Chapter 2

Discussion Problem Solutions

D1. A good example would involve something that is equally likely to occur on each day of the week, such as the birthdays of classmates or the number of shooting stars seen in the night sky. Possible examples are the digits that occur in winning lottery numbers or the last digit of the height in millimeters of members of the class.

D2. Possible examples include the number of automobile accidents or the number of visits to the doctor over the course of a year because both occur more frequently in winter in most places. Shooting stars are more likely in some months than others, a notable meteor shower occurring each August. Because February has fewer days, almost anything should occur less frequently in February. Another possible example is the number of bills paid each day of the month in a given family, which tend to be clustered around given days of the month, such as payday. Additional examples include the frequency of students in the class with that digit as the last digit of their age or the frequency of last digits of prices of items in a clothing store. The number of trips to the beach a person makes in a lifetime on that day or the number of people who attend the local movie theater on that day.

D3. a. The diameter of a typical tennis ball is 65 mm, give or take 2 mm.
b. The weight of a typical penny is about 3.11 g, give or take 0.04 g or so.

D4. a. This distribution is strongly skewed right. Most islands are quite small; Cuba and Hispaniola are comparatively very large. There is a wall at 0 because no island can be smaller than that, but many are close.

b. This distribution should be skewed right because some countries, such as the United States, have a per capita income that is much higher than that of most other countries. There is a wall near 0 because average per capita incomes can't go below that, and for many countries the per capita income is very small.

c. This distribution should not be skewed at all. Lengths like this typically form a distribution that is approximately normal.

d. This distribution probably will be skewed left. There is a wall at 1 hour—no student can take longer than that, and most students will work on an exam for the entire hour or close to it. A few students, however, will leave early.

e. This distribution will be skewed right. Some emperors reigned a long time but most for a moderate number of years. There is a wall at 0 years.

D5. Most distributions will be skewed toward the right, because it is very common to have a wall at 0 since most quantities must be positive.

D6. a. Splitting would be advantageous. College students will have reached their adult height so age differences would not be a factor, however, a split by gender would typically show two distinct normal distributions with different means.

b. Splitting may be advantageous. The gas mileage data would likely consist of three clusters: hybrids and other fuel efficient vehicles, SUVs and other fuel inefficient vehicles, and sedans with midrange fuel efficiency.

c. Splitting would be advantageous. Certain regions of Sudan, for instance Darfur, were regions of conflict and would exhibit more deaths among the young than a peacetime region. Splitting according to conflict would isolate these different distributions.

D7. The "skyline" of the histogram remains the same. Only the scale on the vertical axis changes. The histogram has a vertical scale from 0 to some integer, whereas the relative frequency histogram has a vertical scale from 0 to 1. From a relative frequency histogram, you can tell neither how many cases there are in each bar nor how many total cases. From a frequency histogram, it is harder to judge relative frequency. For large data sets, it is much more convenient to summarize data by stating proportions (relative frequencies) in various bars rather than the large number of cases that might fall into various bars of interest.

D8. The narrower bars cover a smaller interval on the real number line. Thus, you can state more precisely which speeds are in a given bar than you can when they are wider. If you made all histograms with very narrow bars, they would essentially be dot plots and you could have hundreds or thousands of bars. In a histogram, you combine nearby values into bars so that you can have fewer bars, making the overall shape easier to see. On the other hand, if the bars are too wide, you may miss gaps and clusters.

D9. The leftmost column gives the number of values in the stemplot up to and including that row. Below the median, the counting is done from the bottom up. For example, the 8 at the beginning of the seventh row means that by the end of that row, with value 42, there are 8 values so far in the stemplot, counting from the bottom row up. The 2 in parentheses indicates that there are 2 values in the row of the plot that includes the median of the data.

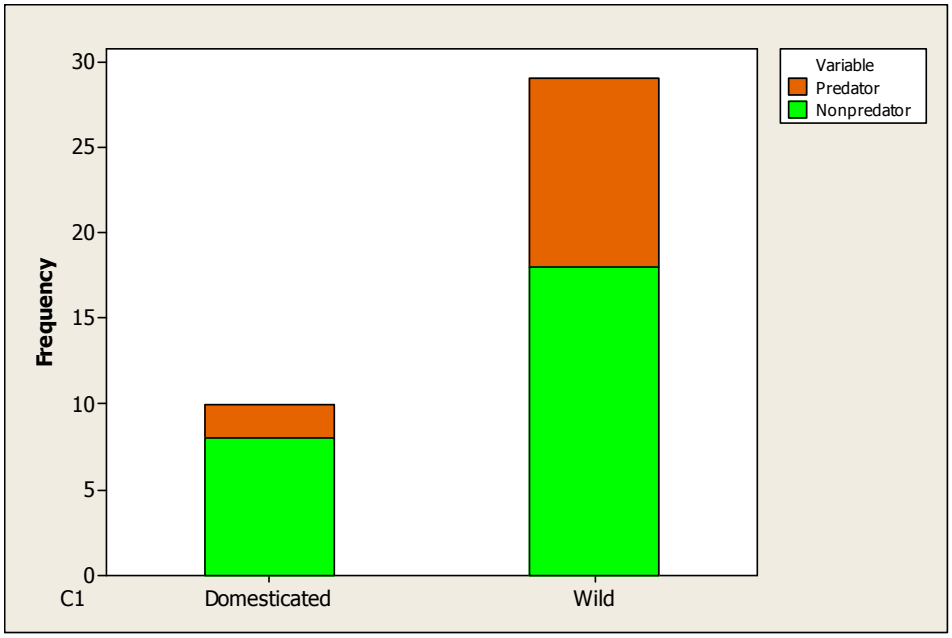
D10. The stemplot could be set up in a few ways. The stem could be made up of two digits (hundreds and tens), so 11|2 means 112 days. Since the number of days ranges from 13 to 660, this would require a long list of stems. A second option would be to have the stem consist of one digit for the hundreds place, then put the tens and ones place in the leaves separated by commas or a space. A third option would be to group values into intervals of ten. In that case, anything from 110 to 119 would be represented by 1|1. With this approach you lose a little information, but shape and approximate values are preserved. This plot was made by Minitab statistical software. "Leaf Unit = 10" means

that the leaf gives the tens place of the number and the stem gives the hundreds place. So, for example, the last line represents one number that falls in the interval 660 to 669. As you can see by the plot, the distribution is highly skewed toward the larger values, with a possible outlier (the elephant) at the high end.

```
Stem-and-leaf of Gestation N = 38
Leaf Unit = 10            N* = 1
  6    0  123334
 13    0  5666699
 17    1  0001
(4)    1  5568
 17    2  02334
 12    2  5588
  8    3  3
  7    3  566
  4    4  02
  2    4  5
  1    5
  1    5
  1    6
  1    6  6
     6|6 represents a number in the
      interval 660-669 days.
```

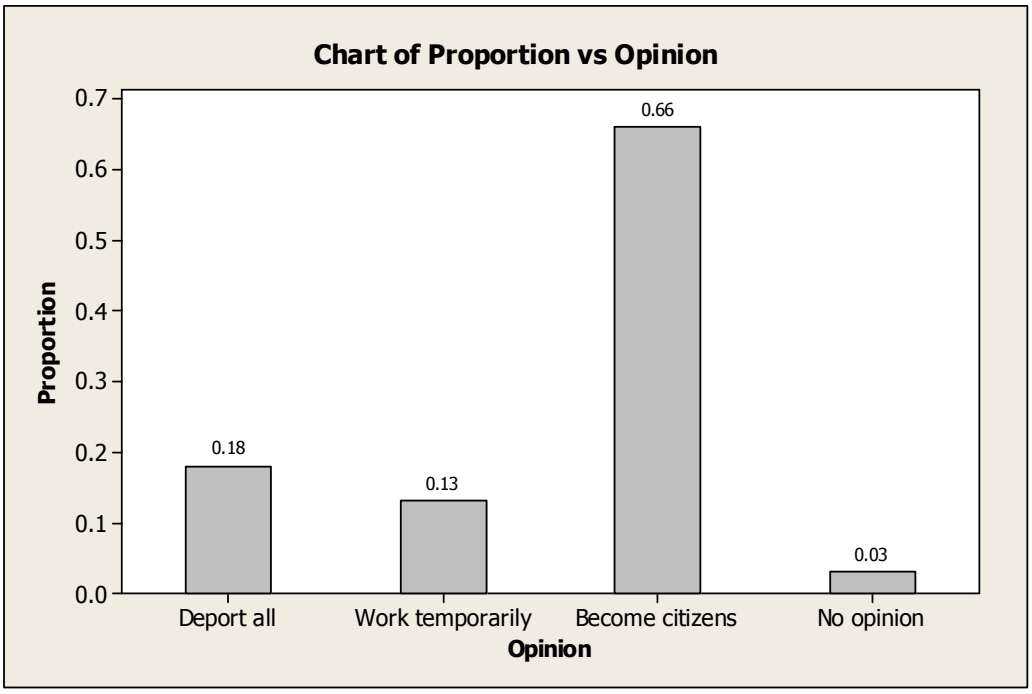
D11. The ordering of the bars in Display 2.23 is completely arbitrary and could have been done in the opposite order. The categories in the education data in Display 2.24, however, represent increasing amounts of education and should be kept in order to see the pattern in the frequencies.

D12.



D13. a. Of the Republicans surveyed, 29% feel that all illegal immigrants currently residing in the United States should be deported.

b.



The order of the bars does not matter.

c. The bar charts described here are used to compare the sizes of the various groups under observation. For example, if one bar is twice as high as another, the group corresponding to the first bar is twice as large as the second. The data in the row labeled "Remain in the U.S. to become citizens" do not indicate the relative sizes of the groups involved. The values of 50 and 60 tell us that 50% of the Republicans surveyed held that opinion while 60% of Independents held the same opinion but there is no way to compare the sizes of the two groups. The total surveyed in each category: Republicans, Independents, and Democrats is needed to construct a meaningful bar chart.

D14. a. mean 2 median 2
b. mean 3 median 2
c. mean 4 median 2
d. mean 100 median 2

The median is unchanged because increasing the largest number doesn't change the fact that 2 is the number in the center. The mean gets larger when any number is increased.

D15. a. As in D14, the mean is more affected by an outlier. To be the balance point, the mean has to move upward with the largest number because increasing the largest number places more weight on the upper end of the distribution. To remain the value in the center, the median doesn't have to change at all.

b. The distribution for the predators is skewed right, with the median smaller than the

mean. The distribution for the nonpredators is skewed left, with the median larger than the mean. Because the nonpredators have generally smaller values to begin with, the means are farther apart than the medians.

c. The distribution has a fairly large gap between ages 38 and 48. When the larger values were removed, the central value or median had to "jump" that gap and became much smaller. This result illustrates that the median can be quite unstable when there are only a few values in a distribution or when there are gaps.

D16. Detective Seymour has received "quite a few" descriptions of the suspect from various eyewitnesses. These descriptions varied so much that Detective Seymour said that the average was useless. For example, suppose he had four eyewitnesses, two of whom said the murderer was 5'7" tall and two of whom said the murderer was 5'8" tall. In this case, it would be perfectly reasonable for Detective Seymour to believe that the murderer was close to 5'7 1/2 " tall. However, if his four eyewitnesses said the murderer was 5', 5'4", 5'9", and 6'5", it wouldn't be reasonable for him to make any conclusion about the murderer's height even though the average is still 5'7 1/2 " .

D17. a. 50%; 25%; 25%

b. A boxplot for a data set that is extremely skewed right would have a lower whisker that is shorter than the bottom half of the box, which is shorter than the top half of the box, which is shorter than the upper whisker. A boxplot for a data set that is extremely skewed left would have a lower whisker that is longer than the lower half of the box, which is longer than the upper half of the box, which is longer than the upper whisker. A boxplot for a data set that is symmetric would have whiskers of equal length, and the two halves of the box would be of equal length.

c. The IQR is equal to the length of the box. The range is the distance from the end of one whisker (or outermost outlier) to the end of the other whisker (or outermost outlier).

d. Yes. There will be no lower whisker, for example, if the minimum and the lower quartile are equal. This set of values has no lower whisker:
{1, 1, 1, 1, 2, 3, 5, 6, 7, 12, 14, 16}

e. A histogram divides the number line into equal intervals and shows how many values fall into each interval. A boxplot divides the data into four equal parts and shows what part of the number line is covered by each portion of the data. From a boxplot, you can see the five-number summary exactly and outliers are clearly marked. These must be estimated from a histogram and can be difficult to estimate. From a histogram, you can estimate the mean by estimating a balance point for the distribution. You cannot do this with a boxplot. A histogram will reveal the frequency of the data within an interval. You do not know the exact values, but you know how many are within the given boundaries. You know a lower and upper bound but not necessarily the exact least and greatest value. You know where there are clusters of data and where there are gaps. With a boxplot, you get a sense of the basic shape of the distribution, but you cannot see clusters or gaps.

D18. a. Those with dark blond hair color have the largest standard deviation, 9.28. The smallest standard deviation, 5.45, is exhibited by the light brunettes.

b. The values of the means for the four groups differ quite a bit and decrease from 59.20 for light blonds to 37.40 for dark brunettes. One conclusion is that dark brunettes are most sensitive to pain while lights blonds are least sensitive.

c. It is difficult to believe that a person's hair color is related to their ability to withstand pain, so one concern is that the observed differences are the result of some other quantity that was not being measured. Another concern is the size of the samples (4 and 5), which are quite small sample sizes for drawing conclusions. A simple concern is that the hair color recorded for a subject may not be the subject's actual hair color.

D19. In both cases, the formulas involve the square root of a sum of squared differences. To compute the standard deviation, find the differences, square, average, and take the square root. To find the distance between two points, find the differences, square, add, and take the square root. For example, suppose that you have a set of three values, {1, 4, 10}, with mean 5. Then, except for dividing by the sample size, the standard deviation is the same thing as the distance in space between the point (1, 4, 10) and the point of the mean (5, 5, 5):

$$\sqrt{(1-5)^2+(4-5)^2+(10-5)^2}$$

Thus, the measure of distance in statistics, the standard deviation, has almost exactly the same form as the measure of distance in Euclidean geometry. Perhaps this reason, above all others, helps convince students that the standard deviation is a natural measure of spread.

D20. Dividing by a slightly smaller number makes the standard deviation a bit larger. The difference between a value that is divided by n and that same value divided by $n-1$ is greater with smaller values of n. For example, let $n = 5$ and the value be 100, then the value divided by n is $\frac{100}{5}$ or 20, while the value divided by $n-1$ is $\frac{100}{4}$ or 25. Now if you let $n = 50$, then $\frac{100}{50}$ is 2, while $\frac{100}{49}$ is about 2.04. So, in the first example, the difference between the two possibilities is 5, in the second example, 0.04. Much smaller.

D21. The formula in the boxed summary uses multiplication as a shortcut for addition. Instead of adding the value $(x-\bar{x})^2$ a total of f times, you can just multiply $(x-\bar{x})^2$ by f. For the example in Display 2.57 on page 60, n is the total number of years, or 18, x is the number of strikes, and f is the number of years that have that particular value. Again, the formula substitutes multiplication for repeated addition.

D22. Multiply the number of houses by the mean value to get the total value of the houses, then multiply by the tax rate. Knowing the mean is equivalent to knowing the total because the total is equal to the number of values times the mean. If you know the

number of houses, you can easily convert between the mean value and the total value.

D23. Income is strongly skewed right. A few people make a lot of money, whereas most people are clustered together toward the low end. Consequently, the mean looks larger than most people think is "typical." The median tells you that half of the residents earn more and half earn less. Another reason the median may be given is that the median income is probably easier to estimate because values of the high incomes would be difficult to obtain.

D24. a. mean $= \$20 \cdot 14.5 \frac{\text{pesos}}{\text{dollar}} = 290$ pesos; $SD = \$5 \cdot 14.5 \frac{\text{pesos}}{\text{dollar}} = 72.5$ pesos

b. median = \$10; Q_1= \$2; Q_3= \$5

D25. median $= \frac{5}{9}(26 - 32) = -3.33$ degrees Celsius;

$SD = \frac{5}{9}(25.6 - 32) = -3.56$ degrees Celsius

D26. Since 1 minute is equal to 60 seconds, and the mean, median, standard deviation, and interquartile range all have units of seconds, convert each statistic to minutes by dividing the computed value by 60. The computed variance would have units of seconds squared. Transform the variance to minutes by dividing by 60^2 or 3600.

D27. a. The mean is affected by the outliers. Because you first add all the values to compute the mean, an unusually large value will increase this sum quite a lot. Outliers tend to have greater influence in small samples than in large ones.

b. The numerical value of an outlier does not affect the median because the size of the largest or smallest value doesn't affect which value is in the center. But removal of an outlier from a data set may cause some change in the median because of a change in the sample size.

D28. a. Yes. The range is greatly affected by an outlier because it is computed by subtracting the largest and smallest values.

b. Yes. Generally, the standard deviation is affected greatly by an outlier. To compute the standard deviation, first square the differences from the mean. If one of these differences is large, squaring it makes it even larger.

c. No. The numerical value of an outlier does not affect the interquartile range because the quartiles aren't affected by the size of the maximum or the minimum. However, removal of an outlier from a data set may cause some change in the quartiles because of a change in the sample size.

D29. a. about the 19th percentile; about 79 years

b. 5 countries

c. Both the plot in Display 2.77 and 2.78 rise as you move to the right of the graph. This would be true of any such percentile graph. The life expectancy data in Display 2.78 represents a much smaller data set than the GMAT data. In Display 2.78, one can see each individual observation in the plot. Sketches for the actual life expectancies in European countries will vary. Life expectancies in an individual country may be skewed left if the mean expectancy is relatively large and skewed right if life expectancies are typically short. It is possible that pooling the life expectancies for all countries results in a distribution approximately normal in shape.

D30. a. 0

b. The lower quartile is the *z*-score that has 25% of the values below it. Looking in the center of the Table A on page 759, the *z*-score with a percentage closest to 25 is -0.67.

c. A percentage of 95 lies right between *z*-scores of 1.64 and 1.65. So it is reasonable to use $z = 1.645$

d. The *z*-score with a percentage closest to 75 is 0.67. The IQR is then $0.67 - (-0.67)$ or 1.34.

D31. a. $\frac{200 - 80}{60} =$ 2 hours

b. Subtract (recenter), and then divide (rescale). That is, how far from the exit? How many hours is that?

D32. You cannot use the normal distribution to solve this problem because the distribution of ages of cars is not approximately normal. In fact, it is strongly skewed right. You can see that because the ages cannot go below 0 and that is only 1.5 standard deviations below the mean. So, there must be quite a few ages far out in the upper tail of the distribution.

D33. From Table A, the area to the left of $z = -3$ is 0.0013. The area to the left of $z = -3$ is 0.9987. Thus, the area between these values is $0.9987 - 0.0013 = 0.9974$.

Practice Problem Solutions

P1. a. The data were collected by a statistics class; a case is a penny; the variable is the age of the penny.

b. The shape is strongly skewed right, with a wall at 0. The median is 8 years, and the spread is quite large, with the middle 50% of ages falling between 3 and 15 years; however, it is not unusual to see a penny that is more than 30 years old.

c. If the same number of pennies is produced each year and if a penny has the same chance of going out of circulation each year, then the height of each column would be a fixed percentage of the previous height. The distribution is skewed right since pennies most recently manufactured would naturally be more readily available in circulation. The older the penny, the less likely it would be for it to arise in a sample simply because it is likely that fewer of them remain in circulation.

P2. a. w: skewed to the left;
x: mound-shaped or approximately normal;
y: two mounds, possibly representing two groups of objects

b. x because it tapers symmetrically on either side of the maximum of the mound, while the other two are either skewed or have two mounds.

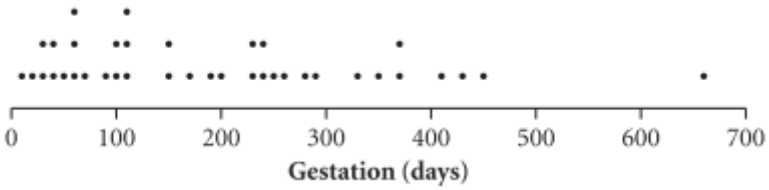
c. The mean is around 52 because most values are equally spaced around 50, with two extra values to the right of 60 that will serve to increase the mean a bit. The standard deviation is around 5.

P3. Student estimates will differ somewhat from the actual means and standard deviations given here.

a. A typical SAT math score is roughly 500, give or take about 100 or so.
b. A typical ACT score is about 20, give or take 5 or so.
c. A typical college-aged woman is about 65 inches tall, give or take 2.5–3 inches or so.
d. A typical professional baseball player had a single-season batting average of about .260 or .270, give or take about .040 or so.

P4. There are 180 dots, so assuming the data are ordered from lowest value to highest value, the lower quartile coincides with the 45^{th} dot, the median is the average of the 90^{th} and 91^{st} dots, and the upper quartile is the 135^{th} dot. As such, we have median = 19; lower quartile = 17; upper quartile = 25, Also, note that the distribution is skewed toward the higher ages.

P5. a.

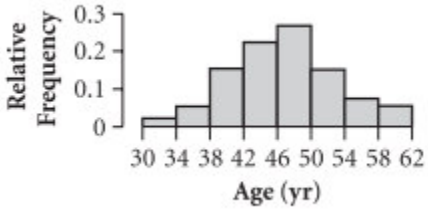


b. The distribution is skewed right with no obvious gaps or clusters. There is a wall at 0 days because no mammal can have a smaller gestation period. The elephant is the only possible outlier.

c. About half of the mammals have a gestation period of more than 160 days, and half have a shorter period. Those in the middle half have gestation periods between 63 and 284. Large mammals have the longer gestation periods.

P6. a. The distribution is approximately normal with mean around 46 and standard deviation around 6.

b.



c. The proportion of people who are at least 50 years old corresponds to the sum of the heights of the three rightmost bars in the histogram in (b), namely about 0.24, or about 24%.

P7. a. The proportion of countries with a life expectancy of less than 50 years is about 0.008 + 0.01 + 0.05 + 0.04 = 0.108, or about 10.8%.

b. The number of countries with a life expectancy of less than 50 years would then be (0.108)(223) = 24.08, or about 24 countries.

c. The shape of this distribution is skewed left toward the smaller values. The median is between 70 and 75 and the middle 50% of the life expectancies are between 60 and 75 years.

P8. a.

Back-to-back stem-and-leaf plot

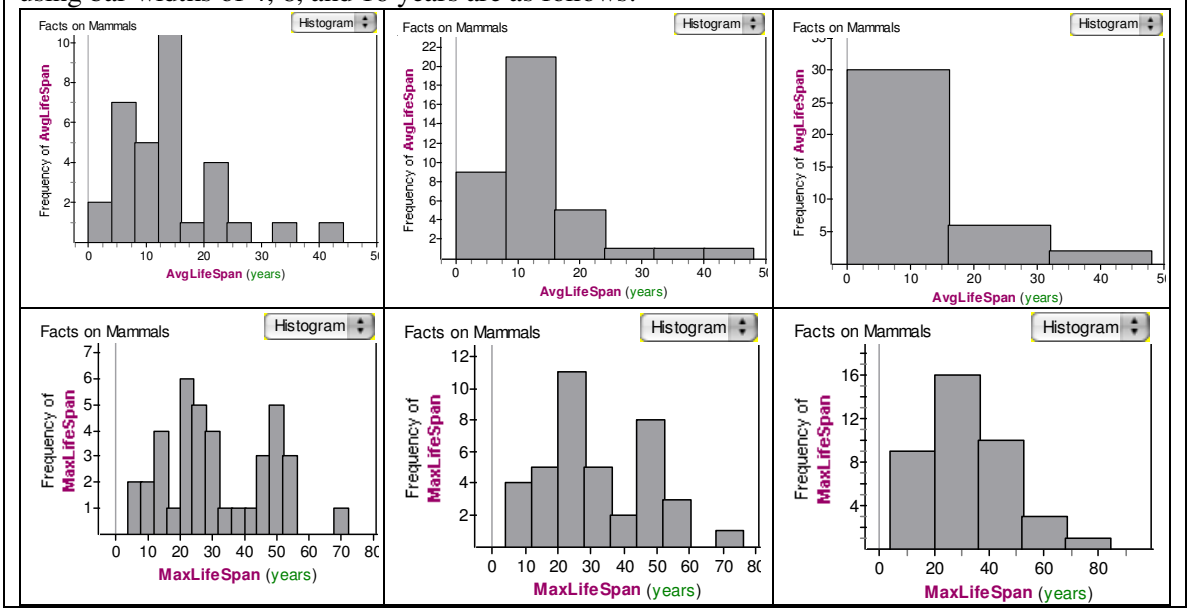
Ave Long		Max Long
431	0	4
88776555	0	588
222222222000	1	3344
65555555	1	8
0000	2	0000334
5	2	67778
	3	0004
5	3	7
1	4	0
	4	557
	5	00000344
	5	
	6	
	6	
	7	0

51216 represents a mammal with an average life span of 25 years and a mammal (not necessarily the same mammal) with a maximum life span of 26 years.

b. The average longevity distribution is skewed right, with two possible outliers, while the distribution of maximum longevity is more uniform but has a peak at 20–30 years and a possible outlier. The center and spread of the distribution of maximum longevity are larger.

c. Maxima are larger than averages by definition; maxima tend to spread out more because of the possibility of extremely large values, not constrained by averaging.

P9. Three pairs of histograms for average longevity (1st row) and maximum longevity (2nd row) using bar widths of 4, 8, and 16 years are as follows:



The shape of the maximum longevity distribution is quite different from that for average longevity. The average longevity distribution is skewed right with two possible outliers at 35–40 and 40–45. The distribution of maximum longevity is more uniform but with a peak at 20–30 years and a possible outlier at 70–80 years. As must be the case, the center of the distribution of maximum longevity is much higher than the center of the distribution of average longevity—about 30 years compared to about 15 years. The spread of the distribution of maximum longevity is also much larger.

The bar width that seems most appropriate in this scenario is 8 years since it preserves the general shape of the graphs using a bar width of 4 years (as compared to distorted graphs obtained using a bar width of 16 years), while not focusing on too much specific detail that is not important to understand the general trend.

P10. The number of deaths per month is fairly uniform, with about 190,000–220,000 per month. Summer months have the smallest numbers of deaths, and winter months the largest.

P11. A case is a student in your class. The quantitative variables are age, number of siblings, and number of miles he or she lives from school. The categorical variables are hair color and gender.

P12. The last digits of social security numbers are essentially random digits, so they should be fairly uniform over the range 0 to 9, as these are. Here we would say that the distribution is approximately uniform with about five–six students with each digit from 0 to 9.

P13. a. The cases are the individual male members of the labor force aged 25 and older, and the variable is their educational attainment. The distribution shows an increasing proportion of males through the first three levels of education with a huge jump at the high school graduation level. After high school, the proportions in each education category decrease regularly with increasing education levels, except for a spike at bachelor's degree.

b. The distributions for males and females have much the same shape and much the same proportions. Females have lower proportions in categories 8 and 9 and higher proportions in categories 4 and 5.

Relative frequency bar charts are better for this comparison because the number of males and the number of females in the labor force are different.

P14. **a.** mean: 2.5 median: 2.5
b. mean: 3 median: 3
c. mean: 3.5 median: 3.5
d. mean: 49.5 median: 49.5
e. mean: 50 median: 50

P15. The new mean height will be about 4 feet 4 inches. The median should not change much because it will still be one of the 3rd graders, who all are about 4 feet tall.

P16. a. The median life expectancies are 55 for Africa and 80 for Europe.
b. For Africa, the median is smaller than the mean (56.5) because of the skewness toward the larger values. For Europe, the mean (79.6) is slightly lower than the median because of the left skew.

P17. a. quartiles: 2 and 5 IQR: 3
b. quartiles: 2 and 6 IQR: 4
c. quartiles: 2.5 and 6.5 IQR: 4
d. quartiles: 2.5 and 7.5 IQR: 5

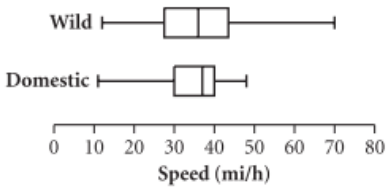
P18. a. The median, lower quartile, and upper quartile (in that order) for predators and nonpredators are given by:

predators: 12, 7 and 15;
nonpredators: 12, 8 and 15

b. The distribution of the average longevity of predators is mound-shaped, centered at about 12 years, with 50% of the values falling between 7 and 15 years. The distribution of the average longevity of nonpredators is centered at exactly the same place and has about the same spread, but it has two outliers on the high side. Its shape is essentially mound-shaped although the two outliers make it appear skewed towards the larger values.

P19. The distribution from the European countries shows a little skewness toward the smaller values with a median around 68, indicating that most countries have predominantly urban populations. Liechtenstein is a small, independent, principality in the Alps between Switzerland and Austria with a population of only 33,400. The distribution from the African countries shows a strong skewness toward the larger values, with a median near 40, indicating that most countries have predominantly more rural populations. The spread in the African distribution is larger than the spread in the European distribution, i.e., there is more variation among the percentages for African countries.

P20. The fact that the median line is not present suggests that it coincides with one of the quartiles. There is much more variability in the average longevity for wild mammals than for domesticated mammals. The side-by-side boxplots for the speeds of wild and domesticated mammals are below:



P21. a. The five-number summary for the average longevity of this set of mammals is

minimum: 1
lower quartile (Q_1): 8
median: 12
upper quartile (Q_3): 15
maximum 41

Once again, software packages may give quartiles slightly different from those done by hand. Here are the results from Minitab:

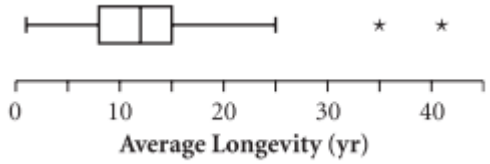
```
Variable N    N*    Mean   Median   TrMean StDev SEMean
Ave long 38   1     13.13  12.00    12.32  8.00  1.30

Variable Min  Max   Q1     Q3
Ave long 1.00 41.00 7.75   15.00
```

b. IQR = 15 – 8 = 7

c. $Q_1 - 1.5 \cdot IQR = 8 - 1.5(7) = -2.5$. There are no outliers on the lower end.

d. $Q_3 + 1.5 \cdot IQR = 15 + 1.5(7) = 25.5$. The life spans of 35 years for the elephant and 41 years for the hippopotamus are outliers. The largest value that isn't an outlier is 25. This is where the upper whisker will end.

e.



P22. The mean is 4.4, and the deviations from the mean are shown in this table. The sum of these deviations is 0. To get the standard deviation, find the (n – 1) average of the squared deviations, $\frac{41.2}{4} = 10.3$, and take the square root to get about 3.21.

Value x	Deviation from Mean: $x - \bar{x}$	Squared Deviations: $(x - \bar{x})^2$
1	$1 - 4.4 = -3.4$	$(-3.4)^2 = 11.56$
2	$2 - 4.4 = -2.4$	$(-2.4)^2 = 5.76$
4	$4 - 4.4 = -0.4$	$(-0.4)^2 = 0.16$
6	$6 - 4.4 = 1.6$	$(1.6)^2 = 2.56$
9	$9 - 4.4 = 4.6$	$(4.6)^2 = 21.16$
Sum	**0**	**41.20**

P23. a. i. 0, because all of the deviations from the mean are 0

b. iii. 0.577

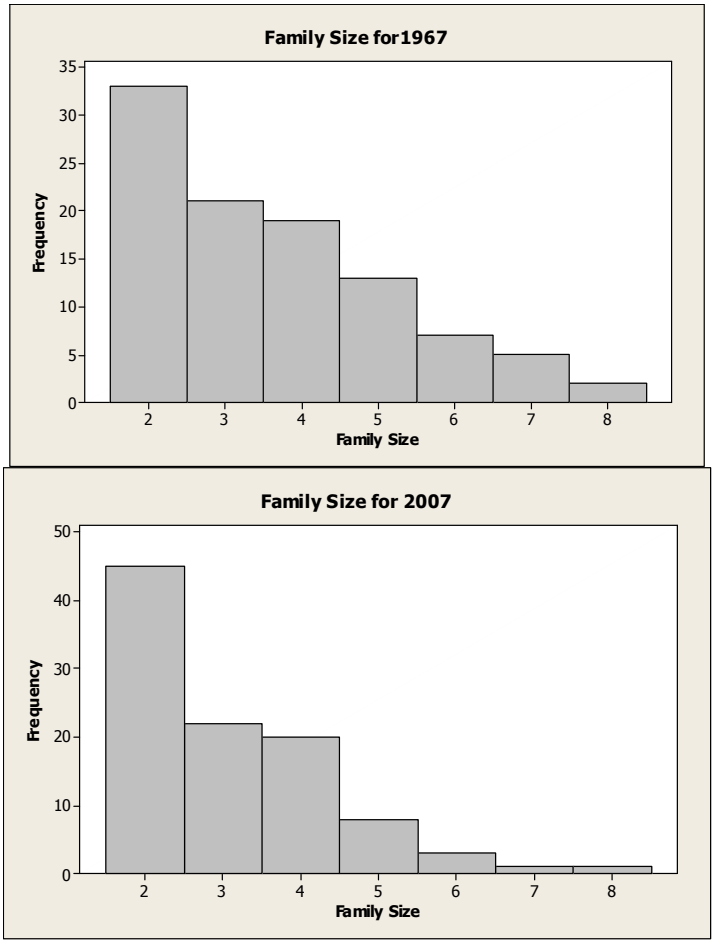
c. iv. 1.581

d. vii. 5.774. Note that the values are 10 times as far from the mean as those in part b, so the standard deviation is 10 times as large.

e. ii. 0.058. Note that the values are one-tenth as far from the mean as those in part b, so the standard deviation is one-tenth as large.

f. v. 3.162. Note that the values are twice as far apart as those in part c, so the standard deviation is twice as large.
g. vi. 3.606. It may be hard for students to distinguish part f from part g. If so, they should compute the standard deviation to check their answer.

P24. a.



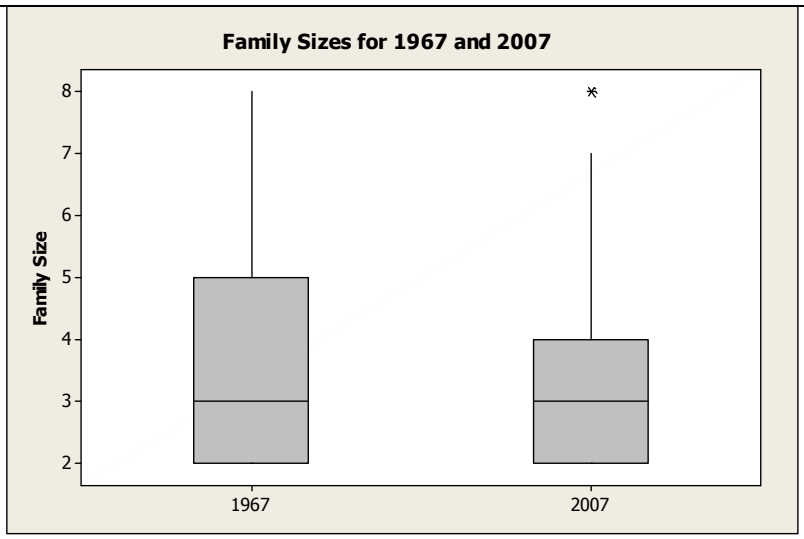
b. 1967: mean = 3.63, SD = 1.61
2007: mean = 3.09, SD = 1.29;

Observe that family sizes got smaller and less variable from 1967 to 2007.

P25. a. The median is the average of the 50^{th} and 51^{st} data value (assuming the data is ordered from lowest to highest); this value is 3 for each year.
b. Note that the lower quartile is the 25^{th} data value, the upper quartile is the 75^{th} data value, and the IQR is the difference of these two. Computing therefore yields

1967: Q_1=2, Q_2=3, Q_3=5, IQR = 3
2007: : Q_1=2, Q_2=3, Q_3=4, IQR = 2

c.



d. Family size has decreased slightly over the 40 year period from 1967 to 2007. The median size is 3 for both years, but the mean size has decreased from 3.63 to 3.09. In addition, there was less variability among family sizes in 2007 than there was in 1967.

P26. a. The mean is larger than the median because house values tend to be skewed right—some houses cost a lot more than most houses in a community. Very few houses cost a lot less.

b. To get the total property taxes, multiply the number of houses by the mean value by the tax rate to get ($392,059)(9,751) (0.0115) = $43,964,124.

c. This is an average of $4,508.68 per house. (This assumes that the assessed value is equal to the price.)

P27. a. Medians were used in this story because the distribution of car ages is strongly skewed right. There are more brand-new cars on the road than cars of any other age (because cars of any other age have been disappearing due to accidents and mechanical problems). A few people drive very old cars.

b. Vehicles are proving more durable, so cars can be driven for a longer time. Also, people are possibly choosing to spend their income on other things or are forced to spend their income on other things.

P28. a. Divide each of the summary statistics by 12 so then the mean is 4 feet, the median is 3.75 feet, the standard deviation is 0.2 feet, and the interquartile range is 0.25 feet.

b. Add 2 to the measures of center so the mean is 50 inches, and the median is 47 inches. The standard deviation and interquartile range do not change.

c. For the measures of center, add 4 then divide by 12 so the mean is $4\frac{1}{3}$ feet, the median

is $4\frac{1}{12}$ feet. The measures of spread are divided by 12 (adding 4 to each height does not change the spread) so the standard deviation is 0.2 feet, and the interquartile range is 0.25 feet.

P29.

a. mean: 12; SD: 1
(Add 10 to the mean from the given example, SD is the same as in the example)

b. mean: 20; SD: 10
(Both the mean and SD are 10 times those in the example)

c. mean: 110; SD: 5
(Each value in the example has been multiplied by 5 and then had 100 added to it. Thus, the mean is 5 times that in the example, plus 100, and the SD is 5 times that in the example.)

d. mean: 900; SD: 100
(Each value in the example has been multiplied by 100 and then had 700 added to it. Thus, the mean is 100 times that in the example, plus 700, and the SD is 100 times that in the example.)

P30. a. Outliers occur above –30 + 1.5(21) = 1.5. Hawaii, at 12, is an outlier.

b. The count will decrease by 1 to become 49.

Summary of Lowest Temperature without Hawaii

No selector

Percentile	25
Count	49
Mean	−41.5
Median	−40
StdDev	16.2
Min	−80
Max	−2
Range	78
Lower ith %tile	−51
Upper ith %tile	−32

The minimum, median, quartiles, and interquartile range should remain the same or about the same. With a sample of size 50, removing one data value will have little effect on the median but could have greater effect on the quartiles as they are, in essence, the medians of samples of size 25. The mean should go down by a bit more than one degree because the difference between Hawaii's temperature and the mean is about –52 degrees and there are 49 states remaining. The standard deviation should go down only slightly, but this is difficult to predict. The range will decrease from 92 degrees to about 80 degrees. The maximum will decrease from 12 to a little less than zero.

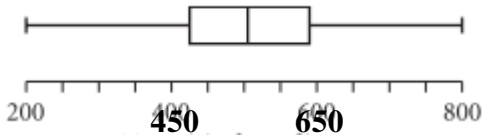
P31. a. 18th

b. The middle 90% is between about 325 and 730, while the middle 95% between about 275 and 750.

c. about 510

P32. 90% of cases lie between the 5th and 95th percentiles (since 5% of cases lies to the left of the 5th percentile and 5% lie to the right of the 95th percentile, by definition). The middle 95% of data is enclosed between the 2.5th percentile and the 97.5th percentile.

P33. The quartiles are about 450 and 650; the median is about 550; the IQR is about 650 – 450 = 200. The minimum of 200 and the maximum of 800 complete the five-number summary. A boxplot is shown here:



450 **650**

GMAT Scores

P34. **a.** 1.29% **b.** 4.75% **c.** 34.46% **d.** 78.81%

P35. **a.** –0.47 **b.** –0.23 **c.** 1.13 **d.** 1.55

P36. **a.** 0.9279 – 0.0721 = 0.8558 or 85.58%.
b. 0.9987 – 0.0013 = 0.9974 or 99.74%.

P37. **a.** The z-score that has 5% of the values below it is –1.645, and the z-score that has 5% of the values above it is 1.645. So the interval is –1.645 to 1.645.
b. –1.96 to 1.96

P38. **a.** The death rate for cancer is more standard deviations below the mean, so it is more extreme.

$$\text{heart disease: } z = \frac{180-219}{46} = -0.8478$$

$$\text{cancer: } z = \frac{151-194}{30} = -1.4333$$

b. The death rate for cancer is more standard deviations above the mean, so it is more extreme.

$$z_{\text{heart}} = \frac{260-219}{46} = 0.8913$$

$$z_{\text{cancer}} = \frac{228-194}{30} = 1.1333$$

c. The death rate for heart disease in Colorado is more extreme than the death rate for cancer in Georgia.

$$z_{\text{CO heart}} = \frac{135-219}{46} = -1.8261$$

$$z_{\text{HI cancer}} = \frac{158-194}{30} = -1.2000$$

P39. a. The z-score for the height 6 feet (or 72 inches) is $\frac{(72-70.4)}{3.0} = 0.53$. The area under the normal curve to the right of this point is 1 – 0.7019 = 0.2981. Thus, about 29.8% of U.S. males between 20 and 29 are taller than 72 inches.

b. The z-score for the 35th percentile is –0.385. The height that corresponds to that z-score is

$$x = mean + z \cdot SD = 65.1 + (-0.385)2.6 \approx 64.1 \text{ inches}.$$

So, a woman would need to be 64.1 inches tall to be at the 35th percentile.

P40. a. 219 ± 1.645(46) or about 143 to 295
b. 219 ± 0.99(46) or about 173 to 265

P41. Recall from P39 that the z-scores are computed as follows:

Male: $z = \frac{x-70.4}{3.0}$ Female: $z = \frac{x-65.1}{2.6}$

a. The z-score is 2.87, which is outside both intervals.
b. The z-score is 1.12, which is not outside either interval.
c. The z-score is -1.80, which is not outside either interval.
d. The z-score is 1.88, which is not outside either interval.

Exercise Solutions

E1. I. Graph D, because on an easy test most people get high scores.
II. Graph A, because the distribution of heights has two modes (mothers and daughters).
III. Graph C, because most countries in the Olympics get no medals at all and only a very small number of countries get multiple medals.
IV. Graph B, because the weights should be mound-shaped. Most chickens will be clustered near a central weight with decreasing numbers having lower or much higher weights.

E2. The distribution is approximately normal, except that it has too many outliers and is a bit too "peaked" to be traced by a normal curve. Because it is roughly symmetric, we can say that the distribution is centered at about 98.

E3. a. A case is one of the approximately 92 officers who attained the rank of colonel in the Royal Netherlands Air Force. There is only one variable: the age at which the officer became a colonel.

b. This distribution is skewed left with no outliers, gaps, or clusters. The median is 52 years, and the quartiles are 50 and 53. So we would say that the middle half of the ages are between 50 and 53, with half above 52 and half below.

c. In questions like this, students aren't expected to "guess" the "right" answer. Instead, they are expected to generate many possibilities that could then be investigated. For example, some military services have an "up or out" rule. It may be the case in the Royal Netherlands Air Force that if you haven't been promoted to colonel by your 55th

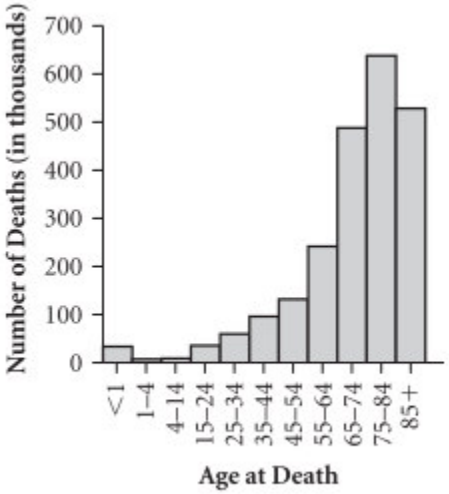
birthday, you must retire. The armed services have very little use for 55-year-old privates. Alternatively, there might be a mandatory retirement at age 55. A third possibility is age discrimination against older people in the service.

E4. a. A case is one of the 101 seasons. The only variable is the number of inches of rain. Note: Rainfall is given by "seasons" rather than by years because Los Angeles has almost no rain from May through October. The rainy season is late fall through early spring. To give rainfall by year would be to break each of these natural seasons into two parts.

b. This distribution is skewed right with no obvious outliers. The three peaks suggest three groups of data points; perhaps weather conditions cause dry, normal, or rainy years. The median is 13, and the quartiles are 10 and 19. So we say that half the values are above 13 and half are below, with the middle half between 10 and 19.

c. The number of inches of rain cannot go below 0, so that is a natural wall. However, it appears that about 4 inches of rain is the effective minimum. There may be some characteristic of the weather that makes it almost impossible to go below that.

E5. a. This distribution is strongly skewed left. The distribution for an actual year follows, and the shape of the distribution is typical. Students will often have the height of the bar for 85+ taller than that for 75–84, confusing actual number of deaths with probability of death. There are fewer people in the 85+ category than in the 75–84 category, so fewer of them die.



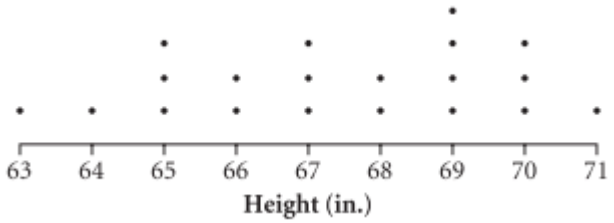
Source: *Statistical Abstract of the United States*, 1997, Table 130.

b. This distribution will be strongly skewed right. Most people get their driver's licenses at the earliest possible age or quite close to it.

c. The distribution of SAT scores for a large number of students should be approximately normal.

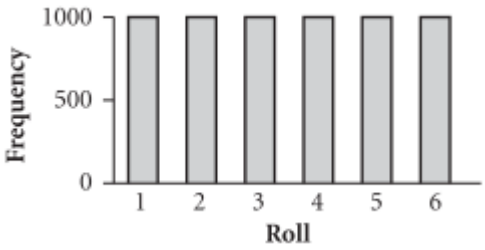
d. Selling prices of new cars should show a few very expensive models (like Corvettes) and a large number of relatively inexpensive (but not cheap!) ones (around $15,000 to $20,000). The distributions should be skewed right (toward the larger values.)

E6. a. skewed right (toward larger values)
b. bimodal; developing countries tend to have higher birth rates than do developed countries.
c. Approximately normal is a good answer. In fact, the distribution is very slightly skewed left. The dot plot shown next gives the heights in inches of the German women's soccer team that won the World Cup in 2003.

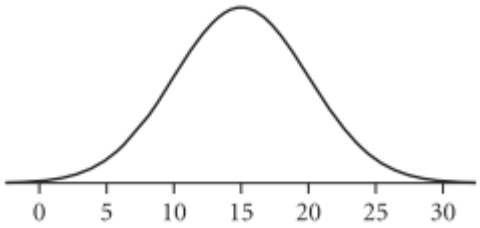


d. roughly uniform
e. skewed left (toward smaller values) with a wall at 50 minutes.

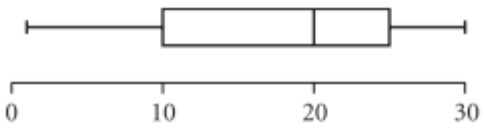
E7. a.



b.

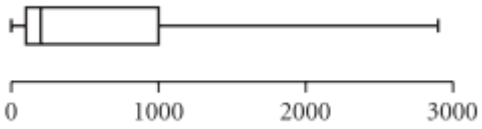


c. There are many possible answers. One such set of numbers is 1, 2, 3, 4, 9, 11, 13, 14, 17, 19, 21, 22, 23, 23, 24, 26, 27, 28, 29, 30. This boxplot represents one possible sketch.

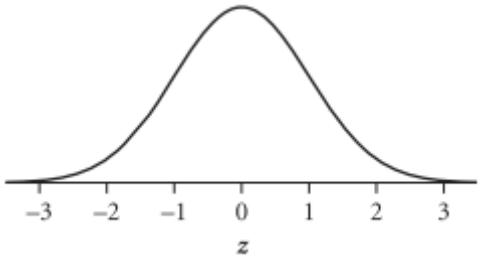


d. There are many possible answers. One such set of numbers is 0, 10, 20, 30, 80, 120, 150, 160, 170, 190, 210, 400, 500, 600, 700, 1300, 1500, 1800, 2400, 2900. This boxplot

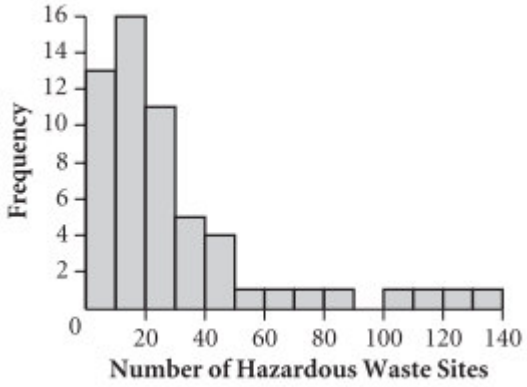
represents one possible sketch.

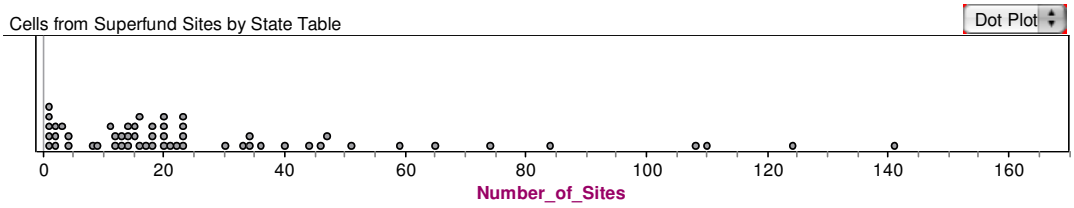


e.

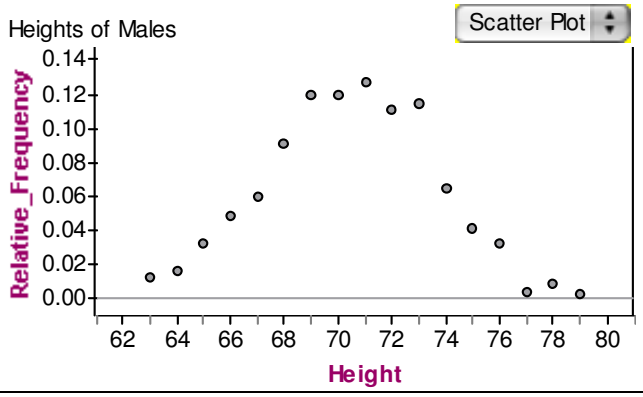


E8. Make a plot that is skewed right. Here are a histogram and dot plot of the actual distribution:



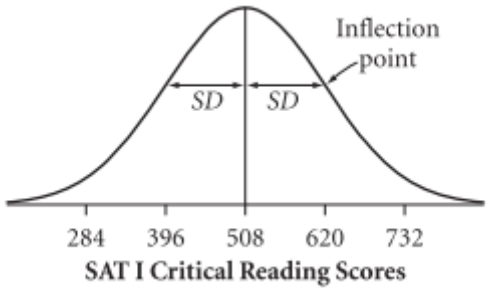


E9. a. Imagine connecting the dots in the scatterplot below by a normal bell curve:

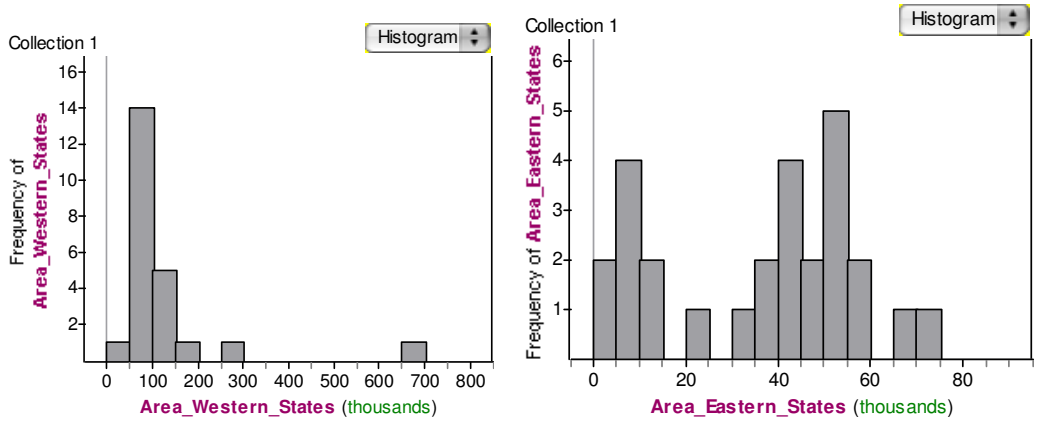


b. The mean is the balance point of the distribution, which occurs at about 71 in. The standard deviation is about 3 in.
c. This is obtained by adding the heights of all vertical bars to the left (and including) the one at 74 inches, which is about 0.90.
d. This is obtained by adding the heights of all vertical bars strictly to the left of the one at 68 inches, which is about 0.10.
e. The distribution isn't smooth and ranges only from about 63 to 79 inches.

E10. a. For the SAT I Math scores in 2004-2005, the mean was 518 and the standard deviation was 114. The mean can be estimated at about 500 because that is near the center of the distribution and that the standard deviation is about 100 because about two-thirds of the values lie between 400 and 600.
b. Roughly 67% of math scores are within one standard deviation of the mean (since the distribution is symmetric and basically normal).
c. Answers will vary based on the estimates from (a). For the response given, we have:



E11. a. The western states tend to be large, the eastern states small.
b. Dividing the data into two groups, one for *western* states and the other for *eastern* states, we have the following two histograms:



Yes; both groups cluster together, although eastern states spread out more.

E12. a. The population density of Vermont is $\frac{621{,}000}{9{,}614} \approx 64.6$

b. New Jersey is certainly a potential outlier, while Rhode Island and Connecticut are less extreme, but still reasonably far away from the bulk of the data points; the same is true, but to a lesser extent, for Massachusetts. However, one must know the mean and

standard deviation to be able to conclude which of these, if any, are outliers for certain.

c. Alaska is not an outlier because it is not far removed from the bulk of the data values.

E13. a. The distribution is wide spread, with two clusters (low GDP and media GDP) and one extremely large value.

b. Norway and Switzerland have the highest per capita GDP; they are not outliers because there is much variability in the data.

c. The higher GDPs belong to Western Europe and North America; the lower GPDs belong to Eastern Europe and Asia.

d. No, neighboring countries tend to have similar economies, producing a clustering effect.

E14. a. This distribution is skewed right. The median is 1 person per room with half the countries having between 0.6 and 1.3 people per room. Again, answers will vary depending on your judgment concerning outliers. In Section 2.3 you'll learn that $1.5 \cdot \text{IQR} = 1.05$, so that any points above 2.35 would be considered outliers. There are two outliers on the right.

b. The two outliers are Pakistan and India. These are not the same two countries that had the highest Per Capita GDP. In fact, these two countries had the lowest Per Capita GDP. This makes sense as poorer nations with large populations (those with lower Per Capita GDP) tend to have more crowded living conditions.

c. The Western European and North American countries are clustered mainly on the left. This too makes sense as richer nations tend to have more spacious living conditions because their population, in general, can afford it.

E15. a. With a perfect uniform distribution on [0, 2], the value 1.0 would divide the values in half.

b. The values 0.5, 1.0, and 1.5 divide the distribution into quarters.

c. The values 0.5 and 1.5 enclose the middle 50% of the data.

d. About 15% of the data lie between 0.4 and 0.7 because the length of this interval is 0.3, and dividing it by the length of the entire interval, namely 2, yields 0.15.

e. The values 0.05 and 1.95 enclose the middle 95% of data values.

E16. a. IV **b.** II **c.** V **d.** III **e.** I

E17. a. Most tend to predict that domesticated animals have greater longevity due to the relative safety and good care provided by their habitat.

b.

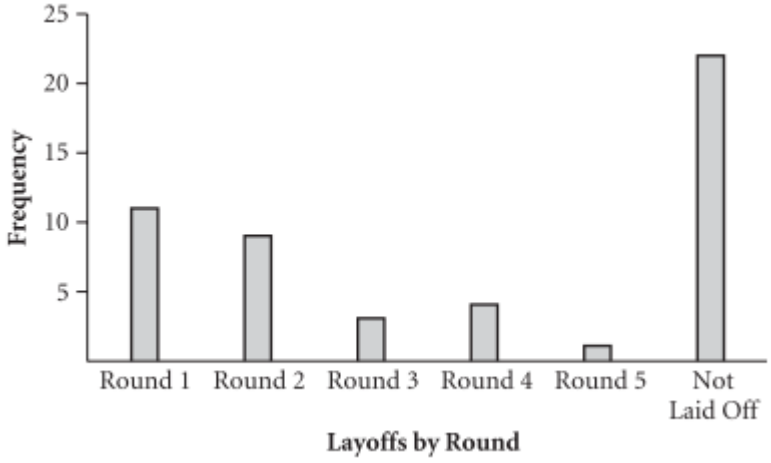
Domestic		Wild
4	0	13
85	0	556778
22220	1	0022222
5	1	5555556
0	2	000
	2	5
	3	
	3	5
	4	1

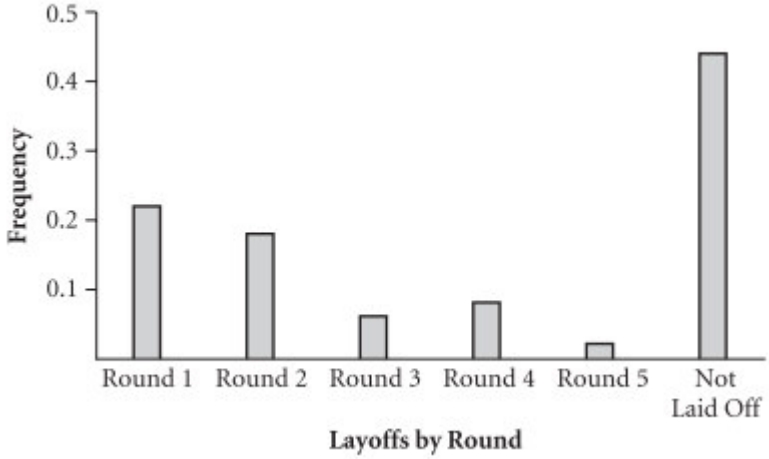
1|5 represents 15 years

c. Both distributions appear to be slightly skewed to the right with possible outliers on the high side. The median of both distributions is 12, but the spread of the distribution for wild mammals is quite a bit larger. The middle half of the domestic mammals have an average longevity approximately between 8 and 12 years. The middle half of the wild mammals have an average longevity between 7.5 and 15.5 years.

E18. For the United States, there are population bulges around the ages of 10–25 and the ages of 35 –50 for both men and women with decreasing percentages in the age groups above 50. For Mexico, the largest segments of the population are young children, with a regularly decreasing pattern in percentage of the population as the age increases. In both countries, there are more infant boys than infant girls. However, this reverses at the oldest ages, especially in the United States, where there are far more older women than older men.

E19. The bar charts are as follows:





From the bar chart, it appears that Westvaco laid off the majority of workers in Rounds one and two. All together, Rounds 3, 4, and 5 make up about the same number of workers as in Round 2 alone.

E20. Quantitative variables are year of birth, year of hire, and age. Categorical variables are row number, job title, round, and pay category (hourly or salaried). Month of birth and month of hire fall somewhat in between and are best called "ordered categories." Months are ordered and can be represented by numbers 1, 2, 3, … , 12 that can be meaningfully compared—they tell you whose birthdays come earlier in the year, for example. On the other hand, they aren't a count of how many or a measure of how much, so they would not be considered a quantitative variable.

E21. a. The heights of the first three bars are the number of nonpredators in the display that fall into the categories Domestic and Wild, and the total number of nonpredators. For example, the first bar shows that there are about 8 nonpredators that are domestic.

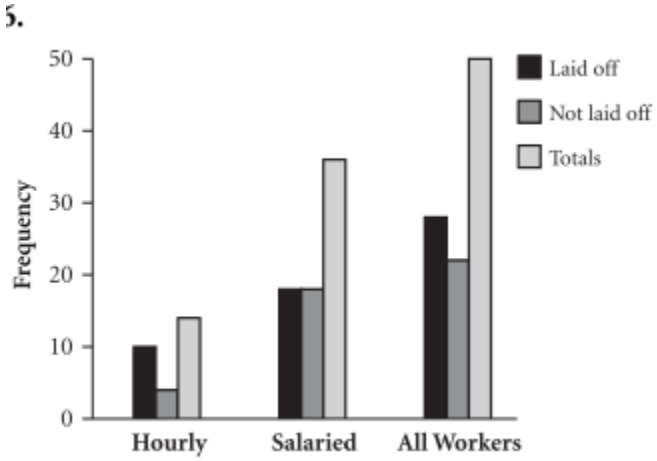
b. Looking at the middle set of bars, for predators, the second bar is taller than the first. Thus, a predator is more likely to be wild than domestic.

c. Looking at the first set of bars, for nonpredators, you can see that a nonpredator is also more likely to be wild than domestic because the second bar is taller than the first bar. However, for nonpredators, the first bar is a larger fraction of the second bar than is the case for predators. Thus, a predator is more likely to be wild than is a nonpredator.

Note: The way the bar chart is set up makes it easy to make the comparison asked for in part b but difficult to make the comparison in part c. You may wish to ask students to make a bar chart that makes it easy to answer part c. A two-way table like the one shown here can be helpful in summarizing the data on different categorical variables before making the bar chart.

	Nonpredator (0)	Predator (1)	Total
Domestic (0)	8	2	**10**
Wild (1)	19	10	**29**
Total	**27**	**12**	**39**

E22.



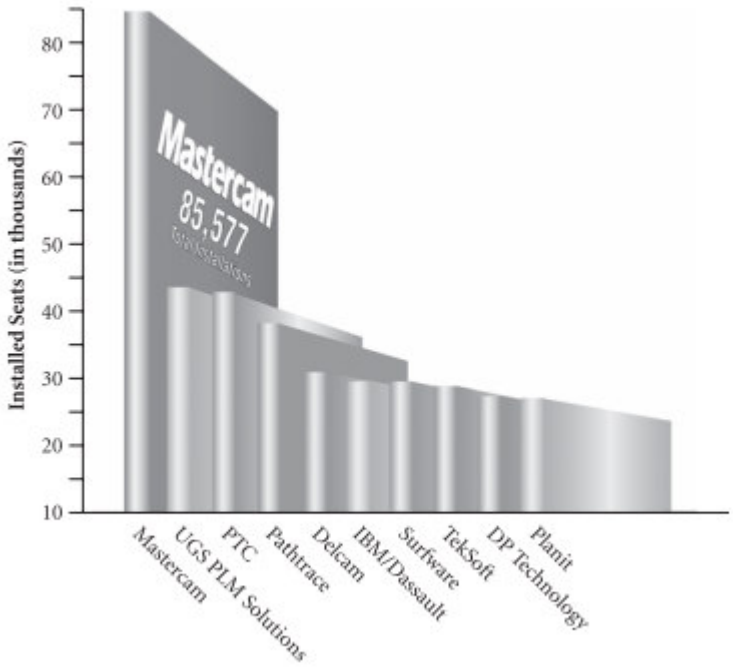
E23. a. Of the 18 mammals for which speeds are given, 12 have speeds that end in either 0 or 5.

b. Two-tenths of the 18 mammals, or 3.6

c. The most likely explanation is that the speeds are actually estimates for the wild mammals. Who is going to measure the speed of a grizzly bear in the wild? The speeds that don't end in 0 or 5 are for the dog, fox, giraffe, horse, pig, and squirrel. For these mammals, with the possible exception of the giraffe, you can see how speed could be measured accurately. (And it certainly is in horse races and dog races.)

E24. You may find examples of bar charts where the scale does not start at 0, making differences in the bars appear larger than they actually are. Graphs of stock market activity often do this. The intent is to emphasize changes, but it can be deceiving to the unaware. You may also find examples of bar charts where the scale does start at 0 and thus obscures differences that are important in the heights of the bars. Or, you may find examples of "picture" graphs where a three-dimensional picture makes, for example, one quantity that is twice as big as another look like it is 8 times as big, and so on. The classic book, *How to Lie with Statistics* by Darrell Huff (Norton, 1993), originally published in 1954, is still informative and entertaining reading. More up-to-date examples can be found in Edward R. Tufte, *The Visual Display of Quantitative Information* (Graphics Press, 1983).

Here's an example of a bar chart that doesn't start at 0, to the advantage of the advertiser.



E25. The seventh value is 10, as you can see from solving

$$25 = \frac{x + 24 + 47 + 34 + 10 + 22 + 28}{7}.$$

E26. $\frac{84}{n} = 6 \Rightarrow n = 14$

E27. a. The middle half of the speeds of domestic mammals are between 30 and 40. The spread for wild mammals is a bit larger—the middle half of the speeds are between 27.5 and 43.5. Half of the domestic mammals have speeds above 37 and half below. The median for wild mammals, again, is almost the same, at 36. Note that in each case the median is closer to the upper quartile than the lower quartile, indicating that the distribution of speeds may be skewed left.

b. The wild mammals are more likely to be predators than are the domestic animals (see E21), and the speeds of predators have the larger IQR.

E28. Of the 36 salaried workers, 18 were kept and 18 were laid off. The 18 salaried workers who were laid off had a median age of 53.5 and quartiles 42 and 61. The 18 salaried workers who were kept had a median age of 48 with quartiles 37 and 55. So the median age of the workers laid off was 5.5 years older, but the distributions had about the same IQR.

Salaried Workers' Ages

Kept		Laid Off
	2	3
9	2	
421	3	012
7	3	
2	4	2
8887	4	9
443	5	0234
975	5	669
10	6	134
	6	69

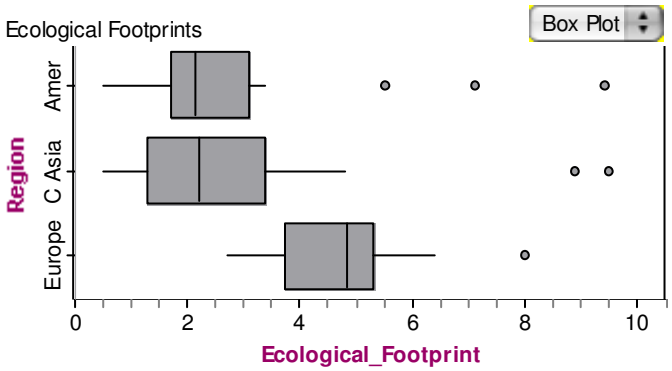
619 represents 69 years

E29. a. A – mound-shaped symmetric; B – skewed right;
C – skewed left; D - uniform

b. Keep in mind the main features of a boxplot, namely the maximum and minimum data values, and the three quartiles. Matching these distinct characteristics to the histograms yields: A – III; B- IV; C – II; D –I

E30. a. boxplot 3 **b.** boxplot 1 **c.** boxplot 2

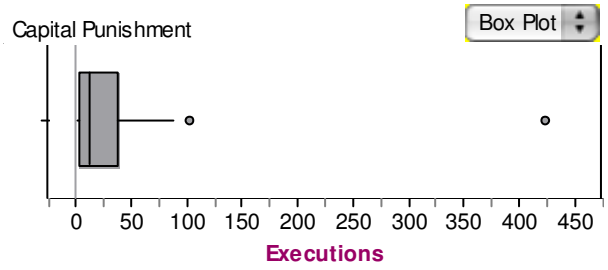
E31. a.



b-c. Distributions for all regions are skewed right, with at least one outlying country in each (United States, Canada and Uruguay; Kuwait and United Arab Emirates; Denmark). Central Asia and the Americas center around 2.5, with Europe centering around 5. Central Asia has the largest spread, followed closely by the Americas.

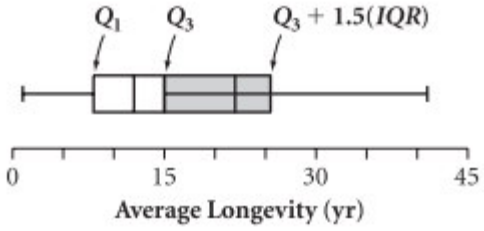
d. About 68%

E32. Begin by ordering the data values from lowest to highest. The mean of these values is 33.3235. The median (the average of the 17th and 18th ordered data values) is 3, the lower quartile (the 8th ordered data value) is 3 and the upper quartile (the 30th ordered data value) is 38. A boxplot summarizing this information is as follows:



E33. The third boxplot cannot be the plot for both classes combined because the minimum test score for the second period is about 10, and that would be the lowest for the combined set also. Students may come up with other possible reasons.

E34. a. The maximum value must be an outlier. The length of the box is the IQR, and Q_3 lies at the top of the box. An outlier lies beyond $Q_3 + 1.5(IQR)$. This point can be estimated from the boxplot by imagining stacking one-and-a-half boxes to the right of the box as illustrated here. The top of these boxes is less than the maximum value.



b. There are two outliers shown.

c. The given display splits the mammals into two groups, wild and domestic. The set of domestic mammals has a much smaller spread than the set of all mammals. The horse, whose average longevity is the outlier for domestic mammals, lives a long time compared to other domestic mammals. However, when comparing to all mammals, an average longevity of 20 years is not unusual. In general there is no reason to expect that the outliers of subsets of a set of data will be the same as the outliers in the set of data itself.

E35. a. II has the largest standard deviation, and III has the smallest.
b. II and III

E36. The second data set is the same as the first except that each value is 4 more. The distances between the values haven't changed, so the second data set has exactly the same spread as the first one. The standard deviation of the second data set, then, is also about 32.

E37. The set of heights of all female NCAA basketball players will have the larger mean because basketball players tend to be taller than other athletes, in general. The set of all female NCAA athletes will have the larger standard deviation because it will include tall, medium, and short athletes, whereas the set of all basketball players will include mostly tall athletes.

E38. a. You would replace the number closest to the mean with one that is furthest from the mean. Replace the 15 with a 1.

b. You would take the number furthest from the mean and replace it with the number closest to the mean. Replace the 32 with a 10.

c. There is no way to get an outlier.

E39. a. With Seinfeld, the midrange is $\frac{(2.32+76.26)}{2} = 39.29$. Without Seinfeld, the midrange is $\frac{(2.32+58.53)}{2} = 30.425$. The midrange is not resistant and is extremely sensitive to outliers because it is computed using only the maximum and minimum.

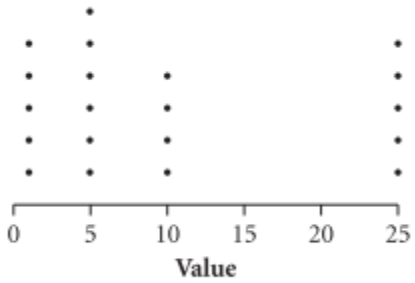
b. The total of the ratings with the Seinfeld episode is (101)11.187 = 1129.887. Subtracting the Seinfeld rating of 76.26 leaves a sum for the remaining programs of 1053.627. So the mean rating is $\frac{1053.627}{100} \approx 10.54$.

E40. a. Students will need to remove the top 2 longevities and the bottom 2 longevities from the data set and compute the mean of the remaining 35 animals. The trimmed mean is 30.5143.

b. Yes, because most outliers are removed before computing the trimmed mean.

E41. a. 3.11 grams
b. 0.04 gram
c. Yes, most of the weights are between 3.07 and 3.15.

E42. a. The mean is about 10.



b. The mean is shown below.

$$\bar{x} = \frac{\sum x \cdot f}{n} = \frac{1 \cdot 5 + 5 \cdot 6 + 10 \cdot 4 + 25 \cdot 5}{20} = \frac{200}{20} = 10$$

c. The standard deviation is closest to 10.

d. The standard deviation is shown below.

$$s = \sqrt{\frac{\sum (x-\bar{x})^2 \cdot f}{n-1}} = \sqrt{\frac{(1-10)^2 \cdot 5 + (5-10)^2 \cdot 6 + (10-10)^2 \cdot 4 + (25-10)^2 \cdot 5}{19}} \approx 9.4$$

E43. a. You can find the total of the scores in each class by multiplying the mean by the sample size, and this allows you to find the mean of the combined groups:

$$\bar{x} = \frac{30(75) + 22(70)}{52} = 72.9$$

This is called the weighted average of the two means.

b. You cannot find the median of the combined groups because to do so requires knowledge of the ordered arrangement of all the data values.

E44. a. Treating 2 million as the population of Kuwait and 200 million as the population of Indonesia, the weighted average of the proportions is given by:

$$\frac{2(.36) + 200(.04)}{202} = 0.043$$

About 4.3% of the population of the two countries thought the terrorist attacks were morally justified.

b. The average of the two percentages is 20%.

E45. $\sum (x-\bar{x}) = \sum x - \sum \bar{x} = \sum x - n\bar{x} = \left(\sum x\right) - \left(\sum x\right) = 0$

E46. $\sum (x-c) = 0 \;\Rightarrow\; \sum x - \sum c = 0 \;\Rightarrow\; \sum x - nc = 0 \;\Rightarrow\; c = \frac{\sum x}{n} = \bar{x}$

E47. a. Either could be used depending on the purpose of computing a measure of center. If, as is typical, there are a few expensive homes mixed in with many modestly priced homes, then the mean price will be larger than the median price. So real estate agents usually report the median price because it is lower and it tells people that half the prices are lower and half are higher. The tax collector would be interested in the mean price because the mean times the tax rate times the number of houses gives the total taxes collected.

b. As always, the answer depends on the purpose for computing a measure of center. Most likely, the reason here is to establish the total crop in Iowa for the year. In that case, it is best to find the mean yield per acre. This mean could be multiplied by the total acres planted in corn to approximate a total yield. An individual farmer probably would want to know the median as it gives the better indication of whether his or her yield was typical.

c. Again, the purpose of computing the measure of center determines which one you would use. Survival times are usually strongly skewed right. Telling a patient only the mean survival time would give too optimistic a picture. The smaller median would inform the person that half the people survive longer and half shorter. On the other hand, if you are the physician and must allocate your time by estimating the total number of

hours you will be caring for your patients with this disease, the mean would be better. You would then multiply the mean number of survival days by the number of patients you have by an estimate of the number of hours each day that each patient takes.

E48. a. The mean length of a generation. You would divide 300 years by the mean length of a generation to get the number of generations.
b. The mean speed. You multiply the mean speed by the time to get the distance.
c. Yes, if you know the number of trees. The average volume of one tree is

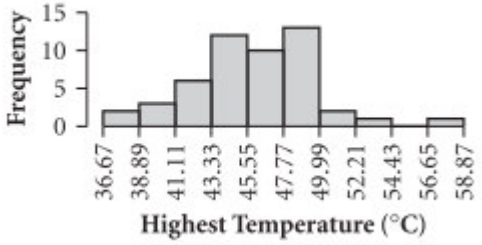
$$\pi r^2 h = \pi \cdot 3^2 \cdot 45 \approx 1272.35 \text{ ft}^3.$$

To get the total volume of wood, multiply by the number of trees.

E49. a. To make the histogram, students need only copy the histogram in the student text and then use the formula

$$C = \frac{5}{9}(F - 32)$$

to convert the numbers on the x-axis from °F to °C. The scale would then go from 36.67 to 58.89.



b. Note that the standard deviation is the tricky one: You just multiply by $\frac{5}{9}$. N stays the same. For each of the others, you subtract 32 and then multiply by $\frac{5}{9}$.

Variable	N	Mean	Median	StDev
Highest	50	45.61	45.56	3.72

Variable	Min	Max	Q1	Q3
Highest	37.78	56.67	43.33	47.78

c. Yes, there is an outlier on the high side. The IQR = 4.45, and 1.5(IQR) = 6.675. So, $Q_3 + 1.5 \cdot \text{IQR} = 54.455$, and the maximum is larger than that—so definitely an outlier. $Q_1 - 1.5 \cdot \text{IQR} = 36.655$, and the minimum is bigger—so none on the lower end.

E50. First, subtract 5478 from every number, leaving

0.1 0.3 0.3 0.9 0.4 0.2

The mean of these numbers is 0.3667, so the mean of the original numbers is 5478.3667. The standard deviation is about 0.280. Because re-centering doesn't change the standard deviation, that's the standard deviation for the original set of numbers, too.

E51. Let the mean of the original set of data be

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

Then the mean of the transformed data is the result of the equation shown below.

$$\frac{(x_1+c)+(x_2+c)+(x_3+c)+(x_4+c)+(x_5+c)}{5} = \frac{x_1+x_2+x_3+x_4+x_5+5c}{5}$$

$$= \frac{x_1+x_2+x_3+x_4+x_5}{5} + \frac{5c}{5}$$

$$\bar{x}+c$$

E52. If $\bar{x}$ is the mean of the original set of data, then the standard deviation is shown below.

You know from E51 that the mean of the transformed data is $\bar{x}+c$. Then the standard deviation of the transformed data is also shown below.

Standard Deviation of Original Data:

$$s = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+(x_4-\bar{x})^2+(x_5-\bar{x})^2}{4}}$$

Standard Deviation of the Transformed Data:

$$\sqrt{\frac{((x_1+c)-(\bar{x}+c))^2+((x_2+c)-(\bar{x}+c))^2+((x_3+c)-(\bar{x}+c))^2+((x_4+c)-(\bar{x}+c))^2+((x_5+c)-(\bar{x}+c))^2}{4}}$$

$$= \sqrt{\frac{(x_1+c-\bar{x}-c)^2+(x_2+c-\bar{x}-c)^2+(x_3+c-\bar{x}-c)^2+(x_4+c-\bar{x}-c)^2+(x_5+c-\bar{x}-c)^2}{4}}$$

$$= \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+(x_4-\bar{x})^2+(x_5-\bar{x})^2}{4}}$$

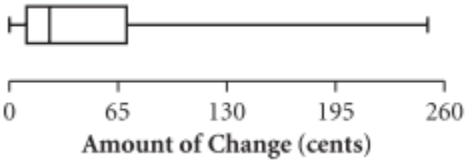
$$= s$$

E53. a. Median: 23 or 24 cents. To estimate this, find the amount of change that corresponds to the 50th percentile.

b. Q_1 is about 10 cents and Q_3 is about 70 cents. To find Q_1 and Q_3 find the amount of change that correspond to the 25th and 75th percentiles, respectively. The IQR is 70 – 10 = 60 cents.

c. The set of data is skewed right. The larger increases toward the left side of the graph indicate larger numbers of students with smaller amounts of change. The large increases correspond to higher bars in a histogram. Or: 75% of the cases are between 0 and 70 cents and the top 25% is spread out from 70 cents to $2.50. So the distribution is skewed toward the larger values.

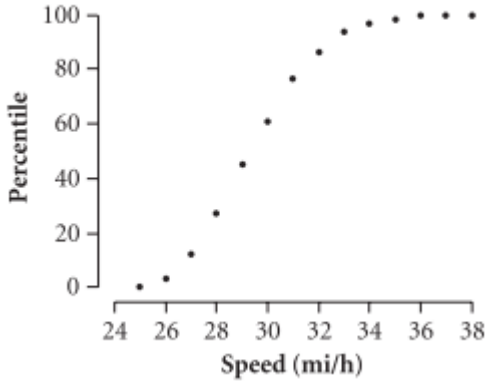
E54. The minimum amount of change is 0 and the maximum amount of change is about $2.50. From E53, Q_1 is about 10 cents, the median is about 23 or 24 cents, and Q_3 is about 70 cents. Since we do not have the individual values we cannot display the outliers, but we do know that there are outliers in the upper values. (Anything above $70+1.5\cdot 60=160$ is an outlier.)



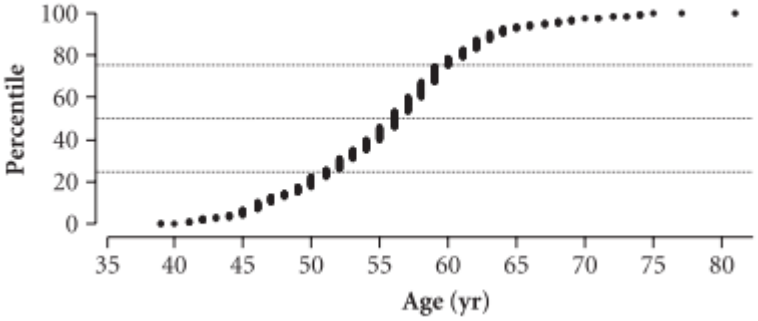
E55. Approximating from the histogram or from the cumulative relative frequencies (or percentiles), the 85% percentile is close to 32 mi/h. Rounding down to the nearest 5 mi/h., the speed limit should be set at 30 mi/h. (See the solution to E56 for the percentile plot.)

E56. Below are the table and plot of the cumulative frequencies and percentiles for the speed data.

Speed	Frequency	Cumulative Frequency	Percentile
25	2	2	0.2
26	31	33	3.3
27	92	125	12.5
28	149	274	27.4
29	178	452	45.2
30	156	608	60.8
31	157	765	76.5
32	99	864	86.4
33	74	938	93.8
34	31	969	96.9
35	16	985	98.5
36	13	998	99.8
37	1	999	99.9
38	1	1000	100



E57. The lower quartile is the 25th percentile, the median is the 50th and the upper quartile is the 75th. From the percentile plot with horizontal lines sketched across at these percentile values, you can see that the lower quartile is a little above 50, the median is a little above 55 and the upper quartile is about 60. So, (a) is the correct choice.



Note: The ages in months are obtained by simply multiplying the ages in years by 12. The shapes of the percentile plot will remain the same, so the median and quartiles also get multiplied by 12. So, the median of the ages in months is 56(12) or 672 months. The first and third quartiles are, respectively, 51(12) = 612 months and 60(12) = 720 months.

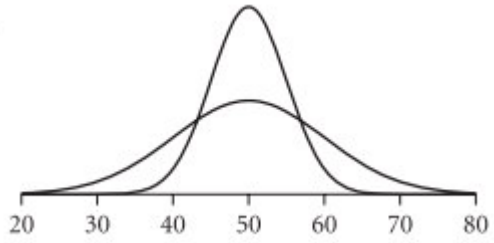
E58. The general idea is the more constant the heights of the histogram are from left to right, the more steadily the cfd (cumulative frequency distribution) rises from left to right. If a histogram is skewed right, the bulk of the data values are smaller, making the cdf rise quickly early on and then taper off toward 1 once the bulk of the data values have past. In such case, you would expect the graph to be concave down (i.e., like an upside down U) because such a function is increasing, but at a slower rate, from left to right. The analysis for a histogram that is skewed left is similar, with the exception that we would expect the cdf to be concave up initially since it is increasing at an increasing rate until the bulk of values have been accounted for, while prior to that point, you would expect the graph to be relatively constant, close to the x-axis.

Based on this analysis, we have the following identifications:

A. IV B. III C. I D. II

E59. **a.** 0.8413 or 84.13%; 0.9943 or 99.43%
b. 0.1587 or 15.87%; 0.0057 or 0.57%
c. 0.9332 or 93.32%
d. 0.6827 or 68.27% (0.6826 using Table A)

E60.



E61. **a.** 2 **b.** 1 **c.** 1.5 **d.** 3 **e.** –1 **f.** -2.5

E62. **a.** 30 **b.** 22 **c.** 85 **d.** –9.5

E63. a. i. 0.6340 (calculator: 0.6319) **ii.** 0.0392 (calculator: 0.0395)
iii. 0.3085 (calculator: 0.3101)

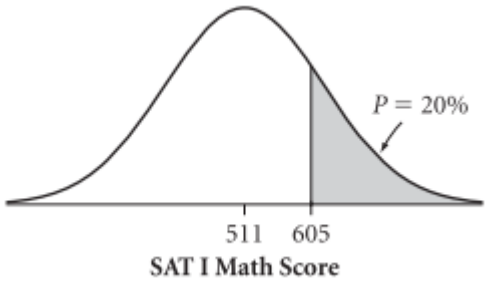
b. The middle 95% of scores range from, approximately,

$$505 - 1.96(111) \approx 287 \text{ to } 505 + 1.96(111) \approx 723.$$

E64. The z-score that has an area of .80 below it is about z = .84.
Unstandardizing,

$$x = mean + z \cdot SD = 511 + .84(112) \approx 605$$

The college should send letters to students who get 605 or more on the exam.



E65. First, find the percentage of men who qualify.

$$z = \frac{x - mean}{SD} = \frac{62 - 70.4}{3.0} = -2.8$$
$$z = \frac{x - mean}{SD} = \frac{72 - 70.4}{3.0} \approx 0.53$$

The area between these z-scores is about 0.4974 + 0.2019 = 0.6993. About 70% of men aged 20 to 29 in the United States meet the height qualifications to be a flight attendant for United Airlines.

Next, find the percentage of women who qualify.

$$z = \frac{x - mean}{SD} = \frac{62 - 65.1}{2.6} \approx -1.19$$
$$z = \frac{x - mean}{SD} = \frac{72 - 65.1}{2.6} \approx 2.65$$

The area between these two z-scores is about 0.3830 + 0.4960 = 0.8790. About 88% of women aged 20 to 29 in the United States meet the height qualifications to be a flight attendant for United Airlines, a higher percentage than men.

E66. a. i. About 0.0021, or 0.21%, are as tall or taller than Shawn Marion. There are only about 39,900 men aged 18 to 24 who are at least as tall.
ii. About 0.2981, or 29.81%, are as tall or taller than Allen Iverson. There are about 5,663,900 men aged 18 to 24 who are at least as tall.
iii. About 0.0000 or 0%, are as tall or taller than Shaquille O'Neal. You would expect to find less than 1 man (or 0.22) this tall in the 18- to 24-year-old age group.

b. The estimates will be too small.

E67. a. 0.1587
b. 8.16
c. Solving $-1.34 = \frac{6-mean}{3}$, you get mean = 10.02.
d. For P = 0.6, z = 0.25. Solving $0.25 = \frac{12-10}{SD}$, you get SD = 8.

E68. a. The z-scores for the quartiles are ±0.67. Thus, $Q_1 = 6.65$ and $Q_3 = 13.35$.
b. The mean must be 150 because it lies midway between the quartiles in a normal distribution. Then SD ≈ 44.78.
c. Solving $-0.67 = \frac{100-mean}{10}$, you get mean = 106.7. Then, because the quartiles are symmetric about the mean, $Q_3 = 113.4$.
d. Because the quartiles are symmetric about the mean, $Q_1 = 9$. Then SD ≈ 1.5.

E69. Make use of the standard rule of thumb for a symmetric normal-like distribution. Doing so, we have the following percentages presented in the order requested in the problem: 68%; 95%; 16%; 84%; 97.5%; 2.5%

E70. $30 \pm 1.96 \cdot 3.6$ or between (roughly) 23 dB and 37 dB.

E71. a. The z-score for the height 68 is $\frac{(68-70.4)}{3.0} = -0.80.$ The area under the normal curve to the left of this point is 0.2119. Thus, about 21.19% of U.S. males between 20 and 29 are less than 68 inches tall.

b. From part a, 0.2119 of the men are below the height of 68 inches. Similarly, the z-score for a height of 67 inches is –1.13, and so 0.1292 of the men are below that height. The proportion in between is 0.2119 – 0.1292 = 0.0827. So, about 0.0827(19,000,000) = 1,571,300 are between the two heights.

c. A percentile of 90 corresponds to a z-score of about 1.28. Hence, the desired height is

$$x = mean + z \cdot SD = 70.4 + 1.28(3.0) = 74.24 \text{ inches}.$$

Equivalently, you could solve the equation:

$$1.28 = \frac{x - 70.1}{3.0}$$

E72. The shape will not change. The mean will be $\frac{70.1}{12} \approx 5.84$ feet, and the SD will be $\frac{2.7}{12} = 0.225$ feet.

E73. a. The distribution is probably skewed right because it's not possible for the length of a reign to be much more than 1 standard deviation below the mean.

b. The z-score for 0 is $\frac{(0-18.5)}{15.4} = -1.20,$ so about 0.1151 of the reigns.

c. If all values in the distribution must be positive and two standard deviations or less below the mean is less than 0, the distribution isn't approximately normal.

E74. a. about 145 points

b. about 25 points

c. From the graph, the middle 95% of the values appear to lie between about 90 and about 200. Using the mean and standard deviation from parts a and b, this interval is about 96 to 195.

d. The z-score for 150 is 0.20, and the area to the right of this point is 0.4207. The z-score for 190 is 1.80, and the area to the right of this point is 0.0359. A weakness here is that next year's teams may not look like a random sample from the set of teams over this 57-year period. Modern teams place more emphasis on scoring than did the teams from an earlier era.

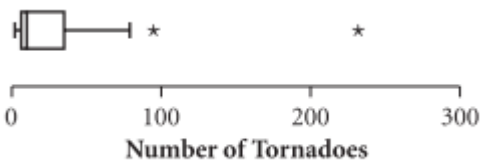
E75. a.

```
Stem-and-leaf of Number N = 51
 Leaf Unit = 1.0

          16    0    0000000111122244
          25    0    566778899
         (5)    1    00013
          21    1    5
          20    2    244
          17    2    66
          15    3    03
          13    3    579
          10    4    2
           9    4    58
           7    5    14
           5    5
           5    6
           5    6
           5    7    2
           4    7    69
 HI 95, 232

                0|5 stands for 5 tornadoes
```

b. Minimum: 0 Q_1: 2 Median: 10 Q_3: 35
Maximum: 232

c. Outliers fall below $2 - 1.5(35 - 2) = -47.5$ or above $35 + 1.5(35 - 2) = 84.5$.
There are two outliers, Florida and Texas.

d.



e. Both plots show the strong skewness in the data and both outliers: Florida and Texas. In the stemplot, you can see that half the states have less than 10 tornadoes. The cleanness of the boxplot makes it clear how much of an outlier Texas actually is. (Of course, it is a very large state, which helps explain why it is an outlier.) However, you can't see from the boxplot at all that so many states have at most 1 tornado. Because the stemplot has a reasonable number of values in it and is consequently easy to read while carrying almost all the complete values, it is reasonable to select it as the most informative.

f. The distribution is strongly skewed right with two outliers and a wall at 0. The median number of tornadoes is 10, with the middle 50% of states having between 2 and 35 tornadoes.

E76. a. The median for special use sites is about 6900 feet. Half of the special use sites have an elevation at or below 6900 feet.

b. An outlier in elevation would likely render that site uninhabitable. As such, such outliers would be in a distribution of habitation sites.

c. The habitation sites have very little variability in elevation, while the seasonal sites have more elevation variability but with the IQR approximately the same as the habitation sites. The special use sites offer even greater variability, but with an IQR even tighter (and included within) the IQR for habitation sites.

E77. a. Outliers would lie below $585 - 1.5(670 - 585) = 457.5$ or above $670 + 1.5(670 - 585) = 797.5$.
b. The bunching up between the lower quartile and the median suggest the distribution is probably skewed right.

E78. Outliers would be any shows with a rating below $4.55 - 1.5(8.25 - 4.55) = -1$, which is impossible, or above $8.25 + 1.5(8.25 - 4.55) = 13.8$. Since the maximum rated show had a rating of 20.7, this show, at least, was an outlier. From the summary of the data we cannot tell if there are other outliers on the high end.

E79. a. Developing countries have lower life expectancies than do developed countries. Thus, Region 1 must be Africa (developing countries, for the most part) and Region 3 must be Europe (developed countries). The Middle East, Region 2, has a mixture of developed and developing countries.

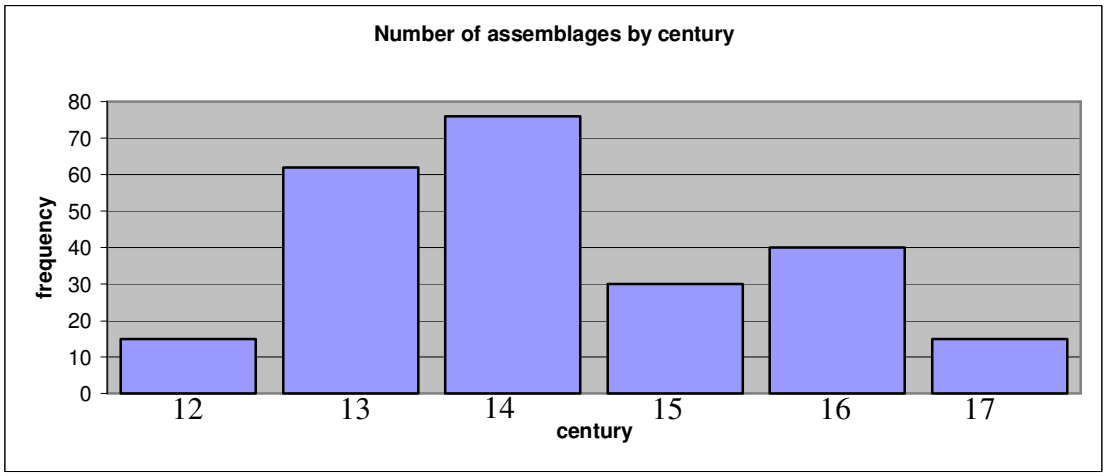
b. C is for Region 1 (Africa); B is for Region 2 (Middle East); A is for Region 3 (Europe).

E80. a. The earliest mean date is 1263, while the latest mean date is 1763.

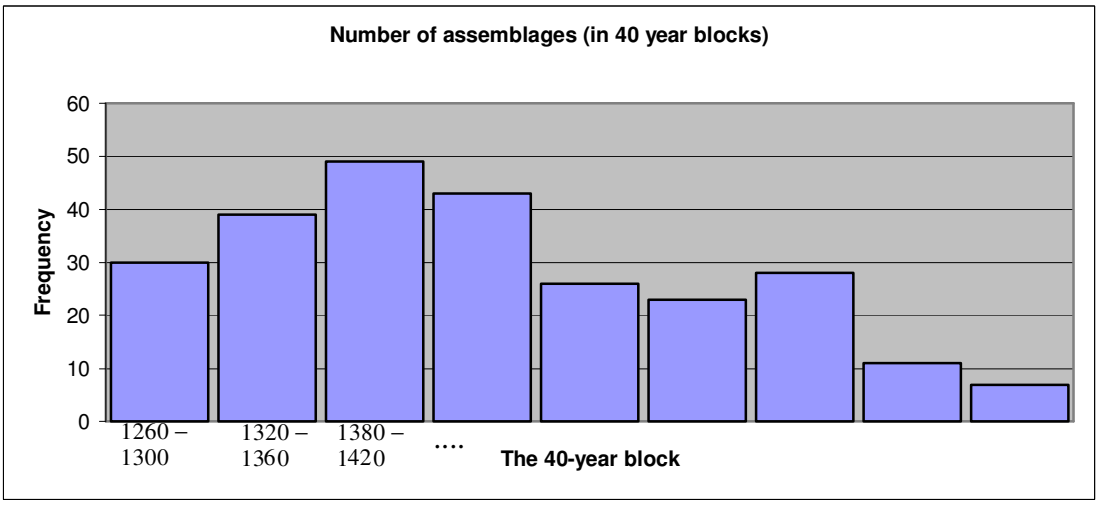
b. Looking at the stem-and-leaf plot on its side (and realizing that the values will now read highest to lowest as you move from left to right) reveals a distribution that is comprised of two mounds. Arguably, the distribution is skewed right since the bulk of the

values occur for earlier years.

c. The suggestion is reasonable based on the distribution provided.



d. Consider the following histogram. Observe that it is still skewed right , so that the impression of the distribution remains the same.



E81. No, this is true only for normal distributions. For example, the set of values {2, 2, 4, 6, 8, 8} is symmetric and has mean and median both equal to 5. The standard deviation is about 2.76. Only two of the values, or about 33% of them, are within one standard deviation of the mean.

E82. I. B II. D III. A IV. C

E83. a. The median number of deaths is 13. Half of the cities have fewer than 13 pedestrian deaths per year, and half have more.

b. The lower quartile is the average of the 10^{th} and 11^{th} data points, namely 8. The upper quartile is the average of the 30^{th} and 31^{st} data points, namely 24.5.

An outlier is a city whose number of deaths exceeds 24.5 + 1.5(24.5-8) = 49.25 or is less than 8 – 1.5 (24.5-8) = –16.75, which is impossible. There are three outliers, Phoenix, Los Angeles, and New York. One explanation is the large populations these cities have; these are the three largest cities in the United States.

c. Plots and explanations will vary. A stemplot is a very good choice. It reveals the three outliers and shows that the distribution is skewed right. It is given by:

0	0134556678888999
1	0002334455679
2	22788
3	06
4	58
5	9
6	
7	
8	
9	9
10	
11	
12	
13	
14	
15	7

d. New York 1.91; Miami 6.68. The rate is adjusted for population size and gives a more accurate picture of pedestrian safety.

E84. a. Guesses may vary. In actual fact, Region 1 is Europe, Region 2 is Asia, Region 3 is South and Central America, and Region 4 is Africa. The outlier for Asia is Bangladesh. The outlier for Africa is Angola (Egypt is the country with 99% in Africa).

b. Distributions 1and 4 are skewed left. Distributions 2 and 3 are symmetrical.

c. The dot plots are in the same order as the boxplots, A being Region 2, B Region 4, C Region 3, and D Region 1.

d. Region 4 does not look skewed in the dot plot even though it appears so in the boxplot. The number of countries plotted is small, and the values vary a lot, so the locations of the quartiles might change quite a bit with small changes in the data. Dot plots give the better picture here.

E85. The mean grade is 2.98 with a standard deviation of 1.33.

E86. There are two 75s, one 74, etc. The mean is

$$\frac{71\cdot1+72\cdot1+73\cdot1+74\cdot1+75\cdot2+76\cdot4+77\cdot4+78\cdot2+79\cdot2+80\cdot3+81\cdot5+82\cdot6+83\cdot7+84\cdot2+85\cdot2}{43}$$

$= 79.581$ years

The standard deviation is

$$\sqrt{\frac{(71-79.581)^2 \cdot 1+(72-79.581)^2 \cdot 1+\cdots+(85-79.581)^2 \cdot 2}{42}} = 3.64 \text{ years}.$$

E87. There are many possible responses. An example is {1, 1, 1, 1, 1, 2, 2, 10}, which has mean 2.375 and standard deviation of about 3.11. One standard deviation below the mean is less than 0.

E88. The standard deviation is 0 because the numbers don't vary.

E89. a. i. 6.325 v. 6.667
ii. 2.000 v. 2.010
iii. 0.632 v. 0.633
b. No, as *n* gets larger, the difference between s and σ goes to 0.

E90. *Important warning in this exercise*: Percentiles are not score/total but a measure of position relative to the number of scores.

a. The mean of the scores on Test I is 15.5, and the standard deviation is 3.028. The score of 19 has a z-score of 1.16 and is at the 90th percentile, relative to the other 9 scores.

b. The mean of the scores on Test II is 6.4, and the standard deviation is 8.708. The score of 18 has a z-score of 1.33 but is only at the 80th percentile, relative to the other 9 scores.

c. Answers will vary, and rightfully so. The student who got a 19 on Test I did better than all but one other student in the class. However, the student who got an 18 on Test II did much better than all but two students in the class.

E91. a. The state with the lowest per capita income in 1980 had an average income per person of $6,573.
b. There is at least one outlier on the high end for both 2000 and 2007, but none for 1980. There are no outliers on the low end for any of the years.
c. There was a noticeable improvement in position as the *z*-scores went from -1.0 in 1980 to -0.86 in 2007.
d. As the years progress, the histograms become more skewed toward the higher values and both the centers and spreads increase.
e. The *z*-score would work best for the 1980 data, as it is more mound-shaped and symmetric then the others.

E92. a. Using the plot, a score of 425 appears to be at about the 20th percentile. Assuming a normal distribution, the z-score for 425 is –0.83, giving a percentile of 20.33. These are close.

b. From the display, the 40th percentile appears to be about 510. The 40th percentile under a standard normal curve has a z-score of –0.25, which translates to a test score of

$$x = mean + z \cdot SD \approx 525.2 - 0.25(120.2) \approx 495$$

Again, these are quite close, which gives some evidence in support of the normality of the scores.

c. Answers will vary somewhat. From the plot, the median appears to be about 540. In a normal distribution, the median should be close to the mean, or 525.2. The mean and median here are the same.

d. The quartiles are about 450 and about 610, giving an IQR of about 610 – 450 = 160. From the standard normal curve, the quartiles have z-scores of approximately –0.67 and +0.67 and the median has a z-score of 0. Thus, the approximate quartiles for the exam scores with mean 525.2 and standard deviation 120.2 are

$$Q_1 = 525.2 - 0.67(120.2) \approx 544.5$$
$$Q_3 = 525.2 + 0.67(120.2) \approx 606$$

giving an IQR of 161. The normal model is looking good.

E93. $176 \pm (1.645)30$ or 126.65 mg/dl and 225.35 mg/dl.

E94. a. Except for a slight bulge around 0.330, the batting averages look quite normal in their distribution.

b. The mean is about 0.270, and the standard deviation is about 0.030.

c. The z-score for 0.300 is about 1. About 16% of the players would have batted over 0.300.

d. The histogram shows about 31 of the 187, or about 16.6% players batted over 0.300. The estimate was quite close.

E95. a. Again, the histogram of batting averages looks quite normal in shape with center at about 0.260. The standard deviation is approximately 0.040.

b. The batting averages in both leagues have distributions that are approximately normal in shape. The American League has a higher mean (by about 0.010) and less spread.

c. The z-score corresponding to 0.300 in the National League is $z = 1$. The corresponding batting average, x, in the American League would still be 1 standard deviation above the mean, or

$$x = 0.270 + (1)(0.030) = 0.300$$

so that batter would be expected to have a similar batting average in the American League.

Concept Review Solutions

C1. B
C2. C
C3. A
C4. D
C5. D
C6. B
C7. E Without knowing whether the distribution of scores is close to normal, you cannot make an accurate assessment of this probability.
C8. C
C9. a. 2,156 **b.** Bacon himself. **c.** For Bacon and Smith, n = 982,586. The mean Bacon number is 2.95 and the mean Smith number is 2.85. So, Smith is the better center. **d.** The Bacon-Smith number and Smith-Bacon number must be equal.