2.3 Measures of Variability

LEARNING OBJECTIVES

1. To learn the concept of the variability of a data set.
2. To learn how to compute three measures of the variability of a data set: the range, the variance, and the standard deviation.

Look at the two data sets in Table 2.1 "Two Data Sets" and the graphical representation of each, called a *dot plot,* in Figure 2.10 "Dot Plots of Data Sets".

Table 2.1 Two Data Sets

| Data Set I: | 40 | 38 | 42 | 40 | 39 | 39 | 43 | 40 | 39 | 40 |
| Data Set II: | 46 | 37 | 40 | 33 | 42 | 36 | 40 | 47 | 34 | 45 |

Figure 2.10 *Dot Plots of Data Sets*

(a) Set I   (b) Set II
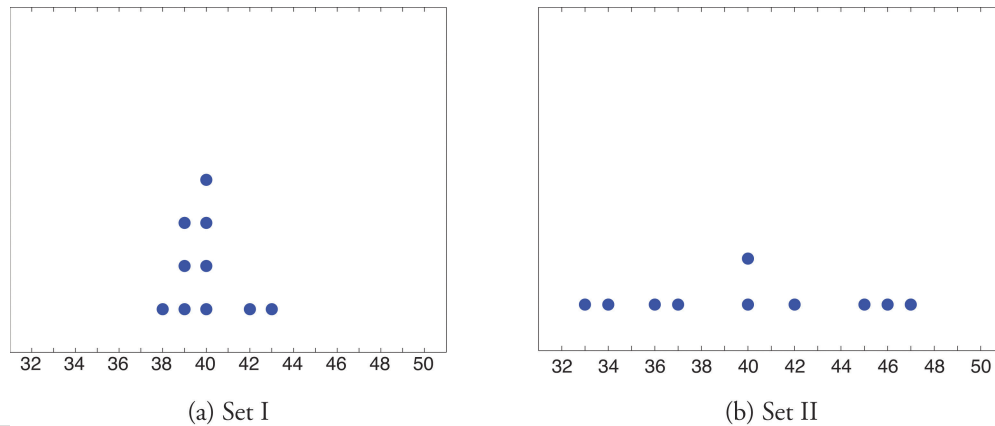
The two sets of ten measurements each center at the same value: they both have mean, median, and mode 40. Nevertheless a glance at the figure shows that they are markedly different. In Data Set I the measurements vary only slightly from the center, while for Data Set II the measurements vary greatly. Just as we have attached numbers to a data set to locate its center, we now wish to associate to each data set numbers that measure quantitatively how the data either scatter away from the center or cluster close to it. These new quantities are called measures of variability, and we will discuss three of them.

The Range

The first measure of variability that we discuss is the simplest.

EXAMPLE 10

Find the range of each data set in Table 2.1 "Two Data Sets".

Solution:

For Data Set I the maximum is 43 and the minimum is 38, so the range is $R = 43 - 38 = 5$.

For Data Set II the maximum is 47 and the minimum is 33, so the range is $R = 47 - 33 = 14$.

The range is a measure of variability because it indicates the size of the interval over which the data points are distributed. A smaller range indicates less variability (less dispersion) among the data, whereas a larger range indicates the opposite.

The Variance and the Standard Deviation

The other two measures of variability that we will consider are more elaborate and also depend on whether the data set is just a sample drawn from a much larger population or is the whole population itself (that is, a census).

Definition

The **sample variance** *of a set of n sample data is the number* s^2 *defined by the formula*

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n-1}$$

which by algebra is equivalent to the formula

$$s^2 = \frac{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}{n-1}$$

The **sample standard deviation** *of a set of n sample data is the square root of the sample variance, hence is the number s given by the formulas*

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}{n-1}}$$

Although the first formula in each case looks less complicated than the second, the latter is easier to use in hand computations, and is called a **shortcut formula**.

EXAMPLE 11

Find the sample variance and the sample standard deviation of Data Set II in Table 2.1 "Two Data Sets".

Solution:

To use the defining formula (the first formula) in the definition we first compute for each observation x its deviation $x - \bar{x}$ from the sample mean. Since the mean of the data is $\bar{x} = 40$

, we obtain the ten numbers displayed in the second line of the supplied table.

x	46	37	40	33	42	36	40	47	34	45
$x - \bar{x}$	6	−3	0	−7	2	−4	0	7	−6	5

Then

$$\Sigma(x - \bar{x})^2 = 6^2 + (-3)^2 + 0^2 + (-7)^2 + 2^2 + (-4)^2 + 0^2 + 7^2 + (-6)^2 + 5^2 = 224$$

so

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n-1} = \frac{224}{9} = 24.\overline{8}$$

and

$$s = \sqrt{24.\overline{8}} \approx 4.99$$

The student is encouraged to compute the ten deviations for Data Set I and verify that their squares add up to 20, so that the sample variance and standard deviation of Data Set I are the much smaller numbers $s^2 = 20/9 = 2.\overline{2}$ and $s = \sqrt{20/9} \approx 1.49$.

EXAMPLE 12

Find the sample variance and the sample standard deviation of the ten GPAs in Note 2.12 "Example 3" in Section 2.2 "Measures of Central Location".

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

Solution:

Since

$$\Sigma x = 1.90 + 3.00 + 2.53 + 3.71 + 2.12 + 1.76 + 2.71 + 1.39 + 4.00 + 3.33 = 26.45$$

and

$$\Sigma\,x^2 \;=\; 1.90^2 + 3.00^2 + 2.53^2 + 3.71^2 + 2.12^2 + 1.76^2$$
$$+2.71^2 + 1.39^2 + 4.00^2 + 3.33^2$$
$$=\; 76.7321$$

the shortcut formula gives

$$s^2 \;=\; \frac{\Sigma\,x^2 - \frac{1}{n}(\Sigma\,x)^2}{n-1} \;=\; \frac{76.7321 - \frac{(26.45)^2}{10}}{10-1} \;=\; \frac{6.77185}{9} \;=\; .752427$$

and

$$s \;=\; \sqrt{.752427} \;\approx\; .867$$

The sample variance has different units from the data. For example, if the units in the data set were inches, the new units would be inches squared, or square inches. It is thus primarily of theoretical importance and will not be considered further in this text, except in passing.

If the data set comprises the whole population, then the *population* standard deviation, denoted σ (the lower case Greek letter sigma), and its square, the *population* variance σ^2, are defined as follows.
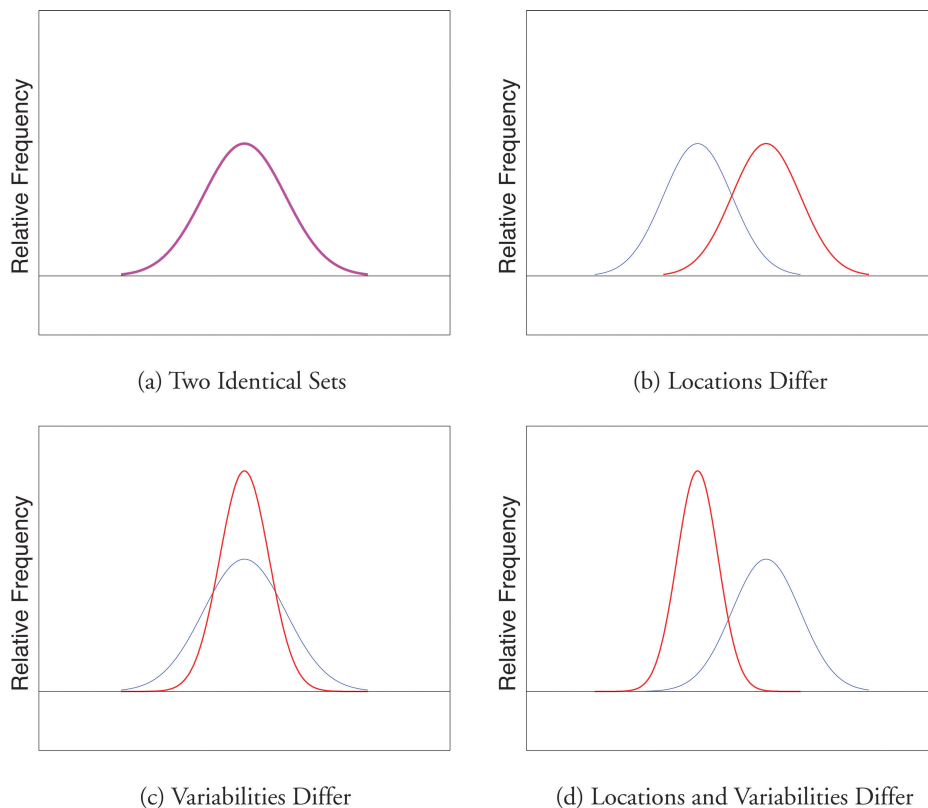

Definition

The **population variance** *and* **population standard deviation** *of a set of N population data are the numbers σ^2 and σ defined by the formulas*

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N} \quad and \quad \sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$

Note that the denominator in the fraction is the full number of observations, not that number reduced by one, as is the case with the sample standard deviation. Since most data sets are samples, we will always work with the sample standard deviation and variance.

Finally, in many real-life situations the most important statistical issues have to do with comparing the means and standard deviations of two data sets. Figure 2.11 "Difference between Two Data Sets" illustrates how a difference in one or both of the sample mean and the sample standard deviation are reflected in the appearance of the data set as shown by the curves derived from the relative frequency histograms built using the data.

Figure 2.11 Difference between Two Data Sets

(a) Two Identical Sets

(b) Locations Differ

(c) Variabilities Differ

(d) Locations and Variabilities Differ

KEY TAKEAWAY

The range, the standard deviation, and the variance each give a quantitative answer to the

question "How variable are the data?"

EXERCISES

BASIC

1. Find the range, the variance, and the standard deviation for the following sample.

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

2. Find the range, the variance, and the standard deviation for the following sample.

$$2 \quad -3 \quad 6 \quad 0 \quad 3 \quad 1$$

3. Find the range, the variance, and the standard deviation for the following sample.

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

4. Find the range, the variance, and the standard deviation for the following sample.

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

5. Find the range, the variance, and the standard deviation for the sample represented by the data frequency table.

x	1	2	7
f	1	2	1

6. Find the range, the variance, and the standard deviation for the sample represented by the data frequency table.

x	-1	0	1	4
f	1	1	3	1

APPLICATIONS

7. Find the range, the variance, and the standard deviation for the sample of ten IQ scores randomly selected from a school for academically gifted students.

$$132 \quad 162 \quad 133 \quad 145 \quad 148$$
$$139 \quad 147 \quad 160 \quad 150 \quad 153$$

8. Find the range, the variance and the standard deviation for the sample of ten IQ scores randomly selected from a school for academically gifted students.

$$142 \ 152 \ 138 \ 145 \ 148$$
$$139 \ 147 \ 155 \ 150 \ 153$$

ADDITIONAL EXERCISES

9. Consider the data set represented by the table

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

a. Use the frequency table to find that $\Sigma x = 1256$ and $\Sigma x^2 = 35{,}926$.

b. Use the information in part (a) to compute the sample mean and the sample standard deviation.

10. Find the sample standard deviation for the data

x	1	2	3	4	5
f	384	208	98	56	28

x	6	7	8	9	10
f	12	8	2	3	1

11. A random sample of 49 invoices for repairs at an automotive body shop is taken. The data are arrayed in the stem and leaf diagram shown. (Stems are thousands of dollars, leaves are hundreds, so that for example the largest observation is 3,800.)

```
3 | 5 6 8
3 | 0 0 1 1 2 4
2 | 5 6 6 7 7 8 8 9 9
2 | 0 0 0 0 1 2 2 4
1 | 5 5 5 6 6 7 7 7 8 8 9
1 | 0 0 1 3 4 4 4
0 | 5 6 8 8
0 | 4
```

For these data, $\Sigma x = 101{,}100$, $\Sigma x^2 = 244{,}830{,}000$.

a. Compute the mean, median, and mode.

b. Compute the range.

c. Compute the sample standard deviation.

12. What must be true of a data set if its standard deviation is 0?

13. A data set consisting of 25 measurements has standard deviation 0. One of the measurements

has value 17. What are the other 24 measurements?

14. Create a sample data set of size $n = 3$ for which the range is 0 and the sample mean is 2.

15. Create a sample data set of size $n = 3$ for which the sample variance is 0 and the sample mean is 1.

16. The sample $\{-1,0,1\}$ has mean $\bar{x} = 0$ and standard deviation $s = 1$. Create a sample data set of size $n = 3$ for which $\bar{x} = 0$ and s is greater than 1.

17. The sample $\{-1,0,1\}$ has mean $\bar{x} = 0$ and standard deviation $s = 1$. Create a sample data set of size $n = 3$ for which $\bar{x} = 0$ and the standard deviation s is less than 1.

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

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

a. Compute the sample standard deviation of Data Set I.

b. Form a new data set, Data Set II, by adding 3 to each number in Data Set I. Calculate the sample standard deviation 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 sample standard deviation 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

19. 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 range and sample standard deviation of the 1,000 SAT scores.

b. Compute the range and sample standard deviation of the 1,000 GPAs.

20. 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 range and population standard deviation σ.

b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.

c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.

21. 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 range and population standard deviation σ.

b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.

c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.

22. 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 range and sample standard deviation of survival time for all mice, without regard to gender.

b. Compute the range and sample standard deviation of survival time for the 65 male mice (separately recorded in Large Data Set 7A).

c. Compute the range and sample standard deviation of survival time for the 75 female mice (separately recorded in Large Data Set 7B). Do you see a difference in the results for male and female mice? Does it appear to be significant?

ANSWERS

1. $R = 3$, $s^2 = 1.7$, $s = 1.3$.

3. $R = 6$, $s^2 = 7.\overline{3}$, $s = 2.7$.

5. $R = 6$, $s^2 = 7.3$, $s = 2.7$.

7. $R = 30$, $s^2 = 103.2$, $s = 10.2$.

9. $\bar{x} = 28.55$, $s = 1.3$.

11. a. $\bar{x} = 2063$, $\tilde{x} = 2000$, mode $= 2000$.

   b. $R = 3400$.

   c. $s = 869$.

13. All are 17.

15. {1,1,1}

17. One example is $\{-.5, 0, .5\}$.

19. a. $R = 1350$ and $s = 212.5455$

   b. $R = 4.00$ and $s = 0.7407$

21. a. $R = 4.00$ and $\sigma = 0.740375$

   b. $R = 3.04$ and $s = 0.808045$

   c. $R = 2.49$ and $s = 0.657843$