

## Chapter 12

# Describing One Numerical Response

### 12.1 Study Suggestions

Chapter 12 stresses the interpretation, rather than the construction and computation, of the visual presentation and numerical summaries of numerical data.

The text suggests a variety of methods for data presentation. The dot plot provides the most faithful presentation of the data because it presents exact data values. Every other method groups data into class intervals. The rationale is that sometimes less detail gives a picture that is easier to interpret and understand. Remember that there is no universal single *best* way to present numerical data. Moreover, for any given set of data, different presentations may have different desirable features.

In many situations it is necessary or desirable to have one or more numerical summaries of a set of data. The two popular measures of center are the mean and median. Resist the temptation to seek rules for deciding whether the mean or median is the better measure of center. Instead, it is important to understand how the mean and median differ. Knowing the mean is equivalent to knowing the total, whereas knowing the median is not. Hence, if the total is important, you probably will be unsatisfied with using the median alone as the measure of center. An example of the distinction is provided by a hypothetical example of the salaries of professional athletes, for example, baseball players. A list of major league baseball players' annual salaries for a given year will be skewed to the right with the mean substantially larger than the median. From the players' point of view the median is a good measure of center because it represents the salary of a "typical" player; about one-half of the players earn more and about one-half

of the players earn less than the median. The owners, however, are concerned with total salary outlay, so the mean is a good measure of center. (This argument is simplified. At the time of this writing U.S. tax law provides tax breaks for employers of highly paid employees. As a result, the owners achieve a tax savings from having very expensive players. Thus, the real cost to an owner of players is a bit more complicated than the total of the salaries.)

The value of the mean is sensitive to one or more outliers, but the median is not.

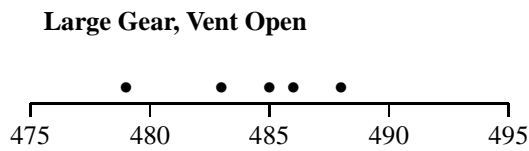
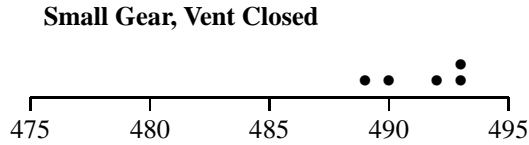
Among the measures of spread, usually the interquartile range is matched with the median and the standard deviation is matched with the mean. The standard deviation is sensitive to outliers and the interquartile range is not. The standard deviation is very important for the theoretical developments of Chapters 13, 15, and 16, while the interquartile range has virtually no theoretical use.

Statisticians have different methods for computing quartiles. I have selected a method that is perhaps conceptually the simplest; namely, the first and third quartiles are the medians of the lower and upper halves of the data set, respectively. Note that my method for computing quartiles is not the method used by Minitab.

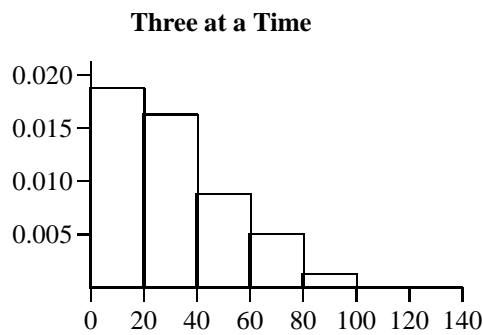
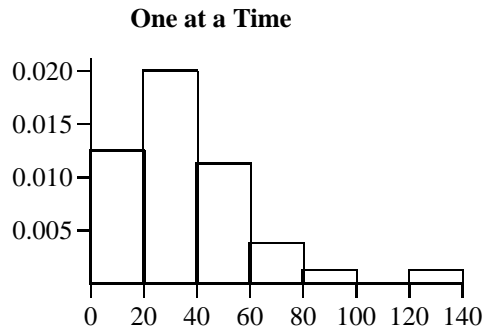
## 12.2 Solutions to Odd-Numbered Exercises

### Solutions for Sections 12.2 and 12.3

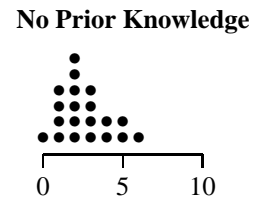
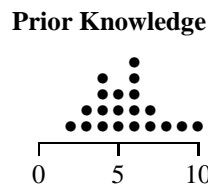
5. (a) The dot plots are below.



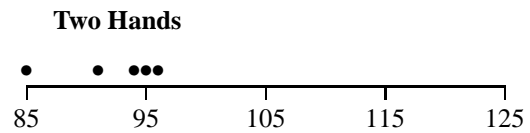
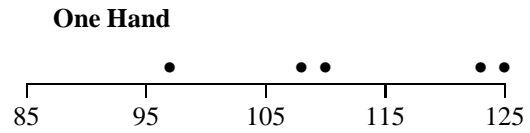
11. (a) The density scale histograms are below.



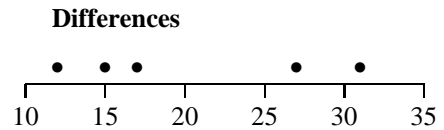
17. (a) The dot plots are below.



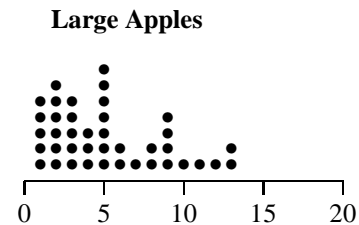
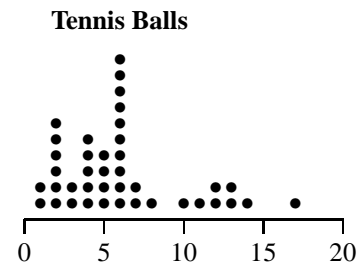
19. (a) The dot plots are below.



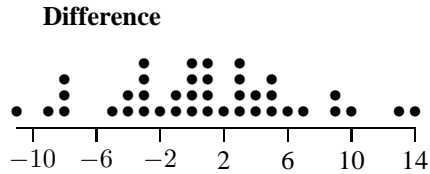
- (b) The dot plot of the differences, the score with one hand minus the score with two hands, is below.



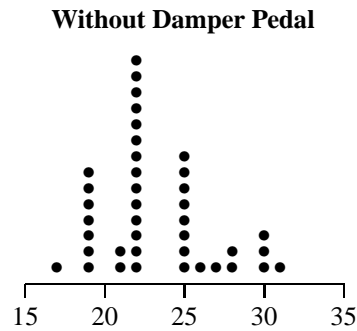
21. (a) The dot plots are below.



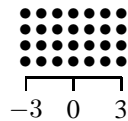
- (b) The dot plot of the differences, the time juggling tennis balls minus the time juggling large apples, is below.



23. (a) The dot plot for the data without the damper pedal is below.



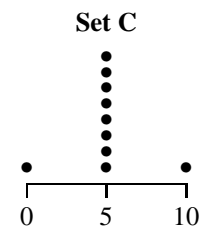
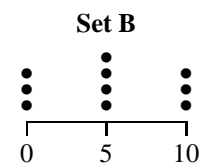
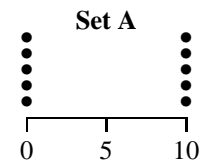
31. (a) The dot plot is below.



**Solutions for Section 12.4.**

1. (a) For the combat boots,  $R = 347 - 321 = 26$  and  $IQR = 337 - 329 = 8$ . For the jungle boots,  $R = 327 - 301 = 26$  and  $IQR = 327 - 316 = 11$ .
3. (a) For the aluminum bat,  $R = 350 - 105 = 245$ , and  $IQR = 187.5 - 141.5 = 46$ . For the wooden bat,  $R = 200 - 86 = 114$ , and  $IQR = 165 - 133 = 32$ .
- (c) After the specified outliers are deleted, for the aluminum bat,  $R = 260 - 105 = 155$  and  $IQR = 186 - 141 = 45$ ; for the wooden bat,  $R = 200 - 114 = 86$  and  $IQR = 166 - 134 = 32$ . The deletions have a big impact on the ranges and little (aluminum) and none (wooden) on the IQRs.
5. (a) For the small gear,  $R = 493 - 489 = 4$  and  $IQR = 493 - 489.5 = 3.5$ ; for the large gear,  $R = 488 - 479 = 9$  and  $IQR = 487 - 481 = 6$ .

7. (a) With no looking allowed,  $R = 164 - 111 = 53$  and  $IQR = 144.5 - 124.5 = 20$ ; with no restrictions,  $R = 177 - 131 = 46$  and  $IQR = 164.5 - 144 = 20.5$ .
9. (a) With no looking allowed,  $R = 120 - 70 = 50$  and  $IQR = 111 - 86.5 = 24.5$ . With no restrictions,  $R = 161 - 105 = 56$  and  $IQR = 154 - 133 = 21$ .
11. (a) For throwing darts one at a time,  $R = 124 - 2 = 122$  and  $IQR = 48 - 20 = 28$ . For throwing darts three at a time,  $R = 85 - 0 = 85$  and  $IQR = 46 - 15 = 31$ .
13. (a) With the 3-wood,  $R = 147 - 22 = 125$  and  $IQR = 128 - 101 = 27$ . With the 3-iron,  $R = 139 - 27 = 112$  and  $IQR = 117 - 83 = 34$ .
15. (a) With the windows open,  $R = 8.34 - 7.34 = 1.00$  and  $IQR = 8.17 - 7.67 = 0.50$ . With the windows closed,  $R = 7.86 - 6.67 = 1.19$  and  $IQR = 7.34 - 6.86 = 0.48$ .
19. (a) The dot plots are below.



- (d) Yes. The standard deviation indicates that data set A has the most spread of the three and data set C has the least. This ranking agrees with our visual assessment of the spread.

21. (a) The proportion of observations between 900 and 1,200 dollars is  $77/152 = 0.507$ .  
 (b) The standardized value of 1,200 is

$$(1,200 - 1,068.5)/276.8 = 0.48,$$

and the standardized value of 900 is

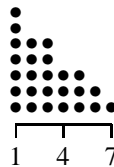
$$(900 - 1,068.5)/276.8 = -0.61.$$

The area under the standard normal curve between  $-0.61$  and  $0.48$  is  $0.7291 - 0.3156 = 0.4135$ . The approximation is poor because the distribution is more peaked in the middle than a bell-shaped curve (see Figure 12.21, on page 412 of the text).

23. (a)  $R = 514.0 - 17.3 = 496.7$   
 (b)  $IQR = 97.2 - 56.5 = 40.7$ .  
 (c) After deleting the largest value,  $R = 320.3 - 17.3 = 303.0$  and  $IQR = 96.25 - 56.5 = 39.75$ .

### Solutions for Section 12.5.

1. The dot plot is below.



3. (a) Combining data sets A and B and sorting yields

0 1 3 5 6 7

which has a sample median of 4.

- (b) Combining data sets A and C and sorting yields

1 3 4 5 6 6

which has a sample median of 4.5.

(c) Part (a) shows that one can combine two data sets of size 3 each with medians of 3 and 5 and obtain a median of 4 for the combined data set. But part (b) shows that one can combine two data sets of size 3 each with medians of 3 and 5 and obtain a median of 4.5 for the combined data set. Thus, there cannot be a rule which allows one to use the individual sample sizes and individual medians to obtain the median of the combined data.

5. It is clear that the answer is absurd. A forecast error is  $f - h$  in degrees Fahrenheit. If  $f$  and  $h$  are converted to Celsius, then

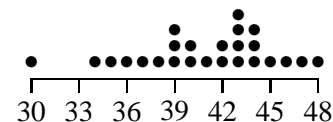
$$f - h = \frac{5}{9}(f - 32) - \frac{5}{9}(h - 32) = \frac{5}{9}(f - h).$$

Thus, the sample mean in degrees Celsius is

$$\frac{5}{9}(-0.712) = -0.396.$$

## 12.3 Exam Questions

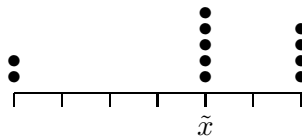
1. A histogram is drawn for a data set.  
 True or false? If the researcher notes that there are no gaps in the histogram, the researcher should conclude that there are no outliers in the data set.
2. Eric plays a round of miniature golf on each of 25 consecutive days. The response is his score, which must be an integer. Below is a dot plot of Eric's data.



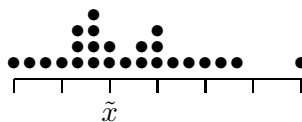
Which score was achieved most often by Eric? How many times was this score achieved?

3. Refer to the dot plot in the previous question. Eric labels one of the observations an outlier. Which one?

4. The sample median is located on the dot plot below. For these data, \_\_\_\_\_.
- (a) the sample mean is larger than the sample median
  - (b) the sample mean equals the sample median
  - (c) the sample mean is smaller than the sample median
  - (d) the relative sizes of the sample mean and sample median cannot be determined without more information



5. The sample median is located on the dot plot below. For these data, \_\_\_\_\_.
- (a) the sample mean is larger than the sample median
  - (b) the sample mean equals the sample median
  - (c) the sample mean is smaller than the sample median
  - (d) the relative sizes of the sample mean and sample median cannot be determined without more information

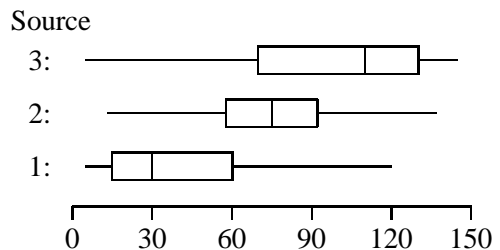


6. Murphy has a data set consisting of 1000 numbers. You are told that the sample mean and sample standard deviation of her numbers are 200 and 50, respectively. You also are told that the distribution of her numbers is approximately bell-shaped. Fill in the blanks to make the statement below correct.
- Approximately 680 of Murphy's numbers lie between \_\_\_\_\_ and \_\_\_\_\_.

7. True or false? For a dot plot with a single peak, the portion of the distribution to the right of the peak is called the right valley of the distribution.
8. Recall the data on lengths of careers of baseball players that were presented in the text. In particular, remember that the sample median was 83.5 games. Kramer states, "I do not believe these data. According to the data one-half of the players had a career of 84 games or longer and one-half had a career of 83 games or less. I have attended many baseball games and nearly all of the players I have seen play had longer careers than 84 games." Which of the following is the best criticism of Kramer's statement?

- (a) 83.5 is the sample median for the data in the textbook; there is no reason to think it would be similar to the sample median for Kramer's experience.
- (b) The games Kramer attended may not be a random sample of baseball games.
- (c) In a random sample of baseball games a player with, for example, a career of 1000 games is more likely to be "seen" than a player with a career of one game.
- (d) Kramer is ignoring the spread in the data in the textbook.

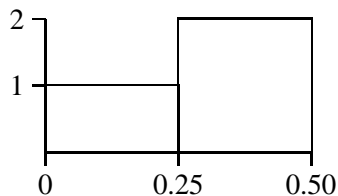
9. Samples of size 100 are obtained from three sources, 1, 2, and 3. Below are their box plots.



Mark each of the following true or false.

- (a) As the source number increases, the minimum value in the sample increases.
- (b) As the source number increases,  $Q_1$  increases.
- (c) As the source number increases,  $\tilde{x}$  increases.

- (d) As the source number increases,  $Q_3$  increases.
- (e) As the source number increases, IQR increases.
- (f) As the source number increases, the maximum value in the sample increases.
- (g) As the source number increases, the sample range increases.
10. Refer to the box plot in the previous exercise. Dot plots reveal that one of these data sets has a symmetric distribution—Which one is it?
- (a) The data from source 1.
- (b) The data from source 2.
- (c) The data from source 3.
11. My dog Casey chewed one of my student's homework papers. On a scrap of paper I found the following part of a histogram.



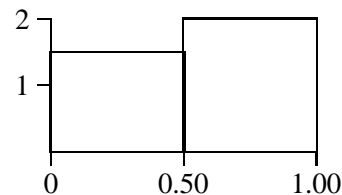
Recall that there are three types of histograms: frequency, relative frequency, and density scale. Answer the three questions below. If you answer “yes,” no explanation of your answer is required. If you answer “no,” however, then to receive any credit for your answer, you must explain why “no” is the correct answer.

- (a) Could this part be from a frequency histogram?
- (b) Could this part be from a relative frequency histogram?
- (c) Could this part be from a density scale histogram?

12. For project number 4, Amy performed a CRD to compare her knitting and purling. A trial consisted of knitting a row of 26 stitches (treatment 1) or purling a row of 26 stitches (treatment 2). The response was the time, to the nearest second, Amy required to complete the trial. Amy's sorted times for knitting are

79 81 84 85 86 86 87 87  
89 91 93 96 100 105 118

- (a) Construct the dot plot of Amy's knitting data.
- (b) Construct the stem and leaf plot of Amy's knitting data that has five leaf values associated with each stem (that is, in the “regular” stem and leaf plot, each stem is split into two).
- (c) Compute the median response for Amy's knitting data.
- (d) Compute the interquartile range for Amy's knitting data.
13. My dog Casey chewed one of my student's homework papers. On a scrap of paper I found the following part of a histogram.



Recall that there are three types of histograms: frequency, relative frequency, and density scale. Answer the three questions on the following page. If you answer “yes,” no explanation of your answer is required. If you answer “no,” however, then to receive any credit for your answer, you must explain why “no” is the correct answer.

- (a) Could this part be from a frequency histogram?

- (b) Could this part be from a relative frequency histogram?
- (c) Could this part be from a density scale histogram?
14. True or false? For a density scale histogram, the area of a rectangle above a class interval is equal to the frequency of the interval.
15. A sample of size  $n = 20$  yields the following data.

21	21	22	22	23
24	25	25	27	27
27	31	33	33	35
37	40	50	52	58

Draw a histogram for these data with class intervals equal to 20–25, 25–35, and 35–60. (Hint: Be careful to choose a histogram that is not misleading.)

16. A sample of size  $n = 20$  yields the following data.

21	21	22	22	23
24	25	25	27	27
27	31	33	33	35
37	40	50	52	58

Present these data in a stem and leaf plot.

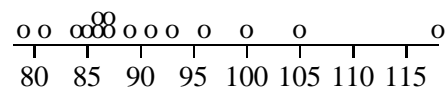
17. Eric plays 18 holes of miniature golf on each of 25 consecutive days. His scores, ordered from smallest to largest, are listed below. (Hint: the sample mean is 40.88, and the sample standard deviation is 4.285.)

30	34	35	36	37
38	39	39	39	40
40	41	42	42	43
43	43	43	44	44
44	45	46	47	48

- (a) Find the proportion of scores between 36 and 42, inclusive.
- (b) Use the normal approximation, without the continuity correction, to approximate the proportion of scores between 36 and 42, inclusive.

## 12.4 Solutions to Exam Questions

1. False.
2. 43; 4.
3. 30.
4. (b) The sample mean equals the sample median.
5. (a) The sample mean is larger than the sample median.
6. 150; 250.
7. False.
8. (c) In a random sample of baseball games a player with, for example, a career of 1000 games is more likely to be “seen” than a player with a career of one game.
9. (a) False.  
(b) True.  
(c) True.  
(d) True.  
(e) False.  
(f) True.  
(g) True.
10. (b) The data from source 2.
11. (a) Yes. Class interval frequencies of 1 and 2 are allowable.  
(b) No. A relative frequency cannot exceed one.  
(c) Yes. The area of the smaller rectangle is 0.25 and the area of the larger rectangle is 0.50. These values are allowable. (Note: When the width of a class interval is smaller than one, as in this case, the height of a density scale rectangle may exceed one.)
12. (a) The dot plot is below.



(b) The stem and leaf plot is below.

```

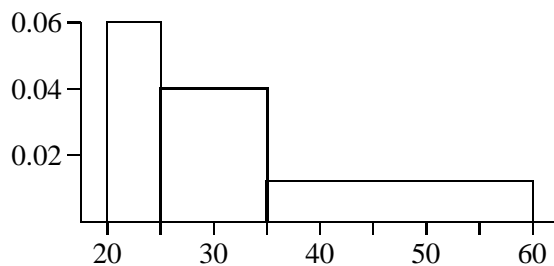
  7 | 9
  8 | 14
  8 | 566779
  9 | 13
  9 | 6
 10 | 0
 10 | 5
 11 |
 11 | 8

```

(c) The median equals 87.

(d)  $Q_1 = 85$ ,  $Q_3 = 96$ , and  $IQR = 96 - 85 = 11$ .

13. (a) No. A frequency may not be a fraction, as in the left rectangle.  
 (b) No. A relative frequency may not exceed one, as in both rectangles.  
 (c) No. The area of the left rectangle is 0.75 and the area of the right rectangle is 1.00, giving a total area of 1.75. The total area under a density scale histogram, however, equals one.
14. False.
15. Because the class intervals do not have a constant width, a density scale histogram is needed.



16. The stem and leaf plot is below.

```

  2 | 11223455777
  3 | 13357
  4 | 0
  5 | 028

```

17. (a)  $11/25 = 0.44$ .

(b)

$$(36 - 40.88)/4.285 = -1.14 \text{ and}$$

$$(42 - 40.88)/4.285 = 0.26.$$

The approximation is the area under the standard normal curve between  $-1.14$  and  $0.26$ ; that is,  $0.4755$ .