

## 9.8: Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

### The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

### The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

### The standard deviation can be used to determine whether a data value is close to or far from the mean

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

#### Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

#### Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because  $5 + (1)(2) = 7$ .

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because  $5 + (-2)(2) = 1$ .

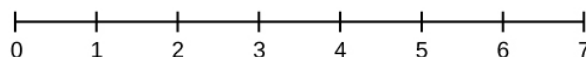


Figure 9.8.1

- In general, a **value** = **mean** + (**#ofSTDEV**)(**standard deviation**)
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because:  $1 = 5 + (-2)(2)$ .

The equation value = mean + (#ofSTDEVs)(standard deviation) can be expressed for a sample and for a population.

- sample:

$$x = \bar{x} + (\text{\#ofSTDEV})(s) \quad (9.8.1)$$

- Population:

$$x = \mu + (\text{\#ofSTDEV})(\sigma) \quad (9.8.2)$$

The lower case letter s represents the sample standard deviation and the Greek letter  $\sigma$  (sigma, lower case) represents the population standard deviation.

The symbol  $\bar{x}$  is the sample mean and the Greek symbol  $\mu$  is the population mean.

### Calculating the Standard Deviation

If  $x$  is a number, then the difference " $x - \text{mean}$ " is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is  $x - \mu$ . For sample data, in symbols a deviation is  $x - \bar{x}$ .

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter  $\sigma$  (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of  $\sigma$ .

To calculate the standard deviation, we need to calculate the variance first. The variance is the **average of the squares of the deviations** (the  $x - \bar{x}$  values for a sample, or the  $x - \mu$  values for a population). The symbol  $\sigma^2$  represents the population variance; the population standard deviation  $\sigma$  is the square root of the population variance. The symbol  $s^2$  represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by  $N$ , the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by  $n - 1$ , one less than the number of items in the sample.

#### Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \quad (9.8.3)$$

or

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} \quad (9.8.4)$$

For the sample standard deviation, the denominator is  $n - 1$ , that is the sample size MINUS 1.

#### Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \quad (9.8.5)$$

or

$$\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}} \quad (9.8.6)$$

For the population standard deviation, the denominator is  $N$ , the number of items in the population.

In Equations 9.8.4 and 9.8.6,  $f$  represents the frequency with which a value appears. For example, if a value appears once,  $f$  is one. If a value appears three times in the data set or population,  $f$  is three.

### Standard Deviation Calculator

Type in the values from the data set separated by commas, for example, 2,4,5,8,11,2, and click Calculate.

- ☐ Population Standard Deviation  
☐ Sample Standard Deviation

**Standard Deviation**

Calculate

### Play the Guess the Standard Deviation Game.

The goal is to guess the standard deviation rounded to the nearest whole number. Try to do this visually without doing any calculations. You score a point if you get it correct, don't lose a point if you are within 1 of the standard deviation, and lose a point if you are off by 2 or more. Enjoy the game!

$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$
--------------	--------------	--------------	--------------

Next	Click Next to start the Game	Score: 0
------	------------------------------	----------

### Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in Section 2.6. How much the statistic varies from one sample to another is known as the sampling variability of a statistic. You typically measure the **sampling variability of a statistic** by its standard error.

The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in Chapter 7. The notation for the standard error of the mean is  $\frac{\sigma}{\sqrt{n}}$  where  $\sigma$  is the standard deviation of the population and  $n$  is the size of the sample.

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation  $\sigma_x$  or  $s_x$  from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you

should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

### Example 9.8.1

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of  $n = 20$  fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525 \quad (9.8.7)$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating  $s$ .

Data	Freq.	Deviations	<i>Deviations</i> <sup>2</sup>	(Freq.)( <i>Deviations</i> <sup>2</sup> )
$x$	$f$	$(x - \bar{x})$	$(x - \bar{x})^2$	$(f)(x - \bar{x})^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times 0.275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times 0.000625 = 0.0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times 0.225625 = 1.35375$
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times 0.950625 = 2.851875$
				The total is 9.7375

The sample variance,  $s^2$ , is equal to the sum of the last column (9.7375) divided by the total number of data values minus one ( $20 - 1$ ):

$$s^2 = \frac{9.7375}{20 - 1} = 0.5125 \quad (9.8.8)$$

The **sample standard deviation**  $s$  is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = 0.715891 \quad (9.8.9)$$

, which is rounded to two decimal places,  $s = 0.72$ .

**Typically, you do the calculation for the standard deviation on your calculator or computer.** The intermediate results are not rounded. This is done for accuracy.

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation on a calculator or computer.
- For a sample:  $x = \bar{x} + (\text{\#ofSTDEVs})(s)$
- For a population:  $x = \mu + (\text{\#ofSTDEVs})\sigma$
- For this example, use  $x = \bar{x} + (\text{\#ofSTDEVs})(s)$  because the data is from a sample
  - Verify the mean and standard deviation on your calculator or computer.
  - Find the value that is one standard deviation above the mean. Find  $(\bar{x} + 1s)$ .
  - Find the value that is two standard deviations below the mean. Find  $(\bar{x} - 2s)$ .
  - Find the values that are 1.5 standard deviations **from** (below and above) the mean.

### Solution

a.

- Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- $\bar{x} = 10.525$
- Use Sx because this is sample data (not a population):  $Sx = 0.715891$

b.  $(\bar{x} + 1s) = 10.53 + (1)(0.72) = 11.25$

c.  $(\bar{x} - 2s) = 10.53 - (2)(0.72) = 9.09$

d. ◦  $(\bar{x} - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$

◦  $(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

### Exercise 2.8.1

On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

### Answer

$$\mu = 30.68$$

$$s = 6.09$$

$$(\bar{x} + 2s = 30.68 + (2)(6.09) = 42.86.$$

### Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is  $-1.525$  for the data value nine. **If you add the deviations, the sum is always zero.** (For Example 9.8.1, there are  $n = 20$  deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by  $n = 20$ , the calculation divided by  $n - 1 = 20 - 1 = 19$  because the data is a sample. For the **sample** variance, we divide by the sample size minus one ( $n - 1$ ). Why not divide by  $n$ ? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by  $(n - 1)$  gives a better estimate of the population variance.

Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number that measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation,  $s$  or  $\sigma$ , is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make  $s$  or  $\sigma$  very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

### Example 9.8.2

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

- Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
- Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
  - The sample mean
  - The sample standard deviation
  - The median
  - The first quartile
  - The third quartile
  - IQR*
- Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

### Answer

- See Table
- The sample mean = 73.5
  - The sample standard deviation = 17.9
  - The median = 73
  - The first quartile = 61
  - The third quartile = 90
  - $IQR = 90 - 61 = 29$
- The  $x$ -axis goes from 32.5 to 100.5;  $y$ -axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is  $(100.5 - 32.5)$  divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5,  $32.5 + 13.6 = 46.1$ ,  $46.1 + 13.6 = 59.7$ ,  $59.7 + 13.6 = 73.3$ ,  $73.3 + 13.6 = 86.9$ ,  $86.9 + 13.6 = 100.5$  = the ending value; No data values fall on an interval boundary.

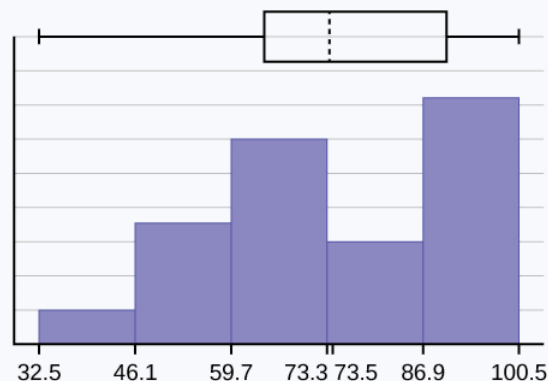
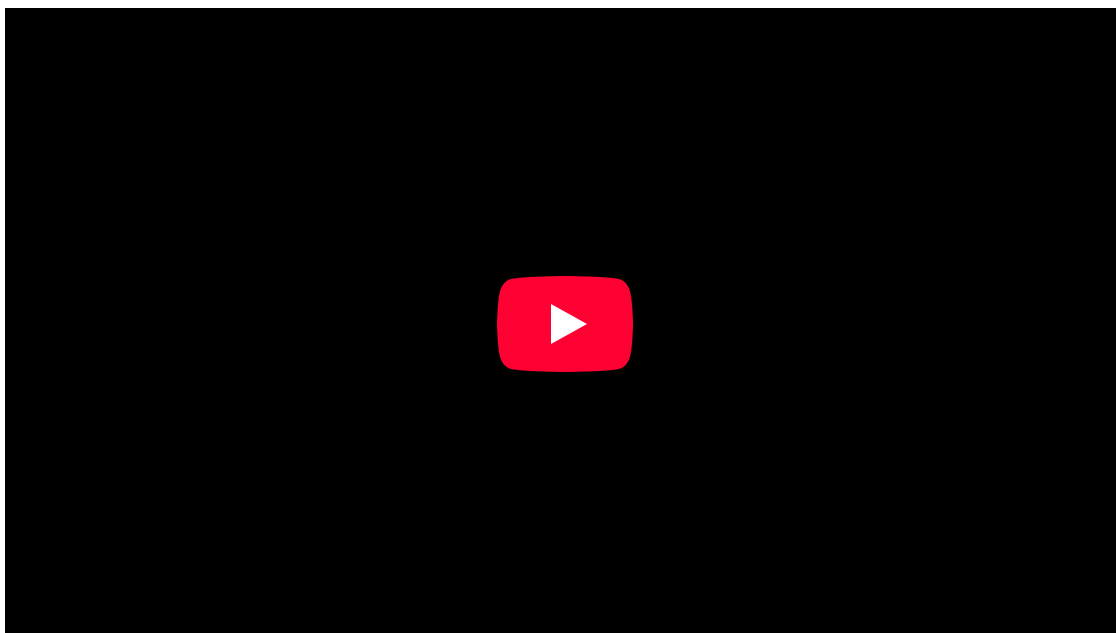


Figure 2.9.2.

The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater ( $73 - 33 = 40$ ) than the spread in the upper 50% ( $100 - 73 = 27$ ). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores ( $IQR = 29$ ) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.



Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

### Exercise 9.8.2

The following data show the different types of pet food stores in the area carry.

6; 6; 6; 6; 7; 7; 7; 7; 8; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;

Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

**Answer**

$\mu = 9.3$  and  $s = 2.2$

### Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \quad (9.8.10)$$

where  $f$  interval frequencies and  $m$  = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how “unusual” individual data is compared to the mean.

### Example 9.8.3

Find the standard deviation for the data in Table 9.8.3.

Table 9.8.3

Class	Frequency, $f$	Midpoint, $m$	$m^2$	$\bar{x}$	$fm^2$	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5



Class	Frequency, $f$	Midpoint, $m$	$m^2$	$\bar{x}$	$fm^2$	Standard Deviation
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

For this data set, we have the mean,  $\bar{x} = 7.58$  and the standard deviation,  $s_x = 3.5$ . This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since  $7.58 - 3.5 - 3.5 = 0.58$ . While the formula for calculating the standard deviation is not complicated,  $s_x = \sqrt{\frac{\sum f(m - \bar{x})^2}{n - 1}}$  where  $s_x$  = sample standard deviation,  $\bar{x}$  = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

Find the standard deviation for the data from the previous example

Class	0-2	3-5	6-8	9-11	12-14	15-17
Frequency, $f$	1	6	10	7	0	2

First, press the **STAT** key and select **1:Edit**

Figure 2.8.3

Input the midpoint values into **L1** and the frequencies into **L2**

Figure 2.8.4

Select **STAT**, **CALC**, and **1: 1-Var Stats**

Figure 2.8.5

Select **2<sup>nd</sup>** then **1** then , **2<sup>nd</sup>** then **2 Enter**

Figure 2.8.6

You will see displayed both a population standard deviation,  $\sigma_x$ , and the sample standard deviation,  $s_x$ .

### Mean and Standard Deviation for grouped frequency Tables Calculator

Type in the midpoints and frequencies below. Put the midpoints in increasing order and do not include any values with zero frequency.

Midpoints

Frequencies

- ☐ Population Standard Deviation  
☐ Sample Standard Deviation

**Mean**

**Standard Deviation**

## Five Point Summary

## Sample Size

Calculate

## Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\text{\#ofSTDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol  $z$ . In symbols, the formulas become:

Sample	$x = \bar{x} + zs$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

## Z-Score Calculator

$x$	$\mu$	$\sigma$
<input type="text"/>	<input type="text"/>	<input type="text"/>

z-score

Enter any three of them, leave the fourth blank, and click Calculate.

Calculate

## Example 9.8.4

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

### Answer

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \# \text{ofSTDEVs} = \left( \frac{\text{value-mean}}{\text{standard deviation}} \right) = \left( \frac{x - \mu}{\sigma} \right)$$

For John,

$$z = \# \text{ofSTDEVs} = \left( \frac{2.85 - 3.0}{0.7} \right) = -0.21$$

For Ali,

$$z = \# \text{ofSTDEVs} = \left( \frac{77 - 80}{10} \right) = -0.3$$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's z-score of  $-0.21$  is higher than Ali's z-score of  $-0.3$ . For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

### Exercise 9.8.4

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

### Answer

For Angie:

$$z = \left( \frac{26.2 - 27.2}{0.8} \right) = -1.25$$

For Beth:

$$z = \left( \frac{27.3 - 30.1}{1.4} \right) = -2$$

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

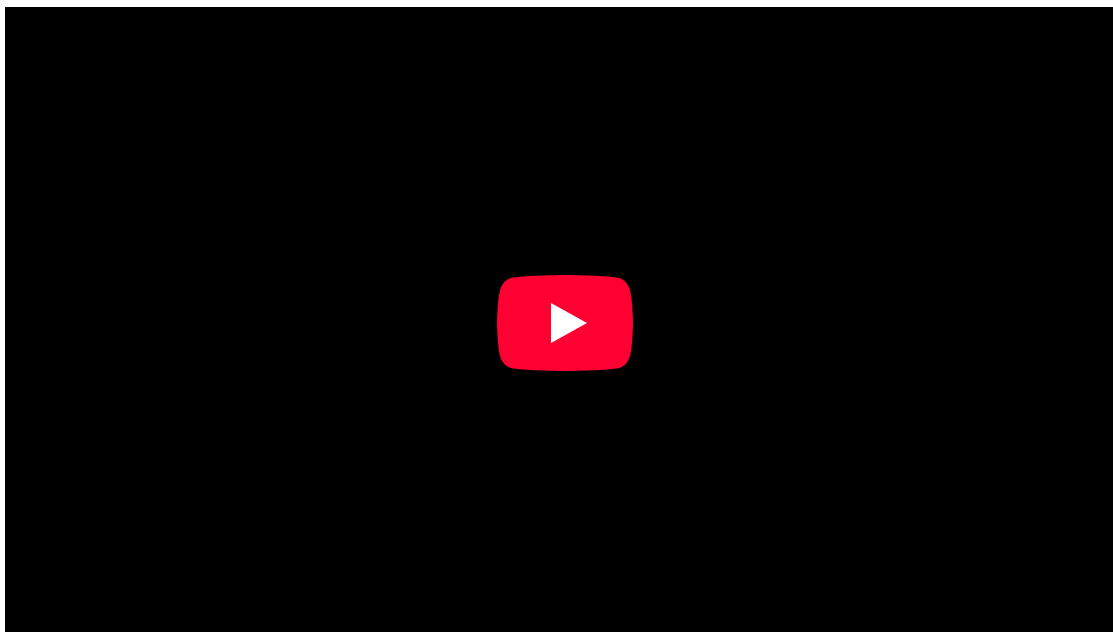
For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as **Chebyshev's Rule**.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.

- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.



## References

- Data from Microsoft Bookshelf.
- King, Bill. "Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at <http://www.ltcc.edu/web/about/institutional-research> (accessed April 3, 2013).

## Chapter Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.

- $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$  or  $s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}}$  is the formula for calculating the standard deviation of a sample. To calculate the

standard deviation of a population, we would use the population mean,  $\mu$ , and the formula  $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$  or

$$\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}}.$$

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} \quad (9.8.11)$$

where  $s_x$  sample standard deviation and  $\bar{x}$  = sample mean

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

#### Exercise 2.8.4

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

**Answer**

$s = 34.5$

#### Exercise 2.8.5

Find the value that is one standard deviation below the mean.

#### Exercise 2.8.6

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

**Answer**

For Fredo:

$$z = \frac{0.158 - 0.166}{0.012} = -0.67$$

For Karl:

$$z = \frac{0.177 - 0.189}{0.015} = -0.8$$

Fredo's z-score of  $-0.67$  is higher than Karl's z-score of  $-0.8$ . For batting average, higher values are better, so Fredo has a better batting average compared to his team.

#### Exercise 2.8.7

Use Table to find the value that is three standard deviations:

- above the mean
- below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

#### Exercise 2.8.5

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3

Grade	Frequency
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

### Answer

$$a. s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$$

$$b. s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$$

$$c. s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$$

## Bringing It Together

### Exercise 2.8.7

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

- Find the sample mean  $\bar{x}$ .
- Find the approximate sample standard deviation,  $s$ .

### Answer

- a. 1.48
- b. 1.12

### Exercise 2.8.8

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let  $X$  = the number of pairs of sneakers owned. The results are as follows:

$X$	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

- a. Find the sample mean  $\bar{x}$
- b. Find the sample standard deviation,  $s$
- c. Construct a histogram of the data.
- d. Complete the columns of the chart.
- e. Find the first quartile.
- f. Find the median.
- g. Find the third quartile.
- h. Construct a box plot of the data.
- i. What percent of the students owned at least five pairs?
- j. Find the 40<sup>th</sup> percentile.
- k. Find the 90<sup>th</sup> percentile.
- l. Construct a line graph of the data
- m. Construct a stemplot of the data

### Exercise 2.8.9


Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

- a. Organize the data from smallest to largest value.
- b. Find the median.
- c. Find the first quartile.
- d. Find the third quartile.
- e. Construct a box plot of the data.
- f. The middle 50% of the weights are from \_\_\_\_\_ to \_\_\_\_\_.
- g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
- h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
- i. Assume the population was the San Francisco 49ers. Find:
  - i. the population mean,  $\mu$ .

- ii. the population standard deviation,  $\sigma$ .
- iii. the weight that is two standard deviations below the mean.
- iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
- j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmitt Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

#### Answer

- a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
- b. 241
- c. 205.5
- d. 272.5
- e.  A box plot with a whisker between 174 and 205.5, a solid line at 205.5, a dashed line at 241, a solid line at 272.5, and a whisker between 272.5 and 302.
- f. 205.5, 272.5
- g. sample
- h. population
- i. i. 236.34  
ii. 37.50  
iii. 161.34  
iv. 0.84 std. dev. below the mean
- j. Young

#### Exercise 2.8.10

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3; 8; -1; 2; 0; 5; -3; 1; -1; 6; 5; -2

- a. What is the mean change score?
- b. What is the standard deviation for this population?
- c. What is the median change score?
- d. Find the change score that is 2.2 standard deviations below the mean.

#### Exercise 2.8.11

Refer to Figure determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

<figure >


 This shows three graphs. The first is a histogram with a mode of 3 and fairly symmetrical distribution between 1 (minimum value) and 5 (maximum value). The second graph is a histogram with peaks at 1 (minimum value) and 5 (maximum value) with 3 having the lowest frequency. The third graph is a box plot. The first whisker extends from 0 to 1. The box begins at the first quartile, 1, and ends at the third quartile, 6. A vertical, dashed line marks the median at 3. The second whisker extends from 6 on.

Figure 2.8.6.

</figure>

- a. The medians for all three graphs are the same.
- b. We cannot determine if any of the means for the three graphs is different.
- c. The standard deviation for graph b is larger than the standard deviation for graph a.
- d. We cannot determine if any of the third quartiles for the three graphs is different.

#### Answer



- a. True
- b. True
- c. True
- d. False

### Exercise 2.8.12

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let  $X$  = the length (in days) of an engineering conference.

- a. Organize the data in a chart.
- b. Find the median, the first quartile, and the third quartile.
- c. Find the 65<sup>th</sup> percentile.
- d. Find the 10<sup>th</sup> percentile.
- e. Construct a box plot of the data.
- f. The middle 50% of the conferences last from \_\_\_\_\_ days to \_\_\_\_\_ days.
- g. Calculate the sample mean of days of engineering conferences.
- h. Calculate the sample standard deviation of days of engineering conferences.
- i. Find the mode.
- j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
- k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

### Exercise 2.8.13

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

- a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
- b. Construct a histogram of the data.
- c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
- d. Calculate the sample mean.
- e. Calculate the sample standard deviation.
- f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

#### Answer

a.	Enrollment	Frequency
	1000-5000	10
	5000-10000	16
	10000-15000	3
	15000-20000	3
	20000-25000	1
	25000-30000	2

- b. Check student's solution.
- c. mode
- d. 8628.74
- e. 6943.88
- f. -0.09

Use the following information to answer the next two exercises.  $X$  = the number of days per week that 100 clients use a particular exercise facility.

$x$	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

#### Exercise 2.8.14

The 80<sup>th</sup> percentile is \_\_\_\_\_

- a. 5
- b. 80
- c. 3
- d. 4

#### Exercise 2.8.15

The number that is 1.5 standard deviations BELOW the mean is approximately \_\_\_\_\_

- a. 0.7
- b. 4.8
- c. -2.8
- d. Cannot be determined

**Answer**

a

#### Exercise 2.8.16

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the Table.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

- a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.

- b. If a data value is identified as an outlier, what should be done about it?
- c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
- d. Do parts a and c of this problem give the same answer?
- e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
- f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

## Glossary

### Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation:  $s$  for sample standard deviation and  $\sigma$  for population standard deviation.

### Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as  $x - \bar{x}$  where  $x$  is a value of the data and  $\bar{x}$  is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

## Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at <http://cnx.org/contents/30189442-699...b91b9de@18.114>.



This page titled [9.8: Measures of the Spread of the Data](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.