Chapter 5
Exploring Data: Distributions

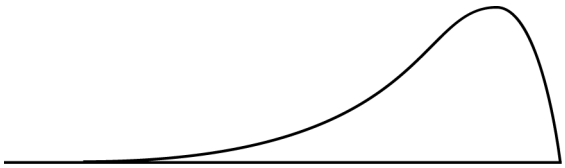
Solutions

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).

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

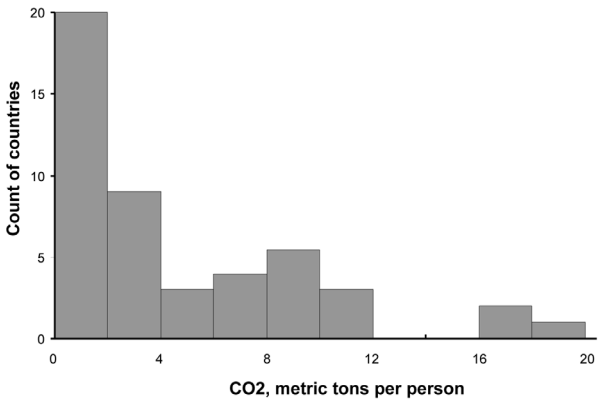


Most coins were minted in recent years, producing a peak at the right (highest-numbered years, like 2004 and 2005). There are few coins from 1990 and even fewer from 1980.

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.

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 24th and 25th in order out of 48, ignoring Alaska and Florida). The spread is from 8.5% to 15.6%.

9. Here is the stemplot.

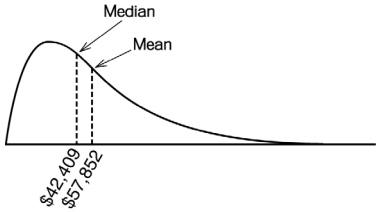
10	139
11	5
12	669
13	77
14	08
15	244
16	55
17	8
18	
19	
20	0

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.

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.

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



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, $\bar{x} = \frac{1+2+3+3+4+17}{6} = \frac{30}{6} = 5.$

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.

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 $\frac{21+1}{2} = \frac{22}{2} = 11^{\text{th}}$ 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.

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^{th} and 25^{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^{th} and 13^{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^{th} and 37^{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.

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.

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

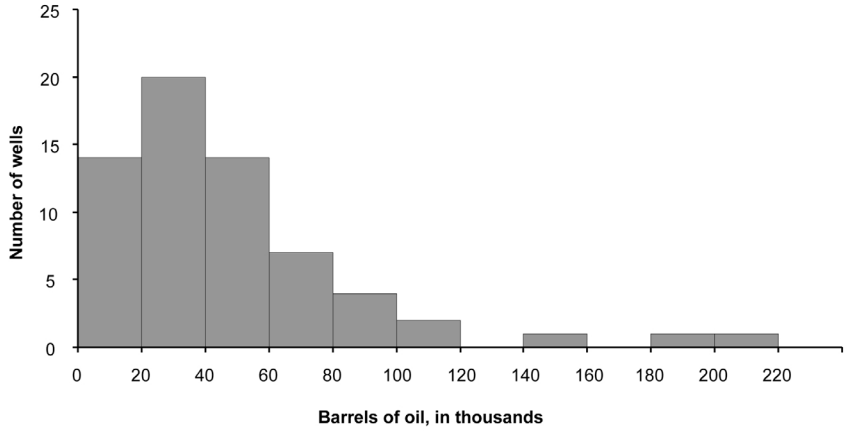
Continued on next page

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 32^{nd} and 33^{rd} 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 16^{th} and 17^{th} 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 48^{th} and 49^{th} 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.

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.

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 $\left(x_i - \bar{x}\right)^2$
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*.

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

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.

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 $\left(x_i - \bar{x}\right)^2$		
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 $\left(x_i - \bar{x}\right)^2$		
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				

Continued on next page

35. continued

Continuing this process we have the following.

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $\left(x_i - \bar{x}\right)^2$		
	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.

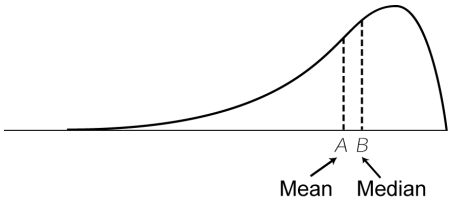
	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $\left(x_i - \bar{x}\right)^2$		
	1000000000000000		0		0	$s^2 = \frac{0}{2}$	0
	1000000000000000		0		0	$s = \sqrt{s^2}$	0
	1000000000000000		0		0		
sum =	**3000000000000000**	**sum =**	**0**	**sum =**	**0**		
$\frac{3,000,000,000,000,000}{3}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

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

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

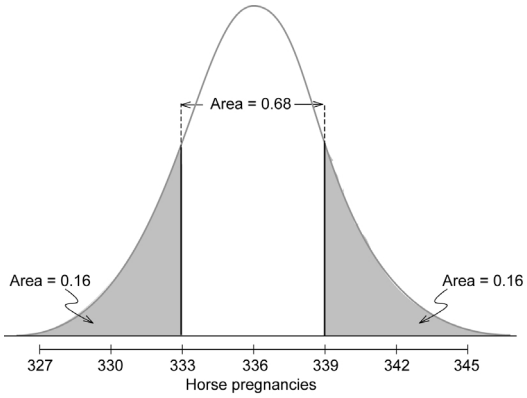
(c) In part (a), the answer is not unique. Any other set of four identical numbers will yield a standard deviation of 0, i.e. the values do not deviate from the mean, which is that repeated number. In part b, the answer is unique. The data are spread out as much as possible, given the constraints.

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



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

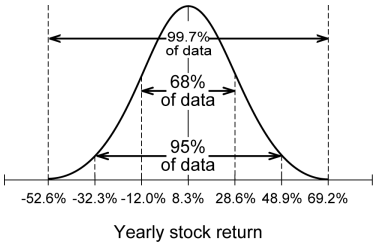
(b) Make a sketch: 339 days is one σ above μ; 68% are with σ of μ.



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

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

45.



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

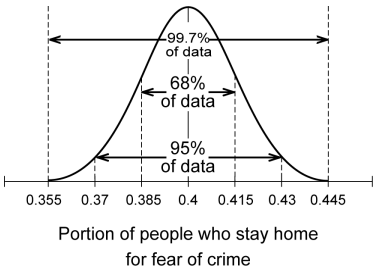
(b) A loss of 32.3% or greater.

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.

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.

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

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 $s \approx \sqrt{3.2355} \approx 1.799$.

Continued on next page

53. continued
Yellow:

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $\left(x_i - \bar{x}\right)^2$
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.

55. The top 2.5% of the distribution lies above

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

The top 16% of the distribution lies above

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

The top 25% of the distribution lies above

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

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.