Third edition

Statistics for Business and Economics

Anderson
Sweeney
Williams
Freeman
Shoesmith

Exercises and Solutions

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter One
Data and Statistics

Textbook Exercises (1-13)
Textbook Exercise Solutions
Supplementary Exercises (14-28)
Supplementary Exercise Solutions

Chapter 1: Data and Statistics

Textbook Exercises:

1. Discuss the differences between statistics as numerical facts and statistics as a discipline or field of study.

2. Every year Condé Nast Traveler conducts an annual survey of subscribers to determine the best new places to stay throughout the world. Table 1.6 shows the ten hotels that were most highly ranked in their 2006 'hot list' survey. Note that (daily) rates quoted are for double rooms and are variously expressed in US dollars, British pounds or euros.
    1. How many elements are in this data set?
    2. How many variables are in this data set?
    3. Which variables are categorical and which variables are quantitative?
    4. What type of measurement scale is used for each of the variables?

3. Refer to Table 1.6.
    a. What is the average number of rooms for the ten hotels?
    b. If €1 = US$1.3149 = £0.8986 compute the average room rate in Euros.

Table 1.6 The ten best new hotels to stay in, in the world

Hot list ranking	Name of property	Country	Room rate	Number of rooms
1	Amangalla, Galle	Sri Lanka	US$574	30
2	Amanwella, Tangalle	Sri Lanka	US$275	30
3	Bairro Alto Hotel, Lisbon	Portugal	€180	55
4	Basico, Playa Del Carmen	Mexico	US$166	15
5	Beit Al Mamlouka	Syria	£75	8
6	Browns Hotel, London	England	£347	117
7	Byblos Art Hotel Villa Amista, Verona	Italy	€270	60
8	Cavas Wine Lodge, Mendoza	Argentina	US$375	14
9	Convento Do Espinheiro Heritage Hotel & Spa, Evora	Portugal	€213	59
10	Cosmopolitan, Toronto	Canada	£150	97

Source: *Condé Nast Traveler*, May 2006 (http://www.cntraveller.co.uk/Special_Features/The_Hot_List_2006/)

    c. What is the percentage of hotels located in Portugal?

d. What is the percentage of hotels with 20 rooms or fewer?

4. Audio systems are typically made up of an MP3 player, a mini disk player, a cassette player, a CD player and separate speakers. The data in Table 1.7 shows the product rating and retail price range for a popular selection of systems. Note that the code Y is used to confirm when a player is included in the system, N when it is not. Output power (watts) details are also provided (Kelkoo Electronics 2006).
   a. How many elements does this data set contain?
   b. What is the population?
   c. Compute the average output power for the sample.

5. Consider the data set for the sample of eight audio systems in Table 1.7.
   a. How many variables are in the data set?
   b. Which of the variables are quantitative and which are categorical?
   c. What percentage of the audio systems has a four star rating or higher?
   d. What percentage of the audio systems includes an MP3 player?

Table 1.7 A sample of eight audio systems

Brand and model	Product rating (# of stars)	Price (£)	MP3 player	Mini disk player	Cassette player	CD (watts) player	Output
Technics SCEH790	1	320–400	Y	N	Y	Y	360
Yamaha M170	3	162–290	N	N	N	Y	50
Panasonic SCPM29	5	188	Y	N	Y	Y	70
Pure Digital DMX50	3	180–230	N	N	N	Y	80
Sony CMTNEZ3	5	60–100	Y	N	Y	Y	30
Philips FWM589	4	143–200	Y	N	N	Y	400
PHILIPS MCM9	5	93–110	Y	N	Y	Y	100
Samsung MM-C6	5	100–130	Y	N	N	Y	40

Source: Kelkoo (http://audiovisual.kelkoo.co.uk)

6. Columbia House provides CDs to its mail-order club members. A Columbia House Music Survey asked new club members to complete an 11-question survey. Some of the questions asked were:

    a. How many CDs have you bought in the last 12 months?

    b. Are you currently a member of a national mail-order book club? (Yes or No)

    c. What is your age?

    d. Including yourself, how many people (adults and children) are in your household?

    e. What kinds of music are you interested in buying? (15 categories were listed, including hard rock, soft rock, adult contemporary, heavy metal, rap and country.)

    Comment on whether each question provides categorical or quantitative data.

7. The Health & Wellbeing Survey ran over a three-week period (ending 19 October 2007) and 389 respondents took part. The survey asked the respondents to respond to the statement, 'How would you describe your own physical health at this time?' (http://inform.glam.ac.uk/news/2007/10/24/health-wellbeing-staff-survey-results/).

    Response categories were strongly agree, agree, neither agree or disagree, disagree, and strongly disagree.

    a. What was the sample size for this survey?

    b. Are the data categorical or quantitative?

    c. Would it make more sense to use averages or percentages as a summary of the data for this question?

    d. Of the respondents, 57 per cent agreed with the statement. How many individuals provided this response?

8. State whether each of the following variables is categorical or quantitative and indicate its measurement scale.

    a. Age.

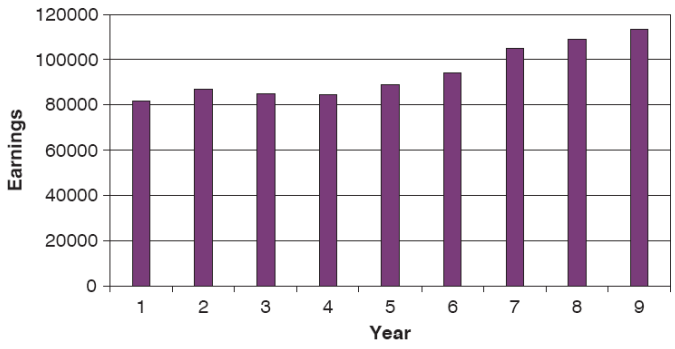
b. Gender.

c. Class rank.

d. Make of car.

e. Number of people favouring closer European integration.

Figure 1.7 Volkswagen's earnings (€m) 1998–2009



9. Figure 1.7 provides a bar chart summarizing the actual earnings for Volkswagen for the years 2000 to 2008 (Source: Volkswagen AG Annual Reports 2001–2008).

    a. Are the data categorical or quantitative?

    b. Are the data times series or cross-sectional?

    c. What is the variable of interest?

    d. Comment on the trend in Volkswagen's earnings over time. Would you expect to see an increase or decrease in 2009?

10. The Hawaii Visitors Bureau collects data on visitors to Hawaii. The following questions were among 16 asked in a questionnaire handed out to passengers during incoming airline flights.

- This trip to Hawaii is my: 1st, 2nd, 3rd, 4th, etc.
- The primary reason for this trip is: (10 categories including vacation, convention, honeymoon)
- Where I plan to stay: (11 categories including hotel, apartment, relatives, camping)

- Total days in Hawaii

a. What is the population being studied?

b. Is the use of a questionnaire a good way to reach the population of passengers on incoming airline flights?

c. Comment on each of the four questions in terms of whether it will provide categorical or quantitative data.

11. A manager of a large corporation recommends a $10,000 raise be given to keep a valued subordinate from moving to another company. What internal and external sources of data might be used to decide whether such a salary increase is appropriate?

12. In a recent study of causes of death in men 60 years of age and older, a sample of 120 men indicated that 48 died as a result of some form of heart disease.
    a. Develop a descriptive statistic that can be used as an estimate of the percentage of men 60 years of age or older who die from some form of heart disease.
    b. Are the data on cause of death categorical or quantitative?
    c. Discuss the role of statistical inference in this type of medical research.

13. In 2007, 75.4 per cent of Economist readers had stayed in a hotel on business in the previous 12 months with 32.4 per cent of readers using first/ business class for travel.
    a. What is the population of interest in this study?
    b. Is class of travel a categorical or quantitative variable?
    c. If a reader had stayed in a hotel on business in the previous 12 months would this be classed as a categorical or quantitative variable?
    d. Does this study involve cross-sectional or time series data?
    e. Describe any statistical inferences The Economist might make on the basis of the survey.

Chapter 1: Data and Statistics

Textbook Exercises Solutions:

1. Statistics can be referred to as numerical facts. In a broader sense, statistics is the field of study dealing with the collection, analysis, presentation and interpretation of data.

2. a. 10
   b. 4
   c. Country is a categorical variable; hot list ranking, number of rooms and room rate are quantitative variables.
   d. Country is nominal; room rate and hot list ranking are ordinal; number of rooms and room rate are ratio.

3. a. Average number of rooms = 485/10 = 48.5 or approximately 49 rooms

   b. Average room rate (€) = 2356.66/10 = 235.67

   c. 2 of 10 are located in Portugal; or 20%

   d. 3 of 10 have 20 rooms or less; or 30%

4. a. 8

   b. All brands of audio systems manufactured.

   c. Average output power = 1130/8 = 141.25 watts

5. a. 7

   b. Product rating, Price, Output are quantitative. MP3 player, Mini Disc player, Cassette player and CD player are categorical.

   c. Number of systems rated 4 stars or higher = 5/8 = 62.5%; approximately 63%.

   d. $\frac{6}{8}(100) = 75\%$

6. Questions a, c, and d provide quantitative data.

   Questions b and e provide categorical data.

7. a. 389

   b. The variable is categorical.

   c. Percentages.

   d. 222 respondents

8. a. Quantitative; ratio

   b. Categorical; nominal

   c. Categorical (Note: Rank is a numeric label that identifies the position of a student in the class. Rank does not indicate how much or how many and is not quantitative.); ordinal

   d. Categorical; nominal

   e. Quantitative; ratio

9. a. Quantitative - Earnings measured in billions of euros.

   b. Time series with 9 observations

   c. Volkswagen's annual earnings.

   d. Time series shows an increase in earnings. An increase would be expected in 2009, but it appears that the rate of increase may be slowing.

10.

   a. All visitors to Hawaii
   b. Yes
   c. First and fourth questions provide quantitative data Second and third questions provide-categorical data

11. Internal data on salaries of other employees can be obtained from the personnel department. External data might be obtained from the Department of Labor or industry associations.

12. a. (48/120)100% = 40% in the sample died from some form of heart disease. This can be used as an estimate of the percentage of all males 60 or older who die of heart disease.
    b. The data on cause of death is categorical.

13. a. All readers of The Economist at the time the survey was conducted.

    b. Categorical

    c. Categorical (stayed in a hotel or not stayed in a hotel)

    d. Cross-sectional - all the data relate to the same time.

    e. Using the sample results, we could infer or estimate that in 2007, 75.4% per cent of the population of *Economist* readers had stayed in a hotel on business in the previous 12 months; also 32.4% of the population of readers used first / business class for travel.

Supplementary Exercises:

14. Statistics released by Emerald, the publisher of *TQM Magazine* (www.brad.ac.uk/acad/management/ec**tqm),** indicate that: the UK provides 30 per cent of contributions to the TQM Magazine; north America and Europe provide 20 per cent; south and east Asia provide 19 per cent; Australasia provides 9 per cent; the Middle East and Africa provide 1 per cent.

    a. Would the geographical source of contributions be described as nominal or ratio data?

    b. What percentage of the applications come from Europe (including the UK) and North America?

15. State whether each of the following variables is categorical or quantitative and indicate its measurement scale.

    a. Annual sales

    b. Soft-drink size (small, medium, large)

    c. European Socio-economic Classification (Class 1 through Class 10)

    d. Earnings per share

    e. Method of payment (cash, cheque, credit card)

16. The Hawaii Visitors Bureau collects data on visitors to Hawaii. The following questions were among 16 asked in a questionnaire handed out to passengers during incoming airline flights in June 2001.

    - This trip to Hawaii is my: 1st, 2nd, 3rd, 4th, etc.

- The primary reason for this trip is: (10 categories including vacation, convention, honeymoon)
- Where I plan to stay: (11 categories including hotel, apartment, relatives, camping)
- Total days in Hawaii

a. What is the population being studied?

b. Is the use of a questionnaire a good way to reach the population of passengers on incoming airline flights?

c. Comment on each of the four questions in terms of whether it will provide categorical or quantitative data.

17. IPFI regularly releases definitive statistics on the global recorded music industry. For 2005, a breakdown of Total Music Sales (physical & digital) by Market (www.ifpi.org/site-content/library/worldsales2005-ff.pdf) was confirmed by IPFI as follows:

	US$m	Local Currency
1 USA	7,012	USD
2 Japan	3,718	JPY
3 UK	2,162	GBP
4 Germany	1,457	EUR
5 France	1,248	EUR
6 Canada	544	CAD
7 Australia	440	AUD
8 Italy	428	EUR
9 Spain	369	EUR
10 Brazil	265	BRL
11 Mexico	263	MXP
12 Netherlands	246	EUR
13 Switzerland	206	CHF
14 Russia	194	RUB
15 Belgium	162	EUR
16 South Africa	159	ZAR
17 Sweden	148	SEK
18 Austria	139	EUR
19 Norway	133	NOK
20 Denmark	113	DKK
Other	1,387	
Total	20,795	

a. Is local currency a categorical or quantitative variable?

b. Construct a bar graph for music sales by country in 2005. Is this graph based on cross-sectional data or time series data?

18. A Business Week North American subscriber study collected data from a sample of 2861 subscribers. Fifty-nine percent of the respondents indicated an annual income of $75,000 or more, and 50% reported having an American Express credit card.

a. What is the population of interest in this study?

b. Is annual income a categorical or quantitative variable?

c. Is ownership of an American Express card a categorical or quantitative variable?

d. Does this study involve cross-sectional or time series data?

e. Describe any statistical inferences Business Week might make on the basis of the survey.

19. A Fall 2002 sample survey of 131 investment managers in Barron's Big Money poll revealed the following (*Barron's, October 28, 2002*):

- 43% of managers classified themselves as bullish or very bullish on the stock market.
- The average expected return over the next 12 months for equities was 11.2%.
- 21% selected health care as the sector most likely to lead the market in the next 12 months.
- When asked to estimate how long it would take for technology and telecom stocks to resume sustainable growth, the managers' average response was 2.5 years.

a. Cite two descriptive statistics.

b. Make an inference about the population of all investment managers concerning the average return expected on equities over the next 12 months.

c. Make an inference about the length of time it will take for technology and telecom stocks to resume sustainable growth.

20. A seven-year medical research study reported that women whose mothers took the drug DES during pregnancy were twice as likely to develop tissue abnormalities that might lead to cancer as were women whose mothers did not take the drug.

a. This study involved the comparison of two populations. What were the populations?

b. Do you suppose the data were obtained in a survey or an experiment?

c. For the population of women whose mothers took the drug DES during pregnancy, a sample of 3980 women showed 63 developed tissue abnormalities that might lead to cancer. Provide a descriptive statistic that could be used to estimate the number of women out of 1000 in this population who have tissue abnormalities.

d. For the population of women whose mothers did not take the drug DES during pregnancy, what is the estimate of the number of women out of 1000 who would be expected to have tissue abnormalities?

e. Medical studies often use a relatively large sample (in this case, 3980). Why?

21. A firm wants to test the advertising effectiveness of a new television commercial. As part of the test, the commercial is shown on a local evening TV news programme in the Czech republic. Two days later, a market research firm conducts a telephone survey to obtain information on recall rates (percentage of viewers who recall seeing the commercial) and impressions of the commercial.

a. What is the population for this study?

b. What is the sample for this study?

c. Why would a sample be used in this situation? Explain.

22. AC Nielsen is the world's leading marketing information company with 21,000 employees worldwide offering services in more than 100 countries. Recently AC Nielsen contributed data

to a study on internet usage in Europe (www.internetworldstats.com/stats4.htm) – selected details from which are summarised below:

EUROPE	Population (m)	% Pop. of World	Internet Users (m)	Penetration %	Usage of World	Use Growth 2000-2006
European Union	462	7.1%	240	51.9%	22.1%	157.5%
EU Candidate Countries	110	1.7%	25	22.7%	2.3%	622.1%
Rest of Europe	235	3.6%	44	18.7%	4.0%	417.5%
TOTAL EUROPE	807	12.4%	309	38.2%	28.4%	193.7%
Rest of World	5,692	87.6%	778	13.7%	71.6%	203.9%
TOTAL WORLD	6,500	100.0%	1,086	16.7%	100.0%	200.9%

a. How many populations were involved in this study and how do you think it / they would be defined?
b. Do you think the internet usage figures here were calculated by census or sample surveys? How do you think such surveys would be carried out?
c. Where is the fastest and slowest growth in internet usage taking place? Any comments on this?

23. The Broadcasters' Audience Research Board (BARB) is responsible for providing estimates of the number of people watching television. This includes which channels and programmes are being watched, at what time, and the type of people who are watching at any one time. BARB provides television audience data on a minute-by-minute basis for channels received within the UK. The data is available for reporting nationally and at ITV and BBC regional level and covers all analogue and digital platforms.

For the years 1990-2005, BARB confirms the Annual % Shares of Viewing (Individuals) to be as follows:

Channel

Year	BBC1	BBC2	ITV 1*	C4	five	Others
1990	37	10	44	9	-	-
1991	34	10	42	10	-	4
1992	34	10	41	10	-	5
1993	33	10	40	11	-	6
1994	32	11	39	11	-	7
1995	32	11	37	11	-	9
1996	33.5	11.5	35.1	10.7	-	10.1
1997	30.8	11.6	32.9	10.6	2.3	11.8
1998	29.5	11.3	31.7	10.3	4.3	12.9
1999	28.4	10.8	31.2	10.3	5.4	14.0
2000	27.2	10.8	29.3	10.5	5.7	16.6
2001	26.9	11.1	26.7	10.0	5.8	19.6
2002	26.2	11.4	24.1	10.0	6.3	22.1
2003	25.6	11.0	23.7	9.6	6.5	23.6
2004	24.7	10.0	22.8	9.7	6.6	26.2
2005	23.3	9.4	21.5	9.7	6.4	29.6

* inc GMTV

Here, the Channel "Others" signifies non-terrestrial channels.

a. Is channel a categorical or quantitative variable?

b. Construct a graph of BBC (BBC1 and BBC2) viewing share over the sixteen year period. Use the horizontal axis to display the year and the vertical axis to display the percentage viewing share. Is this graph based on cross-sectional or time series data?

c. Construct a graph for viewing shares in 2005. Is the graph based on cross-sectional or time series data?

24. In a recent research study (www.springerlink.com/content/3hbfafkg8pnp2uq2/) of TV viewing habits by Greek children, 4876 questionnaires - completed by children with the assistance of their parents - were analysed. Key results were as follows:

- The mean time spent watching TV ranged from 21-32 hours per week.
- The age when children started watching TV correlated with their later educational achievement: good students started watching TV earlier. Bad students, however, watched more TV, as did children from urban areas, and from lower socioeconomic groups.

- Children from households with more than one TV (especially if it was in the child's bedroom) also watched more.
- Children who watched more TV were less compliant with TV restrictions and more likely to imitate TV characters.
- Eating while watching TV was associated with obesity in teenagers.
- Most children watched TV from appropriate distances, with the lights on, and with the sound at medium volume.

a. What do you think the researchers were attempting to measure here?

b. What is the population?

c. Why would a sample be used for this situation?

d. What kinds of decisions or actions are likely to be based on this study?

25. A sample of course percentages for five students showed the following results: 72, 65, 82, 90, 76. Which of the following statements are correct, and which should be challenged as being too generalized?

a. The average course percentage for the sample of five students is 77.

b. The average course percentage for all students who took the exam is 77.

c. An estimate of the average course percentage for all students who took the exam is 77.

d. More than half of the students who take this exam will achieve a percentage of between 70 and 85.

e. If five other students are included in the sample, their course percentage will be between 65 and 90.

26. Recent figures by the European Council on Refugees and Exiles (www.ecre.org/) on illegal immigration into Europe are summarized below.

Country/ region of asylum	2001	2002	2003	2004	2005
Albania	160	110	30	20	30

Austria	30,140	39,350	32,360	24,630	22,470
Belarus	220	160	140	170	210
Belgium	24,550	18,810	16,940	15,360	15,960
Bosnia and Herzegovina	730	580	740	200	150
Bulgaria	2,430	2,890	1,550	1,130	820
Croatia	90	100	60	160	190
Cyprus	1,770	950	4,410	9,860	7,770
Czech Rep.	18,090	8,480	11,400	5,460	4,020
Denmark	12,510	6,070	4,590	3,240	2,260
Estonia	10	10	10	10	10
Finland	1,650	3,440	3,220	3,860	3,560
France	54,290	58,970	59,770	58,550	50,050
Germany	88,290	71,130	50,560	35,610	28,910
Greece	5,500	5,660	8,180	4,470	9,050
Hungary	9,550	6,410	2,400	1,600	1,610
Ireland	10,330	11,630	7,900	4,770	4,320
Italy	9,620	16,020	13,460	9,720	9,500
Latvia	10	30	10	10	20
Liechtenstein	110	100	100	70	50
Lithuania	260	290	180	170	120
Luxembourg	690	1,040	1,550	1,580	800
Malta	120	350	570	1,000	1,170
Moldova, Rep. of	250	110	90	110	110
Netherlands	32,580	18,670	13,400	9,780	12,350
Norway	14,780	17,480	15,960	7,950	5,400
Poland	4,530	5,170	6,910	8,080	5,440
Portugal	230	250	90	110	110
Romania	2,430	1,150	1,080	660	590
Russian Federation	1,680	880	740	910	960
Serbia and Montenegro	150	170	140	60	90
Slovak Republic	8,150	9,700	10,360	11,390	3,490
Slovenia	1,510	700	1,100	1,280	1,600
Spain	9,490	6,310	5,920	5,540	5,260
Sweden	23,520	33,020	31,350	23,160	17,530
Switzerland	20,630	26,130	20,810	14,250	10,060
TfYR Macedonia	200	120	2,280	100	10
Turkey	5,040	3,800	3,950	3,910	3,910
Ukraine	920	460	1,370	1,360	1,740
United Kingdom	91,600	103,080	60,050	40,620	30,460
EU-"Old" (15)	394,990	393,450	309,340	241,000	212,590
EU-"New" (10)	44,000	32,090	31,350	38,860	25,250
EU-Total (25)	438,990	425,540	346,690	279,860	237,840
Nordic countries (5)	52,510	60,130	55,200	38,290	28,840
Western Europe (19)	430,560	437,280	346,290	263,350	228,190
Former Yugoslavia (5)	2,680	1,670	4,320	1,800	2,040
Former USSR (7)	6,620	3,450	3,300	3,620	3,980
Total Europe (44)	492,410	481,740	396,770	312,070	263,210

a. Are these time series or cross-sectional data?

b. What are the elements and variables here?

c. Give an example of an observation in this example.

d. Is the scale of measurement used for the quantitative details here interval or ratio?

e. Comment on any trends revealed by the summary.

27. In January 2005, the European Consumer Centre, Dublin (www.ecic.ie) commissioned Insight Statistical Consulting (ISC), a marketing research company, to undertake research on airline complaints in Ireland. Subsequently a telephone survey of a total of 1067 adults in Ireland was conducted. Amongst the questions asked by ISC were the following:

- Did you take a flight-only journey within the last year i.e. not any part of a package holiday?
- Do you know your rights as an airline passenger?
- Have you heard of the small claims court?

a. What is the population being studied?

b. Do you think the choice of a telephone survey a good way to reach the population of interest?

c. Comment on each of the sample questions in terms of whether it will provide categorical or quantitative data.

28. In a survey in 2001, (www.uri.edu/personal/awel5922/gambling.index.html) nearly 500 customers of the online casinos kennyrogers.com, casinoaustralia.com and goodluck.com were asked by Inland Entertainment Corporation (IEC) to confirm their

Primary language

Gender and

Age range

The questionnaire was used for the survey was in English.

Comment on

a. the quality of the survey design employed by IEC and - following on -

b. the validity / precision of results arising from the survey.

Chapter 1: Data and Statistics

Supplementary Exercises Solutions:

14. a. Nominal
b. 50%

15. a. Quantitative; ratio

b. Categorical; ordinal

c. Categorical; ordinal (assuming employees can be ranked by classification)

d. Quantitative; ratio

e. Categorical; nominal

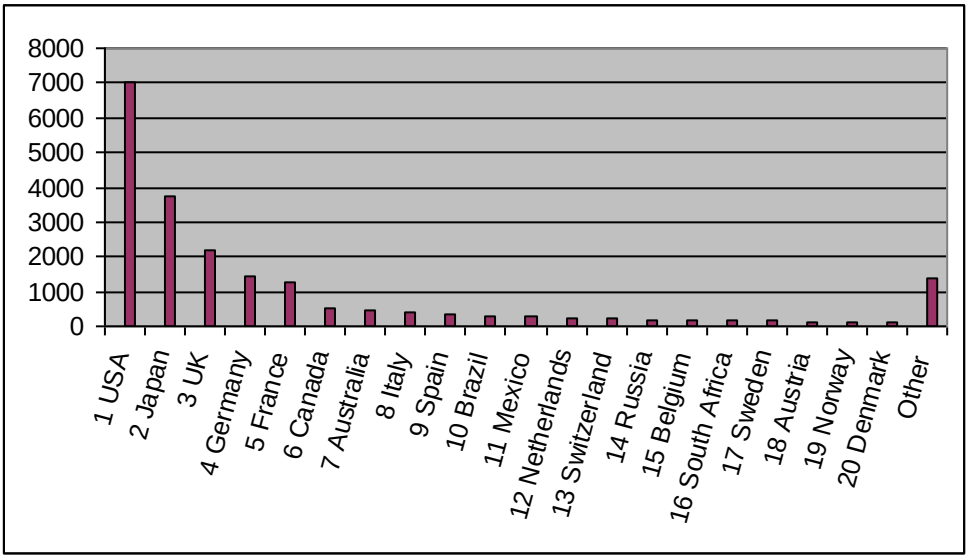
16. a. The population is all visitors coming to the state of Hawaii.

b. Since airline flights carry the vast majority of visitors to the state, the use of questionnaires for passengers during incoming flights is a good way to reach this population. The questionnaire actually appears on the back of a mandatory plants and animals declaration form that passengers must complete during the incoming flight. A large percentage of passengers complete the visitor information questionnaire.

c. Questions 1 and 4 provide quantitative data indicating the number of visits and the number of days in Hawaii. Questions 2 and 3 provide categorical data indicating the categories of reason for the trip and where the visitor plans to stay.

17. a Categorical

b.



Cross-sectional

18. a. All subscribers of Business Week in North America at the time the survey was conducted.

b. Quantitative

c. Categorical (yes or no)

d. Crossectional - all the data relate to the same time.

e. Using the sample results, we could infer or estimate 59% of the population of subscribers have an annual income of $75,000 or more and 50% of the population of subscribers have an American Express credit card.

19. a. 43% of managers were bullish or very bullish.

21% of managers expected health care to be the leading industry over the next 12 months.

b. We estimate the average 12-month return estimate for the population of investment managers to be 11.2%.

c. We estimate the average over the population of investment managers to be 2.5 years.

20. a. The two populations are the population of women whose mothers took the drug DES during pregnancy and the population of women whose mothers did not take the drug DES during pregnancy.

b. It was a survey.

c. 63 / 3.980 = 15.8 women out of each 1000 developed tissue abnormalities.

d. The article reported "twice" as many abnormalities in the women whose mothers had taken DES during pregnancy. Thus, a rough estimate would be 15.8/2 = 7.9 abnormalities per 1000 women whose mothers had *not* taken DES during pregnancy.

e. In many situations, disease occurrences are rare and affect only a small portion of the population. Large samples are needed to collect data on a reasonable number of cases where the disease exists.

21. a. All adult viewers reached by the Czech television station.

b. The viewers contacted in the telephone survey.

c. It would clearly be too costly and time consuming to try to contact all viewers.

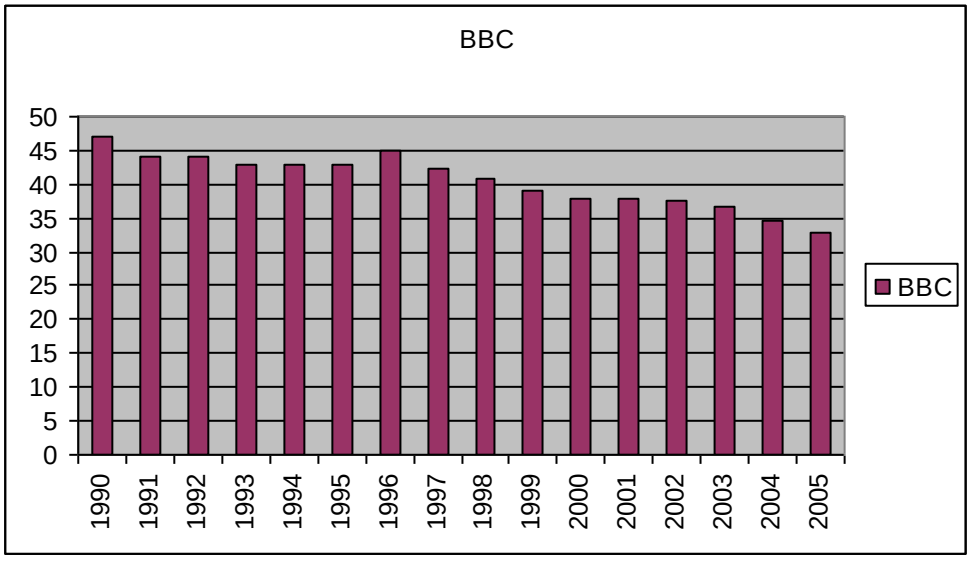
22. a. One. All computers in Europe available to the public for internet usage

b. Sample. A statistically representative panel of internet users.

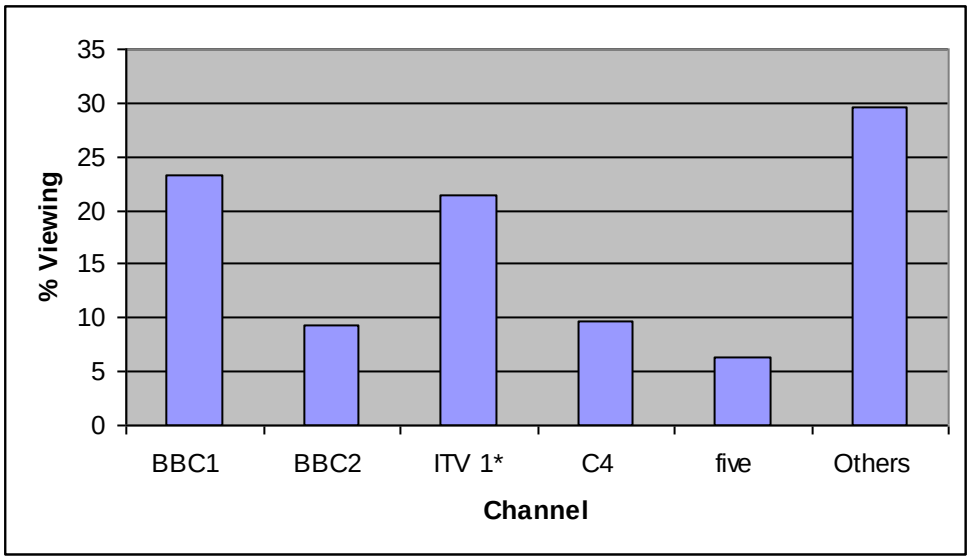
c. EU candidate countries. One measure of the rapid convergence of candidate states with existing EU countries

23. a. Categorical

b.



Time series data

c. 

Cross-sectional

24. a. Percent of television sets that were tuned by Greek children to a particular television show and/or total viewing audience.

b. All television sets in Greece which are available for the children to view. Note this would not include television sets in store displays.

c. A sample is used because it would be too costly to collect data on all television sets in Greece.

d. A demographic understanding of Greek children's TV viewing and possible problems associated with viewing habits.

25. a. This is a statistically correct descriptive statistic for the sample.

b. An incorrect generalization since the data was not collected for the entire population.

c. An acceptable statistical inference based on the use of the word "estimate."

d. While this statement is true for the sample, it is not a justifiable conclusion for the entire population.

e. This statement is not statistically supportable. While it is true for the particular sample observed, it is entirely possible and even very likely that at least some students will be outside the 65 to 90 range of grades.

26. a No

b. Country; annual illegal immigration totals

c. United Kingdom 91,600 103,080 60,050 40,620 30,460

d. Ratio

e. Between 2001 and 2005 illegal immigration into the European countries listed as almost halved.

27. a. The population of Irish airline travellers

b. Yes on the not unreasonable assumption that airline travellers are likely to have phones.

c. All categorical

28. a. Non-English respondents may have difficulty answering a questionnaire in English.

There may also be a bias resulting from the particular casinos chosen for the study.

b. A sample size of 500 is comparatively small by many survey standards so the results are likely to be comparatively imprecise.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Two

Descriptive Statistics – Tabular and Graphical Methods

Textbook Exercises (1-25)
Textbook Exercise Solutions
Supplementary Exercises (26-43)
Supplementary Exercise Solutions

Chapter 2: Descriptive Statistics – Tabular and Graphical Methods

Textbook Exercises:

1 The response to a question has three alternatives: A, B and C. A sample of 120 responses provides 60 A, 24 B and 36 C. Construct the frequency and relative frequency distributions.

2 A partial relative frequency distribution is given below.

Class	Relative frequency
A	0.22
B	0.18
C	0.40
D	

a. What is the relative frequency of class D?

b. The total sample size is 200. What is the frequency of class D?

c. Construct the frequency distribution.

d. Construct the percentage frequency distribution.

3 A questionnaire provides 58 Yes, 42 No and 20 No-opinion answers.

a. In the construction of a pie chart, how many degrees would be in the sector of the pie showing the Yes answers?

b. How many degrees would be in the sector of the pie showing the No answers?

c. Construct a pie chart.

d. Construct a bar chart.

4 CEM4Mobile is a customer experience management company based in Finland. The company does extensive market research in the mobile telecommunications field. Their research shows that the four most popular mobile operating systems in Nordic countries are Apple iOS, Symbian OS, Android and Nokia OS. A sample of 50 page loads from mobile browsing services follows.

Android	Android	Android	Symbian	Apple	Apple	Symbian	Apple	Apple	Android
Android	Symbian	Android	Apple	Nokia	Android	Apple	Apple	Apple	Nokia
Nokia	Apple	Symbian	Apple	Nokia	Symbian	Android	Nokia	Android	Apple
Android	Symbian	Symbian	Apple	Android	Android	Apple	Android	Android	Apple
Apple	Nokia	Symbian	Symbian	Android	Android	Apple	Symbian	Symbian	Android

a. Are these data qualitative or quantitative?

b. Construct frequency and percentage frequency distributions.

c. Construct a bar chart and a pie chart.

d. On the basis of the sample, which mobile operating system was the most popular? Which one was second?

A Wikipedia article listed the six most common last names in Belgium as (in alphabetical order): Jacobs, Janssens, Maes, Mertens, Peeters and Willems. A sample of 50 individuals with one of these last names provided the following data.

Peeters	Peeters	Willems	Janssens	Janssens	Peeters	Jacobs	Maes	Janssens	Mertens
Jacobs	Maes	Peeters	Willems	Jacobs	Maes	Peeters	Janssens	Maes	Maes
Peeters	Maes	Peeters	Maes	Janssens	Janssens	Mertens	Jacobs	Jacobs	Peeters
Mertens	Maes	Peeters	Janssens	Willems	Willems	Peeters	Janssens	Willems	Mertens
Jacobs	Willems	Peeters	Janssens	Mertens	Janssens	Peeters	Mertens	Mertens	Janssens

Summarize the data by constructing the following:

a. Relative and percentage frequency distributions.

b. A bar chart.

c. A pie chart.

d. Based on these data, what are the three most common last names?

6 The flexitime system at Electronics Associates allows employees to begin their working day at 7:00, 7:30, 8:00, 8:30, or 9:00 a.m. The following data represent a sample of the starting times selected by the employees.

7:00	8:30	9:00	8:00	7:30	7:30	8:30	8:30	7:30	7:00
8:30	8:30	8:00	8:00	7:30	8:30	7:00	9:00	8:30	8:00

Summarize the data by constructing the following:

a. A frequency distribution.

b. A percentage frequency distribution.

c. A bar chart.

d. A pie chart.

e. What do the summaries tell you about employee preferences in the flexitime system?

7 A Merrill Lynch Client Satisfaction Survey asked clients to indicate how satisfied they were with their financial consultant. Client responses were coded 1 to 7, with 1 indicating 'not at all satisfied' and 7 indicating 'extremely satisfied'. The following data are from a sample of 60 responses for a particular financial consultant.

5	7	6	6	7	5	5	7	3	6
7	7	6	6	6	5	5	6	7	7
6	6	4	4	7	6	7	6	7	6
5	7	5	7	6	4	7	5	7	6
6	5	3	7	7	6	6	6	6	5
5	6	6	7	7	5	6	4	6	6

a. Comment on why these data are qualitative.

b. Construct a frequency distribution and a relative frequency distribution for the data.

c. Construct a bar chart.

d. On the basis of your summaries, comment on the clients' overall evaluation of the financial consultant.

8 Consider the following data.

14	21	23	21	16	19	22	25	16	16
24	24	25	19	16	19	18	19	21	12
16	17	18	23	25	20	23	16	20	19
24	26	15	22	24	20	22	24	22	20

a. Construct a frequency distribution using classes of 12–14, 15–17, 18–20, 21–23 and 24–26.

b. Construct a relative frequency distribution and a percentage frequency distribution using the classes in (a).

9 Consider the following frequency distribution. Construct a cumulative frequency distribution and a cumulative relative frequency distribution.

Class	Frequency
10–19	10
20–29	14
30–39	17
40–49	7
50–59	2

10 Construct a histogram and an ogive for the data in Exercise 9.

11 Consider the following data.

8.9	10.2	11.5	7.8	10.0	12.2	13.5	14.1	10.0	12.2
6.8	9.5	11.5	11.2	14.9	7.5	10.0	6.0	15.8	11.5

a. Construct a dot plot.

b. Construct a frequency distribution.

c. Construct a percentage frequency distribution.

12 Construct a stem-and-leaf display for the following data.

70	72	75	64	58	83	80	82	76	75	68	65	57	78	85	72

13 Construct a stem-and-leaf display for the following data.

11.3	9.6	10.4	7.5	8.3	10.5	10.0	9.3	8.1	7.7	7.5	8.4	6.3	8.8

14 A doctor's office staff studied the waiting times for patients who arrive at the office with a request for emergency service. The following data with waiting times in minutes were collected over a one-month period.

2	5	10	12	4	4	5	17	11	8	9	8	12	21	6	8	7	13	18	3

Use classes of 0–4, 5–9 and so on in the following:

a. Show the frequency distribution.

b. Show the relative frequency distribution.

c. Show the cumulative frequency distribution.

d. Show the cumulative relative frequency distribution.

e. What proportion of patients needing emergency service wait nine minutes or less?

15 Data for the numbers of units produced by a production employee during the most recent 20 days are shown here.

160	170	181	156	176	148	198	179	162	150
162	156	179	178	151	157	154	179	148	156

Summarize the data by constructing the following:

a. A frequency distribution.

b. A relative frequency distribution.

c. A cumulative frequency distribution.

d. A cumulative relative frequency distribution.

e. An ogive.

16 The closing prices of 40 company shares (in euros) follow.

29.63	34.00	43.25	8.75	37.88	8.63	7.63	30.38
35.25	19.38	9.25	16.50	38.00	53.38	16.63	1.25
48.38	18.00	9.38	9.25	10.00	25.02	18.00	8.00
28.50	24.25	21.63	18.50	33.63	31.13	32.25	29.63
79.38	11.38	38.88	11.50	52.00	14.00	9.00	33.50

a. Construct frequency and relative frequency distributions.

b. Construct cumulative frequency and cumulative relative frequency distributions.

c. Construct a histogram.

d. Using your summaries, make comments and observations about the price of shares.

17 The table below shows the estimated 2013 mid-year population of Kenya, by age group, rounded to the nearest thousand (from the US Census Bureau International Data Base).

Age group	Population (000s)
0 – 9	13 310
10 – 19	9601
20 – 29	7904
30 – 39	5975
40 – 49	3273
50 – 59	2076
60 – 69	1171
70 – 79	555
80+	171

a. Construct a percentage frequency distribution.

b. Construct a cumulative percentage frequency distribution.

c. Construct a cumulative distribution plot (ogive).

d. Using the ogive, estimate the age that divides the population into halves (you will learn in Chapter 3 that this is called the *median*).

18 The Nielsen Home Technology Report provided information about home technology and its usage by individuals aged 12 and older. The following data are the hours of personal computer usage during one week for a sample of 50 individuals.

4.1	1.5	5.9	3.4	5.7	1.6	6.1	3.0	3.7	3.1	4.8	2.0	3.3
11.1	3.5	4.1	4.1	8.8	5.6	4.3	7.1	10.3	6.2	7.6	10.8	0.7
4.0	9.2	4.4	5.7	7.2	6.1	5.7	5.9	4.7	3.9	3.7	3.1	12.1
14.8	5.4	4.2	3.9	4.1	2.8	9.5	12.9	6.1	3.1	10.4		

Summarize the data by constructing the following:

a. A frequency distribution (use a class width of three hours).

b. A relative frequency distribution.

c. A histogram.

d. An ogive.

e. Comment on what the data indicate about personal computer usage at home.

19 The daily high and low temperatures (in degrees Celsius) for 20 cities on one particular day follow.

City	High	Low	City	High	Low
Athens	24	12	Melbourne	19	10
Bangkok	33	23	Montreal	18	11
Cairo	29	14	Paris	25	13
Copenhagen	18	4	Rio de Janeiro	27	16
Dublin	18	8	Rome	27	12
Havana	30	20	Seoul	18	10
Hong Kong	27	22	Singapore	32	24
Johannesburg	16	10	Sydney	20	13
London	23	9	Tokyo	26	15
Manila	34	24	Vancouver	14	6

a. Prepare a stem-and-leaf display for the high temperatures.

b. Prepare a stem-and-leaf display for the low temperatures.

c. Compare the stem-and-leaf displays from parts (a) and (b), and comment on the differences between daily high and low temperatures.

d. Use the stem-and-leaf display from part (a) to determine the number of cities having a high temperature of 25 degrees or above.

e. Provide frequency distributions for both high and low temperature data.

20 The following data are for 30 observations involving two qualitative variables, X and Y. The categories for X are A, B and C; the categories for Y are 1 and 2.

Observation	X	Y	Observation	X	Y
1	A	1	16	B	2
2	B	1	17	C	1
3	B	1	18	B	1
4	C	2	19	C	1
5	B	1	20	B	1
6	C	2	21	C	2
7	B	1	22	B	1
8	C	2	23	C	2
9	A	1	24	A	1
10	B	1	25	B	1
11	A	1	26	C	2
12	B	1	27	C	2
13	C	2	28	A	1
14	C	2	29	B	1
15	C	2	30	B	2

a. Construct a cross-tabulation for the data, with X as the row variable and Y as the column variable.

b. Calculate the row percentages.

c. Calculate the column percentages.

d. What is the relationship, if any, between X and Y?

21 The following 20 observations are for two quantitative variables.

Observation	X	Y	Observation	X	Y
1	−22	22	11	−37	48
2	−33	49	12	34	−29
3	2	8	13	9	−18
4	29	−16	14	−33	31
5	−13	10	15	20	−16
6	21	−28	16	−3	14
7	−13	27	17	−15	18
8	−23	35	18	12	17
9	14	−5	19	−20	−11
10	3	−3	20	−7	−22

a. Construct a scatter diagram for the relationship between X and Y.

b. What is the relationship, if any, between X and Y?

22 Recently, management at Oak Tree Golf Course received a few complaints about the condition of the greens. Several players complained that the greens are too fast. Rather than react to the comments of just a few, the Golf Association conducted a survey of 100 male and 100 female golfers. The survey results are summarized here.

	Male golfers			Female golfers	
	Greens condition			Greens condition	
Handicap	Too fast	Fine	Handicap	Too fast	Fine
Under 15	10	40	Under 15	1	9
15 or more	25	25	15 or more	39	51

and the column labels too fast and fine. Which group shows the highest percentage saying that the greens are too fast?

b. Refer to the initial cross-tabulations. For those players with low handicaps (better players), which group (male or female) shows the highest percentage saying the greens are too fast?

c. Refer to the initial cross-tabulations. For those players with higher handicaps, which group (male or female) shows the highest percentage saying the greens are too fast?

d. What conclusions can you draw about the preferences of men and women concerning the speed of the greens? Are the conclusions you draw from part (a) as compared with parts (b) and (c) consistent? Explain any apparent inconsistencies.

23 The file 'House Sales' on the online platform contains data for a sample of 50 houses advertised for sale in a regional UK newspaper. The first five rows of data are shown for illustration below.

Price (£)	Location	House type	Bedrooms	Reception rooms	Bedrooms + Receptions	Garage capacity
234 995	Town	Detached	4	2	6	1
319 000	Town	Detached	4	2	6	1
154 995	Town	Semi-detached	2	1	3	0
349 950	Village	Detached	4	2	6	2
244 995	Town	Detached	3	2	5	1

a. Prepare a cross-tabulation using sale price (rows) and house type (columns). Use classes of 100 000–199 999, 200 000–299 999, etc. for sale price.

b. Compute row percentages and comment on any relationship between the variables.

24 Refer to the data in Exercise 23.

a. Prepare a cross-tabulation using number of bedrooms and house type.

b. Prepare a frequency distribution for number of bedrooms.

c. Prepare a frequency distribution for house type.

d. How has the cross-tabulation helped in preparing the frequency distributions in parts (b) and (c)?

25 The file ‘OECD 2012’ on the companion website contains data for 33 countries taken from the website of the Organization for Economic Cooperation & Development in mid-2012. The two variables are the Gini coefficient for each country and the percentage of children in the country estimated to be living in poverty. The Gini coefficient is a widely used measure of income inequality. It varies between 0 and 1, with higher coefficients indicating more inequality. The first five rows of data are shown for illustration below.

Country	Child poverty (%)	Income inequality
Australia	14.0	0.336
Austria	7.9	0.261
Belgium	11.3	0.259
Canada	15.1	0.324
Czech Republic	8.4	0.256

a. Prepare a scatter diagram using the data on child poverty and income inequality.

b. Comment on the relationship, if any, between the variables.

Chapter 2

Descriptive Statistics: Tabular and Graphical Methods

Solutions to Textbook Exercises

1.

Class	Frequency	Relative Frequency
A	60	60/120 = 0.50
B	24	24/120 = 0.20
C	36	36/120 = 0.30
	120	1.00

2. a. 1 □ (0.22 + 0.18 + 0.40) = 0.20

   b. 0.20(200) = 40

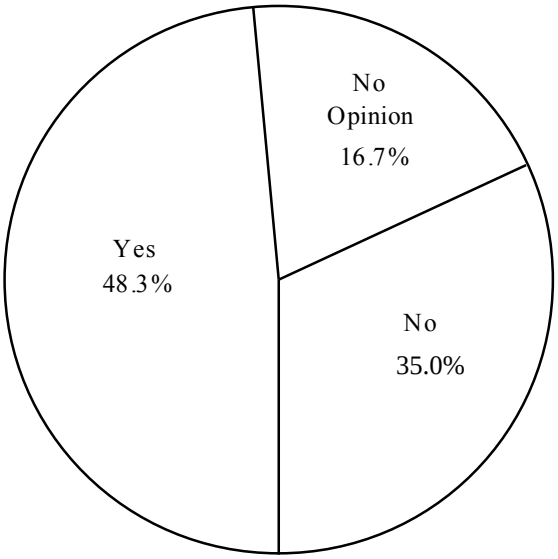
   c/d.

Class	Frequency	Percentage Frequency
A	0.22(200) = 44	22
B	0.18(200) = 36	18
C	0.40(200) = 80	40
D	0.20(200) = 40	20
Total	200	100

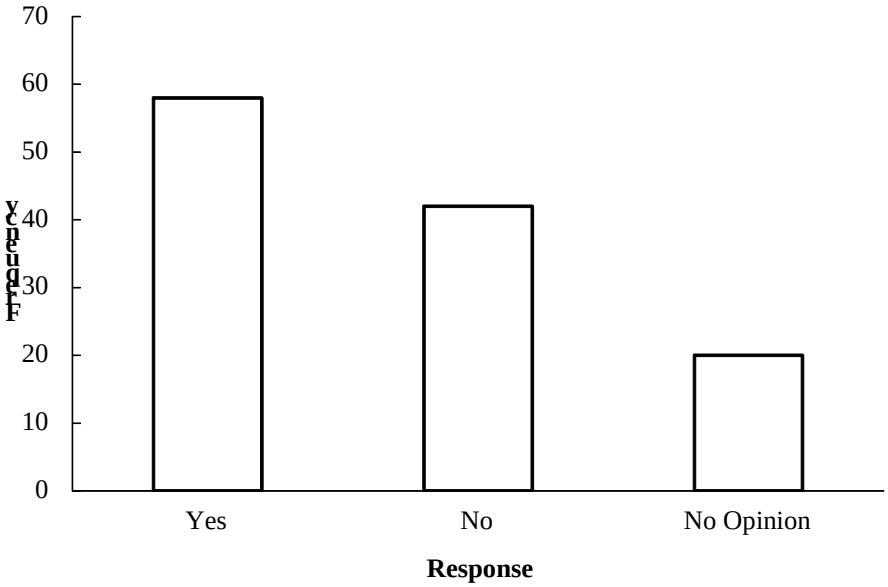
3. a. 360° x 58/120 = 174°

   b. 360° x 42/120 = 126°

c.



d.

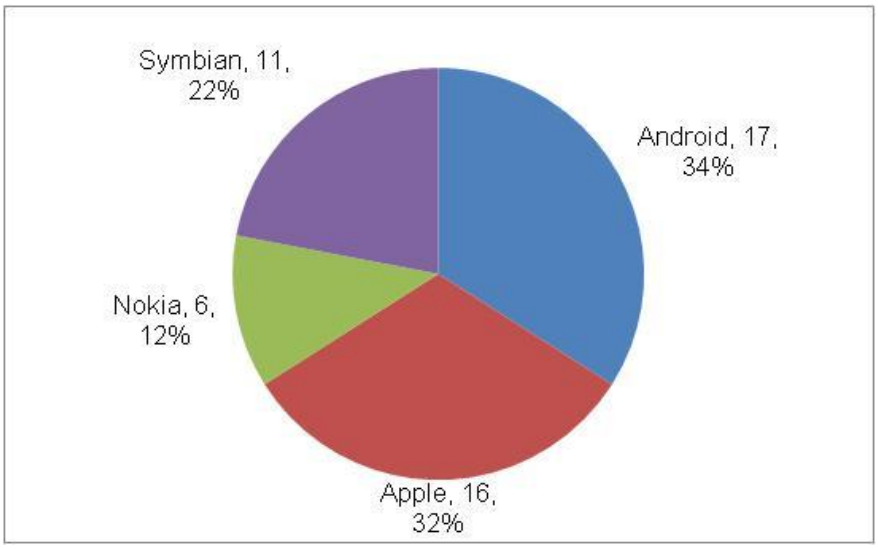


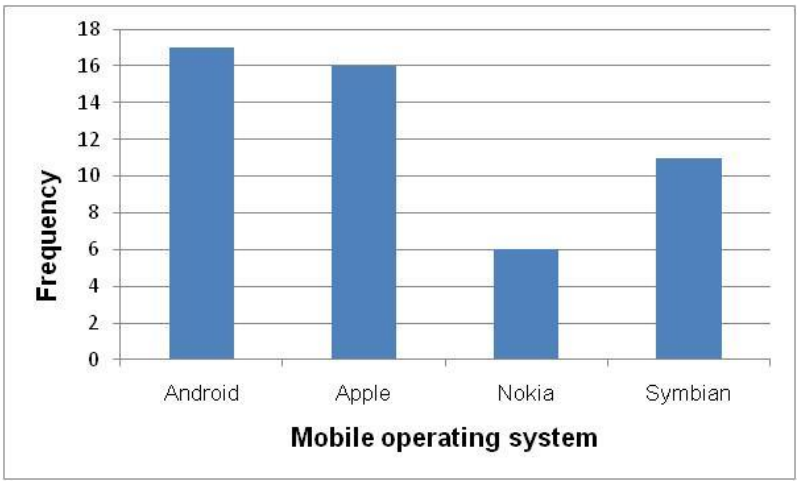
4. a. Qualitative

b.

Operating system	Frequency	Percent frequency
Android	17	34
Apple	16	32
Nokia	6	12
Symbian	11	22

c.



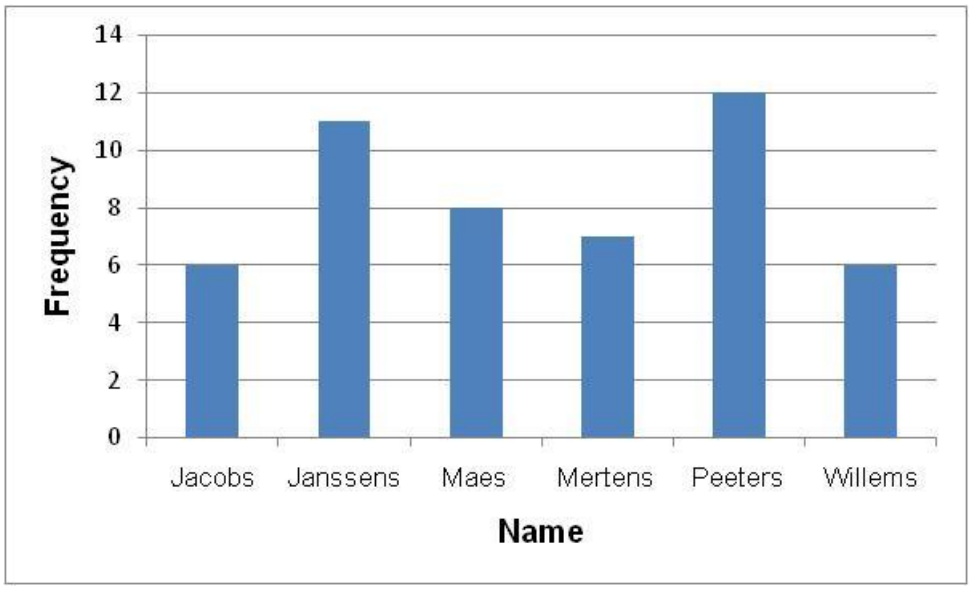


d. Android most popular, Apple close second.

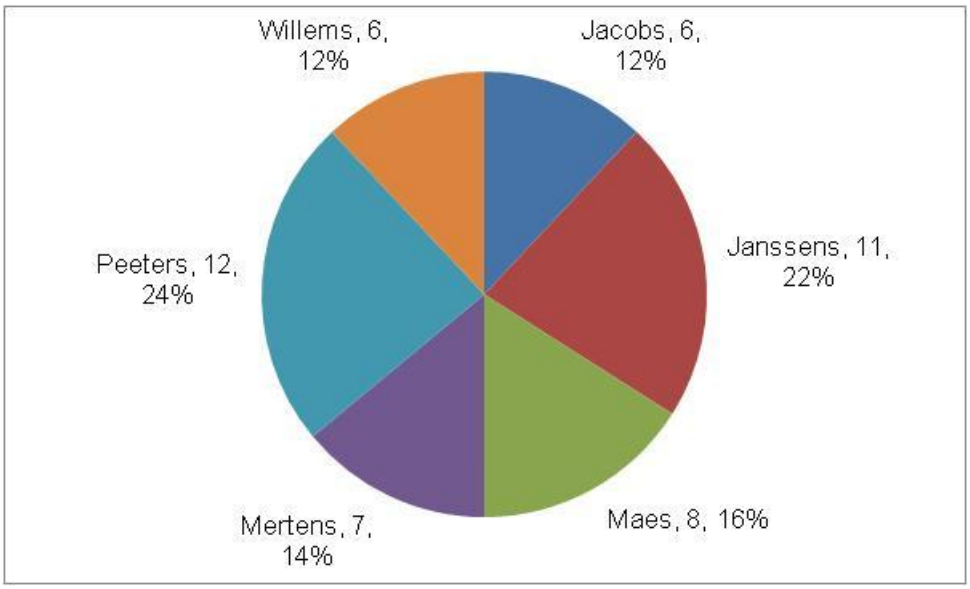
5. a.

Name	Frequency	Relative frequency	Percent frequency
Jacobs	6	0.12	12
Janssens	11	0.22	22
Maes	8	0.16	16
Mertens	7	0.14	14
Peeters	12	0.24	24
Willems	6	0.12	12

b.



c.

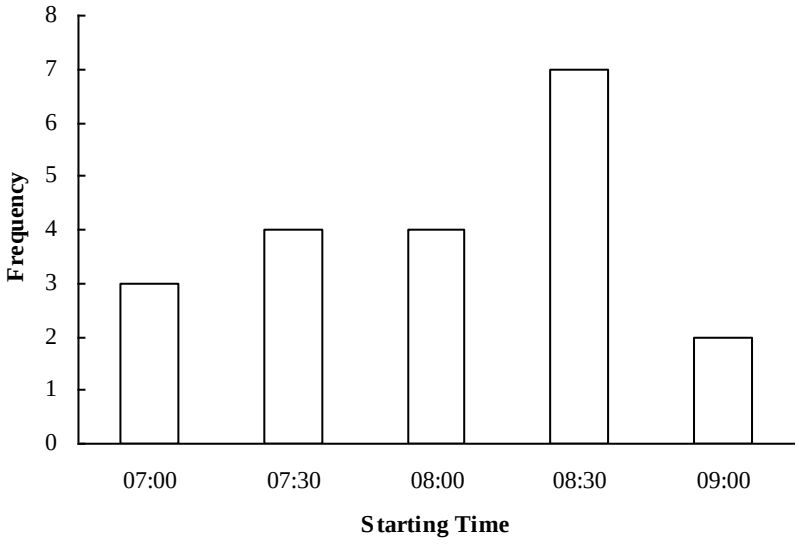


c. Most common is Peeters (24%), next Janssens (22%), then Maes (16%).

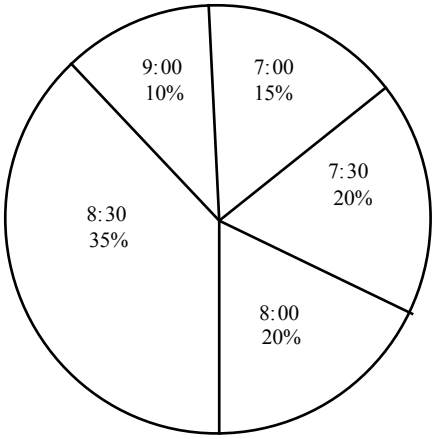
6. a/b.

Starting Time	Frequency	Percentage Frequency
7:00	3	15
7:30	4	20
8:00	4	20
8:30	7	35
9:00	2	10
	20	100

c.



d.



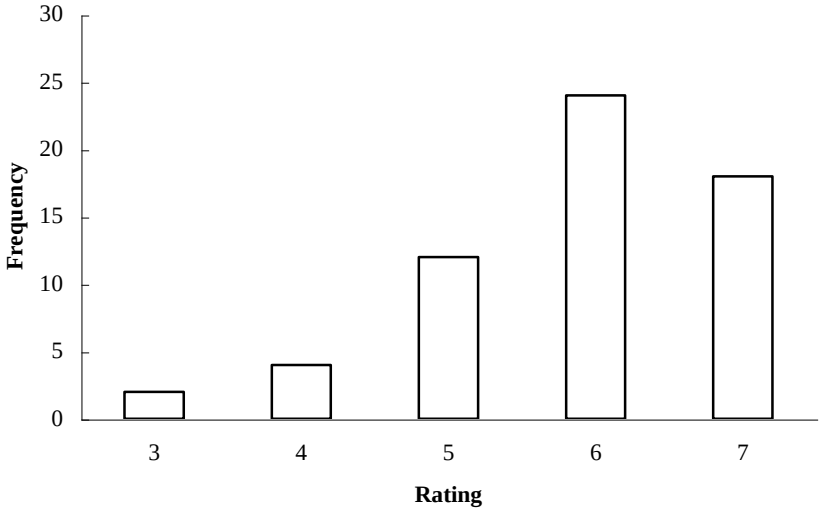
e. The most highly preferred starting time is 8:30 a.m. Starting times of 7:30 and 8:00 a.m. are next.

7. a. The data refer to quality levels from 1 "Not at all Satisfied" to 7 "Extremely Satisfied."

b.

Rating	Frequency	Relative Frequency
3	2	0.03
4	4	0.07
5	12	0.20
6	24	0.40
7	18	0.30
	60	1.00

c. Bar Chart



d. The survey data indicate a high quality of service by the financial consultant. The most common ratings are 6 and 7 (70%) where 7 is extremely satisfied. Only 2 ratings are below the middle scale value of 4. There are no "Not at all Satisfied" ratings.

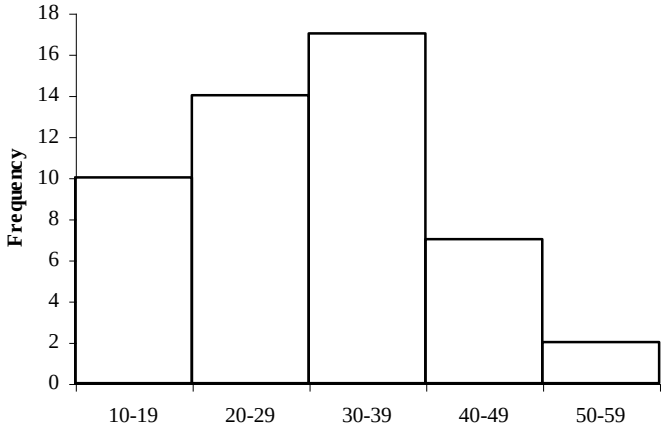
8. a/b

Class	Frequency	Relative Frequency	Percentage Frequency
12-14	2	0.050	5.0
15-17	8	0.200	20.0
18-20	11	0.275	27.5
21-23	10	0.250	25.5
24-26	9	0.225	22.5
Total	40	1.000	100.0

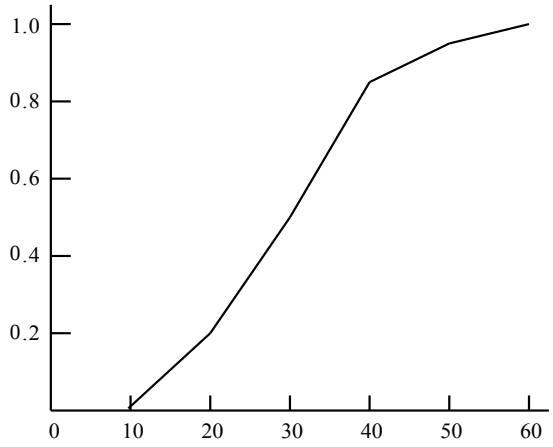
9.

Class	Cumulative Frequency	Cumulative Relative Frequency
less than or equal to 19	10	0.20
less than or equal to 29	24	0.48
less than or equal to 39	41	0.82
less than or equal to 49	48	0.96
less than or equal to 59	50	1.00

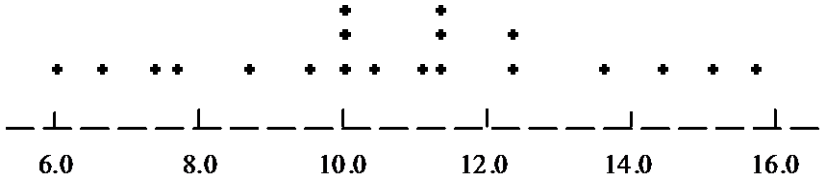
10. Histogram



Ogive



11. a.



b/c.

Class	Frequency	Percentage Frequency
6.0 - 7.9	4	20
8.0 - 9.9	2	10
10.0 - 11.9	8	40
12.0 - 13.9	3	15
14.0 - 15.9	3	15
	20	100

12. Unordered stem-and-leaf:

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

Ordered stem and leaf

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

13. Leaf Unit = 0.1

6	3
7	5 5 7
8	1 3 4 8
9	3 6
10	0 4 5
11	3

14. a/b.

Waiting Time	Frequency	Relative Frequency
0 - 4	4	0.20
5 - 9	8	0.40
10 - 14	5	0.25
15 - 19	2	0.10
20 - 24	1	0.05
Totals	20	1.00

c/d.

Waiting Time	Cumulative Frequency	Cumulative Relative Frequency
Less than or equal to 4	4	0.20
Less than or equal to 9	12	0.60
Less than or equal to 14	17	0.85
Less than or equal to 19	19	0.95
Less than or equal to 24	20	1.00

e. 12/20 = 0.60

15. a/b.

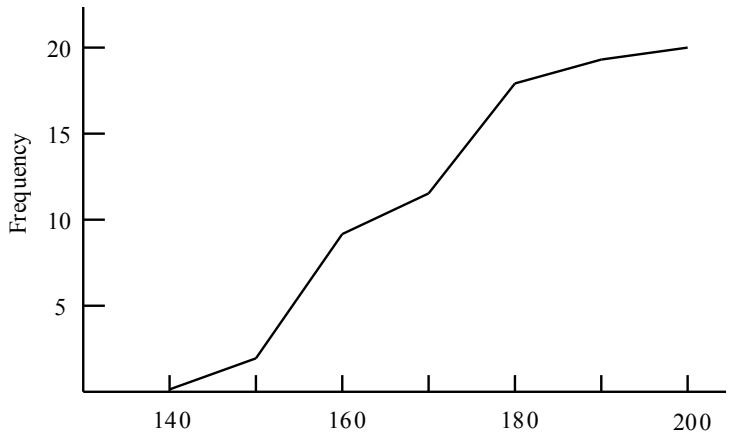
Number	Frequency	Relative Frequency
140 - 149	2	0.10
150 - 159	7	0.35
160 - 169	3	0.15
170 - 179	6	0.30
180 - 189	1	0.05
190 - 199	1	0.05

Totals	20	1.00

c/d.

Number	Cumulative Frequency	Cumulative Relative Frequency
Less than or equal to 149	2	0.10
Less than or equal to 159	9	0.45
Less than or equal to 169	12	0.60
Less than or equal to 179	18	0.90
Less than or equal to 189	19	0.95
Less than or equal to 199	20	1.00

e.



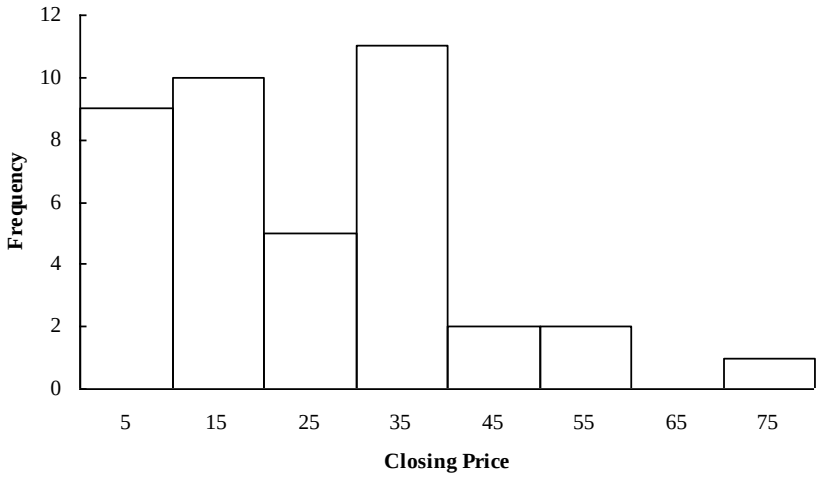
16. a.

Closing Price	Frequency	Relative Frequency
0 - 9.99	9	0.225
10 – 19.99	10	0.250
20 – 29.99	5	0.125
30 – 39.99	11	0.275
40 – 49.99	2	0.050
50 – 59.99	2	0.050
60 – 69.99	0	0.000
70 – 79.99	1	0.025
Totals	40	1.000

b.

Closing Price	Cumulative Frequency	Cumulative Relative Frequency
Less than or equal to 9.99	9	0.225
Less than or equal to 19.99	19	0.475
Less than or equal to 29.99	24	0.600
Less than or equal to 39.99	35	0.875
Less than or equal to 49.99	37	0.925
Less than or equal to 59.99	39	0.975
Less than or equal to 69.99	39	0.975
Less than or equal to 79.99	40	1.000

c.

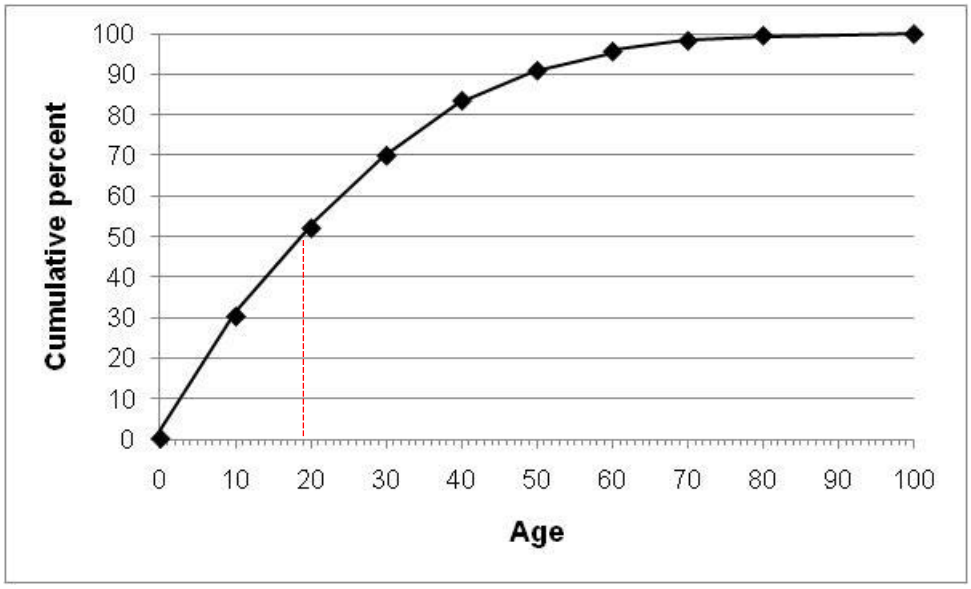


d. Over 87% of shares trade for less than €40 a share and 60% trade for less than €30 per share.

17. a/b.

Age	Frequency	Percent frequency	Cumulative percent frequency
0-9	13310	30.23	30.23
10-19	9601	21.80	52.03
20-29	7904	17.95	69.98
30-39	5975	13.57	83.55
40-49	3273	7.43	90.98
50-59	2076	4.71	95.69
60-69	1171	2.66	98.35
70-79	555	1.26	99.61
80+	171	0.39	100.00
Total	44,036	100.00	

c.

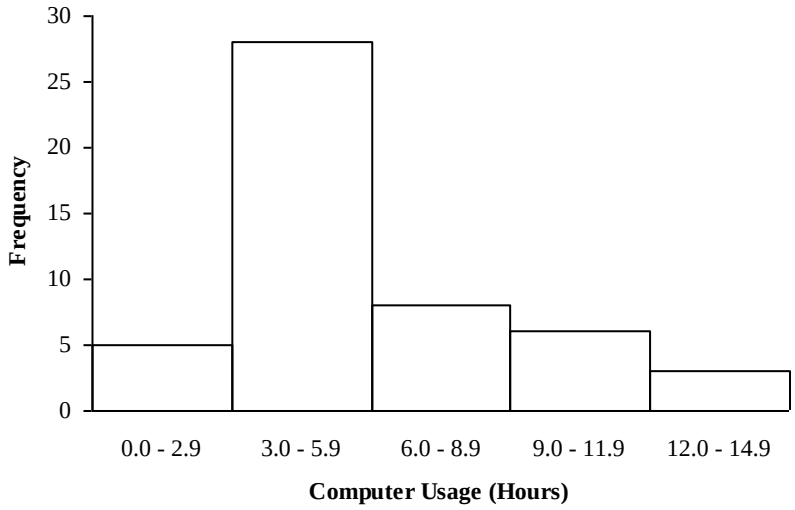


d. Dropping a vertical from the 50% point on the ogive indicates a median age of around 19 years.

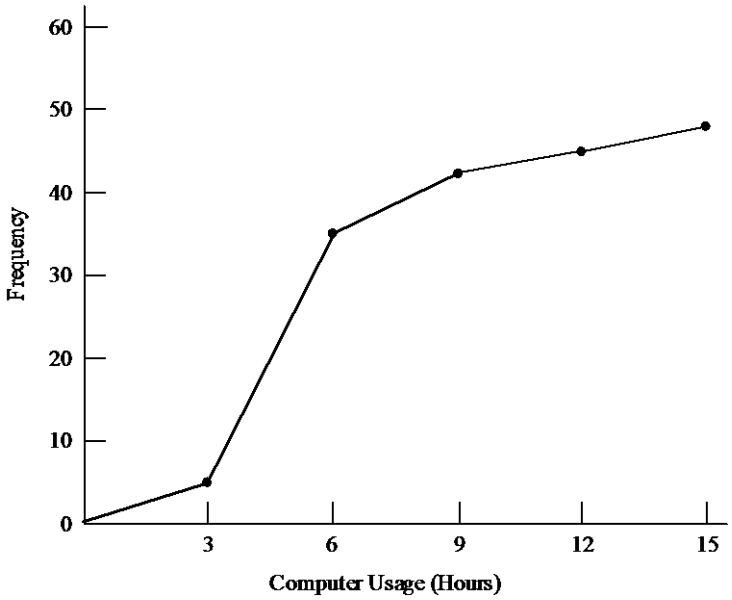
18. a/b.

Computer Usage (Hours)	Frequency	Relative Frequency
0.0 - 2.9	5	0.10
3.0 - 5.9	28	0.56
6.0 - 8.9	8	0.16
9.0 - 11.9	6	0.12
12.0 - 14.9	3	0.06
Total	50	1.00

c.



d.



e. The majority of the computer users are in the 3 to 6 hour range. Usage is somewhat skewed toward the right with 3 users in the 12 to 15 hour range.

19. a/b.

High Temperature		Low Temperature	
0		0	4
0		0	6 8 9
1	4	1	0 0 0 1 2 2 3 3 4
1	6 8 8 8 8 9	1	5 6
2	0 3 4	2	0 2 3 4 4
2	5 6 7 7 7 9	2	
3	0 2 3 4	3	

c. It is clear that the range of low temperatures is below the range of high temperatures. Looking at the stem-and-leaf displays side by side, it appears that the range of low temperatures is about 10 degrees below the range of high temperatures.

d. There are two stems showing high temperatures of 25 degrees or higher. They show 10 cities with high temperatures of 25 degrees or higher.

e.

Temperature	Frequency	
	High Temp.	Low Temp.
0 - 4	0	1
5 - 9	0	3
10 - 14	1	9
15 - 19	6	2
20 - 24	3	5
25 - 29	6	0
30 - 34	4	0
Total	20	20

20 a.

		Y		
		1	2	Total
X	A	5	0	5
	B	11	2	13
	C	2	10	12
	Total	18	12	30

b.

		Y = 1	2	Total
X	A	100.0	0.0	100.0
	B	84.6	15.4	100.0
	C	16.7	83.3	100.0

c.

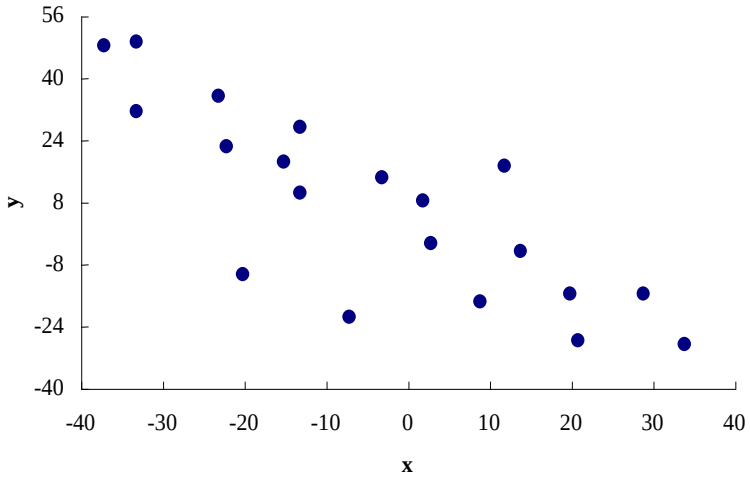
		Y = 1	2
X	A	27.8	0.0
	B	61.1	16.7
	C	11.1	83.3
	Total	100.0	100.0

d. *X* = A values are always paired with *Y* = 1

X = B values are most often paired with *Y* = 1

X = C values are most often paired with *Y* = 2

21. a.



b. There is a negative relationship between *X* and *Y*; *Y* decreases as *X* increases.

22. a. The cross-tabulation of condition of the greens by gender is below.

	Green Condition		
Gender	Too Fast	Fine	Total
Male	35	65	100
Female	40	60	100
Total	75	125	200

The female golfers have the highest percentage saying the greens are too fast: 40%.

b. 10% of the women think the greens are too fast. 20% of the men think the greens are too fast. So, for the low handicappers, the men have a higher percentage who think the greens are too fast.

c. 43% of the woman think the greens are too fast. 50% of the men think the greens are too fast. So, for the high handicappers, the men have a higher percentage who think the greens are too fast.

d. This is an example of Simpson's Paradox. At each handicap level a smaller percentage of the women think the greens are too fast. But, when the cross-tabulations are aggregated, the result is reversed and we find a higher percentage of women who think the greens are too fast. The hidden variable explaining the reversal is handicap level. Fewer people with low handicaps think the greens are too fast, and there are more men with low handicaps than women.

23. a/b.

		House type							
		Detached		Semi-detached		Terraced		Total	
		Count	Row N %	Count	Row N %	Count	Row N %	Count	Row N %
Price class	100,000-199,999	2	25.0%	2	25.0%	4	50.0%	8	100.0%
	200,000-299,999	12	70.6%	2	11.8%	3	17.6%	17	100.0%
	300,000-399,999	10	100.0%	0	.0%	0	.0%	10	100.0%
	400,000-499,999	6	66.7%	0	.0%	3	33.3%	9	100.0%
	500,000-599,999	3	100.0%	0	.0%	0	.0%	3	100.0%
	600,000-699,999	1	100.0%	0	.0%	0	.0%	1	100.0%
	700,000-799,999	1	100.0%	0	.0%	0	.0%	1	100.0%
	800,000-899,999	1	100.0%	0	.0%	0	.0%	1	100.0%
	Total	36	72.0%	4	8.0%	10	20.0%	50	100.0%

b. All houses sold at £500,000 and above were detached houses. In the lower price classes, some were semi-detached or terraced. In the lowest price class (£100,000 - £199,999) half were terraced.

24. a.

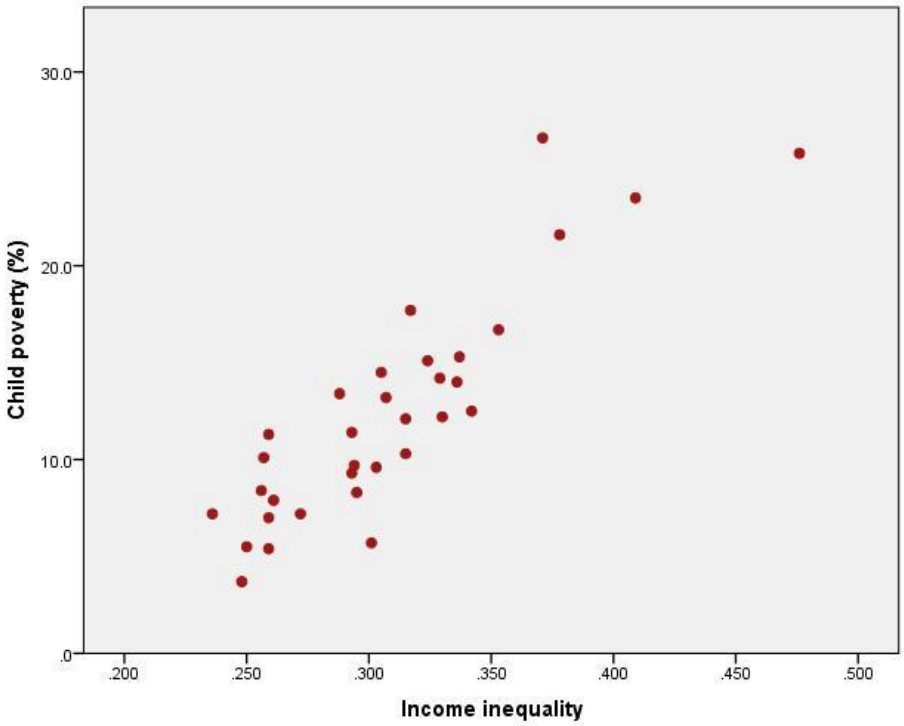
		House type			
		Detached	Semi-detached	Terraced	Total
Bedrooms	2	1	2	3	6
	3	7	1	3	11
	4	22	1	3	26
	5	4	0	1	5
	6	2	0	0	2
Total		36	4	10	50

b. See ‘Total’ column above.

c. See ‘Total’ row above.

d. The frequency distributions in (b) and (c) are formed from the row and column totals respectively of the cross-tabulation.

25. a.



b. There is a strong, positive (approximately linear) relationship between the two variables.

Chapter 2

Descriptive Statistics: Tabular and Graphical Methods

Supplementary Exercises

26. A restaurant/wine bar uses a questionnaire to ask customers how they rate the service, food quality, cocktails, prices, and atmosphere at the restaurant. Each characteristic is rated on a scale of outstanding (O), very good (V), good (G), average (A), and poor (P). Use descriptive statistics to summarize the following data collected on food quality. What is your feeling about the food quality ratings at the restaurant?

G O V G A O V O V G O V A V O P V O G A O
O O G O V V A G O V P V O O G O O V O G A
O V O O G V A G

27. File "BwBooks"

The eight best-selling paperback business books in February 2000 are listed in the table below (*Business Week*, April 3, 2000). A sample of book purchases is provided in the file "BwBooks".

The 7 Habits of Highly Effective People
Investing for Dummies
The Ernst & Young Tax Guide 2000
The Millionaire Next Door
The Motley Fool Investment Guide
Rich Dad, Poor Dad
The Wall Street Guide to Understanding Money and Investing
What Color is Your Parachute? 2000

a. Construct frequency and percentage frequency distributions for the data. Group any books with a frequency of 5% or less in an "other" category.

b. Rank the best-selling books.

c. What percentage of the sales are represented by *The Millionaire Next Door* and *Rich Dad, Poor Dad*?

28. Each of the FTSE 100 companies belongs to a particular category of an industry classification. A sample of 20 companies with their corresponding industry category follows.

Company	**Industry**	**Company**	**Industry**
Alliance Unichem	Pharmaceuticals	Lloyds TSB	Banks
Assoc. Br. Food	Food	Marks & Spencer	Retail
Barclays	Banks	Next	Retail
BP	Oil	Reed Elsevier	Media
BSkyB	Media	Reuters Group	Media
Cadbury Schweppes	Food	Royal Bank of Scot.	Banks
Dixons Group	Retail	Royal Dutch Shell	Oil
Gallaher Group	Tobacco	Severn Trent	Water
Imperial Tobacco	Tobacco	Shire Pharmac.	Pharmaceuticals
Kelda Group	Water	Tate & Lyle	Food

a. Construct a frequency distribution showing the number of companies in each industry category.

b. Construct a percentage frequency distribution.

c. Construct a bar chart for the data.

d. Construct a pie chart for the data.

29. Figures available on the Broadcasters' Audience Research Board website in October 2008 showed that four of the most popular shows broadcast on terrestrial television in the UK were The X Factor, Coronation Street, A Touch of Frost and Strictly Come Dancing. Data indicating the favourite show of a sample of 50 viewers follows.

Strictly	Strictly	X Factor	Coronation	X Factor	X Factor	Coronation	X Factor	X Factor	Strictly
Strictly	Frost	Coronation	X Factor	Coronation	Strictly	X Factor	X Factor	X Factor	Coronation
Coronation	X Factor	Frost	X Factor	Coronation	Frost	Strictly	Coronation	Strictly	X Factor
Strictly	Frost	Frost	X Factor	Strictly	Strictly	X Factor	X Factor	Coronation	X Factor
X Factor	Coronation	Coronation	Coronation	X Factor	Strictly	X Factor	Frost	Frost	Strictly

a. Are these data qualitative or quantitative?

b. Construct frequency and percentage frequency distributions.

c. Construct a bar chart and a pie chart.

d. On the basis of the sample, which television show was the most popular? Which one was second?

30. A Wikipedia article (November 2008) listed the five most common last names in Israel as (in alphabetical order): Biton, Cohen, Levi, Mizrachi and Peretz. A sample of 50 individuals with one of these last names provided the following data.

Cohen	Cohen	Peretz	Cohen	Cohen	Cohen	Levi	Levi	Cohen	Mizrachi
Biton	Levi	Cohen	Peretz	Levi	Levi	Cohen	Cohen	Levi	Levi
Cohen	Cohen	Cohen	Levi	Cohen	Cohen	Mizrachi	Biton	Biton	Cohen
Mizrachi	Levi	Cohen	Cohen	Peretz	Peretz	Cohen	Cohen	Peretz	Mizrachi
Levi	Peretz	Cohen	Cohen	Mizrachi	Cohen	Cohen	Mizrachi	Mizrachi	Cohen

Summarize the data by constructing the following:

a. Relative and percentage frequency distributions.

b. A bar chart.

c. A pie chart.

d. Based on these data, what are the three most common last names?

31. File "Golf"

Golf Magazine's Top 100 Teachers were asked the question, "What is the most critical area that prevents golfers from reaching their potential?" The possible responses were lack of accuracy, poor approach shots, poor mental approach, lack of power, limited practice, poor putting, poor short game, and poor strategic decisions. The data obtained (*Golf Magazine*, February 2002) are in the file "Golf".

a. Construct a frequency distribution and a percentage frequency distribution.

b. Which four critical areas most often prevent golfers from reaching their potential?

32. Consider the following two frequency distributions. The first is an income distribution. The second frequency distribution shows exam scores for students in a college statistics course.

Income (€000s)	**Frequency (000s)**	**Exam score**	**Frequency**
0 – 24	60	Below 30	2
25 – 49	33	30 – 39	5
50 – 74	20	40 – 49	6
75 – 99	6	50 – 59	13
100 – 124	4	60 – 69	32
125 – 149	2	70 –79	78
150 – 174	1	80 – 89	43
175 – 199	1	90 – 99	21
Total	127	Total	200

a. Construct a histogram for the income data. What evidence of skewness does it show? Does this skewness make sense? Explain.

b. Construct a histogram for the exam score data. What evidence of skewness does it show? Explain.

33. The following data are from a sample of 25 households, and show the amount (€) spent in the past year on books and magazines.

280 496 382 202 287 266 119 10 385 135 475 255 379
267 24 42 25 283 110 423 160 123 16 243 363

a. Construct a frequency distribution and relative frequency distribution for the data.
b. Construct a histogram. Comment on the shape of the distribution.
c. Comment on the annual spending on books and magazines for families in the sample.

34. The following data are salaries of 50 senior marketing directors in multi-national companies. Data are in thousands of euros.

145 95 148 112 132 140 162 118 170 144 145 127 148 165 138
173 113 104 141 142 116 178 123 141 138 127 143 134 136 137
155 93 102 154 142 134 165 123 124 124 138 160 157 138 131
114 135 151 138 157

a. What are the lowest and highest salaries?
b. Use a class width of €15,000 and prepare tabular summaries of the annual salary data.
c. What proportion of the annual salaries are €135,000 or less?
d. What percentage of the annual salaries are more than €150,000?
e. Prepare a histogram. Comment on the shape of the distribution.

35 The table below shows the estimated 2009 mid-year population of Zambia, by age group, rounded to the nearest thousand (from the US Census Bureau International Data Base).

Age group	Population (000s)
0 – 4	2005
5 – 9	1749
10 – 14	1591
15 – 19	1440
20 – 24	1253
25 – 29	1022
30 – 34	770
35 – 39	536
40 – 44	369
45 – 49	288
50 – 54	227
55 – 59	186
60 – 64	146
65 – 69	113
70 – 74	83
75 – 79	50
80+	33

a. Construct a percentage frequency distribution.

b. Construct a cumulative percentage frequency distribution.

c. Construct an ogive.

d. Using the ogive, estimate the median age of the population.

36. A psychologist developed a new test of adult intelligence. The test was administered to 20 individuals, and the following data were obtained.

114 99 131 124 117 102 106 127 119 115

98 104 144 151 132 106 125 122 118 118

Construct a stem-and-leaf display for the data.

37. The *American Association of Individual Investors* conducts an annual survey of discount brokers. The following prices charged are from a sample of 24 discount brokers (*AAII Journal*, January 2003). The two types of trades are a broker-assisted trade of 100 shares at $50 per share and an online trade of 500 shares at $50 per share.

	Broker-Assisted	**Online trade**		**Broker-Assisted**	**Online trade**
Accutrade	30.00	29.95	Merrill Lynch Direct	50.00	29.95
Ameritrade	24.99	10.99	Muriel Siebert	45.00	14.95
Banc of America	54.00	24.95	NetVest	24.00	14.00
Brown & Co.	17.00	5.00	Recom Securities	35.00	12.95
Charles Schwab	55.00	29.95	Scottrade	17.00	7.00
CyberTrader	12.95	9.95	Sloan Securities	39.95	19.95
E*TRADE Securities	49.95	14.95	Strong Investments	55.00	24.95
First Discount	35.00	19.75	TD Waterhouse	45.00	17.95
Freedom Investments	25.00	15.00	T. Rowe Price	50.00	19.95
Harrisdirect	40.00	20.00	Vanguard	48.00	20.00
Investors National	39.00	62.50	Wall Street Discount	29.95	19.95
MB Trading	9.95	10.55	York Securities	40.00	36.00

a. Round the trading prices to the nearest dollar and construct a stem-and-leaf display for 100 shares at $50 per share. Comment on what you learned about trading prices.

b. Round the trading prices to the nearest dollar and construct a stretched stem-and-leaf display for 500 shares online at $50 per share. Comment on what you learned about online trading prices.

38. Periodically *Barron's* publishes earnings forecasts for the companies listed in the Dow Jones Industrial Average. The following are the year 2000 forecasts of price/earnings (P/E) ratios for these companies implied by *Barron's* earnings forecasts (*Barron's,* February 14, 2000).

    a. Construct a stem-and-leaf display for the data.

    b. Use the results of the stem-and-leaf display to construct a frequency distribution and a percentage frequency distribution.

Company	**2000 P/E Forecast**	**Company**	**2000 P/E Forecast**
AT&T	23	Honeywell	13
Alcoa	15	IBM	28
American Express	25	Intel	37
Boeing	16	International Paper	14
Caterpillar	13	Johnson & Johnson	23
Citigroup	17	McDonald's	23
Coca-Cola	39	Merck	25
Disney	47	Microsoft	60
Dupont	18	Minnesota Mining	19
Eastman Kodak	11	J. P. Morgan	11
Exxon/Mobil	22	Philip Morris	5
General Electric	37	Proctor & Gamble	26
General Motors	8	SBC Comm.	19
Hewlett-Packard	36	United Technologies	14
Home Depot	48	Wal-Mart	40

39. File "OccupSat"

A study of job satisfaction was conducted for four occupations: lawyer, physical therapist, systems analyst and cabinetmaker. Job satisfaction was measured using an 18-item questionnaire with each question receiving a response score of 1 to 5 and higher scores indicating greater satisfaction. The sum of the 18 scores provides the job satisfaction score for each individual in the sample. The data are in the file "OccupSat".

a. Construct a cross-tabulation of occupation and job satisfaction score.
b. Compute the row percentages for your cross-tabulation in part (a).
c. What observations can you make concerning the level of job satisfaction for these occupations?

40. File "Fortune"

The file "Fortune" provides data on stockholders' equity, market value, and profits for a sample of 50 *Fortune* 500 companies (*Fortune*, April 26, 1999).

a. Prepare a cross-tabulation for the variables Stockholders' Equity and Profit. Use classes of width 200 (starting at zero) for Profit, and classes of width 1200 (starting at zero) for Stockholders' Equity.
b. Compute the row percentages for your cross-tabulation in part (a)
c. What relationship, if any, do you notice between Profit and Stockholders' Equity?

41. File "Fortune"

Refer to the data set in the file "Fortune".

a. Prepare a cross-tabulation for the variables Market Value and Profit.
b. Compute the row percentages for your cross-tabulation in part (a).
c. Comment on any relationship between the variables.

42. File "Fortune"

Refer to the data set in the file "Fortune".

a. Prepare a scatter diagram to show the relationship between the variables Profit and Stockholders' Equity.

b. Comment on any relationship between the variables.

43. The daily high and low temperatures (in degrees Celsius) for 20 cities on one particular day follow.

City	High	Low	City	High	Low
Athens	24	12	Melbourne	19	10
Bangkok	33	23	Montreal	18	11
Cairo	29	14	Paris	25	13
Copenhagen	18	4	Rio de Janeiro	27	16
Dublin	18	8	Rome	27	12
Havana	30	20	Seoul	18	10
Hong Kong	27	22	Singapore	32	24
Johannesburg	16	10	Sydney	20	13
London	23	9	Tokyo	26	15
Manila	34	24	Vancouver	14	6

a. Construct a scatter diagram to show the relationship between the two variables, high temperature and low temperature.

b. Comment on the relationship between high and low temperature.

Chapter 2

Descriptive Statistics: Tabular and Graphical Methods

Supplementary Exercises Solutions

26.

Rating	Frequency	Relative Frequency
Outstanding	19	0.38
Very Good	13	0.26
Good	10	0.20
Average	6	0.12
Poor	2	0.04
	50	1.00

Management should be pleased with these results. 64% of the ratings are very good to outstanding. 84% of the ratings are good or better. Comparing these ratings with previous results will show whether or not the restaurant is making improvements in its ratings of food quality.

27. a.

Book	Frequency	Percentage Frequency
7 Habits	10	16.66
Millionaire	16	26.67
Motley	9	15.00
Dad	13	21.67
WSJ Guide	6	10.00
Other	6	10.00
Total:	60	100.00

The Ernst & Young Tax Guide 2000 with a frequency of 3, *Investing for Dummies* with a frequency of 2, and *What Color is Your Parachute? 2000* with a frequency of 1 are grouped in the "Other" category.

b. The rank order from first to fifth is: Millionaire, Dad, 7 Habits, Motley, and WSJ Guide.

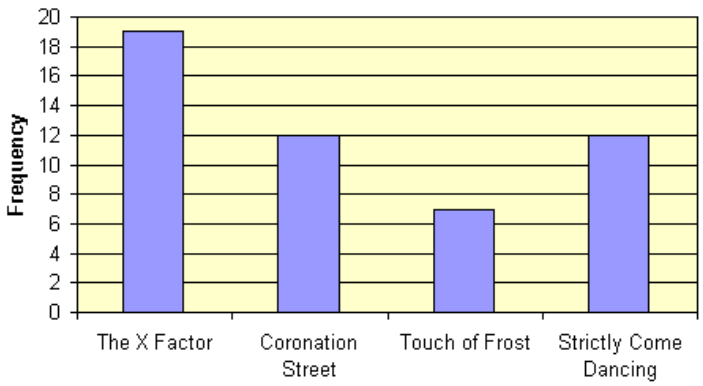
c. The percentage of sales represented by *The Millionaire Next Door* and *Rich Dad, Poor Dad* is 48.33%.

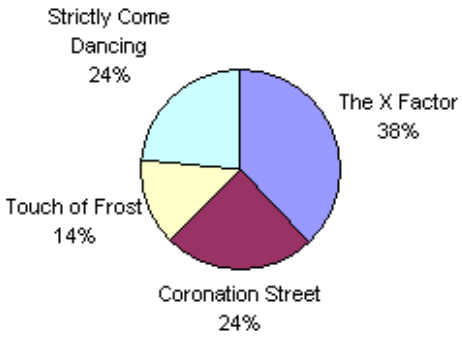
28. a. Qualitative

b.

Show	Frequency	Relative Frequency	Percentage frequency
The X Factor	19	0.38	38
Coronation Street	12	0.24	24
A Touch of Frost	7	0.14	14
Strictly Come Dancing	12	0.24	24
	50	1.00	100%

c.



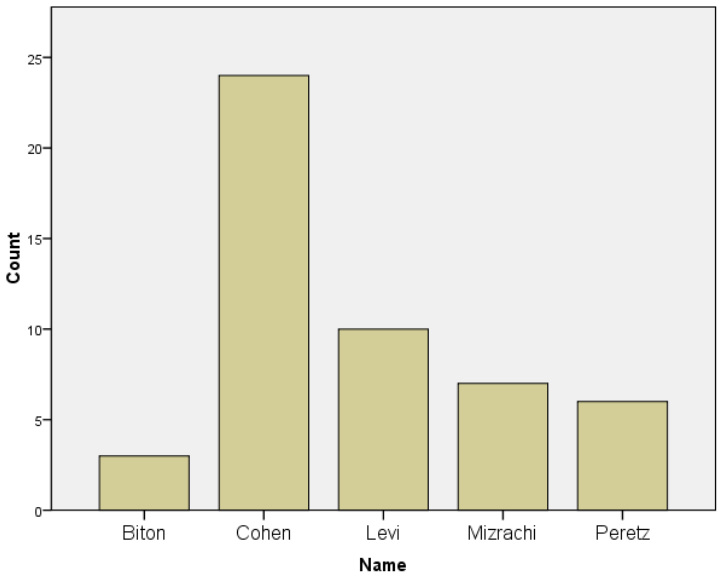


d. *ion Street and Strictly Come Dancing.*

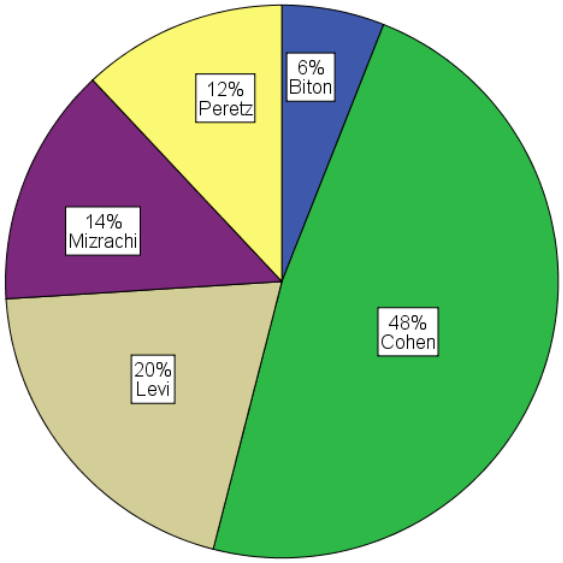
29. a.

Name	Frequency	Relative Frequency	Percentage Frequency
Cohen	24	0.48	48%
Levi	10	0.20	20%
Mizrachi	7	0.14	14%
Peretz	6	0.12	12%
Biton	3	0.06	6%
	50	1.00	100%

b.



c.

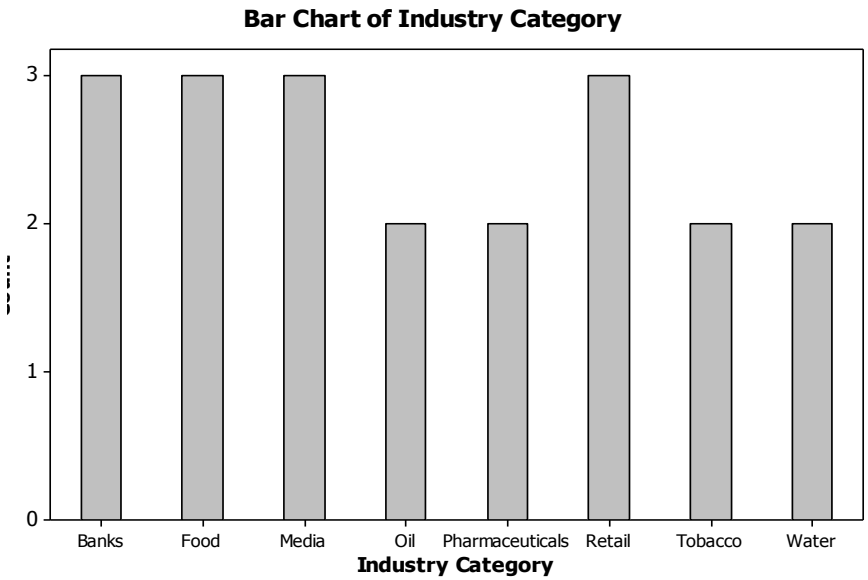


d. Three most common: Cohen, Levi and Mizrachi.

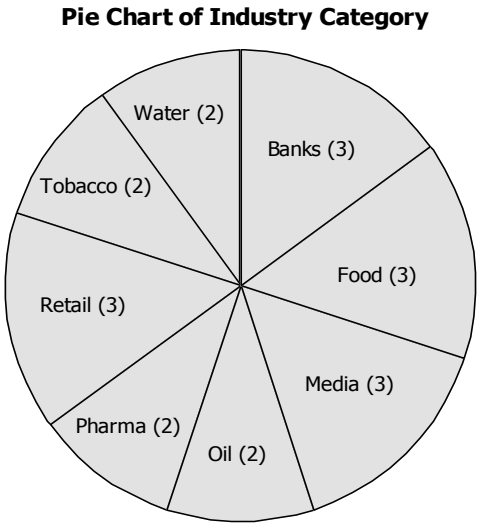
30. a/b.

Industry	Frequency	Percentage Frequency
Banks	3	15
Food	3	15
Media	3	15
Oil	2	10
Pharmaceuticals	2	10
Media	3	15
Retail	2	10
Water	2	10
Total	20	100

c.



d.

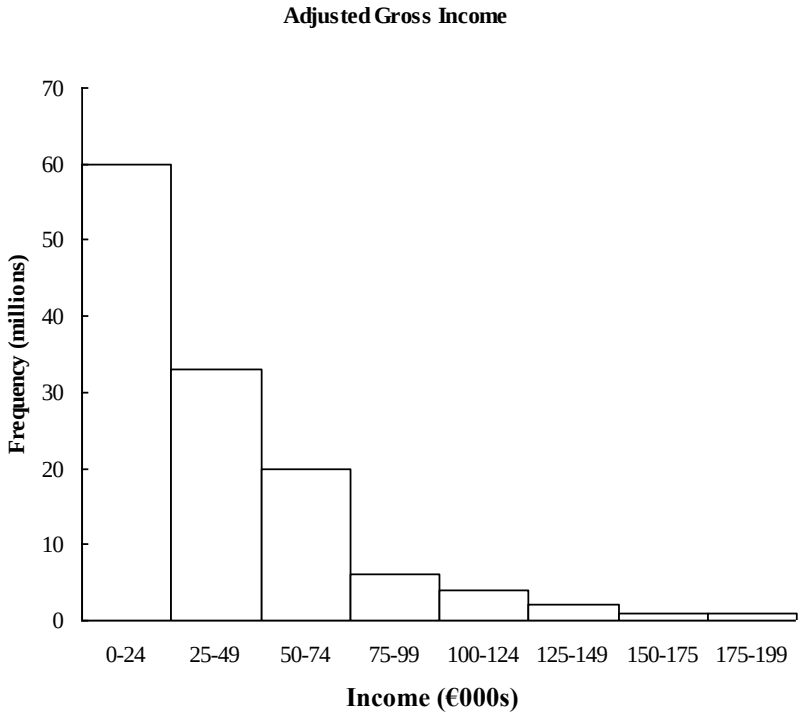


31. a.

Response	Frequency	Percentage Frequency
Accuracy	16	16
Approach Shots	3	3
Mental Approach	17	17
Power	8	8
Practice	15	15
Putting	10	10
Short Game	24	24
Strategic Decisions	7	7
Total	100	100

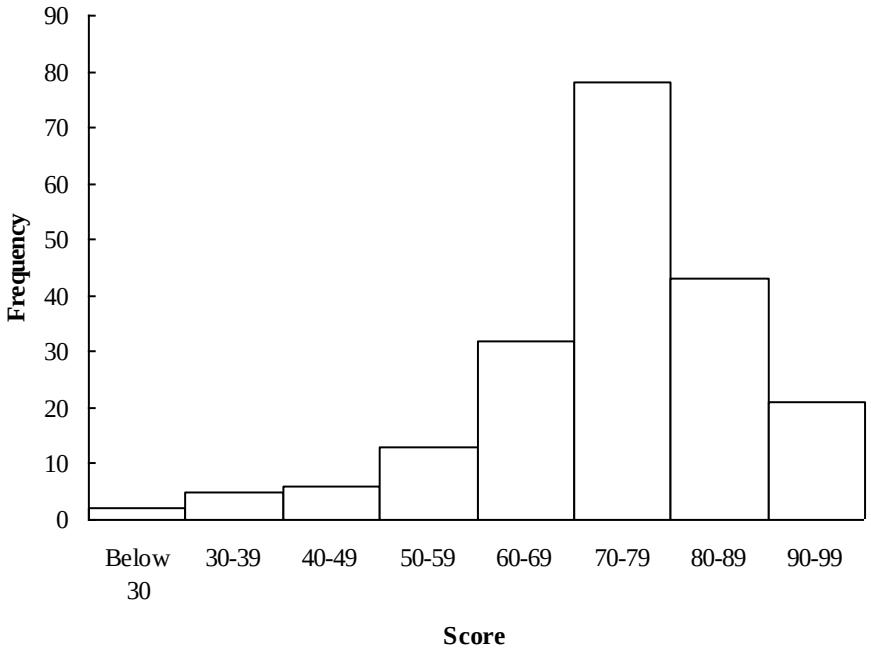
b. Poor short game, poor mental approach, lack of accuracy, and limited practice.

32. a.



The histogram clearly shows that the annual adjusted gross incomes are skewed to the right. And, of course, if annual gross incomes are skewed to the right, so are annual incomes. This makes sense because the vast majority of annual incomes are less than €100,000. But, there are a few individuals with very large incomes.

b.

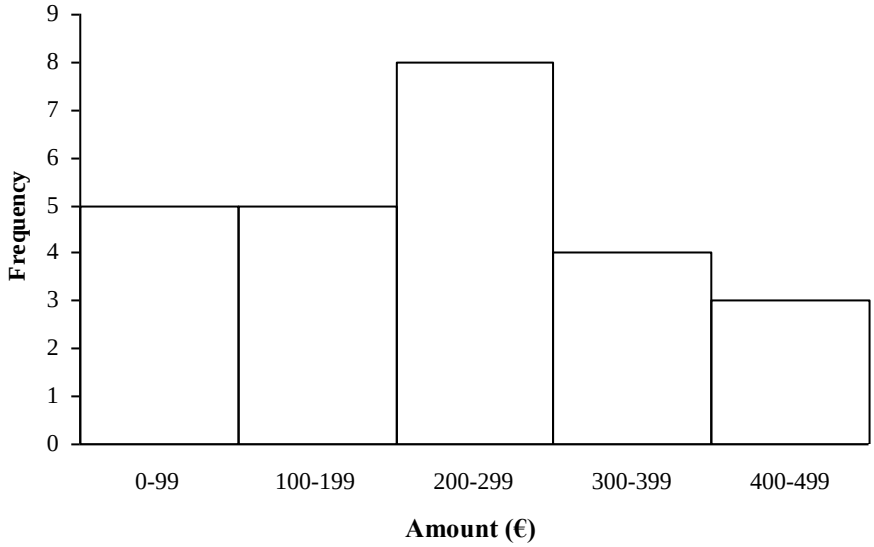


The histogram shows that the distribution of exam scores is skewed to the left. This is to be expected. It is our experience that there are frequently a few very low scores causing such a pattern to appear.

33. a.

Amount	Frequency	Relative Frequency
0-99	5	0.20
100-199	5	0.20
200-299	8	0.32
300-399	4	0.16
400-499	3	0.12

b.



The distribution has a roughly symmetrical shape.

c. The largest group spends €200 €299 per year on books and magazines. There are more in the €0 to €199 range than in the €300 to €499 range.

34. a. Lowest salary: €93,000

Highest salary: €178,000

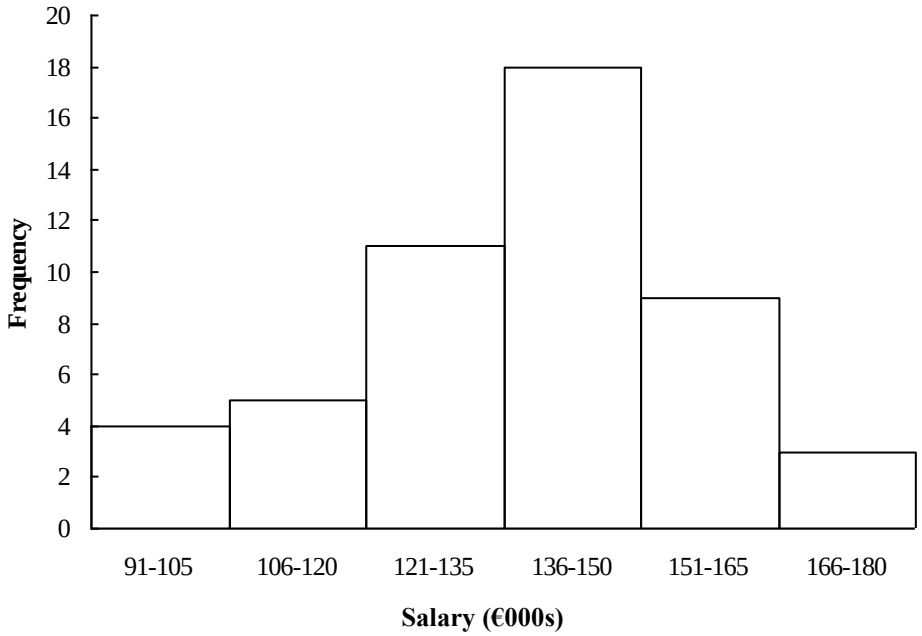
b.

Salary (€000s)	Frequency	Relative Frequency	Percentage Frequency
91-105	4	0.08	8
106-120	5	0.10	10
121-135	11	0.22	22
136-150	18	0.36	36
151-165	9	0.18	18
166-180	3	0.06	6
Total	50	1.00	100

c. Proportion €135,000 or less: 20/50.

d. Percentage more than €150,000: 24%

e.

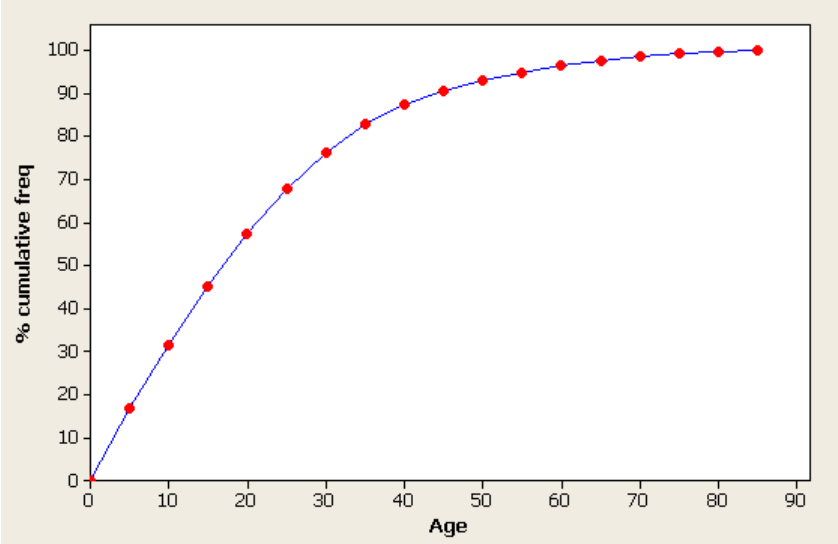


The distribution is skewed slightly to the left.

35. a/b

Age group	Population (000s)	Percentage frequency	Cumulative percentage frequency
0 – 4	2005	16.9	16.9
5 – 9	1749	14.7	31.6
10 – 14	1591	13.4	45.1
15 – 19	1440	12.1	57.2
20 – 24	1253	10.6	67.8
25 – 29	1022	8.6	76.4
30 – 34	770	6.5	82.9
35 – 39	536	4.5	87.4
40 – 44	369	3.1	90.5
45 – 49	288	2.4	92.9
50 – 54	227	1.9	94.8
55 – 59	186	1.6	96.4
60 – 64	146	1.2	97.6
65 – 69	113	1.0	98.6
70 – 74	83	0.7	99.3
75 – 79	50	0.4	99.7
80 +	33	0.3	100.0
Total	11861	100.0	

c.



d. From the ogive, 50% cumulative corresponds to an estimated median of about 17 years.

36.

9	8 9
10	2 4 6 6
11	4 5 7 8 8 9
12	2 4 5 7
13	1 2
14	4
15	1

37. a. 100 shares at $50 per share

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

This stem-and-leaf display shows that the trading prices are closely grouped together. Rotating the stem-and-leaf display counter clockwise shows a histogram that is slightly skewed to the left.

b. 500 shares traded online at $50 per share.

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

This stretched stem-and-leaf display shows that the distribution of online trading prices for most of the brokers for 500 shares are lower than the trading prices for broker assisted trades of 100 shares. There are a couple of outliers. York Securities charges $36 for an online trade and Investors National charges much more than the other brokers: $62.50 for an online trade.

38. a.

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

b.

2000 P/E Forecast	Frequency	Percentage Frequency
5 - 9	2	6.7
10 - 14	6	20.0
15 - 19	6	20.0
20 - 24	6	20.0
25 - 29	2	6.7
30 - 34	0	0.0
35 - 39	4	13.3
40 - 44	1	3.3
45 - 49	2	6.7
50 - 54	0	0.0
55 - 59	0	0.0
60 - 64	1	3.3
Total	30	100.0

39. a.

	Satisfaction Score						
Occupation	30-39	40-49	50-59	60-69	70-79	80-89	Total
Cabinet-maker			2	4	3	1	10
Lawyer	1	5	2	1	1		10
Physical Therapist			5	2	1	2	10
Systems Analyst		2	1	4	3		10
Total	1	7	10	11	8	3	40

b.

Occupation	Satisfaction Score						Total
	30-39	40-49	50-59	60-69	70-79	80-89	
Cabinet-maker			20	40	30	10	100
Lawyer	10	50	20	10	10		100
Physical Therapist			50	20	10	20	100
Systems Analyst		20	10	40	30		100

c. Each row of the percentage cross-tabulation shows a percentage frequency distribution for an occupation. Cabinet-makers seem to have the higher job satisfaction scores while lawyers seem to have the lowest. Fifty per cent of the physical therapists have mediocre scores but the rest are rather high.

40. a. Cross-tabulation for stockholder's equity and profit.

Stockholders' Equity ($000s)	Profits ($000)						Total
	0-200	200-400	400-600	600-800	800-1000	1000-1200	
0-1200	10	1				1	12
1200-2400	4	10			2		16
2400-3600	4	3	3	1	1	1	13
3600-4800					1	2	3
4800-6000		2	3	1			6
Total	18	16	6	2	4	4	50

b. Row Percentages

Stockholders' Equity ($000s)	Profits ($000)						Total
	0-200	200-400	400-600	600-800	800-1000	1000-1200	
0-1200	83.33	8.33	0.00	0.00	0.00	8.33	100
1200-2400	25.00	62.50	0.00	0.00	12.50	0.00	100
2400-3600	30.77	23.08	23.08	7.69	7.69	7.69	100
3600-4800		0.00	0.00	0.00	33.33	66.67	100
4800-6000	0.00	33.33	50.00	16.67	0.00	0.00	100

c. Stockholder's equity and profit seem to be related. As profit goes up, stockholder's equity goes up. The relationship, however, is not very strong.

41. a. Cross-tabulation of market value and profit.

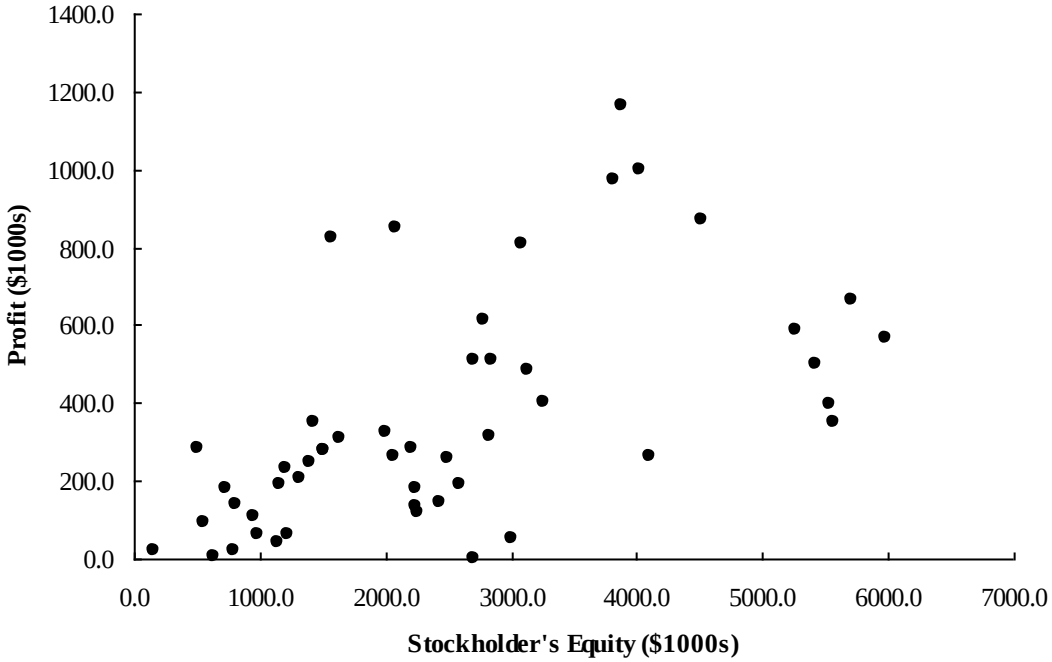
Market Value ($000s)	Profit ($1000s)				Total
	0-300	300-600	600-900	900-1200	
0-8000	23	4			27
8000-16000	4	4	2	2	12
16000-24000		2	1	1	4
24000-32000		1	2	1	4
32000-40000		2	1		3
Total	27	13	6	4	50

b. Row Percentages.

Market Value ($000s)	Profit ($1000s) 0-300	300-600	600-900	900-1200	Total
0-8000	85.19	14.81	0.00	0.00	100
8000-16000	33.33	33.33	16.67	16.67	100
16000-24000	0.00	50.00	25.00	25.00	100
24000-32000	0.00	25.00	50.00	25.00	100
32000-40000	0.00	66.67	33.33	0.00	100

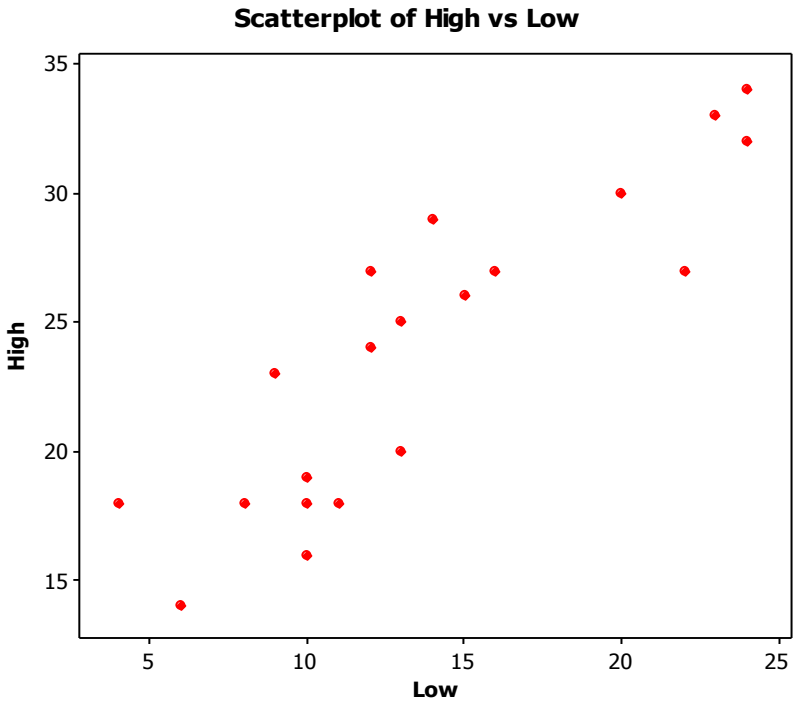
c. There appears to be a positive relationship between Profit and Market Value. As profit goes up, Market Value goes up.

42. a. Scatter diagram of Profit vs. Stockholder's Equity.



b. Profit and Stockholder's Equity appear to be positively related.

43. a.



b. There is clearly a positive relationship between high and low temperature for cities. As one goes up so does the other.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Three

Descriptive Statistics – Numerical Measures

Textbook Exercises (1-39)

Textbook Exercise Solutions

Supplementary Exercises (40-58)

Supplementary Exercise Solutions

Chapter 3: Descriptive Statistics – Numerical Measures

Textbook Exercises:

1 Consider a sample with data values of 10, 20, 12, 17 and 16. Compute the mean and median.

2 Consider a sample with data values of 10, 20, 21, 17, 16 and 12. Compute the mean and median.

3 Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28 and 25. Compute the 20th, 25th, 65th and 75th percentiles.

4 Consider a sample with data values of 53, 55, 70, 58, 64, 57, 53, 69, 57, 68 and 53. Compute the mean, median and mode.

5 A sample of 30 Irish engineering graduates had the following starting salaries. Data are in thousands of euros.

36.8	34.9	35.2	37.2	36.2	35.8	36.8	36.1	36.7	36.6
37.3	38.2	36.3	36.4	39.0	38.3	36.0	35.0	36.7	37.9
38.3	36.4	36.5	38.4	39.4	38.8	35.4	36.4	37.0	36.4

a. What is the mean starting salary?

b. What is the median starting salary?

c. What is the mode?

d. What is the first quartile?

e. What is the third quartile?

6 The following data were obtained for the number of minutes spent listening to recorded music for a sample of 30 individuals on one particular day.

88.3	4.3	4.6	7.0	9.2	0.0	99.2	34.9	81.7	0.0
85.4	0.0	17.5	45.0	53.3	29.1	28.8	0.0	98.9	64.5
4.4	67.9	94.2	7.6	56.6	52.9	145.6	70.4	65.1	63.6

a. Compute the mean.

b. Compute the median.

c. Compute the first and third quartiles.

d. Compute and interpret the 40th percentile.

7 miniRank (www.minirank.com) rates the popularity of websites in most countries of the world, using a points system. The 25 most popular sites in South Africa as listed in July 2012 were as follows (the points scores have been rounded to one decimal place):

Website	Points	Website	Points
www.intoweb.co.za	253.1	www.dweb.co.za	118.2
www.weathersa.co.za	252.3	dweb.co.za	108.5
www.etraffic.co.za	212.4	www.webworx.org.za	107.6
www.gov.za	167.0	www.bacchus.co.za	105.2
www.intowebtraining.co.za	164.6	www.services.gov.za	103.3
www.capewebdesign.co.za	161.7	www.info.gov.za	102.2
www.saweather.co.za,	153.3	www.sars.co.za	95.6
www.web-inn.co.za	136.8	www.sars.gov.za	93.8
www.searchengine.co.za	136.1	www.mwebbusiness.co.za	93.6
saweather.co.za	133.6	www.dti.gov.za,	84.0
www.iol.co.za	132.5	www.jdconsulting.co.za	82.2
www.tradepage.co.za	128.6	www.linx.co.za	81.0
www.proudlysa.co.za	126.9		

a. Compute the mean and median.

b. Do you think it would be better to use the mean or the median as the measure of central location for these data? Explain.

c. Compute the first and third quartiles.

d. Compute and interpret the 85th percentile.

8 Following is a sample of age data for individuals working from home by 'telecommuting'.

18 54 20 46 25 48 53 27 26 37

40 36 42 25 27 33 28 40 45 25

a. Compute the mean and the mode.

b. Suppose the median age of the population of all adults is 35.5 years. Use the median age of the preceding data to comment on whether the at-home workers tend to be younger or older than the population of all adults.

c. Compute the first and third quartiles.

d. Compute and interpret the 32nd percentile.

9 Consider a sample with data values of 10, 20, 12, 17 and 16. Calculate the range and interquartile range.

10 Consider a sample with data values of 10, 20, 12, 17 and 16. Calculate the variance and standard deviation.

11 Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28 and 25. Calculate the range, interquartile range, variance and standard deviation.

12 The number of goals scores in six handball matches was 41, 34, 42, 45, 35 and 37. Using these data as a sample, compute the following descriptive statistics.

a. Range.

b. Variance.

c. Standard deviation.

d. Coefficient of variation.

13 Dinner bill amounts for set menus at a Dubai restaurant, Al Khayam, show the following frequency distribution. The amounts are in AED (Emirati dirham). Compute the mean, variance and standard deviation.

Dinner bill (AED)	Frequency
30	2
40	6
50	4
60	4
70	2
80	2
Total	20

14 The following data were used to construct the histograms of the number of days required to fill orders for Dawson Supply and for J.C. Clark Distributors (see Figure 3.2).

Dawson Supply days for delivery: 11 10 9 10 11 11 10 11 10 10

Clark Distributors days for delivery: 8 10 13 7 10 11 10 7 15 12

Use the range and standard deviation to support the previous observation that Dawson Supply provides the more consistent and reliable delivery times.

15 Police records show the following numbers of daily crime reports for a sample of days during the winter months and a sample of days during the summer months.

Winter: 18 20 15 16 21 20 12 16 19 20

Summer: 28 18 24 32 18 29 23 38 28 18

a. Compute the range and interquartile range for each period.

b. Compute the variance and standard deviation for each period.

c. Compute the coefficient of variation for each period.

d. Compare the variability of the two periods.

16 A production department uses a sampling procedure to test the quality of newly produced items. The department employs the following decision rule at an inspection station: if a sample of 14 items has a variance of more than 0.005, the production line must be shut down for repairs. Suppose the following data have just been collected:

3.43 3.45 3.43 3.48 3.52 3.50 3.39

3.48 3.41 3.38 3.49 3.45 3.51 3.50

Should the production line be shut down? Why or why not?

17 Consider a sample with data values of 10, 20, 12, 17 and 16. Calculate the z-score for each of the five observations.

18 Consider a sample with a mean of 500 and a standard deviation of 100. What are the z-scores for the following data values: 520, 650, 500, 450 and 280?

19 Consider a sample with a mean of 30 and a standard deviation of 5. Use Chebyshev's theorem to determine the percentage of the data within each of the following ranges.

a. 20 to 40

b. 15 to 45

c. 22 to 38

d. 18 to 42

e. 12 to 48

20 Suppose the data have a bell-shaped distribution with a mean of 30 and a standard deviation of 5. Use the empirical rule to determine the percentage of data within each of the following ranges.

a. 20 to 40

b. 15 to 45

c. 25 to 35

21 The results of a survey of 1154 adults showed that on average, adults sleep 6.9 hours per day during the working week. Suppose that the standard deviation is 1.2 hours.

a. Use Chebyshev's theorem to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours per day.

b. Use Chebyshev's theorem to calculate the percentage of individuals who sleep between 3.9 and 9.9 hours per day.

c. Assume that the number of hours of sleep follows a bell-shaped distribution. Use the empirical rule to calculate the percentage of individuals who sleep between

4.5 and 9.3 hours per day. How does this result compare to the value that you obtained using Chebyshev's theorem in part (a)?

22 Suppose that IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15.

a. What percentage of people have an IQ score between 85 and 115?

b. What percentage of people have an IQ score between 70 and 130?

c. What percentage of people have an IQ score of more than 130?

d. A person with an IQ score greater than 145 is considered a genius. Does the empirical rule support this statement? Explain.

23 Suppose the average charge for a 7-day hire of an economy-class car in Kuwait City is KWD (Kuwaiti Dinar) 75.00, and the standard deviation is KWD 20.00.

a. What is the z-score for a car service with an hourly labour cost of KWD 56.00?

b. What is the z-score for a car service with an hourly labour cost of KWD 153.00?

c. Interpret the z-scores in parts (a) and (b). Comment on whether either should be considered an outlier.

24 Consumer Review posts reviews and ratings of a variety of products on the Internet. The following is a sample of 20 speaker systems and their ratings, on a scale of 1 to 5, with 5 being best.

a. Compute the mean and the median.

b. Compute the first and third quartiles.

c. Compute the standard deviation.

d. The skewness of this data is 1.67. Comment on the shape of the distribution.

e. What are the z-scores associated with Allison One and Omni Audio?

f. Do the data contain any outliers? Explain.

Speaker	Rating	Speaker	Rating
Infinity Kappa 6.1	4.00	ACI Sapphire III	4.67
Allison One	4.12	Bose 501 Series	2.14
Cambridge Ensemble II	3.82	DCM KX-212	4.09
Dynaudio Contour 1.3	4.00	Eosone RSF1000	4.17
Hsu Rsch. HRSW12V	4.56	Joseph Audio RM7si	4.88
Legacy Audio Focus	4.32	Martin Logan Aerius	4.26
26 Mission 73li	4.33	Omni Audio SA 12.3	2.32
PSB 400i	4.50	Polk Audio RT12	4.50
Snell Acoustics D IV	4.64	Sunfire True Subwoofer	4.17
Thiel CS1.5	4.20	Yamaha NS-A636	2.17

25 Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28 and 25. Provide the five-number summary for the data.

26 Construct a box plot for the data in Exercise 25.

27 Prepare the five-number summary and the box plot for the following data: 5, 15, 18, 10, 8, 12, 16, 10, 6.

28 A data set has a first quartile of 42 and a third quartile of 50. Compute the lower and upper limits for the corresponding box plot. Should a data value of 65 be considered an outlier?

29 Annual sales, in millions of dollars, for 21 pharmaceutical companies follow.

8408 1374 1872 8879 2459 11413 608

14138 6452 1850 2818 1356 10498 7478

4019 4341 739 2127 3653 5794 8305

a. Provide a five-number summary.

b. Compute the lower and upper limits (for the box plot).

c. Do the data contain any outliers?

d. Johnson & Johnson's sales are the largest on the list at $14 138 million. Suppose a data entry error (a transposition) had been made and the sales had been entered as $41 138 million. Would the method of detecting outliers in part (c) identify this problem and allow for correction of the data entry error?

d. Construct a box plot.

30 A goal of management is to help their company earn as much as possible relative to the capital invested. One measure of success is return on equity – the ratio of net income to stockholders' equity. Return on equity percentages are shown here for 25 companies.

9.0 19.6 22.9 41.6 11.4 15.8 52.7 17.3 12.3 5.1

17.3 31.1 9.6 8.6 11.2 12.8 12.2 14.5 9.2 16.6

5.0 30.3 1 4.7 19.2 6.2

a. Provide a five-number summary.

b. Compute the lower and upper limits (for the box plot).

c. Do the data contain any outliers? How would this information be helpful to a financial analyst?

d. Construct a box plot.

31 In 2008, stock markets around the world lost value. The website www.owneverystock.com listed the following percentage falls in stock market indices between the start of the year and the beginning of October.

Country	% Fall	Country	% Fall
New Zealand	27.05	Brazil	39.59
Canada	27.30	Japan	39.88
Switzerland	28.42	Sweden	40.35
Mexico	29.99	Egypt	41.57
Australia	31.95	Singapore	41.60
Korea	32.18	Italy	42.88
United Kingdom	32.37	Belgium	43.70
Spain	32.69	India	44.16
Malaysia	32.86	Hong Kong	44.52
Argentina	36.83	Netherlands	44.61
France	37.71	Norway	46.98
Israel	37.84	Indonesia	47.13
Germany	37.85	Austria	50.06
Taiwan	38.79	China	60.24

a. What are the mean and median percentage changes for these countries?

b. What are the first and third quartiles?

c. Do the data contain any outliers? Construct a box plot.

d. What percentile would you report for Belgium?

32 Five observations taken for two variables follow.

x_i	4	6	11	3	16
y_i	50	50	40	60	30

a. Construct a scatter diagram with the xi values on the horizontal axis.

b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

c. Compute and interpret the sample covariance.

d. Compute and interpret the sample correlation coefficient.

33 Five observations taken for two variables follow.

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a. Construct a scatter diagram for these data.

b. What does the scatter diagram indicate about a relationship between *X* and *Y*?

c. Compute and interpret the sample covariance.

d. Compute and interpret the sample correlation coefficient.

34 Below are return on investment figures (%) and current ratios (current assets/current liabilities) for 15 German companies, for the year 2011.

Company	Return on investment (%)	Current ratio
Adidas	8.15	1.50
BASF	14.66	1.64
Bayer	6.37	1.50
BMW	5.98	1.04
Continental	7.15	1.06
Daimler	5.70	1.11
Deutsche Bank	0.25	0.82
Deutsche Telekom	2.46	0.65
Fresenius	9.10	1.34
Henkel	9.16	1.58
Linde	5.60	0.89
SAP	20.53	1.54
Siemens	8.87	1.21
Tui	1.53	0.65
Volkswagen	7.46	1.05

a. Construct a scatter diagram with current ratio on the horizontal axis.

b. Is there any relationship between return on investment and current ratio? Explain.

c. Compute and interpret the sample covariance.

d. Compute and interpret the sample correlation coefficient.

e. What does the sample correlation coefficient tell you about the relationship between return on investment and current ratio?

35 Stock markets across the Eurozone tend to have mutual influences on each other. The index levels of the German DAX index and the French CAC 40 index for 10 weeks beginning with 4 June 2012 are shown below (file 'DAX_CAC' on the online platform).

Date	DAX	CAC 40
04-Jun-12	6,130.82	3,051.69
11-Jun-12	6,229.41	3,087.62
18-Jun-12	6,263.25	3,090.90
25-Jun-12	6,416.28	3,196.65
02-Jul-12	6,410.11	3,168.79
09-Jul-12	6,557.10	3,180.81
16-Jul-12	6,630.02	3,193.89
23-Jul-12	6,689.40	3,280.19
30-Jul-12	6,865.66	3,374.19
06-Aug-12	6,967.95	3,453.28

a. Compute the sample correlation coefficient for these data.

b. Are they poorly correlated, or do they have a close association?

36 Consider the following data and corresponding weights.

x_i	Weight
3.2	6
2.0	3
2.5	2
5.0	8

a. Compute the weighted mean.

b. Compute the sample mean of the four data values without weighting. Note the difference in the results provided by the two computations.

37 Consider the sample data in the following frequency distribution.

Class	Midpoint	Frequency
3-7	5	4
8-12	10	7
13-17	15	9
18-22	20	5

a. Compute the sample mean.

b. Compute the sample variance and sample standard deviation.

38 The assessment for a statistics module comprises a multiple-choice test, a data analysis project, an Excel test, and a written examination. Scores for Jil and Ricardo on the four components are show below.

Assessment	**Jil**	**Ricardo**
Multiple-choice test	80%	48%
Data analysis project	60%	78%
Excel test	62%	60%
Written examination	57%	53%

a. Calculate weighted mean scores (%) for Jil and Ricardo assuming the respective weightings for the four components are 20, 20, 30 , 30.

b. Calculate weighted mean scores (%) for Jil and Ricardo assuming the respective weightings for the four components are 10, 25, 15, 50.

39 A petrol station recorded the following frequency distribution for the number of litres of petrol sold per car in a sample of 680 cars.

Petrol (litres)	Frequency
1–15	74
16–30	192
31–45	280
46–60	105
61–75	23
76–90	6
Total	**680**

Compute the mean, variance and standard deviation for these grouped data. If the petrol station expects to serve petrol to about 120 cars on a given day, estimate the total number of litres of petrol that will be sold.

Chapter 3

Descriptive Statistics: Numerical Methods

Textbook Exercise Solutions

1. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{75}{5} = 15$

   10, 12, 16, 17, 20

   Median = 16 (middle value)

2. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{96}{6} = 16$

   10, 12, 16, 17, 20, 21

   Median = $\frac{16+17}{2} = 16.5$

3. 15, 20, 25, 25, 27, 28, 30, 34

   $i = \frac{20}{100}(8) = 1.6$ 2nd position = 20

   $i = \frac{25}{100}(8) = 2$ $\frac{20+25}{2} = 22.5$

   $i = \frac{65}{100}(8) = 5.2$ 6th position = 28

   $i = \frac{75}{100}(8) = 6$ $\frac{28+30}{2} = 29$

4. Mean = $\frac{53+55+70+58+64+57+53+69+57+68+53}{11} = 59.7$

Ordered data set: 53 53 53 55 57 57 58 64 68 69 70

Median is the 6th ordered value = 57

Mode is the most frequent value = 53 (occurs 3 times)

5. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{1106.4}{30} = 36.88$

b. There is an even number of items. Hence, the median is the average of the 15th and 16th items after the data have been placed in rank order.

Median = $\frac{36.6+36.7}{2} = 36.65$

c. Mode = 36.4 This value appears 4 times

d. First Quartile

$$i = \left(\frac{25}{100}\right)30 = 7.5$$

Rounding up, we see that Q_1 is at the 8th position. Q_1 = 36.2

e. Third Quartile

$$i = \left(\frac{75}{100}\right)30 = 22.5$$

Rounding up, we see that Q_3 is at the 23rd position. Q_3 = 37.9

6. a. Mean = $\frac{88.3 + 4.3 + K + 63.6}{30} = 46.0$

b. Ordered data set is: 0.0 0.0 0.0 0.0 4.3 4.4 4.6 7.0 7.6 9.2 17.5 28.8 29.1 34.9 45.0 52.9 53.3 56.6 63.6 64.5 65.1 67.9 70.4 81.7 85.4 88.3 94.2 98.9 99.2 145.6

Median is midway between the 15th and 16th ordered values = $\frac{45.0 + 52.9}{2} = 48.95$

c. Index for lower quartile is

$i = \left(\frac{25}{100}\right)30 = 7.5$, rounded up to 8. Lower quartile = 7.0

Index for upper quartile is $i = \left(\frac{75}{100}\right)30 = 22.5$, rounded up to 23. Upper quartile = 70.4

d. Index for 40th percentile is

$i = \left(\frac{40}{100}\right)30 = 12$, so 40th percentile is midway between the 12th and 13th ordered values = $\frac{28.8 + 29.1}{2} = 28.95$. This implies that at least 40% of values are less than or equal 28.95, and at least 60% are greater than or equal to 29.95.

7. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{3334.1}{25} = 133.4$

Median is the 13th ordered data value = 126.9

b. The data are somewhat positively skewed, with two high outliers. The median would therefore be a better measure of central location.

c. For Q_1, $i = \left(\frac{25}{100}\right)25 = 6.25$

7^{th} ordered data value is $Q_1 = 102.2$

For Q_3, $i = \left(\frac{75}{100}\right)25 = 18.75$

19^{th} ordered data value is $Q_3 = 153.3$

d. $i = \left(\frac{85}{100}\right)25 = 21.25$

Since i is not an integer, we round up to the 22^{nd} position.

85^{th} percentile = 167.0

Approximately 85% of the sites have points scores below 167.0.

8. a. $\Sigma x_i = 695$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{695}{20} = 34.75$$

The modal age is 29; it appears 3 times.

b. Median is average of 10th and 11th items.

$$\text{Median} = \frac{33+36}{2} = 34.5$$

Data suggest at-home workers are slightly younger.

c. For Q_1, $i = \left(\frac{25}{100}\right)20 = 5$

Since i is integer, $Q_1 = \frac{25+26}{2} = 25.5$

For Q_3, $i = \left(\frac{75}{100}\right)20 = 15$

Since i is integer, $Q_3 = \frac{42+45}{2} = 43.5$

d. $i = \left(\frac{32}{100}\right)20 = 6.4$

Since i is not an integer, we round up to the 7th position.

32nd percentile = 27

Approximately 32% are aged under 27.

9. Range 20 – 10 = 10

10, 12, 16, 17, 20

$i = \frac{25}{100}(5) = 1.25$ Q_1 (2nd position) = 12

$i = \frac{75}{100}(5) = 3.75$ Q_3 (4th position) = 17

IQR = $Q_3 - Q_1$ = 17 – 12 = 5

10. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{10 + 20 + 12 + 17 + 16}{5} = \frac{75}{5} = 15$

$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{(-5)^2 + (5)^2 + (-3)^2 + (2)^2 + (1)^2}{4} = \frac{64}{4} = 16$

$s = \sqrt{16} = 4$

11. 15, 20, 25, 25, 27, 28, 30, 34 Range = 34 – 15 = 19

$i = \frac{25}{100}(8) = 2$ $Q_1 = \frac{20 + 25}{2} = 22.5$

$i = \frac{75}{100}(8) = 6$ $Q_3 = \frac{28 + 30}{2} = 29$

IQR = $Q_3 - Q_1$ = 29 – 22.5 = 6.5

$\bar{x} = \frac{\Sigma x_i}{n} = \frac{204}{8} = 25.5$

$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{242}{7} = 34.57$

$s = \sqrt{34.57} = 5.88$

12. a. Range = Maximum – minimum = 45 – 34 = 11

b. Mean = $\frac{\Sigma x_i}{n} = \frac{41 + 34 + 42 + 45 + 35 + 37}{6} = \frac{234}{6} = 39$

Variance = $\frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{(2)^2 + (-5)^2 + (3)^2 + (6)^2 + (4)^2 + (-2)^2}{5} = \frac{94}{5} = 18.8$

c. Standard deviation = $\sqrt{18.8} = 4.34$

d. Mean = $\frac{\sum x_i}{n} = \frac{234}{6} = 39$

Coefficient of variation = $100 \times \frac{4.34}{39} = 11.1\%$

13.

f_i	x_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
2	30	60	-22	484	968
6	40	240	-12	144	864
4	50	200	-2	4	16
4	60	240	8	64	256
2	70	140	18	324	648
2	80	160	28	784	1568
20		1,040			4320

$$\bar{x} = \frac{1040}{20} = 52$$

$$s^2 = \frac{4320}{19} = 227.37 \qquad s = 15.08$$

14. Dawson Supply: Range = 11 – 9 = 2 $\quad s = \sqrt{\frac{4.1}{9}} = 0.67$

J.C. Clark: Range = 15 – 7 = 8 $\quad s = \sqrt{\frac{60.1}{9}} = 2.58$

15. a. Winter

Range = 21 – 12 = 9

IQR = $Q_3 - Q_1$ = 20 – 16 = 4

Summer

Range = 38 – 18 = 20

IQR = $Q_3 - Q_1$ = 29 – 18 = 11

b.

	Variance	Standard Deviation
Winter	8.2333	2.8694
Summer	44.4889	6.6700

c. Winter

Coefficient of Variation = $\left(\frac{s}{\bar{x}}\right)100\% = \left(\frac{2.8694}{17.7}\right)100\% = 16.21\%$

Summer

Coefficient of Variation = $\left(\frac{s}{\bar{x}}\right)100\% = \left(\frac{6.6700}{25.6}\right)100\% = 26.05\%$

d. More variability in the summer months.

16. $s^2 = 0.0021$ Production should not be shut down since the variance is less than 0.005.

17. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{75}{5} = 15$

$s^2 = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{64}{4}} = 4$

10 $z = \frac{10-15}{4} = -1.25$ 20 $z = \frac{20-15}{4} = +1.25$

12 $z = \frac{12-15}{4} = -0.75$ 17 $z = \frac{17-15}{4} = +0.50$

16 $z = \frac{16-15}{4} = +0.25$

18. 520 $z = \frac{520-500}{100} = +0.20$ 650 $z = \frac{650-500}{100} = +1.50$

500 $z = \frac{500-500}{100} = 0.00$ 450 $z = \frac{450-500}{100} = -0.50$

280 $z = \frac{280-500}{100} = -2.20$

19. a. $z = \frac{40-30}{5} = 2$ $1 - \frac{1}{2^2} = 0.75$ At least 75%

b. $z = \frac{45-30}{5} = 3$ $1 - \frac{1}{3^2} = 0.89$ At least 89%

c. $z = \frac{38-30}{5} = 1.6$ $1 - \frac{1}{1.6^2} = 0.61$ At least 61%

d. $z = \frac{42-30}{5} = 2.4$ $1-\frac{1}{2.4^2} = 0.83$ At least 83%

e. $z = \frac{48-30}{5} = 3.6$ $1-\frac{1}{3.6^2} = 0.92$ At least 92%

20. a. Approximately 95%. (20 to 40 is 2 standard deviations either side of the mean.)

b. Almost all. (15 to 45 is 3 standard deviations either side of the mean.)

c. Approximately 68%. (25 to 35 is 1 standard deviation either side of the mean.)

21. a. This is from 2 standard deviations below the mean to 2 standard deviations above the mean.

With $z = 2$, Chebyshev's theorem gives:

$$1-\frac{1}{z^2} = 1-\frac{1}{2^2} = 1-\frac{1}{4} = \frac{3}{4}$$

Therefore, at least 75% of adults sleep between 4.5 and 9.3 hours per day.

b. This is from 2.5 standard deviations below the mean to 2.5 standard deviations above the mean.

With $z = 2.5$, Chebyshev's theorem gives:

$$1-\frac{1}{z^2}=1-\frac{1}{2.5^2}=1-\frac{1}{6.25}=0.84$$

Therefore, at least 84% of adults sleep between 3.9 and 9.9 hours per day.

c. With $z = 2$, the empirical rule suggests that 95% of adults sleep between 4.5and 9.3 hours per day. The percentage obtained using the empirical rule is greater than the percentage obtained using Chebyshev's theorem.

22 a. IQ scores of 85 and 115 are respectively 1 standard below and above the mean. Using the empirical rule, approximately 68% of scores are within 1 standard deviation from the mean.

b. Similarly, approximately 95% of scores are within 2 standard deviations from the mean.

c. Approximately (100% – 95%) / 2 = 2.5% of scores are over 130.

d. An IQ score of 45 is 3 standard deviations above the mean. Yes, the empirical rule almost all IQ scores are less than 145.

23. a. $z = \frac{56 - 75}{20} = -0.95$

b. $z = \frac{153 - 75}{20} = 3.9$

c. ZAR 56 is 0.95 standard deviations below the mean. ZAR 153 is 3.9 standard deviations above the mean. The latter would be considered an outlier.

24. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{79.86}{20} = 3.99$

Median = $\frac{4.17 + 4.20}{2} = 4.185$ (average of 10th and 11th values)

b. $Q_1 = 4.00$ (average of 5th and 6th values)

$Q_3 = 4.50$ (average of 15th and 16th values)

c. $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{12.5080}{19}} = 0.8114$

d. The distribution is significantly skewed to the left (negative skewness).

e. Allison One: $z = \frac{4.12 - 3.99}{0.8114} \approx 0.16$

Omni Audio SA 12.3: $z = \frac{2.32 - 3.99}{0.8114} \approx -2.06$

f. The lowest rating is for the Bose 501 Series. Its *z*-score is:

$$z = \frac{2.14 - 3.99}{0.8114} \approx -2.28.$$

This is not an outlier, so there are no outliers.

25. 15, 20, 25, 25, 27, 28, 30, 34

Smallest = 15

$$i = \frac{25}{100}(8) = 2 \qquad Q_1 = \frac{20 + 25}{2} = 22.5$$

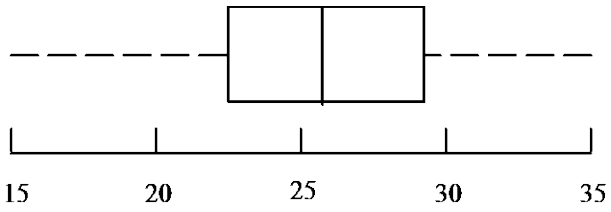
$$\text{Median} = \frac{25 + 27}{2} = 26$$

$$i = \frac{75}{100}(8) = 8 \qquad Q_3 = \frac{28 + 30}{2} = 29$$

Largest = 34

Smallest = 15, Q_1 = 22.5, Median = 26, Q_3 = 29, Largest = 34

26.



27. 5, 6, 8, 10, 10, 12, 15, 16, 18

Smallest = 5

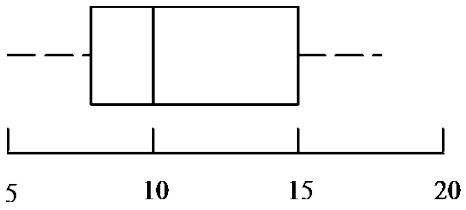
$i = \frac{25}{100}(9) = 2.25$ Q_1 = 8 (3rd position)

Median = 10

$i = \frac{75}{100}(9) = 6.75$ Q_3 = 15 (7th position)

Largest = 18

Smallest = 5, Q_1 = 8, Median = 10, Q_3 = 15, Largest = 18



28. IQR = 50 – 42 = 8

Lower Limit: $Q_1 - 1.5$ IQR $= 42 - 12 = 30$

Upper Limit: $Q_3 + 1.5$ IQR $= 50 + 12 = 62$

65 is an outlier

29. a. Median (11th position) 4019

$i = \frac{25}{100}(21) = 5.25$ Q_1 (6th position) = 1872

$i = \frac{75}{100}(21) = 15.75$ Q_3 (16th position) = 8305

Five number summary: 608, 1872, 4019, 8305, 14138

b. Limits:

IQR $= Q_3 - Q_1 = 8305 - 1872 = 6433$

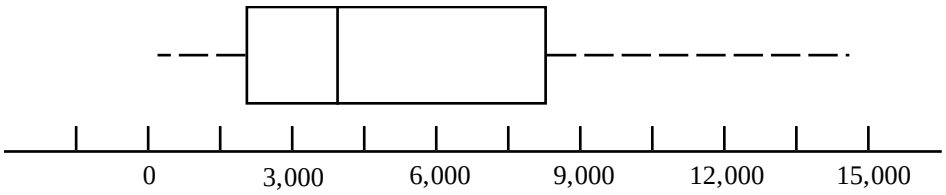
Lower Limit: $Q_1 - 1.5$ (IQR) $= -7777$

Upper Limit: $Q_3 + 1.5$ (IQR) $= 17955$

c. There are no outliers. All data are within the limits.

d. Yes, if the first two digits in Johnson & Johnson's sales were transposed to 41,138, sales would have shown up as an outlier. A review of the data would have enabled the correction of the data.

e.



30. a. Ordered data set is: 5.0 5.1 6.2 8.6 9.0 9.2 9.6 11.2 11.4 12.2 12.3 12.8 14.5 14.7 15.8 16.6 17.3 17.3 19.2 19.6 22.9 30.3 31.1 41.6 52.7

Five number summary: 5 (min) 9.6 (7^{th}) 14.5 (13^{th}) 19.2 (19^{th}) 52.7 (max)

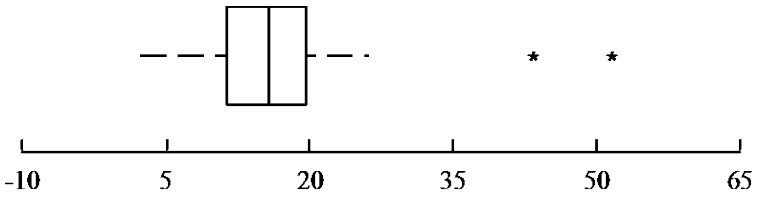
b. IQR = $Q_3 - Q_1$ = 19.2 – 9.6 = 9.6

Lower Limit: $Q_1 - 1.5$ (IQR) = 9.6 – 1.5(9.6) = –4.8

Upper Limit: $Q_3 + 1.5$(IQR) = 19.2 + 1.5(9.6) = 33.6

c. The data value 41.6 is an outlier (larger than the upper limit) and so is the data value 52.7. The financial analyst should first verify that these values are correct. Perhaps a typing error has caused 25.7 to be typed as 52.7 (or 14.6 to be typed as 41.6). If the outliers are correct, the analyst might consider these companies with an unusually large return on equity as good investment candidates.

d.



31. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{1091.10}{28} = 38.97\%$

Median = $\frac{38.79 + 39.59}{2} = 39.19\%$

b. $i = \frac{25}{100}(28) = 7$ $Q_1 = \frac{32.37 + 32.69}{2} = 32.53\%$

$i = \frac{75}{100}(28) = 21$ $Q_3 = \frac{43.70 + 44.16}{2} = 43.93\%$

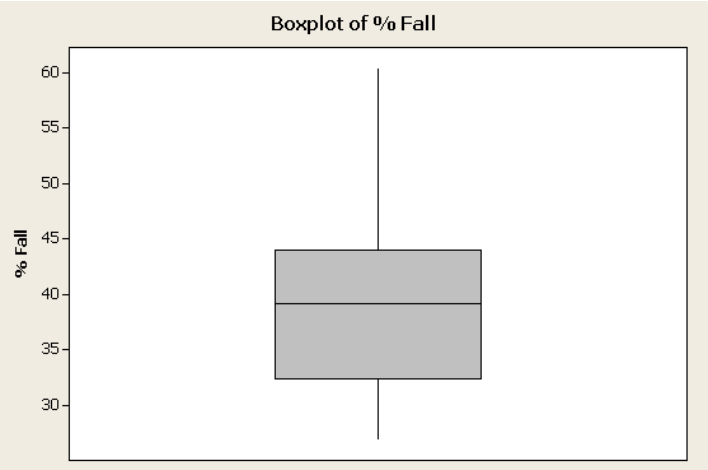
c. IQR = 43.93 – 32.53 = 11.40

Lower Limit:

$Q_1 - 1.5(\text{IQR}) = 32.53 - 1.5(11.40) = 15.43$

Upper Limit:

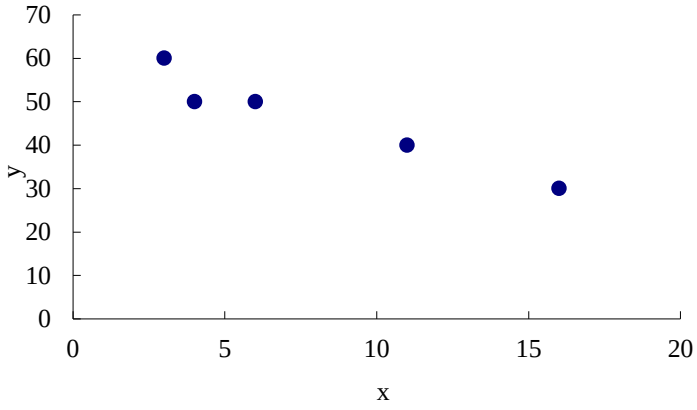
$Q_3 + 1.5(\text{IQR}) = 43.93 + 1.5(11.40) = 61.03$



There are no outliers.

d. Belgium (43.70%) is just below Q_3, i.e. just below the 75th percentile.

32. a.



b. Negative relationship

c/d. $\Sigma x_i = 40 \quad \bar{x} = \frac{40}{5} = 8 \quad \Sigma y_i = 230 \quad \bar{y} = \frac{230}{5} = 46$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = -240 \quad \Sigma(x_i - \bar{x})^2 = 118 \quad \Sigma(y_i - \bar{y})^2 = 520$$

$$s_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{-240}{5-1} = -60$$

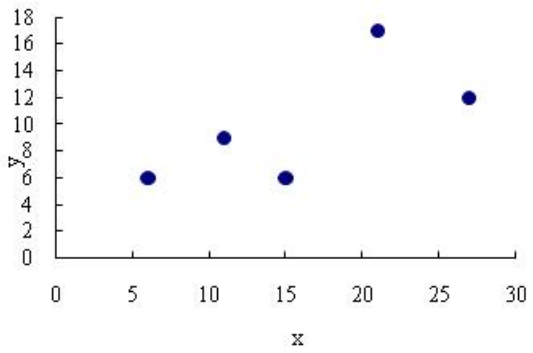
$$s_X = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{118}{5-1}} = 5.4314$$

$$s_Y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{520}{5-1}} = 11.4018$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{-60}{(5.4314)(11.4018)} = -0.969$$

There is a strong negative linear relationship.

33. a.



b. Positive relationship

c/d. $\Sigma x_i = 80 \quad \bar{x} = \frac{80}{5} = 16 \quad \Sigma y_i = 50 \quad \bar{y} = \frac{50}{5} = 10$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 106 \quad \Sigma(x_i - \bar{x})^2 = 272 \quad \Sigma(y_i - \bar{y})^2 = 86$

$$s_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{106}{5-1} = 26.5$$

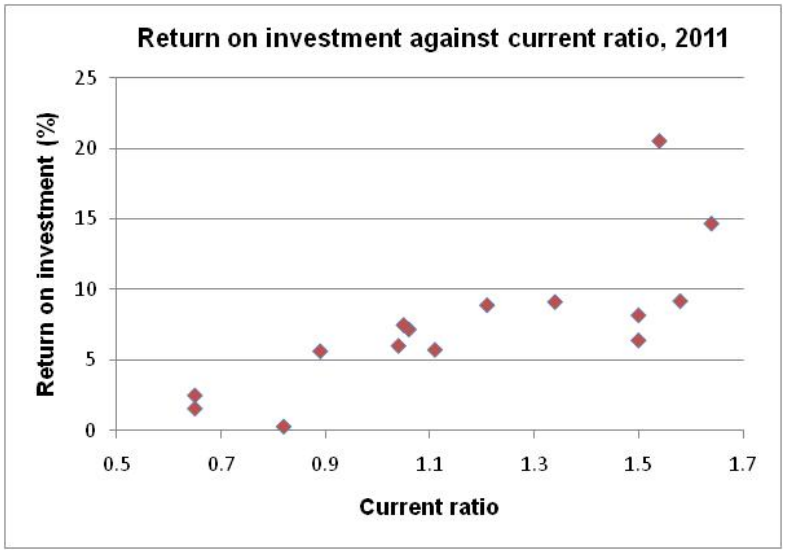
$$s_X = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{272}{5-1}} = 8.2462$$

$$s_Y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{86}{5-1}} = 4.6368$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{26.5}{(8.2462)(4.6368)} = 0.693$$

A positive linear relationship

34. a.



b. Indicates a positive relationship between the two variables.

c. $\Sigma x_i = 17.58 \quad \bar{x} = \dfrac{17.58}{15} = 1.172 \quad \Sigma y_i = 112.97 \quad \bar{y} = \dfrac{112.97}{15} = 7.531$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 18.178$$

$$s_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{18.178}{15-1} = 1.298$$

The covariance is positive, confirming the impression given by the scatter diagram.

d. $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 18.178 \quad \Sigma(x_i - \bar{x})^2 = 1.563 \quad \Sigma(y_i - \bar{y})^2 = 352.81$

$$s_X = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1.563}{15-1}} = 0.3341$$

$$s_Y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{352.81}{15-1}} = 5.020$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{1.298}{0.3341 \times 5.020} = 0.774$$

The correlation coefficient indicates a strong positive linear relationship between the two variables.

35. a. Using DAX as *X* and CAC 40 as *Y*,

$$\mathsf{S}(x_i - \bar{x})(y_i - \bar{y}) = 308{,}343.7 \quad \mathsf{S}(x_i - \bar{x})^2 = 686{,}815.0 \quad \mathsf{S}(y_i - \bar{y})^2 = 148{,}233.6$$

$$s_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{308343.7}{10-1} = 34260.4$$

$$s_X = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{686815.0}{10-1}} = 276.2$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{148233.6}{10-1}} = 128.33$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{34260.4}{(276.2)(128.33)} = 0.967$$

b. There is a very strong positive linear correlation.

36. a. $\bar{x} = \dfrac{\Sigma w_i x_i}{\Sigma w_i} = \dfrac{6(3.2) + 3(2) + 2(2.5) + 8(5)}{6 + 3 + 2 + 8} = \dfrac{70.2}{19} = 3.69$

b. $\dfrac{3.2 + 2 + 2.5 + 5}{4} = \dfrac{12.7}{4} = 3.175$

37. a.

f_i	M_i	$f_i M_i$
4	5	20
7	10	70
9	15	135
5	20	100
25		325

$\bar{x} = \dfrac{\Sigma f_i M_i}{n} = \dfrac{325}{25} = 13$

b.

f_i	M_i	$M_i - \bar{x}$	$(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
4	5	-8	64	256
7	10	-3	9	63
9	15	+2	4	36
5	20	+7	49	245
				600

$$s^2 = \frac{\Sigma f_i (M_i - \bar{x})^2}{n-1} = \frac{600}{24} = 25$$

$$s = \sqrt{25} = 5$$

38. a. Weighted mean for Jil is $\frac{20(80)+20(60)+30(62)+30(57)}{20+20+30+30} = 63.7$

Weighted mean for Ricardo is $\frac{20(48)+20(78)+30(60)+30(53)}{20+20+30+30} = 59.1$

d. Weighted mean for Jil is $\frac{10(80)+25(60)+15(62)+50(57)}{10+25+15+50} = 60.8$

Weighted mean for Ricardo is $\frac{10(48)+25(78)+15(60)+50(53)}{10+25+15+50} = 59.8$

39.

M_i	f_i	$f_i M_i$	$M_i - \bar{x}$	$(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
8	74	592	-26.23	687.90	50904.96
23	192	4416	-11.23	126.07	24204.80
38	280	10,640	3.77	14.23	3983.96
53	105	5565	18.77	352.39	37000.97
68	23	1564	33.77	1140.55	26232.70
83	6	498	48.77	2378.71	14272.28
	680	23,275			156,599.7

$$\bar{x} = \frac{23,275}{680} = 34.2$$

$$s^2 = \frac{156,599.7}{679} = 230.63$$

$$s = 15.2$$

Estimate of total litres sold: (34.2)(120) = 4104

Chapter 3: Descriptive Statistics – Numerical Measures

Supplementary Exercises

40. The American Association of Individual Investors conducts an annual survey of discount brokers. The commissions charged by a sample of 24 discount brokers for two types of trades, a broker-assisted trade of 100 shares at $50 per share and an online trade of 500 shares at $50 per share, are shown in the table below.

	Broker-Assisted	**Online trade**		**Broker-Assisted**	**Online trade**
Accutrade	30.00	29.95	Merrill Lynch Direct	50.00	29.95
Ameritrade	24.99	10.99	Muriel Siebert	45.00	14.95
Banc of America	54.00	24.95	NetVest	24.00	14.00
Brown & Co.	17.00	5.00	Recom Securities	35.00	12.95
Charles Schwab	55.00	29.95	Scottrade	17.00	7.00
CyberTrader	12.95	9.95	Sloan Securities	39.95	19.95
E*TRADE Securities	49.95	14.95	Strong Investments	55.00	24.95
First Discount	35.00	19.75	TD Waterhouse	45.00	17.95
Freedom Investments	25.00	15.00	T. Rowe Price	50.00	19.95
Harrisdirect	40.00	20.00	Vanguard	48.00	20.00
Investors National	39.00	62.50	Wall Street Discount	29.95	19.95
MB Trading	9.95	10.55	York Securities	40.00	36.00

Source: AAII Journal, January 2003.

a. Compute the mean, median and mode for the commission charged on a broker-assisted trade of 100 shares at $50 per share.

b. Compute the mean, median and mode for the commission charged on an online trade of 500 shares at $50 per share.

c. Which costs more, a broker-assisted trade of 100 shares at $50 per share or an online trade of 500 shares at $50 per share?

d. Is the cost of a transaction related to the amount of the transaction?

41. File "Websites"

The data in the file "Websites" show the number of unique visitors to 25 websites.

a. Compute the mean and the median.

b. Do you think it would be better to use the mean or the median as the measure of central location for these data? Explain.

c. Compute the first and third quartiles.

d. Compute and interpret the 85^{th} percentile.

42. File "Cameras"

The data in the file "Cameras" relate to the street price, maximum picture capacity, and battery life (minutes) for 20 digital cameras.

a. Compute the mean price.

b. Compute the mean maximum picture capacity.

c. Compute the mean battery life.

d. If you had to select one camera from this list, what camera would you choose? Explain.

43. miniRank (www.minirank.com) rates the popularity of websites in most countries of the world, using a points system. The 25 most popular sites in Cyprus as listed in November 2008 were as follows (the points scores have been rounded to one decimal place):

Website	Points	Website	Points
www.dart.com.cy	59.2	www.chris-michael.com.cy	8.8
www.dvds.com.cy	21.0	www.music.net.cy	8.7
www.fitness.com.cy	20.5	drivenet.com.cy	8.6
www.airlinetickets.com.cy	20.0	www.prismastore.com.cy	8.6
www.weightloss.com.cy	19.8	www.force.com.cy	8.5
www.cyprus.gov.cy	17.3	www.prisma.com.cy	8.5
www.netcars.com.cy	14.3	www.prismanet.cy	8.5
www.visitcyprus.org.cy	14.3	www.ebos.com.cy	7.3
www.flowershop.com.cy	13.1	www.cytanet.com.cy	6.7
www.netinfo.com.cy	12.5	www.hrdauth.org.cy	6.2
www.interprom.cy	9.5	www.ucy.ac.cy	5.8
www.cyta.com.cy	9.4	www.eplaza.com.cy	5.7
www.drivenet.com.cy	9.1		

a. Compute the mean and median.

b. Do you think it would be better to use the mean or the median as the measure of central location for these data? Explain.

c. Compute the first and third quartiles.

d. Compute and interpret the 85th percentile.

44. According to Forrester Research, Inc. in 2000, approximately 19% of Internet users played games online. The following data show the number of unique users (in thousands) for the month of March for 10 game sites (*The Wall Street Journal*, April 17, 2000).

Site	Unique Users	Site	Unique Users
aolgames.com	9416	prizecentral.com	4899

extremelotto.com	3955	shockwave.com	5582
freelotto.com	12901	speedyclick.com	6628
gamesville.com	4844	uproar.com	8821
iwin.com	7410	webstakes.com	7499

Using these data, compute the mean, median, variance and standard deviation.

45. The American Association of Individual Investors conducts an annual survey of discount brokers (*AAII Journal*, January 2003). The commissions charged by a sample of 24 discount brokers for two types of trades, a broker-assisted trade of 100 shares at $50 per share and an online trade of 500 shares at $50 per share, are shown in the table below.

	Broker-Assisted	**Online trade**		**Broker-Assisted**	**Online trade**
Accutrade	30.00	29.95	Merrill Lynch Direct	50.00	29.95
Ameritrade	24.99	10.99	Muriel Siebert	45.00	14.95
Banc of America	54.00	24.95	NetVest	24.00	14.00
Brown & Co.	17.00	5.00	Recom Securities	35.00	12.95
Charles Schwab	55.00	29.95	Scottrade	17.00	7.00
CyberTrader	12.95	9.95	Sloan Securities	39.95	19.95
E*TRADE Securities	49.95	14.95	Strong Investments	55.00	24.95
First Discount	35.00	19.75	TD Waterhouse	45.00	17.95
Freedom Investments	25.00	15.00	T. Rowe Price	50.00	19.95
Harrisdirect	40.00	20.00	Vanguard	48.00	20.00
Investors National	39.00	62.50	Wall Street Discount	29.95	19.95
MB Trading	9.95	10.55	York Securities	40.00	36.00

Source: AAII Journal, January 2003.

a. Compute the range and interquartile range for each type of trade.

b. Compute the variance and standard deviation for each type of trade.

c. Compute the coefficient of variation for each type of trade.

d. Compare the variability of cost for the two types of trades.

46. *PC World* provided ratings for 15 notebook PCs. A 100-point scale was used to provide an overall rating for each notebook, with higher scores indicating better ratings. The overall ratings for the 15 notebooks are shown here.

Notebook	**Rating**	***Notebook***	**Rating**
AMS Tech Roadster 15CTA380	67	HP Pavillion Notebook PC	83
Compaq Armada M700	78	IBM ThinkPad I Series 1480	78
Compaq Prosignia Notebook 150	79	Micro Express NP7400	77
Dell Inspiron 3700 C466GT	80	Micron TransPort NX PII-400	78
Dell Inspiron 7500 R500VT	84	NEC Versa SX	78
Dell Latitude Cpi A366XT	76	Sceptre Soundx 5200	73
Enpower ENP-313 Pro	77	Sony VAIO PCG-F340	77
Gateway Solo 9300LS	92		

Compute the range, interquartile range, variance, and standard deviation for this sample of notebook PCs.

47. A survey conducted to assess the ability of computer manufacturers to handle problems quickly obtained the following results (*PC Computing*, November 1997).

Company	**Days to Resolve Problems**	**Company**	**Days to Resolve Problems**
Compaq	13	Gateway	21
Packard Bell	27	Digital	27

Quantex	11	IBM	12
Dell	14	Hewlett-Packard	14
NEC	14	AT&T	20
AST	17	Toshiba	37
Acer	16	Micron	17

a. What are the mean and median number of days needed to resolve problems?

b. What are the variance and standard deviation?

c. Which manufacturer holds the best record?

d. What is the *z*-score for Packard Bell?

e. What is the *z*-score for IBM?

f. Do the data contain any outliers?

48. According to a Nielsen Media Research report, young people aged 12 to 17 watch an average of 3 hours of television per day. Suppose that the standard deviation is 1 hour and that the distribution of the time spent watching television has a bell-shaped distribution.

a. What percentage of young people aged 12 to 17 watches television between 2 and 3 hours per day?

b. What percentage of young people aged 12 to 17 watches television between 1 and 4 hours per day?

c. What percentage of young people aged 12 to 17 watches television more than 4 hours per day?

49. The following data are a sample of annual salaries for managers, in thousands of euros.

57.7	64.4	62.1	59.1	71.1	63.0	64.7	61.2	66.8	61.8	64.2	63.3	62.2
61.2	59.4	63.0	66.7	60.3	74.0	62.8	68.7	63.8	59.2	60.3	56.6	59.3
69.5	61.7	58.9	63.1									

a. Compute the mean and standard deviation for the sample data.

b. Using the mean and standard deviation computed in part (a) as estimates of the mean and standard deviation of salary for the population of managers, use Chebyshev's theorem to determine the percentage of managers with an annual salary between €55,000 and €71,000.

c. Construct a histogram for the sample data. Computer software provides 0.97 as the measure of skewness. Does it appear reasonable to assume that the distribution of annual salaries can be approximated by a bell-shaped distribution?

d. Assume that the distribution of annual salary is bell-shaped. Using the mean and standard deviation computed in part (a) as estimates of the mean and standard deviation of salary for the population of managers, use the empirical rule to determine the percentage of managers with an annual salary between €55,000 and €71,000. Compare your answer with the value computed in part (b).

e. Do the sample data contain any outliers?

50. The Highway Loss Data Institute's Injury and Collision Loss Experience report rates car models on the basis of the number of insurance claims filed after accidents. Index ratings near 100 are considered average. Lower ratings are better, indicating a safer car model. Shown are ratings for 20 mid-size cars and 20 small cars.

Mid-size cars:	81	91	93	127	68	81	60	51	58	75
	100	103	119	82	128	76	68	81	91	82
Small cars:	73	100	127	100	124	103	119	108	109	113
	108	118	103	120	102	122	96	133	80	140

Summarize the data for the mid-size and small cars separately.

a. Provide a five-number summary for mid-size cars and for small cars.

b. Construct the box plots.

c. Make a statement about what your summaries indicate about the safety of mid-size cars in comparison to small cars.

51. Public transport and own car are two methods an employee can use to get to work each day. Samples of times recorded for each method are shown. Times are in minutes.

Public Transport:	28	29	32	37	33	25	29	32	41	34
Own Car:	29	31	33	32	34	30	31	32	35	33

a. Compute the sample mean time to get to work for each method.

b. Compute the sample standard deviation for each method.

c. On the basis of your results from parts (a) and (b), which method of transport should be preferred? Explain.

d. Construct a box plot for each method. Does a comparison of the box plots support your conclusion in part (c)?

52. The daily high and low temperatures (in degrees Celsius) for 20 cities on one particular day follow.

City	High	Low	City	High	Low
Athens	24	12	Melbourne	19	10
Bangkok	33	23	Montreal	18	11
Cairo	29	14	Paris	25	13
Copenhagen	18	4	Rio de Janeiro	27	16
Dublin	18	8	Rome	27	12
Havana	30	20	Seoul	18	10
Hong Kong	27	22	Singapore	32	24
Johannesburg	16	10	Sydney	20	13
London	23	9	Tokyo	26	15
Manila	34	24	Vancouver	14	6

What is the correlation between the high and low temperatures?

53. The following data show the trailing 52-weeks primary share earnings and book values as reported by 10 companies (*The Wall Street Journal*, March 13, 2000).

Company	Book value	Earnings
Am Elec	25.21	2.69

Columbia En	23.20	3.01
Con Ed	25.19	3.13
Duke Energy	20.17	2.25
Edison Intn'l	13.55	1.79
Enron Cp	7.44	1.27
Peco	13.61	3.15
Pub Sv Ent	21.86	3.29
Southn C.	8.77	1.86
Unicom	23.22	2.74

a. Construct a scatter diagram for the data with book value on the *x*-axis.

b. What is the sample correlation coefficient, and what does it tell you about the relationship between the earnings per share and the book value?

54. The following data show the media expenditures ($ millions) and shipments in millions of barrels for 10 major brands of beer (*Superbrands '98*, October 20, 1997).

Brand	Media Expenditures ($ millions)	Shipments (millions of barrels)
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9
Coors Light	76.6	13.2

Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Lite	1.4	4.4
Busch Lite	5.3	4.3
Milwaukee's Best	1.7	4.3

a. What is the sample covariance? Does it indicate a positive or negative relationship?

b. What is the sample correlation coefficient?

55. PCWorld provided performance scores and ratings for 15 notebook PCs. The performance score is a measure of how fast a PC can run a mix of common business applications as compared to a baseline machine. For example, a PC with a performance score of 200 is twice as fast as the baseline machine. A 100-point scale was used to provide an overall rating for each notebook tested in the study, with higher scores indicating a better rating. The data are shown below.

Notebook	Performance score	Overall rating
AMS Tech Roadster 15CTA380	115	67
Compaq Armada M700	191	78
Compaq Prosignia Notebook 150	153	79
Dell Inspiron 3700 C466GT	194	80
Dell Inspiron 7500 R500VT	236	84
Dell Latitude Cpi A366XT	184	76
Enpower ENP-313 Pro	184	77
Gateway Solo 9300LS	216	92
HP Pavillion Notebook PC	185	83
IBM ThinkPad I Series 1480	183	78
Micro Express NP7400	189	77
Micron TransPort NX PII-400	202	78
NEC Versa SX	192	78
Sceptre Soundx 5200	141	73
Sony VAIO PCG-F340	187	77

a. Construct a scatter diagram with performance score on the horizontal axis.

b. Is there any relationship between performance score and overall rating? Explain.

c. Compute and interpret the sample covariance.

d. Compute and interpret the sample correlation coefficient.

e. What does the sample correlation coefficient tell you about the relationship between the performance score and the overall rating?

56. The Dow Jones Industrial Average (DJIA) and the Standard & Poor's (S&P) 500 Index are both used as measures of overall movement in the US stock market. The DJIA is based on the price movements of 30 large companies; the S&P 500 is an index composed of 500 stocks. Some say the S&P 500 is a better measure of stock market performance because it is broader based. The index levels of the DJIA and

the S&P 500 for 10 weeks beginning with 1 July 2008 are shown below (file 'DowS&P08' on the accompanying CD).

Date	DJIA	S&P
1 July	11 382.26	1284.91
8 July	11 384.21	1273.70
15 July	10 962.54	1214.91
22 July	11 602.50	1277.00
29 July	11 397.56	1263.20
5 August	11 615.77	1284.88
12 August	11 642.47	1289.59
19 August	11 348.55	1266.69
26 August	11 412.87	1271.51
2 September	11 516.92	1277.58

a. Compute the sample correlation coefficient for these data.

b. Are they poorly correlated, or do they have a close association?

57. The days to maturity for a sample of five money market funds are shown here. The amounts invested in the funds are provided. Use the weighted mean to determine the mean number of days to maturity for money invested in these five funds.

Days to Maturity	**Value (£ millions)**
20	20
12	30
7	10
5	15
6	10

58. Bloomberg Personal Finance (July/August 2001) included the following companies in its recommended investment portfolio. For a portfolio value of €25 000, the recommended euro amounts allocated to each stock are shown.

Company	Portfolio (€)	Estimated growth rate (%)	Dividend yield (%)
Citigroup	3000	15	1.21
General Electric	5500	14	1.48
Kimberley-Clark	4200	12	1.72
Oracle	3000	25	0.00
Pharmacia	3000	20	0.96
SBC Communications	3800	12	2.48
WorldCom	2500	35	0.00

a. Using the portfolio euro amounts as the weights, what is the weighted average estimated growth rate for the portfolio?

b. What is the weighted average dividend yield for the portfolio?

Chapter 3: Descriptive Statistics – Numerical Measures

Supplementary Exercises Solutions

40. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{871.74}{24} = 36.32$

Median is average of 10th and 11th values after arranging in ascending order.

$$\text{Median} = \frac{39.00 + 39.95}{2} = 39.48$$

Data are multimodal

b. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{491.14}{24} = 20.46$

$$\text{Median} = \frac{19.75 + 19.95}{2} = 19.85$$

Data are bimodal: 19.95 (3 brokers), 29.95 (3 brokers)

c. Comparing the measures of central location, we conclude that it costs more to trade 100 shares in a broker-assisted trade than 500 shares online.

d. From the data we have here transaction cost seems more related to whether the trade is broker-assisted or online. The amount of the online transaction is 5 times as great but the cost of the transaction is less. However, if the comparison were restricted to broker-assisted or online trades, we would probably find that larger transactions cost more.

41. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{270,377}{25} = 10,815.08$ Median (Position 13) = 8296

b. Median would be better because of the tail of large data values.

c. i = (25 / 100) 25 = 6.25

Q_1 (Position 7) = 5984

i = (75 / 100) 25 = 18.75

Q_3 (Position 19) = 14,330

d. i = (85/100) 25 = 21.25

85th percentile (position 22) = 15,593. Approximately 85% of the websites have less than 15,593 unique visitors.

42. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{12,780}{20} = \$639$

b. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{1976}{20} = 98.8$ pictures

c. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{2204}{20} = 110.2$ minutes

d. This is not an easy choice because it is a multi-criteria problem. If price was the only criterion, the lowest priced camera (Fujifilm DX-10) would be preferred. If maximum picture capacity was the only criterion, the maximum picture capacity camera (Kodak DC280 Zoom) would be preferred. But, if battery life was the only criterion, the maximum battery life camera (Fujifilm DX10) would be preferred. There are many approaches used to select the best choice in a multi-criteria situation. These approaches are discussed in more specialized books on decision analysis.

43. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{331.9}{25} = 13.28$

Median is the 13^{th} ordered data value = 9.1

b. The data are positively skewed, with one high outlier. The median would therefore be a better measure of central location.

c. For Q_1, $i = \left(\frac{25}{100}\right)25 = 6.25$

7^{th} ordered data value is $Q_1 = 8.5$

For Q_3, $i = \left(\frac{75}{100}\right)25 = 18.75$

19^{th} ordered data value is $Q_3 = 14.3$

d. $i = \left(\frac{85}{100}\right)25 = 21.25$

Since i is not an integer, we round up to the 22^{nd} position.

85^{th} percentile = 20.0

Approximately 85% of the sites have points scores below 20.0.

44. Sample mean = 7195.5

Median = 7019 (average of positions 5 and 6)

Sample variance = 7,165,941

Sample standard deviation = 2676.93

45. a. <u>100 Shares at $50 (Broker-assisted)</u>

Min Value = 9.95 Max Value = 55.00

Range = 55.00 – 9.95 = 45.05

$$Q_1 = \frac{24.99 + 25.00}{2} = 24.995 \qquad Q_3 = \frac{48.00 + 49.95}{2} = 48.975$$

Interquartile range = 48.975 – 24.995 = 23.98

<u>500 Shares at $50 (Online)</u>

Min Value = 5.00 Max Value = 62.50

Range = 62.50 – 5.00 = 57.50

$$Q_1 = \frac{12.95 + 14.00}{2} = 13.475 \qquad Q_3 = \frac{24.95 + 24.95}{2} = 24.95$$

Interquartile range = 24.950 – 13.475 = 11.475

b. 100 Shares at $50 (Broker-assisted)

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = 190.67$$
$$s = \sqrt{s^2} = 13.81$$

500 Shares at $50 (Online)

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = 140.63$$
$$s = \sqrt{s^2} = 11.86$$

c. 100 Shares at $50 (Broker-assisted)

Coefficient of Variation = $\frac{s}{\bar{x}}(100\%) = \frac{13.81}{36.32}(100\%) = 38.02\%$

500 Shares at $50 (Online)

Coefficient of Variation = $\frac{s}{\bar{x}}(100\%) = \frac{11.86}{20.46}(100\%) = 57.97\%$

d. Using the standard deviation as a measure, the variability seems to be greater for the broker-assisted trades. But, using the coefficient of variation as a measure, we see that the relative variability is greater for the online trades.

46. Range = 92 – 67 = 25

IQR = $Q_3 - Q_1 = 80 - 77 = 3$

$\bar{x} = 78.4667$

$\sum (x_i - \bar{x})^2 = 411.7333$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{411.7333}{14} = 29.4095$$

$s = \sqrt{29.4095} = 5.4231$

47. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{260}{14} = 18.57$

Median = 16.5 (Average of 7th and 8th values)

b. $s^2 = 53.49$ $s = 7.31$

c. Quantex has the best record: 11 Days

d. $z = \frac{27 - 18.57}{7.31} = 1.15$

Packard-Bell is 1.15 standard deviations slower than the mean.

e. $z = \frac{12 - 18.57}{7.31} = -0.90$

IBM is 0.9 standard deviations faster than the mean.

f. Check Toshiba:

$z = \frac{37 - 18.57}{7.31} = 2.52$

On the basis of *z*-scores, Toshiba is not an outlier, but it is 2.52 standard deviations slower than the mean.

48. a. 2 hours is 1 standard deviation below the mean. The empirical rule suggests that 68% of the youngsters watch television between 2 and 4 hours per day. Since a

bell-shaped distribution is symmetrical, approximately 34% of the youngsters watch television between 2 and 3 hours per day.

b. 1 hour is 2 standard deviations below the mean. The empirical rule suggests that 95% of the youngsters watch television between 1 and 5 hours per day. Since a bell-shaped distribution is symmetrical, approximately, 47.5% of the youngsters watch television between 1 and 3 hours per day. In part (a) we concluded that approximately 34% of the youngsters watch television between 2 and 3 hours per day. Therefore, approximately 34% of the youngsters watch television between 3 and 4 hours per day. Hence, approximately 47.5% + 34% = 81.5% of youngsters watch television between 1 and 4 hours per day.

c. Since 34% of the youngsters watch television between 3 and 4 hours per day, 50% - 34% = 16% of the youngsters watch television more than 4 hours per day.

49. a. $\overline{x}$ is approximately 63 or €63,000, and s is 4 or €4000

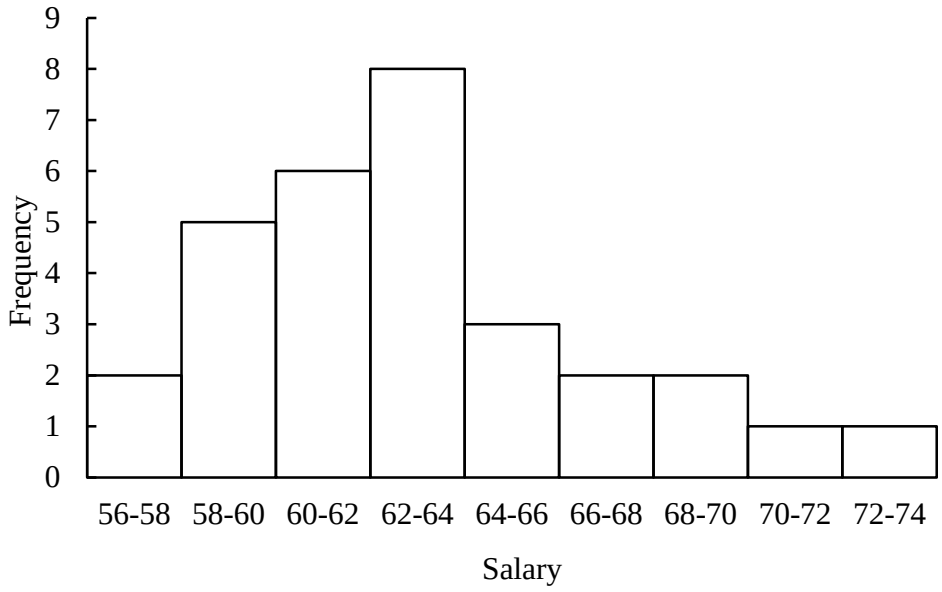
b. This is from 2 standard deviations below the mean to 2 standard deviations above the mean.

With $z = 2$, Chebyshev's theorem gives:

$$1-\frac{1}{z^2}=1-\frac{1}{2^2}=1-\frac{1}{4}=\frac{3}{4}$$

Therefore, at least 75% of managers have an annual salary between €55,000 and €71,000.

c. The histogram of the salary data is shown below:



Visual inspection of the histogram and the skewness measure of 0.97 indicate that it is moderately skewed to the right. Although the distribution is not perfectly bell-shaped, it does appear that the distribution of annual salaries for benefit managers could be approximated by a bell-shaped distribution.

d. With $z = 2$, the empirical rule suggests that 95% of benefits managers have an annual salary between €55,000 and €71,000. The percentage is much higher than obtained using Chebyshev's theorem, but requires the assumption that the distribution of annual salary is bell-shaped.

e. There are no outliers because all the observations are within 3 standard deviations of the mean.

50. a. Five Number Summary (Mid-size)

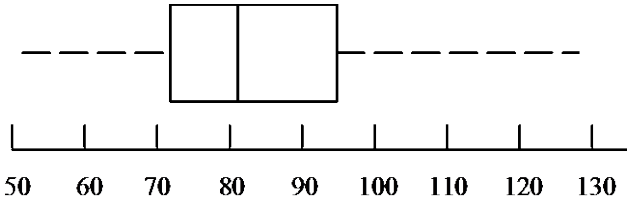
51 71.5 81.5 96.5 128

Five Number Summary (Small)

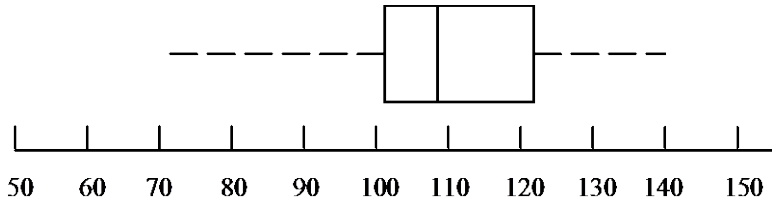
73 101 108.5 121 140

b. Box Plots

Midsize



Small Size



c. The mid-size cars appear to be safer than the small cars.

51. a. Public Transport: $\bar{x} = \frac{320}{10} = 32$

Car: $\bar{x} = \frac{320}{10} = 32$

b. Public Transport: $s = 4.64$

Car: $s = 1.83$

c. Prefer own car. The mean times are the same, but the car has less variability.

d. Data in ascending order:

Public: 25 28 29 29 32 32 33 34 37 41

Car: 29 30 31 31 32 32 33 33 34 35

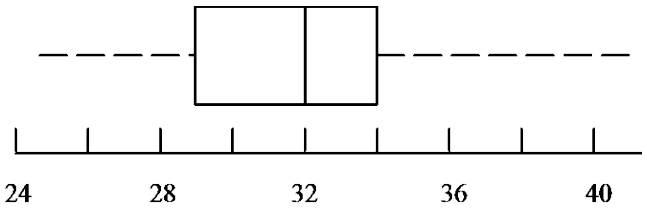
Five number Summaries

Public: 25 29 32 34 41

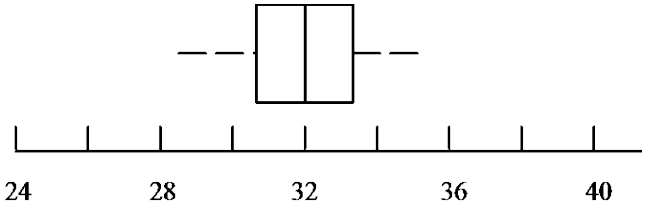
Car: 29 31 32 33 35

Box Plots:

Public:



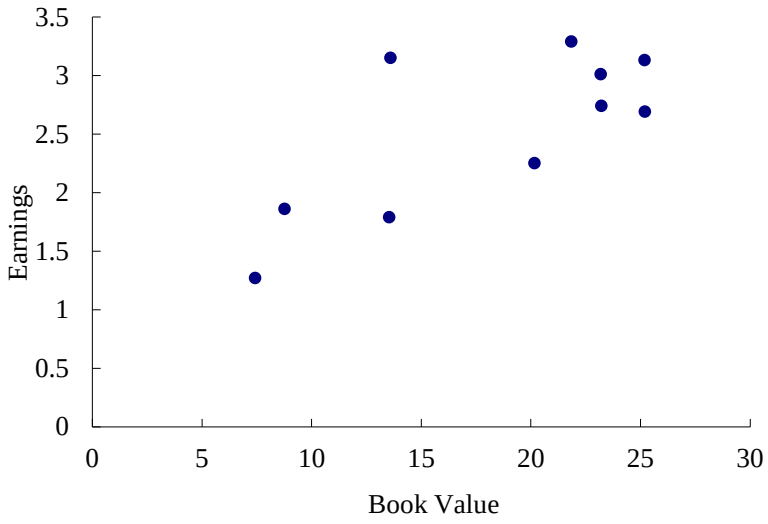
Car:



The box plots do show lower variability with own car and support the conclusion in part c.

52. The sample correlation coefficient is 0.88. This indicates a strong positive linear relationship between the daily high and low temperatures.

53. a.



b. The sample correlation coefficient is 0.75; this indicates a positive linear relationship between book value and earnings.

54. a. Let X = media expenditures ($ millions) and Y = shipments in barrels (millions)

$$\Sigma x_i = 404.1 \quad \bar{x} = \frac{404.1}{10} = 40.41 \quad \Sigma y_i = 119.9 \quad \bar{y} = \frac{119.9}{10} = 11.99$$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 3763.481 \quad \Sigma(x_i - \bar{x})^2 = 19,248.469 \quad \Sigma(y_i - \bar{y})^2 = 939.349$$

$$s_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{3763.481}{10-1} = 418.1646$$

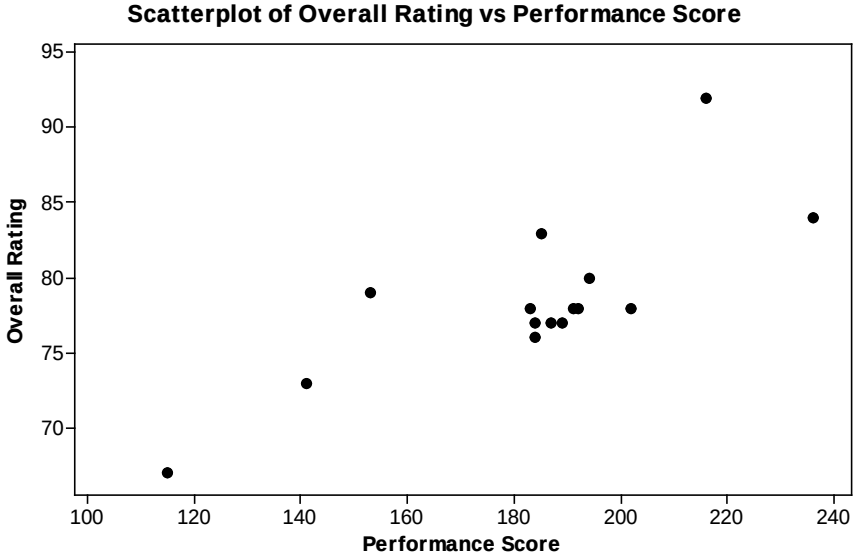
A positive relationship

b. $$s_X = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{19,248.469}{10-1}} = 46.2463$$

$$s_Y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{939.349}{10-1}} = 10.2163$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{418.1646}{(46.2463)(10.2163)} = 0.885$$

55. a.



b. The scatter diagram suggests a positive relationship between the two variables.

c. $\Sigma x_i = 2752 \quad \bar{x} = \frac{2752}{15} = 183.47 \quad \Sigma y_i = 1177 \quad \bar{y} = \frac{1177}{15} = 78.47$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 1723.73$

$s_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1723.73}{15-1} = 123.12$

The covariance is positive, confirming the impression given by the scatter diagram.

d. $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 1723.73 \quad \Sigma(x_i - \bar{x})^2 = 11{,}867.73 \quad \Sigma(y_i - \bar{y})^2 = 411.73$

$s_X = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{11{,}867.73}{15-1}} = 29.12$

$s_Y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{411.73}{15-1}} = 5.423$

$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{123.12}{(29.12)(5.423)} = 0.78$

The correlation coefficient indicates quite a strong positive linear relationship between the two variables.

e. High performance scores tend to be paired with high overall ratings, and low performance scores tend to be paired with high overall ratings.

56. a. Using DJIA as X and S&P as Y,

$\mathrm{S}(x_i - \bar{x})(y_i - \bar{y}) = 34{,}142.01 \quad \mathrm{S}(x_i - \bar{x})^2 = 347{,}721.91 \quad \mathrm{S}(y_i - \bar{y})^2 = 4040.45$

$$s_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{34{,}142.01}{10-1} = 3793.6$$

$$s_X = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{347{,}721}{10-1}} = 196.6$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{4040.445}{10-1}} = 21.188$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{3793.6}{(196.6)(21.188)} = 0.911$$

b. There is a strong positive linear correlation.

57. $$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{20(20) + 30(12) + 10(7) + 15(5) + 10(6)}{20 + 30 + 10 + 15 + 10} = \frac{965}{85} = 11.4 \text{ days}$$

58. a. $$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{3000(15) + 5500(14) + 4200(12) + 3000(25) + 3000(20) + 3800(12) + 2500(35)}{3000 + 5500 + 4200 + 3000 + 3000 + 3800 + 2500}$$

$$= \frac{440{,}500}{25{,}000} = 17.62\%$$

b.

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{3000(1.21) + 5500(1.48) + 4200(1.72) + 3000(0) + 3000(0.96) + 3800(2.48) + 2500(0)}{3000 + 5500 + 4200 + 3000 + 3000 + 3800 + 2500}$$

$$= \frac{31{,}298}{25{,}000} = 1.25\%$$

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Four

Introduction to Probability

Textbook Exercises (1-34)

Textbook Exercise Solutions

Supplementary Exercises (35-49)

Supplementary Exercise Solutions

Chapter 4: Introduction to Probability

Textbook Exercises:

1 An experiment has three steps with three outcomes possible for the first step, two outcomes possible for the second step, and four outcomes possible for the third step. How many experimental outcomes exist for the entire experiment?

2 How many ways can three items be selected from a group of six items? Use the letters A, B, C, D, E, and F to identify the items, and list each of the different combinations of three items.

3 How many permutations of three items can be selected from a group of six? Use the letters A, B, C, D, E, and F to identify the items, and list each of the permutations of items B, D, and F.

4 Consider the experiment of tossing a coin three times.

a. Develop a tree diagram for the experiment.
b. List the experimental outcomes.
c. What is the probability for each experimental outcome?

5 Suppose an experiment has five equally likely outcomes: E1, E2, E3, E4, E5. Assign probabilities to each outcome and show that the requirements in equations (4.3) and (4.4) are satisfied. What method did you use?

6 An experiment with three outcomes has been repeated 50 times, and it was learned that E1 occurred 20 times, E2 occurred 13 times, and E3 occurred 17 times. Assign probabilities to the outcomes. What method did you use?

7 A decision-maker subjectively assigned the following probabilities to the four outcomes of an experiment: P(E1) = 0.10, P(E2) = 0.15, P(E3) = 0.40, and P(E4) = 0.20. Are these probability assignments valid? Explain.

8 Applications for zoning changes in a large metropolitan city go through a two-step process: a review by the planning commission and a final decision by the city council. At step 1 the planning commission reviews the zoning change request and makes a positive or negative recommendation concerning the change. At step 2 the city council reviews the planning commission's recommendation and then votes to approve or to disapprove the zoning change. Suppose the developer of an apartment complex submits an application for a zoning change. Consider the application process as an experiment.

a. How many sample points are there for this experiment? List the sample points.

b. Construct a tree diagram for the experiment.

9 A total of 11 Management students, 4 International Management and American Business Studies (IMABS) and 8 International Management and French Studies (IMF) students have volunteered to take part in an Inter-University tournament.

a. How many different ways can a team consisting of 8 Management students, 2 IMABS and 5 IMF students be selected?

b. If after the team has been selected, 1 Management, 1 IMABS and 2 IMF students are found to be suffering from glandular fever and are unable to play, what is the probability that the team will not have to be changed?

10 A company that franchises coffee houses conducted taste tests for a new coffee product. Four blends were prepared, then randomly chosen individuals were asked to taste the blends and state which one they liked best. Results of the taste test for 100 individuals are given.

Blend	Number choosing
1	20
2	30
3	35
4	15

a. Define the experiment being conducted. How many times was it repeated?

b. Prior to conducting the experiment, it is reasonable to assume preferences for the four blends are equal. What probabilities would you assign to the experimental outcomes prior to conducting the taste test? What method did you use?

c. After conducting the taste test, what probabilities would you assign to the experimental outcomes? What method did you use?

11 A company that manufactures toothpaste is studying five different package designs. Assuming that one design is just as likely to be selected by a consumer as any other design, what selection probability would you assign to each of the package designs? In an actual experiment, 100 consumers were asked to pick the design they preferred. The following data were obtained. Do the data confirm the belief that one design is just as likely to be selected as another? Explain.

Design times	Number of preferred
1	5
2	15
3	30
4	40
5	10

12 An experiment has four equally likely outcomes: E1, E2, E3, and E4.

a. What is the probability that E2 occurs?

b. What is the probability that any two of the outcomes occur (e.g. E1 or E3)?

c. What is the probability that any three of the outcomes occur (e.g. E1 or E2 or E4)?

13 Consider the experiment of selecting a playing card from a deck of 52 playing cards. Each card corresponds to a sample point with a 1/52 probability.

a. List the sample points in the event an ace is selected.

b. List the sample points in the event a club is selected.

c. List the sample points in the event a face card (jack, queen, or king) is selected.

d. Find the probabilities associated with each of the events in parts (a), (b) and (c).

14 Consider the experiment of rolling a pair of dice. Suppose that we are interested in the sum of the face values showing on the dice.

a. How many sample points are possible? (Hint: Use the counting rule for multiple-step experiments.)

b. List the sample points.

c. What is the probability of obtaining a value of 7?

d. What is the probability of obtaining a value of 9 or greater?

e. Because each roll has six possible even values (2, 4, 6, 8, 10 and 12) and only fi ve possible odd values (3, 5, 7, 9 and 11), the dice should show even values more often than odd values. Do you agree with this statement? Explain.

f. What method did you use to assign the probabilities requested?

15 Refer to the KPL sample points and sample point probabilities in Tables 4.2 and 4.3.

a. The design stage (stage 1) will run over budget if it takes four months to complete. List the sample points in the event the design stage is over budget.

b. What is the probability that the design stage is over budget?

c. The construction stage (stage 2) will run over budget if it takes eight months to complete.

List the sample points in the event the construction stage is over budget.

d. What is the probability that the construction stage is over budget?

e. What is the probability that both stages are over budget?

16 Suppose that a manager of a large apartment complex provides the following subjective probability estimates about the number of vacancies that will exist next month. Provide the probability of each of the following events.

Vacancies	Probability
0	0.10
1	0.15
2	0.30
3	0.20
4	0.15
5	0.10

a. No vacancies.

b. At least four vacancies.

c. Two or fewer vacancies.

17 When three marksmen take part in a shooting contest, their chances of hitting the target are 1/2, 1/3 and ¼ respectively. If all three marksmen fire at it simultaneously

a. What is the chance that one and only one bullet will hit the target?

b. What is the chance that two marksmen will hit the target (and therefore one will not)?

c. What is the chance that all three marksmen will hit the target?

18 Suppose that we have a sample space with fi ve equally likely experimental outcomes: E1, E2, E3, E4, E5. Let

$A = \{E_1, E_2\}$
$B = \{E_3, E_4\}$
$C = \{E_2, E_3, E_5\}$

a. Find $P(A)$, $P(B)$, and $P(C)$.
b. Find $P(A \cup B)$. Are A and B mutually exclusive?
c. Find $\overline{A}$, $\overline{C}$, $P(\overline{A})$, and $P(\overline{C})$.
d. Find $A \cup \overline{B}$ and $P(A \cup \overline{B})$.
e. Find $P(B \cup C)$.

19 Suppose that we have a sample space S = {E1, E2, E3, E4, E5, E6, E7}, where E1, E2, . . . , E7 denote the sample points. The following probability assignments apply: P(E1) = 0.05, P(E2) = 0.20, P(E3) = 0.20, P(E4) = 0.25, P(E5) = 0.15, P(E6) = 0.10, and P(E7) = 0.05.
Let

$A = \{E_1, E_2\}$
$B = \{E_3, E_4]$
$C = \{E_2, E_3, E_5\}$

a. Find $P(A)$, $P(B)$, and $P(C)$.
b. Find $A \cup B$ and $P(A \cup B)$.
c. Find $A \cap B$ and $P(A \cap B)$.
d. Are events A and C mutually exclusive?
e. Find $\overline{B}$ and $P(\overline{B})$.

20 A survey of magazine subscribers showed that 45.8 per cent rented a car during the past 12 months for business reasons, 54 per cent rented a car during the past 12 months for personal reasons, and 30 per cent rented a car during the past 12 months for both business and personal reasons.

a. What is the probability that a subscriber rented a car during the past 12 months for business or personal reasons?

b. What is the probability that a subscriber did not rent a car during the past 12 months for either business or personal reasons?

21 Suppose that we have two events, A and B, with P(A) = 0.50, P(B) = 0.60, and P(A ∩ B) = 0.40.

a. Find P(A | B).

b. Find P(B | A).

c. Are A and B independent? Why or why not?

22 Assume that we have two events, A and B, that are mutually exclusive. Assume further that we know P(A) = 0.30 and P(B) = 0.40.

a. What is P(A ∩ B)?

b. What is P(A | B)?

c. A student in statistics argues that the concepts of mutually exclusive events and independent events are really the same, and that if events are mutually exclusive they must be independent. Do you agree with this statement? Use the probability information in this problem to justify your answer.

d. What general conclusion would you make about mutually exclusive and independent events given the results of this problem?

23 A Paris nightclub obtains the following data on the age and marital status of 140 customers.

a. Develop a joint probability table for these data.

	Marital status	
Age	Single	Married
Under 30	77	14
30 or over	28	21

b. Use the marginal probabilities to comment on the age of customers attending the club.

c. Use the marginal probabilities to comment on the marital status of customers attending the club.

d. What is the probability of finding a customer who is single and under the age of 30?

e. If a customer is under 30, what is the probability that he or she is single?

f. Is marital status independent of age? Explain, using probabilities.

24. A slot machine in Melbourne has a hold facility. A gambler experiments with this to see if his success rate is higher when he uses 'hold' compared to when he does not.

The results from 120 plays can be summarised as follows:

	Win	Lose
Hold	14	36
Not hold	10	60

What is the probability that the gambler:

a. Holds?
b. Wins?
c. Wins given that he held?
d. Held and lost?
e. Held given that he won?

25. A sample of convictions and compensation orders issued at a number of Scottish courts was followed up to see whether the offender had paid the compensation to the victim. Details by gender of offender are as follows:

Offender gender	Payment outcome		
	Paid in full	Part paid	Nothing paid
Male	754	62	61
Female	157	7	6

a. What is the probability that no compensation was paid?

b. What is the probability that the offender was not male given that compensation was part paid?

26 A purchasing agent in Haifa placed rush orders for a particular raw material with two different suppliers, A and B. If neither order arrives in four days, the production process must be shut down until at least one of the orders arrives. The probability that supplier A can deliver the material in four days is 0.55. The probability that supplier B can deliver the material in four days is 0.35.

a. What is the probability that both suppliers will deliver the material in four days? Because two separate suppliers are involved, we are willing to assume independence.

b. What is the probability that at least one supplier will deliver the material in four days?

c. What is the probability that the production process will be shut down in four days because of a shortage of raw material (that is, both orders are late)?

27 The prior probabilities for events A1 and A2 are P(A1) = 0.40 and P(A2) = 0.60. It is also known that P(A1 ∩ A2) = 0. Suppose P(B | A1) = 0.20 and P(B | A2) = 0.05.

a. Are A1 and A2 mutually exclusive? Explain.

b. Compute P(A1 ∩ B) and P(A2 ∩ B).

c. Compute P(B).

d. Apply Bayes' theorem to compute P(A1 | B) and P(A2 | B).

28 The prior probabilities for events A1, A2, and A3 are P(A1) = 0.20, P(A2) = 0.50 and P(A3) = 0.30. The conditional probabilities of event B given A1, A2, and A3 are P(B | A1) = 0.50, P(B | A2) = 0.40 and P(B | A3) = 0.30.

a. Compute P(B ∩ A1), P(B ∩ A2) and P(B ∩ A3).

b. Apply Bayes' theorem, equation (4.19), to compute the posterior probability P(A2 | B).

c. Use the tabular approach to applying Bayes' theorem to compute P(A1 | B), P(A2 | B) and P(A3 | B).

29 Records show that for every 100 items produced in a factory during the day shift, two are defective and for every 100 items produced during the night shift, four are defective. What is the prior probability of the bid being successful (that is, prior to the request for additional information)?

a. If during a 24 hour period, 2000 items are produced during the day and 800 at night, what is the probability that an item picked at random from the output over 24 hours came from the night shift if it was defective?

30 A company is about to sell to a new client. It knows from past experience that there is a real possibility that the client may default on payment. As a precaution the company checks with a consultant on the likelihood of the client defaulting in this case and is given an estimate of 20%. Sometimes the consultant gets it wrong. Your own experience of the consultant is that he is correct 70% of the time when he predicts that the client will default but that 20% of clients who he believes will not default actually do.

a. What is the probability that the new client will not default?

31 In 2011, there were 1901 fatalities recorded on Britain's roads, 60 of which were for children (Department of Transport, 2012). Correspondingly, serious injuries totalled 23 122 of which 20 770 were for adults.

a. What is the probability of a serious injury given the victim was a child?

b. What is the probability that the victim was an adult given a fatality occurred?

32 The following cross-tabulation shows industry type and Price/Earnings (P/E) ratio for 100 companies in the consumer products and banking industries.

	P/E ratio					
Industry	5–9	10–14	15–19	20–24	25–29	Total
Consumer	4	10	18	10	8	50
Banking	14	14	12	6	4	50
Total	18	24	30	16	12	100

a. What is the probability that a company had a P/E greater than 9 and belonged to the consumer industry?

b. What is the probability that a company with a P/E in the range 15–19 belonged to the banking industry?

33. A large investment advisory service has a number of analysts who prepare detailed studies of individual companies. On the basis of these studies the analysts make 'buy' or 'sell' recommendations on the companies' shares. The company classes an excellent analyst as one who will be correct 80 per cent of the time, a good analyst as who will be correct 60 per cent of the time, and a poor analyst who will be correct 40 per cent of the time. Two years ago, the advisory service hired Mr Smith who came with considerable experience from the research department of another firm. At the time he was hired it was thought that the probability was 0.90 that he was an excellent analyst, 0.09 that he was a good analyst and 0.01 that he was a poor analyst. In the past two years he has made ten recommendations of which only three have been correct. Assuming that each recommendation is an independent event what probability would you assign to Mr Smith being:

a. An excellent analyst?

b. A good analyst?

c. A poor analyst?

34 An electronic component is produced by four production lines in a manufacturing operation. The components are costly, are quite reliable and are shipped to suppliers in 50-component lots. Because testing is destructive, most buyers of the components test only a small number before deciding to accept or reject lots of incoming components. All four production lines usually only produce 1 per cent defective components which are randomly dispersed in the output. Unfortunately, production line 1 suffered mechanical difficulty and produced 10 per cent defectives during the month of April. This situation became known to the manufacturer after the components had been shipped. A customer received a lot in April and tested five components. Two failed. What is the probability that this lot came from production line 1?

Chapter 4: Introduction to Probability

Textbook Exercises Solutions:

1. Number of experimental Outcomes = (3) (2) (4) = 24

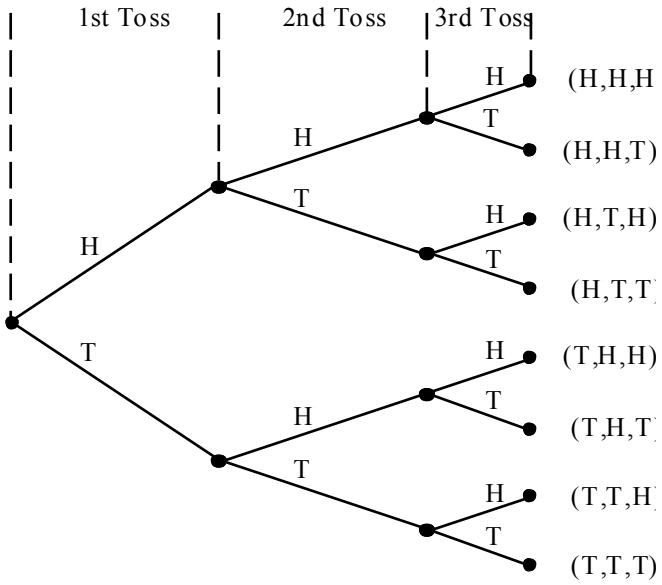
2. $${}^{6}C_{3} = \frac{6!}{3!} = \frac{6*5*4}{3*2*1} = 20$$

ABC	ACE	BCD	BEF
ABD	ACF	BCE	CDE
ABE	ADE	BCF	CDF
ABF	ADF	BDE	CEF
ACD	AEF	BDF	DEF

3. $${}^{6}P_{3} = \frac{6!}{3!} = 6*5*4 = 120$$

BDF BFD DBF DFB FBD FDB

4. a.



b. Let: H be head and T be tail

(H,H,H) (T,H,H)
(H,H,T)(T,H,T)
(H,T,H)(T,T,H)
(H,T,T) (T,T,T)

c. The outcomes are equally likely, so the probability of each outcomes is 1/8.

5. $P(E_i) = 1 / 5$ for i = 1, 2, 3, 4, 5

$P(E_i) \geq 0$ for i = 1, 2, 3, 4, 5
$P(E_1) + P(E_2) + P(E_3) + P(E_4) + P(E_5) = 1 / 5 + 1 / 5 + 1 / 5 + 1 / 5 + 1 / 5 = 1$

The classical method was used.

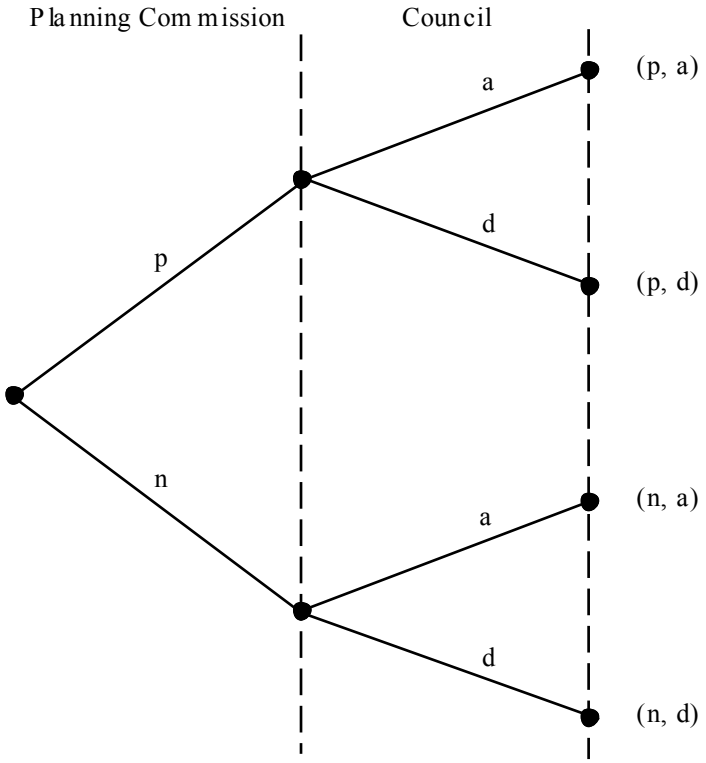
6. $P(E_1) = .40$, $P(E_2) = .26$, $P(E_3) = .34$

The relative frequency method was used.

7. No. Requirement (4.4) is not satisfied; the probabilities do not sum to 1. $P(E_1) + P(E_2) + P(E_3) + P(E_4) = .10 + .15 + .40 + .20 = .85$

8. a. There are four outcomes possible for this 2-step experiment; planning commission positive - council approves; planning commission positive - council disapproves; planning commission negative - council approves; planning commission negative - council disapproves.

b. Let p = positive, n = negative, a = approves, and d = disapproves



9. a. Initially the number of ways = ${}^{11}C_8 * {}^4C_2 * {}^8C_5 = 55440$

b. Afterwards, the number of ways becomes ${}^{10}C_8 * {}^3C_2 * {}^6C_5 = 810$

Required probability = 810/55440 = 0.01461

10. a. Choose a person at random. Have the person taste the four blends and state which is preferred.

b. Assign a probability of 1/4 to each blend. We use the classical method of equally likely outcomes here.

c.

Blend	Probability
1	.20
2	.30
3	.35
4	.15
Total	1.00

The relative frequency method was used.

11. Initially a probability of .20 would be assigned if selection is equally likely. Data does not appear to confirm the belief of equal consumer preference. For example using the relative frequency method we would assign a probability of 5 / 100 = .05 to the design 1 outcome, .15 to design 2, .30 to design 3, .40 to design 4, and .10 to design 5.

12. a. $P(E_2) = 1 / 4$

b. P(any 2 outcomes) = 1 / 4 + 1 / 4 = 1 / 2

c. P(any 3 outcomes) = 1 / 4 + 1 / 4 + 1 / 4 = 3 / 4

13. a. *S* = {ace of clubs, ace of diamonds, ace of hearts, ace of spades}

b. *S* = {2 of clubs, 3 of clubs, . . . , 10 of clubs, J of clubs, Q of clubs, K of clubs, A of clubs}

c. There are 12; jack, queen, or king in each of the four suits.

d. For a: 4 / 52 = 1 / 13 = .08

For b: 13 / 52 = 1 / 4 = .25

For c: 12 / 52 = .23

14. a. (6) (6) = 36 sample points

b.

Die 1 \ Die 2	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8 ← Total for Both
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

c. 6 / 36 = 1 / 6

d. 10 / 36 = 5 / 18

e. No. P(odd) = 18 / 36 = P(even) = 18 / 36 or 1 / 2 for both.

f. Classical. A probability of 1 / 36 is assigned to each experimental outcome.

15. a. (4, 6), (4, 7), (4 , 8)

b. .05 + .10 + .15 = .30

c. (2, 8), (3, 8), (4, 8)

d. .05 + .05 + .15 = .25

e. .15

16. a. P(0) = 0.10

b. P(4 or 5) = 0.25

c. P(0, 1, or 2) = 0.55

17. a. p(one bullet only hits target) = p(marksmen 1 hits the target but not marksmen 2 or 3) +

p(marksmen 2 hits the target but not marksmen 1 or 3) +

p(marksmen 3 hits the target but not marksmen 1 or 2)

$$= \frac{1}{2} * \frac{2}{3} * \frac{3}{4} + \frac{1}{3} * \frac{1}{2} * \frac{3}{4} + \frac{1}{4} * \frac{1}{2} * \frac{2}{3} = \frac{11}{24}$$

since, by independence,

p(marksmen 1 hits the target but not marksmen 2 or 3) =

p(marksmen 1 hits target)*p(marksmen 2 misses target)*p(marksmen 3 misses)

$= \frac{1}{2} * \frac{2}{3} * \frac{3}{4} = \frac{6}{24}$ etc

b. p(two bullets only hit target) = p(marksmen 1 misses the target but not marksmen 2 or 3) +

p(marksmen 2 misses the target but not marksmen 1 or 3) +

p(marksmen 3 misses the target but not marksmen 1 or 2)

$$= \frac{1}{2} * \frac{1}{3} * \frac{1}{4} + \frac{2}{3} * \frac{1}{2} * \frac{1}{4} + \frac{3}{4} * \frac{1}{2} * \frac{1}{3} = \frac{6}{24}$$

c. p(three bullets hit target) =

p(marksmen 1 hits target)*p(marksmen 2 hits target)*p(marksmen 3 hits target) by independence.

$= \frac{1}{2} * \frac{1}{3} * \frac{1}{4} = \frac{1}{24}$

18. a. P(A) = 0.40, P(B) = 0.40, P(C) = 0.60

b. $P(A \cup B) = P(E_1, E_2, E_3, E_4) = 0.80$. Yes $P(A \cup B) = P(A) + P(B)$.

c. $\overline{A} = \{E_3, E_4, E_5\}$ $\overline{C} = \{E_1, E_4\}$ $P(\overline{A}) = 0.60$ $P(\overline{C}) = 0.40$

d. $A \cup \overline{B} = \{E_1, E_2, E_5\}$ $P(A \cup \overline{B}) = 0.60$

e. $P(B \cup C) = P(E_2, E_3, E_4, E_5) = 0.80$

19. a. $P(A) = P(E_1) + P(E_4) + P(E_6) = .05 + .25 + .10 = .40$

$P(B) = P(E_2) + P(E_4) + P(E_7) = .20 + .25 + .05 = .50$

$P(C) = P(E_2) + P(E_3) + P(E_5) + P(E_7) = .20 + .20 + .15 + .05 = .60$

b. $A \cup B = \{E_1, E_2, E_4, E_6, E_7\}$

$P(A \cup B) = P(E_1) + P(E_2) + P(E_4) + P(E_6) + P(E_7)$
$= .05 + .20 + .25 + .10 + .05 = .65$

c. $A \cap B = \{E_4\}$ $P(A \cap B) = P(E_4) = .25$

d. Yes, they are mutually exclusive.

e. $\overline{B} = \{E_1, E_3, E_5, E_6\}$; $P(\overline{B}) = P(E_1) + P(E_3) + P(E_5) + P(E_6)$
$= .05 + .20 + .15 + .10 = .50$

20. Let: B = rented a car for business reasons
P = rented a car for personal reasons

a. $P(B \cup P) = P(B) + P(P) - P(B \cap P)$
$= .54 + .458 - .30 = .698$

b. P(Neither) = 1 - .698 = .302

21. a. $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.40}{.60} = .6667$

b. $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{.40}{.50} = .80$

c. No because $P(A \mid B) \neq P(A)$

22. a. $P(A \cap B) = 0$

b. $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0}{.4} = 0$

c. No. $P(A \mid B) \neq P(A)$; $\therefore$ the events, although mutually exclusive, are not independent.

d. Mutually exclusive events are dependent.

23 a.

	Single	Married	Total
Under 30	.55	.10	.65
30 or over	.20	.15	.35
Total	.75	.25	1.00

b. 65% of the customers are under 30.

c. The majority of customers are single: P(single) = .75.

d. .55

e. Let: A = event under 30
B = event single

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{.55}{.65} = .8462$$

f. $P(A \cap B) = .55$

$P(A)P(B) = (.65)(.75) = .49$

Since $P(A \cap B) \neq P(A)P(B)$, they cannot be independent events; or, since $P(A \mid B) \neq P(B)$, they cannot be independent.

24. a. P(holds) = 50/120

b. P(wins) = 24/120

c. P(wins|he held) = 14/50

d. P(held and lost) = 36/120

e. P(held|won) = 14/24

25 a. P(Nothing paid) = 67/1047 = 0.064

b. P(Female|Part paid) = 7/69 = .101

26. a. $P(A \cap B) = P(A)P(B) = (.55)(.35) = .19$

b. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = .55 + .35 - .19 = .71$

c. $P(\text{shutdown}) = 1 - P(A \cup B) = 1 - .71 = .29$

27. a. Yes, since $P(A_1 \cap A_2) = 0$

b. $P(A_1 \cap B) = P(A_1)P(B \mid A_1) = .40(.20) = .08$

$P(A_2 \cap B) = P(A_2)P(B \mid A_2) = .60(.05) = .03$

c. $P(B) = P(A_1 \cap B) + P(A_2 \cap B) = .08 + .03 = .11$

d. $P(A_1 \mid B) = \frac{.08}{.11} = .7273$

$P(A_2 \mid B) = \frac{.03}{.11} = .2727$

28. a. $P(B \cap A_1) = P(A_1)P(B \mid A_1) = (.20)(.50) = .10$

$P(B \cap A_2) = P(A_2)P(B \mid A_2) = (.50)(.40) = .20$

$P(B \cap A_3) = P(A_3)P(B \mid A_3) = (.30)(.30) = .09$

b. $P(A_2 \mid B) = \frac{.20}{.10 + .20 + .09} = .51$

c.

Events	$P(A_i)$	$P(B \mid A_i)$	$P(A_i \cap B)$	$P(A_i \mid B)$
A_1	.20	.50	.10	.26
A_2	.50	.40	.20	.51
A_3	.30	.30	.09	.23
	1.00		.39	1.00

29. Total defectives on day shift =.2000 *2/100 = 40

Total defectives on night shift =.800 *4/100 = 32

Total defectives = 40 +32 = 72

P(Defective) = P(Defective|Day shift)*P(Day shift) + P(Defective|Night shift)*P(Night shift)

$$= 0.2*\frac{2000}{2800} + 0.4*\frac{800}{2800} = \frac{72}{2800}$$

a. P(Night shift ∩ Defective) = 32/2800 / (72/2800) = 32/72

30. D = default

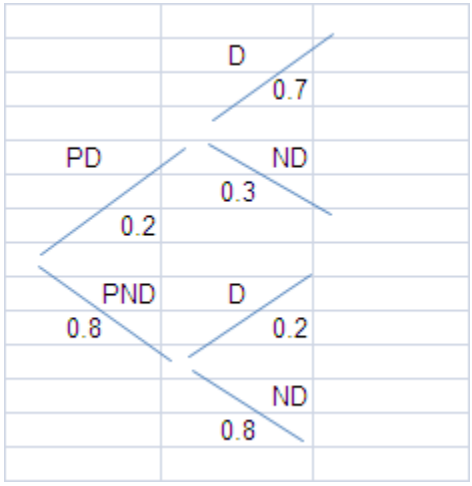
ND = Not default

PD = Predict default

PND = Predict not default

$P(PD) = 0.2$, $P(D|PD) = 0.7$, $P(D|PND) = 0.2$

a.



We require P(ND) which from the above probability tree = 0.3x0.2+0.8x0.8 = 0.7

= P(ND|PD)P(PD) +P(ND|PND)P(PND)

31.

	Fatalities	Serious injuries	Total
Children	60	2352	2412
Adults	1841	20770	22611
Total	1901	23122	25023

a. P(Serious injury|Child) = 2352/2412

b. P(Adult|Fatality) = 1841/1901

32. a. 0.46

b. 12/30 = 0.4

33. P(3 successes from 10 recommendations) is the Binomial probability with x = 3, n = 10 and success probability π = .8 for an excellent analyst, .6 for a good analyst and .4 for a poor analyst (see Chapter 5 for details of the Binomial calculation.) Thus we find:

P(3 |excellent analyst) = .000708

P(3 |good analyst) = .003822

P(3 |poor analyst) = .021499

Hence P(3 successes) = P(excellent rating) P(3 |excellent analyst) + P(good rating) P(3 |good analyst)

+ P(poor rating) P(3 |poor analyst) = (.9)(.000708) + (.09) (.003822) + (.01)(.021499) = .026029

a. P(excellent analyst| 3 successes)

= P(excellent analyst and 3 successes) / p(3 successes) = (.9)(.000708) / .026029 = .027

b. Similarly P(good analyst| 3 successes) = (.09) (.003822)/ .026029 = .147

c. P(poor analyst| 3 successes) = (.01)(.021499)/.026029 = .826

34. Probability (defective) = 0.01 for lines 2, 3 and 4. Thus using the Binomial distribution (see Chapter 5 for calculation details based on n = 5, π = .01)

P(2 defectives from a sample of 5| line i) = .00097. (i = 2,3,4)

Similarly when Probability (defect) = 0.1 for line 1, using the Binomial distribution again:

P(2 defectives from a sample of 5| line 1) = .0729.

Hence P(2 defectives from 5) = P(2 defectives from a sample of 5| line 1)p(line 1) +

P(2 defectives from a sample of 5| line 2)p(line 2) +

P(2 defectives from a sample of 5| line 3)p(line 3) +

P(2 defectives from a sample of 5| line 4)p(line 4)

= (.0729)(.25) + (.00097)(.25) + (.00097)(.25) + (.00097)(.25)

= .01895

Therefore, using Bayes theorem, P(lot came from line 1| 2 defectives from 5)

= P(2 defectives from a sample of 5 from line 1)/ P(2 defectives from 5) = (.0729)(.25)/ .01895 = .961

Supplementary Exercises:

35. A financial manager made two new investments—one in the oil industry and one in municipal bonds. After a one-year period, each of the investments will be classified as either successful or unsuccessful. Consider the making of the two investments as an experiment.

    a. How many sample points exist for this experiment?
    b. Show a tree diagram and list the sample points.
    c. Let O = the event that the oil industry investment is successful and M = the event that the municipal bond investment is successful. List the sample points in O and in M.
    d. List the sample points in the union of the events ($O \cup M$).
    e. List the sample points in the intersection of the events ($O \cap M$).
    f. Are events O and M mutually exclusive? Explain.

36. A telephone survey to determine viewer response to a new television show obtained the following data.

Rating	**Frequency**
Poor	4
Below Average	8
Average	11
Above Average	14
Excellent	13

    a. What is the probability that a randomly selected viewer will rate the new show as average or better?
    b. What is the probability that a randomly selected viewer will rate the new show below average or worse?

37. A survey of new MBA students provided the following background data for 2018 students.

		Applied to More Than One School	
		Yes	**No**
	23 and under	207	201
Age	**24-26**	299	379
Group	**27-30**	185	268
	31-35	66	193
	36 and over	51	169

a. For a randomly selected MBA student, prepare a joint probability table for the experiment consisting of observing the student's age and whether the student applied to one or more schools.

b. What is the probability that a randomly selected applicant is 23 or under?

c. What is the probability that a randomly selected applicant is older than 26?

d. What is the probability that a randomly selected applicant applied to more than one school?

38. Refer again to the survey data in exercise 35.

a. Given that a person applied to more than one school, what is the probability that the person is 24–26 years old?

b. Given that a person is in the 36-and-over age group, what is the probability that the person applied to more than one school?

c. What is the probability that a person is 24–26 years old or applied to more than one school?

d. Suppose a person is known to have applied to only one school. What is the probability that the person is 31 or more years old?

e. Is the number of schools applied to independent of age? Explain.

39. A large consumer goods company ran a television advertisement for one of its soap products. On the basis of a survey that was conducted, probabilities were assigned to the following events.

B = individual purchased the product

S = individual recalls seeing the advertisement

$B \cap S$ = individual purchased the product and recalls seeing the advertisement

The probabilities assigned were $P(B)$ =0. 20, $P(S)$ = 0.40, and $P(B \cap S)$ = 0.12.

a. What is the probability of an individual's purchasing the product given that the individual recalls seeing the advertisement? Does seeing the advertisement increase the probability that the individual will purchase the product? As a decision maker, would you recommend continuing the advertisement (assuming that the cost is reasonable)?

b. Assume that individuals who do not purchase the company's soap product buy from its competitors. What would be your estimate of the company's market share? Would you expect that continuing the advertisement will increase the company's market share? Why or why not?

c. The company also tested another advertisement and assigned it values of $P(S)$ = 0.30 and $P(B \cap S)$ = 0.10. What is $P(B \mid S)$ for this other advertisement? Which advertisement seems to have had the bigger effect on customer purchases?

40. A company studied the number of lost-time accidents occurring at its Faro, Portugal, plant. Historical records show that 6% of the employees suffered lost-time accidents last year. Management believes that a special safety programme will reduce such accidents to 5% during the current year. In addition, it estimates that 15% of employees who had lost-time accidents last year will experience a lost-time accident during the current year.

a. What percentage of the employees will experience lost-time accidents in both years?

b. What percentage of the employees will suffer at least one lost-time accident over the two-year period?

41. An oil company purchased an option on land in Ireland. Preliminary geologic studies assigned the following prior probabilities.

P(high-quality oil) = 0.50

P(medium-quality oil) = 0.20

P(no oil) = 0.30

a. What is the probability of finding oil?

b. After 50 metres of drilling on the first well, a soil test is taken. The probabilities of finding the particular type of soil identified by the test follow.

P(soil | high-quality oil) = 0.50

P(soil | medium-quality oil) = 0.80

$P(\text{soil} \mid \text{no oil}) = 0.20$

How should the firm interpret the soil test? What are the revised probabilities, and what is the new probability of finding oil?

42. A clothes store accepts payment of goods by cash or credit card. Most of its business (70%) is by credit card. Twenty five per cent is for household items, 16% for sportswear, 18% for men's wear and the rest for women's wear. Four percent of women's wear are bought with cash compared to 7% of men's wear and 12% of sports wear. If a sale is selected at random, then what is the probability that it is

    a. by credit card or for women's wear?
    b. for men's wear and not by credit card?
    c. for sports wear given it was for cash?
    d. paid cash given it was for household?

43. A company plans to select a team of five students for a business game competition from a pool of 18 undergraduates. Nine are from the second year management course, five are third year management and the remainder are from outside the management school. What is the probability that:

    a. All five team members are second year management?
    b. No students from outside the management school are selected?

44. A company is about to sell to a new client. It knows from past experience that there is a real possibility that the client may default on payment. As a precaution the company checks with a consultant on the likelihood of the client defaulting in this case and is given an estimate of 20%. Sometimes the consultant gets it wrong. Your own experience of the consultant is that he is correct 70% of the time when he predicts that the client will default but that 20% of clients who he believes will not default actually do. What is the probability that the new client will not default?

45. A firm has agreed to accept delivery of a component shipped in batches of 1000 if, in a random sample of ten, no more than one component is defective. What is the probability that this acceptance sampling rule would lead to a batch with 20 per cent defective items being accepted?

46. Records show that for every 100 items produced in a factory during the day shift, two are defective and for every 100 items produced during the night shift, four are defective. If during a 24 hour period, 2000 items are produced during the day and 800 at night, what is the probability that an item picked at random from the output over 24 hours came from the night shift if it was defective?

47. A sales representative finds that the probability of making a sale on the first visit to a new client is 0.5. On the second visit, the probability of making a sale is 0.6 if a sale was made on the first visit and 0.4 if no sale was made on the first visit.

    a. What is the probability of just one sale resulting from the two visits?
    b. How many sales would be expected?

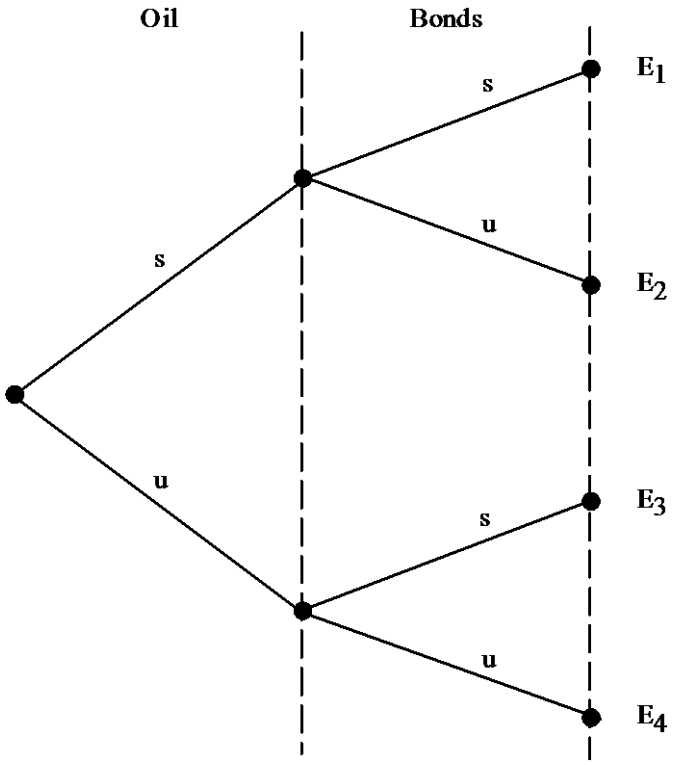
48. It takes at least 9 votes from a 12 member jury to convict a defendant. Suppose that the probability that a juror votes a guilty person innocent is 0.2 whereas the probability that the juror votes an innocent person guilty is 0.1.
    a. If each juror acts independently and 65% of defendants are guilty, what is the probability that the jury renders a correct decision.
    b. What percentage of defendants is convicted?

49. Ten second year statistics students are to be divided into an A team and a B team of 5 each. The A team play in one heat and the B team in another. How many different divisions of the ten students into the two teams are possible? Suppose two particular students insisted on being in the same team. How would this affect your answer?

Supplementary Exercises Solutions:

35. a. (2) (2) = 4

b. Let s = successful
u = unsuccessful



c. O = $\{E_1, E_2\}$

M = $\{E_1, E_3\}$

d. O $\cup$ M = $\{E_1, E_2, E_3\}$

e. O $\cap$ M = $\{E_1\}$

f. No; since O $\cap$ M has a sample point.

36. a. Probability of the event = P(average) + P(above average) + P(excellent)

$$= \frac{11}{50} + \frac{14}{50} + \frac{13}{50} = .22 + .28 + .26 = .76$$

b. Probability of the event = P(poor) + P(below average)

$$= \frac{4}{50} + \frac{8}{50} = .24$$

37. a.

	Yes	No	Total
23 and Under	.1026	.0996	.2022
24 - 26	.1482	.1878	.3360
27 - 30	.0917	.1328	.2245
31 - 35	.0327	.0956	.1283
36 and Over	.0253	.0837	.1090
Total	.4005	.5995	1.0000

b. .2022

c. .2245 + .1283 + .1090 = .4618

d. .4005

38. a. P(24 to 26 | Yes) = .1482 / .4005 = .3700

b. P(Yes | 36 and over) = .0253 / .1090 = .2321

c. .1026 + .1482 + .1878 + .0917 + .0327 + .0253 = .5883

d. P(31 or more | No) = (.0956 + .0837) / .5995 = .2991

e. No, because the conditional probabilities do not all equal the marginal probabilities. For instance,

$$P(24 \text{ to } 26 \mid \text{Yes}) = .3700 \neq P(24 \text{ to } 26) = .3360$$

39. a. $P(B|S) = \dfrac{P(B \cap S)}{P(S)} = \dfrac{.12}{.40} = .30$

We have P(B | S) > P(B).

Yes, continue the ad since it increases the probability of a purchase.

b. Estimate the company's market share at 20%. Continuing the advertisement should increase the market share since P(B | S) = .30.

c. $P(B|S) = \dfrac{P(B \cap S)}{P(S)} = \dfrac{.10}{.30} = .333$

The second ad has a bigger effect.

40. Let A = lost time accident in current year
B = lost time accident previous year

Given: P(B) = .06, P(A) = .05, P(A | B) = .15

a. $P(A \cap B) = P(A \mid B)P(B) = .15(.06) = .009$

b. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = .06 + .05 - .009 = .101$ or 10.1%

41. a. P(Oil) = .50 + .20 = .70

b. Let S = Soil test results

Events	$P(A_i)$	$P(S \mid A_i)$	$P(A_i \cap S)$	$P(A_i \mid S)$
High Quality (A_1)	.50	.20	.10	.31
Medium Quality (A_2)	.20	.80	.16	.50
No Oil (A_3)	.30	.20	.06	.19
	1.00		P(S) = .32	1.00

P(Oil) = .81 which is good; however, probabilities now favor medium quality rather than high quality oil.

42.

Joint probability distribution

	H	**S**	**MW**	**WW**	**All**
CC	0.18	0.04	0.11	0.37	**0.7**
C	0.07	**0.12**	**0.07**	**0.04**	0.3
All	**0.25**	**0.16**	**0.18**	0.41	**1**

where			
	CC=	credit card	Bold figures are given
	C=	cash	Non-bold have to be derived
	H=	household	
	S=	sportswear	
	MW=	mens wear	
	WW=	womens wear	

.

a. $P(CC \cup WW) = P(CC) + P(WW) - P(CC \cap WW) = 0.7 + 0.41 - 0.37 = 0.74$

b. $P(MW \cap C) = 0.07$

c. $P(S|C) = .12/.3 = 0.40$

d. $P(C|H) = .07/.25 = 0.28$

43. $$p(5) = \frac{\binom{9}{5}\binom{9}{0}}{\binom{18}{5}} = \frac{\frac{9!}{5!4!}\frac{9!}{0!9!}}{\frac{18!}{5!13!}} = 0.014706$$

$$p(0) = \frac{\binom{4}{0}\binom{14}{5}}{\binom{18}{5}} = \frac{\frac{4!}{4!0!}\frac{14!}{5!13!}}{\frac{18!}{5!13!}} = 0.23366$$

44. Let D = Default
ND= Not Default
PD = Predict default (by consultant)
PND = Predict not default (by consultant)

Given P(D|PD) = 0.7, P(D|PND)= 0.2 and P(PD)=0.2

Then P(ND) = P(ND|PD)P(PD) + P(ND|PND)P(PND)
= 0.3(0.2) + 0.8(0.8) = 0.7

45. Let X = number of defectives in the sample of 10. Then X has a Binomial distribution based on n = 10 trials and success probability, $\pi = 0.2$.

We require P(accept lot) = P(X=0) + P(X=1) = 0.107 + 0.268 = 0.375

46. Let D = Defective
ND= Not Defective

Given P(D| Day) = 0.02, P(D| Night)= 0.04

Total numbers produced in a day by D and ND:

Shift	D	ND	Total
Day	40	1960	2000
Night	32	768	800
Total	72	2728	2800

P(Night shift |D) = 32/72 = 0.444

47. Let S = Sale
NS= No Sale

Given P(S on first visit) = 0.5, P(S on second visit | S on first visit)= 0.6 and P(S on second visit | NS on first visit)= 0.4.

P(1 sale) = P(sale on first visit and no sale on second visit) +
P(sale on first visit and no sale on second visit) = 0.5(0.4) + 0.5(0.4) = 0.4

Similarly P(no sale) = 0.5*0.6 = 0.3
P(2 sales) = 0.5*0.6 = 0.3

If X = number of sales then expected number of sales = E(X) = Σ XP(X) = 0(0.3) + 1(0.4) + 2(0.3) = 1

48. Given P(individual juror votes person innocent | person guilty) = 0.2
P(individual juror votes person guilty | person not guilty) = 0.1
P(person guilty) = 0.65

P(correct decision) = P(jury votes person guilty | person guilty) p(person guilty)+
P(jury votes person not guilty | person not guilty)p(person not guilty)

P(jury votes person guilty | person guilty) = P(X>=9) where X = number of jurors who vote person guilty

If the person is guilty then X has a Binomial distribution based on n = 12 trials and a success probability of π = 0.8. For this distribution it can be shown the probability P(X>= 9) = 0.794569.

Similarly P(jury votes person not guilty | person not guilty) = P(X<9) where X again has a Binomial distribution based on n = 12 trials and a success probability of π = 0.9. For this distribution it can be shown P(X<9) = 0.999999834

Thus P(correct decision) = 0.794569(0.65) + 0.999999834(0.35) = 0.86647

49. $\binom{10}{5} = \frac{10!}{5!5!} = 252$

$\binom{8}{3} = \frac{8!}{3!5!} = 56$

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Five

Discrete Probability Distributions

Textbook Exercises (1-37)

Textbook Exercise Solutions

Supplementary Exercises (38-89)

Supplementary Exercise Solutions

Chapter 5: Discrete Probability Solutions

Textbook Exercises:

1 Consider the experiment of tossing a coin twice.

a. List the experimental outcomes.

b. Define a random variable that represents the number of heads occurring on the two tosses.

c. Show what value the random variable would assume for each of the experimental outcomes.

d. Is this random variable discrete or continuous?

2 Consider the experiment of a worker assembling a product.

a. Define a random variable that represents the time in minutes required to assemble the product.

b. What values may the random variable assume?

c. Is the random variable discrete or continuous?

3 Three students have interviews scheduled for summer employment. In each case the interview results in either an offer for a position or no offer. Experimental outcomes are defined in terms of the results of the three interviews.

a. List the experimental outcomes.

b. Define a random variable that represents the number of offers made. Is the random variable continuous?

c. Show the value of the random variable for each of the experimental outcomes.

4 Suppose we know home mortgage rates for 12 Danish lending institutions. Assume that the random variable of interest is the number of lending institutions in this group that offers a 30-year fixed rate of 1.5 per cent or less. What values may this random variable assume?

5 To perform a certain type of blood analysis, lab technicians must perform two procedures. The first procedure requires either 1 or 2 separate steps, and the second procedure requires either 1, 2 or 3 steps.

a. List the experimental outcomes associated with performing the blood analysis.

b. If the random variable of interest is the total number of steps required to do the complete analysis (both procedures), show what value the random variable will assume for each of the experimental outcomes.

6 Listed is a series of experiments and associated random variables. In each case, identify the values that the random variable can assume and state whether the random variable is discrete or continuous.

Experiment	Random variable (X)
a. Take a 20-question examination	Number of questions answered correctly
b. Observe cars arriving at a tollbooth for one hour	Number of cars arriving at tollbooth
c. Audit 50 tax returns	Number of returns containing errors
d. Observe an employee's work	Number of non-productive hours in an eight-hour workday
e. Weigh a shipment of goods	Number of kilograms

7 The probability distribution for the random variable X follows.

x	$p(x)$
20	0.20
25	0.15
30	0.25
35	0.40

a. Is this probability distribution valid? Explain.

b. What is the probability that X = 30?

c. What is the probability that X is less than or equal to 25?

d. What is the probability that X is greater than 30?

8 The following data were collected by counting the number of operating rooms in use at a general hospital over a 20-day period. On three of the days only one operating room was used, on five of the days two were used, on eight of the days three were used, and on four days all four of the hospital's operating rooms were used.

a. Use the relative frequency approach to construct a probability distribution for the number of operating rooms in use on any given day.
b. Draw a graph of the probability distribution.
c. Show that your probability distribution satisfies the required conditions for a valid discrete probability distribution.

9 The table below summarizes the joint probability distribution for the percentage monthly return for two ordinary shares 1 and 2. In the case of share 1, the % return X has historically been -1, 0 or 1. Correspondingly, for share 2, the % return Y has been -2, 0 or 2.

Table 5.4 Percent monthly return probabilities for shares 1 and 2

	% Monthly return	share 2 Y		
		-2	0	2
share 1 X	-1	0.1	0.1	0.0
	0	0.1	0.2	0.0
	1	0.0	0.1	0.4

a. Determine E(Y), E(X) Var(X) and Var(Y)
b. Determine the correlation coefficient between X and Y.
c. What do you deduce from b.?

10 A technician services mailing machines at companies in the Berne area. Depending on the type of malfunction, the service call can take 1, 2, 3 or 4 hours. The different types of malfunctions occur at about the same frequency.

a. Develop a probability distribution for the duration of a service call.
b. Draw a graph of the probability distribution.
c. Show that your probability distribution satisfies the conditions required for a discrete probability function.
d. What is the probability a service call will take three hours?
e. A service call has just come in, but the type of malfunction is unknown. It is 3:00 p.m. and service technicians usually get off at 5:00 p.m. What is the probability the service technician will have to work overtime to fi x the machine today?

11 A college admissions tutor subjectively assessed a probability distribution for X, the number of entering students, as follows.

x	$p(x)$
1000	0.15
1100	0.20
1200	0.30
1300	0.25
1400	0.10

a. Is this probability distribution valid? Explain.

b. What is the probability of 1200 or fewer entering students?

12 A psychologist determined that the number of sessions required to obtain the trust of a new patient is either 1, 2 or 3. Let X be a random variable indicating the number of sessions required to gain the patient's trust. The following probability function has been proposed.

$$p(x) = \frac{x}{6}$$

for x = 1, 2, or 3

a. Is this probability function valid? Explain.

b. What is the probability that it takes exactly two sessions to gain the patient's trust?

c. What is the probability that it takes at least two sessions to gain the patient's trust?

13 The following table is a partial probability distribution for the MRA Company's projected profits (X = profit in €'000s) for the first year of operation (the negative value denotes a loss).

x	$p(x)$
−100	0.10
0	0.20
50	0.30
100	0.25
150	0.10
200	

a. What is the proper value for p(200)? What is your interpretation of this value?

b. What is the probability that MRA will be profit table?

c. What is the probability that MRA will make at least €100 000?

14 The following table provides a probability distribution for the random variable X.

x	$p(x)$
3	0.25
6	0.50
9	0.25

a. Compute E(X), the expected value of X.

b. Compute σ 2, the variance of X.

c. Compute σ, the standard deviation of X.

15 The following table provides a probability distribution for the random variable Y.

y	$p(y)$
2	0.20
4	0.30
7	0.40
8	0.10

a. Compute E(Y).

b. Compute Var(Y) and σ.

16 A local ambulance service handles 0 to 5 service calls on any given day. The probability distribution for the number of service calls is as follows.

Number of service calls	Probability
0	0.10
1	0.15
2	0.30
3	0.20
4	0.15
5	0.10

a. What is the expected number of service calls?

b. What is the variance in the number of service calls? What is the standard deviation?

17 A certain machinist works an eight-hour shift. An efficiency expert wants to assess the value of this machinist where value is defined as value added minus the machinist's labour cost. The value added for the work the machinist does is €30 per item and the machinist earns €16 per hour. From past records, the machinist's output per shift is known to have the following probability distribution:

Output/shift	Probability
5	0.2
6	0.4
7	0.3
8	0.1

a. What is the expected monetary value of the machinist to the company per shift?

b. What is the corresponding variance value?

18 A company is contracted to finish a €100,000 project by 31 December. If it does not complete on time a penalty of €8,000 per month (or part of a month) is incurred. The company estimates that if it continues alone there will be a 40 per cent chance of completing on time and that the project may be one, two, three or four months late with equal probability.

Subcontractors can be hired by the firm at a cost of €18,000. If the subcontractors are hired then the probability that the company completes on time is doubled. If the project is still late it will now be only one or two months late with equal probability.

a. Determine the expected profit when
   (i) subcontractors are not used
   (ii) subcontractors are used

b. Which is the better option for the company?

19 The following probability distributions of job satisfaction scores for a sample of information systems (IS) senior executives and IS middle managers range from a low of 1 (very dissatisfied) to a high of 5 (very satisfied).

	Probability	
Job satisfaction score	IS senior executives	IS middle managers
1	0.05	0.04
2	0.09	0.10
3	0.03	0.12
4	0.42	0.46
5	0.41	0.28

a. What is the expected value of the job satisfaction score for senior executives?
b. What is the expected value of the job satisfaction score for middle managers?
c. Compute the variance of job satisfaction scores for executives and middle managers.
d. Compute the standard deviation of job satisfaction scores for both probability distributions.
e. Compare the overall job satisfaction of senior executives and middle managers.

20 The demand for a product of Cobh Industries varies greatly from month to month. The probability distribution in the following table, based on the past two years of data, shows the company's monthly demand.

Unit Demand	Probability
300	0.20
400	0.30
500	0.35
600	0.15

a. If the company bases monthly orders on the expected value of the monthly demand, what should Cobh's monthly order quantity be for this product?
b. Assume that each unit demanded generates €70 in revenue and that each unit ordered costs €50. How much will the company gain or lose in a month if it places an order based on your answer to part (a) and the actual demand for the item is 300 units?

21 Consider a binomial experiment with two trials and $\pi = 0.4$.

a. Draw a tree diagram for this experiment (see Figure 5.3).

b. Compute the probability of one success, p(1).

c. Compute p(0).

d. Compute p(2).

e. Compute the probability of at least one success.

f. Compute the expected value, variance, and standard deviation.

22 Consider a binomial experiment with n = 10 and $\pi = 0.10$.

a. Compute p(0).

b. Compute p(2).

c. Compute $P(x \leq 2)$.

d. Compute $P(x \geq 1)$.

e. Compute E(X).

f. Compute Var(X) and σ.

23 Consider a binomial experiment with n = 20 and $\pi = 0.70$.

a. Compute p(12).

b. Compute p(16).

c. Compute $P(X \geq 16)$.

d. Compute $P(X \leq 15)$.

e. Compute E(X).

f. Compute Var(X) and σ.

24 When a new machine is functioning properly, only 3 per cent of the items produced are defective. Assume that we will randomly select two parts produced on the machine and that we are interested in the number of defective parts found.

a. Describe the conditions under which this situation would be a binomial experiment.

b. Draw a tree diagram similar to Figure 5.3 showing this problem as a two-trial experiment.

c. How many experimental outcomes result in exactly one defect being found?

d. Compute the probabilities associated with finding no defects, exactly one defect, and two defects.

25 It takes at least 9 votes from a 12 member jury to convict a defendant. Suppose that the probability that a juror votes a guilty person innocent is 0.2 whereas the probability that the juror votes an innocent person guilty is 0.1.

a. If each juror acts independently and 65% of defendants are guilty, what is the probability that the jury renders a correct decision.

b. What percentage of defendants is convicted?

26 A firm bills its accounts at a 1 per cent discount for payment within ten days and the full amount is due after ten days. In the past 30 per cent of all invoices have been paid within ten days. If the firm sends out eight invoices during the first week of January, what is the probability that:

a. No one receives the discount?

b. Everyone receives the discount?

c. No more than three receive the discount?

d. At least two receive the discount?

27 In a game of 'Chuck a luck' a player bets on one of the numbers 1 to 6. Three dice are then rolled and if the number bet by the player appears i times (i=1,2,3) the player then wins i units. On the other hand if the number bet by the player does not appear on any of the dice the player loses 1 unit. a. If x is the players' winnings in the game, what is the expected value of X?

28 Consider a Poisson distribution with $\mu = 3$.

a. Write the appropriate Poisson probability function.

b. Compute p(2).

c. Compute p(1).

d. Compute $P(X \geq 2)$.

29 Consider a Poisson distribution with a mean of two occurrences per time period.

a. Write the appropriate Poisson probability function.

b. What is the expected number of occurrences in three time periods?

c. Write the appropriate Poisson probability function to determine the probability of x occurrences in three time periods.

d. Compute the probability of two occurrences in one time period.

e. Compute the probability of six occurrences in three time periods.

f. Compute the probability of fi ve occurrences in two time periods.

30 A certain process produces 100 m long rolls of high quality silk. In order to assess quality a 10 m sample is taken from the end of each roll and inspected for blemishes. The number of blemishes in each sample is thought to follow a Poisson distribution with an average of 2 blemishes per 10 m sample.

31 During the period of time that a local university takes phone-in registrations, calls come in at the rate of one every two minutes.

a. What is the expected number of calls in one hour?

b. What is the probability of three calls in five minutes?

c. What is the probability of no calls in a five-minute period?

32 Airline passengers arrive randomly and independently at the passenger-screening facility at a major international airport. The mean arrival rate is ten passengers per minute.

a. Compute the probability of no arrivals in a one-minute period.

b. Compute the probability that three or fewer passengers arrive in a one-minute period.

c. Compute the probability of no arrivals in a 15-second period.

d. Compute the probability of at least one arrival in a 15-second period.

33 Suppose N = 10 and r = 3. Compute the hypergeometric probabilities for the following values of n and x.

a. n = 4, x = 1.

b. n = 2, x = 2.

c. n = 2, x = 0.

d. n = 4, x = 2.

34 Suppose N = 15 and r = 4. What is the probability of x = 3 for n = 10?

35 Blackjack, or Twenty-one as it is frequently called, is a popular gambling game played in Monte Carlo casinos. A player is dealt two cards. Face cards (jacks, queens and kings) and tens have a point value of ten. Aces have a point value of one or 11. A 52-card deck contains 16 cards with a point value of ten (jacks, queens, kings and tens) and four aces.

a. What is the probability that both cards dealt are aces or ten-point cards?

b. What is the probability that both of the cards are aces?

c. What is the probability that both of the cards have a point value of ten?

d. A blackjack is a ten-point card and an ace for a value of 21. Use your answers to parts (a), (b) and (c) to determine the probability that a player is dealt blackjack. (Hint: Part (d) is not a hypergeometric problem. Develop your own logical relationship as to how the hypergeometric probabilities from parts (a), (b) and (c) can be combined to answer this question.)

36 A company plans to select a team of five students from Gulf university for a business game competition from a pool of 18 undergraduates. Nine are from the second year management course, five are third year management and the remainder are from outside the management school. What is the probability that:

a. All five team members are second year management?

b. No students from outside the management school are selected?

37 Manufactured parts are shipped in lots of 15 items. Four parts are randomly drawn from each lot and tested and the lot is considered acceptable if no defectives are among the four tested.

a. What is the probability that the shipment will be rejected?

Chapter 5: Discrete Probability Solutions

Textbook Exercises Solutions:

Solutions:

1. a. Head, Head (H,H)
   Head, Tail (H,T)
   Tail, Head (T,H)
   Tail, Tail (T,T)

   b. X = number of heads on two coin tosses

   c.

Outcome	Values of *X*
(H,H)	2
(H,T)	1
(T,H)	1
(T,T)	0

   d. Discrete. It may assume 3 values: 0, 1, and 2.

2. a. Let *x* = time (in minutes) to assemble the product.

   b. It may assume any positive value: X > 0.

   c. Continuous

3. Let Y = position is offered
   N = position is not offered

   a. *S* = {(Y,Y,Y), (Y,Y,N), (Y,N,Y), (Y,N,N), (N,Y,Y), (N,Y,N), (N,N,Y), (N,N,N)}

   b. Let N = number of offers made; N is a discrete random variable.

   c.

Experimental Outcome	(Y,Y,Y)	(Y,Y,N)	(Y,N,Y)	(Y,N,N)	(N,Y,Y)	(N,Y,N)	(N,N,Y)	(N,N,N)
Value of N	3	2	2	1	2	1	1	0

4. X = 0, 1, 2, . . ., 12.

5. a. S = {(1,1), (1,2), (1,3), (2,1), (2,2), (2,3)}

b.

Experimental Outcome	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)
Number of Steps Required	2	3	4	3	4	5

6. a. values: 0,1,2,...,20
discrete

b. values: 0,1,2,...
discrete

c. values: 0,1,2,...,50
discrete

d. values: $0 \leq x \leq 8$
continuous

e. values: $x > 0$
continuous

7. a. $p(x) \geq 0$ for all values of x.

$\Sigma p(x) = 1$ Therefore, it is a proper probability distribution.

b. Probability X = 30 is $p(30) = .25$

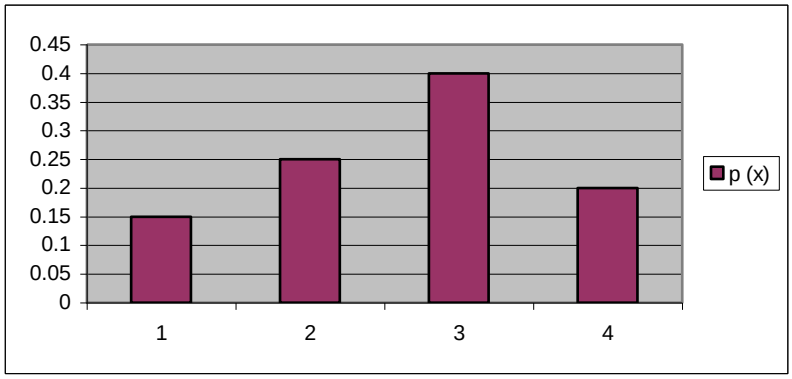
c. Probability $X \leq 25$ is $p(20) + p(25) = .20 + .15 = .35$

d. Probability $X > 30$ is $p(35) = .40$

8. a.

x	$p(x)$
1	3/20 = .15
2	5/20 = .25
3	8/20 = .40
4	4/20 = .20
	Total 1.00

b.



c. $p(x) \geq 0$ for $x = 1,2,3,4$.

$\Sigma p(x) = 1$

9. a. E(X) = 0.3, E(Y) = 0.4, Var(X) = 0.61, Var(Y)= 2.24

b. correlation coefficient = 0.753

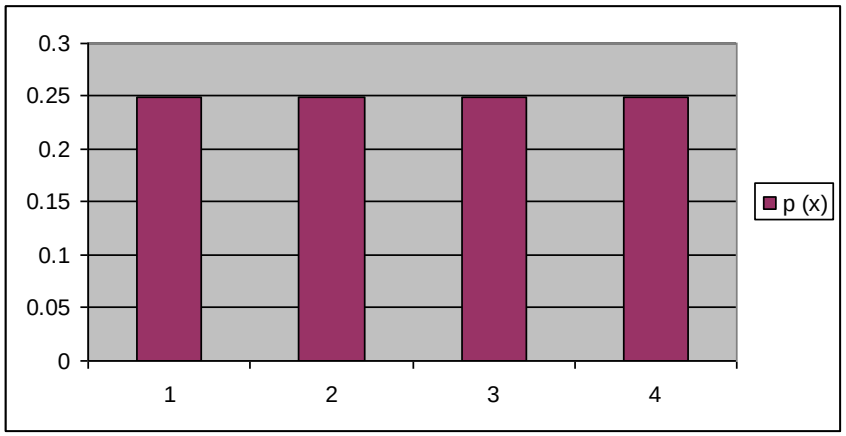
c. X and Y are highly correlated.

10. a.

Duration of Call

x	$p(x)$
1	0.25
2	0.25
3	0.25
4	0.25
	1.00

b.



c. $p(x) \geq 0$ and $p(1) + p(2) + p(3) + p(4) = 0.25 + 0.25 + 0.25 + 0.25 = 1.00$

d. $p(3) = 0.25$

e. P(overtime) $= p(3) + p(4) = 0.25 + 0.25 = 0.50$

11. a. Yes; $p(x) \geq 0$ for all x and $\Sigma p(x) = .15 + .20 + .30 + .25 + .10 = 1$

b. P(1200 or less) $= p(1000) + p(1100) + p(1200)$

$= .15 + .20 + .30 = .65$

12. a. Yes, since $p(x) \geq 0$ for $x = 1,2,3$ and $\Sigma p(x) = p(1) + p(2) + p(3) = 1/6 + 2/6 + 3/6 = 1$

b. $p(2) = 2/6 = .333$

c. $p(2) + p(3) = 2/6 + 3/6 = .833$

13. a. $p(200) = 1 - p(-100) - p(0) - p(50) - p(100) - p(150)$

$= 1 - .95 = .05$

This is the probability MRA will have a €200,000 profit.

b. P(Profit) $= p(50) + p(100) + p(150) + p(200)$

$= .30 + .25 + .10 + .05 = .70$

c. P(at least 100) $= p(100) + p(150) + p(200)$

$= .25 + .10 + .05 = .40$

14. a.

x	$p(x)$	$x\,p(x)$
3	.25	.75
6	.50	3.00
9	.25	2.25
	1.00	6.00

$E(X) = \mu = 6.00$

b.

x	$x - \mu$	$(x - \mu)^2$	p (x)	$(x - \mu)^2$ p (x)
3	-3	9	.25	2.25
6	0	0	.50	0.00
9	3	9	.25	2.25
				4.50

Var (X) = σ^2 = 4.50

c. $\sigma = \sqrt{4.50} = 2.12$

15. a.

y	p (y)	y p (y)
2	.20	.40
4	.30	1.20
7	.40	2.80
8	.10	.80
	1.00	5.20

E(Y) = μ = 5.20

b.

y	$y - \mu$	$(y - \mu)^2$	p (y)	$(y - \mu)^2$ p (y)
2	-3.20	10.24	.20	2.048
4	-1.20	1.44	.30	.432
7	1.80	3.24	.40	1.296
8	2.80	7.84	.10	.784
				4.560

Var (Y) = 4.56 $\quad \sigma = \sqrt{4.56} = 2.14$

16. a. /b.

	Odds	implied probability	probability	winnings	
Phillipe Bois	1/1	0.5	0.455	1	
Gallante Effor	5/2	0.286	0.260	2.5	
Satin Noir	11/2	0.154	0.140	5.5	
Victoire Antheme	9/1	0.1	0.091	9	
Comme Rambleur	16/1	0.059	0.054	16	
TOTAL					1.098 1

For the (fractional) odds x/y here:

$$\text{the implied probability} = \frac{y}{(x+y)}$$

As the sum of all the implied probabilities shown is 1.098 this means that the bookmaker has a 9.8% 'edge'. To convert these implied probabilities to proper probabilities we divide each implied probability above by 1.098.

c. In the winnings column, the amount shown is what the punter would have received if a €1 wager had been made and the horse had won.

Expected (winnings) = 1*0.455 + 2.5*0.260 + …. + 16*0.054 = €3.55

Assuming the €1 wager is also returned to the punter, the bookmaker pays out €4.55.

Yes the bookmaker has collected €5 in wagers for the five horses so the bookmaker's profit is
€5 – 4.55 = €0.45

d. Nothing because now the bookmaker's edge would be zero.

17. a. If X = number of items produced per shift then

X	p(X)	Xp(X)	X - μ	$(X-\mu)^2$	$(X-\mu)^2p(X)$
5	0.2	1	-1.3	1.69	0.338
6	0.4	2.4	-0.3	0.09	0.036
7	0.3	2.1	0.7	0.49	0.147
8	0.1	0.8	1.7	2.89	0.289
		6.3			0.81

Thus $\mu = 6.3$ and $\sigma^2 = 0.81$

The operator's monetary value (€) per shift = 30*X - 16*8 = 30X -128.
So the operator's expected monetary value (€) = E(30X -128)= 30E(X) – 128
= 30*6.3 - 128 =61.

b. The variance of 30X – 130 is Var(30X -128) = 30^2Var(X) = 900*.81 = 729.

18. a. Subcontractors not used

Month	P(x)	x	xP(x)
1	0.15	8000	1200.000
2	0.15	16000	2400.000
3	0.15	24000	3600.000
4	0.15	32000	4800.000
		Total =	€12000.000 = *E* (*X*)

Subcontractors used

Month	P(x)	x	xP(x)
1	0.10	8000	800
2	0.10	16000	1600
		Total =	€2400= *E* (*X*)

Profit (Subcontractors not used) = €100,000 - €12,000 = €88,000

Profit (Subcontractors not used) = €100,000 - €18,000 – €2,400 = €79,600

b. Better not to use subcontractors from the profit results.

19. a. $E(X) = \Sigma x\,p(x) = 0.05(1) + 0.09(2) + 0.03(3) + 0.42(4) + 0.41(5) = 4.05$

b. $E(X) = \Sigma x\,p(x) = 0.04(1) + 0.10(2) + 0.12(3) + 0.46(4) + 0.28(5) = 3.84$

c. Executives: $\sigma^2 = \Sigma (x - \mu)^2 p(x) = 1.2475$

Middle Managers: $\sigma^2 = \Sigma (x - \mu)^2 p(x) = 1.1344$

d. Executives: $\sigma = 1.1169$

Middle Managers: $\sigma = 1.0651$

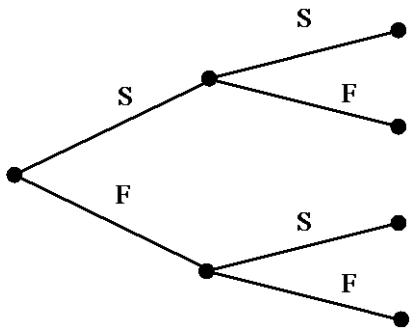
e. The senior executives have a higher average score: 4.05 vs. 3.84 for the middle managers. The executives also have a slightly higher standard deviation.

20. a. $E(X) = \Sigma x\,p(x) = 300\,(.20) + 400\,(.30) + 500\,(.35) + 600\,(.15) = 445$

The monthly order quantity should be 445 units.

b. Cost: 445 @ €50 = €22,250
Revenue: 300 @ €70 = 21,000
€ 1,250 Loss

21. a.



b. $p(1) = \binom{2}{1}(.4)^1(.6)^1 = \frac{2!}{1!1!}(.4)(.6) = .48$

c. $p(0) = \binom{2}{0}.4^0(.6)^2 = \frac{2!}{0!2!}(1)(.36) = .36$

d. $p(2) = \binom{2}{2}.4^2(.6)^0 = \frac{2!}{2!0!}(.16)(1) = .16$

e. $P(X \geq 1) = p(1) + p(2) = .48 + .16 = .64$

f. $E(X) = n\pi = 2(.4) = .8$

$\text{Var}(X) = n\pi(1 - \pi) = 2(.4)(.6) = .48$

$\sigma = \sqrt{.48} = .6928$

22. a. $p(0) = .3487$

b. $p(2) = .1937$

c. $P(X \leq 2) = p(0) + p(1) + p(2) = .3487 + .3874 + .1937 = .9298$

d. $P(X \geq 1) = 1 - p(0) = 1 - .3487 = .6513$

e. $E(X) = n\pi = 10(.1) = 1$

f. $\text{Var}(X) = n\pi(1 - \pi) = 10(.1)(.9) = .9$

$\sigma = \sqrt{.9} = .9487$

23. a. $p(12) = .1144$

b. $p(16) = .1304$

c. $P(X \geq 16) = p(16) + p(17) + p(18) + p(19) + p(20)$

$= .1304 + .0716 + .0278 + .0068 + .0008 = .2374$

d. $P(X \leq 15) = 1 - P(x \geq 16) = 1 - .2374 = .7626$

e. $E(X) = n\pi = 20(.7) = 14$

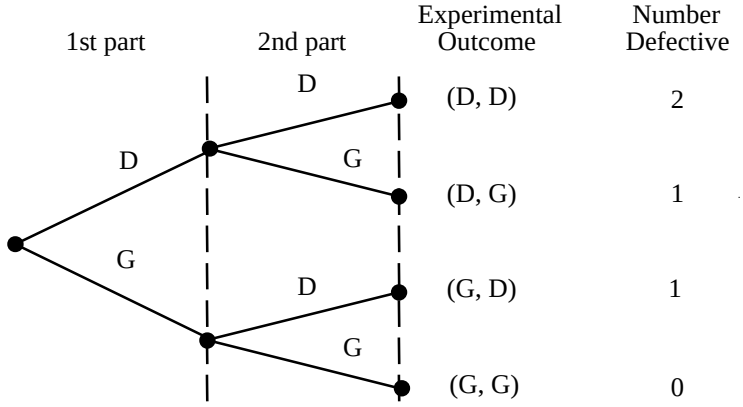
f. $\text{Var}(X) = n\pi(1 - \pi) = 20(.7)(.3) = 4.2$

$\sigma = \sqrt{4.2} = 2.0494$

24. a. Probability of a defective part being produced must be .03 for each part selected; parts must be selected independently.

b. Let:D = defective

G = not defective



c. 2 outcomes result in exactly one defect.

d. P (no defects) = (.97) (.97) = .9409

P (1 defect) = 2 (.03) (.97) = .0582

P (2 defects) = (.03) (.03) = .0009

25. a.

r	P(r vote to convict|guilty)	
9	0.236	
10	0.283	The probabilities here are B(n, π) = where n = 12 and π = 0.8
11	0.206	for r = 9…12
12	0.069	
Total	0.795 = P(Vote to convict|guilty)	

r	P(r vote to convict|innocent)	
9	0.000	
10	0.000	The probabilities here are B(n, π) = where n = 12 and π = 0.1
11	0.000	for r = 9…12
12	0.000	
Total	0.000 = P(Vote to convict|innocent)	

P(correct decision) = 0.65* P(Vote to convict|guilty) + 0.35*P(Vote not to convict|innocent)

$$= 0.65*0.795 + 0.35*(1-0) = 0.866$$

since P(Vote not to convict|innocent) = 1 - P(Vote to convict|innocent)

b. P(convicted) = 0.65* P(Vote to convict|guilty) + 0.35*P(Vote to convict|innocent)

$$= 0.65*0.795 + 0.35*0 = 0.516$$

26. $p(x) = \binom{n}{x}\pi^{x}(1-\pi)^{(n-x)}$

where n = 8, $\pi = 0.3$

a. p (0) = .0576

b. p (8) = .0001

c. $p(X \leq 3) = [\ p(0) + p(1) + p(2) + p(3)\ ] = .8059$

d. $p(X \geq 2) = 1 - p(0) - p(1) = .7447$

27.

a.

i	Payoff	P(Payoff) = Probability (B(3, 1/6)
0	-1	125/216
1	1	75/216
2	2	15/216
3	3	1/216

Expected payoff $= \Sigma$ Payoff * P(Payoff) = $\frac{-17}{216}$

28. a. $p(x) = \dfrac{3^x e^{-3}}{x!}$

b. $p(2) = \dfrac{3^2 e^{-3}}{2!} = \dfrac{9(.0498)}{2} = 0.2241$

c. $p(1) = \dfrac{3^1 e^{-3}}{1!} = 3(.0498) = 0.1494$

d. P (X $\geq$ 2) = 1 - p (0) - p (1) = 1 - .0498 - .1494 = .8008

29. a. $p(x) = \dfrac{2^x e^{-2}}{x!}$

b. μ = 6 for 3 time periods

c. $p(X) = \dfrac{6^x e^{-6}}{x!}$

d. $p(2) = \dfrac{2^2 e^{-2}}{2!} = \dfrac{4(.1353)}{2} = 0.2706$

e. $p(6) = \dfrac{6^6 e^{-6}}{6!} = 0.1606$

f. $p(5) = \dfrac{4^5 e^{-4}}{5!} = 0.1563$

30. a. n = number of blemishes in 30m roll. E(n) = 3*2 = 6 blemishes if μ = 2 for a 10m roll

n	Poisson (6)
0	0.002
1	0.015
2	0.045
3	0.089
4	0.134
5	0.161
6	0.161
7	0.138

Total 0.744 = P(n $\leq$ 7)

Hence P(n > 7) = 0.256

31. a. 30 per hour

b. $\mu = 1\,(5/2) = 5/2$

$$p(3) = \frac{(5/2)^3 e^{-5/2}}{3!} = 0.2138$$

c. $$p(0) = \frac{(5/2)^0 e^{-5/2}}{0!} = e^{-5/2} = 0.0821$$

32. a. $$p(0) = \frac{(10)^0 e^{-10}}{0!} = e^{-10} = 0.000045$$

b. p (0) + p (1) + p (2) + p (3)

p (0) = .000045 (part a)

$$p(1) = \frac{(10)^1 e^{-10}}{1!} = 0.00045$$

Similarly, p (2) = .00225, p (3) = .0075

and p (0) + p (1) + p (2) + p (3) = .010245

c. 2.5 arrivals / 15 sec. period Use $\mu = 2.5$

$$p(0) = \frac{(2.5)^0 e^{-2.5}}{0!} = e^{-5/2} = 0.0821$$

d. 1 - p (0) = 1 - .0821 = .9179

33. a. $$p(1) = \frac{\binom{3}{1}\binom{10-3}{4-1}}{\binom{10}{4}} = \frac{\frac{3!}{1!2!}\frac{7!}{3!4!}}{\frac{10!}{4!6!}} = \frac{3*35}{210} = 0.50$$

b. $$p(2) = \frac{\binom{3}{2}\binom{10-3}{2-2}}{\binom{10}{2}} = \frac{3*1}{45} = 0.067$$

c. $p(0) = \dfrac{\binom{3}{0}\binom{10-3}{2-0}}{\binom{10}{2}} = \dfrac{1*21}{45} = 0.4667$

d. $p(2) = \dfrac{\binom{3}{2}\binom{10-3}{4-2}}{\binom{10}{4}} = \dfrac{3*21}{210} = 0.30$

34. $p(3) = \dfrac{\binom{4}{3}\binom{15-4}{10-3}}{\binom{15}{10}} = \dfrac{4*330}{3003} = 0.4396$

35. Parts a, b & c involve the hypergeometric distribution with $N = 52$ and $n = 2$

a. $r = 20, x = 2$

$$p(2) = \frac{\binom{20}{2}\binom{32}{0}}{\binom{52}{2}} = \frac{190*1}{1326} = 0.1433$$

b. $r = 4, x = 2$

$$p(2) = \frac{\binom{4}{2}\binom{48}{0}}{\binom{52}{2}} = \frac{6*1}{1326} = 0.0045$$

c. $r = 16, x = 2$

$$p(2) = \frac{\binom{16}{2}\binom{36}{0}}{\binom{52}{2}} = \frac{120*1}{1326} = 0.0905$$

d. Part (a) provides the probability of blackjack plus the probability of 2 aces plus the probability of two 10s. To find the probability of blackjack we subtract the probabilities in (b) and (c) from the probability in (a).

P (blackjack) = .1433 - .0045 - .0905 = .0483

36. Hypergeometric distribution applies:

a. $$p(5) = \frac{\binom{9}{5}\binom{9}{0}}{\binom{18}{5}} = 0.014706$$

b. $$p(0) = \frac{\binom{4}{0}\binom{14}{5}}{\binom{18}{5}} = 0.23366$$

37. Hypergeometric distribution applies

a.

$$p(0) = \frac{\binom{4}{0}\binom{11}{4}}{\binom{15}{4}} = 0.241758$$

Supplementary Exercises:

38. Since the shipment is large we can assume that the probabilities do not change from trial to trial and use the binomial probability distribution.

a. $n = 5$

$$p(0) = \binom{5}{0}(0.01)^0(.0.99)^5 = 0.951$$

b.

$$p(1) = \binom{5}{1}(0.01)^1(.0.99)^4 = 0.0480$$

c. 1 - P (0) = 1 - .9510 = .0490

d. No, the probability of finding one or more items in the sample defective when only 1% of the items in the population are defective is small (only 0.0490). I would consider it likely that more than 1% of the items are defective.

39. $\mu = 15$

prob of 20 or more arrivals = p (20) + p (21) + · · ·

= .0418 + .0299 + .0204 + .0133 + .0083 + .0050 + .0029

+ .0016 + .0009 + .0004 + .0002 + .0001 + .0001 = .1249

40. $\mu = 1.5$

prob of 3 or more breakdowns is 1 - [p (0) + p (1) + p (2)].

1 - [p (0) + p (1) + p (2)]

= 1 - [.2231 + .3347 + .2510]

= 1 - .8088 = .1912

41. $\mu = 10$ p (4) = .0189

42. a. $p(3) = \frac{3^3 e^{-3}}{3!} = 0.2240$

b. $p(3) + p(4) + \cdots = 1 - [\, p(0) + p(1) + p(2) \,]$

$p(0) = \frac{3^0 e^{-3}}{0!} = e^{-3} = 0.0498$

Similarly, $p(1) = .1494$, $p(2) = .2240$

$\therefore\ 1 - [\ .0498 + .1494 + .2241\ ] = .5767$

43. Hypergeometric $N = 52$, $n = 5$ and $r = 4$.

a. $\dfrac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} = \dfrac{6(17296)}{2598960} = 0.0399$

b. $\dfrac{\binom{4}{1}\binom{48}{4}}{\binom{52}{5}} = \dfrac{4(194580)}{2598960} = 0.2995$

c. $\dfrac{\binom{4}{0}\binom{48}{5}}{\binom{52}{5}} = \dfrac{1712304}{2598960} = 0.6588$

d. $1 - p(0) = 1 - .6588 = .3412$

44. a. $\mu = 0.05(200) = 10$; $p(5) = \frac{10^5 e^{-10}}{5!} = 0.0378$

b. $p(>2) = 1 - p(0) - p(1) - p(2) = 0.997$ where $p(i) = \frac{10^i e^{-10}}{i!}$ $\quad i = 0, 1, 2$

c. $p(>=5) = 1 - p(0) - p(1) - p(2) - p(3) - p(4) = 0.971$

45. a. P(unprofitable) = p(0) + p(1) + p(2) = 0.604

where $p(i) = \binom{15}{i}(0.15)^{i}(.0.85)^{15-i}$

b. Based on the result of 0.604 for (a) the success rate would need to increase to approximately $\pi = 0.22$

(P(unprofitable) = p(0) + p(1) + p(2) = 0.617

where $p(i) = \binom{10}{i}(0.22)^{i}(.0.78)^{10-i}$)

46. $p(0) = \dfrac{\binom{5}{0}\binom{10}{4}}{\binom{15}{4}} = 0.1538$

47. P(not enough seats) = p(15) + p(16) = 0.284

where $p(i) = \binom{16}{i}(0.85)^{i}(.0.15)^{16-i}$

48. $\mu = 2(30/10) = 6$;

p(>7) = 1 - p(0) - p(1) - p(2) – p(3) – p(4) – p(5) – p(6) – p(7) = 0.256

where $p(i) = \dfrac{6^{i}e^{-6}}{i!}$ i = 0, 1, 2

49. a. 2.81

X	P(X)	XP(X)	$X^2P(X)$
0	0.02	0	0
1	0.09	0.09	0.09
2	0.28	0.56	1.12
3	0.33	0.99	2.97
4	0.24	0.96	3.84
5	0.03	0.15	0.75
6	0.01	0.06	0.36
Total	**1**	**2.81**	**9.13**
		=E(X)	**=E(X^2)**

b. $Var(x) = 9.13 - 2.81^2 = 1.2339$

SD(x) = sqrt(1.2339) = 1.111

50. P(at least 3 ill) = 1- p(0) - p(1) – p(2) = 0.127

where $p(i) = \binom{25}{i}(0.05)^{i}(.0.95)^{25-i}$

51. a. p(0) =0.033

b. p(2) =0.193

c. p(>=3) = 1 - p(0) - p(1) – p(2) = 0.660 where p(i) = $\frac{3.4^{i}e^{-3.4}}{i!}$ i = 0, 1, 2

52. Let X = number of defectives in the sample of 10. Then X has a Binomial distribution based on n = 10 trials and success probability, $\pi = 0.2$.

We require P(accept lot) = P(X=0) + P(X=1) = 0.107 + 0.268 = 0.375

Supplementary Exercises Solutions:

38. Since the shipment is large we can assume that the probabilities do not change from trial to trial and use the binomial probability distribution.

a. $n = 5$

$$p(0) = \binom{5}{0}(0.01)^0(.0.99)^5 = 0.951$$

b.

$$p(1) = \binom{5}{1}(0.01)^1(.0.99)^4 = 0.0480$$

c. 1 - P (0) = 1 - .9510 = .0490

d. No, the probability of finding one or more items in the sample defective when only 1% of the items in the population are defective is small (only 0.0490). I would consider it likely that more than 1% of the items are defective.

39. $\mu = 15$

prob of 20 or more arrivals = p (20) + p (21) + · · ·

= .0418 + .0299 + .0204 + .0133 + .0083 + .0050 + .0029

+ .0016 + .0009 + .0004 + .0002 + .0001 + .0001 = .1249

40. $\mu = 1.5$

prob of 3 or more breakdowns is 1 - [p (0) + p (1) + p (2)].

1 - [p (0) + p (1) + p (2)]

= 1 - [.2231 + .3347 + .2510]

= 1 - .8088 = .1912

41. $\mu = 10$ p (4) = .0189

42. a. $p(3) = \frac{3^3 e^{-3}}{3!} = 0.2240$

b. p (3) + p (4) + · · · = 1 - [p (0) + p (1) + p (2)]

$p(0) = \frac{3^0 e^{-3}}{0!} = e^{-3} = 0.0498$

Similarly, p (1) = .1494, p (2) = .2240

$\therefore$ 1 - [.0498 + .1494 + .2241] = .5767

43. Hypergeometric $N = 52$, $n = 5$ and $r = 4$.

a. $\frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} = \frac{6(17296)}{2598960} = 0.0399$

b. $\frac{\binom{4}{1}\binom{48}{4}}{\binom{52}{5}} = \frac{4(194580)}{2598960} = 0.2995$

c. $\frac{\binom{4}{0}\binom{48}{5}}{\binom{52}{5}} = \frac{1712304}{2598960} = 0.6588$

e. 1 - p (0) = 1 - .6588 = .3412

44. a. $\mu = 0.05(200) = 10$; $p(5) = \frac{10^5 e^{-10}}{5!} = 0.0378$

b. p(>2) = 1 - p(0) - p(1) - p(2) = 0.997 where $p(i) = \frac{10^i e^{-10}}{i!}$ i = 0, 1, 2

c. p(> = 5) = 1 - p(0) - p(1) - p(2) - p(3) - p(4) = 0.971

45. a. P(unprofitable) = p(0) + p(1) + p(2) = 0.604

where $p(i) = \binom{15}{i}(0.15)^{i}(.0.85)^{15-i}$

b. Based on the result of 0.604 for (a) the success rate would need to increase to approximately $\pi = 0.22$

(P(unprofitable) = p(0) + p(1) + p(2) = 0.617

where $p(i) = \binom{10}{i}(0.22)^{i}(.0.78)^{10-i}$)

46. $p(0) = \dfrac{\binom{5}{0}\binom{10}{4}}{\binom{15}{4}} = 0.1538$

47. P(not enough seats) = p(15) + p(16) = 0.284

where $p(i) = \binom{16}{i}(0.85)^{i}(.0.15)^{16-i}$

48. $\mu = 2(30/10) = 6$;

p(>7) = 1 - p(0) - p(1) - p(2) – p(3) – p(4) – p(5) – p(6) – p(7) = 0.256

where $p(i) = \dfrac{6^{i}e^{-6}}{i!}$ i = 0, 1, 2

49. a. 2.81

X	P(X)	XP(X)	$X^2P(X)$
0	0.02	0	0
1	0.09	0.09	0.09
2	0.28	0.56	1.12
3	0.33	0.99	2.97
4	0.24	0.96	3.84
5	0.03	0.15	0.75
6	0.01	0.06	0.36
Total	**1**	**2.81**	**9.13**
		=E(X)	**=E(X^2)**

b. $Var(x) = 9.13 - 2.81^2 = 1.2339$

SD(x) = sqrt(1.2339) = 1.111

50. P(at least 3 ill) = 1- p(0) - p(1) – p(2) = 0.127

where $p(i) = \binom{25}{i}(0.05)^{i}(.0.95)^{25-i}$

51. a. p(0) =0.033

b. p(2) =0.193

c. p(>=3) = 1 - p(0) - p(1) – p(2) = 0.660 where p(i) = $\frac{3.4^{i}e^{-3.4}}{i!}$ i = 0, 1, 2

52. Let X = number of defectives in the sample of 10. Then X has a Binomial distribution based on n = 10 trials and success probability, $\pi = 0.2$.

We require P(accept lot) = P(X=0) + P(X=1) = 0.107 + 0.268 = 0.375

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Six

Continuous Probability Distributions

Textbook Exercises (1-27)

Textbook Exercise Solutions

Supplementary Exercises (28-42)

Supplementary Exercise Solutions

Chapter 6: Continuous Probability Solutions

Textbook Exercises:

1 The random variable X is known to be uniformly distributed between 1.0 and 1.5.

a. Show the graph of the probability density function.

b. Compute P(X = 1.25).

c. Compute P($1.0 \leq X \leq 1.25$).

d. Compute P($1.20 < X < 1.5$).

2 The random variable X is known to be uniformly distributed between 10 and 20.

a. Show the graph of the probability density function.

b. Compute P($X < 15$).

c. Compute P($12 \leq X \leq 18$).

d. Compute E(X).

e. Compute Var(X).

3 A continuous random variable X has probability density function:

$f(x) = kx \qquad 0 < x < 2$

$0 \quad \text{otherwise}$

a. Determine the value of k.

b. Find E(X)and Var(X).

c. What is the probability that X is greater than three standard deviations above the mean? Find the distribution function F(X) and hence the median of X.

4 Most computer languages include a function that can be used to generate random numbers.

In EXCEL, the RAND function can be used to generate random numbers between 0 and 1.

If we let X denote a random number generated using RAND, then X is a continuous random variable with the following probability density function.

$$f(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

a. Graph the probability density function.

b. What is the probability of generating a random number between 0.25 and 0.75?

c. What is the probability of generating a random number with a value less than or equal to 0.30?

d. What is the probability of generating a random number with a value greater than 0.60?

5 Let X denote the number of bricks, a bricklayer will lay in an hour and assume that X takes values in the range 150 to 200 inclusively with equal probability (i.e. has a discrete uniform distribution). If a certain project is 170 bricks short of completion and a further project is waiting to be started as soon as this one is finished, what is the probability that

a. The bricklayer will start the second project within the hour?

b. More than 25 bricks will have been laid on the second project at the end of the next hour?

c. The first project will be more than ten bricks short of completion at the end of the next hour?

d. The bricklayer will lay exactly 175 bricks during the next hour?

6 The label on a bottle of liquid detergent shows contents to be 12 grams per bottle. The production operation fills the bottle uniformly according to the following probability density function.

$$f(x) = \begin{cases} 8 & \text{for } 11.975 \le x \le 12.100 \\ 0 & \text{elsewhere} \end{cases}$$

a. What is the probability that a bottle will be filled with between 12 and 12.05 grams?

b. What is the probability that a bottle will be filled with 12.02 or more grams?

c. Quality control accepts a bottle that is filled to within 0.02 grams of the number of grams shown on the container label. What is the probability that a bottle of this liquid detergent will fail to meet the quality control standard?

7 Suppose we are interested in bidding on a piece of land and we know there is one other bidder. The seller announced that the highest bid in excess of €10 000 will be accepted. Assume that the competitor's bid X is a random variable that is uniformly distributed between €10 000 and €15 000.

a. Suppose you bid €12 000. What is the probability that your bid will be accepted?

b. Suppose you bid €14 000. What is the probability that your bid will be accepted?

c. What amount should you bid to maximize the probability that you get the property?

d. Suppose you know someone who is willing to pay you €16 000 for the property. Would you consider bidding less than the amount in part (c)? Why or why not?

8 Using Figure 6.4 as a guide, sketch a normal curve for a random variable X that has a mean of $\mu = 100$ and a standard deviation of $\sigma = 10$. Label the horizontal axis with values of 70, 80, 90, 100, 110, 120 and 130.

9 A random variable is normally distributed with a mean of $\mu = 50$ and a standard deviation of $\sigma = 5$.

a. Sketch a normal curve for the probability density function. Label the horizontal axis with values of 35, 40, 45, 50, 55, 60 and 65. Figure 6.4 shows that the normal curve almost touches the horizontal axis at three standard deviations below and at three standard deviations above the mean (in this case at 35 and 65).

b. What is the probability the random variable will assume a value between 45 and 55?

c. What is the probability the random variable will assume a value between 40 and 60?

10 Draw a graph for the standard normal distribution. Label the horizontal axis at values of −3, −2, −1, 0, 1, 2 and 3. Then use the table of probabilities for the standard normal distribution to compute the following probabilities.

a. $P(0 \leq Z \leq 1)$

b. $P(0 \leq Z \leq 1.5)$

c. $P(0 < Z < 2)$

d. $P(0 < Z < 2.5)$

11 Given that Z is a standard normal random variable, compute the following probabilities.

a. $P(-1 \leq Z \leq 0)$

b. $P(-1.5 \leq Z \leq 0)$

c. $P(-2 < Z < 0)$

d. $P(-2.5 \leq Z \leq 0)$

e. $P(-3 \leq Z \leq 0)$

12 Given that Z is a standard normal random variable, compute the following probabilities.

a. $P(0 \leq Z \leq 0.83)$

b. $P(-1.57 \leq Z \leq 0)$

c. $P(Z > 0.44)$

d. $P(Z \geq -0.23)$

e. $P(Z < 1.20)$

f. $P(Z \leq -0.71)$

13 Given that Z is a standard normal random variable, compute the following probabilities.

a. $P(-1.98 \leq Z \leq 0.49)$

b. $P(0.52 \leq Z \leq 1.22)$

c. $P(-1.75 \leq Z \leq -1.04)$

14 Given that Z is a standard normal random variable, find z for each situation.

a. The area between 0 and z is 0.4750.

b. The area between 0 and z is 0.2291.

c. The area to the right of z is 0.1314.

d. The area to the left of z is 0.6700.

15 Given that Z is a standard normal random variable, find z for each situation.

a. The area to the left of z is 0.2119.

b. The area between $-z$ and z is 0.9030.

c. The area between $-z$ and z is 0.2052.

d. The area to the left of z is 0.9948.

e. The area to the right of z is 0.6915.

16 Given that Z is a standard normal random variable, find z for each situation.

a. The area to the right of z is 0.01.

b. The area to the right of z is 0.025.

c. The area to the right of z is 0.05.

d. The area to the right of z is 0.10.

17 The Attendance at a rock concert is normally distributed with a mean of 28,000 persons and a standard deviation of 4000 persons. What is the probability, that:

a. more than 28000 persons will attend?

b. less than 14000 persons will attend?

c. between 17000 and 25000 persons will attend?

d. Suppose the number who actually attended was X and the probability of achieving this level of attendance or higher was found to be 5%. What is X?

18 The holdings of clients of a successful on-line stockbroker are normally distributed with a mean of £20 000 and standard deviation of £1 500. To increase its business, the stockbroker is looking to email special promotions to the top 20 per cent of its clientele based on the value of their holdings. What is the minimum holding of this group?

19 A company has been involved in developing a new pesticide. Tests show that the average proportion, p, of insects killed by administration of x units of the insecticide is given by $p = P(X \leq x)$ where the probability $P(X \leq x)$ relates to a normal distribution with unknown mean and standard deviation.

a. Given that $x = 10$ when $p = 0.4$ and that $x = 15$ when $p = 0.9$, determine the dose that will be lethal to 50 per cent of the insect population on average.

b. If a dose of 17.5 units is administered to each of 100 insects, how many will be expected to die?

20 A binomial probability distribution has $\pi = 0.20$ and $n = 100$.

a. What is the mean and standard deviation?

b. Is this a situation in which binomial probabilities can be approximated by the normal probability distribution? Explain.

c. What is the probability of exactly 24 successes?

d. What is the probability of 18 to 22 successes?

e. What is the probability of 15 or fewer successes?

21 Assume a binomial probability distribution has $\pi = 0.60$ and $n = 200$.

a. What is the mean and standard deviation?

b. Is this a situation in which binomial probabilities can be approximated by the normal probability distribution? Explain.

c. What is the probability of 100 to 110 successes?

d. What is the probability of 130 or more successes?

e. What is the advantage of using the normal probability distribution to approximate the binomial probabilities? Use part (d) to explain the advantage.

22 A hotel in Nice has 120 rooms. In the spring months, hotel room occupancy is approximately 75 per cent.

a. What is the probability that at least half of the rooms are occupied on a given day?

b. What is the probability that 100 or more rooms are occupied on a given day?

c. What is the probability that 80 or fewer rooms are occupied on a given day?

23 Consider the following exponential probability density function.

$$f(x) = \frac{1}{8} e^{-x/8} \qquad \text{for } x \geq 0$$

a. Find P(X ≤ 6).

b. Find P(X ≤ 4).

c. Find P(X ≥ 6).

d. Find P(4 ≤ X ≤ 6).

24 Consider the following exponential probability density function.

$$f(x) = \frac{1}{3} e^{-x/3} \qquad \text{for } x \geq 0$$

a. Write the formula for P(X ≤ x0).

b. Find P(X ≤ 2).

c. Find P(X ≥ 3).

d. Find P(X ≤ 5).

e. Find P(2 ≤ X ≤ 5).

25 In a parts store in Mumbai, customers arrive randomly. The cashier's service time is random but it is estimated it takes an average 30 seconds to serve each customer.

a. What is the probability a customer must wait more than 2 minutes for service?

b. Suppose average service time is reduced to 25 seconds. How does this affect the calculation for a. above?

26 The time between arrivals of vehicles at a particular intersection follows an exponential probability distribution with a mean of 12 seconds.

a. Sketch this exponential probability distribution.

b. What is the probability that the arrival time between vehicles is 12 seconds or less?

c. What is the probability that the arrival time between vehicles is 6 seconds or less?

d. What is the probability of 30 or more seconds between vehicle arrivals?

27 According to Barron's 1998 Primary Reader Survey, the average annual number of investment transactions for a subscriber is 30 (www.barronsmag.com, 28 July 2000). Suppose the number of transactions in a year follows the Poisson probability distribution.

a. Show the probability distribution for the time between investment transactions.

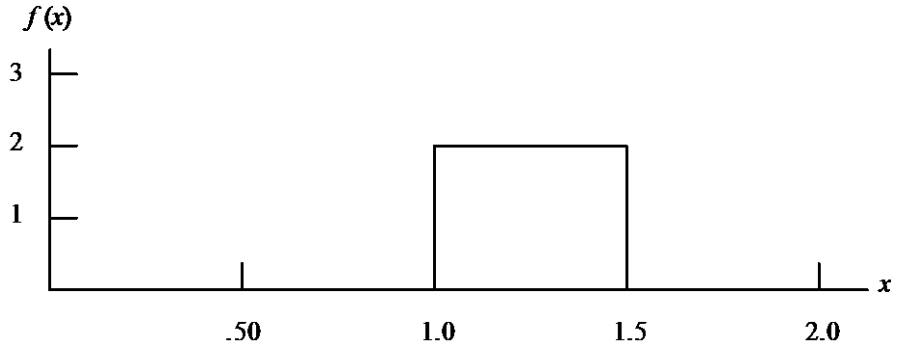
b. What is the probability of no transactions during the month of January for a particular subscriber?

c. What is the probability that the next transaction will occur within the next half month for a particular subscriber?

Chapter 6: Continuous Probability Solutions

Textbook Exercises Solutions:

1. a.

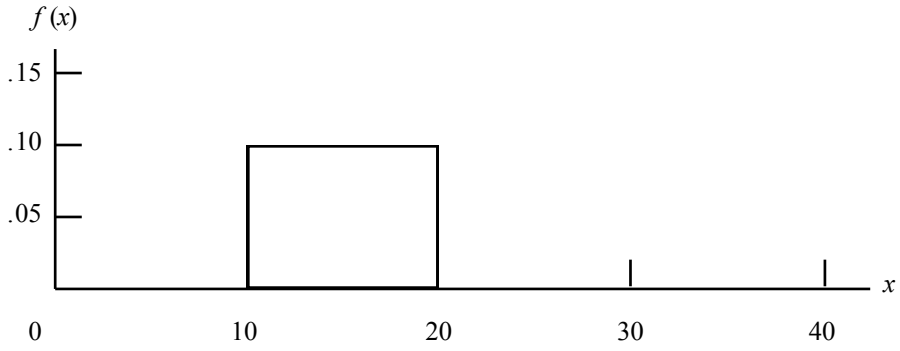


   b. $P(X = 1.25) = 0$. The probability of any single point is zero since the area under the curve above any single point is zero.

   c. $P(1.0 \le X \le 1.25) = 2(.25) = .50$

   d. $P(1.20 < X < 1.5) = 2(.30) = .60$

2. a.



   b. $P(X < 15) = .10(5) = .50$

   c. $P(12 \le X \le 18) = .10(6) = .60$

   d. $E(X) = \frac{10+20}{2} = 15$

   e. $Var(X) = \frac{(20-10)^2}{12} = 8.33$

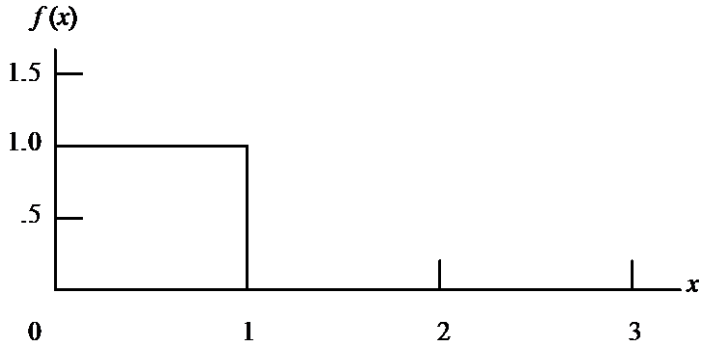
3. a. $k = 0.5$

   b. E(X) = 1, Var(X) = 1/3

   c. $P(X > 1+3*1/3) = P(X>2) = 0$

   d. $F(X) = 0.5x^2$ Median occurs when F(X) = 0.5. Hence x = 1

4. a.



   b. $P(.25 < X < .75) = 1\ (.50) = .50$

   c. $P(X \leq .30) = 1\ (.30) = .30$

   d. $P(X > .60) = 1\ (.40) = .40$

5. f(x) = 1/51

a. P(X > 170) = 30/51

b. P(X > 195) = 5/51

c. P(X < 160) = 10/51

d. P(X = 175) = 1/51

6. a. $P(12 \leq X \leq 12.05) = 8(.05) = .4$

   b. $P(X \geq 12.02) = 8(.08) = .64$

   c. P(X <11.98) + P(X > 12.02) = 8(.005) +8(.08)

   Therefore, the probability is .04 + .64 = .68

7. a. $P(10{,}000 \leq X < 12{,}000) = 2000\,(1 / 5000) = .40$

The probability your competitor will bid lower than you, and you get the bid, is .40.

b. $P(10{,}000 \leq X < 14{,}000) = 4000\,(1 / 5000) = .80$

c. A bid of €15,000 gives a probability of 1 of getting the property.

d. Yes, the bid that maximizes expected profit is €13,000.

The probability of getting the property with a bid of €13,000 is

$$P(10{,}000 \leq X < 13{,}000) = 3000\,(1 / 5000) = 0.60.$$

The probability of not getting the property with a bid of €13,000 is .40.

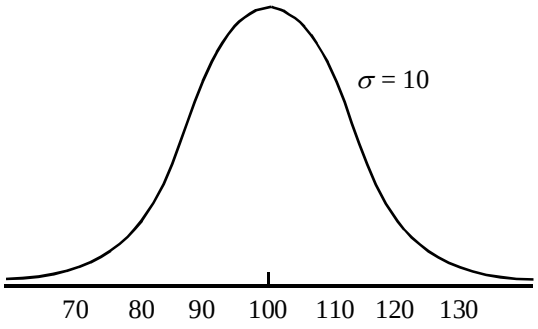
The profit you will make if you get the property with a bid of $13,000 is €3000 = €16,000 - 13,000. So your expected profit with a bid of €13,000 is

$$\text{EP}\,(€13{,}000) = 0.6\,(€3000) + 0.4\,(0) = €1800.$$

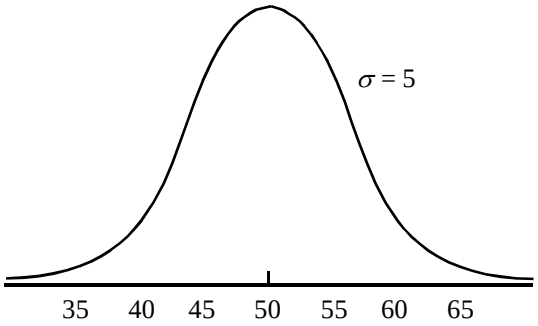
If you bid €15,000 the probability of getting the bid is 1, but the profit if you do get the bid is only €1000 = €16,000 - 15,000. So your expected profit with a bid of €15,000 is

$$\text{EP}\,(€15{,}000) = 1\,(€1000) + 0\,(0) = €1{,}000.$$

8.



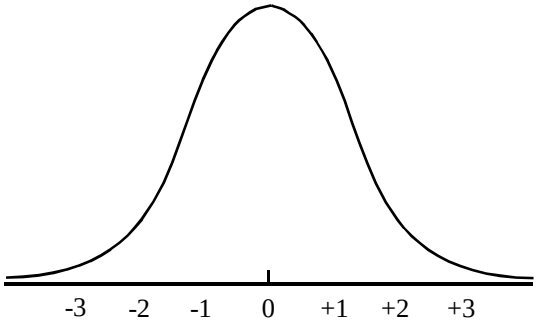
9. a.



b. 6826 since 45 and 55 are within plus or minus 1 standard deviation from the mean of 50.

c. 9544 since 40 and 60 are within plus or minus 2 standard deviations from the mean of 50.

10.



a. .8413 - .5 = .3413

b. .9332 - .5 = .4332

c. .9772 - .5 = .4772

d. .9938 - .5 = .4938

11. a. .5 - .1587 = .3413

b. .5 - .0668 = .4332

c. .5 - .0228 = .4772

d. .5 - .0062 = .4938

e. .5 - .0013 = .4987

12. a. .7967 - .5 = .2967

    b. .5 - .0582 = .4418

    c. 1 - .6700 = .3300

    d. 1 - .4090 = .5910

    e. .8849

    f. .2389

13. a. .6879 - .0239 = .6640

    b. .8888 - .6985 = .1903

    c. .1492 - .0401 = .1091

14. a. Using the table of areas for the standard normal probability distribution, the area of .9750 corresponds to $z = 1.96$.

    b. Using the table, the area of .7291 corresponds to $z = .61$.

    c. Look in the table for an area of 1 - .1314 = .8686. This provides $z = 1.12$.

    d. Look in the table for an area of .6700. This provides $z = .44$.

15. a. Look in the table for an area of .2119 = .2881. This corresponds with $z = -.80$.

    b. Look in the table for an area of .9030 / 2 = .4515; $z = -1.66$.

    c. Look in the table for an area of .2052 / 2 = .1026; $z = -.26$.

    d. Look in the table for an area of .9948; $z = 2.56$.

    e. Look in the table for an area of 1-.6915 = .3085 This corresponds to $z = -.50$.

16. a. Look in the table for an area of 1 - .0100 = .9900. The area value in the table closest to .9900 provides the value $z = 2.33$.

    b. Look in the table for an area of 1 - .0250 = .9750. This corresponds to $z = 1.96$.

    c. Look in the table for an area of 1 - .0500 = .9500. Since .9500 is exactly halfway between .9495 ($z = 1.64$) and .9505 ($z = 1.65$), we select $z = 1.645$. However, $z = 1.64$ or $z = 1.65$ are also acceptable answers.

    d. Look in the table for an area of 1 - .1000 = .9000. The area value in the table closest to .9000 provides the value $z = 1.28$.

17. a. 0.5

b. 0.000

c. 0.224

d. 34,579

18. Area to right = .2. So Area to the left = 0.8. z value = .84 = $z = \dfrac{x - 20000}{1500}$

Hence x = 20000+1500 (.84) = £21262.43 = minimum holding

19. a. $z = \dfrac{x - \mu}{\sigma}$ Hence $x = \mu + z\sigma$

When p = .4, z = -.25. This is when x = 10 so we can write

$x = \mu + z\sigma = 10 = \mu + (-.25)\sigma$ (1)

Similarly when p = .9, z = 1.28. This is when x = 15 so we can write

$15 = \mu + 1.28\sigma$ (2)

Solving equations (1) and (2) simultaneously for μ and σ we have:

$\mu = 10.82$, $\sigma = 3.27$

For the dose lethal to 50% of the insect population p = 0.5 so $x = \mu = 10.82$.

b. When x = 17.5 $z = \dfrac{17.5 - 10.82}{3.27} = 2.05$

Area to left = .9798.

20. a. $\mu = n\pi = 100(.20) = 20$

$\sigma^2 = n\pi(1 - \pi) = 100(.20)(.80) = 16$

$\sigma = \sqrt{16} = 4$

b. Yes since $n\pi = 20$ and $n(1 - \pi) = 80$

c. $P(23.5 \leq X \leq 24.5)$

$$z = \frac{24.5 - 20}{4} = +1.13 \qquad \text{Area} = .3708$$

$$z = \frac{23.5 - 20}{4} = +.88 \qquad \text{Area} = .3106$$

$P(23.5 \leq X \leq 24.5) = .3708 - .3106 = .0602$

d. $P(17.5 \leq X \leq 22.5)$

$$z = \frac{17.5 - 20}{4} = -.63 \qquad \text{Area} = .2357$$

$$z = \frac{22.5 - 20}{4} = +.63 \qquad \text{Area} = .2357$$

$P(17.5 \leq X \leq 22.5) = .2357 + .2357 = .4714$

e. $P(X \leq 15.5)$

$$z = \frac{15.5 - 20}{4} = -1.13 \qquad \text{Area} = .3708$$

$P(X \leq 15.5) = .5000 - .3708 = .1292$

21. a. $\mu = n\pi = 200(.60) = 120$

$\sigma^2 = n\pi(1 - \pi) = 200(.60)(.40) = 48$

$$\sigma = \sqrt{48} = 6.93$$

b. Yes since $n\pi = 120$ and $n(1 - \pi) = 80$

c. $P(99.5 \leq X \leq 110.5)$

$$z = \frac{99.5 - 120}{6.93} = -2.96 \qquad \text{Area} = .4985$$

$$z = \frac{110 - 120}{6.93} = -1.37 \qquad \text{Area} = .4147$$

$P(99.5 \leq X \leq 110.5) = .4985 - .4147 = .0838$

d. $P(X \geq 129.5)$

$$z = \frac{129.5 - 120}{6.93} = +1.37 \quad \text{Area} = .4147$$

$P(X \geq 129.5) = .5000 - .4147 = .0853$

e. Simplifies computation. By direct computation of binomial probabilities we would have to compute

$P(X \geq 130) = p(130) + p(131) + p(132) + p(133) + \ldots$

22. a. $\mu = n\pi = 120(.75) = 90$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{120 * 0.75 * 0.25} = 4.74.$$

The probability at least half the rooms are occupied is the normal probability: $P(X \geq 59.5)$.

At $X = 59.5$

$$z = \frac{59.5 - 90}{4.74} = -6.43$$

Therefore, probability is approximately 1

b. Find the normal probability: $P(X \geq 99.5)$

At $X = 99.5$

$$z = \frac{99.5 - 90}{4.74} = 2.00$$

$P(X \geq 99.5) = P(z \geq 2.00) = .5000 - .4772 = .0228$

c. Find the normal probability: $P(X \leq 80.5)$

At $X = 80.5$

$$z = \frac{80.5 - 90}{4.74} = -2.00$$

$P(X \leq 80.5) = P(z \leq -2.00) = .5000 - .4772 = .0228$

23. a. $P(X \leq 6) = 1 - e^{-6/8} = 1 - .4724 = .5276$

b. $P(X \leq 4) = 1 - e^{-4/8} = 1 - .6065 = .3935$

c. $P(X \geq 6) = 1 - P(X \leq 6) = 1 - .5276 = .4724$

d. $P(4 \leq X \leq 6) = P(X \leq 6) - P(X \leq 4) = .5276 - .3935 = .1341$

24. a. $P(x \leq x_0) = 1 - e^{-x_0/3}$

b. $P(X \leq 2) = 1 - e^{-2/3} = 1 - .5134 = .4866$

c. $P(X \geq 3) = 1 - P(X \leq 3) = 1 - (1 - e^{-3/3}) = e^{-1} = .3679$

d. $P(X \leq 5) = 1 - e^{-5/3} = 1 - .1889 = .8111$

e. $P(2 \leq X \leq 5) = P(X \leq 5) - P(X \leq 2) = .8111 - .4866 = .3245$

25. $\mu = 60/30 = 2$ / min $\quad F(t) = 1 - \exp(-\mu t)$

a. $0.018 = 1\text{-}F(2) = \exp(-4)$

$\mu = 60/25 = 2.4$ / min

b. $0.008 = 1 - F(2) = \exp(-4.8)$

26. a. 50 hours

b. $P(X \leq 25) = 1 - e^{-25/50} = 1 - .6065 = .3935$

c. $P(X \geq 100) = 1 - (1 - e^{-100/50})$
$= .1353$

27. a. If the average number of transactions per year follows the Poisson distribution, the time between transactions follows the exponential distribution. So,

$\mu = \frac{1}{30}$ of a year

and $\frac{1}{\mu} = \frac{1}{1/30} = 30$

then $f(X) = 30\, e^{-30x}$

b. A month is 1/12 of a year so,

$$P\left(x > \frac{1}{12}\right) = 1 - P\left(x \le \frac{1}{12}\right) = 1 - (1 - e^{-30/12}) = e^{-30/12} = .0821$$

The probability of no transaction during January is the same as the probability of no transaction during any month: .0821

c. Since 1/2 month is 1/24 of a year, we compute,

$$P\left(x \le \frac{1}{24}\right) = 1 - e^{-30/24} = 1 - .2865 = .7135$$

Chapter 6: Continuous Probability Solutions

Supplementary Exercises:

28. Motorola used the normal distribution to determine the probability of defects and the number of defects expected in a production process (*APICS—The Performance Advantage,* July 1991). Assume a production process produces items with a mean weight of 10 grams.

    Calculate the probability of a defect and the expected number of defects for a 1000-unit production run in the following situations.

    a. The process standard deviation is 0.15, and the process control is set at plus or minus one standard deviation. Units with weights less than 9.85 or greater than 10.15 grams will be classified as defects.

    b. Through process design improvements, the process standard deviation can be reduced to 0.05. Assume the process control remains the same, with weights less than 9.85 or greater than 10.15 grams being classified as defects.

    c. What is the advantage of reducing process variation thereby setting process control limits at a greater number of standard deviations from the mean?

29. Assume that the test scores from a college admissions test are normally distributed, with a mean of 450 and a standard deviation of 100.

    a. What percentage of the people taking the test score between 400 and 500?

    b. Suppose someone receives a score of 630. What percentage of the people taking the test score better? What percentage score worse?

    c. If a particular university will not admit anyone scoring below 480, what percentage of the persons taking the test would be acceptable to the university?

30. A machine fills containers with a particular product. The standard deviation of filling weights is known from past data to be .6 gram. If only 2% of the containers hold less than 18 grams, what is the mean filling weight for the machine? That is, what must μ equal? Assume the filling weights have a normal distribution.

31. Consider a multiple-choice examination with 50 questions. Each question has four possible answers. Assume that a student who has done the homework and attended lectures has a 75% probability of answering any question correctly.

a. A student must answer 43 or more questions correctly to obtain a grade of A. What percentage of the students who have done their homework and attended lectures will obtain a grade of A on this multiple-choice examination?
b. A student who answers 35 to 39 questions correctly will receive a grade of C. What percentage of students who have done their homework and attended lectures will obtain a grade of C on this multiple-choice examination?
c. A student must answer 30 or more questions correctly to pass the examination. What percentage of the students who have done their homework and attended lectures will pass the examination?
d. Assume that a student has not attended class and has not done the homework for the course. Furthermore, assume that the student will simply guess at the answer to each question. What is the probability that this student will answer 30 or more questions correctly and pass the examination?

32. A blackjack player at a casino in Bad Neuenahr learned that the house will provide a free room if play is for four hours at an average bet of €50. The player's strategy provides a probability of 0.49 of winning on any one hand, and the player knows that there are 60 hands per hour. Suppose the player plays for four hours at a bet of €50 per hand.

a. What is the player's expected payoff?
b. What is the probability the player loses €1000 or more?
c. What is the probability the player wins?
d. Suppose the player starts with €1500. What is the probability of going broke?

33. The time in minutes for which a student uses a computer terminal at the computer centre of a major university follows an exponential probability distribution with a mean of 36 minutes. Assume a student arrives at the terminal just as another student is beginning to work on the terminal.

a. What is the probability that the wait for the second student will be 15 minutes or less?
b. What is the probability that the wait for the second student will be between 15 and 45 minutes?
c. What is the probability that the second student will have to wait an hour or more?

34. The time (in minutes) between telephone calls at an insurance claims office has the following exponential probability distribution.

$$f(x) = 0.50e^{-0.50x} \qquad \text{for } x \geq 0$$

a. What is the mean time between telephone calls?

b. What is the probability of having 30 seconds or less between telephone calls?

c. What is the probability of having 1 minute or less between telephone calls?

d. What is the probability of having 5 or more minutes without a telephone call?

35. The mean life of a make and type of battery is 20 hours with a standard deviation of 0.5 hours.

a. What is the probability, p, that the battery will last no more than 21 hours?

b. How long would we expect a battery to last if p were given as 20%?

36. Attendance at a rock concert is normally distributed with a mean of 28,000 persons and a standard deviation of 4000 persons. What is the probability, that:

a. more than 28000 persons will attend?

b. less than 14000 persons will attend?

c. between 17000 and 25000 persons will attend?

d. Suppose the number who actually attended was X and the probability of achieving this level of attendance or higher was found to be 5%. What is X?

37. Items are manufactured to a mean weight of 3 kg and a standard deviation of 0.5 kg. Cartons each containing nine items are sold on the understanding that the mean weight per item is not less than 2.97 kg. Cartons not meeting this requirement are rejected.

a. What proportion of cartons is rejected?

b. Suppose now that the proportion, rejected, can be reduced by adjusting the process so that the mean weight of all items is increased to 3.01 kg. This adjustment cost 36 cents per carton. If the cost of a rejected carton is €10 determine whether the adjustment is economically justifiable.

38. Examination results for a particular group of students taking an introductory course in statistics are believed to be normally distributed with a mean of 58% and a standard deviation of 10%. What proportion of students has marks

a. exceeding 70%?

b. below 40%?

c. between 40% and 60% ?

39. The holdings of clients of a successful on-line stockbroker are normally distributed with a mean of €20,000 and standard deviation of €1,500. To increase its business, the stockbroker is looking to email special promotions to the top 20% of its clientele based on the value of their holdings. What is the minimum holding of this group?

40. A machine fills 50 kg bags with sand. The actual weight of sand in the bags when the machine operates at its standard speed of 100 bags per hour has a normal distribution with a standard deviation of 0.75 kg. The mean of the distribution depends on the setting of the machine. At what mean weight should the machine be set so that only 5% of the bags are under weight i.e. contain less than 50 kg of sand?

41. The website for the Bed and Breakfast Inns of North America (www.bestinns.net) gets approximately seven visitors per minute (*Time*, September 2001). Suppose the number of website visitors per minute follows a Poisson distribution.

a. What is the mean time between visits to the website?
b. Show the exponential probability density function for the time between website visits.
c. What is the probability no one will access the website in a 1 minute period?
d. What is the probability no one will access the website in a 12 second period?

42. The average travel time to work for New York City residents is 36.5 minutes (Time Almanac, 2001)

a. Assume the exponential probability distribution is applicable and show the probability density function for the travel time to work for a typical New Yorker.
b. What is the probability it will take a typical New Yorker between 20 and 40 minutes to travel to work?
c. What is the probability it will take a typical New Yorker more than 40 minutes to travel to work?

Supplementary Exercises Solutions:

28. a. $P(defect) = 1 - P(9.85 \le X \le 10.15)$

$= 1 - P(-1 \le Z \le 1) = 1 - .6826 = .3174$

Expected number of defects $= 1000(.3174) = 317.4$

b. $P(defect) = 1 - P(9.85 \le X \le 10.15)$

$= 1 - P(-3 \le Z \le 3) = 1 - .9974 = .0026$

Expected number of defects $= 1000(.0026) = 2.6$

c. Reducing the process standard deviation causes a substantial reduction in the number of defects.

29. a. At 400,

$$z = \frac{400 - 450}{100} = -.500$$

Area to left is .3085

At 500,

$$z = \frac{500 - 450}{100} = +.500$$

Area to left is .6915

$P(400 \le X \le 500) = .6915 - .3085 = .3830$

38.3% will score between 400 and 500.

b. At 630,

$$z = \frac{630 - 450}{100} = 1.80$$

96.41% do worse and 3.59% do better .

c. At 480,

$$z = \frac{480 - 450}{100} = .30$$

Area to left is .6179

38.21% are acceptable.

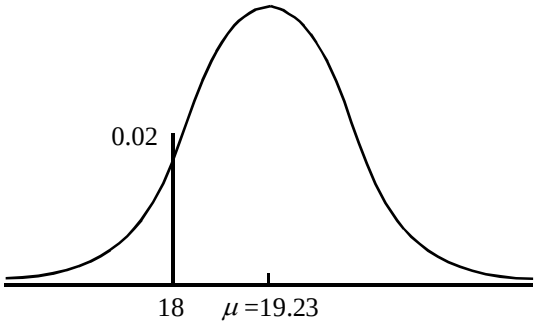
30. σ = .6

At 2%

$$z = -2.05 \quad X = 18$$

$$z = \frac{x - \mu}{\sigma} \qquad \therefore -2.05 = \frac{18 - \mu}{.6}$$

$$\mu = 18 + 2.05\,(.6) = 19.23 \text{ oz.}$$



The mean filling weight must be 19.23 oz.

31. Use normal approximation to binomial.

a. $\mu = n\pi = 50\,(.75) = 37.5$

$$\sigma = \sqrt{np(1-p)} = \sqrt{50(.75)(.25)} = 3.06$$

At X = 42.5

$$z = \frac{x - \mu}{\sigma} = \frac{42.5 - 37.5}{3.06} = 1.63$$

$$P(0 \le Z \le 1.63) = .9448 - .5 = .4484$$

Probability of an A grade = 1 - .9484 = .0516 or 5.16% will obtain an A grade.

b. At X = 34.5

$$z = \frac{34.5 - 37.5}{3.06} = -.98$$

At X = 39.5

$$z = \frac{39.5 - 37.5}{3.06} = .65$$

$P(-.98 \le Z \le .65) = .7422 - .1635 = .5787$

or 57.87% will obtain a C grade.

c. At X = 29.5

$$z = \frac{29.5 - 37.5}{3.06} = -2.61$$

$P(Z \geq -2.61) = 1 - .0045 = .9955$

or 99.55%of the students who have done their homework and attended lectures will pass the examination.

d. $\mu = n\pi = 50\,(.25) = 12.5$ (We use $\pi = .25$ for a guess.)

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{50(0.25)(0.75)} = 3.06$$

At X = 29.5

$$z = \frac{29.5 - 12.5}{3.06} = 5.55$$

$P(Z \geq 5.55) \approx 0$

Thus, essentially no one who simply guesses will pass the examination.

32. a. $\mu = n\pi = (240)(0.49) = 117.6$

Expected number of wins is 117.6

Expected number of losses = 240(0.51) = 122.4

Expected payoff = 117.6(50) - 122.4(50) = (-4.8)(50) = -240.

The player should expect to lose €240.

b. To lose €1000, the player must lose 20 more hands than he wins. With 240 hands in 4 hours, the player must win 110 or less in order to lose €1000. Use normal approximation to binomial.

$\mu = n\pi = (240)(0.49) = 117.6$

$$\sigma = \sqrt{240(0.49)(0.51)} = 7.7444$$

Find $P(X \leq 110.5)$

At X = 110.5

$$z = \frac{110.5 - 117.6}{7.7444} = -.92$$

$P(X \leq 110.5) = 0.1788$

The probability he will lose €1000 or more is 0.1788.

c. In order to win, the player must win 121 or more hands.

Find $P(X \geq 120.5)$

At $X = 120.5$

$$z = \frac{120.5 - 117.6}{7.7444} = .37$$

$P(X \geq 120.5) = 1 - 0.6443 = 0.3557$

The probability that the player will win is 0.3557. The odds are clearly in the house's favor.

d. To lose €1500, the player must lose 30 hands more than he wins. This means he wins 105 or fewer hands.

Find $P(X \leq 105.5)$

At $X = 105.5$

$$z = \frac{105.5 - 117.6}{7.7444} = -1.56$$

$P(X \leq 105.5) = 0.0594$

The probability the player will go broke is 0.0594.

33. a. $P(X \leq 15) = 1 - e^{-15/36} = 1 - .6592 = .3408$

b. $P(X \leq 45) = 1 - e^{-45/36} = 1 - .2865 = .7135$

Therefore $P(15 \leq X \leq 45) = .7135 - .3408 = .3727$

c. $P(X \geq 60) = 1 - P(X < 60)$

$= 1 - (1 - e^{-60/36}) = .1889$

34. a. $\frac{1}{\mu} = 0.5$ therefore μ = 2 minutes = mean time between telephone calls

b. Note: 30 seconds = .5 minutes

$P(X \leq .5) = 1 - e^{-.5/2} = 1 - .7788 = .2212$

c. $P(X \leq 1) = 1 - e^{-1/2} = 1 - .6065 = .3935$

d. $P(X \geq 5) = 1 - P(X < 5) = 1 - (1 - e^{-5/2}) = .0821$

35. a. p(X<= 21) = 0.9772

    b. 19.579 where p(X<= 19.579) = 0.2

36. a. 0.5

    b. 0.000233

    c. 0.224

    d. 34,579 where p(X>= 34579) = 0.05

37. a. 0.429 = p(X<2.97| $\mu = 3, \sigma = 0.5$)

    b. After the adjustment p(reject) = 0.405= p(X<2.97| $\mu = 3.01, \sigma = 0.5$);

    Cost (€) of rejects leaving process alone = 0.429(10) = 4.29
    Cost (€) of rejects after adjusting process = 4.05 (10) + 9(0.36) = 7.29
    So adjustment is not economically justifiable.

38. a. 0.115

    b. 0.036

    c. 0.543

39. €21262 where p(X>= 21262) = 0.20

40. Target= 51.233 = 50 - 0.75($z_{.05}$) where = $z_{.05}$ = -1.645

41. a. Mean time between arrivals = 1/7 minutes

    b. $f(x) = 7e^{-7x}$

    c. $P(X > 1) = 1 - P(X < 1) = 1 - [1 - e^{-7(1)}] = e^{-7} = .0009$

    d. 12 seconds is .2 minutes

    $P(X > .2) = 1 - P(X < .2) = 1 - [1 - e^{-7(.2)}] = e^{-1.4} = .2466$

42. a. $\frac{1}{36.5}e^{-x/36.5} \approx .0274e^{-.0274x}$

b. $P(X < 40) = 1 - e^{-.0274(40)} = 1 - .3342 = .6658$

$P(X < 20) = 1 - e^{-.0274(20)} = 1 - .5781 = .4219$

$P(20 < X < 40) = .6658 - .4219 = .2439$

c. From part (b), $P(X < 40) = .6658$

$P(X > 40) = P(X \geq 40) = 1 - P(X < 40) = 1 - .6658 = .3342$

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Seven

Sampling and Sampling Distributions

Textbook Exercises (1-31)

Textbook Exercise Solutions

Supplementary Exercises (32-51)

Supplementary Exercise Solutions

Chapter 7: Sampling and Sampling Distributions

Textbook Exercises:

1. Consider a finite population with five elements labelled A, B, C, D and E. Ten possible simple random samples of size 2 can be selected.
   a. List the ten samples beginning with AB, AC and so on.
   b. Using simple random sampling, what is the probability that each sample of size 2 is selected?
   c. Assume random number 1 corresponds to A, random number 2 corresponds to B, and so on. List the simple random sample of size 2 that will be selected by using the random digits 8 0 5 7 5 3 2.

2. Assume a finite population has 350 elements. Using the last three digits of each of the following five-digit random numbers (601, 022, 448, . . .), determine the first four elements that will be selected for the simple random sample.

   98601 73022 83448 02147 34229 27553 84147 93289 14209

3. The EURO STOXX 50 share index is calculated using data for 50 blue-chip companies from 12 Eurozone countries. Assume you want to select a simple random sample of five companies from the EURO STOXX 50 list. Use the last two digits in column 9 of Table 7.1, beginning with 54. Read down the column and identify the numbers of the five companies that would be selected.

4. A student union is interested in estimating the proportion of students who favour a mandatory 'pass–fail' grading policy for optional courses. A list of names and addresses of the 645 students enrolled during the current semester is available from the registrar's office. Using three-digit random numbers in row 10 of Table 7.1 and moving across the row from left to right, identify the first ten students who would be selected using simple random sampling. The three-digit random numbers begin with 816, 283 and 610.

5. Assume that we want to identify a simple random sample of 12 of the 372 doctors practising in a particular city. The doctors' names are available from the local health authority. Use the eighth column of five-digit random numbers in Table 7.1 to identify the 12 doctors for the sample. Ignore the first two random digits in each five-digit grouping of the random numbers. This process begins with random number 108 and proceeds down the column of random numbers.

6. Indicate whether the following populations should be considered finite or infinite.
   a. All registered voters in Ireland.
   b. All television sets that could be produced by the Johannesburg factory of the TV-M Company.
   c. All orders that could be processed by a mail-order firm.
   d. All emergency telephone calls that could come into a local police station.
   e. All components that Fibercon plc produced on the second shift on 17 Feb 2013.

7. The following data are from a simple random sample.

   5 8 10 7 10 14

   a. Calculate a point estimate of the population mean.
   b. Calculate a point estimate of the population standard deviation.

8. A survey question for a sample of 150 individuals yielded 75 Yes responses, 55 No responses, and 20 No Opinion responses.
   a. Calculate a point estimate of the proportion in the population who respond Yes.
   b. Calculate a point estimate of the proportion in the population who respond No.

9. A simple random sample of five months of sales data provided the following information:

Month:	1	2	3	4	5
Units sold:	94	100	85	94	92

   a. Calculate a point estimate of the population mean number of units sold per month.
   b. Calculate a point estimate of the population standard deviation.

10. The data set Mutual Fund contains data on a sample of 40 mutual funds. These were randomly selected from 283 funds featured in Business Week. Use the data set to answer the following questions.
    a. Compute a point estimate of the proportion of the Business Week mutual funds that are load funds.
    b. Compute a point estimate of the proportion of the funds that are classified as high risk.
    c. Compute a point estimate of the proportion of the funds that have a below-average risk rating.

11. In a YouGov opinion poll for the *Financial Times* in late June 2012, during the 'Euro crisis', a sample of 1033 German adults was asked "If there were a referendum tomorrow on Germany's membership of the single currency, the euro, how would you vote?" The responses were:

To stay in the single currency	444
To bring back the Deutschmark	424
Would not vote	72
Don't know	93

    Calculate point estimates of the following population parameters:
    (a) The proportion of all adults who would vote to stay in the single currency.
    (b) The proportion of all adults who vote to bring back the Deutschmark.
    (c) The proportion of all adults who would not vote or don't know.

12. Many drugs used to treat cancer are expensive. BusinessWeek reported on the cost per treatment of Herceptin, a drug used to treat breast cancer. Typical treatment costs (in dollars) for Herceptin are provided by a simple random sample of 10 patients.

4376	5578	2717	4920	4495
4798	6446	4119	4237	3814

    a. Calculate a point estimate of the mean cost per treatment with Herceptin.
    b. Calculate a point estimate of the standard deviation of the cost per treatment with Herceptin.

13. A population has a mean of 200 and a standard deviation of 50. A simple random sample of size 100 will be taken and the sample mean will be used to estimate the population mean.
    a. What is the expected value of $\bar{X}$?
    b. What is the standard deviation of $\bar{X}$?
    c. Sketch the sampling distribution of $\bar{X}$?
    d. What does the sampling distribution of $\bar{X}$show?

14. A population has a mean of 200 and a standard deviation of 50. Suppose a simple random sample of size 100 is selected and is used to estimate μ.
    a. What is the probability that the sample mean will be within ±5 of the population mean?
    b. What is the probability that the sample mean will be within ±10 of the population mean?

15. Assume the population standard deviation is σ = 25. Compute the standard error of the mean, $\sigma_{\bar{X}}$, for sample sizes of 50, 100, 150 and 200. What can you say about the size of the standard error of the mean as the sample size is increased?

16. Suppose a simple random sample of size 50 is selected from a population with σ = 25. Find the value of the standard error of the mean in each of the following cases (use the finite population correction factor if appropriate).
    a. The population size is infinite.
    b. The population size is N = 50 000.
    c. The population size is N = 5000.
    d. The population size is N = 500.

17. Refer to the EAI sampling problem. Suppose a simple random sample of 60 managers is used.
    a. Sketch the sampling distribution of $\bar{X}$ when simple random samples of size 60 are used.
    b. What happens to the sampling distribution of $\bar{X}$ if simple random samples of size 120 are used?

c. What general statement can you make about what happens to the sampling distribution of $\bar{X}$ as the sample size is increased? Does this generalization seem logical? Explain.

18. In the EAI sampling problem (see Figure 7.5), we showed that for $n = 30$, there was a 0.5034 probability of obtaining a sample mean within ±€500 of the population mean.
    a. What is the probability that $\bar{X}$ is within €500 of the population mean if a sample of size 60 is used?
    b. Answer part (a) for a sample of size 120.

19. The Automobile Association gave the average price of unleaded petrol in Sweden as 14.63 Swedish Krona (SK) per litre in June 2012. Assume this price is the population mean, and that the population standard deviation is $\sigma = 1$ SK.
    a. What is the probability that the mean price for a sample of 30 petrol stations is within 0.25 SK of the population mean?
    b. What is the probability that the mean price for a sample of 50 petrol stations is within 0.25 SK of the population mean?
    c. What is the probability that the mean price for a sample 100 petrol stations is within 0.25 SK of the population mean?
    d. Would you recommend a sample size of 30, 50 or 100 to have at least a 0.95 probability that the sample mean is within 0.25 SK of the population mean?

20. According to Golf Digest, the average score for male golfers is 95 and the average score for female golfers is 106. Use these values as population means. Assume that the population standard deviation is $\sigma = 14$ strokes for both men and women. A simple random sample of 30 male golfers and another simple random sample of 45 female golfers are taken.
    a. Sketch the sampling distribution of $\bar{X}$ for male golfers.
    b. What is the probability that the sample mean is within 3 strokes of the population mean for the sample of male golfers?
    c. What is the probability that the sample mean is within 3 strokes of the population mean for the sample of female golfers?
    d. In which case is the probability higher (b or c)? Why?

21. A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500.
    a. How large was the sample?
    b. What is the probability that the point estimate was within ±25 of the population mean?

22. To estimate the mean age for a population of 4000 employees, a simple random sample of 40 employees is selected.
    a. Would you use the finite population correction factor in calculating the standard error of the mean? Explain.
    b. If the population standard deviation is $\sigma = 8.2$, compute the standard error both with and without the finite population correction factor. What is the rationale for ignoring the finite population correction factor whenever $n/N \leq 0.05$?
    c. What is the probability that the sample mean age of the employees will be within ±2 years of the population mean age?

23. A simple random sample of size 100 is selected from a population with $\pi = 0.40$.
    a. What is the expected value of P?
    b. What is the standard error of P?
    c. Sketch the sampling distribution of P.

24. Assume that the population proportion is 0.55. Compute the standard error of the sample proportion, σ_P, for sample sizes of 100, 200, 500 and 1000. What can you say about the size of the standard error of the proportion as the sample size is increased?

25. The population proportion is 0.30. What is the probability that a sample proportion will be within ±0.04 of the population proportion for each of the following sample sizes?
    a. $n = 100$
    b. $n = 200$
    c. $n = 500$
    d. $n = 1000$
    e. What is the advantage of a larger sample size?

26. The Chief Executive Officer of Dunkley Distributors plc believes that 30 per cent of the firm's orders come from first-time customers. A simple random sample of 100 orders will be used to estimate the proportion of first-time customers.
    a. Assume that the CEO is correct and $\pi = 0.30$. Describe the sampling distribution of the sample proportion P for this study?
    b. What is the probability that the sample proportion *P* will be between 0.20 and 0.40?
    c. What is the probability that the sample proportion *P* will be between 0.25 and 0.35?

27. Eurostat reported that in 2011, 64 per cent of households in Spain had Internet access. Use a population proportion $\pi = 0.64$ and assume that a sample of 300 households will be selected.
    a. Sketch the sampling distribution of *P*, the sample proportion of households that have Internet access.
    b. What is the probability that the sample proportion *P* will be within ±0.03 of the population proportion?
    c. Answer part (b) for sample sizes of 600 and 1000.

28. Advertisers contract with Internet service providers and search engines to place ads on websites. They pay a fee based on the number of potential customers who click on their ads. Unfortunately, click fraud – i.e. someone clicking on an ad solely for the purpose of driving up advertising revenue – has become a problem. Forty per cent of advertisers claim they have been a victim of click fraud. Suppose a simple random sample of 380 advertisers is taken to learn about how they are affected by this practice.
    a. What is the probability the sample proportion will be within ±0.04 of the population proportion experiencing click fraud?
    b. What is the probability the sample proportion will be greater than 0.45?

29. In April 2012, a Gallup poll amongst a sample of 1074 Egyptian adults reported that 58% thought it would be a bad thing if the military remained involved in politics after the presidential election. Assume that the population proportion was □ = 0.58, and that *P* is the sample proportion in a sample of *n* = 1074.
    a. Sketch the sampling distribution of *P*.
    b. What is the probability that *P* will be within plus or minus 0.02 of □.
    c. Answer part (b) for sample of 2000 adults.

30. A market research firm conducts telephone surveys with a 40 per cent historical response rate. What is the probability that in a new sample of 400 telephone numbers, at least 150 individuals will cooperate and respond to the questions? In other words, what is the probability that the sample proportion will be at least 150/400 = 0.375?

31. Laura Jeffrey is a successful sales representative for a major publisher of university textbooks. Historically, Laura secures a book adoption on 25 per cent of her sales calls. Assume that her sales calls for one month are taken as a sample of all possible sales calls, and that a statistical analysis of the data estimates the standard error of the sample proportion to be 0.0625.
    a. How large was the sample used in this analysis? That is, how many sales calls did Laura make during the month?
    b. Let *P* indicate the sample proportion of book adoptions obtained during the month. Sketch the sampling distribution *P*.
    c. Using the sampling distribution of *P*, compute the probability that Laura will obtain book adoptions on 30 per cent or more of her sales calls during a one-month period.

Chapter 7: Sampling and Sampling Distributions

Textbook Exercise Solutions

1. a. AB, AC, AD, AE, BC, BD, BE, CD, CE, DE

   b. With 10 samples, each has a probability 0.1.

   c. E and C because 8 and 0 do not apply; 5 identifies E; 7 does not apply; 5 is skipped since E is already in the sample; 3 identifies C; 2 is not needed since the sample of size 2 is complete.

2. Using the last 3-digits of each 5-digit grouping provides the random numbers:

   601, 022, 448, 147, 229, 553, 147, 289, 209

   Elements 22, 147, 229, 289 are selected. (The numbers 98601, 83448, 27553 are ignored, because 601, 448 and 553 are out of the 1 to 350 range. 84147 is ignored because 147 has already been selected.)

3. Numbers greater than 50 do not apply, so the simple random sample of five contains numbers 47, 40, 13, 1 and 27.

4. Elements 283, 610, 39, 254, 568, 353, 602, 421, 638, 164 are selected. (The numbers 816, 763, 980, 964 are ignored, because 601, 448 and 553 are out of the 1 to 645 range.)

5. 108, 290, 201, 292, 322, 9, 244, 249, 226, 125, (continuing at the top of column 9) 147, and 113.

6. finite, infinite, infinite, infinite, finite

7. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{54}{6} = 9$

b. $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$

$\Sigma(x_i - \bar{x})^2 = (-4)^2 + (-1)^2 + 1^2\ (-2)^2 + 1^2 + 5^2 = 48$

$s = \sqrt{\frac{48}{6-1}} = 3.1$

8. a. $p = 75/150 = 0.50$

b. $p = 55/150 = 0.3667$

9. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{465}{5} = 93$

b.

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	94	+1	1
	100	+7	49
	85	-8	64
	94	+1	1
	92	-1	1
Totals	465	0	116

$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{116}{4}} = 5.39$

10. a. $p = 18/40 = 0.45$

b. $p = 6/40 = 0.15$

c. $p = (12 + 6)/40 = 0.45$

11. a. $p = 444/1033 = 0.430$

b. $p = 424/1033 = 0.410$

c. $p = (72 + 93)/1033 = 165/1033 = 0.160$

12. a. $\bar{x} = \dfrac{4376 + 5578 + 2717 + 4929 + 4495 + 4798 + 6446 + 4119 + 4237 + 3814}{10} = \dfrac{45500}{10} = 4550$

b. $s = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n - 1}}$

$$= \sqrt{\frac{(-174)^2 + (1028)^2 + (-1833)^2 + (370)^2 + (-55)^2 + (248)^2 + (1896)^2 + (-431)^2 + (-313)^2 + (-736)^2}{9}}$$

$$= \sqrt{\frac{9068620}{9}} = \sqrt{1007624} = 1004$$

13. a. $E(\bar{X}) = m = 200$

b. $s_{\bar{x}} = s/\sqrt{n} = 50/\sqrt{100} = 5$

c. Normal with $E(\bar{X}) = 200$ and $s_{\bar{x}} = 5$

d. It shows the probability distribution of all possible sample means that can be observed with random samples of size 100. This distribution can be used to compute the probability that $\bar{x}$ is within a specified distance of μ.

14. a. The sampling distribution is normal with:

$E(\bar{X}) = \mu = 200$

$s_{\bar{x}} = s/\sqrt{n} = 50/\sqrt{100} = 5$

For ±5, $(\bar{x} - \mu) = 5$,

$z = \frac{\bar{x} - m}{s_{\bar{x}}} = \frac{5}{5} = 1$, cumulative probability for $z = 1$ is 0.8413,

Required probability = 2(0.8413 – 0.5000) = 0.6826

b. For ± 10, $(\bar{x} - \mu) = 10$

$z = \frac{\bar{x} - m}{s_{\bar{x}}} = \frac{10}{5} = 2$, cumulative probability for $z = 2$ is 0.9772,

Required probability = 2(0.9772 – 0.5000) = 0.9544

15. $s_{\bar{x}} = s/\sqrt{n}$

$\sigma_{\bar{x}} = 25/\sqrt{50} = 3.54$

$\sigma_{\bar{x}} = 25/\sqrt{100} = 2.50$

$\sigma_{\bar{x}} = 25/\sqrt{150} = 2.04$

$\sigma_{\bar{x}} = 25/\sqrt{200} = 1.77$

The standard error of the mean decreases as the sample size increases.

16. a. $s_{\bar{x}} = s/\sqrt{n} = 10/\sqrt{50} = 1.41$

b. $n/N = 50/50{,}000 = 0.001$

Use $s_{\bar{x}} = s/\sqrt{n} = 10/\sqrt{50} = 1.41$

c. $n/N = 50/5000 = 0.01$

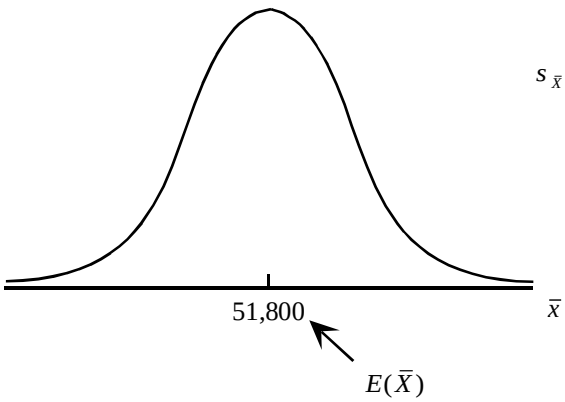
Use $s_{\bar{x}} = s/\sqrt{n} = 10/\sqrt{50} = 1.41$

d. $n/N = 50/500 = 0.10$

Use $s_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\frac{s}{\sqrt{n}} = \sqrt{\frac{500-50}{500-1}}\frac{10}{\sqrt{50}} = 1.34$

Note: Only case (d), where $n/N = 0.10$, requires the use of the finite population correction factor.

17. a.



$s_{\bar{x}} = s/\sqrt{n} = 4000/\sqrt{60} = 516.4$

The normal distribution is based on the Central Limit Theorem.

b. For $n = 120$, $E(\bar{X})$ remains 51,800 and the sampling distribution of $\bar{X}$ can still be approximated by a normal distribution. However, $s_{\bar{x}}$ is reduced to $4000/\sqrt{120} = 365.15$.

c. As the sample size is increased, the standard error of the mean, $\sigma_{\bar{x}}$, is reduced. This appears logical from the point of view that larger samples should tend to provide sample means that are closer to the population mean. Hence, the variability in the sample mean, measured in terms of $\sigma_{\bar{x}}$, should decrease as the sample size is increased.

18. a. $s_{\bar{x}} = s/\sqrt{n} = 4000/\sqrt{60} = 516.40$

$z = \dfrac{52,300 - 51,800}{516.40} = +0.97$, cumulative probability for $z = 0.97$ is 0.8340,

Required probability = 2(0.8340 – 0.5000) =0.6680

b. $s_{\bar{x}} = s/\sqrt{n} = 4000/\sqrt{120} = 365.15$

$z = \dfrac{52,300 - 51,800}{365.15} = +1.37$, cumulative probability for $z = 0.97$ is 0.9147,

Required probability = 2(0.4147 – 0.5000) = 0.8294

19. a. Sampling distribution of $\bar{X}$ is Normal, $E(\bar{X}) = 14.63$, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{1}{\sqrt{30}} = 0.183$

$z = \dfrac{0.25}{0.183} = 1.37$, cumulative probability = 0.9147, required probability = 2(0.9147 – 0.5000) = 0.8296

b. Sampling distribution of $\bar{X}$ is Normal, $E(\bar{X}) = 14.63$, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{50}} = 0.141$

$z = \frac{0.25}{0.141} = 1.77$, cumulative probability = 0.9616, required probability = 2(0.9616 – 0.5000) = 0.923

c. Sampling distribution of $\bar{X}$ is Normal, $E(\bar{X}) = 14.63$, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.100$

$z = \frac{0.25}{0.100} = 2.50$, cumulative probability = 0.9938, required probability = 2(0.9938 – 0.5000) = 0.988

d. Recommend $n = 100$, because the probabilities in a and b are below 0.95, in c above 0.95.

20. a. Normal with $E(\bar{X}) = 95$, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{14}{\sqrt{30}} = 2.556$

b. $z = \frac{3}{2.556} = 1.17$, cumulative probability for $z = 1.17$ is 0.8790,

Required probability = 2(0.8790 – 0.5000) = 0.7580

c. $z = \frac{3}{s/\sqrt{n}} = \frac{3}{14/\sqrt{45}} = \frac{3}{2.087} = 1.44$, cumulative probability for $z = 1.44$ is 0.9251,

Required probability = 2(0.9251 – 0.5000) = 0.8502

d. Probability in (c) is larger because of the larger sample size (smaller standard error).

21. a. $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{500}{\sqrt{n}} = 20$

Solving for n, $\sqrt{n} = \dfrac{500}{20} = 25$, $n = 25^2 = 625$

b. $z = \dfrac{25}{20} = 1.25$, cumulative probability for $z = 1.25$ is 0.8944,

Required probability = 2(0.8944 – 0.5000) = 0.7888

22. a. $n / N = 40 / 4000 = 0.01 < 0.05$; therefore, the finite population correction factor is not necessary.

b. With the finite population correction factor,

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\frac{s}{\sqrt{n}} = \sqrt{\frac{4000-40}{4000-1}}\frac{8.2}{\sqrt{40}} = 1.29$$

Without the finite population correction factor, $s_{\bar{x}} = s/\sqrt{n} = 1.30$

Including the finite population correction factor provides only a slightly different value for $s_{\bar{x}}$ compared to not using the correction factor.

c. $z = \dfrac{\bar{x} - \mu}{1.30} = \dfrac{2}{1.30} = 1.54$, cumulative probability = 0.9382,

required probability = 2(0.9382 – 0.5000) = 0.8764

23. a. $E(P) = \pi = 0.40$

b. $\sigma_P = \sqrt{\dfrac{\pi(1-\pi)}{n}} = \sqrt{\dfrac{0.40(0.60)}{100}} = 0.0490$

c. Normal distribution with $E(P) = 0.40$ and $s_P = 0.0490$

24. $s_P = \sqrt{\frac{p(1- p)}{n}}$

For $n = 100$, $\sigma_p = \sqrt{\frac{(0.55)(0.45)}{100}} = 0.0497$

For $n = 200$, $\sigma_P = \sqrt{\frac{(0.55)(0.45)}{200}} = 0.0352$

For $n = 500$, $\sigma_P = \sqrt{\frac{(0.55)(0.45)}{500}} = 0.0222$

For $n = 1000$, $\sigma_P = \sqrt{\frac{(0.55)(0.45)}{1000}} = 0.0157$

σ_p decreases as n increases

25. a. $\sigma_P = \sqrt{\frac{(0.30)(0.70)}{100}} = 0.0458$

$z = \frac{p - \pi}{\sigma_P} = \frac{0.04}{0.0458} = 0.87$, Required probability = 2(0.8078 – 0.5000) = 0.6156

b. $\sigma_P = \sqrt{\frac{(0.30)(0.70)}{200}} = 0.0324$

$z = \frac{p - \pi}{\sigma_P} = \frac{0.04}{0.0324} = 1.23$, Required probability = 2(0.8907 – 0.5000) = 0.7814

c. $\sigma_P = \sqrt{\frac{(0.30)(0.70)}{500}} = 0.0205$

$z = \frac{p - \pi}{\sigma_P} = \frac{0.004}{0.0205} = 1.95$, Required probability = 2(0.9744 – 0.5000) = 0.9488

d. $\sigma_P = \sqrt{\frac{(0.30)(0.70)}{1000}} = 0.0145$

$z = \frac{p - \pi}{\sigma_P} = \frac{0.04}{0.0145} = 2.76$, Required probability = 2(0.9971 – 0.5000) = 0.9942

e. With a larger sample, there is a higher probability p will be within ± 0.04 of the population proportion π.

26. a. The normal distribution is appropriate because $n\pi = 100(0.30) = 30$ and $n(1 - \pi) = 100(0.70) = 70$ are both greater than 5.

$E(P) = 0.30$, $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.30(0.70)}{100}} = 0.0458$

b. $z = \frac{0.40 - 0.30}{0.0458} = 2.18$, Required probability = 2(0.9854 – 0.5000) = 0.9708

c. $z = \frac{0.35 - 0.30}{0.0458} = 1.09$, Required probability = 2(0.8621 – 0.5000) = 0.7242

27. a. Normal with $E(P) = 0.64$ and $\sigma_P = \sqrt{\frac{0.64 \times 0.36}{300}} = 0.02771$

b. $z = \frac{0.03}{0.02771} = 1.08$, cumulative probability = 0.8599,

Required probability = 2(0.8599 – 0.5000) = 0.720

c. For $n = 600$, $\sigma_P = \sqrt{\frac{0.64 \times 0.36}{600}} = 0.0196$

$z = \frac{0.03}{0.0196} = 1.53$, cumulative probability = 0.9370,

Required probability = 2(0.9370 – 0.5000) = 0.8740

For $n = 1000$, $\sigma_P = \sqrt{\frac{0.64 \times 0.36}{1000}} = 0.0152$

$z = \frac{0.03}{0.0152} = 1.97$, cumulative probability = 0.9756,

Required probability = 2(0.9756 – 0.5000) = 0.951

28. a. Assuming $\pi = 0.40$, the sampling distribution of P is Normal with $E(P) = 0.30$ and $\sigma_P = \sqrt{\frac{0.40 \times 0.60}{380}} = 0.0251$

$z = \frac{0.04}{0.0251} = 1.59$, cumulative probability = 0.9441,

Required probability = 2(0.9441 – 0.5000) = 0.8882

b. $z = \frac{0.45 - 0.40}{0.0251} = 1.99$, cumulative probability = 0.9767,

Required probability = 1 – 0.9767 = 0.0233

29. a. Normal with $E(P) = 0.58$ and $\sigma_P = \sqrt{1074} = 0.0151$

b. $z = \frac{0.02}{0.0151} = 1.32$, Required probability = 2(0.9066 – 0.5000) = 0.813

c. For $n = 1600$, $\sigma_P = \sqrt{\frac{0.58 \times 0.42}{2000}} = 0.0110$, $z = \frac{0.02}{0.0110} = 1.82$,

Required probability = 2(0.9656 – 0.5000) = 0.931

30. $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{(0.40)(0.60)}{400}} = 0.0245$

$z = \frac{0.375 - 0.40}{0.0245} = -1.02$, Cumulative probability for $z = -1.02$ is 0.0.1539

Required probability = 1 – 0.1539 = 0.8461

31. a. $s_P = \sqrt{\frac{p(1- p)}{n}} = \sqrt{\frac{0.25(0.75)}{n}} = 0.0625$

Solving for n, $n = \frac{0.25(0.75)}{(0.0625)^2} = 48$

b. Normal distribution with $E(P) = 0.25$ and $s_P = 0.0625$

c. $z = \frac{0.30 - 0.25}{0.0625} = 0.80$, Cumulative probability = 0.7881

Required probability = 1 – 0.7881 = 0.2119

Chapter 7: Sampling and Sampling Distributions

Supplementary Exercises:

32. The following are 10 of the largest companies listed on the London stock exchange:

    BP, HSBC, Vodafone, GlaxoSmithKline, RBS, Shell, Barclays, HBOS, AstraZeneca, Lloyds TSB

    a. Assume that a random sample of five will be selected for an in-depth study of recent business practices at large, high-volume companies. Beginning with the first random digit in Table 7.1 and reading down the column, use the single-digit random numbers to select a simple random sample of five companies to be used in this study.
    b. How many different simple random samples of size 5 can be selected from the list of 10 companies?

33. The *County and City Data Book,* published by the Census Bureau, lists information on 3139 counties throughout the United States. Assume that a national study will collect data from 30 randomly selected counties. Use four-digit random numbers from the last column of Table 7.1 to identify the numbers corresponding to the first five counties selected for the sample. Ignore the first digits and begin with the four-digit random numbers 9945, 8364, 5702, and so on.

34. *Business Week's* Corporate Scoreboard provides quarterly data on sales, profits, net income, return on equity, price/earnings ratio, and earnings per share for 899 companies. The companies can be numbered 1 to 899 in the order they appear on the Corporate Scoreboard list. Begin at the bottom of the second column of random digits in Table 7.1. Ignoring the first two digits in each group and using three-digit random numbers beginning with 112, read *up* the column to identify the number of (from 1 to 899) the first eight companies to be included in a simple random sample.

35. The Nikkei 225 share index is calculated using data for 225 of the most actively traded stocks on the Tokyo stock exchange. Assume that you want to select a simple random sample of five companies from the Nikkei list. Use the last three digits in column 9 of Table 7.1, beginning with 554. Read down the column and identify the numbers of the five companies that would be selected.

36. A simple random sample of 10 VCRs shows the following useful life in years.

6.5 8.0 6.2 7.4 7.0 8.4 9.5 4.6 5.0 7.4

a. Compute a point estimate of the population mean life expectancy for VCRs.
b. Compute a point estimate of the population standard deviation for life expectancy of VCRs.

37. The proportion of major airline flights that arrive at or before their scheduled arrival times is to be estimated from a sample of 1400 flights. Compute a point estimate of the proportion of all flights that arrive on time, if 1117 of the sample flights arrive on time.

38. In an ICM poll for the Guardian newspaper in October 2008, during the turbulence in the world's financial markets, respondents were asked to what extent they felt they and their families would be affected financially. The opinions of the 1007 adult respondents were:

98 Suffer a great deal
320 Suffer quite a lot
426 Suffer a little
132 Not suffer at all
31 Don't know

Calculate point estimates of the following population parameters.

a. The proportion of all adults who feel they would suffer a little.
b. The proportion of all adults who feel they would not suffer at all.
c. The proportion of all adults who feel they would suffer quite a lot or a great deal.

39. Suppose that the mean annual starting salary for marketing graduates in a particular European country is €34,000, i.e. assume the population mean annual starting salary is μ = 34,000. The population standard deviation is $\sigma = 2000$ (€).

a. What is the probability that a simple random sample of marketing graduates will have a sample mean within $\pm$€250 of the population mean for each of the following sample sizes: 30, 50, 100, 200, and 400?
b. What is the advantage of a larger sample size when attempting to estimate the population mean?

40. Suppose the mean tuition fee at public universities in a particular European country is €4260 per semester. Use this value as the population mean and assume that the population standard deviation is σ = €900. Suppose that a random sample of 50 public universities will be selected.
    a. Show the sampling distribution of $\bar{X}$, the sample mean tuition cost for the 50 public universities.
    b. What is the probability that the simple random sample will provide a sample mean within €250 of the population mean?
    c. What is the probability that the simple random sample will provide a sample mean within €100 of the population mean?

41. The Automobile Association in the UK gave the average price of unleaded petrol as 106.4 p per litre in November 2008. Assume this price is the population mean, and that the population standard deviation is σ = 4.5 p.
    a. What is the probability that the mean price for a sample of 30 petrol stations is within 1.0 p of the population mean?
    b. What is the probability that the mean price for a sample of 50 petrol stations is within 1.0 p of the population mean?
    c. What is the probability that the mean price for a sample 100 petrol stations is within 1.0 p of the population mean?
    d. Would you recommend a sample size of 30, 50 or 100 to have at least a 0.95 probability that the sample mean is within 1.0 p of the population mean?

42. A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500.
    a. How large was the sample used in this survey?
    b. What is the probability that the point estimate was within ± 25 of the population mean?

43. Suppose that the mean television viewing time for teenagers is three hours per day, i.e. assume the population mean is μ = 3. The population standard deviation is σ = 1.2 hours. Suppose a sample of 50 teenagers will be used to monitor television viewing time. Let $\bar{X}$ denote the sample mean viewing time.
    a. Sketch the sampling distribution of $\bar{X}$.
    b. What is the probability the sample mean will be within $\pm$ 0.25 hour of the population mean?

44. Suppose the mean annual salary for government employees is €42,000. Use this figure as the population mean and assume the population standard deviation is σ = €5000. Suppose that a random sample of 50 government employees will be selected from the population.
    a. What is the value of the standard error of the mean?
    b. What is the probability that the sample mean will be more than €42,000?
    c. What is the probability the sample mean will be within €1000 of the population mean?
    d. How would the probability in part (c) change if the sample size were increased to 100?

45. Three firms carry inventories that differ in size. Firm A's inventory contains 2000 items, firm B's inventory contains 5000 items, and firm C's inventory contains 10,000 items. The population standard deviation for the cost of the items in each firm's inventory is σ = 144. A statistical consultant recommends that each firm take a sample of 50 items from its inventory to provide estimates of the average cost per item. Managers of the small firm state that because it has the smallest population, it should be able to make the estimate from a much smaller sample than that required by the larger firms. However, the consultant states that to obtain the same standard error and thus the same precision in the sample results, all firms should use the same sample size regardless of population size.
    a. Using the finite population correction factor, compute the standard error for each of the three firms, given a sample of size 50.
    b. What is the probability that for each firm the sample mean $\overline{X}$ will be within ± 25 of the population mean μ?

46. The proportion of individuals insured by the All-Driver Car Insurance Company who committed at least one motoring offence during a five-year period is 0.15.
    a. Sketch the sampling distribution of the sample proportion *P* if a random sample of 150 insured individuals is used to estimate the proportion having committed at least one offence.
    b. What is the probability that the sample proportion will be within ± 0.03 of the population proportion?

47. The UK Office for National Statistics reported that in 2007, 61 per cent of households in the UK had Internet access. Use a population proportion $\pi = 0.61$ and assume that a sample of 300 households will be selected.
    a. Sketch the sampling distribution of P, the sample proportion of households that have Internet access.

b. What is the probability that the sample proportion P will be within ±0.03 of the population proportion?
c. Answer part (b) for sample sizes of 600 and 1000.

48. Suppose that in a particular adult population, 25% of individuals skip breakfast, i.e. assume the population proportion is $\pi = 0.25$. The sample proportion of adults who skip breakfast based on a sample of 200 adults is *P*.
    a. Show the sampling distribution of *P*.
    b. What is the probability that the sample proportion will be within ± 0.03 of the population proportion π?
    c. What is the probability that the sample proportion will be within ± 0.05 of the population proportion π?

49. In July 2005, the Pew Research Center released the results of a worldwide survey of attitudes towards Islamic extremism. In Pakistan, 52 per cent of respondents considered that Islamic extremism was a threat to their country. Assume the population proportion was $\pi = 0.52$, and that P is the sample proportion expressing approval in a simple random sample of 800 adults from the population.
    a. Sketch the sampling distribution of *P*.
    b. What is the probability that the sample proportion will be within ±0.02 of the population proportion?
    c. Answer part (b) for a sample of 1600 adults.

50. Suppose that in a particular population, 76% of consumers read the ingredients listed on a product's label, i.e. assume the population proportion is $\pi = 0.76$. A sample of 400 consumers is selected from the population.
    a. Show the sampling distribution of the sample proportion *P*, where *P* is the proportion of the sampled consumers who read the ingredients listed on a product's label.
    b. What is the probability that the sample proportion will be within ± 0.03 of the population proportion?
    c. Answer part (b) for a sample of 750 consumers.

51. Suppose that in a particular country, 71% of Internet users connect their computers to the Internet by normal telephone lines, i.e. assume a population proportion $\pi = 0.71$.
    a. What is the probability that the sample proportion from a simple random sample of 350 Internet users will be within ± 0.05 of the population proportion?
    b. What is the probability that the sample proportion from a simple random sample of 350 Internet users will be 0.75 or greater?

Chapter 7: Sampling and Sampling Distributions

Supplementary Exercises Solutions:

32. a. 6, 8, 5, 4, 1

BP, GlaxoSmithKline, RBS, Shell, HBOS

b. $\frac{N!}{n!(N-n)!} = \frac{10!}{5!(10-5)!} = \frac{3,628,500}{(120)(120)} = 252$

33. 2782, 493, 825, 1807, 289

34. 112, 145, 73, 324, 293, 875, 318, 618

35. Numbers greater than 225 do not apply, so the simple random sample of five contains numbers 147, 113, 215, 2 and 33.

36. a. $\bar{x} = \Sigma x_i / n = \frac{70}{10} = 7$ years

b. $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20.2}{10-1}} = 1.5$ years

37. p = 1117/1400 = 0.80

38. a. p = 426/1007 = 0.423

b. p = 132/1007 = 0.131

c. p = (98 + 320)/1007 = 418/1007= 0.415

39. a. $z = \dfrac{\bar{x} - 34{,}000}{\sigma / \sqrt{n}}$

Error = $\bar{x}$ – 34,000 = 250

n = 30 $\quad z = \dfrac{250}{2000 / \sqrt{30}} = 0.68$

Required probability = 2(0.7517 – 0.5000) = 0.5034

n = 50 $\quad z = \dfrac{250}{2000 / \sqrt{50}} = 0.88$

Required probability = 2(0.8106 – 0.5000) = 0.6212

n = 100 $\quad z = \dfrac{250}{2000 / \sqrt{100}} = 1.25$

Required probability = 2(0.8944 – 0.5000) = 0.7888

n = 200 $\quad z = \dfrac{250}{2000 / \sqrt{200}} = 1.77$

Required probability = 2(0.9616 – 0.5000) = 0.9232

n = 400 $\quad z = \dfrac{250}{2000 / \sqrt{400}} = 2.50$

Required probability = 2(0.9938 – 0.5000) = 0.9876

b. A larger sample increases the probability that the sample mean will be within a specified distance from the population mean. In the salary example, the probability of being within ±250 of μ ranges from 0.5036 for a sample of size 30 to 0.9876 for a sample of size 400.

40. a. Normal distribution, $E(\bar{X}) = 4260$

$\sigma_{\bar{X}} = \sigma / \sqrt{n} = 900/ \leq \sqrt{50} = 127.28$

b. Within €250

$P(4010 \leq \bar{X} \leq 4510)$

$z = \dfrac{4510 - 4260}{127.28} = 1.96$ Cumulative probability = 0.9750

Required probability = 2(0.9750 – 0.5000) = 0.95

c. Within €100

$P(4160 \leq \bar{X} \leq 4360)$

$z = \dfrac{4360 - 4260}{127.28} = 0.79$ Cumulative probability = 0.7852

Required probability = 2(0.7852 – 0.5000) = 0.5704

41. a. Sampling distribution of $\bar{X}$ is Normal, $E(\bar{X}) = 106.4$, $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{4.5}{\sqrt{30}} = 0.822$

$z = \dfrac{1.0}{0.822} = 1.22$, cumulative probability = 0.8888,
required probability = 2(0.8888 – 0.5000) = 0.7776

b. Sampling distribution of $\bar{X}$ is Normal, $E(\bar{X}) = 106.4$, $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{4.5}{\sqrt{50}} = 0.636$

$z = \dfrac{1.0}{0.636} = 1.57$, cumulative probability = 0.9418,
required probability = 2(0.9418 – 0.5000) = 0.8836

c. Sampling distribution of $\bar{X}$ is Normal, $E(\bar{X}) = 106.4\ 6$,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4.5}{\sqrt{100}} = 0.450$$

$z = \frac{1.0}{0.450} = 2.22$, cumulative probability = 0.9868,

required probability = 2(0.9868 – 0.5000) = 0.9736

d. Recommend n = 100, because the probabilities in a and b are below 0.95, in c above 0.95.

42. a. $s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{500}{\sqrt{n}} = 20$

$\sqrt{n}$ = 500/20 = 25 and $n = (25)^2 = 625$

b. For ± 25,

$z = \frac{25}{20} = 1.25$ Cumulative probability = 0.8944

Required probability = 2(0.8944 – 0.5000) = 0.7888

43. a. Normal distribution

$E(\bar{X}) = 3$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{50}} = 0.17$$

b. $$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{0.25}{1.2 / \sqrt{50}} = 1.47$$ Cumulative probability = 0.9292

Required probability = 2(0.9292 – 0.5000) = 0.8584

44. $\mu = 42{,}000 \quad \sigma = 5000$

a. $\sigma_{\bar{x}} = 5000 / \sqrt{50} = 707$

b. $$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{0}{707} = 0$$

$P(\bar{X} > 42{,}000) = P(Z > 0) = 0.50$

c. $$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{1000}{707} = 1.41$$

$P(41{,}000 \le \bar{X} \le 43{,}000) = P(-1.41 \le Z \le 1.41) = 0.9207 - 0.0793 = 0.8414$

d. $\sigma_{\bar{x}} = 5000 / \sqrt{100} = 500$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{1000}{500} = 2.00$$

$P(41{,}000 \le \bar{X} \le 43{,}000) = P(-2 \le Z \le 2) = 0.9772 - 0.0228 = 0.9544$

45. a. $s_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\frac{s}{\sqrt{n}}$

$N = 2000$

$$s_{\bar{x}} = \sqrt{\frac{2000-50}{2000-1}}\frac{144}{\sqrt{50}} = 20.11$$

$N = 5000$

$$s_{\bar{x}} = \sqrt{\frac{5000-50}{5000-1}}\frac{144}{\sqrt{50}} = 20.26$$

$N = 10{,}000$

$$s_{\bar{x}} = \sqrt{\frac{10{,}000-50}{10{,}000-1}}\frac{144}{\sqrt{50}} = 20.31$$

Note: With $n / N \leq 0.05$ for all three cases, common statistical practice would be to ignore the finite population correction factor and use $s_{\bar{x}} = 144/\sqrt{50} = 20.36$ for each case.

b. $N = 2000$

$$z = \frac{25}{20.11} = 1.24 \quad \text{Cumulative probability} = 0.8925$$

Required probability = 2(0.8925 – 0.5000) = 0.7850

$N = 5000$

$$z = \frac{25}{20.26} = 1.23 \quad \text{Cumulative probability} = 0.8907$$

Required probability = 2(0.8907 – 0.5000) = 0.7814

$N = 10{,}000$

$$z = \frac{25}{20.31} = 1.23 \quad \text{Cumulative probability} = 0.8907$$

Required probability = 2(0.8907 – 0.5000) = 0.7814

All probabilities are approximately 0.78

46. a. Normal distribution with $E(P) = 0.15$ and

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{(0.15)(0.85)}{150}} = 0.0292$$

b. $z = \frac{0.18 - 0.15}{0.0292} = 1.03$ Cumulative probability = 0.8485

Required probability = 2(0.8485 – 0.5000) = 0.6970

47. a. Normal with $E(P) = 0.61$ and $\sigma_P = \sqrt{\frac{0.61 \times 0.39}{300}} = 0.02816$

b. $z = \frac{0.03}{0.02816} = 1.07$, cumulative probability = 0.8577,

Required probability = 2(0.8577 – 0.5000) = 0.7154

c. <u>For $n = 600$</u>, $\sigma_P = \sqrt{\frac{0.61 \times 0.39}{600}} = 0.0199$

$z = \frac{0.03}{0.0199} = 1.51$, cumulative probability = 0.9345,

Required probability = 2(0.9345 – 0.5000) = 0.8690

<u>For $n = 1000$</u>, $\sigma_P = \sqrt{\frac{0.61 \times 0.39}{1000}} = 0.0154$

$z = \frac{0.03}{0.0154} = 1.95$, cumulative probability = 0.9744,

Required probability = 2(0.9744 – 0.5000) = 0.9488

48. a. Normal distribution

$E(P) = 0.25$

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{(0.25)(0.75)}{200}} = 0.0306$$

b. $z = \frac{0.03}{0.0306} = 0.98$ Cumulative probability = 0.8365

Required probability = 2(0.8365 – 0.5000) = 0.6730

c. $z = \frac{0.05}{0.0306} = 1.63$ Cumulative probability = 0.9484

Required probability = 2(0.9484 – 0.5000) = 0.8968

49. a. Normal with $E(P) = 0.52$ and $\sigma_P = \sqrt{\frac{0.52 \times 0.48}{800}} = 0.0177$

b. $z = \frac{0.02}{0.0177} = 1.13$, Required probability = 2(0.8708 – 0.5000) = 0.7416

c. For $n = 1600$, $\sigma_P = \sqrt{\frac{0.52 \times 0.48}{1600}} = 0.0125$, $z = \frac{0.02}{0.0125} = 1.60$,

Required probability = 2(0.8708 – 0.5000) = 0.8904

50. a. $E(P) = 0.76$

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.76(1-0.76)}{400}} = 0.0214$$

Normal distribution because $n\pi = 400(0.76) = 304$ and $n(1 - \pi) = 400(0.24) = 96$

b. $z = \frac{0.79 - 0.76}{0.0214} = 1.40$ Cumulative probability = 0.9192

$z = \frac{0.73 - 0.76}{0.0214} = -1.40$ Cumulative probability = 0.0808

Required probability = 0.9192 – 0.0808 = 0.8384

c. $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.76(1-0.76)}{750}} = 0.0156$

$z = \frac{0.79-0.76}{0.0156} = 1.92$ Cumulative probability = 0.9726

$z = \frac{0.73-0.76}{0.0156} = -1.92$ Cumulative probability = 0.0274

Required probability = 0.9726 – 0.0274 = 0.9452

51. a. $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{(0.71)(1-0.71)}{350}} = 0.0243$

$z = \frac{p-\pi}{\sigma_P} = \frac{0.05}{0.0243} = 2.06$ Cumulative probability = 0.94803

Required probability = 2(0.9803 – 0.5000) = 0.9606

b. $z = \frac{p-\pi}{\sigma_P} = \frac{0.75-0.71}{0.0243} = 1.65$ Cumulative probability = 0.9505

Required probability = 1 – 0.9505 = 0.0495

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Eight

Interval Estimation

Textbook Exercises (1-34)

Textbook Exercise Solutions

Supplementary Exercises (35-51)

Supplementary Exercise Solutions

Chapter 8: Interval Estimation

Textbook Exercises:

1 A simple random sample of 40 items results in a sample mean of 25. The population standard deviation is $\sigma = 5$.

a. What is the value of the standard error of the mean, $\sigma_{\bar{X}}$?

b. At 95 per cent confidence, what is the margin of error for estimating the population mean?

2 A simple random sample of 50 items from a population with $\sigma = 6$ results in a sample mean of 32.

a. Construct a 90 per cent confidence interval for the population mean.

b. Construct a 95 per cent confidence interval for the population mean.

c. Construct a 99 per cent confidence interval for the population mean.

3 A simple random sample of 60 items results in a sample mean of 80. The population standard deviation is $\sigma = 15$.

a. Compute the 95 per cent confidence interval for the population mean.

b. Assume that the same sample mean was obtained from a sample of 120 items. Construct a 95 per cent confidence interval for the population mean.

c. What is the effect of a larger sample size on the interval estimate?

4 A 95 per cent confidence interval for a population mean was reported to be 152 to 160. If $\sigma = 15$, what sample size was used in this study?

5 In an effort to estimate the mean amount spent per customer for dinner at a Johannesburg restaurant, data were collected for a sample of 49 customers. Assume a population standard deviation of 40 South African Rand (ZAR).

a. At 95 per cent confidence, what is the margin of error?

b. If the sample mean is ZAR186, what is the 95 per cent confidence interval for the population mean?

6 A survey of small businesses with websites found that the average amount spent on a site was €11 500 per year. Given a sample of 60 businesses and a population standard deviation of $\sigma = €4\,000$, what is the margin of error in estimating the population mean spend per year? Use 95 per cent confidence.

7 A survey of 750 university students found they were paying on average €108 per week in accommodation costs. Assume the population standard deviation for weekly accommodation costs is €22.

a. Construct a 90 per cent confidence interval estimate of the population mean.
b. Construct a 95 per cent confidence interval estimate of the population mean.
c. Construct a 99 per cent confidence interval estimate of the population mean.
d. Discuss what happens to the width of the confidence interval as the confidence level is increased. Does this result seem reasonable? Explain.

8 For a *t* distribution with 16 degrees of freedom, find the area, or probability, in each region.
a. To the right of 2.120
b. To the left of 1.337
c. To the left of −1.746
d. To the right of 2.583
e. Between −2.120 and 2.120
f. Between −1.746 and 1.746

9 Find the *t* value(s) for each of the following cases.
a. Upper tail area of 0.025 with 12 degrees of freedom
b. Lower tail area of 0.05 with 50 degrees of freedom
c. Upper tail area of 0.01 with 30 degrees of freedom
d. Where 90 per cent of the area falls between these two t values with 25 degrees of freedom
e. Where 95 per cent of the area falls between these two t values with 45 degrees of freedom

10 The following sample data are from a normal population: 10, 8, 12, 15, 13, 11, 6, 5.
a. What is the point estimate of the population mean?
b. What is the point estimate of the population standard deviation?
c. With 95 per cent confidence, what is the margin of error for the estimation of the population mean?
d. What is the 95 per cent confidence interval for the population mean?

11 A simple random sample with $n = 54$ provided a sample mean of 22.5 and a sample standard deviation of 4.4.
a. Construct a 90 per cent confidence interval for the population mean.
b. Construct a 95 per cent confidence interval for the population mean.
c. Construct a 99 per cent confidence interval for the population mean.
d. What happens to the margin of error and the confidence interval as the confidence level is increased?

12 Sales personnel for Skillings Distributors submit weekly reports listing the customer contacts made during the week. A sample of 65 weekly reports showed a sample mean of 19.5 customer contacts per week. The sample standard deviation was 5.2. Provide 90 per cent and 95 per cent confidence intervals for the population mean number of weekly customer contacts for the sales personnel.

13 Consumption of alcoholic beverages by young women of drinking age is of concern in the UK and some other European countries. Annual consumption data (in litres) are shown below for a sample of 20 European young women.

266	82	199	174	97
170	222	115	130	169
164	102	113	171	0
93	0	93	110	130

Assuming the population is roughly symmetrically distributed; construct a 95 per cent confidence interval for the mean annual consumption of alcoholic beverages by young European women.

14 The International Air Transport Association surveys business travellers to develop quality ratings for international airports. The maximum possible rating is ten. Suppose a simple random sample of business travellers is selected and each traveller is asked to provide a rating for Kuwait International Airport. The ratings obtained from the sample of 50 business travellers follow. Construct a 95 per cent confidence interval estimate of the population mean rating for Kuwait International.

2	1	8	7	3	1	8	1	7	9	2	9	10	9	7	8	9
1	0	3	0	1	6	2	3	1	6	8	7	7	7	7	7	1
2	5	2	1	2	2	0	2	2	7	0	8	7	0	2	8	

15 Suppose a survey of 40 first-time home buyers finds that the mean of annual household income is €40 000 and the sample standard deviation is €15 300.

a. At 95 per cent confidence, what is the margin of error for estimating the population mean household income?

b. What is the 95 per cent confidence interval for the population mean annual household income for first-time home buyers?

16 Thirty fast-food restaurants including McDonald's and Burger King were visited. During each visit, the customer went to the drive-through and ordered a basic meal such as a burger, fries and drink. The time between pulling up to the order kiosk and receiving the filled order was recorded. The times in minutes for the 30 visits are as follows:

0.9	1.0	1.2	2.2	1.9	3.6	2.8	5.2	1.8	2.1	6.8	1.3	3.0	4.5	2.8
2.3	2.7	5.7	4.8	3.5	2.6	3.3	5.0	4.0	7.2	9.1	2.8	3.6	7.3	9.0

a. Provide a point estimate of the population mean drive-through time at fast-food restaurants.

b. At 95 per cent confidence, what is the margin of error?

c. What is the 95 per cent confidence interval estimate of the population mean?

d. Discuss skewness that may be present in this population. What suggestion would you make for a repeat of this study?

17 A survey by Accountemps asked a sample of 200 executives to provide data on the number of minutes per day office workers waste trying to locate mislabelled, misfiled or misplaced items. Data consistent with this survey are contained in the data set 'ActTemps'.

a. Use 'ActTemps' to develop a point estimate of the number of minutes per day office workers waste trying to locate mislabelled, misfiled or misplaced items.

b. What is the sample standard deviation?

c. What is the 95 per cent confidence interval for the mean number of minutes wasted per day?

18 How large a sample should be selected to provide a 95 per cent confidence interval with a margin of error of 10? Assume that the population standard deviation is 40.

19 The range for a set of data is estimated to be 36.

a. What is the planning value for the population standard deviation?

b. At 95 per cent confidence, how large a sample would provide a margin of error of 3?

c. At 95 per cent confidence, how large a sample would provide a margin of error of 2?

20 Refer to the Scheer Industries example in Section 8.2. Use 6.82 days as a planning value for the population standard deviation.

a. Assuming 95 per cent confidence, what sample size would be required to obtain a margin of error of 1.5 days?

b. If the precision statement was made with 90 per cent confidence, what sample size would be required to obtain a margin of error of 2 days?

21 Suppose you are interested in estimating the average cost of staying for one night in a double room in a three-star hotel in France (outside Paris). Using €30.00 as the planning value for the population standard deviation, what sample size is recommended for each of the following cases? Use €3 as the desired margin of error.

a. A 90 per cent confidence interval estimate of the population mean cost.

b. A 95 per cent confidence interval estimate of the population mean cost.

c. A 99 per cent confidence interval estimate of the population mean cost.

d. When the desired margin of error is fixed, what happens to the sample size as the confidence level is increased? Would you recommend a 99 per cent confidence level be used? Discuss.

22 Suppose the price/earnings (P/E) ratio for stocks listed on a European Stock Exchange have a mean value of 35 and a standard deviation of 18. We want to estimate the population mean P/E ratio for all stocks listed on the exchange. How many stocks should be included in the sample if we want a margin of error of 3? Use 95 per cent confidence.

23 Fuel consumption tests are conducted for a particular model of car. If a 98 per cent confidence interval with a margin of error of 0.2 litres per 100 km is desired, how many cars should be used in the test? Assume that preliminary tests indicate the standard deviation is 0.5 litres per 100 km.

24 In developing patient appointment schedules, a medical centre wants to estimate the mean time that a staff member spends with each patient. How large a sample should be taken if the desired margin of error is 2 minutes at a 95 per cent level of confidence? How large a sample should be taken for a 99 per cent level of confidence? Use a planning value for the population standard deviation of 8 minutes.

25 A simple random sample of 400 individuals provides 100 Yes responses.

a. What is the point estimate of the proportion of the population that would provide Yes responses?
b. What is your estimate of the standard error of the sample proportion?
c. Compute a 95 per cent confidence interval for the population proportion.

26 A simple random sample of 800 elements generates a sample proportion $p = 0.70$.

a. Provide a 90 per cent confidence interval for the population proportion.
b. Provide a 95 per cent confidence interval for the population proportion.

27 In a survey, the planning value for the population proportion is $p^* = 0.35$. How large a sample should be taken to provide a 95 per cent confidence interval with a margin of error of 0.05?

28 At 95 per cent confidence, how large a sample should be taken to obtain a margin of error of 0.03 for the estimation of a population proportion? Assume that past data are not available for developing a planning value for p.

29 A survey of 611 office workers investigated telephone answering practices, including how often each office worker was able to answer incoming telephone calls and how often incoming telephone calls went directly to voice mail. A total of 281 office workers indicated that they never need voice mail and are able to take every telephone call.

a. What is the point estimate of the proportion of the population of office workers who are able to take every telephone call?

b. At 90 per cent confidence, what is the margin of error?
c. What is the 90 per cent confidence interval for the proportion of the population of office workers who are able to take every telephone call?

30 The French market research and polling company CSA carried out surveys to investigate job satisfaction among professionally qualified employees of private companies. A total of 629 professionals were involved in the surveys, of whom 195 said that they were dissatisfied with their employer's recognition of their professional experience.

a. What is the point estimate of the proportion of the population of employees who were dissatisfied with their employer's recognition of their professional experience?
b. At 95 per cent confidence, what is the margin of error?
c. What is the 95 per cent confidence interval for the proportion of the population of employees who were dissatisfied with their employer's recognition of their professional experience?

31 In a sample of 162 companies, 104 reported profits that beat prior estimates, 29 matched estimates, and 29 fell short of prior estimates.

a. What is the point estimate of the proportion that fell short of estimates?
b. Determine the margin of error and provide a 95 per cent confidence interval for the proportion that fell short of estimates.
c. How large a sample is needed if the desired margin of error is 0.05?

32 In early December 2008, the Palestinian Center for Policy and Survey Research carried out an opinion poll among adults in the West Bank and Gaza Strip. Respondents were asked their opinion about the chance of an independent Palestinian state being established alongside Israel in the next five years. Among the 1270 respondents, 34.6 per cent felt there was no chance of this happening.

a. Provide a 95 per cent confidence interval for the population proportion of adults who thought there was no chance of an independent Palestinian state being established alongside Israel in the next five years.
b. Provide a 99 per cent confidence interval for the population proportion of adults who thought there was no chance of an independent Palestinian state being established alongside Israel in the next five years.
c. What happens to the margin of error as the confidence is increased from 95 per cent to 99 per cent?

33 In a survey conducted by ICM Research in the UK, 710 out of 1000 adults interviewed said that, if there were to be a referendum, they would vote for the UK not to join the European currency (the euro). What is the margin of error and what is the interval estimate of the population proportion of British adults who would vote for the UK not to join the European currency? Use 95 per cent confidence.

34 A well-known bank credit card firm wishes to estimate the proportion of credit card holders who carry a non-zero balance at the end of the month and incur an interest charge. Assume that the desired margin of error is 0.03 at 98 per cent confidence.

a. How large a sample should be selected if it is anticipated that roughly 70 per cent of the firm's cardholders carry a non-zero balance at the end of the month?
b. How large a sample should be selected if no planning value for the proportion could be specified?

Chapter 8: Interval Estimation

Textbook Exercises Solutions:

1. a. $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{5}{\sqrt{40}} = 0.79$

   b. At the 95% confidence level,

   $$z\frac{\sigma}{\sqrt{n}} = 1.96\left(\frac{5}{\sqrt{40}}\right) = 1.55$$

2. The confidence interval is $\bar{x} \pm z_{a/2}\dfrac{s}{\sqrt{n}}$. Here, $\bar{x} = 32$, $\sigma = 6$, $n = 50$

   a. $z_{\alpha/2} = 1.645$

   The confidence interval is $32 \pm 1.645\,(6/\sqrt{50}) = 32 \pm 1.4$ or 30.6 to 33.4

   b. $z_{\alpha/2} = 1.96$

   $32 \pm 1.96\,(6/\sqrt{50}) = 32 \pm 1.66$ or 30.34 to 33.66

   c. $z_{\alpha/2} = 2.576$

   The confidence interval is $32 \pm 2.576\,(6/\sqrt{50}) = 32 \pm 2.19$ or 29.81 to 34.19

3. a. $80 \pm 1.96\,(15/\sqrt{60})$

   80 ± 3.8 or 76.2 to 83.8

   b. $80 \pm 1.96\,(15/\sqrt{120})$

   80 ± 2.68 or 77.32 to 82.68

   c. Larger sample provides a smaller margin of error.

4. Sample mean $\bar{x} = \dfrac{160 - 152}{2} = 156$

Margin of error = 160 – 156 = 4

$1.96(\sigma / \sqrt{n}) = 4$

$\sqrt{n} = 1.96\sigma / 4 = 1.96(15) / 4 = 7.35$

$n = (7.35)^2 = 54$

5. a. Margin of error = $z_{0.025}(\sigma / \sqrt{n}) = 1.96(40 / \sqrt{49}) = 11.2$

b. 95% confidence interval is 186 ± 11.2 or 174.8 to 197.2

6. The margin of error is $z_{\alpha/2} \dfrac{s}{\sqrt{n}}$. Here, $\sigma = 4000$, $n = 60$

The margin of error is $1.96(4000 / \sqrt{60}) = 1012$

7. a. $\bar{x} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$

108 ± 1.645 $(22 / \sqrt{750})$

108 ± 1.32 or 106.68 to 109.32

b. 108 ± 1.96 $(22 / \sqrt{750})$

108 ± 1.57 or 106.43 to 109.57

c. 108 ± 2.576 $(22 / \sqrt{750})$

108 ± 2.07 or 105.93 to 110.07

d. Width of interval increases. To be more confident of including the true population mean in the interval, the interval must be wider.

8. a. 0.025

b. 1 – 0.10 = 0.90

c. 0.05

d. 0.01

e. 1 – 2(0.025) = 0.95

f. 1 – 2(0.05) = 0.90

9. a. 2.179

b. −1.676

c. 2.457

d. Use 0.05 column, −1.708 and 1.708

e. Use 0.025 column, −2.014 and 2.014

10. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{80}{8} = 10$

b.

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
10	0	0
8	-2	4
12	2	4
15	5	25
13	3	9
11	1	1
6	-4	16
5	-5	25
		84

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{84}{7}} = 3.464$$

c. $t_{0.025}(s/\sqrt{n}) = 2.365(3.464/\sqrt{8}) = 2.9$

d. $\bar{x} \pm t_{0.025}(s/\sqrt{n})$

10 ± 2.9 or 7.1 to 12.9

11. $\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$ $df = 53$

a. $22.5 \pm 1.674\,(4.4/\sqrt{54})$

22.5 ± 1 or 21.5 to 23.5

b. $22.5 \pm 2.006\,(4.4/\sqrt{54})$

22.5 ± 1.2 or 21.3 to 23.7

c. $22.5 \pm 2.672\,(4.4/\sqrt{54})$

22.5 ± 1.6 or 20.9 to 24.1

d. As the confidence level increases, there is a larger margin of error and a wider confidence interval.

12. $\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$

<u>90% confidence</u> $df = 64$ $t_{0.05} = 1.669$

$19.5 \pm 1.669\,(5.2/\sqrt{65})$

19.5 ± 1.08 or 18.42 to 20.58

<u>95% confidence</u> $df = 64$ $t_{0.025} = 1.998$

$19.5 \pm 1.998\,(5.2/\sqrt{65})$

19.5 ± 1.29 or 18.21 to 20.79

13. $\bar{x}$ = 2600/20 = 130.0 and $s = \sqrt{81244/19} = 65.4$

The confidence interval is $\bar{x} \pm t_{0.025}(s/\sqrt{n})$, with $df = 19$ and $t_{0.025} = 2.093$

The confidence interval is $130.0 \pm 2.093\,(65.4/\sqrt{20}) = 130.0 \pm 30.6$ or 99.4 to 160.6

14. Using Minitab, SPSS or Excel, $\bar{x} = 6.34$ and $s = 2.163$

The confidence interval is $\bar{x} \pm t_{0.025}(s/\sqrt{n})$, with $df = 49$ and $t_{0.025} = 2.010$

The confidence interval is $6.34 \pm 2.010\,(2.163/\sqrt{50}) = 6.34 \pm 0.61$ or 5.73 to 6.95

15. a. $\bar{x} \pm t_{0.025}(s/\sqrt{n})$ $df = 39$ $t_{0.025} = 2.023$

Margin of error = $2.023\,(15,300/\sqrt{40}) = 4894$ (€)

b. 95% confidence interval is 40,000 ± 4894 or €35,106 to €44,894

16. Using Minitab, SPSS or Excel, $\bar{x} = 3.8$ and $s = 2.257$

a. $\bar{x} = 3.8$ minutes

b. $t_{0.025}(s/\sqrt{n})$ $df = 29$ $t_{0.025} = 2.045$

$2.045\,(2.257/\sqrt{30}) = 0.84$

c. $\bar{x} \pm t_{0.025}(s/\sqrt{n})$

3.8 ± 0.84 or 2.96 to 4.64

d. There is a modest positive skewness in this data set. This can be expected to exist in the population. While the above results are acceptable, considering a larger sample next time would be a good strategy.

17. a. Using Minitab, SPSS or Excel, $\bar{x} = 49.8$ minutes

b. Using Minitab, SPSS or Excel, $s = 15.99$ minutes

c. $\bar{x} \pm t_{0.025}(s/\sqrt{n})$ $df = 199$ $t_{0.025} \approx 1.96$

$49.8 \pm 1.96\ (15.99/\sqrt{200})$

49.8 ± 2.22 or 47.58 to 52.02

18. The required sample size is given by $n = \dfrac{z_{0.025}^2\sigma^2}{E^2} = \dfrac{(1.96)^2(40)^2}{10^2} = 61.47.$ Use $n = 62$

19. a. Planning value of σ = range/4 = 36/4 = 9

b. $n = \dfrac{z_{0.025}^2\sigma^2}{E^2} = \dfrac{(1.96)^2(9)^2}{3^2} = 34.57.$ Use $n = 35$

c. $n = \dfrac{(1.96)^2(9)^2}{2^2} = 77.79.$ Use $n = 78$

20. a. $n = \dfrac{(1.96)^2(6.82)^2}{(1.5)^2} = 79.41.$ Use $n = 80$

b. $n = \dfrac{(1.645)^2(6.82)^2}{2^2} = 31.47.$ Use $n = 32$

21. a. $n = \dfrac{(1.645)^2(30)^2}{(3)^2} = 270.6.$ Use $n = 271$

b. $n = \dfrac{(1.96)^2(30)^2}{(3)^2} = 384.2.$ Use $n = 385$

c. $n = \dfrac{(2.576)^2(30)^2}{(3)^2} = 663.6.$ Use $n = 664$

d. Sample size increases as confidence level increases. 99% interval requires much larger sample size than 95% interval.

22. $n = \dfrac{(1.96)^2(18)^2}{(3)^2} = 138.3$

Use a sample size of around 139.

23. $z_{0.01} = 2.33$

$n = \dfrac{(2.33)^2(1.0)^2}{(0.4)^2} = 33.9.$ Use $n = 34$

24. 95% level of confidence: $n = \dfrac{(1.96)^2(8)^2}{2^2} = 61.47.$ Use $n = 62$

99% level of confidence: $n = \dfrac{(2.576)^2(8)^2}{2^2} = 106.17.$ Use $n = 107$

25. a. $p = 100/400 = 0.25$

b. $\sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.25(0.75)}{400}} = 0.0217$

c. $p \pm z_{0.025}\sqrt{\dfrac{p(1-p)}{n}}$

$0.25 \pm 1.96\ (0.0217)$

0.25 ± 0.0424 or 0.2076 to 0.2924

26. a. The confidence interval is $p \pm z_{a/2}\sqrt{\frac{p(1- p)}{n}} = 0.70 \pm 1.645\sqrt{\frac{0.70(0.30)}{800}}$

$= 0.70 \pm 0.0267$ or 0.6733 to 0.7267

b. $0.70 \pm 1.96\sqrt{\frac{0.70(0.30)}{800}} = 0.70 \pm 0.0318$ or 0.6682 to 0.7318

27. $n = \frac{z^2_{0.025}p^*(1-p^*)}{E^2} = \frac{(1.96)^2(0.35)(0.65)}{(0.05)^2} = 349.59.$ Use $n = 350$

28. Use planning value $p^* = 0.50$

$n = \frac{(1.96)^2(0.50)(0.50)}{(0.03)^2} = 1067.11$ Use $n = 1068$

29. a. $p = 281/611 = 0.4599$ (46%)

b. $z_{0.05}\sqrt{\frac{p(1-p)}{n}} = 1.645\sqrt{\frac{0.4599(1-0.4599)}{611}} = 0.0332$

c. $p \pm 0.0332$

0.4599 ± 0.0332 or 0.4267 to 0.4931

30. a. $p = 195/629 = 0.310$ (31%)

b. $z_{0.025}\sqrt{\frac{p(1-p)}{n}} = 1.96\sqrt{\frac{0.310(1-0.310)}{629}} = 0.0361$

c. $p \pm 0.0361$

0.310 ± 0.0361or 0.274 to 0.346

31. a. $p = 29/162 = 0.179$

b. Margin of error = $z_{0.025}\sqrt{\frac{p(1-p)}{n}} = 1.96\sqrt{\frac{0.179(1-0.179)}{162}} = 0.059$

95% confidence interval is 0.179 ± 0.059 or 0.120 to 0.238

c. For $E = 0.05$, $n = \frac{(1.96)^2(0.179)(1-0.179)}{(0.05)^2} = 225.8$. Use $n = 226$

32. a. $p \pm z_{0.025}\sqrt{\frac{p(1-p)}{n}} = 0.346 \pm 1.96\sqrt{\frac{0.346 \times 0.654}{1270}} = 0.346 \pm 0.026$, or 0.320 to 0.372

b. $p \pm z_{0.005}\sqrt{\frac{p(1-p)}{n}} = 0.346 \pm 2.576\sqrt{\frac{0.346 \times 0.654}{1270}} = 0.346 \pm 0.034$, or 0.312 to 0.380

c. The margin of error increases.

33. $p = 710/1000 = 0.710$

Margin of error = $z_{0.025}\sqrt{\frac{p(1-p)}{n}} = 1.96\sqrt{\frac{0.710(1-0.710)}{1000}} = 0.028$

95% confidence interval is 0.710 ± 0.028 or 0.682 to 0.738

.

34. The required sample size is given by $n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2}$

a. $n = \frac{(2.33)^2(0.70)(0.30)}{(0.03)^2} = 1266.74$. Use $n = 1267$.

b. $n = \frac{(2.33)^2(0.50)(0.50)}{(0.03)^2} = 1508.03$. Use $n = 1509$.

Chapter 8: Interval Estimation

Supplementary Exercises:

35. Nielsen Media Research reported sample results indicating that the household mean television viewing time during the 8 p.m. to 11 p.m. time period was 8.5 hours per week (*The World Almanac*, 2003). Given a sample size of 300 households and a population standard deviation of σ = 3.5 hours, what is the 95% confidence interval estimate of the mean television viewing time per week during the 8 p.m. to 11 p.m. time period?

36. Suppose that a survey of 10 restaurants in the Pizza Palace chain showed a sample mean customer satisfaction index of 71 (based on a 0 to 100 scale). Past data indicate that the population standard deviation of the index has been relatively stable with σ = 5.
    a. What assumption should the researcher be willing to make if a margin of error needs to be calculated?
    b. Using 95% confidence, what is the margin of error?
    c. What is the margin of error if 99% confidence is required?

37. The UNITE/MORI 2004 survey of student experiences estimated that university students in the UK paid on average £54 per week in accommodation costs. Assume that this average is based on a sample of 755 students and that the population standard deviation for weekly accommodation costs is £11.
    a. Construct a 90 per cent confidence interval estimate of the population mean.
    b. Construct a 95 per cent confidence interval estimate of the population mean.
    c. Construct a 99 per cent confidence interval estimate of the population mean.
    d. Discuss what happens to the width of the confidence interval as the confidence level is increased. Does this result seem reasonable? Explain.

38. Consider a survey of 600 households, which finds that households intend to spend an average of €949 during the Christmas and New Year holiday season. The sample standard deviation was €175.
    a. With 95% confidence, what is the margin of error?
    b. What is the 95% confidence interval estimate of the population mean?
    c. The previous year, the population mean expenditure per household was €932. Discuss the change in holiday season expenditures over the one-year period.

39. The data following are a sample of the times in minutes given over to advertising during 20 half-hour programmes during peak evening viewing time on a commercial TV channel.

6.0	6.6	5.8	7.0	6.3	6.2	7.2	5.7	6.4	7.0
6.5	6.2	6.0	6.5	7.2	7.3	7.6	6.8	6.0	6.2

Assume a normal population and provide a point estimate and a 95% confidence interval for the mean number of advertising minutes on half-hour, peak-time television programmes on this channel.

40. Suppose a preliminary sample of chief executive officers (CEOs) of large European companies shows that the standard deviation of salaries plus bonuses is €675,000. How many CEOs should be in a sample if we want to estimate the population mean annual salary plus bonus with a margin of error of €100,000? Use 95% confidence.

41. The film *Harry Potter and the Sorcerer's Stone* broke the U.S. box office debut record previously held by *The Lost World: Jurassic Park* (*The Wall Street Journal*, November 19, 2001). A sample of 100 cinemas showed that the mean three-day weekend gross was $25,467 per cinema. The sample standard deviation was $4980.
    a. What is the margin of error for this study? Use 95% confidence.
    b. What is the 95% confidence interval estimate for the population mean weekend gross per cinema?
    c. *The Lost World* took in $72.1 million in its first three-day weekend. *Harry Potter and the Sorcerer's Stone* was shown in 3672 cinemas. Calculate an estimate of the total *Harry Potter and the Sorcerer's Stone* took in during its first three-day weekend.
    d. An Associated Press article claimed *Harry Potter* "shattered" the box office debut record held by *The Lost World.* Do your results agree with this claim?

42. Suppose that starting salaries for university graduates with degrees in engineering are expected to be between €30,000 and €45,000. Assume that a 95% confidence interval estimate of the population mean annual starting salary is required. What is the planning value for the population standard deviation? How large a sample should be taken if the required margin of error is
    a. €500?
    b. €200?
    c. €100?
    d. Would you recommend trying to obtain the €100 margin of error? Explain.

43. During the third quarter of 2005, the price/earnings (P/E) ratio for stocks listed on the Hong Kong Stock Exchange generally ranged from 1 to 100. Assume that we want to estimate the population mean P/E ratio for all stocks listed on the exchange. How many stocks should be included in the sample if we want a margin of error of 3? Use 95 per cent confidence.

44. File "Flights"

US Airways conducted a number of studies indicating that substantial savings could be obtained by encouraging Dividend Miles frequent-flyer customers to redeem miles and schedule award flights online (*US Airways Attache*, February 2003). One study collected data on the amount of time required to redeem miles and book an award flight over the telephone. A sample showing the time in minutes required for each of 150 award flights booked by telephone is contained in the data set "Flights". Use Minitab, SPSS or Excel to help answer the following questions.

a. What is the sample mean number of minutes required to book an award flight by telephone?
b. Construct a 95% confidence interval for the population mean time to book an award flight by telephone?
c. Assume a telephone ticket agent works 7.5 hours per day. How many award flights can one ticket agent be expected to handle a day?
d. Discuss why this information supported US Airways plans to use an online system to reduce costs.

45. A survey by the Society for Human Resource Management asked 346 job seekers why employees change jobs so frequently (*The Wall Street Journal*, March 28, 2000). The answer selected most (152 times) was "higher compensation elsewhere."

a. What is the point estimate of the proportion of job seekers who would select "higher compensation elsewhere" as the reason for changing jobs?
b. What is the 95% confidence interval estimate of the population proportion?

46. Consider a survey of 400 users of a sports website, which shows that 26% of the users were women.

a. At 95% confidence, what is the margin of error associated with the estimated proportion of users who are women?
b. What is the 95% confidence interval for the population proportion of the website users who are women?
c. How large a sample should be taken if the desired margin of error is 0.03?

47. In a survey of 369 working parents, 200 said they spend too little time with their children because of work commitments.
    a. What is the point estimate of the proportion of the population of working parents who feel they spend too little time with their children because of work commitments?
    b. At 95% confidence, what is the margin of error?
    c. What is the 95% confidence interval estimate of the population proportion of working parents who feel they spend too little time with their children because of work commitments?

48. Consider a sample survey in which 340 of 500 employees said they would prefer better health insurance to an increase in salary.
    a. What is the point estimate of the population proportion of employees who would prefer better health insurance?
    b. What is the 95% confidence interval estimate of the population proportion?

49. Suppose that the primary purpose of a pre-election poll is to obtain an estimate of the proportion of potential voters who favour each candidate. Assume a planning value of $p^* = 0.50$ and a 95% confidence level.
    a. For $p^* = 0.50$, what is the margin of error for a poll of 500 voters?
    b. Closer to election, better precision and smaller margins of error are desired. Assume the following margins of error are required for surveys to be conducted during the pre-election campaign. Compute the recommended sample size for each survey.

3 months before	0.04
2 months before	0.03
1 month before	0.02
Pre-election day	0.01

50. Suppose that in a sample of $n = 1993$ business travellers who responded to a survey, 618 listed a frequent-flyer program as the most important factor in choosing an airline.
    a. What is the point estimate of the proportion of the population of business travellers who believe a frequent-flyer program is the most important factor when choosing an airline?
    b. Construct a 95% confidence interval estimate of the population proportion.
    c. How large a sample would be required to report the margin of error of 0.01 at 95% confidence?

51. The *2003 Statistical Abstract of the United States* reported the percentage of people 18 years of age and older who smoke. Suppose that a study designed to collect new data on smokers and non-smokers uses a preliminary estimate of the proportion who smoke of 0.30.
    a. How large a sample should be taken to estimate the proportion of smokers in the population with a margin of error of 0.02? Use 95% confidence.
    b. Assume that the study uses your sample size recommendation in part (a) and finds 520 smokers. What is the point estimate of the proportion of smokers in the population?
    c. What is the 95% confidence interval for the proportion of smokers in the population?

Supplementary Exercises Solutions:

35. $\bar{x} \pm z_{0.025}(\sigma / \sqrt{n})$

$8.5 \pm 1.96(3.5/\sqrt{300})$

8.5 ± 0.4 or 8.1 to 8.9

36. a. Since *n* is small, an assumption that the population is at least approximately normal is required.

b. $z_{0.025}(\sigma / \sqrt{n}) = 1.96(5/\sqrt{10}) = 3.1$

c. $z_{0.005}(\sigma / \sqrt{n}) = 2.576(5/\sqrt{10}) = 4.1$

37. a. $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$54 \pm 1.645\,(11/\sqrt{755})$

54 ± 0.66 or 53.34 to 54.66

b. $54 \pm 1.96\,(11/\sqrt{755})$

54 ± 0.78 or 53.22 to 54.78

c. $54 \pm 2.576\,(11/\sqrt{755})$

54 ± 1.03 or 52.97 to 55.03

d. Width of interval increases. To be more confident of including the true population mean in the interval, the interval must be wider.

38. a. $t_{0.025}(s/\sqrt{n})$ $df = 599$

Use ∞ df row, $t_{0.025} = 1.96$

$1.96\,(175/\sqrt{600}) = 14$

b. $\bar{x} \pm t_{0.025}(s/\sqrt{n})$

649 ± 14 or 635 to 663

c. At 95% confidence, the population mean is between €635 and €663. This is slightly above the prior year's €632 level, so holiday spending is increasing. The point estimate of the slight increase is €649 – €632 = €17 or 2.7% per household.

39. $\bar{x} = \Sigma x_i / n = 6.53$ minutes

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = 0.54 \text{ minutes}$$

$\bar{x} \pm t_{0.025}(s/\sqrt{n})$ $df = 19$

$6.53 \pm 2.093\,(0.54/\sqrt{20})$

6.53 ± 0.25 or 6.28 to 6.78

40. $$n = \frac{(1.96)^2(675,000)^2}{100,000^2} = 175.03 \quad \text{Use } n = 176$$

41. a. $t_{0.025}(s/\sqrt{n})$ $df = 99$ $t_{0.025} = 1.984$

$1.984\,(4980/\sqrt{100}) = 998$

b. $\bar{x} \pm 998$

25467 ± 998 or $24,479 to $26,455

c. 3672($25,467) = $93,514,824

d. *Harry Potter* beat *Lost World* by $93.5 – $72.1 = $21.4 million. This is a 21.4/72.1(100) = 30% increase in the first weekend. The words "shatter the record" are justified.

42. Planning value $\sigma = \dfrac{45,000 - 30,000}{4} = 3750$

a. $n = \dfrac{z_{0.025}^2 \sigma^2}{E^2} = \dfrac{(1.96)^2 (3750)^2}{(500)^2} = 216.09$ Use $n = 217$

b. $n = \dfrac{(1.96)^2 (3750)^2}{(200)^2} = 1350.56$ Use $n = 1351$

c. $n = \dfrac{(1.96)^2 (3750)^2}{(100)^2} = 5402.25$ Use $n = 5403$

d. Sampling 5403 college graduates to obtain the €100 margin of error would be viewed as too expensive and too much effort by most researchers.

43. Approximate estimate of σ = range/4 $\approx$ 99/4 = 24.75

$$n = \frac{(1.96)^2 (24.75)^2}{(3)^2} = 261.5$$

Use a sample size of around 265.

44. a. $\bar{x}$ = 14 minutes

b. 13.381 to 14.619

c. 7.5 hours = 7.5(60) = 450 minutes per day. An average of 450/14 = 32 reservations per day if no idle time. Assuming perhaps 15% idle time or time on something other than reservations, this could be reduced to 27 reservations per day.

d. For large airlines, there are many telephone calls such as these per day. Using the online reservations would reduce the telephone reservation staff and payroll. Adding in a reduction in total benefit costs, a change to online reservations could provide a sizeable cost reduction for the airline.

45. a. $p = 152/346 = 0.4393$

b. $\sigma_P = \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.4393(1-0.4393)}{346}} = 0.0267$

$p \pm z_{0.025}\sigma_P = 0.4393 \pm 1.96(0.0267)$

$= 0.4393 \pm 0.0523$ or 0.3870 to 0.4916

46. a. $1.96\sqrt{\dfrac{p(1-p)}{n}} = 1.96\sqrt{\dfrac{(0.26)(0.74)}{400}} = 0.0430$

b. 0.26 ± 0.0430 or 0.2170 to 0.3030

c. $n = \dfrac{1.96^2(0.26)(0.74)}{(0.03)^2} = 821.25$ Use $n = 822$

47. a. $p = 200/369 = 0.5420$

b. $1.96\sqrt{\dfrac{p(1-p)}{n}} = 1.96\sqrt{\dfrac{(0.5420)(0.4580)}{369}} = 0.0508$

c. 0.5420 ± 0.0508 or 0.4912 to 0.5928

48. a. $p = 340/500 = 0.68$

b. $\sigma_P = \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.68(1-0.68)}{500}} = 0.0209$

$p \pm z_{0.025}\sigma_P = 0.68 \pm 1.96(0.0209)$

$= 0.68 \pm 0.0409$ or 0.6391 to 0.7209

49. a. $\sigma_P = \sqrt{\dfrac{p^*(1-p^*)}{n}} = \sqrt{\dfrac{0.50(1-0.50)}{491}} = 0.0226$

$z_{0.025}\sigma_P = 1.96(0.0226) = 0.0442$

b. $n = \dfrac{z^2_{0.025}\, p^*(1-p^*)}{E^2}$

3 months before $n = \dfrac{1.96^2(0.50)(1-0.50)}{0.04^2} = 600.25$ Use $n = 601$

2 months before $n = \dfrac{1.96^2(0.50)(1-0.50)}{0.03^2} = 1067.11$ Use $n = 1068$

1 month before $n = \dfrac{1.96^2(0.50)(1-0.50)}{0.02^2} = 2401$

Pre-Election Day $n = \dfrac{1.96^2(0.50)(1-0.50)}{0.01^2} = 9604$

50. a. $p = 618/1993 = 0.3101$

b. $p \pm 1.96\sqrt{\dfrac{p(1-p)}{1993}} = 0.3101 \pm 1.96\sqrt{\dfrac{(0.3101)(0.6899)}{1993}}$

$= 0.3101 \pm 0.0203$ or 0.2898 to 0.3304

c. $n = \dfrac{z^2 p^*(1-p^*)}{E^2}$

$z = \dfrac{(1.96)^2(0.3101)(0.6899)}{(0.01)^2} = 8218.64$ Use $n = 8219$

51. a. $n = \dfrac{(1.96)^2(0.3)(0.7)}{(0.02)^2} = 2016.84$ Use $n = 2017$

b. $p = 520/2017 = 0.2578$

c. $p \pm 1.96\sqrt{\dfrac{p(1-p)}{n}} = 0.2578 \pm 1.96\sqrt{\dfrac{(0.2578)(0.7422)}{2017}}$

$= 0.2578 \pm 0.0191$ or 0.2387 to 0.2769

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Nine

Hypothesis Tests

Textbook Exercises (1-43)

Textbook Exercise Solutions

Supplementary Exercises (44-61)

Supplementary Exercise Solutions

Chapter 9: Hypothesis Tests

Textbook Exercises:

1 The manager of the Costa Resort Hotel stated that the mean weekend guest bill is €600 or less. A member of the hotel's accounting staff noticed that the total charges for guest bills have been increasing in recent months. The accountant will use a sample of weekend guest bills to test the manager's claim.

a. Which form of the hypotheses should be used to test the manager's claim? Explain.

H_0: $\mu \geq 600$ H_0: $\mu \leq 600$ H_0: $\mu = 600$

H_1: $\mu < 600$ H_1: $\mu > 600$ H_1: $\mu \neq 600$

b. What conclusion is appropriate when H_0 cannot be rejected?

c. What conclusion is appropriate when H_0 can be rejected?

2 The manager of a car dealership is considering a new bonus plan designed to increase sales volume. Currently, the mean sales volume is 14 cars per month. The manager wants to conduct a research study to see whether the new bonus plan increases sales volume. To collect data on the plan, a sample of sales personnel will be allowed to sell under the new bonus plan for a one-month period.

a. Formulate the null and alternative hypotheses most appropriate for this research situation.

b. Comment on the conclusion when H_0 cannot be rejected.

c. Comment on the conclusion when H_0 can be rejected.

3 A production line operation is designed to fill cartons with laundry detergent to a mean weight of 0.75 kg. A sample of cartons is periodically selected and weighed to determine whether under-filling or over-filling is occurring. If the sample data lead to a conclusion of under-filling or over-filling, the production line will be shut down and adjusted to obtain proper filling.

a. Formulate the null and alternative hypotheses that will help in deciding whether to shut down and adjust the production line.

b. Comment on the conclusion and the decision when H_0 cannot be rejected.

c. Comment on the conclusion and the decision when H_0 can be rejected.

4 Because of high production-changeover time and costs, a director of manufacturing must convince management that a proposed manufacturing method reduces costs before the new method can be implemented. The current production method operates with a mean cost of €320 per hour. A research study will measure the cost of the new method over a sample production period.

a. Formulate the null and alternative hypotheses most appropriate for this study.

b. Comment on the conclusion when H_0 cannot be rejected.

c. Comment on the conclusion when H_0 can be rejected.

5 The label on a container of yoghourt claims that the yoghourt contains an average of 1 gram of fat or less. Answer the following questions for a hypothesis test that could be used to test the claim on the label.

a. Formulate the appropriate null and alternative hypotheses.

b. What is the Type I error in this situation? What are the consequences of making this error?

c. What is the Type II error in this situation? What are the consequences of making this error?

6 Carpetland salespersons average €5000 per week in sales. The company's chief executive officer proposes a remuneration plan with new selling incentives. The CEO hopes that the results of a trial selling period will enable him to conclude that the remuneration plan increases the average sales per salesperson.

a. Formulate the appropriate null and alternative hypotheses.

b. What is the Type I error in this situation? What are the consequences of making this error?

c. What is the Type II error in this situation? What are the consequences of making this error?

7 Suppose a new production method will be implemented if a hypothesis test supports the conclusion that the new method reduces the mean operating cost per hour.

a. State the appropriate null and alternative hypotheses if the mean cost for the current production method is €320 per hour.

b. What is the Type I error in this situation? What are the consequences of making this error?

c. What is the Type II error in this situation? What are the consequences of making this error?

Note to student: Some of the exercises that follow ask you to use the p-value approach and others ask you to use the critical value approach. Both methods will provide the same hypothesis testing conclusion. We provide exercises with both methods to give you practice using both. In later sections and in following chapters, we will generally emphasize the *p*-value approach as the preferred method, but you may select either based on personal preference.

8 Consider the following hypothesis test:

$$H_0: \mu \geq 20$$
$$H_1: \mu < 20$$

A sample of 50 gave a sample mean of 19.4. The population standard deviation is 2.

a. Compute the value of the test statistic.

b. What is the *p*-value?

c. Using $\alpha = 0.05$, what is your conclusion?

d. What is the rejection rule using the critical value? What is your conclusion?

9 Consider the following hypothesis test:

$$H_0: \mu = 15$$
$$H_1: \mu \neq 15$$

A sample of 50 provided a sample mean of 14.15. The population standard deviation is 3.

a. Compute the value of the test statistic.

b. What is the *p*-value?

c. At $\alpha = 0.05$, what is your conclusion?

d. What is the rejection rule using the critical value? What is your conclusion?

10 Consider the following hypothesis test:

$$H_0: \mu \leq 50$$
$$H_1: \mu > 50$$

A sample of 60 is used and the population standard deviation is 8. Use the critical value approach to state your conclusion for each of the following sample results. Use $\alpha = 0.05$.

a. $\bar{x} = 52.5$
b. $\bar{x} = 51.0$
c. $\bar{x} = 51.8$

11 Suppose that the mean length of the working week for a population of workers has been previously reported as 39.2 hours. We would like to take a current sample of workers to see whether the mean length of a working week has changed from the previously reported 39.2 hours.

a. State the hypotheses that will help us determine whether a change occurred in the mean length of a working week.
b. Suppose a current sample of 112 workers provided a sample mean of 38.5 hours. Use a population standard deviation $\sigma = 4.8$ hours. What is the *p*-value?
c. At $\alpha = 0.05$, can the null hypothesis be rejected? What is your conclusion?
d. Repeat the preceding hypothesis test using the critical value approach.

12 Suppose the national mean sales price for new two-bedroom houses is £181 900. A sample of 40 new two-bedroom house sales in the north-east of England showed a sample mean of £166 400. Use a population standard deviation of £33 500.

a. Formulate the null and alternative hypotheses that can be used to determine whether the sample data support the conclusion that the population mean sales price for new two-bedroom houses in the north-east is less than the national mean of £181 900.
b. What is the value of the test statistic?
c. What is the *p*-value?
d. At $\alpha = 0.01$, what is your conclusion?

13 Fowler Marketing Research bases charges to a client on the assumption that telephone surveys can be completed in a mean time of 15 minutes or less per interview. If a longer mean interview time is necessary, a premium rate is charged. Suppose a sample of 35 interviews shows a sample mean of 17 minutes. Use $\sigma = 4$ minutes. Is the premium rate justified?

a. Formulate the null and alternative hypotheses for this application.
b. Compute the value of the test statistic.
c. What is the *p*-value?
d. At $\alpha = 0.01$, what is your conclusion?

14 CCN and ActMedia provided a television channel targeted to individuals waiting in supermarket checkout lines. The channel showed news, short features, and advertisements. The length of the programme was based on the assumption that the population mean time a shopper stands in a supermarket checkout line is eight minutes. A sample of actual waiting times will be used to test this assumption and determine whether actual mean waiting time differs from this standard.

a. Formulate the hypotheses for this application.
b. A sample of 120 shoppers showed a sample mean waiting time of eight and a half minutes. Assume a population standard deviation $\sigma = 3.2$ minutes. What is the *p*-value?
c. At $\alpha = 0.05$, what is your conclusion?
d. Compute a 95 per cent confidence interval for the population mean. Does it support your conclusion?

15 During the global economic upheavals in late 2008, research companies affiliated to the Worldwide Independent Network of Market Research carried out polls in 17 countries to assess people's views on the economic outlook. One of the questions asked respondents to rate their trust in their government's management of the financial situation, on a 0 to 10 scale (10 being maximum trust). Suppose the worldwide population mean on this trust question was 5.2, and we are interested in the question of whether the population mean in Germany was different from this worldwide mean.

a. State the hypotheses that could be used to address this question.
b. In the Germany survey, respondents gave the government a mean trust score of 4.0. Suppose the sample size in Germany was 1050, and the population standard deviation score was $\sigma = 2.9$. What is the 95 per cent confidence interval estimate of the population mean trust score for Germany?
c. Use the confidence interval to conduct a hypothesis test. Using $\alpha = 0.05$, what is your conclusion?

16 A production line operates with a mean filling weight of 500 grams per container. Over-filling or under-filling presents a serious problem and when detected requires the operator to shut down the production line to readjust the filling mechanism. From past data, a population standard deviation $\sigma = 25$ grams is assumed. A quality control inspector selects a sample of 30 items every hour and at that time makes the decision of whether to shut down the line for readjustment. The level of significance is $\alpha = 0.05$.

a. State the hypotheses in the hypothesis test for this quality control application.
b. If a sample mean of 510 grams were found, what is the p-value? What action would you recommend?
c. If a sample mean of 495 grams were found, what is the p-value? What action would you recommend?
d. Use the critical value approach. What is the rejection rule for the preceding hypothesis testing procedure? Repeat parts (b) and (c). Do you reach the same conclusion?

17 Consider the following hypothesis test:

$$H_0: \mu \leq 12$$
$$H_1: \mu > 12$$

A sample of 25 provided a sample mean $\bar{x} = 14$ and a sample standard deviation s = 4.32.

a. Compute the value of the test statistic.
b. What does the t distribution table (Table 2 in Appendix B) tell you about the p-value?
c. At $\alpha = 0.05$, what is your conclusion?
d. What is the rejection rule using the critical value? What is your conclusion?

18 Consider the following hypothesis test:

$$H_0: \mu = 18$$
$$H_1: \mu \neq 18$$

A sample of 48 provided a sample mean $\bar{x} = 17$ and a sample standard deviation s = 4.5.

a. Compute the value of the test statistic.
b. What does the t distribution table (Table 2 in Appendix B) tell you about the p-value?
c. At $\alpha = 0.05$, what is your conclusion?
d. What is the rejection rule using the critical value? What is your conclusion?

19 Consider the following hypothesis test:

$$H_0: \mu \geq 45$$
$$H_1: \mu < 45$$

A sample of size 36 is used. Using $\alpha = 0.01$, identify the p-value and state your conclusion for each of the following sample results.

a. $\bar{x} = 44$ and $s = 5.2$
b. $\bar{x} = 43$ and $s = 4.6$
c. $\bar{x} = 46$ and $s = 5.0$

20 Grolsch lager, like some of its competitors, can be bought in handy 300 ml bottles. If a bottle such as Grolsch is marked as containing 300 ml, legislation requires that the production batch from which the bottle came must have a mean fill volume of at least 300 ml.

a. Formulate hypotheses that could be used to determine whether the mean fill volume for a production batch satisfies the legal requirement of being at least 300 ml.

b. Suppose you take a random sample of 30 bottles from a lager-bottling production line and find that the mean fill for the sample of 30 bottles is 299.5 ml, with a sample standard deviation of 1.9 ml. What is the *p*-value?

c. At $\alpha = 0.01$, what is your conclusion?

21 Consider a daily TV programme – like the 10 o'clock news – that over the last calendar year had a mean daily audience of 4.0 million viewers. Assume that for a sample of 40 days during the current year, the daily audience was 4.15 million viewers with a sample standard deviation of 0.45 million viewers.

a. If the TV management company would like to test for a change in mean viewing audience, what statistical hypotheses should be set up?

b. What is the *p*-value?

c. Select your own level of significance. What is your conclusion?

22 A popular pastime amongst football fans is participation in 'fantasy football' competitions. Participants choose a squad of players and a manager, with the objective of increasing the valuation of the squad over the season. Suppose that at the start of the competition, the mean valuation of all available strikers was £4.7 million.

a. Formulate the null and alternative hypotheses that could be used by a football pundit to determine whether mid-fielders have a higher mean valuation than strikers.

b. Suppose a random sample of 30 mid-fielders from the available list had a mean valuation at the start of the competition of £5.80 million, with a sample standard deviation of £2.46 million. On average, by how much did the valuation of mid-fielders exceed that of strikers?

c. At $\alpha = 0.05$, what is your conclusion?

23 Most new models of car sold in the European Union have to undergo an official test for fuel consumption. The test is in two parts: an urban cycle and an extra-urban cycle. The urban cycle is carried out under laboratory conditions, over at total distance of 4 km at an average speed of 19 km per hour. Consider a new car model for which the official fuel consumption figure for the urban cycle is published as 11.8 litres of fuel per 100 km. A consumer affairs organization is interested in examining whether this published figure is truly indicative of urban driving.

a. State the hypotheses that would enable the consumer affairs organization to conclude that the model's fuel consumption is more than the published 11.8 litres per 100 km.
b. A sample of 50 mileage tests with the new model of car showed a sample mean of 12.10 litres per 100 km and a sample standard deviation of 0.92 litre per 100 km. What is the p-value?
c. What conclusion should be drawn from the sample results? Use $\alpha = 0.01$.
d. Repeat the preceding hypothesis test using the critical value approach.

24 Joan's Nursery specializes in custom-designed landscaping for residential areas. The estimated labour cost associated with a particular landscaping proposal is based on the number of plantings of trees, shrubs, and so on to be used for the project. For cost-estimating purposes, managers use two hours of labour time for the planting of a medium-sized tree. Actual times from a sample of ten plantings during the past month follow (times in hours).

1.7	1.5	2.6	2.2	2.4	2.3	2.6	3.0	1.4	2.3

With a 0.05 level of significance, test to see whether the mean tree-planting time differs from two hours.

a. State the null and alternative hypotheses.
b. Compute the sample mean.
c. Compute the sample standard deviation.
d. What is the *p*-value?
e. What is your conclusion?

25 Consider the following hypothesis test:

H_0: $\pi = 0.20$

H_1: $\pi \neq 0.20$

A sample of 400 provided a sample proportion p = 0.175.

a. Compute the value of the test statistic.
b. What is the p-value?
c. At α = 0.05, what is your conclusion?
d. What is the rejection rule using the critical value? What is your conclusion?

26 Consider the following hypothesis test:

H_0: $\pi \geq 0.75$

H_1: $\pi < 0.75$

A sample of 300 items was selected. At $\alpha = 0.05$, compute the *p*-value and state your conclusion for each of the following sample results.

a. $p = 0.68$
b. $p = 0.72$
c. $p = 0.70$
d. $p = 0.77$

27 An airline promotion to business travellers is based on the assumption that two-thirds of business travellers use a laptop computer on overnight business trips.

a. State the hypotheses that can be used to test the assumption.

b. What is the sample proportion from an American Express sponsored survey that found 355 of 546 business travellers use a laptop computer on overnight business trips?

c. What is the p-value?

d. Use $\alpha = 0.05$. What is your conclusion?

28 Eagle Outfitters is a chain of stores specializing in outdoor clothing and camping gear. They are considering a promotion that involves sending discount coupons to all their credit card customers by direct mail. This promotion will be considered a success if more than 10 per cent of those receiving the coupons use them. Before going nationwide with the promotion, coupons were sent to a sample of 100 credit card customers.

a. Formulate hypotheses that can be used to test whether the population proportion of those who will use the coupons is sufficient to go national.
b. The file 'Eagle' contains the sample data. Compute a point estimate of the population proportion.
c. Use $\alpha = 0.05$ to conduct your hypothesis test. Should Eagle go national with the promotion?

29 In an IPSOS South Africa opinion poll in May 2012, a sample of adult South Africans were asked their opinions about the performance of the president, Jacob Zuma. One of the response options was the view that the president was performing 'well'.

a. Formulate the hypotheses that can be used to help determine whether more than 50 per cent of the adult population believes the president was performing well.
b. Suppose that, of the 3565 respondents to the poll, 2140 expressed the view that the president was performing well. What is the sample proportion? What is the *p*-value?
c. At $\alpha = 0.01$, what is your conclusion?

30 A study by Consumer Reports showed that 64 per cent of supermarket shoppers believe supermarket brands to be as good as national name brands. To investigate whether this result applies to its own product, the manufacturer of a national name-brand ketchup asked a sample of shoppers whether they believed that supermarket ketchup was as good as the national brand ketchup.

a. Formulate the hypotheses that could be used to determine whether the percentage of supermarket shoppers who believe that the supermarket ketchup was as good as the national brand ketchup differed from 64 per cent.
b. If a sample of 100 shoppers showed 52 stating that the supermarket brand was as good as the national brand, what is the *p*-value?
c. At $\alpha = 0.05$, what is your conclusion?
d. Should the national brand ketchup manufacturer be pleased with this conclusion? Explain.

31 Microsoft Outlook is the most widely used email manager. A Microsoft executive claims that Microsoft Outlook is used by at least 75 per cent of Internet users. A sample of Internet users will be used to test this claim.

a. Formulate the hypotheses that can be used to test the claim.
b. A Merrill Lynch study reported that Microsoft Outlook is used by 72 per cent of Internet users. Assume that the report was based on a sample size of 300 Internet users. What is the *p*-value?
c. At $\alpha = 0.05$, should the executive's claim of at least 75 per cent be rejected?

32 In the elections in Greece in mid-June 2012, the centre-right New Democracy party polled 29.66% of the vote. About a month before the election, a Public Issue opinion poll had estimated the proportion of support for each party. Did New Democracy's support change during the last month of the election campaign?

a. Formulate the null and alternative hypotheses.
b. Suppose the Public Issue opinion poll in May had a random sample of 1200 potential voters, and that 26.0 per cent expressed support for New Democracy. What is the *p*-value?
c. Using $\alpha = 0.05$, what is your conclusion?

33 Consider the following hypothesis test.

$H_0: \mu \geq 10$

$H_0: \mu < 10$

The sample size is 120 and the population standard deviation is assumed known with $\sigma = 5$. Use $\alpha = 0.05$.

a. If the population mean is 9, what is the probability that the sample mean leads to the conclusion do not reject H_0?
b. What type of error would be made if the actual population mean is 9 and we conclude that H_0: $\mu \geq 10$is true?
c. What is the probability of making a Type II error if the actual population mean is 8?

34 Consider the following hypothesis test.

H_0: $\mu = 20$

H_0: $\mu \neq 20$

A sample of 200 items will be taken and the population standard deviation is $\sigma = 10$. Use $\alpha = 0.05$. Compute the probability of making a Type II error if the population mean is:

a. $\mu = 18.0$
b. $\mu = 22.5$
c. $\mu = 21.0$

35 Fowler Marketing Research bases charges to a client on the assumption that telephone survey interviews can be completed within 15 minutes or less. If more time is required, a premium rate is charged. With a sample of 35 interviews, a population standard deviation of four minutes, and a level of significance of 0.01, the sample mean will be used to test the null hypothesis H_0: $\mu \leq 15$.

a. What is your interpretation of the Type II error for this problem? What is its impact on the firm?
b. What is the probability of making a Type II error when the actual mean time is $\mu =$ 17 minutes?
c. What is the probability of making a Type II error when the actual mean time is $\mu =$ 18 minutes?
d. Sketch the general shape of the power curve for this test.

36 Refer to Exercise 35. Assume the firm selects a sample of 50 interviews and repeat parts (b) and (c). What observation can you make about how increasing the sample size affects the probability of making a Type II error?

37 Young Adult magazine states the following hypotheses about the mean age of its subscribers.

H_0: $\mu = 28$

H_0: $\mu \neq 28$

a. What would it mean to make a Type II error in this situation?
b. The population standard deviation is assumed known at $\sigma = 6$ years and the sample size is 100. With $\alpha = 0.05$, what is the probability of accepting H_0 for μ equal to 26, 27, 29 and 30?
c. What is the power at $\mu = 26$? What does this result tell you?

38 Sparr Investments specializes in tax-deferred investment opportunities for its clients. Recently Sparr offered a payroll deduction investment scheme for the employees of a particular company. Sparr estimates that the employees are currently averaging €100 or less per month in tax-deferred investments. A sample of 40 employees will be used to test Sparr's hypothesis about the current level of investment activity among the population of employees. Assume the employee monthly tax-deferred investment amounts have a standard deviation of €75 and that a 0.05 level of significance will be used in the hypothesis test.

a. What would it mean to make a Type II error in this situation?
b. What is the probability of the Type II error if the actual mean employee monthly investment is €120?
c. What is the probability of the Type II error if the actual mean employee monthly investment is €130?
d. Assume a sample size of 80 employees is used and repeat parts (b) and (c).

39 Consider the following hypothesis test.

H_0: $\mu \geq 10$

H_0: $\mu < 10$

The sample size is 120 and the population standard deviation is 5. Use $\alpha = 0.05$. If the actual population mean is 9, the probability of a Type II error is 0.2912. Suppose the researcher wants to reduce the probability of a Type II error to 0.10 when the actual population mean is 9. What sample size is recommended?

40 Consider the following hypothesis test.

H_0: $\mu = 20$

H_0: $\mu \neq 20$

The population standard deviation is 10. Use $\alpha = 0.05$. How large a sample should be taken if the researcher is willing to accept a 0.05 probability of making a Type II error when the actual population mean is 22?

41 A special industrial battery must have a life of at least 400 hours. A hypothesis test is to be conducted with a 0.02 level of significance. If the batteries from a particular production run have an actual mean use life of 385 hours, the production manager wants a sampling procedure that only 10 per cent of the time would show erroneously that the batch is acceptable. What sample size is recommended for the hypothesis test? Use 30 hours as an estimate of the population standard deviation.

42 Young Adult magazine states the following hypotheses about the mean age of its subscribers.

H_0: $\mu = 28$

H_1: $\mu \neq 28$

If the manager conducting the test will permit a 0.15 probability of making a Type II error when the true mean age is 29, what sample size should be selected? Assume $\sigma = 6$ and a 0.05 level of significance.

43 H_0: $\mu = 120$ and H_1: $\mu \neq 120$ are used to test whether a bath soap production process is meeting the standard output of 120 bars per batch. Use a 0.05 level of significance for the test and a planning value of 5 for the standard deviation.

a. If the mean output drops to 117 bars per batch, the firm wants to have a 98 per cent chance of concluding that the standard production output is not being met. How large a sample should be selected?

b. With your sample size from part (a), what is the probability of concluding that the process is operating satisfactorily for each of the following actual mean outputs: 117, 118, 119, 121, 122 and 123 bars per batch? That is, what is the probability of a Type II error in each case?

Chapter 9: Hypothesis Tests

Textbook Exercises Solutions:

1. a. H_0: $\mu \leq 600$ Manager's claim.
   H_1: $\mu > 600$

   b. We are not able to conclude that the manager's claim is wrong.

   c. The manager's claim can be rejected. We can conclude that $\mu > 600$.

2. a. H_0: $\mu \leq 14$
   H_1: $\mu > 14$ Research hypothesis

   b. There is no statistical evidence that the new bonus plan increases sales volume.

   c. The research hypothesis that $\mu > 14$ is supported. We can conclude that the new bonus plan increases the mean sales volume.

3. a. H_0: $\mu = 0.75$ Specified filling weight
   H_1: $\mu \neq 0.75$ Over-filling or under-filling exists

   b. There is no evidence that the production line is operating outside specifications. Allow the production process to continue.

   c. Conclude $\mu \neq 0.75$ and that over-filling or under-filling exists. Shut down and adjust the production line.

4. a. H_0: $\mu \geq 320$
   H_1: $\mu < 320$ Research hypothesis to see if mean cost is less than €320.

   b. We are unable to conclude that the new method reduces costs.

   c. Conclude $\mu < 320$. Consider implementing the new method based on the conclusion that it lowers the mean cost per hour.

5. a. H_0: $\mu \leq 1$ The label claim or assumption.
H_1: $\mu > 1$

b. Claiming $\mu > 1$ when it is not. This is the error of rejecting the product's claim when the claim is true.

c. Concluding $\mu \leq 1$ when it is not. In this case, we miss the fact that the product is not meeting its label specification.

6. a. H_0: $\mu \leq 5000$

H_1: $\mu > 5000$ Research hypothesis that the plan increases average sales.

b. Claiming $\mu > 5000$ when the plan does not increase sales. A mistake could be implementing the plan when it does not help.

c. Concluding $\mu \leq 5000$ when the plan really would increase sales. This could lead to not implementing a plan that would increase sales.

7. a. H_0: $\mu \geq 320$

H_1: $\mu < 320$

b. Claiming $\mu < 320$ when the new method does not lower costs. A mistake could be implementing the method when it does not help.

c. Concluding $\mu \geq 320$ when the method really would lower costs. This could lead to not implementing a method that would lower costs.

8. a. $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{19.4 - 20}{2 / \sqrt{50}} = -2.12$

b. The cumulative probability for $z = -2.12$ is 0.0170.

p-value = 0.0170

c. p-value < 0.05, reject H_0

d. Reject H_0 if $z \leq -1.645$

$-2.12 < -1.645$, so reject H_0

9. a. $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{14.15 - 15}{3/\sqrt{50}} = -2.00$

b. Cumulative probability = 0.0228

p-value = 2(0.0228) = 0.0456

c. p-value < 0.05, reject H_0

d. Reject H_0 if $z \le -1.96$ or $z \ge 1.96$

$-2.00 < -1.96$, reject H_0

10. Reject H_0 if $z \ge 1.645$

a. $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{52.5 - 50}{8/\sqrt{60}} = 2.42$

$2.42 > 1.645$, reject H_0

b. $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{51 - 50}{8/\sqrt{60}} = 0.97$

$0.97 < 1.645$, do not reject H_0

c. $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{51.8 - 50}{8/\sqrt{60}} = 1.74$

$1.74 > 1.645$, reject H_0

11. a. H_0: $\mu = 39.2$

H_1: $\mu \neq 39.2$

b. $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{38.5 - 39.2}{4.8/\sqrt{112}} = -1.54$

p-value = 2(0.0618) = 0.1236

c. p-value > 0.05, do not reject H_0. We cannot conclude that the mean length of a working week has changed.

d. Reject H_0 if $z \le -1.96$ or $z \ge 1.96$

$z = -1.54$; cannot reject H_0

12. a. H_0: $\mu \geq 181{,}900$

H_1: $\mu < 181{,}900$

b. $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{166{,}400 - 181{,}900}{33{,}500 / \sqrt{40}} = -2.93$

c. p-value = 0.0017

d. p-value < 0.01. Reject H_0. Conclude that the mean for the North-East is less than the national mean.

13. a. H_0: $\mu \leq 15$

H_1: $\mu > 15$

b. $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \dfrac{17 - 15}{4 / \sqrt{35}} = 2.96$

c. p-value = 1 – 0.9985 = 0.0015

d. p-value < 0.01; reject H_0; the premium rate should be charged.

14. a. H_0: $\mu = 8$

H_1: $\mu \neq 8$

b. $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{8.4 - 8}{3.2 / \sqrt{120}} = 1.37$

Cumulative probability for z = 1.37 is 0.9147

p-value = 2(1 – 0.9147) = 0.1706

c. p-value > α = 0.05. Do not reject H_0. Cannot conclude that the population mean waiting time differs from 8 minutes.

15. a. H_0: $\mu = 5.2$

H_1: $\mu \neq 5.2$

b. 95% confidence interval is $\bar{x} \pm z\frac{\sigma}{\sqrt{n}} = 4.0 \pm 1.96\frac{2.9}{\sqrt{1050}} = 4.0 \pm 0.18$, or 3.82 to 4.18.

c. The 95% confidence interval does not include $\mu_0 = 5.2$. Therefore reject H_0 at $\alpha =$ 0.05, and conclude that the population mean for Germany was lower than the worldwide mean.

16. a. H_0: $\mu = 500$

H_1: $\mu \neq 500$

b. $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{510 - 500}{25 / \sqrt{30}} = 2.19$

p-value = 2(1 − 0.9857) = 0.0286.

p-value < 0.05, reject H_0. Over-filling occurring, recommend re-adjustment of the mechanism.

c. $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{495 - 500}{25 / \sqrt{30}} = -1.10$

p-value = 2(0.1357) = 0.2714.

p-value > 0.05, do not reject H_0. No action required.

d. Critical values are:

$500 \pm 1.96\left(\frac{25}{\sqrt{30}}\right) = 500 \pm 8.95$ i.e. 491.05 and 508.95

Reject H_0 if $\bar{x}$ ³ 508.95 or if $\bar{x}$ £ 491.05. Same conclusions as in b and c.

17. a. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{14 - 12}{4.32/\sqrt{25}} = 2.31$

b. Degrees of freedom = $n - 1 = 24$

Using t table, p-value is between 0.01 and 0.025. (Actual p-value = 0.0147.)

c. p-value < 0.05, reject H_0.

d. With $df = 24$, $t_{0.05} = 1.711$

Reject H_0 if $t \geq 1.711$

$2.31 > 1.711$, reject H_0.

18. a. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{17 - 18}{4.5/\sqrt{48}} = -1.54$

b. Degrees of freedom = $n - 1 = 47$

Area in lower tail is between 0.05 and 0.10

Using t table, p-value (two-tailed) is between 0.10 and 0.20. (Actual p-value = 0.1304.)

c. p-value > 0.05, do not reject H_0.

d. With $df = 47$, $t_{0.025} = 2.012$

Reject H_0 if $t \leq -2.012$ or $t \geq 2.012$

$t = -1.54$; do not reject H_0

19. a. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{44 - 45}{5.2/\sqrt{36}} = -1.15$

Degrees of freedom = $n - 1 = 35$

Using t table, p-value is between 0.10 and 0.20. (Actual p-value = 0.1282.)

p-value > 0.01, do not reject H_0

b. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{43 - 45}{4.6/\sqrt{36}} = -2.61$

Using t table, p-value is between 0.005 and 0.01. (Actual p-value = 0.0066.)

p-value < 0.01, reject H_0

c. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{46 - 45}{5/\sqrt{36}} = 1.20$

Using t table, area in upper tail is between 0.10 and 0.20

p-value (lower tail) is between 0.80 and 0.90. (Actual p-value = 0.8809.)

p-value > 0.01, do not reject H_0

20. a. H_0: $\mu \geq 300$

H_1: $\mu < 300$

b. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{299.5 - 300}{1.9/\sqrt{30}} = -1.44$

Degrees of freedom = $n - 1 = 29$

Using t table, p-value is between 0.05 and 0.10. (Actual p-value = 0.080.)

c. p-value > 0.10, do not reject H_0

21. a. H_0: $\mu = 4.0$ (million)

H_1: $\mu \neq 4.0$ (million)

b. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{4.15 - 4.0}{0.45/\sqrt{40}} = 2.11$

$df = n - 1 = 39$

Using t table, area in tail is between 0.01 and 0.025

p-value is between 0.02 and 0.05. (Actual p-value = 0.041.)

c. At $\alpha = 0.05$, H_0 is rejected and we conclude that the mean viewing audience has increased. At $\alpha = 0.01$, the evidence is not sufficiently strong to reach this conclusion.

22. a. H_0: $\mu \leq 4.7$ (£ million)

H_1: $\mu > 4.7$ (£ million)

b. $5.8 - 4.7 = 1.1$ (£ million)

c. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{1.1}{2.46/\sqrt{30}} = 2.45$

$df = n - 1 = 29$

Using t table, p-value is between 0.01 and 0.025 (actual p-value = 0.021).

Reject H_0 at $\alpha = 0.05$, conclude that the mean valuation of mid-fielders is higher than that of strikers.

23. a. H_0: $m£\ 11.8$

H_1: $\mu > 11.8$

b. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{12.10 - 11.8}{0.92/\sqrt{50}} = 2.31$

$df = n - 1 = 49$

Using t table, p-value is between 0.01 and 0.025 (actual p-value = 0.013).

c. p-value > $\alpha = 0.01$. Insufficient evidence to reject H_0 at $\alpha = 0.01$. The consumer affairs organization cannot conclude that the fuel consumption is greater than the published figure.

d. With $df = 49$, $t_{0.025} = 2.405$

Reject H_0 if $t \geq 2.405$

$t = 2.31$; do not reject H_0.

24. a. H_0: $\mu = 2$

H_1: $\mu \neq 2$

b. $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{22}{10} = 2.2$

c. $s = \sqrt{\dfrac{\Sigma (x_i - \bar{x})^2}{n-1}} = 0.516$

d. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{2.2 - 2}{0.516/\sqrt{10}} = 1.22$

Degrees of freedom = $n - 1 = 9$

Using t table, area in tail is between 0.10 and 0.20

p-value is between 0.20 and 0.40. (Actual p-value = 0.2518.)

e. p-value > 0.05; do not reject H_0. No reason to change from the 2 hours for cost-estimating purposes.

25. a. $$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.175 - 0.20}{\sqrt{\frac{0.20(1-0.20)}{400}}} = -1.25$$

b. p-value = 2(0.1056) = 0.2112

c. p-value > 0.05; do not reject H_0

d. $z_{0.025} = 1.96$

Reject H_0 if $z \le -1.96$ or $z \ge 1.96$

$z = -1.25$; do not reject H_0

26. a. $$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.68 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{300}}} = -2.80$$

p-value = 0.0026

p-value ≤ 0.05; reject H_0

b. $$z = \frac{0.72 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{300}}} = -1.20$$

p-value = 0.1151

p-value > 0.05; do not reject H_0

c. $$z = \frac{0.70 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{300}}} = -2.00$$

p-value = 0.0228

p-value < 0.05; reject H_0

d. $$z = \frac{0.77 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{300}}} = 0.80$$

p-value = 0.7881

p-value > 0.05; do not reject H_0

27. a. H_0: $\pi = 0.6667$

H_1: $\pi \neq 0.6667$

b. $p = \dfrac{355}{546} = 0.6502$

c. $z = \dfrac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \dfrac{0.6502 - 0.6667}{\sqrt{\dfrac{0.6667(1-0.6667)}{546}}} = -0.82$

p-value = 2(0.2061) = 0.4122

d. p-value > 0.05; do not reject H_0. Cannot conclude that the population proportion differs from 2/3.

28. a. $H_0 : p \leq 0.10, \quad H_1 : p > 0.10,$ where π is the population proportion of customers who will use the coupons

b. $p = 0.13$

c. $z = \dfrac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \dfrac{0.13 - 0.10}{\sqrt{\dfrac{0.10(1-0.10)}{100}}} = 1.0$

p-value = 1 – 0.8413 = 0.16, do not reject H_0. Eagle should not go national on this evidence.

29. a. H_0: $\pi \leq 0.50$

H_1: $\pi > 0.50$

b. $p = \dfrac{2140}{3565} = 0.6002 \quad (60\%)$

$z = \dfrac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \dfrac{0.6002 - 0.50}{\sqrt{\dfrac{0.50(1-0.50)}{3565}}} = 12.0$

p-value < 0.001.

c. p-value < 0.01, reject H_0. Conclude that more than 50% of population thinks the president was performing well.

30. a. H_0: $\pi = 0.64$
H_1: $\pi \neq 0.64$

b. $p = \dfrac{52}{100} = 0.52$

$$z = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{0.52 - 0.64}{\sqrt{\dfrac{0.64(1-0.64)}{100}}} = -2.50$$

Cumulative probability for z = –2.50 is 0.0062

p-value = 2(0.0062) = 0.0124

c. p-value < 0.05; reject H_0. Proportion differs from the reported 0.64.

d. Yes. Since p = 0.52, it indicates that fewer than 64% of the shoppers believe the supermarket brand is as good as the name brand.

31. a. H_0: $\pi \geq 0.75$
H_1: $\pi < 0.75$

b. $$z = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{0.72 - 0.75}{\sqrt{\dfrac{0.75(1-0.75)}{300}}} = -1.20$$

p-value = 0.1151

c. p-value > 0.05; do not reject H_0. The executive's claim cannot be rejected.

32. a. H_0: $\pi = 0.2966$

H_1: $\pi \neq 0.2966$

b. $$z = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{0.26 - 0.2966}{\sqrt{\dfrac{0.2966(1-0.2966)}{1200}}} = -2.78$$

c. p-value = 0.0054, reject H_0. Conclude that support had changed.

33. $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{5}{\sqrt{120}} = 0.46$

$c = 10 - 1.645\,(5/\sqrt{120}) = 9.25$

Reject H_0 if $\bar{x} \le 9.25$

a. When $\mu = 9$,

$z = \dfrac{9.25-9}{5/\sqrt{120}} = 0.55$

Prob of not rejecting $H_0 = 1 - 0.7088 = 0.2912$

b. Type II error

c. When $\mu = 8$,

$z = \dfrac{9.25-8}{5/\sqrt{120}} = 2.74$

$\beta = (1 - 0.9969) = 0.0031$

34. Reject H_0 if $z \le -1.96$ or if $z \ge 1.96$

$\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{10}{\sqrt{200}} = 0.71$

$c_1 = 20 - 1.96\,(10/\sqrt{200}) = 18.61$

$c_2 = 20 + 1.96\,(10/\sqrt{200}) = 21.39$

a. $\mu = 18$

$z = \dfrac{18.61-18}{10/\sqrt{200}} = 0.86$

$\beta = P(Z > 0.86) = 1 - 0.8051 = 0.1949$

b. $\mu = 22.5$

$z = \dfrac{21.39-22.5}{10/\sqrt{200}} = -1.57$

$\beta = P(Z < -1.57) = 0.0582$

c. $\mu = 21$

$$z = \frac{21.39 - 21}{10/\sqrt{200}} = 0.55$$

$\beta = P(Z < 0.55) = 0.7088$

35. a. H_0: $\mu \leq 15$

H_1: $\mu > 15$

Concluding $\mu \leq 15$ when this is not true. Fowler would not charge the premium rate even though the rate should be charged.

b. Reject H_0 if $z \geq 2.33$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 15}{4/\sqrt{35}} = 2.33$$

Solve for $\bar{x} = 16.58$

Decision Rule:

Accept H_0 if $\bar{x} < 16.58$

Reject H_0 if $\bar{x} \geq 16.58$

For $\mu = 17$,

$$z = \frac{16.58 - 17}{4/\sqrt{35}} = -0.62$$

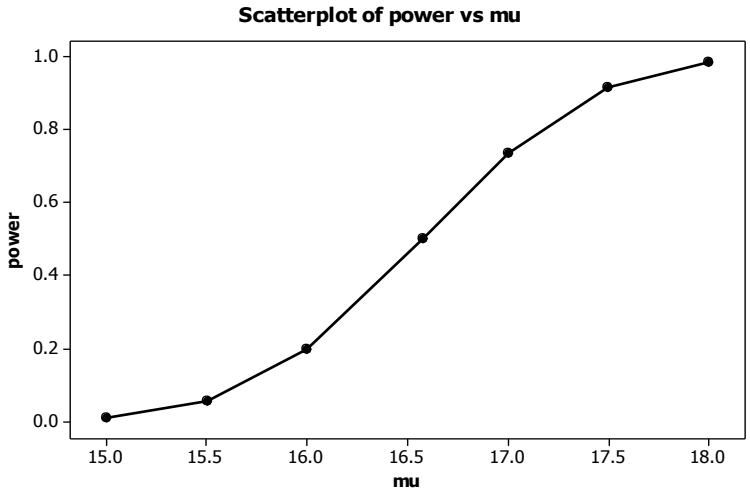
$\beta = 0.2676$

c. For $\mu = 18$,

$$z = \frac{16.58 - 18}{4/\sqrt{35}} = -2.10$$

$\beta = 0.0179$

d.



36. $c = \mu_0 + z_{0.01}\dfrac{\sigma}{\sqrt{n}} = 15 + 2.33\dfrac{4}{\sqrt{50}} = 16.32$

At $\mu = 17$ $\qquad z = \dfrac{16.32 - 17}{4/\sqrt{50}} = -1.20$

$\beta = 0.1151$

At $\mu = 18$ $\qquad z = \dfrac{16.32 - 18}{4/\sqrt{50}} = -2.97$

$\beta = 0.0015$

Increasing the sample size reduces the probability of making a Type II error.

37. a. Accepting H_0 and concluding the mean average age was 28 years when it was not.

b. Reject H_0 if $z \le -1.96$ or if $z \ge 1.96$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 28}{6/\sqrt{100}}$$

Solving for $\bar{x}$, we find

at $z = -1.96$, $\qquad \bar{x} = 26.82$

at $z = +1.96$, $\qquad \bar{x} = 29.18$

Decision Rule:

Accept H_0 if $26.82 < \bar{x} < 29.18$

Reject H_0 if $\bar{x} \leq 26.82$ or if $\bar{x} \geq 29.18$

At $\mu = 26$,

$$z = \frac{26.82 - 26}{6/\sqrt{100}} = 1.37$$

$\beta = 1 - 0.9147 = 0.0853$

At $\mu = 27$,

$$z = \frac{26.82 - 27}{6/\sqrt{100}} = -0.30$$

$\beta = 1 - 0.3821 = 0.6179$

At $\mu = 29$,

$$z = \frac{29.18 - 29}{6/\sqrt{100}} = 0.30$$

$\beta = 0.6179$

At $\mu = 30$,

$$z = \frac{29.18 - 30}{6/\sqrt{100}} = -1.37$$

$\beta = 0.0853$

c. Power = $1 - \beta$

at $\mu = 26$, Power = $1 - 0.0853 = 0.9147$

When $\mu = 26$, there is a 0.9147 probability that the test will correctly reject the null hypothesis that $\mu = 28$.

38. a. A Type II error would be 'accepting' that the mean level of tax-deferred investments is no greater than €100, when in fact it is greater than €100.

b. H_0: $\mu \leq 100$

H_1: $\mu > 100$

Decision Rule:

Accept H_0 $\bar{x} < 100 + 2.326(75/\sqrt{40}) = 127.58$

Reject H_0 if $\bar{x} \geq 127.58$

At $\mu = 120$,

$$z = \frac{127.58 - 120}{75/\sqrt{40}} = 0.64$$

$\beta = 0.7389$

c. At $\mu = 130$,

$$z = \frac{127.58 - 130}{75/\sqrt{40}} = -0.20$$

$\beta = 0.4207$

d. Decision Rule:

Accept H_0 $\bar{x} < 100 + 2.326(75/\sqrt{80}) = 119.50$

Reject H_0 if $\bar{x} \geq 119.50$

At $\mu = 120$,

$$z = \frac{119.50 - 120}{75/\sqrt{80}} = -0.06$$

$\beta = 0.4761$

e. At $\mu = 130$,

$$z = \frac{119.50 - 130}{75/\sqrt{80}} = -1.25$$

$\beta = 0.10567$

39. $$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_a)^2} = \frac{(1.645 + 1.28)^2 (5)^2}{(10 - 9)^2} = 214$$

40. $$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_a)^2} = \frac{(1.96 + 1.645)^2 (10)^2}{(20 - 22)^2} = 325$$

41. At $\mu_0 = 400$, $\alpha = 0.02$. $z_{0.02} = 2.05$

At $\mu_1 = 385$, $\beta = 0.10$. $z_{0.10} = 1.28$

$\sigma = 30$

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(2.05 + 1.28)^2 (30)^2}{(400 - 385)^2} = 44.4 \quad \text{Use } n = 45$$

42. At $\mu_0 = 28$, $\alpha = 0.05$. Note, however, for this two-tailed test, $z_{\alpha/2} = z_{0.025} = 1.96$

At $\mu_1 = 29$, $\beta = 0.15$. $z_{0.15} = 1.04$

$\sigma = 6$

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(1.96 + 1.04)^2 (6)^2}{(28 - 29)^2} = 324$$

43. a. H_0: $\mu = 120$
H_1: $\mu \neq 120$

At $\mu_0 = 120$, $\alpha = 0.05$. With a two-tailed test, $z_{\alpha/2} = z_{0.025} = 1.96$

At $\mu_1 = 117$, $\beta = 0.02$. $z_{0.02} = 2.05$

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(1.96 + 2.05)^2 (5)^2}{(120 - 117)^2} = 44.7. \quad \text{Use } n = 45$$

b. Example calculation for $\mu = 118$

Reject H_0 if $z \leq -1.96$ or if $z \geq 1.96$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{\bar{x} - 120}{5 / \sqrt{45}}$$

Solve for $\bar{x}$ At $z = -1.96$, $\bar{x} = 118.54$

At $z = +1.96$, $\bar{x} = 121.46$

Decision Rule:

Accept H_0 if $118.54 < \bar{x} < 121.46$

Reject H_0 if $\bar{x} \leq 118.54$ or if $\bar{x} \geq 121.46$

For $\mu = 118$,

$$z = \frac{118.54 - 118}{5 / \sqrt{45}} = 0.72$$

$\beta = 1 - 0.2642 = 0.2358$

Other results:

If μ is	z	β
117	2.07	0.0192
118	0.72	0.2358
119	–0.62	0.7291
121	+0.62	0.7291
122	+0.72	0.2358
123	-2.07	0.0192

Supplementary Exercises:

44. Suppose a popular newspaper claims that people spend on average no more than 10 minutes per day reading newspapers. A researcher believes that individuals in management positions spend more time than this reading newspapers. A sample of individuals in management positions will be selected by the researcher. Data on newspaper-reading times will be used to test the following null and alternative hypotheses.

$$H_0: \mu \leq 10$$
$$H_1: \mu > 10$$

a. What is the Type I error in this situation? What are the consequences of making this error?

b. What is the Type II error in this situation? What are the consequences of making this error?

45. Consider the following hypothesis test:

$$H_0: \mu \leq 25$$
$$H_1: \mu > 25$$

A sample of 40 provided a sample mean of 26.4. The population standard deviation is 6.

a. Compute the value of the test statistic.

b. What is the *p*-value?

c. At $\alpha = 0.01$, what is your conclusion?

d. What is the rejection rule using the critical value approach? What is your conclusion?

46. Consider the following hypothesis test:

$$H_0: \mu \geq 80$$
$$H_1: \mu < 80$$

A sample of 100 is used and the population standard deviation is 12. Compute the *p*-value and, using $\alpha = 0.01$, state your conclusion for each of the following sample results.

a. $\bar{x} = 78.5$ b. $\bar{x} = 77$ c. $\bar{x} = 75.5$ d. $\bar{x} = 81$

47. Consider the following hypothesis test:

$$H_0: \mu = 100$$

$$H_1: \mu \neq 100$$

A sample of 65 is used. Calculate the *p*-value and, using $\alpha = 0.05$, state your conclusion for each of the following sample results.

a. $\bar{x} = 103$ and $s = 11.5$

b. $\bar{x} = 96.5$ and $s = 11.0$

c. $\bar{x} = 102$ and $s = 10.5$

48. A new forged titanium golf driver was described as "illegal" because it promised driving distances that exceed the U.S. Golf Association's standard. Suppose a USGA-approved driver has a population mean driving distance of 260 metres. Based on nine test drives, the mean driving distance by the new titanium driver was 266.9 metres. Answer the following questions if the sample standard deviation driving distance was 10 metres.

a. Formulate the null and alternative hypotheses that can be used to determine whether the new titanium driver has a population mean driving distance greater than 260 metres.

b. On average, how many metres further did the golf ball travel with the titanium driver?

c. At $\alpha = 0.05$, what is your conclusion?

49. AOL Time Warner Inc.'s CNN has been the long-time ratings leader of cable television news in the US. Nielsen Media Research indicated that the mean CNN viewing audience was 600 000 viewers per day during 2002 (The Wall Street Journal, 10 March 2003). Assume that for a sample of 40 days during the first half of 2003, the daily audience was 612 000 viewers with a sample standard deviation of 65 000 viewers.

a. What are the hypotheses if CNN management would like information on any change in the CNN viewing audience?

b. What is the *p*-value?

c. Select your own level of significance. What is your conclusion?

d. What recommendation would you make to CNN management in this application?

50. The population mean earnings per share for financial services corporations including American Express, E*TRADE Group, Goldman Sachs, and Merrill Lynch was $3 (*Business Week*, August 14, 2000). In 2001, a sample of 10 financial services corporations provided the following earnings per share data:

1.92 2.16 3.63 3.16 4.02 3.14 2.20 2.34 3.05 2.38

a. Formulate the null and alternative hypotheses that can be used to determine whether the population mean earnings per share in 2001 differ from the $3 reported in 2000.
b. Compute the sample mean.
c. Compute the sample standard deviation.
d. What is the *p*-value?
e. Use $\alpha = 0.05$. What is your conclusion?

51. The chamber of commerce of a developing city advertises that small retail units are available in the city at a mean annual rental of €12,500 or less per unit. Suppose a sample of 32 units provided a sample mean of €13,000 per unit and a sample standard deviation of €1250. Using a 0.05 level of significance, test the validity of the advertising claim.

52. A consumer research group is interested in testing a car manufacturer's claim that a new economy model will travel at least 25 kilometres per litre of diesel fuel (H_0: $\mu \geq 25$).
a. With a 0.02 level of significance and a sample of 30 cars, what is the rejection rule, based on the value of $\bar{X}$, for the test to determine whether the manufacturer's claim should be rejected? Assume that σ is 3 kilometres per litre.
b. What is the probability of committing a Type II error if the actual mean performance is 23 kilometres per litre?
c. What is the probability of committing a Type II error if the actual mean performance is 24 kilometres per litre?
d. What is the probability of committing a Type II error if the actual mean performance is 25.5 kilometres per litre?

53. A production line operation is tested for filling-weight accuracy using the following hypotheses.

H_0: $\mu = 150$ (filling okay; keep production running)
H_1: $\mu \neq 150$ (filling off-standard; stop and adjust machine)

The sample size is 30 and the population standard deviation is $\sigma = 8$. Use $\alpha = 0.05$.
a. What would a Type II error mean in this situation?
b. What is the probability of making a Type II error when the machine is over-filling by 5 grams?
c. What is the power of the statistical test when the machine is over-filling by 5 grams?
d. Show the power curve for this hypothesis test. What information does it contain for the production manager?

54. A petrol economy study on a particular model of car tested the following hypotheses.

H_0: $\mu \geq$ 15 kilometres per litre (manufacturer's claim supported)

H_1: $\mu <$ 15 kilometres per litre (manufacturer's claim rejected; average kilometres per litre less than stated)

For σ = 3 and a 0.02 level of significance, what sample size would be recommended if the researcher wants an 80% chance of detecting that μ is less than 15 kilometres per litre when it is actually 14 kilometres per litre?

55. A funding programme is available to low-income neighbourhoods. To qualify for the funding, a neighbourhood must have a mean household income of less than €15,000 per year. Neighbourhoods with mean annual household income of €15,000 or more do not qualify. Funding decisions are based on a sample of residents in the neighbourhood. A hypothesis test with a 0.02 level of significance is conducted. If the funding guidelines call for a maximum probability of 0.05 of not funding a neighbourhood with a mean annual household income of €14,000, what sample size should be used in the funding decision study? Use σ = €4000 as a planning value.

56. Suppose that a newspaper article about driving practices claims that, in one particular area, 48% of drivers do not stop at stop signs on urban roads. Two months later, a follow-up study collected data in order to see whether this percentage had changed.
    a. Formulate the hypotheses to determine whether the proportion of drivers who did not stop at stop signs had changed.
    b. Assume the study found 360 of 800 drivers did not stop at stop signs. What is the sample proportion? What is the *p*-value?
    c. At α = 0.05, what is your conclusion?

57. Before the Iraqi election in January 2005, an Abu Dhabi TV/Zogby International poll asked a sample of Iraqi adults whether they would prefer an Islamic or a secular government.
    a. Formulate the hypotheses that can be used to help determine whether more than 50 per cent of the adult population would prefer a secular government.
    b. Suppose that, of 805 respondents to the poll, 475 expressed a preference for a secular government. What is the sample proportion? What is the p-value?
    c. At α = 0.01, what is your conclusion?

58. Suppose detailed records show that, in one particular year, 78% of trains run by a particular company arrive on time. The following year, a study finds that 330 of 400 randomly selected trains arrived on time. Does the sample indicate the on-time arrival rate has changed? Test at $\alpha = 0.05$.
    a. What is the sample proportion of trains arriving on time?
    b. What is the *p*-value?
    c. What is your conclusion?

59. Environmental health indicators include air quality, water quality, and food quality. Suppose that 25 years ago, 47% of food samples contained pesticide residues (i.e. take this as a population value). In a recent study, 44 of 125 food samples contained pesticide residues.
    a. State the hypotheses that can be used to test whether the population proportion has declined.
    b. What is the sample proportion?
    c. What is the *p*-value?
    d. Use $\alpha = 0.01$. What is your conclusion?

60. The Heldrich Centre for Workforce Development found in 2000 that 40% of Internet users received more than 10 email messages per day. A similar study on the use of email was repeated in 2002.
    a. Formulate the hypotheses that can be used to determine whether the proportion of Internet users receiving more than 10 email messages per day increased.
    b. If a sample of 425 Internet users found 189 receiving more than 10 email messages per day, what is the *p*-value?
    c. At $\alpha = 0.05$, what is your conclusion?

61. In the Kenyan presidential election in December 2002, Mwai Kibaki, representing Narc, was elected with 63 per cent of the vote ($\pi = 0.63$). A month before the election, an opinion poll had estimated the proportion of support for each candidate. Did Kibaki's support change during the last month of the election campaign?
    a. Formulate the null and alternative hypotheses.
    b. Suppose the November opinion poll had a random sample of 3000 potential voters, and that 68.2 per cent expressed support for Kibaki. What is the *p*-value? Use $\alpha = 0.05$.
    c. What is your conclusion?

Supplementary Exercises Solutions:

44. a. The Type I error is rejecting H_0 when it is true. In this case, this error occurs if the researcher concludes that the mean newspaper-reading time for individuals in management positions is greater than the national average of 8.6 minutes when in fact it is not.

    b. The Type II error is accepting H_0 when it is false. In this case, this error occurs if the researcher concludes that the mean newspaper-reading time for individuals in management positions is less than or equal to the national average of 8.6 minutes when in fact it is greater than 8.6 minutes.

45. a. $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{26.4 - 25}{6 / \sqrt{40}} = 1.48$

    b. p-value = 1 – 0.9306 = 0.0694

    c. p-value > 0.01, do not reject H_0

    d. Reject H_0 if $z \geq 2.33$

    1.48 < 2.33, do not reject H_0

46. a. $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{78.5 - 80}{12 / \sqrt{100}} = -1.25$

    p-value = 0.1056

    p-value > 0.01, do not reject H_0

    b. $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{77 - 80}{12 / \sqrt{100}} = -2.50$

    p-value = = 0.0062

    p-value < 0.01, reject H_0

c. $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{75.5 - 80}{12 / \sqrt{100}} = -3.75$

p-value very close to 0

p-value < 0.01, reject H_0

d. $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{81 - 80}{12 / \sqrt{100}} = 0.83$

p-value = 0.7967

p-value > 0.01, do not reject H_0

47. a. $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} = \dfrac{103 - 100}{11.5 / \sqrt{65}} = 2.10$

Degrees of freedom = $n - 1 = 64$

Using *t* table, area in tail is between 0.01 and 0.025

p-value (two-tailed) is between 0.02 and 0.05

(Actual *p*-value = 0.0394)

p-value < 0.05, reject H_0

b. $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} = \dfrac{96.5 - 100}{11 / \sqrt{65}} = -2.57$

Using *t* table, area in tail is between 0.005 and 0.01

p-value (two-tailed) is between 0.01 and 0.02

(Actual *p*-value = 0.0127)

p-value < 0.05, reject H_0

c. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{102 - 100}{10.5/\sqrt{65}} = 1.54$

Using *t* table, area in tail is between 0.05 and 0.10

p-value (two-tailed) is between 0.10 and 0.20

(Actual *p*-value = 0.1295)

p-value > 0.05, do not reject H_0

48. a. H_0: $\mu \leq 260$

H_1: $\mu > 260$

b. 266.9 – 260 = 6.9 metres

c. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{266.9 - 260}{10/\sqrt{9}} = 2.07$

Degrees of freedom = $n - 1 = 8$

Using *t* table, *p*-value is between 0.025 and 0.05

(Actual *p*-value = 0.0361)

p-value < 0.05; reject H_0. Conclude that the population mean distance for the new driver is greater than the USGA-approved driver.

49. a. H_0: $\mu = 600$

H_1: $\mu \neq 600$

b. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{612 - 600}{65/\sqrt{40}} = 1.17$

$df = n - 1 = 39$

Using *t* table, area in tail is between 0.10 and 0.20

p-value is between 0.20 and 0.40. (Actual *p*-value = 0.2501.)

c. With $\alpha = 0.10$ or less, we cannot reject H_0. We are unable to conclude there has been a change in the mean CNN viewing audience.

d. The sample mean of 612 thousand viewers is encouraging but not conclusive for the sample of 40 days. Recommend additional viewer audience data. A larger sample should help clarify the situation for CNN.

50. a. H_0: $\mu = 3$
H_1: $\mu \neq 3$

b. $\bar{x} = \dfrac{\Sigma x_i}{n} = 2.8$

c. $s = \sqrt{\dfrac{\Sigma (x_i - \bar{x})^2}{n-1}} = 0.70$

d. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{2.8 - 3}{0.70/\sqrt{10}} = -0.90$

Degrees of freedom = 10 – 1 = 9

Using *t* table, area in tail is between 0.10 and 0.20

p-value is between 0.20 and 0.40

(Actual *p*-value = 0.3902)

e. *p*-value > 0.05; do not reject H_0. There is no evidence to conclude a difference compared to previous year.

51. H_0: $\mu \leq 12{,}500$
H_1: $\mu > 12{,}500$

$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{13{,}000 - 12{,}500}{1250/\sqrt{32}} = 2.26$

Degrees of freedom = 32 – 1 = 31

Using *t* table, *p*-value is between 0.01 and 0.025

(Actual *p*-value = 0.0154)
p-value < 0.05; reject H_0. Conclude that the mean cost is greater than €12,500 per unit.

52. a. H_0: $\mu \geq 25$

H_1: $\mu < 25$

Reject H_0 if $z \leq -2.05$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{\bar{x} - 25}{3/\sqrt{30}} = -2.05$$

Solve for $\bar{x} = 23.88$

Decision Rule:

Accept H_0 if $\bar{x} > 23.88$

Reject H_0 if $\bar{x} \leq 23.88$

b. For $\mu = 23$, $z = \frac{23.88 - 23}{3/\sqrt{30}} = 1.61$

$\beta = 1 - 0.9463 = 0.0537$

c. For $\mu = 24$, $z = \frac{23.88 - 24}{3/\sqrt{30}} = -0.22$

$\beta = 1 - 0.4129 = 0.5871$

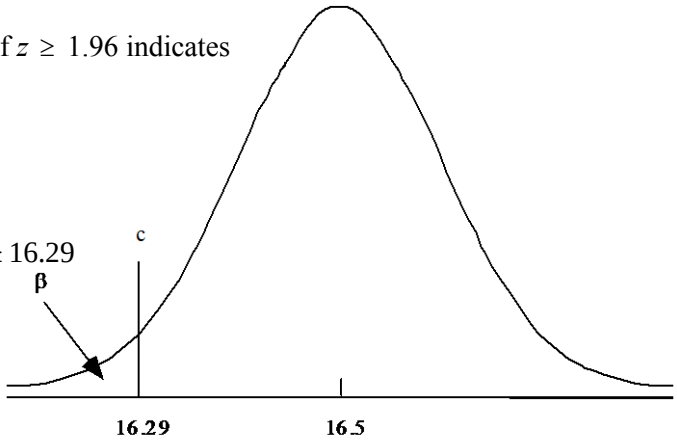
d. The Type II error cannot be made in this case. Note that when $\mu = 25.5$, H_0 is true. The Type II error can only be made when H_0 is false.

53. a. Accepting H_0 and letting the process continue to run when actually over-filling or under-filling exists.

b. Decision Rule: Reject H_0 if $z \leq -1.96$ or if $z \geq 1.96$ indicates

Accept H_0 if $15.71 < \bar{x} < 16.29$

Reject H_0 if $\bar{x} \leq 15.71$ or if $\bar{x} \geq 16.29$



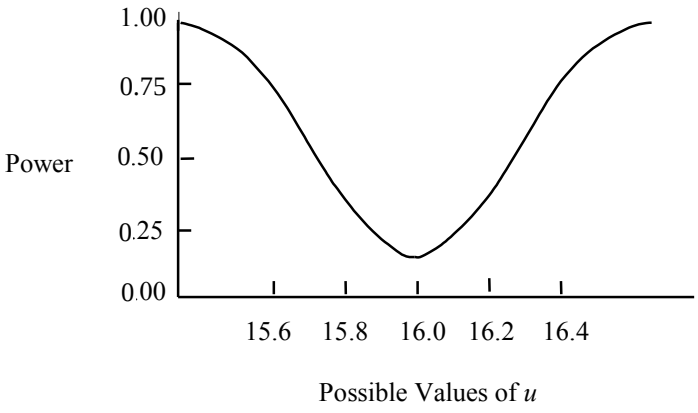
For $\mu = 16.5$

$$z = \frac{16.29 - 16.5}{0.8/\sqrt{30}} = -1.44$$

$\beta = 0.0749$

c. Power = 1 – 0.0749 = 0.9251

d. The power curve shows the probability of rejecting H_0 for various possible values of μ. In particular, it shows the probability of stopping and adjusting the machine under a variety of under-filling and over-filling situations. The general shape of the power curve for this case is



54. At $\mu_0 = 25$, $\alpha = 0.02$, $z_{0.02} = 2.05$

At $\mu_1 = 24$, $\beta = 0.20$, $z_{0.20} = 0.84$

$\sigma = 3$

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(2.05 + .84)^2 (3)^2}{(25 - 24)^2} = 75.2 \qquad \text{Use } n = 76$$

55. H_0: $\mu \geq 15{,}000$

H_1: $\mu < 15{,}000$

At $\mu_0 = 15{,}000$, $\alpha = 0.02$, $z_{0.02} = 2.05$

At $\mu_1 = 14{,}000$, $\beta = 0.05$. $z_{0.10} = 1.645$

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(2.05 + 1.645)^2 (4{,}000)^2}{(15{,}000 - 14{,}000)^2} = 218.5 \qquad \text{Use } n = 219$$

56. a. H_0: $\pi = 0.48$

H_1: $\pi \neq 0.48$

b. $$p = \frac{360}{800} = 0.45$$

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.45 - 0.48}{\sqrt{\frac{0.48(1 - 0.48)}{800}}} = -1.70$$

p-value = 2(0.0446) = 0.0892

c. p-value > 0.05; do not reject H_0. There is no reason to conclude the proportion has changed.

57. a. H_0: $\pi \leq 0.50$

H_1: $\pi > 0.50$

b. $p = \dfrac{2140}{3565} = 0.6002 \quad (60\%)$

$$z = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{0.6002 - 0.50}{\sqrt{\dfrac{0.50(1-0.50)}{3565}}} = 12.0$$

p-value < 0.001.

c. p-value < 0.01, reject H_0. Conclude that more than 50% of population thinks the president was performing well.

58. a. $p = \dfrac{330}{400} = 0.825$

b. H_0: $\pi = 0.78$
H_1: $\pi \neq 0.78$

$$z = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{0.825 - 0.78}{\sqrt{\dfrac{0.78(1-0.78)}{400}}} = 2.17$$

p-value = 2(1 – 0.9850) = 0.03

c. p-value < 0.05; reject H_0. Conclude that the on-time arrival record has improved.

59. a. H_0: $\pi \geq 0.47$
H_1: $\pi < 0.47$

b. $p = \dfrac{44}{125} = 0.352$

c. $$z = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{0.352 - 0.47}{\sqrt{\dfrac{0.47(1-0.47)}{125}}} = -2.64$$

p-value = 0.0041

d. p-value < 0.01; reject H_0. The proportion of foods containing pesticides has declined.

60. a. H_0: $\pi \leq 0.40$

H_1: $\pi > 0.40$

b. $p = \dfrac{189}{425} = 0.45$

$$z = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{0.45 - 0.40}{\sqrt{\dfrac{0.40(1-0.40)}{425}}} = 1.88$$

p-value = 1 – 0.9699 = 0.0301

c. p-value ≤ 0.05; reject H_0. Conclude that more than 40% receive over 10 email messages per day.

61. a. H_0: $\pi = 0.63$

H_1: $\pi \neq 0.63$

b. $$z = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{0.682 - 0.63}{\sqrt{\dfrac{0.63(1-0.63)}{3000}}} = 5.90$$

c. p-value < 0.002, reject H_0. Conclude that support had changed.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Ten

Statistical Inference about Means and Proportions with Two Populations

Textbook Exercises (1-29)

Textbook Exercise Solutions

Supplementary Exercises (30-47)

Supplementary Exercise Solutions

Chapter 10: Statistical Inference about Means and Proportions with Two Populations

---

Textbook Exercises:

1 Consider the following results for two independent random samples taken from two populations.

Sample 1	Sample 2
$n_1 = 50$	$n_2 = 35$
$\bar{x}_1 = 13.6$	$\bar{x}_2 = 11.6$
$\sigma_1 = 2.2$	$\sigma_2 = 3.0$

a. What is the point estimate of the difference between the two population means?

b. Provide a 90 per cent confidence interval for the difference between the two population means.

c. Provide a 95 per cent confidence interval for the difference between the two population means.

2 Consider the following hypothesis test.

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_1: \mu_1 - \mu_2 > 0$$

The following results are for two independent samples taken from the two populations.

Sample 1	Sample 2
$n_1 = 40$	$n_2 = 50$
$\bar{x}_1 = 25.2$	$\bar{x}_2 = 22.8$
$\sigma_1 = 5.2$	$\sigma_2 = 6.0$

a. What is the value of the test statistic?

b. What is the *p*-value?

c. With $\alpha = 0.05$, what is your hypothesis testing conclusion?

3 Consider the following hypothesis test.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

The following results are for two independent samples taken from the two populations.

Sample 1	Sample 2
$n_1 = 80$	$n_2 = 70$
$\bar{x}_1 = 104$	$\bar{x}_2 = 106$
$\sigma_1 = 8.4$	$\sigma_2 = 7.6$

a. What is the value of the test statistic?

b. What is the *p*-value?

c. With $\alpha = 0.05$, what is your hypothesis testing conclusion?

4 A study of wage differentials between men and women reported that one of the reasons wages for men are higher than wages for women is that men tend to have more years of work experience than women. Assume that the following sample summaries show the years of experience for each group.

Men	Women
$n_1 = 100$	$n_2 = 85$
$\bar{x}_1 = 14.9$ years	$\bar{x}_2 = 10.3$ years
$\sigma_1 = 5.2$ years	$\sigma_2 = 3.8$ years

a. What is the point estimate of the difference between the two population means?

b. At 95 per cent confidence, what is the margin of error?

c. What is the 95 per cent confidence interval estimate of the difference between the two population means?

5 The Dublin retailer age study (used as an example above) provided the following data on the ages of customers from independent random samples taken at the two store locations.

Inner-city store	Out-of-town store
$n_1 = 36$	$n_2 = 49$
$\bar{x}_1 = 40$ years	$\bar{x}_2 = 35$ years
$\sigma_1 = 9$ years	$\sigma_2 = 10$ years

a. State the hypotheses that could be used to detect a difference between the population mean ages at the two stores.

b. What is the value of the test statistic?

c. What is the *p*-value

d. At $\alpha = 0.05$, what is your conclusion?

6 Consider the following results from a survey looking at how much people spend on gifts on Valentine's Day (14 February). The average expenditure of 40 males was €135.67, and the average expenditure of 30 females was €68.64. Based on past surveys, the standard deviation for males is assumed to be €35, and the standard deviation for females is assumed to be €20. Do males and females differ in the average amounts they spend?

a. What is the point estimate of the difference between the population mean expenditure for males and the population mean expenditure for females?
b. At 99 per cent confidence, what is the margin of error?
c. Construct a 99 per cent confidence interval for the difference between the two population means.

7 Consider the following results for independent random samples taken from two populations.

Sample 1	Sample 2
$n_1 = 20$	$n_2 = 30$
$\bar{x}_1 = 22.5$	$\bar{x}_2 = 20.1$
$s_1 = 2.5$	$s_2 = 4.8$

a. What is the point estimate of the difference between the two population means?
b. What are the degrees of freedom for the t distribution?
c. At 95 per cent confidence, what is the margin of error?
d. What is the 95 per cent confidence interval for the difference between the two population means?

8 Consider the following hypothesis test.

$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 \neq 0$

The following results are from independent samples taken from two populations.

Sample 1	Sample 2
$n_1 = 35$	$n_2 = 40$
$\bar{x}_1 = 13.6$	$\bar{x}_2 = 10.1$
$s_1 = 5.2$	$s_2 = 8.5$

a. What is the value of the test statistic?

b. What are the degrees of freedom for the t distribution?

c. What is the *p*-value?

d. At $\alpha = 0.05$, what is your conclusion?

9 Consider the following data for two independent random samples taken from two normal populations.

Sample 1	10	7	13	7	9	8
Sample 2	8	7	8	4	6	9

a. Compute the two sample means.

b. Compute the two sample standard deviations.

c. What is the point estimate of the difference between the two population means?

d. What is the 90 per cent confidence interval estimate of the difference between the two population means?

10 The International Air Transport Association surveyed business travellers to determine ratings of various international airports. The maximum possible score was 10. Suppose 50 business travellers were asked to rate airport L and 50 other business travellers were asked to rate airport M. The rating scores follow.

Airport L																									
10	9	6	7	8	7	9	8	10	7	6	5	7	3	5	6	8	7	10	8	4	7	8	6	9	
9	5	3	1	8	9	6	8	5	4	6	10	9	8	3	2	7	9	5	3	10	3	5	10	8	
Airport M																									
6	4	6	8	7	7	6	3	3	8	10	4	8	7	8	7	5	9	5	8	4	3	8	5	5	
4	4	4	8	4	5	6	2	5	9	9	8	4	8	9	9	5	9	7	8	3	10	8	9	6	

Construct a 95 per cent confidence interval estimate of the difference between the mean ratings of the airports L and M.

11 Suppose independent random samples of 15 unionized women and 20 non-unionized women in a skilled manufacturing job provide the following hourly wage rates (€).

Union workers									
22.40	18.90	16.70	14.05	16.20	20.00	16.10	16.30	19.10	16.50
18.50	19.80	17.00	14.30	17.20					
Non-union workers									
17.60	14.40	16.60	15.00	17.65	15.00	17.55	13.30	11.20	15.90
19.20	11.85	16.65	15.20	15.30	17.00	15.10	14.30	13.90	14.50

a. What is the point estimate of the difference between mean hourly wages for the two populations?
b. Develop a 95 per cent confidence interval estimate of the difference between the two population means.
c. Does there appear to be any difference in the mean wage rate for these two groups? Explain.

12 The Scholastic Aptitude Test (SAT) is a commonly used entrance qualification for university. Consider the research hypothesis that students whose parents had attained a higher level of education would on average score higher on the SAT. SAT verbal scores for independent samples of students follow. The first sample shows the SAT verbal test scores for students whose parents are college graduates with a bachelor's degree. The second sample shows the SAT verbal test scores for students whose parents are high school graduates but do not have a college degree.

Student's Parents			
College Grads		High School Grads	
485	487	442	492
534	533	580	478
650	526	479	425
554	410	486	485
550	515	528	390
572	578	524	535
497	448		
592	469		

a. Formulate the hypotheses that can be used to determine whether the sample data support the hypothesis that students show a higher population mean verbal score on the SAT if their parents attained a higher level of education.
b. What is the point estimate of the difference between the means for the two populations?
c. Compute the *p*-value for the hypothesis test.
d. At $\alpha = 0.05$, what is your conclusion?

13 Periodically, Merrill Lynch customers are asked to evaluate Merrill Lynch financial consultants and services. Higher ratings on the client satisfaction survey indicate better service, with 7 the maximum service rating. Independent samples of service

ratings for two financial consultants in the Dubai office are summarized here. Consultant A has ten years of experience while consultant B has one year of experience. Use $\alpha = 0.05$ and test to see whether the consultant with more experience has the higher population mean service rating.

Consultant A	Consultant B
$n_1 = 16$	$n_2 = 10$
$\bar{x}_1 = 6.82$	$\bar{x}_2 = 6.25$
$s_1 = 0.64$	$s_2 = 0.75$

a. State the null and alternative hypotheses.

b. Compute the value of the test statistic.

c. What is the *p*-value?

d. What is your conclusion?

14 Safegate Foods is redesigning the checkouts in its supermarkets throughout the country and is considering two designs. Tests on customer checkout times conducted at two stores where the two new systems have been installed result in the following summary of the data.

System A	System B
$n_1 = 120$	$n_2 = 100$
$\bar{x}_1 = 4.1$ minutes	$\bar{x}_2 = 3.4$ minutes
$s_1 = 2.2$ minutes	$s_2 = 1.5$ minutes

Test at the 0.05 level of significance to determine whether the population mean checkout times of the two systems differ. Which system is preferred?

15 Samples of final examination scores for two statistics classes with different instructors provided the following results.

Instructor A	Instructor B
$n_1 = 12$	$n_2 = 15$
$\bar{x}_1 = 72$	$\bar{x}_2 = 78$
$s_1 = 8$	$s_2 = 10$

With $\alpha = 0.05$, test whether these data are sufficient to conclude that the population mean grades for the two classes differ.

16 Educational testing companies provide tutoring, classroom learning, and practice tests in an effort to help students perform better on tests such as the Scholastic Aptitude Test (SAT). The test preparation companies claim that their courses will improve SAT score performances by an average of 120 points. A researcher is uncertain of this claim and believes that 120 points may be an overstatement in an effort to encourage students to take the test preparation course. In an evaluation study of one test preparation service, the researcher collects SAT score data for 35 students who took the test preparation course and 48 students who did not take the course.

	Course	No course
Sample mean	1058	983
Sample standard deviation	90	105

a. Formulate the hypotheses that can be used to test the researcher's belief that the improvement in SAT scores may be less than the stated average of 120 points.

b. Use $\alpha = 0.05$ and the data above. What is your conclusion?

c. What is the point estimate of the improvement in the average SAT scores provided by the test preparation course? Provide a 95 per cent confidence interval estimate of the improvement.

d. What advice would you have for the researcher after seeing the confidence interval?

17 Consider the following hypothesis test.

$H_0: \mu_d \leq 0$
$H_1: \mu_d > 0$

The following data are from matched samples taken from two populations.

	Population	
Element	1	2
1	21	20
2	28	26
3	18	18
4	20	20
5	26	24

a. Compute the difference value for each element.

b. Compute $\bar{d}$.

c. Compute the standard deviation s_d.

d. Conduct a hypothesis test using $\alpha = 0.05$. What is your conclusion?

18 The following data are from matched samples taken from two populations.

	Population	
Element	1	2
1	11	8
2	7	8
3	9	6
4	12	7
5	13	10
6	15	15
7	15	14

a. Compute the difference value for each element.

b. Compute $\bar{d}$.

c. Compute the standard deviation s_d.

d. What is the point estimate of the difference between the two population means?

e. Provide a 95 per cent confidence interval for the difference between the two population means.

19 In recent years, a growing array of entertainment options competes for consumer time. Researchers used a sample of 15 individuals and collected data on the hours per week spent watching cable television and hours per week spent listening to the radio.

.

Individual	Television	Radio	Individual	Television	Radio
1	22	25	9	21	21
2	8	10	10	23	23
3	25	29	11	14	15
4	22	19	12	14	18
5	12	13	13	14	17
6	26	28	14	16	15
7	22	23	15	24	23
8	19	21			

a. What is the sample mean number of hours per week spent watching cable television? What is the sample mean number of hours per week spent listening to radio? Which medium has the greater usage?
b. Use a 0.05 level of significance and test for a difference between the population mean usage for cable television and radio. What is the *p*-value?

20 A market research firm used a sample of individuals to rate the purchase potential of a particular product before and after the individuals saw a new television commercial about the product. The purchase potential ratings were based on a 0 to 10 scale, with higher values indicating a higher purchase potential. The null hypothesis stated that the mean rating 'after' would be less than or equal to the mean rating 'before'. Rejection of this hypothesis would show that the commercial improved the mean purchase potential rating. Use $\alpha = 0.05$ and the following data to test the hypothesis and comment on the value of the commercial.

	Purchase rating			Purchase rating	
Individual	After	Before	Individual	After	Before
1	6	5	5	3	5
2	6	4	6	9	8
3	7	7	7	7	5
4	4	3	8	6	6

21 Figures on profit margins (%) for 2010 and 2011 are given below for a sample of large French companies. Use the data to comment on differences between profit margins in the two years.

	Profit margin (%)	
Company	2010	2011
BNP Paribus	29.74	23.43
Carrefour	1.29	-1.50
Danone	14.64	12.59
Lafarge	8.83	4.42
L'Oreal	16.17	17.03
Michelin	8.69	9.63
Pernod-Ricard	16.43	17.94
Renault	8.61	6.17
Thales	-2.90	4.61
Vinci	8.04	7.87

a. Use $\alpha = 0.05$ and test for any difference between the population mean profit margins in 2010 and 2011. What is the *p*-value? What is your conclusion?
b. What is the point estimate of the difference between the two mean profit margins?
c. At 95 per cent confidence, what is the margin of error for the estimate in part (b)?

22 A survey was made of Book-of-the-Month-Club members to ascertain whether members spend more time watching television than they do reading. Assume a sample of 15 respondents provided the following data on weekly hours of television watching and weekly hours of reading. Using a 0.05 level of significance, can you conclude that Book-ofthe-Month-Club members spend more hours per week watching television than reading?

Respondent	Television	Reading	Respondent	Television	Reading
1	10	6	9	4	7
2	14	16	10	8	8
3	16	8	11	16	5
4	18	10	12	5	10
5	15	10	13	8	3
6	14	8	14	19	10
7	10	14	15	11	6
8	12	14			

23 Consider the following results for independent samples taken from two populations.

Sample 1	Sample 2
$n_1 = 400$	$n_2 = 300$
$p_1 = 0.48$	$p_2 = 0.36$

a. What is the point estimate of the difference between the two population proportions?

b. Construct a 90 per cent confidence interval for the difference between the two population proportions.

c. Construct a 95 per cent confidence interval for the difference between the two population proportions.

24 Consider the hypothesis test

$$H_0: \pi_1 - \pi_2 \leq 0$$
$$H_1: \pi_1 - \pi_2 > 0$$

The following results are for independent samples taken from the two populations.

Sample 1	Sample 2
$n_1 = 200$	$n_2 = 300$
$p_1 = 0.22$	$p_2 = 0.1$

a. What is the *p*-value?

b. With $\alpha = 0.05$, what is your hypothesis testing conclusion?

25 In November and December 2008, research companies affiliated to the Worldwide Independent Network of Market Research carried out polls in 17 countries to assess people's views on the economic outlook. In the Canadian survey, conducted by Léger Marketing, 61 per cent of the sample of 1511 people thought the economic situation would worsen over the next three months. In the UK survey, conducted by ICM Research, 78 per cent of the sample of 1050 felt that economic conditions would worsen over that period. Provide a 95 per cent confidence interval estimate for the difference between the population proportions in the two countries. What is your interpretation of the interval estimate?

26 In the results of the NUS 2011/12 Student Experience Research, it was reported that 34.3% of students studying Business (n = 2171) said a main reason for choosing their course was that the course was well-regarded by potential employers. The corresponding figure amongst students studying Maths and Computer Science (n = 1180) was 28.1%. Construct a 95 per cent confidence interval for the difference between the proportion of Business students who gave this as main reason and the proportion of Maths and Computer Science students who did likewise.

27 In a test of the quality of two television commercials, each commercial was shown in a separate test area six times over a one-week period. The following week a telephone survey was conducted to identify individuals who had seen the commercials. Those individuals were asked to state the primary message in the commercials. The following results were recorded.

	Commercial A	Commercial B
Number who saw commercial	150	200
Number who recalled message	63	60

a. Use $\alpha = 0.05$ and test the hypothesis that there is no difference in the recall proportions for the two commercials.

b. Compute a 95 per cent confidence interval for the difference between the recall proportions for the two populations.

28 In the UNITE 2007 *Student Experience Report*, it was reported that 49% of 1600 student respondents in UK universities considered the academic reputation of the university an important factor in their choice of university. In the 2012 *Student Experience Report*, 343 out of 488 respondents considered academic reputation to be important. Test the hypothesis $\pi_1 - \pi_2 = 0$ with $\alpha = 0.05$. What is the p-value? What is your conclusion?

29 A large car insurance company selected samples of single and married male policyholders and recorded the number who made an insurance claim over the preceding three-year period.

Single policyholders	Married policyholders
$n_1 = 400$ Number making claims = 76	$n_2 = 900$ Number making claims = 90

a. Use $\alpha = 0.05$ and test to determine whether the claim rates differ between single and married male policyholders.

b. Provide a 95 per cent confidence interval for the difference between the proportions for the two populations.

Chapter 10: Statistical Inference about Means and Proportions with Two Populations

Textbook Exercises Solutions:

1. a. $\bar{x}_1 - \bar{x}_2 = 13.6 - 11.6 = 2$

   b. $z_{\alpha/2} = z_{0.05} = 1.645$

   $$\bar{x}_1 - \bar{x}_2 \pm 1.645\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

   $$= 2 \pm 1.645\sqrt{\frac{(2.2)^2}{50} + \frac{(3)^2}{35}}$$

   $= 2 \pm 0.98$ (1.02 to 2.98)

   c. $z_{\alpha/2} = z_{0.025} = 1.96$

   $$2 \pm 1.96\sqrt{\frac{(2.2)^2}{50} + \frac{(3)^2}{35}}$$

   $= 2 \pm 1.17$ (0.83 to 3.17)

2. a. $$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(25.2 - 22.8) - 0}{\sqrt{\frac{(5.2)^2}{40} + \frac{6^2}{50}}} = 2.03$$

   b. p-value = 1 – 0.9788 = 0.0212

   c. p-value < 0.05, reject H_0.

3. a. $$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(104 - 106) - 0}{\sqrt{\frac{(8.4)^2}{80} + \frac{(7.6)^2}{70}}} = -1.53$$

   b. p-value = 2(0.0630) = 0.1260

   c. p-value > 0.05, do not reject H_0.

4. a. $\bar{x}_1 - \bar{x}_2 = 14.9 - 10.3 = 4.6$ years

b. $z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1.96\sqrt{\frac{(5.2)^2}{100} + \frac{(3.8)^2}{85}} = 1.3$

c. 4.6 ± 1.3 (3.3 to 5.9)

5. a. H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

b. $z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(40-35)-0}{\sqrt{\frac{9^2}{36} + \frac{10^2}{49}}} = 2.41$

c. p-value = 2(1 − 0.9920) = 0.0160

d. p-value < 0.05; reject H_0. There is a difference between the population mean ages at the two stores.

6. a. Point estimate is €(135.67 – 68.64) = €67.03

b. At 99% confidence, margin of error is $z_{0.005}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.576\sqrt{\frac{(35)^2}{40} + \frac{(20)^2}{30}} = €17.08$

c. Confidence interval is €(67.03 ± 17.08) = €49.95 to €84.11.

7. a. $\bar{x}_1 - \bar{x}_2 = 22.5 - 20.1 = 2.4$

b. $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{2.5^2}{20} + \frac{4.8^2}{30}\right)^2}{\frac{1}{19}\left(\frac{2.5^2}{20}\right)^2 + \frac{1}{29}\left(\frac{4.8^2}{30}\right)^2} = 45.8$

Use $df = 45$

c. $t_{0.025} = 2.014$

$t_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.014\sqrt{\frac{2.5^2}{20} + \frac{4.8^2}{30}} = 2.1$

d. 2.4 ± 2.1 (0.3 to 4.5)

8. a. $$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(13.6 - 10.1) - 0}{\sqrt{\frac{5.2^2}{35} + \frac{8.5^2}{40}}} = 2.18$$

b. $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{5.2^2}{35} + \frac{8.5^2}{40}\right)^2}{\frac{1}{34}\left(\frac{5.2^2}{35}\right)^2 + \frac{1}{39}\left(\frac{8.5^2}{40}\right)^2} = 65.7$$

Use $df = 65$.

c. Using t table, area in tail is between 0.01 and 0.025.

Therefore two-tailed p-value is between 0.02 and 0.05. (Actual p-value = 0.0329.)

d. p-value < 0.05, reject H_0.

9. a. $$\bar{x}_1 = \frac{54}{6} = 9 \qquad \bar{x}_2 = \frac{42}{6} = 7$$

b. $$s_1 = \sqrt{\frac{\Sigma(x_i - \bar{x}_1)^2}{n_1 - 1}} = 2.28$$

$$s_2 = \sqrt{\frac{\Sigma(x_i - \bar{x}_2)^2}{n_2 - 1}} = 1.79$$

c. $$\bar{x}_1 - \bar{x}_2 = 9 - 7 = 2$$

d. $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{2.28^2}{6} + \frac{1.79^2}{6}\right)^2}{\frac{1}{5}\left(\frac{2.28^2}{6}\right)^2 + \frac{1}{5}\left(\frac{1.79^2}{6}\right)^2} = 9.5$$

Use $df = 9$, $t_{0.05} = 1.833$.

$$\bar{x}_1 - \bar{x}_2 \pm 1.833\sqrt{\frac{2.28^2}{6} + \frac{1.79^2}{6}}$$

2 ± 2.17 (–0.17 to 4.17)

10. Computer used to obtain the following:

L	M
$n_1 = 50$	$n_2 = 50$
$\bar{x}_1 = 6.72$	$\bar{x}_2 = 6.34$
$s_1 = 2.37$	$s_2 = 2.16$

$\bar{x}_1 - \bar{x}_2 = 6.72 - 6.34 = 0.38$

Airport L is rated slightly more highly.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{2.37^2}{50} + \frac{2.16^2}{50}\right)^2}{\frac{1}{49}\left(\frac{2.37^2}{50}\right)^2 + \frac{1}{49}\left(\frac{2.16^2}{50}\right)^2} = 97.2$$

Use $df = 97$, $t_{0.025} = 1.985$.

$$0.38 \pm 1.985\sqrt{\frac{2.37^2}{50} + \frac{2.16^2}{50}}$$

0.38 ± 0.90 (–0.52 to 1.28)

11. Computer used to obtain the following:

Union	Non-union
$n_1 = 14$	$n_2 = 19$
$\bar{x}_1 = 17.54$	$\bar{x}_2 = 15.36$
$s_1 = 2.24$	$s_2 = 1.99$

a. $\bar{x}_1 - \bar{x}_2 = 17.54 - 15.36 = €2.18$ per hour greater for union workers.

b. $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{2.24^2}{14} + \frac{1.99^2}{19}\right)^2}{\frac{1}{13}\left(\frac{2.24^2}{14}\right)^2 + \frac{1}{18}\left(\frac{1.99^2}{19}\right)^2} = 26.1$$

Use $df = 26$, $t_{0.025} = 2.056$.

$$2.18 \pm 2.056\sqrt{\frac{2.24^2}{14} + \frac{1.99^2}{19}}$$

2.18 ± 1.55 (0.63 to 3.73)

c. Yes, since the interval does not include 0, it can be concluded that union workers have a different (higher) mean wage rate compared with non-union workers.

12. a. H_0: $\mu_1 - \mu_2 \leq 0$

H_1: $\mu_1 - \mu_2 > 0$

where population 1 refers to students whose parents attained a higher level of education

b. $\bar{x}_1 = 525$, $\bar{x}_2 = 487$, point estimate of the difference is $\bar{x}_1 - \bar{x}_2 = 38$

c. $s_1 = 59.4$, $s_2 = 51.7$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{38 - 0}{\sqrt{\frac{(59.4)^2}{16} + \frac{(51.7)^2}{12}}} = 1.80$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{(n_1 - 1)}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{(n_2 - 1)}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{(59.4)^2}{16} + \frac{(51.7)^2}{12}\right)^2}{\frac{1}{15}\left(\frac{(59.4)^2}{16}\right)^2 + \frac{1}{11}\left(\frac{(51.7)^2}{12}\right)^2} = 25$$

$t = 1.80$, $df = 25$

Using t table, p-value is between 0.025 and 0.05 (exact p-value = 0.0420)

d. Reject H_0; conclude higher mean score if parent is a college graduate.

13. a. H_0: $\mu_1 - \mu_2 \leq 0$

H_1: $\mu_1 - \mu_2 > 0$

b. $$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(6.82 - 6.25) - 0}{\sqrt{\frac{0.64^2}{16} + \frac{0.75^2}{10}}} = 1.99$$

c. $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{.64^2}{16} + \frac{.75^2}{10}\right)^2}{\frac{1}{15}\left(\frac{0.64^2}{16}\right)^2 + \frac{1}{9}\left(\frac{0.75^2}{10}\right)^2} = 16.9$$

Use $df = 16$.

Using t table, p-value is between 0.025 and 0.05. (Actual p-value = 0.0318.)

d. p-value < 0.05, reject H_0. The consultant with more experience has a higher population mean rating.

14. H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(4.1 - 3.4) - 0}{\sqrt{\frac{(2.2)^2}{120} + \frac{(1.5)^2}{100}}} = 2.79$$

p-value = 2(1 − 0.9974) = 0.0052.

p-value < 0.05, reject H_0. A difference exists with system B having the lower mean checkout time.

15. H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(72 - 78) - 0}{\sqrt{\frac{8^2}{12} + \frac{10^2}{15}}} = -1.73$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{8^2}{12} + \frac{10^2}{15}\right)^2}{\frac{1}{11}\left(\frac{8^2}{12}\right)^2 + \frac{1}{14}\left(+\frac{10^2}{15}\right)^2} = 25$$

Using t table, area is between 0.025 and 0.05.

Two-tailed p-value is between 0.05 and 0.10. (Actual p-value = 0.0961.)

p-value > 0.05, do not reject H_0. Cannot conclude a difference exists.

16. a. H_0: $\mu_1 - \mu_2 \geq 120$
H_1: $\mu_1 - \mu_2 < 120$

b. $$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(1058 - 983) - 120}{\sqrt{\frac{90^2}{35} + \frac{105^2}{48}}} = -2.10$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{90^2}{35} + \frac{105^2}{48}\right)^2}{\frac{1}{34}\left(\frac{90^2}{35}\right)^2 + \frac{1}{47}\left(\frac{105^2}{48}\right)^2} = 78.8 \qquad \text{Use } df = 78.$$

Using t table, p-value is between 0.01 and 0.025. (Actual p-value = 0.0197.)
p-value < 0.05, reject H_0. The improvement is less than the stated average of 120 points.

c. $\bar{x}_1 - \bar{x}_2 \pm t_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$df = 78$

$(1058 - 983) \pm 1.991\sqrt{\frac{90^2}{35} + \frac{105^2}{48}}$

75 ± 43 (32 to 118)

d. This is a wide interval. A larger sample should be used to reduce the margin of error.

17. a. 1, 2, 0, 0, 2

b. $\bar{d} = \sum d_i / n = 5/5 = 1$

c. $s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{4}{5-1}} = 1$

d. $t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{1-0}{1/\sqrt{5}} = 2.24$

$df = n - 1 = 4$

Using t table, p-value is between 0.025 and 0.05. (Actual p-value = 0.0445.)

Reject H_0; conclude $\mu_d > 0$.

18. a. 11 – 8 = 3, 7 – 8 = -1, 9 – 6 = 3, 12 – 7 = 5, 13 –1 10 = 3, 15 – 15 = 0, 15 – 14 = 1

b. $\bar{d} = \sum d_i / n = 14/7 = 2$

c. $s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{26}{7-1}} = 2.08$

d. $\bar{d} = 2$

e. With 6 degrees of freedom $t_{0.025} = 2.447$

The confidence interval is $2 \pm 2.447\left(2.082/\sqrt{7}\right) =$ 2 ± 1.93, or 0.07 to 3.93.

19. a. Difference = hours radio – hours TV

H_0: $\mu_d = 0$
H_1: $\mu_d \neq 0$

$\bar{d} = 1.2$ and $s_d = 1.97$

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{1.2 - 0}{1.97 / \sqrt{15}} = 2.36$$

$df = n - 1 = 14$

Using t table, p-value (2-tailed) is between 0.02 and 0.05 (actual p-value = 0.0335)

Reject H_0; conclude that there is a significant difference between mean hours listening to radio and mean hours watching cable TV.

b. Sample mean for cable TV is 18.8 hours; sample mean for radio is 20.0 hours. More time on average spent listening to radio.

20. Difference = rating after – rating before

H_0: $\mu_d \leq 0$
H_1: $\mu_d > 0$

$\bar{d} = 0.625$ and $s_d = 1.30$

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{0.625 - 0}{1.30 / \sqrt{8}} = 1.36$$

$df = n - 1 = 7$

Using t table, p-value is between 0.10 and 0.20

Actual p-value = 0.1084

Do not reject H_0; we cannot conclude that seeing the commercial improves the mean potential to purchase.

21. a. Differences (2010 minus 2011) are 6.31, 2.79, 2.05, 4.41, -0.86, -0.94, -1.51, 2.44, -7.51, 0.17

$\bar{d} = \sum d_i / n = 7.35/10 = 0.735$

$s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}} = 3.83$

$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{0.735 - 0}{3.83/\sqrt{10}} = 0.607$

Using *t* table, area in tail is greater than 0.20.

Two-tailed *p*-value is greater than 0.40. (Actual *p*-value = 0.559.)

Cannot reject H_0. No significant difference observed.

b. $\bar{d} = 0.735$. Average profit margin was higher in 2010 than in 2011.

c. $\pm t_{0.025} \frac{s_d}{\sqrt{n}}$ $df = 9$ $t = 2.262$

$\pm 2.262 \times \frac{3.83}{\sqrt{10}} = \pm 2.74$

Given the point estimate of 0.735, the margin of error is large. A larger sample is needed.

22. Using matched samples, the differences are as follows: 4, –2, 8, 8, 5, 6, –4, –2, –3, 0, 11, –5, 5, 9, 5

H_0: $\mu_d \leq 0$
H_1: $\mu_d > 0$

$\bar{d} = 3$ and $s_d = 5.21$

$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{3 - 0}{5.21/\sqrt{15}} = 2.23$

$df = n - 1 = 14$

Using *t* table, *p*-value is between 0.01 and 0.025. (Actual *p*-value = 0.0213.)

Reject H_0. Conclude that the population of readers spends more time, on average, watching television than reading.

23. a. $p_1 - p_2 = 0.48 - 0.36 = 0.12$

b. $$p_1 - p_2 \pm z_{0.05}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$0.12 \pm 1.645\sqrt{\frac{0.48(1-0.48)}{400} + \frac{0.36(1-0.36)}{300}}$$

0.12 ± 0.0614 (0.0586 to 0.1814)

c. $$0.12 \pm 1.96\sqrt{\frac{0.48(1-0.48)}{400} + \frac{0.36(1-0.36)}{300}}$$

0.12 ± 0.0731 (0.0469 to 0.1931)

24. a. $$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{200(0.22) + 300(0.16)}{200 + 300} = 0.1840$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.22 - 0.16}{\sqrt{0.1840(1-0.1840)\left(\frac{1}{200} + \frac{1}{300}\right)}} = 1.70$$

p-value = 1 – 0.9554 = 0.0446

b. p-value < 0.05; reject H_0.

25. $p_1 = 0.61$ $p_2 = 0.78$

95% confidence interval is $(p_2 - p_1) \pm z_{0.025}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

$$= (0.78 - 0.61) \pm 1.96\sqrt{\frac{0.61(1-0.61)}{1511} + \frac{0.78(1-0.78)}{1050}}$$

$= 0.17 \pm 0.035$ (0.135 to 0.205)

A higher proportion of people in the UK thought that economic conditions would worsen, the sample difference between the UK and Canada being 17 percentage points. The confidence interval shows the population difference may be from 13.5 to 20.5 percentage points.

26. $p_1 = 0.343, n_1 = 2171 \qquad p_2 = 0.281, n_2 = 1180$

$$p_1 - p_2 \pm z_{0.025}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$= (0.343 - 0.281) \pm 1.96\sqrt{\frac{0.343(1-0.343)}{2171} + \frac{0.281(1-0.281)}{1180}}$$

$= 0.062 \pm (1.96 \text{ x } 0.0166) = 0.062 \pm 0.0325$ (0.0295 to 0.0945)

The course being well-regarded by potential employers was given as a main reason by a higher proportion of Business students than of Maths & Computer Science students, the sample difference being 6.2 percentage points. The confidence interval shows the difference in the populations may be from 3.0 to 9.5 percentage points, at the 95% confidence level.

27. a. H_0: $\pi_1 - \pi_2 = 0$

H_1: $\pi_1 - \pi_2 \neq 0$

$p_1 = 63/150 = 0.42$

$p_2 = 60/200 = 0.30$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{63 + 60}{150 + 200} = 0.3514$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.42 - 0.30}{\sqrt{0.3514(1-0.3514)\left(\frac{1}{150} + \frac{1}{200}\right)}} = 2.33$$

p-value = 2(1 − 0.9901) = 0.0198

p-value < 0.05, reject H_0. There is a difference between the recall rates for the two commercials.

b. $$p_1 - p_2 \pm z_{0.025}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$0.42 - 0.30 \pm 1.96\sqrt{\frac{0.42(1-0.42)}{150} + \frac{0.30(1-0.30)}{200}}$$

0.12 ± 0.1014 (0.0186 to 0.2214)

Commercial A has the better recall rate.

28. H_0: $\pi_1 - \pi_2 = 0$

H_1: $\pi_1 - \pi_2 \neq 0$

$p_1 = 0.49, \quad n_1 = 1600$

$p_2 = (343/488) = 0.7029, \quad n_2 = 488$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(1600 \times 0.49) + 343}{(1600 + 488)} = 0.54$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.49 - 0.7029}{\sqrt{0.54(1-0.54)\left(\frac{1}{1600} + \frac{1}{488}\right)}} = \frac{-0.2129}{0.02577} = -8.26$$

p-value < 0.002.

p-value < 0.05, reject H_0. There is a difference between the two years in the proportions of students considering academic reputation as an important factor (higher proportion in 2012).

29. a. $p_1 = 76/400 = 0.19$

$p_2 = 90/900 = 0.10$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{76 + 90}{400 + 900} = 0.1277$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.19 - 0.10}{\sqrt{0.1277(1-0.1277)\left(\frac{1}{400} + \frac{1}{900}\right)}} = 4.49$$

p-value is very close to 0

Reject H_0; conclude that there is a difference between claim rates.

b. $$p_1 - p_2 \pm z_{0.025}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$0.19 - 0.10 \pm 1.96\sqrt{\frac{0.19(1-0.19)}{400} + \frac{0.10(1-0.10)}{900}}$$

0.09 ± 0.0432 (0.0468 to 0.1332)

Claim rates are higher for single males.

Chapter 10: Statistical Inference about Means and Proportions with Two Populations

Supplementary Exercises:

30. A sample of 40 petrol stations in city A yielded a mean price for unleaded petrol of €1.54 per litre. A sample of 35 petrol stations in city Y yielded a mean price of €1.22 per litre. Assume that prior studies indicate a population standard deviation of €0.10 in city X and €0.08 in city Y.
    a. Calculate a point estimate of the difference between the population mean prices per litre in city X and city Y.
    b. At 95% confidence, what is the margin of error?
    c. What is the 95% confidence interval estimate of the difference between the population mean prices per litre in the two cities?

31. Consider the following data on mortgage loans. A sample of 270 loans made in 2006 yielded a mean loan of €175,000. A sample of 250 loans made in 2005 had mean of €165,000. Based on historical loan data, the population standard deviations for the loan amounts can be assumed known at €55,000 in 2006 and €50,000 in 2005. Do the sample data indicate an increase in the mean loan amount between 2005 and 2006? Use $\alpha = 0.05$.

32. Starting salaries per annum for individuals with master's and bachelor's degrees in business were collected in two independent random samples. Use the following data to construct a 90% confidence interval estimate of the increase in starting salary that can be expected upon completion of a master's programme.

Master's Degree	**Bachelor's Degree**
$n_1 = 60$	$n_2 = 80$
$\bar{x}_1 = €45{,}000$	$\bar{x}_2 = €35{,}000$
$\sigma_1 = €4000$	$\sigma_2 = €3500$

33. Suppose that for a simple random sample of 50 residents in Bigton the mean number of kilometres travelled to work per day is 22.5, and the standard deviation is 8.4 kilometres. For an independent simple random sample of 40 residents in Smallville the mean is 18.6 kilometres per day and the standard deviation is 7.4 kilometres per day.
    a. What is the point estimate of the difference between the mean number of kilometres that Bigton residents travel to work per day and the mean number of kilometres that Smallville residents travel to work per day?
    b. What is the 95% confidence interval for the difference between the two population means?

34. Data were collected on residents living in the coastal communities as well as on residents living in non-coastal areas throughout the United States (*USA Today*, July 21, 2000). Assume that the following sample results were obtained on the ages of individuals in the two populations:

Coastal areas	**Non-coastal areas**
$n_1 = 150$	$n_2 = 175$
$\bar{x}_1 = 39.3$ years	$\bar{x}_2 = 35.4$ years
$s_1 = 16.8$ years	$s_2 = 15.2$ years

    Test the hypothesis of no difference between the two population means. Use $\alpha = 0.05$.
    a. Formulate the null and alternative hypotheses.
    b. What is the value of the test statistic?
    c. What is the *p*-value?
    d. What is your conclusion?

35. The Educational Testing Service conducted a study to investigate differences between the scores of male and female students on the Scholastic Aptitude Test (SAT). The study identified a random sample of 562 female and 852 male students who achieved the same high score on the mathematics portion of the test. The SAT verbal scores for the two samples are as follows.

Female Students	**Male Students**
$\bar{x}_1 = 547$	$\bar{x}_2 = 525$
$s_1 = 83$	$s_2 = 78$

a. Do the data support the conclusion that, given a population of female students and a population of male students with similarly high mathematical abilities, the female students will score significantly higher on the verbal portion? Test at a 0.01 level of significance. What is your conclusion?
b. Provide a 95% confidence interval for the difference between the population mean scores.

36. File "Mutual"

Mutual funds are classified as *load* or *no-load* funds. Load funds require an investor to pay an initial fee based on a percentage of the amount invested in the fund. The no-load funds do not require this initial fee. Some financial advisors argue that the load mutual funds may be worth the extra fee because these funds provide a higher mean rate of return than the no-load mutual funds. A sample of 30 load mutual funds and a sample of 30 no-load mutual funds were selected from *Barron's Lipper Mutual Funds Quarterly,* January 12, 1998. Data were collected on the annual return for the funds over a five-year period. The data are contained in the data set "Mutual".

a. Formulate H_0 and H_1 such that rejection of H_0 leads to the conclusion that the load mutual funds have a higher mean annual return over the five-year period.
b. Use the 60 mutual funds in the data set Mutual to conduct the hypothesis test. What is the *p*-value? At $\alpha = 0.05$, what is your conclusion?

37. The cost of travel from airport to city centre depends on the method of travel. One-way costs for taxi and bus travel for a sample of 10 major cities follow. Provide a 95% confidence interval for the mean cost increase associated with taxi travel.

City	Taxi (€)	Bus (€)	City	Taxi (€)	Bus (€)
A	15.00	7.00	**F**	16.50	7.50
B	22.00	12.50	**G**	18.00	7.00
C	11.0	5.00	**H**	16.00	8.50
D	15.00	4.50	**I**	20.00	8.00
E	26.0	11.00	**J**	10.00	5.00

38. A manufacturer produces both a deluxe and a standard model of a domestic kettle. Selling prices obtained from a sample of retail outlets follow.

	Model price (€)			Model price (€)	
Retail outlet	**Deluxe**	**Standard**	**Retail outlet**	**Deluxe**	**Standard**
1	39	27	**5**	40	30
2	39	28	**6**	39	34
3	45	35	**7**	35	29
4	38	30			

a. The manufacturer's suggested retail prices for the two models show a €10 price differential. Use a 0.05 level of significance and test that the mean difference between the prices of the two models is €10.

b. What is the 95% confidence interval for the difference between the mean prices of the two models?

39. StreetInsider.com reported 2002 earnings per share data for a sample of major companies (12 February 2003). Prior to 2002, financial analysts predicted the 2002 earnings per share for these same companies (Barron's, 10 September 2001). Use the following data to comment on differences between actual and estimated earnings per share.

Company	Actual	Predicted
AT & T	1.29	0.38
American Express	2.01	2.31
Citigroup	2.59	3.43
Coca Cola	1.60	1.78
DuPont	1.84	2.18
Exxon-Mobil	2.72	2.19
General Electric	1.51	1.71
Johnson & Johnson	2.28	2.18
McDonald's	0.77	1.55
Wal-Mart	1.81	1.74

a. Use $\alpha = 0.05$ and test for any difference between the population mean actual and population mean estimated earnings per share. What is the p-value? What is your conclusion?

b. What is the point estimate of the difference between the two means? Did the analysts tend to underestimate or overestimate the earnings?

c. At 95 per cent confidence, what is the margin of error for the estimate in part (b)? What would you recommend based on this information?

40. The Asian economy faltered during the last few months of 1997. Investors anticipated that the downturn in the Asian economy would have a negative effect on the earnings of companies in the United States during the fourth quarter of 1997. The following sample data show the earnings per share for the fourth quarter of 1996 and the fourth quarter of 1997 (*The Wall Street Journal*, January 28, 1998).
    a. Formulate H_0 and H_1 such that rejection of H_0 leads to the conclusion that the mean earnings per share for the fourth quarter of 1997 are less than for the fourth quarter of 1996.
    b. Use the data to conduct the hypothesis test. At $\alpha = 0.05$, what is your conclusion?

Company	**Earnings 1996**	**Earnings 1997**
Atlantic Richfield	1.16	1.17
Balchem Corp.	0.16	0.13
Black & Decker Corp	0.97	1.02
Dial Corp.	0.18	0.23
Communications	0.15	-0.32
Eastman Chemical	0.77	0.36
Excel Communications	0.28	-0.14
Federal Signal	0.40	0.29
Ford Motor Company	0.97	1.45
GTE Corp	0.81	0.73
ITT Industries	0.59	0.60
Kimberley-Clark	0.61	-0.27
Minnesota Mining & Mfr	0.91	0.89
Proctor & Gamble	0.63	0.71

41. Slot machines are a popular game in casinos. The following sample data show the number of women and number of men who selected slot machines as their favourite casino game.

	Women	**Men**
Sample size	320	250
Favourite game slot machines	256	165

a. What is the point estimate of the proportion of women who say slot machines are their favourite casino game?
b. What is the point estimate of the proportion of men who say slot machines are their favourite casino game?
c. Provide a 95% confidence interval estimate of the difference between the proportion of women and proportion of men who say slot machines are their favourite casino game.

42. Consider the following sample data on airline flight arrival times at a major airport. In January 2005, a sample of 924 flights showed 742 on time. In January 2006, a sample of 841 flights showed 714 on time.
a. What is the point estimate of on-time flights in January 2005?
b. What is the point estimate of on-time flights in January 2006?
c. Let π_1 denote the population proportion of on-time flights in January 2005 and π_2 denote the population proportion of on-time flights in January 2006. State the hypotheses that could be tested to determine whether the airlines improved on-time flight performance during the one-year period.
d. At $\alpha = 0.05$, what is your conclusion? What is the *p*-value?

43. A 2003 *New York Times*/CBS News poll sampled 523 adults who were planning a vacation during the next six months and found that 141 were expecting to travel by air (*New York Times New Service*, March 2, 2003). A similar survey question in a May 1993 *New York Times*/CBS News poll found that of 477 adults who were planning a vacation in the next six months, 81 were expecting to travel by air.
a. State the hypotheses that can be used to determine whether a significant change occurred in the population proportion planning to travel by air over the 10-year period.
b. What is the sample proportion expecting to travel by air in 2003? In 1993?
c. Use $\alpha = 0.01$ and test for a significant difference. What is your conclusion?

44. In a *Business Week*/Harris poll in 2000, adult interviewees were asked how well U.S. companies competed in the global economy. Of 1035 respondents, 704 answered good/excellent. In a similar poll of 1004 adults in 1996, 582 respondents answered good/excellent. Can the sample results be used to conclude that the proportion of adults responding good/excellent has increased over the four years from 1996 to 2000?
    a. State the null and alternative hypotheses.
    b. Compute the *p*-value.
    c. At $\alpha = 0.01$, what is your conclusion?

45. The Anwar Sadat Chair for Peace and Development carried out an opinion poll among adults in six African and Arab states in May 2004. The results show that 69 per cent of 400 respondents in Jordan felt that the war in Iraq had brought less democracy to the country, compared with 57 per cent of 700 respondents in Lebanon who had that view. Construct a 95 per cent confidence interval for the difference between the proportion of Jordanian adults who held this view and the proportion of Lebanese adults who held this view.

46. Medical tests were conducted to learn about drug-resistant tuberculosis. Of 142 cases tested in city X, nine were found to be drug-resistant. Of 268 cases tested in city Y, five were found to be drug-resistant. Do these data suggest a statistically significant difference between the proportions of drug-resistant cases in the two cities? Use a 0.02 level of significance. What is the *p*-value and what is your conclusion?

47. UNITE/MORI published annual 'Student Experience Reports' from 2001 to 2005, based on face-to-face interviews carried out at a sample of UK universities. In 2001, it was reported that 74 per cent of 1103 respondents strongly agreed with the statement that 'going to university is a worthwhile experience'. The 2005 report says that 66 per cent of 1065 respondents strongly agreed with this statement. Test the hypothesis $\pi_1 - \pi_2 = 0$ with $\alpha = 0.05$. What is the *p*-value. What is your conclusion?

Chapter 10: Statistical Inference about Means and Proportions with Two Populations

Supplementary Exercises Solutions:

30. a. $\bar{x}_1 - \bar{x}_2 = 1.54 - 1.22 = 0.32$

b. $z_{0.025}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1.96\sqrt{\frac{(0.10)^2}{40} + \frac{(0.08)^2}{35}} = 1.96(0.0208) = 0.04$

c. 0.32 ± 0.04 or 0.28 to 0.36

31. μ_1 = population mean loan amount for 2006
μ_2 = population mean loan amount for 2005

H_0: $\mu_1 - \mu_2 \leq 0$
H_1: $\mu_1 - \mu_2 > 0$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(175 - 165) - 0}{\sqrt{\frac{55^2}{270} + \frac{50^2}{250}}} = 2.17$$

p-value = 1 – 0.9850 = 0.0150

p-value < 0.05; reject H_0. Conclude that the mean loan amount has increased between 2005 and 2006.

32. $\bar{x}_1 - \bar{x}_2 \pm z_{0.05}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = (45,000 - 35,000) \pm 1.645\sqrt{\frac{(4000)^2}{60} + \frac{(3500)^2}{80}}$

= 10,000 ± 1066 or €8934 to €11,066

33. a. $\bar{x}_1 - \bar{x}_2 = 22.5 - 18.6 = 3.9$

b. $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{8.4^2}{50} + \frac{7.4^2}{40}\right)^2}{\frac{1}{49}\left(\frac{8.4^2}{50}\right)^2 + \frac{1}{39}\left(\frac{7.4^2}{40}\right)^2} = 87.1$$

Use $df = 87$, $t_{0.025} = 1.988$

$$3.9 \pm 1.988\sqrt{\frac{8.4^2}{50} + \frac{7.4^2}{40}} = 3.9 \pm 3.3 \quad \text{or 0.6 to 7.2}$$

34. a. H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

b. $$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(39.3 - 35.4) - 0}{\sqrt{\frac{16.8^2}{150} + \frac{15.2^2}{175}}} = 2.18$$

c. With $n_1 = 150$ and $n_2 = 175$, degrees of freedom will be well over 100. Using last row of the t table, the area in the tail at $t = 2.18$ is between 0.01 and 0.025. Hence, the two-tailed p-value is between 0.02 and 0.05

(Actual p-value = 0.03)

d. p-value < 0.05, reject H_0. Conclude that the population mean ages differ for the coastal and non-coastal areas.

35. a. H_0: $\mu_1 - \mu_2 \leq 0$

H_1: $\mu_1 - \mu_2 > 0$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(547 - 525) - 0}{\sqrt{\frac{83^2}{562} + \frac{78^2}{852}}} = 4.99$$

With a total sample size 562 + 852 = 1414, use infinity row of t distribution table.

Using t table, p-value is less than 0.005

p-value < 0.01, reject H_0. Conclude that females have a higher mean verbal score.

b. $\bar{x}_1 - \bar{x}_2 \pm t_{0.025}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

Use $t_{0.025} = 1.960$

$$(547 - 525) \pm 1.960\sqrt{\frac{83^2}{562} + \frac{78^2}{852}} = 22 \pm 8.6 \qquad \text{or 13.4 to 30.6}$$

36. a. H_0: $\mu_1 - \mu_2 \le 0$
H_1: $\mu_1 - \mu_2 > 0$

b. Using the computer,

$n_1 = 30$ $\qquad$ $n_2 = 30$
$\bar{x}_1 = 16.23$ $\qquad$ $\bar{x}_2 = 15.70$
$s_1 = 3.52$ $\qquad$ $s_2 = 3.31$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{(16.23 - 15.70) - 0}{\sqrt{\dfrac{(3.52)^2}{30} + \dfrac{(3.31)^2}{30}}} = 0.60$$

$$df = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^2}{n_2}\right)^2} = \frac{\left(\dfrac{3.52^2}{30} + \dfrac{3.31^2}{30}\right)}{\dfrac{1}{29}\left(\dfrac{3.52^2}{30}\right)^2 + \dfrac{1}{29}\left(\dfrac{3.31^2}{30}\right)^2} = 57.8$$

Use $df = 57$

Using t table, p-value is greater than 0.20

(Actual p-value = 0.2753)

p-value > 0.05, do not reject H_0. Cannot conclude that the load funds have a greater mean rate of return.

37. Differences: 8, 9.5, 6, 10.5, 15, 9, 11, 7.5, 12, 5
$\bar{d} = 93.5/10 = 9.35$ and $s_d = 2.95$

$t_{0.025} = 2.262$ with $df = n - 1 = 9$

$9.35 \pm 2.262\left(2.95/\sqrt{10}\right) = 9.35 \pm 2.11$

Interval estimate is 7.24 to 11.46

38. a. Difference = Price deluxe – Price Standard

H_0: $\mu_d = 10$

H_1: $\mu_d \neq 10$

$\bar{d} = 8.86$ and $s_d = 2.61$

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{8.86 - 10}{2.61 / \sqrt{7}} = -1.16$$

$df = n - 1 = 6$

Using t table, area is between 0.10 and 0.20

Two-tailed p-value is between 0.20 and 0.40

(Actual p-value = 0.2906)

Do not reject H_0. We cannot reject the hypothesis that a €10 price differential exists.

b. 95% confidence interval for the difference between the mean prices is $8.86 \pm (2.447 \times 2.61) = 8.86 \pm 6.39$ i.e. €2.47 to €15.25 (note that this interval includes the value €10).

39. Differences 0.91, –0.30, –0.84, –0.18, –0.34, 0.53, –0.20, 0.10, –0.78, 0.07

$$\bar{d} = \sum d_i / n = -1.03 / 10 = -0.103$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = 0.54$$

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{-0.103 - 0}{0.54 / \sqrt{10}} = -0.60$$

Using t table, area in tail is greater than 0.20.

Two-tailed p-value is greater than 0.40. (Actual p-value = 0.5602.)

Cannot reject H_0. No significant difference observed.

b. $\bar{d} = -0.103$. Actual was less than predicted. Analysts over-estimated earnings.

c. $\pm t_{0.025} \dfrac{s_d}{\sqrt{n}}$ $df = 9$ $t = 2.262$

$$\pm 2.262 \frac{0.54}{\sqrt{10}} = \pm 0.39$$

Given the point estimate of –0.103, the margin of error is large. A larger sample is needed.

40. a. H_0: $\mu_1 - \mu_2 \leq 0$

H_1: $\mu_1 - \mu_2 > 0$

b. Using difference data,

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{1.74}{14} = 0.124$$

$$s_d = \sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n-1}} = 0.33$$

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{0.124 - 0}{0.33/\sqrt{14}} = 1.42$$

Degrees of freedom = $n - 1 = 13$

Using t table, p-value is between 0.05 and 0.10

(Actual p-value = 0.0889)

p-value > 0.05, do not reject H_0. Data does not support the conclusion that population mean earnings are down in 1997.

41. a. $p_1 = 256/320 = 0.80$

b. $p_2 = 165/250 = 0.66$

c. $p_1 - p_2 = 0.80 - 0.66 = 0.14$

$$0.14 \pm z_{0.025} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$= 0.14 \pm 1.96 \sqrt{\frac{0.80(1-0.80)}{320} + \frac{0.66(1-0.66)}{250}}$$

$= 0.14 \pm 0.0733$ or 0.0667 to 0.2133

42. a. $p_1 = 742/924 = 0.803$

b. $p_2 = 714/841 = 0.849$

c. H_0: $\pi_1 - \pi_2 \geq 0$

H_1: $\pi_1 - \pi_2 < 0$

Support for H_1 will sugest $\pi_2 > \pi_1$ which indicates an improvement in on-time performance.

d. $$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{742 + 714}{924 + 841} = 0.8249$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.803 - 0.849}{\sqrt{0.8249(1-0.8249)\left(\frac{1}{924} + \frac{1}{841}\right)}} = -2.54$$

p-value = 0.0055

Reject H_0. Conclude that on-time performance has improved.

43. a. H_0: $\pi_1 - \pi_2 = 0$

H_1: $\pi_1 - \pi_2 \neq 0$

b. $p_1 = 141/523 = 0.2696$ (27%)

$p_2 = 81/477 = 0.1698$ (17%)

c. $$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{141 + 81}{523 + 477} = 0.2220$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.2696 - 0.1698}{\sqrt{0.222(1-0.222)\left(\frac{1}{523} + \frac{1}{477}\right)}} = 3.79$$

p-value very close to 0

Reject H_0. There is a significant difference in the population proportions. A higher proportion expecting to fly in 2003 is indicated.

44. a. H_0: $\pi_1 - \pi_2 \leq 0$

H_1: $\pi_1 - \pi_2 > 0$

b. $p_1 = 704/1035 = 0.6802$

$p_2 = 582/1004 = 0.5797$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{704 + 582}{1035 + 1004} = 0.6307$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.6802 - 0.579}{\sqrt{0.6307(1-0.6307)\left(\frac{1}{1035} + \frac{1}{1004}\right)}} = 4.70$$

p-value very close to 0

c. p-value < 0.01, reject H_0. Conclude that the proportion indicating good/excellent increased over the four-year period.

45. $p_1 = 0.69$, $n_1 = 400$ $p_2 = 0.57$, $n_2 = 700$

$$p_1 - p_2 \pm z_{0.025}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$0.69 - 0.57 \pm 1.96\sqrt{\frac{0.69(1-0.69)}{400} + \frac{0.57(1-0.57)}{700}}$$

0.12 ± 0.0583 (0.0617 to 0.1783)

The view was held by a higher proportion of respondents in Jordan than in Lebanon, the sample difference being 12 percentage points. The confidence interval shows the difference in the populations may be from 6 to 18 percentage points, at the 95% confidence level.

46. $p_1 = 9/142 = 0.0634$

$p_2 = 5/268 = 0.0187$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{9+5}{142+268} = 0.0341$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.0634 - 0.0187}{\sqrt{0.0341(1-0.0341)\left(\frac{1}{142} + \frac{1}{268}\right)}} = 2.37$$

p-value = 2(1 – 0.9911) = 0.0178

p-value < 0.02, reject H_0. There is a significant difference in drug resistance between the two cities. City X has the higher drug resistance rate.

47. H_0: $\pi_1 - \pi_2 = 0$

H_1: $\pi_1 - \pi_2 \neq 0$

$p_1 = 0.74, \quad n_1 = 1103$

$p_2 = 0.66, \quad n_2 = 1065$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{1519}{2168} = 0.700$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.74 - 0.66}{\sqrt{0.70(1-0.70)\left(\frac{1}{1103} + \frac{1}{1065}\right)}} = 4.06$$

p-value < 0.002.

p-value < 0.05, reject H_0. There is a difference between the proportions of students agreeing with the statement (higher proportion in 2001).

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Eleven

Inferences about Population Variances

Textbook Exercises (1-22)

Textbook Exercise Solutions

Supplementary Exercises (23-42)

Supplementary Exercise Solutions

Chapter 11: Inferences about Population Variances

Textbook Exercises:

1 Find the following chi-squared distribution values from Table 3 of Appendix B.

a. $\chi^2_{0.05}$ with df = 5 b. $\chi^2_{0.025}$ with df = 15 c. $\chi^2_{0.975}$ with df = 20
d. $\chi^2_{0.01}$ with df = 10 e. $\chi^2_{0.95}$ with df = 18

2 A sample of 20 items provides a sample standard deviation of 5.

a. Compute a 90 per cent confidence interval estimate of the population variance.
b. Compute a 95 per cent confidence interval estimate of the population variance.
c. Compute a 95 per cent confidence interval estimate of the population standard deviation.

3 A sample of 16 items provides a sample standard deviation of 9.5. Test the following hypotheses using $\alpha = 0.05$. What is your conclusion? Use both the *p*-value approach and the critical value approach.

$H_0: \sigma^2 \leq 50$
$H_1: \sigma^2 > 50$

4 The variance in drug weights is critical in the pharmaceutical industry. For a specific drug, with weights measured in grams, a sample of 18 units provided a sample variance of $s^2 = 0.36$.

a. Construct a 90 per cent confidence interval estimate of the population variance for the weight of this drug.
b. Construct a 90 per cent confidence interval estimate of the population standard deviation.

5 The table below shows estimated P/E ratios for December 2012, for a sample of eight companies listed on the Tel Aviv stock exchange (Source: Bloomberg, July 2012).

Company	P/E ratio
Avner Oil Exploration	37.69
Bank Hapoalim BM	6.59
Cellcom Israel Ltd	5.30
Delek Group Ltd	14.53
Nice Systems Ltd	14.46
Partner Communications Co Ltd	5.09
Paz Oil Co Ltd	16.13
Teva Pharmaceutical	7.29

a. Compute the sample variance and sample standard deviation for these data.
b. What is the 95 per cent confidence interval for the population variance?

6 Because of staffing decisions, managers of the Worldview Hotel are interested in the variability in the number of rooms occupied per day during a particular season of the year. A sample of 20 days of operation shows a sample mean of 290 rooms occupied per day and a sample standard deviation of 30 rooms.
a. What is the point estimate of the population variance?
b. Provide a 90 per cent confidence interval estimate of the population variance.
c. Provide a 90 per cent confidence interval estimate of the population standard deviation.

7 The CAC 40 is a share index based on the price movements of shares quoted on the Paris stock exchange. The figures below are the quarterly percentage returns for a tracker fund linked to the CAC 40, over the period Jan 2007 to June 2012.

	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2007	6.27	-3.51	1.68	-16.73
2008	2.60	−12.09	−20.61	-14.72
2009	6.25	8.43	5.29	3.65
2010	2.07	-4.55	5.23	4.49
2011	2.53	-10.57	-11.71	1.72
2012	-2.60	-1.12		

a. Compute the mean, variance, and standard deviation for the quarterly returns.
b. Financial analysts often use standard deviation of percentage returns as a measure of risk for stocks and mutual funds. Construct a 95 per cent confidence interval for the population standard deviation of quarterly returns for the CAC 40 tracker fund.

8 In the file 'Travel' on the online platform, there are estimated daily living costs (in euros) for a businessman travelling to 20 major cities. The estimates include a single room at a four star hotel, beverages, breakfast, taxi fares, and incidental costs.
a. Compute the sample mean.
b. Compute the sample standard deviation.
c. Compute a 95 per cent confidence interval for the population standard deviation.

City	Daily living cost	City	Daily living cost
Bangkok	242.87	Madrid	283.56
Bogota	260.93	Mexico City	212.00
Bombay	139.16	Milan	284.08
Cairo	194.19	Paris	436.72
Dublin	260.76	Rio de Janeiro	240.87
Frankfurt	355.36	Seoul	310.41
Hong Kong	346.32	Tel Aviv	223.73
Johannesburg	165.37	Toronto	181.25
Lima	250.08	Warsaw	238.20
London	326.76	Washington DC	250.61

9 Gold Fields Ltd is a South African mining company quoted on several stock exchanges, including NASDAQ Dubai. To analyze the risk, or volatility, associated with investing in Gold Fields Ltd shares, a sample of the monthly percentage return for 12 months was taken using the NASDAQ prices. The returns for the last six months of 2011 and the first six months of are shown here.

Month (2012)	Return (%)	Month (2011)	Return (%)
January	-5.26	July	10.25
February	-6.45	August	6.24
March	− 9.66	September	0.72
April	−9.66	October	−8.15
May	-7.06	November	12.13
June	-9.03	December	−0.58

a. Compute the sample variance and sample standard deviation monthly return for Gold Fields, as measures of volatility.
b. Construct a 95 per cent confidence interval for the population variance.
c. Construct a 95 per cent confidence interval for the population standard deviation.

10 Part variability is critical in the manufacturing of ball bearings. Large variances in the size of the ball bearings cause bearing failure and rapid wear. Production standards call for a maximum variance of 0.0025 when the bearing sizes are measured in millimetres. A sample of 15 bearings shows a sample standard deviation of 0.066 mm.

a. Use $\alpha = 0.10$ to determine whether the sample indicates that the maximum acceptable variance is being exceeded.
b. Compute a 90 per cent confidence interval estimate for the variance of the ball bearings in the population.

11 Suppose that any investment with an annualized standard deviation of percentage returns greater than 20% is classified as ‘high-risk’. The annualized standard deviation of percentage returns for the MSCI Emerging Markets index, based on a sample of size 36, is 25.2 per cent. Construct a hypothesis test that can be used to determine whether an investment based on the movements in the MSCI index would be classified as ‘high-risk’ With a 0.05 level of significance, what is your conclusion?

12 A sample standard deviation for the number of passengers taking a particular airline flight is 8. A 95 per cent confidence interval estimate of the population standard deviation is 5.86 passengers to 12.62 passengers.

a. Was a sample size of 10 or 15 used in the statistical analysis?

b. Suppose the sample standard deviation of $s = 8$ was based on a sample of 25 flights. What change would you expect in the confidence interval for the population standard deviation? Compute a 95 per cent confidence interval estimate of σ with a sample size of 25.

13 Find the following F distribution values from Table 4 of Appendix B.

a. $F_{0.05}$ with degrees of freedom 5 and 10

b. $F_{0.025}$ with degrees of freedom 20 and 15

c. $F_{0.01}$ with degrees of freedom 8 and 12

d. $F_{0.10}$ with degrees of freedom 10 and 20

14 A sample of 16 items from population 1 has a sample variance $s_1^2 = 5.8$ and a sample of 21 items from population 2 has a sample variance $s_2^2 = 2.4$. Test the following hypotheses at the 0.05 level of significance.

$$H_0: \sigma_1^2 \leq \sigma_2^2$$
$$H_1: \sigma_1^2 > \sigma_2^2$$

a. What is your conclusion using the p-value approach?

b. Repeat the test using the critical value approach.

15 Consider the following hypothesis test.

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$

a. What is your conclusion if $n_1 = 21$, $s_1^2 = 8.2$, $n2 = 26$, $s_2^2 = 4.0$? Use $\alpha = 0.05$ and the p-value approach.

b. Repeat the test using the critical value approach.

16 Most individuals are aware of the fact that the average annual repair cost for a car depends on its age. A researcher is interested in finding out whether the variance of the annual repair costs also increases with the age of the car. A sample of 26 cars that were eight years old showed a sample standard deviation for annual repair costs of £170 and a sample of 25 cars that were four years old showed a sample standard deviation for annual repair costs of £100.

a. State the null and alternative hypotheses if the research hypothesis is that the variance in annual repair costs is larger for the older cars.

b. At a 0.01 level of significance, what is your conclusion? What is the *p*-value? Discuss the reasonableness of your findings.

17 On the basis of data provided by a salary survey, the variance in annual salaries for seniors in accounting firms is approximately 2.1 and the variance in annual salaries for managers in accounting firms is approximately 11.1. The salary data were provided in thousands of Euros. Assuming that the salary data were based on samples of 25 seniors and 26 managers, test the hypothesis that the population variances in the salaries are equal. At a 0.05 level of significance, what is your conclusion?

18 For a sample of 100 days in 2012, the euro to US dollars and the British pound to euro exchange rates were recorded. The sample means were 1.2852 US$/€ and 1.2294 €/£. The respective sample standard deviations were 0.03565 US$/€ and 0.02290 €/£. Do a hypothesis test to determine whether there is a difference in variability between the two exchange rates. Use □ = 0.05 as the level of significance. Discuss briefly whether the comparison you have made a 'fair' one?

19 Two new assembly methods are tested and the variances in assembly times are reported. Use $\alpha = 0.10$ and test for equality of the two population variances.

	Method A	Method B
Sample size	$n_1 = 31$	$n_2 = 25$
Sample variation	$s_1^2 = 25$	$s_2^2 = 12$

20 A research hypothesis is that the variance of stopping distances of cars on wet roads is greater than the variance of stopping distances of cars on dry roads. In the research study, 16 cars travelling at the same speeds are tested for stopping distances on wet roads and then tested for stopping distances on dry roads. On wet roads, the standard deviation of stopping distances is ten metres. On dry roads, the standard deviation is five metres.

a. At a 0.05 level of significance, do the sample data justify the conclusion that the variance in stopping distances on wet roads is greater than the variance in stopping distances on dry roads? What is the p-value?
b. What are the implications of your statistical conclusions in terms of driving safety recommendations?

21 The grade point averages of 352 students who completed a college course in financial accounting have a standard deviation of 0.940. The grade point averages of 73 students who dropped out of the same course have a standard deviation of 0.797. Do the data indicate a difference between the variances of grade point averages for students who completed a financial accounting course and students who dropped out? Use a 0.05 level of significance.
Note: $F_{0.025}$ with 351 and 72 degrees of freedom is 1.466.

22 The variance in a production process is an important measure of the quality of the process. A large variance often signals an opportunity for improvement in the process by finding ways to reduce the process variance. The file 'Bags' on the online platform contains data for two machines that fill bags with powder. The file has 25 bag weights for Machine 1 and 22 bag weights for Machine 2. Conduct a statistical test to determine whether there is a significant difference between the variances in the bag weights for the two machines. Use a 0.05 level of significance. What is your conclusion? Which machine, if either, provides the greater opportunity for quality improvements?

Chapter 11: Inferences about Population Variances

Textbook Exercises Solutions:

1. a. 11.070 b. 27.488 c. 9.591

   d. 23.209 e. 9.390

2. $s^2 = 25$

   a. With 19 degrees of freedom $\chi^2_{0.05} = 30.144$ and $\chi^2_{0.95} = 10.117$

   $$\frac{19(25)}{30.144} \le \sigma^2 \le \frac{19(25)}{10.117}$$

   $$15.76 \le \sigma^2 \le 46.95$$

   b. With 19 degrees of freedom $\chi^2_{0.025} = 32.852$ and $\chi^2_{0.975} = 8.907$

   $$\frac{19(25)}{32.852} \le \sigma^2 \le \frac{19(25)}{8.907}$$

   $$14.46 \le \sigma^2 \le 53.33$$

   c. $3.8 \le \sigma \le 7.3$

3. $$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(16-1)(9.5)^2}{50} = 27.08$$

   Degrees of freedom = (16 – 1) = 15

   p-value approach:

   Using χ^2 table, *p*-value is between 0.025 and 0.05. (Actual *p*-value = 0.0281.)

   p-value $\le$ 0.05, reject H_0

   Critical value approach:

   $\chi^2_{0.05} = 24.996$

   Reject H_0 if $\chi^2 \ge 24.996$

   27.08 > 24.996, reject H_0

4. a. $n = 18$

$s^2 = 0.36$

$\chi^2_{0.05} = 27.587$ and $\chi^2_{0.95} = 8.672$ (17 degrees of freedom)

$$\frac{17(.36)}{27.587} \le \sigma^2 \le \frac{17(.36)}{8.672}$$

$$0.22 \le \sigma^2 \le 0.71$$

b. $0.47 \le \sigma \le 0.84$

5. a. $\bar{x} = 13.39$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{818.23}{7} = 116.89$$

$$s = \sqrt{116.89} = 10.81$$

b. $\chi^2_{0.025} = 16.013$ $\chi^2_{0.975} = 1.690$

$$\frac{(8-1)(116.89)}{16.013} \le \sigma^2 \le \frac{(8-1)(116.89)}{1.690}$$

$$51.10 \le \sigma^2 \le 484.16$$

c. $7.14 \le \sigma \le 22.00$

6. a. $s^2 = (30)^2 = 900$

b. $\chi^2_{0.05} = 30.144$ and $\chi^2_{0.95} = 10.117$ (19 degrees of freedom)

$$\frac{(19)(900)}{30.144} \le \sigma^2 \le \frac{(19)(900)}{10.117}$$

$$567 \le \sigma^2 \le 1690$$

c. $23.8 \le \sigma \le 41.1$

7. a. Using Excel, MINITAB or SPSS:

$\bar{x}$ = - 2.18%, s^2 = 72.10 (% points)^2, s = 8.49 (% points)

b. $\chi^2_{0.025}$ = 35.479 and $\chi^2_{0.975}$ = 10.283 (21 degrees of freedom)

$$\sqrt{\frac{(21)(72.10)}{35.479}} \le \sigma \le \sqrt{\frac{(21)(72.10)}{10.283}}$$

$6.53 \le \sigma \le 12.13$

8. a. $\bar{x} = \frac{\Sigma x_i}{n} = 260.16$

b. $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = 4996.8$

$s = \sqrt{4996.79} = 70.69$

c. $\chi^2_{0.025}$ = 32.852 and $\chi^2_{0.975}$ = 8.907 (19 degrees of freedom)

$$\frac{(20-1)(4996.8)}{32.852} \le \sigma^2 \le \frac{(20-1)(4996.8)}{8.907}$$

$2890 \le \sigma^2 \le 10{,}659$

$53.76 \le \sigma \le 103.24$

9. a. Using Excel, MINITAB or SPSS:

s^2 = 62.47 (% points)^2, s = 7.90 (% points)

b. $\chi^2_{0.025}$ = 21.920 and $\chi^2_{0.975}$ = 3.816 (11 degrees of freedom)

$$\frac{(11)(62.47)}{21.92} \le \sigma^2 \le \frac{(11)(62.47)}{3.816}$$

$31.35 \le \sigma^2 \le 180.10$

c. $5.60 \le \sigma \le 13.42$

10. a. H_0: $\sigma^2 \le 0.0025$
H_1: $\sigma^2 > 0.0025$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(15-1)(0.066)^2}{0.0025} = 24.39$$

Degrees of freedom = $n - 1 = 14$

Using χ^2 table, p- value is between 0.025 and 0.05. (Actual p-value = 0.041.)

p-value $\le$ 0.10, reject H_0. Conclude that variance exceeds maximum variance requirement.

b. $\chi^2_{0.05} = 23.685$ and $\chi^2_{0.95} = 6.571$ (14 degrees of freedom)

$$\frac{(14)(0.066)^2}{23.685} \le \sigma^2 \le \frac{(14)(0.066)^2}{6.571}$$

$$0.00257 \le \sigma^2 \le 0.00928$$

11. H_0: s^2 £ 20
H_1: $s^2 > 20$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(36-1)(25.2)^2}{(20)^2} = 55.57$$

Degrees of freedom = $n - 1 = 35$

Using χ^2 table, area in tail (one-tailed p-value) is greater than 0.01 but less than 0.025 (p-value calculated using Excel, MINITAB or SPSS is 0.015).
p-value < 0.05, reject H_0 and conclude that the investment would be classified as 'high-risk'.

12. a. Try n = 15

$\chi^2_{0.025} = 26.119$ and $\chi^2_{0.975} = 5.629$ (14 degrees of freedom)

$$\frac{(14)(64)}{26.119} \le \sigma^2 \le \frac{(14)(64)}{5.629}$$

$$34.3 \le \sigma^2 \le 159.2$$

$$5.86 \le \sigma \le 12.62$$

Therefore a sample size of 15 was used.

b. $n = 25$; expect the width of the interval to be smaller.

$\chi^2_{0.05} = 39.364$ and $\chi^2_{0.975} = 12.401$ (24 degrees of freedom)

$$\frac{(24)(8)^2}{39.364} \le \sigma^2 \le \frac{(24)(8)^2}{12.401}$$

$$39.02 \le \sigma^2 \le 126.86$$

$$6.25 \le \sigma \le 11.13$$

13. a. $F_{0.05} = 3.33$

b. $F_{0.025} = 2.76$

c. $F_{0.01} = 4.50$

d. $F_{0.10} = 1.94$

14. a. $F = \frac{s_1^2}{s_2^2} = \frac{5.8}{2.4} = 2.4$

Degrees of freedom 15 and 20

Using F table, p-value is between 0.025 and 0.05. (Actual p-value = 0.0334.)

p-value $\le$ 0.05, reject H_0. Conclude $\sigma_1^2 > \sigma_2^2$

b. $F_{0.05} = 2.20$

Reject H_0 if $F \ge 2.20$

$2.4 \ge 2.20$, reject H_0. Conclude $\sigma_1^2 > \sigma_2^2$

15. a. Larger sample variance is s_1^2

$$F = \frac{s_1^2}{s_2^2} = \frac{8.2}{4} = 2.05$$

Degrees of freedom 20 and 25

Using F table, area in tail is between 0.025 and 0.05

Two-tailed p-value is between 0.05 and 0.10. (Actual p-value = 0.0904.)

p-value > 0.05, do not reject H_0.

b. Since we have a two-tailed test

$F_{\alpha/2} = F_{0.025} = 2.30$

Reject H_0 if $F \geq 2.30$

$2.05 < 2.30$, do not reject H_0

16. a. Population 1 is 4-year-old automobiles

$H_0: \sigma_1^2 \leq \sigma_2^2$
$H_1: \sigma_1^2 > \sigma_2^2$

b. $F = \dfrac{s_1^2}{s_2^2} = \dfrac{170^2}{100^2} = 2.89$

Degrees of freedom 25 and 24

Using F table, p-value is less than 0.01. (Actual p-value = 0.0057.)

p-value $\leq$ 0.01, reject H_0. Conclude that 4-year-old cars have a larger variance in annual repair costs compared to 2-year-old cars. This is expected due to the fact that older cars are more likely to have more expensive repairs that lead to greater variance in the annual repair costs.

17. $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

$F = \dfrac{s_1^2}{s_2^2} = \dfrac{11.1}{2.1} = 5.29$

Degrees of freedom 25 and 24

Using F table, area in tail is less than 0.01

Two-tailed p-value is less than 0.02. (Actual p-value is very close to 0.)

p-value $\leq$ 0.05, reject H_0. The population variances are not equal for seniors and managers.

18. $F = (0.03565/0.02290)^2 = 2.42$, degrees of freedom 99 and 99

Two-tailed p-value is less than 0.02. (Exact p-value calculated using Excel, Minitab or SPSS is 0.000008)

p-value < 0.05, reject H_0. Conclude that the US$ to € exchange rate was more variable than the € to £ exchange rate.

Question of 'fairness' arises because the US$ to € exchange rate was on average higher than the € to £ exchange rate, so one might expect the variability to be correspondingly higher.

19. $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} = \frac{25}{12} = 2.08$$

Degrees of freedom 30 and 24

Using F table, area in tail is between 0.025 and 0.05

Two-tailed p-value is between 0.05 and 0.10. (Actual p-value = 0.0689.)

p-value $\leq$ 0.10, reject H_0. Conclude that the population variances are not equal.

20. a. Population 1 is wet roads.

$H_0: \sigma_1^2 \leq \sigma_2^2$
$H_1: \sigma_1^2 > \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} = \frac{32^2}{16^2} = 4.00$$

Degrees of freedom 15 and 15

Using F table, p-value is less than 0.01. (Actual p-value = 0.0054.)

p-value $\leq$ 0.05, reject H_0. Conclude that there is greater variability in stopping distances on wet roads.

b. Drive carefully on wet roads because of the uncertainty in stopping distances.

21. $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

Population 1 (students who completed) has the larger sample variance.

$$F = \frac{s_1^2}{s_2^2} = \frac{(0.940)^2}{(0.797)^2} = 1.391$$

Degrees of freedom 351 and 72, $F_{0.025} = 1.466$

Because $1.391 < F_{0.025} = 1.466$, two-tailed p-value) is greater than 0.05 (p-value calculated using Excel, MINITAB or PASW is 0.090).
p-value > 0.05, do not reject H_0. There is no convincing evidence at $\alpha = 0.05$ of a difference between the variances of the grade point averages for those who completed and those who dropped out.

22. $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

$s_1^2 = 0.0489, \quad s_2^2 = 0.0059$

$F = \frac{s_1^2}{s_2^2} = \frac{0.0489}{0.0059} = 8.28$, degrees of freedom 24 and 21

Using F table, area in tail is less than 0.01

Two-tailed p-value is less than 0.02. (Actual p-value very close to 0.)

p-value $\leq$ 0.05, reject H_0. The process variances are significantly different. Machine 1 offers the greater opportunity for process quality improvements.

Note that the sample means are similar with the mean bag weights of approximately 3.3 grams. However, the process variances are significantly different.

Chapter 11: Inferences about Population Variances

Supplementary Exercises:

23. Find the following chi-squared distribution values from Table 3 of Appendix B.

    a. $\chi^2_{0.05}$ with df = 8

    b. $\chi^2_{0.025}$ with df = 35

    c. $\chi^2_{0.975}$ with df = 19

    d. $\chi^2_{0.01}$ with df = 27

    e. $\chi^2_{0.95}$ with df = 50

24. A sample of 30 items provides a sample standard deviation of 10.0.

    a. Compute a 90% confidence interval estimate of the population variance.

    b. Compute a 95% confidence interval estimate of the population variance.

    c. Compute a 95% confidence interval estimate of the population standard deviation.

25. A sample of 25 items provides a sample variance of 150. Test the following hypotheses using $\alpha = 0.05$. What is your conclusion? Use both the *p*-value approach and the critical value approach.

$$H_0 : \sigma \leq 10$$
$$H_1 : \sigma > 10$$

26. The daily car rental rates for a sample of eight cities follow.

City	Daily Car Rental Rate (€)
A	47
B	50
C	53
D	45
E	40
F	43
G	39
H	37

    a. Compute the variance and the standard deviation for these data.

    b. What is the 95% confidence interval estimate of the variance of car rental rates for the population?

    c. What is the 95% confidence interval estimate of the standard deviation for the population?

27. The following data show the annual percentage returns of an investment fund for each of five consecutive years:

10.8 34.2 4.2 9.4 29.4

The standard deviation of the annual returns can be used as a measure of risk, with a larger standard deviation indicating more variation and therefore more uncertainty in the annual returns.

a. Treat the data as a sample of five annual returns and determine the sample standard deviation measure of risk for this fund.

b. What is the 95% confidence interval for the population standard deviation of annual returns for this fund?

28. A group of 12 security analysts provided estimates of the year 2001 earnings per share for Qualcomm, Inc. (*Zacks.com*, June 13, 2000). The data are as follows:

1.40 1.40 1.45 1.49 1.37 1.27 1.40 1.55 1.40 1.42 1.48 1.63

a. Compute the sample variance for the earnings per share estimate.

b. Compute the sample standard deviation for the earnings per share estimate.

c. Provide 95% confidence interval estimates of the population variance and the population standard deviation.

29. The table below shows return-on-equity (ROE) figures for 2007, for a sample of six companies listed on the Tel Aviv stock exchange (Source: Datastream, Thomson Financial).

Company	ROE (%)
Bezeq	25.83
Clal Industries	25.35
Harel Insurance	22.60
Koor Industries	28.72
Mizrahi Bank	17.10
Strauss Group	15.40

a. Compute the sample variance and sample standard deviation for these data.

b. What is the 95 per cent confidence interval for the population variance?

c. What is the 95 per cent confidence interval for the population standard deviation?

30. The filling variance for boxes of breakfast cereal is designed to be 2.0 or less. A sample of 41 boxes of cereal shows a sample standard deviation of 1.6 grams. Use $\alpha = 0.05$ to determine whether the variance in the cereal box fillings is exceeding the design specification.

31. The Fidelity Growth & Income mutual fund received a three-star, or neutral, rating from Morningstar. Shown here are the quarterly percentage returns for the five-year period from 2001 to 2005 (Morningstar Funds 500, 2006).

	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2001	−10.91	5.80	−9.64	6.45
2002	0.83	−10.48	−14.03	5.58
2003	−2.27	10.43	0.85	9.33
2004	1.34	1.11	−0.77	8.03
2005	−2.46	0.89	2.55	1.78

a. Compute the mean, variance, and standard deviation for the quarterly returns.

b. Financial analysts often use standard deviation of percentage returns as a measure of risk for stocks and mutual funds. Construct a 95 per cent confidence interval for the population standard deviation of quarterly returns for the Fidelity Growth & Income mutual fund.

32. City Trucking Company claims consistent delivery times for its routine customer deliveries. A sample of 22 truck deliveries shows a sample variance of 1.5. Test to determine whether H_0: $\sigma^2 \leq 1$ can be rejected. Use $\alpha = 0.10$.

33. To analyze the risk, or volatility, associated with investing in Chevron Corporation common stock, a sample of the monthly total percentage return for 12 months was taken. The returns for the 12 months of 2005 are shown here (Compustat, February 24, 2006). Total return is price appreciation plus any dividend paid.

Month	Return (%)	Month	Return (%)
January	3.60	July	3.74
February	14.86	August	6.62
March	−6.07	September	5.42
April	−10.82	October	−11.83
May	4.29	November	1.21
June	3.98	December	−0.94

a. Compute the sample variance and sample standard deviation as a measure of volatility of monthly total return for Chevron.

b. Construct a 95 per cent confidence interval for the population variance.

c. Construct a 95 per cent confidence interval for the population standard deviation.

34. The average standard deviation for the annual return of large cap stock mutual funds is 18.2 per cent (The Top Mutual Funds, AAII, 2004). The sample standard deviation based on a sample of size 36 for the Vanguard PRIMECAP mutual fund is 22.2 per cent. Construct a hypothesis test that can be used to determine whether the standard deviation for the Vanguard fund is greater than the average standard deviation for large cap mutual funds. With a 0.05 level of significance, what is your conclusion?

35. Find the following *F* distribution values from Table 4 of Appendix B.
    a. $F_{0.05}$ with degrees of freedom 10 and 12
    b. $F_{0.025}$ with degrees of freedom 15 and 10
    c. $F_{0.01}$ with degrees of freedom 9 and 8
    d. $F_{0.10}$ with degrees of freedom 30 and 40

36. A sample of 26 items from population 1 has a sample standard deviation $s_1 = 10.5$ and a sample of 21 items from population 2 has a sample standard deviation e $s_2 = 7.9$. Test the following hypotheses at the 0.05 level of significance.

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

    a. What is your conclusion using the *p*-value approach?
    b. Repeat the test using the critical value approach.

37. The standard deviation of the 12-month earnings per share for 10 companies in the airline industry was 4.27 and the standard deviation of the 12-month earnings per share for 7 companies in the automotive industry was 2.27 (*Business Week*, August 14, 2000). Conduct a test for equal variances at $\alpha = 0.05$. What is your conclusion about the variability in earnings per share for the airline industry and the automotive industry?

38. Fidelity Magellan is a large cap growth mutual fund and Fidelity Small Cap Stock is a small cap growth mutual fund (Morningstar Funds 500, 2006). The standard deviation for both funds was computed based on a sample of size 26. For Fidelity Magellan, the sample standard deviation is 8.89 per cent; for Fidelity Small Cap Stock, the sample standard deviation is 13.03 per cent. Financial analysts often use the standard deviation as a measure of risk. Conduct a hypothesis test to determine whether the small cap growth fund is riskier than the large cap growth fund. Use $\alpha = 0.05$ as the level of significance.

39. The Dow Jones Industrial Average varies as investors buy and sell shares of the 30 stocks that make up the average. Samples of the Dow Jones Industrial Average taken at different times during the first five days of November the first five days of December in the same year are as follows.

November	7493	7525	7760	7499	7555	7690	7668	7600	7516	7711
December	8066	8209	7842	7943	7846	8071	8055	8159	7828	8109

a. Compute the variances of the Dow Jones Industrial Average for the two time periods.

b. Use a 0.05 level of significance and test to determine whether the population variances for the two time periods are equal. What is the *p*-value? What is your conclusion?

40. Each day the major stock markets have a group of leading gainers in price (stocks that go up the most). On one day the standard deviation in the percentage change for a sample of 10 NASDAQ leading gainers was 15.8. On the same day, the standard deviation in the percentage change for a sample of 10 NYSE leading gainers was 7.9 (*USA Today*, September 14, 2000). Conduct a test for equal population variances to see whether it can be concluded that there is a difference in the volatility of the leading gainers on the two exchanges. Use $\alpha = 0.10$. What is your conclusion?

41. The examination scores of 352 students who completed a college course in financial accounting have a standard deviation of 9.40. The examination scores of 73 students who dropped out of the same course have a standard deviation of 7.97. Do the data indicate a difference between the variances of examination scores for students who completed a financial accounting course and students who dropped out? Use a 0.05 level of significance. *Note:* $F_{0.025}$ with 351 and 72 degrees of freedom is 1.466.

42. Suppose the average useful life of video recording machines is 6 years with a standard deviation of 0.75 year. A sample of the useful life of 30 television sets provided a sample standard deviation of 0.95 year. Construct a hypothesis test that can be used to determine whether the standard deviation of the useful life of television sets is significantly greater than the standard deviation of the useful life of video recording machines. With a 0.05 level of significance, what is your conclusion?

Chapter 11: Inferences about Population Variances

Supplementary Exercises Solutions:

23. a. 15.507 b. 53.203
    c. 8.907 d. 46.963
    e. 34.764

24. a. $\chi^2_{0.05} = 42.557$ $\chi^2_{0.95} = 17.708$

$$\frac{(30-1)(10.0)^2}{42.557} \le \sigma^2 \le \frac{(30-1)(10.0)^2}{17.708}$$

$$68.1 \le \sigma^2 \le 163.8$$

b. $\chi^2_{0.025} = 45.722$ $\chi^2_{0.975} = 16.047$

$$\frac{(30-1)(10.0)^2}{45.722} \le \sigma^2 \le \frac{(30-1)(10.0)^2}{16.047}$$

$$63.4 \le \sigma^2 \le 180.7$$

c. $7.96 \le \sigma \le 13.44$

25. H_0: $\sigma \le 10$
    H_1: $\sigma > 10$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(25-1)(150)}{(10)^2} = 36.0$$

Degrees of freedom = $n - 1 = 24$

<u>*p*-value approach</u>

Using χ^2 table, *p*-value is between 0.10 and 0.05. (Actual *p*-value = 0.0549)

p-value > 0.05, do not reject H_0.

<u>Critical value approach</u>

$\chi^2_{0.05} = 36.415$

Reject H_0 if $\chi^2 \ge 36.415$

36.0 < 36.415, do not reject H_0

26. a. $s^2 = \dfrac{\Sigma(x_i - \bar{x})^2}{n-1} = 31.07$

$s = \sqrt{31.07} = 5.57$

b. $\chi^2_{0.025} = 16.013 \qquad \chi^2_{0.975} = 1.690$

$$\frac{(8-1)(31.07)}{16.013} \le \sigma^2 \le \frac{(8-1)(31.07)}{1.690}$$

$13.6 \le \sigma^2 \le 128.7$

c. $3.7 \le \sigma \le 11.3$

27. a. $s^2 = \dfrac{\Sigma(x_i - \bar{x})^2}{n-1} = 176.96$

$s = \sqrt{176.96} = 13.30$

b. $\chi^2_{0.025} = 11.143 \qquad \chi^2_{0.975} = 0.484$

$$\frac{(5-1)(176.96)}{11.143} \le \sigma^2 \le \frac{(5-1)(176.96)}{.484}$$

$63.5 \le \sigma^2 \le 1462.5$

$8.0 \le \sigma \le 38.2$

28. a. $s^2 = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\dfrac{0.0929}{12-1}} = 0.00845$

b. $s = \sqrt{0.00845} = 0.092$

c. 11 degrees of freedom

$\chi^2_{0.025} = 21.920 \qquad \chi^2_{0.975} = 3.816$

$$\frac{(12-1)0.00845}{21.920} \le \sigma^2 \le \frac{(12-1)0.00845}{3.816}$$

$0.0042 \le \sigma^2 \le 0.0244$

$0.065 \le \sigma \le 0.156$

29. a. $\bar{x} = 22.50$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{137.48}{5} = 27.50$$

$$s = \sqrt{27.50} = 5.24$$

b. $\chi^2_{0.025} = 12.832$ $\chi^2_{0.975} = 0.831$

$$\frac{(6-1)(27.50)}{12.832} \le \sigma^2 \le \frac{(6-1)(27.50)}{0.831}$$

$$10.72 \le \sigma^2 \le 165.46$$

c. $3.27 \le \sigma \le 12.86$

30. H_0: $\sigma^2 \le 2.0$
H_1: $\sigma^2 > 2.0$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(41-1)(1.6)^2}{2.0} = 51.20$$

Degrees of freedom = $n - 1 = 40$

Using χ^2 table, p-value is greater than 0.10. (Actual p-value = 0.1104)

p-value > 0.05, do not reject H_0. The population variance does not appear to be exceeding the standard.

31. a. Using Excel, MINITAB or SPSS:

$\bar{x} = 0.221\%$, $s^2 = 47.95$ (% points)^2, $s = 6.92$ (% points)

b. $\chi^2_{0.025} = 32.852$ and $\chi^2_{0.975} = 8.907$ (19 degrees of freedom)

$$\sqrt{\frac{(19)(47.95)}{32.852}} \le \sigma \le \sqrt{\frac{(19)(47.95)}{8.907}}$$

$$5.27 \le \sigma \le 10.11$$

32. H_0: $\sigma^2 \le 1$
H_1: $\sigma^2 > 1$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(22-1)(1.5)}{1} = 31.50$$

Degrees of freedom = $n - 1 = 21$
Using χ^2 table, p-value is between 0.05 and 0.10. (Actual p-value = 0.0657)
p-value < 0.10, reject H_0. Conclude that $\sigma^2 > 1$.

33. a. Using Excel, MINITAB or SPSS:

$s^2 = 57.72$ (% points)^2, $s = 7.60$ (% points)

b. $\chi^2_{0.025} = 21.920$ and $\chi^2_{0.975} = 3.816$ (11 degrees of freedom)

$$\frac{(11)(57.72)}{21.92} \le \sigma^2 \le \frac{(11)(57.72)}{3.816}$$

$$28.97 \le \sigma^2 \le 166.38$$

c. $5.38 \le \sigma \le 12.90$

34. H_0: s^2 £ 18.2
H_1: s^2 > 18.2

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(36-1)(22.2)^2}{(18.2)^2} = 52.08$$

Degrees of freedom = $n - 1 = 35$

Using χ^2 table, area in tail (one-tailed p-value) is greater than 0.025 but less than 0.05 (p-value calculated using Excel, MINITAB or SPSS is 0.032).
p-value < 0.05, reject H_0 and conclude that the standard deviation for the Vanguard fund is higher than the average standard deviation for large cap funds.

35. a. 2.75 b. 3.52
c. 5.91 d. 1.54

36. a. $F = \frac{s_1^2}{s_2^2} = \frac{(10.5)^2}{(7.9)^2} = 1.767$

Degrees of freedom 25 and 20

Using F table, p-value is between 0.10 and 0.05. (Actual p-value = 0.0987.)

p-value > 0.05, do not reject H_0.

b. $F_{0.05} = 2.07$

Reject H_0 if $F \ge 2.07$

1.767 < 2.07, do not reject H_0.

37. $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} = \frac{4.27^2}{2.27^2} = 3.54$$

Degrees of freedom 9 and 6

Using F table, area in tail is between 0.05 and 0.10

Two-tailed p-value is between 0.10 and 0.20. (Actual p-value = 0.1379)

p-value > 0.05, do not reject H_0. Cannot conclude any difference between variances of the two industries.

38. $H_0: \sigma_1^2 \leq \sigma_2^2$
$H_1: \sigma_1^2 > \sigma_2^2$

Population 1 (Fidelity Small Cap Stock) has the larger sample variance.

$$F = \frac{s_1^2}{s_2^2} = \frac{(13.03)^2}{(8.89)^2} = 2.15$$

Degrees of freedom 25 and 25

Using F table, area in tail (one-tailed p-value) is greater than 0.025 but less than 0.05 (p-value calculated using Excel, MINITAB or PASW is 0.031).
p-value < 0.05, reject H_0. Conclude that the small cap growth fund is riskier than the large cap growth fund.

39. a. $$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

$s_{\text{Nov}}^2 = 9664$

$s_{\text{Dec}}^2 = 19{,}238$ (Population 1 since s^2 is larger)

b. $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} = \frac{19{,}238}{9664} = 1.99$$

Degrees of freedom 9 and 9

Using F table, area in tail is greater than 0.10

Two-tailed p-value is greater than 0.20. (Actual p-value = 0.3197)
p-value > 0.05, do not reject H_0. No convincing evidence that the population variances differ.

40. $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

Population 1 is NASDAQ

$$F = \frac{s_1^2}{s_2^2} = \frac{15.8^2}{7.9^2} = 4.00$$

Degrees of freedom 9 and 9

Using F table, area in tail is between 0.025 and 0.05

Two-tailed p-value is between 0.05 and 0.10. (Actual p-value = 0.0510)

p-value > 0.05, do not reject H_0. Cannot conclude that the population variances differ. But with a p-value so close to 0.05, a larger sample is recommended.

41. $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

Using critical value approach, with $F_{0.025} = 1.47$

Reject H_0 if $F \geq 1.47$

$$F = \frac{s_1^2}{s_2^2} = \frac{0.940^2}{0.797^2} = 1.39$$

$F < 1.47$, do not reject H_0. We are not able to conclude that students who complete the course and students who drop out have different variances of grade point averages.

42. $\sigma^2 = (0.75)^2 = 0.5625$

H_0: $\sigma^2 \leq 0.5625$
H_1: $\sigma^2 > 0.5625$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(30-1)(0.95)^2}{0.5625} = 46.53$$

Degrees of freedom = $n - 1 = 29$

Using χ^2 table, p-value is between 0.01 and 0.025. (Actual p-value = 0.0208)

p-value $\leq$ 0.05, reject H_0. The standard deviation for television sets is greater than the standard deviation for VCRs.

Statistics for Business & Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Twelve

Tests of Goodness of Fit and Independence

Textbook Exercises (1-22)

Textbook Exercise Solutions

Supplementary Exercises (23-37)

Supplementary Exercise Solutions

Chapter 12: Tests of Goodness of Fit and Independence

Textbook Exercises:

1 Test the following hypotheses by using the χ^2 goodness of fit test.

H_0: $\pi_A = 0.40$, $\pi_B = 0.40$, $\pi_C = 0.20$

H_1: The population proportions are not $\pi_A = 0.40$, $\pi_B = 0.40$, $\pi_C = 0.20$

A sample of size 200 yielded 60 in category A, 120 in category B, and 20 in category C. Use $\alpha = 0.01$ and test to see whether the proportions are as stated in H_0.

a. Use the *p*-value approach.

b. Repeat the test using the critical value approach.

2 Suppose we have a multinomial population with four categories: A, B, C and D. The null hypothesis is that the proportion of items is the same in every category, i.e. H_0: $\pi_A = \pi_B = \pi_C = \pi_D = 0.25$. A sample of size 300 yielded the following results.

A: 85 B: 95 C: 50 D: 70

Use $\alpha = 0.05$ to determine whether H_0 should be rejected. What is the *p*-value?

3 One of the questions on the Business Week Subscriber Study was, 'When making investment purchases, do you use full service or discount brokerage firms?' Survey results showed that 264 respondents use full service brokerage firms only, 255 use discount brokerage firms only and 229 use both full service and discount firms. Use $\alpha = 0.10$ to determine whether there are any differences in preference among the three service choices.

4 How well do airline companies serve their customers? A study by Business Week showed the following customer ratings: 3 per cent excellent, 28 per cent good, 45 per cent fair and 24 per cent poor. In a follow-up study of service by telephone companies, assume that a sample of 400 adults found the following customer ratings: 24 excellent, 124 good, 172 fair and 80 poor. Taking the figures from the Business Week study as 'population' values, is the distribution of the customer ratings for telephone companies different from the distribution of customer ratings for airline companies? Test with $\alpha = 0.01$. What is your conclusion?

5 In setting sales quotas, the marketing manager of a multinational company makes the assumption that order potentials are the same for each of four sales territories in the Middle East. A sample of 200 sales follows. Should the manager's assumption be rejected? Use α = 0.05.

Sales territories			
1	2	3	4
60	45	59	36

6 A community park will open soon in a large European city. A sample of 210 individuals are asked to state their preference for when they would most like to visit the park. The sample results follow.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
20	30	30	25	35	20	50

In developing a staffing plan, should the park manager plan on the same number of individuals visiting the park each day? Support your conclusion with a statistical test. Use α = 0.05.

7 The results of ComputerWorld's Annual Job Satisfaction Survey showed that 28 per cent of information systems (IS) managers are very satisfied with their job, 46 per cent are somewhat satisfied, 12 per cent are neither satisfied or dissatisfied, 10 per cent are somewhat dissatisfied and 4 per cent are very dissatisfied. Suppose that a sample of 500 computer programmers yielded the following results.

Category	Number of respondents
Very satisfied	105
Somewhat satisfied	235
Neither	55
Somewhat dissatisfied	90
Very dissatisfied	15

Taking the ComputerWorld figures as 'population' values, use α = 0.05 and test to determine whether the job satisfaction for computer programmers is different from the job satisfaction for IS managers.

8 The following 2 × 3 contingency table contains observed frequencies for a sample of 200. Test for independence of the row and column variables using the χ^2 test with α = 0.05.

	Column variable		
Row variable	A	B	C
P	20	44	50
Q	30	26	30

9 The following 3 × 3 contingency table contains observed frequencies for a sample of 240. Test for independence of the row and column variables using the χ^2 test with α = 0.05.

	Column variable		
Row variable	A	B	C
P	20	30	20
Q	30	60	25
R	10	15	30

10 One of the questions on the Business Week Subscriber Study was, 'In the past 12 months, when travelling for business, what type of airline ticket did you purchase most often?' The data obtained are shown in the following contingency table.

	Type of flight	
Type of ticket	Domestic flights	International flights
First class	29	22
Business class	95	121
Economy class	518	135

Use α = 0.05 and test for the independence of type of flight and type of ticket. What is your conclusion?

11 First-destination jobs for business and engineering graduates are classified by industry as shown in the following table.

	Industry			
Degree major	Oil	Chemical	Electrical	Computer
Business	30	15	15	40
Engineering	30	30	20	20

Use α = 0.01 and test for independence of degree major and industry type.

12 Businesses are increasingly placing orders online. The Performance Measurement Group collected data on the rates of correctly filled electronic orders by industry. Assume a sample of 700 electronic orders provided the following results.

	Industry			
Order	Pharmaceutical	Consumer	Computers	Telecommunications
Correct	207	136	151	178
Incorrect	3	4	9	12

a. Test whether order fulfilment is independent of industry. Use $\alpha = 0.05$. What is your conclusion?

b. Which industry has the highest percentage of correctly filled orders?

13 Three suppliers provide the following data on defective parts.

	Part quality		
Supplier	Good	Minor defect	Major defect
A	90	3	7
B	170	18	7
C	135	6	9

Use $\alpha = 0.05$ and test for independence between supplier and part quality. What does the result of your analysis tell the purchasing department?

14 A sample of parts taken in a machine shop in Karachi provided the following contingency table data on part quality by production shift.

Shift	Number good	Number defective
First	368	32
Second	285	15
Third	176	24

Use $\alpha = 0.05$ and test the hypothesis that part quality is independent of the production shift. What is your conclusion?

15 Visa Card studied how frequently consumers of various age groups use plastic cards (debit and credit cards) when making purchases. Sample data for 300 customers shows the use of plastic cards by four age groups.

	Age group			
Payment	18–24	25–34	35–44	45 and over
Plastic	21	27	27	36
Cash or Cheque	21	36	42	90

a. Test for the independence between method of payment and age group. What is the *p*-value? Using $\alpha = 0.05$, what is your conclusion?
b. If method of payment and age group are not independent, what observation can you make about how different age groups use plastic to make purchases?
c. What implications does this study have for companies such as Visa and MasterCard?

16 The following cross-tabulation shows industry type and P/E ratio for 100 companies in the consumer products and banking industries.

	P/E ratio					
Industry	5–9	10–14	15–19	20–24	25–29	Total
Consumer	4	10	18	10	8	50
Banking	14	14	12	6	4	50
Total	18	24	30	16	12	100

Does there appear to be a relationship between industry type and P/E ratio? Support your conclusion with a statistical test using $\alpha = 0.05$.

17 The following data are believed to have come from a normal distribution. Use a goodness of fit test and $\alpha = 0.05$ to test this claim.

17	23	22	24	19	23	18	22	20	13	11	21	18	20	21
21	18	15	24	23	23	43	29	27	26	30	28	33	23	29

18 Data on the number of occurrences per time period and observed frequencies follow. Use $\alpha = 0.05$ and a goodness of fit test to see whether the data fit a Poisson distribution.

Number of occurrences	Observed frequency
0	39
1	30
2	30
3	18
4	3

19 The number of incoming phone calls to a small call centre in Mumbai, during one minute intervals, is believed to have a Poisson distribution. Use $\alpha = 0.10$ and the following data to test the assumption that the incoming phone calls follow a Poisson distribution.

Number of incoming phone calls during a one-minute interval	Observed frequency
0	15
1	31
2	20
3	15
4	13
5	4
6	2
Total	100

20 The weekly demand for a particular product in a white-goods store is thought to be normally distributed. Use a goodness of fit test and the following data to test this assumption. Use $\alpha = 0.10$. The sample mean is 24.5 and the sample standard deviation is 3.0.

18	20	22	27	22	25	22	27	25	24
26	23	20	24	26	27	25	19	21	25
26	25	31	29	25	25	28	26	28	24

21 A random sample of final examination grades for a college course in Middle-East studies follows.

55	85	72	99	48	71	88	70	59	98
80	74	93	85	74	82	90	71	83	60
95	77	84	73	63	72	95	79	51	85
76	81	78	65	75	87	86	70	80	64

Use $\alpha = 0.05$ and test to determine whether a normal distribution should be rejected as being representative of the population's distribution of grades.

22 The number of car accidents per day in a particular city is believed to have a Poisson distribution. A sample of 80 days during the past year gives the following data. Do these data support the belief that the number of accidents per day has a Poisson distribution? Use $\alpha = 0.05$.

Number of accidents	Observed frequency (days)
0	34
1	25
2	11
3	7
4	3

Chapter 12: Tests of Goodness of Fit and Independence

Textbook Exercises Solutions:

1. a. Expected frequencies: $e_1 = 200(0.40) = 80$, $e_2 = 200(0.40) = 80$, $e_3 = 200(0.20) = 40$

Actual frequencies: $f_1 = 60, f_2 = 120, f_3 = 20$

$$\chi^2 = \frac{(60-80)^2}{80} + \frac{(120-80)^2}{80} + \frac{(20-40)^2}{40}$$
$$= \frac{400}{80} + \frac{1600}{80} + \frac{400}{40}$$
$$= 5 + 20 + 10$$
$$= 35$$

$k - 1 = 2$ degrees of freedom

Using chi-squared distribution with $df = 2$, $\chi^2 = 35$ shows p-value very close to 0

p-value < 0.01, reject H_0

b. $\chi^2_{0.01} = 9.210$

Reject H_0 if $\chi^2 \geq 9.210$

$\chi^2 = 35$, reject H_0

2. Expected frequencies: $e_1 = 300(0.25) = 75$, $e_2 = 300(0.25) = 75$

$e_3 = 300(0.25) = 75$, $e_4 = 300(0.25) = 75$

Actual frequencies: $f_1 = 85, f_2 = 95, f_3 = 50, f_4 = 70$

$$\chi^2 = \frac{(85-75)^2}{75} + \frac{(95-75)^2}{75} + \frac{(50-75)^2}{75} + \frac{(70-75)^2}{75}$$
$$= \frac{100}{75} + \frac{400}{75} + \frac{625}{75} + \frac{25}{75}$$
$$= \frac{1150}{75}$$
$$= 15.33$$

$k - 1 = 3$ degrees of freedom

Chi-squared table shows p-value less than 0.005. (Actual p-value = 0.0016.)

p-value < 0.05, reject H_0, conclude that the population proportions are not the same.

3.

Category	Hypothesized Proportion	Observed Frequency (f_i)	Expected Frequency (e_i)	$(f_i - e_i)^2 / e_i$
Full Service	1/3	264	249.33	0.86
Discount	1/3	255	249.33	0.13
Both	1/3	229	249.33	1.66
	Totals:	748		2.65

$k - 1 = 2$ degrees of freedom

Using χ^2 table, χ^2 shows p-value > 0.10. (Actual p-value = 0.2658.)

p-value > 0.10, do not reject H_0. There is no significant difference in preference among the three services.

4. H_0: $\pi_1 = 0.03$, $\pi_2 = 0.28$, $\pi_3 = 0.45$, $\pi_4 = 0.24$

Rating	Observed	Expected	$(f_i - e_i)^2 / e_i$
Excellent	24	0.03(400) = 12	12.00
Good	124	0.28(400) = 112	1.29
Fair	172	0.45(400) = 180	0.36
Poor	80	0.24(400) = 96	2.67
	400	400	$\chi^2 = 16.31$

Degrees of freedom = $k - 1 = 3$

Using χ^2 table, $\chi^2 = 16.31$ shows p-value < 0.005. (Actual p-value = 0.001.)

p-value < 0.01, reject H_0. Conclude that the telephone companies' ratings differ from those of the airline companies. A comparison of observed and expected frequencies shows telephone service is better, with more excellent and good ratings.

5.

Observed	60	45	59	36
Expected	50	50	50	50

$\chi^2 = 8.04$

Degrees of freedom = 4 – 1 = 3

Using χ^2 table, $\chi^2 = 8.04$ shows *p*-value is between 0.025 and 0.05. (Actual *p*-value = 0.0452.)

p-value < 0.05, reject H_0. Conclude that the order potentials are not the same in each sales territory.

6.

Observed	20	30	30	25	35	20	50
Expected	30	30	30	30	30	30	30

$$\chi^2 = \frac{(20-30)^2}{30} + \frac{(30-30)^2}{30} + \frac{(30-30)^2}{30} + \frac{(25-30)^2}{30} + \frac{(35-30)^2}{30} + \frac{(20-30)^2}{30} + \frac{(50-30)^2}{30} = 21.7$$

Degrees of freedom = 7 – 1 = 6

Using χ^2 table, $\chi^2 = 21.7$ shows *p*-value is less than 0.005.

p-value < 0.05, reject H_0. The park manager should not plan on the same number attending each day. Plan on a larger staff for Sundays.

7.

Category	Hypothesized Proportion	Observed Frequency (f_i)	Expected Frequency (e_i)	$(f_i - e_i)^2 / e_i$
Very Satisfied	0.28	105	140	8.75
Somewhat Satisfied	0.46	235	230	0.11
Neither	0.12	55	60	0.42
Somewhat Dissatisfied	0.10	90	50	32.00
Very Dissatisfied	0.04	15	20	1.25
	Totals:	500		42.53

Degrees of freedom = 5 – 1 = 4

Using χ^2 table, $\chi^2 = 42.53$ shows p-value is very close to 0.

p-value < 0.05, reject H_0. Conclude that the job satisfaction for computer programmers is different from the job satisfaction for IS managers.

8. H_0 = The column variable is independent of the row variable
H_1 = The column variable is not independent of the row variable

Expected Frequencies:

	A	B	C
P	28.5	39.9	45.6
Q	21.5	30.1	34.4

$$\chi^2 = \frac{(20-28.5)^2}{28.5} + \frac{(44-39.9)^2}{39.9} + \frac{(50-45.6)^2}{45.6} + \frac{(30-21.5)^2}{21.5} + \frac{(26-30.1)^2}{30.1} + \frac{(30-34.4)^2}{34.4}$$
$$= 7.86$$

Degrees of freedom = (2 – 1)(3 – 1) = 2

Using χ^2 table, $\chi^2 = 7.86$ provides a p-value between 0.01 and 0.025. (Actual p-value = 0.0196.)

p-value < 0.05, reject H_0. Conclude that the column variable is not independent of the row variable.

9. H_0 = The column variable is independent of the row variable

H_1 = The column variable is not independent of the row variable

Expected Frequencies:

	A	B	C
P	17.5000	30.6250	21.8750
Q	28.7500	50.3125	35.9375
R	13.7500	24.0625	17.1875

$$\chi^2 = \frac{(20-17.5000)^2}{17.5000} + \frac{(30-30.6250)^2}{30.6250} + \cdots + \frac{(30-17.1875)^2}{17.1875}$$
$$= 19.77$$

Degrees of freedom = (3 – 1)(3 – 1) = 4

Using χ^2 table, $\chi^2 = 19.77$ shows a p-value less than 0.005. (Actual p-value = 0.0006.)

p-value < 0.05, reject H_0. Conclude that the column variable is not independent of the row variable.

10. H_0 : Type of ticket purchased is independent of the type of flight

H_1: Type of ticket purchased is not independent of the type of flight.

Expected Frequencies:

e_{11} = 35.59 e_{12} = 15.41

e_{21} = 150.73 e_{22} = 65.27

e_{31} = 455.68 e_{32} = 197.32

Ticket	Flight	Observed Frequency (f_i)	Expected Frequency (e_i)	$(f_i - e_i)^2 / e_i$
First	Domestic	29	35.59	1.22
First	International	22	15.41	2.82
Business	Domestic	95	150.73	20.61
Business	International	121	65.27	47.59
Full Fare	Domestic	518	455.68	8.52
Full Fare	International	135	197.32	19.68
	Totals:	920		100.43

Degrees of freedom = (3 – 1)(2 – 1) = 2

Using χ^2 table, $\chi^2 = 100.43$ shows a *p*-value very close to 0.

p-value $\leq$ 0.05, reject H_0. Conclude that the type of ticket purchased is not independent of the type of flight.

11.

	Industry			
Major	Oil	Chemical	Electrical	Computer
Business	30	22.5	17.5	30
Engineering	30	22.5	17.5	30

Note: Values shown above are the expected frequencies.

$\chi^2 = 12.38$

Degrees of freedom = (2 – 1)(4 – 1) = 3

Using χ^2 table, $\chi^2 = 12.38$ shows a *p*-value between 0.005 and 0.01. (Actual *p*-value = 0.0062.)

p-value < 0.01, reject H_0. Conclude that degree major and industry are not independent.

12. a. Observed Frequency (f_{ij})

	Pharm	Consumer	Computer	Telecom	Total
Correct	207	136	151	178	672
Incorrect	3	4	9	12	28
Total	210	140	160	190	700

Expected Frequency (e_{ij})

	Pharm	Consumer	Computer	Telecom	Total
Correct	201.6	134.4	153.6	182.4	672
Incorrect	8.4	5.6	6.4	7.6	28
Total	210	140	160	190	700

Chi-squared $(f_{ij} - e_{ij})^2 / e_{ij}$

	Pharm	Consumer	Computer	Telecom	Total
Correct	0.14	0.02	0.04	0.11	.31
Incorrect	3.47	0.46	1.06	2.55	7.53
					$\chi^2 = 7.85$

Degrees of freedom = (2 – 1)(4 – 1) = 3

Using χ^2 table, $\chi^2 = 7.85$ shows p-value is between 0.025 and 0.05. (Actual p-value = 0.0493.)

p-value < 0.05, reject H_0. Conclude that fulfilment of orders is not independent of industry.

b. The pharmaceutical industry is doing the best with 207 of 210 (98.6%) correctly filled orders.

13. Expected Frequencies:

	Part Quality		
Supplier	Good	Minor Defect	Major Defect
A	88.76	6.07	5.17
B	173.09	11.83	10.08
C	133.15	9.10	7.75

$\chi^2 = 7.71$

Degrees of freedom = $(3 - 1)(3 - 1) = 4$

Using χ^2 table, $\chi^2 = 7.71$ shows p-value is between 0.05 and 0.10. (Actual p-value = 0.1027.)

p-value > 0.05, do not reject H_0. Conclude that the assumption of independence cannot be rejected.

14. Expected Frequencies:

	Quality	
Shift	Good	Defective
1st	368.44	31.56
2nd	276.33	23.67
3rd	184.22	15.78

$\chi^2 = 8.10$

Degrees of freedom = $(3 - 1)(2 - 1) = 2$

Using χ^2 table, $\chi^2 = 8.10$ shows p-value is between 0.01 and 0.025. (Actual p-value = 0.0174.)

p-value < 0.05, reject H_0. Conclude that shift and quality are not independent.

15. a. Expected frequencies:

Payment	18 - 24	25 - 34	35 – 44	45 +
Plastic	15.54	23.31	25.53	46.62
Cash or Cheque	26.46	39.69	43.47	79.38

$\chi^2 = 7.95$

Degrees of freedom = (4 – 1)(2 – 1) = 3

Using χ^2 table, $\chi^2 = 7.95$ shows *p*-value is between 0.025 and 0.05 (Actual *p*-value = 0.047.)

p-value < 0.05, reject H_0, conclude that method of payment and age group are not independent.

b. The figures suggest that the tendency to use plastic is lower the higher the age group. (The percentages are 50%, 43%, 39%, 29% for the four age groups.)

c. Credit and debit card companies might target their marketing to try and increase usage amongst the older age groups.

16. Expected Frequencies:

$$e_{11} = \frac{(50)(18)}{100} = 9, \quad e_{12} = \frac{(50)(24)}{100} = 12, \ \cdots, e_{25} = \frac{(50)(12)}{100} = 6$$

$$\chi^2 = \frac{(4-9)^2}{9} + \frac{(10-12)^2}{12} + \cdots + \frac{(4-6)^2}{6} = 9.76$$

Degrees of freedom = (2 – 1)(5 – 1) = 4

Using χ^2 table, $\chi^2 = 9.76$ shows *p*-value is between 0.025 and 0.05. (Actual *p*-value = 0.0448.)

p-value < 0.05, reject H_0. Banking tends to have lower P/E ratios. We can conclude that industry type and P/E ratio are related.

17. With $n = 30$ we will use six classes, each with the probability of 0.1667.

$\bar{x} = 22.8, \quad s = 6.27$

The z values that create 6 intervals, each with probability 0.1667 are –0.98, –0.43, 0, 0.43, 0.98

z	Cut-off value of x
–0.98	22.8 – 0.98 (6.27) = 16.66
–0.43	22.8 – 0.43 (6.27) = 20.11
0	22.8 + 0 (6.27) = 22.80
0.43	22.8 + 0.43 (6.27) = 25.49
0.98	22.8 + 0.98 (6.27) = 28.94

Interval	Observed Frequency	Expected Frequency	Difference
less than 16.66	3	5	–2
16.66 – 20.11	7	5	2
20.11 – 22.80	5	5	0
22.80 – 25.49	7	5	2
25.49– 28.94	3	5	–2
28.94 and up	5	5	0

$$\chi^2 = \frac{(-2)^2}{5} + \frac{(2)^2}{5} + \frac{(0)^2}{5} + \frac{(2)^2}{5} + \frac{(-2)^2}{5} + \frac{(0)^2}{5} + \frac{16}{5} = 3.20$$

Degrees of freedom = 6 – 2 – 1 = 3

Using χ^2 table, $\chi^2 = 3.20$ shows p-value is greater than 0.10. (Actual p-value = 0.3618.)

p-value > 0.05, do not reject H_0. The claim that the data comes from a normal distribution cannot be rejected.

18. First estimate μ from the sample data. Sample size = 120.

$$\bar{x} = \frac{0(39) + 1(30) + 2(30) + 3(18) + 4(3)}{120} = \frac{156}{120} = 1.3$$

Therefore, we use Poisson probabilities with $\mu = 1.3$ to compute expected frequencies.

x	Observed Frequency	Poisson Probability	Expected Frequency	Difference $(f_i - e_i)$
0	39	0.2725	32.70	6.30
1	30	0.3543	42.51	–12.51
2	30	0.2303	27.63	2.37
3	18	0.0998	11.98	6.02
4	3	0.0431	5.16	– 2.17

$$\chi^2 = \frac{(6.30)^2}{32.70} + \frac{(-12.51)^2}{42.51} + \frac{(2.37)^2}{27.63} + \frac{(6.02)^2}{11.98} + \frac{(-2.17)^2}{5.16} = 9.04$$

Degrees of freedom = 5 – 1 – 1 = 3

Using χ^2 table, $\chi^2 = 9.04$ shows p-value is between 0.025 and 0.05. (Actual p-value = 0.0287.)

p-value < 0.05, reject H_0. Conclude that the data do not follow a Poisson probability distribution.

19. $\bar{x} = \dfrac{0(15)+1(31)+2(20)+3(15)+4(13)+5(4)+6(2)}{100} = 2$

x	Observed	Poisson Probabilities	Expected
0	15	0.1353	13.53
1	31	0.2707	27.07
2	20	0.2707	27.07
3	15	0.1804	18.04
4	13	0.0902	9.02
5 or more	6	0.0527	5.27

$\chi^2 = 4.95$

Degrees of freedom = 6 – 1 – 1 = 4

Using χ^2 table, $\chi^2 = 4.95$ shows p-value is greater than 0.10. (Actual p-value = 0.2929.)

p-value > 0.10, do not reject H_0. The assumption of a Poisson distribution cannot be rejected.

20. $\bar{x} = 24.5$, $s = 3$, $n = 30$. Use 6 classes.

Interval	Observed Frequency	Expected Frequency
less than 21.56	5	5
21.56 – 23.21	4	5
23.21 – 24.50	3	5
24.50 – 25.79	7	5
25.79 – 27.44	7	5
27.41 upwards	4	5

$\chi^2 = 2.8$

Degrees of freedom = 6 – 2 – 1 = 3

Using χ^2 table, $\chi^2 = 2.8$ shows p-value is greater than 0.10. (Actual p-value = 0.4235.)

p-value > 0.10, do not reject H_0. The assumption of a normal distribution cannot be rejected.

21. $\bar{x}$ = 76.83, s = 12.43, n = 40. Use 8 classes.

Interval	Observed Frequency	Expected Frequency
less than 62.54	5	5
62.54 - 68.50	3	5
68.50 - 72.85	6	5
72.85 - 76.83	5	5
76.83 - 80.81	5	5
80.81 - 85.16	7	5
85.16 - 91.12	4	5
91.12 upwards	5	5

$\chi^2 = 2$

Degrees of freedom = 8 – 2 – 1 = 5

Using χ^2 table, $\chi^2 = 2$ shows p-value is greater than 0.10. (Actual p-value = 0.8491.)

p-value > 0.05, do not reject H_0. The assumption of a normal distribution cannot be rejected.

22. $\bar{x} = \dfrac{0(34)+1(25)+2(11)+3(7)+4(3)}{80} = 1.0$

x	Observed	Poisson Probabilities	Expected
0	34	0.3679	29.43
1	25	0.3679	29.43
2	11	0.1839	14.72
3 or more	7	0.0804	6.42

$\chi^2 = 4.30$

Degrees of freedom = 4 – 1 – 1 = 2

Using χ^2 table, $\chi^2 = 4.95$ shows p-value is greater than 0.10. (Actual p-value = 0.116.)

p-value > 0.10, do not reject H_0. The assumption of a Poisson distribution cannot be rejected.

Supplementary Exercises:

23. During the first 13 weeks of the autumn schedules, the Saturday evening 8:00 p.m. to 9:00 p.m. audience proportions were recorded as: BBC1 & 2, 29%; ITV and C4, 28%; Sky channels, 25%; and others, 18%. A sample of 300 homes two weeks after a Saturday night schedule revision yielded the following viewing audience data: BBC1 & 2, 95 homes; ITV and C4, 70 homes; Sky channels, 89 homes; and others, 46 homes. Test with $\alpha = 0.05$ to determine whether the viewing audience proportions changed.

24. Negative appeal is recognized as an effective method of persuasion in advertising. A study in *The Journal of Advertising* reported the results of a content analysis of guilt advertisements in six different types of magazine. An equal number of advertisements were examined in each of the magazine types. The number of ads with guilt appeals follow.

Magazine type	Number of ads with guilt appeals
News and opinion	20
General editorial	15
Family-oriented	30
Business/financial	22
Female-oriented	16
African-American	12

Using $\alpha = 0.10$, test to see whether the proportion of ads with guilt appeals differs among the six types of magazines.

25. Seven percent of mutual fund investors rate corporate stocks "very safe," 58% rate them "somewhat safe," 24% rate them "not very safe," 4% rate them "not at all safe," and 7% are "not sure." A *Business Week*/Harris poll asked 529 mutual fund investors how they would rate corporate bonds on safety. The responses are as follows.

Safety rating	Frequency
Very safe	48
Somewhat safe	323
Not very safe	79
Not at all safe	16
Not sure	63
Total	529

Do mutual fund investors' attitudes toward corporate bonds differ from their attitudes toward corporate stocks? Support your conclusion with a statistical test. Use $\alpha = 0.01$.

26. *The Wall Street Journal* Shareholder Scoreboard tracks the performance of 1000 major U.S. companies. The performance of each company is rated based on the annual total return, including stock price changes and the reinvestment of dividends. Ratings are assigned by dividing all 1000 companies into five groups from A (top 20%), B (next 20%), to E (bottom 20%). Shown here are the one-year ratings for a sample of 60 of the largest companies. Do the largest companies differ in performance from the performance of the 1000 companies in the Shareholder Scoreboard? Use $\alpha = 0.05$.

A	B	C	D	E
5	8	15	20	12

27. A public transport company is concerned about the number of passengers on one of its minibus routes. In setting up the route, the assumption is that the number of riders is the same on every day from Monday to Friday. Using the following data, test with $\alpha = 0.05$ to determine whether the company's assumption is correct.

Day	Number of passengers
Monday	13
Tuesday	16
Wednesday	28
Thursday	17
Friday	16

28. M&M/MARS, makers of M&M® sweets, conducted a national poll in which more than 10 million people indicated their preference for a new colour. The tally of this poll resulted in the replacement of tan-coloured M&Ms with a new blue colour. In a brochure produced by M&M/MARS Consumer Affairs, the distribution of colours for the sweets is as follows:

Brown	Yellow	Red	Orange	Green	Blue
30%	20%	20%	10%	10%	10%

In a study reported in *Chance* (no. 4, 1996), samples of bags were used to determine whether the reported percentages were indeed valid. The following results were obtained for one sample of 506 sweets.

Brown	Yellow	Red	Orange	Green	Blue
177	135	79	41	36	38

Use $\alpha = 0.05$ to determine whether these data support the percentages reported by the company.

29. In a study of brand loyalty in the U.S. car industry, new-car customers were asked whether the make of their new car was the same as the make of their previous car (*Business Week,* May 8, 2000). The breakdown of 600 responses shows the brand loyalty for U.S., European, and Asian cars.

	Manufacturer:		
Bought:	**U.S.**	**European**	**Asian**
Same make	125	55	68
Different make	140	105	107

a. Formulate and test a hypothesis to determine whether brand loyalty is independent of the manufacturer. Use $\alpha = 0.05$. What is your conclusion?
b. If a significant difference is found, which group of manufacturers appears to have the greatest brand loyalty?

30. Negative appeal is recognized as an effective method of persuasion in advertising. A study in *The Journal of Advertising* reported the results of a content analysis of guilt and fear advertisements six different types of magazine. An equal number of advertisements were examined in each of the magazine types. The number of ads with guilt and fear appeals that appeared in selected magazine types follows.

Magazine type	Number of ads with guilt appeals	Number of ads with fear appeals
News and opinion	20	1
General editorial	15	11
Family-oriented	30	19
Business/financial	22	17
Female-oriented	16	14
African-American	12	15

Use the chi-squared test of independence with a 0.01 level of significance to analyze the data. What is your conclusion?

31. A study of educational levels of voters and their political party affiliations yielded the following results.

	Party affiliation:		
Educational level:	**Democratic**	**Republican**	**Independent**
Did not complete high school	40	20	10
High school degree	30	35	15
College degree	30	45	25

Use $\alpha = 0.01$ and determine whether party affiliation is independent of the educational level of the voters.

32. A business magazine subscriber study collected data on the employment status of subscribers. Sample results relating to subscribers of the print and online editions are shown here.

Employment status	Print edition	Online edition
Full-time	1105	574
Part-time	31	15
Self-employed	229	186
Not employed	486	344

Use $\alpha = 0.05$ and test the hypothesis that employment status is independent of the edition. What is your conclusion?

33. Data on the marital status of men and women ages 20 to 29 were obtained as part of a national survey. The results from a sample of 350 men and 400 women follow.

	Marital status:		
Gender:	**Never married**	**Married**	**Divorced**
Men	234	106	10
Women	216	168	16

a. Use $\alpha = 0.01$ and test for independence between marital status and gender. What is your conclusion?

b. Calculate the percentages in each marital status category for men and for women.

34. The following data were collected on the number of emergency ambulance calls for an urban county and a rural county.

		Day of week:							
		Sun	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Total**
County:	**Urban**	61	48	50	55	63	73	43	393
	Rural	7	9	16	13	9	14	10	78
	Total	68	57	66	68	72	87	53	471

Conduct a test for independence using $\alpha = 0.05$. What is your conclusion?

35. A study conducted by Marist Institute for Public Opinion asked men and women to indicate which person is the most difficult to buy holiday gifts for (*USA Today*, December 15, 1997). Suppose that the following data were obtained in a follow-up study consisting of 100 men and 100 women.

		Gender:	
		Men	**Women**
Most Difficult to Buy For:	**Spouse**	37	25
	Parents	28	31
	Children	7	19
	Siblings	8	3
	In-laws	4	10
	Other relatives	16	12

Use $\alpha = 0.05$ and test for independence of gender and the most difficult person to buy for. What is your conclusion?

36. The number of car accidents per day in a particular city is believed to have a Poisson distribution. A sample of 80 days during the past year gives the following data. Do these data support the belief that the number of accidents per day has a Poisson distribution? Use $\alpha = 0.05$.

Number of accidents	**Observed frequency (days)**
0	34
1	25
2	11
3	7
4	3

37. Use $\alpha = 0.01$ and conduct a goodness of fit test to see whether the following sample appears to have been selected from a normal distribution.

55	86	94	58	55	95	55	52	69	95	90	65	87
50	56	55	57	98	58	79	92	62	59	88	65	

After you complete the goodness of fit calculations, construct a histogram of the data. Does the histogram representation support the conclusion reached with the goodness of fit test? (*Note:* $\bar{x} = 71$ and $s = 17$.)

Chapter 12: Tests of Goodness of Fit and Independence

Supplementary Exercises Solutions:

23. H_0 : $p_{BBC} = 0.29, p_{ITV} = 0.28, p_{SKY} = 0.25, p_{OTH} = 0.18$

H_1 : The proportions are not $p_{BBC} = 0.29, p_{ITV} = 0.28, p_{SKY} = 0.25, p_{OTH} = 0.18$

Expected frequencies: 300 (0.29) = 87, 300 (0.28) = 84

300 (0.25) = 75, 300 (0.18) = 54

$e_1 = 87, e_2 = 84, e_3 = 75, e_4 = 54$

Actual frequencies: $f_1 = 95, f_2 = 70, f_3 = 89, f_4 = 46$

$$\chi^2 = \frac{(95-87)^2}{87} + \frac{(70-84)^2}{84} + \frac{(89-75)^2}{75} + \frac{(46-54)^2}{54}$$
$$= 6.87$$

$k - 1 = 3$ degrees of freedom

Using χ^2 table, p-value is between 0.05 and 0.10. (Actual p-value = 0.0762)

p-value > 0.05, do not reject H_0. There has not been a significant change in the viewing audience proportions.

24.

Category	Hypothesized Proportion	Observed Frequency (f_i)	Expected Frequency (e_i)	$(f_i - e_i)^2 / e_i$
News and Opinion	1/6	20	19.17	0.04
General Editorial	1/6	15	19.17	0.91
Family Oriented	1/6	30	19.17	6.12
Business/Financial	1/6	22	19.17	0.42
Female Oriented	1/6	16	19.17	0.52
African-American	1/6	12	19.17	2.68
	Totals:	115		10.69

$k - 1 = 5$ degrees of freedom

Using χ^2 table, $\chi^2 = 10.69$ shows p-value is between 0.05 and 0.10

(Actual p-value = 0.0580)

p-value < 0.10, reject H_0. Conclude that there is a difference in the proportion of ads with guilt appeals among the six types of magazines.

25.

Observed	48	323	79	16	63
Expected	37.03	306.82	126.96	21.16	37.03

$$\chi^2 = \frac{(48-37.03)^2}{37.03} + \frac{(323-306.82)^2}{306.82} + \cdots + \frac{(63-37.03)^2}{37.03} = 41.69$$

Degrees of freedom = 5 – 1 = 4

Using χ^2 table, $\chi^2 = 41.69$ shows p-value very close to 0

p-value < 0.01, reject H_0. Mutual fund investors' attitudes toward corporate bonds differ from their attitudes toward corporate stock.

26. Expected frequencies: 20% each, $n = 60$

$$e_1 = 12,\ e_2 = 12,\ e_3 = 12,\ e_4 = 12,\ e_5 = 12$$

Actual frequencies: $f_1 = 5, f_2 = 8, f_3 = 15, f_4 = 20, f_5 = 12$

$$\chi^2 = \frac{(5-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(15-12)^2}{12} + \frac{(20-12)^2}{12} + \frac{(12-12)^2}{12}$$
$$= 11.50$$

$k – 1 = 4$ degrees of freedom

Using χ^2 table, p-value is between 0.025 and 0.01

(Actual p-value = 0.0215)

Reject H_0. Yes, the largest companies differ in performance from the 1000 companies. In general, the largest companies did not do as well as others. 15 of 60 companies (25%) are in the middle group and 20 of 60 companies (33%) are in the next lower group. These are both greater than the 20% expected. Relative few large companies are in the top A and B categories.

(Note that this result is for the year 2002. This should not be generalized to other years without additional data.)

27.

Observed	13	16	28	17	16
Expected	18	18	18	18	18

$\chi^2 = 7.44$ Degrees of freedom = 5 – 1 = 4

Using χ^2 table, $\chi^2 = 7.44$ shows p-value is greater than 0.10. (Actual p-value = 0.1142)

p-value > 0.05, do not reject H_0. The assumption that the number of riders is uniformly distributed cannot be rejected.

28.

Category	Hypothesized Proportion	Observed Frequency (f_i)	Expected Frequency (e_i)	$(f_i - e_i)^2 / e_i$
Brown	0.30	177	151.8	4.18
Yellow	0.20	135	101.2	11.29
Red	0.20	79	101.2	4.87
Orange	0.10	41	50.6	1.82
Green	0.10	36	50.6	4.21
Blue	0.10	38	50.6	3.14
	Totals:	506		29.51

$k - 1 = 5$ degrees of freedom

Using χ^2 table, $\chi^2 = 29.51$ shows p-value very close to zero 0

p-value < 0.05, reject H_0. Conclude that the percentages reported by the company have changed.

29. a. Observed Frequency (f_{ij})

	Domestic	European	Asian	Total
Same	125	55	68	248
Different	140	105	107	352
Total	265	160	175	600

Expected Frequency (e_{ij})

	Domestic	European	Asian	Total
Same	109.53	66.13	72.33	248
Different	155.47	93.87	102.67	352
Total	265	160	175	600

Chi squared $(f_{ij} - e_{ij})^2 / e_{ij}$

	Domestic	European	Asian	Total
Same	2.18	1.87	0.26	4.32
Different	1.54	1.32	0.18	3.04
				$\chi^2 = 7.36$

Degrees of freedom = (3 – 1)(2 – 1) = 2

Using χ^2 table, $\chi^2 = 7.36$ shows *p*-value is between 0.025 and 0.05

(Actual *p*-value = 0.0252)

p-value < 0.05, reject H_0. Conclude that brand loyalty is not independent of manufacturer.

b. Brand Loyalty

Domestic 125/265 = 0.472 (47.2%) ← Highest
European 55/160 = 0.344 (34.4%)
Asian 68/175 = 0.389 (38.9%)

30.

Magazine	Appeal	Observed Frequency (f_{ij})	Expected Frequency (e_{ij})	$(f_{ij} - e_{ij})^2 / e_{ij}$
News	Guilt	20	17.16	0.47
News	Fear	10	12.84	0.63
General	Guilt	15	14.88	0.00
General	Fear	11	11.12	0.00
Family	Guilt	30	28.03	0.14
Family	Fear	19	20.97	0.18
Business	Guilt	22	22.31	0.00
Business	Fear	17	16.69	0.01
Female	Guilt	16	17.16	0.08
Female	Fear	14	12.84	0.11
African-American	Guilt	12	15.45	0.77
African-American	Fear	15	11.55	1.03
	Totals:	201		3.41

Degrees of freedom = (6 – 1)(2 – 1) = 5

Using χ^2 table, $\chi^2 = 3.41$ shows p-value is greater than 0.10. (Actual p-value = 0.6366)

p-value > 0.01, do not reject H_0. The hypothesis of independence cannot be rejected.

31. Expected Frequencies:

	Party Affiliation		
Education Level	Democratic	Republican	Independent
Did not complete high school	28	28	14
High school degree	32	32	16
College degree	40	40	20

$\chi^2 = 13.42$ Degrees of freedom = (3 – 1)(3 – 1) = 4

Using χ^2 table, $\chi^2 = 13.42$ shows p-value is between 0.005 and 0.01. (Actual p-value = 0.0094)

p-value < 0.01, reject H_0. Conclude that party affiliation is not independent of education level.

32.

Employment	Region	Observed Frequency (f_{ij})	Expected Frequency (e_{ij})	$(f_{ij} - e_{ij})^2 / e_{ij}$
Full-Time	Eastern	1105	1046.19	3.31
Full-time	Western	574	632.81	5.46
Part-Time	Eastern	31	28.66	0.19
Part-Time	Western	15	17.34	0.32
Self-Employed	Eastern	229	258.59	3.39
Self-Employed	Western	186	156.41	5.60
Not Employed	Eastern	485	516.55	1.93
Not Employed	Western	344	312.45	3.19
	Totals:	2969		23.37

Degrees of freedom = (4 – 1)(2 – 1) = 3

Using χ^2 table, $\chi^2 = 23.37$ shows p-value is very close to 0

p-value < 0.05, reject H_0. Conclude that employment status is not independent of region.

33. a. Observed Frequency (f_{ij})

	Never Married	Married	Divorced	Total
Men	234	106	10	350
Women	216	168	16	400
Total	450	274	26	750

Expected Frequency (e_{ij})

	Never Married	Married	Divorced	Total
Men	210	127.87	12.13	350
Women	240	146.13	13.87	400
Total	450	274	26	750

Chi squared $(f_{ij} - e_{ij})^2 / e_{ij}$

	Never Married	Married	Divorced	Total
Men	2.74	3.74	0.38	6.86
Women	2.40	3.27	0.33	6.00
				$\chi^2 = 12.86$

Degrees of freedom = (2 – 1)(3 – 1) = 2

Using χ^2 table, $\chi^2 = 12.86$ shows p-value is less than 0.005. (Actual p-value = 0.0016)

p-value < 0.01, reject H_0. Conclude that marital status is not independent of gender.

b. Marital Status

	Never Married	Married	Divorced
Men	66.9%	30.3%	2.9%
Women	54.0%	42.0%	4.0%

Men 100 – 66.9 = 33.1% have been married
Women 100 – 54.0 = 46.0% have been married

34. Expected Frequencies:

	Days of the Week							
County	Sun	Mon	Tue	Wed	Thur	Fri	Sat	Total
Urban	56.7	47.6	55.1	56.7	60.1	72.6	44.2	393
Rural	11.3	9.4	10.9	11.3	11.9	14.4	8.8	78
Total	**68**	**57**	**66**	**68**	**72**	**87**	**53**	**471**

$\chi^2 = 6.17$ Degrees of freedom = (2 – 1)(7 – 1) = 6

Using χ^2 table, $\chi^2 = 6.17$ shows p-value is greater than 0.10. (Actual p-value = 0.4039)

p-value > 0.05, do not reject H_0. The assumption of independence cannot be rejected.

35.

Most Difficult	Gender	Observed Frequency (f_{ij})	Expected Frequency (e_{ij})	$(f_{ij} - e_{ij})^2 / e_{ij}$
Spouse	Men	37	31.0	1.16
Spouse	Women	25	31.0	1.16
Parents	Men	28	29.5	0.08
Parents	Women	31	29.5	0.08
Children	Men	7	13.0	2.77
Children	Women	19	13.0	2.77
Siblings	Men	8	5.5	1.14
Siblings	Women	3	5.5	1.14
In-Laws	Men	4	7.0	1.29
In-Laws	Women	10	7.0	1.29
Other Relatives	Men	16	14.0	0.29
Other Relatives	Women	12	14.0	0.29
	Totals:	200		13.43

Degrees of freedom = (6 – 1)(2 – 1) = 5

Using χ^2 table, $\chi^2 = 13.43$ shows p-value is between 0.01 and 0.025. (Actual p-value = 0.0197)

p-value < 0.05, reject H_0. Conclude that the most difficult person to buy for is not independent of gender.

36. $\bar{x} = \dfrac{0(34)+1(25)+2(11)+3(7)+4(3)}{80} = 1$

Use Poisson probabilities with $\mu = 1$

x	Observed	Poisson Probabilities	Expected	
0	34	0.3679	29.43	
1	25	0.3679	29.43	
2	11	0.1839	14.71	
3	7	0.0613	4.90	combine into 1
4	3	0.0153	1.22	category of 3 or
5 or more	-	0.0037	0.30	more to make $e_i \geq 5$

$\chi^2 = 4.30$ Degrees of freedom = 4 – 1 – 1 = 2

Using χ^2 table, $\chi^2 = 4.30$ shows p-value is greater than 0.10. (Actual p-value = 0.1162)

p-value > 0.05, do not reject H_0. The assumption of a Poisson distribution cannot be rejected.

37. $\bar{x} = 71$ $s = 17$ $n = 25$ Use 5 classes

Interval	Observed Frequency	Expected Frequency
less than 56.7	7	5
56.7 – 66.4	7	5
66.5 – 74.5	1	5
74.6 – 84.4	1	5
84.5 up	9	5

$\chi^2 = 11.2$ Degrees of freedom = 5 – 1 – 1 = 2

Using χ^2 table, $\chi^2 = 11.2$ shows p-value is greater than 0.005. (Actual p-value = 0.0037)

p-value < 0.01, reject H_0. Conclude the distribution is not a normal distribution.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Thirteen

Analysis of Experiments and Experimental Design

Textbook Exercises (1-30)

Textbook Exercise Solutions

Supplementary Exercises (31-45)

Supplementary Exercise Solutions

Chapter 13: Analysis of Experiments and Experimental Design

Textbook Exercises:

1. The following data are from a completely randomized design.

	Treatment		
	A	**B**	**C**
	162	142	126
	142	156	122
	165	124	138
	145	142	140
	148	136	150
	174	152	128
Sample mean	156	142	134
Sample variance	164.4	131.2	110.4

   a. Compute the sum of squares between treatments.
   b. Compute the mean square between treatments.
   c. Compute the sum of squares due to error.
   d. Compute the mean square due to error.
   e. Set up the ANOVA table for this problem.
   f. At the $\alpha = 0.05$ level of significance, test whether the means for the three treatments are equal.

2. In a completely randomized design, seven experimental units were used for each of the five levels of the factor. Complete the following ANOVA table.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*-value**
Treatments	300				
Error					
Total	460				

3. Refer to exercise 2.
   a. What hypotheses are implied in this problem?
   b. At the $\alpha = 0.05$ level of significance, can we reject the null hypothesis in part (a)? Explain.

4. In an experiment designed to test the output levels of three different treatments, the following results were obtained: SST = 400, SSTR = 150, $n_T = 19$. Set up the ANOVA table and test for any significant difference between the mean output levels of the three treatments. Use $\alpha = 0.05$.

5. In a completely randomized design, 12 experimental units were used for the first treatment, 15 for the second treatment, and 20 for the third treatment. Complete the following analysis of variance. At a 0.05 level of significance, is there a significant difference between the treatments?

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*-value**
Treatments	1200				
Error					
Total	1800				

6. Develop the analysis of variance computations for the following completely randomized design. At $\alpha = 0.05$, is there a significant difference between the treatment means?

	Treatment		
	A	**B**	**C**
	136	107	92
	120	114	82
	113	125	85
	107	104	101
	131	107	89
	114	109	117
	129	97	110
	102	114	120
		104	98
		89	106
$\bar{x}_j$	119	107	100
s_j^2	146.86	96.44	173.78

7. To test whether the mean time needed to mix a batch of material is the same for machines produced by three manufacturers, the Jacobs Chemical Company obtained the following data on the time (in minutes) needed to mix the material. Use these data to test whether the population mean times for mixing a batch of material differ for the three manufacturers.

Use $\alpha = 0.05$.

Manufacturer		
1	2	3
20	28	20
26	26	19
24	31	23
22	27	22

8. Managers at all levels of an organization need adequate information to perform their respective tasks. One study investigated the effect the source has on the dissemination of information. In this particular study the sources of information were a superior, a peer and a subordinate. In each case, a measure of dissemination was obtained, with higher values indicating greater dissemination of information. Use $\alpha = 0.05$ and the following data to test whether the source of information significantly affects dissemination. What is your conclusion, and what does it suggest about the use and dissemination of information?

Superior	Peer	Subordinate
8	6	6
5	6	5
4	7	7
6	5	4
6	3	3
7	4	5
5	7	7
5	6	5

9 A study investigated the perception of corporate ethical values among individuals specializing in marketing. Use $\alpha = 0.05$ and the following data (higher scores indicate higher ethical values) to test for significant differences in perception among the three groups.

Marketing managers	Marketing research	Advertising
6	5	6
5	5	7
4	4	6
5	4	5
6	5	6
4	4	6

10 A study reported in the Journal of Small Business Management concluded that self-employed individuals experience higher job stress than individuals who are not self-employed. In this study job stress was assessed with a 15-item scale designed to measure various aspects of ambiguity and role confl ict. Ratings for each of the 15 items were made using a scale with 1–5 response options ranging from strong agreement to strong disagreement. The sum of the ratings for the 15 items for each individual surveyed is between 15 and 75, with higher values indicating a higher degree of job stress. Suppose that a similar approach, using a 20-item scale with 1–5 response options, was used to measure the job stress of individuals for 15 randomly selected property agents, 15 architects and 15 stockbrokers. The results obtained follow.

Property agent	Architect	Stockbroker
81	43	65
48	63	48
68	60	57
69	52	91
54	54	70
62	77	67
76	68	83
56	57	75
61	61	53
65	80	71
64	50	54
69	37	72
83	73	65
85	84	58
75	58	58

Use $\alpha = 0.05$ to test for any signifi cant difference in job stress among the three professions.

11 Four different paints are advertised as having the same drying time. To check the manufacturer's claims, five samples were tested for each of the paints. The time in minutes until the paint was dry enough for a second coat to be applied was recorded. The following data were obtained.

Paint 1	Paint 2	Paint 3	Paint 4
128	144	133	150
137	133	143	142
135	142	137	135
124	146	136	140
141	130	131	153

At the $\alpha = 0.05$ level of significance, test to see whether the mean drying time is the same for each type of paint.

12 The *Consumer Reports* Restaurant Customer Satisfaction Survey is based upon 148,599 visits to full-service restaurant chains (Consumer Reports website). One of the variables in the study is meal price, the average amount paid per person for dinner and drinks, minus the tip. Suppose a reporter for the *Sun Coast Times* thought that it would be of interest to her readers to conduct a similar study for restaurants located on the Grand Strand section in Myrtle Beach, South Carolina. The reporter selected a sample of eight seafood restaurants, eight Italian restaurants, and eight steakhouses. The following data show the meal prices ($) obtained for the 24 restaurants sampled. Use $\alpha = 0.05$ to test whether there is a significant difference among the mean meal price for the three types of restaurants.

Italian	Seafood	Steakhouse
$12	$16	$24
13	18	19
15	17	23
17	26	25
18	23	21
20	15	22
17	19	27
24	18	31

13 The following data are from a completely randomized design.

	Treatment A	**Treatment B**	**Treatment C**
	32	44	33
	30	43	36
	30	44	35
	26	46	36
	32	48	40
Sample mean	30	45	36
Sample variance	6.00	4.00	6.50

a. At the $\alpha = 0.05$ level of significance, can we reject the null hypothesis that the means of the three treatments are equal?
b. Use Fisher's LSD procedure to test whether there is a significant difference between the means for treatments A and B, treatments A and C, and treatments B and C. Use $\alpha = 0.05$.
c. Use Fisher's LSD procedure to develop a 95% confidence interval estimate of the difference between the means of treatments A and B.

14 The following data are from a completely randomized design. In the following calculations, use $\alpha = 0.05$.

	Treatment 1	**Treatment 2**	**Treatment 3**
	63	82	69
	47	72	54
	54	88	61
	40	66	48
$\bar{x}_j$	51	77	58
s_j^2	96.67	97.34	81.99

a. Use analysis of variance to test for a significant difference among the means of the three treatments.
b. Use Fisher's LSD procedure to determine which means are different.

15 To test whether the mean time needed to mix a batch of material is the same for machines produced by three manufacturers, the Jacobs Chemical Company obtained the following data on the time (in minutes) needed to mix the material.

	Manufacturer	
1	**2**	**3**
20	28	20
26	26	19
24	31	23
22	27	22

a. Use these data to test whether the population mean times for mixing a batch of material differ for the three manufacturers. Use $\alpha = 0.05$.

b. At the $\alpha = 0.05$ level of significance, use Fisher's LSD procedure to test for the equality of the means for manufacturers 1 and 3. What conclusion can you draw after carrying out this test?

16 Refer to exercise 15. Use Fisher's LSD procedure to develop a 95% confidence interval estimate of the difference between the means for manufacturer 1 and manufacturer 2.

17 The following data are from an experiment designed to investigate the perception of corporate ethical values among individuals specializing in marketing (higher scores indicate higher ethical values).

Marketing Managers	**Marketing Research**	**Advertising**
6	5	6
5	5	7
4	4	6
5	4	5
6	5	6
4	4	6

a. Use $\alpha = 0.05$ to test for significant differences in perception among the three groups.

b. At the $\alpha = 0.05$ level of significance, we can conclude that there are differences in the perceptions for marketing managers, marketing research specialists, and advertising specialists. Use the procedures in this section to determine where the differences occur. Use $\alpha = 0.05$.

18 To test for any significant difference in the number of hours between breakdowns for four machines, the following data were obtained.

Machine 1	Machine 2	Machine 3	Machine 4
6.4	8.7	11.1	9.9
7.8	7.4	10.3	12.8
5.3	9.4	9.7	12.1
7.4	10.1	10.3	10.8
8.4	9.2	9.2	11.3
7.3	9.8	8.8	11.5

a. At the $\alpha = 0.05$ level of significance, what is the difference, if any, in the population mean times among the four machines?
b. Use Fisher's LSD procedure to test for the equality of the means for machines 2 and 4. Use a 0.05 level of significance.

19 Refer to exercise 18. Use the Bonferroni adjustment to test for a significant difference between all pairs of means. Assume that a maximum overall experimentwise error rate of 0.05 is desired.

20 Consider the experimental results for the following randomized block design. Make the calculations necessary to set up the analysis of variance table.

		Treatments		
		A	**B**	**C**
	1	10	9	8
	2	12	6	5
Blocks	**3**	18	15	14
	4	20	18	18
	5	8	7	8

Use $\alpha = 0.05$ to test for any significant differences.

21 The following data were obtained for a randomized block design involving five treatments and three blocks: SST = 430, SSTR = 310, SSBL = 85. Set up the ANOVA table and test for any significant differences. Use $\alpha = 0.05$.

22 An experiment has been conducted for four treatments with eight blocks. Complete the following analysis of variance table.

Source of variation	Degrees of freedom	Sum of squares	Mean Square	*F*
Treatments		900		
Blocks		400		
Error				
Total		1800		

Use $\alpha = 0.05$ to test for any significant differences.

23 A car dealer , AfricaDrive, conducted a test to determine if the time in minutes needed to complete a minor engine tune-up depends on whether a computerized engine analyzer or an electronic analyzer is used. Because tune-up time varies among compact, intermediate and full-sized cars, the three types of cars were used as blocks in the experiment. The data obtained follow.

	Analyzer	
Car	Computerized	Electronic
Compact	50	42
Intermediate	55	44
Full-sized	63	46

Use $\alpha = 0.05$ to test for any significant differences.

24 A textile mill produces a silicone proofed fabric for making into rainwear. The chemist in charge thinks that a silicone solution of about 12 per cent strength should yield a fabric with maximum waterproofing-index. He also suspected there may be some batch to batch variation because of slight differences in the cloth. To allow for this possibility five different strengths of solution were used on each of the three different batches of fabric. The following values of water-proofing index were obtained:

[Strength of silicone solution (%)]

		6	9	12	15	18
	1	20.8	20.6	22.0	22.6	20.9
Fabric	2	19.4	21.2	21.8	23.9	22.4
	3	19.9	21.1	22.7	22.7	22.1

Using $\alpha = 0.05$, carry out an appropriate test of these data and comment on the chemist's original beliefs.

25 An important factor in selecting software for word-processing and database management systems is the time required to learn how to use the system. To evaluate three file management systems, a firm designed a test involving five word-processing operators. Because operator variability was believed to be a significant factor, each of the five operators was trained on each of the three file management systems. The data obtained follow.

	System		
Operator	A	B	C
1	6	16	24
2	9	17	22
3	4	13	19
4	3	12	18
5	8	17	22

Use $\alpha = 0.05$ to test for any difference in the mean training time (in hours) for the three systems.

26 A factorial experiment involving two levels of factor A and three levels of factor B resulted in the following data.

		Factor B		
		Level 1	Level 2	Level 3
Factor A	Level 1	135	90	75
		165	66	93
	Level 2	125	127	120
		95	105	136

Test for any significant main effects and any interaction. Use $\alpha = 0.05$.

27 The calculations for a factorial experiment involving four levels of factor A, three levels of factor B, and three replications resulted in the following data: SST = 280, SSA = 26, SSB = 23, SSAB = 175. Set up the ANOVA table and test for any significant main effects and any interaction effect. Use $\alpha = 0.05$.

28 A mail-order catalogue firm designed a factorial experiment to test the effect of the size of a magazine advertisement and the advertisement design on the number of catalogue requests received (data in thousands). Three advertising designs and two different-size advertisements were considered. The data obtained follow.

		Size of advertisement	
		Small	Large
Design	A	8	12
		12	8
	B	22	26
		14	30
	C	10	18
		18	14

Use the ANOVA procedure for factorial designs to test for any signifi cant effects due to type of design, size of advertizement, or interaction. Use $\alpha = 0.05$.

29 An amusement park studied methods for decreasing the waiting time (minutes) for rides by loading and unloading riders more efficiently. Two alternative loading/unloading methods have been proposed. To account for potential differences due to the type of ride and the possible interaction between the method of loading and unloading and the type of ride, a factorial experiment was designed. Use the following data to test for any significant effect due to the loading and unloading method, the type of ride, and interaction. Use $\alpha = 0.05$.

		Type of Ride	
	Roller Coaster	**Screaming Demon**	**Log Flume**
Method 1	41	52	50
	43	44	46
Method 2	49	50	48
	51	46	44

30 As part of a study designed to compare hybrid and similarly equipped conventional vehicles, *Consumer Reports* tested a variety of classes of hybrid and all-gas model cars and sport utility vehicles (SUVs). The following data show the miles-per-gallon rating *Consumer Reports* obtained for two hybrid small cars, two hybrid midsize cars, two hybrid small SUVs, and two hybrid midsized SUVs; also shown are the miles per gallon obtained for eight similarly equipped conventional models (*Consumer Reports,* October 2008).

Make/Model	**Class**	**Type**	**MPG**
Honda Civic	Small Car	Hybrid	37
Honda Civic	Small Car	Conventional	28
Toyota Prius	Small Car	Hybrid	44
Toyota Corolla	Small Car	Conventional	32
Chevrolet Malibu	Midsize Car	Hybrid	27
Chevrolet Malibu	Midsize Car	Conventional	23
Nissan Altima	Midsize Car	Hybrid	32
Nissan Altima	Midsize Car	Conventional	25
Ford Escape	Small SUV	Hybrid	27
Ford Escape	Small SUV	Conventional	21
Saturn Vue	Small SUV	Hybrid	28
Saturn Vue	Small SUV	Conventional	22
Lexus RX	Midsize SUV	Hybrid	23
Lexus RX	Midsize SUV	Conventional	19
Toyota Highlander	Midsize SUV	Hybrid	24
Toyota Highlander	Midsize SUV	Conventional	18

At the $\alpha = 0.05$ level of significance, test for significant effects due to class, type, and interaction.

Chapter 13: Analysis of Experiments and Experimental Design

Textbook Exercises Solutions:

1.

a. $\bar{\bar{x}} = (156 + 142 + 134)/3 = 144$

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$
$$= 6(156 - 144)^2 + 6(142 - 144)^2 + 6(134 - 144)^2$$
$$= 1488$$

b. $$\text{MSTR} = \frac{\text{SSTR}}{k-1} = \frac{1488}{2} = 744$$

$s_1^2 = 164.4, \quad s_2^2 = 131.2, \quad s_3^2 = 110.4$

c. $$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2$$

$$= 5(164.4) + 5(131.2) + 5(110.4)$$
$$= 2030$$

d. $$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{2030}{18 - 3} = 135.3$$

e.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*-value**
Treatments	1488	2	744	5.50	.0162
Error	2030	15	135.3		
Total	3518	17			

f. $$F = \frac{\text{MSTR}}{\text{MSE}} = \frac{744}{135.3} = 5.50$$

From the *F* table (2 numerator degrees of freedom and 15 denominator), *p*-value is between .01 and .025 Using Excel or Minitab, the *p*-value corresponding to $F = 5.50$ is .0162

Because *p*-value $\leq \alpha = 0.05$, we reject the hypothesis that the means for the three treatments are equal

2.

Source of **Variation**	Sum of **Squares**	Degrees of **Freedom**	Mean **Square**	***F***	***p*-value**
Treatments	300	4	75	14.07	.0000
Error	160	30	5.33		
Total	460	34			

3.

a. H_0: $u_1 = u_2 = u_3 = u_4 = u_5$

H_a: Not all the population means are equal

b. Using F table (4 degrees of freedom numerator and 30 denominator), p-value is less than .01

Using Excel or Minitab, the p-value corresponding to $F = 14.07$ is .0000.

Because p-value $\leq \alpha = .05$, we reject H_0

4.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*-value**
Treatments	150	2	75	4.80	.0233
Error	250	16	15.63		
Total	400	18			

Reject H_0 because p-value $\leq \alpha = 0.05$

5.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*	*p*-value
Treatments	1200	2	600	43.99	.0000
Error	600	44	13.64		
Total	1800	46			

Using F table (2 degrees of freedom numerator and 44 denominator), p-value is less than .01

Using Excel or Minitab, the p-value corresponding to $F = 43.99$ is .0000.

Because p-value $\leq \alpha = .05$, we reject the hypothesis that the treatment means are equal.

6. Because *p*-value *p*-value = 0.0082 is less than $\alpha = 0.05$, we reject the null hypothesis that the means of the three treatments are equal

$$\bar{\bar{x}} = (79 + 74 + 66)/3 = 73$$

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 6(79 - 73)^2 + 6(74 - 73)^2 + 6(66 - 73)^2 = 516$$

$$\text{MSTR} = \frac{\text{SSTR}}{k - 1} = \frac{516}{2} = 258$$

$$s_1^2 = 34 \qquad s_2^2 = 20 \qquad s_3^2 = 32$$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 5(34) + 5(20) + 5(32) = 430$$

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{430}{18 - 3} = 28.67$$

7.

	Manufacturer 1	Manufacturer 2	Manufacturer 3
Sample Mean	23	28	21
Sample Variance	6.67	4.67	3.33

$\bar{\bar{x}} = (23 + 28 + 21)/3 = 24$

$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 4(23 - 24)^2 + 4(28 - 24)^2 + 4(21 - 24)^2 = 104$

MSTR = SSTR /(k - 1) = 104/2 = 52

$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 3(6.67) + 3(4.67) + 3(3.33) = 44.01$

MSE = SSE /(n_T - k) = 44.01/(12 - 3) = 4.89

F = MSTR /MSE = 52/4.89 = 10.63

Using F table (2 degrees of freedom numerator and 9 denominator), *p*-value is less than .01

Actual *p*-value = 0.0043

Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the mean time needed to mix a batch of material is the same for each manufacturer.

8.

	Superior	Peer	Subordinate
Sample Mean	5.75	5.5	5.25
Sample Variance	1.64	2.00	1.93

$\bar{\bar{x}} = (5.75 + 5.5 + 5.25)/3 = 5.5$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 8(5.75 - 5.5)^2 + 8(5.5 - 5.5)^2 + 8(5.25 - 5.5)^2 = 1$$

MSTR = SSTR /(k - 1) = 1/2 = 0.5

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 7(1.64) + 7(2.00) + 7(1.93) = 38.99$$

MSE = SSE /(n_T - k) = 38.99/21 = 1.86

F = MSTR /MSE = 0.5/1.86 = 0.27

Using F table (2 degrees of freedom numerator and 21 denominator), p-value is greater than .10

Actual p-value = 0.7660

Because p-value $> \alpha = 0.05$, we cannot reject the null hypothesis that the means of the three populations are equal; thus, the source of information does not significantly affect the dissemination of the information.

9.

	Marketing Managers	Marketing Research	Advertising
Sample Mean	5	4.5	6
Sample Variance	.8	.3	.4

$\bar{\bar{x}} = (5 + 4.5 + 6)/3 = 5.17$

$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 6(5 - 5.17)^2 + 6(4.5 - 5.17)^2 + 6(6 - 5.17)^2 = 7.00$

MSTR = SSTR /(k - 1) = 7.00/2 = 3.5

$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 5(.8) + 5(.3) + 5(.4) = 7.50$

MSE = SSE /(n_T - k) = 7.50/(18 - 3) = 0.5

F = MSTR /MSE = 3.5/.50 = 7.00

Using F table (2 degrees of freedom numerator and 15 denominator), p-value is less than .01

Actual p-value = 0.0071

Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the mean perception score is the same for the three groups of specialists.

10.

	Property Agent	Architect	Stockbroker
Sample Mean	67.73	61.13	65.80
Sample Variance	117.72	180.10	137.12

$\bar{\bar{x}} = (67.73 + 61.13 + 65.80)/3 = 64.89$

$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 15(67.73 - 64.89)^2 + 15(61.13 - 64.89)^2 + 15(65.80 - 64.89)^2 = 345.47$

MSTR = SSTR /(k - 1) = 345.47/2 = 172.74

$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 14(117.72) + 14(180.10) + 14(137.12) = 6089.16$

MSE = SSE /(n_T - k) = 6089.16/(45-3) = 144.98

F = MSTR /MSE = 172.74/144.98 = 1.19

Using F table (2 degrees of freedom numerator and 42 denominator), p-value is greater than .10

Actual p-value = 0.3143

Because p-value $> \alpha = 0.05$, we cannot reject the null hypothesis that the job stress ratings are the same for the three occupations.

11.

	Paint 1	Paint 2	Paint 3	Paint 4
Sample Mean	13.3	139	136	144
Sample Variance	47.5	.50	21	54.5

$\bar{\bar{x}} = (133 + 139 + 136 + 144)/3 = 138$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 5(133 - 138)^2 + 5(139 - 138)^2 + 5(136 - 138)^2 + 5(144 - 138)^2 = 330$$

$\text{MSTR} = \text{SSTR}/(k - 1) = 330/3 = 110$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 4(47.5) + 4(50) + 4(21) + 4(54.5) = 692$$

$\text{MSE} = \text{SSE}/(n_T - k) = 692/(20 - 4) = 43.25$

$F = \text{MSTR}/\text{MSE} = 110/43.25 = 2.54$

Using F table (3 degrees of freedom numerator and 16 denominator), p-value is between .05 and .10

Using Excel or Minitab the p-value corresponding to $F = 2.54$ is .0931.

Because p-value $> \alpha = .05$, we cannot reject the null hypothesis that the mean drying times for the four paints are equal.

12. p-value = 0.0038

Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the mean meal prices are the same for the three types of restaurants

13.

a. $\bar{x} = (30 + 45 + 36)/3 = 37$

$$SSTR = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2 = 5(30 - 37)^2 + 5(45 - 37)^2 + 5(36 - 37)^2 = 570$$

$$MSTR = \frac{SSTR}{k - 1} = \frac{570}{2} = 285$$

$$SSE = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 4(6) + 4(4) + 4(6.5) = 66$$

$$MSE = \frac{SSE}{n_T - k} = \frac{66}{15 - 3} = 5.5$$

$$F = \frac{MSTR}{MSE} = \frac{285}{5.5} = 51.82$$

Using F table (2 numerator degrees of freedom and 12 denominator), p-value is less than .01 Using Excel or Minitab, the p-value corresponding to $F = 51.82$ is .0000 Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the means of the three populations are equal

b.

$$LSD = t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$$= t_{.025}\sqrt{5.5\left(\frac{1}{5} + \frac{1}{5}\right)}$$

$$= 2.179\sqrt{2.2} = 3.23$$

$|\bar{x}_1 - \bar{x}_2| = |30 - 45| = 15 > LSD$; significant difference
$|\bar{x}_1 - \bar{x}_3| = |30 - 36| = 6 > LSD$; significant difference
$|\bar{x}_2 - \bar{x}_3| = |45 - 36| = 9 > LSD$; significant difference

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$(30 - 45) \pm 2.179\sqrt{5.5\left(\frac{1}{5} + \frac{1}{5}\right)}$$

c. $-15 \pm 3.23 = -18.23$ to -11.77

14.

a. Significant; p-value = 0.0106

b. LSD = 15.34

1 and 2; significant

1 and 3; not significant

2 and 3; significant

15.

a.

	Manufacturer 1	Manufacturer 2	Manufacturer 3
Sample Mean	23	28	21
Sample Variance	6.67	4.67	3.33

$$\bar{\bar{x}} = (23 + 28 + 21)/3 = 24$$

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$
$$= 4(23 - 24)^2 + 4(28 - 24)^2 + 4(21 - 24)^2$$
$$= 104$$

$$\text{MSTR} = \frac{\text{SSTR}}{k - 1} = \frac{104}{2} = 52$$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2$$
$$= 3(6.67) + 3(4.67) + 3(3.33) = 44.01$$

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{44.01}{12 - 3} = 4.89$$

$$F = \frac{\text{MSTR}}{\text{MSE}} = \frac{52}{4.89} = 10.63$$

Using F table (2 numerator degrees of freedom and 9 denominator), p-value is less than .01 Using Excel or Minitab, the p-value corresponding to $F = 10.63$ is .0043 Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the mean time needed to mix a batch of material is the same for each manufacturer.

b.

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_3}\right)}$$
$$= t_{.025}\sqrt{4.89\left(\frac{1}{4} + \frac{1}{4}\right)}$$
$$= 2.262\sqrt{2.45} = 3.54$$

Since $|\bar{x}_1 - \bar{x}_3| = |23 - 21| = 2 < 3.54$, there does not appear to be any significant difference between the means for manufacturer 1 and manufacturer 3

16.

$$\bar{x}_1 - \bar{x}_2 \pm \text{LSD}$$
$$23 - 28 \pm 3.54$$
$$-5 \pm 3.54 = -8.54 \text{ to } -1.46$$

17

a.

	Marketing Managers	Marketing Research	Advertising
Sample Mean	5	4.5	6
Sample Variance	.8	.3	.4

$\bar{\bar{x}} = (5 + 4.5 + 6)/3 = 5.17$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 6(5 - 5.17)^2 + 6(4.5 - 5.17)^2 + 6(6 - 5.17)^2 = 7.00$$

$\text{MSTR} = \text{SSTR}/(k - 1) = 7.00/2 = 3.5$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 5(.8) + 5(.3) + 5(.4) = 7.50$$

$\text{MSE} = \text{SSE}/(n_T - k) = 7.50/(18 - 3) = .5$

$F = \text{MSTR}/\text{MSE} = 3.5/.50 = 7.00$

Using F table (2 degrees of freedom numerator and 15 denominator), p-value is less than .01

Using Excel or Minitab, the p-value corresponding to $F = 7.00$ is .0071

Because p-value $\leq \alpha = .05$, we reject the null hypothesis that the mean perception score is the same for the three groups of specialists.

b. Since there are only 3 possible pairwise comparisons we will use the Bonferroni adjustment.

$\alpha = .05/3 = .017$

$t_{.017/2} = t_{.0085}$ which is approximately $t_{.01} = 2.602$

$$\text{BSD} = 2.602\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = 2.602\sqrt{.5\left(\frac{1}{6} + \frac{1}{6}\right)} = 1.06$$

$\left|\bar{x}_1 - \bar{x}_2\right| = \left|5 - 4.5\right| = .5 < 1.06$; no significant difference

$\left|\bar{x}_1 - \bar{x}_3\right| = \left|5 - 6\right| = 1 < 1.06$; no significant difference

$\left|\bar{x}_2 - \bar{x}_3\right| = \left|4.5 - 6\right| = 1.5 > 1.06$; significant difference

18 .

a. Significant; *p*-value = 0.0000

b. Significant; 2.3 > LSD = 1.19

19.

C = 6 [(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)]

$\alpha = .05/6 = .008$ and $\alpha/2 = .004$

Since the smallest value for $\alpha/2$ in the *t* table is .005, we will use $t_{.005} = 2.845$ as an approximation for $t_{.004}$ (20 degrees of freedom)

$$\text{BSD} = 2.845\sqrt{0.97\left(\frac{1}{6}+\frac{1}{6}\right)} = 1.62$$

Thus, if the absolute value of the difference between any two sample means exceeds 1.62, there is sufficient evidence to reject the hypothesis that the corresponding population means are equal.

Means	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)
| Difference |	2	2.8	4.3	0.8	2.3	1.5
Significant ?	Yes	Yes	Yes	No	Yes	No

20.

a. Significant; *p*-value = 0.011

b. Comparing North and South

|7702 – 5566| = 2136 > LSD = 1620.76

significant difference

Comparing North and West

|7702 – 8430| = 728 > LSD = 1620.76

no significant difference

Comparing South and West

|5566 – 8430| = 2864 > LSD = 1775.45

significant difference

21. ***Treatment Means***

$\bar{x}_1 = 13.6, \quad \bar{x}_2 = 11.0, \quad \bar{x}_3 = 10.6$

Block Means

$\bar{x}_{1\cdot} = 9, \ \bar{x}_{2\cdot} = 7.67, \ \bar{x}_{3\cdot} = 15.67, \ \bar{x}_{4\cdot} = 18.67, \ \bar{x}_{5\cdot} = 7.67$

Overall Mean

$\bar{\bar{x}} = 176/15 = 11.73$

Step 1

$$\begin{aligned}\text{SST} &= \sum_i \sum_j (x_{ij} - \bar{\bar{x}})^2 \\ &= (10 - 11.73)^2 + (9 - 11.73)^2 + \cdots + (8 - 11.73)^2 \\ &= 354.93\end{aligned}$$

Step 2

$$\begin{aligned}\text{SSTR} &= b \sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \\ &= 5[(13.6 - 11.73)^2 + (11.0 - 11.73)^2 \\ &\quad + (10.6 - 11.73)^2] = 26.53\end{aligned}$$

Step 3

$$\begin{aligned}\text{SSBL} &= k \sum_j (\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \\ &= 3[(9 - 11.73)^2 + (7.67 - 11.73)^2 \\ &\quad + (15.67 - 11.73)^2 + (18.67 - 11.73)^2 \\ &\quad + (7.67 - 11.73)^2] = 312.32\end{aligned}$$

Step 4

SSE = SST – SSTR – SSBL

= 354.93 – 26.53 – 312.32 = 16.08

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*– value**
Treatments	26.53	2	13.27	6.60	.0203
Blocks	312.32	4	78.08		
Error	16.08	8	2.01		
Total	354.93	14			

From the *F* table (2 numerator degrees of freedom and 8 denominator), *p*-value is between .01 and .025

Actual *p*-value = 0.0203

Because *p*-value ≤ α = 0.05, we reject the null hypothesis that the means of the three treatments are equal

22.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	*p*-value
Treatments	310	4	77.5	17.69	.0005
Blocks	85	2	42.5		
Error	35	8	4.38		
Total	430	14			

Significant; p-value $\leq \alpha = 0.05$

23.

ANOVA

Source of Variation	df	SS	MS	F	P-value
Car (Block)	2	73	36.5	3.476	0.223
Analyzer	1	216	216	20.571	0.045
Error	2	21	10.5		
Total	5	310			

From the pvalues here we reject the null hypothesis that the mean tune-up times for the two analyzer treatments are equal at the 5% level. (Note that it can be similarly deduced there is no significant difference in block means.)

24.

Source of variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Silicone treatment	4	16.103	4.026	9.0
Cloth (Blocks)	2	0.389	0.195	0.4
Error	8	3.577	0.447	
Total	14	20.069		

Using F table (4 degrees of freedom numerator and 8 denominator), $F_{.05}(4,8) = 3.84$
hence p-value is less than .05
Actual p-value = 0.005

(Note that it can be similarly deduced there is no significant difference in block means.)
Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the water index mean for the five silicone solution treatments are equal.

25.

	A	B	C
Sample Mean	119	107	100
Sample Variance	146.89	96.43	173.78

$$\bar{\bar{x}} = \frac{8(119)+10(107)+10(100)}{28} = 107.93$$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 8(119 - 107.93)^2 + 10(107 - 107.93)^2 + 10(100 - 107.93)^2 = 1617.9$$

$$\text{MSTR} = \text{SSTR}/(k - 1) = 1617.9/2 = 809.95$$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 7(146.86) + 9(96.44) + 9(173.78) = 3{,}460$$

$$\text{MSE} = \text{SSE}/(n_T - k) = 3{,}460/(28 - 3) = 138.4$$

$$F = \text{MSTR}/\text{MSE} = 809.95/138.4 = 5.85$$

Using F table (2 degrees of freedom numerator and 25 denominator), p-value is less than .01

Actual p-value = 0.0082

Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the means of the three treatments are equal.

26. *Step 1*

$$\text{SST} = \sum_i \sum_j \sum_k (x_{ijk} - \bar{\bar{x}})^2$$
$$= (135 - 111)^2 + (165 - 111)^2 + \cdots + (136 - 111)^2 = 9028$$

Step 2

$$\text{SSA} = br \sum_i (\bar{x}_{i\cdot} - \bar{\bar{x}})^2$$
$$= 3(2)[(104 - 111)^2 + (118 - 111)^2] = 588$$

Step 3

$$\text{SSB} = ar\sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2$$
$$= 2(2)[(130 - 111)^2 + (97 - 111)^2 + (106 - 111)^2]$$
$$= 2328$$

Step 4

$$\text{SSAB} = r\sum_i\sum_j (\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$$
$$= 2[(150 - 104 - 130 + 111)^2 + (78 - 104 - 97 + 111)^2 + \cdots + (128 - 118 - 106 + 111)^2] = 4392$$

Step 5

SSE = SST – SSA – SSB – SSAB

= 9028 – 588 – 2328 – 4392 = 1720

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	***F***	***p*–value**
Factor A	588	1	588	2.05	.2022
Factor B	2328	2	1164	4.06	.0767
Interaction	4392	2	2196	7.66	.0223
Error	1720	6	286.67		
Total	9028	11			

Factor A: $F = 2.05$

Using F table (1 numerator degree of freedom and 6 denominator), p-value is greater than .10 Using Excel or Minitab, the p-value corresponding to $F = 2.05$ is .2022

Because p-value > $\alpha = 0.05$, Factor A is not significant Factor B: $F = 4.06$

Using F table (2 numerator degrees of freedom and 6 denominator), p-value is between .05 and .10

Using Excel or Minitab, the p-value corresponding to $F = 4.06$ is .0767

Because p-value > $\alpha = 0.05$, Factor B is not significant Interaction: $F = 7.66$

Using F table (2 numerator degrees of freedom and 6 denominator), p-value is between .01 and .025 Using Excel or Minitab, the p-value corresponding to $F = 7.66$ is .0223

Because p-value ≤ $\alpha = 0.05$, interaction is significant

27.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Factor A	26	3	8.67	3.72	.0250
Factor B	23	2	11.50	4.94	.0160
Interaction	175	6	29.17	12.52	.0000
Error	56	24	2.33		
Total	280	35			

Using F table for Factor A (3 degrees of freedom numerator and 24 denominator), p-value is .025

Because p-value $\leq \alpha$ = .05, Factor A is significant.

Using F table for Factor B (2 degrees of freedom numerator and 24 denominator), p-value is between .01 and .025

Using Excel or Minitab, the p-value corresponding to F = 4.94 is .0160.

Because p-value $\leq \alpha$ = .05, Factor B is significant.

Using F table for Interaction (6 degrees of freedom numerator and 24 denominator), p-value is less than .01

Using Excel or Minitab, the p-value corresponding to F = 12.52 is .0000.
Because p-value $\leq \alpha$ = .05, Interaction is significant

28. Design: p-value = 0.0104; significant

Size: p-value = 0.1340; not significant

Interaction: p-value = 0.2519; not significant

29. Incorporate 11e solution (Qn 31) here.

Factor A is method of loading and unloading; Factor B is type of ride.

		Factor B: Roller Coaster	Factor B: Screaming Demon	Factor B: Log Flume	Factor A: Means
Factor A	Method 1	$\bar{x}_{11} = 42$	$\bar{x}_{12} = 48$	$\bar{x}_{13} = 48$	$\bar{x}_{1\cdot} = 46$
	Method 2	$\bar{x}_{21} = 50$	$\bar{x}_{22} = 48$	$\bar{x}_{23} = 46$	$\bar{x}_{2\cdot} = 48$
Factor B	Means	$\bar{x}_{\cdot 1} = 46$	$\bar{x}_{\cdot 2} = 48$	$\bar{x}_{\cdot 3} = 47$	$\bar{\bar{x}} = 47$

Step 1

$$SST = \sum_i \sum_j \sum_k \left(x_{ijk} - \bar{\bar{x}}\right)^2 = (41 - 47)^2 + (43 - 47)^2 + \cdots + (44 - 47)^2 = 136$$

Step 2

$$SSA = br \sum_i \left(\bar{x}_{i\cdot} - \bar{\bar{x}}\right)^2 = 3\,(2)\,[\,(46 - 47)^2 + (48 - 47)^2\,] = 12$$

Step 3

$$SSB = ar \sum_j \left(\bar{x}_{\cdot j} - \bar{\bar{x}}\right)^2 = 2\,(2)\,[\,(46 - 47)^2 + (48 - 47)^2 + (47 - 47)^2\,] = 8$$

Step 4

$$SSAB = r \sum_i \sum_j \left(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}\right)^2 = 2\,[\,(41 - 46 - 46 + 47)^2 + \cdots + (44 - 48 - 47 + 47)^2\,] = 56$$

Step 5

SSE = SST - SSA - SSB - SSAB = 136 - 12 - 8 - 56 = 60

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*	*p*-value
Factor A	12	1	12	12/10 = 1.2	.3153
Factor B	8	2	4	4/10 = .4	.6870
Interaction	56	2	28	28/10 = 2.8	.1384
Error	60	6	10		
Total	136	11			

Using *F* table for Factor A (1 degree of freedom numerator and 6 denominator), *p*-value is greater than .10

Using Excel or Minitab, the *p*-value corresponding to $F = 1.2$ is .3153.

Because *p*-value $> \alpha = .05$, Factor A is not significant

Using *F* table for Factor B (2 degrees of freedom numerator and 6 denominator), *p*-value is greater than .10

Using Excel or Minitab, the *p*-value corresponding to $F = .4$ is .6870.

Because *p*-value $> \alpha = .05$, Factor B is not significant

Using *F* table for Interaction (2 degrees of freedom numerator and 6 denominator), *p*-value is greater than .10

Using Excel or Minitab, the *p*-value corresponding to $F = 2.8$ is .1384.

Because *p*-value $> \alpha = .05$, Interaction is not significant

30**.** Class: *p*-value = 0.0002; significant

Type: *p*-value = 0.0006; significant

Interaction: *p*-value = 0.4229; not significant

Supplementary Exercises:

31. A simple random sample of the asking prices (in thousands of euros) of four houses currently for sale in each of two residential areas resulted in the following data.

Area 1	**Area 2**
92	90
89	102
98	96
105	88

a. Use the ANOVA procedure to test whether the mean asking price is the same. Use $\alpha = 0.05$.

b. Suppose that data were collected for another residential area. The asking prices for the simple random sample from the third area were €81,000, €86,000, €75,000, and €90,000. Is the mean asking price the same for all three areas? Use $\alpha = 0.05$.

32. *Money* magazine reports percentage returns and expense ratios for stock and bond funds. The following data are the expense ratios for 10 medium-sized company (Midcap) funds, 10 small company (small-cap) stock funds, 10 hybrid stock funds, and 10 specialty stock funds (*Money*, March 2003).

Midcap		**Small-Cap**	**Hybrid**	**Specialty**
1.2	2.0	2.0	1.6	
1.1	1.2	2.7	2.7	
1.0	1.7	1.8	2.6	
1.2	1.8	1.5	2.5	
1.3	1.5	2.5	1.9	
1.8	2.3	1.0	1.5	
1.4	1.9	0.9	1.6	
1.4	1.3	1.9	2.7	
1.0	1.2	1.4	2.2	
1.4	1.3	0.3	0.7	

Use $\alpha = 0.05$ to test for any significant difference in the mean expense ratio among the four types of stock funds.

33. A study reported in the *Journal of Small Business Management* concluded that self-employed individuals do not experience higher job satisfaction than individuals who are not self-employed. In this study, job satisfaction is measured using 18 items, each of which is rated using a Likert-type scale with 1–5 response options ranging from strong agreement to strong disagreement. A higher score on this scale indicates a higher degree of job satisfaction.

The sum of the ratings for the 18 items, ranging from 18–90, is used as the measure of job satisfaction. Suppose that this approach was used to measure the job satisfaction for lawyers, physiotherapists, cabinetmakers, and systems analysts. The results obtained for a sample of 10 individuals from each profession follow.

Lawyer	**Physiotherapist**	**Cabinetmaker**	**Systems Analyst**
44	55	54	44
42	78	65	73
74	80	79	71
42	86	69	60
53	60	79	64
50	59	64	66
45	62	59	41
48	52	78	55
64	55	84	76
38	50	60	62

At the $\alpha = 0.05$ level of significance, test for any difference in the job satisfaction among the four professions.

34. To investigate whether there is any difference in the annual compensation for art directors at advertising agencies, suppose that a sample of 10 art directors was selected from each of four regions: West, South, North Central, and Northeast. The base salary (in thousands of euros) for each of the individuals sampled follows.

West	**South**	**North Central**	**Northeast**
60.9	50.8	49.5	65.9
45.9	39.6	42.3	58.6
62.1	44.2	35.5	49.3
66.6	40.0	49.1	52.9
68.0	53.9	56.7	48.5
65.0	45.4	41.4	52.9
49.4	61.1	51.3	52.4

62.3	42.3	49.4	48.1
62.6	38.4	42.1	46.5
57.2	38.3	55.7	45.9

At the $\alpha = 0.05$ level of significance, test whether the mean base salary for art directors is the same for each of the four regions.

35. In a completely randomized experimental design, three brands of paper towels were tested for their ability to absorb water. Equal-size towels were used, with four sections of towels tested per brand. The absorbency rating data follow. At a 0.05 level of significance, does there appear to be a difference in the ability of the brands to absorb water?

Brand		
x	***y***	***z***
91	99	83
100	96	88
88	94	89
89	99	76

36. *Business 2.0*'s first annual employment survey provided data showing the typical annual salary for 97 different jobs. The following data show the annual salary for 30 different jobs in three fields: computer software and hardware, construction, and engineering (*Business 2.0,* March 2003).

Computers		**Construction**		**Engineering**	
Job	**Salary**	**Job**	**Salary**	**Job**	**Salary**
Data Mgr.	94	Administrator	55	Aeronautical	75
Mfg. Mgr.	90	Architect	53	Agricultural	70
Programmer	63	Architect Mgr.	77	Chemical	88
Project Mgr.	84	Const. Mgr.	60	Civil	77
Software Dev.	73	Foreperson	41	Electrical	89
Sr. Design	75	Interior Design	54	Mechanical	85
Staff Systems	94	Landscape Architect	51	Mining	96
Systems Analyst	77	Sr. Estimator	64	Nuclear	105

Use $\alpha = 0.05$ to test for any significant difference in the mean annual salary among the three job fields.

37. Three different assembly methods have been proposed for a new product. A completely

randomized experimental design was chosen to determine which assembly method results in the greatest number of parts produced per hour, and 30 workers were randomly selected and assigned to use one of the proposed methods. The number of units produced by each worker follows.

	Method	
A	**B**	**C**
97	93	99
73	100	94
93	93	87
100	55	66
73	77	59
91	91	75
100	85	84
86	73	72
92	90	88
95	83	86

Use these data and test to see whether the mean number of parts produced is the same with each method. Use $\alpha = 0.05$.

38. Hassels Automotive Parts, wanted to compare the wear for four different types of brake linings. Thirty linings of each type were produced and placed on a fleet of rental cars. The number of kilometres that each brake lining lasted until it no longer met the required federal safety standard was recorded, and an average value was computed for each type of lining. The following data were obtained.

Type	**Sample Size**	**Sample Mean**	**Standard Deviation**
A	30	32,000	1450
B	30	27,500	1525
C	30	34,200	1650
D	30	30,300	1400

Test to see whether the corresponding population means are equal. Use $\alpha = 0.05$.

39. A manufacturer of batteries for electronic toys and calculators is considering three new battery designs. Data were collected to determine whether the mean lifetime in hours is the same for each of the three designs.

Design A	Design B	Design C
78	112	115
98	99	101
88	101	100
96	116	120

Test to see whether the population means are equal. Use $\alpha = 0.05$.

40. In a study conducted to investigate browsing activity by shoppers, each shopper was initially classified as a non-browser, light browser, or heavy browser. For each shopper, the study obtained a measure to determine how comfortable the shopper was in a store. Higher scores indicated greater comfort. Suppose the following data were collected.

Non-browser	Light Browser	Heavy Browser
4	5	5
5	6	7
6	5	5
3	4	7
3	7	4
4	4	6
5	6	5
4	5	7

a. Use $\alpha = 0.05$ to test for differences among comfort levels for the three types of browsers.

b. Use Fisher's LSD procedure to compare the comfort levels of non-browsers and light browsers. Use $\alpha = 0.05$. What is your conclusion?

41. A research firm tests the kilometres-per-litre characteristics of three brands of petrol. Because of different petrol performance characteristics in different brands of cars, five brands of cars are selected and treated as blocks in the experiment; that is, each brand of car is tested with each type of petrol. The results of the experiment (in kilometres per litre) follow.

		Petrol Brands		
		I	**II**	**III**
	A	**8**	**9**	**9**
	B	**10**	**11**	**11**
Cars	**C**	**13**	**12**	**14**
	D	**9**	**11**	**10**
	E	**9**	**10**	**10**

a. At $\alpha = 0.05$, is there a significant difference in the mean kilometres-per-litre characteristics of the three brands of petrol?

b. Analyze the experimental data using the ANOVA procedure for completely randomized designs. Compare your findings with those obtained in part (a). What is the advantage of attempting to remove the block effect?

42. Each month *Internet Magazine* accesses more than 100 Internet service providers (ISPs) in order to check the availability of the ISP and test the speed of the connection by measuring the time (seconds) it takes to download a number of popular Web pages. The following data show the download time for 22 free ISPs for Web sites located in the United Kingdom, United States, and Europe (*Internet Magazine*, January 2000).

ISP Name	**U.K.**	**U.S.**	**Europe**
Abel Gratis	10.62	14.64	17.08
Breathe	11.67	14.14	19.86
btclick.com	12.12	16.43	21.30
Bun	11.13	14.09	15.83
Cable & Wireless Life	9.99	13.07	18.43
conX	12.63	15.97	22.12
Freebeeb	11.71	15.52	19.57
Free-Online	13.77	13.98	23.35
Freeserve	10.65	13.62	25.56
FreeUK	12.20	14.96	18.95
Icom-Web	9.62	11.66	15.91
IPNet	13.82	16.70	22.86
I-way Soho	14.86	12.86	19.32
LineOne	12.01	17.82	21.88
Madasafish	13.38	15.59	19.61
NetDirect Online	11.71	15.52	19.57
Netscape Online	10.84	12.66	16.52
Screaming.net (BT Line)	13.23	15.91	23.08
Telinco Internet Services	12.83	15.34	18.76
UK Online	10.39	13.28	21.04
UKPeople	13.79	19.82	19.76
Virgin Net	12.17	15.47	21.94

At $\alpha = 0.05$, is there a significant difference in the mean download time for Web sites located in the United Kingdom, United States, and Europe?

43. A factorial experiment was designed to test for any significant differences in the time needed to perform English to foreign language translations with two computerized language translators. Because the type of language translated was also considered a significant factor, translations were made with both systems for three different languages: Spanish, French, and German. Use the following data for translation time in hours.

		Language	
	Spanish	**French**	**German**
System 1	8	10	12
	12	14	16
System 2	6	14	16
	10	16	22

a. Test for any significant differences due to language translator, type of language, and interaction. Use $\alpha = 0.05$.

44. A manufacturing company designed a factorial experiment to determine whether the number of defective parts produced by two machines differed and if the number of defective parts produced also depended on whether the raw material needed by each machine was loaded manually or by an automatic feed system. The following data give the numbers of defective parts produced.

	Loading System	
	Manual	**Automatic**
Machine 1	30	30
	34	26
Machine 2	20	24
	22	28

a. Use $\alpha = 0.05$ to test for any significant effect due to machine, loading system, and interaction.

45. A factorial experiment involved measurement of average fuel consumption for 24 long journeys for three different types of vehicle and four different types of fuel additive. The data obtained were as follows:

	Fuel additive			
Vehicle type	1	2	3	4
A	34.0	30.1	29.8	29.0
	32.7	32.8	26.7	28.9
B	32.0	30.2	28.7	27.6
	33.2	29.8	28.1	27.8
C	28.4	27.3	29.7	28.8
	29.3	28.9	27.3	29.3

a. Perform an appropriate analysis of these data.

b. What are your conclusions?

Supplementary Exercises Solutions:

31 . a.

	Area 1	Area 2
Sample Mean	96	94
Sample Variance	50	40

$\bar{\bar{x}} = (96 + 94)/2 = 95$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 4(96 - 95)^2 + 4(94 - 95)^2 = 8$$

$\text{MSTR} = \text{SSTR}/(k - 1) = 8/1 = 8$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 3(50) + 3(40) = 270$$

$\text{MSE} = \text{SSE}/(n_T - k) = 270/(8 - 2) = 45$

$F = \text{MSTR/MSE} = 8/45 = 0.18$

Using F table (1 degree of freedom numerator and 6 denominator), p-value is greater than .10

Actual p-value = 0.6862

Because p-value $> \alpha = 0.05$, the means are not significantly different.

b.

	Area 1	Area 2	Area 3
Sample Mean	96	94	83
Sample Variance	50	40	42

$\bar{\bar{x}} = (96 + 94 + 83)/3 = 91$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 4(96 - 91)^2 + 4(94 - 91)^2 + 4(83 - 91)^2 = 392$$

MSTR = SSTR /(k - 1) = 392 /2 = 196

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 3(50) + 3(40) + 3(42) = 396$$

MSTR = SSE /(n_T - k) = 396 /(12 - 3) = 44

F = MSTR /MSE = 196 /44 = 4.45

Using F table (2 degrees of freedom numerator and 6 denominator), p-value is between .05 and .10

Actual p-value = 0.0653

Because p-value > α = 0.05, we cannot reject the null hypothesis that the mean asking prices for all three areas are equal.

32. The Minitab output is shown below:

```
      Analysis of Variance
      Source     DF        SS        MS        F        P
      Factor      3     2.603     0.868     2.94    0.046
      Error      36    10.612     0.295
      Total      39    13.215
                                          Individual 95% CIs
For Mean
                                          Based on Pooled
StDev
      Level       N      Mean     StDev  -------+---------
+---------+---------
      Midcap     10    1.2800    0.2394  (--------*--------
)
      Smallcap   10    1.6200    0.3795           (-------
*--------)
      Hybrid     10    1.6000    0.7379          (--------
*--------)
      Specialt   10    2.0000    0.6583
(--------*--------)
                                          -------+---------
+---------+---------
      Pooled StDev =   0.5429               1.20
1.60      2.00
```

Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the mean expense ratios are equal.

33.

	Lawyer	Physical Therapist	Cabinet Maker	Systems Analyst
Sample Mean	50.0	63.7	69.1	61.2
Sample Variance	124.22	164.68	105.88	136.62

$$\bar{\bar{x}} = \frac{50.0 + 63.7 + 69.1 + 61.2}{4} = 61$$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 10(50.0 - 61)^2 + 10(63.7 - 61)^2 + 10(69.1 - 61)^2 +$$

$10(61.2 - 61)^2 = 1939.4$

MSTR = SSTR /(k - 1) = 1939.4 /3 = 646.47

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 9(124.22) + 9(164.68) + 9(105.88) + 9(136.62) = 4{,}782.60$$

MSE = SSE /(n_T - k) = 4782.6 /(40 - 4) = 132.85

F = MSTR /MSE = 646.47 /132.85 = 4.87

Using F table (3 degrees of freedom numerator and 36 denominator), p-value is less than .01

Actual p-value = 0.0061

Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the mean job satisfaction rating is the same for the four professions.

34. The Minitab output is shown below:

```
      ANALYSIS OF VARIANCE
      SOURCE      DF        SS         MS         F         p
      FACTOR       3    1271.0      423.7     8.74     0.000
      ERROR       36    1744.2       48.4
      TOTAL       39    3015.2
                                             INDIVIDUAL 95 PCT
CI'S FOR MEAN
                                             BASED ON POOLED
STDEV
      LEVEL        N      MEAN      STDEV  --+---------+-----
----+---------+----
      West        10    60.000      7.218
(------*-----)
      South       10    45.400      7.610  (------*-----)
      N.Cent      10    47.300      6.778     (------*-----)
      N.East      10    52.100      6.152            (-----*-
-----)
                                             --+---------+-----
----+---------+----
      POOLED STDEV =     6.961           42.0      49.0
56.0      63.0
```

Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that that the mean base salary for art directors is the same for each of the four regions.

35.

	x	*y*	*z*
Sample Mean	92	97	84
Sample Variance	30	6	35.33

$\bar{\bar{x}} = (92 + 97 + 44)/3 = 91$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 4(92 - 91)^2 + 4(97 - 91)^2 + 4(84 - 91)^2 = 344$$

$\text{MSTR} = \text{SSTR}/(k - 1) = 344/2 = 172$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 3(30) + 3(6) + 3(35.33) = 213.99$$

$\text{MSE} = \text{SSE}/(n_T - k) = 213.99/(12 - 3) = 23.78$

$F = \text{MSTR}/\text{MSE} = 172/23.78 = 7.23$

Using *F* table (2 degrees of freedom numerator and 9 denominator), *p*-value is between .01 and .025

Actual *p*-value = 0.0134

Because *p*-value $\leq \alpha = 0.05$, we reject the null hypothesis that the mean absorbency ratings for the three brands are equal.

36. The Minitab output is shown below:

```
      Analysis of Variance
      Source     DF        SS        MS        F        P
      Factor      2      3840      1920    15.64    0.000
      Error      21      2578       123
      Total      23      6418
                                          Individual 95% CIs
For Mean
                                          Based on Pooled
StDev
      Level       N      Mean     StDev  --------+---------
+---------+--------
      Computer    8     81.25     11.13
(----*-----)
      Constr.     8     56.88     10.55  (-----*----)
      Engineer    8     85.63     11.54
(----*-----)
                                          --------+---------
+---------+--------
      Pooled StDev =    11.08                  60
75        90
```

Because p-value $\leq \alpha = 0.05$, we reject the null hypothesis that the mean salary is the same for the three jobs.

37.

	Method A	Method B	Method C
Sample Mean	90	84	81
Sample Variance	98.00	168.44	159.78

$\bar{\bar{x}} = (90 + 84 + 81)/3 = 85$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 10(90 - 85)^2 + 10(84 - 85)^2 + 10(81 - 85)^2 = 420$$

$\text{MSTR} = \text{SSTR}/(k - 1) = 420/2 = 210$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 9(98.00) + 9(168.44) + 9(159.78) = 3{,}836$$

$\text{MSE} = \text{SSE}/(n_T - k) = 3{,}836/(30 - 3) = 142.07$

$F = \text{MSTR}/\text{MSE} = 210/142.07 = 1.48$

Using F table (2 degrees of freedom numerator and 27 denominator), p-value is greater than .10

Actual p-value = 0.2455

Because p-value $> \alpha = 0.05$, we can not reject the null hypothesis that the means are equal.

38.

	Type A	Type B	Type C	Type D
Sample Mean	32,000	27,500	34,200	30,300
Sample Variance	2,102,500	2,325,625	2,722,500	1,960,000

$\bar{\bar{x}} = (32{,}000 + 27{,}500 + 34{,}200 + 30{,}000)/4 = 31{,}000$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 30(32{,}000 - 31{,}000)^2 + 30(27{,}500 - 31{,}000)^2 +$$

$$30(34{,}200 - 31{,}000)^2 +$$

$$30(30{,}300 - 31{,}000)^2 = 719{,}400{,}000$$

$\text{MSTR} = \text{SSTR}/(k - 1) = 719{,}400{,}000/3 = 239{,}800{,}000$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 29(2{,}102{,}500) + 29(2{,}325{,}625) + 29(2{,}722{,}500) +$$

$$29(1{,}960{,}000)$$

$$= 264{,}208{,}125$$

$\text{MSE} = \text{SSE}/(n_T - k) = 264{,}208{,}125/(120 - 4) = 2{,}277{,}656.25$

$F = \text{MSTR}/\text{MSE} = 239{,}800{,}000/2{,}277{,}656.25 = 105.28$

Using F table (3 degrees of freedom numerator and 116 denominator), p-value is less than .01

Actual p-value = 0.0000

Because $p\text{-value} \leq \alpha = 0.05$, we reject the null hypothesis that the population means are equal.

39.

	Design A	Design B	Design C
Sample Mean	90	107	109
Sample Variance	82.67	68.67	100.67

$\bar{\bar{x}} = (90 + 107 + 109)/3 = 102$

$$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 4(90 - 102)^2 + 4(107 - 102)^2 + (109 - 102)^2 = 872$$

$\text{MSTR} = \text{SSTR}/(k - 1) = 872/2 = 436$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 3(82.67) + 3(68.67) + 3(100.67) = 756.03$$

$\text{MSE} = \text{SSE}/(n_T - k) = 756.03/(12 - 3) = 84$

$F = \text{MSTR}/\text{MSE} = 436/84 = 5.19$

Using F table (2 degrees of freedom numerator and 9 denominator), p-value is between .025 and .05

Actual p-value = 0.0317

Because $p\text{-value} \leq \alpha = 0.05$, we reject the null hypothesis that the mean lifetime in hours is the same for the three designs.

40. a.

	Nonbrowser	Light Browser	Heavy Browser
Sample Mean	4.25	5.25	5.75
Sample Variance	1.07	1.07	1.36

$\bar{\bar{x}} = (4.25 + 5.25 + 5.75)/3 = 5.08$

$\text{SSTR} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 8(4.25 - 5.08)^2 + 8(5.25 - 5.08)^2 + 8(5.75 - 5.08)^2 =$ 9.33

$\text{MSB} = \text{SSB}/(k - 1) = 9.33/2 = 4.67$

$\text{SSW} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 7(1.07) + 7(1.07) + 7(1.36) = 24.5$

$\text{MSW} = \text{SSW}/(n_T - k) = 24.5/(24 - 3) = 1.17$

$F = \text{MSB}/\text{MSW} = 4.67/1.17 = 3.99$

Using F table (2 degrees of freedom numerator and 21 denominator), p-value is between .025 and .05

Actual p-value = 0.0340

Because $p\text{-value} \le \alpha = 0.05$, we reject the null hypothesis that the mean comfort scores are the same for the three groups.

b. $\text{LSD} = t_{\alpha/2}\sqrt{\text{MSW}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = 2.080\sqrt{1.17\left(\frac{1}{8} + \frac{1}{8}\right)} = 1.12$

Since the absolute value of the difference between the sample means for nonbrowsers and light browsers is $|4.25 - 5.25| = 1$, we cannot reject the null hypothesis that the two population means are equal.

41. a. Treatment Means:

Brands	Mean
I	$9.8 = \bar{x}_{\cdot 1}$
II	$10.6 = \bar{x}_{\cdot 2}$
III	$10.8 = \bar{x}_{\cdot 3}$

Block Means:

Cars	Mean
A	$8.6667 = \bar{x}_{1\cdot}$
B	$10.6667 = \bar{x}_{2\cdot}$
C	$13.0000 = \bar{x}_{3\cdot}$
D	$10.0000 = \bar{x}_{4\cdot}$
E	$9.6667 = \bar{x}_{5\cdot}$

Overall Mean:

$\bar{\bar{x}} = 156 / 15 = 10.4$

The Minitab output for these data is shown below:

Two-way ANOVA: kml versus Cars, Brands

Source	DF	SS	MS	F	P
Cars	4	31.6	7.9	19.75	0.000
Brands	2	2.8	1.4	3.50	0.081
Error	8	3.2	0.4		
Total	14	37.6			

S = 0.6325 R-Sq = 91.49% R-Sq(adj) = 85.11%

Because *p*-value (Brands) > $\alpha = 0.05$, we accept the null hypothesis that the mean kms per litre ratings for the three brands of petrol are equal.

b. The Minitab output for these data is shown below:

One-way ANOVA: kml versus Brands

Source	DF	SS	MS	F	P
Brands	2	2.80	1.40	0.48	0.629
Error	12	34.80	2.90		
Total	14	37.60			

S = 1.703 R-Sq = 7.45% R-Sq(adj) = 0.00%

Again because *p*-value $> \alpha = 0.05$, we cannot reject the null hypothesis that the mean miles per gallon ratings for the three brands of petrol are equal.

Note however that by removing the block effect the pvalue of 0.081has increased to 0.629. So whereas with the randomized block design, Brands are close to having significantly different means (and there is actually significance at the 10% level) when the blocks are ignored, there is not the slightest hint of significance between Brands.

Note that SSE for the completely randomized design is the sum of SSBL (31.6) and SSE (3.2) for the randomized block design. This illustrates that the effect of blocking is to remove the block effect from the error sum of squares; thus, the estimate of σ^2 for the randomized block design is substantially smaller than it is for the completely randomized design.

42. The Minitab output for these data is shown below:

```
Analysis of Variance
Source     DF        SS        MS        F        P
Factor      2    731.75    365.88    93.16    0.000
Error      63    247.42      3.93
Total      65    979.17
                                        Individual 95% CIs
For Mean
                                        Based on Pooled
StDev
Level       N      Mean     StDev  ---+---------+----
-----+---------+---
UK         22    12.052     1.393  (--*--)
US         22    14.957     1.847            (--*--)
Europe     22    20.105     2.536
(--*--)
                                        ---+---------+----
-----+---------+---
Pooled StDev =    1.982             12.0      15.0
18.0       21.0
```

Because the p-value $\leq \alpha = 0.05$, we can reject the null hypothesis that the mean download time is the same for Web sites located in the three countries. Note that the mean download time for Web sites located in the United Kingdom (12.052 seconds) is less than the mean download time for Web sites in the United States (14.957) and Web sites located in Europe (20.105).

43.

		Factor B			Factor A
		Spanish	French	German	Means
Factor A	System 1	$\bar{x}_{11} = 10$	$\bar{x}_{12} = 12$	$\bar{x}_{13} = 14$	$\bar{x}_{1\cdot} = 12$
	System 2	$\bar{x}_{21} = 8$	$\bar{x}_{22} = 15$	$\bar{x}_{23} = 19$	$\bar{x}_{2\cdot} = 14$
Factor B	Means	$\bar{x}_{\cdot 1} = 9$	$\bar{x}_{\cdot 2} = 13.5$	$\bar{x}_{\cdot 3} = 16.5$	$\bar{\bar{x}} = 13$

Step 1

$$\text{SST} = \sum_i \sum_j \sum_k \left(x_{ijk} - \bar{\bar{x}}\right)^2 = (8 - 13)^2 + (12 - 13)^2 + \cdots + (22 - 13)^2 = 204$$

Step 2

$$\text{SSA} = br \sum_i \left(\bar{x}_{i\cdot} - \bar{\bar{x}}\right)^2 = 3\,(2)\,[\,(12 - 13)^2 + (14 - 13)^2\,] = 12$$

Step 3

$$\text{SSB} = ar \sum_j \left(\bar{x}_{\cdot j} - \bar{\bar{x}}\right)^2 = 2\,(2)\,[\,(9 - 13)^2 + (13.5 - 13)^2 + (16.5 - 13)^2\,] = 114$$

Step 4

$$\text{SSAB} = r \sum_i \sum_j \left(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}\right)^2 = 2\,[(8 - 12 - 9 + 13)^2 + \cdots + (22 - 14 - 16.5 + 13)^2] = 26$$

Step 5

SSE = SST - SSA - SSB - SSAB = 204 - 12 - 114 - 26 = 52

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	*F*
Factor A	1	12	12	1.38
Factor B	2	114	57	6.57
Interaction	2	26	12	1.50
Error	6	52	8.67	
Total	11	204		

Factor A: Actual *p*-value = 0.2846. Because *p*-value $> \alpha = 0.05$, Factor A (translator) is not significant.

Factor B: Actual *p*-value = 0.0308. Because *p*-value $\leq \alpha = 0.05$, Factor B (language translated) is significant.

Interaction: Actual *p*-value = 0.2963. Because *p*-value $> \alpha = 0.05$, Interaction is not significant.

44.

		Factor B		Factor B
		Manual	Automatic	Means
Factor A	Machine 1	$\bar{x}_{11} = 32$	$= 28$	$= 36$
	Machine 2	$\bar{x}_{21} = 21$	$= 26$	$= 23.5$
Factor B Means		$= 26.5$	$= 27$	$= 26.75$

Step 1

$$\text{SST} = \sum_i \sum_j \sum_k \left(x_{ijk} - \bar{\bar{x}}\right)^2 = (30 - 26.75)^2 + (34 - 26.75)^2 + \cdots + (28 - 26.75)^2 =$$

151.5

Step 2

$$\text{SSA} = br\sum_i \left(\bar{x}_{i\square} - \bar{\bar{x}}\right)^2 = 2\ (2)\ [\ (30 - 26.75)^2 + (23.5 - 26.75)^2\ ] = 84.5$$

Step 3

$$\text{SSB} = ar\sum_j \left(\bar{x}_{\square j} - \bar{\bar{x}}\right)^2 = 2\ (2)\ [\ (26.5 - 26.75)^2 + (27 - 26.75)^2\ ] = 0.5$$

Step 4

$$\text{SSAB} = r\sum_i \sum_j \left(\bar{x}_{ij} - \bar{x}_{i\square} - \bar{x}_{\square j} + \bar{\bar{x}}\right)^2 = 2[(30 - 30 - 26.5 + 26.75)^2 + \cdot\ \cdot\ \cdot + (28 - 23.5 - 27 + 26.75)^2]$$

$$= 40.5$$

Step 5

SSE = SST - SSA - SSB - SSAB = 151.5 - 84.5 - 0.5 - 40.5 = 26

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	*F*
Factor A	1	84.5	84.5	13
Factor B	1	.5	.5	.08
Interaction	1	40.5	40.5	6.23
Error	4	26	6.5	
Total	7	151.5		

Factor A: Actual *p*-value = 0.0226. Because *p*-value $\leq \alpha$ = 0.05, Factor A (machine) is significant.

Factor B: Actual *p*-value = 0.7913. Because *p*-value $> \alpha$ = 0.05, Factor B (loading system) is not significant.

Interaction: Actual *p*-value = 0.0671. Because *p*-value $> \alpha$ = 0.05, Interaction is not significant.

45 Output from EXCEL for these data is shown below:

Source		df	SS	MS	F	pvalue
Vehicles		2	14.130	7.065	5.656	0.019
Additives	3		39.663	13.221	10.584	0.001
Interaction	6		22.737	3.789	3.034	0.048
Error		12	14.990	1.249		
Total		23	91.520			

In every case, p-value $\leq \alpha$ = 0.05 so we infer both individual factors are significant as well as the interaction term (but only just).

Because of the significant interaction, we must focus on treatments (combinations of factor levels) rather than main effects of factors:

Treatment means can be derived as follows:

Vehicle type	1	2	3	4
A	33.35	31.45	28.25	28.95
B	32.6	30	28.4	27.7
C	28.85	28.1	28.5	29.05

From which it can be deduced the combination A1 (car type A, additive 1) appears most uneconomical, combination B4 (car type B, additive 4) appears most economical etc

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Fourteen

Simple Linear Regression

Textbook Exercises (1-30)

Textbook Exercise Solutions

Supplementary Exercises (31-45)

Supplementary Exercise Solutions

Chapter 14: Simple Linear Regression

Textbook Exercises:

1 Given are five observations for two variables, X and Y

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Develop a scatter diagram for these data.

b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

c. Try to approximate the relationship between X and Y by drawing a straight line through the data.

d. Develop the estimated regression equation by computing the values of b0 and b1 using equations (14.6) and (14.7).

e. Use the estimated regression equation to predict the value of Y when X = 4.

2 Given are five observations for two variables, X and Y.

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. Develop a scatter diagram for these data.

b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

c. Try to approximate the relationship between X and Y by drawing a straight line through the data.

d. Develop the estimated regression equation by computing the values of b0 and b1 using equations (14.6) and (14.7).

e. Use the estimated regression equation to predict the value of Y when X = 6.

3 Given are five observations collected in a regression study on two variables.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

a. Develop a scatter diagram for these data.

b. Develop the estimated regression equation for these data.

c. Use the estimated regression equation to predict the value of Y when X = 4.

4 The following data were collected on the height (cm) and weight (kg) of women swimmers.

Height	173	163	157	165	168
Weight	60	49	46	52	58

a. Develop a scatter diagram for these data with height as the independent variable.

b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

c. Try to approximate the relationship between height and weight by drawing a straight line through the data.

d. Develop the estimated regression equation by computing the values of b0 and b1.

e. If a swimmer's height is 160 cm, what would you estimate her weight to be?

5 The Dow Jones Industrial Average (DJIA) and the Standard & Poor's 500 (S&P) indexes are both used as measures of overall movement in the stock market. The DJIA is based on the price movements of 30 large companies; the S&P 500 is an index composed of 500 stocks. Some say the S&P 500 is a better measure of stock market performance because it is broader based. The closing prices for the DJIA and the S&P 500 for ten weeks, beginning with 11 February 2009, follow (uk.fi nance.yahoo.com, 21 April 2009).

Date	DJIA	S&P
11 Feb 09	7939.53	833.74
18 Feb 09	7555.63	788.42
25 Feb 09	7270.89	764.90
03 Mar 09	6726.02	696.33
10 Mar 09	6926.49	719.60
17 Mar 09	7395.70	778.12
24 Mar 09	7660.21	806.12
31 Mar 09	7608.92	797.87
07 Apr 09	7789.56	815.55
14 Apr 09	7920.18	841.50

a. Develop a scatter diagram for these data with DJIA as the independent variable.

b. Develop the least squares estimated regression equation.

c. Suppose the closing price for the DJIA is 8000. Estimate the closing price for the S&P 500.

6 The following table shows the observations of transportation time and distance for a sample of 10 rail shipments made by a motor parts supplier.

Delivery time (days)	Distance (kilometres)
5	210
7	290
6	350
11	480
8	490
11	730
12	780
8	850
15	920
12	1010

a. Develop a scatter diagram for these data with distance as the independent variable.
b. Develop an estimated regression equation that can be used to predict annual sales given the years of experience.
c. Use the estimated regression equation to predict delivery time for a customer situated 600 kilometres from the company.

7 The data from exercise 1 follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

The estimated regression equation for these data is $\hat{y} = 0.20 + 2.60x$.

a. Compute SSE, SST and SSR using equations (14.8), (14.9) and (14.10).

b. Compute the coefficient of determination r2. Comment on the goodness of fi t.

c. Compute the sample correlation coefficient.

8 The data from exercise 2 follow.

x_i	2	3	5	1	8
y_i	25	25	20	30	16

The estimated regression equation for these data is $\hat{y} = 30.33 - 1.88x$.

a. Compute SSE, SST and SSR.

b. Compute the coefficient of determination r2. Comment on the goodness of fit.

c. Compute the sample correlation coefficient.

9 The data from exercise 3 follow.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

The estimated regression equation for these data is $\hat{y} = 0.75 + 0.51x$. What percentage of the total sum of squares can be accounted for by the estimated regression equation? What is the value of the sample correlation coefficient?

10 The estimated regression equation for the data in exercise 5 can be shown to be $\hat{y} = -75.586 + 0.115x$. What percentage of the total sum of squares can be accounted for by the estimated regression equation?
Comment on the goodness of fi t. What is the sample correlation coefficient?

11 An investment manager studying haulage companies calculates for a random sample of six such firms, the percentage capital investment in vehicles and the profit before tax as a percentage of turnover with the following results:

% Capital investment, vehicles	37	47	10	22	41	25
% Profit	14	21	-5	16	19	8

a. Calculate the coefficient of determination. What percentage of the variation in total cost can be explained by production volume?
b. Carry out a linear regression analysis for the data.
c. Hence estimate the % profit when the % Capital investment, vehicles is

(i) 30%

(ii) 90%

12 PCWorld provided details for ten of the most economical laser printers (PCWorld, April 2009).

The following data show the maximum printing speed in pages per minute (ppm) and the price (in euros including 15 per cent value added tax) for each printer.

Name	Speed (ppm)	Price (€)
Brother HL 2035	18	61.35
HP Laserjet P1005	15	70.13
Samsung ML-1640	16	77.39
HP Laserjet P1006	17	82.93
Brother HL-2140	22	92.34
Brother DCP7030	22	96.04
HP Laserjet P1009	16	99.52
HP Laserjet P1505	24	119.10
Samsung 4300	18	121.64
Epson EPL-6200 Mono	20	133.53

a. Develop the estimated regression equation with speed as the independent variable.

b. Compute r2. What percentage of the variation in cost can be explained by the printing speed?

c. What is the sample correlation coefficient between speed and price? Does it reflect a strong or weak relationship between printing speed and cost?

13 The data from exercise 1 follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Compute the mean square error using equation (14.15).

b. Compute the standard error of the estimate using equation (14.16).

c. Compute the estimated standard deviation of b1 using equation (14.18). d. Use the t test to test the following hypotheses ($\alpha = 0.05$):

$H0: \beta 1 = 0$

$H1: \beta 1 \neq 0$

e. Use the F test to test the hypotheses in part (d) at a 0.05 level of significance. Present the results in the analysis of variance table format.

14 The data from exercise 2 follow.

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. Compute the mean square error using equation (14.15).

b. Compute the standard error of the estimate using equation (14.16).

c. Compute the estimated standard deviation of b1 using equation (14.18).

d. Use the t test to test the following hypotheses ($\alpha = 0.05$):

$H0: \beta 1 = 0$

$H1: \beta 1 \neq 0$

e. Use the F test to test the hypotheses in part (d) at a 0.05 level of significance. Present the results in the analysis of variance table format.

15 The data from exercise 3 follow.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

a. What is the value of the standard error of the estimate?

b. Test for a significant relationship by using the t test. Use $\alpha = 0.05$.

c. Use the F test to test for a significant relationship. Use $\alpha = 0.05$. What is your conclusion?

16 The Supplies Office of a local authority is reviewing its policy for the replacement of photocopiers. For the 10 photocopiers in use within the local authority, the number of breakdowns during the past year has been recorded.

Photocopier	A	B	C	D	E	F	G	H	I	J
No of breakdowns	11	9	13	10	18	13	15	8	16	10
Age (years)	6	4	6	2	9	4	8	1	7	3

The Supplies Offices wishes to determine how the number of breakdowns depends upon the age of the photocopier.
Use $\alpha = 0.05$ to test whether number of breakdowns is significantly related to the age.
Show the ANOVA table. What is your conclusion?

17 Refer to exercise 12 where the data were used to determine whether the price of a printer is related to the speed for plain text printing (PC World, April 2009). Does the evidence indicate a significant relationship between printing speed and price? Conduct the appropriate statistical test and state your conclusion. Use $\alpha = 0.05$.

18 The data from exercise 1 follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Use expression (14.22) to develop a 95 per cent confidence interval for the expected value of Y when X = 4.

b. Use expression (14.23) to develop a 95 per cent prediction interval for Y when X = 4.

19 The data from exercise 2 follow.

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. Estimate the standard deviation of $\hat{y}_p$ when X = 3.

b. Develop a 95 per cent confidence interval for the expected value of Y when X = 3.

c. Estimate the standard deviation of an individual value of Y when X = 3.

d. Develop a 95 per cent prediction interval for Y when X = 3.

20 The data from exercise 3 follow.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

Develop the 95 per cent confidence and prediction intervals when X = 3. Explain why these two intervals are different.

21 A company that manufactures ball-point pens has a cost function of the form:

$$T=T_0 + kx^2$$

where T_0 is a constant value linked to the production method used and x is the quantity of pens (in thousands) manufactured. During the last year, the total costs of the company were recorded as follows (where pens were recorded in thousands and costs are recorded in €'000s)

Month	# of pens (x)	Total cost (T)
Jan	5.5	80.1
Feb	4.2	80.4
Mar	6.4	58.0
Apr	3.3	90.1
May	7.2	47.2
Jun	8.6	27.0
Jul	9.2	17.4
Aug	3.9	82.8
Sep	6.8	53.8
Oct	8.3	33.1
Nov	5.9	63.2
Dec	8.2	32.8

a. Derive least squares estimates of T_0 and k.
b. Hence determine a 95% interval estimate of Total Cost when 6000 pens are manufactured.

22 The commercial division of the Supreme real estate firm in Cyprus is conducting a regression analysis of the relationship between *X*, annual gross rents (in thousands of euros), and *Y*, selling price (in thousands of euros) for apartment buildings. Data were collected on several properties recently sold and the following computer selective output was obtained.

```
The regression equation is
Y = 20.0 + 7.21 X

Predictor       Coef    SE Coef       T
Constant      20.000     3.2213    6.21
X              7.210     1.3626    5.29

Analysis of Variance

SOURCE            DF         SS
Regression         1    41587.3
Residual Error     7
Total              8    51984.1
```

a. How many apartment buildings were in the sample?

b. Write the estimated regression equation.

c. What is the value of s b 1 ?

d. Use the F statistic to test the significance of the relationship at a 0.05 level of significance.

e. Estimate the selling price of an apartment building with gross annual rents of €50 000.

23 Following is a portion of the computer output for a regression analysis relating Y = maintenance expense (euros per month) to X = usage (hours per week) of a particular brand of computer terminal.

```
The regression equation is
Y = 6.1092 + .8951 X

Predictor       Coef     SE Coef
Constant      6.1092      0.9361
X             0.8951      0.1490

Analysis of Variance

SOURCE            DF         SS         MS
Regression         1    1575.76    1575.76
Residual Error     8     349.14      43.64
Total              9    1924.90
```

a. Write the estimated regression equation.

b. Use a t test to determine whether monthly maintenance expense is related to usage at the 0.05 level of significance.

c. Use the estimated regression equation to predict mean monthly maintenance expense for any terminal that is used 25 hours per week.

24 A regression model relating X, number of salespersons at a branch office, to Y, annual sales at the office (in thousands of euros) provided the following computer output from a regression analysis of the data.

```
The regression equation is
Y = 80.0 + 50.00 X

Predictor       Coef     SE Coef       T
Constant        80.0      11.333    7.06
X               50.0       5.482    9.12

Analysis of Variance

SOURCE             DF          SS          MS
Regression          1      6828.6      6828.6
Residual Error     28      2298.8        82.1
Total              29      9127.4
```

a. Write the estimated regression equation.

b. How many branch offices were involved in the study?

c. Compute the F statistic and test the significance of the relationship at a 0.05 level of significance.

d. Predict the annual sales at the Marseilles branch office. This branch employs 12 salespersons.

25 Given are data for two variables, X and Y.

x_i	6	11	15	18	20
y_i	6	8	12	20	30

a. Develop an estimated regression equation for these data.

b. Compute the residuals.

c. Develop a plot of the residuals against the independent variable X. Do the assumptions about the error terms seem to be satisfied?

d. Compute the standardized residuals.

e. Develop a plot of the standardized residuals against $\hat{y}$. What conclusions can you draw from this plot?

26 The following data were used in a regression study.

Observation	x_i	y_i	Observation	x_i	y_i
1	2	4	6	7	6
2	3	5	7	7	9
3	4	4	8	8	5
4	5	6	9	9	11
5	7	4			

a. Develop an estimated regression equation for these data.

b. Construct a plot of the residuals. Do the assumptions about the error term seem to be satisfied?

27 A doctor has access to historical data as follows:

	Vehicles per 100 population	Road death per 100 000 population
Great Britain	31	14
Belgium	32	29
Denmark	30	22
France	47	32
Germany	30	25
Irish Republic	19	20
Italy	36	21
Netherlands	40	22
Canada	47	30
USA	58	35

a. First identifying the X and Y variables appropriately, use the method of least squares to develop a straight line approximation of the relationship between the two variables.
b. Test whether vehicles and road deaths are related at a 0.05 level of significance.
c. Prepare a residual plot of $y - \hat{y}$ versus $\hat{y}$. Use the result from part (a) to obtain the values of $\hat{y}$.
d. What conclusions can you draw from residual analysis? Should this model be used, or should we look for a better one?

28 Consider the data concerning years of experience (Y) and annual sales (X) for different salespersons:

Salesperson	Years of experience	Annual sales (€000s)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a. Compute the residuals and construct a residual plot for this problem.

b. Do the assumptions about the error terms seem reasonable in light of the residual plot?

29 Consider the following data for two variables, X and Y.

x_i	135	110	130	145	175	160	120
y_i	145	100	120	120	130	130	110

a. Compute the standardized residuals for these data. Do there appear to be any outliers in the data? Explain.

b. Plot the standardized residuals against ^y . Does this plot reveal any outliers?

c. Develop a scatter diagram for these data. Does the scatter diagram indicate any outliers in the data? In general, what implications does this fi nding have for simple linear regression?

30 Consider the following data for two variables, X and Y.

x_i	4	5	7	8	10	12	12	22
y_i	12	14	16	15	18	20	24	19

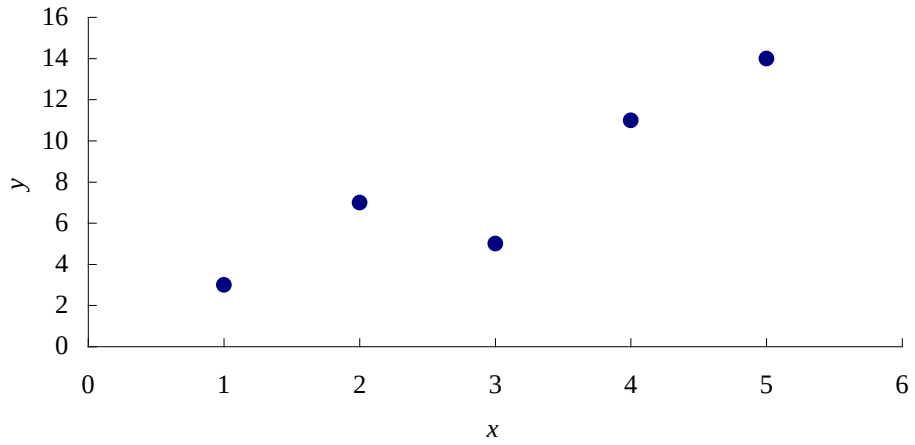
a. Compute the standardized residuals for these data. Do there appear to be any outliers in the data? Explain.

b. Compute the leverage values for these data. Do there appear to be any influential observations in these data? Explain.

c. Develop a scatter diagram for these data. Does the scatter diagram indicate any influential observations? Explain.

Chapter 14: Simple Linear Regression

Textbook Exercises Solutions:



1 a.

b. There appears to be a linear relationship between *x* and y.

c. Many different straight lines can be drawn to provide a linear approximation of the relationship between *x and* y; in part d we will determine the equation of a straight line that "best" represents the relationship according to the least squares criterion.

d. Summations needed to compute the slope and y-intercept are:

$$\Sigma x_i = 15 \quad \Sigma y_i = 40 \quad \Sigma(x_i - \bar{x})(y_i - \bar{y}) = 26 \quad \Sigma(x_i - \bar{x})^2 = 10$$

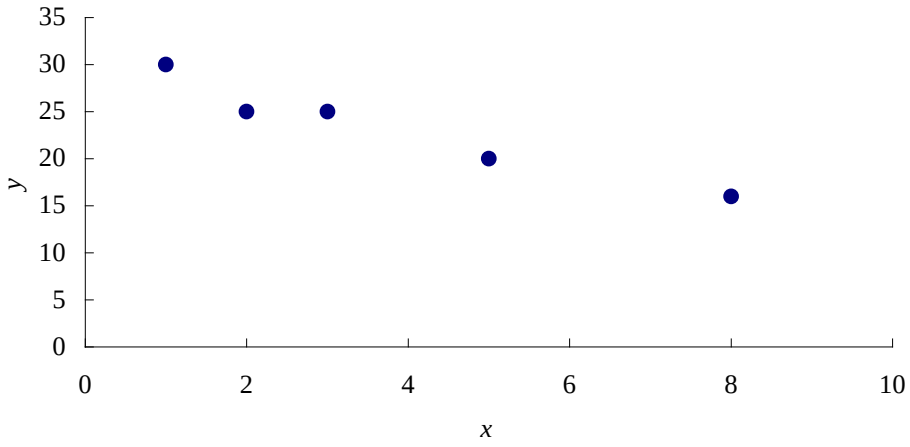
$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{26}{10} = 2.6$$

$$b_0 = \bar{y} - b_1\bar{x} = 8 - (2.6)(3) = 0.2$$

$$\hat{y} = 0.2 + 2.6x$$

e. $\hat{y} = 0.2 + 2.6(4) = 10.6$

2. a.



b. There appears to be a linear relationship between *x* and *y*.

c. Many different straight lines can be drawn to provide a linear approximation of the relationship between *x and y*; in part d we will determine the equation of a straight line that "best" represents the relationship according to the least squares criterion.

d. Summations needed to compute the slope and *y*-intercept are:

$$\Sigma x_i = 19 \quad \Sigma y_i = 116 \quad \Sigma(x_i - \bar{x})(y_i - \bar{y}) = -57.8 \quad \Sigma(x_i - \bar{x})^2 = 30.8$$

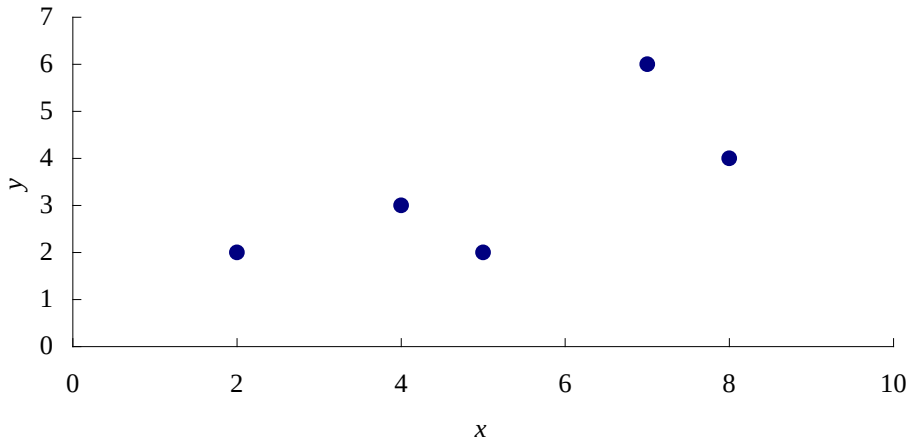
$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{-57.8}{30.8} = -1.8766$$

$$b_0 = \bar{y} - b_1\bar{x} = 23.2 - (-1.8766)(3.8) = 30.3311$$

$$\hat{y} = 30.33 - 1.88x$$

e. $\hat{y} = 30.33 - 1.88(6) = 19.05$

3. a.



b. Summations needed to compute the slope and *y*-intercept are:

$$\Sigma x_i = 26 \quad \Sigma y_i = 17 \quad \Sigma(x_i - \bar{x})(y_i - \bar{y}) = 11.6 \quad \Sigma(x_i - \bar{x})^2 = 22.8$$

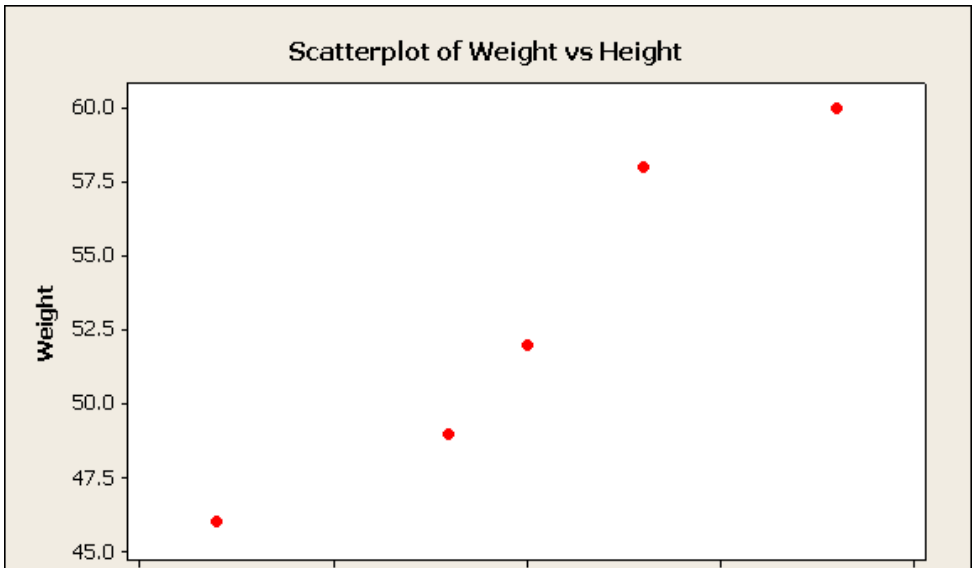
$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{11.6}{22.8} = 0.5088$$

$$b_0 = \bar{y} - b_1\bar{x} = 3.4 - (0.5088)(5.2) = 0.7542$$

$$\hat{y} = 0.75 + 0.51x$$

c. $\hat{y} = 0.75 + 0.51(4) = 2.79$

4. a.



b. There appears to be a linear relationship between *x* and *y*.

c. Many different straight lines can be drawn to provide a linear approximation of the relationship between *x and y*; in part d we will determine the equation of a straight line that "best" represents the relationship according to the least squares criterion.

d. Summations needed to compute the slope and *y*-intercept are:

$$\sum x_i = 826 \quad \sum y_i = 265 \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = 135 \quad \sum (x_i - \bar{x})^2 = 140.8$$

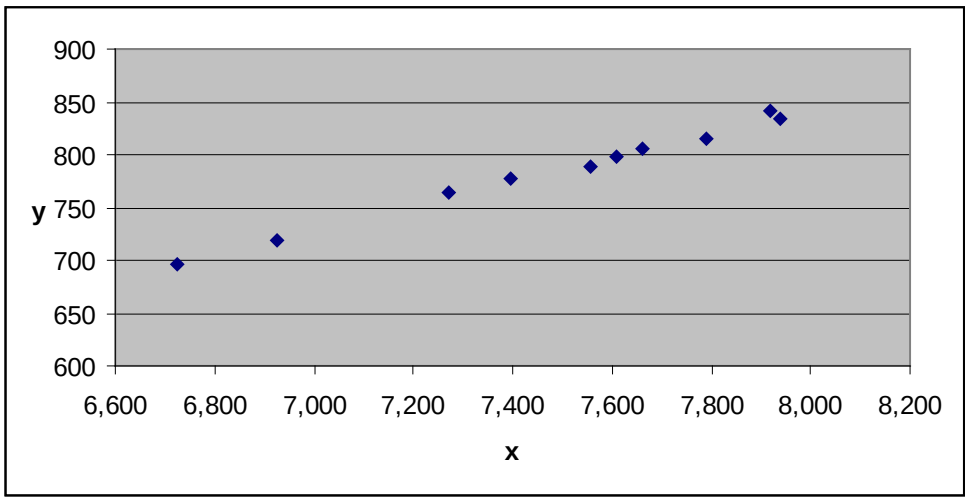
$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{135}{140.8} = 0.9588$$

$$b_0 = \bar{y} - b_1 \bar{x} = 53 - (0.9588)(165.2) = -105.39$$

$$\hat{y} = -105.39 + 0.9588x$$

e. $\hat{y} = -105.39 + 0.9588x = -105.39 + 0.9588(160) = 48.02\,kg$

5. a.



b. Let x = DJIA and y = S&P. Summations needed to compute the slope and y-intercept are:

$$\sum x_i = 74{,}793.13 \quad \sum y_i = 7{,}842.15 \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = 170{,}281.27$$

$$\sum (x_i - \bar{x})^2 = 1{,}481{,}257.48$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{170281.27}{1481257.48} = .1149572$$

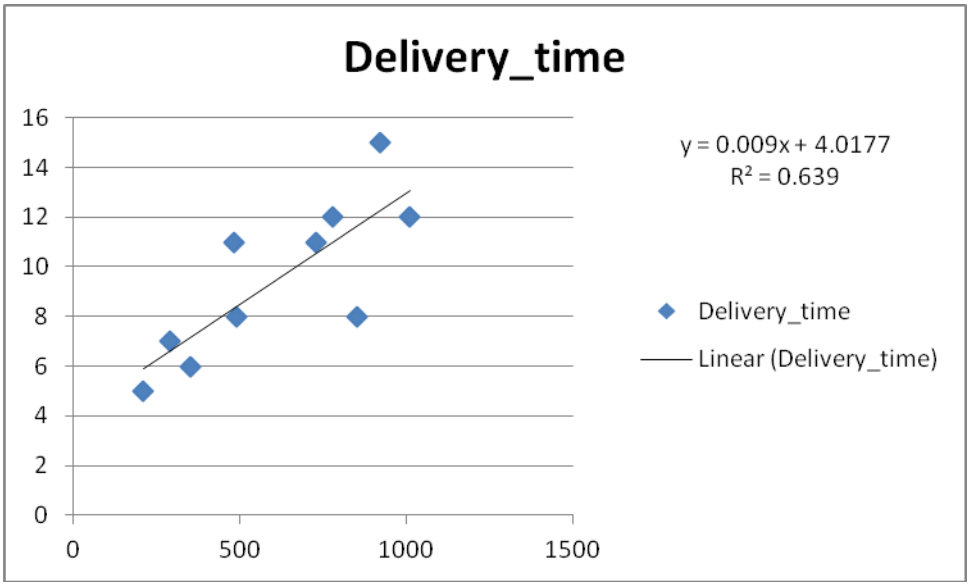
$$b_0 = \bar{y} - b_1 \bar{x} = 784.22 - (.1149572)(7479.31) = -75.586$$

$$\hat{y} = -75.586 + .11496x$$

c. $\hat{y} = -75.586 + .11496(1100) = 1188.94$ or approximately 1189

6.

a./b.



c. When distance = 600, $\hat{y}$ =0.009x +4.0177 = 5.4 + 4.0177 = 9.4177 days

7. a. The estimated regression equation and the mean for the dependent variable are:

$\hat{y}_i = 0.2 + 2.6x_i \qquad \bar{y} = 8$

The sum of squares due to error and the total sum of squares are

$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = 12.40 \qquad \text{SST} = \sum(y_i - \bar{y})^2 = 80$

Thus, SSR = SST - SSE = 80 - 12.4 = 67.6

b. r^2 = SSR/SST = 67.6/80 = .845

The least squares line provided a very good fit; 84.5% of the variability in *y* has been explained by the least squares line.

c. $r = \sqrt{.845} = +.9192$

8. a. The estimated regression equation and the mean for the dependent variable are:

$\hat{y}_i = 30.33 - 1.88x \qquad \bar{y} = 23.2$

The sum of squares due to error and the total sum of squares are

$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = 6.33 \qquad \text{SST} = \sum(y_i - \bar{y})^2 = 114.80$

Thus, SSR = SST - SSE = 114.80 - 6.33 = 108.47

b. r^2 = SSR/SST = 108.47/114.80 = .945

The least squares line provided an excellent fit; 94.5% of the variability in *y* has been explained by the estimated regression equation.

c. $r = \sqrt{.945} = -.9721$

Note: the sign for *r* is negative because the slope of the estimated regression equation is negative.

(b_1 = -1.88)

9. The estimated regression equation and the mean for the dependent variable are:

$\hat{y}_i = .75 + .51x \qquad \bar{y} = 3.4$

The sum of squares due to error and the total sum of squares are

$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = 5.3 \qquad \text{SST} = \sum(y_i - \bar{y})^2 = 11.2$

Thus, SSR = SST - SSE = 11.2 - 5.3 = 5.9

r^2 = SSR/SST = 5.9/11.2 = .527

We see that 52.7% of the variability in *y* has been explained by the least squares line.

$r = \sqrt{.527} = +.7259$

10. The estimated regression equation is:

$\hat{y} = -75.586 + .11496x$

The sum of squares due to error and the total sum of squares are

SSE = 134.1735 SST = 19,709.24

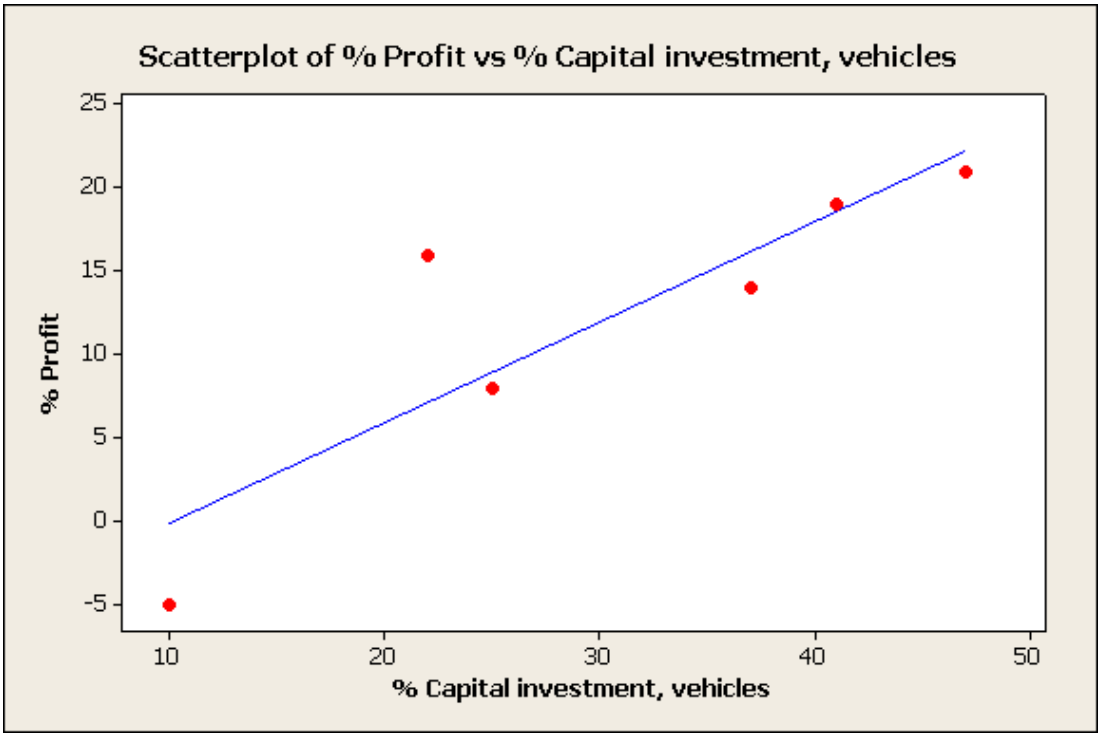
Thus, SSR = SST - SSE = 19,709.24 – 134.1735 = 19575.07

r^2 = SSR/SST = 19,575.07/19,709.24 = .993

We see that 99.3% of the variability in *y* has been explained by the least squares line. This is a very high goodness of fit.

$r = \sqrt{.993} = .997$

11. a.



b./c.

Regression Analysis: % Profit versus % Capital investment, vehicles

The regression equation is
% Profit = - 6.14 + 0.603 % Capital investment, vehicles

Predictor	Coef	SE Coef	T	P
Constant	-6.138	5.591	-1.10	0.334
% Capital investment, vehicles	0.6034	0.1703	3.54	0.024

S = 5.24076 R-Sq = 75.8% R-Sq(adj) = 69.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	344.97	344.97	12.56	0.024
Residual Error	4	109.86	27.47		
Total	5	454.83			

Predicted Values for New Observations

X=30

New Obs	Fit	SE Fit	95% CI	95% PI
1	11.97	2.14	(6.02, 17.91)	(-3.75, 27.68)

X=90

New Obs	Fit	SE Fit	95% CI	95% PI
1	48.17	10.38	(19.35, 77.00)	(15.88, 80.46)XX

XX denotes a point that is an extreme outlier in the predictors.

12. a. Let x = speed (ppm) and y = price (€)

The summations needed in this problem are:

$$\sum x_i = 188 \quad \sum y_i = 953.97 \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = 324.864$$

$$\sum (x_i - \bar{x})^2 = 83.6$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{324.864}{83.6} = 3.8859$$

$$b_0 = \bar{y} - b_1 \bar{x} = 95.397 - (3.8859)(18.8) = 22.341$$

$$\hat{y} = 22.341 + 3.8859x$$

b. The sum of squares due to error and the total sum of squares are:

SSE = 3746.309 SST = 5008.708

Thus, SSR = SST - SSE = 5008.708– 3746.309= 1262.4

r^2 = SSR/SST = 1262.4/5008.708 = .252

We see that 25.2% of the variability in price has been explained by the speed. This is a relatively poor goodness of fit.

c. $r = \sqrt{.252} = +.502$

13. a. s^2 = MSE = SSE / (n - 2) = 12.4 / 3 = 4.133

b. $s = \sqrt{\text{MSE}} = \sqrt{4.133} = 2.033$

c. $\Sigma(x_i - \bar{x})^2 = 10$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{2.033}{\sqrt{10}} = 0.643$$

d. $$t = \frac{b_1}{s_{b_1}} = \frac{2.6}{.643} = 4.04$$

Using t table (3 degrees of freedom), area in tail is between .01 and .025

p-value is between .02 and .05

Actual p-value = .0272

Because p-value $\leq \alpha$, we reject H_0: $\beta_1 = 0$

e. MSR = SSR / 1 = 67.6

F = MSR / MSE = 67.6 / 4.133 = 16.36

Using F table (1 degree of freedom numerator and 3 denominator), p-value is between .025 and .05

Actual p-value = .0272

Because p-value $\leq \alpha$, we reject H_0: $\beta_1 = 0$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Regression	1	67.6	67.6	16.36
Error	3	12.4	4.133	
Total	4	80.0		

14. a. s^2 = MSE = SSE / (n - 2) = 6.33 / 3 = 2.11

b. $s = \sqrt{\text{MSE}} = \sqrt{2.11} = 1.453$

c. $\Sigma(x_i - \bar{x})^2 = 30.8$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{1.453}{\sqrt{30.8}} = 0.262$$

d. $$t = \frac{b_1}{s_{b_1}} = \frac{-1.88}{.262} = -7.18$$

Using t table (3 degrees of freedom), area in tail is less than .005; p-value is less than .01

Actual p-value = .0056

Because p-value $\leq \alpha$, we reject H_0: $\beta_1 = 0$

e. MSR = SSR / 1 = 8.47

F = MSR / MSE = 108.47 / 2.11 = 51.41

Using F table (1 degree of freedom numerator and 3 denominator), p-value is less than .01

Actual p-value = .0056

Because p-value $\leq \alpha$, we reject H_0: $\beta_1 = 0$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Regression	1	108.47	108.47	51.41
Error	3	6.33	2.11	
Total	4	114.80		

15. a. s^2 = MSE = SSE / (n - 2) = 5.30 / 3 = 1.77

$$s = \sqrt{\text{MSE}} = \sqrt{1.77} = 1.33$$

b. $\Sigma(x_i - \bar{x})^2 = 22.8$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{1.33}{\sqrt{22.8}} = 0.28$$

$$t = \frac{b_1}{s_{b_1}} = \frac{.51}{.28} = 1.82$$

Using t table (3 degrees of freedom), area in tail is between .05 and .10

p-value is between .10 and .20

Actual p-value = .1664

Because p-value $> \alpha$, we cannot reject H_0: $\beta_1 = 0$; x and y do not appear to be related.

c. MSR = SSR/1 = 5.90 /1 = 5.90

F = MSR/MSE = 5.90/1.77 = 3.33

Using F table (1 degree of freedom numerator and 3 denominator), p-value is greater than .10

Actual p-value = .1664

Because p-value $> \alpha$, we cannot reject H_0: $\beta_1 = 0$; x and y do not appear to be related.

16.

	df	SS	MS	F	Significance F
Regression	1	49.542	49.542	31.814	0.000
Residual	8	12.458	1.557		
Total	9	62			

Because p-value $\leq \alpha$, we reject H_0: $\beta_1 = 0$. Number of breakdowns and age are related.

17. Using the computations from Exercise 12,

SSE = 3746.309 SST = 5008.708 SSR = 1262.4

s^2 = MSE = SSE/(n-2) = 3746.309/8 = 468.2886

$$s = +\sqrt{468.2886} = 21.64$$

$$\sum (x_i - \bar{x})^2 = 83.6$$

$$s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{21.64}{\sqrt{83.6}} = 2.37$$

$$t = \frac{b_1}{s_{b_1}} = \frac{3.8859}{2.37} = 1.64$$

Using t table (8 degree of freedom), area in tail is greater than .05

Actual p-value = .139
Because p-value$\geq \alpha$ we cannot reject H_0: $\beta_1 = 0$

There is not a significant relationship between x and y.

18. a. s = 2.033

$$\bar{x} = 3 \quad \Sigma (x_i - \bar{x})^2 = 10 \quad \hat{y} = 0.2 + 2.6x = 0.2 + 2.6(4) = 10.6$$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}} = 2.033\sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}} = 1.11$$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$$

10.6 ± 3.182 (1.11) = 10.6 ± 3.53

or 7.07 to 14.13

b. $\hat{y}_p \pm t_{\alpha/2} s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}} = \sqrt{1+\frac{1}{5}+\frac{(4-3)^2}{10}}$

10.6 ± 3.182 (2.32) = 10.6 ± 7.38

or 3.22 to 17.98

19. a. $s = 1.453$

b. $\bar{x} = 3.8 \quad \Sigma(x_i-\bar{x})^2 = 30.8$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}} = 1.453\sqrt{\frac{1}{5}+\frac{(3-3.8)^2}{30.8}} = .068$$

$$\hat{y} = 30.33 - 1.88x = 30.33 - 1.88(3) = 24.69$$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$$

24.69 ± 3.182 (.68) = 24.69 ± 2.16

or 22.53 to 26.85

c. $s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}} = \sqrt{1+\frac{1}{5}+\frac{(3-3.8)^2}{30.8}} = 1.61$

d. $\hat{y}_p \pm t_{\alpha/2}(1.61) =$

24.69 ± 3.182 (1.61) = 24.69 ± 5.12

or 19.57 to 29.81

20. $s = 1.33$

$\bar{x} = 5.2 \quad \Sigma(x_i - \bar{x})^2 = 22.8$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = 1.33\sqrt{\frac{1}{5} + \frac{(3-5.2)^2}{22.8}} = 0.85$$

$$\hat{y} = 0.75 + 0.51x = 0.75 + 0.51(3) = 2.28$$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$$

2.28 ± 3.182 (.85) = 2.28 ± 2.70

or -.40 to 4.98

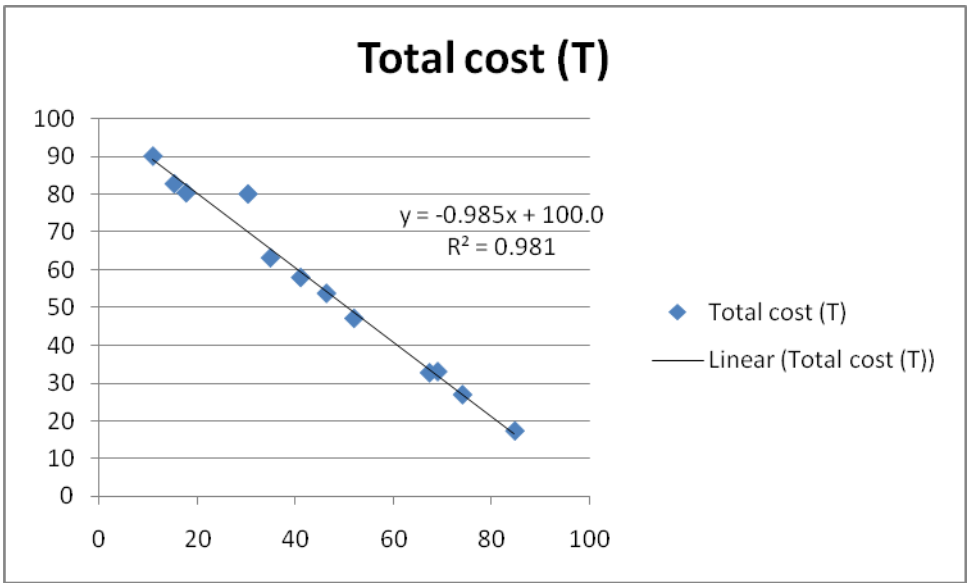
$$s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = \sqrt{1 + \frac{1}{5} + \frac{(3-5.2)^2}{22.8}} = 1.58$$

$$\hat{y}_p \pm t_{\alpha/2}(1.58) =$$

2.28 ☐ 3.182 (1.58) = 2.28 ☐ 5.03

or -2.27 to 7.31

21. a.



b. Corresponding to 6000 pens, x = 6 and $\hat{T} = 100 - 0.985 * 6^2 = 64.57$ **so estimated costs = £64,570.**

Relevant formula interval estimate formula is:

$$\hat{T}_p \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_p^2 - \bar{x}^2)^2}{\sum(x^2 - \bar{x}^2)^2}} \text{ where } s^2 = \frac{\sum e^2}{n-2} = \frac{\sum(T - \hat{T})^2}{n-2}$$

here $x_p = 67$ $\hat{T}_p = 64.54$ and $t_{.025}(10) = 2.228$ (Note that [illegible] 2.344)

Yielding the 95% confidence interval: 64.57+/- 2.43 = (62.14, 67.0)

22. a. 9

b. $\hat{y} = 20.0 + 7.21x$

c. 1.3626

d. SSE = SST - SSR = 51,984.1 - 41,587.3 = 10,396.8

MSE = 10,396.8/7 = 1,485.3

F = MSR / MSE = 41,587.3 /1,485.3 = 28.00

Using *F* table (1 degree of freedom numerator and 7 denominator), *p*-value is less than .01

Actual *p*-value = .0011

Because *p*-value $\leq \alpha$, we reject H_0: $B_1 = 0$.

e. $\hat{y} = 20.0 + 7.21(50) = 380.5$ or €380,500

23. a. $\hat{y} = 6.1092 + .8951x$

b. $t = \dfrac{b_1 - B_1}{s_{b_1}} = \dfrac{.8951 - 0}{.149} = 6.01$

Using *t* table (1 degree of freedom numerator and 8 denominator), area in tail is less than .005

p-value is less than .01

Actual *p*-value = .0004

Because *p*-value $\leq \alpha$, we reject H_0: $\beta_1 = 0$

c. $\hat{y} = 6.1092 + .8951(25) = 28.49$ or €28.49 per month

24 a. $\hat{y} = 80.0 + 50.0x$

b. 30

c. *F* = MSR / MSE = 6828.6/82.1 = 83.17

Using *F* table (1 degree of freedom numerator and 28 denominator), *p*-value is less than .01

Actual *p*-value = .0001

Because *p*-value $< \alpha$ = .05, we reject H_0: $\beta_1 = 0$.

Branch office sales are related to the salespersons.

d. $\hat{y} = 80 + 50\ (12) = 680$ or €680,000

25. a. $\Sigma x_i = 70 \quad \Sigma y_i = 76 \quad \Sigma(x_i - \bar{x})(y_i - \bar{y}) = 200 \quad \Sigma(x_i - \bar{x})^2 = 126$

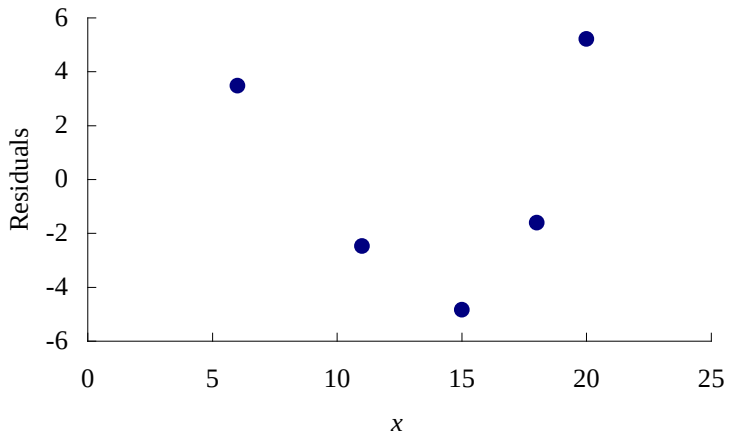
$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{200}{126} = 1.5873$$

$$b_0 = \bar{y} - b_1\bar{x} = 15.2 - (1.5873)(14) = -7.0222$$

$$\hat{y} = -7.02 + 1.59x$$

b. The residuals are 3.48, -2.47, -4.83, -1.6, and 5.22

c.



With only 5 observations it is difficult to determine if the assumptions are satisfied. However, the plot does suggest curvature in the residuals that would indicate that the error term assumptions are not satisfied. The scatter diagram for these data also indicates that the underlying relationship between *x* and *y* may be curvilinear.

d. $s^2 = 23.78$

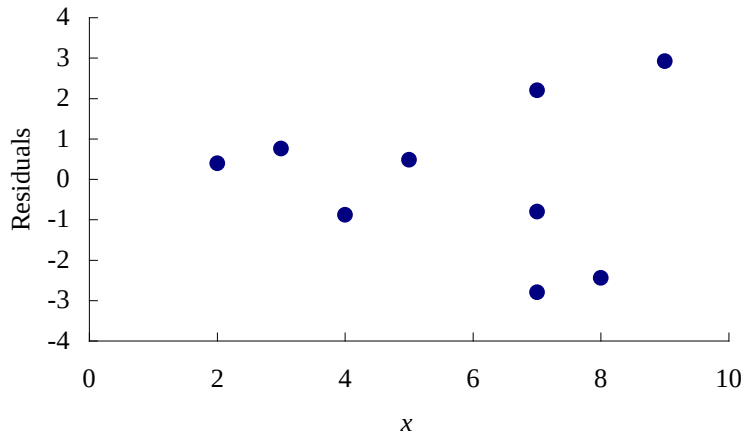
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} = \frac{1}{5} + \frac{(x_i - 14)^2}{126}$$

The standardized residuals are 1.32, -.59, -1.11, -.40, 1.49.

e. The standardized residual plot has the same shape as the original residual plot. The curvature observed indicates that the assumptions regarding the error term may not be satisfied.

26.a. $\hat{y} = 2.32 + .64x$

b.



The assumption that the variance is the same for all values of x is questionable. The variance appears to increase for larger values of x.

27. a. /b.

The regression equation is
Deaths = 9.27 + 0.425 Vehicles

Predictor	Coef	SE Coef	T	P
Constant	9.273	5.194	1.79	0.112
Vehicles	0.4250	0.1349	3.15	0.014

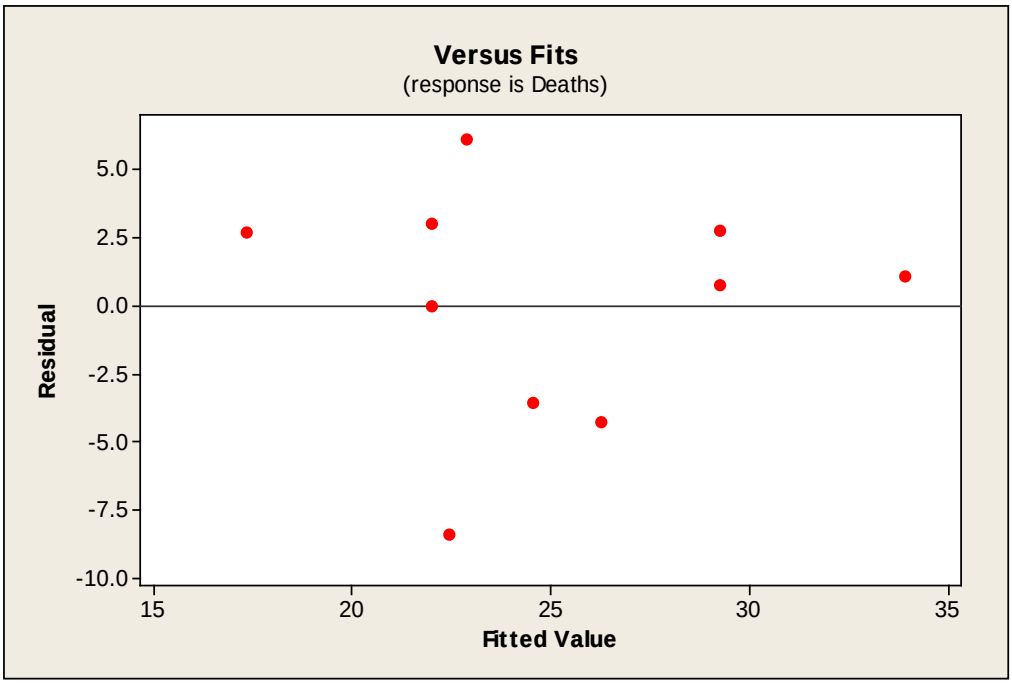
S = 4.54325 R-Sq = 55.4% R-Sq(adj) = 49.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	204.87	204.87	9.93	0.014
Residual Error	8	165.13	20.64		
Total	9	370.00			

As the pvalue for Vehicles is 0.014 < 0.05 = α we deduce the model is significant.

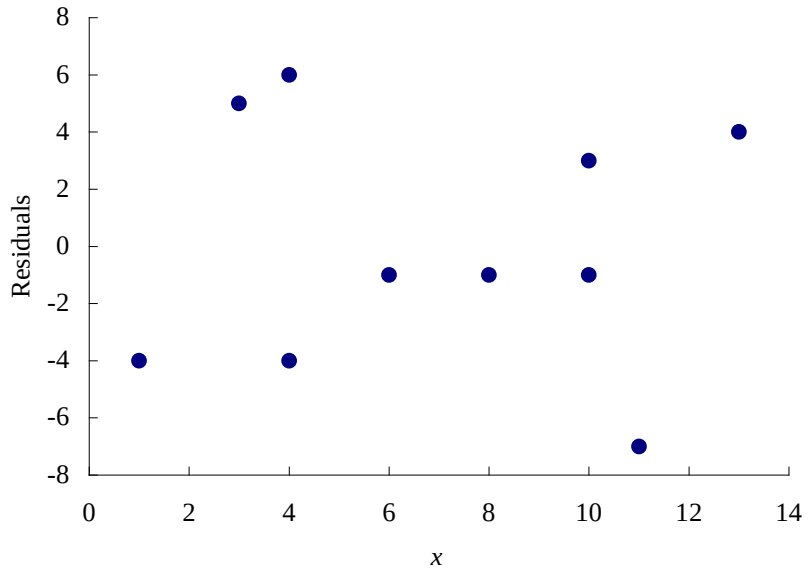
b.



c. The above residual plot looks archetypically random so yes this model can be used.

28. a. $\hat{y} = 80 + 4x$

Observation	Predicted Y	Residual
1	84	-4
2	92	5
3	96	-4
4	96	6
5	104	-1
6	112	-1
7	120	-1
8	120	3
9	124	-7
10	132	4



b. The assumptions concerning the error term appear reasonable.

29. a. The MINITAB output is shown below:

The regression equation is
Y = 66.1 + 0.402 X

Predictor	Coef	SE Coef	T	p
Constant	66.10	32.06	2.06	0.094
X	0.4023	0.2276	1.77	0.137

S = 12.62 R-sq = 38.5% R-sq(adj) = 26.1%
Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	497.2	497.2	3.12	0.137
Residual Error	5	795.7	159.1		
Total	6	1292.9			

Unusual Observations

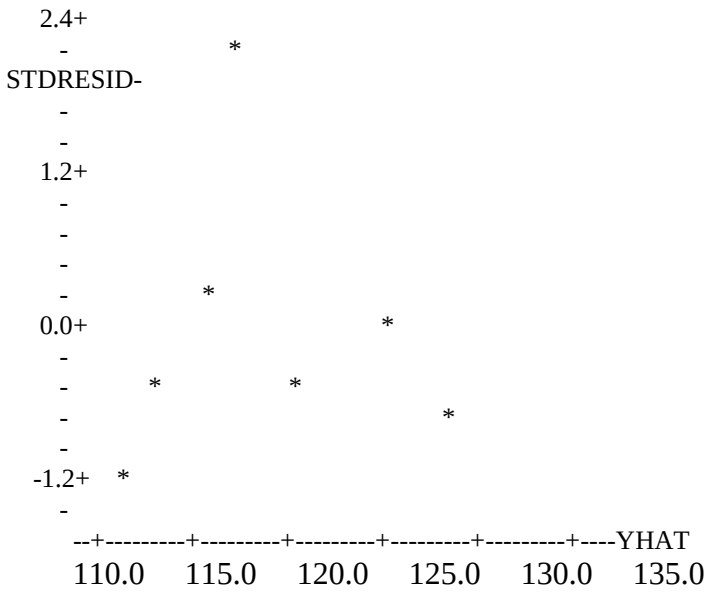
Obs.	X	Y	Fit	SEFit	Residual	St.Resid
1	135	145.00	120.42	4.87	24.58	2.11R

R denotes an observation with a large standardized residual.

The standardized residuals are: 2.11, -1.08, .14, -.38, -.78, -.04, -.41

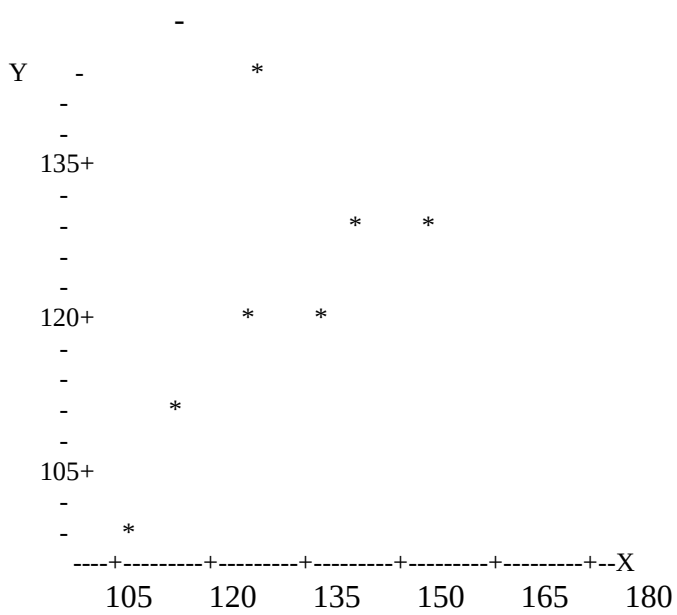
The first observation appears to be an outlier since it has a large standardized residual.

b.



The standardized residual plot indicates that the observation $x = 135$, $y = 145$ may be an outlier; note that this observation has a standardized residual of 2.11.

c. The scatter diagram is shown below



The scatter diagram also indicates that the observation $x = 135$, $y = 145$ may be an outlier; the implication is that for simple linear regression an outlier can be identified by looking at the scatter diagram.

30. a. The Minitab output is shown below:

The regression equation is
Y = 13.0 + 0.425 X

Predictor	Coef	SE Coef	T	p
Constant	13.002	2.396	5.43	0.002
X	0.4248	0.2116	2.01	0.091

S = 3.181 R-sq = 40.2% R-sq(adj) = 30.2%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	40.78	40.78	4.03	0.091
Residual Error	6	60.72	10.12		
Total	7	101.50			

Unusual Observations

Obs.	X	Y	Fit	Stdev.Fit	Residual	St.Resid
7	12.0	24.00	18.10	1.20	5.90	2.00R
8	22.0	19.00	22.35	2.78	-3.35	-2.16RX

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

The standardized residuals are: -1.00, -.41, .01, -.48, .25, .65, -2.00, -2.16

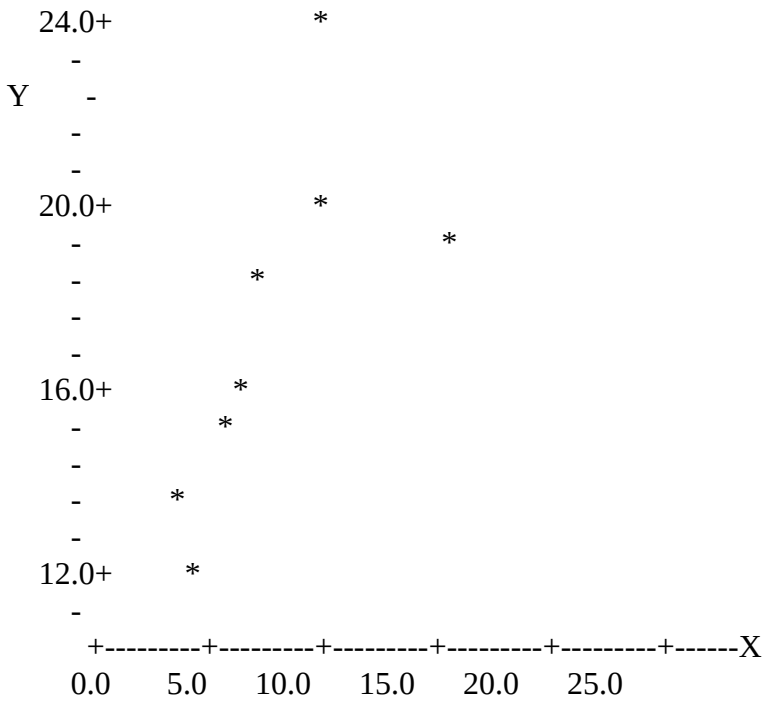
The last two observations in the data set appear to be outliers since the standardized residuals for these observations are 2.00 and -2.16, respectively.

b. Using MINITAB, we obtained the following leverage values:

.28, .24, .16, .14, .13, .14, .14, .76

MINITAB identifies an observation as having high leverage if $h_i > 6/n$; for these data, $6/n = 6/8 = .75$. Since the leverage for the observation $x = 22$, $y = 19$ is .76, MINITAB would identify observation 8 as a high leverage point. Thus, we conclude that observation 8 is an influential observation.

c.



The scatter diagram indicates that the observation $x = 22, y = 19$ is an influential observation.

Supplementary Exercises:

31. Does a high value of r^2 imply that two variables are causally related? Explain.
32. In your own words, explain the difference between an interval estimate of the mean value of y for a given x and an interval estimate for an individual value of y for a given x.

33. What is the purpose of testing whether $\beta_1 = 0$? If we reject $\beta_1 = 0$, does it imply a good fit?

34. *Bloomberg Personal Finance* (July/August 2001) reported the market beta for Texas Instruments was 1.46. Market betas for individual stocks are determined by simple linear regression. For each stock, the dependent variable is its quarterly percentage return (capital appreciation plus dividends) minus the percentage return that could be obtained from a risk-free investment (the Treasury Bill rate is used as the risk-free rate). The independent variable is the quarterly percentage return (capital appreciation plus dividends) for the stock market (S&P 500) minus the percentage return from a risk-free investment. An estimated regression equation is developed with quarterly data; the market beta for the stock is the slope of the estimated regression equation (b_1). The value of the market beta is often interpreted as a measure of the risk associated with the stock. Market betas greater than 1 indicate that the stock is more volatile than the market average; market betas less than 1 indicate that the stock is less volatile than the market average. Suppose that the following figures are the differences between the percentage return and the risk-free return for 10 quarters for the S&P 500 and Horizon Technology.

S&P 500	Horizon
1.2	-0.7
-2.5	-2.0
-3.0	-5.5
2.0	4.7
5.0	1.8
1.2	4.1
3.0	2.6
-1.0	2.0
.5	-1.3
2.5	5.5

a. Develop an estimated regression equation that can be used to determine the market beta for Horizon Technology. What is Horizon Technology's market beta?

b. Test for a significant relationship at the 0.05 level of significance.

c. Did the estimated regression equation provide a good fit? Explain.

d. Use the market betas of Texas Instruments and Horizon Technology to compare the risk associated with the two stocks.

35. File "HighLow"

The daily high and low temperatures for 20 cities follow (*USA Today*, May 9, 2000).

	Low	High
Athens	54	75
Bangkok	74	92
Cairo	57	84
Copenhagen	39	64
Dublin	46	64
Havana	68	86
Hong Kong	72	81
Johannesburg	50	61
London	48	73
Manila	75	93
Melbourne	50	66
Montreal	52	64
Paris	55	77
Rio de Janeiro	61	80
Rome	54	81
Seoul	50	64
Singapore	75	90
Sydney	55	68
Tokyo	59	79
Vancouver	43	57

a. Develop a scatter diagram with low temperature on the horizontal axis and high temperature on the vertical axis.

b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

c. Develop the estimated regression equation that could be used to predict the high temperature given the low temperature.

d. Test for a significant relationship at the 0.05 level of significance.

e. Did the estimated regression equation provide a good fit? Explain.

f. What is the value of the sample correlation coefficient?

36. Jensen Tyre & Auto is in the process of deciding whether to purchase a maintenance contract for its new computer wheel alignment and balancing machine. Managers feel that maintenance expense should be related to usage, and they collected the following information on weekly usage (hours) and annual maintenance expense (in hundreds of euros).

Weekly Usage	**Annual Maintenance Expense**
13	17.0
10	22.0
20	30.0
28	37.0
32	47.0
17	30.5
24	32.5
31	39.0
40	51.5
38	40.0

a. Develop the estimated regression equation that relates annual maintenance expense to weekly usage.

b. Test the significance of the relationship in part (a) at a 0.05 level of significance.

c. Jensen expects to use the new machine 30 hours per week. Develop a 95% prediction interval for the company's annual maintenance expense.

d. If the maintenance contract costs €3000 per year, would you recommend purchasing it? Why or why not?

37. In a manufacturing process, the assembly line speed (metre per minute) was thought to affect the number of defective parts found during the inspection process. To test this theory, managers devised a situation in which the same batch of parts was inspected visually at a variety of line speeds. They collected the following data.

Line Speed	Number of Defective Parts Found
20	21
20	19
40	15
30	16
60	14
40	17

a. Develop the estimated regression equation that relates line speed to the number of defective parts found.

b. At a 0.05 level of significance, determine whether line speed and number of defective parts found are related.

c. Did the estimated regression equation provide a good fit to the data?

d. Develop a 95% confidence interval to predict the mean number of defective parts for a line speed of 50 metres per minute.

38. A sociologist was hired by a large city hospital to investigate the relationship between the number of unauthorized days that employees are absent per year and the distance (kilometres) between home and work for the employees. A sample of 10 employees was chosen, and the following data were collected.

Distance to Work	Number of Days Absent
1	8
3	5
4	8
6	7
8	6
10	3
12	5
14	2
14	4
18	2

a. Develop a scatter diagram for these data. Does a linear relationship appear reasonable? Explain.

b. Develop the least squares estimated regression equation.

c. Is there a significant relationship between the two variables? Use $\alpha = 0.05$.

d. Did the estimated regression equation provide a good fit? Explain.

e. Use the estimated regression equation developed in part (b) to develop a 95% confidence interval for the expected number of days absent for employees living 5 kilometres from the company.

39. The regional transit authority for a major metropolitan area wants to determine whether there is any relationship between the age of a bus and the annual maintenance cost. A sample of 10 buses resulted in the following data.

Age of Bus (years)	**Maintenance Cost (€)**
1	350
2	370
2	480
2	520
2	590
3	550
4	750
4	800
5	790
5	950

a. Develop the least squares estimated regression equation.

b. Test to see whether the two variables are significantly related with $\alpha = 0.05$.

c. Did the least squares line provide a good fit to the observed data? Explain.

d. Develop a 95% prediction interval for the maintenance cost for a specific bus that is 4 years old.

40. A marketing professor at Givens College is interested in the relationship between hours spent studying and total points earned in a course. Data collected on 10 students who took the course last quarter follow.

Hours Spent Studying	**Total Points Earned**
45	40
30	35
90	75
60	65
105	90
65	50
90	90
80	80
55	45
75	65

a. Develop an estimated regression equation showing how total points earned is related to hours spent studying.

b. Test the significance of the model with $\alpha = 0.05$.

c. Predict the total points earned by Pat Sweeney. He spent 95 hours studying.

d. Develop a 95% prediction interval for the total points earned by Pat Sweeney.

41. In an effort to determine the relationship between annual wages for employees and the number of days absent from work due to sickness, a large corporation studied the personnel records for a random sample of employees. The resultant data are as follows:

Employee	Annual wages (€000)	Days absent
1	15.7	4
2	17.2	3
3	13.8	6
4	24.2	5
5	15.0	3
6	12.7	12
7	13.8	5
8	18.7	1
9	10.8	12
10	11.8	11
11	25.4	2
12	17.2	4

Carry out an appropriate analysis of this information. What inferences do you draw?

42. The sales director of a chain of hardware stores would like to investigate the relationship between floor area of its stores and their annual sales in the previous year. Data have been collected from a random sample of 12 stores as follows:

Annual sales (€'00,000's)	Sales area (000's m^2)
118	9
116	7
165	15
157	13
165	14
130	7
138	11
132	14
150	10
220	25
170	12.5
180	15

a. Plot a scattergram for these data. Following on, calculate an appropriate least squares regression line. By superimposing the line over the scattergram, comment on the quality of the fit.

b. Determine the sample correlation coefficient and corresponding coefficient of determination. How would you interpret these values?

c. Use the regression line in a) to predict the annual sales for a store whose floor area is 16,000 m^2. Calculate a 99% confidence interval for your prediction.

43. File "Shrinkage"

Data on textile shrinkage are available as follows:

Percentage shrinkage in samples of cloth after washing, in directions along and across the cloth

Along (x)	Across (y)	Along (x)	Across (y)
12	5	7	5
4	2	12	7
10	5	18	10
10	8	14	7
11	6	14	8
10	8	8	4
6	3	11	6
6	4	17	8
6	3	21	11
13	5	12	9

a. Draw a scattergram for the data. Calculate the least squares regression line of x on y.

By superimposing the line over the plot, comment on the effectiveness of the latter modelling approach.

b. Determine the sample correlation coefficient and corresponding coefficient of determination. How would you interpret these values?

c. A roll of cloth is sampled by cutting a narrow test strip across the roll. The strip proves to have a percentage shrinkage of 7. Use your model in i) to predict the percentage shrinkage for this sample along the cloth. Calculate a 95% confidence interval for your prediction.

d. There is a requirement for a piece to be cut from the roll of cloth which would be expected to shrink to 10 cm square after washing. Describe how this piece should be cut.

44. A doctor has access to historical data as follows:

	Vehicles per 100 population	Road deaths per 100 000 population
Great Britain	31	14
Belgium	32	29
Denmark	30	22
France	47	32
Germany	30	25
Irish Republic	19	20
Italy	36	21
Netherlands	40	22
Canada	47	30
USA	58	35

a. Carry out an appropriate analysis of this information stating all relevant assumptions. b. What inferences do you draw?

45. File “CarIncome”

Data are available for two variables over a twelve year period as follows:

Cars	**Income**
2016.2	385796
2210.2	405997
2304.5	423510
2005.2	439384
1600.2	446102
1598.0	462638
1776.5	476613
1902.0	482708
1938.1	495337
2018.3	506145
2157.1	525321
2288.5	525507

Source: Central Statistical Office

(Here the **Cars** variable is defined as New registrations of cars in thousands as a monthly average and **Income** as Real household disposable income (£m) at 1995 prices.)

a. What sort of causal relationship would you expect between these variables?
b. Draw a scattergram for the data and comment on any patterns that this might reveal. Does it support your earlier expectation?
c. Calculate the coefficient of determination between Cars and Income. How would you interpret this? See below
d. Estimate a linear regression model for the data. By graphing and / or otherwise, comment on its value as a forecasting aid.

Supplementary Exercises Solutions:

31. No. Regression or correlation analysis can never prove that two variables are causally related.

32. The estimate of a mean value is an estimate of the average of all *y* values associated with the same *x*. The estimate of an individual *y* value is an estimate of only one of the *y* values associated with a particular *x*.

33. The purpose of testing whether $\beta_1 = 0$ is to determine whether or not there is a significant relationship between *x* and *y*. However, rejecting $\beta_1 = 0$ does not necessarily imply a good fit. For example, if $\beta_1 = 0$ is rejected and r^2 is low, there is a statistically significant relationship between *x* and *y* but the fit is not very good.

34. a. The Minitab output is shown below:

The regression equation is
Horizon = 0.275 + 0.950 S&P 500

Predictor	Coef	SE Coef	T	P
Constant	0.2747	0.9004	0.31	0.768
S&P 500	0.9498	0.3569	2.66	0.029

S = 2.664 R-Sq = 47.0% R-Sq(adj) = 40.3%

Analysis of Variance

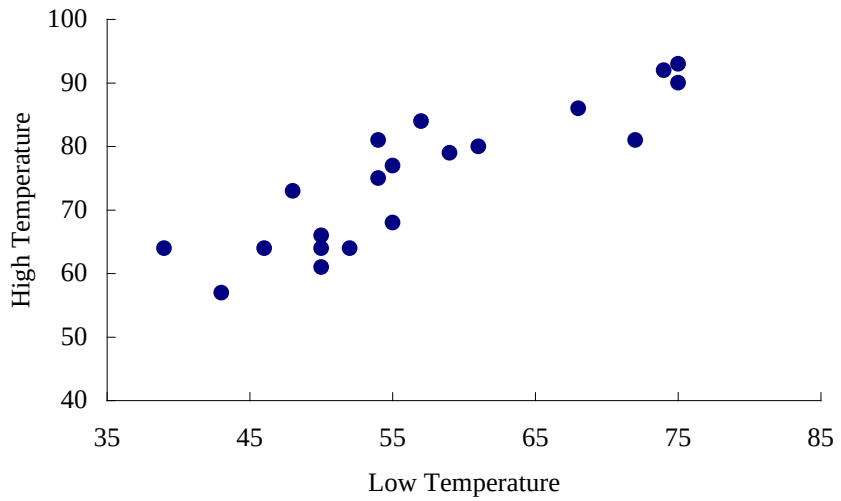
Source	DF	SS	MS	F	P
Regression	1	50.255	50.255	7.08	0.029
Residual Error	8	56.781	7.098		
Total	9	107.036			

b. Since the *p*-value = 0.029 is less than α = .05, the relationship is significant.

c. r^2 = .470. The least squares line does not provide a very good fit.

d. Texas Instruments has higher risk with a market beta of 1.25.

35. a.



b. It appears that there is a positive linear relationship between the two variables.

c. The Minitab output is shown below:

The regression equation is
High = 23.9 + 0.898 Low

Predictor	Coef	SE Coef	T	P
Constant	23.899	6.481	3.69	0.002
Low	0.8980	0.1121	8.01	0.000

S = 5.285 R-Sq = 78.1% R-Sq(adj) = 76.9%

Analysis of Variance

Source	DF	SS	MS	F	P

Regression	1	1792.3	1792.3	64.18	0.000
Residual Error	18	502.7	27.9		
Total	19	2294.9			

d. Since the *p*-value corresponding to $F = 64.18 = .000 < \alpha = .05$, the relationship is significant.

e. $r^2 = .781$; a good fit. The least squares line explained 78.1% of the variability in high temperature.

f. $r = \sqrt{.781} = +.88$

36. The Minitab output is shown below:

The regression equation is
Y = 10.5 + 0.953 X

Predictor	Coef	SE Coef	T	p
Constant	10.528	3.745	2.81	0.023
X	0.9534	0.1382	6.90	0.000

S = 4.250 R-sq = 85.6% R-sq(adj) = 83.8%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	860.05	860.05	47.62	0.000
Residual Error	8	144.47	18.06		
Total	9	1004.53			

Fit	Stdev.Fit	95% C.I.	95% P.I.
39.13	1.49	(35.69, 42.57)	(28.74, 49.52)

a. $\hat{y} = 10.5 + .953\,x$

b. Since the *p*-value corresponding to $F = 47.62 = .000 < \alpha = .05$, we reject $H_0: \beta_1 = 0$.

c. The 95% prediction interval is 28.74 to 49.52 or €2874 to €4952

d. Yes, since the expected expense is €3913.

37. a. The Minitab output is shown below:

The regression equation is
Defects = 22.2 - 0.148 Speed

Predictor	Coef	SE Coef	T	P
Constant	22.174	1.653	13.42	0.000
Speed	-0.14783	0.04391	-3.37	0.028

S = 1.489 R-Sq = 73.9% R-Sq(adj) = 67.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	25.130	25.130	11.33	0.028
Residual Error	4	8.870	2.217		
Total	5	34.000			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	14.783	0.896	(12.294, 17.271)	(9.957, 19.608)

b. Since the *p*-value corresponding to $F = 11.33 = .028 < \alpha = .05$, the relationship is significant.

c. $r^2 = .739$; a good fit. The least squares line explained 73.9% of the variability in the number of defects.

d. Using the Minitab output in part (a), the 95% confidence interval is 12.294 to 17.271.

38. a. There appears to be a negative linear relationship between distance to work and number of days absent.

b. The MINITAB output is shown below:

```
The regression equation is
Y = 8.10 - 0.344 X

Predictor       Coef     SE Coef        T       p
Constant      8.0978      0.8088    10.01   0.000
X          -0.34420     0.07761    -4.43   0.002

S = 1.289       R-sq = 71.1%     R-sq(adj) = 67.5%

Analysis of Variance

SOURCE        DF         SS         MS        F       p
Regression     1     32.699     32.699    19.67   0.002
Residual Error 8     13.301      1.663
Total          9     46.000

    Fit  Stdev.Fit        95% C.I.        95% P.I.
  6.377      0.512   (  5.195,  7.559)  (  3.176,  9.577)
```

c. Since the *p*-value corresponding to $F = 419.67$ is .002 $<$ □ = .05. We reject $H_0 : \beta_1 = 0$.

d. $r^2 = .711$. The estimated regression equation explained 71.1% of the variability in *y*; this is a reasonably good fit.

e. The 95% confidence interval is 5.195 to 7.559 or approximately 5.2 to 7.6 days.

39. a. Let *X* = the age of a bus and *Y* = the annual maintenance cost.

The Minitab output is shown below:

```
The regression equation is
Y = 220 + 132 X

Predictor       Coef     SE Coef        T       p
Constant      220.00       58.48     3.76   0.006
X             131.67       17.80     7.40   0.000

S = 75.50       R-sq = 87.3%     R-sq(adj) = 85.7%

Analysis of Variance

SOURCE        DF         SS         MS        F       p
Regression     1     312050     312050    54.75   0.000
Residual Error 8      45600       5700
Total          9     357650

    Fit  Stdev.Fit        95% C.I.        95% P.I.
  746.7       29.8   (  678.0,  815.4)  (  559.5,  933.9)
```

b. Since the *p*-value corresponding to $F = 54.75$ is .000 < α = .05, we reject H_0: $\beta_1 = 0$.

c. $r^2 = .873$. The least squares line provided a very good fit.

d. The 95% prediction interval is 559.5 to 933.9 or €559.50 to €933.90

40. a.Let X = hours spent studying and Y = total points earned

The Minitab output is shown below:

The regression equation is
Y = 5.85 + 0.830 X

Predictor	Coef	SE Coef	T	p
Constant	5.847	7.972	0.73	0.484
X	0.8295	0.1095	7.58	0.000

S = 7.523 R-sq = 87.8% R-sq(adj) = 86.2%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	3249.7	3249.7	57.42	0.000
Residual Error	8	452.8	56.6		
Total	9	3702.5			

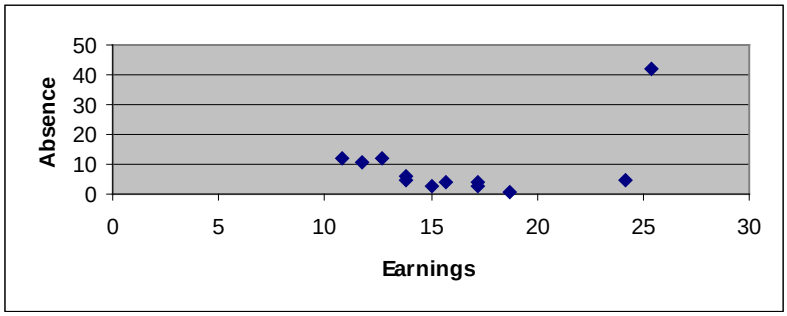
Fit	Stdev.Fit	95% C.I.	95% P.I.
84.65	3.67	(76.19, 93.11)	(65.35, 103.96)

b. Since the *p*-value corresponding to $F = 57.42$ is .000 < α = .05, we reject H_0: $\beta_1 = 0$.

c. 84.65 points

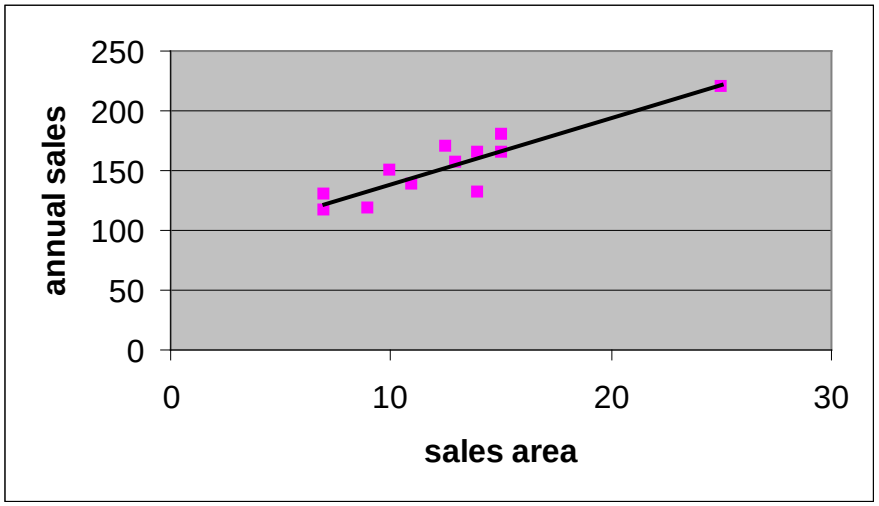
d. The 95% prediction interval is 65.35 to 103.96

41. From EXCEL:



A linear model does not look very convincing for the data from this scatter plot. Correspondingly the sample correlation coefficient r = 0.417. Again this indicates a relatively poor linear association between Absence and Earnings.

42. a. From EXCEL



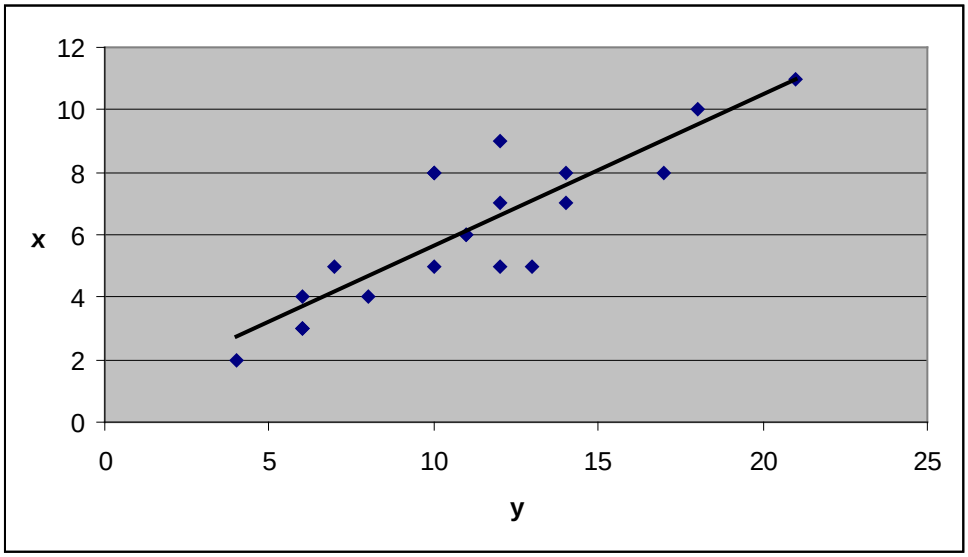
$\hat{y}$ = 5.5263x + 83.186 where y = Annual sales, x = Sales area

The scattergram suggests the least squares regression line provides a very good fit.

b. The coefficient of determination is $r^2 = 0.8082$. We deduce the estimated regression equation explained 80.82% of the variability in *y*; again this indicates a very good fit. The sample correlation r = $+\sqrt{0.8082}$ =0.899 (since the regression line has a positive slope). So y and x are highly positively correlated.

c. For a store whose Sales area is 16,000 m^2, Annual sales (€'00,000's) would be predicted as 171.61 or €17,161,000. The 99% confidence interval correspondingly is €15,635,100 to €18,686,900

43. a. From EXCEL



$$\hat{x} = 1.537 + 1.542y$$

b. The coefficient of determination is $r^2 = 0.7443$. We deduce the estimated regression equation explained 74.43% of the variability in *y*; again this indicates a very good fit. The sample correlation r = $+\sqrt{0.7443} = 0.863$ (since the regression line has a positive slope). So y and x are highly positively correlated.

c.

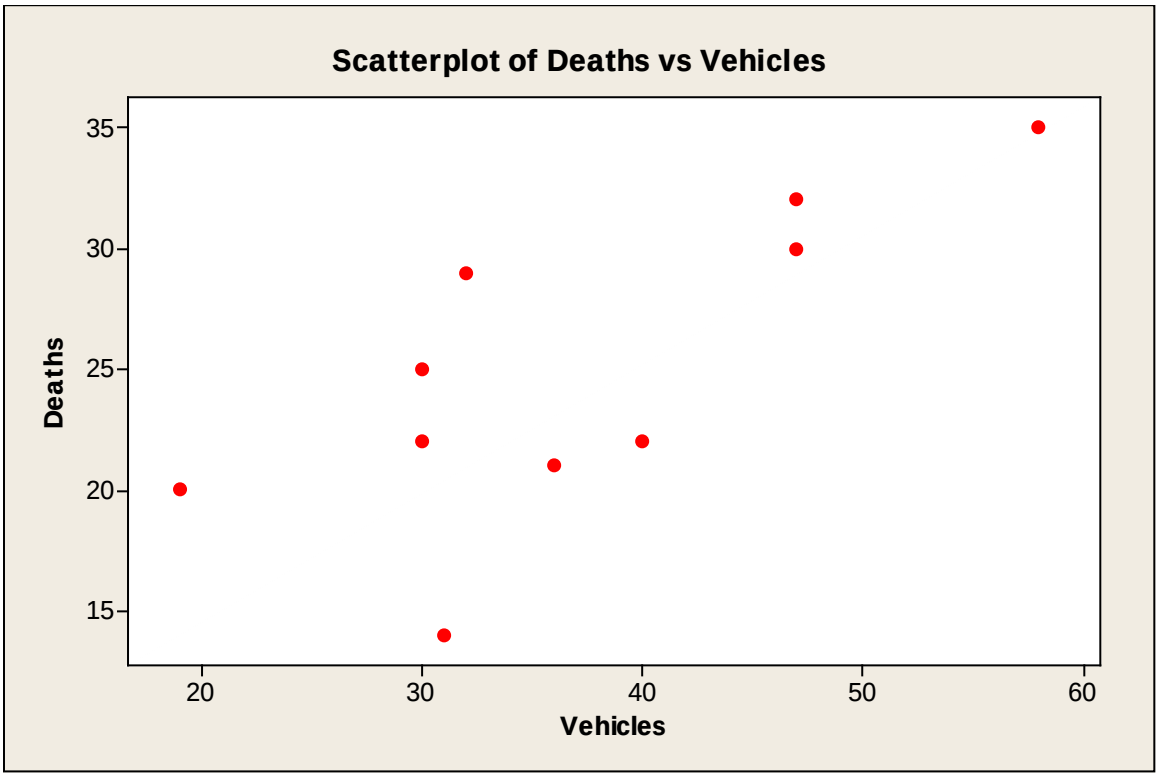
i. 12.334

ii. (11.210, 13.457)

d. Average shrinkage across the cloth is $\bar{y} =$ 6.2% Similarly $\bar{x} =$ 11.1% which can be predicted from the regression equation in (a). Thus the

cloth should initially be cut with sides measuring across 10 *100/ (100-6.2) = 10.66 cm and along 10*100/ (100-11.1) = 11.25 cm

44. a. From MINITAB, a scattergram of Deaths against Vehicles is as follows:



This plot suggests linear regression modelling may prove productive.

Relevant MINITAB output is shown below:

The regression equation is
Deaths = 9.27 + 0.425 Vehicles

Predictor	Coef	SE Coef	T	P
Constant	9.273	5.194	1.79	0.112
Vehicles	0.4250	0.1349	3.15	0.014

S = 4.54325 R-Sq = 55.4% R-Sq(adj) = 49.8%

Analysis of Variance

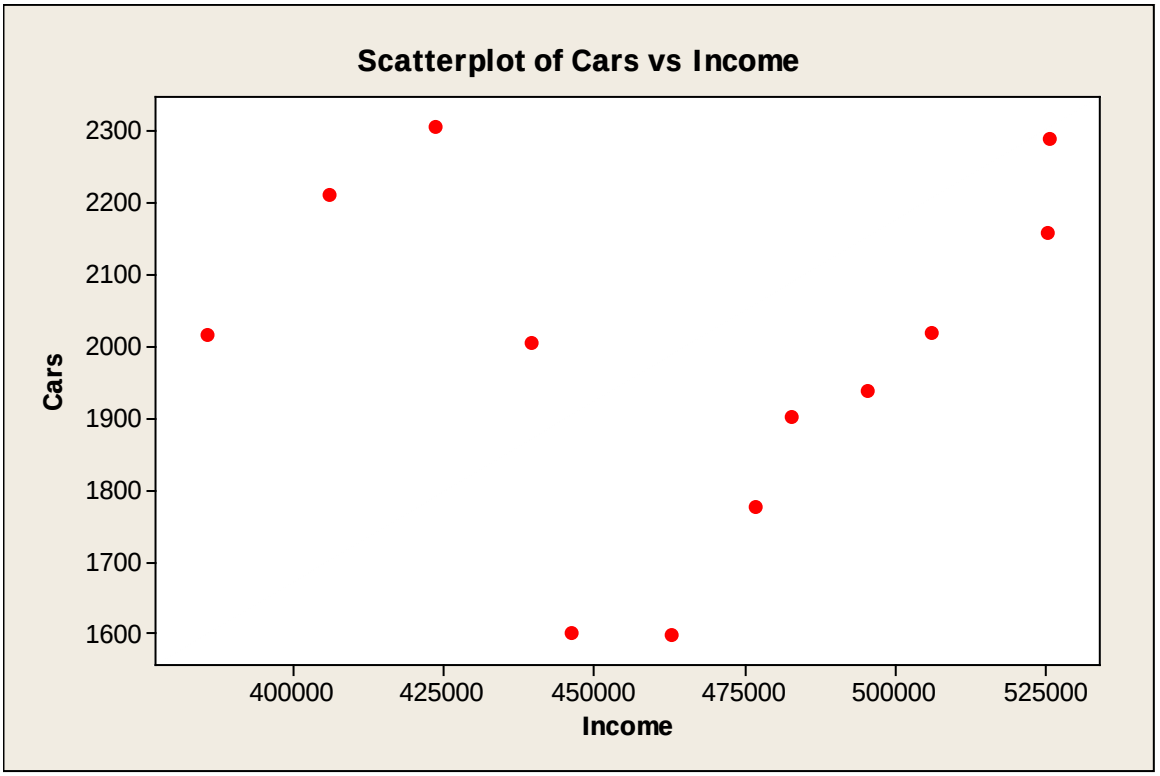
Source	DF	SS	MS	F	P
Regression	1	204.87	204.87	9.93	0.014
Residual Error	8	165.13	20.64		
Total	9	370.00			

b. A significant linear model exists for the data. This is borne out by the pvalue for the F statistic of 0.014 which is less than $\alpha = 0.05$. Hence we reject H_0: $\beta_1 = 0$.

$r^2 = 55.4\%$. The estimated regression equation explained 55.4% of the variability in *y*; this is a reasonable fit.

45. a. As Income gets higher new Car registrations should increase and vice versa. Possibly this would take place at a uniform rate.

b.



The plot appears fairly random and the expectation from (a) is not supported.

c $R^2 = 0.0008$. This value suggests the least squares linear regression model (see (d)) has little explanatory value. More specifically, only 0.08% of the variability in Cars can be explained by the linear relationship between the Income and Cars.

d. $\hat{y}$ = 0.0002x + 1914.4 where y = Cars, x = Income. The regression line superimposed on the scattergram below is almost horizontal confirming the lack of a linear relationship.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Fifteen

Multiple Regression

Textbook Exercises (1-33)

Textbook Exercise Solutions

Supplementary Exercises (34-48)

Supplementary Exercise Solutions

Chapter 15: Multiple Regression

Textbook Exercises:

1. The estimated regression equation for a model involving two independent variables and ten observations follows.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

a. Interpret b_1 and b_2 in this estimated regression equation.

b. Estimate Y when $X_1 = 180$ and $X_2 = 310$.

2. Consider the following data for a dependent variable Y and two independent variables, X_1 and X_2.

x_1	x_2	y
30	12	94
47	10	108
25	17	112
51	16	178
40	5	94
51	19	175
74	7	170
36	12	117
59	13	142
76	16	211

a. Develop an estimated regression equation relating Y to X_1.

Estimate Y if $X_1 = 45$.

b. Develop an estimated regression equation relating Y to X_2.

Estimate Y if $X_2 = 15$.

c. Develop an estimated regression equation relating Y to X_1 and X_2.

Estimate Y if $X_1 = 45$ and $X_2 = 15$.

3. In a regression analysis involving 30 observations, the following estimated regression equation was obtained.

$$\hat{y} = 17.6 + 0\ 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

a. Interpret b_1, b_2, b_3, and b_4 in this estimated regression equation.

b. Estimate Y when $X_1 = 10$, $X_2 = 5$, $X_3 = 1$, and $X_4 = 2$.

4. The stack loss plant data of Brownlee (1965) contains 21 days of measurements from a plant's oxidation of ammonia to nitric acid. The nitric oxide pollutants are captured in an absorption tower. Details of variables are as follows:

Y = LOSS = 10 times the percentage of ammonia going into the plant that escapes from the absorption column

X_1 = AIRFLOW = Rate of operation of the plant

X_2 = TEMP = Cooling water temperature in the absorption tower

X_3 = ACID = Acid concentration of circulating acid minus 50 times.

The following estimated regression equation relating LOSS to AIRFLOW and TEMP was given.

$$\hat{y} = -50.359 + 0.671x_1 + 1.295x_2$$

a. Estimate sales resulting from an AIRFLOW of 60 and a TEMP of 20.
b. Interpret b_1 and b_2 in this estimated regression equation.

5. The owner of Toulon Theatres would like to estimate weekly gross revenue as a function of advertising expenditures. Historical data for a sample of eight weeks follow.

Weekly gross revenue (€000s)	Television advertising (€000s)	Newspaper advertising (€000s)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

a. Develop an estimated regression equation with the amount of television advertising as the independent variable.

b. Develop an estimated regression equation with both television advertizing and newspaper advertising as the independent variables.

c. Is the estimated regression equation coeffi cient for television advertising expenditures the same in part (a) and in part (b)? Interpret the coeffi cient in each case.

d. What is the estimate of the weekly gross revenue for a week when €3500 is spent on television advertising and €1800 is spent on newspaper advertising?

6. The following table gives the annual return, the safety rating (0 = riskiest, 10 = safest), and the annual expense ratio for 20 foreign funds.

	Annual safety rating	Expense ratio (%)	Annual return (%)
Accessor Int'l Equity 'Adv'	7.1	1.59	49
Aetna 'I' International	7.2	1.35	52
Amer Century Int'l Discovery 'Inv'	6.8	1.68	89
Columbia International Stock	7.1	1.56	58
Concert Inv 'A' Int'l Equity	6.2	2.16	131
Dreyfus Founders Int'l Equity 'F'	7.4	1.80	59
Driehaus International Growth	6.5	1.88	99
Excelsior 'Inst' Int'l Equity	7.0	0.90	53
Julius Baer International Equity	6.9	1.79	77
Marshall International Stock 'Y'	7.2	1.49	54
MassMutual Int'l Equity 'S'	7.1	1.05	57
Morgan Grenfell Int'l Sm Cap 'Inst'	7.7	1.25	61
New England 'A' Int'l Equity	7.0	1.83	88
Pilgrim Int'l Small Cap 'A'	7.0	1.94	122
Republic International Equity	7.2	1.09	71
Sit International Growth	6.9	1.50	51
Smith Barney 'A' Int'l Equity	7.0	1.28	60
State St Research 'S' Int'l Equity	7.1	1.65	50
Strong International Stock	6.5	1.61	93
Vontobel International Equity	7.0	1.50	47

a. Develop an estimated regression equation relating the annual return to the safety rating and the annual expense ratio.

b. Estimate the annual return for a firm that has a safety rating of 7.5 and annual expense ratio of 2.

7. In exercise 1, the following estimated regression equation based on ten observations was presented.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

The values of SST and SSR are 6724.125 and 6216.375, respectively.

a. Find SSE.
b. Compute R_2.
c. Compute *Adj* R_2.
d. Comment on the goodness of fit.

8. In exercise 2, ten observations were provided for a dependent variable *Y* and two independent variables X_1 and X_2; for these data SST = 15 182.9, and SSR = 14 052.2.
a. Compute R_2.
b. Compute *Adj* R_2.
c. Does the estimated regression equation explain a large amount of the variability in the data? Explain.

9. In exercise 3, the following estimated regression equation based on 30 observations was presented.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

The values of SST and SSR are 1805 and 1760, respectively.

a. Compute R_2.
b. Compute *Adj* R_2.
c. Comment on the goodness of fit.

10. In Exercise 4, the following estimated regression equation relating LOSS (Y) to AIRLOW (X_1) and TEMP (X_2) was given.

$$\hat{y} = -50.359 + 0.671x_1 + 1.295x_2$$

For these data SST = 2069.238 and SSR = 1880.443.

a. For the estimated regression equation given, compute R^2.

b. Compute *Adj* R^2.

c. Does the model appear to explain a large amount of variability in the data? Explain.

11. In exercise 5, the owner of Toulon Theatres used multiple regression analysis to predict gross revenue (Y) as a function of television advertising (X_1) and newspaper advertising (X_2). The estimated regression equation was

$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

The computer solution provided SST = 25.5 and SSR = 23.435.

a. Compute and interpret R_2 and *Adj* R_2.

b. When television advertising was the only independent variable, $R2 = 0.653$ and *Adj* $R_2 = 0.595$. Do you prefer the multiple regression results? Explain.

12. In exercise 1, the following estimated regression equation based on ten observations was presented.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

Here SST = 6724.125, SSR = 6216.375, $s_{b_1} = 0.0813$ and $s_{b_2} = 0.0567$.

a. Compute MSR and MSE.

b. Compute F and perform the appropriate F test. Use $\alpha = 0.05$.

c. Perform a t test for the significance of β_1. Use $\alpha = 0.05$.

d. Perform a t test for the significance of β_2. Use $\alpha = 0.05$.

13. Refer to the data presented in exercise 2. The estimated regression equation for these data is:

$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

Here SST = 15 182.9, SSR = 14 052.2, $s_{b_1} = 0.2471$ and $s_{b_2} = 0.9484$.

a. Test for a significant relationship among X_1, X_2 and Y. Use $\alpha = 0.05$.

b. Is β_1 significant? Use $\alpha = 0.05$.

c. Is β_2 significant? Use $\alpha = 0.05$.

14. The following estimated regression equation was developed for a model involving two independent variables.

$y = 40.7 + 8.63x_1 + 2.71x_2$

After X_2 was dropped from the model, the least squares method was used to obtain an estimated regression equation involving only X_1 as an independent variable.

$y = 42.0 + 9.01x_1$

a. Give an interpretation of the coefficient of $X1$ in both models.

b. Could multicollinearity explain why the coefficient of $X1$ differs in the two models? If so, how?

15. In Exercise 4, the following estimated regression equation relating LOSS (Y) to AIRFLOW (X_1) and TEMP (X_2) was given.

$$\hat{y} = -50.359 + 0.671x_1 + 1.295x_2$$

For these data SST = 2069.238 and SSR = 1880.443.

Compute SSE, MSE and MSR.

a. Use an F test and a 0.05 level of significance to determine whether there is a relationship among the variables.

16. Refer to exercise 5.

a. Use $\alpha = 0.01$ to test the hypotheses

$H_0: \beta_1 = \beta_2 = 0$
$H_1: \beta_1$ and/or β_2 is not equal to zero

for the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where

X_1 = television advertising (€1000s)
X_2 = newspaper advertising (€1000s)

b. Use $\alpha = 0.05$ to test the significance of β_1. Should X_1 be dropped from the model?
c. Use $\alpha = 0.05$ to test the significance of β_2. Should X_2 be dropped from the model?

17. In exercise 1, the following estimated regression equation based on ten observations was presented.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

a. Develop a point estimate of the mean value of Y when $X_1 = 180$ and $X_2 = 310$.
b. Develop a point estimate for an individual value of Y when $X_1 = 180$ and $X_2 = 310$.

18. Refer to the data in exercise 2. The estimated regression equation for those data is

$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

a. Develop a 95 per cent confidence interval for the mean value of Y when $X_1 = 45$ and $X_2 = 15$.
b. Develop a 95 per cent prediction interval for Y when $X_1 = 45$ and $X_2 = 15$.

19. In exercise 5, the owner of Toulon Theatres used multiple regression analysis to predict gross revenue (Y) as a function of television advertising (X_1) and newspaper advertising (X_2).The estimated regression equation was:

$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

a. What is the gross revenue expected for a week when €3500 is spent on television advertising ($X_1 = 3.5$) and €1800 is spent on newspaper advertising ($X_2 = 1.8$)?

b. Provide a 95 per cent confidence interval for the mean revenue of all weeks with the expenditures listed in part (a).

c. Provide a 95 per cent prediction interval for next week's revenue, assuming that the advertising expenditures will be allocated as in part (a).

20. Consider a regression study involving a dependent variable Y, a quantitative independent variable X_1 and a qualitative variable with two levels (level 1 and level 2).

a. Write a multiple regression equation relating X_1 and the qualitative variable to Y.

b. What is the expected value of Y corresponding to level 1 of the qualitative variable?

c. What is the expected value of Y corresponding to level 2 of the qualitative variable?

d. Interpret the parameters in your regression equation.

21. Consider a regression study involving a dependent variable Y, a quantitative independent variable X_1, and a qualitative independent variable with three possible levels (level 1, level 2 and level 3).

a. How many dummy variables are required to represent the qualitative variable?

b. Write a multiple regression equation relating X_1 and the qualitative variable to Y.

c. Interpret the parameters in your regression equation.

22. Management proposed the following regression model to predict the effect of physical exercise on pulse in an experiment involving 92 participants

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

where

Y = Pulse 2 = second pulse reading taken at end of experiment

x_1 = Pulse 1= initial (resting) pulse reading

x_2 = Ran = 1 if individual ran on the spot for 1 minute, 2 if they did not (this was decided randomly)

x_3 = Sex = 1 if male, 2 if female

The following estimated regression equation was developed using MINITAB:

$$\hat{y} = 42.62 + 0.812x_1 - 20.1x_2 + 7.8x_3$$

a. What is the amount of the expected value of Pulse 2 attributable to x_3?
b. Predict Pulse 2 for a female who ran on the spot for 1 minute and had an initial pulse reading of 70 bpm.
c. Predict Pulse 2 for a male who did not run on the spot for 1 minute and had an initial pulse reading of 60 bpm.

23. Refer to the Johansson Filtration problem introduced in this section. Suppose that in addition to information on the number of months since the machine was serviced and whether a mechanical or an electrical failure had occurred, the managers obtained a list showing which engineer performed the service. The revised data follow.

Repair time in hours	Months since last service	Type of repair	Engineer
2.9	2	Electrical	Heinz Kolb
3.0	6	Mechanical	Heinz Kolb
4.8	8	Electrical	Wolfgang Linz
1.8	3	Mechanical	Heinz Kolb
2.9	2	Electrical	Heinz Kolb
4.9	7	Electrical	Wolfgang Linz
4.2	9	Mechanical	Wolfgang Linz
4.8	8	Mechanical	Wolfgang Linz
4.4	4	Electrical	Wolfgang Linz
4.5	6	Electrical	Heinz Kolb

a. Ignore for now the months since the last maintenance service (X_1) and the engineer who performed the service. Develop the estimated simple linear regression equation to predict the repair time (Y) given the type of repair (X_2). Recall that $X_2 = 0$ if the type of repair is mechanical and 1 if the type of repair is electrical.

b. Does the equation that you developed in part (a) provide a good fit for the observed data? Explain.

c. Ignore for now the months since the last maintenance service and the type of repair associated with the machine. Develop the estimated simple linear regression equation to predict the repair time given the engineer who performed the service. Let $X_3 = 0$ if Heinz Kolb performed the service and $X_3 = 1$ if Wolfgang Linz performed the service.

d. Does the equation that you developed in part (c) provide a good fit for the observed data? Explain.

24. In a multiple regression analysis by McIntyre (1994), Tar, Nicotine and Weight are considered as possible predictors of Carbon Monoxide (CO) content for 25 different brands of cigarette. Details of variables and data follow.

Brand	The cigarette brand
Tar	The tar content (in mg)
Nicotine	The nicotine content (in mg)
Weight	The weight (in g)
CO	The carbon monoxide (CO) content (in mg)

Brand	Tar	Nicotine	Weight	CO
Alpine	14.1	0.86	.9853	13.6
Benson&Hedges	16.0	1.06	1.0938	16.6
BullDurham	29.8	2.03	1.1650	23.5
CamelLights	8.0	0.67	0.9280	10.2
Carlton	4.1	0.40	0.9462	5.4
Chesterfield	15.0	1.04	0.8885	15.0
GoldenLights	8.8	0.76	1.0267	9.0
Kent	12.4	0.95	0.9225	12.3
Kool	16.6	1.12	0.9372	16.3

Brand	Tar	Nicotine	Weight	CO
L&M	14.9	1.02	0.8858	15.4
LarkLights	13.7	1.01	0.9643	13.0
Marlboro	15.1	0.90	0.9316	14.4
Merit	7.8	0.57	0.9705	10.0
MultiFilter	11.4	0.78	1.1240	10.2
NewportLights	9.0	0.74	0.8517	9.5
Now	1.0	0.13	0.7851	1.5
OldGold	17.0	1.26	0.9186	18.5
PallMallLight	12.8	1.08	1.0395	12.6
Raleigh	15.8	0.96	0.9573	17.5
SalemUltra	4.5	0.42	0.9106	4.9
Tareyton	14.5	1.01	1.0070	15.9
True	7.3	0.61	0.9806	8.5
ViceroyRichLight	8.6	0.69	0.9693	10.6
VirginiaSlims	15.2	1.02	0.9496	13.9
WinstonLights	12.0	0.82	1.1184	14.9

a. Examine correlations between variables in the study and hence assess the possibility of problems of multicollinearity affecting any subsequent regression model involving independent variables Tar and Nicotine.

b. Thus develop an estimated multiple regression equation using an appropriate number of the independent variables featured in the study.

c. Are your predictors statistically significant? Use $\alpha = 0.05$. What explanation can you give for the results observed?

25. The data below (Dunn, 2007) come from a study investigating a new method of measuring body composition. Body fat percentage, age and gender is given for 18 adults aged between 23 and 61.

Age	Percent.Fat	Gender
23	9.5	M
23	27.9	F
27	7.8	M
27	17.8	M
39	31.4	F
41	25.9	F
45	27.4	M
49	25.2	F
50	31.1	F
53	34.7	F
53	42	F
54	29.1	F
56	32.5	F
57	30.3	F
58	33	F
58	33.8	F
60	41.1	F
61	34.5	F

a. Develop an estimated regression equation that relates Age and Gender to Percent.Fat

b. Is Age a significant factor in predicting Percent.Fat? Explain. Use $\alpha = 0.05$.

c. What is the estimated body fat percentage for a female aged 45?

26. Data for two variables, *X* and *Y*, follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Develop the estimated regression equation for these data.

b. Plot the standardized residuals versus *y*. Do there appear to be any outliers in these data? Explain.

c. Compute the studentized deleted residuals for these data. At the 0.05 level of significance, can any of these observations be classified as an outlier? Explain.

27. Data for two variables, *X* and *Y*, follow.

x_i	22	24	26	28	40
y_i	12	21	31	35	70

a. Develop the estimated regression equation for these data.

b. Compute the studentized deleted residuals for these data. At the 0.05 level of significance, can any of these observations be classified as an outlier? Explain.

c. Compute the leverage values for these data. Do there appear to be any influential observations in these data? Explain.

d. Compute Cook's distance measure for these data. Are any observations influential? Explain.

28. Data collected by Montgomery and Peck (see Hawkins, 1991) concern the three variables:

Y, the time taken to service a vending machine, X_1, the number of items stocked by the machine and X_2, the distance travelled to reach it.

X1	X2	Y
7	560	16.68
3	220	11.5
3	340	12.03
4	80	14.88
6	150	13.75
7	330	18.11
2	110	8
7	210	17.83
30	1460	79.24
5	605	21.5
16	688	40.33
10	215	21
4	255	13.5
6	462	19.75
9	448	24
10	776	29
6	200	15.35
7	132	19
3	36	9.5
17	770	35.1
10	140	17.9
26	810	52.32
9	450	18.75
8	635	19.83
4	150	10.75

a. Find an estimated regression equation relating the time taken to service a vending machine to the number of items stocked by the machine and the distance travelled to reach it.

b. Plot the standardized residuals against $\hat{y}$. Does the residual plot support the assumptions about ε? Explain.

c. Check for any outliers in these data. What are your conclusions?

d. Are there any influential observations? Explain.

29. Data (Tufte, 1974) on male deaths per million in 1950 for lung cancer (Y) and *per capita* cigarette consumption in 1930 (X) are given below:

Country	y	x	Country	y	x
Ireland	58	220	Norway	90	250
Sweden	115	310	Canada	150	510
Denmark	165	380	Australia	170	455
USA	190	1280	Holland	245	460
Switzerland	250	530	Finland	350	1115
GB	465	1145			

Results from a simple regression analysis of this information are as follows:

Regression Analysis: y versus x

The regression equation is
y = 65.7 + 0.229 x

Predictor	Coef	SE Coef	T	P
Constant	65.75	48.96	1.34	0.212
x	0.22912	0.06921	3.31	0.009

S = 84.1296 R-Sq = 54.9% R-Sq(adj) = 49.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	77554	77554	10.96	0.009
Residual Error	9	63700	7078		
Total	10	141255			

Unusual Observations

Obs	x	y	Fit	SE Fit	Residual	St Resid
4	1280	190.0	359.0	53.2	-169.0	-2.59R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 2.07188

Corresponding lererage and cook distance details are as follows.

HI1	COOK1
0.191237	0.06985
0.149813	0.00694
0.125175	0.00172
0.399306	2.23320
0.094716	0.03222
0.288283	0.75365
0.176211	0.02001
0.097018	0.00893
0.106139	0.00000
0.105140	0.05060
0.266962	0.02909

Carry out any further statistical tests you deem appropriate, otherwise comment on the effectiveness of the linear modes.

30. Refer to the Stamm Stores example introduced in this section. The dependent variable is coded as $Y = 1$ if the customer makes a purchase and 0 if not. Suppose that the only information available to help predict whether the customer will make a purchase is the customer's credit card status, coded as $X = 1$ if the customer has a Stamm credit card and $X = 0$ if not.

a. Write the logistic regression equation relating X to Y.
b. What is the interpretation of $E(Y)$ when $X = 0$?
c. For the Stamm data in Table 15.11, use MINITAB to compute the estimated logit.
d. Use the estimated logit computed in part (c) to compute an estimate of the probability of making a purchase for customers who do not have a Stamm credit card and an estimate of the probability of making a purchase for customers who have a Stamm credit card.
e. What is the estimate of the odds ratio? What is its interpretation?

31. In Table 15.12 we provided estimates of the probability of a purchase in the Stamm Stores catalogue promotion. A different value is obtained for each combination of values for the independent variables.

a. Compute the odds in favour of a purchase for a customer with annual spending of €4000 who does not have a Stamm credit card ($X_1 = 4$, $X_2 = 0$).
b. Use the information in Table 15.12 and part (a) to compute the odds ratio for the Stamm credit card variable X_2 holding annual spending constant at $X_1 = 4$.
c. In the text, the odds ratio for the credit card variable was computed using the information in the €2000 column of Table 15.12. Did you get the same value for the odds ratio in part (b)?

32. Community Bank would like to increase the number of customers who use payroll direct deposit. Management is considering a new sales campaign that will require each branch manager to call each customer who does not currently use payroll direct deposit. As an incentive to sign up for payroll direct deposit, each customer contacted will be offered free banking for two years. Because of the time and cost associated with the new campaign, management would like to focus their efforts on customers who have the highest probability of signing up for payroll direct deposit. Management believes that the average monthly balance in a customer's current account may be a useful predictor of whether the customer will sign up for direct payroll deposit. To investigate the relationship between these two variables, Community Bank tried the new campaign using a sample of 50 current account customers that do not currently use payroll direct deposit. The sample data show the average monthly current account balance (in hundreds of euros) and whether the customer contacted signed up for payroll direct deposit (coded 1 if the customer signed up for payroll direct deposit and 0 if not). The data are contained in the data set named Bank; a portion of the data follows.

Customer	X Monthly balance	Y Direct deposit
1	1.22	0
2	1.56	0
3	2.10	0
4	2.25	0
5	2.89	0
6	3.55	0
7	3.56	0
8	3.65	1
.	.	.
.	.	.
.	.	.
48	18.45	1
49	24.98	0
50	26.05	1

a. Write the logistic regression equation relating X to Y.

b. For the Community Bank data, use MINITAB to compute the estimated logistic regression equation.

c. Conduct a test of significance using the G test statistic. Use $\alpha = 0.05$.

d. Estimate the probability that customers with an average monthly balance of €1000 will sign up for direct payroll deposit.

e. Suppose Community Bank only wants to contact customers who have a 0.50 or higher probability of signing up for direct payroll deposit. What is the average monthly balance required to achieve this level of probability?

f. What is the estimate of the odds ratio? What is its interpretation?

33. Prior to the *Challenger* tragedy on January 28, 1986 after each launch of the space shuttle the solid rocket boosters were recovered from the ocean and inspected. Of the previous 24 shuttle launches, 7 had incidents of damage to the joints, 16 had no incidents of damage and 1 was unknown because the boosters were not recovered after launch.

In trying to explain the damage to joints it was thought that temperature at the time of launch could be a contributing factor.

For the data that follow, a 1 represents damage to field joints, and a 0 represents no damage.

Temp	Damage	Temp	Damage	Temp	Damage
66	0	57	1	70	0
70	1	63	1	81	0
69	0	70	1	76	0
68	0	78	0	79	0
67	0	67	0	75	1
72	0	53	1	76	0
73	0	67	0	58	1
70	0	75	0		

a. Fit a logistic regression model to these data and obtain a plot of the data and fitted curve.

b. Conduct a test of significance using the G test statistic. Use $\alpha = 0.05$

c. Estimate the probability of damage for a temperature of 50

d. What is the estimate of the odds ratio? How would you interpret it?

Chapter 15: Multiple Regression

Textbook Exercises Solutions:

1. a. $b_1 = .5906$ is an estimate of the change in y corresponding to a 1 unit change in x_1 when x_2 is held constant.

   $b_2 = .4980$ is an estimate of the change in y corresponding to a 1 unit change in x_2 when x_1 is held constant.

2. a. The estimated regression equation is

   $\hat{y} = 45.06 + 1.94x_1$

   An estimate of y when $X_1 = 45$ is

   $\hat{y} = 45.06 + 1.94(45) = 132.36$

   b. The estimated regression equation is

   $\hat{y} = 85.22 + 4.32x_2$

   An estimate of y when $X_2 = 15$ is

   $\hat{y} = 85.22 + 4.32(15) = 150.02$

   c. The estimated regression equation is

   $\hat{y} = -18.37 + 2.01x_1 + 4.74x_2$

   An estimate of y when $X_1 = 45$ and $X_2 = 15$ is

   $\hat{y} = -18.37 + 2.01(45) + 4.74(15) = 143.18$

3. a. $b_1 = 3.8$ is an estimate of the change in y corresponding to a 1 unit change in x_1 when x_2, x_3, and x_4 are held constant.

$b_2 = -2.3$ is an estimate of the change in y corresponding to a 1 unit change in x_2 when x_1, x_3, and x_4 are held constant.

$b_3 = 7.6$ is an estimate of the change in y corresponding to a 1 unit change in x_3 when x_1, x_2, and x_4 are held constant.

$b_4 = 2.7$ is an estimate of the change in y corresponding to a 1 unit change in x_4 when x_1, x_2, and x_3 are held constant.

4. a. $\hat{y} = -50.359 + 0.671(60) + 1.295(20) = 15.801$=estimated LOSS

b. LOSS can be expected to increase by 0.671 for every unit increase in AIRFLOW when TEMP is held constant. LOSS can be expected to increase by 1.295for every unit increase in TEMP when AIRFLOW is held constant.

5. a. The Minitab output is shown below:

```
The regression equation is
Revenue = 88.6 + 1.60 TVAdv

Predictor        Coef     SE Coef          T        P
Constant       88.638       1.582      56.02    0.000
TVAdv          1.6039      0.4778       3.36    0.015

S = 1.215       R-Sq = 65.3%     R-Sq(adj) = 59.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      16.640      16.640     11.27    0.015
Residual Error     6       8.860       1.477
Total              7      25.500
```

b. The Minitab output is shown below:

```
The regression equation is
Revenue = 83.2 + 2.29 TVAdv + 1.30 NewsAdv

Predictor        Coef     SE Coef          T        P
Constant       83.230       1.574      52.88    0.000
TVAdv          2.2902      0.3041       7.53    0.001
NewsAdv        1.3010      0.3207       4.06    0.010

S = 0.6426      R-Sq = 91.9%     R-Sq(adj) = 88.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      23.435      11.718     28.38    0.002
Residual Error     5       2.065       0.413
Total              7      25.500

Source       DF      Seq SS
TVAdv         1      16.640
NewsAdv       1       6.795
```

c. No, it is 1.60 in part (a) and 2.29 above. In part (b) it represents the marginal change in revenue due to an increase in television advertising with newspaper advertising held constant.

d. Revenue = 83.2 + 2.29(3.5) + 1.30(1.8) = €93.56 or €93,560

6. a. The Minitab output is shown below:

```
The regression equation is
Return = 247 - 32.8 Safety + 34.6 ExpRatio

Predictor        Coef     SE Coef          T        P
Constant        247.4       110.4       2.24    0.039
Safety         -32.84       13.95      -2.35    0.031
ExpRatio        34.59       14.13       2.45    0.026

S = 16.98       R-Sq = 58.2%     R-Sq(adj) = 53.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      6823.2      3411.6     11.84    0.001
Residual Error    17      4899.7       288.2
Total             19     11723.0
```

b. $\hat{y} = 247 - 32.8(7.5) + 34.6(2) = 70.2$

7. a. SSE = SST - SSR = 6,724.125 - 6,216.375 = 507.75

b. $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{6,216.375}{6,724.125} = .924$

c. $R_a^2 = 1-(1-R^2)\frac{n-1}{n-p-1} = 1-(1-.924)\frac{10-1}{10-2-1} = .902$

d. The estimated regression equation provided an excellent fit.

8. a. $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{14,052.2}{15,182.9} = .926$

b.

adj R^2 = 1 - (1 - R^2) $\frac{n-1}{n-p-1}$ = 1 - (1 – 0.926) $\frac{10-1}{10-2-1}$ = 0.905

c. Yes; after adjusting for the number of independent variables in the model, we see that 90.5% of the variability in *y* has been accounted for.

9. a. $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{1760}{1805} = .975$

b.

adj R^2 = 1 - (1 - R^2) $\frac{n-1}{n-p-1}$ = 1 - (1 – 0.975) $\frac{30-1}{30-4-1}$ = 0.971

c. The estimated regression equation provided an excellent fit.

10.

a. $R^2 = \frac{1880.443}{2069.238} = 0.909$

b.

adj R^2 = 1 - (1 - R^2) $\frac{n-1}{n-p-1}$ = 1 - 0.091 $\frac{20}{17}$ = 0.893

c. The adjusted coefficient of determination shows that 89.3% of the variability has been explained by the two independent variables; thus, we conclude that the model does explain a large amount of variability.

11. a. $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{23.435}{25.5} = .919$

$$\textbf{adj } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - (1 - 0.919)\frac{8-1}{8-2-1} = 0.887$$

b. Multiple regression analysis is preferred since both R^2 and **adj** R^2 show an increased percentage of the variability of Y explained when both independent variables are used.

12. a. MSR = SSR/p = 6,216.375/2 = 3,108.188

$$\text{MSE} = \frac{\text{SSE}}{n-p-1} = \frac{507.75}{10-2-1} = 72.536$$

b. F = MSR/MSE = 3,108.188/72.536 = 42.85

Using F table (2 degrees of freedom numerator and 7 denominator), p-value is less than .01

Because p-value $\leq \alpha$ = .05, the overall model is significant.

c. t = .5906/.0813 = 7.26

Using t table (7 degrees of freedom), area in tail is less than .005; p-value is less than .01

Because p-value $\leq \alpha$, β_1 is significant.

d. t = .4980/.0567 = 8.78

Using t table (7 degrees of freedom), area in tail is less than .005; p-value is less than .01

Because p-value $\leq \alpha$, β_2 is significant.

13. A portion of the Minitab output is shown below.

```
The regression equation is
Y = - 18.4 + 2.01 X1 + 4.74 X2

Predictor        Coef     SE Coef          T        P
Constant       -18.37       17.97      -1.02    0.341
X1             2.0102      0.2471       8.13    0.000
X2             4.7378      0.9484       5.00    0.002

S = 12.71       R-Sq = 92.6%     R-Sq(adj) = 90.4%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2     14052.2      7026.1     43.50    0.000
Residual Error     7      1130.7       161.5
Total              9     15182.9
```

a. Since the *p*-value corresponding to $F = 43.50$ is .000 < $\alpha = .05$, we reject H_0: $\beta_1 = \beta_2 = 0$; there is a significant relationship.

b. Since the *p*-value corresponding to $t = 8.13$ is .000 < $\alpha = .05$, we reject H_0: $\beta_1 = 0$; β_1 is significant.

c. Since the *p*-value corresponding to $t = 5.00$ is .002 < $\alpha = .05$, we reject H_0: $\beta_2 = 0$; β_2 is significant.

14. a. In the two independent variable case the coefficient of x_1 represents the expected change in y corresponding to a one unit increase in x_1 when x_2 is held constant. In the single independent variable case the coefficient of x_1 represents the expected change in y corresponding to a one unit increase in x_1.

b. Yes. If X_1 and X_2 are correlated one would expect a change in X_1 to be accompanied by a change in X_2.

15. a. SSE = SST - SSR = 2069.238 – 1880.443 = 188.795

$$s^2 = \frac{SSE}{n-p-1} = \frac{188.795}{7} = 26.971$$

$$MSR = \frac{SSR}{p} = \frac{1880.443}{2} = 940.222$$

b. F = MSR/MSE = 940.222/26.971 = 34.861

Using F table (2 degrees of freedom numerator and 7 denominator), p-value is less than .01

Because p-value $\leq \alpha$, we reject H_0. There is a significant relationship among the variables.

16. a. $F = 28.38$

Using F table (2 degrees of freedom numerator and 7 denominator), p-value is less than .01

Actual p-value = .002

Because p-value $\leq \alpha$, there is a significant relationship.

b. $t = 7.53$

Using t table (7 degrees of freedom), area in tail is less than .005; p-value is less than .01

Actual p-value = .001

Because p-value $\leq \alpha, \beta_1$ is significant and X_1 should not be dropped from the model.

c. $t = 4.06$

Actual p-value = .010

Because p-value $\leq \alpha$, β_2 is significant and x_2 should not be dropped from the model.

17. a. $\hat{y} = 29.1270 + .5906(180) + .4980(310) = 289.8150$

b. The point estimate for an individual value is $\hat{y} = 289.8150$, the same as the point estimate of the mean value.

18. a. Using Minitab, the 95% confidence interval is 132.16 to 154.16.

b. Using Minitab, the 95% prediction interval is 111.13 to 175.18.

19. a. $\hat{y} = 83.2 + 2.29(3.5) + 1.30(1.8) = 93.555$ or €93,555

Note: In Exercise 5b, the Minitab output also shows that $b_0 = 83.230$, $b_1 = 2.2902$,
and $b_2 = 1.3010$; hence, $\hat{y} = 83.230 + 2.2902x_1 + 1.3010x_2$. Using this estimated regression equation, we obtain

$\hat{y} = 83.230 + 2.2902(3.5) + 1.3010(1.8) = 93.588$ or €93,588

The difference (€93,588 - €93,555 = €33) is simply due to the fact that additional significant digits are used in the computations. From a practical point of view, however, the difference is not enough to be concerned about. In practice, a computer software package is always used to perform the computations and this will not be an issue.

The Minitab output is shown below:

Fit	Stdev.Fit	95% C.I.	95% P.I.
93.588	0.291	(92.840, 94.335)	(91.774, 95.401)

Note that the value of FIT ($\hat{y}$) is 93.588.

b. Confidence interval estimate: 92.840 to 94.335 or €92,840 to €94,335

c. Prediction interval estimate: 91.774 to 95.401 or €91,774 to €95,401

20. a. $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where

$x_2 = 0$ if level 1 and 1 if level 2

b. $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1$

c. $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2$

d. $\beta_2 = E(Y \mid \text{level 2}) - E(Y \mid \text{level 1})$

β_1 is the change in $E(Y)$ for a 1 unit change in x_1 holding x_2 constant.

21. a. two

b. $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ where

x_2	x_3	Level
0	0	1
1	0	2
0	1	3

c. $E(Y \mid \text{level 1}) = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 x_1$

$E(Y \mid \text{level 2}) = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) = \beta_0 + \beta_1 x_1 + \beta_2$

$E(Y \mid \text{level 3}) = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 x_1 + \beta_3$

$\beta_2 = E(Y \mid \text{level 2}) - E(Y \mid \text{level 1})$

$\beta_3 = E(Y \mid \text{level 3}) - E(Y \mid \text{level 1})$

β_1 is the change in $E(Y)$ for a 1 unit change in x_1 holding x_2 and x_3 constant.

22. a. 7.8

b. Estimate of sales = 42.62 + 0.812(70) - 20.1(1) + 7.8(2) = 94.96

c. Estimate of sales = 42.62 + 0.812(60) - 20.1(2) + 7.8(1) = 58.94

23. a. Let Type = 0 if a mechanical repair
Type = 1 if an electrical repair

The Minitab output is shown below:

```
The regression equation is
Time = 3.45 + 0.617 Type

Predictor        Coef     SE Coef          T        P
Constant       3.4500      0.5467       6.31    0.000
Type           0.6167      0.7058       0.87    0.408

S = 1.093       R-Sq = 8.7%      R-Sq(adj) = 0.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1       0.913       0.913      0.76    0.408
Residual Error     8       9.563       1.195
Total              9      10.476
```

b. The estimated regression equation did not provide a good fit. In fact, the *p*-value of .408 shows that the relationship is not significant for any reasonable value of α.

c. Person = 0 if Heinz Kolb performed the service and Person = 1 if Wolfgang Linz performed the service. The Minitab output is shown below:

```
The regression equation is
Time = 4.62 - 1.60 Person

Predictor        Coef     SE Coef          T        P
Constant       4.6200      0.3192      14.47    0.000
Person        -1.6000      0.4514      -3.54    0.008

S = 0.7138      R-Sq = 61.1%     R-Sq(adj) = 56.2%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      6.4000      6.4000     12.56    0.008
Residual Error     8      4.0760      0.5095
Total              9     10.4760
```

d. We see that 61.1% of the variability in repair time has been explained by the repair person that performed the service; an acceptable, but not good, fit.

24. a. The SPSS output is shown below:

Correlations

		Tar	Nicotine	Weight	CO
Tar	Pearson Correlation	1.000	.977**	.491*	.957**
	Sig. (2-tailed)		.000	.013	.000
	N	25.000	25	25	25
Nicotine	Pearson Correlation	.977**	1.000	.500*	.926**
	Sig. (2-tailed)	.000		.011	.000
	N	25	25.000	25	25
Weight	Pearson Correlation	.491*	.500*	1.000	.464*
	Sig. (2-tailed)	.013	.011		.019
	N	25	25	25.000	25
CO	Pearson Correlation	.957**	.926**	.464*	1.000
	Sig. (2-tailed)	.000	.000	.019	
	N	25	25	25	25.000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Clearly the variables are significantly correlated with each other suggesting multicollinearity is likely to be a major problem in any related multiple regression analysis. .

b. Relevant SPSS output from a Stepwise regression analysis is as follows:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.957[a]	.917	.913	1.3967

a. Predictors: (Constant), Tar

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	494.281	1	494.281	253.370	.000[a]
	Residual	44.869	23	1.951		
	Total	539.150	24			

a. Predictors: (Constant), Tar

b. Dependent Variable: CO

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.743	.675		4.063	.000		
	Tar	.801	.050	.957	15.918	.000	1.000	1.000

a. Dependent Variable: CO

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Nicotine	-.198[a]	-.699	.492	-.147	.046	21.627	.046
	Weight	-.008[a]	-.111	.913	-.024	.759	1.317	.759

a. Predictors in the Model: (Constant), Tar

b. Dependent Variable: CO

Clearly both Nicotine and Weight have been omitted as predictors from the model. As Tar is so highly correlated with Nicotine this is hardly surprising for the latter variable. If we try to fit a model with Tar and Weight as predictors, it can be shown that Weight is not statistically significant:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.114	3.416		.912	.372		
	Tar	.804	.059	.961	13.622	.000	.759	1.317
	Weight	-.423	3.813	-.008	-.111	.913	.759	1.317

a. Dependent Variable: CO

since p-value = 0.913 $> \alpha = .05$.

c. The *p*-value corresponding to Tar for the one predictor model in b. is $.000 < \alpha = .05$; thus Tar is a significant predictor here. In terms of its R square characteristics this is the best one predictor model available. Alternative models allowing for two or three predictors are not possible because of multicollinearity.

25. a. Recoding values of the Gender variable so that F = 0 and M = 1 we obtain relevant SPSS output below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	15.071	6.224		2.421	.029
	Age	.339	.120	.490	2.835	.013
	Gender	-9.791	3.697	-.458	-2.649	.018

a. Dependent Variable: Percent.Fat

$$\text{i.e. } \hat{y} = 15.071 + .339\,\text{Age} - 9.791\,\text{Gender}$$

b. Since the *p*-value corresponding to the Age predictor is .013 < α = .05, we deduce that Age is indeed a significant factor in predicting Percent.Fat.

c. For a female aged 45, Gender = 0 and Age = 45. Hence Percent.Fat can be predicted as:

$$\hat{y} = 15.071 + .339\,\text{Age} - 9.791\,\text{Gender}$$

$$= 15.071 + .339\,(45) - 9.791\,(0) = 30.326\%$$

26. a. The Minitab output is shown below:

```
The regression equation is
Y = 0.20 + 2.60 X

Predictor        Coef     SE Coef          T        P
Constant        0.200       2.132       0.09    0.931
X              2.6000      0.6429       4.04    0.027

S = 2.033       R-Sq = 84.5%     R-Sq(adj) = 79.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      67.600      67.600     16.35    0.027
Residual Error     3      12.400       4.133
Total              4      80.000
```

b. Using Minitab we obtained the following values:

x_i	y_i	$\hat{y}_i$	Standardized Residual
1	3	2.8	.16
2	7	5.4	.94
3	5	8.0	-1.65
4	11	10.6	.24
5	14	13.2	.62

The point (3,5) does not appear to follow the trend of remaining data; however, the value of the standardized residual for this point, -1.65, is not large enough for us to conclude that (3, 5) is an outlier.

c. Using Minitab, we obtained the following values:

x_i	y_i	Studentized Deleted Residual
1	3	.13
2	7	.91
3	5	- 4.42
4	11	.19
5	14	.54

$t_{.025} = 4.303$ ($n - p - 2 = 5 - 1 - 2 = 2$ degrees of freedom)

Since the studentized deleted residual for (3, 5) is -4.42 < -4.303, we conclude that the 3rd observation is an outlier.

27. a. The Minitab output is shown below:

```
The regression equation is
Y = -53.3 + 3.11 X

Predicator             Coef     SE Coef           T         p
Constant            -53.280       5.786       -9.21     0.003
X                     3.1100      0.2016       15.43     0.001

S = 2.851    R-sq = 98.8%  R-sq (adj) = 98.3%

Analysis of Variance

SOURCE              DF          SS          MS         F         p
Regression           1      1934.4      1934.4    238.03     0.001
Residual Error       3        24.4         8.1
Total                4      1598.8
```

b. Using the Minitab we obtained the following values:

x_i	y_i	Studentized Deleted Residual
22	12	-1.94
24	21	-.12
26	31	1.79
28	35	.40
40	70	-1.90

$t_{.025} = 4.303$ ($n - p - 2 = 5 - 1 - 2 = 2$ degrees of freedom)

Since none of the studentized deleted residuals are less than -4.303 or greater than 4.303, none of the observations can be classified as an outlier.

c. Using Minitab we obtained the following values:

x_i	y_i	h_i
22	12	.38
24	21	.28
26	31	.22
28	35	.20
40	70	.92

The critical value is

$$\frac{3(p+1)}{n} = \frac{3(1+1)}{5} = 1.2$$

Since none of the values exceed 1.2, we conclude that there are no influential observations in the data.

d. Using Minitab we obtained the following values:

x_i	y_i	D_i
22	12	.60
24	21	.00
26	31	.26
28	35	.03
40	70	11.09

Since $D_5 = 11.09 > 1$ (rule of thumb critical value), we conclude that the fifth observation is influential.

28. a. The Minitab output appears in the solution to part (b) of Exercise 5; the estimated regression equation is:

Revenue = 83.2 + 2.29 TVAdv + 1.30 NewsAdv

b. Using Minitab we obtained the following values:

$\hat{y}_i$	Standardized Residual
96.63	-1.62
90.41	-1.08
94.34	1.22
92.21	- .37
94.39	1.10
94.24	- .40
94.42	-1.12
93.35	1.08

With the relatively few observations, it is difficult to determine if any of the assumptions regarding the error term have been violated. For instance, an argument could be made that there does not appear to be any pattern in the plot; alternatively an argument could be made that there is a curvilinear pattern in the plot.

c. The values of the standardized residuals are greater than -2 and less than +2; thus, using test, there are no outliers. As a further check for outliers, we used Minitab to compute the following studentized deleted residuals:

Observation	Studentized Deleted Residual
1	-2.11
2	-1.10
3	1.31
4	- .33
5	1.13
6	- .36
7	-1.16
8	1.10

$t_{.025} = 2.776$ ($n - p - 2 = 8 - 2 - 2 = 4$ degrees of freedom)

Since none of the studentized deleted residuals is less than -2.776 or greater than 2.776, we conclude that there are no outliers in the data.

d. Using Minitab we obtained the following values:

Observation	h_i	D_i
1	.63	1.52
2	.65	.70
3	.30	.22
4	.23	.01
5	.26	.14
6	.14	.01
7	.66	.81
8	.13	.06

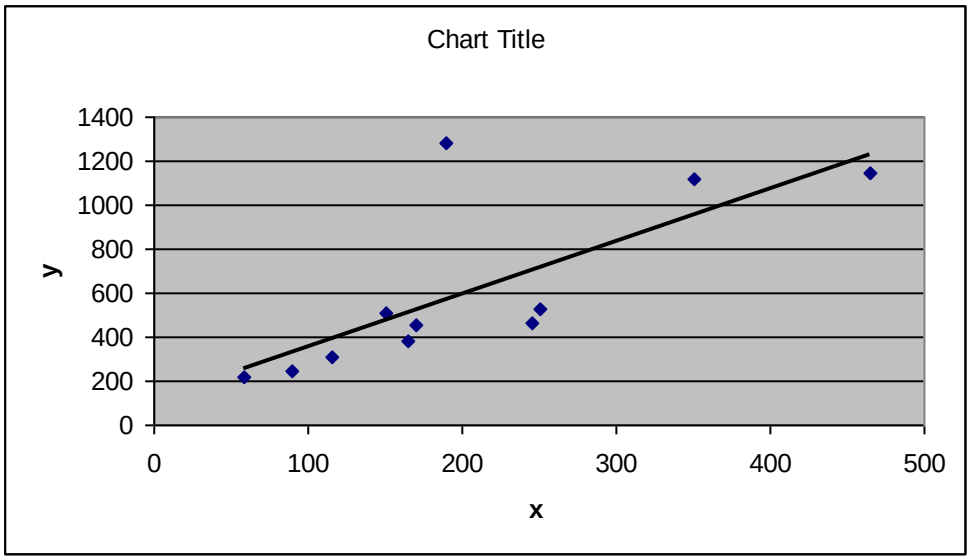
The critical average value is

$$\frac{3(p+1)}{n} = \frac{3(2+1)}{8} = 1.125$$

Since none of the values exceed 1.125, we conclude that there are no influential observations.

However, using Cook's distance measure, we see that $D_1 > 1$ (rule of thumb critical value); thus, we conclude the first observation is influential. Final Conclusion: observations 1 is an influential observation.

29.



None of the hi values are greater than 6/n = 6/11 so influence does not seem to be a problem with the dataset. However, observation 4 (for the USA) has a Cook distance > 1 which suggests in the absence of influence that the corresponding residual is large. From the scattergram, the outlier for the USA is clearly very evident. Yet even with this, a significant linear regression model (pvalue=.0009 < .05) has been obtained confirming there is a significant correlation between cigarette consumption and the incidence of lung cancer. Note that because of the outlier, the coefficient of determination is only 54.9%.

30. a. $E(Y) = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x_1}}$

b. It is an estimate of the probability that a customer that does not have a Stamm credit card will make a purchase.

c. A portion of the Minitab binary logistic regression output follows:

```
Logistic Regression Table

                                                    Odds         95% CI
Predictor       Coef    SE Coef        Z      P    Ratio    Lower    Upper
Constant     -0.9445     0.3150    -3.00 0.003
Card          1.0245     0.4235     2.42 0.016     2.79     1.21     6.39

Log-Likelihood = -64.265
Test that all slopes are zero: G = 6.072, DF = 1, P-Value = 0.014
```

Thus, the estimated logit is $\hat{g}(x) = -0.9445 + 1.0245x$

d. For customers that do not have a Stamm credit card ($x = 0$)

$$\hat{g}(0) = -0.9445 + 1.245(0) = 0.9445$$

and

$$\hat{y} = \frac{e^{\hat{g}(0)}}{1+e^{\hat{g}(0)}} = \frac{e^{-0.9445}}{1+e^{-0.9445}} = \frac{0.3889}{1+0.3889} = 0.28$$

For customers that have a Stamm credit card ($x = 1$)

$$\hat{g}(1) = -0.9445 + 1.245(1) = 0.0800$$

and

$$\hat{y} = \frac{e^{\hat{g}(1)}}{1+e^{\hat{g}(1)}} = \frac{e^{0.08}}{1+e^{0.08}} = \frac{1.0833}{1+1.0833} = 0.52$$

e. Using the Minitab output shown in part (c), the estimated odds ratio is 2.79. We can conclude that the estimated odds of making a purchase for customers who have a Stamm credit card are 2.79 times greater than the estimated odds of making a purchase for customers that do not have a Stamm credit card.

31. a. $\text{odds} = \dfrac{.3413}{1-.3413} = .4584$

b. $\text{odds}_1 = \dfrac{.5790}{1-.5790} = 1.3753$

$\text{odds}_0 = .4584$ (from part (a))

$\text{odds ratio} = \dfrac{\text{odds}_1}{\text{odds}_0} = \dfrac{1.3753}{.4584} = 3.00$

c. The odds ratio for x_2 computed holding annual spending constant at €2000 is also 3.00. This shows that the odds ratio for x_2 is independent of the value of x_1.

32. a. $E(Y) = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x_1}}$

b. A portion of the Minitab binary logistic regression output follows:

```
Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant     -2.6335     0.7985    -3.30 0.001
Balance      0.22018    0.09002     2.45 0.014     1.25     1.04     1.49

Log-Likelihood = -25.813
Test that all slopes are zero: G = 9.460, DF = 1, P-Value = 0.002
```

Thus, the estimated logistic regression equation is

$$E(y) = \frac{e^{-2.6355+0.22018x}}{1+e^{-2.6355+0.22018x}}$$

c. Significant result: the *p*-value corresponding to the *G* test statistic is 0.0002.

d. For an average monthly balance of €1000, $x = 10$

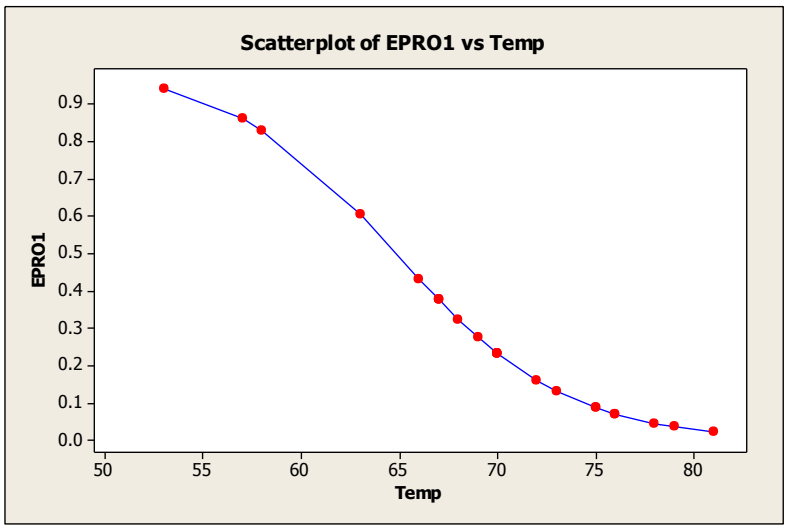
$$E(y) = \frac{e^{-2.6355+0.22018x}}{1+e^{-2.6355+0.22018x}} = \frac{e^{-2.6355+0.22018(10)}}{1+e^{-2.6355+0.22018(10)}} = \frac{e^{-0.4317}}{1+e^{-0.4317}} = \frac{0.6494}{1.6494} = 0.39$$

Thus, an estimate of the probability that customers with an average monthly balance of €1000 will sign up for direct payroll deposit is 0.39.

e. Repeating the calculations in part (d) using various values for x, a value of $x = 12$ or an average monthly balance of approximately €1200 is required to achieve this level of probability.

f. Using the Minitab output shown in part (b), the estimated odds ratio is 1.25. Because values of x are measured in hundreds of euros, the estimated odds of signing up for payroll direct deposit for customers that have an average monthly balance of €600 is 1.25 times greater than the estimated odds of signing up for payroll direct deposit for customers that have an average monthly balance of €500. Moreover, this interpretation is true for any one hundred euro increment in the average monthly balance.

33. a.



A portion of the Minitab binary logistic regression output follows:

```
Logistic Regression Table

                                                  Odds      95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant     15.0429    7.37862   2.04  0.041
Temp       -0.232163   0.108236  -2.14  0.032   0.79   0.64   0.98


Log-Likelihood = -10.158
Test that all slopes are zero: G = 7.952, DF = 1, P-Value = 0.005
```

Hence estimated logistic equation is:

$$E(Y) = P(y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{e^{15.0429 - 0.232163x}}{1 + e^{15.0429 - 0.232163x}}$$

b. From the above output, the pvalue associated with G = 7.952 is $0.005 < .05 = \alpha$. Hence the model is significant overall.

c. Predicted Event Probabilities for New Observations

New Obs	Prob	SE Prob	95% CI
1	0.968774	0.0612053	(0.370358, 0.999389)

Hence, the Probability when Temp = 50 is 0.968774

d. Using the Minitab output shown in part (b), the estimated odds ratio is 0.79. The estimated odds of damage is 0.79 times less for every degree rise in temperature.

Supplementary Exercises:

34. The personnel director for Electronics Associates developed the following regression equation relating an employee's score on a job satisfaction test to his or her length of service and wage rate

$$\hat{y} = 14.4 - 8.69x_1 + 13.5x_2$$

where X_1 = length of service (years)

X_2 = wage rate (euros)

Y = job satisfaction test score (higher scores indicate greater job satisfaction)

a. Interpret the coefficients in this estimated regression equation.

b. Develop an estimate of the job satisfaction test score for an employee who has four years of service and earns €6.50 per hour.

35. A partial computer output from a regression analysis follows.

```
The regression equation is
Y = 8.103 + 7.602 X1 + 3.111 X2


Predictor       Coef   SE Coef      T
Constant     ________    2.667   _______
X1           ________    2.105   _______
X2           ________    0.613   _______

S = 3.335     R-Sq = 92.3%  R-Sq(adj) = _______%


Analysis of Variance

Source          DF      SS      MS      F
Regression   _____    1612  ______  _____
Residual Error   12  _____  ______
Total        ______  _____
```

a. Compute the appropriate *t*-ratios.

b. Test for the significance of β_1 and β_2 at $\alpha = 0.05$.

c. Compute the entries in the DF, SS, and MS columns.

d. Compute **adj** R^2.

36. Recall that in exercise 34 the personnel director for Electronics Associates estimated a regression equation relating an employee's score on a job satisfaction test to length of service and wage rate. A portion of the Minitab computer output follows.

```
The regression equation is
Y = 14.4 - 8.69 X1 + 13.52 X2

Predictor       Coef   SE Coef      T
Constant     14.448     8.191   1.76
X1           _______     1.555  ______
X2           13.517     2.085   ______

S = 3.335    R-Sq = _______%     R-Sq(adj) = _______%


Analysis of Variance

Source          DF      SS      MS      F
Regression       2   ______   _____   _____
Residual Error ______ 71.17   _____
Total            7   720.0
```

a. Complete the missing entries in this output.

b. Compute *F* and test using □ = 0.05 to see whether a significant relationship is present.

c. Did the estimated regression equation provide a good fit to the data? Explain.

d. Use the *t* test and □ = 0.05 to test H_0: $\square_1 = 0$ and H_0: $\square_2 = 0$.

37. Data (Sen & Srivistava 1990) on employees' relationships with their supervisors have been collected as follows:

Y	satisfaction with supervisor/job
X_1	social contact with supervisor
X_2	supervisor's personal interest in employee's personal life
X_3	supervisor's level of support
X_4	supervisor's drive (employee's perception)
X_5	supervisor's drive (self-assessment)

The maximum value possible with Y is 20 and for the regression analysis described below the dependent variable was taken as

$$y = -\ln(1-Y/20)$$

MODEL 1

```
The regression equation is
y = 13.1 - 0.37 x1 - 1.50 x2 + 0.69 x3 - 0.149 x4 + 0.739 x5

Predictor       Coef       Stdev    t-ratio        p       VIF
Constant      13.063       1.362       9.59    0.000
x1            -0.374       2.775      -0.13    0.894     665.9
x2            -1.501       1.261      -1.19    0.246     118.0
x3             0.695       2.068       0.34    0.740     715.5
x4           -0.1487      0.7648      -0.19    0.847     100.1
x5            0.7386      0.6300       1.17    0.253     116.8

s = 0.9903      R-sq = 41.2%     R-sq(adj) = 28.9%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    5     16.4909      3.2982      3.36    0.019
Error        24     23.5357      0.9807
Total        29     40.0267

SOURCE       DF      SEQ SS
x1            1     15.0663
x2            1      0.0451
x3            1      0.0073
x4            1      0.0243
x5            1      1.3480

Unusual Observations
Obs.       x1        y        Fit  Stdev.Fit   Residual    St.Resid
  4      6.6    13.300     15.561      0.446     -2.261      -2.56R
  6      4.7    12.600     15.211      0.469     -2.611      -2.99R

R denotes an obs. with a large st. resid.

     Fit  Stdev.Fit         95% C.I.             95% P.I.
  14.320      3.632   (   6.822,  21.819)  (   6.548,  22.092) XX

X  denotes a row with X values away from the center
XX denotes a row with very extreme X values
```

(Note that the prediction output here relates to a case or data point with the values:

$x_1 = 6$
$x_2 = 8$
$x_3 = 10$
$x_4 = 12$
$x_5 = 14$

In contrast, output for a model involving the single predictor x_1 is as follows:

MODEL 2

```
The regression equation is
y = 13.2 + 0.421 x1

Predictor       Coef       Stdev    t-ratio        p
Constant     13.1576      0.7757      16.96    0.000
x1            0.4215      0.1025       4.11    0.000

s = 0.9442      R-sq = 37.6%     R-sq(adj) = 35.4%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      15.066      15.066     16.90    0.000
Error        28      24.960       0.891
Total        29      40.027

Unusual Observations
Obs.       x1          y        Fit  Stdev.Fit   Residual    St.Resid
  4      6.6    13.300     15.939      0.190     -2.639      -2.85R
  6      4.7    12.600     15.139      0.324     -2.539      -2.86R
R denotes an obs. with a large st. resid.
```

a. Comment on the effectiveness of MODELS 1 & 2.

b. Which model do you prefer and why?

38. A study, referred to by Gunst & Mason (1980), explores possible relationships between reaction times (to a visual stimulus) and physiological and fitness characteristics for police applicants. Details of the particular variables involved and the subsequent multiple regression analysis are as follows:

REACT	Reaction Time (seconds)
HEIGHT	Height of Applicant (cm)
WEIGHT	Weight of Applicant (kg)
SHLDR	Shoulder Width (cm)
PELVIC	Pelvic Width (cm)
CHEST	Minimum Chest Circumference (cm)
THIGH	Thigh Skinfold Thickness (mm)
PULSE	Pulse Rate (count per min)
DIAST	Diastolic Blood Pressure
CHNUP	Number of Chinups Applicant can complete
BREATH	Maximnum Breathing Capacity (litres)
RECVR	Pulse Rate (after 5 mins recovery from treadmill)
SPEED	Maximum Treadmill speed
ENDUR	Treadmill Endurance Time (min)
FAT	Total Body fat measurement

(Note that predictor variables were appropriately standardised before the analysis.)

Correlation coefficients

	REACT
HEIGHT	0.222
WEIGHT	0.056
SHLDR	-0.094
PELVIC	-0.056
CHEST	-0.032
THIGH	0.132
PULSE	0.163
DIAST	0.147
CHNUP	-0.158
BREATH	0.160
RECVR	-0.076
SPEED	-0.149
ENDUR	-0.053
FAT	0.165

Regression Analysis: react versus height, weight, ...

```
The regression equation is
react = 0.651 + 0.00404 height - 0.00106 weight + 0.00235 shldr
       -  0.0165 pelvic - 0.00351 chest - 0.00218 thigh + 0.000308 pulse
       + 0.00146 diast + 0.00150 chnup + 0.000279 breath - 0.00207 recvr -
       0.0493 speed - 0.0035 endur + 0.00583 fat
```

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.6508	0.4138	1.57	0.125	
height	0.004035	0.001728	2.34	0.025	3.4
weight	-0.001062	0.002278	-0.47	0.644	17.1
shldr	0.002350	0.007894	0.30	0.768	3.9
pelvic	-0.016542	0.008004	-2.07	0.046	3.3
chest	-0.003514	0.003164	-1.11	0.274	8.9
thigh	-0.002183	0.002603	-0.84	0.407	6.3
pulse	0.0003079	0.0006803	0.45	0.654	1.9
diast	0.0014623	0.0009369	1.56	0.128	1.4
chnup	0.001504	0.002696	0.56	0.581	3.5
breath	0.0002792	0.0003576	0.78	0.440	2.1
recvr	-0.0020655	0.0009233	-2.24	0.032	2.3
speed	-0.04930	0.02700	-1.83	0.076	2.4
endur	-0.00353	0.01327	-0.27	0.792	1.4
fat	0.005829	0.003294	1.77	0.085	14.6

```
S = 0.0442831   R-Sq = 39.7%   R-Sq(adj) = 15.6%
```

```
Analysis of Variance

Source          DF        SS        MS     F      P
Regression      14  0.045273  0.003234  1.65  0.114
Residual Error  35  0.068635  0.001961
Total           49  0.113908


Source  DF    Seq SS
height   1  0.005631
weight   1  0.001381
shldr    1  0.010110
pelvic   1  0.002695
chest    1  0.000760
thigh    1  0.000490
pulse    1  0.002687
diast    1  0.004749
chnup    1  0.000902
breath   1  0.000013
recvr    1  0.005248
speed    1  0.003863
endur    1  0.000601
fat      1  0.006143


Unusual Observations

Obs  height    react      Fit   SE Fit  Residual  St Resid
  9     184  0.42500  0.32760  0.02055   0.09740      2.48R

R denotes an observation with a large standardized residual.
```

a. Describe and interpret the results here.
b. How effective do you think the model is as a predictive aid?

39. *Consumer Reports* conducted a taste test on 19 brands of boxed chocolates. The following data show the price per serving, based on the PDA serving size of 1.4 ounces, and the quality rating for the 19 chocolates tested (*Consumer Reports, February 2002*).

Manufacturer	**Price**	**Rating**
Bernard Callebaut	3.17	Very Good
Candinas	3.58	Excellent
Fannie May	1.49	Good
Godiva	2.91	Very Good
Hershey's	0.76	Good
L.A. Burdick	3.70	Very Good
La Maison du Chocolate	5.08	Excellent
Leonidas	2.11	Very Good
Lindt	2.20	Good
Marline's	4.76	Excellent
Michael Recchiuti	7.05	Very Good
Neuchatel	3.36	Good
Neuchatel Sugar Free	3.22	Good
Richard Donnelly	6.55	Very Good
Russell Stover	0.70	Good
See's	1.06	Very Good
Teuscher Lake of Zurich	4.66	Very Good
Whitman's	0.70	Fair
Whitman's Sugar Free	1.21	Fair

Suppose that you would like to determine whether products that cost more rate higher in quality. For the purpose of this exercise, use the following binary dependent variable:

Y = 1 if the quality rating is very good or excellent and 0 if good or fair

a. Write the logistic regression equation relating X = price per serving to Y.

b. Use Minitab to compute the estimated logit.

c. Use the estimated logit computed in part (b) to compute an estimate of the probability a chocolate that has a price per serving of $4.00 will have a quality rating of very good or excellent.

d. What is the estimate of the odds ratio? What is its interpretation?

40. A college admissions officer developed the following estimated regression equation relating the final college performance GPA to the students SAT mathematics score and secondary education level GPA.

$$\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_2$$

where X_1 = secondary education level GPA

X_2 = SAT mathematics score

Y = final college performance GPA

a. Interpret the coefficients in this estimated regression equation.

b. Estimate the final college GPA for a student who had a secondary level GPA of 84 and a score of 540 on the SAT mathematics test.

41. Recall that in exercise 40 the college admissions officer developed the following estimated regression equation relating the final college performance GPA to the students SAT mathematics score and secondary education level GPA.

$$\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_2$$

where X_1 = secondary education level GPA

X_2 = SAT mathematics score

Y = final college performance GPA

A portion of the Minitab computer output follows.

```
The regression equation is
Y = -1.41 + 0.0235 X1 + 0.00486 X2


Predictor       Coef    SE Coef       T
Constant     -1.04053   0.4848    1.76
X1            0.023467  0.008666 ______
X2           ________    2.085    ______

S = 0.1298    R-Sq = _______%     R-Sq(adj) = _______%


Analysis of Variance

Source          DF      SS      MS      F
Regression       2   1.76209   _____   _____
Residual Error _____ _______   _____
Total            9    1.88000
```

a. Complete the missing entries in this output.
b. Compute F and test using $\alpha = 0.05$ to see whether a significant relationship is present.
c. Did the estimated regression equation provide a good fit to the data? Explain.
d. Use the t test and $\alpha = 0.05$ to test H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$.

42. *Auto Rental News* provided the following data, which show the number of cars in service (1000s), the number of locations, and the rental revenue ($ millions) for 15 car rental companies (*The Wall Street Journal Almanac 1998*).

Company	Cars	Locations	Revenue
Alamo	130	171	1180
Avis	190	1130	1500
Budget	126	1052	1500
Dollar	63.5	450	560
Enterprise	315.1	2636	2060
FRCS(Ford)	55.25	1784	312.5
Hertz	250	1200	2400
National	135	935	1200
Payless	15	100	47
PROP (Chrysler)	27	1500	160
Rent-A-Wreck	10.9	460	78
Snappy	15.5	259	85
Thrifty	34	480	340
U-Save	13.5	500	95
Value	18	45	150.1

a. Determine the estimated regression equation that can be used to predict the rental revenue given the number of cars in service.

b. Provide an interpretation for the slope of the estimated regression equation developed in part (a).

c. Determine the estimated regression equation that can be used to predict the rental revenue given the number of cars in service and the number of locations.

43. *Barron's* conducts an annual review of online brokers, including both brokers that can be accessed via a Web browser, as well as direct-access brokers that connect customers directly with the broker's network server. Each broker's offerings and performance are evaluated in six areas, using a point value of 0-5 in each category. The results are weighted to obtain an overall score, and a final star rating, ranging from zero to five stars, is assigned to each broker. Trade execution, ease of use, and range of offerings are three of the areas evaluated. A point value of 5 in the trade execution area means the order entry and execution process flowed easily from one step to the next. A value of 5 in the ease of use area means that the site was easy to use and can be tailored to show what the user wants to see. A value of 5 in the range offerings area means that all of the investment transactions can be executed online. The following data show the point values for trade execution, ease of use, range of offerings, and the star rating for a sample of 10 of the online brokers that Barren's evaluated (*Barron 's, March 10, 2003*).

Broker	Trade Execution	Use	Range	Rating
Wall St. Access	3.7	4.5	4.8	4.0
E*TRADE (Power)	3.4	3.0	4.2	3.5
E*TRADE (Standard)	2.5	4.0	4.0	3 5
Preferred Trade	4.8	3.7	3.4	3.5
my Track	4.0	3.5	3.2	3.5
TD Waterhouse	3.0	3.0	4.6	3.5
Brown & Co.	2.7	2.5	3.3	3.0
Brokerage America	1.7	3.5	3.1	3.0
Men-ill Lynch Direct	2.2	2.7	3.0	2.5
Strong Funds	1.4	3.6	2.5	2.0

a. Determine the estimated regression equation that can be used to predict the star rating given the point values for execution, ease of use, and range of offerings.

b. Use the F test to determine the overall significance of the relationship. What is the conclusion at the 0.05 level of significance?

c. Use the t test to determine the significance of each independent variable. What is your conclusion at the 0.05 level of significance?

d. Remove any independent variable that is not significant from the estimated regression equation. What is your recommended estimated regression equation? Compare the R^2 with the value of R^2 from part (a). Discuss the differences.

44. In exercise 42 an estimated regression equation was developed relating the rental revenue to the number of cars in service and the number of locations.

a. Test for a significant relationship between the dependent variable and the two independent variables. Use $\alpha = 0.05$.

b. Is the number of cars in service significant? Use $\alpha = 0.05$.

c. Is the number of locations significant? Use $\alpha = 0.05$.

45. The following table reports the horsepower, curb weight, and the speed at ¼ mile for 16 sports and GT cars (*1998 Road & Track Sports & GT Cars*).

Sports & GT Car	Curb Weight (lb.)	Horsepower	Speed at ¼ mile (mph)
Acura Integra Type R	2577	195	90.7
Acura NSX-T	3066	290	108.0
BMW Z3 2.8	2844	189	93.2
Chevrolet Camaro Z28	3439	305	103.2
Chevrolet Corvette Convertible	3246	345	102.1
Dodge Viper RT/10	3319	450	116.2
Ford Mustang GT	3227	225	91.7
Honda Prelude Type SH	3042	195	89.7
Mercedes-Benz CLK320	3240	215	93.0
Mercedes-Benz SLK230	3025	185	92.3
Mitsubishi 3000GT VR-4	3737	320	99.0
Nissan 240SX SE	2862	155	84.6
Pontiac Firebird Trans Am	3455	305	103.2
Porsche Boxster	2822	201	93.2
Toyota Supra Turbo	3505	320	105.0
VolvoC70	3285	236	97.0

a. Use curb weight as the independent variable and the speed at ¼ mile as the dependent variable. What is the estimated regression equation?

b. Use curb weight and horsepower as two independent variables and the speed at ¼ mile as the dependent variable. What is the estimated regression equation?

c. The 1999 Porsche 911 Carrera has been advertised as having a curb weight of 2910 pounds and an engine with 296 horsepower. Use the results in part (b) to predict the speed at ¼ mile for the Porsche 911.

46. Designers of backpacks use exotic material such as supernylon Delrin, high-density polyethylene, aircraft aluminum, and thermomoulded foam to make packs that fit comfortably and distribute weight to eliminate pressure points. The following data show the capacity (cubic inches), comfort rating, and price for 10 backpacks tested by *Outside Magazine*. Comfort was measured using a rating from 1 to 5, with a rating of 1 denoting average comfort and a rating of 5 denoting excellent comfort (*Outside Buyer's Guide, 2001*).

Manufacturer and Model	**Capacity**	**Comfort**	**Price ($)**
Camp Trails Paragon II	4330	2	19C
EMS 5500	5500	3	219
Lowe Alpomayo 90+20	5500	4	249
Marmot Muir	4700	3	249
Kelly Bigfoot 5200	5200	4	250
Gregory Whitney	5500	4	340
Osprey 75	4700	4	389
Arc'Teryx Bora 95	5500	5	395
Dana Design Terraplane LTW	5800	5	439
The Works @ Mystery Ranch Jazz	5000	5	525

a. Determine the estimated regression equation that can be used to predict the price of a backpack given the capacity and the comfort rating.

b. Interpret b_1 and b_2.

c. Predict the price for a backpack with a capacity of 4500 cubic inches and a comfort rating of 4.

47. In exercise 45, data were given on curb weight, horsepower, and speed at ¼ mile for 16 sports and GT cars (*1998 Road and Track Sports & GT Cars*).

a. Estimate the ¼ mile speed of a 1999 Porsche 911 Carrera that has a curb weight of 2910 pounds and a horsepower of 296.

b. Provide a 95% confidence interval for the ¼-mile speed of all sports and GT cars with the characteristics listed in part (a).

c. Provide a 95% prediction interval for the 1999 Porsche 911 Carrera described in part (a).

48. Over the past few years the percentage of students who leave Euroland College at the end of the first year has increased. Last year Euroland started a voluntary one-week orientation programme to help first-year students adjust to campus life. If Euroland is able to show that the orientation programme has a positive effect on retention, they will consider making the programme a requirement for all first-year students. Euroland's administration also suspects that students with lower GPAs have a higher probability of leaving Euroland at the end of the first year. In order to investigate the relation of these variables to retention, Euroland selected a random sample of 100 students from last year's entering class. The data are contained in the data set named Euroland; a portion of the data follows.

Student	**GPA**	**Programme**	**Return**
1	3.78	1	1
2	2.38	0	1
3	1.30	0	0
4	2.19	1	0
5	3.22	1	1
6	2.68	1	1
.	.	.	.
.	.	.	.
.	.	.	.
98	2.57	1	1
99	1.70	1	1
100	3.85	1	1

The dependent variable was coded as Y = 1 if the student returned to Euroland for the first year and y = 0 if not. The two independent variables are:

X_1 = GPA at the end of the first semester

X_2 = 0 if the student did not attend the orientation programme
1 if the student attended the orientation programme

a. Write the logistic regression equation relating x_1 and x_2 to y.

b. What is the interpretation of E(Y) when $X_2 = 0$?

c. Use both independent variables and Minitab to compute the estimated logit.

d. Conduct a test for overall significance using $\alpha = 0.05$.

e. Use $\alpha = 0.05$ to determine whether each of the independent variables is significant.

f. Use the estimated logit computed in part (c) to compute an estimate of the probability that students with a 2.5 grade point average who did not attend the orientation programme will return to Euroland for their first year. What is the estimated probability for students with a 2.5 grade point average who attended the orientation programme?

g. What is the estimate of the odds ratio for the orientation programme? Interpret it.

h. Would you recommend making the orientation programme a required activity? Why or why not?

Supplementary Exercises Solutions:

33. a. Job satisfaction can be expected to decrease by 8.69 units with a one unit increase in length of service if the wage rate does not change. A dollar increase in the wage rate is associated with a 13.5 point increase in the job satisfaction score when the length of service does not change.

b. $\hat{y} = 14.4 - 8.69(4) + 13.5(6.5) = 67.39$

34. a. The computer output with the missing values filled in is as follows:

```
The regression equation is

Y = 8.103 + 7.602 X1 + 3.111 X2


Predictor        Coef     SE Coef          T
Constant        8.103       2.667       3.04
X1              7.602       2.105       3.61
X2              3.111       0.613       5.08

S = 3.35      R-sq = 92.3%      R-sq (adj) = 91.0%

Analysis of Variance

SOURCE             DF          SS          MS         F
Regression          2        1612         806     71.82
Residual Error     12      134,67     11.2225
Total              14     1746.67
```

b. Using *t* table (12 degrees of freedom), area in tail corresponding to $t = 3.61$ is less than .005; *p*-value is less than .01

Because *p*-value $\leq \alpha$, reject $H_0 : \beta_1 = 0$

Using *t* table (12 degrees of freedom), area in tail corresponding to $t = 5.08$ is less than .005; *p*-value is less than .01

Because *p*-value $\leq \alpha$, reject $H_0 : \beta_2 = 0$

c. See computer output.

d. $R_a^2 = 1-(1-.923)\dfrac{14}{12} = .91$

35. a. The regression equation is

```
Y = 14.4 - 8.69 X1 + 13.52 X2

Predictor         Coef       SE Coef            T
Constant       14.448         8.191         1.76
X1              -8.69         1.555        -5.59
X2             13.517         2.085         6.48

S = 3.773     R-sq = 90.1%      R-sq (adj) = 86.1%

Analysis of Variance

SOURCE             DF          SS           MS          F
Regression          2      648.83      324.415      22.79
Residual Error      5       71.17       14.234
Total               7      720.00
```

b. $F_{.05} = 5.79$ (5 DF)

$F = 22.79 > F_{.05}$; significant relationship.

c. $R^2 = \dfrac{\text{SSR}}{\text{SST}} = .901$

$R_a^2 = 1-(1-.901)\dfrac{7}{5} = .861$

good fit

d. $t_{.025} = 2.571$ (5 DF)

for β_1: $t = -5.59 < -2.571$; reject $H_0 : \beta_1 = 0$

for β_2: $t = 6.48 > 2.571$; reject $H_0 : \beta_2 = 0$

36. a. MODEL 1:

There are a number of problems with this model – principally in terms of multicollinearity as indicated by the very high VIF values (all much greater than 10) for every predictor variable. In effect, the predictor variables are so inter-correlated it is infeasible to use them together in the same regression model. With multicollinearity the t tests become unreliable and should be ignored. Not so the F statistic (with a pvalue = 0.019 < α = 0.05) from the ANOVA table which indicates significance at the 5% level. This suggests that modelling might be more productive if only a subset of the original five predictors is considered.

The set of x values used to forecast y bears little relation to the sample in general (see sample means below) – making the results obtained effectively unusable.

	MEAN
x1	7.377
x2	7.157
x3	10.293
x4	10.943
x5	15.130

Amongst the residuals there are two outliers. For a sample size 30, this number may not seem as serious as it first appears since we would expect about 1 in 20 residuals to be outliers even if the dataset was well-behaved. However because of the relatively high values (in absolute value terms) of standardized residuals, these do look problematic and should be investigated.

MODEL 2:
This is a simple regression model. As such – unlike MODEL 1 – there is no possibility of problems of multicollinearity. The zero pvalues indicate the model is highly significant. In particular we would reject $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. Again observations 4 and 6 are associated with markedly outlying residuals so as before this should be carefully checked.

b. From the previous discussion it is clear MODEL 2 is technically sounder than MODEL 1 and would therefore be preferred.

37. a. The model is affected by multicollinearity according to the VIF values for variables weight and fat in particular. This means the t test results for individual predictors cannot be relied on. The F test from the ANOVA table is not significant at the 5% level (pvalue = 0.114 > α = 0.05) and the **adj** R^2 is a very unimpressive 15.6%. One residual – that for observation 9 – is also highlighted as an outlier.

b. From (a) this does not look a very effective predictive modeling aid. Even if the multicollinearity problem could be addressed by dropping correlated predictors, it would still appear to hold little promise.

38. a. $E(y) = \dfrac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$

b. A portion of the Minitab binary logistic regression output follows:

```
Logistic Regression Table
                                                           Odds
95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower
Upper
Constant      -2.805      1.432    -1.96 0.050
Price         1.1492     0.5143     2.23 0.025     3.16     1.15
8.65

Log-Likelihood = -8.200
Test that all slopes are zero: G = 9.465, DF = 1, P-Value = 0.002
```

Thus, the estimated logit is $\hat{g}(x) = -2.805 + 1.1492x$

c. For chocolates that have a price per serving of $4.00

$\hat{g}(4) = -2.805 + 1.1492(4) = 1.7918$

and

$$\hat{y} = \frac{e^{\hat{g}(4)}}{1 + e^{\hat{g}(4)}} = \frac{e^{1.7918}}{1 + e^{1.7918}} = \frac{6.0002}{1 + 6.0002} = 0.86$$

d. Using the Minitab output shown in part (b), the estimated odds ratio is 3.16. We can conclude that the estimated odds of having a quality rating of very good or excellent for a chocolate that has a price of $4.00 per serving is 3.16 times greater than the estimated odds for a chocolate with a price of $3.00 per serving. Moreover, this interpretation is true for any one dollar difference in the price per serving.

39. a. The expected increase in final college grade point average corresponding to a one point increase in high school grade point average is .0235 when SAT mathematics score does not change. Similarly, the expected increase in final college grade point average corresponding to a one point increase in the SAT mathematics score is .00486 when the high school grade point average does not change.

b. $\hat{y}$ = -1.41 + .0235(84) + .00486(540) = 3.19

40. a. The regression equation is

```
Y = -1.41 + .0235 X1 + .00486 X2

Predictor         Coef       SE Coef            T
Constant       -1.4053        0.4848        -2.90
X1            0.023467      0.008666         2.71
X2              .00486      0.001077         4.51

S = 0.1298    R-sq = 93.7%      R-sq (adj) = 91.9%

Analysis of Variance

SOURCE             DF          SS           MS          F
Regression          2     1.76209         .881      52.44
Residual Error      7       .1179        .0168
Total               9     1.88000
```

b. Using *F* table (2 degrees of freedom numerator and 7 degrees of freedom denominator), *p*-value is less than .01

Because *p*-value $\leq \alpha$, there is a significant relationship.

c. $R^2 = \dfrac{\text{SSR}}{\text{SST}} = .937$

$R_a^2 = 1-(1-.937)\dfrac{9}{7} = .919$

good fit

d. $t_{.025} = 2.365$ (7 DF)

for β_1: p-value is between .02 and .05; reject H_0: $\beta_1 = 0$

for β_2: p-value is less than .01; reject H_0: $\beta_2 = 0$

41. a. The Minitab output is shown below:

```
The regression equation is
Revenue = 33.3 + 7.98 Cars

Predictor       Coef     SE Coef          T        P
Constant       33.34       83.08       0.40    0.695
Cars          7.9840      0.6323      12.63    0.000

S = 226.7       R-Sq = 92.5%     R-Sq(adj) = 91.9%


Analysis of Variance

Source          DF          SS          MS          F
P
Regression       1     8192067     8192067     159.44
0.000
Residual Error  13      667936       51380
Total           14     8860003
```

b. An increase of 1000 cars in service will result in an increase in revenue of $7.98 million.

c. The Minitab output is shown below:

```
The regression equation is
Revenue = 106 + 8.94 Cars - 0.191 Location

Predictor       Coef     SE Coef          T        P
Constant      105.97       85.52       1.24    0.239
Cars          8.9427      0.7746      11.55    0.000
Location     -0.1914      0.1026      -1.87    0.087

S = 207.7       R-Sq = 94.2%     R-Sq(adj) = 93.2%

Analysis of Variance

Source          DF          SS          MS          F
P
Regression       2     8342186     4171093      96.66
0.000
Residual Error  12      517817       43151
Total           14     8860003
```

42. a. The Minitab output is shown below:

```
The regression equation is
Rating = 0.345 + 0.255 TradeEx + 0.132 Use + 0.459 Range

Predictor        Coef     SE Coef          T        P
Constant       0.3451      0.5307       0.65    0.540
TradeEx       0.25482     0.08556       2.98    0.025
Use            0.1325      0.1404       0.94    0.382
Range          0.4585      0.1232       3.72    0.010

S = 0.2431      R-Sq = 88.6%     R-Sq(adj) = 82.8%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     2.74541     0.91514     15.49    0.003
Residual Error     6     0.35459     0.05910
Total              9     3.10000
```

b. Because the *p*-value = .003 < α = .05, there is a significant relationship.

c. For TradeEx: Because the *p*-value = .025 < α = .05, TradeEx is significant.

For Use: Because the *p*-value = .382 > α = .05, Use is not significant.

For Range: Because the *p*-value = .010 < α = .05, Range is significant.

The Minitab output after removing Use is shown below:

```
The regression equation is
Rating = 0.672 + 0.264 TradeEx + 0.485 Range

Predictor        Coef     SE Coef          T        P
Constant       0.6718      0.3989       1.68    0.136
TradeEx       0.26406     0.08432       3.13    0.017
Range          0.4853      0.1189       4.08    0.005

S = 0.2412      R-Sq = 86.9%     R-Sq(adj) = 83.1%


Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      2.6928      1.3464     23.15    0.001
Residual Error     7      0.4072      0.0582
Total              9      3.1000
```

The coefficient of determination for the estimated regression equation developed in part (a) is .886. After the removal of Use, the coefficient of determination is .869. There is very little difference in the fit provided by the two estimated regression equations. But, because Use is not significant, this result is as expected.

43. Note: The Minitab output is shown with the solution to Exercise 10.

a. Since the *p*-value corresponding to $F = 96.66$ is $0.000 < \alpha = .05$, there is a significant relationship among the variables.

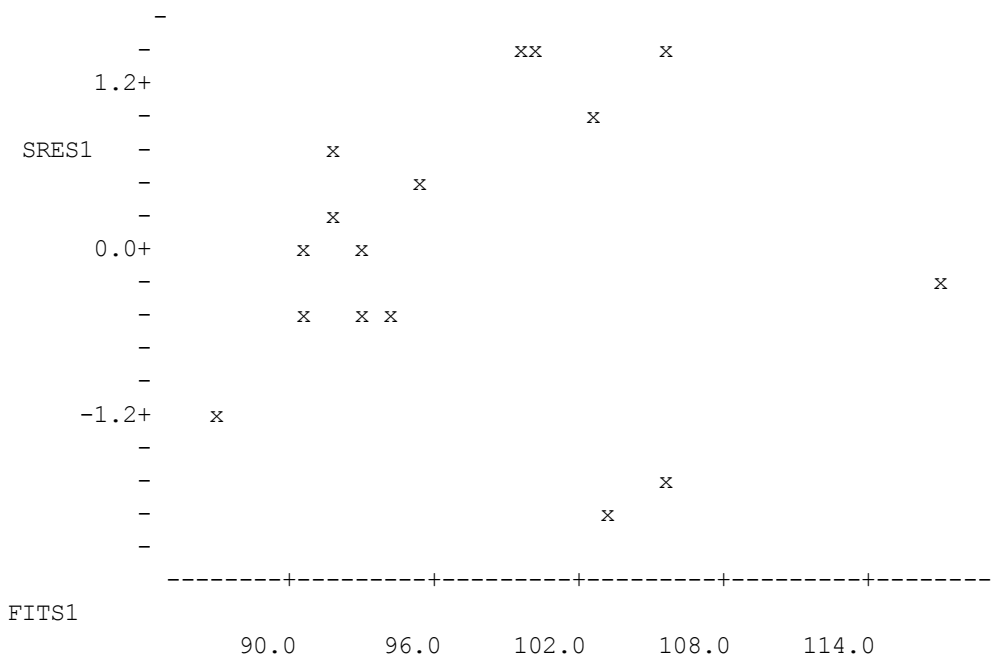
b. For Cars: Since the *p*-value = $0.000 < \alpha = 0.05$, Cars is significant

c. For Location: Since the *p*-value = $0.087 > \alpha = 0.05$, Location is not significant

44. a. The Minitab output is shown below:

```
The regression equation is
Speed = 71.3 + 0.107 Price + 0.0845 Horsepwr

Predictor        Coef     SE Coef          T        P
Constant       71.328       2.248      31.73    0.000
Price         0.10719     0.03918       2.74    0.017
Horsepwr     0.084496    0.009306       9.08    0.000

S = 2.485       R-Sq = 91.9%     R-Sq(adj) = 90.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      915.66      457.83     74.12    0.000
Residual Error    13       80.30        6.18
Total             15      995.95

Source       DF      Seq SS
Price         1      406.39
Horsepwr      1      509.27

Unusual Observations
Obs      Price     Speed      Fit     SE Fit   Residual   St
Resid
  2       93.8   108.000  105.882      2.007      2.118     1.45
X

X denotes an observation whose X value gives it large influence.
```

b. The standardized residual plot is shown below. There appears to be a very unusual trend in the standardized residuals.



c. The Minitab output shown in part (a) did not identify any observations with a large standardized residual; thus, there does not appear to be any outliers in the data.

d. The Minitab output shown in part (a) identifies observation 2 as an influential observation.

45. a. The Minitab output is shown below:

```
The regression equation is
Price = 356 - 0.0987 Capacity + 123 Comfort

Predictor        Coef     SE Coef          T        P
Constant        356.1       197.2       1.81    0.114
Capacity     -0.09874     0.04588      -2.15    0.068
Comfort        122.87       21.80       5.64    0.001

S = 51.14       R-Sq = 83.2%     R-Sq(adj) = 78.4%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2       90548       45274     17.31    0.002
Residual Error     7       18304        2615
Total              9      108852
```

b. $b_1 = -.0987$ is an estimate of the change in the price with respect to a 1 cubic inch change in capacity with the comfort rating held constant. $b_2 = 123$ is an estimate of the change in the price with respect to a 1 unit change in the comfort rating with the capacity held constant.

c. $\hat{y} = 356 - .0987(4500) + 123\,(4) = 404$

46. a. Since weight is not statistically significant (see Exercise 24), we will use an estimated regression equation which uses only Horsepower to predict the speed at 1/4 mile. The Minitab output is shown below:

```
The regression equation is
Speed = 72.6 + 0.0968 Horsepwr

Predictor        Coef     SE Coef          T        P
Constant       72.650       2.655      27.36    0.000
Horsepwr     0.096756    0.009865       9.81    0.000

S = 3.006       R-Sq = 87.3%     R-Sq(adj) = 86.4%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      869.43      869.43     96.21    0.000
Residual Error    14      126.52        9.04
Total             15      995.95

Unusual Observations
Obs   Horsepwr    Speed      Fit    SE Fit    Residual    St
Resid
  2        290  108.000  100.709     0.814       7.291
2.52R
  6        450  116.200  116.190     2.036       0.010      0.00
X

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

The output shows that the point estimate is a speed of 101.290 miles per hour.

b. The 95% confidence interval is 99.490 to 103.089 miles per hour.

c. The 95% prediction interval is 94.596 to 107.984 miles per hour.

47. a. $E(y) = \dfrac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$

b. For a given GPA, it is an estimate of the probability that a student who did not attend the orientation program will return to Lakeland for the sophomore year.

c. A portion of the Minitab binary logistic regression output follows:

```
Logistic Regression Table
                                                           Odds
95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower
Upper
Constant      -6.893      1.747    -3.94 0.000
GPA           2.5388     0.6729     3.77 0.000    12.66     3.39
47.35
Program       1.5608     0.5631     2.77 0.006     4.76     1.58
14.36

Log-Likelihood = -40.169
Test that all slopes are zero: G = 47.869, DF = 2, P-Value =
0.000
```

Thus, the estimated logit is $\hat{g}(x_1, x_2) = -6.893 + 2.5388x_1 + 1.5608x_2$

d. Significant result: the *p*-value corresponding to the *G* test statistic is 0.0000.

e. Both variables are significant at $\alpha = .01$: the *p*-value for X_1 is 0.000 and the *p*-value for X_2 is 0.006

f. For $X_1 = 2.5$ and $X_2 = 0$

$\hat{g}$ (2.5, 0) = -6.893 + 2.5388(2.5) + 1.5608(0) = -0.5460

and

$$\hat{y} = \frac{e^{\hat{g}(2.5,0)}}{1+e^{\hat{g}(2.5,0)}} = \frac{e^{-0.5460}}{1+e^{-0.5460}} = \frac{0.5793}{1+0.5793} = 0.37$$

For $X_1 = 2.5$ and $X_2 = 1$

$\hat{g}$ (2.5, 1) = -6.893 + 2.5388(2.5) + 1.5608(1) = 1.0148

and

$$\hat{y} = \frac{e^{\hat{g}(2.5,1)}}{1+e^{\hat{g}(2.5,1)}} = \frac{e^{1.0148}}{1+e^{1.0148}} = \frac{2.7588}{1+2.7588} = 0.73$$

g. From the Minitab output in part (c) we see that the estimated odds ratio is 4.76 for the orientation program. This means that the odds of students who attended the orientation program continuing are 4.76 times greater than for students who did not attend the program.

h. We recommend making the orientation program required. From part (e), we see that the odds of continuing are much higher for students who have attended the orientation program.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Sixteen

Regression Analysis: Model Building

Textbook Exercises (1-18)

Textbook Exercise Solutions

Supplementary Exercises (19-33)

Supplementary Exercise Solutions

Chapter 16: Regression Analysis: Model Building

Textbook Exercises:

1. Consider the following data for two variables, X and Y.

x	22	24	26	30	35	40
y	12	21	33	35	40	36

a. Develop an estimated regression equation for the data of the form $\hat{y} = b_0 + b_1x$.

b. Use the results from part (a) to test for a significant relationship between X and Y. Use $\alpha = 0.05$.

c. Develop a scatter diagram for the data. Does the scatter diagram suggest an estimated regression equation of the form $\hat{y} = b_0 + b_1x + b_2x^2$? Explain.

d. Develop an estimated regression equation for the data of the form $\hat{y} = b_0 + b_1x + b_2x^2$.

e. Refer to part (d). Is the relationship between X, X^2, and Y significant? Use $\alpha = 0.05$.

f. Predict the value of Y when $X = 25$.

2. Consider the following data for two variables, X and Y.

x	9	32	18	15	26
y	10	20	21	16	22

a. Develop an estimated regression equation for the data of the form $\hat{y} = b_0 + b_1x$. Comment on the adequacy of this equation for predicting Y.

b. Develop an estimated regression equation for the data of the form $\hat{y} = b_0 + b_1x + b_2x_2$.

Comment on the adequacy of this equation for predicting Y.

c. Predict the value of Y when $X = 20$.

3. Consider the following data for two variables, *X* and *Y*.

x	2	3	4	5	7	7	7	8	9
y	4	5	4	6	4	6	9	5	11

a. Does there appear to be a linear relationship between *X* and *Y*? Explain.

b. Develop the estimated regression equation relating *X* and *Y*.

c. Plot the standardized residuals versus for the estimated regression equation developed in part (b). Do the model assumptions appear to be satisfied? Explain.

d. Perform a logarithmic transformation on the dependent variable *Y*. Develop an estimated regression equation using the transformed dependent variable. Do the model assumptions appear to be satisfied by using the transformed dependent variable? Does a reciprocal transformation work better in this case? Explain.

4. The table below lists the total estimated numbers of AIDS cases, by year of diagnosis from 1999 to 2003 in the United States (Source: US Dept. of Health and Human Services, Centers for Disease Control and Prevention, HIV/AIDS Surveillance, 2003.)

Year	**AIDS Cases**
1999	41,356
2000	41,267
2001	40,833
2002	41,289
2003	43,171

a. Plot the data, letting x = 0 correspond to the year 1998,

Find a linear function $\hat{y} = b_0 + b_1 x$ that models the data,

b. Plot the function on the graph with the data and determine how well the graph fits the data.

5. In working further with the problem of exercise 4, statisticians suggested the use of the following curvilinear estimated regression equation.

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

a. Use the data of exercise 4 to determine estimated regression equation.

b. Use $\alpha = 0.01$ to test for a significant relationship.

6. An international study of life expectancy by Ross (1994) covers variables

LifeExp	Life expectancy in years
People.per.TV	Average number of people per TV
People.per.Dr	Average number of people per physician
LifeExp.Male	Male life expectancy in years
LifeExp.Female	Female life expectancy in years

With data details as follows:

	LifeExp	People. per.TV	People. per.Dr	LifeExp. Male	LifeExp. Female
Argentina	70.5	4	370	74	67
Bangladesh	53.5	315	6 166	53	54
Brazil	65	4	684	68	62
Canada	76.5	1.7	449	80	73
China	70	8	643	72	68
Colombia	71	5.6	1 551	74	68
Egypt	60.5	15	616	61	60

	LifeExp	People. per.TV	People. per.Dr	LifeExp. Male	LifeExp. Female
Ethiopia	51.5	503	36 660	53	50
France	78	2.6	403	82	74
Germany	76	2.6	346	79	73
India	57.5	44	2 471	58	57
Indonesia	61	24	7 427	63	59
Iran	64.5	23	2 992	65	64
Italy	78.5	3.8	233	82	75
Japan	79	1.8	609	82	76
Kenya	61	96	7 615	63	59
Korea.North	70	90	370	73	67
Korea.South	70	4.9	1 066	73	67
Mexico	72	6.6	600	76	68
Morocco	64.5	21	4 873	66	63
Burma	54.5	592	3 485	56	53
Pakistan	56.5	73	2 364	57	56
Peru	64.5	14	1 016	67	62
Philippines	64.5	8.8	1 062	67	62
Poland	73	3.9	480	77	69
Romania	72	6	559	75	69
Russia	69	3.2	259	74	64
South.Africa	64	11	1 340	67	61
Spain	78.5	2.6	275	82	75
Sudan	53	23	12 550	54	52
Taiwan	75	3.2	965	78	72
Tanzania	52.5	NA	25 229	55	50
Thailand	68.5	11	4 883	71	66
Turkey	70	5	1 189	72	68
Ukraine	70.5	3	226	75	66
UK	76	3	611	79	73
USA	75.5	1.3	404	79	72
Venezuela	74.5	5.6	576	78	71
Vietnam	65	29	3 096	67	63
Zaire	54	NA	23 193	56	52

(Note that the average number of people per TV is not given for Tanzania and Zaire.)

(Note that the average number of people per TV is not given for Tanzania and Zaire.)

a. Develop scatter diagrams for these data, treating LifeExp as the dependent variable.

b. Does a simple linear model appear to be appropriate? Explain.

c. Estimate simple regression equations for the data accordingly. Which do you prefer and why?

7. To assess the reliability of computer media, *Choice* magazine (www.choice.com.au) has obtained data by:

price (A$)	Paid in April 2005
pack	the number of disks in the pack
media	one of CD (CD), DVD (DVD-R) or DVDRW (DVD+/-RW)

with details as follows:

Price	Pack	Media	Price	Pack	Media
0.48	50	CD	1.85	10	DVD
0.60	25	CD	0.72	25	DVD
0.64	25	CD	2.28	10	DVD
0.50	50	CD	2.34	5	DVD
0.89	10	CD	2.40	10	DVD
0.89	10	CD	1.49	5	DVD
1.20	10	CD	3.60	5	DVDRW
1.30	10	CD	5.00	10	DVDRW
1.29	10	CD	2.79	5	DVDRW
0.50	10	CD	2.79	10	DVDRW
0.57	50	DVD	4.37	5	DVDRW
2.60	10	DVD	1.50	10	DVDRW
1.59	10	DVD	2.50	5	DVDRW
1.85	10	DVD	3.90	10	DVDRW

a. Develop scatter diagrams for these data with pack and media as potential independent variables.

b. Does a simple or multiple linear regression model appear to be appropriate?

c. Develop an estimated regression equation for the data you believe will best explain the relationship between these variables.

8. In Europe the number of Internet users varies widely from country to country. In 1999, 44.3 per cent of all Swedes used the Internet, while in France the audience was less than 10 per cent. The disparities are expected to persist even though Internet usage is expected to grow dramatically over the next several years. The following table shows the number of Internet users in 1999 and in 2011 for selected European countries. (http://www.internetworldstats.com/top25.htm)

File "INTERNET2011"

	% Internet users	
	1999	2011
Austria	12.6	74.8
Belgium	24.2	81.4
Denmark	40.4	89.0
Finland	40.9	88.6
France	9.7	77.2
Germany	15.0	82.7
Ireland	12.1	66.8
Netherlands	18.6	89.5
Norway	38.0	97.2
Spain	7.4	65.6
Sweden	44.3	92.9
Switzerland	28.1	84.2
UK	23.6	84.5

a. Develop a scatter diagram of the data using the 1999 Internet user percentage as the independent variable. Does a simple linear regression model appear to be appropriate? Discuss.
b. Develop an estimated multiple regression equation with X = the number of 1999 Internet users and X^2 as the two independent variables.
c. Consider the nonlinear relationship shown by equation (16.6). Use logarithms to develop an estimated regression equation for this model.
d. Do you prefer the estimated regression equation developed in part (b) or part (c)? Explain.

9. In a regression analysis involving 27 observations, the following estimated regression equation was developed.

$$\hat{y} = 25.2 + 5.5x_1$$

For this estimated regression equation SST = 1550 and SSE = 520.

a. At $\alpha = 0.05$, test whether X_1 is significant.

Suppose that variables X_2 and X_3 are added to the model and the following regression equation is obtained.

$$\hat{y} = 16.3 + 2.3x_1 + 12.1x_2 - 5.8x_3$$

For this estimated regression equation SST = 1550 and SSE = 100.

b. Use an F test and a 0.05 level of significance to determine whether X_2 and X_3 contribute significantly to the model.

10. In a regression analysis involving 30 observations, the following estimated regression equation was obtained.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

For this estimated regression equation SST = 1805 and SSR = 1760.

a. At $\alpha = 0.05$, test the signifi cance of the relationship among the variables.

Suppose variables X_1 and X_4 are dropped from the model and the following estimated regression equation is obtained.

$$\hat{y} = 11.1 - 3.6x_2 + 8.1x_3$$

For this model SST = 1805 and SSR = 1705.

b. Compute SSE(x_1, x_2, x_3, x_4).

c. Compute SSE(x_2, x_3).

d. Use an F test and a 0.05 level of significance to determine whether X_1 and $X4$ contribute significantly to the model.

11. In an experiment involving measurements of Heat Production (calories) at various Body Masses (kgs) and Work levels (Calories/hour) on a stationary bike, the following results were obtained:

Body Mass (M)	Work level (W)	Heat production (H)
43.7	19	177
43.7	43	279
43.7	56	346
54.6	13	160
54.6	19	193
54.6	43	280
54.6	56	335
55.7	13	169
55.7	26	212
55.7	34.5	244
55.7	43	285
58.8	13	181
58.8	43	298
60.5	19	212
60.5	43	317
60.5	56	347
61.9	13	186
61.9	19	216
61.9	34.5	265
61.9	43	306
61.9	56	348
66.7	13	209
66.7	43	324
66.7	56	352

File “MUSCLE”

a. Develop an estimated regression equation that can be used to predict Heat production for a given Body Mass and Work level.
b. Consider adding an independent variable to the model developed in part (a) for the interaction between Body Mass and Work level. Develop an estimated regression equation using these three independent variables.
c. At a 0.05 level of significance, test to see whether the addition of the interaction term contributes significantly to the estimated regression equation developed in part (a).

12. Failure data obtained in the course of the development of a silver-zinc battery for a NASA programme were analyzed by Sidik, Leibecki and Bozek in 1980. Relevant variables were as follows:

$x1$	charge rate (amps):
$x2$	discharge rate (amps)
$x3$	depth of discharge (% of rated ampere – hours)
$x4$	temperature (°C)
$x5$	end of charge voltage (volts)
y	cycles to failure

Adopting ln(y) as the response variable, a number of regression models were estimated for the data using MINITAB:

Regression Analysis: lny versus x1, x2, x3, x4, x5

The regression equation is
lny = 63.7 - 0.459 x1 - 0.327 x2 - 0.0111 x3 + 0.116 x4 + 33.8 x5

Predictor	Coef	SE Coef	T	P	VIF
Constant	-63.68	51.18	-1.24	0.234	
x1	-0.4593	0.5493	-0.84	0.417	1.1
x2	-0.3267	0.1761	-1.85	0.085	1.0
x3	-0.01113	0.01699	-0.66	0.523	1.1
x4	0.11577	0.02499	4.63	0.000	1.0
x5	33.81	25.59	1.32	0.208	1.0

S = 1.070 R-Sq = 66.3% R-Sq(adj) = 54.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	31.578	6.316	5.52	0.005
Residual Error	14	16.032	1.145		
Total	19	47.610			

Source	DF	Seq SS
x1	1	1.464
x2	1	4.512
x3	1	0.291
x4	1	23.311
x5	1	1.999

Unusual Observations

Obs	x1	lny	Fit	StDev Fit	Residual	St Resid
1	0.38	4.615	6.708	0.651	-2.093	-2.46R

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 1.72

Regression Analysis: lny versus x4

The regression equation is lny = 1.78 + 0.114 x4

Predictor	Coef	SE Coef	T	P
Constant	1.7777	0.5660	3.14	0.006
x4	0.11395	0.02597	4.39	0.000

S = 1.130 R-Sq = 51.7% R-Sq(adj) = 49.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	24.607	24.607	19.26	0.000
Residual Error	18	23.002	1.278		
Total	19	47.610			

Unusual Observations

Obs	x4	lny	Fit	StDev Fit	Residual	St Resid
12	10.0	0.693	2.917	0.353	-2.224	-2.07R

R denotes an observation with a large standardized residual

a. Explain this computer output, carrying out any additional tests you think necessary or appropriate.

b. Is the first model significantly better than the second?

c. Which model do you prefer and why?

13. A section of MINITAB output from an analysis of data relating to truck exhaust emissions under different atmospheric conditions (Hare and Bradow, 1977) is as follows:

Regression Analysis: nox versus humi, temp, HT

```
The regression equation is
nox = 1.61 - 0.0146 humi - 0.00681 temp + 0.000150 HT


Predictor        Coef     SE Coef      T      P
Constant       1.6104      0.2287   7.04  0.000
humi        -0.014572    0.003091  -4.71  0.000
temp        -0.006806    0.002889  -2.36  0.023
HT         0.00014985  0.00003733   4.01  0.000


S = 0.0595096   R-Sq = 71.5%   R-Sq(adj) = 69.4%


Analysis of Variance

Source          DF       SS       MS      F      P
Regression       3  0.35544  0.11848  33.46  0.000
Residual Error  40  0.14166  0.00354
Total           43  0.49710


Source  DF   Seq SS
humi     1  0.28446
temp     1  0.01392
HT       1  0.05706


Unusual Observations

Obs  humi      nox      Fit   SE Fit  Residual  St Resid
  6    13  1.11000  1.09407  0.03316   0.01593      0.32 X
 14    11  1.10000  0.99555  0.03105   0.10445      2.06R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.


Durbin-Watson statistic = 1.63335
```

Variables used in this analysis are defi ned as follows:

Nox	Nitrous oxides, NO and NO_2, (grams/km)
Humi	Humidity (grains H_2O/lbm dry air)
Temp	emperature (°F)
HT	humi × temp

a. Provide a descriptive summary of this information, carrying out any further calculations or statistical tests you think relevant or necessary.

b. It has been argued that the inclusion of quadratic terms

$$HH = humi \times humi$$

$$TT = temp \times temp$$

on the right hand side of the model will lead to a significantly improved *R*-square outcome. Details of the revised analysis are shown below. Is the claim justified?

Regression Analysis: nox versus humi, temp, HT, HH, TT

The regression equation is
nox = 2.69 - 0.0102 humi - 0.0371 temp + 0.000057 HT + 0.000022 HH
+ 0.000222 TT

Predictor	Coef	SE Coef	T	P
Constant	2.685	1.306	2.06	0.047
humi	-0.010167	0.003015	-3.37	0.002
temp	-0.03714	0.03414	-1.09	0.284
HT	0.00005662	0.00004073	1.39	0.173
HH	0.00002209	0.00000592	3.73	0.001
TT	0.0002221	0.0002224	1.00	0.324

S = 0.0515260 R-Sq = 79.7% R-Sq(adj) = 77.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	0.396213	0.079243	29.85	0.000
Residual Error	38	0.100887	0.002655		
Total	43	0.497100			

Source	DF	Seq SS
humi	1	0.284462
temp	1	0.013924
HT	1	0.057058
HH	1	0.038121
TT	1	0.002648

Unusual Observations

Obs	humi	nox	Fit	SE Fit	Residual	St Resid
5	10	0.99000	1.08738	0.02324	-0.09738	-2.12R
14	11	1.10000	1.08224	0.03654	0.01776	0.49 X
40	139	0.70000	0.82314	0.01759	-0.12314	-2.54R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 1.77873

14. A sample of 16 companies taken from the *Stock Investor Pro* database was used to obtain the following data on the price/earnings (P/E) ratio, the gross profit margin, and the sales growth for each company (*Stock Investor Pro*, American Association of Individual Investors, 21 August 1997). The data in the Industry column are codes used to define the industry for each company: 1 = energy-international oil; 2 = health-drugs; and 3 = other.

Firm	P/E ratio	Gross profit margin (%)	Sales growth (%)	Industry
Abbott Laboratories	22.3	23.7	10.0	2
American Home Products	22.6	21.1	5.3	2
Amoco	16.7	11.0	16.5	1
Bristol Meyers Squibb Co.	25.9	26.6	9.4	2
Chevron	18.3	11.6	18.4	1
Exxon	18.7	9.8	8.3	1
General Electric Company	13.1	13.4	13.1	3
Hewlett-Packard	23.3	9.7	21.9	3
IBM	17.3	11.5	5.6	3
Merck & Co. Inc.	26.2	25.6	18.9	2
Mobil	18.7	8.2	8.1	1
Pfizer	34.6	25.1	12.8	2
Pharmacia & Upjohn, Inc.	22.3	15.0	2.7	2
Procter & Gamble Co.	5.4	14.9	5.4	3
Texaco	12.3	7.3	23.7	1
Travelers Group Inc.	28.7	17.8	28.7	3

Develop an estimated regression equation that can be used to predict price/earnings ratio. Briefly discuss the process you used to develop a recommended estimated regression equation for these data.

15. A sales executive is interested in predicting sales of a newly released record (Field, 2005). Details are available for 200 individual past recordings as follows:

airplay	=	*number of times a record is played on Radio 1*
sales	=	*record sales (thousands)*
advert	=	*advertizing budget (£000s)*
attract	=	*attractiveness rating (1–10) of recording act*

Selective modelling details using MINITAB are given below:

Correlations: adverts, sales, airplay, attract

	adverts	sales	airplay
sales	0.578		
	0.000		
airplay	0.102	0.599	
	0.151	0.000	
attract	0.081	0.326	0.182
	0.256	0.000	0.010

Cell Contents: Pearson correlation
P-Value

Stepwise Regression: sales versus adverts, airplay, attract

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is sales on 3 predictors, with N = 200

Step	1	2	3
Constant	84.87	41.12	-26.61
airplay	3.94	3.59	3.37
T-Value	10.52	12.51	12.12
P-Value	0.000	0.000	0.000
adverts		0.0869	0.0849
T-Value		11.99	12.26
P-Value		0.000	0.000

```
attract                          11.1
T-Value                          4.55
P-Value                         0.000

S              64.8     49.4     47.1
R-Sq          35.87    62.93    66.47
R-Sq(adj)     35.55    62.55    65.95
Mallows Cp   178.8     22.7      4.0
```

Regression Analysis: sales versus adverts, airplay, attract

```
The regression equation is
sales = - 26.6 + 0.0849 adverts + 3.37 airplay + 11.1 attract


Predictor       Coef    SE Coef      T       P    VIF
Constant     -26.61      17.35   -1.53   0.127
adverts    0.084885   0.006923   12.26   0.000   1.015
airplay      3.3674     0.2778   12.12   0.000   1.043
attract      11.086      2.438    4.55   0.000   1.038


S = 47.0873   R-Sq = 66.5%   R-Sq(adj) = 66.0%


Analysis of Variance

Source            DF       SS       MS       F       P
Regression         3   861377   287126  129.50   0.000
Residual Error   196   434575     2217
Total            199  1295952


Source   DF  Seq SS
adverts   1  433688
airplay   1  381836
attract   1   45853


Unusual Observations

Obs  adverts    sales      Fit  SE Fit  Residual  St Resid
  1       10   330.00   229.92   10.23    100.08      2.18R
  2      986   120.00   228.95    4.21   -108.95     -2.32R
  7      472    70.00    91.87   14.21    -21.87     -0.49 X
 10      174   300.00   200.47    5.85     99.53      2.13R
 12      611    70.00   114.81   11.92    -44.81     -0.98 X
 47      103    40.00   154.97    5.90   -114.97     -2.46R
 52      406   190.00    92.60    8.05     97.40      2.10R
 55     1542   190.00   304.12    7.61   -114.12     -2.46R
 61      579   300.00   201.19    3.44     98.81      2.10R
 68       57    70.00   180.42    5.90   -110.42     -2.36R
100     1000   250.00   152.71    7.85     97.29      2.10R
138       30    60.00    81.34   14.79    -21.34     -0.48 X
164        9   120.00   241.32    9.34   -121.32     -2.63R
169      146   360.00   215.87    6.79    144.13      3.09R
181      179    70.00    63.65   14.33      6.35      0.14 X
184     2272   320.00   326.06   12.97     -6.06     -0.13 X
200      786   110.00   207.21    7.07    -97.21     -2.09R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 1.94982
```

Best Subsets Regression: sales versus adverts, airplay, attract

Response is sales

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	adverts	airplay	attract
1	35.9	35.5	178.8	64.787		X	
1	33.5	33.1	192.9	65.991	X		
2	62.9	62.6	22.7	49.383	X	X	
2	41.3	40.7	149.0	62.129	X		X
3	66.5	66.0	4.0	47.087	X	X	X

a. How would you interpret this information?

b. Which of the various models shown here do you favour and why?

16. In a study of car ownership in 24 countries, data (OECD, 1982) have been collected on the following variables:

ao	cars per person
pop	population (millions)
den	population density
gdp	*per capita* income ($)
pr	petrol price (cents per litre)
con	petrol consumption (tonnes per car per year)
tr	bus and rail use (passenger km per person)

Selective results from a linear modelling analysis (ao is the dependent variable) are as follows:

Best Subsets Regression: ao versus pop, den, gdp, pr, con, tr

```
Response is ao

                                                  p d g   c
                          Mallows                 o e d p o t
Vars  R-Sq  R-Sq(adj)       Cp         S         p n p r n r
   1  53.0       50.9     41.2  0.085534           X
   1  10.7        6.7     96.4   0.11791             X
   2  67.8       64.7     24.0  0.072526           X X
   2  67.3       64.2     24.6  0.073035           X     X
   3  72.5       68.4     19.8  0.068579           X X   X
   3  72.1       68.0     20.3  0.069090  X        X     X
   4  83.0       79.5      8.1  0.055298           X X X X
   4  77.1       72.3     15.8  0.064197  X        X   X X
   5  86.2       82.4      6.0  0.051208  X        X X X X
   5  83.2       78.5      9.9  0.056611    X      X X X X
   6  87.0       82.4      7.0  0.051270  X X      X X X X
```

Correlations: ao, pop, den, gdp, pr, con

```
         ao      pop      den      gdp       pr      con
pop   0.278
      0.188

den  -0.042    0.109
      0.846    0.612

gdp   0.728    0.057    0.193
      0.000    0.791    0.365

pr   -0.327   -0.437    0.338    0.076
      0.118    0.033    0.106    0.724

con   0.076    0.342   -0.357   -0.085   -0.723
      0.723    0.101    0.087    0.694    0.000

tr   -0.119   -0.025    0.397    0.328    0.483   -0.602
      0.581    0.906    0.055    0.118    0.017    0.002


Cell Contents: Pearson correlation
               P-Value
```

Regression Analysis: ao versus pop, gdp, pr, con, tr

```
The regression equation is
ao = 0.472 + 0.000521 pop + 0.0319 gdp - 0.00429 pr - 0.104 con - 0.0735 tr


Predictor       Coef    SE Coef       T      P
Constant     0.47190    0.09081    5.20  0.000
pop        0.0005211  0.0002556    2.04  0.056
gdp          0.031889   0.003423    9.32  0.000
pr          -0.004289   0.001245   -3.44  0.003
con          -0.10449    0.02626   -3.98  0.001
tr           -0.07354    0.01733   -4.24  0.000


S = 0.0512085   R-Sq = 86.2%   R-Sq(adj) = 82.4%
```

```
Analysis of Variance

Source           DF        SS        MS      F      P
Regression        5  0.295364  0.059073  22.53  0.000
Residual Error   18  0.047202  0.002622
Total            23  0.342565


Source  DF    Seq SS
pop      1  0.026535
gdp      1  0.174352
pr       1  0.033131
con      1  0.014144
tr       1  0.047202


Unusual Observations

Obs  pop      ao     Fit  SE Fit  Residual  St Resid
 11   57  0.3000  0.1914  0.0235    0.1086      2.39R
 23  218  0.5300  0.5178  0.0454    0.0122      0.52 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

a. Which of the various model options considered here do you prefer and why?

b. Corresponding stepwise output from MINITAB terminates after two stages, gdp being the first independent variable selected and pr the second. How does this latest information reconcile with that summarized earlier?

c. Does it alter in any way, your inferences for a.? If so, why and if not, why not?

17. In an analysis of the effects of rainfall, temperature and time of exposure on the ret loss of flax, the following MINITAB output has been obtained:

(Note: X_1 = Mean daily rainfall (0.01 inches per day)

X_2 = Retting period (days)

X_3 = Mean maximum daily temperature (°F)

Y = per cent ret loss of flax

Regression Analysis: y versus x1, x2, x3

The regression equation is
y = 10.8 + 1.81 x1 + 0.109 x2 + 0.0926 x3

Predictor	Coef	SE Coef	T	P	VIF
Constant	10.819	7.258	1.49	0.150	
x1	1.8101	0.5451	3.32	0.003	1.2
x2	0.10887	0.05858	1.86	0.076	1.5
x3	0.09263	0.09296	1.00	0.329	1.7

S = 2.197 R-Sq = 42.3% R-Sq(adj) = 34.7%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	3	81.285	27.095	5.61	0.005
Error	23	111.045	4.828		
Total	26	192.330			

SOURCE	DF	SEQ SS
x1	1	37.060
x2	1	39.430
x3	1	4.795

Unusual Observations

Obs.	x1	y	Fit	SE Fit	Residual	St. Resid
21	4.80	29.500	34.004	1.013	-4.504	-2.31R
24	5.40	38.900	34.050	0.890	4.850	2.41R

R denotes an obs. with a large st. resid.

Durbin-Watson statistic = 1.64

Stepwise Regression: y versus x1, x2, x3

Stepwise regression of y on 3 predictors, with N = 27

STEP	1	2
CONSTANT	27.39	16.42
x1	1.36	1.59
T-RATIO	2.44	3.20
x2		0.141
T-RATIO	2.86	
S	2.49	2.20
R-SQ	19.27	39.77

Best Subsets Regression: y versus x1, x2, x3

Best Subsets
Regression of y

Vars	R-sq	Adj. R-sq	C-p	s	x 1	x 2	x 3
1	19.3	16.0	9.2	2.4921	X		
1	14.1	10.7	11.2	2.5700		X	
2	39.8	34.8	3.0	2.1970	X	X	
2	33.6	28.1	5.5	2.3069	X		X
3	42.3	34.7	4.0	2.1973	X	X	X

a. How would you interpret this information?

b. Confirm details of any tests you carry out to support your inferences.

c. Which is your preferred model of those covered here?

18. A senior police manager is reviewing manpower allocation of police officers to a number of geographical districts which fall under her responsibility (Wisniewski, 2002). Detailed regression analysis results have been obtained involving the following variables:

Crimes	number of reported crimes
Officers	number of full-time equivalent police officers
Support	number of civilian support staff
Unemployment	unemployment rate (%) for the area
Retired	percentage of the local population who are retired

Selected MINITAB output is given below:

Correlations: Crimes, Officers, Support, Unemployment, Retired

	Crimes	Officers	Support	Unemployment
Officers	-0.735 0.000			
Support	0.259 0.202	-0.345 0.085		
Unemployment	0.760 0.000	-0.434 0.027	0.128 0.535	
Retired	-0.867 0.000	0.655 0.000	-0.138 0.501	-0.661 0.000

Cell Contents: Pearson correlation
P-Value

Best Subsets Regression: Crimes versus Officers, Support, ...

Response is Crimes

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	Officers	Support	Unemployment	Retired
1	75.1	74.0	16.8	100.55				X
1	57.7	55.9	43.9	131.05		X		
2	81.3	79.7	9.2	89.041		X		X
2	80.0	78.3	11.1	91.970	X			X
3	86.2	84.3	3.5	78.170	X	X		X
3	82.9	80.6	8.6	86.946		X	X	X
4	86.5	83.9	5.0	79.085	X	X	X	X

Stepwise Regression: Crimes versus Officers, Support, ...

Backward elimination. Alpha-to-Remove: 0.1

Response is Crimes on 4 predictors, with N = 26

Step	1	2
Constant	1344	1411
Officers	-14.1	-15.5
T-Value	-2.36	-2.80
P-Value	0.028	0.010
Support	10	
T-Value	0.70	
P-Value	0.490	
Unemployment	17.0	17.1
T-Value	3.06	3.14
P-Value	0.006	0.005
Retired	-20.6	-20.0
T-Value	-3.64	-3.62
P-Value	0.002	0.002
S	79.1	78.2
R-Sq	86.52	86.20
R-Sq(adj)	83.95	84.32
Mallows Cp	5.0	3.5

If a new variable *Total staff = Officers + Support* is created and a further analysis undertaken, the following results are obtained.

Regression Analysis: Crimes versus Unemployment, Retired, Total staff

```
The regression equation is
Crimes = 1433 + 17.4 Unemployment - 21.5 Retired - 13.4 Total staff

Predictor        Coef  SE Coef      T      P
Constant       1433.0    164.6   8.71  0.000
Unemployment   17.412    5.786   3.01  0.006
Retired       -21.511    5.883  -3.66  0.001
Total staff   -13.398    6.205  -2.16  0.042

S = 82.7009   R-Sq = 84.6%   R-Sq(adj) = 82.4%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       3  823591  274530  40.14  0.000
Residual Error  22  150468    6839
Total           25  974059

Source        DF  Seq SS
Unemployment   1  561901
Retired        1  229809
Total staff    1   31881

Durbin-Watson statistic = 2.22341
```

a. Explain this computer output, carrying out any additional tests you think necessary or appropriate.

b. Is the last model a significant improvement on the corresponding two predictor model (best subsets option with R2 _ 81.3 per cent) for which details were summarized earlier?

c. Which of the various models shown do you prefer and why?

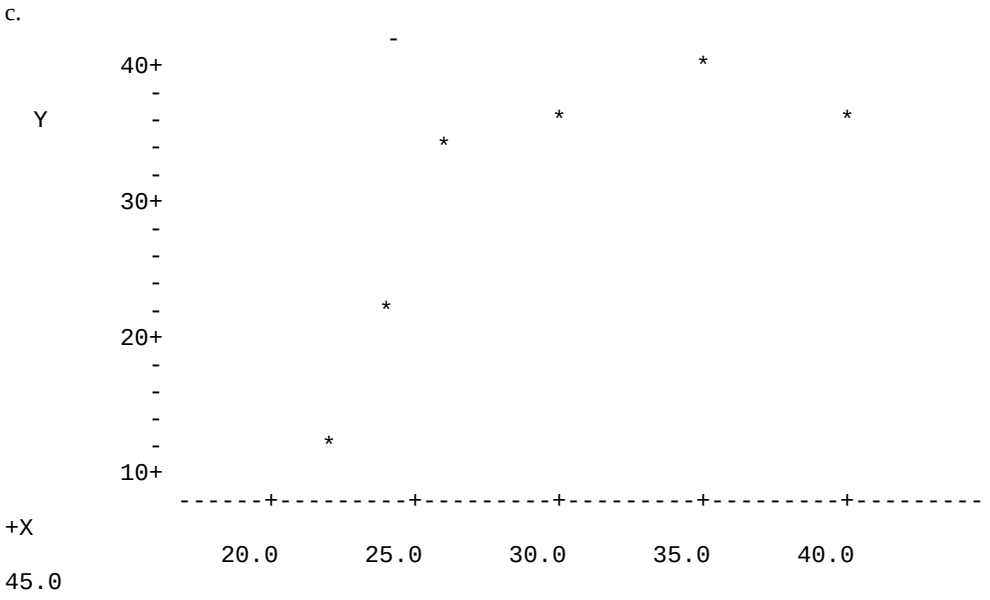
Textbook Exercises Solutions:

1. a. The Minitab output is shown below:

```
The regression equation is
Y = - 6.8 + 1.23 X

Predictor       Coef     SE Coef          T        p
Constant       -6.77       14.17      -0.48    0.658
X             1.2296      0.4697       2.62    0.059

S = 7.269       R-sq = 63.1%     R-sq(adj) = 53.9%

Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     1      362.13      362.13      6.85    0.059
Residual Error 4      211.37       52.84
Total          5      573.50
```

b. Since the *p*-value corresponding to $F = 6.85$ is $0.59 >$ [illegible] the relationship is not significant.

c.



The scatter diagram suggests that a curvilinear relationship may be appropriate.

d. The Minitab output is shown below:

```
The regression equation is
Y = - 169 + 12.2 X - 0.177 XSQ

Predictor       Coef     SE Coef          T        p
Constant     -168.88       39.79      -4.24    0.024
X             12.187       2.663       4.58    0.020
XSQ          -0.17704     0.04290      -4.13    0.026

S = 3.248       R-sq = 94.5%     R-sq(adj) = 90.8%


Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     2      541.85      270.92     25.68    0.013
Residual Error 3       31.65       10.55
Total          5      573.50
```

e. Since the *p*-value corresponding to $F = 25.68$ is $.013 < \alpha = .05$, the relationship is significant.

f. $\hat{y} = -168.88 + 12.187(25) - 0.17704(25)^2 = 25.145$

2. a. The Minitab output is shown below:

```
The regression equation is
Y = 9.32 + 0.424 X

Predictor       Coef     SE Coef          T        p
Constant       9.315       4.196       2.22    0.113
X             0.4242      0.1944       2.18    0.117

S = 3.531       R-sq = 61.4%     R-sq(adj) = 48.5%

Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     1       59.39       59.39      4.76    0.117
Residual Error 3       37.41       12.47
Total          4       96.80
```

The high *p*-value (.117) indicates a weak relationship; note that 61.4% of the variability in *y* has been explained by *x*.

b. The Minitab output is shown below:

```
The regression equation is
```

```
Y = - 8.10 + 2.41 X - 0.0480 XSQ

Predictor       Coef     SE Coef          T        p
Constant      -8.101       4.104      -1.97    0.187
X             2.4127      0.4409       5.47    0.032
XSQ         -0.04797     0.01050      -4.57    0.045

S = 1.279       R-sq = 96.6%     R-sq(adj) = 93.2%

Analysis of Variance

SOURCE        DF          SS          MS         F
p
Regression     2      93.529      46.765     28.60
0.034
Residual Error 2       3.271       1.635
Total          4      96.800
```

At the .05 level of significance, the relationship is significant; the fit is excellent.

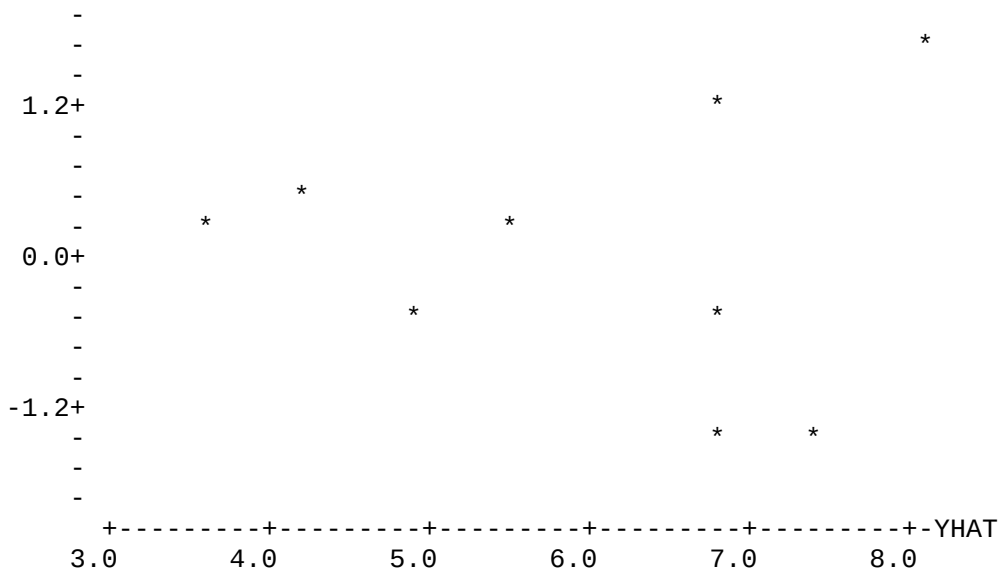
c. $\hat{y} = -8.101 + 2.4127(20) - 0.04797(20)^2 = 20.965$

3. a. The scatter diagram shows some evidence of a possible linear relationship.

b. The Minitab output is shown below:

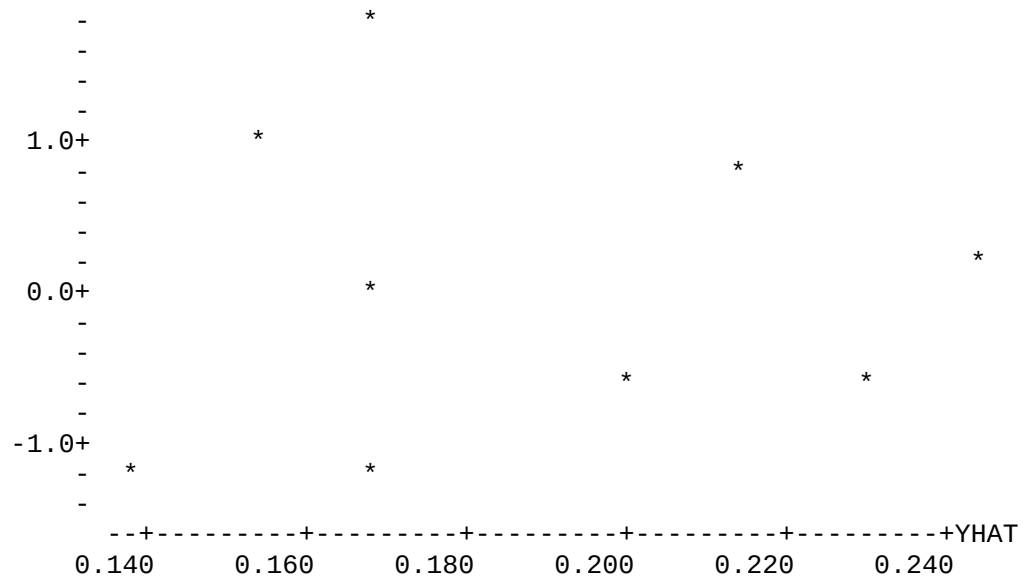
```
The regression equation is
Y = 2.32 + 0.637 X

Predictor       Coef     SE Coef          T        p
Constant       2.322       1.887       1.23    0.258
X             0.6366      0.3044       2.09    0.075

S = 2.054       R-sq = 38.5%     R-sq(adj) = 29.7%

Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     1      18.461      18.461      4.37    0.075
Residual Error 7      29.539       4.220
Total          8      48.000
```

c. The following standardized residual plot indicates that the constant variance assumption is not satisfied.

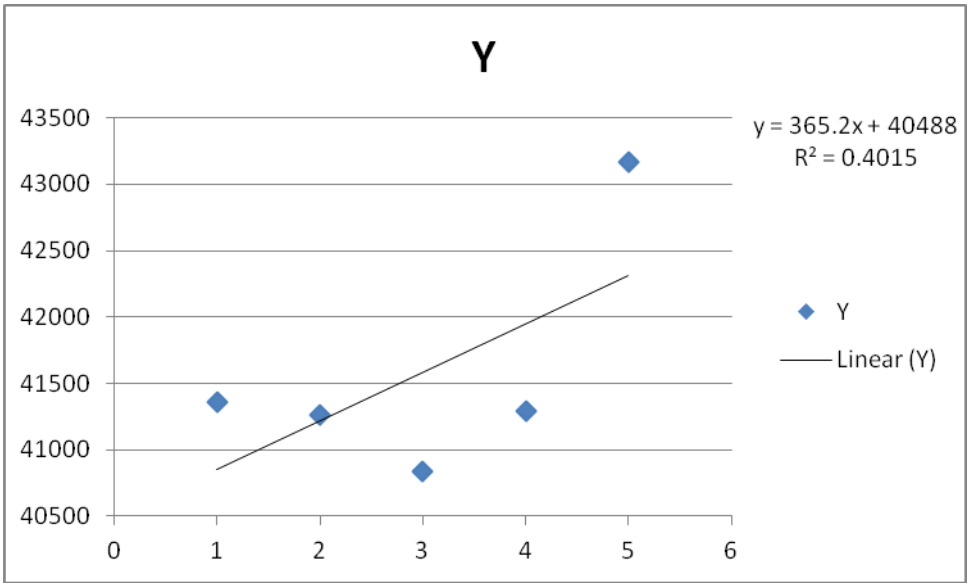


d. The logarithmic transformation does not appear to eliminate the wedged-shaped pattern in the above residual plot. The reciprocal transformation does, however, remove the wedge-shaped pattern. Neither transformation provides a good fit. The Minitab output for the reciprocal transformation and the corresponding standardized residual pot are shown below.

```
The regression equation is
1/Y = 0.275 - 0.0152 X

Predictor       Coef     SE Coef          T        p
Constant     0.27498     0.04601       5.98    0.000
X          -0.015182    0.007421      -2.05    0.080

S = 0.05009     R-sq = 37.4%     R-sq(adj) = 28.5%


Analysis of Variance

SOURCE        DF          SS          MS         F
p
Regression     1    0.010501    0.010501      4.19
0.080
Residual Error 7    0.017563    0.002509
Total          8    0.028064
```



4. a./b. The Minitab output is shown below:



The proposed linear function does not look to be a particularly good fit from the plot above. The relatively low R^2 of 40.15% would appear to corroborate this viewpoint.

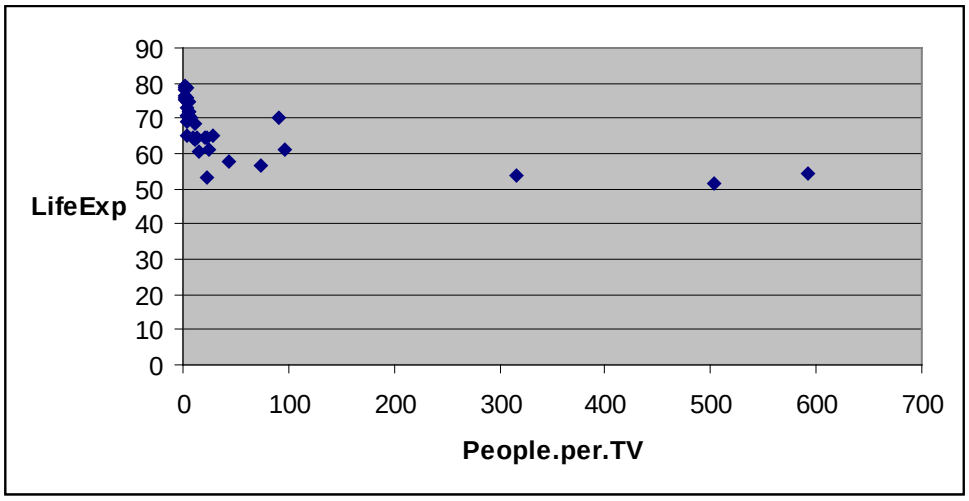
5. The Minitab output is shown below:

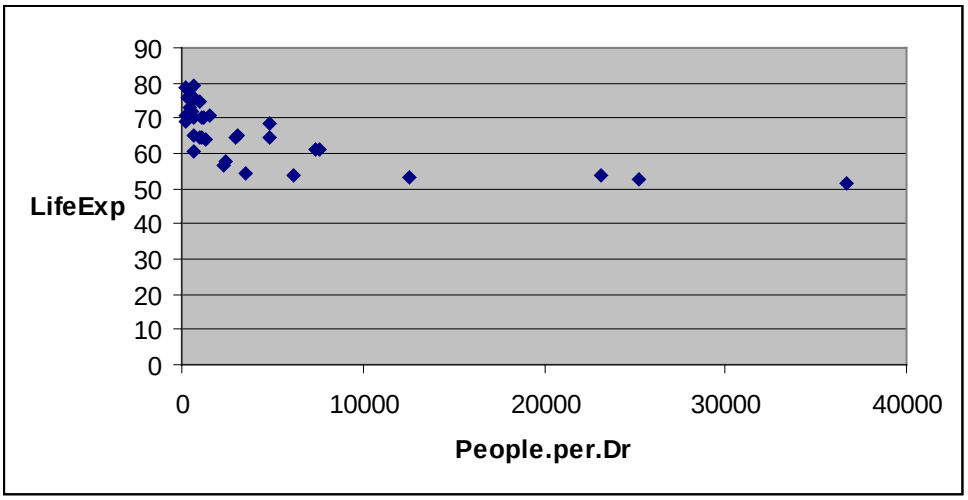
```
The regression equation is
Y = 433 + 37.4 X - 0.383 XSQ

Predictor       Coef     SE Coef          T        p
Constant       432.6       141.2       3.06    0.055
X             37.429       7.807       4.79    0.017
XSQ          -0.3829      0.1036      -3.70    0.034

S = 15.83       R-sq = 98.0%     R-sq(adj) = 96.7%


Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     2       36643       18322     73.15    0.003
Residual Error 3         751         250
Total          5       37395
```

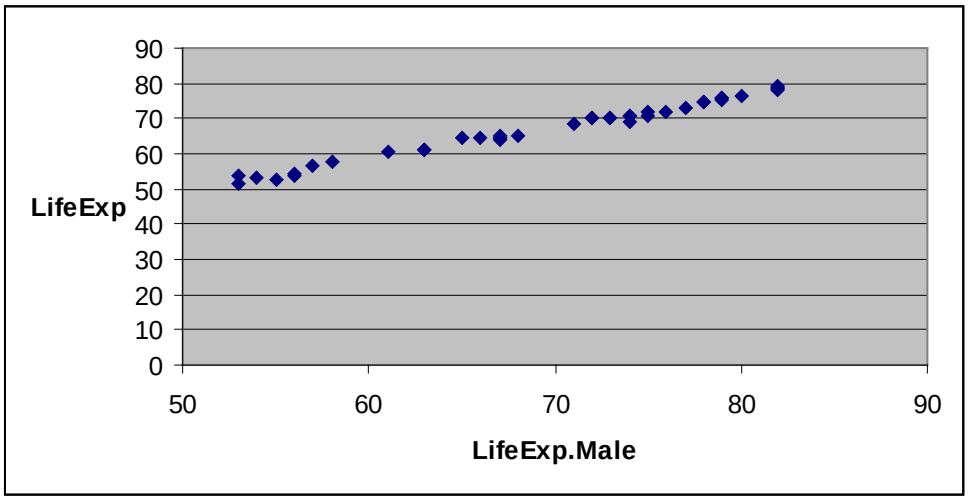
b. Since the linear relationship was significant (Exercise 4), this relationship must be significant. Note also that since the *p*-value of .003 $< \alpha = .05$, we can reject H_0.

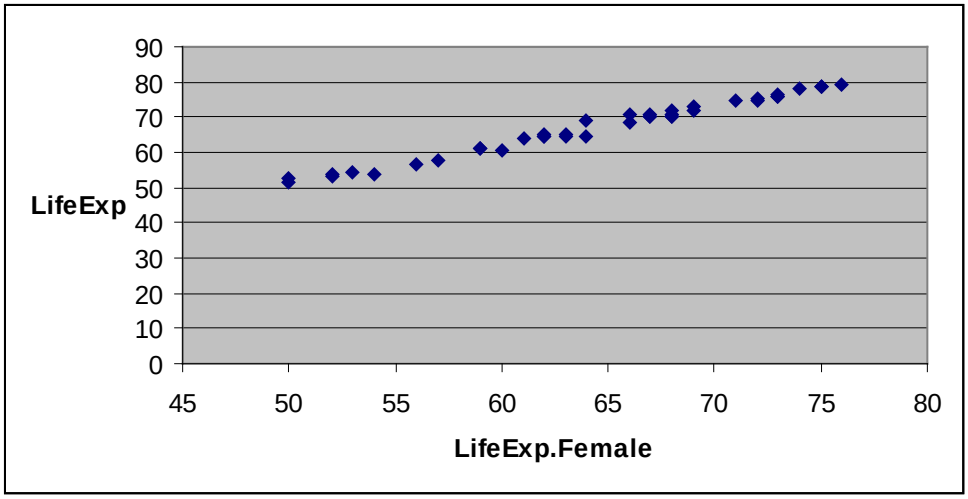
c. The fitted value is 1302.01, with a standard deviation of 9.93. The 95% confidence interval is 1270.41 to 1333.61; the 95% prediction interval is 1242.55 to 1361.47.

6. a. The scatter diagrams are shown below:









b. The relationship between LifeExp and LifeExp.Male is almost perfectly linear. Ditto the relationship between LifeExp and LifeExp.Female. This is

only to be expected since the values of the LifeExp.Male and LifeExp.Female values directly make up the corresponding LifeExp ones. In these circumstances using either LifeExp.Female and LifeExp.Male as predictors of LifeExp would make no real sense from a causal regression standpoint.

The other two variables (People.per.TV and People.per.Dr) of LifeExp – from the first two scattergrams above – do not look wholly convincing predictors in linear modelling terms (a hyperbolic or negative exponential fit might be more convincing). Nevertheless as part c. shows, significant linear regression models can be obtained in each case.

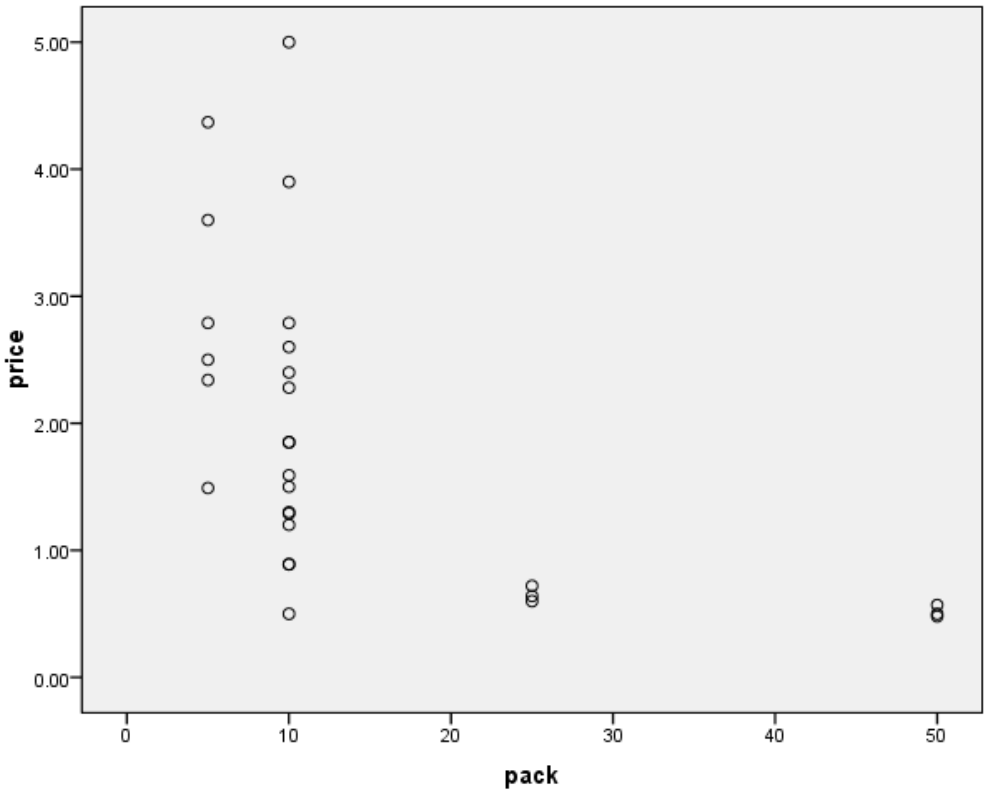
c.

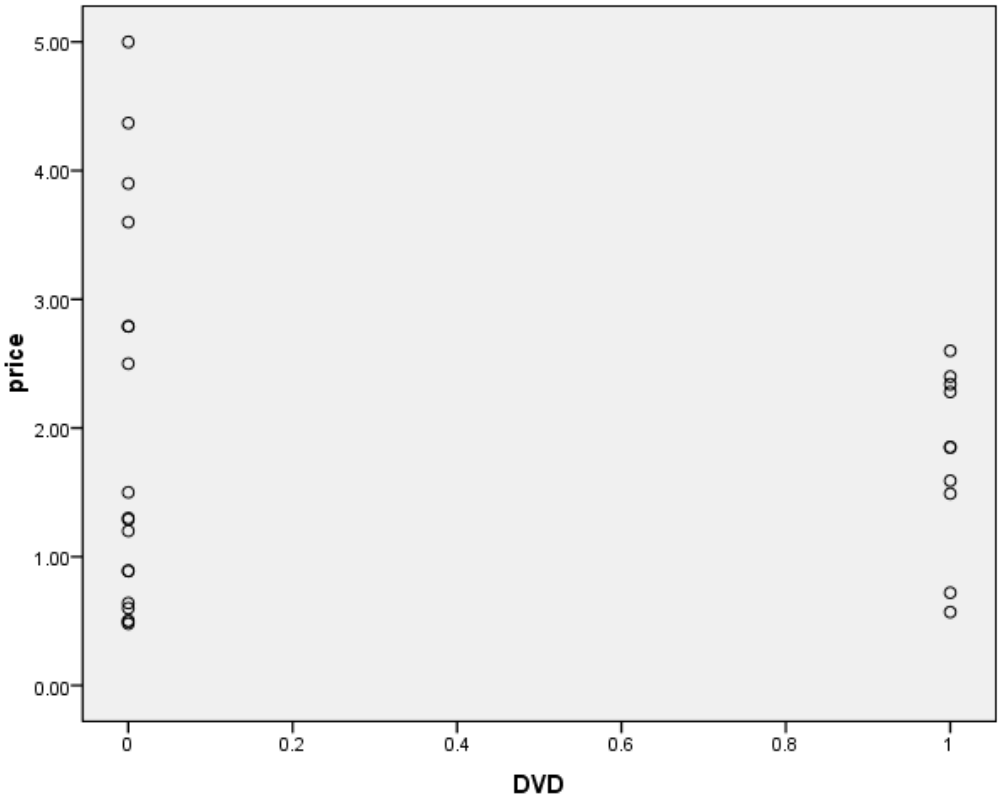
$\hat{LifeExp}$ = 69.648 - .036 People.per.TV $R^2 = 0.367$

$\hat{LifeExp}$ = 69.902 - .0007 People.per.Dr $R^2 = 0.444$

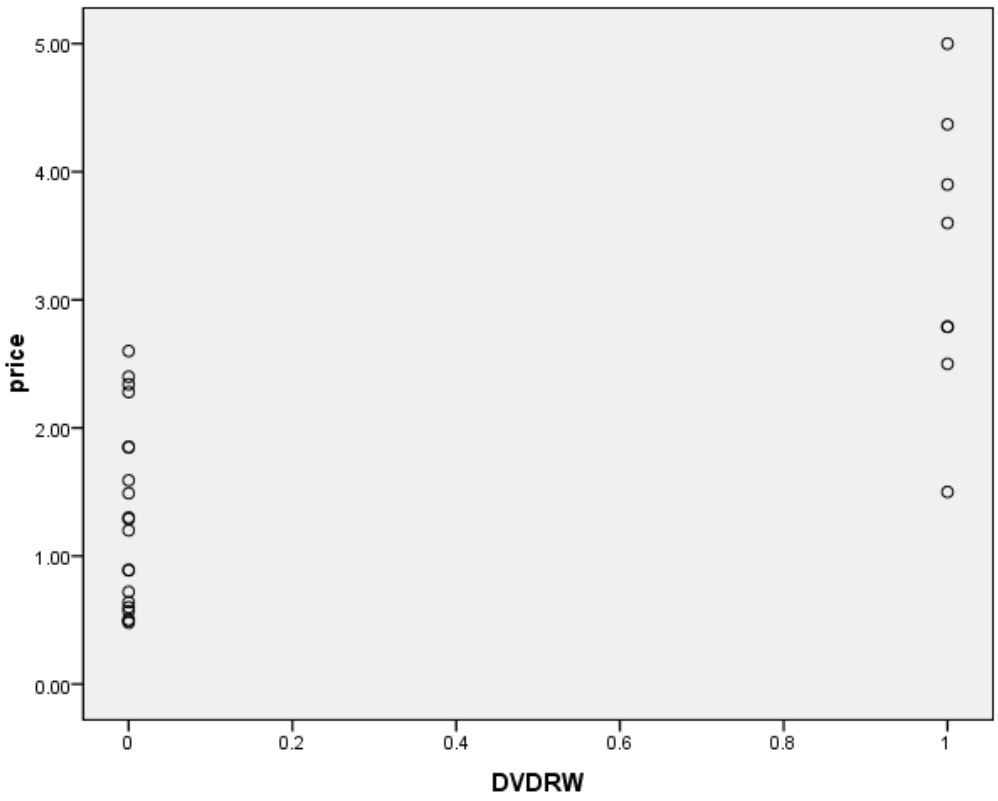
The latter equation looks marginally better from an R square point of view. However, neither model looks particularly impressive against their respective scattergrams shown earlier.

7. a. The scatter diagrams are shown below :





Note that for CD, the dummy variable DVD = 0 and DVDRW = 0; for DVD, DVD = 1 and DVDRW = 0



For dummy variable DVDRW, DVD = 0 and DVDRW = 1

b. Yes, the scattergrams suggests a regression model is likely to hold according to the scattergrams in a.

c. From the selective SPSS (stepwise regression) output below, a number of significant regression models can be found for the data:

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.743[a]	.553	.535	.84684	
2	.812[b]	.659	.631	.75439	
3	.845[c]	.715	.679	.70375	2.356

a. Predictors: (Constant), DVDRW

b. Predictors: (Constant), DVDRW, DVD

c. Predictors: (Constant), DVDRW, DVD, pack

d. Dependent Variable: price

Coefficients[a]

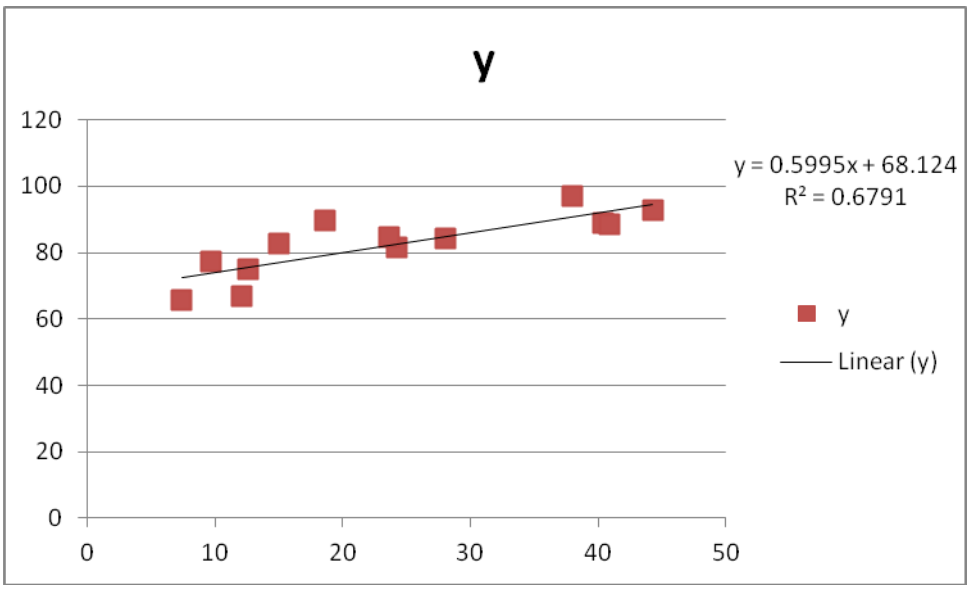
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1.299	.189		6.860	.000		
	DVDRW	2.007	.354	.743	5.666	.000	1.000	1.000
2	(Constant)	.829	.239		3.475	.002		
	DVDRW	2.477	.358	.917	6.923	.000	.778	1.286
	DVD	.940	.337	.369	2.786	.010	.778	1.286
3	(Constant)	1.327	.319		4.155	.000		
	DVDRW	2.157	.365	.799	5.912	.000	.651	1.536
	DVD	.786	.323	.309	2.436	.023	.740	1.351
	pack	-.024	.011	-.259	-2.174	.040	.837	1.195

a. Dependent Variable: price

Particularly, the 3 predictor model:

$\hat{price} = 1.327 + 2.157\ \text{DVDRW} + .786\ \text{DVD} - 0.24\ \text{pack}$

8. a. The scatter diagram is shown below:



It appears that a simple linear regression model may be appropriate from this plot and the fairly high coefficient of determination result of 67.9%.

b.

	Coefficients	Standard Error	t Stat	P-value
Intercept	57.713	7.375	7.825	0.000
x	1.645	0.678	2.425	0.036
x squared	-0.020	0.013	-1.565	0.149

ANOVA

	df	SS	MS	F	Significance F
Regression	2	801.672	400.836	14.399	0.001
Residual	10	278.381	27.838		
Total	12	1080.052			

Though the overall regression is significant according to the F value in the ANOVA table, the coefficient for X squared from the preceding output is not significant (pvalue = 0.149 > α = 0.05).

c. Regression Statistics

Multiple R	0.861
R Square	0.742
Adjusted R Square	0.718
Standard Error	0.063
Observations	13

	Coefficients	Standard Error	t Stat	P-value
Intercept	3.886	0.094	41.143	0.000
ln(x)	0.172	0.031	5.624	0.000

From the above the regression of log(y) on log(x) yields a significant fit with (pvalue < 0.05) for both regression coefficients.

d. The estimated regression in part (c) is preferred because having an Rsquare of 0.742, it explains a higher percentage of the variability in the dependent variable.

9. a. SSR = SST - SSE = 1030

MSR = 1030 MSE = 520/25 = 20.8 F = 1030/20.8 = 49.52

$F_{.05}$ = 4.24 (25 DF)

Since 49.52 > 4.24 we reject H_0: [illegible] and conclude that x_1 is significant.

b. $F = \dfrac{(520-100)/2}{100/23} = 48.3$

$F_{.05}$ = 3.42 (2 degrees of freedom numerator and 23 denominator)

Since 48.3 > 3.42 the addition of variables x_2 and x_3 is statistically significant

10. a. SSE = SST - SSR = 1805 - 1760 = 45

MSR = 1760/4 = 440 MSE =45/25 = 1.8

F = 440/1.8 = 244.44

$F_{.05}$ = 2.76 (4 degrees of freedom numerator and 25 denominator)

Since 244.44 > 2.76, variables x_1 and x_4 contribute significantly to the model

b. SSE(x_1, x_2, x_3, x_4) = 45

c. SSE(x_2, x_3) = 1805 - 1705 = 100

d. $F = \dfrac{(100-45)/2}{1.8} = 15.28$

$F_{.05}$ = 3.39 (2 numerator and 25 denominator DF)

Since 15.28 > 3.39 we conclude that x_1 and x_3 contribute significantly to the model.

11. a. The Minitab output is shown below:

```
The regression equation is
H = 127 + 2.21 W + 0.0297 M


Predictor     Coef  SE Coef      T      P     VIF
Constant   126.588    6.393  19.80  0.000
W           2.2115   0.6129   3.61  0.002  13.007
M          0.02974  0.01024   2.90  0.009  13.007


S = 13.3340   R-Sq = 96.3%   R-Sq(adj) = 95.9%


Analysis of Variance

Source          DF     SS     MS       F      P
Regression       2  96157  48079  270.41  0.000
Residual Error  21   3734    178
Total           23  99891


Source  DF  Seq SS
W        1   94659
M        1    1499


Unusual Observations

Obs     W       H     Fit  SE Fit  Residual  St Resid
  3  56.0  346.00  323.21    9.15     22.79      2.35RX
 22  13.0  209.00  181.12    4.66     27.88      2.23R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

Ominously the VIF's above are both greater than 10 indicating a potential multicollinearity problem. Corresponding correlation results below support this assessment.

b. The Minitab output is shown below:

```
The regression equation is
H = 121 + 2.33 W + 0.0354 M - 0.000114 M*W


Predictor        Coef    SE Coef      T      P     VIF
Constant       120.76      14.73   8.20  0.000
W              2.3308     0.6810   3.42  0.003  15.442
M             0.03538    0.01652   2.14  0.045  32.539
M*W        -0.0001141  0.0002587  -0.44  0.664  36.221


S = 13.5973   R-Sq = 96.3%   R-Sq(adj) = 95.7%

Analysis of Variance

Source          DF     SS     MS       F      P
Regression       3  96193  32064  173.43  0.000
Residual Error  20   3698    185
Total           23  99891


Source  DF  Seq SS
W        1   94659
M        1    1499
M*W      1      36


Unusual Observations

Obs     W       H     Fit  SE Fit  Residual  St Resid
  3  56.0  346.00  322.25    9.58     23.75      2.46R
 22  13.0  209.00  180.46    4.99     28.54      2.26R

R denotes an observation with a large standardized residual.
```

Again the VIF's are a problem here – in fact even more so.

c. Stepwise Regression: H versus M, W, M*W

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

Response is H on 3 predictors, with N = 24

Step            1      2
Constant    126.6  126.6

W            3.92   2.21
T-Value     19.95   3.61
P-Value     0.000  0.002

M                  0.030
T-Value             2.90
P-Value            0.009

S            15.4   13.3
R-Sq        94.76  96.26
R-Sq(adj)   94.52  95.91
Mallows Cp    8.3    2.2
```

Because the interaction term M*W has not been loaded into the model in the latter Stepwise Regression we deduce it does not contribute significantly to the model.

12 For the first model featuring the five predictors x1, x2, x3, x4 and x5, the significant F ratio from the ANOVA table (pvalue = 0.005 < α = 0.05) suggests that the overall model is a significant fit to the data. Yet none of the individual t tests associated with each of the regression slopes beforehand are significant except that for x4 (pvalue = 0.005 < α = 0.05). From the VIF's which are all close to 1, multicollinearity would not appear to be a problem for the data. The R Square of 66.3% indicates that the multiple regression model explains 66.3% of the variation in the response variable and this might be regarded as quite favourable. On the down side the model suffers from a single outlier according to MINITAB but for a sample of size 20 this does not seem unreasonable. The Durbin-Watson statistic of 1.72 but for a two-sided Durbin Watson test the relevant d_L and d_U values (based on n = 20 and k = 5 predictors) are 0.70 and 1.87. As $d_L < 1.72 < d_U$ we deduce the test is inconclusive.

The second model is a simple regression with just x4 as the predictor. The model is significant according to both the overall F test and the specific t tests associated with the regression slope for x4. As would be expected the R square value has dropped – in this case to 51.7%. Again there is an outlier (albeit for observation 12 now instead of observation 1 previously but with a corresponding standardized residual of -2.07 this does not look too serious.)

To check if the earlier five predictor model is a significant improvement on this one predictor model, a partial F test can be undertaken. The relevant calculation using equation (16,11) is as follows (note that p = 5, q = 1):

$$F = \frac{\dfrac{\text{SSE}(x_1, x_2, \ldots, x_q) - \text{SSE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{\text{SSE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}}$$

$$= \frac{\dfrac{23.002 - 16.032}{5 - 1}}{\dfrac{16.032}{14}}$$

$$= 1.52 < 3.11 = F_{.95}(4,14)$$

Hence we are not able to reject H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ at a 5% significance level and deduce that the five predictor model is not a significant improvement on the corresponding 1 predictor equivalent.

Note that because of the 'ln' transformation on y the relationship between y and x4 has an essentially exponential character despite the fact that we have effectively used a linear modelling formulation for the analysis.

13 a. For the first model, the F ratio from the ANOVA table (p value = 0.000 < α = 0.05) is highly significant which suggests the overall model offers a significant fit to the data. Ignoring the constant, t tests for the regression slopes corresponding to the humi, temp and HT variables are all significant (have a p value < 5%). The R Square (coefficient of determination) of 71.5% is favourable and suggests the multiple regression model explains the variation in the response quite well. There is one outlier but given the sample size is 44 this does not seem to be especially problematic. Observation 6 is categorized as influential and this should be investigated further. The Durbin-Watson statistic of 1.63 but for a two-sided Durbin Watson test the relevant d_L and d_U values (based on n = 44 and k = 3 predictors) are 1.29 and 1.58. As $1.58 = d_U < 1.63 < 4\text{-}d_U = 2.42$ we deduce no evidence of first order serial correlation of residuals is present.

b. Again according to the F ratio details provided, the second model too is significant. However from corresponding t tests only the slopes for humi and HH can be considered significantly different from zero. In this case however the R square is an impressive 79.7%. There are two outliers and one (different) influential observation with this model. The outliers do not look serious but as before the influential observation should be investigated. The Durbin-Watson statistic is 1.78. For n = 44 and k = 5, d_L and d_U are 1.29 and 1.58 respectively. As $1.58 = d_U < 1.78 < 4 - d_U = 2.42$ we deduce there is no problem with residuals suffering from first order serial correlation.

To check if the earlier five predictor model is a significant improvement on this one predictor model, a partial F test can be undertaken. The relevant calculation using equation (16,11) is as follows :

$$F = \frac{\dfrac{SSE(x_1, x_2, \ldots, x_q) - SSE(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{SSE(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}}$$

$$= \frac{\dfrac{0.14166 - 0.100887}{5-3}}{\dfrac{0.100887}{38}}$$

$$= 26.68 > 3.25 = F_{.95}(2,38)$$

Hence we reject H_0: $\beta_4 = \beta_5 = 0$ at the 5% significance level and deduce that the five predictor model is a significant improvement on the corresponding 3 predictor alternative.

14. Let Health = 1 if health-drugs
Health = 0 if energy-international or other

```
The regression equation is
P/E = 10.8 + 0.430 Sales% + 10.6 Health

Predictor       Coef     SE Coef          T        P
Constant      10.817       3.143       3.44    0.004
Sales%        0.4297      0.1813       2.37    0.034
Health        10.600       2.750       3.85    0.002

S = 5.012       R-Sq = 55.4%     R-Sq(adj) = 48.5%

Analysis of Variance

Source          DF          SS          MS         F        P
Regression       2      405.69      202.85      8.07    0.005
Residual Error  13      326.59       25.12
Total           15      732.28

Source       DF      Seq SS
Sales%        1       32.36
Health        1      373.33
```

15. From the correlation matrix provided, the sales response is significantly correlated with all of the three predictors listed. However the *attract* variable is also significantly correlated with that for *airplay* suggesting potential problems of multicollinearity if both variables are fitted together in a linear model

    Interestingly, when a Stepwise regression analysis is undertaken, all three predictors are successfully entered into the model and as the full regression details that appear immediately afterwards indicates – according to the VIF figures (< 10) at least – multicollinearity is not actually an issue.

    For the latter model, all predictors are significant at the 5% level (p-value = .000 in each case) and the R Square for the model is 65.95%. Thus the three predictor model explains some 66% of the variation in sales.

    Reassuringly, the three predictor model also seems to outperform the alternatives shown in the Best Subsets Regression summary.

    Assessing the significance of the Durbin-Watson statistic is not straightforward for a sample size of 200 using the tables provided in the text. However from http://www-leland.stanford.edu/~clint/bench/dw05b.htm it can be shown that at a = 5%, $d_L = 1.74833$ and as $d_L < 1.95 < 4 - d_L$ the test would be at best inconclusive.

    The main problem with the model is the presence of many significant outliers – but more importantly the existence of a number of influential observations.

16 a. From the best subsets regression summary, the 5 predictor model with an R Square of 86.2% is almost as good on all measures as the full 6 predictor model represented by the bottom line of the table. The same five predictor model is described in detail after the correlation matrix and can be seen from the ANOVA F statistic to be significant overall. Corresponding t statistics are also significant (though technically the p value (of 0.054) associated with the regression slope for the pop variable is just slightly above the test size of 5%).

b. Clearly multicollinearity is a problem here. This is informed by significant correlations between predictor variables e.g. pr and con. Also the sign of the coefficient of the con predictor in the detailed regression output is opposite to that of the corresponding correlation between con and ao.

c. Yes. In these circumstances the two predictor model from Stepwise now looks technically more appealing.

17 a. In the regression summary at the beginning, only x1 with an associated p value of 0.003 can be considered as a significant predictor of y though the regression model overall from the ANOVA F result can be judged significant. Note that x2 is close to significance because of the p value of 0.076 (being only slightly over the test size of 0.05). The stepwise procedure reveals that there are two significant predictors of y, firstly in terms of x1 and after x1, x2. Corresponding best subsets output shows that the two predictor model with an R Square of 39.8 is on all other diagnostics here better than the full three predictor model.

b. To evaluate the two (x1, x2) predictor model ,properly a full regression analysis needs to be conducted.

c. Only when b. is done and relevant tests performed, can the two predictor model be adequately assessed. Nevertheless this would appear to offer the most promise initially.

18 a. Because of the significant correlations between predictors in the summary table at the beginning of the output the possibility of multicollinearity looms large for any subsequent regression modelling. From the Best Subsets table the three predictor model with an R Square of 86.2% compares well with the four predictor baseline model. This is essentially the same model picked out by the Stepwise (backward elimination) analysis afterwards. According to this, all predictors accept Support look as if they could be usefully retained in for regression modelling analysis. Following on, the detailed output for a three predictor regression model shows that Retired, Unemployment and Total Staff are all significant predictors of the response variable, Crimes. Because of relatively low VIF values the model does not seem to suffer from multicollinearity problems

mentioned earlier. With an R Square of 84.6% it compares with the three predictor model described earlier fairly well. The Durbin Watson statistic of 2.22 is not problematic since if n = 26 and k = 3 then $d_L = 1.04$, $d_U = 1.54$. As $1.54 = d_U < 2.22 < 4\text{-}d_U = 2.46$ we deduce there is no evidence of first order serial correlation of residuals present.

b. For the relevant two predictor model the root mean square error s = 89.041. Corresponding the error sums of squares would be $23 * 89.041^2 = 182350.9$. By comparing this model with the last 3 predictor model in a. a partial F test statistic from equation (16.11) can be calculated as follows:

$$F = \frac{\dfrac{SSE(x_1, x_2) - SSE(x_1, x_2, x_3)}{p - q}}{\dfrac{SSE(x_1, x_2, x_3)}{n-p-1}}$$

$$= \frac{\dfrac{182350.9-150468}{3-2}}{\dfrac{150468}{22}}$$

$$= 26.66 > 4.30 = F_{.95}(1,22)$$

Hence we reject $H_0: \beta_3 = 0$ at the 5% significance level and deduce that the three predictor model is a significant improvement on the corresponding two predictor one.

c. Following on from b. the three predictor model based on Retired, Unemployment and Total Staff would be preferred.

Supplementary Exercises:

19. A study investigated the relationship between audit delay (Delay), the length of time from a company's fiscal year-end to the date of the auditor's report, and variables that describe the client and the auditor. Some of the independent variables that were included in this study follow.

Industry	A dummy variable coded 1 if the firm was an industrial company or 0 if the firm was a bank, savings and loan, or insurance company.
Public	A dummy variable coded 1 if the company was traded on an organized exchange or over the counter; otherwise coded 0.
Quality	A measure of overall quality of internal controls, as judged by the auditor, a five-point scale ranging from "virtually none" (1) to "excellent" (5)
Finished	A measure ranging from 1 to 4, as judged by the auditor, where 1 indicates "all work performed subsequent to year-end" and 4 indicates "most work performed prior to year-end."

File “Audit”

A sample of 40 companies provided the following data.

Delay	Industry	Public	Quality	Finished
62	0	0	3	1
45	0	1	3	3
54	0	0	2	2
71	0	1	1	2
91	0	0	1	1
62	0	0	4	4
61	0	0	3	2
69	0	1	5	2
80	0	0	1	1
52	0	0	5	3
47	0	0	3	2
65	0	1	2	3
60	0	0	1	3
81	1	0	1	2
73	1	0	2	2
89	1	0	2	1
71	1	0	5	4
76	1	0	2	2
68	1	0	1	2
68	1	0	5	2
86	1	0	2	2
76	1	1	3	1
67	1	0	2	3
57	1	0	4	2
55	1	1	3	2
54	1	0	5	2
69	1	0	3	3
82	1	0	5	1
94	1	0	1	1
74	1	1	5	2
75	1	1	4	3
69	1	0	2	2
71	1	0	4	4
79	1	0	5	2
90	1	0	1	4
91	1	0	4	1
92	1	0	1	4
46	1	1	4	3
72	1	0	5	2
85	1	0	5	1

a. Develop the estimated regression equation using all of the independent variables

b. How well does the estimated regression equation developed in part (a) represent the data?

c. Develop a scatter diagram showing Delay as a function of Finished. What does this scatter diagram indicate about the relationship between Delay and Finished.

d. On the basis of your observations about the relationship between Delay and Finished develop an alternative estimated regression equation to the one developed in (a) to explain as much of the variability in Delay as possible.

20. Annual data published by Conrad (1989) over a 21 year period features the following variables:

Y = consumption of tobacco goods
X_1 = real personal disposable income per capita
X_2 = real price of tobacco goods
HSA, HSB, HSC = health scare dummy variables
(where HSA = 1 for years 8 and 9, 0 otherwise
HSB = 1 for years 10 and 11, 0 otherwise
HSC = 1 for years 17 and 18, 0 otherwise)

Corresponding MINITAB modeling output is as follows:

```
The regression equation is
lnY = 5.63 + 0.0478 HSA + 0.0149 HSB - 0.0535 HSC - 0.0126 lnX1 +
0.001 lnX2

Predictor      Coef  SE Coef      T      P
Constant     5.6307   0.2093  26.90  0.000
HSA         0.04785  0.02932   1.63  0.124
HSB         0.01485  0.03131   0.47  0.642
HSC        -0.05352  0.03057  -1.75  0.100
lnX1       -0.01261  0.03274  -0.39  0.706
lnX2         0.0009   0.1133   0.01  0.994

S = 0.0382038   R-Sq = 34.8%   R-Sq(adj) = 13.1%

Analysis of Variance

Source          DF        SS        MS     F      P
Regression       5  0.011700  0.002340  1.60  0.219
Residual Error  15  0.021893  0.001460
Total           20  0.033593


Source  DF    Seq SS
HSA      1  0.004986
HSB      1  0.000710
HSC      1  0.005775
lnX1     1  0.000229
lnX2     1  0.000000

Unusual Observations

Obs   HSA      lnY      Fit   SE Fit  Residual  St Resid
7    0.00  5.62654  5.55090  0.01145   0.07564      2.08R

R denotes an observation with a large standardized residual.
```

Stepwise Regression: lnY versus HSA, HSB, HSC, lnX1, lnX2

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is lnY on 5 predictors, with N = 21

Step	1	2
Constant	5.556	5.551
HSC	-0.064	-0.059
T-Value	-2.30	-2.23
P-Value	0.033	0.039
HSA		0.046
T-Value		1.75
P-Value		0.096
S	0.0372	0.0353
R-Sq	21.81	33.23
R-Sq(adj)	17.70	25.82
Mallows C-p	1.0	0.4

Best Subsets Regression: lnY versus HSA, HSB, HSC, lnX1, lnX2

Response is lnY

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	HSA	HSB	HSC	lnX1	lnX2
1	21.8	17.7	1.0	0.037181			X		
1	14.8	10.4	2.6	0.038802		X			
2	33.2	25.8	0.4	0.035299	X		X		
2	22.8	14.2	2.8	0.037965			X		X
3	34.1	22.5	2.2	0.036073	X	X	X		
3	33.7	22.0	2.3	0.036200	X		X	X	
4	34.8	18.5	4.0	0.036991	X	X	X	X	
4	34.2	17.7	4.1	0.037173	X	X	X		X
5	34.8	13.1	6.0	0.038204	X	X	X	X	X

a. Comment on the effectiveness of the various models here carrying out any statistical tests or additional analysis you think appropriate.

b. How would you advise the Tobacco Research Council who sourced the data?

21. Refer to the data in exercise 19. Consider a model in which only Industry is used to predict Delay. At a 0.01 level of significance, test for any positive autocorrelation in the data.

22. Refer to the data in exercise 19.

a. Develop an estimated regression equation that can be used to predict Delay by using Industry and Quality.

b. Plot the residuals obtained from the estimated regression equation developed in part (a) as a function of the order in which the data are presented. Does any autocorrelation appear to be present in the data? Explain.

c. At the 0.05 level of significance, test for any positive autocorrelation in the data.

23. A regression analysis of heart disease by country (Cooper & Weekes, 1983) is based on the following variables:

sug	sugar consumption
tdp	total dairy products consumption
agemp	percentage employment in agriculture, fishing and forestry
ihdmr	ischaemic heart disease mortality rate (**RESPONSE variable**)

Relevant MINITAB output for two contrasting models is given below:

MODEL 1

Regression Analysis

The regression equation is
ihdmr = 1.6 + 1.41 sug + 0.178 tdp - 2.12 agemp

Predictor	Coef	StDev	T	P
Constant	1.62	86.16	0.02	0.985
sug	1.4070	0.9752	1.44	0.167
tdp	0.1775	0.1214	1.46	0.162
agemp	-2.124	2.137	-0.99	0.334

S = 58.38 R-Sq = 68.0% R-Sq(adj) = 62.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	123022	41007	12.03	0.000
Residual Error	17	57950	3409		
Total	20	180972			

Source	DF	Seq SS
sug	1	114848
tdp	1	4806
agemp	1	3369

Unusual Observations

Obs	sug	ihdmr	Fit	SE Fit	Residual	St Resid
6	115	240.9	288.2	46.3	-47.3	-1.33 X
19	117	105.3	227.2	15.5	-121.9	-2.17R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

MODEL 2

Regression Analysis

The regression equation is
ihdmr = - 84.4 + 2.73 sug

Predictor	Coef	StDev	T	P
Constant	-84.35	50.10	-1.68	0.109
sug	2.7255	0.4744	5.74	0.000

S = 58.99 R-Sq = 63.5% R-Sq(adj) = 61.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	114848	114848	33.00	0.000
Residual Error	19	66124	3480		
Total	20	180972			

Unusual Observations

Obs	sug	ihdmr	Fit	SE Fit	Residual	St Resid
19	117	105.3	234.3	14.7	-129.0	-2.26R

R denotes an observation with a large standardized residual

Explain this computer output, carrying out any additional tests you think necessary or appropriate. Is the first model significantly better than the second? Which model do you prefer and why?

24. A regression analysis of UK imports (Barrow, 2001) is based on the following variables:

lnimport	natural log of UK imports in real prices (£bn) **(RESPONSE variable)**
lngdp	natural log of UK Gross Domestic Product in real value terms (£bn)
lnprice	natural logarithm of the unit value index of imports
laglnimport	lagged value of lnimport variable

Relevant MINITAB output is given below:

Regression Analysis: lnimport versus lngdp, lnprice

The regression equation is
lnimport = - 4.08 + 1.76 lngdp - 0.292 lnprice

Predictor	Coef	SE Coef	T	P	VIF
Constant	-4.078	1.514	-2.69	0.015	
lngdp	1.7625	0.1644	10.72	0.000	4.9
lnprice	-0.2917	0.1280	-2.28	0.035	4.9

S = 0.04093 R-Sq = 97.8% R-Sq(adj) = 97.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1.35348	0.67674	403.88	0.000
Residual Error	18	0.03016	0.00168		
Total	20	1.38364			

Source	DF	Seq SS
lngdp	1	1.34478
lnprice	1	0.00870

Unusual Observations

Obs	lngdp	lnimport	Fit	SE Fit	Residual	St Resid
2	5.55	4.32281	4.22598	0.01726	0.09683	2.61R

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 1.09

Correlations: lnimport, lngdp, lnprice, laglnimport

	lnimport	lngdp	lnprice
lngdp	0.986		
	0.000		
lnprice	-0.916	-0.893	
	0.000	0.000	
laglnimp	0.977	0.954	-0.939
	0.000	0.000	0.000

Cell Contents: Pearson correlation
P-Value

Stepwise Regression: lnimport versus lngdp, lnprice, laglnimport

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is lnimport on 3 predictors, with N =   20
N(cases with missing observations) =   1 N(all cases) =   21

Step	1	2
Constant	-7.605	-4.881
lngdp	2.134	1.332
T-Value	26.48	7.21
P-Value	0.000	0.000
laglnimp		0.408
T-Value		4.55
P-Value		0.000
S	0.0430	0.0297
R-Sq	97.50	98.87
R-Sq(adj)	97.36	98.74
C-p	19.6	2.1

a. Explain this computer output, carrying out any additional tests you think necessary or appropriate.

b. Which of the various models shown do you prefer and why?

25. Tony runs a used car business. He would like to predict monthly sales. Tony believes that sales, y (in €'0,000s) is directly related to the number of sales-people employed (x_1) and the average number of cars on the lot for sale (x_2). The following data were collected over a period of 10 months:

File "CarSales"

y	x_1	x_2
5.8	4	20
8.1	4	25
7.5	5	15
13.3	8	30
11.4	7	25
15.0	9	35
7.0	3	17
8.3	5	20
5.1	2	18
6.8	4	23

a. Calculate the correlations between y, x_1 and x_2. What do these suggest?

b. Estimate the regression equation of y on x_1 and x_2 and calculate the corresponding VIF's for the independent variables.

c. What do you infer from (b)?

d. Plot the residuals by $\hat{y}$, x_1 and x_2 and comment on the validity of the theoretical assumptions for regression in this case.

26. For the data in Exercise 25, use MINITAB to carry out a
Stepwise and
best subsets analysis.

a. Interpret the resulting computer outputs.
b. Which of the various models covered by these outputs do you most prefer and why?

27. File "Proposals"

The CEO of a computer firm is interested in funding research proposals by graduate students who wish to perform experiments in the firm's advanced technology laboratory during the summer months. The CEO receives 18 proposals and sends these proposals to the director of the laboratory for evaluation. The director rates the proposals on two different criteria and gives a score between zero and ten for each criterion, with 10 representing the best score possible. (The variables x_1 and x_2 represent these two scores. The variable y (in €000s) is the level of funding that the CEO grants for the proposal.) The collected data are given below:

y	**x_1**	**x_2**
9.5	8.7	9.2
7.3	8.1	8.0
6.5	7.4	7.7
8.4	8.4	8.6
8.0	8.3	8.0
6.1	7.0	7.3
8.5	8.6	8.8
7.2	8.3	7.8
5.8	6.7	7.0
6.3	7.3	7.5
9.0	8.6	9.0
6.4	7.7	7.5
7.0	7.9	7.9
7.4	8.2	8.0
8.3	8.5	8.4
8.2	8.6	7.9
5.3	6.6	6.9
6.7	7.8	7.5

The director tries to work out what the CEO will grant, given how he scores a proposal.

a. Find a 90% confidence interval for the mean value of y at $x_1 = 8.0$ and $x_2 = 7.8$.

b. Find a 90% confidence interval for the value of Y at $x_1 = 8.0$ and $x_2 = 7.8$.

c. Interpret each of these confidence intervals: what is the difference between them?

28. For the data in Exercise 27, use MINITAB to carry out a

Stepwise and

best subsets analysis.

a. Interpret the resulting computer outputs.

b. Which of the various models covered by these outputs do you most prefer and why?

29. Consider the following dataset for 12 growth-orientated companies. Y represents the growth rate of a company for the current year. X_1 represents the growth rate of the company for the previous year. X_2 represents the percentage of the market that does not use the company's product or a similar product, and X_3 represents the current growth rate for the industry sector to which the company belongs. (All values are percentages.)

File "Growth"

Y	**X_1**	**X_2**	**X_3**
20	10	30	2.8
30	15	60	3.4
24	12	35	5.6
36	42	38	2.8
18	15	25	10.1
47	45	40	6.2
33	30	40	2.8
35	32	32	7.9
27	19	32	3.4
28	24	31	10.1
20	24	20	7.9
32	20	50	2.8

a. Using MINITAB, derive the sample correlations for the variables and estimate the regression equation of Y on X_1, X_2 and X_3. Test the significance of X_1, X_2 and X_3 in the model. What do you deduce?

b. Perform a stepwise analysis of the data using the backward elimination procedure. Comment on the results obtained and compare these with the outputs from (a). Are they consistent?

30. Chatterjee and Price (1977) present attitude data for clerical staff towards their supervisors within a large commercial organization. Details of the variables involved in the study and of the predictive model obtained using the MINITAB package, are as follows:

y	: Overall rating of job being done by supervisor
cmplain	: Handles staff complaints
prvileg	: Does not allow special privileges
learn	: Opportunity to learn new things
rises	: Rises based on performances
critcal	: Too critical of poor performances
advance	: Rate of advancing to better jobs

```
y - 10.8  +  0.613 cmplain - 0.073 prvileg
            (0.161)         (0.136)
         vif  2.7            1.6

                   +  0.320 learn +  0. 082 rises
                     (0.169)         (0.222)
                vif   2.3             3.1

                   +  0.038 critcal - 0.217 advance
                     (0.147)         (0.178)
                 vif   1.2            2.0
```

Note that the figures in brackets here are the standard errors of associated estimated regression coefficients. Also that the total (corrected) sum of squares on y = 4296.97 and the sample size = 30

The sample correlations for the data are as follows:

	y	cmplain	prvileg	learn	rises	critcal
cmplain	0.825					
prvileg	0.426	0.558				
learn	0.624	0.597	0.493			
rises	0.590	0.669	0.445	0.640		
critcal	0.156	0.188	0.147	0.116	0.377	
advance	0.155	0.225	0.343	0.532	0.574	0.283

Results from running the MINITAB's best subsets procedure for the data are given below:

Best Subsets Regression of y

VARS	R-sq	Adj. R-sq	C-p	s	c m p 1 a i n	p r v i e g	1 e a r n	r i s e	c r i t a l	a d v a n c e
1	68.1	67.0	1.4	6.9933	X					
1	38.9	36.7	26.6	9.6835			X			
2	70.8	68.6	1.1	6.8168	X		X			
2	68.4	66.0	3.2	7.0927	X			X		
3	72.6	69.4	1.6	6.7343	X		X			X
3	71.5	68.2	2.5	6.8630	X	X	X			
4	72.9	68.6	3.3	6.8206	X	X	X			X
4	72.9	68.5	3.4	6.8310	X		X	X		X
5	73.2	67.6	5.1	6.9294	X	X	X	X		X
5	73.1	67.5	5.1	6.9396	X	X	X		X	X
6	73.3	66.3	7.0	7.0680	X	X	X	X	X	X

a. Given the evidence provided here and making any additional calculations and / or statistical tests you think necessary, how would you interpret this information?

b. What is your overall view of the model's effectiveness?

31. Pre-employment tests are widely used in many large corporations as an approach for estimating likely job performance. In a published study, separate regression analyses (see MODEL 2 below) were conducted for white and minority sections of a recruitment sample. The results, given, contrast with those from a pooled analysis of the entire sample (MODEL 1):

jperf : Job Performance
test : Pro-employment test
race : 1 if a minority applicant, 0 if a white applicant
racetest : race X test

MODEL 1

```
jperf =  1.03  +  2.36  test
           (0.868)   (0.538)
```

ANOVA

SOURCE	df	SS	MS	F
Regression	1	48.723	48.723	19.25
Error	18	45.568	2.532	
Itotal	19	94.291		

Note that figures in brackets above denote the standard errors of corresponding regression slope estimates, Corresponding to the 'test' variable taking the value 2.5 we have predicted jperf value, confidence and prediction intervals as follows:-

Fit	95% C.I.	95% P.I.
6.936	(5.554, 8.319)	(3.3I8, 10.554)

MODEL 2

```
jperf =  2.01  - 1.91  race +  1.31  test +  2.00  racetest
               (1.540)        (0.670)        (0.954)

Error SS  = 31.655
```

For this alternative formulation the predicted value of jperf, corresponding to the value of 2.5 for 'test', confidence and prediction intervals shown separately for white and minority employees are as follows:-

	Fit	95% C.I.	95% P.I.
Minority	8.374	(6.681,10.068)	(4.945, 11.804)
White	5.294	(3.491, 7.097)	(1.809, 8.779)

a. Interpret these results, carrying out any additional calculations, tests etc. you think necessary.

b. Is MODEL 2 significantly better than MODEL 1? Depending on your answer here, what would you say this signifies in terns of the two groups?

32. Data relating to import activity in the French economy have been analysed by Malinvaud (1966). Details of a multiple regression model developed from these data appear below:

import : Imports
doprod : Domestic Production
stock : Stock Formation
consum : Domestic Consumption

The sample correlations for these data are as follows:

	import	doprod	stock
import			
deprod	0.984		
stock	0.266	0.215	
consum	0.985	0.999	0.214

```
import  = - 19.7  +  0.032 doprod +  0.414 stock
                    (0.187)         (0.322)
                vif  469.7             1.0

                   +  0.243 consum
                      (0.285)
                vif   469.4
```

(Note the figures in brackets here are the standard errors of the corresponding regression slope estimates.)

Estimated Error Variance, s^2	= 5.10
Sample Size	= 18
R Square	= 97.3%
Durbin-Watson statistic	= 0.24

a. Interpret these results, carrying out any additional calculations, tests etc. you think necessary.

The VIF values here reveal major problems with multicollinearity. Thus estimated coefficients in the regression model as well as corresponding t tests are likely to be very dubious.

b. What is your overall view of the model as a technology for predicting French Imports? What improvements (if any) are necessary, in your opinion, before implementation of the model is finally considered?

33. A regression analysis of data from a cloud-seeding experiment (Woodley et al; 1977) yields the following results:-

MODEL 1

$\hat{y}$ = 4.654 + 1.013 x1 - 0.032 x2 - 0.911 x3

(3.337) (1.203) (0.029) (0.751)

+ 0.006 x4 + 1.844 x5 + 2.168 x6

(.115) (2.758) (1.579)

where y = amount of rain (cubic metres x 10^7) that fell in target area for a 6 hour period on each day seeding was suitable.

x1 =1, seeding or 0, no seeding

x2 = number of days since the experiment began

x3 = seeding suitability factor

x4 = per cent cloud cover

x5 = total rainfall on target area one hour before seeding

x6 = 1, moving radar echo, or 2, a stationary radar

(Note the bracketed figures are the standard errors of the estimated regression coefficients)

R^2 = 0.385, Durbin-Watson statistic = 1.448

s^2 = 8.044, Sample size = 24

Correlations

	x1	x2	x3	x4	x5	x6	y
x1	1	.03	.177	.062	-.030	-.103	.076
x2		1	.451	-.350	-.269	-.218	-.496
x3			1	-.151	.040	-.186	-.408
x4				1	.648	-.019	.270
x5					1	-.257	.174
x6						1	.332
y							1

A second analysis of the data yields the alternative model:

MODEL 2

$\hat{y}$ = - 3.499 + 16.245x1 - 0.045x2 + 0.420x3 + 0.388x4

(4.063) (5.522) (0.025) (.845) (.218)

+ 4•108x5+ 3.153x6 - 3.200x1x3- 0.486x1x4

(3.601) (1.933) (1.267) (0.241)

- 2.557x1x5 -0.526x1x6

(4.481) (2.643)

$R^2 = 0.72$

$s^2 = 4.86$

a. Is the second model significantly more effective than the first?

b. From a broad comparison of the two models, which would you choose for forecasting rainfall in the target area? Give your reasons, for and against.

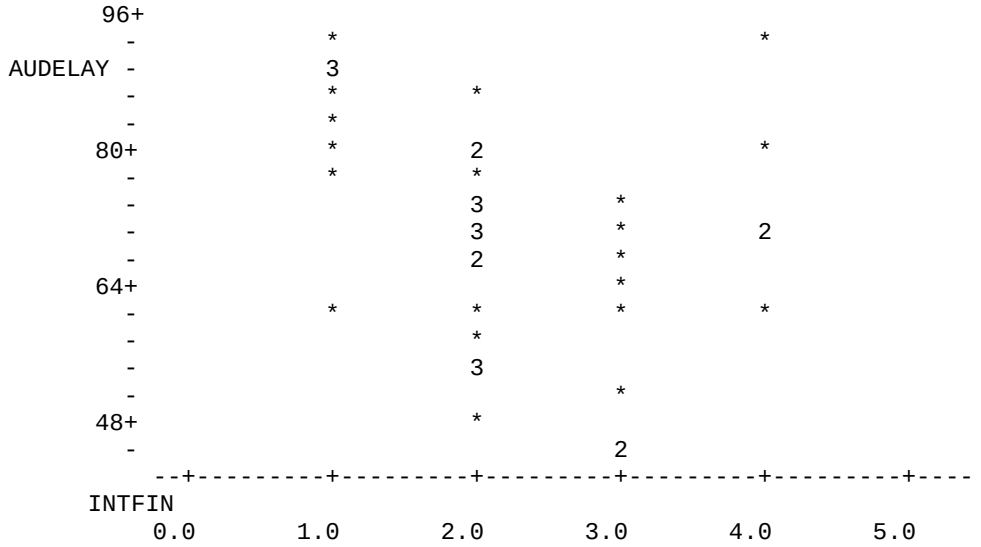
Supplementary Exercises Solutions:

19. a. The Minitab output is shown below:

```
The regression equation is
AUDELAY = 80.4 + 11.9 INDUS - 4.82 PUBLIC - 2.62 ICQUAL - 4.07
INTFIN

Predictor       Coef     SE Coef          T        p
Constant      80.429       5.916      13.60    0.000
INDUS         11.944       3.798       3.15    0.003
PUBLIC        -4.816       4.229      -1.14    0.263
ICQUAL        -2.624       1.184      -2.22    0.033
INTFIN        -4.073       1.851      -2.20    0.035

S = 10.92       R-sq = 38.3%     R-sq(adj) = 31.2%

Analysis of Variance

SOURCE          DF          SS          MS         F        p
Regression       4      2587.7       646.9      5.42    0.002
Residual Error  35      4176.3       119.3
Total           39      6764.0
```

b. The low value of the adjusted coefficient of determination (31.2%) does not indicate a good fit.

c. The scatter diagram is shown below:



The scatter diagram suggests a curvilinear relationship between these two variables.

d. The output from the stepwise procedure is shown below, where INTFINSQ is the square of INTFIN.

```
Response is AUDELAY  on  5 predictors, with N =   40

    Step          1        2
Constant      112.4    112.8

INDUS          11.5     11.6
T-Value        3.67     3.80
P-Value       0.001    0.001

PUBLIC         -1.0
T-Value       -0.29
P-Value       0.775

ICQUAL        -2.45    -2.49
T-Value       -2.51    -2.60
P-Value       0.017    0.014

INTFIN        -36.0    -36.6
T-Value       -4.61    -4.91
P-Value       0.000    0.000

INTFINSQ        6.5      6.6
T-Value        4.17     4.44
P-Value       0.000    0.000

S              9.01     8.90
R-Sq          59.15    59.05
R-Sq(adj)     53.14    54.37
C-p             6.0      4.1
```

20. a. The results from the Stepwise procedure indicate that lnY can be significantly explained in terms of the dummy variables HSC and HSA. At the same time, the R^2 and **adj** R^2 values for this model (33.23%, 25.82% respectively) are not particularly high. The same model features in the Best Subsets output (it corresponds with the first of the two predictor alternatives of models in the list) and technically appears the have the edge on its eight competitors. However, one practical problem with the HSA variable is that the sign of the estimated regression coefficient is positive, suggesting that the health scare in year 8 actually resulted in a growth rather than a decline in tobacco consumption.

b. From the various comments in (a) the linear formulation adopted for analysing the data does not seem to have been helpful or productive. The absence of lnX1 or lnX2 as predictors in any of the models is a particular indictment so much so that one wonders why this approach was ever investigated in the first place.

21. The computer output is shown below:

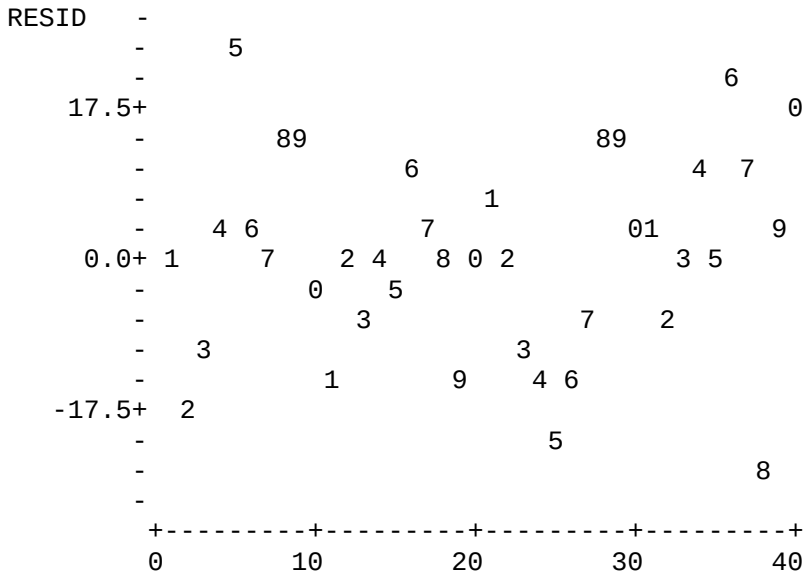
```
The regression equation is
AUDELAY = 63.0 + 11.1 INDUS

Predictor       Coef     SE Coef          T        p
Constant      63.000       3.393      18.57    0.000
INDUS         11.074       4.130       2.68    0.011

S = 12.23       R-sq = 15.9%     R-sq(adj) = 13.7%

Analysis of Variance

SOURCE          DF          SS          MS         F
p
Regression       1      1076.1      1076.1      7.19
0.011
Residual Error  38      5687.9       149.7
Total           39      6764.0

Unusual Observations
Obs.   INDUS   AUDELAY       Fit Stdev.Fit  Residual
St.Resid
  5     0.00     91.00     63.00      3.39     28.00
2.38R
 38     1.00     46.00     74.07      2.35    -28.07     -
2.34R

Durbin-Watson statistic = 1.55
```

At the .05 level of significance, $d_L = 1.44$ and $d_U = 1.54$. Since $d = 1.55 > d_U$, there is no significant positive autocorrelation.

22. a. The Minitab output is shown below:

```
The regression equation is
AUDELAY = 70.6 + 12.7 INDUS - 2.92 ICQUAL

Predictor       Coef     SE Coef          T        p
Constant      70.634       4.558      15.50    0.000
INDUS         12.737       3.966       3.21    0.003
ICQUAL        -2.919       1.238      -2.36    0.024

S = 11.56       R-sq = 26.9%     R-sq(adj) = 22.9%

Analysis of Variance

SOURCE          DF          SS          MS         F        p
Regression       2      1818.6       909.3      6.80    0.003
Residual Error  37      4945.4       133.7
Total           39      6764.0

SOURCE       DF      SEQ SS
INDUS         1      1076.1
ICQUAL        1       742.4

Unusual Observations
Obs.   INDUS   AUDELAY       Fit Stdev.Fit  Residual   St.Resid
  5     0.00     91.00     67.71      3.78     23.29      2.13R
 38     1.00     46.00     71.70      2.44    -25.70     -2.27R

R denotes an obs. with a large st. resid.

Durbin-Watson statistic = 1.43
```

b. The residual plot as a function of the order in which the data are presented is shown below:



There is no obvious pattern in the data indicative of positive autocorrelation.

c. At the .05 level of significance, $d_L = 1.44$ and $d_U = 1.54$. Since $d = 1.43 > d_U$, there is no significant positive autocorrelation.

23.

MODEL 1

This is a particularly flawed model. None of the predictors here are significant according to their individual p values yet the F statistic has a pvalue of $0.000 < \alpha = 0.05$ indicating that the model, overall, is significant. Because of this there is a strong suspicion that multicollinearity is present. The R^2 of 68% for the model is relatively good and there is a single outlying residual and as well as an influential observation. The outlier does not look serious because of its standardized residual value but observation 6's influence needs to be carefully checked out.

MODEL 2

This is a much simpler model which not surprisingly overcomes many of the problems of MODEL 1. The single predictor used in the model is significant. Observation 6 is no longer influential. Observation 19 is still associated with an outlying residual but this is hardly any worse than before.

Is MODEL 1 despite its various difficulties an improvement on MODEL 2 though?

To find out we note the following Error SS details:

Model	DF	Error SS
1	17	57950
2	19	66124

Based on formula (16.11) the relevant test statistic is:

$$F = \frac{(66124 - 57950)/(19-17)}{57950/17} = 1.20$$

Under the hypothesis H_0: $\beta_2 = \beta_3 = 0$ F has an F distribution on 2 and 17 degrees of freedom. Since the 5% critical value for this distribution is 3.59 we cannot reject H_0 and deduce therefore that MODEL 1 is not a significant improvement on MODEL 2. This is the clincher and so MODEL 2 would be preferred

24. a. The model is significant overall with all its predictors significant also. This is borne out by the pvalues for the F and t statistics which are < α = 0.05 without exception. A two-sided Durbin-Watson test (α = 0.05) yields an inconclusive result since dL =1.01 < d= 1.09 < dU = 1.41 There is a single outlier but this does not appear to be too extreme according to its standardized residual value. The main problem with the model is multicollinearity as evidenced by the high correlations between all variables – and which was somehow played down by previous VIF values. The earlier t test results are therefore likely to be very dubious.

The Stepwise output features a new predictor laglnimp which happens to be selected for the final step 2 model. The problem is that this model too is likely to suffer from multicollinearity.

b. Hence the preferred model of all those considered is the Stepwise (step 1) simple regression model with the lngdp predictor. This model had a very high R square (97.5%) and is highly significant according to the pvalue result (0.000) provided by Stepwise.

25. a.

```
        y      x1
x1   0.964
     0.000

x2   0.873   0.815
     0.001   0.004

Cell Contents: Pearson correlation
               P-Value
```

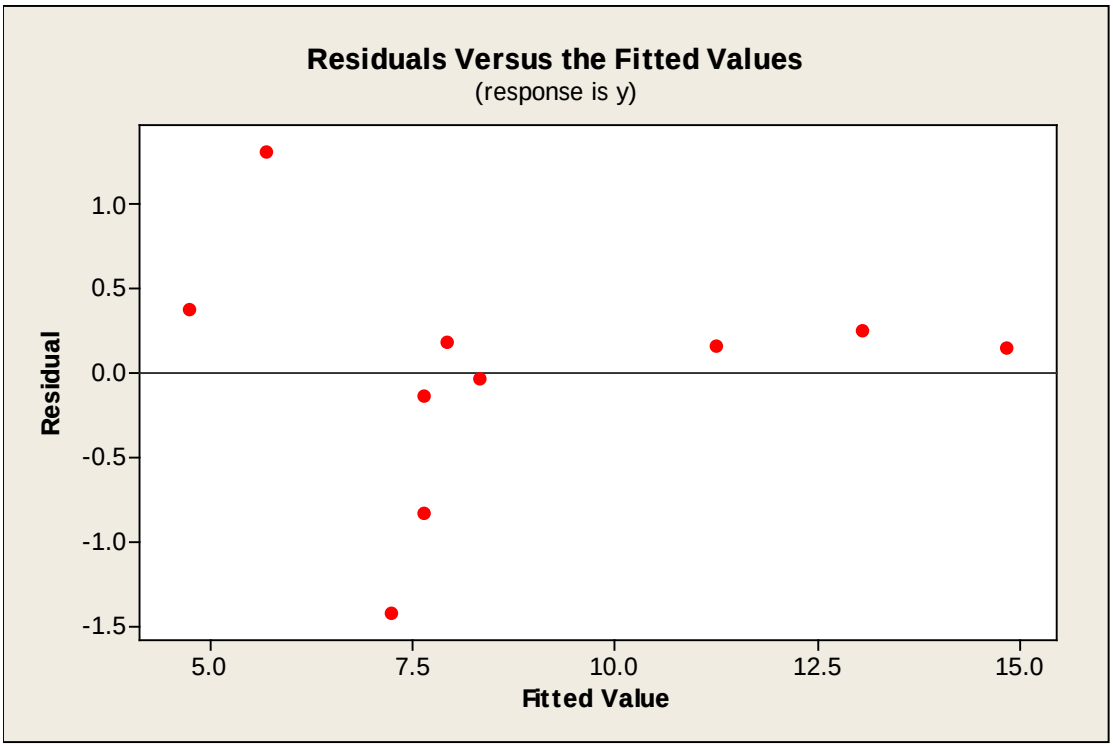
All the correlations here are very high as well as being highly significant (p value < 0.01)

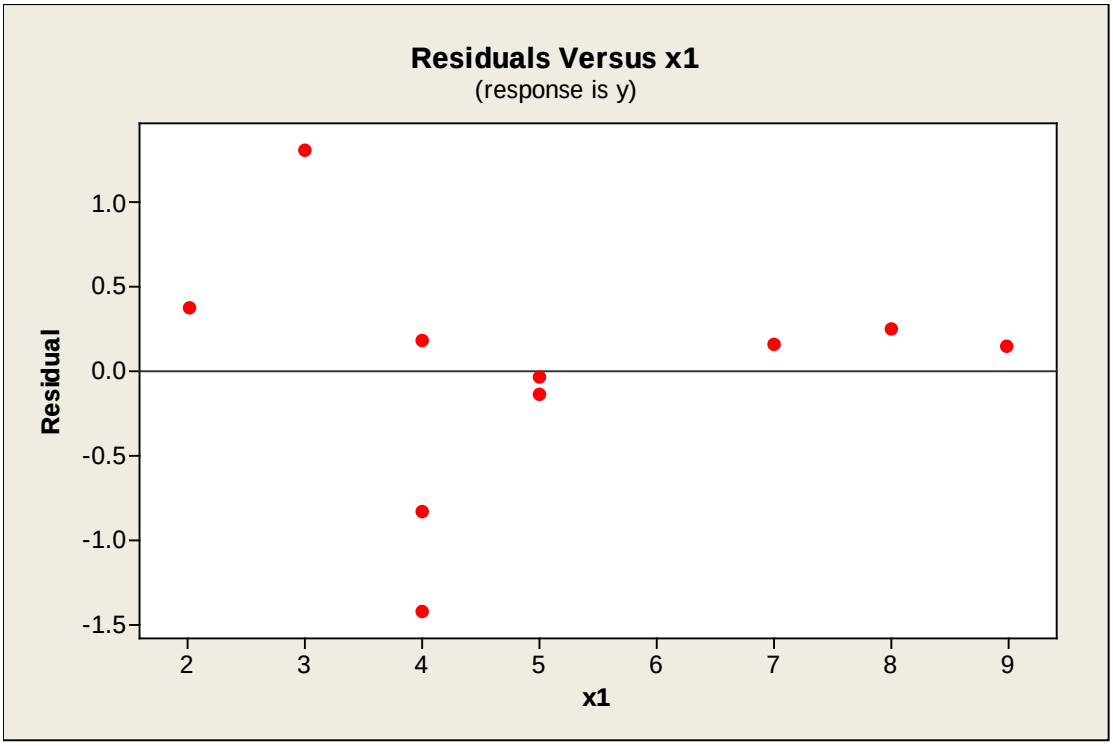
b.

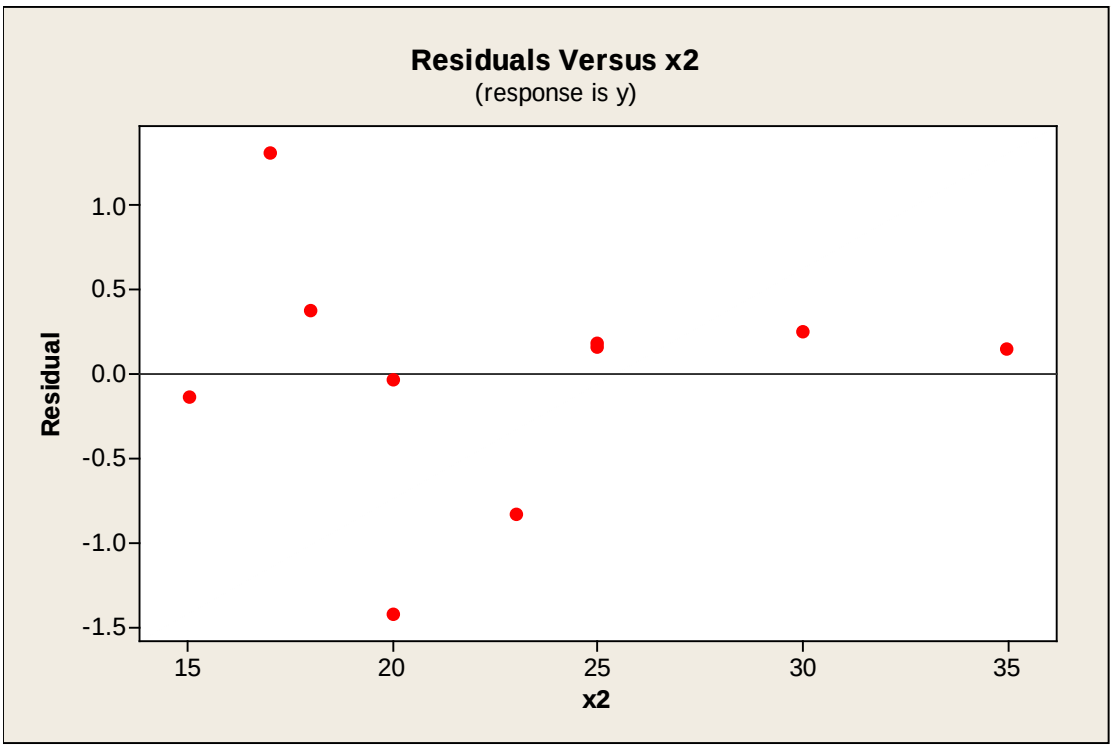
```
The regression equation is
y = 0.01 + 1.11 x1 + 0.139 x2


Predictor     Coef  SE Coef     T      P  VIF
Constant     0.009    1.098  0.01  0.994
x1          1.1102   0.2116  5.25  0.001  3.0
x2         0.13855  0.07648  1.81  0.113  3.0


S = 0.821927   R-Sq = 95.2%   R-Sq(adj) = 93.8%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       2  93.072  46.536  68.88  0.000
Residual Error   7   4.729   0.676
Total            9  97.801


Source  DF  Seq SS
x1       1  90.855
x2       1   2.217
```

c. The VIF values in (b) do not suggest problems of multicollinearity are possible despite the significant correlation between x_1 and x_2 found in (a). Clearly however there <u>are</u> problems since x_2 (with a p value = 0.113 > 0.05 = α) is not a significant predictor whereas x_1 (with a p value = 0.001 < 0.05 = α) is. (According to the correlations both should be significant predictors.)

d. The relevant plots are as follows:







None of these plots seem to be out of line with theoretical assumptions but the sample size is relatively small so this is not altogether unexpected.

26. a.

Stepwise Regression: y versus x1, x2

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

Response is y on 2 predictors, with N = 10

Step                1         2
Constant     1.575278  0.008928

x1               1.42      1.11
T-Value         10.23      5.25
P-Value         0.000     0.001

x2                         0.139
T-Value                     1.81
P-Value                    0.113

S               0.932     0.822
R-Sq            92.90     95.16
R-Sq(adj)       92.01     93.78
Mallows C-p       4.3       3.0
```

Best Subsets Regression: y versus x1, x2

Response is y

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	x 1	x 2
1	92.9	92.0	4.3	0.93182	X	
1	76.1	73.2	28.5	1.7078		X
2	95.2	93.8	3.0	0.82193	X	X

Stepwise seems to favour the full two predictor model which also corresponds to model described on the bottom line of the Best Subsets output. Yet the **adj** R2 value (of 92.0%) for the single x_1 predictor model (step 1) is almost the same as that for the full model (93.8%). Note that the corresponding difference between root mean square error values is slightly more pronounced.

b. Despite this, the single x_1 alternative might be favoured given earlier concerns about hidden multicollinearity.

27. Relevant regression output is as follows:

```
The regression equation is
y = - 6.93 + 0.698 x1 + 1.10 x2


Predictor     Coef  SE Coef      T      P
Constant   -6.9262   0.7543  -9.18  0.000
x1          0.6979   0.1835   3.80  0.002
x2          1.0978   0.1911   5.74  0.000


S = 0.249288   R-Sq = 96.0%   R-Sq(adj) = 95.5%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       2  22.344  11.172  179.77  0.000
Residual Error  15   0.932   0.062
Total           17  23.276


Source  DF  Seq SS
x1       1  20.294
x2       1   2.050


Unusual Observations

Obs    x1       y     Fit  SE Fit  Residual  St Resid
 16  8.60  8.2000  7.7481  0.1434    0.4519      2.22R

R denotes an observation with a large standardized residual.

a.

New
Obs     Fit  SE Fit       90% CI
  1  7.2196  0.0709  (7.0953, 7.3439)


b.
New
Obs     Fit  SE Fit      90% PI
  1  7.2196  0.0709  (6.7653, 7.6740)
```

c. The (confidence) interval in a. corresponds to any proposal with the scores x_1 = 8.0 and x_2 = 7.8 whereas the (prediction) interval in b. corresponds to a specific proposal with the scores x_1 = 8.0 and x_2 = 7.8

28. a.

Stepwise Regression: y versus x1, x2

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15


Response is y on 2 predictors, with N = 18


Step             1       2
Constant     -6.436  -6.926

x2             1.73    1.10
T-Value       13.69    5.74
P-Value       0.000   0.000

x1                     0.70
T-Value                3.80
P-Value               0.002

S             0.338   0.249
R-Sq          92.13   96.00
R-Sq(adj)     91.64   95.46
Mallows C-p    15.5     3.0
```

Best Subsets Regression: y versus x1, x2

```
Response is y

                        Mallows           x x
Vars  R-Sq  R-Sq(adj)      C-p        S  1 2
   1  92.1       91.6     15.5  0.33834    X
   1  87.2       86.4     34.0  0.43170  X
   2  96.0       95.5      3.0  0.24929  X X
```

From the Stepwise and Best Subsets output it is clear the full two predictor model is most favoured. Both predictors X1 and X2 contribute very significantly to the model according to the relevant T ratios and pvalues. The root mean square error value is also markedly better for this model than the alternatives.

b. The two predictor model is conspicuously better than either single predictor alternatives for representing the data.

29. a. Relevant output is as follows:

Correlations: Y, X1, X2, X3

```
          Y      X1      X2
X1    0.828
      0.001

X2    0.492   0.041
      0.104   0.899

X3   -0.235   0.012  -0.578
      0.463   0.969   0.049


Cell Contents: Pearson correlation
               P-Value
```

From the correlations here, it can be seen X1 and X2 are respectively the most correlated with y.

Regression Analysis: Y versus X1, X2, X3

```
The regression equation is
Y = 1.41 + 0.589 X1 + 0.364 X2 + 0.087 X3


Predictor     Coef  SE Coef     T      P
Constant     1.409    5.717  0.25  0.811
X1         0.58919  0.08290  7.11  0.000
X2          0.3641   0.1066  3.42  0.009
X3          0.0870   0.3965  0.22  0.832


S = 3.10495   R-Sq = 89.7%   R-Sq(adj) = 85.8%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        3  670.54  223.51  23.18  0.000
Residual Error    8   77.13    9.64
Total            11  747.67

Source  DF  Seq SS
X1       1  513.14
X2       1  156.94
X3       1    0.46
```

The output above shows a significant linear model has been fitted (the pvalue for the F ratio is .000 < 0.05 = α). X1 and X2 are significant predictors of Y (for each of the T ratios pvalue < 0.05 = α).

b.

Stepwise Regression: Y versus X1, X2, X3

Backward elimination. Alpha-to-Remove: 0.1

Response is Y on 3 predictors, with N = 12

Step	1	2
Constant	1.409	2.356
X1	0.589	0.590
T-Value	7.11	7.53
P-Value	0.000	0.000
X2	0.364	0.351
T-Value	3.42	4.27
P-Value	0.009	0.002
X3	0.09	
T-Value	0.22	
P-Value	0.832	
S	3.10	2.94
R-Sq	89.68	89.62
R-Sq(adj)	85.82	87.32
Mallows C-p	4.0	2.0

The best model according to this procedure is the one featuring the two predictors X1 and X2. This is essentially what we would have expected following the regression analysis in (a).

30. The initial model does not seem to affected by multicollinearity from the VIF values yet the sample correlations between predictors do look potentially problematic in places e.g. the correlation between cmplian and rises of 0.666. T ratios for the model are given below:

Predictor	**Coefficient**	**Standard Error**	**t**
cmplain	0.613	0.161	3.81
prvileg	-0.073	0.136	-0.54
learn	0.32	0.169	1.89
rises	0.082	0.222	0.37
critcal	0.038	0.147	0.26
advance	-0.217	0.178	-1.22

Only the t ratio for cmplain is significant since under H_0: $\beta_i = 0$ i= 1, 2, 3.. 6 t is distributed on 23 = n – p – 1 = 30 – 6 -1 degrees of freedom. And apart from cmplain none of the ratios above are > $t_{.025}(23) = 2.069$ or < - $t_{.025}(23) = -2.069$.

Using the Total sum of squares result of 4296.97 and the root mean square error value = 7.0680 from the bottom line of the Best Subsets output, the ANOVA table for the model can be constructed as follows:

	df	**SS**	**MS**	**F**
Regression	6	3147.968	524.661	10.502
Error	23	1149.002	49.957	
Total	29	4296.97		

The F statistic here is significant (10.502 > 2.53 = $F_{.05}$ for an F distribution on 6 and 23 degrees of freedom. Thus we would reject H_0: $\beta_1 = \beta_2 = \ldots \beta_6 = 0$ and deduce the model is significant.

From the Best subsets output two models stand out - namely the first of the two predictor models and the first of the three predictor models listed. The three predictor model features the advance predictor which is not strongly correlated with y. In this sense the two predictor one might therefore be preferred. The full six predictor model falls well short of either of these alternatives.

31. a. MODEL 1

This is a significant regression model based on the F ratio result of 19.25. (Under H_0: $\beta_1 = 0$, F has an F distribution on 1 and 18 degrees of freedom. The 5% critical value for this distribution is $F_{.05} = 4.41$. Since $F = 19.25 > F_{.05} = 4.41$ we would therefore reject H_0.

MODEL 2

From the Error SS information provided we can recreate the corresponding ANOVA table as follows:

	df	SS	MS	F
Regression	3	62.636	20.879	10.553
Error	16	31.655	1.978	
Total	19	94.291		

(since the TOTAL SS remains the same however many predictors we choose for the modelling). The F statistic here is significant ($10.553 > 3.24 = F_{.05}$ for an F distribution on 3 and 16 degrees of freedom. Thus we would reject H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ and deduce the model is significant.

Corresponding t ratios are as follows:

	Coefficient	standard error	t
race	-1.91	1.54	-1.24
test	1.31	0.67	1.96
raceXtest	2	0.954	2.10

The relevant t distribution under H_0: $\beta_i = 0$ i= 1, 2, 3 is t on 16 degrees of freedom. As none of the ratios above are $> t_{.025}(16) = 2.12$ or $< - t_{.025}(17) = -2.12$ we cannot reject H_0 for i= 0, 1, 2,3

b. To test if MODEL 2 is an improvement over MODEL 1 we note the following error sums of squares details:

Model	Error SS
1	45.568
2	31.655

Based on formula (16.11) the relevant test statistic is:

$$F = \frac{(45.568 - 31.655)/2}{31.655/(20-3-1)} = 3.52$$

Under the hypothesis H_0: $\beta_2 = \beta_3 = 0$ F has an F distribution on 2 and 16 degrees of freedom. Since the 5% critical value for this distribution is 3.63 we cannot reject H_0 and deduce MODEL 2 is not a significant improvement on MODEL 1. Hence introducing the race variable into the model either explicitly or through an interaction does not seem to have improved the model's performance. The fact that both the confidence and prediction intervals overlap for minority and white candidates for MODEL 2 suggests that the classification has no significant value for the exercise. MODEL 1 would therefore be preferred.

32. The VIF values here reveal a major problem of multicollinearity. Thus estimated coefficients for the regression model as well as corresponding t tests are likely to be very dubious. From the correlation matrix the source of the multicollinearity seems to be between the doprod and consum predictors. With a correlation of 0.999 they would be regarded mathematically by MINITAB as being in essence, identical variables. One of them needs to be dropped from the model – it is up to the analyst to decide which. Whether the stock predictor is worth retaining is another issue and could be investigated using stepwise procedures.

The $R^2 = 97.3\%$ result is impressive and corresponds with an F value for the

ANOVA table of

$$F = \frac{R^2(n-p-1)}{(1-R^2)p} = \frac{0.973(18-3-1)}{(1-0.973)3} = 168.17$$

Under H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ F has an F distribution on 3 and 14 degrees of freedom. The 5% critical value for this distribution is $F_{.05}$ =3.34. Since F = 168.17 > $F_{.05}$ =3.34 we would reject H_0 and deduce the model is significant.

From the Durbin Watson tables provided, it can be shown for n = 18 p = 3 and $\alpha = 0.025$ that $d_L = 0.82$ and $d_U = 1.56$. Based on a two sided test approach we deduce the test result of d = 0.24 indicates significant positive autocorrelation of errors exists.

b. The model is very problematic as it stands. Both the multicollinearity and first order serial correlation of errors problems need to be resolved before it can be seriously considered as statistical prediction tool.

33. MODEL 1

T statistics can be calculated for the estimated model by calculating the regression coefficient / standard error ratios as follows:

	t
Constant	1.39
x1	0.84
x2	-1.10
x3	-1.21
x4	0.05
x5	0.69
x6	1.37

Given:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$ i= 0, 1, 2, …6

and α = .05 the above ratios under H_0 have a t distribution on 17 degrees of

freedom where 17 = n – p - 1 where n = 24 and p = the number of independent variables = 6. As none of the ratios above are > $t_{.025}(17) = 2.11$ or < - $t_{.025}(17) = -2.11$ we cannot reject H_0 for i= 0, 1, 2, …6.

From the R^2 value of 0.385, the F statistic for the ANOVA table can be calculated as follows:

$$F = \frac{R^2(n-p-1)}{(1-R^2)p} = \frac{0.385(24-6-1)}{(1-0.385)6} = 1.78$$

Under H_0: $\beta_1 = \beta_2 = \ldots\ldots = \beta_6 = 0$ F has an F distribution on 6 and 17 degrees of freedom. It can be shown that for this distribution $F_{.05} = 2.92$. As $F = 1.78 < 2.92$ H_0 cannot be rejected and we deduce the multiple regression model is not significant.

Superficially the sample correlations provided do not indicate problems of multicollinearity (though it would have helped to have been provided with corresponding pvalues).

The Durbin Watson statistic cannot strictly be tested using the critical values provided in the book (only a maximum of 5 predictors is catered for) but given for n = 24 p = 5 and
$\alpha = 0.025$ that $d_L = 0.83$ and $d_U = 1.79$ a two-sided test is likely to be inconclusive.

MODEL 2

The second model features an additional four interaction terms. Their presence seems to considerably improve the R^2 result and it can be shown based on the t ratios below that x1and x3x1 are significant predictors:

	t
Constant	-0.86
x1	2.94
x2	1.8
x3	0.50
x4	1.78
x5	1.14
x6	1.63
x3x1	-2.53
x1x4	-2.01
x1x5	-0.57
x1x6	-0.21

(As before

H_0: $\beta_i = 0$

H_1: $\beta_i = 0$ i= 0, 1, 2, …10

and $\alpha = .05$. The above ratios under H_0 have a t distribution on 13 degrees of freedom where 13 = n – p – 1, n = 24 and p = the number of independent variables = 10. The relevant critical values in this case are $t_{.025}(17) = 2.17$ and $-t_{.025}(17) = -2.17$).

a. To test if MODEL 2 is an improvement over MODEL 1 we calculate the error sums of squares = mean square error X degrees of freedom for each model = $s^2(n-p-1)$.

From the details provided we have:

Model 1	s^2	(n-p-1)	Error SS
1	8.044	17	136.748
2	4.86	13	63.18

Based on formula (16.11) the relevant test statistic is:

$$F = \frac{(136.748 - 63.18)/4}{63.18/13} = 3.78$$

Under the hypothesis H_0: $\beta_7 = \beta_8 = \ldots\ldots = \beta_{10} = 0$ F has an F distribution on 4 and 13 degrees of freedom. Since the 5% critical value for this distribution is 3.18 we reject H_0 in favour of

H_1: One or more of the parameters is not equal to zero

and deduce MODEL 2 is a significant improvement on MODEL 1.

b. Both models are problematic but MODEL 2 is at least a significant improvement on MODEL 1 so this would be preferred. (MODEL 1 is not significant in any way.) Many terms in MODEL 2 are not significant and needlessly clutter it so ideally these should be eliminated using, for example, the stepwise technique before the model would actually come into operation.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Seventeen

Forecasting

Textbook Exercises (1-31)

Textbook Exercise Solutions (1-31)

Supplementary Exercises (32-46)

Supplementary Exercise Solutions

Chapter 17: Forecasting

Textbook Exercises:

1. Consider the following time series data.

Week	1	2	3	4	5	6
Value	18	13	16	11	17	14

   Using the naive method (most recent value) as the forecast for the next week, compute the following measures of forecast accuracy.

   a. Mean absolute error.
   b. Mean squared error.
   c. Mean absolute percentage error.
   d. What is the forecast for week 7?

2. Refer to the time series data in exercise 1. Using the average of all the historical data as a forecast for the next period, compute the following measures of forecast accuracy.

   a. Mean absolute error.
   b. Mean squared error.
   c. Mean absolute percentage error.
   d. What is the forecast for week 7?

3. Exercises 1 and 2 used different forecasting methods. Which method appears to provide the more accurate forecasts for the historical data? Explain.

4. Consider the following time series data.

Month	1	2	3	4	5	6	7
Value	24	13	20	12	19	23	15

   a. Compute MSE using the most recent value as the forecast for the next period. What is the forecast for month 8?
   b. Compute MSE using the average of all the data available as the forecast for the next period. What is the forecast for month 8?
   c. Which method appears to provide the better forecast?

5. Consider the following time series data.

Week	1	2	3	4	5	6
Value	18	13	16	11	17	14

   a. Construct a time series plot. What type of pattern exists in the data?
   b. Develop the three-week moving average forecasts for this time series. Compute MSE and a forecast for week 7.
   c. Use $\alpha = 0.2$ to compute the exponential smoothing forecasts for the time series. Compute MSE and a forecast for week 7.
   d. Compare the three-week moving average approach with the exponential smoothing approach using $\alpha = 0.2$. Which appears to provide more accurate forecasts based on MSE? Explain.
   e. Use a smoothing constant of $\alpha = 0.4$ to compute the exponential smoothing forecasts. Does a smoothing constant of 0.2 or 0.4 appear to provide more accurate forecasts based on MSE? Explain.

6. Consider the following time series data.

Month	1	2	3	4	5	6	7
Value	24	13	20	12	19	23	15

   Construct a time series plot. What type of pattern exists in the data?
   a. Develop the three-week moving average forecasts for this time series. Compute MSE and a forecast for week 8.
   b. Use $\alpha = 0.2$ to compute the exponential smoothing forecasts for the time series. Compute MSE and a forecast for week 8.
   c. Compare the three-week moving average approach with the exponential smoothing approach using $\alpha = 0.2$. Which appears to provide more accurate forecasts based on MSE?
   d. Use a smoothing constant of $\alpha = 0.4$ to compute the exponential smoothing forecasts. Does a smoothing constant of 0.2 or 0.4 appear to provide more accurate forecasts based on MSE? Explain.

7. Refer to the petrol sales time series data in Table 17.1.
   a. Compute four-week and five-week moving averages for the time series.
   b. Compute the MSE for the four-week and five-week moving average forecasts.

c. What appears to be the best number of weeks of past data (three, four, or five) to use in the moving average computation? Recall that MSE for the three-week moving average is 10.22.

8. Refer again to the petrol sales time series data in Table 17.1.
   a. Using a weight of 1/2 for the most recent observation, 1/3 for the second most recent observation, and 1/6 for third most recent observation, compute a three-week weighted moving average for the time series.
   b. Compute the MSE for the weighted moving average in part (a). Do you prefer this weighted moving average to the unweighted moving average? Remember that the MSE for the unweighted moving average is 10.22.
   c. Suppose you are allowed to choose any weights as long as they sum to 1. Could you always find a set of weights that would make the MSE at least as small for a weighted moving average than for an unweighted moving average? Why or why not?

9. With the petrol time series data from Table 17.1, show the exponential smoothing forecasts using $\alpha = 0.1$.
   a. Applying the MSE measure of forecast accuracy, would you prefer a smoothing constant of $\alpha = 0.1$ or $\alpha = 0.2$ for the petrol sales time series?
   b. Are the results the same if you apply MAE as the measure of accuracy?
   c. What are the results if MAPE is used?

10. With a smoothing constant of $\alpha = 0.2$, equation (17.2) shows that the forecast for week 13 of the petrol sales data from Table 17.1 is given by $F_{13} = 0.2Y_{12} + 0.8F_{12}$. However, the forecast for week 12 is given by $F_{12} = 0.2Y_{11} + 0.8F_{11}$. Thus, we could combine these two results to show that the forecast for week 13 can be written

$$F_{13} = 0.2Y_{12} + 0.8(0.2Y_{11} + 0.8F_{11}) = 0.2Y_{12} + 0.16Y_{11} + 0.64F_{11}$$

   a. Making use of the fact that $F_{11} = 0.2Y_{10} + 0.8F_{10}$ (and similarly for F_{10} and F_9), continue to expand the expression for F13 until it is written in terms of the past data values Y_{12}, Y_{11}, Y_{10}, Y_9, Y_8, and the forecast for period 8.
   b. Refer to the coefficients or weights for the past values Y_{12}, Y_{11}, Y_{10}, Y_9, Y_8. What observation can you make about how exponential smoothing weights

past data values in arriving at new forecasts? Compare this weighting pattern with the weighting pattern of the moving averages method.

11. For SIS Cargo Services in Dubai, the monthly percentages of all shipments received on time over the past 12 months are 80, 82, 84, 83, 83, 84, 85, 84, 82, 83, 84, and 83.
    a. Construct a time series plot. What type of pattern exists in the data?
    b. Compare the three-month moving average approach with the exponential smoothing approach for $\alpha = 0.2$. Which provides more accurate forecasts using MSE as the measure of forecast accuracy?
    c. What is the forecast for next month?

12. The values of Austrian building contracts (in millions of euros) for a 12-month period follow.

    240 350 230 260 280 320 220 310 240 310 240 230

    a. Construct a time series plot. What type of pattern exists in the data?
    b. Compare the three-month moving average approach with the exponential smoothing forecast using $\alpha = 0.2$. Which provides more accurate forecasts based on MSE?
    c. What is the forecast for the next month?

13 The following data represent indices for the seasonally adjusted merchandise trade volumes for New Zealand from 2005–2008.

Year	Quarter	Index	Year	Quarter	Index
2005	Mar	999	2007	Mar	1046
	Jun	998		Jun	1057
	Sep	981		Sep	1052
	Dec	1007		Dec	1157
2006	Mar	993	2008	Mar	1111
	Jun	1004		Jun	1068
	Sep	1062		Sep	1043
	Dec	1005			

a. Compute three- and four-quarter moving averages for this time series. Which moving average provides the better forecast for the fourth quarter of 2008?

b. Plot the data. Do you think the exponential smoothing model would be appropriate for forecasting in this case?

14. Consider the following time series data.

t	1	2	3	4	5
Y_t	6	11	9	14	15

a. Construct a time series plot. What type of pattern exists in the data?

b. Develop the linear trend equation for this time series.

c. What is the forecast for $t = 6$?

15 Refer to the time series in exercise 14. Use Holt's linear exponential smoothing method with $\alpha = 0.3$ and $\beta = 0.5$ to develop a forecast for $t = 6$.

16. Consider the following time series.

t	1	2	3	4	5	6	7
Y_t	120	110	100	96	94	92	88

a. Construct a time series plot. What type of pattern exists in the data?

b. Develop the linear trend equation for this time series.

c. What is the forecast for $t = 8$?

17. Consider the following time series.

t	1	2	3	4	5	6	7
Y_t	82	60	44	35	30	29	35

a. Construct a time series plot. What type of pattern exists in the data?

b. Using Minitab or Excel, develop the quadratic trend equation for the time series.

c. What is the forecast for $t = 8$?

18 Car sales at Perez Motors provided the following ten-year time series.

Year	Sales	Year	Sales
1	400	6	260
2	390	7	300
3	320	8	320
4	340	9	340
5	270	10	370

Plot the time series and comment on the appropriateness of a linear trend. What type of functional form do you believe would be most appropriate for the trend pattern of this time series?

19 Numbers of overseas visitors to Ireland (000s) estimated by the Central Statistics Office for the years 2001–2007 are as follows:

2001	2002	2003	2004	2005	2006	2007
5990	6065	6369	6574	6977	7709	8012

a. Graph the data and assess its suitability for linear trend projection.

b. Use a linear trend projection to forecast this time series for 2008–2009.

20 GDP (Singapore $) for 1990–2007 are tabulated below (*Statistics Singapore, 2009*).

Year	S$	Year	S$
1990	66 778	1999	140 022
1991	74 570	2000	159 840
1992	80 984	2001	153 398
1993	93 971	2002	158 047
1994	107 957	2003	162 288
1995	119 470	2004	184 508
1996	130 502	2005	199 375
1997	142 341	2006	216 995
1998	137 902	2007	243

a. Graph this time series. Does a linear trend appear to be present?

b. Develop a linear trend equation for this time series.

c. Use the trend equation to estimate the GDP for the years 2008–2010.

21 Gross revenue data (in millions of euros) for Hispanic Airlines for a ten-year period follow.

Year	Revenue	Year	Revenue
1	2428	6	4264
2	2951	7	4738
3	3533	8	4460
4	3618	9	5318
5	3616	10	6915

a. Develop a linear trend equation for this time series. Comment on what the equation tells about the gross revenue for Hispanic airlines for the ten-year period.

b. Provide the forecasts for gross revenue for years 11 and 12.

22 Consider the following time series.

Quarter	**Year 1**	**Year 2**	**Year 3**
1	71	68	62
2	49	41	51
3	58	60	53
4	78	81	72

a. Construct a time series plot. What type of pattern exists in the data?

b. Use the following dummy variables to develop an estimated regression equation to account for seasonal effects in the data: Qtr1 = 1 if Quarter 1, 0 otherwise; Qtr2 = 1 if Quarter 2, 0 otherwise; Qtr3 = 1 if Quarter 3, 0 otherwise.

c. Compute the quarterly forecasts for next year.

23 Consider the following time series data.

Quarter	**Year 1**	**Year 2**	**Year 3**
1	4	6	7
2	2	3	6
3	3	5	6
4	5	7	8

a. Construct a time series plot. What type of pattern exists in the data?

b. Use the following dummy variables to develop an estimated regression equation to account for any seasonal and linear trend effects in the data: Qtr1 = 1 if Quarter 1, 0 otherwise; Qtr2 = 1 if Quarter 2, 0 otherwise; Qtr3 = 1 if Quarter 3, 0 otherwise.

c. Compute the quarterly forecasts for next year.

24 The quarterly sales data (number of copies sold) for a college textbook over the past three years follow.

Quarter	Year 1	Year 2	Year 3
1	1690	1800	1850
2	940	900	1100
3	2625	2900	2930
4	2500	2360	2615

a. Construct a time series plot. What type of pattern exists in the data?

b. Use the following dummy variables to develop an estimated regression equation to account for any seasonal effects in the data: Qtr1 = 1 if Quarter 1, 0 otherwise; Qtr2 = 1 if Quarter 2, 0 otherwise; Qtr3 = 1 if Quarter 3, 0 otherwise.

c. Compute the quarterly forecasts for next year.

d. Let $t = 1$ to refer to the observation in quarter 1 of year 1; $t = 2$ to refer to the observation in quarter 2 of year 1; ... and $t = 12$ to refer to the observation in quarter 4 of year 3. Using the dummy variables defined in part (b) and t, develop an estimated regression equation to account for seasonal effects and any linear trend in the time series. Based upon the seasonal effects in the data and linear trend, compute the quarterly forecasts for next year.

25 Air pollution control specialists in northern Poland monitor the amount of ozone, carbon dioxide, and nitrogen dioxide in the air on an hourly basis. The hourly time series data exhibit seasonality, with the levels of pollutants showing patterns that vary over the hours in the day. On July 15, 16, and 17, the following levels of nitrogen dioxide were observed for the 12 hours from 6:00 A. M. to 6:00 P. M.

July 15:	25	28	35	50	60	60	40	35	30	25	25	20
July 16:	28	30	35	48	60	65	50	40	35	25	20	20
July 17:	35	42	45	70	72	75	60	45	40	25	25	25

a. Construct a time series plot. What type of pattern exists in the data?

b. Use the following dummy variables to develop an estimated regression equation to account for the seasonal effects in the data.

   Hour1 = 1 if the reading was made between 6:00 A.M. and 7:00 A.M.; 0 otherwise

   Hour2 = 1 if if the reading was made between 7:00 A.M. and 8:00 A.M.; 0 otherwise

   .

   .

   .

   Hour11 = 1 if the reading was made between 4:00 P.M. and 5:00 P.M., 0 otherwise.

Note that when the values of the 11 dummy variables are equal to 0, the observation corresponds to the 5:00 P.M. to 6:00 P.M. hour.

c. Using the estimated regression equation developed in part (a), compute estimates of the levels of nitrogen dioxide for July 18.

d. Let $t = 1$ to refer to the observation in hour 1 on July 15; $t = 2$ to refer to the observation in hour 2 of July 15;0... and $t = 36$ to refer to the observation in hour 12 of July 17. Using the dummy variables defined in part (b) and t, develop an estimated regression equation to account for seasonal effects and any linear trend in the time series. Based upon the seasonal effects in the data and linear trend, compute estimates of the levels of nitrogen dioxide for July 18.

26 Consider the following time series data.

Quarter	Year 1	Year 2	Year 3
1	4	6	7
2	2	3	6
3	3	5	6
4	5	7	8

a. Construct a time series plot. What type of pattern exists in the data?
b. Show the four-quarter and centred moving average values for this time series.
c. Compute seasonal indices and adjusted seasonal indices for the four quarters.

27 Refer to exercise 26.

a. Deseasonalize the time series using the adjusted seasonal indices computed in part (c) of exercise 35.
b. Using Minitab or Excel, compute the linear trend regression equation for the deseasonalized data.
c. Compute the deseasonalized quarterly trend forecast for Year 4.
d. Use the seasonal indices to adjust the deseasonalized trend forecasts computed in part (c).

28 The quarterly sales data (number of copies sold) for a college textbook over the past three years follow.

Quarter	Year 1	Year 2	Year 3
1	1690	1800	1850
2	940	900	1100
3	2625	2900	2930
4	2500	2360	2615

a. Construct a time series plot. What type of pattern exists in the data?
b. Show the four-quarter and centred moving average values for this time series.
c. Compute the seasonal and adjusted seasonal indices for the four quarters.
d. When does the publisher have the largest seasonal index? Does this result appear reasonable? Explain.
e. Deseasonalize the time series.

f. Compute the linear trend equation for the deseasonalized data and forecast sales using the linear trend equation.

g. Adjust the linear trend forecasts using the adjusted seasonal indices computed in part (c).

29 Quarterly sales data for the number of houses sold over the past four years or so by a national chain are as follows:

Year	Ql	Q2	Q3	Q4
1	200	212	229	207
2	195	204	216	202
3	201	209	221	205
4	208	217	231	213
5	218			

a. Decompose the series into trend, seasonal and random components using a multiplicative model.

b. Hence derive forecasts of the number of houses that will be sold in the next four quarters.

c. Comment on the quality of your modelling results.

30 The following table shows the number of passengers per quarter (in thousands) who flew with MBI Junior for the first quarter of this year and the three years preceding:

Year	Q1	Q2	Q3	Q4
1	44	92	156	68
2	60	112	180	80
3	64	124	200	104
4	76			

a. Decompose the series into trend, seasonal and random components using an additive model.

b. Hence derive forecasts of the passenger numbers in the next four quarters.

c. Comment on the quality of your modelling.

31 The data below relates to the UK and shows the number of marriages (‘000’s) over a recent four year period.

Year	Quarter	Marriages	Year	Quarter	Marriages
1	1	52.9	3	1	41.7
	2	114.3		2	100.5
	3	138.7		3	138.5
	4	62.7		4	60.9
2	1	45.6	4	1	41.7
	2	101.9		2	100.5
	3	146.2		3	138.5
	4	62.3		4	60.9

a. Using the decomposition method, forecast marriages for the next four quarters in the series.

Chapter 17: Forecasting

Textbook Exercises Solutions:

1.

a.

Week	Time-Series Value	Forecast	Forecast Error	$(\text{Error})^2$
1	8			
2	13			
3	15			
4	17	12	5	25
5	16	15	1	1
6	9	16	-7	49
				75

Forecast for week 7 is (17 + 16 + 9) / 3 = 14

b. MSE = 75 / 3 = 25

c. Smoothing constant = .3.

Week t	Time-Series Value Y_t	Forecast F_t	Forecast Error $Y_t - F_t$	Squared Error $(Y_t - F_t)^2$
1	8			
2	13	8.00	5.00	25.00
3	15	9.00	6.00	36.00
4	17	10.20	6.80	46.24
5	16	11.56	4.44	19.71
6	9	12.45	-3.45	11.90
				138.85

Forecast for week 7 is .2(9) + .8(12.45) = 11.76

d. For the $\alpha = .2$ exponential smoothing forecast MSE = 138.85 / 5 = 27.77. Since the three-week moving average has a smaller MSE, it appears to provide the better forecasts.

e. Smoothing constant = .4.

Week t	Time-Series Value Y_t	Forecast F_t	Forecast Error $Y_t - F_t$	Squared Error $(Y_t - F_t)^2$
1	8			
2	13	8.0	5.0	25.00
3	15	10.0	5.0	25.00
4	17	12.0	5.0	25.00
5	16	14.0	2.0	4.00
6	9	14.8	-5.8	33.64
				112.64

MSE = 112.64 / 5 = 22.53. A smoothing constant of .4 appears to provide better forecasts.

Forecast for week 7 is .4(9) + .6(14.8) = 12.48

2.

a.

Week	Time-Series Value	4-Week Moving Average Forecast	(Error)^2	5-Week Moving Average Forecast	(Error)^2
1	17				
2	21				
3	19				
4	23				
5	18	20.00	4.00		
6	16	20.25	18.06	19.60	12.96
7	20	19.00	1.00	19.40	0.36
8	18	19.25	1.56	19.20	1.44
9	22	18.00	16.00	19.00	9.00
10	20	19.00	1.00	18.80	1.44
11	15	20.00	25.00	19.20	17.64
12	22	18.75	10.56	19.00	9.00
			77.18		51.84

b. MSE(4-Week) = 77.18 / 8 = 9.65
MSE(5-Week) = 51.84 / 7 = 7.41

c. For the limited data provided, the 5-week moving average provides the smallest MSE.

3. a.

Week	Time-Series Value	Weighted Moving Average Forecast	Forecast Error	$(\text{Error})^2$
1	17			
2	21			
3	19			
4	23	19.33	3.67	13.47
5	18	21.33	-3.33	11.09
6	16	19.83	-3.83	14.67
7	20	17.83	2.17	4.71
8	18	18.33	-0.33	0.11
9	22	18.33	3.67	13.47
10	20	20.33	-0.33	0.11
11	15	20.33	-5.33	28.41
12	22	17.83	4.17	17.39
				103.43

b. MSE = 103.43 / 9 = 11.49
Prefer the unweighted moving average here.

c. You could always find a weighted moving average at least as good as the unweighted one. Actually the unweighted moving average is a special case of the weighted ones where the weights are equal.

4.

Week	Time-Series Value	Forecast	Error	$(\text{Error})^2$
1	17			
2	21	17.00	4.00	16.00
3	19	17.40	1.60	2.56
4	23	17.56	5.44	29.59
5	18	18.10	-0.10	0.01
6	16	18.09	-2.09	4.37
7	20	17.88	2.12	4.49
8	18	18.10	-0.10	0.01
9	22	18.09	3.91	15.29
10	20	18.48	1.52	2.31
11	15	18.63	-3.63	13.18
12	22	18.27	3.73	13.91
				101.72

MSE = 101.72 / 11 = 9.25

$\alpha = .2$ provided a lower MSE; therefore $\alpha = .2$ is better than $\alpha = .1$

5. a. $F_{13} = .2Y_{12} + .16Y_{11} + .64(.2Y_{10} + .8F_{10}) = .2Y_{12} + .16Y_{11} + .128Y_{10} + .512F_{10}$

$F_{13} = .2Y_{12} + .16Y_{11} + .128Y_{10} + .512(.2Y_9 + .8F_9) = .2Y_{12} + .16Y_{11} + .128Y_{10} + .1024Y_9 + .4096F_9$

$F_{13} = .2Y_{12} + .16Y_{11} + .128Y_{10} + .1024Y_9 + .4096(.2Y_8 + .8F_8) = .2Y_{12} + .16Y_{11} + .128Y_{10} + .1024Y_9 + .08192Y_8 + .32768F_8$

b. The more recent data receives the greater weight or importance in determining the forecast. The moving averages method weights the last *n* data values equally in determining the forecast.

6. a.

Month	Y_t	3-Month Moving Averages Forecast	$(\text{Error})^2$	$\alpha = 2$ Forecast	$(\text{Error})^2$
1	80				
2	82			80.00	4.00
3	84			80.40	12.96
4	83	82.00	1.00	81.12	3.53
5	83	83.00	0.00	81.50	2.25
6	84	83.33	0.45	81.80	4.84
7	85	83.33	2.79	82.24	7.62
8	84	84.00	0.00	82.79	1.46
9	82	84.33	5.43	83.03	1.06
10	83	83.67	0.45	82.83	0.03
11	84	83.00	1.00	82.86	1.30
12	83	83.00	0.00	83.09	0.01
			11.12		39.06

MSE(3-Month) = 11.12 / 9 = 1.24
MSEα= .2) = 39.06 / 11 = 3.55
Use 3-month moving averages.

b. (83 + 84 + 83) / 3 = 83.3

7. a.

Month	Time-Series Value	3-Month Moving Average Forecast	$(\text{Error})^2$	4-Month Moving Average Forecast	$(\text{Error})^2$
1	9.5				
2	9.3				
3	9.4				
4	9.6	9.40	0.04		
5	9.8	9.43	0.14	9.45	0.12
6	9.7	9.60	0.01	9.53	0.03
7	9.8	9.70	0.01	9.63	0.03
8	10.5	9.77	0.53	9.73	0.59
9	9.9	10.00	0.01	9.95	0.00
10	9.7	10.07	0.14	9.98	0.08
11	9.6	10.03	0.18	9.97	0.14
12	9.6	9.73	0.02	9.92	0.10
			1.08		1.09

MSE(3-Month) = 1.08 / 9 = .12

MSE(4-Month) = 1.09 / 8 = .14

Use 3-Month moving averages.

b. Forecast = (9.7 + 9.6 + 9.6) / 3 = 9.63

8. a.

Month	Time-Series Value	3-Month Moving Average Forecast	$(\text{Error})^2$	$\alpha = .2$ Forecast	$(\text{Error})^2$
1	240				
2	350			240.00	12100.00
3	230			262.00	1024.00
4	260	273.33	177.69	255.60	19.36
5	280	280.00	0.00	256.48	553.19
6	320	256.67	4010.69	261.18	3459.79
7	220	286.67	4444.89	272.95	2803.70
8	310	273.33	1344.69	262.36	2269.57
9	240	283.33	1877.49	271.89	1016.97
10	310	256.67	2844.09	265.51	1979.36
11	240	286.67	2178.09	274.41	1184.05
12	230	263.33	1110.89	267.53	1408.50
			17,988.52		27,818.49

MSE(3-Month) = 17,988.52 / 9 = 1998.72
MSE(α = .2) = 27,818.49 / 11 = 2528.95

Based on the above MSE values, the 3-month moving averages appears better. However, exponential smoothing was penalized by including month 2 which was difficult for any method to forecast. Using only the errors for months 4 to 12, the MSE for exponential smoothing is:

MSE(α = .2) = 14,694.49 / 9 = 1632.72

Thus, exponential smoothing was better considering months 4 to 12.

b. Using exponential smoothing,

$$F_{13} = \alpha Y_{12} + (1 - \alpha)F_{12} = .20(230) + .80(267.53) = 260$$

9. a. Smoothing constant = .3.

Month t	Time-Series Value Y_t	Forecast F_t	Forecast Error $Y_t - F_t$	Squared Error $(Y_t - F_t)^2$
1	105			
2	135	105.00	30.00	900.00
3	120	114.00	6.00	36.00
4	105	115.80	-10.80	116.64
5	90	112.56	-22.56	508.95
6	120	105.79	14.21	201.92
7	145	110.05	34.95	1221.50
8	140	120.54	19.46	378.69
9	100	126.38	-26.38	695.90
10	80	118.46	-38.46	1479.17
11	100	106.92	-6.92	47.89
12	110	104.85	5.15	26.52
			Total	5613.18

MSE = 5613.18 / 11 = 510.29
Forecast for month 13: F_{13} = .3(110) + .7(104.85) = 106.4

b. Smoothing constant = .5

Month t	Time-Series Value Y_t	Forecast F_t	Forecast Error $Y_t - F_t$	Squared Error $(Y_t - F_t)^2$
1	105			
2	135	105	30.00	900.00
3	120	.5(135) + .5(105) = 120	0.00	0.00
4	105	.5(120) + .5(120) = 120	-15.00	225.00
5	90	.5(105) + .5(120) = 112.50	-22.50	506.25
6	120	.5(90) + .5(112.5) = 101.25	18.75	351.56
7	145	.5(120) + .5(101.25) =110.63	34.37	1181.30
8	140	.5(145) + .5(110.63) = 127.81	12.19	148.60
9	100	.5(140) + .5(127.81) = 133.91	-33.91	1149.89
10	80	.5(100) + .5(133.91) = 116.95	-36.95	1365.30
11	100	.5(80) + .5(116.95) = 98.48	1.52	2.31
12	110	.5(100) + .5(98.48) = 99.24	10.76	115.78
				5945.99

MSE = 5945.99 / 11 = 540.55
Forecast for month 13: F_{13} = .5(110) + .5(99.24) = 104.62

Conclusion: a smoothing constant of .3 is better than a smoothing constant of .5 since the MSE is less for 0.3.

10. a./b.

Week	Time-Series Value	$\alpha = .2$ Forecast	$(\text{Error})^2$	$\alpha = .3$ Forecast	$(\text{Error})^2$
1	7.35				
2	7.40	7.35	.0025	7.35	.0025
3	7.55	7.36	.0361	7.36	.0361
4	7.56	7.40	.0256	7.42	.0196
5	7.60	7.43	.0289	7.46	.0196
6	7.52	7.46	.0036	7.50	.0004
7	7.52	7.48	.0016	7.51	.0001
8	7.70	7.48	.0484	7.51	.0361
9	7.62	7.53	.0081	7.57	.0025
10	7.55	7.55	.0000	7.58	.0009
			.1548		.1178

c. MSE($\alpha = .2$) = .1548 / 9 = .0172

MSE($\alpha = .3$) = .1178 / 9 = .0131
Use $\alpha = .3$.

$F_{11} = .3Y_{10} + .7F_{10} = .3(7.55) + .7(7.58) = 7.57$

11. a.

Method	Forecast	MSE
3-Quarter	80.73	2.53
4-Quarter	80.55	2.81

The 3-quarter moving average forecast is better because it has the smallest MSE.

b.

Method	Forecast	MSE
$\alpha = .4$	80.40	2.40
$\alpha = .5$	80.57	2.01

The $\alpha = .5$ smoothing constant is better because it has the smallest MSE.

c. The $\alpha = .5$ is better because it has the smallest MSE.

12. The following values are needed to compute the slope and intercept:

$$\sum t = 15 \quad \sum t^2 = 15 \quad \sum Y_t = 55 \quad \sum tY_t = 186$$

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{186 - (15)(55)/5}{55 - (15)^2/5} = 2.1$$

$$b_0 = \bar{Y} - b_1\bar{t} = 11 - 2.1(3) = 4.7$$

$$T_t = 4.7 + 2.1t$$

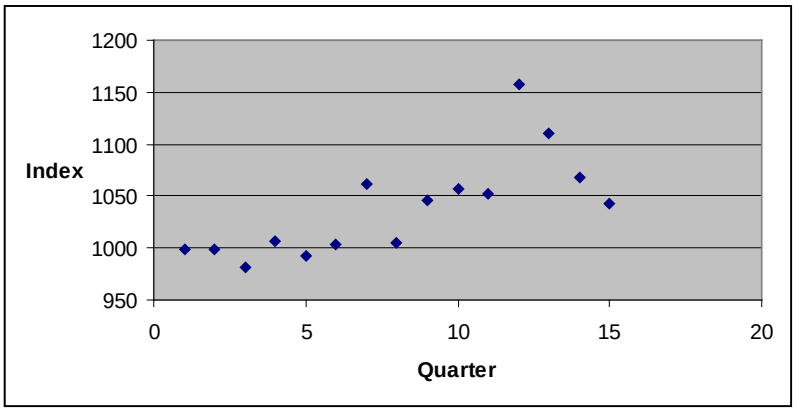
Forecast: $T_6 = 4.7 + 2.1(6) = 17.3$

13. a.

Method	Forecast	MSE
3-Quarter	1074	28.44
4-Quarter	1094.75	10.29

The 4-quarter moving average forecast is better because it has the smaller MSE.

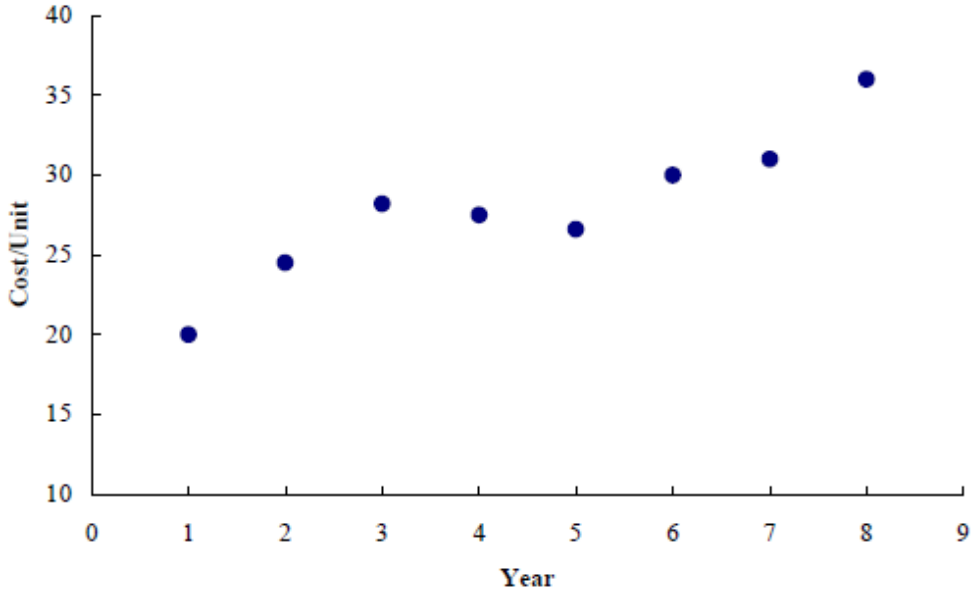
b.



.

To be able to apply the simple exponential smoothing model we need to assume the data are stationary. From the plot above, this does not appear to be the case.

14. a.



A linear trend appears to be reasonable.

b. The following values are needed to compute the slope and intercept:

$\sum t = 36 \quad \sum t^2 = 204 \quad \sum Y_t = 223.8 \quad \sum tY_t = 1081.6$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{1081.6 - (36)(223.8)/8}{204 - (36)^2/8} = 1.7738$$

Computation of intercept:

$$b_0 = \bar{Y} - b_1\bar{t} = 27.975 - 1.7738(4.5) = 19.993$$

Equation for linear trend: $T_t = 19.993 + 1.774\,t$

Conclusion: The firm has been realizing an average cost increase of \$1.77 per unit per year.

15. a.

Rural group:

$\sum t = 15 \quad \sum t^2 = 55 \quad \sum Y_t = 58 \quad \sum tY_t = 226$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{226 - (15)(58)/5}{55 - (15)^2/5} = 5.2$$

Computation of intercept:

$$b_0 = \bar{Y} - b_1\bar{t} = 11.6 - 5.2(3) = -4$$

Equation for linear trend: $T_t = -4 + 5.2\,t$

Urban group:

$\sum t = 15 \quad \sum t^2 = 55 \quad \sum Y_t = 115 \quad \sum tY_t = 414$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{414 - (15)(115)/5}{55 - (15)^2/5} = 6.9$$

Computation of intercept:

$$b_0 = \bar{Y} - b_1\bar{t} = 23 - 6.9(3) = 2.3$$

Equation for linear trend: $T_t = 2.3 + 6.9\,t$

Suburban group:

$\sum t = 15 \quad \sum t^2 = 55 \quad \sum Y_t = 118 \quad \sum tY_t = 428$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{428 - (15)(118)/5}{55 - (15)^2/5} = 7.4$$

Computation of intercept:

$$b_0 = \bar{Y} - b_1\bar{t} = 23.6 - 7.4(3) = 1.4$$

Equation for linear trend: $T_t = 1.4 + 7.4\,t$

b. Over the past five years the percentage increase is as follows:

Rural	5.2%
Urban	6.9%
Suburban	7.4%

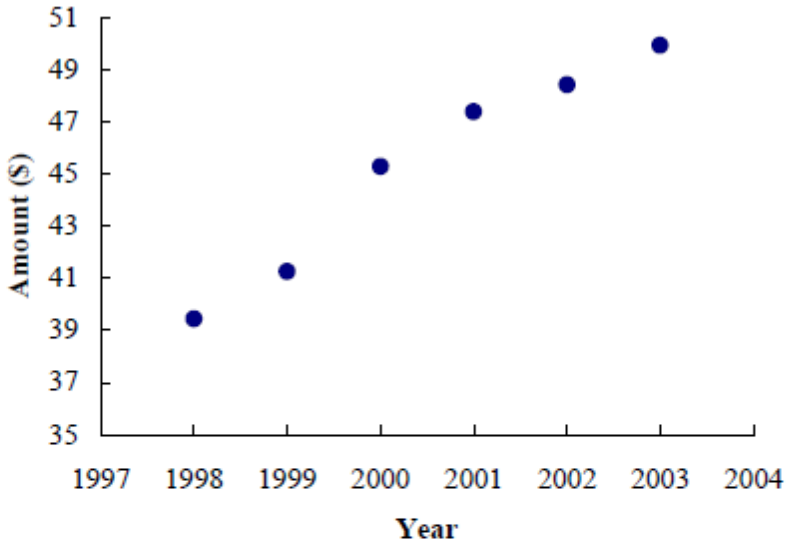
c. Forecast for Rural: $T_t = -4 + 5.2\,t = -4 + 5.2(6) = 27.2\%$

Forecast for Urban: $T_t = 2.3 + 6.9\,t = 2.3 + 6.9(6) = 43.7\%$

Forecast for Suburban: $T_t = 1.4 + 7.4\,t = 1.4 + 7.4(6) = 45.8\%$

16.

a.



The graph shows a linear trend.

b. The following values are needed to compute the slope and intercept:

$$\sum t = 21 \qquad \sum t^2 = 91 \qquad \sum Y_t = 271.62 \qquad \sum tY_t = 988.66$$

Computation of slope:

$$b_1 = \frac{\sum tY_t - \left(\sum t \sum Y_t\right)/n}{\sum t^2 - \left(\sum t\right)^2/n} = \frac{988.66 - (21)(271.62)/6}{91 - (21)^2/6} = 2.1709$$

Computation of intercept:

$$b_0 = \bar{Y} - b_1\bar{t} = 45.27 + 2.1709(3.5) = 37.6719$$

Equation for linear trend: $T_t = 37.67 + 2.17t$

c. $T_t = 37.67 + 2.17t = 37.67 + 2.17(7) = 52.86$

17.

a. The following values are needed to compute the slope and intercept:

$$\sum t = 55 \quad \sum t^2 = 385 \quad \sum Y_t = 41841 \quad \sum tY_t = 262,923$$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{262,923 - (55)(41,841)/10}{385 - (55)^2/10} = 397.545$$

Computation of intercept:

$$b_0 = \bar{Y} - b_1\bar{t} = 4184.1 - 397.545(5.5) = 1997.6$$

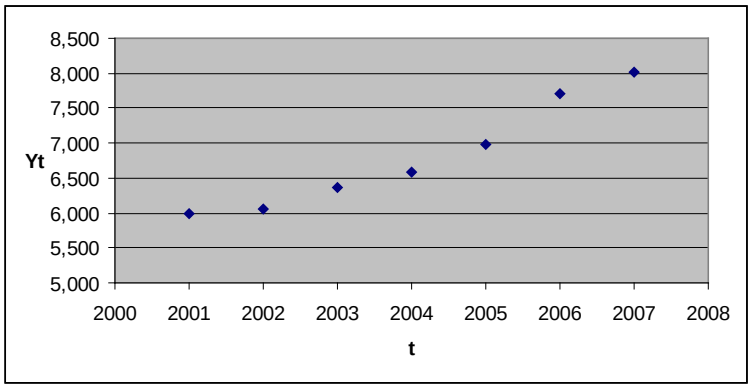
Equation for linear trend: $T_t = 1997.6 + 397.545\ t$

b. $T_{11} = 1997.6 + 397.545(11) = 6371$

$T_{12} = 1997.6 + 397.545(12) = 6768$

18. A linear trend model is not appropriate. A nonlinear model would provide a better approximation.

19. a. From the graph of the data below, trend projection does indeed look appropriate.



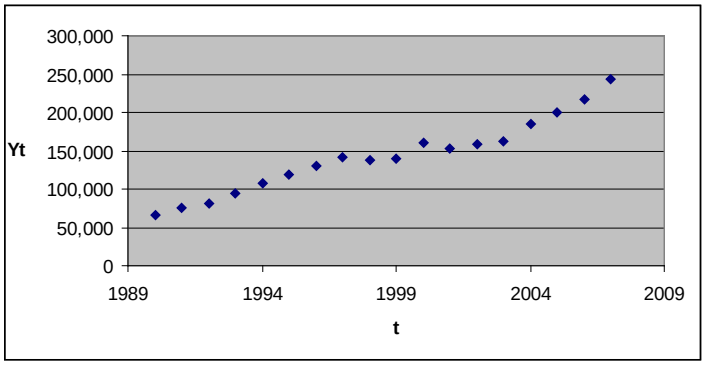
From EXCEL

Equation for linear trend: T_t = -706181+ 355.79 *t*

Forecast: T_{2008} = -706181+ 355.79 (*2008)* = 8245.32

T_{2009} = -706181+ 355.79 (*2009)* = 8601.11

20. a. From the graph below, a linear trend is evidently present.



b. From EXCEL

Equation for linear trend: T_t = 67473 + 8873.2 x where x = Year - 1990

Forecast: T_{2008} = 67473 + 8873.2 (18) = 227190.6

T_{2009} = 67473 + 8873.2 (19) = 236063.8

T_{2010} = 67473 + 8873.2 (20) = 244937

21. a. The following values are needed to compute the slope and intercept:

$\sum t = 55 \quad \sum t^2 = 385 \quad \sum Y_t = 41841 \quad \sum tY_t = 262,923$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{262,923 - (55)(41,841)/10}{385 - (55)^2/10} = 397.545$$

Computation of intercept:

$b_0 = \bar{Y} - b_1\bar{t} =$ 4184.1 - 397.545(5.5) = 1997.6

Equation for linear trend: T_t = 1997.6 + 397.545 t

b. T_{11} = 1997.6 + 397.545(11) = 6371

T_{12} = 1997.6 + 397.545(12) = 6768

22 a. The time series plot shows a horizontal pattern, but there is a seasonal pattern in the data; for instance, in each year the lowest value occurs in quarter 2 and the highest value occurs in quarter 4

b. A portion of the Minitab regression output is shown;

The regression equation is

Value = 77.0 – 10.0 Qtr1 – 30.0

Qtr2 – 20.0 Qtr3

c. The quarterly forecasts for next year are as follows:

Quarter 1 forecast = 77.0 – 10.0(1) – 30.0(0) – 20.0(0) = 67

Quarter 2 forecast = 77.0 – 10.0(0) – 30.0(1) – 20.0(0) = 47

Quarter 3 forecast = 77.0 – 10.0(0) – 30.0(0) –20.0(1) = 57

Quarter 4 forecast = 77.0 – 10.0(0) – 30.0(0) – 20.0(0) = 77

23 a.

Year	Quarter	Y_t	Four-Quarter Moving Average	Centred Moving Average
1	1	4		
	2	2		
			3.50	
	3	3		3.750
			4.00	
	4	5		4.125
			4.25	
2	1	6		4.500
			4.75	
	2	3		5.000
			5.25	
	3	5		5.375
			5.50	
	4	7		5.875
			6.25	
3	1	7		6.375
			6.50	
	2	6		6.625
			6.75	
	3	6		
	4	8		

b.

Year	Quarter	Y_t	Centred Moving Average	Seasonal-Irregular Component
1	1	4		
	2	2		
	3	3	3.750	0.8000
	4	5	4.125	1.2121
2	1	6	4.500	1.3333
	2	3	5.000	0.6000
	3	5	5.375	0.9302
	4	7	5.875	1.1915
3	1	7	6.375	1.0980
	2	6	6.625	0.9057
	3	6		
	4	8		

Quarter	Seasonal-Irregular Component Values	Seasonal Index	Adjusted Seasonal Index
1	1.3333,1.0980	1.2157	1.2050
2	.60000,.9057	0.7529	0.7463
3	.80000,.9032	0.8651	0.8675
4	1.2121,1.1915	1.2018	1.1912
		4.0355	

Note: Adjustment for seasonal index = 4.000 / 4.0355 = 0.9912

24. a. Four quarter moving averages beginning with

(1690 + 940 + 2625 + 2500) / 4 = 1938.75

Other moving averages are

1966.25	2002.50
1956.25	2052.50
2025.00	2060.00
1990.00	2123.75

b.

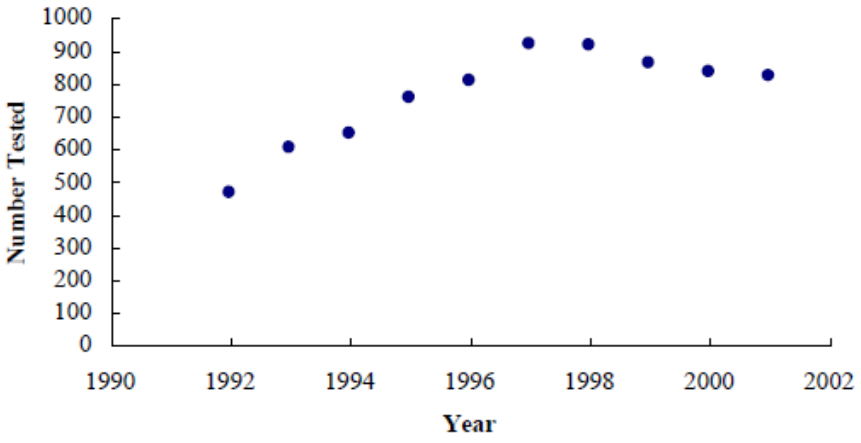
Quarter	Seasonal-Irregular Component Values		Seasonal Index	Adjusted Seasonal Index
1	0.904	0.900	0.9020	0.900
2	0.448	0.526	0.4970	0.486
3	1.344	1.453	1.3985	1.396
4	1.275	1.164	1.2195	1.217
			4.0070	

Note: Adjustment for seasonal index = 4.000 / 4.007 = 0.9983

c. The largest seasonal effect is in the third quarter which corresponds to the back-to-school demand during July, August, and September of each year.

25.

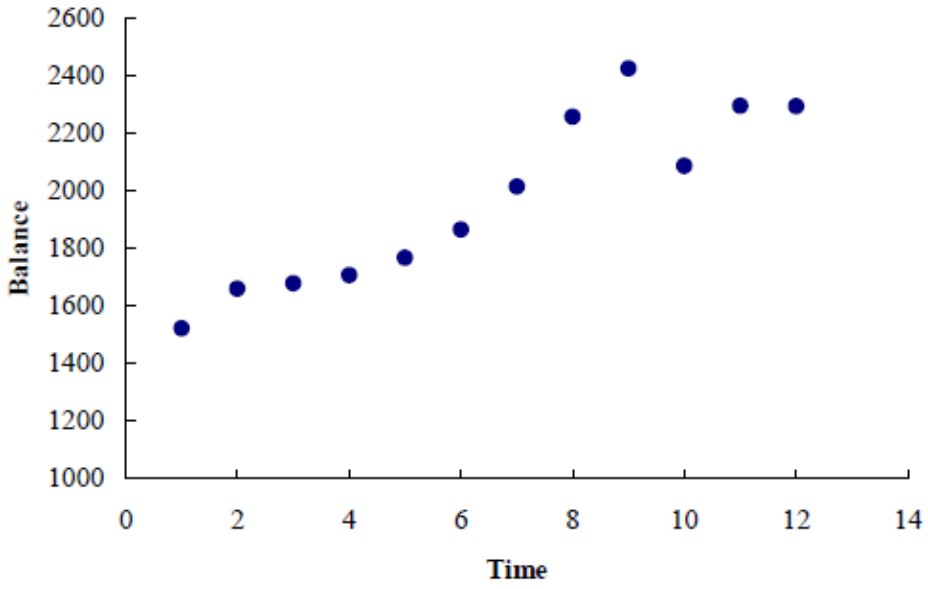
A graph of the time series is shown below:



A linear trend does not seem appropriate. The plot indicates some type of curvilinear relationship over time such as $T_t = b_0 + b_1 t + b_2 t^2$.

26.

a.



The graph shows a linear trend.

b. The following values are needed to compute the slope and intercept:

$$\sum t = 78 \qquad \sum t^2 = 650 \qquad \sum Y_t = 23,577.54 \qquad \sum tY_t = 164,400$$

Computation of slope:

$$b_1 = \frac{\sum tY_t - \left(\sum t \sum Y_t\right)/n}{\sum t^2 - \left(\sum t\right)^2/n} = \frac{164,400 - (78)(23,577.54)/12}{650 - (78)^2/12} = 77.944$$

Computation of intercept:

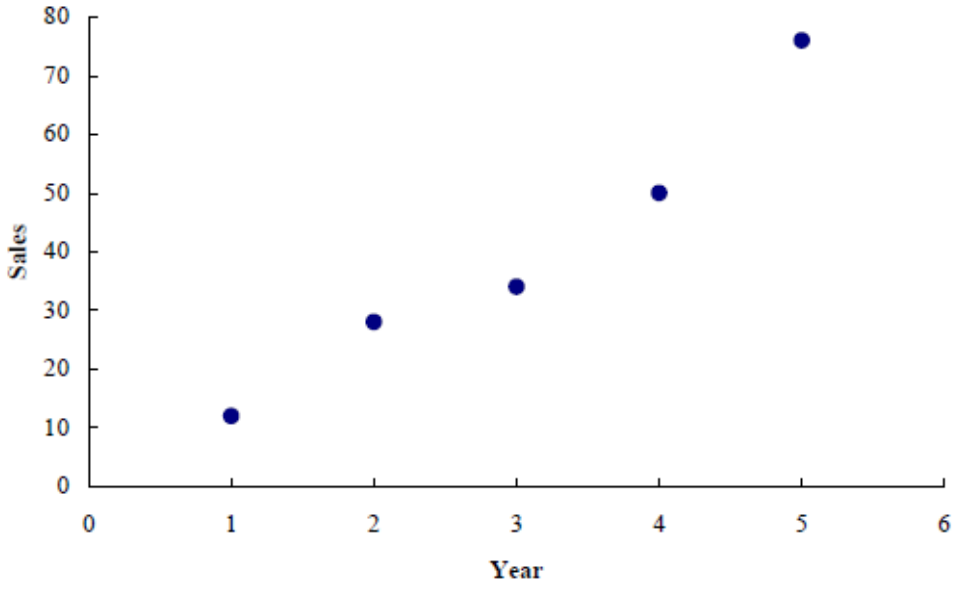
$$b_0 = \bar{Y} - b_1\bar{t} = 1964.795 - 77.944(6.5) = 1458.159$$

Equation for linear trend: $T_t = 1458.159 + 77.944\,t$

c. $T_t = 1458.159 + 1964.795\,t = 1458.159 + 77.944\,(13) = 2471.43$

27.

a.



A linear trend appears to be present.

b The following values are needed to compute the slope and intercept:

$\sum t = 15 \quad \sum t^2 = 55 \quad \sum Y_t = 200 \quad \sum tY_t = 750$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{750 - (15)(200)/5}{55 - (15)^2/5} = 15$$

Computation of intercept:

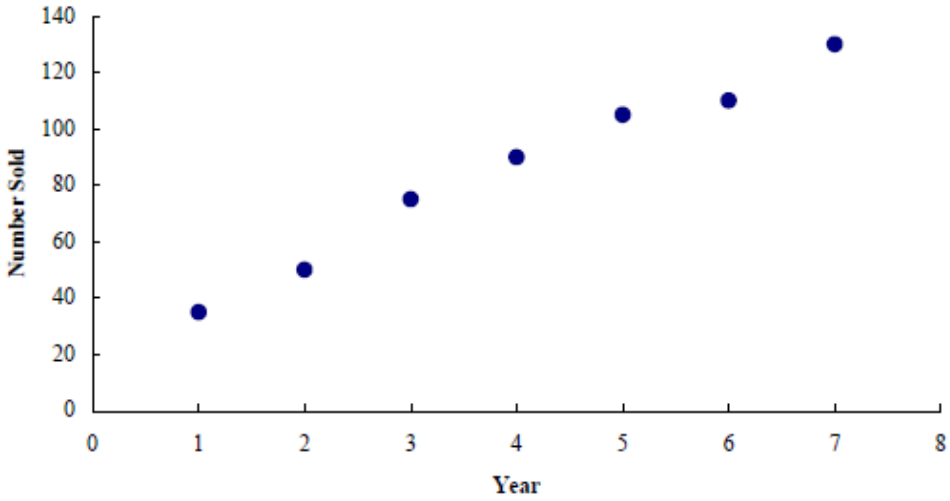
$b_0 = \overline{Y} - b_1\overline{t} = 40 - 15(3) = -5$

Equation for linear trend: $T_t = -5 + 15t$

Conclusion: average increase in sales is 15 units per year

28.

a.



A linear trend appears to be present.

b The following values are needed to compute the slope and intercept:

$\sum t = 28 \quad \sum t^2 = 140 \quad \sum Y_t = 595 \quad \sum tY_t = 2815$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{2815 - (28)(595)/7}{140 - (28)^2/7} = 15.5357$$

Computation of intercept:

$b_0 = \bar{Y} - b_1\bar{t} = 85 - 15.5357(4) = 22.857$

Equation for linear trend: $T_t = 22.857 + 15.536t$

c. Forecast: $T_8 = 22.857 + 15.536(8) = 147.15$

29 Multiplicative Model

Data sales
Length 17
NMissing 0

Fitted Trend Equation

Yt = 204.59 + 0.762*t

Seasonal Indices

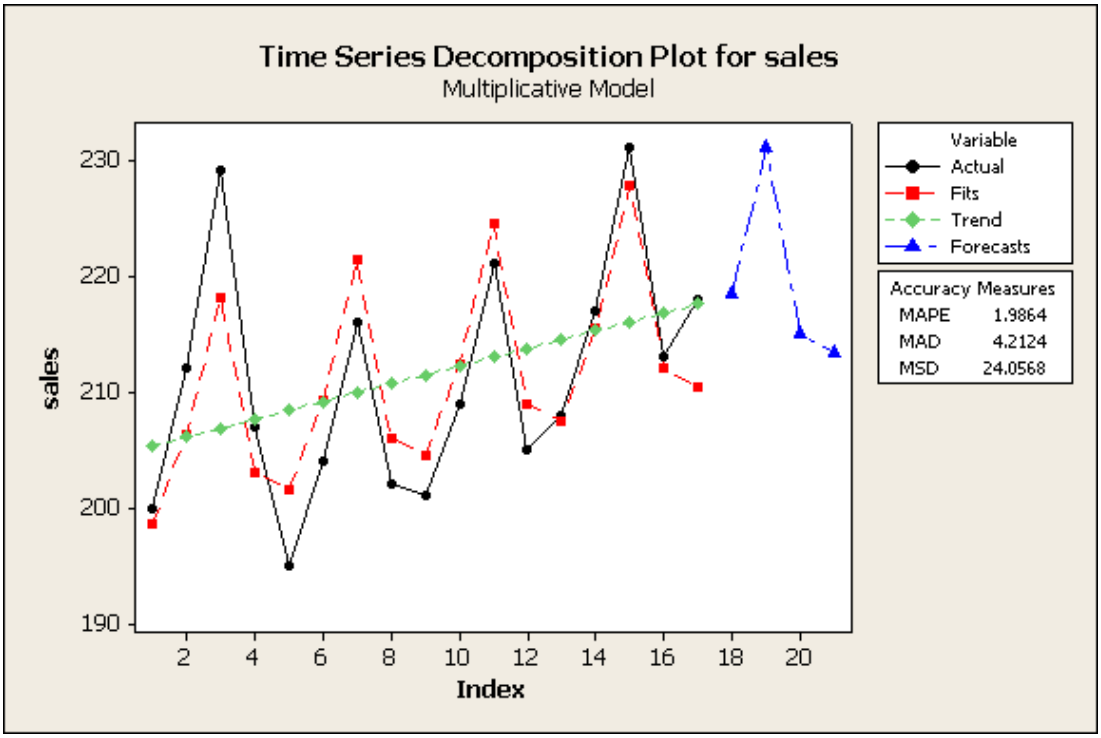
Period	Index
1	0.96709
2	1.00077
3	1.05435
4	0.97779

Accuracy Measures

MAPE 1.9864
MAD 4.2124
MSD 24.0568

Forecasts

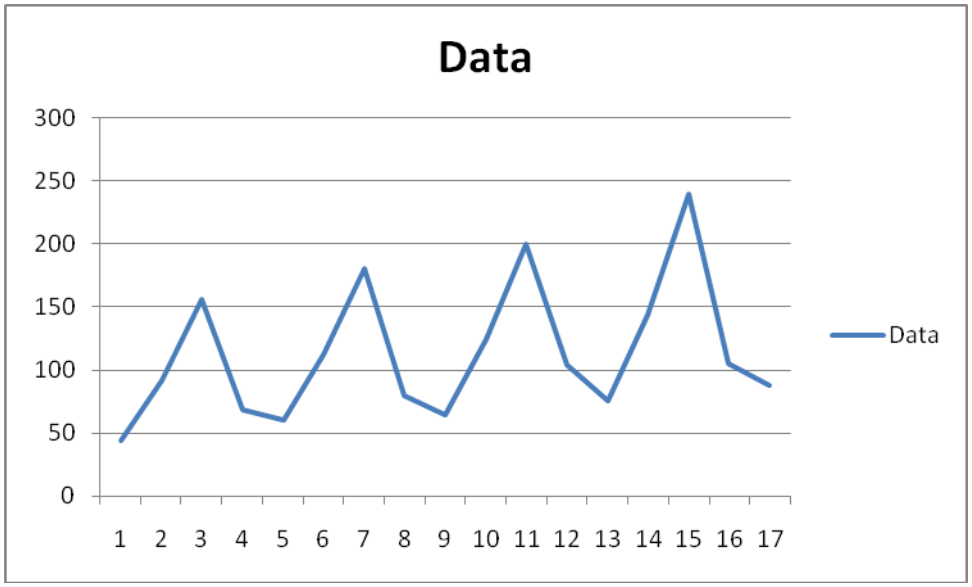
Period	Forecast
18	218.476
19	230.977
20	214.951
21	213.336



30 **See EXCEL Summary below (The forecasts correspond with the last four points of the graph following.) The modelling undertaken seems both appropriate and successful.**

Multiplicative Model									
		Data	MA	Centred MA		Seasonal I	Forecast	Irregular	Deseasona
1	1	44				0.5725			76.86
	2	92				1.0405			88.42
	3	156	90.00	92.00	1.6957	1.6741	154.02	1.98	93.18
	4	68	94.00	96.50	0.7047	0.7129	68.79	-0.79	95.39
2	1	60	99.00	102.00	0.5882	0.5725	58.40	1.61	104.80
	2	112	105.00	106.50	1.0516	1.0405	110.81	1.19	107.64
	3	180	108.00	108.50	1.6590	1.6741	181.64	-1.64	107.52
	4	80	109.00	110.50	0.7240	0.7129	78.78	1.22	112.22
3	1	64	112.00	114.50	0.5590	0.5725	65.55	-1.55	111.79
	2	124	117.00	120.00	1.0333	1.0405	124.86	-0.86	119.17
	3	200	123.00	124.50	1.6064	1.6741	208.43	-8.43	119.47
	4	104	126.00			0.7129			145.88
4	1	76				0.5725			132.75
	2	144				1.0405			138
	3	239				1.6741			143
	4	106				0.7129			148
5	1	88				0.5725			153

	1	2	3	4
1			1.6957	0.7047
2	0.5882	1.0516	1.6590	0.7240
3	0.5590	1.0333	1.6064	
	0.5736	1.0425	1.6773	0.7143
	0.5725	1.0405	1.6741	0.7129



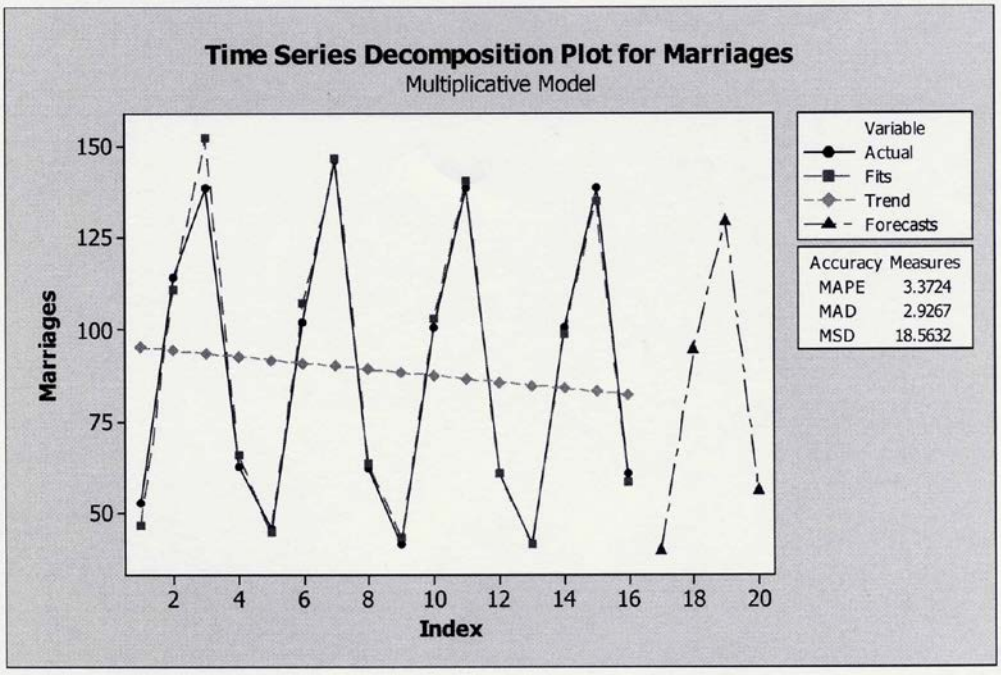
31

Solution

Time Series Decomposition for Marriages

Multiplicative Model

Data Marriages
Length 16
NMissing 0



Forecasts

7

Period	Forecast
17	39.778
18	94.647
19	129.284
20	55.913

Supplementary Exercises:

32. Moving averages often are used to identify movements in stock prices. Daily closing prices (in dollars per share) for SanDisk for August 16, 2002, through September 3, 2002, follow (http://www.finance.yahoo.com).

Day	Price ($)	Day	Price ($)
August 16	14.45	August 26	16.45
August 19	15.75	August 27	15.60
August 20	16.45	August 28	15.09
August 21	17.40	August 29	16.42
August 22	17.32	August 30	16.21
August 23	15.96	September 3	15.22

a. Use a five-month moving average to smooth the time series. Forecast the closing price for September 4, 2002.

b. Use a four-month weighted moving average to smooth the time series. Use a weight of 0.4 for the most recent period, 0.3 for the next period back, 0.2 for the third period back, and 0.1 for the fourth period back. Forecast the closing price for September 4, 2002.

c. Use exponential smoothing with a smoothing constant of $\alpha = 0.7$ to smooth the time series. Forecast the closing price for September 4, 2002.

d. Which of the three methods do you prefer? Why?

33. A chain of grocery stores noted the weekly demand (in cases) reported in the following table for a particular brand of automatic dishwasher detergent. Use exponential smoothing with $\alpha = 0.2$ to develop a forecast for week 11.

Week	Demand	Week	Demand
1	22	6	24
2	18	7	20
3	23	8	19
4	21	9	18
5	17	10	21

34. European Dairies supplies milk to several independent grocers throughout the Netherlands. Managers at European Dairies want to develop a forecast of the number of half-litres of milk sold per week. Sales data for the past 12 weeks follow.

Week	Sales	Week	Sales
1	2750	7	3300
2	3100	8	3100
3	3250	9	2950
4	2800	10	3000
5	2900	11	3200
6	3050	12	3150

Use exponential smoothing with $\alpha = 0.4$ to develop a forecast of demand for week 13.

35. The Garden Avenue Seven sells tapes of its musical performances. The following table reports sales (in units) for the past 18 months. The group's manager wants an accurate method for forecasting future sales.

Month	Sales	Month	Sales	Month	Sales
1	293	7	381	13	549
2	283	8	431	14	544
3	322	9	424	15	601
4	355	10	433	16	587
5	346	11	470	17	644
6	379	12	481	18	660

a. Use exponential smoothing with $\alpha = 0.3$, 0.4, and 0.5. Which value of α provides the best forecasts?

b. Use trend projection to provide a forecast. What is the value of MSE?

c. Which method of forecasting would you recommend to the manager? Why?

36. The Mayer Department Store in Cologne, Germany is trying to determine the amount of sales lost while it was shut down during July and August because of damage caused by floods by River Rhine. Sales data for January through June follow.

Month	**Sales (€1000s)**	**Month**	**Sales (€1000s)**
January	185.72	April	210.36
February	167.84	May	255.57
March	205.11	June	261.19

a. Use exponential smoothing, with $\alpha = 0.4$, to develop a forecast for July and August. (*Hint:* Use the forecast for July as the actual sales in July in developing the August forecast.) Comment on the use of exponential smoothing for forecasts more than one period into the future.

b. Use trend projection to forecast sales for July and August.

c. Mayer's insurance company proposed a settlement based on lost sales of €240,000 in July and August. Is this amount fair? If not, what amount would you recommend as a counteroffer?

37. Canton Produits is a service firm that employs approximately 100 individuals. Managers of Canton Produits are concerned about meeting monthly cash obligations and want to develop a forecast of monthly cash requirements. Because of a recent change in operating policy, only the past seven months of data are considered to be relevant. With the following historical data, use trend projection to develop a forecast of cash requirements for each of the next two months.

Month	1	2	3	4	5	6	7
Cash Required (€1000s)	205	212	218	224	230	240	246

38. The Costello Music Company has been in business for five years. During that time, sales of electric organs increased from 12 units in the first year to 76 units in the most recent year. Seamus Costello, the firm's owner, wants to develop a forecast of organ sales for the coming year. The historical data follow.

Year	1	2	3	4	5
Sales	12	28	34	50	76

a. Show a graph of this time series. Does a linear trend appear to be present?

b. Develop the equation for the linear trend component for the time series. What is the average increase in sales that the firm has been realizing per year?

39. Arno Marina has been an authorized dealer for C&D marine radios for the past seven years. The following table reports the number of radios sold each year.

Year	1	2	3	4	5	6	7
Number Sold	35	50	75	90	105	110	130

a. Show a graph of this time series. Does a linear trend appear to be present?

b. Develop the equation for the linear trend component of the time series.

c. Use the linear trend developed in part (b) to develop a forecast for annual sales in year 8.

40. Refer to the Arno Marina problem in exercise 26. Suppose the quarterly sales values for the seven years of historical data are as follow.

Year	**Quarter 1**	**Quarter 2**	**Quarter 3**	**Quarter 4**	**Total Yearly Sales**
1	6	15	10	4	35
2	10	18	15	7	50
3	14	26	23	12	75
4	19	28	25	18	90
5	22	34	28	21	105
6	24	36	30	20	110
7	28	40	35	27	130

a. Show the four-quarter moving average values for this time series. Plot both the original time series and the moving average series on the same graph.

b. Compute the seasonal indices for the four quarters.

c. When does Arno Marina experience the largest seasonal effect? Does this result seem reasonable? Explain.

41. Consider the Costello Music Company problem in exercise 25. The quarterly sales data follow.

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4	Total Yearly Sales
1	4	2	1	5	12
2	6	4	4	14	28
3	10	3	5	16	34
4	12	9	7	22	50
5	18	10	13	35	76

a. Compute the seasonal indices for the four quarters.

b. When does Costello Music experience the largest seasonal effect? Does this result appear reasonable? Explain.

42. Refer to the Arno Marina data in exercise 26.

a. Deseasonalize the data and use the deseasonalized time series to identify the trend.

b. Use the results of part (a) to develop a quarterly forecast for next year based on trend.

c. Use the seasonal indices developed in exercise 27 to adjust the forecasts developed in part (b) to account for the effect of season.

43. Consider the Costello Music Company time series in exercise 25.

a. Deseasonalize the data and use the deseasonalized time series to identify the trend.

b. Use the results of part (a) to develop a quarterly forecast for next year based on trend.

c. Use the seasonal indices developed in exercise 28 to adjust the forecasts developed in part (b) to account for the effect of season.

44. The table below shows the number of passengers per quarter (in thousands) who flew with a charter airline during the years 2003-2005 and the first quarter of 2006.

Year	Quarter	Passengers	Year	Quarter	Passengers
2003	1	44	2005	1	64
	2	92		2	124
	3	156		3	200
	4	68		4	104
2004	1	60	2006	1	76
	2	112			
	3	180			
	4	80			

a. Derive a multiplicative model for the data and use it to estimate the next three observations in the series.
b. Graph your results. How would you rate the success of your modelling?

45. A company producing torches is negotiating with a company for the supply of torch bulbs. The bulb company therefore needs to plan its production to meet the needs of the torch company and thus uses that company's quarterly sales figures over the past three years to forecast future demand.
The sales figures are as follows:

	Quarterly sales figures (000's)			
Year	Ql	Q2	Q3	Q4
1	349.4	295.5	196.9	389.3
2	447.5	418	324.1	456.4
3	550.6	528.6	415.2	615.3

a. Using a multiplicative model, estimate trend values and seasonal indices for the series. Fit a least squares regression line to the trend values. Then use this trend regression line and your estimates of the seasonal variation factors to forecast future demand for the four quarters of year 4 for torches.

b. Graph your results and hence comment on the quality of your modelling.

46. Data on road casualties in Great Britain (Wizniewski, 2002) for children (aged under sixteen) by quarter over four years are as follows:

Year	Ql	Q2	Q3	Q4
1	8853	12107	13233	10642
2	9338	12138	12527	10543
3	9121	12019	12109	10196
4	8971	11344	12053	9683

a. Decompose the series into trend, seasonal and random components (using a multiplicative model). Hence derive quarterly forecasts of road casualties for children in Britain in year 5.

b. By plotting actual and estimated results by quarter make a judgment on the effectiveness of your modelling.

Supplementary Exercises Solutions:

32. a.

Day	Time-Series Value	5-Day Moving Average Forecast	Forecast Error	$(\text{Error})^2$
1	14.45			
2	15.75			
3	16.45			
4	17.40			
5	17.32			
6	15.96	16.27	-0.31	0.10
7	16.45	16.58	-0.13	0.02
8	15.60	16.72	-1.12	1.25
9	15.09	16.55	-1.46	2.12
10	16.42	16.08	0.34	0.11
11	16.21	15.90	0.31	0.09
12	15.22	15.95	-0.73	0.54

Note: MSE = 4.23/7 = 0.60

Forecast for September 4 is (15.60 + 15.09 + 16.42 + 16.21 + 15.22)/5 = 15.71

b. The weighted moving average forecasts for days 5-12 are 16.49, 17.01, 16.71, 16.57, 16.10, 15.60, 15.09, 16.42, 16.21 and 15.22

Note: MSE = 5.21/8 = 0.65

Forecast for September 4 is 0.1(15.09) + 0.2(16.42) + 0.3(16.21) + 0.4(15.22) = 15.74

c. The exponential smoothing forecasts for days 2-12 are 14.45, 15.36, 16.12, 17.02, 17.23, 16.34, 16.42, 15.85, 15.32, 16.09 and 16.17

Note: MSE = 9.57/11 = 0.87

Forecast for September 4 is 0.7(15.22) + 0.3(16.17) = 15.51

d.

Method	MSE
Moving Averages	0.60
Weighted Moving Average	0.65
Exponential Smoothing	0.87

Moving Averages is the best of the three approaches because it has the smallest MSE.

33.

Week t	Time-Series Value Y_t	Forecast F_t	Forecast Error $Y_t - F_t$	Squared Error $(Y_t - F_t)^2$
1	22			
2	18	22.00	-4.00	16.00
3	23	21.20	1.80	3.24
4	21	21.56	-0.56	0.31
5	17	21.45	-4.45	19.80
6	24	20.56	3.44	11.83
7	20	21.25	-1.25	1.56
8	19	21.00	-2.00	4.00
9	18	20.60	-2.60	6.76
10	21	20.08	0.92	0.85
			Total	64.35

MSE = 64.35 / 9 = 7.15

Forecast for week 11:

F_{11} = 0.2(21) + 0.8(20.08) = 20.26

34.

t	Y_t	F_t	$Y_t - F_t$	$(Y_t - F_t)^2$
1	2,750			
2	3,100	2,750.00	350.00	122,500.00
3	3,250	2,890.00	360.00	129,600.00
4	2,800	3,034.00	-234.00	54,756.00
5	2,900	2,940.40	-40.40	1,632.16
6	3,050	2,924.24	125.76	15,815.58
7	3,300	2,974.54	325.46	105,924.21
8	3,100	3,104.73	-4.73	22.37
9	2,950	3,102.84	-152.84	23,260.07
10	3,000	3,041.70	-41.70	1,738.89
11	3,200	3,025.02	174.98	30,618.00
12	3,150	3,095.01	54.99	3,023.90
			Total:	488,991.18

MSE = 488,991.18 / 11 = 44,453.74

Forecast for week 13: F_{13} = 0.4(3,150) + 0.6(3,095.01) = 3,117.01

35. a.

Smoothing Constant	MSE
α = .3	4,492.37
α = .4	2,964.67
α = .5	2,160.31

The α = .5 smoothing constant is better because it has the smallest MSE.

b. $T_t = 244.778 + 22.088t$

MSE = 357.81

c. Trend projection provides much better forecasts because it has the smallest MSE. The reason MSE is smaller for trend projection is that sales are increasing over time; as a result, exponential smoothing continuously underestimates the value of sales. If you look at the forecast errors for exponential smoothing you will see that the forecast errors are positive for periods 2 through 18.

36. a. Forecast for July is 236.97

Forecast for August, using forecast for July as the actual sales in July, is 236.97.

Exponential smoothing provides the same forecast for every period in the future. This is why it is not usually recommended for long-term forecasting.

b. $T_t = 149.719 + 18.451t$

Forecast for July is 278.88

Forecast for August is 297.33

c. The proposed settlement is not fair since it does not account for the upward trend in sales. Based upon trend projection, the settlement should be based on forecasted lost sales of $278,880 in July and $297,330 in August.

37. The following values are needed to compute the slope and intercept:

$$\sum t = 28 \quad \sum t^2 = 140 \quad \sum Y_t = 1575 \quad \sum tY_t = 6491$$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{6491 - (28)(1575)/7}{140 - (28)^2/7} = 6.8214$$

Computation of intercept:

$$b_0 = \bar{Y} - b_1\bar{t} = 225 - 6.8214(4) = 197.714$$

Equation for linear trend: $T_t = 197.714 + 6.821t$

Forecast: $T_8 = 197.714 + 6.821(8) = 252.28$

$T_9 = 65.025 + 4.735(9) = 259.10$

38.

a. A graph of these data shows a linear trend.

b The following values are needed to compute the slope and intercept:

$$\sum t = 15 \quad \sum t^2 = 55 \quad \sum Y_t = 200 \quad \sum tY_t = 750$$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{750 - (15)(200)/5}{55 - (15)^2/5} = 15$$

Computation of intercept:

$b_0 = \bar{Y} - b_1\bar{t} = 40 - 15(3) = -5$

Equation for linear trend: $T_t = -5 + 15t$

Conclusion: average increase in sales is 15 units per year

39. a. Yes, a linear trend appears to exist.

b The following values are needed to compute the slope and intercept:

$\sum t = 28 \quad \sum t^2 = 140 \quad \sum Y_t = 595 \quad \sum tY_t = 2815$

Computation of slope:

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} = \frac{2815 - (28)(595)/7}{140 - (28)^2/7} = 15.5357$$

Computation of intercept:

$b_0 = \bar{Y} - b_1\bar{t} = 85 - 15.5357(4) = 22.857$

Equation for linear trend: $T_t = 22.857 + 15.536t$

c. Forecast: $T_8 = 22.857 + 15.536(8) = 147.15$

40. a.

t	Sales	Centred Moving Average	Seasonal-Irregular Component
1	6		
2	15		
3	10	9.250	1.081
4	4	10.125	0.395
5	10	11.125	0.899
6	18	12.125	1.485
7	15	13.000	1.154
8	7	14.500	0.483
9	14	16.500	0.848
10	26	18.125	1.434
11	23	19.375	1.187
12	12	20.250	0.593
13	19	20.750	0.916
14	28	21.750	1.287
15	25	22.875	1.093
16	18	24.000	0.750
17	22	25.125	0.876
18	34	25.875	1.314
19	28	26.500	1.057
20	21	27.000	0.778
21	24	27.500	0.873
22	36	27.625	1.303
23	30	28.000	1.071
24	20	29.000	0.690
25	28	30.125	0.929
26	40	31.625	1.265
27	35		
28	27		

b.

Quarter	Seasonal-Irregular Component Values		Seasonal Index
1	0.899, 0.848, 0.916, 0.876, 0.873, 0.929		0.890
2	1.485, 1.434, 1.287, 1.314, 1.303, 1.265		1.348
3	1.081, 1.154, 1.187, 1.093, 1.057, 1.071		1.107
4	0.395, 0.483, 0.593, 0.750, 0.778, 0.690		0.615
		Total	3.960

Quarter	Adjusted Seasonal Index
1	0.899
2	1.362
3	1.118
4	0.621

Note: Adjustment for seasonal index =
4.00 / 3.96 = 1.0101

c. Hudson Marine experiences the largest seasonal increase in quarter 2. Since this quarter occurs prior to the peak summer boating season, this result seems reasonable.

41. a.

t	Sales	Centred Moving Average	Seasonal-Irregular Component
1	4		
2	2		
3	1	3.250	0.308
4	5	3.750	1.333
5	6	4.375	1.371
6	4	5.875	0.681
7	4	7.500	0.533
8	14	7.875	1.778
9	10	7.875	1.270
10	3	8.250	0.364
11	5	8.750	0.571
12	16	9.750	1.641
13	12	10.750	1.116
14	9	11.750	0.766
15	7	13.250	0.528
16	22	14.125	1.558
17	18	15.000	1.200
18	10	17.375	0.576
19	13		
20	35		

Quarter	Seasonal-Irregular Component Values	Seasonal Index
1	1.371, 1.270, 1.116, 1.200	1.239
2	0.681, 0.364, 0.776, 0.576	0.597
3	0.308, 0.533, 0.571, 0.528	0.485
4	1.333, 1.778, 1.641, 1.558	1.578
	Total	3.899

Quarter	Adjusted Seasonal Index
1	1.271
2	0.613
3	0.498
4	1.619

Note: Adjustment for seasonal index = 4 / 3.899 = 1.026

b. The largest effect is in quarter 4; this seems reasonable since retail sales are generally higher during October, November, and December.

42. a. Note: To simplify the calculations the seasonal indexes calculated in problem 27 have been rounded to two decimal places.

Year	Quarter	Sales Y_t	Seasonal Factor S_t	Deseasonalized Sales $Y_t / S_t = T_tI_t$
1	1	6	0.90	6.67
	2	15	1.36	11.03
	3	10	1.12	8.93
	4	4	0.62	6.45
2	1	10	0.90	11.11
	2	18	1.36	13.24
	3	15	1.12	13.39
	4	7	0.62	11.29
3	1	14	0.90	15.56
	2	26	1.36	19.12
	3	23	1.12	20.54
	4	12	0.62	19.35
4	1	19	0.90	21.11
	2	28	1.36	20.59
	3	25	1.12	22.32
	4	18	0.62	29.03
5	1	22	0.90	24.44
	2	34	1.36	25.00
	3	28	1.12	25.00
	4	21	0.62	33.87
6	1	24	0.90	26.67
	2	36	1.36	26.47
	3	30	1.12	26.79
	4	20	0.62	32.26
7	1	28	0.90	31.11
	2	40	1.36	29.41
	3	35	1.12	31.25
	4	27	0.62	43.55

t	Y_t (deseasonalized)	tY_t	t^2
1	6.67	6.67	1
2	11.03	22.06	4
3	8.93	26.79	9
4	6.45	25.80	16
5	11.11	55.55	25
6	13.24	79.44	36
7	13.39	93.73	49
8	11.29	90.32	64
9	15.56	140.04	81
10	19.12	191.20	100
11	20.54	225.94	121
12	19.35	232.20	144
13	21.11	274.43	169
14	20.59	288.26	196
15	22.32	334.80	225
16	29.03	464.48	256
17	24.44	415.48	289
18	25.00	450.00	324
19	25.00	475.00	361
20	33.87	677.40	400
21	26.67	560.07	441
22	26.47	582.34	484
23	26.79	616.17	529
24	32.26	774.24	576
25	31.11	777.75	625
26	29.41	764.66	676
27	31.25	843.75	729
28	43.55	1,219.40	784
406	605.55	10,707.34	7,714

$\bar{t} = 14.5$ $\bar{Y} = 21.627$ $b_1 = 1.055$ $b_0 = 6.329$ $T_t = 6.329 + 1.055t$

b/c.

t	Trend Forecast
29	36.92
30	37.98
31	39.03
32	40.09

Year	Quarter	Trend Forecast	Seasonal Index	Quarterly Forecast
8	1	36.92	0.90	33.23
	2	37.98	1.36	51.65
	3	29.03	1.12	43.71
	4	40.09	0.62	24.86

43. a Note: To simplify the calculations the seasonal indexes in problem 28 have been rounded to two decimal places.

Year	Quarter	Sales Y_t	Seasonal Factor S_t	Deseasonalized Sales $Y_t / S_t = T_tI_t$
1	1	4	1.27	3.15
	2	2	0.61	3.28
	3	1	0.50	2.00
	4	5	1.62	3.09
2	1	6	1.27	4.72
	2	4	0.61	6.56
	3	4	0.50	8.00
	4	14	1.62	8.64
3	1	10	1.27	7.87
	2	3	0.61	4.92
	3	5	0.50	10.00
	4	16	1.62	9.88
4	1	12	1.27	9.45
	2	9	0.61	14.75
	3	7	0.50	14.00
	4	22	1.62	13.58
5	1	18	1.27	14.17
	2	10	0.61	16.39
	3	13	0.50	26.00
	4	35	1.62	21.60

t	Y_t (deseasonalized)	tY_t	t^2
1	3.15	3.15	1
2	3.28	6.56	4
3	2.00	6.00	9
4	3.09	12.36	16
5	4.72	23.60	25
6	6.56	39.36	36
7	8.00	56.00	49
8	8.64	69.12	64
9	7.87	70.83	81
10	4.92	49.20	100
11	10.00	110.00	121
12	9.88	118.56	144
13	9.45	122.85	169
14	14.75	206.50	196
15	14.00	210.00	225
16	13.58	217.28	256
17	14.17	240.89	289
18	16.39	295.02	324
19	26.00	494.00	361
20	21.60	432.00	400
210	202.05	2783.28	2870

$\bar{t} = 10.5 \quad \bar{Y} = 10.1025 \quad b_1 = .995 \quad b_0 = -.345 \quad T_t = -.345 + .995t$

b.

y	Trend Forecast
21	20.55
22	21.55
23	22.54
24	23.54

c.

Year	Quarter	Trend Forecast	Seasonal Index	Quarterly Forecast
6	1	20.55	1.27	26.10
	2	21.55	0.61	13.15
	3	22.54	0.50	11.27
	4	23.54	1.62	38.13

44. From EXCEL we have

Year	Quarter	Passengers Y_t	Centred Moving Average	Seasonal Irregular Component	Seasonal Factor S_t	Deseasonalised Passengers
2003	1	44			0.5759	76.40
	2	92			1.0466	87.90
	3	156	92.00	1.6957	1.6603	93.96
	4	68	96.50	0.7047	0.7172	94.81
2004	1	60	102.00	0.5882	0.5759	104.18
	2	112	106.50	1.0516	1.0466	107.01
	3	180	108.50	1.6590	1.6603	108.41
	4	80	110.50	0.7240	0.7172	111.54
2005	1	64	114.50	0.5590	0.5759	111.13
	2	124	120.00	1.0333	1.0466	118.48
	3	200	124.50	1.6064	1.6603	120.46
	4	104			0.7172	145.01
2006	1	76			0.5759	131.97

where

		Quarter				
		1	2	3	4	
	2003			1.6957	0.7047	
Year	2004	0.5882	1.0516	1.6590	0.7240	
	2005	0.5590	1.0333	1.6064		
Seasonal Index		0.5736	1.0425	1.6537	0.7143	3.9841
Adjusted Seasonal Index		0.5759	1.0466	1.6603	0.7172	4.0000

$\bar{t} = 7$ $\overline{Y} = 104.6154$ $b_1 = 4.4744$ $b_0 = 77.239$ $T_t = 77.239 + 4.4744t$

y	Trend Forecast
14	139.88
15	144.36
16	148.83

Year	Quarter	Trend Forecast	Seasonal Index	Quarterly Forecast
2006	2	139.88	1.0466	146
	3	144.36	1.6603	240
	4	148.83	0.7172	107

45. a. From EXCEL we have

Year	Quarter	Sales Y_t	Centred Moving Average	Seasonal Irregular Component	Seasonal Factor S_t	Deseasonalised Sales
1	1	349.4			1.1795	296.23
	2	295.5			1.0489	281.72
	3	196.9	320.04	0.6152	0.6959	282.94
	4	389.3	347.61	1.1199	1.0758	361.87
2	1	447.5	378.83	1.1813	1.1795	379.40
	2	418	403.11	1.0369	1.0489	398.51
	3	324.1	424.39	0.7637	0.6959	465.73
	4	456.4	451.10	1.0117	1.0758	424.24
3	1	550.6	476.31	1.1560	1.1795	466.81
	2	528.6	507.56	1.0414	1.0489	503.96
	3	415.2			0.6959	596.64
	4	615.3			1.0758	571.95

where

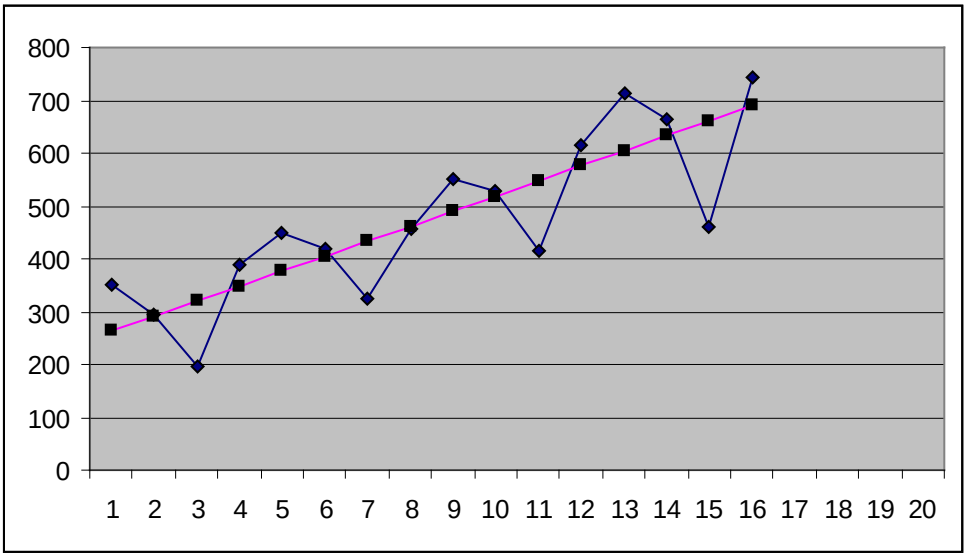
		Quarter				
		1	2	3	4	
	1			0.6152	1.1199	
Year	2	1.1813	1.0369	0.7637	1.0117	
	3	1.1560	1.0414			
Seasonal Index		1.1686	1.0392	0.6895	1.0658	3.9631
Adjusted Seasonal Index		1.1795	1.0489	0.6959	1.0758	4.0000

$\bar{t} = 6.5$ $\overline{Y} = 419.17$ $b_1 = 28.464$ $b_0 = 234.15$ $T_t = 234.15 + 28.464t$

y	Trend Forecast
13	604.18
14	632.65
15	661.11
16	689.57

Year	Quarter	Trend Forecast	Seasonal Index	Quarterly Forecast
4	1	604.18	1.1795	712.63
	2	632.65	1.0489	663.58
	3	661.11	0.6959	460.07
	4	689.57	1.0758	741.84

b. The model seems to be of high quality from the plot below.



46. From EXCEL we have

Year	Quarter	Casualties Y_t	Centred Moving Average	Seasonal Irregular Component	Seasonal Factor S_t	Deseasonalised Casualties
1	1	8853			0.8322	10638.07
	2	12107			1.0856	11152.36
	3	13233	11269.38	1.1742	1.1372	11636.48
	4	10642	11333.88	0.9390	0.945	11261.38
2	1	9338	11249.50	0.8301	0.8322	11220.86
	2	12138	11148.88	1.0887	1.0856	11180.91
	3	12527	11109.38	1.1276	1.1372	11015.65
	4	10543	11067.38	0.9526	0.945	11156.61
3	1	9121	11000.25	0.8292	0.8322	10960.11
	2	12019	10904.63	1.1022	1.0856	11071.30
	3	12109	10842.50	1.1168	1.1372	10648.08
	4	10196	10739.38	0.9494	0.945	10789.42
4	1	8971	10648.00	0.8425	0.8322	10779.86
	2	11344	10576.88	1.0725	1.0856	10449.52
	3	12053			1.1372	10598.84
	4	9683			0.945	10246.56

where

		Quarter				
		1	2	3	4	
Year	1			1.1742	0.9390	
	2	0.8301	1.0887	1.1276	0.9526	
	3	0.8292	1.1022	1.1168	0.9494	
	4	0.8425	1.0725			
Seasonal Index		0.8339	1.0878	1.1396	0.9470	4.0083
Adjusted Seasonal Index		0.8322	1.0856	1.1372	0.9450	4.0000

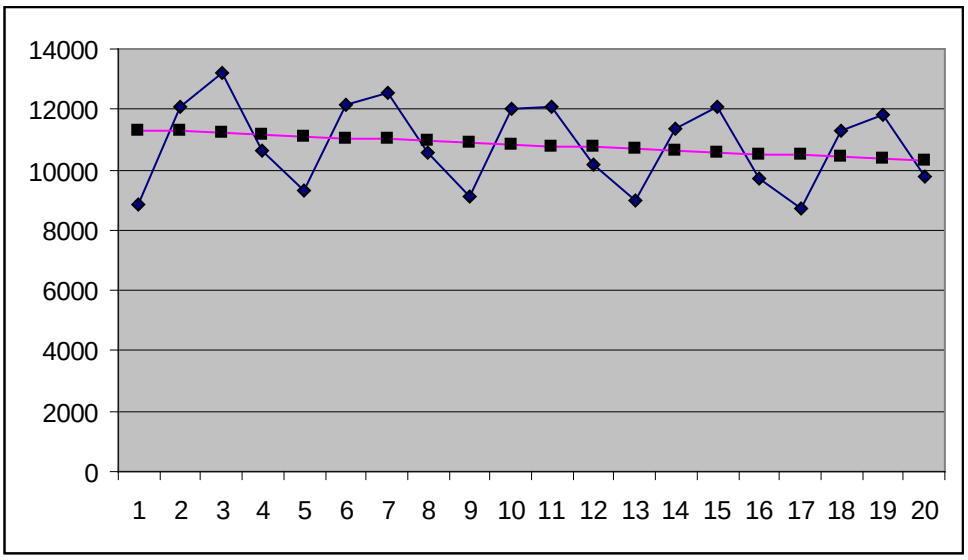
$\bar{t} = 8.5$ $\bar{Y} = 10925.38$ $b_1 = -53.19$ $b_0 = 11378$ $T_t = 11378 - 53.195t$

y	Trend Forecast
17	10473.69
18	10420.49
19	10367.30
20	10314.10

Year	Quarter	Trend Forecast	Seasonal Index	Quarterly Forecast
5	1	10473.69	0.8322	8716
	2	10420.49	1.0856	11312
	3	10367.30	1.1372	11790
	4	10314.10	0.945	9747

From the EXCEL plot below the model looks very effective.

b.



.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Eighteen

Non-parametric Methods

Textbook Exercises (1-26)
Textbook Exercise Solutions
Supplementary Exercises (27-40)
Supplementary Exercise Solutions

Textbook Exercises:

1 The following table lists the preferences indicated by ten individuals in taste tests involving two brands of a product.

Individual	Brand A versus Brand B	Individual	Brand A versus Brand B
1	+	6	+
2	+	7	–
3	+	8	+
4	–	9	–
5	+	10	+

With α = 0.05, test for a significant difference in the preferences for the two brands. A plus indicates a preference for brand A over brand B.

2 The following hypothesis test is to be conducted.

H_0: Median $\leq$ 150
H_1: Median $>$ 150

A sample of size 30 yields 22 cases in which a value greater than 150 is obtained, three cases in which a value of exactly 150 is obtained, and five cases in which a value less than 150 is obtained. Use α = 0.01 and conduct the hypothesis test.

3 A poll asked 1253 adults a series of questions about the state of the economy and their children's future. One question was, 'Do you expect your children to have a better life than you have had, a worse life, or a life about as good as yours?' The responses were 34 per cent better, 29 per cent worse, 33 per cent about the same and 4 per cent not sure. Use the sign test and a 0.05 level of significance to determine whether more adults feel their children will have a better future than feel their children will have a worse future. What is your conclusion?

4 SNL Securities studied stock splits in the banking industry over an 18-month period and found that stock splits tended to increase the value of an individual's stock holding. Assume that of a sample of 20 recent stock splits, 14 led to an increase in value, four led to a decrease in value, and two resulted in no change. Suppose a sign test is to be used to determine whether stock splits continue to be beneficial for holders of bank stocks.

a. What are the null and alternative hypotheses?
b. With α = 0.05, what is your conclusion?

5 An opinion survey asked the following question regarding a proposed educational policy. 'Do you favour or oppose providing tax-funded vouchers or tax deductions to parents who send their children to private fee-paying schools?' Of the 2010 individuals surveyed, 905 favoured the support, 1045 opposed the support, and 60 offered no opinion. Do the data indicate a significant tendency towards favouring or opposing the proposed policy? Use a 0.05 level of significance.

6. Suppose a national survey in France has shown that the median annual income adults say would make their dreams come true is €152 000. Suppose further that of a sample of 225 individuals in Calais, 122 individuals report that the amount of income needed to make their dreams come true is less than €152 000 and 103 report that the amount needed is more than €152 000. Test the null hypothesis that the median amount of annual income needed to make dreams come true in Calais is €152 000. Use $\alpha = 0.05$. What is your conclusion?

7 The median number of part-time employees at fast-food restaurants in a particular city was known to be 15 last year. The city council thinks the use of part-time employees may have increased this year. A sample of nine fast-food restaurants showed that more than 15 part-time employees worked at seven of the restaurants, one restaurant had exactly 15 part-time employees, and one had fewer than 15 part-time employees. Test at $\alpha = 0.05$ to see whether the median number of part-time employees has increased.

8 Land Registry figures for late 2011 show the median selling price of houses in England as £185 000. Assume that the following data were obtained for sales of houses in Greater Manchester and in Oxfordshire.

	Greater than £185 000	Equal to £185 000	Less than £185 000
Greater Manchester	11	2	32
Oxfordshire	27	1	13

a. Is the median selling price in Greater Manchester lower than the national median of £185 000? Use a statistical test with $\alpha = 0.05$ to support your conclusion.
b. Is the median selling price in Oxfordshire higher than the national median of £185 000? Use a statistical test with $\alpha = 0.05$ to support your conclusion.

9 Two fuel additives are tested to determine their effect on litres of fuel consumed per 100 kilometres travelled, for passenger cars. Test results for 12 cars follow. Each car was tested with both fuel additives. Use $\alpha = 0.05$ and the Wilcoxon signed-rank test to see whether there is a significant difference in the additives.

	Additive			Additive	
Car	1	2	Car	1	2
1	7.02	7.82	7	8.74	8.21
2	6.00	6.49	8	7.62	9.43
3	6.41	6.26	9	6.46	7.05
4	7.37	8.28	10	5.83	6.68
5	6.65	6.65	11	6.09	6.20
6	5.70	5.93	12	5.65	5.96

10 A sample of ten men was used in a study to test the effects of a relaxant on the time required to fall asleep for male adults. Data for ten subjects showing the number of minutes required to fall asleep with and without the relaxant follow. Use a 0.05 level of significance to determine whether the relaxant reduces the time required to fall asleep. What is your conclusion?

Participant	Without relaxant	With relaxant	Participant	Without relaxant	With relaxant
1	15	10	6	7	5
2	12	10	7	8	10
3	22	12	8	10	7
4	8	11	9	14	11
5	10	9	10	9	6

11 A test was conducted of two overnight mail delivery services. Two samples of identical deliveries were set up so that both delivery services were notifi ed of the need for a delivery at the same time. The hours required to make each delivery follow. Do the data shown suggest a difference in the delivery times for the two services? Use a 0.05 level of significance for the test.

	Service	
Delivery	1	2
1	24.5	18.0
2	26.0	25.5
3	28.0	32.0
4	21.0	20.0
5	18.0	19.5
6	36.0	28.0
7	25.0	29.0
8	21.0	22.0
9	24.0	23.5
10	26.0	29.5
11	31.0	30.0

12 Ten test-market cities in France were selected as part of a market research study designed to evaluate the effectiveness of a particular advertising campaign. The sales in euros for each city were recorded for the week prior to the promotional programme. Then the campaign was conducted for two weeks and new sales data were collected for the week immediately after the campaign. The two sets of sales data (in thousands of euros) follow.

City	Pre-campaign sales	Post-campaign sales
Bordeaux	130	160
Strasbourg	100	105
Nantes	120	140
St Etienne	95	90
Lyon	140	130
Rennes	80	82
Le Havre	65	55
Amiens	90	105
Toulouse	140	152
Marseilles	125	140

Use $\alpha = 0.05$. What conclusion would you draw about the value of the advertising programme?

13 Two fuel additives are being tested to determine their effect on petrol consumption. Seven cars were tested with additive 1 and nine cars were tested with additive 2. The following data show the litres of fuel used per 100 kilometres with the two additives. Use $\alpha = 0.05$ and the MWW test to see whether there is a significant difference in petrol consumption for the two additives.

Additive 1	Additive 2
8.20	7.52
7.69	7.94
7.41	6.62
8.47	6.71
7.75	6.41
7.58	7.52
8.06	7.14
	6.80
	6.99

14 A company's price/earnings (P/E) ratio is the company's current stock price divided by the latest 12 months' earnings per share. Listed below are the P/E ratios for a sample of ten Japanese and 12 US companies. Is the difference in P/E ratios between the two countries significant? Use the MWW test and $\alpha = 0.01$ to support your conclusion.

Japan		US	
Company	P/E ratio	Company	P/E ratio
Sumitomo Corp.	153	Gannet	19
Kinden	21	Motorola	24
Heiwa	18	Schlumberger	24
NCR Japan	125	Oracle Systems	43
Suzuki Motor	31	Gap	22
Fuji Bank	213	Winn-Dixie	14
Sumitomo Chemical	64	Ingersoll-Rand	21
Seibu Railway	666	American Electric Power	14
Shiseido	33	Hercules	21
Toho Gas	68	Times Mirror	38
		WellPoint Health	15
		Northern States Power	14

15 Samples of starting annual salaries for individuals entering the public accounting and financial planning professions follow. Annual salaries are shown in thousands of euros.

Public Accountant	Public Accountant	Financial Planner	Financial Planner
45.2	50.0	44.0	48.6
53.8	45.9	44.2	44.7
51.3	54.5	48.1	48.9
53.2	52.0	50.9	46.8
49.2	46.9	46.9	43.9

a. Use $\alpha = 0.05$ level of significance and test the hypothesis that there is no difference between the starting annual salaries of public accountants and financial planners. What is your conclusion?
b. What are the sample mean annual salaries for the two professions?

16 A confederation of house builders provided data on the cost (in £) of the most popular home re-modelling projects. Use the Mann-Whitney-Wilcoxon test to see whether it can be concluded that the cost of kitchen re-modelling differs from the cost of master bedroom re-modelling. Use a 0.05 level of significance.

Kitchen	Master Bedroom
13 200	6 000
5 400	10 900
10 800	14 400
9 900	12 800
7 700	14 900
11 000	5 800
7 700	12 600
4 900	9 000
9 800	
11 600	

17 The gap between the earnings of men and women with equal education is narrowing but has not closed. Sample data for seven men and seven women with bachelor's degrees are as follows. Data of earnings are shown in thousands of euros.

Men	30.6	75.5	45.2	62.2	38.2	49.9	55.3
Women	44.5	35.4	27.9	40.5	25.8	47.5	24.8

a. What is the median salary for men? For women?
b. Use $\alpha = 0.05$ and conduct the hypothesis test for equal populations. What is your conclusion?

18 Three college admission test preparation programmes are being evaluated. The scores obtained by a sample of 20 people who used the test preparation programmes provided the following data. Use the Kruskal-Wallis test to determine whether there is a significant difference among the three test preparation programmes. Use $\alpha = 0.01$.

Programme		
A	B	C
540	450	600
400	540	630
490	400	580
530	410	490
490	480	590
610	370	620
	550	570

19 Forty-minute workouts of one of the following activities three days a week will lead to a loss of weight. The following sample data show the number of calories burned during 40-minute workouts for three different activities. Do these data indicate differences in the amount of calories burned for the three activities? Use a 0.05 level of significance. What is your conclusion?

Swimming	Tennis	Cycling
408	415	385
380	485	250
425	450	295
400	420	402
427	530	268

20 *Condé Nast Traveler* magazine conducts an annual survey of its readers in order to rate the top 80 cruise ships in the world. With 100 the highest possible rating, the overall ratings for a sample of ships from the Holland America, Princess, and Royal Caribbean cruise lines are shown here. Use the Kruskal-Wallis test with $\alpha = 0.05$ to determine whether the overall ratings among the three cruise lines differ significantly.

Holland America		Princess		Royal Caribbean	
Ship	Rating	Ship	Rating	Ship	Rating
Amsterdam	84.5	Coral	85.1	Adventure	84.8
Maasdam	81.4	Dawn	79.0	Jewel	81.8
Ooterdam	84.0	Island	83.9	Mariner	84.0
Volendam	78.5	Princess	81.1	Navigator	85.9
Westerdam	80.9	Star	83.7	Serenade	87.4

21 Course-evaluation ratings for four instructors follow. Use $\alpha = 0.05$ and the Kruskal-Wallis procedure to test for a significant difference in teaching abilities.

Instructor	Course-evaluation rating								
Black	88	80	79	68	96	69			
Jennings	87	78	82	85	99	99	85	94	81
Swanson	88	76	68	82	85	82	84	83	
Wilson	80	85	56	71	89	87			

22 Consider the following set of rankings for a sample of ten elements.

Element	x_i	y_i	Element	x_i	y_i
1	10	8	6	2	7
2	6	4	7	8	6
3	7	10	8	5	3
4	3	2	9	1	1
5	4	5	10	9	9

a. Compute the Spearman rank-correlation coefficient for the data.
b. Use $\alpha = 0.05$ and test for significant rank correlation. What is your conclusion?

23 Consider the following two sets of rankings for six items.

Case One			Case Two		
Item	First ranking	Second ranking	Item	First ranking	Second ranking
A	1	1	A	1	6
B	2	2	B	2	5
C	3	3	C	3	4
D	4	4	D	4	3
E	5	5	E	5	2
F	6	6	F	6	1

Note that in the first case the rankings are identical, whereas in the second case the rankings are exactly opposite. What value should you expect for the Spearman rank correlation coefficient for each of these cases? Explain. Calculate the rank-correlation coefficient for each case.

24 The following two lists show how ten IT companies ranked in a national survey, in terms of reputation and percentage of respondents who said they would purchase the company's shares. A positive rank correlation is anticipated because it seems reasonable to expect that a company with a higher reputation would be a more desirable purchase.

Company	Reputation	Probable purchase
Microsoft	1	3
Intel	2	4
Dell	3	1
Lucent	4	2
Texas Instruments	5	9
Cisco Systems	6	5
Hewlett-Packard	7	10
IBM	8	6
Motorola	9	7
Yahoo	10	8

a. Compute the rank correlation between reputation and probable purchase.
b. Test for a significant positive rank correlation. What is the *p*-value?
c. At $\alpha = 0.05$, what is your conclusion?

25 A student organization surveyed both recent graduates and current students to obtain information on the quality of teaching at a particular university. An analysis of the responses provided the following teaching-ability rankings. Do the rankings given by the current students agree with the rankings given by the recent graduates? Use $\alpha = 0.10$ and test for a significant rank correlation.

Professor	Ranking by	
	Current students	Recent graduates
A	4	6
B	6	8
C	8	5
D	3	1
E	1	2
F	2	3
G	5	7
H	10	9
J	7	4
K	9	10

26 A sample of 15 students received the following rankings on mid-term and final examinations in a statistics course.

Rank		Rank		Rank	
Mid-term	Final	Mid-term	Final	Mid-term	Final
1	4	6	2	11	14
2	7	7	5	12	15
3	1	8	12	13	11
4	3	9	6	14	10
5	8	10	9	15	13

Compute the Spearman rank-correlation coefficient for the data and test for a significant correlation, with $\alpha = 0.10$.

Chapter 18: Non-parametric Methods

Textbook Exercises Solutions:

1. Binomial Probabilities for $n = 10$, $\pi = 0.50$.

x	Probability	x	Probability
0	0.0010	6	0.2051
1	0.0098	7	0.1172
2	0.0439	8	0.0439
3	0.1172	9	0.0098
4	0.2051	10	0.0010
5	0.2461		

Number of plus signs is 7.

$$P(X \geq 7) = P(7) + P(8) + P(9) + P(10)$$
$$= 0.1172 + 0.0439 + 0.0098 + 0.0010$$
$$= 0.1719$$

Two-tailed p-value = 2(0.1719) = 0.3438

p-value > 0.05, do not reject H_0. There is no indication that a difference exists.

2. There are $n = 27$ cases in which a value different from 150 is obtained.

Use the normal approximation with $\mu = n\pi = 0.5(27) = 13.5$ and

$$\sigma = \sqrt{0.25n} = \sqrt{0.25(27)} = 2.6$$

Use $x = 22$ as the number of plus signs and obtain the following test statistic:

$$z = \frac{x - \mu}{\sigma} = \frac{22 - 13.5}{2.6} = 3.27$$

Last entry in normal distribution table is $z > 3.09$, cumulative probability = 0.9990.

For $z = 3.27$, p-value is less than (1 – 0.9990) = 0.0010. p-value < 0.01, reject H_0.

Conclusion: The median is greater than 150.

3. We need to determine the number who said better and the number who said worse. The sum of the two is the sample size used for the study.

$$n = 0.34(1253) + 0.29(1253) = 789.4$$

Use the large sample test involving the normal distribution. This means the value of n ($n = 789.4$ above) need not be integer. Hence,

$$\mu = 0.5\, n = 0.5(789.4) = 394.7$$

$$\sigma = \sqrt{0.25n} = \sqrt{0.25(789.4)} = 14.05$$

Let π = proportion of adults who feel children will have a better future.

H_0: $\pi \leq 0.50$

H_1: $\pi > 0.50$

With $x = 0.34(1253) = 426$

$$z = \frac{x - \mu}{\sigma} = \frac{426 - 394.7}{14.05} = 2.23$$

p-value = (1 – 0.9871) = 0.0129

p-value $\leq$ 0.05, reject H_0

Conclusion: More adults feel their children will have a better future.

4. a. Let π = probability the shares held will be worth more after the split

H_0: $\pi \leq 0.50$
H_1: $\pi > 0.50$

If H_0 cannot be rejected, there is no evidence to conclude stock splits continue to add value to stock holdings.

b. Let X be the number of plus signs (increases in value).

Use the binomial probability tables with n = 18 (there were 2 ties in the 20 cases)

$$\begin{aligned} P(X \geq 14) &= P(14) + P(15) + P(16) + P(17) + P(18) \\ &= 0.0117 + 0.0031 + 0.0006 + 0.0001 + 0.0000 \\ &= 0.0155 \end{aligned}$$

p-value = 0.0155

p-value < 0.05, reject H_0. The results support the conclusion that stock splits are beneficial for shareholders.

5. $n = 905 + 1045 = 1950$

Use the large sample test involving the normal distribution.

$\mu = 0.5\ n = 0.5(1950) = 975$

$$\sigma = \sqrt{0.25n} = \sqrt{0.25(1950)} = 22.08$$

Let π = proportion who favour the support.

H_0: $\pi = 0.50$

H_1: $\pi \neq 0.50$

With $x = 905$

$$z = \frac{x - \mu}{\sigma} = \frac{905 - 975}{22.08} = -3.17$$

p-value < 2(0.0010) = 0.002

p-value < 0.05, reject H_0

Conclusion: There is a significant tendency towards opposing the support.

6. H_0: Median = 152,000
H_1: Median ≠ 152,000

$\mu = 0.5n = 0.5(225) = 112.5$

$\sigma = \sqrt{0.25n} = \sqrt{0.25(225)} = 7.5$

For $x = 122$

$$z = \frac{202 - 112.5}{7.5} = 1.27$$

Area in tail = 1 − 0.8980 = 0.1020

p-value = 2(0.1020) = 0.2040

p-value > 0.05, do not reject H_0. We are unable to conclude that the median annual income needed differs from that reported in the survey.

7. H_0: Median ≤ 15
H_1: Median > 15

Use binomial probabilities with $n = 8$ and $\pi = 0.50$:

$P(X \geq 7) = P(7) + P(8)$
$= 0.0313 + 0.0039 = 0.0352$

p-value = 0.0352

p-value < 0.05, reject H_0. Data does enable us to conclude that there has been an increase in the median number of part-time employees.

8. a. H_0: Median $\geq$ 185,000
H_1: Median < 185,000

$n = 11 + 32 = 43$

Use the large sample test involving the normal distribution. Let x be the number of houses with prices less than £185,000.

$\mu = 0.5n = 0.5(43) = 21.5$

$\sigma = \sqrt{0.25n} = \sqrt{0.25(43)} = 3.279$

$$z = \frac{x - \mu}{\sigma} = \frac{32 - 21.5}{3.279} = 3.202$$

p-value < 1 – 0.9990 = 0.001

p-value < 0.05, reject H_0. Conclude that the median selling price in Manchester is less than the national median.

b. H_0: Median $\leq$ 185,000
H_1: Median > 185,000

$n = 27 + 13 = 40$

Use the large sample test involving the normal distribution. Let x be the number of houses with prices greater than £185,000.

$\mu = 0.5n = 0.5(40) = 20$

$\sigma = \sqrt{0.25n} = \sqrt{0.25(43)} = 3.162$

$$z = \frac{x - \mu}{\sigma} = \frac{27 - 20}{3.162} = 2.214$$

p-value < 1 – 0.9864 = 0.0136

p-value < 0.05, reject H_0. Conclude that the median selling price in Oxfordshire is greater than the national median.

9. H_0: The populations are identical
H_1: The populations are not identical

Additive 1	Additive 2	Difference	Absolute Value	Rank	Signed Rank
7.02	7.82	0.82	0.82	8	+8
6.00	6.49	0.49	0.49	5	+5
6.41	6.26	-0.15	0.15	2	-2
7.37	8.28	0.91	0.91	10	+10
6.65	6.65	0	0		
5.70	5.93	0.23	0.23	3	+3
8.74	8.21	-0.53	0.53	6	-6
7.62	9.43	1.81	1.81	11	+11
6.46	7.05	0.59	0.59	7	+7
5.83	6.68	0.85	0.85	9	+9
6.09	6.20	0.11	0.11	1	+1
5.65	5.96	0.41	0.41	4	+4
					Total 50

$\mu_T = 0$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{11(12)(23)}{6}} = 22.5$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{50 - 0}{22.5} = 2.22$$

p-value = 2(1 – 0.9868) = 0.0262

p-value < 0.05, reject H_0.

Conclusion: There is a significant difference between the additives.

10.

Without Relaxant	With Relaxant	Difference	Rank of Absolute Difference	Signed Rank
15	10	5	9	9
12	10	2	3	3
22	12	10	10	10
8	11	-3	6.5	-6.5
10	9	1	1	1
7	5	2	3	3
8	10	-2	3	-3
10	7	3	6.5	6.5
14	11	3	6.5	6.5
9	6	3	6.5	6.5
				Total 36

$\mu_T = 0$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{10(11)(21)}{6}} = 19.62$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{36}{19.62} = 1.83$$

One-tailed test.

p-value = 1 – 0.9664 = 0.0336

p-value < 0.05, reject H_0.

Conclusion: There is a significant difference, in favour of the relaxant.

11.

Service #1	Service #2	Difference	Rank of Absolute Difference	Signed Rank
24.5	28.0	-3.5	7.5	-7.5
26.0	25.5	0.5	1.5	1.5
28.0	32.0	-4.0	9.5	-9.5
21.0	20.0	1.0	4	4.0
18.0	19.5	-1.5	6	-6.0
36.0	28.0	8.0	11	11.0
25.0	29.0	-4.0	9.5	-9.5
21.0	22.0	-1.0	4	-4.0
24.0	23.5	0.5	1.5	1.5
26.0	29.5	-3.5	7.5	-7.5
31.0	30.0	1.0	4	4.0
				T = -22.0

$\mu_T = 0$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{11(12)(23)}{6}} = 22.49$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{-22}{22.49} = -0.98$$

p-value = 2(0.1635) = 0.3270

p-value > 0.05, do not reject H_0. There is no significant difference.

12.

City	Pre-campaign sales	Post-campaign sales	Difference	Rank of Absolute Difference	Signed Rank
Bordeaux	130	160	30	10	10
Strasbourg	100	105	5	2.5	2.5
Nantes	120	140	20	9	9
St Etienne	95	90	-5	2.5	-2.5
Lyon	140	130	-10	4.5	-4.5
Rennes	80	82	2	1	1
Le Havre	65	55	-10	4.5	-4.5
Amiens	90	105	15	7.5	7.5
Toulouse	140	152	12	6	6
Marseilles	125	140	15	7.5	7.5
					$T = 32$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{10(11)(21)}{6}} = 19.62$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{32}{19.62} = 1.63$$

p-value = 1 – 0.9484 = 0.0516

p-value > 0.05, do not reject H_0. Insufficient evidence to conclude that the campaign had an effect on sales.

13. Rank the combined samples and find the rank sum for each sample.

This is a small sample test since $n_1 = 7$ and $n_2 = 9$

Additive 1		Additive 2	
litres per 100 km	Rank	litres per 100 km	Rank
8.20	15	7.52	8.5
7.69	11	7.94	13
7.41	7	6.62	2
8.47	16	6.71	3
7.75	12	6.41	1
7.58	10	7.52	8.5
8.06	14	7.14	6
	85	6.80	4
		6.99	5
			51

$T = 85$

With $\alpha = 0.05$, $n_1 = 7$ and $n_2 = 9$

$T_L = 41$ and $T_U = 7(7 + 9 + 1) - 41 = 78$

Since $T = 85 > 78$, we reject H_0

Conclusion: There is a significant difference in petrol consumption.

14. H_0: There is no difference in the distributions of P/E ratios
H_1: There is a difference between the distributions of P/E ratios

Japan			United States		
Company	P/E Ratio	Rank	Company	P/E Ratio	Rank
Sumitomo Corp.	153	20	Gannet	19	6
Kinden	21	8	Motorola	24	11.5
Heiwa	18	5	Schlumberger	24	11.5
NCR Japan	125	19	Oracle Systems	43	16
Suzuki Motor	31	13	Gap	22	10
Fuji Bank	213	21	Winn-dixie	14	2
Sumitomo Chemical	64	17	Ingersoll-Rand	21	8
Seibu Railway	666	22	Am. Elec. Power	14	2
Shiseido	33	14	Hercules	21	8
Toho Gas	68	18	Times Mirror	38	15
	Total	157	WellPoint Health	15	4
			No. States Power	14	2
				Total	96

$$\mu_T = \frac{1}{2} n_1 (n_1 + n_2 + 1) = \frac{1}{2} 10(10 + 12 + 1) = 115$$

$$\sigma_T = \sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)} = \sqrt{\frac{1}{12} (10)(12)(10 + 12 + 1)} = 15.17$$

$T = 157$

$$z = \frac{157 - 115}{15.17} = 2.77$$

p-value = 2(1 − 0.9972) = 0.0056

p-value < 0.01, reject H_0.

We conclude that there is a significant difference in P/E ratios for the two countries.

15. a. Rank the combined samples and find the rank sum for each sample.

This is a small sample test since $n_1 = 10$ and $n_2 = 10$

Public Accountant		Financial Planners	
Salary (€ 000s)	Rank	Salary (€ 000s)	Rank
45.2	5	44.0	2
53.8	19	44.2	3
51.3	16	48.1	10
53.2	18	50.9	15
49.2	13	46.9	8.5
50.0	14	48.6	11
45.9	6	44.7	4
54.5	20	48.9	12
52.0	17	46.8	7
46.9	8.5	43.9	1
Total	136.5	Total	73.5

Use $T = 136.5$

With $\alpha = 0.05$, $n_1 = 10$ and $n_2 = 10$

$T_L = 79$ and $T_U = 10(10 + 10 + 1) - 79 = 131$

Since $T = 136.5 > 131$, we reject H_0

There is a significant difference between the starting salaries of the two professions.

b. Sample mean for Public Accountants is €50,200

Sample mean for Financial Planners is €46,700

16. Rank the combined samples and find the rank sum for each sample.

This is a small sample test since $n_1 = 10$ and $n_2 = 8$

Kitchen		Master Bedroom	
Cost in £	Rank	Cost in £	Rank
13,200	16	6,000	4
5,400	2	10,900	11
10,800	10	14,400	17
9,900	9	12,800	15
7,700	5.5	14,900	18
11,000	12	5,800	3
7,700	5.5	12,600	14
4,900	1	9,000	7
9,800	8		89
11,600	13		
	82		

$T = 82$

With $\alpha = 0.05$, $n_1 = 10$ and $n_2 = 8$

$T_L = 76$ and $T_U = 10(10 + 8 + 1) - 76 = 114$

Since $T = 82$ lies between T_L and T_U, we cannot reject H_0

Conclusion: There is a significant difference between the costs.

17. a. Median salary for men is €49,900 (4^{th} ordered value)

Median salary for men is €35,400 (4^{th} ordered value)

b. This is a small sample test since $n_1 = 7$ and $n_2 = 7$

Men		Women	
30.6	4	44.5	8
75.5	14	35.4	5
45.2	9	27.9	3
62.2	13	40.5	7
38.2	6	25.8	2
49.9	11	47.5	10
55.3	12	24.8	1
	69		36

$T = 69$

With $\alpha = 0.05$, $n_1 = 7$ and $n_2 = 7$

$T_L = 37$ and $T_U = 7(7 + 7 + 1) - 37 = 68$

Since $T = 69 > T_U$, we reject H_0.

There is a significant difference between the earnings of men and women.

18.

A	B	C
11.5	5.0	17.0
2.5	11.5	20.0
8.0	2.5	15.0
10.0	4.0	8.0
8.0	6.0	16.0
18.0	1.0	19.0
	13.0	14.0
58.0	43.0	109.0

$$W = \frac{12}{(20)(21)}\left[\frac{(58)^2}{6} + \frac{(43)^2}{7} + \frac{(109)^2}{7}\right] - 3(21) = 9.06$$

Degrees of freedom = 2

Using χ^2 table, $\chi^2 = 9.06$ shows p-value is between 0.01 and 0.025 (actual p-value = 0.0108)

p-value > 0.01, do not reject H_0. We cannot conclude that there is a significant difference in test preparation programs.

19.

Swimming	Rank	Tennis	Rank	Cycling	Rank
408	8	415	9	385	5
380	4	485	14	250	1
425	11	450	13	295	3
400	6	420	10	402	7
427	12	530	15	268	2
Sum	41		61		18

$$W = \frac{12}{(15)(15+1)}\left[\frac{(41)^2}{5} + \frac{(61)^2}{5} + \frac{(18)^2}{5}\right] - 3(15+1) = 9.26$$

Degrees of freedom = 2

Using χ^2 table, $\chi^2 = 9.26$ shows p-value is between 0.005 and 0.01

Actual p-value = 0.0098

p-value < 0.05, reject H_0. Conclude that there is a significant difference in calories among the three activities.

20.

Holland America	Rank	Princess	Rank	Royal Caribbean	Rank
84.5	11	85.1	13	84.8	12
81.4	5	79.0	2	81.8	6
84.0	9.5	83.9	8	84.0	9.5
78.5	1	81.1	4	85.9	14
80.9	3	83.7	7	87.4	15
Sum	29.5		34		56.5

$$W = \frac{12}{(15)(15+1)}\left[\frac{(29.5)^2}{5} + \frac{(34)^2}{5} + \frac{(56.5)^2}{5}\right] - 3(15+1) = 4.185$$

Degrees of freedom = 2

Using χ^2 table, $\chi^2 = 8.03$ shows p-value is greater than 0.10 (actual p-value = 0.123)

p-value > 0.05, do not reject H_0. We cannot conclude that there is a significant difference between the ratings of the three cruise lines.

21.

Black	Jennings	Swanson	Wilson
22.5	20.5	22.5	9.5
9.5	27.0	6.0	17.5
8.0	7.0	2.5	1.0
2.5	17.5	12.5	5.0
26.0	28.5	17.5	24.0
4.0	28.5	12.5	20.5
	17.5	15.0	
	25.0	14.0	
		11.0	
72.5	171.5	113.5	77.5

$$W = \frac{12}{(29)(30)}\left[\frac{(72.5)^2}{6} + \frac{(171.5)^2}{8} + \frac{(113.5)^2}{9} + \frac{(77.5)^2}{6}\right] - 3(30) = 6.34$$

Degrees of freedom = 3

Using χ^2 table, $\chi^2 = 6.34$ shows p-value is between 0.05 and 0.10

Actual p-value = 0.0962

p-value > 0.05, do not reject H_0. We cannot conclude that there is a significant difference among the course evaluation ratings for the 4 instructors.

22. a. $\Sigma d_i^2 = 52$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(52)}{10(99)} = 0.68$$

b. $$\sigma_{r_S} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{9}} = 0.33$$

$$z = \frac{r_S - 0}{\sigma_{r_S}} = \frac{0.68}{0.33} = 2.05$$

p-value = 2(1 – 0.9798) = 0.0404

p-value < 0.05, reject H_0. Conclude that significant rank correlation exists.

23. Case 1:

$\Sigma d_i^2 = 0$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(0)}{6(36-1)} = 1$$

Case 2:

$\Sigma d_i^2 = 70$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(70)}{6(36-1)} = -1$$

With perfect agreement, $r_S = 1$.

With exact opposite ranking, $r_S = -1$.

24. a. $\Sigma d_i^2 = 54$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(54)}{10(10^2-1)} = 0.67$$

b. $H_0 : \rho_S \leq 0$
$H_1 : \rho_S > 0$

$$\sigma_{r_S} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{10-1}} = 0.33$$

$$z = \frac{r_S - \mu_{r_S}}{\sigma_{r_S}} = \frac{0.67 - 0}{0.33} = 2.02$$

p-value = 2(1 – 0.9783) = 0.0434

c. p-value < 0.05, reject H_0. Conclude a significant positive rank correlation.

25. $\Sigma d_i^2 = 38$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(38)}{10(99)} = 0.77$$

$$\mu_{r_S} = 0$$

$$\sigma_{r_S} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{9}} = 0.33$$

$$z = \frac{r_S - 0}{\sigma_{r_S}} = \frac{0.77}{0.33} = 2.31$$

p-value = 2(1 − 0.9896) = 0.0208

p-value < 0.10, reject H_0. There is a significant rank correlation between current students and recent graduates.

26. $\Sigma d_i^2 = 136$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(136)}{15(224)} = 0.76$$

$$\sigma_{r_S} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{14}} = 0.27$$

$$z = \frac{r_S - \mu_{r_S}}{\sigma_{r_S}} = \frac{0.76}{0.27} = 2.83$$

p-value = 2(1 − 0.9977) = 0.0046

p-value < 0.10, reject H_0. Conclude that there is a significant rank correlation between the two exams.

Supplementary Exercises:

27. *Coronation Street* and *Eastenders* are two of the top-rated U.K. television programmes. Assume that in a local television preference poll, 410 individuals were asked to indicate their favourite television programme: 185 selected *Coronation Street*, 165 selected *Eastenders*, and 60 selected other television programmes. Use a 0.05 level of significance to test the hypothesis that *Coronation Street* and *Eastenders* have the same level of preference. What is your conclusion?

28. In 1996, Packard Bell and Compaq were the leaders in the consumer PC market (*USA Today*, July 21, 1997). A sample of 500 purchases showed 202 Packard Bell computers, 158 Compaq computers, and 140 other computers including IBM, Apple, and NEC. Use a 0.05 level of significance to test the hypothesis that Packard Bell and Compaq had the same share of the consumer PC market. What is your conclusion?

29. In a sample of 150 university football games that ended with a win for one or other of the teams, the home team won 98 games. Test to see whether the data support the claim of a home-team advantage in university football. Use a 0.05 level of significance. What is your conclusion?

30. The median annual income of subscribers to *Barron's* magazine is $131,000 (barronsmag.com, July 28, 2000). Assume a sample of 300 subscribers to *The Wall Street Journal* found 165 subscribers with an income over $131,000 and 135 subscribers with an income under $131,000. Can you conclude that there is any difference between the median incomes of the two subscriber groups? At $\alpha = 0.05$, what is your conclusion?

31. File "Chicago"

    The U.S. Bureau of Labor Statistics reported the 1996 median weekly earnings for women in administrative and managerial positions to be US$585. A sample of women working in administrative and managerial posts in the Chicago area provided income figures for 50 women. These data are in the file "Chicago". Use the sample data to test H_0: median $\leq$ 585, H_1: median > 585 for the population of administrative and managerial working women in Chicago. Use a 0.05 level of significance. What is your conclusion?

32. The 1997 price/earnings ratios for a sample of 12 stocks are shown in the following list (*Barron's,* December 8, 1997). Assume that a financial analyst provided the estimated price/earnings ratio for 1998. Using a 0.05 level of significance, what is your conclusion about the differences between the price/earnings ratios for 1997 and the estimates for 1998?

Stock	1997 P/E ratio	1998 P/E ratio (Est.)
Coca-Cola	40	32
Du Pont	24	22
Eastman Kodak	21	23
General Electric	30	23
General Mills	25	19
IBM	19	19
McDonald's	20	17
Merck	29	19
Motorola	35	20
Philip Morris	17	18
Walt Disney	33	27
Xerox	20	16

33. Samples of annual starting salaries for individuals entering the public accounting and financial planning professions follow (*Fortune,* June 26, 1995). Annual salaries are shown in thousands of dollars.

Public Accountant	Financial planner	Public Accountant	Financial planner
25.2	24.0	30.0	28.6
33.8	24.2	25.9	24.7
31.3	28.1	34.5	28.9
33.2	30.9	31.7	26.8
29.2	26.9	26.9	23.9

a. What are the sample mean annual salaries for the two professions?

b. Use a 0.05 level of significance and test the hypothesis that there is no difference between the starting annual salaries of public accountants and financial planners. What is your conclusion?

34. Sample data for seven men and seven women with bachelor's degrees are as follows. Data are shown in thousands of euros.

Men	30.6	75.5	45.2	62.2	38.2	49.9	55.3
Women	44.5	35.4	27.9	40.5	25.8	47.5	24.8

a. What is the sample median salary for men? For women?

b. Use $\alpha = 0.05$ and conduct a hypothesis test to examine whether these samples are drawn from identical populations. What is your conclusion?

35. Fuel consumption tests were conducted for two models of car. Twelve cars of each model were selected randomly and a litres per 100 kilometres rating for each model was measured on the basis of 1000 kilometres of extra-urban driving. The data follow.

	Model 1		**Model 2**
Car	**Litres per 100 km**	**Car**	**Litres per 100 km**
1	10.6	1	11.3
2	9.9	2	7.6
3	8.6	3	7.4
4	8.9	4	8.5
5	8.8	5	9.7
6	10.2	6	11.1
7	11.0	7	7.3
8	10.5	8	8.8
9	9.8	9	7.8
10	9.8	10	6.9
11	9.2	11	8.0
12	10.5	12	10.1

Use $\alpha = 0.10$ and test for a significant difference in the populations of litres-per-100 kilometres ratings for the two models.

36. A certain brand of TFT-screen TV was priced (in €) at 10 stores in city X and 13 stores in city Y. The data follow. Use a 0.05 level of significance and test whether prices for the TV oven are the same in the two cities.

City X	445	489	405	485	439	449	436	420	430	405			
City Y	460	451	435	479	475	445	429	434	410	422	425	459	430

37. Three products received the following performance ratings by a panel of 15 consumers. Use the Kruskal-Wallis test and $\alpha = 0.05$ to determine whether there is a significant difference in the performance ratings for the products.

	Product	
A	**B**	**C**
50	80	60
52	95	45
75	98	30
48	87	58
65	90	57

38. The better-selling sweets and chocolates tend to be high in calories. Assume that the following data show the calorie content from samples of M&Ms, Kit Kat and Milky Way II. Test for significant differences in the calorie content of these three products. At a 0.05 level of significance, what is your conclusion?

	Product	
M&Ms	**Kit Kat**	**Milky Way II**
230	225	200
210	205	208
240	245	202
250	235	190
230	220	180

39. For a sample of 11 cities, the following table gives the ranks for pupil-teacher ratio (1 = lowest, 11 = highest) and expenditure per pupil (1 = highest, 11 = lowest).

City	Pupil-Teacher Ratio	Expenditure per Pupil	City	Pupil-Teacher Ratio	Expenditure per Pupil
A	10	9	G	1	1
B	8	5	H	2	7
C	6	4	I	7	8
D	11	2	J	5	10
E	4	66	K	9	3
F	3	11			

At the $\alpha = 0.05$ level, does there appear to be a relationship between expenditure per pupil and pupil-teacher ratio?

40. The 1996 rankings of a sample of professional golfers for driving distance and for putting follow (*Golf Digest,* January 1997). What is the rank correlation between driving distance and putting rankings? Use a 0.10 level of significance.

Golfer	Driving distance	Putting
Fred Couples	1	5
David Duval	5	6
Ernie Els	4	10
Nick Faldo	9	2
Tom Lehman	6	7
Justin Leonard	10	3
Davis Love III	2	8
Phil Mickelson	3	9
Greg Norman	7	4
Mark O'Meara	8	1

Supplementary Exercises Solutions:

27. $n = 185 + 165 = 350$

$\mu = 0.5n = 0.5(350) = 175$

$\sigma = \sqrt{0.25n} = \sqrt{0.25(350)} = 9.35$

Using *Coronation Street*, $x = 185$

$$z = \frac{185 - 175}{9.35} = 1.07$$

Area in tail = 1 − 0.8577 = 0.1423

p-value = 2(0.1423) = 0.2846

p-value > 0.05, do not reject H_0. Cannot conclude there is a difference in preference for the two programmes.

28. $n = 202 + 158 = 360$

$\mu = 0.5n = 0.5(360) = 180$

$\sigma = \sqrt{0.25n} = \sqrt{0.25(360)} = 9.49$

Using Packard Bell, $x = 202$

$$z = \frac{202 - 180}{9.49} = 2.32$$

Area in tail = 1 − 0.9898 = 0.0102

p-value = 2(0.0102) = 0.0204

p-value < 0.05, reject H_0. Conclude Packard Bell and Compaq had different market shares.

29. $\mu = 0.5n = 0.5(150) = 75$

$\sigma = \sqrt{0.25n} = \sqrt{0.25(150)} = 6.12$

For 98 + signs

$$z = \frac{98 - 75}{6.12} = 3.76$$

For $z = 3.76$, area in tail is less than 0.001

With two tails, p-value is less than 0.002

Using Excel, p-value = 0.0002

p-value < 0.05, reject H_0. Conclude that a home team advantage exists.

30. $\mu = 0.5n = 0.5(300) = 150$

$$\sigma = \sqrt{0.25n} = \sqrt{0.25(300)} = 8.66$$

$$z = \frac{165 - 150}{8.66} = 1.73$$

p-value = 2(1 − 0.9582) = 0.0836

p-value > 0.05, do not reject H_0; we are unable to conclude that the median annual incomes differ.

31. $n = 50$

$\mu = 0.5n = 0.5(50) = 25$

$$\sigma = \sqrt{0.25n} = \sqrt{0.25(50)} = 3.54$$

$x = 33$ had wages greater than \$585

$$z = \frac{33 - 25}{3.54} = 2.26$$

p-value = 1 − 0.9881 = 0.0119

p-value < 0.05, reject H_0. Conclude that the median weekly wage is greater than \$585.

32.

1997 P/E Ratio	Est. 1998 P/E Ratio	Difference	Rank	Signed Rank
40	32	8	9	9
24	22	2	2.5	2.5
21	23	–2	2.5	–2.5
30	23	7	8	8
25	19	6	6.5	6.5
19	19	0	0	0
20	17	3	4	4
29	19	10	10	10
35	20	15	11	11
17	18	–1	1	–1
33	27	6	6.5	6.5
20	16	4	5	5
				$T = 59$

$n = 11$ (discarding the 0)

$\mu_T = 0$ $\qquad$ $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{11(12)(23)}{6}} = 22.49$

$$z = \frac{59 - 0}{22.49} = 2.62$$

p-value = 2(1– 0.9956) = 0.0088

p-value < 0.05, reject H_0. Conclude a difference exists between the 1997 P/E ratios and the estimated 1998 P/E ratios.

33. a. Public Accountant: $30,200
Financial Planner: $26,700

b.

Public Accountant	Rank	Financial Planner	Rank
25.2	5	24.0	2
33.8	19	24.2	3
31.3	16	28.1	10
33.2	18	30.9	15
29.2	13	26.9	8.5
30.0	14	28.6	11
25.9	6	24.7	4
34.5	20	28.9	12
31.7	17	26.8	7
26.9	8.5	23.9	1
	136.5		73.5

$T = 136.5$

$$\mu_T = \frac{1}{2} n_1 (n_1 + n_2 + 1) = \frac{1}{2} 10(10 + 10 + 1) = 105,$$

$$\sigma_T = \sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)} = \sqrt{\frac{1}{12} (10)(10)(21)} = 13.23$$

$$z = \frac{136.5 - 105}{13.23} = 2.38$$

p-value = 2(1 − 0.9913) = 0.0174

p-value < 0.05, reject H_0. Salaries differ significantly for the two professions.

34. a. Median → 4th salary for each
Men 49.9 Women 35.4

b.

Men	Rank	Women	Rank
30.6	4	44.5	8
75.5	14	35.4	5
45.2	9	27.9	3
62.2	13	40.5	7
38.2	6	25.8	2
49.9	11	47.5	10
55.3	12	24.8	1
			$T = 36$

From tables $T_L = 37$

$T < T_L$, reject H_0. Conclude that the populations differ. Men show higher salaries.

35. Sum of ranks (Model 1) = 185.5

Sum of ranks (Model 2) = 114.5

Use $T = 185.5$

$$\mu_T = \frac{1}{2} n_1 (n_1 + n_2 + 1) = \frac{1}{2} 12(12 + 12 + 1) = 150$$

$$\sigma_T = \sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)} = \sqrt{\frac{1}{12} (12)(12)(25)} = 17.32$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{185.5 - 150}{17.32} = 2.05$$

p-value = 2(1 – 0.9798) = 0.0404

p-value < 0.10, reject H_0.

Conclusion: there is a significant difference between the populations.

36. Sum of ranks (City X) = 116 Sum of ranks (City Y) = 160

Use $T = 116$

$$\mu_T = \frac{1}{2} n_1 (n_1 + n_2 + 1) = \frac{1}{2} 10(24) = 120$$

$$\sigma_T = \sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)} = \sqrt{\frac{1}{12} (10)(13)(24)} = 16.12$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{116 - 120}{16.12} = -0.25$$

p-value = 2(0.4013) = 0.8026

p-value > 0.05, do not reject H_0. There is no convincing evidence to conclude that there is a difference.

37.

A	B	C
4	11	7
8	14	2
10	15	1
3	12	6
9	13	5
34	65	21

$$W = \frac{12}{(15)(16)}\left[\frac{(34)^2}{5} + \frac{(65)^2}{5} + \frac{(21)^2}{5}\right] - 3(16) = 10.22$$

Degrees of freedom = 2

Using χ^2 table, $\chi^2 = 10.22$ shows p-value is between 0.005 and 0.01

(Actual p-value = 0.006)

p-value < 0.05, reject H_0. Conclude that the ratings for the products differ.

38.

M&Ms	Kit Kat	Milky Way II
10.5	9	3
7	5	6
13	14	4
15	12	2
10.5	8	1
56	48	16

$$W = \frac{12}{(15)(16)}\left[\frac{(56)^2}{5} + \frac{(48)^2}{5} + \frac{(16)^2}{5}\right] - 3(16) = 8.96$$

Degrees of freedom = 2

Using χ^2 table, $\chi^2 = 8.96$ shows p-value is between 0.01 and 0.025

(Actual p-value = 0.0113)

p-value < 0.05, reject H_0. There are significant differences in calorie content among the three products.

39. $\Sigma d_i^2 = 250$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(250)}{11(120)} = -0.136$$

$$\sigma_{r_S} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{10}} = 0.32$$

$$z = \frac{r_S - 0}{\sigma_{r_S}} = \frac{-0.136}{0.32} = -0.43$$

p-value = 2(0.3336) = 0.6672

p-value > 0.05, do not reject H_0. Conclude that there is not a significant relationship between the rankings.

40.

Driving Distance	Putting	d_i	d_i^2
1	5	-4	16
5	6	-1	1
4	10	-6	36
9	2	7	49
6	7	-1	1
10	3	7	49
2	8	-6	36
3	9	-6	36
7	4	3	9
8	1	7	49
			$\Sigma d_i^2 = 282$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(282)}{10(100-1)} = -0.71$$

$$\mu_{r_S} = 0$$

$$\sigma_{r_S} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{9}} = 0.33$$

$$z = \frac{-0.71 - 0}{0.33} = -2.13$$

p-value = 2(0.0166) = 0.0332

p-value < 0.10, reject H_0. There is a significant negative rank correlation between driving distance and putting.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Nineteen

Index Numbers

Textbook Exercises (1-25)
Textbook Exercise Solutions
Supplementary Exercises (26-40)
Supplementary Exercise Solutions

Chapter 19: Index Numbers

Textbook Exercises:

1. The following table reports prices and quantities used for two items in 2010 and 2013.

	Quantity		Unit price (€)	
Item	2010	2013	2010	2013
A	1500	1800	7.50	7.75
B	2	1	630.00	1500.00

a. Compute price relatives for each item in 2013 using 2010 as the base period.
b. Compute an unweighted aggregate price index for the two items in 2013 using 2010 as the base period.
c. Compute a weighted aggregate price index for the two items using the Laspeyres method.
d. Compute a weighted aggregate price index for the two items using the Paasche method.

2. An item with a price relative of 132 cost €10.75 in 2013. The base year was 2005.
a. What was the percentage increase or decrease in cost of the item over the 8-year period?
b. What did the item cost in 2005?

3. The average selling prices for an apartment in London, England for the years 2010, 2011 and the first half of 2012 and 2008, by quarter, were as follows (Source: Nationwide Building Society).

Quarter & Year	Price (£)
Q1 2010	229 422
Q2 2010	236 890
Q3 2010	239 188
Q4 2010	234 194
Q1 2011	233 259
Q2 2011	244 706
Q3 2011	237 535
Q4 2011	245 437
Q1 2012	239 309
Q2 2012	267 101

a. Use Q1 2010 as the base period and construct a price index for apartments in London over this 2.5-year period.
b. Use Q1 2011 as the base year and construct a price index for apartments in Yorkshire over this 2.5-year period.

4. A large manufacturer purchases an identical component from three independent suppliers that differ in respect of unit price and quantity supplied. The relevant data for 2010 and 2013 are given here.

		Unit price (€)	
Supplier	Quantity (2010)	2010	2013
A	150	5.45	6.00
B	200	5.60	5.95
C	120	5.50	6.20

a. Compute the 2013 price relatives for each of the component suppliers separately, using 2010 as the base year. Compare the price increases by the suppliers over the two-year period.
b. Compute an unweighted aggregate price index for the component part in 2013, using 2010 as the base year.
c. Compute a 2013 weighted aggregate price index for the component part, using 2010 as the base year. What is the interpretation of this index value for the manufacturing firm?

5. Tipple Limited provides a complete range of beer, wine and soft drink products for distribution through retail outlets in Ireland. Below are unit price data for 2008 and 2013, and quantities sold (cases of 12 one-litre bottles) for 2008.

		Unit price (€)	
Product	2008 quantity (cases)	2008	2013
Beer	35 000	21.50	23.25
Wine	5 000	85.70	91.50
Soft drink	60 000	14.00	14.30

Compute a weighted aggregate price index for Tipple Limited products in 2013, with 2008 as the base period.

6. The data below refer to four products produced by a company. Under the LIFO (Last In, First Out) inventory valuation method, a price index for inventory must be established for tax purposes with quantity weights based on year-ending inventory levels. Using the beginning-of-the-year price per unit as the base-period price, construct a weighted aggregate price index for total inventory at the end of the year. What type of index number is this?

		Unit price (€)	
Product	Year-end inventory	Beginning	Ending
A	500	0.15	0.19
B	50	1.60	1.80
C	100	4.50	4.20
D	40	12.00	12.20

7. Exports Europe Limited has the following data on quantities exported and unit costs of shipping for each of its four products:

		Mean freight cost per unit (€)	
Products	Quantities exported in 2009	2009	2013
A	2000	0.50	15.90
B	5000	6.25	32.00
C	6500	2.20	17.40
D	2500	20.00	35.50

Using 2009 as the base year, compute a Laspeyres price index that reflects the freight cost change over the four-year period.

8. With 2009 as the base year, use the price data in exercise 7 to compute a Paasche index for the freight cost in 2013, if 2013 quantities are 4000, 3000, 7500 and 3000 for each of the four products.

9. Price relatives for three items, along with base-period prices and usage, are shown in the following table. Compute a weighted aggregate price index for the current period.

	Current period	Base period	
Item	price relative	Price (€ per unit)	Usage (number of units)
A	150	22.00	20
B	90	5.00	50
C	120	14.00	40

10. The Europa Chemical Company produces a special industrial chemical that is a blend of three chemical ingredients. The beginning-of-year cost per pound, the end-of-year cost per kilogram and the blend proportions follow.

	Cost per kilogram (€)		
Ingredient	Beginning	Ending	Quantity (kg) per 100 kg of product
A	2.50	3.95	25
B	8.75	9.90	15
C	0.99	0.95	60

a. Compute a price relative for end-of-year, using beginning-of-year as base period, for each of the three ingredients.
b. Compute a weighted average of the price relatives to develop an end-of-year cost index for raw materials used in the product. What is your interpretation of this index value?

11. An investment portfolio consists of shares in four companies. The purchase price, current price, and number of shares are reported in the following table.

Company	Purchase price per share (£)	Current price per share (£)	Number of shares
Euro Leisure	1.55	1.70	5000
UK Utilities	1.85	2.05	2000
Scand Gas	2.65	2.60	5000
PQ Domestic	4.25	4.55	3000

Construct a weighted average of price relatives as an index of the performance of the portfolio to date. Interpret this price index.

12. Compute the price relatives for the Tipple Limited products in exercise 5. Use a weighted average of price relatives to show that this method provides the same index as the weighted aggregate method.

13. Using 2009 as the base year, compute the 2013 price relatives for the four products making up the index in exercise 7. Use the weighted aggregates of price relatives method to compute a value for the 2013 index.

14. On 1 April 1999 when the UK National Minimum Wage was introduced, the minimum rate for workers aged 21 and over was £3.60 per hour. In April 2012, the rate was £6.08 per hour. The RPI in April 1999 was 165.2; in April 2012 it was 242.5 (1 January 1987 = 100).

a. Deflate the hourly wage rates in 1999 and 2012 to find the real wage rates at January 1987 price levels.
b. What is the percentage change from 1999 to 2012 in actual minimum hourly wage rates?
c. What is the percentage change from 1999 to 2012 in real wage rates at January 1987 price levels?

15. Statistics Sweden report the following exports of furniture from Sweden over the period 2005 to 2011, and a relevant Export Price Index (EPI) for furniture.

Year	Exports (Swedish Krone, millions)	EPI (2005 = 100)
2005	13 103	100.0
2006	14 706	100.7
2007	16 090	103.0
2008	16 622	108.4
2009	15 567	111.2
2010	15 513	111.4
2011	16 131	112.4

Use the Export Price Index figures to deflate the export values and comment on the pattern shown by the 'real' export values.

16. Inditex Group, based in Spain, owns the fashion chain Zara. Sales figures for Inditex for 2006 to 2011, in millions of euros, are shown in the following table. Also shown are annual values for the Harmonized Index of Consumer Prices for Spain, based on 2005 = 100. Deflate the Inditex sales figures on the basis of 2005 constant euros, and comment on the firm's sales volumes in terms of deflated euros.

Year	Retail sales (€ million)	HICP (2005 = 100)
2006	8196	103.56
2007	9435	106.51
2008	10407	110.91
2009	11048	110.64
2010	12527	112.90
2011	13793	116.35

17. Annual salaries in England and Wales for school teachers on the bottom point of the 'leadership' cale were as shown in the following table. Use the CPI to deflate the salary data to constant values. Comment on the trend in school teachers' salaries in England and Wales as indicated by these data.

Year	Salary (£)	CPI (base year 2005)
2005	33 249	100.0
2006	34 083	102.3
2007	34 938	104.7
2008	35 794	108.5
2009	36 618	110.8
2010	37 461	114.5
2011	37 461	119.6

18. Data on quantities of three items sold in 2005 and 2013 are given here along with the sales prices of the items in 2005. Compute a weighted aggregate quantity index for 20139.

	Quantity sold		
Item	2005	2013	Price per unit 2005 (€)
A	350	300	18.00
B	220	400	4.90
C	730	850	15.00

19. A freight company handles four commodities for a particular distributor. Total shipments for the commodities in 2000 and 2013, measured in terms of standard containers, as well as the 2000 prices, are reported in the following table.

	Containers shipped		
Commodity	2000	2013	Price per shipment 2000 (€)
A	120	95	1200
B	86	75	1800
C	35	50	2000
D	60	70	1500

Calculate a weighted aggregate quantity index number for 2013 with a 2000 base. Comment on the growth or decline in shipments over the 2000–2013 period.

20. A car dealer reports the 2008 and 2013 sales for three models in the following table. Compute quantity relatives and use them to develop a weighted aggregate quantity index for 2013 using the two years of data.

	Sales		
Model	2008	2013	Mean price per sale 2008 (£)
Two-door hatchback	200	170	15 200
Two-door cabriolet	100	80	17 000
Four-door saloon	75	60	16800

21. A major manufacturing company reports the quantity and product value information for 2005 and 2013 in the table that follows. Compute a weighted aggregate quantity index for the data. Comment on what this quantity index means.

	Quantities		
Product	2005	2013	Value (£)
A	800	1200	30.00
B	600	500	20.00
C	200	500	25.00

Chapter 19: Index Numbers

Textbook Exercise Solutions:

1. a.

Item	Price Relative
A	103 = (7.75/7.50)
B	238 = (1500/630)

b. $I_{2013} = \dfrac{7.75 + 1500.00}{7.50 + 630.00}(100) = \dfrac{1507.75}{637.50}(100) = 237$

c. $I_{2013} = \dfrac{7.75(1500) + 1500.00(2)}{7.50(1500) + 630.00(2)}(100) = \dfrac{14,625.00}{12,510.00}(100) = 117$

d. $I_{2013} = \dfrac{7.75(1800) + 1500.00(1)}{7.50(1800) + 630.00(1)}(100) = \dfrac{15,450.00}{14,130.00}(100) = 109$

2. a. From the price relative we see the percentage increase was 32%.

b. Divide the current cost by the price relative and multiply by 100.

$$2005 \text{ cost} = \frac{€10.75}{132}(100) = €8.14$$

3. a/b

Quarter & Year	Price (£)	Price index (Q1 2010 = 100)	Price index (Q1 2011 = 100)
Q1 2010	229,422	100.0	98.4
Q2 2010	236,890	103.3	101.6
Q3 2010	239,188	104.3	102.5
Q4 2010	234,194	102.1	100.4
Q1 2011	233,259	101.7	100.0
Q2 2011	244,706	106.7	104.9
Q3 2011	237,535	103.5	101.8
Q4 2011	245,437	107.0	105.2
Q1 2012	239,309	104.3	102.6
Q2 2012	267,101	116.4	114.5

4. a. Price Relatives A = (6.00 / 5.45) 100 = 110

B = (5.95 / 5.60) 100 = 106

C = (6.20 / 5.50) 100 = 113

b. $I_{2013} = \dfrac{6.00 + 5.95 + 6.20}{5.45 + 5.60 + 5.50}(100) = 110$

c. $I_{2013} = \dfrac{6.00(150) + 5.95(200) + 6.20(120)}{5.45(150) + 5.60(200) + 5.50(120)}(100) = 109$

9% increase over the three-year period.

5.

Product	2008 Quantity	2008 Price	2013 Price	$P_{2008} \times Q_{2008}$	$P_{2013} \times Q_{2008}$
Beer	35,000	21.50	23.25	752,500	813,750
Wine	5,000	85.70	91.50	428,500	457,500
Soft drink	60,000	14.00	14.30	840,000	858,000
				2,021,000	2,129,250

$$I = 100 \times \frac{2{,}129{,}250}{2{,}021{,}000} = 105.4$$

6. $$I = \frac{0.19(500) + 1.80(50) + 4.20(100) + 13.20(40)}{0.15(500) + 1.60(50) + 4.50(100) + 12.00(40)}(100) = 104$$

Paasche index

7. $$I = \frac{15.90(2000) + 32.00(5000) + 17.40(6500) + 35.50(2500)}{10.50(2000) + 16.25(5000) + 12.20(6500) + 20.00(2500)}(100) = 170$$

8. $$I = \frac{15.90(4000) + 32.00(3000) + 17.40(7500) + 35.50(3000)}{10.50(4000) + 16.25(3000) + 12.20(7500) + 20.00(3000)}(100) = 164$$

9.

Item	Price Relative	Base Period Price	Base Period Usage	Weight	Weighted Price Relatives
A	150	22.00	20	440	66,000
B	90	5.00	50	250	22,500
C	120	14.00	40	560	67,200
				1250	155,700

$$I = \frac{155{,}700}{1250} = 125$$

10. a. Price Relatives A = (3.95 / 2.50) 100 = 158

B = (9.90 / 8.75) 100 = 113

C = (0.95 /0.99) 100 = 96

b.

Item	Price Relatives	Base-period Price	Quantity	Weight $P_{i0}Q_i$	Weighted Price Relatives
A	158	2.50	25	62.5	9875
B	113	8.75	15	131.3	14837
C	96	0 .99	60	59.4	5702
				253.2	30414

$$I = \frac{30414}{253.2} = 120$$

Cost of raw materials is up 20% for the chemical.

11.

Stock	Base-period Price	Current Price	Price Relatives	Quantity	Weight	Weighted Price Relatives
Euro Leisure	1.55	1.70	109.68	5000	7750	850020
UK Utilities	1.85	2.05	110.81	2000	3700	409997
Scand Gas	2.65	2.60	98.11	5000	13250	1299957.5
PQ Domestic	4.25	4.55	107.06	3000	12750	1365015
					37450	3924989.5

$$I = \frac{3924989.5}{37450} = 104.8$$

Portfolio up by nearly 5%

12.

Product	2005 Quantity	2005 Price	2009 Price	Price Relative	Weight	Weighted Price Relative
Beer	35,000	21.50	23.25	108.14	752,500	81,375,000
Wine	5,000	85.70	91.50	106.77	428,500	45,750,000
Soft						
	60,000	14.00	14.30	102.14	840,000	85,800,000
					2,021,000	212,925,000

$$I = \frac{212,925,000}{2,021,000} = 105.4$$

13.

Product	2009 Quantity	2009 Price	2013 Price	Price relative	Weight	Weighted price relative
A	2000	10.50	15.90	151.4	21,000	3,180,000
B	5000	16.25	32.00	196.9	81,250	16,000,000
C	6500	12.20	17.40	142.6	79,300	11,310,000
D	2500	20.00	35.50	177.5	50,000	8,875,000
					231,550	39,365,000

$$I = \frac{39,365,000}{231,550} = 170.0$$

14. a. Deflated wage rate for 1999 = $3.60 \times \frac{100}{165.2} = 2.179$ (£2.18)

Deflated wage rate for 2008 = $5.73 \times \frac{100}{217.7} = 2.632$ (£2.63)

b. % change in actual wage rates = $100 \times \frac{(6.08 - 3.60)}{3.60} = 68.9\%$

c. % change in deflated wage rates = $100 \times \frac{(2.507 - 2.179)}{2.179} = 15.1\%$

15.

Year	Exports (Swedish Krone, millions)	EPI (2005 = 100)	Exports deflated using EPI
2005	13,103	100.0	131.0
2006	14,706	100.7	146.0
2007	16,090	103.0	156.2
2008	16,622	108.4	153.3
2009	15,567	111.2	140.0
2010	15,513	111.4	139.3
2011	16,131	112.4	143.5

See table above. 2008 deflated value, for example, is: $16622 \times \frac{100.0}{108.4} = 153.3$

The 'real' export values show rises for two years up to 2007, then falls over the next three years. The 2011 figure is higher than 2010, but lower than the figure in 2006.

16.

Year	Retail sales (€ million)	HICP (2005 = 100)	Deflated retail sales (€ million, 2005 prices)
2006	8196	103.56	7914
2007	9435	106.51	8858
2008	10407	110.91	9383
2009	11048	110.64	9986
2010	12527	112.90	11096
2011	13793	116.35	11855

See table above. 2008 deflated value, for example, is: $10407 \times \frac{100}{110.91} = 9383$

Year-on-year increases of greater than 5% in every year. Particularly large rises 2006 – 2007 and 2009 – 2010, both over 11%.

17.

Year	Salary (£)	CPI (year 2005 = 100)	Deflated salary (£, 2005 values)
2005	33249	100	33249
2006	34083	102.3	33317
2007	34938	104.7	33370
2008	35794	108.5	32990
2009	36618	110.8	33049
2011	37461	114.5	32717

See table above. 2007 deflated value, for example, is: $34938 \times \frac{100}{104.7} = 33370$

Trend is slightly upward, year on year, for the first two years. There are then falls in two of the other three years. The 2011 figure is below the 2005 figure.

18. $I = \frac{300(18.00) + 400(4.90) + 850(15.00)}{350(18.00) + 220(4.90) + 730(15.00)}(100) = \frac{20,110}{18,328}(100) = 110$

19. $I = \frac{95(1200) + 75(1800) + 50(2000) + 70(1500)}{120(1200) + 86(1800) + 35(2000) + 60(1500)}(100) = 99$

Quantities are down slightly (by 1%).

20.

Model	Quantity Relatives	Base Quantity	Price (£)	Weight	Weighted Quantity Relatives
2-door hatchback	100(170/200) = 85	200	15,200	3,040,000	258,400,000
2-door cabriolet	100(89/100) = 80	100	17,000	1,700,000	136,000,000
4-door saloon	100(60/75) = 80	75	16,800	1,260,000	100,800,000
				6,000,000	495,200,000

$I = \frac{495,200,000}{6,000,000} = 83$

21. $I = \frac{1200(30) + 500(20) + 500(25)}{800(30) + 600(20) + 200(25)}(100) = 143$

Quantity is up 43%.

Supplementary Exercises:

22. Suppose that median selling prices for new single-family houses for the years 2006–2009 are as follows.

Year	Price (€000s)
2006	152.5
2007	161.0
2008	169.0
2009	175.2

a. Use 2006 as the base year and calculate a price index series for new single-family homes over this four-year period.
b. Use 2007 as the base year and calculate a price index series for new single-family homes over this four-year period.

23. The average selling prices for an apartment in Yorkshire, England for the years 2007 and 2008, by quarter, were as follows (Source: Nationwide Building Society).

Quarter & Year	Price (£)
Q1 2007	132 383
Q2 2007	133 063
Q3 2007	133 762
Q4 2007	131 401
Q1 2008	121 205
Q2 2008	121 647
Q3 2008	116 224
Q4 2008	121 916

a. Use Q1 2007 as the base period and construct a price index for apartments in Yorkshire over this two-year period.
b. Use Q1 2008 as the base year and construct a price index for apartments in Yorkshire over this two-year period.

24. Brahms & Liszt Beverages provides a complete line of beer, wine and soft drink products for distribution through retail outlets. Unit price data for 2004 and 2009, and quantities sold in cases for 2004, follow.

		Unit price (€)	
	2004 Quantity (cases)	**2004**	**2009**
Beer	30,000	15.00	16.25
Wine	15,000	60.00	64.00
Soft drink	60,000	9.80	10.00

Compute 2009 price relatives for the three types of product, using 2004 as the base period.

25. Refer to the data for exercise 24. Compute a weighted aggregate index for the Brahms & Liszt Beverage sales in 2009, with 2004 as the base period.

26. Refer to the data for exercise 24. Use a weighted average of price relatives to show that this method provides the same index as the weighted aggregate method.

27. Boran Stockbrokers selects four stocks for the purpose of developing its own index of stock market behaviour. Prices per share for a 2001 base period, January 2003, and March 2003 follow. Base-year quantities are set on the basis of historical volumes for the four stocks. Use the 2001 base period to compute the Boran index for January 2003 and March 2003. Comment on what the index tells you about what was happening in the stock market.

			Price per share ($)		
Stock	**Industry**	**2001 Quantity**	**2001 Base**	**Jan 2003**	**Mar 2003**
A	Oil	100	31.50	22.75	22.50
B	Computer	150	65.00	49.00	47.50
C	Steel	75	40.00	32.00	29.50
D	Property	50	18.00	6.50	3.75

28. Compute the price relatives for the four stocks making up the Boran index in exercise 27. Use the weighted average of price relatives approach to compute the January 2003 and March 2003 Boran indexes.

29. Consider the following price relatives and quantity information for grain production in Iowa (*Statistical Abstract of the United States*, 2002).

Product	1991 Quantities (millions of bushels)	Base Price per Bushel ($)	1991-2001 Price Relatives
Corn	1427	2.30	91
Soybeans	350	5.51	78

What is the 2001 weighted aggregate price index for the Iowa grains?

30. Fresh fruit price and quantity data for the years 1988 and 2001 follow. Quantity data reflect per capita consumption in kilograms and prices are per kilogram.
    a. Compute a price relative for each product.
    b. Compute a weighted aggregate price index for fruit products. Comment on the change in fruit prices over the 13-year period.

Fruit	1988 Per Capita Consumption (kg)	1988 Price (€ per kg)	2001 Price (€ per kg)
Bananas	24.3	0.41	0.51
Apples	19.9	0.71	0.87
Oranges	13.9	0.56	0.71
Pears	3.2	0.64	0.98

31. A family has kept records of their expenditures on music, cinema and theatre. Below are details for 2005 and 2009.

	Unit price (€)		Quantities per month	
	2005	2009	2005	2009
Purchase of CDs	15.50	13.50	7	10
Purchase of DVDs	22.00	21.00	4	7
Visits to concerts	40.00	55.00	2	1
Visits to cinema	9.50	12.50	3	2

    a. Construct a weighted aggregate price index value for 2009 for this set of items, with 2005 as the base period, using the Laspeyres method.
    b. Construct a weighted aggregate price index value for 2009 for this set of items, with 2005 as the base period, using the Paasche method.
    c. Comment on any difference between the answers to a and b.

32. Suppose that average hourly wages for manufacturing workers in 1985 were €7.27; in 2004, they were €14.36. Suppose further that the CPI in 1985 was 82.4 (base period 1990); in 2004 it was 172.2.
    a. Deflate the hourly wage rates in 1985 and 2004 to find the real wage rates at 1990 price levels.
    b. What is the percentage change in nominal hourly wages from 1985 to 2004?
    c. What is the percentage change in real wages from 1985 to 2004?

33. On 1 April 1999 when the UK National Minimum Wage was introduced, the minimum rate for workers aged 22 and over was £3.60 per hour. In October 2008, the rate was £5.73 per hour. The RPI in April 1999 was 165.2; in October 2008 it was 217.7 (1 January 1987 = 100).
    a. Deflate the hourly wage rates in 1999 and 2008 to find the real wage rates at January 1987 price levels.
    b. What is the percentage change from 1999 to 2008 in actual minimum hourly wage rates?
    c. What is the percentage change from 1999 to 2008 in real wage rates at January 1987 price levels?

34. Average hourly wages for workers in service industries for the four years from 2002 to 2005 are reported here. Use the Consumer Price Index information to deflate the wages series. Calculate the percentage increase in real wages and salaries from 2002 to 2005.

Year	**Hourly wages (€)**	**CPI (1990 base)**
2002	8.13	163.0
2003	8.45	166.6
2004	8.65	172.2
2005	9.15	177.1

35. Statistics Sweden report the following exports of motor vehicles from Sweden over the period 2000 to 2008, and a relevant Export Price Index (EPI) for motor vehicles.

    Use the Export Price Index figures to deflate the export values and comment on the pattern shown by the 'real' export values.

Year	Exports (Swedish Krone, millions)	EPI (1990 = 100)
2000	97.69	123.7
2001	100.26	128.5
2002	100.14	128.5
2003	113.77	121.7
2004	128.95	115.1
2005	133.65	115.0
2006	146.55	116.4
2007	154.27	116.9
2008	143.02	119.1

36. Marks & Spencer UK retail sales figures for 2000 to 2008, in millions of pounds sterling, are shown in the following table. Also shown are annual values for the RPI, based at 1 January 1987. Deflate the Marks & Spencer sales figures on the basis of 1987 constant pounds sterling, and comment on the firm's sales volumes in terms of deflated pounds.

Year	Retail sales (£ million)	RPI (1 Jan 1987 = 100)
2000	6483	170.3
2001	6293	173.4
2002	6575	176.2
2003	7027	181.3
2004	7160	186.7
2005	7035	192.0
2006	7275	198.1
2007	7978	206.6
2008	8309	214.8

37. The U.S Census Bureau reported the following total manufacturing shipments for the three years from 1999 to 2001.

Year	**Manufacturing shipments ($ billions)**
1999	4032
2000	4218
2001	3971

a. The CPI for 1999–2001 is given below. Use this information to deflate the manufacturing shipments series and comment on the pattern of manufacturers' shipments in terms of constant dollars.

b. The following Producer Price Indexes (finished consumer goods) are for 1999 to 2001, with 1982 as the base year. Use the PPI to deflate the series.

c. Do you feel that the CPI or the PPI is more appropriate to use as a deflator for manufacturing shipments?

Year	CPI (1982-1984 Base)	PPI (1982 = 100)
1999	166.6	133.0
2000	172.2	138.0
2001	177.1	140.7

38. Dooley Retail Outlets' total retail sales volumes for selected years since 1982 is shown in the following table. Also shown is the CPI with the index base of 1982–1984. Deflate the sales volume figures on the basis of 1982–1984 constant dollars, and comment on the firm's sales volumes in terms of deflated dollars.

Year	Retail Sales ($)	CPI (1982-1984 Base)
1982	380,000	96.5
1987	520,000	113.6
1992	700,000	140.3
1997	870,000	160.5
2002	940,000	179.9

39. Starting faculty salaries for assistant professors of business administration at a major university follow. Use the CPI to deflate the salary data to constant euros. Comment on the trend in salaries in higher education as indicated by these data.

Year	Salary (€)	CPI (1980 base)
1970	14,000	38.8
1975	17,500	53.8
1980	23,000	82.4
1985	37,000	107.6
1990	53,000	130.7
1995	65,000	152.4
2000	80,000	172.2

40. The five-year historical prices per share for a particular stock and the Consumer Price Index with a 1982–1984 base period follow.

Year	Price per Share ($)	CPI
1998	51.00	163.0
1999	54.00	166.6
2000	58.00	172.2
2001	59.50	177.1
2002	59.00	179.9

Deflate the stock price series and comment on the investment aspects of this stock.

Chapter 19: Index Numbers

Supplementary Exercises Solutions:

22. a/b.

	Price Index	
Year	2006 Base	2007 Base
2006	100.0	94.7
2007	105.6	100.0
2008	110.8	105.0
2009	114.9	108.8

23. a/b

Quarter & Year	Price (£)	Price index (Q1 2007 = 100)	Price index (Q1 2008 = 100)
Q1 2007	132,383	100.0	109.2
Q2 2007	133,063	100.5	109.8
Q3 2007	133,762	101.0	110.4
Q4 2007	131,401	99.3	108.4
Q1 2008	121,205	91.6	100.0
Q2 2008	121,647	91.9	100.4
Q3 2008	116,224	87.8	95.9
Q4 2008	121,916	92.1	100.6

24. Price relatives for 2009 (2004 base)

Beer	100(16.25/15.00) = 108.33
Wine	100(64.00/60.00) = 106.67
Soft drink	100(10.00/9.80) = 102.04

25. $$I_{2009} = \frac{16.25(30,000) + 64.00(15,000) + 10.00(60,000)}{15.00(30,000) + 60.00(15,000) + 9.80(60,000)}(100) = 105.65$$

26. Weights (2004 base)

Beer	30,000 x 15.00 = 450,000
Wine	15,000 x 60.00 = 900,000
Soft drink	60,000 x 9.80 = 588,000

$$I_{2009} = \frac{(108.33 \times 450,000) + (106.67 \times 900,000) + (102.04 \times 588,000)}{(450,000 + 900,000 + 588,000)} = 105.65$$

27. $$I_{\text{Jan}} = \frac{22.75(100) + 49(150) + 32(75) + 6.5(50)}{31.50(100) + 65(150) + 40(75) + 18(50)}(100) = 73.5$$

$$I_{\text{Mar}} = \frac{22.50(100) + 47.5(150) + 29.5(75) + 3.75(50)}{31.50(100) + 65(150) + 40(75) + 18(50)}(100) = 70.1$$

Market was down compared to 2001.

28.

	Price Relatives		Weights
	Jan	Mar	
Oil	(22.75 / 31.50) (100) = 72.2	71.4	100 x 31.50 = 3150
Computer	(49.00 / 65.00) (100) = 75.4	73.1	150 x 65.00 = 9750
Steel	(32.00 / 40.00) (100) = 80.0	73.8	75 x 40.00 = 3000
Property	(6.50 / 18.00) (100) = 36.1	20.8	50 x 18.00 = 900

$$I_{\text{Jan}} = \frac{(72.2 \times 3150) + (75.4 \times 9750) + (80.0 \times 3000) + (36.1 \times 900)}{(3150 + 9750 + 3000 + 9000)} = 73.5$$

$$I_{\text{Mar}} = \frac{(71.4 \times 3150) + (73.1 \times 9750) + (73.8 \times 3000) + (20.8 \times 900)}{(3150 + 9750 + 3000 + 9000)} = 70.1$$

29.

Product	Relatives	Base Price	Quantity	Weight	Weighted Price Relatives
Corn	91	2.30	1427	3282.1	298,671.1
Soybeans	78	5.51	350	1928.5	150,423.0
				5210.6	449,094.1

$$I = \frac{449,094.1}{5210.6} = 86.2$$

30. a.

Fruit	Price Relatives
Bananas	(0.51/0.41) (100) = 124.4
Apples	(0.87/0.71) (100) = 122.5
Oranges	(0.71/0.56) (100) = 126.8
Pears	(0.98/0.64) (100) = 153.1

b.

	Weights ($P_{io}Q_{io}$)	Price Relative	Weighted price relative
	9.963	124.4	1239.3
	14.129	122.5	1731.3
	7.784	126.8	986.9
	2.048	153.1	313.6
Totals	33.924		4271.1

$$I = \frac{4271.1}{33.924} = 125.9$$

Fruit prices have increased by 25.9% over the 13-year period according to the index.

31. a. $I_{2009} = \dfrac{(13.50\times 7)+(21.00\times 4)+(55.00\times 2)+(12.50\times 3)}{(15.50\times 7)+(22.00\times 4)+(40.00\times 2)+(9.50\times 3)}(100) = \dfrac{326}{305}(100) = 106.88$

b. $I_{2009} = \dfrac{(13.50\times 10)+(21.00\times 7)+(55.00\times 1)+(12.50\times 2)}{(15.50\times 10)+(22.00\times 7)+(40.00\times 1)+(9.50\times 2)}(100) = \dfrac{368}{362}(100) = 101.66$

c. Paasche index indicates a smaller increase in prices overall than the Laspeyres index. CDs and DVDs have fallen in price, whereas visits to cinema and theatre have increased in price. The family has shifted its expenditure towards CDs and DVDs, and away from visits to cinema and theatre.

32. a. Deflated 1985 wages: $\dfrac{€7.27}{82.4}(100) = €8.82$

Deflated 2004 wages: $\dfrac{€14.36}{172.2}(100) = €8.34$

b. $\dfrac{14.36}{7.27}(100) = 197.5$. The percentage increase in nominal wages is 97.5%.

c. $\dfrac{8.34}{8.82}(100) = 94.6$. The change in real wages is a decrease of 5.4%.

33. a. Deflated wage rate for 1999 = $3.60\times\dfrac{100}{165.2} = 2.179$ (£2.18)

Deflated wage rate for 2008 = $5.73\times\dfrac{100}{217.7} = 2.632$ (£2.63)

b. % change in actual wage rates = $100\times\dfrac{(5.73-3.60)}{3.60} = 59.2\%$

c. % change in deflated wage rates = $100\times\dfrac{(2.632-2.179)}{2.179} = 20.8\%$

34.

2002	8.13 (100/163.0)	=	4.99
2003	8.45 (100/166.6)	=	5.07
2004	8.65 (100/172.2)	=	5.02
2005	9.15 (100/177.1)	=	5.17

$(100)\dfrac{5.17}{4.99} = 103.6$. The increase in real wages and salaries from 2002 to 2005 is 3.6%.

35.

Year	Exports (Swedish Krone, millions)	EPI (1990 = 100)	Exports deflated using EPI
2000	97.69	123.7	79.0
2001	100.26	128.5	78.0
2002	100.14	128.5	77.9
2003	113.77	121.7	93.5
2004	128.95	115.1	112.0
2005	133.65	115.0	116.2
2006	146.55	116.4	125.9
2007	154.27	116.9	132.0
2008	143.02	119.1	120.1

See table above. 2000 deflated value, for example, is: $97.69 \times \frac{100}{123.7} = 79.0$

Real export values have same general pattern as the original values because the EPI has shown relatively modest changes. The real export values show marked rises 2002-2003, 2003-2004, 2005-2006, 2006-2007. Large fall 2007-2008.

36.

Year	Retail sales (£ million)	RPI (1 Jan 87 = 100)	Deflated retail sales (£ million, Jan 87 prices)
2000	6483	170.3	3806.8
2001	6293	173.3	3631.3
2002	6575	176.2	3731.6
2003	7027	181.3	3875.9
2004	7160	186.7	3835.0
2005	7035	192.0	3664.1
2006	7275	198.1	3672.4
2007	7978	206.6	3861.6
2008	8309	214.8	3868.2

See table above. 2000 deflated value, for example, is: $6483 \times \frac{100}{170.3} = 3806.8$

Sales volume falls noticeably 2000–2001 and 2004–2005, rises 2001–2002, 2002–2003 and 2006–2007. Sales volume in 2008 round about the 2003 and 2004 levels.

37. a.

1999	4032 (100/166.6)	=	2420
2000	4218 (100/172.2)	=	2449
2001	3971 (100/177.1)	=	2242

Manufacturers' shipments have decreased slightly in constant dollars when deflated using the CPI.

b.

1999	4032 (100/133.0)	=	3032
2000	4218 (100/138.0)	=	3057
2001	3971 (100/140.7)	=	2822

c. The PPI is a better deflator since manufacturing shipments reflect prices paid by manufacturers.

38.

Year	Retail Sales ($)	CPI	Deflated Retail Sales ($)
1982	380,000	96.5	393,782
1987	520,000	113.6	457,746
1992	700,000	140.3	498,931
1997	870,000	160.5	542,056
2002	940,000	179.9	522,513

In terms of constant dollars, the firm's sales were increasing moderately until 2002, then they decreased a little.

39. Salaries in constant (1980) euros are computed as follows:

1970 €14,000 (100 / 38.8) = €36,082
1975 €17,500 (100 / 53.8) = €32,528
1980 €23,000 (100 / 82.4) = €27,913
1985 €37,000 (100 / 107.6) = €34,387
1990 €53,000 (100 / 130.7) = €40,551
1995 €65,000 (100 / 152.4) = €42,651
2000 €80,000 (100 / 172.2) = €46,458

In constant value terms, real starting salaries have increased about 29% over this period.

40. The stock market prices in constant (1982–84) dollars are computed as follows:

1998 $51.00 (100 / 163.0) = $31.29
1999 $54.00 (100 / 166.6) = $32.41
2000 $58.00 (100 / 172.2) = $33.68
2001 $59.50 (100 / 177.1) = $33.60
2002 $59.00 (100 / 179.9) = $32.80

The value of the stock, in real dollars, is only slightly more in 2002 than it is in 1998. Of course, if the stock paid a high dividend it may still have been a good investment over this period.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Twenty

Statistical Methods for Quality Control

Textbook Exercises (1-15)
Textbook Exercise Solutions
Supplementary Exercises (16-30)
Supplementary Exercise Solutions

Chapter 20: Statistical Methods for Quality Control

Textbook Exercises:

1. A process that is in control has a mean of $\pi = 2.5$ and a standard deviation of $\sigma = 0.8$.

a. Construct an $\overline{X}$ chart if samples of size 4 are to be used.
b. Repeat part (a) for samples of size 8 and 16.
c. What happens to the limits of the control chart as the sample size is increased? Discuss why this is reasonable.

2. Twenty-five samples, each of size 5, were selected from a process that was in control. The sum of all the data collected was 307.3 kg.

a. What is an estimate of the process mean (in terms of kg per unit) when the process is in control?
b. Develop the control chart for this process if samples of size 5 will be used. Assume that the process standard deviation is 0.5 when the process is in control, and that the mean of the process is the estimate developed in part (a).

3. Twenty-five samples of 100 items each were inspected when a process was considered to be operating satisfactorily. In the 25 samples, a total of 135 items were found to be defective.

a. What is an estimate of the proportion defective when the process is in control?
b. What is the standard error of the proportion if samples of size 100 will be used for statistical process control?
c. Compute the upper and lower control limits for the control chart.

4. A process sampled 20 times with a sample of size 8 resulted in $\bar{\bar{x}} = 28.5$ and $\bar{R} = 1.6$.

Compute the upper and lower control limits for the $\overline{X}$ and R charts for this process.

5. Temperature is used to measure the output of a production process. When the process is in control, the mean of the process is $\mu = 128.5$ and the standard deviation is $\sigma = 0.4$.

a. Construct an $\overline{X}$ chart if samples of size 6 are to be used.
b. Is the process in control for a sample providing the following data?

128.8	128.2	129.1	128.7	128.4	129.2

c. Is the process in control for a sample providing the following data?

129.3	128.7	128.6	129.2	129.5	129.0

6. Thirty samples of size 3 are taken of an electrical component. The details are given below.

Sample			
1	207	194	201
2	204	191	203
3	198	201	196
4	195	199	181
5	199	221	218
6	200	200	207
7	222	195	205
8	215	186	181
9	188	199	191
10	171	200	201
11	200	201	192
12	204	207	194
13	191	215	200
14	201	204	200
15	198	212	206
16	231	188	223
17	202	210	219
18	187	190	205
19	194	196	207
20	196	199	190
21	198	208	200
22	185	206	201
23	209	225	226
24	199	199	208
25	214	203	195
26	208	202	199
27	202	205	185
28	195	210	199
29	206	191	200
30	205	198	202

a. Compute trial limits for $\bar{x}$ and R.
b. Hence determine if the process is in control.
c. If not, compute revised control limits after eliminating samples that appear to be affected by 'assignable' causes.

7. The Guttman Tyre and Rubber Company periodically tests its tyres for tread wear under simulated road conditions. To study and control the manufacturing process, 20 samples, each containing three radial tyres, were chosen from different shifts over several days of operation, with the following results. Assuming that these data were collected when the manufacturing process was believed to be operating in control, develop the *R* and $\bar{X}$ charts.

Sample	Tread wear (mm)		
1	8	11	7
2	7	5	9
3	6	8	9
4	4	6	5
5	10	7	9
6	10	11	9
7	5	4	7
8	8	7	7
9	10	9	8
10	7	4	8
11	7	8	10

Sample	Tread wear (mm)		
12	6	5	6
13	4	6	8
14	11	9	4
15	5	6	7
16	8	11	8
17	7	9	8
18	10	7	8
19	5	7	7
20	6	9	7

8. A company is concerned with monitoring the pH value of a liquid. Measurements are taken at intervals, three times per day so that over a 24 hour data period we have data as follows:

	Measurement				Measurement		
Sample	1	2	3	Sample	1	2	3
1	6.0	5.8	6.1	13	6.1	6.9	7.4
2	5.2	6.4	6.9	14	6.2	5.2	6.8
3	5.5	5.8	5.2	15	4.9	6.6	6.6
4	5.0	5.7	6.5	16	7.0	6.4	6.1
5	6.7	6.5	5.5	17	5.4	6.5	6.7
6	5.8	5.2	5.0	18	6.6	7.0	6.8
7	5.6	5.1	5.2	19	4.7	6.2	7.1
8	6.0	5.8	6.0	20	6.7	5.4	6.7
9	5.5	4.9	5.7	21	6.8	6.5	5.2
10	4.3	6.4	6.3	22	5.9	6.4	6.0
11	6.2	6.9	5.0	23	6.7	6.3	4.6
12	6.7	7.1	6.2	24	7.4	6.8	6.3

Compute R and $\overline{X}$ charts for the data and hence determine if the underlying production process is in control.

9. Samples of size 50 are randomly selected each day from a continuous process. The number of defectives observed in the first 28 samples is shown below.

Sample	Number of Defectives	Sample	Number of Defectives
1	5	**15**	6
2	7	**16**	9
3	11	**17**	10
4	9	**18**	11
5	14	**19**	13
6	21	**20**	30
7	25	**21**	26
8	18	**22**	13
9	10	**23**	8
10	8	**24**	23
11	18	**25**	34
12	19	**26**	25
13	6	**27**	18
14	8	**28**	12

a. Compute $\overline{p}$ 1 and the trial control limits.

b. Indicate which, if any, observations fall outside the trial limits.

10. For an acceptance sampling plan with $n = 25$ and $c = 0$, find the probability of accepting a lot that has a detect rate of 2 per cent. What is the probability of accepting the lot it the detect rate is 6 per cent?

11. Consider an acceptance sampling plan with $n = 20$ and $c = 0$. Compute the producer's risk for each of the following cases.

a. The lot has a detect rate of 2 per cent.
b. The lot has a detect rate of 6 per cent.

12. A production company is considering two possible plans for the acceptance sampling of some raw material. Both are attribute inspection plans, the first specifying sample size 10 and acceptance if the number of substandard items is no greater than 1, and the second specifying sample size 25 and acceptance if the number of substandard items is no greater than 3. Plot on the same graph the operating characteristic for each plan. If the production process can

work well on raw material 5 per cent substandard, but cannot work if the proportion gets close to 15 per cent, which plan should the company choose?

13. Refer to the KALI problem presented in this section. The quality control manager requested a producer's risk of 0.10 when p_0 was 0.03 and a consumer's risk of 0.20 when p_1 was 0.15. Consider the acceptance sampling plan based on a sample size of 20 and an acceptance number of 1. Answer the following questions.

a. What is the producer's risk for the $n = 20$, $c = 1$ sampling plan?
b. What is the consumer's risk for the $n = 20$, $c = 1$ sampling plan?
c. Does the $n = 20$, $c = 1$ sampling plan satisfy the risks requested by the quality control manager? Discuss.

14. A domestic manufacturer of watches purchases quartz crystals from a Swiss firm. The crystals are shipped in lots of 1000. The acceptance sampling procedure uses 20 randomly selected crystals.

a. Construct operating characteristic curves for acceptance numbers of 0, 1 and 2.
b. It p_0 is 0.01 and $p_1 = 0.08$, what are the producer's and consumer's risks for each sampling plan in part (a)?

15. A company wishes to design and implement a single attribute sampling plan so that a good lot with a defective rate of 2 per cent will be accepted 95 per cent of the time while a bad lot with a defective rate of 10 per cent will be accepted 15 per cent of the time. How would you advise the company?

Textbook Exercises Solutions:

1. a. For $n = 4$

$$\text{UCL} = \mu + 3(\sigma/\sqrt{n}) = 12.5 + 3(.8/\sqrt{4}) = 13.7$$
$$\text{LCL} = \mu - 3(\sigma/\sqrt{n}) = 12.5 - 3(.8/\sqrt{4}) = 11.3$$

b. For $n = 8$

$$\text{UCL} = \mu + 3(.8/\sqrt{8}) = 13.35$$
$$\text{LCL} = \mu - 3(.8/\sqrt{8}) = 11.65$$

For $n = 16$

$$\text{UCL} = \mu + 3(.8/\sqrt{16}) = 13.10$$
$$\text{LCL} = \mu - 3(.8/\sqrt{16}) = 11.90$$

c. UCL and LCL become closer together as *n* increases. If the process is in control, the larger samples should have less variance and should fall closer to 12.5.

2. a. $\mu = \dfrac{677.5}{25(5)} = 5.42$

b. $$\text{UCL} = \mu + 3(\sigma/\sqrt{n}) = 5.42 + 3(.5/\sqrt{5}) = 6.09$$
$$\text{LCL} = \mu - 3(\sigma/\sqrt{n}) = 5.42 - 3(.5/\sqrt{5}) = 4.75$$

3. a. $p = \dfrac{135}{25(100)} = 0.0540$

b. $\sigma_p = \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.0540(0.9460)}{100}} = 0.0226$

c. $$\text{UCL} = p + 3\sigma_{\bar{p}} = 0.0540 + 3(0.0226) = 0.1218$$
$$\text{LCL} = p - 3\sigma_{\bar{p}} = 0.0540 - 3(0.0226) = -0.0138$$

Use LCL = 0

4. R Chart:

$$\text{UCL} = \bar{R}D_4 = 1.6(1.864) = 2.98$$
$$\text{LCL} = \bar{R}D_3 = 1.6(0.136) = 0.22$$

$\bar{x}$ Chart:

$$\text{UCL} = \bar{\bar{x}} + A_2\bar{R} = 28.5 + 0.373(1.6) = 29.10$$
$$\text{LCL} = \bar{\bar{x}} - A_2\bar{R} = 28.5 - 0.373(1.6) = 27.90$$

5. a. $$\text{UCL} = \mu + 3(\sigma/\sqrt{n}) = 128.5 + 3(.4/\sqrt{6}) = 128.99$$

$$LCL = \mu - 3(\sigma / \sqrt{n}) = 128.5 - 3(.4 / \sqrt{6}) = 128.01$$

b. $\bar{x} = \Sigma x_i / n = \frac{772.4}{6} = 128.73$ in control

c. $\bar{x} = \Sigma x_i / n = \frac{774.3}{6} = 129.05$ out of control

6. a./b./c. $\bar{\bar{x}} = 201.27$ = Estimate of Process Mean2

n = 3

$$\sigma = 10.36 = \sqrt{\frac{\sum_{i=1}^{30}\sum_{j=1}^{3} x_{ij}^2 - \frac{\left[\left(\sum_{i=1}^{30}\sum_{j=1}^{3} x_{ij}\right)^2\right]}{90}}{89}}$$

3Warning lines $\bar{\bar{x}} \pm \frac{2\sigma}{\sqrt{n4}}$

correspond with limits (189.31, 213.23)

Similarly

5Action lines $\bar{\bar{x}} \pm \frac{3\sigma}{\sqrt{n6}}$

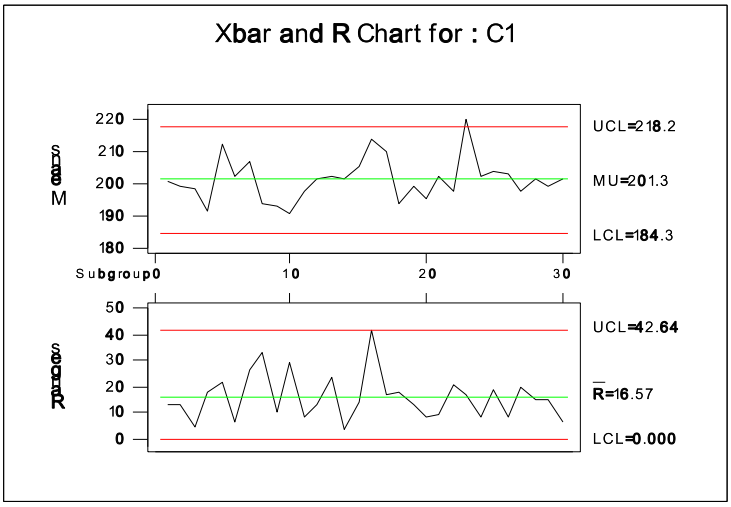
correspond with limits (183.33, 219.21)

The 23rd sample mean can be shown to fall just outside the upper Action line ($\bar{x}_{23} = 220$)7

If we drop this sample from the analysis we obtain the revised intervals:

Warning limits (189.29, 211.95)
Action limits (183.63, 217.61)

Corresponding MINITAB outputs are as follows:



7.

Sample number		Observations		$\bar{x}_j$	R_j
1	8	11	7	8.67	4
2	7	5	9	7.00	4
3	6	8	9	7.67	3
4	4	6	5	5.00	2
5	10	7	9	8.67	3
6	10	11	9	10.00	2
7	5	4	7	5.33	3
8	8	7	7	7.33	1
9	10	9	8	9.00	2
10	7	4	8	6.33	4
11	7	8	10	8.33	3
12	6	5	6	5.67	1
13	4	6	8	6.00	4
14	11	9	4	8.00	7
15	5	6	7	6.00	2
16	8	11	8	9.00	3
17	7	9	8	8.00	2
18	10	7	8	8.33	3
19	5	7	7	6.33	2
20	6	9	7	7.33	3

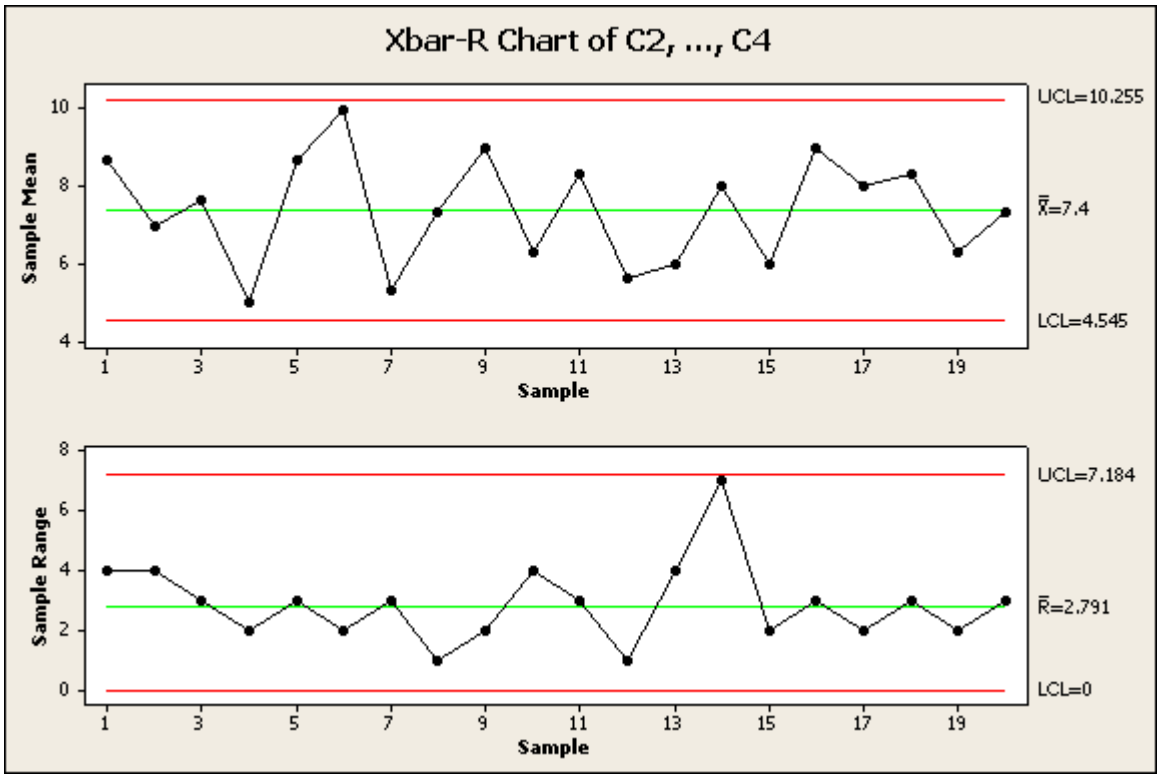
$\bar{R} = 2.79$ and $\bar{\bar{x}} = 7.40$

R Chart:

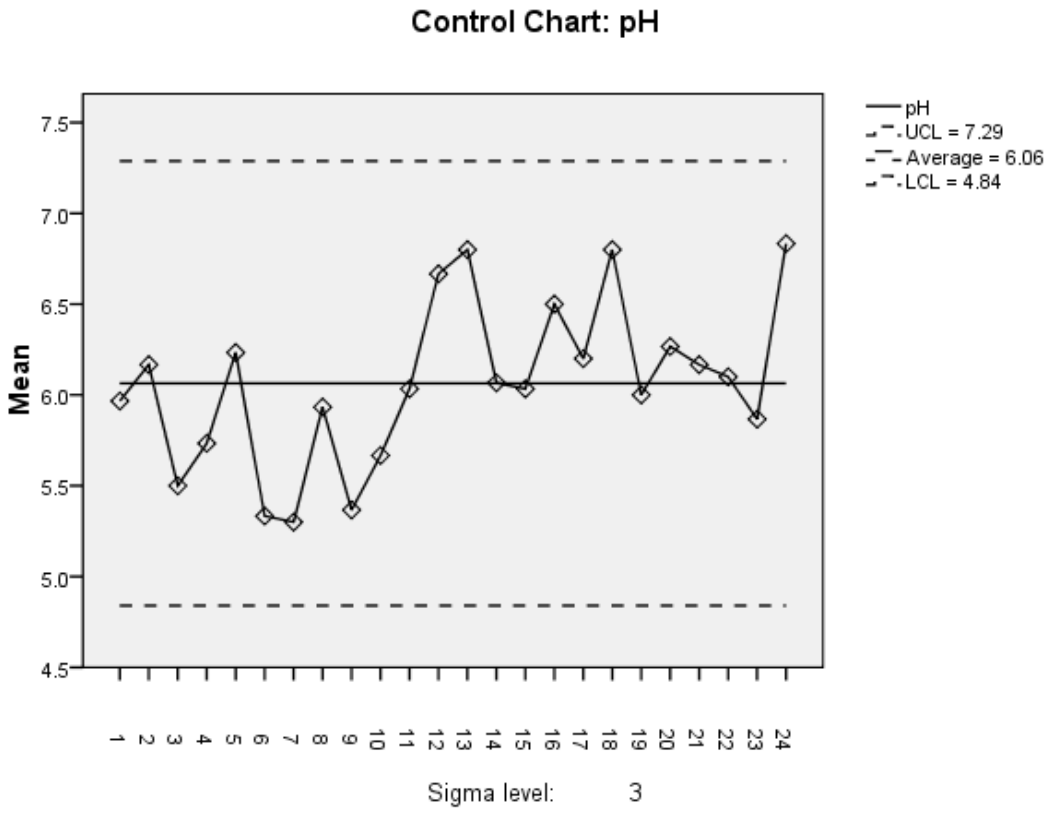
$$\text{UCL} = \bar{R}D_4 = 2.79(2.575) = 7.18$$
$$\text{LCL} = \bar{R}D_3 = 2.90(0) = 0$$

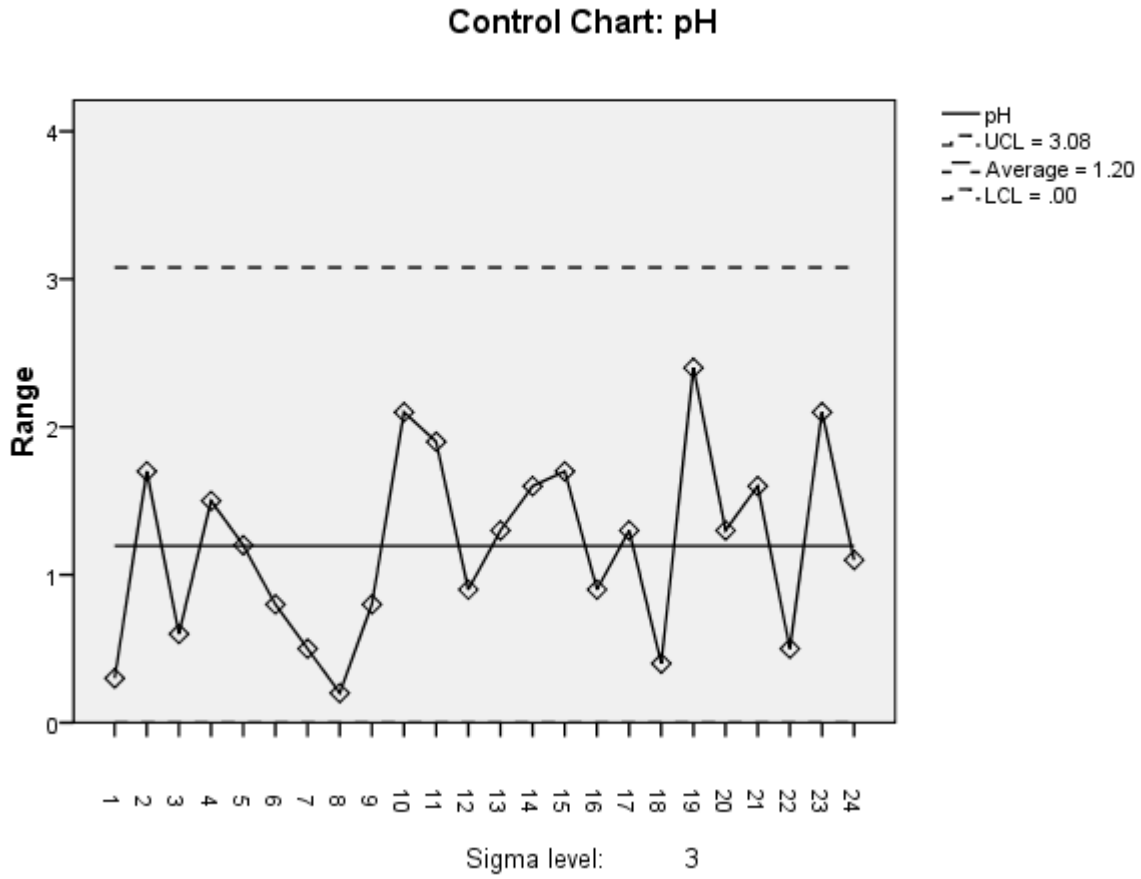
$\bar{x}$ Chart:

$$\text{UCL} = \bar{\bar{x}} + A_2\bar{R} = 7.40 + 1.023(2.79) = 10.26$$
$$\text{LCL} = \bar{\bar{x}} - A_2\bar{R} = 7.40 - 1.023(2.79) = 4.54$$



8. From the SPSS output below it can be seen that the process is in control in respect of both the $\bar{x}$ and R charts:





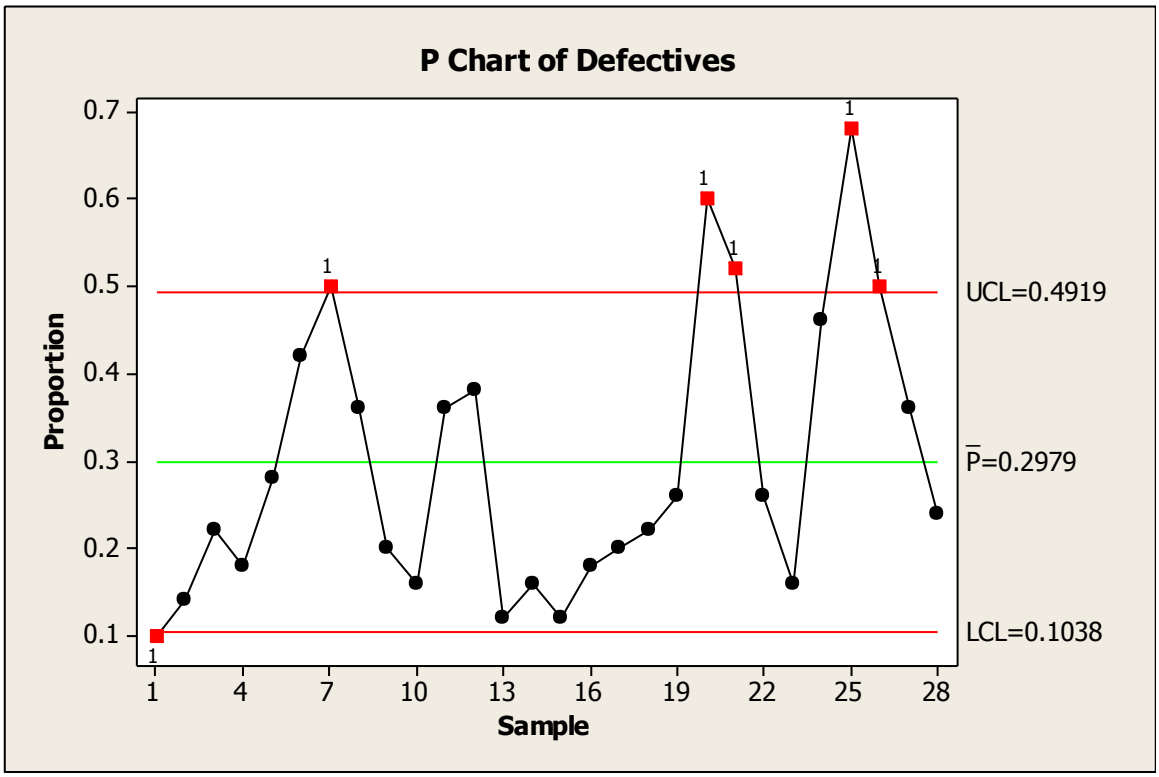
9. a. 8

$\overline{p} = 0.298$, n = 50

Control limits $\overline{p} \pm 3\sqrt{\frac{\overline{p}\,(1-\overline{p}\,)}{n}} = 0.298 \pm 0.194 = (0.104, 0.492)$

9

Observations 1, 7, 20, 21, 25, 26 fall outside these.



10. $$p(x) = \frac{n!}{x!\,(n-x)!}\pi^{x}(1-\pi)^{(n-x)}$$

When π = .02, the probability of accepting the lot is

$$p(0) = \frac{25!}{0!\,(25-0)!}0.02^{0}(1-0.02)^{25} = 0.6035$$

When π = .06, the probability of accepting the lot is

$$p(0) = \frac{25!}{0!\,(25-0)!}0.06^{0}(1-0.06)^{25} = 0.2129$$

11. a. Using binomial probabilities with n = 20 and p_0 = .02.

P (Accept lot) = p (0) = .6676

Producer's risk: α = 1 - .6676 = .3324

b. P (Accept lot) = p (0) = .2901

Producer's risk: α = 1 - .2901 = .7099

12. Suppose $p_0 = .01$ (i.e. 1% defective) , then under the first plan $n = 10$ and $c = 1$and we have

P (Accept lot) = p (0) + p (1) = 0.9044 + 0.0913= 0.9957

Producer's risk: $\alpha = 1 - .9957 = .0043$

Similarly if $p_0 = .01$, under the second plan $n = 25$ and $c = 3$ we have:

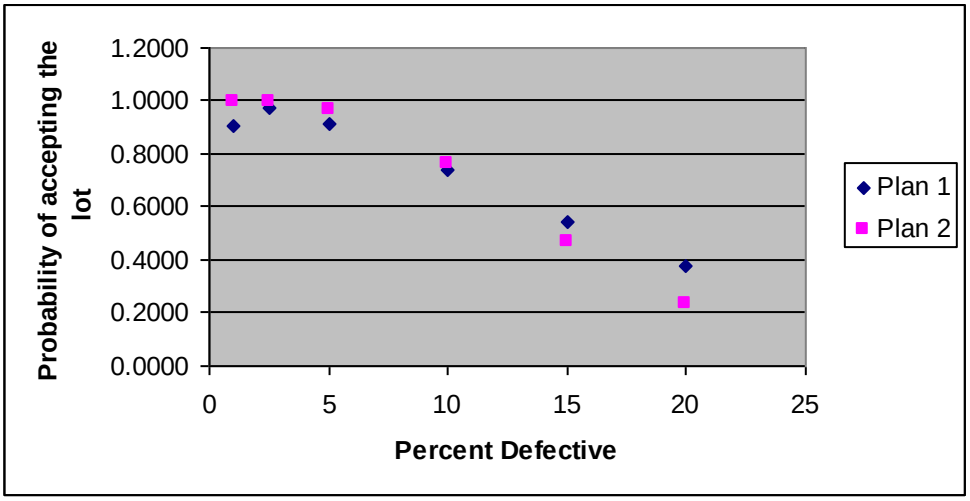
P (Accept lot) = p (0) + p (1) + p(2) + p(3) = .9999

Producer's risk: $\alpha = 1 - .9999 = .0001$

Allowing for percentage defective ($100p_0\%$) ranging from 1% to 20% in the above calculations we obtain the acceptance lot probabilities tabulated below:

	P(Accept lot)	
% Defective	Plan 1	Plan 2
1	0.9044	0.9999
2.5	0.9754	0.9968
5	0.9139	0.9659
10	0.7361	0.7636
15	0.5443	0.4711
20	0.3758	0.2340

Corresponding operating characteristic curves are as follows:



As we would like the lot acceptance probability to be large for $p_0 = .05$ and small for $p_0 = .15$ it is clear that the second plan is better on both counts. This is not surprising given the larger sample size involved.

13. a. Using binomial probabilities with $n = 20$ and $p_0 = .03$.

P(Accept lot) = p (0) + p (1)

= .5438 + .3364 = .8802

Producer's risk: $\alpha = 1 - .8802 = .1198$

b. With $n = 20$ and $p_1 = .15$.

P(Accept lot) = p (0) + p (1)

= .0388 + .1368 = .1756

Consumer's risk: $\beta = .1756$

c. The consumer's risk is acceptable; however, the producer's risk associated with the $n = 20$, $c = 1$ plan is a little larger than desired.

14. a. P (Accept) shown for π values below:

c	$\pi = .01$	$\pi = .05$	$\pi = .08$	$\pi = .10$	$\pi = .15$
0	.8179	.3585	.1887	.1216	.0388
1	.9831	.7359	.5169	.3918	.1756
2	.9990	.9246	.7880	.6770	.4049

The operating characteristic curves would show the P (Accept) versus p for each value of c.

b. P (Accept)

c	At $p_0 = .01$	Producer's Risk	At $p_1 = .08$	Consumer's Risk
0	.8179	.1821	.1887	.1887
1	.9831	.0169	.5169	.5169
2	.9990	.0010	.7880	.7880

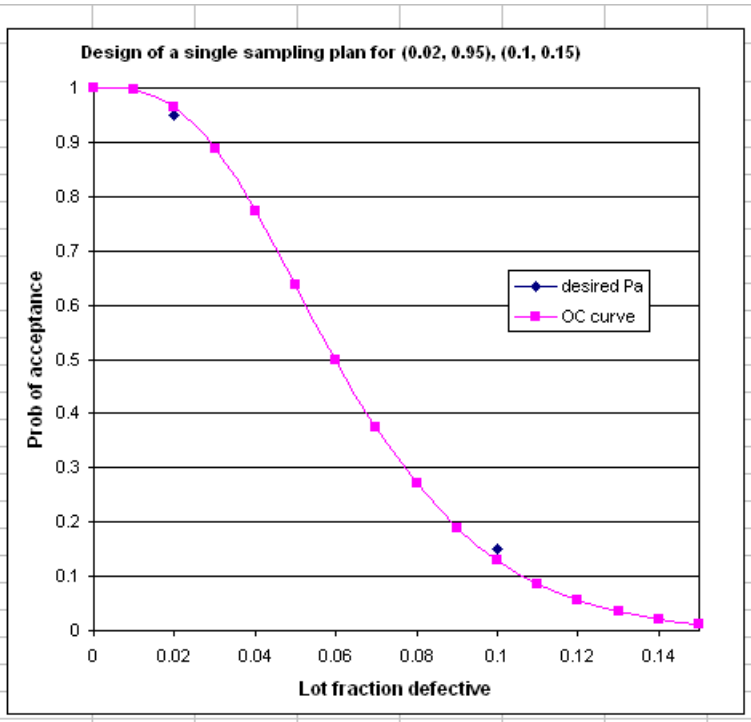
15. We wish to find an attribute sampling plan whose OC curve passes through the two points (0.02, 0.95) and (0.1, 0.15). How can we design a plan (i.e., choose appropriate values of n and c) that satisfies these criteria?

Spreadsheet simulation provides a useful facility for achieving this: see http://mathdl.maa.org/images/upload_library/4/vol3/ng/risk.xls for details.

Single sampling plan

	n	c
	61	3

Lot fraction defective	Desired Pa	Bin Pa
0		1
0.01		0.9966837
0.02	0.95	0.9660749
0.03		0.8893742
0.04		0.7727932
0.05		0.6357887
0.06		0.4982396
0.07		0.3741883
0.08		0.2707628
0.09		0.1896091
0.1	0.15	0.1289632
0.11		0.0854427
0.12		0.0552725
0.13		0.0349783
0.14		0.021688
0.15		0.0131923



By using the Spin Button to change the values of n and c it can be shown that the n = 61 c = 3 scheme comes very close to meeting the stated requirements. Out of interest, the acceptance number can be seen to have much greater effect on the probability of acceptance and hence the shape of the OC curve than the sample size.

Supplementary Exercises:

16. Samples of size 5 provided the following 20 sample means for a production process that is believed to be in control.

95.72	95.24	95.18	95.46
95.44	95.46	95.32	95.72
95.60	95.78	94.82	95.04
95.40	95.44	95.08	95.22
95.50	95.80	95.22	95.56

a. Based on these data, what is an estimate of the mean when the process is in control?

b. Assume that the process standard deviation is $\sigma = 0.50$ and develop a control chart for this production process. Assume that the mean of the process is the estimate developed in part (a).

c. Do any of the 20 sample means indicate that the process was out of control?

17. Product filling weights are normally distributed with a mean of 350 grams and a standard deviation of 15 grams.

a. Develop the control limits for samples of size 10, 20, and 30.

b. What happens to the control limits as the sample size is increased?

c. What happens when a Type I error is made?

d. What happens when a Type II error is made?

e. What is the probability of a Type I error for samples of size 10, 20, and 30?

f. What is the advantage of increasing the sample size for control chart purposes? What error probability is reduced as the sample size is increased?

18. Twenty-five samples of size 5 resulted in $\bar{\bar{x}} = 5.42$ and $\bar{R} = 2.0$. Compute control limits for the $\bar{x}$ and R charts, and estimate the standard deviation of the process.

19. The following are quality control data for a manufacturing process at Kafka Chemical Company. The data show the temperature in degrees centigrade at five points in time during a manufacturing cycle. The company is interested in using control charts to monitor the temperature of its manufacturing process.

File "Kafka"

Construct the $\bar{x}$ chart and *R* chart. What conclusions can be made about the quality of the process?

Sample	$\bar{x}$	***R***	**Sample**	$\bar{x}$	***R***
1	95.72	1.0	11	95.80	0.6
2	95.24	0.9	12	95.22	0.2
3	95.18	0.8	13	95.56	1.3
4	95.44	0.4	14	95.22	0.5
5	95.46	0.5	15	95.04	0.8
6	95.32	1.1	16	95.72	1.1
7	95.40	0.9	17	94.82	0.6
8	95.44	0.3	18	95.46	0.5
9	95.08	0.2	19	95.60	0.4
10	95.50	0.6	20	95.74	0.6

20. The following were collected for the Maestro Coffee production process. The data show the filling weights based on samples of 1.5 kg cans of coffee. Use these data to construct the $\bar{x}$ and *R* chart. What conclusions can be made about the quality of the production process?

File "Maestro"

	Observations				
Sample 1	**2**	**3**	**4**	**5**	
1	1.52	1.54	1.53	1.55	1.55
2	1.56	1.53	1.52	1.55	1.55
3	1.53	1.52	1.56	1.55	1.55
	1.54	1.54	1.54	1.54	1.53
5	1.55	1.53	1.53	1.53	1.54
6	1.54	1.55	1.56	1.51	1.53
7	1.53	1.53	1.54	1.55	1.54
8	1.55	1.54	1.53	1.53	1.53
9	1.54	1.54	1.54	1.53	1.54
10	1.53	1.55	1.53	1.54	1.53

21. Consider the following situations. Comment on whether the situation might cause concern about the quality of the process.

a. A *p* chart has LCL = 0 and UCL = 0.068. When the process is in control, the proportion defective is .033. Plot the following seven sample results: 0.035, 0.062, 0.055, 0.049, 0.058, 0.066, and 0.055. Discuss.

b. An $\bar{x}$ chart has LCL = 22.2 and UCL = 24.5. The mean is μ = 23.35 when the process is in control. Plot the following seven sample results: 22.4, 22.6, 22.65, 23.2, 211.9, 23.85, and 24.1. Discuss.

22. Thirty samples of size 3 are taken of an electrical component. The details are given below.

File "Electrical"

Sample			
1	207	194	201
2	204	191	203
3	198	201	196
4	195	199	181
5	199	221	218
6	200	200	207
7	222	195	205
8	215	186	181
9	188	199	191
10	171	200	201
11	200	201	192
12	204	207	194
13	191	215	200
14	201	204	200
15	198	212	206
16	231	188	223
17	202	210	219
18	187	190	205
19	194	196	207
20	196	199	190
21	198	208	200
22	185	206	201
23	209	225	226
24	199	199	208
25	214	203	195
26	208	202	199
27	202	205	185
28	195	210	199
29	206	191	200
30	205	198	202

Compute trial control limits for $\bar{x}$ and R. Hence determine if the process is in control. If not, compute revised control limits after eliminating samples that appear to be affected by 'assignable' causes.

23. Samples of size 50 are randomly selected each day from a continuous process. The number of defectives observed in the first 28 samples is shown below.

File "Defectives1"

Sample	Number of Defectives
1	5
2	7
3	11
4	9
5	14
6	21
7	25
8	18
9	10
10	8
11	18
12	19
13	6
14	8
15	6
16	9
17	10
18	11
19	13
20	30
21	26
22	13
23	8
24	23
25	34
26	25
27	8
28	12

a. Compute p10 and the trial control limits.

b. Indicate which, if any, observations fall outside the trial limits.

24. An engineering company produces 20 cm high tensile bolts. Samples of four are taken at thirty minute intervals and tested to destruction. From the individual breaking strengths recorded (in kg / cm^2), sample means and ranges are then calculated - with details for the first thirty samples shown below:

File "Bolts"

Sample	$\bar{x}$	***R***	**Sample**	$\bar{x}$	***R***
1	200.2	3.2	16	200.8	3.0
2	199.3	2.5	17	198.3	1.0
3	201.0	1.0	18	198.9	1.0
4	200.1	2.8	19	199.2	1.5
5	198.3	2.7	20	200.8	2.3
6	201.5	1.3	21	201.0	3.0
7	201.2	1.4	22	199.0	2.6
8	199.6	3.0	23	201.4	1.5
9	199.9	2.2	24	198.8	2.3
10	200.5	2.0	25	199.2	3.4
11	200.1	3.1	26	201.0	2.5
12	198.8	2.7	27	201.3	2.6
13	198.9	0.8	28	198.7	3.1
14	201.5	1.1	29	199.0	2.8
15	200.0	2.2	30	201.7	1.4

a. Construct mean and range charts for these data.

b. The bolts are intended to have a tensile strength of 200 ± 5 kg / cm^2. Is the process capable of producing to this specification?

25. Samples of 200 are tested periodically for defectives. After 20 samples the number of defectives in each sample was found to be:

File "Defectives2"

Sample	Number of Defectives
1	18
2	21
3	10
4	35
5	30
6	27
7	14
8	16
9	8
10	21
11	23
12	14
13	18
14	21
15	32
16	15
17	20
18	25
19	13
20	19

a. Construct a control chart for the proportion (or number) of defectives.

b. What do you deduce about the background production process?

26. Managers of 1200 different retail outlets make twice-a-month restocking orders from a central warehouse. Past experience shows that 4% of the orders result in one or more errors such as wrong item shipped, wrong quantity shipped, and item requested but not shipped. Random samples of 200 orders are selected monthly and checked for accuracy.

a. Construct a control chart for this situation.

b. Six months of data show the following numbers of orders with one or more errors: 10, 15, 6, 13, 8, and 17. Plot the data on the control chart. What does your plot indicate about the order process?

27. An $n = 10$, $c = 2$ acceptance sampling plan is being considered; assume that $p_0 = 0.05$ and $p_1 = 0.20$.

a. Compute both producer's and consumer's risk for this acceptance sampling plan.

b. Would the producer, the consumer, or both be unhappy with the proposed sampling plan?

c. What change in the sampling plan, if any, would you recommend?

28. An acceptance sampling plan with $n = 15$ and $c = 1$ has been designed with a producer's risk of 0.075.

a. Was the value of p_0 .01, .02, .03, .04, or .05? What does this value mean?

b. What is the consumer's risk associated with this plan if p_1 is .25?

29. A manufacturer produces lots of a canned food product. Let denote the proportion of the lots that do not meet the product quality specifications. An $n = 25$, $c = 0$ acceptance sampling plan will be used.

a. Compute points on the operating characteristic curve when □ = 0.01, 0.03, 0.10, and 0.20.

b. Plot the operating characteristic curve.

c. What is the probability that the acceptance sampling plan will reject a lot containing 0.01 defective?

30. Sometimes an acceptance sampling plan will be based on a large sample. In this case, the normal approximation to the binomial distribution can be used to compute the producer's and the consumer's risk associated with the plan. Referring to Chapter 6, we know that the normal distribution used to approximate binomial probabilities has a mean of n□ and a standard deviation of $\sqrt{n\pi(1-\pi)}$.

Assume that an acceptance sampling plan is $n = 250$, $c = 10$.

a. What is the producer's risk if p_0 is 0.02? As discussed in Chapter 6, a continuity correction factor should be used in this case. Thus, the probability of acceptance is based on the normal probability of the random variable being less than or equal to 10.5.

b. What is the consumer's risk if p_1 is 0.08?

c. What is an advantage of a large sample size for acceptance sampling? What is a disadvantage?

Supplementary Exercises Solutions:

16. a. $\mu = \frac{\Sigma \bar{x}}{20} = \frac{1908}{20} = 95.4$

b.

$$\text{UCL} = \mu + 3(\sigma / \sqrt{n}) = 95.4 + 3(.50 / \sqrt{5}) = 96.07$$
$$\text{LCL} = \mu - 3(\sigma / \sqrt{n}) = 95.4 - 3(.50 / \sqrt{5}) = 94.73$$

c. No; all were in control

17. a. For $n = 10$

$$\text{UCL} = \mu + 3(\sigma / \sqrt{n}) = 350 + 3(15 / \sqrt{10}) = 364.23$$
$$\text{LCL} = \mu - 3(\sigma / \sqrt{n}) = 350 - 3(15 / \sqrt{10}) = 335.77$$

For $n = 20$

$$\text{UCL} = 350 + 3(15 / \sqrt{20}) = 360.06$$
$$\text{LCL} = 350 - 3(15 / \sqrt{20}) = 339.94$$

For $n = 30$

$$\text{UCL} = 350 + 3(15 / \sqrt{30}) = 358.22$$
$$\text{LCL} = 350 - 3(15 / \sqrt{30}) = 343.78$$

b. Both control limits come closer to the process mean as the sample size is increased.

c. The process will be declared out of control and adjusted when the process is in control.

d. The process will be judged in control and allowed to continue when the process is out of control.

e. All have $z = 3$ where area $= .4986$

$$P(\text{Type I}) = 1 - 2(.4986) = .0028$$

18. *R* Chart:

$$\text{UCL} = \bar{R}D_4 = 2(2.115) = 4.23$$
$$\text{LCL} = \bar{R}D_3 = 2(0) = 0$$

$\bar{x}$ Chart:

$$\text{UCL} = \bar{\bar{x}} + A_2\bar{R} = 5.42 + 0.577(2) = 6.57$$
$$\text{LCL} = \bar{\bar{x}} - A_2\bar{R} = 5.42 - 0.577(2) = 4.27$$

Estimate of Standard Deviation:

$$\hat{\sigma} = \frac{\bar{R}}{d_2} = \frac{2}{2.326} = 0.86$$

19. $\bar{R} = 0.665$ $\bar{\bar{x}} = 95.398$

$\bar{x}$ Chart:

$$\text{UCL} = \bar{\bar{x}} + A_2\bar{R} = 95.398 + 0.577(0.665) = 95.782$$
$$\text{LCL} = \bar{\bar{x}} - A_2\bar{R} = 95.398 - 0.577(0.665) = 95.014$$

R Chart:

$$\text{UCL} = \bar{R}D_4 = 0.665(2.115) = 1.406$$
$$\text{LCL} = \bar{R}D_3 = 0.665(0) = 0$$

The *R* chart indicated the process variability is in control. All sample ranges are within the control limits. However, the process mean is out of control. Sample 11 ($\bar{x} = 95.80$) and Sample 17 ($\bar{x} = 94.82$) fall outside the control limits.

20. $\bar{R} = .053$ $\bar{\bar{x}} = 3.082$

$\bar{x}$ Chart:

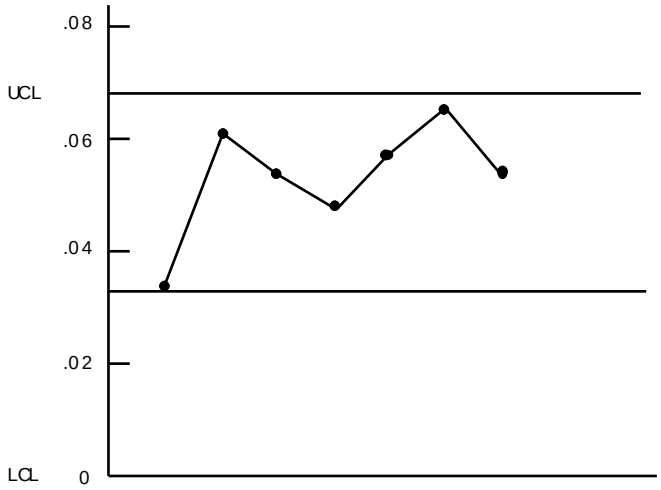
$$\text{UCL} = \bar{\bar{x}} + A_2\bar{R} = 3.082 + 0.577(0.053) = 3.112$$
$$\text{LCL} = \bar{\bar{x}} - A_2\bar{R} = 3.082 - 0.577(0.053) = 3.051$$

R Chart:

$$\text{UCL} = \bar{R}D_4 = 0.053(2.115) = 0.1121$$
$$\text{LCL} = \bar{R}D_3 = 0.053(0) = 0$$

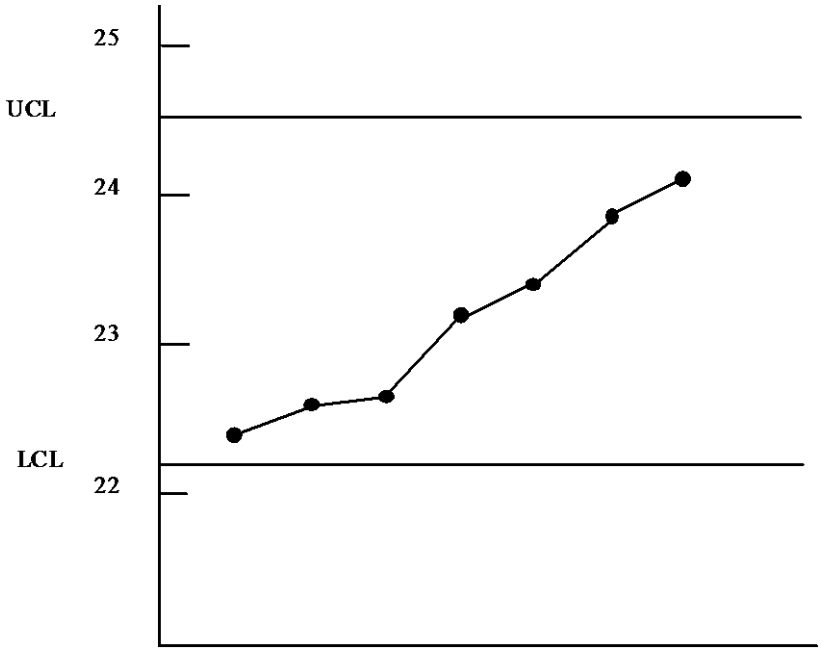
All data points are within the control limits for both charts.

21. a.



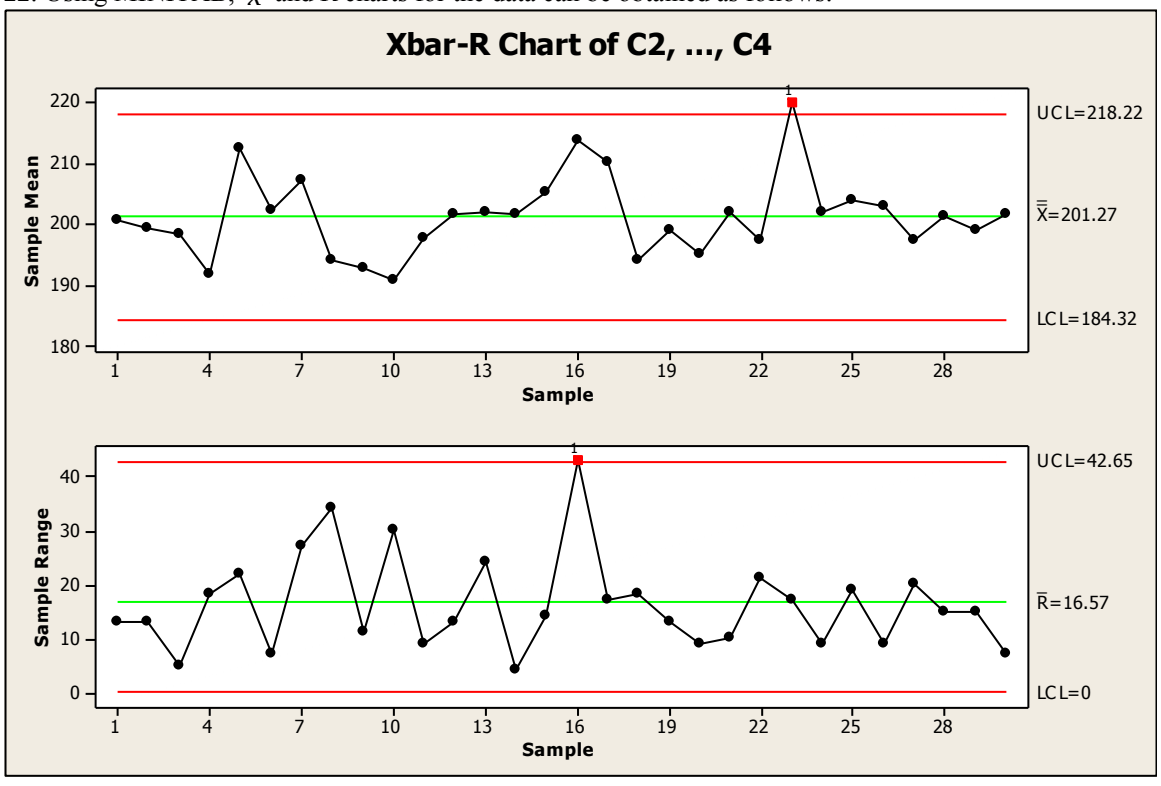
Warning: Process should be checked. All points are within control limits; however, all points are also greater than the process proportion defective.

b.



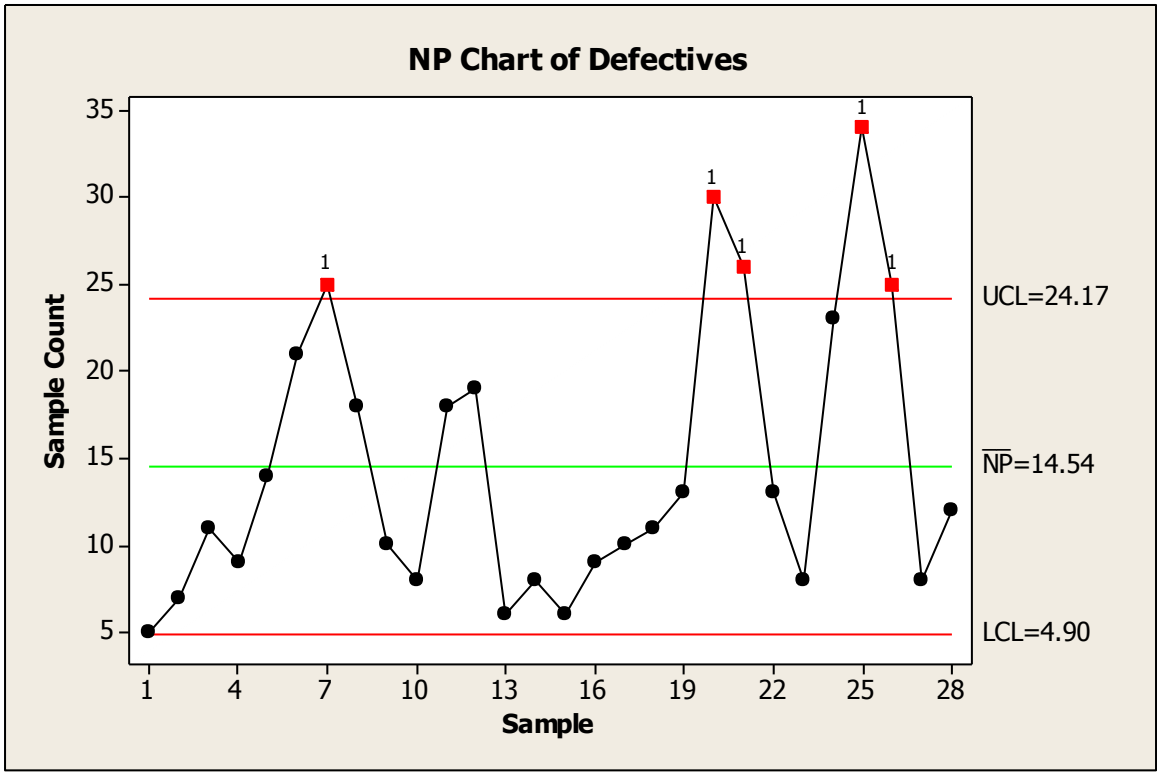
Warning: Process should be checked. All points are within control limits yet the trend in points show a movement or shift toward UCL out-of-control point.

22. Using MINITAB, $\bar{x}$ and R charts for the data can be obtained as follows:



Technically, both charts here reveal two problem points. In the case of the $\bar{x}$ chart, the 23rd sample mean of 220 is just above the UCL of 218.22. Correspondingly the 16th sample with a range of 43 is just actionable according to the R chart. However, because of the closeness of these points to the respective control limits and the fact that the plots are relatively well-behaved after each of them we would deduce the process appears to be back in control. Nevertheless it would be wise to monitor both plots very carefully in the immediate future.

23. Using MINITAB an np chart of defectives can be obtained as follows:



From this chart we deduce the process is seriously out of control, problem samples being 7, 20, 21, 25 and 26 (which all have defective counts higher than the UCL of 24.17). A possible pattern with the plot suggests the situation - if anything – is becoming worse. So corrective action should be taken as soon as practicable.

24. a. It can be shown $\bar{\bar{x}} = 200$ and $\bar{R} = 2.2$

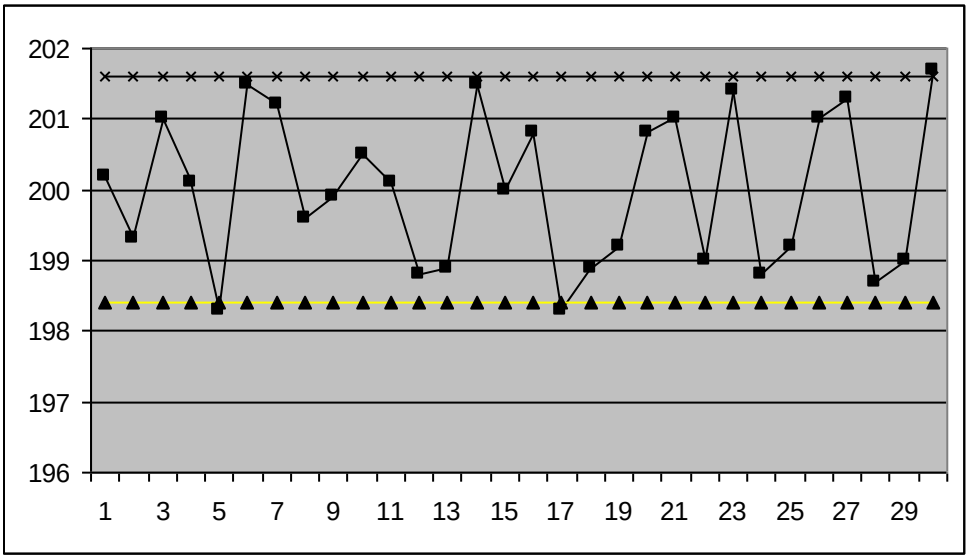
Control limits for the $\bar{x}$ chart are given by $\bar{\bar{x}} \pm A_2 \bar{R}$

where for a sample size, n = 4 , $A_2 = 0.729$

Thus UCL = 200 + 0.729(2.2) = 201.6

LCL = 200 - 0.729(2.2) = 198.4

The corresponding $\bar{x}$ chart from EXCEL is as follows:



Control limits for the R chart are given by:

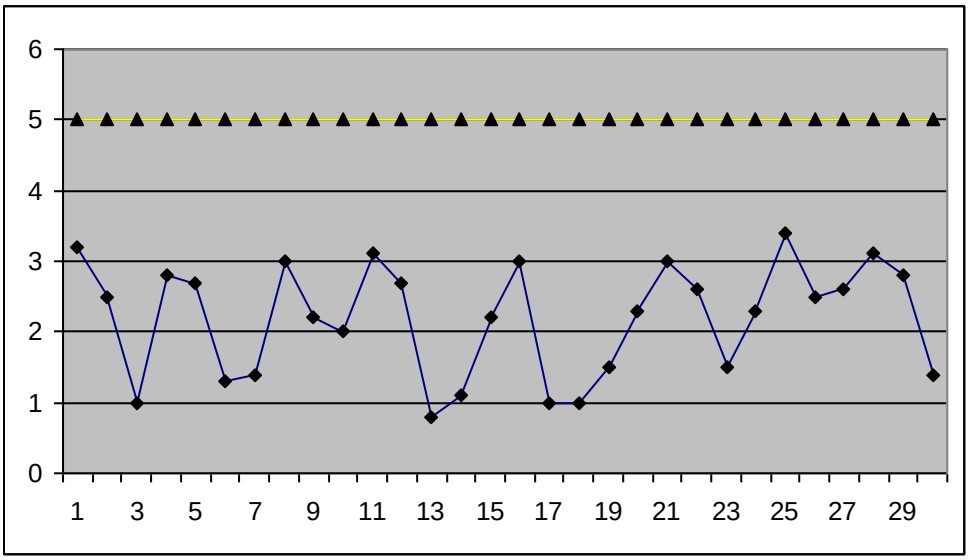
UCL = $\overline{R}\,D_4$ LCL = $\overline{R}\,D_3$

where for a sample size n = 4 we have $D_3 = 0$, $D_4 = 2.282$

Thus UCL = 2.2(2.282) = 5.0

LCL = 2.2(0) = 0

The R chart chart from EXCEL is as follows:



This R chart can be seen to be in control. Not so the corresponding $\overline{x}$ chart which shows points 5, 17 and 30 to be just outside the control limits.

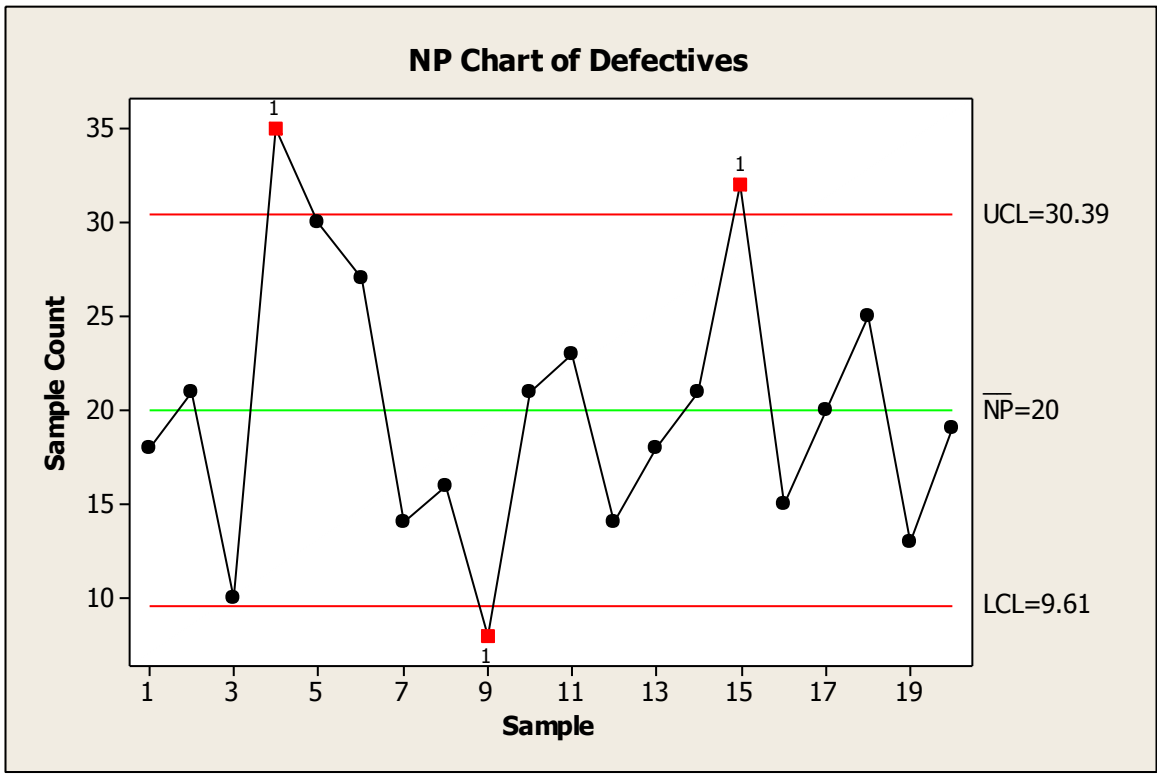
b. The estimated standard deviation of $\bar{x}$ values from MINITAB is 1.08 = $\hat{\sigma}_{\bar{x}}$.

As $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ our estimate $\hat{\sigma}$ of σ, is $\hat{\sigma}_{\bar{x}}(\sqrt{n}) = 1.08(\sqrt{4}) = 2.16$

Thus the specification limits for an individual bolt are $\bar{\bar{x}} \pm 3\hat{\sigma}$ = 200 ± 3(2.16) = 200 ± 6.48.

We therefore deduce the process is <u>not</u> capable of producing to a 200 ± 5 kg / cm^2 tensile strength specification.

25. Using MINITAB an np chart of defectives can be obtained as follows:



We deduce the process is out of control, problem samples being 4, 9, and 15. (The first and third of these samples have defective counts higher than the UCL of 30.39, the second, a count below the LCL of 9.61.)

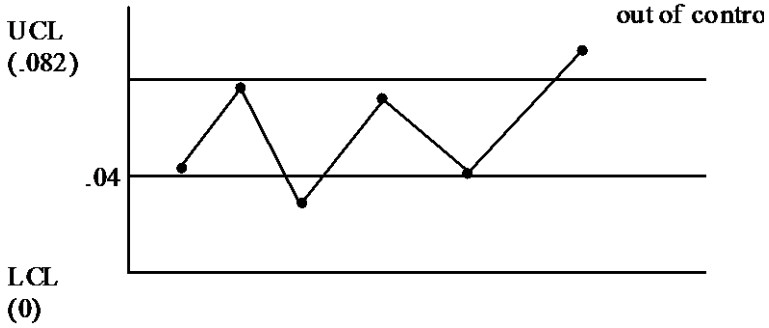
26. a. π = .04

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{(0.04)(0.96)}{200}} = 0.0139$$

UCL = $p + 3\sigma_p$ = 0.04 + 3(0.0139) = 0.0817
LCL = $p - 3\sigma_p$ = 0.04 - 3(0.0139) = -0.0017

Use LCL = 0

b.



For month 1 p= 10/200 = 0.05. Other monthly values are .075, .03, .065, .04, and .085. Only the last month with p = 0.085 is an out-of-control situation.

27. a. Use binomial probabilities with $n = 10$.

At $p_0 = .05$,

P(Accept lot) = p (0) + p (1) + p (2)

= .5987 + .3151 + .0746 = .9884

Producer's Risk: $\alpha = 1 - .9884 = .0116$

At $p_1 = .20$,

P(Accept lot) = p (0) + p (1) + p (2)

= .1074 + .2684 + .3020 = .6778

Consumer's risk: $\beta = .6778$

b. The consumer's risk is unacceptably high. Too many bad lots would be accepted.

c. Reducing c would help, but increasing the sample size appears to be the best solution.

28. a. P (Accept) are shown below: (Using $n = 15$)

	$\pi = .01$	$\pi = .02$	$\pi = .03$	$\pi = .04$	$\pi = .05$
p (0)	.8601	.7386	.6333	.5421	.4633
p(1)	.1303	.2261	.2938	.3388	.3658
	.9904	.9647	.9271	.8809	.8291
$\alpha = 1 - P$ (Accept)	.0096	.0353	.0729	.1191	.1709

Using $p_0 = .03$ since α is close to .075. Thus, .03 is the fraction defective where the producer will tolerate a .075 probability of rejecting a good lot (only .03 defective).

b.

	$\pi = .25$
p (0)	.0134
p (1)	.0668
$\beta =$	.0802

29. a. P (Accept) when $n = 25$ and $c = 0$. Use the binomial probability function with

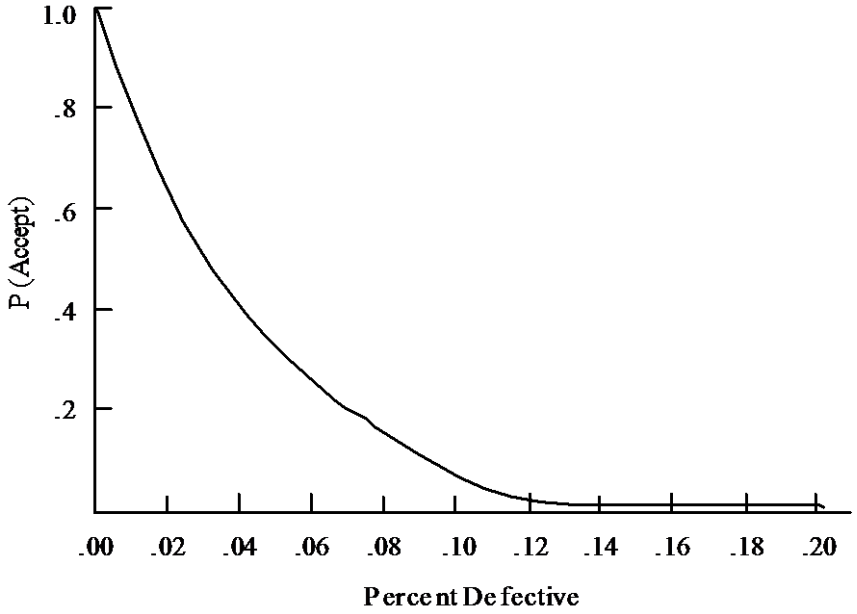
$$p(x) = \frac{n!}{x!(n-x)!}\pi^{x}(1-\pi)^{(n-x)}$$

or

$$p(0) = \frac{25!}{0!25!}\pi^{0}(1-\pi)^{25} = (1-\pi)^{25}$$

If	p (0)
$\pi = .01$	.7778
$\pi = .03$	.4670
$\pi = .10$	.0718
$\pi = .20$	.0038

b.



c. $1 - p(0) = 1 - .778 = .222$

30. a. $\mu = n\pi = 250(.02) = 5$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{250(0.02)(0.98)} = 2.21$$

P (Accept) $= P(x \leq 10.5)$

$$z = \frac{10.5-5}{2.21} = 2.49$$

P (Accept) $= .5000 + .4936 = .9936$

Producer's Risk: $\alpha = 1 - .9936 = .0064$

b. $\mu = n\pi = 250\,(.08) = 20$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{250(0.08)(0.92)} = 4.29$$

P (Accept) = $P\,(x \leq 10.5)$

$$z = \frac{10.5 - 20}{4.29} = -2.21$$

P (Accept) = 1 - .4864 = .0136

Consumer's Risk: β = .0136

c. The advantage is the excellent control over the producer's and the consumer's risk. The disadvantage is the cost of taking a large sample.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Twenty-One

Decision Analysis

Textbook Exercises (1-17)
Textbook Exercise Solutions (1-17)
Supplementary Exercises (18-31)
Supplementary Exercise Solutions

Textbook Exercises:

1. The following payoff table shows profi t for a decision analysis problem with two decision alternatives and three states of nature.

	States of nature		
Decision alternative	s_1	s_2	s_3
d_1	250	100	25
d_2	100	100	75

a) Construct a decision tree for this problem.
b) Suppose that the decision-maker obtains the probabilities $P(s_1) = 0.65$, $P(s_2) = 0.15$ and $P(s_3) = 0.20$. Use the expected value approach to determine the optimal decision.

2. A decision-maker faced with four decision alternatives and four states of nature develops the following profit payoff table.

	States of nature			
Decision alternative	s_1	s_2	s_3	s_4
d_1	14	9	10	5
d_2	1	10	8	7
d_3	9	10	10	11
d_4	8	10	11	13

The decision-maker obtains information that enables the following probabilities assessments:

$P(s_1) = 0.5$, $P(s_2) = 0.2$, $P(s_3) = 0.2$, and $P(s_4) = 0.1$.

a) Use the expected value approach to determine the optimal solution.
b) Now assume that the entries in the payoff table are costs. Use the expected value approach to determine the optimal decision.

3 Consider the payoff matrix, below:

	s_1	s_2	s_3
d_1	100,000	40,000	-60,000
d_2	50,000	20,000	-30,000
d_3	20,000	20,000	-10,000
d_4	40,000	20,000	-60,000

Suppose the probabilities associated with the states of nature here are as follows:

$P(s_1) = 0.1$, $P(s_2) = 0.3$ and $P(s_3) = 0.6$

a. What is the best decision available, for the data, using the expected value (EV) criterion?

4. Holland Corporation is considering three options for managing its data processing operation: continue with its own staff, hire an outside vendor to do the managing (referred to as *outsourcing*) or use a combination of its own staff and an outside vendor. The cost of the operation depends on future demand. The annual cost of each option (in thousands of euros) depends on demand as follows.

	Demand		
Staffing options	High	Medium	Low
Own staff	650	650	600
Outside vendor	900	600	300
Combination	800	650	500

a) If the demand probabilities are 0.2, 0.5 and 0.3, which decision alternative will minimize the expected cost of the data processing operation? What is the expected annual cost associated with your recommendation?
b) What is the expected value of perfect information?

5. Magyar Air Express decided to offer direct service from Budapest to Prague. Management must decide between a full price service using the company's new fleet of jet aircraft and a discount service using smaller capacity commuter planes. It is clear that the best choice depends on the market reaction to the service Magyar Air offers. Management developed estimates of the contribution to profit for each type of service based upon two possible levels of demand for service to Prague: strong and weak. The following table shows the estimated quarterly profits (in thousands of euros).

	Demand for service	
Service	Strong	Weak
Full price	960	−490
Discount	670	320

a) What is the decision to be made, what is the chance event, and what is the consequence for this problem? How many decision alternatives are there? How many outcomes are there for the chance event?
b) Suppose that management of Magyar Air Express believes that the probability of strong demand is 0.7 and the probability of weak

demand is 0.3. Use the expected value approach to determine an optimal decision.

c) Suppose that the probability of strong demand is 0.8 and the probability of weak demand is 0.2. What is the optimal decision using the expected value approach?

6. A South African company is considering whether it should tender for two contracts (MS1 and MS2) on offer from a government department for the supply of certain components. The company has three options: (i) tender for MS1 only (ii) tender for MS2 only (ii) or tender for both MS1 and MS2.

If tenders are to be submitted the company will incur additional costs. These costs will have to be entirely recouped from the contract price. The risk, of course, is that if a tender is unsuccessful the company will have made a loss.

The cost of tendering for contract MS1 only is 500 000 RAND. The component supply cost if the tender is successful would be 180 000 RAND.

The cost of tendering for contract MS2 only is 140 000 RAND. The component supply cost if the tender is successful would be 120 000 RAND.

The cost of tendering for both contract MS1 and contract MS2 is 550 000 RAND. The component supply cost if the tender is successful would be 240 000 RAND.

For each contract, possible tender prices have been determined. In addition, subjective assessments have been made of the probability of getting the contract with a particular tender price as shown below. Note here that the company can only submit one tender and cannot, for example, submit two tenders (at different prices) for the same contract.

Option	Possible tender prices (RAND)	Probability of getting contract
MS1 only	1 300 000	0.20
	1 150 000	0.85
MS2 only	700 000	0.15
	650 000	0.80
	600 000	0.95
MS1 and MS2	1 900 000	0.05
	1 400 000	0.65

In the event that the company tenders for both MS1 and MS2 it will either win both contracts (at the price shown above) or no contract at all.

a) What do you suggest the company should do and why?
b) What is the downside and the upside of the suggested course of action?
c) A consultant has approached the company with an offer that in return for 200 000 RAND in cash she will ensure that if you tender 600 000 RAND for contract MS2 only your tender is guaranteed to be successful. Should you accept her offer or not and why?

7. An Australian government committee is considering the economic benefits of a programme of preventative flu vaccinations. If vaccinations are not introduced then the estimated cost to the government if flu strikes in the next year is A$7 m with probability 0.1, A$10 m with probability 0.3 and A$15 m with probability 0.6. It is estimated that such a programme will cost A$7 m and that the probability of flu striking in the next year is 0.75. One alternative open to the committee is to institute an 'early-warning' monitoring scheme (costing A$3 m) which will enable it to detect an outbreak of fl u early and hence institute a rush vaccination program (costing A$10 m because of the need to vaccinate quickly before the outbreak spreads).

a) What recommendations should the committee make to the government if their objective is to maximize expected monetary value (EMV)?
b) The committee has also been informed that there are alternatives to using EMV. What are these alternatives and would they be appropriate in this case?

8. Consider a variation of the PDC decision tree shown in Figure 21.5. The company must first decide whether to undertake the market research study. If the market research study is conducted, the outcome will either be favourable (*F*) or unfavourable (*U*). Assume there are only two decision alternatives d_1 and d_2 and two states of nature s_1 and s_2. The payoff table showing profit is as follows:

	State of nature	
Decision alternative	s_1	s_2
d_1	100	300
d_2	400	200

a) Show the decision tree.
b) Use the following probabilities. What is the optimal decision strategy?

$P(F) = 0.56$ $P(s_1|F) = 0.57$ $P(s_1|U) = 0.18$ $P(s_1) = 0.40$
$P(U) = 0.44$ $P(s_2|F) = 0.43$ $P(s_2|U) = 0.82$ $P(s_2) = 0.60$

9. A quality control procedure involves 100% inspection of parts received from a supplier. Historical records show the following defective rates have been observed :

% Defective	Probability
0	0.15
1	0.25
2	0.40
3	0.20

The cost of 100% inspection is €250 for each shipment of 500 parts. If the shipment is not 100% inspected, defective parts will cause rework problems later in production process. The rework is €25 for each defective part.

The plant manager is considering eliminating the inspection process in order to save the €250 inspection cost per shipment.

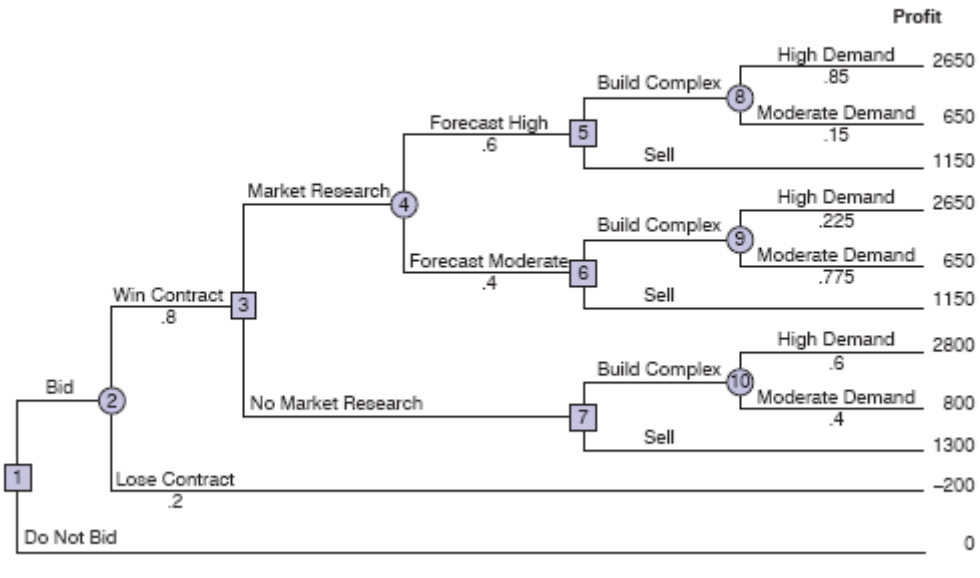
a. Provide a decision matrix representation for this problem.

b. Would you support the plant manager's view as regards eliminating the inspection process? Explain.

c. Represent the problem as a decision tree.

10. Dante Development Corporation is considering bidding on a contract for a new office building complex. Figure 21.9 shows the decision tree prepared by one of Dante's analysts. At node 1, the company must decide whether to bid on the contract. The cost of preparing the bid is €200 000. The upper branch from node 2 shows that the company has a 0.8 probability of winning the contract if it submits a bid. If the company wins the bid, it will have to pay €2 000 000 to become a partner in the project. Node 3 shows that the company will then consider doing a market research study to forecast demand for the office units prior to beginning construction. The cost of this study is €150 000. Node 4 is a chance node showing the possible outcomes of the market research study.

Figure 21.9 Decision tree for the Dante Development Corporation



Nodes 5, 6 and 7 are similar in that they are the decision nodes for Dante to either build the office complex or sell the rights in the project to another developer. The decision to build the complex will result in an income of €5 000 000 if demand is high and €3 000 000 if demand is moderate. If Dante chooses to sell its rights in the project to another developer, income from the sale is estimated to be €3 500 000. The probabilities shown at nodes 4, 8 and 9 are based on the projected outcomes of the market research study.

a) Verify Dante's profit projections shown at the ending branches of the decision tree by calculating the payoffs of €2 650 000 and €650 000 for the first two outcomes.
b) What is the optimal decision strategy for Dante, and what is the expected profit for this project?
c) What would the cost of the market research study have to be before Dante would change its decision about conducting the study?

11. A Turkish company is trying to decide whether to bid for a certain contract or not. They estimate that merely preparing the bid will cost YTL10 000. If their company bid then they estimate that there is a 50 per cent chance that their bid will be put on the 'short-list', otherwise their bid will be rejected. Once 'short-listed' the company will have to supply further detailed information (entailing costs estimated at YTL5000). After this stage their bid will either be accepted or rejected.

The company estimate that the labour and material costs associated with the contract are YTL127 000. They are considering three possible bid prices, namely YTL155 000, YTL170 000 and YTL190 000. They estimate that the probability of these bids being accepted (once they have been short-listed) is 0.90, 0.75 and 0.35 respectively.

What should the company do and what is the expected monetary value of your suggested course of action?

12. Hale's TV Productions is considering producing a pilot for a comedy series in the hope of selling it to a major television network. The network may decide to reject the series, but it may also decide to purchase the rights to the series for either one or two years. At this point in time, Hale may either produce the pilot and wait for the network's decision or transfer the rights for the pilot and series to a competitor for €100 000. Hale's decision alternatives and profits (in thousands of euros) are as follows:

	State of nature		
Decision alternative	Reject, s_1	1 Year, s_2	2 Years, s_3
Produce pilot, d_1	−100	50	150
Sell to competitor, d_2	100	100	100

The probabilities for the states of nature are $P(s_1) = 0.2$, $P(s_2) = 0.3$, and $P(s_3) = 0.5$.
For a consulting fee of €5000, an agency will review the plans for the comedy series and indicate the overall chances of a favourable network reaction to the series. Assume that the agency review will result in a favourable (F) or an unfavourable (U) review and that the following probabilities are relevant.

$P(F) = 0.69$	$P(s_1\|F) = 0.09$	$P(s_2\|F) = 0.26$	$P(s_3\|F) = 0.65$
$P(U) = 0.31$	$P(s_1\|U) = 0.45$	$P(s_2\|U) = 0.39$	$P(s_3\|U) = 0.16$

a) Construct a decision tree for this problem.
b) What is the recommended decision if the agency opinion is not used? What is the expected value?
c) What is the expected value of perfect information?
d) What is Hale's optimal decision strategy assuming the agency's information is used?
e) What is the expected value of the agency's information?
f) Is the agency's information worth the €5000 fee? What is the maximum that Hale should be willing to pay for the information?
g) What is the recommended decision?

13. Larson's Department Store faces a buying decision for a seasonal product for which demand can be high, medium or low. The purchaser for Larson's can order 1, 2 or 3 lots of the product before the season begins but cannot reorder later. Profit projections (in thousands of euros) are shown.

	State of nature		
	High demand s_1	Medium demand s_2	Low demand s_3
Order 1 lot, d_1	60	60	50
Order 2 lots, d_2	80	80	30
Order 3 lots, d_3	100	70	10

a) If the prior probabilities for the three states of nature are 0.3, 0.3 and 0.4, respectively, what is the recommended order quantity?
b) At each pre-season sales meeting, the head of sales provides a personal opinion regarding potential demand for this product. Because of the CEO's enthusiasm and optimistic nature, the predictions of market conditions have always been either 'excellent' (*E*) or 'very good' (*V*). Probabilities are as follows. What is the optimal decision strategy?

$P(E) = 0.7$	$P(s_1\|E) = 0.34$	$P(s_2\|E) = 0.32$	$P(s_3\|E) = 0.34$
$P(V) = 0.3$	$P(s_1\|V) = 0.20$	$P(s_2\|V) = 0.26$	$P(s_3\|V) = 0.54$

c) Compute EVPI and EVSI. Discuss whether the fi rm should consider a consulting expert who could provide independent forecasts of market conditions for the product.

14. Suppose that you are given a decision situation with three possible states of nature: s_1, s_2, and s_3. The prior probabilities are $P(s_1) = 0.2$, $P(s_2) = 0.5$ and $P(s_3) = 0.3$. With sample information I, $P(I|s_1) = 0.1$, $P(I|s_2) = 0.05$ and $P(I|s_3) = 0.2$. Compute the revised or posterior probabilities: $P(s_1|I)$, $P(s_2|I)$, and $P(s_3|I)$.

15. In the following profit payoff table for a decision problem with two states of nature and three decision alternatives, the prior probabilities for s_1 and s_2 are $P(s_1) = 0.8$ and $P(s_2) = 0.2$.

	State of nature	
Decision alternative	s_1	s_2
d_1	15	10
d_2	10	12
d_3	8	20

a) What is the optimal decision?
b) Find the EVPI.
c) Suppose that sample information I is obtained, with $P(I|s_1) = 0.20$ and $P(I|s_2) = 0.75$. Find the posterior probabilities $P(s_1|I)$ and $P(s_2|I)$. Recommend a decision alternative based on these probabilities.

16. To save on expenses, Rene and Jacques agreed to form a carpool for travelling to and from work. Rene preferred to use the somewhat longer but more consistent Boulevard Peripherique. Although Jacques preferred the quicker expressway, he agreed with Rene that they should take Boulevard Peripherique if the expressway had a traffi c jam. The following payoff table provides the one-way time estimate in minutes for travelling to and from work.

	State of nature	
Decision alternative	Expressway open s_1	Expressway jammed s_2
Boulevard Peripherique, d_1	30	30
Expressway, d_2	25	45

Based on their experience with traffic problems, Rene and Jacques agreed on a 0.15 probability that the expressway would be jammed. In addition, they agreed that weather seemed to affect the traffic conditions on the expressway.

Let

C = clear
O = overcast
R = rain

The following conditional probabilities apply.

$P(C\|s_1) = 0.8$	$P(O\|s_1) = 0.2$	$P(R\|s_1) = 0.0$
$P(C\|s_2) = 0.1$	$P(O\|s_2) = 0.3$	$P(R\|s_2) = 0.6$

a) Use Bayes' theorem for probability revision to compute the probability of each weather condition and the conditional probability of the expressway open s_1 or jammed s_2 given each weather condition.
b) Show the decision tree for this problem.
c) What is the optimal decision strategy, and what is the expected travel time?

17. The Granaldi Manufacturing Company must decide whether to manufacture a component part at its Milan plant or purchase the component part from a supplier. The resulting profit is dependent upon the demand for the product. The following payoff table shows the projected profi t (in thousands of euros).

	State of nature		
Decision alternative	Low demand s_1	Medium demand s_2	High demand s_3
Manufacture, d_1	−20	40	100
Purchase, d_2	10	45	70

The state-of-nature probabilities are $P(s_1) = 0.35$, $P(s_2) = 0.35$, and $P(s_3) = 0.30$.

a) Use a decision tree to recommend a decision.
b) Use EVPI to determine whether Granaldi should attempt to obtain a better estimate of demand.
c) A test market study of the potential demand for the product is expected to report either a favourable (*F*) or unfavourable (*U*) situation. The relevant conditional probabilities are as follows:

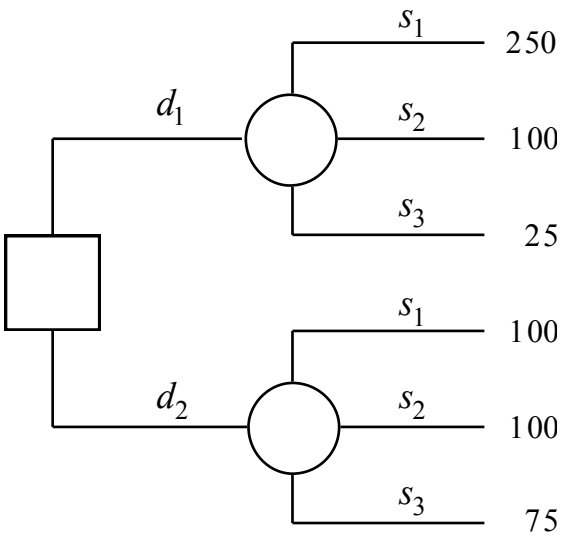
$P(F|s_1) = 0.10$ $P(U|s_1) = 0.90$
$P(F|s_2) = 0.40$ $P(U|s_2) = 0.60$
$P(F|s_3) = 0.60$ $P(U|s_3) = 0.40$

What is the probability that the market research report will be favourable?
d) What is Granaldi's optimal decision strategy?
e) What is the expected value of the market research information?

Chapter 21: Decision Analysis

Textbook Exercises Solutions:

1. a.



b. EV(d_1) = .65(250) + .15(100) + .20(25) = 182.5
EV(d_2) = .65(100) + .15(100) + .20(75) = 95

The optimal decision is d_1

2. a. EV(d_1) = 0.5(14) + 0.2(9) + 0.2(10) + 0.1(5) = 11.3
EV(d_2) = 0.5(11) + 0.2(10) + 0.2(8) + 0.1(7) = 9.8
EV(d_3) = 0.5(9) + 0.2(10) + 0.2(10) + 0.1(11) = 9.6
EV(d_4) = 0.5(8) + 0.2(10) + 0.2(11) + 0.1(13) = 9.5

Recommended decision: d_1

b. The best decision in this case is the one with the smallest expected value; thus, d_4, with an expected cost of 9.5, is the recommended decision.

3. EMV (d_1)= 0.1 (100k) + 0.3 (40k) + 0.6 (-60k) = -14k = -14,000

EMV (d_2) = 0.1 (50k) + 0.3 (20k) + 0.6 (-30k)
= -7k = -7,000

EMV (d_3)= 0.1 (20k) + 0.3 (20k) + 0.6 (-10k) = 2k = 2,000

Hence d_4 is optimal. Note the EMV for d_4 is not a consideration because d_4 is DOMINATED by d_2 i.e. the payoff is better (or no worse) for d_2 than d_4 whatever the state of nature.

4 a. EV(own staff) = 0.2(650) + 0.5(650) + 0.3(600) = 635
EV(outside vendor) = 0.2(900) + 0.5(600) + 0.3(300) = 570
EV(combination) = 0.2(800) + 0.5(650) + 0.3(500) = 635

The optimal decision is to hire an outside vendor with an expected annual cost of €570,000.

b. Expected value with perfect information = .2(650) + .5(600) + .3(300) = 520
EVPI = $|520 - 570| = 50$ or €50,000

4. a. The decision to be made is to choose the type of service to provide. The chance event is the level of demand for the Magyar Air service. The consequence is the amount of quarterly profit. There are two decision alternatives (full price and discount service). There are two outcomes for the chance event (strong demand and weak demand).

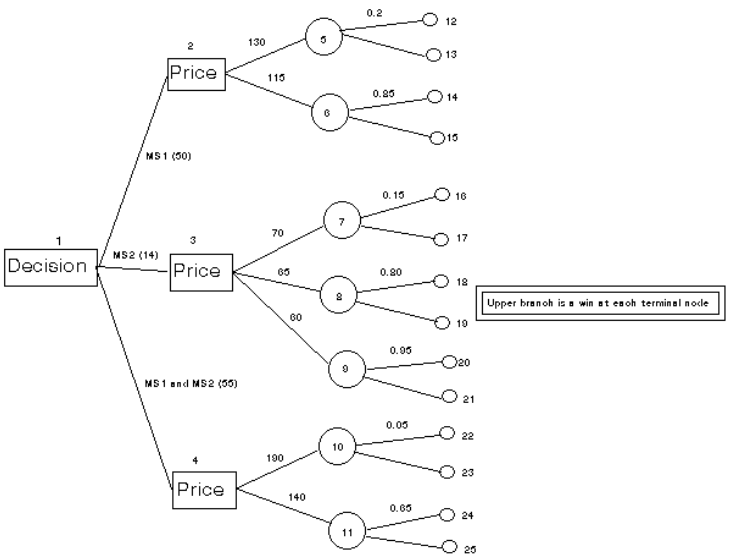
b. EV(Full) = 0.7(960) + 0.3(-490) = 525
EV (Discount) = 0.7(670) + 0.3(320) = 565

Optimal Decision: Discount service

c. EV(Full) = 0.8(960) + 0.2(-490) = 670
EV (Discount) = 0.8(670) + 0.2(320) = 600

Optimal Decision: Full price service

5. a. The decision tree (see http://people.brunel.ac.uk/~mastjjb/jeb/or/decmore.html) for the problem is shown below.



Below we carry out step 1 of the decision tree solution procedure which (for this example) involves working out the total profit for each of the paths from the initial node to the terminal node (all figures in '0,000 Rand).

Step 1

- path to terminal node 12, we tender for MS1 only (cost 50), at a price of 130, and win the contract, so incurring component supply costs of 18, total profit 130-50-18 = 62

- path to terminal node 13, we tender for MS1 only (cost 50), at a price of 130, and lose the contract, total profit -50

- path to terminal node 14, we tender for MS1 only (cost 50), at a price of 115, and win the contract, so incurring component supply costs of 18, total profit 115-50-18 = 47

- path to terminal node 15, we tender for MS1 only (cost 50), at a price of 115, and lose the contract, total profit -50

- path to terminal node 16, we tender for MS2 only (cost 14), at a price of 70, and win the contract, so incurring component supply costs of 12, total profit 70-14-12 = 44

- path to terminal node 17, we tender for MS2 only (cost 14), at a price of 70, and lose the contract, total profit -14

- path to terminal node 18, we tender for MS2 only (cost 14), at a price of 65, and win the contract, so incurring component supply costs of 12, total profit 65-14-12 = 39

- path to terminal node 19, we tender for MS2 only (cost 14), at a price of 65, and lose the contract, total profit -14

- path to terminal node 20, we tender for MS2 only (cost 14), at a price of 60, and win the contract, so incurring component supply costs of 12, total profit 60-14-12 = 34

- path to terminal node 21, we tender for MS2 only (cost 14), at a price of 60, and lose the contract, total profit -14

- path to terminal node 22, we tender for MS1 and MS2 (cost 55), at a price of 190, and win the contract, so incurring component supply costs of 24, total profit 190-55-24=111

- path to terminal node 23, we tender for MS1 and MS2 (cost 55), at a price of 190, and lose the contract, total profit -55

- path to terminal node 24, we tender for MS1 and MS2 (cost 55), at a price of 140, and win the contract, so incurring component supply costs of 24, total profit 140-55- 24=61

- path to terminal node 25, we tender for MS1 and MS2 (cost 55), at a price of 140, and lose the contract, total profit -55

Hence we can arrive at the table below indicating for each branch the total profit involved in that branch from the initial node to the terminal node.

Terminal node	Total profit ‘0,000 Rand
12	62
13	-50
14	47
15	-50
16	44
17	-14
18	39
19	-14
20	34
21	-14
22	111
23	-55
24	61
25	-55

We can now carry out the second step of the decision tree solution procedure where we work from the right-hand side of the diagram back to the left-hand side.

Step 2

For chance node 5 the EMV is 0.2(62) + 0.8(-50) = -27.6
For chance node 6 the EMV is 0.85(47) + 0.15(-50) = 32.45
Hence the best decision at decision node 2 is to tender at a price of 115 (EMV=32.45).

For chance node 7 the EMV is 0.15(44) + 0.85(-14) = -5.3
For chance node 8 the EMV is 0.80(39) + 0.20(-14) = 28.4
For chance node 9 the EMV is 0.95(34) + 0.05(-14) = 31.6
Hence the best decision at decision node 3 is to tender at a price of 60 (EMV=31.6).

For chance node 10 the EMV is 0.05(111) + 0.95(-55) = -46.7
For chance node 11 the EMV is 0.65(61) + 0.35(-55) = 20.4
Hence the best decision at decision node 4 is to tender at a price of 140 (EMV=20.4).

Hence at decision node 1 have three alternatives:

tender for MS1 only EMV=32.45
tender for MS2 only EMV=31.6
tender for both MS1 and MS2 EMV = 20.4

Hence the best decision is to tender for MS1 only (at a price of 115) as it has the highest expected monetary value of 32.45 ('0,000 RAND).

b. The downside is a loss of 50 and the upside is a profit of 47.

c. With regard to the consultants offer then, ignoring ethical considerations, we could of course, tender 60 for MS2 only without her help and if we were to do that we would have a 0.95 probability of having our tender accepted. Hence there are essentially three options:

as before,

- tender for MS1 only at a price of 115: EMV 32.45, downside -50 (probability 0.15), upside 47 (probability 0.85)
- tender for MS2 only at a price of 60, unaided by the consultant: EMV 31.6, downside -14 (probability 0.05), upside 34 (probability 0.95)
- tender for MS2 only at a price of 60, with the consultants help, then (assuming she can fulfil her promise of guaranteeing we will be successful), we have a certain outcome with a profit of 34 (terminal node 20) - 20 (cash paid to the consultant) = 14

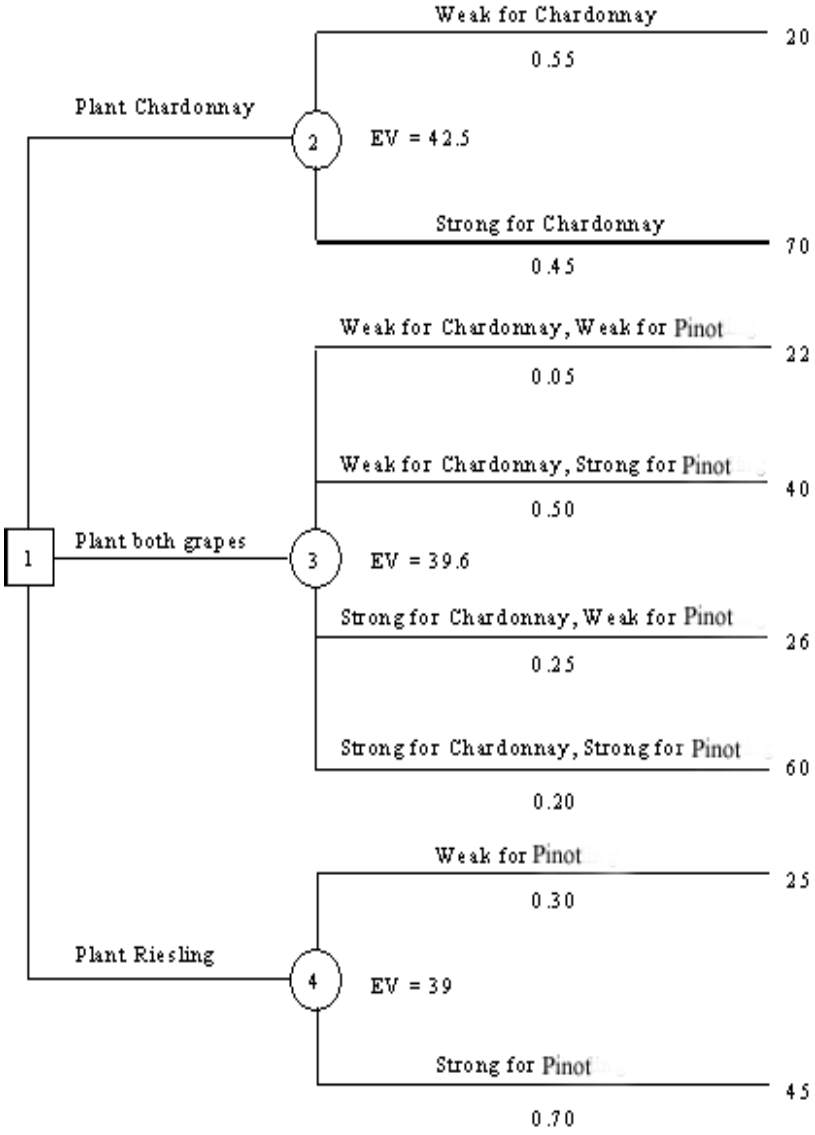
On an EMV basis we would still support our original decision. Looking at the risks (probabilities) of losing money, and considering tendering for MS2 only at 60, we would essentially be paying the consultant 20 to avoid a 0.05 chance of loosing 14, the downside of tendering unaided.

Paying 20 to guarantee not incurring a loss of 14 which will occur with a probability of 0.05 (one in twenty) does not seem like an awfully good investment and so we should reject her offer (or offer her a smaller sum of money in return for her guarantee!).

6. a. The decision is to choose what type of grapes to plant, the chance event is demand for the wine and the consequence is the expected annual profit contribution. There are three decision alternatives (Chardonnay, Pinot and both). There are four chance outcomes: (W,W); (W,S); (S,W); and (S,S). For instance, (W,S) denotes the outcomes corresponding to weak demand for Chardonnay and strong demand for Pinot.

b. In constructing a decision tree, it is only necessary to show two branches when only a single grape is planted. But, the branch probabilities in these cases are the sum of two probabilities. For example, the probability that demand for Chardonnay is strong is given by:

P (Strong demand for Chardonnay) = P(S,W) + P(S,S)
= 0.25 + 0.20
= 0.45



c. EV (Plant Chardonnay) = 0.55(20) +0.45(70) = 42.5
EV (Plant both grapes) = 0.05(22) + 0.50(40) + 0.25(26) + 0.20(60) = 39.6
EV (Plant Pinot) = 0.30(25) + 0.70(45) = 39.0

Optimal decision: Plant Chardonnay grapes only.

d. This changes the expected value in the case where both grapes are planted and when Pinot only is planted.

EV (Plant both grapes) = 0.05(22) + 0.50(40) +0.05(26) + 0.40(60) = 46.4

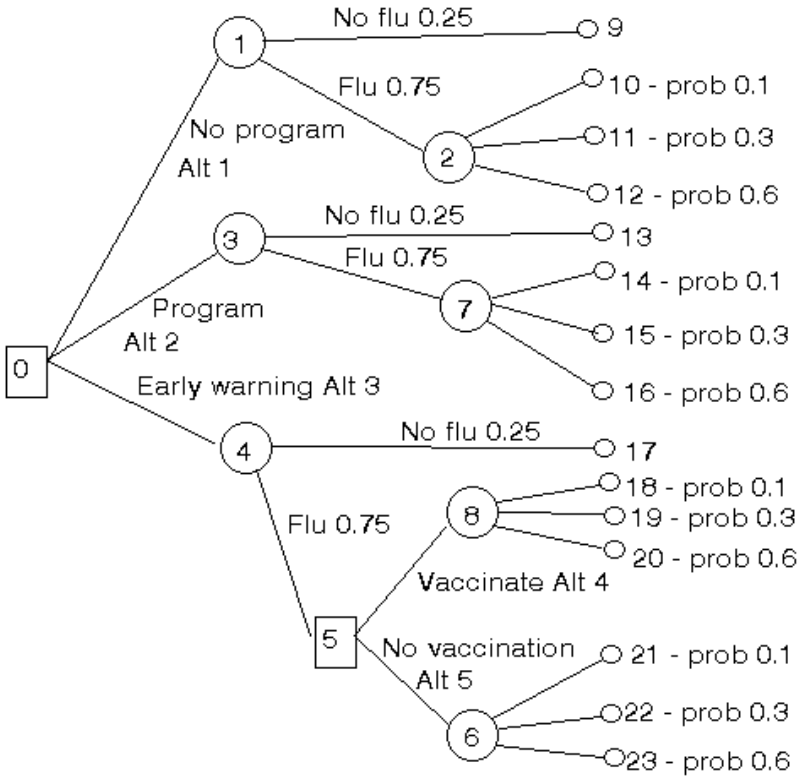
EV (Plant Pinot) = 0.10(25) + 0.90(45) = 43.0
We see that the optimal decision is now to plant both grapes. The optimal decision is sensitive to this change in probabilities.

e. Only the expected value for node 2 in the decision tree needs to be recomputed.

EV (Plant Chardonnay) = 0.55(20) + 0.45(50) = 33.5

This change in the payoffs makes planting Chardonnay only less attractive. It is now best to plant both types of grapes. The optimal decision is sensitive to a change in the payoff of this magnitude.

7. a. The decision tree (see http://people.brunel.ac.uk/~mastjjb/jeb/or/decmore.html) for the problem is shown below.



Below we carry out step 1 of the decision tree solution procedure which (for this example) involves working out the total profit for each of the paths from the initial node to the terminal nodes.

Step 1

- path to terminal node 9 - we carry out no program and flu does not strike

  Total revenue = 0

  Total cost = 0

  Total profit = 0

- path to terminal node 10 - we carry out no program and flu strikes costing the government A$7m

  Total revenue = 0

  Total cost = 7

  Total profit = -7 (all figures in A$m)

- path to terminal nodes 11 and 12 similar to the case above giving a total profit of -10 and -15 respectively

- path to terminal node 13 - we carry out a program costing A$7m and flu does not strike

  Total revenue = 0

  Total cost = 7

  Total profit = -7

- path to terminal node 14 - we carry out a program costing A$7m and flu strikes. Now we would have lost A$7m with this flu outbreak but because of the program (which we assume to be 100% effective) we do not.

The key here is to regard the A$7m paid for the program as "insurance" which reimburses the government for whatever losses are suffered as a result of flu striking. Hence we have

Total revenue = 7 (reimbursement)

Total cost = 7 (cost of program) + 7 (loss due to flu striking)

Total profit = -7

It is clear from the above calculation that since (in this case) the reimbursement always exactly equals the amount lost the total profit will just be the cost of the "insurance" (-A$7m).

The situation with the vaccination program is very similar to household insurance where a single payment guarantees replacement of any losses suffered. Whatever happens the effect of the insurance will be "as if" nothing had occurred. Under these circumstances the only expense (in effect) is the cost of the insurance.

- path to terminal nodes 15 and 16 similar to the case above where we carry out a program costing A$7m and this insures us against losses. Hence

  Total profit = -7 terminal node 15

  Total profit = -7 terminal node 16

- path to terminal node 17 - we carry out an early warning program costing A$3m and flu does not strike giving

  Total revenue = 0

  Total cost = 3

  Total profit = -3

- path to terminal nodes 18, 19 and 20 - we carry out an early warning program costing A$3m, flu strikes and we decide to vaccinate costing A$10m. Hence for a total cost of A$13m we are insured against losses so that we have

  Total profit = -13 terminal node 18

  Total profit = -13 terminal node 19

  Total profit = -13 terminal node 20

- path to terminal nodes 21, 22 and 23 - we carry out an early warning program costing A$3m, flu strikes but we decide not to vaccinate, leading to costs of A$7m, A$10m and A$15m. Hence

Total profit = -10 terminal node 21

Total profit = -13 terminal node 22

Total profit = -18 terminal node 23

Hence we can form the table below indicating for each branch the total profit involved in that branch from the initial node to the terminal node.

Terminal node	Total profit (A$m)
9	0
10	-7
11	-10
12	-15
13	-7
14	-7
15	-7
16	-7
17	-3
18	-13
19	-13
20	-13
21	-10
22	-13
23	-18

We can now carry out the second step of the decision tree solution procedure where we work from the right-hand side of the diagram back to the left-hand side.

Step 2

Consider chance node 2 (with branches to terminal nodes 10, 11 and 12 emanating from it). The expected monetary value (EMV) for this chance node is given by 0.1 x (-7) + 0.3 x (-10) + 0.6 x (-15) = -12.7

Hence the EMV for chance node 1 is given by 0.25 x (0) + 0.75 x (-12.7) = -9.525

Similarly the EMV for chance node 7 is given by 0.1 x (-7) + 0.3 x (-7) + 0.6 x (-7) = -7

which leads to an EMV for chance node 3 of 0.25 x (-7) + 0.75 x (-7) = -7

The EMV for chance node 8 is 0.1 x (-13) + 0.3 x (-13) + 0.6 x (-13) = -13
and the EMV for chance node 6 is 0.1 x (-10) + 0.3 x (-13) + 0.6 x (-18) = -15.7

Hence for decision node 5 we have the two alternatives:

(4) vaccinate EMV = -13

(5) no vaccination EMV = -15.7

Hence the best alternative here is to vaccinate (alternative 4) with an EMV of -13.

The EMV for chance node 4 is therefore 0.25 x (-3) + 0.75 x (-13) = -10.5

and at the initial decision node (node 0) we have the three alternatives:

(1) no program EMV = -9.525

(2) program EMV = -7

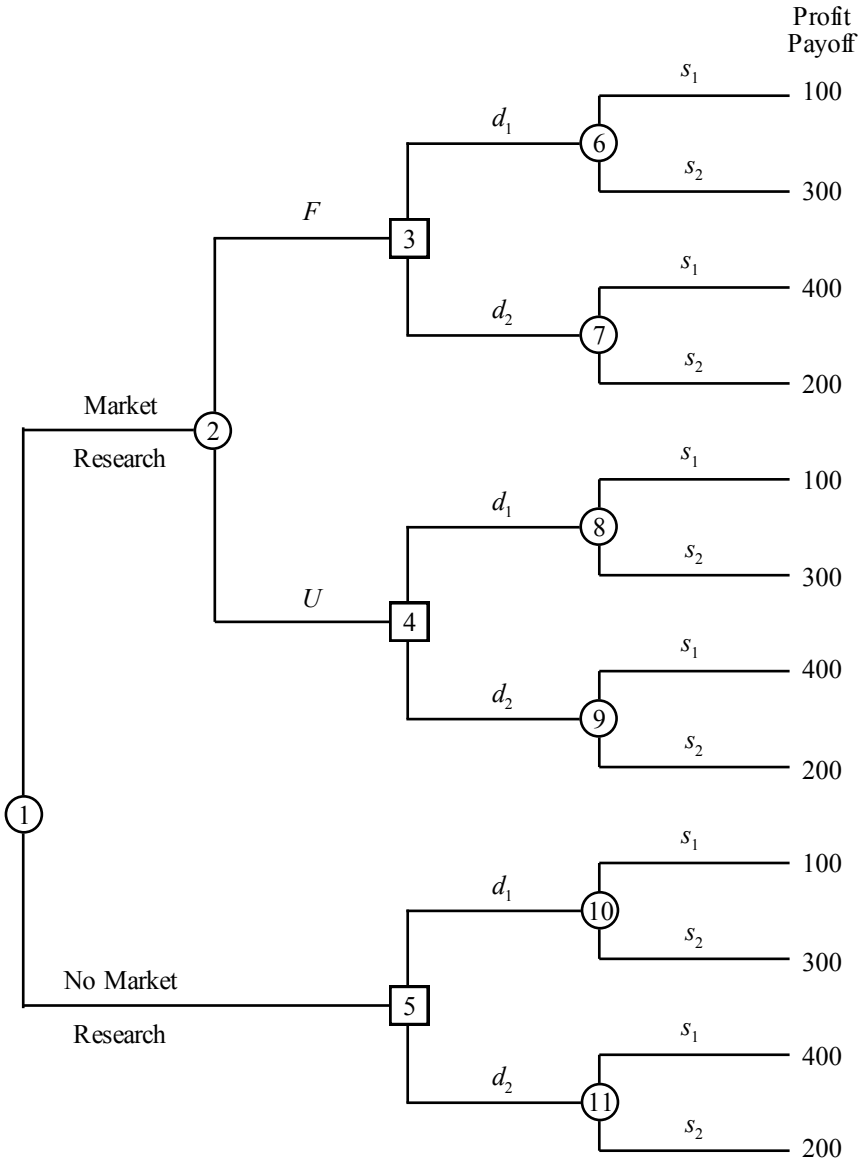
(3) early warning EMV = -10.5

Hence the best alternative is alternative 2, institute a program costing A$7m, leading to an EMV of -A$7m.

Note here that it is clear that the concept of the vaccination program being an insurance against all possible losses could have enabled us to have drawn a much simpler decision tree (e.g. chance node 3 could be transformed into a "terminal" node of cost -A$7m and nodes 7,13,14,15 dropped altogether (similarly for nodes 8,18,19,20)). However, for clarity, we have presented the decision tree as given above.

b. With respect to the last part of the question mention discounting, alternative value for a chance node (other than EMV), changing the decision node ("choose highest EMV alternative") rule and briefly discuss whether appropriate/inappropriate.

8. a.



b. EV (node 6) = 0.57(100) + 0.43(300) = 186
EV (node 7) = 0.57(400) + 0.43(200) = 314
EV (node 8) = 0.18(100) + 0.82(300) = 264

EV (node 9) = 0.18(400) + 0.82(200) = 236
EV (node 10) = 0.40(100) + 0.60(300) = 220
EV (node 11) = 0.40(400) + 0.60(200) = 280

EV (node 3) = Max(186,314) = 314 d_2
EV (node 4) = Max(264,236) = 264 d_1
EV (node 5) = Max(220,280) = 280 d_2

EV (node 2) = 0.56(314) + 0.44(264) = 292
EV (node 1) = Max(292,280) = 292

∴ Market Research
If Favourable, decision d_2
If Unfavourable, decision d_1

9. a. The appropriate decision matrix is:

	defective rate			
	0%	1%	2%	3%
d_1: 100% inspection	250	250	250	250
d_2: No inspection	0	125	250	375

b. EMV (d_1)= 250

EMV (d_2)= .5 (0) + .25 (125) + .40 (250) + .20 (375)
= 206.25

As the `payoff's' tabulated are actually costs, we choose the EMV which is smaller here i.e. d_2 is the optimum decision and the manager's view is supported.

iii.

10. a. Outcome 1 (€ in 000s)

Bid	-€200
Contract	-2000
Market Research	-150
High Demand	+5000
	€2650

Outcome 2 (€ in 000s)

Bid	-€200
Contract	-2000
Market Research	-150
Moderate Demand	+3000
	€650

b. EV (node 8) = 0.85(2650) + 0.15(650) = 2350 — S_1

EV (node 5) = Max(2350, 1150) = 2350 Decision: Build — S_2

EV (node 9) = 0.225(2650) + 0.775(650) = 1100 — S_3

EV (node 6) = Max(1100, 1150) = 1150 Decision: Sell — S_4

EV (node 10) = 0.6(2800) + 0.4(800)= 2000 — S_1

EV (node 7) = Max(2000, 1300) = 2000 Decision: Build — S_2

EV (node 4) = 0.6 EV(node 5) + 0.4 EV(node 6) = 0.6(2350) + 0.4(1150) = 1870 — S_3

EV (node 3) = MAX (EV(node 4), EV (node 7)) = Max (1870, 2000) = 2000
Decision: No Market Research — S_4

EV (node 2) = 0.8 EV(node 3) + 0.2 (-200) = 0.8(2000) + 0.2(-200) = 1560

EV (node 1) = MAX (EV(node 2), 0) = Max (1560, 0) = 1560
Decision: Bid on Contract

Decision Strategy:

Bid on the Contract
Do not do the Market Research
Build the Complex

Expected Value is €1,560,000

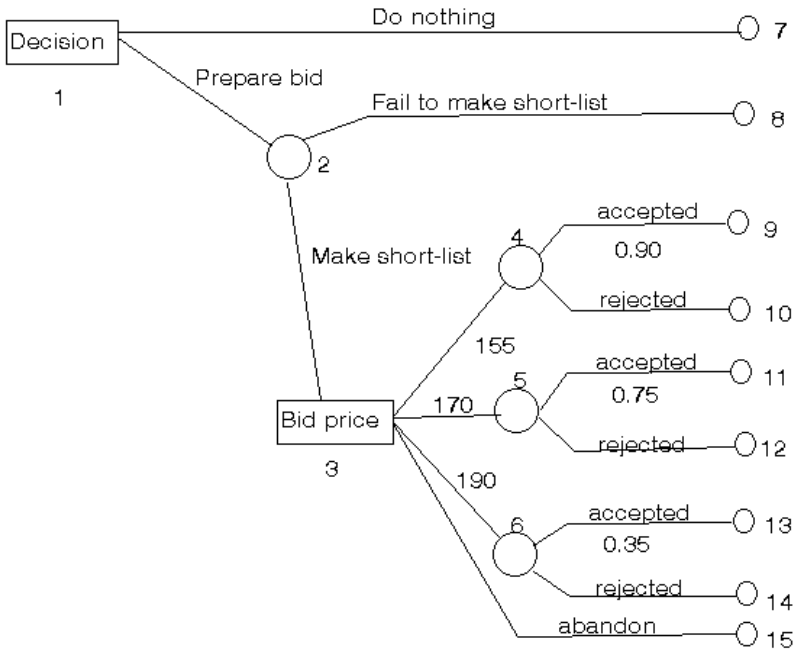
c. Compare Expected Values at nodes 4 and 7.

EV(node 4) = 1870 Includes €150 cost for research
EV (node 7) = 2000

Difference is 2000 - 1870 = €130

Market research cost would have to be lowered €130,000 to €20,000 or less to make undertaking the research desirable.

11. a. The decision tree (see http://people.brunel.ac.uk/~mastjjb/jeb/or/decmore.html) for the problem is shown below.



Below we carry out step 1 of the decision tree solution procedure which (for this example) involves working out the total profit for each of the paths from the initial node to the terminal node (all figures in YTL'000).

Step 1

- path to terminal node 7 - the company do nothing

Total profit = 0

- path to terminal node 8 - the company prepare the bid but fail to make the short-list

Total cost = 10 Total profit = -10

- path to terminal node 9 - the company prepare the bid, make the short-list and their bid of YTL155K is accepted

Total cost = 10 + 5 + 127 Total revenue = 155 Total profit = 13

- path to terminal node 10 - the company prepare the bid, make the short-list but their bid of YTL155K is unsuccessful

Total cost = 10 + 5 Total profit = -15

- path to terminal node 11 - the company prepare the bid, make the short-list and their bid of YTL170K is accepted

Total cost = 10 + 5 + 127 Total revenue = 170 Total profit = 28

- path to terminal node 12 - the company prepare the bid, make the short-list but their bid of YTL170K is unsuccessful

Total cost = 10 + 5 Total profit = -15

- path to terminal node 13 - the company prepare the bid, make the short-list and their bid of YTL190K is accepted

Total cost = 10 + 5 + 127 Total revenue = 190 Total profit = 48

- path to terminal node 14 - the company prepare the bid, make the short-list but their bid of YTL190K is unsuccessful

Total cost = 10 + 5 Total profit = -15

- path to terminal node 15 - the company prepare the bid and make the short-list and then decide to abandon bidding (an implicit option available to the company)

Total cost = 10 + 5 Total profit = -15

Hence we can arrive at the table below indicating for each branch the total profit involved in that branch from the initial node to the terminal node.

Terminal node	Total profit YTL
7	0
8	-10
9	13
10	-15
11	28
11	-15
13	48
14	-15
15	-15

We can now carry out the second step of the decision tree solution procedure where we work from the right-hand side of the diagram back to the left-hand side.

Step 2

Consider chance node 4 with branches to terminal nodes 9 and 10 emanating from it. The expected monetary value for this chance node is given by 0.90(13) + 0.10(-15) = 10.2

Similarly the EMV for chance node 5 is given by 0.75(28) + 0.25(-15) = 17.25

The EMV for chance node 6 is given by 0.35(48) + 0.65(-15) = 7.05

Hence at the bid price decision node we have the four alternatives

(1) bid YTL155K EMV = 10.2

(2) bid YTL170K EMV = 17.25

(3) bid YTL190K EMV = 7.05

(4) abandon the bidding EMV = -15

Hence the best alternative is to bid YTL170K leading to an EMV of 17.25

Hence at chance node 2 the EMV is given by 0.50(17.25) + 0.50(-10) = 3.625

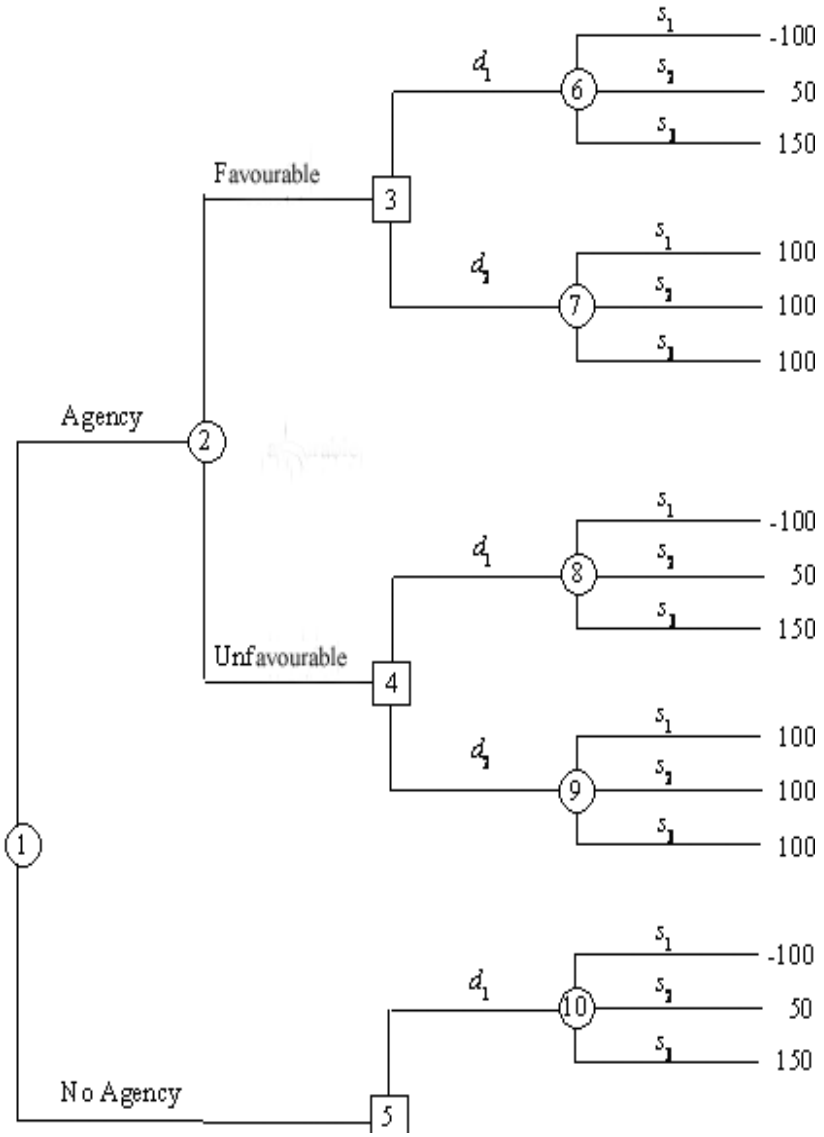
Hence at the initial decision node we have the two alternatives

(1) prepare bid EMV = 3.625

(2) do nothing EMV = 0

Hence the best alternative is to prepare the bid leading to an EMV of YTL3625. In the event that the company is short-listed then (as discussed above) it should bid YTL170,000.

12. a.



b. Using node 5,

EV (node 10) = 0.20(-100) + 0.30(50) + 0.50(150) = 70
EV (node 11) = 100

Decision Sell Expected Value = €100

c. Expected value with perfect information = 0.20(100) + 0.30(100) + 0.50(150) = €125

EVPI = €125 - €100 = €25

d. EV (node 6) = 0.09(-100) + 0.26(50) + 0.65(150) = 101.5
EV (node 7) = 100
EV (node 8) = 0.45(-100) + 0.39(50) + 0.16(150) = -1.5
EV (node 9) = 100

EV (node 3) = Max(101.5,100) = 101.5 Produce
EV (node 4) = Max(-1.5,100) = 100 Sell

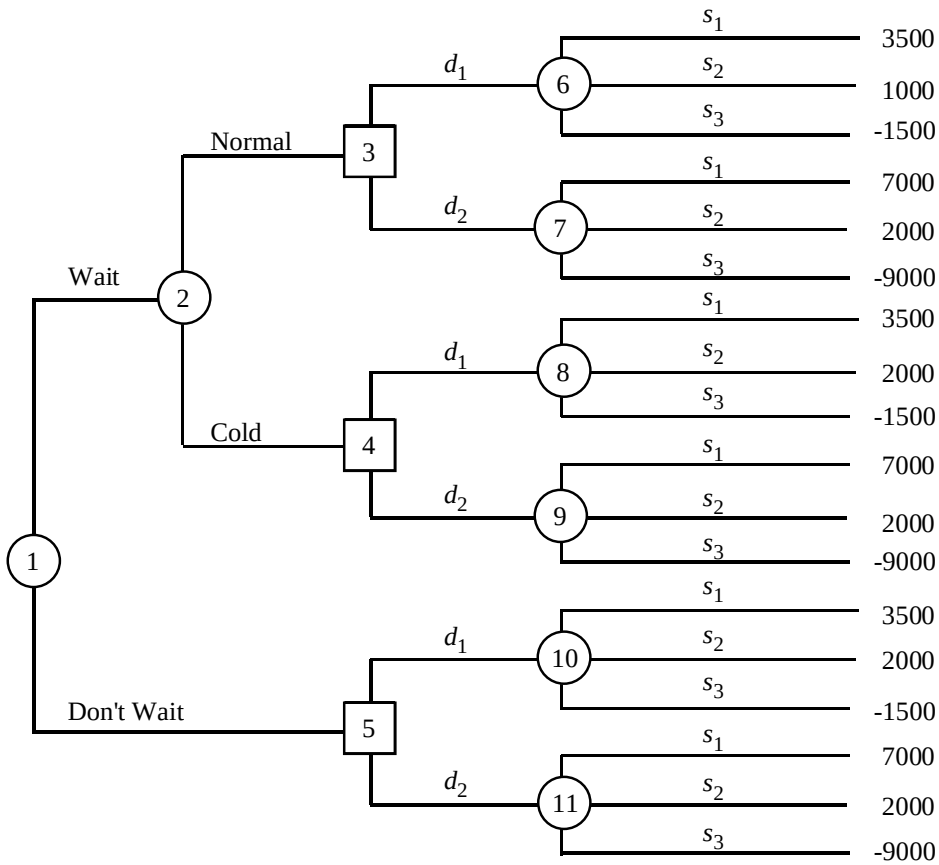
EV (node 2) = 0.69(101.5) + 0.31(100) = 101.04

If Favourable, Produce
If Unfavourable, Sell EV = €101.04

e. EVSI = €101.04 - 100 = €1.04 or €1,040.

f. No, maximum Hale should pay is €1,040.

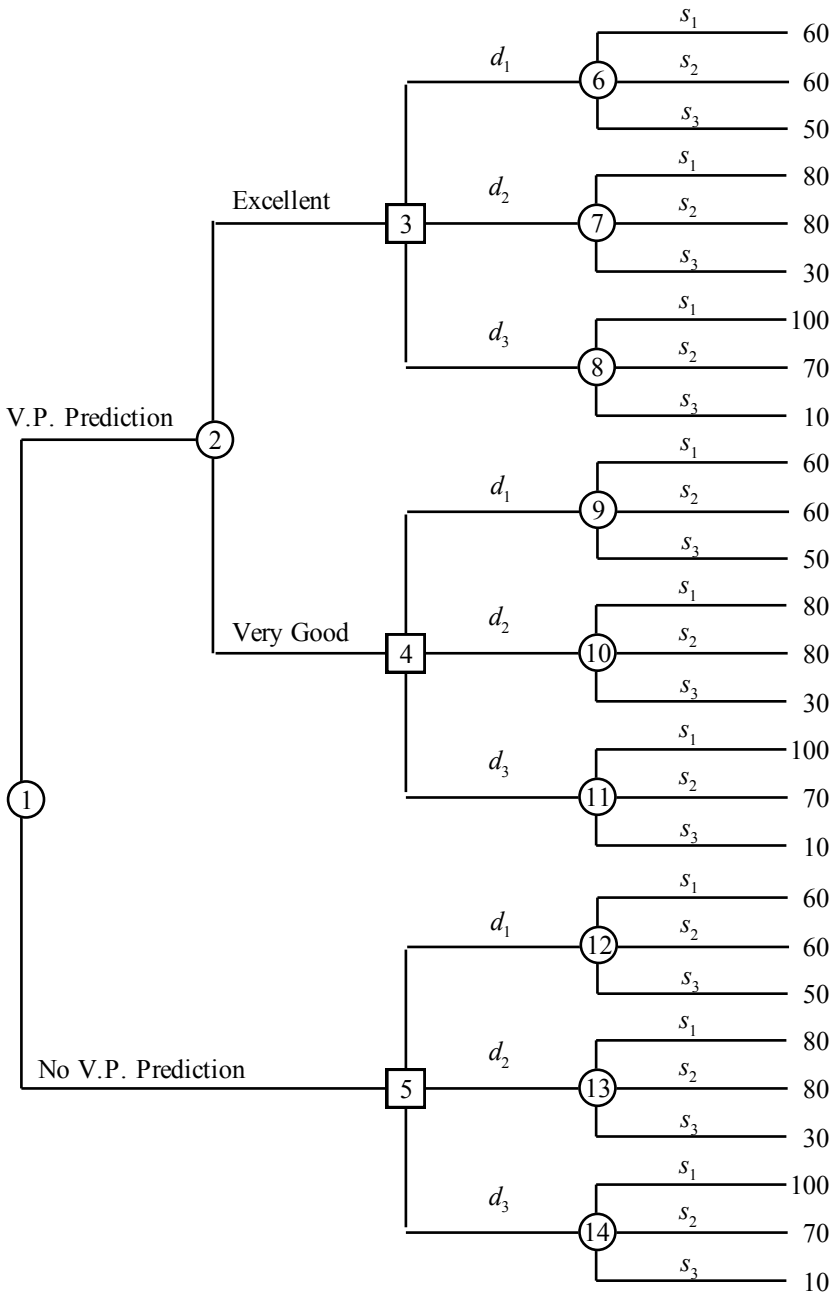
g. No agency; sell the pilot.



13. a. EV (1 lot) = 0.3(60) + 0.3(60) + 0.4(50) = 56
EV (2 lots) = 0.3(80) + 0.3(80) + 0.4(30) = 60
EV (3 lots) = 0.3 (100) + 0.3(70) + 0.4(10) = 55

Decision: Order 2 lots Expected Value €60,000

b. The following decision tree applies.



Calculations

EV (node 6) = 0.34(60) + 0.32(60) + 0.34(50) = 56.6
EV (node 7) = 0.34(80) + 0.32(80) + 0.34(30) = 63.0
EV (node 8) = 0.34(100) + 0.32(70) + 0.34(10) = 59.8
EV (node 9) = 0.20(60) + 0.26(60) + 0.54(50) = 54.6
EV (node 10) = 0.20(80) + 0.26(80) + 0.54(30) = 53.0
EV (node 11) = 0.20(100) + 0.26(70) + 0.54(10) = 43.6
EV (node 12) = 0.30(60) + 0.30(60) + 0.40(50) = 56.0
EV (node 13) = 0.30(80) + 0.30(80) + 0.40(30) = 60.0
EV (node 14) = 0.30(100) + 0.30(70) + 0.40(10) = 55.0

EV (node 3) = Max(56.6,63.0,59.8) = 63.0 2 lots
EV (node 4) = Max(54.6,53.0,43.6) = 54.6 1 lot
EV (node 5) = Max(56.0,60.0,55.0) = 60.0 2 lots

EV (node 2) = 0.70(63.0) + 0.30(54.6) = 60.5
EV (node 1) = Max(60.5,60.0) = 60.5 Prediction

Optimal Strategy:
If prediction is excellent, 2 lots
If prediction is very good, 1 lot

c. Expected value with perfect information = 0.3(100) + 0.3(80) + 0.4(50) = 74

EVPI = 74 - 60 = 14
EVSI = 60.5 - 60 = 0.5

The EVPI is €14,000, but the V.P's recommendation is only valued at EVSI = €500. This indicates additional information is probably worthwhile. The ability of the consultant to forecast market conditions should be considered.

14.

State of Nature	$P(s_j)$	$P(I \mid s_j)$	$P(I \cap s_j)$	$P(s_j \mid I)$
s_1	0.2	0.10	0.020	0.1905
s_2	0.5	0.05	0.025	0.2381
s_3	0.3	0.20	0.060	0.5714
	1.0	$P(I)$ =	0.105	1.0000

15. a. EV (d_1) = 0.8(15) + 0.2(10) = 14.0
EV (d_2) = 0.8(10) + 0.2(12) = 10.4
EV (d_3) = 0.8(8) + 0.2(20) = 10.4

Decision d_1 Expected Value 14

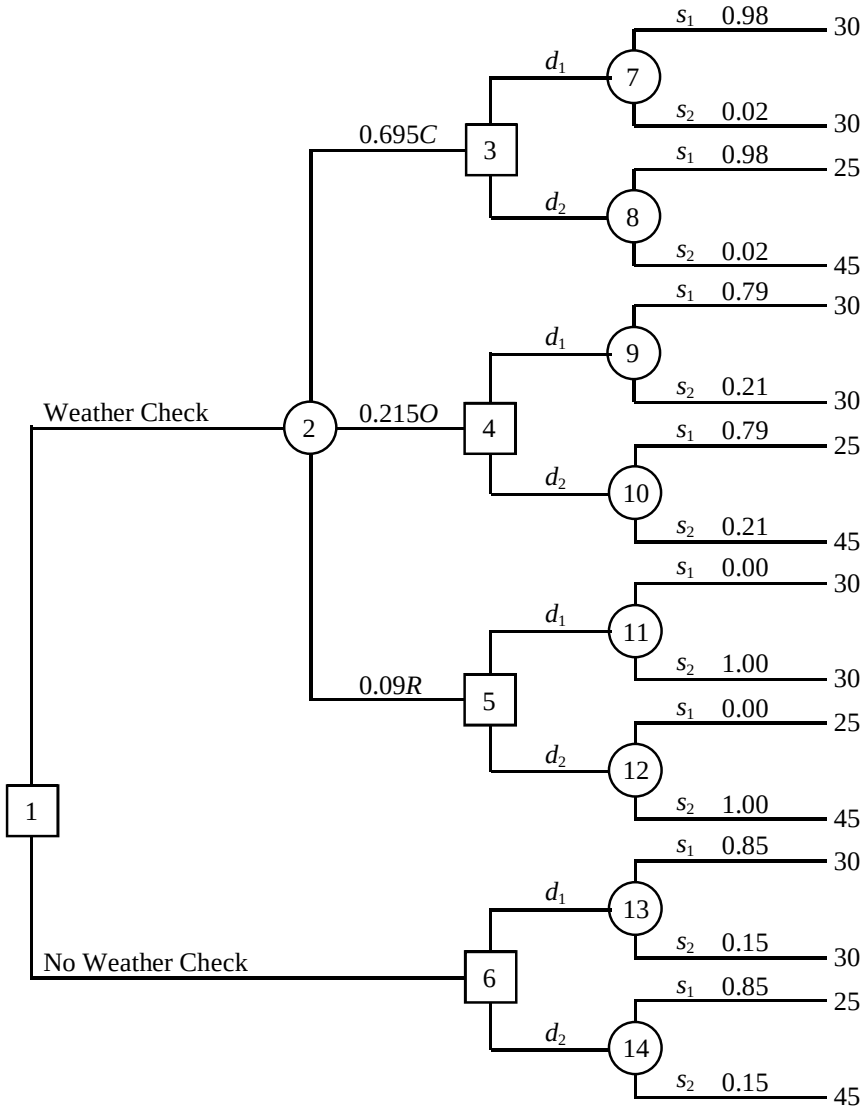
b. Expected value with perfect information = 0.8(15) + 0.2(20) = 16
EVPI = 16 - 14 = 2

c. Indicator I

State of Nature	Prior Probabilities	Conditional Probabilities	Joint Probabilities	Posterior Probabilities
State s1	0.8	0.20	0.16	0.52
State s2	0.2	0.75	0.15	0.48
		P(I) =	0.31	1.00

EV (d_1) = 0.5161(15) + 0.4839(10) = 12.6
EV (d_2) = 0.5161(10) + 0.4839(12) = 11.0
EV (d_3) = 0.5161(8) + 0.4839(20) = 13.8
If indicator I occurs, decision d_3 is recommended.

16. a,b. The revised probabilities are shown on the branches of the decision tree.



EV (node 7) = 30
EV (node 8) = 0.98(25) + 0.02(45) = 25.4
EV (node 9) = 30
EV (node 10) = 0.79(25) + 0.21(45) = 29.2
EV (node 11) = 30
EV (node 12) = 0.00(25) + 1.00(45) = 45.0
EV (node 13) = 30
EV (node 14) = 0.85(25) + 0.15(45) = 28.0

EV (node 3) = Min(30,25.4) = 25.4 Expressway
EV (node 4) = Min(30,29.2) = 29.2 Expressway
EV (node 5) = Min(30,45) = 30.0 Boulevard Periphique
EV (node 6) = Min(30,28) = 28.0 Expressway

EV (node 2) = 0.695(25.4) + 0.215(29.2) + 0.09(30.0) = 26.6
EV (node 1) = Min(26.6,28) = 26.6 Weather Check

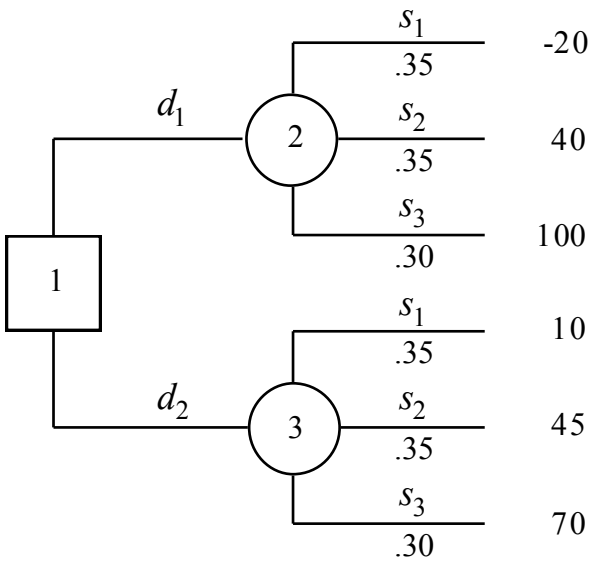
c. Strategy:

Check the weather, take the expressway unless there is rain. If rain, take Boulevard Periphique.

Expected time: 26.6 minutes.

17. a. d_1 = Manufacture component
d_2 = Purchase component

s_1 = Low demand
s_2 = Medium demand
s_3 = High demand



EV(node 2) = (0.35)(-20) + (0.35)(40) + (0.30)(100) = 37

EV(node 3) = (0.35)(10) + (0.35)(45) + (0.30)(70) = 40.25

Recommended decision: d_2 (purchase component)

b. Optimal decision strategy with perfect information:

If s_1 then d_2

If s_2 then d_2

If s_3 then d_1

Expected value of this strategy is 0.35(10) + 0.35(45) + 0.30(100) = 49.25
EVPI = 49.25 - 40.25 = 9 or €9,000

c. If F - Favourable

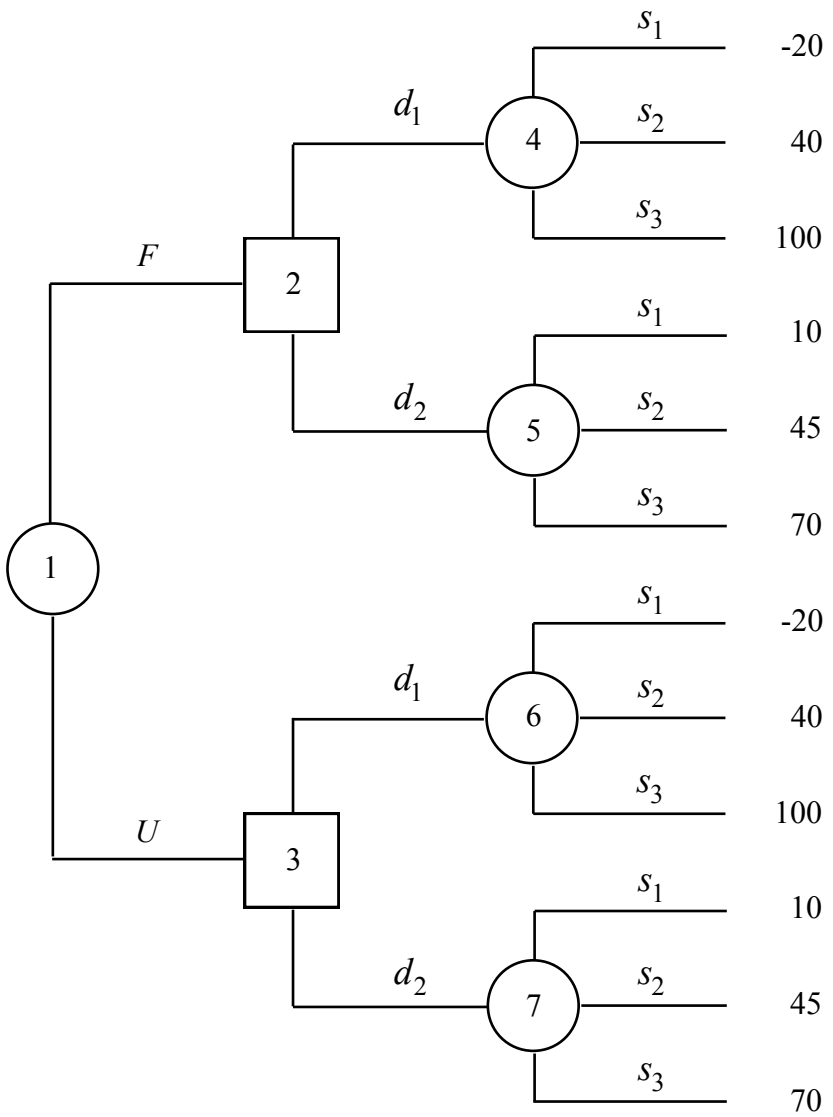
State of Nature	$P(s_j)$	$P(F \mid s_j)$	$P(F \cap s_j)$	$P(s_j \mid F)$
s_1	0.35	0.10	0.035	0.0986
s_2	0.35	0.40	0.140	0.3944
s_3	0.30	0.60	0.180	0.5070
		$P(F)$ =	0.355	

If U - Unfavourable

State of Nature	$P(s_j)$	$P(U \mid s_j)$	$P(U \cap s_j)$	$P(s_j \mid U)$
s_1	0.35	0.90	0.315	0.4884
s_2	0.35	0.60	0.210	0.3256
s_3	0.30	0.40	0.120	0.1860
		$P(U) =$	0.645	

The probability the report will be favourable is $P(F) = 0.355$

d. Assuming the test market study is used, a portion of the decision tree is shown below.



Summary of Calculations

Node	Expected Value
4	64.51
5	54.23
6	21.86
7	32.56

Decision strategy:

If F then d_1 since EV(node 4) > EV(node 5)

If U then d_2 since EV(node 7) > EV(node 6)

EV(node 1) = 0.355(64.51) + 0.645(32.56) = 43.90

e. With no information:

EV(d_1) = 0.35(-20) + 0.35(40) + 0.30(100) = 37

EV(d_2) = 0.35(10) + 0.35(45) + 0.30(70) = 40.25

Recommended decision: d_2

f. Optimal decision strategy with perfect information:

If s_1 then d_2
If s_2 then d_2
If s_3 then d_1

Expected value of this strategy is 0.35(10) + 0.35(45) + 0.30(100) = 49.25
EVPI = 49.25 - 40.25 = 9 or €9,000
Efficiency = (3650 / 9000)100 = 40.6%

Supplementary Exercises:

18. Consider the payoff (€) table, below.

	States of Nature		
Decisions	s_1	s_2	s_3
d_1	100,000	40,000	-60,000
d_2	50,000	20,000	-30,000
d_3	20,000	20,000	-10,000
d_4	40,000	20,000	-60,000

The probabilities associated with the states of nature here are as follows:

$P(s_1) = 0.1$, $P(s_2) = 0.3$ and $P(s_3) = 0.6$

What is the decision-maker's best option using the Expected Value (EV) criterion?

19. The owner of SBL is trying to decide whether or not to undertake one of two contracts M and N that have been offered to him. Profits are dependent on the economic conditions that prevail and three different scenarios (states of nature) are thought possible in each case. Probabilities of these occurring and corresponding profits (in €'000's) are as follows:

Contract M		Contract N	
Probability	Profit	Probability	Profit
0.6	80	0.5	50
0.1	10	0.3	30
0.3	-30	0.2	-10

Use the EV criterion to determine which of the two contracts SBL should opt for.

20. A firm is offered a contract to develop a special turbine engine. The contract stipulates that if the engine is not developed within two years from the date of the contract, the contract is void. If the engine is developed in time, the firm's profit is €0.25 million; if not, the cost to the firm is €1.25 million. The firm's current assets are €1.50 million. The R&D department assesses the probability of successful development within the two-year period as 0.9.

If the firm uses expected profit as its criterion, what is the optimal action?

21. A team of advisers has just completed an analysis of the prospects next year for five well-known stocks. The prospects depend upon the strength of the economy. The team have supposed that the economy will be in one of four states:

decline (s_1),

holding steady (s_2)

slight improvement (s_3) and

major expansion (s_4)

The percentage sterling growth for the stocks corresponding to each state of the economy is forecast as:

Stock	State of the economy			
	s_1	s_2	s_3	s_4
A	-35	-15	15	60
B	-25	-5	0	35
C	-15	0	5	20
D	-5	5	10	20
E	0	5	5	10

The probabilities associated with the different states of the economy occurring have been estimated as follows:

$p(s_1) = 0.1$ $p(s_2) = 0.3$ $p(s_3) = 0.4$ $p(s_4) = 0.2$

a. Which stock is preferable according to the EV criterion?

b. What is the expected value of perfect information (EVPI) for these data? How would you interpret the EVPI?

22. A large engineering firm has invented a new product which it is not presently equipped to manufacture itself. It has three choices a) to acquire the necessary machinery to carry out the manufacture itself b) sell the rights to the product to another firm and c) sell rights on a royalty basis to another firm. There is a 60% chance that the product will catch on and a 40% chance it will flop. Based on the relevant payoffs tabulated below (in €000s), what is the firm's best strategy?

	Product	
	Catches on	Flops
Manufacture	100	-175
Sell all rights	10	10
Sell on royalty basis	55	5

23. Following on from Exercise 22, if the firm opts to manufacture the product itself, there is the possibility of improving the design of the product so that in the event of the design being successful, an extra €150,000 profit will be made but if it fails there will be a €100,000 loss. (Note that the probability of the new design failing is estimated at 20%.) What is the firm's optimal strategy in this case?

24. A quality control procedure involves 100% inspection of parts received from a supplier. Historical records show the following defective rates have been observed:

% Defective	Probability
0	0.15
1	0.25
2	0.40
3	0.20

The cost of 100% inspection is €250 for each shipment of 500 parts. If the shipment is not 100% inspected, defective parts will cause rework problems later in production process. The rework is €25 for each defective part.

The plant manager is considering eliminating the inspection process in order to save the €250 inspection cost per shipment.

a. Provide a payoff table representation for this problem.

b. Would you support the plant manager's view as regards eliminating the inspection process? Explain.

c. Represent the problem as a decision tree.

25. A company has a new product which it could sell through its retail outlets. The company currently feels that the probability that the product would be a success is 0.2. Prior to putting the product in the shops, the company can conduct a market research survey which would provide the company with a decision to

Sell the product

Don't sell the product

The market researchers have kept statistics on the performance of similar ventures and have found

p(success | survey says sell & the company sells) = 0.6

p(success | survey says don't sell & the company sells) = 0.1

Given the probability of the survey saying sell = 0.2, cost of product launch = €75,000, costs of survey = €15,000 and expected revenue over planning horizon is €338,000, what should the company do?

26. An independent operator in the oil business has to decide whether or not to buy a tract of land for the possibility that a reservoir of crude oil lies underneath. A preliminary geological survey of the surface suggests the following prior probabilities for the possible quantities of crude oil:

State	Prior probability
No oil	0.8
l million barrels	0.1
2 million barrels	0.1

Before making his decision the operator has the option of carrying out a seismic experiment which will yield either a 'high' or a 'low' reading. Past experience suggests the following conditional probabilities for the outcome of the experiment:

	'high' reading	'low' reading
No oil	0.2	0.8
1 million barrels	0.4	0.6
2 million barrels	0.9	0.1

If the operator buys the land, he loses €100,000 if it contains no oil, but gains €400,000 or €800,000 if it yields 1 million or 2 million barrels respectively. The cost of the seismic experiment is €10,000.

Should the operator make his decision immediately without experimentation, or should he carry out the seismic experiment and base his decision on the outcome of the experiment?

27. A TV production firm is considering producing a pilot for a comedy series for a major television network. The network may reject the pilot and series (s_1), but it may also purchase the program for 1 (s_2) or 2 years (s_3). The firm may decide to produce the pilot or transfer the rights for the series to a competitor for €150,000. The firm's profits are summarised in the following profits (in thousands of euros) payoff table.

Payoff Table

Decision Alternative	State of Nature		
	Reject s_1	1 Year s_2	2 Years s_3
Produce pilot, d_1	-150	75	225
Sell to competitor, d_2	150	150	150

The probabilities for the states of nature are $P(s_1) = .3$, $P(s_2) = .3$, and $P(s_3) = .4$.

For a consulting fee of €37500, an agency will review the plans for the comedy series and indicates the overall chances of a favourable network reaction to the series. If the special agency review results in a favourable (F) or an unfavourable (U) evaluation, what should the firm decision strategy be? Suppose the firm believes that the following conditional probabilities are realistic appraisals of the agency's evaluation accuracy.

$$p(F/s_1) = 0.3 \qquad p(U/s_1) = 0.7$$
$$p(F/s_2) = 0.6 \qquad p(U/s_2) = 0.4$$
$$p(F/s_3) = 0.9 \qquad p(U/s_3) = 0.1$$

a. Show the decision tree for this new problem.

b. Calculate the posterior probabilities.

c. What is the recommended decision strategy and the expected value, assuming that the agency information is obtained?

28. Northern Equipment Ltd has developed a new product. The company can choose to produce the product or sell its right for €12000. If it chooses to produce the product, the profitability of the venture depends on the company's ability to market the product. It has sufficient access to retail outlets so that it can guarantee sales of 10 units. On the other hand, if this product catches on, it can sell 100 units. The company believes that both sales alternatives are equally likely and that all other alternatives are negligible.

The cost of setting up the assembly line is €3000. The difference between the selling price and the variable cost is €200 per unit. Market research can be performed at a cost of €1500 to determine which of the two levels of demand is more realistic. Previous experience indicates that whether demand is high or low such market research can provide correct forecasts on 66.5% of occasions.

a. Draw and properly label a decision tree for the problem.

b. Calculate the posterior probabilities.

c. Evaluate the decision tree and determine the optimal policy that can maximise the expected total profit. State the optimal policy and the maximum profit.

29. The Shetlands Oil Refinery plans to produce a new product (PN). The company can choose to sell PN through its own outlets or sell a fixed amount of PN to a wholesaler for a fixed profit of £800,000 every week. If Shetlands chooses to sell the product through its own outlets, the profitability of the sales depends on the company's ability to market the product. Shetlands is sure that it can sell 2000 tonnes of PN per week and the potential sales could be as high as 3500 tonnes of PN per week. Shetlands believes that the two sales alternatives are equally likely and that all

other alternatives are negligible. The unit profit is £330 per tonne of PN. Before deciding how to sell PN,

Shetlands can choose to conduct a market survey to identify which of the two levels of demand is more realistic. Market research can be performed at a cost of £112,500.

Previous experience indicates that whether demand is high or low such market research can provide correct forecasts 70% of time.

a. Draw and properly label a decision tree for the problem.

b. Use Bayes theorem to calculate the posterior probabilities.

c. Evaluate the decision tree and determine the optimal policy that can maximise the expected total profit. State the optimal policy and the maximum profit.

30. ToyKing Ltd has designed a new type of toy: Robocar. ToyKing can choose to manufacture the product or sell the rights to the product to another manufacturer for €200,000. ToyKing developed and manufactured similar products before and anticipates that the demand for the new product could be low or high; the probabilities and profits for low and high demands are estimated as shown in the following table. The company objective is to maximise expected profit.

	Low Demand	High Demand
Probability	0.45	0.55
Profit (€)	75,000	300,000

a. Draw and properly label a decision tree for the problem.

b. Evaluate the decision tree and identify the optimal policy for the company.

The company can also choose to conduct a market survey at a cost of €15,000 before deciding whether to sell the rights or produce the product itself. The market research may lead to an "optimistic" forecast or a "pessimistic" forecast. Previous experience shows that when the demand turned out to be high, the forecast had been "optimistic" on 90% of occasions; and when the demand turned out to be low, the forecast had been "pessimistic" on 40% of occasions.

c. Draw and properly label the new decision tree.

d. Calculate the posterior probabilities.

e. Evaluate the new decision tree and determine the new optimal policy.

31. The Pine Furniture Company has designed a new type of luxurious furniture.

The company can choose to produce the furniture or sell its right for €37,500. If Pine chooses to produce the furniture, the profitability of the venture depends on the company's ability to market the furniture. Pine can guarantee sales of 10 units and the potential sales could be as high as 100 units. Pine believes that both sales alternatives are equally likely and that the probabilities of all other sales outcomes are negligible. The cost of setting up the production line is €5,000. The difference between the selling price and the variable cost (exclusive of setting up cost) is €900 per unit. Market research can be performed at a cost of €7,000 to determine which of the two levels of demand is more realistic. Previous experience indicates that whether demand turns out to be high or low, such market research can provide correct forecasts 80% of the time.

a. Draw and properly label a decision tree for the problem.

b. Calculate the posterior probabilities.

c. Evaluate the decision tree and determine the optimal policy that can maximise the expected total profit. State the optimal policy and the maximum profit.

Supplementary Exercises Solutions:

18. Note that the payoff values used in calculations that follow are in €000's.

EV(d_1) = 0.1(100) + 0.3(40) + 0.6(-60) = -14
EV(d_2) = 0.1(50) + 0.3(20) + 0.6(-30) = -7
EV(d_3) = 0.1(20) + 0.3(20) + 0.6(-10) = 2
EV(d_4) = 0.1(40) + 0.3(20) + 0.6(-60) = -26

Recommended decision: d_3

19. Note that the payoff values used in calculations that follow are in €000's.

EV(M) = 0.6(80) + 0.1(10) + 0.3(-30) = 40
EV(N) = 0.5(50) + 0.3(30) + 0.2(-10) = 32

So SBL should opt for contract M

20. Note that the payoff values used in calculations that follow are in €m's.

EV(accept contract) = 0.9(.25) + 0.1(1.25) = 1
EV(not accept project) = 0

So the firm should accept the contract.

21. a.
EV(A) = 0.1(-35) + 0.3(-15) + 0.4(15) +0.2(60) = 10
EV(B) = 0.1(-25) + 0.3(-5) + 0.4(0) +0.2(35) = 3
EV(C) = 0.1(-15) + 0.3(0) + 0.4(5) +0.2(20) = 4.5
EV(D) = 0.1(-5) + 0.3(5) + 0.4(10) +0.2(20) = 9
EV(E) = 0.1(0) + 0.3(5) + 0.4(5) +0.2(10) = 5.5

So stock A is preferred.

b. EVPI = 0.1(0)+0.3(5)+0.4(15)+0.2(60) – 10 =9.5; This is the amount in terms of percentage sterling growth we would be prepared to pay to know which of the four states of nature was certain to arise.

22. Note that the payoff values used in calculations that follow are in €000's.

EV(Manufacture) = 0.6(100) + 0.4(-175) = -10
EV(Sell all rights) = 0.6(10) + 0.4(10) = 10
EV(Sell on royalty basis) = 0.6(55) + 0.4(5) = 35

The firm's best strategy is to: Sell on royalty basis

23 If the firm opts to manufacture the product and further, to improve it successfully, the payoff (€000's) will be 100 + 150 = 250. If the improvement is not successful however the payoff will be -100. Given the p(improvement not successful) = 0.2 the

EV (improvement) = 0.8(250)+0.2(-100) = 180

As this value exceeds the 100 payoff for the Manufacture choice based on the assumption that the product takes off the firm would therefore opt to improve the product. We now have the revised:

EV(Manufacture) = 0.6(180) + 0.4(-175) = 38

The firm's best strategy is now to: Manufacture and improve the product.

24 a. There are two decision options: 100% inspection and Not 100% inspection
The states of nature correspond to the four % defective outcomes.

Following on, the payoff (€) table takes the form:

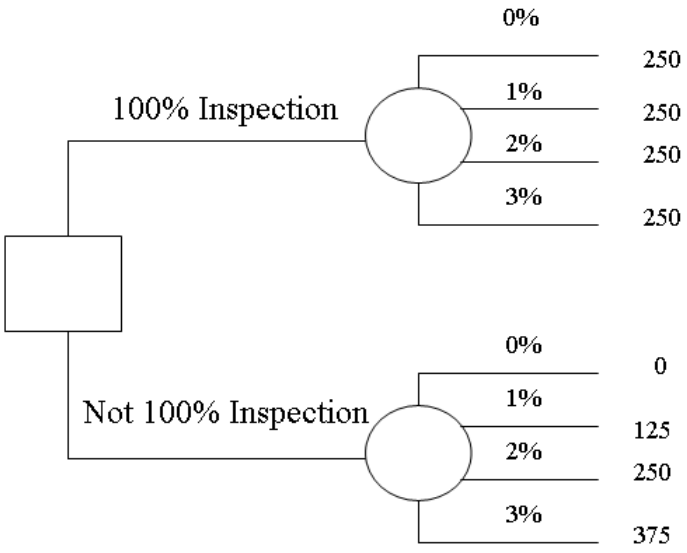
	% Defectives			
	0	1	2	3
100% inspection	250	250	250	250
Not 100% inspection	0	125	250	250

where for example, the €125 payoff on the bottom line corresponds with a 1% defective rate and 1% of a 500 part shipment = 5 defective parts, each costing €25 to rework. Note that in the case of the 100% inspection decision option, the payoff is €250 irrespective of the % defective rate detected.

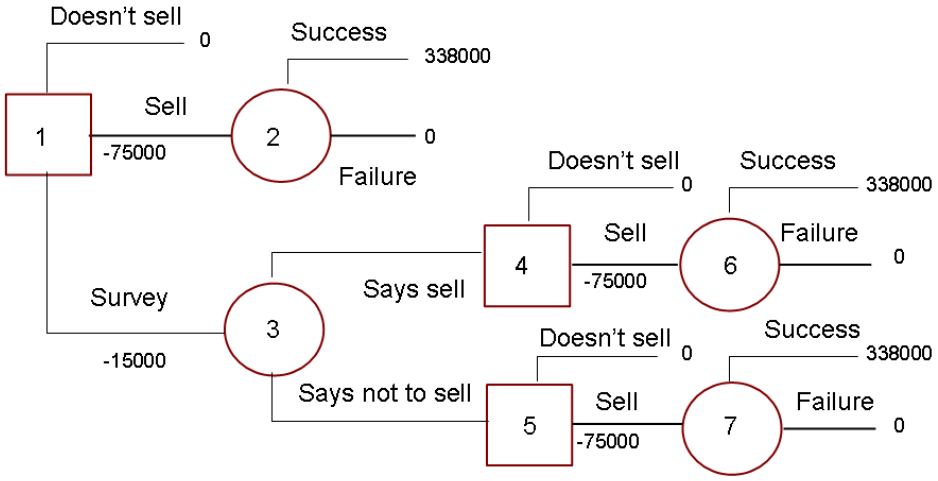
b. EV(100% inspection) = 250
EV(Not 100% inspection) = 0.15(0) + 0.25(125) + 0.4(250) + 0.2(375) = 206.25.

Thus the manager is right that on average it will cost less not to carry out a 100% inspection of shipments.

c.



25. A decision tree representation of the problem is as follows:



Starting the backward pass calculations to compute expected values from nodes 7 to 2 provides the following results.

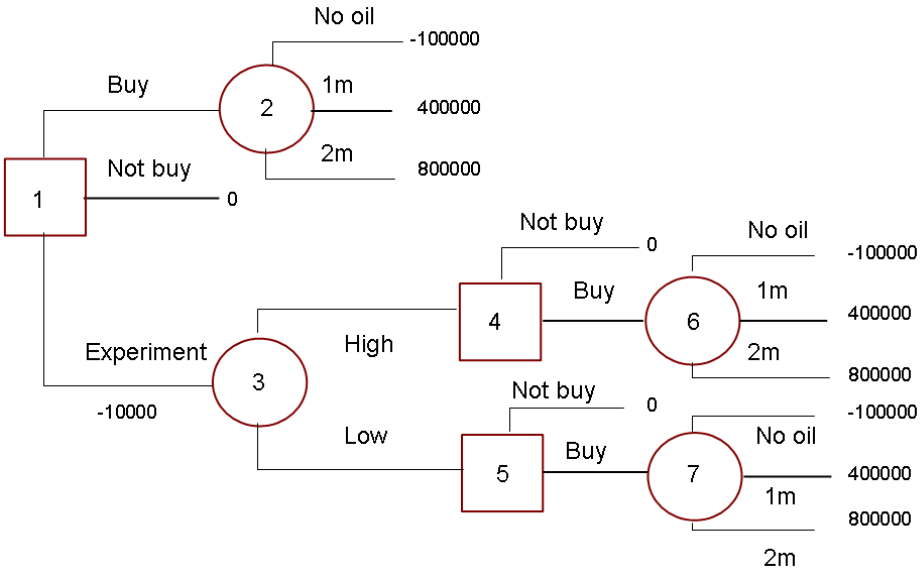
EV(Node 7)= 0.1(338000) + 0.9(0) = 33800
EV(Node 6) = 0.6(338000) + 0.9(0) = 202800
EV(Node 5) = max(0, 33800 - 75000) = 0
EV(Node 4) = max(0, 202800 - 75000) = 127800
EV(Node 3) = 0.2(127800) + .0.8(0) = 25560
EV(Node 2) = 0.2(338000) + 0.8(0) = 67600

At node 1 the EVs are therefore:

EV (doesn't sell) = 0
EV(sell) = 67600 - 75000 = -7400
EV(survey) = 25560

So the firm's best strategy is to carry out the survey and to sell if the survey says so but not sell if it says not to.

26.



The conditional probabilities provided take the form:

P(High |No oil) = 0.2
P(Low |No oil) = 0.8
P(High | 1m) = 0.4
P(Low | 1m) = 0.6
P(High | 2m) = 0.9
P(Low | 2m) = 0.1

Using the prior probabilities P(No oil) = 0.8, P(1m) = 0.1 and P(2m) = 0.1 then

P(High) = P(High |No oil) P(No oil) + P(High | 1m) P(1m) + P(High | 2m) P(2m)
= 0.2(0.8) + 0.4(0.1) + 0.9(0.1) = 0.29.

Therefore P(Low) = 0.71 = 1- P(High)
Thus:

P(No oil |High) = 0.2(0.8)/0.29 = 0.55
P(1m |High) = 0.4(0.1)/0.29 = 0.14
P(2m |High) = 0.9(0.1)/0.29 = 0.31

Similarly it can be found
PNo oil |High) = 0.90
P(1m |High) = 0.08
P(2m |High) = 0.02

Using these last results for the backward pass calculations of expected values from nodes 7 to 2 provides the following results.

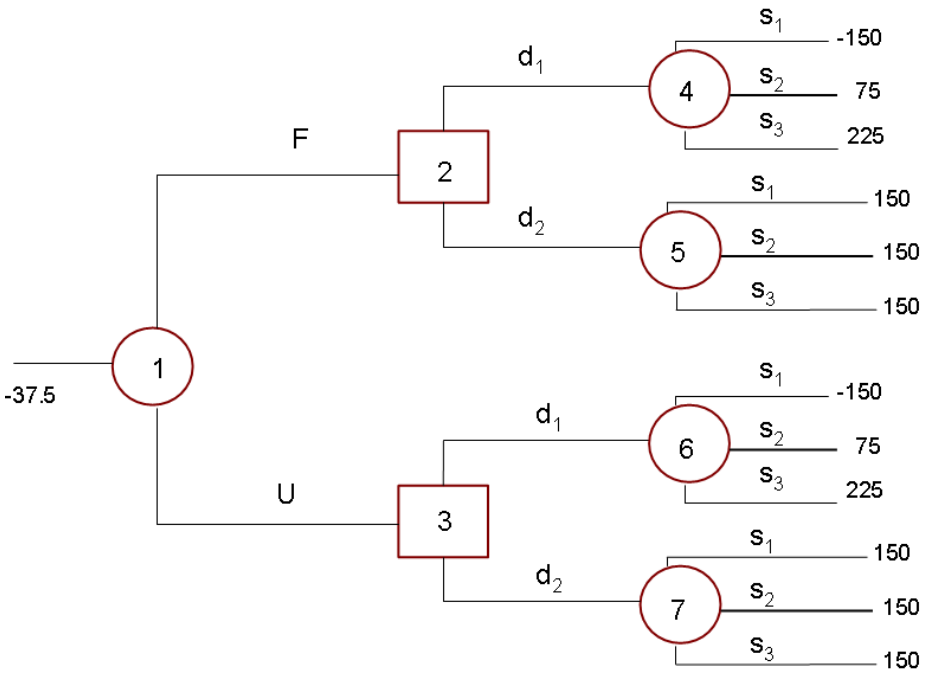
EV(Node 7)= 0.9(-100000) + 0.08(400000) + 0.02(800000) = -42000
EV(Node 6) = 0.9(-100000) + 0.08(400000) + 0.02(800000) = 249000
EV(Node 5) = max(0, -42000) = 0
EV(Node 4) = max(0, 249000) = 249000
EV(Node 3) = 0.29(249000) + .0.8(0) = 72210
EV(Node 2) = 0.8(-100000) + 0.1(400000) + 0.1(800000) = 40000

At node 1 the EVs are therefore:

EV(Buy) = 40000
EV (Not buy) = 0
EV(Experiment) = 72210 – 10000 = 62210

So the firm's best strategy is to carry out the experiment and to buy the land if a high reading is obtained but not buy the land if a low reading is obtained.

27 a.



b. $P(F) = P(F|s_1)\,P(s_1) + P(F|s_2)\,P(s_2) + P(F|s_3)\,P(s_3)$
$= 0.3(0.3) + 0.6(0.3) + 0.9(0.4) = 0.63.$

Therefore $P(U) = 0.37 = 1 - P(F)$

Thus:

$Ps_1|F) = 0.3(0.3)/0.63 = 0.14$
$P(s_2|F) = 0.6(0.3)/0.63 = 0.29$
$P(s_3|F) = 0.9(0.4)/0.63 = 0.57$

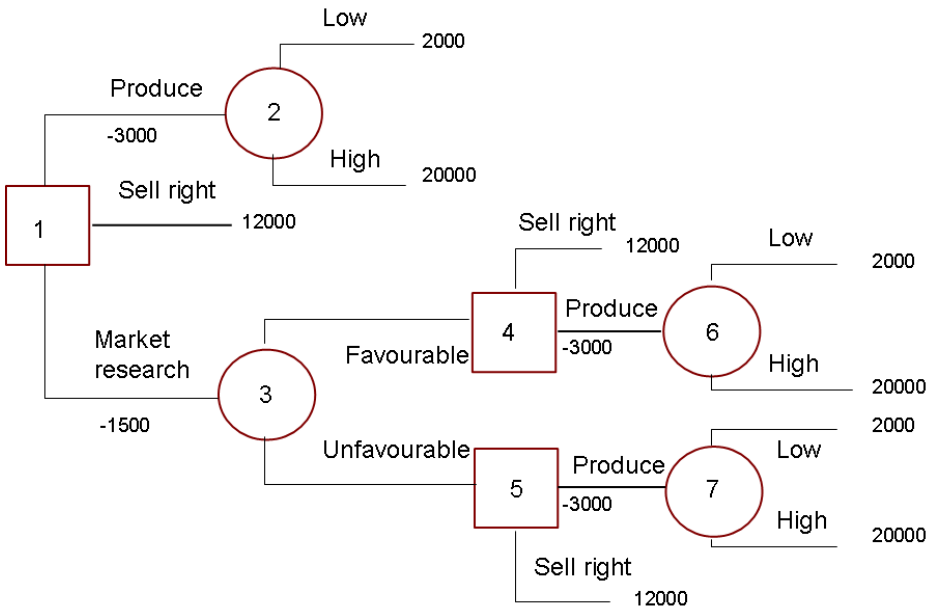
Similarly it can be found
$P(s_1|U) = 0.57$
$P(s_2|U) = 0.32$
$P(s_3|U) = 0.11$

Using these last results for the backward pass calculations of expected values from nodes 7 to 2 provides the following results.

EV(Node 7) = 0.57(150) + 0.32(150) + 0.11(150) = 150
EV(Node 6) = 0.57(-150) + 0.32(75) + 0.11(225) = -36.75
EV(Node 3) = max(150, -36.75) = 150
EV(Node 5) = 0.14(150) + 0.29(150) + 0.57(150) = 150
EV(Node 4) = 0.14(-150) + 0.29(75) + 0.57(225) = 129
EV(Node 2) = 0.37(150) + 0.63(150) = 150

The EV at node 1 is therefore 150 - 37.5 = 112.5 and irrespective of whether the survey result is favourable or unfavourable the firm should opt to sell to the competitor (decision d_2).

28. a. The decision tree is as follows:



Note that the payoffs here correspond with the relevant level of demand X €200, the difference between the selling price and the variable cost per unit

b. Let L = Low demand, H = High demand
F= favourable market research outcome (high demand)
U = unfavourable market research outcome (low demand)

Then we are given P(L) = P(H) = 0.5

$$P(F|H) = 0.665 = P(U|L)$$

Thus P(U|H) = P(F|L) = 0.335
Consequently P(F) = P(F|H)P(H) + P(F|L)P(L) = 0.665(0.5) + 0.335(0.5) = 0.5

Therefore P(U) = 0.5.

It follows the posterior probabilities are:

P(H|F) = 0.665(0.5)/0.5 = 0.665
P(L|F) = 0.335(0.5)/0.5 = 0.335

Similarly:

P(H|U) = 0.335(0.5)/0.5 = 0.335
P(L|U) = 0.665(0.5)/0.5 = 0.665

Using these last results for the backward pass calculations of expected values from nodes 7 to 2 provides the following results.

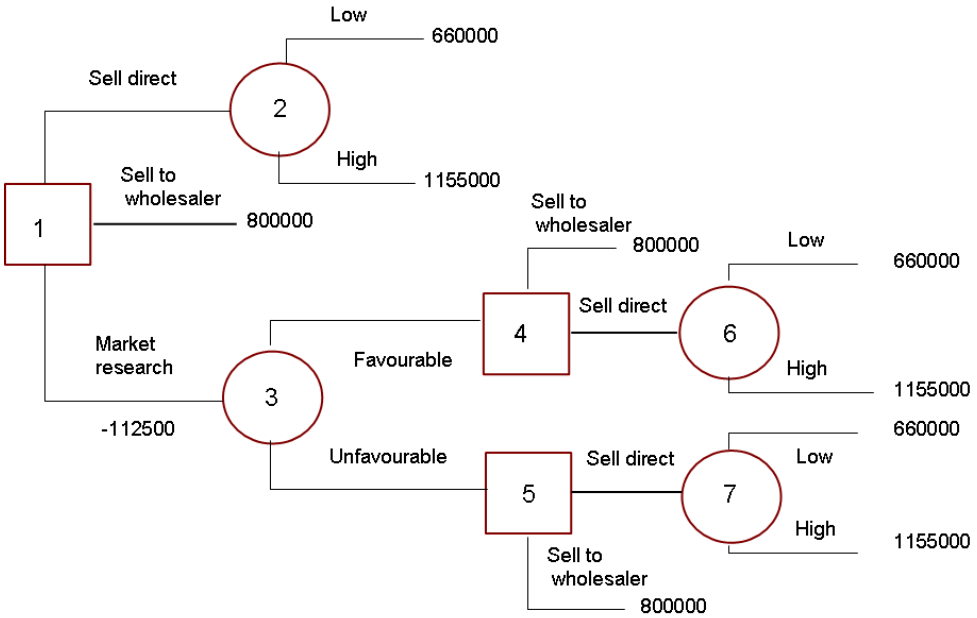
EV(Node 7) = 0.665(2000) + 0.335(20000) = 8030
EV(Node 5) = max(12000, 8030-3000) = 12000
EV(Node 6) = 0.335(2000) + 0.665(20000) = 13970
EV(Node 4) = max(12000, 13970-3000) = 12000
EV(Node 3) = 0.5(12000) + 0.5(12000) = 12000
EV(Node 2) = 0.5(2000) + 0.5(20000) = 11000

At node 1 the EV's are:

EV(Produce) = 11000-3000 = 8000
EV(Sell right) = 12000
EV(Market research) = 12000-1500 = 10500

Therefore Northern Equipment Ltd should sell the right to the product.

29.



Note that the payoffs here correspond with the relevant level of demand X £330, the unit profit.

b. Let L = Low demand (2000 tonnes), H = High demand (3500 tonnes)
F= favourable market research outcome (high demand)
U = unfavourable market research outcome (low demand)

Then we are given P(L) = P(H) = 0.5

$$P(F|H) = 0.7 = P(U|L)$$

Thus P(U|H) = P(F|L) = 0.3

Consequently P(F) = P(F|H)P(H) + P(F|L)P(L) = 0.7(0.5) + 0.3(0.5) = 0.5

Therefore P(U) = 0.5.

It follows the posterior probabilities are:

P(H|F) = 0.7(0.5)/0.5 = 0.7
P(L|F) = 0.3(0.5)/0.5 = 0.3

Similarly:

P(H|U) = 0.3(0.5)/0.5 = 0.3
P(L|U) = 0.7(0.5)/0.5 = 0.7

Using these last results for the backward pass calculations of expected values from nodes 7 to 2 provides the following results.

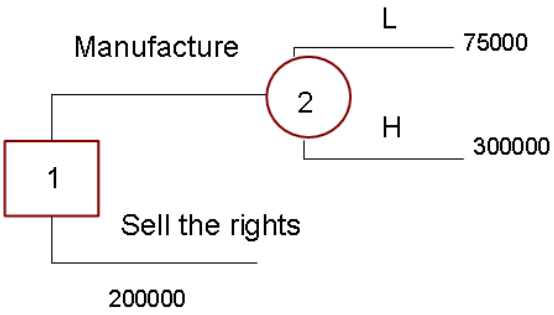
EV(Node 7) = 0.7(660000) + 0.3(1155000) = 808500
EV(Node 5) = max(808500, 800000) = 808500
EV(Node 6) = 0.3(660000) + 0.7(1155000) = 1006500
EV(Node 4) = max(1006500, 800000) = 1006500
EV(Node 3) = 0.5(808500) + 0.5(1006500) = 907500
EV(Node 2) = 0.5(660000) + 0.5(1155000) = 907500

At node the EV's are:

EV(Sell direct) = 907500
EV(Sell to wholesaler) = 800000
EV(Market research) = 907500-112500 = 795000

So Shetland Oil Refinery should sell direct i.e. through its own retail outlets..

30. a.

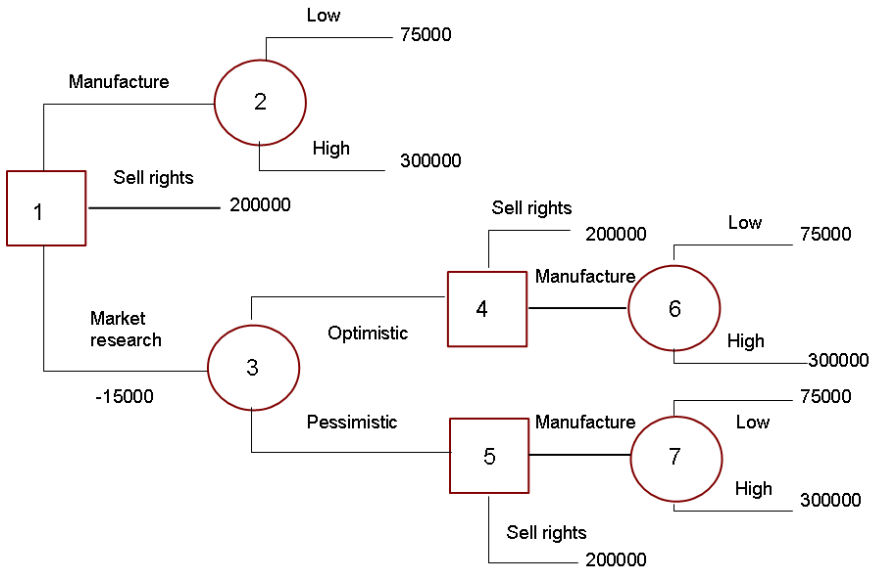


b. EV(Node 2) = 0.45(75000) + 0.55(300000) = 198750

EV(node 1) = max (198750, 200000) = 200000

So ToyKing should sell the rights to the Robocar product.

c.



Let L = Low demand, H = High demand
O= optimistic market research outcome (high demand)
P = pessimistic market research outcome (low demand)

Then we are given P(L) = 0.45
P(H) = 0.55

P(O|H) = 0.9
P(P|L) = 0.4

Thus P(P|H) = 0.1
P(O|L) = 0.6

Consequently P(O) = P(O|H)P(H) + P(O|L)P(L) = 0.9(0.55) + 0.6(0.45) = 0.765

Therefore P(P) = 0.235.

It follows the posterior probabilities are:

P(H|O) = 0.9(0.55)/0.765 = 0.647
P(L|O) = 0.6(0.45)/0.765 = 0.353

Similarly:

P(H|P) = 0.1(0.55)/0.235 = 0.234
P(L|P) = 0.4(0.45)/0.235 = 0.766

Using these last results for the backward pass calculations of expected values from nodes 7 to 2 provides the following results.

EV(Node 7) = 0.766(75000) + 0.234(300000) = 127650
EV(Node 5) = max(127650, 200000) = 200000
EV(Node 6) = 0.353(75000) + 0.647(300000) = 220575
EV(Node 4) = max(220575, 200000) = 220575
EV(Node 3) = 0.765(220575) + 0.235(200000) = 215740

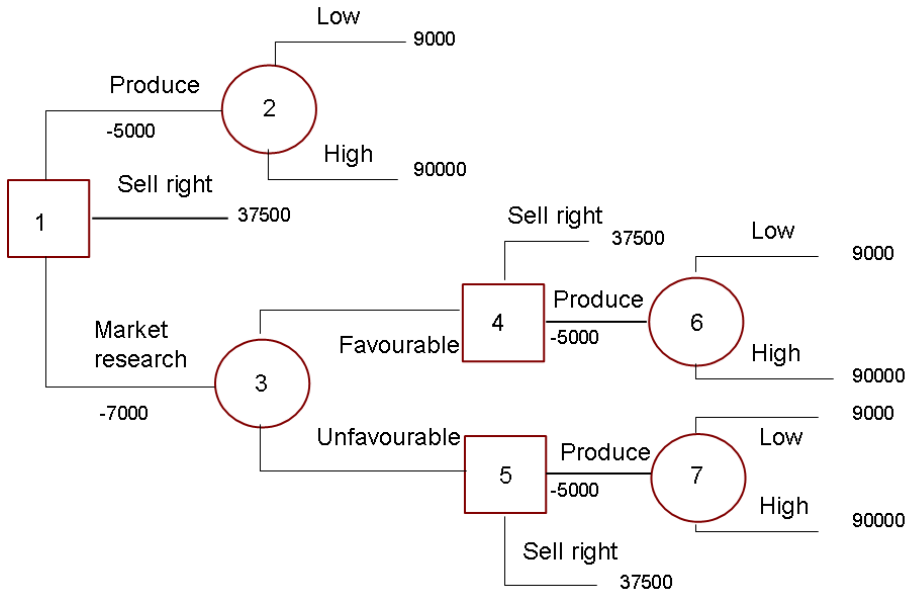
EV(Node 2) = 0.45(75000) + 0.55(300000) = 198750

At node 1 the EV's are:

EV(Manufacture) = 198750
EV(Sell right) = 200000
EV(Market research) = 215740 - 15000= 200740

Therefore ToyKing Ltd should carry out the market research; if the outcome is optimistic it should manufacture the product otherwise if pessimistic it should sell the rights to the product.

31.



Note that the payoffs here correspond with the relevant level of demand X €900, the difference between the selling price and the variable cost per unit

b. Let L = Low demand, H = High demand
F= favourable market research outcome (high demand)
U = unfavourable market research outcome (low demand)

Then we are given P(L) = P(H) = 0.5

$$P(F|H) = 0.9 = P(U|L)$$

Thus P(U|H) = P(F|L) = 0.1
Consequently P(F) = P(F|H)P(H) + P(F|L)P(L) = 0.9(0.5) + 0.1(0.5) = 0.5

Therefore P(U) = 0.5.

It follows the posterior probabilities are:

P(H|F) = 0.9(0.5)/0.5 = 0.9
P(L|F) = 0.1(0.5)/0.5 = 0.1

Similarly:

P(H|U) = 0.1(0.5)/0.5 = 0.1
P(L|U) = 0.9(0.5)/0.5 = 0.9

Using these last results for the backward pass calculations of expected values from nodes 7 to 2 provides the following results.

EV(Node 7) = 0.9(9000) + 0.1(90000) = 17100
EV(Node 5) = max(37500, 17100-5000) = 37500
EV(Node 6) = 0.1(9000) + 0.9(90000) = 13970
EV(Node 4) = max(37500, 81900-5000) = 76900
EV(Node 3) = 0.5(37500) + 0.5(76900) = 57200
EV(Node 2) = 0.5(9000) + 0.5(90000) = 49500

At node 1 the EV's are:

EV(Produce) = 49500-5000 = 44500
EV(Sell right) = 37500
EV(Market research) = 76900-7000 = 69900

Therefore Pine Furniture Company should commission a market research survey. If the results from this are favourable it should produce the new furniture itself. Otherwise it should sell the right to the new furniture design. The maximum expected profit applying this strategy would be €69900.

Statistics for Business and Economics 3e

Anderson, Sweeney, Williams, Freeman, Shoesmith

Chapter Twenty-Two

Sample Surveys

Textbook Exercises (1-17)
Textbook Exercise Solutions (1-17)
Supplementary Exercises (18-31)
Supplementary Exercise Solutions

Chapter 22: Sample Surveys

Textbook Exercises:

1 Simple random sampling was used to obtain a sample of $n = 50$ elements from a population of $N = 800$. The sample mean was $\bar{x} = 215$, and the sample standard deviation was found to be $s = 20$.

a) Estimate the population mean.
b) Estimate the standard error of the mean.
c) Construct an approximate 95 per cent confidence interval for the population mean.

2 Simple random sampling was used to obtain a sample of $n = 80$ elements from a population of $N = 400$. The sample mean was $\bar{x} = 75$, and the sample standard deviation was found to be $s = 8$.

a) Estimate the population total.
b) Estimate the standard error of the population total.
c) Construct an approximate 95 per cent confidence interval for the population total.

3 Simple random sampling was used to obtain a sample of $n = 100$ elements from a population of $N = 1000$. The sample proportion was $p = 0.30$.

a) Estimate the population proportion.
b) Estimate the standard error of the proportion.
c) Construct an approximate 95 per cent confidence interval for the population proportion.

4 A sample is to be taken to develop an approximate 95 per cent confidence interval estimate of the population mean. The population consists of 450 elements, and a pilot study resulted in $s = 70$. How large must the sample be if we want to construct an approximate 95 per cent confidence interval with a width of 30?

5 There are 376 district and unitary local authorities in England and Wales. Suppose you take a simple random sample of 50 of them and find that the mean number of people living in these localities is 135 210, with a sample standard deviation of 93 030. You find that 29 of the sampled local authorities have populations of more than 100 000.

a) Construct an approximate 95 per cent confidence interval for the mean number of residents in all 376 localities.
b) Construct an approximate 95 per cent confidence interval for the total population of all 376 localities.
c) Construct an approximate 95 per cent confidence interval for the proportion of all local authorities that have populations of more than 100 000.

6 *The Wall Street Journal* conducted a survey of subscribers to its interactive edition. One question asked the 504 respondents whether they used a laptop computer when travelling; 55 per cent said they did. Another question asked respondents whether they used an express or package service when travelling; 31 per cent said they did.

a) Calculate an estimate of the standard error for the proportion that uses a laptop computer.
b) Calculate an estimate of the standard error for the proportion that uses an express or package service.
c) Are the estimates of the standard error the same in parts (a) and (b)? If they differ, explain why.
d) Construct an approximate 95 per cent confidence interval for the proportion that uses a laptop computer.
e) Construct an approximate 95 per cent confidence interval for the proportion that uses an express or package service.

7 A quality of life survey was conducted with employees of a manufacturing firm. Of the firm's 3000 employees, a sample of 300 was sent questionnaires. Two hundred usable questionnaires were obtained, giving a response rate of 67 per cent.

a) The mean annual salary for the sample was £23 200 with s = £3000. Construct an approximate 95 per cent confidence interval for the mean annual salary of the population.
b) Use the information in part (a) to construct an approximate 95 per cent confidence interval for the total salary of all 3000 employees.
c) There were 73 per cent of the respondents who reported that they were 'generally satisfied' with their job. Construct an approximate 95 per cent confidence interval for the population proportion.
d) Comment on whether you think the results in part (c) might be biased. Would your opinion change if you knew the respondents were guaranteed anonymity?

8 A stratified random sample was taken with the following results.

Stratum (h)	$\bar{x}_h$	s_h	p_h	N_h	n_h
1	138	30	0.50	200	20
2	103	25	0.78	250	30
3	210	50	0.21	100	25

a) Construct an estimate of the population mean for each stratum.
b) Construct an approximate 95 per cent confidence interval for the population mean in each stratum.
c) Construct an approximate 95 per cent confidence interval for the overall population mean.

9 Reconsider the sample results in exercise 8.

a) Construct an estimate of the population total for each stratum.
b) Construct a point estimate of the total for all 550 elements in the population.
c) Construct an approximate 95 per cent confidence interval for the population total.

10 Reconsider the sample results in exercise 8.

a) Construct an approximate 95 per cent confidence interval for the proportion in each stratum.
b) Construct a point estimate of the population proportion for the 550 elements in the population.
c) Estimate the standard error the point estimator of the population proportion.
d) Construct an approximate 95 per cent confidence interval for the population proportion.

11 A population was divided into three strata with $N_1 = 300$, $N_2 = 600$, and $N_3 = 500$. From a past survey, the following estimates for the standard deviations in the three strata are available: $s_1 = 150$, $s_2 = 75$, $s_3 = 100$.

a) Suppose an estimate of the population mean with a bound on the error of estimate of $B = 20$ is required. How large must the sample be? How many elements should be allocated to each stratum?
b) Suppose a bound of $B = 10$ is desired. How large must the sample be? How many elements should be allocated to each stratum?
c) Suppose an estimate of the population total with a bound of $B = 15\,000$ is requested. How large must the sample be? How many elements should be allocated to each stratum?

12 An accounting firm works for a number of clients in the banking, insurance, and share-dealing sectors: $N_1 = 50$ banks, $N_2 = 38$ insurance companies, and $N_3 = 35$ share-dealing firms. A marketing research firm has been hired to survey the accounting firm's clients in these three sectors. The survey will ask a variety of questions about both the clients' businesses and their satisfaction with services provided by the accounting firm. Suppose an approximate 95 per cent confidence interval is requested for the mean number of employees for the 123 clients, with a bound on the error of estimation of $B = 30$.

a) Suppose a pilot study finds $s_1 = 80$, $s_2 = 150$, and $s_3 = 45$. Choose a total sample size, and explain how the sample size should be allocated to the three strata.
b) Suppose the pilot study is called into question and a decision is made to assume the
c) stratum standard deviations are all equal to 100 in choosing the sample size. Choose a total sample size, and determine how many elements should be sampled in each stratum.

13 A stratified simple random sample is to be taken of a bank's customers to learn about a variety of attitudinal and demographic issues. The stratification is to be based on savings account balances as of 30 June 2009. A frequency distribution follows showing the number of accounts in each stratum, together with the standard deviation of account balances by stratum.

Stratum (£)	Accounts	Standard deviation of account balances (£)
0.00–1000.00	3000	80
1000.01–2000.00	600	150
2000.01–5000.00	250	220
5000.00–10 000.00	100	700
Over 10 000	50	3000

a) Assuming the cost per unit sampled is approximately equal across strata, determine the total number of persons to include in the sample. Assume we want a bound on the error of estimate of the population mean for savings account balances of B = £20.
b) Use the Neyman allocation procedure to determine the number to be sampled for each stratum.

14 A sample of four clusters is to be taken from a population with $N = 25$ clusters and $M = 300$ elements. The values of M_i, t_i and a_i for each cluster in the sample follow.

Cluster (i)	M_i	t_i	a_i
1	7	95	1
2	18	325	6
3	15	190	6
4	10	140	2
Totals	50	750	15

a) Construct point estimates of the population mean, total, and proportion.
b) Estimate the standard errors for the estimates in part (a).
c) Construct an approximate 95 per cent confidence interval for the population mean.
d) Construct an approximate 95 per cent confidence interval for the population total.
e) Construct an approximate 95 per cent confidence interval for the population proportion.

15 A sample of six clusters is to be taken from a population with $N = 30$ clusters and $M = 600$ elements. The following table shows values of M_i, t_i and a_i for each cluster in the sample.

Cluster (i)	M_i	t_i	a_i
1	35	3 500	3
2	15	965	0
3	12	960	1
4	23	2 070	4
5	20	1 100	3
6	25	1 805	2
Totals	130	10 400	13

a) Construct point estimates of the population mean, total, and proportion.
b) Construct an approximate 95 per cent confidence interval for the population mean.

c) Construct an approximate 95 per cent confidence interval for the population total.
d) Construct an approximate 95 per cent confidence interval for the population proportion.

16 A public utility is conducting a survey of mechanical engineers to learn more about the factors influencing the choice of heating, ventilation, and air conditioning HVAC) equipment for new commercial buildings. A total of 120 firms in the utility's service area are engaged in designing HVAC systems. The sampling plan is to use cluster sampling with each firm representing a cluster. For each firm in the sample, all of the mechanical engineers will be interviewed. Approximately 500 mechanical engineers are believed to be employed by the 120 firms. A sample of ten firms was taken. Among other things, the age of each respondent was recorded as well as whether the respondent had attended the local university.

Cluster (*i*)	M_i	*Total of respondents' ages*	*Number attending local university*
1	12	520	8
2	1	33	0
3	2	70	1
4	1	29	1
5	6	270	3
6	3	129	2
7	2	102	0
8	1	48	1
9	9	337	7
10	13	462	12
Totals	50	2000	35

a) Estimate the mean age of mechanical engineers engaged in this type of work.
b) Estimate the proportion of mechanical engineers in the utility's service area who attended the local university.
c) Construct an approximate 95 per cent confidence interval for the mean age of mechanical engineers designing HVAC systems for commercial buildings.
d) Construct an approximate 95 per cent confidence interval for the proportion of mechanical engineers in the utility's service area who attended the local university.

17 A public agency is interested in learning more about the people living in nursing homes in a particular city. A total of 100 nursing homes are caring for 4800 people in the city and a cluster sample of six homes has been taken. Each person in the six homes has been interviewed. A portion of the sample results follows.

Home	Number of residents	Average age of residents	Number of disabled residents
1	14	61	12
2	7	74	2
3	96	78	30
4	23	69	8
5	71	73	10
6	29	84	22

a) Calculate an estimate of the mean age of nursing home residents in this city.
b) Construct an approximate 95 per cent confidence interval for the proportion of disabled persons in the city's nursing homes.
c) Estimate the total number of disabled persons residing in nursing homes in this city.

Chapter 22: Sample Surveys

Textbook Exercises Solutions:

1. a. $\bar{x}$ = 215 is an estimate of the population mean.

   b. $s_{\bar{x}} = \sqrt{\frac{800-50}{800}}\frac{20}{\sqrt{50}} = 2.7386$

   c. 215 $\pm$ (2 x 2.7386) = 209.5 to 220.5

2. a. Estimate of population total = $N\bar{x}$ = 400(75) = 30,000

   b. Estimate of Standard Error = $Ns_{\bar{x}}$

   $$Ns_{\bar{x}} = 400\sqrt{\frac{400-80}{400}}\left(\frac{8}{\sqrt{80}}\right) = 320$$

   c. 30,000 $\pm$ 2(320) or 29,360 to 30,640

3. a. p = 0.30 is an estimate of the population proportion

   b. $s_p = \sqrt{\frac{1000-100}{1000}}\sqrt{\frac{0.3\times 0.7}{100}} = 0.04347$

   c. 0.30 $\pm$ (2 x 0.04347) = 0.213 to 0.387

4. B = 15

   $$n = \frac{450(70)^2}{450\frac{(15)^2}{4} + (70)^2} = 72.98$$

   A sample size of 73 will provide an approximate 95% confidence interval of width 30.

5. a. $\bar{x}$ = 135,210 and s = 93,030

   $$s_{\bar{x}} = \sqrt{\frac{376-50}{376}}\left(\frac{93,030}{\sqrt{50}}\right) = 12,250.47$$

   Approximate 95% confidence interval: 135,210 $\pm$ 2(12,250.47) or 110,700 to 159,700

   b. $\hat{\tau} = N\bar{x}$ = 376(135,210) = 50,838,960

   $s_{\hat{\tau}} = N s_{\bar{x}}$ = 376(12,250.47) = 4,606,176.7

   Approximate 95% confidence interval: 50,838,960 $\pm$ 2(4,606,176.7) or 41.63 millions to 60.05 millions

c. $p = 29/50 = 0.58$ and $s_P = \sqrt{\left(\frac{376-50}{376}\right)\left(\frac{(0.58)(0.42)}{50}\right)} = 0.0650$

Approximate 95% confidence interval: $0.58 \pm 2(0.0650)$ or 0.450 to 0.710

This is a rather large interval; sample sizes must be large to obtain tight confidence intervals on a population proportion.

6. a. Assume $(N - n) / N \approx 1$

$p = 0.55$

$$s_P = \sqrt{\frac{(0.55)(0.45)}{504}} = 0.0222$$

b. $p = 0.31$

$$s_P = \sqrt{\frac{(0.31)(0.69)}{504}} = 0.0206$$

c. The estimate of the standard error in part (a) is larger because p is closer to 0.50.

d. Approximate 95% confidence interval:

$0.55 \pm 2(0.0222)$ or 0.506 to 0.594

e. Approximate 95% confidence interval:

$0.31 \pm 2(0.0206)$ or 0.269 to 0.351

7. a. $s_{\bar{x}} = \sqrt{\frac{3000-200}{3000}}\frac{3000}{\sqrt{200}} = 204.9390$

Approximate 95% confidence Interval for mean annual salary:

$23{,}200 \pm 2(204.9390)$ or £22,790 to £23,610

b. $\hat{t} = N\bar{x} = 3000(23{,}200) = 69{,}600{,}000$

$s_{\hat{\tau}} = 3000\ (204.9390) = 614{,}817$

Approximate 95% confidence interval for population total salary:

$69{,}600{,}000 \pm 2(614{,}817)$ or £68,370,366 to £70,829,634

c. $p = 0.73$

$$s_P = \sqrt{\left(\frac{3000-200}{3000}\right)\left(\frac{(0.73)(0.27)}{200}\right)} = 0.0303$$

Approximate 95% confidence interval for population proportion that are "generally satisfied":

$0.73 \pm 2(0.0303)$ or 0.669 to 0.791

d. If management administered the questionnaire and anonymity was not guaranteed we would expect a definite upward bias in the percentage reporting they were "generally satisfied" with their job. A procedure for guaranteeing anonymity should reduce the bias.

8. a. Stratum 1: $\bar{x}_1 = 138$

Stratum 2: $\bar{x}_2 = 103$

Stratum 3: $\bar{x}_3 = 210$

b. Stratum 1

$\bar{x}_1 = 138$

$$s_{\bar{x}_1} = \left(\frac{30}{\sqrt{20}}\right)\sqrt{\frac{200-20}{200}} = 6.3640$$

Approximate 95% confidence interval is: 138 $\pm$ 2(6.3640) or 125.3 to 150.7

Stratum 2

$\bar{x}_2 = 103$

$$s_{\bar{x}_2} = \left(\frac{25}{\sqrt{30}}\right)\sqrt{\frac{250-30}{250}} = 4.2817$$

Approximate 95% confidence interval is: 103 $\pm$ 2(4.2817) or 94.4 to 111.6

Stratum 3

$\bar{x}_3 = 210$

$$s_{\bar{x}_3} = \left(\frac{50}{\sqrt{25}}\right)\sqrt{\frac{100-25}{100}} = 8.6603$$

Approximate 95% confidence interval is: 210 $\pm$ 2(8.6603) or 192.7 to 227.3

c. $$\bar{x}_{st} = \left(\frac{200}{550}\right)138 + \left(\frac{250}{550}\right)103 + \left(\frac{100}{550}\right)210$$

$= 50.1818 + 46.8182 + 38.1818$

$= 135.18$

$$s_{\bar{x}_{st}} = \sqrt{\left(\frac{1}{(550)^2}\right)\left(200(180)\frac{(30)^2}{20} + 250(220)\frac{(25)^2}{30} + 100(75)\frac{(50)^2}{25}\right)}$$

$$= \sqrt{\left(\frac{1}{(550)^2}\right)3{,}515{,}833.3} = 3.4092$$

Approximate 95% confidence interval is: 135.1818 $\pm$ 2(3.4092) or 128.4 to 142.0

9. a. Stratum 1: $N_1\bar{x}_1 = 200(138) = 27{,}600$

Stratum 2: $N_2\bar{x}_2 = 250(103) = 25{,}750$

Stratum 3: $N_3\bar{x}_3 = 100(210) = 21{,}000$

b. $N\bar{x}_{st} = 27{,}600 + 25{,}750 + 21{,}000 = 74{,}350$

Note: the sum of the estimates for each stratum total equals $N\bar{x}_{st}$

c. $s_{N\bar{x}_{st}} = 550(3.4092) = 1875.06$ (see 8c)

Approximate 95% confidence interval is: 74,350 $\pm$ 2(1875.06) or 70,600 to 78,100

10. a. Stratum 1

$p_1 = 0.50$

$$s_{p_1} = \sqrt{\left(\frac{200-20}{200}\right)\left(\frac{(0.50)(0.50)}{20}\right)} = 0.1061$$

Approximate 95% confidence interval is: 0.50 $\pm$ 2(0.1061) or 0.288 to 0.712

Stratum 2

$p_2 = 0.78$

$$s_{p_2} = \sqrt{\left(\frac{250-30}{250}\right)\left(\frac{(0.78)(0.22)}{30}\right)} = 0.0709$$

Approximate 95% confidence interval is: 0.78 $\pm$ 2(0.0709) or 0.638 to 0.922

Stratum 3

$p_3 = 0.21$

$$s_{p_3} = \sqrt{\left(\frac{100-25}{100}\right)\left(\frac{(0.21)(0.79)}{25}\right)} = 0.0705$$

Approximate 95% confidence interval is: 0.21 $\pm$ 2(0.0705) or 0.069 to 0.351

b. $$p_{st} = \frac{200}{550}(0.50) + \frac{250}{550}(0.78) + \frac{100}{550}(0.21) = 0.5745$$

c. $$s_{p_{st}} = \sqrt{\left(\frac{1}{(550)^2}\right)\left(200(180)\frac{(0.5)(0.5)}{20} + 250(220)\frac{(0.78)(0.22)}{30} + 100(75)\frac{(0.21)(0.79)}{25}\right)}$$

$$= \sqrt{\left(\frac{1}{(550)^2}\right)(450 + 314.6 + 49.77)} = 0.0519$$

d. Approximate 95% confidence interval is: 0.5745 $\pm$ 2(0.0519) or 0.471 to 0.678

11. a.

$$n = \frac{\left[300(150) + 600(75) + 500(100)\right]^2}{(1400)^2\left(\frac{(20)^2}{2}\right) + \left[300(150)^2 + 600(75)^2 + 500(100)^2\right]} = \frac{(140,000)^2}{196,000,000 + 15,125,000} = 92.8359$$

Rounding up we choose a total sample of 93.

$$n_1 = 93\left(\frac{300(150)}{140,000}\right) = 30$$

$$n_2 = 93\left(\frac{600(75)}{140,000}\right) = 30$$

$$n_3 = 93\left(\frac{500(100)}{140,000}\right) = 33$$

b. With $B = 10$, the first term in the denominator in the formula for n changes.

$$n = \frac{(140,000)^2}{(1400)^2\left(\frac{(10)^2}{4}\right) + 15,125,000} = \frac{(140,000)^2}{49,000,000 + 15,125,000} = 305.6530$$

Rounding up, we see that a sample size of 306 is needed to provide this level of precision.

$$n_1 = 306\left(\frac{300(150)}{140,000}\right) = 98$$

$$n_2 = 306\left(\frac{600(75)}{140,000}\right) = 98$$

$$n_3 = 306\left(\frac{500(100)}{140,000}\right) = 109$$

Due to rounding, the total of the allocations to each strata only add to 305. Note that even though the sample size is larger, the proportion allocated to each stratum has not changed.

c. $$n = \frac{(140,000)^2}{\frac{(15,000)^2}{4} + 15,125,000} = \frac{(140,000)^2}{56,250,000 + 15,125,000} = 274.6060$$

Rounding up, we see that a sample size of 275 will provide the desired level of precision.

The allocations to the strata are in the same proportion as for parts a and b.

$$n_1 = 275\left(\frac{300(150)}{140,000}\right) = 98$$

$$n_2 = 275\left(\frac{600(75)}{140,000}\right) = 88$$

$$n_3 = 275\left(\frac{500(100)}{140,000}\right) = 98$$

Again, due to rounding, the stratum allocations do not add to the total sample size. Another item could be sampled from, say, stratum 3 if desired.

12. $$n = \frac{\left[50(80)+38(150)+35(45)\right]^2}{(123)^2\left(\frac{(30)^2}{4}\right)+\left[50(80)^2+38(150)^2+35(45)^2\right]} = \frac{(11,275)^2}{3,404,025+1,245,875} = 27.3394$$

Rounding up we see that a sample size of 28 is necessary to obtain the desired precision.

$$n_1 = 28\left(\frac{50(80)}{11,275}\right) = 10$$

$$n_2 = 28\left(\frac{38(150)}{11,275}\right) = 14$$

$$n_3 = 28\left(\frac{35(45)}{11,275}\right) = 4$$

b. $$n = \frac{\left[50(100)+38(100)+35(100)\right]^2}{(123)^2\left(\frac{(30)^2}{4}\right)+\left[50(100)^2+38(100)^2+35(100)^2\right]} = \frac{\left[123(100)\right]^2}{3,404,025+123(100)^2} = 33$$

$$n_1 = 33\left(\frac{50(100)}{12,300}\right) = 13$$

$$n_2 = 33\left(\frac{38(100)}{12,300}\right) = 10$$

$$n_3 = 33\left(\frac{35(100)}{12,300}\right) = 9$$

This is the same as proportional allocation. Note that for each stratum

$$n_h = n\left(\frac{N_h}{N}\right)$$

13. a. $$n = \frac{\left[3000(80)+600(150)+250(220)+100(700)+50(3000)\right]^2}{(4000)^2\left(\frac{(20)^2}{4}\right)+3000(80)^2+600(150)^2+250(220)^2+100(700)^2+50(3000)^2}$$

$$= \frac{366,025,000,000}{1,600,000,000+543,800,000} = 170.7365$$

Rounding up, we need a sample size of 171 for the desired precision.

b. $$n_1 = 171\left(\frac{3000(80)}{605,000}\right) = 68$$

$$n_2 = 171\left(\frac{600(150)}{605,000}\right) = 25$$

$$n_3 = 171\left(\frac{250(220)}{605,000}\right) = 16$$

$$n_4 = 171\left(\frac{100(700)}{605,000}\right) = 20$$

$$n_5 = 171\left(\frac{50(3000)}{605,000}\right) = 42$$

14. a. $\bar{x}_c = \frac{\sum t_i}{\sum M_i} = \frac{750}{50} = 15$

$\hat{\tau} = M\,\bar{x}_c = 300(15) = 4500$

$p_c = \frac{\sum a_i}{\sum M_i} = \frac{15}{50} = 0.30$

b. $\sum(t_i - \bar{x}_c M_i)^2 = [95 - 15(7)]^2 + [325 - 15(18)]^2 + [190 - 15(15)]^2 + [140 - 15(10)]^2$
$= (-10)^2 + (55)^2 + (-35)^2 + (-10)^2$
$= 4450$

$$s_{\bar{x}_c} = \sqrt{\left(\frac{25-4}{(25)(4)(12)^2}\right)\left(\frac{4450}{3}\right)} = 1.4708$$

$s_{\hat{\tau}} = M s_{\bar{x}_c} = 300(1.4708) = 441.24$

$\sum(a_i - p_c M_i)^2 = [1 - 0.3(7)]^2 + [6 - 0.3(18)]^2 + [6 - 0.3(15)]^2 + [2 - 0.3(10)]^2$
$= (-1.1)^2 + (0.6)^2 + (1.5)^2 + (-1)^2$
$= 4.82$

$$s_{p_c} = \sqrt{\left(\frac{25-4}{(25)(4)(12)^2}\right)\left(\frac{4.82}{3}\right)} = 0.0484$$

c. Approximate 95% confidence interval for population mean: $15 \pm 2(1.4708)$ or 12.1 to 17.9

d. Approximate 95% confidence interval for population total: $4500 \pm 2(441.24)$ or 3618 to 5382

e. Approximate 95% confidence interval for population proportion: $0.30 \pm 2(0.0484)$ or 0.203 to 0.397

15. a. $\bar{x}_c = \dfrac{10,400}{130} = 80$

$\hat{\tau} = M\bar{x}_c = 600(80) = 48,000$

$p_c = \dfrac{13}{130} = 0.10$

b. $\sum(t_i - \bar{x}_c M_i)^2 = [3500 - 80(35)]^2 + [965 - 80(15)]^2 + [960 - 80(12)]^2 + [2070 - 80(23)]^2 + [1100 - 80(20)]^2 + [1805 - 80(25)]^2$
$= (700)^2 + (-235)^2 + (0)^2 + (230)^2 + (-500)^2 + (-195)^2$
$= 886,150$

$s_{\bar{x}_c} = \sqrt{\left(\dfrac{30-6}{(30)(6)(20)^2}\right)\left(\dfrac{886,150}{5}\right)} = 7.6861$

Approximate 95% confidence interval for population mean: 80 $\pm$ 2(7.6861) or 64.6 to 95.4

c. $s_{\hat{\tau}} = 600(7.6861) = 4611.66$

Approximate 95% confidence interval for population total: 48,000 $\pm$ 2(4611.66) or 38,777 to 57,223

d. $\sum(a_i - p_c M_i)^2 = [3 - 0.1(35)]^2 + [0 - 0.1(15)]^2 + [1 - 0.1(12)]^2 + [4 - 0.1(23)]^2 + [3 - 0.1(20)]^2 + [2 - 0.1(25)]^2$
$= (-0.5)^2 + (-1.5)^2 + (-0.2)^2 + (1.7)^2 + (1)^2 + (-0.5)^2$
$= 6.68$

$s_{P_c} = \sqrt{\left(\dfrac{30-6}{(30)(6)(20)^2}\right)\left(\dfrac{6.68}{5}\right)} = 0.0211$

Approximate 95% confidence interval for population proportion: 0.10 $\pm$ 2(0.0211) or 0.058 to 0.142

16. a. $\bar{x}_c = \dfrac{2000}{50} = 40$

Estimate of mean age of mechanical engineers: 40 years

b. $p_c = \dfrac{35}{50} = 0.70$

Estimate of proportion attending local university: 0.70

c. $\sum(t_i - \bar{x}_c M_i)^2 = [520 - 40(12)]^2 + \cdots + [462 - 40(13)]^2$
$= (40)^2 + (-7)^2 + (-10)^2 + (-11)^2 + (30)^2 + (9)^2 + (22)^2 + (8)^2 + (-23)^2 + (-58)^2$
$= 7292$

$s_{\bar{x}_c} = \sqrt{\left(\dfrac{120-10}{(120)(10)(50/12)^2}\right)\left(\dfrac{7292}{9}\right)} = 2.0683$

Approximate 95% confidence interval for mean age: 40 $\pm$ 2(2.0683) or 35.9 to 44.1

d. $\sum(a_i - p_c M_i)^2 = [8 - 0.7(12)]^2 + \cdots + [12 - 0.7(13)]^2$

$= (-0.4)^2 + (-0.7)^2 + (-0.4)^2 + (0.3)^2 + (-1.2)^2 + (-0.1)^2 + (-1.4)^2 + (0.3)^2$

$+ (0.7)^2 + (2.9)^2$

$= 13.3$

$$s_{P_c} = \sqrt{\left(\frac{120-10}{(120)(10)(50/12)^2}\right)\left(\frac{13.3}{9}\right)} = 0.0883$$

Approximate 95% confidence interval for proportion attending local university: $0.70 \pm 2(0.0883)$ or 0.523 to 0.877

17. a. $$\bar{x}_c = \frac{14(61)+7(74)+96(78)+23(69)+71(73)+29(84)}{14+7+96+23+71+29} = \frac{18,066}{240} = 75.275$$

Estimate of mean age is approximately 75 years old.

b. $$p_c = \frac{12+2+30+8+10+22}{14+7+96+23+71+29} = \frac{84}{240} = 0.35$$

$\sum(a_i - p_c M_i)^2 = [12 - 0.35(14)]^2 + [2 - 0.35(7)]^2 + [30 - 0.35(96)]^2$

$+ [8 - 0.35(23)]^2 + [10 - 0.35(71)]^2 + [22 - 0.35(29)]^2$

$= (7.1)^2 + (-0.45)^2 + (-3.6)^2 + (-0.05)^2 + (-14.85)^2 + (11.85)^2$

$= 424.52$

$$s_{P_c} = \sqrt{\left(\frac{100-6}{(100)(6)(48)^2}\right)\left(\frac{424.52}{5}\right)} = 0.0760$$

Approximate 95% confidence interval: $0.35 \pm 2(0.0760)$ or 0.198 to 0.502

c. $\hat{\tau} = 4800(0.35) = 1680$

Estimate of total number of disabled persons is 1680.

Supplementary Exercises:

18. Suppose a sample was taken of 50 small businesses that have received government loans. The sample was selected from 771 businesses that received government loans. The sample mean loan is €149,670 with a standard deviation of €73,420.
    a. Construct an approximate 95% confidence interval for the population mean loan value.
    b. Construct an approximate 95% confidence interval for the total value of all 771 loans.

19. Refer to exercise 18. Suppose that 18 of the businesses in the sample were manufacturing companies. Construct an approximate 95% confidence interval for the proportion of all loans given to manufacturing companies.

20. Suppose that in a particular tax district, 724 corporate tax returns were filed. The mean annual income reported was €161,220 with a standard deviation of €31,300. How large a sample will be necessary next year to construct an approximate 95% confidence interval for mean annual corporate income, if the precision required is an interval width of no more than €5000.

21. To assess consumer acceptance of a new series of ads for Miller Lite Beer, Louis Harris conducted a nationwide poll of 363 adults who had seen the Miller Lite ads (*USA Today*, November 17, 1997). The following responses are based on that survey. (*Note:* Because the survey sampled only a small fraction of all adults, assume $(N - n)/N = 1$ in any formulae involving the standard error.)
    a. Nineteen percent of all respondents indicated they liked the ads a lot. Construct a 95% confidence interval for the population proportion of adults who like the ads a lot.
    b. Thirty-one percent of the respondents disliked the new ads. Construct a 95% confidence interval for the population proportion of adults who dislike the ads.
    c. Seventeen percent of the respondents felt the ads are very effective. Construct a 95% confidence interval for the population proportion of adults who think the ads are very effective.

22. Refer to exercise 21.
    a. Louis Harris reported that the "margin of error is five percentage points." What does this statement mean and how do you think they arrived at this number?
    b. How might non-sampling error bias the results of such a survey?

23. A pharmacy chain has stores in four cities: 38 stores in city A, 45 in city B, 80 in city C, and 70 in city D. Pharmacy sales in the four cities vary considerably because of the competition. The following sales data (in thousands of euros) are available from a sample survey. Each of the cities was considered a separate stratum, and a stratified random sample was taken.

City A	City B	City C	City D
50.3	48.7	16.7	14.7
41.2	59.8	38.4	88.3
15.7	28.9	51.6	94.2
22.5	36.5	42.7	76.8
26.7	89.8	45.0	35.1
20.8	96.0	59.7	48.2
	77.2	80.0	57.9
	81.3	27.6	18.8
			22.0
			74.3

    a. Estimate the mean pharmacy sales per store for each city (stratum).
    b. Construct an approximate 95% confidence interval for the mean pharmacy sales per store in each city.

24. Reconsider the sample survey results in exercise 23.
    a. Estimate the proportion of stores with sales of €50,000 or more.
    b. Construct an approximate 95% confidence interval for the proportion of stores with sales of €50,000 or more.

25. Reconsider the sample survey results in exercise 23.
    a. Estimate the population total pharmacy sales for city C.
    b. Estimate the population total pharmacy sales for city A.

26. Reconsider the sample survey results in exercise 23.
    a. For the chain of stores, construct an approximate 95% confidence interval for mean pharmacy sales per store.
    b. For the chain of stores, construct an approximate 95% confidence interval for population total pharmacy sales.

27. This exercise concerns small business start-ups in three large cities. In city A, city B, and city C, the number of start-ups was, respectively, 380, 760, and 260. Suppose a stratified random sample with the following results was taken to learn more about the character and success of the start-ups.

Stratum	Sample size	Unsuccessful*	No government financial support	Service sector
City A	30	10	9	21
City B	45	19	12	34
City C	25	7	11	15

* Judged after one year of operation.

    a. Construct an approximate 95% confidence interval for the proportion of unsuccessful start-ups in city A.
    b. Construct an estimate for the total number of successful start-ups in city B.
    c. Construct an approximate 95% confidence interval for the proportion of unsuccessful start-ups in city B.

28. Refer again to the data in exercise 27. Construct an approximate 95% confidence interval for the proportion of unsuccessful start-ups across all three cities.

29. Refer again to the data in exercise 27.
    a. Compute an estimate of the total number of start-ups across all three cities that received no government financial support.
    b. Construct an approximate 95% confidence interval for the proportion of start-ups across all three cities that received no government financial support.

30. Refer again to the data in exercise 27.
    a. Construct an approximate 95% confidence interval for the proportion of start-ups across all three cities in the service sector.
    b. Construct an estimate of the total number of start-ups across all three cities in the service sector.

31. A national estate agency has just acquired a smaller firm with 150 offices and 6000 agents. The national firm conducted a sample survey to learn about attitudes and other characteristics of its new employees. In a sample of eight offices, all of the agents completed the questionnaire. Results of the survey for the eight offices follow.

Office	Agents	Mean age	University graduates	Male agents
1	17	37	3	4
2	35	32	14	12
3	26	36	8	7
4	66	30	38	28
5	43	41	18	12
6	12	52	2	6
7	48	35	20	17
8	57	44	25	26

    a. Estimate the mean age of the agents.
    b. Construct an approximate 95% confidence interval for the mean age of the agents.

32. Reconsider the sample survey results in exercise 31.
    a. Estimate the proportion of agents who are university graduates and the proportion who are male.
    b. Construct an approximate 95% confidence interval for the proportion of agents who are university graduates.
    c. Construct an approximate 95% confidence interval for the proportion of agents who are male.

Supplementary Exercises Solutions:

18. a. $\bar{x} = 149{,}670$ and $s = 73{,}420$

$$s_{\bar{x}} = \sqrt{\frac{771-50}{771}}\left(\frac{73{,}420}{\sqrt{50}}\right) = 10{,}040.83$$

Approximate 95% confidence interval: 149,670 $\pm$ 2(10,040.83) or €129,588 to €169,752

b. $\hat{\tau} = N\bar{x} = 771(149{,}670) = 115{,}395{,}570$

$s_{\hat{\tau}} = N s_{\bar{x}} = 771(10{,}040.83) = 7{,}741{,}479.93$

Approximate 95% confidence interval: 115,395,570 $\pm$ 2(7,741,479.93)
or €99,912,610 to €130,878,530

19. $p =$ 18/50 = 0.36 and $s_p = \sqrt{\left(\frac{771-50}{771}\right)\left(\frac{(0.36)(0.64)}{50}\right)} = 0.0656$

Approximate 95% confidence interval: 0.36 $\pm$ 2(0.0656) or 0.229 to 0.491

This is a rather large interval; sample sizes must be large to obtain tight confidence intervals on a population proportion.

20. B = 5000/2 = 2500.

Use the value of s for the previous year in the formula to determine the necessary sample size.

$$n = \frac{(31.3)^2}{\frac{(2.5)^2}{4} + \frac{(31.3)^2}{724}} = \frac{979.69}{2.9157} = 336.0051$$

A sample size of 337 will provide an approximate 95% confidence interval of width no larger than €5000.

21. a. $p =$ 0.19

$$s_p = \sqrt{\frac{(0.19)(0.81)}{363}} = 0.0206$$

Approximate 95% confidence interval: 0.19 $\pm$ 2(0.0206) or 0.1488 to 0.2312

b. $p =$ 0.31

$$s_P = \sqrt{\frac{(0.31)(0.69)}{363}} = 0.0243$$

Approximate 95% confidence interval: 0.31 $\pm$ 2(0.0243) or 0.2615 to 0.3585

c. $p = 0.17$

$$s_P = \sqrt{\frac{(0.17)(0.83)}{373}} = 0.0197$$

Approximate 95% confidence interval: 0.17 $\pm$ 2(0.0197) or 0.1306 to 0.2094

22. a. The largest standard error is when $\pi = 0.50$.

At $\pi = 0.50$, we get

$$s_P = \sqrt{\frac{(0.5)(0.5)}{363}} = 0.0262$$

Multiplying by 2, we get a bound of $B = 2(0.0262) = 0.0525$

For a sample of 363, then, they know that in the worst case ($\pi = 0.50$), the bound will be approximately 5%.

b. If the poll was conducted by calling people at home during the day the sample results would only be representative of adults not working outside the home. It is likely that the Louis Harris organization took precautions against this and other possible sources of bias.

23. a. $\bar{x}_1 = 29.5333$ $\bar{x}_2 = 64.775$

$\bar{x}_3 = 45.2125$ $\bar{x}_4 = 53.0300$

b. City A

$$29.533 \pm 2\left(\frac{13.3603}{\sqrt{6}}\right)\sqrt{\frac{38-6}{38}}$$

= 29.533 $\pm$ 10.9086(0.9177) or 19.52 to 39.54

City B

$$64.775 \pm 2\left(\frac{25.0666}{\sqrt{8}}\right)\sqrt{\frac{45-8}{45}}$$

= 64.775 $\pm$ 17.7248(0.9068) or 48.70 to 80.85

City C

$$45.2125 \pm 2\left(\frac{19.4084}{\sqrt{8}}\right)\sqrt{\frac{80-8}{80}}$$

45.2125 $\pm$ (13.7238) (0.9487) or 32.19 to 58.23

City D

$$53.0300 \pm 2\left(\frac{29.6810}{\sqrt{10}}\right)\sqrt{\frac{70-10}{70}}$$

= 53.03 $\pm$ 18.7719(0.9258) or 35.65 to 70.41

24. a. $$p_{st} = \left(\frac{38}{233}\right)\left(\frac{1}{6}\right)+\left(\frac{45}{233}\right)\left(\frac{5}{8}\right)+\left(\frac{80}{233}\right)\left(\frac{3}{8}\right)+\left(\frac{70}{233}\right)\left(\frac{5}{10}\right) = 0.4269$$

b. $$N_1(N_1-n_1)\frac{p_1(1-p_1)}{n_1} = 38(32)\frac{\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)}{6} = 28.15$$

$$N_2(N_2-n_2)\frac{p_2(1-p_2)}{n_2} = 45(37)\frac{\left(\frac{5}{8}\right)\left(\frac{3}{8}\right)}{8} = 48.78$$

$$N_3(N_3-n_3)\frac{p_3(1-p_3)}{n_3} = 80(72)\frac{\left(\frac{3}{8}\right)\left(\frac{5}{8}\right)}{8} = 168.75$$

$$N_4(N_4-n_4)\frac{p_4(1-p_4)}{n_4} = 70(60)\frac{\left(\frac{5}{10}\right)\left(\frac{5}{10}\right)}{10} = 105$$

$$s_{P_{st}} = \sqrt{\left(\frac{1}{(233)^2}\right)[28.15+48.78+168.75+105]} = \sqrt{\frac{1}{(233)^2}(350.68)} = 0.0804$$

Approximate 95% confidence interval: 0.4269 $\pm$ 2(0.0804) or 0.266 to 0.589

25. a. City C total = $N_3\bar{x}_3$ = 80(45.2125) = 3617

In euros: €3,617,000

b. City A total = $N_1\bar{x}_1$ = 38(29.5333) = 1122.27

In euros: €1,122,270

26. a. $\bar{x}_{st} = \left(\frac{38}{233}\right)29.5333 + \left(\frac{45}{233}\right)64.775 + \left(\frac{80}{233}\right)45.2125 + \left(\frac{70}{233}\right)53.0300 = 48.7821$

$$N_1(N_1 - n_1)\frac{s_1^2}{n_1} = 38(32)\frac{(13.3603)^2}{6} = 36{,}175.517$$

$$N_2(N_2 - n_2)\frac{s_2^2}{n_2} = 45(37)\frac{(25.0666)^2}{8} = 130{,}772.1$$

$$N_3(N_3 - n_3)\frac{s_3^2}{n_3} = 80(72)\frac{(19.4084)^2}{8} = 271{,}213.91$$

$$N_4(N_4 - n_4)\frac{s_4^2}{n_4} = 70(60)\frac{(29.6810)^2}{10} = 370{,}003.94$$

$$s_{\bar{x}_{st}} = \sqrt{\left(\frac{1}{(233)^2}\right)\left[36{,}175.517 + 130{,}772.1 + 271{,}213.91 + 370{,}003.94\right]}$$

$$= \sqrt{\frac{1}{(233)^2}(808{,}165.47)} = 3.8583$$

Approximate 95% confidence interval: $\bar{x}_{st} \pm 2s_{\bar{x}_{st}}$ = 48.7821 $\pm$ 2(3.8583)
or 41.07 to 56.50

In euros: €41,070 to €56,500

b. Approximate 95% confidence interval: $N\bar{x}_{st} \pm 2Ns_{\bar{x}_{st}}$ = 233(48.7821) $\pm$ 2(233)(3.8583)
= 11,366.229 $\pm$ 1797.9678
or 9,568.26 to 13,164.20

In euros: €9,568,260 to €13,164,200

27. a. $p_A = 1/3$

$$s_{P_A} = \sqrt{\left(\frac{380-30}{380}\right)\left(\frac{(1/3)(2/3)}{30}\right)} = 0.0826$$

Approximate 95% confidence interval: $0.3333 \pm 2(0.0826)$ or 0.1681 to 0.4985

b. Estimated total number of successful start-ups in City B = 760(19/45) = 320. 9

c. $p_B = 19/45 = 0.4222$

$$s_{P_B} = \sqrt{\left(\frac{760-45}{760}\right)\left(\frac{(19/45)(26/45)}{45}\right)} = 0.0714$$

Approximate 95% confidence interval: $0.4222 \pm 2(0.0714)$ or 0.2794 to 0.5650

28. $$p_{st} = \left(\frac{380}{1400}\right)\left(\frac{10}{30}\right) + \left(\frac{760}{1400}\right)\left(\frac{19}{45}\right) + \left(\frac{260}{1400}\right)\left(\frac{7}{25}\right) = 0.3717$$

$$\sum N_h(N_h - n_h)\left[\frac{p_h(1-p_h)}{n_h}\right] =$$

$$380(350)\frac{(1/3)(2/3)}{30} + 760(715)\frac{(19/45)(26/45)}{45} + 260(235)\frac{(7/25)(18/25)}{25}$$

$$= 985.19 + 2945.84 + 492.71 = 4423.73$$

$$s_{P_{st}} = \sqrt{\left(\frac{1}{(1400)^2}\right)4423.73} = 0.0475$$

Approximate 95% confidence interval: $0.3717 \pm 2(0.0475)$ or 0.2767 to 0.4667

29. a. Estimated total number of start-ups that received no government financial support =
380 (9/30) + 760 (12/45) + 260 (11/25) = 431.0667

Estimate approximately 431 start-ups that received no government financial support.

b. $p_{st} = \left(\frac{380}{1400}\right)\left(\frac{9}{30}\right) + \left(\frac{760}{1400}\right)\left(\frac{12}{45}\right) + \left(\frac{260}{1400}\right)\left(\frac{11}{25}\right) = 0.3079$

$$\sum N_h (N_h - n_h)\frac{[p_h(1-p_h)]}{n_h} = (380)(380-30)(9/30)(21/30)/30 + (760)(760-45)(12/45)(33/45)/45 + (260)(260-25)(11/25)(14/25)/25 = 3894.6$$

$$s_{P_{st}} = \sqrt{\left(\frac{1}{(1400)^2}\right)3894.6} = 0.04458$$

Approximate 95% Confidence Interval: $0.3079 \pm 2(0.0446)$ or 0.2187 to 0.3971

30 a. $p_{st} = \left(\frac{380}{1400}\right)\left(\frac{21}{30}\right) + \left(\frac{760}{1400}\right)\left(\frac{34}{45}\right) + \left(\frac{260}{1400}\right)\left(\frac{15}{25}\right) = 0.7116$

$$\sum N_h (N_h - n_h)\frac{[p_h(1-p_h)]}{n_h} = (380)(380-30)(21/30)(9/30)/30 + (760)(760-45)(34/45)(11/45)/45 + (260)(260-25)(15/25)(10/25)/25 = 3747.8$$

$$s_{P_{st}} = \sqrt{\left(\frac{1}{(1400)^2}\right)3747.8} = 0.04373$$

Approximate 95% confidence interval: $0.7116 \pm 2(0.0437)$ or 0.6242 to 0.7990

b. Estimated total number of start-ups in the service sector = $1400(0.7116) = 996.24$

Estimate of total number of start-ups in the service sector ≈ 996

31. a. $\bar{x}_c = \frac{17(37) + 35(32) + \cdots + 57(44)}{17 + 35 + \cdots + 57} = \frac{11,240}{304} = 36.9737$

Estimate of mean age: 36.97 years

b. $$\sum (t_i - \bar{x}_c M_i)^2 = [\,17\,(37) - (36.9737)\,(17)\,]^2 + \cdots + [\,57\,(44) - (36.9737)\,(44)\,]^2$$
$$= (0.4471)^2 + (-174.0795)^2 + (-25.3162)^2 + (-460.2642)^2 + (173.1309)^2 + (180.3156)^2 + (-94.7376)^2 + (400.4991)^2$$
$$= 474,650.68$$

$$s_{\bar{X}_c} = \sqrt{\left(\frac{150-8}{(150)(8)(40)^2}\right)\left(\frac{474,650.68}{7}\right)} = 2.2394$$

Approximate 95% confidence interval for mean age of agents: $36.9737 \pm 2(2.2394)$ or 32.49 to 41.45

32. a. Proportion of university graduates: 128/304 = 0.4211

Proportion of males: 112/304 = 0.3684

b. $\sum(a_i - p_c M_i)^2 = [\ 3 - 0.4211\ (17)\ ]^2 + \cdots + [\ 25 - 0.4211\ (57)\ ]^2$

$= (-4.1587)^2 + (-0.7385)^2 + (-2.9486)^2 + (10.2074)^2 + (-0.1073)^2 + (-3.0532)^2$

$+ (-0.2128)^2 + (0.9973)^2$

$= 141.0989$

$$s_{P_c} = \sqrt{\left(\frac{150-8}{(150)(8)(40)^2}\right)\left(\frac{141.0989}{7}\right)} = 0.0386$$

Approximate 95% confidence interval for proportion of agents who are university graduates:

$0.4211 \pm 2(0.0386)$ or 0.3439 to 0.4983

c. $\sum(a_i - p_c M_i)^2 = [\ 4 - 0.3684\ (17)\ ]^2 + \cdots + [\ 26 - 0.3684\ (57)\ ]^2$

$= (-2.2628)^2 + (-0.8940)^2 + (-2.5784)^2 + (3.6856)^2 + (-3.8412)^2 + 1.5792)^2$

$+ (-0.6832)^2 + (5.0012)^2$

$= 68.8787$

$$s_{P_c} = \sqrt{\left(\frac{150-8}{(150)(8)(40)^2}\right)\left(\frac{68.8787}{7}\right)} = 0.0270$$

Approximate 95% confidence interval for proportion of agents who are male:

$0.3684 \pm 2(0.0270)$ or 0.3144 to 0.4224