Unit 6: Standard Deviation

SUMMARY OF VIDEO

The video begins with a tale of two cities, Portland, Oregon, and Montreal, Quebec. The average monthly precipitation in these two cities is similar – Portland's average precipitation is 3.32 inches per month and Montreal's is 3.4 inches per month. Not much difference there! However, the average (or mean) monthly precipitation for these two cities is not the whole story. As can be seen in Figure 6.1, Montreal's precipitation is relatively consistent from month to month, while Portland's is far more variable – more rain in the winter months and little rain in the summer months.

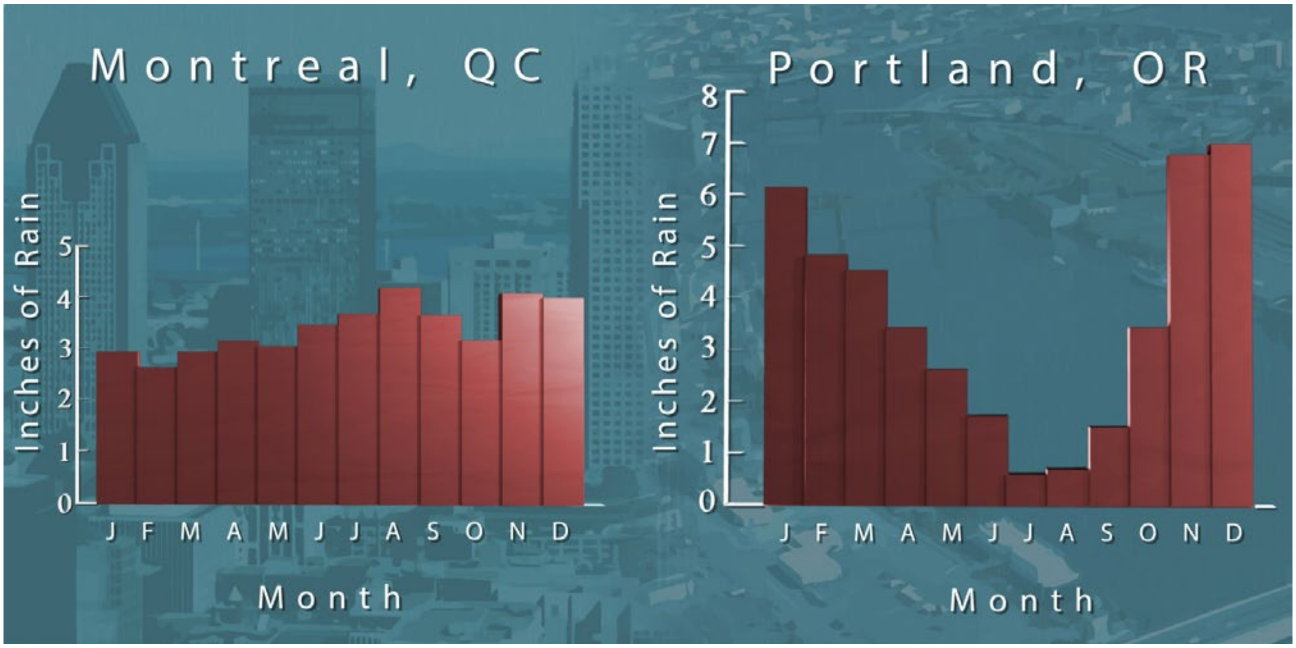
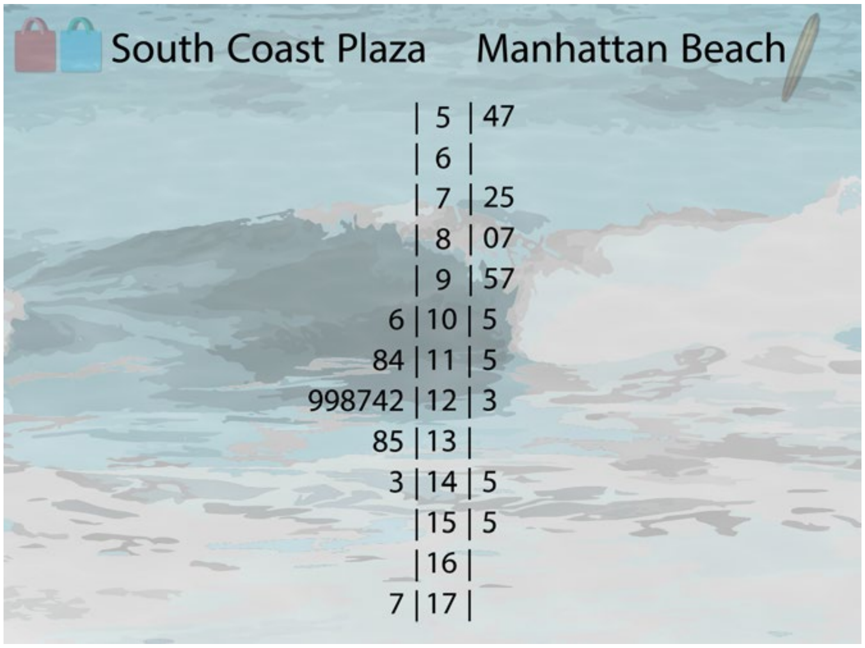


Figure 6.1. Monthly precipitation for Montreal and Portland.

We've hit on a case where a measure of center does not provide all the information we need. To express the climate difference between these two cities, we need a measure of how much spread or variability there is in month-to-month precipitation.

Next, the video switches to two locations in California – the sites of two Wahoo's Fish Taco restaurants, one located at Manhattan Beach and the other in the South Coast Plaza. Wing Lam and his two brothers founded their first Wahoo's as a single taco shop on the beach. Now they have more than 50 Wahoo's Fish Taco restaurants in their chain. Knowing what to expect, based on how busy each month has been in the past, lets managers plan inventory orders and staff schedules appropriate to each location.

The mean sales per four-week period for the South Coast Plaza and Manhattan Beach locations are around $130,000 and $97,000, respectively. But that's not all that differs between these two locations. South Coast Plaza is a great store because it is indoors in a shopping mall, and therefore, sales are relatively unaffected by the weather. On the other hand, the Manhattan Beach restaurant is on the beach and so the weather has greater impact on its sales. Back-to-back stemplots in Figure 6.2 show the sales from these two restaurants over four-week periods.



South Coast Plaza		Manhattan Beach
	5	47
	6	
	7	25
	8	07
	9	57
6	10	5
84	11	5
998742	12	3
85	13	
3	14	5
	15	5
	16	
7	17	

Figure 6.2. Back-to-back stemplots of sales.

Because the South Coast Plaza restaurant is indoors, its sales are pretty stable. A stemplot of the South Coast Plaza restaurant data is roughly symmetric, with one outlier at $177,000 (for December). On the other hand, the sales data for the Manhattan Beach location are more variable than the South Coast Plaza location. The range for the Manhattan Beach location is over $100,000 compared to a range of $70,000 for the South Coast Plaza location. Boxplots in Figure 6.3 based on these sales figures make the variation obvious. Check out the interquartile ranges, represented by the widths of the boxes. The interquartile range for the Manhattan Beach location is wider than for the South Coast Plaza location.

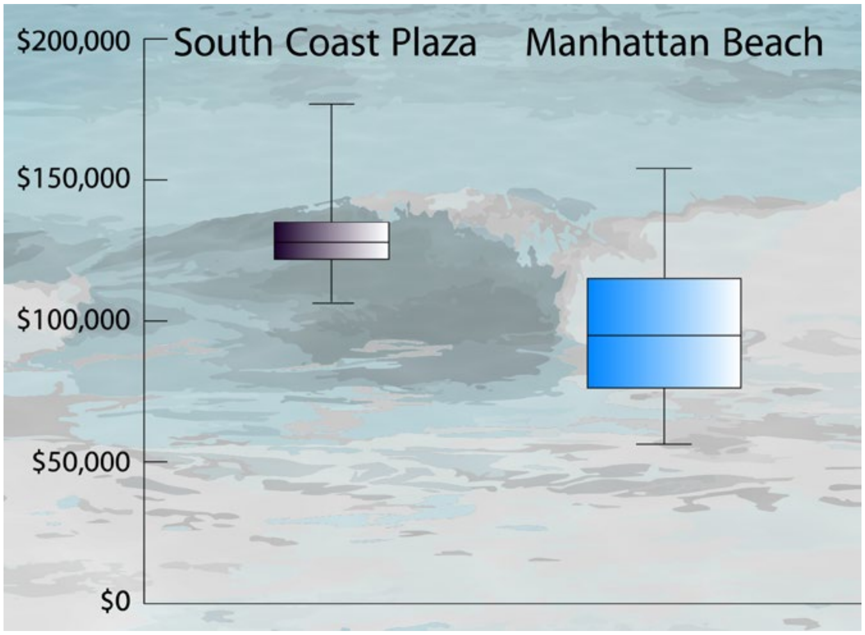


Figure 6.3. Boxplots of the sales for two Wahoo's Fish Taco restaurants.

Another way to quantify the spread is to measure how far observations are from their mean. This is the idea behind measures of spread called the variance and standard deviation.
To find the variance of the Manhattan Beach and South Coast Plaza data, we use the following formula

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

The standard deviation is the square root of the variance. The standard deviations for the Manhattan Beach and South Coast Plaza Wahoo's are $31,075 and $17,000, respectively. This difference in variability makes sense given the locations of these two Wahoo's. At the beach, sales are affected by weather – warm, sunny weather brings people to the beach along with good sales, while cold rainy days keep people away resulting in poor sales. However, at the indoor mall location, Wahoo's sales are fairly steady all year long.

STUDENT LEARNING OBJECTIVES

A. Know that the sample standard deviation, *s,* is the measure of spread most commonly used when the mean, $\overline{x}$, is used as the measure of center.

B. Be able to calculate the standard deviation *s* from the formula for small data sets (say $n \leq 10$).

C. Know the basic properties of the standard deviation:

- $s \geq 0$, and $s = 0$ only when all data values are identical.
- *s* increases as the spread about $\overline{x}$ increases.
- *s,* like $\overline{x}$, is strongly influenced by outliers.

D. Know that the standard deviation is most useful for symmetric distributions and, in particular, for normal distributions.

E. Know that adding the same constant **a** to all the observations increases the value of $\overline{x}$ by **a**. However, adding the same constant **a** to all the observations does not change the value of *s.* That's because adding a constant **a t**o all data values shifts the location of the data but does not affect its spread.

F. Know that multiplying all data values by a constant amount *k* changes $\overline{x}$ and *s* by a factor of *k.*

CONTENT OVERVIEW

Although the variance and standard deviation are the most common measures of spread or variability in statistical practice, they are tedious to calculate from their formulas and somewhat difficult to interpret. Their common use arises from the fact that the standard deviation is the natural measure of spread for normal distributions, in which data tend to be mound-shaped and symmetric. (Normal distributions will be covered in Units 7 – 9.) For describing distributions of data, the five-number summary is generally more useful than $\overline{x}$ and s particularly for distributions that are not roughly mound-shaped and symmetric.

If we decide to use the mean $\overline{x}$ to describe the center of a set of data, then it makes sense to use the deviations from the mean, $x - \overline{x}$, as the basis for a measure of spread. Clearly, if all the deviations from the mean are small, then all data values lie close to $\overline{x}$ and there is little spread. The only problem is that some of the deviations from the mean are positive and some are negative. If you sum the deviations from the mean, $\sum(x - \overline{x})$, the result is always exactly zero. However, the sum of the squared deviations will be positive and so the “average squared deviation from the mean” makes sense as a measure of spread. This is the idea behind the **variance**. The **standard deviation** is the square root of the variance.

The variance is an average of the squared deviations from the mean:

$$s^2 = \frac{\sum\left(x - \overline{x}\right)^2}{n-1}$$, where *n* is the sample size

You are probably wondering why we divide by $n-1$ in computing the average rather than by *n.* Here is an explanation. Because the sum of the deviations, $\sum(x - \overline{x})$, is always zero, the last deviation can be found once we know the first $n-1$. That means that only $n-1$ of the squared deviations can vary freely, and so, the average is found by dividing by $n-1$.

Because the variance involves squaring the deviations, it does not have the same units of measurement as the original data values. For example, lengths measured in centimeters have a variance measured in squared centimeters. Taking the square root remedies this, so that the standard deviation $s = \sqrt{s^2}$ measures dispersion about the mean in the original scale.

Calculations of s or s^2 using the formula can be tedious even for relatively small data sets. As an example, we calculate the mean and standard deviation for the sample of five data values given below.

90 40 85 69 79

First, we calculate the sample mean as follows:

$$\bar{x} = \frac{90+40+85+69+79}{5} = \frac{363}{5} = 72.6$$

Now, we are ready to calculate the variance. Here are the steps:

- For each data value, x, we calculate its deviation from the mean, x - 72.6, and enter that value into the second column of Table 6.1.
- Next, we square each of these deviations and enter them into the third column.
- Then we sum the entries in the third column.

x	x - 72.6	$(x - 72.6)^2$
90	17.4	302.76
40	-32.6	1062.76
85	12.4	153.76
69	-3.6	12.96
79	6.4	40.96
	Sum =	1573.2

Table 6.1. Calculating the Variance.

For the final step, we divide the sum by 5 – 1, or 4:

$$s^2 = \frac{1573.2}{5-1} = 393.3$$

To compute the standard deviation, we take the square root of the variance:

$$s = \sqrt{393.3} \approx 19.8$$

Although calculating the standard deviation from the formula is reasonable when the sample size is small, for larger data sets it is better to use graphing calculators or software (such as Minitab, SPSS, and Excel). We conclude this overview with a list of properties of the standard deviation.

- The standard deviation, *s*, measures spread about the mean, and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise $s > 0$, and *s* increases as the observations move farther apart (and hence farther from the mean).
- Like the mean, *s* is strongly affected by a few extreme observations. In fact, the use of squared deviations renders *s* even more sensitive to a few extreme observations than is the mean.
- Adding a constant value to each data value shifts the data affecting its location but not its spread. Therefore, *s* does not change.
- Multiplying each data value by a constant does affect the spread of data. If *s* is the standard deviation of a data set, and the data are modified by multiplying each data value by a constant *k,* then the standard deviation of the modified data set is $k \cdot s$.

KEY TERMS

Given a data set, one measure of center is the mean, $\bar{x}$. One way to judge the spread of the data is to look at the **deviations from the mean**, $x - \bar{x}$.

The **variance** is a measure of variability that is based on the square of the deviations from the mean. The formula for computing variance is:

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

Because the units for variance are the square of the units for the original data, we generally take the square root of the variance, which gives us the standard deviation:

$$s = \sqrt{s^2}$$

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. In comparing monthly precipitation for Portland, Oregon, and Montreal, Canada, why was comparing the mean monthly precipitation rates insufficient?

2. Why don't we measure spread about the mean by simply averaging $x - \bar{x}$, the deviations of individual data values from their mean?

3. What did the standard deviation of four-week sales data tell you about the two Wahoo's Taco locations, Manhattan Beach and South Coast Plaza?

4. Can the standard deviation of a set of observations be $s = -1.5$? Explain.

UNIT ACTIVITY:

VISUALIZING STANDARD DEVIATION

The standard deviation is a measure of spread that is based on the deviations from the mean. For some insight into deviations from the mean, we start with the following data set: 6, 6, 2, 8, 3. We calculate the mean of these data:

$$\bar{x} = \frac{6+6+2+8+3}{5} = 5 .$$

A dotplot of the 5 data values is shown in Figure 6.4. A vertical line has been drawn at the mean, $\bar{x} = 5$. The horizontal line segments represent the deviations of each data value from the mean. The lowest horizontal line segment represents the deviation of the first data value, 6, from the mean; this deviation has length 1, since 6 – 5 = 1.

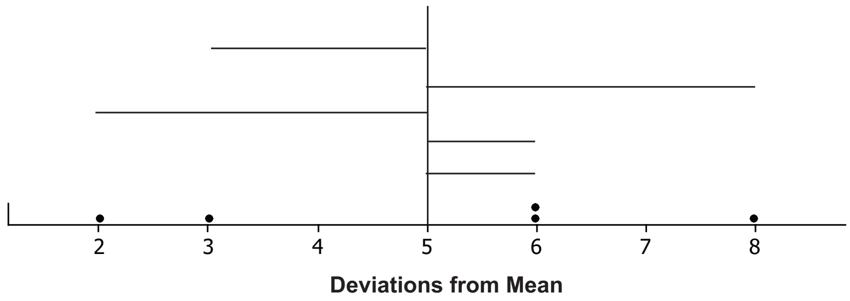


Figure 6.4. Dotplot with horizontal line segments.

1. a. List the deviations from the mean for each of the 5 data values.

b. What is the sum of the five deviations from the mean listed in (a)?

c. Some of the deviations from the mean are positive and some are negative. To keep the deviations from cancelling each other out, square each of the deviations listed in (a). What is the sum of the square of the deviations?

d. To find the variance, s^2, divide your answer to (c) by n – 1, which in this case is 4. What is the variance? The standard deviation is the square root of the variance. What is the standard deviation? (Round to two decimals.)

2. Consider two more data sets:

Data Set X: 2, 4, 3, 4, 5, 3
Data Set Y: 3, 2, 3, 10, 5, 4

a. Calculate the means for Data Sets *X* and *Y*.

b. Draw dotplots of the two data sets – use the same scale for both plots. Draw a vertical line at the mean and then represent the deviations from the mean as horizontal line segments.

c. Based on your plots in (b), which data set do you think will have the larger standard deviation? Explain your reasoning.

d. Calculate the standard deviations for Data Sets *x* and *Y*. (Round answers to two decimals.) Do your calculations confirm your answer to (c)?

Figures 6.5 – 6.9 show histograms of five data sets. Your task in questions 3 and 4 will be to determine which data sets have larger standard deviations based on histograms of the data.

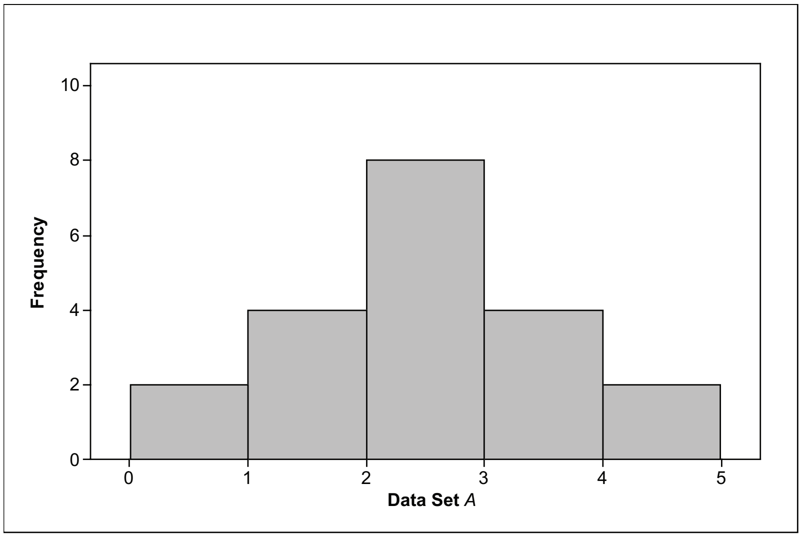


Figure 6.5. Histogram of Data Set A.

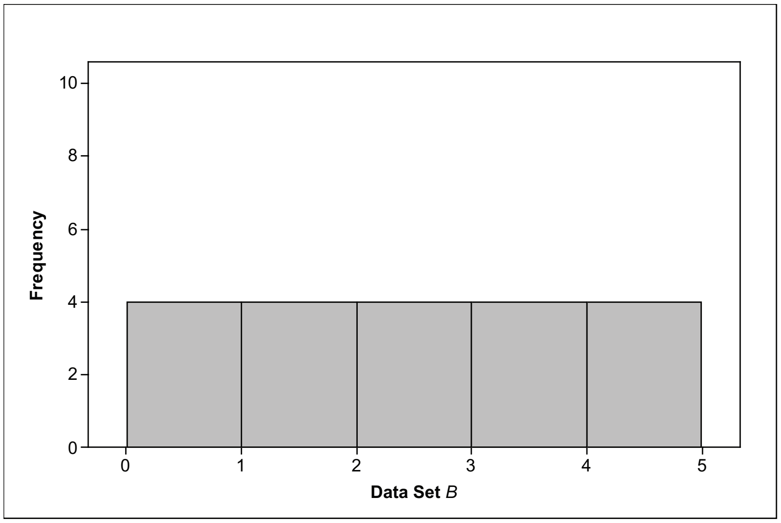


Figure 6.6. Histogram of Data Set B.

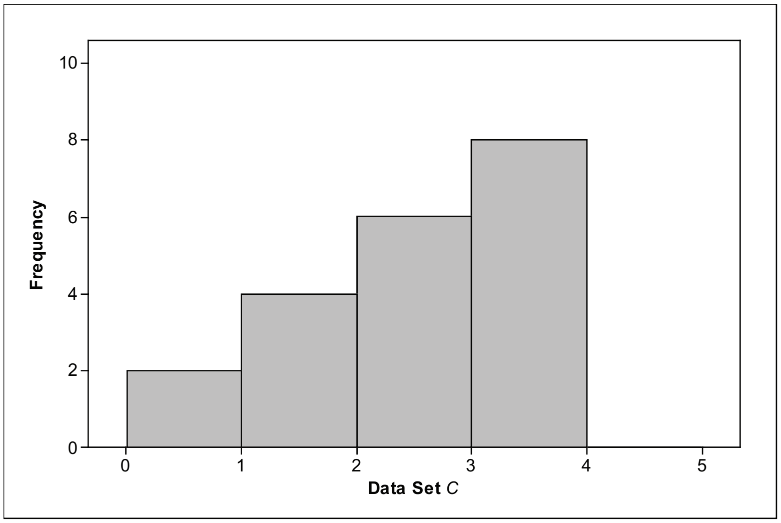


Figure 6.7. Histogram of Data Set C.

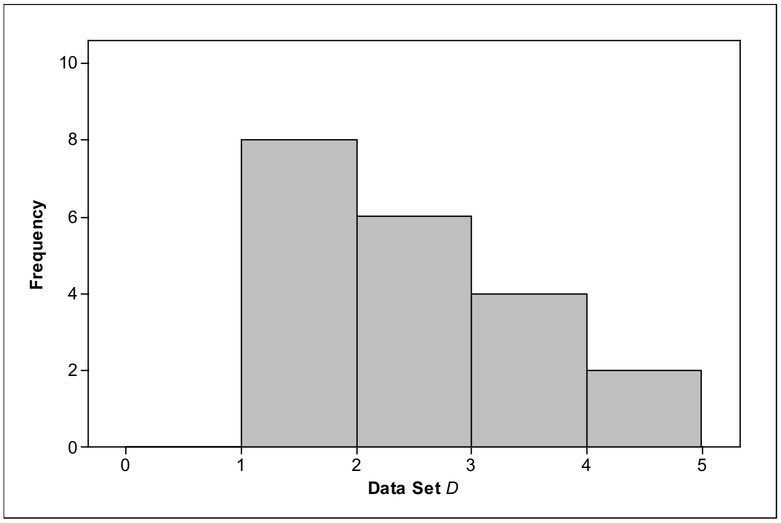


Figure 6.8. Histogram of Data Set D.

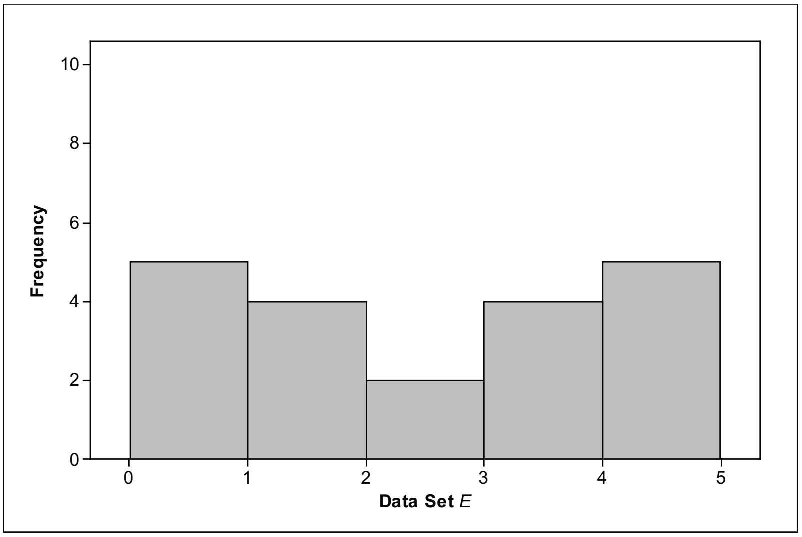


Figure 6.9. Histogram of Data Set E.

3. Data Sets *A* – *E* displayed graphically in Figures 6.5 – 6.9 all have means of 2.5. So, the mean does not provide any information that could be used to distinguish one data set from another. In parts (a) – (c), determine from the histograms which of the two data sets has the larger standard deviation, or if the standard deviations are about the same. In each case, give a justification of your answer.

a. Data Set A and Data Set B.

b. Data Set C and Data Set D.

c. Data Set D and Data Set E.

4. The standard deviations of the Data Sets *A* – *E* are given below in random order. Match each standard deviation with its data set.

1.589   1.026   1.124   1.026   1.451

EXERCISES

1. SAT Math scores in recent years have had means around 490 and standard deviations around 100. What is the variance of SAT Math scores?

2. Six ninth-grade students and six 12th-grade students were asked: How many movies have you seen this month? Here are their responses.

Ninth-grade students:	5, 1, 2, 5, 3, 8
12th-grade students:	4, 2, 0, 2, 3, 1

a. Calculate the mean, variance, and standard deviation of each of these data sets. Which is more spread out, the ninth-grade or 12th-grade data set?

b. Make a graph of both data sets. Which of these data sets appears more spread out? Does your answer agree with your conclusion in part (a)?

3. Suppose we add 2 to each of the numbers in the ninth-grade data in question 2. That modification produces the following data: 7, 3, 4, 7, 5, 10.

a. Find the mean and the standard deviation of the modified data.

b. Compare your answers from (a) with the mean and standard deviation from the original ninth-graders' data in question 2. How did adding 2 to each data value change the mean? How did it change the standard deviation?

c. Without doing the calculations, guess what will happen to the mean and standard deviation of the 12th-graders' data from question 2 if we add 10 to each data value.

4. Return to the SAT data from Table 2.1, Unit 2.

a. Use technology to calculate the standard deviation for the critical reading, mathematics, and writing SAT scores.

b. Based on the standard deviations, which of the three SAT exams has the largest spread over the 50 states and the District of Columbia?

REVIEW QUESTIONS

1. a. If two distributions have exactly the same mean and standard deviation, must their histograms look exactly alike? Explain.

b. If two distributions have the same five-number summary, must their histograms be identical? Explain.

2. The Army wants to describe the distribution of head sizes of its soldiers in order to plan orders of helmets. Here are the head sizes in inches of 30 male soldiers, which were obtained by putting a tape measure around each soldier's forehead.

23.0 22.2 21.7 22.0 22.3 22.6
22.7 21.5 22.7 24.9 20.8 23.3
24.2 23.5 23.9 23.4 20.8 21.5
23.0 24.0 22.7 22.6 23.9 21.8
23.1 21.9 21.0 22.4 23.5 22.5

a. Give a graphical description of these data. Is there any aspect of the distribution that would discourage the use of $\bar{x}$ and s to measure center and spread?

b. Find $\bar{x}$ and s for these data. Be sure to include the units in which these numbers are measured.

c. What percentage of the data is within one standard deviation of the mean?

3. a. The head-size data in question 2 was measured in inches. There are 2.54 centimeters per inch. Change the head-size data from inches to centimeters by multiplying each data value by 2.54.

b. Calculate the standard deviation of the head-size data from (a).

c. How is the standard deviation from (a), for head sizes measured in centimeters, related to the standard deviation in 2(b), for head sizes measured in inches?

4. Two types of wire, a 12 ½-gauge low-carbon wire and a thinner 14-gauge high-tensile wire, both used in barbed-wire fencing, are tested to see which wire is stronger. Data on the breaking strengths, in pounds, of both types of wire are given below.

12 ½-gauge, low-carbon wire:

455	455	495	490	410	470	475	450	480	460
480	445	435	405	450	500	435	430	460	480

14-gauge, high-tensile wire:

780	780	775	770	780	780	790	780	770	780
780	785	800	775	760	770	800	780	790	790

a. Make comparative boxplots for the breaking strengths of the types of wire. Would you describe the shapes of the distributions as symmetric or skewed?

b. Determine the mean breaking strengths for the two types of wire. Which type of wire has the larger mean breaking strength?

c. The two types of wires have different properties that affect the consistency of the wire's breaking strength. Find the standard deviation of breaking strengths for each type of wire. Which wire has the more variable breaking strength?