LESSON 9: MEASURES OF VARIATION

How Spread Out is the Data?

PART A: FOUR MEASURES OF SPREAD

Here is a list of 100 test scores. How would you measure the spread of the scores?

70  40  60  93  86  87  13  86  66  80
71  65  95  75  69  47  72  100 39  99
90  63  81  88  82  51  89  58  46  95
50  15  81  52  47  76  90  70  89  44
46  53  46  52  37  38  54  64  70  40
94  60  63  56  92  92  63  86  56  87
56  51  83  79  91  87  53  97  84  49
91  81  69  82  77  91  71  90  51  92
68  88  38  84  54  47  50  14  48  29
61  97  79  36  52  93  46  92  28  38

Let’s make things easier. Let’s go from 100 scores to five.

Example 1 (Five Test Scores)

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

80  76  100  83  100

How can we find a single number that tells us how spread out the scores are?

Let’s look at four possibilities for measuring the spread (or variation) of a quantitative data set.

WARNING 1: Nonnegativity. All four measures are nonnegative in value. If you ever get a negative value, you have made a mistake!
1) **Range**

\[ \text{Range} = \text{Max} - \text{Min} \]

= highest value – lowest value \( \text{in the data set} \)

In Example 1

\[ \text{Range} = \text{Max} - \text{Min} \]

\[ = 100 - 76 \]

\[ = 24 \text{ points} \]

**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.

The next two measures … 2) and 3) below … 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)**

\[ \text{SD} = \sqrt{\text{VAR}} \]

This is the most commonly used measure of spread.

And finally …

4) **Interquartile Range (IQR)**

We will discuss this in Lesson 10.
PART B: NOTATION

- Let $N$ = the population size for a population data set.

- Let $n$ = the sample size for a sample data set.

- We typically use Greek letters to denote population parameters.

- We typically use Roman / English letters to denote sample statistics.

- $\sigma$ is lowercase sigma. The summation operator $\sum$ is uppercase sigma.

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>VAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population (Size $N$)</td>
<td>$\mu$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>$\downarrow$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample (Size $n$)</td>
<td>$\bar{x}$</td>
<td>$s$</td>
</tr>
</tbody>
</table>

- The purpose of the sample mean ($\bar{x}$) is really to estimate the population mean, $\mu$.

- Likewise, the purpose of the sample SD ($s$) is really to estimate the population SD, $\sigma$. The purpose of the sample VAR ($s^2$) is really to estimate the population variance, ($\sigma^2$).
PART C: POPULATION DATA

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

Idea

Focus on the mean as a reference point; imagine 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:

(same range; higher $\sigma^2$, $\sigma$)
### Recipe

Given data: $x_1, x_2, \ldots, x_N$.

<table>
<thead>
<tr>
<th>Steps</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1) Find the population <strong>mean</strong>.</td>
<td>$\mu = \frac{\sum x}{N}$</td>
</tr>
<tr>
<td>Step 2) Find the <strong>deviations from the mean</strong> by subtracting the mean from all the data values.</td>
<td>$(x - \mu)$ values</td>
</tr>
<tr>
<td>Step 3) <strong>Square</strong> the deviations from Step 2).</td>
<td>$(x - \mu)^2$ values</td>
</tr>
<tr>
<td>Step 4) <strong>VAR</strong>, or $\sigma^2 =$ the <strong>average</strong> of the squared deviations from Step 3)</td>
<td>$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$</td>
</tr>
<tr>
<td>Step 5) <strong>SD</strong>, or $\sigma = \sqrt{\text{VAR}}$</td>
<td>$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$</td>
</tr>
</tbody>
</table>
Back to Example 1 (Five Test Scores)

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

<table>
<thead>
<tr>
<th>Data</th>
<th>Step 2</th>
<th>Step 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x) values</td>
<td>Deviations: (x - \mu) values</td>
<td>Squared Deviations: ((x - \mu)^2) values</td>
</tr>
<tr>
<td>80</td>
<td>-7.8</td>
<td>60.84</td>
</tr>
<tr>
<td>76</td>
<td>-11.8</td>
<td>139.24</td>
</tr>
<tr>
<td>100</td>
<td>12.2</td>
<td>148.84</td>
</tr>
<tr>
<td>83</td>
<td>-4.8</td>
<td>23.04</td>
</tr>
<tr>
<td>100</td>
<td>12.2</td>
<td>148.84</td>
</tr>
<tr>
<td>(\text{Step 1: } \mu = 87.8) points</td>
<td></td>
<td>Sum = 520.8</td>
</tr>
<tr>
<td></td>
<td>See Note 2 below.</td>
<td>Do Steps 4, 5.</td>
</tr>
</tbody>
</table>

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

\[
\begin{align*}
80 & \quad \text{Subtract 87.8} \quad -7.8 \quad \text{Square} \quad 60.84 \\
\end{align*}
\]

Note 2: \textit{Check} that the \textit{squared deviations} under \textit{Step 3} are \textit{nonnegative}.

Note 3: \textbf{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 \textbf{square the deviations before taking an average}. What if we were to take the \textit{absolute values} of the deviations before taking an average? The result would be the \textit{mean absolute deviation (MAD)}, which is not as commonly used.

Note 4 on Rounding: We were fortunate that exact values were easily written in the table. See Lesson 7, page 7.02 on rounding.
Step 4:

VAR, or $\sigma^2 = \text{the average of the squared deviations}$

\[
\frac{520.8}{5} = 104.16 \text{ square points}
\]

- The **population VAR** of the data is about 104.16 square points.

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

Step 5:

SD, or $\sigma = \sqrt{\text{VAR}}$

\[
\sqrt{104.16} \approx 10.2 \text{ points}
\]

- The **population SD** of the data is about 10.2 points.

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

Many **calculators** have a $\sigma$ or $\sigma_N$ button that allows you to compute the **population SD** of inputed 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 \( s \).

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:

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

How and why do these formulas differ from the formulas for \( \sigma^2 \) and \( \sigma \)?

- The population mean, \( \mu \), is presumably unknown, so we replace it with the sample mean, \( \bar{x} \).

- We also replace the population size, \( N \), with \( n - 1 \). 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, \( \mu \), not \( x \). 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 \( \sigma^2 \), the population variance, we inflate our estimate by dividing by \( n - 1 \) instead of \( n \).

\[
\left( \text{Remember, for example: } \frac{1}{4} > \frac{1}{5} \right)
\]

Then, \( s^2 \) will be an unbiased estimator of \( \sigma^2 \); it does not tend to consistently over- or underestimate \( \sigma^2 \).
Example 2 (Example 1 Data as Sample Data)

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.

<table>
<thead>
<tr>
<th>Data (x) values</th>
<th>Step 2 Deviations: (x - \bar{x}) values</th>
<th>Step 3 Squared Deviations: (x - \bar{x})^2 values</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>-7.8</td>
<td>60.84</td>
</tr>
<tr>
<td>76</td>
<td>-11.8</td>
<td>139.24</td>
</tr>
<tr>
<td>100</td>
<td>12.2</td>
<td>148.84</td>
</tr>
<tr>
<td>83</td>
<td>-4.8</td>
<td>23.04</td>
</tr>
<tr>
<td>100</td>
<td>12.2</td>
<td>148.84</td>
</tr>
<tr>
<td><strong>Step 1:</strong> ( \bar{x} = 87.8 ) points</td>
<td>Sum = 520.8</td>
<td>\text{Do Steps 4, 5.}</td>
</tr>
</tbody>
</table>

§ Solution

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 \( \mu \) to denote the mean of the data.
Step 4:

VAR, or $s^2$ = the "tilted" average of the squared deviations

$$= \frac{520.8}{4}$$

$$= 130.2 \text{ square points}$$

• The sample **VAR** of the data is 130.2 square points.

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

Step 5:

SD, or $s = \sqrt{\text{VAR}}$

$$= \sqrt{130.2}$$

$$\approx 11.4 \text{ points}$$

• The sample **SD** of the data is about 11.4 points.

• Observe that this is higher than 10.2 points, the population SD for the five students from Example 1. The “tilted” average we used in Step 4 here inflated our value for the sample SD to better estimate the population SD ($\sigma$) for the 1000 students’ scores in Example 2.

§

Many **calculators** have an $s$ or $\sigma_{n-1}$ button that allows you to compute the **sample SD** of inputed data.
PART E: APPLICATIONS OF SD

How can we use a population SD … or a sample SD assuming that it does, in fact, do a decent job of estimating the population SD?

1) Comparing populations.

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

• For example, if Professor Dum’s evening class has a much higher SD on a standardized test than Professor Dee’s morning class does, then that could reflect a greater variety in the students in Professor Dum’s class.

2) Finding how close to the mean we expect “most” of the population values to be.

• In Part F, we will use Chebyshev’s Theorem and the “68-95-99.7%” Empirical Rule to help us understand how population values are clustered about the mean.

3) Finding an interval of usual values for the population.

• In Part G, we will see what values in the population would be typical (think: “usual”) and which would be unusual.
PART F: CHEBYSHEV’S THEOREM and
THE “68-95-99.7%” EMPIRICAL RULE

Chebyshev’s Theorem and the “68-95-99.7%” Empirical Rule help us understand how population values are clustered about the mean. They both imply that, in terms of SDs, **we can’t have too many values too far away from the mean.**

**Chebyshev’s Theorem** applies to any distribution shape (with finite SD) and is thus robust, or distribution-free. (For details, see Footnote 2.) According to the theorem, …

- … at least $\frac{3}{4}$ (that is, at least 75%) of the population values must lie **within two SDs** of the mean.

- … at least $\frac{8}{9}$ (that is, at least 89%) of the population values must lie **within three SDs** of the mean.
The “68-95-99.7%” Empirical Rule only applies to distributions that are approximately normal. (Normal distributions are an important category of bell-shaped distributions.) According to the rule, …

• … about 68% of the population values must lie within one SD of the mean.

• … about 95% of the population values must lie within two SDs of the mean.

• … about 99.7% of the population values must lie within three SDs of the mean.

Look at the normal distribution in the figure below.

• The inflection points on the curve are the points where the curve changes from concave up (think: “curving up”) to concave down (think: “curving down”), or vice-versa.

• The “mean line” is the vertical line through the maximum point (at the mean, $\mu$). If we draw vertical lines through the two inflection points, then the SD, $\sigma$, can be interpreted as the horizontal distance between either of those two lines and the “mean line.”

Note: The rule is called the “Empirical Rule” because approximately normal distributions are often seen in practice.
Example 3 (Applying Chebyshev’s Theorem and the “68-95-99.7% Empirical Rule)

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

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$\sigma$</th>
<th>$\sigma$</th>
<th>$\sigma$</th>
<th>$\sigma$</th>
<th>$\sigma$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>$\mu - 3\sigma$</td>
<td>$\mu - 2\sigma$</td>
<td>$\mu - \sigma$</td>
<td>$\mu$</td>
<td>$\mu + \sigma$</td>
<td>$\mu + 2\sigma$</td>
<td>$\mu + 3\sigma$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What proportion of the values are within 1 SD of the mean (between 40 and 60 points)?</th>
<th>By Chebyshev’s Theorem (distribution-free)</th>
<th>By the Empirical Rule (assumes normality)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(no information)</td>
<td>About 68%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What proportion of the values are within 2 SDs of the mean (between 30 and 70 points)?</th>
<th>By Chebyshev’s Theorem (distribution-free)</th>
<th>By the Empirical Rule (assumes normality)</th>
</tr>
</thead>
<tbody>
<tr>
<td>At least $\frac{3}{4}$ (At least 75%)</td>
<td>About 95%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What proportion of the values are within 3 SDs of the mean (between 20 and 80 points)?</th>
<th>By Chebyshev’s Theorem (distribution-free)</th>
<th>By the Empirical Rule (assumes normality)</th>
</tr>
</thead>
<tbody>
<tr>
<td>At least $\frac{8}{9}$ (At least 89%)</td>
<td>About 99.7%</td>
<td></td>
</tr>
</tbody>
</table>

- Observe that the results from the Empirical Rule are consistent with the results from Chebyshev’s Theorem. The powerful assumption of normality allows us to more precisely pinpoint a proportion, instead of just providing a lower bound (such as 75%), as Chebyshev’s Theorem does.
PART G: INTERVALS OF USUAL VALUES

Both Chebyshev’s Theorem and the Empirical Rule imply that the “vast majority” of the population values must lie within two SDs of the mean.

<table>
<thead>
<tr>
<th>The &quot;Two SD&quot; (2σ) Rule for Usual Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>An appropriate interval of usual values is given by ((\mu - 2\sigma, \mu + 2\sigma)).</td>
</tr>
<tr>
<td>If only sample data is available, we could use ((\bar{x} - 2s, \bar{x} + 2s)).</td>
</tr>
</tbody>
</table>

Example 4 (Interval of Usual Values: Revisiting Example 3)

Let’s revisit Example 3. The scores on a test have mean \(\mu = 50\) points and SD \(\sigma = 10\) points. What is the interval of usual values, based on the “2σ Rule”?

§ Solution

The interval of usual values is given by: \((\mu - 2\sigma, \mu + 2\sigma)\), which is \((30\) points, 70 points) here.

• This means that scores such as 31 points and 69 points would be typical (or “usual”) for the class, but that scores such as 25 points and 75 points would be unusual.

§
PART H: ESTIMATING SD

Example 5 (Estimating SD: Revisiting Example 3)

Let’s say the professor browses through the 1000 test scores and sees that the “vast majority” of scores are between 30 points and 70 points. If the professor assumes that the distribution is approximately symmetric, then the professor could estimate that the mean $\mu = 50$ points, the average of 30 points and 70 points. How could the professor then estimate the SD, $\sigma$?

Solution

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:

The "Two SD" (2$\sigma$) Rule for Estimating $\sigma$

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

Here, we obtain:

$\sigma \approx \frac{40}{4} \approx 10$ points

We estimate that the SD, $\sigma$, is about 10 points.
Example 6 (Estimating SD: Textbook Prices)

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

§ Solution

$$\sigma \approx \text{Realistic range} / 4$$
$$\sigma \approx (180 - 20) / 4$$
$$\sigma \approx 160 / 4$$
$$\sigma \approx 40$$

Our estimate for the SD, $\sigma$, is about $40.

• 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 investment portfolio is that risk is spread out across many stocks, bonds, etc. No single investment will destroy you if it sinks. Your hope is that your portfolio will at least keep up with inflation and the generally (hopefully) upward trend of the market over time.

PART J: EXCEL COMMANDS

Microsoft Excel uses the following commands:

• AVERAGE for a mean
• MEDIAN for a median
• STDEV.P for a population SD
• STDEV.S for a sample SD
#1) **Alternate, “one-pass” formula for** $s$. The alternate formula for $s$ below 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.

#2) **Chebyshev’s Theorem.** (“Chebyshev” can be spelled in various ways; some spellings begin with a “T.”) If $k > 1$, then the proportion of values in the population that must lie within $k$ SDs of the mean is at least $1 - \frac{1}{k^2}$.

- If $k = 2$, then $1 - \frac{1}{k^2} = 1 - \frac{1}{4} = \frac{3}{4}$, so at least $\frac{3}{4}$ (at least 75%) of the population values must lie within **two** SDs of the mean.

- If $k = 3$, then $1 - \frac{1}{k^2} = 1 - \frac{1}{9} = \frac{8}{9}$, so at least $\frac{8}{9}$ (at least 89%) of the population values must lie within **three** SDs of the mean.