Part IV

Complete Solutions

Chapter 1 Getting Started

Section 1.1

1. (a) The variable is the response regarding frequency of eating at fast-food restaurants.
   (b) The variable is qualitative. The categories are the number of times one eats in fast-food restaurants.
   (c) The implied population is responses for all adults in the U.S.

2. (a) The variable is the miles per gallon.
   (b) The variable is quantitative because arithmetic operations can be applied the mpg values.
   (c) The implied population is gasoline mileage for all new 2001 cars.

3. (a) The variable is student fees.
   (b) The variable is quantitative because arithmetic operations can be applied to the fee values.
   (c) The implied population is student fees at all colleges and universities in the U.S.

4. (a) The variable is the shelf life.
   (b) The variable is quantitative because arithmetic operations can be applied to the shelf life values.
   (c) The implied population is the shelf life of all Healthy Crunch granola bars.

5. (a) The variable is the time interval between check arrival and clearance.
   (b) The variable is quantitative because arithmetic operations can be applied to the time intervals.
   (c) The implied population is the time interval between check arrival and clearance for all companies in the five-state region.

6. Form B would be better. Statistical methods can be applied to the ordinal data obtained from Form B, but not to the answers obtained from Form A.

7. (a) *Length of time to complete an exam* is a ratio level of measurement. The data may be arranged in order, differences and ratios are meaningful, and a time of 0 is the starting point for all measurements.
   (b) *Time of first class* is an interval level of measurement. The data may be arranged in order and differences are meaningful.
   (c) *Class categories* is a nominal level of measurement. The data consists of names only.
   (d) *Course evaluation scale* is an ordinal level of measurement. The data may be arranged in order.
   (e) *Score on last exam* is a ratio level of measurement. The data may be arranged in order, differences and ratios are meaningful, and a score of 0 is the starting point for all measurements.
   (f) *Age of student* is a ratio level of measurement. The data may be arranged in order, differences and ratios are meaningful, and an age of 0 is the starting point for all measurements.

8. (a) *Salesperson's performance* is an ordinal level of measurement. The data may be arranged in order.
   (b) *Price of company's stock* is a ratio level of measurement. The data may be arranged in order, differences and ratios are meaningful, and a price of 0 is the starting point for all measurements.
   (c) *Names of new products* is a nominal level of measurement. The data consist of names only.
   (d) *Room temperature* is an interval level of measurement. The data may be arranged in order and differences are meaningful.

(e) *Gross income* is a ratio level of measurement. The data may be arranged in order, differences and ratios are meaningful, and an income of 0 is the starting point for all measurements.

(f) *Color of packaging* is a nominal level of measurement. The data consist of names only.

9. (a) *Species of fish* is a nominal level of measurement. Data consist of names only.
   (b) *Cost of rod and reel* is a ratio level of measurement. The data may be arranged in order, differences and ratios are meaningful, and a cost of 0 is the starting point for all measurements.
   (c) *Time of return home* is an interval level of measurement. The data may be arranged in order and differences are meaningful.
   (d) *Guidebook rating* is an ordinal level of measurement. Data may be arranged in order.
   (e) *Number of fish* caught is a ratio level of measurement. The data may be arranged in order, differences and ratios are meaningful, and 0 fish caught is the starting point for all measurements.
   (f) *Temperature of the water* is an interval level of measurement. The data may be arranged in order and differences are meaningful.

Section 1.2

1. Essay

2. Answers vary. Use groups of 3 digits.

3. Answers vary. Use groups of 4 digits.

4. Answers vary. Use groups of 3 digits.

5. (a) Assign a distinct number to each subject. Then use a random number table. Group assignment methods vary.
   (b) Repeat part (a) for 22 subjects.
   (c) Answers vary.

6. Answers vary. Use single digits with odd corresponding to heads and even to tails.

7. (a) Yes, it is appropriate that the same number appears more than once because the outcome of a die roll can repeat. The outcome of the 4th roll is 2.
   (b) No, we do not expect the same sequence because the process is random.

8. Answers vary. Use groups of 3 digits.

9. (a) Reasons may vary. For instance, the first four students may make a special effort to get to class on time.
   (b) Reasons may vary. For instance, four students who come in late might all be nursing students enrolled in an anatomy and physiology class that meets the hour before in a far-away building. They may be more motivated than other students to complete a degree requirement.
   (c) Reasons may vary. For instance, four students sitting in the back row might be less inclined to participate in class discussions.
   (d) Reasons may vary. For instance, the tallest students might all be male.

10. In all cases, assign distinct numbers to the items, and use a random-number table.

11. In all cases, assign distinct numbers to the items, and use a random-number table.

12. Answers vary. Use single digits with even corresponding to true and odd corresponding to false.

13. Answers vary. Use single digits with correct answer placed in corresponding position.

14. (a) This technique is stratified sampling. The population was divided into strata (4 categories of length of hospital stay), then a simple random sample was drawn from each stratum.
    (b) This technique is simple random sampling. Every sample of size n from the population has an equal chance of being selected and every member of the population has an equal chance of being included in the sample.
    (c) This technique is cluster sampling. There are 5 geographic regions and a random sample of hospitals is selected from each region. Then, for each selected hospital, all patients on the discharge list are surveyed to create the patient satisfaction profiles. Within each hospital, the degree of satisfaction varies patient to patient. The sampling units (the hospitals) are clusters of individuals who will be studied.
    (d) This technique is systematic sampling. Every k^{th} element is included in the sample.
    (e) This technique is convenience sampling. This technique uses results or data that are conveniently and readily obtained.

15. (a) This technique is simple random sampling. Every sample of size n from the population has an equal chance of being selected and every member of the population has an equal chance of being included in the sample.
    (b) This technique is cluster sampling. The state, Hawaii, is divided into regions using, say, the first 3 digits of the Zip code. Within each region a random sample of 10 Zip code areas is selected using, say, all 5 digits of the Zip code. Then, within each selected Zip codes, all businesses are surveyed. The sampling units, defined by 5 digit Zip codes, are clusters of businesses, and within each selected Zip code, the benefits package the businesses offer their employees differs business to business.
    (c) This technique is convenience sampling. This technique uses results or data that are conveniently and readily obtained.
    (d) This technique is systematic sampling. Every k^{th} element is included in the sample.
    (e) This technique is stratified sampling. The population was divided into strata (10 business types), then a simple random sample was drawn from each stratum.

Section 1.3

1. (a) This is an observational study because observations and measurements of individuals are conducted in a way that doesn't change the response or the variable being measured.
    (b) This is an experiment because a treatment is deliberately imposed on the individuals in order to observe a possible change in the response or variable being measured.
    (c) This is an experiment because a treatment is deliberately imposed on the individuals in order to observe a possible change in the response or variable being measured.
    (d) This is an observational study because observations and measurements of individuals are conducted in a way that doesn't change the response or the variable being measured.

2. (a) A census was used because data for all the games were used.
    (b) An experiment was used. A treatment is deliberately imposed on the individuals in order to observe change in the response or variable being measured.
    (c) A simulation was used because computer imaging of runners was used.
    (d) Sampling was used because measurements from a representative part of the population were used.

3. (a) Sampling was used because measurements from a representative part of the population were used.
    (b) A simulation was used because computer programs that mimic actual flight were used.
    (c) A census was used because data for all scores are available.

(d) An experiment was used. A treatment is deliberately imposed on the individuals in order to observe change in the response or variable being measured.

4. (a) No, "over the last few years" could mean the last three years to some and the last five years to others, etc.; answers vary.
   (b) Yes. The response to doubling fines would be affected by whether the responder had ever run a stop sign.
   (c) Answers vary.

5. (a) Use random selection to pick 10 calves to inoculate. Then test all calves to see if there is a difference in resistance to infection between the two groups. There is no placebo being used.
   (b) Use random selection to pick 9 schools to visit. Then survey all the schools to see if there is a difference in views between the two groups. There is no placebo being used.
   (c) Use random selection to pick 40 volunteers for skin patch with drug. Then survey all volunteers to see if a difference exists between the two groups. A placebo for the remaining 35 volunteers in the second group is used.

6. (a) Use random selection to pick 25 cars for high-temperature bond tires. Then examine tires of all the cars to see if a difference exists between the two groups. This is a double-blind experiment because neither the individuals in the study nor the observers know which subjects are receiving the new tires.
   (b) Use random selection to pick 10 bags. Then send all bags through the security check. This is not a double-blind experiment because the agent carrying the bag knows whether or not the bag contains a weapon.
   (c) Use random selection to pick 35 patients for new eye drops. Then measure eye pressure for all patients to see if a difference exists between the two groups. This is a double-blind experiment because neither the patients nor the doctors know which subjects are receiving the new drops.

Chapter 1 Review

1. Answers vary.

2. The implied population is the opinions of all the listeners. The variable is the opinion of a caller. There is probably bias in the selection of the sample because those with the strongest opinions are most likely to call in.

3. Essay

4. Name, social security number, color of hair and eyes, address, phone number, place of birth, and college major are all nominal because the data consist of names or qualities only. Letter grade on test is ordinal because the data may be arranged in order. Year of birth is interval because the data may be arranged in order and differences are meaningful. Height, age, and distance from home to college are ratio because the data may be arranged in order, differences and ratios are meaningful, and 0 is the starting point for all measurements.

5. In the random number table use groups of 2 digits. Select the first six distinct groups of 2 digits that fall in the range from 01 to 42. Choices vary according to the starting place in the random number table.

6. (a) Cluster sampling was used because a random sample of 10 telephone prefixes was selected and all households in the selected prefixes were included in the sample.
   (b) Convenience sampling was used because it uses results or data that are conveniently and readily obtained.
   (c) Systematic sampling was used because every k^{th} element is included in the sample.

(d) Random sampling was used because every sample of size 30 from the population has an equal chance of being selected and every member of the population has an equal chance of being included.

(e) Stratified sampling was used because the population was divided into strata (three age categories), then a simple random sample was drawn from each stratum.

7. (a) This is an observational study because observations and measurements of individuals are conducted in a way that doesn't change the response or the variable being measured.

(b) This is an experiment because a treatment is deliberately imposed on the individuals in order to observe a possible change in the response or variable being measured.

8. (a) Use random selection to pick half to solicit by mail. Then compute the percentage of donors in each group. Compare the results. No placebo was used.

(b) Use random selection to pick 43 volunteers to be given whitening gel. Evaluate tooth whiteness for all participants. Compare the results. A placebo was used with the remaining 42 in the second group. The experiment could be double-blind if the observers did not know which subjects were receiving the tooth whitening chemicals.

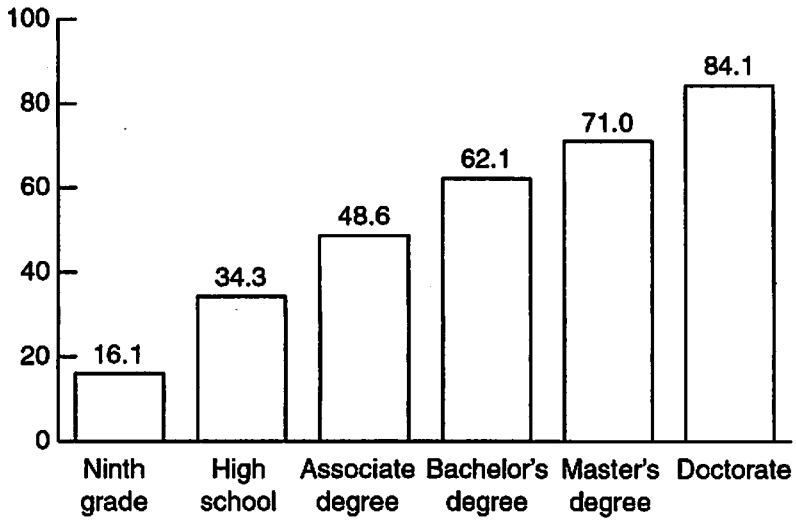
9. This is a good problem for class discussion. Some items such as age and grade point average might be sensitive information. You could ask the class to design a data form that can be filled out anonymously. Other issues to discuss involve the accuracy and honesty of the responses.

10. Students may easily spend several hours at this Web site.

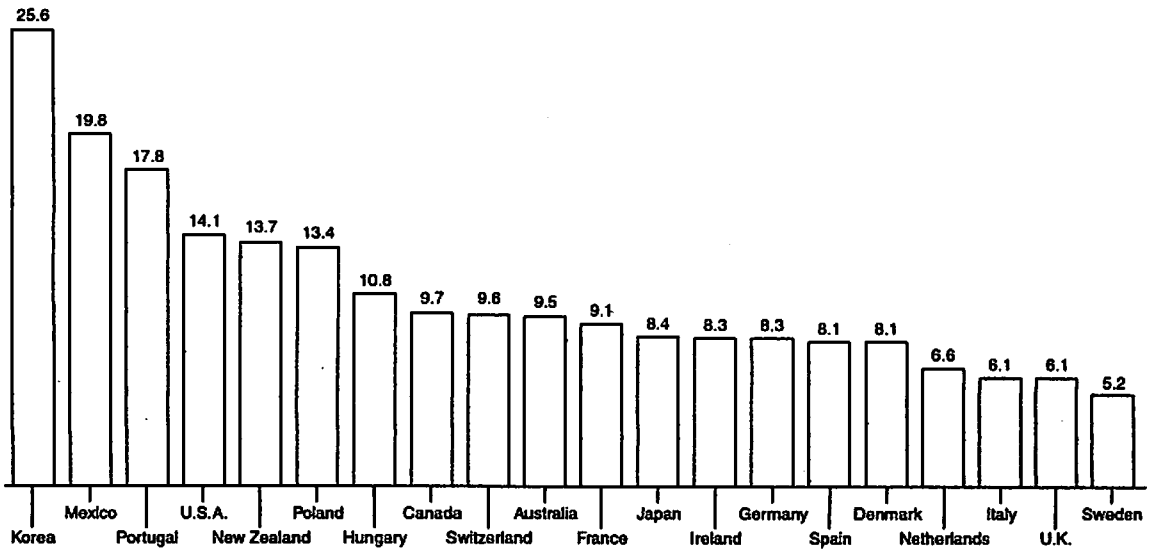
Chapter 2 Organizing Data

Section 2.1

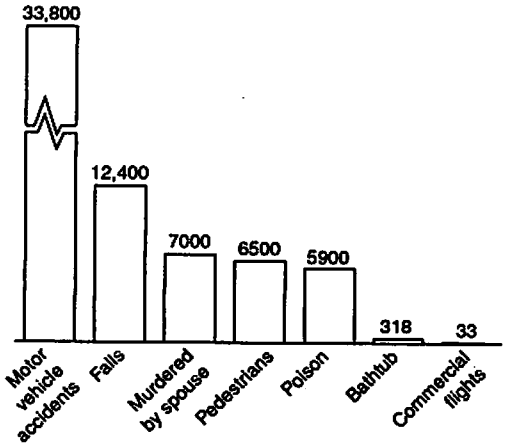
1. Highest Level of Education and Average Annual Household Income (in thousands of dollars)



2. Annual Number of Deaths from Injuries per 100,000 Children (Ages 1 to 14)

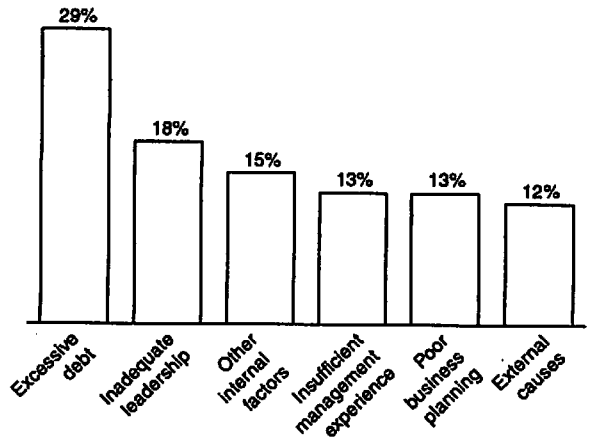


3. Number of People Who Died in a Calendar Year from Listed Causes—Pareto Chart



4. (a) Since 88% of those surveyed cited internal problems, 100% – 88% = 12% cited external factors as the leading cause of business failure. Among the internal causes, 88% – 13% – 13% – 18% – 29% = 15% must have listed various other internal factors for the leading cause of business failure.

Causes for Business Failure—Pareto Chart



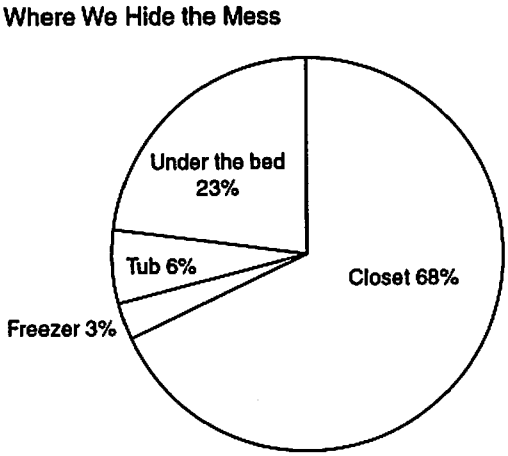
(b) As shown in part (a), 15% of those interviewed cited other internal factors as the leading cause of business failure. Excessive debt was the most commonly cited (internal and overall) cause for business failure.

Cause of Business Failure	Percentage	Frequency
Insufficient management experiences	13%	13% × 1300 = 169
Poor business planning	13%	13% × 1300 = 169
Inadequate leadership	18%	18% × 1300 = 234
Excessive debt	29%	29% × 1300 = 377
Other internal factors	15%	15% × 1300 = 195
External factors	12%	12% × 1300 = 156
Total	100%	1300

5.

Hiding place	Percentage	Number of Degrees
In the closet	68%	$68\% \times 360° \approx 245°$
Under the bed	23%	$23\% \times 360° \approx 83°$
In the bathtub	6%	$6\% \times 360° \approx 22°$
In the freezer	3%	$3\% \times 360° \approx 11°$
Total	100%	361°*

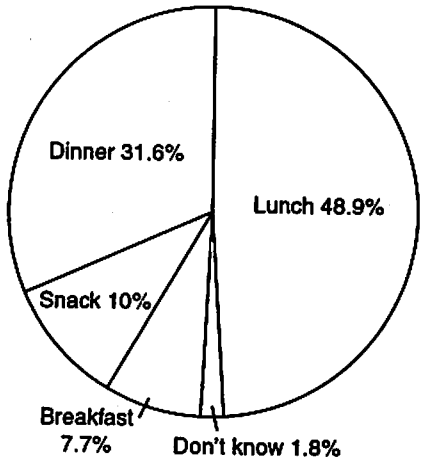
*Total does not add to 360° due to rounding.



6.

Meal	Percentage	Number of Degrees
Lunch	48.9%	$48.9\% \times 360° \approx 176°$
Breakfast	7.7%	$7.7\% \times 360° \approx 28°$
Dinner	31.6%	$31.6\% \times 360° \approx 114°$
Snack	10.0%	$10.0\% \times 360° = 36°$
Don't know	1.8%	$1.8\% \times 360° \approx 6°$
Total	100.0%	360°

Meals We Are Most Likely to Eat in a Fast-Food Restaurant

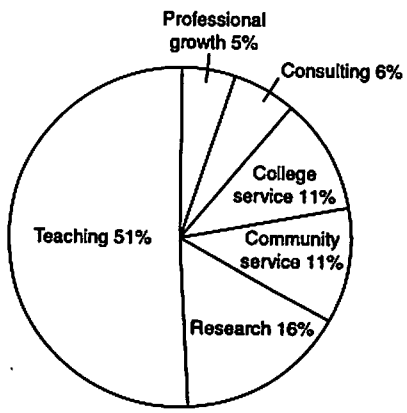


7.

Professional Activity	Percentage	Number of Degrees
Teaching	51%	$51\% \times 360^\circ \approx 184^\circ$
Research	16%	$16\% \times 360^\circ \approx 58^\circ$
Professional growth	5%	$5\% \times 360^\circ = 18^\circ$
Community service	11%	$11\% \times 360^\circ \approx 40^\circ$
Service to the college	11%	$11\% \times 360^\circ \approx 40^\circ$
Consulting outside the college	6%	$6\% \times 360^\circ \approx 22^\circ$
Total	100%	362°*

* Total does not add to 360° due to rounding.

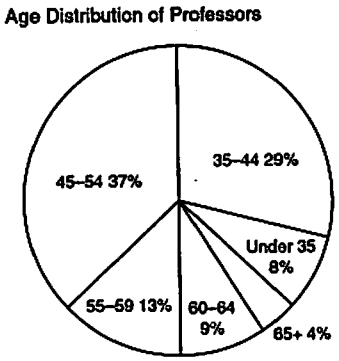
How College Professors Spend Time



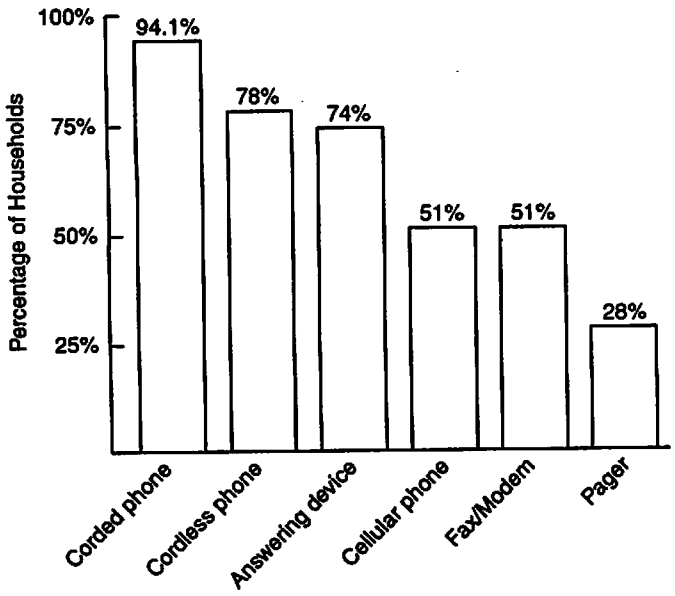
8.

Age	Percentage	Number of Degrees
Under 35 years	8%	$8\% \times 360° \approx 29°$
35–44 years	29%	$29\% \times 360° \approx 104°$
45–54 years	37%	$37\% \times 360° \approx 133°$
55–59 years	13%	$13\% \times 360° \approx 47°$
60–64 years	9%	$9\% \times 360° \approx 32°$
65 years and over	4%	$4\% \times 360° \approx 14°$
Total	100%	359°*

*Total does not add to 360° due to rounding.

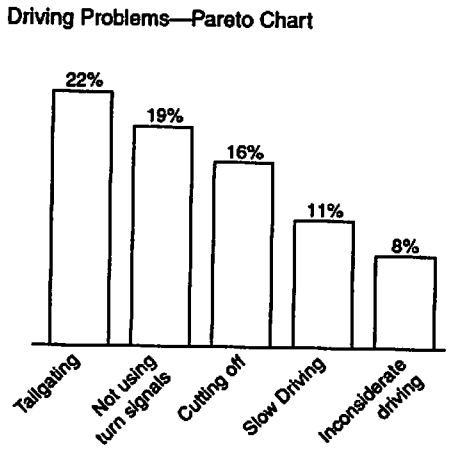


9. Percentage of Households with Telephone Gadgets



No. Since household can report having more than one telephone gadget, the percentages will not add to 100%.

10. The following Pareto Chart shows the percentage of drivers for each stated complaint.

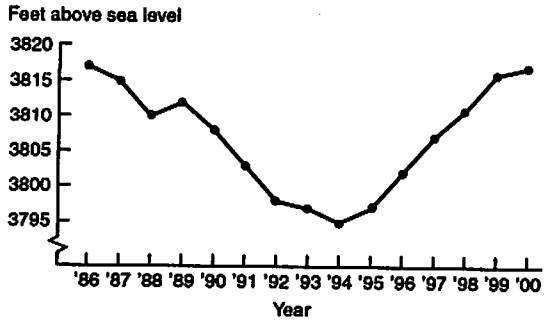


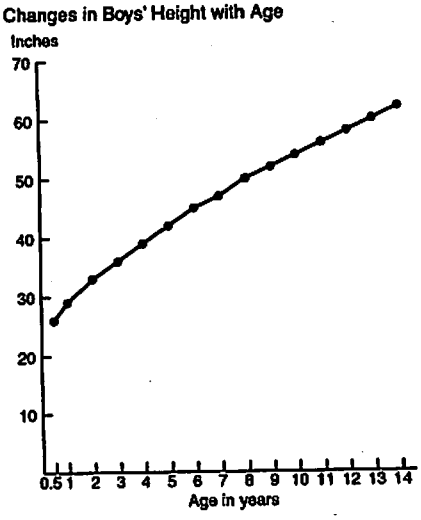
By subtraction, 100% – 22% – 19% – 16% – 11% – 8% = 24% of the respondents citied other bad habits.

Bad Habit	Percentage	Frequency
Tailgating	22%	22% × 500 = 110
Not using turn signals	19%	19% × 500 = 95
Cutting off other drivers	16%	16% × 500 = 80
Driving too slowly	11%	11% × 500 = 55
Being inconsiderate	8%	8% × 500 = 40
Other	24%	24% × 500 = 120
Total	100%	500

As reported, the percentages add to 76%, not the 100% needed for a circle graph. However, if there was only one response per person, knowing that the company surveyed 500 drivers tells us that 120 drivers, or 24%, had other bad driving complaints. Using this fact, a circle graph could be used.

11. Elevation of Pyramid Lake Surface—Time Plot



12. 

Section 2.2

1. (a) largest data value = 360
smallest data value = 236
number of classes specified = 5
class width $= \frac{360-236}{5} = 24.8$, increased to next whole number, 25

(b) The lower class limit of the first class in the smallest value, 236.
The lower class limit of the next class is the previous class's lower class limit plus the class width; for the second class, this is 236 + 25 = 261.
The upper class limit is one value less than lower class limit of the next class; for the first class, the upper class limit is 261 – 1 = 260.
The class boundaries are the halfway points between (i.e., the average of) the (adjacent) upper class limit of one class and the lower class limit of the next class. The lower class boundary of the first class is the lower class limit minus one-half unit. The upper class boundary for the last class is the upper class limit plus one-half unit. For the first class, the class boundaries are $236 - \frac{1}{2} = 235.5$ and $\frac{260+261}{2} = 260.5$. For the last class, the class boundaries are $\frac{335+336}{2} = 335.5$ and $360 + \frac{1}{2} = 360.5$.
The class mark or midpoint is the average of the class limits for that class. For the first class, the midpoint is $\frac{236+260}{2} = 248$.
The class frequency is the number of data values that belong to that class; call this value f.
The relative frequency of a class is the class frequency, f, divided by the total number of data values, i.e., the overall sample size, n.
For the first class, $f = 4$, $n = 57$, and the relative frequency is $f/n = \frac{4}{57} \approx 0.07$.

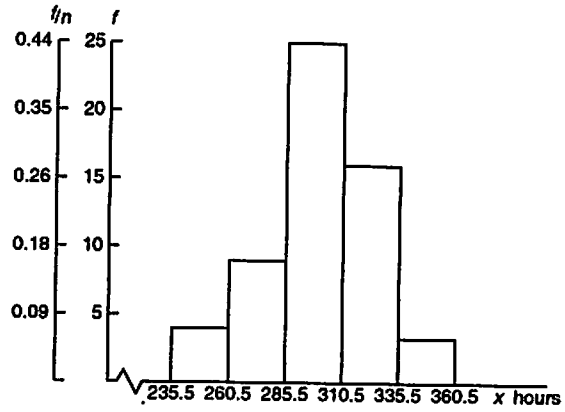
Class Limits	Boundaries	Midpoint	Frequency	Relative Frequency
236–260	235.5–260.5	248	4	0.07
261–285	260.5–285.5	273	9	0.16
286–310	285.5–310.5	298	25	0.44
311–335	310.5–335.5	323	16	0.28
336–360	335.5–360.5	348	3	0.05

(c) The histogram plots the class frequencies on the y-axis and the class boundaries on the x-axis. Since adjacent classes share boundary values, the bars touch each other. [Alternatively, the bars may be centered over the class marks (midpoints).

(d) The relative frequency histogram is exactly the same shape as the frequency histogram, but the vertical scale is relative frequency, f/n, instead of actual frequency, f.

The following figure shows the histogram and relative-frequency histogram for (c), and (d). Note that two vertical scales are shown.

Hours to Complete the Iditarod
Histogram and Relative-Frequency Histogram



2. (a) largest data value = 65

smallest data value = 20

number of classes specified = 5

class width $= \dfrac{65-20}{5} = 9$, increased to next whole number, 10

(b) The lower class limit of the first class in the smallest value, 20.

The lower class limit of the next class is the previous class's lower class limit plus the class width; for the second class, this is 20 + 10 = 30.

The upper class limit is one value less than lower class limit of the next class; for the first class, the upper class limit is 30 – 1 = 29.

The class boundaries are the halfway points between (i.e., the average of) the (adjacent) upper class limit of one class and the lower class limit of the next class. The lower class boundary of the first class is the lower class limit minus one-half unit. The upper class boundary for the last class is the upper class limit plus one-half unit. For the first class, the class boundaries are $20-\frac{1}{2}=19.5$ and $\frac{29+30}{2}=29.5$. For the last class, the class boundaries are $\frac{59+60}{2}=59.5$ and $69+\frac{1}{2}=69.5$.

The class mark or midpoint is the average of the class limits for that class. For the first class, the midpoint is $\frac{20+29}{2}=24.5$.

The class frequency is the number of data values that belong to that class; call this value f.

The relative frequency of a class is the class frequency, f, divided by the total number of data values, i.e., the overall sample size, n.

For the first class, $f = 3$, $n = 35$, and the relative frequency is $f/n = 3/35 \approx 0.0857$.

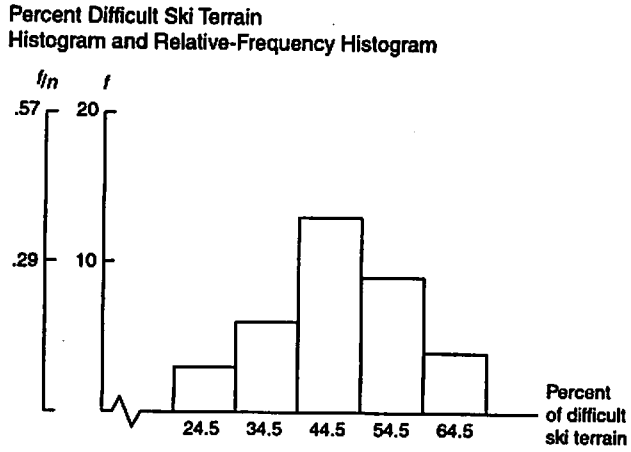
Percent Difficult Ski Terrain

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency
20–29	19.5–29.5	24.5	3	0.0857
30–39	29.5–39.5	34.5	6	0.1714
40–49	39.5–49.5	44.5	13	0.3714
50–59	49.5–59.5	54.5	9	0.2571
60–69	59.5–69.5	64.5	4	0.1143

(c) The histogram plots the class frequencies on the y-axis and the class boundaries on the x-axis. Since adjacent classes share boundary values, the bars touch each other. [Alternatively, the bars may be centered over the class marks (midpoints).]

(d) The relative frequency histogram is exactly the same shape as the frequency histogram, but the vertical scale is relative frequency, f/n, instead of actual frequency, f.

The following figure shows the histogram and relative-frequency histogram for (c) and (d). Note that two vertical scales are shown.



3. (a) largest data value = 53

smallest data value = 5

number of classes specified = 7

class width $= \frac{53-5}{7} \approx 6.86$, increased to next whole number, 7

(b) The lower class limit of the first class in the smallest value, 5.

The lower class limit of the next class is the previous class's lower class limit plus the class width; for the second class, this is 5 + 7 = 12.

The upper class limit is one value less than lower class limit of the next class; for the first class, the upper class limit is 12 – 1 = 11.

The class boundaries are the halfway points between (i.e., the average of) the (adjacent) upper class limit of one class and the lower class limit of the next class. The lower class boundary of the first class is the lower class limit minus one-half unit. The upper class boundary for the last class is the upper class limit plus one-half unit. For the first class, the class boundaries are $5-\frac{1}{2}=4.5$ and $\frac{11+12}{2}=11.5$. For the last class, the class boundaries are $\frac{46+47}{2}=46.5$ and $53+\frac{1}{2}=53.5$.

The class mark or midpoint is the average of the class limits for that class. For the first class, the midpoint is $\frac{5+11}{2}=8$.

The class frequency is the number of data values that belong to that class; call this value f.

The relative frequency of a class is the class frequency, f, divided by the total number of data values, i.e., the overall sample size, n.

For the first class, $f = 4$, $n = 50$, and the relative frequency is $f/n = 4/50 \approx 0.08$.

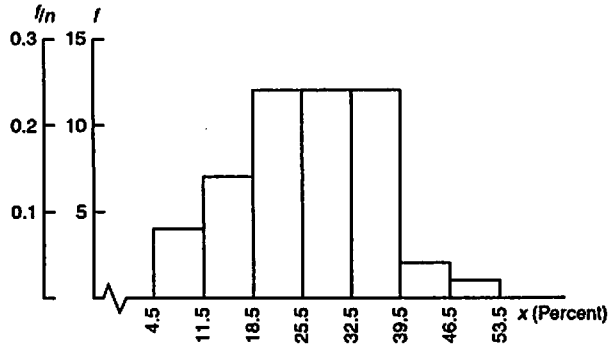
Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency
5–11	4.5–11.5	8	4	0.08
12–18	11.5–18.5	15	7	0.14
19–25	18.5–25.5	22	12	0.24
26–32	25.5–32.5	29	12	0.24
33–39	32.5–39.5	36	12	0.24
40–46	39.5–46.5	43	2	0.04
47–53	46.5–53.5	50	1	0.02

(c) The histogram plots the class frequencies on the *y*-axis and the class boundaries on the *x*-axis. Since adjacent classes share boundary values, the bars touch each other. [Alternatively, the bars may be centered over the class marks (midpoints).]

(d) The relative frequency histogram is exactly the same shape as the frequency histogram, but the vertical scale is relative frequency, f/n, instead of actual frequency, f.

The following figure shows the histogram and relative-frequency histogram for (c) and (d). Note that two vertical scales are shown.

Percentage of Children in Neighorhood
Histogram and Relative-Frequency Histogram



4. (a) largest data value = 75

smallest data value = 5

number of classes specified = 5

class width $= \dfrac{75-5}{5} = 14$, increased to next whole number, 15

(b) The lower class limit of the first class in the smallest value, 5.

The lower class limit of the next class is the previous class's lower class limit plus the class width; for the second class, this is 5 + 15 = 20.

The upper class limit is one value less than lower class limit of the next class; for the first class, the upper class limit is 20 – 1 = 19.

The class boundaries are the halfway points between (i.e., the average of) the (adjacent) upper class limit of one class and the lower class limit of the next class. The lower class boundary of the first class is the lower class limit minus one-half unit. The upper class boundary for the last class is the upper class limit plus one-half unit. For the first class, the class boundaries are $5-\frac{1}{2}=4.5$ and $\frac{19+20}{2}=19.5$. For the last class, the class boundaries are $\frac{64+65}{2}=64.5$ and $79+\frac{1}{2}=79.5$.

The class mark or midpoint is the average of the class limits for that class. For the first class, the midpoint is $\frac{5+19}{2}=12$.

The class frequency is the number of data values that belong to that class; call this value f.

The relative frequency of a class is the class frequency, f, divided by the total number of data values, i.e., the overall sample size, n.

For the first class, $f = 21$, $n = 63$, and the relative frequency is $f/n = 21/63 \approx 0.3333$.

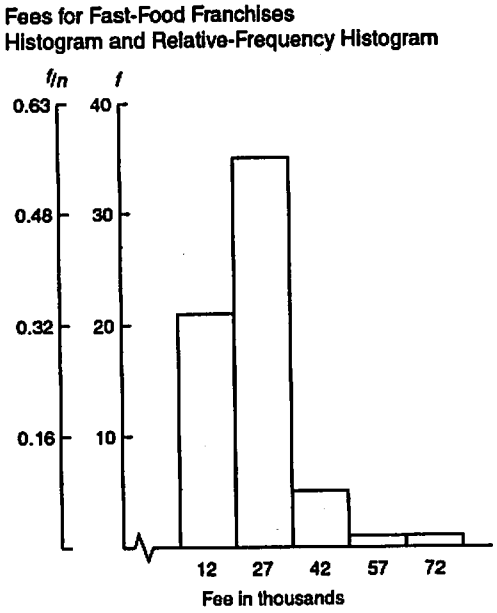
Fast-Food Franchise Fees (in thousands)

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency
5–19	4.5–19.5	12	21	0.3333
20–34	19.5–34.5	27	35	0.5556
35–49	34.5–49.5	42	5	0.0794
50–64	49.5–64.5	57	1	0.0159
65–79	64.5–79.5	72	1	0.0159

(c) The histogram plots the class frequencies on the y-axis and the class boundaries on the x-axis. Since adjacent classes share boundary values, the bars touch each other. [Alternatively, the bars may be centered over the class marks (midpoints).]

(d) The relative frequency histogram is exactly the same shape as the frequency histogram, but the vertical scale is relative frequency, f/n, instead of actual frequency, f.

The following figure shows the histogram and relative-frequency histogram for (c) and (d). Note that two vertical scales are shown.



5. (a) largest data value = 102

smallest data value = 18

number of classes specified = 5

class width $= \frac{102-18}{5} = 16.8$, increased to next whole number, 17

(b) The lower class limit of the first class in the smallest value, 18.

The lower class limit of the next class is the previous class's lower class limit plus the class width; for the second class, this is 18 + 17 = 35.

The upper class limit is one value less than lower class limit of the next class; for the first class, the upper class limit is 35 – 1 = 34.

The class boundaries are the halfway points between (i.e., the average of) the (adjacent) upper class limit of one class and the lower class limit of the next class. The lower class boundary of the first class is the lower class limit minus one-half unit. The upper class boundary for the last class is the upper class limit plus one-half unit. For the first class, the class boundaries are $18-\frac{1}{2}=17.5$ and $\frac{34+35}{2}=34.5$. For the last class, the class boundaries are $\frac{85+86}{2}=85.5$ and $102+\frac{1}{2}=102.5$.

The class mark or midpoint is the average of the class limits for that class. For the first class, the midpoint is $\frac{18+34}{2}=26$.

The class frequency is the number of data values that belong to that class; call this value f.

The relative frequency of a class is the class frequency, f, divided by the total number of data values, i.e., the overall sample size, n.

For the first class, $f = 1$, $n = 35$, and the relative frequency is $f/n = 1/35 \approx 0.03$.

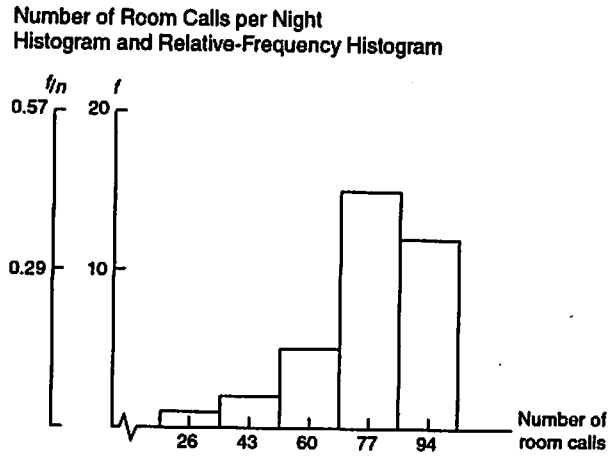
Number of Room Calls per Night

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency
18–34	17.5–34.5	26	1	0.03
35–51	34.5–51.5	43	2	0.06
52–68	51.5–68.5	60	5	0.14
69–85	68.5–85.5	77	15	0.43
86–102	85.5–102.5	94	12	0.34

(c) The histogram plots the class frequencies on the y-axis and the class boundaries on the x-axis. Since adjacent classes share boundary values, the bars touch each other. [Alternatively, the bars may be centered over the class marks (midpoints).]

(d) The relative frequency histogram is exactly the same shape as the frequency histogram, but the vertical scale is relative frequency, f/n, instead of actual frequency, f.

The following figure shows the histogram and relative-frequency histogram for (c) and (d). Note that two vertical scales are shown.



6. (a) largest data value = 43

smallest data value = 0

number of classes specified = 8

class width $= \dfrac{43-0}{8} = 5.375$, increased to next whole number, 6

(b) The lower class limit of the first class in the smallest value, 0.

The lower class limit of the next class is the previous class's lower class limit plus the class width; for the second class, this is $0 + 6 = 6$.

The upper class limit is one value less than lower class limit of the next class; for the first class, the upper class limit is $6 - 1 = 5$.

The class boundaries are the halfway points between (i.e., the average of) the (adjacent) upper class limit of one class and the lower class limit of the next class. The lower class boundary of the first class is the lower class limit minus one-half unit. The upper class boundary for the last class is the upper class limit plus one-half unit. For the first class, the class boundaries are $0-\frac{1}{2}=-0.5$ and $\frac{5+6}{2}=5.5$.

For the last class, the class boundaries are $\frac{41+42}{2}=41.5$ and $47+\frac{1}{2}=47.5$.

The class mark or midpoint is the average of the class limits for that class. For the first class, the midpoint is $\frac{0+5}{2}=2.5$.

The class frequency is the number of data values that belong to that class; call this value f.

The relative frequency of a class is the class frequency, f, divided by the total number of data values, i.e., the overall sample size, n.

For the first class, $f = 13$, $n = 55$, and the relative frequency is $f/n = 13/55 \approx 0.24$.

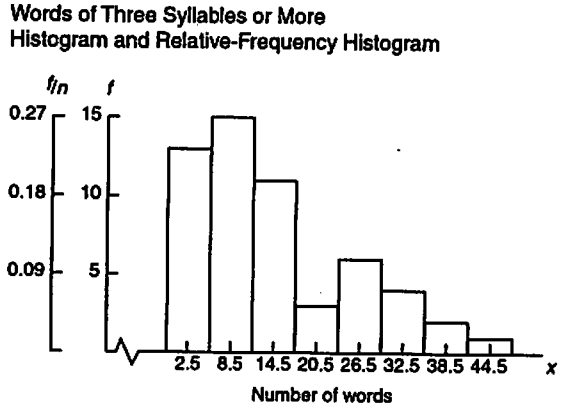
Words of Three Syllables or More

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency
0–5	0.5–5.5	2.5	13	0.24
6–11	5.5–11.5	8.5	15	0.27
12–17	11.5–17.5	14.5	11	0.20
18–23	17.5–23.5	20.5	3	0.05
24–29	23.5–29.5	26.5	6	0.11
30–35	29.5–35.5	32.5	4	0.07
36–41	35.5–41.5	38.5	2	0.04
42–47	41.5–47.5	44.5	1	0.02

(c) The histogram plots the class frequencies on the y-axis and the class boundaries on the x-axis. Since adjacent classes share boundary values, the bars touch each other. [Alternatively, the bars may be centered over the class marks (midpoints).]

(d) The relative frequency histogram is exactly the same shape as the frequency histogram, but the vertical scale is relative frequency, f/n, instead of actual frequency, f.

The following figure shows the histogram and relative-frequency histogram for (c) and (d). Note that two vertical scales are shown.



7. (a)

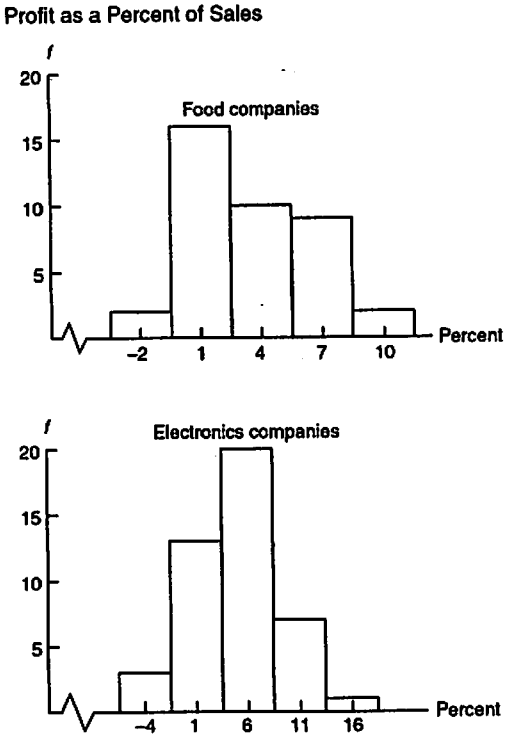
	Largest value	Smallest value	Class width
Food Companies	11	−3	$\frac{11-(-3)}{5} = 2.8$; use 3
Electronic Companies	16	−6	$\frac{16-(-6)}{5} = 4.4$; use 5

Profit as Percent of Sales—Food Companies

Class	Frequency	Midpoint
−3 to −1	2	−2
0–2	16	1
3–5	10	4
6–8	9	7
9–11	2	10

Profit as Percent of Sales—Electronic Companies

Class	Frequency	Midpoint
−6 to −2	3	−4
−1 to 3	13	1
4–8	20	6
9–13	7	11
14–18	1	16



(b) Because the classes and class widths are different for the two company types, it is difficult to compare profits as a percentage of sales. We can notice that for the electronic companies the 16 profits as a percentage of sales extends as high as 18, while for the food companies the highest profit as a percentage of sales is 11. On the other hand, some of the electronic companies also have greater losses than the food companies. Had we made the class limits the same for both company types and overlaid the histograms, it would be easier to compare the data.

8. (a)

	Largest value	Smallest value	Class width
Miami Dolphins	295	175	$\frac{295-175}{6} = 20$; use 21
San Diego Charges	310	119	$\frac{310-119}{6} \approx 31.8$; use 32

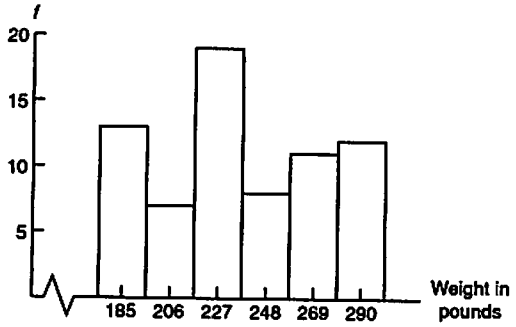
Weights of Football Players:
Miami Dolphins

Class	Midpoint	Frequency
175–195	185	13
196–216	206	7
217–237	227	19
238–258	248	8
259–279	269	11
280–300	290	12

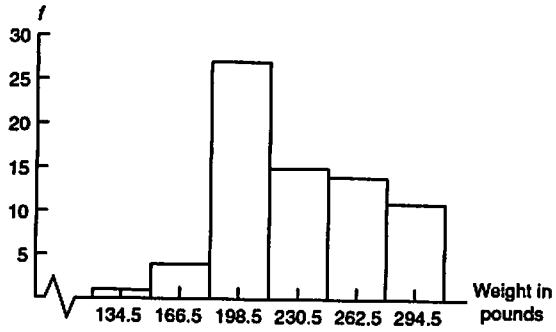
San Diego Chargers

Class	Midpoint	Frequency
119–150	134.5	1
151–182	166.5	4
183–214	198.5	27
215–246	230.5	15
247–278	262.5	14
279–310	294.5	11

Weights of Football Players—Miami Dolphins



Weights of Football Players—San Diego Chargers



(b) Because the class widths are different, it is difficult to compare the histograms. However, San Diego has 4 players who are smaller than the smallest Miami player, and 4 players who are larger than the largest player.

It would be easier to compare the teams' weights if the histograms had common classes and were overlaid.

9. (a) Uniform is rectangular, symmetric looks like mirror images on each side of the middle, bimodal has two modes (peaks), and skewed distributions have long tails on one side, and are skewed in the direction of the tail ("skew, few"). (Note that uniform distributions are also symmetric, but "uniform" is more descriptive.)

(a) skewed left; (b) uniform, (c) symmetric, (d) bimodal, (e) skewed right.

(b) Answers vary. Students would probably like (a) since there are many high scores and few low scores. Students would probably dislike (e) since there are few high scores but lots of lows scores. (b) is designed to give approximately the same number of As, Bs, etc. (d) has more Bs and Ds, say. (c) is the way many tests are designed: As and Fs for the exceptionally high and low scores with most students receiving Cs.

10. (a) Uniform is rectangular, symmetric looks like mirror images on each side of the middle, bimodal has two modes (peaks), and skewed distributions have long tails on one side, and are skewed in the direction of the tail ("skew, few"). (Note that uniform distributions are also symmetric, but "uniform" is more descriptive.)

(a) uniform, (b) skewed right, (c) bimodal, (d) bimodal, (e) symmetric. [Note that (c) has a major and a minor mode. "Tails" in a distribution's shape "tail off," i.e., get thinner, and do not have "bumps" in them as (c) does.]

(b) Answers vary. Ads should target the largest number of potential buyers, so ads should be aimed at the income levels with the greatest concentration (frequency) of households.

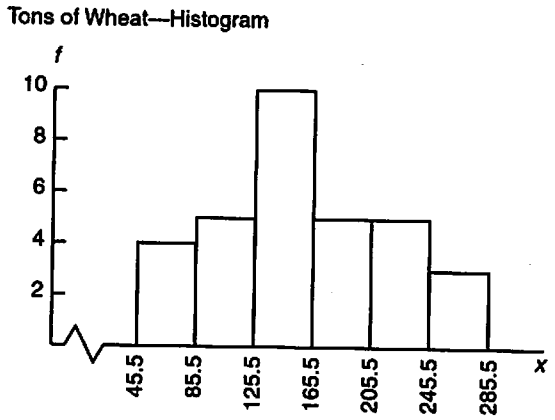
(c) Answers vary. Since warranty/registration cards are returned voluntarily, the income data are most likely not representative of the buying public in general, and probably are not even representative of those buying the specific product. Also, people tend to inflate their income levels on most forms, except those sent to the IRS.

11. (a) $2.71 \times 100 = 271$, $1.62 \times 100 = 162, \ldots, 0.70 \times 100 = 70$.

(b) largest value = 282, smallest value = 46

$$\text{class width} = \frac{282 - 46}{6} \approx 39.3; \text{ use } 40$$

Class Limits	Class Boundaries	Midpoint	Frequency
46–85	45.5–85.5	65.5	4
86–125	85.5–125.5	105.5	5
126–165	125.5–165.5	145.5	10
166–205	165.5–205.5	185.5	5
206–245	205.5–245.5	225.5	5
246–285	245.5–285.5	265.5	3



(c) class width is $\frac{40}{100} = 0.40$

Class Limits	Class Boundaries	Midpoint	Frequency
0.46–0.85	0.455–0.855	0. 655	4
0.86–1.25	0.855–1.255	1.055	5
1.26–1.65	1.255–1.655	1.455	10
1.66–2.05	1.655–2.055	1.855	5
2.06–2.45	2.055–2.455	2.255	5
2.46–2.85	2.455–2.855	2.655	3

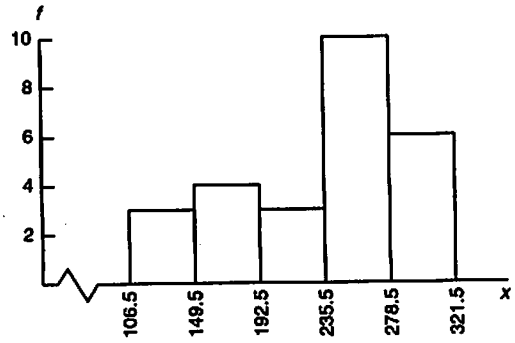
12. (a) $0.194\times1000 = 194, 0.258\times1000 = 258, \ldots, 0.200\times1000 = 200.$

(b) largest value = 317, smallest value = 107

class width = $\frac{317-107}{5} = 42$, use 43

Class Limits	Class Boundaries	Midpoint	Frequency
107–149	106.5–149.5	128	3
150–192	149.5–192.5	171	4
193–235	192.5–235.5	214	3
236–278	235.5–278.5	257	10
279–321	278.5–321.5	300	6

Baseball Batting Averages—Histogram



(c) class width = $\frac{43}{1000} = 0.043$

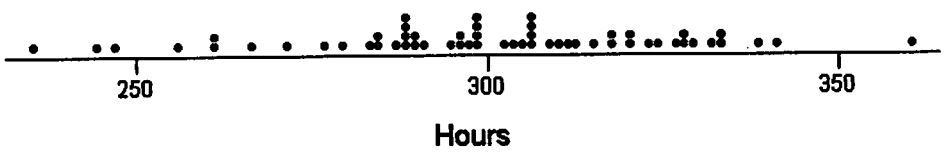
Class Limits	Class Boundaries	Midpoint	Frequency
0.107–0.149	0.1065–0.1495	0.128	3
0.150–0.192	0.1495–0.1925	0.171	4
0.193–0.235	0.1925–0.2355	0.214	3
0.236–0.278	0.2355–0.2785	0.257	10
0.279–0.321	0.2785–0.3215	0.300	6

13. **(a)** There is one dot below 600, so 1 state has 600 or fewer licensed drivers per 1000 residents.

(b) 5 values are close to 800; $\frac{5}{51} \approx 0.0980 \approx 9.8\%$

(c) 9 values below 650
37 values between 650 and 750
5 values above 750
From either the counts or the dotplot, the interval from 650 to 750 licensed drivers per 1000 residents has the most "states."

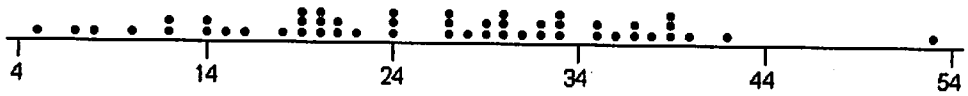
14. The dotplot shows some of the characteristics of the histogram such as more dot density from, say 280 to 340, corresponding roughly to the histogram bars of heights 25 and 16.
However, they are somewhat difficult to compare since the dotplot can be thought of as a histogram with one value, the class mark, i.e., the data value, per class.
Because the definitions of the classes and, therefore, the class widths, differ, it is difficult to compare the two figures.



15. The dotplot shows some of the characteristics of the histogram, such as the concentration of most of the data from, say, 20 to 40; this corresponds roughly to the 3 histogram bars of height 12. There are more data (dots) below 20 than above 40, which corresponds to the histogram bars of heights 4 and 7, and the bars of heights 2 and 1, respectively.

However, they are somewhat difficult to compare since the dotplot can be thought of as a histogram with one value, the class mark, i.e., the data value, per class.

Because the definitions of the classes and, therefore, the class widths, differ, it is difficult to compare the two figures.



Percentage

Section 2.3

1. (a) The smallest value is 47 and the largest is 97, so we need stems 4, 5, 6, 7, 8, and 9. Use the tens digit as the stem and the ones digit as the leaf.

Longevity of Cowboys

4	7 = 47 years
4	7
5	2 7 8 8
6	1 6 6 8 8
7	0 2 2 3 3 5 6 7
8	4 4 4 5 6 6 7 9
9	0 1 1 2 3 7

(b) Yes, certainly these cowboys lived long lives, as evidenced by the high frequency of leaves for stems 7, 8, and 9 (i.e., 70-, 80-, and 90-year olds).

2. The largest value is 91 (percent of wetlands lost) and the smallest value is 9 (percent), which is coded as 09. We need stems 0 to 9. Use the tens digit as the stem and the ones digit as the leaf. The percentages are concentrated from 20 to 50 percent. The distribution is asymmetrical but not skewed because of the "bump" in the 80s. If we smoothed the shape, we might consider this bimodal. There is a gap showing none of the lower 48 states has lost from 10 to 19% of its wetlands.

Percent of Wetlands Lost

4	0 = 40%
0	9
1	
2	0 3 4 7 7 8
3	0 1 3 5 5 5 6 7 8 8 9
4	2 2 6 6 6 8 9 9
5	0 0 0 2 2 4 6 6 9 9
6	0 7
7	2 3 4
8	1 5 7 7 9
9	0 1

3. The longest average length of stay is 11.1 days in North Dakota and the shortest is 5.2 days in Utah. We need stems from 5 to 11. Use the digit(s) to the left of the decimal point as the stem, and the digit to the right as the leaf.

Average Length of Hospital Stay

5	2 = 5.2 days
5	2 3 5 5 6 7
6	0 2 4 6 6 7 7 8 8 8 8 9 9
7	0 0 0 0 0 0 1 1 1 2 2 2 3 3 3 3 4 4 5 5 6 6 8
8	4 5 7
9	4 6 9
10	0 3
11	1

The distribution is skewed right.

4. Number of Hospitals per State

0	8 = 8 hospitals		
0	8	15	
1	1 2 5 6 9	16	2
2	1 7 7	17	5
3	5 7 8	18	
4	1 2 7	19	3
5	1 2 3 9	20	9
6	1 6 8	21	
7	1	22	7
8	8	23	1 6
9	0 2 6 8		
10	1 2 7	42	1
11	3 3 7 9	43	
12	2 3 9	44	0
13	3 3 6		
14	8		

Texas and California have the highest number of hospitals, 421 and 440, respectively. Both states have large populations and large areas. The four largest states by area are Alaska, Texas, California, and Montana; however, both Alaska and Montana have small populations, but the population tends to cluster at their largest cities, thus reducing the number of hospitals needed.

5. (a) The longest time during 1961–1980 is 23 minutes (i.e., 2:23) and the shortest time is 9 minutes (2:09). We need stems 0, 1, and 2, which we'll write as 0*, 0˙, 1*, 1˙, 2*, and 2˙. (We can eliminate 0* since no time was 2:04 or less and 2˙ because no winning time was 2:25 or more. We'll use the tens digit as the stem and the ones digit as the leaf, placing leaves 0, 1, 2, 3, and 4 on the "* stem" and leaves 5, 6, 7, 8, and 9 on the "˙ stem."

Minutes Beyond 2 Hours (1961-1980)

0	9 = 9 minutes past 2 hours
0•	9 9
1*	0 0 2 3 3
1•	5 5 6 6 7 8 8 9
2*	0 2 3 3

(b) The longest time during the period 1981–2000 was 14 (2:14), and the shortest was 7 (2:07), so we'll need stems 0• and 1* only.

Minutes Beyond 2 Hours (1981-2000)

0	7 = 7 minutes past 2 hours
0•	7 7 7 8 8 8 8 9 9 9 9 9 9 9 9
1*	0 0 1 1 4

(c) In more recent times, the winning times have been closer to 2 hours, with all 20 times between 7 and 14 minutes over two hours. In the earlier period, more than half the times (12 or 20) were more than 2 hours and 14 minutes.

6. (a) The largest (worst) score in the first round was 75; the smallest (best) score was 65. We need stems 6• and both 7* and 7•; leaves 0 to 4 go on the "* stem" and leaves 5–9 belong on the "• stem."

First Round Scores

6	5 = score of 65
6•	5 6 7 7
7*	0 1 1 1 1 1 1 1 1 1 1 2 2 2 3 3 3 3 4 4 4
7•	5 5 5 5 5 5 5

(b) the largest score in the fourth round was 74 and the smallest was 68. Here we need stems 6• and 7*, we don't need 7• because no scores were over 74.

Fourth Round Scores

6	8 = score of 68
6•	8 9 9 9 9 9
7*	0 0 0 0 1 1 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 4 4 4

(c) Scores are lower in the fourth round. In the first round both the low and high scores were more extreme than in the fourth round.

7. The largest value in the data is 29.8 mg. Of tar per cigarette smoked, and the smallest value is 1.0. We will need stems from 1 to 29, and we will use the numbers to the right of the decimal point as the leaves.

Milligrams of Tar per Cigarette

1	0 = 1.0 mg tar
1	0
2	
3	
4	1 5
5	
6	
7	3 8
8	0 6 8
9	0
10	
11	4
12	0 4 8
13	7
14	1 5 9
15	0 1 2 8
16	0 6
17	0
29	8

8. The largest value in the data set is 23.5 mg Carbon monoxide per cigarette smoked, and the smallest is 1.5. We need stems from 1 to 23, and we'll use the numbers to the right of the decimal point as leaves.

Milligrams of Carbon Monoxide

1	5 = 1.5 mg CO
1	5
2	
3	
4	9
5	4
6	
7	
8	5
9	0 5
10	0 2 2 6
11	
12	3 6
13	0 6 9
14	4 9
15	0 4 9
16	3 6
17	5
18	5
23	5

9. The largest value in the data set is 2.03 mg nicotine per cigarette smoked. The smallest value is 0.13. We will need stems 0*, 0•, 1*, 1•, and 2*. Leaves 0 to 4 belong on the * stems and leaves 5 to 9 belong on the • stems. We will use the number to the left of the decimal point as the stem and the first number to the right of the decimal point as the leaf. The number 2 places to the right of the decimal point (the hundredths digit) will be truncated (chopped off; not rounded off).

Milligrams of Nicotine per Cigarette

0	1 = 0.1 milligram
0*	1 4 4
0•	5 6 6 6 7 7 7 8 8 9 9 9
1*	0 0 0 0 0 0 0 1 2
1•	
2*	0

10. **(a)** For Site I, read the values in Figure 2-27 from the center (stem) to the left to find the least depth is 25 cm and the greatest depth is 110 cm. For Site II, read the values from the center (stem) to the right to find the least depth is 20 cm and the greatest depth is 125 cm.

(b) The Site I depth distribution is, smoothed out, fairly symmetrical around approximately 70 cm. Site II, however, is fairly uniform in shape except that it has a huge gap with no artifacts from about 70 to 100 cm.

(c) It would appear that Site II was probably unoccupied during the time period associated with 70 cm to 100 cm.

11. **(a)** Average salaries in California range from \$49,000 to \$126,000. Salaries in New York range from \$45,000 to \$120,000.

(b) New York has a greater number of average salaries in the \$60,000 than California, but California has more average salaries than New York in the \$70,000 range.

(c) The California data appear to be similar in shape to the New York data, but California's distribution has been shifted up approximately \$10,000. It is also heavier in the upper tail and shows no gap in average salaries, unlike New York which has no salaries in the \$110,000 range. California has higher average salaries.

Chapter 2 Review

1. Figure 2-1 (a) (in the text) is essentially a bar graph with a "horizontal" axis showing years and a "vertical" axis showing miles per gallon. However, in depicting the data as a highway and showing it in perspective, the ability to correctly compare bar heights visually has been lost. For example, determining what would appear to be the bar heights by measuring from the white line on the road to the edge of the road along a line drawn from the year to its mpg value, we get the bar height for 1983 to be approximately 7/8 inch and the bar height for 1985 to be approximately 1 3/8 inches (i.e., 11/8 inches). Taking the ratio of the given bar heights, we see that the bar for 1985 should be $\frac{27.5}{26} \approx 1.06$ times the length of the 1983 bar. However, the measurements show a ratio of $\frac{\frac{11}{8}}{\frac{7}{8}} = \frac{11}{7} \approx 1.60$, i.e., the 1985 bar is (visually) 1.6 times the length of the 1983 bar. Also, the years are evenly spaced numerically, but the figure shows the more recent years to be more widely spaced due to the use of perspective.

Figure 2-1(b) is a time plot, showing the years on the x-axis and miles per gallon on the y-axis. Everything is to scale and not distorted visually by the use of perspective. It is easy to see the mpg standards for each year, and you can also see how fuel economy standards for new cars have changed over the eight years shown (i.e., a steep increase in the early years and a leveling off in the later years).

2. **(a)** By reading the y-coordinate of the dot associated with the year, we estimate the 1980 prison population at approximately 140 prisoners per 100,000, and the 1997 population at approximately 440 prisoners per 100,000 people.

   (b) The number of inmates per 100,000 increased.

   (c) The population 266,574,000 is $2{,}665.74 \times 100{,}000$, and 444 per 100,000 is $\frac{444}{100{,}000}$.

   So $\frac{444}{100{,}000} \times (2{,}665.74 \times 100{,}000) \approx 1{,}183{,}589$ prisoners.

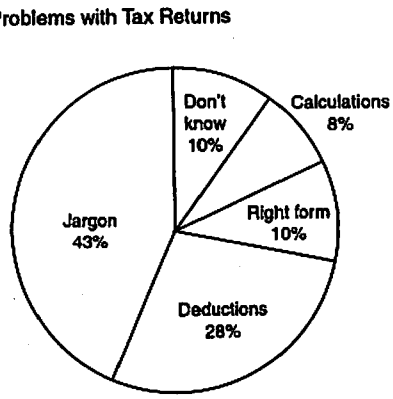
   The projected 2020 population is 323,724,000, or $3{,}237.24 \times 100{,}000$.

   So $\frac{444}{100{,}000} \times (3{,}237.24 \times 100{,}000) \approx 1{,}437{,}335$ prisoners.

3.

Most Difficult Task	Percentage	Degrees
IRS jargon	43%	$0.43 \times 360° \approx 155°$
Deductions	28%	$0.28 \times 360° \approx 101°$
Right form	10%	$0.10 \times 360° = 36°$
Calculations	8%	$0.08 \times 360° \approx 29°$
Don't know	10%	$0.10 \times 360° = 36°$

Note: Degrees do not total 360° due to rounding.



4. (a) Since the ages are two digit numbers, use the tens digit as the stem and the ones digit as the leaf.

Age of DUI Arrests

1	6 = 16 years
1	6 8
2	0 1 1 2 2 2 3 4 4 5 6 6 6 7 7 7 9
3	0 0 1 1 2 3 4 4 5 5 6 7 8 9
4	0 0 1 3 5 6 7 7 9 9
5	1 3 5 6 8
6	3 4

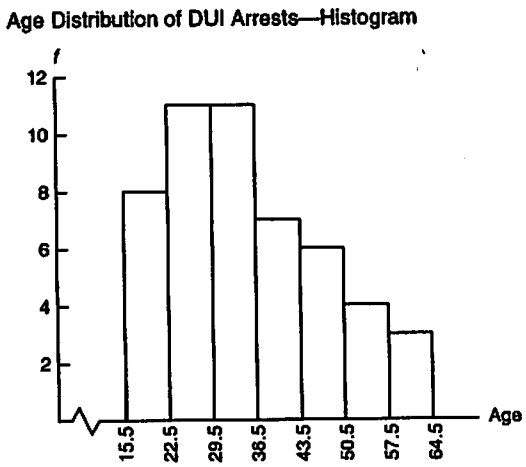
(b) The largest age is 64 and the smallest is 16, so the class with for 7 classes is $\frac{64-16}{7} \approx 6.86$; use 7. The lower class limit for the first class is 16; the lower class limit for the second class is 16 + 7 = 23. The total number of data points is 50, so calculate the relative frequency by dividing the class frequency by 50.

Age Distribution of DUI Arrests

Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency	Cumulative Frequency
16–22	15.5–22.5	19	8	0.16	8
23–29	22.5–29.5	26	11	0.22	19
30–36	29.5–36.5	33	11	0.22	30
37–43	36.5–43.5	40	7	0.14	37
44–50	43.5–50.5	47	6	0.12	43
51–57	50.5–57.5	54	4	0.08	47
58–64	57.5–64.5	61	3	0.06	50

The class boundaries are the average of the upper class limit of the next class. The midpoint is the average of the class limits for that class.

(c) The class boundaries are shown in (b).



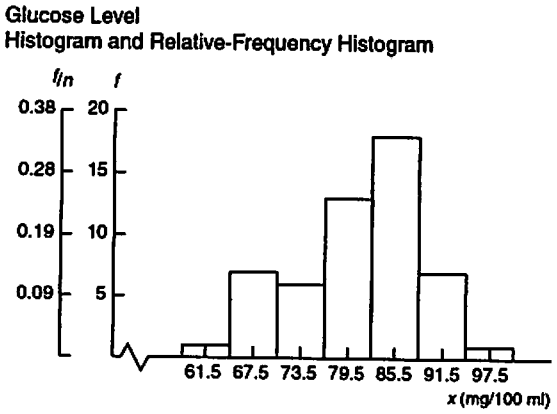
5. (a) The largest value is 96 mg of glucose per 100 ml of blood, and the smallest value is 59. For 7 classes we need a class width of $\frac{96-59}{7} \approx 5.3$; use 6. The lower class limit of the first class is 59, and the lower class limit of the second class is 59 + 6 = 65.

The class boundaries are the average of the upper class limit of one class and the lower class limit of the next higher class. The midpoint is the average of the class limits for that class. There are 53 data values total so the relative frequency is the class frequency divided by 53.

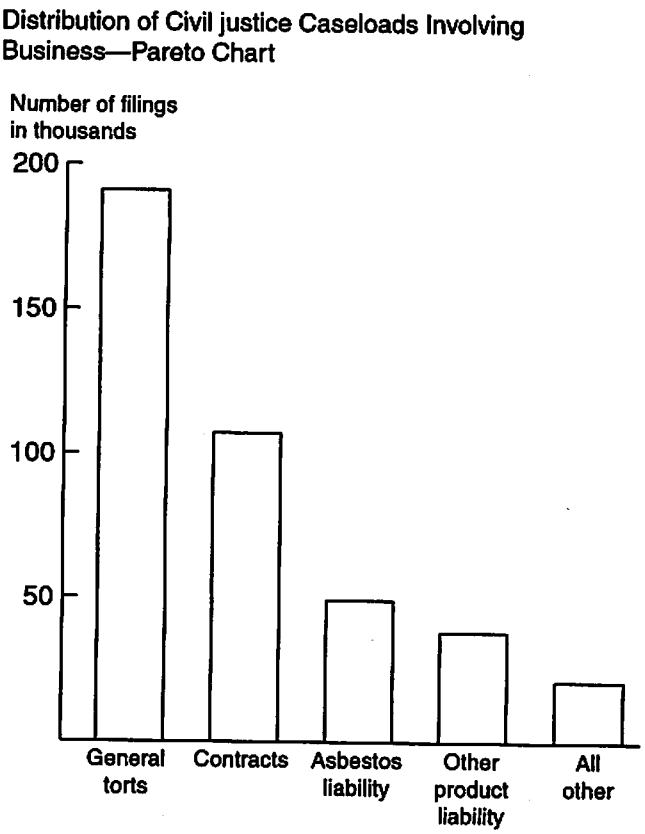
Class Limits	Class Boundaries	Midpoint	Frequency	Relative Frequency
59–64	58.5–64.5	61.5	1	0.02
65–70	64.5–70.5	67.5	7	0.13
71–76	70.5–76.5	73.5	6	0.11
77–82	76.5–82.5	79.5	13	0.25
83–88	82.5–88.5	85.5	18	0.34
89–94	88.5–94.5	91.5	7	0.13
95–100	94.5–100.5	97.5	1	0.02

(b) The histogram shows the bars centered over the midpoints of each class.

(c) The frequency histogram and the relative frequency histogram are the same except in the latter, the vertical scale is relative frequency, not frequency.



6. (a) A pareto chart is similar to a bar chart, except the bars are in decreasing order by frequency.



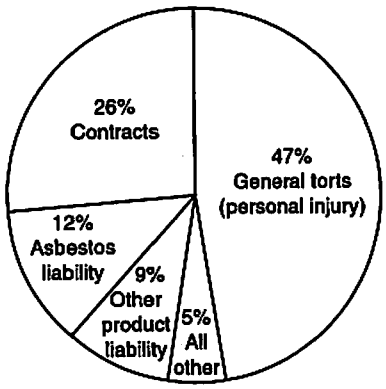
The general torts (personal injury) lawsuits occur with the greatest frequency.

(b) The total number of filings shown is 406 (thousand).

Case Type	Percentage	Degrees
Contracts	107/406 ≈ 26%	0.26 × 360° ≈ 94°
General torts	191/406 ≈ 47%	0.47 × 360° ≈ 169°
Asbestos liability	49/406 ≈ 12%	0.12 × 360° ≈ 43°
Other product liability	38/406 ≈ 9%	0.09 × 360° ≈ 32°
All other	21/406 ≈ 5%	0.05 × 360° = 18°

Note: Percentages do not add to 100% due to rounding. Similarly, the degrees do not add to 360° due to rounding.

Distribution of Civil justice Caseloads Involving Business—Pie Chart



7. (a) To determine the decade which contained the most samples, count both rows (if shown) of leaves; recall leaves 0–4 belong on the first line and 5–9 belong on the second line when two lines per stem are used. The greatest number of leaves is found on stem 124, i.e., the 1240s (the 40s decade in the 1200s), with 40 samples.

(b) The number of samples with tree ring dates 1200 A.D. to 1239 A.D. is 28 + 3 + 19 + 25 = 75.

(c) The dates of the longest interval with no sample values are 1204 through 1211 A.D. This might mean that for these eight years, the pueblo was unoccupied (thus no new or repaired structures) or that the population remained stable (no new structures needed) or that, say, weather conditions were favorable these years so existing structures didn't need repair. If relatively few new structures were built or repaired during this period, their tree rings might have been missed during sample selection.

8. (a) It has a long tail on the left, so it is skewed left.

(b) The class width is the difference between any two adjacent midpoints. Here, for example, the class width is 4 – 3.5 = 0.5 grade points. The average of any two adjacent midpoints is the boundary value between the two midpoints classes*. So, for midpoints 1 and 1.5, the boundary value is $1+\frac{1.5}{26}=1.25$. The difference between any two adjacent boundary values is also the class width, so the other class boundary values within the histogram are 1.25 + 0.5 = 1.75, 1.75 + 0.5 = 2.25, 2.25 + 0.5 = 2.75, 2.75 + 0.5 = 3.25, and 3.25 + 0.5 = 3.75; 3.75 is the lower class boundary for the class, so its upper class boundary is 3.75 + 0.5 = 4.25. Similarly, the upper class boundary of the first class was 1.25, so its lower class boundary is 1.25 – 0.5 = 0.75. The class boundaries are, therefore, 0.75, 1.25, 1.75, 2.25, 2.75, 3.25, 3.75 and 4.25 (from left to right).

*Recall that the average of a and b is $\frac{a+b}{2}$ which is also the value halfway between a and b.

(c) The relative frequencies are f/n, so if we multiply this decimal value by 100, we have the relative frequency expressed as a percent. The relative frequencies, expressed as percents, are 1%, 1%, 2%, 8%, 17%, 27%, and 44%, from left to right. The GPA of 3.25 is a boundary value, so to find the percentage of college graduates who had high school GPAs less than 3.25 is the sum of the relative frequency percentages for bars at or below 3.25: 1% + 1% + 2% + 8% + 17% = 29%. A high school GPA of 3.75 is the next boundary value above 3.25, so if we take the percentage of students with GPAs less 3.25 (29%), and add the percentage of students with GPAs between 3.25 and 3.75 (27%), we find 29% + 27% = 56% of college graduates had high school GPAs of less than 3.75. (Recall that, technically, boundary values are not values the data can take on. They are values between the upper class limit of one class and the lower class limit of the next class, and the class limits specify the largest and smallest data values, respectively, that can be put in those classes. Traditionally, the boundary values are specified to one more decimal place than the data, and that is the case here: the data are reported to one decimal place, but the boundaries are reported to two decimal places.)

Class Midpoints	Class Boundaries	Relative Frequency	Relative Frequency
1	0.75–1.25	0.01	1%
1.5	1.25–1.75	0.01	1%
2	1.75–2.25	0.02	2%
2.5	2.25–2.75	0.08	8%
3	2.75–3.25	0.17	17%
3.5	3.25–3.75	0.27	27%
4	3.75–4.25	0.44	44%

Chapter 3 Averages and Variation

Section 3.1

1. Mean $= \bar{x} = \dfrac{\Sigma x}{n} = \dfrac{156+161+152+\cdots+157}{12}$

$= \dfrac{1876}{12}$

$= 156.33$

The mean is 156.33.

Organize the data from smallest to largest.

144	148	152	153	156	157
157	157	161	161	162	168

To find the median, add the two middle values and divide by 2 since there is an even number of values.

$$\text{Median} = \frac{157+157}{2} = 157$$

The median is 157.

The mode is 157 because it is the value that occurs most frequently.

A gardener in Colorado should look at seed and plant descriptions to determine if the plant can thrive and mature in the designated number of frost-free days. The mean, median, and mode are all close. About half the locations have 157 or fewer frost-free days.

2. Mean $= \bar{x} = \dfrac{\Sigma x}{n} = \dfrac{11+29+54+\cdots+46}{12}$

$= \dfrac{542}{12}$

$= 45.17$

The mean is 45.17.

Organize the data from smallest to largest.

11	29	41	46	46	46
47	49	54	54	59	60

To find the median, add the two middle values and divide by 2 since there is an even number of values.

$$\text{Median} = \frac{46+47}{2} = 46.5$$

The median is 46.5.

The mode is 46 because it is the value that occurs most frequently.

3. $\text{Mean} = \bar{x} = \dfrac{\Sigma x}{n} = \dfrac{146+152+168+\cdots+144}{14}$

$= \dfrac{2342}{14}$

$= 167.3$

The mean is 167.3°F.

Organize the data from smallest to largest.

144 146 152 152 165 168 168
174 178 178 178 179 180 180

To find the median, add the two middle values and divide by 2 since there is an even number of values.

$$\text{Median} = \frac{168+174}{2} = 171$$

The median is 171° F.

The mode is 178° F because it is the value that occurs most frequently.

4. (a) $\text{Mean} = \bar{x} = \dfrac{\Sigma x}{n} = \dfrac{2723}{20} = 136.15$

The mean is \$136.15.

The median is \$66.50.

The mode is \$60.

(b) 5% of 20 is 1. Eliminate one data value from the bottom and one from the top of the ordered data. In this case eliminate \$40 and \$500.

$$\text{Mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{2183}{18} \approx 121.28$$

The 5% trimmed mean is \$121.28.

Yes, the trimmed mean more accurately reflects the general level of the daily rental cost, but is still higher than the median.

(c) Median. The low and high prices would be helpful also.

5. First organize the data from smallest to largest. Then compute the mean, median, and mode.

(a) Upper Canyon

1	1	1	2	3	3	3	3	4	6	9

$\text{Mean} = \bar{x} = \dfrac{\Sigma x}{n} = \dfrac{36}{11} \approx 3.27$

Median = 3 (middle value)

Mode = 3 (occurs most frequently)

(b) Lower Canyon

0	0	1	1	1	1	2	2	3	6	7	8	13	14

$$\text{Mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{59}{14} \approx 4.21$$

$$\text{Median} = \frac{2+2}{2} = 2$$

Mode = 1 (occurs most frequently)

(c) The mean for the Lower Canyon is greater than that of the Upper Canyon. However, the median and mode for the Lower Canyon are less than those of the Upper Canyon.

(d) 5% of 14 is 0.7 which rounds to 1. So, eliminate one data value from the bottom of the list and one from the top. Then compute the mean of the remaining 12 values.

$$5\% \text{ trimmed mean} = \frac{\Sigma x}{n} = \frac{45}{12} = 3.75$$

Now this value is closer to the Upper Canyon mean.

6. (a) First arrange the data from smallest to largest. Then compute the mean, median, and mode..

$$\text{Mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{1050}{40} \approx 26.3$$

The mean is 26.3 yr.

$$\text{Median} = \frac{25+26}{2} = 25.5$$

The median is 25.5 yr.

Mode = 25

The mode is 25 yr.

(b) The median may represent the age most accurately. The answers are very close.

7. (a) $\text{Mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{93+80+15+\cdots+13}{12}$

$$= \frac{346}{12}$$

$$\approx 28.83$$

The mean is 28.83 thousand dollars.

(b) $\text{Median} = \frac{18+19}{2} = 18.5$

The median is 18.5 thousand dollars.

The median best describes the salary of the majority of employees, since the mean is influenced by the high salaries of the president and vice president.

(c) $\text{Mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{15+25+14+\cdots+13}{10}$

$= \frac{173}{10}$

$= 17.3$

The mean is 17.3 thousand dollars.

$\text{Median} = \frac{16+18}{2} = 17$

The median is 17 thousand dollars.

(d) Without the salaries for the two executives, the mean and the median are closer, and both reflect the salary of most of the other workers more accurately. The mean changed quite a bit, while the median did not, a difference that indicates that the mean is more sensitive to the absence or presence of extreme values.

8. (a) Since this data is at the ratio level of measurement, the mean, median, and mode (if it exists) can be used to summarize the data.

(b) Since this data is at the nominal level of measurement, only the mode (if it exists) can be used to summarize the data.

(c) Since this data is at the ratio level of measurement, the mean, median, and mode (if it exists) can be used to summarize the data.

9. (a) Since this data is at the nominal level of measurement, only the mode (if it exists) can be used to summarize the data.

(b) Since this data is at the ratio level of measurement, the mean, median, and mode (if it exists) can be used to summarize the data.

(c) The mode can be used (if it exists). If a 24-hour clock is used, then the data is at the ratio level of measurement, so the mean and median may be used as well.

10. Discussion question.

11. (a) If the largest data value is *replaced* by a larger value, the mean will increase because the sum of the data values will increase, but the number of them will remain the same. The median will not change. The same value will still be in the eighth position when the data are ordered.

(b) If the largest value is replaced by a value that is smaller (but still higher than the median), the mean will decrease because the sum of the data values will decrease. The median will not change. The same value will be in the eighth position in increasing order.

(c) If the largest value is replaced by a value that is smaller than the median, the mean will decrease because the sum of the data values will decrease. The median also will decrease because the former value in the eighth position will move to the ninth position in increasing order. The median will be the new value in the eighth position.

12. Answers will vary according to data collected.

Section 3.2

1. (a) Range = largest value − smallest value
$= 58 - 4 = 54$

The range is 54 deer/km^2.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{251}{12} \approx 20.9$$

The sample mean is 20.9 deer/km^2.

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1} = \frac{2474.9}{11} \approx 225.0$$

The sample variance is 225.0.

$$s = \sqrt{s^2} = \sqrt{225.0} = 15.0$$

The sample standard deviation is 15.0 deer/km^2.

(b) $CV = \frac{s}{\bar{x}} \cdot 100 = \frac{15.0}{20.9} \cdot 100 \approx 71.8\%$

s is 71.8% of $\bar{x}$.

Since the standard deviation is about 71.8% of the mean, there is considerable variation in the distribution of deer from one part of the park to another.

2. (a) Range = largest value − smallest value
$= 78.6 - 17.8 = 60.8$

The range is 60.8%.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{540.8}{10} \approx 54.1$$

The mean is 54.1%.

(b) $s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1} = \frac{3400}{9} \approx 377.78$

The sample variance is 377.78.

$$s = \sqrt{s^2} = \sqrt{377.78} \approx 19.44$$

The standard deviation is 19.44%.

(c) $CV = \frac{s}{\bar{x}} \cdot 100 = \frac{19.44}{54.1} \cdot 100 \approx 35.9\%$

s is 35.9% of $\bar{x}$.

3. **(a)** Range $= 90.3 - 12.7 = 77.6$

The range is 77.6%.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{556.7}{10} \approx 55.7$$

The mean is 55.7%.

(b) $s^2 = \dfrac{\Sigma (x - \bar{x})^2}{n-1} = \dfrac{4833}{9} \approx 537$

The sample variance is 537.

$$s = \sqrt{s^2} = \sqrt{537} \approx 23.17$$

The standard deviation is 23.17%.

(c) $CV = \dfrac{s}{\bar{x}} \cdot 100 = \dfrac{23.17}{55.7} \cdot 100 \approx 41.6\%$

s is 41.6% of $\bar{x}$.

This CV is larger than the CV for geese. So, nesting success rates for ducks have greater relative variability.

4. **(a)** Range $= 14.1 - 6.8 = 7.3$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{63.7}{7} = 9.1$$

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1} = \frac{53.28}{6} = 8.88$$

$$s = \sqrt{s^2} = \sqrt{8.88} \approx 2.98$$

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{2.98}{9.1} \cdot 100 = 32.7\%$$

(b) Range $= 31.0 - 19.1 = 11.9$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{182.9}{7} = 26.1$$

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1} = \frac{118.71}{6} \approx 19.79$$

$$s = \sqrt{s^2} = \sqrt{19.79} \approx 4.45$$

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{4.45}{26.1} \cdot 100 = 17.0\%$$

(c) More relatively consistent productivity at a higher average level.

5. **(a)** Pax $CV = \dfrac{s}{\bar{x}} \cdot 100 = \dfrac{11.56}{11.69} \cdot 100 \approx 98.9\%$

Vanguard $CV = \dfrac{s}{\bar{x}} \cdot 100 = \dfrac{12.50}{5.61} \cdot 100 \approx 222.8\%$

Pax World Balanced seems less risky.

(b) Pax: $\bar{x}-2s=11.69-2(11.56)=-11.43$

$\bar{x}+2s=11.69+2(11.56)=34.81$

At least 75% of the data fall in the interval −11.43% to 34.81%.

Vanguard: $\bar{x}-2s=5.61-2(12.50)=-19.39$

$\bar{x}+2s=5.61+2(12.50)=30.61$

At least 75% of the data fall in the interval −19.39% to 30.61%.

The performance range for Pax seems better than for Vanguard (based on these historical data).

6. (a) Results round to answers given.

(b) $\bar{x}-2s=730-2(172)=386$

$\bar{x}+2s=730+2(172)=1074$

We expect at least 75% of the years to have between 386 and 1074 tornados.

(c) $\bar{x}-3s=730-3(172)=214$

$\bar{x}+3s=730+3(172)=1246$

We expect at least 88.9% of the years to have between 214 and 1246 tornados.

7. (a) $\text{Range}=956-219=737$

$$\bar{x}=\frac{\Sigma x}{n}=\frac{3968}{7}\approx 566.9$$

(b) $$s^2=\frac{\Sigma(x-\bar{x})^2}{n-1}=\frac{427{,}213}{6}\approx 71{,}202$$

$$s=\sqrt{s^2}=\sqrt{71{,}202}\approx 266.8$$

(c) $$CV=\frac{s}{\bar{x}}\cdot 100=\frac{266.8}{566.9}\cdot 100=47.1\%$$

s is 47.1% of $\bar{x}$.

(d) $\bar{x}-2s=566.9-2(266.8)\approx 33$

$\bar{x}+2s=566.9+2(266.8)\approx 1100$

We expect at least 75% of the artifact counts for all such excavation sites to fall in the interval 33 to 1100.

8. $$CV=\frac{s}{\bar{x}}\cdot 100$$

$$\frac{\bar{x}\cdot CV}{100}=s$$

$$s=\frac{\bar{x}\cdot CV}{100}$$

$$s=\frac{2.2(1.5)}{100}$$

$$s=0.033$$

9. (a) Students verify results.

(b) Wal-Mart $CV = \frac{s}{\bar{x}} \cdot 100 = \frac{1.06}{52.03} \cdot 100 \approx 2\%$

Disney $CV = \frac{s}{\bar{x}} \cdot 100 = \frac{0.98}{32.23} \cdot 100 \approx 3\%$

Yes, since the CV's are approximately equal, they appear to be equally attractive.

(c) Wal-Mart:

$\bar{x} - 3s = 52.03 - 3(1.06) = 48.85$

$\bar{x} + 3s = 52.03 + 3(1.06) = 55.21$

Disney:

$\bar{x} - 3s = 32.23 - 3(0.98) = 29.29$

$\bar{x} + 3s = 32.23 + 3(0.98) = 35.17$

The support is \$48.85 and resistance is \$55.21 for Wal-Mart.
The support is \$29.29 and the resistance is \$35.17 for Disney.

10. Answers vary.

11.

Class	f	x	xf	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
21–30	260	25.5	6630	−10.3	106.09	27,583.4
31–40	348	35.5	12,354	−0.3	0.09	31.3
41 and over	287	45.5	13,058.5	9.7	94.09	27,003.8
	$n = \Sigma f = 895$		$\Sigma xf = 32{,}042.5$			$\Sigma(x-\bar{x})^2 f = 54{,}619$

$\bar{x} = \frac{\Sigma xf}{n} = \frac{32{,}042.5}{895} \approx 35.80$

$s^2 = \frac{\Sigma(x-\bar{x})^2 \cdot f}{n-1} = \frac{54{,}619}{894} \approx 61.1$

$s = \sqrt{61.1} \approx 7.82$

12.

Class	f	x	xf	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
1–10	34	5.5	187	−10.6	112.36	3820.24
11–20	18	15.5	279	−0.6	0.36	6.48
21–30	17	25.5	433.5	9.4	88.36	1502.12
31 and over	11	35.5	390.5	19.4	376.36	4139.96
	$n = \Sigma f = 80$		$\Sigma xf = 1290$			$\Sigma(x-\bar{x})^2 f = 9468.8$

$\bar{x} = \frac{\Sigma xf}{n} = \frac{1290}{80} \approx 16.1$

$s^2 = \frac{\Sigma(x-\bar{x})^2 f}{n-1} = \frac{9468.8}{79} \approx 119.9$

$s = \sqrt{119.9} \approx 10.95$

13.

Class	f	x	xf	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
8.6–12.5	15	10.55	158.25	−5.05	25.502	382.537
12.6–16.5	20	14.55	291.00	−1.05	1.102	22.050
16.6–20.5	5	18.55	92.75	2.95	8.703	43.513
20.6–24.5	7	22.55	157.85	6.95	48.303	338.118
24.6–28.5	3	26.55	79.65	10.95	119.903	359.708
	$n=\Sigma f=50$		$\Sigma xf=779.5$			$\Sigma(x-\bar{x})^2 f=1145.9$

$$\bar{x}=\frac{\Sigma xf}{n}=\frac{779.5}{50}\approx 15.6$$

$$s^2=\frac{\Sigma(x-\bar{x})^2 f}{n-1}=\frac{1145.9}{49}\approx 23.4$$

$$s=\sqrt{23.4}\approx 4.8$$

14.

Class	f	x	xf	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
18–24	78	21.0	1638.0	−18.12	328.33	25610.1
25–34	75	29.5	2212.5	−9.62	92.54	6940.8
35–44	48	39.5	1896.0	0.38	0.14	6.9
45–54	33	49.5	1633.5	10.38	107.74	3555.6
55–64	33	59.5	1963.5	20.38	415.34	13706.4
65–80	33	72.5	2392.5	33.38	1114.22	36769.4
	$n=\Sigma f=300$		$\Sigma xf=11{,}736$			$\Sigma(x-\bar{x})^2 f=86{,}589$

$$\bar{x}=\frac{\Sigma xf}{n}=\frac{11{,}736}{300}=39.12$$

$$s=\sqrt{\frac{\Sigma(x-\bar{x})^2 f}{n-1}}=\sqrt{\frac{86{,}589}{299}}\approx 17.02$$

$$CV=\frac{s}{\bar{x}}\cdot 100=\frac{17.02}{39.12}\cdot 100\approx 43.5\%$$

15.

x	f	xf	$x^2 f$
3.5	2	7	24.5
4.5	2	9	40.5
5.5	4	22	121.0
6.5	22	143	929.5
7.5	64	480	3600.0
8.5	90	765	6502.5
9.5	14	133	1263.5
10.5	2	21	220.5
	$\Sigma f=200$	$\Sigma xf=1580$	$\Sigma x^2 f=12{,}702$

$$\bar{x} = \frac{\Sigma xf}{n} = \frac{1580}{200} = 7.9$$

$$SS_x = \Sigma x^2 f - \frac{(\Sigma xf)^2}{n} = 12{,}702 - \frac{(1580)^2}{200} = 220$$

$$s = \sqrt{\frac{SS_x}{n-1}} = \sqrt{\frac{220}{199}} \approx 1.05$$

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{1.05}{7.9} \cdot 100 \approx 13.29\%$$

Section 3.3

1. 82% or more of the scores were at or below her score. 100% –82% = 18% or less of the scores were above her score. Note: This answer is correct, but it relies on a more precise definition than that given in the text on page 124. An adequate answer, matching the definition in the text would be: 82% of the scores were at or below her score, and (100 – 82)% = 18% of the scores were at or above her score.

2. The upper quartile is the 75th percentile. Therefore, the minimal percentile rank must be the 75th.

3. No, the score 82 might have a percentile rank less than 70.

4. Timothy performed better because a percentile rank of 72 is greater than a percentile rank of 70.

5. Order the data from smallest to largest.

   Lowest value = 2
   Highest value = 42

   There are 20 data values.

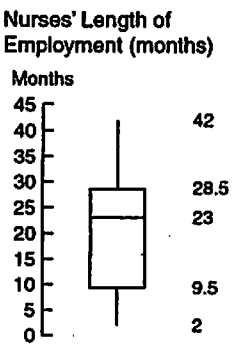
$$\text{Median} = \frac{23+23}{2} = 23$$

There are 10 values less than the Q_2 position and 10 values greater than the Q_2 position.

$$Q_1 = \frac{8+11}{2} = 9.5$$

$$Q_3 = \frac{28+29}{2} = 28.5$$

$$IQR = Q_3 - Q_1 = 28.5 - 9.5 = 19$$



6. (a) Order the data from smallest to largest.

Lowest value = 3
Highest value = 72

There are 20 data values.

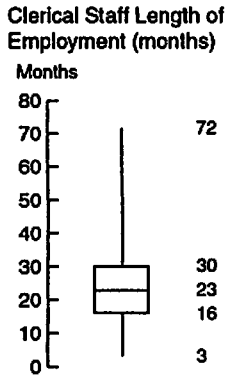
$$\text{Median} = \frac{22+24}{2} = 23$$

There are 10 values less than the median and 10 values greater than the median.

$$Q_1 = \frac{15+17}{2} = 16$$

$$Q_3 = \frac{29+31}{2} = 30$$

$$IQR = Q_3 - Q_1 = 30 - 16 = 14$$



(b) The medians are the same (23) and the *IQR*'s are similar. However, the distances from Q_1 to the minimum value and from Q_3 to the maximum value are greater here than in Problem 7.

7. (a) Order the data from smallest to largest.

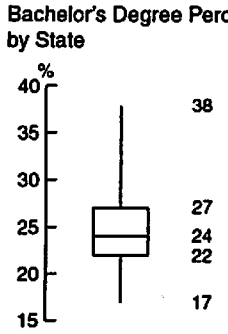
Lowest value = 17
Highest value = 38

There are 50 data values.

$$\text{Median} = \frac{24+24}{2} = 24$$

There are 25 values above and 25 values below the Q_2 position.

$$Q_1 = 22$$
$$Q_3 = 27$$
$$IQR = 27 - 22 = 5$$



(b) 26% is in the 3rd quartile, since it is between the median and Q_3.

8. (a) Order the data from smallest to largest.

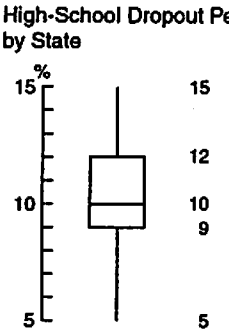
Lowest value = 5
Highest value = 15

There are 50 data values.

$$\text{Median} = \frac{10+10}{2} = 10$$

There are 25 values above and 25 values below the Q_2 position.

$Q_1 = 9$
$Q_3 = 12$
$IQR = 12 - 9 = 3$



(b) 7% is in the 1st quartile, since it is below Q_1.

9. (a) California has the lowest premium since its left whisker is farthest to the left. Pennsylvania has the highest premium since its right whisker is farthest to the right.

(b) Pennsylvania has the highest median premium since its line in the middle of the box is farthest to the right.

(c) California has the smallest range of premiums since the distance between the ends of the whiskers is the smallest. Texas has the smallest interquartile range since the distance between the ends of the boxes is the smallest.

(d) Based on the answers to (a)-(c) above, we can determine that part (a) of Figure 3-13 is for Texas, part (b) of Figure 3-13 is for Pennsylvania, and part (c) of Figure 3-13 is for California.

10. (a) Order the data from smallest to largest.

Lowest value = 4
Highest value = 80

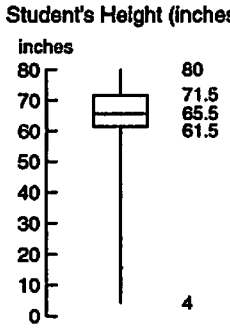
There are 24 data values.

$$\text{Median} = \frac{65+66}{2} = 65.5$$

There are 12 values above and 12 values below the median.

$$Q_1 = \frac{61+62}{2} = 61.5$$

$$Q_3 = \frac{71+72}{2} = 71.5$$



(b) $IQR = Q_3 - Q_1 = 71.5 - 61.5 = 10$

(c) $1.5(10) = 15$
Lower limit: $Q_1 - 1.5(IQR) = 61.5 - 15 = 46.5$
Upper limit: $Q_3 + 1.5(IQR) = 71.5 + 15 = 86.5$

(d) Yes, the value 4 is below the lower limit and so is an outlier; it is probably an error. Our guess is that one of the students is 4 feet tall and listed height in feet instead of inches. There are no values above the upper limit.

11. (a) Assistant had the smallest median percentage salary increase since the bar in the middle of the box is the lowest. Associate had the single highest salary increase since it has the highest asterisk.

(b) Instructor had the largest spread between the first and third quartiles since the distance between the ends of the box is greatest.

(c) Assistant had the smallest spread for the lower 50% of the percentage salary increases since the distance between the bar in the box and the maximum value is the smallest.

(d) Professor had the most symmetric percentage salary increases because there are no outliers and the bar representing the median is close to the center of the box.
Yes, if the outliers for the associate professors were omitted, that distribution would appear to be symmetric.

(e) Associate professor:

$$IQR = 5.075 - 2.350 = 2.725$$
$$Q_3 + 1.5(IQR) = 5.075 + 1.5(2.725) \approx 9.16$$

Yes, since 17.7 is greater than 9.16, there is at least one outlier.

Instructor:

$$IQR = 5.800 - 2.850 = 2.950$$
$$Q_3 + 1.5(IQR) = 5.800 + 1.5(2.950) \approx 10.23$$

Yes, since 13.4 is greater than 10.23, there is at least one outlier.

Chapter 3 Review

1. (a) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{876}{8} = 109.5$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{7044}{7}} = \sqrt{1006.3} \approx 31.7$$

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{31.7}{109.5} \cdot 100 \approx 28.9\%$$

range = maximum value − minimum value
$= 142 - 73 = 69$

(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{881}{8} = 110.125$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{358.87}{7}} \approx 7.2$$

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{7.2}{110.125} \cdot 100 \approx 6.5\%$$

range = maximum value − minimum value
$= 120 - 100 = 20$

(c) The means are about the same. The first distribution has greater spread. The standard deviation, *CV*, and range for the first set of measurements are greater than those for the second set of measurements.

2. (a) Mean $= \bar{x} = \dfrac{\Sigma x}{n} = \dfrac{1.9 + 2.8 + \cdots + 7.2}{8}$

$$= \frac{36.2}{8}$$

$$= 4.525$$

Order the data from smallest to largest.

1.9 1.9 2.8 3.9 4.2 5.7 7.2 8.6

Median $= \dfrac{3.9 + 4.2}{2} = 4.05$

The mode is 1.9 because it is the value that occurs most frequently.

(b) $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{42.395}{7}} \approx 2.46$

$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{2.46}{4.525} \cdot 100 \approx 54.4\%$

Range $= 8.6 - 1.9 = 6.7$

3. (a) Order the data from smallest to largest.

Lowest value = 31
Highest value = 68

There are 60 data values.

$$\text{Median} = \frac{45+45}{2} = 45$$

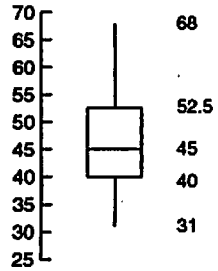
There are 30 values above and 30 values below the Q_2 position.

$$Q_1 = \frac{40+40}{2} = 40$$

$$Q_3 = \frac{52+53}{2} = 52.5$$

$$IQR = 52.5 - 40 = 12.5$$

Percentage of Democratic Vote by Counties in Georgia



(b) Class width = 8

Class	x Midpoint	f	xf	x^2f
31–38	34.5	11	379.5	13,092.8
39–46	42.5	24	1020	43,350.0
47–54	50.5	15	757.5	38,253.8
55–62	58.5	7	409.5	23,955.8
63–70	66.5	3	199.5	13,266.8
		$n = \Sigma f = 60$	$\Sigma xf = 2766$	$\Sigma x^2 f = 131{,}919$

$$\bar{x} = \frac{\Sigma xf}{n} = \frac{2766}{60} = 46.1$$

$$SS_x = \Sigma x^2 f - \frac{(\Sigma xf)^2}{n} = 131{,}919 - \frac{(2766)^2}{60} = 4406.4$$

$$s = \sqrt{\frac{SS_x}{n-1}} = \sqrt{\frac{4406.4}{59}} \approx 8.64$$

$\bar{x} - 2s = 46.1 - 2(8.64) = 28.82$
$\bar{x} + 2s = 46.1 + 2(8.64) = 63.38$
We expect at least 75% of the data to fall in the interval 28.82 to 63.38.

(c) $\bar{x} = 46.15,\ s \approx 8.63$

4. (a) Order the data from smallest to largest.

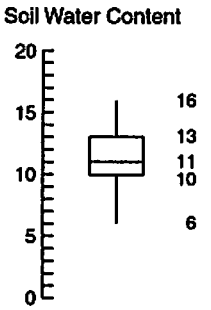
Lowest value = 6
Highest value = 16

There are 50 data values.

$$\text{Median} = \frac{11+11}{2} = 11$$

There are 25 values above and 25 values below the Q_2 position.

$Q_1 = 10$
$Q_3 = 13$
$IQR = Q_3 - Q_1 = 13 - 10 = 3$



(b)

Class	x Midpoint	f	xf	x^2f
6–8	7	4	28	196
9–11	10	24	240	2400
12–14	13	15	195	2535
15–17	16	7	112	1792
		$n = \Sigma f = 50$	$\Sigma xf = 575$	$\Sigma x^2 f = 6923$

$$\bar{x} = \frac{\Sigma xf}{n} = \frac{575}{50} = 11.5$$

$$SS_x = \Sigma x^2 f - \frac{(\Sigma xf)^2}{n} = 6923 - \frac{(575)^2}{50} = 310.5$$

$$s = \sqrt{\frac{SS_x}{n-1}} = \sqrt{\frac{310.5}{49}} \approx 2.52$$

$$\bar{x} - 2s = 11.5 - 2(2.52) = 6.46$$
$$\bar{x} + 2s = 11.5 + 2(2.52) = 16.54$$

We expect at least 75% of the data to fall in the interval 6.46 to 16.54.

(c) $\bar{x} \approx 11.48$; $s \approx 2.44$

5. (a) Mean $= \bar{x} = \frac{\Sigma x}{n} = \frac{10.1 + 6.2 + \cdots + 5.7}{6} = \frac{47}{6} \approx 7.83$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{26.913}{5}} \approx 2.32$$

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{2.32}{7.83} \cdot 100 \approx 29.6\%$$

Range = largest value − smallest value
$= 10.1 - 5.3 = 4.8$

(b) Mean $= \bar{x} = \frac{\Sigma x}{n} = \frac{10.2 + 9.7 + \cdots + 10.1}{6} = \frac{59.7}{6} = 9.95$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{0.415}{5}} \approx 0.29$$

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{0.29}{9.95} \cdot 100 \approx 2.9\%$$

Range = largest value − smallest value
$= 10.3 - 9.6 = 0.7$

(c) Second line has more consistent performance as reflected by the smaller standard deviation, *CV*, and range.

6. Order the data from smallest to largest.

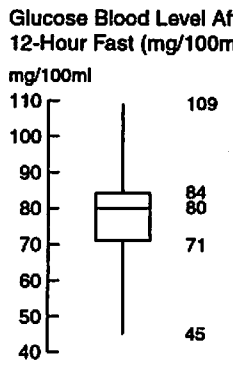
Lowest value = 45
Highest value = 109

There are 70 data values.

$$\text{Median} = \frac{80+80}{2} = 80$$

There are 35 values above and 35 values below the Q_2 position.

$$Q_1 = 71$$
$$Q_3 = 84$$
$$IQR = 84 - 71 = 13$$



7. Mean weight $= \frac{2500}{16} = 156.25$

The mean weight is 156.25 lb.

8. (a) It is possible for the range and the standard deviation to be the same. For instance, for data values that are all the same, such as 1, 1, 1, 1, 1, the range and standard deviation are both 0.

(b) It is possible for the mean, median, and mode to be all the same. For instance, the data set 1, 2, 3, 3, 3, 4, 5 has mean, median, and mode all equal to 3. The averages can all be different, as in the data set 1, 2, 3, 3. In this case, the mean is 2.25, the median is 2.5, and the mode is 3.

Chapter 4 Regression and Correlation

Section 4.1

1. The points seem close to a straight line, so there is moderate or low linear correlation.

2. No straight line is realistically a good fit, so there is no linear correlation.

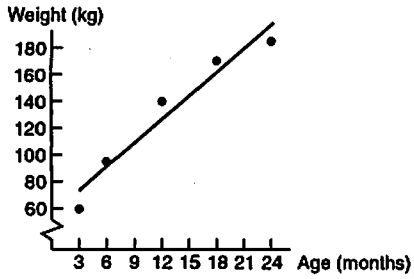
3. The points seem very close to a straight line, so there is high linear correlation.

4. The points seem close to a straight line, so there is moderate or low linear correlation.

5. The points seem very close to a straight line, so there is high linear correlation.

6. No straight line is realistically a good fit, so there is no linear correlation.

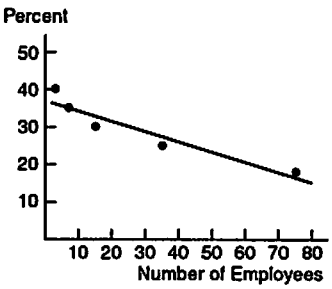
7. (a) Ages and Average Weights of Shetland Ponies



(b) Draw the line you think fits best. (Method to find equation is in Section 4.2.)

(c) Since the points are very close to a straight line, the correlation is high.

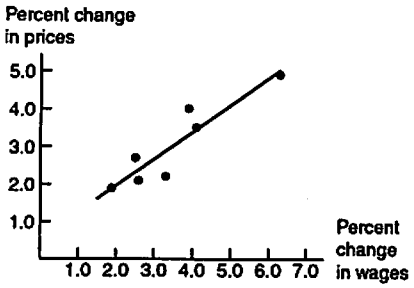
8. (a) Group Health Insurance Plans: Average Number of Employees versus Administrative Costs as a Percentage of Claims



(b) Draw the line you think fits best. (Method to find equation is in Section 4.2.)

(c) Since the points are fairly close to a straight line, the correlation is moderate.

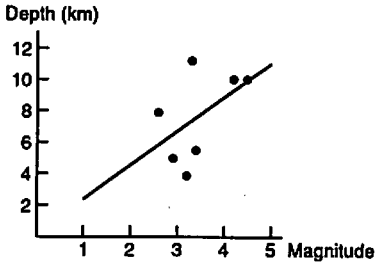
9. (a) Change in Wages and in Consumer Prices in Various Countries (%)



(b) Draw the line you think fits best. (Method to find equation is in Section 4.2.)

(c) Since the points are fairly close to a straight line, the correlation is moderate.

10. (a) Magnitude (Richter Scale) and Depth (km) of Earthquakes

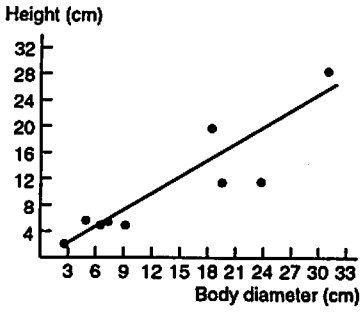


(b) Draw the line you think fits best. (Method to find equation is in Section 4.2.)

(c) Since the points are not close to a straight line, the correlation is low.

Note: One possible reason why there appears to be little, if any, linear relationship is that the Richter scale is logarithmic. An increase of 1 on the Richter scale represents a 60-fold increase in energy.

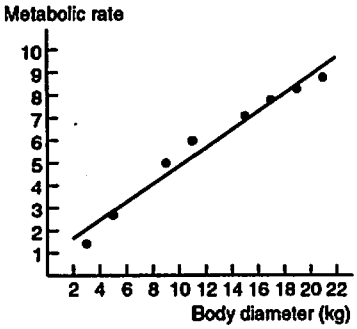
11. (a) Body Diameter and Weight of Prehistoric Pottery



(b) Draw the line you think fits best. (Method to find equation is in Section 4.2.)

(c) Since the points are fairly close to a straight line, the correlation is moderate.

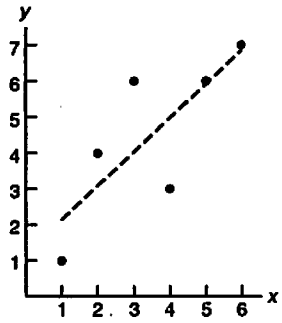
12. (a) Body Weight and Metabolic Rate of Children



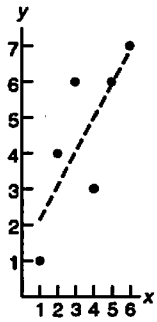
(b) Draw the line you think fits best. (Method to find equation is in Section 4.2.)

(c) Since the points are very close to a straight line, the correlation is high.

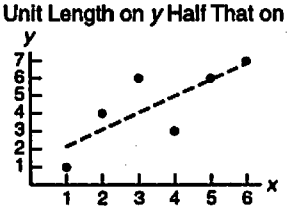
13. (a) Unit Length on *y* Same as That on *x*



(b) Unit Length on *y* Twice That on *x*



(c) Unit Length on *y* Half That on *x*

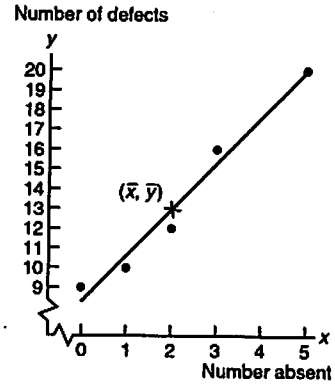


(d) Draw the lines you think best fit the data points.
Stretching the scale on the *y*-axis makes the line appear steeper. Shrinking the scale on the *y*-axis makes the line appear flatter. The slope of the line does not change. Only the appearance (visual impression) of slope changes as the scale of the *y*-axis changes.

Section 4.2

Note: In this section and the next two, answers may vary slightly, depending on how many significant digits are used throughout the calculations.

1. (a) Absenteeism and Number of Assembly Line Defects



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{11}{5} = 2.2$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{67}{5} = 13.4$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{34.6}{14.8} = 2.3378$

$a = \bar{y} - b\bar{x} = 13.4 - 2.3378(2.2) = 8.26$

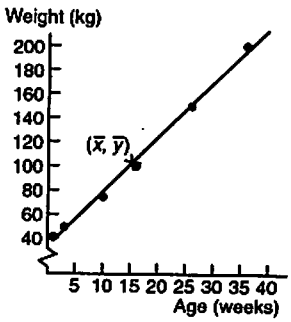
$y = a + bx$ or $y = 8.26 + 2.338x$

(c) See figure of part (a).

(d) Use $x = 4$.

$y_p = 8.26 + 2.338(4) = 17.6$ defects.

2. (a) Age and Weight of Healthy Calves



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{92}{6} = 15.33$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{617}{6} = 102.83$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{4181.3}{927.3} = 4.509$

$a = \bar{y} - b\bar{x} = 102.83 - 4.509(15.33) = 33.70$

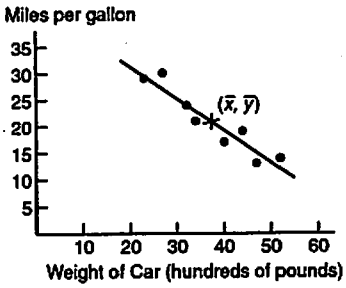
$y = a + bx$ or $y = 33.70 + 4.51x$

(c) See figure of part (a).

(d) Use $x = 12$.

$y_p = 33.70 + 4.51(12) = 87.8$ kg

3. (a) Weight of Cars and Gasoline Mileage



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{299}{8} = 37.375$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{167}{8} = 20.875$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{-427.625}{711.875} = -0.6007$

$a = \bar{y} - b\bar{x} = 20.875 - (-0.6007)(37.375) = 43.3263$

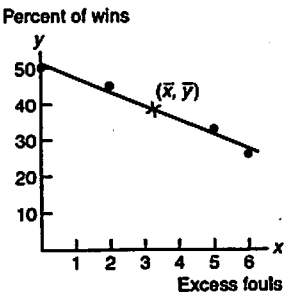
$y = a + bx$ or $y = 43.3263 - 0.6007x$

(c) See figure of part (a).

(d) Use $x = 38$.

$y_p = 43.3263 - 0.6007(38) = 20.5$ mpg

4. (a) Fouls and Basketball Losses



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{13}{4} = 3.25$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{154}{4} = 38.5$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{-89.5}{22.75} = -3.934$

$a = \bar{y} - b\bar{x} = 38.5 - (-3.934)(3.25) = 51.29$

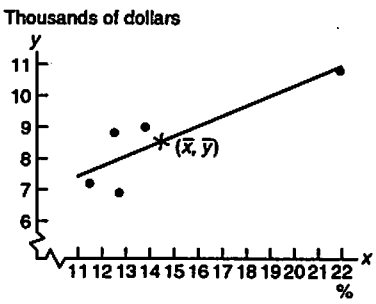
$y = a + bx$ or $y = 51.29 - 3.934x$

(c) See figure of part (a).

(d) Use $x = 4$.

$y_p = 51.29 - 3.934(4) = 35.55\%$

5. (a) Education and Income in Small Cities



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{72.4}{5} = 14.48$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{42.7}{5} = 8.54$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{22.854}{71.448} = 0.320$

$a = \bar{y} - b\bar{x} = 8.54 - 0.320(14.48) = 3.91$

$y = a + bx$ or $y = 3.91 + 0.320x$

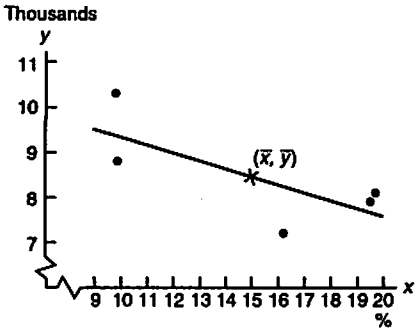
(c) See figure of part (a).

Note that the regression line would be much steeper if (21.9, 10.8) were eliminated from the data set [which would also affect $(\bar{x}, \bar{y})$]. Not all outliers (this point is an outlier in both x (probably) and y) have this effect; however, when the parameter estimates a and b depend heavily on a particular observation, as is the case here, the point is called "influential," and conclusions drawn are shaky at best when influential observations remain in the data. For further information, refer to a more advanced textbook such as Applied Regression Analysis by Draper and Smith.

(d) Use $x = 20$.

$y_p = 3.91 + 0.320(20) = 10.31$ i.e., 10.31 thousand dollars

6. (a) Percentage of 16 to 19-Year-Olds Not in School and per Capita Income (thousands of dollars)



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{75.1}{5} = 15.02$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{42.3}{5} = 8.46$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{-17.026}{96.828} = -0.1758$

$a = \bar{y} - b\bar{x} = 8.46 - (-0.1758)15.02 = 11.10$

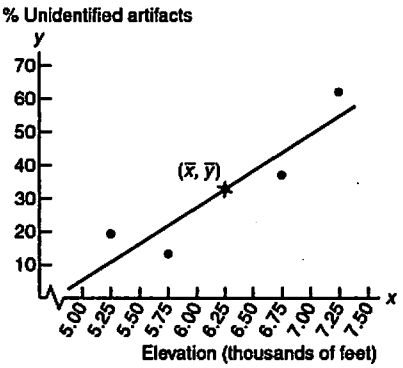
$y = a + bx$ or $y = 11.10 - 0.176x$

(c) See figure of part (a).

(d) Use $x = 17$.

$y_p = 11.10 - 0.176(17) = 8.11$ thousand dollars

7. (a) Cultural Affiliation and Elevation of Achaeological Sites



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{31.25}{5} = 6.25$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{164}{5} = 32.8$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{55}{2.5} = 22.0$

$a = \bar{y} - b\bar{x} = 32.8 - 22.0(6.25) = -104.7$

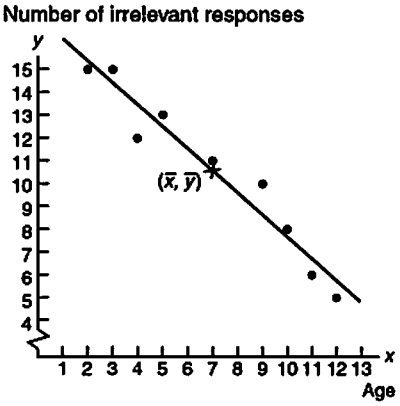
$y = a + bx$ or $y = -104.7 + 22.0x$

(c) See figure of part (a).

(d) Use $x = 6.5$.

$y_p = -104.7 + 22.0(6.5) = 38.3$ percent

8. (a) Ages of Children and Their Responses to Questions



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{63}{9} = 7.0$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{95}{9} = 10.56$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{-104}{108} = -0.96296$

$a = \bar{y} - b\bar{x} = 10.56 - (-0.96296)(7.0) = 17.30$

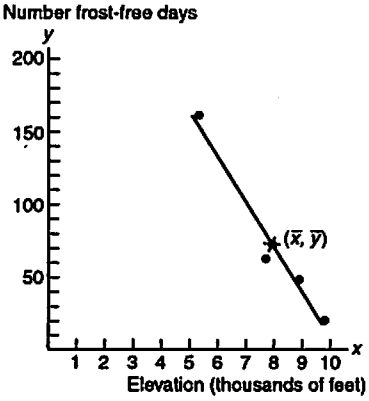
$y = a + bx$ or $y = 17.30 - 0.963x$

(c) See figure of part (a).

(d) Use $x = 9.5$.

$y_p = 17.30 - 0.963(9.5) = 8.15$ irrelevant responses

9. (a) Elevation and the Number of Frost-Free Days



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{39.6}{5} = 7.92$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{368}{5} = 73.6$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{-352.26}{11.408} = -30.8783$

$a = \bar{y} - b\bar{x} = 73.6 - (-30.8783)(7.92) = 318.16$

$y = a + bx$ or $y = 318.16 - 30.878x$

(c) See figure of part (a).

Note: Compare this figure to that in Problem 5 above, the point (5.3, 162) is an outlier (possibly in x, definitely in y) but it is more or less along the regression line that would be drawn if it were eliminated from the data set. Thus, this is not an "influential" observation.

(d) Use $x = 6$.

$y_p = 318.16 - 30.878(6) = 132.89$ days

10. (a) Results checks.

(b) Results checks.

(c) Yes.

(d) $y = 0.143 + 1.071x$

$y - 0.143 = 1.071x$

$\dfrac{y - 0.143}{1.071} = x$

$\dfrac{1}{1.071}y - \dfrac{0.143}{1.071} = x$

or

$x = 0.9337y - 0.1335$

The equation $x = 0.9337y - 0.1335$ does not match part (b), with the symbols x and y exchanged.

(e) In general, switching x and y values produces a different least-squares equation. It is important that when you perform a linear regression, you know which variable is the explanatory variable and which is the response variable.

Section 4.3

1. (a) No, high positive correlation does not mean causation.

(b) An increase in the population is a third factor that might cause traffic accidents and the number of safety stickers to increase together.

2. (a) No, high positive correlation does not mean causation.

(b) There is an increase in buying power due to increase in salaries.

3. (a) No, strong negative correlation does not mean causation.

(b) Better medical treatment is a third factor that might be decreasing infant mortalities and at the same time increasing life span.

4. (a) No, strong positive correlation does not mean causation.

(b) An increase in population could account for increases both in consumption of soda pop and in number of traffic accidents.

5. (a) Number of Jobs (in hundreds)

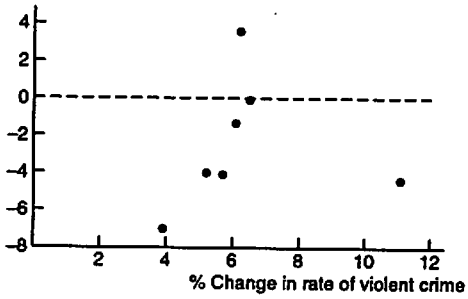
(b) r should be close to 1 because the points seem to be clustered fairly close to a straight line going up from left to right.

(c) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{153.\overline{3}}{\sqrt{953.\overline{3}(33.\overline{3})}} = 0.860$

$r^2 = (0.860)^2 = 0.740$

This means that 74.0% of the variation in y = number of entry-level jobs can be explained by the corresponding variation in x = total number of jobs using the least squares line. 100% – 74.0% = 26.0% of the variation is unexplained.

6. (a) % Change in rate of imprisonment



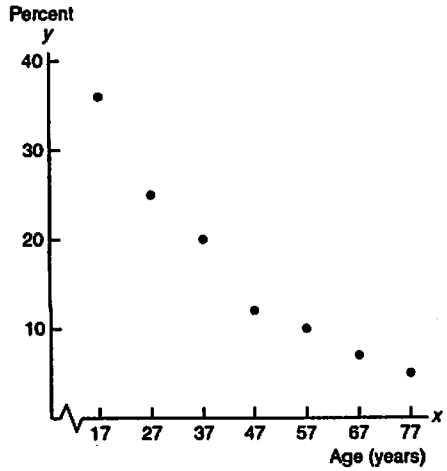
(b) r should be close to 0 because the points are not all clustered around a straight line, due to (11.1, –4.4) (which is an influential observation).

(c) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{3.9314}{\sqrt{30.4086(72.8486)}} = 0.084$

$r^2 = (0.084)^2 = 0.007$

This means that 0.7% of the variation in y = percent change in the rate of imprisonment can be explained by the corresponding variation in x = percent change in the rate of violent crime using the least squares line. 100% – 0.7% = 99.3% of the variation is unexplained.

7. (a) Drivers' Ages and Percent Fatal Accidents Due to Speeding



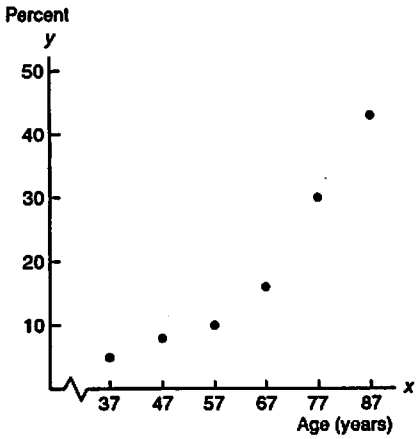
(b) r should be closer to −1 because the points are clustered very close to a straight line going down from left to right. (Note also that the data values fall nicely on a curve.)

(c) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{-1390}{\sqrt{2800(749.714)}} = -0.959$

$r^2 = (-0.959)^2 = 0.920$

This means that 92% of the variation in y = percentage of all fatal accidents due to speeding can be explained by the corresponding variation in x = age in years of a licensed automobile driver using the least squares line. 100% − 92% = 8% of the variation is unexplained.

8. (a) Driver's Ages and Percent Fatal Accidents Due to Not Yielding



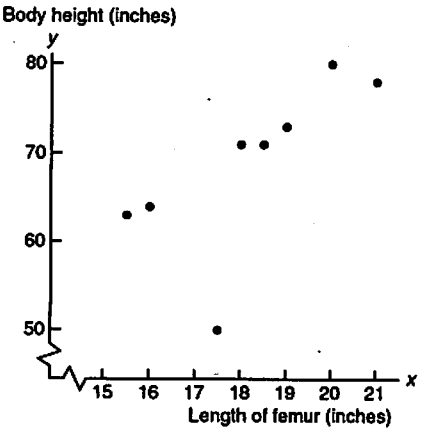
(b) r should be closer to 1 because the points are clustered very close to a straight line going up from left to right. (Note also that the data follow a curve.)

(c) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{1310}{\sqrt{1750(1103.\overline{3})}} = 0.943$

$r^2 = (0.943)^2 = 0.889$

This means that 88.9% of the variation in y = percentage of fatal accidents due to failure to yield the right of way can be explained by the corresponding variation in x = age of a licensed driver in years using the least squares line. 100% – 88.9% = 11.1% of the variation is unexplained.

9. (a) Body Height and Bone Size



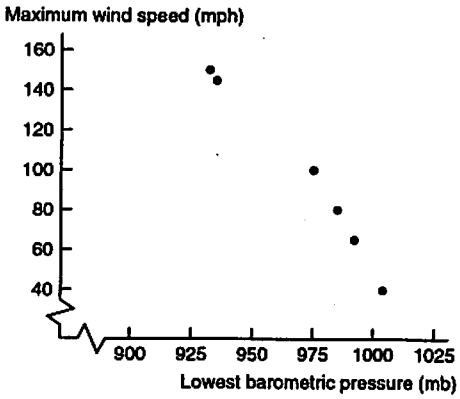
(b) r should be closer to 1 because the points are clustered close to a straight line going up from left to right.

(c) $$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{88.875}{\sqrt{24.4688(647.5)}} = 0.7061$$

$r^2 = (0.7061)^2 = 0.499$

This means that 49.9% of the variation in y = body height can be explained by the corresponding variation in x = length of femur using the least squares line. 100% – 49.9% = 50.1% of the variation is unexplained.

10. (a) Lowest Barometric Pressure and Maximum Wind Speed for Tropical Cyclones



(b) r should be closer to –1 because the points are clustered very close to a straight line going down from left to right.

(c) $$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{-6575}{\sqrt{4557.5(9683.\overline{3})}} = -0.9897$$

$r^2 = (-0.9897)^2 = 0.9795$ or 0.98

This means that 98% of the variation in y = maximum wind speed of the cyclone can be explained by the corresponding variation in x = lowest pressure as a cyclone approaches using the least-squares line. 100% – 98% = 2% of the variation is unexplained.

11. (a) We get the same result.
$SS_{xy} = SS_{yx}$

(b) We get the same result.

(c) We get the same result.

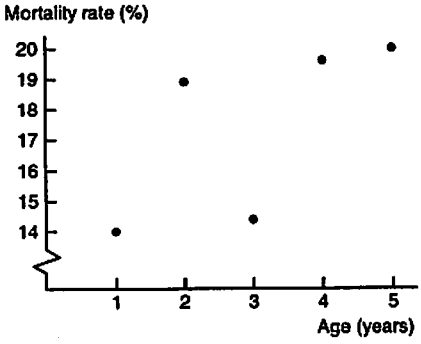
(d) First set: $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{5}{\sqrt{4.\overline{6}(14)}} = 0.618590$

Second set: $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{5}{\sqrt{14(4.\overline{6})}} = 0.618590$

$r = 0.618590$ in both cases.
The least-squares equations are not necessarily the same.

Chapter 4 Review Problems

1. (a) Age and Mortality Rate for Bighorn Sheep



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{15}{5} = 3$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{86.9}{5} = 17.38$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{12.7}{10} = 1.27$

$a = \bar{y} - b\bar{x} = 17.38 - 1.27(3) = 13.57$

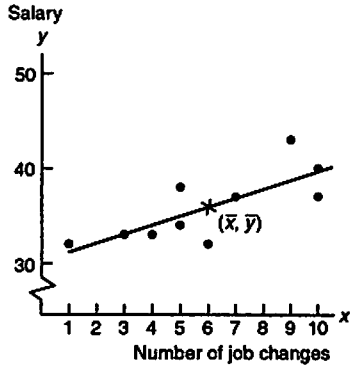
$y = a + bx$ or $y = 13.57 + 1.27x$

(c) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{12.7}{\sqrt{10(34.408)}} = 0.685$

$r^2 = (0.685)^2 = 0.469$

The correlation coefficient r measures the strength of the linear relationship between a bighorn sheep's age and the mortality rate. The coefficient of determination, r^2, means that 46.9% of the variation in y = mortality rate in this age groups can be explained by the corresponding variation in x = age of a bighorn sheep using the least-squares line.

2. (a) Annual Salary (thousands) and Number of Job Changes



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{60}{10} = 6.0$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{359}{10} = 35.9$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{77}{82} = 0.939024$

$a = \bar{y} - b\bar{x} = 35.9 - 0.939024(6.0) = 30.266$

$y = a + bx$ or $y = 30.266 + 0.939x$

(c) See the figure in part (a).

(d) Let $x = 2$.

$y_p = 30.266 + 0.939(2) = 32.14$

The predicted salary is \$32,140.

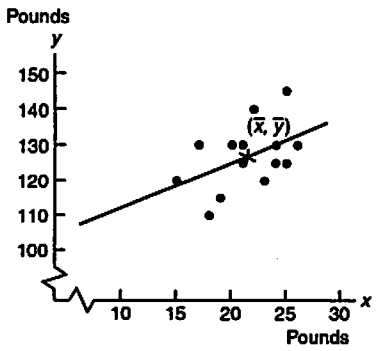
(e) The correlation coefficient will be positive because the points are clustered around a straight line going up from left to right.

(f) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{77}{\sqrt{82(124.9)}} = 0.761$

$r^2 = (0.761)^2 = 0.579$

This means that 57.9% of the variation in y = salary can be explained by the corresponding variation in x = number of job changes using the least-squares line.

3. (a) Weight of One-Year-Old versus Weight of Adult



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{300}{14} = 21.43$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{1775}{14} = 126.79$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{184.2857}{143.4286} = 1.285$

$a = \bar{y} - b\bar{x} = 126.79 - (1.285)(21.43) = 99.25$

$y = a + bx$ or $y = 99.25 + 1.285x$

(c) See the figure in part (a).

(d) Let $x = 20$.

$y_p = 99.25 + 1.285(20) = 124.95$

The predicted weight is 124.95 pounds.

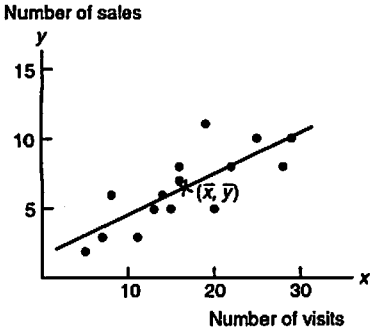
(e) The correlation coefficient will be positive because the points are clustered around a straight line going up from left to right.

(f) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{184.2857}{\sqrt{143.4286(1080.36)}} = 0.468$

$r^2 = (0.468)^2 = 0.219$

The correlation coefficient r measures the strength of the linear relationship between a woman's weight at age 1 and at age 30. The coefficient of determination r^2 means that 21.9% of the variation in y = weight of a mature adult (30 years old) can be explained by the corresponding variation in x = weight of a 1-year-old baby using the least-squares line.

4. (a) Number of Insurance Sales and Number of Visits



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{248}{15} = 16.5\overline{3} \approx 16.53$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{97}{15} = 6.4\overline{6} \approx 6.47$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{221.2\overline{6}}{755.7\overline{3}} = 0.292784$

$a = \bar{y} - b\bar{x} = 6.4\overline{6} - 0.292784(16.5\overline{3}) = 1.626$

$y = a + bx$ or $y = 1.626 + 0.293x$

(c) See the figure in part (a).

(d) Let $x = 18$.

$y_p = 1.626 + 0.293(18) = 6.9$

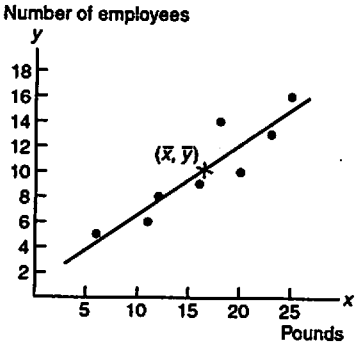
The predicted number of sales is 6.9 or 7.

(e) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{221.2\overline{6}}{\sqrt{755.7\overline{3}(103.7\overline{3})}} = 0.790$

$r^2 = (0.790)^2 = 0.624$

This means that 62.4% of the variation in y = number of people who bought insurance that week can be explained by the corresponding variation in x = number of visits made each week using the least-squares line.

5. (a)



(b) $\bar{x} = \dfrac{\sum x}{n} = \dfrac{131}{8} = 16.375 \approx 16.38$

$\bar{y} = \dfrac{\sum y}{n} = \dfrac{81}{8} = 10.125 \approx 10.13$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{160.625}{289.875} = 0.554118 \approx 0.554$

$a = \bar{y} - b\bar{x} = 10.125 - 0.554118(16.375) = 1.051$

$y = a + bx$ or $y = 1.051 + 0.544x$

(c) See the figure in part (a).

(d) Use $x = 15$.

$y_p = 1.051 + 0.544(15) = 9.36$

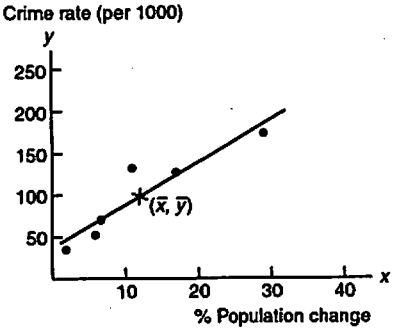
About 9 or 10 employees should be assigned mail duty.

(e) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{160.625}{\sqrt{289.875(106.875)}} = 0.913$

$r^2 = (0.913)^2 = 0.834$

The correlation coefficient r measures the strength of the linear association between weight of incoming mail and number of employees assigned to answer it. The coefficient of determination, r^2, means that 83.4% of the variation in y = number of employees can be explained by the corresponding variation in x = weight of incoming mail using the least-squares line.

6. (a) Percent Population Change



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{72}{6} = 12.0$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{589}{6} = 98.1\overline{6} \approx 98.17$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{2431}{476} = 5.1071 \approx 5.11$

$a = \bar{y} - b\bar{x} = 98.1\overline{6} - 5.1071(12.0) = 36.881 \approx 36.9$

$y = a + bx$ or $y = 36.9 + 5.11x$

See the figure in part (a).

(c) Let $x = 12$

$y = 36.9 + 5.11(12) = 98.2$

The predicted crime rate is 98.2 crimes per thousand.

(d) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{2431}{\sqrt{476(14456.8\overline{3})}} = 0.927$

$r^2 = (0.927)^2 = 0.859$

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{0.927\sqrt{6-2}}{\sqrt{1-(0.927)^2}} = 4.94$

$d.f. = n - 2 = 6 - 2 = 4$

At 1% level of significance, $t_0 = \pm 4.604$.

Since $4.94 > 4.604$, we reject H_0 and conclude that the sample evidence supports a significant correlation coefficient.

(e) High correlation does not guarantee a "cause-and-effect" situation. Before causation is established, more work needs to be done taking other variables into account.
High correlation is simply an indication of a mathematical relationship between variables.

Chapter 5 Elementary Probability Theory

Section 5.1

1. Answers vary. Probability is a number between 0 and 1, inclusive, that expresses the likelihood that a specific event will occur. Three ways to find or assign a probability to an event are (1) through intuition (subjective probability), (2) by considering the long-term relative frequency of recurrence of an event in repeated independent trials (empirical probability), and (3) by computing the ratio of the number of favorable outcomes to the total number of possible outcomes, assuming all outcomes are equally likely (classical probability).

2. Answers vary. Probability in business: market research; in medicine: drug tests to determine if a new drug is more effective than the standard treatment; in social science: determining which characteristics to use in creating a profile to detect terrorists; in natural sciences: predicting the likely path and location of landfall for a hurricane.
Statistics is the science of collecting, analyzing, and interpreting quantitative data in such a way that the reliability of the conclusions based on the data can be evaluated objectively. Probability is used in determining the reliability of the results.

3. These are not probabilities: (b) because it is greater than 1, (d) because it is less than zero (negative), (h) 150% = 1.50, because it is greater than 1.

4. Remember $0 \le$ probability of an event ≤ 1
   (a) $-0.41 < 0$
   (b) $1.21 > 1$
   (c) $120\% = 1.2 > 1$
   (d) yes, $0 \le 0.56 \le 1$

5. Answers vary. The result is a sample, although not necessarily a good one, showing the relative frequency of people able to wiggle their ears.

6. Answers vary. The results are one example (not necessarily a good one) of the relative frequency of occurrence of raising one eyebrow.

7. (a) $P(\text{no similar preferences}) = P(0) = \frac{15}{375}$, $P(1) = \frac{71}{375}$, $P(2) = \frac{124}{375}$, $P(3) = \frac{131}{375}$, $P(4) = \frac{34}{375}$
   (b) $\frac{15+71+124+131+34}{375} = \frac{375}{375} = 1$, yes
   Personality types were classified into 4 main preferences; all possible numbers of shared preferences were considered. The sample space is 0, 1, 2, 3, and 4 shared preferences.

8. (a) $P(\text{couple not engaged}) = \frac{200}{1000} = 0.20$, $P(\text{dated less than 1 year}) = \frac{240}{1000} = 0.24$, $P(\text{dated 1 to 2 years}) = \frac{210}{1000} = 0.21$, $P(\text{dated more than 2 years}) = \frac{350}{1000} = 0.35$, based on the number of favorable outcomes divided by the total number of outcomes (1000 couples' engagement status)
   (b) $\frac{200+240+210+350}{1000} = \frac{1000}{1000} = 1$, yes
   They should add to 1 because all possible outcomes were considered. The sample space is never engaged, engaged less than 1 year, engaged 1 to 2 years, engaged more than 2 years.

9. **(a)** Note: "includes the left limit but not the right limit" means 6 A.M. $\le$ time $t <$ noon, noon $\le t <$ 6 P.M., 6 P.M. $\le t <$ midnight, midnight $\le t <$ 6 A.M. P(best idea 6 A.M.–12 noon) $= \dfrac{290}{966} \approx 0.30$; P(best idea 12 noon–6 P.M.) $= \dfrac{135}{966} \approx 0.14$; P(best idea 6 P.M.–12 midnight) $\dfrac{319}{966} \approx 0.33$; P(best idea from 12 midnight to 6 A.M.) $= \dfrac{222}{966} \approx 0.23$.

(b) The probabilities add up to 1. They should add up to 1 provided that the intervals do not overlap and each inventor chose only one interval. The sample space is the set of four time intervals.

10. **(a)** $P(\text{germinate}) = \dfrac{\text{number germinated}}{\text{number planted}} = \dfrac{2430}{3000} = 0.81$

(b) $P(\text{not germinate}) = \dfrac{3000 - 2430}{3000} = \dfrac{570}{3000} = 0.19$

(c) The sample space is 2 outcomes, germinate and not germinate.
P(germinate) + P(not germinate) = 0.81 + 0.19 = 1
The probabilities of all the outcomes in the sample space should and do sum to 1.

(d) no; P(germinate) = 0.81, P(not germinate) = 0.19
If they were equally likely, each would have probability $\dfrac{1}{2} = 0.5$.

11. Make a table showing the information known about the 127 people who walked by the store: [Example 6 in Section 4.2 uses this technique.]

	Buy	Did not buy	Row Total
Came into the store	25	58 – 25 = 33	58
Did not come in	0	69	127 – 58 = 69
Column Total	25	102	127

If 58 came in, 69 didn't; 25 of the 58 bought something, so 33 came in but didn't buy anything. Those who did not come in, couldn't buy anything. The row entries must sum to the row totals; the column entries must sum to the column totals; and the row totals, as well as the column totals, must sum to the overall total, i.e., the 127 people who walked by the store. Also, the four inner cells must sum to the overall total: 25 + 33 + 0 + 69 = 127.

This kind of problem relies on formula (2), $P(\text{event } A) = \dfrac{\text{number outcomes favorable to } A}{\text{total number of outcomes}}$. The "trick" is to decide what belongs in the denominator *first*. If the denominator is a row total, stay in that row. If the denominator is a column total, stay in that column. If the denominator is the overall total, the numerator can be a row total, a column total, or the number in any one of the four "cells" inside the table.

(a) total outcomes: people walking by, overall total, 127
favorable outcomes: enter the store, row total, 58 (that's all we know about them)
$P(A) = \dfrac{58}{127} \approx 0.46$

(b) total outcomes: people who walk into the store, row total 58
favorable outcomes: staying in the row, those who buy: 25
$P(A) = \dfrac{25}{58} \approx 0.43$

(c) total outcomes: people walking by, overall total 127
favorable outcomes: people coming in *and* buying, the cell at the *intersection* of the "coming in" row and the "buying" column (the upper left corner), 25 (Recall from set theory that "and" means both things happen, that the two sets *intersect*: >)

$$P(A) = \frac{25}{127} \approx 0.20$$

(d) total outcomes: people coming into the store, row total, 58
favorable outcomes: staying in the row, those who do not buy, 33

$$P(A) = \frac{33}{58} \approx 0.57$$

$$\left(\text{alternate method: this is the complement to (b): } P(A) = 1 - \frac{25}{58} = \frac{33}{58} \approx 0.57\right)$$

Section 5.2

1. (a) Green and blue are mutually exclusive because each M&M candy is only 1 color.
P(green *or* blue) = P(green) + P(blue) = 10% + 10% = 20%

   (b) Yellow and red are mutually exclusive, again, because each candy is only one color, and if the candy is yellow, it can't be red, too.
P(yellow *or* red) = P(yellow) + P(red) = 20% + 20% = 40%

   (c) It is faster here to use the complementary event rule than to add up the probabilities of all the colors except purple.
P(*not* purple) = 1 – P(purple) = 1 – 0.20 = 0.80, or 80%

2. (a) Green and blue are mutually exclusive because each M&M candy is only 1 color.
P(green *or* blue) = P(green) + P(blue) = 20% + 10% = 30%

   (b) Yes, mutually exclusive colors
P(yellow *or* red) = P(yellow) + P(red) = 20% + 20% = 40%

   (c) P(*not* purple) = 1 – P(purple) = 1 – 0.20 = 0.80 = 80%
Since the percentage of green M&Ms is 10% for plain and 20% for almond, I expect the results for part (a) to be different. Parts (b) and (c) should be the same because the percentages for these colors are the same.

3. (a) Green and blue are mutually exclusive because each M&M candy is only 1 color.
P(green *or* blue) = P(green) + P(blue) = 16.6% + 16.6% = 33.2%

   (b) Mutually exclusive: P(yellow *or* red) = P(yellow) + P(red) = 16.6% + 16.6% = 33.2%

   (c) P(*not* purple) = 1 – P(purple) = 1 – 0 = 1 = 100% (no purple)
Since the color distributions differ for plain and Dulce de Leche~Carmel M&Ms, I expect the results for all parts to be different. If the answers were the same, it would only be by coincidence.

4. The total number of arches tabled is 288. Arch heights are mutually exclusive because if the height is 12 feet, it can't be 42 feet as well.

   (a) $P(3 \text{ to } 9) = \frac{111}{288}$

   (b) $P(30 \textit{ or taller}) = P(30 \text{ to } 49) + P(50 \text{ to } 74) + P(75 \text{ and higher}) = \frac{30}{288} + \frac{33}{288} + \frac{18}{288} = \frac{81}{288}$

   (c) $P(3 \text{ to } 49) = P(3-9) + P(10-29) + P(30-49) = \frac{111}{288} + \frac{96}{288} + \frac{30}{288} = \frac{237}{288}$

   (d) $P(10 \text{ to } 74) = P(10-29) + P(30-49) + P(50-74) = \frac{96}{288} + \frac{30}{288} + \frac{33}{288} = \frac{159}{288}$

(e) $P(75 \text{ or taller}) = \frac{18}{288}$

Hint for Problems 5–8: Refer to Figure 4–1 if necessary. (Without loss of generality, let the red die be the first die and the green die be the second die in Figure 4–1.) Think of the outcomes as an (x, y) ordered pair. Then, without loss of generality, (1, 6) means 1 on the red die and 6 on the green die. (We are "ordering" the dice for convenience only — which is first and which is second have no bearing on this problem.) The only important fact is that they are distinguishable outcomes, so that (1 on red, 2 on green) is different from (2 on red, 1 on green).

5. **(a)** Yes, the outcome of the red die does not influence the outcome of the green die.

   (b) $P(5 \text{ on green } \textit{and } 3 \text{ on red}) = P(5 \text{ on green}) \cdot P(3 \text{ on red}) = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36} \approx 0.028$ because they are independent.

   (c) $P(3 \text{ on green } \textit{and } 5 \text{ on red}) = P(3 \text{ on green}) \cdot P(5 \text{ on red}) = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36} \approx 0.028$

   (d) $P[(5 \text{ on green } \textit{and } 3 \text{ on red}) \textit{ or } (3 \text{ on green } \textit{and } 5 \text{ on red})]$
   $= P(5 \text{ on green } \textit{and } 3 \text{ on red}) + P(3 \text{ on green } \textit{and } 5 \text{ on red})]$
   $= \frac{1}{36} + \frac{1}{36} = \frac{2}{36} = \frac{1}{18} \approx 0.056$ [because they are mutually exclusive outcomes]

6. **(a)** Yes, the outcome of the red die does not influence the outcome of the green die.

   (b) $P(1 \text{ on green } \textit{and } 2 \text{ on red}) = P(1 \text{ on green}) \cdot P(2 \text{ on red}) = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$

   (c) $P(2 \text{ on green } \textit{and } 1 \text{ on red}) = P(2 \text{ on green}) \cdot P(1 \text{ on red}) = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$

   (d) $P[(1 \text{ on green } \textit{and } 2 \text{ on red}) \textit{ or } (2 \text{ on green } \textit{and } 1 \text{ on red})]$
   $= P(1 \text{ on green } \textit{and } 2 \text{ on red}) + P(2 \text{ on green } \textit{and } 1 \text{ on red})]$
   $= \frac{1}{36} + \frac{1}{36} = \frac{2}{36} = \frac{1}{18}$ [because they are mutually exclusive outcomes]

7. **(a)** $1 + 5 = 6, 2 + 4 = 6, 3 + 3 = 6, 4 + 2 = 6, 5 + 1 = 6$
   $P(\text{sum} = 6) = P[(1, 5) \textit{ or } (2, 4) \textit{ or } (3 \text{ on red}, 3 \text{ on green}) \textit{ or } (4, 2) \textit{ or } (5, 1)]$
   $= P(1, 5) + P(2, 4) + P(3, 3) + P(4, 2) + P(5, 1)$
   since the (red, green) outcomes are mutually exclusive
   $= \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)$
   because the red die outcome is independent of the green die outcome
   $= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{5}{36}$

   (b) $1 + 3 = 4, 2 + 2 = 4, 3 + 1 = 4$
   $P(\text{sum is } 4) = P[(1, 3) \textit{ or } (2, 2) \textit{ or } (3, 1)]$
   $= P(1, 3) + P(2, 2) + P(3, 1)$
   because the (red, green) outcomes are mutually exclusive
   $= \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)$
   because the red die outcome is independent of the green die outcome
   $= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{3}{36} = \frac{1}{12}$

   (c) Since a sum of six can't simultaneously be a sum of 4, these are mutually exclusive events;
   $P(\text{sum of 6 } \textit{or } 4) = P(\text{sum of 6}) + P(\text{sum of 4}) = \frac{5}{36} + \frac{3}{36} = \frac{8}{36} = \frac{2}{9}$

8. (a) $1+6=7, 2+5=7, 3+4=7, 4+3=7, 5+2=7, 6+1=7$

$P(\text{sum is } 7) = P[(1, 6) \textit{ or } (2, 5) \textit{ or } (3, 4) \textit{ or } (4, 3) \textit{ or } (5, 2) \textit{ or } (6, 1)]$

$= P(1, 6) + P(2, 5) + P(3, 4) + P(4, 3) + P(5, 2) + P(6, 1)$

because the (red, green) outcomes are mutually exclusive

$$= \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)$$

because the red die outcome is independent of the green die outcome

$$= \frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}=\frac{6}{36}=\frac{1}{6}$$

(b) $5+6=11, 6+5=11$

$P(\text{sum is } 11) = P[(5, 6) \textit{ or } (6, 5)]$

$= P(5, 6) + P(6, 5)$

because the (red, green) outcomes are mutually exclusive

$$= \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)$$

because the red die outcome is independent of the green die outcome

$$= \frac{1}{36}+\frac{1}{36}=\frac{2}{36}=\frac{1}{18}$$

(c) Since a sum of can't be both 7 and 11, they are mutually exclusive

$$P(\text{sum is } 7 \textit{ or } 11) = P(\text{sum is } 7) + P(\text{sum is } 11) = \frac{6}{36}+\frac{2}{36}=\frac{8}{36}=\frac{2}{9}$$

9. (a) No, the key idea here is "without replacement," which means the draws are dependent, because the outcome of the second card drawn depends on what the first card drawn was. Let the card draws be represented by an (x, y) ordered pair. For example, (K, 6) means the first card drawn was a king and the second card drawn was a 6. Here the order of the cards *is* important.

(b) $P(\text{ace on 1st } \textit{and} \text{ king on second}) = P(\text{ace, king}) = \left(\frac{4}{52}\right)\left(\frac{4}{51}\right)=\frac{16}{2652}=\frac{4}{663}$

There are 4 aces and 4 kings in the deck. Once the first card is drawn and not replaced, there are only 51 cards left to draw from, but all the kings are still there.

(c) $P(\text{king, ace}) = \left(\frac{4}{52}\right)\left(\frac{4}{51}\right)=\frac{16}{2652}=\frac{4}{663}$

There are 4 kings and 4 aces in the deck. Once the first card is drawn and not replaced, there are only 51 cards left to draw from, but all the aces are still there.

(d) $P(\text{ace } \textit{and} \text{ king in either order})$

$= P[(\text{ace, king}) \textit{ or } (\text{king, ace})]$

$= P(\text{ace, king}) + P(\text{king, ace})$ because these two outcomes are mutually exclusive

$$= \frac{16}{2652}+\frac{16}{2652}=\frac{32}{2652}=\frac{8}{663}$$

10. (a) No, the key idea here is "without replacement," which means the draws are dependent, because the outcome of the second card drawn depends on what the first card drawn was. Let the card draws be represented by an (x, y) ordered pair. For example, (K, 6) means the first card drawn was a king and the second card drawn was a 6. Here the order of the cards *is* important.

(b) $P(3, 10) = P[(3 \text{ on 1st}) \textit{ and } (10 \text{ on 2nd, } \textit{given } 3 \text{ on 1st})]$

$= P(3 \text{ on 1st}) \cdot P(10 \text{ on 2nd, } \textit{given } 3 \text{ on 1st})$

$$= \left(\frac{4}{52}\right)\left(\frac{4}{51}\right)=\frac{16}{2652}=\frac{4}{663}\approx 0.006$$

(c) $P(10, 3) = P[(10 \text{ on 1st}) \textit{ and } (3 \text{ on 2nd, } \textit{given } 10 \text{ on 1st})]$
$= P(10 \text{ on 1st}) \cdot P(3 \text{ on 2nd, } \textit{given } 10 \text{ on 1st})$
$= \left(\frac{4}{52}\right)\left(\frac{4}{51}\right) = \frac{16}{2652} = \frac{4}{663} \approx 0.006$

(d) $P[(3, 10) \textit{ or } (10, 3)] = P(3, 10) + P(10, 3)$ since these 2 outcomes are mutually exclusive
$= \frac{4}{663} + \frac{4}{663} = \frac{8}{663} \approx 0.012$

11. (a) Yes; the key idea here is "with replacement." When the first card drawn is replaced, the sample space is the same when the second card is drawn as it was when the first card was drawn and the second card is in no way influenced by the outcome of the first draw; in fact, it is possible to draw the same card twice. Let the card draws be represented by an (x, y) ordered pair; for example (K, 6) means a king was drawn first, replaced, and then the second card, a "6," was drawn independently of the first.

(b) $P(A, K) = P(A) \cdot P(K)$ because they are independent
$= \left(\frac{4}{52}\right)\left(\frac{4}{52}\right) = \frac{16}{2704} = \frac{1}{169}$

(c) $P(K, A) = P(K) \cdot P(A)$ because they are independent
$= \left(\frac{4}{52}\right)\left(\frac{4}{52}\right) = \frac{16}{2704} = \frac{1}{169}$

(d) $P[(A, K) \textit{ or } (K, A)] = P(A, K) + P(K, A)$ since the 2 outcomes are mutually exclusive when we consider the order
$= \frac{1}{169} + \frac{1}{169} = \frac{2}{169}$

12. (a) Yes; the key idea here is "with replacement." When the first card drawn is replaced, the sample space is the same when the second card is drawn as it was when the first card was drawn and the second card is in no way influenced by the outcome of the first draw; in fact, it is possible to draw the same card twice. Let the card draws be represented by an (x, y) ordered pair; for example (K, 6) means a king was drawn first, replaced, and then the second card, a "6," was drawn independently of the first.

(b) $P(3, 10) = P(3) \cdot P(10)$ because draws are independent
$= \left(\frac{4}{52}\right)\left(\frac{4}{52}\right) = \frac{16}{2704} = \frac{1}{169} \approx 0.0059$

(c) $P(10, 3) = P(10) \cdot P(3)$ because of independence
$= \left(\frac{4}{52}\right)\left(\frac{4}{52}\right) = \frac{16}{2704} = \frac{1}{169} \approx 0.0059$

(d) $P[(3, 10) \textit{ or } (10, 3)] = P(3, 10) + P(10, 3)$ because these outcomes are mutually exclusive
$= \frac{1}{169} + \frac{1}{169} = \frac{2}{169} \approx 0.0118$

13. (a) $P(6 \textit{ or } \text{older}) = P[(6 \text{ to } 9) \textit{ or } (10 \text{ to } 12) \textit{ or } (13 \text{ and over})]$
$= P(6-9) + P(10-12) + P(13+)$ because they are mutually exclusive age groups – no child is both 7 and 11 years old.
$= 27\% + 14\% + 22\% = 63\% = 0.63$

(b) $P(12 \text{ or younger}) = 1 - P(13 \text{ and over}) = 1 - 0.22 = 0.78$

(c) $P(\text{between 6 and 12}) = P[(6 \text{ to } 9) \textit{ or } (10 \text{ to } 12)]$
$= P(6 \text{ to } 9) + P(10 \text{ to } 12)$ because the age groups are mutually exclusive
$= 27\% + 14\% = 41\% = 0.41$

(d) P(between 3 and 9)=P[(3 to 5) *or* (6 to 9)]
$=P$(3 to 5) + P(6 to 9) because age categories are mutually exclusive
$= 22\% + 27\% = 49\% = 0.49$
Answers vary; however, category 10–12 years covers only 3 years while 13 and over covers many more years and many more people, including adults who buy toys for themselves.

14. What we know: P(seniors get flu) = 0.14,
P(younger people get flu) = 0.24
P(senior) = 0.125
Let S denote seniors, so *not* S denotes younger people. Let F denote flu and *not* F denote did not get the flu. So P(F, *given* S) = 0.14, P(F, *given not* S) = 0.24 and P(S) = 0.125 so P(*not* S) = 1 – 0.125 = 0.875. Note the phrases 14% *of* seniors, i.e., they were already seniors, so this is a given condition; and 24% *of* people under 65, i.e., these people were already under 65, so under 65 (younger) is a given condition.

(a) P(person is senior *and* will get flu) = P(S *and* F)
$= P(S) \cdot P(F,\ given\ S) = (0.125)(0.14) = 0.0175$
conditional probability rule

(b) P(person is *not* senior *and* will get flu) = P[(*not* S) *and* F]
$= P(not\ S) \cdot P(F,\ given\ not\ S) = 0.875(0.24) = 0.21$

(c) Here, P(S) = 0.95 so P(*not* S) = 1 – 0.95 = 0.05
(a) P(S *and* F) = P(S) · P(F, *given* S) = (0.95)(0.14) = 0.133
(b) P(*not* S *and* F) = P(*not* S) · P(F, *given not* S) = (0.05)(0.24) = 0.012

(d) Here, P(S) = P(*not* S) = 0.50
(a) P(S *and* F) = P(S) · P(F, *given* S) = 0.50(0.14) = 0.07
(b) P(*not* S *and* F) = P(*not* S) · P(F, *given not* S) = 0.50(0.24) = 0.12

15. What we know: P(polygraph says "lying" when person is lying) = 72%
P(polygraph says "lying" when person is not lying) = 7%
Let L denote that the polygraph results show lying and *not* L denote that the polygraph results show the person is not lying. Let T denote that the person is telling the truth and let *not* T denote that the person is not telling the truth, so P(L, *given not* T) = 72%
P(L, *given* T) = 7%.
We are told whether the person is telling the truth or not; what we know is what the polygraph results are, given the case where the person tells the truth, and given the situation where the person is not telling the truth.

(a) P(T) = 0.90 so P(*not* T) = 0.10
P(polygraph says lying and person tells truth)
=P(L *and* T) = P(T) · P(L, *given* T)
=(0.90)(0.07) = 0.063 = 6.3%

(b) P(*not* T) = 0.10 so P(T) = 0.90
P(polygraph says lying and person is not telling the truth)
=P(L *and not* T) = P(*not* T) · P(L, *given not* T)
=(0.10)(0.72) = 0.072 = 7.2%

(c) P(T) = P(*not* T) = 0.50
(a) P(L *and* T) = P(T) · P(L, *given* T)
= (0.50)(0.07) = 0.035 = 3.5%
(b) P(L *and not* T) = P(*not* T) · P(L, *given not* T)
=(0.50)(0.72) = 0.36 = 36%

(d) $P(\text{T}) = 0.15$ so $P(\textit{not}\ \text{T}) = 1 - P(\text{T}) = 1 - 0.15 = 0.85$

(a) $P(\text{L}\ \textit{and}\ \text{T}) = P(\text{T}) \cdot P(\text{L},\ \textit{given}\ \text{T})$
$= (0.15)(0.07) = 0.0105 = 1.05\%$

(b) $P(\text{L}\ \textit{and not}\ \text{T}) = P(\textit{not}\ \text{T}) \cdot P(\text{L},\ \textit{given not}\ \text{T})$
$= (0.85)(0.72) = 0.612 = 61.2\%$

16. What we know: P(polygraph says "lying" when person is lying) = 72%
P(polygraph says "lying" when person is not lying) = 7%
Let L denote that the polygraph results show lying and *not* L denote that the polygraph results show the person is not lying. Let T denote that the person is telling the truth and let *not* T denote that the person is not telling the truth, so $P(\text{L},\ \textit{given not}\ \text{T}) = 72\%$
$P(\text{L},\ \textit{given}\ \text{T}) = 7\%$
We are told whether the person is telling the truth or not; what we know is what the polygraph results are, given the case where the person tells the truth, and given the situation where the person is not telling the truth.

(a) P(polygraph reports "lying") $= P(\text{L}) = 30\%$
We want to find P(person is lying) $= P(\textit{not}\ \text{T})$
There are two possibilities when the polygraph says the person is lying: either the polygraph is right, or the polygraph is wrong. If the polygraph is right, the polygraph results show "lying" and the person is not telling the truth, i.e., $P(\text{L}\ \textit{and not}\ \text{T})$. If the polygraph is wrong, then the polygraph results show "lying" but, in fact, the person is telling the truth, i.e., $P(\text{L}\ \textit{and}\ \text{T})$. (This is the basic "trick" to this problem, and the idea comes directly from set theory.)

So $P(\text{L}) = P(\text{L}\ \textit{and not}\ \text{T}) + P(\text{L}\ \textit{and}\ \text{T})$
$= [P(\textit{not}\ \text{T}) \cdot P(\text{L},\ \textit{given not}\ \text{T})] + [P(\text{T}) \cdot P(\text{L},\ \textit{given}\ \text{T})]$
using conditional probability rules
$= [P(\textit{not}\ \text{T}) \cdot P(\text{L},\ \textit{given not}\ \text{T})] + \{[1 - P(\textit{not}\ \text{T})] \cdot P(\text{L},\ \textit{given}\ \text{T})\}$
using the complementary event rule to rewrite $P(\text{T})$ as $1 - P(\textit{not}\ \text{T})$
$0.30 = [P(\textit{not}\ \text{T})] \cdot (0.72)] + [(1 - P(\textit{not}\ \text{T})] \cdot (0.07)$
substituting in the known values as given in # 15, and as given above
$= (0.72) \cdot P(\textit{not}\ \text{T}) + [0.07 - (0.07) \cdot P(\textit{not}\ \text{T})]$
$0.30 - 0.07 = (0.72) \cdot P(\textit{not}\ \text{T}) - (0.07) \cdot P(\textit{not}\ \text{T})$
$0.23 = P(\textit{not}\ \text{T})(0.72 - 0.07) = P(\textit{not}\ \text{T})(0.65)$
$\dfrac{0.23}{0.65} = P(\textit{not}\ \text{T})$, or $P(\textit{not}\ \text{T}) \approx 0.354 = 35.4\%$

(b) Here, $P(\text{L}) = 70\% = 0.70$
This is the same as (a) except for the new $P(\text{L})$. Starting from the step in (a) just before we substituted in the numerical values we knew:
$P(\text{L}) = [P(\textit{not}\ \text{T}) \cdot P(\text{L},\ \textit{given not}\ \text{T})] + \{[1 - P(\textit{not}\ \text{T})] \cdot P(\text{L},\ \textit{given}\ \text{T})\}$
$0.70 = P(\textit{not}\ \text{T}) \cdot (0.72) + [1 - P(\textit{not}\ \text{T})] \cdot (0.07)$
$0.70 = (0.72) \cdot P(\textit{not}\ \text{T}) + [0.07 - 0.07 \cdot P(\textit{not}\ \text{T})]$
$0.70 - 0.07 = (0.72 - 0.07) \cdot P(\textit{not}\ \text{T})$
$0.63 = 0.65P(\textit{not}\ \text{T})$
so $P(\textit{not}\ \text{T}) = \dfrac{0.63}{0.65} \approx 0.969 = 96.9\%$.

17. We have $P(A) = \dfrac{580}{1160}$, $P(Pa) = \dfrac{580}{1160} = P(\textit{not}\ A)$, $P(S) = \dfrac{686}{1160}$, $P(N) = \dfrac{474}{1160} = P(\textit{not}\ S)$

(a) $P(S) = \dfrac{686}{1160}$

$P(S,\ \textit{given}\ A) = \dfrac{270}{580}$ (*given A* means stay in the *A*, aggressive row)

$P(S,\ \textit{given}\ Pa) = \dfrac{416}{580}$ (staying in row *Pa*)

(b) $P(S) = \dfrac{686}{1160} = \dfrac{343}{580}$

$P(S, \textit{given } Pa) = \dfrac{416}{580}$

They are not independent since the probabilities are not the same.

(c) $P(A \textit{ and } S) = P(A) \cdot P(S, \textit{given } A)$

$$= \left(\frac{580}{1160}\right)\left(\frac{270}{580}\right) = \frac{270}{1160}$$

$P(Pa \textit{ and } S) = P(Pa) \cdot P(S, \textit{given } Pa)$

$$= \left(\frac{580}{1160}\right)\left(\frac{416}{580}\right) = \frac{416}{1160}$$

(d) $P(N) = \dfrac{474}{1160}$

$P(N, \textit{given } A) = \dfrac{310}{580}$ (stay in the A row)

(e) $P(N) = \dfrac{474}{1160} = \dfrac{237}{580}$

$P(N, \textit{given } A) = \dfrac{310}{580}$

Since the probabilities are not the same, N and A are not independent.

(f) $P(A \textit{ or } S) = P(A) + P(S) - P(A \textit{ and } S)$

$$= \frac{580}{1160} + \frac{686}{1160} - \frac{270}{1160} = \frac{996}{1160}$$

18. (a) $P(+, \textit{given}$ condition present$) = \dfrac{110}{130}$ (stay in "condition present" row)

(b) $P(-, \textit{given}$ condition present$) = \dfrac{20}{130}$ (stay in "condition present" row)

[(a) and (b) are complementary events]

(c) $P(-, \textit{given}$ condition absent$) = \dfrac{50}{70}$ (stay in the row or column of the "*given*")

(d) $P(+, \textit{given}$ condition absent$) = \dfrac{20}{70}$

(e) P(condition present *and* +) $= P$(condition present) $\cdot$ $P(+, \textit{given}$ condition present$)$

$$= \left(\frac{130}{200}\right)\left(\frac{110}{130}\right) = \frac{110}{200}$$

(f) P(condition present *and* –) $= P$(condition present) $\cdot$ $P(-, \textit{given}$ condition present$)$

$$= \left(\frac{130}{200}\right)\left(\frac{20}{130}\right) = \frac{20}{200}$$

19. Let C denote the condition is present, and *not* C denote the condition is absent.

(a) $P(+, \textit{given}$ C$) = \dfrac{72}{154}$ (stay in C column)

(b) $P(-, \textit{given}$ C$) = \dfrac{82}{154}$ (stay in C column)

(c) $P(-, \textit{given not}$ C$) = \dfrac{79}{116}$ (stay in *not* C column)

(d) $P(+, \textit{given not}$ C$) = \dfrac{37}{116}$ (stay in *not* C column)

(e) $P(\text{C and } +) = P(\text{C}) \cdot P(+, \textit{given C}) = \left(\frac{154}{270}\right)\left(\frac{72}{154}\right) = \frac{72}{270}$

(f) $P(\text{C and } -) = P(\text{C}) \cdot P(-, \textit{given C}) = \left(\frac{154}{270}\right)\left(\frac{82}{154}\right) = \frac{82}{270}$

20. First determine the denominator. If it is a row or column total, the numerator will be in the body (inside) of the table in that same row or column. If the denominator is the grand total the numerator can be one or more row totals, one or more column totals, or a body-of-the-table cell entry. A cell entry is usually indicated when the problem mentions both a row category and a column category, in which case the desired cell is the one where the row and column intersect.

(a) customer at random, denominator is grand total, 2008; loyal 10–14 years, numerator is column total, 291; $\frac{291}{2008}$

(b) given: customer is from the East, so denominator is row total, 452; loyal 10–14 years: the cell entry in that row, 77; $\frac{77}{452}$

(c) no qualifiers on the customers, so denominator is grand total, 2008; at least 10 years: need entry for 10–14 years and 15+ years, so numerator is the sum of these 2 column totals, 291 + 535 = 826; $\frac{826}{2008}$

(d) given: from the West means the denominator is the West row total, 373; loyal at least 10 years means we sum the numbers in the West row for 10–14 and 15+ years, 45 + 86 = 131; $\frac{131}{373}$

(e) given: loyal less than 1 year means the denominator is column total, 157; from the West means the numerator is the cell entry for West in that column, 41; $\frac{41}{157}$

(f) given: loyal < 1 year, so denominator is 157; from South, so numerator is at the intersection of the < 1 year column and the South row, 53; $\frac{53}{157}$

(g) given: East, so denominator is 452; loyal 1+ years: either add up all the entries except < 1 year in the East row, or use the complementary event rule (less work!);
$P(\text{loyal 1+ years}, \textit{given} \text{ East}) = 1 - P(\text{loyal} < 1 \text{ year}, \textit{given} \text{ East})$

$$= 1 - \frac{32}{452} = \frac{420}{452}$$

(h) given: West, so denominator is West row total, 373; loyal 1+ years is either the sum of all the West row entries except < 1 year, or apply the complementary event rule in the probability calculation;
$P(\text{loyal 1+ years}, \textit{given} \text{ West}) = 1 - P(\text{loyal} < 1 \text{ year}, \textit{given} \text{ West})$

$$= 1 - \frac{41}{373} = \frac{332}{373}$$

(i) $P(\text{East}) = \frac{452}{2008} \approx 0.2251$

$P(\text{loyal 15+ years}) = \frac{535}{2008} \approx 0.2664$

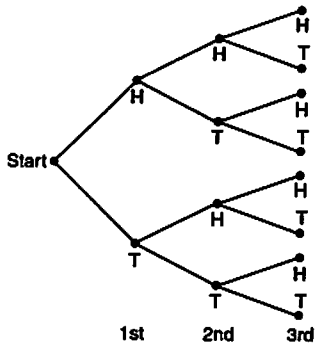
$P(\text{East}, \textit{given} \text{ 15+ years}) = \frac{118}{535} \approx 0.2206$

$P(\text{loyal 15+ years}, \textit{given} \text{ East}) = \frac{118}{452} \approx 0.2611$

If they are independent, $P(\text{East}) = P(\text{East}, \textit{given}$ 15+ years) but $0.2251 \neq 0.2206$, and if they are independent, $P(\text{loyal 15+ years}) = P(\text{loyal 15+ years}, \textit{given}$ East) but $0.2664 \neq 0.2611$, so they aren't independent. (If you use decimal approximations, and the 2 probabilities are quite close, it's time to reduce fractions or use the least common denominator to get an accurate comparison.)
Note: independence is symmetric, i.e., if A is independent of B, then B is independent of A; this means you don't have to do *both* independence checks; one is sufficient.

Section 5.3

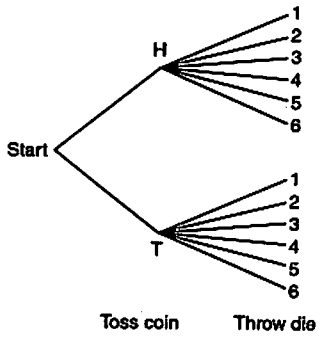
1. (a) Outcomes for Tossing a Coin Three Times



(b) HHT, HTH, THH: 3

(c) 8 possible outcomes, 3 with exactly 2 Hs: $\frac{3}{8}$

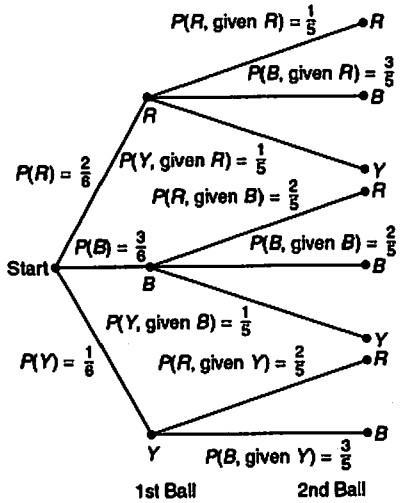
2. (a) Outcomes of Tossing a Coin and Throwing a Die



(b) outcomes with H and > 4
H5, H6:2

(c) 12 outcomes, two with H and > 4: $\frac{2}{12} = \frac{1}{6}$

3. (a) Outcomes for Drawing Two Balls (without replacement)



Because we drew without replacement the number of available balls drops to 5 and one of the colors drops by 1. Note that if the yellow ball is drawn first, there are only two possibilities for the second draw: red and blue; the yellow balls are exhausted.

(b) $P(R, R) = \left(\frac{2}{6}\right)\left(\frac{1}{5}\right) = \frac{2}{30} = \frac{1}{15}$

$P(R, B) = \left(\frac{2}{6}\right)\left(\frac{3}{5}\right) = \frac{6}{30} = \frac{1}{5}$

$P(R, Y) = \left(\frac{2}{6}\right)\left(\frac{1}{5}\right) = \frac{2}{30} = \frac{1}{15}$

$P(B, R) = \left(\frac{3}{6}\right)\left(\frac{2}{5}\right) = \frac{6}{30} = \frac{1}{5}$

$P(B, B) = \left(\frac{3}{6}\right)\left(\frac{2}{5}\right) = \frac{6}{30} = \frac{1}{5}$

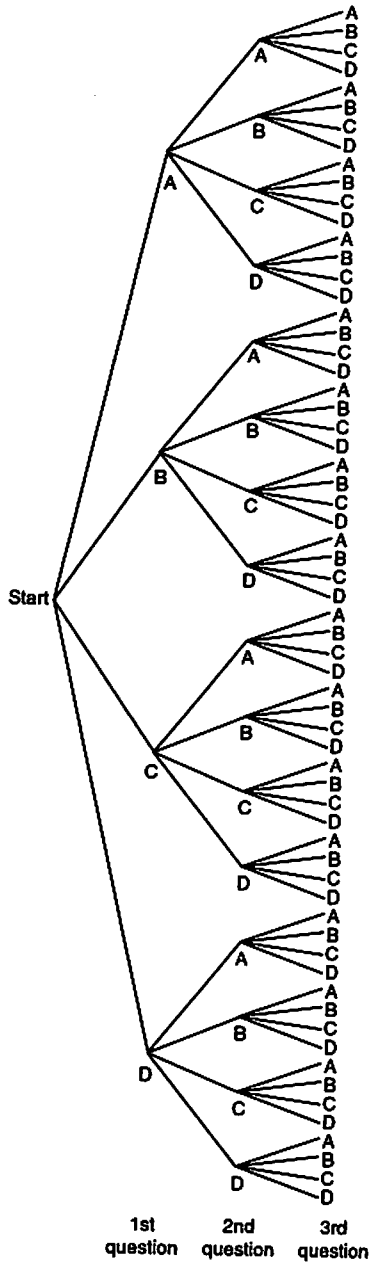
$P(B, Y) = \left(\frac{3}{6}\right)\left(\frac{1}{5}\right) = \frac{3}{30} = \frac{1}{10}$

$P(Y, R) = \left(\frac{1}{6}\right)\left(\frac{2}{5}\right) = \frac{2}{30} = \frac{1}{15}$

$P(Y, B) = \left(\frac{1}{6}\right)\left(\frac{3}{5}\right) = \frac{3}{30} = \frac{1}{10}$

where $P(x, y)$ is the probability the first ball is color x, and the second ball is color y. Multiply the branch probability values along each branch from start to finish. Observe the sum of the probabilities is 1.

4. (a) Outcomes of Three Multiple-Choice Questions



This is a gaudier version of problems 1 and 5 where there are 3 questions, but now there are 4 responses (A, B, C, D) for each question at each step.

(b) If the outcomes are equally likely, then $P(\text{all 3 correct}) = \left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{64}$.

5. 4 wire choices for the first leaves 3 wire choices for the second, 2 for the third, and only 1 wire choice for the fourth wire connection: $4 \cdot 3 \cdot 2 \cdot 1 = 4! = 24$.

6. 4 choices for his first stop, 3 for the second, 2 for the third, and only 1 city for his (last) fourth stop: $4 \cdot 3 \cdot 2 \cdot 1 = 4! = 24$. This problem is identical to problem 7 except wires were changed to cities.

7. **(a)** Choose 1 card from each deck. The number of pairs (one card from the first deck and one card from the second) is $52 \cdot 52 = 52^2 = 2704$.

(b) There are 4 kings in the first deck and four in the second, so $4 \cdot 4 = 16$.

(c) There are 16 ways to draw a king from each deck, and 2704 ways to draw a card from each deck, so $\frac{16}{2704} = \frac{1}{169} \approx 0.006$.

8. **(a)** The die rolls are independent, so multiply the 6 ways the first die can land by the 6 ways the second die can land: $6 \cdot 6 = 36$.

(b) Even numbers are 2, 4, and 6, three possibilities per die, so $3 \cdot 3 = 9$.

(c) $P(\text{even, even}) = \frac{9}{36} = \frac{1}{4} = 0.25$

using $P(\text{event}) = \frac{\text{number of favorable outcomes}}{\text{total number of outcomes}}$

Problems 9–12 deal with permutations, $P_{n,r} = \frac{n!}{(n-r)!}$. This counts the number of ways r objects can be selected from n when the order of the result is important. For example, if we choose two people from a group, the first of which is to be the group's chair, and the second, the assistant chair, then (John, Mary) is distinct from (Mary, John).

9. $P_{5,2}$: $n = 5$, $r = 2$

$$P_{5,2} = \frac{5!}{(5-2)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3!} = 20$$

10. $P_{8,3}$: $n = 8$, $r = 3$

$$P_{8,3} = \frac{8!}{(8-3)!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{5!} = \frac{8 \cdot 7 \cdot 6 \cdot 5!}{5!} = 336$$

11. $P_{7,7}$: $n = r = 7$

$$P_{7,7} = \frac{7!}{(7-7)!} = \frac{7!}{0!} = 7! = 5040 \text{ (recall } 0! = 1)$$

In general, $P_{n,n} = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1} = n!$.

12. $P_{9,9}$: $n = r = 9$

$$P_{9,9} = \frac{9!}{(9-9)!} = \frac{9!}{0!} = \frac{9!}{1} = 362{,}880$$

In general, $P_{n,n} = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1} = n!$.

Problems 13–16 deal with combination, $C_{n,r} = \frac{n!}{r!(n-r)!}$. This counts the number of ways r items can be selected from among n items when the order of the result doesn't matter. For example, when choosing two people from an office to pick up coffee and doughnuts, (John, Mary) is the same as (Mary, John) — both get to carry the goodies back to the office.

13. $C_{5,2}$: $n = 5$, $r = 2$

$$C_{5,2} = \frac{5!}{2!(5-2)!} = \frac{5!}{2!3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1} = \frac{20}{2} = 10$$

14. $C_{8,3}$: $n = 8$, $r = 3$

$$C_{8,3} = \frac{8!}{3!(8-3)!} = \frac{8!}{3!5!} = \frac{8 \cdot 7 \cdot 6 \cdot 5!}{3 \cdot 2 \cdot 1 \cdot 5!} = 56$$

15. $C_{7,7}$: $n = r = 7$

$$C_{7,7} = \frac{7!}{7!(7-7)!} = \frac{7!}{7!0!} = \frac{7!}{7!(1)} = 1 \text{ (recall } 0! = 1)$$

In general, $C_{n,n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!0!} = \frac{n!}{n!(1)} = 1$. There is only 1 way to choose all n objects without regard to order.

16. $C_{8,8}$: $n = r = 8$

$$P_{9,9} = \frac{8!}{8!(8-8)!} = \frac{8!}{8!0!} = \frac{8!}{8!(1)} = 1 \text{ (recall } 0! = 1)$$

In general, $C_{n,n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!0!} = \frac{n!}{n!(1)} = 1$. There is only 1 way to choose all n objects without regard to order.

17. Since the order matters (first is day supervisor, second is night supervisor, and third is coordinator), this is a permutation of 15 nurse candidates to fill 3 positions.

$$P_{15,3} = \frac{15!}{(15-3)!} = \frac{15!}{12!} = \frac{15 \cdot 14 \cdot 13 \cdot 12!}{12!} = 2730$$

18. The order of the software packages selected doesn't matter, since all three are going home with the customer. (Assume the software packages are of equal interest to the customer.)

$$C_{10,3} = \frac{10!}{3!(10-3)!} = \frac{10!}{3!7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{3!7!} = \frac{720}{6} = 120$$

19. The order of trainee selection doesn't matter, since they are all going to be trained the same.

$$C_{15,5} = \frac{15!}{5!(15-5)!} = \frac{15!}{5!10!} = \frac{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11 \cdot 10!}{5!10!} = \frac{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 3003$$

20. It doesn't matter in which order the professor grades the problems, the 5 selected problems all get graded.

(a) $C_{12,5} = \frac{12!}{5!(12-5)!} = \frac{12!}{5!7!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7!}{5!7!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 792$

(b) Jerry must have the very same 5 problems as the professor selected to grade, so

$P(\text{Jerry chose the right problems}) = \frac{1}{792} \approx 0.001$. (Jerry is pushing his luck.)

(c) Silvia did seven problems, which have $C_{7,5}$ subsets of 5 problems which would be among 792 subsets of 5 the professor selected from.

$$C_{7,5} = \frac{7!}{5!(7-5)!} = \frac{7!}{5!2!} = \frac{7\cdot 6\cdot 5!}{5!2!} = \frac{7\cdot 6}{2\cdot 1} = \frac{42}{2} = 21$$

$$P(\text{Silvia lucked out}) = \frac{21}{792} \approx 0.027$$

(Silvia is pushing her luck, too, but she increased her chances by a factor of 21, compared to Jerry, just by doing two more problems. Now, if these two had just done all the problems, or even split them half and half, ...)

21. (a) Six applicants are selected from among 12 without regard to order.

$$C_{12,6} = \frac{12!}{6!6!} = \frac{479{,}001{,}600}{(720)^2} = 924$$

(b) There are 7 women and 5 men. This problem really asks, in how many ways can 6 women be selected from among 7, and zero men be selected from 5?

$$(C_{7,6})(C_{5,0}) = \left(\frac{7!}{6!(7-6)!}\right)\left(\frac{5!}{0!(5-0)!}\right) = \frac{7!}{6!1!}\cdot\frac{5!}{0!5!} = 7$$

Since the zero men are "selected" be default, all positions being filled. This problem reduces to, in how many ways can 6 applicants be selected from 7 women?

(c) $$P(\text{event A}) = \frac{\text{number of favorable outcomes}}{\text{total number of outcomes}}$$

$$P(\text{all hired are women}) = \frac{7}{924} = \frac{1}{132} \approx 0.008$$

22. It doesn't matter in which order you or the state select the 6 numbers each. It only matters that you and the state pick the *same* six numbers. While you spend your zillion dollars, you can always reorder your numbers if you want to.

(a) $$C_{42,6} = \frac{42!}{6!(42-6)!} = \frac{42!}{6!36!} = \frac{42\cdot 41\cdot 40\cdot 39\cdot 38\cdot 37\cdot 36!}{6!36!} = \frac{3{,}776{,}965{,}920}{720} = 5{,}245{,}786$$

(Most calculators will handle numbers through 69! But if you hate to see numbers like 1.771×10^{98}, cancel out the common factorial factors, such as 36! here.)

(b) This problem asks, what is the chance you choose the very same 6 numbers the state chose.

$$P(\text{winning ticket}) = \frac{1}{5{,}245{,}786} \approx 0.000000191$$

(c) What is the chance one of your 10 tickets is the winning ticket? (We'll assume each ticket is different from the other 9 you have, but, it really doesn't matter much ...)

$$P(\text{win}) = \frac{10}{5{,}245{,}786} = \frac{5}{2{,}622{,}893} \approx 0.0000019$$

Chapter 5 Review

1. $P(\text{asked}) = 24\% = 0.24$
   $P(\text{received, } \textit{given} \text{ asked}) = 45\% = 0.45$
   $P(\text{asked } \textit{and} \text{ received}) = P(\text{asked}) \cdot P(\text{received, } \textit{given} \text{ asked}) = (0.24)(0.45) = 0.108 = 10.8\%$

2. $P(\text{asked}) = 20\% = 0.20$
   $P(\text{received, } \textit{given} \text{ asked}) = 59\% = 0.59$
   $P(\text{asked } \textit{and} \text{ received}) = P(\text{asked}) \cdot P(\text{received, } \textit{given} \text{ asked}) = (0.20)(0.59) = 0.118 = 11.8\%$

3. (a) If the first card is replaced before the second is chosen (sampling with replacement), they are independent. If the sampling is without replacements they are dependent.

   (b) $P(\text{heart}) = \dfrac{13}{52} = \dfrac{1}{4}$

   with replacement, independent

   $P(\text{H on both}) = \left(\dfrac{1}{4}\right)\left(\dfrac{1}{4}\right) = \dfrac{1}{16} = 0.0625 \approx 0.063$

   (c) without replacement, dependent

   $P(\text{H on first } \textit{and} \text{ H on second}) = \dfrac{13}{52} \cdot \dfrac{12}{51} = \dfrac{156}{2652} \approx 0.059$

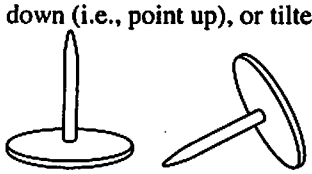
4. (a) There are 11 other outcomes besides 3H. The sample space is the 12 outcomes shown.

1H	1T
2H	2T
(3H)	3T
4H	4T
5H	5T
6H	6T

   (b) Yes; the die and the coin are independent. Each outcome has probability $\left(\dfrac{1}{6}\right)\left(\dfrac{1}{2}\right) = \dfrac{1}{12}$.

   (c) $P(\text{H } \textit{and} \text{ number} < 3) = P[\text{H } \textit{and} \text{ (1 } \textit{or} \text{ 2)}] = P(\text{1H } \textit{or} \text{ 2H}) = P(\text{1H}) + P(\text{2H}) = \dfrac{1}{12} + \dfrac{1}{12} = \dfrac{2}{12} = \dfrac{1}{6} \approx 0.167$

5. (a) Throw a large number of similar thumbtacks, or one thumbtack a large number of times, and record the frequency of occurrence of the various outcomes. Assume the thumbtack falls either flat side down (i.e., point up), or tilted (with the point down, resting on the edge of the flat side). (We will

   assume these are the only two ways the tack can land.) To estimate the probability the tack lands on its flat side with the point up, find the relative frequency of this occurrence, dividing the number of times this occurred by the total number of thumbtack tosses.

   (b) The sample space is the two outcomes flat side down (point up) and tilted (point down).

   (c) $P(\text{flat side down, point up}) = \dfrac{340}{500} = 0.68$

   $P(\text{tilted, point down}) = 1 - 0.68 = 0.32$

6. (a) $P(N) = \dfrac{470}{1000} = 0.470$

$P(M) = \dfrac{390}{1000} = 0.390$

$P(S) = \dfrac{140}{1000} = 0.140$

(b) $P(N, \text{given } W) = \dfrac{420}{500} = 0.840$

$P(S, \text{given } W) = \dfrac{20}{500} = 0.040$

(c) $P(N, \text{given } A) = \dfrac{50}{500} = 0.100$

$P(S, \text{given } A) = \dfrac{120}{500} = 0.240$

(d) $P(N \text{ and } W) = P(W) \cdot P(N, \text{ given } W)$

$$= \left(\frac{500}{1000}\right)(0.840) = 0.420$$

$P(M \text{ and } W) = P(W) \cdot P(M, \text{ given } W)$

$$= \left(\frac{500}{1000}\right)\left(\frac{60}{500}\right) = 0.060$$

(e) $P(N \text{ or } M) = P(N) + P(M)$ if mutually exclusive

$$= \left(\frac{470}{1000}\right) + \left(\frac{390}{1000}\right) = \frac{860}{1000} = 0.860$$

They are mutually exclusive because the reactions are defined into 3 distinct, mutually exclusive categories, a reaction can't be both mild and non-existent.

(f) If N and W were independent, $P(N \text{ and } W) = P(N) \cdot P(W) = (0.470)(0.500) = 0.235$. However, from (d), we have $P(N \text{ and } W) = 0.420$. They are not independent.

7. (a) possible values for x, the sum of the two dice faces, is 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12

(b) 2:1 and 1
3:1 and 2, or 2 and 1
4:1 and 3, 2 and 2, 3 and 1
5:1 and 4, 2 and 3, 3 and 2, 4 and 1
6:1 and 5, 2 and 4, 3 and 3, 4 and 2, 5 and 1
7:1 and 6, 2 and 5, 3 and 4, 4 and 3, 5 and 2, 6 and 1
8:2 and 6, 3 and 5, 4 and 4, 5 and 3, 6 and 2
9:3 and 6, 4 and 5, 5 and 4, 6 and 3
10:4 and 6, 5 and 5, 6 and 4
11:5 and 6, 6 and 5
12:6 and 6

x	$P(x)$
2	$\frac{1}{36} \approx 0.028$
3	$\frac{2}{36} \approx 0.056$
4	$\frac{3}{36} \approx 0.083$
5	$\frac{4}{36} \approx 0.111$
6	$\frac{5}{36} \approx 0.139$
7	$\frac{6}{36} \approx 0.167$
8	$\frac{5}{36} \approx 0.139$
9	$\frac{4}{36} \approx 0.111$
10	$\frac{3}{36} \approx 0.083$
11	$\frac{2}{36} \approx 0.056$
12	$\frac{1}{36} \approx 0.028$

Where there are $(6)(6) = 36$ possible, equally likely outcomes (the sums, however, are not equally likely).

8. $P(\text{pass } 101) = 0.77$
$P(\text{pass } 102, \textit{given} \text{ pass } 101) = 0.90$
$P(\text{pass } 101 \textit{ and} \text{ pass } 102) = P(\text{pass } 101) \cdot P(\text{pass } 102, \textit{given} \text{ pass } 101) = 0.77(0.90) = 0.693$

9. $C_{8,2} = \frac{8!}{2!6!} = \frac{8 \cdot 7 \cdot 6!}{(2 \cdot 1)6!} = \frac{56}{2} = 28$

10. (a) $P_{7,2} = \frac{7!}{(7-2)!} = \frac{7!}{5!} = 7(6) = 42$

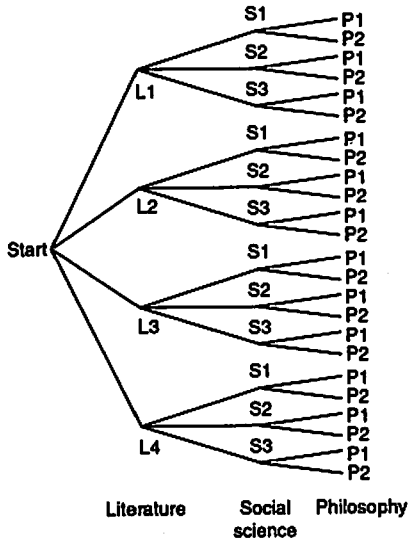
(b) $C_{7,2} = \frac{7!}{2!5!} = \frac{7 \cdot 6}{2} = 21$

(c) $P_{3,3} = \dfrac{3!}{(3-3)!} = \dfrac{3!}{0!} = 6$

(d) $C_{4,4} = \dfrac{4!}{4!(4-4)!} = \dfrac{4!}{4!0!} = 1$

11. $3 \cdot 2 \cdot 1 = 6$

12. Ways to Satisfy Literature, Social Science, and Philosophy Requirements



Let Li, $i = 1, \cdots, 4$ denote the 4 literature courses.
Let Si, $i = 1, 2, 3$ denote the 3 social science courses.
Let Pi, $i = 1, 2$ denote the 2 philosophy courses.
There are $4 \cdot 3 \cdot 2 = 24$ possible course combinations.

13. 5 multiple choice questions, each with 4 possible answers (A, B, C, or D), so 4 answers for first question; and for each of those, 4 answers for the second question; and for each of those, 4 answers for the third question; and for each of those, 4 answers for the fifth question. There are $4 \cdot 4 \cdot 4 \cdot 4 \cdot 4 = 4^5 = 1024$ possible sequences, such as A, D, B, B, C or C, B, A, D, D, etc.

P(getting the correct sequence) $= \dfrac{1}{1024} \approx 0.00098$

14. Two possible outcomes per coin toss; 6 tosses to get a sequence such as THTHHT
$2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^6 = 64$ possible sequences.