AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

© 1992 by Wadsworth, Inc.

© 2006 by Shelley Rasmussen

Permission is granted by the author for non-for-profit educational use.

Shelley_Rasmussen@uml.edu

Minitab is a statistical package, a computer program that performs many statistical procedures. The versions of Minitab now available for use on personal computers are menu-driven and much easier to use than the main-frame version originally discussed in this text. Those sections are not included in this online edition of the text. At this time, the most recent version is Minitab 15, available at very reasonable prices for purchase or rental from:

www.e-academy.com/minitab

System Requirements

Processor:	PC with a 1 GHz 32- or 64-bit processor
Memory:	512 MB or more of available RAM
Disk Space:	125 MB free space available
Operating System:	Microsoft Windows 2000, XP, or Vista.
Display:	A display capable of 1024 X 768 or higher resolution
Software:	Adobe Acrobat Reader 5.0 or higher for Meet Minitab

ANSWER BOOK FOR RASMUSSEN'S
AN INTRODUCTION TO

STATISTICS
WITH
DATA ANALYSIS

Lee Panas
Shelley Rasmussen

Brooks/Cole Publishing Company
Pacific Grove, California

Contents

To the instructor: Most numerical solutions were obtained using Minitab. Student answers may differ somewhat, especially if calculated by hand. For instance, Minitab's algorithm for finding quartiles differs from that described in the text. The test statistics shown for Wilcoxon signed rank tests and Wilcoxon-Mann-Whitney tests are those printed by Minitab; students may use these to calculate the test statistics in forms described in the text.

Part I

Chapter 2

Exercise 2-1

A frequency plot can be created in Minitab by plotting the frequency (or percent) versus the qualitative variable, in a scatterplot. For this plot, the percent of total could be shown on the vertical axis and group (coded with numbers) on the horizontal axis. A finished frequency plot is obtained by drawing a vertical line segment from each plotted point to the horizontal axis. The student is asked to express his/her own preference between the table and the plot. Since there are not very many numbers, the table might be preferred. The plot might be preferred since it allows visual comparisons of the percents for the six groups.

Exercise 2-2

(a) A stem-and-leaf plot of the 147 toilet flush volumes shows a distribution with a major peak between 10 and 12 liters. Students may differ in descriptions of shape and symmetry.

(b) The distribution is unimodal, having one peak.

(c) The distribution is not exactly symmetrical, but can't really be described as negatively skewed or positively skewed, either.

(d) Among other things, there is a large concentration of average volumes in the interval from 7 to 12 liters; the vast majority of average volumes are in the interval from 6 to 14 liters.

(e) There would tend to be many 0 leaves if human chart readers tend to round to the nearest liter.

Exercise 2-3

(a) A stem-and-leaf plot of the 129 shower flow rates shows a distribution with a major peak in the interval from 6 to 8 liters/minute.

(b) Some students might describe the distribution as unimodal; others might describe it as bimodal, with a major peak over the interval 6-7 and a minor peak over the interval 9-10 liters/minute.

(c) The distribution might be described as slightly positively skewed.

(d) Among other things, there is a heavy concentration of average shower flow rates in the interval from 5 to 11 liters/minute; the vast majority of average shower flow rates are in the interval from 3 to 12 liters/minute.

(e) To use this sample to draw inferences about a larger population, we would have to consider the sample to be representative of that population. We would feel most comfortable about making such inferences if we knew the sample was a random sample from the population (that is, each member of the population had an equal and independent chance of being included in the sample).

Exercise 2-4

(a) I prefer interval widths of 5 and 10 °F over interval width 20 °F. With interval widths of 20, the histogram seems too compact; there are too few intervals to get a feel for the distribution of extreme cold temperatures among the 50 states.

(b) The distribution seems reasonably symmetrical.
(c) From the histogram with interval widths of 5 °F, the distribution might be considered bimodal, with peaks around −32.5 °F and −47.5 °F. From the histogram with interval widths of 10 °F, the distribution appears unimodal, with a peak around −35 °F.

Exercise 2-5
(a) I prefer interval widths of .1 and .2 unit over interval width .5 unit. With interval widths of .5 unit, the histogram seems too compact; there are too few intervals to get a feel for the distribution of stress loads at which graphite beams fractured.
(b) The distribution appears to be unimodal, having one peak.
(c) A box plot looks very symmetrical.
(d) The distribution appears reasonably symmetrical.
(e) The box plot is a simple plot, giving an idea of extreme values and center, and an impression of symmetry (or skewness). The first histogram (with interval widths .1 unit, having more intervals than the others) gives more information on the actual values of stress loads than do the other two histograms or the box plot.

Exercise 2-6
(a) and (b) The finished stem-and-leaf plot has a major peak in the interval from 9 to 10.5 psi.
(c) The distribution of yarn strengths is unimodal, with one peak.
(d) The distribution of yarn strengths is reasonably symmetrical.

Exercise 2-7
(a) Median survival for this group of 43 patients was 702 days.
(b) The first quartile is 440 and the third quartile is 1,367 days. A box graph would have whiskers extending to the 10th percentile (125.5, the average of 74 and 177) and the 90th percentile (2,050.5, the average of 2,045 and 2,056). The points less than the 10th percentile (7, 47, 58 and 74) and the points greater than the 90th percentile (2,056, 2,260, 2,429 and 2,509) would also the plotted.
(c) From a box plot or box graph, we cannot assess the shape of a distribution. The median survival time was 702 days (about 1.9 years), with an interquartile range (middle 50% of values) from 440 days (about 1.2 years) to 1,367 days (about 3.7 years). The shortest survival time was 7 days and the longest was 2,509 days (almost seven years). The distribution appears to be positively skewed.

Exercise 2-8
The distribution appears roughly symmetrical. The distribution might be described as unimodal, with a peak around 102.5 pregnancies per 1,000 girls. Students might see another peak around 82.5 pregnancies per 1,000 girls 15-19 years of age.

Exercise 2-9
A histogram of the shear strengths shows a shape similar to that shown in the stem-and-leaf plot. I prefer the stem-and-leaf plot for these data since it provides at least as much information as the histogram about the shape of the distribution, in addition to listing all the data values. A student might describe the distribution as unimodal, with a peak in the interval from 2,260 to 2,370 pounds per weld, or bimodal, with a peak near 2,275 and a peak near 2,350 pounds per weld. The distribution might be described as roughly symmetrical or slightly negatively skewed.

Exercise 2-10
A stem-and-leaf plot of these light intensity measurements from Minitab does not show the tenths place. There is a fairly even distribution of values from about 2 to about 13 candela per square meter. The distribution is roughly symmetrical. The minimum value is 1.5, the maximum is 17.3, the first quartile is 4.575, the median is 9.1 and the third quartile is 13.1 candela per square meter.

Exercise 2-11

(a) The median is 222, the first quartile is 79 and the third quartile is 445 acre-feet. In a box graph, whiskers would extend to the 10th percentile (12.6, the average of 7.7 and 17.5) and the 90th percentile (1,676.9, the average of 1,656 and 1,697.8). In addition, values less than the 10th percentile (4.1 and 7.7) and values greater than the 90th percentile (1,697.8 and 2,745.6) would be plotted.

(b) The rainfall measurement range from 4.1 acre-feet to 2,745.6 acre-feet. The middle 50% of the values are in the interquartile range from 79 to 445 acre-feet. The median measurement is 222 acre-feet. The distribution appears to be positively skewed.

Exercise 2-12

(a) A dot plot and a box plot of the sulfur dioxide concentrations both suggest symmetry of the distribution of values, and provide information on the center and spread of the values. The dot plot shows all 24 values, while the box plot just summarizes the distribution with five summary values.

(b) I would say the distribution is unimodal, with a center between 50 and 55 micrograms per cubic meter.

(c) The distribution seems to be roughly symmetrical.

(d) The sulfur dioxide concentration is much higher in this forest than in other parts of the country. No causal relationship can be inferred from this information. It certainly suggests a need for investigation into possible causes of the damage in this forest.

Exercise 2-13

(a) Both a dot plot and a box graph suggest symmetry in the middle portion of the observations. The dot plot shows that the distribution is quite symmetrical, except for the one extremely small (outlying) value. A box graph provides similar information, while the box plot provides less.

(b) The distribution is unimodal, having a major peak around 150 pounds per weld.

(c) The distribution is quite symmetrical, except for the one extremely small value.

Exercise 2-14

(a) Both a dot plot and a box graph suggest reasonable symmetry of the distribution, and provide information on center and variation of the values. The dot plot obviously provides more information, since the location of each value is indicated.

(b) The distribution is reasonably symmetrical. There is a concentration of values in the range from 50 to 60%, with percentages ranging from a little greater than 30% to almost 90%. A smaller concentration of values is in the range from 70 to 84%.

Exercise 2-15

(a) Both a dot plot and a box plot of these errors suggest reasonable symmetry of the distribution, and provide information on center and variation of the values. The dot plot provides more information, since each value is plotted.

(b) The distribution is reasonably symmetrical, with a median a little greater than 0. Errors range from −.1 to .1 meters. There is a fairly even spread of values across the range of the distribution.

Exercise 2-16

(a) A box plot suggests a reasonably symmetrical distribution, summarizing center and spread of the values. A stem-and-leaf plot shows all the values, showing the distribution to be unimodal.

(b) The center of these observations is between 24 and 25%. The values range from 21.9% to 27.8%. The distribution is symmetrical and unimodal.

(c) The X-ray microanalysis seemed to provide percentages generally less than the theoretical value of 26.6%.

Exercise 2-17
The distribution is positively skewed and unimodal, with a peak at the leftmost interval of the histogram. The lifespans range from 1,625 to 3,200, with a majority of values less than 2,150 kilometers.

Exercise 2-18
A frequency plot can be created in Minitab by plotting the frequency or count on the vertical axis of a scatterplot and the number of defects category on the horizontal axis. A student can then use vertical line segments to connect the plotted points with the horizontal axis. Clearly, the most frequent number of defects is 1, with 0 defects being the next most frequent, followed by 2 defects.

Exercise 2-19
The distribution is fairly symmetrical. The distribution is unimodal, with a peak around $65,000.

Exercise 2-20
The distribution is negatively skewed, with a peak in the interval from $15,000 to $18,000 (this interval contains a large majority of the values).

Exercise 2-21
The distribution of number of full-time employees for these 265 restaurants is positively skewed, the most frequent (modal) value being 0 full-time employees. In any survey situation in which there is not 100% response, we must be careful interpreting results. We have to worry about how responding restaurants differ from the nonrespondents, especially with respect to number of full-time employees (the variable of interest here).

Exercise 2-22
The distribution of length of stay is positively skewed, with a peak at the modal value of 1 day.

Exercise 2-23
A frequency plot for males can be constructed as a scatterplot with site (coded with numbers) on the vertical axis and number of males on the horizontal axis. Similarly for females. These frequency plots might be completed by using horizontal line segments to connect each plotted point with the vertical axis. To compare the two plots, we would probably want to use percentages instead of frequencies and use the same vertical scale for each plot.

Exercise 2-24
A frequency plot for these data can be constructed as a scatterplot with number of women on the vertical axis and length of fertile period on the horizontal axis. A student can complete the plot by using vertical line segments to connect each plotted point with the horizontal axis. The median is also the modal value, 8 days. This distribution is unimodal, with a peak in the interval from 8 to 9 days. The values 5, 6 and 7 are the next most frequently occurring values.

Exercise 2-25
The distribution appears to be somewhat negatively skewed. The percentages range from 33.8 to 93.6. The greatest concentration of values is in the interval from about 70 to 85%.

Chapter 3

Exercise 3-1
The mean of these eight salaries is $135,000; the median is $10,000; and the 15% trimmed mean is $12,000. The median and the 15% trimmed mean provide a more correct impression of a typical salary; the mean is drastically inflated by the president's huge salary.

Exercise 3-2
(a) The mean life expectancy for these 19 countries is 76.053 years.
(b) The 10% trimmed mean life expectancy based on 15 countries is 76.067 years.
(c) The weighted (by population size) mean life expectancy is 76.313 years.
(d) These three measures of central tendency are very similar. The weighted mean is somewhat larger than the mean because several of the more populous countries have life expectancies greater than 76.
(e) The United States has a life expectancy of 76, which is the same as each of the three measures of central tendency in parts (a)-(c), when rounded to a whole number.

Exercise 3-3
(a) The mean life expectancy for these 35 countries is 50.2 years.
(b) The 15% trimmed mean life expectancy based on 25 countries is 49.0 years.
(c) The weighted (by population size) life expectancy is 60.6 years.
(d) The trimmed mean is slightly smaller than the mean because several relatively large values are excluded from its calculation. The median value of 49 equals the 15% trimmed mean. The weighted mean is larger than the other measures of central tendency because of the longer life expectancies of several of the more populous countries.
(e) The mean life expectancy for China and India is 62.5 years, the weighted (by population size) mean life expectancy for China and India is 63.49 years. The weighted mean is larger since China has a larger population and a longer life expectancy than India.
(f) The mean life expectancy for the 33 countries excluding China and India is 49.45 years, the weighted (by population size) life expectancy for these 33 countries is 52.14 years. Again, the weighted mean is larger than the mean because of longer life expectancies for some more populous countries.
(g) The mean and weighted mean in part (f) are closer in value than those in parts (a) and (c) because the influence of the large countries with longer life expectancies (China and India) has been removed.

Exercise 3-4
(a) The median is 3, the mean is 10.04, the 5% trimmed mean is 8.15 and the 15% trimmed mean is 4.60 percent.
(b) The weighted (by population size) mean contraception use is 44.84 percent.
(c) Clearly the weighted mean is much larger than the other measures of central tendency because of the much greater contraception use reported by some of the most populous countries. The mean is also larger than the median and trimmed means, inflated by a few large values of contraception use.

Exercise 3-5
(a) A dot plot of these ozone measurements shows the values fairly evenly distributed from 0.08 to 0.35 ppm ozone.
(b) The mean is 0.2142 ppm, the 15% trimmed mean is 0.2138 ppm and the median is 0.2050 ppm, these three measures of central tendency are very similar in value.
(c) The range spans all the values, from 0.08 to 0.35 ppm, for a difference of 0.27 ppm. The interquartile range spans the middle 50% of the values, from 0.155 to 0.3025 ppm, for a difference of

0.1475. The standard deviation of 0.0854 ppm measures variation about the mean.
(d) The distribution is fairly symmetrical and evenly spread over the range of values.

Exercise 3-6
(a) A dot plot, like the list, shows most values in the range from 23 to 102, with a few much larger values.
(b) and (c) The distribution appears unimodal (although with so few points it is hard to say) and positively skewed.
(d) The mean is 163.4 hours, the 15% trimmed mean is 127.9 hours, and the median is 97 hours. Since the distribution is positively skewed, the median and trimmed mean are less than the mean.
(e) The range spans all the values from 23 to 487, for a difference of 464 hours. The interquartile range spans the middle 50% of the values, from 52.5 to 276.5, for a difference of 224 hours. The standard deviation of 154.4 hours measures variation of the values about the mean.

Exercise 3-7
(a) A plot shows the distribution to be roughly unimodal and somewhat positively skewed.
(b) The mean is 0.0036 inches, the 15% trimmed mean is 0.003471 inches, the median is 0.003 inches. These measures a very similar. Because the distribution is somewhat positively skewed, the mean is larger than the median.
(c) The range spans the errors from 0.001 to 0.008 inches, for a difference of 0.007 inches. The interquartile range spans the middle 50% of the values, from 0.002 to 0.005, for a difference of 0.003 inches. The standard deviation of 0.0019 inches measures variation of the values about their mean.

Exercise 3-8
(a) and (d) A plot shows the distribution to be unimodal and roughly symmetrical, centered round a value of about 13.3 seconds.
(b) The mean is 13.447, the 15% trimmed mean is 13.40, and the median is 13.3 seconds. Because of the approximate symmetry of the distribution, these three measures of central tendency are very similar.
(c) The range spans the values from 11.6 to 15.3, for a difference of 3.7 seconds. The interquartile range spans the middle 50% of the values, from 12.9 to 14.2, for a difference of 1.3 seconds.
(e) Since 11 of the 15 values are less than 14 seconds, it certainly appears that a byte of memory on this chip *can* to reprogrammed in less than 14 seconds. Since 4 of the 15 values exceed 14 seconds, the reprogramming time is not always less than 14 seconds.

Exercise 3-9
(a) and (b) A dot plot, like the list, shows three small values similar in magnitude, with the other three larger values more spread out. There are too few values to assess number of modes. The distribution might be described as positively skewed.
(c) The mean is 9.65, the 15% trimmed mean is 9.23 and the median is 8.9 mg/g fresh weight. The mean is larger than the median because of the spread in the three larger values.
(d) The range spans the values from 4.9 to 16.1, for a difference of 11.2 mg/g. The interquartile range spans the middle 50% of the values, from 5.1 to 14 mg/g, for a difference of 8.9 mg/g. The standard deviation of 4.91 mg/g measures variation of the values about their mean.

Exercise 3-10
(a) and (d) The values are fairly evenly spread from 167.9 to 208.4 and from 226.3 to 258.6. Because of this gap in the values, the distribution might be described as bimodal. Other descriptions would depend on the student.
(b) The mean is 209.75, the 5% trimmed mean is 209.25 and the median is 207.05 cm^3/cm H_2O.

These three measures of central tendency are very similar in value.
(c) The range is from 167.9 to 258.6, for a difference of 90.7 $cm^3/cm\ H_2O$. The interquartile range is from 191.05 to 228.30, for a difference of 37.25 $cm^3/cm\ H_2O$.

Exercise 3-11
(a) and (b) The middle four points are clustered together, with one extreme value on either side. The distribution might be described as slightly negatively skewed.
(c) The mean is 2.710, the 15% trimmed mean is 2.82 and the median is 2.82 mg/100 ml. The mean is slightly less than the others because of the slight negative skewness of the distribution.
(d) The range spans the values from 1.44 to 3.54, for a difference of 2.1 mg/100 ml. The interquartile range spans the middle 50% of the values, from 2.385 to 3.09, for a difference of 0.705 mg/100 ml. The standard deviation of 0.690 mg/100 ml measures variation about the mean.

Exercise 3-12
(a) and (b) The distribution might be described as unimodal and roughly symmetrical.
(c) The mean is 0.484, the 10% trimmed mean is 0.475 and the median is 0.49 milliroentgens per hour. These three measures of central tendency are very similar in value because of symmetry.
(d) The range spans the values from 0.15 to 0.89, for a difference of 0.74; the interquartile range spans the middle 50% of the values, from 0.31 to 0.65, for a difference of 0.34 milliroentgens per hour. The standard deviation of 0.2393 milliroentgens per hours measures variation about the mean.

Exercise 3-13
(a) Dot plots show similar variation for the four sets of values. Average distance hit increases for bat/ball combinations from top to bottom in the order listed below.
(b) The mean, standard deviation and range, respectively, are listed here (in feet) for each set of data. Wooden bat, Dudley Thunder: 241, 8.25, 20. Wooden bat, Worth Red Dot: 265.5, 7.05, 17. Aluminum bat, Dudley Thunder: 273.5, 7.51, 17. Aluminum bat, Worth Red Dot: 305, 11.94, 28.
(c) 24.5 feet
(d) 31.5 feet
(e) The advantage using the Worth Red Dot ball appears somewhat greater with the aluminum bat.
(f) 39.5 feet
(g) 32.5 feet
(h) The advantage using the aluminum bat appears somewhat greater with the Worth Red Dot ball.
(i) The variation in distances is very similar for the first three conditions; somewhat greater variation with the aluminum bat and Worth Red Dot ball.

Chapter 4

Exercise 4-1
(a) This scatterplot puts all four points in the upper-right part of the graph; the rest is wasted blank space.
(b) This graph is better than the one in part (a) because there is less wasted space. However, axes that meet can create an incorrect impression of an origin or (0,0) point.
(c) The range frame scatterplot does not waste space; since axes do not meet, the impression of an origin is not given.

Exercise 4-2
The student could create a dot chart in Minitab by plotting category (coded as numbers) on the vertical axis and cost on the horizontal axis, using the PLOT command. The student could use

horizontal lines to connect plotted points with the vertical axis. To obtain a dot chart showing cost and percentage, the student must use a different computer package or draw the graph by hand.

Exercise 4-3
The proportion of responses increases slowly with dose to a dose of 3% sodium saccharin in diet, then increases more sharply with dose after that.

Exercise 4-4
Proportion with retinopathy seems to increase in roughly a linear fashion with duration of patient's diabetes in years.

Exercise 4-5
The mean and standard deviation, respectively, are shown here for each method. Method 1: 16.24, 4.52; Method 2: 22.41, 8.94; Method 3: 25.254, 4.465.

Exercise 4-6
(a) The distribution of percent successful point-after-touchdown attempts is unimodal and negatively skewed. Percent successful field-goal attempts is more evenly distributed.
(b) A scatterplot of percent successful field-goal attempts versus percent successful point-after-touchdown attempts shows positive association, with much scatter.

Exercise 4-7
(a) Dot plots show both distributions to be fairly symmetrical and evenly spread over the corresponding range of values.
(b) A scatterplot of LSAT versus grades shows positive association, with scatter.

Exercise 4-8
(a) The student will find the distribution of second measurements shifted slightly to the left, toward smaller values; the shapes of the two distributions are similar.
(b) A scatterplot of second measurement versus first measurement shows that second measurements tend to be less than corresponding first measurements.

Exercise 4-9
Scores tend to be lower and less spread out for the published than for the unpublished articles. The mean and standard deviation are, respectively, 2.42 and 0.23 for unpublished reports; 1.75 and 0.18 for published articles.

Exercise 4-10
Both distributions are roughly symmetrical and even spread over the corresponding range of values. Bacterial counts tend to be somewhat higher for carpeted rooms. The mean and standard deviation are, respectively, 11.2 and 2.7 for carpeted rooms; 9.79 and 3.2 for uncarpeted rooms.

Exercise 4-11
(a) The mean, median and standard deviation, respectively, are 10.03, 10.05 and 5.08 for carbon monoxide concentration; 3.37, 1.85 and 3.055 for benzo(a)pyrene concentration.
(b) A scatterplot of carbon monoxide concentration versus benzo(a)pyrene concentration shows slight positive association.

Exercise 4-12
(a) Both distributions are roughly symmetrical and evenly spread over the corresponding range of values. The mean, median and standard deviation are, respectively, 46.5, 47.95 and 11.99 for sodium; 9.94, 9.5 and 2.116 for potassium.
(b) A scatterplot of sodium versus potassium shows slight negative association.

Exercise 4-13
Listed are, respectively, sample size, mean, median, range, interquartile range and sample standard deviation for each sport. Wrestling: 5, 57.58, 58.30, 13.6, 6.6, 5.36; weightlifting: 6, 45.12, 44.45, 10.6, 8, 4.39; shot/discus: 4, 45.60, 45.15, 6.9, 5.8, 3.45; ice hockey: 3, 56.57, 54.60, 7.9, 7.9, 4.30; cross country skiing: 4, 72.28, 73.45, 14.4, 7.65, 6.04.

Exercise 4-14
Listed are, respectively, sample size, mean, median, range and sample standard deviation for each sport. Basketball: 3, 44.93, 42.90, 7.3, 4.05; swimming: 3, 43.37, 43.4, 5.7, 2.85; distance running: 3, 57.17, 57.5, 12.4, 6.21; volleyball: 4, 47.95, 47.05, 14.3, 6.6.

Exercise 4-15
The plot in part (a) shows more nonsmokers in each cycle category. The plot in part (b) shows a greater percentage of nonsmokers in the first two cycle categories. If the first cycle category is excluded, there is still a greater percentage of nonsmokers than smokers who got pregnant in the second cycle.

Exercise 4-16
The results suggest that the 64K autoantibody might be useful as an early warning of Type I diabetes.

Exercise 4-17
(a) 74.4%
(b) 49.6%
(c) We might worry about the extent of agreement between animal studies and carcinogenicity in humans.

Exercise 4-18
A student may find it informative to look at row and column percentages in the two-way frequency table.

Exercise 4-19
HTLV-I antibodies were found in 59% of 17 patients with tropical spastic paraparesis and 4% of 303 people without the disease. Since HTLV-I is a virus associated with some lymphomas and leukemias, these results might suggest that the same virus is associated with tropical spastic paraparesis.

Exercise 4-20
Comparing percentages of trials in which predator spider pounced on prey suggests that the snowberry fly is less vulnerable to attack by the zebra spider than are the housefly and the blackened wing snowberry fly, but more vulnerable than a spider. The markings may be protective.

Exercise 4-21
There appears to be a relationship between sex and science career interest, since 19% of prepared females expressed an interest in science careers, compared with 64% of prepared males in this sample.

Exercise 4-22
In this group of cats there is an association between sex and presence of the virus, since 8% of females tested positive, compared with 20% of males.

Exercise 4-23
The results suggest abnormal liver function tests may be associated with obesity.

Exercise 4-24
A student may find it useful to compare percentages of males and females across the injury sites.

Exercise 4-25
Use of smokeless tobacco was reported by 14% of 10th graders, 21% of 11th graders and 24% of 12th graders. Other observations are left to the student.

Exercise 4-26
All four scatterplots show an increasing relationship between Down's ratio and mother's age as age increases from 17.5 to 32.5; the increasing relationship is much steeper over the ages from 37.5 to 47.5. The plots in parts (b) and (d) look like two straight lines, the one for older mothers steeper than for younger mothers.

Exercise 4-27
A scatterplot of number of beetles responding within 60 seconds versus dose rate looks like a straight line with positive slope.

Exercise 4-28
A student should find a two-way frequency table (cause of death versus sex) with row and column percentages helpful.

Chapter 5

Exercise 5-1
Male and female mice on the same treatment seem to have similar survival distributions. Survival is shorter for urethane and the highest dose of DDT.

Exercise 5-2
Yes, the data do suggest a possible synergistic antitumor effect of these two forms of immunotherapy.

Exercise 5-3
The distribution of body weights and the distribution of kidney weights for diabetic mice are shifted toward greater values than those for normal mice. The positive association between body weight and kidney weight is stronger for diabetic mice than for normal mice.

Exercise 5-4
Forearm tremor frequency decreases with increasing weight for each volunteer. Plots of frequency versus weight for each volunteer show that the over level of frequency differs across volunteers; the shapes of the profiles differ somewhat also.

Exercise 5-5
There is a fairly strong positive association between minority enrollment and minority teachers, and

between minority enrollment and minority new hires. The positive association between minority teachers and minority new hires is less strong.

Exercise 5-6
The range of values for each variable is small for these small cars. There is a fairly strong positive association between city gas and expressway gas. There is weaker positive association between city gas and weight, and between expressway gas and weight, with much scatter. Engine displacement shows negligible association with gas mileage.

Exercise 5-7
For Shannon River Valley, there is a strong positive association between depth and date, with the last point appearing as an outliner. For Bluff Cave, there is a stronger positive association between depth and date, with the point for depth 53 cm appearing as an outlier. The relationship between depth and radiocarbon age is very similar for the two sites.

Exercise 5-8
There is a positive association (not linear) between Ca and Cl. There is a strong (fairly linear) positive association between Na and Cl.

Exercise 5-9
After-hatch-year ducks tend to have greater body weight, greater wing length and somewhat greater condition index than hatch-year ducks. There is a positive association between weight and wing length, a very strong positive association between weight and condition index, and little association between wing length and condition index for both groups of ducks. Descriptive statistics are given in the answers at the back of the text.

Exercise 5-10
Cry counts and IQ scores are similar for girls and boys. The positive relationship between cry count and IQ score is similar for girls and boys.

Exercise 5-11
Obesity measurements tend to be greater for females than for males. Blood pressure measurements are similar for females and males. The positive association between obesity and blood pressure is stronger for females than for males.

Exercise 5-12
There is a positive (nonlinear) association between HVA and each of the variables IQ, MQ and IQ−MQ; a single outlying point strongly contributes to this impression of a positive association. There is a strong positive association between IQ and MQ. There is little association between IQ and IQ−MQ and between MQ and IQ−MQ.

Exercise 5-13
For all four experiments, there is a positive association between weight and requirements; the association is not quite as strong for Experiment 3 as for the others. Plotting requirements versus weight, the slope is somewhat steeper for Experiments 2 and 4 than for 1 and 3.

Exercise 5-14
The three plots are parallel, with relatively steep increase in recovery to week 2, flatter recovery over weeks 4-6. The overall recovery plot for sitting balance is greater than that for standing up, which is greater than the plot for independent walking.

Exercise 5-15
The male and female basketball players have similar values of maximal oxygen uptake. The males have greater values than females in the sports speed skating, cross-country skiing and distance running.

Exercise 5-16
There is a strong negative association between age and maximal oxygen capacity. There is an apparent positive association between age and height, age and weight, height and weight; a single outlying point creates this impression. There is negligible association between height and capacity, and between weight and capacity.

Exercise 5-17
More mosquitos were caught by human bait than by the electrocuting device at both sites; the two sites have similar distributions for these two variables. There is a stronger positive association between number caught by the device and number caught by the human for site 1 than for site 2 (the weaker association for site 2 is caused by the "outlying" trial 3.

Exercise 5-18
HIV and risk are associated for Burkina Faso and Ivory Coast.

Exercise 5-19 to Exercise 5-26
The student may find it helpful to calculate row and column percentages in frequency tables.

Exercise 5-27
There is a strong negative association between percent smokers and year for all specialties. The shape of the plot of percent smokers versus year varies with specialty.

Exercise 5-28
There is a fairly strong positive association between polio and water. There is fairly strong negative association between water and breastfeeding, between polio and breastfeeding, and between polio and low weight.

Exercise 5-28 to Exercise 5-35
Exploratory data analysis projects are left to the student.

Part II

Chapter 6

Exercise 6-1
(a) Frequencies are listed by smoking category, for sex categories in this order: girl, boy and total.
Never smoked: 379, 322, 701
Tried a few: 114, 187, 301
Regular smoker: 7, 20, 27
Former smoker: 15, 16, 31
Total: 515, 545, 1060
(b) Outcomes in the sample space are (girl, never smoked), (boy, never smoked), (girl, tried a few), (boy, tried a few), (girl, regular smoker), (boy, regular smoker), (girl, former smoker), (boy, former smoker). The corresponding probabilities are, respectively, .3575, .3038, .1075, .1764, .0066, .0189, .0142, .0151.

(c) P(girl)=.4858, P(never smoked)=.6613, P(tried a few)=.2840, P(regular smoker)=.0255, P(boy, regular smoker)=.0189, P(girl, never smoked)=.3575
(d) P(girl|regular smoker) = 7/27 = .2593, P(regular smoker|girl) = 7/515 = .0136, P(boy|tried a few) = 187/301 = .6213, P(tried a few|boy) = 187/545 = .3431
(e) Dependent; sex and smoking status seem to be associated characteristics here.

Exercise 6-2
(a) Frequencies are listed by combat exposure category, for career categories in this order: nurse, not a nurse, total.
Exposed: 94, 67, 161
Not exposed: 302, 1570, 1872
Total: 396, 1637, 2033
(b) Outcomes in the sample space are (nurse, exposed), (not a nurse, exposed), (nurse, not exposed), (not a nurse, not exposed). The corresponding probabilities are, respectively, .0462, .0330, .1485, .7723.
(c) P(exposed|nurse)=94/396=.2374; P(exposed)=.0792; dependent; combat and job status seem to be associated characteristics here.

Exercise 6-3
(a) The outcomes in the sample space are (boy, none), (boy, little), (boy, lots), (girl, none), (girl, little), (girl, lots). The corresponding probabilities are, respectively, .1275, .2550, .1275, .1225, .2450, .1225.
(b) Yes; independent

Exercise 6-4
(a) The probabilities for the outcomes (day, not acceptable), (day, ready), (evening, not acceptable), (evening, ready), (night, not acceptable), (night, ready) are, respectively, .0111, .3222, .0333, .3, .0667, .2667.
(b) No; not independent

Exercise 6-5
(a) The probabilities for the outcomes (boy, Rh positive), (boy, Rh negative), (girl, Rh positive), (girl, Rh negative) are, respectively, .4335, .0765, .4165, .0735.
(b) P(boy)=.51, P(Rh positive)=.85, P(boy, Rh positive)=.4335
(c) odds of a girl=.49/.51=.9608; odds of Rh positive=.85/.15=5.6667, odds of Rh positive girl=.4165/.5835=.7138
(d) P(girl|Rh positive)=.4165/.85=.49; P(girl)=.49; the probabilities are equal because sex and Rh factor are independent characteristics
(e) P(Rh positive|girl)=.4165/.49=.85; P(Rh positive)=.85; equal because of independence

Exercise 6-6
(a) The probabilities for the outcomes (boy, colorblind), (girl, colorblind), (boy, not colorblind), (girl, not colorblind) are, respectively, .031, .002, .479, .488.
(b) P(girl)=.49, P(colorblind)=.033, P(girl, colorblind)=.002
(c) odds of colorblindness=.033/.967=.0341; odds of a boy=.51/.49=1.0408; odds of a boy who is colorblind=.031/.969=.0320
(d) P(boy|colorblind)=31/33=.9394; P(colorblind|boy)=31/510=.0608
(e) for a girl, odds of colorblindness= 2/488=.0041; for a boy, odds of colorblindness=31/479=.0647; for a colorblind baby, odds of being a boy=31/2=15.5; for a noncolorblind baby, odds of being a boy=479/488=.9816
(f) P(colorblind)=.033; P(colorblind|boy)=.0608; colorblindness and sex are dependent or associated characteristics, so colorblindness is a sex-linked characteristic.

Exercise 6-7

(a) The probabilities for the outcomes (A fails, B fails), (A fails, B okay), (A okay, B fails), (A okay, B okay) are, respectively, .0005, .0095, .0495, .9405.

(b) P(unit functions)=.9995, P(unit fails)=.0005

(c) P(unit functions)=.9405, P(unit fails)=.0595

Exercise 6-8

(a) .5405; (b) .5455

(c) and (d) Let an outcome be denoted by win or loss for Bruins, Red Sox, Celtics, in that order. The possible outcomes are (WWW), (WWL), (WLW), (LWW), (LLW), (LWL), (WLL), (LLL) with probabilities .1622, .1327, .1351, .1379, .1149, .1128, .1105, .0940, respectively.

(d) Independence here implies that the results of one game would not in any way affect the outcome of the other games.

(e) P(WWW)=.1622; P(at least one wins)=.906; P(no wins)=.0940

Exercise 6-9

(a) P(AIDS|positive test)=.0054/.1545=.0350; false positive, P(no AIDS|positive test)=.9650

(b) P(No AIDS|negative test)=.8449/.8455=.9993; false negative, P(AIDS|negative test)=.0007

(c) A test with lower sensitivity and lower specificity can increase the chance of false positives and false negatives.

Exercise 6-10

(a) P(AIDS|positive test)=.00594/.01588=.3741; false positive, P(no AIDS|positive test)=.6259

(b) P(No AIDS|negative test)=.98406/.98412=.999939;
false negative, P(AIDS|negative test)=.000061

(c) A test with greater sensitivity and greater specificity can decrease the chance of false positives and false negatives.

Exercise 6-11

(a) P(AIDS|positive test)=.098/.161=.6087; false positive, P(no AIDS|positive test)=.3913

(b) P(No AIDS|negative test)=.837/.839=.9976; false negative, P(AIDS|negative test)=.0024

(c) With greater risk, the chance of a false positive decreases, the chance of a false negative increases.

Exercise 6-12

(a) P(AIDS|positive test)=.09/.225=.4; false positive, P(no AIDS|positive test)=.6

(b) P(No AIDS|negative test)=.765/.775=.9871; false negative, P(AIDS|negative test)=.0129

Exercise 6-13

(a) Assume independence of results in repeated testing, so P(++|AIDS) = P(+|AIDS)P(+|AIDS) = (.98)(.98) = .9604 and P(Not ++|AIDS) = 1 − .9604 = .0396. Similarly, P(++|No AIDS) = P(+|No AIDS)P(+|No AIDS) = (.07)(.07) = .0049 and P(Not ++|No AIDS) = 1 − .0049 = .9951. Then, use Bayes' Rule to show that P(AIDS|++) = .00576/.01063 = .54 and P(No AIDS|++) = .46 (false positives).

(b) P(No AIDS|No ++) = .9891/.9894 = .99976 and P(AIDS|No ++) = .00024 (false negatives).

(c) With retesting, the rate of false positives is decreased.

Exercise 6-14

(a) P(lung cancer|smoker)=.004504/.3=.015; odds=.004504/.295496=.0152

(b) P(lung cancer|nonsmoker)=.003496/.7=.00499; odds=.00502

(c) The odds are very similar to the conditional probabilities in both (a) and (b). The odds of lung

cancer is more than 3 times greater for smokers than nonsmokers.
(d) We assume that the same probabilities apply to all smokers and nonsmokers; we ignore other contributing factors.

Exercise 6-15
(a) P(birth defect|positive test)=.00095/.03092=.0307; false positive, P(no birth defect|positive test) = .9693
(b) P(no defect|negative test)=.96903/.96908=.999948; false negative, P(birth defect|negative result) = .000052
(c) The procedure has a low rate of false negatives, but very high rate of false positives.

Exercise 6-16
(a) (i) .9826, .0174; (ii) .4433, .5567; (b) (i) .1628, .8372; (ii) .9957, .0043
(c) The probability of a false positive increases as the proportion guilty decreases. The probability of a false negative decreases as the proportion guilty decreases.
(d) As in all experimental situations, the makeup of the control group can have a determining effect on our conclusions.

Exercise 6-17
(a) and (b) The probability of SSSS is .0001; the probability of each of the outcomes SSSF, SSFS, SFSS, FSSS is .0009; the probability of each of the outcomes SSFF, SFFS, FFSS, SFSF, SFFS, FSFS is .0081; the probability of each of the outcomes SFFF, FSFF, FFSF, FFFS is .0729; the probability of FFFF is .6561.
(c) .0037; .9477
(d) The probability that X takes the values 0, 1, 2, 3, 4 is, respectively, .6561, .2916, .0486, .0036, .0001; E(X)=.4; var(X)=.36; SD(X)=.6

Exercise 6-18
(a) The probability of four successes is .4096; the probability is .1024 for each of the four 3-success outcomes; the probability is .0256 for each of the six 2-success outcomes; the probability is .0064 for each of the four 1-success outcomes; the probability of four failures is .0016.
(b) .8192; .0272
(c) The probability that X takes the values 0, 1, 2, 3, 4 is, respectively, .0016, .0256, .1536, .4096, .4096; E(X)=3.2; var(X)=.64; SD(X)=.8

Exercise 6-19
(a) The probability of SSSS is .00077; the probability for each of the outcomes SSSF, SSFS, SFSS, FSSS is .00386; the probability for each of the outcomes SFFS, FFSS, SFSF, FSSF, FSFS, SSFF is .01929; the probability for each of the outcomes SFFF, FSFF, FFSF and FFFS is .09645; the probability of FFFF is .48225.
(b) The probability that W equals 0, 1, 2, 3 and 4 is, respectively, .48225, .38580, .11574, .01544 and .00077.
(c) E(W)=.66667 correct answers; var(W)=.55556 (correct answers)^2; SD(W)=.745 correct answers.
(d) .13195

Exercise 6-20
(a) .6; (b) .25; (c) .6; (d) E(Y)=3.7 points, Var(Y)=3.01 points^2, SD(Y)=1.7 points

Exercise 6-21

(a) The sample space is S={1, 2, 3, 4, 5, 6} and the probability function assigns the value 1/6 to each outcome in S.
(b) .5/.5=1
(c) The values W takes on are .5, 1.5, 2.5, −.5, −1.5, and −2.5, each with probability 1/6.
(d) 0.5
(e) 0 dollars
(f) Var(W)=2.92 dollars^2, SD(W)=1.71 dollars

Exercise 6-22
(a) The sample space consists of the outcomes 1, 2, 3, 4, 5 and 6 dots with probabilities 1/21, 2/21, 3/21, 4/21, 5/21 and 6/21, respectively.
(b) 6/15=0.4
(c) W takes values 0.5, 1.5, 2.5, −0.5, −1.5, −2.5 with probabilities 4/21, 5/21, 6/21, 3/21, 2/21, and 1/21, respectively.
(d) 15/21 = .7143
(e) 0.8333 dollars
(f) Var(W)=2.22 dollars^2, SD(W)=1.49 dollars

Exercise 6-23
(a) Y takes values 1, 2, 3, 4, 5 and 6, each with probability 1/6.
(b) 1
(c) 3.5 dollars
(d) Var(Y)=2.92 dollars^2 and SD(Y)=1.71 dollars

Exercise 6-24
(a) Y takes values 1, 2, 3, 4, 5 and 6, with probabilities 1/21, 2/21, 3/21, 4/21, 5/21 and 6/21, respectively.
(b) 1
(c) 4.33 dollars
(d) Var(Y)=2.22 dollars^2 and SD(Y)=1.49 dollars

Exercise 6-25
(a) If P(X=c)=1, then E(X)=cP(X=c)=c $\times$ 1=c and Var(X)=(c−c)^{2}P(X=c)=0 $\times$ 1=0.
(b) Certainly the variance of X can never be less than 0 since it is a sum of terms that are either positive or 0. If X takes at least two distinct values with positive probability, then at least one of these values, call it x, does not equal E(X). Therefore, the term (x−E(X))^{2}P(X=x) is positive, so Var(X) must be positive, not 0.

Exercise 6-26
Define X to be 0 with probability 2/3, to be 600 with probability 1/3. Then E(X)=0 $\times$ 2/3 + 600 $\times$ 1/3 = 200 disease victims saved.

Exercise 6-27
Define Y to be 600 with probability 2/3 and 0 with probability 1/3. Then E(Y)=600 $\times$ 2/3 + 0 $\times$ 1/3 = 400 deaths.

Exercise 6-28
(a) The sample space contains the 36 ordered pairs (i, j) where i and j are integers from 1 to 6. Each outcome in the sample space has probability 1/36.
(b) P(A)=1/2, P(B)=1/2, P(C)=4/36=1/9, P(A and B)=1/4, P(A and C)=1/18, P(B and C)=1/18, P(A and B and C)=0

(c) Pairwise independence is easily checked; note that $P(A)\times P(B)\times P(C)=4/144=1/36$ which does not equal P(A and B and C)=0.
(d) $P(A)=1/2$, $P(C)=1/9$, $P(D)=1/2$, P(A and C)=1/18, P(A and D)=9/36=1/4, P(C and D)=3/36=1/12, P(A and C and D)=1/36
(e) Note that P(A and C and D)=1/36=$P(A)\times P(C)\times P(D)$ but P(C and D)=1/12 does not equal $P(C)\times P(D)=1/18$.

Exercise 6-29
P(no hits)=$(.675)^4$=.2076; P(at least 1 hit)=1−.2076=.7924; discussion of assumptions is left to the student.

Exercise 6-30
(a) In a hybrid cross the sample space of possible genotypes for offspring has four outcomes: BB, Bb, bB and bb. Under the model of independent selection, each of these outcomes has probability 1/4. Since three of these genotypes represent the dominant phenotype and the other genotype the recessive phenotype, the dominant phenotype has probability 3/4 and the recessive phenotype has probability 1/4. Therefore, under the model of independent selection the odds in favor of the dominant phenotype is 3:1 (the expected ratio of dominant to recessive phenotypes).
(b) In a dihybrid cross, the sample space contains the 16 possible genotypes BBCC, BBCc, BBcC, BBcc, BbCC, BbCc, BbcC, Bbcc, bBCC, bBCc, bBcC, bBcc, bbCC, bbCc, bbcC, bbcc. Under the model of completely independent selection with equal probabilities, each genotype has probability 1/16. The student will note that the event corresponding to the BC phenotype contains 9 of the 16 genotypes, so this event has probability 9/16. Similarly, the events corresponding to phenotypes Bc, bC and bc have probabilities, respectively, 3/16, 3/16 and 1/16 under the independent selection model.

Chapter 7

Exercise 7-1 4!=24

Exercise 7-2 5!=120; 1/5

Exercise 7-3 $\binom{8}{5}=56$

Exercise 7-4 $\binom{9}{6}=84$; 6/9=2/3

Exercise 7-5 $\binom{36}{6}=1,947,792$; 1/1,947,792=.00000051

Exercise 7-6 There are $2\binom{3}{3}=2$ ways the series can end in 4 games, $2\binom{4}{3}=8$ ways the series can end in 5 games, $2\binom{5}{3}=20$ ways the series can end in 6 games, $2\binom{6}{3}=40$ ways the series can end in 7 games, for a total of 70 ways the series can end.

Exercise 7-7
(a) The results of one game does not in any way influence the chance of winning another game.
(b) Let n be the number of games played and p the (constant) probability you win a single game. Then your probability of winning the series is .3174 for n=5 and p=.4, .5 for n=5 and p=.5, .6826 for n=5 and p=.6, .2898 for n=7 and p=.4, .5 for n=7 and p=.5, .7102 for n=7 and p=.6. If the teams are evenly favored, you chance of winning the series is the same for 5 and 7 games. If you are favored, a 7 game series is to your advantage. If your opponent is favored, a 5 game series is to your advantage.

Exercise 7-8
(a) 2/20=1/10; (b) 1/2; (c) 1/5; (d) 19/20

Exercise 7-9
(a) With this decision rule, the probability of accepting the lot is 1, no matter what the proportion defectives in the lot. The graph of probability of accepting (vertical axis) versus proportion defective p (horizontal axis) is the horizontal line with probability of accepting always 1.
(b) If p = proportion defective in the lot, then P(accept the lot) = 1−p. For proportion defective p equal to 0, .1, .2, .3, .4, .5, .6, .7, .8, .9 and 1, respectively, the probability of accepting the lot is 1, .9, .8, .7, .6, .5, .4, .3, .2, .1 and 0. The graph of probability of accepting versus proportion defective is a straight line connecting the points (0, 1) and (1, 0).
(c) For proportion defective p, P(accept the lot) = $\binom{10-10p}{2}/\binom{10}{2}$.

For proportion defective p equal to 0, .1, .2, .3, .4, .5, .6, .7, .8, .9 and 1, respectively, the probability of accepting the lot is 1, .8, .62, .47, .33, .22, .13, .07, .02, .00 and 00. The graph of probability of accepting versus proportion defective p is a curve always below the line in (b).

(d) For proportion defective p, P(accept the lot) = $\binom{10-10p}{5}/\binom{10}{5}$.

For proportion defective p equal to 0, .1, .2, .3, .4, .5, .6, .7, .8, .9 and 1, respectively, the probability of accepting the lot is 1, .5, .22, .08, .02, .004, 0, 0, 0, 0, 0. The graph of probability of accepting versus proportion defective p is a curve always below the curve in (c).
(e) For proportion defective p, P(accept the lot) equals 1 if p=0 and equals 0 if p>0. With this plan, we accept the lot only if all 10 cars are acceptable.
(f) Sampling costs increase from (a) to (e), while the chance of shipping defective cars decreases from (a) to (e).
(g) For proportion defective p, P(accept the lot) = $\binom{10-10p}{2}/\binom{10}{2} + \binom{10p}{1}\binom{10-10p}{1}/\binom{10}{2}$.

For proportion defective p equal to 0, .1, .2, .3, .4, .5, .6, .7, .8, .9 and 1, respectively, the probability of accepting the lot is 1, 1, .98, .94, .86, .78, .66, .54, .38, .20 and 0. The graph of probability of accepting versus proportion defective p is a curve always above the line in (b).

(h) For proportion defective p, P(accept the lot) = $\binom{10-10p}{5}/\binom{10}{5} + \binom{10p}{1}\binom{10-10p}{4}/\binom{10}{5}$.

For proportion defective p equal to 0, .1, .2, .3, .4, .5, .6, .7, .8, .9 and 1, respectively, the probability of accepting the lot is 1, 1, .78, .5, .26, .1, .02, 0, 0, 0, 0. This curve is always above that for (d) and always below that for (g). It crosses (b) before p=.2 and crosses (c) before p=.4. Plan (h) is better than (b) and (c) for greater values of p, worse for smaller values.

Exercise 7-10
(a) The sex of one child does not in any way influence the probability that the other child is a girl.
(b) The four outcomes MM, MF, FM and FF have probabilities, respectively, $(1-p)^2$, $p(1-p)$, $(1-p)p$ and p^2 where M denotes a male child and F denotes a female child.
(c) Y has values 0, 1 and 2 with probabilities, respectively, $(1-p)^2$, $2p(1-p)$ and p^2.
(d) $2p(1-p)$; p^2; $(1-p)^2$; $(1-p)^2+p^2$
(e) 24/25; 1; 26/25

Exercise 7-11
(a) For N=10, n=3, m_1=5 and m_2=5, P(X=k) = $\binom{5}{k}\binom{5}{3-k}/\binom{10}{3}$ for k=0, 1, 2, 3

(b) For N=10, n=3, m_1=2 and m_2=8, P(X=k) = $\binom{2}{k}\binom{8}{3-k}/\binom{10}{3}$ for k=0, 1, 2

(c) For N=10, n=3, m_1=8 and m_2=2, $P(X=k) = \binom{8}{k}\binom{2}{3-k}/\binom{10}{3}$ for k=1, 2, 3

Exercise 7-12

(a) The results of any events do not in any way affect the probability of her winning another event.

(b) Let Y denote a win and N a loss. The 8 possible outcomes, corresponding to her results in the three distinct events, are YYY, YYN, YNY, NYY, YNN, NNY, NYN, NNN, with probabilities, respectively, p^3, $p^2(1-p)$, $p^2(1-p)$, $p^2(1-p)$, $p(1-p)^2$, $p(1-p)^2$, $p(1-p)^2$, $(1-p)^3$.

(c) $P(X=0)=(1-p)^3$, $P(X=1)=3p(1-p)^2$, $P(X=2)=3p^2(1-p)$, $P(X=3)=p^3$

(d) For p=.4 we have P(X=0)=.216, $P(X \geq 1)$=.784, P(X=2 or 3)=.352, P(X=3)=.064, E(X)=1.2; for p=.5 we have P(X=0)=.125, $P(X \geq 1)$=.875, P(X=2 or 3)=.5, P(X=3)=.125, E(X)=1.5; for p=.6 we have P(X=0)=.064, $P(X \geq 1)$=.936, P(X=2 or 3)=.648, P(X=3)=.216, E(X)=1.8.

Exercise 7-13

(a) $\binom{10}{2}\binom{10}{2}/\binom{20}{4} + \binom{10}{3}\binom{10}{1}/\binom{20}{4} + \binom{10}{4}\binom{10}{0}/\binom{20}{4} = .7090$

(b) $\binom{5}{2}\binom{10}{2}/\binom{15}{4} + \binom{5}{3}\binom{10}{1}/\binom{15}{4} + \binom{5}{4}\binom{10}{0}/\binom{15}{4} = .4066$

(c) $\binom{3}{2}\binom{9}{2}/\binom{12}{4} + \binom{3}{3}\binom{9}{1}/\binom{12}{4} = .2364$

(d) (.7090)(.4066)(.2364)=.0681

(e) If selection was completely random, there is less than a 7% chance that 2 or more women would be selected from each shift. This is a fairly rare event. The women might be justified in thinking the selection was not random.

(f) The probability of 3 or more women from each shift being selected if selection was completely random and independent across shifts is (.2910)(.0769)(.0182)=.0004. This event is so rare under the random selection hypothesis, that the women would be justified in rejecting that hypothesis.

(g) The probability of 1 or more women from each shift being selected if selection was completely random and independent across shifts is (.9567)(.8462)(.7455)=.6. This probability is large and does not discredit the random selection claim.

Exercise 7-14

(a) and (d) These are for the student to discuss.

(b) Binomial(n,p)

(c) (i) .9219; (ii) .5256; (iii) .9984; (iv) .9289; (v) .9591; (vi) .4148; (vii) .9999; (viii) .9568

Exercise 7-15

(a) When n=2, a binomial random variable has values 0, 1 and 2 with probabilities, respectively: .81, .18, .01 when p=1/10; .5625, .3750, .0625 when p=1/4; .25, .5, .25 when p=1/2; .0625, .3750, .5625 when p=3/4; .01, .18, .81 when p=9/10.

When n=3, a binomial random variable has values 0, 1, 2 and 3 with probabilities, respectively: .729, .243, .027, .001 when p=1/10; .4219, .4219, .1406, .0156 when p=1/4; .125, .375, .375, .125 when p=1/2; .0156, .1406, .4219, .4219 when p=3/4; .001, .027, .243, .729 when p=9/10.

When n=4, a binomial random variable has values 0, 1, 2, 3 and 4 with probabilities, respectively: .6561, .2916, .0486, .0036, .0001 when p=1/10; .3164, .4219, .2109, .0469, .0039 when p=1/4; .0625, .25, .375, .25, .0625 when p=1/2; .0039, .0469, .2109, .4219, .3164 when p=3/4; .0001, .0036, .0486, .2916, .6561 when p=9/10.

When n=5, a binomial random variable has values 0, 1, 2, 3, 4 and 5 when probabilities, respectively: .5905, .3281, .0729, .0081, .0005, .0000 when p=1/10; .2373, .3955, .2637, .0879, .0146, .001 when p=1/4; .0313, .1562, .3125, .3125, .1562, .0313 when p=1/2; .001, .0146, .0879, .2637, .3955, .2373 when p=3/4; .0000, .0005, .0081, .0729, .3281, .5905 when p=9/10.
(b) Students can use the answers to part (a) to construct these graphs.
(c) The expected value and variance of a binomial random variable and shown below for probabilities p equal to: 1/10; 1/4; 1/2; 3/4; and 9/10, respectively:
n=2: .2, .18; .5, .375; 1, .5; 1.5, .375; 1.8, .18
n=3: .3, .27; .75, .5625; 1.5, .75; 2.25, .5625; 2.7, .27
n=4: .4, .36; 1, .75; 2, 1; 3, .75; 3.6, .36
n=5: .5, .45; 1.25, .9375; 2.5, 1.25; 3.75, .9375; 4.5, .45

Chapter 8

Exercise 8-1 (a) .0606; (b) .0082; (c) .6247; (d) .2417; (e) .0054

Exercise 8-2 (a) .0475; (b) .9992; (c) .0471; (d) .7486; (e) .0471; (f) .9544; (g) .1587; (h) .1587; (i) .0038

Exercise 8-3 The distribution is fairly symmetrical with a modal interval from .8 to 1.4 ppm.

Exercise 8-4 The distribution is somewhat positively skewed.

Exercise 8-5 The distribution for control patients is somewhat positively skewed (mean=4.09, standard deviation=.439). The distribution for obese patients is fairly symmetrical (mean=4.53, standard deviation=.803).

Exercise 8-6 These 16 values are fairly evenly distributed between the minimum (71.69) and maximum (151) values.

Exercise 8-7 (a) The distribution is reasonably symmetrical and unimodal.
(b) This distribution is positively skewed, with a modal interval of .3 to .4 μmoles per 100 ml urine.

Exercise 8-8 The plot on page 31 of the text does not look Gaussian; it looks multimodal.

Exercise 8-9 The plot on page 47 of the text does not look Gaussian; it is bimodal.

Exercise 8-10 The plot on page 38 of the text is positively skewed.

Exercise 8-11 The plot on page 45 of the text is positively skewed.

Exercise 8-12 The plot on page 45 of the text is positively skewed, with a modal value of 0.

Exercise 8-13
(b) The mean is 264.02 yards, the standard deviation is 7.92 yards.
(c) (i) .1779; (ii) .2428; (iii) .2286; (iv) .1271; (v) .2236

(d) (i) 3/19=.1579; (ii) 4/19=.2105; (iii) 5/19=.2632; (iv) 3/19=.1579; (v) 4/19=.2105
(e) The plot of average drive is reasonably symmetrical and unimodal.
(f) The plot of earnings is positively skewed.

Exercise 8-14
(b) The mean is 71.38 percent and the standard deviation is 26.05 percent.
(c) (i) .0559; (ii) .2389; (iii) .4232
(d) (i) 4/29=.1379; (ii) 10/29=.3448; (iii) 8/29=.2759
(e) The distribution is negatively skewed, with a modal interval of 90-98 percent.

Exercise 8-15
(b) The mean is $67,055 and the standard deviation is $14,102.
(c) (i) .2668; (ii) .1525; (iii) .1788; (iv) .1131
(d) (i) 20/50=.4; (ii) 11/50=.22; (iii) 8/50=.16; (iv) 5/50=.1

Exercise 8-16 The plot on page 255 of the text is positively skewed and multimodal.

Exercise 8-17
(a) The mean is 12.5, the variance is 6.25, the standard deviation is 2.5.
(b) (i) .7704; (ii) .2122; (iii) .2122; (iv) .9568; (v) .0216; (vi) .0216
(c) (i) .6826; (ii) .1587; (iii) .1587; (iv) .9282; (v) .0139; (vi) .0139
(d) The approximation is not too bad.

Exercise 8-18
(a) The mean is 2.5, the variance is 2.25, the standard deviation is 1.5.
(b) (i) .8302; (ii) .2364; (iii) .2712; (iv) .9666;; (v) .0334
(c) (i) .6826; (ii) .1587; (iii) .1587; (iv) .9050; (v) .0099
(d) The approximation is not very good.

Exercise 8-19
(a) and (c) The greater the number of tosses, the more the frequency plot resembles a Gaussian distribution.
(b) For the version of the Central Limit Theorem in the text, we must assume that the tosses were all independent (within and between students) and that the probability of heads was the same for each toss.

Part III

Chapter 9

Exercise 9-1 (c)

Exercise 9-2 (f)

Exercise 9-3 (d)

Exercise 9-4 (c)

Exercise 9-5 (b)

Exercise 9-6

(a) See definition of significance level in text.

(b) .8891, .5583, .0862, .0862, .5583, .8891

Exercise 9-7

(a) The significance level is .006, with interpretation as discussed in the text.

(b) .6590, .2749, .0199, .0199, .2749, .6590

Exercise 9-8

(a) The significance level is .0004, with interpretation as discussed in the text.

(b) .2824, .0687, .0022, .0022, .0687, .2824

Exercise 9-9

(a) The significance level is .2188, with interpretation as discussed in the text.

(b) .8858, .657, .2743, .2743, .657, .8858

Exercise 9-10

(a) The significance level is .0312, with interpretation as discussed in the text.

(b) .5314, .2622, .0508, .0508, .2622, .5314

Exercise 9-11

(a) The p-value equals $\binom{4}{0}\binom{4}{4}/\binom{8}{4} + \binom{4}{4}\binom{4}{0}/\binom{8}{4} = 2/70 = .0286$

(b) The p-value equals $\binom{8}{0}\binom{8}{8}/\binom{16}{8} + \binom{8}{8}\binom{8}{0}/\binom{16}{8} = 2/12{,}870 = .000155$

(c) In Example 9-2 with 2 runs under each condition, p-value = .33. Increasing sample size increases the power of the test.

Chapter 10

Exercise 10-1
(b) A 95% confidence interval for the population mean: (103.61, 133.39) μmol/liter
(c) An approximate 95% confidence interval for the population mean: (102.5, 133.5) μmol/liter
(d) An approximate 95% confidence interval for the population median: (100.9, 139.4) μmol/liter

Exercise 10-2
(b) Test the null hypothesis that the median difference is 0 versus the alternative hypothesis that the median difference is greater than 0. Using a t test with 4 degrees of freedom, the test statistic equals 4.4, with a p-value of .006. Using a Wilcoxon signed rank test for sample size 5, the test statistic equals 15 (or 0), with a p-value of .03. Using a sign test for sample size 5, the test statistic equals 0, with a p-value of .03. The patent claim seems to be justified based on these experimental results.

Exercise 10-3
(b) Test the null hypothesis that the mean fasting blood sugar level for women during their third trimester of pregnancy equals 80 mg/100 ml, versus the two-sided alternative. Using a large sample test based on the standard Gaussian distribution, the test statistic equals -7.4 with a p-value less than .0001. Similar results are found with a t test, a Wilcoxon signed rank test and a sign test.
(c) Large sample approximate 99% confidence interval for the population mean based on a standard Gaussian distribution: (66.7, 73.6) mg/100 ml.

Exercise 10-4
(a) Test the null hypothesis that the proportion of divers with a history of decompression sickness who that this heart defect equals .05, versus the two-sided alternative. The large sample test statistic equals 7.96, with a p-value less than .0001. Using the Binomial(30, .05) distribution for the small sample test, the p-value is also less than .0001.
(b) Large sample approximate 95% confidence interval for the proportion of interest is (.19, .54).

Exercise 10-5
(b) Use the sign test to test the null hypothesis that the median time to failure is 20 hours versus the two-sided alternative. We have 18 failures before 20 hours and 17 failures after 20 hours, for a p-value of 1. The experimental results are consistent with the null hypothesis.
(c) An approximate 95% confidence interval for the median time to failure based on a binomial distribution is (17.8, 25.5) hours.

Exercise 10-6
(a) Let p denote the proportion of people with coronary artery disease who are clinically depressed. Test the null hypothesis that p equals 1/3 versus the two-sided alternative. The large sample test statistic equals -2.45 with a p-value of .01.
(b) Large sample approximate 95% confidence interval for p is (.07, .28). These study suggests that the proportion p of people with coronary artery disease who are clinically depressed is less than 1/3.