# PRINCIPLES OF MEDICAL STATISTICS

## X—THE COEFFICIENT OF CORRELATION

A PROBLEM with which the statistician is frequently faced is the measurement of the *degree* of relationship between two, or more, characteristics of a population. For instance in a particular area the air temperature is recorded at certain times and the mean air temperature of each week is computed ; the number of deaths registered as due to bronchitis and pneumonia is put alongside it. Suppose the following results are reached :—

| Mean temperature of week in °F. | Number of weeks with given mean temperature. | Mean number of deaths registered as due to bronchitis and pneumonia in these weeks. | Range in weekly deaths. |
|---|---|---|---|
| 35– | 5 | 253 | 186–284 |
| 38– | 7 | 205 | 147–238 |
| 41– | 10 | 130 | 94–180 |
| 44–47 | 4 | 87 | 69–112 |

There is clearly some relationship between these two measurements. As the mean weekly temperature rises there is a decrease in the average weekly number of deaths from bronchitis and pneumonia, which fall from an average of 253 in the 5 coldest weeks to an average of 87 in the 4 warmest weeks. On the other hand there is at the same mean temperature a considerable variability in the number of deaths registered in each week, as shown in the ranges given in the right-hand column. In individual weeks there were sometimes, for instance, more deaths registered when the temperature was 38°–41° than when it was 35°–38°. In measuring the closeness of the relationship between temperature and registered deaths this variability must be taken into account.

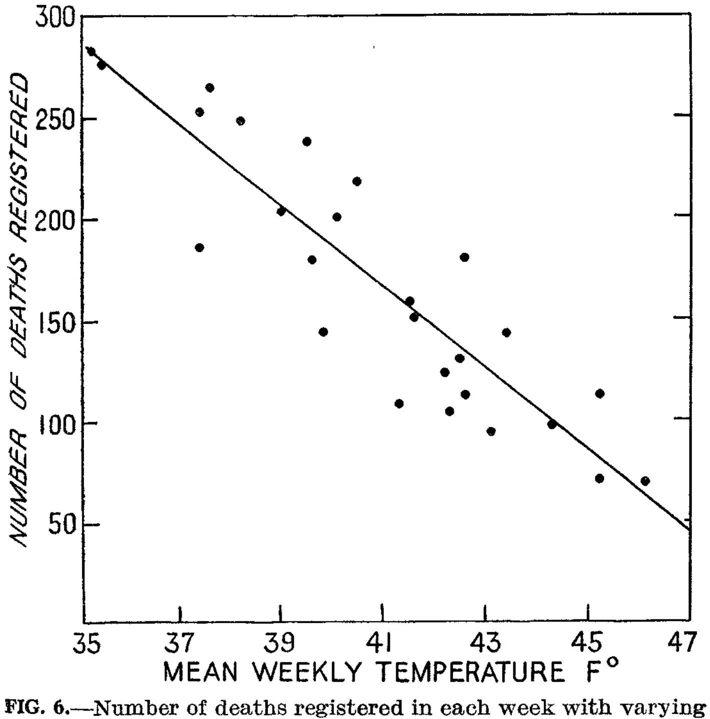

FIG. 6.—Number of deaths registered in each week with varying mean temperatures. A scatter diagram.

The declining number of deaths as the mean temperature rises, and also the variability of this number in weeks of about the same temperature, are shown clearly in Fig. 6—known as a *scatter diagram.* It appears from the distribution of the points (each of which represents the mean temperature of and the deaths registered in one week) that the relationship between the temperature and the deaths could be reasonably described by a straight line, such as the line drawn through them on the diagram. The points are widely scattered round the line in this instance but their downward trend follows the line and shows no tendency to be curved. The means in the table above very nearly fall on the line. In such instances a satisfactory measure of the degree of relationship between the two characteristics is the coefficient of correlation, the advantage of which is that it gives in a single figure an assessment of the degree of the relationship which is more vaguely shown in the table and diagram (both of which are, however, perfectly valid and valuable ways of showing associations).

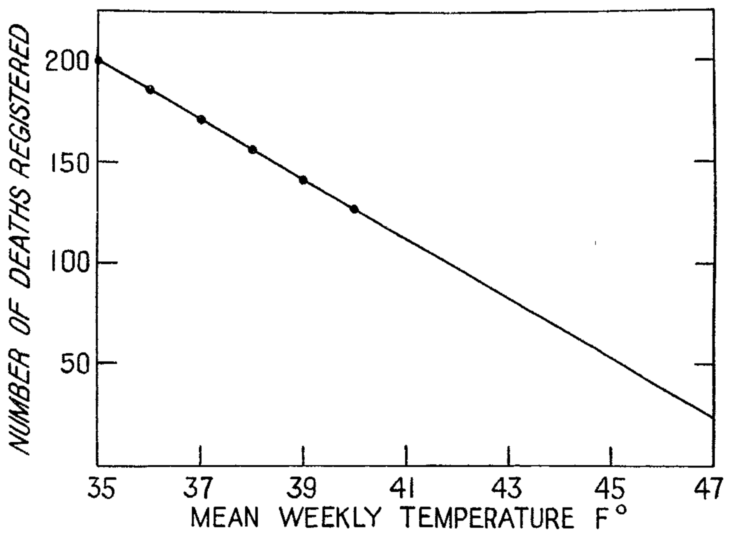

FIG. 7.—Number of deaths registered in each week with varying mean temperatures. Hypothetical case of complete correlation.

### Dependent and Independent Characteristics

There are various ways of considering this coefficient. The following is perhaps the simplest.

(i) Let us suppose first that deaths from pneumonia and bronchitis are dependent upon the temperature of the week and upon no other factor and also follow a straight line relationship—i.e., for each temperature there can be only one total for the deaths and this total falls by the same amount as the temperature increases each further degree. Then our scatter diagram reduces to a series of points lying exactly upon a straight line. For instance, in Fig. 7 the deaths total 200 at 35°, 185 at 36°, 170 at 37°. For each weekly temperature there is only a single value for the deaths, the number of which falls by 15 as the temperature rises one degree. If we know the temperature we can state precisely the number of deaths. No error can be made for there is no scatter round the line.

(ii) Now let us suppose that the deaths are completely independent of the temperature, but fluctuate from week to week for quite other reasons. When the temperature is low there is then no reason why we should observe a larger number of deaths than when the temperature is high, or vice versa. If we had a very large number of weekly records we should observe at each temperature all kinds of totals of deaths. The scatter diagram would take the form shown in Fig. 8 (only roughly, of course, in practice). At 35° there were, for example, weeks in which were recorded 50, 75, 100, 125, 150, 175, and 200 deaths ; at 36° we see the same totals, and similarly at each higher temperature. The average of the weekly totals of deaths observed is the same at each temperature and is equal to the average of all the weekly observations put together, for the two characteristics being independent there is no tendency for these averages to move up or down as the temperature changes. If we know the temperature of the week we obviously cannot state the number of deaths in that week with any accuracy. We can, however, attempt to do so and we can measure the amount of our error. The best estimate of the deaths we can make at any weekly temperature is the average number of deaths taking place in all the weeks put together—for as pointed out we have no reason to suppose that the deaths recorded in a specified week will be more or less numerous than the average, merely because the temperature was high or low. Our error in a specified week is, therefore, the difference between the mean number of deaths in all weeks and the number of deaths that actually occurred in that particular week; e.g., the mean number of deaths in all the weeks recorded in Fig. 8 was 125; in one week with a mean temperature of 36° the number recorded was 50, and our error of estimation for this week is $125-50=75$. We can compute this error for each week in turn and find the average size of our error, or, preferably, we can find the average of the squared errors. This latter value will, in fact, be the square of the standard deviation of the weekly numbers of deaths (which, as shown previously, is the mean squared deviation of the observations from their average). Our weekly errors between estimation and observation can therefore be measured by $\sigma^2$.

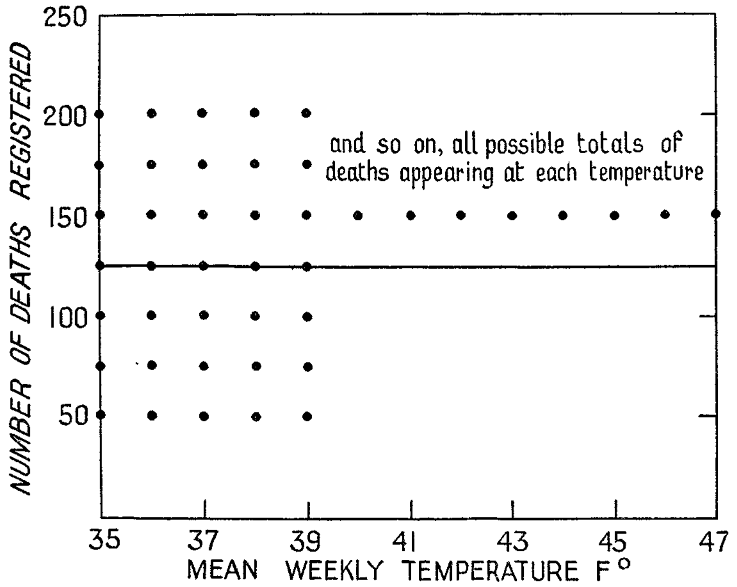

FIG. 8.—Number of deaths registered in each week with varying mean temperatures. Hypothetical case of complete absence of correlation.

(iii) Now let us suppose that neither of the extreme cases is present—i.e., neither complete dependence nor complete independence—but that we have something between the two, as in Fig. 6, where the deaths certainly decline as the temperature rises but show some variability at each temperature. How precisely can we now state the number of deaths when we know the temperature? Let us draw through the points a line which represents, broadly, their trend. From that line we can read off the expected number of deaths at each temperature and compare it with the observed number in that week. The difference will be the error we make in using this line. We can calculate this error for each week and the average of these squared errors we can call $S^2$.

Are we any better off in our estimations of the actual weekly deaths by the use of this line than when we say that in each week we expect to see the average number of deaths that took place in all the weeks? We can measure our relative success by comparing $\sigma^2$ with $S^2$. If the two characteristics are entirely *independent* of one another (as in Fig. 8) the line we draw can have no slope at all and will pass at each temperature through the average number of deaths in all the weeks (there is nothing to make it higher or lower than the average at different temperatures). $S^2$ and $\sigma^2$ will then be the same. If the two characteristics are completely *dependent*, as in Fig. 7, then there is no scatter round the line at all and $S^2$ becomes 0. In practice we use as a measure of the degree of association $\sqrt{1-\frac{S^2}{\sigma^2}}$ which is known as $r$ or the correlation coefficient. If no association at all exists $S^2$ and $\sigma^2$ are, as pointed out, equal and $r$ equals 0. If there is complete dependence $S^2$ is 0 and $r$ equals 1. For any other degree of association $r$ must lie between 0 and 1, being low as its value approaches 0 and high as it approaches 1.

## The Use and Meaning of the Correlation Coefficient

The actual mode of calculation of the correlation coefficient is fully described in numerous statistical text-books, and attention will be confined here to its use and meaning. It is calculated in such a way that its value may be either positive or negative, between $+1$ and $-1$. Either plus or minus 1 indicates complete dependence of one characteristic upon the other, the sign showing whether the association is direct or inverse; a positive value shows that the two characteristics rise and fall together—e.g., age and height of school-children, a negative value that one falls as the other rises, as in our example of deaths and temperature. In the latter instance the value of the coefficient is, in fact, $-0{\cdot}90$. This figure shows that there is in this short series of observations a very high degree of relationship between the temperature and the deaths, but, as it and the graph make obvious, not a complete relationship. Other factors are influencing the number of deaths as well as the temperature. If we knew the equation to the line we could certainly predict the number of deaths that would take place in a particular week with a given mean temperature with more accuracy than would be possible without that information; but the diagram shows that in individual weeks we might still be a long way out in our prediction. It is the original variability in the number of deaths at *each* temperature than makes it impossible for the prediction to be accurate for an individual week, though we might be able to predict very closely the *average* number of deaths in a group of weeks of the same temperature. The advantage of the coefficient is, as previously pointed out, that it gives in a single figure a measure of the amount of relationship. For instance, we might calculate two such coefficients between, say, mean weekly temperature and number of deaths from bronchitis and pneumonia at ages 0–5, and between mean weekly temperature and number of these deaths at ages 65 and over, and thus determine in which of the two age-groups are deaths from these causes more closely associated with temperature level. We can also pass beyond the coefficient of correlation and find the equation to the straight line that we have drawn through the points. Reading from the diagram the straight line shows that at a weekly temperature of 39° F. the estimated number of deaths is about 205; at a weekly temperature of 40° F. the deaths become about 185; at a temperature of 41° F. they become 165. For each rise of 1° F. in the mean weekly temperature the deaths will fall, according to this line, by some 20 deaths. In practice the method of calculating the coefficient of correlation ensures that this line is drawn through the points in such a way as to make the sum of the squares of the differences between the actual observations at given temperatures and the corresponding values predicted from the line for those temperatures, have the smallest possible value. No other line drawn through the points could make the sum of the squared errors of the estimates have a smaller value, so that on this criterion our estimates are the best possible.

### THE REGRESSION EQUATION

The equation to the line can be found from the following formula:—

$$\text{Deaths} - \text{Mean number of deaths} = \begin{array}{l}\text{Correlation}\\ \text{coefficient}\end{array} \times \frac{\text{Standard deviation of deaths}}{\text{Standard deviation of temperature}}$$

$$\times\ (\text{Temperature} - \text{Mean of the weekly temperatures})$$

where the two means and standard deviations are those of all the weekly values taken together.

Writing in the values we know from the data under study this becomes—

$$\text{Deaths} - 167{\cdot}038 = -0{\cdot}9022\ \frac{64{\cdot}310}{2{\cdot}867}\ (\text{Temperature} - 40{\cdot}908)$$

The fraction on the right hand side of the equation equals 20·24 so that we have—

$$\text{Deaths} = -20{\cdot}24\ (\text{Temperature} - 40{\cdot}908) + 167{\cdot}038$$

Removing the parentheses and multiplying the terms within by −20·24 this becomes—

$$\text{Deaths} = -20{\cdot}24\ \text{Temperature} + 827{\cdot}978 + 167{\cdot}038$$

or, finally,

$$\text{Deaths} = -20{\cdot}24\ \text{Temperature} + 995{\cdot}02.$$

The figure −20·24 shows, as we saw previously from the diagram, that for each rise of 1° F. in the temperature the deaths decline by about 20. (For when the temperature is, say, 40° F., the deaths estimated from the line are 995·02−(20·24) 40 = 185·4 and when the temperature goes up 1° F. to 41° F., the estimated deaths are 995·02−(20·24) 41 = 165·2.)

The figure −20·24 is known as the *regression coefficient*; as seen above it shows the change that, according to the line, takes place in one characteristic for a unit change in the other. The equation is known as the *regression equation*. As far as we have gone our conclusions from the example taken is that deaths from bronchitis and pneumonia are in a certain area closely associated with the weekly air temperature and that a rise of 1° F. in the latter leads, on the average, to a fall of 20 in the former.

## Precautions in Use and Interpretation

In using and interpreting the correlation coefficient certain points must be observed.

### THE RELATIONSHIP MUST BE REPRESENTABLE BY A STRAIGHT LINE

(1) In calculating this coefficient we are, as has been shown, presuming that the relationship between the two factors with which we are dealing is one which a straight line adequately describes. If that is not approximately true then this measure of association is not an efficient one. For instance, we may suppose the absence of a vitamin affects some measurable characteristic of the body. As administration of the vitamin increases a favourable effect on this body measurement is observed, but this favourable effect may continue only up to some optimum point. Further administration leads, let us suppose, to an unfavourable effect. We should then have a distinct *curve* of relationship between vitamin administration and the measurable characteristic of the body, the latter first rising and then falling. The graph of the points would be shaped roughly like an inverted U and no straight line could possibly describe it. Efficient methods of measuring that type of relationship have been devised—e.g., the correlation ratio—and the correlation coefficient should not be used. Plotting the observations, as in Fig. 6 relating to temperature and deaths from bronchitis and pneumonia, is a rough but reasonably satisfactory way of determining whether a straight line will adequately describe the observations. If the number of observations is large it would be a very heavy test to plot the individual records, but one may then plot the *means of columns* in place of the individual observations—e.g., the mean height of children aged 6–7 years was so many inches, of children aged 7–8 years so many inches—and see whether those means lie approximately on a straight line.

### THE LINE MUST NOT BE UNDULY EXTENDED

(2) If the straight line is drawn and the regression equation found, it is dangerous to extend that line beyond the range of the actual observations upon which it is based. For example, in school-children height increases with age in such a way that a straight line describes the relationship reasonably well. But to use that line to predict the height of adults would be ridiculous. If, for instance, at school ages height increases each year by an inch and a half, that increase must cease as adult age is reached. The regression equation gives a measure of the relationship between certain observations; to presume that the same relationship holds beyond the range of those observations would need justification on other grounds.

### ASSOCIATION IS NOT NECESSARILY CAUSATION

(3) The correlation coefficient is a measure of *association* and in interpreting its meaning one must not confuse association with causation. Proof that A and B are associated is not proof that a change in A is directly responsible for a change in B or vice versa. There may be some common factor C which is responsible for their associated movements. For instance, in a series of towns it might be shown that the phthisis death-rate and overcrowding were correlated with one another, the former being high where the latter was high and vice versa. This is not necessarily evidence that phthisis is due to overcrowding. Possibly, and probably, towns with a high degree of overcrowding are also those with a low standard of living and nutrition. This third factor may be the one which is responsible for the level of the phthisis rate, and overcrowding is only indirectly associated with it. It follows that the meaning of correlation coefficients must always be considered with care, whether the relationship is a simple direct one or due to the interplay of other common factors. In statistics we are invariably trying to disentangle a chain of causation and several factors are likely to be involved. Time correlations are particularly difficult to interpret but are particularly frequent in use as evidence of causal relationships—e.g., the recorded increase in the death-rate from cancer is attributed to the increase in the

consumption of tinned foods. Clearly such concomitant movement might result from quite unrelated causes and the two characteristics actually have no relationship whatever with one another except in time. Merely to presume that the relationship is one of cause and effect is fatally easy; to secure satisfactory proof or disproof, if it be possible at all, is often a task of very great complexity.

### THE STANDARD ERROR

(4) As with all statistical values the correlation coefficient must be regarded from the point of view of sampling errors. In taking a sample of individuals from a universe it was shown that the mean and other statistical characteristics would vary from one sample to another. Similarly if we have *two* measures for each individual the correlation between those measures will differ from one sample to another. For instance, if the correlation between the age and weight in all school-children were 0·85, we should not always observe that value in samples of a few hundred children; the observed values will fluctuate around it. We need a measure of that fluctuation, or, in other words, the standard error of the correlation coefficient. Similarly if two characteristics are not correlated at all so that the coefficient would, if we could measure all the individuals in the universe, be 0, we shall not necessarily reach a coefficient of exactly 0 in relatively small samples of those individuals. The coefficient observed in such a sample may have some positive or negative value. In practice we have to answer this question: could the value of the coefficient we have reached have arisen quite easily by chance in taking a sample, of the size observed, from a universe in which there is no correlation at all between the two characteristics? For example in a sample of 145 individuals we find the correlation between two characteristics to be +0·32. Is it likely that these two characteristics are not really correlated at all, that if we had taken a very much larger sample of observations the coefficient would be 0 or approximately 0? It can be proved that if the value of the correlation coefficient in the universe is 0, then (*a*) the *mean* value of the coefficients that will be actually observed if we take a series of samples from that universe will be 0, but (*b*) the separate coefficients will be scattered round that mean, with a standard deviation, or standard error, of $1/\sqrt{N-1}$, where $N$ is the number of individuals in each sample. In the example above the standard error will, therefore, be $1/\sqrt{144} = 0{\cdot}083$. Values which deviate from the expected mean value of 0 by more than twice the standard deviation are, we have previously seen, relatively rare. Hence if we observe a coefficient that is more than twice its standard error we conclude that it is unlikely that we are sampling a universe in which the two characters are really not correlated at all. In the present case the coefficient of 0·32 is nearly four times its standard error; with a sample of 145 individuals we should only very rarely observe a coefficient of this magnitude if the two characters are not correlated at all in the universe. We may conclude that there is a "significant" correlation between them—i.e., more than is likely to have arisen by chance due to sampling errors. If on the other hand the size of the sample had been only 26 the standard error of the coefficient would have been $1/\sqrt{25} = 0{\cdot}2$. As the coefficient is only 1·6 times its standard error we should conclude that a coefficient of this magnitude might have arisen merely by chance in taking a sample of this size, and that in fact the two characters may not be correlated at all. We should need more evidence before drawing any but very tentative conclusions. This test of "significance" should be applied to the correlation coefficient before any attempt is made to interpret it. Somewhat more intricate methods are needed to test whether one coefficient differs "significantly" from another—e.g., whether deaths from bronchitis and pneumonia are more closely correlated with air temperature at ages 0–5 years than at ages 65 and over (see, for instance, The Methods of Statistics. By L. H. C. Tippett. London: Williams and Norgate Ltd. 1931. 15*s*.).

## Summary

The correlation coefficient is a useful measure of the degree of association between two characteristics, but only when their relationship is adequately described by a straight line. The equation to this line, the regression equation, allows the value of one characteristic to be estimated when the value of the other characteristic is known. The error of this estimation may be very large even when the correlation is very high. Evidence of association is not necessarily evidence of causation, and the possible influence of other common factors must be remembered in interpreting correlation coefficients. It is possible to bring a series of characteristics into the equation, so that, for instance, we may estimate the weight of a child from a knowledge of his age, height, and chest measurement, but the methods are beyond the limited scope of these articles.

A. B. H.

CORRIGENDA.—In last week's article the formula in line 17 of the second column of p. 527 should read $n=(c-1)(r-1)$ and the numerator of the fourfold table on p. 528 should read $(ad-bc)^2\,(a+b+c+d)$.

KING'S COLLEGE HOSPITAL: NEW WING.—On Feb. 23rd Viscount Wakefield, vice-president of the hospital, opened a wing for private patients which has been presented to the hospital by the Stock Exchange Dramatic and Operatic Society and by other friends on the Stock Exchange. The building is 266 feet long and 80 feet wide and runs parallel with the main corridors of the hospital. The offices are on the side nearest to the hospital so that accommodation for the patients is separated from sounds connected with its work, but there is direct access to the X ray and other special departments in the main building. Above the entrance rises a tower 70 feet high in which are situated the suite of rooms for the resident medical officer and a muniment room for the safe custody of records; it is a memorial to the flight of Mr. Giles Guthrie from England to Johannesburg erected by his father, Sir Connop Guthrie. On the first floor there are eighteen single rooms for which the charge will be 8 guineas a week. The patient makes his own arrangement for the payment of fees for professional attendance, including the services of the pathologist and radiologist. In close proximity to the floor is an operating theatre specially provided for visiting surgeons. On the floor above is a complete maternity unit where accommodation is provided at varying charges ranging from 7 to 10 guineas in rooms with one, two, and four beds. The services of the resident medical officer will be available for antenatal attention as well as for the confinement, and a comprehensive fee can be arranged to cover both. On the floor above is a maternity isolation unit. In order to provide the additional staff for the new wing it has been necessary to extend the accommodation for resident medical officers, nurses, and maids. On these extensions there is a debt outstanding of £15,000. The sum of £8000 would enable the ground floor to be fitted up as a dental department readily accessible from the out-patient department, which would release the general ward of the hospital where it is at present housed for occupation by ordinary patients.