Calculating Standard Deviation

A simple set of numeric data…

i	1	2	3	4	5	6	7	8	9	10
x_i	1	11	13	14	15	16	16	17	18	29

Find…

a) The number of pieces of data (this is n)

b) The mean (referred to as $\bar{x}$).

c) The range

d) The interquartile range

The mean ($\bar{x}$) together with the median and mode, are measures of location which indicate whereabouts on the numberline this set of data is centred around.

The range and the interquartile range are measures of spread which indicate how spread out the data is. However, both of these measures have issues with them. The range is determined only by the lowest and highest numbers in a set of data and does not take any other of the other data into account. It is affected by outliers. The IQR is better than the range as it is not affected so much by outliers but neither does it take the value of each and every piece of data into account.

Introducing standard deviation…

This is a measure of the average distance from the mean of every piece of data in the data set. i.e., it's the *standardised amount of deviation from the mean* of the data. Importantly, every piece of data contributes to and affects the standard deviation so it's a good summary statistic of the whole data set.

Complete the table below…

i	1	2	3	4	5	6	7	8	9	10
x_i	1	11	13	14	15	16	16	17	18	29
$x_i - \bar{x}$										
$(x_i - \bar{x})^2$										

Find…

e) $\sum x_i - \bar{x}$

f) $\sum (x_i - \bar{x})^2$

g) $\frac{1}{n}\sum (x_i - \bar{x})^2$

The formula and value found in part (g) above is the variance of the data. This is the *spread of the squares of the deviations from the mean*. It therefore makes sense to square-root this figure to obtain the *spread of deviations from the mean*, and it is this which is the standard deviation of a data set.

Standard deviation is given the notation σ (lower case sigma) and therefore variance has the notation σ^2.

$$\text{Variance} = \sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$$

$$\text{Standard deviation} = \sigma = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$$

There are alternate, and more useful, formats of these formulae which are…

$$\text{Variance} = \sigma^2 = \frac{1}{n}\sum(x_i)^2 - \bar{x}^2$$

$$\text{Standard deviation} = \sigma = \sqrt{\frac{1}{n}\sum(x_i)^2 - \bar{x}^2}$$

The measurement of standard deviation is the same as that of the data. I.e. if the data is age in years then the standard deviation will be age in years etc.

Getting more technical…

When dealing with grouped data, such as with a grouped frequency table, the formula for variance becomes

$$\text{Variance} = \sigma^2 = \frac{1}{n}(\sum fx^2) - \left\{\frac{(\sum fx)}{n}\right\}^2$$

The formulae for calculating standard deviation of a sample is subtly different to when calculating that of a population and involves dividing by $n - 1$ instead of n.

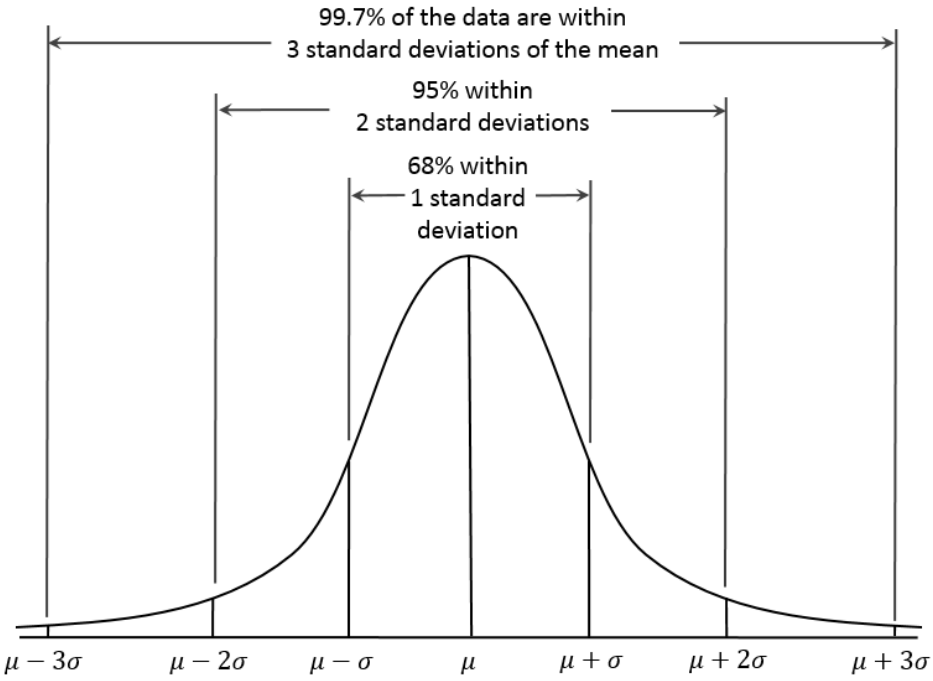
Sample	Population
$s^2 = \frac{1}{n-1}\sum(x - \bar{x})^2$	$\sigma^2 = \frac{1}{n}\sum(x - \mu)^2$
$s^2 = \frac{1}{n-1}\left\{\sum x^2 - \frac{(\sum x)^2}{n}\right\}$	$\sigma^2 = \frac{1}{n}\left(\sum x^2\right) - \mu^2$
$s^2 = \frac{1}{n-1}\left\{\sum fx^2 - \frac{(\sum fx)^2}{n}\right\}$	$\sigma^2 = \frac{1}{n}\left(\sum fx^2\right) - \left\{\frac{(\sum fx)}{n}\right\}^2$

When using standard deviation as part of a normally distributed set of data then the empirical rule states that…

1 standard deviation includes 68% of data

2 standard deviations include 95% of data

3 standard deviations include 99.7% (‘almost all’) data



Things to do now…

1. Practice using this formulae
2. Learn how your calculator can work out standard deviation and variance for you
3. See how the alternate formulae can be derived from the originals.
4. Read about the Empirical Rule.