Previous Section Table of Contents Next Section

2.2 Measures of Central Location

LEARNING OBJECTIVES

1. To learn the concept of the “center” of a data set.
2. To learn the meaning of each of three measures of the center of a data set—the mean, the median, and the mode—and how to compute each one.

This section could be titled “three kinds of averages of a data set.” Any kind of “average” is meant to be an answer to the question “Where do the data center?” It is thus a measure of the central location of the data set. We will see that the nature of the data set, as indicated by a relative frequency histogram, will determine what constitutes a good answer. Different shapes of the histogram call for different measures of central location.

The Mean

The first measure of central location is the usual “average” that is familiar to everyone. In the formula in the following definition we introduce the standard summation notation Σ, where Σ is the capital Greek letter sigma. In general, the notation Σ followed by a second mathematical symbol means to add up all the values that the second symbol can take in the context of the problem. Here is an example to illustrate this.

EXAMPLE 1

Find Σx, Σx^2, and $\Sigma (x-1)^2$ for the data set

$$1 \quad 3 \quad 4$$

Solution:

$$\begin{aligned} \Sigma x &= 1+3+4=8 \\ \Sigma x^2 &= 1^2+3^2+4^2=1+9+16=26 \\ \Sigma(x-1)^2 &= (1-1)^2+(3-1)^2+(4-1)^2=0^2+2^2+3^2=13 \end{aligned}$$

In the definition we follow the convention of using lowercase *n* to denote the number of measurements in a sample, which is called the **sample size**.

Definition

The ***sample mean*** *of a set of n sample data is the number* $\bar{x}$ *defined by the formula*

$$\bar{x} = \frac{\Sigma x}{n}$$

EXAMPLE 2

Find the mean of the sample data

$$2 \quad -1 \quad 0 \quad 2$$

Solution:

$$\bar{x} = \frac{\Sigma\, x}{n} = \frac{2 + (-1) + 0 + 2}{4} = \frac{3}{4} = 0.75$$

EXAMPLE 3

A random sample of ten students is taken from the student body of a college and their GPAs are recorded as follows.

$$1.90 \quad 3.00 \quad 2.53 \quad 3.71 \quad 2.12 \quad 1.76 \quad 2.71 \quad 1.39 \quad 4.00 \quad 3.33$$

Find the sample mean.

Solution:

$$\begin{aligned} \bar{x} &= \frac{\Sigma\, x}{n} = \frac{1.90+3.00+2.53+3.71+2.12+1.76+2.71+1.39+4.00+3.33}{10} \\ &= \frac{26.45}{10} = 2.645 \end{aligned}$$

EXAMPLE 4

A random sample of 19 women beyond child-bearing age gave the following data, where *x* is the number of children and *f* is the *frequency* of that value, the number of times it occurred in the data set.

x	0	1	2	3	4
f	3	6	6	3	1

Find the sample mean.

Solution:

In this example the data are presented by means of a data frequency table, introduced in Chapter 1 "Introduction". Each number in the first line of the table is a number that appears in the data set; the number below it is how many times it occurs. Thus the value 0 is observed three times, that is, three of the measurements in the data set are 0, the value 1 is observed six times, and so on. In the context of the problem this means that three women in the sample have had no children, six have had exactly one child, and so on. The explicit list of all the observations in this data set is therefore

$$0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 4$$

The sample size can be read directly from the table, without first listing the entire data set, as the sum of the frequencies: $n = 3 + 6 + 6 + 3 + 1 = 19$. The sample mean can be computed directly from the table as well:

$$\overline{x} = \frac{\Sigma x}{n} = \frac{0 \times 3 + 1 \times 6 + 2 \times 6 + 3 \times 3 + 4 \times 1}{19} = \frac{31}{19} = 1.6316$$

In the examples above the data sets were described as samples. Therefore the means were sample means, denoted by $\overline{x}$. If the data come from a census, so that there is a measurement for every element of the population, then the mean is calculated by exactly the same process of summing all the measurements and dividing by how many of them there are, but it is now the *population mean* and is denoted by μ, the lower case Greek letter mu.

Definition

The **population mean** *of a set of N population data is the number μ defined by the formula*

$$\mu = \frac{\Sigma x}{N}$$

The mean of two numbers is the number that is halfway between them. For example, the average of the numbers 5 and 17 is (5 + 17) / 2 = 11, which is 6 units above 5 and 6 units below 17. In this sense the average 11 is the “center” of the data set {5,17}. For larger data sets the mean can similarly be regarded as the “center” of the data.

The Median

To see why another concept of average is needed, consider the following situation. Suppose we are interested in the average yearly income of employees at a large corporation. We take a random sample of seven employees, obtaining the sample data (rounded to the nearest hundred dollars, and expressed in thousands of dollars).

$$24.8 \quad 22.8 \quad 24.6 \quad 192.5 \quad 25.2 \quad 18.5 \quad 23.7$$

The mean (rounded to one decimal place) is $\overline{x} = 47.4$, but the statement “the average income of employees at this corporation is $47,400” is surely misleading. It is approximately twice what six of the seven employees in the sample make and is nowhere near what any of them makes. It is easy to see what went wrong: the presence of the one executive in the sample, whose salary is so large compared to everyone else’s, caused the numerator in the formula for the sample mean to be far too large, pulling the mean far to the right of where we think that the average “ought” to be, namely around $24,000 or $25,000. The number 192.5 in our data set is called an **outlier**, a number that is far removed from most or all of the remaining measurements. Many times an outlier is the result of some sort of error, but not always, as is the case here. We would get a better measure of the “center” of the data if we were to arrange the data in numerical order,

$$18.5 \;\; 22.8 \;\; 23.7 \;\; 24.6 \;\; 24.8 \;\; 25.2 \;\; 192.5$$

then select the middle number in the list, in this case 24.6. The result is called the *median* of the data set, and has the property that roughly half of the measurements are larger than it is, and roughly half are smaller. In this sense it locates the center of the data. If there are an even number of measurements in the data set, then there will be two middle elements when all are lined up in order, so we take the mean of the middle two as the median. Thus we have the following definition.

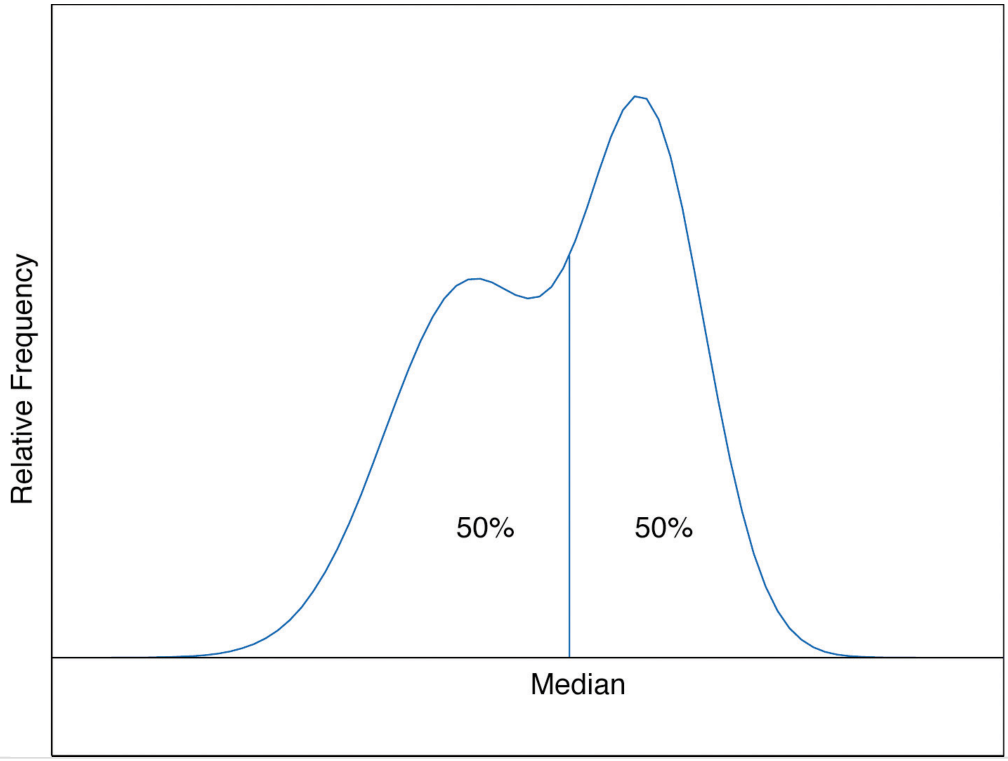
Definition

The **sample median** *$\tilde{x}$ of a set of sample data for which there are an odd number of measurements is the middle measurement when the data are arranged in numerical order. The sample median $\tilde{x}$ of a set of sample data for which there are an even number of measurements is the mean of the two middle measurements when the data are arranged in numerical order.*

The population median is defined in a similar way, but we will not have occasion to refer to it again in this text.

The median is a value that divides the observations in a data set so that 50% of the data are on its left and the other 50% on its right. In accordance with Figure 2.6 "A Very Fine Relative Frequency Histogram", therefore, in the curve that represents the distribution of the data, a vertical line drawn at the median divides the area in two, area 0.5 (50% of the total area 1) to the left and area 0.5 (50% of the total area 1) to the right, as shown in Figure 2.7 "The Median". In our income example the median, $24,600, clearly gave a much better measure of the middle of the data set than did the mean $47,400. This is typical for situations in which the distribution is skewed. (Skewness and symmetry of distributions are discussed at the end of this subsection.)

Figure 2.7 The Median



EXAMPLE 5

Compute the sample median for the data of Note 2.11 "Example 2".

Solution:

The data in numerical order are −1, 0, 2, 2. The two middle measurements are 0 and 2, so $\tilde{x} = (0 + 2)/2 = 1$.

EXAMPLE 6

Compute the sample median for the data of Note 2.12 "Example 3".

Solution:

The data in numerical order are

$$1.39 \quad 1.76 \quad 1.90 \quad 2.12 \quad 2.53 \quad 2.71 \quad 3.00 \quad 3.33 \quad 3.71 \quad 4.00$$

The number of observations is ten, which is even, so there are two middle measurements, the fifth and sixth, which are 2.53 and 2.71. Therefore the median of these data is $\tilde{x} = (2.53 + 2.71)/2 = 2.62$.

EXAMPLE 7

Compute the sample median for the data of Note 2.13 "Example 4".

Solution:

The data in numerical order are

$$0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 4$$

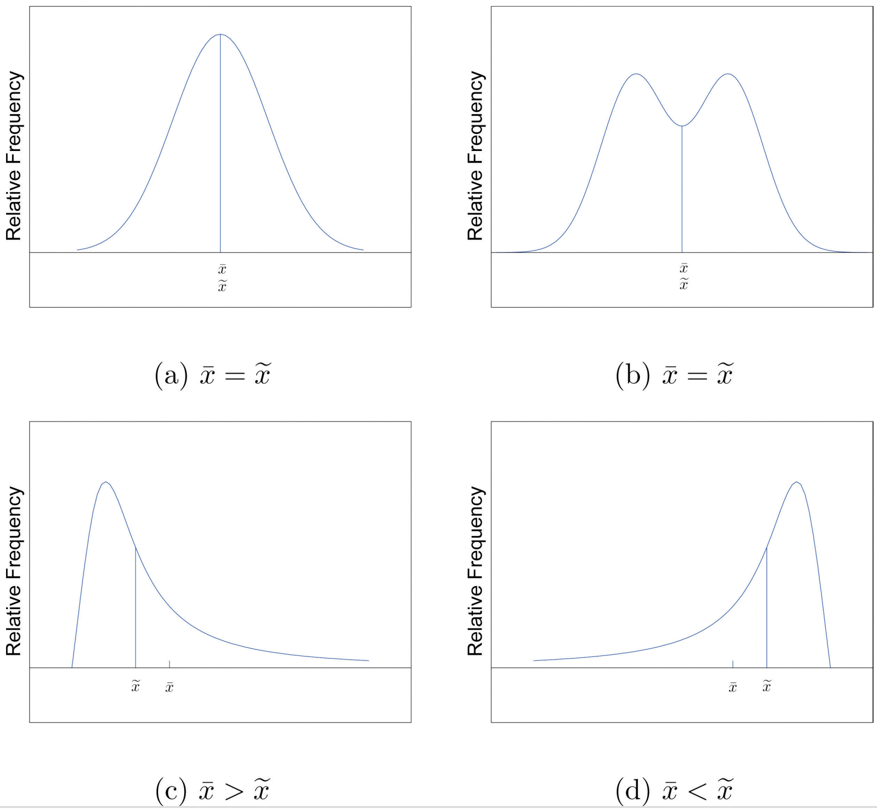
The number of observations is 19, which is odd, so there is one middle measurement, the tenth. Since the tenth measurement is 2, the median is $\tilde{x} = 2$.

It is important to note that we could have computed the median without first explicitly listing all the observations in the data set. We already saw in Note 2.13 "Example 4" how to find the number of observations directly from the frequencies listed in the table: $n = 3 + 6 + 6 + 3 + 1 = 19$. As just above we figure out that the median is the tenth observation. The second line of the table in Note 2.13 "Example 4" shows that when the data are listed in order there will be three 0s followed by six 1s, so the tenth observation is a 2. The median is therefore 2.

The relationship between the mean and the median for several common shapes of distributions is shown in Figure 2.8 "Skewness of Relative Frequency Histograms". The distributions in panels (a) and (b) are said to be *symmetric* because of the symmetry that they exhibit. The distributions in the remaining two panels are said to be *skewed*. In each distribution we have drawn a vertical line that divides the area under the curve in half, which in accordance with Figure 2.7 "The Median" is located at the median. The following facts are true in general:

a. When the distribution is symmetric, as in panels (a) and (b) of Figure 2.8 "Skewness of Relative Frequency Histograms", the mean and the median are equal.
b. When the distribution is as shown in panel (c) of Figure 2.8 "Skewness of Relative Frequency Histograms", it is said to be *skewed right*. The mean has been pulled to the right of the median by the long “right tail” of the distribution, the few relatively large data values.
c. When the distribution is as shown in panel (d) of Figure 2.8 "Skewness of Relative Frequency Histograms", it is said to be *skewed left*. The mean has been pulled to the left of the median by the long “left tail” of the distribution, the few relatively small data values.

Figure 2.8 *Skewness of Relative Frequency Histograms*



(a) $\bar{x} = \widetilde{x}$

(b) $\bar{x} = \widetilde{x}$

(c) $\bar{x} > \widetilde{x}$

(d) $\bar{x} < \widetilde{x}$

The Mode

Perhaps you have heard a statement like “The average number of automobiles owned by households in the United States is 1.37,” and have been amused at the thought of a fraction of an automobile sitting in a driveway. In such a context the following measure for central location might make more sense.

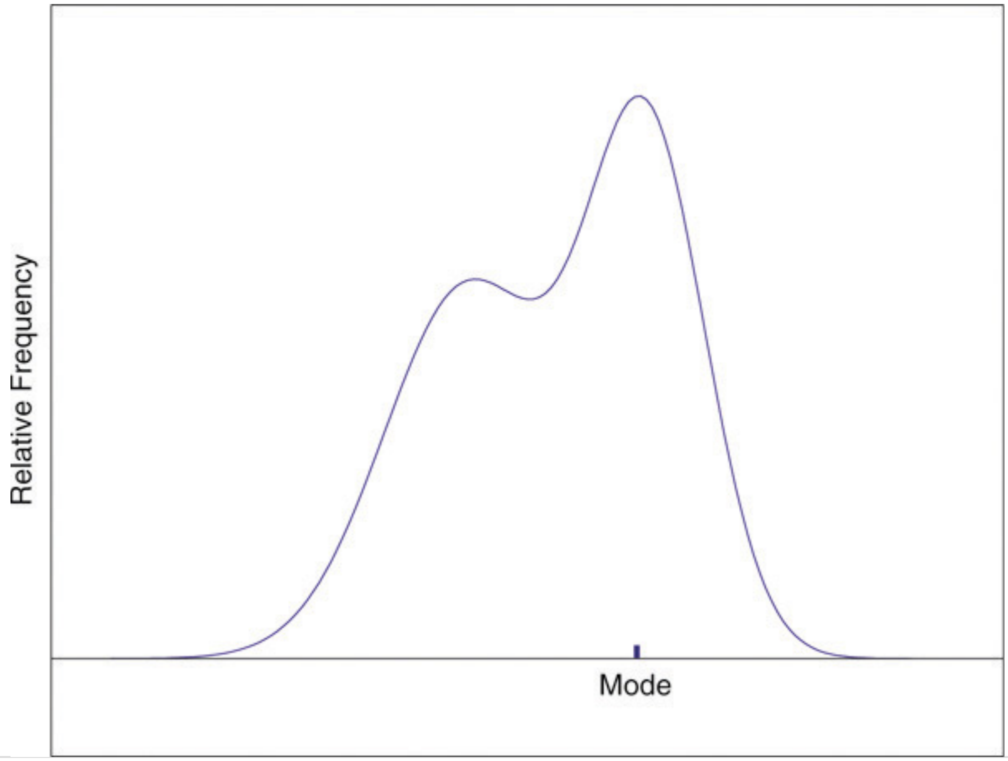
Definition

The **sample mode** *of a set of sample data is the most frequently occurring value.*

The population mode is defined in a similar way, but we will not have occasion to refer to it again in this text.

On a relative frequency histogram, the highest point of the histogram corresponds to the mode of the data set. Figure 2.9 "Mode" illustrates the mode.

Figure 2.9 Mode



For any data set there is always exactly one mean and exactly one median. This need not be true of the mode; several different values could occur with the highest frequency, as we will see. It could even happen that every value occurs with the same frequency, in which case the concept of the mode does not make much sense.

EXAMPLE 8

Find the mode of the following data set.

$$-1 \quad 0 \quad 2 \quad 0$$

Solution:

The value 0 is most frequently observed and therefore the mode is 0.

EXAMPLE 9

Compute the sample mode for the data of Note 2.13 "Example 4".

Solution:

The two most frequently observed values in the data set are 1 and 2. Therefore mode is a set of two values: {1,2}.

The mode is a measure of central location since most real-life data sets have more observations near the center of the data range and fewer observations on the lower and upper ends. The value with the highest frequency is often in the middle of the data range.

KEY TAKEAWAY

The mean, the median, and the mode each answer the question “Where is the center of the data set?” The nature of the data set, as indicated by a relative frequency histogram, determines which one gives the best answer.

EXERCISES

BASIC

1. For the sample data set $\{1,2,6\}$ find

   a. Σx
   b. Σx^2
   c. $\Sigma(x-3)$
   d. $\Sigma(x-3)^2$

2. For the sample data set $\{-1,0,1,4\}$ find

   a. Σx
   b. Σx^2
   c. $\Sigma(x-1)$
   d. $\Sigma(x-1)^2$

3. Find the mean, the median, and the mode for the sample

$$1 \quad 2 \quad 3 \quad 4$$

4. Find the mean, the median, and the mode for the sample

$$3 \quad 3 \quad 4 \quad 4$$

5. Find the mean, the median, and the mode for the sample

$$2 \quad 1 \quad 2 \quad 7$$

6. Find the mean, the median, and the mode for the sample

$$-1 \quad 0 \quad 1 \quad 4 \quad 1 \quad 1$$

7. Find the mean, the median, and the mode for the sample data represented by the table

x	1	2	7
f	1	2	1

8. Find the mean, the median, and the mode for the sample data represented by the table

x	−1	0	1	4
f	1	1	3	1

9. Create a sample data set of size $n = 3$ for which the mean $\bar{x}$ is greater than the median $\tilde{x}$.

10. Create a sample data set of size $n = 3$ for which the mean $\bar{x}$ is less than the median $\tilde{x}$.

11. Create a sample data set of size $n = 4$ for which the mean $\bar{x}$, the median $\tilde{x}$, and the mode are all identical.

12. Create a data set of size $n = 4$ for which the median $\tilde{x}$ and the mode are identical but the mean $\bar{x}$ is different.

APPLICATIONS

13. Find the mean and the median for the LDL cholesterol level in a sample of ten heart patients.

$$\begin{array}{ccccc} 132 & 162 & 133 & 145 & 148 \\ 139 & 147 & 160 & 150 & 153 \end{array}$$

14. Find the mean and the median, for the LDL cholesterol level in a sample of ten heart patients on a special diet.

$$\begin{array}{ccccc} 127 & 152 & 138 & 110 & 152 \\ 113 & 131 & 148 & 135 & 158 \end{array}$$

15. Find the mean, the median, and the mode for the number of vehicles owned in a survey of 52 households.

x	0	1	2	3	4	5	6	7
f	2	12	15	11	6	3	1	2

16. The number of passengers in each of 120 randomly observed vehicles during morning rush hour was recorded, with the following results.

x	1	2	3	4	5
f	84	29	3	3	1

Find the mean, the median, and the mode of this data set.

17. Twenty-five 1-lb boxes of 16d nails were randomly selected and the number of nails in each box was counted, with the following results.

x	47	48	49	50	51
f	1	3	18	2	1

Find the mean, the median, and the mode of this data set.

ADDITIONAL EXERCISES

18. Five laboratory mice with thymus leukemia are observed for a predetermined period of 500 days. After 500 days, four mice have died but the fifth one survives. The recorded survival times for the five mice are

$$493 \quad 421 \quad 222 \quad 378 \quad 500^*$$

where 500^* indicates that the fifth mouse survived for at least 500 days but the survival time (i.e., the exact value of the observation) is unknown.

    a. Can you find the sample mean for the data set? If so, find it. If not, why not?
    b. Can you find the sample median for the data set? If so, find it. If not, why not?

19. Five laboratory mice with thymus leukemia are observed for a predetermined period of 500 days. After 450 days, three mice have died, and one of the remaining mice is sacrificed for analysis. By the end of the observational period, the last remaining mouse still survives. The recorded survival times for the five mice are

$$222 \quad 421 \quad 378 \quad 450^* \quad 500^*$$

where * indicates that the mouse survived for at least the given number of days but the exact value of the observation is unknown.

    a. Can you find the sample mean for the data set? If so, find it. If not, explain why not.
    b. Can you find the sample median for the data set? If so, find it. If not, explain why not.

20. A player keeps track of all the rolls of a pair of dice when playing a board game and obtains the following data.

x	2	3	4	5	6	7	8	9	10	11	12
f	10	29	40	56	68	77	67	55	39	28	11

Find the mean, the median, and the mode.

21. Cordelia records her daily commute time to work each day, to the nearest minute, for two months, and obtains the following data.

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

a. Based on the frequencies, do you expect the mean and the median to be about the same or markedly different, and why?

b. Compute the mean, the median, and the mode.

22. An ordered stem and leaf diagram gives the scores of 71 students on an exam.

10	0 0
9	1 1 1 1 2 3
8	0 1 1 2 2 3 4 5 7 8 8 9
7	0 0 0 1 1 2 4 4 5 6 6 6 7 7 7 8 8 9
6	0 1 2 2 2 3 4 4 5 7 7 7 7 8 8
5	0 2 3 3 4 4 6 7 7 8 9
4	2 5 6 8 8
3	9 9

a. Based on the shape of the display, do you expect the mean and the median to be about the same or markedly different, and why?

b. Compute the mean, the median, and the mode.

23. A man tosses a coin repeatedly until it lands heads and records the number of tosses required. (For example, if it lands heads on the first toss he records a 1; if it lands tails on the first two tosses and heads on the third he records a 3.) The data are shown.

x	1	2	3	4	5	6	7	8	9	10
f	384	208	98	56	28	12	8	2	3	1

a. Find the mean of the data.

b. Find the median of the data.

24. a. Construct a data set consisting of ten numbers, all but one of which is above average, where the average is the mean.

b. Is it possible to construct a data set as in part (a) when the average is the median? Explain.

25. Show that no matter what kind of average is used (mean, median, or mode) it is impossible for all members of a data set to be above average.

26. a. Twenty sacks of grain weigh a total of 1,003 lb. What is the mean weight per sack?

b. Can the median weight per sack be calculated based on the information given? If not, construct two data sets with the same total but different medians.

27. Begin with the following set of data, call it Data Set I.

$$5 \quad -2 \quad 6 \quad 14 \quad -3 \quad 0 \quad 1 \quad 4 \quad 3 \quad 2 \quad 5$$

a. Compute the mean, median, and mode.

b. Form a new data set, Data Set II, by adding 3 to each number in Data Set I. Calculate the mean, median, and mode of Data Set II.

c. Form a new data set, Data Set III, by subtracting 6 from each number in Data Set I. Calculate the mean, median, and mode of Data Set III.

d. Comparing the answers to parts (a), (b), and (c), can you guess the pattern? State the general principle that you expect to be true.

LARGE DATA SET EXERCISES

28. Large Data Set 1 lists the SAT scores and GPAs of 1,000 students.

http://www.flatworldknowledge.com/sites/all/files/data1.xls

a. Compute the mean and median of the 1,000 SAT scores.

b. Compute the mean and median of the 1,000 GPAs.

29. Large Data Set 1 lists the SAT scores of 1,000 students.

http://www.flatworldknowledge.com/sites/all/files/data1.xls

a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population mean μ.

b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample mean $\overline{x}$ and compare it to μ.

c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample mean $\overline{x}$ and compare it to μ.

30. Large Data Set 1 lists the GPAs of 1,000 students.

http://www.flatworldknowledge.com/sites/all/files/data1.xls

a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population mean μ.

b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample mean $\bar{x}$ and compare it to μ.

c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample mean $\bar{x}$ and compare it to μ.

31. Large Data Sets 7, 7A, and 7B list the survival times in days of 140 laboratory mice with thymic leukemia from onset to death.

    http://www.flatworldknowledge.com/sites/all/files/data7.xls

    http://www.flatworldknowledge.com/sites/all/files/data7A.xls

    http://www.flatworldknowledge.com/sites/all/files/data7B.xls

    a. Compute the mean and median survival time for all mice, without regard to gender.
    b. Compute the mean and median survival time for the 65 male mice (separately recorded in Large Data Set 7A).
    c. Compute the mean and median survival time for the 75 female mice (separately recorded in Large Data Set 7B).

ANSWERS

1. a. 9.
   b. 41.
   c. 0.
   d. 14.

3. $\bar{x} = 2.5, \tilde{x} = 2.5, \text{mode} = \{1,2,3,4\}$.

5. $\bar{x} = 3, \tilde{x} = 2, \text{mode} = 2$.

7. $\bar{x} = 3, \tilde{x} = 2, \text{mode} = 2$.

9. {0,0,3}.

11. {0,1,1,2}.

13. $\bar{x} = 146.9$, $\tilde{x} = 147.5$

15. $\bar{x} = 2.6$, $\tilde{x} = 2$, $\text{mode} = 2$

17. $\bar{x} = 48.96$, $\tilde{x} = 49$, $\text{mode} = 49$

19. a. No, the survival times of the fourth and fifth mice are unknown.

    b. Yes, $\tilde{x} = 421$.

21. $\bar{x} = 28.55$, $\tilde{x} = 28$, $\text{mode} = 28$

23. $\bar{x} = 2.05$, $\tilde{x} = 2$, $\text{mode} = 1$

25. Mean: $nx_{\text{min}} \le \Sigma x$ so dividing by *n* yields $x_{\text{min}} \le \bar{x}$, so the minimum value is not above average. Median: the middle measurement, or average of the two middle measurements, $\tilde{x}$, is at least as large as x_{min}, so the minimum value is not above average. Mode: the mode is one of the measurements, and is not greater than itself.

27. a. $\bar{x} = 3.\overline{18}$, $\tilde{x} = 3$, $\text{mode} = 5$.

    b. $\bar{x} = 6.\overline{18}$, $\tilde{x} = 6$, $\text{mode} = 8$.

    c. $\bar{x} = -2.\overline{81}$, $\tilde{x} = -3$, $\text{mode} = -1$.

    d. If a number is added to every measurement in a data set, then the mean, median, and mode all change by that number.

29. a. $\mu = 1528.74$

    b. $\bar{x} = 1502.8$

    c. $\bar{x} = 1535.2$

31. a. $\bar{x} = 553.4286$ and $\tilde{x} = 552.5$

    b. $\bar{x} = 665.9692$ and $\tilde{x} = 667$

    c. $\bar{x} = 455.8933$ and $\tilde{x} = 448$

Previous Section Table of Contents Next Section