3.7: Practice SD Formula and Interpretation

You may or may not understand the importance of calculating and understanding the variation of your 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.

Answering Questions

There are a couple common kinds of questions that standard deviations can answer, in addition being foundational for later statistical analyses. First, a standard deviation helps understand the shape of a distribution. Second, a standard deviation can show if a score is extreme.

Describing the Shape of a Distribution

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. Distributions with small standard deviations have a tall and narrow line graph. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation. Distributions with large standard deviations may have a wide and flat lin graph, or they may be skewed (with the outlier(s) making the standard deviation bigger).

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.

Identifying Extreme Scores

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 3.7.1- Scale from 0 to 7 (CC-BY by Barbara Illowsky & Susan Dean (De Anza College) from OpenStax)

- 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 numbers in parentheses that touch should be multiplied)

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

- sample: $x=\bar{x}+(\#\text{ofSTDEV})\ \backslash\text{times}\ (\text{s})$
- Population: $x=\mu+(\#\text{ofSTDEV})\ \backslash\text{times}\ (\text{s})$

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

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

Calculating the Standard Deviation

If x is a number, then the difference "x – 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 σ (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 σ.

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 σ^2 represents the population variance; the population standard deviation σ 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

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

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

Practice!

Example 3.7.1

In a fifth grade class at a private school, 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 in Table 3.7.1, but first let's talk about the context.

1. Who was the sample? Who could this sample represent (population)?

The sample is the 20 fifth graders from a private school. The population could be all fifth graders from private schools?

2. What was measured?

Age, in years, was measured. This is the DV, the outcome variable.

Table 3.7.1- Ages of a sample of 20 fifth graders

Ages of Sample Fifth Graders
9
9.5
9.5
10
10
10
10
10.5
10.5
10.5
10.5
11
11
11
11
11
11
11.5
11.5
11.5

3. What is the mean?

$$\bar{x} = \frac{(9+9.5+9.5+10+10+10+10+10.5+10.5+10.5+10.5+11+11+11+11+11+11+11.5+11.5+11.5)}{20} = 10.525 = 10.53$$

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

4. What is the standard deviation?

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.

Table 3.7.1- Ages of One Fifth Grade Class

Data	Deviations	*Deviations*^{2}
x	$(X - \bar{X})$	$(X - \bar{X})^2$
9	$9 - 10.525 = -1.525$	$(-1.525)^2 = (-1.525 \times -1.525) = 2.325625$
9.5	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = (-1.025 \times -1.025) = 1.050625$
9.5	$9.5 - 10.525 = -1.0.25$	$(-1.025)^2 = 1.050625$
10	$10 - 10.525 = -0.525$	$(-0.525)^2 = (-0.525 \times -0.525) = 0.275625$
10	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$
10	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$
10	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$
10.5	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = (-0.025 \times -0.025) = 0.000625$
10.5	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$
10.5	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$
10.5	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = (0.475 \times 0.475) = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$
11.5	$11.5 - 10.525 = 0.975$	$(0.975)^2 = (0.975 \times 0.975) = 0.950625$
11.5	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$
11.5	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$
$\sum X$	**0 (basically)**	$\sum =$ **9.7375**

The first column in Table 3.7.1 has the data, the second column has has deviations (each score minus the mean), the third column has deviations squared. The first row is the row's title, the second row is the symbols for that column, the rest of the rows are the scores until the bottom row, which is the sum of each of the rows.

Take the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$$\frac{9.7375}{20-1} = 0.5125$$

The **sample standard deviation** s is the square root of $\frac{SS}{df}$:

$$s = \sqrt{0.5125} = 0.715891$$

and this is rounded to two decimal places, $s = 0.72$. The standard deviation of the sample fo 20 fifth graders is 0.72 years.

Typically, you do the calculation for the standard deviation on your calculator or computer. When calculations are completed on devices, the intermediate results are not rounded so the results are more accurate. It's also darned easier. So why are spending time learning this outdated formula? So that you can see what's happening. We are finding the difference between *each score* and the mean to see how varied the distribution of data is around the center, dividing it by the sample size minus

one to make it like an average, then square rooting it to get the final answer back into the units that we started with (age in years).

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**.
- For a sample: $x = \bar{x}$ + (#ofSTDEVs)(s)
- For a population: $x = \mu$ + (#ofSTDEVs)σ
- For this example, use $x = \bar{x}$ + (#ofSTDEVs)(s) because the data is from a sample

5. Verify the mean and standard deviation on your own.
6. Find the value that is one standard deviation above the mean. Find ($\bar{x}$ + 1s).
7. Find the value that is two standard deviations below the mean. Find ($\bar{x} - 2$s).
8. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

a. You should get something close to 0.72 years, but anything from 0.70 to 0.74 shows that you have the general idea.
b. $(\bar{x} + 1s) = 10.53 + (1)(0.72) = 11.25$
c. $(\bar{x} - 2s) = 10.53 - (2)(0.72) = 9.09$
d.
   - $(\bar{x} - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$
   - $(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

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, we divide by the sample size minus one ($n - 1$). The sample variance is an estimate of the population variance. After countless replications, it turns out that when the formula division by only N (the size of the sample) is used on a sample to infer the population's variance, it always under-estimates the variance of the population.

Which one has the bigger solution, the one with the smaller denominator or the larger denominator?

- $\frac{10}{2} =$
- $\frac{10}{5} =$

Smaller denominators make the resulting product **larger.** To solve our problem of using the population's variance formula on a sample under-estimating the variance, we make the denominator of our equation smaller when calculating variance for a sample. In other words, based on the mathematics that lies behind these calculations, dividing by ($n - 1$) gives a better estimate of the population.

What does it mean?

The deviations show how spread out the data are about the mean. From Table 3.7.1, 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 (age, in this case) is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean (that particular student is younger than the average age of the class) . The deviation is –1.525 for the data value nine. If you add the deviations, the sum is always zero, 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. But the variance is a squared measure and does not have the same units as the data. No one knows what 9.7375 years squared *means*. Taking the square root solves the problem! The standard deviation measures the spread in the same units as the data.

The standard deviation, s or σ, 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 σ very large.

? Exercise 3.7.1

Scenario: Using one baseball professional team as a sample for all professional baseball teams, the ages of each of the players are shown in Table 3.7.2.

Table 3.7.2- One Baseball Team's Ages

Data	Deviations	*Deviations*^{2}
x	$(x - \bar{x})$	$(x - \bar{x})^2$
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		
$\sum X$ = 767	$\sum X$ should be 0 (basically)	$\sum X$ = ?

If you get stuck after the table, don't forget that: $s = \sqrt{\frac{\sum(X - \overline{X})^2}{N - 1}}$

All of your answers should be complete sentences, not just one word or one number. Behavioral statistics is about research, not math.

1. Who was the sample? Who could this sample represent (population)?
2. What was measured?
3. What is the mean? (Get in the practice of including the units of measurement when answering questions; a number is usually not a complete answer).

4. What is the standard deviation?

$$s = \sqrt{\frac{\sum(X - \overline{X})^2}{N-1}} = \sqrt{\frac{SS}{df}}$$

5. Find the value that is two standard deviations above the mean, and determine if there are any players that are more than two standard deviations above the mean.

Answer

1. The sample is 25 players from a professional baseball team. They were chosen to represent all professional baseball players (it says so in the scenario description!).
2. Age, in years, was measured.
3. The mean of the sample ($\bar{X}$ was 30.68 years.
4. The standard deviation was 6.09 years ($s = 6.09$), although due to rounding differences you could get something from about 6.05 to 6.12. Don't worry too much if you don't get exactly 6.09; if you are close, then you did the formula correctly!
5. The age that is two standard deviations above the mean is 42.86 years, and none of the players are older than that.

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

.

What standard deviation show us can seem unclear at first. Especially when you are unfamiliar (and maybe nervous) about using the formula. By graphing your data, you can get a better "feel" for what a standard deviation can show you. You will find that in symmetrical distributions, the standard deviation can be very helpful. Because numbers can be confusing, **always graph your data**.

Summary

The standard deviation can help you calculate the spread of data.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$ is the formula for calculating the standard deviation of a sample.

Contributors and Attributions

- 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.
- Dr. MO (Taft College)
