

Statistics

1. Using Formulas

There are many formulas used in statistical calculations. Here we will go through some of the most common.

CALCULATING THE MEAN

You probably learned fairly early on about finding the average of a set of numbers. To calculate this *average*, you add up all the numbers then divide that total by how many numbers you added together. This average is called the *mean*.

You can find the mean of 2 numbers or of 1000 numbers or of any set of numbers. We often refer to sets of numbers as *data*, and the numbers in the set *data points*, or *values*, or even *scores*. So, to find the mean of a set of scores (numbers), add all the scores together and divide the total by the number of scores.

A formula is a shorthand way of writing the method of calculating whatever you want to calculate. In general, we can label our data by the variable x . We can then use a subscript to tell us which score in the data we are looking at. For example, let's take a very simple set of scores: 11, 13, 16, 18, 19. The first score in this set is 11, so we could write $x_1 = 11$. The second score is 13, so $x_2 = 13$, and so on until $x_5 = 19$.

To find the mean of this set of scores, the first step would be to add 11, 13, 16, 18, and 19 together. That is, we would add $x_1, x_2, x_3, x_4,$ and x_5 . Now, imagine writing all these out, particularly if there were over 100 scores – it would be very long! There is a much quicker way to write a sum, and that is by using the Greek letter \sum (pronounced “sigma”).

$$\sum_{i=1}^n$$

The symbol means add up the following with i starting at 1 and ending at n . n is used to represent the number of things we are adding (e.g. the number of values in our data set) and i is used as a counter, it tells us how many things we are adding and can also be used to change what we are adding. For example:

$$\sum_{i=1}^4 (2i + 1)$$

Means add up $2i + 1$ starting with $i = 1$ and ending with $i = 4$. i.e.

$$\begin{aligned} \sum_{i=1}^4 (2i + 1) &= \overset{i=1}{(2 \times 1 + 1)} + \overset{i=2}{(2 \times 2 + 1)} + \overset{i=3}{(2 \times 3 + 1)} + \overset{i=4}{(2 \times 4 + 1)} \\ &= 3 + 5 + 7 + 9 \\ &= 24 \end{aligned}$$

In our example,

$$\sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

and so

$$\sum_{i=1}^5 x_i = 11 + 13 + 16 + 18 + 19 = 77$$

Notice how much easier it is to write the sum using the sigma notation. We can also start and end our counter at whatever numbers we like. Sometimes you just see $\sum x$, which means add up all the data, and sometimes you see $\sum x_i$, which again means add up all the data values (sometimes when all possible i values are to be used its start and end values are omitted). You may even see $\sum_{i=1}^n x_i$, where the start and end values of the counter are written as sub and super scripts, this is just done to save vertical space.



To find the mean of our scores, now we have to divide our total by 5. Our mean would be $77 \div 5$, which can also be written as $\frac{77}{5}$, and is equal to 15.4. In symbols, we have added up all the scores ($\sum_{i=1}^n x_i$) and divided by the number of scores (n), so the formula for finding the mean of a set of scores is:

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

(Note, the mean can either be called μ or \bar{x} , depending on the type of scores involved (population or sample, respectively). μ is the Greek letter pronounced “mu”, and we say “ex-bar” for “ \bar{x} ”.)

There is another way to work out n and the mean. Instead of all the scores being different, as they are in the above example, we might have some that are the same. For example, here is another set of scores: 11, 11, 11, 16, 18, 18. Here there are three 11s, one 16 and two 18s. The frequency of each score is as follows: 11: 3; 16: 1; 18: 2. We usually summarise the frequencies in a table:

Score (x)	Frequency (f)
11	3
16	1
18	2

If we add the frequencies, we get the same as if we had just counted each score in turn – there are 6 scores in total, and so:

$$n = \sum_{i=1}^3 f_i = 6$$

Our table then becomes:

Score (x)	Frequency (f)
11	3
16	1
18	2
	$\sum_{i=1}^3 f_i = 6$

To find the mean from the table, we would see that there are three 11s, which would make 33 when we add all the scores together, there is one 16 and two 18s (which would make 36), so we would add 33, 16, and 36 to get 85. We can do this using a new column in the table:

Score (x)	Frequency (f)	$f \times x$
11	3	33
16	1	16
18	2	36
	$\sum_{i=1}^3 f_i = 6$	$\sum_{i=1}^3 f_i \times x_i = 85$

Note that in mathematics the multiplication symbol is often omitted, so that last cell could be written as $\sum_{i=1}^3 f_i x_i$.

We have found the same totals as if we had just added 11, 11, 11, 16, 18, and 18 (the 85) and found how many scores there are (6). The means is therefore $85 \div 6 \approx 14.17$ (we have rounded to two decimal places and used the wavy equals sign to indicate this is only approximate).



So the formula for the mean can also be written as:

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

CALCULATING THE STANDARD DEVIATION

We can also use the sigma notation to give a formula for the standard deviation of a set of scores. The standard deviation is a measure of the variability among a set of scores. There are few formulas and different methods used to calculate the standard deviation. One of the formulas is:

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

Here the standard deviation is represented by the lower case Greek letter “sigma”, σ (as opposed to the upper case “sigma” Σ). This formula is for the population standard deviation. For a sample standard deviation you replace the population mean, μ , by the sample mean, \bar{x} , and the n in the denominator by $n - 1$ (some texts still use n as the denominator, check your lecture notes!).

To calculate the standard deviation using this formula, first we have to find the mean by adding up all the scores and dividing by the number of scores in the set. Then we subtract the mean from each score and square the result of each subtraction, then we add all the squares together. We then divide the result by the number of scores in the set, before finally taking the square root. It is easiest to use a table to carry out each calculation in turn (unless you use your calculator and simply type in everything before pressing the equals button).

Here is a small set of scores that we will find the mean and standard deviation of. 4, 8, 9, 11, 16. Firstly, we see that there are five scores, so $n = 5$. Now let’s set up a table:

Score (x)	Score – Mean ($x - \mu$)	$(x - \mu)^2$
4		
8		
9		
11		
16		
$\sum_{i=1}^5 x_i = 48$		

The last row will be used as our summation line. In this case we have been able to fill in the whole first column. The mean can now be calculated $\mu = 48 \div 5 = 9.6$. Now we can fill in the rest of the table!

Score (x)	Score – Mean ($x - \mu$)	$(x - \mu)^2$
4	$4 - 9.6 = -5.6$	$(-5.6)^2 = 31.36$
8	$8 - 9.6 = -1.6$	$(-1.6)^2 = 2.56$
9	$9 - 9.6 = -0.6$	$(-0.6)^2 = 0.36$
11	$11 - 9.6 = 1.4$	$(1.4)^2 = 1.96$
16	$16 - 9.6 = 6.4$	$(6.4)^2 = 40.96$
$\sum_{i=1}^5 x_i = 48$		$\sum_{i=1}^5 (x_i - \mu)^2 = 77.18$



Note that all numbers in the final column should be positive. If you do these calculations on your calculator you will need to put the brackets around the negative numbers! Also we have not added up the values in the middle column, what do you get if you do? Why?

We now need to divide the sum of squares by n : $77.18 \div 5 = 15.436$ (this is called the variance), and we now take the square root to find σ :

$$\sigma = \sqrt{15.436} = 3.9288675 \dots$$

So the standard deviation of our set is approximately 3.93.

Here is a set of scores for you to try. You can check your answer with the solutions at the end of the resource.

EXERCISES

Find the mean and the standard deviation for the following data:

28, 39, 25, 16, 40, 23, 35, 36, 27, 30.

If you need help with any of the maths covered in this resource (or any other maths topic), you can make an appointment with Learning Development through reception: phone (02) 4221 3977, or Level 2 (top floor), Building 11, or through your campus.



SOLUTIONS TO EXERCISES

We will use our table method.

Score (x)	Score – Mean ($x - \mu$)	$(x - \mu)^2$
28	$28 - 29.9 = -1.9$	$(-1.9)^2 = 3.61$
39	$39 - 29.9 = 9.1$	$(9.1)^2 = 82.81$
25	$25 - 29.9 = -4.9$	$(-4.9)^2 = 24.01$
16	$16 - 29.9 = -13.9$	$(-13.9)^2 = 193.21$
40	$40 - 29.9 = 10.1$	$(10.1)^2 = 102.01$
23	$23 - 29.9 = -6.9$	$(-6.9)^2 = 47.61$
35	$35 - 29.9 = 5.1$	$(5.1)^2 = 26.01$
36	$36 - 29.9 = 6.1$	$(6.1)^2 = 37.21$
27	$27 - 29.9 = -2.9$	$(-2.9)^2 = 8.41$
30	$30 - 29.9 = 0.1$	$(0.1)^2 = 0.01$
$\sum_{i=1}^{10} x_i = 299$		$\sum_{i=1}^5 (x_i - \mu)^2 = 524.9$

So the mean is $299 \div 10 = 29.9$. This was then used to calculate the second and third columns. The standard deviation is the $\sqrt{524.9 \div 10} \approx 7.24$.

