

CHAPTER 2

Description of Samples and Populations

2.1.1 (a) i) Molar width

- ii) Continuous variable
- iii) A molar
- iv) 36

(b) i) Birthweight, date of birth, and race

- ii) Birthweight is continuous, date of birth is discrete (although one might say categorical and ordinal), and race is categorical
- iii) A baby
- iv) 65

• 2.1.2 (a) i) Height and weight

- ii) Continuous variables
- iii) A child
- iv) 37

(b) i) Blood type and cholesterol level

- ii) Blood type is categorical, cholesterol level is continuous
- iii) A person
- iv) 129

2.1.3 (a) i) Number of leaves

- ii) Discrete variable
- iii) A plant
- iv) 25

(b) i) Number of seizures

- ii) Discrete variable
- iii) A patient
- iv) 20

2.1.4 (a) i) Type of weather and number of parked cars

- ii) Weather is categorical and ordinal, cars is discrete
- iii) A day
- iv) 18

(b) i) pH and sugar content

- ii) Both variables are continuous
- iii) A barrel of wine
- iv) 7

2.1.5 (a) i) body mass and sex

- ii) body mass is continuous, sex is categorical
- iii) A blue jay
- iv) 123

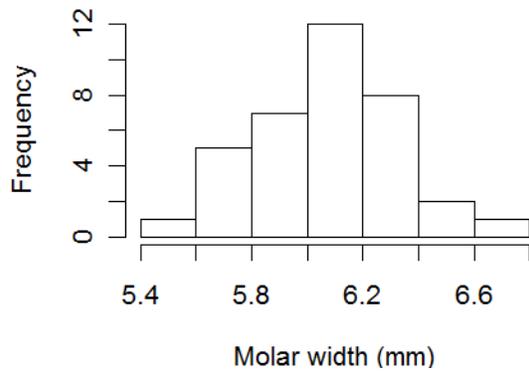
(b) i) lifespan, thorax length, and percent of time spent sleeping

28 Solutions to Exercises

- ii) Lifespan is discrete, thorax length and sleeping time are continuous
- iii) A fruit fly
- iv) 125

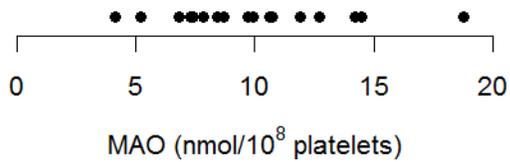
• 2.2.1 (a) There is no single correct answer. One possibility is:

Molar width	Frequency (no. specimens)
[5.4, 5.6)	1
[5.6, 5.8)	5
[5.8, 6.0)	7
[6.0, 6.2)	12
[6.2, 6.4)	8
[6.4, 6.6)	2
[6.6, 6.8)	1
Total	36



(b) The distribution is fairly symmetric.

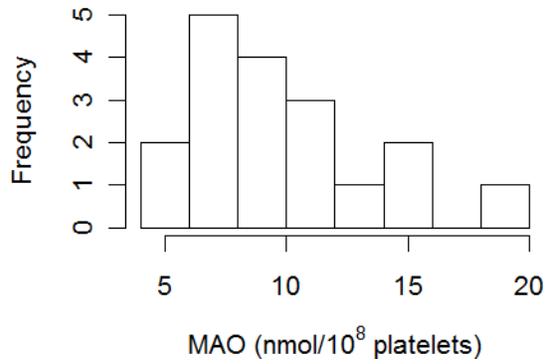
2.2.2



2.2.3 There is no single correct answer. One possibility is

MAO	Frequency (no. patients)
4.0-5.9	2
6.0-7.9	5
8.0-9.9	4
10.0-11.9	3
12.0-13.9	1
14.0-15.9	2

16.0-17.9	0
18.0-19.9	1
Total	18

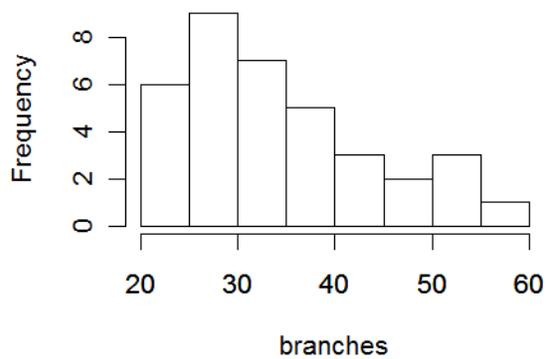


2.2.4



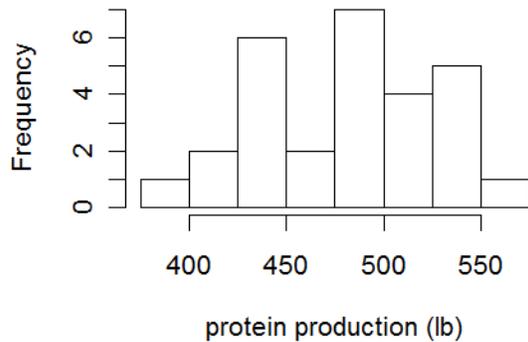
2.2.5 There is no single correct answer. One possibility is

Branches	Frequency (no. cells)
20-24	6
25-29	9
30-34	7
35-39	5
40-44	3
45-49	2
50-54	3
55-59	1
Total	36



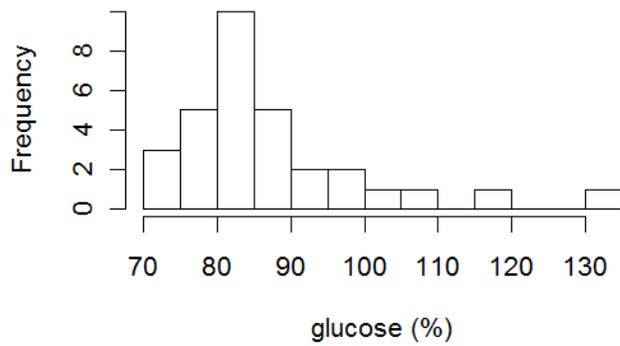
2.2.6 There is no single correct answer. One possibility is

Protein production	Frequency (no. cows)
375-399	1
400-424	2
425-449	6
450-474	2
475-499	7
500-524	4
525-549	5
550-574	1
Total	28

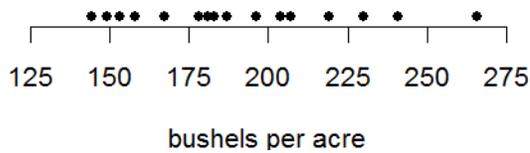


• 2.2.7 There is no single correct answer. One possibility is

Glucose (%)	Frequency (no. of dogs)
70-74	3
75-79	5
80-84	10
85-89	5
90-94	2
95-99	2
100-104	1
105-109	1
110-114	0
115-119	1
120-124	0
125-129	0
130-134	1
Total	31



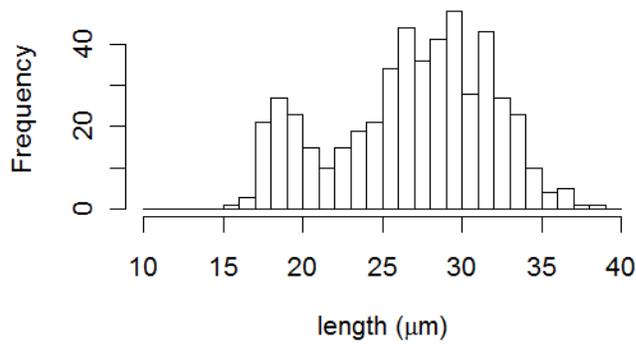
2.2.8 (a)



(b)

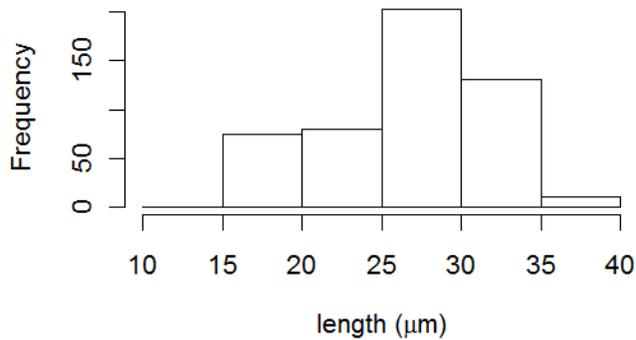
The distribution is very slightly skewed to the right.

2.2.9 (a)



(b) The distribution is bimodal.

(c) The histogram with only 6 classes obscures the bimodal nature of the distribution.



32 Solutions to Exercises

• **2.3.1** Any sample with $\Sigma y_i = 100$ would be a correct answer. For example: 18, 19, 20, 21, 22.

2.3.2 Any sample with $\Sigma y_i = 100$ and median 15 would be a correct answer. For example: 13, 14, 15, 28, 30.

2.3.3 $\bar{y} = \Sigma y_i/n = \frac{6.3 + 5.9 + 7.0 + 6.9 + 5.9}{5} = 6.40$ nmol/gm. The median is the 3rd largest value (i.e., the third observation in the *ordered* array of 5.9 5.9 6.3 6.9 7.0), so the median is 6.3 nmol/gm.

2.3.4 Yes, the data are consistent with the claim that the typical liver tissue concentration is 6.3 nmol/gm. The value of 6.3 fits comfortably near the center of the sample data.

• **2.3.5** $\bar{y} = 293.8$ mg/dl; median = 283 mg/dl.

• **2.3.6** $\bar{y} = 309$ mg/dl; median = 292 mg/dl.

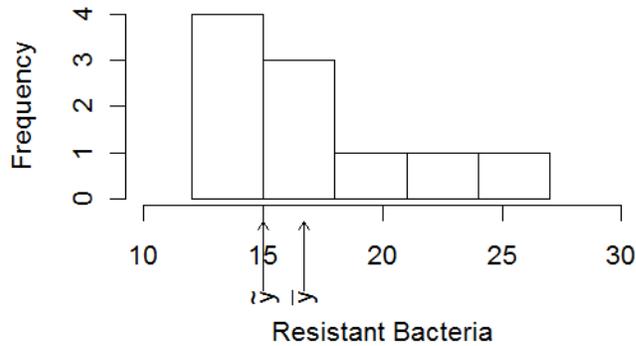
2.3.7 $\bar{y} = 3.492$ lb; median = 3.36 lb.

2.3.8 Yes, the data are consistent with the claim that, in general, steers gain 3.5 lb/day; the value of 3.5 fits comfortably near the center of the sample data. However, the data do not support the claim that 4.0 lb/day is the typical amount that steers gain. Both the mean and the median are less than 4.0; indeed, the maximum in the sample is less than 4.0.

2.3.9 $\bar{y} = 3.389$ lb; median = 3.335 lb.

2.3.10 There is no single correct answer. One possibility is

Resistant bacteria	Frequency (no. aliquots)
12-14	4
15-17	3
18-20	1
21-23	1
24-26	1
Total	10



(b) $\bar{y} = 16.7$, median = $\frac{15+15}{2} = 15$.

- **2.3.11** The median is the average of the 18th and 19th largest values. There are 18 values less than or equal to 10 and 18 values that are greater than or equal to 11. Thus, the median is

$$\frac{10+11}{2} = 10.5 \text{ piglets.}$$

2.3.12 $\bar{y} = 375/36 = 10.4$.

- **2.3.13** The distribution is fairly symmetric so the mean and median are roughly equal. It appears that half of the distribution is below 50 and half is above 50. Thus, mean \approx median \approx 50.

2.3.14 Mean \approx 35, median \approx 40

2.4.1 (a) Putting the data in order, we have

13 13 14 14 15 15 16 20 21 26

The median is the average of observations 5 and 6 in the ordered list. Thus, the median is $\frac{15+15}{2} = 15$. The lower half of the distribution is

13 13 14 14 15

The median of this list is the 3rd largest value, which is 14. Thus, the first quartile of the distribution is $Q_1 = 14$. Likewise, the upper half of the distribution is

15 16 20 21 26

The median of this list is the 3rd largest value, which is 20. Thus, the third quartile of the distribution is $Q_3 = 20$.

(b) $IQR = Q_3 - Q_1 = 20 - 14 = 6$

34 Solutions to Exercises

(c) To be an outlier at the upper end of the distribution, an observation would have to be larger than $Q_3 + 1.5(IQR) = 20 + 1.5(6) = 20 + 9 = 29$, which is the upper fence.

• **2.4.2 (a)** The median is the average of the 9th and 10th largest observations. The ordered list of the data is

4.1 5.2 6.8 7.3 7.4 7.8 7.8 8.4 8.7 9.7 9.9 10.6 10.7 11.9 12.7 14.2 14.5 18.8

Thus, the median is $\frac{8.7 + 9.7}{2} = 9.2$.

To find Q_1 we consider only the lower half of the data set:

4.1 5.2 6.8 7.3 7.4 7.8 7.8 8.4 8.7 9.7

Q_1 is the median of this half (i.e., the 5th largest value), which is 7.4.

To find Q_3 we consider only the upper half of the data set:

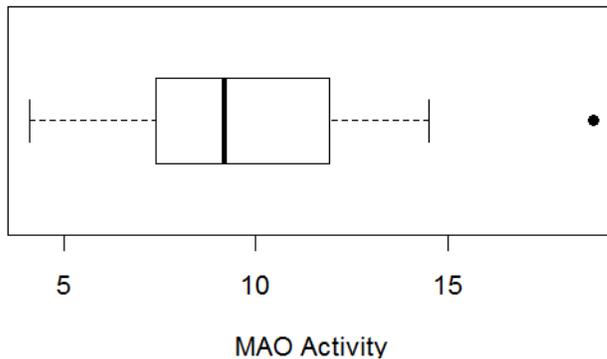
9.7 9.9 10.6 10.7 11.9 12.7 14.2 14.5 18.8.

Q_3 is the median of this half (i.e., the 5th largest value in this list), which is 11.9.

(b) $IQR = Q_3 - Q_1 = 11.9 - 7.4 = 4.5$.

(c) Upper fence = $Q_3 + 1.5 \times IQR = 11.9 + 6.75 = 18.65$.

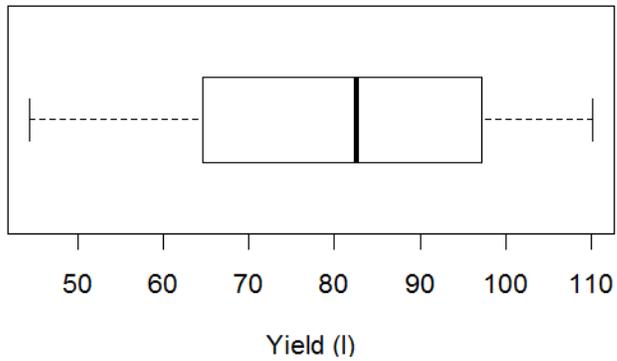
(d)



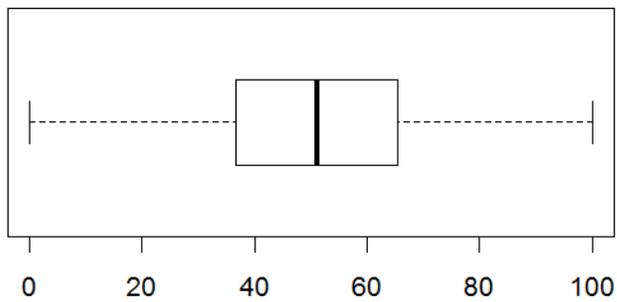
2.4.3 (a) Median = 82.6, $Q_1 = 63.7$, $Q_3 = 102.9$.

(b) $IQR = 39.2$.

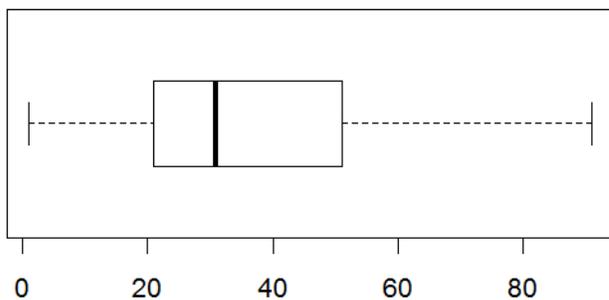
(c)



2.4.4 (a) $Q_1 = 35$, median = 50, $Q_3 = 65$.



(b) $Q_1 = 20$, median = 35, $Q_3 = 50$.



2.4.5 The histogram is centered at 40. The minimum of the distribution is near 25 and the maximum is near 65. Thus, boxplot (d) is the right choice.

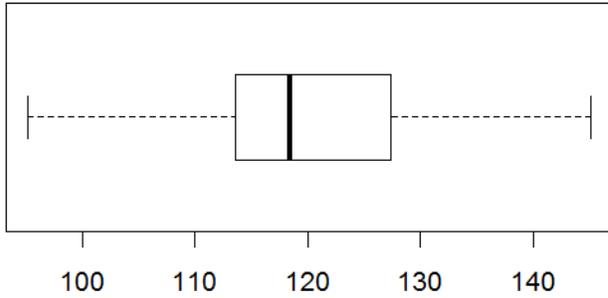
2.4.6 Yes, it is possible that there is no data value of exactly 42. The first quartile does not need to equal a data value. For example, it could be that the first quartile is the average of two points that have the values 41 and 43.

2.4.7 (a) The IQR is $127.42 - 113.59 = 13.83$.

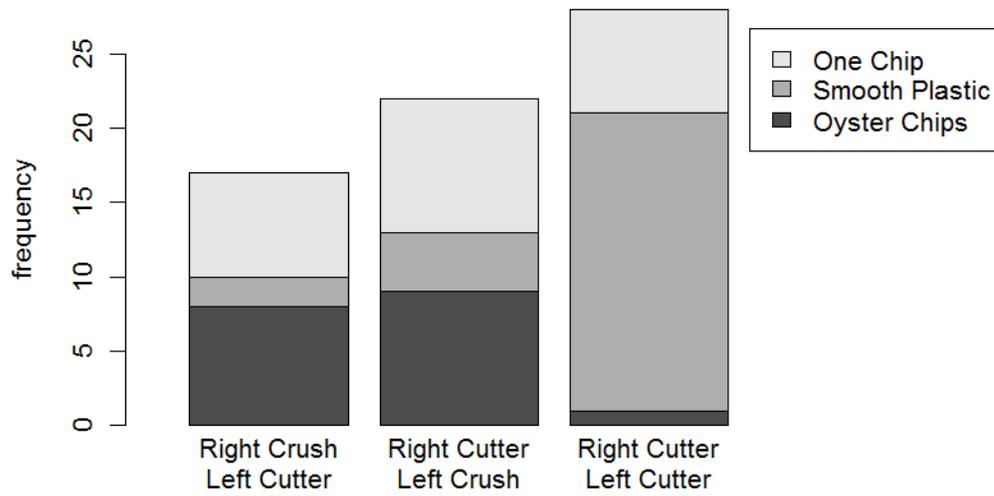
36 Solutions to Exercises

(b) For a point to be an outlier it would have to be less than $113.59 - 1.5 \cdot 13.83 = 92.845$ or else greater than $127.42 + 1.5 \cdot 13.83 = 148.165$. But the minimum is 95.16 and the maximum is 145.11, so there are no outliers present.

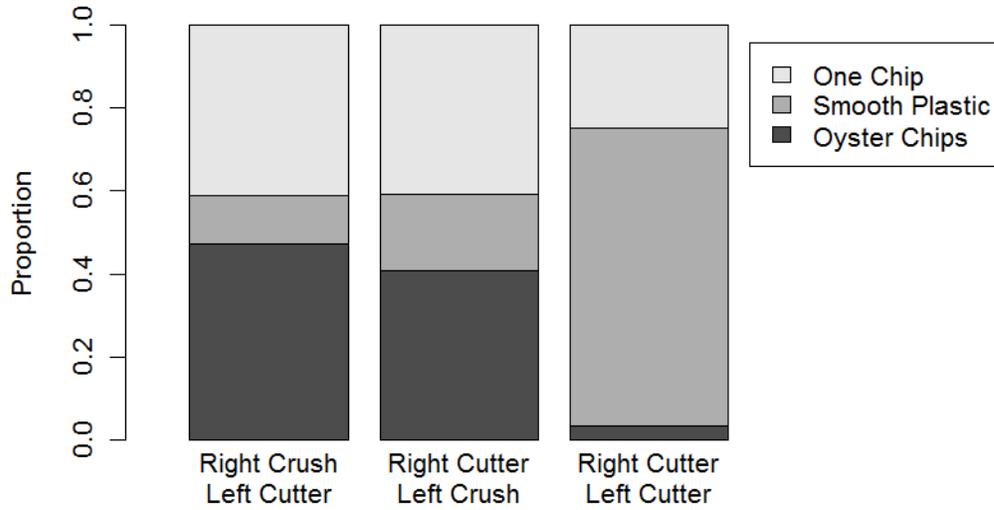
2.4.8



2.5.1 (a)

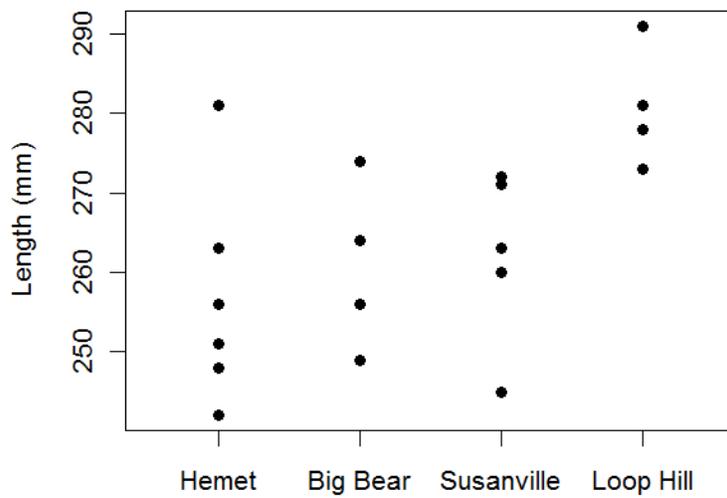


(b)

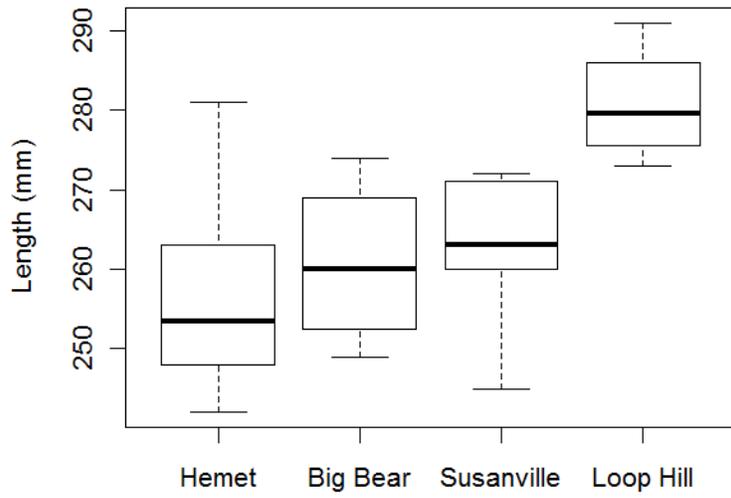


(c) The relative frequency chart in part (b) is more useful because it accounts for the different samples sizes for each claw configuration.

2.5.2 (a)

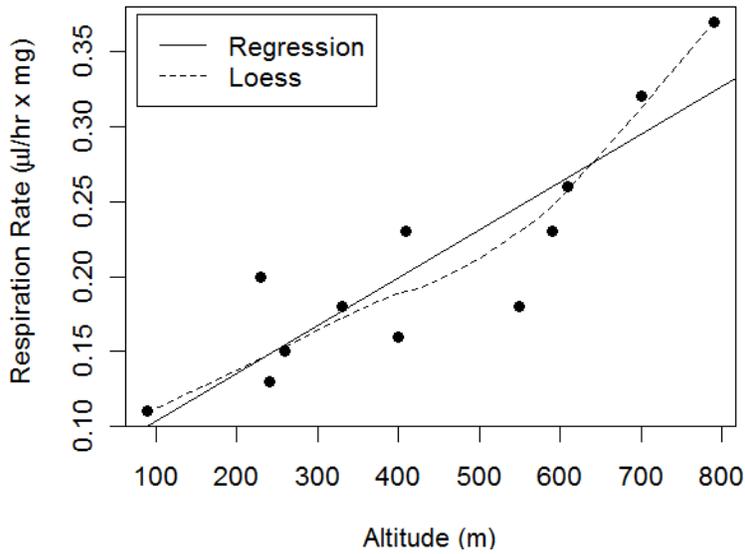


(b)

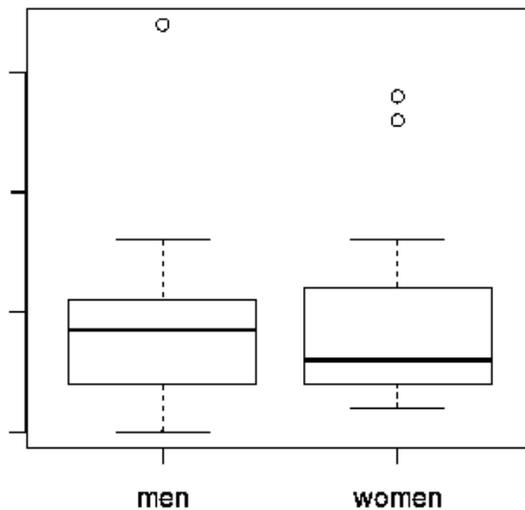


(c) Because of the small sample size, the dotplot is a simple and adequate way to represent these data.

2.5.3



2.5.4 (a)



(b) The women's distribution is skewed to the right, with the first quartile close to the minimum and the median close to the first quartile, but the third quartile farther away and the two largest values being outliers. The distribution for the men is close to being symmetric except for a high outlier. The median for the men is greater than that for the women but the first and third quartiles are similar for the sexes. The men have a lower minimum and a higher maximum.

• 2.6.1 (a) $\bar{y} = 15$, $\sum (y_i - \bar{y})^2 = 18$, $s = \sqrt{18/3} = 2.45$.

(b) $\bar{y} = 35$, $\sum (y_i - \bar{y})^2 = 44$, $s = \sqrt{44/4} = 3.32$.

(c) $\bar{y} = 1$, $\sum (y_i - \bar{y})^2 = 24$, $s = \sqrt{24/3} = 2.83$.

(d) $\bar{y} = 3$, $\sum (y_i - \bar{y})^2 = 28$, $s = \sqrt{28/4} = 2.65$.

2.6.2 (a) $\bar{y} = 7$, $\sum (y_i - \bar{y})^2 = 16$, $s = \sqrt{16/4} = 2$.

(b) $\bar{y} = 5$, $\sum (y_i - \bar{y})^2 = 6$, $s = \sqrt{6/3} = 1.4$.

(c) $\bar{y} = 6$, $\sum (y_i - \bar{y})^2 = 26$, $s = \sqrt{26/4} = 2.5$.

2.6.3 Any sample for which the deviations $(y_i - \bar{y})$ are equal to -3, -1, 0, 2, 2 would be a correct answer. Example: 17, 19, 20, 22, 22; in this case $\bar{y} = 20$.

40 Solutions to Exercises

• **2.6.4** $\bar{y} = 33.10$ lb; $s = 3.444$ lb.

2.6.5 $\bar{y} = 1.190$; $s = 0.1840$.

2.6.6 The data are -13, -29, -7, 2, -10, -43, 4, 15, -13, -30. $\bar{y} = -12.4$ mm Hg; $s = 17.6$ mm Hg.

2.6.7 (a) $\bar{y} = 6.343$; $s = 0.7020$.

(b) Median = 6.2; $Q_1 = 5.9$, $Q_3 = 6.8$, IQR = $6.8 - 5.9 = .9$.

(c) New $\bar{y} = 6.77$; new $s = 1.68$; new median = 6.2, new IQR = 0.9. The median and interquartile range display resistance in that they do not change. The standard deviation changes greatly, showing its lack of resistance. The mean changes a modest amount.

2.6.8 (a) The first quartile is the 4th largest observation, which is 26.4. The third quartile is the 12th largest observation, which is 37.5. The interquartile range is $37.5 - 26.4 = 11.1$.

(b) The range is $45.5 - 18.4 = 27.1$.

• **2.6.9 (a)** $\bar{y} \pm s$ is 32.23 ± 8.07 , or 24.16 to 40.30; this interval contains 10/15 or 67% of the observations.

(b) $\bar{y} \pm 2s$ is 16.09 to 48.37; this interval contains 15/15 or 100% of the observations.

2.6.10 According to the Empirical Rule, we expect 68% of the data to be within one SD of the mean; this is quite close to the observed 67%. We expect 95% of the data to be within two SDs of the mean. In fact, 100% of the observations are in this interval.

2.6.11 (a) $\bar{y} \pm s$ is 57.9 to 138.7; this interval contains 26/36 or 72% of the observations.

(b) $\bar{y} \pm 2s$ is 17.5 to 179.1; this interval contains 34/36 or 94% of the observations.

(c) $\bar{y} \pm 3s$ is -22.9 to 219.5; this interval contains 36/36 or 100% of the observations.

2.6.12 According to the Empirical Rule, we expect 68% of the data to be within one SD of the mean; this is close to the observed 72%. We expect 95% of the data to be within two SDs of the mean; this is quite close to the observed 94%. We expect over 99% of the data to be within three SDs of the mean; in fact, 100% of the observations are in this interval.

2.6.13 We would expect the coefficient of variation for weight to change more from age 2 to age 9. This is due to the fact that genetic factors tend to influence height more than weight and these genetic factors do not change in the course of life. On the other hand, environmental factors such as food, etc., can vary significantly.

• **2.6.14** Coefficient of variation = $\frac{s}{\bar{y}} = \frac{6.8}{166.3} = 0.04$ or 4%.

• **2.6.15** The mean is about 45. The length of the interval that covers the middle 95% of the data is approximately equal to $70 - 20 = 50$. An estimate of s is $(\text{length of interval})/4 = 50/4 \approx 12$.

2.6.16 The mean is about 100. The length of the interval that covers the middle 95% of the data is approximately equal to $150 - 60 = 90$. An estimate of s is $(\text{length of interval})/4 = 90/4 \approx 22$.

2.6.17 (a) The two SDs should be about the same. (The sample size doesn't affect the SD.)

(b) (ii) The sample of basketball players will include mostly tall men while the sample of college men will include tall, short, and average height men, leading to a larger SD.

(c) (ii) The jockeys will all be close to the same height, leading to a small SD for them.

• **2.7.1** $y' = (y - 7) \times 100$. Thus, the mean of y' is $(\bar{y} - 7) \times 100 = (7.373 - 7) \times 100 = 37.3$. The SD of y' is $s \times 100 = 0.129 \times 100 = 12.9$.

2.7.2 (a) Mean = $(36.497)(1.8) + 32 = 97.695$; SD = $(0.172)(1.8) = 0.310$

(b) Coefficient of variation = $.310/97.695 = 0.003$ or 0.3%. [Remark: This is not the same as the original coefficient of variation, which is $0.172/36.497 = 0.005$ or 0.5%.]

2.7.3 (a) Mean = $(1.461)(2.20) = 3.214$ lb/day; SD = $(0.178)(2.20) = 0.392$ lb/day.

(b) (i) Coefficient of variation = $0.178/1.461 = 0.122$ or 12.2%

(ii) Coefficient of variation = $0.392/3.214 = 0.122$ or 12.2%

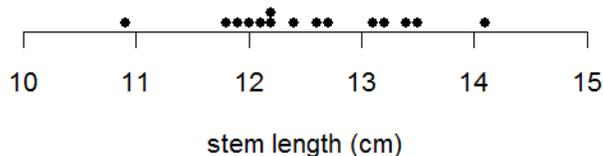
2.7.4 New mean = $(1.461-1)(100) = 46.1$; new SD = $(.178)(100) = 17.8$.

2.7.5 Both a log transformation and a square root transformation will pull in the right-hand tail and push out the left-hand tail, but a log transformation is more severe than is a square root transformation. Thus, the first histogram, (i) is for the square root transformation and (ii) is the histogram that results from a log transformation.

2.7.6 Both $\log(Y)$ and $1/\sqrt{Y}$ do a good job, with $1/\sqrt{Y}$ being slightly better (but either of these two could be considered acceptable).

2.S.1 If the mean of five observations is 181, then the sum of the five observations is $5 \times 181 = 905$. The sum of the first four observations is $180 + 182 + 179 + 176 = 717$. Thus, the height of the fifth student must be $905 - 717 = 188$ cm.

2.S.2 (a)



(b) IQR = $13.2 - 12 = 1.2$

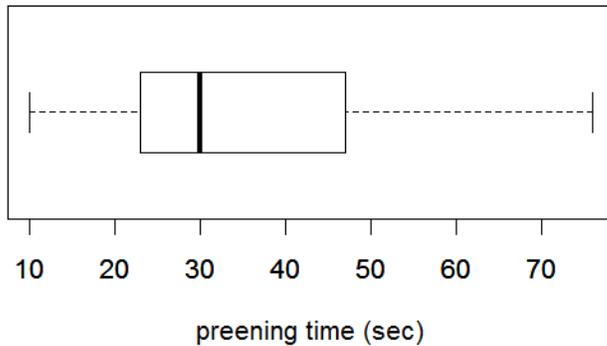
42 Solutions to Exercises

(c) To be a low outlier a value must be less than $12 - 1.5 \cdot 1.2 = 12 - 1.8 = 10.2$, so there are no outliers on the low end. To be a high outlier a value must be greater than $13.2 + 1.5 \cdot 1.2 = 13.2 + 1.8 = 15$, so there are no outliers on the high end.

2.S.3 (a) The median is $(29 + 31)/2 = 30$. $Q_1 = (22 + 24)/2 = 23$; $Q_3 = (46 + 48)/2 = 47$.

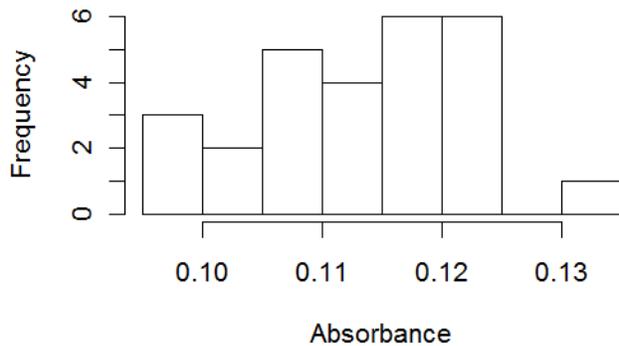
(b) $IQR = 47 - 23 = 24$.

(c) Upper fence = $47 + (1.5)(24) = 47 + 36 = 83$. Lower fence = $23 - (1.5)(24) = 23 - 36 = -13$.



2.S.4 (a) There is no single correct answer. One possibility is

Absorbance	Frequency
0.095-0.099	3
0.100-0.104	2
0.105-0.109	5
0.110-0.114	4
0.115-0.119	6
0.120-0.124	6
0.125-0.129	0
0.130-0.134	1
Total	27



2.S.5 (a) Median = 0.114; $Q_1 = 0.107$; $Q_3 = 0.120$; $IQR = 0.120 - 0.107 = 0.013$.

(b) To be an outlier on the high end of the distribution, an observation must be greater than the upper fence of $0.120 + (1.5)(0.013) = 0.120 + 0.0195 = 0.1395$.

2.S.6 The midrange is not robust, because changes in the extreme values – the minimum or the maximum – result in changes in the midrange.

2.S.7 (a) Median = $(0 + 1)/2 = 0.5$.

(b) Mean = 2.75.

(c)



(d) The distribution is strongly bimodal, with no data near the mean.

• 2.S.8 The bars in the histogram that correspond to observations less than 45 represent roughly one-third of the total area. Thus, about 30% of the observations are less than 45.

2.S.9 The bars in the histogram that correspond to observations greater than 25 represent roughly nine-tenths of the total area. Thus, about 90% of the observations are greater than 25.

2.S.10 (a) $\bar{y} = 8$, $\sum (y_i - \bar{y})^2 = 30$, $s = \sqrt{30/4} = 2.74$.

(b) $\bar{y} = 24$, $\sum (y_i - \bar{y})^2 = 36$, $s = \sqrt{36/4} = 3.00$.

(c) $\bar{y} = 2$, $\sum (y_i - \bar{y})^2 = 54$, $s = \sqrt{54/4} = 3.67$.

2.S.11 (a) (i) $\bar{y} \pm s$ is 0.76 to 3.70; this interval contains $34 + 50 + 18 = 102$ of the observations. Thus, $102/144$ or 71% of the observations are within 1 SD of the mean.

(ii) $\bar{y} \pm 2s$ is -0.71 to 5.17; this interval contains $13 + 34 + 50 + 18 + 16 + 10 = 141$ of the observations. Thus, $141/144$ or 98% of the observations are within 2 SDs of the mean.

(b) The total number of larvae is $(0)(13) + (1)(34) + (2)(50) + \dots + (7)(1) = 321$. The mean is $\bar{y} = 321/144$.

44 Solutions to Exercises

(c) The median is the average of the 72nd and 73rd largest observations. Both of these observations are in the class "2", so the median is $(2 + 2)/2 = 2$.

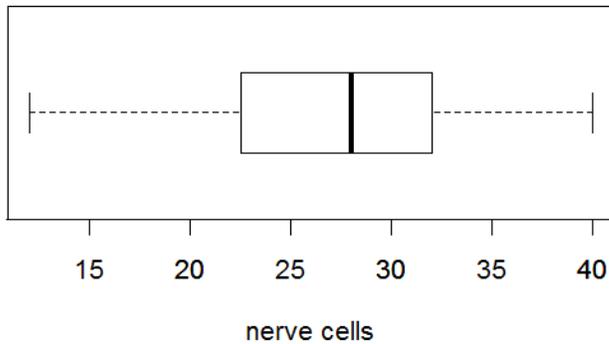
2.S.12 (a) $\bar{y} = 0.19$; $s = 4.22$.

(b) Median = 1.0.

(c) New mean = 1.44; new SD = 2.08; new median = $(1.0 + 1.5)/2 = 1.25$. The median displays resistance. The mean and the SD change greatly, showing lack of resistance.

2.S.13 (a) Median = 28; $Q_1 = 22$; $Q_3 = 33$; IQR = $33 - 22 = 11$.

(b) Upper fence = $33 + (1.5)(11) = 49.5$; lower fence = $22 - (1.5)(11) = 5.5$.

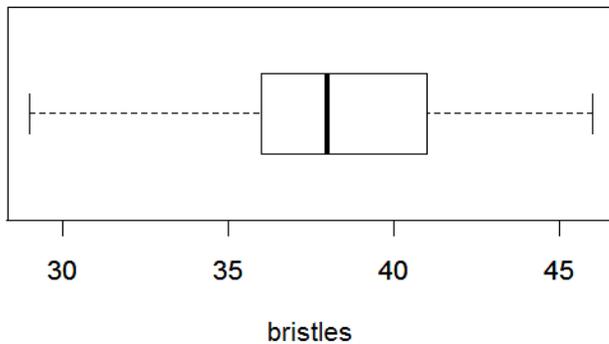


2.S.14 Yes, the distribution appears to be reasonably symmetric and mound-shaped.

• 2.S.15 (a) $n = 119$, so the median is the 60th largest observation. There are 32 observations less than or equal to 37 and 44 observations less than or equal to 38. Thus, the median is 38.

(b) The first quartile is the 30th largest observation, which is 36. The third quartile is the 90th largest observation, which is 41.

(c)



(d) The mean is 38.45 and the SD is 3.20. Thus, the interval $\bar{y} \pm s$ is 38.45 ± 3.20 , which is 35.25 to 41.65. This interval includes 36, 37, 28, 29, 40, and 41. The number of flies with 36 to 41 bristles

is $11 + 12 + 18 + 13 + 10 + 15 = 79$. Thus, the percentage of observations that fall within one standard deviation of the mean is $79/119 * 100\% = 0.664 * 100\% = 66.4\%$.

(e) The quartiles are 36 and 41. Between 36 and 41 (inclusive) there are 79 observations out of 119 total so 66.4%.

2.S.16 (a) Mean increase = 3.870; SD of increase = 1.274.

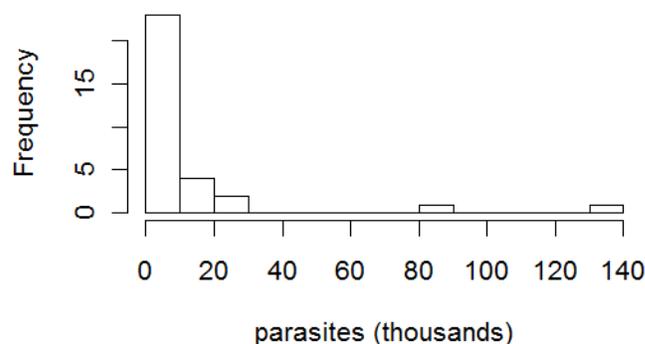
(b) Mean before = 1.680, mean after = 5.550. Yes; $5.550 - 1.680 = 3.870$.

(c) The median increase is $(3.5 + 3.5)/2 = 3.5$.

(d) Before: The median increase is $(1.5 + 1.5)/2 = 1.5$.
After: The median increase is $(5.2 + 5.8)/2 = 5.5$.
No; $5.5 - 1.5 \neq 3.5$.

2.S.17 (a)

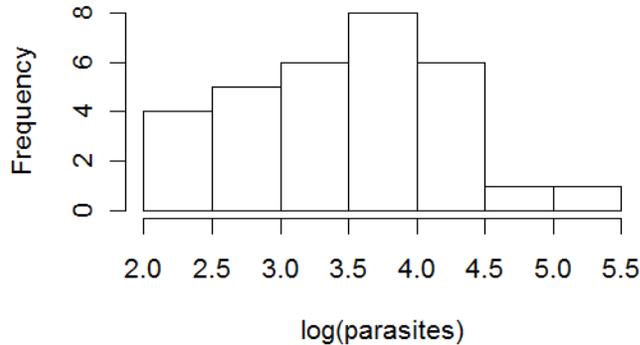
Number of Parasites	Frequency
0-9,999	23
10,000-19,999	4
20,000-29,999	2
30,000-39,999	0
40,000-49,999	0
50,000-59,999	0
60,000-69,999	0
70,000-79,999	0
80,000-89,999	1
90,000-99,999	0
100,000-109,999	0
110,000-119,999	0
120,000-129,999	0
130,000-139,999	1
Total	31



(b)

Log (Number of Parasites)	Frequency
---------------------------	-----------

2.000-2.499	4
2.500-2.999	5
3.000-3.499	6
3.500-3.999	8
4.000-4.499	6
4.500-4.999	1
5.000-5.499	1
Total	31



The original distribution is highly skewed; the distribution of the logs is much less skewed.

(c) The original mean is 12,889.6; the mean of the logs is 3.4854. The log of the original mean is $\log(12,889.6) = 4.1102 \neq 3.4854$. No, they are not equal.

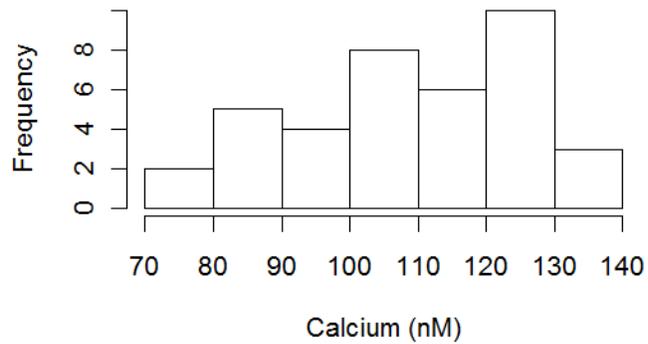
(d) The original median is 3672; the median of the logs is 3.5649. The log of the original median is $\log(3673) = 3.5649$. Yes, they are equal.

2.S.18 The first histogram is a rough histogram for the data, because the data imply that the distribution is skewed to the right. The average is 3.6 and the SD is 1.6. If the distribution were symmetric and mound shaped, like the middle histogram, then there would be data throughout the interval $\bar{y} \pm 2s$, which is 3.6 ± 3.2 , or 0.4 to 6.8. However, the minimum is 1.2, which is greater than 0.4. The situation for the third histogram is even worse (i.e., for this histogram there are values much lower than the mean).

2.S.19 The fourth computer output has the largest SD and has the mean and median nearly equal. This corresponds to histogram (b), which has the most spread. Histogram (c) has a mean which exceeds the mean. It follows that this histogram corresponds to the first computer output. Histogram (a) has the smallest SD and the mean and median are nearly equal, so the second computer output corresponds to histogram (a). The third computer output has a lower mean than median, so it is not used.

2.S.20 The low volume distribution is symmetric, centered at 20, with a minimum of 0 and a maximum of 40. The high volume distribution is shifted down from the low volume distribution, with a median of about 18 and a maximum of 30, which is the third quartile for the low volume distribution. Thus, one-fourth of the low volume hospitals have mortality rates greater than the highest mortality rate among high volume hospitals.

2.S.21 The mean is 107.87, the median is 108.5, and the SD is 16.08.



The histogram is fairly symmetric.

2.S.22 The median is 28, which is consistent with any of the histograms. Likewise the minimum and maximum values agree for all three histograms. However, the boxplot shows that the distribution has a small IQR and is skewed to the right, which means that histogram (a) is correct.

2.S.23 The mean is pulled down by the long left-hand tail, which results in the mean being less than the median.

2.S.24 Approximately 15% of the area in the histogram is to the left of 40.