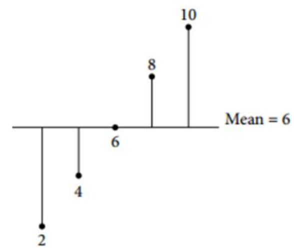
STANDARD DEVIATION AS A MEASURE OF SPREAD



Consider the following very simple data set: 2, 4, 6, 8, 10. Here, the mean is 6.

The differences between each data value and the mean, in order from left to right, are: −4, −2, 0, 2, 4. The sum of these differences is always 0. To avoid this you square the deviations to make them all positive: 16, 4, 0, 4, 16.

The average of these squared deviations is: $\frac{16+4+0+4+16}{5} = \frac{40}{5} = 8$

Now find the square root of this value (to undo the squaring you did earlier): $\sqrt{8} = 2.83$.

This value -2.83 - is called the standard deviation.

The **standard deviation** is a measure of spread. It uses all the data values, so is not influenced by outliers as much as the range. The range only uses the highest and lowest score, so is affected by outliers. The standard deviation, as a measure of spread, is often paired with the mean, a measure of central tendency. Statisticians like to use a measure of central tendency combined with a measure of spread to describe a data set.

The **population standard deviation**, represented by the Greek letter sigma σ, uses all data values in the population.

The **sample standard deviation**, represented by *s*, uses only a sample taken from the population. To find the sample standard deviation you divide by one less than the number of results. This means the sample standard deviation will always be greater than the population standard deviation. This is reasonable because taking a sample has introduced a degree of uncertainty to our calculations.

In either case, the bigger the value of the standard deviation, then the more spread out the data is considered to be.

If you are unsure whether you have a population or a sample, then calculate the sample standard deviation so that you do not draw conclusions that are unjustified by the statistical evidence.

Population standard deviation: $\sigma = \sigma_x = \sqrt{\frac{\Sigma(x-\mu)^2}{n}}$, where x refers to the individual scores, μ is the population mean and n is the number of values.

Sample standard deviation: $s = s_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$, where x refers to the individual scores, $\bar{x}$ is the sample mean and n is the number of values.

Calculating the standard deviation by hand is quite complicated so you should calculate it using technology.

Example 21

The results of a survey show the number of brothers and sisters for each member of the class.

1, 3, 4, 2, 1, 0, 3, 1, 2, 4, 0, 1, 0, 2, 1, 0, 1, 2, 2, 3

Calculate the standard deviation.

Solution

$n = 20$ the size of the population

$\mu = \frac{33}{20} = 1.65$ the population mean

To calculate the standard deviation, either a frequency table, scientific calculator or spreadsheet needs to be used.

Method 1—frequency table

x	f	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
0	4	−1.65	2.7225	10.89
1	6	−0.65	0.4225	2.535
2	5	0.35	0.1225	0.6125
3	3	1.35	1.8225	5.4675
4	2	2.35	5.5225	11.045
	$\sum f = 20$			$\sum f(x-\mu)^2 = 30.55$

$$\sigma_x = \sqrt{\frac{\Sigma(x-\mu)^2}{n}} = \sqrt{\frac{30.55}{20}} = 1.2359 \approx 1.24$$

Method 2—scientific calculator

Set your calculator in Statistics mode.

Enter the data score by score.

Use $\bar{x}$ key to find the mean: $\bar{x} = 1.65$

Use σ_x to find the standard deviation: $\sigma_x = 1.2359 \approx 1.24$

Use the s_x if you need the sample standard deviation: $s_x \approx 1.27$

Note: Each calculator is different so make sure you learn how to use your model.

Method 3—spreadsheet

x	f	xf	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
0	4	0	−1.65	2.7225	10.89
1	6	6	−0.65	0.4225	2.535
2	5	10	0.35	0.1225	0.6125
3	3	9	1.35	1.8225	5.4675
4	2	8	2.35	5.5225	11.045
Σ	20	33			30.55

$\mu = 1.65$

$s_x = 1.268$

$\sigma_x = 1.2359$

Method 4: use EXCEL

Example 22

A sample of members at a running club is asked to keep track of the distance, to the nearest whole kilometre, covered in training over the course of a week. The frequency table summarises the data collected.

Find the mean and the standard deviation for the data set shown.

Distance covered (nearest km)	Number of runners
21–30	3
31–40	5
41–50	4
51–60	2
61–70	1
71–80	1

Solution

Add a column for the midpoint of each class interval, x_m, and enter the data into a spreadsheet.

x	f	x_m	$x_m f$	$x_m - \mu$	$(x_m - \mu)^2$	$f(x_m - \mu)^2$
21–30	3	25.5	76.5	−17.5	306.25	918.75
31–40	5	35.5	177.5	−7.5	56.25	281.25
41–50	4	45.5	182	2.5	6.25	25
51–60	2	55.5	111	12.5	156.25	312.5
61–70	1	65.5	65.5	22.5	506.25	506.25
71–80	1	75.5	75.5	32.5	1056.25	1056.25
Σ	16		688			3100

$\mu = 43$
$\sigma_x = 13.9194$
$s_x = 14.3759$

The mean is 43 and the standard deviation is 13.92, correct to 2 decimal places.

The sample standard deviation is 14.38.

Limitations of using the standard deviation

The standard deviation uses every piece of data, so it can be affected by outliers in the same way that those values can affect the mean.

Example 23

The number of goals Eloise scored in 15 games of a netball competition was recorded.

22, 27, 32, 48, 29, 36, 29, 41, 44, 29, 38, 26, 39, 42, 40

(a) Find the mean and standard deviation for the data. Assume the data is a sample of her goals for the season. Give answers rounded to 2 decimal places if necessary.

(b) In the following week Eloise is injured early in the game and scores only 4 goals. Find the new mean and standard deviation.

Solution

(a) Using a scientific calculator, $n = 15$: $\bar{x} = \dfrac{522}{15} = 34.8$

$\sigma_x = 7.3774 \approx 7.38$
$s_x = 7.6364 \approx 7.64$
Considering the data as a sample of the goals for the season, use $s_x = 7.64$ as the standard deviation.
Mean is 34.8.

(b) All that has to be done is to add the new data value into your calculator and then obtain the answers:

$n = 16$

$\bar{x} = \dfrac{526}{16} = 32.875 \approx 32.88$

$\sigma_x = 10.3252 \approx 10.33$
$s_x = 10.6638 \approx 10.66$
Considering the data as a sample of the goals for the season, use $s_x = 10.66$ as the standard deviation.
Mean is 32.88.

Example 24

A teacher recorded the results for a test, marked out of 50, as follows:

12, 17, 22, 23, 26, 26, 27, 29, 30, 31, 31, 32, 32, 35, 35, 37, 38, 40, 41, 42, 42, 49.

After the tests were returned, the student who had scored 12 pointed out that several pages had not been corrected. Their mark increased to 36.

Which of the following statements correctly describes what would happen to the class statistics? (Do not use technology for this question.)

- **A** The standard deviation and mean would both decrease.
- **B** The standard deviation and mean would both increase.
- **C** The standard deviation would increase and the mean would decrease.
- **D** The standard deviation would decrease and the mean would increase.
- **E** The median would not change.

Solution

The median is 31.5 and will become 32 once the 12 is removed and the 36 inserted. This means **E** is incorrect.

The lowest value has been removed and replaced by a value nearer the mean. This means the standard deviation must be smaller. So, **B** and **C** can both be eliminated.

A smaller value has been replaced by a bigger value; the mean must increase. So, **A** is incorrect.

D is the correct response.