A DEMONSTRATION OF THE COEFFICIENT OF CORRELATION, FOR ELEMENTARY STUDENTS OF PLANT BREEDING.

By HERBERT F. ROBERTS,
Kansas State Agricultural College.

It is often profitable, in teaching the subject of genetics to elementary students, to make tolerably free use of biometric methods, especially in the earlier stages of the course. The writer has found, in his own experience at least, that the careful marshalling and rigorous analysis of data required in the calculation of the chief constants, is of value in training students in precise and accurate methods of work.

It is the purpose of the present paper to discuss a method of presenting the coefficient of correlation. Although this coefficient is little used in inheritance studies today, it nevertheless has a distinct and practical value in demonstrating the general trend of evolution within a group. It deals, of course, only with large numbers and general tendencies, and takes no account necessarily of the genetic conditions existing in any particular case. Within a genetically related population, however, it is able to forecast the incidence of group evolution with respect to the characters in question. It shows, on the average, in a related population, the extent or degree of interdependence of any two given phenotypic (i. e., visible somatic) characters. However, for purposes of breeding, which takes account, not of populations, but of genotypes, the coefficient, as ordinarily found, may have no essential meaning.

Where a high coefficient exists between two correlatives, there must, of course, be a physiological reason involving their common presence in the organism, although the actual proof of the genetic linkage of factors, remains to be determined by the usual process of factorial analysis of genotypes. Nevertheless it is true, that for work in plant improvement to be carried on by farmers, the knowledge of the existence of a high percentage of correlation between two somatic characters may be of substantial practical benefit in commencing the practice of selection. Even though the coefficient is merely of a statistical nature, and has no necessarily direct relation to inheritance, it nevertheless may frequently point the way to the existence of a possible linkage of characters, determinable through genetic analysis.

Pearl (3, p. 15), sums the matter up satisfactorily in the statement:

Heredity is not the sole cause which can lead statistically to a significant correlation between parent and offspring. (p. 69.) The coefficient merely indicates the existence of association, but does not state for us upon what basis that association rests. . . . It may be inheritance. . . . it may be local environmental differences, or it may be anything whatever, so far as the correlation method *per se* helps us.

However, the determination of the existence of such association, in the case of somatic characters, in genetically related individuals, may serve as a reconnaissance basis for actual selection of genotypes for breeding purposes.

Quoting Pearl again (p. 70):

There can, of course be no valid objection to the study, in and for itself, of the correlation existing between genetically related individuals in respect of somatic characters. Such studies may, indeed, for one reason or another, have a high intrinsic interest . . . but, in dealing with such correlatives, one should always keep closely in mind that he is not dealing directly and primarily with phenomena of inheritance, but only indirectly and secondarily. . . . The only way to determine whether the "differences" indicated by the correlative method are really heritable is to apply the method of individual pedigree analysis to the complex heterogeneous material of the table. If it is possible to isolate and propagate distinct genotypes from the material, then it may be concluded that the primary basis of differentiation or heterogeneity detected by the correlation coefficient was inheritance.

The coefficient has been recently utilized by Rugg (4, pp. 246, 261), in demonstrating the degree of association, irrespective of causal relationship, existing between such measurable educational data as "ability in mathematics," "ability in languages," "ability in drawing," "ability in shop practice."

For elementary students in genetics, at any rate, a certain amount of work in determining the correlation coefficient, provided its meaning and range of application are made clear, is of value for training in the orderly and systematic arrangement and classification of data, indispensable to future work in breeding.

The graphic method of illustrating the association of variable biological factors supposedly interdependent was first used by Francis Galton in his studies of inheritance of human stature. It seems, however, according to Rugg (p. 350), that Bravais in 1846, was the first to suggest that the degree of association between two factors could be represented in terms of the sum of the products of the deviations of the variates from their respective means. Bravais, however, did not, it appears, express the degree of this relationship in the form of an equation or a coefficient. Pearson, in 1896, furnished the statement of the equation for the regression curve, and of the coefficient of correlation.

For the purpose of illustrating a method of demonstrating

the nature of the coefficient of correlation to students, which the writer has found satisfactory, a concrete case may be taken, of the correlation between length of ears of corn in inches, and their weight in ounces, as given by Davenport (2, p. 458). For convenience, the table, somewhat modified, is here given

y	X	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
		8.65	7.65	6.65	5.65	4.65	3.65	2.65	1.65	0.65	0.35	1.35	2.35	3.35	4.35	5.35	6.35	7.35	8.35	9.35	10.35	
3.0	4.85	1	2		1																	4
3.5	4.35		4		1																	5
4.0	3.85	3	5	5	1																	14
4.5	3.35		6	5	4			1														16
5.0	2.85		2	4	7	2	4															19
5.5	2.35		2	9	15	14	8	4	1													53
6.0	1.85		1	2	12	16	13	13	6	1												64
6.5	1.35			1	6	11	26	11	8	6	1											70
7.0	0.85			1	2	2	12	18	12	12	11	4	1									75
7.5	0.35				1	2	4	20	12	13	21	11	6	6	1	1						98
8.0	0.15						3	7	19	25	17	22	17	3	1							113
8.5	0.65						1	1	12	9	23	30	26	26	5	1						134
9.0	1.15								1	7	10	23	35	26	24	12	1	2	1			142
9.5	1.65									1	4	14	19	29	17	10	1	3		1	1	100
10.0	2.15									1	1	3	8	18	10	6	4	2				53
10.5	2.65												2	3	6	7	2	5	1			26
11.0	3.15													1	1				2	1		5
11.5	3.65																	1				1
		4	22	27	50	47	71	75	71	75	88	107	114	112	65	37	8	13	4	2	1	993

TABLE I—CORRELATION TABLE BETWEEN LENGTH OF EARS OF CORN IN INCHES (SUBJECT), AND WEIGHT OF GRAIN (RELATIVE), IN OUNCES.

First (left hand) vertical column, length of ears in inches; second vertical column, deviations from the mean. Upper horizontal array, weight of grain in ounces; lower horizontal array, deviations from the mean. Constants of the table: A, $Y = 7.85 \pm 0.03$. A, $X = 10.65 \pm 0.08$, $\sigma\ Y = 1.57 \pm 0.02$; σ. $X = 3.63 \pm 0.05$, $r = 0.87 \pm 0.005$.

The constants are as follows: $_{\sigma}x = 3.63$ oz.; $_{\sigma}y = 1.57$ in. $r = .87$. In order to let the student understand the significance of the correlation table and the coefficient, the means are now found for the vertical and the horizontal columns, and these are plotted as in Figure 1. We have here the means of the horizontal columns (average weight of ears in ounces), corresponding to the units of difference in length of the ears in inches. Plotting the best fitting line for the moments, we have the graph GH, the equation for which is $y = r\ y/x.\ x = 1.88x$. Correspondingly, we have the means of the vertical columns (average lengths of ears in inches), corresponding to the units of difference (ounces) in the weight of the ears. Plotting the

best fitting line for the moments, we have the graph IK, for which the equation is $x = r\, x/y.\, y = 2.01y$. The solution for these equations has sometimes been termed the "regression coefficient."

Following are the data for the moments:

Length of Ears, Inches	Average Wt. of Ears, Ounces	Weight of Ears, Ounces	Average Length of Ears, Inches
4.85	7.40	6.65	4.10
4.35	6.12	7.65	3.57
3.85	7.36	6.65	2.99
3.35	6.52	5.65	2.33
2.85	5.54	4.65	1.81
2.35	5.02	3.65	1.43
1.85	3.99	2.65	1.01
1.35	3.33	1.65	0.69
0.85	2.10	0.65	0.62
0.35	1.80	0.35	0.63
0.15	1.47	1.35	0.80
0.65	1.90	2.35	0.70
1.15	2.87	3.35	1.44
1.65	3.38	4.35	1.53
2.15	3.77	5.35	1.69
2.65	5.25	6.35	2.08
3.15	6.75	7.35	1.90
3.65	7.35	8.35	2.52
		9.35	2.40
		10.35	3.65

The student should have carefully explained to him the difference between the regression lines or graphs, GH (regression of x on y), and IK (regression of y on x), on the one hand, and the graph of the correlation coefficient, EF, on the other. It is well to remind the student that in the three equations,

$y = 1.88x$ (GH),

$x = 2.01y$ (IK),

the coefficients of regression, on the one hand, and

$x = .87y$ (EF),

the coefficient of correlation on the other, are to be held in mind as representing two distinct concepts.

The ratio $x/y = .87$ is taken as the tangent of the angle that the graph EF, of the correlation coefficient, makes with the horizontal. Instead of leaving the student with the simple statement that with perfect correlation, the coefficient is 1.0, and that tan. $\alpha = 1$, when $\alpha = 45$, and that then we have CD as the correlation curve, it is well to develop the idea further, upon the basis of the graphical meaning of the tangent in reference to the coefficient of correlation. The value of the tangent of the angle α at 45° is 1. (Figure 2.) The value of the correlation coefficient is always the value of the tangent of the angle which the correlation line forms with the axis of X, where it

passes through the point of intersection of the axes X and Y. In the figure, CD is the correlation line, where the value of the correlation coefficient is 1; i. e., where the angle COX or α is 45°, and the tangent of that angle is 1. GH is the correlation line, for example, when the coefficient of correlation (r), is .74.

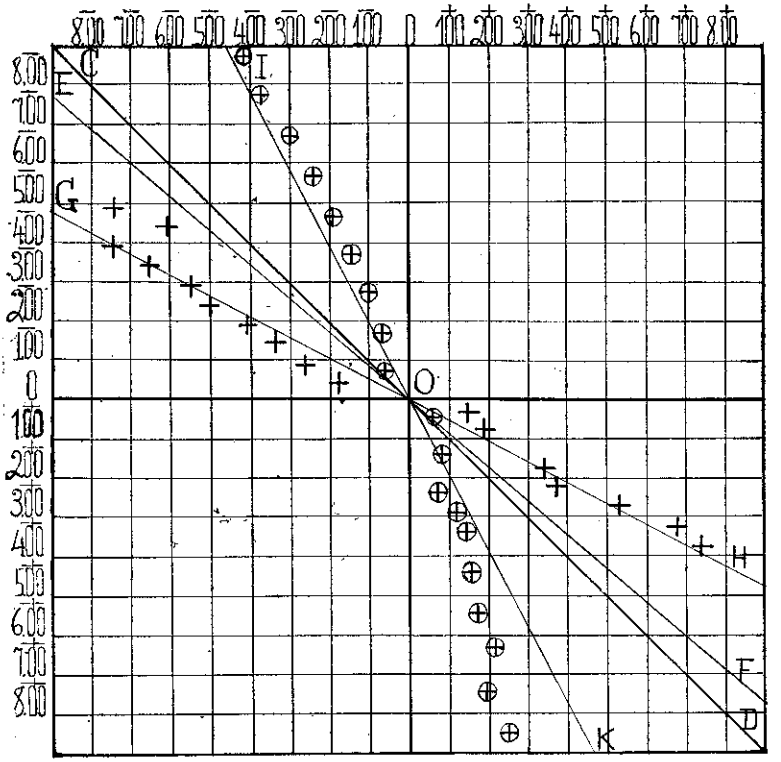

FIGURE 1.
Regression table, showing regression of X on Y (GH), or Y on X (IK) and correlation graph, EF.

Referring to a trigonometric table of natural tangents, we find that .74 is the tangent of the angle 36° 30′. We accordingly take GO to form the angle γ, or GOX, of 36° 30′.

In the case in hand, the correlation coefficient of length to weight of ears of corn, .87, is practically identical with the tangent .8703 of the angle 41° 2′. We accordingly draw the line EF through O (Figure 2), forming the angle EOX, or β, of 41° 2′ with the axis X. This, then, represents the line or graph corresponding to the correlation coefficient, .87.

The relation of the tangent to the correlation coefficient, can, however, be made clearer to students by presenting the

subject from the standpoint of the numerical value of the tangent as a line; i. e., as the line tangent.

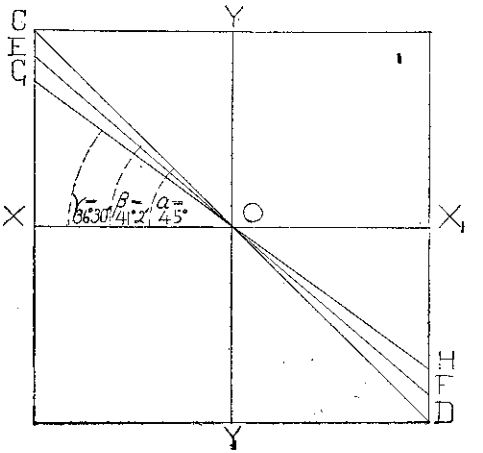

FIGURE 2.
Graph for $r=0.87$ (E F), and supposed case for $r=0.74$ (G H), for purpose of demonstrating the angle for the tangents represented by r.

About O, describe a circle with a radius equal to 1. Such a circle is called a unit circle. (Figure 3.) Let $X-X_1$ be a diameter of such a circle, and $Y-Y_1$ be another diameter erected perpendicular to $X-X_1$. Let CX be a geometrical tangent to the circle O, drawn perpendicular to the diameter

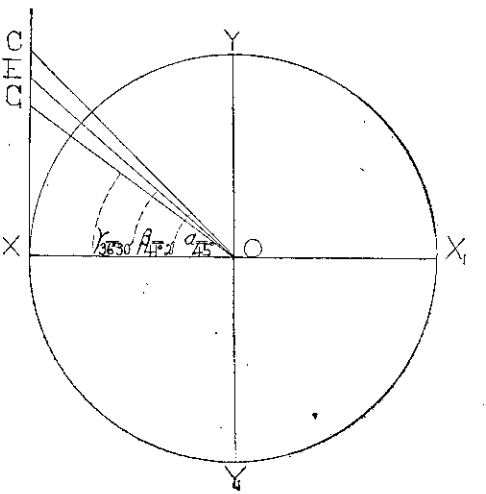

FIGURE 3.
Unit circle, for purpose of demonstrating the numerical or line value of the tangent CX to the angle α (45°), and the value of the tangents for the angles β and γ, where $r=0.87$ (E O), and 0.74 (G O).

$X-X_1$, and CO a line bisecting the angle YOX, and intersecting the tangent CX at C. Then the triangle CXO will be a right-angled triangle, of which the angles COX and OCX, each by construction, equals 45° and the sides CX and OX are equal. But, by construction, the side OX equals 1. The line CX therefore also equals 1, and since the line CX is tangent to the circle at X, then the numerical value of such a tangent, limited by the point of intersection of a line forming an angle of 45° with the horizontal axis OX, is 1.

From O, draw the lines OE and OG, intersecting CX at E and G, and forming the angles 41° 2′ (β), and 36° 30′ (γ), respectively. Now, by simple mensuration, the length of the line CX is found to be 100 mm., and of the lines EX and GX, 87 and 74 mm., respectively. Since CX by construction equals 1, then the values of EX and GX as percentages of 1, are found by taking EX/CX and GX/CX = 87/100 and 74/100, or .87 and .74, respectively.

We thus have the idea of the tangent presented from the geometrical standpoint, and represented as the numerical value of the line tangent in terms of percentages of its limiting value, and this is expressive of the value of the correlation coefficient, in that it states the value of the tangents of the angles which correlation graphs such as EO, GO, etc., form with the horizontal. Since the correlation coefficient is simply the coefficient of y in the equation $x = ay$ the numerical value of the coefficient determines the slope of the correlation curve.

The same idea may then be presented to the student in the usual language of trigonometry with greater facility. Since CX and OX are equal by construction, and the numerical value of the tangent, expressed as a line, being, let us say, EX/CX or 87/100 = .87, it will also be true that EX/OX, or the opposite leg, /adjacent leg, of the triangle COX, = .87 by mensuration. By making the angle COX or α, an angle from 45° to 0° we have the line CX becoming constantly shorter than the line OX, and therefore the value of CX/OX becoming continuously less and less, so that when the angle α becomes as small as 1°, the value of the tangent, which began as 1.0000, has diminished to 0.0003, and finally when $\alpha = 0°$, CX/OX = 0.

By thus treating the tangent from the geometrical standpoint, its value and its relation to the correlation curve are made clearer than by merely referring to it as a trigonometrical function. The graphical exposition and meaning of the coefficient

should be made as clear and definite as possible, where the plotting of the correlation tables forms a part of the laboratory work.

The fact of the correlation coefficient being simply the coefficient of the equation $x = ay$, should be impressed by an example for which the following may serve as a means of demonstration:

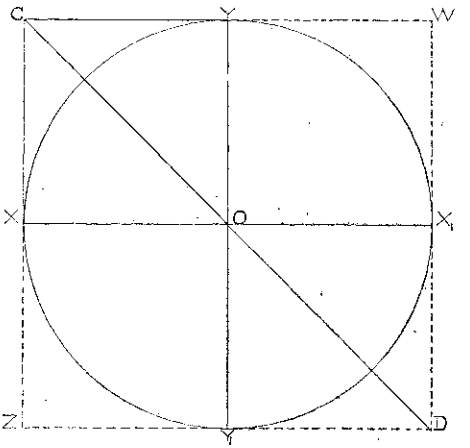

FIGURE 4.
Demonstration of principle of the construction of the regression table upon the basis of the unit circle.

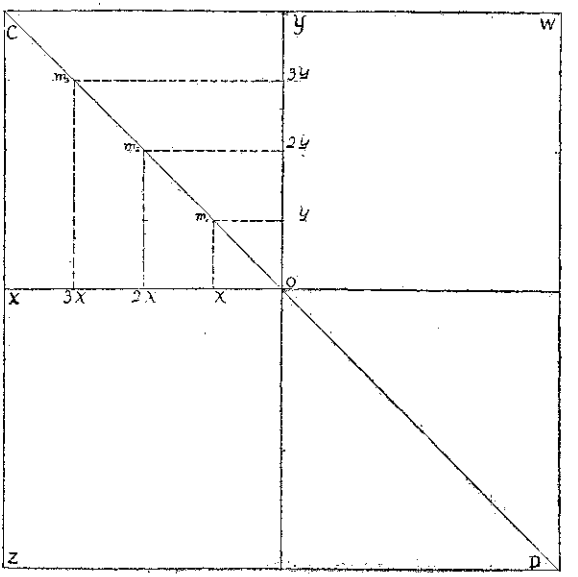

FIGURE 5.
Regression table showing m_1, m_2, and m_3; plotting of moments of curve, when tan. α or $r = 1$, i. e., when $X = Y$.

In the circle O (Figure 4), we may regard the diameters $X-X_1$ and $Y-Y_1$, as a system of coordinate axes, independent of the circle, and which we may call X and Y, respectively. Let us take the axis Y parallel with, and equal to, CX, joining C and Y parallel to XO, likewise completing the rectangle CWDZ, with the line CO projected to intersect WD and ZD at D. The side CX, therefore, of the triangle COX, is by construction, parallel to and equal to the axis Y. Now we have tan. α = CX/OX, or, transposing, CX = OX tan. α. But, substituting axis Y for CX, and axis X for OX, we have $y = x$ tan. α. Applying this equation to the location of points along the line CO, where tan. $\alpha = 1$, we have $y = x$, $2y = 2x$, $3y = 3x$, etc. (Figure 5.)

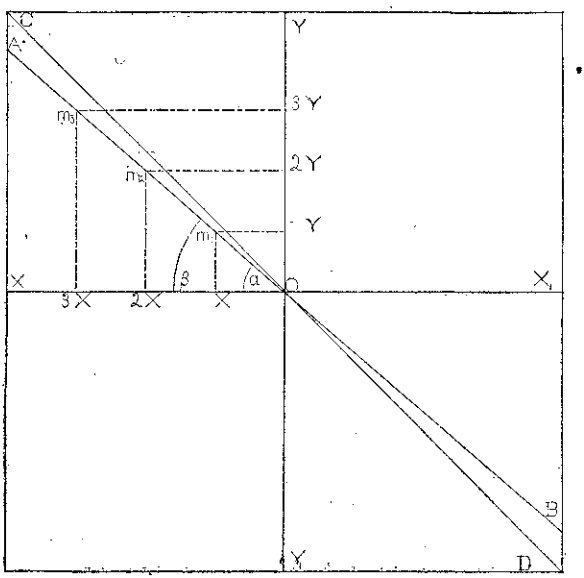

FIGURE 6.
Regression table, showing plotting of moments of curve, m_1, m_2 and m_3 when $x = 0.87\ y$, i. e., when tan. α or $r = 0.87$.

Now, if we plot equal distances on Y and X in any unit whatsoever, and erect perpendicular lines on X from x, $2x$, and $3x$, and on Y from y, $2y$, and $3y$, they will intersect at the points m_1, m_2, and m_3. If we continue to plot units on the axis Y until we reach Y, and on the axis X until we reach X, we find the perpendiculars erected at Y and X intersecting at C. We therefore have the points C . . . . m_3, m_2, m_1, on a line which, by a corresponding course of reasoning, can be shown to intersect the axes Y and X at O. This line C . . . . . m_3, m_2, m_1

. . . . . O, is therefore the curve, or graph, expressing the equation $y = x$ tan. α, where tan. $\alpha = 1$.

But if such a curve can be plotted for the equation, when tan. $\alpha = 1$, it can also be plotted when tan α equals any decimal fraction from 0.9999 down to 0.0000, or, in other words, the value of the tangent. The value of the tangent being a concrete number, it may be expressed in general by a, and the equation $y = ax$, stands as the equation of any straight line intersecting a pair of coordinate axes, X and Y.

In the present case, the ears of corn employed for illustration, where the correlation coefficient is .87, the tangent value of .87 represents an angle of 41° 2′—the line EO in Figure 3. Taking the equation $y = ax$, or $y =$ tan. αx, we have the following values for y with respect to their corresponding abscissas:

$$x = .87\,y$$
$$2x = 1.74\,y$$
$$3x = 2.61\,y$$

Plotting these, we have the moments m_1, m_2, m_3, of the line AB, (Figure 6), which represents the curve of correlation when the correlation coefficient = .87.

It is hoped that this method of demonstrating the correlation coefficient will be found useful, when biometric work forms a part of the schedule in agricultural classes in plant breeding.

BIBLIOGRAPHY.

[1]Davenport, C. B. Statistical methods with special reference to biological variation New York and London, 1904.
[2]Davenport, E., *Principles of Breeding*, New York, 1907.
[3]Pearl, Raymond, *Modes of Research in Genetics*, New York, 1915.
[4]Rugg, Harold O., *Statistical Methods Applied to Education*, New York, 1917.

DECREASE IN COMMON BRICK.

The production of common brick, until 1917 the most valuable product of the clay-working industry, showed decrease in both quantity and value. The quantity produced in 1918 was 3,450,612,000, which, when compared with the quantity produced in 1917, 5,864,909,000, shows a decrease of 2,414,297,000, or forty-one per cent. The value of common brick marketed in 1918 was $37,208,000, which, when compared with the value in 1917, $47,936,344, shows a decrease of $10,728,000, or twenty-two per cent. The production in 1918 was the smallest ever recorded by the United States Geological Survey, and the value is the smallest since 1898. Common brick is used principally in the structural industries, which, so far as general operations were concerned, were almost paralyzed in 1918. It is true that the number of Government buildings erected in that year was abnormally large, but the number of brick used in these buildings was comparatively small ,and their use by the Government contributed little to offset the losses in the general demand for common brick. The average price per thousand in 1918 compared with 1917 increased $2.61, or to $10.78, the highest average price reached in recent years in the United States for common brick, and nearly twice as great as it was ten years ago.