOLUTION: We estimate p by the sample point estimate

$$\hat{p} = \frac{r}{n} = \frac{600}{800} = 0.75$$

We estimate q by

$$\hat{q} = 1 - \hat{p} = 1 - 0.75 = 0.25$$

- (c) **Check Requirements** Would it seem that the number of trials is large enough to justify a normal approximation to the binomial?

SOLUTION: Since $n = 800$, $p \approx 0.75$, and $q \approx 0.25$, then

$$np \approx (800)(0.75) = 600 > 5 \quad \text{and} \quad np \approx (800)(0.25) = 200 > 5$$

A normal approximation is certainly justified.

- (d) Find a 99% confidence interval for p .

SOLUTION:

$$z_{0.99} = 2.58 \text{ (see Table 7-2 or Table 5(b) of Appendix II)}$$

$$E \approx z_{0.99} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \approx 2.58 \sqrt{\frac{(0.75)(0.25)}{800}} \approx 0.0395$$

The 99% confidence interval is then

$$\hat{p} - E < p < \hat{p} + E$$

$$0.75 - 0.0395 < p < 0.75 + 0.0395$$

$$0.71 < p < 0.79$$

Interpretation We are 99% confident that the probability a flu shot will be effective for a student selected at random is between 0.71 and 0.79.

GUIDED EXERCISE 4**Confidence interval for p**

A random sample of 188 books purchased at a local bookstore showed that 66 of the books were murder mysteries. Let p represent the proportion of books sold by this store that are murder mysteries.

- (a) What is a point estimate for p ?

$$\Rightarrow \hat{p} = \frac{r}{n} = \frac{66}{188} = 0.35$$

Continued

GUIDED EXERCISE 4 *continued*

- (b) Find a 90% confidence interval for p .

$$\Rightarrow E = z_c \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$= 1.645 \sqrt{\frac{(0.35)(1 - 0.35)}{188}} \approx 0.0572$$

The confidence interval is

$$\hat{p} - E < p < \hat{p} + E$$

$$0.35 - 0.0572 < p < 0.35 + 0.0572$$

$$0.29 < p < 0.41$$

- (c) **Interpretation** What does the confidence interval you just computed mean in the context of this application?

\Rightarrow If we had computed the interval for many different sets of 188 books, we would have found that about 90% of the intervals actually contained p , the population proportion of mysteries. Consequently, we can be 90% confident that our interval is one of the intervals that contain the unknown value p .

- (d) **Check Requirements** To compute the confidence interval, we used a normal approximation. Does this seem justified?

$$\Rightarrow n = 188; p \approx 0.35; q \approx 0.65$$

Since $np \approx 65.8 > 5$ and $nq \approx 122.2 > 5$, the approximation is justified.

It is interesting to note that our sample point estimate $\hat{p} = r/n$ and the confidence interval for the population proportion p do not depend on the size of the population. In our bookstore example, it made no difference how many books the store sold. On the other hand, the size of the sample does affect the accuracy of a statistical estimate. At the end of this section, we will study the effect of sample size on the reliability of our estimate.



TECH NOTES

The TI-84Plus/TI-83Plus/TI-nspire calculators and Minitab provide confidence intervals for proportions.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Press the STAT key, select TESTS, and choose option A:1-PropZInt. The letter x represents the number of successes r . The TI-84Plus/TI-83Plus/TI-nspire output shows the results for Guided Exercise 4.

```
1-PropZInt
(.29381,.40832)
̂p=.3510638298
n=188
```

Minitab Use the menu selections Stat ▶ Basic Statistics ▶ 1 Proportion. In the dialogue box, select Summarized Data and fill in the number of trials and the number of successes. Under Options, select a confidence interval. Minitab uses the binomial distribution directly unless Normal is checked. The Minitab output shows the results for Guided Exercise 4. Information from Chapter 8 material is also shown.

Test and Confidence Interval for One Proportion (Using Binomial)

Test of $p = 0.5$ vs $p \neq 0.5$

Sample	X	N	Sample p	90.0 % CI	P-Value
1	66	188	0.351064	(0.293222, 0.412466)	0.000

Test and Confidence Interval for One Proportion (Using Normal)						
	Test of $p = 0.5$ vs $p \neq 0.5$					
Sample	X	N	Sample p	90.0 % CI	Z-value	P-value
1	66	188	0.351064	(0.293805, 0.408323)	-4.08	0.000

Interpreting Results from a Poll



Newspapers frequently report the results of an opinion poll. In articles that give more information, a statement about the margin of error accompanies the poll results. In most polls, the margin of error is given for a 95% *confidence interval*.

General interpretation of poll results

- When a poll states the results of a survey, the proportion reported to respond in the designated manner is \hat{p} , the sample estimate of the population proportion.
- The *margin of error* is the maximal error E of a 95% confidence interval for p .
- A 95% confidence interval for the population proportion p is
poll report $\hat{p} - \text{margin of error } E < p < \text{poll report } \hat{p} + \text{margin of error } E$

Margin of error for polls

COMMENT Leslie Kish, a statistician at the University of Michigan, was the first to apply the term *margin of error*. He was a pioneer in the study of population sampling techniques. His book *Survey Sampling* is still widely used all around the world.

Some articles clarify the meaning of the margin of error further by saying that it is an error due to sampling. For instance, the following comments accompany results of a political poll reported in an issue of the *Wall Street Journal*.

How Poll Was Conducted

The *Wall Street Journal/NBC News* poll was based on nationwide telephone interviews of 1508 adults conducted last Friday through Tuesday by the polling organizations of Peter Hart and Robert Teeter.

The sample was drawn from 315 randomly selected geographic points in the continental United States. Each region was represented in proportion to its population. Households were selected by a method that gave all telephone numbers . . . an equal chance of being included.

One adult, 18 years or older, was selected from each household by a procedure to provide the correct number of male and female respondents.

Chances are 19 of 20 that if all adults with telephones in the United States had been surveyed, the findings would differ from these poll results by no more than 2.6 percentage points in either direction.

GUIDED EXERCISE 5

Reading a poll

Read the last paragraph of the article excerpt “How Poll Was Conducted.”

- (a) What confidence level corresponds to the phrase $\rightarrow \frac{19}{20} = 0.95$
“chances are 19 of 20 that if . . .”?

A 95% confidence interval is being discussed.

Continued

GUIDED EXERCISE 5 *continued*

(b) The complete article indicates that everyone in the sample was asked the question “Which party, the Democratic Party or the Republican Party, do you think would do a better job handling . . . education?” Possible responses were “Democrats,” “neither,” “both,” or “Republicans.” The poll reported that 32% of the respondents said, “Democrats.” Does 32% represent the sample statistic \hat{p} or the population parameter p for the proportion of adults responding, “Democrats”?

(c) Continue reading the last paragraph of the article. It goes on to state, “. . . if all adults with telephones in the U.S. had been surveyed, the findings would differ from these poll results by no more than 2.6 percentage points in either direction.” Use this information, together with parts (a) and (b), to find a 95% confidence interval for the proportion p of the specified population who responded, “Democrats” to the question.



32% represents a sample statistic \hat{p} because 32% represents the percentage of the adults in the *sample* who responded, “Democrats.”



The value 2.6 percentage points represents the margin of error. Since the margin of error is for a 95% confidence interval, the confidence interval is

$$32\% - 2.6\% < p < 32\% + 2.6\% \\ 29.4\% < p < 34.6\%$$

The poll indicates that at the time of the poll, between 29.4% and 34.6% of the specified population thought Democrats would do a better job handling education.

Sample Size for Estimating p

Suppose you want to specify the maximal margin of error in advance for a confidence interval for p at a given confidence level c . What sample size do you need? The answer depends on whether or not you have a preliminary estimate for the population probability of success p in a binomial distribution.

PROCEDURE

HOW TO FIND THE SAMPLE SIZE n FOR ESTIMATING A PROPORTION p

$$n = p(1 - p)\left(\frac{z_c}{E}\right)^2 \text{ if you have a preliminary estimate for } p \quad (21)$$

$$n = \frac{1}{4}\left(\frac{z_c}{E}\right)^2 \text{ if you do } \textit{not} \text{ have a preliminary estimate for } p \quad (22)$$

where E = specified maximal error of estimate

z_c = critical value from the normal distribution for the desired confidence level c . Commonly used value of z_c can be found in Table 5(b) of Appendix II.

If n is not a whole number, increase n to the next higher whole number. Also, if necessary, increase the sample size n to ensure that both $np > 5$ and $nq > 5$. Note that n is the minimal sample size for a specified confidence level and maximal error of estimate.

COMMENT To obtain Equation (21), simply solve the formula that gives the maximal error of estimate E of p for the sample size n . When you don’t have an estimate for p , a little algebra can be used to show that the maximum value of $p(1 - q)$ is 1/4. See Problem 27.

 EXAMPLE 7SAMPLE SIZE FOR ESTIMATING p

A company is in the business of selling wholesale popcorn to grocery stores. The company buys directly from farmers. A buyer for the company is examining a large amount of corn from a certain farmer. Before the purchase is made, the buyer wants to estimate p , the probability that a kernel will pop.

Suppose a random sample of n kernels is taken and r of these kernels pop. The buyer wants to be 95% sure that the point estimate $\hat{p} = r/n$ for p will be in error either way by less than 0.01.

- (a) If no preliminary study is made to estimate p , how large a sample should the buyer use?

SOLUTION: In this case, we use Equation (22) with $z_{0.95} = 1.96$ (see Table 7-2) and $E = 0.01$.

$$n = \frac{1}{4} \left(\frac{z_c}{E} \right)^2 = \frac{1}{4} \left(\frac{1.96}{0.01} \right)^2 = 0.25(38,416) = 9604$$

The buyer would need a sample of $n = 9604$ kernels.

- (b) A preliminary study showed that p was approximately 0.86. If the buyer uses the results of the preliminary study, how large a sample should he use?

SOLUTION: In this case, we use Equation (21) with $p \approx 0.86$. Again, from Table 7-2, $z_{0.95} = 1.96$, and from the problem, $E = 0.01$.

$$n = p(1 - p) \left(\frac{z_c}{E} \right)^2 = (0.86)(0.14) \left(\frac{1.96}{0.01} \right)^2 = 4625.29$$

The sample size should be at least $n = 4626$. This sample is less than half the sample size necessary without the preliminary study.


VIEWPOINT

"Band-Aid Surgery"

Faster recovery time and less pain. Sounds great! An alternate surgical technique called laparoscopic ("Band-Aid") surgery involves small incisions through which tiny video cameras and long surgical instruments are maneuvered. Instead of a 10-inch incision, surgeons might use four little stabs of about $\frac{1}{2}$ -inch in length. However, not every such surgery is successful. An article in the Health Section of the Wall Street Journal recommends using a surgeon who has done at least 50 such surgeries. Then the prospective patient should ask about the rate of conversion—that is, the proportion p of times the surgeon has been forced by complications to switch in midoperation to conventional surgery. A confidence interval for the proportion p would be useful patient information!

**SECTION 7.3
PROBLEMS**

For all these problems, carry at least four digits after the decimal in your calculations. Answers may vary slightly due to rounding.

1. **Statistical Literacy** For a binomial experiment with r successes out of n trials, what value do we use as a point estimate for the probability of success p on a single trial?
2. **Statistical Literacy** In order to use a normal distribution to compute confidence intervals for p , what conditions on np and nq need to be satisfied?

3. **Critical Thinking** Results of a poll of a random sample of 3003 American adults showed that 20% did not know that caffeine contributes to dehydration. The poll was conducted for the Nutrition Information Center and had a margin of error of $\pm 1.4\%$.
- Does the margin of error take into account any problems with the wording of the survey question, interviewer errors, bias from sequence of questions, and so forth?
 - What does the margin of error reflect?
4. **Critical Thinking** You want to conduct a survey to determine the proportion of people who favor a proposed tax policy. How does increasing the sample size affect the size of the margin of error?
5. **Critical Thinking** Jerry tested 30 laptop computers owned by classmates enrolled in a large computer science class and discovered that 22 were infected with keystroke-tracking spyware. Is it appropriate for Jerry to use his data to estimate the proportion of all laptops infected with such spyware? Explain.
6. **Critical Thinking: Brain Teaser** A requirement for using the normal distribution to approximate the \hat{p} distribution is that both $np > 5$ and $nq > 5$. Since we usually do not know p , we estimate p by \hat{p} and q by $\hat{q} = 1 - \hat{p}$. Then we require that $n\hat{p} > 5$ and $n\hat{q} > 5$. Show that the conditions $n\hat{p} > 5$ and $n\hat{q} > 5$ are equivalent to the condition that out of n binomial trials, both the number of successes r and the number of failures $n - r$ must exceed 5. Hint: In the inequality $n\hat{p} > 5$, replace \hat{p} by r/n and solve for r . In the inequality $n\hat{q} > 5$, replace \hat{q} by $(n - r)/n$ and solve for $n - r$.
7. **Basic Computation: Confidence Interval for p** Consider $n = 100$ binomial trials with $r = 30$ successes.
- Check Requirements** Is it appropriate to use a normal distribution to approximate the \hat{p} distribution?
 - Find a 90% confidence interval for the population proportion of successes p .
 - Interpretation** Explain the meaning of the confidence interval you computed.
8. **Basic Computation: Confidence Interval for p** Consider $n = 200$ binomial trials with $r = 80$ successes.
- Check Requirements** Is it appropriate to use a normal distribution to approximate the \hat{p} distribution?
 - Find a 95% confidence interval for the population proportion of successes p .
 - Interpretation** Explain the meaning of the confidence interval you computed.
9. **Basic Computation: Sample Size** What is the minimal sample size needed for a 95% confidence interval to have a maximal margin of error of 0.1
- if a preliminary estimate for p is 0.25?
 - if there is no preliminary estimate for p ?
10. **Basic Computation: Sample Size** What is the minimal sample size needed for a 99% confidence interval to have a maximal margin of error of 0.06
- if a preliminary estimate for p is 0.8?
 - if there is no preliminary estimate for p ?
11. **Myers–Briggs: Actors** Isabel Myers was a pioneer in the study of personality types. The following information is taken from *A Guide to the Development and Use of the Myers–Briggs Type Indicator* by Myers and McCaulley (Consulting Psychologists Press). In a random sample of 62 professional actors, it was found that 39 were extroverts.
- Let p represent the proportion of all actors who are extroverts. Find a point estimate for p .
 - Find a 95% confidence interval for p . Give a brief interpretation of the meaning of the confidence interval you have found.
 - Check Requirements** Do you think the conditions $np > 5$ and $nq > 5$ are satisfied in this problem? Explain why this would be an important consideration.

12. **Myers–Briggs: Judges** In a random sample of 519 judges, it was found that 285 were introverts (see reference of Problem 11).
- Let p represent the proportion of all judges who are introverts. Find a point estimate for p .
 - Find a 99% confidence interval for p . Give a brief interpretation of the meaning of the confidence interval you have found.
 - Check Requirements** Do you think the conditions $np > 5$ and $nq > 5$ are satisfied in this problem? Explain why this would be an important consideration.
13. **Navajo Lifestyle: Traditional Hogans** A random sample of 5222 permanent dwellings on the entire Navajo Indian Reservation showed that 1619 were traditional Navajo hogans (*Navajo Architecture: Forms, History, Distributions* by Jett and Spencer, University of Arizona Press).
- Let p be the proportion of all permanent dwellings on the entire Navajo Reservation that are traditional hogans. Find a point estimate for p .
 - Find a 99% confidence interval for p . Give a brief interpretation of the confidence interval.
 - Check Requirements** Do you think that $np > 5$ and $nq > 5$ are satisfied for this problem? Explain why this would be an important consideration.
14. **Archaeology: Pottery** Santa Fe black-on-white is a type of pottery commonly found at archaeological excavations in Bandelier National Monument. At one excavation site a sample of 592 potsherds was found, of which 360 were identified as Santa Fe black-on-white (*Bandelier Archaeological Excavation Project: Summer 1990 Excavations at Burnt Mesa Pueblo and Casa del Rito*, edited by Kohler and Root, Washington State University).
- Let p represent the population proportion of Santa Fe black-on-white potsherds at the excavation site. Find a point estimate for p .
 - Find a 95% confidence interval for p . Give a brief statement of the meaning of the confidence interval.
 - Check Requirements** Do you think the conditions $np > 5$ and $nq > 5$ are satisfied in this problem? Why would this be important?
15. **Health Care: Colorado Physicians** A random sample of 5792 physicians in Colorado showed that 3139 provide at least some charity care (i.e., treat poor people at no cost). These data are based on information from *State Health Care Data: Utilization, Spending, and Characteristics* (American Medical Association).
- Let p represent the proportion of all Colorado physicians who provide some charity care. Find a point estimate for p .
 - Find a 99% confidence interval for p . Give a brief explanation of the meaning of your answer in the context of this problem.
 - Check Requirements** Is the normal approximation to the binomial justified in this problem? Explain.
16. **Law Enforcement: Escaped Convicts** Case studies showed that out of 10,351 convicts who escaped from U.S. prisons, only 7867 were recaptured (*The Book of Odds* by Shook and Shook, Signet).
- Let p represent the proportion of all escaped convicts who will eventually be recaptured. Find a point estimate for p .
 - Find a 99% confidence interval for p . Give a brief statement of the meaning of the confidence interval.
 - Check Requirements** Is use of the normal approximation to the binomial justified in this problem? Explain.
17. **Fishing: Barbless Hooks** In a combined study of northern pike, cutthroat trout, rainbow trout, and lake trout, it was found that 26 out of 855 fish died when caught and released using barbless hooks on flies or lures. All hooks were removed from the fish (Source: *A National Symposium on Catch and Release Fishing*, Humboldt State University Press).

- (a) Let p represent the proportion of all pike and trout that die (i.e., p is the mortality rate) when caught and released using barbless hooks. Find a point estimate for p .
- (b) Find a 99% confidence interval for p , and give a brief explanation of the meaning of the interval.
- (c) **Check Requirements** Is the normal approximation to the binomial justified in this problem? Explain.
18. **Physicians: Solo Practice** A random sample of 328 medical doctors showed that 171 have a solo practice (Source: *Practice Patterns of General Internal Medicine*, American Medical Association).
- (a) Let p represent the proportion of all medical doctors who have a solo practice. Find a point estimate for p .
- (b) Find a 95% confidence interval for p . Give a brief explanation of the meaning of the interval.
- (c) **Interpretation** As a news writer, how would you report the survey results regarding the percentage of medical doctors in solo practice? What is the margin of error based on a 95% confidence interval?
19. **Marketing: Customer Loyalty** In a marketing survey, a random sample of 730 women shoppers revealed that 628 remained loyal to their favorite supermarket during the past year (i.e., did not switch stores) (Source: *Trends in the United States: Consumer Attitudes and the Supermarket*, The Research Department, Food Marketing Institute).
- (a) Let p represent the proportion of all women shoppers who remain loyal to their favorite supermarket. Find a point estimate for p .
- (b) Find a 95% confidence interval for p . Give a brief explanation of the meaning of the interval.
- (c) **Interpretation** As a news writer, how would you report the survey results regarding the percentage of women supermarket shoppers who remained loyal to their favorite supermarket during the past year? What is the margin of error based on a 95% confidence interval?
20. **Marketing: Bargain Hunters** In a marketing survey, a random sample of 1001 supermarket shoppers revealed that 273 always stock up on an item when they find that item at a real bargain price. See reference in Problem 19.
- (a) Let p represent the proportion of all supermarket shoppers who always stock up on an item when they find a real bargain. Find a point estimate for p .
- (b) Find a 95% confidence interval for p . Give a brief explanation of the meaning of the interval.
- (c) **Interpretation** As a news writer, how would you report the survey results on the percentage of supermarket shoppers who stock up on real-bargain items? What is the margin of error based on a 95% confidence interval?
21. **Lifestyle: Smoking** In a survey of 1000 large corporations, 250 said that, given a choice between a job candidate who smokes and an equally qualified nonsmoker, the nonsmoker would get the job (*USA Today*).
- (a) Let p represent the proportion of all corporations preferring a nonsmoking candidate. Find a point estimate for p .
- (b) Find a 0.95 confidence interval for p .
- (c) **Interpretation** As a news writer, how would you report the survey results regarding the proportion of corporations that hire the equally qualified nonsmoker? What is the margin of error based on a 95% confidence interval?
22. **Opinion Poll: Crime and Violence** A *New York Times/CBS* poll asked the question, “What do you think is the most important problem facing this country today?” Nineteen percent of the respondents answered, “Crime and violence.” The margin of sampling error was plus or minus 3 percentage points. Following the convention that the margin of error is based on a 95% confidence interval, find

- a 95% confidence interval for the percentage of the population that would respond, “Crime and violence” to the question asked by the pollsters.
23. **Medical: Blood Type** A random sample of medical files is used to estimate the proportion p of all people who have blood type B.
- If you have no preliminary estimate for p , how many medical files should you include in a random sample in order to be 85% sure that the point estimate \hat{p} will be within a distance of 0.05 from p ?
 - Answer part (a) if you use the preliminary estimate that about 8 out of 90 people have blood type B (Reference: *Manual of Laboratory and Diagnostic Tests* by F. Fischbach).
24. **Business: Phone Contact** How hard is it to reach a businessperson by phone? Let p be the proportion of calls to businesspeople for which the caller reaches the person being called on the *first* try.
- If you have no preliminary estimate for p , how many business phone calls should you include in a random sample to be 80% sure that the point estimate \hat{p} will be within a distance of 0.03 from p ?
 - The Book of Odds* by Shook and Shook (Signet) reports that businesspeople can be reached by a single phone call approximately 17% of the time. Using this (national) estimate for p , answer part (a).
25. **Campus Life: Coeds** What percentage of your campus student body is female? Let p be the proportion of women students on your campus.
- If no preliminary study is made to estimate p , how large a sample is needed to be 99% sure that a point estimate \hat{p} will be within a distance of 0.05 from p ?
 - The Statistical Abstract of the United States*, 112th Edition, indicates that approximately 54% of college students are female. Answer part (a) using this estimate for p .
26. **Small Business: Bankruptcy** The National Council of Small Businesses is interested in the proportion of small businesses that declared Chapter 11 bankruptcy last year. Since there are so many small businesses, the National Council intends to estimate the proportion from a random sample. Let p be the proportion of small businesses that declared Chapter 11 bankruptcy last year.
- If no preliminary sample is taken to estimate p , how large a sample is necessary to be 95% sure that a point estimate \hat{p} will be within a distance of 0.10 from p ?
 - In a preliminary random sample of 38 small businesses, it was found that six had declared Chapter 11 bankruptcy. How many *more* small businesses should be included in the sample to be 95% sure that a point estimate \hat{p} will be within a distance of 0.10 from p ?
27. **Brain Teaser: Algebra** Why do we use $1/4$ in place of $p(1 - p)$ in formula (22) for sample size when the probability of success p is unknown?
- Show that $p(1 - p) = 1/4 - (p - 1/2)^2$.
 - Why is $p(1 - p)$ never greater than $1/4$?
28. **Expand Your Knowledge: Plus Four Confidence Interval for a Single Proportion** One of the technical difficulties that arises in the computation of confidence intervals for a single proportion is that the exact formula for the maximal margin of error requires knowledge of the population proportion of success p . Since p is usually not known, we use the sample estimate $\hat{p} = r/n$ in place of p . As discussed in the article “How Much Confidence Should You Have in Binomial Confidence Intervals?” appearing in issue no. 45 of the magazine *STATS* (a publication of the American Statistical Association), use of \hat{p} as an estimate for p means that the actual confidence level for the intervals may in fact be smaller than the specified level c . This problem arises even when n is large, especially if p is not near $1/2$.



A simple adjustment to the formula for the confidence intervals is the *plus four estimate*, first suggested by Edwin Bidwell Wilson in 1927. It is also called the Agresti–Coull confidence interval. This adjustment works best for 95% confidence intervals.

The plus four adjustment has us add two successes and two failures to the sample data. This means that r , the number of successes, is increased by 2, and n , the sample size, is increased by 4. We use the symbol \tilde{p} , read “ p tilde,” for the resulting sample estimate of p . So, $\tilde{p} = (r + 2)/(n + 4)$.

PROCEDURE

HOW TO COMPUTE A PLUS FOUR CONFIDENCE INTERVAL FOR p

Requirements

Consider a binomial experiment with n trials, where p represents the population probability of success and $q = 1 - p$ represents the population probability of failure. Let r be a random variable that represents the number of successes out of the n binomial trials.

The plus four point estimates for p and q are

$$\tilde{p} = \frac{r + 2}{n + 4} \quad \text{and} \quad \tilde{q} = 1 - \tilde{p}$$

The number of trials n should be at least 10.

Approximate confidence interval for p

$$\tilde{p} - E < p < \tilde{p} + E$$

$$\text{where } E \approx z_c \sqrt{\frac{\tilde{p}\tilde{q}}{n+4}} = z_c \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

c = confidence level ($0 < c < 1$)

z_c = critical value for confidence level c based on the standard normal distribution

- (a) Consider a random sample of 50 trials with 20 successes. Compute a 95% confidence interval for p using the plus four method.
- (b) Compute a traditional 95% confidence interval for p using a random sample of 50 trials with 20 successes.
- (c) Compare the lengths of the intervals obtained using the two methods. Is the point estimate closer to 1/2 when using the plus four method? Is the margin of error smaller when using the plus four method?

SECTION 7.4

Estimating $\mu_1 - \mu_2$ and $p_1 - p_2$

FOCUS POINTS

- Distinguish between independent and dependent samples.
- Compute confidence intervals for $\mu_1 - \mu_2$ when σ_1 and σ_2 are known.
- Compute confidence intervals for $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown.
- Compute confidence intervals for $p_1 - p_2$ using the normal approximation.
- Interpret the meaning and implications of an all-positive, all-negative, or mixed confidence interval.

Independent Samples and Dependent Samples

How can we tell if two populations are different? One way is to compare the difference in population means or the difference in population proportions. In this section, we will use samples from two populations to create confidence intervals for the difference between population parameters.

To make a statistical estimate about the difference between two population parameters, we need to have a sample from each population. Samples may be *independent* or *dependent* according to how they are selected.

Independent samples

Two samples are **independent** if sample data drawn from one population are completely unrelated to the selection of sample data from the other population.

Dependent samples

Two samples are **dependent** if each data value in one sample can be paired with a corresponding data value in the other sample.

Dependent samples and data pairs occur very naturally in “before and after” situations in which the *same object* or item is *measured twice*. We will devote an entire section (8.4) to the study of dependent samples and paired data. However, in this section, we will confine our interest to independent samples.

Independent samples occur very naturally when we draw *two random samples*, one from the first population and one from the second population. Because *both* samples are random samples, there is no pairing of measurements between the two populations. All the examples of this section will involve independent random samples.

GUIDED EXERCISE 6

Distinguishing between independent and dependent samples

For each experiment, categorize the sampling as independent or dependent, and explain your choice.

- (a) In many medical experiments, a sample of subjects is randomly divided into two groups. One group is given a specific treatment, and the other group is given a placebo. After a certain period of time, both groups are measured for the same condition. Do the measurements from these two groups constitute independent or dependent samples?



Since the subjects are *randomly assigned* to the two treatment groups (one receives a treatment, the other a placebo), the resulting measurements form independent samples.

- (b) In an accountability study, a group of students in an English composition course is given a pretest. After the course, the same students are given a posttest covering similar material. Are the two groups of scores independent or dependent?



Since the pretest scores and the posttest scores are from the same students, the samples are dependent. Each student has both a pretest score and a posttest score, so there is a natural pairing of data values.

Confidence Intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 known)

The $\bar{x}_1 - \bar{x}_2$ sampling distribution

In this section, we will use probability distributions that arise from a difference of means (or proportions). How do we obtain such distributions? Suppose we have two statistical variables x_1 and x_2 , each with its own distribution. We take *independent* random samples of size n_1 from the x_1 distribution and of size n_2 from the x_2 distribution. Then we compute the respective means \bar{x}_1 and \bar{x}_2 . Now consider the difference $\bar{x}_1 - \bar{x}_2$. This expression represents a difference of means. If we repeat this sampling process over and over, we will create lots of $\bar{x}_1 - \bar{x}_2$ values. Figure 7-7 illustrates the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

The values of $\bar{x}_1 - \bar{x}_2$ that come from repeated (independent) sampling of populations 1 and 2 can be arranged in a relative-frequency table and a relative-frequency histogram (see Section 2.1). This would give us an experimental idea of the theoretical probability distribution of $\bar{x}_1 - \bar{x}_2$.

Fortunately, it is not necessary to carry out this lengthy process for each example. The results have been worked out mathematically. The next theorem presents the main results.

THEOREM 7.1 Let x_1 and x_2 have normal distributions with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 , respectively. If we take independent random samples of size n_1 from the x_1 distribution and of size n_2 from the x_2 distribution, then the variable $\bar{x}_1 - \bar{x}_2$ has

1. a normal distribution
2. mean $\mu_1 - \mu_2$
3. standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

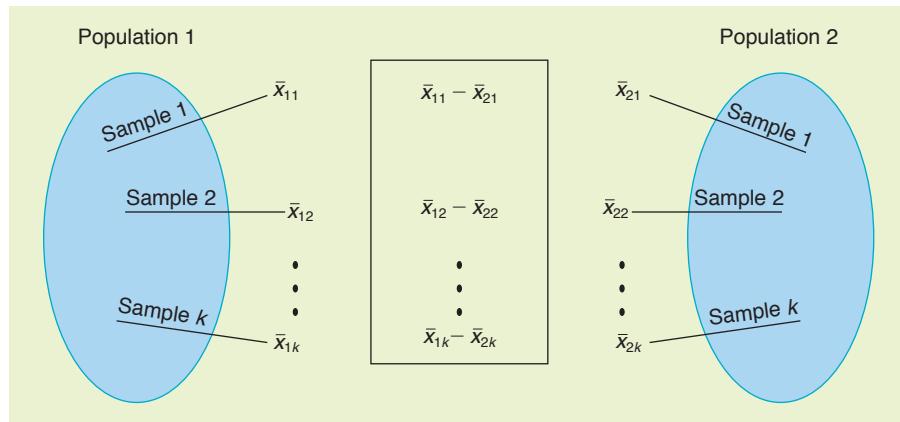
COMMENT The theorem requires that x_1 and x_2 have *normal* distributions. However, if *both* n_1 and n_2 are 30 or larger, then the central limit theorem (Section 6.5) assures us that \bar{x}_1 and \bar{x}_2 are approximately normally distributed. In this case, the conclusions of the theorem are again valid even if the original x_1 and x_2 distributions are not exactly normal.

Confidence intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 known)

If we use Theorem 7.1, then a discussion similar to that of Section 7.1 gives the following information.

FIGURE 7-7

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$



PROCEDURE**HOW TO FIND A CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$ WHEN BOTH σ_1 AND σ_2 ARE KNOWN****Requirements**

Let σ_1 and σ_2 be the population standard deviations of populations 1 and 2. Obtain two *independent* random samples from populations 1 and 2, where

- \bar{x}_1 and \bar{x}_2 are sample means from populations 1 and 2
- n_1 and n_2 are sample sizes from populations 1 and 2

If you can assume that both population distributions 1 and 2 are normal, any sample sizes n_1 and n_2 will work. If you cannot assume this, then use sample sizes $n_1 \geq 30$ and $n_2 \geq 30$.

Confidence interval for $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

$$\text{where } E = z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

c = confidence level ($0 < c < 1$)

z_c = critical value for confidence level c based on the standard normal distribution (See Table 5(b) of Appendix II for commonly used values.)



Problem 28 of this section shows how to determine minimal sample sizes for a specified maximal error of estimate E .

EXAMPLE 8**CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$, σ_1 AND σ_2 KNOWN**

In the summer of 1988, Yellowstone National Park had some major fires that destroyed large tracts of old timber near many famous trout streams. Fishermen were concerned about the long-term effects of the fires on these streams. However, biologists claimed that the new meadows that would spring up under dead trees would produce a lot more insects, which would in turn mean better fishing in the years ahead. Guide services registered with the park provided data about the daily catch for fishermen over many years. Ranger checks on the streams also provided data about the daily number of fish caught by fishermen.

Yellowstone Today (a national park publication) indicates that the biologists' claim is basically correct and that Yellowstone anglers are delighted by their average increased catch.

Suppose you are a biologist studying fishing data from Yellowstone streams before and after the fire. Fishing reports include the number of trout caught per day per fisherman. A random sample of $n_1 = 167$ reports from the period before the fire showed that the average catch was $\bar{x}_1 = 5.2$ trout per day. Assume that the standard deviation of daily catch per fisherman during this period was $\sigma_1 = 1.9$. Another random sample of $n_2 = 125$ fishing reports 5 years after the fire showed that the average catch per day was $\bar{x}_2 = 6.8$ trout. Assume that the standard deviation during this period was $\sigma_2 = 2.3$.

- (a) **Check Requirements** For each sample, what is the population? Are the samples dependent or independent? Explain. Is it appropriate to use a normal distribution for the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.

SOLUTION: The population for the first sample is the number of trout caught per day by fishermen before the fire. The population for the second sample is the number of trout caught per day after the fire. Both samples were random



Yellowstone National Park

samples taken in their respective time periods. There was no effort to pair individual data values. Therefore, the samples can be thought of as independent samples.

A normal distribution is appropriate for the $\bar{x}_1 - \bar{x}_2$ distribution because sample sizes are sufficiently large and we know both σ_1 and σ_2 .

- (b) Compute a 95% confidence interval for $\mu_1 - \mu_2$, the difference of population means.

SOLUTION: Since $n_1 = 167$, $\bar{x}_1 = 5.2$, $\sigma_1 = 1.9$, $n_2 = 125$, $\bar{x}_2 = 6.8$, $\sigma_2 = 2.3$, and $z_{0.95} = 1.96$ (see Table 7-2), then

$$\begin{aligned} E &= z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= 1.96 \sqrt{\frac{(1.9)^2}{167} + \frac{(2.3)^2}{125}} \approx 1.96 \sqrt{0.0639} \approx 0.4955 \approx 0.50 \end{aligned}$$

The 95% confidence interval is

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E \\ (5.2 - 6.8) - 0.50 < \mu_1 - \mu_2 < (5.2 - 6.8) + 0.50 \\ -2.10 < \mu_1 - \mu_2 < -1.10 \end{aligned}$$

- (c) **Interpretation** What is the meaning of the confidence interval computed in part (b)?

SOLUTION: We are 95% confident that the interval -2.10 to -1.10 fish per day is one of the intervals containing the population difference $\mu_1 - \mu_2$, where μ_1 represents the population average daily catch before the fire and μ_2 represents the population average daily catch after the fire. Put another way, since the confidence interval contains only *negative values*, we can be 95% sure that $\mu_1 - \mu_2 < 0$. This means we are 95% sure that $\mu_1 < \mu_2$. In words, we are 95% sure that the average catch before the fire was less than the average catch after the fire.



COMMENT In the case of large samples ($n_1 \geq 30$ and $n_2 \geq 30$), it is not unusual to see σ_1 and σ_2 approximated by s_1 and s_2 . Then Theorem 7.1 is used as a basis for approximating confidence intervals for $\mu_1 - \mu_2$. In other words, when samples are large, sample estimates for σ_1 and σ_2 can be used together with the standard normal distribution to find confidence intervals for $\mu_1 - \mu_2$. However, in this text, we follow the more common convention of using a Student's t distribution whenever σ_1 and σ_2 are unknown.

Confidence Intervals for $\mu_1 - \mu_2$ When σ_1 and σ_2 Are Unknown

When σ_1 and σ_2 are unknown, we turn to a Student's t distribution. As before, when we use a Student's t distribution, we require that our populations be normal or approximately normal (mound-shaped and symmetric) when the sample sizes n_1 and n_2 are less than 30. We also replace σ_1 by s_1 and σ_2 by s_2 . Then we consider the approximate t value attributed to Welch (*Biometrika*, Vol. 29, pp. 350–362).

$$t \approx \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Unfortunately, this approximation is *not exactly* a Student's t distribution. However, it will be a good approximation provided we adjust the degrees of freedom by one of the following methods.



Problem 30 of this section shows the formula for Satterthwaite's approximation for degrees of freedom.

Confidence intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 unknown)

1. The adjustment for the degrees of freedom is calculated from sample data. The formula, called *Satterthwaite's approximation*, is rather complicated. Satterthwaite's approximation is used in statistical software packages such as Minitab and in the TI-84Plus/TI-83Plus/TI-nspire calculators. See Problem 30 for the formula.
2. An alternative method, which is much simpler, is to approximate the degrees of freedom using the *smaller* of $n_1 - 1$ and $n_2 - 1$.

For confidence intervals, we take the degrees of freedom *d.f.* to be the smaller of $n_1 - 1$ and $n_2 - 1$. This commonly used choice for the degrees of freedom is more conservative than Satterthwaite's approximation in the sense that the former produces a slightly larger margin of error. The resulting confidence interval will be *at least* at the c level, or a little higher.

Applying methods similar to those used to find confidence intervals for μ when σ is unknown, and using the Welch approximation for t , we obtain the following results.

PROCEDURE

HOW TO FIND A CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$ WHEN σ_1 AND σ_2 ARE UNKNOWN

Requirements

Obtain two independent random samples from populations 1 and 2, where

\bar{x}_1 and \bar{x}_2 are sample means from populations 1 and 2

s_1 and s_2 are sample standard deviations from populations 1 and 2

n_1 and n_2 are sample sizes from populations 1 and 2

If you can assume that both population distributions 1 and 2 are normal or at least mound-shaped and symmetric, then any sample sizes n_1 and n_2 will work. If you cannot assume this, then use sample sizes $n_1 \geq 30$ and $n_2 \geq 30$.

Confidence interval for $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

$$\text{where } E \approx t_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

c = confidence level ($0 < c < 1$)

t_c = critical value for confidence level c

$d.f.$ = *smaller* of $n_1 - 1$ and $n_2 - 1$. Note that statistical software gives a more accurate and larger d.f. based on Satterthwaite's approximation.

EXAMPLE 9

CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$, σ_1 AND σ_2 UNKNOWN

Alexander Borbely is a professor at the Medical School of the University of Zurich, where he is director of the Sleep Laboratory. Dr. Borbely and his colleagues are experts on sleep, dreams, and sleep disorders. In his book *Secrets of Sleep*, Dr. Borbely discusses brain waves, which are measured in hertz, the number of oscillations per second. Rapid brain waves (wakefulness) are in the range of 16 to 25 hertz. Slow brain waves (sleep) are in the range of 4 to 8 hertz. During normal sleep, a person goes through several cycles (each cycle is about 90 minutes) of brain waves, from rapid to slow and back to rapid. During deep sleep, brain waves are at their slowest.

In his book, Professor Borbely comments that alcohol is a *poor* sleep aid. In one study, a number of subjects were given 1/2 liter of red wine before they went

to sleep. The subjects fell asleep quickly but did not remain asleep the entire night. Toward morning, between 4 and 6 A.M., they tended to wake up and have trouble going back to sleep.

Suppose that a random sample of 29 college students was randomly divided into two groups. The first group of $n_1 = 15$ people was given 1/2 liter of red wine before going to sleep. The second group of $n_2 = 14$ people was given no alcohol before going to sleep. Everyone in both groups went to sleep at 11 P.M. The average brain wave activity (4 to 6 A.M.) was determined for each individual in the groups. Assume the average brain wave distribution in each group is mound-shaped and symmetric. The results follow:

Group 1 (x_1 values): $n_1 = 15$ (with alcohol)

Average brain wave activity in the hours 4 to 6 A.M.

16.0	19.6	19.9	20.9	20.3	20.1	16.4	20.6
20.1	22.3	18.8	19.1	17.4	21.1	22.1	

For group 1, we have the sample mean and standard deviation of

$$\bar{x}_1 \approx 19.65 \quad \text{and} \quad s_1 \approx 1.86$$

Group 2 (x_2 values): $n_2 = 14$ (no alcohol)

Average brain wave activity in the hours 4 to 6 A.M.

8.2	5.4	6.8	6.5	4.7	5.9	2.9
7.6	10.2	6.4	8.8	5.4	8.3	5.1

For group 2, we have the sample mean and standard deviation of

$$\bar{x}_2 \approx 6.59 \quad \text{and} \quad s_2 \approx 1.91$$

- (a) **Check Requirements** Are the samples independent or dependent? Explain. Is it appropriate to use a Student's t distribution to approximate the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.

SOLUTION: Since the original random sample of 29 students was randomly divided into two groups, it is reasonable to say that the samples are independent. A Student's t distribution is appropriate for the $\bar{x}_1 - \bar{x}_2$ distribution because both original distributions are mound-shaped and symmetric. We don't know population standard deviations, but we can compute s_1 and s_2 .

- (b) Compute a 90% confidence interval for $\mu_1 - \mu_2$, the difference of population means.

SOLUTION: First we find $t_{0.90}$. We approximate the degrees of freedom *d.f.* by using the smaller of $n_1 - 1$ and $n_2 - 1$. Since n_2 is smaller, *d.f.* = $n_2 - 1 = 14 - 1 = 13$. This gives us $t_{0.90} \approx 1.771$. The margin of error is then

$$E \approx t_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.771 \sqrt{\frac{1.86^2}{15} + \frac{1.91^2}{14}} \approx 1.24$$

The c confidence interval is

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E \\ (19.65 - 6.59) - 1.24 < \mu_1 - \mu_2 < (19.65 - 6.59) + 1.24 \\ 11.82 < \mu_1 - \mu_2 < 14.30 \end{aligned}$$

After further rounding we have

$$11.8 \text{ hertz} < \mu_1 - \mu_2 < 14.3 \text{ hertz}$$

- (c) **Interpretation** What is the meaning of the confidence interval you computed in part (b)?

SOLUTION: μ_1 represents the population average brain wave activity for people who drank 1/2 liter of wine before sleeping. μ_2 represents the population

average brain wave activity for people who took no alcohol before sleeping. Both periods of measurement are from 4 to 6 A.M. We are 90% confident that the interval between 11.8 and 14.3 hertz is one that contains the difference $\mu_1 - \mu_2$. It would seem reasonable to conclude that people who drink before sleeping might wake up in the early morning and have trouble going back to sleep. Since the confidence interval from 11.8 to 14.3 contains only *positive values*, we could express this by saying that we are 90% confident that $\mu_1 - \mu_2$ is *positive*. This means that $\mu_1 - \mu_2 > 0$. Thus, we are 90% confident that $\mu_1 > \mu_2$ (that is, average brain wave activity from 4 to 6 A.M. for the group drinking wine was more than average brain wave activity for the group not drinking).



There is another method of constructing confidence intervals for $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown. Suppose the sample values s_1 and s_2 are sufficiently close and there is reason to believe that $\sigma_1 = \sigma_2$. Methods shown in Section 10.4 use sample standard deviations s_1 and s_2 to determine if $\sigma_1 = \sigma_2$. When you can assume that $\sigma_1 = \sigma_2$, it is best to use a *pooled standard deviation* to compute the margin of error. The $\bar{x}_1 - \bar{x}_2$ distribution has an *exact Student's t* distribution with $d.f. = n_1 + n_2 - 2$. Problem 31 of this section gives the details.



The formulas and requirements for using a pooled standard deviation to compute a confidence interval for $\mu_1 - \mu_2$ can be found in Problem 31 at the end of this section.



Sometimes we want several confidence intervals to hold concurrently. Problem 19 in the Chapter Review Problems discusses this situation.

Summary

What should a person do? You have independent random samples from two populations. You can compute \bar{x}_1 , \bar{x}_2 , s_1 , and s_2 and you have the sample sizes n_1 and n_2 . In any case, a confidence interval for the difference $\mu_1 - \mu_2$ of population means is

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

where E is the margin of error. How do you compute E ? The answer depends on how much you know about the x_1 and x_2 distributions.

Situation I (the usual case)

You simply don't know the population values of σ_1 and σ_2 . In this situation, you use a *t* distribution with margin of error

$$E = t_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where a conservative estimate for the degrees of freedom is

$$d.f. = \text{minimum of } n_1 - 1 \text{ and } n_2 - 1$$

Like a good friend, the *t* distribution has a reputation for being robust and forgiving. Nevertheless, some guidelines should be observed. If n_1 and n_2 are both less than 30, then x_1 and x_2 should have distributions that are mound-shaped and approximately symmetric (or, even better, normal). If both n_1 and n_2 are 30 or more, the central limit theorem (Chapter 6) implies that these restrictions can be relaxed.

Situation II (almost never happens)

You actually know the population values of σ_1 and σ_2 . In addition, you know that x_1 and x_2 have normal distributions. If you know σ_1 and σ_2 but are not sure about the x_1 and x_2 distributions, then you must have $n_1 \geq 30$ and $n_2 \geq 30$. In this situation, you use a *z* distribution with margin of error

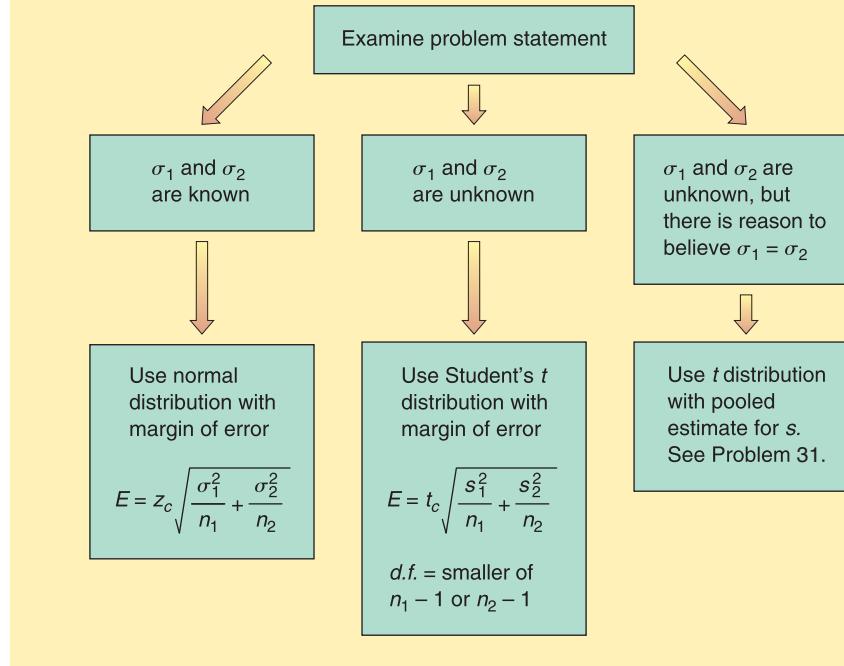
$$E = z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Continued

Situation III (yes, this does sometimes occur)

You don't know σ_1 and σ_2 , but the sample values s_1 and s_2 are close to each other and there is reason to believe that $\sigma_1 = \sigma_2$. This can happen when you make a slight change or alteration to a known process or method of production. The standard deviation may not change much, but the outputs or means could be very different. In this situation, you are advised to use a t distribution with a pooled standard deviation. See Problem 31 at the end of this section.

Which distribution should you use for $\bar{x}_1 - \bar{x}_2$

**Estimating the Difference of Proportions $p_1 - p_2$**

We conclude this section with a discussion of confidence intervals for $p_1 - p_2$, the difference of two proportions from binomial probability distributions. The main result on this topic is the following theorem.

THEOREM 7.2 Consider two binomial probability distributions

Distribution 1

n_1 = number of trials

r_1 = number of successes out of n_1 trials

p_1 = probability of success on each trial

$q_1 = 1 - p_1$ = probability of failure on each trial

$\hat{p}_1 = \frac{r_1}{n_1}$ = point estimate for p_1

$\hat{q}_1 = 1 - \frac{r_1}{n_1}$ = point estimate for q_1

Distribution 2

n_2 = number of trials

r_2 = number of successes out of n_2 trials

p_2 = probability of success on each trial

$q_2 = 1 - p_2$ = probability of failure on each trial

$\hat{p}_2 = \frac{r_2}{n_2}$ = point estimate for p_2

$\hat{q}_2 = 1 - \frac{r_2}{n_2}$ = point estimate for q_2

For most practical applications, if the four quantities

$$n_1\hat{p}_1 \quad n_1\hat{q}_1 \quad n_2\hat{p}_2 \quad n_2\hat{q}_2$$

are all larger than 5 (see Section 6.6), then the following statements are true about the random variable $\frac{r_1}{n_1} - \frac{r_2}{n_2}$:

1. $\frac{r_1}{n_1} - \frac{r_2}{n_2}$ has an approximately normal distribution.
2. The mean is $p_1 - p_2$.
3. The standard deviation is approximately

$$\hat{\sigma} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Confidence intervals for $p_1 - p_2$

Based on the preceding theorem, we can find confidence intervals for $p_1 - p_2$ in the following way:

PROCEDURE

HOW TO FIND A CONFIDENCE INTERVAL FOR $p_1 - p_2$

Requirements

Consider two independent binomial experiments.

Binomial Experiment 1

n_1 = number of trials

r_1 = number of successes out of n_1 trials

$$\hat{p}_1 = \frac{r_1}{n_1}; \hat{q}_1 = 1 - \hat{p}_1$$

p_1 = population probability of success

Binomial Experiment 2

n_2 = number of trials

r_2 = number of successes out of n_2 trials

$$\hat{p}_2 = \frac{r_2}{n_2}; \hat{q}_2 = 1 - \hat{p}_2$$

p_2 = population probability of success

The number of trials should be sufficiently large so that all four of the following inequalities are true:

$$n_1\hat{p}_1 > 5; \quad n_1\hat{q}_1 > 5; \quad n_2\hat{p}_2 > 5; \quad n_2\hat{q}_2 > 5$$

Confidence interval for $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) - E \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + E$$

where

$$E = z_c \hat{\sigma} = z_c \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

c = confidence level, $0 < c < 1$

z_c = critical value for confidence level c based on the standard normal distribution (See Table 5(b) of Appendix II for commonly used values.)



Problem 29 shows how to compute minimal sample sizes for a specified margin of error E .

EXAMPLE 10

CONFIDENCE INTERVAL FOR $p_1 - p_2$

In his book *Secrets of Sleep*, Professor Borbely describes research on dreams in the Sleep Laboratory at the University of Zurich Medical School. During normal sleep, there is a phase known as REM (rapid eye movement). For most people, REM sleep occurs about every 90 minutes or so, and it is thought that dreams occur just before or during the REM phase. Using electronic equipment in the Sleep Laboratory, it is possible to detect the REM phase in a sleeping person. If a person is wakened immediately after the REM phase, he or she usually can

describe a dream that has just taken place. Based on a study of over 650 people in the Zurich Sleep Laboratory, it was found that about one-third of all dream reports contain feelings of fear, anxiety, or aggression. There is a conjecture that if a person is in a good mood when going to sleep, the proportion of “bad” dreams (fear, anxiety, aggression) might be reduced.

Suppose that two groups of subjects were randomly chosen for a sleep study. In group I, before going to sleep, the subjects spent 1 hour watching a comedy movie. In this group, there were a total of $n_1 = 175$ dreams recorded, of which $r_1 = 49$ were dreams with feelings of anxiety, fear, or aggression. In group II, the subjects did not watch a movie but simply went to sleep. In this group, there were a total of $n_2 = 180$ dreams recorded, of which $r_2 = 63$ were dreams with feelings of anxiety, fear, or aggression.

- (a) **Check Requirements** Why could groups I and II be considered independent binomial distributions? Why do we have a “large-sample” situation?

SOLUTION: Since the two groups were chosen randomly, it is reasonable to assume that neither group’s responses would be related to the other’s. In both groups, each recorded dream could be thought of as a trial, with success being a dream with feelings of fear, anxiety, or aggression.

$$\hat{p}_1 = \frac{r_1}{n_1} = \frac{49}{175} = 0.28 \quad \text{and} \quad \hat{q}_1 = 1 - \hat{p}_1 = 0.72$$

$$\hat{p}_2 = \frac{r_2}{n_2} = \frac{63}{180} = 0.35 \quad \text{and} \quad \hat{q}_2 = 1 - \hat{p}_2 = 0.65$$

Since

$$n_1\hat{p}_1 = 49 > 5 \quad n_1\hat{q}_1 = 126 > 5$$

$$n_2\hat{p}_2 = 63 > 5 \quad n_2\hat{q}_2 = 117 > 5$$

large-sample theory is appropriate.

- (b) What is $p_1 - p_2$? Compute a 95% confidence interval for $p_1 - p_2$.

SOLUTION: p_1 is the population proportion of successes (bad dreams) for all people who watched comedy movies before bed. Thus, p_1 can be thought of as the percentage of bad dreams for all people who were in a “good mood” when they went to bed. Likewise, p_2 is the percentage of bad dreams for the population of all people who just went to bed (no movie). The difference $p_1 - p_2$ is the population difference.

To find a confidence interval for $p_1 - p_2$, we need the values of z_c , $\hat{\sigma}$, and then E . From Table 7-2, we see that $z_{0.95} = 1.96$, so

$$\hat{\sigma} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = \sqrt{\frac{(0.28)(0.72)}{175} + \frac{(0.35)(0.65)}{180}}$$

$$\approx \sqrt{0.0024} \approx 0.0492$$

$$E = z_c \hat{\sigma} = 1.96(0.0492) \approx 0.096$$

$$(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E$$

$$(0.28 - 0.35) - 0.096 < p_1 - p_2 < (0.28 - 0.35) + 0.096$$

$$- 0.166 < p_1 - p_2 < 0.026$$

- (c) **Interpretation** What is the meaning of the confidence interval constructed in part (b)?

SOLUTION: We are 95% sure that the interval between -16.6% and 2.6% is one that contains the percentage difference of “bad” dreams for group I and group II. Since the interval -0.166 to 0.026 is not all negative (or all positive), we cannot say that $p_1 - p_2 < 0$ (or $p_1 - p_2 > 0$). Thus, at the 95% confidence level, we *cannot* conclude that $p_1 < p_2$ or $p_1 > p_2$. The comedy

movies before bed helped some people reduce the percentage of “bad” dreams, but at the 95% confidence level, we cannot say that the *population difference* is reduced.



Interpreting Confidence Intervals for Differences

As we have seen in the preceding examples, at the c confidence level we can determine how two means or proportions from independent random samples are related. The next procedure summarizes the results.

PROCEDURE

HOW TO INTERPRET CONFIDENCE INTERVALS FOR DIFFERENCES

Suppose we construct a $c\%$ confidence interval for $\mu_1 - \mu_2$ (or $p_1 - p_2$). Then three cases arise:

1. The $c\%$ confidence interval contains only *negative values* (see Example 8). In this case, we conclude that $\mu_1 - \mu_2 < 0$ (or $p_1 - p_2 < 0$), and we are therefore $c\%$ confident that $\mu_1 < \mu_2$ (or $p_1 < p_2$).
2. The $c\%$ confidence interval contains only *positive values* (see Example 9). In this case, we conclude that $\mu_1 - \mu_2 > 0$ (or $p_1 - p_2 > 0$), and we can be $c\%$ confident that $\mu_1 > \mu_2$ (or $p_1 > p_2$).
3. The $c\%$ confidence interval contains *both positive and negative values* (see Example 10). In this case, we cannot at the $c\%$ confidence level conclude that either μ_1 or μ_2 (or p_1 or p_2) is larger. However, if we *reduce* the confidence level c to a *smaller value*, then the confidence interval will, in general, be shorter (explain why). Another approach (when possible) is to increase the sample sizes n_1 and n_2 . This would also tend to make the confidence interval shorter (explain why). A shorter confidence interval *might* put us back into case 1 or case 2 above (again, explain why).

GUIDED EXERCISE 7

Interpreting a confidence interval

- (a) A study reported a 90% confidence interval for the difference of means to be

$$10 < \mu_1 - \mu_2 < 20$$

For this interval, what can you conclude about the respective values of μ_1 and μ_2 ?



At a 90% level of confidence, we can say that the difference $\mu_1 - \mu_2$ is positive, so $\mu_1 - \mu_2 > 0$ and $\mu_1 > \mu_2$.

- (b) A study reported a 95% confidence interval for the difference of proportions to be

$$-0.32 < p_1 - p_2 < 0.16$$

For this interval, what can you conclude about the respective values of p_1 and p_2 ?



At the 95% confidence level, we see that the difference of proportions ranges from negative to positive values. We cannot tell from this interval if p_1 is greater than p_2 or p_1 is less than p_2 .

TECH NOTES

The TI-84Plus/TI-83Plus/TI-nspire calculators and Minitab supply confidence intervals for the difference of means and for the difference of proportions.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Use the STAT key and highlight TESTS. The choice 9:2-SampZInt finds confidence intervals for differences of means when σ_1 and σ_2 are known. Choice 0:2-SampTInt finds confidence intervals for differences of means when σ_1 and σ_2 are unknown. In general, use No for Pooled. However, if $\sigma_1 \approx \sigma_2$, use Yes for Pooled. Choice B:2-PropZInt provides confidence intervals for proportions.

Minitab Use the menu choice STAT > Basic Statistics > 2 sample t or 2 proportions. Minitab always uses the Student's *t* distribution for $\mu_1 - \mu_2$ confidence intervals. If the variances are equal, check "assume equal variances."

VIEWPOINT**What's the Difference?**

Will two 15-minute piano lessons a week significantly improve a child's analytical reasoning skills? Why piano? Why not computer keyboard instruction or maybe voice lessons? Professor Frances Rauscher, University of Wisconsin, and Professor Gordon Shaw, University of California at Irvine, claim there is a difference! How could this be measured? A large number of piano students were given complicated tests of mental ability. Independent control groups of other students were given the same tests. Techniques involving the study of differences of means were used to draw the conclusion that students taking piano lessons did better on tests measuring analytical reasoning skills. (Reported in The Denver Post.)

**SECTION 7.4
PROBLEMS**

Answers may vary slightly due to rounding.

1. **Statistical Literacy** When are two random samples independent?
2. **Statistical Literacy** When are two random samples dependent?
3. **Critical Thinking** Josh and Kendra each calculated a 90% confidence interval for the difference of means using a Student's *t* distribution for random samples of size $n_1 = 20$ and $n_2 = 31$. Kendra followed the convention of using the smaller sample size to compute $d.f. = 19$. Josh used his calculator and Satterthwaite's approximation and obtained $d.f. \approx 36.3$. Which confidence interval is shorter? Which confidence interval is more conservative in the sense that the margin of error is larger?
4. **Critical Thinking** If a 90% confidence interval for the difference of means $\mu_1 - \mu_2$ contains all positive values, what can we conclude about the relationship between μ_1 and μ_2 at the 90% confidence level?
5. **Critical Thinking** If a 90% confidence interval for the difference of means $\mu_1 - \mu_2$ contains all negative values, what can we conclude about the relationship between μ_1 and μ_2 at the 90% confidence level?
6. **Critical Thinking** If a 90% confidence interval for the difference of proportions contains some positive and some negative values, what can we conclude about the relationship between p_1 and p_2 at the 90% confidence level?
7. **Basic Computation: Confidence Interval for $\mu_1 - \mu_2$** Consider two independent normal distributions. A random sample of size $n_1 = 20$ from the first distribution showed $\bar{x}_1 = 12$ and a random sample of size $n_2 = 25$ from the second distribution showed $\bar{x}_2 = 14$.

- (a) **Check Requirements** If σ_1 and σ_2 are known, what distribution does $\bar{x}_1 - \bar{x}_2$ follow? Explain.
- (b) Given $\sigma_1 = 3$ and $\sigma_2 = 4$, find a 90% confidence interval for $\mu_1 - \mu_2$.
- (c) **Check Requirements** Suppose σ_1 and σ_2 are both unknown, but from the random samples, you know $s_1 = 3$ and $s_2 = 4$. What distribution approximates the $\bar{x}_1 - \bar{x}_2$ distribution? What are the degrees of freedom? Explain.
- (d) With $s_1 = 3$ and $s_2 = 4$, find a 90% confidence interval for $\mu_1 - \mu_2$.
- (e) If you have an appropriate calculator or computer software, find a 90% confidence interval for $\mu_1 - \mu_2$ using degrees of freedom based on Satterthwaite's approximation.
- (f) **Interpretation** Based on the confidence intervals you computed, can you be 90% confident that μ_1 is smaller than μ_2 ? Explain.
8. **Basic Computation: Confidence Interval for $\mu_1 - \mu_2$** Consider two independent distributions that are mound-shaped. A random sample of size $n_1 = 36$ from the first distribution showed $\bar{x}_1 = 15$, and a random sample of size $n_2 = 40$ from the second distribution showed $\bar{x}_2 = 14$.
- (a) **Check Requirements** If σ_1 and σ_2 are known, what distribution does $\bar{x}_1 - \bar{x}_2$ follow? Explain.
- (b) Given $\sigma_1 = 3$ and $\sigma_2 = 4$, find a 95% confidence interval for $\mu_1 - \mu_2$.
- (c) **Check Requirements** Suppose σ_1 and σ_2 are both unknown, but from the random samples, you know $s_1 = 3$ and $s_2 = 4$. What distribution approximates the $\bar{x}_1 - \bar{x}_2$ distribution? What are the degrees of freedom? Explain.
- (d) With $s_1 = 3$ and $s_2 = 4$, find a 95% confidence interval for $\mu_1 - \mu_2$.
- (e) If you have an appropriate calculator or computer software, find a 95% confidence interval for $\mu_1 - \mu_2$ using degrees of freedom based on Satterthwaite's approximation.
- (f) **Interpretation** Based on the confidence intervals you computed, can you be 95% confident that μ_1 is larger than μ_2 ? Explain.
9. **Basic Computation: Confidence Interval for $p_1 - p_2$** Consider two independent binomial experiments. In the first one, 40 trials had 10 successes. In the second one, 50 trials had 15 successes.
- (a) **Check Requirements** Is it appropriate to use a normal distribution to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
- (b) Find a 90% confidence interval for $p_1 - p_2$.
- (c) **Interpretation** Based on the confidence interval you computed, can you be 90% confident that p_1 is less than p_2 ? Explain.
10. **Basic Computation: Confidence Interval for $p_1 - p_2$** Consider two independent binomial experiments. In the first one, 40 trials had 15 successes. In the second one, 60 trials had 6 successes.
- (a) **Check Requirements** Is it appropriate to use a normal distribution to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
- (b) Find a 95% confidence interval for $p_1 - p_2$.
- (c) **Interpretation** Based on the confidence interval you computed, can you be 95% confident that p_1 is more than p_2 ? Explain.
11. **Archaeology: Ireland** Inorganic phosphorous is a naturally occurring element in all plants and animals, with concentrations increasing progressively up the food chain (fruit < vegetables < cereals < nuts < corpse). Geochemical surveys take soil samples to determine phosphorous content (in ppm, parts per million). A high phosphorous content may or may not indicate an ancient burial site, food storage site, or even a garbage dump. The Hill of Tara is a very important archaeological site in Ireland. It is by legend the seat of Ireland's ancient high kings (Reference: *Tara, An Archaeological Survey* by Conor Newman, Royal Irish Academy, Dublin). Independent random samples from two regions in Tara gave the following phosphorous measurements (in ppm). Assume the population distributions of phosphorous are mound-shaped and symmetric for these two regions.

Region I: $x_1; n_1 = 12$

540	810	790	790	340	800
890	860	820	640	970	720

Region II: $x_2; n_2 = 16$

750	870	700	810	965	350	895	850
635	955	710	890	520	650	280	993

- (a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 747.5$, $s_1 \approx 170.4$, $\bar{x}_2 \approx 738.9$, and $s_2 \approx 212.1$.
- (b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2 . Find a 90% confidence interval for $\mu_1 - \mu_2$.
- (c) **Interpretation** Explain what the confidence interval means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 90% level of confidence, is one region more interesting than the other from a geochemical perspective?
- (d) **Check Requirements** Which distribution (standard normal or Student's t) did you use? Why?

12. **Archaeology: Ireland** Please see the setting and reference in Problem 11. Independent random samples from two regions (not those cited in Problem 11) gave the following phosphorous measurements (in ppm). Assume the distribution of phosphorous is mound-shaped and symmetric for these two regions.

Region I: $x_1; n_1 = 15$

855	1550	1230	875	1080	2330	1850	1860
2340	1080	910	1130	1450	1260	1010	

Region II: $x_2; n_2 = 14$

540	810	790	1230	1770	960	1650	860
890	640	1180	1160	1050	1020		

- (a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 1387.3$, $s_1 \approx 498.3$, $\bar{x}_2 \approx 1039.3$, and $s_2 \approx 346.7$.
- (b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2 . Find an 80% confidence interval for $\mu_1 - \mu_2$.
- (c) **Interpretation** Explain what the confidence interval means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 80% level of confidence, is one region more interesting than the other from a geochemical perspective?
- (d) **Check Requirements** Which distribution (standard normal or Student's t) did you use? Why?

13. **Large U.S. Companies: Foreign Revenue** For large U.S. companies, what percentage of their total income comes from foreign sales? A random sample of technology companies (IBM, Hewlett-Packard, Intel, and others) gave the following information.

Technology companies, % foreign revenue: $x_1; n_1 = 16$

62.8	55.7	47.0	59.6	55.3	41.0	65.1	51.1
53.4	50.8	48.5	44.6	49.4	61.2	39.3	41.8

Another independent random sample of basic consumer product companies (Goodyear, Sarah Lee, H.J. Heinz, Toys "R" Us) gave the following information.

Basic consumer product companies, % foreign revenue: $x_2; n_2 = 17$

28.0	30.5	34.2	50.3	11.1	28.8	40.0	44.9
40.7	60.1	23.1	21.3	42.8	18.0	36.9	28.0
		32.5					

(Reference: *Forbes Top Companies*.) Assume that the distributions of percentage foreign revenue are mound-shaped and symmetric for these two company types.

- Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 51.66$, $s_1 \approx 7.93$, $\bar{x}_2 \approx 33.60$, and $s_2 \approx 12.26$.
- Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2 . Find an 85% confidence interval for $\mu_1 - \mu_2$.
- Interpretation** Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 85% level of confidence, do technology companies have a greater percentage foreign revenue than basic consumer product companies?
- Check Requirements** Which distribution (standard normal or Student's *t*) did you use? Why?



14.

Pro Football and Basketball: Weights of Players Independent random samples of professional football and basketball players gave the following information (References: *Sports Encyclopedia of Pro Football* and *Official NBA Basketball Encyclopedia*). Note: These data are also available for download at the Online Study Center. Assume that the weight distributions are mound-shaped and symmetric.

Weights (in lb) of pro football players: x_1 ; $n_1 = 21$

245	262	255	251	244	276	240	265	257	252	282
256	250	264	270	275	245	275	253	265	270	

Weights (in lb) of pro basketball players: x_2 ; $n_2 = 19$

205	200	220	210	191	215	221	216	228	207
225	208	195	191	207	196	181	193	201	

Richard Paul Kane, 2010/Used under license from Shutterstock.com



15.

Pro Football and Basketball: Heights of Players Independent random samples of professional football and basketball players gave the following information (References: *Sports Encyclopedia of Pro Football* and *Official NBA Basketball Encyclopedia*). Note: These data are also available for download at the Online Study Center.

Heights (in ft) of pro football players: x_1 ; $n_1 = 45$

6.33	6.50	6.50	6.25	6.50	6.33	6.25	6.17	6.42	6.33
6.42	6.58	6.08	6.58	6.50	6.42	6.25	6.67	5.91	6.00
5.83	6.00	5.83	5.08	6.75	5.83	6.17	5.75	6.00	5.75
6.50	5.83	5.91	5.67	6.00	6.08	6.17	6.58	6.50	6.25
6.33	5.25	6.67	6.50	5.83					

Heights (in ft) of pro basketball players: x_2 ; $n_2 = 40$

6.08	6.58	6.25	6.58	6.25	5.92	7.00	6.41	6.75	6.25
6.00	6.92	6.83	6.58	6.41	6.67	6.67	5.75	6.25	6.25
6.50	6.00	6.92	6.25	6.42	6.58	6.58	6.08	6.75	6.50
6.83	6.08	6.92	6.00	6.33	6.50	6.58	6.83	6.50	6.58

- (a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 6.179$, $s_1 \approx 0.366$, $\bar{x}_2 \approx 6.453$, and $s_2 \approx 0.314$.
- (b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2 . Find a 90% confidence interval for $\mu_1 - \mu_2$.
- (c) **Interpretation** Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 90% level of confidence, do professional football players tend to have a higher population mean height than professional basketball players?
- (d) **Check Requirements** Which distribution (standard normal or Student's t) did you use? Why? Do you need information about the height distributions? Explain.



16. **Botany: Iris** The following data represent petal lengths (in cm) for independent random samples of two species of iris (Reference: E. Anderson, *Bulletin American Iris Society*). Note: These data are also available for download at the Online Study Center.

Petal length (in cm) of *Iris virginica*: x_1 ; $n_1 = 35$

5.1 5.8 6.3 6.1 5.1 5.5 5.3 5.5 6.9 5.0 4.9 6.0 4.8 6.1 5.6 5.1
5.6 4.8 5.4 5.1 5.1 5.9 5.2 5.7 5.4 4.5 6.1 5.3 5.5 6.7 5.7 4.9
4.8 5.8 5.1

Petal length (in cm) of *Iris setosa*: x_2 ; $n_2 = 38$

1.5 1.7 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.4 1.7 1.0 1.7 1.9 1.6 1.4
1.5 1.4 1.2 1.3 1.5 1.3 1.6 1.9 1.4 1.6 1.5 1.4 1.6 1.2 1.9 1.5
1.6 1.4 1.3 1.7 1.5 1.7

- (a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 5.48$, $s_1 \approx 0.55$, $\bar{x}_2 \approx 1.49$, and $s_2 \approx 0.21$.
- (b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2 . Find a 99% confidence interval for $\mu_1 - \mu_2$.
- (c) **Interpretation** Explain what the confidence interval means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 99% level of confidence, is the population mean petal length of *Iris virginica* longer than that of *Iris setosa*?
- (d) **Check Requirements** Which distribution (standard normal or Student's t) did you use? Why? Do you need information about the petal length distributions? Explain.

17. **Myers–Briggs: Marriage Counseling** Isabel Myers was a pioneer in the study of personality types. She identified four basic personality preferences, which are described at length in the book *A Guide to the Development and Use of the Myers–Briggs Type Indicator* by Myers and McCaulley (Consulting Psychologists Press). Marriage counselors know that couples who have none of the four preferences in common may have a stormy marriage. Myers took a random sample of 375 married couples and found that 289 had two or more personality preferences in common. In another random sample of 571 married couples, it was found that only 23 had no preferences in common. Let p_1 be the population proportion of all married couples who have two or more personality preferences in common. Let p_2 be the population proportion of all married couples who have no personality preferences in common.

- (a) **Check Requirements** Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
- (b) Find a 99% confidence interval for $p_1 - p_2$.
- (c) **Interpretation** Explain the meaning of the confidence interval in part (a) in the context of this problem. Does the confidence interval contain all positive, all negative, or both positive and negative numbers? What does this tell

you (at the 99% confidence level) about the proportion of married couples with two or more personality preferences in common compared with the proportion of married couples sharing no personality preferences in common?

18. **Myers–Briggs: Marriage Counseling** Most married couples have two or three personality preferences in common (see reference in Problem 17). Myers used a random sample of 375 married couples and found that 132 had three preferences in common. Another random sample of 571 couples showed that 217 had two personality preferences in common. Let p_1 be the population proportion of all married couples who have three personality preferences in common. Let p_2 be the population proportion of all married couples who have two personality preferences in common.
- Check Requirements** Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
 - Find a 90% confidence interval for $p_1 - p_2$.
 - Interpretation** Examine the confidence interval in part (a) and explain what it means in the context of this problem. Does the confidence interval contain all positive, all negative, or both positive and negative numbers? What does this tell you about the proportion of married couples with three personality preferences in common compared with the proportion of couples with two preferences in common (at the 90% confidence level)?
19. **Yellowstone National Park: Old Faithful Geyser** The U.S. Geological Survey compiled historical data about Old Faithful Geyser (Yellowstone National Park) from 1870 to 1987. Some of these data are published in the book *The Story of Old Faithful*, by G. D. Marler (Yellowstone Association Press). Let x_1 be a random variable that represents the time interval (in minutes) between Old Faithful's eruptions for the years 1948 to 1952. Based on 9340 observations, the sample mean interval was $\bar{x}_1 = 63.3$ minutes. Let x_2 be a random variable that represents the time interval in minutes between Old Faithful's eruptions for the years 1983 to 1987. Based on 25,111 observations, the sample mean time interval was $\bar{x}_2 = 72.1$ minutes. Historical data suggest that $\sigma_1 = 9.17$ minutes and $\sigma_2 = 12.67$ minutes. Let μ_1 be the population mean of x_1 and let μ_2 be the population mean of x_2 .
- Check Requirements** Which distribution, normal or Student's t , do we use to approximate the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.
 - Compute a 99% confidence interval for $\mu_1 - \mu_2$.
 - Interpretation** Comment on the meaning of the confidence interval in the context of this problem. Does the interval consist of positive numbers only? negative numbers only? a mix of positive and negative numbers? Does it appear (at the 99% confidence level) that a change in the interval length between eruptions has occurred? Many geologic experts believe that the distribution of eruption times of Old Faithful changed after the major earthquake that occurred in 1959.
20. **Psychology: Parental Sensitivity** “Parental Sensitivity to Infant Cues: Similarities and Differences Between Mothers and Fathers” by M. V. Graham (*Journal of Pediatric Nursing*, Vol. 8, No. 6) reports a study of parental empathy for sensitivity cues and baby temperament (higher scores mean more empathy). Let x_1 be a random variable that represents the score of a mother on an empathy test (as regards her baby). Let x_2 be the empathy score of a father. A random sample of 32 mothers gave a sample mean of $\bar{x}_1 = 69.44$. Another random sample of 32 fathers gave $\bar{x}_2 = 59$. Assume that $\sigma_1 = 11.69$ and $\sigma_2 = 11.60$.
- Check Requirements** Which distribution, normal or Student's t , do we use to approximate the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.
 - Let μ_1 be the population mean of x_1 and let μ_2 be the population mean of x_2 . Find a 99% confidence interval for $\mu_1 - \mu_2$.

- (c) **Interpretation** Examine the confidence interval and explain what it means in the context of this problem. Does the confidence interval contain all positive, all negative, or both positive and negative numbers? What does this tell you about the relationship between average empathy scores for mothers compared with those for fathers at the 99% confidence level?
21. **Navajo Culture: Traditional Hogans** S. C. Jett is a professor of geography at the University of California, Davis. He and a colleague, V. E. Spencer, are experts on modern Navajo culture and geography. The following information is taken from their book *Navajo Architecture: Forms, History, Distributions* (University of Arizona Press). On the Navajo Reservation, a random sample of 210 permanent dwellings in the Fort Defiance region showed that 65 were traditional Navajo hogans. In the Indian Wells region, a random sample of 152 permanent dwellings showed that 18 were traditional hogans. Let p_1 be the population proportion of all traditional hogans in the Fort Defiance region, and let p_2 be the population proportion of all traditional hogans in the Indian Wells region.
- Check Requirements** Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
 - Find a 99% confidence interval for $p_1 - p_2$.
 - Interpretation** Examine the confidence interval and comment on its meaning. Does it include numbers that are all positive? all negative? mixed? What if it is hypothesized that Navajo who follow the traditional culture of their people tend to occupy hogans? Comment on the confidence interval for $p_1 - p_2$ in this context.
22. **Archaeology: Cultural Affiliation** “Unknown cultural affiliations and loss of identity at high elevations.” These words are used to propose the hypothesis that archaeological sites tend to lose their identity as altitude extremes are reached. This idea is based on the notion that prehistoric people tended *not* to take trade wares to temporary settings and/or isolated areas (Source: *Prehistoric New Mexico: Background for Survey*, by D. E. Stuart and R. P. Gauthier, University of New Mexico Press). As elevation zones of prehistoric people (in what is now the state of New Mexico) increased, there seemed to be a loss of artifact identification. Consider the following information.
- | Elevation Zone | Number of Artifacts | Number Unidentified |
|----------------|---------------------|---------------------|
| 7000–7500 ft | 112 | 69 |
| 5000–5500 ft | 140 | 26 |
- Let p_1 be the population proportion of unidentified archaeological artifacts at the elevation zone 7000–7500 feet in the given archaeological area. Let p_2 be the population proportion of unidentified archaeological artifacts at the elevation zone 5000–5500 feet in the given archaeological area.
- Check Requirements** Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
 - Find a 99% confidence interval for $p_1 - p_2$.
 - Interpretation** Explain the meaning of the confidence interval in the context of this problem. Does the confidence interval contain all positive numbers? all negative numbers? both positive and negative numbers? What does this tell you (at the 99% confidence level) about the comparison of the population proportion of unidentified artifacts at high elevations (7000–7500 feet) with the population proportion of unidentified artifacts at lower elevations (5000–5500 feet)? How does this relate to the stated hypothesis?
23. **Wildlife: Wolves** David E. Brown is an expert in wildlife conservation. In his book *The Wolf in the Southwest: The Making of an Endangered Species* (University of Arizona Press), he lists the following weights of adult gray wolves from two regions in Old Mexico.

Chihuahua region: x_1 variable in pounds

86	75	91	70	79
80	68	71	74	64

Durango region: x_2 variable in pounds

68	72	79	68	77	89	62	55	68
68	59	63	66	58	54	71	59	67

- (a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 = 75.80$ pounds, $s_1 = 8.32$ pounds, $\bar{x}_2 = 66.83$ pounds, and $s_2 = 8.87$ pounds.
- (b) **Check Requirements** Assuming that the original distribution of the weights of wolves are mound-shaped and symmetric, what distribution can be used to approximate the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.
- (c) Let μ_1 be the mean weight of the population of all gray wolves in the Chihuahua region. Let μ_2 be the mean weight of the population of all gray wolves in the Durango region. Find an 85% confidence interval for $\mu_1 - \mu_2$.
- (d) **Interpretation** Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 85% level of confidence, what can you say about the comparison of the average weight of gray wolves in the Chihuahua region with the average weight of gray wolves in the Durango region?
24. **Medical: Plasma Compress** At Community Hospital, the burn center is experimenting with a new plasma compress treatment. A random sample of $n_1 = 316$ patients with minor burns received the plasma compress treatment. Of these patients, it was found that 259 had no visible scars after treatment. Another random sample of $n_2 = 419$ patients with minor burns received no plasma compress treatment. For this group, it was found that 94 had no visible scars after treatment. Let p_1 be the population proportion of all patients with minor burns receiving the plasma compress treatment who have no visible scars. Let p_2 be the population proportion of all patients with minor burns not receiving the plasma compress treatment who have no visible scars.
- (a) **Check Requirements** Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
- (b) Find a 95% confidence interval for $p_1 - p_2$.
- (c) **Interpretation** Explain the meaning of the confidence interval found in part (b) in the context of the problem. Does the interval contain numbers that are all positive? all negative? both positive and negative? At the 95% level of confidence, does treatment with plasma compresses seem to make a difference in the proportion of patients with visible scars from minor burns?
25. **Psychology: Self-Esteem** Female undergraduates in randomized groups of 15 took part in a self-esteem study (“There’s More to Self-Esteem than Whether It Is High or Low: The Importance of Stability of Self-Esteem,” by M. H. Kernis et al., *Journal of Personality and Social Psychology*, Vol. 65, No. 6). The study measured an index of self-esteem from the points of view competence, social acceptance, and physical attractiveness. Let x_1 , x_2 , and x_3 be random variables representing the measure of self-esteem through x_1 (competence), x_2 (social acceptance), and x_3 (attractiveness). Higher index values mean a more positive influence on self-esteem.

Variable	Sample Size	Mean \bar{x}	Standard Deviation s	Population Mean
x_1	15	19.84	3.07	μ_1
x_2	15	19.32	3.62	μ_2
x_3	15	17.88	3.74	μ_3

- (a) Find an 85% confidence interval for $\mu_1 - \mu_2$.
 (b) Find an 85% confidence interval for $\mu_1 - \mu_3$.
 (c) Find an 85% confidence interval for $\mu_2 - \mu_3$.
 (d) **Interpretation** Comment on the meaning of each of the confidence intervals found in parts (a), (b), and (c). At the 85% confidence level, what can you say about the average differences in influence on self-esteem between competence and social acceptance? between competence and attractiveness? between social acceptance and attractiveness?

26. **Focus Problem: Wood Duck Nests** In the Focus Problem at the beginning of this chapter, a study was described comparing the hatch ratios of wood duck nesting boxes. Group I nesting boxes were well separated from each other and well hidden by available brush. There were a total of 474 eggs in group I boxes, of which a field count showed about 270 had hatched. Group II nesting boxes were placed in highly visible locations and grouped closely together. There were a total of 805 eggs in group II boxes, of which a field count showed about 270 had hatched.
- (a) Find a point estimate \hat{p}_1 for p_1 , the proportion of eggs that hatched in group I nest box placements. Find a 95% confidence interval for p_1 .
 (b) Find a point estimate \hat{p}_2 for p_2 , the proportion of eggs that hatched in group II nest box placements. Find a 95% confidence interval for p_2 .
 (c) Find a 95% confidence interval for $p_1 - p_2$. Does the interval indicate that the proportion of eggs hatched from group I nest boxes is higher than, lower than, or equal to the proportion of eggs hatched from group II nest boxes?
 (d) **Interpretation** What conclusions about placement of nest boxes can be drawn? In the article discussed in the Focus Problem, additional concerns are raised about the higher cost of placing and maintaining group I nest box placements. Also at issue is the cost efficiency per successful wood duck hatch.

27. **Critical Thinking: Different Confidence Levels**

- (a) Suppose a 95% confidence interval for the difference of means contains both positive and negative numbers. Will a 99% confidence interval based on the same data necessarily contain both positive and negative numbers? Explain. What about a 90% confidence interval? Explain.
 (b) Suppose a 95% confidence interval for the difference of proportions contains all positive numbers. Will a 99% confidence interval based on the same data necessarily contain all positive numbers as well? Explain. What about a 90% confidence interval? Explain.



28. **Expand Your Knowledge: Sample Size, Difference of Means** What about sample size? If we want a confidence interval with maximal margin of error E and level of confidence c , then Section 7.1 shows us which formulas to apply for a *single* mean μ and Section 7.3 shows us formulas for a *single* proportion p .
- (a) How about a *difference of means*? When σ_1 and σ_2 are known, the margin of error E for a $c\%$ confidence interval is

$$E = z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Let us make the simplifying assumption that we have *equal sample sizes* n so that $n = n_1 = n_2$. We also assume that $n \geq 30$. In this context, we get

$$E = z_c \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}} = \frac{z_c}{\sqrt{n}} \sqrt{\sigma_1^2 + \sigma_2^2}$$

Solve this equation for n and show that

$$n = \left(\frac{z_c}{E}\right)^2 (\sigma_1^2 + \sigma_2^2)$$

- (b) In Problem 15 (football and basketball player heights), suppose we want to be 95% sure that our estimate $\bar{x}_1 - \bar{x}_2$ for the difference $\mu_1 - \mu_2$ has a margin of error $E = 0.05$ foot. How large should the sample size be (assuming equal sample size—i.e., $n = n_1 = n_2$)? Since we do not know σ_1 or σ_2 and $n \geq 30$, use s_1 and s_2 , respectively, from the preliminary sample of Problem 15.
- (c) In Problem 16 (petal lengths of two iris species), suppose we want to be 90% sure that our estimate $\bar{x}_1 - \bar{x}_2$ for the difference $\mu_1 - \mu_2$ has a margin of error $E = 0.1$ cm. How large should the sample size be (assuming equal sample size—i.e., $n = n_1 = n_2$)? Since we do not know σ_1 or σ_2 and $n \geq 30$, use s_1 and s_2 , respectively, from the preliminary sample of Problem 16.



29.

Expand Your Knowledge: Sample Size, Difference of Proportions What about the sample size n for confidence intervals for the difference of proportions $p_1 - p_2$? Let us make the following assumptions: *equal sample sizes* $n = n_1 = n_2$ and *all four quantities* $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$, and $n_2\hat{q}_2$ *are greater than 5*. Those readers familiar with algebra can use the procedure outlined in Problem 28 to show that if we have preliminary estimates \hat{p}_1 and \hat{p}_2 and a given maximal margin of error E for a specified confidence level c , then the sample size n should be at least

$$n = \left(\frac{z_c}{E} \right)^2 (\hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2)$$

However, if we have no preliminary estimates for \hat{p}_1 and \hat{p}_2 , then the theory similar to that used in this section tells us that the sample size n should be at least

$$n = \frac{1}{2} \left(\frac{z_c}{E} \right)^2$$



30.

- (a) In Problem 17 (Myers–Briggs personality type indicators in common for married couples), suppose we want to be 99% confident that our estimate $\hat{p}_1 - \hat{p}_2$ for the difference $p_1 - p_2$ has a maximal margin of error $E = 0.04$. Use the preliminary estimates $\hat{p}_1 = 289/375$ for the proportion of couples sharing two personality traits and $\hat{p}_2 = 23/571$ for the proportion having no traits in common. How large should the sample size be (assuming equal sample size—i.e., $n = n_1 = n_2$)?
- (b) Suppose that in Problem 17 we have no preliminary estimates for \hat{p}_1 and \hat{p}_2 and we want to be 95% confident that our estimate $\hat{p}_1 - \hat{p}_2$ for the difference $p_1 - p_2$ has a maximal margin of error $E = 0.05$. How large should the sample size be (assuming equal sample size—i.e., $n = n_1 = n_2$)?

Expand Your Knowledge: Software Approximation for Degrees of Freedom Given x_1 and x_2 distributions that are normal or approximately normal with unknown σ_1 and σ_2 , the value of t corresponding to $\bar{x}_1 - \bar{x}_2$ has a distribution that is approximated by a Student's t distribution. We use the convention that the degrees of freedom are approximately the smaller of $n_1 - 1$ and $n_2 - 1$. However, a more accurate estimate for the appropriate degrees of freedom is given by *Satterthwaite's formula*

$$d.f. \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2} \right)^2}$$

where s_1 , s_2 , n_1 , and n_2 are the respective sample standard deviations and sample sizes of independent random samples from the x_1 and x_2 distributions. This is the approximation used by most statistical software. When both n_1 and n_2 are 5 or larger, it is quite accurate. The degrees of freedom computed from this formula are either truncated or not rounded.

- (a) Use the data of Problem 14 (weights of pro football and pro basketball players) to compute $d.f.$ using the formula. Compare the result to 36, the value generated by Minitab. Did Minitab truncate?
- (b) Compute a 99% confidence interval using $d.f. \approx 36$. (Using Table 6 requires using $d.f. = 35$.) Compare this confidence interval to the one you computed in Problem 14. Which $d.f.$ gives the longer interval?



31.

Expand Your Knowledge: Pooled Two-Sample Procedures Under the condition that both populations have equal standard deviations ($\sigma_1 = \sigma_2$), we can pool the standard deviations and use a Student's t distribution with degrees of freedom $d.f. = n_1 + n_2 - 2$ to find the margin of error of a c confidence interval for $\mu_1 - \mu_2$. This technique demonstrates another commonly used method of computing confidence intervals for $\mu_1 - \mu_2$.

PROCEDURE

HOW TO FIND A CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$ WHEN $\sigma_1 = \sigma_2$

Requirements

Consider two *independent* random samples, where

\bar{x}_1 and \bar{x}_2 are sample means from populations 1 and 2

s_1 and s_2 are sample standard deviations from populations 1 and 2

n_1 and n_2 are sample sizes from populations 1 and 2

If you can assume that both population distributions 1 and 2 are normal or at least mound-shaped and symmetric, then any sample sizes n_1 and n_2 will work. If you cannot assume this, then use sample sizes $n_1 \geq 30$ and $n_2 \geq 30$.

Confidence interval for $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2$

$$(\hat{x}_1 - \hat{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

where

$$E = t_c s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (\text{pooled standard deviation})$$

c = confidence level ($0 < c < 1$)

t_c = critical value for confidence level c and degrees of freedom

$d.f. = n_1 + n_2 - 2$ (See Table 6 of Appendix II.)

Note: With statistical software, select pooled variance or equal variance options.

- (a) There are many situations in which we want to compare means from populations having standard deviations that are equal. The pooled standard deviation method applies even if the standard deviations are known to be only approximately equal. (See Section 10.4 for methods to test that $\sigma_1 = \sigma_2$.) Consider Problem 23 regarding weights of grey wolves in two regions. Notice that $s_1 = 8.32$ pounds and $s_2 = 8.87$ pounds are fairly close. Use the method of pooled standard deviation to find an 85% confidence interval for the difference in population mean weights of grey wolves in the Chihuahua region compared with those in the Durango region.
- (b) Compare the confidence interval computed in part (a) with that computed in Problem 23. Which method has the larger degrees of freedom? Which method has the longer confidence interval?



SUMMARY

Chapter Review

How do you get information about a population by looking at a random sample? One way is to use point estimates and confidence intervals.

- Point estimates and their corresponding parameters are
 \bar{x} for μ $\bar{x}_1 - \bar{x}_2$ for $\mu_1 - \mu_2$
 \hat{p} for p $\hat{p}_1 - \hat{p}_2$ for $p_1 - p_2$
- Confidence intervals are of the form
 $\text{point estimate} - E < \text{parameter} < \text{point estimate} + E$
- E is the maximal margin of error. Specific values of E depend on the parameter, level of confidence, whether population standard deviations are known, sample size, and the shapes of the original population distributions.
For μ : $E = z_c \frac{\sigma}{\sqrt{n}}$ when σ is known;
 $E = t_c \frac{s}{\sqrt{n}}$ with d.f. = $n - 1$ when σ is unknown

For p : $E = z_c \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$ when $n\hat{p} > 5$ and $n\hat{q} > 5$.

For $\mu_1 - \mu_2$: $E = z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ when σ_1 and σ_2 are known

$E = t_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ when σ_1 or σ_2 is unknown with d.f. = smaller of $n_1 - 1$ or $n_2 - 1$

Software uses Satterthwaite's approximation for d.f.

For $p_1 - p_2$: $E = z_c \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$ for sufficiently large n

- Confidence intervals have an associated probability c called the confidence level. For a given sample size, the proportion of all corresponding confidence intervals that contain the parameter in question is c .

IMPORTANT WORDS & SYMBOLS

Section 7.1

- Point estimate for μ 335
Confidence level c 335
Critical values z_c 335
Maximal margin of error E 337
 c confidence interval 338
Sample size for estimating μ 342

Section 7.2

- Student's t distribution 347
Degrees of freedom d.f. 348
Critical values t_c 349

Section 7.3

- Point estimate for p , \hat{p} 360

- Confidence interval for p 362
Margin of error for polls 365
Sample size for estimating p 366

Section 7.4

- Independent samples 373
Dependent samples 373
 $\bar{x}_1 - \bar{x}_2$ sampling distribution 374
Confidence interval for $\mu_1 - \mu_2$ (σ_1 and σ_2 known) 375
Confidence interval for $\mu_1 - \mu_2$ (σ_1 and σ_2 unknown) 377
Confidence interval for $p_1 - p_2$ 381
Satterthwaite's Formula for d.f. 393

VIEWPOINT

All Systems Go?

On January 28, 1986, the space shuttle Challenger caught fire and blew up only seconds after launch. A great deal of good engineering had gone into the design of the Challenger. However, when a system has several confidence levels operating at once, it can happen, in rare cases, that risks will increase rather than cancel out. (See Chapter Review Problem 19.) Diane Vaughn is a professor of sociology at Boston College and author of the book *The Challenger Launch Decision*

(University of Chicago Press). Her book contains an excellent discussion of risks, the normalization of deviants, and cost/safety tradeoffs. Vaughn's book is described as "a remarkable and important analysis of how social structures can induce consequential errors in a decision process" (Robert K. Merton, Columbia University).

CHAPTER REVIEW PROBLEMS

- Statistical Literacy** In your own words, carefully explain the meanings of the following terms: *point estimate*, *critical value*, *maximal margin of error*, *confidence level*, and *confidence interval*.
- Critical Thinking** Suppose you are told that a 95% confidence interval for the average price of a gallon of regular gasoline in your state is from \$3.15 to \$3.45. Use the fact that the confidence interval for the mean has the form $\bar{x} - E$ to $\bar{x} + E$ to compute the sample mean and the maximal margin of error E .
- Critical Thinking** If you have a 99% confidence interval for μ based on a simple random sample,
 - is it correct to say that the *probability* that μ is in the specified interval is 99%? Explain.
 - is it correct to say that in the long run, if you computed many, many confidence intervals using the prescribed method, about 99% of such intervals would contain μ ? Explain.

For Problems 4–19, categorize each problem according to the parameter being estimated: proportion p , mean μ , difference of means $\mu_1 - \mu_2$, or difference of proportions $p_1 - p_2$. Then solve the problem.

- Auto Insurance: Claims** Anystate Auto Insurance Company took a random sample of 370 insurance claims paid out during a 1-year period. The average claim paid was \$1570. Assume $\sigma = \$250$. Find 0.90 and 0.99 confidence intervals for the mean claim payment.
- Psychology: Closure** Three experiments investigating the relationship between need for cognitive closure and persuasion were reported in "Motivated Resistance and Openness to Persuasion in the Presence or Absence of Prior Information" by A. W. Kruglanski (*Journal of Personality and Social Psychology*, Vol. 65, No. 5, pp. 861–874). Part of the study involved administering a "need for closure scale" to a group of students enrolled in an introductory psychology course. The "need for closure scale" has scores ranging from 101 to 201. For the 73 students in the highest quartile of the distribution, the mean score was $\bar{x} = 178.70$. Assume a population standard deviation of $\sigma = 7.81$. These students were all classified as high on their need for closure. Assume that the 73 students represent a random sample of all students who are classified as high on their need for closure. Find a 95% confidence interval for the population mean score μ on the "need for closure scale" for all students with a high need for closure.
- Psychology: Closure** How large a sample is needed in Problem 5 if we wish to be 99% confident that the sample mean score is within 2 points of the population mean score for students who are high on the need for closure?
- Archaeology: Excavations** The Wind Mountain archaeological site is located in southwestern New Mexico. Wind Mountain was home to an ancient culture of prehistoric Native Americans called Anasazi. A random sample of excavations at Wind Mountain gave the following depths (in centimeters) from present-day surface grade to the location of significant archaeological artifacts (Source: *Mimbres Mogollon Archaeology*, by A. Woosley and A. McIntyre, University of New Mexico Press).

85	45	120	80	75	55	65	60
65	95	90	70	75	65	68	

- (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 74.2$ cm and $s \approx 18.3$ cm.
 (b) Compute a 95% confidence interval for the mean depth μ at which archaeological artifacts from the Wind Mountain excavation site can be found.
8. **Archaeology: Pottery** Shards of clay vessels were put together to reconstruct rim diameters of the original ceramic vessels found at the Wind Mountain archaeological site (see source in Problem 7). A random sample of ceramic vessels gave the following rim diameters (in centimeters):
- | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| 15.9 | 13.4 | 22.1 | 12.7 | 13.1 | 19.6 | 11.7 | 13.5 | 17.7 | 18.1 |
|------|------|------|------|------|------|------|------|------|------|
- (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 15.8$ cm and $s \approx 3.5$ cm.
 (b) Compute an 80% confidence interval for the population mean μ of rim diameters for such ceramic vessels found at the Wind Mountain archaeological site.
9. **Telephone Interviews: Survey** The National Study of the Changing Work Force conducted an extensive survey of 2958 wage and salaried workers on issues ranging from relationships with their bosses to household chores. The data were gathered through hour-long telephone interviews with a nationally representative sample (*The Wall Street Journal*). In response to the question “What does success mean to you?” 1538 responded, “Personal satisfaction from doing a good job.” Let p be the population proportion of all wage and salaried workers who would respond the same way to the stated question. Find a 90% confidence interval for p .
10. **Telephone Interviews: Survey** How large a sample is needed in Problem 9 if we wish to be 95% confident that the sample percentage of those equating success with personal satisfaction is within 1% of the population percentage? Hint: Use $p \approx 0.52$ as a preliminary estimate.
11. **Archaeology: Pottery** Three-circle, red-on-white is one distinctive pattern painted on ceramic vessels of the Anasazi period found at the Wind Mountain archaeological site (see source for Problem 7). At one excavation, a sample of 167 potsherds indicated that 68 were of the three-circle, red-on-white pattern.
 (a) Find a point estimate \hat{p} for the proportion of all ceramic potsherds at this site that are of the three-circle, red-on-white pattern.
 (b) Compute a 95% confidence interval for the population proportion p of all ceramic potsherds with this distinctive pattern found at the site.
12. **Archaeology: Pottery** Consider the three-circle, red-on-white pattern discussed in Problem 11. How many ceramic potsherds must be found and identified if we are to be 95% confident that the sample proportion \hat{p} of such potsherds is within 6% of the population proportion of three-circle, red-on-white patterns found at this excavation site? Hint: Use the results of Problem 11 as a preliminary estimate.
13. **Agriculture: Bell Peppers** The following data represent soil water content (percent water by volume) for independent random samples of soil taken from two experimental fields growing bell peppers (Reference: *Journal of Agricultural, Biological, and Environmental Statistics*). Note: These data are also available for download at the Online Study Center.
- Soil water content from field I: x_1 ; $n_1 = 72$**
- | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|
| 15.1 | 11.2 | 10.3 | 10.8 | 16.6 | 8.3 | 9.1 | 12.3 | 9.1 | 14.3 |
| 10.7 | 16.1 | 10.2 | 15.2 | 8.9 | 9.5 | 9.6 | 11.3 | 14.0 | 11.3 |
| 15.6 | 11.2 | 13.8 | 9.0 | 8.4 | 8.2 | 12.0 | 13.9 | 11.6 | 16.0 |
| 9.6 | 11.4 | 8.4 | 8.0 | 14.1 | 10.9 | 13.2 | 13.8 | 14.6 | 10.2 |
| 11.5 | 13.1 | 14.7 | 12.5 | 10.2 | 11.8 | 11.0 | 12.7 | 10.3 | 10.8 |
| 11.0 | 12.6 | 10.8 | 9.6 | 11.5 | 10.6 | 11.7 | 10.1 | 9.7 | 9.7 |
| 11.2 | 9.8 | 10.3 | 11.9 | 9.7 | 11.3 | 10.4 | 12.0 | 11.0 | 10.7 |
| 8.8 | 11.1 | | | | | | | | |



Soil water content from field II: x_2 ; $n_2 = 80$

12.1	10.2	13.6	8.1	13.5	7.8	11.8	7.7	8.1	9.2
14.1	8.9	13.9	7.5	12.6	7.3	14.9	12.2	7.6	8.9
13.9	8.4	13.4	7.1	12.4	7.6	9.9	26.0	7.3	7.4
14.3	8.4	13.2	7.3	11.3	7.5	9.7	12.3	6.9	7.6
13.8	7.5	13.3	8.0	11.3	6.8	7.4	11.7	11.8	7.7
12.6	7.7	13.2	13.9	10.4	12.8	7.6	10.7	10.7	10.9
12.5	11.3	10.7	13.2	8.9	12.9	7.7	9.7	9.7	11.4
11.9	13.4	9.2	13.4	8.8	11.9	7.1	8.5	14.0	14.2

- (a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 11.42$, $s_1 \approx 2.08$, $\bar{x}_2 \approx 10.65$, and $s_2 \approx 3.03$.
- (b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2 . Find a 95% confidence interval for $\mu_1 - \mu_2$.
- (c) **Interpretation** Explain what the confidence interval means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 95% level of confidence, is the population mean soil water content of the first field higher than that of the second field?
- (d) Which distribution (standard normal or Student's t) did you use? Why? Do you need information about the soil water content distributions?
14. **Stocks: Retail and Utility** How profitable are different sectors of the stock market? One way to answer such a question is to examine profit as a percentage of stockholder equity. A random sample of 32 retail stocks such as Toys "R" Us, Best Buy, and Gap was studied for x_1 , profit as a percentage of stockholder equity. The result was $\bar{x}_1 = 13.7$. A random sample of 34 utility (gas and electric) stocks such as Boston Edison, Wisconsin Energy, and Texas Utilities was studied for x_2 , profit as a percentage of stockholder equity. The result was $\bar{x}_2 = 10.1$ (Source: *Fortune 500*, Vol. 135, No. 8). Assume that $\sigma_1 = 4.1$ and $\sigma_2 = 2.7$.
- (a) Let μ_1 represent the population mean profit as a percentage of stockholder equity for retail stocks, and let μ_2 represent the population mean profit as a percentage of stockholder equity for utility stocks. Find a 95% confidence interval for $\mu_1 - \mu_2$.
- (b) **Interpretation** Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 95% level of confidence, does it appear that the profit as a percentage of stockholder equity for retail stocks is higher than that for utility stocks?
15. **Wildlife: Wolves** A random sample of 18 adult male wolves from the Canadian Northwest Territories gave an average weight $\bar{x}_1 = 98$ pounds, with estimated sample standard deviation $s_1 = 6.5$ pounds. Another sample of 24 adult male wolves from Alaska gave an average weight $\bar{x}_2 = 90$ pounds, with estimated sample standard deviation $s_2 = 7.3$ pounds (Source: *The Wolf* by L. D. Mech, University of Minnesota Press).
- (a) Let μ_1 represent the population mean weight of adult male wolves from the Northwest Territories, and let μ_2 represent the population mean weight of adult male wolves from Alaska. Find a 75% confidence interval for $\mu_1 - \mu_2$.
- (b) **Interpretation** Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 75% level of confidence, does it appear that the average weight of adult male wolves from the Northwest Territories is greater than that of the Alaska wolves?
16. **Wildlife: Wolves** A random sample of 17 wolf litters in Ontario, Canada, gave an average of $\bar{x}_1 = 4.9$ wolf pups per litter, with estimated sample standard deviation $s_1 = 1.0$. Another random sample of 6 wolf litters in Finland gave an average of $\bar{x}_2 = 2.8$ wolf pups per litter, with sample standard deviation $s_2 = 1.2$ (see source for Problem 15).

- (a) Find an 85% confidence interval for $\mu_1 - \mu_2$, the difference in population mean litter size between Ontario and Finland.
- (b) **Interpretation** Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 85% level of confidence, does it appear that the average litter size of wolf pups in Ontario is greater than the average litter size in Finland?
17. **Survey Response: Validity** The book *Survey Responses: An Evaluation of Their Validity* by E. J. Wentland and K. Smith (Academic Press), includes studies reporting accuracy of answers to questions from surveys. A study by Locander et al. considered the question “Are you a registered voter?” Accuracy of response was confirmed by a check of city voting records. Two methods of survey were used: a face-to-face interview and a telephone interview. A random sample of 93 people were asked the voter registration question face-to-face. Seventy-nine respondents gave accurate answers (as verified by city records). Another random sample of 83 people were asked the same question during a telephone interview. Seventy-four respondents gave accurate answers. Assume the samples are representative of the general population.
- (a) Let p_1 be the population proportion of all people who answer the voter registration question accurately during a face-to-face interview. Let p_2 be the population proportion of all people who answer the question accurately during a telephone interview. Find a 95% confidence interval for $p_1 - p_2$.
- (b) **Interpretation** Does the interval contain numbers that are all positive? all negative? mixed? Comment on the meaning of the confidence interval in the context of this problem. At the 95% level, do you detect any difference in the proportion of accurate responses from face-to-face interviews compared with the proportion of accurate responses from telephone interviews?
18. **Survey Response: Validity** Locander et al. (see reference in Problem 17) also studied the accuracy of responses on questions involving more sensitive material than voter registration. From public records, individuals were identified as having been charged with drunken driving not less than 6 months or more than 12 months from the starting date of the study. Two random samples from this group were studied. In the first sample of 30 individuals, the respondents were asked in a face-to-face interview if they had been charged with drunken driving in the last 12 months. Of these 30 people interviewed face-to-face, 16 answered the question accurately. The second random sample consisted of 46 people who had been charged with drunken driving. During a telephone interview, 25 of these responded accurately to the question asking if they had been charged with drunken driving during the past 12 months. Assume the samples are representative of all people recently charged with drunken driving.
- (a) Let p_1 represent the population proportion of all people with recent charges of drunken driving who respond accurately to a face-to-face interview asking if they have been charged with drunken driving during the past 12 months. Let p_2 represent the population proportion of people who respond accurately to the question when it is asked in a telephone interview. Find a 90% confidence interval for $p_1 - p_2$.
- (b) **Interpretation** Does the interval found in part (a) contain numbers that are all positive? all negative? mixed? Comment on the meaning of the confidence interval in the context of this problem. At the 90% level, do you detect any differences in the proportion of accurate responses to the question from face-to-face interviews as compared with the proportion of accurate responses from telephone interviews?



19.

Expand Your Knowledge: Two Confidence Intervals What happens if we want several confidence intervals to hold at the same time (concurrently)? Do we still have the same level of confidence we had for *each* individual interval?

- (a) Suppose we have two independent random variables x_1 and x_2 with respective population means μ_1 and μ_2 . Let us say that we use sample data to construct two 80% confidence intervals.

Confidence Interval	Confidence Level
$A_1 < \mu_1 < B_1$	0.80
$A_2 < \mu_2 < B_2$	0.80

Now, what is the probability that *both* intervals hold at the same time? Use methods of Section 4.2 to show that

$$P(A_1 < \mu_1 < B_1 \text{ and } A_2 < \mu_2 < B_2) = 0.64$$

Hint: You are combining independent events. If the confidence is 64% that both intervals hold concurrently, explain why the risk that at least one interval does not hold (i.e., fails) must be 36%.

- (b) Suppose we want *both* intervals to hold with 90% confidence (i.e., only 10% risk level). How much confidence c should each interval have to achieve this combined level of confidence? (Assume that each interval has the same confidence level c .)

$$\begin{aligned} \text{Hint: } P(A_1 < \mu_1 < B_1 \text{ and } A_2 < \mu_2 < B_2) &= 0.90 \\ P(A_1 < \mu_1 < B_1) \times P(A_2 < \mu_2 < B_2) &= 0.90 \\ c \times c &= 0.90 \end{aligned}$$

Now solve for c .

- (c) If we want *both* intervals to hold at the 90% level of confidence, then the individual intervals must hold at a *higher* level of confidence. Write a brief but detailed explanation of how this could be of importance in a large, complex engineering design such as a rocket booster or a spacecraft.

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

Judy Griesedieck/Encyclopedia/Corbis



Digging clams

1. Garrison Bay is a small bay in Washington state. A popular recreational activity in the bay is clam digging. For several years, this harvest has been monitored and the size distribution of clams recorded. Data for lengths and widths of little neck clams (*Protothaca staminea*) were recorded by a method of systematic sampling in a study done by S. Scherba and V. F. Gallucci ("The Application of Systematic Sampling to a Study of Infaunal Variation in a Soft Substrate Intertidal Environment," *Fishery Bulletin*, Vol. 74, pp. 937–948). The data in Tables 7-4 and 7-5 give lengths and widths for 35 little neck clams.
 - (a) Use a calculator to compute the sample mean and sample standard deviation for the lengths and widths. Compute the coefficient of variation for each.
 - (b) Compute a 95% confidence interval for the population mean length of all Garrison Bay little neck clams.
 - (c) How many more little neck clams would be needed in a sample if you wanted to be 95% sure that the sample mean length is within a maximal margin of error of 10 mm of the population mean length?

TABLE 7-4 Lengths of Little Neck Clams (mm)

530	517	505	512	487	481	485	479	452	468
459	449	472	471	455	394	475	335	508	486
474	465	420	402	410	393	389	330	305	169
91	537	519	509	511					

TABLE 7-5 Widths of Little Neck Clams (mm)

494	477	471	413	407	427	408	430	395	417
394	397	402	401	385	338	422	288	464	436
414	402	383	340	349	333	356	268	264	141
77	498	456	433	447					

- (d) Compute a 95% confidence interval for the population mean width of all Garrison Bay little neck clams.
- (e) How many more little neck clams would be needed in a sample if you wanted to be 95% sure that the sample mean width is within a maximal margin of error of 10 mm of the population mean width?
- (f) The *same* 35 clams were used for measures of length and width. Are the sample measurements length and width independent or dependent? Why?
2. Examine Figure 7-8, “Fall Back.”
- (a) Of the 1024 adults surveyed, 66% were reported to favor daylight saving time. How many people in the sample preferred daylight saving time? Using the statistic $\hat{p} = 0.66$ and sample size $n = 1024$, find a 95% confidence interval for the proportion of people p who favor daylight saving time. How could you report this information in terms of a margin of error?
- (b) Look at Figure 7-8 to find the sample statistic \hat{p} for the proportion of people preferring standard time. Find a 95% confidence interval for the population proportion p of people who favor standard time. Report the same information in terms of a margin of error.
3. Examine Figure 7-9, “Coupons: Limited Use.”
- (a) Use Figure 7-9 to estimate the percentage of merchandise coupons that were redeemed. Also estimate the percentage dollar value of the coupons that were redeemed. Are these numbers approximately equal?

FIGURE 7-8

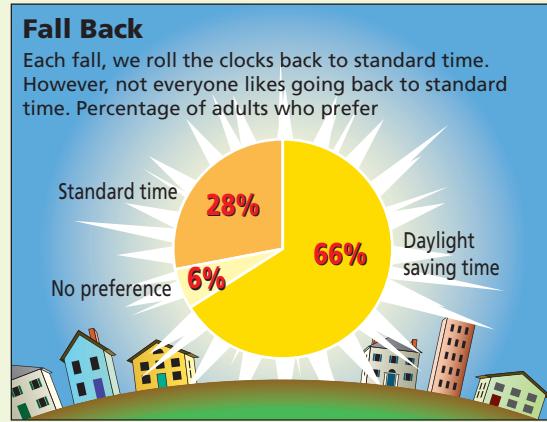
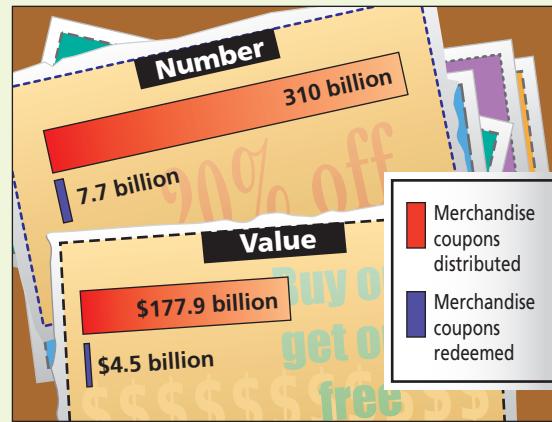


FIGURE 7-9 Coupons: Limited Use



- (b) Suppose you are a marketing executive working for a national chain of toy stores. You wish to estimate the percentage of coupons that will be redeemed for the toy stores. How many coupons should you check to be 95% sure that the percentage of coupons redeemed is within 1% of the population proportion of all coupons redeemed for the toy store?
- (c) Use the results of part (a) as a preliminary estimate for p , the percentage of coupons that are redeemed, and redo part (b).
- (d) Suppose you sent out 937 coupons and found that 27 were redeemed. Explain why you could be 95% confident that the proportion of such coupons redeemed in the future would be between 1.9% and 3.9%.
- (e) Suppose the dollar value of a collection of coupons was \$10,000. Use the data in Figure 7-9 to find the expected value and standard deviation of the dollar value of the redeemed coupons. What is the probability that between \$225 and \$275 (out of the \$10,000) is redeemed?

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. In this chapter, we have studied confidence intervals. Carefully read the following statements about confidence intervals:
 - (a) Once the endpoints of the confidence interval are numerically fixed, the parameter in question (either μ or p) does or does not fall inside the “fixed” interval.
 - (b) A given fixed interval either does or does not contain the parameter μ or p ; therefore, the probability is 1 or 0 that the parameter is in the interval.
 Next, read the following statements. Then discuss all four statements in the context of what we actually mean by a confidence interval.
- (c) Nontrivial probability statements can be made only about variables, not constants.
- (d) The confidence level c represents the proportion of all (fixed) intervals that would contain the parameter if we repeated the process many, many times.
2. Throughout Chapter 7, we have used the normal distribution, the central limit theorem, and the Student’s t distribution.
 - (a) Give a brief outline describing how confidence intervals for means use the normal distribution or Student’s t distribution in their basic construction.
 - (b) Give a brief outline describing how the normal approximation to the binomial distribution is used in the construction of confidence intervals for a proportion p .
 - (c) Give a brief outline describing how the sample size for a predetermined error tolerance and level of confidence is determined from the normal distribution.
3. When the results of a survey or a poll are published, the sample size is usually given, as well as the margin of error. For example, suppose the *Honolulu Star Bulletin* reported that it surveyed 385 Honolulu residents and 78% said they favor mandatory jail sentences for people convicted of driving under the influence of drugs or alcohol (with margin of error of 3 percentage points in either direction). Usually the confidence level of the interval is not given, but it is standard practice to use the margin of error for a 95% confidence interval when no other confidence level is given.

- (a) The paper reported a point estimate of 78%, with margin of error of $\pm 3\%$. Write this information in the form of a confidence interval for p , the population proportion of residents favoring mandatory jail sentences for people convicted of driving under the influence. What is the assumed confidence level?
- (b) The margin of error is simply the error due to using a sample instead of the entire population. It does not take into account the bias that might be introduced by the wording of the question, by the truthfulness of the respondents, or by other factors. Suppose the question was asked in this fashion: “Considering the devastating injuries suffered by innocent victims in auto accidents caused by drunken or drugged drivers, do you favor a mandatory jail sentence for those convicted of driving under the influence of drugs or alcohol?” Do you think the wording of the question would influence the respondents? Do you think the population proportion of those favoring mandatory jail sentences would be accurately represented by a confidence interval based on responses to such a question? Explain your answer.

If the question had been “Considering existing overcrowding of our prisons, do you favor a mandatory jail sentence for people convicted of driving under the influence of drugs or alcohol?” do you think the population proportion of those favoring mandatory sentences would be accurately represented by a confidence interval based on responses to such a question? Explain.

USING TECHNOLOGY

Application 1

Finding a Confidence Interval for a Population Mean μ

Cryptanalysis, the science of breaking codes, makes extensive use of language patterns. The frequency of various letter combinations is an important part of the study. A letter combination consisting of a single letter is a monograph, while combinations consisting of two letters are called digraphs, and those with three letters are called trigraphs. In the English language, the most frequent digraph is the letter combination TH.

The *characteristic rate* of a letter combination is a measurement of its rate of occurrence. To compute the characteristic rate, count the number of occurrences of a given letter combination and divide by the number of letters in the text. For instance, to estimate the characteristic rate of the digraph TH, you could select a newspaper text and pick a random starting place. From that place, mark off 2000 letters and count the number of times that TH occurs. Then divide the number of occurrences by 2000.

The characteristic rate of a digraph can vary slightly depending on the style of the author, so to estimate an overall characteristic frequency, you want to consider several samples of newspaper text by different authors. Suppose you did this with a random sample of 15 articles and found the characteristic rates of the digraph TH in the articles. The results follow.

0.0275	0.0230	0.0300	0.0255
0.0280	0.0295	0.0265	0.0265
0.0240	0.0315	0.0250	0.0265
0.0290	0.0295	0.0275	

- (a) Find a 95% confidence interval for the mean characteristic rate of the digraph TH.
- (b) Repeat part (a) for a 90% confidence interval.
- (c) Repeat part (a) for an 80% confidence interval.
- (d) Repeat part (a) for a 70% confidence interval.
- (e) Repeat part (a) for a 60% confidence interval.
- (f) For each confidence interval in parts (a)–(e), compute the length of the given interval. Do you notice a

relation between the confidence level and the length of the interval?

A good reference for cryptanalysis is a book by Sinkov: Sinkov, Abraham. *Elementary Cryptanalysis*. New York: Random House.

In the book, other common digraphs and trigraphs are given.

Application 2

Confidence Interval Demonstration

When we generate different random samples of the same size from a population, we discover that \bar{x} varies from sample to sample. Likewise, different samples produce different confidence intervals for μ . The endpoints $\bar{x} \pm E$ of a confidence interval are statistical variables. A 90% confidence interval tells us that if we obtain lots of confidence intervals (for the same sample size), then the proportion of all intervals that will turn out to contain μ is 90%.

- (a) Use the technology of your choice to generate 10 large random samples from a population with a known mean μ .
- (b) Construct a 90% confidence interval for the mean for each sample.
- (c) Examine the confidence intervals and note the percentage of the intervals that contain the population mean μ . We have 10 confidence intervals. Will exactly 90% of 10 intervals always contain μ ? Explain. What if we have 1000 intervals?

Technology Hints for Confidence Interval Demonstration

TI-84Plus/TI-83Plus/TI-nspire

The TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) generates random samples from uniform, normal, and binomial distributions. Press the MATH key and select PRB. Choice 5:randInt(lower, upper, sample size n) generates random samples of size n from the integers between the specified lower and upper values. Choice 6:randNorm(μ , σ , sample size n) generates random samples of size n from a normal distribution with specified mean and standard deviation.

Choice 7:randBin(number of trials, p , sample size) generates samples of the specified size from the designated binomial distribution. Under STAT, select EDIT and highlight the list name, such as L1. At the = sign, use the MATH key to access the desired population distribution. Finally, use Zinterval under the TESTS option of the STAT key to generate 90% confidence intervals.

Excel 2007

On the Home screen, click the Data tab. Then in the Analysis Group, click Data Analysis. In the resulting dialogue box, select Random Number Generator. In that dialogue box, the number of variables refers to the number of samples. The number of random numbers refers to the number of data values in each sample. Select the population distribution (uniform, normal, binomial, etc.). When you click OK the data appear in columns on a spreadsheet, with each sample appearing in a separate column. Click on the Insert function f_x . In the dialogue box, select Statistical for the category and then select Confidence. In the dialogue box for Confidence, $\alpha = 1 - c$, so for a 90% confidence interval, enter 0.10 for alpha. Then enter the population standard deviation σ , and the sample size. The resulting output gives the value of the maximal margin of error E for the confidence interval for the mean μ . Note that if you use the population standard deviation σ in the function, the value of E will be the same for all samples of the same size. Next, find the sample mean \bar{x} for each sample (use Insert function f_x with Statistical for category in the dialogue box and select Average). Finally, construct the endpoints $\bar{x} \pm E$ of the confidence interval for each sample.

Minitab

Minitab provides options for sampling from a variety of distributions. To generate random samples from a specific distribution, use the menu selection Calc \rightarrow Random Data \rightarrow and then select the population distribution. In the dialogue box, the *number of rows of data* represents the *sample size*. The *number of samples* corresponds to the number of columns selected for data storage. For example, C1–C10 in data storage produces 10 different random samples of the specified size. Use the menu selection Stat \rightarrow Basic Statistics \rightarrow 1 sample z to generate confidence intervals for the mean μ

from each sample. In the variables box, list all the columns containing your samples. For instance, using C1–C10 in the variables list will produce confidence intervals for each of the 10 samples stored in columns C1 through C10.

The Minitab display shows 90% confidence intervals for 10 different random samples of size 50 taken from a normal distribution with $\mu = 30$ and $\sigma = 4$. Notice that, as expected, 9 out of 10 of the intervals contain $\mu = 30$.

Minitab Display

z Confidence Intervals (Samples from a Normal Population with $\mu = 30$ and $\sigma = 4$)					
Variable	N	Mean	StDev	SE Mean	90.0 % CI
C1	50	30.265	4.300	0.566	(29.334, 31.195)
C2	50	31.040	3.957	0.566	(30.109, 31.971)
C3	50	29.940	4.195	0.566	(29.010, 30.871)
C4	50	30.753	3.842	0.566	(29.823, 31.684)
C5	50	30.047	4.174	0.566	(29.116, 30.977)
C6	50	29.254	4.423	0.566	(28.324, 30.185)
C7	50	29.062	4.532	0.566	(28.131, 29.992)
C8	50	29.344	4.487	0.566	(28.414, 30.275)
C9	50	30.062	4.199	0.566	(29.131, 30.992)
C10	50	29.989	3.451	0.566	(29.058, 30.919)

SPSS

SPSS uses a Student's *t* distribution to generate confidence intervals for the mean and difference of means. Use the menu choices Analyze \rightarrow Compare Means and then One-Sample T Test or Independent-Sample T Tests for confidence intervals for a single mean or difference of means, respectively. In the dialogue box, use 0 for the test value. Click Options... to provide the confidence level.

To generate 10 random samples of size $n = 30$ from a normal distribution with $\mu = 30$ and $\sigma = 4$, first enter consecutive integers from 1 to 30 in a column of the data editor. Then, under variable view, enter the variable names Sample1 through Sample10. Use the menu choices Transform \rightarrow Compute Variable. In the dialogue box, use Sample1 for the target variable. In the function group select Random Numbers. Then select the function Rv.Normal. Use 30 for the mean and 4 for the standard deviation. Continue until you have 10 samples. To sample

from other distributions, use appropriate functions in the Compute dialogue box.

The SPSS display shows 90% confidence intervals for 10 different random samples of size 30 taken from a normal distribution with $\mu = 30$ and $\sigma = 4$. Notice that, as expected, 9 of the 10 intervals contain the population mean $\mu = 30$.

SPSS Display

90% t-confidence intervals for random samples of size n = 30 from a normal distribution with $\mu = 30$ and $\sigma = 4$.						
	t	df	Sig(2-tail)	Mean	Lower	Upper
SAMPLE1	42.304	29	.000	29.7149	28.5214	30.9084
SAMPLE2	43.374	29	.000	30.1552	28.9739	31.3365
SAMPLE3	53.606	29	.000	31.2743	30.2830	32.2656
SAMPLE4	35.648	29	.000	30.1490	28.7120	31.5860
SAMPLE5	47.964	29	.000	31.0161	29.9173	32.1148
SAMPLE6	34.718	29	.000	30.3519	28.8665	31.8374
SAMPLE7	34.698	29	.000	30.7665	29.2599	32.2731
SAMPLE8	39.731	29	.000	30.2388	28.9456	31.5320
SAMPLE9	44.206	29	.000	29.7256	28.5831	30.8681
SAMPLE10	49.981	29	.000	29.7273	28.7167	30.7379

Application 3

Bootstrap Demonstration

Bootstrap can be used to construct confidence intervals for μ when traditional methods cannot be used. For example, if the sample size is small and the sample shows extreme outliers or extreme lack of symmetry, use of the Student's t distribution is inappropriate. Bootstrap makes no assumptions about the population.

Consider the following random sample of size 20:

12 15 21 2 6 3 15 51 22 18
37 12 25 19 33 15 14 17 12 27

A stem-and-leaf display shows that the data are skewed with one outlier.

0		2	represents 2
0		236	
1		2224555789	
2		1257	
3		37	
4			
5		1	

We can use Minitab to model the bootstrap method for constructing confidence intervals for μ . (The Professional edition of Minitab is required because of spreadsheet size and other limitations of the Student edition.) This demonstration uses only 1000 samples. Bootstrap uses many thousands.

Step 1: Create 1000 new samples, each of size 20, by sampling *with replacement* from the original data. To do this in Minitab, we enter the original 20 data values in column C1. Then, in column C2, we place equal probabilities of 0.05 beside each of the original data values. Use the menu choices **Calc** ► **Random Data** ► **Discrete**. In the dialogue box, fill in 1000 as the number of rows, store the data in columns C11–C30, and use column C1 for values and column C2 for probabilities.

Step 2: Find the sample mean of each of the 1000 samples. To do this in Minitab, use the menu choices **Calc** ► **Row Statistics**. In the dialogue box, select **mean**. Use columns C11–C30 as the input variables and store the results in column C31.

Step 3: Order the 1000 means from smallest to largest. In Minitab, use the menu choices **Manip** ► **Sort**. In the dialogue box, indicate C31 as the column to be sorted. Store the results in column C32. Sort by values in column C31.

Step 4: Create a 95% confidence interval by finding the boundaries for the middle 95% of the data. In other words, you need to find the values of the 2.5 percentile ($P_{2.5}$) and the 97.5 percentile ($P_{97.5}$).

Since there are 1000 data values, the 2.5 percentile is the data value in position 25, while the 97.5 percentile is the data value in position 975. The confidence interval is $P_{2.5} < \mu < P_{97.5}$.

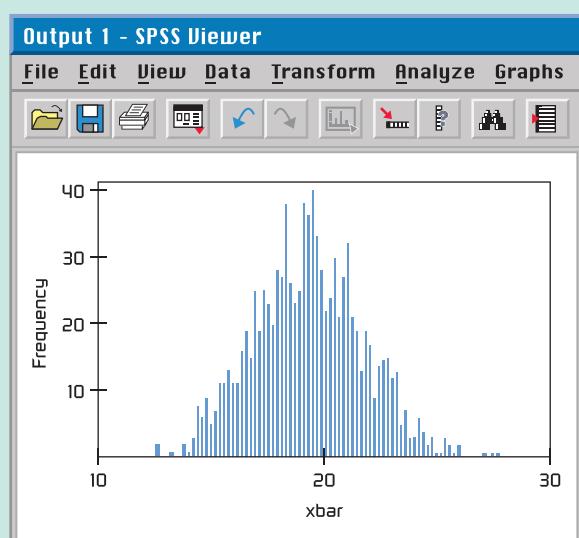
Demonstration Results

Figure 7-10 shows a histogram of the 1000 \bar{x} values from one bootstrap simulation. Three bootstrap simulations produced the following 95% confidence intervals.

- 13.90 to 23.90
- 14.00 to 24.15
- 14.05 to 23.8

Using the t distribution on the sample data, Minitab produced the interval 13.33 to 24.27. The results of the bootstrap simulations and the t distribution method are quite close.

FIGURE 7-10 Bootstrap Simulation, \bar{x} Distribution





8

- 8.1 Introduction to Statistical Tests
- 8.2 Testing the Mean μ
- 8.3 Testing a Proportion p
- 8.4 Tests Involving Paired Differences
(Dependent Samples)
- 8.5 Testing $\mu_1 - \mu_2$ and $p_1 - p_2$
(Independent Samples)



Sam Abell/National Geographic/Getty Images

"Would you tell me,
please, which way I
ought to go from
here?"

"That depends a good
deal on where you want to get to," said
the Cat.

"I don't much care where—" said Alice.

"Then it doesn't matter which way you
go," said the Cat.

—LEWIS CARROLL
ALICE'S ADVENTURES IN WONDERLAND



Mary Evans Picture Library/Arthur Rackham/
The Image Works

Charles Lutwidge Dodgson (1832–1898) was an English mathematician who loved to write children's stories in his free time. The dialogue between Alice and the Cheshire Cat occurs in the masterpiece *Alice's Adventures in Wonderland*, written by Dodgson under the pen name Lewis Carroll. These lines relate to our study of hypothesis testing. Statistical tests cannot answer all of life's questions. They cannot always tell us "where to go," but after this decision is made on other grounds, they can help us find the best way to get there.

For online student resources, visit the Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>.

HYPOTHESIS TESTING

PREVIEW QUESTIONS

Many of life's questions require a yes or no answer. When you must act on incomplete (sample) information, how do you decide whether to accept or reject a proposal? (SECTION 8.1)

What is the P-value of a statistical test? What does this measurement have to do with performance reliability? (SECTION 8.1)

How do you construct statistical tests for μ ? Does it make a difference whether σ is known or unknown? (SECTION 8.2)

How do you construct statistical tests for the proportion p of successes in a binomial experiment? (SECTION 8.3)

What are the advantages of pairing data values? How do you construct statistical tests for paired differences? (SECTION 8.4)

How do you construct statistical tests for differences of independent random variables? (SECTION 8.5)



Comstock Images/Jupiter Images

FOCUS PROBLEM

Benford's Law: The Importance of Being Number 1

Benford's Law states that in a wide variety of circumstances, numbers have "1" as their first nonzero digit disproportionately often. Benford's Law applies to such diverse topics as the drainage areas of rivers; properties of chemicals; populations of towns; figures in newspapers, magazines, and government reports; and the half-lives of radioactive atoms!

Specifically, such diverse measurements begin with "1" about 30% of the time, with "2" about 18% of time, and with "3" about 12.5% of the time. Larger digits occur less often. For example, less than 5% of the numbers in circumstances such as these begin with the digit 9. This is in dramatic contrast to a random sampling situation, in which each of the digits 1 through 9 has an equal chance of appearing.

The first nonzero digits of numbers taken from large bodies of numerical records such as tax returns, population studies, government records, and so forth, show the probabilities of occurrence as displayed in the table on the next page.



Corbis

First nonzero digit	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

More than 100 years ago, the astronomer Simon Newcomb noticed that books of logarithm tables were much dirtier near the fronts of the tables. It seemed that people were more frequently looking up numbers with a low first digit. This was regarded as an odd phenomenon and a strange curiosity. The phenomenon was rediscovered in 1938 by physicist Frank Benford (hence the name *Benford's Law*).

More recently, Ted Hill, a mathematician at the Georgia Institute of Technology, studied situations that might demonstrate Benford's Law. Professor Hill showed that such probability distributions are likely to occur when we have a "distribution of distributions." Put another way, large random collections of random samples tend to follow Benford's Law. This seems to be especially true for samples taken from large government data banks, accounting reports for large corporations, large collections of astronomical observations, and so forth. For more information, see *American Scientist*, Vol. 86, pp. 358–363, and *Chance*, American Statistical Association, Vol. 12, No. 3, pp. 27–31.

Can Benford's Law be applied to help solve a real-world problem? Well, one application might be accounting fraud! Suppose the first nonzero digits of the entries in the accounting records of a large corporation (such as Enron or WorldCom) do not follow Benford's Law. Should this set off an accounting alarm for the FBI or the stockholders? How "significant" would this be? Such questions are the subject of statistics.

In Section 8.3, you will see how to use sample data to test whether the proportion of first nonzero digits of the entries in a large accounting report follows Benford's Law. Problems 7 and 8 of Section 8.3 relate to Benford's Law and accounting discrepancies. In one problem, you are asked to use sample data to determine if accounting books have been "cooked" by "pumping numbers up" to make the company look more attractive or perhaps to provide a cover for money laundering. In the other problem, you are asked to determine if accounting books have been "cooked" by artificially lowered numbers, perhaps to hide profits from the Internal Revenue Service or to divert company profits to unscrupulous employees. (See Problems 7 and 8 of Section 8.3.)

SECTION 8.1

Introduction to Statistical Tests

FOCUS POINTS

- Understand the rationale for statistical tests.
- Identify the null and alternate hypotheses in a statistical test.
- Identify right-tailed, left-tailed, and two-tailed tests.
- Use a test statistic to compute a P -value.
- Recognize types of errors, level of significance, and power of a test.
- Understand the meaning and risks of rejecting or not rejecting the null hypothesis.

In Chapter 1, we emphasized the fact that one of a statistician's most important jobs is to draw inferences about populations based on samples taken from the populations. Most statistical inference centers around the parameters of a population (often the mean or probability of success in a binomial trial). Methods for drawing inferences about parameters are of two types: Either we make decisions concerning the value of the parameter, or we actually estimate the value of the

Hypothesis testing

Hypothesis

Null hypothesis H_0

Alternate hypothesis H_1

parameter. When we estimate the value (or location) of a parameter, we are using methods of estimation such as those studied in Chapter 7. Decisions concerning the value of a parameter are obtained by *hypothesis testing*, the topic we shall study in this chapter.

Students often ask which method should be used on a particular problem—that is, should the parameter be estimated, or should we test a *hypothesis* involving the parameter? The answer lies in the practical nature of the problem and the questions posed about it. Some people prefer to test theories concerning the parameters. Others prefer to express their inferences as estimates. Both estimation and hypothesis testing are found extensively in the literature of statistical applications.

Stating Hypotheses

Our first step is to establish a working hypothesis about the population parameter in question. This hypothesis is called the *null hypothesis*, denoted by the symbol H_0 . The value specified in the null hypothesis is often a historical value, a claim, or a production specification. For instance, if the average height of a professional male basketball player was 6.5 feet 10 years ago, we might use a null hypothesis $H_0: \mu = 6.5$ feet for a study involving the average height of this year's professional male basketball players. If television networks claim that the average length of time devoted to commercials in a 60-minute program is 12 minutes, we would use $H_0: \mu = 12$ minutes as our null hypothesis in a study regarding the average length of time devoted to commercials. Finally, if a repair shop claims that it should take an average of 25 minutes to install a new muffler on a passenger automobile, we would use $H_0: \mu = 25$ minutes as the null hypothesis for a study of how well the repair shop is conforming to specified average times for a muffler installation.

Any hypothesis that differs from the null hypothesis is called an *alternate hypothesis*. An alternate hypothesis is constructed in such a way that it is the hypothesis to be accepted when the null hypothesis must be rejected. The alternate hypothesis is denoted by the symbol H_1 . For instance, if we believe the average height of professional male basketball players is taller than it was 10 years ago, we would use an alternate hypothesis $H_1: \mu > 6.5$ feet with the null hypothesis $H_0: \mu = 6.5$ feet.

Null hypothesis H_0 : This is the statement that is under investigation or being tested. Usually the null hypothesis represents a statement of “no effect,” “no difference,” or, put another way, “things haven’t changed.”

Alternate hypothesis H_1 : This is the statement you will adopt in the situation in which the evidence (data) is so strong that you reject H_0 . A statistical test is designed to assess the strength of the evidence (data) against the null hypothesis.

EXAMPLE 1

NULL AND ALTERNATE HYPOTHESES

A car manufacturer advertises that its new subcompact models get 47 miles per gallon (mpg). Let μ be the mean of the mileage distribution for these cars. You assume that the manufacturer will not underrate the car, but you suspect that the mileage might be overrated.

(a) What shall we use for H_0 ?

SOLUTION: We want to see if the manufacturer’s claim that $\mu = 47$ mpg can be rejected. Therefore, our null hypothesis is simply that $\mu = 47$ mpg. We denote the null hypothesis as

$$H_0: \mu = 47 \text{ mpg}$$

(b) What shall we use for H_1 ?

SOLUTION: From experience with this manufacturer, we have every reason to believe that the advertised mileage is too high. If μ is not 47 mpg, we are sure it is less than 47 mpg. Therefore, the alternate hypothesis is

$$H_1: \mu < 47 \text{ mpg}$$



GUIDED EXERCISE 1

Null and alternate hypotheses

A company manufactures ball bearings for precision machines. The average diameter of a certain type of ball bearing should be 6.0 mm. To check that the average diameter is correct, the company formulates a statistical test.

(a) What should be used for H_0 ? (*Hint:* What is the company trying to test?)



If μ is the mean diameter of the ball bearings, the company wants to test whether $\mu = 6.0$ mm. Therefore, $H_0: \mu = 6.0$ mm.

(b) What should be used for H_1 ? (*Hint:* An error either way, too small or too large, would be serious.)



An error either way could occur, and it would be serious. Therefore, $H_1: \mu \neq 6.0$ mm (μ is either smaller than or larger than 6.0 mm).

COMMENT: NOTATION REGARDING THE NULL HYPOTHESIS In statistical testing, the null hypothesis H_0 always contains the equals symbol. However, in the null hypothesis, some statistical software packages and texts also include the inequality symbol that is opposite that shown in the alternate hypothesis. For instance, if the alternate hypothesis is “ μ is less than 3” ($\mu < 3$), then the corresponding null hypothesis is sometimes written as “ μ is greater than or equal to 3” ($\mu \geq 3$). The mathematical construction of a statistical test uses the null hypothesis to assign a specific number (rather than a range of numbers) to the parameter μ in question. The null hypothesis establishes a single fixed value for μ , so we are working with a single distribution having a specific mean. In this case, H_0 assigns $\mu = 3$. So, when $H_1: \mu < 3$ is the alternate hypothesis, we follow the commonly used convention of writing the null hypothesis simply as $H_0: \mu = 3$.

Types of Tests

The null hypothesis H_0 always states that the parameter of interest *equals* a specified value. The alternate hypothesis H_1 states that the parameter is *less than*, *greater than*, or simply *not equal to* the same value. We categorize a statistical test as *left-tailed*, *right-tailed*, or *two-tailed* according to the alternate hypothesis.

Types of statistical tests

A statistical test is:

left-tailed if H_1 states that the parameter is less than the value claimed in H_0

right-tailed if H_1 states that the parameter is greater than the value claimed in H_0

two-tailed if H_1 states that the parameter is different from (or not equal to) the value claimed in H_0

Left-tailed test

Right-tailed test

Two-tailed test

TABLE 8-1 The Null and Alternate Hypotheses for Tests of the Mean μ

Null Hypothesis	Alternate Hypotheses and Type of Test		
Claim about μ or historical value of μ $H_0: \mu = k$	You believe that μ is less than value stated in H_0 . $H_1: \mu < k$ Left-tailed test	You believe that μ is more than value stated in H_0 . $H_1: \mu > k$ Right-tailed test	You believe that μ is different from value stated in H_0 . $H_1: \mu \neq k$ Two-tailed test

In this introduction to statistical tests, we discuss tests involving a population mean μ . However, you should keep an open mind and be aware that the methods outlined apply to testing other parameters as well (e.g., p , σ , $\mu_1 - \mu_2$, $p_1 - p_2$, and so on). Table 8-1 shows how tests of the mean μ are categorized.

Hypothesis Tests of μ , Given x Is Normal and σ Is Known

Once you have selected the null and alternate hypotheses, how do you decide which hypothesis is likely to be valid? Data from a simple random sample and the sample test statistic, together with the corresponding sampling distribution of the test statistic, will help you decide. Example 2 leads you through the decision process.

First, a quick review of Section 6.4 is in order. Recall that a population *parameter* is a numerical descriptive measurement of the entire population. Examples of population parameters are μ , p , and σ . It is important to remember that for a given population, the parameters are *fixed* values. They do not vary! The null hypothesis H_0 makes a statement about a population parameter.

A *statistic* is a numerical descriptive measurement of a sample. Examples of statistics are \bar{x} , \hat{p} , and s . Statistics usually *vary* from one sample to the next. The probability distribution of the statistic we are using is called a *sampling distribution*.

For hypothesis testing, we take a *simple* random sample and compute a *sample test statistic* corresponding to the parameter in H_0 . Based on the sampling distribution of the statistic, we can assess how compatible the sample test statistic is with H_0 .

In this section, we use hypothesis tests about the mean to introduce the concepts and vocabulary of hypothesis testing. In particular, let's suppose that x has a *normal distribution* with mean μ and standard deviation σ . Then, Theorem 6.1 tells us that \bar{x} has a *normal distribution* with mean μ and standard deviation σ/\sqrt{n} .

Sample test statistic for μ , given x normal and σ known

PROCEDURE

Requirements The x distribution is *normal* with known standard deviation σ .

Then \bar{x} has a normal distribution. The standardized test statistic is

$$\text{test statistic} = z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where \bar{x} = mean of a simple random sample

μ = value stated in H_0 .

n = sample size

EXAMPLE 2

STATISTICAL TESTING PREVIEW

Rosie is an aging sheep dog in Montana who gets regular checkups from her owner, the local veterinarian. Let x be a random variable that represents Rosie's resting heart rate (in beats per minute). From past experience, the vet knows that



x has a normal distribution with $\sigma = 12$. The vet checked the *Merck Veterinary Manual* and found that for dogs of this breed, $\mu = 115$ beats per minute.

Over the past six weeks, Rosie's heart rate (beats/min) measured

93 109 110 89 112 117

The sample mean is $\bar{x} = 105.0$. The vet is concerned that Rosie's heart rate may be slowing. Do the data indicate that this is the case?

SOLUTION:

- (a) Establish the null and alternate hypotheses.

If "nothing has changed" from Rosie's earlier life, then her heart rate should be nearly average. This point of view is represented by the null hypothesis

$$H_0: \mu = 115$$

However, the vet is concerned about Rosie's heart rate slowing. This point of view is represented by the alternate hypothesis

$$H_1: \mu < 115$$

- (b) Are the observed sample data compatible with the null hypothesis?

Are the six observations of Rosie's heart rate compatible with the null hypothesis $H_0: \mu = 115$? To answer this question, we need to know the *probability* of obtaining a sample mean of 105.0 or less from a population with true mean $\mu = 115$. If this probability is small, we conclude that $H_0: \mu = 115$ is not the case. Rather, $H_1: \mu < 115$ and Rosie's heart rate is slowing.

- (c) How do we compute the probability in part (b)?

Well, you probably guessed it! We use the sampling distribution for \bar{x} and compute $P(\bar{x} < 105.0)$. Figure 8-1 shows the \bar{x} distribution and the corresponding standard normal distribution with the desired probability shaded.

Check Requirements Since x has a normal distribution, \bar{x} will also have a normal distribution for any sample size n and given σ (see Theorem 6.1).

Note that using $\mu = 115$ from H_0 , $\sigma = 12$, and $n = 6$ the sample $\bar{x} = 105.0$ converts to

$$\text{test statistic} = z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{105.0 - 115}{12/\sqrt{6}} \approx -2.04$$

Using the standard normal distribution table, we find that

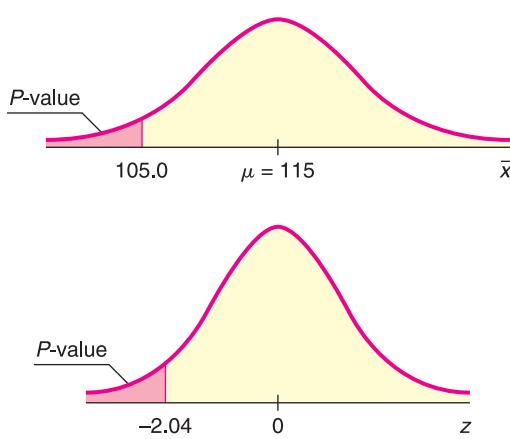
$$P(\bar{x} < 105.0) = P(z < -2.04) = 0.0207$$

P-value

The area in the left tail that is more extreme than $\bar{x} = 105.0$ is called the *P-value* of the test. In this example, $P\text{-value} = 0.0207$. We will learn more about *P*-values later.

FIGURE 8-1

Sampling Distribution for \bar{x} and Corresponding z Distribution



- (d) **Interpretation** What conclusion can be drawn about Rosie's average heart rate?

If $H_0: \mu = 115$ is in fact true, the probability of getting a sample mean of $\bar{x} \leq 105.0$ is only about 2%. Because this probability is small, we reject $H_0: \mu = 115$ and conclude that $H_1: \mu < 115$. Rosie's average heart rate seems to be slowing.

- (e) Have we proved $H_0: \mu = 115$ to be false and $H_1: \mu < 115$ to be true?

No! The sample data do not *prove* H_0 to be false and H_1 to be true! We do say that H_0 has been "discredited" by a small *P*-value of 0.0207. Therefore, we abandon the claim $H_0: \mu = 115$ and adopt the claim $H_1: \mu < 115$.



The *P*-value of a Statistical Test

Rosie the sheep dog has helped us to "sniff out" an important statistical concept.

P-value

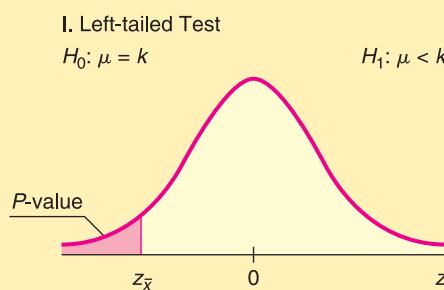
Assuming H_0 is true, the *probability* that the test statistic will take on values as extreme as or more extreme than the observed test statistic (computed from sample data) is called the ***P*-value** of the test. The smaller the *P*-value computed from sample data, the stronger the evidence against H_0 .

The *P*-value, sometimes called the *probability of chance*, can be thought of as the probability that the results of a statistical experiment are due only to chance. The lower the *P*-value, the greater the likelihood of obtaining the same (or very similar) results in a repetition of the statistical experiment. Thus, a low *P*-value is a good indication that your results are not due to random chance alone.

The *P*-value associated with the observed test statistic takes on different values depending on the alternate hypothesis and the type of test. Let's look at *P*-values and types of tests when the test involves the mean and standard normal distribution. Notice that in Example 2, part (c), we computed a *P*-value for a left-tailed test. Guided Exercise 3 asks you to compute a *P*-value for a two-tailed test.

P-values and types of tests

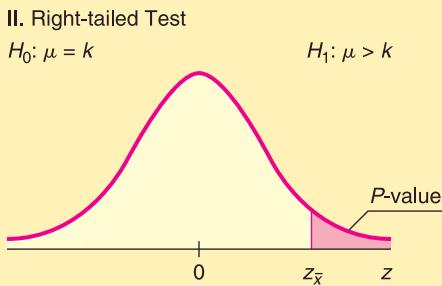
Let $z_{\bar{x}}$ represent the standardized sample test statistic for testing a mean μ using the standard normal distribution. That is, $z_{\bar{x}} = (\bar{x} - \mu)/(\sigma/\sqrt{n})$.



$$P\text{-value} = P(z < z_{\bar{x}})$$

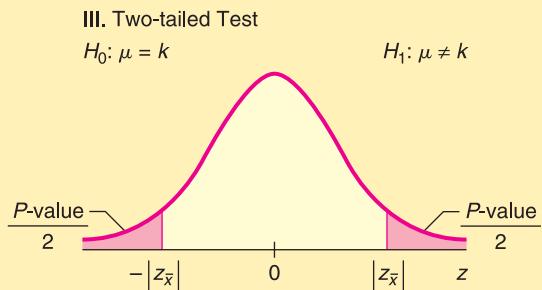
This is the probability of getting a test statistic as low as or lower than $z_{\bar{x}}$.

Continued



$$P\text{-value} = P(z > z_{\bar{x}})$$

This is the probability of getting a test statistic as high or higher than $z_{\bar{x}}$



$$\frac{P\text{-value}}{2} = P(z > |z_{\bar{x}}|); \text{ therefore,}$$

$$P\text{-value} = 2P(z > |z_{\bar{x}}|)$$

This is the probability of getting a test statistic either lower than $-|z_{\bar{x}}|$ or higher than $|z_{\bar{x}}|$.

Types of Errors

Type I error
Type II error

If we *reject the null hypothesis when it is, in fact, true*, we have made an error that is called a *type I error*. On the other hand, if we *accept the null hypothesis when it is, in fact, false*, we have made an error that is called a *type II error*. Table 8-2 indicates how these errors occur.

For tests of hypotheses to be well constructed, they must be designed to minimize possible errors of decision. (Usually, we do not know if an error has been made, and therefore, we can talk only about the probability of making an error.) Usually, for a given sample size, an attempt to reduce the probability of one type of error results in an increase in the probability of the other type of error. In practical applications, one type of error may be more serious than another. In such a case, careful attention is given to the more serious error. If we increase the sample size, it is possible to reduce both types of errors, but increasing the sample size may not be possible.

Good statistical practice requires that we announce in advance how much evidence against H_0 will be required to reject H_0 . The probability with which we are willing to risk a type I error is called the *level of significance* of a test. The level of significance is denoted by the Greek letter α (pronounced “alpha”).

Level of significance α

The level of significance α is the probability of rejecting H_0 when it is true.
This is the probability of a type I error.

TABLE 8-2 Type I and Type II Errors

Truth of H_0	Our Decision	
	And if we do not reject H_0	And if we reject H_0
If H_0 is true	Correct decision; no error	Type I error
If H_0 is false	Type II error	Correct decision; no error

TABLE 8-3

Probabilities Associated with a Statistical Test

Truth of H_0	Our Decision	
	And if we accept H_0 as true	And if we reject H_0 as false
If H_0 is true	Correct decision, with corresponding probability $1 - \alpha$	Type I error, with corresponding probability α , called the <i>level of significance of the test</i>
If H_0 is false	Type II error, with corresponding probability β	Correct decision; with corresponding probability $1 - \beta$, called the <i>power of the test</i>

Probability of a type II error β

The *probability of making a type II error* is denoted by the Greek letter β (pronounced “beta”).

Methods of hypothesis testing require us to choose α and β values to be as small as possible. In elementary statistical applications, we usually choose α first.

Power of a test ($1 - \beta$)

The quantity $1 - \beta$ is called the *power of a test* and represents the probability of rejecting H_0 when it is, in fact, false.

For a given level of significance, how much power can we expect from a test? The actual value of the power is usually difficult (and sometimes impossible) to obtain, since it requires us to know the H_1 distribution. However, we can make the following general comments:

1. The power of a statistical test increases as the level of significance α increases. A test performed at the $\alpha = 0.05$ level has more power than one performed at $\alpha = 0.01$. This means that the less stringent we make our significance level α , the more likely we will be to reject the null hypothesis when it is false.
2. Using a larger value of α will increase the power, but it also will increase the probability of a type I error. Despite this fact, most business executives, administrators, social scientists, and scientists use *small* α values. This choice reflects the conservative nature of administrators and scientists, who are usually more willing to make an error by failing to reject a claim (i.e., H_0) than to make an error by accepting another claim (i.e., H_1) that is false. Table 8-3 summarizes the probabilities of errors associated with a statistical test.

COMMENT Since the calculation of the probability of a type II error is treated in advanced statistics courses, we will restrict our attention to the probability of a type I error.

GUIDED EXERCISE 2**Types of errors**

Let's reconsider Guided Exercise 1, in which we were considering the manufacturing specifications for the diameter of ball bearings. The hypotheses were

$$H_0: \mu = 6.0 \text{ mm} \text{ (manufacturer's specification)}$$

$$H_1: \mu \neq 6.0 \text{ mm} \text{ (cause for adjusting process)}$$

- (a) Suppose the manufacturer requires a 1% level of significance. Describe a type I error, its consequence, and its probability.



A type I error is caused when sample evidence indicates that we should reject H_0 when, in fact, the average diameter of the ball bearings being produced is 6.0 mm. A type I error will cause a needless

Continued

GUIDED EXERCISE 2 *continued*

(b) Discuss a type II error and its consequences.



adjustment and delay of the manufacturing process. The probability of such an error is 1% because $\alpha = 0.01$.

A type II error occurs if the sample evidence tells us not to reject the null hypothesis $H_0: \mu = 6.0$ mm when, in fact, the average diameter of the ball bearing is either too large or too small to meet specifications. Such an error would mean that the production process would not be adjusted even though it really needed to be adjusted. This could possibly result in a large production of ball bearings that do not meet specifications.

Concluding a Statistical Test

Usually, α is specified in advance before any samples are drawn so that results will not influence the choice for the level of significance. To conclude a statistical test, we compare our α value with the P -value computed using sample data and the sampling distribution.

PROCEDURE

Statistical significance

HOW TO CONCLUDE A TEST USING THE P -VALUE AND LEVEL OF SIGNIFICANCE α

If P -value $\leq \alpha$, we **reject** the null hypothesis and say the data are **statistically significant** at the level α .

If P -value $> \alpha$, we **do not reject** the null hypothesis.

In what sense are we using the word *significant*? *Webster's Dictionary* gives two interpretations of *significance*: (1) having or signifying *meaning*; or (2) being important or momentous.

In statistical work, significance does not necessarily imply momentous importance. For us, “significant” at the α level has a special meaning. It says that at the α level of risk, the evidence (sample data) against the null hypothesis H_0 is sufficient to discredit H_0 , so we adopt the alternate hypothesis H_1 .

In any case, we do not claim that we have “proved” or “disproved” the null hypothesis H_0 . We can say that the probability of a type I error (rejecting H_0 when it is, in fact, true) is α .

Basic components of a statistical test

A statistical test can be thought of as a package of five basic ingredients.

1. Null hypothesis H_0 , alternate hypothesis H_1 , and preset level of significance α

If the evidence (sample data) against H_0 is strong enough, we reject H_0 and adopt H_1 . The level of significance α is the probability of rejecting H_0 when it is, in fact, true.

2. Test statistic and sampling distribution

These are mathematical tools used to measure compatibility of sample data and the null hypothesis.

3. P-value

This is the probability of obtaining a test statistic from the sampling distribution that is as extreme as, or more extreme (as specified by H_1) than, the sample test statistic computed from the data under the assumption that H_0 is true.

4. Test conclusion

If $P\text{-value} \leq \alpha$, we reject H_0 and say that the data are significant at level α . If $P\text{-value} > \alpha$, we do not reject H_0 .

5. Interpretation of the test results

Give a simple explanation of your conclusions in the context of the application.

GUIDED EXERCISE 3**Constructing a statistical test for μ (normal distribution)**

The Environmental Protection Agency has been studying Miller Creek regarding ammonia nitrogen concentration. For many years, the concentration has been 2.3 mg/l. However, a new golf course and new housing developments are raising concern that the concentration may have changed because of lawn fertilizer. Any change (either an increase or a decrease) in the ammonia nitrogen concentration can affect plant and animal life in and around the creek (Reference: *EPA Report 832-R-93-005*). Let x be a random variable representing ammonia nitrogen concentration (in mg/l). Based on recent studies of Miller Creek, we may assume that x has a normal distribution with $\sigma = 0.30$. Recently, a random sample of eight water tests from the creek gave the following x values.

2.1 2.5 2.2 2.8 3.0 2.2 2.4 2.9

The sample mean is $\bar{x} \approx 2.51$.

Let us construct a statistical test to examine the claim that the concentration of ammonia nitrogen has changed from 2.3 mg/l. Use level of significance $\alpha = 0.01$.

- (a) What is the null hypothesis? What is the alternate hypothesis? What is the level of significance α ?

$$\Rightarrow H_0: \mu = 2.3$$

$$H_1: \mu \neq 2.3$$

$$\alpha = 0.01$$

- (b) Is this a right-tailed, left-tailed, or two-tailed test?

\Rightarrow Since $H_1: \mu \neq 2.3$, this is a two-tailed test.

- (c) **Check Requirements** What sampling distribution shall we use? Note that the value of μ is given in the null hypothesis, H_0 .

\Rightarrow Since the x distribution is normal and σ is known, we use the standard normal distribution with

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - 2.3}{\frac{0.3}{\sqrt{8}}}$$

- (d) What is the sample test statistic? Convert the sample mean \bar{x} to a standard z value.

\Rightarrow The sample of eight measurements has mean $\bar{x} = 2.51$. Converting this measurement to z , we have

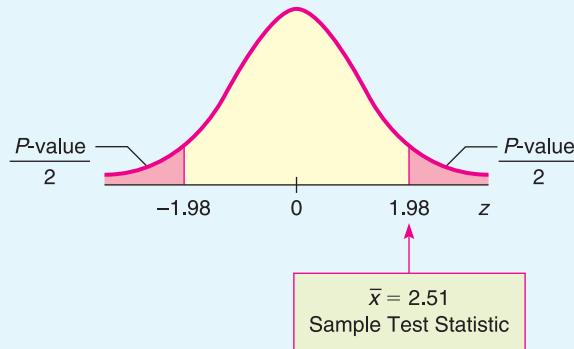
$$\text{test statistic} = z = \frac{2.51 - 2.3}{\frac{0.3}{\sqrt{8}}} \approx 1.98$$

Continued

GUIDED EXERCISE 3 *continued*

- (e) Draw a sketch showing the P -value area on the standard normal distribution. Find the P -value.

$$\Rightarrow P\text{-value} = 2P(z > 1.98) = 2(0.0239) = 0.0478$$

FIGURE 8-2 P -value

- (f) Compare the level of significance α and the P -value. What is your conclusion?
 (g) *Interpret* your results in the context of this problem.

- \Rightarrow Since P -value $0.0478 \geq 0.01$, we see that P -value $> \alpha$. We fail to reject H_0 .
- \Rightarrow The sample data are not significant at the $\alpha = 1\%$ level. At this point in time, there is not enough evidence to conclude that the ammonia nitrogen concentration has changed in Miller Creek.

Interpretation of level of significance

In most statistical applications, the level of significance is specified to be $\alpha = 0.05$ or $\alpha = 0.01$, although other values can be used. If $\alpha = 0.05$, then we say we are using a 5% level of significance. This means that in 100 similar situations, H_0 will be rejected 5 times, on average, when it should not have been rejected.

Meaning of accepting H_0

Using Technology at the end of this chapter shows a simulation of this phenomenon.

When we accept (or fail to reject) the null hypothesis, we should understand that we are *not proving the null hypothesis*. We are saying only that the sample evidence (data) is not strong enough to justify rejection of the null hypothesis. The word *accept* sometimes has a stronger meaning in common English usage than we are willing to give it in our application of statistics. Therefore, we often use the expression *fail to reject H_0* instead of *accept H_0* . “*Fail to reject the null hypothesis*” simply means that the evidence in favor of rejection was not strong enough (see Table 8-4). Often, in the case that H_0 cannot be rejected, a confidence interval is used to estimate the parameter in question. The confidence interval gives the statistician a range of possible values for the parameter.

TABLE 8-4 Meaning of the Terms *Fail to Reject H_0* and *Reject H_0*

Term	Meaning
<i>Fail to reject H_0</i>	Fail to reject H_0 There is not enough evidence in the data (and the test being used) to justify a rejection of H_0 . This means that we retain H_0 with the understanding that we have not proved it to be true beyond all doubt.
<i>Reject H_0</i>	Reject H_0 There is enough evidence in the data (and the test employed) to justify rejection of H_0 . This means that we choose the alternate hypothesis H_1 with the understanding that we have not proved H_1 to be true beyond all doubt.

COMMENT Some comments about P -values and level of significance α should be made. The level of significance α should be a fixed, prespecified value. Usually, α is chosen before any samples are drawn. The level of significance α is the probability of a type I error. So, α is the probability of rejecting H_0 when, in fact, H_0 is true.

Interpreting the P -value of a test statistic

The P -value should *not* be interpreted as the probability of a type I error. The level of significance (in theory) is set in advance before any samples are drawn. The P -value cannot be set in advance, since it is determined from the random sample. The P -value, together with α , should be regarded as tools used to conclude the test. If P -value $\leq \alpha$, then reject H_0 , and if P -value $> \alpha$, then do not reject H_0 .

In most computer applications and journal articles, only the P -value is given. It is understood that the person using this information will supply an appropriate level of significance α . From an historical point of view, the English statistician F. Y. Edgeworth (1845–1926) was one of the first to use the term *significant* to imply that the sample data indicate a “meaningful” difference from a previously held view.

In this book, we are using the most popular method of testing, which is called the *P-value method*. At the end of the next section, you will learn about another (equivalent) method of testing called the *critical region method*. An extensive discussion regarding the P -value method of testing versus the critical region method can be found in *The American Statistician*, Vol. 57, No. 3, pp. 171–178, American Statistical Association.

VIEWPOINT

Lovers, Take Heed!!!

If you are going to whisper sweet nothings to your sweetheart, be sure to whisper them in the left ear. Professor Sim of Sam Houston State University (Huntsville, Texas) found that emotionally loaded words have a higher recall rate when spoken into a person's left ear, not the right. Professor Sim presented his findings at the British Psychology Society European Congress. He told the Congress that his findings are consistent with the hypothesis that the brain's right hemisphere has more influence in the processing of emotional stimuli. (The left ear is controlled by the right side of the brain.) Sim's research involved statistical tests like the ones you will study in this chapter.

SECTION 8.1 PROBLEMS

1. **Statistical Literacy** Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you answer the following questions.
 - (a) What is a null hypothesis H_0 ?
 - (b) What is an alternate hypothesis H_1 ?
 - (c) What is a type I error? a type II error?
 - (d) What is the level of significance of a test? What is the probability of a type II error?
2. **Statistical Literacy** In a statistical test, we have a choice of a left-tailed test, a right-tailed test, or a two-tailed test. Is it the null hypothesis or the alternate hypothesis that determines which type of test is used? Explain your answer.
3. **Statistical Literacy** If we fail to reject (i.e., “accept”) the null hypothesis, does this mean that we have *proved* it to be true beyond *all* doubt? Explain your answer.
4. **Statistical Literacy** If we reject the null hypothesis, does this mean that we have *proved* it to be false beyond *all* doubt? Explain your answer.

5. **Statistical Literacy** What terminology do we use for the probability of rejecting the null hypothesis when it is true? What symbol do we use for this probability? Is this the probability of a type I or a type II error?
6. **Statistical Literacy** What terminology do we use for the probability of rejecting the null hypothesis when it is, in fact, false?
7. **Statistical Literacy** If the P -value in a statistical test is greater than the level of significance for the test, do we reject or fail to reject H_0 ?
8. **Statistical Literacy** If the P -value in a statistical test is less than or equal to the level of significance for the test, do we reject or fail to reject H_0 ?
9. **Statistical Literacy** Suppose the P -value in a right-tailed test is 0.0092. Based on the same population, sample, and null hypothesis, what is the P -value for a corresponding two-tailed test?
10. **Statistical Literacy** Suppose the P -value in a two-tailed test is 0.0134. Based on the same population, sample, and null hypothesis, and assuming the test statistic z is negative, what is the P -value for a corresponding left-tailed test?
11. **Basic Computation: Setting Hypotheses** Suppose you want to test the claim that a population mean equals 40.
 - (a) State the null hypothesis.
 - (b) State the alternate hypothesis if you have no information regarding how the population mean might differ from 40.
 - (c) State the alternate hypothesis if you believe (based on experience or past studies) that the population mean may exceed 40.
 - (d) State the alternate hypothesis if you believe (based on experience or past studies) that the population mean may be less than 40.
12. **Basic Computation: Setting Hypotheses** Suppose you want to test the claim that a population mean equals 30.
 - (a) State the null hypothesis.
 - (b) State the alternate hypothesis if you have no information regarding how the population mean might differ from 30.
 - (c) State the alternate hypothesis if you believe (based on experience or past studies) that the population mean may be greater than 30.
 - (d) State the alternate hypothesis if you believe (based on experience or past studies) that the population mean may not be as large as 30.
13. **Basic Computation: Find Test Statistic, Corresponding P -value, and Conclude Test** A random sample of size 20 from a normal distribution with $\sigma = 4$ produced a sample mean of 8.
 - (a) **Check Requirements** Is the \bar{x} distribution normal? Explain.
 - (b) Compute the sample test statistic z under the null hypothesis $H_0: \mu = 7$.
 - (c) For $H_1: \mu \neq 7$, estimate the P -value of the test statistic.
 - (d) For a level of significance of 0.05 and the hypotheses of parts (b) and (c), do you reject or fail to reject the null hypothesis? Explain.
14. **Basic Computation: Find the Test Statistic and Corresponding P -value** A random sample of size 16 from a normal distribution with $\sigma = 3$ produced a sample mean of 4.5.
 - (a) **Check Requirements** Is the \bar{x} distribution normal? Explain.
 - (b) Compute the sample test statistic z under the null hypothesis $H_0: \mu = 6.3$.
 - (c) For $H_1: \mu < 6.3$, estimate the P -value of the test statistic.
 - (d) For a level of significance of 0.01 and the hypotheses of parts (b) and (c), do you reject or fail to reject the null hypothesis? Explain.
15. **Veterinary Science: Colts** The body weight of a healthy 3-month-old colt should be about $\mu = 60$ kg (Source: *The Merck Veterinary Manual*, a standard reference manual used in most veterinary colleges).
 - (a) If you want to set up a statistical test to challenge the claim that $\mu = 60$ kg, what would you use for the null hypothesis H_0 ?

- (b) In Nevada, there are many herds of wild horses. Suppose you want to test the claim that the average weight of a wild Nevada colt (3 months old) is less than 60 kg. What would you use for the alternate hypothesis H_1 ?
- (c) Suppose you want to test the claim that the average weight of such a wild colt is greater than 60 kg. What would you use for the alternate hypothesis?
- (d) Suppose you want to test the claim that the average weight of such a wild colt is *different* from 60 kg. What would you use for the alternate hypothesis?
- (e) For each of the tests in parts (b), (c), and (d), would the area corresponding to the P -value be on the left, on the right, or on both sides of the mean? Explain your answer in each case.
16. **Marketing: Shopping Time** How much customers buy is a direct result of how much time they spend in a store. A study of average shopping times in a large national housewares store gave the following information (Source: *Why We Buy: The Science of Shopping* by P. Underhill):
- Women with female companion: 8.3 min.
- Women with male companion: 4.5 min.
- Suppose you want to set up a statistical test to challenge the claim that a woman with a female friend spends an average of 8.3 minutes shopping in such a store.
- (a) What would you use for the null and alternate hypotheses if you believe the average shopping time is less than 8.3 minutes? Is this a right-tailed, left-tailed, or two-tailed test?
- (b) What would you use for the null and alternate hypotheses if you believe the average shopping time is different from 8.3 minutes? Is this a right-tailed, left-tailed, or two-tailed test?
- Stores that sell mainly to women should figure out a way to engage the interest of men—perhaps comfortable seats and a big TV with sports programs! Suppose such an entertainment center was installed and you now wish to challenge the claim that a woman with a male friend spends only 4.5 minutes shopping in a housewares store.
- (c) What would you use for the null and alternate hypotheses if you believe the average shopping time is more than 4.5 minutes? Is this a right-tailed, left-tailed, or two-tailed test?
- (d) What would you use for the null and alternate hypotheses if you believe the average shopping time is different from 4.5 minutes? Is this a right-tailed, left-tailed, or two-tailed test?
17. **Meteorology: Storms** *Weatherwise* magazine is published in association with the American Meteorological Society. Volume 46, Number 6 has a rating system to classify Nor'easter storms that frequently hit New England states and can cause much damage near the ocean coast. A *severe* storm has an average peak wave height of 16.4 feet for waves hitting the shore. Suppose that a Nor'easter is in progress at the severe storm class rating.
- (a) Let us say that we want to set up a statistical test to see if the wave action (i.e., height) is dying down or getting worse. What would be the null hypothesis regarding average wave height?
- (b) If you wanted to test the hypothesis that the storm is getting worse, what would you use for the alternate hypothesis?
- (c) If you wanted to test the hypothesis that the waves are dying down, what would you use for the alternate hypothesis?
- (d) Suppose you do not know whether the storm is getting worse or dying out. You just want to test the hypothesis that the average wave height is *different* (either higher or lower) from the severe storm class rating. What would you use for the alternate hypothesis?
- (e) For each of the tests in parts (b), (c), and (d), would the area corresponding to the P -value be on the left, on the right, or on both sides of the mean? Explain your answer in each case.

18. **Chrysler Concorde: Acceleration** *Consumer Reports* stated that the mean time for a Chrysler Concorde to go from 0 to 60 miles per hour is 8.7 seconds.
- If you want to set up a statistical test to challenge the claim of 8.7 seconds, what would you use for the null hypothesis?
 - The town of Leadville, Colorado, has an elevation over 10,000 feet. Suppose you wanted to test the claim that the average time to accelerate from 0 to 60 miles per hour is longer in Leadville (because of less oxygen). What would you use for the alternate hypothesis?
 - Suppose you made an engine modification and you think the average time to accelerate from 0 to 60 miles per hour is reduced. What would you use for the alternate hypothesis?
 - For each of the tests in parts (b) and (c), would the P -value area be on the left, on the right, or on both sides of the mean? Explain your answer in each case.

For Problems 19–24, please provide the following information.

- What is the level of significance? State the null and alternate hypotheses. Will you use a left-tailed, right-tailed, or two-tailed test?
- Check Requirements** What sampling distribution will you use? Explain the rationale for your choice of sampling distribution. Compute the value of the sample test statistic.
- Find (or estimate) the P -value. Sketch the sampling distribution and show the area corresponding to the P -value.
- Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis? Are the data statistically significant at level α ?
- Interpret** your conclusion in the context of the application.

19. **Dividend Yield: Australian Bank Stocks** Let x be a random variable representing dividend yield of Australian bank stocks. We may assume that x has a normal distribution with $\sigma = 2.4\%$. A random sample of 10 Australian bank stocks gave the following yields.

5.7 4.8 6.0 4.9 4.0 3.4 6.5 7.1 5.3 6.1

The sample mean is $\bar{x} = 5.38\%$. For the entire Australian stock market, the mean dividend yield is $\mu = 4.7\%$ (Reference: *Forbes*). Do these data indicate that the dividend yield of all Australian bank stocks is higher than 4.7%? Use $\alpha = 0.01$.

20. **Glucose Level: Horses** Gentle Ben is a Morgan horse at a Colorado dude ranch. Over the past 8 weeks, a veterinarian took the following glucose readings from this horse (in mg/100 ml).

93 88 82 105 99 110 84 89

The sample mean is $\bar{x} \approx 93.8$. Let x be a random variable representing glucose readings taken from Gentle Ben. We may assume that x has a normal distribution, and we know from past experience that $\sigma = 12.5$. The mean glucose level for horses should be $\mu = 85$ mg/100 ml (Reference: *Merck Veterinary Manual*). Do these data indicate that Gentle Ben has an overall average glucose level higher than 85? Use $\alpha = 0.05$.

21. **Ecology: Hummingbirds** Bill Alther is a zoologist who studies Anna's hummingbird (*Calypte anna*) (Reference: *Hummingbirds* by K. Long and W. Alther). Suppose that in a remote part of the Grand Canyon, a random sample of six of these birds was caught, weighed, and released. The weights (in grams) were

3.7 2.9 3.8 4.2 4.8 3.1

The sample mean is $\bar{x} = 3.75$ grams. Let x be a random variable representing weights of Anna's hummingbirds in this part of the Grand Canyon. We assume that x has a normal distribution and $\sigma = 0.70$ gram. It is known that for the population of all Anna's hummingbirds, the mean weight is $\mu = 4.55$ grams. Do the data indicate that the mean weight of these birds in this part of the Grand Canyon is less than 4.55 grams? Use $\alpha = 0.01$.

22. **Finance: P/E of Stocks** The price-to-earnings (P/E) ratio is an important tool in financial work. A random sample of 14 large U.S. banks (J.P. Morgan, Bank of America, and others) gave the following P/E ratios (Reference: *Forbes*).

24	16	22	14	12	13	17
22	15	19	23	13	11	18

The sample mean is $\bar{x} \approx 17.1$. Generally speaking, a low P/E ratio indicates a “value” or bargain stock. A recent copy of the *Wall Street Journal* indicated that the P/E ratio of the entire S&P 500 stock index is $\mu = 19$. Let x be a random variable representing the P/E ratio of all large U.S. bank stocks. We assume that x has a normal distribution and $\sigma = 4.5$. Do these data indicate that the P/E ratio of all U.S. bank stocks is less than 19? Use $\alpha = 0.05$.

23. **Insurance: Hail Damage** Nationally, about 11% of the total U.S. wheat crop is destroyed each year by hail (Reference: *Agricultural Statistics*, U.S. Department of Agriculture). An insurance company is studying wheat hail damage claims in Weld County, Colorado. A random sample of 16 claims in Weld County gave the following data (% wheat crop lost to hail).

15	8	9	11	12	20	14	11
7	10	24	20	13	9	12	5

The sample mean is $\bar{x} = 12.5\%$. Let x be a random variable that represents the percentage of wheat crop in Weld County lost to hail. Assume that x has a normal distribution and $\sigma = 5.0\%$. Do these data indicate that the percentage of wheat crop lost to hail in Weld County is different (either way) from the national mean of 11%? Use $\alpha = 0.01$.

24. **Medical: Red Blood Cell Volume** Total blood volume (in ml) per body weight (in kg) is important in medical research. For healthy adults, the red blood cell volume mean is about $\mu = 28$ ml/kg (Reference: *Laboratory and Diagnostic Tests* by F. Fischbach). Red blood cell volume that is too low or too high can indicate a medical problem (see reference). Suppose that Roger has had seven blood tests, and the red blood cell volumes were

32	25	41	35	30	37	29
----	----	----	----	----	----	----

The sample mean is $\bar{x} \approx 32.7$ ml/kg. Let x be a random variable that represents Roger’s red blood cell volume. Assume that x has a normal distribution and $\sigma = 4.75$. Do the data indicate that Roger’s red blood cell volume is different (either way) from $\mu = 28$ ml/kg? Use a 0.01 level of significance.

SECTION 8.2

Testing the Mean μ

FOCUS POINTS

- Review the general procedure for testing using P -values.
- Test μ when σ is known using the normal distribution.
- Test μ when σ is unknown using a Student’s t distribution.
- Understand the “traditional” method of testing that uses critical regions and critical values instead of P -values.

In this section, we continue our study of testing the mean μ . The method we are using is called the P -value method. It was used extensively by the famous statistician R. A. Fisher and is the most popular method of testing in use today. At the end of this section, we present another method of testing called the *critical region method* (or *traditional method*). The critical region method was used extensively

by the statisticians J. Neyman and E. Pearson. In recent years, the use of this method has been declining. It is important to realize that for a fixed, preset level of significance α , both methods are logically equivalent.

In Section 8.1, we discussed the vocabulary and method of hypothesis testing using P -values. Let's quickly review the basic process.

1. We first state a proposed value for a population parameter in the null hypothesis H_0 . The alternate hypothesis H_1 states alternative values of the parameter, either $<$, $>$, or \neq the value proposed in H_0 . We also set the level of significance α . This is the risk we are willing to take of committing a type I error. That is, α is the probability of rejecting H_0 when it is, in fact, true.
2. We use a corresponding sample statistic from a simple random sample to challenge the statement made in H_0 . We convert the sample statistic to a test statistic, which is the corresponding value of the appropriate sampling distribution.
3. We use the sampling distribution of the test statistic and the type of test to compute the P -value of this statistic. Under the assumption that the null hypothesis is true, the P -value is the probability of getting a sample statistic as extreme as or more extreme than the observed statistic from our random sample.
4. Next, we conclude the test. If the P -value is very small, we have evidence to reject H_0 and adopt H_1 . What do we mean by "very small"? We compare the P -value to the preset level of significance α . If the P -value $\leq \alpha$, then we say that we have evidence to reject H_0 and adopt H_1 . Otherwise, we say that the sample evidence is insufficient to reject H_0 .
5. Finally, we *interpret* the results in the context of the application.

Knowing the sampling distribution of the sample test statistic is an essential part of the hypothesis testing process. For tests of μ , we use one of two sampling distributions for \bar{x} : the standard normal distribution or a Student's t distribution. As discussed in Chapters 6 and 7, the appropriate distribution depends upon our knowledge of the population standard deviation σ , the nature of the x distribution, and the sample size.

Part I: Testing μ When σ Is Known

In most real-world situations, σ is simply not known. However, in some cases a preliminary study or other information can be used to get a realistic and accurate value for σ .

PROCEDURE

HOW TO TEST μ WHEN σ IS KNOWN

Requirements

Let x be a random variable appropriate to your application. Obtain a simple random sample (of size n) of x values from which you compute the sample mean \bar{x} . The value of σ is already known (perhaps from a previous study). If you can assume that x has a normal distribution, then any sample size n will work. If you cannot assume this, then use a sample size $n \geq 30$.

Procedure

1. In the context of the application, state the *null and alternate hypotheses* and set the *level of significance* α .
2. Use the known σ , the sample size n , the value of x from the sample, and μ from the null hypothesis to compute the standardized *sample test statistic*.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

3. Use the standard normal distribution and the type of test, one-tailed or two-tailed, to find the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If $P\text{-value} \leq \alpha$, then reject H_0 . If $P\text{-value} > \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.

In Section 8.1, we examined *P*-value tests for normal distributions with relatively small sample sizes ($n < 30$). The next example does not assume a normal distribution, but has a large sample size ($n \geq 30$).

EXAMPLE 3

TESTING μ, σ KNOWN

Sunspots have been observed for many centuries. Records of sunspots from ancient Persian and Chinese astronomers go back thousands of years. Some archaeologists think sunspot activity may somehow be related to prolonged periods of drought in the southwestern United States. Let x be a random variable representing the average number of sunspots observed in a four-week period. A random sample of 40 such periods from Spanish colonial times gave the following data (Reference: M. Waldmeir, *Sun Spot Activity*, International Astronomical Union Bulletin).

12.5	14.1	37.6	48.3	67.3	70.0	43.8	56.5	59.7	24.0
12.0	27.4	53.5	73.9	104.0	54.6	4.4	177.3	70.1	54.0
28.0	13.0	6.5	134.7	114.0	72.7	81.2	24.1	20.4	13.3
9.4	25.7	47.8	50.0	45.3	61.0	39.0	12.0	7.2	11.3

The sample mean is $\bar{x} \approx 47.0$. Previous studies of sunspot activity during this period indicate that $\sigma = 35$. It is thought that for thousands of years, the mean number of sunspots per four-week period was about $\mu = 41$. Sunspot activity above this level may (or may not) be linked to gradual climate change. Do the data indicate that the mean sunspot activity during the Spanish colonial period was higher than 41? Use $\alpha = 0.05$.

SOLUTION:

- (a) Establish the null and alternate hypotheses.

Since we want to know whether the average sunspot activity during the Spanish colonial period was higher than the long-term average of $\mu = 41$,

$$H_0: \mu = 41 \quad \text{and} \quad H_1: \mu > 41$$

- (b) **Check Requirements** What distribution do we use for the sample test statistic?

Compute the test statistic from the sample data. Since $n \geq 30$ and we know σ , we use the standard normal distribution. Using $\bar{x} = 47$ from the sample, $\sigma = 35$, $\mu = 41$ from H_0 , and $n = 40$,

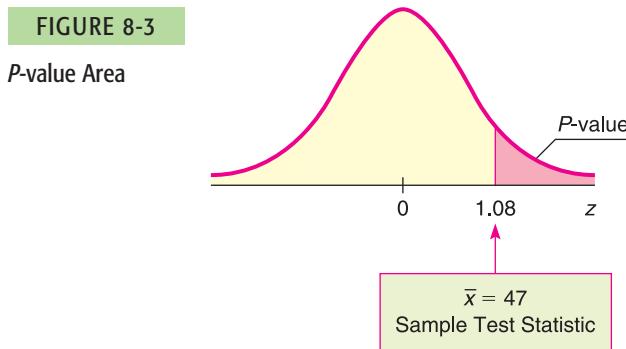
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{47 - 41}{35/\sqrt{40}} \approx 1.08$$

- (c) Find the *P*-value of the test statistic.

Figure 8-3 shows the *P*-value. Since we have a right-tailed test, the *P*-value is the area to the right of $z = 1.08$ shown in Figure 8-3. Using Table 5 of Appendix II, we find that

$$P\text{-value} = P(z > 1.08) \approx 0.1401.$$





(d) Conclude the test.

Since the P -value of $0.1401 > 0.05$ for α we do not reject H_0 .

(e) **Interpretation** Interpret the results in the context of the problem.

At the 5% level of significance, the evidence is not sufficient to reject H_0 . Based on the sample data, we do not think the average sunspot activity during the Spanish colonial period was higher than the long-term mean.



Part II: Testing μ When σ Is Unknown

In many real-world situations, you have only a random sample of data values. In addition, you may have some limited information about the probability distribution of your data values. Can you still test μ under these circumstances? In most cases, the answer is yes!

PROCEDURE

HOW TO TEST μ WHEN σ IS UNKNOWN

Requirements

Let x be a random variable appropriate to your application. Obtain a simple random sample (of size n) of x values from which you compute the sample mean \bar{x} and the sample standard deviation s . If you can assume that x has a normal distribution or simply a mound-shaped and symmetric distribution, then any sample size n will work. If you cannot assume this, use a sample size $n \geq 30$.

Procedure

1. In the context of the application, state the *null and alternate hypotheses* and set the *level of significance* α .
2. Use \bar{x} , s , and n from the sample, with μ from H_0 , to compute the *sample test statistic*.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad \text{with degrees of freedom } d.f. = n - 1$$
3. Use the Student's t distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If P -value $\leq \alpha$, then reject H_0 . If P -value $> \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.

d.f. for testing μ when σ unknown

Using the Student's t table to estimate P -values

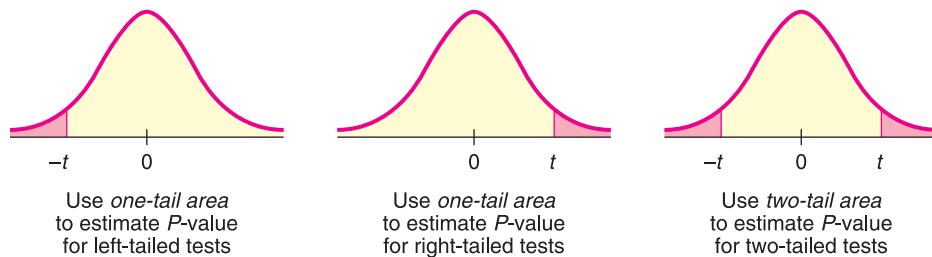
In Sections 7.2 and 7.4, we used Table 6 of Appendix II, Student's t Distribution, to find critical values t_c for confidence intervals. The critical values are in the body of the table. We find P -values in the rows headed by "one-tail area" and "two-tail area," depending on whether we have a one-tailed or two-tailed test. If the test statistic t for the sample statistic \bar{x} is negative, look up the P -value for the corresponding *positive* value of t (i.e., look up the P -value for $|t|$).

Note: In Table 6, areas are given in *one tail* beyond positive t on the right or negative t on the left, and in *two tails* beyond $\pm t$. Notice that in each column, two-tail area = 2(one-tail area). Consequently, we use *one-tail areas* as endpoints of the interval containing the P -value for *one-tailed tests*. We use *two-tail areas* as endpoints of the interval containing the P -value for *two-tailed tests*. (See Figure 8-4.)

Example 4 and Guided Exercise 4 show how to use Table 6 of Appendix II to find an interval containing the P -value corresponding to a test statistic t .

FIGURE 8-4

P -value for One-Tailed Tests and for Two-Tailed Tests



EXAMPLE 4

TESTING μ , σ UNKNOWN

The drug 6-mP (6-mercaptopurine) is used to treat leukemia. The following data represent the remission times (in weeks) for a random sample of 21 patients using 6-mP (Reference: E. A. Gehan, University of Texas Cancer Center).

10	7	32	23	22	6	16	34	32	25	11
20	19	6	17	35	6	13	9	6	10	

The sample mean is $\bar{x} \approx 17.1$ weeks, with sample standard deviation $s \approx 10.0$. Let x be a random variable representing the remission time (in weeks) for all patients using 6-mP. Assume the x distribution is mound-shaped and symmetric. A previously used drug treatment had a mean remission time of $\mu = 12.5$ weeks. Do the data indicate that the mean remission time using the drug 6-mP is different (either way) from 12.5 weeks? Use $\alpha = 0.01$.

SOLUTION:

- (a) Establish the null and alternate hypotheses.

Since we want to determine if the drug 6-mP provides a mean remission time that is different from that provided by a previously used drug having $\mu = 12.5$ weeks,

$$H_0: \mu = 12.5 \text{ weeks} \quad \text{and} \quad H_1: \mu \neq 12.5 \text{ weeks}$$

- (b) **Check Requirements** What distribution do we use for the sample test statistic t ? Compute the sample test statistic from the sample data.

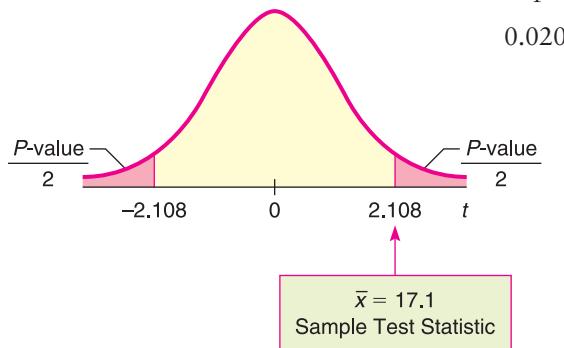
The x distribution is assumed to be mound-shaped and symmetric. Because we don't know σ , we use a Student's t distribution with $d.f. = 20$. Using $\bar{x} \approx 17.1$ and $s \approx 10.0$ from the sample data, $\mu = 12.5$ from H_0 , and $n = 21$,

$$t \approx \frac{\bar{x} - \mu}{s/\sqrt{n}} \approx \frac{17.1 - 12.5}{10.0/\sqrt{21}} \approx 2.108$$

- (c) Find the P -value or the interval containing the P -value.

Figure 8-5 shows the P -value. Using Table 6 of Appendix II, we find an interval containing the P -value. Since this is a two-tailed test, we use entries from the row headed by *two-tail area*. Look up the t value in the row headed by $d.f. = n - 1 = 21 - 1 = 20$. The sample statistic $t = 2.108$ falls between 2.086 and 2.528. The P -value for the sample t falls between the corresponding two-tail areas 0.050 and 0.020. (See Table 8-5.)

FIGURE 8-5

 P -value

$$0.020 < P\text{-value} < 0.050$$

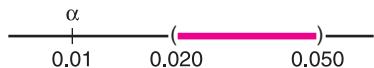
TABLE 8-5 Excerpt from Student's t Distribution (Table 6, Appendix II)

one-tail area
✓ two-tail area	0.050	0.020
$d.f. = 20$	2.086	2.528

↑
Sample $t = 2.108$

- (d) Conclude the test.

The following diagram shows the interval that contains the single P -value corresponding to the test statistic. Note that there is just one P -value corresponding to the test statistic. Table 6 of Appendix II does not give that specific value, but it does give a range that contains that specific P -value. As the diagram shows, the entire range is greater than α . This means the specific P -value is greater than α , so we cannot reject H_0 .



Note: Using the raw data, computer software gives P -value ≈ 0.048 . This value is in the interval we estimated. It is larger than the α value of 0.01, so we do not reject H_0 .

- (e) **Interpretation** Interpret the results in the context of the problem.

At the 1% level of significance, the evidence is not sufficient to reject H_0 . Based on the sample data, we cannot say that the drug 6-mP provides a different average remission time than the previous drug.



GUIDED EXERCISE 4

Testing μ , σ unknown

Archaeologists become excited when they find an anomaly in discovered artifacts. The anomaly may (or may not) indicate a new trading region or a new method of craftsmanship. Suppose the lengths of projectile points (arrowheads) at a certain archaeological site have mean length $\mu = 2.6$ cm. A random sample of 61 recently discovered projectile points in an adjacent cliff dwelling gave the following lengths (in cm) (Reference: A. Woosley and A. McIntyre, *Mimbres Mogollon Archaeology*, University of New Mexico Press).

3.1	4.1	1.8	2.1	2.2	1.3	1.7	3.0	3.7	2.3	2.6	2.2	2.8	3.0
3.2	3.3	2.4	2.8	2.8	2.9	2.9	2.2	2.4	2.1	3.4	3.1	1.6	3.1
3.5	2.3	3.1	2.7	2.1	2.0	4.8	1.9	3.9	2.0	5.2	2.2	2.6	1.9
4.0	3.0	3.4	4.2	2.4	3.5	3.1	3.7	3.7	2.9	2.6	3.6	3.9	3.5
1.9	4.0	4.0	4.6	1.9									

Continued

GUIDED EXERCISE 4 *continued*

The sample mean is $\bar{x} \approx 2.92$ cm and the sample standard deviation is $s \approx 0.85$, where x is a random variable that represents the lengths (in cm) of all projectile points found at the adjacent cliff dwelling site. Do these data indicate that the mean length of projectile points in the adjacent cliff dwelling is longer than 2.6 cm? Use a 1% level of significance.

- (a) State H_0 , H_1 , and α .
- (b) **Check Requirements** What sampling distribution should you use? What is the value of the sample test statistic t ?

- (c) When you use Table 6, Appendix II, to find an interval containing the P -value, do you use one-tail or two-tail areas? Why? Sketch a figure showing the P -value. Find an interval containing the P -value.

TABLE 8-6 Excerpt from Student's t Table

✓ one-tail area	... 0.005	0.0005
two-tail area	... 0.010	0.0010
$d.f. = 60$... 2.660	3.460
	↑	
	Sample $t = 2.940$	

- (d) Do we reject or fail to reject H_0 ?

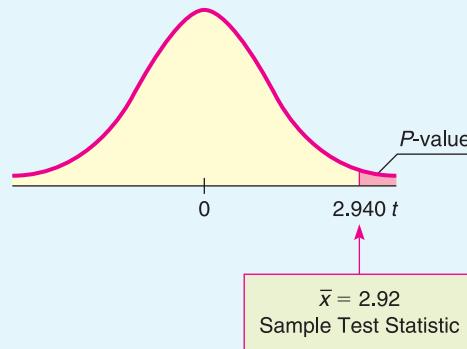
- (e) **Interpretation** Interpret your results in the context of the application.

→ $H_0: \mu = 2.6$ cm; $H_1: \mu > 2.6$ cm; $\alpha = 0.01$

→ Because $n \geq 30$ and σ is unknown, use the Student's t distribution with $d.f. = n - 1 = 61 - 1 = 60$. Using $\bar{x} \approx 2.92$, $s \approx 0.85$, $\mu = 2.6$ from H_0 and $n = 61$,

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{2.92 - 2.6}{0.85/\sqrt{61}} \approx 2.940$$

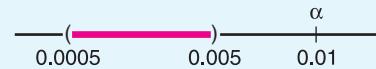
→ This is a right-tailed test, so use a one-tail area.

FIGURE 8-6 P -value

Using $d.f. = 60$, we find that the sample $t = 2.940$ is between the critical values 2.660 and 3.460. The sample P -value is then between the one-tail areas 0.005 and 0.0005.

$$0.0005 < P\text{-value} < 0.005$$

→ Since the interval containing the P -value lies to the left of $\alpha = 0.01$, we reject H_0 .



Note: Using the raw data, computer software gives $P\text{-value} \approx 0.0022$. This value is in our estimated range and is less than $\alpha = 0.01$ so we reject H_0 .

→ At the 1% level of significance, sample evidence is sufficiently strong to reject H_0 and conclude that the average projectile point length at the adjacent cliff dwelling site is longer than 2.6 cm.



TECH NOTES

The TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab all support testing of μ using the standard normal distribution. The TI-84Plus/TI-83Plus/TI-nspire and Minitab support testing of μ using a Student's t distribution. All the technologies return a P -value for the test.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) You can select to enter raw data (Data) or summary statistics (Stats). Enter the value of μ_0 used in the null hypothesis $H_0: \mu = \mu_0$. Select the symbol used in the alternate hypothesis ($\neq \mu_0, < \mu_0, > \mu_0$). To test μ using the standard normal distribution, press Stat, select Tests, and use option 1:Z-Test. The value for σ is required. To test μ using a Student's t distribution, use option 2:T-Test. Using data from Example 4 regarding remission times, we have the following displays. The P -value is given as p .

```
T-Test
Inpt:Data Stats
μ₀:12.5
List:L1
Freq:1
μ: $\neq$ μ₀ <μ₀ >μ₀
Calculate Draw
```

```
T-Tests
μ≠12.5
t=2.105902924
p=.0480466063
x̄=17.0952381
Sx=9.999523798
n=21
```

Excel 2007 In Excel, the ZTEST function finds the P -values for a right-tailed test. Click the ribbon choice Insert Function (f_x) . In the dialogue box, select Statistical for the category and ZTEST for the function. In the next dialogue box, give the cell range containing your data for the array. Use the value of μ stated in H_0 for x . Provide σ . Otherwise, Excel uses the sample standard deviation computed from the data.

Minitab Enter the raw data from a sample. Use the menu selections Stat \blacktriangleright Basic Stat \blacktriangleright 1-Sample z for tests using the standard normal distribution. For tests of μ using a Student's t distribution, select 1-Sample t.



Part III: Testing μ Using Critical Regions (Traditional Method)

Critical region method

The most popular method of statistical testing is the P -value method. For that reason, the P -value method is emphasized in this book. Another method of testing is called the *critical region method* or *traditional method*.

For a fixed, preset value of the level of significance α , both methods are logically equivalent. Because of this, we treat the traditional method as an “optional” topic and consider only the case of testing μ when σ is known.

Consider the null hypothesis $H_0: \mu = k$. We use information from a random sample, together with the sampling distribution for \bar{x} and the level of significance α , to determine whether or not we should reject the null hypothesis. The essential question is, “How much can \bar{x} vary from $\mu = k$ before we suspect that $H_0: \mu = k$ is false and reject it?”

The answer to the question regarding the relative sizes of \bar{x} and μ , as stated in the null hypothesis, depends on the sampling distribution of \bar{x} , the alternate hypothesis H_1 , and the level of significance α . If the sample test statistic \bar{x} is sufficiently different from the claim about μ made in the null hypothesis, we reject the null hypothesis.

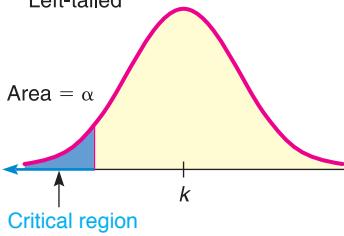
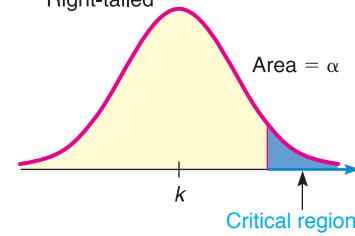
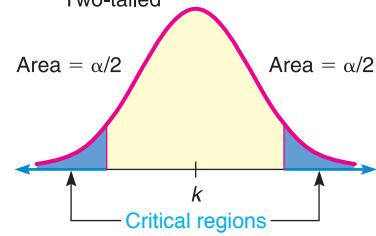
The values of \bar{x} for which we reject H_0 are called the *critical region* of the \bar{x} distribution. Depending on the alternate hypothesis, the critical region is located on the left side, the right side, or both sides of the \bar{x} distribution. Figure 8-7 shows the relationship of the critical region to the alternate hypothesis and the level of significance α .

Notice that the total area in the critical region is preset to be the level of significance α . This is *not* the P -value discussed earlier! In fact, you cannot set the P -value in advance because it is determined from a random sample. Recall that the level of significance α should (in theory) be a fixed, preset number assigned before drawing any samples.



Another method for concluding two-tailed tests involves the use of confidence intervals. Problems 25 and 26 at the end of this section discuss the confidence interval method.

FIGURE 8-7

Critical Regions for $H_0: \mu = k$ (a) $H_1: \mu < k$
Left-tailed(b) $H_1: \mu > k$
Right-tailed(c) $H_1: \mu \neq k$
Two-tailed

Critical values

The most commonly used levels of significance are $\alpha = 0.05$ and $\alpha = 0.01$. Critical regions of a standard normal distribution are shown for these levels of significance in Figure 8-8. *Critical values* are the boundaries of the critical region. Critical values designated as z_0 for the standard normal distribution are shown in Figure 8-8. For easy reference, they are also included in Table 5 of Appendix II, Areas of a Standard Normal Distribution.

The procedure for hypothesis testing using critical regions follows the same first two steps as the procedure using *P*-values. However, instead of finding a *P*-value for the sample test statistic, we check if the sample test statistic falls in the critical region. If it does, we reject H_0 . Otherwise, we do not reject H_0 .

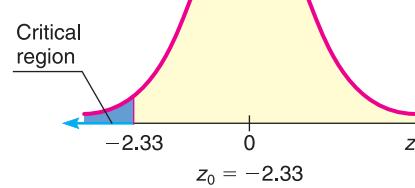
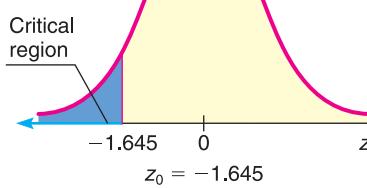
FIGURE 8-8

Critical Values z_0 for Tests Involving a Mean (Large Samples)

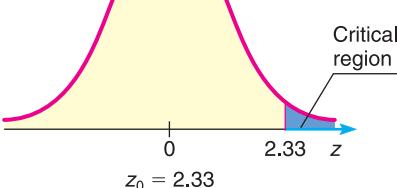
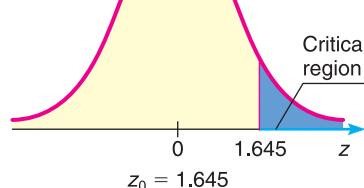
Level of significance

 $\alpha = 0.05$ $\alpha = 0.01$

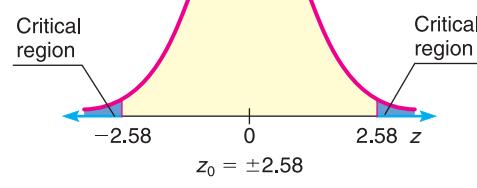
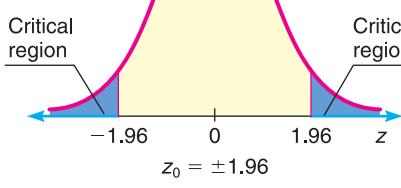
For a left-tailed test
 $H_1: \mu < k$
Critical value z_0
Critical region:
all $z < z_0$



For a right-tailed test
 $H_1: \mu > k$
Critical value z_0
Critical region:
all $z > z_0$



For a two-tailed test
 $H_1: \mu \neq k$
Critical value $\pm z_0$
Critical regions:
all $z < -z_0$ together
with all $z > +z_0$



PROCEDURE**HOW TO TEST μ WHEN σ IS KNOWN (CRITICAL REGION METHOD)**

Let x be a random variable appropriate to your application. Obtain a simple random sample (of size n) of x values from which you compute the sample mean \bar{x} . The value of σ is already known (perhaps from a previous study). If you can assume that x has a normal distribution, then any sample size n will work. If you cannot assume this, use a sample size $n \geq 30$. Then \bar{x} follows a distribution that is normal or approximately normal.

1. In the context of the application, state the *null and alternate hypotheses* and set the *level of significance* α . We use the most popular choices, $\alpha = 0.05$ or $\alpha = 0.01$.
2. Use the known σ , the sample size n , the value of \bar{x} from the sample, and μ from the null hypothesis to compute the standardized sample *test statistic*.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

3. Show the *critical region* and *critical value(s)* on a graph of the sampling distribution. The level of significance α and the alternate hypothesis determine the locations of critical regions and critical values.
4. *Conclude* the test. If the test statistic z computed in Step 2 is in the critical region, then reject H_0 . If the test statistic z is not in the critical region, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.

**EXAMPLE 5****CRITICAL REGION METHOD OF TESTING μ**

Consider Example 3 regarding sunspots. Let x be a random variable representing the number of sunspots observed in a four-week period. A random sample of 40 such periods from Spanish colonial times gave the number of sunspots per period. The raw data are given in Example 3. The sample mean is $\bar{x} \approx 47.0$. Previous studies indicate that for this period, $\sigma = 35$. It is thought that for thousands of years, the mean number of sunspots per four-week period was about $\mu = 41$. Do the data indicate that the mean sunspot activity during the Spanish colonial period was higher than 41? Use $\alpha = 0.05$.

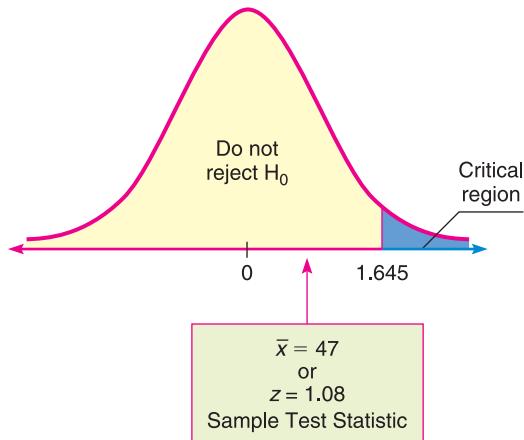
SOLUTION:

- (a) Set the null and alternate hypotheses.
As in Example 3, we use $H_0: \mu = 41$ and $H_1: \mu > 41$.
 - (b) Compute the sample test statistic.
As in Example 3, we use the standard normal distribution, with $\bar{x} = 47$, $\sigma = 35$, $\mu = 41$ from H_0 , and $n = 40$.
- $$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{47 - 41}{35/\sqrt{40}} \approx 1.08$$
- (c) Determine the critical region and critical value based on H_1 and $\alpha = 0.05$.
Since we have a right-tailed test, the critical region is the rightmost 5% of the standard normal distribution. According to Figure 8-8, the critical value is $z_0 = 1.645$.

(d) Conclude the test.

We conclude the test by showing the critical region, critical value, and sample test statistic $z = 1.08$ on the standard normal curve. For a right-tailed test with $\alpha = 0.05$ the critical value is $z_0 = 1.645$. Figure 8-9 shows the critical region. As we can see, the sample test statistic does not fall in the critical region. Therefore, we fail to reject H_0 .

FIGURE 8-9
Critical Region, $\alpha = 0.05$



(e) **Interpretation** Interpret the results in the context of the application.

At the 5% level of significance, the sample evidence is insufficient to justify rejecting H_0 . It seems that the average sunspot activity during the Spanish colonial period was the same as the historical average.

(f) How do results of the critical region method compare to the results of the P -value method for a 5% level of significance?

The results, as expected, are the same. In both cases, we fail to reject H_0 .

The critical region method of testing as outlined applies to tests of other parameters. As with the P -value method, you need to know the sampling distribution of the sample test statistic. Critical values for distributions are usually found in tables rather than in computer software outputs. For example, Table 6 of Appendix II provides critical values for Student's t distributions.

The critical region method of hypothesis testing is very general. The following procedure box outlines the process of concluding a hypothesis test using the critical region method.

PROCEDURE

HOW TO CONCLUDE TESTS USING THE CRITICAL REGION METHOD

1. Compute the sample test statistic using an appropriate sampling distribution.
2. Using the same sampling distribution, find the critical value(s) as determined by the level of significance α and the nature of the test: right-tailed, left-tailed, or two-tailed.
3. Compare the sample test statistic to the critical value(s).
 - (a) For a right-tailed test,
 - i. if sample test statistic \geq critical value, reject H_0 .
 - ii. if sample test statistic $<$ critical value, fail to reject H_0 .

Continued

- (b) For a left-tailed test,
 - i. if sample test statistic \leq critical value, reject H_0 .
 - ii. if sample test statistic $>$ critical value, fail to reject H_0 .
- (c) For a two-tailed test,
 - i. if sample test statistic lies at or beyond critical values, reject H_0 .
 - ii. if sample test statistic lies between critical values, fail to reject H_0 .

VIEWPOINT**Predator or Prey?**

Consider animals such as the arctic fox, gray wolf, desert lion, and South American jaguar. Each animal is a predator. What are the total sleep time (hours per day), maximum life span (years), and overall danger index from other animals? Now consider prey such as rabbits, deer, wild horses, and the Brazilian tapir (a wild pig). Is there a statistically significant difference in average sleep time, life span, and danger index? What about other variables such as the ratio of brain weight to body weight or the sleep exposure index (sleeping in a well-protected den or out in the open)? How did prehistoric humans fit into this picture? Scientists have collected a lot of data, and a great deal of statistical work has been done regarding such questions. For more information, see the web site <http://lib.stat.cmu.edu/> and follow the links to Datasets and then Sleep.

**SECTION 8.2
PROBLEMS**

1. | **Statistical Literacy** For the same sample data and null hypothesis, how does the P -value for a two-tailed test of μ compare to that for a one-tailed test?
2. | **Statistical Literacy** To test μ for an x distribution that is mound-shaped using sample size $n \geq 30$, how do you decide whether to use the normal or the Student's t distribution?
3. | **Statistical Literacy** When using the Student's t distribution to test μ , what value do you use for the degrees of freedom?
4. | **Critical Thinking** Consider a test for μ . If the P -value is such that you can reject H_0 at the 5% level of significance, can you always reject H_0 at the 1% level of significance? Explain.
5. | **Critical Thinking** Consider a test for μ . If the P -value is such that you can reject H_0 for $\alpha = 0.01$, can you always reject H_0 for $\alpha = 0.05$? Explain.
6. | **Critical Thinking** If sample data is such that for a one-tailed test of μ you can reject H_0 at the 1% level of significance, can you always reject H_0 for a two-tailed test at the same level of significance? Explain.
7. | **Basic Computation: P -value Corresponding to t Value** For a Student's t distribution with $d.f. = 10$ and $t = 2.930$,
 - (a) find an interval containing the corresponding P -value for a two-tailed test.
 - (b) find an interval containing the corresponding P -value for a right-tailed test.
8. | **Basic Computation: P -value Corresponding to t Value** For a Student's t distribution with $d.f. = 16$ and $t = -1.830$,
 - (a) find an interval containing the corresponding P -value for a two-tailed test.
 - (b) find an interval containing the corresponding P -value for a left-tailed test.

9. **Basic Computation: Testing μ , or Unknown** A random sample of 25 values is drawn from a mound-shaped and symmetric distribution. The sample mean is 10 and the sample standard deviation is 2. Use a level of significance of 0.05 to conduct a two-tailed test of the claim that the population mean is 9.5.
- Check Requirements** Is it appropriate to use a Student's t distribution? Explain. How many degrees of freedom do we use?
 - What are the hypotheses?
 - Compute the sample test statistic t .
 - Estimate the P -value for the test.
 - Do we reject or fail to reject H_0 ?
 - Interpret** the results.
10. **Basic Computation: Testing μ , or Unknown** A random sample has 49 values. The sample mean is 8.5 and the sample standard deviation is 1.5. Use a level of significance of 0.01 to conduct a left-tailed test of the claim that the population mean is 9.2.
- Check Requirements** Is it appropriate to use a Student's t distribution? Explain. How many degrees of freedom do we use?
 - What are the hypotheses?
 - Compute the sample test statistic t .
 - Estimate the P -value for the test.
 - Do we reject or fail to reject H_0 ?
 - Interpret** the results.
- Please provide the following information for Problems 11–22.
- What is the level of significance? State the null and alternate hypotheses.
 - Check Requirements** What sampling distribution will you use? Explain the rationale for your choice of sampling distribution. Compute the value of the sample test statistic.
 - Find (or estimate) the P -value. Sketch the sampling distribution and show the area corresponding to the P -value.
 - Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis? Are the data statistically significant at level α ?
 - Interpret** your conclusion in the context of the application.
- Note:* For degrees of freedom $d.f.$ not given in the Student's t table, use the closest $d.f.$ that is *smaller*. In some situations, this choice of $d.f.$ may increase the P -value by a small amount and therefore produce a slightly more "conservative" answer.
11. **Meteorology: Storms** *Weatherwise* is a magazine published by the American Meteorological Society. One issue gives a rating system used to classify Nor'easter storms that frequently hit New England and can cause much damage near the ocean. A severe storm has an average peak wave height of $\mu = 16.4$ feet for waves hitting the shore. Suppose that a Nor'easter is in progress at the severe storm class rating. Peak wave heights are usually measured from land (using binoculars) off fixed cement piers. Suppose that a reading of 36 waves showed an average wave height of $\bar{x} = 17.3$ feet. Previous studies of severe storms indicate that $\sigma = 3.5$ feet. Does this information suggest that the storm is (perhaps temporarily) increasing above the severe rating? Use $\alpha = 0.01$.
12. **Medical: Blood Plasma** Let x be a random variable that represents the pH of arterial plasma (i.e., acidity of the blood). For healthy adults, the mean of the x distribution is $\mu = 7.4$ (Reference: *Merck Manual*, a commonly used reference in medical schools and nursing programs). A new drug for arthritis has been developed. However, it is thought that this drug may change blood pH. A random sample of 31 patients with arthritis took the drug for 3 months. Blood tests showed that $\bar{x} = 8.1$ with sample standard deviation $s = 1.9$. Use a 5% level of significance to test the claim that the drug has changed (either way) the mean pH level of the blood.

13. **Wildlife: Coyotes** A random sample of 46 adult coyotes in a region of northern Minnesota showed the average age to be $\bar{x} = 2.05$ years, with sample standard deviation $s = 0.82$ years (based on information from the book *Coyotes: Biology, Behavior and Management* by M. Bekoff, Academic Press). However, it is thought that the overall population mean age of coyotes is $\mu = 1.75$. Do the sample data indicate that coyotes in this region of northern Minnesota tend to live longer than the average of 1.75 years? Use $\alpha = 0.01$.
14. **Fishing: Trout** Pyramid Lake is on the Paiute Indian Reservation in Nevada. The lake is famous for cutthroat trout. Suppose a friend tells you that the average length of trout caught in Pyramid Lake is $\mu = 19$ inches. However, the *Creel Survey* (published by the Pyramid Lake Paiute Tribe Fisheries Association) reported that of a random sample of 51 fish caught, the mean length was $\bar{x} = 18.5$ inches, with estimated standard deviation $s = 3.2$ inches. Do these data indicate that the average length of a trout caught in Pyramid Lake is less than $\mu = 19$ inches? Use $\alpha = 0.05$.
15. **Investing: Stocks** Socially conscious investors screen out stocks of alcohol and tobacco makers, firms with poor environmental records, and companies with poor labor practices. Some examples of “good,” socially conscious companies are Johnson and Johnson, Dell Computers, Bank of America, and Home Depot. The question is, are such stocks overpriced? One measure of value is the P/E, or price-to-earnings, ratio. High P/E ratios may indicate a stock is overpriced. For the S&P stock index of all major stocks, the mean P/E ratio is $\mu = 19.4$. A random sample of 36 “socially conscious” stocks gave a P/E ratio sample mean of $\bar{x} = 17.9$, with sample standard deviation $s = 5.2$ (Reference: *Morningstar*, a financial analysis company in Chicago). Does this indicate that the mean P/E ratio of all socially conscious stocks is different (either way) from the mean P/E ratio of the S&P stock index? Use $\alpha = 0.05$.
16. **Agriculture: Ground Water** Unfortunately, arsenic occurs naturally in some ground water (Reference: *Union Carbide Technical Report K/UR-1*). A mean arsenic level of $\mu = 8.0$ parts per billion (ppb) is considered safe for agricultural use. A well in Texas is used to water cotton crops. This well is tested on a regular basis for arsenic. A random sample of 37 tests gave a sample mean of $\bar{x} = 7.2$ ppb arsenic, with $s = 1.9$ ppb. Does this information indicate that the mean level of arsenic in this well is less than 8 ppb? Use $\alpha = 0.01$.
17. **Medical: Red Blood Cell Count** Let x be a random variable that represents red blood cell (RBC) count in millions of cells per cubic millimeter of whole blood. Then x has a distribution that is approximately normal. For the population of healthy female adults, the mean of the x distribution is about 4.8 (based on information from *Diagnostic Tests with Nursing Implications*, Springhouse Corporation). Suppose that a female patient has taken six laboratory blood tests over the past several months and that the RBC count data sent to the patient’s doctor are
- | | | | | | |
|-----|-----|-----|-----|-----|-----|
| 4.9 | 4.2 | 4.5 | 4.1 | 4.4 | 4.3 |
|-----|-----|-----|-----|-----|-----|
- Use a calculator with sample mean and sample standard deviation keys to verify that $\bar{x} = 4.40$ and $s \approx 0.28$.
 - Do the given data indicate that the population mean RBC count for this patient is lower than 4.8? Use $\alpha = 0.05$.
18. **Medical: Hemoglobin Count** Let x be a random variable that represents hemoglobin count (HC) in grams per 100 milliliters of whole blood. Then x has a distribution that is approximately normal, with population mean of about 14 for healthy adult women (see reference in Problem 17). Suppose that a female patient has taken 10 laboratory blood tests during the past year. The HC data sent to the patient’s doctor are
- | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 15 | 18 | 16 | 19 | 14 | 12 | 14 | 17 | 15 | 11 |
|----|----|----|----|----|----|----|----|----|----|

- i. Use a calculator with sample mean and sample standard deviation keys to verify that $\bar{x} = 15.1$ and $s \approx 2.51$.
ii. Does this information indicate that the population average HC for this patient is higher than 14? Use $\alpha = 0.01$.
19. **Ski Patrol: Avalanches** Snow avalanches can be a real problem for travelers in the western United States and Canada. A very common type of avalanche is called the slab avalanche. These have been studied extensively by David McClung, a professor of civil engineering at the University of British Columbia. Slab avalanches studied in Canada have an average thickness of $\mu = 67$ (Source: *Avalanche Handbook* by D. McClung and P. Schaefer). The ski patrol at Vail, Colorado, is studying slab avalanches in its region. A random sample of avalanches in spring gave the following thicknesses (in cm):
- | | | | | | | | |
|----|----|----|----|----|----|----|----|
| 59 | 51 | 76 | 38 | 65 | 54 | 49 | 62 |
| 68 | 55 | 64 | 67 | 63 | 74 | 65 | 79 |
- i. Use a calculator with mean and standard deviation keys to verify that $\bar{x} \approx 61.8$ and $s \approx 10.6$ cm.
ii. Assume the slab thickness has an approximately normal distribution. Use a 1% level of significance to test the claim that the mean slab thickness in the Vail region is different from that in Canada.
20. **Longevity: Honolulu** USA Today reported that the state with the longest mean life span is Hawaii, where the population mean life span is 77 years. A random sample of 20 obituary notices in the *Honolulu Advertiser* gave the following information about life span (in years) of Honolulu residents:
- | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 72 | 68 | 81 | 93 | 56 | 19 | 78 | 94 | 83 | 84 |
| 77 | 69 | 85 | 97 | 75 | 71 | 86 | 47 | 66 | 27 |
- i. Use a calculator with mean and standard deviation keys to verify that $\bar{x} = 71.4$ years and $s \approx 20.65$ years.
ii. Assuming that life span in Honolulu is approximately normally distributed, does this information indicate that the population mean life span for Honolulu residents is less than 77 years? Use a 5% level of significance.
21. **Fishing: Atlantic Salmon** Homser Lake, Oregon, has an Atlantic salmon catch and release program that has been very successful. The average fisherman's catch has been $\mu = 8.8$ Atlantic salmon per day (Source: *National Symposium on Catch and Release Fishing*, Humboldt State University). Suppose that a new quota system restricting the number of fishermen has been put into effect this season. A random sample of fishermen gave the following catches per day:
- | | | | | | | |
|----|---|----|----|---|----|----|
| 12 | 6 | 11 | 12 | 5 | 0 | 2 |
| 7 | 8 | 7 | 6 | 3 | 12 | 12 |
- i. Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} = 7.36$ and $s \approx 4.03$.
ii. Assuming the catch per day has an approximately normal distribution, use a 5% level of significance to test the claim that the population average catch per day is now different from 8.8.
22. **Archaeology: Tree Rings** Tree-ring dating from archaeological excavation sites is used in conjunction with other chronologic evidence to estimate occupation dates of prehistoric Indian ruins in the southwestern United States. It is thought that Burnt Mesa Pueblo was occupied around 1300 A.D. (based on evidence from potsherds and stone tools). The following data give tree-ring dates (A.D.) from adjacent archaeological sites (*Bandelier Archaeological Excavation Project: Summer 1990 Excavations at Burnt Mesa Pueblo*, edited by T. Kohler, Washington State University Department of Anthropology, 1992):
- | | | | | |
|------|------|------|------|------|
| 1189 | 1267 | 1268 | 1275 | 1275 |
| 1271 | 1272 | 1316 | 1317 | 1230 |

- i. Use a calculator with mean and standard deviation keys to verify that $\bar{x} = 1268$ and $s \approx 37.29$ years.
- ii. Assuming the tree-ring dates in this excavation area follow a distribution that is approximately normal, does this information indicate that the population mean of tree-ring dates in the area is different from (either higher or lower than) that in 1300 A.D.? Use a 1% level of significance.

23. **Critical Thinking: One-Tailed versus Two-Tailed Tests**

- (a) For the same data and null hypothesis, is the P -value of a one-tailed test (right or left) larger or smaller than that of a two-tailed test? Explain.
- (b) For the same data, null hypothesis, and level of significance, is it possible that a one-tailed test results in the conclusion to reject H_0 while a two-tailed test results in the conclusion to fail to reject H_0 ? Explain.
- (c) For the same data, null hypothesis, and level of significance, if the conclusion is to reject H_0 based on a two-tailed test, do you also reject H_0 based on a one-tailed test? Explain.
- (d) If a report states that certain data were used to reject a given hypothesis, would it be a good idea to know what type of test (one-tailed or two-tailed) was used? Explain.

24. **Critical Thinking: Comparing Hypothesis Tests with U.S. Courtroom System**

Compare statistical testing with legal methods used in a U.S. court setting. Then discuss the following topics in class or consider the topics on your own. Please write a brief but complete essay in which you answer the following questions.

- (a) In a court setting, the person charged with a crime is initially considered to be innocent. The claim of innocence is maintained until the jury returns with a decision. Explain how the claim of innocence could be taken to be the null hypothesis. Do we assume that the null hypothesis is true throughout the testing procedure? What would the alternate hypothesis be in a court setting?
- (b) The court claims that a person is innocent if the evidence against the person is not adequate to find him or her guilty. This does not mean, however, that the court has necessarily *proved* the person to be innocent. It simply means that the evidence against the person was not adequate for the jury to find him or her guilty. How does this situation compare with a statistical test for which the conclusion is “do not reject” (i.e., accept) the null hypothesis? What would be a type II error in this context?
- (c) If the evidence against a person is adequate for the jury to find him or her guilty, then the court claims that the person is guilty. Remember, this does not mean that the court has necessarily *proved* the person to be guilty. It simply means that the evidence against the person was strong enough to find him or her guilty. How does this situation compare with a statistical test for which the conclusion is to “reject” the null hypothesis? What would be a type I error in this context?
- (d) In a court setting, the final decision as to whether the person charged is innocent or guilty is made at the end of the trial, usually by a jury of impartial people. In hypothesis testing, the final decision to reject or not reject the null hypothesis is made at the end of the test by using information or data from an (impartial) random sample. Discuss these similarities between statistical hypothesis testing and a court setting.
- (e) We hope that you are able to use this discussion to increase your understanding of statistical testing by comparing it with something that is a well-known part of our American way of life. However, all analogies have weak points, and it is important not to take the analogy between statistical hypothesis testing and legal court methods too far. For instance, the judge does not set a level of significance and tell the jury to determine a verdict that is wrong only 5% or 1% of the time. Discuss some of these weak points in the analogy between the court setting and hypothesis testing.



25.

- Expand Your Knowledge: Confidence Intervals and Two-Tailed Hypothesis Tests** Is there a relationship between confidence intervals and two-tailed hypothesis tests? Let c be the level of confidence used to construct a confidence interval from sample data. Let α be the level of significance for a two-tailed hypothesis test. The following statement applies to hypothesis tests of the mean.

For a two-tailed hypothesis test with level of significance α and null hypothesis $H_0: \mu = k$, we *reject* H_0 whenever k falls *outside* the $c = 1 - \alpha$ confidence interval for μ based on the sample data. When k falls within the $c = 1 - \alpha$ confidence interval, we do not reject H_0 .

(A corresponding relationship between confidence intervals and two-tailed hypothesis tests also is valid for other parameters, such as p , $\mu_1 - \mu_2$, and $p_1 - p_2$, which we will study in Sections 8.3 and 8.5.) Whenever the value of k given in the null hypothesis falls *outside* the $c = 1 - \alpha$ confidence interval for the parameter, we *reject* H_0 . For example, consider a two-tailed hypothesis test with $\alpha = 0.01$ and

$$H_0: \mu = 20 \quad H_1: \mu \neq 20$$

A random sample of size 36 has a sample mean $\bar{x} = 22$ from a population with standard deviation $\sigma = 4$.



26.

- Confidence Intervals and Two-Tailed Hypothesis Tests** Change the null hypothesis of Problem 25 to $H_0: \mu = 21$. Repeat parts (a) and (b).

27.

- Critical Region Method: Standard Normal** Solve Problem 11 using the critical region method of testing (i.e., traditional method). Compare your conclusion with the conclusion obtained by using the *P*-value method. Are they the same?

28.

- Critical Region Method: Student's *t*** Table 6 of Appendix II gives critical values for the Student's *t* distribution. Use an appropriate *d.f.* as the row header. For a *right-tailed* test, the column header is the value of α found in the *one-tail area* row. For a *left-tailed* test, the column header is the value of α found in the *one-tail area* row, but you must change the sign of the critical value t to $-t$. For a *two-tailed* test, the column header is the value of α from the *two-tail area* row. The critical values are the $\pm t$ values shown. Solve Problem 12 using the critical region method of testing. Compare your conclusion with the conclusion obtained by using the *P*-value method. Are they the same?

29.

- Critical Region Method: Student's *t*** Solve Problem 13 using the critical region method of testing. *Hint:* See Problem 28. Compare your conclusion with the conclusion obtained by using the *P*-value method. Are they the same?

30.

- Critical Region Method: Student's *t*** Solve Problem 14 using the critical region method of testing. *Hint:* See Problem 28. Compare your conclusion with the conclusion obtained by using the *P*-value method. Are they the same?

SECTION 8.3**Testing a Proportion p** **FOCUS POINTS**

- Identify the components needed for testing a proportion.
- Compute the sample test statistic.
- Find the P -value and conclude the test.

Many situations arise that call for tests of proportions or percentages rather than means. For instance, a college registrar may want to determine if the proportion of students wanting 3-week intensive courses has increased.

How can we make such a test? In this section, we will study tests involving proportions (i.e., percentages or proportions). Such tests are similar to those in Sections 8.1 and 8.2. The main difference is that we are working with a distribution of proportions.

Throughout this section, we will assume that the situations we are dealing with satisfy the conditions underlying the binomial distribution. In particular, we will let r be a binomial random variable. This means that r is the number of successes out of n independent binomial trials (for the definition of a binomial trial, see Section 5.2). We will use $\hat{p} = r/n$ as our estimate for p , the population probability of success on each trial. The letter q again represents the population probability of failure on each trial, and so $q = 1 - p$. We also assume that the samples are large (i.e., $np > 5$ and $nq > 5$).

For large samples, $np > 5$ and $nq > 5$, the distribution of $\hat{p} = r/n$ values is well approximated by a *normal curve* with mean μ and standard deviation σ as follows:

$$\begin{aligned}\mu &= p \\ \sigma &= \sqrt{\frac{pq}{n}}\end{aligned}$$

Tests for a single proportion

Criteria for using normal approximation to binomial
 $np > 5$ and $nq > 5$

Hypothesis for testing p **Sample test statistic \hat{p}**

The null and alternate hypotheses for tests of proportions are

Left-Tailed Test	Right-Tailed Test	Two-Tailed Test
$H_0: p = k$	$H_0: p = k$	$H_0: p = k$
$H_1: p < k$	$H_1: p > k$	$H_1: p \neq k$

depending on what is asked for in the problem. Notice that since p is a probability, the value k must be between 0 and 1.

For tests of proportions, we need to convert the sample test statistic \hat{p} to a z value. Then we can find a P -value appropriate for the test. The \hat{p} distribution is approximately normal, with mean p and standard deviation $\sqrt{pq/n}$. Therefore, the conversion of \hat{p} to z follows the formula

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

where $\hat{p} = r/n$ is the sample test statistic

n = number of trials

p = proportion specified in H_0

$q = 1 - p$

Using this mathematical information about the sampling distribution for \hat{p} , the basic procedure is similar to tests you have conducted before.

PROCEDURE**HOW TO TEST A PROPORTION p** **Requirements**

Consider a binomial experiment with n trials, where p represents the population probability of success and $q = 1 - p$ represents the population probability of failure. Let r be a random variable that represents the number of successes out of the n binomial trials. The number of trials n should be sufficiently large so that both $np > 5$ and $nq > 5$ (use p from the null hypothesis). In this case, $\hat{p} = r/n$ can be approximated by the normal distribution.

Procedure

1. In the context of the application, state the *null and alternate hypotheses* and set the *level of significance* α .
2. Compute the standardized *sample test statistic*

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

where p is the value specified in H_0 and $q = 1 - p$.

3. Use the standard normal distribution and the type of test, one-tailed or two-tailed, to find the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If $P\text{-value} \leq \alpha$, then reject H_0 . If $P\text{-value} > \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.

EXAMPLE 6

Ned Frisk/Blend Images/
Jupiter Images

**TESTING p**

A team of eye surgeons has developed a new technique for a risky eye operation to restore the sight of people blinded from a certain disease. Under the old method, it is known that only 30% of the patients who undergo this operation recover their eyesight.

Suppose that surgeons in various hospitals have performed a total of 225 operations using the new method and that 88 have been successful (i.e., the patients fully recovered their sight). Can we justify the claim that the new method is better than the old one? (Use a 1% level of significance.)

SOLUTION:

- (a) Establish H_0 and H_1 and note the level of significance.

The level of significance is $\alpha = 0.01$. Let p be the probability that a patient fully recovers his or her eyesight. The null hypothesis is that p is still 0.30, even for the new method. The alternate hypothesis is that the new method has improved the chances of a patient recovering his or her eyesight. Therefore,

$$H_0: p = 0.30 \quad \text{and} \quad H_1: p > 0.30$$

- (b) **Check Requirements** Is the sample sufficiently large to justify use of the normal distribution for \hat{p} ? Find the sample test statistic \hat{p} and convert it to a z value, if appropriate.

Using p from H_0 we note that $np = 225(0.3) = 67.5$ is greater than 5 and that $nq = 225(0.7) = 157.5$ is also greater than 5, so we can use the normal distribution for the sample statistic \hat{p} .

$$\hat{p} = \frac{r}{n} = \frac{88}{225} \approx 0.39$$

The z value corresponding to \hat{p} is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \approx \frac{0.39 - 0.30}{\sqrt{\frac{0.30(0.70)}{225}}} \approx 2.95$$

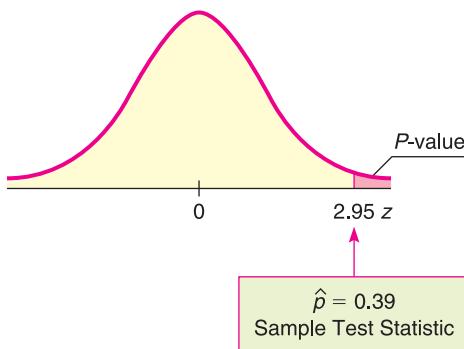
In the formula, the value for p is from the null hypothesis. H_0 specifies that $p = 0.30$, so $q = 1 - 0.30 = 0.70$.

- (c) Find the P -value of the test statistic.

Figure 8-10 shows the P -value. Since we have a right-tailed test, the P -value is the area to the right of $z = 2.95$. Using the normal distribution (Table 5 of Appendix II), we find that P -value = $P(z > 2.95) \approx 0.0016$.

FIGURE 8-10

P-value Area



- (d) Conclude the test.

Since the P -value of $0.0016 \leq 0.01$ for α , we reject H_0 .

- (e) **Interpretation** Interpret the results in the context of the problem.

At the 1% level of significance, the evidence shows that the population probability of success for the new surgery technique is higher than that of the old technique.



GUIDED EXERCISE 5

Testing p

A botanist has produced a new variety of hybrid wheat that is better able to withstand drought than other varieties. The botanist knows that for the parent plants, the proportion of seeds germinating is 80%. The proportion of seeds germinating for the hybrid variety is unknown, but the botanist claims that it is 80%. To test this claim, 400 seeds from the hybrid plant are tested, and it is found that 312 germinate. Use a 5% level of significance to test the claim that the proportion germinating for the hybrid is 80%.

- (a) Let p be the proportion of hybrid seeds that will germinate. Notice that we have no prior knowledge about the germination proportion for the hybrid plant. State H_0 and H_1 . What is the required level of significance?

$$\Rightarrow H_0: p = 0.80; H_1: p \neq 0.80; \alpha = 0.05$$

- (b) **Check Requirements** Using the value of p in H_0 , are both $np > 5$ and $nq > 5$? Can we use the normal distribution for \hat{p} ?

$$\Rightarrow \begin{aligned} \text{From } H_0, p &= 0.80 \text{ and } q = 1 - p = 0.20 \\ np &= 400(0.8) = 320 > 5 \\ nq &= 400(0.2) = 80 > 5 \end{aligned}$$

So, we can use the normal distribution for \hat{p} .

Continued

GUIDED EXERCISE 5 *continued*

Calculate the sample test statistic \hat{p} .

- (c) Next, we convert the sample test statistic $\hat{p} = 0.78$ to a z value. Based on our choice for H_0 , what value should we use for p in our formula? Since $q = 1 - p$, what value should we use for q ? Using these values for p and q , convert \hat{p} to a z value.

The number of trials is $n = 400$, and the number of successes is $r = 312$. Thus,

$$\hat{p} = \frac{r}{n} = \frac{312}{400} = 0.78$$



According to H_0 , $p = 0.80$. Then $q = 1 - p = 0.20$. Using these values in the following formula gives

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.78 - 0.80}{\sqrt{\frac{0.80(0.20)}{400}}} = -1.00$$

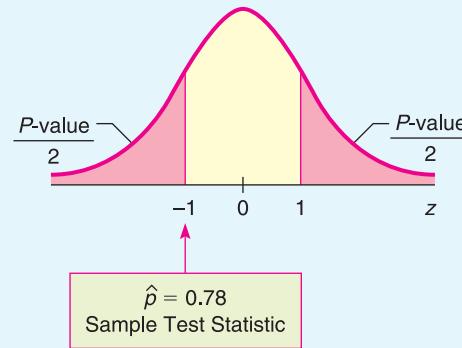
CALCULATOR NOTE If you evaluate the denominator separately, be sure to carry at least four digits after the decimal.

- (d) Is the test right-tailed, left-tailed, or two-tailed? Find the P -value of the sample test statistic and sketch a standard normal curve showing the P -value.

For a two-tailed test, using the normal distribution (Table 5 of Appendix II), we find that

$$P\text{-value} = 2P(z < -1.00) = 2(0.1587) = 0.3174$$

FIGURE 8-11 P -value



- (e) Do we reject or fail to reject H_0 ?



We fail to reject H_0 because P -value of 0.3171 > 0.05 for α .

- (f) **Interpretation** Interpret your conclusion in the context of the application.



At the 5% level of significance, there is insufficient evidence to conclude that the germination rate is not 80%.

Critical region method

Since the \hat{p} sampling distribution is approximately normal, we use Table 5, "Areas of a Standard Normal Distribution," in Appendix II to find critical values.



EXAMPLE 7

CRITICAL REGION METHOD FOR TESTING p

Let's solve Guided Exercise 5 using the critical region approach. In that problem, 312 of 400 seeds from a hybrid wheat variety germinated. For the parent plants, the proportion of germinating seeds is 80%. Use a 5% level of significance to test the claim that the population proportion of germinating seeds from the hybrid wheat is different from that of the parent plants.

SOLUTION:

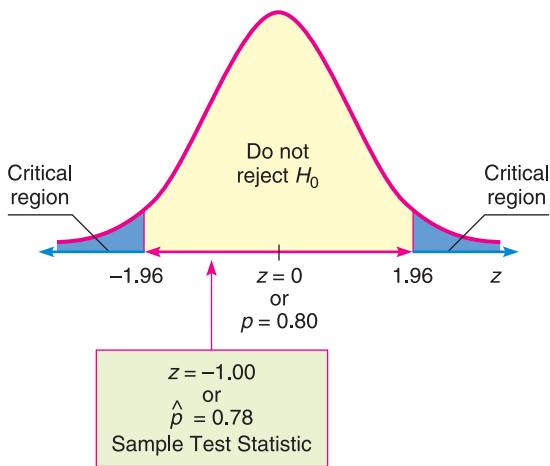
- (a) As in Guided Exercise 5, we have $\alpha = 0.05$, $H_0: p = 0.80$, and $H_1: p \neq 0.80$. The next step is to find \hat{p} and the corresponding sample test statistic z . This



was done in Guided Exercise 5, where we found that $\hat{p} = 0.78$, with corresponding $z = -1.00$.

- (b) Now we find the critical value z_0 for a two-tailed test using $\alpha = 0.05$. This means that we want the total area 0.05 divided between two tails, one to the right of z_0 and one to the left of $-z_0$. As shown in Figure 8-8 of Section 8.2, the critical value(s) are ± 1.96 (See also Table 5, part (c), of Appendix II for critical values of the z distribution.)
- (c) Figure 8-12 shows the critical regions and the location of the sample test statistic.

FIGURE 8-12
Critical Regions, $\alpha = 0.05$



- (d) Finally, we conclude the test and compare the results to Guided Exercise 5. Since the sample test statistic does not fall in the critical region, we fail to reject H_0 and conclude that, at the 5% level of significance, the evidence is not strong enough to reject the botanist's claim. This result, as expected, is consistent with the conclusion obtained by using the P -value method.



TECH NOTES

The TI-84Plus/TI-83Plus/TI-nspire calculators and Minitab support tests of proportions. The output for both technologies includes the sample proportion \hat{p} and the P -value of \hat{p} . Minitab also includes the z value corresponding to \hat{p} .

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Press STAT, select TESTS, and use option 5:1-PropZTest. The value of p_0 is from the null hypothesis $H_0: p = p_0$. The number of successes is the value for x .

Minitab Menu selections: Stat > Basic Statistics > 1 Proportion. Under options, set the test proportion as the value in H_0 . Choose to use the normal distribution.



CRITICAL THINKING

Issues Related to Hypothesis Testing

Through our work with hypothesis tests of μ and p , we've gained experience in setting up, performing, and interpreting results of such tests.

We know that different random samples from the same population are very likely to have sample statistics \bar{x} or \hat{p} that differ from their corresponding parameters μ or p . Some values of a statistic from a random sample will be close to the corresponding parameter. Others may be farther away simply because we happened to draw a random sample of more extreme data values.

The central question in hypothesis testing is whether or not you think the value of the sample test statistic is too far away from the value of the population parameter proposed in H_0 to occur by chance alone.

This is where the P -value of the sample test statistic comes into play. The P -value of the sample test statistic tells you the probability that you would get a sample statistic as far away as, or farther from, the value of the parameter as stated in the null hypothesis H_0 .

If the P -value is very small, you reject H_0 . But what does “very small” mean? It is customary to define “very small” as smaller than the preset level of significance α .

When you reject H_0 , are you absolutely certain that you are making a correct decision? The answer is no! You are simply willing to take a chance that you are making a mistake (a type I error). The level of significance α describes the chance of making a mistake if you reject H_0 when it is, in fact, true.

Several issues come to mind:

1. What if the P -value is so close to α that we “barely” reject or fail to reject H_0 ? In such cases, researchers might attempt to clarify the results by
 - increasing the sample size.
 - controlling the experiment to reduce the standard deviation.Both actions tend to increase the magnitude of the z or t value of the sample test statistic, resulting in a smaller corresponding P -value.
2. How reliable is the study and the measurements in the sample?
 - When reading results of a statistical study, be aware of the source of the data and the reliability of the organization doing the study.
 - Is the study sponsored by an organization that might profit or benefit from the stated conclusions? If so, look at the study carefully to ensure that the measurements, sampling technique, and handling of data are proper and meet professional standards.

VIEWPOINT

Who Did What?

Art, music, literature, and science share a common need to classify things:

Who painted that picture? Who composed that music? Who wrote that document? Who should get that patent? In statistics, such questions are called classification problems. For example, the Federalist Papers were published anonymously in 1787–1788 by Alexander Hamilton, John Jay, and James Madison. But who wrote what? That question is addressed by F. Mosteller (Harvard University) and D. Wallace (University of Chicago) in the book Statistics: A Guide to the Unknown, edited by J. M. Tanur. Other scholars have studied authorship regarding Plato’s Republic and Plato’s Dialogues, including the Symposium. For more information on this topic, see the source in Problems 15 and 16 of this exercise set.

SECTION 8.3 PROBLEMS

1. **Statistical Literacy** To use the normal distribution to test a proportion p , the conditions $np > 5$ and $nq > 5$ must be satisfied. Does the value of p come from H_0 , or is it estimated by using \hat{p} from the sample?
2. **Statistical Literacy** Consider a binomial experiment with n trials and r successes. For a test for a proportion p , what is the formula for the sample test statistic? Describe each symbol used in the formula.

3. **Critical Thinking** In general, if sample data are such that the null hypothesis is rejected at the $\alpha = 1\%$ level of significance based on a two-tailed test, is H_0 also rejected at the $\alpha = 1\%$ level of significance for a corresponding one-tailed test? Explain.
4. **Critical Thinking** An article in a newspaper states that the proportion of traffic accidents involving road rage is higher this year than it was last year, when it was 15%. Reconstruct the information of the study in terms of a hypothesis test. Discuss possible hypotheses, possible issues about the sample, possible levels of significance, and the “absolute truth” of the conclusion.
5. **Basic Computation: Testing p** A random sample of 30 binomials trials resulted in 12 successes. Test the claim that the population proportion of successes does not equal 0.50. Use a level of significance of 0.05.
 - Check Requirements** Can a normal distribution be used for the \hat{p} distribution? Explain.
 - State the hypotheses.
 - Compute \hat{p} and the corresponding standardized sample test statistic.
 - Find the P -value of the test statistic.
 - Do you reject or fail to reject H_0 ? Explain.
 - Interpretation** What do the results tell you?
6. **Basic Computation: Testing p** A random sample of 60 binomials trials resulted in 18 successes. Test the claim that the population proportion of successes exceeds 18%. Use a level of significance of 0.01.
 - Check Requirements** Can a normal distribution be used for the \hat{p} distribution? Explain.
 - State the hypotheses.
 - Compute \hat{p} and the corresponding standardized sample test statistic.
 - Find the P -value of the test statistic.
 - Do you reject or fail to reject H_0 ? Explain.
 - Interpretation** What do the results tell you?

For Problems 7–21, please provide the following information.

- What is the level of significance? State the null and alternate hypotheses.
- Check Requirements** What sampling distribution will you use? Do you think the sample size is sufficiently large? Explain. Compute the value of the sample test statistic.
- Find the P -value of the test statistic. Sketch the sampling distribution and show the area corresponding to the P -value.
- Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis? Are the data statistically significant at level α ?
- Interpret** your conclusion in the context of the application.

7. **Focus Problem: Benford's Law** Please read the Focus Problem at the beginning of this chapter. Recall that Benford's Law claims that numbers chosen from very large data files tend to have “1” as the first nonzero digit disproportionately often. In fact, research has shown that if you randomly draw a number from a very large data file, the probability of getting a number with “1” as the leading digit is about 0.301 (see the reference in this chapter's Focus Problem).

Now suppose you are an auditor for a very large corporation. The revenue report involves millions of numbers in a large computer file. Let us say you took a random sample of $n = 215$ numerical entries from the file and $r = 46$ of the entries had a first nonzero digit of 1. Let p represent the population proportion of all numbers in the corporate file that have a first nonzero digit of 1.

- Test the claim that p is less than 0.301. Use $\alpha = 0.01$.
- If p is in fact less than 0.301, would it make you suspect that there are not enough numbers in the data file with leading 1's? Could this indicate that the books have been “cooked” by “pumping up” or inflating the numbers? Comment from the viewpoint of a stockholder. Comment from the perspective

- of the Federal Bureau of Investigation as it looks for money laundering in the form of false profits.
- iii. Comment on the following statement: “If we reject the null hypothesis at level of significance α , we have not *proved* H_0 to be false. We can say that the probability is α that we made a mistake in rejecting H_0 .” Based on the outcome of the test, would you recommend further investigation before accusing the company of fraud?
8. **Focus Problem: Benford’s Law** Again, suppose you are the auditor for a very large corporation. The revenue file contains millions of numbers in a large computer data bank (see Problem 7). You draw a random sample of $n = 228$ numbers from this file and $r = 92$ have a first nonzero digit of 1. Let p represent the population proportion of all numbers in the computer file that have a leading digit of 1.
- i. Test the claim that p is more than 0.301. Use $\alpha = 0.01$.
 - ii. If p is in fact larger than 0.301, it would seem there are too many numbers in the file with leading 1’s. Could this indicate that the books have been “cooked” by artificially lowering numbers in the file? Comment from the point of view of the Internal Revenue Service. Comment from the perspective of the Federal Bureau of Investigation as it looks for “profit skimming” by unscrupulous employees.
 - iii. Comment on the following statement: “If we reject the null hypothesis at level of significance α , we have not *proved* H_0 to be false. We can say that the probability is α that we made a mistake in rejecting H_0 .” Based on the outcome of the test, would you recommend further investigation before accusing the company of fraud?
9. **Sociology: Crime Rate** Is the national crime rate really going down? Some sociologists say yes! They say that the reason for the decline in crime rates in the 1980s and 1990s is demographics. It seems that the population is aging, and older people commit fewer crimes. According to the FBI and the Justice Department, 70% of all arrests are of males aged 15 to 34 years (Source: *True Odds* by J. Walsh, Merritt Publishing). Suppose you are a sociologist in Rock Springs, Wyoming, and a random sample of police files showed that of 32 arrests last month, 24 were of males aged 15 to 34 years. Use a 1% level of significance to test the claim that the population proportion of such arrests in Rock Springs is different from 70%.
10. **College Athletics: Graduation Rate** Women athletes at the University of Colorado, Boulder, have a long-term graduation rate of 67% (Source: *Chronicle of Higher Education*). Over the past several years, a random sample of 38 women athletes at the school showed that 21 eventually graduated. Does this indicate that the population proportion of women athletes who graduate from the University of Colorado, Boulder, is now less than 67%? Use a 5% level of significance.
11. **Highway Accidents: DUI** The U.S. Department of Transportation, National Highway Traffic Safety Administration, reported that 77% of all fatally injured automobile drivers were intoxicated. A random sample of 27 records of automobile driver fatalities in Kit Carson County, Colorado, showed that 15 involved an intoxicated driver. Do these data indicate that the population proportion of driver fatalities related to alcohol is less than 77% in Kit Carson County? Use $\alpha = 0.01$.
12. **Preference: Color** What is your favorite color? A large survey of countries, including the United States, China, Russia, France, Turkey, Kenya, and others, indicated that most people prefer the color blue. In fact, about 24% of the population claim blue as their favorite color (Reference: Study by J. Bunge and A. Freeman-Gallant, Statistics Center, Cornell University). Suppose a random sample of $n = 56$ college students were surveyed and $r = 12$ of them said that blue is their favorite color. Does this information imply that the color preference of all college students is different (either way) from that of the general population? Use $\alpha = 0.05$.

13. **Wildlife: Wolves** The following is based on information from *The Wolf in the Southwest: The Making of an Endangered Species* by David E. Brown (University of Arizona Press). Before 1918, the proportion of female wolves in the general population of all southwestern wolves was about 50%. However, after 1918, southwestern cattle ranchers began a widespread effort to destroy wolves. In a recent sample of 34 wolves, there were only 10 females. One theory is that male wolves tend to return sooner than females to their old territories where their predecessors were exterminated. Do these data indicate that the population proportion of female wolves is now less than 50% in the region? Use $\alpha = 0.01$.
14. **Fishing: Northern Pike** Athabasca Fishing Lodge is located on Lake Athabasca in northern Canada. In one of its recent brochures, the lodge advertises that 75% of its guests catch northern pike over 20 pounds. Suppose that last summer 64 out of a random sample of 83 guests did, in fact, catch northern pike weighing over 20 pounds. Does this indicate that the population proportion of guests who catch pike over 20 pounds is different from 75% (either higher or lower)? Use $\alpha = 0.05$.
15. **Plato's Republic: Syllable Patterns** Prose rhythm is characterized by the occurrence of five-syllable sequences in long passages of text. This characterization may be used to assess the similarity among passages of text and sometimes the identity of authors. The following information is based on an article by D. Wishart and S. V. Leach appearing in *Computer Studies of the Humanities and Verbal Behavior* (Vol. 3, pp. 90–99). Syllables were categorized as long or short. On analyzing Plato's *Republic*, Wishart and Leach found that about 26.1% of the five-syllable sequences are of the type in which two are short and three are long. Suppose that Greek archaeologists have found an ancient manuscript dating back to Plato's time (about 427–347 B.C.). A random sample of 317 five-syllable sequences from the newly discovered manuscript showed that 61 are of the type two short and three long. Do the data indicate that the population proportion of this type of five-syllable sequence is different (either way) from the text of Plato's *Republic*? Use $\alpha = 0.01$.
16. **Plato's Dialogues: Prose Rhythm** *Symposium* is part of a larger work referred to as Plato's *Dialogues*. Wishart and Leach (see source in Problem 15) found that about 21.4% of five-syllable sequences in *Symposium* are of the type in which four are short and one is long. Suppose an antiquities store in Athens has a very old manuscript that the owner claims is part of Plato's *Dialogues*. A random sample of 493 five-syllable sequences from this manuscript showed that 136 were of the type four short and one long. Do the data indicate that the population proportion of this type of five-syllable sequence is higher than that found in Plato's *Symposium*? Use $\alpha = 0.01$.
17. **Consumers: Product Loyalty** USA Today reported that about 47% of the general consumer population in the United States is loyal to the automobile manufacturer of their choice. Suppose Chevrolet did a study of a random sample of 1006 Chevrolet owners and found that 490 said they would buy another Chevrolet. Does this indicate that the population proportion of consumers loyal to Chevrolet is more than 47%? Use $\alpha = 0.01$.

18. **Supermarket: Prices** *Harper's Index* reported that 80% of all supermarket prices end in the digit 9 or 5. Suppose you check a random sample of 115 items in a supermarket and find that 88 have prices that end in 9 or 5. Does this indicate that less than 80% of the prices in the store end in the digits 9 or 5? Use $\alpha = 0.05$.
19. **Medical: Hypertension** This problem is based on information taken from *The Merck Manual* (a reference manual used in most medical and nursing schools). Hypertension is defined as a blood pressure reading over 140 mm Hg systolic and/or over 90 mm Hg diastolic. Hypertension, if not corrected, can cause long-term health problems. In the college-age population (18–24 years), about 9.2% have hypertension. Suppose that a blood donor program is taking place in a college dormitory this week (final exams week). Before each student gives blood, the nurse takes a blood pressure reading. Of 196 donors, it is found that 29 have hypertension. Do these data indicate that the population proportion of students with hypertension during final exams week is higher than 9.2%? Use a 5% level of significance.
20. **Medical: Hypertension** Diltiazem is a commonly prescribed drug for hypertension (see source in Problem 19). However, diltiazem causes headaches in about 12% of patients using the drug. It is hypothesized that regular exercise might help reduce the headaches. If a random sample of 209 patients using diltiazem exercised regularly and only 16 had headaches, would this indicate a reduction in the population proportion of patients having headaches? Use a 1% level of significance.
21. **Myers–Briggs: Extroverts** Are most student government leaders extroverts? According to Myers–Briggs estimates, about 82% of college student government leaders are extroverts (Source: *Myers–Briggs Type Indicator Atlas of Type Tables*). Suppose that a Myers–Briggs personality preference test was given to a random sample of 73 student government leaders attending a large national leadership conference and that 56 were found to be extroverts. Does this indicate that the population proportion of extroverts among college student government leaders is different (either way) from 82%? Use $\alpha = 0.01$.
22. **Critical Region Method: Testing Proportions** Solve Problem 9 using the critical region method of testing. Since the sampling distribution of \hat{p} is the normal distribution, you can use critical values from the standard normal distribution as shown in Figure 8–8 or part (c) of Table 5, Appendix II. Compare your conclusions with the conclusions obtained by using the P -value method. Are they the same?
23. **Critical Region Method: Testing Proportions** Solve Problem 11 using the critical region method of testing. Hint: See Problem 22. Compare your conclusions with the conclusions obtained by using the P -value method. Are they the same?
24. **Critical Region Method: Testing Proportions** Solve Problem 17 using the critical region method of testing. Hint: See Problem 22. Compare your conclusions with the conclusions obtained by using the P -value method. Are they the same?

SECTION 8.4**Tests Involving Paired Differences (Dependent Samples)****FOCUS POINTS**

- Identify paired data and dependent samples.
- Explain the advantages of paired data tests.
- Compute differences and the sample test statistic.
- Estimate the P -value and conclude the test.

Paired data**Paired data****Creating data pairs**

Many statistical applications use *paired data* samples to draw conclusions about the difference between two population means. *Data pairs* occur very naturally in “before and after” situations, where the *same* object or item is measured both before and after a treatment. Applied problems in social science, natural science, and business administration frequently involve a study of matching pairs. Psychological studies of identical twins; biological studies of plant growth on plots of land matched for soil type, moisture, and sun exposure; and business studies on sales of matched inventories are examples of paired data studies.

When working with paired data, it is very important to have a definite and uniform method of creating data pairs that clearly utilizes a natural matching of characteristics. The next example and Guided Exercise demonstrate this feature.

EXAMPLE 8**PAIRED DATA**

A shoe manufacturer claims that among the general population of adults in the United States, the average length of the left foot is longer than that of the right. To compare the average length of the left foot with that of the right, we can take a random sample of 15 U.S. adults and measure the length of the left foot and then the length of the right foot for each person in the sample. Is there a natural way of pairing the measurements? How many pairs will we have?

SOLUTION: In this case, we can pair each left-foot measurement with the same person's right-foot measurement. The person serves as the “matching link” between the two distributions. We will have 15 pairs of measurements.

GUIDED EXERCISE 6**Paired data**

A psychologist has developed a series of exercises called the Instrumental Enrichment (IE) Program, which he claims is useful in overcoming cognitive deficiencies in mentally retarded children. To test the program, extensive statistical tests are being conducted. In one experiment, a random sample of 10-year-old students with IQ scores below 80 was selected. An IQ test was given to these students before they spent 2 years in an IE Program, and an IQ test was given to the same students after the program.

(a) On what basis can you pair the IQ scores?



Take the “before and after” IQ scores of each individual student.

(b) If there were 20 students in the sample, how many data pairs would you have?



Twenty data pairs. Note that there would be 40 IQ scores, but only 20 pairs.

COMMENT To compare two populations, we cannot always employ paired data tests, but when we can, what are the advantages? Using matched or paired data often can reduce the danger of introducing extraneous or uncontrollable factors into our sample measurements because the matched or paired data have essentially the *same* characteristics except for the *one* characteristic that is being measured. Furthermore, it can be shown that pairing data has the theoretical effect of reducing measurement variability (i.e., variance), which increases the accuracy of statistical conclusions.

When we wish to compare the means of two samples, the first item to be determined is whether or not there is a natural pairing between the data in the two samples. Again, data pairs are created from “before and after” situations, or from matching data by using studies of the same object, or by a process of taking measurements of closely matched items.

When testing *paired* data, we take the difference d of the data pairs *first* and look at the mean difference \bar{d} . Then we use a test on \bar{d} . Theorem 8.1 provides the basis for our work with paired data.

THEOREM 8.1 Consider a random sample of n data pairs. Suppose the differences d between the first and second members of each data pair are (approximately) normally distributed, with population mean μ_d . Then the t values

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

where \bar{d} is the sample mean of the d values, n is the number of data pairs, and

$$s_d = \sqrt{\frac{\sum(d - \bar{d})^2}{n - 1}}$$

is the sample standard deviation of the d values, follow a Student's t distribution with degrees of freedom $d.f. = n - 1$.

Hypotheses for testing the mean of paired differences

When testing the mean of the differences of paired data values, the null hypothesis is that there is no difference among the pairs. That is, the mean of the differences μ_d is zero.

$$H_0: \mu_d = 0$$

The alternate hypothesis depends on the problem and can be

$H_1: \mu_d < 0$ (left-tailed)	$H_1: \mu_d > 0$ (right-tailed)	$H_1: \mu_d \neq 0$ (two-tailed)
-----------------------------------	------------------------------------	-------------------------------------

Sample test statistic

For paired difference tests, we make our decision regarding H_0 according to the evidence of the sample mean \bar{d} of the differences of measurements. By Theorem 8.1, we convert the sample test statistic \bar{d} to a t value using the formula

$$t = \frac{\bar{d} - \mu_d}{(s_d / \sqrt{n})} \text{ with } d.f. = n - 1$$

where s_d = sample standard deviation of the differences d

n = number of data pairs

$\mu_d = 0$, as specified in H_0

P-values from Table 6 of Appendix II

To find the P -value (or an interval containing the P -value) corresponding to the test statistic t computed from \bar{d} , we use the Student's t distribution table

(Table 6, Appendix II). Recall from Section 8.2 that we find the test statistic t (or, if t is negative, $|t|$) in the row headed by $d.f. = n - 1$, where n is the number of data pairs. The P -value for the test statistic is the column entry in the *one-tail area* row for one-tailed tests (right or left). For two-tailed tests, the P -value is the column entry in the *two-tail area* row. Usually the exact test statistic t is not in the table, so we obtain an interval that contains the P -value by using adjacent entries in the table. Table 8-7 gives the basic structure for using the Student's t distribution table to find the P -value or an interval containing the P -value.

TABLE 8-7 Using Student's t Distribution Table for P -values

For one-tailed tests:	one-tail area	P -value	P -value
For two-tailed tests:	two-tail area	P -value	P -value
Use row header	$d.f. = n - 1$	↑ Find t value	

With the preceding information, you are now ready to test paired differences. First let's summarize the procedure.

PROCEDURE

HOW TO TEST PAIRED DIFFERENCES USING THE STUDENT'S t DISTRIBUTION

Requirements

Obtain a simple random sample of n matched data pairs A, B . Let d be a random variable representing the difference between the values in a matched data pair. Compute the sample mean \bar{d} and sample standard deviation s_d . If you can assume that d has a normal distribution or simply has a mound-shaped, symmetric distribution, then any sample size n will work. If you cannot assume this, then use a sample size $n \geq 30$.

Procedure

1. Use the *null hypothesis of no difference*, $H_0: \mu_d = 0$. In the context of the application, choose the *alternate hypothesis* to be $H_1: \mu_d > 0$, $H_1: \mu_d < 0$, or $H_1: \mu_d \neq 0$. Set the *level of significance* α .
2. Use \bar{d} , s_d , the sample size n , and $\mu_d = 0$ from the null hypothesis to compute the *sample test statistic*

$$t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d}\sqrt{n}}{s_d}$$

with degrees of freedom $d.f. = n - 1$.

3. Use the Student's t distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If P -value $\leq \alpha$, then reject H_0 . If P -value $> \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.



Problem 22 at the end of this section shows how to find confidence intervals for μ_d .

EXAMPLE 9

PAIRED DIFFERENCE TEST

A team of heart surgeons at Saint Ann's Hospital knows that many patients who undergo corrective heart surgery have a dangerous buildup of anxiety before their scheduled operations. The staff psychiatrist at the hospital has started a new



The Image Works

counseling program intended to reduce this anxiety. A test of anxiety is given to patients who know they must undergo heart surgery. Then each patient participates in a series of counseling sessions with the staff psychiatrist. At the end of the counseling sessions, each patient is retested to determine anxiety level. Table 8-8 indicates the results for a random sample of nine patients. Higher scores mean higher levels of anxiety. Assume the distribution of differences is mound-shaped and symmetric.

From the given data, can we conclude that the counseling sessions reduce anxiety? Use a 0.01 level of significance.

SOLUTION: Before we answer this question, let us notice two important points: (1) We have a *random sample* of nine patients, and (2) we have a *pair* of measurements taken on the same patient before and after counseling sessions. In our problem, the sample size is $n = 9$ pairs (i.e., patients), and the d values are found in the fourth column of Table 8-8.

(a) Note the level of significance and set the hypotheses.

In the problem statement, $\alpha = 0.01$. We want to test the claim that the counseling sessions reduce anxiety. This means that the anxiety level before counseling is expected to be higher than the anxiety level after counseling. In symbols, $d = B - A$ should tend to be positive, and the population mean of differences μ_d also should be positive. Therefore, we have

$$H_0: \mu_d = 0 \quad \text{and} \quad H_1: \mu_d > 0$$

(b) **Check Requirements** Is it appropriate to use a Student's t distribution for the sample test statistic? Explain. What degrees of freedom are used?

We have a random sample of paired differences d . Under the assumption that the d distribution is mound-shaped and symmetric, we use a Student's t distribution with $d.f. = n - 1 = 9 - 1 = 8$.

(c) Find the sample test statistic \bar{d} and convert it to a corresponding test statistic t . First we need to compute \bar{d} and s_d . Using formulas or a calculator and the d values shown in Table 8-8, we find that

$$\bar{d} \approx 33.33 \quad \text{and} \quad s_d \approx 22.92$$

TABLE 8-8

Patient	B Score Before Counseling	A Score After Counseling	$d = B - A$ Difference
Jan	121	76	45
Tom	93	93	0
Diane	105	64	41
Barbara	115	117	-2
Mike	130	82	48
Bill	98	80	18
Frank	142	79	63
Carol	118	67	51
Alice	125	89	36

Using these values together with $n = 9$ and $\mu_d = 0$, we have

$$t = \frac{\bar{d} - 0}{(s_d/\sqrt{n})} \approx \frac{33.33}{22.92/\sqrt{9}} \approx 4.363$$

(d) Find the P -value for the test statistic and sketch the P -value on the t distribution. Since we have a right-tailed test, the P -value is the area to the right

of $t = 4.363$, as shown in Figure 8-13. In Table 6 of Appendix II, we find an interval containing the P -value. Use entries from the row headed by $d.f. = n - 1 = 9 - 1 = 8$. The test statistic $t = 4.363$ falls between 3.355 and 5.041. The P -value for the sample t falls between the corresponding one-tail areas 0.005 and 0.0005. (See Table 8-9, excerpt from Table 6, Appendix II.)

$$0.0005 < P\text{-value} < 0.005$$

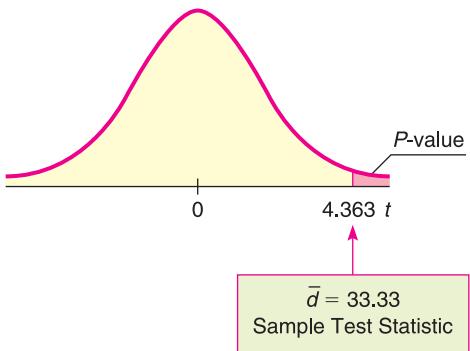
**TABLE 8-9 Excerpt from Student's t Distribution Table
(Table 6, Appendix II)**

✓ one-tail area	0.005	0.0005
two-tail area	0.010	0.0010
$d.f. = 8$	3.355	5.041

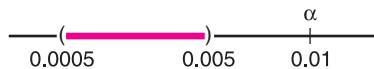
↑
Sample $t = 4.363$

FIGURE 8-13

P -value



(e) Conclude the test.



Since the interval containing the P -value lies to the left of $\alpha = 0.01$, we reject H_0 . Note: Using the raw data and software, P -value ≈ 0.0012 .

- (f) **Interpretation** Interpret the results in the context of the application.
At the 1% level of significance, we conclude that the counseling sessions reduce the average anxiety level of patients about to undergo corrective heart surgery.



The problem we have just solved is a paired difference problem of the “before and after” type. The next guided exercise demonstrates a paired difference problem of the “matched pair” type.

GUIDED EXERCISE 7

Paired difference test

Do educational toys make a difference in the age at which a child learns to read? To study this question, researchers designed an experiment in which one group of preschool children spent 2 hours each day (for 6 months) in a room well supplied with “educational” toys such as alphabet blocks, puzzles, ABC readers, coloring books featuring letters, and so forth. A control group of children spent 2 hours a day for 6 months in a “noneducational” toy room.

Continued

GUIDED EXERCISE 7 *continued*

It was anticipated that IQ differences and home environment might be uncontrollable factors unless identical twins could be used. Therefore, six pairs of identical twins of preschool age were randomly selected. From each pair, one member was randomly selected to participate in the experimental (i.e., educational toy room) group and the other in the control (i.e., noneducational toy room) group. For each twin, the data item recorded is the age in months at which the child began reading at the primary level (Table 8-10). Assume the distribution of differences is mound-shaped and symmetric.

TABLE 8-10 Reading Ages for Identical Twins (in Months)

Twin Pair	Experimental Group $B = \text{Reading Age}$	Control Group $A = \text{Reading Age}$	Difference $d = B - A$
1	58	60	-2
2	61	64	-3
3	53	52	1
4	60	65	-5
5	71	75	-4
6	62	63	-1

- (a) Compute the entries in the $d = B - A$ column of Table 8-10. Using formulas for the mean and sample standard deviation or a calculator with mean and sample standard deviation keys, compute \bar{d} and s_d .

Pair	$d = B - A$
1	-2
2	-3
3	1
4	-5
5	-4
6	-1

$$\bar{d} \approx -2.33$$

$$s_d \approx 2.16$$

- (b) What is the null hypothesis?
 (c) To test the claim that the experimental group learned to read at a *different age* (either younger or older), what should the alternate hypothesis be?
 (d) **Check Requirements** What distribution does the sample test statistic t follow? Find the degrees of freedom.

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

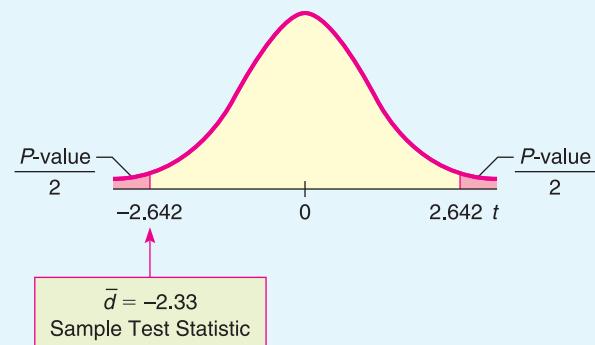
- (e) Convert the sample test statistic \bar{d} to a t value.
 (f) When we use Table 6 of Appendix II to find an interval containing the P -value, do we use one-tail or two-tail areas? Why? Sketch a figure showing the P -value. Find an interval containing the P -value.

Because the d distribution is mound-shaped and symmetric and we have a random sample of $n = 6$ paired differences, the sample test statistic follows a Student's t distribution with $d.f. = n - 1 = 6 - 1 = 5$.

Using $\mu_d = 0$ from H_0 , $\bar{d} = -2.33$, $n = 6$ and $s_d = 2.16$, we get

$$t = \frac{\bar{d} - \mu_d}{(s_d/\sqrt{n})} \approx \frac{-2.33 - 0}{(2.16/\sqrt{6})} \approx -2.642$$

This is a two-tailed test, so we use two-tail areas.

FIGURE 8-14 P -valueTABLE 8-11 Excerpt from Student's t Table

one-tail area	0.025	0.010
✓ two-tail area	0.050	0.020
$d.f. = 5$	2.571	3.365
	↑ Sample $t = 2.642$	

Continued

GUIDED EXERCISE 7 *continued*(g) Using $\alpha = 0.05$, do we reject or fail to reject H_0 ? The sample t is between 2.571 and 3.365. $0.020 < P\text{-value} < 0.050$ *Note:* Using the raw data and software,
 $P\text{-value} \approx 0.0457$.(h) **Interpretation** What do the results mean in the context of this application? Since the interval containing the P -value has values that are all smaller than 0.05, we reject H_0 .(h) **Interpretation** What do the results mean in the context of this application?

At the 5% level of significance, the experiment indicates that educational toys make a difference in the age at which a child learns to read.

LOOKING FORWARD

The test for paired differences used in this section is called a *parametric test*. Such tests usually require certain assumptions, such as a normal distribution or a large sample size. In Section 11.1, you will find the sign test for matched pairs. This is a *nonparametric test*. Such tests are useful when you cannot make assumptions about the population distribution. The disadvantage of nonparametric tests is that they tend to accept the null hypothesis more often than they should. That is, they are less sensitive tests. Which type of test should you use? If it is reasonable to assume that the underlying population is normal (or at least mound-shaped and symmetric), or if you have a large sample size, then use the more powerful parametric test described in this section.



TECH NOTES

Both Excel 2007 and Minitab support paired difference tests directly. On the TI-84Plus/TI-83Plus/TI-nspire calculators, construct a column of differences and then do a t test on the data in that column. For each technology, be sure to relate the alternate hypothesis to the “before and after” assignments. All the displays show the results for the data of Guided Exercise 7.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Enter the “before” data in column L1 and the “after” data in column L2. Highlight L3, type L1 – L2, and press Enter. The column L3 now contains the $B - A$ differences. To conduct the test, press STAT, select TESTS, and use option 2:T-Test. Note that the letter x is used in place of d .

L1	L2	L3	3
58	60	-2	
61	64	-3	
53	52	1	
60	65	-5	
71	75	-4	
62	63	-1	
L3(7) =			

```
T-Test
μ≠0
t=-2.645751311
p=.0456591238
̄x=-2.333333333
Sx=2.160246899
n=6
```

Excel 2007 Enter the data in two columns. On the home ribbon, click the Data tab. In the Analysis Group, select Data Analysis. In the dialogue box, select t-Test: Paired Two-Sample for Means. Fill in the dialogue box with the hypothesized mean difference of 0. Set alpha.

B	C	D
t-Test: Paired Two Sample for Means		
	Variable 1	Variable 2
Mean	60.83333	63.16666667
Variance	34.96667	55.76666667
Observations	6	6
Pearson Correlation	0.974519	
Hypothesized Mean Difference	0	
df	5	
t Stat	-2.64575	
P(T<=t) one-tail	0.02283	
t Critical one-tail	2.015049	
P(T<=t) two-tail	0.045659	
t Critical two-tail	2.570578	

P-value →

Minitab Enter the data in two columns. Use the menu selection **Stat ▶ Basic Statistics ▶ Paired t**. Under Options, set the null and alternate hypotheses.

Paired T-Test and Confidence Interval

Paired T for B - A

	N	Mean	StDevSE	Mean
B	6	60.83	5.91	2.41
A	6	63.17	7.47	3.05
Difference	6	-2.333	2.160	0.882
95% CI for mean difference: (-4.601, -0.066)				
T-Test of mean difference = 0 (vs not = 0):				
T-Value = -2.65 P-Value = 0.046				

EXAMPLE 10

CRITICAL REGION METHOD

Let's revisit Guided Exercise 7 regarding educational toys and reading age and conclude the test using the critical region method. Recall that there were six pairs of twins. One twin of each set was given educational toys and the other was not. The difference d in reading ages for each pair of twins was measured, and $\alpha = 0.05$.

SOLUTION: From Guided Exercise 7, we have

$$H_0: \mu_d = 0 \quad \text{and} \quad H_1: \mu_d \neq 0$$

We computed $\bar{d} \approx -2.33$ with corresponding sample test statistic $t \approx -2.642$.

- (a) Find the critical values for $\alpha = 0.05$.

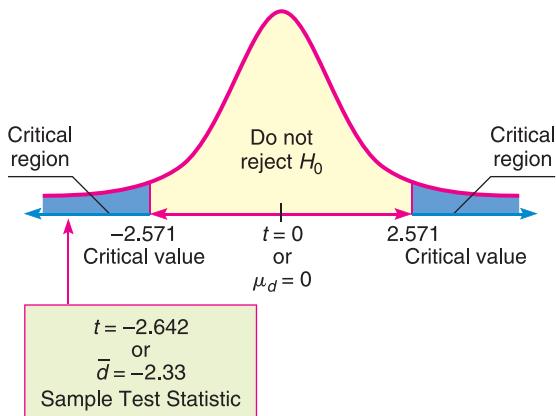
Since the number of pairs is $n = 6$, $d.f. = n - 1 = 5$. In the Student's t distribution table (Table 6, Appendix II), look in the row headed by 5. To find the column, locate $\alpha = 0.05$ in the *two-tail area* row, since we have a two-tailed test. The critical values are $\pm t_0 = \pm 2.571$.

- (b) Sketch the critical regions and place the t value of the sample test statistic \bar{d} on the sketch. Conclude the test. Compare the result to the result given by the *P-value* method of Guided Exercise 7.

Since the sample test statistic falls in the critical region (see Figure 8-15), we reject H_0 at the 5% level of significance. At this level, educational toys seem to make a difference in reading age. Notice that this conclusion is consistent with the conclusion obtained using the *P-value*.

FIGURE 8-15

Critical Region, with $\alpha = 0.05$,
 $d.f. = 5$



VIEWPOINT

DUI

DUI usually means “driving under the influence” of alcohol, but driving under the influence of sleep loss can be just as dangerous. Researchers in Australia have found that after staying awake for 24 hours straight, a person will be about as impaired as if he or she had had enough alcohol to be legally drunk in most U.S. states (Source: Rocky Mountain News). Using driver-simulation exams and statistical tests (paired difference tests) found in this section, it is possible to show that the null hypothesis $H_0: \mu_d = 0$ cannot be rejected. Or, put another way, the average level of impairment for a given individual from alcohol (at the DUI level) is about the same as the average level of impairment from sleep loss (24 hours without sleep).

SECTION 8.4 PROBLEMS

- Statistical Literacy** Are data that can be paired independent or dependent?
- Statistical Literacy** Consider a set of data pairs. What is the first step in processing the data for a paired differences test? What is the formula for the sample test statistic t ? Describe each symbol used in the formula.
- Statistical Literacy** When testing the difference of means for paired data, what is the null hypothesis?
- Statistical Literacy** When conducting a paired differences test, what is the value of n ?
- Statistical Literacy** When using a Student’s t distribution for a paired differences test with n data pairs, what value do you use for the degrees of freedom?
- Critical Thinking** Alisha is conducting a paired differences test for a “before (B score) and after (A score)” situation. She is interested in testing whether the average of the “before” scores is higher than that of the “after” scores.
 - To use a right-tailed test, how should Alisha construct the differences between the “before” and “after” scores?
 - To use a left-tailed test, how should she construct the differences between the “before” and “after” scores?
- Basic Computation: Paired Differences Test** For a random sample of 36 data pairs, the sample mean of the differences was 0.8. The sample standard deviation of the differences was 2. At the 5% level of significance, test the claim that the population mean of the differences is different from 0.
 - Check Requirements** Is it appropriate to use a Student’s t distribution for the sample test statistic? Explain. What degrees of freedom are used?
 - State the hypotheses.

- (c) Compute the sample test statistic.
 (d) Estimate the P -value of the sample test statistic.
 (e) Do we reject or fail to reject the null hypothesis? Explain.
 (f) **Interpretation** What do your results tell you?
8. **Basic Computation: Paired Differences Test** For a random sample of 20 data pairs, the sample mean of the differences was 2. The sample standard deviation of the differences was 5. Assume that the distribution of the differences is mound-shaped and symmetric. At the 1% level of significance, test the claim that the population mean of the differences is positive.
- (a) **Check Requirements** Is it appropriate to use a Student's t distribution for the sample test statistic? Explain. What degrees of freedom are used?
 (b) State the hypotheses.
 (c) Compute the sample test statistic.
 (d) Estimate the P -value of the sample test statistic.
 (e) Do we reject or fail to reject the null hypothesis? Explain.
 (f) **Interpretation** What do your results tell you?

For Problems 9–21 assume that the distribution of differences d is mound-shaped and symmetric.

Please provide the following information for Problems 9–21.

- (a) What is the level of significance? State the null and alternate hypotheses. Will you use a left-tailed, right-tailed, or two-tailed test?
 (b) **Check Requirements** What sampling distribution will you use? What assumptions are you making? Compute the value of the sample test statistic.
 (c) Find (or estimate) the P -value. Sketch the sampling distribution and show the area corresponding to the P -value.
 (d) Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis? Are the data statistically significant at level α ?
 (e) **Interpret** your conclusion in the context of the application.

In these problems, assume that the distribution of differences is approximately normal.

Note: For degrees of freedom $d.f.$ not in the Student's t table, use the closest $d.f.$ that is *smaller*. In some situations, this choice of $d.f.$ may increase the P -value by a small amount and therefore produce a slightly more "conservative" answer.

9. **Business: CEO Raises** Are America's top chief executive officers (CEOs) really worth all that money? One way to answer this question is to look at row *B*, the annual company percentage increase in revenue, versus row *A*, the CEO's annual percentage salary increase in that same company (Source: *Forbes*, Vol. 159, No. 10). A random sample of companies such as John Deere & Co., General Electric, Union Carbide, and Dow Chemical yielded the following data:

<i>B: Percent increase for company</i>	24	23	25	18	6	4	21	37
<i>A: Percent increase for CEO</i>	21	25	20	14	-4	19	15	30

Do these data indicate that the population mean percentage increase in corporate revenue (row *B*) is different from the population mean percentage increase in CEO salary? Use a 5% level of significance.

10. **Fishing: Shore or Boat?** Is fishing better from a boat or from the shore? Pyramid Lake is located on the Paiute Indian Reservation in Nevada. Presidents, movie stars, and people who just want to catch fish go to Pyramid Lake for really large cutthroat trout. Let row *B* represent hours per fish caught fishing from the shore, and let row *A* represent hours per fish caught using a boat. The following data

are paired by month from October through April (Source: *Pyramid Lake Fisheries*, Paiute Reservation, Nevada).

	Oct.	Nov.	Dec.	Jan.	Feb.	March	April
B: Shore	1.6	1.8	2.0	3.2	3.9	3.6	3.3
A: Boat	1.5	1.4	1.6	2.2	3.3	3.0	3.8

Use a 1% level of significance to test if there is a difference in the population mean hours per fish caught using a boat compared with fishing from the shore.

11. **Ecology: Rocky Mountain National Park** The following is based on information taken from *Winter Wind Studies in Rocky Mountain National Park* by D. E. Glidden (Rocky Mountain Nature Association). At five weather stations on Trail Ridge Road in Rocky Mountain National Park, the peak wind gusts (in miles per hour) for January and April are recorded below.

Weather Station	1	2	3	4	5
January	139	122	126	64	78
April	104	113	100	88	61

Does this information indicate that the peak wind gusts are higher in January than in April? Use $\alpha = 0.01$.

12. **Wildlife: Highways** The western United States has a number of four-lane interstate highways that cut through long tracts of wilderness. To prevent car accidents with wild animals, the highways are bordered on both sides with 12-foot-high woven wire fences. Although the fences prevent accidents, they also disturb the winter migration pattern of many animals. To compensate for this disturbance, the highways have frequent wilderness underpasses designed for exclusive use by deer, elk, and other animals.

In Colorado, there is a large group of deer that spend their summer months in a region on one side of a highway and survive the winter months in a lower region on the other side. To determine if the highway has disturbed deer migration to the winter feeding area, the following data were gathered on a random sample of 10 wilderness districts in the winter feeding area. Row B represents the average January deer count for a 5-year period before the highway was built, and row A represents the average January deer count for a 5-year period after the highway was built. The highway department claims that the January population has not changed. Test this claim against the claim that the January population has dropped. Use a 5% level of significance. Units used in the table are hundreds of deer.

Wilderness District	1	2	3	4	5	6	7	8	9	10
B: Before highway	10.3	7.2	12.9	5.8	17.4	9.9	20.5	16.2	18.9	11.6
A: After highway	9.1	8.4	10.0	4.1	4.0	7.1	15.2	8.3	12.2	7.3

13. **Wildlife: Wolves** In environmental studies, sex ratios are of great importance. Wolf society, packs, and ecology have been studied extensively at different locations in the United States and foreign countries. Sex ratios for eight study sites in northern Europe are shown below (based on *The Wolf* by L. D. Mech, University of Minnesota Press).

Gender Study of Large Wolf Packs

Location of Wolf Pack	% Males (Winter)	% Males (Summer)
Finland	72	53
Finland	47	51
Finland	89	72
Lapland	55	48
Lapland	64	55
Russia	50	50
Russia	41	50
Russia	55	45

It is hypothesized that in winter, “loner” males (not present in summer packs) join the pack to increase survival rate. Use a 5% level of significance to test the claim that the average percentage of males in a wolf pack is higher in winter.

14. **Demographics: Birthrate and Death Rate** In the following data pairs, A represents birthrate and B represents death rate per 1000 resident population. The data are paired by counties in the Midwest. A random sample of 16 counties gave the following information (Reference: *County and City Data Book*, U.S. Department of Commerce).

A:	12.7	13.4	12.8	12.1	11.6	11.1	14.2	15.1
B:	9.8	14.5	10.7	14.2	13.0	12.9	10.9	10.0

A:	12.5	12.3	13.1	15.8	10.3	12.7	11.1	15.7
B:	14.1	13.6	9.1	10.2	17.9	11.8	7.0	9.2

Do the data indicate a difference (either way) between population average birthrate and death rate in this region? Use $\alpha = 0.01$.

15. **Navajo Reservation: Hogans** The following data are based on information taken from the book *Navajo Architecture: Forms, History, Distributions* by S. C. Jett and V. E. Spencer (University of Arizona Press). A survey of houses and traditional hogans was made in a number of different regions of the modern Navajo Indian Reservation. The following table is the result of a random sample of eight regions on the Navajo Reservation.

Area on Navajo Reservation	Number of Inhabited Houses	Number of Inhabited Hogans
Bitter Springs	18	13
Rainbow Lodge	16	14
Kayenta	68	46
Red Mesa	9	32
Black Mesa	11	15
Canyon de Chelly	28	47
Cedar Point	50	17
Burnt Water	50	18

Does this information indicate that the population mean number of inhabited houses is greater than that of hogans on the Navajo Reservation? Use a 5% level of significance.

16. **Archaeology: Stone Tools** The following is based on information taken from *Bandelier Archaeological Excavation Project: Summer 1990 Excavations at Burnt Mesa Pueblo and Casa del Rito*, edited by T. A. Kohler (Washington State University, Department of Anthropology). The artifact frequency for an excavation of a kiva in Bandelier National Monument gave the following information.

Stratum	Flaked Stone Tools	Nonflaked Stone Tools
1	7	3
2	3	2
3	10	1
4	1	3
5	4	7
6	38	32
7	51	30
8	25	12

Does this information indicate that there tend to be more flaked stone tools than nonflaked stone tools at this excavation site? Use a 5% level of significance.

17. **Economics: Cost of Living Index** In the following data pairs, A represents the cost of living index for housing and B represents the cost of living index for groceries. The data are paired by metropolitan areas in the United States. A random sample of 36 metropolitan areas gave the following information (Reference: *Statistical Abstract of the United States*, 121st Edition).

A:	132	109	128	122	100	96	100	131	97
B:	125	118	139	104	103	107	109	117	105
A:	120	115	98	111	93	97	111	110	92
B:	110	109	105	109	104	102	100	106	103
A:	85	109	123	115	107	96	108	104	128
B:	98	102	100	95	93	98	93	90	108
A:	121	85	91	115	114	86	115	90	113
B:	102	96	92	108	117	109	107	100	95

- i. Let d be the random variable $d = A - B$. Use a calculator to verify that $\bar{d} \approx 2.472$ and $s_d \approx 12.124$.
- ii. Do the data indicate that the U.S. population mean cost of living index for housing is higher than that for groceries in these areas? Use $\alpha = 0.05$.

18. **Economics: Cost of Living Index** In the following data pairs, A represents the cost of living index for utilities and B represents the cost of living index for transportation. The data are paired by metropolitan areas in the United States. A random sample of 46 metropolitan areas gave the following information (Reference: *Statistical Abstract of the United States*, 121st Edition).

A:	90	84	85	106	83	101	89	125	105
B:	100	91	103	103	109	109	94	114	113
A:	118	133	104	84	80	77	90	92	90
B:	120	130	117	109	107	104	104	113	101

A:	106	95	110	112	105	93	119	99	109
B:	96	109	103	107	103	102	101	86	94
<hr/>									
A:	109	113	90	121	120	85	91	91	97
B:	88	100	104	119	116	104	121	108	86
<hr/>									
A:	95	115	99	86	88	106	80	108	90
B:	100	83	88	103	94	125	115	100	96

- i. Let d be the random variable $d = A - B$. Use a calculator to verify that $\bar{d} \approx -5.739$ and $s_d \approx 15.910$.

ii. Do the data indicate that the U.S. population mean cost of living index for utilities is less than that for transportation in these areas? Use $\alpha = 0.05$.

19. **Golf: Tournaments** Do professional golfers play better in their first round? Let row B represent the score in the fourth (and final) round, and let row A represent the score in the first round of a professional golf tournament. A random sample of finalists in the British Open gave the following data for their first and last rounds in the tournament (Source: *Golf Almanac*).

B: Last	73	68	73	71	71	72	68	68	74
A: First	66	70	64	71	65	71	71	71	71

Do the data indicate that the population mean score on the last round is higher than that on the first? Use a 5% level of significance.

20. ***Psychology: Training Rats*** The following data are based on information from the Regis University Psychology Department. In an effort to determine if rats perform certain tasks more quickly if offered larger rewards, the following experiment was performed. On day 1, a group of three rats was given a reward of one food pellet each time they ran a maze. A second group of three rats was given a reward of five food pellets each time they ran the maze. On day 2, the groups were reversed, so the first group now got five food pellets for running the maze and the second group got only one pellet for running the same maze. The average times in seconds for each rat to run the maze 30 times are shown in the following table.

Rat	A	B	C	D	E	F
Time with one food pellet	3.6	4.2	2.9	3.1	3.5	3.9
Time with five food pellets	3.0	3.7	3.0	3.3	2.8	3.0

Do these data indicate that rats receiving larger rewards tend to run the maze in less time? Use a 5% level of significance.

21. **Psychology: Training Rats** The same experimental design discussed in Problem 20 was used to test rats trained to climb a sequence of short ladders. Times in seconds for eight rats to perform this task are shown in the following table.

Rat	A	B	C	D	E	F	G	H
Time 1 pellet	12.5	13.7	11.4	12.1	11.0	10.4	14.6	12.3
Time 5 pellets	11.1	12.0	12.2	10.6	11.5	10.5	12.9	11.0

Do these data indicate that rats receiving larger rewards tend to climb the ladder in less time? Use a 5% level of significance.



22. ***Expand Your Knowledge: Confidence Intervals for μ_d*** Using techniques from Section 7.2, we can find a confidence interval for μ_d . Consider a random sample of n matched data pairs A, B . Let $d = B - A$ be a random variable representing the difference between the values in a matched data pair. Compute the sample mean \bar{d} of the differences and the sample standard deviation s_d . If d has a normal distribution or is mound-shaped, or if $n \geq 30$, then a confidence interval for μ_d is

$$\bar{d} - E < \mu_d < \bar{d} + E$$

$$\text{where } E = t_c \frac{s_d}{\sqrt{n}}$$

c = confidence level ($0 < c < 1$)

t_c = critical value for confidence level c and $d.f. = n - 1$

- (a) Using the data of Problem 9, find a 95% confidence interval for the mean difference between percentage increase in company revenue and percentage increase in CEO salary.
- (b) Use the confidence interval method of hypothesis testing outlined in Problem 25 of Section 8.2 to test the hypothesis that population mean percentage increase in company revenue is different from that of CEO salary. Use a 5% level of significance.
23. ***Critical Region Method: Student's t*** Solve Problem 9 using the critical region method of testing. Compare your conclusions with the conclusion obtained by using the P -value method. Are they the same?
24. ***Critical Region Method: Student's t*** Solve Problem 11 using the critical region method of testing. Compare your conclusions with the conclusion obtained by using the P -value method. Are they the same?

SECTION 8.5

Testing $\mu_1 - \mu_2$ and $p_1 - p_2$ (Independent Samples)

FOCUS POINTS

- Identify independent samples and sampling distributions.
- Compute the sample test statistic and P -value for $\mu_1 - \mu_2$ and conclude the test.
- Compute the sample test statistic and P -value for $p_1 - p_2$ and conclude the test.

Independent Samples

Many practical applications of statistics involve a comparison of two population means or two population proportions. In Section 8.4, we considered tests of differences of means for *dependent samples*. With dependent samples, we could pair the data and then consider the difference of data measurements d . In this section, we will turn our attention to tests of differences of means from *independent samples*. We will see new techniques for testing the difference of means from independent samples.

First, let's consider independent samples. We say that two sampling distributions are *independent* if there is no relationship whatsoever between specific values of the two distributions.

Independent Samples

EXAMPLE 11**INDEPENDENT SAMPLE**

A teacher wishes to compare the effectiveness of two teaching methods. Students are randomly divided into two groups: The first group is taught by method 1; the second group, by method 2. At the end of the course, a comprehensive exam is given to all students, and the mean score \bar{x}_1 for group 1 is compared with the mean score \bar{x}_2 for group 2. Are the samples independent or dependent?

SOLUTION: Because the students were *randomly* divided into two groups, it is reasonable to say that the \bar{x}_1 distribution is independent of the \bar{x}_2 distribution.

EXAMPLE 12**DEPENDENT SAMPLE**

In Section 8.4, we considered a situation in which a shoe manufacturer claimed that for the general population of adult U.S. citizens, the average length of the left foot is longer than the average length of the right foot. To study this claim, the manufacturer gathers data in this fashion: Sixty adult U.S. citizens are drawn at random, and for these 60 people, both their left and right feet are measured. Let \bar{x}_1 be the mean length of the left feet and \bar{x}_2 be the mean length of the right feet. Are the \bar{x}_1 and \bar{x}_2 distributions independent for this method of collecting data?

SOLUTION: In this method, there is only *one* random sample of people drawn, and both the left and right feet are measured from this sample. The length of a person's left foot is usually related to the length of the person's right foot, so in this case the \bar{x}_1 and \bar{x}_2 distributions are *not* independent. In fact, we could pair the data and consider the distribution of the differences, left foot length minus right foot length. Then we would use the techniques of paired difference tests as found in Section 8.4.

GUIDED EXERCISE 8**Independent sample**

Suppose the shoe manufacturer of Example 12 gathers data in the following way: Sixty adult U.S. citizens are drawn at random and their left feet are measured; then another 60 adult U.S. citizens are drawn at random and their right feet are measured. Again, \bar{x}_1 is the mean of the left foot measurements and \bar{x}_2 is the mean of the right foot measurements.

Are the \bar{x}_1 and \bar{x}_2 distributions independent for this method of collecting data?



For this method of gathering data, two random samples are drawn: one for the left-foot measurements and one for the right-foot measurements. The first sample is not related to the second sample. The \bar{x}_1 and \bar{x}_2 distributions are independent.

Part I: Testing $\mu_1 - \mu_2$ When σ_1 and σ_2 Are Known**Properties of $\bar{x}_1 - \bar{x}_2$ distribution,
 σ_1 and σ_2 known**

In this part, we will use distributions that arise from a difference of means from independent samples. How do we obtain such distributions? If we have two statistical variables x_1 and x_2 , each with its own distribution, we take independent random samples of size n_1 from the x_1 distribution and of size n_2 from the x_2 distribution. Then we can compute the respective means \bar{x}_1 and \bar{x}_2 . Consider the difference $\bar{x}_1 - \bar{x}_2$. This represents a difference of means. If we repeat the sampling process over and over, we will come up with lots of $\bar{x}_1 - \bar{x}_2$ values. These values can be arranged in a frequency table, and we can make a histogram for the

distribution of $\bar{x}_1 - \bar{x}_2$ values. This will give us an experimental idea of the theoretical distribution of $\bar{x}_1 - \bar{x}_2$.

Fortunately, it is not necessary to carry out this lengthy process for each example. The results have already been worked out mathematically. The next theorem presents the main results.

THEOREM 8.2 Let x_1 have a normal distribution with mean μ_1 and standard deviation σ_1 . Let x_2 have a normal distribution with mean μ_2 and standard deviation σ_2 . If we take independent random samples of size n_1 from the x_1 distribution and of size n_2 from the x_2 distribution, then the variable $\bar{x}_1 - \bar{x}_2$ has

1. A normal distribution
2. Mean $\mu_1 - \mu_2$
3. Standard deviation

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

COMMENT Theorem 8.2 requires that x_1 and x_2 have normal distributions. However, if both n_1 and n_2 are 30 or larger, then for most practical applications, the central limit theorem assures us that \bar{x}_1 and \bar{x}_2 are approximately normally distributed. In this case, the conclusions of the theorem are again valid, even if the original x_1 and x_2 distributions are not normal.

Hypotheses for testing difference of means

When testing the difference of means, it is customary to use the null hypothesis:

$$H_0: \mu_1 - \mu_2 = 0 \text{ or, equivalently, } H_0: \mu_1 = \mu_2$$

As mentioned in Section 8.1, the null hypothesis is set up to see if it can be rejected. When testing the difference of means, we first set up the hypothesis H_0 that there is no difference. The alternate hypothesis could then be any of the ones listed in Table 8-12. The alternate hypothesis and consequent type of test used depend on the particular problem. Note that μ_1 is always listed first.

TABLE 8-12 Alternate Hypotheses and Type of Test: Difference of Two Means

	H_1	Type of Test
$H_1: \mu_1 - \mu_2 < 0$	or equivalently	Left-tailed test
$H_1: \mu_1 - \mu_2 > 0$	or equivalently	Right-tailed test
$H_1: \mu_1 - \mu_2 \neq 0$	or equivalently	Two-tailed test

Using Theorem 8.2 and the central limit theorem (Section 6.5), we can summarize the procedure for testing $\mu_1 - \mu_2$ when both σ_1 and σ_2 are known.

PROCEDURE

HOW TO TEST $\mu_1 - \mu_2$ WHEN BOTH σ_1 AND σ_2 ARE KNOWN

Requirements

Let σ_1 and σ_2 be the known population standard deviations of populations 1 and 2. Obtain two independent random samples from populations 1 and 2, where

\bar{x}_1 and \bar{x}_2 are sample means from populations 1 and 2

n_1 and n_2 are sample sizes from populations 1 and 2

If you can assume that both population distributions 1 and 2 are normal, any sample sizes n_1 and n_2 will work. If you cannot assume this, then use sample sizes $n_1 \geq 30$ and $n_2 \geq 30$.

Procedure

1. In the context of the application, state the *null and alternate hypotheses* and set the *level of significance* α . It is customary to use $H_0: \mu_1 - \mu_2 = 0$.

2. Use $\mu_1 - \mu_2 = 0$ from the null hypothesis together with $\bar{x}_1, \bar{x}_2, \sigma_1, \sigma_2, n_1$, and n_2 to compute the *sample test statistic*.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

3. Use the standard normal distribution and the type of test, one-tailed or two-tailed, to find the *P-value* corresponding to the sample test statistic.
4. *Conclude the test.* If $P\text{-value} \leq \alpha$, then reject H_0 . If $P\text{-value} > \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.

EXAMPLE 13

Bob Winsett/CORBIS

TESTING THE DIFFERENCE OF MEANS (σ_1 AND σ_2 KNOWN)

A consumer group is testing camp stoves. To test the heating capacity of a stove, it measures the time required to bring 2 quarts of water from 50° F to boiling (at sea level). Two competing models are under consideration. Ten stoves of the first model and 12 stoves of the second model are tested. The following results are obtained.

Model 1: Mean time $\bar{x}_1 = 11.4$ min; $\sigma_1 = 2.5$ min; $n_1 = 10$

Model 2: Mean time $\bar{x}_2 = 9.9$ min; $\sigma_2 = 3.0$ min; $n_2 = 12$

Assume that the time required to bring water to a boil is normally distributed for each stove. Is there any difference (either way) between the performances of these two models? Use a 5% level of significance.

SOLUTION:

- (a) State the null and alternate hypotheses and note the value of α .

Let μ_1 and μ_2 be the means of the distributions of times for models 1 and 2, respectively. We set up the null hypothesis to state that there is no difference:

$$H_0: \mu_1 = \mu_2 \quad \text{or} \quad H_0: \mu_1 - \mu_2 = 0$$

The alternate hypothesis states that there is a difference:

$$H_1: \mu_1 \neq \mu_2 \quad \text{or} \quad H_1: \mu_1 - \mu_2 \neq 0$$

The level of significance is $\alpha = 0.05$.

- (b) **Check Requirements** What distribution does the sample test statistic follow?

The sample test statistic z follows a standard normal distribution because the original distributions from which the samples are drawn are normal. In addition, the population standard deviations of the original distributions are known and the samples are independent.

- (c) Compute $\bar{x}_1 - \bar{x}_2$ and then convert it to the sample test statistic z . We are given the values $\bar{x}_1 = 11.4$ and $\bar{x}_2 = 9.9$. Therefore, $\bar{x}_1 - \bar{x}_2 = 11.4 - 9.9 = 1.5$. To convert this to a z value, we use the values $\sigma_1 = 2.5$, $\sigma_2 = 3.0$, $n_1 = 10$, and $n_2 = 12$. From the null hypothesis, $\mu_1 - \mu_2 = 0$

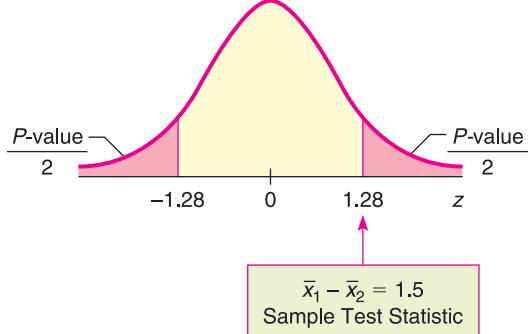
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{1.5}{\sqrt{\frac{2.5^2}{10} + \frac{3.0^2}{12}}} \approx 1.28$$

- (d) Find the P -value and sketch the area on the standard normal curve. Figure 8-16 shows the P -value. Use the standard normal distribution (Table 5 of Appendix II) and the fact that we have a two-tailed test.

$$P\text{-value} \approx 2(0.1003) = 0.2006$$

FIGURE 8-16

P -value



- (e) Conclude the test.

The P -value is 0.2006 and $\alpha = 0.05$. Since $P\text{-value} > \alpha$, do not reject H_0 .

- (f) **Interpretation** Interpret the results.

At the 5% level of significance, the sample data do not indicate any difference in the population mean times for boiling water for the two stove models.



GUIDED EXERCISE 9

Testing the difference of means (σ_1 and σ_2 known)

Let us return to Example 11 at the beginning of this section. A teacher wishes to compare the effectiveness of two teaching methods for her students. Students are randomly divided into two groups. The first group is taught by method 1; the second group, by method 2. At the end of the course, a comprehensive exam is given to all students.

The first group consists of $n_1 = 49$ students, with a mean score of $\bar{x}_1 = 74.8$ points. The second group has $n_2 = 50$ students, with a mean score of $\bar{x}_2 = 81.3$ points. The teacher claims that the second method will increase the mean score on the comprehensive exam. Is this claim justified at the 5% level of significance? Earlier research for the two methods indicates that $\sigma_1 = 14$ points and $\sigma_2 = 15$ points.

Let μ_1 and μ_2 be the mean scores of the distribution of all scores using method 1 and method 2, respectively.

- (a) What is the null hypothesis? $\rightarrow H_0: \mu_1 = \mu_2$ or $H_0: \mu_1 - \mu_2 = 0$
- (b) **Check Requirements** What distribution does the sample test statistic follow? Explain. \rightarrow The sample test statistic follows a standard normal distribution because the samples are large, both σ_1 and σ_2 are known, and the samples are independent.
- (c) To examine the validity of the teacher's claim, what should the alternate hypothesis be? What is α ? $\rightarrow H_1: \mu_1 < \mu_2$ (the second method gives a higher average score) or $H_1: \mu_1 - \mu_2 < 0$. $\alpha = 0.05$
- (d) Compute $\bar{x}_1 - \bar{x}_2$. $\rightarrow \bar{x}_1 - \bar{x}_2 = 74.8 - 81.3 = -6.5$
- (e) Convert $\bar{x}_1 - \bar{x}_2 = -6.5$ to the sample test statistic z . \rightarrow Using $\sigma_1 = 14$, $\sigma_2 = 15$, $n_1 = 49$, $n_2 = 50$, and $\mu_1 - \mu_2 = 0$ from H_0 , we have

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{-6.5 - 0}{\sqrt{\frac{14^2}{49} + \frac{15^2}{50}}} \approx -2.23$$

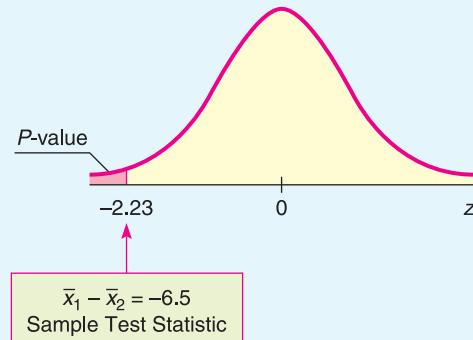
Continued

GUIDED EXERCISE 9 *continued*

- (f) Find the P -value and sketch the area on the standard normal curve.



Figure 8-17 shows the P -value. It is the area to the left of $z = -2.23$. Using Table 5 of Appendix II, we find $P\text{-value} = P(z < -2.23) \approx 0.0129$.

FIGURE 8-17 P -value

- (g) Conclude the test.



Since P -value of $0.0129 \leq 0.05$ for α , reject H_0 .

- (h) *Interpret* the results in the context of the application.



At the 5% level of significance, there is sufficient evidence to show that the second teaching method increased the population mean score on the exam.

Part II: Testing $\mu_1 - \mu_2$ When σ_1 and σ_2 Are Unknown

To test $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown, we use distribution methods similar to those in Chapter 7 for estimating $\mu_1 - \mu_2$. In particular, if the two distributions are normal or approximately mound-shaped, or if both sample sizes are large (≥ 30), we use a Student's t distribution. Let's summarize the method of testing $\mu_1 - \mu_2$.

PROCEDURE

HOW TO TEST $\mu_1 - \mu_2$ WHEN σ_1 AND σ_2 ARE UNKNOWN

Requirements

Obtain two independent random samples from populations 1 and 2, where

\bar{x}_1 and \bar{x}_2 are sample means from populations 1 and 2

s_1 and s_2 are sample standard deviations from populations 1 and 2

n_1 and n_2 are sample sizes from populations 1 and 2

If you can assume that both population distributions 1 and 2 are normal or at least mound-shaped and symmetric, then any sample sizes n_1 and n_2 will work. If you cannot assume this, then use sample sizes $n_1 \geq 30$ and $n_2 \geq 30$.

Procedure

- In the context of the application, state the *null and alternate hypotheses* and set the *level of significance* α . It is customary to use $H_0: \mu_1 - \mu_2 = 0$.
- Use $\mu_1 - \mu_2 = 0$ from the null hypothesis together with $\bar{x}_1, \bar{x}_2, s_1, s_2, n_1$, and n_2 to compute the *sample test statistic*.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



Problem 27 at the end of this section gives the formula for Satterthwaite's approximation for degrees of freedom.

Continued

d.f. for testing $\mu_1 - \mu_2$ when σ_1 and σ_2 unknown

The sample test statistic distribution is approximately that of a Student's *t* with *degrees of freedom d.f. = smaller of $n_1 - 1$ and $n_2 - 1$* .

Note that statistical software gives a more accurate and larger *d.f.* based on Satterthwaite's approximation (see Problem 27).

3. Use a Student's *t* distribution and the type of test, one-tailed or two-tailed, to find the *P-value* corresponding to the sample test statistic.
4. *Conclude the test.* If *P-value* $\leq \alpha$, then reject H_0 . If *P-value* $> \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.

EXAMPLE 14

TESTING THE DIFFERENCE OF MEANS (σ_1 AND σ_2 UNKNOWN)

Two competing headache remedies claim to give fast-acting relief. An experiment was performed to compare the mean lengths of time required for bodily absorption of brand A and brand B headache remedies.

Twelve people were randomly selected and given an oral dosage of brand A. Another 12 were randomly selected and given an equal dosage of brand B. The lengths of time in minutes for the drugs to reach a specified level in the blood were recorded. The means, standard deviations, and sizes of the two samples follow.

$$\text{Brand A: } \bar{x}_1 = 21.8 \text{ min; } s_1 = 8.7 \text{ min; } n_1 = 12$$

$$\text{Brand B: } \bar{x}_2 = 18.9 \text{ min; } s_2 = 7.5 \text{ min; } n_2 = 12$$

Past experience with the drug composition of the two remedies permits researchers to assume that both distributions are approximately normal. Let us use a 5% level of significance to test the claim that there is no difference between the two brands in the mean time required for bodily absorption. Also, find or estimate the *P-value* of the sample test statistic.

SOLUTION:

(a) $\alpha = 0.05$. The null hypothesis is

$$H_0: \mu_1 = \mu_2 \quad \text{or} \quad H_0: \mu_1 - \mu_2 = 0$$

Since we have no prior knowledge about which brand is faster, the alternate hypothesis is

$$H_1: \mu_1 \neq \mu_2 \quad \text{or} \quad H_1: \mu_1 - \mu_2 \neq 0$$

(b) **Check Requirements** What distribution does the sample test statistic follow?

We use a Student's *t* distribution with *d.f. = smaller sample size - 1 = 12 - 1 = 11*. Note that both samples are of size 12. Degrees of freedom can be computed also by Satterthwaite's approximation. A Student's *t* distribution is appropriate because the original populations are approximately normal, the population standard deviations are not known, and the samples are independent.

(c) Compute the sample test statistic.

We're given $\bar{x}_1 = 21.8$ and $\bar{x}_2 = 18.9$, so the sample difference is $\bar{x}_1 - \bar{x}_2 = 21.8 - 18.9 = 2.9$. Using $s_1 = 8.7$, $s_2 = 7.5$, $n_1 = 12$, $n_2 = 12$, and $\mu_1 - \mu_2 = 0$ from H_0 , we compute the sample test statistic.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{2.9}{\sqrt{\frac{8.7^2}{12} + \frac{7.5^2}{12}}} \approx 0.875$$

(d) Estimate the P -value and sketch the area on a t graph.

Figure 8-18 shows the P -value. The degrees of freedom are $d.f. = 11$ (since both samples are of size 12). Because the test is a two-tailed test, the P -value is the area to the right of 0.875 together with the area to the left of -0.875 . In the Student's t distribution table (Table 6 of Appendix II), we find an interval containing the P -value. Find 0.875 in the row headed by $d.f. = 11$. The test statistic 0.875 falls between the entries 0.697 and 1.214. Because this is a two-tailed test, we use the corresponding P -values 0.500 and 0.250 from the *two-tail area* row (see Table 8-13, Excerpt from Table 6). The P -value for the sample t is in the interval

$$0.250 < P\text{-value} < 0.500$$

(e) Conclude the test.



Since the interval containing the P -value lies to the right of $\alpha = 0.05$, we fail to reject H_0 .

Note: Using the raw data and a calculator with Satterthwaite's approximation for the degrees of freedom $d.f. \approx 21.53$, the P -value ≈ 0.3915 . This value is in the interval we computed.

FIGURE 8-18

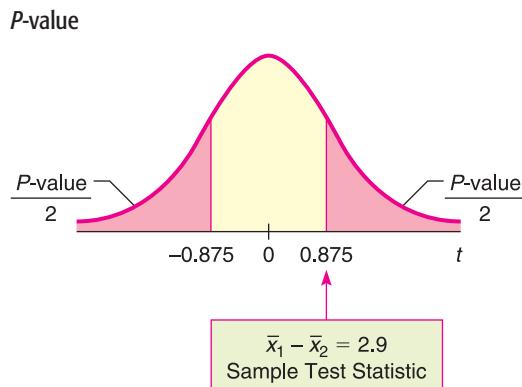


TABLE 8-13 Excerpt from Table 6, Appendix II

one-tail area	0.250	0.125
✓ two-tail area	0.500	0.250
$d.f. = 11$	0.697	1.214
	↑ Sample $t = 0.875$	

(f) *Interpretation* Interpret the results.

At the 5% level of significance, there is insufficient evidence to conclude that there is a difference in mean times for the remedies to reach the specified level in the bloodstream.

GUIDED EXERCISE 10

Testing the difference of means (σ_1 and σ_2 unknown)

Suppose the experiment to measure the times in minutes for the headache remedies to enter the bloodstream (Example 14) yielded sample means, sample standard deviations, and sample sizes as follows:

Brand A: $\bar{x}_1 = 20.1$ min; $s_1 = 8.7$ min; $n_1 = 12$

Brand B: $\bar{x}_2 = 11.2$ min; $s_2 = 7.5$ min; $n_2 = 8$

Brand B claims to be faster. Is this claim justified at the 1% level of significance? (Use the following steps to obtain the answer.)

Continued

GUIDED EXERCISE 10 *continued*

- (a) What is α ? State H_0 and H_1 .

→ $\alpha = 0.01$

$H_0: \mu_1 = \mu_2$ or $H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 > \mu_2$ (or $H_1: \mu_1 - \mu_2 > 0$). This states that the mean time for brand B is less than the mean time for brand A.

- (b) **Check Requirements** What distribution does the sample test statistic follow? Explain

→ Student's t distribution with $d.f. n_2 - 1 = 8 - 1 = 7$ since $n_2 < n_1$. From Example 14, we know that the original distributions are approximately normal. Both σ_1 and σ_2 are unknown and the samples are independent.

- (c) Compute $\bar{x}_1 - \bar{x}_2$ and the corresponding sample test statistic t .

→ $\bar{x}_1 - \bar{x}_2 = 20.1 - 11.2 = 8.9$. Using $s_1 = 8.7$, $s_2 = 7.5$, $n_1 = 12$, $n_2 = 8$, and $\mu_1 - \mu_2 = 0$ from H_0 , we have

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{8.9}{\sqrt{\frac{8.7^2}{12} + \frac{7.5^2}{8}}} \approx 2.437$$

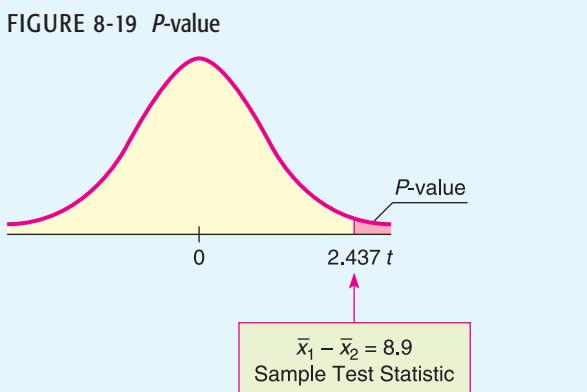
- (d) What degrees of freedom do you use? To find an interval containing the P -value, do you use one-tail or two-tail areas of Table 6, Appendix II? Sketch a figure showing the P -value. Find an interval for the P -value.

→ Since $n_2 < n_1$, $d.f. = n_2 - 1 = 8 - 1 = 7$. Use *one-tail area* of Table 6.

TABLE 8-14 Excerpt from Student's t Table

✓ one-tail area	0.025	0.010
two-tail area	0.050	0.020
$d.f. = 7$	2.365	2.998

↑
Sample $t = 2.437$

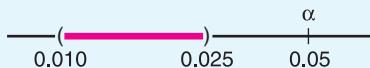


The sample t is between 2.365 and 2.998.

$0.010 < P\text{-value} < 0.025$

- (e) Do we reject or fail to reject H_0 ?

→ Since the interval containing the P -value has values that are all less than 0.05, we reject H_0 .



Note: On the calculator with Satterthwaite's approximation for $d.f.$, we have $d.f. = 16.66$ and $P\text{-value} \approx 0.013$.

- (f) **Interpret** the results in the context of the application.

→ At the 5% level of significance, there is sufficient evidence to conclude that the mean time for brand B to enter the bloodstream is less than that for brand A.

Alternate method using pooled standard deviation

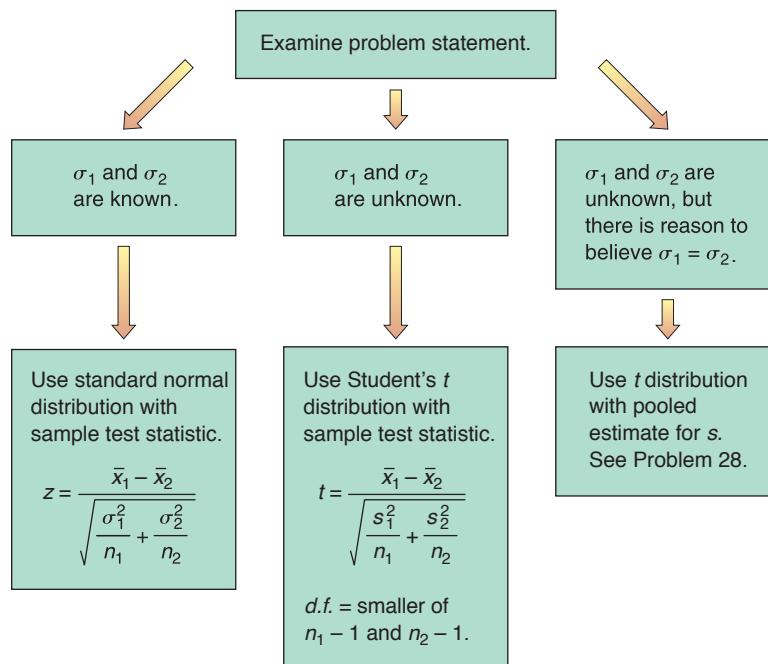


Problem 28 at the end of this section discusses the pooled standard deviation method for testing the difference of means when population standard deviations are equal. Section 10.4 has techniques for testing whether two standard deviations or two variances are equal.

Pooled standard deviation

There is another method of testing $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown. Suppose that the sample values s_1 and s_2 are sufficiently close and that there is reason to believe $\sigma_1 = \sigma_2$ (or the standard deviations are approximately equal). This situation can happen when you make a slight change or alteration to a known process or method of production. The standard deviation may not change much, but the outputs or means could be very different. When there is reason to believe that $\sigma_1 = \sigma_2$, it is best to use a *pooled standard deviation*. The sample test statistic $\bar{x}_1 - \bar{x}_2$ has a corresponding t variable with an *exact Student's t* distribution and degrees of freedom $d.f. = n_1 + n_2 - 2$. Problem 28 at the end of the section provides the details.

Under the null hypothesis $H_0: \mu_1 = \mu_2$ which distribution should you use for the sample test statistic $\bar{x}_1 - \bar{x}_2$?



Part III: Testing a Difference of Proportions $p_1 - p_2$

Suppose we have two independent binomial experiments. That is, outcomes from one binomial experiment are in no way paired with outcomes from the other. We use the notation

Binomial Experiment 1

n_1 = number of trials

r_1 = number of successes

p_1 = population probability
of success on a single trial

Binomial Experiment 2

n_2 = number of trials

r_2 = number of successes

p_2 = population probability
of success on a single trial

For large values of n_1 and n_2 , the distribution of sample differences

$$\hat{p}_1 - \hat{p}_2 = \frac{r_1}{n_1} - \frac{r_2}{n_2}$$

is closely approximated by a *normal distribution* with mean μ and standard deviation σ as shown:

$$\mu = p_1 - p_2 \quad \sigma = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Pooled estimate of proportion \bar{p}

For most practical problems involving a comparison of two binomial populations, the experimenters will want to test the null hypothesis $H_0: p_1 = p_2$. Consequently, this is the type of test we shall consider. Since the values of p_1 and p_2 are unknown, and since specific values are not assumed under the null hypothesis $H_0: p_1 = p_2$, the best estimate for the common value is the total number of successes ($r_1 + r_2$) divided by the total number of trials ($n_1 + n_2$). If we denote this *pooled estimate of proportion* by \bar{p} (read “p bar”), then

$$\bar{p} = \frac{r_1 + r_2}{n_1 + n_2}$$

This formula gives the best sample estimate \bar{p} for p_1 and p_2 *under the assumption that $p_1 = p_2$* . Also, $\bar{q} = 1 - \bar{p}$.

Criteria for using the normal approximation to the binomial**Sample test statistic**

COMMENT For most practical applications, the sample sizes n_1 and n_2 are considered large samples if each of the four quantities

$$n_1\bar{p} \quad n_1\bar{q} \quad n_2\bar{p} \quad n_2\bar{q}$$

is larger than 5 (see Section 6.6).

As stated earlier, the sample statistic $\hat{p}_1 - \hat{p}_2$ has a normal distribution with mean $\mu = p_1 - p_2$ and standard deviation $\sigma = \sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}$. Under the null hypothesis, we assume that $p_1 = p_2$ and then use the pooled estimate \bar{p} in place of each p . Using all this information, we find that the *sample test statistic* is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

$$\text{where } \bar{p} = \frac{r_1 + r_2}{n_1 + n_2} \quad \text{and} \quad \bar{q} = 1 - \bar{p}$$

$$\hat{p}_1 = \frac{r_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{r_2}{n_2}$$

Using this information, we summarize the procedure for testing $p_1 - p_2$.

PROCEDURE**HOW TO TEST A DIFFERENCE OF PROPORTIONS $P_1 - P_2$** **Procedure**

Consider two independent binomial experiments.

Binomial Experiment 1

n_1 = number of trials

r_1 = number of successes
out of n_1 trials

$$\hat{p}_1 = \frac{r_1}{n_1}$$

p_1 = population probability of
success on a single trial

Binomial Experiment 2

n_2 = number of trials

r_2 = number of successes
out of n_2 trials

$$\hat{p}_2 = \frac{r_2}{n_2}$$

p_2 = population probability of
success on a single trial

1. Use the *null hypothesis* of no difference, $H_0: p_1 - p_2 = 0$. In the context of the application, choose the *alternate hypothesis*. Set the *level of significance* α .

2. The null hypothesis claims that $p_1 = p_2$; therefore, *pooled best estimates* for the population probabilities of success and failure are

$$\bar{p} = \frac{r_1 + r_2}{n_1 + n_2} \quad \text{and} \quad \bar{q} = 1 - \bar{p}$$

Requirements

The number of trials should be sufficiently large so that each of the four quantities $n_1\bar{p}$, $n_1\bar{q}$, $n_2\bar{p}$, and $n_2\bar{q}$ is larger than 5.

In this case, you compute the *sample test statistic*

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

3. Use the standard normal distribution and a type of test, one-tailed or two-tailed, to find the *P-value* corresponding to the sample test statistic.
4. *Conclude the test.* If $P\text{-value} \leq \alpha$, then reject H_0 . If $P\text{-value} > \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.

EXAMPLE 15

TESTING THE DIFFERENCE OF PROPORTIONS

The Macek County Clerk wishes to improve voter registration. One method under consideration is to send reminders in the mail to all citizens in the county who are eligible to register. As part of a pilot study to determine if this method will actually improve voter registration, a random sample of 1250 potential voters was taken. Then this sample was randomly divided into two groups.

Group 1: There were 625 people in this group. No reminders to register were sent to them. The number of potential voters from this group who registered was 295.

Group 2: This group also contained 625 people. Reminders were sent in the mail to each member in the group, and the number who registered to vote was 350.

The county clerk claims that the proportion of people who registered was significantly greater in group 2. On the basis of this claim, the clerk recommends that the project be funded for the entire population of Macek County. Use a 5% level of significance to test the claim that the proportion of potential voters who registered was greater in group 2, the group that received reminders.

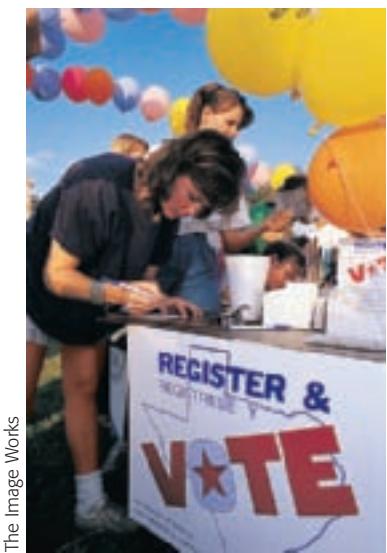
SOLUTION:

(a) Note that $\alpha = 0.05$. Let p_1 be the proportion of voters from group 1 who registered, and let p_2 be the proportion from group 2 who registered. The null hypothesis is that there is no difference in proportions, so,

$$H_0: p_1 = p_2 \quad \text{or} \quad H_0: p_1 - p_2 = 0$$

The alternate hypothesis is that the proportion of voters who registered was greater in the group that received reminders.

$$H_1: p_1 < p_2 \quad \text{or} \quad H_1: p_1 - p_2 < 0$$



The Image Works

(b) Calculate the *pooled estimates* \bar{p} and \bar{q} .

Under the null hypothesis that $p_1 = p_2$ we find

$$\bar{p} = \frac{r_1 + r_2}{n_1 + n_2} = \frac{295 + 350}{625 + 625} \approx 0.516 \quad \text{and} \quad \bar{q} = 1 - \bar{p} \approx 0.484$$

(c) **Check Requirements** What distribution does the sample test statistic follow?

The sample test statistic follows a standard normal distribution because the number of binomial trials is large enough that each of the products $n_1\bar{p}$, $n_1\bar{q}$, $n_2\bar{p}$, $n_2\bar{q}$ exceeds 5.

(d) Compute $\hat{p}_1 - \hat{p}_2$ and convert it to the sample statistic z .

CALCULATOR NOTE Carry the values for \hat{p}_1 , \hat{p}_2 , and the pooled estimates \bar{p} and \bar{q} out to at least three places after the decimal. Then round the z value of the corresponding test statistic to two places after the decimal.

For the first group, the number of successes is $r_1 = 295$ out of $n_1 = 625$ trials. For the second group, there are $r_2 = 350$ successes out of $n_2 = 625$ trials. Since

$$\hat{p}_1 = \frac{r_1}{n_1} = \frac{295}{625} = 0.472 \quad \text{and} \quad \hat{p}_2 = \frac{r_2}{n_2} = \frac{350}{625} = 0.560$$

then

$$\hat{p}_1 - \hat{p}_2 = 0.472 - 0.560 = -0.088$$

We computed the pooled estimate for \bar{p} and \bar{q} in part (b).

$$\bar{p} = 0.516 \quad \text{and} \quad \bar{q} = 1 - \bar{p} = 0.484$$

Using these values, we find the sample test statistic

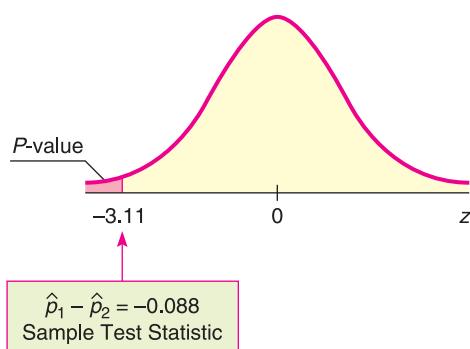
$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} = \frac{-0.088}{\sqrt{\frac{(0.516)(0.484)}{625} + \frac{(0.516)(0.484)}{625}}} \approx -3.11$$

(e) Find the *P-value* and sketch the area on the standard normal curve.

Figure 8-20 shows the *P-value*. This is a left-tailed test, so the *P-value* is the area to the left of -3.11 . Using the standard normal distribution (Table 5 of Appendix II), we find $P\text{-value} = P(z < -3.11) \approx 0.0009$.

FIGURE 8-20

P-value



(f) Conclude the test.

Since P -value of $0.0009 \leq 0.05$ for α , we reject H_0 .

(g) **Interpretation** Interpret the results in the context of the application.

At the 5% level of significance, the data indicate that the population proportion of potential voters who registered was greater in group 2, the group that received reminders.

GUIDED EXERCISE 11

Testing the difference of proportions

In Example 15 about voter registration, suppose that a random sample of 1100 potential voters was randomly divided into two groups.

Group 1: 500 potential voters; no registration reminders sent; 248 registered to vote

Group 2: 600 potential voters; registration reminders sent; 332 registered to vote

Do these data support the claim that the proportion of voters who registered was greater in the group that received reminders than in the group that did not? Use a 1% level of significance.

(a) What is α ? State H_0 and H_1 .

➡ $\alpha = 0.01$. As before, $H_0: p_1 = p_2$ and $H_1: p_1 < p_2$.

(b) Under the null hypothesis $p_1 = p_2$, calculate the *pooled estimates* \bar{p} and \bar{q}

➡ $n_1 = 500, r_1 = 248; n_2 = 600, r_2 = 332$

$$\bar{p} = \frac{r_1 + r_2}{n_1 + n_2} = \frac{248 + 332}{500 + 600} \approx 0.527$$

$$\bar{q} = 1 - \bar{p} \approx 1 - 0.527 \approx 0.473$$

(c) **Check Requirements** What distribution does the sample test statistic follow?

➡ The sample test statistic z follows a standard normal distribution because the number of binomial trials is large enough that each of the products $n_1\bar{p}, n_1\bar{q}, n_2\bar{p}, n_2\bar{q}$ exceeds 5.

(d) Compute $\hat{p}_1 - \hat{p}_2$.

$$\hat{p}_1 = \frac{r_1}{n_1} = \frac{248}{500} = 0.496 \quad \hat{p}_2 = \frac{r_2}{n_2} = \frac{332}{600} \approx 0.553 \\ \hat{p}_1 - \hat{p}_2 \approx -0.057$$

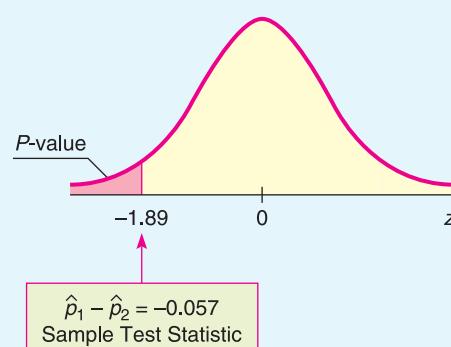
(e) Find the sample test statistic.

$$\hat{z} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}} = \frac{-0.057}{\sqrt{\frac{(0.527)(0.473)}{500} + \frac{(0.572)(0.473)}{600}}} \\ \approx -1.89$$

(f) Find the P -value and sketch the area on the standard normal curve.

➡ Figure 8-21 shows the P -value. It is the area to the left of $z = -1.89$. Using Table 5 of Appendix II, we find P -value = $P(z < -1.89) = 0.0294$.

FIGURE 8-21 P -value



Continued

GUIDED EXERCISE 11 *continued*

(g) Conclude the test.

Since P -value of $0.0294 > 0.01$ for α , we cannot reject H_0 .

(h) *Interpret* the results in the context of the application.

At the 1% level of significance, the data do not support the claim that the reminders increased the proportion of registered voters.

 TECH NOTES

The TI-84Plus/TI-83Plus/TI-nspire calculators, Excel 2007, and Minitab all support testing the difference of means for independent samples. The TI-84Plus/TI-83Plus/TI-nspire calculators and Minitab also support testing the difference of proportions. When testing the difference of means using the normal distribution, the technologies require the population standard deviations for the distributions of the two samples. When testing the difference of means using a Student's t distribution, the technologies give the option of using the pooled standard deviation. As discussed in Problem 28 of this section, the pooled standard deviation is appropriate when the standard deviations of the two populations are approximately equal. If the pooled standard deviation option is not selected, the technologies compute the sample test statistic for $\bar{x}_1 - \bar{x}_2$ using the procedures described in this section. However, Satterthwaite's approximation for the degrees of freedom is used.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) Enter the data. Press STAT and select TESTS. Options 3: 2-SampZTest, 4: 2-SampTTest, 6: 2-PropZTest perform tests for the difference of means using the normal distribution, difference of means using a Student's t distribution, and difference of proportions, respectively. The calculator uses the symbol \hat{p} to designate the pooled estimate for p .

Excel 2007 Enter the data in columns. On the home ribbon, click the Data tab. Then in the Analysis Group, select Data Analysis. The choice z-Test Two Sample Means conducts a test for the difference of means using the normal distribution. The choice t-Test: Two-Sample Assuming Unequal Variances conducts a test using a Student's t distribution with Satterthwaite's approximation for the degrees of freedom. The choice t-Test: Two-Sample Assuming Equal Variances conducts a test using the pooled standard deviation.

Minitab Enter the data in two columns. Use the menu choices Stat ▶ Basic Statistics. The choice 2 sample t tests the difference of means using a Student's t distribution. In the dialogue box, leaving the box Assume equal variances unchecked produces a test using Satterthwaite's approximation for the degrees of freedom. Checking the box Assume equal variances produces a test using the pooled standard deviation. Minitab does not support testing the difference of means using the normal distribution. The menu item 2 Proportions tests the difference of proportions. Under Options, select the null and alternate hypotheses.


Part IV: Testing $\mu_1 - \mu_2$ and $p_1 - p_2$ Using Critical Regions

For a fixed, preset level of significance α , the P -value method of testing is logically equivalent to the critical region method of testing. This book emphasizes the P -value method because of its great popularity and because it is readily compatible with most computer software. However, for completeness, we provide an optional example utilizing the critical region method.

Recall that the critical region method and the P -value method of testing share a number of steps. Both methods use the same *null and alternate hypotheses*, the same *sample test statistic*, and the same *sampling distribution* for the test statistic. However, instead of computing the P -value and comparing it to the level of significance α , the critical region method compares the sample test statistic to a *critical value* from the sampling distribution that is based on α and the alternate hypothesis (left-tailed, right-tailed, or two-tailed).

Test conclusions based on critical values

For a *right-tailed test*, if the *sample test statistic* \geq *critical value*, reject H_0 .

For a *left-tailed test*, if the *sample test statistic* \leq *critical value*, reject H_0 .

For a *two-tailed test*, if the *sample test statistic* lies at or beyond the *critical values* (that is, \leq negative critical value or \geq positive critical value), reject H_0 .

Otherwise, in each case, do not reject H_0 .

Critical values z_0 for tests using the standard normal distribution can be found in Table 5(c) of Appendix II. Critical values t_0 for tests using a Student's t distribution are found in Table 6 of Appendix II. Use the row headed by the appropriate degrees of freedom and the column that includes the value of α (level of significance) in the *one-tail area* row for one-tailed tests or the *two-tail area* row for two-tailed tests.

EXAMPLE 16

CRITICAL REGION METHOD

Use the critical region method to solve the application in Example 14 (testing $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown).

SOLUTION: Example 14 involves testing the difference in average time for two headache remedies to reach the bloodstream. For brand A, $\bar{x}_1 = 21.8$ min, $s_1 = 8.7$ min, and $n_1 = 12$; for brand B, $\bar{x}_2 = 18.9$ min, $s_2 = 7.5$ min, and $n_2 = 12$. The level of significance α is 0.05. The Student's t distribution is appropriate because both populations are approximately normal.

- (a) To use the critical region method to test for a difference in average time, we use the same hypotheses and the same sample test statistic as in Example 14.

$$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2;$$

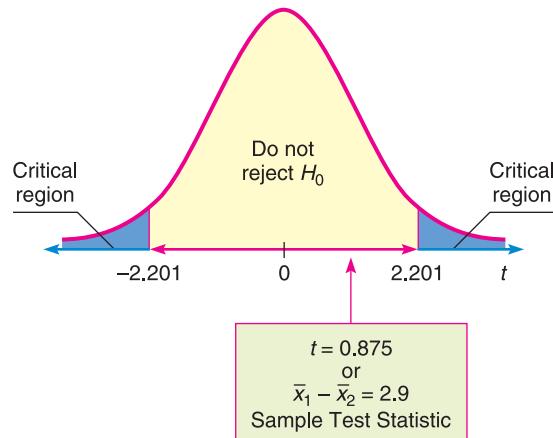
$$\bar{x}_1 - \bar{x}_2 = 2.9 \text{ min; sample test statistic } t = 0.875$$

- (b) Instead of finding the P -value of the sample test statistic, we use α and H_1 to find the critical values in Table 6 of Appendix II. We have $d.f. = 11$ (since both samples are of size 12). We find $\alpha = 0.05$ in the *two-tail area* row, since we have a two-tailed test. The critical values are $\pm t_0 = \pm 2.201$.

Next compare the sample test statistic $t = 0.875$ to the critical values. Figure 8-22 shows the critical regions and the sample test statistic. We see that the sample test statistic falls in the "do not reject H_0 " region. At the 5% level of significance, sample evidence does not show a difference in the time for the drugs to reach the bloodstream. The result is consistent with the result obtained by the P -value method of Example 14.

FIGURE 8-22

Critical Regions $\alpha = 0.05$; $d.f. = 11$



VIEWPOINT**Temper! Temper!**

In her book *Red Ink Behaviors*, Jean Hollands discusses inappropriate, problem behaviors of professional employees in the corporate business world. Temper tantrums, flaming e-mails, omitting essential information, sabotaging fellow workers, and the arrogant opinion that others are “dumb and dispensable” create personnel problems that cost companies a lot in the form of wasted time, reduced productivity, and lost revenues. A study of major industries in the Silicon Valley area gave Hollands data for estimating just how much time and money are wasted by such “red ink behaviors.” For more information, see Problems 25 and 26 in this section.

**SECTION 8.5
PROBLEMS**

1. **Statistical Literacy** Consider a hypothesis test of difference of means for two independent populations x_1 and x_2 .
 - (a) What does the null hypothesis say about the relationship between the two population means?
 - (b) In the case that the sample test statistic is z , give the formula for z .
 - (c) In the case that the sample test statistic is t , give the formula for t .
2. **Statistical Literacy** Consider a hypothesis test of difference of means for two independent populations x_1 and x_2 . Suppose that both sample sizes are greater than 30 and that you know σ_1 but not σ_2 . Is it standard practice to use the normal distribution or a Student’s t distribution?
3. **Statistical Literacy** Consider a hypothesis test of difference of means for two independent populations x_1 and x_2 . What are two ways of expressing the null hypothesis?
4. **Statistical Literacy** Consider a hypothesis test of difference of proportions for two independent populations. Suppose random samples produce r_1 successes out of n_1 trials for the first population and r_2 successes out of n_2 trials for the second population.
 - (a) What does the null hypothesis claim about the relationship between the proportions of successes in the two populations?
 - (b) What is the formula for the sample test statistic?
5. **Statistical Literacy** Consider a hypothesis test of difference of proportions for two independent populations. Suppose random samples produce r_1 successes out of n_1 trials for the first population and r_2 successes out of n_2 trials for the second population. What is the best pooled estimate \bar{p} for the population probability of success using $H_0: p_1 = p_2$?
6. **Critical Thinking** Consider use of a Student’s t distribution to test the difference of means for independent populations using random samples of sizes n_1 and n_2 .
 - (a) Which process gives the larger degrees of freedom, Satterthwaite’s approximation or using the smaller of $n_1 - 1$ and $n_2 - 1$? Which method is more conservative? What do we mean by “conservative”? Note that computer programs and other technologies commonly use Satterthwaite’s approximation.
 - (b) Using the same hypotheses and sample data, is the P -value smaller for larger degrees of freedom? How might a larger P -value impact the significance of a test?
7. **Critical Thinking** When conducting a test for the difference of means for two independent populations x_1 and x_2 , what alternate hypothesis would indicate that the mean of the x_2 population is smaller than that of the x_1 population? Express the alternate hypothesis in two ways.

8. **Critical Thinking** When conducting a test for the difference of means for two independent populations x_1 and x_2 , what alternate hypothesis would indicate that the mean of the x_2 population is larger than that of the x_1 population? Express the alternate hypothesis in two ways.
9. **Basic Computation: Testing $\mu_1 - \mu_2$** A random sample of 49 measurements from one population had a sample mean of 10, with sample standard deviation 3. An independent random sample of 64 measurements from a second population had a sample mean of 12, with sample standard deviation 4. Test the claim that the population means are different. Use level of significance 0.01.
(a) **Check Requirements** What distribution does the sample test statistic follow? Explain.
(b) State the hypotheses.
(c) Compute $\bar{x}_1 - \bar{x}_2$ and the corresponding sample test statistic.
(d) Estimate the P -value of the sample test statistic.
(e) Conclude the test.
(f) **Interpret** the results.
10. **Basic Computation: Testing $\mu_1 - \mu_2$** Two populations have mound-shaped, symmetric distributions. A random sample of 16 measurements from the first population had a sample mean of 20, with sample standard deviation 2. An independent random sample of 9 measurements from the second population had a sample mean of 19, with sample standard deviation 3. Test the claim that the population mean of the first population exceeds that of the second. Use a 5% level of significance.
(a) **Check Requirements** What distribution does the sample test statistic follow? Explain.
(b) State the hypotheses.
(c) Compute $\bar{x}_1 - \bar{x}_2$ and the corresponding sample test statistic.
(d) Estimate the P -value of the sample test statistic.
(e) Conclude the test.
(f) **Interpret** the results.
11. **Basic Computation: Testing $\mu_1 - \mu_2$** A random sample of 49 measurements from a population with population standard deviation 3 had a sample mean of 10. An independent random sample of 64 measurements from a second population with population standard deviation 4 had a sample mean of 12. Test the claim that the population means are different. Use level of significance 0.01.
(a) **Check Requirements** What distribution does the sample test statistic follow? Explain.
(b) State the hypotheses.
(c) Compute $\bar{x}_1 - \bar{x}_2$ and the corresponding sample test statistic.
(d) Find the P -value of the sample test statistic.
(e) Conclude the test.
(f) **Interpret** the results.
12. **Basic Computation: Testing $\mu_1 - \mu_2$** Two populations have normal distributions. The first has population standard deviation 2 and the second has population standard deviation 3. A random sample of 16 measurements from the first population had a sample mean of 20. An independent random sample of 9 measurements from the second population had a sample mean of 19. Test the claim that the population mean of the first population exceeds that of the second. Use a 5% level of significance.
(a) **Check Requirements** What distribution does the sample test statistic follow? Explain.
(b) State the hypotheses.
(c) Compute $\bar{x}_1 - \bar{x}_2$ and the corresponding sample test statistic.
(d) Find the P -value of the sample test statistic.
(e) Conclude the test
(f) **Interpret** the results.

13. **Basic Computation: Test $p_1 - p_2$** For one binomial experiment, 75 binomial trials produced 45 successes. For a second independent binomial experiment, 100 binomial trials produced 70 successes. At the 5% level of significance, test the claim that the probabilities of success for the two binomial experiments differ.
- Compute the pooled probability of success for the two experiments.
 - Check Requirements** What distribution does the sample test statistic follow? Explain.
 - State the hypotheses.
 - Compute $\hat{p}_1 - \hat{p}_2$ and the corresponding sample test statistic.
 - Find the P -value of the sample test statistic.
 - Conclude the test.
 - Interpret** the results.
14. **Basic Computation: Test $p_1 - p_2$** For one binomial experiment, 200 binomial trials produced 60 successes. For a second independent binomial experiment, 400 binomial trials produced 156 successes. At the 5% level of significance, test the claim that the probability of success for the second binomial experiment is greater than that for the first.
- Compute the pooled probability of success for the two experiments.
 - Check Requirements** What distribution does the sample test statistic follow? Explain.
 - State the hypotheses.
 - Compute $\hat{p}_1 - \hat{p}_2$ and the corresponding sample test statistic.
 - Find the P -value of the sample test statistic.
 - Conclude the test.
 - Interpret** the results.

Please provide the following information for Problems 15–26 and 29–35.

- What is the level of significance? State the null and alternate hypotheses.
- Check Requirements** What sampling distribution will you use? What assumptions are you making? Compute the sample test statistic.
- Find (or estimate) the P -value. Sketch the sampling distribution and show the area corresponding to the P -value.
- Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis? Are the data statistically significant at level α ?
- Interpret** your conclusion in the context of the application.

Note: For degrees of freedom $d.f.$ not in the Student's t table, use the closest $d.f.$ that is *smaller*. In some situations, this choice of $d.f.$ may increase the P -value a small amount, and therefore produce a slightly more "conservative" answer.

Answers may vary due to rounding.

15. **Medical: REM Sleep** REM (rapid eye movement) sleep is sleep during which most dreams occur. Each night a person has both REM and non-REM sleep. However, it is thought that children have more REM sleep than adults (Reference: *Secrets of Sleep* by Dr. A. Borbely). Assume that REM sleep time is normally distributed for both children and adults. A random sample of $n_1 = 10$ children (9 years old) showed that they had an average REM sleep time of $\bar{x}_1 = 2.8$ hours per night. From previous studies, it is known that $\sigma_1 = 0.5$ hour. Another random sample of $n_2 = 10$ adults showed that they had an average REM sleep time of $\bar{x}_2 = 2.1$ hours per night. Previous studies show that $\sigma_2 = 0.7$ hour. Do these data indicate that, on average, children tend to have more REM sleep than adults? Use a 1% level of significance.
16. **Environment: Pollution Index** Based on information from *The Denver Post*, a random sample of $n_1 = 12$ winter days in Denver gave a sample mean pollution index of $\bar{x}_1 = 43$. Previous studies show that $\sigma_1 = 21$. For Englewood (a suburb of Denver), a random sample of $n_2 = 14$ winter days gave a sample mean pollution index of $\bar{x}_2 = 36$. Previous studies show that $\sigma_2 = 15$. Assume the pollution index is normally distributed in both Englewood and

- Denver. Do these data indicate that the mean population pollution index of Englewood is different (either way) from that of Denver in the winter? Use a 1% level of significance.
17. **Survey: Outdoor Activities** A Michigan study concerning preference for outdoor activities used a questionnaire with a 6-point Likert-type response in which 1 designated “not important” and 6 designated “extremely important.” A random sample of $n_1 = 46$ adults were asked about fishing as an outdoor activity. The mean response was $\bar{x}_1 = 4.9$. Another random sample of $n_2 = 51$ adults were asked about camping as an outdoor activity. For this group, the mean response was $\bar{x}_2 = 4.3$. From previous studies, it is known that $\sigma_1 = 1.5$ and $\sigma_2 = 1.2$. Does this indicate a difference (either way) regarding preference for camping versus preference for fishing as an outdoor activity? Use a 5% level of significance. Note: A *Likert scale* usually has to do with approval of or agreement with a statement in a questionnaire. For example, respondents are asked to indicate whether they “strongly agree,” “agree,” “disagree,” or “strongly disagree” with the statement.
18. **Generation Gap: Education** Education influences attitude and lifestyle. Differences in education are a big factor in the “generation gap.” Is the younger generation really better educated? Large surveys of people age 65 and older were taken in $n_1 = 32$ U.S. cities. The sample mean for these cities showed that $\bar{x}_1 = 15.2\%$ of the older adults had attended college. Large surveys of young adults (ages 25–34) were taken in $n_2 = 35$ U.S. cities. The sample mean for these cities showed that $\bar{x}_2 = 19.7\%$ of the young adults had attended college. From previous studies, it is known that $\sigma_1 = 7.2\%$ and $\sigma_2 = 5.2\%$ (Reference: *American Generations* by S. Mitchell). Does this information indicate that the population mean percentage of young adults who attended college is higher? Use $\alpha = 0.05$.
19. **Crime Rate: FBI** A random sample of $n_1 = 10$ regions in New England gave the following violent crime rates (per million population).
- x_1 : New England Crime Rate
- | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 3.5 | 3.7 | 4.0 | 3.9 | 3.3 | 4.1 | 1.8 | 4.8 | 2.9 | 3.1 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
- Another random sample of $n_2 = 12$ regions in the Rocky Mountain states gave the following violent crime rates (per million population).
- x_2 : Rocky Mountain States
- | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 3.7 | 4.3 | 4.5 | 5.3 | 3.3 | 4.8 | 3.5 | 2.4 | 3.1 | 3.5 | 5.2 | 2.8 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
- (Reference: *Crime in the United States*, Federal Bureau of Investigation.) Assume that the crime rate distribution is approximately normal in both regions.
- Use a calculator to verify that $\bar{x}_1 \approx 3.51$, $s_1 \approx 0.81$, $\bar{x}_2 \approx 3.87$, and $s_2 \approx 0.94$.
 - Do the data indicate that the violent crime rate in the Rocky Mountain region is higher than that in New England? Use $\alpha = 0.01$.
20. **Medical: Hay Fever** A random sample of $n_1 = 16$ communities in western Kansas gave the following information for people under 25 years of age.
- x_1 : Rate of hay fever per 1000 population for people under 25
- | | | | | | | | |
|-----|----|-----|-----|----|-----|-----|----|
| 98 | 90 | 120 | 128 | 92 | 123 | 112 | 93 |
| 125 | 95 | 125 | 117 | 97 | 122 | 127 | 88 |
- A random sample of $n_2 = 14$ regions in western Kansas gave the following information for people over 50 years old.
- x_2 : Rate of hay fever per 1000 population for people over 50
- | | | | | | | |
|----|-----|-----|----|-----|----|-----|
| 95 | 110 | 101 | 97 | 112 | 88 | 110 |
| 79 | 115 | 100 | 89 | 114 | 85 | 96 |
- (Reference: National Center for Health Statistics.)
- Use a calculator to verify that $\bar{x}_1 \approx 109.50$, $s_1 \approx 15.41$, $\bar{x}_2 \approx 99.36$, and $s_2 \approx 11.57$.

- ii. Assume that the hay fever rate in each age group has an approximately normal distribution. Do the data indicate that the age group over 50 has a lower rate of hay fever? Use $\alpha = 0.05$.
21. ***Education: Tutoring*** In the journal *Mental Retardation*, an article reported the results of a peer tutoring program to help mildly mentally retarded children learn to read. In the experiment, the mildly retarded children were randomly divided into two groups: the experimental group received peer tutoring along with regular instruction, and the control group received regular instruction with no peer tutoring. There were $n_1 = n_2 = 30$ children in each group. The Gates–MacGintie Reading Test was given to both groups before instruction began. For the experimental group, the mean score on the vocabulary portion of the test was $\bar{x}_1 = 344.5$, with sample standard deviation $s_1 = 49.1$. For the control group, the mean score on the same test was $\bar{x}_2 = 354.2$, with sample standard deviation $s_2 = 50.9$. Use a 5% level of significance to test the hypothesis that there was no difference in the vocabulary scores of the two groups before the instruction began.
22. ***Education: Tutoring*** In the article cited in Problem 21, the results of the following experiment were reported. Form 2 of the Gates–MacGintie Reading Test was administered to both an experimental group and a control group after 6 weeks of instruction, during which the experimental group received peer tutoring and the control group did not. For the experimental group $n_1 = 30$ children, the mean score on the vocabulary portion of the test was $\bar{x}_1 = 368.4$, with sample standard deviation $s_1 = 39.5$. The average score on the vocabulary portion of the test for the $n_2 = 30$ subjects in the control group was $\bar{x}_2 = 349.2$, with sample standard deviation $s_2 = 56.6$. Use a 1% level of significance to test the claim that the experimental group performed better than the control group.
23. ***Wildlife: Fox Rabies*** A study of fox rabies in southern Germany gave the following information about different regions and the occurrence of rabies in each region (Reference: B. Sayers et al., “A Pattern Analysis Study of a Wildlife Rabies Epizootic,” *Medical Informatics*, Vol. 2, pp. 11–34). Based on information from this article, a random sample of $n_1 = 16$ locations in region I gave the following information about the number of cases of fox rabies near that location.
- | | | | | | | | | |
|----------------------|---|---|---|---|---|---|---|---|
| $x_1:$ Region I data | 1 | 8 | 8 | 8 | 7 | 8 | 8 | 1 |
| | 3 | 3 | 3 | 2 | 5 | 1 | 4 | 6 |
- A second random sample of $n_2 = 15$ locations in region II gave the following information about the number of cases of fox rabies near that location.
- | | | | | | | | | |
|-----------------------|---|---|---|---|---|---|---|---|
| $x_2:$ Region II data | 1 | 1 | 3 | 1 | 4 | 8 | 5 | 4 |
| | 4 | 4 | 2 | 2 | 5 | 6 | 9 | |
- i. Use a calculator with sample mean and sample standard deviation keys to verify that $\bar{x}_1 = 4.75$ with $s_1 \approx 2.82$ in region I and $\bar{x}_2 \approx 3.93$ with $s_2 \approx 2.43$ in region II.
- ii. Does this information indicate that there is a difference (either way) in the mean number of cases of fox rabies between the two regions? Use a 5% level of significance. (Assume the distribution of rabies cases in both regions is mound-shaped and approximately normal.)
24. ***Agriculture: Bell Peppers*** The pathogen *Phytophthora capsici* causes bell peppers to wilt and die. Because bell peppers are an important commercial crop, this disease has undergone a great deal of agricultural research. It is thought that too much water aids the spread of the pathogen. Two fields are under study. The first step in the research project is to compare the mean soil water content for the two fields (Source: *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 2, No. 2). Units are percent water by volume of soil.

Field A samples, x_1 :

10.2	10.7	15.5	10.4	9.9	10.0	16.6
15.1	15.2	13.8	14.1	11.4	11.5	11.0

Field B samples, x_2 :

8.1	8.5	8.4	7.3	8.0	7.1	13.9	12.2
13.4	11.3	12.6	12.6	12.7	12.4	11.3	12.5

- i. Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 12.53$, $s_1 \approx 2.39$, $\bar{x}_2 \approx 10.77$, and $s_2 = 2.40$.
- ii. Assuming the distribution of soil water content in each field is mound-shaped and symmetric, use a 5% level of significance to test the claim that field A has, on average, a higher soil water content than field B.

25. **Management: Lost Time** In her book *Red Ink Behaviors*, Jean Hollands reports on the assessment of leading Silicon Valley companies regarding a manager's lost time due to inappropriate behavior of employees. Consider the following independent random variables. The first variable x_1 measures a manager's hours per week lost due to hot tempers, flaming e-mails, and general unproductive tensions:

$$x_1: \quad 1 \quad 5 \quad 8 \quad 4 \quad 2 \quad 4 \quad 10$$

The variable x_2 measures a manager's hours per week lost due to disputes regarding technical workers' superior attitudes that their colleagues are "dumb and dispensable":

$$x_2: \quad 10 \quad 5 \quad 4 \quad 7 \quad 9 \quad 4 \quad 10 \quad 3$$

- i. Use a calculator with sample mean and standard deviation keys to verify that $\bar{x}_1 \approx 4.86$, $s_1 \approx 3.18$, $\bar{x}_2 = 6.5$, and $s_2 \approx 2.88$.
- ii. Does the information indicate that the population mean time lost due to hot tempers is different (either way) from population mean time lost due to disputes arising from technical workers' superior attitudes? Use $\alpha = 0.05$. Assume that the two lost-time population distributions are mound-shaped and symmetric.

26. **Management: Intimidators and Stressors** This problem is based on information regarding productivity in leading Silicon Valley companies (see reference in Problem 25). In large corporations, an "intimidator" is an employee who tries to stop communication, sometimes sabotages others, and, above all, likes to listen to him- or herself talk. Let x_1 be a random variable representing productive hours per week lost by peer employees of an intimidator.

$$x_1: \quad 8 \quad 3 \quad 6 \quad 2 \quad 2 \quad 5 \quad 2$$

A "stressor" is an employee with a hot temper that leads to unproductive tantrums in corporate society. Let x_2 be a random variable representing productive hours per week lost by peer employees of a stressor.

$$x_2: \quad 3 \quad 3 \quad 10 \quad 7 \quad 6 \quad 2 \quad 5 \quad 8$$

- i. Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 = 4.00$, $s_1 \approx 2.38$, $\bar{x}_2 = 5.5$, and $s_2 \approx 2.78$.
- ii. Assuming the variables x_1 and x_2 are independent, do the data indicate that the population mean time lost due to stressors is greater than the population mean time lost due to intimidators? Use a 5% level of significance. (Assume the population distributions of time lost due to intimidators and time lost due to stressors are each mound-shaped and symmetric.)



27. **Expand Your Knowledge: Software Approximation for Degrees of Freedom** Given x_1 and x_2 distributions that are normal or approximately normal with unknown σ_1 and σ_2 , the value of t corresponding to $\bar{x}_1 - \bar{x}_2$ has a distribution that is approximated by a Student's t distribution. We use the convention that

the degrees of freedom are approximately the smaller of $n_1 - 1$ and $n_2 - 1$. However, a more accurate estimate for the appropriate degrees of freedom is given by Satterthwaite's formula:

$$d.f. \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2} \right)^2}$$

where s_1 , s_2 , n_1 , and n_2 are the respective sample standard deviations and sample sizes of independent random samples from the x_1 and x_2 distributions. This is the approximation used by most statistical software. When both n_1 and n_2 are 5 or larger, it is quite accurate. The degrees of freedom computed from this formula are either truncated or not rounded.

- (a) In Problem 19, we tested whether the population average crime rate μ_2 in the Rocky Mountain region is higher than that in New England, μ_1 . The data were $n_1 = 10$, $\bar{x}_1 \approx 3.51$, $s_1 \approx 0.81$, $n_2 = 12$, $\bar{x}_2 \approx 3.87$ and $s_2 \approx 0.94$. Use Satterthwaite's formula to compute the degrees of freedom for the Student's t distribution.
- (b) When you did Problem 19, you followed the convention that degrees of freedom $d.f. = \text{smaller of } n_1 = 1 \text{ and } n_2 = 1$. Compare this value of $d.f.$ with that found with Satterthwaite's formula.



28.

Expand Your Knowledge: Pooled Two-Sample Procedure Consider independent random samples from two populations that are normal or approximately normal, or the case in which both sample sizes are at least 30. Then, if σ_1 and σ_2 are unknown but we have reason to believe that $\sigma_1 = \sigma_2$, we can pool the standard deviations. Using sample sizes n_1 and n_2 , the sample test statistic $\bar{x}_1 - \bar{x}_2$ has a Student's t distribution, where

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{with degrees of freedom } d.f. = n_1 + n_2 - 2$$

and where the pooled standard deviation s is

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Note: With statistical software, select the pooled variance or equal variance options.

- (a) There are many situations in which we want to compare means from populations having standard deviations that are equal. This method applies even if the standard deviations are known to be only approximately equal (see Section 10.4 for methods to test that $\sigma_1 = \sigma_2$). Consider Problem 23 regarding average incidence of fox rabies in two regions. For region I, $n_1 = 16$, $\bar{x}_1 = 4.75$, and $s_1 \approx 2.82$ and for region II, $n_2 = 15$, $\bar{x}_2 \approx 3.93$, and $s_2 \approx 2.43$. The two sample standard deviations are sufficiently close that we can assume $\sigma_1 = \sigma_2$. Use the method of pooled standard deviation to redo Problem 23, where we tested if there was a difference in population mean average incidence of rabies at the 5% level of significance.
- (b) Compare the t value calculated in part (a) using the pooled standard deviation with the t value calculated in Problem 23 using the unpooled standard deviation. Compare the degrees of freedom for the sample test statistic. Compare the conclusions.

29. **Federal Tax Money: Art Funding** Would you favor spending more federal tax money on the arts? This question was asked by a research group on behalf of *The National Institute* (Reference: *Painting by Numbers*, J. Wypijewski, University of California Press). Of a random sample of $n_1 = 220$ women, $r_1 = 59$ responded yes. Another random sample of $n_2 = 175$ men showed that $r_2 = 56$ responded yes. Does this information indicate a difference (either way) between the population proportion of women and the population proportion of men who favor spending more federal tax dollars on the arts? Use $\alpha = 0.05$.
30. **Art Funding: Politics** Would you favor spending more federal tax money on the arts? This question was asked by a research group on behalf of *The National Institute* (Reference: *Painting by Numbers*, J. Wypijewski, University of California Press). Of a random sample of $n_1 = 93$ politically conservative voters, $r_1 = 21$ responded yes. Another random sample of $n_2 = 83$ politically moderate voters showed that $r_2 = 22$ responded yes. Does this information indicate that the population proportion of conservative voters inclined to spend more federal tax money on funding the arts is less than the proportion of moderate voters so inclined? Use $\alpha = 0.05$.
31. **Sociology: High School Dropouts** This problem is based on information taken from *Life in America's Fifty States* by G. S. Thomas. A random sample of $n_1 = 153$ people ages 16 to 19 was taken from the island of Oahu, Hawaii, and 12 were found to be high school dropouts. Another random sample of $n_2 = 128$ people ages 16 to 19 was taken from Sweetwater County, Wyoming, and 7 were found to be high school dropouts. Do these data indicate that the population proportion of high school dropouts on Oahu is different (either way) from that of Sweetwater County? Use a 1% level of significance.
32. **Political Science: Voters** A random sample of $n_1 = 228$ voters registered in the state of California showed that 141 voted in the last general election. A random sample of $n_2 = 216$ registered voters in the state of Colorado showed that 125 voted in the most recent general election. (See reference in Problem 31.) Do these data indicate that the population proportion of voter turnout in Colorado is higher than that in California? Use a 5% level of significance.
33. **Extraterrestrials: Believe It?** Based on information from *Harper's Index*, $r_1 = 37$ people out of a random sample of $n_1 = 100$ adult Americans who did not attend college believe in extraterrestrials. However, out of a random sample of $n_2 = 100$ adult Americans who did attend college, $r_2 = 47$ claim that they believe in extraterrestrials. Does this indicate that the proportion of people who attended college and who believe in extraterrestrials is higher than the proportion who did not attend college but believe in extraterrestrials? Use $\alpha = 0.01$.
34. **Art: Politics** Do you prefer paintings in which the people are fully clothed? This question was asked by a professional survey group on behalf of the National Arts Society (see reference in Problem 30). A random sample of $n_1 = 59$ people who are conservative voters showed that $r_1 = 45$ said yes. Another random sample of $n_2 = 62$ people who are liberal voters showed that $r_2 = 36$ said yes. Does this indicate that the population proportion of conservative voters who prefer art with fully clothed people is higher than that of liberal voters? Use $\alpha = 0.05$.
35. **Sociology: Trusting People** Generally speaking, would you say that most people can be trusted? A random sample of $n_1 = 250$ people in Chicago ages 18–25 showed that $r_1 = 45$ said yes. Another random sample of $n_2 = 280$ people in Chicago ages 35–45 showed that $r_2 = 71$ said yes (based on information from the *National Opinion Research Center*, University of Chicago). Does this indicate that the population proportion of trusting people in Chicago is higher for the older group? Use $\alpha = 0.05$.

36. **Critical Region Method: Testing** Redo Problem 15
using the critical region method and compare your results to those obtained using the *P*-value method.
37. **Critical Region Method: Testing** Redo Problem 19
using the critical region method and compare your results to those obtained using the *P*-value method.
38. **Critical Region Method: Testing $p_1 - p_2$** Redo Problem 29 using the critical region method and compare your results to those obtained using the *P*-value method.



Chapter Review

SUMMARY

Hypothesis testing is a major component of inferential statistics. In hypothesis testing, we propose a specific value for the population parameter in question. Then we use sample data from a random sample and probability to determine whether or not to reject this specific value for the parameter.

Basic components of a hypothesis test are:

- The *null hypothesis* H_0 states that a parameter equals a specific value.
- The *alternate hypothesis* H_1 states that the parameter is greater than, less than, or simply not equal to the value specified in H_0 .
- The *level of significance* α of the test is the probability of rejecting H_0 when it is true.
- The *sample test statistic* corresponding to the parameter in H_0 is computed from a random sample and appropriate sampling distribution.
- Assuming H_0 is true, the probability that a sample test statistic will take on a value as extreme as, or more extreme than, the observed sample test statistic is the *P-value* of the test. The *P*-value is computed by using the sample test statistic, the corresponding sampling distribution, H_0 , and H_1 .
- If P -value $\leq \alpha$, we reject H_0 . If P -value $> \alpha$, we fail to reject H_0 .

- We say that sample data are *significant* if we can reject H_0 .

An alternative way to conclude a test of hypotheses is to use critical regions based on the alternate hypothesis and α . Critical values z_0 are found in Table 5(c) of Appendix II. Critical values t_0 are found in Table 6 of Appendix II. If the sample test statistic falls beyond the critical values—that is, in the critical region—we reject H_0 .

The methods of hypothesis testing are very general, and we will see them used again in later chapters. In this chapter, we looked at tests involving

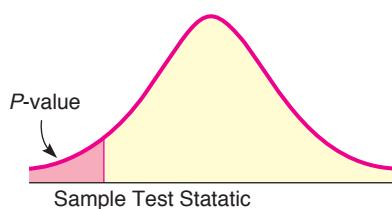
- Parameter μ . Use standard normal or Student's *t* distribution. See procedure displays in Section 8.2.
- Parameter p . Use standard normal distribution. See procedure displays in Section 8.3.
- Paired difference test for difference of means from dependent populations. Use Student's *t* distribution. See procedure displays in Section 8.4.
- Parameter $\mu_1 - \mu_2$ from independent populations. Use standard normal or Student's *t* distribution. See procedure displays in Section 8.5.
- Parameter $p_1 - p_2$ from independent populations. Use standard normal distribution. See procedure displays in Section 8.5.

FINDING THE P-VALUE CORRESPONDING TO A SAMPLE TEST STATISTIC

Use the appropriate sampling distribution as described in procedure displays for each of the various tests.

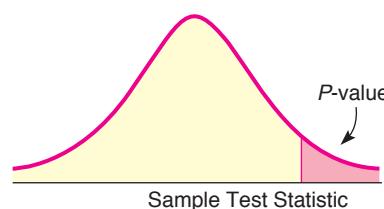
Left-Tailed Test

P -value = area to the left of the sample test statistic.



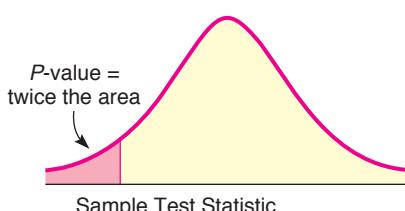
Right-Tailed Test

P -value = area to the right of the sample test statistic.

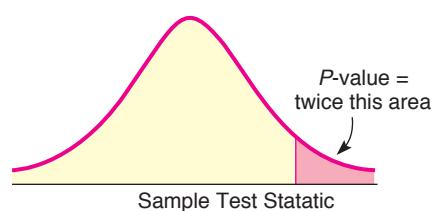


Two-Tailed Test

Sample test statistic lies to *left* of center P -value
= twice the area to the left of sample test statistic



Sample test statistic lies to *right* of center P -value
= twice area to the right of sample test statistic



Sampling Distributions for Inferences Regarding μ or p

Parameter	Condition	Sampling Distribution
μ	σ is known and x has a normal distribution or $n \geq 30$	Normal Distribution
μ	σ is not known and x has a normal or mound-shaped symmetric distribution or $n \geq 30$	Student's <i>t</i> Distribution with $d.f. = n - 1$
p	$np > 5$ and $n(1 - p) > 5$	Normal Distribution

IMPORTANT WORDS & SYMBOLS

Section 8.1

- Hypothesis testing 411
- Hypothesis 411
- Null hypothesis H_0 411
- Alternate hypothesis H_1 411
- right-tailed test 412
- two-tailed test 412
- left-tailed test 412
- sample test statistic 413
- P -value 414
- Type I error 416
- Type II error 416
- Level of significance of a test and the probability of a type I error, α 416
- Probability of a type II error, β 417
- Power of a test ($1 - \beta$) 417
- statistical significance 418

Section 8.2

- $d.f.$ for testing μ when σ is unknown 428
- Critical region method 432
- Critical value 433

Section 8.3

- Criteria for using normal approximation to binomial, $np > 5$ and $nq > 5$ 442

Section 8.4

- Paired data 452
- Dependent samples 452

Section 8.5

- Independent samples 466
- $d.f.$ for testing $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown 472
- Pooled standard deviation 475
- Pooled estimate of proportion \bar{p} 476

VIEWPOINT**Will It Rain?**

Do cloud-seed experiments ever work? If you seed the clouds, will it rain? If it does rain, who will benefit? Who will be displeased by the rain? If you seed the clouds and nothing happens, will taxpayers (who support the effort) complain or rejoice? Maybe this should be studied over a remote island—such as Tasmania (near Australia). Using what you already know about statistical testing, you can conduct your own tests, given the appropriate data. Remember, there are sociological questions (pleased/displeased with result) as well as technical questions (number of inches of rain produced). For data regarding cloud-seeding experiments over Tasmania, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to DASL, the Carnegie Mellon University Data and Story Library. From the DASL site, look under Datasets for Cloud.

**CHAPTER REVIEW
PROBLEMS**

1. **Statistical Literacy** When testing μ or the difference of means $\mu_1 - \mu_2$ from independent populations, how do we decide whether to use the standard normal distribution or a Student's t distribution?
2. **Statistical Literacy** What do we mean when we say a test is *significant*? Does this necessarily mean the results are important?
3. **Critical Thinking** All other conditions being equal, does a larger sample size increase or decrease the corresponding magnitude of the z or t value of the sample test statistic?
4. **Critical Thinking** All other conditions being equal, does a z or t value with larger magnitude have a larger or smaller corresponding P -value?

Before you solve each problem below, first categorize it by answering the following question: Are we testing a single mean, a difference of means, a paired difference, a single proportion, or a difference of proportions? Assume underlying population distributions are mound-shaped and symmetric for problems with small samples that involve testing a mean or difference of means. Then provide the following information for Problems 5–18.

- (a) What is the level of significance? State the null and alternate hypotheses.
- (b) **Check Requirements** What sampling distribution will you use? What assumptions are you making? What is the value of the sample test statistic?
- (c) Find (or estimate) the P -value. Sketch the sampling distribution and show the area corresponding to the P -value.
- (d) Based on your answers in parts (a) to (c), will you reject or fail to reject the null hypothesis? Are the data statistically significant at level α ?
- (e) **Interpret** your conclusion in the context of the application.

Note: For degrees of freedom $d.f.$ not in the Student's t table, use the closest $d.f.$ that is *smaller*. In some situations, this choice of $d.f.$ may increase the P -value by a small amount and therefore produce a slightly more “conservative” answer. Answers may vary due to rounding.

5. **Vehicles: Mileage** Based on information in *Statistical Abstract of the United States* (116th Edition), the average annual miles driven per vehicle in the United States is 11.1 thousand miles, with $\sigma \approx 600$ miles. Suppose that a random sample of 36 vehicles owned by residents of Chicago showed that the average mileage driven last year was 10.8 thousand miles. Does this indicate that the average miles driven per vehicle in Chicago is different from (higher or lower than) the national average? Use a 0.05 level of significance.
6. **Student Life: Employment** Professor Jennings claims that only 35% of the students at Flora College work while attending school. Dean Renata thinks that

- the professor has underestimated the number of students with part-time or full-time jobs. A random sample of 81 students shows that 39 have jobs. Do the data indicate that more than 35% of the students have jobs? (Use a 5% level of significance.)
7. **Toys: Electric Trains** The Toylot Company makes an electric train with a motor that it claims will draw an average of only 0.8 ampere (A) under a normal load. A sample of nine motors was tested, and it was found that the mean current was $\bar{x} = 1.4\text{ A}$, with a sample standard deviation of $s = 0.41\text{ A}$. Do the data indicate that the Toylot claim of 0.8 A is too low? (Use a 1% level of significance.)
8. **Highways: Reflective Paint** The highway department is testing two types of reflecting paint for concrete bridge end pillars. The two kinds of paint are alike in every respect except that one is orange and the other is yellow. The orange paint is applied to 12 bridges, and the yellow paint is applied to 12 bridges. After a period of 1 year, reflectometer readings were made on all these bridge end pillars. (A higher reading means better visibility.) For the orange paint, the mean reflectometer reading was $\bar{x}_1 = 9.4$, with standard deviation $s_1 = 2.1$. For the yellow paint, the mean was $\bar{x}_2 = 6.9$, with standard deviation $s_2 = 2.0$. Based on these data, can we conclude that the yellow paint has less visibility after 1 year? (Use a 1% level of significance.)
9. **Medical: Plasma Compress** A hospital reported that the normal death rate for patients with extensive burns (more than 40% of skin area) has been significantly reduced by the use of new fluid plasma compresses. Before the new treatment, the mortality rate for extensively burned patients was about 60%. Using the new compresses, the hospital found that only 40 of 90 patients with extensive burns died. Use a 1% level of significance to test the claim that the mortality rate has dropped.
10. **Bus Lines: Schedules** A comparison is made between two bus lines to determine if arrival times of their regular buses from Denver to Durango are off schedule by the same amount of time. For 51 randomly selected runs, bus line A was observed to be off schedule an average time of 53 minutes, with standard deviation 19 minutes. For 60 randomly selected runs, bus line B was observed to be off schedule an average of 62 minutes, with standard deviation 15 minutes. Do the data indicate a significant difference in average off-schedule times? Use a 5% level of significance.
11. **Matches: Number per Box** The Nero Match Company sells matchboxes that are supposed to have an average of 40 matches per box, with $\sigma = 9$. A random sample of 94 Nero matchboxes shows the average number of matches per box to be 43.1. Using a 1% level of significance, can you say that the average number of matches per box is more than 40?
12. **Magazines: Subscriptions** A study is made of residents in Phoenix and its suburbs concerning the proportion of residents who subscribe to *Sporting News*. A random sample of 88 urban residents showed that 12 subscribed, and a random sample of 97 suburban residents showed that 18 subscribed. Does this indicate that a higher proportion of suburban residents subscribe to *Sporting News*? (Use a 5% level of significance.)
13. **Archaeology: Arrowheads** The Wind Mountain archaeological site is in southwest New Mexico. Prehistoric Native Americans called Anasazi once lived and hunted small game in this region. A stemmed projectile point is an arrowhead that has a notch on each side of the base. Both stemmed and stemless projectile points were found at the Wind Mountain site. A random sample of $n_1 = 55$ stemmed projectile points showed the mean length to be $\bar{x}_1 = 3.0\text{ cm}$, with sample standard deviation $s_1 = 0.8\text{ cm}$. Another random sample of $n_2 = 51$ stemless projectile

points showed the mean length to be $\bar{x}_2 = 2.7$ cm, with $s_2 = 0.9$ cm (Source: *Mimbres Mogollon Archaeology*, by A. I. Woosley and A. J. McIntyre, University of New Mexico Press). Do these data indicate a difference (either way) in the population mean length of the two types of projectile points? Use a 5% level of significance.

14. **Civil Service: College Degrees** The Congressional Budget Office reports that 36% of federal civilian employees have a bachelor's degree or higher (*The Wall Street Journal*). A random sample of 120 employees in the private sector showed that 33 have a bachelor's degree or higher. Does this indicate that the percentage of employees holding bachelor's degrees or higher in the private sector is less than that in the federal civilian sector? Use $\alpha = 0.05$.
15. **Vending Machines: Coffee** A machine in the student lounge dispenses coffee. The average cup of coffee is supposed to contain 7.0 ounces. Eight cups of coffee from this machine show the average content to be 7.3 ounces with a standard deviation of 0.5 ounce. Do you think that the machine has slipped out of adjustment and that the average amount of coffee per cup is different from 7 ounces? Use a 5% level of significance.
16. **Psychology: Creative Thinking** Six sets of identical twins were randomly selected from a population of identical twins. One child was taken at random from each pair to form an experimental group. These children participated in a program designed to promote creative thinking. The other child from each pair was part of the control group that did not participate in the program to promote creative thinking. At the end of the program, a creative problem-solving test was given, with the results shown in the following table:

Twin pair	A	B	C	D	E	F
Experimental group	53	35	12	25	33	47
Control group	39	21	5	18	21	42

Higher scores indicate better performance in creative problem solving. Do the data support the claim that the program of the experimental group did promote creative problem solving? (Use $\alpha = 0.01$.)

17. **Marketing: Sporting Goods** A marketing consultant was hired to visit a random sample of five sporting goods stores across the state of California. Each store was part of a large franchise of sporting goods stores. The consultant taught the managers of each store better ways to advertise and display their goods. The net sales for 1 month before and 1 month after the consultant's visit were recorded as follows for each store (in thousands of dollars):

Store	1	2	3	4	5
Before visit	57.1	94.6	49.2	77.4	43.2
After visit	63.5	101.8	57.8	81.2	41.9

Do the data indicate that the average net sales improved? (Use $\alpha = 0.05$.)

18. **Sports Car: Fuel Injection** The manufacturer of a sports car claims that the fuel injection system lasts 48 months before it needs to be replaced. A consumer group tests this claim by surveying a random sample of 10 owners who had the fuel injection system replaced. The ages of the cars at the time of replacement were (in months):

29 42 49 48 53 46 30 51 42 52

- i. Use your calculator to verify that the mean age of a car when the fuel injection system fails is $\bar{x} = 44.2$ months, with standard deviation $s \approx 8.61$ months.

- ii. Test the claim that the fuel injection system lasts less than an average of 48 months before needing replacement. Use a 5% level of significance.

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

1. “With Sampling, There Is Too a Free Lunch”—This is a headline that appeared in the *Wall Street Journal*. The article is about food product samples available at grocery stores. Giving out food samples is expensive and labor-intensive. It clogs supermarket aisles. It is risky. What if a customer tries an item and spits it out on the floor or says the product is awful? It creates litter. Some customers drop toothpicks or small paper cups on the floor or spill the product. However, the budget that companies are willing to spend to have their products sampled is growing. The director of communications for Bigg’s “hypermarket” (a combination grocery and general-merchandise store) says that more than 60% of customers sample products and about 37% of those who sample buy the product.
 - (a) Let’s test the hypothesis that 60% of customers sample a particular product. What is the null hypothesis? Do you believe that the percentage of customers who sample products is less than, more than, or just different from 60%? What will you use for the alternate hypothesis?
 - (b) Choose a level of significance α .
 - (c) Go to a grocery store when special products are being sampled (not just the usual in-house store samples often available at the deli or bakery). Count the number of customers going by the display when a sample is available and the number of customers who try the sample. Be sure the number of customers n is large enough to use the normal distribution to approximate the binomial.
 - (d) Using your sample data, conclude the hypothesis test. What is your conclusion?
 - (e) Do you think different food products might have a higher or lower percentage of customers trying them? For instance, does a higher percentage of customers try samples of pizza than samples of yogurt? How could you use statistics to justify your answer?
 - (f) Do you want to include young children in your sample? Do they pick up items to include in the customer’s basket, or do they just munch the samples?
2. “Sweets May Not Be Culprit in Hyper Kids”—This is a *USA Today* headliner reporting results of a study that appeared in the *New England Journal of Medicine*. In this study, the subjects were 25 normal preschoolers, aged 3 to 5, and 23 kids, aged 6 to 10, who had been described as “sensitive to sugar.” The kids and their families were put on three different diets for 3 weeks each. One

diet was high in sugar, one was low in sugar and contained aspartame, and one was low in sugar and contained saccharin. The diets were all free of additives, artificial food coloring, preservatives, and chocolate. All food in the household was removed, and the meals were delivered to the families. Researchers gathered information about the kids' behavior from parents, babysitters, and teachers. In addition, researchers tested the kids for memory, concentration, reading, and math skills. The result: "We couldn't find any difference in terms of their behavior or their learning on any of the three diets," says Mark Wolraich, professor of pediatrics at Vanderbilt University Medical Center who oversaw the project. In another interview, Dr. Wolraich is quoted as saying, "Our study would say there is no evidence sugar has an adverse effect on children's behavior."

- (a) This research involved comparing several means, not just two. (An introduction to such methods, called *analysis of variance*, is found in Chapter 10.) However, let us take a simplified view of the problem and consider the difference of behavior when children consumed the diet with sugar compared with their behavior when they consumed the diet with aspartame and low sugar. List some variables that might be measured to reflect the behavior of the children.
- (b) Let's assume that the general null hypothesis was that there is no difference in children's behavior when they have a diet high in sugar. Was the evidence sufficient to allow the researchers to reject the null hypothesis and conclude that there are differences in children's behavior when they have a diet high in sugar? When we cannot reject H_0 , have we *proved* that H_0 is true? In your own words, paraphrase the comments made by Dr. Wolraich.

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

The most important questions in life usually cannot be answered with absolute certainty. Many important questions are answered by giving an estimate and a measure of confidence in the estimate. This was the focus of Chapter 7. However, sometimes important questions must be answered in a more straightforward manner by a simple yes or no. Hypothesis testing is the statistical process of answering questions with a straightforward yes or no *and* providing an estimate of the risk in accepting the answer.

1. Review and discuss type I and type II errors associated with hypothesis testing.
2. Review and discuss the level of significance and power of a statistical test.
3. The following statements are very important. Give them some careful thought and discuss them.
 - (a) When we fail to reject the null hypothesis, we do not claim that it is absolutely true. We simply claim that at the given level of significance, the data were not sufficient to reject the null hypothesis.
 - (b) When we accept the alternate hypothesis, we do not claim that the null hypothesis is absolutely false. We do claim that at the given level of significance, the data presented enough evidence to reject the null hypothesis.
4. In the text, it is said that a statistical test is a package of five basic ingredients. List these ingredients, discuss them in class, and write a short description of how these ingredients relate to the above discussion questions.
5. As access to computers becomes more and more prevalent, we see the *P*-value reported in hypothesis testing more frequently. Review the use of the *P*-value in hypothesis testing. What is the difference between the level of significance of a test and the *P*-value? Considering both the *P*-value and level of significance, under what conditions do we reject or fail to reject the null hypothesis?

USING TECHNOLOGY

Simulation

Recall that the level of significance α is the probability of mistakenly rejecting a true null hypothesis. If $\alpha = 0.05$, then we expect to mistakenly reject a true null hypothesis about 5% of the time. The following simulation conducted with Minitab demonstrates this phenomenon.

We draw 40 random samples of size 50 from a population that is normally distributed with mean $\mu = 30$ and standard deviation $\sigma = 2.5$. The display shows the results of a hypothesis test with

$$H_0: \mu = 30 \quad H_1: \mu > 30$$

for each of the 40 samples labeled C1 through C40. Because each of the 40 samples is drawn from a population with mean $\mu = 30$, the null hypothesis $H_0: \mu = 30$ is true for the test based on each sample. However, as the display shows, for some samples we reject the true null hypothesis.

- How many of the 40 samples have a sample mean \bar{x} above $\mu = 30$? below $\mu = 30$?
- Look at the P -value of the sample statistic \bar{x} in each of the 40 samples. How many P -values are less than or equal to $\alpha = 0.05$? What percent of the P -values are less than or equal to α ? What percent of the samples have us reject H_0 when, in fact, each of the samples was drawn from a normal distribution with $\mu = 30$, as hypothesized in the null hypothesis?
- If you have access to computer or calculator technology that creates random samples from a normal distribution with a specified mean and standard deviation, repeat this simulation. Do you expect to get the same results? Why or why not?

Minitab Display: Random samples of size 50 from a normal population with $\mu = 30$ and $\sigma = 2.5$

z-Test						
	Test of mu = 30.000 vs. mu > 30.000	Variable	N	Mean	StDev	SE Mean
C1	The assumed sigma = 2.50	C1	50	30.002	2.776	0.354
C2		C2	50	30.120	2.511	0.354
C3		C3	50	30.032	2.721	0.354
C4		C4	50	30.504	2.138	0.354
C5		C5	50	29.901	2.496	0.354
C6		C6	50	30.059	2.836	0.354
C7		C7	50	30.443	2.519	0.354
C8		C8	50	29.775	2.530	0.354
C9		C9	50	30.188	2.204	0.354
C10		C10	50	29.907	2.302	0.354
C11		C11	50	30.036	2.762	0.354
C12		C12	50	30.656	2.399	0.354
C13		C13	50	30.158	2.884	0.354
C14		C14	50	29.830	3.129	0.354
C15		C15	50	30.308	2.241	0.354
C16		C16	50	29.751	2.165	0.354
C17		C17	50	29.833	2.358	0.354
C18		C18	50	29.741	2.836	0.354
C19		C19	50	30.441	2.194	0.354
C20		C20	50	29.820	2.156	0.354
C21		C21	50	29.611	2.360	0.354
C22		C22	50	30.569	2.659	0.354
C23		C23	50	30.294	2.302	0.354
C24		C24	50	29.978	2.298	0.354
C25		C25	50	29.836	2.438	0.354
C26		C26	50	30.102	2.322	0.354
C27		C27	50	30.066	2.266	0.354
C28		C28	50	29.071	2.219	0.354
C29		C29	50	30.597	2.426	0.354
C30		C30	50	30.092	2.296	0.354
C31		C31	50	29.803	2.495	0.354
C32		C32	50	29.546	2.335	0.354
C33		C33	50	29.702	1.902	0.354
C34		C34	50	29.233	2.657	0.354
C35		C35	50	30.097	2.472	0.354
C36		C36	50	29.733	2.588	0.354
C37		C37	50	30.379	2.976	0.354
C38		C38	50	29.424	2.827	0.354
C39		C39	50	30.288	2.396	0.354
C40		C40	50	30.195	3.051	0.354

Technology Hints

TI-84Plus/TI-83Plus/TI-nspire

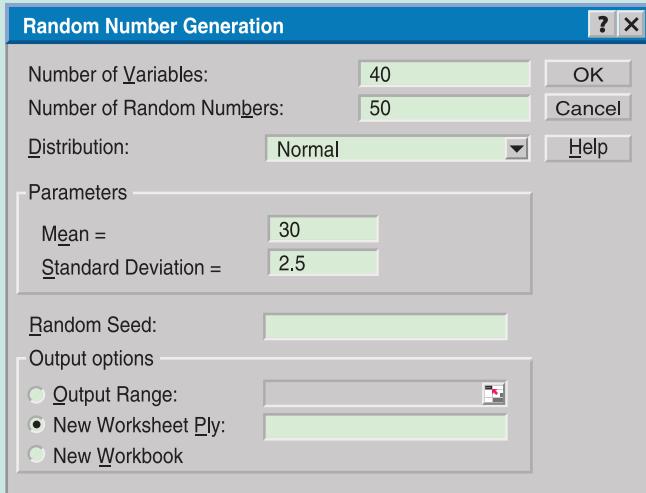
(with TI-84Plus keypad)

Press STAT and select EDIT. Highlight the list name, such as L1. Then press MATH, select PRB, and highlight 6:randNorm(μ , σ , sample size). Press enter. Fill in the values of $\mu = 30$, $\sigma = 2.5$, and sample size = 50. Press enter. Now list L1 contains a random sample from the normal distribution specified.

To test the hypothesis $H_0: \mu = 30$ against $H_1: \mu > 30$, press STAT, select TESTS, and use option 1:Z-Test. Fill in the value 30 for μ_0 , 2.5 for σ , and $> \mu_0$. The output provides the value of the sample statistic \bar{x} , its corresponding z value, and the P -value of the sample statistic.

Excel 2007

To draw random samples from a normal distribution with $\mu = 30$ and $\sigma = 2.5$, click the Data tab on the home ribbon and select Data Analysis in the Analysis Group. In the dialogue box, select Random Number Generator. In the first dialogue box, select Normal for the Distribution. In the resulting dialogue box shown below, the number of variables is the number of samples. Fill in the rest of the dialogue box as shown.



To conduct a hypothesis test of $H_0: \mu = 30$ against $H_1: \mu > 30$, click the Insert Function (f_x) on the home screen. For the category, select Statistical and then select the function ZTEST. Fill in the dialogue box with the array containing the random numbers, 30 for X, the value in H_0 , and 2.5 for sigma.

Minitab

To generate random samples from a normal distribution, use the menu choices Calc ► Random Data ► Normal. In the dialogue box, the number of rows refers to the sample size. Use 50 rows. Then designate the columns for the samples. Using C1–C40 will generate 40 random samples and put the samples in columns C1 through C40.

To test the hypothesis $H_0: \mu = 30$ against $H_1: \mu > 30$, use the menu choices Stat ► Basic Statistics ► 1-SampleZ. Use columns C1–C40 as the variables. Fill in 30 for the test mean, use “greater than” for the alternate hypothesis, and use 2.5 for sigma.

SPSS

SPSS uses a Student's t distribution to test the mean and difference of means. SPSS uses the sample standard deviation s even if the population σ is known. Use the menu choices Analyze ► Compare Means and then One-Sample T Test or Independent-Sample T Tests for tests of a single mean or a difference of means, respectively. In the dialogue box, fill in the test value of the null hypothesis.

To generate 40 random samples of size $n = 50$ from a normal distribution with $\mu = 30$ and $\sigma = 2.5$, first enter consecutive integers from 1 to 50 in a column of the data editor. Then, under variable view, enter the variable names Sample1 through Sample40. Use the menu choices Transform ► Compute Variable. In the dialogue box, use Sample1 for the target variable, select Random Numbers for the Function Group and then select the function Rv.Normal(mean, stddev). Use 30 for the mean and 2.5 for the standard deviation. Continue until you have 40 samples. To sample from other distributions, use appropriate functions in the Compute dialogue box.

The SPSS display shows the test results ($H_0: \mu = 30$; $H_1: \mu \neq 30$) for a sample of size $n = 50$ drawn from a normal distribution with $\mu = 30$ and $\sigma = 2.5$. The P -value is given as the significance for a two-tailed test.

For a one-tailed test, divide the significance by 2. In the display, the significance is 0.360 for a two-tailed test. So, for a one-tailed test, the P -value is $0.360/2$, or 0.180.

SPSS Display

T-Test				
One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
SAMPLE1	50	30.3228	2.46936	.34922

One-Sample Test						
	Test Value = 30					90% Confidence Interval of the Difference
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
SAMPLE1	.924	49	.360	.3228	-.2627	.9083



9

- 9.1 Scatter Diagrams and Linear Correlation
- 9.2 Linear Regression and the Coefficient of Determination
- 9.3 Inferences for Correlation and Regression
- 9.4 Multiple Regression



Portrait of René Descartes c. 1649 (after) Frans Hals/Louvre/Giraudon/The Bridgeman Art Library

When it is not in our power to determine what is true, we ought to follow what is most probable.

—RENÉ DESCARTES



Asia Images Group/Getty Images

It is important to realize that statistics and probability do not deal in the realm of certainty. If there is any realm of human knowledge where genuine certainty exists, you may be sure that our statistical methods are not needed there. In most human endeavors, and in almost all of the natural world around us, the element of chance happenings cannot be avoided. When we cannot expect something with true certainty, we must rely on probability to be our guide. In this chapter, we will study regression, correlation, and forecasting. One of the tools we use is a scatter plot. René Descartes (1596–1650) was the first mathematician to systematically use rectangular coordinate plots. For this reason, such a coordinate axis is called a Cartesian axis.

For online student resources, visit the Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>

CORRELATION AND REGRESSION

PREVIEW QUESTIONS

How can you use a scatter diagram to visually estimate the degree of linear correlation of two random variables? (SECTION 9.1)

How do you compute the correlation coefficient and what does it tell you about the strength of the linear relationship between two random variables? (SECTION 9.1)

What is the least-squares criterion? How do you find the equation of the least-squares line? (SECTION 9.2)

What is the coefficient of determination and what does it tell you about explained variation of y in a random sample of data pairs (x, y) ? (SECTION 9.2)

How do you determine if the sample correlation coefficient is statistically significant? (SECTION 9.3)

How do you find a confidence interval for predictions based on the least-squares model? (SECTION 9.3)

How do you test the slope β of the population least-squares line? How do you construct a confidence interval for β ? (SECTION 9.3)

What if you have more than two random variables? How do you construct a linear regression model for three, four, or more random variables? (SECTION 9.4)

David Young-Wolff/PhotoEdit



FOCUS PROBLEM

Changing Populations and Crime Rate

Is the crime rate higher in neighborhoods where people might not know each other very well? Is there a relationship between crime rate and population change? If so, can we make predictions based on such a relationship? Is the relationship statistically significant? Is it possible to predict crime rates from population changes?

Denver is a city that has had a lot of growth and consequently a lot of population change in recent years. Sociologists studying population changes and crime rates could find a wealth of information in Denver statistics. Let x be a random variable representing percentage change in neighborhood population in the past few years, and let y be a random variable representing crime rate (crimes per 1000 population). A random sample of

ifoto, 2010/Used under license from Shutterstock.com



six Denver neighborhoods gave the following information (Source: *Neighborhood Facts*, The Piton Foundation). To find out more about the Piton Foundation, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Piton Foundation.

x	29	2	11	17	7	6
y	173	35	132	127	69	53

Using information presented in this chapter, you will be able to analyze the relationship between the variables x and y using the following tools.

- Scatter diagram
- Sample correlation coefficient and coefficient of determination
- Least-squares line equation
- Predictions for y using the least-squares line
- Tests of population correlation coefficient and of slope of least-squares line
- Confidence intervals for slope and for predictions

(See Problem 10 in the Chapter Review Problems.)

SECTION 9.1

Scatter Diagrams and Linear Correlation

FOCUS POINTS

- Make a scatter diagram.
- Visually estimate the location of the “best-fitting” line for a scatter diagram.
- Use sample data to compute the sample correlation coefficient r .
- Investigate the meaning of the correlation coefficient r .

Paired data values

Studies of correlation and regression of two variables usually begin with a graph of *paired data values* (x, y). We call such a graph a *scatter diagram*.

Scatter diagram

A **scatter diagram** is a graph in which data pairs (x, y) are plotted as individual points on a grid with horizontal axis x and vertical axis y . We call x the **explanatory variable** and y the **response variable**.

Explanatory variable
Response variable

By looking at a scatter diagram of data pairs, you can observe whether there seems to be a linear relationship between the x and y values.



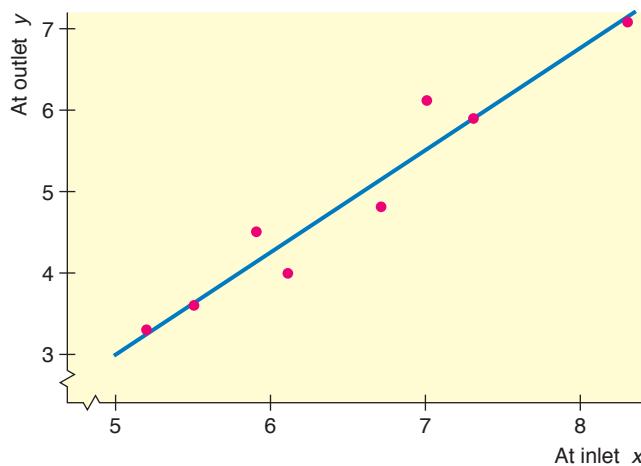
EXAMPLE 1

SCATTER DIAGRAM

Phosphorous is a chemical used in many household and industrial cleaning compounds. Unfortunately, phosphorous tends to find its way into surface water, where it can kill fish, plants, and other wetland creatures. Phosphorous-reduction programs are required by law and are monitored by the Environmental Protection Agency (EPA) (Reference: *EPA Case Study 832-R-93-005*).

A random sample of eight sites in a California wetlands study gave the following information about phosphorous reduction in drainage water. In this study, x is a random variable that represents phosphorous concentration (in 100 mg/l) at the inlet of a passive biotreatment facility, and y is a random

FIGURE 9-1
Phosphorous Reduction (100 mg/l)



variable that represents total phosphorous concentration (in 100 mg/l) at the outlet of the passive biotreatment facility.

<i>x</i>	5.2	7.3	6.7	5.9	6.1	8.3	5.5	7.0
<i>y</i>	3.3	5.9	4.8	4.5	4.0	7.1	3.6	6.1

(a) Make a scatter diagram for these data.

SOLUTION: Figure 9-1 shows points corresponding to the given data pairs. These plotted points constitute the scatter diagram. To make the diagram, first scan the data and decide on an appropriate scale for each axis. Figure 9-1 shows the scatter diagram (points) along with a line segment showing the basic trend. Notice a “jump scale” on both axes.

(b) *Interpretation* Comment on the relationship between x and y shown in Figure 9-1.

SOLUTION: By inspecting the figure, we see that smaller values of x are associated with smaller values of y and larger values of x tend to be associated with larger values of y . Roughly speaking, the general trend seems to be reasonably well represented by an upward-sloping line segment, as shown in the diagram.

Cindy Kassab/Corbis Edge/Corbis



Introduction to linear correlation

No linear correlation

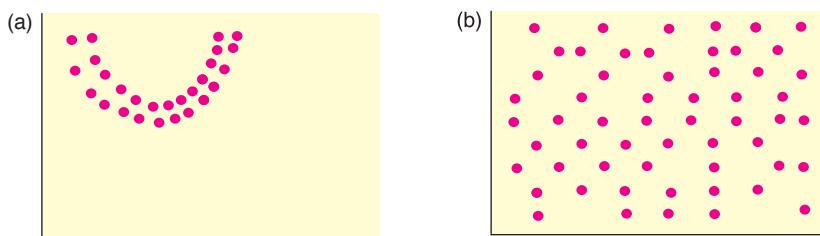
Of course, it is possible to draw many curves close to the points in Figure 9-1, but a straight line is the simplest and most widely used in elementary studies of paired data. We can draw many lines in Figure 9-1, but in some sense, the “best” line should be the one that comes closest to each of the points of the scatter diagram. To single out one line as the “best-fitting line,” we must find a mathematical criterion for this line and a formula representing the line. This will be done in Section 9.2 using the *method of least squares*.

Another problem precedes that of finding the “best-fitting line.” That is the problem of determining how well the points of the scatter diagram are suited for fitting *any* line. Certainly, if the points are a very poor fit to any line, there is little use in trying to find the “best” line.

If the points of a scatter diagram are located so that *no* line is realistically a “good” fit, we then say that the points possess *no linear correlation*. We see some examples of scatter diagrams for which there is no linear correlation in Figure 9-2.

FIGURE 9-2

Scatter Diagrams with No Linear Correlation



GUIDED EXERCISE 1

Scatter diagram

A large industrial plant has seven divisions that do the same type of work. A safety inspector visits each division of 20 workers quarterly. The number x of work-hours devoted to safety training and the number y of work-hours lost due to industry-related accidents are recorded for each separate division in Table 9-1.



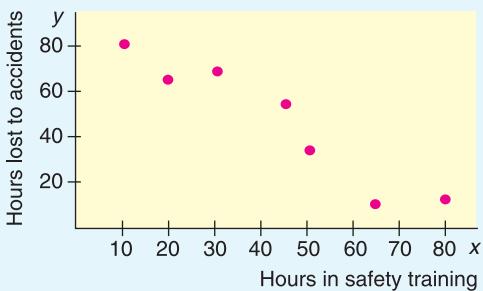
TABLE 9-1 Safety Report

Division	x	y
1	10.0	80
2	19.5	65
3	30.0	68
4	45.0	55
5	50.0	35
6	65.0	10
7	80.0	12

- (a) Make a scatter diagram for these pairs. Place the x values on the horizontal axis and the y values on the vertical axis.



FIGURE 9-3 Scatter Diagram for Safety Report



- (b) As the number of hours spent on safety training increases, what happens to the number of hours lost due to industry-related accidents?



In general, as the number of hours in safety training goes up, the number of hours lost due to accidents goes down.

- (c) **Interpretation** Does a line fit the data reasonably well?



A line fits reasonably well.

- (d) Draw a line that you think "fits best."



Use a downward-sloping line that lies close to the points. Later, you will find the equation of the line that is a "best fit."

Perfect linear correlation

If the points seem close to a straight line, we say the linear correlation is moderate to high, depending on how close the points lie to the line. If all the points do, in fact, lie on a line, then we have *perfect linear correlation*. In Figure 9-4, we see some diagrams with perfect linear correlation. In statistical applications, perfect linear correlation almost never occurs.

Positive correlation

The variables x and y are said to have *positive correlation* if low values of x are associated with low values of y and high values of x are associated with high