  
Overview  
(/study/app/  
122-  
cid-  
754029/k

TOPIC 4  
PROBABILITY AND STATISTICS



(https://intercom.help/kognity)



  
Table of  
contents

  
Notebook

  
Glossary

  
Reading  
assistance

4.3.0 The big picture

4.3.1 Measures of central tendency

4.3.2 Measures of dispersion

4.3.3 Data transformation

4.3.4 Checklist

4.3.5 Investigation

  
Student  
view



Show all topics





Overview  
(/study/ap  
122-  
cid-  
754029/k

Teacher view

## Index

The big picture  
Measures of central tendency  
Measures of dispersion  
Data transformation  
Checklist  
Investigation

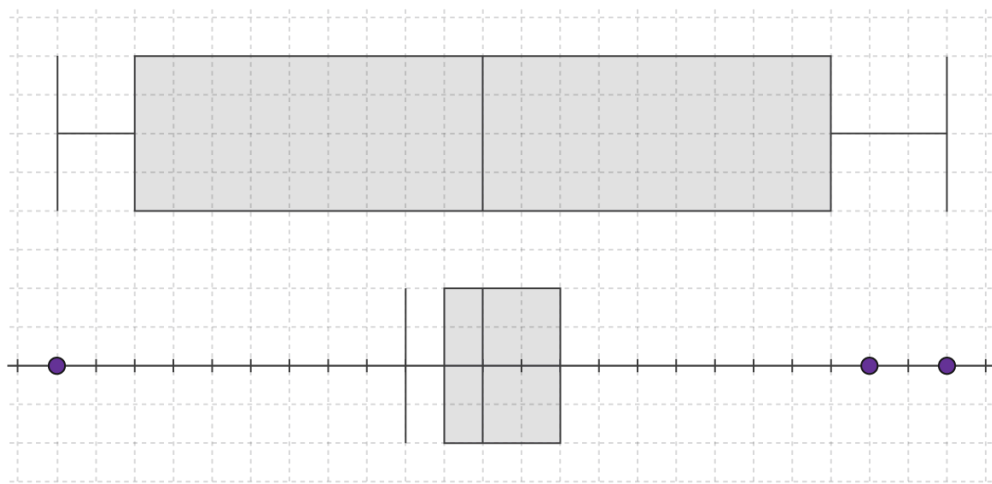
4. Probability and statistics / 4.3 Measures of central tendency and dispersion

# The big picture

Consider the following two sets of data and their box-and-whisker plots:

Set A: {3, 4, 5, 13, 13, 14, 18, 21, 23, 25, 26}

Set B: {3, 12, 13, 13, 13, 14, 15, 16, 16, 24, 26}




Box-and-whisker plots for Set A (upper) and Set B (lower).

More information

The image shows two box-and-whisker plots against a grid background. The upper plot represents Set A, illustrating a wide spread with a central box that describes the interquartile range. The median is shown as a line inside the box. Whiskers extend from the box to indicate the full range of the data, with outliers marked as points beyond the



Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/  
k

whiskers. The lower plot represents Set B, showing a narrower interquartile range compared to Set A. Its median is similarly marked, with smaller whiskers indicative of data spread. Both plots allow visual comparison of data distribution and range between the two datasets.

[Generated by AI]

When you look at the two sets, and especially their respective box-and-whisker plots, it appears that they have little in common. However, when you calculate various statistics for the sets, you see a very different story:

Statistics for Set A and Set B.		
	Set A	Set B
Mean	15	15
Median	14	14
Mode	13	13
Range	23	23

How can this happen? How could two sets with such different data values have the same statistics? What makes the difference? The answer is that the data in Set A is more widely **dispersed** than the data in Set B. The range of each set is 23, but the range considers only the maximum and minimum values. How is the rest of the data spread within that range? A box-and-whisker plot shows us the quartiles, range and interquartile range (IQR), but in this subtopic we will delve deeper and explore other statistics to help us interpret the data.

First, we will review the various measures of central tendency– the mean, median and mode – and explore how to estimate the mean for data grouped into classes. Then we will introduce two new measures of dispersion– variance and standard deviation. We will also consider how these statistics are affected when the data in a set change uniformly.



Student  
view



Overview  
(/study/app/  
122-  
cid-  
754029/k



## Concept

In this subtopic, we will take the opportunity to explore **relationships** between different statistical measures, such as the mean and standard deviation, as well as the impact that various **changes** have on those measures. We will use other strategies to **approximate** these measures when dealing with data that is grouped into ranges (called classes). How do these statistical measures help us understand the data better?

4. Probability and statistics / 4.3 Measures of central tendency and dispersion

# Measures of central tendency

If you had to represent a data set with just one value, which one would you use? It should be a value around which the entire data set is well distributed. This first measure with which we summarise the data is a value for the **central tendency**, the one number which best represents the entire data set. In everyday language, this is referred to as the **average**. There are three common measures for the central tendency : mode, median and mean.

## Mode

In [section 4.2.1 \(/study/app/m/sid-122-cid-754029/book/grouped-data-and-quartiles-id-26228/\)](/study/app/m/sid-122-cid-754029/book/grouped-data-and-quartiles-id-26228/) you learned that for data grouped into intervals or classes, the class with the greatest frequency is called the **modal class**. The mode of a data set is the *value* with the greatest frequency (the word 'mode' is used for what is fashionable, what is 'popular'). For data grouped into intervals, the mode is estimated as the mid-interval value of the modal class. It is possible for a data set to have more than one mode – we call a data set like this bimodal, trimodal, and so on. If no values are repeated, or all of the values in a set are repeated the same number of times, then there is *no mode* .



## Exam tip

Class intervals do not have to be equal, but for your work in this course, they will be.



Student  
view



# Median, $m$

Overview

(/study/app/m/sid-122-cid-754029/book/box-and-whisker-plots-id-26229/)you  
122-  
cid-  
754029/c

In [section 4.2.2](#) (/study/app/m/sid-122-cid-754029/book/box-and-whisker-plots-id-26229/) you used **quartiles** to present data as a box-and-whisker plot and learned that the second quartile is also called the median. The median of a data set is the middle value when the data is arranged in order. It divides the data set in two, with half the data below and half the data above its value. If the data set has an even number of data points, say  $n$ , the median is estimated as the average of the data values of the  $\frac{n}{2}$ th and  $\left(\frac{n}{2} + 1\right)$ th. If the data is grouped into classes, the median is estimated as the mid-interval of the class containing the middle value.

## Mean, $\mu$

The measure of mean refers to the **arithmetic mean** of the data: the sum of the numerical data divided by the number of data points. The mean is often what is meant when the word 'average' is used. Thus, for a data set given by  $n$  data values  $x_i$ , the mean is given by

$$\mu = \frac{\sum_{i=1}^n x_i}{n}.$$

If the data is given by a frequency distribution table, then the mean is given by

$$\mu = \frac{\sum_{i=1}^k f_i x_i}{n},$$

where  $f_i$  is the frequency of the data value  $x_i$  and  $n = \sum_{i=1}^k f_i$  is the number of data points in the data set.

If the data is grouped into  $k$  classes, we can adapt this formula to *estimate* the mean:

$$\mu = \frac{\sum_{i=1}^k f_i x_i}{n},$$

where  $f_i$  is the frequency of the  $i$ th class and  $x_i$  is the mid-interval value of the  $i$ th class.

Section

Student... (0/0)

Feedback



Print (/study/app/m/sid-122-cid-

Assign

754029/book/the-big-picture-id-26233/print/)

Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k



## Making connections

Consider the difference that extreme values could have on these measures of centre. The mode would probably not change at all, and the median might only change a little, but what about the mean?

People working in fields where financial data are used must often consider which of these measures should be used in order to communicate information accurately and ethically. For example, median income is often used because it is not affected very much by one very rich person living in the area, whereas the mean income could drastically increase. Likewise, median home values are often used because one unkempt house in a neighbourhood could bring the mean value down, affecting the value of other homes.

How do you think people could use data like this unethically in order to mislead others?

### Example 1



The table below shows the data found by asking 20 students how many different countries they have travelled to in their lifetime, including the one in which they live. Find the mean, median and mode of the number of countries visited by the group of students.

Number of countries	Frequency
1	4
2	6
4	3
5	4
7	2
10	1



Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k



How many countries have you travelled to?

Kat72 Getty Images

More information

The image is a world map showing continents and countries in various colors. There are numerous colored map pins placed across different countries. The map spans from North and South America on the left to Europe, Africa, and Asia on the right. The Atlantic and Pacific Oceans are labeled in blue, and countries are marked with black borders. The map pins are placed in clusters over North America, Europe, and parts of the Middle East, indicating locations that might represent traveled destinations or points of interest. The pins vary in color, potentially symbolizing different categories or frequencies.

[Generated by AI]

To find the mean  $\mu$  given a frequency table, use the formula

$$\mu = \frac{\sum_{i=1}^n f_i x_i}{n}.$$

In the numerator, multiply each value by its frequency and add the products. In the denominator, add the frequencies (or notice from the problem that there are 20 students total).



Student  
view

$$\mu = \frac{1 \times 4 + 2 \times 6 + 4 \times 3 + 5 \times 4 + 7 \times 2 + 10 \times 1}{4 + 6 + 3 + 4 + 2 + 1}$$

Simplify the expression to find the value of the mean.

$$\mu = \frac{72}{20} = 3.6$$

To find the median  $m$  from 20 students, find the mean of the 10th and 11th values.

$$m = \frac{2 + 4}{2} = 3.$$

To find the mode, note that the highest frequency is 6, for the value 2.

$$\text{mode} = 2.$$

Which measure of central tendency do you think best represents the data in this example?

Example 2




The table below shows grouped data for the heights of 51 people. Estimate the mean, median and mode of the data.

Grouped height data.

Height (cm)	Frequency $f_i$
150—152	3
153—155	8
156—158	6
159—161	12
162—164	7





Overview  
(/study/app/  
122-  
cid-  
754029/k

Height (cm)	Frequency $f_i$
165—167	4
168—170	2
171—173	4
174—176	5
Total	51

The table below shows the data along with the mid-interval,  $x_i$ , and the product of the frequency with the mid-interval values,  $f_i x_i$ .

Calculations to determine the mean of grouped data.

Height (cm)	Mid-interval (cm) $x_i$	Frequency $f_i$	$f_i x_i$
150—152	151	3	453
153—155	154	8	1232
156—158	157	6	942
159—161	160	12	1920
162—164	163	7	1141
165—167	166	4	664
168—170	169	2	338
171—173	172	4	688
174—176	175	5	875



Student  
view



Overview  
(/study/app/  
122-  
cid-  
754029/k

Height (cm)	Mid-interval (cm) $x_i$	Frequency $f_i$	$f_i x_i$
	<b>Total</b>	51	8253

To estimate the mean, sum the frequencies and the products:

$$\mu = \frac{\sum_{i=1}^9 f_i x_i}{n} = \frac{8253}{51} \approx 161.82 \text{ cm.}$$

To estimate the median height of 51 people, look for the 26th value. In this case, the 26th value is in the interval 159–161, so the estimated median is the mid-interval value 160.

To estimate the mode, take the mid-interval value of the modal class. The interval 159–161 is the modal class, so the estimated mode is 160.

### ✓ Important

Calculating the mean from grouped data will, in general, not give the same result as calculating from ungrouped (raw) data. Why do you think using mid-interval values provides only an estimate of the mean? How does that differ from the mean we found in Example 1? Why are the calculations different for grouped and ungrouped data?

### 🔗 Making connections

While something as mundane as measuring heights can be sufficiently accurate with the ranges of data in Example 2, some disciplines require more precision than others. For example, a medical researcher testing the results of an experimental drug may use minute intervals when compiling data related to different dosages. On the other hand, an economist might look at financial data with ranges spanning several millions. Think about a field of study that you might want to study someday. Would research in that field require data in minute intervals, or would less precise data be sufficient?



Student  
view



Overview  
(/study/app/  
122-  
cid-  
754029/)



## Theory of Knowledge

Consider the way language is used in mathematics. When finding measures of central tendency, we define **centre** in three different ways. The mean is central in relation to its distance from the other data. The median is central in its position among the other data. The mode could be considered the centre of attention, or popularity, given that it is the value that more of the data are equal to. Can you think of any other mathematical concepts that are interpreted differently based on how we define certain terms?

## 3 section questions ▾

4. Probability and statistics / 4.3 Measures of central tendency and dispersion

# Measures of dispersion

## Range and interquartile range

You have already encountered the **range** and **interquartile range** (IQR), which are two basic examples of measures of dispersion. Dispersion has to do with how much data is spread out or clustered together. In this section, we will examine how measures of dispersion tell us more about the data. To explore this, we will revisit the two sets of data from section 4.3.0 (/study/app/m/sid-122-cid-754029/book/the-big-picture-id-26233/):

Set A: {3, 4, 5, 13, 13, 14, 18, 21, 23, 25, 26}


Set B: {3, 12, 13, 13, 13, 14, 15, 16, 16, 24, 26}

When we examined this data before, we recognised that the mean, median, mode and range were all identical, but even a brief look at the data indicates that something is different.

Identical statistics do not necessarily indicate identical data.



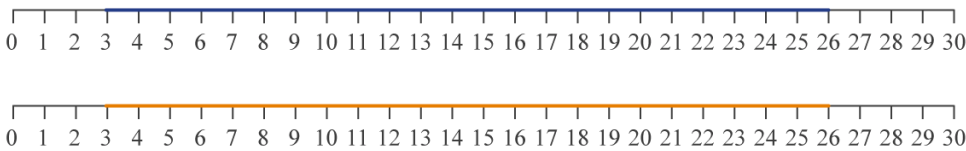
Student  
view




Overview  
(/study/ap  
122-  
cid-  
754029/k

	Set A	Set B
Mean	15	15
Median	14	14
Mode	13	13
Range	23	23

The range is the distance between the highest value and the lowest; that is, **maximum – minimum**. While this does tell you something, you don’t learn very much about the set of data.



Number lines for Set A (blue) and Set B (orange)

 More information

The image contains two number lines. The first number line on top is labeled for Set A and is colored blue. It spans horizontally from 0 to 30, with a solid blue line stretching from 5 to 23. The second number line below it is labeled for Set B and is colored orange. It also spans from 0 to 30, with a solid orange line extending from 13 to 16. These number lines visually represent the ranges of two different data sets, Set A and Set B, showing their respective minimum and maximum values on the number line.

[Generated by AI]

The IQR is somewhat more useful, in that it tells us the distance between the first and third quartiles, that is,  $Q_3 - Q_1$ , which is the range of the middle 50% of the data. This is  $23 - 5 = 18$  for Set A and  $16 - 13 = 3$  for Set B. The middle 50% of Set A is spread out six

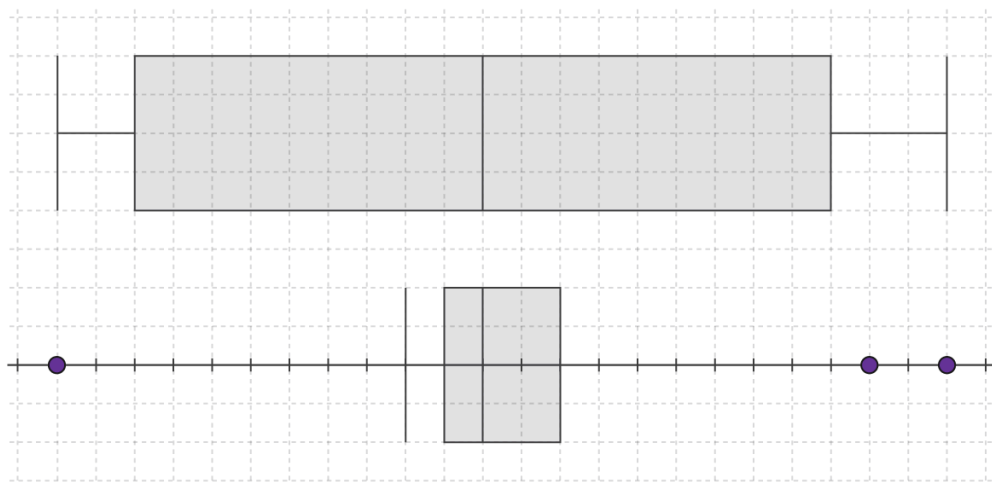


Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k

times as much as the middle 50% of Set B. This gives more information, but it could have been estimated simply by looking at the box-and-whisker plot. How could you describe the dispersion of the data even more specifically?



Box-and-whisker plots for Set A and Set B.

More information

The image shows box-and-whisker plots for two different data sets, Set A and Set B, displayed over a grid. Both plots illustrate the distribution of data, with each plot featuring a box that represents the interquartile range (IQR).

- **Set A:** The box for Set A is significantly larger, indicating a wider IQR. The lower and upper quartiles are marked by the edges of the box which stretches horizontally from the first quartile (Q1) to the third quartile (Q3). The line inside the box represents the median of Set A. Whiskers extend from the box to depict the smallest and largest values within a specified range.
- **Set B:** Comparatively, the box for Set B is much smaller, indicating a narrow IQR. The edges of the box mark the lower and upper quartiles. A line within this box shows the median value for Set B. The whiskers for Set B also extend outward but cover a shorter range than those of Set A.

Overall, Set A demonstrates more variance in data distribution compared to Set B, highlighting the different spreads within the two data sets.

[Generated by AI]



Student  
view



# Variance, $\sigma^2$ , and standard deviation $\sigma$

Overview

(/study/ap  
122-  
cid-  
754029/k

Two more advanced measures of dispersion are variance and standard deviation. These measure how much the data set *varies* or *deviates* from the mean. The higher the value, the more spread out the data is.

For example, first calculate the difference between each data point and the mean. Then add up all these differences and divide by the number of data points. Is this a good measure of dispersion? Can you think of any problems with this method? Think about what happens when you sum positive and negative values.

To make sure that positive and negative differences don't cancel each other out, we **square** the differences when calculating the variance. The variance, written Var or often  $\sigma^2$ , is therefore a measure of the spread of the data around the mean. The standard deviation,  $\sigma$ , is the **square root** of the variance. Thus, variance and standard deviation measure the spread in the same manner, but one is the square of the other.

In this course, you will not need to calculate the variance or standard deviation except with your GDC, which you will see how to do later in this section. Thus, while **you will not be asked to apply the following formulae**, it is worth seeing how it is done.

We define the variance as

$$\sigma^2 = \frac{\sum_{i=1}^k f_i(x_i - \mu)^2}{n},$$

where  $k$  is the number of distinct values or classes. The standard deviation is therefore

$$\sigma = \sqrt{\frac{\sum_{i=1}^k f_i(x_i - \mu)^2}{n}}.$$

Now we can determine the standard deviation of each set of data to see the difference in the dispersion as it relates to distance from the mean.

The variance can also be expressed as

$$\sigma^2 = \frac{\sum_{i=1}^k f_i x_i^2}{n} - \mu^2$$



Student  
view



because

Overview  
(/study/ap  
122-  
cid-  
754029/k

$$\begin{aligned}\frac{\sum_{i=1}^k f_i(x_i - \mu)^2}{n} &= \frac{\sum_{i=1}^k f_i x_i^2}{n} - 2\mu \frac{\sum_{i=1}^k f_i x_i}{n} + \mu^2 \frac{\sum_{i=1}^k f_i}{n} \\ &= \frac{\sum_{i=1}^k f_i x_i^2}{n} - 2\mu^2 + \mu^2 \\ &= \frac{\sum_{i=1}^k f_i x_i^2}{n} - \mu^2.\end{aligned}$$

This formula can be easier to remember as it is ‘the mean of the squares minus the square of the mean’.



### Making connections

If the interquartile range goes together with the median, the standard deviation, which is associated with the variance of the data, goes together with the mean. Why do you think this is?

## Example 1



Find the variance and standard deviation of the data from Set A ( $\mu = 15$ ). Since only one value is repeated, you do not need to use a frequency table.

To find the variance, start with the formula, replacing  $f_i$  with 1 and  $k$  with  $n$ :

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

Since there are several calculations for each element of data, it is helpful to use a table to organise your calculations as we do in the table below.

Calculations to determine the variance of the data in Set A.

$x_i$	$x_i - \mu$	$(x_i - \mu)^2$
3	3 - 15	144



Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k

$x_i$	$x_i - \mu$	$(x_i - \mu)^2$
4	$4 - 15$	121
5	$5 - 15$	100
13	$13 - 15$	4
13	$13 - 15$	4
14	$14 - 15$	1
18	$18 - 15$	9
21	$21 - 15$	36
23	$23 - 15$	64
25	$25 - 15$	100
26	$26 - 15$	121
		Total = 704

Evaluating the formula with these values gives you

$$\sigma^2 = \frac{704}{11} = 64.$$

Since standard deviation is the square root of variance,

$$\sigma = \sqrt{64} = 8.$$

## Example 2



Find the variance and standard deviation of the data from Set B ( $\mu = 15$ ). Since several values are repeated, use a frequency table.



Student  
view



To find the variance, start with the formula:

$$\sigma^2 = \frac{\sum_{i=1}^k f_i(x_i - \mu)^2}{n}$$

The table below shows the data from Set B arranged with frequencies and the calculations to find the variance .

Calculations to determine the variance of the data in Set B.

$x_i$	$f_i$	$x_i - \mu$	$f_i(x_i - \mu)^2$
3	1	3 – 15	144
12	1	12 – 15	9
13	3	13 – 15	12
14	1	14 – 15	1
15	1	15 – 15	0
16	2	16 – 15	2
24	1	24 – 15	81
26	1	26 – 15	121
Total	11		370

Evaluating the formula with these values gives you

$$\sigma^2 = \frac{370}{11} \approx 33.64.$$

Since standard deviation is the square root of variance,

$$\sigma \approx \sqrt{33.64} \approx 5.80.$$





Overview  
(/study/ap  
122-  
cid-  
754029/k

Based on our measures of the mean and standard deviation, we can summarise the data from each set by quoting its mean and standard deviation in the form:

$$\text{Set A: } \mu \pm \sigma = 15 \pm 8$$

$$\text{Set B: } \mu \pm \sigma = 15 \pm 5.8$$

The smaller standard deviation of Set B means that, on average, the data in Set B is closer to the mean than the data in Set A.

## Theory of Knowledge

You will notice above that there are actually two formulae for finding variance.

$$\sigma^2 = \frac{\sum_{i=1}^n f_i(x_i - \mu)^2}{n} \text{ and}$$

$$\sigma^2 = \frac{\sum_{i=1}^k f_i x_i^2}{n} - \mu^2.$$

In this case we saw how the second was derived from the first, but in other cases it may not be evident how two formulae that find the same value are related. If two different formulae can consistently give you the same value, what does this tell us about the concept of truth in mathematics?

## Calculating the mean, standard deviation and variance on a calculator

Data may consist of lots of numbers. Statistical analysis consists of applying various formulae to all these numbers. Using technology, we can compute statistics for large sets of data with a speed and accuracy that would be impossible for us to achieve on our own.

### Exam tip


Knowing how to use your GDC is essential to success on the final exam. Make sure you learn how to use your calculator efficiently so you do not waste time during the exam trying to figure out where functions are and how they work.



Student  
view

Overview (/study/ap  
122-  
cid-  
754029/k

In subtopic 4.2 (/study/app/m/sid-122-cid-754029/book/the-big-picture-id-26227/)you learned how to find the five-number summary of a data set using your GDC. The instructions below show how to use your calculator to find the mean, variance and standard deviation of a data set.

Step	Explanation
Press 2 to open the statistics mode.	



Student  
view



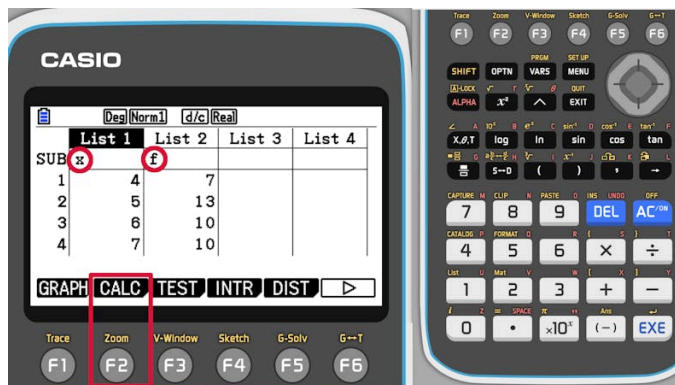
Overview  
(/study/ap  
122-  
cid-  
754029/k

## Step

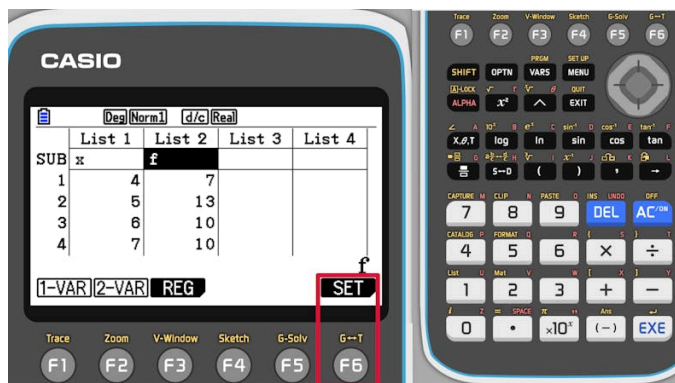
## Explanation

Enter the data. Note, that you can also give a name to your data. Remember that you used List1 and List2 for storing the data, you will need this later.

Once done entering the data, press F2.



First you need to set some parameters, so press F6.



Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k

## Step

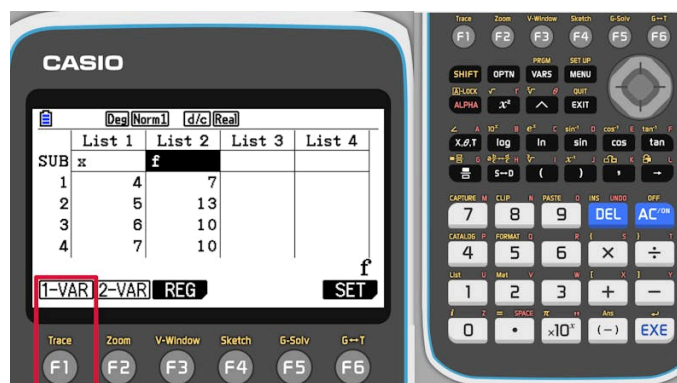
Remember, you stored the  $x$ -values in List1 and the frequencies in List2. Use F1 and the number keys to tell this to the calculator. You are interested in the settings for 1-variable statistics now.

Once done, press EXE.

## Explanation



Press now F1 to view the 1-variable statistics summary.



Student  
view

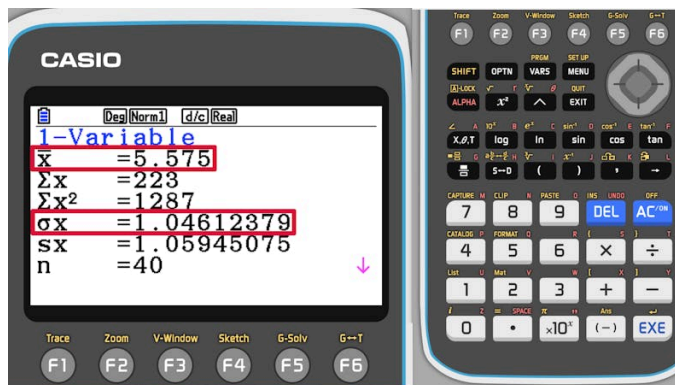


Overview  
(/study/ap  
122-  
cid-  
754029/k

## Step

## Explanation

In the summary, the mean is denoted by  $\bar{x}$  and the standard deviation is  $\sigma x$ . The variance is the square of the standard deviation, it is not given in the summary directly.



## Step

## Explanation

Choose the 1-variable statistics application.



Student  
view

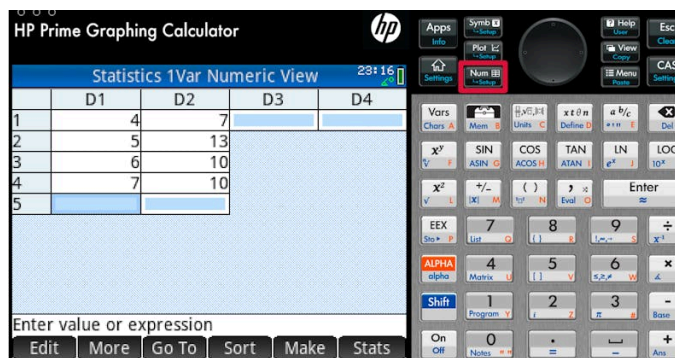


Overview  
(/study/ap  
122-  
cid-  
754029/k

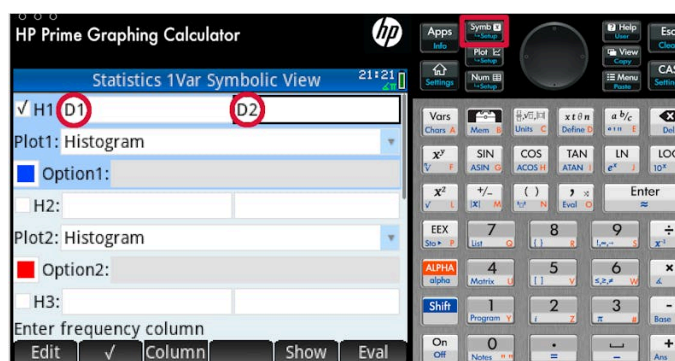
## Step

In the numeric view of the application, enter the data. Remember that you used D1 and D2 for storing the data, you will need this in the next step.

## Explanation



In the symbolic view you can tell the calculator where you stored the data. The first field is for the  $x$ -values (so enter D1). Enter D2 in the second field to tell the calculator where the frequencies are.



Student  
view



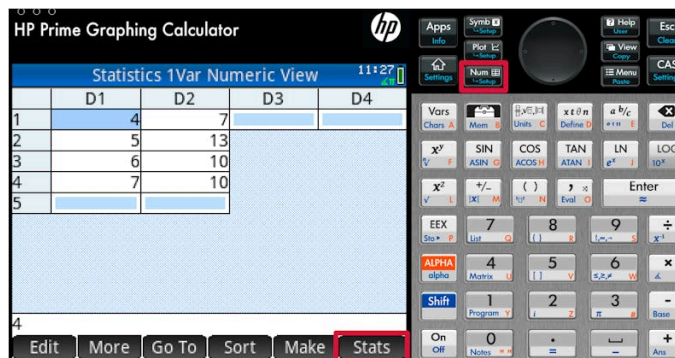


Overview  
(/study/ap  
122-  
cid-  
754029/k

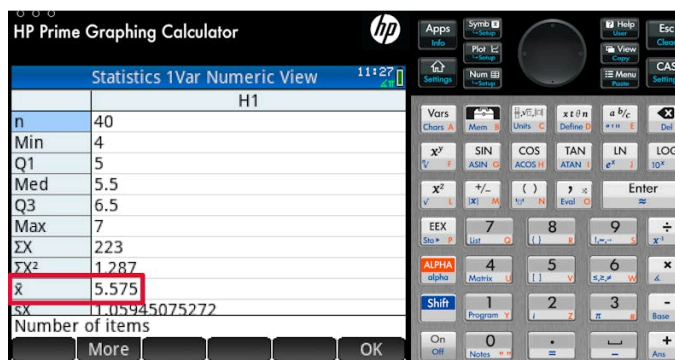
## Step

Go back in the numeric view and press stats to get the 1-variable statistics summary.

## Explanation



On the first screen (among other information) you can see the mean (denoted by  $\bar{x}$ ).



Student  
view



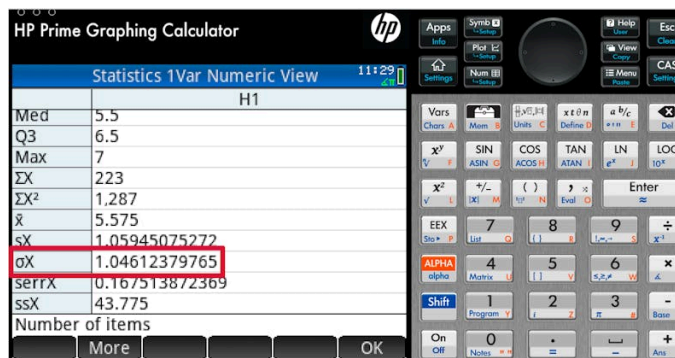


Overview  
(/study/ap  
122-  
cid-  
754029/k

Step

Explanation

To see the standard deviation (which is denoted by  $\sigma X$ ), you need to scroll down. The variance is the square of the standard deviation, it is not given in the summary directly.



Step

Explanation

Press stat to start working with data sets.



Section

Student... (0/0)


Feedback

Print (/study/app/m/sid-122-cid-754029/book/measures-of-dispersion-id-26235/print/)

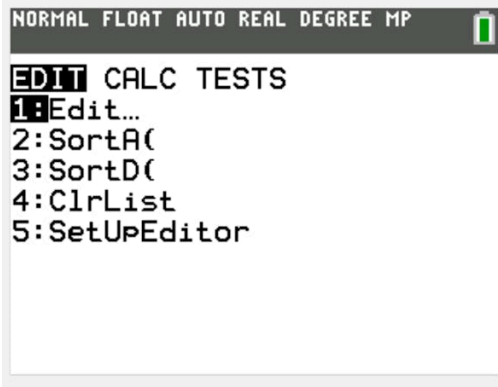


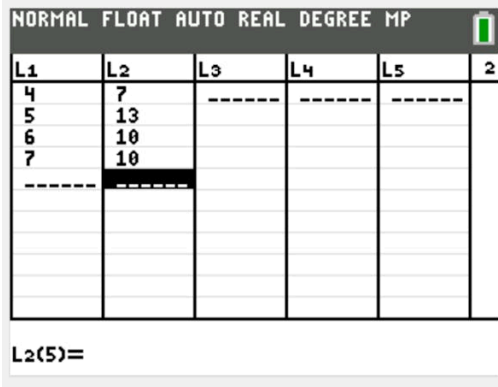


Assign



Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k

Step	Explanation
First you need to enter your data, so choose to bring up the editing screen.	<div></div> <div></div> <div></div>
Enter the data. Remember that you used L1 and L2 for storing the data, you will need this later.  Once done entering the data, press stat again.	<div></div> <div></div> <div></div>

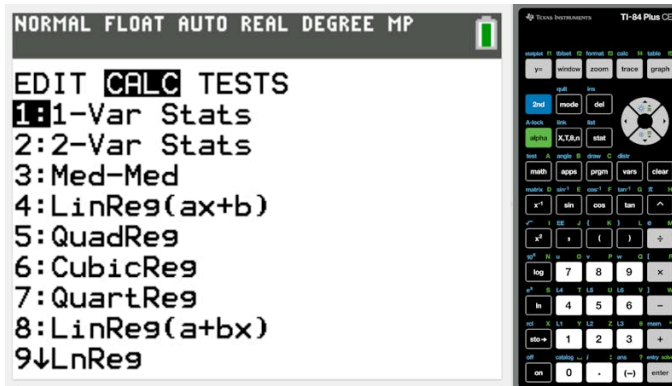


Overview  
(/study/ap  
122-  
cid-  
754029/k

## Step

## Explanation

Choose the option to calculate the 1-variable statistics.



Remember, you stored the  $x$ -values in L1 and the frequencies in L2. Use 2nd and the number keys to tell this to the calculator.

Once done, move down to the last row and press enter.



Student  
view

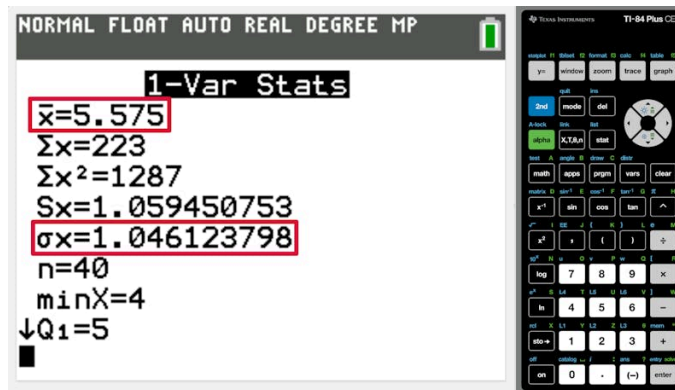


Overview  
(/study/ap  
122-  
cid-  
754029/k

### Step

In the summary, the mean is denoted by  $\bar{x}$  and the standard deviation is  $\sigma x$ . The variance is the square of the standard deviation, it is not given in the summary directly.

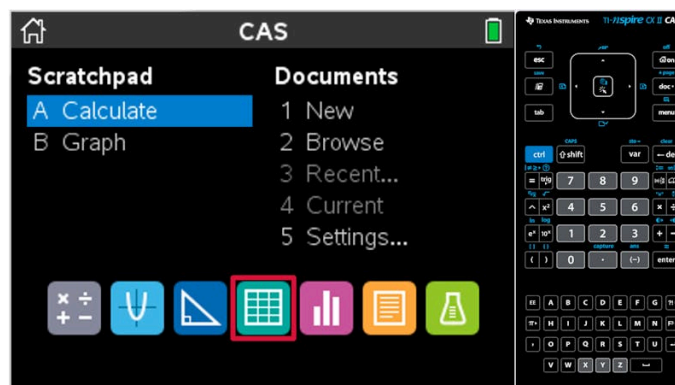
### Explanation



### Step

To start, open a spreadsheet document.

### Explanation



Student  
view

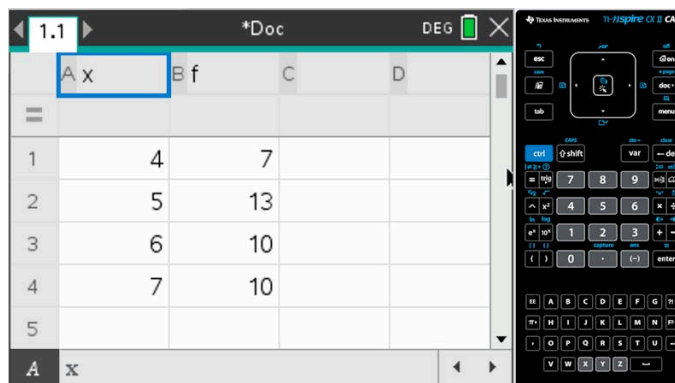


Overview  
(/study/ap  
122-  
cid-  
754029/k

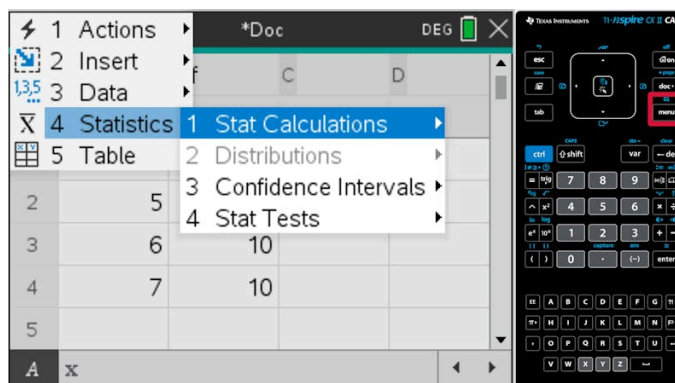
## Step

Enter the data. Note, that you can also give a name to your data. Remember these names, you will need this later.

## Explanation



To see the mean and standard deviation, open the menu...



Student  
view

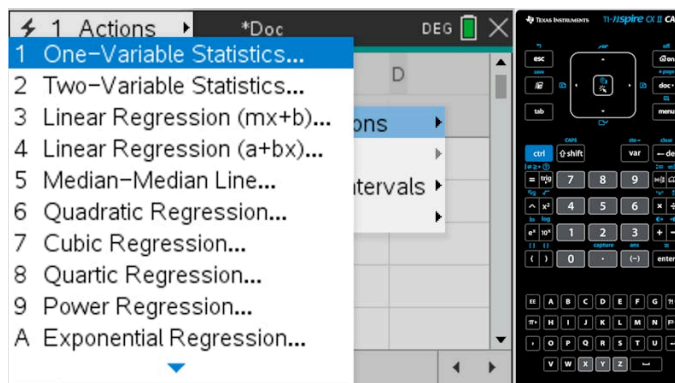


Overview  
(/study/ap  
122-  
cid-  
754029/k

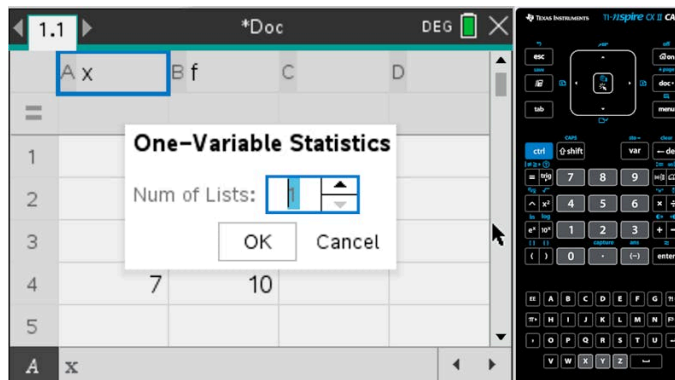
Step

Explanation

... and choose one-variable statistics.



You have one list (you will specify the frequencies later).



Student  
view



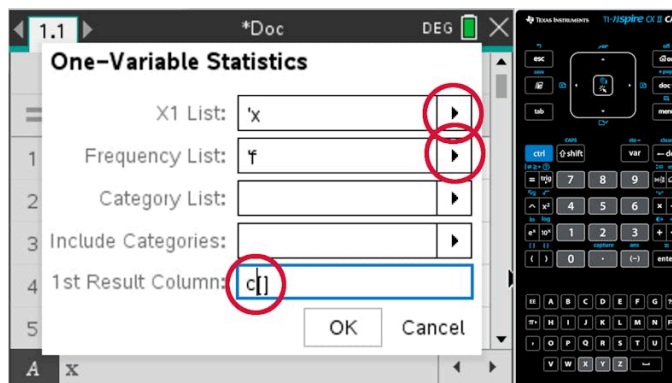
Overview  
(/study/ap  
122-  
cid-  
754029/k

## Step

Choose these names you used to store the  $x$ -values and the frequencies from the pull-down menu.

The results will be stored in the spreadsheet. Since the data is stored in the first two columns (a and b), choose a different column for the result.

## Explanation



On the first screen (among other information) you can see the mean (denoted by  $\bar{x}$ ).

	x	f		
1	4	7	Title	One-Va...
2	5	13	$\bar{x}$	5.575
3	6	10	$\Sigma x$	223.
4	7	10	$\Sigma x^2$	1287.
5			$s_x := s_n - \dots$	1.05945



Student  
view

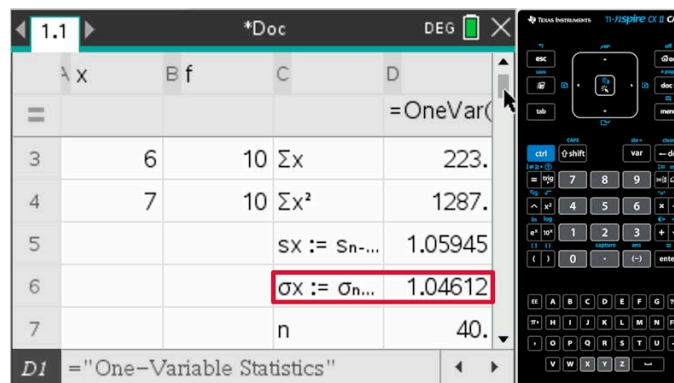


Overview  
(/study/ap  
122-  
cid-  
754029/k

## Step

To see the standard deviation (which is denoted by  $\sigma_X$ ), you need to scroll down. The variance is the square of the standard deviation, it is not given in the summary directly.

## Explanation



	x	f	C	D
=				=OneVar
3	6	10	$\Sigma x$	223.
4	7	10	$\Sigma x^2$	1287.
5			$s_x := s_{n-...}$	1.05945
6			$\sigma_x := \sigma_{n-...}$	1.04612
7			n	40.

D1 = "One-Variable Statistics"

**Be aware**

It is easy to make a mistake in data entry when using your GDC to calculate statistical measures. Always double check your input. As well, once the analysis is done by the GDC, check that the number of data points  $n$  matches the number of data points in the question, which may be mentioned in the question or is, otherwise, easy to find.

**International Mindedness**

Imagine you are an economist researching market trends across the globe, or a political strategist trying to determine the impact of a trade agreement on the several countries involved. While two countries may be drastically different in population and economics, statistics can still be used to make comparisons and determine how certain factors impact the country. How do you think measures of



Student  
view





Overview  
(/study/ap  
122-  
cid-  
754029/k

centre and dispersion might be used to compare or contrast very large countries with smaller countries? Or how might prosperous countries be compared with less prosperous countries?

## Theory of Knowledge

Statistical analysis in regard to significance as well as measures of dispersion is a key area within the human sciences where mathematics is applied. It is this use of analysis, and drawing conclusions from the mathematical analysis, that contributes to subject areas such as psychology and economics being considering a science.

Psychologists and economists have often had ‘statistically significant’ results of various studies, though the **application** of their findings do not always apply to all people in all contexts. How can this be? If the mathematics was accurate and valid then why would mathematical modelling lack ecological validity?

## 3 section questions ▾

4. Probability and statistics / 4.3 Measures of central tendency and dispersion

# Data transformation

Section

Student... (0/0)



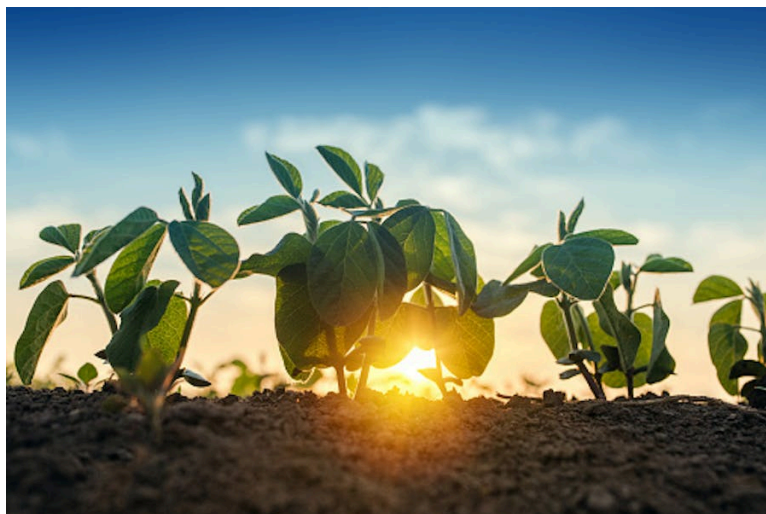
Feedback



Print (/study/app/m/sid-122-cid-

Assign

754029/book/data-transformation-id-26236/print/)



Growing plants

Credit: stevanovicigor Getty Images



Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k

Imagine you have been collecting agricultural data by measuring the heights of several plants at various stages throughout their growth. After completing your entire experiment, you realise that the measuring stick you used has been cut off and is missing the first 3 cm. Is all your work wasted? Do you need to plant all new plants and collect all the data again?

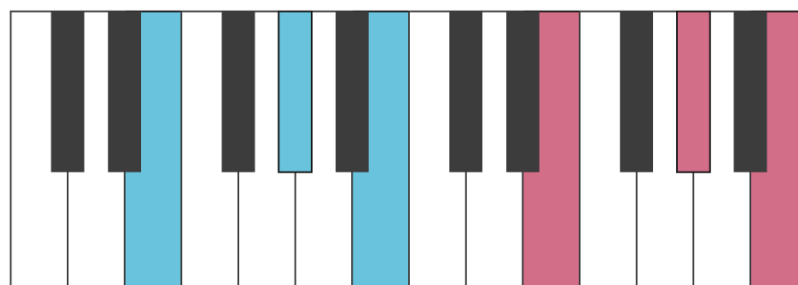
Thankfully, no. All you need to do is to subtract 3 cm from all your measurements. But what will that do to the statistics you've already calculated? Will the mean change? What about the standard deviation? Will it get bigger or smaller, or will it stay the same?

### ✓ Important

When *adding or subtracting* the same value to all the data in a set, the mean will increase or decrease by the same amount, but the standard deviation will remain the same. This is because the distance between the data values has not changed. They have not got any further apart or closer together; they have just shifted up or down.

### 🔗 Making connections

The change described here is similar to what a pianist does when playing a chord in different octaves. The chord sounds the same, because the notes are all still the same distance apart. This pitch is different, because they are playing in a different octave.



Student  
view

More information



Overview  
(/study/app/  
122-  
cid-  
754029/k

This image is a diagram of a piano keyboard, featuring several keys highlighted with different colors. The layout follows the standard pattern of a piano, alternating between black and white keys. Several white keys are colored in blue and pink, indicating specific notes or keys of interest. White keys are between sets of two and three black keys, mimicking the natural arrangement. The blue keys are second and fourth in the sequence, and pink keys are sixth and eighth, correlating with certain notes or functions on the keyboard. This visual representation may be used to demonstrate a musical concept or lesson related to piano playing or music theory.

[Generated by AI]

Now suppose the measuring stick was printed with the wrong scale, and all the measurements need to be multiplied by 2. How will this change the mean? Will it affect the variance or the standard deviation, and if so, how?

### ✓ Important

If each data point is *multiplied or divided* by a constant factor, then the spread *does* change. This is because as values are multiplied, the distances between them get multiplied as well.

So in our example, the mean height will be doubled and the distance between each height and the mean height will also be doubled. What will be the effect of this on the variance and the standard deviation? Remember that the variance is based on the square of the distance from each value to the mean, and the standard deviation is the square root of the variance.

For a data set  $\{x\}$  with mean  $\mu_x$ , standard deviation  $\sigma_x$  and variance  $\sigma_x^2$ , if every value in  $\{x\}$  is multiplied by a constant  $c$ , then:

- the mean of the transformed data set is  $c\mu_x$
- the standard deviation of the transformed data set is  $|c| \sigma_x$
- the variance of the transformed data set is  $c^2 \sigma_x^2$ .

## Example 1



Student  
view





Overview  
(/study/ap  
122-  
cid-  
754029/k

A set of data has a mean of 11 and a standard deviation of 2.3. If every value in the set is decreased by 5, write down the mean and standard deviation of the new set of data.

When every value in a set of data is increased or decreased by the same amount, the mean shifts by the same amount, but all the values will be the same distance apart. Since the data were all decreased by 5, the mean will now be  $11 - 5 = 6$  and the standard deviation will stay at 2.3.

## Example 2



The same set of data, with a mean of 11 and a standard deviation of 2.3, is modified so that every value in the set is multiplied by 4. Find the mean and standard deviation of the new set of data.

When every value in a set of data is multiplied by the same number, the mean is also multiplied by that amount, as well as the distances between each of the data. Since the data were all multiplied by 4, the mean will now be  $11 \times 4 = 44$  and the standard deviation will increase to  $2.3 \times 4 = 9.2$ .

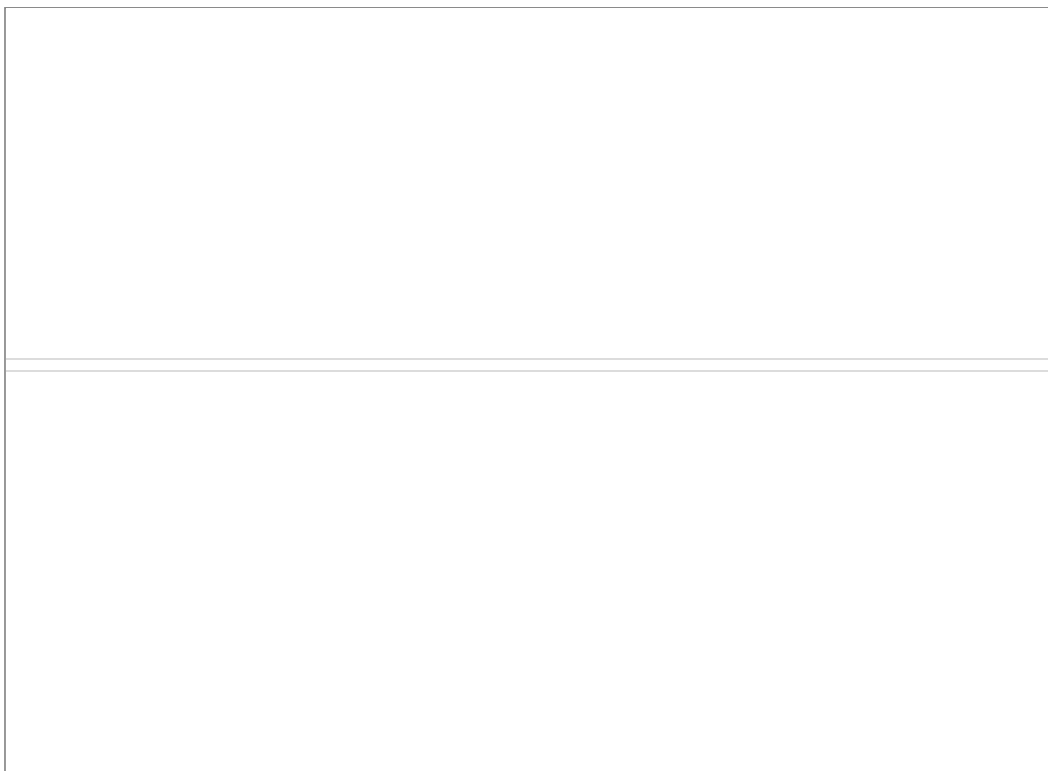
You can see this relationship for yourself by examining how the mean and standard deviation change using the applet below. Can you predict the new mean and standard deviation before finding the values with the applet?



Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k



### Interactive 1. Mean and Standard Deviation.

Credit: GeoGebra  (<https://www.geogebra.org/m/UcXWCHfW>) Maths Learning Centre University of Adelaide

 More information for interactive 1

This interactive allows users to explore how linear transformations affect a data set by adjusting the values of constants 'a' (ranging from -5 to 5) and 'b' (ranging from -10 to 10). Also users can move red points on the first graph, which dynamically changes the dataset's mean and standard deviation. A second graph displays a target distribution shown as a purple point.

Users can apply transformations of the form  $aX + b$  to the data set, using the 'a' and 'b' sliders to apply linear transformations, the user's goal is to match the transformed data's position to the purple point's distribution, effectively recreating the target statistical properties through proper scaling and shifting adjustments.

For example, we start with an original dataset (X) having a mean of 2 and a standard deviation of 3, as shown on the first number line. When we apply the linear transformation  $aX + b$  with  $a = 1.1$  (scaling factor) and  $b = -0.6$  (shifting factor), the tool dynamically updates the statistics:

The new mean becomes 1.6 (calculated as:

$$a \times \text{original mean} + b = 1.1 \times 2 + (-0.6) = 1.6).$$

The new standard deviation becomes 3.3 (calculated as:  $|a| \times \text{original sd} = 1.1 \times 3 = 3.3$ ).

The second graph displays the transformed distribution, visually demonstrating how: Scaling (a): Stretches ( $a > 1$ ) or compresses ( $0 < a < 1$ ) the spread of data, affecting the standard deviation and Shifting (b): Moves the entire



Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k

distribution left ( $b < Q$ ) or right ( $b > Q$ ), altering the mean without changing the spread.

Users can further experiment by adjusting  $a$  and  $b$  to observe real-time changes in the statistical measures and their visual representation on the number line.

## 3 section questions ▾

4. Probability and statistics / 4.3 Measures of central tendency and dispersion

# Checklist

Section

Student... (0/0)



Feedback



Print (/study/app/m/sid-122-cid-

754029/book/checklist-id-26237/print/)

Assign



## What you should know

By the end of this subtopic you should be able to:

- calculate the mean of a frequency distribution by hand or with a calculator
- use mid-interval values to estimate the mean of data grouped in intervals
- find the variance and standard deviation of a frequency distribution with a calculator
- find the new mean, variance and standard deviation when a set of data is transformed by addition, subtraction, multiplication or division.

4. Probability and statistics / 4.3 Measures of central tendency and dispersion

# Investigation

Section

Student... (0/0)



Feedback



Print (/study/app/m/sid-122-cid-

754029/book/investigation-id-26238/print/)

Assign

Here's a problem-solving challenge you can do with a friend.



Student  
view



Overview  
(/study/ap  
122-  
cid-  
754029/k

---

1. Create a set of data, with or without frequencies depending on the level of challenge you want.
2. Calculate the mean and standard deviation of the set and check your values with the GDC.
3. Roll dice, use a random number generator, or have a friend choose a new mean and standard deviation for you to use as your target.
4. Transform your set of data by adding, subtracting, multiplying or dividing the set by constants until the mean and standard deviation equal the target.
5. You and a friend can make it a game to see who can work it out faster or who can come up with the most original ways to reach the target.

### Rate subtopic 4.3 Measures of central tendency and dispersion

Help us improve the content and user experience.



Student  
view