PRINCIPLES OF MEDICAL STATISTICS

XVII—CALCULATION OF THE CORRELATION COEFFICIENT

THE correlation coefficient, r, is most easily calculated from the formula r = mean of the values of (observation of x minus mean of the observations of x) × (corresponding observation of y minus mean of the observations of y) ÷ standard deviation of x × standard deviation of y. Or in the symbols previously used,

$$r_{xy} = \frac{\text{Sum of values of } (x - \bar{x})(y - \bar{y})}{n\sigma_x \sigma_y}$$

The Ungrouped Series

Suppose, for instance, we have measured on twelve persons their pulse-rate and their stature, and wish to measure the degree of relationship, if any, between the two by means of the correlation coefficient. The twelve observations are given in columns (2) and (3) of Table X A.

TABLE X A

Individual No.	Resting pulse-rate in beats per minute.	Height in inches.	x^2	y^2	x − Mean x.	y − Mean y.	(6) × (7).	$x \times y$ or (2) × (3).
	x	y						
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	62	68	3,844	4,624	−10	−1	+10	4,216
2	74	65	5,476	4,225	+ 2	−4	− 8	4,810
3	80	73	6,400	5,329	+ 8	+4	+32	5,840
4	59	70	3,481	4,900	−13	+1	−13	4,130
5	65	69	4,225	4,761	− 7	0	0	4,485
6	73	66	5,329	4,356	+ 1	−3	− 3	4,818
7	78	69	6,084	4,761	+ 6	0	0	5,382
8	86	70	7,396	4,900	+14	+1	+14	6,020
9	64	72	4,096	5,184	− 8	+3	−24	4,608
10	68	71	4,624	5,041	− 4	+2	− 8	4,828
11	75	68	5,625	4,624	+ 3	−1	− 3	5,100
12	80	67	6,400	4,489	+ 8	−2	−16	5,360
Total 12	864	828	62,980	57,194	0	0	−19	59,597
Mean values	72	69	5248·33	4766·17	—	—	−1·58	4966·42

The standard deviations can be found, as shown before, by squaring each observation, finding the mean of these squares, subtracting from it the square of the mean, and taking the square root of the resulting figure. The standard deviation of the pulse-rates is therefore $\sqrt{5248{\cdot}33 - (72)^2} = 8{\cdot}02$ and of the height is $\sqrt{4766{\cdot}17 - (69)^2} = 2{\cdot}27$. The deviation of each individual's pulse-rate from the mean pulse-rate of the twelve persons is given in column (6) and of each height from the mean height in column (7). If there is any substantial (and direct) correlation between the two measurements, then a person with a pulse-rate below the mean pulse-rate ought to have a stature below the mean height, one with a pulse-rate above the mean rate ought to have a stature above the mean height. (If the association is inverse positive signs in one will be associated with negative signs in the other.) Inspection of the figures suggests very little correlation between the characteristics. For the numerator of the correlation coefficient formula we need the product of the two deviations shown by each person. These are given in column (8). Their sum is −19 and their mean is therefore −19/12 = −1·58. The coefficient is reached by dividing this mean product value by the product of the two standard deviations—namely, 8·02 × 2·27, and gives a value of −0·09. In other words in these twelve individuals the pulse-rate and stature are not related to one another.

In the example taken this is a satisfactory mode of calculation because the mean pulse-rate and the mean height are whole numbers and also the original measurements are whole numbers; it is, then, easy to calculate the deviation of each observation from its mean. But if decimals had been involved the deviations would have been troublesome to calculate. In that case it is easier to avoid altogether using deviations and to multiply directly the pulse-rate of each person by his stature, as in column (9), applying a correction at the end to the resulting mean value. The correction necessary is this: from the mean value of the products thus found we must subtract the product of the two means—i.e., the product of the distances between the points from which we chose to measure the deviations and the points from which we ought to have measured them. In the example taken this gives 4966·42 − 72 × 69 = − 1·58, or the same value as was previously reached by working with the real deviations from the means.

This is the simplest and best method of calculating the two standard deviations and the correlation coefficient between the characteristics in anything up to 50–60 observations. With a larger series of observations, finding the individual squares and products becomes progressively more laborious and it is better to construct a grouped correlation table.

The Grouped Series

As an example, we may take for each of a number of large towns in England and Wales (1) a measure of the amount of overcrowding present in a given year, and (2) the infant mortality-rate in the same year; we wish to see whether in towns with much overcrowding the infant mortality-rate tends to be higher than in towns with less overcrowding. We must first construct a table which shows not only how many towns there were with different degrees of overcrowding but also their associated infant mortality-rates.

Table X B gives this information. The town with least overcrowding had only 1·5 per cent. of its population living more than 2 persons to a room (this being used as the criterion of overcrowding); the percentage for the town with most overcrowding was 17·5. The lowest infant mortality-rate was 37 deaths under 1 per 1000 live births and the highest was 110. Reasonably narrow groups have been adopted to include those maxima and minima and each town is placed in the appropriate "cell"—e.g., there were 5 towns in which the overcrowding index lay between 1·5 and 4·5 and in which the infant mortality-rates were between 36 and 46, there were 2 in which the overcrowding index lay between 10·5 and 13·5 and in which the infant mortality-rates were between 86 and 96. (If a very large number of observations is involved it is best to make a separate card for each town, person, or whatever may have been measured, putting the observed measurements on the card always in the same order; the cards are first sorted into their proper groups for one characteristic (overcrowding), and then each of those packs of different (overcrowding) levels is sorted into groups for the other characteristic (infant mortality). The cards in each small pack then relate to a particular cell of the table.)

Table X B shows at once that there is some association between overcrowding and the infant mortality-rate, for towns with the least overcrowding tend, on the average, to show relatively low mortality-rates, while towns with much overcrowding tend to show high mortality-rates. The table is, in fact, a form of scatter diagram.

TABLE X B

Overcrowding and Infant Mortality. Example of Correlation Table

Infant mortality-rate.	Percentage of population in private families living more than two persons per room.						Total.
	1·5–	4·5–	7·5–	10·5–	13·5–	16·5–19·5	
36– ..	5	..	..	..	..	..	5
46– ..	9	1	..	..	..	..	10
56– ..	10	4	1	..	..	1	16
66– ..	4	7	5	2	..	..	18
76– ..	2	5	4	1	1	..	13
86– ..	..	2	2	2	..	1	7
96– ..	..	1	2	2	1	1	7
106–116 ..	..	1	..	1	..	..	2
Total ..	30	21	14	8	2	3	78

To calculate the coefficient of correlation we need (1) the mean and standard deviation of the overcrowding index; (2) similar figures for the infant mortality-rate; and (3) for each town the product of its two deviations from the means—i.e. (overcrowding index in town A minus mean overcrowding index) × (infant mortality-rate in town A minus mean infant mortality-rate). In other words we wish to see whether a town that is abnormal (far removed from the average) in its level of overcrowding is also abnormal in the level of its infant mortality-rate. In calculating the means and standard deviations of the two distributions we can entirely ignore the centre of the table; we have to work on the totals in the horizontal and vertical margins. The method is shown in Table X C.

TABLE X C

Overcrowding and Infant Mortality. Calculation of Correlation Coefficient

Infant mortality-rate.		Percentage of population in private families living more than two persons per room.						Total.
		R.U.: 1·5–	4·5–	7·5–	10·5–	13·5–	16·5–19·5	
R.U.	W.U.	W.U. −2	−1	0	+1	+2	+3	
36–	−3	5 (+30)	..	..	..	..	..	5
46–	−2	9 (+36)	1 (+2)	..	..	..	..	10
56–	−1	10 (+20)	4 (+4)	1 (0)	..	..	1 (−3)	16
66–	0	4 (0)	7 (0)	5 (0)	2 (0)	..	..	18
76–	+1	2 (−4)	5 (−5)	4 (0)	1 (+1)	1 (+2)	..	13
86–	+2	..	2 (−4)	2 (0)	2 (+4)	..	1 (+6)	7
96–	+3	..	1 (−3)	2 (0)	2 (+6)	1 (+6)	1 (+9)	7
106–116	+4	..	1 (−4)	..	1 (+4)	..	..	2
Total	..	30	21	14	8	2	3	78

R.U. = Real units; W.U. = Working units.

For instance, we see from the right-hand totals that in 5 towns the infant mortality-rate was between 36 and 46, in 10 between 46 and 56, in 16 between 56 and 66, and so on. Of this distribution we want the mean and standard deviation. As shown previously these sums are more easily carried out in "working units" instead of in the real, and larger, units. In these units we have 5 towns with an infant mortality-rate of −3, 10 with a rate of −2, and so on. In working units, therefore, the sum of the rates is +5 and the mean rate is +5/78 = +0·06. The mean in real units is, then, 71 (the centre of the group opposite 0) + 0·06 × 10 (10 being the unit of grouping) = 71·6. To reach the standard deviation we continue to work in these units. Measuring the deviations from the 0 value instead of from the mean there are 5 towns with a squared deviation of $(-3)^2$ and these contribute 45 to the sum of squared deviations; there are 10 towns with a squared deviation of $(-2)^2$ and these contribute 40 to the sum of squared deviations.

In working units the sum of squared deviations from 0 is thus found to be 237. The mean squared deviation from 0 is 237/78 = 3·0385 and from this, as correction for having measured the deviations from 0, we must subtract the square of 0·06, the value from which we ought to have measured the deviations. The standard deviation of the rates in working units is therefore the square root of $3{\cdot}0385 - (0{\cdot}06)^2 = 1{\cdot}742$, and in real units is 1·742 × 10 = 17·42.

We can now work in just the same way on the distribution of overcrowding—there are 30 towns whose overcrowding index in working units was −2, 21 whose index was −1, and so on. This gives a mean and standard deviation in working units of −0·77 and 1·329, and in real units of 6·7 and 3·99. In this there is nothing new; the process was given in full in Section XVI.

We now need the product of the deviations from the means for the numerator of the correlation coefficient. This is easily reached by continuing to measure the deviations in working units from the 0 values and making a correction as usual at the end.

For instance, there are 5 towns the deviation of which is −2 in overcrowding and −3 in infant mortality. The product deviation is therefore +6, and as there are 5 such towns the contribution to the product deviation sum is +30. Each of these values can be written in the appropriate cell (they are the figures in parentheses in Table X C). Their sum is +107 and the mean product value is +107/78 = +1·3718. These deviations in working units were measured from the two 0 values whereas they ought to have been measured from the two mean values, +0·06 and −0·77; therefore as correction we must subtract from +1·3718 the product of +0·06 and −0·77 (as in the ungrouped series, the correction is the product of the distances between the points from which we chose to measure the deviations and the points from which we ought to have measured them). The numerator to the coefficient is therefore +1·3718 − (0·06 × −0·77), and the denominator is the product of the two standard deviations; so that

$$r = \frac{+1{\cdot}3718 + 0{\cdot}0462}{1{\cdot}742 \times 1{\cdot}329} = +0{\cdot}61.$$

(It may be noted that as the numerator is in working units, *the standard deviations must be inserted in their working units.*)

There is we see a fair degree of correlation between overcrowding and the infant mortality-rate, but at the same time Table X B shows that with towns of the same degree of overcrowding there are considerable differences between the infant mortality-rates. The standard error of the coefficient is $1/\sqrt{n-1} = 1/\sqrt{78-1} = 0{\cdot}11$; as the coefficient is more than five times its standard error it may certainly be accepted as "significant."

The calculation is very much speedier with the observations thus grouped and little change has been

made in the values reached, as the following figures show :—

—	Grouped series of Table X B.	Same 78 observations ungrouped.
Means—		
Overcrowding	6·7	6·6
Infant mortality ..	71·6	71·0
Standard deviations—		
Overcrowding	3·99	3·74
Infant mortality ..	17·4	17·3
Correlation coefficient ..	+0·61	+0·59

The regression equation is—

Infant mortality minus 71·6 = $+0{\cdot}61 \frac{17{\cdot}4}{3{\cdot}99}$ (overcrowding index minus 6·7) which reduces to—

Infant mortality = 2·66 overcrowding index + 53·78. (It must be noted that the values in *real* units must be inserted in this equation.) In other words the infant mortality rises, according to these data, by 2·66 per 1000 as the percentage of the population overcrowded increases by 1.

A. BRADFORD HILL.

SPECIAL ARTICLES

BRITISH HEALTH RESORTS ASSOCIATION

A MEETING of the British Health Resorts Association was held at Skegness on Saturday last. The conference was well attended and the various advantages that might be derived by the sick, the convalescent, and the public from the facilities offered at the different centres in this country were brought out. The congressists were the guests of the municipality.

Lord Meston, the president of the association, spoke of the inception of the movement and of its development from a winter-in-England movement into a body not only interested in presenting the claims of British resorts on to the attention of the medical profession and the public, but also in promoting the study of the climatic and other conditions which made these resorts suitable ; or perhaps unsuitable, as health seekers at different times of the year must have resorts selected for them. He pleaded for greater support by local authorities, whose interests the association was unselfishly serving, but noted the steady improvement in hotels, to which he thought the action of the association had contributed. Speaking from personal experience in many countries, he held that the British hotels " had nothing to fear from comparison if like were compared with like."

The chairman of the Skegness urban district council, Mr. J. Crawshaw, presided at the first session, dealing with

Industry and the Health Resort

Mr. A. L. Peterson, managing director of the Spirella Company of Letchworth, spoke as an employer in a firm which had made complete arrangements for the welfare of its employees. As most of the users of health resorts, he said, were connected with industry it was important for employers who desired to be progressive should be told more of the advantages of these resorts for the workers. He stressed the point that, with payment for holidays, a movement which was advancing, there would be great opportunities for health resorts placed in proximity to the centres of industry. He said that there should be talks in the factories on how holidays could be used so that holiday-makers could get the best value out of them.

Mr. Ernest Bevin, general secretary of the Transport and General Workers Union, said that the provision of holidays with full pay for the workers was one of the principal struggles of the movement he represented. The different treatment of staff, public servants, and others employed in more favourable occupations, as against the actual workers, formed a serious grievance. Since the war the more enlightened employers were realising that the granting of holidays with pay was not only an advantage to the working person and his family but to industry itself. Unpaid unemployment, moreover, had risen periodically so that the whole question of what might be termed the contractual period for labour had been brought within the realm of practical politics and the Government had now set up a committee, of which he was himself a member, to consider the whole problem.

THE AMENITIES OF HEALTH RESORTS

There was need for the creation of an industry to cater for the holiday needs of the workers ; there should be a scientific study of the whole matter and the provision of holidays for the workers would open an avenue of employment for thousands of others. He held that the British health resorts had not taken a sufficiently enlightened view to meet the requirements that this new development would entail. It had been demonstrated that a great set-off against the cost of holidays to industry generally was the decrease in sick leave and absenteeism. The English climate was not an easy one for which to cater, but the kind of shelter put up along sea fronts, with a glass partition in the centre, was quite inadequate to meet the weather changes and yet these shelters were an absolute necessity if visitors were to derive the benefit of the sea air. And when holiday resorts were planned there must be a real drive to secure a better standard of accommodation. Although " hot and cold running water " in every room was so readily advertised, in thousands of houses, which the workers now have to use, the accommodation was quite out of date ; yet from a health point of view, good accommodation and bright surroundings were an even bigger contribution to recuperation than medicine. Taking the British seaside resorts generally, the municipalities had spent more money and were superior to continental resorts in the arrangements made to cater for the pleasures of the people. But in accommodation and cuisine they lagged behind. Here was a great opportunity for municipal enterprise which would be called for, since the extension of holidays over a lengthened season would cause the question of holiday centres or homes to be dealt with. These could not be erected in every health resort, and if the millions were to be catered for some development must occur. In many industries it was impracticable for the whole of the workers to take their holidays during the summer months and there were three other periods in the year which offered opportunity for catering, if correctly handled—i.e., late October, Christmas, and Easter. The maintenance of our health resorts would be largely dependent upon catering for the masses. Amongst the so-called