On the Theory of Correlation
Author(s): G. Udny Yule
Source: *Journal of the Royal Statistical Society*, Vol. 60, No. 4 (Dec., 1897), pp. 812-854
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2979746
Accessed: 11/10/2014 05:21

On the Theory *of* Correlation. *By* G. Udny Yule.

CONTENTS:

I.—*On Correlation in Genəral. Introductory.*

The investigation of causal relations between economic phenomena presents many problems of peculiar difficulty, and offers many opportunities for fallacious conclusions. Since the statistician can seldom or never make experiments for himself, he has to accept the data of daily experience, and discuss as best he can the relations of a whole group of changes; he cannot, like the physicist, narrow down the issue to the effect of one variation at a time. The problems of statistics are in this sense far more complex than the problems of physics.

In view of this complexity it will, I think, be generally allowed that the statistical methods hitherto employed are frequently inadequate. No apology therefore is necessary for bringing before the Society a method that, whatever its difficulties, exceeds in completeness and generality any other that covers the same ground.

Few of the results given in the sequel are entirely new, but they have not hitherto been collected together or fully illustrated by numerical examples. The majority have at best found notice in this *Journal* in notices or abstracts of papers by Professor Edgeworth, Mr. Francis Galton, or Professor Pearson.

Before proceeding to the development of formulæ, it may be well to define a few terms that are possibly unfamiliar. The quantities—necessarily *numerical quantities*—whose relations it is desired to investigate will be spoken of as the *variables*, since their magnitude varies. Instead of speaking of "causal relation," "causally related quantities," we will use the terms "correlation," "correlated quantities." This definition is provisional, subject to numerical measures given later.

The ordinary table of double entry which gives the frequency of occurrence of different pairs of two associated or correlated variables will be called a *frequency table* or *correlation table*. The marriage tables that appear in every Annual Report of the Registrar-General, showing the number of marriages between men and women of different ages, are perhaps the most familiar instances of correlation tables. The different frequencies given in such a table may be represented by verticals of varying height; if the tops of these verticals be joined up, we obtain a *frequency surface* or *correlation surface*, representing the whole table to the eye. The model of Perozzo's Italian marriage surface in the rooms of the Society will be familiar to many members.

Either a column or a row of a correlation table is conveniently called an *array*, the middle value of the variable with which the row is associated being called its *type*, to use two terms introduced by Professor Pearson. Thus, in the Registrar-General's marriage tables, the age-groups run, 25—30, 30—35, 35—40, and so on; the types of rows or columns are consequently, 27·5, 32·5, 37·5, &c. In the general case we may speak of an x-array of type y, or a y-array of type x, the x-arrays being generally understood to be rows, and the y-arrays columns. Such tables of double entry form the basis of our discussion of correlation.

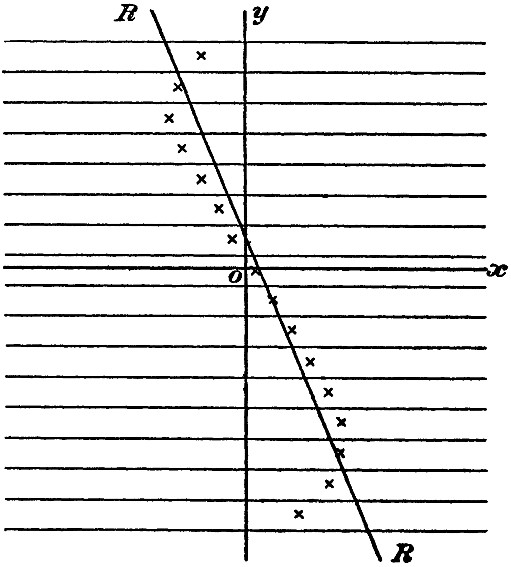

Fig. 1.

Let the diagram of fig. 1[1] represent a general correlation table,

[1] The whole of the first part of the paper up to p. 838 is an expansion, with the addition of worked-out arithmetical examples (pp. 822—831), of a paper by me "On the Significance of Bravais' formulæ for Regression, &c., in the case of Skew Correlation.—"Proceedings Royal Society, 1897," vol. lx, p. 477.

and let the points marked × be the means of successive x-arrays, *e.g.*, in the case we have taken, the mean ages of wives who married husbands of a given age. It is a fact attested by statistical experience that these means do not lie chaotically all over the table, but range themselves more or less closely round a smooth curve, which we will name the *curve of regression*[2] *of x on y.* There is of course a second curve of regression, of y on x, given by the means of the y-arrays. In many cases this curve does not diverge very seriously from a straight line; in a few cases it may be said to be a straight line within the limits of probable error; we will then speak of the *line of regression.* Now suppose that we take a straight line, RR, and fit it to the curve, subjecting the distances of the means from the line to some minimal condition. If the slope of RR be positive, we may say that large values of x are on the whole associated with large values of y. If it is negative, large values of x are associated with small values of y, and *vice versâ.* Further, the slope of RR to the vertical is a measure of a rough practical kind of the shift of the mean of an x-array, corresponding to a given shift of its type y. The equation to the line RR consequently gives a concise and definite answer to two most important statistical questions: Can we say that large values of x are on the whole associated with either large values of y or small values of y? And, What is the average shift of the mean of an x-array corresponding to a shift of unity in its type? If RR be vertical, we may call the two variables *uncorrelated;* so long as it slope to the vertical we may call them correlated. Also, it is evident that if the means of arrays do lie on a straight line, as in the case of normal correlation, discussed in Part II of this paper, RR will be that straight line; *i.e.*, the equation to RR will be the equation to the line of regression. The lines RR we may in general call the characteristic lines.

Let n be the number of correlated pairs, or total frequency, in any x-array, and let d be the horizontal distance of the mean of this array from the line RR. We shall subject the line to the condition that the sum of all quantities like nd^2 shall be a minimum; *i.e.*, we shall use the condition of least squares. This is done solely for convenience of analysis. Using S to denote "the sum of all quantities like" we might make $S(nd)$ a minimum, counting d always as positive or $S(nd^4)$, but the first condition is the easiest to use, gives good results, and is a well known method.

[2] The term regression was introduced by Mr. Galton. He was dealing with the correlation of the heights of sons with their fathers' heights. In this case it is found that the mean height of the sons of parents of a given type is nearer to the mean height of the general population than their parents' height is. That is to say, there is a "stepping back" or "regression" of the sons toward the population-mean.

Now this minimal condition may be regarded from a different point of view. Let O in fig. (1) be the mean of the whole table, *e.g.*, the point corresponding to mean age of husband and mean age of wife, and let us measure our variables not from absolute zero but from O. Let x and y then be a pair of associated deviations, let σ be the standard deviation[3] of any array about its own mean, and let

$$X=a+bY$$

be the equation to RR. Then for any one array[4]

$$S\{x-(a+by)\}^2=S\{x-(a+bY)\}^2=n\sigma^2+nd^2$$

Hence extending the meaning of S to summation over the whole table,

$$S(nd^2)=S\{x-(a+by)\}^2-S(n\sigma^2).$$

But in this expression $S(n\sigma^2)$ is independent of a and b; it is in fact a characteristic of the table. Therefore making $S(nd^2)$ a minimum is equivalent to making

$$S\{x-(a+by)\}^2$$

a minimum. That is to say, we form a *single valued relation*

$$x=a+by$$

between a pair of associated deviations, such that the sum of the squares of our errors in estimating any one x from its y by the equation is a minimum. This single valued relation, which we may call the characteristic relation, is simply the equation to the line of regression RR. There will be two such equations to be formed corresponding to the two lines of regression.

The idea of the method may at once be extended to the case of correlation between several variables. Let x_1, x_2, x_3, &c., be now a group of associated deviations. Let n be the number of observations in an array of x_1's associated with fixed types X_2, X_3, X_4, &c., of the remaining variables, let σ_1 be the standard deviation of this array, and let d be the difference of its mean from the value given by a regression equation

$$X_1=a_{11}+a_{12}\,X_2+a_{13}\,X_3+a_{14}\,X_4+\ .\ .\ .\ ,$$

then as before we shall determine the coefficients a_{11}, a_{12}, &c., so as to make $S(nd^2)$ a minimum. But this is again equivalent to making

$$S\{x_1-(a_{11}+a_{12}\,x_2+a_{13}\,x_3+\ .\ .\ .)\}^2$$

a minimum, for

$$S\{x_1-(a_{11}+a_{12}\,x_2+a_{13}\,x_3+\ .\ .\ .)\}^2=S(n\sigma_1{}^2)+S(nd^2).$$

[3] Standard deviation = Gauss's "mean error." It is the square root of the mean (deviation)² of the array from its mean, *i.e.*, $\sigma^2=S\ (x^2)/n$, n being the total frequency of the array.

[4] The second sum in the expression below is the standard deviation of the array about a point differing by $(a+by)$ or d from the mean, and we have used the usual transformation (*cf.* below p. 822).

Hence we may say that we solve for a single valued relation

$$x_1 = a_{11} + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + \ldots$$

between the associated deviations of our variables; the relation being such that the sum of the squares of the errors made in estimating any one x_1 from its associated x_2, x_3, x_4, &c., is the least possible. We have laid some stress throughout on the fact that the relation is *single valued*, *i.e.*, it gives only one value of x_1 for a given set of values of x_2, x_3, x_4, &c. In actual statistics we know that we will find not one value of x_1 but a whole series, fluctuating round some *mean*. It is this *mean* value that is given by the estimated x_1, to a greater or less degree of approximation, according as the actual regressional relation approximates more or less closely to the linear form. If the regressional relation be *strictly* linear, *i.e.*, if the mean of the array be strictly a linear function of its associated types, then the mean x_1 is given strictly by the characteristic equation.

If we call deviations in excess of the mean positive, and those in defect negative, the characteristic tells us not only *whether a positive or negative* x_1 may on the average be expected for given values of x_2, x_3, x_4, &c., but also *how great* the value of x_1 will be on the average. The method is not only qualitative but also quantitative.

It might appear a more natural proceeding to form a "characteristic relation" between the absolute values of the variables, and not their deviations from the mean, for it is the absolute values that must in general be regarded as related. This may however be most conveniently done by working with the mean as origin until the characteristic is obtained, and then transferring the equation to zero as origin. The work would be much more laborious, and would only lead to the same final result, if zero were used as origin from the commencement. The procedure is illustrated by the arithmetical examples given later on (pp. 822, *et seq.*).

It must be understood that we only take a *linear* characteristic relation because it is the simplest possible form to calculate. We could on precisely the same principles solve for some more general equation such as—

$$\begin{aligned} x_1 = a_{11} & \\ & + a_{12}x_2 + b_{12}x_2^2 + c_{12}x^3{}_2 + \ldots \\ & + a_{13}x_3 + b_{13}x_3^2 + c_{13}x_3^3 + \ldots \\ & + a_{14}x_4 + b_{14}x_4^2 + c_{14}x_4^3 + \ldots \\ & + \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot + \ldots \end{aligned}$$

which would doubtless give the mean of the x_1-array with a greater degree of accuracy. But I think it will be found that the amount of arithmetic involved in arriving at the linear characteristic for four or five variables will be quite sufficient to content the most enthusiastic statistician. In point of fact few statistics would seem worth the labour of calculating any characteristic more complex than the linear. In most cases the deviation from the linear character, at all events near the middle of the table where frequencies are greatest, does not appear to be very serious even though well defined.

Reviewing briefly the contents of this section, we may say that the present method consists in forming a linear equation between any one variable x_1 of a group, and the other variables x_2, x_3, x_4, &c.; this equation being so formed that the sum of the squares of the errors made in estimating x_1 from its associated variables x_2, x_3, x_4, &c., is the least possible. This relation is termed the characteristic relation. It is evident that if there are n variables, n characteristics can be formed between them, expressing each one in turn in terms of the others. The magnitude and sign of the coefficients of the x's on the right of such an equation show in what direction and to what extent the average of x_1 will be altered when x_2, x_3, x_4, &c., undergo alterations of any given magnitude and sign.

We may now proceed to the detailed discussion of the special cases of two, three, or more variables.

(1) Case of two variables.

Let x and y be a pair of associated deviations measured from the mean X and the mean Y. Then using S to denote summation over the whole series of observations, we have[5]

$$\mathrm{S}(x) = \mathrm{S}(y) = 0.$$

The characteristic or regression equations which we have to find are of the form

$$\left.\begin{aligned} x &= a_1 + b_1 y \\ y &= a_2 + b_2 x \end{aligned}\right\} \quad \ldots \quad \ldots \quad \ldots \quad (1)$$

Taking the equation for x first, the two normal equations[6] for determining a_1 and b_1 so that

$$\mathrm{S}\{x - (a_1 + b_1 y)\}^2$$

[5] In case this property of the mean may not be familiar, we may prove it. The ordinary definition of the mean is—

$$\mathrm{M} = \mathrm{S}(\mathrm{X})/\mathrm{N}$$

X being the absolute values of the variables. Now measure the X's, not from zero, but from a new origin m, so that

$$\mathrm{X} = m + \xi.$$

Then

$$\begin{aligned} \mathrm{S}\,(\mathrm{X}) &= m\,.\,\mathrm{N} + \mathrm{S}(\xi) \\ \mathrm{M} &= m + \mathrm{S}(\xi)/\mathrm{N}. \end{aligned}$$

Therefore if m and M are identical, S(ξ) must be zero.

[6] We cannot digress here into the proof of the method for forming the normal equations in the method of least squares, but must refer to the text-books

shall be a minimum, are

$$\left.\begin{aligned} S(x) &= Na_1 + b_1 S(y) \\ S(xy) &= a_1 S(y) + b_1 S(y_2) \end{aligned}\right\} \quad \ldots \quad \ldots \quad (2)$$

where N is the whole number of observations, or rather of correlated pairs. The first equation gives us at once

$$a_1 = 0.$$

From the second we have

$$b_1 = \frac{S(xy)}{S(y^2)}.$$

To simplify our notation let us write

$$S(x^2) = N\sigma_1^2. \qquad S(y^2) = N\sigma_2^2.$$
$$S(xy) = Nr\sigma_1\sigma_2.$$

σ_1 and σ_2 are then the two standard deviations or errors of mean square, measuring the degree of scatter of the X's and Y's round their mean values. Rewriting the value of b_1 in terms of these symbols, we have

$$b_1 = r\frac{\sigma_1}{\sigma_2} \quad \ldots \quad \ldots \quad \ldots \quad (3)$$

Similarly if we take the second equation, expressing y in terms of x, we get

$$a_2 = 0$$
$$b_2 = r\frac{\sigma_2}{\sigma_1} \quad \ldots \quad \ldots \quad \ldots \quad (4)$$

That is to say, the two characteristics are

$$\left.\begin{aligned} x &= r\frac{\sigma_1}{\sigma_2}y \\ y &= r\frac{\sigma_2}{\sigma_1}x \end{aligned}\right\} \quad \ldots \quad \ldots \quad \ldots \quad (5)$$

Since the two constants on the right hand side have vanished, it follows that the two characteristic lines must both pass through the mean of the whole table.

The standard deviations σ_1 and σ_2 are both essentially positive quantities. The number r may be either positive or negative, according as x's of either signs are generally associated with y's of

(*e.g.*, Merriman's "Method of Least Squares," 6th edit., New York, 1894). The rule may, however, be stated. Let

$$x_0 = a_1x_1 + a_2x_2 + a_3x_3 + \ldots + a_nx_n$$

be an "observation equation." Multiply all through by the coefficient of a_{n-m} (the a's being the unknown constants to be found), and add together all the equations so obtained, giving

$$\begin{aligned} S(x_0x_{n-m}) = a_1\, &S(x_1 \,.\, x_{n-m}) + a_2S(x_2 \,.\, x_{n-m}) \\ &+ \;.\;.\;.\;.\;.\;.\;.\;.\;.\;.\;.\;. \\ &+ a_{n-m}\, S(x^2_{n-m}) + \;.\;.\;.\;. \\ &+ a_nS(x_{n-m}x_n) \end{aligned}$$

This is the $(n—m)$th normal equation,

the same or of opposite sign. Hence the numbers b_1 and b_2, that is $r\sigma_1/\sigma_2$ and $r\sigma_2/\sigma_1$, may also be either positive or negative. This is obviously necessary when we remember the physical meaning of the b's; they are measures of the shift of one variable corresponding on an average to a given shift of the other, and these shifts may be either of the same sign or of opposite sign. The b's are conveniently termed the coefficients of regression, or simply the regressions.

It must be noted that if the regression be really linear, *i.e.*, if the curves of regression of fig. 1 become straight lines, the equations (5) are quite strictly the equations to these straight lines; $b_1 y$ then gives the mean of the x-array of type y, $b_2 x$ gives the mean of the y-array of type x. This theorem admits of a direct and simple geometrical proof. Let n be the total frequency in any one x-array (fig. 2), and let θ be the angle that the line of regression makes with the axis of y. We have to show that

$$AB = b_1 \,.\, OA,$$

or

$$\tan \theta = b_1 \,.$$

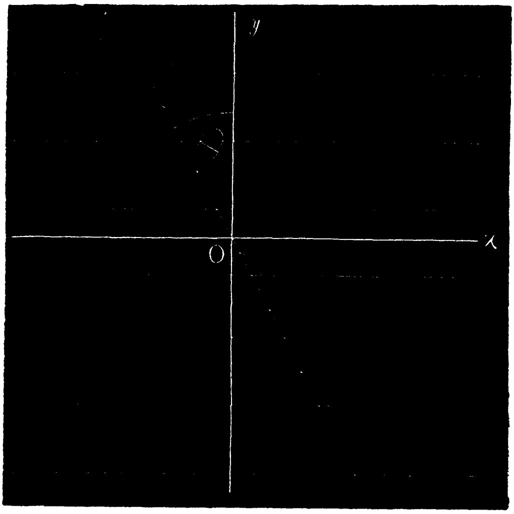

Fig. 2.

We have for a single array

$$S(xy) = yS(x) = ny^2 \tan \theta,$$

or extending the significance of S to summation over the whole surface,

$$S(xy) = \tan \theta \,.\, S(ny^2) = \tan \theta \,.\, N\sigma_2^2,$$

that is

$$\tan \theta = \frac{S(xy)}{N\sigma_2^2} = r\frac{\sigma_1}{\sigma_2}\,.$$

Although the regressions may be either positive or negative, they must both have the same sign, viz., the sign of r. This quantity r is most important. Suppose that instead of measuring x and y in arbitrary units, we measure each in terms of its own standard deviation as unit. Then let us write

$$\left(\frac{x}{\sigma_1}\right) = \rho\left(\frac{y}{\sigma_2}\right) \quad \ldots \quad \ldots \quad \ldots \quad (6),$$

and solve for ρ by the method of least squares. We have omitted a constant on the right hand side, since it would vanish as before. We have at once—

$$\frac{S(xy)}{\sigma_1\sigma_2} = \rho\,\frac{S(y^2)}{\sigma_2^{\,2}},$$

$$\rho = \frac{S(xy)}{N\sigma_1\sigma_2} = r \quad \ldots \quad \ldots \quad \ldots \quad (7).$$

But r is symmetrical with regard to x and y, that is to say, we would have got the same value for ρ from the equation

$$\left(\frac{y}{\sigma_2}\right) = \rho\left(\frac{x}{\sigma_1}\right)$$

by again using the method of least squares. Hence—*if we measure x and y each in terms of its own standard deviation, r becomes at once the regression of x on y, and the regression of y on x, these two regressions being then identical.*[7]

Again, let us form the sums of the squares of residual errors in equations (5) and (6). Inserting the values of b_1, b_2, and ρ, we get—

$$\left.\begin{aligned} S(x - b_1y)^2 &= N\sigma_1^2(1 - r^2) \\ S(y - b_2x)^2 &= N\sigma_2^2(1 - r^2) \\ S\left(\frac{x}{\sigma_1} - \rho\,\frac{y}{\sigma_2}\right)^2 = S\left(\frac{y}{\sigma_2} - \rho\,\frac{x}{\sigma_1}\right)^2 &= N(1 - r^2) \end{aligned}\right\} \ldots \quad (8).$$

All these quantities, being the sums of series of squares, must necessarily be positive. *Hence, r cannot be numerically greater than unity.* Further, if r be equal to ± 1, all the above three sums become zero. But

$$S\left(\frac{x}{\sigma_1} \pm \frac{y}{\sigma_2}\right)^2$$

can only vanish if

$$\frac{x}{\sigma_1} \pm \frac{y}{\sigma_2} = 0$$

in every case, or if the relation hold good

$$\frac{x_1}{y_1} = \frac{x_2}{y_2} = \frac{x_3}{y_3} = \ldots\ldots = \pm\frac{\sigma_1}{\sigma_2} \quad \ldots \quad \ldots \quad (9)$$

[7] This property of r has been used as the fundamental one by Mr. Francis Galton for the case of normal correlation. *Vide* "Correlations and their Measure-"ment."—"Proceedings Royal Society, 1888," vol. xlv, p. 135.

the sign of the last term being the sign of r. That is to say, *when the value of r is unity, all pairs of deviations bear the same ratio to one another, or the values of the two variables are related by a simple linear law.* The ordinary scattered correlation table has collapsed into a distribution of frequency along a straight line. The greater the value of r the more nearly does this theorem hold good, as evidenced by the expression (8). Hence the number r is termed the *coefficient of correlation.* It must be borne in mind that if the true regression be not linear r can never become unity, though we know from experience that it can approach pretty closely to that value. If the regression be very far from linear some caution must evidently be used in employing r to compare two different distributions.[8] When r is unity we may say that the two variables are perfectly correlated, but when it is zero we cannot say that they are strictly uncorrelated, although both regressions are zero. For the fact that $r=0$ does not *in general* imply that the variables are strictly independent in the sense that the chance of getting a given pair of deviations is equal to the product of the chances of getting either separately. The condition $r = 0$ is necessary but is not sufficient.

Referring again to the expressions (8), we see that $\sigma_1\sqrt{1-r^2}$ is the standard error made in estimating x from the characteristic

$$x = r\frac{\sigma_1}{\sigma_2}\,.\,y,$$

and similarly $\sigma_2\sqrt{1-r_2}$ is the standard error made in estimating y from the second characteristic

$$y = r\frac{\sigma_2}{\sigma_1}\,.\,x.$$

If the form of the correlation be such that the regression is linear, and also the standard deviations of all parallel arrays are equal, then $\sigma_1\sqrt{1-r^2}$ is the standard deviation of every x-array, and $\sigma_2\sqrt{1-r^2}$ is the standard deviation of every y-array. This theorem is of use for the special case of "normal correlation," to be dealt with separately later. To facilitate the calculation of these important quantities we have given in the Appendix (Table I) a table of the values of $\sqrt{1-r^2}$ for values of r progressing by hundredths. This will suffice for most ordinary calculations.

[8] One seems almost to require a generalised correlation coefficient, measuring the approach of the distribution towards a single-valued law *of any form,* for the case of general correlation. It would be easy to get a limit to such a coefficient by finding the standard deviation from the actual line of means.

Examples of the Arithmetical Work Involved in Calculating Standard Deviations, Regressions, and Coefficients of Correlation.

It must be remembered that

$$\sigma_1^2=\frac{S(x^2)}{N} \qquad \sigma_2^2=\frac{S(y^2)}{N} \qquad r=\frac{S(xy)}{N\sigma_1\sigma_2}$$

x and y being deviations measured from their respective means, so that

$$S(x)=0. \qquad S(y)=0.$$

The first thing to be done is consequently to calculate the means. The work for this may often be greatly abbreviated by measuring the variables from an arbitrary origin instead of from zero. Thus let X be the absolute magnitude of any one of the x-variables. Then in the ordinary method of calculating

$$\text{mean } X=\frac{S(X)}{N},$$

N being the total number of them.

But now suppose we measure the X's from an arbitrary origin D, so that

$$X=\xi+D;$$

then

$$S(X)=S(\xi)+N.D,$$

or

$$\text{mean } X=D+\frac{S(\xi)}{N} \quad \dots \quad \dots \quad \dots \quad (I)$$

Now let us find the standard deviation round the mean in terms of the standard deviation round the arbitrary origin D Let us write for brevity

$$\frac{S(\xi)}{N}=d.$$

Then if x be as before a deviation from the mean,

$$\xi=x+d,$$

and

$$S(\xi^2)=S(x+d)^2$$
$$=S(x^2)+2dS(x)+Nd^2$$

or as $S(x)=0$

$$\frac{S(x^2)}{N}=\frac{S(\xi^2)}{N}-d^2 \quad \dots \quad \dots \quad \dots \quad \dots \quad (II.)$$

That is, using *sd* as a convenient abbreviation for standard deviation, *the* $(sd)^2$ *round the mean is equal to the* $(sd)^2$ *round any other origin less the square of the distance between the mean and that origin.*

Finally let us take the value of the product sum $S(xy)$. If $\xi\eta$ be the deviations as measured from the arbitrary origin from which the coordinates of the mean are $\bar{x}\,\bar{y}$ (which must of course be given their proper sign),

$$\xi=x+\bar{x}, \quad \eta=y+\bar{y}$$

therefore

$$S(\xi\eta)=S(x+\bar{x})(y+\bar{y})$$

or, expanding and dropping the sums that vanish,

$$S(xy)=S(\xi\eta)-N\overline{xy} \quad \ldots \quad \ldots \quad \text{(III)}$$

As an example of the work let us take the correlation table below,[9] which exhibits the correlation between the per-

Number Relieved Outdoors to One Indoors.	Percentage of Males over 65 in Receipt of Relief.									
	0—5.	5—10.	10—15.	15—20.	20—25.	25—30.	30—35.	35—40.	40—45.	Total.
0— 1........	0·5 **9**	6·0 **6**	9·0 **3**	1·0 **0**	— —	— —	— —	1·0 **12**	— —	17·5 —
1— 2........	3·5 **6**	13·0 **4**	10·0 **2**	14·0 **0**	5·0 **2**	— —	— —	— —	— —	45·5 —
2— 3........	1·0 **3**	4·5	13·0 **1**	13·5 **0**	14·0 **1**	2·0 **2**	— —	— —	— —	48·0 —
3— 4........	1·0 **0**	4·5 **0**	7·5 **0**	14·0 **0**	14·0 **0**	3·0 **0**	— —	— —	— —	44·0 —
4— 5........	— —	1·0 **2**	6·0 **1**	11·5 **0**	8·5 **1**	1·0 **2**	— —	— —	— —	28·0 —
5— 6........	— —	— —	3·5 **2**	3·0 **0**	4·5 **2**	2·0 **4**	— —	— —	— —	13·0 —
6— 7........	— —	1·0 **6**	2·0 **3**	1·0 **0**	2·0 **3**	4·0 **6**	1·0 **9**	— —	— —	11·0 —
7— 8........	— —	0·5 **8**	1·0 **4**	1·0 **0**	3·0 **4**	— —	— —	— —	— —	5·5 —
8— 9........	— —	0·5 **10**	1·0 **5**	1·0 **0**	1·0 **5**	4·0 **10**	— —	— —	— —	7·5 —
9—10........	— —	1·0 **12**	— —	2·0 **0**	4·0 **6**	— —	— —	— —	— —	7·0 —
10—11..........	—	—	—	—	—	—	—	—	—	—
11—12........	— —	— —	— —	— —	2·0 **8**	— —	— —	— —	— —	2·0 —
12—13........	— —	— —	1·0 **9**	— —	— —	— —	— —	— —	— —	1·0 —
13—14........	— —	1·0 **20**	— —	— —	— —	— —	— —	— —	— —	1·0 —
14—15..........	—	—	—	—	—	—	—	—	—	—
15—16........	— —	— —	— —	1·0 **0**	— —	1·0 **24**	— —	— —	— —	2·0 —
16—17..........	—	—	—	—	—	—	—	—	—	—
17—18........	— —	— —	— —	— —	— —	1·0 **28**	— —	— —	— —	1·0 —
18—19........	— —	— —	— —	— —	1·0 **15**	— —	— —	— —	— —	1·0 —
Totals	6·0	33·0	54·0	63·0	59·0	18·0	1·0	1·0	—	235·0

[9] "Economic Journal," vol. vi (1896), p. 623, Table VII. I regret to have discovered several errors in completely reworking the nine correlation coefficients of p. 618 of this article. For 0·13 *read* 0·23, for 0·29 *read* 0·25, for 0·40 *read* 0·34. The second error is due to a blunder in standard deviation; for 4·97 *read* 5·75. The third is the present example. None of these errors at all affect the conclusion of the article.

centage of males over 65 in receipt of relief in a certain group of 235 "mostly rural" unions, and the ratio of the number in receipt of out-door to those in receipt of in-door relief. The frequencies[10] given are consequently numbers of unions. It is evident that a slight difficulty in grouping arises if an out-relief ratio happens to work out to a whole number. In such a case that union has to be divided between two rows, and hence the half unions of the table. A union with the out-relief ratio 4, for example, was divided between the rows 3—4 and 4—5, and so on. Exact integral pauperism did not occur, or at all events not so as to affect the table. All the unions in any one compartment of the table are, for the purposes of calculation, assumed to have the out-relief ratio and pauperism corresponding to the centre of that compartment; the fourteen unions, for instance, in row 3—4 Col. 15—20, are assumed to have an out-relief ratio 3·5 and (male) old age pauperism 17·5.

The calculations we will take in the following order—(1) Mean and standard deviation of pauperism. (2) Mean and standard deviation of out-relief ratio. (3) The product-sum and the coefficient of correlation.

The frequencies of the different groups of pauperism are given in the first column of the table below, being taken from the bottom row of the correlation table. The value of pauperism 17·5 per cent. was taken as arbitrary origin, corresponding to the centre of the group with frequency 63, so that ξ for this group is zero. 5 per cent. of pauperism is then taken as the unit, so that the ξ's of successive groups are (+ or −) 1, 2, 3, &c. Calling any frequency f, the third column contains the products $f\xi$; multiplying the figures of this column by the ξ's again, we get the products $f\xi^2$.

Mean and Standard Deviation of Pauperism.

1	2	3	4
f.	ξ.	$f\xi$.	$f\xi^2$.
6	− 3	18	54
33	− 2	66	132
54	− 1	54	54
63	0	− 138	
59	+ 1	59	59
18	+ 2	36	72
1	+ 3	3	9
1	+ 4	4	16
235	....	+ 102 − 138	396
		− 36	

[10] The frequencies are the ordinary figures. The *heavy type* figures are only the values of $\xi\eta$, and serve in reckoning $S(xy)$.

Hence

$$S(\xi) = -36. \qquad S(\xi^2) = 396$$

$\therefore$

$$d = -\frac{36}{235} = -{\cdot}1532 \text{ unit}$$
$$= -{\cdot}766 \text{ per cent.}$$

$\therefore$ Mean pauperism $= 17{\cdot}5 - {\cdot}77$
$= 16{\cdot}73$ per cent.

$$\frac{S(\xi^2)}{N} = \frac{396}{235} = 1{\cdot}6851$$

$\therefore$

$$\sigma_1{}^2 = 1{\cdot}6851 - d^2$$
$$= 1{\cdot}6616$$

$\therefore$

$$\sigma_1 = \sqrt{1{\cdot}6616} = 1{\cdot}289 \text{ units,}$$

or

S.D. of pauperism $= 6{\cdot}445$ per cent.

Proceeding now to the Out-Relief Ratio, and taking the total frequencies from the "Total" column on the right of our correlation-table we get the table below, choosing as origin the group 3 and 4 with frequency 44:—

Mean and Standard Deviation of Out-Relief Ratio.

1 *f*.	2 ξ.	3 *f*ξ.	4 *f*ξ².
17·5	− 3	52·5	157·5
45·5	− 2	91	182
48	− 1	48	48
44	0	− 191·5	....
28	+ 1	28	28
13	+ 2	26	52
11	+ 3	33	99
5·5	+ 4	22	88
7·5	+ 5	37·5	187·5
7	+ 6	42	252
....	+ 7	....	....
2	+ 8	16	128
1	+ 9	9	81
1	+ 10	10	100
....	+ 11	....	
2	+ 12	24	288
....	+ 13	....	....
1	+ 14	14	196
1	+ 15	15	225
235	....	+ 276·5	2,112
		− 191·5	
		+ 85	

That is to say we have

$$S(\xi) = +\,85. \qquad S(\xi^2) = 2112$$

$$\therefore \qquad d = +\frac{85}{235} = {\cdot}3617$$

$$\therefore \qquad \text{Mean out-relief ratio} = 3{\cdot}5 + {\cdot}36$$
$$= \underline{3{\cdot}86}$$

$$\frac{S(\xi^2)}{N} = \frac{2112}{235} = 8{\cdot}9872$$

$$\therefore \qquad \sigma_2{}^2 = 8{\cdot}9872 - d^2$$
$$= 8{\cdot}8564$$

$$\therefore \qquad \sigma_2 = \sqrt{8{\cdot}8564} = 2{\cdot}976$$

As we have grouped by units in this case, this is the standard deviation of the out-relief ratio.

The only work now left is the calculation of the product sum.

The table is now divided into four quadrants by the rows $x = 0$, $y = 0$ (pauperism 15 — 20, and out-relief 3 — 4).

Counting deviations in excess positive, and in defect negative, the products $(\xi\eta)$ will be positive in the upper left hand and lower right hand quadrants, and negative in the two others.

1	2	3	4	5
			$f \times (\xi\eta)$.	
$(\xi\eta)$.	Frequencies.	Total.	Positive.	Negative.
1	8 5 + 13 − 14 − 6	+ 1·5	1·5	....
2	4·5 + 1 + 10 + 4·5 − 2 − 5 − 1´ − 3·5	+ 8·5	17	....
3	2 + 9 + 1 − 2	+ 10	30	....
4	3 + 2 + 13 − 1	+ 17	68	....
5	1 − 1	0	....	....
6	4 + 4 + 6 + 3·5 − 1	+ 16·5	99	....
8	2 − 0·5	+ 1·5	12	....
9	1·5 − 1	+ 0·5	4·5	....
10	4 − 0·5	+ 3·5	35	....
12	− 2	− 2	....	24
15	+ 1	+ 1	15	....
20	− 1	− 1	....	20
24	+ 1	+ 1	24	....
28	+ 1	+ 1	28	....
Totals	....	....	334 44	44
			+ 290	

The whole table must now be gone through, and the value of $(\xi\eta)$ for each compartment entered below its frequency; these figures have been printed in *heavy type* in our table. Finally, all the frequencies of compartments having the same value of $(\xi\eta)$ must be collected together, as in the second column of the table above, and their sum entered as in the third column, counting as negative the frequencies in the negative quadrants. Finally the products of these "total frequencies" with their corresponding values of $(\xi\eta)$ are formed and entered in Cols. 4 and 5, the positive in 4, and the negative in 5. These two columns are then added up separately and the negative deducted from the positive total, leaving the value of $S(\xi\eta)$ 290 in the present case.

It must be remembered that our units have been throughout the units of grouping. Therefore $\bar{x}$ and $\bar{y}$ (equation III) are $-{\cdot}1532$ and $+{\cdot}3617$ respectively. Therefore

$$\begin{aligned} S(xy) &= S(\xi\eta) - N\bar{x}\bar{y} \\ &= 290 + (235 \times {\cdot}3617 \times {\cdot}1532 \\ &= 290 + 13{\cdot}02 = 303{\cdot}02 \end{aligned}$$

$$\begin{aligned} r &= \frac{S(xy)}{N\,\sigma_1\sigma_2} \\ &= \frac{303{\cdot}02}{235 \times 1{\cdot}289 \times 2{\cdot}976} \\ &= \frac{303{\cdot}2}{9014{\cdot}6} = +{\cdot}336. \end{aligned}$$

If we wish to proceed to the determination of the regressions and characteristic equations, we have

Regression of pauperism on out-relief

$$\begin{aligned} &= r\frac{\sigma_1}{\sigma_2} = {\cdot}336 \times \frac{6{\cdot}445}{2{\cdot}976} \\ &= {\cdot}728 \text{ per cent.,} \end{aligned}$$

i.e., the percentage of males over 65 in receipt of relief increases by ·728, on the average, for every increase of unity in the out-relief ratio.

The characteristic equation between pauperism and out-relief ratio is consequently

$$x = {\cdot}728y,$$

where x is the deviation of the pauperism from the average, and y the deviation of the out-relief. But if X and Y are the actual values (not deviations)

$$\begin{aligned} X &= x + \text{mean} = x + 16{\cdot}73 \\ Y &= y + \text{mean} = y + 3{\cdot}86. \end{aligned}$$

Therefore we may write the above

$$\begin{aligned} (X - 16{\cdot}73) &= {\cdot}728(Y - 3{\cdot}86) \\ X &= 13{\cdot}92 + {\cdot}728Y, \end{aligned}$$

or in words—

The percentage of males over 65 in receipt of relief = 13·92 per cent. + ·728 times the out-relief ratio.

The average standard deviation of the array of pauperism (from this line of regression), or the standard error made in using this relation for estimating the pauperism of any union from its out-relief ratio is $\sigma_1\sqrt{1-r^2}$, or,

$$6{\cdot}445\sqrt{{\cdot}8871}$$
$$= 6{\cdot}445 \times {\cdot}9419 = 6{\cdot}071.$$

If the distribution be approximately "normal," the mean error is ·674498 . . . time this, or, 4·095, *i.e.*, one may expect to make an error of 4 per cent. pauperism as often as not.

The second characteristic, between out-relief ratio and pauperism is

$$y = {\cdot}155x\,;$$

or,

$$(\mathrm{Y} - 3{\cdot}86) = {\cdot}155(\mathrm{X} - 16{\cdot}73)$$
$$\mathrm{Y} = 1{\cdot}27 + {\cdot}155\mathrm{X}.$$

This is the relation that would have to be used if it were desired to estimate out-relief ratio from pauperism; it is not of so much interest as the first. The value of $\sigma_2\sqrt{1-r^2}$ is 2·803.

To complete the work we have given (fig. 3) a small

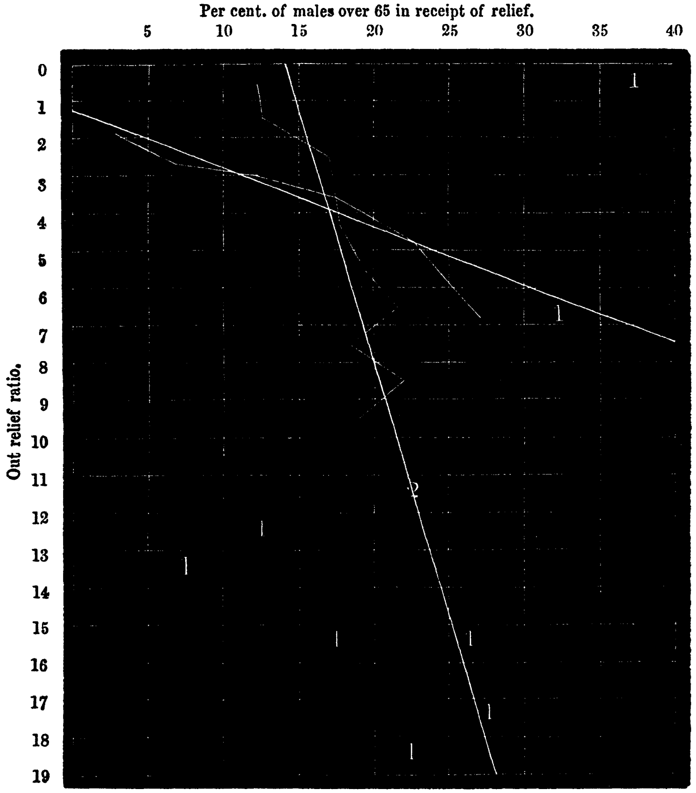

Fig. 3.

Lines of regression (out-relief ratio and pauperism). 235 Unions.

diagram of the lines of regression and the actual means of rows and columns. At the extremities of the figure where there are only one or two observations to an array we have not entered the means, but only the observations. It must be remembered that we have taken an example based on a comparatively small number of observations (235), in order to avoid unwieldy arithmetic in the text, and consequently the results are decidedly irregular.[11]

Unfortunately in practice one has frequently to be content with a much smaller number of observations than two hundred; with so few, in fact, that there is not sufficient material for forming a table of double entry. In this case the form of the arithmetic is somewhat altered. We have thought it worth while to give a special example.

In the table below we have tabulated: (1) the estimated earnings[12] of agricultural labourers in the thirty-eight unions investigated by the Royal Commission on Labour; (2) the pauperism of those unions on 1st January, 1891. Required, to find the correlation between these two quantities, and the regression of pauperism on earnings.

The mean earnings and mean pauperism we have now found by the ordinary straightforward method, adding up Cols. 2 and 3 and dividing the totals by 38.

$$\text{Mean earnings} = \frac{605{\cdot}67}{38} = 15{\cdot}94 \text{ shillings.}$$

$$\text{Mean pauperism} = \frac{139{\cdot}62}{38} = 3{\cdot}67 \text{ per cent.}$$

Having obtained these means, two new columns are formed, giving the deviation of each observation from the mean of its column, taking care to affix to each deviation its proper sign. Thus the numbers in Col. 4 are earnings minus 15·94, and those in Col. 5 pauperism minus 3·67; these deviations are the x's and y's of our previous notation. In Cols. 6 and 7 are entered the squares of the deviations x^2 and y^2, and these two columns are then added up. As we have been working from the mean to a sufficiently close approximation, we have at once without any correction

$$\sigma_1 = \sqrt{\frac{111{\cdot}2997}{38}} = \sqrt{2{\cdot}9289} = 1{\cdot}711 \text{ shillings.}$$

$$\sigma_2 = \sqrt{\frac{63{\cdot}0556}{38}} = \sqrt{1{\cdot}6594} = 1{\cdot}288 \text{ per cent.}$$

Finally in Cols. 8 and 9 are collected the successive

[11] For probable errors of correlation coefficients, see p. 847.

[12] "Earnings from Royal Commission on Labour. The Agricultural Labourer; General Report. By Mr. W. C. Little. [C–6894—xxv] 1894; last column of table on p. 80. Pauperism (percentage of population in receipt of relief of any

products (xy); the negative products in 8 and the positive in 9, so that they can be added up separately. We have finally

$$S(xy) = -55{\cdot}6274.$$

Hence—

$$r = \frac{S(xy)}{N\sigma_1\sigma_2} = -\frac{55{\cdot}6274}{38 \times 1{\cdot}711 \times 1{\cdot}288}$$
$$= -{\cdot}664$$

The regressions are both negative, as r is negative, *i.e.*, high earnings correspond to low pauperism and *vice versâ*, as one would hope. The regression of pauperism on earnings is—

$$r\frac{\sigma_2}{\sigma} = -{\cdot}664 \times \frac{1{\cdot}288}{1{\cdot}711} = -{\cdot}500,$$

i.e., a rise of a shilling in earnings corresponds to a fall of ·5 in average pauperism. Hence the characteristic equation between pauperism and earnings may be written

$$(\text{Pauperism} - 3{\cdot}67) = -{\cdot}5\ (\text{Earnings} - 15{\cdot}94),$$

or

$$\text{Pauperism} = 11{\cdot}64 \text{ per cent.} - {\cdot}5 \text{ earnings in shillings.}$$

The standard error made in estimating the pauperism by this relation is $\sigma_2\sqrt{1-r^2}$, or, interpolating in Table I of the appendix,

$$1{\cdot}288 \times {\cdot}7477$$
$$= {\cdot}96 \text{ per cent. pauperism.}$$

In all work of this kind an arithmometer is of course perfectly invaluable, but in working such an example as the last tables are quicker. Barlow's tables[13] can be used for simply writing down the values of x^2 and y^2, and subsequently for evaluating the square roots. Crelle's[14] multiplication table similarly gives the products (xy) at once by only looking out the factors.

kind) from B Return for 1st January, 1891. We have expressed the estimated earnings in decimals of a shilling; it would have been shorter to have worked in pence throughout, or the last figure might obviously have been omitted.

[13] Barlow's "Tables of Squares, Cubes, Square Roots, Cube Roots, and Reci-"procals of all Integer Numbers up to 10,000." 8vo. (6*s*.) E. and F. N. Spon.

[14] Crelle's multiplication table, giving the products of all numbers up to 1,000 × 1,000. Folio. Nutt, or Williams and Norgate, 21*s*.

1	2	3	4	5	6	7	8	9
Union.	Earnings in Shillings.	Pauperism Per Cent.	x.	y.	x^2.	y^2.	Products.	
							Negative.	Positive.
Glendale	20·75	2·40	+4·81	−1·27	23·1361	1·6129	6·1087	....
Wigton	20·25	2·29	+4·31	−1·38	18·5761	1·9044	5·9478	....
Garstang	19·67	1·39	+3·73	−2·28	13·9129	5·1984	8·5044	....
Belper	18·50	1·92	+2·56	−1·75	6·5536	3·0625	4·5800	....
Nantwich	17·67	2·98	+1·73	−0·69	2·9929	0·4761	1·1937	....
Atcham	17·50	1·17	+1·56	−2·50	2·4336	6·2500	3·9000	....
Driffield	17·08	3·79	+1·14	+0·12	1·2996	0·0144	....	0·1368
Uttoxeter	17·00	3·01	+1·06	−0·66	1·1236	0·4356	0·6996	....
Wetherby	17·00	2·39	+1·06	−1·28	1·1236	1·6384	1·3568	....
Easingwold	16·93	2·78	+0·99	−0·89	0·9801	0·7921	0·8811	....
Southwell	16·50	3·09	+0·56	−0·58	0·3136	0·3364	0·3248	....
Hollingbourn	16·33	2·78	+0·39	−0·89	0·1521	0·7921	0·3471	....
Melton Mowbray	16·25	2·61	+0·31	−1·06	0·0961	1·1236	0·3286	....
Truro	16·25	4·33	+0·31	+0·66	0·0961	0·4356	....	0·2046
Godstone	16·00	3·02	+0·06	−0·65	0·0036	0·4225	0·0390	....
Louth	16·00	4·20	+0·06	+0·53	0·0036	0·2809	....	0·0318
Brixworth	15·75	1·29	−0·19	−2·38	0·0361	5·6644	....	0·4522
Crediton	15·67	5·16	−0·27	+1·49	0·0729	2·2201	0·4023	....
Holbeach	15·50	4·75	−0·44	+1·08	0·1936	1·1664	0·4752	....
Maldon	15·50	4·64	−0·44	+0·97	0·1936	0·9409	0·4268	....
Monmouth	15·33	4·26	−0·61	+0·59	0·3721	0·3481	0·3599	....
St. Neots	15·25	1·66	−0·69	−2·01	0·4761	4·0401	....	1·3869
Swaffham	15·00	5·37	−0·94	+1·70	0·8836	2·8900	1·5980	....
Thakeham	15·00	3·38	−0·94	−0·29	0·8836	0·0841	....	0·2726
Thame	15·00	5·84	−0·94	+2·17	0·8836	4·7089	2·0398	....
Thingoe	15·00	4·63	−0·94	+0·96	0·8836	0·9216	0·9024	....
Basingstoke	15·00	3·93	−0·94	+0·26	0·8836	0·0676	0·2444	....
Cirencester	15·00	4·54	−0·94	+0·87	0·8836	0·7569	0·8178	....
North Witchford	14·83	3·42	−1·11	−0·25	1·2321	0·0625	....	0·2775
Pewsey	14·75	5·88	−1·19	+2·21	1·4161	4·8841	2·6299	....
Bromyard	14·75	4·36	−1·19	+0·69	1·4161	0·4761	0·8211	....
Wantage	14·75	3·85	−1·19	+0·18	1·4161	0·0324	0·2142	....
Stratford-on-Avon	14·58	3·92	−1·36	+0·25	1·8496	0·0625	0·3400	....
Dorchester	14·50	4·48	−1·44	+0·81	2·0736	0·6561	1·1664	....
Woburn	14·50	5·67	−1·44	+2·00	2·0736	4·0000	2·8800	...
Buntingford	14·33	4·91	−1·61	+1·24	2·5921	1·5376	1·9964	....
Pershore	13·50	4·34	−2·44	+0·67	5·9536	0·4489	1·6348	....
Langport	12·50	5·19	−3·44	+1·52	11·8336	2·3104	5·2288	....
Totals	605·67	139·62	....	....	111·2997	63·0556	58·3898	2·7624
	....	....	....	....	....	....	2·7624	....
	....	....	....	....	....	....	55·6274	....

(2) Case of Three Variables.

Let the three variables be X_1, X_2, X_3, and let x_1, x_2, x_3, denote deviations of these variables from their respective means. Let us write for brevity, in notation similar to that of the last section,

$$S(x_1^2)=N\,.\,\sigma_1^2. \qquad S(x_2^2)=N\sigma_2^2.$$
$$S(x_3^2)=N\sigma_3^2.$$
$$S(x_1x_2)=Nr_{12}\sigma_1\sigma_2. \qquad S(x_2x_3)=Nr_{23}\sigma_2\sigma_3.$$
$$S(x_3x_1)=Nr_{31}\sigma_3\sigma_1.$$

The characteristic or regression equation to be determined will now be of the form—

$$x_1 = b_{12}x_2 + b_{13}x_3 \quad \ldots \quad \ldots \quad \ldots \quad (9)$$

b_{12} and b_{13} being the unknowns to be found from the observations by the method of least squares. A constant term on the right hand side may be at once omitted, since its least square value will be zero as before. The two normal equations for b_{12} b_{13} are

$$\begin{aligned} S(x_1x_2) &= b_{12}\, S(x_2{}^2) + b_{13}\, S(x_2x_3) \\ S(x_1x_3) &= b_{12}\, S(x_2x_3) + b_{13}\, S(x_3{}^2) \end{aligned}$$

or, replacing the sums by the symbols defined above, and simplifying,

$$\left.\begin{aligned} r_{12}\sigma_1 &= b_{12}\sigma_2 + b_{13}r_{23}\sigma_3 \\ r_{13}\sigma_1 &= b_{12}r_{23}\sigma_2 + b_{13}\sigma^3 \end{aligned}\right\} \quad \ldots \quad \ldots \quad (10).$$

Whence

$$\left.\begin{aligned} b_{12} &= \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}{}^2} \cdot \frac{\sigma_1}{\sigma_2} \\ b_{13} &= \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}{}^2} \cdot \frac{\sigma_1}{\sigma_3} \end{aligned}\right\} \quad \ldots \quad \ldots \quad (11).$$

That is, the characteristic relation between x_1 and x_2 x_3 is

$$x_1 = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}{}^2} \cdot \frac{\sigma_1}{\sigma_2}\, x_2 + \frac{r_{13} - r_{12}r_{23}\sigma_1}{1 - r_{23}{}^2\,\sigma_2}\, x_3 \quad \ldots \quad (12)$$

There are of course two other characteristic relations of this form, expressing x_2 and x_3 respectively in terms of the remaining pair of variables. The value of any b in terms of the r's can be written down from the expressions (11) by simply interchanging the suffixes. Thus b_{23} could be written down by simply writing 2 for 1 and 3 for 2 all through the expression for b_{12}.

Let

$$v = x_1 - (b_{12}x_2 + b_{13}x_3)\,;$$

i.e., let v be an error made in estimating x_1 from relation (12) or a deviation of x_1 from the value $(b_{12}x_2 + b_{13}x_3)$. Then relation (12) has been so formed that

$$S(v^2) = S[x_1 - (b_{12}x_2 + b_{13}x_3)]^2$$

is the least possible. Inserting in this expression the values of b_{12} and b_{13} from (11) we get, after some reduction,

$$\begin{aligned} S(v^2) &= N\sigma_1{}^2\left(1 - \frac{r_{12}{}^2 + r_{13}{}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}{}^2}\right) \\ &= N\sigma_1{}^2\,(1 - R_1{}^2) \quad \ldots \quad \ldots \quad \ldots \quad (13) \end{aligned}$$

say. Hence $\sigma_1\sqrt{1 - R_1{}^2}$ is the standard error made in estimating x_1 from its associated variables x_2 and x_3 by relation (12).

If we select an array of x_1's for given types, say h_2 and h_3 of x_2 and x_3, then the equation

$$h_1 = b_{12}h_2 + b_{13}h_3$$

will give the mean of this X_1-array, *so long as h_1 is really a linear function of h_2 and h_3; whatever be the distribution of frequency round the means.* If the mean of the x_1-array is not, however, a linear function of its types, the regressional relation will only give its mean to a greater or less degree of approximation. Again, if the standard deviation of all x_1-arrays be equal, as well as the means being linear functions of the types, $\sigma_1\sqrt{1 - R_1^2}$ is the standard deviation of *every array.*

The quantity R_1 is of some interest, as it exactly takes the place of r in the residual-expressions (7). R_1 may in fact be regarded as a coefficient of correlation between x_1 and (x_2, x_3). It can only rise to the value unity if the linear relation (9) or (12) hold good in every single case.

The quantities b_{12}, b_{13}, &c., may be termed the *net* or *partial* regressions of x_1 on x_2, x_1 on x_3, &c. If we write 1 for 2 and 2 for 1 all through the expression for b_{12}, we have

$$b_{21} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{13}^2} \cdot \frac{\sigma_2}{\sigma_1}$$

b_{21} being the net regression of x_2 on x_1.

In correlation of two variables we called the geometrical mean of the regressions the coefficient of correlation. Now we shall call the quantity

$$\rho_{12} = \sqrt{b_{12} \cdot b_{21}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad \ldots \quad (14),$$

where the sign of the root in the denominator is to be taken as positive, the *net* coefficient of correlation[15] between x_1 and x_2 . ρ_{12} is the coefficient of correlation for any group of X_1, X_2's associated with a single type of X_3's so long as the conditions hold that the means of all arrays are linear functions of their types. In general correlation where the net coefficient will change as we pass from one type of X_3 to another, ρ_{12} can only retain an average significance. In any case it retains three of the chief properties of the ordinary coefficients; (1) it can only be zero if both net regressions are zero; (2) it is a symmetrical function of the variables (*i.e.*, $\rho_{12} = \rho_{21}$); (3) it cannot be greater than unity; for by (13)

$$r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31} \text{ is not} > 1 - r_{23}^2,$$

[15] My quantities b_{12} b_{13}, &c., were termed by Professor Pearson ("Regression," &c., "Phil. Trans." A (1896), vol. clxxxvii, p. 287), "coefficients of double regression," and quantities like b_{12} σ_2/σ_1, &c., "coefficients of double correlation." The ρ's he did not use. Having named the ρ's net correlations, it seemed most natural to rename the b's net regressions, since the b's and ρ's are corresponding quantities. The expressions $b_{12}\sigma_2/\sigma_1$, &c., are not symmetrical functions of the variables, and may be numerically greater than unity. The quantities like R_1 might well be called coefficients of double correlation.

or adding $r_{13}^2 r_{23}^2$ to both sides, and transferring r_{13}^2 to the right hand side,

$$(r_{12} - r_{13}r_{23})^2 \text{ is not } > (1 - r_{13}^2)(1 - r_{23}^2),$$

that is, the numerator of ρ_{13} cannot be greater than its denominator.

The inequality we have used here is one of considerable interest, since if two coefficients, say r_{12} and r_{13}, be known, it gives us limits for the value of the third. Throwing it into the form

$$(r_{23} - r_{12}\, r_{13})^2 \text{ is not } > 1 + r_{12}^2\, r_{13}^2 - r_{12}^2 - r_{13}^2,$$

we have r_{23} must lie between the limits

$$r_{12}\, r_{13} \pm (\sqrt{1 + r_{12}^2\, r_{13}^2 - r_{12}^2 - r_{13}^2}).$$

The values of these limits for some special cases are collected in the following table:—

Values of r_{12} and r_{13}.	Limits of r_{23}.
$r_{12} = r_{13} = 0$	0
$r_{12} = r_{13} = \pm 1$	$+1$
$r_{12} = +1,\ r_{13} = -1$	-1
$r_{12} = 0,\ r_{13} = \pm 1$	0
$r_{12} = 0,\ r_{13} = \pm r$	$\pm\sqrt{1 - r^2}$
$r_{12} = r_{13} = \pm r$	1 and $2r^2 - 1$
$r_{12} = +r,\ r_{13} = -r$	$2r^2 - 1$ and -1
$r_{12} = r_{13} = \pm\sqrt{0{\cdot}5} = 0{\cdot}707$	0 and 1
$r_{12} = +\sqrt{0{\cdot}5}\ r_{12} = -\sqrt{0{\cdot}5}$	0 „ -1

One is rather prone to argue that if A be correlated with B, and B with C, A will be correlated with C. But the above table shows that this is quite unnecessary. A may be positively correlated with B, and B positively correlated with C, but yet A may, in general, be negatively correlated with C. Only when the (AB) and (BC) coefficients are both greater than $\sqrt{0{\cdot}5}$ or ·707 . . . can one ascribe the correct sign to the (AC) correlation, and ·707 . . . is an unusually high coefficient in practice.

There is one further point in three-variable correlation that may be cleared up. One expects in general to make a smaller standard error in estimating x_1 from two associated variables x_2 and x_3 than in estimating it from one only, say x_2. But under what conditions is this strictly true? The necessary condition is evidently simply

$$(1 - R_1^2) < (1 - r_{12}^2)$$
$$R_1^2 > r_{12}^2,$$

or

$$r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23} > r_{12}^2 - r_{12}^2\, r_{23}^2,$$

or

$$(r_{13} - r_{12}\, r_{23})^2 > 0.$$

But $(r_{13} - r_{12}r_{23})$ is the numerator of ρ_{13}, the net coefficient of correlation between x_1 and x_3. Hence the standard error in the second case will always be less than in the first so long as ρ_{13} is not zero.

This condition is somewhat interesting, leading, as it does, to unexpected results. To take an arithmetical example, suppose one had in some actual case

$$r_{12} = \pm 0{\cdot}8,$$
$$r_{23} = +0{\cdot}5, \qquad r_{13} = +0{\cdot}4,$$

one might very naturally imagine that the introduction of the third variable with a fairly high correlation-coefficient (0·4)with the first variable, would considerably lessen the standard deviation of the x_1-array, *i.e.*, increase the accuracy of estimation of x_1; but this is not so, for

$$\rho_{13} = \frac{0{\cdot}4 - (0{\cdot}5 \times 0{\cdot}8)}{\sqrt{0{\cdot}75 \times 0{\cdot}36}} = 0,$$

so the third variable would be of no assistance. Again, if $r_{13} = 0$, one cannot conclude that the third variable is of no use, for ρ_{13} will not be zero unless r_{23} also $= 0$.

Little need be said on the carrying out of the arithmetic for the case of three-variable correlation. The three standard deviations σ_1, σ_2, σ_3 must be calculated, and then the three-product sums, giving r_{12}, r_{13}, r_{23}. The net regressions can then be reckoned directly from the expressions (11) and their analogues, or, if preferred, the equations (10) (most simply in their primitive form) may be written down and solved for the b's. The last method is very quick if an arithmometer is available. The net correlation coefficients ρ may also be calculated directly, or, if all six regressions have been already evaluated, may be obtained by using the relations

$$\rho_{12} = \sqrt{b_{12}b_{21}},$$
$$\rho_{13} = \sqrt{b_{13}b_{31}},$$
$$\rho_{23} = \sqrt{b_{23}b_{32}}.$$

Finally, the residual sums like (13) should be obtained, or rather the standard deviations of arrays $\sigma_1\sqrt{1 - R_1^2}$, $\sigma_2\sqrt{1 - R_2^2}$, $\sigma_3\sqrt{1 - R_3^2}$.

All the quantities mentioned will not always be required. The nature of the problem in hand will point to the most suitable form in which to state the results. But in every case the means, standard deviations, and (gross) correlation coefficients r_{12}, r_{13}, r_{23} should be tabulated in a published work, as they enable readers of the paper to evaluate any of the omitted functions for themselves.

(3) Case of Four Variables—general case :—

The general method of procedure in correlation of several variables will have been evident from the last case. We will only ndicate briefly the results for the case of four variables, which will serve to illustrate the very rapid growth in the complexity of formulæ and arithmetic as the number of variables increases.

If x_1, x_2, x_3, x_4 be associated deviations of the four variables from their respective means, the characteristic equation will be of the form

$$x_1 = b_{12}x_2 + b_{13}x_3 + b_{14}x_4 \qquad \dots \qquad \dots \quad (14).$$

The normal equations for the b's are in our previous notation—

$$\left.\begin{aligned} r_{12}\sigma_1 &= b_{12}\sigma_2 + b_{13}r_{23}\sigma_3 + b_{14}r_{24}\sigma_4 \\ r_{13}\sigma_1 &= b_{12}r_{23}\sigma_2 + b_{13}\sigma_3 + b_{14}r_{34}\sigma_4 \\ r_{14}\sigma_1 &= b_{12}r_{24}\sigma_2 + b_{13}r_{34}\sigma_3 + b_{14}\sigma_4 \end{aligned}\right\}\dots \qquad \dots \quad (15).$$

Hence

$$b_{12} = \frac{r_{12}(1 - r_{34}^2) + r_{13}(r_{34}r_{24} - r_{23}) + r_{14}(r_{23}r_{34} - r_{24})}{(1 - r_{34}^2) + r_{23}(r_{34}r_{24} - r_{23}) + r_{24}(r_{23}r_{34} - r_{24})} \qquad (16),$$

and so on for the others. There are twelve such net-regressions altogether, but all of them will not as a rule be required. As before, we may call

$$\rho_{12} = \sqrt{b_{12}b_{21}} \qquad \dots \qquad \dots \qquad \dots \quad (17)$$

the net coefficient of correlation between x_1 and x_2. There are six such net coefficients, viz., ρ_{12}, ρ_{13}, ρ_{14}, ρ_{23}, ρ_{24}, ρ_{34}. On the assumption of truly linear regression, the ρ's become strictly the coefficients of correlation of the sub-groups, ρ_{12} being the coefficient of correlation for a group of the x_1's and x_2's associated with fixed types of the x_3's and x_4's, and so on.

If we write

$$u = x_1 - (b_{12}x_2 + b_{13}x_3 + b_{14}x_4),$$

we have, after some rather lengthy reduction,

$$\mathrm{S}(u^2) = \mathrm{N}\sigma_1^2(1 - \mathrm{R}_1^2) \qquad \dots \qquad \dots \quad (18),$$

where

$$\mathrm{R}_1^2 = \frac{\left\{\begin{aligned} & r_{12}^2 + r_{13}^2 + r_{14}^2 - r_{12}^2r_{34}^2 - r_{23}^2r_{14}^2 - r_{13}^2r_{24}^2 \\ & - 2(r_{13}r_{14}r_{34} + r_{12}r_{14}r_{24} + r_{12}r_{13}r_{23}) + 2(r_{12}r_{14}r_{23}r_{34} + r_{13}r_{14}r_{23}r_{24} + r_{12}r_{13}r_{24}r_{34}. \end{aligned}\right\}}{1 - r_{23}^2 - r_{34}^2 - r_{24}^2 + 2r_{23}r_{34}r_{24}} (19).$$

Hence, $\sigma_1\sqrt{1 - \mathrm{R}_1^2}$ is the standard error made in estimating x_1 by equation (14) from its associated variables x_2, x_3, and x_4. The quantity R may be considered as a coefficient of correlation, as in the case of three variables. It can only range between the values ± 1, and can only attain these limiting values if the linear relation (14) hold good in each individual case.

We showed at the end of the last section that the standard error made in estimating x_1 from the relation

$$x_1 = b_{12}x_2 + b_{13}x_3$$

was always less than the standard error when only x_2 was taken into account unless

$$\rho_{13} = 0.$$

We may now prove the similar theorem that where three variables are used on which to base the estimate, the standard error will be again decreased unless

$$\rho_{14} = 0.$$

The condition that $S(u^2)$ in the present case shall be less than $S(v^2)$ in the last is in fact

$$\left\{\begin{array}{l} \quad r_{12}^2 + r_{13}^2 + r_{14}^2 - r_{12}^2 r_{34}^2 - r_{23}^2 r_{14}^2 - r_{13}^2 r_{24}^2 \\ -2(r_{13}r_{14}r_{34} + r_{12}r_{14}r_{24} + r_{12}r_{13}r_{23}) \\ +2(r_{12}r_{14}r_{23}r_{34} + r_{13}r_{14}r_{23}r_{24} + r_{12}r_{13}r_{24}r_{34}) \end{array}\right\}(1 - r_{23}^2)$$
$$> (r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23})(1 - r_{23}^2 - r_{24}^2 - r_{34}^2 + 2r_{23}r_{24}r_{34}).$$

This may be finally reduced to

$$(r_{14} - r_{13}r_{34} - r_{12}r_{24} - r_{14}r_{23}^2 + r_{13}r_{23}r_{24} - r_{12}r_{23}r_{34})^2 > 0,$$

or,

$$\rho_{14}^2 > 0,$$

which proves the theorem. In practically working out a case of four-variables correlation, the *six* correlation tables (12, 13, 14, 23, 24, 34) have first to be formed, and the four standard deviations and six product sums calculated. In most cases all the four regression equations will not be required, but only one or two. If this be the case, it will probably be quickest simply to write down the equations (15) and solve them straightforwardly for the b's, without using the general solutions like (16). Similarly, it may be quicker to evaluate $S(u^2)$ in the elementary form.

$$\left.\begin{array}{l} S(x_1^2) + b_{12}S(x_2^2) + b_{13}S(x_3^2) + b_{14}S(x_4^2) \\ -2b_{12}S(x_1x_2) - 2b_{13}S(x_1x_3) - 2b_{14}S(x_1x_4) \\ +2b_{12}b_{13}S(x_2x_3) + 2b_{12}b_{14}S(x_2x_4) + 2b_{13}b_{14}S(x_3x_4) \end{array}\right\} \qquad (20),$$

instead of using the general expressions (18) and (19) in terms of the r's.

It is important to notice this last step because the general formula in the form of (19) may be difficult to arrive at in the cases of more variables than four, whereas there is never any difficulty in writing down the elementary form above. In the general cases of n variables there will be n means and standard deviations to be reckoned; and as many tables to be formed and product sums calculated as there are numbers of combinations of n things two together, or

$$\frac{n(n-1)}{2}.$$

Thus the amount of labour involved grows very rapidly with the number of variables taken into account, as is illustrated by the following table:—

Number of Variables.	Number of Correlation Coefficients.
2	1
3	3
4	6
5	10
6	15
7	21
8	28

Thus if we wished to discuss the causality of pauperism on the basis of as many as eight variables, the work involved would be something like *twenty-eight times* as much as that necessary for the example taken on pp. 824—831. The labour would, in fact, be almost impossible for a single individual. Hence the remark in the introduction (p. 817), that "the labour involved in arriving at "the linear characteristic for four or five variables, will be quite "sufficient to content the most enthusiastic statistician," without attempting to obtain any more complex form of regressional relation.

When the coefficients of correlation and standard deviations have been obtained, the general equations may be written down in the form (15) and solved for the b's. There is no further difficulty in the general case save that arising from the length of arithmetic.

We may note that although the sum of squares of residuals, $S(u^2)$ or the standard deviation of the array will in general get less and less the more variables we introduce, the decrease may be very slight, although the (gross) correlation coefficients $r_{12}\,r_{13}$, &c., are all large. The condition for the decrease is, as we have pointed out and exemplified (pp. 834 and 837), that the *net correlations* should not be zero. But the net correlations may well be zero (or very small) even though the r's are large.

Professor Pearson points out to me that as a limiting case the standard deviation of the array may vanish. In the case of two variables for instance Σ_1 will vanish if (*vide* (13))

$$r_{12}^2+r_{13}^2+r_{23}^2=1+2r_{12}r_{13}r_{23}$$

If the left hand side is greater than the right Σ_1 becomes imaginary, so the condition is a limiting one. Evidently however it is a desirable condition to approximate to if the regression equation is to be used for estimating.[16]

[16] The property is actually made use of in a Paper by Professor Pearson and Miss Lee, now printing for the Phil. Trans., for estimating the barometric height at one station from a knowledge of its value at two others. The estimate may be made extremely accurate by a proper choice of stations.

II. *On Normal Correlation.*[17]

(1.) Case of Two Variables.

In the whole of the preceding work we have made no postulates as to the form of the distribution of frequency. The whole problem lay in obtaining the characteristic relation and its interpretation.

But the much more general problem of obtaining an expression completely describing the frequency-distribution is one that may sometimes become of importance. It has only been solved for the case corresponding to the "normal curve" in the distribution of a single variable.

Suppose the distribution of one variable X_1 round its mean to be given by the normal curve

$$y=y_0e^{-\frac{x_1^2}{2\sigma_1^2}} \quad \ldots \quad \ldots \quad \ldots \quad (1)$$

Where σ_1 is the standard deviation of the variable, and x_1 a deviation from the mean. Similarly, suppose the distribution of a second variable x_2 to be given by

$$y'=y_0'\,e^{-\frac{x_2^2}{2\sigma_2^2}} \quad \ldots \quad \ldots \quad \ldots \quad (2)$$

Then, *if* X_1 *and* X_2 *be uncorrelated*, the chance of a pair of observed deviations lying between $x_1-\frac{1}{2}\delta x_1$, $x_1+\frac{1}{2}\delta x_1$ and $x_2-\frac{1}{2}\delta x_2$, $x_2+\frac{1}{2}\delta x_2$ must be proportional to

$$z=y_0y_0'\,e^{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2}+\frac{x_2^2}{\sigma_2^2}\right)} \quad \ldots \quad \ldots \quad (3)$$

This is what we may term a special case of normal correlation surface, the correlation being nil. If x_1 and x_2 be taken as coordinates in the horizontal plane, the contour lines of the surface are the similar and similarly situated ellipses

$$\frac{x_1^2}{\sigma_1^2}+\frac{x_2^2}{\sigma_2^2}=\text{const.}$$

Any section of the surface taken by a vertical plane parallel to the x_1 axis (*i.e.*, the distribution of any x_1-array) is of the form

$$z=z_0e^{-\frac{x_1^2}{2\sigma_1^2}}$$

and is consequently a normal curve of standard deviation σ_1, the mean lying on the x_1 axis. Similar statements hold of course for sections parallel to the x_2 axis. The elliptic contour lines lie in fact as in fig. 4.

[17] The first discussion of the normal theory of frequency for more than one variable was given by Bravais ("Mémoires présentés par Divers Savants," ix, 1846).

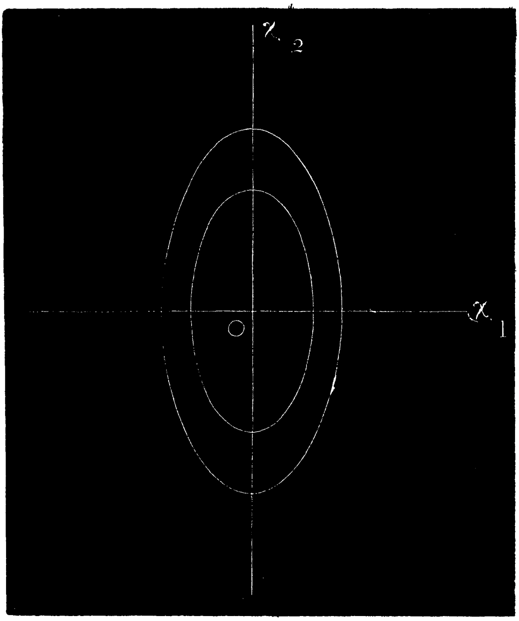

Fig. 4.

But the two variables X_1 X_2 will not, in general, be independent, and we cannot take the chance of getting a given pair of

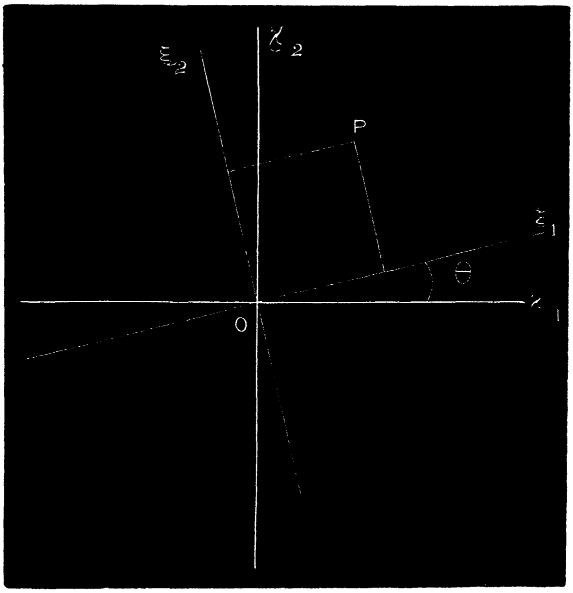

Fig. 5.

deviations to be the product of the chances of getting either separately. To illustrate what may happen in this case, suppose that having got the normal surface for x_1 and x_2, we now take a new pair of rectangular axes $O\xi_1$, $O\xi_2$, making an angle θ with the first pair, and suppose deviations to be now measured from these axes (fig. 5). Then our new variables ξ_1 and ξ_2 are connected with the old by the relations

$$\xi_1 = x_1 \cos\theta + x_2 \sin\theta$$
$$\xi_2 = x_2 \cos\theta - x_1 \sin\theta$$

and hence ξ_1 and ξ_2 *are not independent*, both being functions of x_1 and x_2. Using the converse transformation equations,

$$\left.\begin{aligned} x_1 &= \xi_1 \cos\theta - \xi_2 \sin\theta \\ x_2 &= \xi_1 \sin\theta + \xi_2 \cos\theta \end{aligned}\right\},$$

the frequency surface becomes

$$z = z_0 e^{-\left\{\frac{(\xi_1 \cos\theta - \xi_2 \sin\theta)^2}{2\sigma_1^2} + \frac{(\xi_1 \sin\theta + \xi_2 \cos\theta)^2}{2\sigma_2^2}\right\}},$$

or, reducing, the surface is of the form

$$z = z_0 e^{-(g_1\xi_1^2 + g_2\xi_2^2 - 2h\xi_1\xi_2)} \quad \dots \quad \dots \quad (4),$$

i.e, it differs from the uncorrelated surface by the introduction of the product term $\xi_1 \xi_2$ into the exponent.

The values of g_1 g_2 and h can be obtained without the use of the calculus from the work in Part I. The distribution of an array of type t is given by

$$z = z_0{}^{-(g_1\xi_1^2 + g_2t^2 - 2ht\xi_1)},$$

or transforming the exponent into a perfect square,

$$z = z_0{}' e^{-g_1\left\{\xi_1 - \frac{ht}{g_1}\right\}^2} \quad \dots \quad \dots \quad (5)$$

This is a normal distribution, the mean of which differs by ht/g_1 from the mean of the whole surface, with a standard deviation $1/\sqrt{2g_1}$. It follows that (1) the deviation of the mean of the array, is directly proportional to the type, or the regression is truly linear; (2) the standard deviations of all parallel arrays are equal and independent of their types. Hence if we write

$$S(\xi_1^2) = N \,.\, c_1^2,$$
$$S(\xi_2^2) = N \,.\, c_2^2,$$
$$S(\xi_1\xi_2) = N \,.\, rc_1c_2,$$

it follows directly from the formulæ for general correlation that

$$\frac{1}{2g_1} = c_1^2(1 - r^2)$$

or

$$g_1 = \frac{1}{2c_1^2(1 - r^2)} \quad \dots \quad \dots \quad \dots \quad (6)$$

and

$$\frac{h}{g_1} = r\frac{c_1}{c_2}$$

$$h = \frac{1}{2c_1{}^2(1-r^2)}\,r\frac{c_1}{c_2}$$

$$= \frac{r}{2c_1c_2\,(1-r^2)} \quad \ldots \quad \ldots \quad \ldots \quad (7)$$

By symmetry

$$g_2 = \frac{1}{2c_2{}^2\,(1-r^2)} \quad \ldots \quad \ldots \quad \ldots \quad (8)$$

Hence the equation to the frequency surface may be written in the form

$$z = z_0 e^{-\left\{\frac{\xi_1{}^2}{2c_1{}^2\,(1-r^2)} + \frac{\xi_2{}^2}{2c_2{}^2\,(1-r^2)} - \frac{r\xi_1\xi_2}{c_1c_2\,(1-r^2)}\right\}} \qquad (9)$$

where c_1 and c_2 are the S.D.'s of ξ_1 and ξ_2, and r is the coefficient of correlation between them. This then is the general form of normal correlation between two dependent variables ξ_1 and ξ_2. The most important properties of such a surface we have already mentioned, viz. :—

(1.) Not only must the total distributions of the two variables ξ_1 and ξ_2 be normal, but the distribution of every array must also be normal.

(2.) The regressions are truly linear, *i.e.*, the means of parallel arrays lie on straight lines.

(3.) The standard deviations of all parallel arrays are equal.

(4.) The contour lines are a system of similar and similarly situated ellipses, the centres coinciding with the mean of the whole surface.

The accompanying figure (6) shows the form of the contour lines. M M′ M M′ are the two lines of means or lines of regression, the equations to which, in the notation of Part I, are of the form

$$x = r\frac{\sigma_1}{\sigma_2}y \qquad\qquad y = r\frac{\sigma_2}{\sigma_1}x.$$

The two lines of regression are conjugates respectively to the horizontal and vertical.

The axes coinciding with the major and minor axes of the ellipses are termed the *principal axes* of the surface. But with reference to the principal axes the coefficient of correlation is zero. Hence, if $x_1\,x_2$ be two *correlated* variables with S.D.'s $\sigma_1\,\sigma_2$, it follows that we can always determine two new variables,

$$ax_1 + bx_2$$
$$cx_1 - dx_2$$

that are *uncorrelated*, by simply referring the surface to its principal

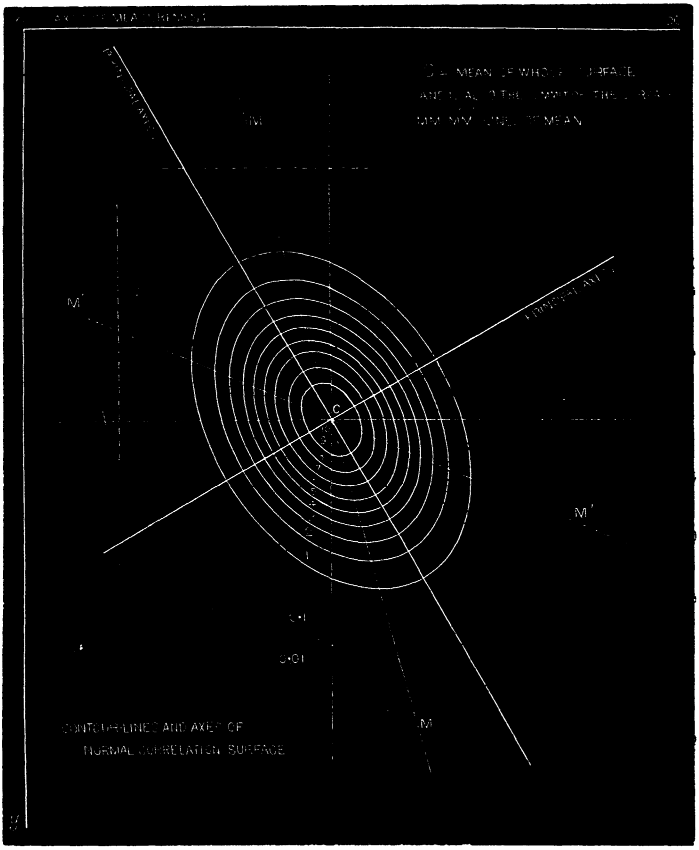
C = MEAN OF WHOLE SURFACE
MM M'M' LINES OF MEAN
PRINCIPAL AXIS
PRINCIPAL AXIS
M
M'
M'
M
C
0·1
0·01
CONTOUR-LINES AND AXES OF
NORMAL CORRELATION SURFACE

Fig. 6.

axes; if in fact the principal axes make an angle θ with the axes of measurement,

$$X_1 = x_1 \cos\theta + x_2 \sin\theta$$
$$X_2 = x_2 \cos\theta - x_1 \sin\theta$$

are uncorrelated. Professor Edgeworth[18] has recently proposed to turn these "uncorrelated functions" to practical account, for convenience in sorting anthropometric measurements. The angle θ is given by

$$\tan 2\theta = \frac{2r\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\ ^{19} \qquad \ldots \qquad \ldots \qquad (9),$$

[18] *Journal of the Royal Statistical Society.* Vol. lix, 1896, p. 534.

[19] This may be proved thus: Multiply together the two transformation equations above. This gives

$$X_1X_2 \doteq (x_2^2 - x_1^2)\frac{\sin 2\theta}{2} + x_1x_2 \cos 2\theta.$$

where $\sigma_1\,\sigma_2$ are the S.D.'s of the correlated variables, and r their correlation coefficient.

The coefficient of correlation with reference to principal axes being zero, and with reference to other axes *something*, there must be some pair of axes for which r is a maximum. r^2 is in fact greatest, or r numerically greatest without regard to sign, for those axes that make 45° with the principal axes.

To complete the numerical data for the normal surface we require the value of z_0, the ordinate of the surface at the mean, or the maximum ordinate. This will be found by integrating out the volume of the surface and equating to the total frequency N. Its value is

$$z_0 = \frac{\mathrm{N}}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \quad \ldots \quad \ldots \quad (10).$$

Suppose Σ_1 and Σ_2 to be the standard deviations of the given distribution *about the principal axes.* If $\sigma_1\;\sigma_2$ are known Σ_1 and Σ_2 may be also taken as known, since by the equations just quoted above for transforming coordinates,

$$\mathrm{X}_1^2 + \mathrm{X}_2^2 = x_1^2 + x_2^2$$

and therefore, summing,

$$\Sigma_1^2 + \Sigma_2^2 = \sigma_1^2 + \sigma_2^2 \quad \ldots \quad \ldots \quad (11)$$

Again, from the converse transformation equations,

$$\sigma_1^2 - \sigma_2^2 = (\Sigma_1^2 - \Sigma_2^2)\cos 2\theta.$$

Squaring, and eliminating θ from (9).

$$\Sigma_1^4 + \Sigma_2^4 - 2\Sigma_1^2\Sigma_2^2 = \sigma_1^4 + \sigma_2^4 - 2\sigma_1^2\sigma_2^2(1-2r^2).$$

Squaring (11) and subtracting this equation from it,

$$\Sigma_1^2\Sigma_2^2 = \sigma_1^2\sigma_2^2(1-r^2)\ldots \quad \ldots \quad \ldots \quad (12).$$

(11) and (12) determine Σ_1, Σ_2 given $\sigma_1\sigma_2$. Hence, referring the surface to its principal axes, the expression for the frequency may be written

$$\mathrm{F} = \frac{\mathrm{N}}{2\pi\Sigma_1\Sigma_2}\,e^{-\frac{1}{2}\left(\frac{\mathrm{X}_1^2}{\Sigma_1^2}+\frac{\mathrm{X}_2^2}{\Sigma_2^2}\right)} \quad \ldots \quad \ldots \quad (13).$$

Any elliptic contour line is given by

$$\frac{\mathrm{X}_1^2}{\Sigma_1^2} + \frac{\mathrm{X}_2^2}{\Sigma_2^2} = c^2,$$

Now sum, remembering

$$\mathrm{S}(\mathrm{X}_1\mathrm{X}_2) = 0,\ \mathrm{S}(x_1x_2) = \mathrm{N}r\sigma_1\sigma_2,$$

and we have

$$0 = (\sigma_2^2 - \sigma_1^2)\frac{\sin 2\theta}{2} + r\sigma_1\sigma_2\cos 2\theta.$$

Therefore

$$\tan 2\theta = \frac{2r\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}.$$

c being a constant. Let us find the total frequency between two adjacent ellipses c, and $c + \delta c$. The area between the two is

$$\pi\Sigma_1\Sigma_2\{(c+\delta c)^2-c^2\}$$
$$=2\pi\Sigma_1\Sigma_2 c\,.\,\delta c\,.$$

Hence the total frequency contained on the element is

$$\mathrm{N}c\,.\,\delta c\,.\,e^{-\frac{1}{2}c^2}$$

Hence the total frequency contained *inside* the ellipse c is

$$\mathrm{N}\int_0^c c\,.\,e^{-\frac{1}{2}c^2}dc=\mathrm{N}\,{}_0^c\!\left[-e^{-\frac{1}{2}c^2}\right]$$

or

$$\mathrm{N}\left(1-e^{-\frac{1}{2}c^2}\right).$$

Hence the probability that an observation lie *outside* the ellipse c is

$$e^{-\frac{1}{2}c^2},$$

a quantity that is given by any of the ordinary tables of the exponential function, or the small table below.[20] This theorem is due to Bravais, *loc. cit.*, on p. 839, note 17.

c.	$e^{-\frac{1}{2}c^2}$.	c.	$e^{-\frac{1}{2}c^2}$.
0·0	1	3·2	0·0060
0·2	0·9802	3·4	0·0031
0·4	0·9231	3·6	0·0015
0·6	0·8353	3·8	0·00073
0·8	0·7262	4·0	0·00034
1·0	0·6065	4·2	0·00015
1·2	0·4868	4·4	0·000062
1·4	0·3753	4·6	0·000028
1·6	0·2780	4·8	0·000011
1·8	0·1979	5·0	0·0000037
2·0	0·1353	5·2	0·0000013
2·2	0·0889	5·4	0·00000047
2·4	0·0561	5·6	0·00000015
2·6	0·0340	5·8	0·000000050
2·8	0·0198	6·0	0·000000015
3·0	0·0111		

The ellipse $c=1$, whose axes are Σ_1 and Σ_2, may be called the "standard ellipse." It very nearly coincides with contour line 6 of the plate, since

$$e^{-\frac{1}{2}}=\cdot6053\ldots$$

The ellipse for which

$$e^{-\frac{1}{2}c^2}=0\cdot5$$
$$c=1\cdot17741\ldots$$

may be called (by analogy from the term "probable error") the "probable ellipse," or better the quartile ellipse. It is ellipse 5

[20] Extended from a set of tables issued by Professor Karl Pearson in 1893 for his Gresham Lectures on Chance.

of the figure. In the diagram of figure (6) in fact 10 per cent. of the whole frequency lies outside ellipse 1, 20 per cent. outside ellipse 2, and so on.

Only about 1 per cent. of the whole frequency lies outside the ellipse $c=3$, for which

$$e^{-\frac{1}{2}c^2}=0{\cdot}0111\ .\ .\ .$$

Consequently this ellipse will, roughly speaking, cover the whole range of small series of observations.

Various Values of the Correlation Coefficient.

All through the present section we have used the value of r obtained by the method of least squares in Sec. I, viz. :—

$$r = \frac{\mathrm{S}(xy)}{\mathrm{N}\sigma_1\sigma_2}, \quad \ldots \quad \ldots \quad \ldots \quad (\mathrm{A})$$

but it is evident that if the correlation surface be strictly normal r may be evaluated in many other ways; *e.g.*, the S.D. of an array, σ'_1, may be calculated and compared with the S.D. σ_1 of the totality of such arrays, when we have

$$\sigma'^2_1 = \sigma_1^2(1 - r^2),$$

or,

$$r^2 = 1 - \frac{\sigma'^2_1}{\sigma_1^2} \quad \ldots \quad \ldots \quad \ldots \quad (\mathrm{B}),$$

Or, again, the mean of an array of type x may be found, whence, if the deviation of its mean be $\bar{y}$,

$$\bar{y} = r\,\frac{\sigma_2}{\sigma_1}x\,.$$

$$r = \frac{\sigma_1}{\sigma_2} \times \frac{\bar{y}}{x} \quad \ldots \quad \ldots \quad \ldots \quad (\mathrm{C}).$$

The question may consequently be put, what is the best value to give r, *i.e.*, what value of r will make the probability of an observed set of deviations a maximum? The value (A) of r was given by Bravais[21] in 1846; it has recently been shown to be the best value by Professor Pearson.[22] Thus, let $(x_1\,y_1)$, $(x_2\,y_2)$, &c., be a series of observed deviations. The chance of one pair occurring varies as

$$\frac{1}{\sqrt{1-r^2}}e^{-\frac{1}{2(1-r^2)}\left\{\frac{x_1^2}{\sigma_1^2}+\frac{y_1^2}{\sigma_2^2}-\frac{2x_1y_1r}{\sigma_1\sigma_2}\right\}}$$

or the chance of the whole system occurring varies as the product of all such quantities, that is

$$\frac{1}{(1-r^2)^{n/2}}e^{-\frac{1}{2(1-r^2)}\left\{\frac{\mathrm{S}(x^2)}{\sigma_1^2}+\frac{\mathrm{S}(y^2)}{\sigma_2^2}-\frac{2\mathrm{S}(xy)\,.\,r}{\sigma_1\sigma_2}\right\}}$$

[21] "Mémoires présentés par divers Savants," II$^{\mathrm{e}}$ Série, ix, 1846.

[22] "Phil. Trans." (A), vol. clxxxvii, p. 253, 1896.

Remembering

$$\sigma_1^2 = \mathrm{S}(x^2)/n, \qquad \sigma_2^2 = \mathrm{S}(y^2)/n,$$

and writing

$$\frac{\mathrm{S}(xy)}{n\sigma_1\sigma_2} = p,$$

this quantity becomes

$$\frac{1}{(1-r^2)^{n/2}} e^{-n\frac{1-pr}{1-r^2}}$$

The differential of the logarithm of this with regard to r is

$$\frac{n(1+r^2)(p-r)}{(1-r^2)^2},$$

and equating it to zero gives at once

$$r = p$$

as the best value of r, that is Bravais's value is the best.

Probable Errors.

Professor Pearson next proceeds, in the paper cited, to evaluate the probable error of r, and arrives at the result

$$\text{probable error of } r = \cdot 674489\ldots\frac{1-r^2}{\sqrt{n(1+r^2)}}.$$

This value, he has recently concluded, is slightly in error, being only the probable error of an array of r's corresponding to given values of, σ_1 σ_2. The corrected value is[23]

$$\text{probable error of } r = \cdot 674489\ldots\frac{1-r^2}{\sqrt{n}}.$$

The difference from the first value given is always small, and vanishes altogether for $r = 0$ and $r = 1$. In Table I at the end will be found the values of $(1 - r^2)$ for all values of r increasing by hundredths, and in Table II is given a table of the actual probable errors for values of r rising by tenths, and values of n 25, 50, 75, 100, 200, &c., to 1,000 From the table can be seen by inspection the approximate probable error of any correlation coefficient where the number of observations does not exceed 1,000.

In the case of normal variation the probable error of the mean is

$$\cdot 674489\ldots\frac{\sigma}{\sqrt{n}}.$$

and the probable error of the standard deviation

$$\cdot 674489\ldots\frac{\sigma}{\sqrt{2n}}$$

These are all the probable errors concerned.

[23] "On the probable errors of Frequency Constants," &c., by Prof. Karl Pearson, F.R.S., and L. N. G. Filon, B.A., Abstract in Proceedings, Royal Society. Read 18th October, 1897. An important paper by Mr. W. F. Sheppard bearing partly on probable errors was also read at the same meeting.

(2.) Case of Three Variables.

We contented ourselves, in the preceding case, with giving a deduction of the general expression for the normal frequency surface based solely on a transformation of coordinates from the case of "no correlation" ($r = 0$). This proceeding was of course a suggestive artifice and not a strict proof; there may not always be two directions of independent variation such as were practically postulated. The same method may however be employed to *suggest* the "normal surface" for correlation of n variables; it will lead simply to the general form

$$\text{frequency} = \text{const.} \times e^{-(\text{general quadratic function of the variables})}.$$

In the case of three variables, for example, the exponent is (dropping the minus sign),

$$g_1x_1^2 + g_2x_2^2 + g_3x_3^2 - 2h_1x_2x_3 - 2h_2x_1x_3 - 2h_3x_1x_2.$$

The g's and h's can be determined as in the last case from the work in Part I.

Taking special types of x_2 and x_3, the standard deviations of the x_1-array is $1/\sqrt{2g_1}$, and so on for the others. Hence writing for brevity

$$\Delta = 1 - r_{12}^2 - r_{23}^2 - r_{13}^2 + 2r_{12}r_{23}r_{13},$$

we have

$$\left.\begin{aligned} g_1 &= \frac{1 - r_{23}^2}{2\Delta\sigma_1^2} \\ g_2 &= \frac{1 - r_{13}^2}{2\Delta\sigma_2^2} \\ g_3 &= \frac{1 - r_{12}^2}{2\Delta\sigma_3^2} \end{aligned}\right\}$$

To find the values of the h's, let t_2 and t_3 be the types assigned to x_2 and x_3. Then the distribution of the x_1-array is of the form

$$\begin{aligned} z &= z_0' e^{-g_1\left\{x_1^2 - \frac{2}{g_1}x_1(h_2t_3 + h_3t_2)\right\}} \\ &= z_0'' e^{-g_1\left(x_1 - \frac{h_2t_3 + h_3t_2}{g_1}\right)^2} \end{aligned}$$

The regression is, as before, truly linear, the mean of the x_1-array differing from the mean of all x_1's by the quantity

$$\frac{h_2t_3 + h_3t_2}{g_1}.$$

Hence, putting $t_2 = 0$, we have at once (part I, (2), p. 832)

$$\frac{h_2}{g_1} = b_{13} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \cdot \frac{\sigma_1}{\sigma_3}.$$

Therefore

$$\begin{aligned} h_2 &= \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \cdot \frac{1 - r_{23}^2}{2\Delta\sigma_1^2} \cdot \frac{\sigma_1}{\sigma_3} \\ &= \frac{r_{13} - r_{12}r_{23}}{2\Delta\sigma_1\sigma_3} \end{aligned}$$

Similarly

$$h_3 = \frac{r_{12} - r_{13}r_{23}}{2\Delta\sigma_1\sigma_2}$$

$$h_1 = \frac{r_{23} - r_{12}r_{13}}{2\Delta\sigma_2\sigma_3}$$

Finally, the constant outside the exponential, or the maximum ordinate, will be required. Its value is found as before by integrating out the whole volume of the surface and equating to the total frequency N, which gives

$$\frac{\mathrm{N}}{(2\pi)^{3/2}\sigma_1\sigma_2\sigma_3\sqrt{\Delta}}.$$

In the case of two variables the whole frequency-distribution could be represented by a family of similar and similarly situated ellipses; in the present case we must take a family of similar and similarly situated ellipsoids. Just as in the former case we could refer the *ellipses* to their principal axes, so in this case we can refer the *ellipsoids* to their principal axes. Since the product-terms in the exponent must vanish for these axes, we must have

$$r_{23} = r_{12}r_{13},$$
$$r_{13} = r_{12}r_{23},$$
$$r_{12} = r_{13}r_{23}.$$

From the first two of these equations

$$r_{13}r_{23} = r_{12}^2 r_{13}r_{23},$$

and from the last

$$r_{12} = r_{13}r_{23},$$

which can only be satisfied by

$$r_{12} = r_{13} = r_{23}\,0 \text{ or } 1.$$

But the first is the only admissible solution (as for the second Δ becomes zero), so that there is no correlation round the principal axes. Hence three linear functions of x_1, x_2, x_3, can always be formed between which the correlation is zero.

If the surface be referred to its principal axes the equation to it is

$$\mathrm{F} = \frac{\mathrm{N}}{(2\pi)^{3/2}\Sigma_1\Sigma_2\Sigma_3}\, e^{-\frac{1}{2}\left\{\frac{\mathrm{X}_1^2}{\Sigma_1^2}+\frac{\mathrm{X}_2^2}{\Sigma_2^2}+\frac{\mathrm{X}_3^2}{\Sigma_3^3}\right\}}$$

The volume of the ellipsoid

$$\frac{\mathrm{X}_1^2}{\Sigma_1^2}+\frac{\mathrm{X}_2^2}{\Sigma_2^2}+\frac{\mathrm{X}_3^2}{\Sigma_3^2}=c^2$$

is

$$\frac{4}{3}\pi c^3\Sigma_1\Sigma_2\Sigma_3.$$

Therefore the volume between ellipsoids c and $c+\delta c$ is

$$4\pi\,.\,\Sigma_1\Sigma_2\Sigma_3c^2\delta c.$$

3 K 2

Hence the total frequency contained between these two ellipsoids is

$$N\sqrt{\frac{2}{\pi}}c^2e^{-\frac{1}{2}c^2}\delta c$$

Therefore the chance that an observation falls inside the ellipsoid c is

$$\sqrt{\frac{2}{\pi}}\int_0^c c^2e^{-\frac{1}{2}c^2}dc$$

$$= \sqrt{\frac{2}{\pi}}\left\{-ce^{-\frac{1}{2}c^2}+\int_0^c e^{-\frac{1}{2}c^2}dc\right\}$$

$$= \Phi, \text{ say.}$$

Of the two terms in the bracket, the first can be readily calculated from our previous table of the exponential function (p. 845), and the second can be obtained from the tables of the probability integral given in most text-books.[24] The small table below gives the values of the chance $(1 - \Phi)$ for a certain range of the values of c, *i.e.*, the chance that the observation shall lie *outside* the ellipsoid :—[25]

c.	$1 - \Phi$.	c.	$1 - \Phi$.
0·2	0·9979	2·2	0·1839
0·4	0·9838	2·4	0·1238
0·6	0·9484	2·6	0·0799
0 8	0·8873	2·8	0·0493
1·0	0·8012	3·0	0·0293
1·2	0·6962	3 2	0·0167
1·4	0·5807	3·4	0·0091
1·6	0·4645	3·6	0 0046
1·8	0·3561	3·8	0·0024
2·0	0·2614	4·0	0·0011

Since the regression is truly linear, all the expressions for net regressions and net coefficients of correlation (given in Part I) are immediately and definitely interpretable without any limitations such as are necessary in the general case. In the present case of three variables, for example, the coefficient of correlation between x_1 and x_2 *in any group for which x_3 is constant* is

$$\frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13})^2(1 - r_{23}^2)}}$$

[24] The majority of such tables appear to be derived from Kramp's "Analyse "des Réfractions," Strasburg, 1798. There are convenient tables in Galloway's "Probability" (Edinburgh, 1839), and in De Morgan's "Encyclopædia Metro-"politana," "Treatise on the Theory of Probabilities" (1837).

[25] I cannot see any error in the values here given, but they differ from some tabulated by Czuber, *Theorie der Beobachtungsfehler*, 1891, p. 404. These theorems on the chance of an observation lying within a given ellipse or ellipsoid are due to Bravais, *loc. cit.*

as given on p. 833, and similar expressions hold good for the coefficients of correlation in other sub-groups.

It is hardly necessary to proceed to deal in detail with the cases of four or more variables, as the preceding examples of two and three variables are sufficient to show the general character of normal correlation. Consideration of the exponent as in the previous cases shows that the properties must always hold that (1) the standard deviations of all x_n-arrays are equal, (2) the regression is strictly linear, (3) the frequency is constant over a quadratic surface, (4) the net coefficients between all sub-groups of $x_n\, x_m$, corresponding to given types of the remaining variables, are equal.

In the majority of cases of economic interest the correlation is not normal, *e.g.*, in cases involving variations of pauperism, prices, rateable values, or quantities varying with age such as marriage-rates. In most economic examples in fact the amount of the variation is a considerable fraction of the mean, and it is a necessary condition for the occurrence of normal variation that deviations should only be small compared with the mean; otherwise in fact an absolute physical limit might be close to the mean on one side, a limit for which normal variation cannot allow. But in some instances of great importance, *e.g.*, anthropometry, the normal law holds very approximately, and in any case is so frequently useful for obtaining approximations and theoretical results (in probable errors and so forth), that the statistician should be familiar with it.

For the exposition of the normal theory in the general case, *vide* Edgeworth, "Correlated Averages," "Phil. Mag.," 1892, vol. xxxiv, p. 190, or Pearson, "Regression, Heredity, &c., "Phil. Trans." (A), 1896, vol. clxxxvii, p. 253.

APPENDIX.

TABLE I.—*Giving the Values of* $1 - r^2$ *and* $\sqrt{1 - r^2}$ *and their first Differencсs, for all Values of r between* 0 *and* 1, *proceeding by Hundredths.*

r.	$1 - r^2$.	Δ_1.	$\sqrt{1 - r^2}$.	Δ_1.
0·01	0·9999	3	0·99995	15
0·02	0·9996	5	0·99980	25
0·03	0·9991	7	0·99955	35
0·04	0·9984	9	0·99920	45
0·05	0·9975	11	0·99875	55
0·06	0·9464	13	0·99820	65
0·07	0·9951	15	0·99755	75
0·08	0·9936	17	0·99679	85
0·09	0·9919	19	0·99594	95
0·10	0·9900	21	0·99499	106
0·11	0·9879	23	0·99393	116
0·12	0·9856	25	0·99277	126
0·13	0·9831	27	0·99151	136
0·14	0·9804	29	0·99015	146
0·15	0·9775	31	0·98869	157
0·16	0·9744	33	0·98712	167
0·17	0·9711	35	0·98544	177
0·18	0·9676	37	0·98367	188
0·19	0·9639	39	0·98178	199
0·20	0·9600	41	0·97980	209
0·21	0·9559	43	0·97770	220
0·22	0·9516	45	0·97550	231
0 23	0·9471	47	0·97319	242
0·24	0·9424	49	0·97077	253
0·25	0·9375	51	0·96825	264
0·26	0·9324	53	0·96561	275
0·27	0·9271	55	0·96286	286
0·28	0·9216	57	0·96000	297
0·29	0·9159	59	0·95703	309
0·30	0·9100	61	0·95394	320
0·31	0·9039	63	0·95074	332
0·32	0·8976	65	0·94742	344
0·33	0·8911	67	0·94398	356
0·34	0·8844	69	0·94043	368
0·35	0·8775	71	0·93675	380
0·36	0·8704	73	0·93295	392
0·37	0·8631	75	0·02903	405
0·38	0·8556	77	0 92498	417
0·39	0·8479	79	0·92081	430
0·40	0·8400	81	0·91651	443
0·41	0·8319	83	0·91209	456
0·42	0·8236	85	0·90752	470
0·43	0·8151	87	0·90283	483
0·44	0·8064	89	0·89800	497

TABLE I.—*Giving the Values of* $1 - r^2$ *and* $\sqrt{1 - r^2}$—*Contd.*

r.	$1 - r^2$.	Δ_1.	$\sqrt{1 - r^2}$.	Δ_1.
0·45	0·7975	91	0·89303	511
0 46	0·7884	93	0·88792	525
0·47	0·7791	95	0·88267	540
0·48	0·7696	97	0·87727	555
0·49	0·7599	99	0·87172	570
0·50	0·7500	101	0·86603	585
0·51	0·7399	103	0·86017	601
0·52	0·7296	105	0·85417	617
0·53	0·7191	107	0·84800	633
0·54	0·7084	109	0·84167	650
0·55	0·6975	111	0·83516	667
0·56	0·6864	113	0·82849	685
0·57	0·6751	115	0·82164	703
0·58	0·6636	117	0·81462	721
0·59	0·6519	119	0·80740	740
0·60	0·6400	121	0·80000	760
0·61	0·6279	123	0·79240	780
0·62	0·6156	125	0·78460	801
0·63	0·6031	127	0·77660	822
0·64	0·5904	129	0·76837	844
0·65	0·5775	131	0·75993	867
0·66	0·5644	133	0·75127	890
0·67	0·5511	135	0·74236	915
0·68	0·5376	137	0·73321	940
0·69	0·5239	139	0·72381	967
0·70	0·5100	141	0·71414	994
0·71	0·4959	143	0·70420	1023
0·72	0·4816	145	0·69397	1053
0·73	0·4671	147	0·68345	1084
0·74	0·4524	149	0·67261	1117
0·75	0·4375	151	0·66144	1151
0·76	0·4224	153	0·64992	1188
0·77	0·4071	155	0·63804	1226
0·78	0·3916	157	0·62578	1267
0·79	0·3759	159	0·61311	1311
0·80	0·3600	161	0·60000	1357
0·81	0·3439	163	0·58643	1407
0·82	0·3276	165	0·57236	1460
0·83	0·3111	167	0·55776	1518
0·84	0·2944	169	0 54259	1580
0·85	0·2775	171	0·52678	1649
0·86	0·2604	173	0·51029	1724
0·87	0·2431	175	0·49305	1808
0·88	0·2256	177	0·47497	1901
0·89	0·2079	179	0·45596	2007
0·90	0·1900	181	0·43589	2128
0·91	0·1719	183	0·41461	2269
0·92	0·1536	185	0·39192	2436
0·93	0·1351	187	0·36756	2639
0·94	0·1164	189	0·34117	2892
0·95	0·0975	191	0·31225	3225
0·96	0·0784	193	0·28000	3690
0·97	0·0591	195	0·24310	4411
0·98	0·0396	197	0·19900	5793
0·99	0·0199	199	0·14107	14107
1·00	0·0000	—	0·00000	—

TABLE II.—*Giving the Probable Errors of the Correlation Coefficient r for various Numbers of Observations.*

Number of Observations.	Correlation Coefficient *r*.										
	0·0.	0·1.	0·2.	0·3.	0·4.	0·5.	0·6.	0·7.	0·8.	0·9.	1·0.
25....	0·1349	0·1335	0·1295	0·1228	0·1133	0·1012	0·0863	0·0688	0·0486	0·0256	o
50....	0·0954	0·0944	0·0916	0·0868	0 0801	0·0715	0·0610	0·0486	0·0343	0·0181	o
75....	0·0779	0·0771	0·0748	0·0709	0·0654	0·0584	0·0498	0·0397	0·0280	0·0148	o
100....	0·0674	0·0668	0·0648	0·0614	0·0567	0·0506	0·0432	0·0344	0·0243	0·0128	o
200....	0·0478	0·0473	0·0459	0·0435	0·0402	0·0359	0·0306	0·0244	0·0172	0·0091	o
300....	0·0389	0·0386	0·0374	0·0354	0·0327	0·0292	0·0249	0·0199	0·0140	0·0074	o
400....	0·0337	0·0334	0·0324	0·0307	0·0283	0·0253	0·0216	0·0172	0·0121	0·0064	o
500....	0·0302	0·0299	0·0290	0·0274	0·0253	0·0226	0·0193	0·0154	0·0109	0·0057	o
600....	0·0275	0·0273	0·0264	0·0251	0·0231	0·0207	0·0176	0·0140	0·0099	0·0052	o
700....	0·0255	0·0252	0·0245	0·0232	0·0214	0·0191	0·0163	0·0130	0·0092	0·0048	o
800....	0·0238	0·0236	0·0229	0·0217	0·0200	0·0179	0·0153	0·0122	0·0086	0·0045	o
900....	0·0225	0·0223	0·0216	0·0205	0·0189	0·0169	0·0144	0·0115	0·0081	0·0043	o
1,000....	0·0213	0·0211	0·0205	0·0194	0·0179	0·0160	0·0137	0·0109	0·0077	0·0041	o