Measures of center are very important numerical summaries for a data set; another critical kind of numerical statistic includes measures of variation, spread, dispersion, or variability. The two most common such statistics are the range and the standard deviation.

The **range** of a set of data is the difference between the highest and lowest values.

$$RANGE = MAX - MIN$$

Since the range is highly impacted by an outlier, there is another measure of variability which is more commonly used. The **standard deviation** is an average of differences between data values and the mean and is often used along with the mean to summarize a given data set.

As a reminder, with the mean, whether the entire population or a sample is involved is really an insignificant point. Both calculations involve the total of the data values divided by the number of data values.

| Population Mean | Sample Mean |
|---|---|
| $\mu = \dfrac{\Sigma X}{N}$ | $\overline{X} = \dfrac{\Sigma X}{n}$ |

However, with the standard deviation, the population and sample standard deviation formulas differ slightly. As you decide which formula to use, think about whether you have a complete set of data or data from the entire population of interest. For example, we use the population standard deviation for the number of home runs Henry Aaron hit through every season of his career or for the ages or GPAs of all students at our college.

On the other hand, if the data set is incomplete or the data was gathered from a sample or subset of the population, we use the sample standard deviation formula. Because of logistics, this is much more common. For example, we use the sample standard deviation for a current or active sports player or for weights of babies born in a given time period at the local regional medical center.

| Population Standard Deviation ($\sigma$) | Sample Standard Deviation (s) |
|---|---|
| $\sigma = \sqrt{\dfrac{\Sigma(X-\mu)^2}{N}}$   OR   $\sigma = \sqrt{\dfrac{\Sigma X^2}{N} - \left(\dfrac{\Sigma X}{N}\right)^2}$ | $s = \sqrt{\dfrac{\Sigma(X-\overline{X})^2}{n-1}}$   OR   $s = \sqrt{\dfrac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}}$ |

Note:  The major difference between the population and sample calculations is the n – 1 in the sample formula. Think of it as a compensation for the fact that $\overline{X}$ is only an estimate of $\mu$. The first formulas are more "definitional", showing averaging of differences between data values and the mean. In each case, the second formula is best (and easiest) to use because it avoids the rounding issue; these two "shortcut formulas" just require two sums — the sum of the data values ($\Sigma X$) and the sum of the squares of the data values ($\Sigma X^2$).

Other fairly popular statistics include the **population variance** ($\sigma^2$) and the **sample variance** ($s^2$) which are easily calculated by squaring the corresponding standard deviation.

Technology notes:

> In a TI graphing calculator, after the data values have been entered into a list, you can calculate 1-variable statistics, and the TI calculator uses the symbol "$\sigma_X$" for the population standard deviation, "$s_x$" for the sample standard deviation, "minX" for the minimum value, "maxX" for the maximum value, and you can then calculate the range.

> The computer spreadsheet software Excel uses the formula structure "=STDEV(first cell:last cell)" for the sample standard deviation, "=STDEVP(first cell:last cell)" for the population standard deviation, and "=MAX(first cell:last cell) – MIN(first cell:last cell)" for the range.

Examples:

(1)     These are quiz scores on a 10-question My Math Lab quiz:

90     100     80     80     40     70     60     90     100     80

The range is $100 - 40 = 60$.

Assuming this is **population data**, we have two formulas for calculating the standard deviation.

(a) In order to use $\sigma = \sqrt{\dfrac{\Sigma(X - \mu)^2}{N}}$ , we need 3 columns:

| X | X − μ | (X − μ)² |
|---|---|---|
| 90 | $90 - 79 = 11$ | $11^2 = 121$ |
| 100 | 21 | 441 |
| 80 | 1 | 1 |
| 80 | 1 | 1 |
| 40 | -39 | 1521 |
| 70 | -9 | 81 |
| 60 | -19 | 361 |
| 90 | 11 | 121 |
| 100 | 21 | 441 |
| 80 | 1 | 1 |

The calculation of the population mean μ is the sum of the test score values divided by the number of values ($790 \div 10 = 79$).

Then the standard deviation calculation uses the sum of the third column and N = 10.

$\Sigma(X - \mu)^2 = 3{,}090$            $\sigma = \sqrt{\dfrac{3090}{10}} \approx 17.6$

(b) In order to use $\sigma = \sqrt{\dfrac{\Sigma X^2}{N} - \left(\dfrac{\Sigma X}{N}\right)^2}$ , we need only 2 columns:

| X | X² |
|---|---|
| 90 | 8100 |
| 100 | 10000 |
| 80 | 6400 |
| 80 | 6400 |
| 40 | 1600 |
| 70 | 4900 |
| 60 | 3600 |
| 90 | 8100 |
| 100 | 10000 |
| 80 | 6400 |

Then the calculation uses the sums of the two columns and N = 10.

$\Sigma X = 790$ and $\Sigma X^2 = 65500$

$\sigma = \sqrt{\dfrac{65500}{10} - \left(\dfrac{790}{10}\right)^2} \approx 17.6$

Note: Whenever the mean has several decimal places (or repeats), this is clearly the better option, as there is no possibility of roundoff error before the final calculation. The only possible rounding for the method used in part (b) is at the very end.

Assuming this is **sample data**, we have two formulas for calculating the standard deviation.

(c) In order to use $s = \sqrt{\dfrac{\Sigma\left(X - \overline{X}\right)^2}{n-1}}$ , we would need the same 3 columns as in part (a)

with the slight adjustment from μ to $\overline{X}$ :

| X | $X - \overline{X}$ | $\left(X - \overline{X}\right)^2$ |
|---|---|---|

The calculation of the sample mean, $\overline{X}$, is the sum of the test score values divided by the number of values (790 ÷ 10 = 79).

Then, as before, the calculation uses the sum of the third column and n.

$\Sigma(X - \overline{X})^2 = 3{,}090$ and $n - 1 = 9$ $\qquad s = \sqrt{\dfrac{3090}{9}} \approx 18.5$

(d) In order to use $s = \sqrt{\dfrac{\Sigma X^2 - \left(\Sigma X\right)^2/n}{n-1}}$ , we need only 2 columns as in part (b):

Then the calculation uses the same sums of the two columns as in part (b):
$\Sigma X = 790$ and $\Sigma X^2 = 65500$

$s = \sqrt{\dfrac{65500 - 790^2/10}{9}} \approx 18.5$

Note: Whenever the mean has several decimal places (or repeats), this is clearly the better option, as there is no possibility of roundoff error before the final calculation.


(2)     Henry Aaron played 23 seasons in the Major Leagues (1954-1976) for the Milwaukee and Atlanta Braves, returning to Milwaukee to play for the Brewers his last few seasons. He is known for his prolific (and consistent!) home run hitting.

| Season | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|--------|----|----|----|----|----|----|----|----|----|----|----|----|
| # HR | 13 | 27 | 26 | 44 | 30 | 39 | 40 | 34 | 45 | 44 | 24 | 32 |

| Season | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | Total |
|--------|----|----|----|----|----|----|----|----|----|----|----|-------|
| # HR | 44 | 39 | 29 | 44 | 38 | 47 | 34 | 40 | 20 | 12 | 10 | 755 |

Solutions:     Range = 47 – 10 = 37

Standard deviation (population):
$$\sum X = 13 + 27 + 26 + 44 + \ldots + 12 + 10 = 755$$
$$\sum X^2 = 13^2 + 27^2 + 26^2 + 44^2 + \ldots + 12^2 + 10^2 = 27{,}535$$

$$\sigma = \sqrt{\frac{27535}{23} - \left(\frac{755}{23}\right)^2} = \sqrt{119.6219282} \approx 10.9\,\text{PPG}$$

[By the way, the value under the radical sign in this calculation is the population variance, $\sigma^2 \approx 119.6$.]


(3)     Michael Jordan played 15 seasons of his NBA career (1984-1998, 2001-2003), and ended each regular season with these points per game (PPG) averages. Ignore the two seasons in which he only played in 17 and 18 games (notice the strikethroughs); the rest of the seasons, he played at least 60 games of the 82-game season. Compute the range and standard deviation for his scoring averages through these 13 typical seasons in his career.

28.2   ~~22.7~~   37.1   35.0   32.5   33.6   31.5   30.1   32.6   ~~26.9~~
30.4   29.6   28.7   22.9   20.0

Solutions:     Range = 37.1 – 20.0 = 17.1 PPG

Standard deviation (population):
$$\sum X = 28.2 + 37.1 + 35.0 + \ldots + 22.9 + 20.0 = 392.2$$
$$\sum X^2 = 28.2^2 + 37.1^2 + 35.0^2 + \ldots + 22.9^2 + 20.0^2 = 12{,}091.3$$

$$\sigma = \sqrt{\frac{12091.3}{13} - \left(\frac{392.2}{13}\right)^2} = \sqrt{19.91751479} \approx 4.46\,\text{PPG}$$

(4)　In an experiment on the effect of a drug on reaction time, a patient is asked to depress a button whenever a light flashes. The reaction time (in milliseconds) for 15 trials is:

| 102 | 135 | 123 | 143 | 134 | 130 | 122 | 123 | 103 | 98 | 93 | 112 |
| 117 | 105 | 128 | | | | | | | | | |

Compute the range and standard deviation for these reaction times.

Solutions:　Range = 143 – 93 = 50 ms

Standard deviation (for a sample):
$$\sum X = 102 + 135 + 123 + \ldots + 105 + 128 = 1{,}768$$
$$\sum X^2 = 102^2 + 135^2 + 123^2 + \ldots + 105^2 + 128^2 = 211{,}580$$

Substituting these two calculated values, we have

$$s = \sqrt{\frac{211580 - 1768^2/15}{15-1}} = \sqrt{227.9809524} \approx 15.1\,\text{ms}$$

[By the way, the value under the radical sign in this calculation is the sample variance, $s^2 \approx 228.0\,\text{ms}^2$.]

(5)　The following data represent the numbers of years of teaching experience of a random sample of 10 faculty members from a state college

| 7 | 5 | 10 | 26 | 17 | 13 | 4 | 4 | 12 | 16 |

Compute the range and standard deviation for the faculty experience at this college.

Solutions:　Range = 26 – 4 = 22 years

Standard deviation (for a sample):
$$\sum X = 7 + 5 + 10 + 26 + 17 + 13 + 4 + 4 + 12 + 16 = 114$$
$$\sum X^2 = 7^2 + 5^2 + 10^2 + 26^2 + 17^2 + 13^2 + 4^2 + 4^2 + 12^2 + 16^2 = 1{,}740$$

$$s = \sqrt{\frac{1740 - 114^2/10}{10-1}} = \sqrt{48.93333333} \approx 7.0\,\text{years}$$

To prove the calculator shortcut formula from the more definitional formula, let's start with the sample variance formula (to avoid repeated square root symbols!).

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

For the data set, $X_1, X_2, X_3, \ldots, X_n$, this is equivalent to

$$s^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}$$

"FOIL"ing the perfect squares yields

$$s^2 = \frac{(X_1^2 - 2X_1\overline{X} + \overline{X}^2) + (X_2^2 - 2X_2\overline{X} + \overline{X}^2) + (X_3^2 - 2X_3\overline{X} + \overline{X}^2) + \ldots + (X_n^2 - 2X_n\overline{X} + \overline{X}^2)}{n-1}$$

Rearranging terms we have

$$s^2 = \frac{(X_1^2 + X_2^2 + X_3^2 + \ldots + X_n^2) - 2X_1\overline{X} - 2X_2\overline{X} - 2x_3\overline{X} - \ldots - 2X_n\overline{X} + (\overline{X}^2 + \overline{X}^2 + \overline{X}^2 + \ldots + \overline{X}^2)}{n-1}$$

The first grouping ($X_1^2 + X_2^2 + X_3^2 + \ldots + X_n^2$) is equivalent to $\sum X^2$. We can factor out $-2\overline{X}$ from all of the middle terms and replace the sum ($X_1 + X_2 + X_3 + \ldots + X_n$) by $\sum X$. Lastly, there are n identical $\overline{X}^2$ terms in the third grouping. Putting these all together, this simplifies to

$$s^2 = \frac{\sum X^2 - 2\overline{X}\sum X + n\overline{X}^2}{n-1}$$. Then replacing $\overline{X}$ with $(\sum X)/n$ and simplifying, we have

$$s^2 = \frac{\sum X^2 - 2\frac{\sum X}{n}\sum X + n\left(\frac{\sum X}{n}\right)^2}{n-1} = \frac{\sum X^2 - 2\frac{(\sum X)^2}{n} + \frac{(\sum X)^2}{n}}{n-1} = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$$

Multiplying both numerator and denominator by n, we have an alternate formula.

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

For the population variance formula, if we replace each n and n-1 with N in the sample formula, we have

$$\sigma^2 = \frac{N\sum x^2 - (\sum x)^2}{N^2}$$ OR $$\frac{\sum x^2}{N} - \frac{(\sum x)^2}{N^2} = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2$$ which leads directly to the population standard deviation calculator-ready formula, as shown below.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$

Exercises    [In general, round mean and standard deviation results to 1 more decimal place than the given data values.]:

1.    For the following golf scores, compute the mean and standard deviation. Then compare the players' performances using statistical language. By the way, Andy and Barney have played golf together before these 5 rounds, and they have played since these 5 rounds.

Andy:      78    83    81    84    80
Barney:    82    92    74    100    78

2.    The number of consecutive hours a light bulb will last before it burns out is tested. The number of hours for 24 bulbs is shown below. Compute the range and standard deviation for the life of a light bulb.

402    405    409    389    456    423    432    440    425    436    423    430
421    423    408    411    434    448    392    400    439    430    412    409

3.    Here are Ted Williams' home run totals by season (1939-1960):

31    23    37    36    0*    0*    0*    38    32    25    43
28    30    1**    13**    29    28    24    38    26    10    29
            *Active service, WWII        **Active service, Korean War

Compute the range and standard deviation for his home runs (excluding the 5 service year outliers).

4.    A dietician obtains the amounts of sugar (in grams) from 1 gram from each of 16 different cereals, including Cheerios, Corn Flakes, Fruit Loops, Trix, and 12 other brands. Find the range and standard deviation of these values.

0.03    0.24    0.30    0.47    0.43    0.07    0.47    0.13
0.44    0.39    0.48    0.17    0.13    0.09    0.45    0.43