2 Making Sense of Your World with Statistics

2.1 Summarizing Data with a Few Good Numbers

Exercise 2.1.1. The mean is $28/7 = 4$, the median is the fourth data piece, so 4 and the mode is also 4.

Exercise 2.1.2.

data	39	47	52	55	57	59	60	63	64	66	67	70	72	75	77	78	80
frequency	1	1	1	2	4	1	2	1	1	3	1	2	2	3	1	1	1

Exercise 2.1.3. The mean is the sum of our data, which is 1,792, divided by the total number of pieces of data, which is 28. Thus the mean is $1{,}792/28 = 64$. The median is average of the 14th and 15th pieces of data, so it is $(64 + 65)/2 = 65$ and the mode is 57.

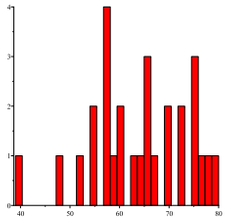
Exercise 2.1.4. (a) There are 28 pieces of data, 67 is the 18th data piece and $18/28 = 0.643$, hence the 64.3 th percentile. (b) 77 is the 26th piece of data and $26/28 = 0.929$, hence in the 92.9th percentile.

Exercise 2.1.5. The 25th percentile is the number in the 7th position (since $7/28 = 1/4$).

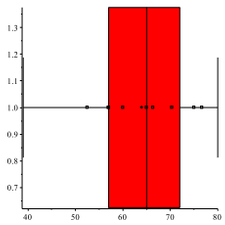
Exercise 2.1.6. We will use linear interpolation. After reordering the number in increasing order, the 19th number, 70, corresponds to the 67.9 percentile, while the 20th number, also 70, corresponds to the 71.4 percentile. The number 70 is $70 - 67.9 = 2.1$ away from 67.9, out of a spread of $71.4 - 67.9 = 3.5$ between the two percentiles, so 70 is $2.1/3/5 = 0.6$ of this spread. Thus, using linear interpolation we get that the 70th percentile is at $70 + 0.6\ times 70 = 70$.

Exercise 2.1.7. (a) The frequency plot is

K20611_SM_Cover.indd 17 16/04/15 6:05 pm



(b) The boxplot is



Exercise 2.1.8. (a) Q_1 is the median of the lower half of the data, so we get $(57+57)/2 = 57$. Q_3 is the median of the upper half of the data, so we get $(72+72)/2 = 72$.
(b) The range is the difference between the maximum and minimum values, so $80 - 39 = 41$. The IQR is $Q_3 - Q_1 = 72 - 57 = 15$.

Exercise 2.1.9. 39, 57, 65, 72, 80

Exercise 2.1.10. The lower fence (the leftmost vertical line) for outliers is $Q_1 - 1.5\times$ IQR $= 57 - 1.5(15) = 34.5$ and the upper fence (the rightmost vertical line) is $Q_3 + 1.5\times$ IQR $= 94.5$. Since our data falls within these values, we have no outliers.

Exercise 2.1.11. (a) $[64 - 9.72, 64 + 9.72] = [54.28, 73.72]$. Since 16 numbers fall within this range, the proportion of data is $16/28 = 57.1\%$.
(b) $[64 - 2(9.72), 64 + 2(9.72)] = [44.56, 83.44]$. Since 25 numbers fall within this range, the proportion of data is $25/28 = 89.3\%$.
(c) $[64 - 3(9.72), 64 + 3(9.72)] = [34.84, 93.16]$. Since 27 numbers fall within this range, the proportion of data is $27/28 = 96.4\%$.

Exercise 2.1.12. The mean is the sum of all the data, 2,622, divided by the number of pieces of data, 7. So the mean is $2{,}622/7 = 374.57$, the variance is

$(323 - 374.57)^2 + (384 - 374.57)^2 + \ldots + (407 - 374.57)^2 = 3{,}318.3$

, and the standard deviation is the square root of the variance, namely $\sqrt{3{,}318.8} = 57.61$.

Exercise 2.1.13. The mean is the sum of all the data, 2,682, divided by the number of pieces of data, 7, giving a mean of $2{,}682/7 = 383.14$, the variance is

$$(323 - 383.14)^2 + (389 - 383.14)^2 + \ldots + (507 - 383.14)^2 = 6{,}808.98,$$

and the standard deviation is the square root of the variance, namely $\sqrt{6{,}808.98} = 82.5$.

Exercise 2.1.14. First, convert the heights from feet to inches.
(a) The mean is 79.65 inches (6 feet 7.65 inches).
(b) The median height is 80 inches (6 foot 8 inches).
(c) The standard deviation is 2.83 inches.

Exercise 2.1.15. (a) The mean number of accidents is 11,020,000. (b) The standard deviation is 906,000.

Exercise 2.1.16. (a) The mean of the mean scores is 516.18. (b) The standard deviations is 2.08. (c) The mean score each year is over 500, so the mean throughout the years will likely be higher than 500.

Exercise 2.1.17. The standard deviation is smaller since the removed values had a large deviation from the mean, and the standard deviation is not resistant to outliers.

Exercise 2.1.18. The winning percent is $77^2/(77^2 + 63^2) = 0.599$. Since 59.9% of 45 games is 26.95 we would expect them to win approximately 27 games.

Exercise 2.1.19. Winning Percent is computed as $S^2/(S^2 + A)$. We know that the winning percent is $36/45 = 0.8$ and that 51 goals were allowed, so we have that $S^2/(S^2 + 51^2) = 0.8$. Thus we have that $S^2 = 0.8S^2 + 2{,}080.8$ Solving for S, we find that $S = 102$, so about 102 goals were scored.

Exercise 2.1.20. (a)

song title	duration (sec)
A Hard Day's Night	154
I Should Have Known Better	163
If I Fell	139
I'm Happy Just to Dance with You	116
And I Love Her	150
Tell Me Why	129
Can't Buy Me Love	132
Any Time at All	131
I'll Cry Instead	106
Things We Said Today	155
When I Get Home	137
You Can't Do That	155
I'll Be Back	144

(b) The mean duration of a song is 139.3 seconds, or 2 minutes 19 seconds.

K20611_SM_Cover.indd 19 16/04/15 6:05 pm

(c) The median duration of a song is 139 seconds.
(d) The standard deviation of a song's duration is 5.9 seconds.

Exercise 2.1.21. (a)

song title	duration (sec)
Sgt. Pepper's Lonely Hearts Club Band	122
With a Little Help from My Friends	164
Lucy in the Sky with Diamonds	208
Getting Better	168
Fixing a Hole	156
She's Leaving Home	215
Being for the Benefit of Mr. Kite!	157
Within You Without You	304
When I'm Sixty-Four	156
Lovely Rita	162
Good Morning Good Morning	161
Sgt. Pepper's Lonely Hearts Club Band (Reprise)	79
A Day in the Life	339

(b) The mean duration of a song is 184 seconds (3 minutes 4seconds)
(c) The median duration of a song is 162 seconds (2 minutes 40 seconds)
(d) The standard deviation of a song's duration is 67.22 seconds.
(e) Some of these songs are substantially longer than the previous album, hence a larger mean. They also have more variance within the lengths of the songs, hence a larger standard deviation.

Exercise 2.1.22. The standard deviation is 0 exactly when all the data values are equal, since the deviations from the mean needs to be 0 for each piece of data.

Exercise 2.1.23. If an outlier is added to a data set, the mean, range and standard deviation will most likely be affected, as they are not resistant to outliers. For example, for the data $1, 2, 3, 4, 5, 6$ the five number summary is $1, 2, 3.5, 6$ and the mean is 3.5 and standard deviation is 1.87. The range is 5 and IQR is 3. For the data set $1, 2, 3, 4, 5, 6, 6{,}000$, the five number summary is $1, 2.5, 4, 5.5, 6{,}000$ and the mean is 860.14 and standard deviation is 2,266.5. The range is 5,999 and IQR is 3.

Exercise 2.1.24. Michael Jackson (music), Nelson Mandela (leadership), Wayne Gretzky (hockey).

2.2 *Estimating Unknowns*

Exercise 2.2.1. Simple Random Sample

Exercise 2.2.2. Cluster Sampling

Exercise 2.2.3. Stratified Sampling

Exercise 2.2.4. Systematic Sampling

Exercise 2.2.5. Cluster Sampling

Exercise 2.2.6. Simple Random Sampling

Exercise 2.2.7. Population Statistic

Exercise 2.2.8. Sample Statistic

Exercise 2.2.9. Population Statistic

Exercise 2.2.10. Sample Statistic

Exercise 2.2.11. (a) $\hat{p} = 11/50 = 0.22$
(b) The margin of error is $z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.64\sqrt{\frac{0.22(1-0.22)}{50}} = 0.096$. The 90% confidence interval is $\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = [0.22 - 0.096, 0.22 + 0.096] = [0.124, 0.316]$.
(c) The margin of error is $z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.22(1-0.22)}{50}} = 0.115$. The 95% confidence interval is $\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = [0.22 - 0.115, 0.22 + 0.115] = [0.105, 0.335]$.
(d) The margin of error is $z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2.57\sqrt{\frac{0.22(1-0.22)}{50}} = 0.151$. The 95% confidence interval is $\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = [0.22 - 0.151, 0.22 + 0.151] = [0.069, 0.371]$.

Exercise 2.2.12. At all three levels computed above, as the 90%, 95% and 99% confidence intervals contain 0.181.

Exercise 2.2.13. Very Well: $[0.24, 0.37]$
Fairly Well: $[0.47, 0.55]$
Not Very Well: $[0.1, 0.18]$
Not at All: $[-0.02, 0.06]$ (or just $[0, 0.06]$, as a proportion can't be negative)
No Opinion: $[-0.04, 0.04]$ (or just $[0, 0.04]$, as a proportion can't be negative)

K20611_SM_Cover.indd 21 16/04/15 6:05 pm

Exercise 2.2.14. The margin of error, E is given by

$$E = z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

so using a bit of algebra, we find that

$$n = \left(\frac{z}{E}\right)^2 \times \hat{p}(1-\hat{p}),$$

and any larger n will certainly not increase the margin of error. Moreover, as in the text, $\hat{p}(1-\hat{p})$ is always biggest when $\hat{p} = 1/2$, so that a value of n that will do is any such n that satisfies

$$n \geq \left(\frac{z}{E}\right)^2 \times 0.5 \times (1-0.5) = 0.25 \times \left(\frac{z}{E}\right)^2.$$

Now for $z = 1.96$ and $E = 0.02$, we have

$$n \geq = 0.25 \times \left(\frac{1.96}{0.02}\right)^2 (1/2)(1/2) = 2{,}401,$$

so 2,401 people would suffice. (Note: should our answer had not been a whole number, we always round up.)

Exercise 2.2.15. As in the solution to the previous problem,

$$n \geq 0.25 \times \left(\frac{z}{E}\right)^2 = 0.25 \times \left(\frac{1.64}{0.02}\right)^2 = 1{,}681$$

so 1,681 people would suffice.

Exercise 2.2.16. As in the solution to Exercise 2.2.14,

$$n \geq 0.25 \times \left(\frac{z}{E}\right)^2 = 0.25 \times \left(\frac{2.57}{0.02}\right)^2 (1/2)(1/2) = 4{,}128.06,$$

so 4,129 people would suffice.

Exercise 2.2.17. (a) The margin of error is $1.64\sqrt{\frac{\hat{p}(1-\hat{p})}{1{,}500}}$. The biggest this can be occurs when $\hat{p} = 1/2$. We then get a margin of error of approximately 0.021 or 2.1%.
(b) The margin of error is $1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{1{,}500}}$. The biggest this can be occurs when $\hat{p} = 1/2$. We then get a margin of error of approximately 0.025 or 2.5%.
(c) The margin of error is $2.57\sqrt{\frac{\hat{p}(1-\hat{p})}{1{,}500}}$. The biggest this can be occurs when $\hat{p} = 1/2$. We then get a margin of error of approximately 0.033 or 3.3%.

K20611_SM_Cover.indd 22 16/04/15 6:05 pm

Exercise 2.2.18. Our alternative hypothesis would be that the drug works better than the placebo. Based on our p-value we have the following conclusions:
(a) Virtually no evidence for the alternative hypothesis.
(b) Moderate evidence for the alternative hypothesis.
(c) Good evidence for the alternative hypothesis.
(d) Very good evidence for the alternative hypothesis.

Exercise 2.2.19. From the data you compute, the curve should look bell-curved.

Exercise 2.2.20. From the data you compute, the curve should also look bell-curved.

Exercise 2.2.21. (a) Turn the values into inches, to obtain $64''$, $69''$, $63''$, $66''$, $59''$, $67''$. The sample mean is 64.67 inches and the sample standard deviation is 3.502 inches.
(b) The margin of error is $\frac{2.571 \times 3.502}{\sqrt{6}} = 3.676$. The confidence interval is $[64.47 - 3.676, 64.47 + 3.676] = [60.994, 68.346]$.

Exercise 2.2.22. (a) The sample mean is 187.73 and the sample standard deviation is 16.85.
(b) The margin of error is $t \times \frac{s}{\sqrt{n}} = 3.169 \times \frac{16.85}{\sqrt{11}} = 16.10$. So the 99% confidence interval is $[187.73 - 16.10, 187.73 - 16.10] = [171.63, 203.83]$.

Exercise 2.2.23. (a) The margin of error is $t \times \frac{s}{\sqrt{n}} = 1.676 \times \frac{14.02}{\sqrt{51}} = 3.29$. So the 90% confidence interval is $[53.57 - 3.29, 53.57 + 3.29] = [50.28, 56.86]$.
(b) The margin of error is $t \times \frac{s}{\sqrt{n}} = 2.009 \times \frac{14.02}{\sqrt{51}} = 3.94$. So the 95% confidence interval is $[53.57 - 3.94, 53.57 + 3.94] = [49.57, 57.51]$.
(c) The margin of error is $t \times \frac{s}{\sqrt{n}} = 2.678 \times \frac{14.02}{\sqrt{51}} = 5.26$. So the 99% confidence interval is $[53.57 - 5.26, 53.57 + 5.26] = [48.31, 58.83]$.

Exercise 2.2.24. For 90% the margin of error is

$$t \times \frac{s}{\sqrt{n}} = 1.676 \times \frac{14.02}{\sqrt{51}} = 3.29.$$

For 95% the margin of error is

$$t \times \frac{s}{\sqrt{n}} = 2.009 \times \frac{14.02}{\sqrt{51}} = 3.94.$$

For 99% the margin of error is

$$t \times \frac{s}{\sqrt{n}} = 2.678 \times \frac{14.02}{\sqrt{51}} = 5.26.$$

Exercise 2.2.25. At the 90% and 95% confidence levels we could not agree, as \$58.05 is not in the confidence intervals, but at the 99% confidence level we could agree, since \$580.05 is in the confidence interval.

K20611_SM_Cover.indd 23 16/04/15 6:05 pm

Exercise 2.2.26. At the 90%, 95% and 99% levels of confidence we could agree, as \$54.45 is in those confidence intervals.

Exercise 2.2.27. Since the margin of error is $t \times \frac{s}{\sqrt{n}}$, we have that $4.21 = 1.646 \times \frac{s}{\sqrt{1001}}$. Solving for s, we get $s = 80.923$.

2.3 *Leading You Down the Garden Path with Statistics*

Exercise 2.3.1. Non-representative sample

Exercise 2.3.2. Self-selection bias

Exercise 2.3.3. Participants are likely to lie

Exercise 2.3.4. Non-representative sample

Exercise 2.3.5. Self-selection bias

Exercise 2.3.6. Mean is not the appropriate statistic in this case

Exercise 2.3.7. Non-representative sample

Exercise 2.3.8. Non-representative sample

Exercise 2.3.9. Quota bias

Exercise 2.3.10. Interview bias

Exercise 2.3.11. Median is resistant to outliers, and may not be the appropriate statistic in this case

Exercise 2.3.12. Mean is not resistant to outliers, and may not be the appropriate statistic in this case

Exercise 2.3.13. Non-representative sample

Exercise 2.3.14. Correlation does not imply causation

Exercise 2.3.15. Checking Facebook while driving could cause accidents! You could create an experiment in which you take some randomly selected people who don't use Facebook and randomly assign some participants to use Facebook. You then track accidents over a certain time frame.

Exercise 2.3.16. No significance level was given.

K20611_SM_Cover.indd 24 16/04/15 6:05 pm

Exercise 2.3.17. Since the p-value is very small, there is very good evidence for the alternative hypothesis.

Exercise 2.3.18. You might not pay exorbitant salaries to sons and daughters of top players until they prove themselves.

Exercise 2.3.19. It is difficult to think of any instances where the children were as extreme!

Exercise 2.3.20. Yes; perhaps the manager should not get overly excited by a player's performance over half a season, thinking a trend has developed.

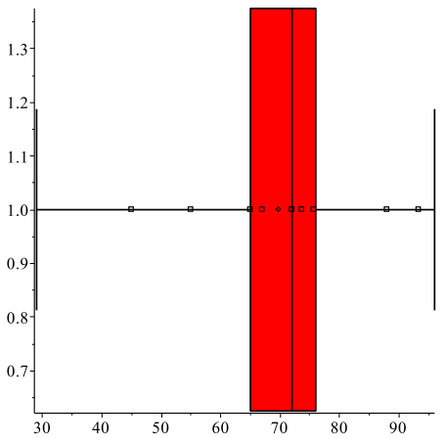
2.4 *Review Exercises*

Exercise 2.4.1. The mean is 3.875, the median is 4 and the mode is 4.

Exercise 2.4.2. The mean is 141.567, the median is 138.1 and there is no mode, as all the values appear once.

Exercise 2.4.3. (a) The mean is 501.9. (b) The standard deviation is 4.23. (c) Potentially, as 500 is within one standard deviation of the mean.

Exercise 2.4.4. (a) The mean is 69.769, the median is 72 and the mode is 65.
(b) There are 13 pieces of data and 68 is the 6th piece. Since $6/13 = 0.462$, 68 is the 46.th percentile.
(c) 75 is the 9th piece of data. Since $9/13 = 0.692$, 75 in the 69th percentile.
(d)



(e) Q1 is the median of the lower half of the data, hence is it 65. Q3 is the median of the

upper half of the data, hence is it 76.
(f) The range is Max $-$ Min $= 96 - 29 = 67$. The IQR $=$ Q3 $-$ Q1 $= 76 - 65 = 11$.
(g) 29, 65, 72, 76, 96
(h) The lower fence (the leftmost vertical line) is $65 - 1.5(11) = 48.5$, the upper fence (the rightmost vertical line) is $76 + 1.5(11) = 92.5$. Since 29 is lower than our lower fence and 96 is larger than the upper fence, these values are outliers.

Exercise 2.4.5. (a) Cluster (b) Systematic (c) Stratified

Exercise 2.4.6. (a) $64/100 = 0.64$
(b) The margin of error is $1.64\sqrt{\frac{(0.64)(0.36)}{100}} = 0.0787$. The confidence interval is $[0.5613, 0.7187]$.
(c) The margin of error is $1.96\sqrt{\frac{(0.64)(0.36)}{100}} = 0.0941$. The confidence interval is $[0.5459, 0.7341]$.
(d) The margin of error is $2.57\sqrt{\frac{(0.64)(0.36)}{100}} = 0.1234$. The confidence interval is $[0.5166, 0.7634]$.

Exercise 2.4.7. The appropriate confidence interval for the proportion who believe that the temperature increase is due to human causes is $0.57 \pm 0.04 = (0.53, 0.61)$. Our confidence interval has values all above 0.5, so it seems unlikely that the population proportions of those who think the increases is due to human activities and those who think it is due to other causes are equal (both 50%).

Exercise 2.4.8. The alternative hypothesis is that the new arthritis medicine is better than the placebo.
(a) Moderate evidence for the alternative hypothesis.
(b) Very good evidence for the alternative hypothesis.
(c) Virtually no evidence for the alternative hypothesis.

Exercise 2.4.9. (a) Self-selection bias
(b) Non-representative sample
(c) Mean is not the appropriate statistic in this case

Exercise 2.4.10. We cannot conclude that hearing loss requires the use of a cane, since correlation does not imply causation. Likely, both are due to old age!

K20611_SM_Cover.indd 26 16/04/15 6:05 pm