Schoolworkout Maths

Standard Deviation and Variance (raw data)

In statistics it is convenient to summarise a set of data by highlighting some key features. It is common to summarise data using an *average* (such as the mean or median) but it is also helpful to have a measure of the *spread* of the data. Two simple measures of spread are the *range* (i.e. the difference between the largest and smallest values in the data) and the *inter-quartile range* (i.e. the difference between the lower and upper quartiles).

Standard deviation is another measure of spread which is widely used in statistics. The standard deviation gives a measure of how far the data tends to be from the mean value. One formula for the standard deviation is:

$$\text{s.d.} = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

Note that: 1) $\bar{x}$ is the notation used for the mean of a set of data

2) The symbol Σ is the Greek letter sigma – it is used in maths to mean "add up".

The *variance* is also sometimes used. The variance is the square of the standard deviation and so is given by the formula:

$$\text{variance} = \frac{\sum (x-\bar{x})^2}{n}$$

The example below shows how these formulae are used.

Introduction:

Snow White timed each of the seven dwarfs running a race. Their times (in seconds) were as follows:

Dopey: 35 seconds
Doc: 39 seconds
Bashful: 43 seconds
Sleepy: 47 seconds
Grumpy: 41 seconds
Happy: 49 seconds
Sneezy: 40 seconds

The mean of these 7 times is: $\bar{x} = \frac{\sum x}{n} = \frac{35+...+47}{7} = \frac{294}{7} = 42$ seconds.

To find the standard deviation, we can draw up a table:

Data, x	$x-\bar{x} = x - 42$	$(x-\bar{x})^2$
35	-7	49
41	-1	1
39	-3	9
49	7	49
43	1	1
40	-2	4
47	5	25
$\sum x = 294$	$\sum (x-\bar{x}) = 0$	$\sum (x-\bar{x})^2 = 138$

The variance of the dwarfs' times is therefore:

$$\text{variance} = \frac{\sum (x-\bar{x})^2}{n} = \frac{138}{7} = 19.714$$

So the standard deviation is: s.d. = $\sqrt{\text{variance}} = \sqrt{19.714} = 4.44\,\text{sec}$.

Note: Standard deviation is measured in the same units as the original data whereas variance is measured in squared units.

A more useful formula…

There are alternative formulae which are usually simpler to use in order to find the variance or the standard deviation. These are:

$$\text{variance} = \frac{\sum x^2}{n} - \bar{x}^2$$

and

$$\text{s.d.} = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

The steps involved to find the standard deviation therefore are as follows:

Step 1: Square each piece of data
Step 2: Add up these squares (to get $\sum x^2$)
Step 3: Divide by the number of values (to get $\frac{\sum x^2}{n}$)
Step 4: Subtract the square of the mean (to get the variance)
Step 5: Square root (to get the standard deviation)

If we apply these steps to the dwarfs' race times (from page 1) we get:

Step 1:

$35^2 = 1225$ $41^2 = 1681$ $39^2 = 1521$ $49^2 = 2401$ $43^2 = 1849$ $40^2 = 1600$ $47^2 = 2209$

Step 2: So, $\sum x^2 = 12486$

Step 3: Therefore, $\frac{\sum x^2}{n} = \frac{12486}{7} = 1783.714...$

Step 4: So variance = $\frac{\sum x^2}{n} - \bar{x}^2 = 1783.714... - 42^2 = 19.714...$

Step 5: Consequently, standard deviation = $\sqrt{19.714...} = 4.44$ secs (as before)

Usually we show less working as the following example demonstrates:

Worked example

A class sat tests in Statistics and in Pure Mathematics. Their results (expressed as percentages) were as follows:

Statistics mark, x:	45	72	63	59	78	64	51	67
Pure mark, y:	49	85	64	41	73	53	32	55

a) Calculate the mean and standard deviation for each test.
b) Compare the results obtained in Statistics and Pure Maths.

Solution:

a)

For Statistics:

$$\bar{x} = \frac{45 + 72 + ... + 67}{8} = \frac{499}{8} = 62.375$$

To find the standard deviation, the key value is $\sum x^2$:

$$\sum x^2 = 45^2 + 72^2 + ... + 67^2 = 31929$$

So the standard deviation is given by:

$$s.d. = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{31929}{8} - 62.375^2} = \sqrt{100.48...} = 10.0$$

For Pure:

$$\bar{y} = \frac{49 + 85 + ... + 55}{8} = \frac{452}{8} = 56.5$$

The sum of the squares is

$$\sum y^2 = 49^2 + ... + 55^2 = 27590$$

So the standard deviation is:

$$s.d. = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2} = \sqrt{\frac{27590}{8} - 56.5^2} = \sqrt{256.5} = 16.0$$

b) When comparing two sets of data, it is important to compare the values of both the mean and the standard deviation using the context of the question.

In this case, we can conclude that:

i) students generally achieved higher marks in Statistics (as shown by the higher mean);
ii) the standard deviation was higher for the Pure marks indicating that there was greater variation in the students' performances in the Pure test than in the Statistics test.