SUPPLEMENTARY EXERCISES

for

THE PRACTICE OF STATISTICS FOR BUSINESS AND ECONOMICS

Third Edition

David S. Moore
George P. McCabe
Layth C. Alwan
Bruce A. Craig
and
William M. Duckworth

These exercises appeared in the first, or second editions of *The Practice of Business Statistics*. They do not appear in the third edition, but they remain high-quality exercises that supplement those in the text.

CHAPTER 1

Section 1.1

1.1 Motor vehicle fuel economy. Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 2001 Model Year motor vehicles:

Make and model	Vehicle type	Transmission type	Number of cylinders	City Mpg	Highway Mpg
⋮					
BMW 330CI	Subcompact	Automatic	6	19	27
BMW 330CI	Subcompact	Manual	6	21	30
Buick Century	Midsize	Automatic	6	20	29
Chevrolet Blazer	Four-wheel drive	Automatic	6	15	20
⋮					

(a) What are the individuals in this data set?
(b) For each individual, what variables are given? Which of these variables are categorical and which are quantitative?

1.2 Data from a pharmaceutical company medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?
(a) Gender (female or male)
(b) Age (years)
(c) Race (Asian, black, white, or other)
(d) Smoker (yes or no)
(e) Systolic blood pressure (millimeters of mercury)
(f) Level of calcium in the blood (micrograms per milliliter)

1.3 Television-viewing habits. You are preparing to study the television-viewing habits of college students. Describe two categorical variables and two quantitative variables that you might measure for each student. Give the units of measurement for the quantitative variables. Also, what will the individuals in your data set be?

1.4 Protecting wood surfaces. How can we help wood surfaces resist weathering, especially when restoring historic wooden buildings? A study of this question prepared wooden panels and then exposed them to the weather. Here are some of the variables recorded. Which of these variables are categorical, and which are quantitative?
(a) Type of wood (yellow poplar, pine, cedar)
(b) Water repellent (solvent-based, water-based)
(c) Paint thickness (millimeters)

(d) Paint color (white, gray, light blue)
(e) Weathering time (months)

1.5 Undergraduate majors. Here are data on the percent of undergraduate majors for various colleges within a university (data from the Iowa State University Fact Book based on fall semester enrollment for 1999. Found online at `public.iastate.edu/~inst_res_info/factbk.html`):

Agriculture	13.9%
Business	17.0%
Design	8.2%
Education	8.5%
Engineering	21.0%
Liberal Arts and Sciences	29.9%

(a) Present these data in a well-labeled bar graph.
(b) Would it also be correct to use a pie chart to display these data? Explain your answer.

1.6 Occupational deaths. In 1999 there were 6023 job-related deaths in the United States. Among these were 807 deaths in agricultural-related jobs (including forestry and fishing), 121 in mining, 1190 in construction, 719 in manufacturing, 1006 in transportation and public utilities, 237 in wholesale trade, 507 in retail trade, 105 in finance-related jobs (including insurance and real estate), 732 in service-related jobs, and 562 in government jobs. (Occupational fatalities data from the Bureau of Labor Statistics Web site, `bls.gov`.)
(a) Find the percent of occupational deaths for each of these job categories, rounded to the nearest percent. What percent of job-related deaths were in categories not listed above?
(b) Make a well-labeled bar graph of the distribution of occupational deaths. Be sure to include an "other occupations" bar.
(c) Make a well-labeled Pareto chart of these data. What percent of all occupational deaths are accounted for by the first 3 categories in your Pareto chart?
(d) Would it also be correct to use a pie chart to display these data? Explain your answer.

1.7 Automobile fuel economy. Environmental Protection Agency regulations require automakers to give the city and highway gas mileages for each model of car. The following table gives the highway mileages (miles per gallon) for 32 midsize 2006 model cars. (Data from the U.S. Department of Energy, *Model Year 2006 Fuel Economy Guide*, `epa.gov/OMSWWW/`.) Make a histogram of the highway mileages of these cars.

Highway gas mileage for 2006 model midsize cars

Model	Mpg	Model	Mpg
Acura 3.5RL	26	Lexus GS300	30
Audi A6 Quattro	27	Lincoln-Mercury LS	25
BMW 550I	23	Lincoln-Mercury Milan	32
Buick Allure	30	Mazda 626	32
Cadillac CTS	27	Mercedes-Benz E320	37
Cadillac STS AWD	25	Mercedes-Benz E350	27
Chevrolet Malibu	32	Mercedes-Benz E55 AMG	21
Chrysler Sebring	30	Mitsubishi Galant	30
Dodge Stratus	30	Nissan Altima	29
Honda Accord	34	Nissan Maxima	28
Hyundai Elantra	32	Pontiac Grand Prix	28
Jaguar S	28	Rolls-Royce Phantom	19
Jaguar S/R	23	Suzuki Verona	27
Jaguar X	26	Toyota Camry	34
Kia Optima	30	Toyota Prius	51
Kia Spectra	34	Volvo S80	30

Indianapolis architectural firms

Name	1998 Total billings ($ millions)	1998 Arch. billings ($ millions)	1997 Arch. billings ($ millions)	Architects	Engineers	Staff
Schmidt Associates	11.5	11.5	5.0	19	7	111
CSO Architects	12.6	9.3	8.8	29	12	126
BSA Design	13.8	9.0	7.5	31	21	155
InterDesign Group	6.4	6.4	4.7	19	3	57
Browning Day Mullins	8.5	6.2	3.9	24	0	70
Ratio Architects	7.8	6.2	5.6	21	0	68
Odle McGuire	8.3	4.2	4.1	9	2	62
Gibraltar	5.8	3.6	4.2	12	4	52
American Consulting	12.0	3.5	3.3	5	23	131
Fanning/Howey	5.3	3.5	3.8	12	4	61
HNTB Corporation	15.0	3.4	3.0	10	35	110
Schenkel Schultz	2.7	2.7	2.5	5	0	22
Simmons & Associates	2.2	2.2	2.4	2	1	13
Paul I. Cripe	7.5	2.1	1.7	5	13	115
Plus4 Architects	2.7	2.1	2.0	5	0	15
Architectural Alliance	2.0	2.0	1.2	4	0	14
Blackburn Architects	2.6	1.8	1.5	8	1	24
Snapp & Associates	1.8	1.8	1.4	3	1	7
Sebree & Associates	1.7	1.7	1.0	3	0	15
Armstrong & Associates	9.1	1.6	2.4	3	23	96
Lamson & Condon	1.6	1.6	2.0	4	0	17
RQAW	6.3	1.6	2.3	6	14	72
Woollen Molzan	1.6	1.6	1.3	5	0	15
United Consulting	7.0	1.3	0.7	2	12	70
URS Greiner Woodward	1.7	1.3	0.9	5	1	17

1.8 Architectural firms. The preceeding table contains data describing firms engaged in commercial architecture in the Indianapolis, Indiana area. (From a table entitled "Largest Indianapolis-Area Architectural Firms," *Indianapolis Business Journal*, September 20–26, 1999.) One of the variables is the count of full-time staff members employed by each firm. Make a histogram of the staff counts.

1.9 Automobile fuel economy. The table in Exercise 1.7 gives data on the fuel economy of 2006 model midsize cars. Based on a histogram of these data:
(a) Describe the main features (shape, center, spread, outliers) of the distribution of highway mileage.
(b) The government imposes a "gas guzzler" tax on cars with low gas mileage. Which of these cars do you think are subject to the gas guzzler tax?

1.10 Architectural firms. The table in Exercise 1.8 gives the number of full-time staff employed by Indianapolis architectural firms. Make a stemplot of the staff counts. What are the main features of the shape of this distribution?

1.11 Supermarket shoppers. A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (in dollars), arranged from smallest to largest:

3.11	8.88	9.26	10.81	12.69	13.78	15.23	15.62	17.00	17.39
18.36	18.43	19.27	19.50	19.54	20.16	20.59	22.22	23.04	24.47
24.58	25.13	26.24	26.26	27.65	28.06	28.08	28.38	32.03	34.98
36.37	38.64	39.16	41.02	42.97	44.08	44.67	45.40	46.69	48.65
50.39	52.75	54.80	59.07	61.22	70.32	82.70	85.76	86.37	93.34

Round these amounts to the nearest dollar and then make a stemplot of these data. About where is the center of the distribution? Are there any outliers? What is the spread of the values (ignoring any outliers)? Is the distribution symmetric, skewed left, or skewed right? Make a second stemplot of the data by splitting the stems as described in this section.

1.12 Yields of Treasury bills. Treasury bills are short-term borrowing by the U.S. government. They are important in financial theory because the interest rate for Treasury bills is a "risk-free rate" that says what return investors can get while taking (almost) no risk. More risky investments should in theory offer higher returns in the long run. Here are the annual returns on Treasury bills from 1970 to 2004 (data from the Web site of Professor Kenneth French of Dartmouth, `mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`):

Year	Rate	Year	Rate	Year	Rate	Year	Rate
1970	6.52	1979	10.38	1988	6.36	1997	5.25
1971	4.39	1980	11.26	1989	8.38	1998	4.85
1972	3.84	1981	14.72	1990	7.84	1999	4.69
1973	6.93	1982	10.53	1991	5.60	2000	5.88
1974	8.01	1983	8.80	1992	3.50	2001	3.86
1975	5.80	1984	9.84	1993	2.90	2002	1.63
1976	5.08	1985	7.72	1994	3.91	2003	1.02
1977	5.13	1986	6.16	1995	5.60	2004	1.19
1978	7.19	1987	5.47	1996	5.20		

(a) Make a time plot of the returns paid by Treasury bills in these years.
(b) Interest rates, like many economic variables, show **cycles,** clear but irregular up-and-down movements. In which years did the interest rate cycle reach temporary peaks?
(c) A time plot may show a consistent trend underneath cycles. When did interest rates reach their overall peak during these years? Has there been a general trend downward since that year?

1.13 Vehicle colors. Favorite vehicle colors differ among types of vehicle. Here are data on the most popular colors in 2003 for luxury cars and for SUVs, trucks, and vans. (Press release for the 2003 DuPont Automotive Color Survey, at `automotive.dupont.com`.) The entry "—" means "less than 1%."

Color	Luxury car percent	SUV/truck/van percent
Black	10.9	11.6
Light brown	—	6.3
Medium/dark blue	3.8	9.3
Medium/dark gray	23.3	8.8
Medium/dark green	—	7.0
Medium red	3.9	6.2
White	30.4	22.3
Silver	18.8	17.0

(a) Make a bar graph for the luxury car percents.
(b) Make a bar graph for the SUV/truck/van percents.
(c) Now, be creative: make *one* bar graph that compares the two vehicle types as well as comparing colors. Arrange your graph so that it is easy to compare the two types of vehicle.

1.14 Types of email spam. Email spam is the curse of the Internet. Here is a compilation of the most common types of spam (Robyn Greenspan, "The deadly duo: spam and viruses, October 2003," at
`cyberatlas.internet.com`):

Type of spam	Percent
Adult	14.5
Financial	16.2
Health	7.3
Leisure	7.8
Products	21.0
Scams	14.2

(a) Make a bar graph of these percents with the bars ordered alphabetically (as in the table).
(b) Make another bar graph of these percents with the bars ordered from tallest to shortest.
(c) What is the name given to the bar graph in part (b)? What is the advantage of this type of bar graph?

1.15 Property damage by tornadoes. The states differ greatly in the kinds of severe weather that afflict them. The following table shows the average property damage caused by tornadoes per year over the period from 1950 to 1999 in each of the 50 states and Puerto Rico. (National Climactic Data Center, storm events database. Found at `sciencepolicy.colorado.edu/sourcebook/tornadoes.html`.) To adjust for the changing buying power of the dollar over time, all damages were restated in 1999 dollars.
(a) What are the top five states for tornado damage? The bottom five? (Include Puerto Rico, though it is not a state.)
(b) Make a histogram of the data, by hand or using software, with classes "$0 \leq$ damage < 10," "$10 \leq$ damage < 20," and so on. Describe the shape, center, and spread of the distribution. Which states may be outliers? (To understand the outliers, note that most tornadoes in largely rural states such as Kansas cause little property damage. Damage to crops is not counted as property damage.)
(c) If you are using software, also display the "default" histogram that your software makes when you give it no instructions. How does this compare with your graph in (b)?

Average property damage per year due to tornadoes

State	Damage ($ millions)	State	Damage ($ millions)	State	Damage ($ millions)
Alabama	51.88	Louisiana	27.75	Ohio	44.36
Alaska	0.00	Maine	0.53	Oklahoma	81.94
Arizona	3.47	Maryland	2.33	Oregon	5.52
Arkansas	40.96	Massachusetts	4.42	Pennsylvania	17.11
California	3.68	Michigan	29.88	Puerto Rico	0.05
Colorado	4.62	Minnesota	84.84	Rhode Island	0.09
Connecticut	2.26	Mississippi	43.62	South Carolina	17.19
Delaware	0.27	Missouri	68.93	South Dakota	10.64
Florida	37.32	Montana	2.27	Tennessee	23.47
Georgia	51.68	Nebraska	30.26	Texas	88.60
Hawaii	0.34	Nevada	0.10	Utah	3.57
Idaho	0.26	New Hampshire	0.66	Vermont	0.24
Illinois	62.94	New Jersey	2.94	Virginia	7.42
Indiana	53.13	New Mexico	1.49	Washington	2.37
Iowa	49.51	New York	15.73	West Virginia	2.14
Kansas	49.28	North Carolina	14.90	Wisconsin	31.33
Kentucky	24.84	North Dakota	14.69	Wyoming	1.78

Carbon dioxide emissions (metric tons per person)

Country	CO_2	Country	CO_2	Country	CO_2
Algeria	2.3	Iran	3.8	Poland	8.0
Argentina	3.9	Iraq	3.6	Romania	3.9
Australia	17.0	Italy	7.3	Russia	10.2
Bangladesh	0.2	Japan	9.1	Saudi Arabia	11.0
Brazil	1.8	Kenya	0.3	South Africa	8.1
Canada	16.0	Korea, North	9.7	Spain	6.8
China	2.5	Korea, South	8.8	Sudan	0.2
Colombia	1.4	Malaysia	4.6	Tanzania	0.1
Congo	0.0	Mexico	3.7	Thailand	2.5
Egypt	1.7	Morocco	1.0	Turkey	2.8
Ethiopia	0.0	Myanmar	0.2	Ukraine	7.6
France	6.1	Nepal	0.1	United Kingdom	9.0
Germany	10.0	Nigeria	0.3	United States	19.9
Ghana	0.2	Pakistan	0.7	Uzbekistan	4.8
India	0.9	Peru	0.8	Venezuela	5.1
Indonesia	1.2	Philippines	0.9	Vietnam	0.5

1.16 Carbon dioxide emissions. Burning fuels in power plants or motor vehicles emits carbon dioxide (CO_2), which contributes to global warming. The preceeding table displays CO_2 emissions per person from countries with populations of at least 20 million. (Found online at `earthtrends.wri.org`.)

(a) Why do you think we choose to measure emissions per person rather than total CO_2 emissions for each country?

(b) Display the data in a graph. Describe the shape, center, and spread of the distribution. Which countries are outliers?

1.17 Demographic: 65 and older. The population of the United States is aging, though less rapidly than in other developed countries. Here is a stemplot of the percents of residents aged 65 and older in the 50 states, according to the 2000 census. The stems are whole percents and the leaves are tenths of a percent.

5	7
6	
7	
8	5
9	679
10	6
11	02233677
12	0011113445789
13	00012233345568
14	034579
15	36
16	
17	6

(a) There are two outliers: Alaska has the lowest percent of older residents, and Florida has the highest. What are the percents for these two states?
(b) Ignoring Alaska and Florida, describe the shape, center, and spread of this distribution.

1.18 Demographic: 65 and older. Make another stemplot of the percent of residents aged 65 and older in the states other than Alaska and Florida by splitting stems 8 to 15 in the plot from the previous exercise. Which plot do you prefer? Why?

1.19 Tracking quality. The J. D. Power Initial Quality Study polls more than 50,000 buyers of new motor vehicles 90 days after their purchase. A two-page questionnaire asks about "things gone wrong." Here are data on problems per 100 vehicles for vehicles made by Toyota and by General Motors in recent years. Toyota has been the industry leader in quality. Make two time plots in the same graph to compare Toyota and GM. What are the most important conclusions you can draw from your graph?

Year:	1998	1999	2000	2001	2002	2003	2004
GM	187	179	164	147	130	134	120
Toyota	156	134	116	115	107	115	101

1.20 Mass layoffs. The Bureau of Labor Statistics says that a "mass layoff" occurs when 50 or more people from the same employer file initial claims for unemployment insurance in a five-week period. The following table gives data on the number of mass layoffs in the United States over a 10-year period. (From the Bureau of Labor Statistics Web site, `www.bls.gov/data`.)

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1995				1431	1194	1512	1502	680	706	1070	1282	2058
1996	1763	947	994	1030	942	924	1534	918	513	1289	1433	1824
1997	2139	755	783	1269	1152	1238	1899	973	548	1414	1156	1634
1998	2360	970	762	1253	1180	1208	2220	617	637	1553	1368	1776
1999	2421	1067	880	1270	1032	1140	1741	698	717	1098	1336	1509
2000	1934	1045	986	924	984	1597	1333	751	936	874	1697	2677
2001	1522	1501	1527	1450	1434	2107	2117	1490	1327	1831	2721	2440
2002	2146	1382	1460	1506	1723	1584	2042	1248	1062	1497	2153	2474
2003	2315	1363	1207	1581	1703	1691	2087	1258	868	1523	1438	1929
2004	2428	941	920	1458	988	1379	2094	809				

(a) Make a time plot of these data.
(b) There is strong seasonal variation, regular up-and-down movements that occur at about the same time each year. At what time of year are mass layoffs most common? Least common?
(c) Is there also a long-term trend in the number of mass layoffs? Explain your conclusion.

1.21 Motor vehicle fuel economy. Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 2001 model motor vehicles:

Make and model	Vehicle type	Transmission type	Number of cylinders	City Mpg	Highway Mpg
⋮					
BMW 330CI	Subcompact	Automatic	6	19	27
BMW 330CI	Subcompact	Manual	6	21	30
Buick Century	Midsize	Automatic	6	20	29
Chevrolet Blazer	Four-wheel drive	Automatic	6	15	20
⋮					

(a) What are the individuals in this data set?
(b) For each individual, what variables are given? Which of these variables are categorical and which are quantitative?

1.22 Pharmaceutical study. Data from a pharmaceutical company medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?
(a) Gender (female or male)
(b) Age (years)
(c) Race (Asian, black, white, or other)
(d) Smoker (yes or no)
(e) Systolic blood pressure (millimeters of mercury)
(f) Level of calcium in the blood (micrograms per milliliter)

1.23 Mutual funds. The following is a small part of a data set that describes mutual funds available to the public:

Fund	Category	Net assets ($ millions)	Year-to-date return	Largest holding
⋮				
Fidelity Low-Priced Stock	Small value	6,189	4.56%	Dallas Semiconductor
Price International Stock	International stock	9,745	−0.45%	Vodafone
Vanguard 500 Index	Large blend	89,394	3.45%	General Electric
⋮				

(a) What individuals does this data set describe?
(b) In addition to the fund's name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?
(c) What are the units of measurement for each of the quantitative variables?

1.24 Insurance: cause of death. Cause of death is an important issue for life insurance companies—especially identifying the most likely causes of death for various demographic subgroups of the population. The number of deaths among persons aged 15 to 24 years in the United States in 1997 due to the seven leading causes of death for this age group were accidents, 12,958; homicide, 5793; suicide, 4146; cancer, 1583; heart disease, 1013; congenital defects, 383; AIDS, 276. (Centers for Disease Control and Prevention, *Births and Deaths: Preliminary Data for 1997*, Monthly Vital Statistics Reports, 47, No. 4, 1998.)
(a) Make a bar graph to display these data.
(b) What additional information do you need to make a pie chart?

1.25 Location of a new facility. A company deciding where to build a new facility will rank potential locations in terms of how desirable it is to operate a business in each location. Describe five variables that you would measure for each location if you were designing a study to help your company rank locations. Give reasons for each of your choices.

1.26 The changing age distribution of the United States. The distribution of the ages of a nation's population has a strong influence on economic and social conditions. The following table shows the age distribution of U.S. residents in 1950 and 2075, in millions of people. The 1950 data come from that year's census, while the 2075 data are projections made by the Census Bureau.
(a) Because the total population in 2075 is much larger than the 1950 population, comparing percents in each age group is clearer than comparing counts. Make a table of the percent of the total population in each age group for both 1950 and 2075.
(b) Make a histogram with vertical scale in percents of the 1950 age distribution. Describe the main features of the distribution. In particular, look at the percent of children relative to the rest of the population.
(c) Make a histogram with vertical scale in percents of the projected age distribution

for the year 2075. Use the same scales as in (b) for easy comparison. What are the most important changes in the U.S. age distribution projected for the years between 1950 and 2075?

Age distribution in the United States, 1950 and 2075 (in millions of persons)

Age group	1950	2075
Under 10 years	29.3	53.3
10–19 years	21.8	53.2
20–29 years	24.0	51.2
30–39 years	22.8	50.5
40–49 years	19.3	47.5
50–59 years	15.5	44.8
60–69 years	11.0	40.7
70–79 years	5.5	30.9
80–89 years	1.6	21.7
90–99 years	0.1	8.8
100–109 years	—	1.1
Total	151.1	403.7

1.27 Reliability of household appliances. Always ask whether a particular variable is really a suitable measure for your purpose. You are writing an article for a consumer magazine based on a survey of the magazine's readers on the reliability of their household appliances. Of 13,376 readers who reported owning Brand A dishwashers, 2942 required a service call during the past year. Only 192 service calls were reported by the 480 readers who owned Brand B dishwashers.
(a) Why is the count of service calls (2942 versus 192) not a good measure of the reliability of these two brands of dishwashers?
(b) Use the information given to calculate a suitable measure of reliability. What do you conclude about the reliability of Brand A and Brand B?

1.28 Bear markets. Investors speak of a "bear market" when stock prices drop substantially. The following table gives data on all declines of at least 10% in the Standard & Poor's 500 stock index between 1940 and 2002. The data show how far the index fell from its peak and how long the decline in stock prices lasted.
(a) Make a stemplot of the percent declines in stock prices during these bear markets. Make a second stemplot, splitting the stems. Which graph do you prefer? Why?
(b) The shape of this distribution is irregular, but we could describe it as somewhat skewed. Is the distribution skewed to the right or to the left?
(c) Describe the center and spread of the data. What would you tell an investor about how far stocks fall in a bear market?

Year	Decline (percent)	Duration (months)	Year	Decline (percent)	Duration (months)
1940–1942	42	28	1968–1970	36	18
1946	27	5	1973–1974	48	21
1950	14	1	1981–1982	26	19
1953	15	8	1983–1984	14	10
1955	10	1	1987	34	3
1956–1957	22	15	1990	20	3
1959–1960	14	15	1998	19	3
1962	26	6	2000–2002	47	26
1966	22	8			

1.29 "The Fortune 500." Each year *Fortune* magazine lists the top 500 companies in the United States, ranked according to their total annual sales in dollars. Describe three other variables that could reasonably be used to measure the "size" of a company.

1.30 Salary distributions in a factory. A manufacturing company is reviewing the salaries of its full-time employees below the executive level at a large plant. The clerical staff is almost entirely female, while a majority of the production workers and technical staff is male. As a result, the distributions of salaries for male and female employees may be quite different. The table below gives the counts and percents of women and men in each salary class. Make histograms from these data, choosing the type that is most appropriate for comparing the two distributions. Then describe the overall shape of each salary distribution and the chief differences between them.

Salary distributions of female and male workers in a large factory

	Women		Men	
Salary ($1000)	Number	%	Number	%
10–15	89	11.8	26	1.1
15–20	192	25.4	221	9.0
20–25	236	31.2	677	27.9
25–30	111	14.7	823	33.6
30–35	86	11.4	365	14.9
35–40	25	3.3	182	7.4
40–45	11	1.5	91	3.7
45–50	3	0.4	33	1.4
50–55	2	0.3	19	0.8
55–60	0	0.0	11	0.4
60–65	0	0.0	0	0.0
65–70	1	0.1	3	0.1
Total	756	100.1	2451	100.0

1.31 The cost of Internet access. How much do users pay for Internet service? Here are the monthly fees (in dollars) paid by a random sample of 50 users of commercial Internet service providers in August 2000 (data from the August 2000

supplement to the Current Population Survey, from the Census Bureau Web site, `census.gov`):

20	40	22	22	21	21	20	10	20	20
20	13	18	50	20	18	15	8	22	25
22	10	20	22	22	21	15	23	30	12
9	20	40	22	29	19	15	20	20	20
20	15	19	21	14	22	21	35	20	22

Make a stemplot of these data. Briefly describe the pattern you see. About how much do you think America Online and its larger competitors were charging in August 2000? Which members of the sample may have been early adopters of fast access via cable modems or DSL lines?

1.32 Architects and engineers. The table for Exercise 1.8 gives the numbers of architects and engineers employed by Indianapolis architectural firms. A **back-to-back stemplot** helps us compare these two distributions. Write the stems as usual, but with a vertical line both to their left and to their right. On the right, put leaves for architects. On the left, put the leaves for engineers. Arrange the leaves on each stem in increasing order out from the stem. Now write a brief comparison of the distributions.

1.33 Watch those scales! The impression that a time plot gives depends on the scales you use on the two axes. If you stretch the vertical axis and compress the time axis, change appears to be more rapid. Compressing the vertical axis and stretching the time axis make change appear slower. Make two time plots of the first 12 years of data in Exercise 1.12, one that makes rates appear to increase very rapidly and one that suggests only a modest increase. The moral of this exercise is: pay close attention to the scales when you look at a time plot.

Section 1.2

1.34 Private consumption. The success of companies expanding to developing regions of the world depends in part on growing private consumption in those regions. Here are World Bank data on the growth of per capita private consumption (percent per year) for the period 1990 to 1997 in countries in Asia (outside Japan).

Country	Growth
Bangladesh	2.3
China	8.8
Hong Kong, China	3.9
India	4.1
Indonesia	6.4
Korea (South)	5.9
Malaysia	4.2
Pakistan	2.9
Philippines	1.3
Singapore	5.1
Thailand	5.6
Vietnam	6.2

(a) Make a stemplot of the data. Note the high outlier.
(b) Find the mean and median growth rates. How does the outlier explain the difference between your two results?
(c) Find the mean and median growth rates without the outlier. How does comparing your results in (b) and (c) illustrate the resistance of the median and the lack of resistance of the mean?

1.35 Private consumption. Refer to the previous exercise. The World Bank data also looked at 13 Eastern European countries. Here is the growth of per capita private consumption for each of these countries:

Country	Growth
Albania	6.0
Belarus	−5.2
Bulgaria	−1.0
Croatia	3.5
Czech Republic	3.2
Estonia	−1.7
Hungary	−1.5
Latvia	1.0
Poland	4.9
Romania	1.4
Russian Federation	7.0
Slovenia	3.7
Ukraine	−12.1

(a) Find the five-number summary for each group of countries (Asian and Eastern European).
(b) Make side-by-side boxplots to compare the growth of per capita private consumption for the two groups of countries. What do you conclude?

1.36 Supermarket shoppers. Here are the amounts spent (in dollars) by 50 consecutive shoppers at a supermarket, arranged in increasing order:

3.11	8.88	9.26	10.81	12.69	13.78	15.23	15.62	17.00	17.39
18.36	18.43	19.27	19.50	19.54	20.16	20.59	22.22	23.04	24.47
24.58	25.13	26.24	26.26	27.65	28.06	28.08	28.38	32.03	34.98
36.37	38.64	39.16	41.02	42.97	44.08	44.67	45.40	46.69	48.65
50.39	52.75	54.80	59.07	61.22	70.32	82.70	85.76	86.37	93.34

(a) Find the mean amount spent from the formula for the mean. Then enter the data into your calculator or software and use the mean function to obtain the mean. Verify that you get the same result.
(b) A stemplot suggests that the largest four values may be a cluster of outliers. Find the mean for the 46 observations that remain when you drop these outliers. How do the outliers change the mean?

1.37 Supermarket shoppers. Refer to the previous exercise where you found the mean amount spent by 50 consecutive shoppers at a supermarket. Now find the median of these amounts. Is the median smaller or larger than the mean? Explain why this is so.

1.38 The richest 1%. The distribution of individual incomes in the United States is strongly skewed to the right. In 1997, the mean and median incomes of the top 1% of Americans were $330,000 and $675,000. Which of these numbers is the mean and which is the median? Explain your reasoning.

1.39 Privately held restaurant companies. The Forbes 500 list of the largest privately held companies includes six restaurant-industry companies. Here they are, here with annual revenues in millions of dollars:

Company Name	Revenue ($millions)
Metromedia	1600
Domino's Pizza	1157
Buffets	937
Ilitch Ventures	800
AFC Enterprises	707
RTM Restaurant Group	700

(Only Domino's Pizza has a name familiar to consumers. You can check `www.forbes.com` to learn the brand names under which the other companies operate restaurants.) A graph of only 6 observations gives little information, so we proceed to compute the mean and standard deviation.
(a) Find the mean from its definition. That is, find the sum of the 6 observations and divide by 6.
(b) Find the standard deviation from its definition. That is, find the deviations of each observation from the mean, square the deviations, then obtain the variance and the standard deviation.
(c) Now enter the data into your calculator and use the mean and standard deviation buttons to obtain $\overline{x}$ and s. Do the results agree with your hand calculations?

1.40 Meaningful mean? A discussion of extreme weather says: "In most states, hurricanes occur infrequently. Yet, when a hurricane hits, the losses can be catastrophic. Average annual losses are not a meaningful measure of damage from rare

but potentially catastrophic events." (*Extreme Weather Sourcebook 2001,* at `sciencepolicy.colorado.edu/sourcebook`.) Why is this true?

1.41 Household assets. A report on the assets of American households says that the median net worth of households headed by someone younger than age 35 is $11,600. The mean net worth of these same young households is $90,700. (Douglas Fore, "Do we have a retirement crisis in America?" TIAA-CREF Institute *Research Dialogue* No. 77, 2003. The data are for the year 2001.) What explains the difference between these two measures of center?

1.42 Property damage by tornadoes. The table in Exercise 1.15 shows the average property damage caused by tornadoes over a 50-year period in each of the states and Puerto Rico. The distribution is strongly skewed to the right.
(a) Give the five-number summary. Explain why you can see from these five numbers that the distribution is right-skewed.
(b) A histogram or stemplot suggests that a few states are outliers. Show that there are *no* suspected outliers according to the $1.5 \times IQR$ rule. You see once again that a rule is not a substitute for plotting your data.
(c) Find the mean property damage. Explain why the mean and median differ so greatly for this distribution.

1.43 Carbon dioxide emissions. The table in Exercise 1.16 gives carbon dioxide (CO_2) emissions per person for countries with population at least 20 million. The distribution is strongly skewed to the right. The United States and several other countries appear to be high outliers.
(a) Give the five-number summary. Explain why this summary suggests that the distribution is right-skewed.
(b) Which countries are outliers according to the $1.5 \times IQR$ rule? Make a stemplot or histogram of the data. Do you agree with the rule's suggestions about which countries are and are not outliers?

1.44 Stock performance. How well have stocks done over the past generation? The Standard & Poor's 500 stock index describes the average performance of the stocks of 500 leading companies. Because each return is weighted by the total market value of each company's stock, the index emphasizes larger companies. Here are the real (that is, adjusted for the changing buying power of the dollar) returns on the S&P 500 for the years from 1971 to 2003:

Year	Return	Year	Return	Year	Return
1971	10.691	1982	17.033	1993	7.127
1972	15.070	1983	18.075	1994	−1.316
1973	−21.522	1984	2.253	1995	34.167
1974	−34.540	1985	26.896	1996	19.008
1975	28.353	1986	17.390	1997	31.138
1976	18.177	1987	0.783	1998	26.534
1977	−12.992	1988	11.677	1999	17.881
1978	−2.264	1989	25.821	2000	−12.082
1979	4.682	1990	−8.679	2001	−13.230
1980	17.797	1991	26.594	2002	−23.909
1981	−12.710	1992	4.584	2003	26.311

(a) Make a histogram of the real returns. Describe the shape of the distribution.
(b) Carefully make a stemplot of the real returns. Describe how negative values must be handled in order to maintain the overall process for constructing stemplots.

1.45 Stock performance. Continue with the S&P 500 data from the previous exercise.
(a) Find the mean and median. How does the shape of the distribution explain the relationship between the two measures of center?
(b) Returns on stocks vary a lot: they range from a loss of more than 34% to a gain of more than 34%. Are any of these years suspected outliers by the $1.5 \times IQR$ rule?

1.46 The Platinum Gasaver. National Fuelsaver Corporation manufactures the Platinum Gasaver, a device they claim "may increase gas mileage by 22%." In an advertisement published in the *Des Moines Register*, the gas mileages with and without the device were presented for 15 "identical" 5-liter vehicles. The percent changes in gas mileage for the vehicles were calculated and are presented here:

48.3 46.9 46.8 44.6 40.2 38.5 34.6 33.7
28.7 28.7 24.8 10.8 10.4 6.9 −12.4

The 12.4% *decrease* in gas mileage is an outlier in this data set.
(a) Find the mean $\overline{x}$ and the standard deviation s.
(b) Find $\overline{x}$ and s for the 14 observations that remain when you leave out the outlier. How does the outlier affect the values of $\overline{x}$ and s?
(c) What do you think the advertisement means when it calls these vehicles "identical"?

1.47 Education and income. Each March, the Bureau of Labor Statistics (BLS) records the incomes of all adults in a sample of 50,000 American households. We are interested in how income varies with the highest education level a person has reached. Computer software applied to the data from the March, 2000 survey gives the following results for people aged 25 or over:

Education	N	5%	25%	Median	75%	95%
High school diploma	31970	0	7800	17000	29600	56294
Some college, no degree	18797	0	8083	19600	34800	70026

Bachelor's degree	14705	500	17501	34150	55307	110086
Master's degree	4918	3300	27043	45069	68500	132560
Professional degree	1229	3000	33922	65850	118992	236967

It is common to make boxplots of large data sets using the 5% and 95% points in place of the minimum and maximum. The highest income among the 31,970 people with only a high school education, for example, is $425,510. It is more informative to see that 95% of this group earned less than $56,294. The 5% and 95% points contain between them the middle 90% of the observations.
(a) Use this output to make boxplots that compare the income distributions for the five education groups.
(b) Write a brief summary of the relationship between education and income. For example, do people who start college but don't get a degree do much better than people with only a high school education?

1.48 Education and income. Refer to the previous exercise. The output in the previous exercise shows that the data set contains information about 31,970 people with only a high school diploma and 1229 people with a professional degree. What are the positions of the median, the two quartiles, and the 5th and 95th percentiles in the ordered list of incomes for each of these groups?

1.49 Wealth of *Forbes* readers. The business magazine *Forbes* estimates that the "average" household wealth of its readers is either about $800,000 or about $2.2 million, depending on which "average" it reports. Which of these numbers is the mean wealth and which is the median wealth? Explain your answer.

1.50 A hot stock? It is usual in the study of investments to use the mean and standard deviation to summarize and compare investment returns. The following table gives the monthly returns on Philip Morris stock for the period from June 1990 to July 2001. (The return on an investment consists of the change in its price plus any cash payments made, given here as a percent of its price at the start of each month.)
(a) Make either a histogram or a stemplot of these data. How did you decide which graph to make?
(b) There are two clear outliers. What are the values of these observations? (The most extreme observation is explained by news of action against smoking, which depressed this tobacco company stock.) Describe the shape, center, and spread of the data after you omit the two outliers. (Both outliers are negative rates of return; however, there is one positive rate of return that is almost as separated from the neighboring rates of return in the positive direction as the two outliers on the negative side. For the sake of this exercise, we will not consider this highest positive rate of return to be an outlier.)
(c) Find the mean monthly return and the standard deviation of the returns (include the outliers). If you invested $100 in this stock at the beginning of a month and got the mean return, how much would you have at the end of the month?
(d) The distribution can be described as "symmetric and single-peaked, with two low outliers." If you invested $100 in this stock at the beginning of the worst month in the data (the more extreme outlier), how much would you have at the end of the

month? Find the mean and standard deviation again, this time leaving out the two low outliers. How much did these two observations affect the summary measures? Would leaving out these two observations substantially change the median? The quartiles? How do you know, without actual calculation? (Returns over longer periods of time, or returns on portfolios containing several investments, tend to follow a Normal distribution more closely than these monthly returns do. So use of the mean and standard deviation is better justified for such data.)

Monthly percent returns on Philip Morris stock from June 1990 to July 2001

3.0	−5.7	1.2	4.1	3.2	7.3	7.5	18.7	3.7	−1.8
2.4	−6.5	6.7	9.4	−2.0	−2.8	−3.4	19.2	−4.8	0.5
−0.6	2.8	−0.5	−4.5	8.7	2.7	4.1	−10.3	4.8	−2.3
−3.1	−10.2	−3.7	−26.6	7.2	−2.4	−2.8	3.4	−4.6	17.2
4.2	0.5	8.3	−7.1	−8.4	7.7	−9.6	6.0	6.8	10.9
1.6	0.2	−2.4	−2.4	3.9	1.7	9.0	3.6	7.6	3.2
−3.7	4.2	13.2	0.9	4.2	4.0	2.8	6.7	−10.4	2.7
10.3	5.7	0.6	−14.2	1.3	2.9	11.8	10.6	5.2	13.8
−14.7	3.5	11.7	1.5	2.0	−3.2	−3.9	−4.7	9.8	4.9
−8.3	4.8	−3.2	−10.9	0.7	6.4	11.3	−5.1	12.3	10.5
9.4	−3.6	−12.4	−16.5	−8.9	−0.4	10.0	5.4	−7.3	0.5
−7.4	−22.9	−0.5	−10.6	−9.2	−3.3	5.2	5.4	19.4	3.5
−4.9	17.8	0.7	24.4	4.3	16.6	0.0	9.5	−0.4	5.6
2.6	−2.7	−8.1	4.2						

1.51 Initial public offerings. During the stock market boom of the 1990s, initial public offerings (IPOs) of the stock of new companies often produced enormous gains for people who bought the stocks when they first became available. At least that's what legend says. A study of all 4567 companies that went public in the years 1990 to 2000 (excluding very small IPOs) found that on the average their stock prices had either *risen* 111% or *declined* 31% by the end of the year 2000. (Scott DeCarlo, Michael Schubach, and Vladimir Naumovski, "A decade of new issues," *Forbes*, March 5, 2001.) One of these numbers is the mean change in price and one is the median change. Which is which, and how can you tell?

1.52 Highly paid athletes. A news article reports that of the 411 players on National Basketball Association rosters in February 1998, only 139 "made more than the league average salary" of $2.36 million. Is $2.36 million the mean or median salary for NBA players? How do you know?

1.53 Mean or median? Which measure of center, the mean or the median, should you use in each of the following situations?
(a) Middletown is considering imposing an income tax on citizens. The city government wants to know the average income of citizens so that it can estimate the total tax base.
(b) In a study of the standard of living of typical families in Middletown, a sociologist estimates the average family income in that city.

Section 1.3

1.54 Use Table A to find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.
(a) $z < 2.85$
(b) $z > 2.85$
(c) $z > -1.66$
(d) $-1.66 < z < 2.85$

1.55 Gas mileage. The miles per gallon ratings for 2001 model vehicles vary according to an approximately Normal distribution with mean $\mu = 21.22$ miles per gallon and standard deviation $\sigma = 5.36$ miles per gallon.
(a) What percent of vehicles have miles per gallon ratings greater than 30?
(b) What percent of vehicles have miles per gallon ratings between 30 and 35?
(c) What percent of vehicles have miles per gallon ratings less than 12.45?

1.56 Use Table A to find the value z of a standard Normal variable that satisfies each of the following conditions. (Use the value of z from Table A that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of z marked on the axis.
(a) The point z with 25% of the observations falling below it.
(b) The point z with 40% of the observations falling above it.

1.57 GMAT scores. Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. (From *GMAT Examinee Score Interpretation Guide*, Graduate Management Admissions Council, 2000. Found online at `www.gmac.com`.) Total scores on the GMAT for the more than 500,000 people who took the exam between April 1997 and March 2000 are roughly Normally distributed with mean $\mu = 527$ and standard deviation $\sigma = 112$.
(a) What percent of test takers have scores above 500?
(b) What GMAT scores fall in the lowest 25% of the distribution?
(c) How high a GMAT score is needed to be in the highest 5%?

1.58 The Environmental Protection Agency requires that the exhaust of each model of motor vehicle be tested for the level of several pollutants. The level of oxides of nitrogen (NOX) in the exhaust of one light truck model was found to vary among individual trucks according to a Normal distribution with mean $\mu = 1.45$ grams per mile driven and standard deviation $\sigma = 0.40$ grams per mile. Sketch the density curve of this Normal distribution, with the scale of grams per mile marked on the horizontal axis. Also, give an interval that contains the middle 95% of NOX levels in the exhaust of trucks using this Normal model.

1.59 Stock performance. The 99.7 part of the 68–95–99.7 rule says that in practice Normal distributions are about six standard deviations wide. The table in Exercise 1.44 gives data on the real returns for the S&P 500 stock index over a

33-year period. The shape of the distribution is not close to Normal.
(a) Determine the mean and standard deviation of the real returns.
(b) What are the values three standard deviations above and below the mean, which would span the distribution if it were Normal?
(c) How do these values compare with the actual lowest and highest returns? (Remember that the 68–95–99.7 rule applies only to Normal distributions.)

1.60 Variability of pollutants. The Environmental Protection Agency requires that the exhaust of each model of motor vehicle be tested for the level of several pollutants. The level of oxides of nitrogen (NOX) in the exhaust of one light truck model was found to vary among individual trucks according to a Normal distribution with mean $\mu = 1.45$ grams per mile driven and standard deviation $\sigma = 0.40$ grams per mile. Sketch the density curve of this Normal distribution, with the scale of grams per mile marked on the horizontal axis. Also, using this Normal model, give an interval that contains the middle 95% of NOX levels in the exhaust of trucks.

1.61 NCAA rules for athletes. The National Collegiate Athletic Association (NCAA) requires Division I athletes to score at least 820 on the combined Mathematics and Verbal parts of the SAT to compete in their first college year. (Higher scores are required for students with poor high school grades.) In 2000, the scores of the 1,260,000 students taking the SATs were approximately Normal with mean 1019 and standard deviation 209. What percent of all students had scores less than 820?

1.62 More NCAA rules. The NCAA considers a student a "partial qualifier" eligible to practice and receive an athletic scholarship, but not to compete, if the combined SAT score is at least 720. Use the information in the previous exercise to find the percent of all SAT scores that are less than 720.

1.63 The stock market. The yearly rate of return on stock indexes (which combine many individual stocks) is approximately Normal. Between 1950 and 2000, U.S. common stocks had a mean yearly return of about 13%, with a standard deviation of about 17%. Take this Normal distribution to be the distribution of yearly returns over a long period.
(a) In what range do the middle 95% of all yearly returns lie?
(b) The market is down for the year if the return is less than zero. In what percent of years is the market down?
(c) In what percent of years does the index gain 25% or more?

Chapter 1 Review Exercises

1.64 Household and family income. In government data, a household contains all people who live together in a residence. A family consists of two or more people living together and related by blood, marriage, or adoption. In 1999, the mean and median of household incomes were \$40,816 and \$54,842. The mean and median of family incomes were \$48,950 and \$62,636.
(a) Which of each pair is the mean and which is the median? How do you know?
(b) Why are the mean and median incomes higher for families than for households?

1.65 Yankee salaries. Few companies release their employees' salaries. The baseball players' union, however, makes player salaries public. The following table contains the salaries of the New York Yankees as of opening day of the 2001 season. (New York Yankees salaries found online at `espn.go.com/mlb/`.) Describe the distribution of Yankee salaries, giving a graph and numerical measures to back up your description.

2001 salaries for the New York Yankees baseball team

Player	Salary	Player	Salary	Player	Salary
Derek Jeter	$12,600,000	Jorge Posada	$4,050,000	Shane Spencer	$320,000
Bernie Williams	$12,357,143	Mike Stanton	$2,450,000	Todd Williams	$320,000
Roger Clemens	$10,300,000	Orlando Hernandez	$2,050,000	Carlos Almanzar	$270,000
Mike Mussina	$10,000,000	Allen Watson	$1,700,000	Clay Bellinger	$230,000
Mariano Rivera	$ 9,150,000	Ramiro Mendoza	$1,600,000	Darrell Einertson	$206,000
David Justice	$ 7,000,000	Joe Oliver	$1,100,000	Randy Choate	$204,750
Andy Pettitte	$ 7,000,000	Henry Rodriguez	$ 850,000	Michael Coleman	$204,000
Paul O'Neill	$ 6,500,000	Alfonso Soriano	$ 630,000	D'Angelo Jimenez	$200,000
Chuck Knoblauch	$ 6,000,000	Luis Sojo	$ 500,000	Christian Parker	$200,000
Tino Martinez	$ 6,000,000	Brian Boehringer	$ 350,000	Scott Seabol	$200,000
Scott Brosius	$ 5,250,000				

1.66 Stock returns. The table for Exercise 1.50 gives the monthly percent returns on Philip Morris stock for the period from June 1990 to July 2001. The data appear in time order reading from left to right across each row in turn, beginning with the 3.0% return in June 1990. Make a time plot of the data. This was a period of increasing action against smoking, so we might expect a trend toward lower returns. But it was also a period in which stocks in general rose sharply, which would produce an increasing trend. What does your time plot show?

1.67 Do SUVs waste gas? The table in Exercise 1.7 gives the highway fuel consumption (in miles per gallon) for 32 midsize cars. Here are the highway mileages for 19 four-wheel-drive sport utility vehicles (data from the U.S. Department of Energy, *Model Year 2006 Fuel Economy Guide*, `epa.gov/OMSWWW/`):

Model	MPG	Model	MPG
Acura MDX	23	Jeep Wrangler	19
Chevrolet Blazer	22	Land Rover	15
Chevrolet Tahoe	18	Mazda Tribute	24
Dodge Durango	17	Mercedes-Benz ML320	21
Ford Expedition	18	Mitsubishi Montero	20
Ford Explorer	19	Nissan Pathfinder	18
Honda Passport	20	Suzuki Vitara	25
Infiniti QX4	19	Toyota RAV4	27
Isuzu Trooper	19	Toyota 4Runner	19
Jeep Grand Cherokee	19		

(a) Give a graphical and numerical description of highway fuel consumption for SUVs. What are the main features of the distribution?

(b) Make boxplots to compare the highway fuel consumption of midsize cars and SUVs. What are the most important differences between the two distributions?

1.68 California counties. You are planning a sample survey of households in California. You decide to select households separately within each county and to choose more households from the more populous counties. To aid in the planning, the following table gives the populations of California counties from the 2000 census. (From the Census Bureau Web site, `census.gov`.) Examine the distribution of county populations both graphically and numerically, using whatever tools are most suitable. Write a brief description of the main features of this distribution. Sample surveys often select households from all of the most populous counties but from only some of the less populous. How would you divide California counties into three groups according to population, with the intent of including all of the first group, half of the second, and a smaller fraction of the third in your survey?

Population of California counties, 2000

Alameda	1,443,741	Marin	247,289	San Mateo	707,161
Alpine	1,208	Mariposa	17,130	Santa Barbara	399,347
Amador	35,100	Mendocino	86,265	Santa Clara	1,682,585
Butte	203,171	Merced	210,554	Santa Cruz	255,602
Calaveras	40,554	Modoc	9,449	Shasta	163,256
Colusa	18,804	Mono	12,853	Sierra	3,555
Contra Costa	948,816	Monterey	401,762	Siskiyou	44,301
Del Norte	27,507	Napa	124,279	Solano	394,542
El Dorado	156,299	Nevada	92,033	Sonoma	458,614
Fresno	799,407	Orange	2,846,289	Stanislaus	446,997
Glenn	26,453	Placer	248,399	Sutter	78,930
Humboldt	126,518	Plumas	20,824	Tehama	56,039
Imperial	142,361	Riverside	1,545,387	Trinity	13,022
Inyo	17,945	Sacramento	1,223,499	Tulare	368,021
Kern	661,645	San Benito	53,234	Tuolumne	54,501
Kings	129,461	San Bernardino	1,709,434	Ventura	753,197
Lake	58,309	San Diego	2,813,833	Yolo	168,660
Lassen	33,828	San Francisco	776,733	Yuba	60,219
Los Angeles	9,519,338	San Joaquin	563,598		
Madera	123,109	San Luis Obispo	246,681		

CHAPTER 2

Section 2.1

2.1 Independent and dependent variables. In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?
(a) The amount of time a student spends studying for a statistics exam and the grade on the exam
(b) The weight and height of a person
(c) The amount of yearly rainfall and the yield of a crop
(d) An employee's salary and number of sick days used
(e) The economic class of a father and of a son

2.2 Stock prices. How well does a stock's market price at the beginning of the year predict its price at the end of the year? To find out, record the prices of a large group of stocks at the beginning of the year, wait until the end of the year, then record their prices again. What are the explanatory and response variables here? Are these variables categorical or quantitative?

2.3 Hand wipes. Antibacterial hand wipes can irritate the skin. A company wants to compare two different formulas for new wipes. Investigators choose two groups of adults at random. Each group uses one type of wipes. After several weeks, a doctor assesses whether or not each person's skin appears abnormally irritated. What are the explanatory and response variables? Are they categorical or quantitative variables?

2.4 Architectural firms. The table in Example 1.8 contains data describing firms engaged in commercial architecture in the Indianapolis area.
(a) We want to examine the relationship between number of full-time staff members employed and total billings. Which is the explanatory variable?
(b) Make a scatterplot of these data. (Be sure to label the axes with the variable names, not just x and y.) What does the scatterplot show about the relationship between these variables?

2.5 More on architectural firms. Refer to the previous exercise where you made a scatterplot of number of full-time staff members employed and total billings.
(a) Describe the direction of the relationship. Are the variables positively or negatively associated?
(b) Describe the form of the relationship. Is it linear?
(c) Describe the strength of the relationship. Can the total billings of a firm be predicted accurately by the number of full-time staff members? If a firm maintains a full-time staff of 75 members, approximately what will its total billings be from year to year?

2.6 Does fast driving waste fuel? How does the fuel consumption of a car change as its speed increases? Here are data for a British Ford Escort. Speed is measured in kilometers per hour, and fuel consumption is measured in liters of gasoline used per 100 kilometers traveled. (Based on T. N. Lam, "Estimating fuel consumption from engine size," *Journal of Transportation Engineering*, 111 (1985), pp. 339–357.) The data for 10 to 50 km/hr are measured; those for 60 and higher are calculated from a model given in the paper and are therefore smoothed.

Speed (km/h)	Fuel used (liters/100 km)	Speed (km/h)	Fuel used (liters/100 km)
10	21.00	?90	?7.57
20	13.00	100	?8.27
30	10.00	110	?9.03
40	?8.00	120	?9.87
50	?7.00	130	10.79
60	?5.90	140	11.77
70	?6.30	150	12.83
80	?6.95		

2.7 Do heavier people burn more energy? In judging the effectiveness of popular diet and exercise programs, researchers wish to take into account metabolic rate, the rate at which the body consumes energy. The following table gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.

(a) Make a scatterplot of the data for the female subjects. Which is the explanatory variable?

(b) Is the association between these variables positive or negative? What is the form of the relationship? How strong is the relationship?

(c) Now add the data for the male subjects to your graph, using a different color or a different plotting symbol. Does the pattern of the relationship that you observed in (b) hold for men also? How do the male subjects as a group differ from the female subjects as a group?

Lean body mass and metabolic rate

Subject	Sex	Mass (kg)	Rate (cal)	Subject	Sex	Mass (kg)	Rate (cal)
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

2.8 Sector fund returns. Fidelity Investments, like other large mutual funds companies, offers many "sector funds," which concentrate their investments in narrow segments of the stock market. These funds often rise or fall by much more than the market as a whole. We can group them by broader market sector to compare returns. Here are percent total returns for 23 Fidelity "Select Portfolios" funds for the year 2003, a year in which stocks rose sharply:

Market sector	Fund returns (percent)						
Consumer	23.9	14.1	41.8	43.9	31.1		
Financial services	32.3	36.5	30.6	36.9	27.5		
Natural resources	22.9	7.6	32.1	28.7	29.5	19.1	
Technology	26.1	62.7	68.1	71.9	57.0	35.0	59.4

(a) Make a plot of total return against market sector (space the four market sectors equally on the horizontal axis). Compute the mean return for each sector, add the means to your plot, and connect the means with line segments.
(b) Based on the data, which of these market sectors were the best places in which to invest in 2003? Hindsight is wonderful.
(c) Does it make sense to speak of a positive or negative association between market sector and total return?

2.9 Sector fund returns. The data for 2003 in the previous exercise make sector funds look attractive. Stocks rose sharply in 2003, after falling sharply in 2002 (and also in 2001 and 2000). Let's look at the percent returns for 2003 and 2002 for these same 23 funds.

2002 returns	2003 returns	2002 returns	2003 returns	2002 returns	2003 returns
−17.1	23.9	−0.7	36.9	−37.8	59.4
−6.7	14.1	−5.6	27.5	−11.5	22.9
−21.1	41.8	−26.9	26.1	−0.7	36.9
−12.8	43.9	−42.0	62.7	64.3	32.1
−18.9	31.1	−47.8	68.1	−9.6	28.7
−7.7	32.3	−50.5	71.9	−11.7	29.5
−17.2	36.5	−49.5	57.0	−2.3	19.1
−11.4	30.6	−23.4	35.0		

Do a careful graphical analysis of these data: side-by-side comparison of the distributions of returns in 2002 and 2003 and also a description of the relationship between the returns of the same funds in these two years. What are your most important findings? (The outlier is Fidelity Gold Fund.)

2.10 Online stock brokerages. How closely does the number of trading accounts relate to the total assets of the 10 largest online stock brokerages? The following table presents data on these brokerages. The number of accounts is in thousands, and assets are in billions of dollars.

Brokerage	Accounts (1000)	Assets ($ billions)
Ameritrade	428	19.5
Charles Schwab	2500	219.0
Datek	205	5.5
Discover	134	5.9
DLJ Direct	590	11.2
E* Trade	909	21.1
Fidelity	2300	160.0
National Discount Brokers	125	6.8
Suretrade	130	1.3
TD Waterhouse	615	38.8

(a) Does the wording of the initial question in this exercise indicate which variable should be treated as the response?
(b) Construct a scatterplot with Assets on the vertical axis and Accounts on the horizontal axis.
(c) Comment on the form, direction, and strength of the relationship between Assets and Accounts.
(d) Two brokerages stand out from the rest of the data. Identify these two brokerages by putting their names near their respective data points in your scatterplot.

2.11 Employment numbers. Where do most Americans work? Here are the numbers of paid employees in 1997 and in 2002 in the 10 kinds of business that employ the most people, according to a government census of business (ddata from the March 2002 annual demographic supplement to the Current Population Survey, downloaded from `bls.census.gov/cps`):

Business type	2002 paid employees	1997 paid employees
Accommodation and food services	10,836,365	9,451,226
Administrative and support	8,002,725	7,066,658
Construction	6,943,582	5,664,853
Finance and insurance	6,663,714	5,835,214
Health care and social assistance	15,346,996	13,561,579
Information	3,845,583	3,066,167
Manufacturing	14,543,338	16,805,127
Professional, scientific, technical	7,508,866	5,361,210
Retail trade	15,029,339	13,991,103
Wholesale trade	6,034,579	5,796,557

(a) Construct a scatterplot with the number of paid employees in 2002 on the vertical axis and the number of paid employees in 1997 on the horizontal axis. Comment on the form, direction, and strength of the relationship between the two sets of values in the table. Also, identify any potential outliers in the data.
(b) Construct a scatterplot with the number of paid employees in 1997 on the vertical axis and the number of paid employees in 2002 on the horizontal axis. Comment on the form, direction, and strength of the relationship between the two sets of values in the table. Also, identify any potential outliers in the data.
(c) Are your comments on form, direction, and strength different in parts (a) and

(b)? Clearly and carefully describe at least one scenario in which it would make sense to use a 1997 value to predict an unknown 2002 value in this context. Then, describe at least one scenario in which it would make sense to use a 2002 value to predict an unknown 1997 value.

2.12 Stocks and bonds. How is the flow of investors' money into stock mutual funds related to the flow of money into bond mutual funds? Here are data on the net new money flowing into stock and bond mutual funds in the years 1985 to 2000, in billions of dollars. (Net cash flow data from Sean Collins, *Mutual Fund Assets and Flows in 2000*, Investment Company Institute, 2001. Found online at `ici.org`. The raw data were converted to real dollars using annual average values of the Consumer Price Index.) "Net" means that funds flowing out are subtracted from those flowing in. If more money leaves than arrives, the net flow will be negative. To eliminate the effect of inflation, all dollar amounts are in "real dollars" with constant buying power equal to that of a dollar in the year 2000. Make a plot (with stocks on the horizontal axis) and describe the pattern it reveals.

Year	1985	1986	1987	1988	1989	1990	1991	1992
Stocks	12.8	34.6	28.8	−23.3	8.3	17.1	50.6	97.0
Bonds	100.8	161.8	10.6	−5.8	−1.4	9.2	74.6	87.1

Year	1993	1994	1995	1996	1997	1998	1999	2000
Stocks	151.3	133.6	140.1	238.2	243.5	165.9	194.3	309.0
Bonds	84.6	−72.0	−6.8	3.3	30.0	79.2	−6.2	−48.0

2.13 Calories and salt in hot dogs. Are hot dogs that are high in calories also high in salt? Here are data for calories and salt content (milligrams of sodium) in 17 brands of meat hot dogs (data from *Consumer Reports*, June 1986, pp. 366–367):

Brand	Calories	Sodium (mg)	Brand	Calories	Sodium (mg)
1	173	458	10	136	393
2	191	506	11	179	405
3	182	473	12	153	372
4	190	545	13	107	144
5	172	496	14	195	511
6	147	360	15	135	405
7	146	387	16	140	428
8	139	386	17	138	339
9	175	507			

(a) Make a scatterplot of these data, with calories on the horizontal axis. Describe the overall form, direction, and strength of the relationship. Are hot dogs that are high in calories generally also high in salt?
(b) One brand, "Eat Slim Veal Hot Dogs," is made of veal rather than beef and pork and positions itself as a diet brand. Which brand in the table do you think this is?

2.14 How many corn plants are too many? Midwestern farmers make many business decisions before planting. Data can help. For example, how much corn per acre should a farmer plant to obtain the highest yield? Too few plants will give a low yield. On the other hand, if there are too many plants, they will compete with each other for moisture and nutrients, and yields will fall. To find the best planting rate, plant at different rates on several plots of ground and measure the harvest. (Be sure to treat all the plots the same except for the planting rate.) Here are data from such an experiment (W. L. Colville and D. P. McGill, "Effect of rate and method of planting on several plant characters and yield of irrigated corn," *Agronomy Journal*, 54 (1962), pp. 235–238):

Plants per acre	Yield (bushels per acre)			
12,000	150.1	113.0	118.4	142.6
16,000	166.9	120.7	135.2	149.8
20,000	165.3	130.1	139.6	149.9
24,000	134.7	138.4	156.1	
28,000	119.0	150.5		

(a) Is yield or planting rate the explanatory variable?
(b) Make a scatterplot of yield and planting rate.
(c) Describe the overall pattern of the relationship. Is it linear? Is there a positive or negative association, or neither?
(d) Find the mean yield for each of the five planting rates. Plot each mean yield against its planting rate on your scatterplot and connect these five points with lines. This combination of numerical description and graphing makes the relationship clearer. What planting rate would you recommend to a farmer whose conditions were similar to those in the experiment?

2.15 Business starts and failures. The following table lists the number of businesses started and the number of businesses that failed by state for one year. We might expect an association to exist between these economic measures. (a) Make a scatterplot of business failures against business starts. Take business starts as the explanatory variable.
(b) The plot shows a positive association between the two variables. Why do we say that the association is positive?
(c) Find the point for Florida in the scatterplot and circle it.
(d) There is an outlier at the upper right of the plot. Which state is this?
(e) We wonder about clusters and gaps in the data display. There is a relatively clear cluster of states at the lower left of the plot. Four states are outside this cluster. Which states are these? Are they mainly from one part of the country?

Business starts and failures

State	Starts	Failures	State	Starts	Failures
AL	2,645	546	MT	397	201
AK	271	177	NE	565	383
AZ	2,868	1,225	NV	1,465	677
AR	1,091	748	NH	708	322
CA	21,582	17,679	NJ	6,412	2,024
CO	3,041	2,483	NM	887	585
CT	2,069	580	NY	13,403	4,233
DE	508	28	NC	4,371	846
DC	537	75	ND	229	144
FL	13,029	2,047	OH	4,829	2,524
GA	5,471	800	OK	1,367	990
HI	593	781	OR	1,823	1,109
ID	639	441	PA	5,525	2,641
IL	5,542	3,291	RI	544	150
IN	2,611	473	SC	2,023	410
IA	1,020	244	SD	281	275
KS	967	1,140	TN	2,835	1,369
KY	1,824	270	TX	10,936	6,785
LA	1,849	377	UT	1,417	388
ME	577	259	VT	261	80
MD	3,139	1,283	VA	3,502	860
MA	3,425	1,200	WA	2,956	2,528
MI	4,293	1,551	WV	623	305
MN	2,111	1,711	WI	2,357	1,005
MS	1,347	177	WY	213	166
MO	2,163	1,321			

2.16 Transforming data. Data analysts often look for a **transformation** of data that simplifies the overall pattern. Here is an example of how transforming the response variable can simplify the pattern of a scatterplot. The data show the growth of world crude-oil production between 1880 and 1990. (Data from the Energy Information Administration, recorded in Robert H. Romer, *Energy: An Introduction to Physics*, W. H. Freeman, 1976, for 1880 to 1970, and in the *Statistical Abstract of the United States* for more recent years.)

Year	1880	1890	1900	1910	1920	1930	1940
Millions of barrels	30	77	149	328	689	1412	2150

Year	1950	1960	1970	1980	1990
Millions of barrels	3803	7674	16,690	21,722	22,100

(a) Make a scatterplot of production (millions of barrels) against year. Briefly describe the pattern of world crude-oil production growth.
(b) Now take the logarithm of the production in each year (use the `log` button on your calculator). Plot the logarithms against year. What is the overall pattern on this plot?

2.17 Categorical explanatory variable. A scatterplot shows the relationship between two quantitative variables. Here is a similar plot to study the relationship between a categorical explanatory variable and a quantitative response variable.

Fidelity Investments, like other large mutual funds companies, offers many "sector funds," which concentrate their investments in narrow segments of the stock market. These funds often rise or fall by much more than the market as a whole. We can group them by broader market sector to compare returns. Here are percent total returns for 23 Fidelity "Select Portfolios" funds for the year 2000, a year in which stocks declined and technology shares were especially hard hit (compiled from Fidelity data by the *Fidelity Insight* newsletter, found online at `fidelity.kobren.com`):

Market sector	Fund returns (percent)						
Consumer	−9.3	29.8	−24.4	−23.1	−11.3		
Financial services	18.3	28.1	28.5	50.2	53.3		
Technology	−1.6	−30.4	−28.8	−17.5	−20.2	−32.3	−37.4
Utilities and natural resources	31.8	50.3	−18.1	71.3	30.4	−13.5	

(a) Make a plot of total return against market sector (space the four market sectors equally on the horizontal axis). Compute the mean return for each sector, add the means to your plot, and connect the means with line segments.
(b) Based on the data, which market sectors were good places to invest in 2000? Hindsight is wonderful.
(c) Does it make sense to speak of a positive or negative association between market sector and total return?

Section 2.2

2.18 Four-wheel-drive minivans. The city and highway miles per gallon for all three 2001 model four-wheel-drive minivans are

City Mpg	18.6	19.2	18.8
Highway Mpg	28.7	29.3	29.2

(a) Make a scatterplot with appropriately labeled axes.
(b) Find the correlation r step-by-step. First, find the mean and standard deviation of the city mileages and of the highway mileages. (Use a calculator.) Then find the three standardized values for each variable and use the formula for r.
(c) Comment on the direction and strength of the relationship.

2.19 The price of a gigabyte. Here are data on the size in gigabytes (GB) and the wholesale price in dollars of several internal hard drive models from one manufacturer:

Size (GB)	80	60	180	300	200	100	30
Price ($)	31	29	50	80	47	33	15

(a) Make a scatterplot. Be sure to label the axes appropriately. Comment on the form, direction, and strength of the relationship.
(b) Find the correlation r step-by-step. First, find the mean and standard deviation

of the sizes and of the prices (use your calculator). Then find the seven standardized values for these variables and use the formula for r.
(c) Enter these data into your calculator and use the calculator's correlation function to find r. Check that you get the same result as in (b).

2.20 Brand names and generic products. (a) If a store always prices its generic "store brand" products at 90% of the brand name products' prices, what would be the correlation between the prices of the brand name products and the store brand products? (*Hint:* Draw a scatterplot for several prices.)
(b) If the store always prices its generic products \$1 less than the corresponding brand name products, then what would be the correlation between the prices of the brand name products and the store brand products?

2.21 Strong association but no correlation. The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

Speed	20	30	40	50	60
Mpg	24	28	30	28	24

Make a scatterplot of mileage versus speed. Show that the correlation between speed and mileage is $r = 0$. Explain why the correlation is 0 even though there is a strong relationship between speed and mileage.

2.22 Driving speed and fuel consumption. The data in the previous exercise were made up to create an example of a strong curved relationship for which, nonetheless, $r = 0$. Exercise 2.6 gives actual data on gas used versus speed for a small car. Make a scatterplot if you did not do so in that exercise. Calculate the correlation, and explain why r is close to 0 despite a strong relationship between speed and gas used.

2.23 Sector fund returns. Refer to the data in Exercise 2.9 on the returns from 23 Fidelity "sector funds" in 2002 (a down-year for stocks) and 2003 (an up-year).
(a) Make a scatterplot. Fidelity Gold Fund, the only fund with a positive return in both years, is an extreme outlier.
(b) To demonstrate that correlation is not resistant, find r for all 23 funds and then find r for the 22 funds other than Gold. Explain from Gold's position in your plot why omitting this point makes r more negative.

2.24 How many calories? A food industry group asked 3368 people to guess the number of calories in each of several common foods. The following table displays the averages of their guesses and the correct number of calories. (Data from a survey by the Wheat Industry Council reported in *USA Today*, October 20, 1983.) (a) We think that how many calories a food actually has helps explain people's guesses of how many calories it has. With this in mind, make a scatterplot of these data.
(b) Find the correlation r. Explain why your r is reasonable based on the scatterplot.
(c) The guesses are all higher than the true calorie counts. Does this fact influence the correlation in any way? How would r change if every guess were 100 calories

higher?
(d) The guesses are much too high for spaghetti and the snack cake. Circle these points on your scatterplot. Calculate r for the other eight foods, leaving out these two points. Explain why r changed in the direction that it did.

Guessed and true calories in 10 foods

Food	Guessed calories	Correct calories	Food	Guessed calories	Correct calories
8 oz whole milk	196	159	2-oz candy bar	364	260
5 oz spaghetti with tomato sauce	394	163	Saltine cracker	74	12
5 oz macaroni with cheese	350	269	Medium-size apple	107	80
One slice wheat bread	117	61	Medium-size potato	160	88
One slice white bread	136	76	Cream-filled snack cake	419	160

2.25 Investment diversification. A mutual funds company's newsletter says, "A well-diversified portfolio includes assets with low correlations." The newsletter includes a table of correlations between the returns on various classes of investments. For example, the correlation between municipal bonds and large-cap stocks is 0.50, and the correlation between municipal bonds and small-cap stocks is 0.21. (*T. Rowe Price Report*, Winter 1997, p. 4.) (a) Rachel invests heavily in municipal bonds. She wants to diversify by adding an investment whose returns do not closely follow the returns on her bonds. Should she choose large-cap stocks or small-cap stocks for this purpose? Explain your answer.
(b) If Rachel wants an investment that tends to increase when the return on her bonds drops, what kind of correlation should she look for?

2.26 Sloppy writing about correlation. Each of the following statements contains a blunder. Explain in each case what is wrong.
(a) "There is a high correlation between the gender of American workers and their income."
(b) "We found a high correlation ($r = 1.09$) between students' ratings of faculty teaching and ratings made by other faculty members."
(c) "The correlation between planting rate and yield of corn was found to be $r = 0.23$ bushel."

Section 2.3

2.27 Architectural firms. The table in Example 1.8 contains data describing firms engaged in commercial architecture in the Indianapolis area. The regression line for predicting total billings from number of full-time staff members employed is

$$\text{billings} = 0.2821 + (0.0917 \times \text{employed})$$

(a) Make a scatterplot and draw this regression line on the plot. Using the regression equation, predict the total billings of an architectural firm in Indianapolis with 111 full-time staff members.
(b) Compare the observed total billings for the firm with 111 full-time staff members

with your prediction from part (a) by calculating the prediction error. How accurate was your prediction?

2.28 Review of straight lines. A company manufactures batteries for cell phones. The overhead expenses of keeping the factory operational for a month—even if no batteries are made—total $500,000. Batteries are manufactured in lots (1000 batteries per lot) costing $10,000 to make. In this scenario, $500,000 is the *fixed* cost associated with producing cell phone batteries, and $10,000 is the *marginal* (or *variable*) cost of producing each lot of batteries. The total monthly cost y of producing x lots of cell phone batteries is given by the equation

$$y = 500000 + 10000x$$

(a) Draw a graph of this equation. (Choose two values of x, such as 0 and 10. Compute the corresponding values of y from the equation. Plot these two points on graph paper and draw the straight line joining them.)
(b) What will it cost to produce 20 lots of batteries (20,000 batteries)?
(c) If each lot cost $20,000 instead of $10,000 to produce, what is the equation that describes total monthly cost for x lots produced?

2.29 Review of straight lines. A local consumer electronics store sells exactly 4 DVD players of a particular model each week. The store expects no more shipments of this particular model, and they have 96 such units in their current inventory.
(a) Give an equation for the number of DVD players of this particular model in inventory after x weeks. What is the slope of this line?
(b) Draw a graph of this line between now (Week 0) and Week 10.
(c) Would you be willing to use this line to predict the inventory after 25 weeks? Do the prediction and think about the reasonableness of the result.

2.30 Review of straight lines. A cellular telephone company offers two plans. Plan A charges $20 a month for up to 75 minutes of airtime and $0.45 per minute above 75 minutes. Plan B charges $30 a month for up to 250 minutes and $0.40 per minute above 250 minutes.
(a) Draw a graph of the Plan A charge against minutes used from 0 to 250 minutes.
(b) How many minutes a month must the user talk in order for Plan B to be less expensive than Plan A?

2.31 Sector fund returns. Exercise 2.9 gives the returns of 23 Fidelity "sector funds" for the years 2002 and 2003. These mutual funds invest in narrow segments of the stock market. They often rise faster than the overall market in up-years such as 2003 and fall faster than the market in down-years such as 2002. A scatterplot shows that Fidelity Gold Fund—the only fund that went up in 2002—is an outlier. In Exercise 2.23, you showed that this outlier has a strong effect on the correlation. The least-squares line, like the correlation, is not resistant.
(a) Find the equations of two least-squares lines for predicting 2003 return from 2002 return, one for all 23 funds and one omitting Fidelity Gold Fund. Make a scatterplot with both lines drawn on it. The two lines are very different.
(b) Starting with the least-squares idea, explain why adding Fidelity Gold Fund to the other 22 funds moves the line in the direction that your graph shows.

2.32 Stock index returns. How well have stocks done over the past generation? The Standard & Poor's 500 stock index describes the average performance of the stocks of 500 leading companies. Because the average is weighted by the total market value of each company's stock, the index emphasizes larger companies. Here are the real (that is, adjusted for the changing buying power of the dollar) returns on the S&P 500 for the years from 1971 to 2003:

Year	Return	Year	Return	Year	Return
1971	10.691	1982	17.033	1993	7.127
1972	15.070	1983	18.075	1994	−1.316
1973	−21.522	1984	2.253	1995	34.167
1974	−34.540	1985	26.896	1996	19.008
1975	28.353	1986	17.390	1997	31.138
1976	18.177	1987	0.783	1998	26.534
1977	−12.992	1988	11.677	1999	17.881
1978	−2.264	1989	25.821	2000	−12.082
1979	4.682	1990	−8.679	2001	−13.230
1980	17.797	1991	26.594	2002	−23.909
1981	−12.710	1992	4.584	2003	26.311

(a) Plot real returns against time. To bring out the pattern in these highly variable returns, use a vertical scale from −60% to 60%. Describe this pattern.
(b) Because of the curved pattern and the large amount of scatter about the pattern, a straight line doesn't describe the data well. Regress real return on year, and add the regression line to your plot. What is the value of r^2, and what does this value tell you about the usefulness of the regression line?

2.33 Florida condo sales. Florida reappraises real estate every year, and the county appraiser's Web site lists the current "fair market value" of each piece of property. Property usually sells for somewhat more than the appraised market value. Here are the appraised market values and actual selling prices of condominium units sold in a beachfront building over a 19-month period (in thousands of dollars):

Selling price	Appraised value	Month	Selling price	Appraised value	Month
850	758	0	790	605.9	13
900	812.7	1	700	483.8	14
625	504	2	715	585.8	14
1075	956.7	2	825	707.6	14
890	747.9	8	675	493.9	17
810	717.7	8	1050	802.6	17
650	576.6	9	1325	1031.8	18
845	648.3	12	845	586.7	19

(a) Make a scatterplot. It appears that appraised market value can be used to predict selling price.
(b) Find the least-squares line for predicting selling price from appraised value. Add this line to your scatterplot. Another unit in this building has appraised value $802,600. What do you predict that it will sell for?

2.34 Florida condo sales. Prices for beachfront property were rising rapidly during the period for which the previous exercise gives data. Because property is reassessed just once a year, selling prices might pull away from appraised values over time. The data in the previous exercise are in order by date of the sale, and we have given the number of months from the start of the data period. The residuals from the regression of selling price on appraised value are (rounded):

−70.60	−77.85	−29.76	−53.56	−20.03	−68.42	−80.75	39.21
28.59	66.38	−25.38	−42.85	30.81	82.72	117.83	103.68

(a) Plot the residuals against the explanatory variable (appraised value). To make the pattern clearer, use vertical limits −200 to 200. Explain why the pattern you see implies that a straight line is a good summary of these data.
(b) Next, plot the residuals against month. Explain why the pattern fits the fact that selling prices were rising rapidly.

2.35 Florida condo sales. Consider the Florida condo sales data that you analyzed in the previous two exercises.
(a) Make a scatterplot with selling prices on the vertical axis and appraised values on the horizontal axis if you have not already done so. Does the plot suggest a linear relationship?
(b) Report the correlation for these data. From the correlation, calculate r^2 and interpret its value in the context of these data.
(c) Find the least-squares line for predicting selling price from appraised value. Add this line to your scatterplot.
(d) If another unit in this building is appraised at $688,750, what do you predict that it will sell for?
(e) Calculate the average of the selling prices and compare this with your answer in part (d). What do you now know about the value $688,750?

2.36 Florida condo sales. Consider the Florida condo sales data that you analyzed in the previous two exercises.
(a) Use statistical software to calculate the residuals from the least-squares regression line for predicting selling prices from appraised values.
(b) Construct a residual plot. Does the residual plot indicate that the line fits the data well? Explain your response.

2.37 Online stock brokerages. The table in Exercise 2.10 presents data on the number of accounts and the total assets of 10 online stock brokerages. Our goal is to predict total assets based on number of accounts. Part of this exercise requires that you calculate the slope and intercept of the least-squares regression line "by hand" using several summary statistics. Least-squares calculations like these are best done by statistical software. However, the connections among means, standard deviations, correlation, slope, and intercept are often lost when the computer *always* does the calculations for us.
(a) Use the following summary statistics and the formulas of this section to calculate the least-squares slope and intercept for these data. Be sure to show at least one intermediate step in your calculations.

$$\begin{array}{rclrcl} \bar{x} & = & 793.6 & \bar{y} & = & 48.9 \\ s_x & = & 886.3337 & s_y & = & 76.1639 \end{array}$$
$$r = 0.9683$$

(b) Demonstrate that the least-squares line you calculated in part (a) passes through the point $(\bar{x}, \bar{y})$.
(c) Calculate and interpret the value of r^2 in the context of these data.

2.38 Online stock brokerages. Continuing with the brokerage data in the previous exercise, use statistical software to complete this exercise.
(a) Report the least-squares regression line from your statistical software and compare the estimated intercept and slope to the estimates from the previous exercise.
(b) Report the 10 residuals from the model fitted in part (a) and provide a residual plot of these values.
(c) Does the residual plot display a systematic pattern of any kind? If so, describe the pattern in a sentence or two. If not, then what does the lack of systematic pattern indicate?
(d) Identify the two unusual points on the far right of the residual plot by brokerage name. Are these points mostly unusual because of their residual value or their x-value (number of accounts)?

2.39 Online stock brokerages. Refer to the data and model in the previous exercise. Show at least one intermediate step in all calculations.
(a) Use the equation of the least-squares regression line to calculate a prediction of total assets for DLJ Direct.
(b) Is the prediction for DLJ Direct an overestimate or an underestimate of the actual total assets of this brokerage? How far off is the prediction from the actual total assets of DLJ Direct?

2.40 Online stock brokerages. Refer to the data and model in the previous two exercises. Use statistical software to answer these questions.
(a) Provide a scatterplot of total assets (vertical axis) versus number of accounts (horizontal axis).
(b) Fit the least-squares regression line to these data, and plot the fitted line on the scatterplot from part (a).
(c) Repeat part (b) after excluding the data for Charles Schwab and Fidelity. How do the two regression lines compare?
(d) Calculate the absolute change and the percent change in both the intercept and the slope when Charles Schwab and Fidelity were excluded from the analysis.
(e) Do these two points appear to be influential observations? Explain your response based on your work in this exercise.

2.41 Employment numbers. Refer to the data on number of paid employees presented in Exercise 2.11.
(a) Use statistical software to calculate the least-squares regression line for predicting the number of paid employees in 2002 from the number of paid employees in 1997. Report the estimated intercept, the estimated slope, and the value of r^2.
(b) Now report the number of paid employees with the units "millions of employees" rather than just a count of employees (for example, 10,836,365 employees is

10.836365 million employees). Using these values, repeat part (a).
(c) Are all three values (intercept, slope, r^2) affected by the change in units from "employees" to "millions of employees"? Compare and comment on the differences in the three values in parts (a) and (b).

2.42 Employment numbers. Refer to the data on number of paid employees presented in Exercise 2.11.
(a) Use statistical software to calculate the residuals from the least-squares regression line for predicting the number of paid employees in 2002 from the number of paid employees in 1997.
(b) Construct a residual plot and be sure to draw a horizontal line at zero on the residual plot.
(c) If the least-squares regression line captures the overall pattern in the data, then the residual plot should have no systematic pattern. Using this criterion, do you conclude that the regression line captures the overall pattern in the data? Explain your response.

2.43 Moving in step? One reason to invest abroad is that markets in different countries don't move in step. When American stocks go down, foreign stocks may go up. So an investor who holds both bears less risk. That's the theory. Now we read, "The correlation between changes in American and European share prices has risen from 0.4 in the mid-1990s to 0.8 in 2000." ("Dancing in step," *Economist*, March 22, 2001.) Explain to an investor who knows no statistics why this fact reduces the protection provided by buying European stocks.

2.44 Interpreting correlation. Refer to the previous exercise. The same article that claims that the correlation between changes in stock prices in Europe and the United States was 0.8 in 2000 goes on to say, "Crudely, that means that movements on Wall Street can explain 80% of price movements in Europe." Is this true? What is the correct percent explained if $r = 0.8$?

2.45 Investing at home and overseas. Investors ask about the relationship between returns on investments in the United States and on investments overseas. The following table gives the total returns on U.S. and overseas common stocks over a 30-year period. (The total return is change in price plus any dividends paid, converted into U.S. dollars. Both returns are averages over many individual stocks.) (The U.S. returns are for the Standard & Poor's 500 stock index. The overseas returns are for the Morgan Stanley Europe, Australasia, Far East (EAFE) index.)
(a) Make a scatterplot suitable for predicting overseas returns from U.S. returns.
(b) Find the correlation and r^2. Describe the relationship between U.S. and overseas returns in words, using r and r^2 to make your description more precise.
(c) Find the least-squares regression line of overseas returns on U.S. returns. Draw the line on the scatterplot. Are you confident that predictions using the regression line will be quite accurate? Why?
(d) Circle the point that has the largest residual (either positive or negative). What year is this? Are there any points that seem likely to be very influential?

Annual total return on overseas and U.S. stocks

Year	Overseas % return	U.S. % return	Year	Overseas % return	U.S. % return
1971	29.6	14.6	1986	69.4	18.6
1972	36.3	18.9	1987	24.6	5.1
1973	−14.9	−14.8	1988	28.5	16.8
1974	−23.2	−26.4	1989	10.6	31.5
1975	35.4	37.2	1990	−23.0	−3.1
1976	2.5	23.6	1991	12.8	30.4
1977	18.1	−7.4	1992	−12.1	7.6
1978	32.6	6.4	1993	32.9	10.1
1979	4.8	18.2	1994	6.2	1.3
1980	22.6	32.3	1995	11.2	37.6
1981	−2.3	−5.0	1996	6.4	23.0
1982	−1.9	21.5	1997	2.1	33.4
1983	23.7	22.4	1998	20.3	28.6
1984	7.4	6.1	1999	27.2	21.0
1985	56.2	31.6	2000	−14.0	−9.1

2.46 Investing at home and overseas. Refer to the previous exercise where you examined the relationship between returns on U.S. and overseas stocks. Investors also want to know what typical returns are and how much year-to-year variability (called *volatility* in finance) there is. Regression and correlation do not answer these questions.
(a) Find the five-number summaries for both U.S. and overseas returns, and make side-by-side boxplots to compare the two distributions.
(b) Were returns generally higher in the United States or overseas during this period? Explain your answer.
(c) Were returns more volatile (more variable) in the United States or overseas during this period? Explain your answer.

2.47 How many calories? The table in Exercise 2.24 gives data on the true calories in 10 foods and the average guesses made by a large group of people. In that exercise, we explored the influence of two outlying observations on the correlation.
(a) Make a scatterplot suitable for predicting guessed calories from true calories. Circle the points for spaghetti and snack cake on your plot. These points lie outside the linear pattern of the other 8 points.
(b) Find the least-squares regression line of guessed calories on true calories. Do this twice, first for all 10 data points and then leaving out spaghetti and snack cake.
(c) Plot both lines on your graph. (Make one dashed so you can tell them apart.) Are spaghetti and snack cake, taken together, influential observations? Explain your answer.

2.48 What's my grade? In Professor Friedman's economics course, the correlation between the students' total scores before the final examination and their final examination scores is $r = 0.6$. The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final-exam scores have mean 75 and standard deviation 8. Professor Friedman has lost Julie's final exam but knows that

her total before the exam was 300. He decides to predict her final-exam score from her pre-exam total.
(a) What is the slope of the least-squares regression line of final-exam scores on pre-exam total scores in this course? What is the intercept?
(b) Use the regression line to predict Julie's final-exam score.
(c) Julie doesn't think this method accurately predicts how well she did on the final exam. Calculate r^2 and use the value you get to argue that her actual score could have been much higher (or much lower) than the predicted value.

2.49 Missing work. Data on number of days of work missed and annual salary increase for a company's employees show that in general employees who missed more days of work during the year received smaller raises than those who missed fewer days. Number of days missed explained 64% of the variation in salary increases. What is the numerical value of the correlation between number of days missed and salary increase?

2.50 Will I bomb the final? We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Gary Smith of Pomona College looked at the exam scores of all 346 students who took his statistics class over a 10-year period. (Gary Smith, "Do statistics test scores regress toward the mean?" *Chance*, 10, No. 4 (1997), pp. 42–45.) The least-squares line for predicting final-exam score from midterm exam score was $\hat{y} = 46.6 + 0.41x$.

Octavio scores 10 points above the class mean on the midterm. How many points above the class mean do you predict that he will score on the final? (*Hint:* Use the fact that the least-squares line passes through the point $(\bar{x}, \bar{y})$ and the fact that Octavio's midterm score is $\bar{x} + 10$. This is an example of the phenomenon that gave "regression" its name: students who do well on the midterm will on the average do less well, but still above average, on the final.)

2.51 Four-wheel-drive minivans. Exercise 2.18 gives data on all three 2001 model four-wheel-drive minivans. The least-squares regression line for these data is

$$\hat{y} = -4.5742 + 0.8065x$$

(a) If the minivan with a highway mileage of 28.7 had a city mileage of 19.3 rather than 18.6, how would the least-squares regression line change? Find the least-squares line for the altered data. In words, describe the change in the line resulting from the change in this one observation. What name is given to an observation like this one?
(b) If a fourth observation were added with a highway mileage equal to the average of the highway mileages of the three minivans ($\bar{x} = 29.0667$) and a city mileage of 20, the least-squares regression line for the data set with four observations would be $\hat{y} = -4.2909 + 0.8065x$. In words, describe the change in the line resulting from adding this particular "new" observation. (This illustrates the effect an outlier has on the least-squares regression line.)

Section 2.4

2.52 The declining farm population. The number of people living on American farms has declined steadily during the twentieth century. Here are data on farm population (millions of persons) from 1935 to 1980:

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

(a) Make a scatterplot of these data and find the least-squares regression line of farm population on year.
(b) According to the regression line, how much did the farm population decline each year on the average during this period? What percent of the observed variation in farm population is accounted for by linear change over time?
(c) Use the regression equation to predict the number of people living on farms in 1990. Is this result reasonable? Why?

2.53 Stock market indexes. The Standard & Poor's 500 stock index is an average of the price of 500 stocks. There is a moderately strong correlation (roughly $r = 0.6$) between how much this index changes in January and how much it changes during the entire year. If we looked instead at data on all 500 individual stocks, we would find a quite different correlation. Would the correlation be higher or lower? Why?

2.54 Do firefighters make fires worse? Someone says, "There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage." Explain why this reasoning is wrong.

2.55 How's your self-esteem? People who do well tend to feel good about themselves. Perhaps helping people feel good about themselves will help them do better in their jobs and in life. For a time, raising self-esteem became a goal in many schools and companies. Can you think of explanations for the association between high self-esteem and good performance other than "Self-esteem causes better work"?

2.56 Are big hospitals bad for you? A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds x) and the median number of days y that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital? Why?

2.57 Employment numbers. Here are the numbers of paid employees in 1997 and in 2002 in the 10 kinds of business that employ the most people, according to a government census of business.

Business type	2002 paid employees	1997 paid employees
Accommodation and food services	10,836,365	9,451,226
Administrative and support	8,002,725	7,066,658
Construction	6,943,582	5,664,853
Finance and insurance	6,663,714	5,835,214
Health care and social assistance	15,346,996	13,561,579
Information	3,845,583	3,066,167
Manufacturing	14,543,338	16,805,127
Professional, scientific, technical	7,508,866	5,361,210
Retail trade	15,029,339	13,991,103
Wholesale trade	6,034,579	5,796,557

(a) Extrapolation is commonly associated with using a model to predict the response for values of the explanatory variable that are *greater* than the x-values in the data. Why is considering extrapolation of this type meaningless in this particular case?
(b) Would you consider the estimated intercept of the least-squares regression line to be an example of extrapolation for these data? Explain your response.

2.58 Calories and salt in hot dogs. Refer to Exercise 2.13 which gives data on the calories and sodium content for 17 brands of meat hot dogs. A scatterplot suggests that there is one outlier.
(a) Calculate two least-squares regression lines for predicting sodium from calories, one using all of the observations and the other omitting the outlier. Draw both lines on a scatterplot. Does a comparison of the two regression lines show that the outlier is influential? Explain your answer.
(b) A new brand of meat hot dog has 150 calories per frank. How many milligrams of sodium do you estimate that one of these hot dogs contains?

2.59 Is math the key to success in college? Here is the opening of a newspaper account of a College Board study of 15,941 high school graduates:

> *Minority students who take high school algebra and geometry succeed in college at almost the same rate as whites, a new study says.*
>
> *The link between high school math and college graduation is "almost magical," says College Board President Donald Stewart, suggesting "math is the gatekeeper for success in college."*
>
> *"These findings," he says, "justify serious consideration of a national policy to ensure that all students take algebra and geometry."*

(From a Gannett News Service article appearing in the *Lafayette (Indiana) Journal and Courier*, April 23, 1994.) What lurking variables might explain the association between taking several math courses in high school and success in college? Explain why requiring algebra and geometry may have little effect on who succeeds in college.

2.60 Beef consumption. The following table gives data on the amount of beef consumed (pounds per person) and average retail price of beef (dollars per pound) in the United States for the years 1970 to 1993. Because all prices were generally rising during this period, the prices given are "real prices" in 1993 dollars. These

are dollars with the buying power that a dollar had in 1993.
(a) Economists expect consumption of an item to fall when its real price rises. Make a scatterplot of beef consumption y against beef price x. Do you see a relationship of the type expected?
(b) Find the equation of the least-squares line and draw the line on your plot. What proportion of the variation in beef consumption is explained by regression on beef price?
(c) Although it appears that price helps explain consumption, the scatterplot seems to show some nonlinear patterns. Find the residuals from your regression in (b) and plot them against time. Connect the successive points by line segments to help see the pattern. Are there systematic effects of time remaining after we regress consumption on price? (A partial explanation is that beef production responds to price changes only after some time lag.)

Price and consumption of beef, 1970–1993

Year	Price per pound (1993 dollars)	Consumption (lb/capita)	Year	Price per pound (1993 dollars)	Consumption (lb/capita)
1970	3.721	84.62	1982	3.570	77.03
1971	3.789	83.93	1983	3.396	78.64
1972	4.031	85.27	1984	3.274	78.41
1973	4.543	80.51	1985	3.069	79.19
1974	4.212	85.58	1986	2.989	78.83
1975	4.106	88.15	1987	3.032	73.84
1976	3.698	94.36	1988	3.057	72.65
1977	3.477	91.76	1989	3.096	69.34
1978	3.960	87.29	1990	3.107	67.78
1979	4.423	78.10	1991	3.059	66.79
1980	4.098	76.56	1992	2.931	66.48
1981	3.731	77.26	1993	2.934	65.06

2.61 Seafood prices. The price of seafood varies with species and time. The following table gives the prices in cents per pound received in 1970 and 1980 by fishermen and vessel owners for several species.
(a) Plot the data with the 1970 price on the x axis and the 1980 price on the y axis.
(b) Describe the overall pattern. Are there any outliers? If so, circle them on your graph. Do these unusual points have large residuals from a fitted line? Are they influential in the sense that removing them would change the fitted line?
(c) Compute the correlation for the entire set of data. What percent of the variation in 1980 prices is explained by the 1970 prices?
(d) Recompute the correlation discarding the cases that you circled in (b). Do these observations have a strong effect on the correlation? Explain why or why not.
(e) Does the correlation provide a good measure of the relationship between the 1970 and 1980 prices for this set of data? Explain your answer.

Seafood price per pound, 1970 and 1980

Species	1970 price ($)	1980 price ($)	Species	1970 price ($)	1980 price ($)
Cod	13.1	27.3	Tuna, albacore	26.7	80.1
Flounder	15.3	42.4	Clams, soft	47.5	150.7
Haddock	25.8	38.7	Clams, blue, hard	6.6	20.3
Menhaden	1.8	4.5	Lobsters, American	94.7	189.7
Ocean perch	4.9	23.0	Oysters, eastern	61.1	131.3
Salmon, chinook	55.4	166.3	Sea scallops	135.6	404.2
Salmon, coho	39.3	109.7	Shrimp	47.6	149.0

Section 2.5

2.62 Marital status. Refer to the following table. Give the marginal distribution of marital status (in percents) for men.

Marital status and job level

	Marital Status				
Job grade	Single	Married	Divorced	Widowed	Total
1	58	874	15	8	955
2	222	3927	70	20	4239
3	50	2396	34	10	2490
4	7	533	7	4	551
Total	337	7730	126	42	8235

2.63 The top jobs. Refer to the previous exercise. Find the percent of men in each marital status group who have Grade 4 jobs. Draw a bar graph that compares these percents. Explain briefly what the data show.

2.64 Who holds the top jobs? Refer to the previous two exercises. Find the conditional distribution of marital status among men with Grade 4 jobs. (To do this, look only at the "Job grade 4" row in the table.)

2.65 Smoking by students and their parents. Advertising by tobacco companies is believed to encourage students to smoke and use other tobacco products. Another source of "encouragement" may be the example set by parents with respect to tobacco use. Here are data from eight high schools on smoking among students and among their parents (S. V. Zagona (ed.), *Studies and Issues in Smoking Behavior*, University of Arizona Press, 1967, pp. 157–180):

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	1168	1823	1380
Student smokes	188	?416	400

(a) How many students do these data describe?
(b) What percent of these students smoke?

(c) Give the marginal distribution of parents' smoking behavior, both in counts and in percents.

2.66 Smoking by students. Refer to the previous exercise. Calculate the percent of students in each group (neither parent smokes, one parent smokes, both parents smoke) who smoke. One might believe that parents' smoking increases smoking in their children. Do the data support that belief? Briefly explain your response.

2.67 Majors for men and women in business. To study the career plans of young women and men, questionnaires were sent to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded (F. D. Blau and M. A. Ferber, "Career plans and expectations of young women and men," *Journal of Human Resources*, 26 (1991), pp. 581–607):

	Female	Male
Accounting	68	56
Administration	91	40
Economics	?5	?6
Finance	61	59

(a) Find the two conditional distributions of major, one for women and one for men. Based on your calculations, describe the differences between women and men with a graph and in words.
(b) What percent of the students did not respond to the questionnaire? The nonresponse weakens conclusions drawn from these data.

2.68 Here are the row and column totals for a two-way table with two rows and two columns:

a	b	50
c	d	50
60	40	100

Find *two different* sets of counts a, b, c, and d for the body of the table that give these same totals. This shows that the relationship between two variables cannot be obtained from the two individual distributions of the variables.

2.69 Advertising to divers. How far in advance do scuba divers plan diving trips to the Florida Keys? The answer will help in planning advertising to divers. Here are counts from a survey of people on the mailing list of a diving agency:

	Months in Advance			
	Less than 1	1 to 3	4 to 6	7 or more
Men	11	18	10	4
Women	4	9	11	3

(a) Give the marginal distribution (in percents) of how long before a trip divers do their planning. If you want to mail a brochure advertising a diving expedition that will take place in April, when would you mail it?

(b) Find the conditional distributions of advance planning for men and for women. What are the most important differences between the sexes?

2.70 Impulse purchasing. We might expect that shoppers are more likely to use credit cards for "impulse purchases," which they decide to make on the spot, as opposed to purchases they had in mind when they went to the store. Stop every third person leaving a department store with a purchase. A few questions allow us to classify the purchase as impulse or not. Here are the data on how the customer paid (Karen M. Herbert, "Does impulse buying vary by mode of payment?" MS thesis, Purdue University, 1994):

	Payment Method		
	Cash	Check	Credit card
Impulse purchases	14	4	13
Planned purchases	20	11	35

Describe the relationship between type of purchase and payment method by finding and commenting on several percents. What do you think might explain the choice of payment method for impulse purchases?

College undergraduates. The following five exercises are based on the following table. This two-way table reports data on all undergraduate students enrolled in U.S. colleges and universities in the fall of 1997 whose age was known. (*Digest of Education Statistics 2000*, accessed on the National Center for Education Statistics Web site, `nces.ed.gov`.)

Undergraduate college enrollment, fall 1997 (thousands of students)

Age	2-year full-time	2-year part-time	4-year full-time	4-year part-time
Under 18	45	170	83	55
18 to 24	1478	1202	4759	562
25 to 39	421	1344	1234	1273
40 and up	121	748	236	611
Total	2065	3464	6312	2501

2.71 College undergraduates. (a) How many undergraduate students were enrolled in colleges and universities?
(b) What percent of all undergraduate students were 18 to 24 years old in the fall of the academic year?
(c) Find the percent of the undergraduates enrolled in each of the four types of programs who were 18 to 24 years old. Make a bar graph to compare these percents.
(d) The 18 to 24 group is the traditional age group for college students. Briefly summarize what you have learned from the data about the extent to which this group predominates in different kinds of college programs.

2.72 Two-year college students. (a) An association of two-year colleges asks: "What percent of students enrolled part-time at 2-year colleges are 25 to 39 years old?" Find the percent.

(b) A bank that makes education loans to adults asks: "What percent of all 25- to 39-year-old students are enrolled part-time at 2-year colleges?" Find the percent.

2.73 Students' ages. (a) Find the marginal distribution of age among all undergraduate students, first in counts and then in percents. Make a bar graph of the distribution in percents.
(b) Find the conditional distribution of age (in percents) among students enrolled part-time in 2-year colleges and make a bar graph of this distribution.
(c) Briefly describe the most important differences between the two age distributions.
(d) The sum of the entries in the "2-year full-time" column is not the same as the total given for that column. Why is this?

2.74 Older students. Call students aged 40 and up "older students." Compare the presence of older students in the four types of programs with numbers, a graph, and a brief summary of your findings.

2.75 Nontraditional students. With a little thought, you can extract from the table information other than marginal and conditional distributions. The traditional college age group is ages 18 to 24 years.
(a) What percent of all undergraduates fall in this age group?
(b) What percent of students at 2-year colleges fall in this age group?
(c) What percent of part-time students fall in this group?

2.76 Helping cocaine addicts. Cocaine addiction is hard to break. Addicts need cocaine to feel any pleasure, so perhaps giving them an antidepressant drug will help. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine (manufactured by Hoechst-Marion-Roussel) with lithium and a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a dummy drug, used so that the effect of being in the study but not taking any drug can be seen.) One-third of the subjects, chosen at random, received each drug. Here are the results (D. M. Barnes, "Breaking the cycle of addiction," *Science*, 241 (1988), pp. 1029–1030):

	Desipramine	Lithium	Placebo
Relapse	10	18	20
No relapse	14	?6	?4
Total	24	24	24

(a) Compare the effectiveness of the three treatments in preventing relapse. Use percents and draw a bar graph.
(b) Do you think that this study gives good evidence that desipramine actually *causes* a reduction in relapses?

2.77 Employee performance. Four employees are responsible for handling the cash register at your store. One item that needs to be recorded for each sale is the type of payment: cash, check, or credit card. For a number of transactions this information is missing, though. Are certain employees responsible, or is everyone equally guilty in forgetting to record this information? Below is a table summarizing the last 3501 transactions. (Data provided by Duck Worth Wearing of Ames, Iowa.)

	Not recorded	Recorded
Employee 1	68	?897
Employee 2	62	?679
Employee 3	90	1169
Employee 4	39	?497

(a) What percent of all transactions do not have the payment terms recorded?
(b) Compare the reliability of the four employees in recording the payment terms for each transaction they handle. Use percents and draw a bar graph.
(c) Do you think these data provide good evidence that a certain employee (or subset of employees) is causing the high percent of transactions without payment terms recorded?

Chapter 2 Review Exercises

2.78 Faculty salaries. Data on the salaries of full professors in an engineering department at a large midwestern university are given below. The salaries are for the academic years 2002–2003 and 2005–2006.

2002 salary ($)	2005 salary ($)
95,600	113,400
93,800	114,500
104,200	113,300
119,000	134,000
100,600	115,000
108,200	128,000
92,100	100,800
106,700	115,700
102,300	144,200
112,500	133,100
113,100	131,100
122,500	134,400
121,100	130,000
124,000	133,700
153,300	172,800
138,500	144,500

(a) Construct a scatterplot with the 2005 salaries on the vertical axis and the 2002 salaries on the horizontal axis.
(b) Comment on the form, direction, and strength of the data.
(c) Identify the point in the scatterplot corresponding to the ninth row in the data set, and mark it with a different plotting symbol than what you used to mark the points in part (a). Do any other professors in the data set have a 2002 salary like this one? Do any other professors in the data set have a 2005 salary like this one? Does this point follow the same overall pattern as the other points? Describe precisely what is unusual about this point. Is this what we describe as an outlier?
(d) Repeat part (c) for the fifteenth row of the data set.

2.79 Florida condo sales. Florida reappraises real estate every year, and the county appraiser's Web site lists the current "fair market value" of each piece of property. Property usually sells for somewhat more than the appraised market value. Here are the appraised market values and actual selling prices of condominium units sold in a beachfront building over a 19-month period (in thousands of dollars):

Selling price	Appraised value	Month	Selling price	Appraised value	Month
850	758	0	790	605.9	13
900	812.7	1	700	483.8	14
625	504	2	715	585.8	14
1075	956.7	2	825	707.6	14
890	747.9	8	675	493.9	17
810	717.7	8	1050	802.6	17
650	576.6	9	1325	1031.8	18
845	648.3	12	845	586.7	19

(a) Make a scatterplot using the default ranges that your software chooses for the axes.
(b) Make a scatterplot with both axes ranging from 0 to 1500.
(c) Make a scatterplot with both axes ranging from 0 to 3000.
(d) Comment on the effects of changing the range of values represented by the axes.

2.80 Florida condo sales. Refer to the Florida condo sales data in the previous exercise.
(a) Make a scatterplot using the default ranges that your software chooses for the axes.
(b) Make a scatterplot with the horizontal axis ranging from 0 to 1500 and the vertical axis ranging from 0 to 5000.
(c) Make a scatterplot with the horizontal axis ranging from 0 to 5000 and the vertical axis ranging from 0 to 1500.
(d) Comment on the effects of changing the range of values represented by the axes.

2.81 Florida condo sales. Consider the Florida condo sales data in the two previous exercises.
(a) Calculate the correlation, r^2, and the least-squares regression line for these data. Also, create a residual plot.
(b) Add one more condo's data and redo part (a). The additional condo has an appraised value of \$688,750 and a selling price of \$1,248,125.
(c) What do you conclude about the additional condo by inspecting the residual plot from part (b)?
(d) Describe how the additional data point affected the correlation, r^2, the estimated intercept, and the estimated slope.
(e) If the additional data point was for a condo with an appraised value of \$688,750 and a selling price of \$448,125, then how would you expect the correlation, r^2, the estimated intercept, and the estimated slope to change?

2.82 Employment numbers. Refer to the data on number of paid employees presented in Exercise 2.11 to answer the following questions. Use statistical software

for the calculations.
(a) Calculate the correlation between the number of paid employees in 1997 and 2002.
(b) Report the values in the units "millions of employees" rather than as a count of employees (for example, 10,836,365 employees is reported as 10.836365 million employees) and calculate the correlation for the new values.
(c) Now, round the values from part (b) to one decimal place (for example, 10.836365 rounds to 10.8) and calculate the correlation for the rounded data.
(d) How do the calculated correlations in parts (a) and (b) compare? Is this a special case or an example of a general fact? Explain your response.
(e) How do the calculated correlations in parts (b) and (c) compare? Rounding data can sometimes have a considerable effect on subsequent calculations. Comment on the effect of rounding in calculating the correlation for this particular data set.

2.83 Faculty salaries. Data on the salaries of full professors in an engineering department at a large midwestern university were presented in Exercise 2.78. Use statistical software to perform the calculations required in this exercise.
(a) Make a scatterplot of the salary data with the 2005 salaries on the vertical axis and the 2002 salaries on the horizontal axis.
(b) Call the original data set DATA-A, and consider two variations of DATA-A. First, call DATA-B the data set that results from deleting row 9 from DATA-A. Second, call DATA-C the data set that results from deleting row 15 from DATA-A. Now, let r_A be the correlation for DATA-A, r_B the correlation for DATA-B, and r_C the correlation for DATA-C. Before doing any calculations, attempt to put the three correlations in order from smallest to largest.
(c) Calculate the three correlations described in part (b) and report them in order from smallest to largest. Compare your answer to part (b) with the calculated correlations in part (c). How well did you guess?

2.84 Faculty salaries. Refer to the previous exercise.
(a) Report the least-squares line for predicting 2005 salaries from 2002 salaries. In addition to the equation of the line, report s and r^2.
(b) Interpret the value of the estimated slope within the context of these data.
(c) Interpret the value of r^2 within the context of these data.

2.85 Faculty salaries. Continue with the least-squares model from the previous exercise.
(a) With the mathematical definition of the y intercept in mind, interpret the value of the estimated intercept. Is your interpretation a "practical" one given the context of the data? Explain your response.
(b) One might believe that the intercept in this context should be zero. Make a scatterplot of these data with the least-squares regression line superimposed. Adjust the horizontal and vertical axes so that both include the value zero; that is, the point $(0, 0)$ should be visible on your scatterplot. Now, sketch the line one obtains if you replace the least-squares intercept with zero (and leave the slope as it is).
(c) Comment on the fit of the "zero intercept" line compared to the least-squares line. What does the value of the estimated intercept accomplish in addition to its interpretation (or in spite of the interpretation, whichever the case may be)?

2.86 Faculty salaries. Refer to the previous three exercises. Consider fitting a line to the data with the 2005 salaries as the response variable and the 2002 salaries as the explanatory variable.
(a) Calculate the residuals from the least-squares regression and use them to create a residual plot. Also, report the values of s and r^2 for the model.
(b) Row 9 appears to be an outlier. Delete this point from the data set and redo part (a).
(c) Contrast the residual plots in parts (a) and (b).
(d) Describe the change in the values of s and r^2 when row 9 is deleted.

2.87 Faculty salaries. Continue your analysis of the faculty salary data in the previous four exercises.
(a) Row 15 has the potential to be an influential observation because of its unusual x-value. Report the least-squares line for these data with and without row 15. Also, report s and r^2 for both regressions.
(b) Would you classify row 15 as an influential observation? Explain your response.

2.88 A hot stock? It is usual in finance to describe the returns from investing in a single stock by regressing the stock's returns on the returns from the stock market as a whole. This helps us see how closely the stock follows the market. We analyzed the monthly percent total return y on Philip Morris common stock and the monthly return x on the Standard & Poor's 500 stock index, which represents the market, for the period between July 1990 and May 1997. Here are the results:

$$\begin{array}{lll} \overline{x} = 1.304 & s_x = 3.392 & r = 0.5251 \\ \overline{y} = 1.878 & s_y = 7.554 & \end{array}$$

A scatterplot shows no very influential observations.
(a) Find the equation of the least-squares line from this information. What percent of the variation in Philip Morris stock is explained by the linear relationship with the market as a whole?
(b) Explain carefully what the slope of the line tells us about how Philip Morris stock responds to changes in the market. This slope is called "beta" in investment theory.
(c) Returns on most individual stocks have a positive correlation with returns on the entire market. That is, when the market goes up, an individual stock tends also to go up. Explain why an investor should prefer stocks with beta > 1 when the market is rising and stocks with beta < 1 when the market is falling.

2.89 Beta. The previous exercise introduced the financial concept of a stock's "beta." Beta measures the volatility of a stock relative to the overall market. A beta of less than 1 indicates lower risk than the market; a beta of more than 1 indicates higher risk than the market. In February 2001, the information on Apple Computer, Inc., at `finance.yahoo.com` listed a beta of 1.24; the information on Apple Computer, Inc., at `www.nasdaq.com` listed a beta of 1.69. Both Web sites use the S&P 500 to measure changes in the overall market. Using the 60 monthly returns from March 1996 to February 2001, we find the least-squares regression line $\hat{y} = 0.30 + 1.35x$ (y is Apple return, x is S&P 500 return).
(a) The correlation for our 60 months of data is 0.3481. Interpret this correlation

in terms of using the movement of the overall market to predict movement in Apple stock.
(b) We have three different beta values for Apple stock: 1.24, 1.35, and 1.69. What is a likely explanation for this discrepancy?

2.90 Heating a business. Joan is concerned about the amount of energy she uses to heat her small business. She keeps a record of the natural gas she consumes each month over one year's heating season. Here are Joan's data (provided by Robert Dale, Purdue University):

	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
Degree-days per day	15.6	26.8	37.8	36.4	35.5	18.6	15.3	7.9	0.0
Gas consumed per day	520	610	870	850	880	490	450	250	110

(a) Make a scatterplot of these data. There is a strong linear pattern with no outliers.
(b) Find the equation of the least-squares regression line for predicting gas use from degree-days. Draw this line on your graph. Explain in simple language what the slope of the regression line tells us about how gas use responds to outdoor temperature.
(c) Joan adds some insulation during the summer, hoping to reduce her gas consumption. Next February averages 40 degree-days per day, and her gas consumption is 870 cubic feet per day. Predict from the regression equation how much gas the business would have used at 40 degree-days per day last winter before the extra insulation. Did the insulation reduce gas consumption?

2.91 Heating a business. Find the mean and standard deviation of the degree-day and gas consumption data in the previous exercise. Find the correlation between the two variables. Use these five numbers to find the equation of the regression line for predicting gas use from degree-days. Verify that your work agrees with your previous results. Use the same five numbers to find the equation of the regression line for predicting degree-days from gas use. What units does each of these slopes have?

2.92 Size and selling price of houses. The following table provides information on a random sample of 50 houses sold in Ames, Iowa, in the year 2000. (Data provided by the Ames City Assessor, Ames, Iowa.)
(a) Describe the distribution of selling price with a graph and a numerical summary. What are the main features of this distribution?
(b) Make a scatterplot of selling price versus square feet and describe the relationship between these two variables.
(c) Calculate the least-squares regression line for these data. On average, how much does each square foot add to the selling price of a house?
(d) What would you expect the selling price of a 1600-square-foot house in Ames to be?
(e) What percent of the variability in these 50 selling prices can be attributed to differences in square footage?

Houses sold in Ames, Iowa

Selling price ($)	Square footage	Age (years)	Selling price ($)	Square footage	Age (years)
268,380	1897	1	169,900	1686	35
131,000	1157	15	180,000	2054	34
112,000	1024	35	127,000	1386	50
112,000	935	35	242,500	2603	10
122,000	1236	39	152,900	1582	3
127,900	1248	32	171,600	1790	1
157,600	1620	33	195,000	1908	6
135,000	1124	33	83,100	1378	72
145,900	1248	35	125,000	1668	55
126,000	1139	39	60,500	1248	100
142,000	1329	40	85,000	1229	59
107,500	1040	45	117,000	1308	60
110,000	951	42	57,000	892	90
187,000	1628	1	110,000	1981	72
94,000	816	43	127,250	1098	70
99,500	1060	24	119,000	1858	80
78,000	800	68	172,500	2010	60
55,790	492	79	123,000	1680	86
70,000	792	80	161,715	1670	1
53,600	980	62	179,797	1938	1
157,000	1629	3	117,250	1120	36
166,730	1889	0	116,500	914	4
340,000	2759	6	117,000	1008	23
195,000	1811	3	177,500	1920	32
215,850	2400	27	132,000	1146	37

2.93 Age and selling price of houses. (a) Using the data from the table above, calculate the least-squares regression line for predicting selling price from age at time of sale, that is, in 2000.
(b) What would you expect for the selling price of a house built in 2000? 1999? 1998? 1997? Describe specifically how age relates to selling price.
(c) Would you trust this regression line for predicting the selling price of a house that was built in 1900? 1899? 1850? Explain your responses. (What would this line predict for a house built in 1850?)
(d) Calculate and interpret the correlation between selling price and age.

CHAPTER 3

Section 3.1

3.1 Sampling employees. A firm wants to understand the attitudes of its minority managers toward its system for assessing management performance. Below is a list of all the firm's managers who are members of minority groups. Use Table B at line 139 to choose 6 to be interviewed in detail about the performance appraisal system.

Agarwal	Dewald	Huang	Puri
Anderson	Fernandez	Kim	Richards
Baxter	Fleming	Liao	Rodriguez
Bowman	Garcia	Mourning	Santiago
Brown	Gates	Naber	Shen
Castillo	Goel	Peters	Vega
Cross	Gomez	Pliego	Wang

3.2 Focus group sampling. Thirty individuals in your target audience have been using a new product. Each person has filled out short evaluations of the product periodically during the test period. At the end of the period, you decide to select 4 of the individuals at random for a lengthy interview. The list of participants appears below. Choose an SRS of 4, using the table of random digits, beginning at line 145.

Armstrong	Gonzalez	Kempthorne	Robertson
Aspin	Green	Laskowsky	Sanchez
Bennett	Gupta	Liu	Sosa
Bock	Gutierrez	Montoya	Tran
Breiman	Harter	Patnaik	Trevino
Collins	Henderson	Pirelli	Wu
Dixon	Hughes	Rao	
Edwards	Johnson	Rider	

3.3 Sampling retail locations. You must choose an SRS of 10 of the 440 retail outlets in New York that sell your company's products. How would you label this population? Use Table B, starting at line 105, to choose your sample.

3.4 Campaign contributions. Here are two wordings for the same question. The first question was asked by presidential candidate Ross Perot, and the second by a *Time*/CNN Poll, both in March 1993. (Mitofsky, "Mr. Perot, you're no pollster," *New York Times*, March 27, 1993.)

A. Should laws be passed to eliminate all possibilities of special interests giving huge sums of money to candidates?

B. Should laws be passed to prohibit interest groups from contributing to campaigns, or do groups have a right to contribute to the candidates they support?

One of these questions drew 40% favoring banning contributions; the other drew 80% with this opinion. Which question produced the 40% and which got 80%? Explain why the results were so different.

3.5 Make it an experiment! In the following observational studies, describe changes that could be made to the data collection process that would result in an experiment rather than an observation study. Also, offer suggestions about unseen biases or lurking variables that may be present in the studies as they are described here.
(a) In a sample of 50 members of a local health club, you find that 12 of these members meet weekly with a physical fitness trainer and that the average body mass index (BMI) of these 12 members is less than the average BMI of the other 38 club members in your sample.
(b) In a sample of 12 bank tellers at a local branch office, the 7 tellers who have completed the advanced training program offered by the bank have a lower average error rate in the processing of transactions than the remaining 5 tellers.

3.6 Aren't voluntary response samples convenient? You are asked to develop an ad campaign for a headache relief product. You decide to collect information from potential customers on what their most common causes of headaches are, so that you might gear the ads toward these causes and thus connect with the potential customers.
(a) Describe how you could obtain a voluntary response sample from the population of potential customers. What bias might your sample have as a result of your data collection method?
(b) Describe how you could obtain a convenience sample from the population of potential customers. What bias might your sample have as a result of your data collection method?

3.7 Table B or not Table B. Numbering the members of a population in sequence (001, 002, 003, ...) and using Table B to select a random sample of integers corresponding to members of the numbered population is only one way of choosing an SRS. Most statistical software programs have random number generators, and some will select a random sample from any list of numbers or text that you provide. Note that not all random number generators and sampling programs are equally good at their task. Be sure to investigate your software's reputation for random number generation. Another option is to use the random number generation features available on some Web sites. The same caution applies here, though: be sure to investigate a site's reputation for random number generation before depending on the site for random samples.
(a) Inspect the Web site www.random.org. Write a short description, in your own words, of how Random.org generates random numbers.
(b) Does Random.org offer any evidence of how well it generates random numbers? Describe what evidence you find at the site.
(c) Use the site to generate a random sample of size 5 from a population of 30 items. Report the labels of the 5 items in your random sample.

(d) Repeat part (c) and report this second random sample of 5 items. Is there any overlap between your two random samples?

3.8 Table B or not Table B. Another Web site for random numbers is www.randomizer.org. Randomizer.org uses a different method of generating random numbers than the one used by Random.org.
(a) Inspect the Web site www.randomizer.org. Write a short description, in your own words, of how Randomizer.org generates random numbers.
(b) Does Randomizer.org offer any evidence of how well it generates random numbers? Describe what evidence you find at the site.
(c) Use the site to generate a random sample of size 5 from a population of 30 items. Report the labels of the 5 items in your random sample.
(d) Repeat part (c) and report your second random sample of items. Is there any overlap between your two random samples?

3.9 Instant opinion. The Excite Poll can be found online at `poll.excite.com`. The question appears on the screen, and you simply click buttons to vote "Yes," "No," or "Not sure." On January 25, 2000, the question was "Should female athletes be paid the same as men for the work they do?" In all, 13,147 respondents (44%) said "Yes," another 15,182 (50%) said "No," and the remaining 1448 said "Don't know."
(a) What is the sample size for this poll?
(b) That's a much larger sample than standard sample surveys. In spite of this, we can't trust the result to give good information about any clearly defined population. Why?
(c) It is still true that more men than women use the Web. How might this fact affect the poll results?

3.10 Mail to Congress. You are on the staff of a member of Congress who is considering a bill that would require all employers to provide health insurance for their employees. You report that 1128 letters dealing with the issue have been received, of which 871 oppose the legislation. "I'm surprised that most of my constituents oppose the bill. I thought it would be quite popular," says the congresswoman. Are you convinced that a majority of the voters opposes the bill? State briefly how you would explain the statistical issue to the congresswoman.

3.11 Quality control sampling. A manufacturer of chemicals chooses 3 containers from each lot of 25 containers of a reagent to test for purity and potency. Below are the control numbers stamped on the bottles in the current lot. Use Table B at line 111 to choose an SRS of 3 of these bottles.

A1096	A1097	A1098	A1101	A1108
A1112	A1113	A1117	A2109	A2211
A2220	B0986	B1011	B1096	B1101
B1102	B1103	B1110	B1119	B1137
B1189	B1223	B1277	B1286	B1299

3.12 Design your own bad sample. Your college wants to gather student opinion about parking for students on campus. It isn't practical to contact all students.

(a) Give an example of a way to choose a sample of students that is poor practice because it depends on voluntary response.
(b) Give an example of a bad way to choose a sample that doesn't use voluntary response.

3.13 Random digits. Which of the following statements are true of a table of random digits, and which are false? Briefly explain your answers.
(a) There are exactly four 0s in each row of 40 digits.
(b) Each pair of digits has chance 1/100 of being 00.
(c) The digits 0000 can never appear as a group, because this pattern is not random.

3.14 Systematic random samples. The last stage of the Current Population Survey chooses clusters of households within small areas called blocks. The method used is **systematic random sampling.** An example will illustrate the idea of a systematic sample. Suppose that we must choose 4 clusters out of 100. Because $100/4 = 25$, we can think of the list as four lists of 25 clusters. Choose 1 of the first 25 at random, using Table B. The sample will contain this cluster and the clusters 25, 50, and 75 places down the list from it. If 13 is chosen, for example, then the systematic random sample consists of the clusters numbered 13, 38, 63, and 88.
(a) Use Table B to choose a systematic random sample of 5 clusters from a list of 200. Enter the table at line 120.
(b) Like an SRS, a systematic sample gives all individuals the same chance to be chosen. Explain why this is true, then explain carefully why a systematic sample is nonetheless *not* an SRS.

3.15 Is this an SRS? A company employs 2000 male and 500 female engineers. A stratified random sample of 50 female and 200 male engineers gives each engineer one chance in 10 to be chosen. This sample design gives every individual in the population the same chance to be chosen for the sample. Is it an SRS? Explain your answer.

3.16 A stratified sample. A company employs 2000 male and 500 female engineers. The human resources department wants to poll the opinions of a random sample of engineers about the company's performance review system. To give adequate attention to female opinion, you will choose a stratified random sample of 200 males and 200 females. You have alphabetized lists of female and male engineers. Explain how you would assign labels and use random digits to choose the desired sample. Enter Table B at line 122 and give the labels of the first 5 females and the first 5 males in the sample.

3.17 Do the people want a tax cut? During the 2000 presidential campaign, the candidates debated what to do with the large government surplus. The Pew Research Center asked two questions of random samples of adults. Both said that Social Security would be "fixed." Here are the uses suggested for the remaining surplus:

> *Should the money be used for a tax cut, or should it be used to fund new government programs?*

Should the money be used for a tax cut, or should it be spent on programs for education, the environment, health care, crime-fighting and military defense?

One of these questions drew 60% favoring a tax cut; the other, only 22%. Which wording pulls respondents toward a tax cut? Why?

Section 3.2

3.18 Sickle-cell disease. Sickle-cell disease is an inherited disorder of the red blood cells that in the United States affects mostly blacks. It can cause severe pain and many complications. Bristol-Myers Squibb markets Hydrea—a brand name for the drug hydroxyurea—to treat sickle-cell disease. Federal regulations allow the sale of a new drug only after statistically designed experiments (called clinical trials in medical language) show that the drug is safe and effective. A clinical trial at the National Institutes of Health gave hydroxyurea to 150 sickle-cell sufferers and a placebo (a dummy medication) to another 150. The researchers then counted the episodes of pain reported by each subject. What are the subjects, the factors, the treatments, and the response variables?

3.19 Sealing food packages. A manufacturer of food products uses package liners that are sealed at the top by applying heated jaws after the package is filled. The customer peels the sealed pieces apart to open the package. What effect does the temperature of the jaws have on the force needed to peel the liner? To answer this question, engineers prepare 20 pairs of pieces of package liner. They seal 5 pairs of each at 250°F, 275°F, 300°F, and 325°F. Then they measure the force needed to peel each seal.
(a) What are the individuals studied?
(b) There is one factor (explanatory variable). What is it, and what are its levels?
(c) What is the response variable?

3.20 Sealing food packages. Use a diagram to describe a completely randomized experimental design for the package liner experiment of the previous exercise. (Show the size of the groups, the treatment each group receives, and the response variable.) Use software or Table B, starting at line 120, to do the randomization required by your design.

3.21 An industrial experiment. A chemical engineer is designing the production process for a new product. The chemical reaction that produces the product may have higher or lower yield, depending on the temperature and the stirring rate in the vessel in which the reaction takes place. The engineer decides to investigate the effects of combinations of two temperatures (50°C and 60°C) and three stirring rates (60 rpm, 90 rpm, and 120 rpm) on the yield of the process. She will process two batches of the product at each combination of temperature and stirring rate.
(a) What are the individuals and the response variable in this experiment?
(b) How many factors are there? How many treatments? Use a diagram to lay out the treatments.
(c) How many individuals are required for the experiment?

3.22 Does child care help recruit employees? Will providing child care for employees make a company more attractive to women, even those who are unmarried? You are designing an experiment to answer this question. You prepare recruiting material for two fictitious companies, both in similar businesses in the same location. Company A's brochure does not mention child care. There are two versions of Company B's material, identical except that one describes the company's on-site child care facility. Your subjects are 40 unmarried women who are college seniors seeking employment. Each subject will read recruiting material for both companies and choose the one she would prefer to work for. You will give each version of Company B's brochure to half the women. You expect that a higher percent of those who read the description that includes child care will choose Company B.
(a) Outline an appropriate design for the experiment.
(b) The names of the subjects appear below. Use Table B, beginning at line 131, to do the randomization required by your design. List the subjects who will read the version that mentions child care.

Abrams	Danielson	Gutierrez	Lippman	Rosen
Adamson	Durr	Howard	Martinez	Sugiwara
Afifi	Edwards	Hwang	McNeill	Thompson
Brown	Fluharty	Iselin	Morse	Travers
Cansico	Garcia	Janle	Ng	Turing
Chen	Gerson	Kaplan	Quinones	Ullmann
Cortez	Green	Kim	Rivera	Williams
Curzakis	Gupta	Lattimore	Roberts	Wong

3.23 Comparing weight-loss treatments. Twenty overweight females have agreed to participate in a study of the effectiveness of 4 weight-loss treatments: A, B, C, and D. The company researcher first calculates how overweight each subject is by comparing the subject's actual weight with her "ideal" weight. The subjects and their excess weights in pounds are

Birnbaum	35	Hernandez	25	Moses	25	Smith	29
Brown	34	Jackson	33	Nevesky	39	Stall	33
Brunk	30	Kendall	28	Obrach	30	Tran	35
Cruz	34	Loren	32	Rodriguez	30	Wilansky	42
Deng	24	Mann	28	Santiago	27	Williams	22

The response variable is the weight lost after 8 weeks of treatment. Because a subject's excess weight will influence the response, a block design is appropriate.
(a) Arrange the subjects in order of increasing excess weight. Form 5 blocks of 4 subjects each by grouping the 4 least overweight, then the next 4, and so on.
(b) Use Table B to randomly assign the 4 subjects in each block to the 4 weight-loss treatments. Be sure to explain exactly how you used the table.

3.24 Cash bonuses for the unemployed. Will cash bonuses speed the return to work of unemployed people? The Illinois Department of Employment Security designed an experiment to find out. The subjects were 10,065 people aged 20 to 54 who were filing claims for unemployment insurance. Some were offered $500 if they

found a job within 11 weeks and held it for at least 4 months. Others could tell potential employers that the state would pay the employer $500 for hiring them. A control group got neither kind of bonus. (Based on Stephen A. Woodbury and Robert G. Spiegelman, "Bonuses to workers and employers to reduce unemployment: randomized trials in Illinois," *American Economic Review,* 77 (1987), pp. 513–530.)
(a) Suggest a few response variables of interest to the state and outline the design of the experiment.
(b) How will you label the subjects for random assignment? Use Table B at line 127 to choose the first 3 subjects for the first treatment.

3.25 Poultry-processing plants. The air in poultry-processing plants often contains fungus spores. Inadequate ventilation can affect the health of the workers. The problem is most serious during the summer. To measure the presence of spores, air samples are pumped to an agar plate, and "colony-forming units (CFUs)" are counted after an incubation period. Here are data from two locations in a plant that processes 37,000 turkeys per day, taken on four days in the summer. The units are CFUs per cubic meter of air. (Michael W. Peugh, "Field Investigation of ventilation and air quality in duck and turkey slaughter plants," MS thesis, Purdue University, 1996.)

	Day 1	Day 2	Day 3	Day 4
Kill room	3175	2526	1763	1090
Processing	529	141	362	224

(a) Explain carefully why these are matched pairs data.
(b) The spore count is clearly higher in the kill room. Give sample means to demonstrate the difference for these data.

3.26 Clinical trial basics. Fizz Laboratories, a pharmaceutical company, has developed a new pain-relief medication. Three hundred patients suffering from arthritis and needing pain relief are available. Each patient will be treated and asked an hour later, "About what percent of pain relief did you experience?"
(a) Why should Fizz not simply administer the new drug and record the patients' responses?
(b) Outline the design of an experiment to compare the drug's effectiveness with that of aspirin and of a placebo.
(c) Should patients be told which drug they are receiving? How would this knowledge probably affect their reactions?
(d) If patients are not told which treatment they are receiving, the experiment is single-blind. Should this experiment be double-blind also? Explain.

3.27 Treating prostate disease. A large study used records from Canada's national health care system to compare the effectiveness of two ways to treat prostate disease. The two treatments are traditional surgery and a new method that does not require surgery. The records described many patients whose doctors had chosen each method. The study found that patients treated by the new method were significantly more likely to die within 8 years. (Based on Christopher Anderson, "Measuring what works in health care," *Science*, 263 (1994), pp. 1080–1082.)
(a) Further study of the data showed that this conclusion was wrong. The extra

deaths among patients who got the new method could be explained by lurking variables. What lurking variables might be confounded with a doctor's choice of surgical or nonsurgical treatment?
(b) You have 300 prostate patients who are willing to serve as subjects in an experiment to compare the two methods. Use a diagram to outline the design of a randomized comparative experiment.

3.28 Prayer and meditation. Not all effective medical treatments come from pharmaceutical companies. You read in a magazine that "nonphysical treatments such as meditation and prayer have been shown to be effective in controlled scientific studies for such ailments as high blood pressure, insomnia, ulcers, and asthma." Explain in simple language what the article means by "controlled scientific studies" and why such studies can show that meditation and prayer are effective treatments for some medical problems.

3.29 Sickle-cell disease. Sickle-cell disease is an inherited disorder of the red blood cells that in the United States affects mostly blacks. It can cause severe pain and many complications. Bristol-Myers Squibb markets Hydrea—a brand name for the drug hydroxyurea—to treat sickle-cell disease. Federal regulations allow the sale of a new drug only after statistically designed experiments (called clinical trials in medical language) show that the drug is safe and effective. A clinical trial at the National Institutes of Health gave hydroxyurea to 150 sickle-cell sufferers and a placebo (a dummy medication) to another 150. The researchers then counted the episodes of pain reported by each subject.
(a) Use a diagram to outline the design of this experiment.
(b) Use of a placebo is considered ethical if there is no effective standard treatment to give the control group. It might seem humane to give all the subjects hydroxyurea in the hope that it will help them. Explain clearly why this would not provide information about the effectiveness of the drug. (In fact, the experiment was stopped ahead of schedule because the hydroxyurea group had only half as many pain episodes as the control group. Ethical standards required stopping the experiment as soon as significant evidence became available.)

3.30 Absorption of a drug. Researchers at a pharmaceutical company studying the absorption of a drug into the bloodstream inject the drug (the treatment) into 25 people (the subjects). The response variable is the concentration of the drug in a subject's blood, measured 30 minutes after the injection. This experiment has a single factor with only one level. If three different doses of the drug are injected, there is still a single factor (the dosage of the drug), now with three levels. The three levels of the single factor are the treatments that the experiment compares. Use a diagram to outline a completely randomized design for this experiment.

3.31 Absorption of a drug: another step. The drug studied in the previous exercise can be administered by injection, by a skin patch, or by intravenous drip. Concentration in the blood may depend both on the dose and on the method of administration. Make a sketch that describes the treatments formed by combining dosage and method. Then use a diagram to outline a completely randomized design for this two-factor experiment.

3.32 Potatoes. A horticulturist is comparing two methods (call them A and B) of growing potatoes. Standard potato cuttings will be planted in small plots of ground. The response variables are number of tubers per plant and fresh weight (weight when just harvested) of vegetable growth per plant. There are 20 plots available for the experiment. Sketch a rectangular field divided into 5 rows of 4 plots each. Then diagram the experimental design and do the required randomization. (If you use Table B, start at line 145.) Mark on your sketch which growing method you will use in each plot.

Section 3.3

3.33 Unlisted telephone numbers. A telemarketing firm in Los Angeles uses a device that dials residential telephone numbers in that city at random. Of the first 100 numbers dialed, **43** are unlisted. This is not surprising, because **52%** of all Los Angeles residential phones are unlisted. Which of the bold numbers is a parameter and which is a statistic?

3.34 Indianapolis voters. Voter registration records show that **68%** of all voters in Indianapolis are registered as Republicans. To test a random digit dialing device, you use the device to call 150 randomly chosen residential telephones in Indianapolis. Of the registered voters contacted, **73%** are registered Republicans. Which of the bold numbers is a parameter and which is a statistic?

3.35 Making movies, making money. During the 1990s, a total of 1986 movies were released in the United States. We used statistical software to choose several SRSs of size 25 from this population. Our first sample of 25 movies had mean domestic gross sales 33.916 million dollars, standard deviation 50.2697 million dollars, and median 16.1 million dollars. A second random sample of 25 movies had mean 38.712 million dollars, standard deviation 58 million dollars, and median 14.1 million dollars. The following table lists the members of a third SRS. (The gross domestic sales of all U.S.-released movies for the 1990s were obtained from `worldwideboxoffice.com`.)
(a) Calculate the mean domestic gross $\bar{x}$ for the sample in the table. How does it compare with the means of the first two samples?
(b) Calculate the standard deviation s. How does it compare with those of the first two samples?
(c) Calculate the median M. How does it compare with the previous two medians?

Sample of 25 movies released in the 1990s

Movie name (year)	Domestic gross ($ millions)
Aces: Iron Eagle III (1992)	2.5
BASEketball (1998)	7.0
Body of Evidence (1993)	13.3
Car 54, Where Are You? (1994)	1.2
City of Angels (1998)	78.9
Coneheads (1993)	21.3
Days of Thunder (1990)	82.7
Death Warrant (1990)	16.9
Desperate Hours (1990)	2.7
Ernest Scared Stupid (1991)	14.1
Executive Decision (1996)	68.8
For Love of the Game (1999)	35.2
I Come in Peace (1990)	4.3
Jumanji (1995)	100.2
Kika (1993)	2.1
Miami Rhapsody (1995)	5.2
Mighty, The (1998)	2.6
Perez Family, The (1995)	2.8
Revenge (1990)	15.7
Shine (1996)	35.8
So I Married an Axe Murderer (1993)	11.6
Thinner (1996)	15.2
Wedding Singer, The (1998)	80.2
Wing Commander (1999)	11.6
Xizao (1999)	1.2

3.36 Canada's national health care. The Ministry of Health in the Canadian province of Ontario wants to know whether the national health care system is achieving its goals in the province. Much information about health care comes from patient records, but that source doesn't allow us to compare people who use health services with those who don't. So the Ministry of Health conducted the Ontario Health Survey, which interviewed a probability sample of 61,239 people who live in Ontario. (Warren McIsaac and Vivek Goel, "Is access to physician services in Ontario equitable?" Institute for Clinical Evaluative Sciences in Ontario, October 18, 1993.)

(a) What is the population for this sample survey? What is the sample?

(b) The survey found that 76% of males and 86% of females in the sample had visited a general practitioner at least once in the past year. Do you think these estimates are close to the truth about the entire population? Why?

Section 3.4

How successful are customer support chats? The following six exercises deal with the following scenario. Some companies now offer customer support via online live chats with customer support staff. The records of the last 96,621 online chats

at one company's online customer support Web site are to be sampled and studied to estimate the success rate of these chats in satisfying the customers who use them. The method that will eventually be used to estimate p, the proportion of chats that are successful, is as follows:

1. Obtain an SRS of 150 chats from the 96,621 available chat records, and
2. count the number of successful chats out of the 150 (call this x), then
3. calculate the proportion of successful chats in the sample as $\hat{p} = x/150$.

We should not think that our $\hat{p}$ will equal p. The probability of this is zero. The real question to be addressed is how far off *could* our $\hat{p}$ be from p? This question gets at the heart of how well the method of estimating p described above works.

3.37 Step 1: Worst possible cases. The value of p (whatever it happens to be) is a number between 0 and 1, including both of those values as possibilities. The same statement is true about *any* $\hat{p}$-value that we might obtain from an application of the method described above.
(a) For the moment, take only the fact that both p and $\hat{p}$ are numbers between 0 and 1 and explain how far off $\hat{p}$ could be from p. Provide a sketch to support your conclusion.
(b) Now, imagine that, instead of the method described above, we simply take $\hat{p}$ to *always* be 0.5. Using this method, how far off could $\hat{p}$ be from p? Provide a sketch to support your conclusion.

3.38 Step 2: What-if cases. (a) What if the proportion of successful chats is (unknown to us) actually 0.2? Taking the fact that any $\hat{p}$-value we end up with is a number between 0 and 1, explain how far off $\hat{p}$ could be from p. Provide a sketch to support your conclusion.
(b) If we were to use the "always use 0.5 as your estimate" approach described in part (b) of Step 1, how far off could $\hat{p}$ be from p? Provide a sketch to support your conclusion.

3.39 Step 3: More what-if cases.
(a) Repeat parts (a) and (b) of Step 2 assuming that p is actually 0.7.
(b) Repeat parts (a) and (b) of Step 2 assuming that p is actually 0.5.

3.40 Step 4: Finally, some data. You decide that you have a handle on how estimating p works based on the previous exercises, so you decide to apply the method to the 96,621 chat records. You obtain an SRS of 150 chats and find that 122 of them were successful.
(a) Calculate the value of $\hat{p}$ for this sample.
(b) How far off *could* this $\hat{p}$ be from p? Explain your calculation and/or your reasoning.
(c) How far off *is* this $\hat{p}$ from p? Explain your calculation and/or your reasoning.

3.41 Step 5: How well did it work? Your work so far should convince you that evaluating *how well a method works* is best done by applying the method in a case where the answer is known. In this series of exercises, this means applying

the method in a case where p is known. This is an important point. If we already know the value of p, then we are (obviously) *not* applying the method to *estimate* p. Rather, we are applying the method to see how well the method works, and to see how well the method works we must apply it in a case where we already know the answer.

The task of inspecting all 96,621 chat records is undertaken. The end result is that 83,384 of the chats are found to be successful; that is, $p = 83,384/96,621 = 0.863$ (rounded to three decimal places). Notice that if our *only* goal is to estimate p, then we are done now. In fact, we've done better than estimate p; we have actually *calculated* p. However, our goal is to see how well the prescribed method works in general, not just to estimate p for this one scenario.
(a) For the SRS obtained in Step 4, how well did the method of estimating p work?
(b) Another SRS of size 150 is obtained from the 96,621 chat records. For this SRS, the number of successful chats is 133. Calculate $\hat{p}$ for this sample. How well did the method work for this sample?
(c) You now have accomplished two realizations of the method of estimating p. Can you say how well the method of estimating p works *in general?* Why or why not? If not, what would you need in order to evaluate how well the method works in general?

3.42 Step 6: How well does it work? The method of estimating p is applied to the 96,621 chat records a total of 1200 times. The resulting 1200 $\hat{p}$-values have the following characteristics:

- The average of the 1200 $\hat{p}$-values is 0.8634.
- The standard deviation of the 1200 $\hat{p}$-values is 0.0281.
- The histogram of the 1200 $\hat{p}$-values is distinctly symmetric and mound-shaped.

(a) Sketch the histogram of the 1200 $\hat{p}$-values.
(b) Given the shape of the distribution of the 1200 $\hat{p}$-values, we can apply the 68–95–99.7 rule to these $\hat{p}$-values. Approximately 95% of the $\hat{p}$-values are inside what interval?
(c) Explain how the interval in part (b) provides a reasonable response to the question "How far off could $\hat{p}$ be from p?"

3.43 Unemployment. The Bureau of Labor Statistics announces that last month it interviewed all members of the labor force in a sample of 55,000 households; **6.2%** of the people interviewed were unemployed. Is the bold number a parameter or a statistic? Why?

3.44 Acceptance sampling. A carload lot of ball bearings has a mean diameter of **2.503** centimeters (cm). This is within the specifications for acceptance of the lot by the purchaser. The inspector happens to inspect 100 bearings from the lot with a mean diameter of **2.515** cm. This is outside the specified limits, so the lot is mistakenly rejected. Is each of the bold numbers a parameter or a statistic? Explain your answers.

3.45 Margin of error. A *New York Times* opinion poll on women's issues contacted a sample of 1025 women and 472 men by randomly selecting telephone numbers. The *Times* publishes descriptions of its polling methods. Here is part of the description for this poll:

> *In theory, in 19 cases out of 20 the results based on the entire sample will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all adult Americans.*
>
> *The potential sampling error for smaller subgroups is larger. For example, for men it is plus or minus five percentage points.*

(From the *New York Times*, August 21, 1989.) Explain why the margin of error is larger for conclusions about men alone than for conclusions about all adults.

Chapter 3 Review Exercises

3.46 In your own words. A friend mentions a report about high rates of the disease lupus appearing to coincide with high rates of discarded petroleum products in several nearby counties. Using this scenario, write a paragraph or two explaining to your friend the concepts of *observational study, experiment,* and *confounding.*

3.47 Coupon effect? A researcher studying the effect of coupons on consumers' expectations makes up two different series of ads for a hypothetical brand of cola for the past year. Students in a family science course view one or the other sequence of ads on a computer. Some students see a sequence of ads with no coupon offered on the cola, while others see regular coupon offerings that effectively lower the price of the cola temporarily. Next, the students are asked what price they would expect to pay for the cola.
(a) Is this study an experiment? Why?
(b) What are the explanatory and response variables?

3.48 California area codes. A poll of opinion in California uses random digit dialing to choose telephone numbers at random. Numbers are selected separately within each California area code. The size of the sample in each area code is proportional to the population living there.
(a) What is the name for this kind of sampling design?
(b) California area codes, in rough order from north to south, are

530	707	916	209	415	925	510	650	408	831	805	559	760
661	818	213	626	323	562	709	310	949	909	858	619	

Another California survey does not call numbers in all area codes but starts with an SRS of 10 area codes. Choose such an SRS. If you use Table B, start at line 111.

3.49 Sampling entrepreneurs. A group of business school researchers wanted information on "why some firms survive while other firms with equal economic performance do not." Here are some bare facts about the study's data. Write a brief discussion of the difficulty of gathering data from this population, using this study as an example. Include nonresponse rates in percents.

> *The researchers sent approximately 13,000 questionnaires to members of the National Federation of Independent Businesses who reported that they had recently become business owners. Responses were obtained from 4814 entrepreneurs, of whom 2294 had become owners during the preceding 17 months. Follow-up questionnaires were sent to these 2294 people after one year and after two years. After two years, 963 replied that their firms had survived and 171 firms had been sold. From other sources, the authors could identify 611 firms that no longer existed. The remaining businesses did not respond and could not be identified as discontinued or sold. The study then compared the 963 surviving firms with the 611 known to have failed.*

(Javier Gimeno et al., "Survival of the fittest? Entrepreneurial human capital and the persistence of under-performing firms," *Administrative Science Quarterly*, 42, No. 4 (1997).)

3.50 The price of digital cable TV. A cable television provider plans to introduce "digital cable" to its current market. Digital cable offers more channels and features than standard cable television, but at a higher price. The provider wants to gauge how many consumers would sign up for digital service. You must plan a sample survey of households in the company's service area. You feel that these groups may respond differently:

- Households that currently subscribe to standard cable television service.
- Households that currently subscribe to some other form of television service such as satellite television.
- Households that do not subscribe to any pay television service.

Briefly discuss the design of your sample.

3.51 Did you vote? When the Current Population Survey asked the adults in its sample of 55,000 households if they had voted in the 1996 presidential election, 54% said they had. In fact, only 49% of the adult population voted in that election. Why do you think the CPS result missed by much more than its margin of error?

3.52 Do antioxidants prevent cancer? People who eat lots of fruits and vegetables have lower rates of colon cancer than those who eat little of these foods. Fruits and vegetables are rich in "antioxidants" such as vitamins A, C, and E. Will taking antioxidants help prevent colon cancer? A clinical trial studied this question with 864 people who were at risk of colon cancer. The subjects were divided into four groups: daily beta-carotene, daily vitamins C and E, all three vitamins every day, and daily placebo. After four years, the researchers were surprised to find no significant difference in colon cancer among the groups. (The study is described in G. Kolata, "New study finds vitamins are not cancer preventers," *New York Times*, July 21, 1994. Look in the *Journal of the American Medical Association* of the same date for the details.)

(a) What are the explanatory and response variables in this experiment?
(b) Outline the design of the experiment. Use your judgment in choosing the group sizes.
(c) Assign labels to the 864 subjects and use Table B, starting at line 118, to choose the first 5 subjects for the beta-carotene group.
(d) The study was double-blind. What does this mean?
(e) What does "no significant difference" mean in describing the outcome of the study?
(f) Suggest some lurking variables that could explain why people who eat lots of fruits and vegetables have lower rates of colon cancer. The experiment suggests that these variables, rather than the antioxidants, may be responsible for the observed benefits of fruits and vegetables.

3.53 McDonald's versus Wendy's. The food industry uses taste tests to improve products and for comparison with competitors. Do consumers prefer the taste of a cheeseburger from McDonald's or from Wendy's in a blind test in which neither burger is identified? Describe briefly the design of a matched pairs experiment to investigate this question. How will you be sure the comparison is "blind"?

CHAPTER 4

Section 4.1

4.1 Simulating an opinion poll. A recent opinion poll showed that about 65% of the American public have a favorable opinion of the software company Microsoft. Suppose that this is exactly true. Choosing a person at random then has probability 0.65 of getting one who has a favorable opinion of Microsoft. Use the *What Is Probability?* applet or your statistical software to simulate choosing many people independently. (In most software, the key phrase to look for is "Bernoulli trials." This is the technical term for independent trials with Yes/No outcomes. Our outcomes here are "favorable" or not.)
(a) Simulate drawing 20 people, then 80 people, then 320 people. What proportion have a favorable opinion of Microsoft in each case? We expect (but because of chance variation we can't be sure) that the proportion will be closer to 0.65 in larger runs.
(b) Simulate drawing 20 people 10 times and record the percents in each trial who have a favorable opinion of Microsoft. Then simulate drawing 320 people 10 times and again record the 10 percents. Which set of 10 results is less variable? We expect the results of larger trials to be more predictable (less variable) than the results of smaller trials. That is "long-run regularity" showing itself.

Section 4.2

4.2 Causes of death. Government data on job-related deaths assign a single occupation for each such death that occurs in the United States. The data on occupational deaths in 1999 show that the probability is 0.134 that a randomly chosen death was agriculture–related, and 0.119 that it was manufacturing–related. What is the probability that a death was either agriculture–related or manufacturing–related? What is the probability that the death was related to some other occupation?

4.3 Rating the economy. A Gallup Poll (June 11–17, 2001) interviewed a random sample of 1004 adults (18 years or older). The people in the sample were asked how they would rate economic conditions in the United States today. Here are the results:

Outcome	Probability
Excellent	0.03
Good	0.39
Fair	?
Poor	0.12
No opinion	0.01

These proportions are probabilities for the random phenomenon of choosing an adult at random and asking the person's opinion on current economic conditions.
(a) What must be the probability that the person chosen says "fair"? Why?
(b) The official press release focused on the percent of adults giving the economy a "positive" rating where positive is defined as "good" or "excellent." What is this probability?

4.4 Self-employed workers. Draw a self-employed worker at random and record the industry in which the person works. "At random" means that we give every such person the same chance to be the one we choose. That is, we choose an SRS of size 1. The probability of any industry is just the proportion of all self-employed workers who work in that industry—if we drew many such workers, this is the proportion we would get. Here is the probability model:

Industry	Probability
Agriculture	0.130
Construction	0.147
Finance, insurance, real estate	0.059
Manufacturing	0.042
Mining	0.002
Services	0.419
Trade	0.159
Transportation, public utilities	0.042

(a) Show that this is a legitimate probability model.
(b) What is the probability that a randomly chosen worker does not work in agriculture?
(c) What is the probability that a randomly chosen worker works in construction, manufacturing, or mining?

4.5 Job satisfaction. We can use the results of a poll on job satisfaction to give a probability model for the job satisfaction rating of a randomly chosen employed (full-time or part-time) American. Here is the model:

Rating	Completely satisfied	Somewhat satisfied	Somewhat dissatisfied	Completely dissatisfied
Probability	?	0.47	0.12	0.02

(a) What is the probability of a randomly selected employed American being completely satisfied with his or her job? Why?
(b) What is the probability that a randomly selected employed American will be dissatisfied with his or her job?

4.6 Miles per gallon for 2001 vehicles. The Normal distribution with mean $\mu = 21.2$ miles per gallon and standard deviation $\sigma = 5.4$ miles per gallon is an approximate model for the city gas mileage of 2001 model year vehicles. Let X be the city miles per gallon of one 2001 vehicle chosen at random.
(a) Write the event "the vehicle chosen has a city miles per gallon of 32 or higher" in terms of X.
(b) Find the probability of this event.

4.7 NAEP math scores. Scores on the National Assessment of Educational Progress 12th-grade mathematics test for the year 2000 were approximately Normal with mean 300 points (out of 500 possible) and standard deviation 35 points. Let Y stand for the score of a randomly chosen student. Express each of the following events in terms of Y and use the 68–95–99.7 rule to give the approximate probability.

(a) The student has a score above 300.
(b) The student's score is above 370.

4.8 Filtering junk mail. A majority of email messages are now "spam," or junk mail. Email providers, including colleges, businesses, Internet service providers (ISPs), and free email providers (like Yahoo and Google), are engaged in fighting the onslaught of junk mail. Some kinds of spam messages are more bothersome to email users than others. Consider choosing a spam email message at random. Here is the distribution of topics:

Topic	Adult	Financial	Health	Leisure	Products	Scams
Probability	0.145	0.162	0.073	0.078	0.210	0.142

(a) What is the probability that a spam email does not concern one of these topics?
(b) Corinne is particularly annoyed by spam offering "adult" content (that is, pornography) and scams. What is the probability that a randomly chosen spam email falls into one or the other of these categories?
(c) What are the chances that a randomly chosen spam message is not selling a product?

4.9 World Internet usage. According to information summarized by `InternetWorldStats.com`, approximately 16.6% of the world's population uses the Internet (as of January 2007). Furthermore, a randomly chosen Internet user has the following probabilities of being from the given region of the world:

Region	Asia	Europe	North America	Latin America/Caribbean	Other
Probability	0.356	0.286	0.212	0.081	?

(a) What is the probability of a randomly chosen Internet user not being from one of the four regions explicitly listed in this table?
(b) What is the probability of a randomly chosen Internet user living in either Asia or Europe?
(c) What is the probability of a randomly chosen Internet user not living in North America?

4.10 Broadband preferences. According to information presented by the National Telecommunications and Information Administration (NTIA), approximately 19.9% of all U.S. households have high-speed (broadband) Internet (as of October 2003). Of these broadband Internet households, the following broadband technologies are in use, with the specified percents:

Technology	Cable	DSL	MMDS	Satellite
Percent of broadband users	56.4	41.6	01.1	00.9

(a) Do all broadband households appear to be accounted for by this table? Explain your response.
(b) What is the probability of a randomly chosen broadband household having neither cable nor DSL Internet access?
(c) What is the probability of a randomly chosen U.S. household having broadband Internet access?

4.11 Car colors. Choose a new car or light truck at random and note its color. Here are the probabilities of the most popular colors for cars made in North America in 2000:

Color	Silver	White	Black	Dark green	Dark blue	Medium red
Probability	0.176	0.172	0.113	0.089	0.088	0.067

What is the probability that the car you choose has any color other than the six listed? What is the probability that a randomly chosen car is either silver or white?

4.12 Colors of M&M's. The colors of candies such as M&M's are carefully chosen to match consumer preferences. The color of an M&M drawn at random from a bag has a probability distribution determined by the proportions of colors among all M&M's of that type.
(a) Here is the distribution for plain M&M's:

Color	Brown	Red	Yellow	Green	Orange	Blue
Probability	0.3	0.2	0.2	0.1	0.1	?

What must be the probability of drawing a blue candy?
(b) The probabilities for peanut M&M's are a bit different. Here they are:

Color	Brown	Red	Yellow	Green	Orange	Blue
Probability	0.2	0.2	0.2	0.1	0.1	?

What is the probability that a peanut M&M chosen at random is blue?
(c) What is the probability that a plain M&M is any of red, yellow, or orange? What is the probability that a peanut M&M has one of these colors?

4.13 Crispy M&M's. Exercise 4.12 gives the probabilities that an M&M candy is each of brown, red, yellow, green, orange, and blue. "Crispy Chocolate" M&M's are equally likely to be any of these colors. What is the probability of any one color?

Section 4.3

4.14 Selling mobile phones. You own a mobile phone store with two locations, one in the local mall and one downtown on Main Street. Let X be the number of phones sold during the next month at the mall location, and let Y be the number of phones sold during the next month at the downtown location. For the mall location, you estimate the following probability distribution:

Phones sold	200	300	400
Probability	0.4	0.4	0.2

For the downtown location, your estimated distribution is

Phones sold	100	200	300	400
Probability	0.3	0.5	0.15	0.05

(a) Calculate μ_X and μ_Y.
(b) Your rental cost is higher at the mall location. You make $25 profit on each

phone sold at the mall location and $35 profit on each phone sold at the downtown location. Calculate the mean profit for each location.
(c) Calculate the mean profit for both locations combined.

4.15 Mutual funds. The addition rule for means extends to sums of any number of random variables. Let's look at a portfolio containing three mutual funds. The monthly returns on Fidelity Magellan Fund, Fidelity Real Estate Fund, and Fidelity Japan Fund for the 36 months ending in December 2000 had approximately these means:

W = Magellan monthly return	$\mu_W = 1.14\%$
X = Real Estate monthly return	$\mu_X = 0.16\%$
Y = Japan monthly return	$\mu_Y = 1.59\%$

What is the mean monthly return for a portfolio consisting of 50% Magellan, 30% Real Estate, and 20% Japan?

4.16 Selling mobile phones. Exercise 4.14 gives the distribution of weekly mobile phone sales in two locations.
(a) Calculate the variance and the standard deviation of the number of phones sold at the mall location.
(b) Calculate σ_Y^2 and σ_Y for the downtown location.

4.17 Exercise 4.14 gives the distributions of X, the number of mobile phones sold at the mall location of a store during the next month, and Y, the number of mobile phones sold at the downtown location during the next month. You did some useful variance calculations in Exercise 4.16. Each phone sold at the mall location results in $25 profit, and each phone sold at the downtown location results in $35 profit.
(a) Calculate the standard deviation of the profit for each location using Rule 1 for variances.
(b) Assuming phone sales at the two locations are independent, calculate the standard deviation for total profit of both locations combined.
(c) Assuming $\rho = 0.8$, calculate the standard deviation for total profit of both locations combined.
(d) Assuming $\rho = 0$, calculate the standard deviation for total profit of both locations combined. How does this compare with your result in part (b)? In part (c)?
(e) Assuming $\rho = -0.8$, calculate the standard deviation for total profit of both locations combined. How does this compare with your result in part (b)? In part (c)? In part (d)?

4.18 How many rooms? Furniture makers and others are interested in how many rooms housing units have, because more rooms can generate more sales. Here are the distributions of the number of rooms for owner-occupied units and renter-occupied units in San Jose, California:

Rooms	1	2	3	4	5	6	7	8	9	10
Owned	0.003	0.002	0.023	0.104	0.210	0.224	0.197	0.149	0.053	0.035
Rented	0.008	0.027	0.287	0.363	0.164	0.093	0.039	0.013	0.003	0.003

(a) Make probability histograms of these two distributions, using the same scales. What are the most important differences between the distributions for owner-occupied and rented housing units?
(b) Find the mean number of rooms for both types of housing units. How do the means reflect the differences you found in (a)?

4.19 How many rooms? Which of the two distributions for room counts in Exercise 4.18 appears more spread out in the probability histograms? Why? Find the standard deviation for both distributions. The standard deviation provides a numerical measure of spread.

Section 4.4

4.20 Personal income. The Annual Demographic Supplement to the Current Population Survey interviewed more than 128,000 people in March 2001. The Census Bureau reports that the mean income of the Hispanic males interviewed was **$22,771.** The mean income for the non-Hispanic white males interviewed was **$41,798.** Is each of the bold numbers a parameter or a statistic?

4.21 Roulette. A roulette wheel has 38 slots, of which 18 are black, 18 are red, and 2 are green. When the wheel is spun, the ball is equally likely to come to rest in any of the slots. One of the simplest wagers chooses red or black. A bet of $1 on red returns $2 if the ball lands in a red slot. Otherwise, the player loses his dollar. When gamblers bet on red or black, the two green slots belong to the house. Because the probability of winning $2 is 18/38, the mean payoff from a $1 bet is twice 18/38, or 94.7 cents. Explain what the law of large numbers tells us about what will happen if a gambler makes a large number of bets on red.

4.22 Cotton clothing treatment. "Durable press" cotton fabrics are treated to improve their recovery from wrinkles after washing. Unfortunately, the treatment also reduces the strength of the fabric. The breaking strength of untreated fabric is Normally distributed with mean 58 pounds and standard deviation 2.3 pounds. The same type of fabric after treatment has Normally distributed breaking strength with mean 30 pounds and standard deviation 1.6 pounds. A clothing manufacturer tests 5 specimens of each fabric. All 10 strength measurements are independent.
(a) What is the probability that the mean breaking strength of the 5 untreated specimens exceeds 50 pounds?
(b) What is the probability that the mean breaking strength of the 5 untreated specimens is at least 25 pounds greater than the mean strength of the 5 treated specimens?

4.23 Bottling cola. A bottling company uses a filling machine to fill plastic bottles with cola. The bottles are supposed to contain 300 milliliters (ml). In fact, the contents vary according to a Normal distribution with mean $\mu = 298$ ml and standard deviation $\sigma = 3$ ml.
(a) What is the probability that an individual bottle contains less than 295 ml?
(b) What is the probability that the mean contents of the bottles in a six-pack is less than 295 ml?

4.24 Glucose testing and medical costs. Shelia's doctor is concerned that she may suffer from gestational diabetes (high blood glucose levels during pregnancy). There is variation both in the actual glucose level and in the blood test that measures the level. A patient is classified as having gestational diabetes if the glucose level is above 140 milligrams per deciliter one hour after a sugary drink is ingested. Shelia's measured glucose level one hour after ingesting the sugary drink varies according to the Normal distribution with $\mu = 125$ mg/dl and $\sigma = 10$ mg/dl.
(a) If a single glucose measurement is made, what is the probability that Shelia is diagnosed as having gestational diabetes?
(b) If measurements are made instead on 4 separate days and the mean result is compared with the criterion 140 mg/dl, what is the probability that Shelia is diagnosed as having gestational diabetes?
(c) If Shelia is incorrectly diagnosed with gestational diabetes, then she (and her insurance company) will incur unnecessary additional expenses (insulin, needles, doctor visits) for treating the condition during the remainder of her pregnancy. These additional expenses are greater than the cost of repeating the test on three additional days as suggested in (b). Considering the probabilities you calculated in (a) and (b), comment on which testing method seems more appropriate given that Shelia has an acceptable mean glucose level of 125 mg/dl.

4.25 Pollutants in auto exhausts. The level of nitrogen oxides (NOX) in the exhaust of a particular car model varies with mean 0.9 grams per mile (g/mi) and standard deviation 0.15 g/mi. A company has 125 cars of this model in its fleet.
(a) What is the approximate distribution of the mean NOX emission level $\overline{x}$ for these cars?
(b) What is the level L such that the probability that $\overline{x}$ is greater than L is only 0.01? (*Hint:* This requires a backward Normal calculation.)

4.26 Glucose testing and medical costs. In Exercise 4.24, Shelia's measured glucose level one hour after ingesting the sugary drink varies according to the Normal distribution with $\mu = 125$ mg/dl and $\sigma = 10$ mg/dl. Find the level L such that there is a probability of only 0.05 that the mean glucose level of 4 test results falls above L for Shelia's glucose level distribution. What is the value of L? (*Hint:* This requires a backward Normal calculation.)

CHAPTER 5

Section 5.1

5.1 Telemarketing. Telephone marketers and opinion polls use random digit dialing equipment to call residential telephone numbers at random. The telephone polling firm Zogby International reports that the probability that a call reaches a live person is 0.2. Calls are independent.
(a) A telemarketer places 5 calls. What is the probability that none of them reaches a person?
(b) When calls are made to New York City, the probability of reaching a person is only 0.08. What is the probability that none of 5 calls made to New York City reaches a person?

Blood types. *All human blood can be "ABO-typed" as one of O, A, B, or AB, but the distribution of the types varies a bit among groups of people. Here is the distribution of blood types for a randomly chosen person in the United States:*

Blood type	O	A	B	AB
U.S. probability	0.45	0.40	0.11	0.04

Choose a married couple at random. It is reasonable to assume that the blood types of husband and wife are independent and follow this distribution. Exercises 5.2 to 5.5 concern this setting.

5.2 Is transfusion safe? Someone with type B blood can safely receive transfusions only from persons with type B or type O blood. What is the probability that the husband of a woman with type B blood is an acceptable blood donor for her?

5.3 Same type? What is the probability that a wife and husband share the same blood type?

5.4 Blood types, continued. What is the probability that the wife has type A blood and the husband has type B? What is the probability that one of the couple has type A blood and the other has type B?

5.5 Don't forget Rh. Human blood is typed as O, A, B, or AB and also as Rh-positive or Rh-negative. ABO type and Rh-factor type are independent because they are governed by different genes. In the American population, 84% of people are Rh-positive. Give the probability distribution of blood type (ABO and Rh together) for a randomly chosen person.

Section 5.2

5.6 College degrees. Here are the counts (in thousands) of earned degrees in the United States in the 2001–2002 academic year, classified by level and by the gender of the degree recipient:

	Bachelor's	Master's	Professional	Doctorate	Total
Female	645	227	32	18	922
Male	505	161	40	26	732
Total	1150	388	72	44	1654

(a) If you choose a degree recipient at random, what is the probability that the person you choose is a woman?
(b) What is the conditional probability that you choose a woman, given that the person chosen received a professional degree?
(c) Are the events "choose a woman" and "choose a professional degree recipient" independent? How do you know?

5.7 Income tax returns. In 2000, the Internal Revenue Service received 129,075,000 individual tax returns. Of these, 10,855,000 reported an adjusted gross income of at least $100,000, and 240,000 reported at least $1 million.
(a) What is the probability that a randomly chosen individual tax return reports an income of at least $100,000? At least $1 million?
(b) If you know that the return chosen shows an income of $100,000 or more, what is the conditional probability that the income is at least $1 million?

Section 5.3

5.8 Births. You observe the sex of the next 20 children born at a local hospital; X is the number of girls among them.

5.9 First girl. A couple decides to continue to have children until their first girl is born; X is the total number of children the couple has.

5.10 Inheriting blood type. Genetics says that children receive genes from their parents independently. Each child of a particular pair of parents has probability 0.25 of having type O blood. If these parents have 5 children, the number who have type O blood is the count X of successes in 5 independent trials with probability 0.25 of a success on each trial. So X has the binomial distribution with $n = 5$ and $p = 0.25$.
(a) What are the possible values of X?
(b) Find the probability of each value of X. Draw a probability histogram to display this distribution. (Because probabilities are long-run proportions, a histogram with the probabilities as the heights of the bars shows what the distribution of X would be in very many repetitions.)

Chapter 5 Review Exercises

Working. *In the language of government statistics, you are "in the labor force" if you are available for work and either working or actively seeking work. The unemployment rate is the proportion of the labor force (not of the entire population) who are unemployed. Here are data from the Current Population Survey for the civilian population aged 25 years and over. The table entries are counts in thousands of people. Exercises 5.11 to 5.13 concern these data.*

Highest education	Total population	In labor force	Employed
Did not finish high school	27,325	12,073	11,139
High school but no college	57,221	36,855	35,137
Less than bachelor's degree	45,471	33,331	31,975
College graduate	47,371	37,281	36,259

5.11 Unemployment rates. Find the unemployment rate for people with each level of education. How does the unemployment rate change with education? Explain carefully why your results show that level of education and being employed are not independent.

5.12 Education and work.
(a) What is the probability that a randomly chosen person 25 years of age or older is in the labor force?
(b) If you know that the person chosen is a college graduate, what is the conditional probability that he or she is in the labor force?
(c) Are the events "in the labor force" and "college graduate" independent? How do you know?

5.13 Education and work, continued. You know that a person is employed. What is the conditional probability that he or she is a college graduate? You know that a second person is a college graduate. What is the conditional probability that he or she is employed?

CHAPTER 6

Section 6.1

6.1 Business mergers. A Gallup Poll asked 1004 adults about mergers between companies. One question was, "Do you think the result is usually good for the economy or bad for the economy?" Forty-three percent of the sample thought mergers were good for the economy. The Gallup press release added

> *For results based on this sample, one can say with 95 percent confidence that the maximum error attributable to sampling and other random effects is plus or minus 3 percentage points. In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.*

The Gallup Poll uses a complex multistage sample design, but the sample percent has approximately a Normal sampling distribution.
(a) The announced poll result was $43\% \pm 3\%$. Can we be certain that the true population percent falls in this interval?
(b) Explain to someone who knows no statistics what the announced result $43\% \pm 3\%$ means.
(c) This confidence interval has the same form we have met earlier:

$$\text{estimate} \pm z^* \sigma_{\text{estimate}}$$

What is the standard deviation σ_{estimate} of the estimated percent?
(d) Does the announced margin of error include errors due to practical problems such as undercoverage and nonresponse?

6.2 Median household income. When the statistic that estimates an unknown parameter has a Normal distribution, a confidence interval for the parameter has the form

$$\text{estimate} \pm z^* \sigma_{\text{estimate}}$$

In a complex sample survey design, the appropriate unbiased estimate of the population mean and the standard deviation of this estimate may require elaborate computations. But when the estimate is known to have a Normal distribution and its standard deviation is given, we can calculate a confidence interval for μ from complex sample designs without knowing the formulas that led to the numbers given.

A report based on the Current Population Survey estimates the 1999 median annual earnings of households as \$40,816 and also estimates that the standard deviation of this estimate is \$191. The Current Population Survey uses an elaborate multistage sampling design to select a sample of about 50,000 households. The sampling distribution of the estimated median income is approximately Normal. Give a 95% confidence interval for the 1999 median annual earnings of households.

6.3 Household income by state. The previous problem reports data on the median household income for the entire United States. In a detailed report based

on the same sample survey, you find that the estimated median income for four-person families in Michigan is $65,467. Is the margin of error for this estimate with 95% confidence greater or less than the margin of error for the national median? Why?

6.4 Bank assets. The 110 banks in an American Bank Association survey had mean assets of 220 million dollars. The standard deviation of their assets was 161. Assume that the sample standard deviation can be used in place of the population standard deviation. Give a 95% confidence interval for μ, the mean assets for all community banks.

6.5 Is the margin of error larger or smaller? In the setting of the previous exercise, would the margin of error for 99% confidence be larger or smaller? Verify your answer by performing the calculations.

6.6 Fuel efficiency. Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the mpg were recorded each time the gas tank was filled, and the computer was then reset. Here are the mpg values for a random sample of 20 of these records:

15.8	13.6	15.6	19.1	22.4	15.6	22.5	17.2	19.4	22.6
19.4	18.0	14.6	18.7	21.0	14.8	22.6	21.5	14.3	20.9

Suppose that the standard deviation of the population is known to be $\sigma = 2.9$ mpg.
(a) What is $\sigma_{\overline{x}}$, the standard deviation of $\overline{x}$?
(b) Give a 95% confidence interval for μ, the mean mpg for this vehicle.

6.7 Speed. Refer to the previous exercise. Here are the values of the average speed in miles per hour (mph) for the same sample:

21.0	19.0	18.7	39.2	45.8	19.8	48.4	21.0	29.1	35.7
31.6	49.0	16.0	34.6	36.3	19.0	43.3	37.5	16.5	34.5

Assume that the standard deviation is 10.3 mph. Estimate the mean speed at which this vehicle was driven with a margin of error for 95% confidence.

6.8 Convert to metric. In the previous exercise you found an estimate with a margin of error for the average speed expressed in miles per hour. Convert your estimate and margin of error to the metric units kilometers per hour (kph). To change mph to kph, multiply by 1.6.

6.9 Bank workers. We have previously examined the wages of a random sample of National Bank black female hourly workers. Here are the data (in dollars per year):

16,015	17,516	17,274	16,555	20,788
19,312	17,124	18,405	19,090	12,641
17,813	18,206	19,338	15,953	16,904

Find a 95% confidence interval for the mean earnings of all black female hourly workers at National Bank. Use $1900 for the standard deviation.

Section 6.2

6.10 Compare student loan debt for different groups of students. One purpose of the National Student Loan Survey is to compare the debt of different subgroups of students. For example, the 525 borrowers who last attended a private four-year college had a mean debt of $21,200, while those who last attended a public four-year college had a mean debt of $17,100. The difference of $4100 is fairly large, but we know that these numbers are estimates of the true means. If we took a different sample, we would get different estimates. Can we conclude from these data that the private four-year students have greater debt than public four-year borrowers? We answer this question by computing the probability of obtaining a difference as large or larger than the observed $4100 assuming that, in fact, there is no difference in the true means. This probability is 0.17. What do you conclude? Illustrate the probability result with a sketch and write a short paragraph explaining your answer.

6.11 Compare student loan debt in two different years. Another purpose of the National Student Loan Survey is to look for changes over time. For example, in 1997, the survey found that the mean debt for undergraduate study was $11,400. How does this compare with the value of $18,900 in the 2002 study? The difference is $7500. To determine if the data provide evidence that there is an increase in borrowing we calculate the probability of observing an increase in mean debt that is $7500 or more under the assumption that there is no difference in the true means. The calculated probability is 0.00004. What do you conclude? Illustrate the probability result with a sketch and write a short paragraph explaining your answer.

6.12 Blood pressure and calcium. A randomized comparative experiment examined whether a calcium supplement in the diet reduces the blood pressure of healthy men. The subjects received either a calcium supplement or a placebo for 12 weeks. The statistical analysis was quite complex, but one conclusion was that "the calcium group had lower seated systolic blood pressure ($P = 0.008$) compared with the placebo group." Explain this conclusion, especially the P-value, as if you were speaking to a doctor who knows no statistics.

6.13 California brushfires. We often see televised reports of brushfires threatening homes in California. Some people argue that the modern practice of quickly putting out small fires allows fuel to accumulate and so increases the damage done by large fires. A detailed study of historical data suggests that this is wrong—the damage has risen simply because there are more houses in risky areas. As usual, the study report gives statistical information tersely. Here is the summary of a regression of number of fires on decade (9 data points, for the 1910s to the 1990s):

> *Collectively, since 1910, there has been a highly significant increase ($r^2 = 0.61$, $P < 0.01$) in the number of fires per decade.*

How would you explain this statement to someone who knows no statistics? Include an explanation of both the description given by r^2 and the statistical significance.

Section 6.4

6.14 Net income of banks. Earlier in the chapter, we asked if the net income of community banks has changed in the last year. If μ is the mean percent change for all such banks, the hypotheses are H_0: $\mu = 0$ and H_a: $\mu \neq 0$. The data come from an SRS of 110 banks and we assume that the population standard deviation is $\sigma = 26.4$. The test rejects H_0 at the 1% level of significance when $z \geq 2.576$ or $z < -2.576$, where

$$z = \frac{\overline{x} - 0}{26.4/\sqrt{110}}$$

Is this test sufficiently sensitive to usually detect a 5% change in net income? Answer this question by calculating the power of the test for the alternative $\mu = 5$.

Chapter 6 Review Exercises

6.15 Company cash flow and investment. How much a company invests in its business depends on how much cash flow it has. What factors influence the relationship between cash flow and investment? Here's a clever suggestion: If an industry has an active market in used equipment, investment will be less sensitive to cash flow ("lower elasticity" in economic jargon). Companies in these industries can borrow easily because lenders know they can sell the company's equipment if it defaults. A study of 270 manufacturing industries measured SHRUSED, the proportion of secondhand equipment in the industry's total investment. The study found that "industries with SHRUSED values above the median have smaller cash flow elasticities than those with lower SHRUSED values; the difference is significant at the 5% level in the full sample." Explain to someone who knows no statistics why this study gives good reason to think that an active used-equipment market really does change the relationship between cash flow and investment.

6.16 Foreign investment and exchange rates. We might suspect that foreign direct investment (FDI), in which U.S. companies buy or build facilities overseas, depends on the rate at which the dollar can be exchanged with foreign currencies. A study of 3036 FDI transactions found that "there is no statistically significant relationship between the level of the exchange rate and foreign investment relative to domestic investment." (a) Explain this conclusion to someone who knows no statistics.
(b) We are reasonably confident that, if there were a relationship between FDI and exchange rate that is large enough to be of interest, this study would have found it. Why?

6.17 Annual household income. A government report gives a 90% confidence interval for the 1999 median annual household income as $\$40,816 \pm \$314$. This result was calculated by advanced methods from the Current Population Survey, a multistage random sample of about 50,000 households.
(a) Would a 95% confidence interval be wider or narrower? Explain your answer.
(b) Would the null hypothesis that the 1999 median household income was $40,000 be rejected at the 10% significance level in favor of the two-sided alternative?

6.18 Annual household income. Refer to the previous exercise. Give a 90% confidence interval for the 1999 median *weekly* household income. Use 52.14 as the number of weeks in a year.

CHAPTER 7

Section 7.1

7.1 The Platinum Gasaver. National Fuelsaver Corporation manufactures the Platinum Gasaver, a device they claim "may increase gas mileage by 22%." Here are the percent changes in gas mileage for 15 identical vehicles, as presented in one of the company's advertisements:

48.3	46.9	46.8	44.6	40.2	38.5	34.6	33.7
28.7	28.7	24.8	10.8	10.4	6.9	−12.4	

Would you recommend use of a t confidence interval to estimate the mean fuel savings in the population of all such vehicles? Explain your answer.

7.2 The cost of Internet access. How much do users pay for Internet service? Here are the monthly fees (in dollars) paid by a random sample of 50 users of commercial Internet service providers in August 2000:

20	40	22	22	21	21	20	10	20	20
20	13	18	50	20	18	15	8	22	25
22	10	20	22	22	21	15	23	30	12
9	20	40	22	29	19	15	20	20	20
20	15	19	21	14	22	21	35	20	22

(a) Make a stemplot of the data. Also make a Normal quantile plot if your software permits. The data are not Normal: there are stacks of observations taking the same values, and the distribution is more spread out in both directions and somewhat skewed to the right. The t procedures are nonetheless approximately correct because $n = 50$ and there are no extreme outliers.
(b) Give a 95% confidence interval for the mean monthly cost of Internet access in August 2000.

7.3 The cost of Internet access. The data in the previous exercise show that many people paid $20 per month for Internet access, presumably because major providers such as AOL charged this amount. Do the data give good reason to think that the mean cost for all Internet users differs from $20 per month?

7.4 Economic impact of the Internet. Refer to the previous two exercises. The Census Bureau estimates that 44 million households had Internet access in 2000. Use your confidence interval from Exercise 7.2 to give a 95% confidence interval for the total amount these households paid in Internet access fees. This is one aspect of the national economic impact of the Internet.

7.5 Does the product lose value when cooked? Cooking can cause loss of vitamin C. In Haiti, gruel ("bouillie" in Creole) is made from wheat-soy blend, sugar, milk, bananas, and seasonings. The researchers collected specimens of gruel prepared in Haitian households and measured the vitamin C content before and after cooking. Here are the results (milligrams per 100 grams of blend, dry basis):

Household	1	2	3	4	5
Before	73	79	86	88	78
After	20	27	29	36	17

It is not possible for cooking to increase the amount of vitamin C. State appropriate hypotheses and carry out a matched pairs t test for these data.

7.6 How much loss? The previous exercise demonstrates that cooking reduces the vitamin C content of food made with wheat-soy blend. This fact is neither new nor surprising. The real question is how much vitamin C is lost. Use the data in the previous exercise to give a 95% confidence interval for the amount of vitamin C lost in preparing and cooking gruel in Haiti.

7.7 Does the product lose value when cooked? Exercise 7.5 gives data on the amount of vitamin C in gruel made from wheat-soy blend in 5 Haitian households before and after cooking. Is there evidence that the median amount of vitamin C is less after cooking? State hypotheses, carry out a sign test, and report your conclusion.

7.8 Oil wells. Data were collected from a sample of oil wells in the Devonian Richmond Dolomite area of the Michigan basin. (Data from J. Marcus Jobe and Hutch Jobe, "A statistical approach for additional infill development," *Energy Exploration and Exploitation*, 18 (2000), 89–103.) The variable reported here is the estimated total amount of oil, in thousands of barrels, that could be recovered from each well.

21.7	43.4	79.5	82.2	56.4
36.6	12.0	53.2	204.9	30.5

Would you recommend use of a t confidence interval to estimate the mean amount of oil in the population of all such oil wells? Use a graphical summary to support your answer.

7.9 t is robust. A manufacturer of small appliances employs a market research firm to estimate retail sales of its products. Here are last month's sales of electric can openers from an SRS of 50 stores in the Midwest sales region:

19	19	16	19	25	26	24	63	22	16
13	26	34	10	48	16	20	14	13	24
34	14	25	16	26	25	25	26	11	79
17	25	18	15	13	35	17	15	21	12
19	20	32	19	24	19	17	41	24	27

(a) Make a stemplot of the data. The distribution is skewed to the right and has several high outliers. The *bootstrap* is a modern computer-intensive tool for getting accurate confidence intervals without the Normality condition. Three bootstrap simulations, each with 10,000 repetitions, give these 95% confidence intervals for mean sales in the entire region: (20.42, 27.26), (20.40, 27.18), and (20.48, 27.28).
(b) Find the 95% t confidence interval for the mean. It is essentially the same as the bootstrap intervals. The lesson is that for sample sizes as large as $n = 50$, t procedures are very robust.

7.10 Fuel efficiency. Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the mpg were recorded each time the gas tank was filled, and the computer was then reset (the vehicle is a 1997 Pontiac Transport van). Here are the mpg values for a random sample of 20 of these records:

15.8	13.6	15.6	19.1	22.4	15.6	22.5	17.2	19.4	22.6
19.4	18.0	14.6	18.7	21.0	14.8	22.6	21.5	14.3	20.9

Give a 95% confidence interval for μ, the mean mpg for this vehicle.

7.11 Speed. Refer to the previous exercise. Here are the values of the average speed in miles per hour (mph) for the same sample:

21.0	19.0	18.7	39.2	45.8	19.8	48.4	21.0	29.1	35.7
31.6	49.0	16.0	34.6	36.3	19.0	43.3	37.5	16.5	34.5

Estimate the mean speed at which this vehicle was driven with a margin of error for 95% confidence.

7.12 Clothing for runners. Your company sells exercise clothing and equipment on the Internet. To design the clothing, you collect data on the physical characteristics of your different types of customers. Here are the weights (in kilograms) for a sample of 24 male runners. Assume that these runners can be viewed as a random sample of your potential male customers.

68.7	61.8	63.2	53.1	62.3	59.7	55.4	58.9
60.9	69.2	63.7	67.8	65.6	65.5	56.0	57.8
66.0	62.9	53.6	65.0	55.8	60.4	69.3	62.1

(a) What is $\sigma_{\overline{x}}$, the standard deviation of $\overline{x}$?
(b) Give a 95% confidence interval for μ, the mean of the population from which the sample is drawn.
(c) Will the interval contain the weights of approximately 95% of similar runners? Explain your answer.

7.13 Pounds versus kilograms. Suppose that the weights of the runners in the previous exercise were recorded in pounds rather than kilograms. Use your answers to the previous exercise, and the fact that 1 kilogram equals 2.2 pounds, to answer these questions.
(a) What is the mean weight of these runners?
(b) What is the standard deviation of the mean weight?
(c) Give a 95% confidence interval for the mean weight of the population of runners that these runners represent.

7.14 Earnings of bank workers. Here are the annual earnings (dollars) of a random sample of 20 workers from a bank.

28,258	22,029	18,233	29,606	32,346
35,176	29,863	18,904	26,477	21,102
17,002	23,596	28,885	26,780	18,698
23,308	27,576	26,497	22,612	23,757

(a) Display the data in a stemplot and, if your software permits, a Normal quantile plot. The distribution shows no strong departures from Normality.
(b) Give a 95% confidence interval for the mean annual earnings of workers at this bank.

7.15 Earnings of bank workers. Does the sample in the previous exercise give good evidence that the mean annual earnings of workers at this bank are greater than $20,000?

7.16 Loss of vitamin C, reconsidered. Exercise 7.5 gives these data on the amount of vitamin C in gruel made from wheat-soy blend in 5 Haitian households before and after cooking:

Household	1	2	3	4	5
Before	73	79	86	88	78
After	20	27	29	36	17

The units are milligrams per 100 grams of blend, dry basis. In Exercise 7.6, you were asked to give a confidence interval for the amount of vitamin C lost. The result is hard to interpret without some background information. A loss of 50 mg/100 g would be of little concern if we started with 5000 mg/100 g but would be serious if we started with 75 mg/100 g. In fact, the specifications call for the blend to contain 98 mg of vitamin C/100 g. The "before" values differ from the specifications due to variation in manufacturing and in handling the product before it was used to prepare gruel in Haiti. Express the "after" data as percent of specifications and give a 95% confidence interval for the mean percent.

7.17 Piano lessons. Do piano lessons improve the spatial-temporal reasoning of preschool children? Neurobiological arguments suggest that this may be true. A study designed to test this hypothesis measured the spatial-temporal reasoning of 34 preschool children before and after six months of piano lessons (data from F. H. Rauscher et al., "Music training causes long-term enhancement of preschool children's spatial-temporal reasoning," *Neurological Research*, 19 (1997), pp. 2–8). The study also included children who took computer lessons and a control group; but we are not concerned with those here. The changes in the reasoning scores are

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

(a) Display the data and summarize the distribution.
(b) Find the mean, the standard deviation, and the standard error of the mean.
(c) Give a 95% confidence interval for the mean improvement in reasoning scores.

7.18 Are the results statistically significant? Using the data in the previous exercise, test the null hypothesis that there is no improvement versus the alternative suggested by the neurobiological arguments. State the hypotheses, and give the test statistic with degrees of freedom and the P-value. What do you conclude? From your answer to part (c) of the previous exercise, what can be concluded from this significance test?

7.19 Effect of storage and shipment on a product. The researchers in the project described in Exercise 7.5 are also interested in whether some of the vitamin C content is destroyed as a result of storage and shipment of the product. The researchers marked a sample of bags at the factory and tested each to determine the vitamin C content. Five months later in Haiti they found the marked bags and again tested their contents. The data consist of two vitamin C measures for each bag, one at the time of production in the factory and the other five months later in Haiti. The units are milligrams of vitamin C per 100 grams of WSB. Here are the data:

Factory	Haiti	Factory	Haiti	Factory	Haiti
44	40	45	38	39	43
50	37	32	40	52	38
48	39	47	35	45	38
44	35	40	38	37	38
42	35	38	34	38	41
47	41	41	35	44	40
49	37	43	37	43	35
50	37	40	34	39	38
39	34	37	40	44	36

(a) Set up hypotheses to examine the question of interest to these researchers.
(b) Perform the significance test and summarize your results.
(c) Find 95% confidence intervals for the mean at the factory, for the mean five months later in Haiti, and for the change.

Section 7.2

7.20 Piano lessons. Do piano lessons improve the spatial-temporal reasoning of preschool children? We examined this question in Exercises 7.17 and 7.18 by analyzing the change in spatial-temporal reasoning of 34 preschool children after six months of piano lessons. Here we examine the same question by comparing the changes of those students with the changes of 44 children in a control group. Here are the data for the children who took piano lessons:

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

The control group scores are

1	–1	0	1	–4	0	0	1	0	–1
0	1	1	–3	–2	4	–1	2	4	2
2	2	–3	–3	0	2	0	–1	3	–1
5	–1	7	0	4	0	2	1	–6	0
2	–1	0	–2						

(a) Display the data and summarize the distributions.
(b) Make a table with the sample size, the mean, the standard deviation, and the standard error of the mean for each of the two groups.

(c) Translate the question of interest into hypotheses, test them, and summarize your conclusions.

7.21 Piano lessons, continued. The previous exercise gives data from a study of the effects of piano lessons. Give a 95% confidence interval that describes the comparison between the children who took piano lessons and the controls.

7.22 Comparing several statistical approaches. Exercises 7.17 and 7.18 (page xxx) and Exercises 7.20 and 7.21 all address the effects of piano lessons on spatial-temporal reasoning. Discuss the relative merits of each approach. (You need not actually do all four exercises.)

7.23 Compare the effectiveness of two treatments. In a study of cereal leaf beetle damage to oats, researchers measured the number of beetle larvae per stem in small plots of oats after randomly applying one of two treatments: no pesticide or Malathion at the rate of 0.25 pound per acre. Here are the data (based on M. C. Wilson et al., "Impact of cereal leaf beetle larvae on yields of oats," *Journal of Economic Entomology*, 62 (1969), pp. 699–702):

Control	2	4	3	4	2	3	3	5	3	2	6	3	4
Treatment	0	1	1	2	1	2	1	1	2	1	1	1	

(a) Is there significant evidence at the 1% level that the mean number of larvae per stem is reduced by Malathion? Be sure to state H_0 and H_a.
(b) These data are far from Normal. Why? The researchers nonetheless used t procedures. Although we might prefer a different approach, t procedures are probably reasonably accurate here. Why?

7.24 Cocaine use and low birth weight. Does cocaine use by pregnant women cause their babies to have low birth weight? To study this question, birth weights of babies of women who tested positive for cocaine during a drug-screening test were compared with the birth weights for women who either tested negative or were not tested, a group we call "other" (data from a study conducted at the Medical University of South Carolina in 1989). Here are the summary statistics. The birth weights are measured in grams.

Group	n	$\overline{x}$	s
Positive test	134	2733	599
Other	5974	3118	672

(a) Formulate appropriate hypotheses and carry out the test of significance for these data.
(b) Give a 95% confidence interval for the mean difference in birth weights.
(c) Discuss the limitations of the study design. What do you believe can be concluded from this study?

Section 7.3

7.25 Piano lessons. In Exercise 7.20 we examined the effect of piano lessons on spatial-temporal reasoning. Do the data provide evidence that would cause us

to suspect that the standard deviation of the children who took piano lessons is different from that of the controls? Set up the hypotheses, perform the significance test, and summarize the results.

7.26 Cocaine use and birth weight: power. Exercise 7.24 (page xxx) summarizes data on cocaine use and birth weight. The study has been criticized because of several design problems. Suppose that you are designing a new study. Based on the results in Exercise 7.24, you think that the true difference in mean birth weights may be about 300 grams (g). A difference this large is clinically important. For planning purposes assume that you will have 100 women in each group and that the common standard deviation is 650 g, a guess that is between the two standard deviations in Exercise 7.24. If you use a pooled two-sample t test with significance level 0.05, what is the power of the test for this design?

Chapter 7 Review Exercises

7.27 Comparing earnings of bank employees. Banks employ many workers paid by the hour as tellers and data clerks and in other capacities. The table below presents the annual earnings for a random sample of hourly workers at National Bank. Suppose that we are interested only in the question of whether or not there is an apparent difference in the salaries of men and women. We will therefore combine the sample data for the two genders. Use the data in the table for an analysis that compares the earnings of men and women. Include a graphical summary, the results of a significance test, and a confidence interval. Summarize your conclusions. Does the finding of a statistically significant difference mean that the bank discriminates?

Annual earnings of hourly workers at National Bank

Black females	Black males	White females	White males	Black females	Black males	White females	White males
$16,015	$18,365	$25,249	$15,100	$17,813	$29,347	$18,002	$17,194
$17,516	$17,755	$19,029	$22,346	$18,206	$19,028	$21,596	$30,383
$17,274	$16,890	$17,233	$22,049	$19,338		$26,885	$18,364
$16,555	$17,147	$26,606	$26,970	$15,953		$24,780	$18,245
$20,788	$18,402	$28,346	$16,411	$16,904		$14,698	$23,531
$19,312	$20,972	$31,176	$19,268			$19,308	
$17,124	$24,750	$18,863	$28,336			$17,576	
$18,405	$16,576	$15,904	$19,007			$24,497	
$19,090	$16,853	$22,477	$22,078			$20,612	
$12,641	$21,565	$19,102	$19,977			$17,757	

7.28 Behavior of pet owners. On the morning of March 5, 1996, a train with 14 tankers of propane derailed near the center of the small Wisconsin town of Weyauwega. Six of the tankers were ruptured and burning when the 1700 residents were ordered to evacuate the town. Researchers study disasters like this so that effective relief efforts can be designed for future disasters. About half of the households with pets did not evacuate all of their pets. A study conducted after the derailment focused on problems associated with retrieval of the pets after the

evacuation and characteristics of the pet owners (data provided by Professor Sebastian Heath, School of Veterinary Medicine, Purdue University). One of the scales measured "commitment to adult animals." The people who evacuated some or all of their pets were compared with those who did not evacuate any of their pets. Higher scores indicate that the pet owner is more likely to take actions that benefit the pet. Here are the data summaries:

Group	n	$\overline{x}$	s
Evacuated all or some pets	116	7.95	3.62
Did not evacuate any pets	125	6.26	3.56

Analyze the data and prepare a short report describing the results.

7.29 Weight-loss programs. In a study of the effectiveness of weight-loss programs, 47 subjects who were at least 20% overweight took part in a group support program for 10 weeks. Private weighings determined each subject's weight at the beginning of the program and 6 months after the program's end. A t test was used to assess the significance of the average weight loss. The paper reporting the study said, "The subjects lost a significant amount of weight over time, $t(46) = 4.68$, $p < 0.01$." (Based loosely on D. R. Black et al., "Minimal interventions for weight control: a cost-effective alternative," *Addictive Behaviors*, 9 (1984), pp. 279–285.) It is common to report the results of statistical tests in this abbreviated style.
(a) Which t test did the study use?
(b) Explain to someone who knows no statistics but is interested in weight-loss programs what the practical conclusion is.
(c) The paper follows the tradition of reporting significance only at fixed levels such as $\alpha = 0.01$. In fact, the results are more significant than "$p < 0.01$" suggests. What can you say about the P-value of the test?

7.30 Preservatives in meat products. Nitrites are often added to meat products as preservatives. In a study of the effect of these chemicals on bacteria, the rate of uptake of a radiolabeled amino acid was measured for a number of cultures of bacteria, some growing in a medium to which nitrites had been added. Here are the summary statistics from this study:

Group	n	$\overline{x}$	s
Nitrite	30	7880	1115
Control	30	8112	1250

Carry out a test of the research hypothesis that nitrites decrease amino acid uptake, and report your results.

7.31 Conditions for inference. The data in Exercise 1.70 gives the populations of all 58 counties in the state of California. Is it proper to apply the one-sample t method to these data to give a 95% confidence interval for the mean population of a California county? Explain your answer.

7.32 Male and female CS students. Is there a difference between the average SAT scores of male and female computer science students? The CSDATA data set, described in the Data Appendix, gives the Math (SATM) and Verbal (SATV) scores

for a group of 224 computer science majors. The variable SEX indicates whether each individual is male or female.

(a) Compare the two distributions graphically, and then use software to compare the average SATM scores of males and females. Is it appropriate to use the pooled t test for this comparison? Write a brief summary of your results and conclusions. Refer to both versions of the t test and also to the F test for equality of standard deviations. Include a 99% confidence interval for the difference in the means.

(b) The students in the CSDATA data set are all computer science majors who entered a major university during a particular year. To what extent do you think that your results would generalize to (i) computer science students entering in different years, (ii) computer science majors at other colleges and universities, and (iii) college students in general?

CHAPTER 8

Section 8.1

8.1 Free throws. Leroy, a starting player for a major college basketball team, made only 38.4% of his free throws last season. During the summer he worked on developing a softer shot in the hope of improving his free-throw accuracy. In the first eight games of this season Leroy made 25 free throws in 40 attempts. Let p be his probability of making each free throw he shoots this season.
(a) State the null hypothesis H_0 that Leroy's free-throw probability has remained the same as last year and the alternative H_a that his work in the summer resulted in a higher probability of success.
(b) Calculate the z statistic for testing H_0 versus H_a.
(c) Do you accept or reject H_0 for $\alpha = 0.05$? Find the P-value.
(d) Give a 90% confidence interval for Leroy's free-throw success probability for the new season. Are you convinced that he is now a better free-throw shooter than last season?
(e) What assumptions are needed for the validity of the test and confidence interval calculations that you performed?

8.2 Student employment. You want to estimate the proportion of students at your college or university who are employed for 10 or more hours per week while classes are in session. You plan to present your results by a 95% confidence interval. Using the guessed value $p^* = 0.35$, find the sample size required if the interval is to have an approximate margin of error of $m = 0.05$.

8.3 Will the upgrade be profitable? To profitably produce a planned upgrade of a software product you make, you must charge customers \$100. Are your customers willing to pay this much? You contact a random sample of 50 customers and find that 17 would pay \$100 for the upgrade. Find a 95% confidence interval for the proportion of all of your customers (the population) who would be willing to buy the upgrade for \$100.

8.4 Will the upgrade be profitable? Refer to the previous exercise. Compute the plus four 95% confidence interval and compare this interval with the one that you obtained in that exercise.

8.5 Holiday shopping. A poll of 811 adults aged 18 or older asked about purchases that they intended to make for the upcoming holiday season. One of the questions asked about what kind of gift they intended to buy for the person on whom they would spend the most. Clothing was the first choice of 487 people. Give a 95% confidence interval for the proportion of people in this population who intend to buy clothing as their first choice.

8.6 New-product sales. Yesterday, your top salesperson called on 5 customers and obtained orders for your new product from all 5. Suppose that it is reasonable to view these 5 customers as a random sample of all of her customers.
(a) Give the plus four estimate of the proportion of her customers who would buy

the new product. Notice that we don't estimate that all customers will buy, even though all 5 in the sample did.
(b) Give the margin of error for 95% confidence. (You may see that the upper endpoint of the confidence interval is greater than 1. In that case, take the upper endpoint to be 1.)
(c) Do the results apply to all of your sales force? Explain why or why not.

8.7 A profitable upgrade? In the exercise titled "Will the upgrade be profitable?," we found that 17 customers from a random sample of 50 would be willing to buy a software upgrade that costs $100. If the upgrade is to be profitable, you will need to sell it to more than 20% of your customers. Do the sample data give good evidence that more than 20% are willing to buy?
(a) Formulate this problem as a hypothesis test. Give the null and alternative hypotheses. Will you use a one-sided or a two-sided alternative? Why?
(b) Carry out the significance test. Report the test statistic and the P-value.
(c) Should you proceed with plans to produce and market the upgrade?

8.8 Yes or no? In a survey of 100 employees, 68 answered "Yes" to a question on work stress. That is, 32 of the sample of 100 answered "No."
(a) Find the 95% confidence interval for the proportion of all employees who would answer "Yes." Give the 95% confidence interval for the proportion who would answer "No." Explain carefully differences between your two results.
(b) Test the null hypothesis that the proportion of "Yes" among all employees is 0.75. Then, carry out a two-sided test of the null hypothesis that the proportion of "No" is 0.25. Explain carefully differences between your two results.

8.9 Do students report Internet sources? The National Survey of Student Engagement found that 87% of students report that their peers at least "sometimes" copy information from the Internet in their papers without crediting the source. Assume that the sample size is 430,000.
(a) Find the margin of error for 99% confidence.
(b) Here are some facts from the report that summarizes the survey. More than 430,000 students from 730 four-year colleges and universities participated. The average response rate was 43% and ranged from 15% to 89%. Institutions paid a participation fee of between $3000 and $7500 based on the size of their undergraduate enrollment. Discuss these as sources of error in this study. How do you think these errors would compare with the error that you calculated in part (a)?

8.10 Student loans. A survey of 1280 student loan borrowers found that 448 had loans totaling more than $20,000 for their undergraduate education. Give a 95% confidence interval for the proportion of all student loan borrowers who have loans of $20,000 or more for their undergraduate education.

8.11 Did the borrowers get a degree? In the survey described in the previous exercise, there were 1050 borrowers whose total debt was $10,000 or more. Of these, 192 left school without completing a degree. Consider the population to be borrowers whose total debt was $10,000 or more. Find a 95% confidence interval

for the proportion of borrowers who left school without completing a degree in this population.

8.12 How would the confidence interval change? Refer to Exercise 8.10. Would a 99% confidence interval be wider or narrower than the one that you found in that exercise? Verify your results by computing the interval.

8.13 How would the confidence interval change? Refer to Exercise 8.11. Would a 90% confidence interval be wider or narrower than the one that you found in that exercise? Verify your results by computing the interval

8.14 Income of your potential customers. In the study described in the previous exercise, 1434 subjects out of a total of 2533 reported that their annual income was $25,000 or more.
(a) Give a 95% confidence interval for the true proportion of subjects in this population with incomes above $25,000.
(b) Do you think that some respondents might not give truthful answers to a question about their income? Discuss the possible effects on your estimate and confidence interval.
(c) In the previous exercise, the question analyzed concerned pet ownership. Compare this question with the income question with respect to the possibility that the respondents were not truthful.

8.15 Should young people have credit cards? In a survey of teens aged 12 to 19 years, 38% said that they thought that credit cards should be limited to adult use.
(a) The survey summary stated that approximately 2000 teens were surveyed. Use this information to construct a 95% confidence interval for the population proportion who think that credit card use should be limited in this way.
(b) Comment on the age range for the survey. Do you think that there might be considerable variation in the responses based on age? Discuss and propose an alternative way to conduct the survey or to report the results.

8.16 Do job applicants lie? When trying to hire managers and executives, companies sometimes verify the academic credentials described by the applicants. One company that performs these checks summarized its findings for a six-month period. Of the 84 applicants whose credentials were checked, 15 lied about having a degree.
(a) Find the sample proportion of applicants who lie about having a degree and the standard error of this estimate.
(b) Consider these data to be a random sample of credentials from a large collection of similar applicants. Give a 95% confidence interval for the true proportion of applicants who lie about having a degree.

8.17 Can you make this inference? Suppose that 9 of the 84 applicants checked in the previous exercise lied about their major. Can we conclude that a total of $24 = 15 + 9$ applicants lied about having a degree or about their major? Explain your answer.

8.18 Alcohol abuse on campus. College presidents have described alcohol abuse as the number one problem on campus. How common is it? A survey of 17,096 students in U.S. four-year colleges collected information on drinking behavior and alcohol-related problems. The researchers defined "frequent binge drinking" as having five or more drinks in a row three or more times in the past two weeks. According to this definition, 3314 students were classified as frequent binge drinkers. Find a 99% confidence interval for the proportion of frequent binge drinkers in this population.

8.19 Bicycle accidents and alcohol. In the United States approximately 900 people die in bicycle accidents each year. One study examined the records of 1711 bicyclists aged 15 or older who were fatally injured in bicycle accidents between 1987 and 1991 and were tested for alcohol. Of these, 542 tested positive for alcohol (blood alcohol concentration of 0.01% or higher).
(a) Summarize the data with appropriate descriptive statistics.
(b) To do statistical inference for these data, we think of p as the probability that a tested bicycle rider is positive for alcohol. Find a 95% confidence interval for p.
(c) Can you conclude from your statistical analysis of this study that alcohol causes fatal bicycle accidents?

8.20 What proportion were legally drunk? The study mentioned in the previous exercise found that 386 bicyclists had blood alcohol levels above 0.10%, a level defined as legally drunk in all states. Give a 95% confidence interval for the proportion who were legally drunk according to this criterion.

8.21 Insect infestations. An entomologist samples a field for egg masses of a harmful insect by placing a yard-square frame at random locations and carefully examining the ground within the frame. An SRS of 70 locations selected from a county's pastureland found egg masses in 14 locations. Give a 95% confidence interval for the proportion of all possible locations that are infested.

Section 8.2

8.22 Did the Yankees have a home field advantage? In the 1996 regular baseball season, the World Series Champion New York Yankees played 80 games at home and 82 games away. They won 49 of their home games and 43 of the games played away. We can consider these games as samples from potentially large populations of games played at home and away. How much advantage does the Yankee home field provide?
(a) Find the proportion of wins for the home games. Do the same for the away games.
(b) Find the standard error needed to compute a confidence interval for the difference in the proportions.
(c) Compute a 90% confidence interval for the difference between the probability that the Yankees win at home and the probability that they win when on the road. Are you convinced that the 1996 Yankees were more likely to win at home?

8.23 Is it easier to win at home? Return to the New York Yankees data in the previous exercise.
(a) Most people think that it is easier to win at home than away. State null and alternative hypotheses to test this supposition.
(b) Combining all of the games played, what proportion did the Yankees win?
(c) Find the standard error needed for testing that the probability of winning is the same at home and away.
(d) Compute the z statistic and its P-value. What conclusion do you draw?

8.24 What about the Mets? In the 2000 World Series the New York Yankees played the New York Mets. An exercise earlier examines the Yankees' home and away victories. During the regular season the Mets won 55 of the 84 home games that they played and 39 of the 81 games that they played away. Perform the same analysis for the Mets as you did in the previous exercise, and write a short summary comparing these results with those you found for the Yankees.

8.25 Who gets stock options? Different kinds of companies compensate their key employees in different ways. Established companies may pay higher salaries, while new companies may offer stock options that will be valuable if the company succeeds. Do high-tech companies tend to offer stock options more often than other companies? One study looked at a random sample of 200 companies. Of these, 91 were listed in the *Directory of Public High Technology Corporations* and 109 were not listed. Treat these two groups as SRSs of high-tech and non-high-tech companies. Seventy-three of the high-tech companies and 75 of the non-high-tech companies offered incentive stock options to key employees. Give a 95% confidence interval for the difference in the proportions of the two types of companies that offer stock options.

8.26 Are the high-techs different? In the previous exercise we looked at whether or not companies offered stock options to their employees. There, we compared high-tech companies with other companies using a 95% confidence interval. Let's now test the null hypothesis that the two types of companies are equally likely to offer this kind of benefit to their employees. In the sample, 73 of the 91 high-tech companies and 75 of the 109 other companies offered incentive stock options to key employees. State appropriate null and alternative hypotheses, compute the test statistic, and report the P-value. Give a brief statement of your conclusion.

8.27 Unhappy HMO customers. A study was designed to find reasons why patients leave a health maintenance organization (HMO). Patients were classified as to whether or not they had filed a complaint with the HMO. We want to compare the proportion of complainers who leave the HMO with the proportion of those who do not file complaints. In the year of the study, 639 patients filed complaints, and 54 of these patients left the HMO voluntarily. For comparison, the HMO chose an SRS of 743 patients who had not filed complaints. Twenty-two of these patients left voluntarily. Give an estimate of the difference in the two proportions with a 95% confidence interval.

8.28 Are the unhappy HMO customers likely to leave? In the previous exercise we examined data from a study designed to find reasons why patients leave a health maintenance organization (HMO). There we compared the proportion of complainers who leave the HMO with the proportion of noncomplainers who leave. In the year of the study, 639 patients filed complaints and 54 of these patients left the HMO voluntarily. For comparison, the HMO chose an SRS of 743 patients who had not filed complaints. Twenty-two of those patients left voluntarily. We expect a higher proportion of complainers to leave. Do the data support this expectation? State hypotheses, find the test statistic and its P-value, and state your conclusion.

8.29 Credit cards and impulse shopping. We might expect that shoppers are more likely to use credit cards for "impulse purchases" (that is, those they decide to make on the spot), as opposed to purchases they had in mind when they went to the store. Stop every third person leaving a department store with a purchase. (This is in effect a random sample of people who buy at that store.) A few questions allow us to classify the purchase as impulse or not. Here are the data on how the customer paid:

	Credit Card?	
	Yes	No
Impulse purchases	13	18
Planned purchases	35	31

(a) Is the difference in credit card use between impulse and planned purchases statistically significant?
(b) Give a 95% confidence interval for the difference.

8.30 Brand loyalty. It's hard to persuade consumers to abandon a product with which they are familiar. One experiment gave consumers free samples of a new laundry detergent and also of a standard detergent. After some time, ask the subjects which detergent they prefer. Among the 48 customers who normally used the standard product, 19 preferred the new product; among the 56 customers who did not use the standard product, 29 preferred the new product. Are current users of the standard detergent less likely than nonusers to prefer the new detergent? Summarize the data and do a test of significance.

8.31 Consider other factors. The reaction of nonusers to the new detergent, reported in the previous exercise, looks promising. Let's consider only people who don't currently use the standard detergent and look at other factors that might influence which detergent they prefer. There were 116 customers who used soft water and 110 customers who used hard water. In the soft-water group, 53 customers preferred the new product, while 42 customers in the hard-water group expressed this preference.

	Soft water	Hard water
Prefer standard product	63	68
Prefer new product	53	42

Compare the preferences of people with hard water and soft water, including a test of significance. Did you learn anything that might be used in advertising the new detergent?

8.32 Support for a marketing program. Many states have "corn checkoff" programs that divert a small part of the sale price of corn (typically one-half cent per bushel). These funds are used to develop new markets, to promote renewable fuels and new technologies that use corn, and to access federal government money for these activities. Some checkoff programs are voluntary and some are mandatory. The Indiana Department of Agriculture asked random samples of corn producers in each county whether they favored a mandatory program. In Tippecanoe County, 263 farmers were in favor of the program and 252 were not. In neighboring Benton County, 260 were in favor and 377 were not.
(a) Find the proportions of farmers in favor of the program in each of the two counties.
(b) Find the standard error needed to compute a confidence interval for the difference in the proportions.
(c) Compute a 99% confidence interval for the difference between the proportions of farmers favoring the program in Tippecanoe County and in Benton County. Do you think opinions differed in the two counties?

8.33 Perform a significance test. Refer to the survey of farmers described in the previous exercise.
(a) Formulate null and alternative hypotheses for comparing the proportions of farmers in the two counties who favor mandatory checkoff.
(b) Combine the two samples and find the overall proportion of farmers who favor the corn checkoff program.
(c) Find the standard error needed for testing the hypotheses you stated in (a).
(d) Compute the z statistic and its P-value. What conclusion do you draw?

8.34 A hazardous work environment. The power takeoff driveline on farm tractors can be a serious hazard to farmers. A shield covers the driveline on new tractors, but for a variety of reasons, the shield is often missing on older tractors. Two types of shield are the bolt-on and the flip-up. A study initiated by the National Safety Council took a sample of older tractors to examine the proportions of shields removed. The study found that 35 shields had been removed from the 83 tractors having bolt-on shields and that 15 had been removed from the 136 tractors with flip-up shields.
(a) Test the null hypothesis that there is no difference between the proportions of the two types of shields removed. Give the z statistic and the P-value. State your conclusion in words.
(b) Give a 90% confidence interval for the difference in the proportions of removed shields for the bolt-on and the flip-up types. Based on the data, what recommendation would you make about the type of shield to be used on new tractors?

8.35 Bicycle accidents, alcohol, and gender. In a previous exercise titled "Bicycle accidents and alcohol," we examined the percent of fatally injured bicyclists tested for alcohol who tested positive. Here are the same data broken down by gender:

Gender	n	X(tested positive)
Female	191	27
Male	1520	515

(a) Summarize the data by giving the two population proportions and a 95% confidence interval for their difference.
(b) The standard error SE_D contains a contribution from each sample, $\hat{p}_1(1-\hat{p}_1)/n_1$ and $\hat{p}_2(1-\hat{p}_2)/n_2$. Which of these contributes the larger amount to the standard error of the difference? Explain why.

8.36 Are the gender differences statistically significant? The proportions of fatally injured female and male bicyclists were compared with a confidence interval in the previous exercise. Examine the same data with a test of significance.

8.37 Lying by job applicants. Is lying about credentials by job applicants changing? In a previous exercise titled "Do job applicants lie?" we looked at the proportion of applicants who lied about having a degree in a six-month period. To see if there is a change over time, we can compare that period with the following six months. Here are the data:

Period	n	X(lied)
1	84	15
2	106	21

Use a 95% confidence interval to address the question of interest.

8.38 Have the lies increased? Data on the proportion of applicants who lied about having a degree in two consecutive six-month periods appear in the previous exercise. Is there evidence of a change over time? State hypotheses, carry out a significance test, and summarize the results.

8.39 Aspirin and stroke. A clinical trial examined the effectiveness of aspirin in the treatment of cerebral ischemia (stroke). Patients were randomized into treatment and control groups. The study was double-blind in the sense that neither the patients nor the physicians who evaluated the patients knew which patients received aspirin and which the placebo tablet. After six months of treatment, the attending physicians evaluated each patient's progress as either favorable or unfavorable. Of the 78 patients in the aspirin group, 63 had favorable outcomes; 43 of the 77 control patients had favorable outcomes.
(a) Compute the proportions of patients having favorable outcomes in the two samples.
(b) Give a 95% confidence interval for the difference between the favorable proportions in the populations.
(c) The physicians conducting the study believed from previous research that aspirin was likely to increase the chance of a favorable outcome. Carry out a significance test to confirm this conclusion. State hypotheses, find the P-value, and write a summary of your results.

8.40 Chromosomes and crime. A study of chromosomal abnormalities and criminality examined data on 4124 Danish males born in Copenhagen. The study used the penal registers maintained in the offices of the local police chiefs and classified each man as having a criminal record or not. Each was also classified as having the normal male XY chromosome pair or one of the abnormalities XYY or XXY. Of

the 4096 men with normal chromosomes, 381 had criminal records, while 8 of the 28 men with chromosomal abnormalities had criminal records. Some experts believe that chromosomal abnormalities are associated with increased criminality. Do these data lend support to this belief? Report your analysis and draw a conclusion.

8.41 Effect of the sample size. A previous exercise titled "Summer employment of college students" looked at undergraduate student summer employment. Similar results from a smaller number of students may not have the same statistical significance. Specifically, suppose that 71 of 78 men surveyed were employed and 62 of 71 women surveyed were employed. The sample proportions are essentially the same as in the earlier exercise.
(a) Compute the z statistic for these data and report the P-value. What do you conclude?
(b) Compare the results of this significance test with your results in the previous exercise. What do you observe about the effect of the sample size on the results of these significance tests?

8.42 Products to control cockroaches. The pesticide diazinon is in common use to treat infestations of the German cockroach, *Blattella germanica*. A study investigated the persistence of this pesticide on various types of surfaces. Researchers applied a 0.5% emulsion of diazinon to glass and plasterboard. After 14 days, they placed 18 cockroaches on each surface and recorded the number that died within 48 hours. On glass, 9 cockroaches died, while on plasterboard, 13 died.
(a) Find a 90% confidence interval for the difference in the two population proportions of dead cockroaches.
(b) Chemical analysis of the residues of diazinon suggests that it may persist longer on plasterboard than on glass because it binds to the paper covering on the plasterboard. The researchers therefore expected the mortality rate to be greater on plasterboard than on glass. Conduct a significance test to assess the evidence that this is true.

8.43 More cockroaches. Suppose that the experiment in the previous exercise placed more cockroaches on each surface and observed similar mortality rates. Specifically, suppose that 36 cockroaches were placed on each surface and that 26 died on the plasterboard, while 18 died on the glass.
(a) Compute the z statistic for these data and report its P-value. What do you conclude?
(b) Compare the results of this significance test with those you gave in the previous exercise. What do you observe about the effect of the sample size on the results of these significance tests?

Chapter 8 Review Exercises

8.44 Gender and top students. Many colleges that once enrolled only male or only female students have become coeducational. Some administrators and alumni were concerned that the academic standards of the institutions would decrease with

the change. One formerly all-male college undertook a study of the first class to contain women. The class consisted of 851 students, 214 of whom were women. An examination of first-semester grades revealed that 15 of the top 30 students were female.
(a) What is the proportion of women in the class? Call this value p_0.
(b) Assume that the number of females in the top 30 is approximately a binomial random variable with $n = 30$ and unknown probability p of success. In this case success corresponds to the student being female. What is the value of $\hat{p}$?
(c) Are women more likely to be top students than their proportion in the class would suggest? State hypotheses that ask this question, carry out a significance test, and report your conclusion.

8.45 Race and diet. In a study on blood pressure and diet, a random sample of Seventh Day Adventists were interviewed at a national meeting. Because many people who belong to this denomination are vegetarians, they are a very useful group for studying the effects of a meatless diet. Blacks in the population as a whole have a higher average blood pressure than whites. A study of this type should therefore take race into account in the analysis. The 312 people in the sample were categorized by race and whether or not they were vegetarians. The data are given in the following table (data provided by Chris Melby and David Goldflies, Department of Physical Education, Health, and Recreation Studies, Purdue University):

	Black	White
Vegetarian	42	135
Not vegetarian	47	88

Are the proportions of vegetarians the same among all black and white Seventh Day Adventists who attended this meeting? Analyze the data, paying particular attention to this question. Summarize your analysis and conclusions. What can you infer about the proportions of vegetarians among black and white Seventh Day Adventists in general? What about blacks and whites in general?

8.46 Marketing travel on the Internet. To devise effective marketing strategies it is helpful to know the characteristics of your customers. A study compared demographic characteristics of people who use the Internet for travel arrangements and of people who do not. Of 1132 Internet users, 643 had completed college. Among the 852 nonusers, 349 had completed college.
(a) Do users and nonusers differ significantly in the proportion of college graduates?
(b) Give a 95% confidence interval for the difference in the proportions.

8.47 Income of the customers. The study mentioned in the previous exercise also asked about income. Among Internet users, 493 reported income of less than \$50,000 and 378 reported income of \$50,000 or more. (Not everyone answered the income question.) The corresponding numbers for nonusers were 477 and 200. Perform a significance test to compare the incomes of users with nonusers and also give an estimate of the difference in proportions with a 95% margin of error.

8.48 Nonresponse for the income question. Refer to the previous two exercises. Give the total number of users and the total number of nonusers for the

analysis of education. Do the same for the analysis of income. The difference is due to respondents who chose "Rather not say" for the income question. Give the proportions of "Rather not say" individuals for users and nonusers. Perform a significance test to compare these and give a 95% confidence interval for the difference. People are often reluctant to provide information about their income. Do you think that this amount of nonresponse for the income question is a serious limitation for this study?

8.49 Credit cards. A Gallup Poll used telephone interviews to survey a sample of 1025 U.S. residents over the age of 18 regarding their use of credit cards. The poll reported that 76% of Americans said that they had at least one credit card. Give the 95% margin of error for this estimate.

8.50 Do they pay off the monthly balance? The Gallup Poll in the previous exercise reported that 41% of those who have credit cards do not pay the full balance each month. Find the number of people in the survey who said that they had at least one credit card, using the information in the previous exercise. Combine this number with the reported 41% to give a margin of error for the proportion of credit card holders who do not pay their full balance.

8.51 Ability of children to distinguish new products. Many new products are targeted toward children. The choice behavior of children with regard to new products is of interest to companies that design marketing strategies for these products. As part of one study, children in different age groups were compared on their ability to sort new products into the correct product category (milk or juice in this case).[1] Here are some of the data:

Age group	n	Number who sorted correctly
4- to 5-year-olds	50	10
6- to 7-year-olds	53	28

Test the null hypothesis that the two age groups are equally skilled at sorting. Justify your choice of an alternative hypothesis. Also, give a 90% confidence interval for the difference. Summarize your results with a short paragraph.

8.52 Binge drinking by men and women. In a previous exercise titled "Alcohol abuse on campus," we estimated the proportion of college students who engage in frequent binge drinking. Are there student characteristics related to this behavior? For example, how similar is frequent binge drinking among men and women? Here are counts of frequent binge drinkers by gender:

Population	n	X
1 (men)	7,180	1,630
2 (women)	9,916	1,684
Total	17,096	3,314

Write a short report on the size and the statistical significance of the male-female difference in frequent binge drinking.

8.53 Blood pressure and the risk of death. In an example given in a previous edition of this text, we discussed a study that examined the association between

high blood pressure and increased risk of death from cardiovascular disease. There were 2676 men with low blood pressure and 3338 men with high blood pressure. In the low-blood-pressure group, 21 men died from cardiovascular disease; in the high-blood-pressure group, 55 died.
(a) Calculate a 95% confidence interval for the difference in proportions.
(b) Do the study data confirm that death rates are higher among men with high blood pressure? State hypotheses, carry out a significance test, and give your conclusions.

8.54 Aspirin and heart disease. A large experiment evaluated the effects of aspirin on cardiovascular disease. The subjects were 5139 male British medical doctors. The doctors were randomly assigned to two groups. One group of 3429 doctors took one aspirin daily, and the other group did not take aspirin. After 6 years, there were 148 deaths from heart attack or stroke in the first group and 79 in the second group. A similar experiment used male American medical doctors as subjects. These doctors were also randomly assigned to one of two groups. The 11,037 doctors in the first group took one aspirin every other day, and the 11,034 doctors in the second group took no aspirin. After nearly 5 years, there were 104 deaths from heart attacks in the first group and 189 in the second. Analyze the data from these two studies and summarize the results. How do the conclusions of the two studies differ, and why?

CHAPTER 9

Chapter 9 Review Exercises

9.1 Wine and music. Music influences our mood and behavior. Marketers try to choose background music that will influence consumers. Does the type of music we hear have an effect on the purchases we make? One study of this question varied the type of background music and observed the effect on wine sales. The percent of French wine among all bottles sold was 52% when French music was playing and 36% when other types of music were playing. (C. M. Ryan, C. A. Northrup-Clewes, B. Knox, and D. I. Thurnham, "The effect of in-store music on consumer choice of wine," *Proceedings of the Nutrition Society*, 57 (1998), p. 1069A.)
(a) Write down the outline of a two-way table in which you would record the data for this study. Your outline will show the number of rows and columns and their labels but no actual data in the cells.
(b) Would you view one variable as a response variable and the other as an explanatory variable? Give a reason for your answer. How did this influence the form of your table outline?

9.2 Admission to business and law school. Mountain View University has professional schools in business and law. Here is a three-way table of applicants to these professional schools, categorized by gender, school, and admission decision. (Francine D. Blau and Marianne A. Ferber, "Career plans and expectations of young women and men," *Journal of Human Resources*, 26 (1991), pp. 581–607.)

Business			Law		
	Admit			Admit	
Gender	Yes	No	Gender	Yes	No
Male	400	200	Male	90	110
Female	200	100	Female	200	200

(a) Make a two-way table of gender by admission decision for the combined professional schools by summing entries in the three-way table.
(b) From your two-way table, compute separately the percents of male and female applicants admitted. Male applicants are admitted to Mountain View's professional schools at a higher rate than female applicants.
(c) Perform the significance test for the combined two-way table. We justify the use of a significance test in this setting by viewing the admissions process that generated these data as the object of our study.
(d) Now compute separately the percents of male and female applicants admitted by the business school and by the law school.
(e) Perform a significance test for the business school data. Do the same for the law school data.
(f) Explain carefully, as if speaking to a skeptical reporter, how it can happen that Mountain View appears to favor males when this is not true within each of the professional schools.

9.3 Construct a similar table. Refer to the previous exercise. Make up a similar table for a hypothetical university having four different schools that illustrates the same point. Carefully summarize your table with the appropriate percents.

9.4 Students change majors. A task force set up to examine retention of students in the majors that they chose when starting college examined data on transfers to other majors. (Data provided by Susan Prohofsky, from her PhD dissertation, "Selection of undergraduate major: the influence of expected costs and expected benefits," Purdue University, 1991.) Here are some data giving counts of students classified by initial major and the area that they transferred to:

	Area transferred to				
Initial major	Engineering	Management	Liberal arts	Other	Total
Biology	13	25	158		398
Chemistry	16	15	19		114
Mathematics	3	11	20		72
Physics	9	5	14		61

Complete the table by computing the values for the "Other" column. Write a short paragraph explaining what conclusions you can draw about the relationship between initial major and area transferred to. Be sure to include numerical and graphical summaries as well as the details of your significance test.

9.5 Majors for men and women in business. A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded (Francine D. Blau and Marianne A. Ferber, "Career plans and expectations of young women and men," *Journal of Human Resources*, 26 (1991), pp. 581–607):

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

(a) Test the null hypothesis that there is no relation between the gender of students and their choice of major. Give a P-value and state your conclusion.
(b) Describe the differences between the distributions of majors for women and men with percents, with a graph, and in words.
(c) Which two cells have the largest terms in the sum that makes up the X^2 statistic? How do the observed and expected counts differ in these cells? (This should strengthen your conclusions in (b).)
(d) Two of the observed cell counts are small. Do the study data satisfy our guidelines for safe use of the chi-square test?
(e) What percent of the students did not respond to the questionnaire? The nonresponse weakens conclusions drawn from these data.

9.6 Survey response rates. To study the export activity of manufacturing firms, researchers mailed questionnaires to an SRS of firms in each of five industries that

export many of their products. The response rate was only 12.5%, because private companies don't like to fill out long questionnaires from academic researchers. Here are data on the planned sample sizes and the actual number of responses received from each industry (Erdener Kaynak and Wellington Kang-yen Kuan, "Environment, strategy, structure, and performance in the context of export activity: an empirical study of Taiwanese manufacturing firms," *Journal of Business Research*, 27 (1993), pp. 33–49):

	Sample size	Responses
Metal products	185	17
Machinery	301	35
Electrical equipment	552	75
Transportation equipment	100	15
Precision instruments	90	12

If the response rates differ greatly, comparisons among the industries may be difficult. Is there good evidence of unequal response rates among the five industries? (Start by creating a two-way table of response or nonresponse by industry.)

9.7 Secondhand stores. Shopping at secondhand stores is becoming more popular and has even attracted the attention of business schools. A study of customers' attitudes toward secondhand stores interviewed samples of shoppers at two secondhand stores of the same chain in two cities. The breakdown of the respondents by gender is as follows: (William D. Darley, "Store-choice behavior for pre-owned merchandise," *Journal of Business Research*, 27 (1993), pp. 17–31.)

	City 1	City 2
Men	38	68
Women	203	150
Total	241	218

Is there a significant difference between the proportions of women customers in the two cities?
(a) State the null hypothesis, find the sample proportions of women in both cities, do a two-sided z test, and give a P-value using Table A.
(b) Calculate the X^2 statistic and show that it is the square of the z statistic. Show that the P-value from Table F agrees (up to the accuracy of the table) with your result from (a).
(c) Give a 95% confidence interval for the difference between the proportions of women customers in the two cities.

9.8 More secondhand stores. The study of shoppers in secondhand stores cited in the previous exercise also compared the income distributions of shoppers in the two stores. Here is the two-way table of counts:

Income	City 1	City 2
Under \$10,000	70	62
\$10,000 to \$19,999	52	63
\$20,000 to \$24,999	69	50
\$25,000 to \$34,999	22	19
\$35,000 or more	28	24

A statistical calculator gives the chi-square statistic for this table as $X^2 = 3.955$. Is there good evidence that customers at the two stores have different income distributions? (Give the degrees of freedom, the P-value, and your conclusion.)

9.9 Child care workers. A large study of child care used samples from the data tapes of the Current Population Survey over a period of several years. The result is close to an SRS of child care workers. The Current Population Survey has three classes of child care workers: private household, nonhousehold, and preschool teacher. Here are data on the number of blacks among women workers in these three classes (David M. Blau, "The child care labor market," *Journal of Human Resources,* 27 (1992), pp. 9–39):

	Total	Black
Household	2455	172
Nonhousehold	1191	167
Teachers	659	86

(a) What percent of each class of child care workers is black?
(b) Make a two-way table of class of worker by race (black or other).
(c) Can we safely use the chi-square test? What null and alternative hypotheses does X^2 test?
(d) The chi-square statistic for this table is $X^2 = 53.194$. What are its degrees of freedom? Use Table F to approximate the P-value.
(e) What do you conclude from these data?

9.10 Mail survey response rate. Can you increase the response rate for a mail survey by contacting the respondents before they receive the survey? A study designed to address this question compared three groups of subjects. (James E. Stafford, "Influence of preliminary contact on mail returns," *Journal of Marketing Research*, 3 (1966), pp. 410-411.) The first group received a preliminary letter about the survey, the second group was phoned, and the third received no preliminary contact. A positive response was defined as returning the survey within two weeks. Here are the counts:

	Intervention		
Response	Letter	Phone call	None
Yes	171	146	118
No	220	68	455
Total	391	214	573

(a) For each intervention find the proportion of positive responses.
(b) Translate the question of interest into appropriate null and alternative hypotheses for this problem.
(c) Give the test statistic, degrees of freedom, and the P-value for the significance test. What do you conclude?

9.11 Planning a study. The survey in the Exercise 9.10 was conducted in 1966 on subjects who were college students in Houston, Texas, and asked about clothing preferences. The data in following exercise were collected in 1989, and the survey questions concerned pregnancy among resident physicians. You are planning a study

to be conducted next month to assess the needs in your community for a new Internet access provider. Discuss whether and how you would use the results of these two surveys to design your study.

9.12 Nonresponse among physicians. Does a prenotification letter affect the response rate of physicians chosen for a sample survey? A study response rate of those who received the letters was compared with that of a control group who did not. Here are the data (Patricia H. Shiono and Mark A. Klebanoff, "The effect of two mailing strategies on the response to a survey of physicians," *American Journal of Epidemiology*, 134 (1991), pp. 539-542):

Response	Letter	No letter
Yes	2570	2645
No	2448	2384
Total	5018	5029

(a) Give the percents of positive responses for those who received letters and those who did not.
(b) Analyze the two-way table. State the hypotheses and give the test statistic, degrees of freedom, P-value, and your conclusion.

9.13 Persistence of fund performance. If the performance of a stock fund is due to the skill of the manager, then we would expect a fund that does well this year to perform well next year also. This is called persistence of fund performance. One study classified funds as losers or winners depending on whether their rate of return was less than or greater than the median of all funds. (Burton G. Malkeil, "Returns from investing in equity mutual funds, 1971 to 1991," *Journal of Finance*, 50 (1995), pp. 549–572.) To examine the question of interest, we form a two-way table that classifies each fund as a loser or winner in each of two successive years. Here is one such table:

	Next year	
This year	Winner	Loser
Winner	85	35
Loser	37	83

Is there evidence in favor of persistence of fund performance in this table? Support your conclusion with a complete analysis of the data.

9.14 Rerun comparing proportions. Rerun the analysis in the previous exercise using the method for comparing two proportions of Section 8.2. Verify that the X^2 statistic is the square of the z statistic and that the P-values for both analyses are the same.

9.15 Retrospective and prospective studies. We have already remarked that the chi-square analysis does not depend on the study design that led to a particular two-way table. Let's verify that this is also true for the "compare two proportions" analysis of 2×2 tables.

Return to the data used in the previous two exercises. These data might result from either of two designs. If we draw separate random samples of winners and losers

this year and record the outcome next year, this is a **prospective study** (forward looking). We would compare the percents of winners and losers next year for the two "this year" groups. You did this in Exercise 9.13. On the other hand, if we draw separate random samples of winners and losers "next year" and look back to see if they were winners or losers in the previous year, we have a **retrospective study** (backward looking). You would now compare the percents of winners and losers this year for the two "next year" groups. Thus, the two designs lead to two different sets of sample proportions. Verify that you nonetheless get the same value of z (and therefore the same P-value) using either the prospective or the retrospective approach.

9.16 Do conclusions generalize? If we find evidence in favor of an effect in one set of circumstances, it is natural to want to conclude that it holds in others. Unfortunately, this is not always true. For example, here is additional data from the study described in the previous four exercises:

	Next year	
This year	Winner	Loser
Winner	96	148
Loser	145	99

Analyze these data. What do you conclude? This set of data is for 1987 to 1988; in the previous exercises, the years were 1977 to 1978. Many things change with time.

9.17 Retention of graduate students. Are there gender differences in the progress of students in doctoral programs? A major university classified all students entering PhD programs in a given year by their status six years later. The categories used were as follows: completed the degree, still enrolled, and dropped out. Here are the data:

Status	Men	Women
Completed	423	98
Still enrolled	134	33
Dropped out	238	98

Assume that these data can be viewed as a random sample giving us information on student progress. Describe the data using whatever percents are appropriate. State and test a null hypothesis and alternative that address the question of gender differences. Summarize your conclusions. What factors not given might be relevant to this study?

9.18 Student loans. A study of 865 college students found that 42.5% had student loans. The students were randomly selected from the approximately 30,000 undergraduates enrolled in a large public university. The overall purpose of the study was to examine the effects of student-loan burdens on the choice of a career. (Data provided by Susan Prohofsky, from her PhD dissertation, "Selection of undergraduate major: the influence of expected costs and expected benefits," Purdue University, 1991.) A student with a large debt may be more likely to choose a field where starting salaries are high so that the loan can more easily be repaid. The following table classifies the students by field of study and whether or not they have a loan.

	Student loan	
Field of study	Yes	No
Agriculture	32	35
Child development and family studies	37	50
Engineering	98	137
Liberal arts and education	89	124
Management	24	51
Science	31	29
Technology	57	71

Carry out a complete analysis of the association between having a loan and field of study, including a description of the association and an assessment of its statistical significance.

9.19 Compare fields of study on the PEOPLE score. In the study described in the previous exercise, students were asked to respond to some questions regarding their interests and attitudes. Some of these questions form a scale called PEOPLE that measures altruism, or an interest in the welfare of others. Each student was classified as low, medium, or high on this scale. Is there an association between PEOPLE score and field of study? Here are the data:

	PEOPLE score		
Field of study	Low	Medium	High
Agriculture	5	27	35
Child development and family studies	1	32	54
Engineering	12	129	94
Liberal arts and education	7	77	129
Management	3	44	28
Science	7	29	24
Technology	2	62	64

Analyze the data and summarize your results. Are there some fields of study that have very large or very small proportions of students in the high-PEOPLE category?

9.20 Women pharmacy students. The proportion of women entering many professions has undergone considerable change in recent years. A study of students enrolled in pharmacy programs describes the changes in this field. A random sample of 700 students in their third or higher year of study at colleges of pharmacy was taken in each of nine years. The following table gives the numbers of women in each of these samples (Data provided by Dr. Susan Meyer, Senior Vice President of the American Association of Colleges of Pharmacy):

Year	1970	1972	1974	1976	1978	1980	1982	1984	1986
Women	164	195	226	283	302	342	369	385	412

Use the chi-square test to assess the change in the percent of women pharmacy students over time, and summarize your results. (You will need to calculate the number of male students for each year using the fact that the sample size each year was 700.) Plot the percent of women versus year. Describe the plot. Is it roughly

linear? Find the least-squares line that summarizes the relation between time and the percent of women pharmacy students.

9.21 Does the trend continue? Refer to the previous exercise. Here are the actual percents of women pharmacy students for the years 1987 to 2000:

Year	1987	1988	1989	1990	1991	1992	1993
Women	60.0%	60.6%	61.6%	62.4%	63.0%	63.4%	63.2%
Year	1994	1995	1996	1997	1998	1999	2000
Women	63.3%	63.4%	63.8%	64.2%	64.4%	64.9%	65.9%

Plot these percents versus year and summarize the pattern. Using your analysis of the data in this and the previous exercise, write a report summarizing the changes that have occurred in the percent of women pharmacy students from 1970 to 2000. Include an estimate of the percent for the year 2004 with an explanation of why you chose this estimate.

9.22 Dissatisfied customers. Customers who are dissatisfied with a product often discard it. Marketers need to understand factors related to dissatisfaction. Unfortunately, many pet owners are dissatisfied customers. Euthanasia of healthy but unwanted pets by animal shelters is believed to be the leading cause of death for cats and dogs. A study designed to find factors associated with bringing a cat to an animal shelter compared data on cats that were brought to the Humane Society of Saint Joseph County in Mishawaka, Indiana, with controls, cats from the same county that were not brought in. (Gary J. Patronek et al., "Risk factors for relinquishment of cats to an animal shelter," *Journal of the American Veterinary Medical Association*, 209 (1996), pp. 582-588.) One of the factors examined was the source of the cat; the categories were private owner or breeder, pet store, and other (includes born in home, stray, and obtained from a shelter). This kind of study is called a **case-control study** by epidemiologists. Here are the data:

	Source		
Group	Private	Pet store	Other
Cases (brought to shelter)	124	16	76
Controls (not brought)	219	24	203

(a) Should you use row or column percents to describe these data? Give reasons for your answer and summarize the pattern that you see in these percents.
(b) Use a significance test to examine the association between the two variables. Summarize the results.

9.23 What about dogs? The investigators responsible for the study of cats in the previous exercise did a similar study for dogs. (Gary J. Patronek et al., "Risk factors for relinquishment of dogs to an animal shelter," *Journal of the American Veterinary Medical Association*, 209 (1996), pp. 572-581.) Here are the data:

	Source		
Group	Private	Pet store	Other
Cases	188	7	90
Controls	518	68	142

Analyze the data and write a short report explaining your conclusions.

9.24 Compare the sources of dogs and cats. The studies described in the previous two exercises contain data on where people got their pets. The control group data (but not the "cases") were obtained by a random digit dialing telephone survey and can be considered an SRS of households with a cat or a dog in this geographic area. Compare the sources of cats and of dogs. Write a short report on your analysis; include appropriate descriptive statistics, the results of a significance test, and your conclusion.

9.25 More categories from the same data. The "Other" category for the source of the pet in the previous exercises includes born in home, stray, and obtained from a shelter. The following two-way table lists these categories separately for cats:

	Source				
Group	Private	Pet store	Home	Stray	Shelter
Cases	124	16	20	38	18
Controls	219	24	38	116	49

Here are the results for dogs:

	Source				
Group	Private	Pet store	Home	Stray	Shelter
Cases	188	7	11	23	56
Controls	518	68	20	55	67

Analyze these 2×5 tables and compare the significance of the chi-square test with your results for the 2×3 tables in Exercises 9.22 and 9.23. With a large number of cells, the chi-square test sometimes does not have very much power.

9.26 Sports goals. Knowing why different groups of customers participate in an activity or purchase a product can be very useful information in designing a marketing strategy. One study looked at why students participate in recreational sports and compared the profiles of men and women participants. (This study is reported in Joan L. Duda, "The relationship between goal perspectives, persistence and behavioral intensity among male and female recreational sport participants," *Leisure Sciences*, 10 (1988), pp. 95–106.) One goal of people who participate in sports is social comparison—the desire to win or to do better than other people. Another is mastery—the desire to improve one's skills or to try one's best. Data were collected from 67 male and 67 female undergraduates at a large university. Each student was classified into one of four categories based on his or her responses to a questionnaire about sports goals. The four categories were high social comparison–high mastery (HSC-HM), high social comparison–low mastery (HSC-LM), low social comparison–high mastery (LSC-HM), and low social comparison–low mastery (LSC-LM). Here are the data displayed in a two-way table:

Observed counts for sports goals

	Sex		
Goal	Female	Male	Total
HSC-HM	14	31	45
HSC-LM	7	18	25
LSC-HM	21	5	26
LSC-LM	25	13	38
Total	67	67	134

(a) Analyze this 4×2 table and summarize your conclusions.
(b) Construct a 2×2 table by summing over the mastery variable. (For example, there will be $14 + 7 = 21$ females in the HSC group.) Analyze this table and summarize your results.
(c) Perform a similar analysis by summing over the social comparison variable.
(d) Write a report summarizing your conclusions.

9.27 Credit cards and household income. A Gallup Poll used telephone interviews to ask 1025 adults aged 18 and over about credit card ownership. (From an article by David W. Moore, "Only one in five Americans without a credit card," describing the results of a poll conducted April 6–8, 2001. The article was found on the Gallup Web site, `www.gallup.com`.) A report based on the poll described the relationship between ownership of a credit card and income. The report states that 54% of people with incomes less than $20,000 have no credit cards. The percents for other income levels are 24% for $20,000 to $30,000, 18% for $30,000 to $50,000, and 7% for over $50,000.
(a) Make a graph that describes the relationship between credit card ownership and income.
(b) Describe the relationship in a short paragraph.
(c) If possible, perform the statistical significance test that addresses the question of whether or not there is a relationship. If it is not possible to do this, explain what additional information you would need to perform the test.

9.28 Web users. To design effective marketing strategies, you need to know your customers. What are the characteristics of people who use the World Wide Web to collect information on travel, and how do they differ from those who use other sources? A survey that collected data to address this question examined the responses of 1401 Web users (*www*) and 1080 people who used other sources for this information (Other). (From K. Weber and W. S. Roehl, "Profiling people searching for and purchasing travel products on the world wide web," *Journal of Travel Research*, 37 (1999), pp. 291–298. The Web site `www.gvu.gatech.edu/user_surveys` has more information about this and similar surveys.) The following tables give counts of *www* and Other for various demographic characteristics. Note that the marginal sums are sometimes less than 1401 and 1080 because of missing data. Use the methods of this chapter to compare the two groups. Include graphical and numerical summaries along with the results of your significance tests. In some cases you may want to combine some categories for the demographic variables. Be sure to include a discussion of missing values. Write a report summarizing your work.

Age in years		
	www	Other
Under 18	22	24
18–25	160	161
26–35	328	184
36–45	277	189
46–55	224	164
Over 55	101	109

Gender		
	www	Other
Female	709	561
Male	423	291

Education		
	www	Other
Grammar school	4	15
High school	85	125
Vocational training	54	53
Some college	336	293
College	357	218
Postgraduate	259	114
Professional	27	17
Other	10	17

Occupational category		
	www	Other
Management	167	87
Professional	264	156
Educator/student	175	164
Computer related	309	164
Other	217	281

Household income (U.S. $)		
	www	Other
Less than 10,000	43	65
10,000–19,999	58	78
20,000–29,999	116	102
30,000–39,999	149	127
40,000–49,999	127	105
50,000–74,999	259	129
75,000–99,999	119	71
Over 100,000	134	47
Rather not say	127	128

Race		
	www	Other
Caucasian/white	1001	772
African American	20	16
Asian/Pacific Islander	35	16
Hispanic/Latino	20	15
Other	56	33

9.29 Start-up businesses in the United States and Korea. What are the characteristics of successful start-up businesses in the United States, and how do they differ from similar businesses in Korea? A study designed to address these questions examined characteristics of 62 U.S. firms and 53 Korean counterparts. The tables below give counts for various characteristics. Analyze these data using the methods of this chapter, and write a report summarizing your findings. Include numerical as well as graphical summaries with the results of your significance tests.

Gender of owner/manager		
	U.S.	Korea
Female	19	14
Male	43	39

Age of owner/manager (in years)		
	U.S.	Korea
Under 30	2	2
31–40	12	7
41–50	17	33
Over 50	31	11

Major of owner/manager in college		
	U.S.	Korea
Business/economics	12	12
Engineering	12	29
Other	38	12

Education of owner/manager		
	U.S.	Korea
High school	3	3
Undergraduate degree	26	37
Master's degree	20	6
Doctoral degree	13	7

Previous area of work of owner/manager		
	U.S.	Korea
Technical	4	4
Administrative	14	12
Marketing	11	14
Research and development	6	6
Other	27	17

Previous job position of owner/manager		
	U.S.	Korea
Owner	4	3
CEO	8	11
Department manager	16	10
Department director	14	14
Employee	20	15

Years of experience of owner/manager in current business		
	U.S.	Korea
Less than 1	1	31
1–2	5	14
3–4	14	2
5 or more	42	6

Type of business		
	U.S.	Korea
General (opportunistic)	33	30
Technical (craftsman)	29	23

Ownership type		
	U.S.	Korea
Sole proprietorship	1	10
Partnership	4	2
Corporation	56	36
Other	1	5

Type of site		
	U.S.	Korea
General (free) location	30	43
Industrial complex	17	2
Other	15	8

CHAPTER 10

Section 10.1

10.1 Agricultural productivity. The productivity of a process or an industry is defined as output per unit of input. We can measure the productivity of land used to grow corn by the yield of corn in bushels per acre. Improvements in other inputs (seed, fertilizers, pesticides, and so on) have led to great increases in the productivity of land. Here are the average corn yields in the United States in the middle of four successive decades: (Data provided by Robert Dale, Purdue University.)

Year	1966	1976	1986	1996
Yield	73.1	88.0	119.4	127.1

(a) Make a scatterplot that shows the increase in yield over time. Does the plot suggest a linear relationship between yield and time?
(b) Find the equation of the least-squares regression line for predicting yield from year. (Use a calculator or software.) Add this line to your scatterplot.
(c) Find by hand the residuals of the four observations from the regression line. Use these residuals to calculate the standard error s.
(d) Write the regression model for this setting. What are your estimates of the unknown parameters in this model?
(*Comment:* These are *time series data*. Simple regression is often a good fit to time series data over a limited span of time. See Chapter 13 for methods designed specifically for use with time series.)

Chapter 10 Review Exercises

Age of houses and their selling prices. *The following table describes a random sample of 50 houses sold in Ames, Iowa, in 2000. We have already seen that selling price is related to size in square feet. Now we ask if selling price is related to age in years. Exercises 10.1 to 10.6 use this information. (Data provided by the Ames City Assessor, Ames, Iowa.)*

Houses sold in Ames, Iowa

Selling price ($)	Square footage	Age (years)	Selling price ($)	Square footage	Age (years)
268,380	1897	1	169,900	1686	35
131,000	1157	15	180,000	2054	34
112,000	1024	35	127,000	1386	50
112,000	935	35	242,500	2603	10
122,000	1236	39	152,900	1582	3
127,900	1248	32	171,600	1790	1
157,600	1620	33	195,000	1908	6
135,000	1124	33	83,100	1378	72
145,900	1248	35	125,000	1668	55
126,000	1139	39	60,500	1248	100
142,000	1329	40	85,000	1229	59
107,500	1040	45	117,000	1308	60
110,000	951	42	57,000	892	90
187,000	1628	1	110,000	1981	72
94,000	816	43	127,250	1098	70
99,500	1060	24	119,000	1858	80
78,000	800	68	172,500	2010	60
55,790	492	79	123,000	1680	86
70,000	792	80	161,715	1670	1
53,600	980	62	179,797	1938	1
157,000	1629	3	117,250	1120	36
166,730	1889	0	116,500	914	4
340,000	2759	6	117,000	1008	23
195,000	1811	3	177,500	1920	32
215,850	2400	27	132,000	1146	37

10.2 Price versus age. Describe the relationship between price and age by a description based on the plot, the least-squares line, and r^2. Why would you expect the slope of the regression line to be negative? State in simple language what the numerical value of the slope says about how the selling price of houses changes with their age. What is the interpretation of the intercept of the regression line in this setting?

10.3 How strong is the relationship?
(a) Give a 95% confidence interval for the slope of the regression line of price on age in the population of all houses sold in Ames in the year 2000.
(b) Is there strong evidence that this slope is negative? (State the hypotheses, give the test statistic and its P-value, and state your conclusion.)

10.4 Predicting price from age.
(a) A developer built a number of houses in Ames in 1990. What should these houses sell for in 2000, on the average? Give a 95% confidence interval for the mean selling price in 2000 of houses built in 1990.
(b) Chong and Mei-Ling bought a newly built house in Ames in 1990. Give a 95%

interval that predicts the selling price of their house in 2000.
(c) Explain to someone who knows no statistics why it is more difficult to predict the price of one house when we know only its age than to predict the average price of all houses that old.

10.5 Price versus age. Plot the least-squares regression line for ages between 0 and 50 years. Above each of the ages 10, 20, 30, and 40 years plot the predicted price (this point lies on the line) and the upper and lower end of the 95% confidence interval for the mean price of houses this old. Then draw a vertical line between each pair of endpoints to show the four confidence intervals. Your graph gives a good picture of how mean price declines with age and of the considerable uncertainty about the mean at each age.

10.6 Predicting price at 30 years. Find a 90% confidence interval for the mean price of houses in Ames that were 30 years old in 2000.

10.7 House prices: size and age. We have now looked at predicting the selling prices of houses in a community first from their sizes and then from their ages. It may occur to you that *newer houses tend to be larger*. In that case, using age to predict price works in part because age is a stand-in for size. Yet we also expect a newer house to generally sell for more than an older house of the same size. To untangle these relationships, we need *multiple regression* using both age and size together to predict price. Multiple regression is the topic of Chapter 11. We can, however, use software to make a start now.
(a) Describe the relationship between age and size for the houses. Use both a graph and numerical descriptions and summarize your findings in words. Do newer houses tend to be larger?
(b) Regress selling price on size and save the residuals. The residuals represent the variation in price that is not explained by size. Plot these residuals against age and find the correlation between these variables. A clear linear relationship tells us that age can help explain selling price even after the effect of size has been removed. Does it appear that age is a useful predictor of price if we already know the size of a house?

10.8 The effect of bank size. The following table gives the wages and months of service for a sample of 59 married women who hold customer service jobs in Indiana banks. The table also notes whether each woman worked at a large bank (100 or more workers) or at a smaller bank. Make a scatterplot of these variables using different symbols for the 34 women who work in large banks and the 25 women who work in small banks. Which group appears to have generally higher wages? For which group does it appear that regressing wages on length of service will do a better job of explaining wages?

Bank wages and length of service

Wages	LOS	Size	Wages	LOS	Size	Wages	LOS	Size
48.3355	94	Large	64.1026	24	Large	41.2088	97	Small
49.0279	48	Small	54.9451	222	Small	67.9096	228	Small
40.8817	102	Small	43.8095	58	Large	43.0942	27	Large
36.5854	20	Small	43.3455	41	Small	40.7000	48	Small
46.7596	60	Large	61.9893	153	Large	40.5748	7	Large
59.5238	78	Small	40.0183	16	Small	39.6825	74	Small
39.1304	45	Large	50.7143	43	Small	50.1742	204	Large
39.2465	39	Large	48.8400	96	Large	54.9451	24	Large
40.2037	20	Large	34.3407	98	Large	32.3822	13	Small
38.1563	65	Small	80.5861	150	Large	51.7130	30	Large
50.0905	76	Large	33.7163	124	Small	55.8379	95	Large
46.9043	48	Small	60.3792	60	Large	54.9451	104	Large
43.1894	61	Small	48.8400	7	Large	70.2786	34	Large
60.5637	30	Large	38.5579	22	Small	57.2344	184	Small
97.6801	70	Large	39.2760	57	Large	54.1126	156	Small
48.5795	108	Large	47.6564	78	Large	39.8687	25	Large
67.1551	61	Large	44.6864	36	Large	27.4725	43	Small
38.7847	10	Small	45.7875	83	Small	67.9584	36	Large
51.8926	68	Large	65.6288	66	Large	44.9317	60	Small
51.8326	54	Large	33.5775	47	Small	51.5612	102	Large

10.9 Verifying the effect of bank size. Refer to the previous exercise. Do a two-sample t test comparing the mean wages at small and large banks. Wages are significantly higher at large banks. Do another two-sample t test comparing the mean length of service at small and large banks. These means do not differ significantly.

10.10 Separate regressions. The plot and tests of the previous two exercises seem to confirm that bank size does influence the relationship between wages and length of service. Do separate regressions of wages on length of service for large banks ($n = 34$) and for small banks ($n = 25$). Report the tests for significance of regression in both cases. Write a careful summary of your findings. In particular, for which of the two groups of banks might we use length of service to predict wages?

The Leaning Tower of Pisa. *The Leaning Tower of Pisa was reopened to the public late in 2001 after being closed for almost 12 years while engineers took steps to prevent the tower from collapsing. Data on the lean of the tower over time show why it was in danger of collapse. The following table gives measurements for the years 1975 to 1987. The variable "lean" represents the difference between where a point near the top of the tower would be if the tower were straight and where it actually is. The data are coded as tenths of a millimeter in excess of 2.9 meters, so that the 1975 lean, which was 2.9642 meters, appears in the table as 642. Only the last two digits of the year were entered into the computer* (G. Geri and B. Palla, "Considerazioni sulle più recenti osservazioni ottiche alla Torre Pendente di Pisa," *Estratto dal Bollettino della Società Italiana di Topografia e Fotogrammetria*,

2 (1988), pp. 121–135. Professor Julia Mortera of the University of Rome provided valuable assistance with the translation).

Year	Lean	Year	Lean	Year	Lean
75	642	80	688	84	717
76	644	81	696	85	725
77	656	82	698	86	742
78	667	83	713	87	757
79	673				

The next three exercises ask you to analyze these data.

10.11 The lean of the Leaning Tower.
(a) Plot the lean of the tower against time. Is the trend linear? That is, is the tower's lean increasing at a fixed rate?
(b) What is the equation of the least-squares line for predicting lean? What percent of the variation in lean is explained by this line?
(c) Give a 95% confidence interval for the average rate of change (tenths of a millimeter per year) of the lean.

10.12 Looking into the past.
(a) In 1918 the lean was 2.9071 meters. (The coded value is 71.) Using the least-squares equation for the years 1975 to 1987, calculate a predicted value for the lean in 1918. (Note that you must use the coded value 18 for year.)
(b) Although the least-squares line gives an excellent fit to the data for 1975 to 1987, this pattern does not extrapolate to 1918. Write a short statement explaining why this conclusion follows from the information available. Use numerical and graphical summaries to support your explanation.

10.13 Looking to the future.
(a) The engineers working on the tower were most interested in how much the tower would lean if no corrective action was taken. Use the least-squares equation to predict the tower's lean in the year 2000 if no corrective action had been taken.
(b) To give a margin of error for the lean in 2000, would you use a confidence interval for a mean response or a prediction interval? Explain your choice.

CHAPTER 11

Section 11.2

11.1 Faculty salaries. Data on the salaries of full professors in an engineering department at a large midwestern university are given below. The salaries are for the academic years 2002–2003 and 2005–2006. The data also include years in rank as a full professor.

Years in rank	2002 salary ($)	2005 salary ($)
3	95,600	113,400
4	93,800	114,500
5	104,200	113,300
6	119,000	134,000
6	100,600	115,000
7	108,200	128,000
7	92,100	100,800
8	106,700	115,700
10	102,300	144,200
10	112,500	133,100
11	113,100	131,100
13	122,500	134,400
15	121,100	130,000
15	124,000	133,700
35	153,300	172,800
36	138,500	144,500

(a) Write the model that you would use for a multiple regression to predict salary in 2005 from salary in 2002 and years in rank.
(b) What are the parameters of your model?
(c) Run the multiple regression and give the estimates of the model parameters.
(d) Test the hypothesis that the coefficients for salary in 2002 and years in rank are both zero. Give the test statistic with degrees of freedom and the P-value. What do you conclude?
(e) What is the value of R^2?
(f) Give the results of the hypothesis test for the coefficient for salary in 2002. Include the test statistic, degrees of freedom, and the P-value. Do the same for years in rank. Summarize your conclusions from these two tests.

11.2 Examine the assumptions. We now ask whether the data of the previous exercise meet the requirements of the multiple regression model. Find the residuals.
(a) Examine the distribution of the residuals. Summarize what you have found.
(b) Plot the residuals versus each of the explanatory variables. Describe the plots. Does your analysis suggest that the model assumptions may not be reasonable for this problem? Why?

11.3 Compare regression coefficients. Using the faculty salary data in Exercise 11.1, do the regression to predict salary in 2005 using only years in rank.
(a) Give the fitted regression equation.
(b) Summarize the results of the significance test for the coefficient for years in rank.
(c) In Exercise 11.1 you found a coefficient for years in rank and reported the results of a significance test for this coefficient. Give those results here and explain why they differ from what you found in parts (a) and (b) of this exercise.

Section 11.3

11.4 Professors' salaries. Refer to the data on salaries of a collection of engineering full professors given in Exercise 11.1. A plot of salary for 2005 versus years in rank suggests that the relationship may be slightly curved. Examine this question by running a regression to predict the 2005 salary using years in rank and the square of years in rank. Report the relevant test statistic with its degrees of freedom and P-value, and summarize your conclusion.

11.5 Compare the models. Refer to the previous exercise. We can view this analysis in the framework of testing a hypothesis about a collection of regression coefficients that we studied in Section 11.2. The first model includes years in rank and the square of years in rank, while the second includes only years in rank.
(a) Run both regressions and find the value of R^2 for each.
(b) Find the F statistic for comparing the models based on the difference in the values of R^2. Carry out the test and report your conclusion.
(c) Verify that the square of the t statistic that you found in the previous exercise for testing the coefficient of the quadratic term is equal to the F statistic that you found for this exercise.

11.6 Predict the quality of a product. Refer to the CHEESE data set in the Data Appendix. Here the measure of quality is a variable called Taste and the explanatory variables are the concentrations of three chemicals found in cheese. These are acetic acid, hydrogen sulfide, and lactic acid. With three explanatory variables, there are three models that have a single explanatory variable, three that have two explanatory variables, and one with all three included. Run these 7 regressions and make a table giving the regression coefficients and the value of R^2 for each regression. (If an explanatory variable is not included in a particular regression, enter a value of 0 for its coefficient in the table.) Mark coefficients that are statistically significant at the 5% level with an asterisk (*). Summarize your results and state which model you prefer.

Chapter 11 Review Exercises

11.7 Vitamin-enriched bread. Does bread lose its vitamins when stored? Small loaves of bread were prepared with flour that was fortified with a fixed amount of vitamins. After baking, the vitamin C content of two loaves was measured. Another two loaves were baked at the same time, stored for one day, and then the vitamin C content was measured. In a similar manner, two loaves were stored for three,

five, and seven days before measurements were taken. The units are milligrams per hundred grams of flour (mg/100 g) (data provided by Helen Park; see H. Park et al., "Fortifying bread with each of three antioxidants," *Cereal Chemistry,* 74 (1997), pp. 202–206). Here are the data:

Condition	Vitamin C (mg/100 g)	
Immediately after baking	47.62	49.79
One day after baking	40.45	43.46
Three days after baking	21.25	22.34
Five days after baking	13.18	11.65
Seven days after baking	8.51	8.13

We will use some regression models to examine the relationship between the vitamin C concentration (VitC) and days after baking (Days) using these data.
(a) Run a linear regression to predict VitC using Days. Report the regression equation and the test statistic for the coefficient of Days with its degrees of freedom and P-value, and summarize your conclusion.
(b) Plot VitC versus Days. Also plot the residuals from your model in part (a) versus Days. Interpret these plots.
(c) Run a multiple regression using Days and the square of Days to predict VitC. Summarize the results of all of the relevant significance tests associated with this analysis.
(d) Write a short summary comparing the results of parts (a) and (c) of this exercise. Be sure to mention the values of R^2 for both regressions.

11.8 What about vitamins A and E? Refer to the previous exercise. Measurements of the amounts of vitamin A (beta-carotene) and vitamin E in each loaf are given below. Using the parts of the previous exercise as a guide, analyze the data for each of these vitamins. Write a report summarizing what happens to the three vitamins in bread when it is stored. Be sure to note similarities and differences in the three sets of results.

Condition	Vitamin A (mg/100 g)		Vitamin E (mg/100 g)	
Immediately after baking	3.36	3.34	94.6	96.0
One day after baking	3.28	3.20	95.7	93.2
Three days after baking	3.26	3.16	97.4	94.3
Five days after baking	3.25	3.36	95.0	97.7
Seven days after baking	3.01	2.92	92.3	95.1

Exercises 11.9 to 11.11 use the bank wages data given in Exercise 11.1. For these exercises we code the size of the bank as 1 if it is large and 0 if it is small. There is one outlier in this data set. Delete it and use the remaining 59 observations in these exercises.

11.9 Length of service and wages. Use length of service (LOS) to predict wages with a simple linear regression. Write a short summary of your results and conclusions.

11.10 Size of the bank and wages. Predict wages using the size of the bank as the explanatory variable. Use the coded values 0 and 1 for this model.

(a) Summarize the results of your analysis. Include a statement of all hypotheses, test statistics with degrees of freedom, P-values, and conclusions.
(b) Calculate the t statistic for comparing the mean wages for the large and the small banks assuming equal standard deviations. Give the degrees of freedom. Verify that this t is the same as the t statistic for the coefficient of size in the regression. Explain why this makes sense.
(c) Plot the residuals versus LOS. What do you conclude?

11.11 Use both variables to predict wages. Use a multiple linear regression to predict wages from LOS and the size of the bank. Include an interaction term in your model. Write a report summarizing your work. Include graphs and the results of significance tests.

CHAPTER 12

Section 12.2

12.1 Special causes. Is each of the following examples of a special cause most likely to first result in (i) one point out on the s or R chart; (ii) one point out on the $\overline{x}$ chart; or (iii) a run on the $\overline{x}$ chart? In each case, briefly explain your reasoning.
(a) An etching solution deteriorates as more items are etched.
(b) Buildup of dirt reduces the precision with which parts are placed for machining.
(c) A new customer service representative for a Spanish-language help line is not a native speaker and has difficulty understanding customers.
(d) A data entry employee grows less attentive as her shift continues.

12.2 Mixtures. Here is an artificial situation that illustrates an unusual control chart pattern. Invoices are processed and paid by two clerks, one very experienced and the other newly hired. The experienced clerk processes invoices quickly. The new hire must often refer to a handbook and is much slower. Both are quite consistent, so that their times vary little from invoice to invoice. It happens that each sample of invoices comes from one of the clerks, some samples from one and some from the other. Sketch the $\overline{x}$ chart pattern that will result.

12.3 Special causes. Is each of the following examples of a special cause most likely to first result in (i) a sudden change in level on the s or R chart; (ii) a sudden change in level on the $\overline{x}$ chart; or (iii) a gradual drift up or down on the $\overline{x}$ chart? In each case, briefly explain your reasoning.
(a) An airline pilots' union puts pressure on management during labor negotiations by asking its members to "work to rule" in doing the detailed checks required before a plane can leave the gate.
(b) Measurements of part dimensions that were formerly made by hand are now made by a very accurate laser system. (The process producing the parts does not change—measurement methods can also affect control charts.)
(c) Inadequate air conditioning on a hot day allows the temperature to rise during the afternoon in an office that prepares a company's invoices.

12.4 The Boston Marathon. The Boston Marathon has been run each year since 1897. Winning times were highly variable in the early years, but control improved as the best runners became more professional. A clear downward trend continued until the 1980s. Rick plans to make a control chart for the winning times from 1950 to the present. The first few times are 153 148 152 139 141 138 . . . .Calculation from the winning times from 1950 to 2002 gives

$$\overline{x} = 134.906 \text{ minutes} \qquad \text{and} \qquad s = 6.523 \text{ minutes}$$

Rick draws a center line at $\overline{x}$ and control limits at $\overline{x} \pm 3s$ for a plot of individual winning times. Explain carefully why these control limits are too wide to effectively signal unusually fast or slow times.

12.5 Is Joe's weight stable? Joe has recorded his weight, measured at the gym after a workout, for several years. The mean is 162 pounds and the standard

deviation 1.5 pounds, with no signs of lack of control. An injury keeps Joe away from the gym for several months. The data below give his weight, measured once each week for the first 16 weeks after he returns from the injury:

Week	1	2	3	4	5	6	7	8
Weight	168.7	167.6	165.8	167.5	165.3	163.4	163.0	165.5
Week	9	10	11	12	13	14	15	16
Weight	162.6	160.8	162.3	162.7	160.9	161.3	162.1	161.0

Joe wants to plot these individual measurements on a control chart. When each "sample" is just one measurement, short-term variation is estimated by advanced techniques. The short-term variation in Joe's weight is estimated to be about $\sigma = 1.3$ pounds. Joe has a target of $\mu = 162$ pounds for his weight. Make a control chart for his measurements, using control limits $\mu \pm 2\sigma$. It is common to use these narrower limits on an "individuals chart." Comment on individual points out of control and on runs. Is Joe's weight stable or does it change systematically over this period?

12.6 Meaning of capability indexes. Explain why C_p is often referred to as a *potential capability* index, whereas C_{pk} is referred to as an *actual capability* index.

Section 12.3

12.7 Lost baggage. The Department of Transportation reports that about 1 of every 200 passengers on domestic flights of the 10 largest U.S. airlines files a report of mishandled baggage. Starting with this information, you plan to sample records for 1000 passengers per day at a large airport to monitor the effects of efforts to reduce mishandled baggage. What are the initial center line and control limits for a chart of the daily proportion of mishandled baggage reports? (You will find that LCL < 0. Because proportions $\hat{p}$ are always 0 or positive, take LCL $= 0$.)

12.8 Doctors' prescriptions. A regional chain of retail pharmacies finds that about 1% of prescriptions it receives from doctors are incorrect or illegible. The chain puts in place a secure online system that doctors' offices can use to enter prescriptions directly. It hopes that fewer prescriptions entered online will be incorrect or illegible. A p chart will monitor progress. Use information about past prescriptions to set initial center line and control limits for the proportion of incorrect or illegible prescriptions on a day when the chain fills 75,000 online prescriptions. What are the center line and control limits for a day when only 50,000 online prescriptions are filled?

Chapter 12 Review Exercises

12.9 What type of chart? What type of control chart or charts would you use as part of efforts to improve each of following performance measures in a corporate personnel office? Explain your choices.
(a) Time to get security clearance.
(b) Percent of job offers accepted.
(c) Employee participation in voluntary health screening.

12.10 What type of chart? What type of control chart or charts would you use as part of efforts to improve each of following performance measures in a corporate information systems department? Explain your choices.
(a) Computer system availability.
(b) Time to respond to requests for help.
(c) Percent of programming changes not properly documented.

CHAPTER 13

Section 13.2

13.1 JCPenney sales. The table below contains retail sales for JCPenney in millions of dollars. The data are quarterly, beginning with the first quarter of 1996 and ending with the fourth quarter of 2001.

Year-quarter	Sales	Year-quarter	Sales
1996-1st	4452	1999-1st	7339
1996-2nd	4507	1999-2nd	7104
1996-3rd	5537	1999-3rd	7639
1996-4th	8157	1999-4th	9661
1997-1st	6481	2000-1st	7528
1997-2nd	6420	2000-2nd	7207
1997-3rd	7208	2000-3rd	7538
1997-4th	9509	2000-4th	9573
1998-1st	6755	2001-1st	7522
1998-2nd	6483	2001-2nd	7211
1998-3rd	7129	2001-3rd	7729
1998-4th	9072	2001-4th	9542

(a) Before plotting these data, inspect the values in the table. Do you see any interesting features of JCPenney quarterly sales?
(b) Now, make a time plot of the data. Be sure to connect the points in your plot to highlight patterns.
(c) Is there an obvious trend in JCPenney quarterly sales? If so, is the trend positive or negative?
(d) Is there an obvious repeating pattern in the data? If so, clearly describe the repeating pattern.

13.2 JCPenney sales. In Exercise 13.1, you are provided with a table of JCPenney sales data. Use statistical software to further investigate the JCPenney sales data.
(a) Find the least-squares line for the sales data. Use 1, 2, 3, ... as the values for the explanatory variable, with $x = 1$ corresponding to the first quarter of 1996, $x = 2$ corresponding to the second quarter of 1996, etc.
(b) The intercept is a prediction of sales for what quarter?
(c) Interpret the slope in the context of JCPenney quarterly sales.

13.3 JCPenney sales. In Exercise 13.1, you made a time plot of the sales data. Sales seem to follow a pattern of ups and downs that repeats every four quarters. Add indicator variables for first, second, and third quarters to the linear trend model fitted in previous exercise. Call these indicator variables $X1$, $X2$, and $X3$, respectively.
(a) Write down the estimated trend-and-season model.
(b) Explain why no indicator variable is needed for fourth quarters.
(c) Does the intercept still predict sales for a specific quarter? If so, what quarter?

Compare the estimated intercept of this model with that of the trend-only model (see Exercise 13.2(b)). Given the pattern of seasonal variation, which appears to be the better estimate?

13.4 JCPenney sales. In Exercise 13.2, you fitted a linear trend-only model to the JCPenney sales data. Starting with this trend model, we want to incorporate seasonality factors to account for the pattern that repeats every four quarters.
(a) Calculate the seasonality factor for each quarter.
(b) Average the four seasonality factors. Is this average close to 1? If so, interpret the seasonality factor for fourth quarters.
(c) Make a scatterplot of seasonality factor versus quarter with the seasonality factors on the vertical axis and the quarters on the horizontal axis. Connect the points to see the general pattern of seasonal variation. Also, draw a horizontal line at the average of the four seasonality factors.

13.5 JCPenney sales. A linear trend-only model was fitted to the JCPenney sales data in Exercise 13.2. Use this model to answer the following:
(a) On a time plot of the sales data, draw the least-squares line. Comment on any pattern of over- or underprediction if we were to use this trend-only model for predicting sales.
(b) Using the equation of the least-squares line, forecast sales for the first quarter of 2002 and for the fourth quarter of 2002.
(c) Which forecast in part (b) do you believe will be more accurate when compared to actual JCPenney sales? Why?

13.6 JCPenney sales. A trend-and-season model with indicator variables was fitted to the JCPenney sales data in Exercise 13.3.
(a) Using the equation of the trend-and-season model, forecast sales for the first quarter of 2002 and for the fourth quarter of 2002.
(b) Compare your forecasts with the same forecasts based on the trend-only model of the previous exercise.

13.7 JCPenney sales. A trend-and-season model using seasonality factors was calculated in Exercise 13.4.
(a) Using the linear trend-only model and the seasonality factors, forecast sales for the first quarter of 2002 and for the fourth quarter of 2002.
(b) Compare your forecasts with the same forecasts based on the trend-only model of Exercise 13.5.
(c) Compare your forecasts with the same forecasts based on the trend-and-season model of Exercise 13.6.

13.8 JCPenney sales. The model in part (a) of Exercise 13.3 can be viewed as four models—one for each quarter—by setting the indicator variables to 0 or 1 for each of the four quarters (our seasons for this data set).
(a) Write down the four models (one for each quarter) that are derived from the trend-and-season model in Exercise 13.3.
(b) Each of the models in part (a) is linear. What do you notice about the four slopes?

(c) On a time plot of the sales data, sketch each of the four lines. What geometric property do the lines possess?

13.9 Comparing models for JCPenney sales. Compare the trend-only model of Exercise 13.2 with the trend-and-season model of Exercise 13.3.
(a) Report the value of R^2 for each model and comment on the difference.
(b) Report the value of s for each model and comment on the difference.
(c) Make a time plot of the original JCPenney sales figures. On this plot, overlay both the trend-only predictions and the trend-and-season predictions.
(d) Taking into account parts (a), (b), and (c), is the trend-and-season model a big improvement over the trend-only model?

13.10 Seasonally adjusted JCPenney sales. In Exercise 13.4, you calculated seasonality factors for the JCPenney quarterly sales data. Using these factors, complete the following:
(a) Calculate the seasonally adjusted value of the time series for the fourth quarter of 2001. Do the calculation by hand.
(b) Using statistical software, calculate the seasonally adjusted JCPenney sales time series. Make a time plot of the original sales data with the seasonally adjusted sales data superimposed.
(c) Did seasonally adjusting the JCPenney sales data smooth the time series greatly? What does this imply about the strength of the seasonal pattern in these two time series?

13.11 Autocorrelation in the JCPenney time series. In Exercise 13.2, a linear trend-only model was fitted to the JCPenney sales data. Using the residuals from this model, look for evidence of autocorrelation.
(a) Make a time plot of the residuals. Describe any pattern you see in this plot.
(b) Make a lagged residual plot and calculate the correlation between successive residuals e_{t-1} and e_t. Do we have evidence of autocorrelation?

13.12 Autocorrelation in the JCPenney time series. Repeat the previous exercise using the residuals from the trend-and-season model of Exercise 13.3.

Section 13.4

13.13 Philip Morris returns.
Following are the monthly percent returns on Philip Morris stock from June 1990 to July 2001:

Monthly percent returns on Philip Morris stock

from June 1990 to July 2001

3.0	−5.7	1.2	4.1	3.2	7.3	7.5	18.7	3.7	−1.8
2.4	−6.5	6.7	9.4	−2.0	−2.8	−3.4	19.2	−4.8	0.5
−0.6	2.8	−0.5	−4.5	8.7	2.7	4.1	−10.3	4.8	−2.3
−3.1	−10.2	−3.7	−26.6	7.2	−2.4	−2.8	3.4	−4.6	17.2
4.2	0.5	8.3	−7.1	−8.4	7.7	−9.6	6.0	6.8	10.9
1.6	0.2	−2.4	−2.4	3.9	1.7	9.0	3.6	7.6	3.2
−3.7	4.2	13.2	0.9	4.2	4.0	2.8	6.7	−10.4	2.7
10.3	5.7	0.6	−14.2	1.3	2.9	11.8	10.6	5.2	13.8
−14.7	3.5	11.7	1.5	2.0	−3.2	−3.9	−4.7	9.8	4.9
−8.3	4.8	−3.2	−10.9	0.7	6.4	11.3	−5.1	12.3	10.5
9.4	−3.6	−12.4	−16.5	−8.9	−0.4	10.0	5.4	−7.3	0.5
−7.4	−22.9	−0.5	−10.6	−9.2	−3.3	5.2	5.4	19.4	3.5
−4.9	17.8	0.7	24.4	4.3	16.6	0.0	9.5	−0.4	5.6
2.6	−2.7	−8.1	4.2						

(a) Using the Philip Morris returns, calculate the first 4 forecasts $\widehat{y}_1, \ldots, \widehat{y}_4$ using an exponential smoothing model with $w = 0.1$. Do not use statistical software for these calculations.
(b) Now, use software to fit a simple exponential smoothing model with $w = 0.1$. Use the forecasts provided by the software to verify your hand calculations in part (a). Are your forecasts the same as those provided by your software?
(c) Provide a forecast for the August 2001 Philip Morris return based on your exponential smoothing model with $w = 0.1$.
(d) Write down the forecast equation for the September 2001 Philip Morris return based on the exponential smoothing model with $w = 0.1$.

CHAPTER 14

Chapter 14 Review Exercises

14.1 Pooling variances. An experiment was run to compare four groups. The sample sizes were 30, 22, 180, and 25, and the corresponding estimated standard deviations were 20, 21, 10, and 19.
(a) Is it reasonable to use the assumption of equal standard deviations when we analyze these data? Give a reason for your answer.
(b) Give the values of the variances for the four groups.
(c) Find the pooled variance.
(d) What is the value of the pooled standard deviation?
(e) Explain why your answer in part (c) is much closer to the standard deviation for the third group than to any of the other standard deviations.

14.2 Do poets die young? According to William Butler Yeats, "She is the Gaelic muse, for she gives inspiration to those she persecutes. The Gaelic poets die young, for she is restless, and will not let them remain long on earth." One study designed to investigate this issue examined the age at death for writers from different cultures and genders. (The data were provided by James Kaufman. The study is described in James C. Kaufman, "The cost of the muse: poets die young," *Death Studies,* 27 (2003), pp. 813–821. The quote from Yeats appears in this article.) Three categories of writers examined were novelists, poets, and nonfiction writers. The ages at death for female writers in these categories from North America are given in the following table. Most of the writers are from the United States, but Canadian and Mexican writers are also included.

Age at death for women writers

Type	Age at death									
Novels ($n = 67$)	57	90	67	56	90	72	56	90	80	74
	73	86	53	72	86	82	74	60	79	80
	79	77	64	72	88	75	79	74	85	71
	78	57	54	50	59	72	60	77	50	49
	73	39	73	61	90	77	57	72	82	54
	62	74	65	83	86	73	79	63	72	85
	91	77	66	75	90	35	86			
Poems ($n = 32$)	88	69	78	68	72	60	50	47	74	36
	87	55	68	75	78	85	69	38	58	51
	72	58	84	30	79	90	66	45	70	48
	31	43								
Nonfiction ($n = 24$)	74	86	87	68	76	73	63	78	83	86
	40	75	90	47	91	94	61	83	75	89
	77	86	66	97						

(a) Use graphical and numerical methods to describe the data.
(b) Examine the assumptions necessary for ANOVA. Summarize your findings.
(c) Run the ANOVA and report the results.

(d) Use a contrast to compare the poets with the two other types of writers. Do you think that the quote from Yeats justifies the use of a one-sided alternative for examining this contrast? Explain your answer.
(e) Use another contrast to compare the novelists with the nonfiction writers. Explain your choice for an alternative hypothesis for this contrast.
(f) Use a multiple-comparisons procedure to compare the three means. How do the conclusions from this approach compare with those using the contrasts?

14.3 Does a product lose value when stored? Does bread lose its vitamins when stored? Small loaves of bread were prepared with flour that was fortified with a fixed amount of vitamins. After baking, the vitamin C content of two loaves was measured. Another two loaves were baked at the same time, stored for one day, and then the vitamin C content was measured. In a similar manner, two loaves were stored for three, five, and seven days before measurements were taken. The units are milligrams per hundred grams of flour (mg/100 g) (these data were provided by Helen Park; see H. Park et al., "Fortifying bread with each of three antioxidants," *Cereal Chemistry*, 74 (1997), pp. 202–206). Here are the data:

Condition	Vitamin C (mg/100 g)	
Immediately after baking	47.62	49.79
One day after baking	40.45	43.46
Three days after baking	21.25	22.34
Five days after baking	13.18	11.65
Seven days after baking	8.51	8.13

(a) Give a table of the sample sizes, means, and standard deviations for the five conditions.
(b) Perform a one-way ANOVA for these data. State hypotheses and give the test statistic, its degrees of freedom, and the P-value.
(c) Summarize the data and the means with a plot. Use the plot and the ANOVA results to write a short summary of your conclusions.

14.4 Compare the means. Refer to the previous exercise. Use the Bonferroni or another multiple-comparisons procedure to compare the group means. Summarize the results.

14.5 Two other vitamins in the product. Refer to Exercise 14.3. Measurements of the amounts of vitamin A (beta-carotene) and vitamin E in each loaf are given below. Use an analysis of variance to study the data for each of these vitamins.

Condition	Vitamin A (mg/100 g)		Vitamin E (mg/100 g)	
Immediately after baking	3.36	3.34	94.6	96.0
One day after baking	3.28	3.20	95.7	93.2
Three days after baking	3.26	3.16	97.4	94.3
Five days after baking	3.25	3.36	95.0	97.7
Seven days after baking	3.01	2.92	92.3	95.1

14.6 Multiple comparisons. Refer to the previous exercise.
(a) Explain why it is inappropriate to perform a multiple-comparisons analysis for

the vitamin E data.
(b) Perform the Bonferroni or another multiple-comparisons procedure for the vitamin A data and summarize the results.

14.7 Write a report. In Exercises 14.3 to 14.6 you have studied vitamin loss in bread stored after baking. Write a report summarizing the overall findings. Include appropriate statistical inference results and graphs.

14.8 Piano lessons and spatial-temporal reasoning. Do piano lessons improve the spatial-temporal reasoning of preschool children? In Exercise 7.20 we examined this question by comparing the change scores (after treatment minus before treatment) of 34 children who took piano lessons with the scores of 44 children who did not. The latter group actually contained three groups of children: 10 were given singing lessons, 20 had some computer instruction, and 14 received no extra lessons. The data appear in the table below.

Piano lesson data

Lessons	Scores									
Piano	2	5	7	–2	2	7	4	1	0	7
	3	4	3	4	9	4	5	2	9	6
	0	3	6	–1	3	4	6	7	–2	7
	–3	3	4	4						
Singing	1	–1	0	1	–4	0	0	1	0	–1
Computer	0	1	1	–3	–2	4	–1	2	4	2
	2	2	–3	–3	0	2	0	–1	3	–1
None	5	–1	7	0	4	0	2	1	–6	0
	2	–1	0	–2						

(a) Make a table giving the sample size, mean, and standard deviation for each group.
(b) Analyze the data using one-way analysis of variance. State the null and alternative hypotheses, the test statistic with degrees of freedom, the P-value, and your conclusion.

14.9 Compare the means. Refer to the previous exercise. Use the Bonferroni or another multiple-comparisons procedure to compare the group means. Summarize the results and support your conclusions with a graph of the means.

14.10 A contrast. The researchers in Exercise 14.8 based their research on a biological argument for a causal link between music and spatial-temporal reasoning. It is therefore natural to test the contrast that compares the mean of the piano lesson group with the average of the three other means. Construct this contrast, perform the significance test, and summarize the results. Note that this is not the test that we performed in Exercise 7.20 (page xxx), where we did not differentiate among the three groups who did not receive piano instruction.

14.11 Cooking pots and dietary iron. Iron-deficiency anemia is the most common form of malnutrition in developing countries, affecting about 50% of children

and women and 25% of men. Iron pots for cooking foods had traditionally been used in many of these countries, but they have been largely replaced by aluminum pots, which are cheaper and lighter. Some research has suggested that food cooked in iron pots will contain more iron than food cooked in other types of pots. One study designed to investigate this issue compared the iron content of some Ethiopian foods cooked in aluminum, clay, and iron pots (based on A. A. Adish et al., "Effect of consumption of food cooked in iron pots on iron status and growth of young children: a randomised trial," *The Lancet*, 353 (1999), pp. 712–716). One of the foods was *yesiga wet'*, beef cut into small pieces and prepared with several Ethiopian spices. The iron content of four samples of *yesiga wet'* cooked in each of the three types of pots is given below. The units are milligrams of iron per 100 grams of cooked food.

Type of pot	Iron (mg/100 g food)			
Aluminum	1.77	2.36	1.96	2.14
Clay	2.27	1.28	2.48	2.68
Iron	5.27	5.17	4.06	4.22

(a) Make a table giving the sample size, mean, and standard deviation for each type of pot. Is it reasonable to pool the variances? Note that with the small sample sizes in this experiment, we expect a large amount of variability in the sample standard deviations. For this reason, we will proceed with the ANOVA.
(b) Carry out the analysis of variance. Report the F statistic with its degrees of freedom and P-value. What do you conclude?

14.12 Residuals and multiple comparisons. Refer to the previous exercise.
(a) Examine the residuals. Is the Normality assumption reasonable for these data?
(b) Use the Bonferroni or another multiple-comparisons procedure to determine which pairs of means differ significantly. Summarize your results in a short report. Be sure to include a graph.

14.13 What colors attract beetles? The presence of harmful insects in farm fields is detected by erecting boards covered with a sticky material and then examining the insects trapped on the board. To investigate which colors are most attractive to cereal leaf beetles, researchers placed six boards of each of four colors in a field of oats in July (modified from M. C. Wilson and R. E. Shade, "Relative attractiveness of various luminescent colors to the cereal leaf beetle and the meadow spittlebug," *Journal of Economic Entomology*, 60 (1967), pp. 578–580). The table below gives data on the number of cereal leaf beetles trapped:

Color	Insects trapped					
Lemon yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

(a) Make a table of means and standard deviations for the four colors, and plot the means.
(b) State H_0 and H_a for an ANOVA on these data, and explain in words what ANOVA tests in this setting.

(c) Using computer software, run the ANOVA. What is the value of s_p? What are the F statistic and its P-value? What do you conclude?

14.14 Multiple comparisons. Return to the previous exercise. For the Bonferroni procedure with $\alpha = 0.05$, the value of t^{**} is 2.61. Use this multiple-comparisons procedure to decide which pairs of colors are significantly different. Summarize your results. Which color would you recommend for attracting cereal leaf beetles?