Exercices tirés du livre
Introduction to probability and statistics 13th edition
de Mendenhall, Beaver et Beaver.
STT1700
(Automne 2008)

Chapitre 1:

1.1, 1.2, 1.3, 1.5, 1.8, 1.11, 1.13, 1.14, 1.18, 1.19, 1.25, 1.26, 1.30, 1.33, 1.39, 1.41, 1.51, 1.61, 1.67, 2.2, 2.3, 2.4, 2.8, 2.21, 2.24, 2.26, 2.37, 2.38, 2.42, 2.47, 2.51, 2.54, 2.55, 2.65, 2.79, 2.80, 3.11, 3.16, 3.39.

1: Describing Data with Graphs

1.1 **a** The experimental unit, the individual or object on which a variable is measured, is the student.

b The experimental unit on which the number of errors is measured is the exam.

c The experimental unit is the patient.

d The experimental unit is the azalea plant.

e The experimental unit is the car.

1.2 **a** "Time to assemble" is a *quantitative* variable because a numerical quantity (1 hour, 1.5 hours, etc.) is measured.

b "Number of students" is a *quantitative* variable because a numerical quantity (1, 2, etc.) is measured.

c "Rating of a politician" is a *qualitative* variable since a quality (excellent, good, fair, poor) is measured.

d "State of residence" is a *qualitative* variable since a quality (CA, MT, AL, etc.) is measured.

1.3 **a** "Population" is a *discrete* variable because it can take on only integer values.

b "Weight" is a *continuous* variable, taking on any values associated with an interval on the real line.

c "Time" is a *continuous* variable.

d "Number of consumers" is integer-valued and hence *discrete*.

1.5 **a** The experimental unit, the item or object on which variables are measured, is the vehicle.

b Type (qualitative); make (qualitative); carpool or not? (qualitative); one-way commute distance (quantitative continuous); age of vehicle (quantitative continuous)

c Since five variables have been measured, this is *multivariate data*.

1.8 **a-b** The variable "survival time" is a quantitative continuous variable.

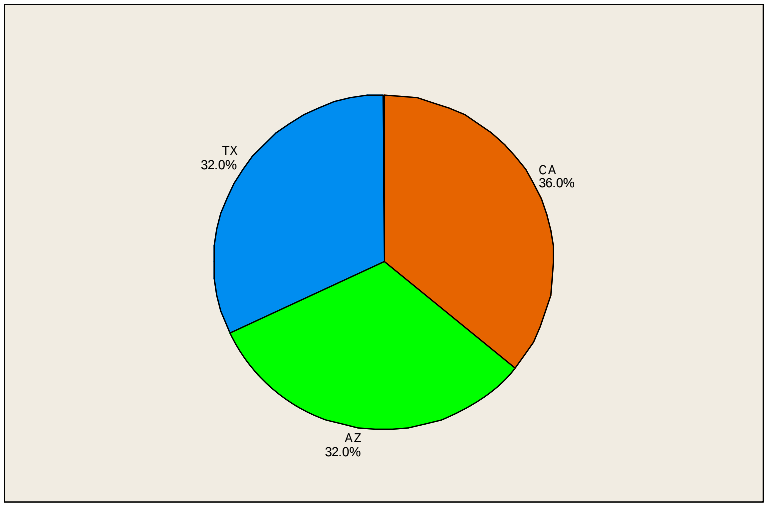
c The population of interest is the population of survival times for all patients having a particular type of cancer and having undergone a particular type of radiotherapy.

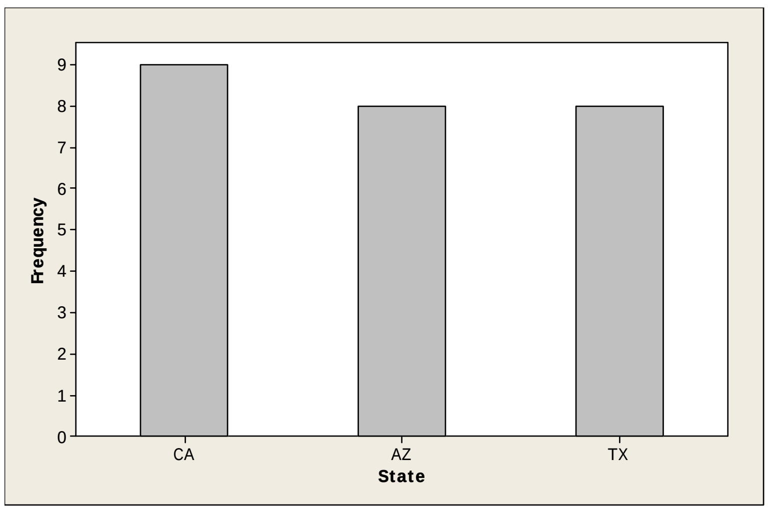
d-e Note that there is a problem with sampling in this situation. If we sample from all patients having cancer and radiotherapy, some may still be living and their survival time will not be measurable. Hence, we cannot sample directly from the population of interest, but must arrive at some reasonable alternate population from which to sample.

1.11 **a-b** The experimental unit is the pair of jeans, on which the qualitative variable "state" is measured.

c-d Construct a statistical table to summarize the data. The pie and bar charts are shown in the figures below.

State	Frequency	Fraction of Total	Sector Angle
CA	9	.36	129.6
AZ	8	.32	115.2
TX	8	.32	115.2





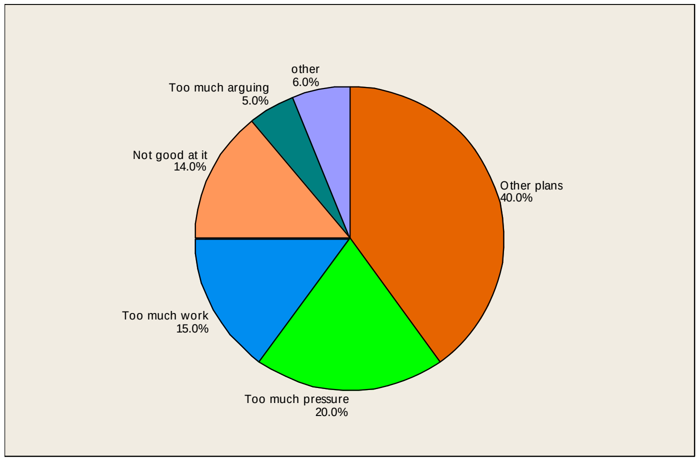
e From the table or the chart, Texas produced $8/25 = 0.32$ of the jeans.

f The highest bar represents California, which produced the most pairs of jeans.

g Since the bars and the sectors are almost equal in size, the three states produced roughly the same number of pairs of jeans.

1.13 **a** The percentages given in the exercise only add to 94%. We should add another category called "Other", which will account for the other 6% of the responses.

b Either type of chart is appropriate. Since the data is already presented as percentages of the whole group, we choose to use a pie chart, shown in the figure below.

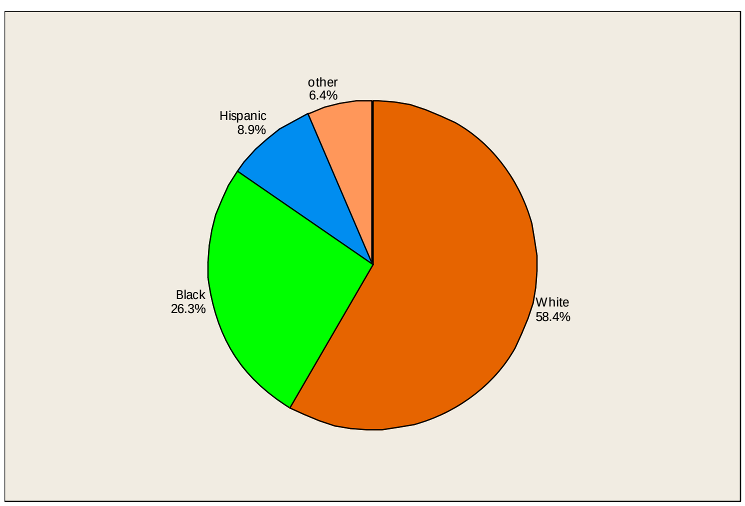


c-d Answers will vary.

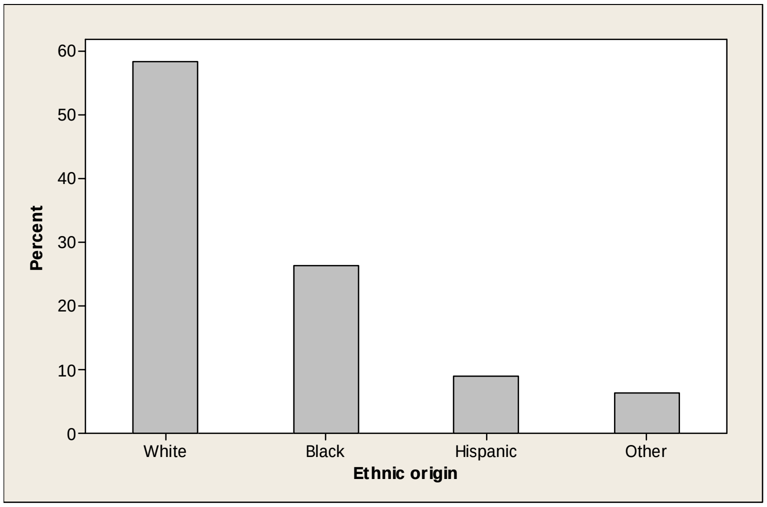
1.14 **a-b** The variable being measured is a qualitative variable, which would be described as "ethnic origin."

c The numbers represent the percentages of Army and Air Force members who fall in each of the four categories.

d-e The percentages falling in each of the four categories have already been calculated, and the pie chart and bar charts are shown in the figures below.



Army



Air Force

f Use the pie chart for the Army. The appropriate percentage for the Army is $26.3\% + 8.9\% + 6.4\% = 41.6\%$. For the Air Force (the bar chart), the percentage of minorities is $16.2\% + 5.0\% + 3.3\% = 24.5\%$.

1.18 The most obvious choice of a stem is to use the ones digit. The portion of the observation to the right of the ones digit constitutes the leaf. Observations are classified by row according to stem and also within each stem according to relative magnitude. The stem and leaf display is shown below.

```
1   6 8
2   1 2 5 5 5 7 8 8 9 9
3   1 1 4 5 5 6 6 6 7 7 7 7 8 9 9 9        leaf digit = 0.1
4   0 0 0 1 2 2 3 4 5 6 7 8 9 9 9          1 2  represents 1.2
5   1 1 6 6 7
6   1 2
```

a The stem and leaf display has a mound shaped distribution.

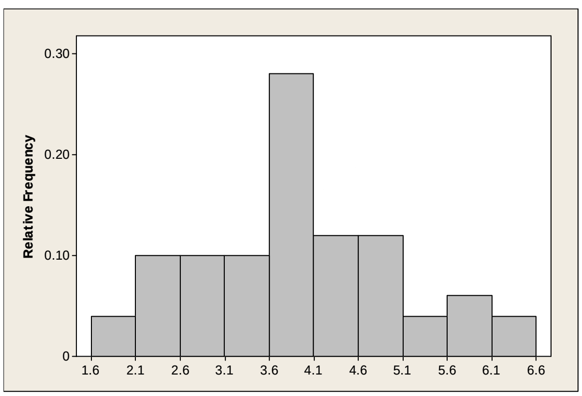
b From the stem and leaf display, the smallest observation is 1.6 (1 6).

c The eight and ninth largest observations are both 4.9 (4 9).

1.19 **a** For $n = 5$, use between 8 and 10 classes.

b

Class i	**Class Boundaries**	**Tally**	f_i	**Relative frequency, f_i/n**
1	1.6 to < 2.1	11	2	.04
2	2.1 to < 2.6	11111	5	.10
3	2.6 to < 3.1	11111	5	.10
4	3.1 to < 3.6	11111	5	.10
5	3.6 to < 4.1	11111 11111 1111	14	.28
6	4.1 to < 4.6	11111 11	7	.14
7	4.6 to < 5.1	11111	5	.10
8	5.1 to < 5.6	11	2	.04
9	5.6 to < 6.1	111	3	.06
10	6.1 to < 6.6	11	2	.04



c From **b**, the fraction less than 5.1 is that fraction lying in classes 1-7, or

$$(2+5+\cdots+7+5)/50 = 43/50 = 0.86$$

d From **b**, the fraction larger than 3.6 lies in classes 5-10, or.

$$(14+7+\cdots+3+2)/50 = 33/50 = 0.66$$

e The stem and leaf display has a more peaked mound-shaped distribution than the relative frequency histogram because of the smaller number of groups.

1.25 **a** The test scores are graphed using a stem and leaf plot generated by ***Minitab***.

Stem-and-Leaf Display: Scores

Stem-and-leaf of Scores N = 20

Leaf Unit = 1.0

```
 2   5  57
 5   6  123
 8   6  578
 9   7  2
(2)  7  56
 9   8  24
 7   8  6679
 3   9  134
```

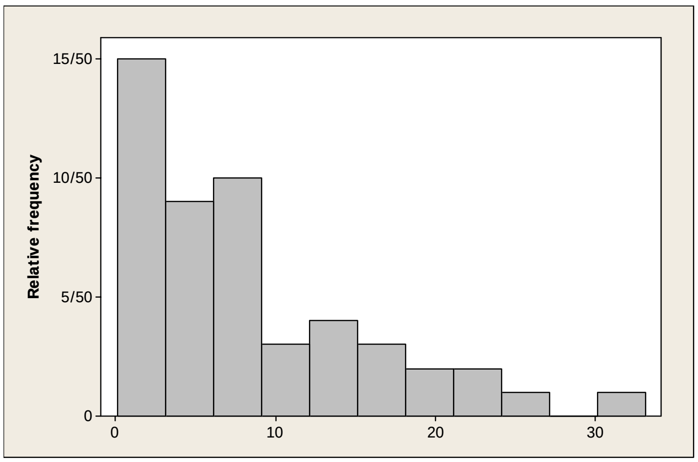
b-c The distribution is not mound-shaped, but is rather bimodal with two peaks centered around the scores 65 and 85. This might indicate that the students are divided into two groups – those who understand the material and do well on exams, and those who do not have a thorough command of the material.

1.26 **a** The range of the data $32.3 - 0.2 = 32.1$. We choose to use eleven class intervals of length 3 ($32.1/11 = 2.9$, which when rounded to the next largest integer is 3). The subintervals 0.1 to < 3.1, 3.1 to < 6.1, 6.1 to < 9.1, and so on, are convenient and the tally is shown below.

Class i	**Class Boundaries**	**Tally**	f_i	**Relative frequency,** f_i/n
1	0.1 to < 3.1	11111 11111 11111	15	15/50
2	3.1 to < 6.1	11111 1111	9	9/50
3	6.1 to < 9.1	11111 11111	10	10/50
4	9.1 to < 12.1	111	3	3/50
5	12.1 to < 15.1	1111	4	4/50
6	15.1 to < 18.1	111	3	3/50
7	18.1 to < 21.1	11	2	2/50

8	21.1 to < 24.1	11	2	2/50
9	24.1 to < 37.1	1	1	1/50
10	27.1 to < 30.1		0	0/50
11	30.1 to < 33.1	1	1	1/50

The relative frequency histogram is shown on the next page.



b The data is skewed to the right, with a few unusually large measurements.

c Looking at the data, we see that 36 patients had a disease recurrence within 10 months. Therefore, the fraction of recurrence times less than or equal to 10 is $36/50 = 0.72$.

1.30 **a** Use the ones digit as the stem, and the portion to the right of the ones digit as the leaf, dividing each stem into two parts.

```
0 | 2 2 3 3 3 4 4 4
0 | 5 5 6 6 6 6 7 7 7 8 8 8 8 9 9
1 | 0 0 1 1 1 1 1 1 1 2 2 2 3 3 3 4 4
1 | 6 6 7 7 8 8 8 8 9 9
2 | 1 2 3
2 | 5 8                    leaf digit = 0.1
3 | 1 1                    1  2 represents 1.2
3 | 6
4 |
4 | 5
5 | 2
```

b Looking at the original data, we see that 25 customers waited one minute or less. Therefore, the fraction of service times less than or equal to one is $25/60 = 0.4167$.

c The smallest measurement is 0 2 which is translated as 0.2.

1.33 **a** Answers will vary.

b The stem and leaf plot is constructed using the tens place as the stem and the ones place as the leaf. ***Minitab*** divides each stem into two parts to create a better descriptive picture. Notice that the distribution is roughly mound-shaped.

Stem-and-Leaf Display: Ages

Stem-and-leaf of Ages N = 38

Leaf Unit = 1.0

2	4	69
3	5	3
7	5	6678
13	6	003344
19	6	567778
19	7	0111234
12	7	7889
8	8	013
5	8	58
3	9	003

c Three of the five youngest presidents – Kennedy, Lincoln and Garfield – were assassinated while in office. This would explain the fact that their ages at death were in the lower tail of the distribution.

1.39 To determine whether a distribution is likely to be skewed, look for the likelihood of observing extremely large or extremely small values of the variable of interest.

a The distribution of non-secured loan sizes might be skewed (a few extremely large loans are possible).

b The distribution of secured loan sizes is not likely to contain unusually large or small values.

c Not likely to be skewed.

d Not likely to be skewed.

e If a package is dropped, it is likely that all the shells will be broken. Hence, a few large number of broken shells is possible. The distribution will be skewed.

f If an animal has one tick, he is likely to have more than one. There will be some “0”s with uninfected rabbits, and then a larger number of large values. The distribution will not be symmetric.

1.41 **a** Weight is continuous, taking any positive real value.

b Body temperature is continuous, taking any real value.

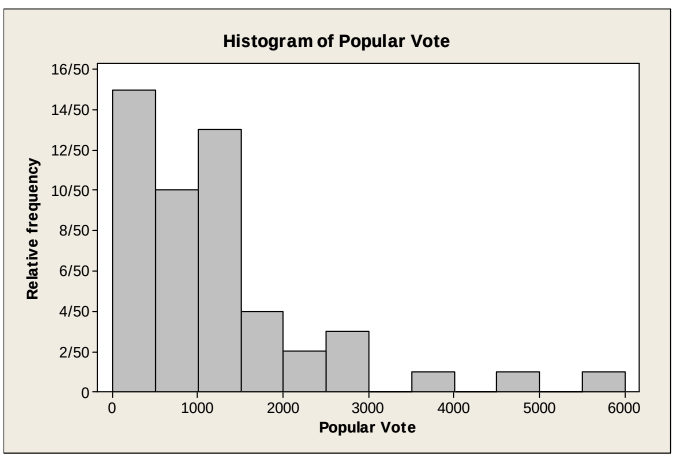
c Number of people is discrete, taking the values 0, 1, 2, …

d Number of properties is discrete.

e Number of claims is discrete.

1.51 **a** The popular vote within each state should vary depending on the size of the state. Since there are several very large states (in population) in the United States, the distribution should be skewed to the right.

b-c Histograms will vary from student to student, but should resemble the histogram generated by ***Minitab*** in the figure below. The distribution is indeed skewed to the right, with three "outliers" – California, Florida and Texas.



1.61 Answers will vary from student to student. Students should notice that both distributions are skewed left. The higher peak with a low bar to its left in the laptop group may indicate that students who would generally receive average scores (65-75) are scoring higher than usual. This may or may not be *caused* by the fact that they used laptop computers.

1.67 **a-b** The distribution is approximately mound-shaped, with one unusual measurement, in the class with midpoint at 100.8°. Perhaps the person whose temperature was 100.8 has some sort of illness coming on?

c The value 98.6° is slightly to the right of center.

2: Describing Data with Numerical Measures

2.2 **a** The mean is

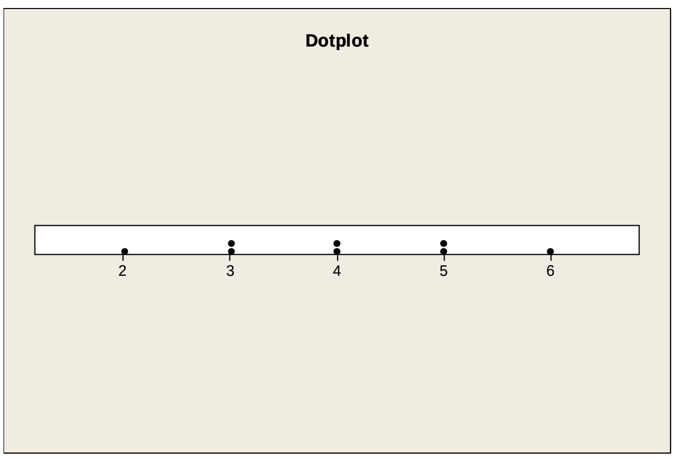
$$\bar{x} = \frac{\sum x_i}{n} = \frac{3+2+\cdots+5}{8} = \frac{32}{8} = 4$$

b To calculate the median, the observations are first ranked from smallest to largest:

2, 3, 3, 4, 4, 5, 5, 6

Since $n = 8$ is even, the position of the median is $0.5(n+1) = 4.5$, and the median is the average of the 4th and 5th measurements, or $m = (4+4)/2 = 4$.

c Since the mean and the median are equal, we conclude that the measurements are symmetric. The dotplot shown below confirms this conclusion.



2.3 **a** $\bar{x} = \frac{\sum x_i}{n} = \frac{58}{10} = 5.8$

b The ranked observations are: 2, 3, 4, 5, 5, 6, 6, 8, 9, 10. Since $n = 10$, the median is halfway between the 5th and 6th ordered observations, or $m = (5+6)/2 = 5.5$.

c There are two measurements, 5 and 6, which both occur twice. Since this is the highest frequency of occurrence for the data set, we say that the set is *bimodal* with modes at 5 and 6.

2.4 **a** $\bar{x} = \frac{\sum x_i}{n} = \frac{9455}{4} = 2363.75$ **b** $\bar{x} = \frac{\sum x_i}{n} = \frac{8280}{4} = 2070$

c The average premium cost in several different cities is not as important to the consumer as the average cost for a variety of consumers in his or her geographical area.

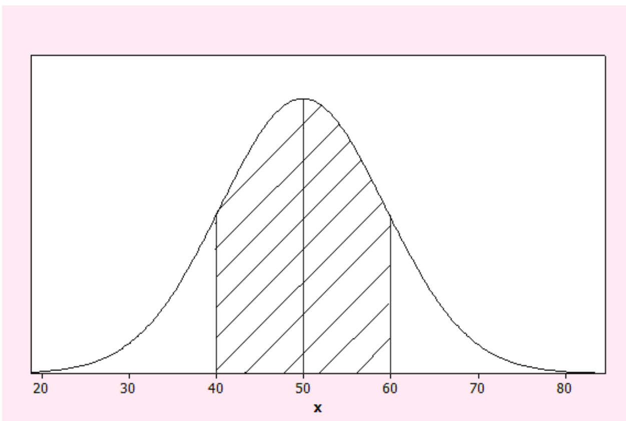
2.8 **a** Similar to previous exercises. The mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{0.99 + 1.92 + \cdots + 0.66}{14} = \frac{12.55}{14} = 0.896$$

b To calculate the median, rank the observations from smallest to largest. The position of the median is $0.5(n+1) = 7.5$, and the median is the average of the 7th and 8th ranked measurement or $m = (0.67 + 0.69)/2 = 0.68$.

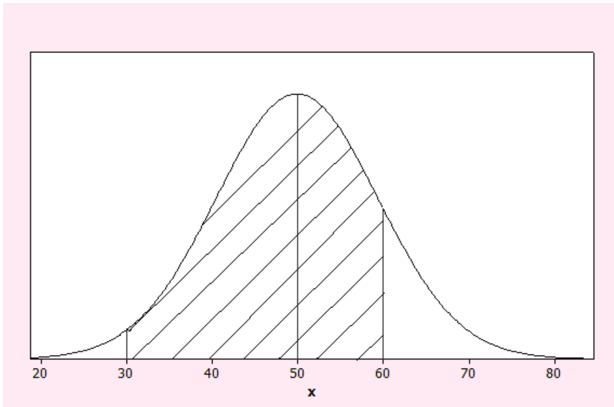
c Since the mean is slightly larger than the median, the distribution is slightly skewed to the right.

2.21 **a** The interval from 40 to 60 represents $\mu \pm \sigma = 50 \pm 10$. Since the distribution is relatively mound-shaped, the proportion of measurements between 40 and 60 is 68% according to the Empirical Rule and is shown below.



b Again, using the Empirical Rule, the interval $\mu \pm 2\sigma = 50 \pm 2(10)$ or between 30 and 70 contains approximately 95% of the measurements.

c Refer to the figure below.



Since approximately 68% of the measurements are between 40 and 60, the symmetry of the distribution implies that 34% of the measurements are between 50 and 60. Similarly, since 95% of the measurements are between 30 and 70, approximately 47.5% are between 30 and 50. Thus, the proportion of measurements between 30 and 60 is

$$0.34 + 0.475 = 0.815$$

d From the figure in part **a**, the proportion of the measurements between 50 and 60 is 0.34 and the proportion of the measurements which are greater than 50 is 0.50. Therefore, the proportion that are greater than 60 must be

$$0.5 - 0.34 = 0.16$$

2.24 **a** The stem and leaf plot generated by ***Minitab*** shows that the data is roughly mound-shaped. Note however the gap in the center of the distribution and the two measurements in the upper tail.

Stem-and-Leaf Display: Weight

Stem-and-leaf of Weight N = 27

Leaf Unit = 0.010

1	7	5
2	8	3
6	8	7999
8	9	23
13	9	66789
13	10	
(3)	10	688
11	11	2244
7	11	788
4	12	4
3	12	8
2	13	
2	13	8
1	14	1

b Calculate $\sum x_i = 28.41$ and $\sum x_i^2 = 30.6071$, the sample mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{28.41}{27} = 1.052$$

and the standard deviation of the sample is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{30.6071 - \frac{(28.41)^2}{27}}{26}} = 0.166$$

c The following table gives the actual percentage of measurements falling in the intervals $\bar{x} \pm ks$ for $k = 1, 2, 3$.

k	$\bar{x} \pm ks$	Interval	Number in Interval	Percentage
1	1.052 ± 0.166	0.866 to 1.218	21	78%
2	1.052 ± 0.332	0.720 to 1.384	26	96%
3	1.052 ± 0.498	0.554 to 1.550	27	100%

d The percentages in part **c** do not agree too closely with those given by the Empirical Rule, especially in the one standard deviation range. This is caused by the lack of mounding (indicated by the gap) in the center of the distribution.

e The lack of any one-pound packages is probably a marketing technique intentionally used by the supermarket. People who buy slightly less than one-pound would be drawn by the slightly lower price, while those who need exactly one-pound of meat for their recipe might tend to opt for the larger package, increasing the store's profit.

2.26 **a** The stem and leaf plots are shown below. The second set has a slightly higher location and spread.

Stem-and-Leaf Display: Method 1, Method 2

Stem-and-leaf of Method 1 N = 10	Stem-and-leaf of Method 2 N = 10
Leaf Unit = 0.00010	Leaf Unit = 0.00010
1 10 0	
3 11 00	1 11 0
4 12 0	3 12 00
(4) 13 0000	5 13 00
2 14 0	5 14 0
1 15 0	4 15 00
	2 16 0
	1 17 0

b *Method 1*: Calculate $\sum x_i = 0.125$ and $\sum x_i^2 = 0.001583$. Then $\bar{x} = \frac{\sum x_i}{n} = 0.0125$ and

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{0.001583 - \frac{(0.125)^2}{10}}{9}} = 0.00151$$

Method 2: Calculate $\sum x_i = 0.138$ and $\sum x_i^2 = 0.001938$. Then $\bar{x} = \frac{\sum x_i}{n} = 0.0138$ and

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}} = \sqrt{\frac{0.001938 - \frac{(0.138)^2}{10}}{9}} = 0.00193$$

The results confirm the conclusions of part **a**.

2.37 **a** Calculate $n = 15$, $\sum x_i = 21$ and $\sum x_i^2 = 49$. Then $\bar{x} = \frac{\sum x_i}{n} = \frac{21}{15} = 1.4$ and

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{49 - \frac{(21)^2}{15}}{14} = 1.4$$

b Using the frequency table and the grouped formulas, calculate

$$\sum x_i f_i = 0(4) + 1(5) + 2(2) + 3(4) = 21$$

$$\sum x_i^2 f_i = 0^2(4) + 1^2(5) + 2^2(2) + 3^2(4) = 49$$

Then, as in part **a**,

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{21}{15} = 1.4$$

$$s^2 = \frac{\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n}}{n-1} = \frac{49 - \frac{(21)^2}{15}}{14} = 1.4$$

2.38 Use the formulas for grouped data given in Exercise 2.37. Calculate $n = 17$, $\sum x_i f_i = 79$, and $\sum x_i^2 f_i = 393$. Then,

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{79}{17} = 4.65$$

$$s^2 = \frac{\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n}}{n-1} = \frac{393 - \frac{(79)^2}{17}}{16} = 1.6176 \text{ and } s = \sqrt{1.6176} = 1.27$$

2.42 The ordered data are:

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

a With $n = 12$, the median is in position $0.5(n+1) = 6.5$, or halfway between the 6th and 7th observations. The lower quartile is in position $0.25(n+1) = 3.25$ (one-fourth of the way between the 3rd and 4th observations) and the upper quartile is in position $0.75(n+1) = 9.75$ (three-fourths of the way between the 9th and 10th observations). Hence,
$m = (5+6)/2 = 5.5,\ Q_1 = 3 + 0.25(4-3) = 3.25$ and $Q_3 = 6 + 0.75(7-6) = 6.75$. Then the five-number summary is

Min	Q_1	Median	Q_3	Max
0	3.25	5.5	6.75	8

and

$$IQR = Q_3 - Q_1 = 6.75 - 3.25 = 3.50$$

b Calculate $n = 12$, $\sum x_i = 57$ and $\sum x_i^2 = 337$. Then $\bar{x} = \frac{\sum x_i}{n} = \frac{57}{12} = 4.75$ and the sample standard deviation is

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{337 - \frac{(57)^2}{12}}{11}} = \sqrt{6.022727} = 2.454$$

c For the smaller observation, $x = 0$,

$$z\text{-score} = \frac{x - \bar{x}}{s} = \frac{0 - 4.75}{2.454} = -1.94$$

and for the largest observation, $x = 8$,

$$z\text{-score} = \frac{x - \bar{x}}{s} = \frac{8 - 4.75}{2.454} = 1.32$$

Since neither *z*-score exceeds 2 in absolute value, none of the observations are unusually small or large.

2.47 **a** The ordered data are shown below:

1.70	101.00	209.00	264.00	316.00	445.00
1.72	118.00	218.00	278.00	318.00	481.00
5.90	168.00	221.00	286.00	329.00	485.00
8.80	180.00	241.00	314.00	397.00	
85.40	183.00	252.00	315.00	406.00	

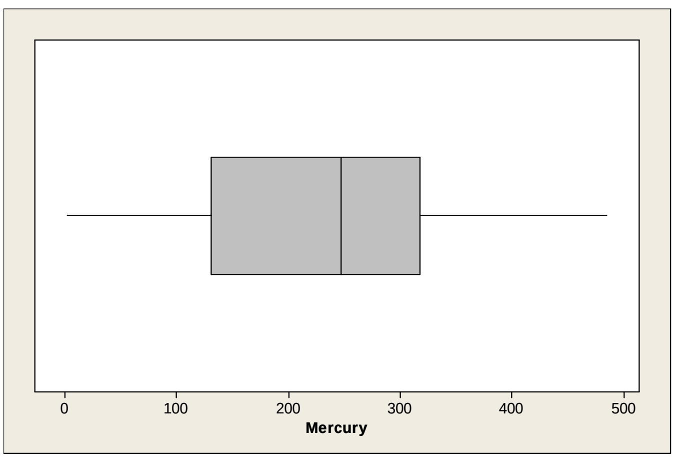
For $n = 28$, the position of the median is $0.5(n+1) = 14.5$ and the positions of the quartiles are $0.25(n+1) = 7.25$ and $0.75(n+1) = 21.75$. The lower quartile is ¼ the way between the 7th and 8th measurements or $Q_1 = 118 + 0.25(168-118) = 130.5$ and the upper quartile is ¾ the way between the 21st and 22nd measurements or $Q_3 = 316 + 0.75(318-316) = 317.5$. Then the five-number summary is

Min	Q_1	Median	Q_3	Max
1.70	130.5	246.5	317.5	485

b Calculate $IQR = Q_3 - Q_1 = 317.5 - 130.5 = 187$. Then the *lower and upper fences* are:

$$Q_1 - 1.5IQR = 130.5 - 280.5 = -150$$
$$Q_3 + 1.5IQR = 317.5 + 280.5 = 598$$

The box plot is shown below. Since there are no outliers, the whiskers connect the box to the minimum and maximum values in the ordered set.



c-d The boxplot does not identify any of the measurements as outliers, mainly because the large variation in the measurements cause the IQR to be large. However, the student should notice the extreme difference in the magnitude of the first four observations taken on young dolphins. These animals have not been alive long enough to accumulate a large amount of mercury in their bodies.

2.51 **a** Just by scanning through the 20 measurements, it seems that there are a few unusually small measurements, which would indicate a distribution that is skewed to the left.

b The position of the median is $0.5(n+1) = 0.5(25+1) = 10.5$ and $m = (120+127)/2 = 123.5$.

The mean is
$$\bar{x} = \frac{\sum x_i}{n} = \frac{2163}{20} = 108.15$$

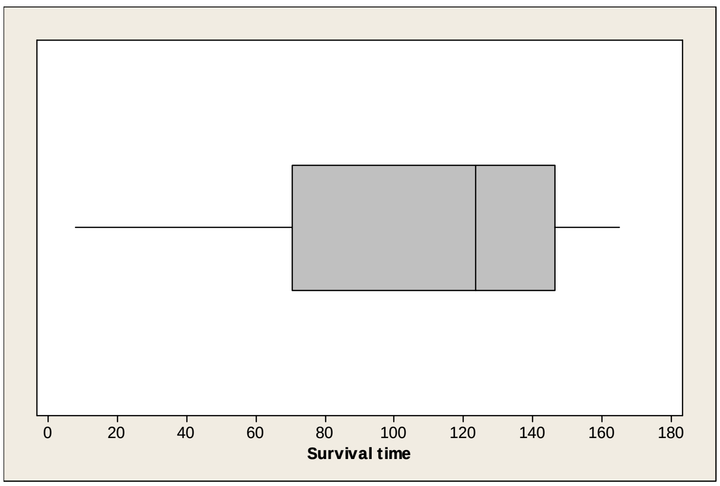
which is smaller than the median, indicate a distribution skewed to the left.

c The positions of the quartiles are $0.25(n+1)=5.25$ and $0.75(n+1)=15.75$, so that $Q_1=65-.25(87-65)=70.5$, $Q_3=144+.75(147-144)=146.25$, and $IQR=146.25-70.5=75.75$.

The *lower and upper fences* are:

$$Q_1-1.5IQR=70.5-113.625=-43.125$$
$$Q_3+1.5IQR=146.25+113.625=259.875$$

The box plot is shown below. There are no outliers. The long left whisker and the median line located to the right of the center of the box indicates that the distribution that is skewed to the left.



2.54 **a** Calculate $n=14$, $\sum x_i=367$ and $\sum x_i^2=9641$. Then $\bar{x}=\frac{\sum x_i}{n}=\frac{367}{14}=26.214$ and

$$s=\sqrt{\frac{\sum x_i^2-\frac{(\sum x_i)^2}{n}}{n-1}}=\sqrt{\frac{9641-\frac{(367)^2}{14}}{13}}=1.251$$

b Calculate $n=14$, $\sum x_i=366$ and $\sum x_i^2=9644$. Then $\bar{x}=\frac{\sum x_i}{n}=\frac{366}{14}=26.143$ and

$$s=\sqrt{\frac{\sum x_i^2-\frac{(\sum x_i)^2}{n}}{n-1}}=\sqrt{\frac{9644-\frac{(366)^2}{14}}{13}}=2.413$$

c The centers are roughly the same; the Sunmaid raisins appear slightly more variable.

2.55 **a** The ordered sets are shown below:

Generic					Sunmaid				
24	25	25	25	26	22	24	24	24	24
26	26	26	26	27	25	25	27	28	28
27	28	28	28		28	28	29	30	

For $n = 14$, the position of the median is $0.5(n+1) = 0.5(14+1) = 7.5$ and the positions of the quartiles are $0.25(n+1) = 3.75$ and $0.75(n+1) = 11.25$, so that

Generic: $m = 26,\ Q_1 = 25,\ Q_3 = 27.25,$ and $IQR = 27.25 - 25 = 2.25$

Sunmaid: $m = 26,\ Q_1 = 24,\ Q_3 = 28,$ and $IQR = 28 - 24 = 4$

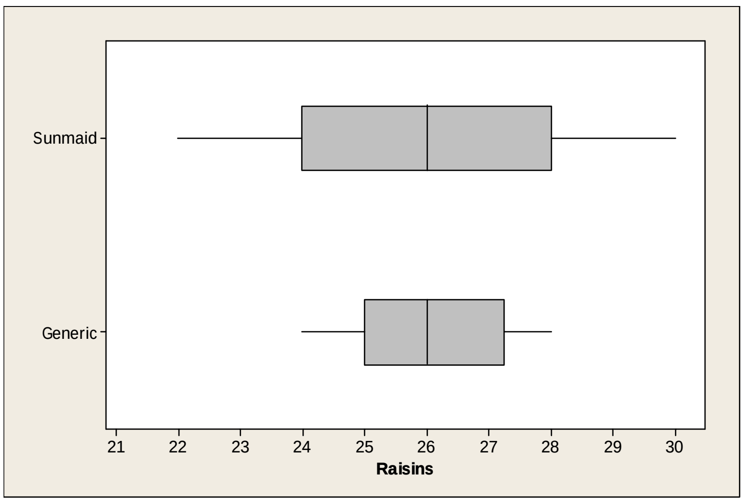
b **Generic:** *Lower and upper fences* are:

$$Q_1 - 1.5IQR = 25 - 3.375 = 21.625$$
$$Q_3 + 1.5IQR = 27.25 + 3.375 = 30.625$$

Sunmaid: *Lower and upper fences* are:

$$Q_1 - 1.5IQR = 24 - 6 = 18$$
$$Q_3 + 1.5IQR = 28 + 6 = 34$$

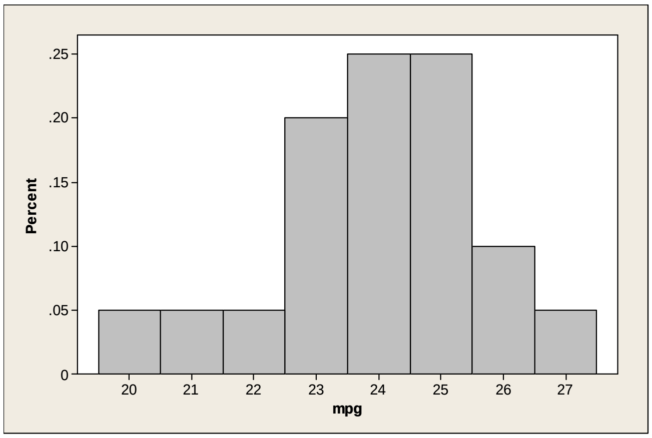
The box plots are shown below. There are no outliers.



d If the boxes are not being underfilled, the average size of the raisins is roughly the same for the two brands. However, since the number of raisins is more variable for the Sunmaid brand, it would appear that some of the Sunmaid raisins are large while others are small. The individual sizes of the generic raisins are not as variable.

2.65 **a** Max = 27, Min = 20.2 and the range is $R = 27 - 20.2 = 6.8$.

b Answers will vary. A typical histogram is shown below. The distribution is slightly skewed to the left.



c Calculate $n = 20$, $\sum x_i = 479.2$, $\sum x_i^2 = 11532.82$. Then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{479.2}{20} = 23.96$$

$$s = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}} = \sqrt{\frac{11532.82 - \frac{(479.2)^2}{20}}{19}} = \sqrt{2.694} = 1.641$$

d The sorted data is shown below:

20.2	21.3	22.2	22.7	22.9
23.1	23.2	23.6	23.7	24.2
24.4	24.4	24.6	24.7	24.7
24.9	25.3	25.9	26.2	27.0

The *z*-scores for *x* = 20.2 and *x* = 27 are

$$z = \frac{x - \bar{x}}{s} = \frac{20.2 - 23.96}{1.641} = -2.29 \text{ and } z = \frac{x - \bar{x}}{s} = \frac{27 - 23.96}{1.641} = 1.85$$

Since neither of the *z*-scores are greater than 3 in absolute value, the measurements are not judged to be outliers.

e The position of the median is $0.5(n+1) = 10.5$ and the median is $m = (24.2 + 24.4)/2 = 24.3$

f The positions of the quartiles are $0.25(n+1) = 5.25$ and $0.75(n+1) = 15.75$. Then $Q_1 = 22.9 + 0.25(23.1 - 22.9) = 22.95$ and $Q_3 = 24.7 + 0.75(24.9 - 24.7) = 24.85$.

2.79 Notice that two (Sosa and McGuire) of the four players have relatively symmetric distributions. The whiskers are the same length and the median line is close to the middle of the box. The variability of the distributions is similar for all four players, but Barry Bonds has a distribution with a long right whisker, meaning that there may be an unusually large number of homers during one of his seasons. The distribution for Babe Ruth is slightly different from the others. The median line to the right of middle indicates a distribution skewed to the left; that there were a few seasons in which his homerun total was unusually low. In fact, the median number of homeruns for the other three players are all about 34-35, while Babe Ruth's median number of homeruns is closer to 40.

2.80 **a** Use the information in the exercise. For 2001, $IQR = 16.5$, and the upper fence is

$$Q_3 + 1.5IQR = 41.50 + 24.75 = 66.25$$

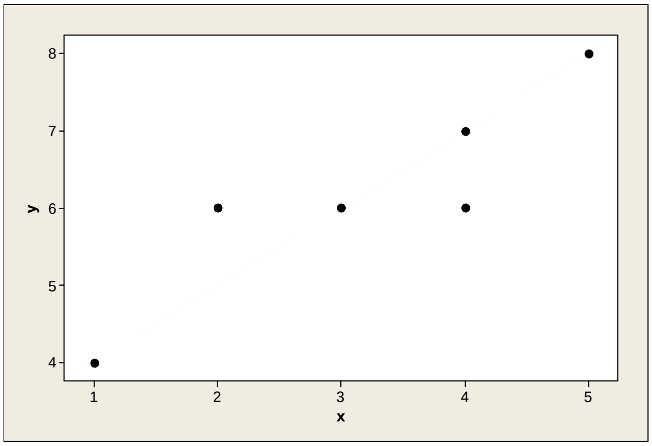
For 2006, $IQR = 20$, and the upper fence is

$$Q_3 + 1.5IQR = 45.00 + 30.00 = 75.00$$

b The upper fence is different in 2006, so that the record number of homers, $x = 73$ is no longer an outlier, although it is still the most homers ever hit in a single season!

3: Describing Bivariate Data

3.11 **a** The first variable (x) is the first number in the pair and is plotted on the horizontal axis, while the second variable (y) is the second number in the pair and is plotted on the vertical axis. The scatterplot is shown in the figure below.



b There appears to be a positive relationship between *x* and *y*; that is, as *x* increases, so does *y*.

c Use your scientific calculator to calculate the sums, sums of square sand sum of cross products for the pairs (x_i, y_i).

$$\sum x_i = 19;\ \sum y_i = 37;\ \sum x_i^2 = 71;\ \sum y_i^2 = 237;\ \sum x_i y_i = 126$$

Then the covariance is

$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1} = \frac{126 - \frac{(19)37}{6}}{5} = 1.76667$$

and the sample standard deviations are

$$s_x = \sqrt{\frac{\sum x_i - \frac{(\sum x_i)^2}{n}}{n-1}} = \frac{71 - \frac{(19)^2}{6}}{5} = 1.472 \text{ and } s_y = \sqrt{\frac{\sum y_i - \frac{(\sum y_i)^2}{n}}{n-1}} = \sqrt{\frac{237 - \frac{(37)^2}{6}}{5}} = 1.329$$

The correlation coefficient is

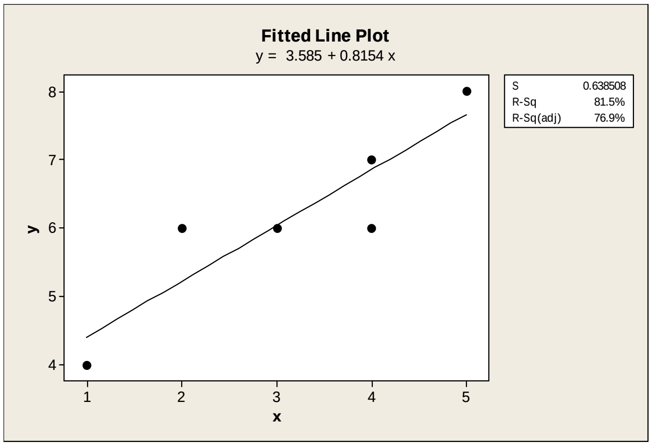
$$r = \frac{s_{xy}}{s_x s_y} = \frac{1.76667}{(1.472)(1.329)} = 0.902986 \approx 0.903$$

d The slope and *y*-intercept of the regression line are

$$b = r\frac{s_y}{s_x} = 0.902986\left(\frac{1.329}{1.472}\right) = 0.81526 \text{ and } a = \bar{y} - b\bar{x} = \frac{37}{6} - 0.81526\left(\frac{19}{6}\right) = 3.58$$

and the equation of the regression line is $y = 3.58 + 0.815x$.

The graph of the data points and the best fitting line is shown below. The line fits through the data points.



3.16 Similar to previous exercises. Calculate
$n = 12;\ \sum x_i = 20,980;\ \sum y_i = 4043.5;\ \sum x_i^2 = 37,551,600;$

$\sum y_i^2 = 1,401,773.75;\ \sum x_i y_i = 7,240,383$. Then the covariance is

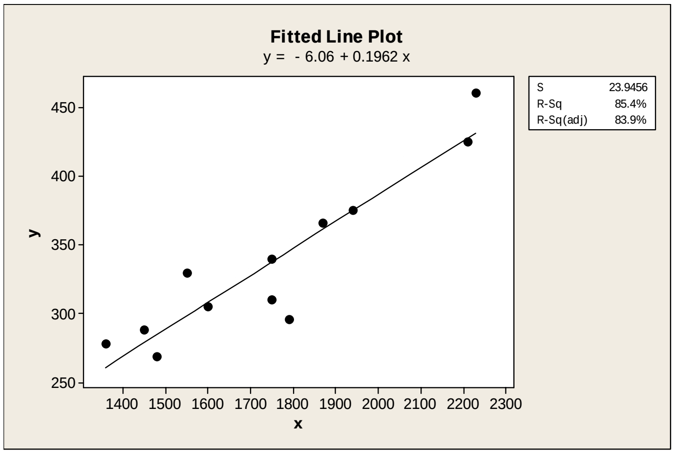
$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1} = 15,545.19697$$

The sample standard deviations are $s_x = 281.48416$ and $s_y = 59.75916$ so that $r = 0.9241$. Then

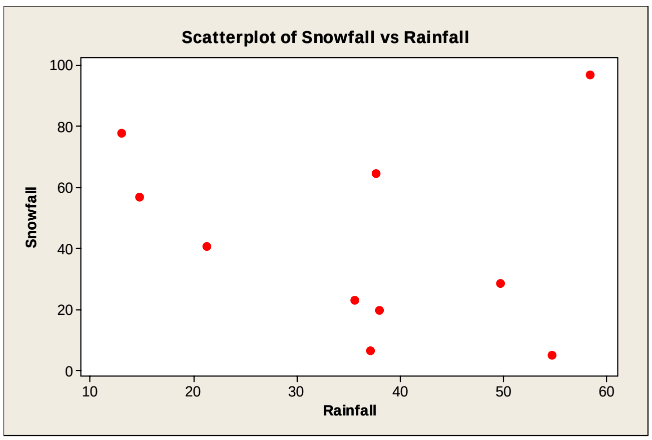
$$b = r\frac{s_y}{s_x} = 0.19620 \text{ and } a = \overline{y} - b\overline{x} = 336.95833 - 0.19620(1748.333) = -6.06$$

and the equation of the regression line is $y = -6.06 + 0.1962x$.

The graph of the data points and the best fitting line is shown below.



3.39 **a** The scatterplot is shown below.



b
Calculate
$n=10;\ \sum x_i = 359.87;\ \sum y_i = 420.30;\ \sum x_i^2 = 15{,}189.9963;\ \sum y_i^2 = 26{,}344.81;\ \sum x_i y_i = 14{,}323.507$.

Then the covariance is

$$s_{xy} = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{n-1} = -89.092122$$

The sample standard deviations are $s_x = 15.7739399$ and $s_y = 31.054792$ so that $r = -0.182$.
There is a weak negative relationship between rainfall and snowfall.

c There is one outlier in the scatterplot—corresponding to Juneau, Alaska. Juneau has a lot of rain **and** a /lot of snow!

d With the outlier removed, calculate

$n = 9;\ \sum x_i = 301.54;\ \sum y_i = 323.3;\ \sum x_i^2 = 11,787.3074;\ \sum y_i^2 = 16,935.81;\ \sum x_i y_i = 8665.497$.
Then the covariance is

$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1} = -270.811236$$

The sample standard deviations are $s_x = 14.511535$ and $s_y = 25.7928177$ so that $r = -0.724$.

The correlation between rainfall and snowfall now shows a fairly strong linear relationship.

Exercices tirés du livre
Introduction to probability and statistics 13th edition
de Mendenhall, Beaver et Beaver.
STT1700
(Automne 2008)

Chapitre 2:

4.1, 4.2, 4.6, 4.7, 4.8, 4.10, 4.42, 4.43, 4.44, 4.49, 4.50, 4.51, 4.52, 4.54, 4.55, 4.56, 4.60, 4.62, 4.63, 4.66, 4.72, 4.73, 4.74, 4.79, 4.101, 4.109, 4.114.

4: Probability and Probability Distributions

4.1 **a** This experiment involves tossing a single die and observing the outcome. The sample space for this experiment consists of the following simple events:

E_1: Observe a 1	E_4: Observe a 4
E_2: Observe a 2	E_5: Observe a 5
E_3: Observe a 3	E_6: Observe a 6

b Events A through F are compound events and are composed in the following manner:

A: (E_2)	D: (E_2)
B: (E_2, E_4, E_6)	E: (E_2, E_4, E_6)
C: (E_3, E_4, E_5, E_6)	F: contains no simple events

c Since the simple events E_i, $i = 1, 2, 3, \ldots, 6$ are equally likely, $P(E_i) = 1/6$.

d To find the probability of an event, we sum the probabilities assigned to the simple events in that event. For example,

$$P(A) = P(E_2) = \frac{1}{6}$$

Similarly, $P(D) = 1/6; P(B) = P(E) = P(E_2) + P(E_4) + P(E_6) = \frac{3}{6} = \frac{1}{2}$; and $P(C) = \frac{4}{6} = \frac{2}{3}$. Since event F contains no simple events, $P(F) = 0$.

4.2 **a** It is given that $P(E_1) = P(E_2) = 1.5$ and $P(E_3) = .40$. Since $\sum_S P(E_i) = 1$, we know that

$$P(E_4) + P(E_5) = 1 - .15 - .15 - .40 = .30 \qquad \text{(i)}$$

Also, it is given that

$$P(E_4) = 2P(E_5) \qquad \text{(ii)}$$

We have two equations in two unknowns which can be solved simultaneously for $P(E_4)$ and $P(E_5)$. Substituting equation (ii) into equation (i), we have

$$2P(E_5) + P(E_5) = .3$$
$$3P(E_5) = .3 \text{ so that } P(E_5) = .1$$

Then from (i), $P(E_4) + .1 = .3$ and $P(E_4) = .2$.

b To find the necessary probabilities, sum the probabilities of the simple events:

$$P(A) = P(E_1) + P(E_3) + P(E_4) = .15 + .4 + .2 = .75$$
$$P(B) = P(E_2) + P(E_3) = .15 + .4 = .55$$

c-d The following events are in either A or B or both: { E_1, E_2, E_3, E_4}. Only event E_3 is in both A and B.

4.6 **a** The experiment consists of selecting one of 25 students and recording the student's gender as well as whether or not the student had gone to preschool.

b The experiment is accomplished in two stages, as shown in the tree diagram below.

Gender	Preschool	Simple Events	Probability
Male	Yes	E_1: Male, Preschool	8/25
	No	E_2: Male, No preschool	6/25
Female	Yes	E_3: Female, Preschool	9/25
	No	E_4: Female, No preschool	2/25

c Since each of the 25 students are equally likely to be chosen, the probabilities will be proportional to the number of students in each of the four gender-preschool categories. These probabilities are shown in the last column of the tree diagram above.

d $P(\text{Male}) = P(E_1) + P(E_2) = \frac{8}{25} + \frac{6}{25} = \frac{14}{25}$; $P(\text{Female and no preschool}) = P(E_4) = \frac{2}{25}$

4.7 Label the five balls as R_1, R_2, R_3, Y_1 and Y_2. The selection of two balls is accomplished in two stages to produce the simple events in the tree diagram on the next page.

First Ball	Second Ball	Simple Events
R_1	R_2	R_1R_2
	R_3	R_1R_3
	Y_1	R_1Y_1
	Y_2	R_1Y_2
R_2	R_1	R_2R_1
	R_3	R_2R_3
	Y_1	R_2Y_1
	Y_2	R_2Y_2
R_3	R_1	R_3R_1
	R_2	R_3R_2
	Y_1	R_3Y_1
	Y_2	R_3Y_2

First Ball	Second Ball	Simple Events
Y_1	R_1	Y_1R_1
	R_2	Y_1R_2
	R_3	Y_1R_3
	Y_2	Y_1Y_2
Y_2	R_1	Y_2R_1
	R_2	Y_2R_2
	R_3	Y_2R_3
	Y_1	Y_2Y_1

4.8 When the first ball is replaced before the second ball is drawn, five additional simple events become possible:

$$R_1R_1,\ R_2R_2,\ R_3R_3,\ Y_1Y_1, \text{ and } Y_2Y_2$$

4.10 **a** There are 38 simple events, each corresponding to a single outcome of the wheel's spin. The 38 simple events are indicated below.

E_1: Observe a 1

E_2: Observe a 2

⋮

E_{36}: Observe a 36

E_{37}: Observe a 0

E_{38}: Observe a 00

b Since any pocket is just as likely as any other, $P(E_i) = 1/38$.

c The event A contains two simple events, E_{37} and E_{38}. Then

$$P(A) = P(E_{37}) + P(E_{38}) = 2/38 = 1/19.$$

d Define event B as the event that you win on a single spin. Since you have bet on the numbers 1 through 18, event B contains 18 simple events, E_1, E_2, …, E_{18}. Then

$$P(B) = P(E_1) + P(E_2) + \cdots + P(E_{18}) = 18/38 = 9/19$$

4.42 Each simple event is equally likely, with probability $1/5$.

a $A^C = \{E_2, E_4, E_5\}$ $\quad P\left(A^C\right) = 3/5$

b $A \cap B = \{E_1\}$ $\quad P\left(A \cap B\right) = 1/5$

c $B \cap C = \{E_4\}$ $\quad P\left(B \cap C\right) = 1/5$

d $A \cup B = S = \{E_1, E_2, E_3, E_4, E_5\}$ $\quad P\left(A \cup B\right) = 1$

e $B | C = \{E_4\}$ $\quad P(B | C) = 1/2$

f $A | B = \{E_1\}$ $\quad P(A | B) = 1/4$

g $A \cup B \cup C = S$ $\quad P\left(A \cup B \cup C\right) = 1$

h $\left(A \cap B\right)^C = \{E_2, E_3, E_4, E_5\}$ $\quad P\left(A \cap B\right)^C = 4/5$

4.43 **a** $P\left(A^C\right) = 1 - P\left(A\right) = 1 - \frac{2}{5} = \frac{3}{5}$

b $P\left(A \cap B\right)^C = 1 - P\left(A \cup B\right) = 1 - \frac{1}{5} = \frac{4}{5}$

4.44 **a** $P\left(A | B\right) = \frac{P(A \cap B)}{P(B)} = \frac{1/5}{4/5} = \frac{1}{4}$

b $P\left(B | C\right) = \frac{P(B \cap C)}{P(C)} = \frac{1/5}{2/5} = \frac{1}{2}$

4.49 **a** Since A and B are independent, $P(A \cap B) = P(A)P(B) = .4(.2) = .08$.

b $P(A \cup B) = P(A) + P(B) - P(A \cap B) = .4 + .2 - (.4)(.2) = .52$

4.50 **a** Since A and B are mutually exclusive, $P(A \cap B) = 0$.

b $P(A \cup B) = P(A) + P(B) - P(A \cap B) = .3 + .5 - 0 = .8$

4.51 **a** Use the definition of conditional probability to find

$$P\left(B \mid A\right) = \frac{P(A \cap B)}{P(A)} = \frac{.12}{.4} = .3$$

b Since $P(A \cap B) \neq 0$, A and B are not mutually exclusive.

c If $P(B) = .3$, then $P(B) = P(B \mid A)$ which means that A and B are independent.

4.52 **a** The event A will occur whether event B occurs or not (B^C). Hence,

$$P(A) = P(A \cap B) + P(A \cap B^C) = .34 + .15 = .49$$

b Similar to part **a**. $P(B) = P(A \cap B) + P(A^C \cap B) = .34 + .46 = .80$.

c The contents of the cell in the first row and first column is $P(A \cap B) = .34$.

d $P(A \cup B) = P(A) + P(B) - P(A \cap B) = .49 + .80 - .34 = .95$.

e Use the definition of conditional probability

$$P\left(A \mid B\right) = \frac{P(A \cap B)}{P(B)} = \frac{.34}{.80} = .425$$

f Similar to part **e**.

$$P\left(B \mid A\right) = \frac{P(A \cap B)}{P(A)} = \frac{.34}{.49} = .694$$

4.54 Define the following events:

P: test is positive for drugs

N: test is negative for drugs

D: employee is a drug user

It is given that, on a given test, $P(P \mid D) = .98$, $P(N \mid D^C) = .98$.

a For a given test, $P(P \mid D^C) = 1 - .98 = .02$. Since the tests are independent

$$P(\text{fail both test} \mid D^C) = (.02)(.02) = .0004$$

b $P(\text{detection} \mid D) = P(N \cap P \mid D) + P(P \cap N \mid D) + P(P \cap P \mid D)$

$$= (.98)(.02) + (.02)(.98) + (.98)(.98) = .9996$$

c $P(\text{pass both} \mid D) = P(N \cap N \mid D) = (.02)(.02) = .0004$

4.55 Define the following events:

A: project is approved for funding

D: project is disapproved for funding

For the first group, $P(A_1) = .2$ and $P(D_1) = .8$. For the second group, $P[\text{same decision as first group}] = .7$ and $P[\text{reversal}] = .3$. That is,

$$P(A_2 \mid A_1) = P(D_2 \mid D_1) = .7 \text{ and } P(A_2 \mid D_1) = P(D_2 \mid A_1) = .3.$$

a $P(A_1 \cap A_2) = P(A_1)P(A_2 \mid A_1) = .2(.7) = .14$

b $P(D_1 \cap D_2) = P(D_1)P(D_2 \mid D_1) = .8(.7) = .56$

c $P(D_1 \cap A_2) + P(A_1 \cap D_2) = P(D_1)P(A_2 \mid D_1) + P(A_1)P(D_2 \mid A_1) = .8(.3) + .2(.3) = .30$

4.56 The two-way table in the text gives probabilities for events A, A^C, B, B^C in the column and row marked "Totals". The interior of the table contains the four two-way intersections as shown below.

$A \cap B$	$A \cap B^C$
$A^C \cap B$	$A^C \cap B^C$

The necessary probabilities can be found using various rules of probability if not directly from the table.

a $P(A) = .4$

b $P(B) = .37$

c $P(A \cap B) = .10$

d $P\left(A\cup B\right)=.4+.37-.10=.67$

e $P(A^C)=1-.4=.6$

f $P\left(A\cup B\right)^C=1-P(A\cup B)=1-.67=.33$

g $P\left(A\cap B\right)^C=1-P(A\cap B)=.90$

h $P\left(A\,|\,B\right)=P(A\cap B)/P(B)=.1/.37=.27$

i $P\left(B\,|\,A\right)=P(A\cap B)/P(A)=.1/.4=.25$

4.60 Define the following events:

S: student chooses Starbucks

P: student chooses Peets

C: student orders a café mocha

Then $P(S)=.7$; $P(P)=.3$; $P(C\,|\,S)=P(C\,|\,P)=.60$

a Using the given probabilities, $P(S\cap C)=P(S)P(C\,|\,S)=.7(.6)=.42$

b Since $P(C)=.6$ regardless of whether the student visits Starbucks or Peets, the two events are independent.

c $P(P\,|\,C)=\dfrac{P(P\cap C)}{P(C)}=\dfrac{P(P)P(C\,|\,P)}{P(C)}=P(P)=.3$

d $P(S\cup C)=P(S)+P(C)-P(S\cap C)=.7+.6-(.7)(.6)=.88$

4.62 Define the events: D: person dies

S: person smokes

It is given that $P(S)=.2$, $P(D)=.006$, and $P(D\,|\,S)=10P(D\,|\,S^C)$. The probability of interest is $P(D\,|\,S)$. The event D, whose probability is given, can be written as the union of two mutually exclusive intersections. That is,

$$D=(D\cap S)\cup(D\cap S^C)$$

Then, using the Addition and Multiplication Rules,

$$P(D) = P(D \cap S) + P(D \cap S^C) = P(D \mid S)P(S) + P(D \mid S^C)P(S^C) = P(D \mid S)(.2) + \left[(1/10)P(D \mid S)\right](.8)$$

Since $P(D) = .006$, the above equation can be solved for $P(D \mid S)$

$$P(D \mid S)(.2 + .08) = .006$$
$$P(D \mid S) = .006/.28 = .0214$$

4.63 Define A: smoke is detected by device A

B: smoke is detected by device B

If it is given that $P(A) = .95,\ P(B) = .98,$ and $P(A \cap B) = .94$.

a $P(A \cup B) = P(A) + P(B) - P(A \cap B) = .95 + .98 - .94 = .99$

b $P(A^C \cap B^C) = 1 - P(A \cup B) = 1 - .99 = .01$

4.66 Similar to Exercise 4.56.

a $P(F) = .35 + .36 = .71$

b $P(G) = .20 + .09 = .29$

c $P(F \mid M) = \dfrac{P(F \cap M)}{P(M)} = \dfrac{.35}{.55} = .63$

d $P(F \mid W) = \dfrac{P(F \cap W)}{P(W)} = \dfrac{.36}{.45} = .80$

e $P(M \mid F) = \dfrac{P(M \cap F)}{P(F)} = \dfrac{.35}{.71} = .49$

e $P(W \mid G) = \dfrac{P(W \cap G)}{P(G)} = \dfrac{.09}{.29} = .31$

4.72 Define the following events: V: crime is violent

R: crime is reported

It is given that $P(V) = .2,\ P(V^C) = .8,\ P(R \mid V) = .9,\ P(R \mid V^C) = .7$

a The overall reporting rate for crimes is

$$P(R) = P(V)P(R \mid V) + P(V^C)P(R \mid V^C) = .2(.9) + .8(.7) = .74$$

b Use Bayes' Rule: $P(V \mid R) = \dfrac{P(V)P(R \mid V)}{P(R)} = \dfrac{.2(.9)}{.74} = .24$

and $P(V^C \mid R) = \dfrac{P(V^C)P(R \mid V^C)}{P(R)} = \dfrac{.8(.7)}{.74} = .76$

c Notice that the proportion of non-violent crimes (.8) is much larger than the proportion of violent crimes (.2). Therefore, when a crime is reported, it is more likely to be a non-violent crime.

4.73 Define A: machine produces a defective item

B: worker follows instructions

Then $P(A|B)=.01,\ P(B)=.90,\ \ P(A|B^C)=.03,\ P(B^C)=.10$. The probability of interest is

$$\begin{aligned} P(A) &= P(A\cap B)+P(A\cap B^C) \\ &= P(A|B)P(B)+P(A|B^C)P(B^C) \\ &= .01(.90)+.03(.10)=.012 \end{aligned}$$

4.74 Define the following events: A: passenger uses airport A

B: passenger uses airport B

C: passenger uses airport C

D: a weapon is detected

Suppose that a passenger is carrying a weapon. It is given that

$P(D|A)=.9$ $\qquad P(A)=.5$
$P(D|B)=.8$ $\qquad P(B)=.3$
$P(D|C)=.85$ $\qquad P(C)=.2$

The probability of interest is

$$P(A|D)=\frac{P(A)P(D|A)}{P(A)P(D|A)+P(B)P(D|B)+P(C)P(D|C)}=\frac{.5(.9)}{.5(.9)+.3(.8)+.2(.85)}=.5233$$

Similarly,

$$P(C|D)=\frac{.2(.85)}{.5(.9)+.3(.8)+.2(.85)}=\frac{.17}{.86}=.1977$$

4.79 **a** Using the probability table,

$P(D)=.08+.02=.10$ $\qquad P(D^C)=1-P(D)=1-.10=.90$

$P(N|D^C)=\dfrac{P(N\cap D^C)}{P(D^C)}=\dfrac{.85}{.90}=.94$ $\qquad P(N|D)=\dfrac{P(N\cap D)}{P(D)}=\dfrac{.02}{.10}=.20$

b Using Bayes' Rule,

$$P(D|N)=\frac{P(D)P(N|D)}{P(D)P(N|D)+P(D^C)P(N|D^C)}=\frac{.10(.20)}{.10(.20)+.90(.94)}=.023$$

c Using the definition of conditional probability,

$$P(D|N)=\frac{P(N\cap D)}{P(N)}=\frac{.02}{.87}=.023$$

d $P(\text{false positive})=P(P|D^C)=\frac{P(P\cap D^C)}{P(D^C)}=\frac{.05}{.90}=.056$

e $P(\text{false negative})=P(N|D)=\frac{P(N\cap D)}{P(D)}=\frac{.02}{.10}=.20$

f The probability of a false negative is quite high, and would cause concern about the reliability of the screening method.

4.101 Define the following events:

A: worker fails to report fraud

B: worker suffers reprisal

It is given that $P(B|A^C)=.23$ and $P(A)=.69$. The probability of interest is

$$P(A^C\cap B)=P(B|A^C)P(A^C)=.23(.31)=.0713$$

4.109 **a** Similar to Exercise 4.15. $P(\text{cold})=\frac{49+43+34}{276}=\frac{126}{276}=.4565$

b Define: F: person has four or five relationships

S: person has six or more relationships

Then for the two people chosen from the total 276,

$$P(\text{one F and one S})=P(F\cap S)+P(S\cap F)$$
$$=\left(\frac{100}{276}\right)\left(\frac{96}{275}\right)+\left(\frac{96}{276}\right)\left(\frac{100}{275}\right)=.2530$$

c $P(\text{Three or fewer}\mid\text{cold})=\frac{P(\text{three or fewer}\cap\text{cold})}{P(\text{cold})}=\frac{49/276}{126/276}=\frac{49}{126}=.3889$

4.114 Define the following events: B: man takes the bus

S: man takes the subway

L: the man is late

It is given that $P(B) = .3,\ P(S) = .7,\ P(L \mid B) = .3,\ P(L \mid S) = .2$. Using Bayes' Rule,

$$P(B \mid L) = \frac{P(L \mid B)P(B)}{P(L \mid B)P(B) + P(L \mid S)P(S)} = \frac{(.3)(.3)}{(.3)(.3) + (.2)(.7)} = \frac{.09}{.23} = .3913$$

Exercices tirés du livre
Introduction to probability and statistics 13th edition
de Mendenhall, Beaver et Beaver.
STT1700
(Automne 2008)

Chapitre 3:

4.80, 4.81, 4.82, 4.86, 4.89, 4.95, 4.96, 4.98, 4.99, 4.107, 4.108, 4.124, ~~5.3, 5.4, 5.11, 5.20, 5.21, 5.22, 5.23, 5.25, 5.30, 5.63.~~

4: Probability and Probability Distributions

4.80 **a** The number of points scored is a discrete random variable taking the countably infinite number of values, 0, 1, 2, …

b Shelf life is a continuous random variable, since it can take on any positive real value.

c Height is a continuous random variable, taking on any positive real value.

d Length is a continuous random variable, taking on any positive real value.

e Number of near collisions is a discrete random variable, taking the values 0, 1, 2, …

4.81 **a** The increase in length of life achieved by a cancer patient as a result of surgery is a continuous random variable, since an increase in life (measured in units of time) can take on any of an infinite number of values in a particular interval.

b The tensile strength, in pounds per square inch, of one-inch diameter steel wire cable is a continuous random variable.

c The number of deer killed per year in a state wildlife preserve is a discrete random variable taking the values 0,1, 2, …

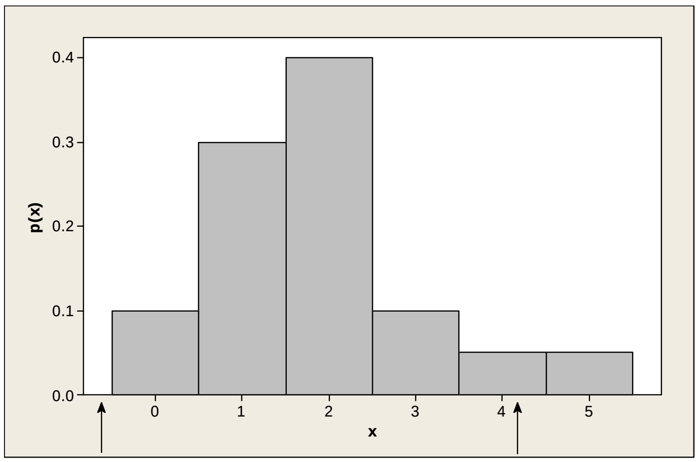
d The number of overdue accounts in a department store at a particular point in time is a discrete random variable, taking the values 0, 1, 2, ….

e Blood pressure is a continuous random variable.

4.82 **a** Since one of the requirements of a probability distribution is that $\sum_x p(x) = 1$, we need

$p(4) = 1 - (.1 + .3 + .4 + .1 + .05) = 1 - .95 = .05$

b The probability histogram is shown in the figure on the next page.



c For the random variable *x* given here,

$$\mu = E(x) = \sum xp(x) = 0(.1) + 1(.3) + \cdots + 5(.05) = 1.85$$

The variance of *x* is defined as

$$\sigma^2 = E\left[(x-\mu)^2\right] = \sum (x-\mu)^2 \, p(x) = (0-1.85)^2(.1) + (1-1.85)^2(.3) + \cdots + (5-1.85)^2(.05) = 1.4275$$

and $\sigma = \sqrt{1.4275} = 1.19$.

d The interval of interest is $\mu \pm 2\sigma = 1.85 \pm 2.38$ or –.53 to 4.23. This interval is shown on the probability histogram above. Then $P[-.53 \le x \le 4.23] = P[0 \le x \le 4] = .95$.

e Since the probability that *x* falls in the interval $\mu \pm 2\sigma$ is .95 from part **d**, we would expect <u>most</u> of the observations to fall in this interval.

4.86 **a** Define

D: person prefers David Letterman

J: person prefers Jay Leno

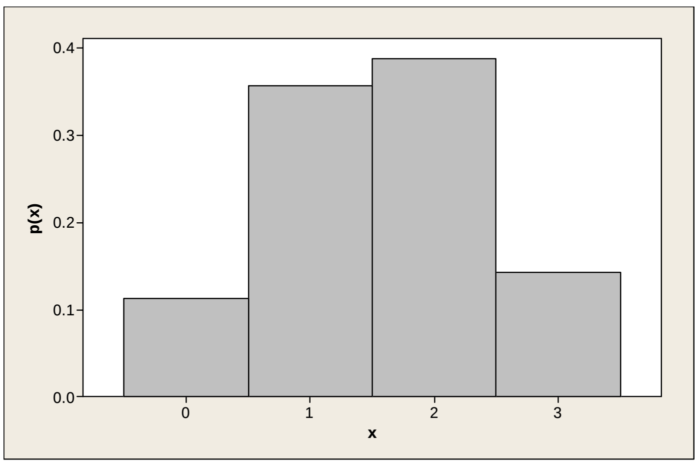
There are eight simple events in the experiment:

DDD	DDJ
DJJ	DJD
JDJ	JDD
JJD	JJJ

and the probabilities for x = number who prefer Jay Leno = 0, 1, 2, 3 are shown below.

$$P(x=0)=P(DDD)=(.48)^3=.1106$$
$$P(x=1)=P(DDJ)+P(DJD)+P(JDD)=3(.52)(.48)^2=.3594$$
$$P(x=2)=P(DJJ)+P(JJD)+P(JDJ)=3(.52)^2(.48)=.3894$$
$$P(x=3)=P(JJJ)=(.52)^3=.1406$$

b The probability histogram is shown below.



c $P(x=1)=.3594$

d The average value of x is

$$\mu=E(x)=\sum xp(x)=0(.1106)+1(.3594)+2(.3894)+3(.1406)=1.56$$

The variance of *x* is

$$\sigma^2 = E\left[(x-\mu)^2\right] = \sum (x-\mu)^2 \, p(x)$$
$$= (0-1.56)^2 (.1106) + (1-1.56)^2 (.3594) + (2-1.56)^2 (.3894) + (3-1.56)^2 (.1406)$$
$$= .7488$$

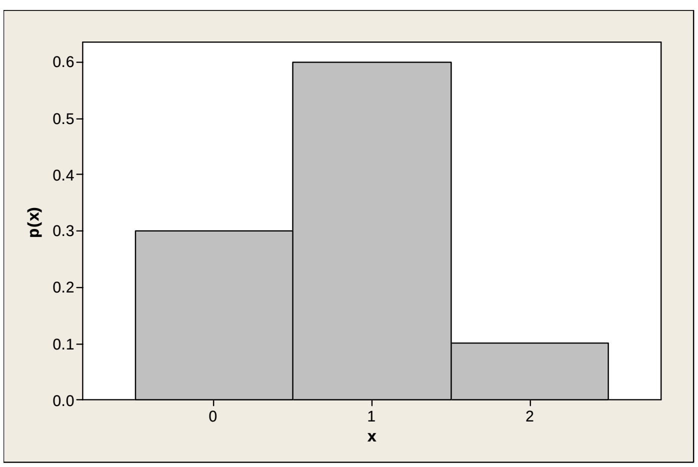
and

$$\sigma = \sqrt{.7488} = .865 .$$

4.89 **a-b** Let W_1 and W_2 be the two women while M_1, M_2 and M_3 are the three men. There are 10 ways to choose the two people to fill the positions. Let *x* be the number of women chosen. The 10 equally likely simple events are:

E_1: W_1W_2 $(x=2)$	E_6: W_2M_2 $(x=1)$
E_2: W_1M_1 $(x=1)$	E_7: W_2M_3 $(x=1)$
E_3: W_1M_2 $(x=1)$	E_8: M_1M_2 $(x=0)$
E_4: W_1M_3 $(x=1)$	E_9: M_1M_3 $(x=0)$
E_5: W_2M_1 $(x=1)$	E_{10}: M_2M_3 $(x=0)$

The probability distribution for *x* is then $p(0) = 3/10$, $p(1) = 6/10$, $p(2) = 1/10$. The probability histogram is shown below.



4.95 The random variable G, total gain to the insurance company, will be D if there is no theft, but D – 50,000 if there is a theft during a given year. These two events will occur with probability .99 and .01, respectively. Hence, the probability distribution for G is given below.

G	*p*(G)
D	.99
D – 50,000	.01

The expected gain is

$$E(G) = \sum Gp(G) = .99D + .01(D - 50,000)$$
$$= D - 50,000$$

In order that $E(G) = 1000$, it is necessary to have $1000 = D - 500$ or $D = \$1500$.

4.96 **a** $\mu = E(x) = \sum xp(x) = 3(.03) + 4(.05) + \cdots + 13(.01) = 7.9$

b $\sigma^2 = \sum (x-\mu)^2 \; p(x) = (3-7.9)^2(.03) + (4-7.9)^2(.05) + \cdots + (13-7.9)^2(.01) = 4.73$ and $\sigma = \sqrt{4.73} = 2.1749$.

c Calculate $\mu \pm 2\sigma = 7.9 \pm 4.350$ or 3.55 to 12.25. Then, referring to the probability distribution of *x*, $P[3.55 \le x \le 12.25] = P[4 \le x \le 12] = 1 - p(3) - p(13) = 1 - .04 = .96$.

4.98 The company will either gain ($15.50 – 14.80) if the package is delivered on time, or will lose $14.80 if the package is not delivered on time. We assume that, if the package is not delivered within 24 hours, the company does not collect the $15.50 delivery fee. Then the probability distribution for *x*, the company's gain is

x	*p*(*x*)
.70	.98
–14.80	.02

and

$$\mu = E(x) = .70(.98) - 14.80(.02) = .39.$$

The expected gain per package is $0.39.

4.99 We are asked to find the premium that the insurance company should charge in order to break even. Let *c* be the unknown value of the premium and *x* be the gain to the insurance company caused by marketing the new product. There are three possible values for *x*. If the product is a failure or moderately successful, *x* will be negative; if the product is a success, the insurance company will gain the amount of the premium and *x* will be positive. The probability distribution for *x* follows:

x	$p(x)$
c	.94
$-800{,}000 + c$	.01
$-250{,}000 + c$	.05

In order to break even,

$E(x) = \sum xp(x) = 0$.

Therefore,

$$.94(c) + .01(-800,000 + c) + (.05)(-250,000 + c) = 0$$
$$-8000 - 12,500 + (.01 + .05 + .94)c = 0$$
$$c = 20,500$$

Hence, the insurance company should charge a premium of $20,500.

4.107 The random variable *x*, defined as the number of householders insured against fire, can assume the values 0, 1, 2, 3 or 4. The probability that, on any of the four draws, an insured person is found is .6; hence, the probability of finding an uninsured person is .4. Note that each numerical event represents the intersection of the results of four independent draws.

1 $P[x = 0] = (.4)(.4)(.4)(.4) = .0256$, since all four people must be uninsured.

2 $P[x = 1] = 4(.6)(.4)(.4)(.4) = .1536$ (Note: the 4 appears in this expression because $x = 1$ is the union of four mutually exclusive events. These represent the 4 ways to choose the single insured person from the fours.)

3 $P[x = 2] = 6(.6)(.6)(.4)(.4) = .3456$, since the two insured people can be chosen in any of 6 ways.

4 $P[x = 3] = 4(.6)^3(.4) = .3456$ and $P[x = 4] = (.6)^4 = .1296$.

Then

$$P[\text{at least three insured}] = p(3) + p(4) = .3456 + .1296 = .4752$$

4.108 In this exercise, *x* may take the values 0, 1, 2, or 3, and the probabilities associated with *x* are evaluated as in Exercise 4.107.

a $P[x = 0] = (.2)^3 = .008$

$P[x = 1] = 3(.8)(.2)^2 = .096$

$P[x = 2] = 3(.8)^2(.2) = .384$

$P[x = 3] = (.8)^3 = .512$

The reader may verify that the probabilities sum to one and that $0 \le p(x) \le 1$ for $x = 0, 1, 2,$ and 3. The requirements for a probability distribution have been satisfied.

b The alarm will function if $x = 1, 2,$ or 3. Hence,

$$P[\text{alarm functions}] = p(1) + p(2) + p(3) = .096 + .384 + .512 = .992$$

c $\mu = E(x) = \Sigma xp(x) = 0(.008) + 1(.096) + 2(.384) + 3(.512) = 2.4$

$$\sigma^2 = \sum (x-\mu)^2 p(x) = (0-2.4)^2(.008) + (1-2.4)^2(.096) + (2-2.4)^2(.384) + (3-2.4)^2(.512) = .48$$

4.124 Let *y* represent the value of the premium which the insurance company charges and let *x* be the insurance company's gain. There are four possible values for *x*. If no accident occurs or if an accident results in no damage to the car, the insurance company gains *y* dollars. If an accident occurs and the car is damaged, the company will gain either $y - 22,000$ dollars, $y - .6(22,000)$ dollars, or $y - .2(22,000)$ dollars, depending upon whether the damage to the car is total, 60% of market value, or 20% of market value, respectively. The following probabilities are known.

$P[\text{accident occurs}] = .15$ $\qquad$ $P[\text{total loss | accident occurs}] = .08$

$P[\text{60\% loss | accident occurs}] = .12$ $\qquad$ $P[\text{20\% loss | accident occurs}] = .80$

Hence,

$$P[x = y - 22,000] = P[\text{accident}]P[\text{total loss | accident}] = .15(.08) = .012$$

Similarly,

$$P[x = y - 13,200] = .15(.12) = .018 \text{ and } P[x = y - 4400] = .15(.80) = .12$$

The gain *x* and its associated probability distribution are shown below. Note that *p*(*y*) is found by subtraction.

x	$p(x)$
$y - 22,000$	.012
$y - 13,200$	.018
$y - 4400$	.12
y	.85

Letting the expected gain equal zero, the value of the premium is obtained.

$$
\begin{aligned}
E(x) &= \sum xp(x) = .012(y - 22,000) + .018(y - 13,200) + .12(y - 4400) + .85y \\
E(x) &= y - (264 + 237.6 + 528) = y - 1029.6 \\
y &= \$1029.6
\end{aligned}
$$