Measures of dispersion: Range, variance, and standard deviation

Recently, you learned that **measures of dispersion** let you describe the spread of your dataset, or the amount of variation in your data values. Measures of dispersion like standard deviation can give you an initial understanding of the distribution of your data, and help you determine what statistical methods to apply to your data.

In this reading, you'll learn more about three measures of dispersion: the range, variance, and standard deviation. This reading focuses on the foundational concept of standard deviation. As a data professional, you'll frequently calculate the standard deviation of your data, and use standard deviation as part of more complex data analysis.

Measures of dispersion

Let's examine out the definition of each measure of dispersion: the range, variance, and standard deviation.

Range

The **range** is the difference between the largest and smallest value in a dataset.

For example, imagine you're a biology teacher and you have data on scores for the final exam. The highest score is 99/100, or 99%. The lowest score is 62/100, or 62%. To calculate the range, subtract the lowest score from the highest score.

99 - 62 = 37

The range is 37 percentage points.

The range is a useful metric because it's easy to calculate, and it gives you a very quick understanding of the overall spread of your dataset.

Variance

Another measure of spread is called the **variance,** which is the average of the squared difference of each data point from the mean. Basically, it's the square of the standard deviation. You'll learn more about variance and how to use it in a later course.

Standard deviation

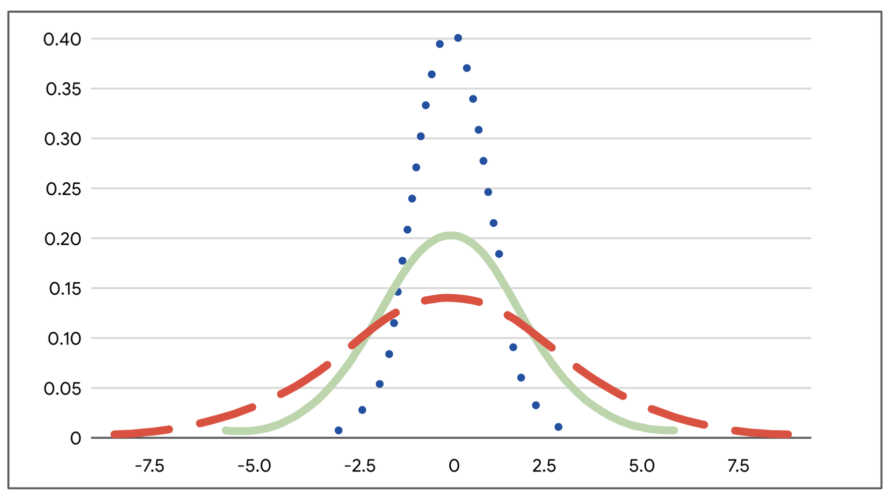
To get a better understanding of the concept of standard deviation, let's explore its definition, visualization, and statistical formula.

Definition

Standard deviation measures how spread out your values are from the mean of your dataset. It calculates the typical distance of a data point from the mean. The larger the standard deviation, the more spread out your values are from the mean. The smaller the standard deviation, the less spread out your values are from the mean.

Visualization

Let's examine the plots of three normal probability distributions to get a better idea of spread. Later on, you'll learn about distributions, which map all the values in a dataset. For now, just know that the mean is the highest point on each curve, right in the center.



Each curve has the same mean and a different standard deviation. The standard deviation of the blue dotted curve is 1, the green solid curve is 2, and the red dashed curve is 3. The blue dotted curve has the least spread since most of its data values fall close to the mean. Therefore, the blue dotted curve has the smallest standard deviation. The red dashed curve has the most spread since most of its data values fall farther away from the mean. Therefore, the red dashed curve has the largest standard deviation.

Formula

Now let's discuss how you calculate standard deviation using a formula.

There are different formulas to calculate the standard deviation for a population and a sample. As a reminder, data professionals typically work with sample data, and they make inferences about populations based on the sample. So, let's review the formula for sample standard deviation:

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

In the formula, n is the total number of data values in your sample, x is each individual data value, and x̄ (pronounced "x-bar") is the mean of your data values. The Greek letter Sigma is a symbol that means sum.

Note: As a data professional, you'll typically use a computer for calculations. Being able to perform calculations is important for your future career, but being familiar with the concepts behind the calculations will help you apply statistical methods to workplace problems.

To better understand the different parts of the formula, let's calculate the sample standard deviation of a small dataset: 2, 3, 10.

You can do this in five steps:

1. Calculate the mean, or average, of your data values.

(2 + 3 +10) ÷ 3 = 15 ÷ 3 = 5

2. Subtract the mean from each value.

2 - 5 = -3

3 - 5 = -2

10 - 5 = 5

3. Square each result.

-3 * -3 = 9

-2 * -2 = 4

5 * 5 = 25

4. Add up the squared results and divide this sum by one less than the number of data values. This is the variance.

(9 + 4 + 25) ÷ (3 -1) = 38 ÷ 2 = 19

5. Finally, find the square root of the variance.

√19 = 4.36

The sample standard deviation is 4.36.

Now that you know more about the concept of standard deviation, let's check out an example of its practical application.

Example: Real estate prices

Imagine you're a data professional working for a real estate company. The real estate agents on your team like to inform their clients about the variation in rental prices in different residential areas. Part of your job is calculating the standard deviation of monthly rental prices for apartments in specific neighborhoods, and sharing this information with your team. Let's say you have sample data on monthly rental prices for one-bedroom apartments in two different neighborhoods: Emerald Woods and Rock Park. Assume you calculate the mean and standard deviation for each dataset.

Emerald Woods

Apartment	#1	#2	#3	#4	#5
Monthly Rent	$900	$950	$1,000	$1,050	$1,100

Mean: $1,000

Standard deviation: $79.05

Rock Park

Apartment	#1	#2	#3	#4	#5
Monthly Rent	$500	$650	$1,000	$1,350	$1,500

Mean: $1,000

Standard deviation: $431.56

Both neighborhoods have the same mean rental price of $1,000 per month. However, the standard deviation for rental prices in Rock Park ($431.56) is much higher than the standard deviation for rental prices in Emerald Woods ($79.05). This means that there is a lot more variation in rental prices in Rock Park. This is useful information for your agents. For example, they can tell clients that it may be easier for them to find a more affordable apartment in Rock Park that is far below the mean of $1,000. Standard deviation helps you quickly understand the variation in prices in any given neighborhood.

Key takeaways

Data professionals use standard deviation to measure variation in many types of data like ad revenues, stock prices, employee salaries, and more. Measures of dispersion like the standard deviation, variance, and range let you quickly identify the variation in your data values, and get a better understanding of the basic structure of your data.

Resources for more information

To learn more about measures of dispersion like the range, variance, and standard deviation, explore the following resources:

- This article from Statistics Canada provides a helpful summary of variance and standard deviation, and discusses the usefulness of standard deviation as a measure of dispersion.

Mark as completed