

9

Standard deviation

We have seen that the interquartile range indicates the variation of data where the median is the measure of central tendency. Standard deviation is used where this measure is the *mean*. It indicates the difference between a group of values and their mean, taking *all* of the data into account. Although this means that it may be influenced by extreme values, the standard deviation plays an important role in many tests of statistical significance (which will be described in later chapters). The larger the standard deviation, the more the values differ from the mean, and therefore the more widely they are spread out.

For example, one small group of patients in a particular outpatient clinic may wait for a mean time of 11 minutes to be seen by a doctor, and the standard deviation from the mean for this group is 5.701. Individual waiting times vary widely – from 7 minutes up to 21 minutes. There is wide variation between these waiting times, and they are quite widely spread out from their mean. These waiting times are therefore *heterogeneous* or dissimilar.

On another day, another group of patients from the same clinic may also have a mean waiting time of 11 minutes, but their standard deviation is 0.707. This is much less than the first group's standard deviation of 5.701. Looking at this group's actual waiting times, it can be seen that they only vary from 10 to 12 minutes. Waiting times for the second group are more *homogeneous* – that is, the data are more similar to each other. They are less widely spread out around their mean than the first group.

Let us look at the actual waiting times recorded for each group, as shown in Table 9.1.

You can see that the data in Group 1 are much more spread out than those in Group 2. This difference in standard deviations can be explained by the fact that, although most patients in Group 1 waited a very short time, one patient had to wait for a long time (21 minutes). Although this

Table 9.1: Waiting times and standard deviation for each patient group

Group	Time 1	Time 2	Time 3	Time 4	Time 5	Mean	Standard deviation
1	10	7	8	9	21	11	5.701
2	11	11	10	11	12	11	0.707

one 'outlier' waiting time is not representative of the whole group, it has a large effect on the overall results, and it strongly affects the mean and standard deviation. Several patients from Group 2 actually waited longer than Group 1 patients, although the difference between the waiting times in Group 2 is very slight.

Although the abbreviations *SD* or *s.d.* are used to represent standard deviation generally, *s* is used to represent standard deviation for *samples*, and σ is used to represent standard deviation for *populations*.

The most usual formula for standard deviation is as follows:

$$\sqrt{\sum (x - \bar{x})^2 / (n - 1)}$$

where x = individual value, \bar{x} = sample mean and n = number of values.

The above equation is only suitable for a sample (or *population estimate*). This will usually be the case, since we rarely know the true population value (which in this case is the mean).

The following steps are used to work out a standard deviation.

- 1 Find the mean of the group.
- 2 Subtract this from every value in the group individually – this shows the deviation from the mean, for every value.
- 3 Work out the square (x^2) of every deviation (that is, multiply each deviation by itself, e.g. $5^2 = 5 \times 5 = 25$) – this produces a squared deviation for every value.
- 4 Add up all of the squared deviations.
- 5 Add up the number of observed values, and subtract 1.
- 6 Divide the sum of squared deviations by this number, to produce the *sample variance*.
- 7 Work out the square root of the variance.

If you have to work out a standard deviation by hand, it is helpful to use a grid like the one shown in Table 9.2. We can use this to work out the standard deviation of the data for Group 1 from Table 9.1.

Table 9.2: Grid showing preliminary calculations for standard deviation

<i>Value number</i>	<i>Time (a)</i>	<i>Mean time (b)</i>	<i>Deviation from the mean (a – b)</i>	<i>Squared deviation (a – b)²</i>
1	10	11	–1	1
2	7	11	–4	16
3	8	11	–3	9
4	9	11	–2	4
5	21	11	10	100
				Total = 130

- 1 We already know the mean is 11 (*see* page 30).
- 2 Subtract each time value from the mean. Note each result in the ‘Deviation from the mean’ column.
- 3 Multiply each deviation by itself, and write each result in the ‘Squared deviation’ column (e.g. $-4^2 = -4 \times -4 = 16$) (note that multiplying *minus* numbers produces *positive* ones).
- 4 Adding all of the squared deviations ($1 + 16 + 9 + 4 + 100$) gives a value of 130.
- 5 There are five separate values in the group. Subtract 1, and you get 4.
- 6 Divide the sum of squared deviations by 4, to produce the variance ($130/4 = 32.5$).
- 7 Use a calculator to determine the square root of the variance (32.5) – that is $\sqrt{32.5} = 5.701$.

Of course, calculating standard deviation by hand like this is not practical if you have a large number of values. Moreover, the mean is unlikely to be a whole number as it is in the above example. Calculators and computer programs are an invaluable aid to this process, and are readily available.

Other uses of standard deviation are discussed under normal distribution (*see* Chapter 11).

