Unit 5: Boxplots

PREREQUISITES

Students should be able to use a stemplots (Unit 2) to aid in ordering data from smallest to largest. Given an ordered set of data, they should be able to compute the median, which was covered in Unit 4, Measures of Center.

ACTIVITY DESCRIPTION

The Unit 5 activity returns to the data collected in the survey for Unit 2's activity. In Unit 2's activity, the sample data, used for the sample solutions, were created rather than collected from a class, with the exception of the height data. If you chose not to collect the data from your class, let students use the sample data from Unit 2 to complete Unit 5's activity.

Students can compare the stemplots that they created for Unit 2's activity to the boxplots from this activity. If students are drawing the boxplots by hand, the stemplots will help them order the data from smallest to largest. Students are asked to complete around 11 boxplots for this activity. So, it may be best to use software to create the boxplots. Otherwise, students should work in small groups so that they can split up the work of constructing the boxplots among group members.

THE VIDEO SOLUTIONS

1. The different brands of hot dogs were compared by their calories.

2. The one-quarter point is called the first quartile.

3. The values in a five-number summary are the minimum, first quartile (Q_1), median, third quartile (Q_3), and maximum.

4. The interquartile range or IQR = $Q_3 - Q_1$.

5. The median of the poultry hot dogs is below the minimum for the beef hot dogs. So, half of the brands of poultry hot dogs have fewer calories than the lowest calorie brand of all-beef hot dogs.

UNIT ACTIVITY:

USING BOXPLOTS TO ANALYZE DATA SOLUTIONS

Unit 5 solutions are based on sample data from Unit 2's activity given in Table T5.1.

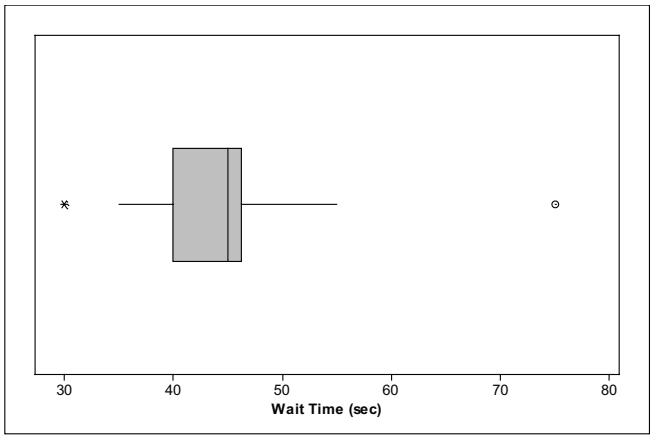
Question 1 Wait (sec)	Question 2 Coins (cents)	Question 3 Height (in)	Question 4 Study (min)	Question 5 Exercise (min)	Question 6 Gender
40	77	68	30	75	Male
40	62	67	60	20	Female
75	175	73	30	0	Male
40	189	72	20	45	Male
50	120	71	15	90	Male
40	54	68	45	75	Female
45	26	66	60	30	Female
45	145	75	30	30	Male
40	0	69	120	0	Female
35	0	71	45	60	Male
45	35	72	30	30	Male
45	47	64	15	45	Female
45	125	72	20	90	Male
45	55	71	30	0	Male
40	35	69	60	45	Male
45	78	63	45	90	Female
55	157	65	45	20	Male
40	225	62	75	40	Female
40	92	64	30	60	Female
50	85	62	60	0	Female
35	35	64	45	30	Female
45	59	66	45	90	Female
50	145	60	30	45	Female
30	137	70	30	30	Male
50	142	69	20	45	Male
45	62	69	30	60	Male

Table T5.1. Sample data for Unit 2 Activity survey questions.

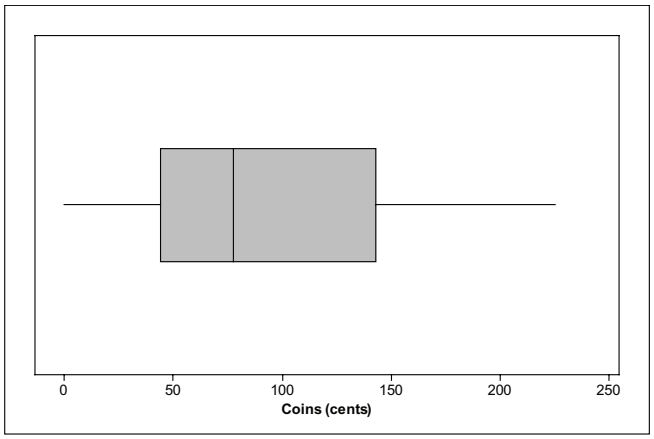
Minitab was used to create the boxplots for the sample answers. Hand drawn boxplots may differ slightly.

1. Sample answers:

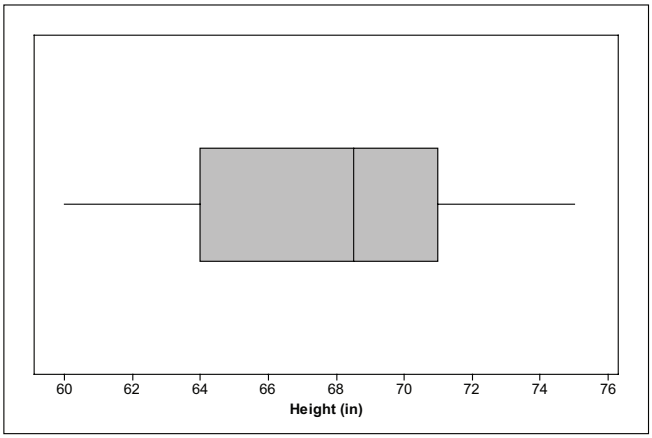
Question 1 – Wait Time: There were two outliers, a mild outlier on the low side and an extreme outlier on the high side. The times in the third quarter of the data are really concentrated compared to the times in the first, second and fourth quarter.



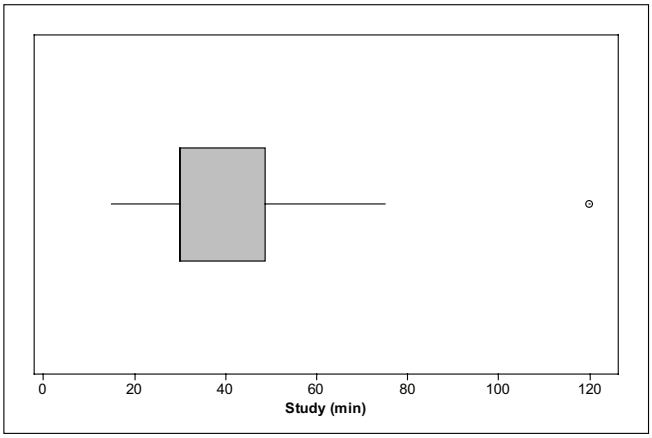
Question 2 – Coins: The amount of money in coins ranged from 0 cents to 225 cents. The data appear to be right skewed.



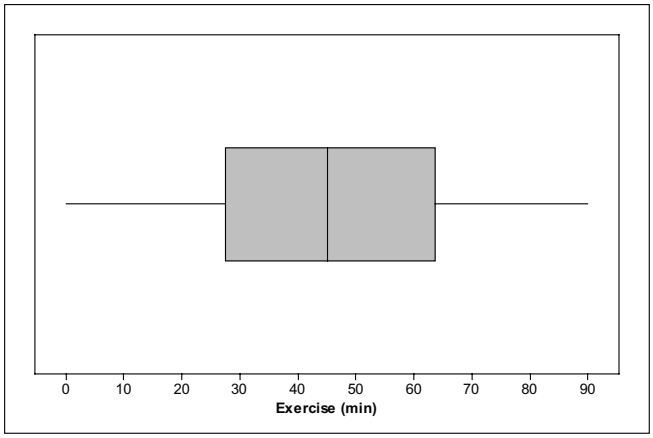
Question 3 – Height: The lower quarter of the heights and the upper quarter of the heights appear to have similar spread. However, the second quarter of the heights appear to be about twice as spread out as the third quarter of the heights.



Question 4 – Study Time: There was one person in class who claimed to study, on average, for 120 minutes per exam. That turned out to be an extreme outlier. However, there doesn't appear to be any dividing line in the box. That is because the first quartile and median were both 30. In fact, there were 9 students who claimed that they studied, on average, 30 minutes for an exam.

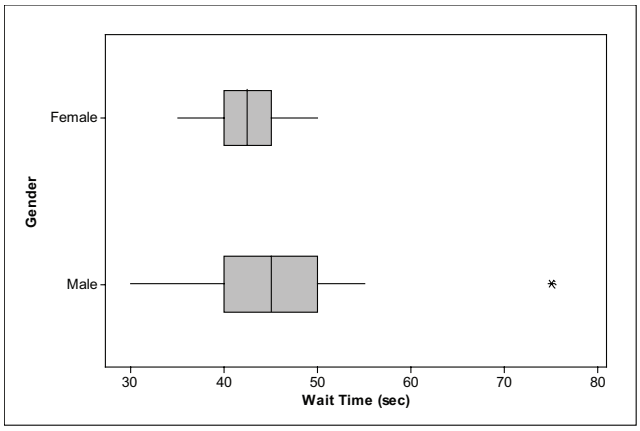


Question 5 – Exercise: The exercise times look pretty symmetric. Even though some people exercised on average for 90 minutes and others did not exercise, neither extreme turned out to be an outlier.

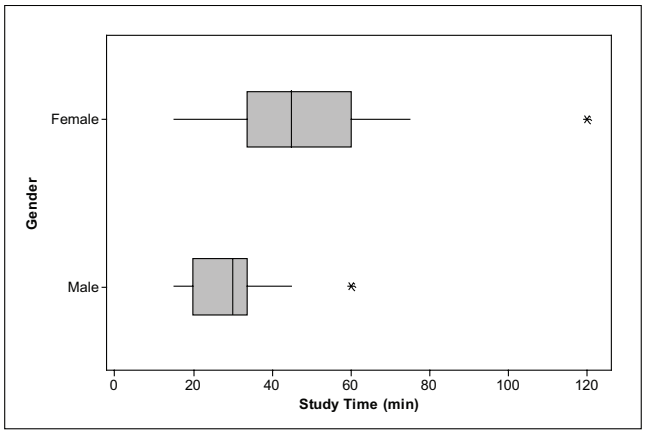


2. Comparative plots for how long female students felt they waited compared to male students are shown below. The outlier for the males turns out to be a mild outlier. The median time for the females was lower than for males. The data for the males were considerably more spread out than for the females, particularly if the range was used to describe the spread.

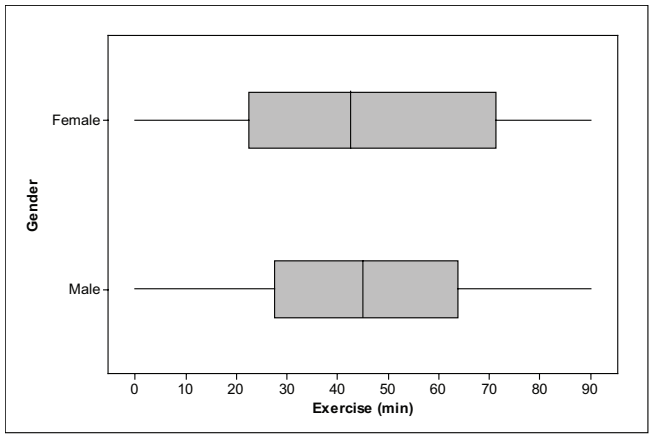
(See boxplot on next page...)



3. The study times for females were more spread out than for males. Using Minitab's algorithm for computing quartiles, the first quartile for the females equaled the third quartile for the males. Therefore, three-quarters of the females spent longer studying than three-quarters of the males. Both of the outliers are mild outliers, even though we had thought that the study time outlier for females would be an extreme outlier. That is most likely due to the larger IQR for the female study times.



4. The overall spread of the two data sets is the same, with a range of 90 minutes. The inner 50% of the data for the males is more concentrated than for females. The median for the males is very slightly higher than for females. Both distributions are roughly symmetric.



EXERCISE SOLUTIONS

1. a. First, the data need to be ordered from smallest to largest:

111	131	132	135	139	141	148	149	149	152
153	157	158	175	176	181	184	186	190	190

The median is computed by averaging the 10th and 11th data values:

median = (152 + 153)/2 = 152.5

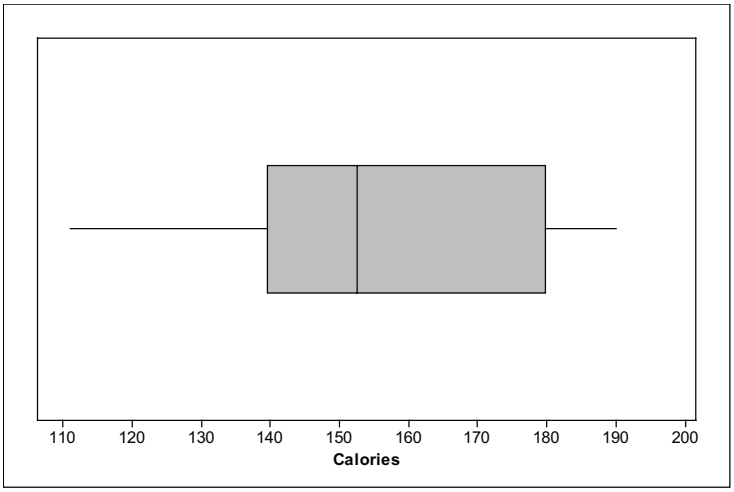
Q_1 is the median of the lower half of the data (top row): Q_1 = (139 + 141)/2 = 140
Q_3 is the median of the upper half of the data (bottom row): Q_3= (176 + 181)/2 = 178.5
5-number summary: 111, 140, 152.5, 178.5, 190

(Note: If quartiles are computed using Minitab, Q_1 = 139.50 and Q_3= 179.75. If Excel is used to compute the quartiles, Q_1= 140.5 and Q_3= 177.25. In all three cases, 5 data values or 25% of the data fall below Q_1. Similarly, 75% of the data fall below Q_3, regardless of whether Minitab's, Excel's or the hand-calculated value of is used.)

b. range = 190 – 111 = 79; IQR = 178.5 – 140 = 38.5. The range gives the spread between the minimum and maximum data value. The IQR tells how spread out the middle half of the data are.

c. No; 175 is below Q_3, the cutoff for the top quarter of the data.

2. a.



b. The second quarter (represented by left section of box) and the fourth quarter (represented by right whisker) appear close in length. Data values in the second quarter lie between 140 and 152.5, a distance of 12.5; data values in the fourth quarter lie between 178.5 and 190, a distance of 11.5. So, the data in the fourth quarter are the most concentrated.

c. The first quarter (represented by the left whisker) and third quarter (represented by right section of box) appear equally spread. Data values in the first quarter lie between 111 and 140, a spread of 29; data values in the third quarter lie between 152.5 and 178.5, a spread of 26. Hence, data in the first quarter exhibit the most spread.

d. $1.5 \times 38.5 = 57.57$; $Q_1 - 57.75 = 82.25$ and $Q_3 + 57.57 = 236.25$. None of the beef hot dogs in the sample had calories below 82.25 or above 236.25. Therefore, there are no outliers.

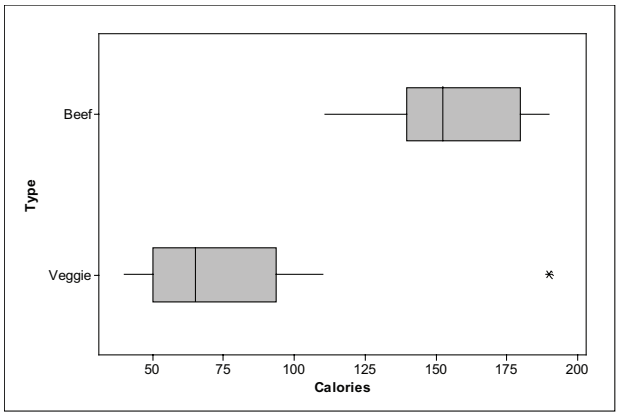
3. The stemplot appears below. Notice that there was one brand of beef hot dog that had low calories compared to the other brands. According to the 1.5 × IQR rule, this value was not sufficiently small compared to the rest of the data to be classified as an outlier. Ignoring that value, it appears that the beef hot dogs fall into two categories separated by a gap. The lower-calorie hot dogs have between 131 and 158 calories; the higher-calorie hot dogs have between 175 and 190 calories. This gap that appears to divide the beef hot dogs into two categories on the stemplot is not visible in the boxplot.

11	1
12	
13	1259
14	1899
15	2378
16	
17	56
18	146
19	00

4. a. Five-number summary: 40, 50, 65, 92.5, 190

b. For the veggie dogs: $IQR = Q_3 - Q_1 = -42.5$; step $= 1.5 \times 42.5 = 63.75$
Inner upper fence: $Q_3 + 1$ step $= 92.5 + 63.75 = 156.25$
Outer upper fence: $Q_3 + 2$ steps $= 220$
Hence, 190 is a mild outlier.
See (c) for the modified boxplot. (The upper whisker ends at 110.)

c. Boxplots comparing calories of beef and veggie dogs.



d. As a group, the veggie dogs appear to have fewer calories than the beef dogs. The third quartile (right end of the box) for the veggie dogs is below the minimum calories for the beef dogs. Hence, at least three-quarters of the veggie dogs have fewer calories than the beef dogs. However, there is one veggie dog that has the same number of calories as the highest calorie beef hot dog. So, if you are trying to limit calories, you need to read the label to make sure you are getting a low-calorie veggie dog.

5. a. Ordered career home runs data is shown below. The 26th, 52th, 53rd, and 79th positions have been highlighted.

13	18	24	27	27	33	33	34	36	37
38	39	40	41	42	42	45	46	47	48
49	51	52	53	54	58	58	62	63	64
65	65	67	68	69	69	75	80	83	83
83	92	93	96	97	101	101	102	102	103
105	106	106	106	106	110	113	116	117	117
118	119	127	128	132	135	136	137	138	154
164	181	183	184	196	202	205	207	219	227
229	238	240	248	254	300	300	301	307	309
312	331	354	359	361	383	449	474	475	493
521	534	555	714						

b. Five-number summary: 13, 58, 106, 219, 714
(Keep in mind that statistical software may use a different algorithm for computing the first and third quartiles. For example, Minitab gives $Q_3 = 207 + 0.75(219 – 207) = 216$.)

c. Calculations of fences:

IQR = 219 – 58 = 161

step = (1.5)(161) = 241.5

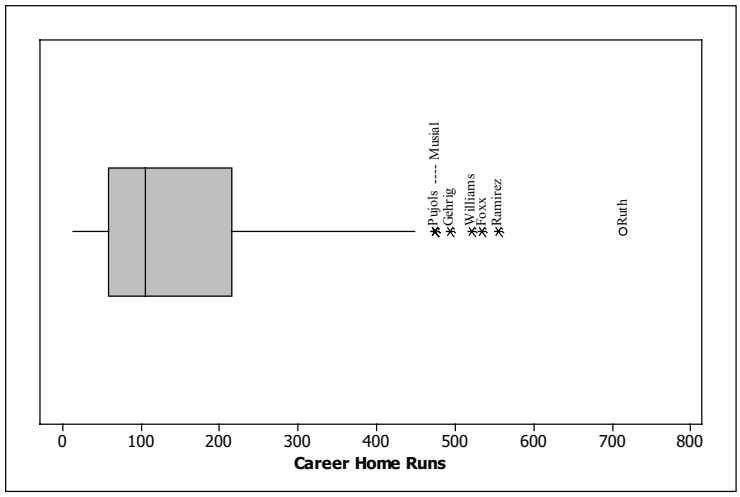
upper inner fence: 219 + 241.5 = 460.5

upper outer fence: 219 + 2(241.5) = 702

Mild outliers: 474 475 493 521 534 555

Extreme outlier: 714

d. The distribution of career home runs is skewed to the right. The right tail (fourth quarter) of the data is much more spread out than the left tail (first quarter) of the data.



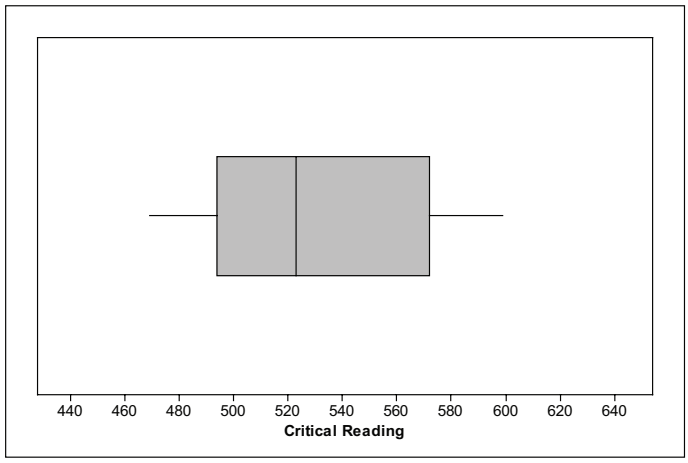
REVIEW QUESTIONS SOLUTIONS

1. a. 469, 494, 523, 572, 599

b. California's average SAT Critical Reading score does not fall in the top half of the states' average Critical Reading scores because 499 is below the median score of 523. It does fall above the bottom quarter because 499 is greater than the first quartile, which is 494.

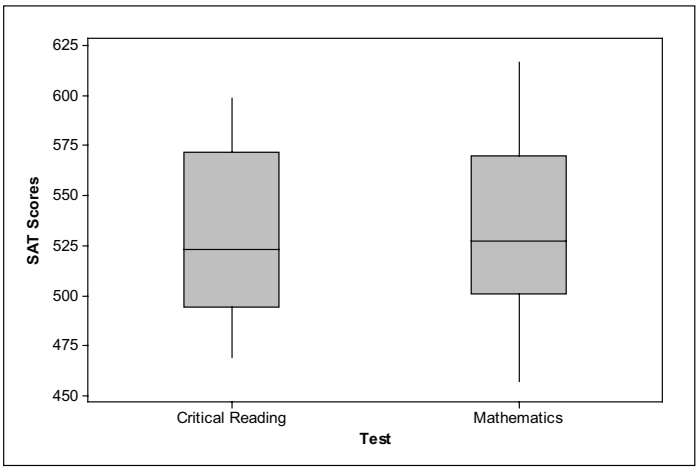
c. About 25% of the states; 12 states. That is because Wyoming's score is the third quartile.

d. In the boxplot below, the third quarter of the data appears to be more spread out than the other quarters of the data.



2. a. 457, 501, 527, 570, 617

b. Comparative boxplots appear below. (Note: Boxplots can be oriented either horizontally or vertically.)

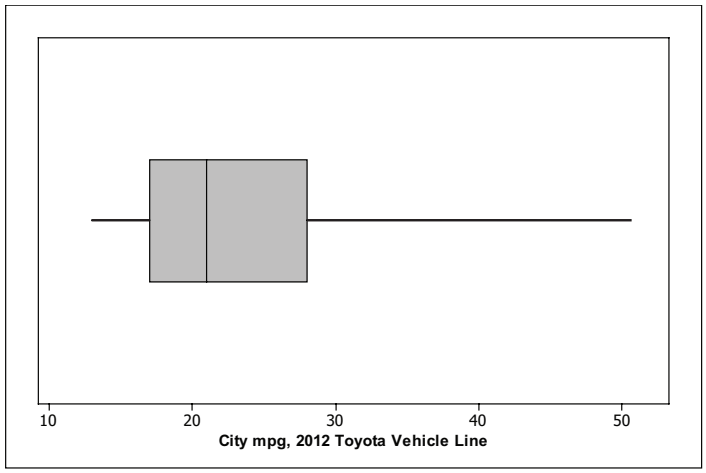


c. Sample answer: The states' average SAT Math scores are centered slightly higher than the SAT Critical Reading scores; the median for the Critical Reading scores is 523 and the median for the Math scores is 527. The middle half of the Critical Reading scores data is more spread out than the middle half of the Math scores; the IQR for Critical Reading scores is 78 compared to only 69 for the Math scores.

However, based on the length of the whiskers, the first and fourth quarters of the Math scores are more spread out than the first and fourth quarters of the Critical Reading scores. The length of the whiskers contributed to the range of the Math scores, which was 30 points higher than the range of the Critical Reading scores (range Math scores = 160 and range of Critical Reading scores = 130).

3. a. 13, 17 ,21, 28 , 51

b.

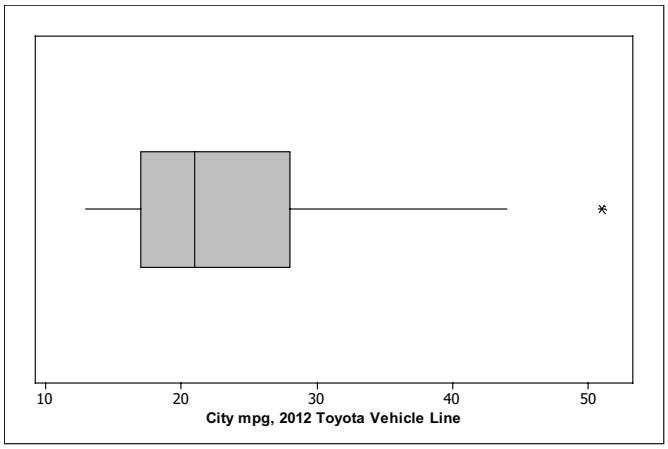


c. There appear to be three potential outliers: 43, 44, 51.

d. Calculations for fences:

IQR = 28 – 17 = 11; step = (1.5)(11) = 16.5
Lower inner fence = 17 – 16.5 = 0.5; no data values lie below this fence.
Upper inner fence = 28 + 16.5 = 44.5
Upper outer fence = 28 + 2(16.5) = 61

Only one data value, 51, is an outlier and it is a mild outlier because it falls between the two upper fences (see below).


10
20
30
40
50
City mpg, 2012 Toyota Vehicle Line