

LESSON 6: MEASURES OF SPREAD OR VARIATION **(SECTION 3-3)**

PART A: THREE MEASURES

Example 1 (same as in [Lesson 5](#))

The five students in a class take a test. Their scores in points are as follows:

80 76 100 83 100

Let's look at three possibilities for measuring the spread or variation of a data set.

Note: All three of these measures are nonnegative in value.

1) Range

$$\begin{aligned}\text{Range} &= \text{Max} - \text{Min} \\ &= \text{highest value} - \text{lowest value (in the data set)}\end{aligned}$$

In Example 1

$$\begin{aligned}\text{Range} &= \text{Max} - \text{Min} \\ &= 100 - 76 \\ &= 24 \text{ points}\end{aligned}$$

Pros: The range is quick and easy to find, and it seems like a natural measure of spread.

Cons: The range uses only two of the data values (excluding ties), and it is extremely sensitive to outliers.

We will focus on the following, which use all of the data values and are used in many formulas. They are much harder to compute manually, though:

2) Variance (VAR)

3) Standard Deviation (SD)

$$SD = \sqrt{VAR} .$$

This is the most commonly used measure of spread.

PART B: NOTATION

We typically use Greek letters to denote population parameters.

σ is lowercase sigma; remember that the summation operator Σ is uppercase sigma.

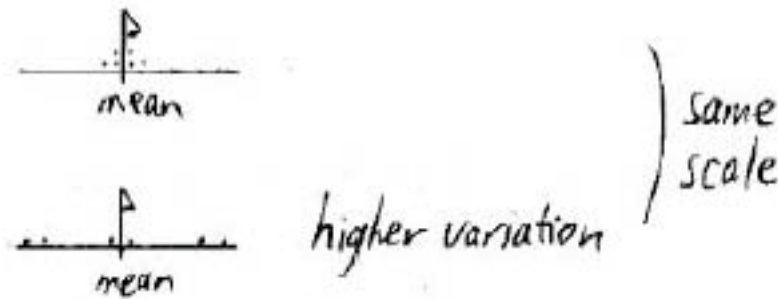
	<u>Mean</u>	<u>SD</u>	<u>VAR</u>
Population (Size N)	μ	σ	σ^2
↓			
Sample (Size n)	\bar{x}	s	s^2

PART C: POPULATION DATA

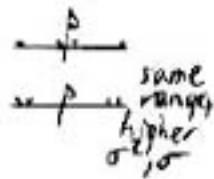
What are the population variance, σ^2 , and the population standard deviation, σ , of a population data set?

Idea

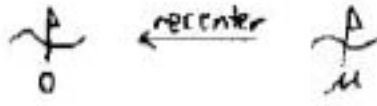
Focus on the mean as a reference point; envision planting a flag there on the real number line. We want to measure the spread of the data values around the mean.



Let's compare another pair of data sets using the same scale:



RecipeGiven data: x_1, x_2, \dots, x_N .

Steps	Notation
Step 1) Find the mean.	$\mu = \frac{\sum x}{N}$
Step 2) Find the deviations from the mean by subtracting the mean from all the data values. 	$(x - \mu)$ values
Step 3) Square the deviations from Step 2).	$(x - \mu)^2$ values
Step 4) VAR, or σ^2 = the average of the squared deviations from Step 3)	$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$
Step 5) SD, or $\sigma = \sqrt{\text{VAR}}$	$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$

Back to Example 1

Find the population VAR and SD of the given data set.

Show all work on exams!

Data (x)	Step 2 Deviations: $(x - \mu)$ values	Step 3 Squared Deviations: $(x - \mu)^2$ values
80	-7.8	60.84
76	-11.8	139.24
100	12.2	148.84
83	-4.8	23.04
100	12.2	148.84
Step 1: $\mu = 87.8$ points	See Note 2 below.	Sum = 520.8 Do Steps 4, 5.

Note 1: You should fill out the above table row by row. For example, take the “80”, subtract off the mean, and then square the result:

$$80 \xrightarrow{\text{Subtract } 87.8} -7.8 \xrightarrow{\text{Square}} 60.84$$

Note 2: Why not use the sum or average of the deviations as a measure of spread? It is always 0 for any data set, so it is meaningless as a measure of spread. The deviations effectively cancel each other out. This reflects the fact that the new mean of our recentered data set is 0. For this and for other theoretical reasons, we square the deviations before taking an average. What if we were to take the absolute values of the deviations before taking an average? The result would be the mean absolute deviation (MAD), which is discussed on [pp.102-103 of Triola](#).

Note 3 on Rounding: We were fortunate that exact values were easily written in the table. See [Notes 3.02](#) for a reminder of our rounding rules for [Chapter 3](#).

Step 4:

$$\begin{aligned}\text{VAR, or } \sigma^2 &= \text{the average of the squared deviations} \\ &= \frac{520.8}{5} \\ &= 104.16 \text{ square points}\end{aligned}$$

One reason why we often prefer the SD over the VAR is that units like “square points” are not natural to us.

Step 5:

$$\begin{aligned}\text{SD, or } \sigma &= \sqrt{\text{VAR}} \\ &= \sqrt{104.16} \\ &\approx 10.2 \text{ points}\end{aligned}$$

Observe that the SD shares the same units as the original data values.

Many calculators have a σ or σ_N button that allows you to compute the population SD of inputted data.

PART D: SAMPLE DATA

What are the sample variance, s^2 , and the sample standard deviation, s , of a sample data set?

We want the sample variance to estimate the population variance of the population from which the sample was drawn. We will use modified versions of the formula and the recipe in [Part C](#) to compute s^2 and then s .

	<u>Mean</u>	<u>SD</u>	<u>VAR</u>
Population (Size N)	μ	σ	σ^2
↓			
Sample (Size n)	\bar{x}	s	s^2

The formula for sample variance is given by:

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

The formula for sample SD is then given by:

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

Note 1: How does this differ from the formula for σ ? The population mean, μ , is presumably unknown, so we replace it with the sample mean, \bar{x} . We also replace the population size, N , with $n - 1$. But

Note 2: Why $n - 1$, not n ? Why is s the square root of a “tilted” average of the squared deviations from the sample mean, \bar{x} ? The appropriate reference point is still the population mean, μ , not \bar{x} . The sample data values are more naturally clustered around their sample mean than around the population mean. In order to make s^2 a better estimate for σ^2 , the population variance, we inflate our estimate by dividing by $n - 1$ instead of n . (Remember, for example: $\frac{1}{4} > \frac{1}{5}$). Then, s^2 will be an unbiased estimator of σ^2 in that it does not have an automatic tendency to consistently over- or underestimate σ^2 .

Note 3: [Triola on p.94](#) provides the following alternate formula, which is much harder to remember but is more often used in computers and calculators:

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

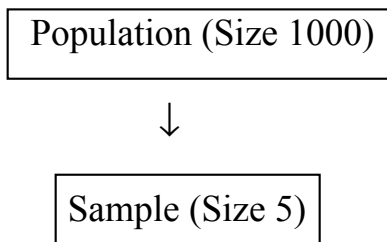
This formula is algebraically equivalent to the previous one. It is a one-pass formula in that a computer or calculator does not need to use the data values more than once. In particular, the sample mean does not have to be computed first. (The previous formula is a two-pass formula.) This latter formula is preferred for maximum accuracy in that it does more to avoid roundoff errors.

Example 2

We will modify Example 1. Let's say 1000 students in a large lecture class have taken a test. Five of the tests are randomly selected, and their scores are as follows:

80 76 100 83 100

Find the sample VAR and SD of the given data set.



Solution to Example 2

Observe that these are the same five scores as in Example 1, but we are treating them as sample data this time. The computations in the table are exactly the same as in Example 1, although we now use \bar{x} instead of μ to denote the mean of the data.

Data (x)	Step 2 Deviations: $(x - \bar{x})$ values	Step 3 Squared Deviations: $(x - \bar{x})^2$ values
80	-7.8	60.84
76	-11.8	139.24
100	12.2	148.84
83	-4.8	23.04
100	12.2	148.84
Step 1: $\bar{x} = 87.8$ points		Sum = 520.8 Do Steps 4, 5.

Step 4:

$$\begin{aligned} \text{VAR, or } s^2 &= \text{the "tilted" average of the squared deviations} \\ &= \frac{520.8}{4} \\ &= 130.2 \text{ square points} \end{aligned}$$

Here, we differ from Example 1 in that the sum from Column 3, 520.8, is divided not by 5, but by 4 (i.e., $n - 1$).

Step 5:

$$\begin{aligned} \text{SD, or } \sigma &= \sqrt{\text{VAR}} \\ &= \sqrt{130.2} \\ &\approx 11.4 \text{ points} \end{aligned}$$

Observe that this is higher than 10.2 points, the SD from Example 1. The “tilted” average we used in Step 4 inflated our value for the sample SD.

Many calculators have an s or σ_{n-1} button that allows you to compute the sample SD of inputted data.

PART E: APPLICATIONS

If the population SD of a data set is 10.2 points, what is the usefulness of that?

If we have different populations (for example, men and women) for which the same measure (such as age or height using the same units) is being taken, then we could compare their SDs.

In [Parts F, G, and H](#), we will use the SD to provide information about how data values are distributed. To avoid confusion, let’s say we’re dealing with the population SD of a population data set.

PART F: CHEBYSHEV’S THEOREM

Chebyshev’s Theorem

Let k be a real number such that $k > 1$.

What fraction of the values in a data set must lie within k SDs of the mean?

The answer is: at least $1 - \frac{1}{k^2}$.

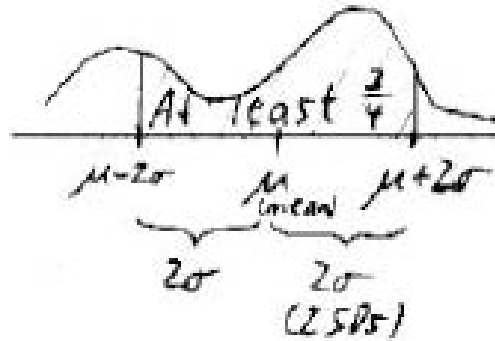
Pro: This theorem applies to **any** distribution shape and is thus distribution free.

Con: $1 - \frac{1}{k^2}$ might not be close to our desired fraction; it only provides a lower bound on what the desired fraction could be.

Case ($k = 2$)

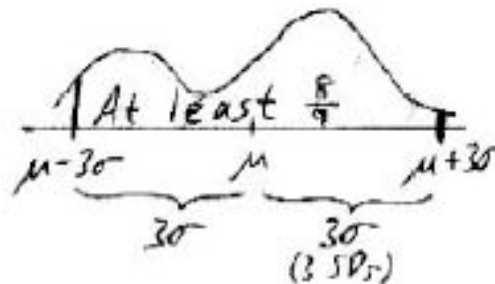
$$\begin{aligned}
 1 - \frac{1}{k^2} &= 1 - \frac{1}{(2)^2} \\
 &= 1 - \frac{1}{4} \\
 &= \frac{3}{4}
 \end{aligned}$$

Therefore, at least $\frac{3}{4}$ (i.e., at least 75%) of the data values must lie within two SDs of the mean.

Case ($k = 3$)

$$\begin{aligned}
 1 - \frac{1}{k^2} &= 1 - \frac{1}{(3)^2} \\
 &= 1 - \frac{1}{9} \\
 &= \frac{8}{9}
 \end{aligned}$$

Therefore, at least $\frac{8}{9}$ (i.e., at least 88.8%) of the data values must lie within three SDs of the mean.



PART H: RANGE RULE OF THUMB

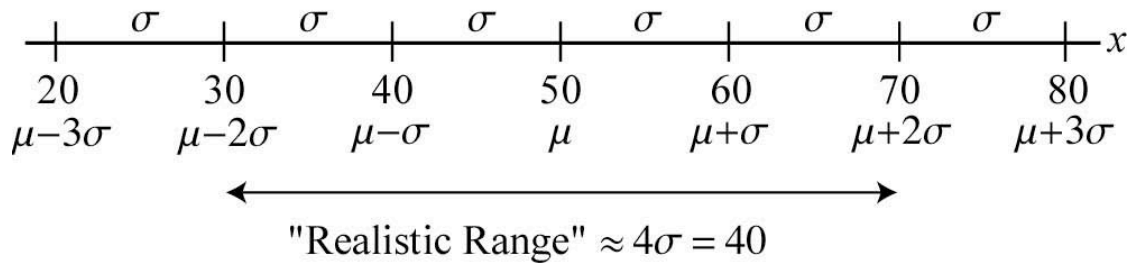
Both Chebyshev's Theorem and the Empirical Rule imply that, in terms of SDs, we can't have too much of the data too far away from the mean. In particular, the "vast majority" of the data must lie within two SDs of the mean. (We could also use three SDs, for example.)

Range Rule of Thumb for Interpreting σ

The vast majority of the values in a population data set will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.

Example 3 revisited

The scores on a test have mean $\mu = 50$ points and SD $\sigma = 10$ points.



The "vast majority" of the scores must lie between 30 and 70 points.

Let's say that the "realistic range" of the data set is $70 - 30 = 40$ points. The idea is that few of the scores will be less than 30 points or greater than 70 points.

In general, Realistic range $\approx 4\sigma$. As a result, we obtain:

Range Rule of Thumb for Estimating σ

$$\sigma \approx \frac{\text{Realistic range}}{4}$$

Example 4

We believe that the “vast majority” of textbooks in a campus bookstore have prices between \$20 and \$180. Estimate σ , the SD of textbook prices in the bookstore.

Solution to Example 4

$$\begin{aligned}\sigma &\approx \frac{\text{Realistic range}}{4} \\ \sigma &\approx \frac{\$180 - \$20}{4} \\ \sigma &\approx \frac{\$160}{4} \\ \sigma &\approx \$40\end{aligned}$$

Again, this is a very rough estimate. We at least expect it to be on the correct order of magnitude, as opposed to, say, \$4 or \$400.

PART I: INVESTING

The benefit of a diverse portfolio is that risk is spread out across many stocks. No single stock will destroy you. Your hope is that your stock portfolio will at least keep up with the general upward trend of the market over time.

[See the margin essay “More Stocks, Less Risk” on p.99.](#)