Lesson 5: Measuring Variability for Symmetrical Distributions

Student Outcomes

- Students calculate the standard deviation for a set of data.
- Students interpret the standard deviation as a typical distance from the mean.

Lesson Notes

In this lesson, students calculate standard deviation for the first time and examine the process for its calculation more closely. Through questioning and discussion, students link each step in the process to its meaning in the context of the problem and explore the many questions about the rationale behind the development of the formula. Guiding questions and responses to facilitate this discussion are provided as the closing discussion for this lesson. However, it is recommended that the discussion be allowed to occur at any point in the lesson when students are asking questions about the calculation of standard deviation.

Classwork

Example 1 (12 minutes): Calculating the Standard Deviation

Discuss the following points with students using the dot plot and students' previous results in Lesson 4.

- In Lesson 4, we looked at what might be a typical deviation from the mean. We will now develop a way to use the deviations from the mean to calculate a measure of variability called the *standard deviation*.
- Let's return to the battery lifetimes of Brand A from Lesson 4. Look at the dot plot of the lives of the Brand A batteries.

> **Example 1: Calculating the Standard Deviation**
>
> **Here is a dot plot of the lives of the Brand A batteries from Lesson 4.**
>
>
>
> Battery life (hours)
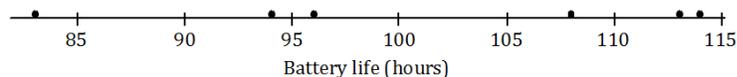
- The mean was 101 hours. Mark the location of the mean on the dot plot above.
- What is a typical distance or deviation from the mean for these Brand A batteries?
  - *A typical deviation is around* 10 *hours.*
- Now let's explore a more common measure of deviation from the mean—the standard deviation.

Walk students through the steps in their lesson resources for Example 1.

How do you measure variability of this data set? One way is by calculating standard deviation.

- First, find each deviation from the mean.
- Then, square the deviations from the mean. For example, when the deviation from the mean is -18, the squared deviation from the mean is $(-18)^2 = 324$.

Life (Hours)	83	94	96	106	113	114
Deviation from the Mean	-18	-7	-5	5	12	13
Squared Deviations from the Mean	324	49	25	25	144	169

- Add up the squared deviations:

    $324 + 49 + 25 + 25 + 144 + 169 = 736$.

    This result is the *sum* of the squared deviations.

The number of values in the data set is denoted by n. In this example, n is 6.

- You divide the sum of the squared deviations by $n - 1$, which here is $6 - 1$, or 5.

$$\frac{736}{5} = 147.2$$

- Finally, you take the square root of 147.2, which to the nearest hundredth is 12.13.

That is the standard deviation! It seems like a very complicated process at first, but you will soon get used to it.

We conclude that a typical deviation of a Brand A battery lifetime from the mean battery lifetime for Brand A is 12.13 hours. The unit of standard deviation is always the same as the unit of the original data set. So, the standard deviation to the nearest hundredth, with the unit, is 12.13 hours. How close is the answer to the typical deviation that you estimated at the beginning of the lesson?

- How close is the answer to the typical deviation that you estimated at the beginning of the lesson?
    - *It is fairly close to the typical deviation of around* 10 *hours.*

The value of 12.13 could be considered to be reasonably close to the earlier estimate of 10. The fact that the standard deviation is a little larger than the earlier estimate could be attributed to the effect of the point at 83. The standard deviation is affected more by values with comparatively large deviations from the mean than, for example, is the mean absolute deviation that students learned about in Grade 6.

This is a good time to mention precision when calculating the standard deviation. Encourage students, when calculating the standard deviation, to use several decimal places in the value that they use for the mean. Explain that students might get somewhat varying answers for the standard deviation depending on how far they round the value of the mean.

Exercises 1–5 (8 minutes)

Have students work independently and confirm answers with a neighbor or the group. Discuss any conflicting answers as needed.

Exercises 1–5

Now you can calculate the standard deviation of the lifetimes for the eight Brand B batteries. The mean was 100.5. We already have the deviations from the mean.

Life (Hours)	73	76	92	94	110	117	118	124
Deviation from the Mean	-27.5	-24.5	-8.5	-6.5	9.5	16.5	17.5	23.5
Squared Deviation from the Mean	756.25	600.25	72.25	42.25	90.25	272.25	306.25	552.25

1. Write the squared deviations in the table.

    See table above.

2. Add up the squared deviations. What result do you get?

    The sum is $2,692$.

3. What is the value of n for this data set? Divide the sum of the squared deviations by $n-1$, and write your answer below. Round your answer to the nearest thousandth.

    $$n = 8; \quad \frac{2692}{7} \approx 384.571$$

4. Take the square root to find the standard deviation. Record your answer to the nearest hundredth.

    $$\sqrt{384.571} \approx 19.61$$

5. How would you interpret the standard deviation that you found in Exercise 4? (Remember to give your answer in the context of this question. Interpret your answer to the nearest hundredth.)

    The standard deviation, 19.61 hours, is a typical deviation of a Brand B battery lifetime from the mean battery lifetime for Brand B.

MP.3

▪ So, now we have computed the standard deviation of the data on Brand A and of the data on Brand B. Compare the two, and describe what you notice in the context of the problem.

  ▫ *The fact that the standard deviation for Brand B is greater than the standard deviation for Brand A tells us that the battery life of Brand B had a greater spread (or variability) than the battery life of Brand A. This means that the Brand B battery lifetimes tended to vary more from one battery to another than the battery lifetimes for Brand A.*

Exercises 6–7 (8 minutes)

Have students work independently and confirm answers with a neighbor or the group. Discuss any conflicting answers as needed.

Exercises 6–7

Jenna has bought a new hybrid car. Each week for a period of seven weeks, she has noted the fuel efficiency (in miles per gallon) of her car. The results are shown below.

$$45 \quad 44 \quad 43 \quad 44 \quad 45 \quad 44 \quad 43$$

6. Calculate the standard deviation of these results to the nearest hundredth. Be sure to show your work.

The mean is 44.

The deviations from the mean are $1,\ 0,\ -1,\ 0,\ 1,\ 0,\ -1$.

The squared deviations from the mean are $1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1$.

The sum of the squared deviations is 4.

$$n = 7;\ \frac{4}{6} \approx 0.667$$

The standard deviation is $\sqrt{0.667}$ which is approximately 0.82 miles per gallon.

7. What is the meaning of the standard deviation you found in Exercise 6?

The standard deviation, 0.82 miles per gallon, is a typical deviation of a weekly fuel efficiency value from the mean weekly fuel efficiency.

Closing (7 minutes)

- What result would we get if we just added the deviations from the mean?
  - *Zero. This value highlights the fact that the mean is the balance point for the original distribution.*
- Why do you suppose that we square each deviation?
  - *This is one way to avoid the numbers adding to zero. By squaring the deviations, we make sure that all the numbers are positive.*

Students might also ask why we square the deviations and then take the square root at the end. Why not just find the average of the absolute values of the deviations from the mean? The answer is that the two approaches give different answers. (The square root of an average of squares of positive numbers is different from the average of the original set of numbers.) However, this idea of finding the mean of the absolute values of the deviations is a perfectly valid measure of the variability of a data set. This measure of spread is known as the mean absolute deviation (MAD), and students used this measure in previous years. The reason that variance and standard deviation are used more commonly than the mean absolute deviation is that the variance (and, therefore, the standard deviation) turns out to behave very nicely, mathematically, and is useful for developing relatively straightforward techniques of statistical analysis.

- ▪ Why do we take the square root?
    - □ *Before taking the square root, we have a typical squared deviation from the mean. It is easier to interpret a typical deviation from the mean than a typical squared deviation from the mean because a typical deviation has the same units as the original data. For example, the typical deviation from the mean for the battery life data is expressed in hours rather than hours[2].*
- ▪ Why did we divide by $n - 1$ instead of n?
    - □ *We only use $n - 1$ when we are calculating the standard deviation using sample data. Careful study has shown that using $n - 1$ gives the best estimate of the standard deviation for the entire population. If we have data from an entire population, we would divide by n instead of $n - 1$. (See note below for a detailed explanation.)*

NOTE: This topic is addressed throughout a study of statistics:

More information on why to divide by $n - 1$ and not n:

It is helpful to first explore the variance. The variance of a set of values is the square of the standard deviation. So, to calculate the variance, the same process is used, but the square root is not taken at the end.

Suppose, for a start, that there is a very large population of values, for example, the heights of all the people in a country. One can think of the variance of this population as being calculated using a division by n (although, since the population is very large, the difference between using n or $n - 1$ for the population is extremely small).

Imagine now taking a random sample from the population, such as taking a random sample of people from the country and measuring their heights. Use the variance of the sample as an estimate of the variance of the population. If division by n is used in calculating the variance of the sample, the result tends to be a little too small as an estimate of the population variance. (To be a little more precise about this, the sample variance would sometimes be smaller and sometimes larger than the population variance. But, on average, over all possible samples, the sample variance is a little too small.)

So, something has to be done about the formula for the sample variance in order to fix this problem of its tendency to be too small as an estimate of the population variance. It turns out, mathematically, that replacing the n with $n - 1$ has exactly the desired effect. Now, when dividing by $n - 1$, rather than n, even though the sample variance is sometimes greater and sometimes less than the population variance, on average, the sample variance is correct as an estimator of the population variance.

- ▪ What does standard deviation measure? How can we summarize what we are attempting to compute?
    - □ *The value of the standard deviation is close to the average distance of observations from the mean. It can be interpreted as a typical deviation from the mean.*
- ▪ How does the spread of the distribution relate to the value of the standard deviation?
    - □ *The larger the spread of the distribution, the larger the standard deviation.*
- ▪ Who can write a formula for standard deviation, s?
    - □ *Encourage students to attempt to write the formula without assistance, perhaps comparing their results with their peers.*

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

In this formula,

- x is a value from the original data set;
- $x - \bar{x}$ is a deviation of the value, x, from the mean, $\bar{x}$;
- $(x - \bar{x})^2$ is a squared deviation from the mean;
- $\sum(x - \bar{x})^2$ is the sum of the squared deviations;
- $\dfrac{\sum(x-\bar{x})^2}{n-1}$ is the result of dividing the sum of the squared deviations by $n - 1$;
- So, $\sqrt{\dfrac{\sum(x-\bar{x})^2}{n-1}}$ is the standard deviation.

Lesson Summary

- **The standard deviation measures a typical deviation from the mean.**
- **To calculate the standard deviation,**
  1. **Find the mean of the data set;**
  2. **Calculate the deviations from the mean;**
  3. **Square the deviations from the mean;**
  4. **Add up the squared deviations;**
  5. **Divide by $n - 1$ (if working with data from a sample, which is the most common case);**
  6. **Take the square root.**
- **The unit of the standard deviation is always the same as the unit of the original data set.**
- **The larger the standard deviation, the greater the spread (variability) of the data set.**
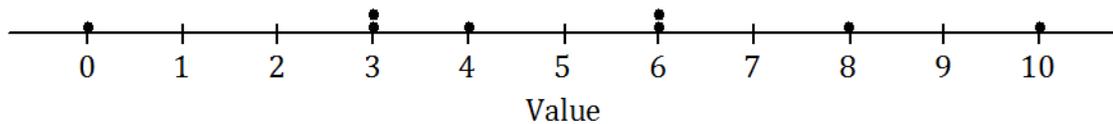

Exit Ticket (10 minutes)

Name _____   Date _____

Lesson 5: Measuring Variability for Symmetrical Distributions
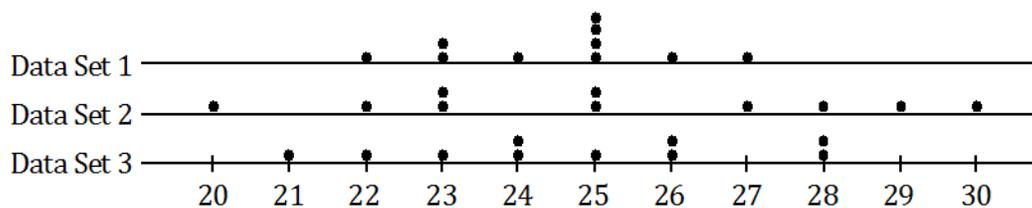
Exit Ticket

1. Look at the dot plot below.

a. *Estimate* the mean of this data set.

b. Remember that the standard deviation measures a typical deviation from the mean. The standard deviation of this data set is either 3.2, 6.2, or 9.2. Which of these values is correct for the standard deviation?

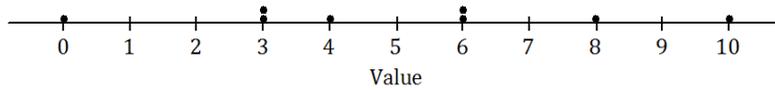2. Three data sets are shown in the dot plots below.

a. Which data set has the smallest standard deviation of the three? Justify your answer.

b. Which data set has the largest standard deviation of the three? Justify your answer.

Exit Ticket Sample Solutions

1. **Look at the dot plot below.**

Value

a. *Estimate* the mean of this data set.

The mean of the data set is 5, so any number above 4 and below 6 would be acceptable as an estimate of the mean.

b. **Remember that the standard deviation measures a typical deviation from the mean. The standard deviation of this data set is either $3.2, 6.2,$ or 9.2. Which of these values is correct for the standard deviation?**

The greatest deviation from the mean is 5 (found by calculating $10 - 5$ or $0 - 5$), and a typical deviation from the mean must be less than 5. So, 3.2 must be chosen as the standard deviation.
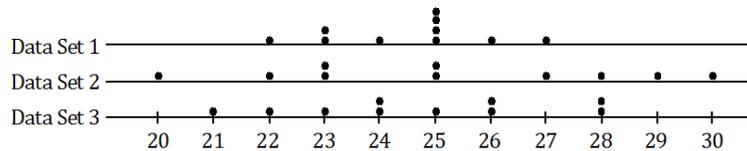
2. **Three data sets are shown in the dot plots below.**

a. **Which data set has the smallest standard deviation of the three? Justify your answer.**

Data Set 1

The points are clustered more closely together in Data Set 1. The deviations or the distances of the points to the left and to the right of the mean position would be smaller for this data set. As a result, the sum of the squares of the distances would also be smaller; thus, the standard deviation would be smaller.

b. **Which data set has the largest standard deviation of the three? Justify your answer.**

Data Set 2

The points of this data set are the most spread out. The deviations or the distances of the points to the left and to the right of the mean position would be larger for this data set. The sum of the squares of the distances would also be larger. As a result, the standard deviation would be larger.

Problem Set Sample Solutions

1. A small car dealership tests the fuel efficiency of sedans on its lot. It chooses 12 sedans for the test. The fuel efficiency (mpg) values of the cars are given in the table below. Complete the table as directed below.

Fuel Efficiency (miles per gallon)	29	35	24	25	21	21	18	28	31	26	26	22
Deviation from the Mean	3.5	9.5	-1.5	-0.5	-4.5	-4.5	-7.5	2.5	5.5	0.5	0.5	-3.5
Squared Deviation from the Mean	12.25	90.25	2.25	0.25	20.25	20.25	56.25	6.25	30.25	0.25	0.25	12.25

a. Calculate the mean fuel efficiency for these cars.

Mean $= 25.5$

b. Calculate the deviations from the mean, and write your answers in the second row of the table.

See table above.

c. Square the deviations from the mean, and write the squared deviations in the third row of the table.

See table above.

d. Find the sum of the squared deviations.

The sum of the squared deviations is 251.

e. What is the value of n for this data set? Divide the sum of the squared deviations by $n - 1$.

$n = 12$

$\dfrac{251}{11} = 22.818$, *to the nearest thousandth.*

f. Take the square root of your answer to part (e) to find the standard deviation of the fuel efficiencies of these cars. Round your answer to the nearest hundredth.

$\sqrt{22.818} = 4.78$

The standard deviation of the fuel efficiencies of these cars is 4.78 *miles per gallon.*

2. The same dealership decides to test fuel efficiency of SUVs. It selects six SUVs on its lot for the test. The fuel efficiencies (in miles per gallon) of these cars are shown below.

$$21 \quad 21 \quad 21 \quad 30 \quad 28 \quad 24$$

Calculate the mean and the standard deviation of these values. Be sure to show your work, and include a unit in your answer.

Mean $= 24.17$ miles per gallon; standard deviation $= 3.97$ miles per gallon

Note: Students might get somewhat varying answers for the standard deviation depending on how far they round the value of the mean. Encourage students, when calculating the standard deviation, to use several decimal places in the value that they use for the mean.

3. Consider the following questions regarding the cars described in Problems 1 and 2.

   a. What is the standard deviation of the fuel efficiencies of the cars in Problem 1? Explain what this value tells you.

      The standard deviation for the cars in Problem 1 is 4.78 miles per gallon. This is a typical deviation from the mean for the fuel efficiencies of the cars in Problem 1.

   b. You also calculated the standard deviation of the fuel efficiencies for the cars in Problem 2. Which of the two data sets (Problem 1 or Problem 2) has the larger standard deviation? What does this tell you about the two types of cars (sedans and SUVs)?

      The standard deviation is greater for the cars in Problem 1. This tells us that there is greater variability in the fuel efficiencies of the cars in Problem 1 (the sedans) than in the fuel efficiencies of the cars in Problem 2 (the SUVs). This means that the fuel efficiency varies more from car to car for sedans than for SUVs.