PRINCIPLES OF MEDICAL STATISTICS

XVI—CALCULATION OF THE STANDARD DEVIATION *

In Table III, which appeared in the article on the variability of observations and is here reprinted, there are given twenty observations of systolic blood pressure of which the mean value was found to be 128. The variability of these observations was measured by means of the standard deviation.

TABLE III

(*Reprinted from* THE LANCET, *Jan. 23rd, 1937, p. 219*)

Twenty observations of systolic blood pressure.	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

This value was calculated by (1) finding by how much each observation differed from the mean, (2) squaring each of those differences, (3) adding up these squares, and finding their mean by dividing by the number of observations, (4) taking the square root of this number. Putting this in symbols, if the number of observations is n, each observation is designated by x, and the mean of them by $\bar{x}$, then the standard deviation equals

$$\sqrt{\frac{\text{Sum of values of } (x - \bar{x})^2}{n}}.$$

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—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

Instead of measuring the differences between the observations and their mean we can first take those differences from some other point, *any* point which makes the calculation simpler, and make a correction at the end for having done so. For instance, taking the figures of Table III, instead of calculating the differences between the observations and their mean value, 128, let us measure the differences between the observations and 100. These differences are given in column (2) of Table III A and their squares in column (3). The sum of the squared deviations from 100 is 19,354 and the mean squared difference is, therefore, 19,354 ÷ 20 = 967·7. To this value we must now make a correction for having measured the deviations from 100 instead of from the mean of 128. The correction is to subtract from this mean square value of 967·7, the square of the distance between the value from which we chose to measure the deviations (100 in this case) and the value from which we ought to have measured them (128). Thus we have 967·7 minus $(128-100)^2$, or $(28)^2$,= 967·7 minus 784, which gives 183·7. The standard deviation is, then, $\sqrt{183{\cdot}7} = 13{\cdot}55$, the value we reached before by taking the deviations from the mean itself.

TABLE III A

Calculation of Standard Deviation : Ungrouped Series

Twenty observations of systolic blood pressure.	Deviation of each observation from 100.	(Deviation.)²	Square of observation.
(1)	(2)	(3)	(4)
98	− 2	4	9,604
160	+60	3,600	25,600
136	+36	1,296	18,496
128	+28	784	16,384
130	+30	900	16,900
114	+14	196	12,996
123	+23	529	15,129
134	+34	1,156	17,956
128	+28	784	16,384
107	+ 7	49	11,449
123	+23	529	15,129
125	+25	625	15,625
129	+29	841	16,641
132	+32	1,024	17,424
154	+54	2,916	23,716
115	+15	225	13,225
126	+26	676	15,876
132	+32	1,024	17,424
136	+36	1,296	18,496
130	+30	900	16,900
Sum 2560	—	19,354	331,354

If the observations all lie near one hundred this is a convenient method of working, for the deviations are thus reduced to a size which it is easy to handle and the squares can often be done in one's head. On the other hand one has to make subtractions from 100 to obtain the deviations. Even this step can be eliminated by measuring the deviations of the observations from zero—i.e., by squaring the observations themselves, as is done in column (4). The squares can be taken from a book of tables (e.g., Barlow's Tables of Squares, Cubes, Square Roots, &c. London : E. and F. Spon. 1930. 7*s.* 6*d.*)

This obviates finding any deviations at all.

The sum of these squares is 331,354, and the mean square is 331,354 ÷ 20 = 16,567·7. In using the squares of the observations themselves we have measured their deviations from 0 instead of from the mean value of 128. Therefore the distance between the value from which we chose to measure the deviations and the value from which we ought to have measured them is 128 ; as correction we must, then, subtract $(128)^2$ from our mean square value. This gives 16,567·7 minus 16,384 = 183·7, and the standard deviation is $\sqrt{183{\cdot}7} = 13{\cdot}55$ as before. To calculate the standard deviation in a short ungrouped series of figures the procedure is, then, as follows : (1) find the mean of the observations ; (2) square each

* In accordance with many requests, I am adding to this series, of which the main argument was concluded in THE LANCET of last week, two additional articles on the calculation of (1) the standard deviation, and (2) the correlation coefficient. The latter will appear next week.

observation; (3) sum these squares and find their mean; (4) from this mean square subtract the square of the mean; (5) the square root of this last value is the standard deviation.

The standard deviation therefore equals :—

$$\sqrt{\frac{\text{sum of squares of observations}}{\text{number of observations}} \text{ minus } (\text{mean of observations})^2}$$

or in symbols is $\sqrt{\frac{\text{sum of } (x)^2}{n} - (\bar{x})^2}$.

(The proof of the correction is quite simple but the worker who wishes to apply the method has no need to worry about it.)

The Grouped Series

With a large number of observations this method of squaring each observation 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 II (see *Lancet*, Jan. 23rd, p. 219) of the ages at death from diseases of the Fallopian tube. This distribution is given again in column (2) of Table III B.

TABLE III B

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	−6	− 6	36
5–	—	−5	—	—
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
70–75	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 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 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 $(2{\cdot}5 \times 1) + (12{\cdot}5 \times 1) + (17{\cdot}5 \times 7) + (22{\cdot}5 \times 12) + \ldots\ldots + (62{\cdot}5 \times 5) + (67{\cdot}5 \times 1) + (72{\cdot}5 \times 2) = 7670{\cdot}0$ 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 14·7 (37·2 minus 22·5), the square of which 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 :—

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

The standard deviation is, therefore,

$$\sqrt{26{,}310{\cdot}54/206} = \sqrt{127{\cdot}72} = 11{\cdot}30.$$

SHORT METHOD, WITH GROUPED SERIES

This is a possible method of working but, it will be observed, a somewhat laborious way. In practice a much shorter method is adopted. The principle of this method is that 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, 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 be of equal size; 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 III B. 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 the deviations of the observations from the 0 value instead of from the mean for simplicity. The squares of the deviations in these units are merely 1, 4, 9, 16, &c., 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 simpler process still of reaching the same result is to multiply column (4) by column (3), (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 and their mean is 1237/206 = 6·0049.

These deviations in working units have been measured round the 0 value, whereas they ought to have been measured round the mean (in working units) of + 0·947. The correction, as stated before, is to subtract the square of the distance between the value round which the deviations ought to have been measured and the value round which they were in fact measured; in this case the distance is 0 − 0·947 = −0·947. The standard deviation in working units is therefore $\sqrt{6{\cdot}0049 - (-0{\cdot}947)^2} = 2{\cdot}26$.

We have now to translate the mean, +0·947, and the standard deviation, 2·26, 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, the *centre* of the group against which we placed the 0, not its beginning, a mistake which is somewhat easy to make). The real mean must therefore be 32·5 + 5 (0·947) = 37·2—which is the same as the value we found by the long method using real units throughout.

The multiplier 5 is arrived at thus : the mean is found to be 0·947 above the 0 value when the groups differ in their distances from one another's centres by unity—e.g., from −1 to −2; but in the real distribution their distance from one another's centres is 5—e.g., from 27·5 to 22·5; therefore the mean in real units must be 5 times 0·947 above 32·5 (if the mean in working units had been

+1 clearly the real mean would be 37·5, for the latter is the value for which +1 was the substitute—i.e., 32·5 + 5 (1)).

The *rule* then is this. Having found the mean in working units, multiply its value by the original unit of grouping (4, 5, 10, or whatever it may be) and add the resulting figure (or subtract it according to its sign) to the value of the *centre* of the group against which the 0 was originally placed. That gives the real mean value. 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 2·26 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 × 2·26 = 11·30.

CHECKING THE ARITHMETIC

As regards the final result 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 best to place it centrally so that the multipliers may be kept small. For the sake of demonstration the calculations for Table III B are repeated in Table III C taking another position for 0. This, in practice, is a good method of checking the arithmetic.

TABLE III C

Calculation of Standard Deviation: Grouped Series

Age in years.	Number of deaths in each age-group.	Age in working units.	(2) × (3).	(4) × (3).
(1)	(2)	(3)	(4)	(5)
0–	1	−8	− 8	64
5–	—	−7	—	—
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
70–75	2	+6	+ 12	72
Total ..	206	—	−217	1281

From the calculations in Table III C we have:

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

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

(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).

Mean squared deviation in working units round 0 = 1281/206 = 6·2184

∴ standard deviation in working units is

$$\sqrt{6{\cdot}2184 - (1{\cdot}053)^2} = 2{\cdot}26$$

(1·053 is the distance between the value of 0 from which we measured the deviations and the value from which we ought to have measured them; note that the correction is *subtracted* whatever the sign of the mean in working units).

∴ the real standard deviation is 2·26 × 5 = 11·30. These values agree with those previously found.

The Standard Deviation in Small Samples

Finally it may be noted that 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. 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. A simple method of making this change is to calculate the standard deviation in the way just described and multiply the result by $\sqrt{\frac{n}{n-1}}$ e.g., the standard deviation of the 20 observations of blood pressure in Table III would be $13{\cdot}55 \times \sqrt{\frac{20}{19}} = 13{\cdot}90$. This correction should be applied if the number of observations is less than about 30, especially if tests of "significance" are to be applied.

A. B. H.

THE FIGHT AGAINST LEPROSY

THE fourfold objective of the British Empire Leprosy Relief Association was outlined by Dr. Ernest Muir, its medical secretary, at the annual meeting held at the India Office on April 15th. The Association is concerned with the study of leprosy and of the conditions under which it exists and spreads. It endeavours also to help the leper, by care, treatment, and training; to combating leprosy with a view to its final control; and to interest, rouse, and educate the British public in the problem of leprosy. Dr. Muir said that since the inception of the Association 13 years ago a much more accurate idea had been obtained of the widespread distribution of leprosy and of the various factors which govern its incidence. Study of the disease itself had shown that while most lepers are not infectious, a few highly infectious cases can spread the disease to many others, and thus one generation infects the next. Those infected in childhood furnish most of the serious infectious cases. As to treatment, it was now recognised that though medicines are of value, the main remedy lies in healthy occupation and sound nutrition. Compulsory segregation and treatment were generally worse than useless. The leper must be led, not driven; without his coöperation neither effective treatment nor limitation of the infection could be secured. Segregation by itself would never do more than touch the fringe of leprosy control—at least in poor and densely populated countries; but well-equipped and staffed settlements could be used as centres for an educative campaign, and indeed their chief function should be to act as a centre of training and enlightenment in the district.

The annual report of the Association emphasises the fact that leprosy is a problem of colonial development. At present two types of leper institution are to be found; one is a refuge where patients crippled and deformed and often non-infective are concentrated, while infectious cases, not recognised as such, mix freely with the community. The other progressive type is that to which cheerful and energetic lepers are attracted mainly by the hope of recovery. It is to the organisation and multiplication of this latter type of settlement that the Association is devoting its energies and as well as to the education of all people in the nature of leprosy and the means of its prevention. Sir William Peel has succeeded Sir Edward Gait as chairman of the executive committee of the Association. He made an urgent plea for more support from the British public for the maintenance and extension of its activities.