Chapter 5
Exploring Data: Distributions

Chapter Outline

Chapter Summary

The description of data is an important link between its collection and its interpretation. Exploratory data analysis combines numerical summary with graphical display in an attempt to discern patterns in data. *Data* are often values of a numeric variable. The pattern of these values is called a distribution. In general, one expects chance deviation from a smooth distribution. Important deviations are called outliers.

In analyzing data, it is wise to proceed from the simple to the complex, to analyze variables one at a time and then to seek relations between variables. Visual displays are useful tools in data analysis because of the eye's uncanny ability to see pattern as well as deviation from pattern. *Histograms* and *stemplots* can be extremely suggestive. *Boxplots* provide both a visual and numeric summary of data and can be useful in comparing distributions for several variables.

Numerical summary of data should give a measure of *center* and of spread for the data. Typically used measures of center are the *mean* (arithmetic average) and *median* (midpoint of the data set). Measures of spread or *variability* include the range of the data set (difference between the high and low values), quartiles (the values that have 25% and 75% of the data lying below them), standard deviation, and *variance*. The *five-number summary* of a set of data provides the median, the upper and lower quartiles, and the high and low values for the data set. A boxplot is a visual display of the five-number summary.

Another and very important visual display is given by normal distributions. All *normal curves* are symmetric and bell-shaped. These are the main characteristics of a normal distribution. A particular normal curve is completely specified by its mean μ (center point) and standard deviation *s* (spread). Any normal distribution satisfies the *68–95–99.7 rule*. Also, the first and third quartiles of any normal distribution can be easily calculated given its mean and standard deviation.

Skill Objectives

1. Construct a histogram for a small data set.
2. List and describe two types of distribution for a histogram.
3. Identify from a histogram possible outliers of a data set.
4. Construct a stemplot for a small data set.
5. Calculate the mean of a set of data.
6. Sort a set of data from smallest to largest and then determine its median.
7. Determine the upper and lower quartiles for a data set.
8. Calculate the five-number summary for a data set.
9. Construct the diagram of a boxplot from the data set's five-number summary.
10. Calculate the standard deviation of a small data set.
11. Describe a normal curve.
12. Given the mean and standard deviation of a normally distributed data set, compute the first and third quartiles.
13. Explain the 68–95–99.7 rule.
14. Sketch the graph of a normal curve given its mean and standard deviation.
15. Given the mean and standard deviation of a normally distributed data set, compute the intervals in which the data set fall into a given percentage by applying the 68–95–99.7 rule.

Teaching Tips

1. You may wish to give handouts that include a few different data sets when the chapter is introduced for students to become familiar with. These data sets can be used for exploring the different types of statistical analysis in the chapter. By having these data sets already prepared, you can forgo having to write them on the board and having students copy them during the lecture.
2. Students can easily get the terms skewed right and skewed left confused. Emphasize that the location of the "tail" indicates the direction.
3. When constructing a histogram, students often need help in determining appropriate classes for data of a given range. When listing intervals for a histogram, students are often confused about how to handle data points that occur at the endpoints of the interval. It may be helpful to explain that these are normally done in an arbitrary fashion. If, for example, an interval goes from 5.10 to 5.20, one approach is to end that interval at 5.19 (assuming measurements are done to the nearest hundredth) and start the next at 5.20. An equally desirable approach would be to end the first interval at 5.20 and then begin the next at 5.21. Without attempting to apply sophisticated statistical guidelines, you may want to examine different data sets and discuss appropriate classes.

4. Be clear to your students as to your expectations regarding the use of calculators/spreadsheets in their work. If technology is emphasized, let students know what is required regarding intermediate calculations.

5. A helpful approach to determining both the median and lower and upper quartiles is to work out examples when the number of data points is even and a multiple of 4, when it is even but not a multiple of 4, and when it is odd. Subdividing the set of an odd number of data points results in eliminating the median from the set when analyzing the lower and upper halves, whereas subdividing for an even number places each one of the two middle points in a different half.

6. Using the formula $\frac{n+1}{2}$ to determine the location of the median within a sorted data set sometimes creates confusion, because students don't understand the difference between a cardinal and an ordinal number. Reporting the answer to this calculation as the "*n*th" location within the set helps the student understand this distinction.

7. Although sorting the data points in a set can be done either from lowest to highest or from highest to lowest, the natural progression for the sort, when written in a vertical fashion, is to go from low to high. Then the concepts of lower and upper quartiles flow naturally. In a hierarchical order, "lower" comes before "higher."

8. An example that emphasizes to students that mean and median are measures of describing center and that the mean is sensitive to outliers is the following.

A town advertises that its average salary is $100,000. Would you want to move there?

As it turns out, there are 11 people in that town. One makes $1,000,000 and the rest make $10,000.

You may also choose to briefly mention *mode* at this point to simply emphasize that the term "average" depends on the form of measurement.

9. Exercise 12 has an outcome of the gap between the median and the mean being cut about in half when removing the high outlier. This data set can also be used in discussing standard deviation. In this data set, the standard deviation drops from 13.2 to 9.9.

10. When introducing the concept of the normal curve, you may wish to concentrate first on the base line because of its similarity to the number line. The mean plays the role of the origin, and the standard deviation markings, the units. Inserting a specific mean and standard deviation then becomes an exercise in relabeling the mean and calculating the units, which are three standard deviations from the mean. Some students will need a lot of practice in labeling normal curves with given means and standard deviations.

Research Paper

Exercise 18 deals with the popular vote by the successful candidate in the presidential elections from 1948 to 2004. Students can write a paper using this data to determine the winning candidates in each election. Students can indicate which parties were involved and determine if any other statistical data may be of interest (such as if a candidate was incumbent). You may also ask students to further research presidential statistical data such as age when taking office, age when leaving office, or number of days in office.

Collaborative Learning

Data Analysis

Suppose you are the president of CBS, and you are preparing to submit a bid on televising the World Series. You have to decide on how much to bid, depending on how much advertising revenue you expect to receive. However, there is a good deal of uncertainty in this matter, since the length of the World Series is not known in advance. The series is played on a best 4-out-of-7 basis, so that the series can last anywhere from 4 to 7 games. To assist you in your calculations, you decide to look to the past. The following table tells you the length of the series in each of the years from 1905 on. (1903, 1919, 1920, and 1921 are left out because those series were played on a best 5-out-of-9 basis. Also, the series was not played in 1904 and 1994.)

Using the data, do the following.

1. Construct a frequency chart of the number of games in a World Series.

2. Draw a histogram of the number of games in a World Series.

3. Find the mean number of games in a World Series.

4. Find the standard deviation of the number of games in a World Series.

1905 5	1924 7	1940 7	1956 7	1972 7	1988 5
1906 6	1925 7	1941 5	1957 7	1973 7	1989 4
1907 4	1926 7	1942 5	1958 7	1974 5	1990 4
1908 5	1927 4	1943 5	1959 6	1975 7	1991 7
1909 7	1928 4	1944 6	1960 7	1976 4	1992 6
1910 5	1929 5	1945 7	1961 5	1977 6	1993 6
1911 6	1930 6	1946 7	1962 7	1978 6	1995 6
1912 7	1931 7	1947 7	1963 4	1979 7	1996 6
1913 5	1932 4	1948 6	1964 7	1980 6	1997 7
1914 4	1933 5	1949 5	1965 7	1981 6	1998 4
1915 5	1934 7	1950 4	1966 4	1982 7	1999 4
1916 5	1935 6	1951 6	1967 7	1983 5	2000 5
1917 6	1936 6	1952 7	1968 7	1984 5	2001 7
1918 6	1937 5	1953 6	1969 5	1985 7	2002 7
1922 4	1938 4	1954 4	1970 5	1986 7	2003 6
1923 6	1939 4	1955 7	1971 7	1987 7	2004 4

Solutions

Skills Check:

1. a	2. a	3. b	4. b	5. c	6. a	7. b	8. c	9. b	10. b
11. c	12. c	13. b	14. a	15. b	16. a	17. a	18. a	19. b	20. c

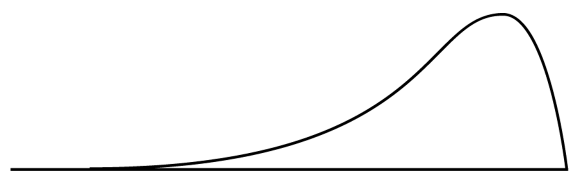
Exercises:

1. (a) The individuals in the data set are the make and model of 2004 motor vehicles.

(b) The variables are vehicle type, transmission type, number of cylinders, city MPG, and highway MPG. Histograms would be helpful for cylinders (maybe), and the two MPGs (certainly).

2. The distribution is skewed to the right. Shakespeare uses many words of two, three, and four letters and a few long words (10, 11, and 12 letters). Most English prose will show a similarly shaped distribution of word lengths because so many common words (I, you, him, her, and, or, not,…) are short.

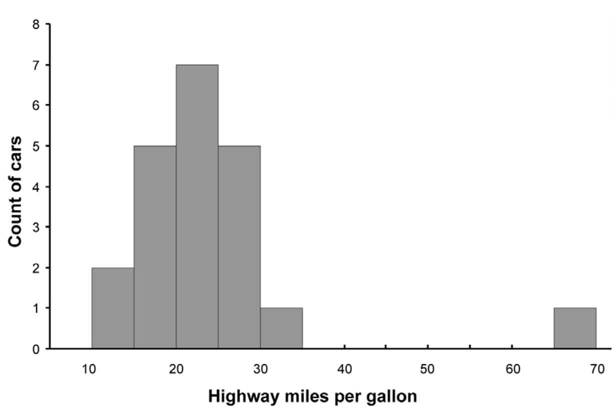
3. Draw a histogram with a peak at the right and lower bars trailing out to the left of the peak.



Most coins were minted in recent years, producing a peak at the right (highest-numbered years, like 2004 and 2005). There are few coins from 1990 and even fewer from 1980.

4. Here is a histogram that includes the outlier. The histogram shows the shape (roughly symmetric, plus the outlier) better than the stemplot.

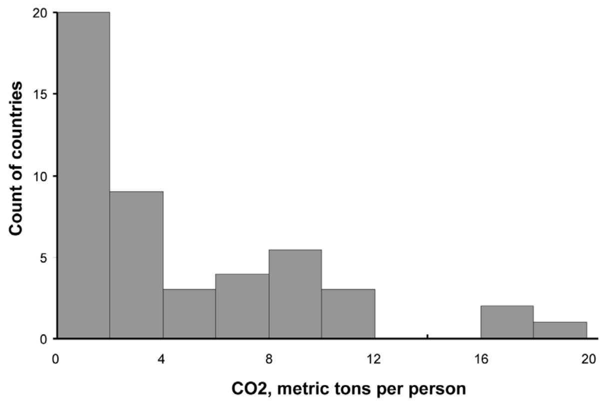
Class	Count
6 – 10	0
11 – 15	2
16 – 20	5
21 – 25	7
26 – 30	5
31 – 35	1
36 – 40	0
41 – 45	0
46 – 50	0
51 – 55	0
56 – 60	0
61 – 65	0
66 – 70	1



5. (a) Otherwise, big countries would top the list even if they had low emissions for their size.

(b) Using class widths of 2 metric tons per person, we have the following.

Class	Count
0.0 – 1.9	20
2.0 – 3.9	9
4.0 – 5.9	3
6.0 – 7.9	4
8.0 – 9.9	6
10.0 – 11.9	3
12.0 – 13.9	0
14.0 – 15.9	0
16.0 – 17.9	2
18.0 – 19.9	1



The distribution is skewed to the right. There appear to be three high outliers: Canada, Australia, and the United States.

6. From top left, (a) is study time: right-skewed; (b) is handedness: most people are right-handed; (c) is gender: majority female but more balanced than handedness; (d) is height, single-peaked and roughly symmetric.

7. (a) Alaska is 5.7% and Florida is 17.6%.

(b) The distribution is single-peaked and roughly symmetric. The center is near 12.7% (12.7% and 12.8% are the 24^{th} and 25^{th} in order out of 48, ignoring Alaska and Florida). The spread is from 8.5% to 15.6%.

8. Rounding the data to the nearest 10, we have the following.

140	160	110	150	130	100	100	**0**80	150
170	200	270	100	170	360	150	150	260

Truncating the zero as described, we have the following.

14	16	11	15	13	10	10	08	15
17	20	27	10	17	36	15	15	26

There is one high outlier (359) which we omit from the stemplot. The distribution is single-peaked and slightly right-skewed. Control of glucose is poor: all but five are above 130. Here is the stemplot:

0	8
1	000134
1	5555677
2	0
2	67

9. Here is the stemplot.

10	139
11	5
12	669
13	77
14	08
15	244
16	55
17	8
18	
19	
20	0

There is one high outlier, 200. The center of the 17 observations other than the outlier is 137 (9th of 17). The spread is 101 to 178.

10. Here is the stemplot.

48	8
49	
50	7
51	0
52	6799
53	04469
54	2467
55	03578
56	12358
57	59
58	5

The distribution is roughly symmetric and single-peaked except for one low observation, which may be an outlier.

11. (a) $\bar{x} = \frac{154+109+137+115+152+140+154+178+101+103+126+126+137+165+165+129+200+148}{18} = \frac{2539}{18} \approx 141.1.$

(b) Without the outlier, $\bar{x} = \frac{154+109+137+115+152+140+154+178+101+103+126+126+137+165+165+129+148}{17} = \frac{2339}{17} \approx 137.6.$ The high outlier pulls the mean up.

12. (a) Here is the stemplot, with the outlier.

1	69
2	455
3	334477
4	025669
5	
6	
7	3

(b) For all 18 years, $\bar{x} = \frac{16+25+24+19+33+25+34+46+37+33+42+40+37+34+49+73+46+45}{18} = \frac{658}{18} \approx 36.56.$ To determine the median, we must put the data in order from smallest to largest.

16 19 24 25 25 33 33 34 **34** **37** 37 40 42 45 46 46 49 73

Since there are 18 pieces of data, the median is the mean of the 9th and 10th pieces of data.

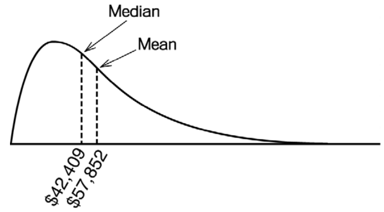
$$M = \frac{34+37}{2} = \frac{71}{2} = 35.5.$$

For the 17 years other than 2001, $\bar{x} = \frac{16+25+24+19+33+25+34+46+37+33+42+40+37+34+49+46+45}{17} = \frac{585}{17} \approx 34.41.$ Taking out 73 as a piece of data, we have now 17 pieces of data.

16 19 24 25 25 33 33 34 **34** 37 37 40 42 45 46 46 49

The median is the $\frac{17+1}{2} = \frac{18}{2} = 9^{th}$ piece of data, namely $M = 34.$ Removing the high outlier cuts the gap between the median and the mean about in half.

13. The distribution of incomes is strongly right-skewed, so the mean is much higher than the median. Thus, \$57,852 is the mean.



14. The distribution of household assets is very right-skewed. Most young households have few assets, but a few wealthy households have large assets. The strong skewness pulls the mean up.

15. Examples will vary.
One high outlier will do it. For example, 1, 2, 3, 3, 4, 17. These data have third quartile 4 and mean, $\bar{x} = \frac{1+2+3+3+4+17}{6} = \frac{30}{6} = 5$.

16. Examples will vary.
We want 10 to be the 3rd (in increasing order) of the 5 observations. To get mean 7, the sum must be 35. So, for example, 1, 1, 10, 10, 13 will do.

17. The five-number summary is Minimum, Q_1, M, Q_3, Maximum.
The minimum is 5.7 and maximum is 17.6. The median is the mean of the 25th and 26th pieces of data, namely $\frac{12.7+12.8}{2} = \frac{25.5}{2} = 12.75$. There are 25 pieces of data below the median, thus Q_1 is the 13th piece of data, 11.7. There are 25 pieces of data above the median, thus Q_3 is the 38th piece of data, namely 13.5. Thus, the five-number summary for these 50 observations is 5.7, 11.7, 12.75, 13.5, 17.6.

18. (a) Round to full percents.

Year	1948	1952	1956	1960	1964	1968	1972	1976
Percent	50	55	75	50	61	43	61	50
Year	1980	1984	1988	1992	1996	2000	2004	
Percent	51	59	54	43	49	48	51	

All percents are on stems 4, 5, and 6. The stemplot is

```
4 | 3389
5 | 000114579
6 | 11
```

b) To determine the median, we must put the data in order from smallest to largest.

43.2 43.4 47.9 49.2 49.6 49.7 50.1 **50.7** 51.2 53.9 55.1 57.4 58.8 60.7 61.1

Since there are 15 pieces of data, the 8th piece of data is the median, namely 50.7%.

c) Since there are 7 pieces of data above the median, the third quartile is the 12th piece of data, namely 57.4%. So, the 1956, 1964, 1972, and 1984 elections were landslides.

19. To determine the minimum, maximum, and median, we must put the 21 pieces of data in order from smallest to largest.

13 15 16 16 17 19 20 22 23 23 **23** 24 25 25 26 28 28 28 29 32 66

The minimum is 13 and the maximum is 66. The median is the $\frac{21+1}{2} = \frac{22}{2} = 11^{th}$ piece of data, namely 23. Since there are 10 observations to the left of the median, Q_1 is the mean of the 5th and 6th pieces of data, namely $\frac{17+19}{2} = \frac{36}{2} = 18$. Since there are 10 observations to the right of the median, Q_3 is the mean of the 16th and 17th pieces of data, namely $\frac{28+28}{2} = \frac{56}{2} = 28$.

Thus, the five-number summary is 13, 18, 23, 28, 66.

20. To determine the minimum, maximum, and median, we must put the 21 pieces of data in order from smallest to largest.

4.88 5.07 5.10 5.26 5.27 5.29 5.29 5.30 5.34 5.34 5.36 5.39 5.42 5.44 5.46
5.47 5.50 5.53 5.55 5.57 5.58 5.61 5.62 5.63 5.65 5.68 5.75 5.79 5.85

The minimum is 4.88 and the maximum is 5.85. The median is the $\frac{29+1}{2}=\frac{30}{2}=15^{\text{th}}$ piece of data, namely 5.46. Since there are 14 observations to the left of the median, Q_1 is the mean of the 7$^{\text{th}}$ and 8$^{\text{th}}$ pieces of data, namely $\frac{5.29+5.30}{2}=\frac{10.59}{2}=5.295$. Since there are 14 observations to the right of the median, Q_3 is the mean of the 22$^{\text{th}}$ and 23$^{\text{th}}$ pieces of data, namely $\frac{5.61+5.62}{2}=\frac{11.23}{2}=5.615$.

Thus, the five-number summary is 4.88, 5.295, 5.46, 5.615, 5.85. The quartiles are roughly equidistant from the median (symmetry). The minimum (a possible outlier) is farther from the median than is the maximum.

21. To determine the minimum, maximum, and median, we should put the 48 pieces of data in order from smallest to largest. It may be easier, however, to create a stemplot.

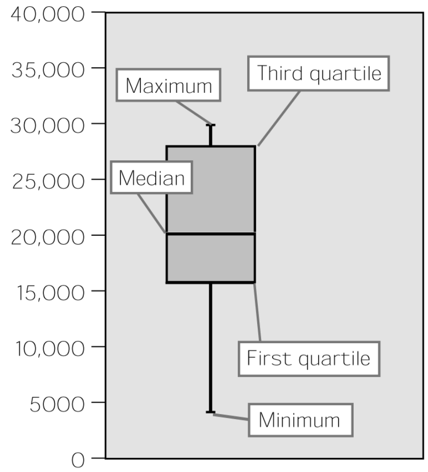
0	001122223357899
1	02478
2	3558
3	67899
4	68
5	1
6	18
7	36
8	018
9	017
10	02
11	0
12	
13	
14	
15	
16	0
17	0
18	
19	9

The minimum is 0.0 and the maximum is 19.9. Since there are 48 pieces of data, the median is the mean of the 24$^{\text{th}}$ and 25$^{\text{th}}$ pieces of data, namely $\frac{2.8+3.6}{2}=\frac{6.4}{2}=3.2$. Since there are 24 observations to the left of the median, Q_1 is the mean of the 12$^{\text{th}}$ and 13$^{\text{th}}$ piece, of data, namely $\frac{0.7+0.8}{2}=\frac{1.5}{2}=0.75$. Since there are 24 observations to the right of the median, Q_3 is the mean of the 36$^{\text{th}}$ and 37$^{\text{th}}$ pieces of data, namely $\frac{7.6+8.0}{2}=\frac{15.6}{2}=7.8$.

Thus, the five-number summary is 0.0, 0.75, 3.2, 7.8, 19.9. The third quartile and maximum are much farther from the median that the first quartile and minimum, showing that the right side of the distribution is more spread out than the left side.

22. The minimum is 4123 and the maximum is 29,875. Since there are 56 pieces of data, the median is the mean of the 28th and 29th pieces of data, namely $\frac{19{,}910+20{,}234}{2}=\frac{40{,}144}{2}=20{,}072.$ Since there are 28 observations to the left of the median, Q_1 is the mean of the 14th and 15th pieces of data, namely $\frac{15{,}500+15{,}934}{2}=\frac{31{,}434}{2}=15{,}717.$ Since there are 28 observations to the right of the median, Q_3 is the mean of the 42nd and 43rd pieces of data, namely $\frac{27{,}904+28{,}011}{2}=\frac{55{,}915}{2}=27{,}957.5.$

Thus, the five-number summary is 4123, 15717, 20072, 27957.5, 29875. The boxplot displays these five numbers.



It cannot, however, show the clusters in the data.

23. The income distribution for bachelor's degree holders is generally higher than for high school graduates: the median for bachelor's is greater than the third quartile for high school. The bachelor's distribution is very much more spread out, especially at the high-income end but also between the quartiles.

24. (a) The median is at the $\frac{n+1}{2}=\frac{14{,}959+1}{2}=\frac{14{,}960}{2}=7480^{\text{th}}$ position in the list. The value of the median is (from the data file) \$46,000.

(b) Q_1 is the median of the 7479 observations to the left of the median, position 3740. Similarly, Q_3 has position $7480+3740=11{,}220$. Although student answers may vary slightly, the values of the quartiles are $Q_1=\$31{,}000$ and $Q_3=\$65{,}000$.

25. (a) Placing the data in order from smallest to largest, we have the following.

2.0	2.5	3.0	7.1	10.1	10.3	12.0	12.1
12.9	14.7	14.8	17.6	18.0	18.5	20.1	21.3
21.7	24.9	26.9	28.3	29.1	30.5	31.4	32.5
32.9	33.7	34.6	34.6	35.1	36.6	37.0	37.7
37.9	38.6	42.7	43.4	44.5	44.9	46.4	47.6
49.4	50.4	51.9	53.2	54.2	56.4	57.4	58.8
61.4	63.1	64.9	65.6	69.5	69.8	79.5	81.1
82.2	92.2	97.7	103.1	118.2	156.5	196.0	204.9

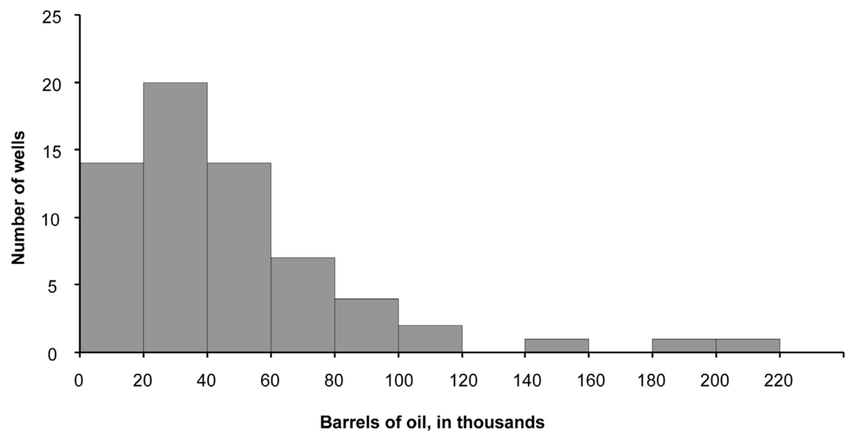
Continued on next page

25. (a) continued

Either a histogram or a stemplot will do. By rounding to the nearest whole number and using the ones digit as leaves, we have the following stemplot, with three high outliers (156.5, 196.0, 204.9) omitted.

0	2337
1	0022355889
2	0125789
3	1133455577889
4	3355689
5	0234679
6	1356
7	00
8	012
9	28
10	3
11	8

Here is the histogram using class widths of 20 (thousand) barrels of oil.



The distribution is right-skewed with high outliers.

(b) $\bar{x} = \frac{3087.9}{64} \approx 48.25$ and since there are 64 pieces of data, the median is the mean of the 32nd and 33rd pieces of data, namely $M = \frac{37.7+37.9}{2} = \frac{75.6}{2} = 37.8$. The long right tail pulls the mean up.

(c) The minimum is 2.0 and the maximum is 204.9. As found in part b, the median is 37.8. Since there are 32 observations to the left of the median, Q_1 is the mean of the 16th and 17th piece of data, namely $\frac{21.3+21.7}{2} = \frac{43}{2} = 21.5$. Since there are 32 observations to the right of the median, Q_3 is the mean of the 48th and 49th piece of data, namely $\frac{58.8+61.4}{2} = \frac{120.2}{2} = 60.1$.

Thus, the five-number summary is 2.0, 21.5, 37.8, 60.1, 204.9. The third quartile and maximum are much farther above the median that the first quartile and minimum are below it, showing that the right side of the distribution is much more spread out than the left side.

26. Each bar represents a single word length. The bar heights appear to be roughly 5%, 17%, 22%, 24%, 12%, 7%, 6%, 3%, and four shorter bars at the right. Q_1 is the 25% point, which is in the 3rd bar, so $Q_1 = 3$. M is the 50% point, so $M = 4$, and similarly $Q_3 = 5$. The five-number summary is 1, 3, 4, 5, 12.

27. For the data in Table 5.1, $M = 4.7\%$, $Q_1 = 2.1\%$, and $Q_3 = 8.7\%$. So $IQR = 8.7 - 2.1 = 6.6$ and $1.5IQR = 1.5(6.6) = 9.9$. Values less than $2.1 - 9.9 = -7.8$ or greater than $8.7 + 9.9 = 18.6$ are suspected outliers. By this criterion, there are 5 high outliers: Arizona, California, Nevada, New Mexico, and Texas.

28. From Exercise 21, the five-number summary is 0.0, 0.75, 3.2, 7.8, 19.9. So $IQR = 7.8 - 0.75 = 7.05$ and $1.5IQR = 1.5(7.05) = 10.575$. Observations above $7.8 + 10.575 = 18.375$ are suspected outliers. Only the United States meets this criterion. The histogram in Exercise 5 suggests that there are three clear outliers, not just one.

29. (a) Placing the data in order (not required, but helpful), we have the following hand calculations.

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $(x_i - \bar{x})^2$
	4.88		– 0.57		0.3226
	5.07		– 0.38		0.1429
	5.10		– 0.35		0.1211
	5.26		– 0.19		0.0353
	5.27		– 0.18		0.0317
	5.29		– 0.16		0.0250
	5.29		– 0.16		0.0250
	5.30		– 0.15		0.0219
	5.34		– 0.11		0.0117
	5.34		– 0.11		0.0117
	5.36		– 0.09		0.0077
	5.39		– 0.06		0.0034
	5.42		– 0.03		0.0008
	5.44		– 0.01		0.0001
	5.46		0.01		0.0001
	5.47		0.02		0.0005
	5.50		0.05		0.0027
	5.53		0.08		0.0067
	5.55		0.10		0.0104
	5.57		0.12		0.0149
	5.58		0.13		0.0174
	5.61		0.16		0.0262
	5.62		0.17		0.0296
	5.63		0.18		0.0331
	5.65		0.20		0.0408
	5.68		0.23		0.0538
	5.75		0.30		0.0912
	5.79		0.34		0.1170
	5.85		0.40		0.1616
sum =	**157.99**	**sum =**	**0.00**	**sum =**	**1.3669**

$\bar{x} = \frac{157.99}{29} \approx 5.448$ and $s^2 = \frac{1.3669}{29\text{-}1} = \frac{1.3669}{28} \approx 0.0488$ which implies $s \approx \sqrt{0.0488} \approx 0.221$.

(b) The median is the $\frac{29+1}{2} = \frac{30}{2} = 15^{\text{th}}$ piece of data, namely 5.46 (From Exercise 10). The mean and median are close, but the one low observation pulls $\bar{x}$ slightly below M.

30. (a) $\bar{x} = \frac{5.6+5.2+4.6+4.9+5.7+6.4}{6} = \frac{32.4}{6} = 5.4.$

(b) Placing the data in order (not required, but helpful), we have the following hand calculations.

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $\left(x_i - \bar{x}\right)^2$
	4.6		– 0.8		0.64
	4.9		– 0.5		0.25
	5.2		– 0.2		0.04
	5.6		0.2		0.04
	5.7		0.3		0.09
	6.4		1.0		1.00
sum =	**157.99**	**sum =**	**0.00**	**sum =**	**2.06**

Thus, $s^2 = \frac{2.06}{6-1} = \frac{2.06}{5} = 0.412$ and $s = \sqrt{0.412} \approx 0.642.$

31. Since the standard deviation, s, is 15 we have the variance $s^2 = 15^2 = 225$.

32. (a), (b), (d) are in the units of the original observations, years for (a) and (d) and seconds for (b). The correlation (c) has no units.

33. For both data, $\bar{x} = 7.50$ and $s = 2.03$ (to two decimal places). Data A has two low outliers:

3	1
4	7
5	
6	1
7	3
8	1178
9	113

and Data B has one high outlier:

5	368
6	69
7	079
8	58
9	
10	
11	
12	5

Additional comments may vary.

34. (a) Strong skewness and high outliers, so prefer the five-number summary.

(b) Symmetric (and close to normal, as we will see), so use $\bar{x}$ and s.

(c) Irregular with clusters, so no numerical summary does well.

35. With most non-graphing calculators, this goes quickly. Some calculators, e.g., the Sharp EL-509S, already give a wrong answer for three central zeros. The TI-30Xa handles four zeros but is wrong for five central zeros. In both cases, the calculators report $s = 0$, so an alert user knows the result is wrong. With a spreadsheet program such as Excel, we have the following results.

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $(x_i - \bar{x})^2$		
	1001		– 1		1	$s^2 = \frac{2}{2}$	1
	1002		0		0	$s = \sqrt{s^2}$	1
	1003		1		1		
sum =	**3006**	**sum =**	**0**	**sum =**	**2**		
$\frac{3006}{3} =$	1002						

Altering the data as described be have the following.

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $(x_i - \bar{x})^2$		
	10001		– 1		1	$s^2 = \frac{2}{2}$	1
	10002		0		0	$s = \sqrt{s^2}$	1
	10003		1		1		
sum =	**30006**	**sum =**	**0**	**sum =**	**2**		
$\frac{30006}{3} =$	10002						

Continuing this process we have the following.

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $(x_i - \bar{x})^2$		
	100000000000001		– 1		1	$s^2 = \frac{2}{2}$	1
	100000000000002		0		0	$s = \sqrt{s^2}$	1
	100000000000003		1		1		
sum =	**300000000000006**	**sum =**	**0**	**sum =**	**2**		
$\frac{300{,}000{,}000{,}000{,}006}{3} =$	100000000000002						

At the next stage, we have the following. Note that the observations are as recorded from Excel.

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $(x_i - \bar{x})^2$		
	1000000000000000		0		0	$s^2 = \frac{0}{2}$	0
	1000000000000000		0		0	$s = \sqrt{s^2}$	0
	1000000000000000		0		0		
sum =	**3000000000000000**	**sum =**	**0**	**sum =**	**0**		
$\frac{3{,}000{,}000{,}000{,}000{,}000}{3} =$	1000000000000000						

36. The samples are small, so graphs mainly protect against outliers and other deviations. The stemplots are as follows.

Group 1

0	
1	699
2	01247789
3	3

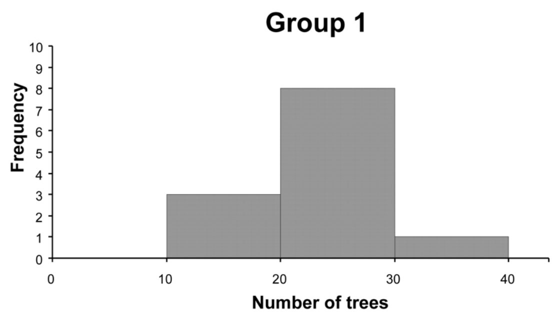
Group 2

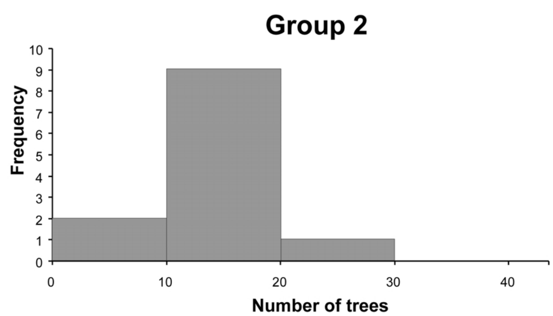
0	29
1	224457789
2	0
3	

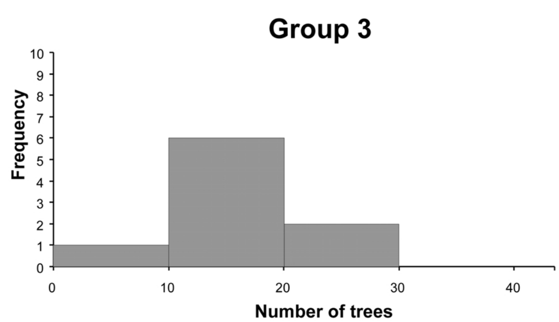
Group 3

0	4
1	225889
2	22
3	

Here is the histogram using class widths of 10 trees per 0.1 hectare in area.







All three samples are at least roughly single-peaked and symmetric, so that use of $\bar{x}$ and s is justified.

Continued on next page

36. continued

Group 1:

	Observations x_i		Deviations $x_i - \overline{x}$		Squared deviations $(x_i - \overline{x})^2$
	16		– 7.75		60.0625
	19		– 4.75		22.5625
	19		– 4.75		22.5625
	20		– 3.75		14.0625
	21		– 2.75		7.5625
	22		– 1.75		3.0625
	24		0.25		0.0625
	27		3.25		10.5625
	27		3.25		10.5625
	28		4.25		18.0625
	29		5.25		27.5625
	33		9.25		85.5625
sum =	**285**	**sum =**	**0**	**sum =**	**282.25**

$\overline{x} = \frac{285}{12} = 23.75$ and $s^2 = \frac{282.25}{12-1} = \frac{282.25}{11} \approx 25.6591$ which implies $s \approx \sqrt{25.6591} \approx 5.065$.

Group 2:

	Observations x_i		Deviations $x_i - \overline{x}$		Squared deviations $(x_i - \overline{x})^2$
	2		– 12.083333		146.00694
	9		– 5.083333		25.84027
	12		– 2.083333		4.34028
	12		– 2.083333		4.34028
	14		– 0.083333		0.00694
	14		– 0.083333		0.00694
	15		0.916667		0.84028
	17		2.916667		8.50695
	17		2.916667		8.50695
	18		3.916667		15.34028
	19		4.916667		24.17361
	20		5.916667		35.00695
sum =	**169**	**sum =**	**0.000004**	**sum =**	**272.91667**

$\overline{x} = \frac{169}{12} \approx 14.08$ (we used $\overline{x} \approx 14.083333$ in the deviations calculations for better accuracy and rounded to five decimal places in the calculation of squared deviations) and $s^2 \approx \frac{272.91667}{12-1} = \frac{272.91667}{11} \approx 24.8106$ which implies $s \approx \sqrt{24.8106} \approx 4.981$.

Continued on next page

36. continued

Group 3:

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $\left(x_i - \bar{x}\right)^2$
4	– 11.777778	138.71605
12	– 3.777778	14.27161
12	– 3.777778	14.27161
15	– 0.777778	0.60494
18	2.222222	4.93827
18	2.222222	4.93827
19	3.222222	10.38271
22	6.222222	38.71605
22	6.222222	38.71605
sum = 142	**sum = – 0.000002**	**sum = 265.55556**

$\bar{x} = \frac{142}{9} \approx 15.78$ (we used $\bar{x} \approx 15.777778$ in the deviations calculations for better accuracy and rounded to five decimal places in the calculation of squared deviations) and $s^2 \approx \frac{265.55556}{9-1} = \frac{265.55556}{8} \approx 33.1944$ which implies $s \approx \sqrt{33.1944} \approx 5.761$.

In summary, we have the following calculations.

Group 1	Group 2	Group 3
$\bar{x} = 23.75$	$\bar{x} = 14.08$	$\bar{x} = 15.78$
$s = 5.065$	$s = 4.981$	$s = 5.761$

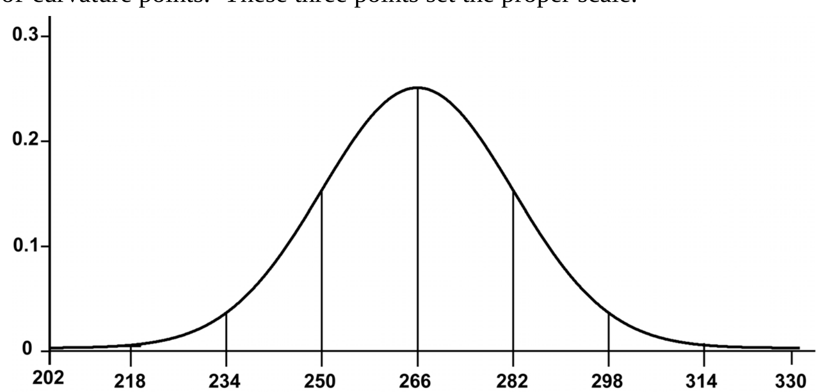
The effect of logging is clear: Group 1 (never logged) has many more species than Groups 2 and 3. Waiting 8 years (Group 3) has little effect on species richness. All three groups have similar standard deviations, so that comparisons of means are straightforward.

37. (a) $s = 0$ is smallest possible: 1, 1, 1, 1

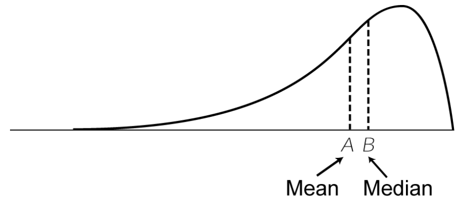
(b) Largest possible spread: 0, 0, 10, 10.

(c) In part (a), the answer is not unique. Any other set of four identical numbers will yield a standard deviation of 0, i.e. the values do not deviate from the mean, which is that repeated number. In part b, the answer is unique. The data are spread out as much as possible, given the constraints.

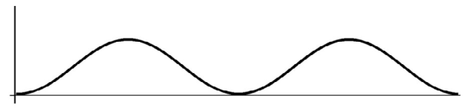
38. Sketch a normal curve, mark the axis with 266 at the center of the curve and 250 and 282 at the change-of-curvature points. These three points set the proper scale.



39. Left-skewed, so the mean is pulled toward the long left tail: A = mean and B = median.

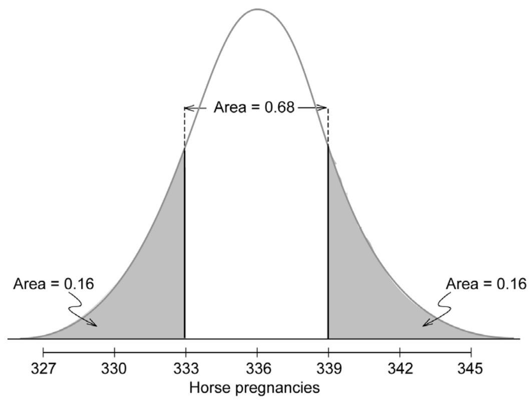


40. Think of the letter M.



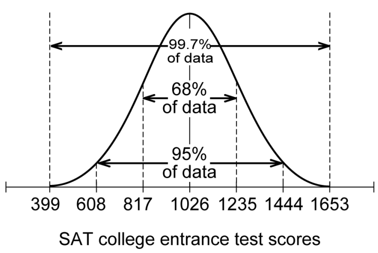
41. (a) $\mu \pm 3\sigma = 336 \pm 3(3) = 336 \pm 9$, or 327 to 345 days.

(b) Make a sketch: 339 days is one σ above μ; 68% are with σ of μ.



32% lie farther from μ. Thus, half of these, or 16%, lie above 339.

42.



(a) $\mu \pm \sigma = 1026 \pm 209$, or 817 to 1235.

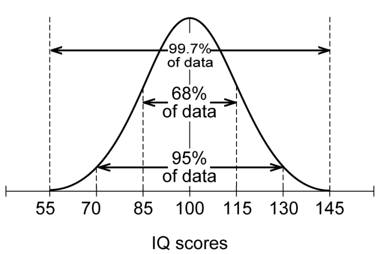
(b) 95% of scores lie within 2 standard deviations from the mean. Thus, the top 2.5% is 2 standard deviations above the mean. Given the following:

$$\sigma + 2\sigma = 1026 + 2(209)$$
$$= 1026 + 418 = 1444,$$

we have 2.5% lie above 1444.

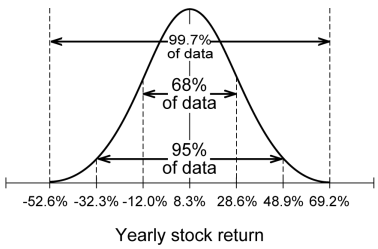
43. The quartiles are $\mu \pm 0.67\sigma = 1026 \pm 0.67(209) \approx 1026 \pm 140$, or $Q_1 = 886$ and $Q_3 = 1166$.

44.



Since $\sigma - 2\sigma = 100 - 2(15) = 100 - 30 = 70$, 70 is 2σ below μ; because 95% of values lie within 2σ of μ, 2.5% of people have IQ scores below 70.

45.



(a) $\mu \pm 2\sigma = 8.3 \pm 2(20.3) = 8.3 \pm 40.6,$ or -32.3% to 48.9%.

(b) A loss of 32.3% or greater.

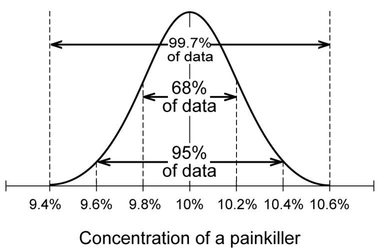
46. The quartiles mark the middle 50% of the distribution.
They are located at $\mu \pm 0.67\sigma = 8.3 \pm 13.6,$ or -5.3% to 21.9%.

47. (a) Normal curves are symmetric, so median = mean = 10%.

(b) Because 95% of values lie within 2σ of μ, $\mu \pm 2\sigma = 10 \pm 2(0.2) = 10 \pm 0.4$ implies 9.6% to 10.4% is the range of concentrations the cover the middle 95% of all the capsules.

(c) The range between the two quartiles covers the middle half of all capsules. Thus, $\mu \pm 0.67\sigma = 10 \pm 0.67(0.2) = 10 \pm 0.134$ implies 9.866% to 10.134% is the desired range.

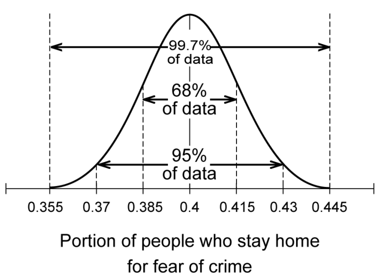
48.



(a) 10.4% is 2σ above μ, so 2.5% of capsules lie above there.

(b) 10.6% is 3σ above μ. Because 99.7% of capsules lie within 3σ of μ, 0.3% lie outside and half of these, 0.15% (0.0015) lie above.

49.



(a) 50% above 0.4, because of the symmetry of normal curves; 0.43 is 2σ above μ, so 2.5%.

(b) $\mu \pm 2\sigma = 0.4 \pm 2(0.015) = 0.4 \pm 0.03,$ or 0.37 to 0.43.

50. (a) Jermaine's standard score is $\frac{27-20.8}{4.8} = 1.29$.

(b) Tonya's standard score is $\frac{1318-1026}{209} = 1.40$.

(c) Tonya stands higher in the distribution of scores, at 1.4 standard deviations above the mean.

51. Lengths of red flowers are somewhat right-skewed, with no outliers:

37	489
38	00112289
39	268
40	67
41	5799
42	02
43	1

Lengths of yellow flowers are quite symmetric, with no outliers:

34	66
35	247
36	0015788
37	01
38	1

52. For the red variety, the $n=23$ ordered lengths are

37.40	37.78	37.87	37.97	38.01	38.07	38.10	38.20	38.23
38.79	38.87	39.16	39.63	39.78	40.57	40.66	41.47	41.69
41.90	41.93	42.01	42.18	43.09				

The minimum is 37.40, and the maximum is 43.09. Since there are 23 pieces of data, the median is the $\frac{23+1}{2}=\frac{24}{2}=12^{\text{th}}$ piece of data, namely 39.16. Since there are 11 observations to the left of the median, Q_1 is the $\frac{11+1}{2}=\frac{12}{2}=6^{\text{th}}$ piece of data, namely 38.07. Since there are 11 observations to the right of the median, Q_3 is the $12+6=18^{\text{th}}$ piece of data, namely 41.69.

Thus, the five-number summary is 37.40, 38.07, 39.16, 41.69, 43.09.

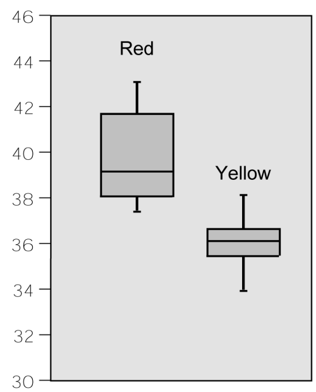
The $n=15$ yellow lengths in order are

34.57	34.63	35.17	35.45	35.68	36.03	36.03	36.11	36.52
36.66	36.78	36.82	37.02	37.10	38.13			

The minimum is 34.57, and the maximum is 38.13. Since there are 15 pieces of data, the median is the $\frac{15+1}{2}=\frac{16}{2}=8^{\text{th}}$ piece of data, namely 36.11. Since there are 7 observations to the left of the median, Q_1 is the $\frac{7+1}{2}=\frac{8}{2}=4^{\text{th}}$ piece of data, namely 35.45. Since there are 7 observations to the right of the median, Q_3 is the $8+4=12^{\text{th}}$ piece of data, namely 36.82.

Thus, the five-number summary is 34.57, 35.45, 36.11, 36.82, 38.13.

From these numbers, draw the boxplots.



Most of the red flowers are longer than all the yellow flowers. The greater variability and the right skewness of the red distribution are also visible.

53. Red:

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $(x_i - \bar{x})^2$
	37.40		– 2.311304		5.34213
	37.78		– 1.931304		3.72994
	37.87		– 1.841304		3.39040
	37.97		– 1.741304		3.03214
	38.01		– 1.701304		2.89444
	38.07		– 1.641304		2.69388
	38.10		– 1.611304		2.59630
	38.20		– 1.511304		2.28404
	38.23		– 1.481304		2.19426
	38.79		– 0.921304		0.84880
	38.87		– 0.841304		0.70779
	39.16		– 0.551304		0.30394
	39.63		– 0.081304		0.00661
	39.78		0.068696		0.00472
	40.57		0.858696		0.73736
	40.66		0.948696		0.90002
	41.47		1.758696		3.09301
	41.69		1.978696		3.91524
	41.90		2.188696		4.79039
	41.93		2.218696		4.92261
	42.01		2.298696		5.28400
	42.18		2.468696		6.09446
	43.09		3.378696		11.41559
sum =	**913.36**	**sum =**	**0.000008**	**sum =**	**71.18206**

$\bar{x} = \frac{913.36}{23} \approx 39.71$ (we used $\bar{x} \approx 39.711304$ in the deviations calculations for better accuracy and rounded to five decimal places in the calculation of squared deviations). We therefore have $s^2 \approx \frac{71.18206}{23-1} = \frac{71.18206}{22} \approx 3.2355$ which implies $s \approx \sqrt{3.2355} \approx 1.799.$

Yellow:

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $(x_i - \bar{x})^2$
	34.57		– 1.61		2.5921
	34.63		– 1.55		2.4025
	35.17		– 1.01		1.0201
	35.45		– 0.73		0.5329
	35.68		– 0.50		0.2500
	36.03		– 0.15		0.0225
	36.03		– 0.15		0.0225
	36.11		– 0.07		0.0049
	36.52		0.34		0.1156
	36.66		0.48		0.2304
	36.78		0.60		0.3600
	36.82		0.64		0.4096
	37.02		0.84		0.7056
	37.10		0.92		0.8464
	38.13		1.95		3.8025
sum =	**542.70**	**sum =**	**0.00**	**sum =**	**13.3176**

$\bar{x} = \frac{542.70}{15} = 36.18$ and $s^2 = \frac{13.31766}{15-1} = \frac{13.31766}{14} \approx 0.9513$ which implies $s \approx \sqrt{0.9513} \approx 0.975.$

The mean and standard deviation are better suited to the symmetrical yellow distribution.

54. Take $\mu = 36.18$ and $\sigma = 0.975$ millimeters.

(a) The middle 50% is spanned by the quartiles, which are 0.67σ on either side of the mean. So the range is from

$36.18 - 0.67(0.975) = 35.53$ millimeters to $36.18 + 0.67(0.975) = 36.83$ millimeters

(b) The 95 part of 68-95-99.7 rule says that the middle 95% of the distribution lies within 2σ of μ. So the range is from

$36.18 - 2(0.975) = 34.23$ millimeters to $36.18 + 2(0.975) = 38.13$ millimeters.

55. The top 2.5% of the distribution lies above

$$36.18 + 2(0.975) = 38.13 \text{ millimeters.}$$

The top 16% of the distribution lies above

$$36.18 + 1(0.975) = 37.155 \text{ millimeters.}$$

The top 25% of the distribution lies above

$$36.18 + 0.67(0.975) = 36.83 \text{ millimeters.}$$

The value 37.4 is between 37.155 and 38.13, so between 2.5% and 16% of yellow flowers are longer that 37.4 millimeters.

Word Search Solution

I	Y	E	Q	T	E	L	L	I	A	T	S	O	P	M	S	T	S	Y	V	G	N	I	I	A	S
M	U	E	R	E	R	F	J	U	O	R	M	T	E	I	T	Z	C	N	F	P	E	N	G	I	A
G	R	G	E	N	I	P	Q	G	V	O	E	I	J	R	P	S	O	G	Q	S	E	I	S	N	K
S	R	J	R	D	P	T	F	O	E	I	A	K	N	E	H	A	T	E	T	P	H	I	K	B	T
I	O	H	I	S	T	O	G	R	A	M	T	C	O	Z	F	L	I	A	L	C	A	M	E	L	H
X	A	S	C	S	J	Y	O	E	L	B	A	I	R	A	V	C	N	I	K	R	Q	H	W	L	D
S	J	Y	F	T	O	Z	H	T	X	S	P	G	M	W	O	D	H	L	O	I	A	P	E	C	E
O	N	M	A	E	U	D	R	G	F	L	N	T	A	N	A	T	O	L	P	X	O	B	D	U	G
E	I	M	I	M	T	E	G	R	S	E	I	K	I	R	E	A	M	O	L	J	I	S	D	A	K
E	O	E	X	P	L	O	R	A	T	O	R	Y	D	A	T	A	A	N	A	L	Y	S	I	S	E
T	F	T	U	L	I	A	M	M	S	G	K	D	I	P	E	Q	V	S	D	H	M	J	S	S	F
M	T	R	O	O	E	E	L	E	E	S	E	Y	S	T	H	E	H	L	I	E	E	T	T	X	P
X	Q	I	I	T	R	T	O	D	I	V	Z	D	T	E	C	N	E	A	S	A	A	Q	R	G	E
A	D	C	I	I	S	Z	N	I	I	D	E	N	R	M	E	S	A	U	T	O	N	H	I	F	O
S	F	D	D	T	X	N	T	A	T	A	T	F	I	L	C	F	R	D	R	E	E	A	B	E	F
Y	I	I	E	W	K	V	T	N	R	D	F	R	B	A	E	S	T	I	I	L	C	A	U	E	N
O	T	S	R	K	G	I	Z	X	A	H	E	K	U	F	H	R	O	V	B	G	N	A	T	A	J
Y	E	T	O	T	O	Z	L	E	U	Z	C	A	T	S	I	X	T	I	U	V	A	A	I	D	I
E	M	R	V	N	N	I	F	X	Q	M	I	O	I	O	U	R	H	D	T	E	I	E	O	A	E
I	Y	I	E	N	L	A	P	T	E	K	R	U	O	N	S	O	L	N	I	O	R	Z	N	N	S
V	D	B	S	N	E	E	A	Q	F	S	O	E	N	M	H	C	X	I	O	J	A	C	H	H	W
L	O	U	A	D	N	M	Y	R	A	M	M	U	S	R	E	B	M	U	N	E	V	I	F	J	I
A	I	T	Y	L	F	O	S	T	D	B	I	O	O	S	C	M	A	F	N	N	F	U	I	G	S
K	S	I	P	N	K	D	S	S	Q	G	D	S	W	H	O	U	G	N	A	R	F	R	O	M	X
M	N	O	M	U	A	Z	S	I	O	E	Y	Q	K	N	T	Q	S	O	H	O	H	S	L	M	F
P	X	N	S	A	I	Q	P	U	R	S	R	R	D	R	Z	E	Q	S	A	E	K	R	R	R	H