8 Calculation of the Standard Deviation

Twenty observations of systolic blood pressure were given in Table 9 (reproduced on p. 87) and their mean value was found to be 128 mm. The variability of these observations was measured by means of the standard deviation. This value, it may be noted, is sometimes referred to in short as the S.D. and sometimes designated by the Greek letter sigma (σ).* Its actual value was calculated in Chapter 7 by (1) finding by how much each observation differed from the mean, (2) squaring each of these differences, (3) adding up these squares, and dividing this total by the number of observations minus one, (4) taking the square root of this number (or, taking the processes in reverse order, 'the root mean square deviation,' an old name for the standard deviation). This method of calculation would have been much more laborious if the mean blood pressure had not been a whole number – e.g. if it had been 128·4 – and if each of the original observations had been taken to one decimal place (presuming such a degree of accuracy to be possible) – e.g. the first had been 98·7. The differences between the observations and their mean, and the squares of these values, would then have been less simple to calculate. But in such cases the necessary arithmetic can still be kept simple by a slight change of method.

The Ungrouped Series

Observations from Table 9

Twenty Observations of Systolic Blood Pressure in mm	Deviation of each Observation from the Mean (Mean = 128)	Square of each Deviation from the Mean
(1)	(2)	(3)
98	− 30	900
160	+ 32	1024
136	+ 8	64
128	0	0
130	+ 2	4
114	− 14	196
123	− 5	25
134	+ 6	36
128	0	0
107	− 21	441
123	− 5	25
125	− 3	9
129	+ 1	1
132	+ 4	16
154	+ 26	676
115	− 13	169
126	− 2	4
132	+ 4	16
136	+ 8	64
130	+ 2	4
Sum 2560	0	3674

If we call each individual observation x and the mean of all 20 we call $\bar{x}$ (pronounced x-bar), then by the method of Table 9, on the opposite page, we must first find each separate deviation from the mean, $(x - \bar{x})$, as in column (2), and then we must calculate the square of each of these deviations, $(x - \bar{x})^2$, as in column (3). The required sum of all the squared deviations in column (3) can then be computed (3674) and may be described as *Sum* $(x - \bar{x})^2$. We can, however, reach this sum ***without calculating any deviations at all*** by means of the relationship:—

$$\textit{Sum}\ (x - \bar{x})^2 = \textit{Sum}\ (x^2) - (\textit{Sum}\ x)^2/n.$$

Thus in practice we square each observation, x, as it stands, as in column (2) of Table 11, and we find the sum of these squares; thus *Sum* $(x^2) = 331\ 354$. In calculating the mean we have already found the sum of the 20 observations themselves; thus, from column (1), *Sum* $x = 2560$. Our required sum of squared deviations round the mean, *Sum* $(x - \bar{x})^2$, is therefore $331\ 354 - (2560)^2/20 = 3674$. The standard deviation is then, as before, $\sqrt{3674 \div 19} = 13{\cdot}91$ mm (using for the reasons given previously $n - 1$ as the divisor).

Thus to calculate the standard deviation in a short ungrouped series

* In some circumstances σ is reserved for the unknown standard deviation of the universe sampled and s is used for the standard deviation estimated from the actual sample of observations under discussion.

TABLE 11

CALCULATION OF STANDARD DEVIATION: UNGROUPED SERIES

Twenty Observations of Systolic Blood Pressure in mm	Square of each Observation
(1)	(2)
98	9 604
160	25 600
136	18 496
128	16 384
130	16 900
114	12 996
123	15 129
134	17 956
128	16 384
107	11 449
123	15 129
125	15 625
129	16 641
132	17 424
154	23 716
115	13 225
126	15 876
132	17 424
136	18 496
130	16 900
Sum 2560	331 354

of figures we have five steps: (*a*) Square the *individual observations* themselves and find the sum of these squares. (*b*) Square the *sum* of the observations themselves and divide this by the total numbers of observations available. (*c*) Subtracting (*b*) from (*a*) gives the required sum of the squared deviations of the observations *around their own mean.* (*d*) Divide the value by $n-1$ to reach the variance and (*e*) take the square root of the variance to reach the standard deviation.

A warning is necessary on the use of this method if the observations vary so little from one another that the standard deviation will be very small in comparison with the mean. In these circumstances *Sum* (x^2) is very nearly equal to $(\textit{Sum}\ x)^2/n$ and the accuracy of the calculation will be lost when the subtraction is made; for most of the digits will cancel out. If the calculations are being made by 'hand' the difficulty will probably be observed but the user of an electronic calculator may fail to realise that anything untoward has occurred, and the user of a computer certainly cannot realise it. To overcome this trouble by 'hand' or with a calculator the actual deviations should be used (as on p.87) or by means of a scale of working units with its zero placed close to the mean of the distribution. For a computer program other methods exist.

The Grouped Series

With a large number of observations this method of squaring each observation separately would be very laborious. A shorter method which will give very nearly the same result can be adopted. The observations must first be grouped in a frequency distribution. As an example we may take the distribution given in Table 8 (p. 75) of the ages at death from diseases of the Fallopian tube. This distribution is given again in column (2) of Table 12.

TABLE 12

CALCULATION OF STANDARD DEVIATION: GROUPED SERIES OF AGES AT DEATH FROM DISEASES OF THE FALLOPIAN TUBE

Age in Years	Number of Deaths in each Age-group	Age in Working Units	(2) × (3)	(3) × (4)
(1)	(2)	(3)	(4)	(5)
0–	1	− 6	− 6	36
5–	0	− 5	0	0
10–	1	− 4	− 4	16
15–	7	− 3	− 21	63
20–	12	− 2	− 24	48
25–	35	− 1	− 35	35
30–	42	0	—	—
35–	33	+ 1	+ 33	33
40–	24	+ 2	+ 48	96
45–	27	+ 3	+ 81	243
50–	10	+ 4	+ 40	160
55–	6	+ 5	+ 30	150
60–	5	+ 6	+ 30	180
65–	1	+ 7	+ 7	49
70–74	2	+ 8	+ 16	128
Total	206	—	+ 195	1237

To reach the mean age at death we could add up the 206 individually recorded ages and divide by 206. But at the risk of making only an immaterial error we can, as shown in Chapter 6, shorten this process by presuming that the individuals belonging to each 5-yearly age-group died at the centre age of that group – e.g. that the 42 women dying at ages between 30 and 35 all died at age 32·5. Some will have died between 30 and 32·5, some, perhaps, at exactly 32·5, some between 32·5 and 35. If the distribution is fairly symmetrical, then, as previously stated, the positive and negative errors we make by this assumption will nearly balance out. The sum of the 206 ages at death will then be $(2{\cdot}5 \times 1) + (12{\cdot}5 \times 1) + (17{\cdot}5 \times 7) + (22{\cdot}5 \times 12) + \ldots + (62{\cdot}5 \times 5) + (67{\cdot}5 \times 1) + (72{\cdot}5 \times 2) = 7670{\cdot}0$ and the mean age at death is $7670{\cdot}0 \div 206 = 37{\cdot}2$ years. Having found the mean in this way the standard deviation could be found by calculating how much the observations in each group deviate from it and taking the square of this value. For instance the 12 individuals in the age-group 20–25 died, on our assumption, at age 22·5. They differ from the mean, therefore, by (37·2 minus 22·5) or 14·7; the square of this is 216·09, and this value we must take 12 times as there are 12 individuals with that deviation.

Following this procedure, we should reach for the squares of the deviations of the individuals from their mean the following values:—

$$\begin{aligned}
&(-34{\cdot}7)^2 \times 1 + (-24{\cdot}7)^2 \times 1 + (-19{\cdot}7)^2 \times 7 + \\
&(-14{\cdot}7)^2 \times 12 + (-9{\cdot}7)^2 \times 35 + (-4{\cdot}7)^2 \times 42 + \\
&(0{\cdot}3)^2 \times 33 + (5{\cdot}3)^2 \times 24 + (10{\cdot}3)^2 \times 27 + \\
&(15{\cdot}3)^2 \times 10 + (20{\cdot}3)^2 \times 6 + (25{\cdot}3)^2 \times 5 + \\
&(30{\cdot}3)^2 \times 1 + (35{\cdot}3)^2 \times 2 = 26\,310{\cdot}54.
\end{aligned}$$

The sum of these calculations is 26 310·54 and the standard deviation is therefore

$$\sqrt{26\,310{\cdot}54/205} \doteq \sqrt{128{\cdot}34} = 11{\cdot}33 \text{ years.}$$

Short Method, with Grouped Series

This is a possible method of working but, it will be observed, a laborious way. In practice a considerably shorter method is adopted. The principle of this method is merely an extension of that used in Chapter 6 for finding the mean, i.e. instead of working in the real, and cumbersome, units of measurement we translate them arbitrarily into smaller and more convenient units, work the sums in those smaller units, and translate the results back again into the real units at the end.

Let us, for instance, in this case replace 32·5 by 0, 27·5 by −1, 22·5 by −2, and so on, 37·5 by +1, 42·5 by +2, and so on. (The original groups must have, it will be remembered, intervals of equal width; they were all 5-yearly in our example.) Now instead of having to multiply −27·5 by 35, for example, we have the simpler task of multiplying −1 by 35. These multiplications are made in column (4) of Table 12. Their sum, taking the sign into account (as must be done), is +195. The mean in these units is therefore

$$+195/206 = +0{\cdot}947.$$

The standard deviation can be found in these same small units, measuring, for simplicity, the deviations of the observations from the 0 value instead of from the mean. The squares of the deviations in these units are merely 1, 4, 9, 16, etc., and these have to be multiplied by the number of individuals with the particular deviation – e.g. 7 × 9 for the −3 group, 24 × 4 for the +2 group, and so forth. A still simpler process of reaching the same result is to multiply column (4) by column (3), i.e. instead of multiplying 7 by 9 we multiply (7 × −3) by −3. This gives the figures of column (5). The sum of these squared deviations is, then, 1237.

These deviations in working units have, however, been measured round the 0 value, whereas they ought to have been measured round the mean (in working units) of +0·947. The correction is again made by the formula given on p. 87, namely that the required $\mathit{Sum}\ (x - \bar{x})^2$ equals $\mathit{Sum}\ (x^2) - (\mathit{Sum}\ x)^2/n$. Therefore from the values in Table 12 we can calculate $\mathit{Sum}\ (x - \bar{x})^2$ to be $1237 - (195)^2/206 = 1052{\cdot}41$. The standard deviation in working units is, therefore, $\sqrt{1052{\cdot}41/205} = 2{\cdot}265$.

We have now to translate the mean, +0·947, and the standard deviation, 2·265, back into the real units. This is simply done. The mean in working units is +0·947 – i.e. 0·947 working units above our 0. In real units our 0 is equivalent to 32·5, for that is the substitution we made (note, once more, the *centre* of the group against which we placed the 0, not its beginning). The real mean must therefore be 32·2 + 5 (0·947) which equals 37·2 years, the same as the value we found by the long method using real units throughout. (The multiplier 5, it will be remembered, comes from the size of the interval of the original group).

To reach the real standard deviation, all that has to be done is to multiply the standard deviation as found in working units by the original units of grouping – in this case by 5. For if this measure of the scatter of the observations is 2·265 when the range is only 14 units (from −6 to +8) it must be 5 times as much when the range is really 70 units (from 2·5 to 72·5). The real standard deviation is therefore 5 × 2·265 = 11·33 years. (It should be noted that if the original units are *smaller* than the working units then the standard deviation will be smaller in the real units, e.g. the multiplier will be 0·25 if that is the original group interval).

TABLE 13

CALCULATION OF STANDARD DEVIATION: GROUPED SERIES

Age in Years	Number of Deaths in each Age-group	Age in Working Units	(2) × (3)	(3) × (4)
(1)	(2)	(3)	(4)	(5)
0–	1	− 8	− 8	64
5–	0	− 7	0	0
10–	1	− 6	− 6	36
15–	7	− 5	− 35	175
20–	12	− 4	− 48	192
25–	35	− 3	− 105	315
30–	42	− 2	− 84	168
35–	33	− 1	− 33	33
40–	24	0	—	—
45–	27	+ 1	+ 27	27
50–	10	+ 2	+ 20	40
55–	6	+ 3	+ 18	54
60–	5	+ 4	+ 20	80
65–	1	+ 5	+ 5	25
70–74	2	+ 6	+ 12	72
Total	206	—	− 217	1281

Checking the Arithmetic

As regards the final result for the standard deviation, as well as the mean, it is immaterial where the 0 is placed; the same answers in *real* units must be reached. From the point of view of the arithmetic it is usually best to place it centrally so that the multipliers may be kept small. For the sake of demonstration the calculations for Table 12 are repeated in Table 13 taking another position for 0. This, in practice, is a good method of checking the arithmetic.

From the calculations in Table 13 we have:

Mean in working units = −217/206 = −1·053,

∴ mean in real units = 42·5 − 5 (1·053) = 37·2 years

(42·5 is the centre of the group against which the 0 was placed; note that the correction has now to be subtracted, for the sign of the mean in working units is negative).

Sum of squared deviations in working units round the mean is $1281 - (217)^2/206 = 1052\cdot41$ and, therefore, as before, the standard deviation in working units is $\sqrt{1052\cdot41/205} = 2\cdot265$ and in true units $2\cdot265 \times 5 = 11\cdot33$ years.

These values agree with those previously found.

The Standard Deviation in Small Samples

As already pointed out, the standard deviation found for a set of observations is an estimate of the variability of the observations in the population, or universe, that has been sampled and on the average a slightly better estimate is reached by dividing the sum of the squared deviations from the mean by $n - 1$ instead of by n (where n is the number of observations). If the number of observations is large the difference is immaterial; if it is small, some difference results.

An arithmetical demonstration of the advantage, on the average, of basing the calculation upon one less than the total number of observations is given in Table 14. One hundred samples, each containing 5 individuals, or observations, were drawn from a 'universe.' In this 'universe' the 'persons' could have any value from 0 to 9. We might imagine, as in the next chapter, that the value denoted the number of colds a person had had in the previous 12 months.

The 'universe' used was composed of *Random Sampling Numbers,* such as are given on pp. 305 to 312. Within such a universe the numbers 0, 1, 2, 3 up to 9 should occur with equal frequency, and it can, therefore, be calculated that its mean is 4·50 and its standard deviation 2·87. If a *large* sample be drawn from it, it will (almost certainly) be found that the mean and standard deviation of that sample do not differ appreciably from those values. With small samples they may differ appreciably. This question is discussed in detail in the next chapter, and the present issue is merely whether the standard deviation of the sample is likely to be nearer the truth when it is based upon one less than the number of observations than when it is based upon the total number. Table 14 shows the number of times a standard deviation of a given size was seen to occur in a sample of 5 'persons.' In column (2) the divisor was based upon all 5 observations, and it will be seen that of the 100 S.D.'s 69 fell below the real value of 2·87 and only 31 were larger than the real value (including here one which was exactly the correct value). In other words, there is a greater chance that the sample S.D. will be too low than too high when it is based upon the total number of observations. The average value for these hundred S.D.'s is 2·50, i.e. somewhat below the real value of 2·87.

On the other hand, when the divisor is based upon 4 observations, or one less than the total number of 5, the distribution of S.D.'s becomes much more equally spread on either side of the true value, 48 now being below and 52 above the real value. The average value for the 100 S.D.'s has become 2·81, i.e. very close to the real value.

Two points should, however, be noted.

(1) The improvement in the sample S.D. as an estimate of the universe S.D. when the former is based upon $n-1$ is *only an average improvement.* If the S.D. of the sample is based upon *all* the observations and, already in this form, is *larger* than the real S.D. of the universe, then basing the estimate upon $n-1$ must make it still larger and therefore still more distant from the truth. The point is that there are more values too low than too high, so that we have an average improvement by increasing all the values. Also, for the tests of significance discussed in subsequent chapters it is, on the whole, wiser to have an over-stated than an under-stated standard deviation.

(2) The second point worth noting is that, with such a very small sample, the standard deviation may, naturally, differ greatly from the real value.

TABLE 14

A COMPARISON OF THE STANDARD DEVIATIONS OCCURRING IN 100 SAMPLES OF 5 OBSERVATIONS

When based upon

(*a*) The Number of Observations in the Sample, and
(*b*) One Less than the Number of Observations

Size of the S.D. observed in the Sample	Number of Times an S.D. of the given size occurred when the calculation was based upon—			
	The Number of Observations = 5		One Less than the Number of Observations = 4	
(1)	(2)		(3)	
0·62–	2		2	
0·87–	2		0	
1·12–	0		1	
1·37–	4		1	
1·62–	11		4	
1·87–	11		10	
2·12–	13		11	
2·37–	9		13	
2·62–	17	69	6	48
2·87–	13		17	
3·12–	8		12	
3·37–	7		8	
3·62–	2		8	
3·87–	1		5	
4·21–4·37	0	31	2	52
Total		100		100
Average S.D. Value	2·50		2·81	