FURTHER ILLUSTRATIONS OF THE APPLICABILITY OF A COEFFICIENT MEASURING THE CORRELATION BETWEEN A VARIABLE AND THE DEVIATION OF A DEPENDENT VARIABLE FROM ITS PROBABLE VALUE

J. ARTHUR HARRIS

Station for Experimental Evolution, Cold Spring Harbor, New York

[Received November 1, 1917]

TABLE OF CONTENTS

	PAGE
INTRODUCTORY	328
Earlier applications	331
Further illustrations	332
Illustration 1. Proportionality of parts in <i>Paramecium</i>	332
Illustration 2. Absence of relationship between size of litter and sex in swine	336
Illustration 3. Proportion of pistillate and hermaphrodite flowers in the inflorescence of the composite <i>Homogyne</i>	336
Illustration 4. Fertility of capsules and viability of seed in carnation crosses	337
Illustration 5. Relationship between the total number of pedicels and the number of abnormal pedicels in <i>Spiraea Vanhouttei</i>	337
Illustration 6. Interrelationship of cotyledons and primordial leaves in a race of <i>Phaseolus vulgaris</i> highly variable in seedling characters.....	340
Illustration 7. Change in proportion of parts in developing trout.....	344
Illustration 8. Relationship between total solids and sucrose content in the juice of sugar beets.....	346
Illustration 9. Relationship between total number of spikelets and number of sterile spikelets in wheat.....	348
Illustration 10. Viability of dominants and recessives in F_2 generation of Mendelian hybrids	349
RECAPITULATION	350
LITERATURE CITED	351

INTRODUCTORY

Eight years ago I pointed out (HARRIS 1909 a) that when y is some fraction of x the correlation between them, r_{xy} , while of descriptive value,

does not give all of the information which is required concerning the interrelationship of these two variables, and that a coefficient showing whether the value of y becomes relatively larger or smaller with increasing values of x , would have considerable analytical value.

I showed¹ then that if $z = px$, where $p = \bar{y}/\bar{x}$, the bars denoting population means,

$$r_{xz} = \frac{r_{xy} - v_x v_y}{\sqrt{1 - r_{xy}^2 + (r_{xy} - v_x v_y)^2}}.$$

The purpose of this paper is to illustrate the range of usefulness of this coefficient by noting biological progress which has been made by its use, and by actually applying it to series of data which have not been heretofore fully analyzed by the higher statistical methods.

Before passing to actual illustrations of applicability of the coefficient, some questions of method should be considered.

In cases in which the coefficient r_{xz} is desired the correlation table has usually been formed to determine r_{vy} , where $v + y = x$. The mean and standard deviation for x made up of any number of components (HARRIS 1917 c) are well known (PEARL 1909, HARRIS 1918 a); when $x = v + y$

$$\bar{x} = \bar{v} + \bar{y}, \quad \sigma_x = \sqrt{\sigma_v^2 + \sigma_y^2 + 2r_{vy}\sigma_v\sigma_y}$$

If one has made his summations about zero as origin, as suggested elsewhere (HARRIS 1910 b) it is quite easy to determine the necessary moments from those already available.

$$\begin{aligned} \bar{x} &= \Sigma(x)/n = [\Sigma(v) + \Sigma(y)]/n \\ \sigma_x^2 &= [\Sigma(v^2) + \Sigma(y^2) + 2\Sigma(vy)]/n - \bar{x}^2 \end{aligned}$$

The product moments are

$$\begin{aligned} \Sigma(xv) &= \Sigma(v^2) + \Sigma(vy) \\ \Sigma(xy) &= \Sigma(y^2) + \Sigma(vy) \end{aligned}$$

$\Sigma(vy)$ is in each case the product moment already obtained in the calculation of r_{vy} .

In the original paper (HARRIS 1909 a), the method of determining the value of σ_z , which is essential for calculating the equation for the regression of z on x , was not indicated.

¹ The deduction of a usable formula for the problem then in hand was, as I have elsewhere pointed out, due to Professor PEARSON who, with characteristic generosity to his pupils, insisted that my name only appear in the paper.

The first and second moment of y for each grade of x , i.e., $\Sigma(y_x)$ and $\Sigma(y_x^2)$, are easily determined. The first of these will have been obtained in computing the population product moment, $S(xy)$, by the method cited above.

Remembering that $\Sigma(y_x)$ and $\Sigma(y_x^2)$ are the first and second moments about zero as origin for the several y arrays of x , the moments of z for each array are

$$\Sigma(z_x) = \Sigma(y_x) - n_x px$$

From these the means of arrays which are required for plotting the regression curve are given at once by

$$\bar{z}_x = [\Sigma(y_x) - n_x px] / n_x$$

For the whole population the first moment of z about 0 as origin is

$$S\Sigma(z_x) = S[\Sigma(y_x) - n_x px] = 0$$

The second moments for the individual arrays are

$$\Sigma(z_x^2) = \Sigma(y_x^2) - 2px[\Sigma(y_x) - n_x px] - n_x(px)^2,$$

and for the population

$$S\Sigma(z_x^2) = S\{\Sigma(y_x^2) - 2px[\Sigma(y_x) - n_x px] - n_x(px)^2\},$$

or in some cases more conveniently for actual computation,

$$S(z^2) = S(y^2) - 2pS(xy) + S[n_x(px)^2],$$

where S denotes summation of values of y arrays of x , or throughout the population. Thus

$$\sigma_z = [S(z^2)/N]^{1/2}$$

It is quite possible to determine the correlation between x and z , the deviation of y from its probable value, directly.

Remembering that $[\Sigma(y_x) - n_x px]$ is the first moment about zero as origin of z for any array, the product moment for the population is

$$Sx[\Sigma(y_x) - n_x px],$$

where S denotes summation of the values of the y arrays of x .

Practically it is more convenient to determine the product moment from

$$S(xz) = S(xy) - pS(x^2),$$

where $S(xy)$ and $S(x^2)$ are the product moments of x and y and the second moment of x for the population.

EARLIER APPLICATIONS

The method has been most extensively applied to problems of fertility and fecundity. Thus the relationship between the number of ovaries formed and the number of ovaries developing into fruits has been investigated in the inflorescence of *Staphylea* (HARRIS 1909), *Celastrus* (HARRIS 1909 b) and *Crinum* (HARRIS 1912). In *Staphylea* and *Crinum* inflorescences which produce larger numbers of flowers mature relatively fewer fruits. In *Celastrus* there is apparently no relationship between the number of flowers formed and the capacity of the inflorescence for maturing the ovaries into fruits.

In the fruit, the relationship between the total number of ovules laid down and the deviation of the number of seeds matured from their probable number has been investigated in *Sanguinaria* (HARRIS 1910 a).

For *Phaseolus vulgaris* a first study (HARRIS 1913) of 53 series comprising 166,130 pods and a supplemental investigation of 16 series comprising 56,698 pods (HARRIS 1917 a) leave no doubt that the pods with the larger number of ovules mature relatively fewer of their ovules into seeds. The same relationship holds in the arborescent legume, *Cercis canadensis*, as is shown by studies based on massed data (HARRIS 1914 a) and on series from individual trees (HARRIS 1914 b).

The relationship found in *Cercis* and *Phaseolus* is not universal for the Leguminosae. In a series of 1427 pods of *Robinia* (HARRIS 1909 a), the pods with larger numbers of ovules mature a relatively higher proportion of their ovules into seeds. The correlation between the actual number of ovules formed and the actual number of seeds developing is $r_{os} = .693 \pm .009$, while that between the number of ovules formed and the deviation of the number of seeds matured from their probable value is $r_{oz} = .365 \pm .015$.

That this result represents a real biological relationship is indicated by the correlations, hitherto unpublished, for the individual trees. Only three of the twelve constants in table 1 are negative in sign. No one of these can be regarded as statistically significant when the probable error is taken into consideration, while seven of the nine positive coefficients must be looked upon as statistically trustworthy.

The formula has also been advantageously applied to the problem of the interrelationship of the number of male and female flowers in the inflorescence of the aroid *Arisarum* (HARRIS 1916 a) and that of the interdependence of numbers of stamens and pistils in the ranunculaceous genus, *Ficaria* (HARRIS 1918). In *Arisarum* the relative number

TABLE I

Relationship between seed and ovule number in Robinia.

Number of tree	Number of pods	r_{os}	r_{oz}	$r_{oz}/E_{r_{oz}}$
1	122	.524±.044	-.055±.061	(—)0.90
2	64	.802±.030	.468±.066	7.09
3	111	.478±.049	.074±.064	1.16
4	102	.430±.054	.105±.066	1.59
5	122	.671±.034	.338±.054	6.26
6	120	.533±.044	.272±.057	4.77
7	120	.259±.057	-.115±.061	(—)1.89
8	159	.614±.033	.216±.051	4.24
9	128	.590±.040	.350±.052	6.73
10	78	.507±.057	.225±.073	3.08
11	105	.714±.032	-.044±.066	(—)0.67
12	196	.797±.018	.489±.037	13.22
All trees	1427	.693±.009	.365±.015	24.33

of pistillate flowers increases as the total number of flowers per inflorescence increases. In *Ficaria* the relative number of pistils increases as the total number of sporophylls becomes larger.

Dr. BLAKESLEE and I (1918) have applied this coefficient to the determination of the relationship between the total annual egg production and the monthly egg production of White Leghorn fowl. We have there shown by means of this coefficient that the winter months, November, December, January and February, and the following autumn months, August, September and October, show an increase over their theoretical quota of eggs when the annual total egg production rises above the normal. That is, r_{ez} , the correlation between total annual egg production and the deviation of the monthly production from its probable value, is on the whole significantly and substantially positive. The spring and summer months, April, May, June and July, show negative values of r_{ez} , that is, they make a lower relative contribution to the annual total than might be expected when the total varies in the direction of an increase above the normal egg production of the flock as a whole.

FURTHER ILLUSTRATIONS

Illustration 1. Proportionality of parts in Paramecium

JENNINGS (1911) in his masterly investigation of assortative mating in *Paramecium*, has given data for determining the relationships between

(a) distance from the anterior end of the organism to the posterior margin of the mouth, (b) distance from the posterior margin of the mouth to the posterior end of the organism, and (c) the total length in series of conjugant and non-conjugant *Paramecia*.

He has calculated and discussed for a purpose which does not concern us here the correlations between certain of these dimensions. All the correlations between the absolute measurements, calculated from his data, are given in table 2.

TABLE 2
Relationship between total length, l, and anterior length, a, and between anterior length and posterior length, p, in Paramecium.

Series*	r_{ap}	r_{la}	r_{lz}	$r_{lz}/E_{r_{lz}}$
Lot 7, table 40	.246±.050	.735±.025	-.733±.025	3.0
Lot 7, table 41	.570±.047	.893±.014	-.427±.057	7.5
Lot 19, table 51	.382±.044	.832±.016	-.434±.042	10.4
Lot 19, table 52	.671±.034	.939±.074	-.196±.060	3.3
Lot 22, table 55	-.403±.026	.485±.049	-.282±.060	4.7
Lot 22, table 56	.620±.025	.906±.073	-.405±.034	12.0
Lot 24, table 64	.277±.040	.820±.014	-.282±.040	7.1
Lot 24, table 65	.488±.031	.885±.089	-.288±.038	7.6

* Lot 7, table 40, conjugants of wild cultures. 2. Lot 7, table 41, non-conjugants of wild cultures. 3. Lot 19, table 51, conjugants of race *g*. 4. Lot 19, table 52, non-conjugants of race *g*. 5. Lot 22, table 55, wild culture conjugants not yet separated. 6. Lot 22, table 56, wild culture conjugants about twelve hours after separation. 7. Lot 24, table 64, unseparated conjugants of race *k*. 8. Lot 24, table 65, conjugants of race *k* about twelve hours after separation.

In this table the constants for anterior and posterior length are arranged in pairs of conjugants and non-conjugants or ex-conjugants. In every instance the correlation between the anterior and posterior portions of conjugants is lower than that between the same dimensions in non-conjugants or ex-conjugants.²

All of these values are low, as JENNINGS has noted. In the case of lot 22 the coefficient for the conjugants is actually negative in sign.

The correlation between total length and the length of the section anterior to the mouth is high. In every case the value of r_{la} is higher in non-conjugants or ex-conjugants than in conjugants. The additional relationship to be brought out by the formulae here under discussion is

² See in this connection the discussion by JENNINGS (1911, pp. 65-66, 71-73).

that between the total length of the organism and the relative length of either anterior or posterior element.

JENNINGS (1911, p. 63) has emphasized the high variability of the post-oral dimension. Table 3, in which all the coefficients of variation

TABLE 3
Coefficients of variation for anterior and posterior fractions of length in Paramecium.

Series	Total length	Anterior length	Posterior length	$V_p - V_a$
Lot 7, table 40	6.90	5.79	13.86	+ 8.07
Lot 7, table 41	9.68	8.76	14.53	+ 5.77
Lot 19, table 51	8.55	7.78	14.31	+ 6.53
Lot 19, table 52	12.18	12.08	15.13	+ 3.05
Lot 22, table 55	5.48	7.38	14.29	+ 6.91
Lot 22, table 56	7.96	7.28	11.20	+ 3.92
Lot 24, table 64	6.34	6.42	9.98	+ 3.56
Lot 24, table 65	6.62	6.45	9.35	+ 2.90

are laid side by side, fully confirms his conclusion in this regard. Utilizing these coefficients of variation we obtain the values for the correlation between total length and the deviation of the anterior length from its probable value, given in the fourth column of table 2.

These constants are negative in sign throughout, and while variable in magnitude all may reasonably be considered statistically significant in comparison with their probable errors.

Thus when *Paramecium* varies in length both anterior and posterior fractions of the body contribute to this variation, but as length increases the anterior portion becomes relatively shorter.

For one series, the unseparated conjugants of race *k*, I have determined the regression of the anterior length on total length and the regression of the deviation of the anterior length from its probable value on the total length of the organism. The equations are

$$a = 3.0359 + .4977 l$$

$$\sigma_x = .7616, \quad z_a = 3.3743 - .1135 l$$

The equations and empirical means are represented graphically in diagrams 1 and 2. In both cases the relationships are sensibly linear.

DIAGRAM 1.—Relationship of anterior length to total length in Paramecium. Compare diagram 2.

DIAGRAM 2.—Relationship of deviation of anterior length from its probable value to total length in Paramecium. Compare diagram 1.

Illustration 2. Absence of relationship between size of litter and sex in swine

PARKER and BULLARD (1913) have discussed the possible relationship between the size of the litter and sex in the contents of 1000 uteri of swine. From a simple percentage table they state that the relative numbers of males and females are "even in the extreme cases so nearly uniform that we may conclude with reasonable assurance that there is no intimate relation between sex and the size of the litters."

The correlations between the total numbers of pigs in the litter, l , and the number of males, m , and females, f , may be deduced from their data. They are

$$r_{lm} = .6833 \pm .0114$$

$$r_{lf} = .6875 \pm .0112$$

From these and the three coefficients of variation one may deduce

$$\text{For males } r_{lz} = - .0177 \pm .0213$$

$$\text{For females } r_{lz} = + .0177 \pm .0213$$

The correlation is sensibly zero, with regard to its probable error. This method of analysis therefore fully confirms the conclusion drawn by PARKER and BULLARD.

Illustration 3. Proportion of pistillate and hermaphrodite flowers in the inflorescence of the composite Homogyne

LUDWIG (1901) has given data for the correlation between the number of pistillate and the number of hermaphrodite flowers in the inflorescence of Homogyne. From his data we deduce

For hermaphrodite flowers, h ,

$$\bar{h} = 31.8333, \sigma_h = 7.3981, V_h = 23.240$$

For pistillate flowers, p ,

$$\bar{p} = 10.5370, \sigma_p = 2.6460, V_p = 25.112$$

For total flowers, f ,

$$\bar{f} = 42.3704, \sigma_f = 8.7749, V_f = 20.7099$$

For hermaphrodite and pistillate flowers,

$$r_{hp} = .3899 \pm .0449$$

For total flowers and hermaphrodite flowers,

$$r_{fh} = .9607 \pm .0049, r_{fz_h} = .2429 \pm .0499$$

For total flowers and pistillate flowers,

$$r_{fp} = .6303 \pm .0339, r_{fz_p} = - .2429 \pm .0499$$

It follows, therefore, that in the larger heads the purely pistillate flowers are relatively less, and the hermaphrodite flowers relatively more, numerous.

Illustration 4. Fertility of capsules and viability of seed in carnation crosses

STUART (1912) has recorded the number of seeds obtained and the number germinated, planted into the field, and producing flowers in various carnation crosses. Our problem is to determine whether the seeds which come from capsules producing a large number of seeds are relatively more (or less) viable than those from capsules producing small numbers.

Using his two larger tables of data, tables 3 and 6,³ and confining attention to the relationship between number of seeds per capsule, and the number which germinated, I find

For commercial $\times$ commercial, STUART's table 3, $N = 23$,

$$r_{sg} = .775 \pm .056, \quad r_{sz} = -.072 \pm .141$$

For single flower $\times$ double flower, STUART's table 6, $N = 32$,

$$r_{sg} = .649 \pm .069, \quad r_{sz} = -.118 \pm .118.$$

The signs are both negative, indicating a relatively higher failure to germinate among the seeds which are produced many in a capsule. With regard to their probable errors, the constants are untrustworthy. Because so few observations are available, no biological significance is attached to these two series, which serve merely as another illustration of the kind of problems to which the method may be applied.

Illustration 5. Relationship between the total number of pedicels and the number of abnormal pedicels in Spiraea Vanhouttei

In *Spiraea Vanhouttei* the pedicels of the umbel-like raceme normally produce but a single flower each. An abnormal condition in which one or more pedicels may bear a relatively large number of flowers is frequently observed (HARRIS 1917 d).

Let x be the total pedicels in an inflorescence and a the number which are abnormal. Then if abnormality be distributed purely at random among the pedicels one would expect material values of r_{xa} . The correlation r_{xz} meets our requirements since it shows whether inflorescences with a large number of rays have relatively more or fewer of their rays abnormal than those with a small number.

³ In table 6 the cases in which the seeds are not normally developed are omitted.

During the last fifty years a great deal has been said about the influence of nutrition, vegetative vigor, etc., upon the development of anomalies. If a larger number of rays indicates greater vigor or better nutrition one might *a priori* expect larger inflorescences to have a proportionately higher number of branched rays, providing of course, that the classic theories are true.

The constants for a short series of data collected in 1906 were published in 1909. Since then a large number of determinations have been made on a general sample of inflorescences from a number of shrubs in 1909 and from three large individual shrubs in 1913.

In the latter series the data have been analyzed in two ways. First, the inflorescences which contain at least a single abnormal pedicel have been used as the basis of the correlations. These are designated as the abnormal inflorescences. Second, the normal inflorescences from the same plants have been included and counted as zero in the distribution of number of abnormal rays.

The results are:

$$\begin{aligned}\text{For 1906}^4 \quad r_{xa} &= +.121 \pm .034 \\ r_{xz} &= -.071 \pm .034\end{aligned}$$

For 1909. Massed statistics. Inflorescences producing some abnormal pedicels ($N = 785$),

$$\begin{aligned}r_{xa} &= +.1542 \pm .0235 \\ r_{xz} &= -.0915 \pm .0239\end{aligned}$$

For 1909. Massed statistics. All 2040 inflorescences,

$$\begin{aligned}r_{xa} &= +.1584 \pm .0146 \\ r_{xz} &= +.0370 \pm .0149.\end{aligned}$$

For 1913. Individual plants. Inflorescences producing some abnormal pedicels,

Plant 1. $N = 747$ inflorescences.

$$\begin{aligned}r_{xa} &= +.0880 \pm .0244 \\ r_{xz} &= -.2846 \pm .0227\end{aligned}$$

Plant 2. $N = 641$ inflorescences.

$$\begin{aligned}r_{xa} &= +.1148 \pm .0263 \\ r_{xz} &= -.3855 \pm .0227\end{aligned}$$

⁴ Since in the 1906 series only synanthies were observed, the total pedicels (including those with synanthies) were counted and the correlation is between the total pedicels and the pedicels with synanthies. In the 1909 material in which synanthies were rare, the correlation was determined for total pedicels and abnormal pedicels, abnormality being synanthi or any degree of branching.

Plant 3. $N = 548$ inflorescences.

$$r_{xa} = + .0941 \pm .0285$$

$$r_{xz} = - .2849 \pm .0265$$

For 1913. Individual plants. All inflorescences,

Plant 1. $N = 1135$ inflorescences.

$$r_{xa} = - .0067 \pm .0200$$

$$r_{xz} = - .2125 \pm .0191$$

Plant 2. $N = 975$ inflorescences.

$$r_{xa} = - .0821 \pm .0214$$

$$r_{xz} = - .3681 \pm .0187$$

Plant 3. $N = 912$ inflorescences.

$$r_{xa} = - .0360 \pm .0223$$

$$r_{xz} = - .2342 \pm .0211$$

For all the samples of inflorescences in which there is at least one abnormal pedicel the correlations between the total number of pedicels and the number of normal pedicels is positive in sign and perhaps statistically significant, but low in actual magnitude. Thus the number of abnormal pedicels increases on the average as the total number of pedicels per inflorescence becomes larger. The relationships are, however, very slight indeed.

For these five series the correlation between the total number of pedicels and the deviation of the abnormal pedicels from their probable value, is negative in sign. Thus the larger inflorescences have a relatively smaller proportion of abnormal pedicels than do those with a smaller total number of pedicels.

In the four series in which the wholly normal inflorescences are included, the correlations between total number of pedicels and number of abnormal pedicels is positive in 1909 but negative throughout and insignificant in magnitude in 1913. The three series from individual shrubs studied in 1913 show low but significantly negative correlations between the total number of pedicels per inflorescence and the deviation of the number of abnormal inflorescences from their probable value. The constant for the heterogeneous data of 1909 is positive but insignificant.

Taking the data altogether, there can be no reasonable doubt that the relative number of abnormal pedicels decreases as the total number of abnormal pedicels increases.

This is shown in diagram 3, which represents the regression of the deviation of the number of abnormal rays from their probable value on

the total number of rays in the series showing the lowest, and in one of these showing the highest, correlation.

DIAGRAM 3.—Regression of the deviation of the number of abnormal pedicels from their probable value on the total number of pedicels in *Spiraea*.

The standard deviations are:

For 1909, $\sigma_z = 2.4691$.

For 1913, plant 1, $\sigma_z = 3.0527$.

The regression equations are:

For 1909, $z = 1.3186 - .0609 x$.

For 1913, $z = 3.5036 - .2212 x$.

Illustration 6. Interrelationship of cotyledons and primordial leaves in a race of Phaseolus vulgaris highly variable in seedling characters

The ontogeny of the flowering plant is usually divided into a primary period of dependent development and a sharply separated period of independent growth. The leaf homologs which are developed during the first period are usually of a very definite form and number, and frequently

highly differentiated from those developed after growth is resumed as an independent organism.

Those rare cases in which the number of leaf homologs which are laid down during the development of the seed is highly variable seem to offer especially favorable opportunities for the morphologist to learn something of the interrelationship of the two forms which may be assumed by homologous organs.

Suitable material for such work is furnished by a race of *Phaseolus vulgaris* (HARRIS 1916 b) having a modal number of four cotyledons and four primordial leaves, but highly variable in the number of both of these organs.

The correlations and regression equations are:⁵

Correlation between number of cotyledons and number of leaves:

Less mature series, $r = .1170 \pm .0320$.

$$c = 3.4759 + 0.0421 l$$

$$l = 3.2919 + 0.3252 c$$

More mature series, $r = .1568 \pm .0236$.

$$c = 3.5128 + 0.0499 l$$

$$l = 3.0039 + 0.4922 c$$

Combined series, $r = .1386 \pm .0193$.

$$c = 3.5030 + 0.0464 l$$

$$l = 3.3565 + 0.4142 c$$

The correlations between number of cotyledons and number of primordial leaves indicate slender interrelationships in this newly originated race between characters which are usually thought of as highly correlated.

Note from the equations that there is on the average a change of .3 to .5 of a leaf for each variation of one cotyledon, but that there is a change of only .04 of a cotyledon for a variation of one leaf.

The results for the combined series are represented graphically in diagram 4. Regression is for each character very nearly linear.

Correlation between total leaf homologs and number of cotyledons:⁶

Less mature series, $r = .4328 \pm .0264$.

$$c = 2.4595 + 0.1413 h$$

⁵ The plants were graded in a less mature and a more mature series as a precaution against recording as primordial leaves those really developed subsequently to the maturation of the seed.

⁶ The following correlations are published here for the first time.

DIAGRAM 4.—Relationship between number of cotyledons and number of primordial leaves in a tetracotyledonous race of *Phaseolus vulgaris*.

More mature series, $r = .4337 \pm .0197$.

$$c = 2.6704 + 0.1260 h$$

Combined series, $r = .4312 \pm .0084$.

$$c = 2.5954 + 0.1314 h$$

The straight line showing the regression of number of cotyledons, c , on total leaf homologs, h , is shown with the empirical means in diagram 5.

Correlation between total leaf homologs and number of leaves:

Less mature series, $r = .9459 \pm .0034$.

$$l = -2.4595 + 0.8586 h$$

DIAGRAM 5.—Relationship between number of cotyledons and number of leaves and total number of leaf homologs (cotyledons + leaves) in Phaseolus.

More mature series, $r = .9579 \pm .0020$

$$l = -2.6704 + 0.8740 h$$

Combined series, $r = .9533 \pm .0018$.

$$l = -2.5953 + 0.8685 h$$

Diagram 5 gives the regression of number of leaves on the number of leaf homologs (cotyledons + leaves). The solid dots representing the empirical mean leaf number lie practically on the theoretical line.

Correlation between total leaf homologs and the deviation of the number of leaves from their probable value:

Less mature series, $r = .6936 \pm .0226$.

More mature series, $r = .7642 \pm .0185$.

Combined series, $r = .7378 \pm .0089$.

The correlations between total leaf homologs and the deviations of the cotyledons from their probable value are numerically identical with the foregoing but negative in sign.

The results show a high degree of consistency of the two series. These final constants show that when the total number of leaf homologs increases, the variation is due to a far greater extent to the laying down of a greater number of primordial leaves than to the formation of a larger number of cotyledons. The regression equations for both the deviation of the number of leaves from their probable value (z_l) and the deviation of the number of cotyledons from their probable value (z_c) are given by

$$\sigma_z = .874557$$

$$z_c = 2.5954 - 0.3007 h$$

$$z_l = -2.5954 + 0.3007 h$$

and represented in diagram 6. The results are clearly linear.

Illustration 7. Change in proportion of parts in developing trout

JENKINSON (1912) has given data for total length and length of head for three growth stages in the American rainbow trout. His constants are:

	Total length, l CV	Head length, h CV	Total length and head length r_{lh}
Stage 1	13.11	19.68	.945
Stage 3	5.11	6.57	.729
Stage 5	7.74	6.55	.848

DIAGRAM 6.—Regression of the deviation of the number of cotyledons and of the number of leaves from their probable numbers on the total number of leaf homologs in *Phaseolus*. Note that the theoretical lines and the empirical means are identical but opposite in sign.

Whence we deduce:

$$\text{Stage 1, } r_{12} = +.649 \pm .027.$$

$$\text{Stage 3, } r_{12} = -.071 \pm .047.$$

$$\text{Stage 5, } r_{12} = -.534 \pm .038.$$

These results bring out especially well the analytical value of the method here proposed. The correlations between total length and head length are *roughly* the same in the three stages of development. The correlation between total length and deviation of head length is at first strongly positive, then sensibly zero in the third stage and finally becomes negative and of the order $-.5$ in the fifth stage.

Thus while in each stage the larger individuals have the larger heads, the relationships between the total length of the body and the proportional length of the head changes greatly during development.

Illustration 8. Relationship between total solids and sucrose content in the juice of sugar beets

Sucrose content of the juice of the sugar beet has been one of the classic examples of variation in text-books on genetics. Nevertheless our knowledge of the problems are, as indicated by a recent review (HARRIS 1917 b), far from complete. Relatively little is known, for example, of the correlation between total solids and sucrose content. Yet this problem is not merely of physiological and genetical interest but of very great practical importance as well, since the coefficient of purity of the juice is an important factor in sugar manufacture.

Table 4 shows the correlation between total solids and sucrose content in 475 Nevada sugar beets. Nominally this series is composed of a number of commercial varieties, but since it has been shown elsewhere (HARRIS and GORTNER 1913) that the correlation between weight and composition of juice is not greatly influenced by the combination of these nominal varieties, such heterogeneity as exists probably does not influence materially the correlations to be deduced. For purposes of comparison the constants for a series of 61 determinations from Washington are added.

The Nevada series gives:

For total solids,

$$\text{Mean} = 18.021, \sigma = 3.7224, \text{CV} = 20.656$$

For sucrose,

$$\text{Mean} = 14.989, \sigma = 4.0727, \text{CV} = 27.171$$

For total solids and sucrose,

$$r = .882 \pm .007$$

For total solids and deviation of sucrose from its probable value,

$$r_{sz} = + .251 \pm .029$$

For the Washington series the results are:

For total solids,

$$\text{Mean} = 16.445, \sigma = 1.5896, \text{CV} = 9.666$$

For sucrose,

$$\text{Mean} = 11.838, \sigma = 1.9098, \text{CV} = 16.133$$

For total solids and sucrose,

$$r = .866 \pm .022$$

TABLE 4

Total solids and sucrose content in sugar beets.

	Sucrose content																				Totals
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
6	—	—	2	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3
7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
9	—	1	—	1	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	3
10	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	1
11	—	—	—	—	—	1	—	2	—	—	—	—	—	—	—	—	—	—	—	—	7
12	—	—	—	1	—	3	3	5	2	—	—	—	—	—	—	—	—	—	—	—	14
13	—	—	—	—	—	—	1	4	3	10	2	1	—	—	—	—	—	—	—	—	21
14	—	—	—	—	1	1	1	3	8	6	8	6	1	—	—	—	1	—	—	—	36
15	—	—	—	1	—	—	1	—	7	3	15	9	5	—	—	—	—	—	—	—	41
16	—	—	—	—	—	1	1	1	—	1	6	9	13	4	1	—	—	—	—	—	37
17	—	1	—	—	—	—	—	1	—	1	1	5	9	16	5	1	—	—	—	—	40
18	—	—	—	—	—	—	—	1	—	—	2	3	7	10	20	9	1	—	—	—	53
19	—	—	—	—	1	—	—	1	—	—	2	1	3	4	7	17	6	—	—	—	42
20	—	—	—	—	—	—	—	—	—	1	1	1	—	1	4	18	28	3	—	1	58
21	—	—	—	—	—	—	—	—	—	—	—	2	—	4	2	2	8	5	10	1	34
22	—	—	—	—	—	—	—	—	—	—	—	1	—	3	1	3	5	9	5	3	31
23	—	—	—	—	—	—	—	—	1	—	—	—	—	—	1	2	1	1	5	8	20
24	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	1	3	9	22
25	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	4
26	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4
27	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	2
34	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
Totals	1	2	2	4	2	6	13	18	21	22	37	38	38	42	40	53	51	19	23	24	475

Total solids

For total solids and deviation of sucrose from its probable value,

$$r_{sz} = .471 \pm .067$$

Thus in both cases the correlation between total solids and sucrose is moderately high, as is necessarily the case because of the fact that total solids are to a considerable extent made up of sucrose.

Quantity of sucrose is in both cases more variable than the total amount of solids. The correlation between total solids and the relative amount of sucrose is in both cases positive and of a substantial order of magnitude. Beet juice with higher total solids contains both absolutely more and relatively more sugar than that of beets with low total solid content.

Another way of approaching the problem is to calculate the correlations between total solids and the coefficient of purity, which is often determined.

Nevada series,

$$r_{sp} = .340 \pm .027$$

Washington series,

$$r_{sp} = .529 \pm .062$$

These values are also positive and substantial, but their interpretation involves difficulties due to spurious correlation.

Illustration 9. Relationship between total number of spikelets and number of sterile spikelets in wheat

GRANTHAM and GROFF (1916) have given data for the correlation between the total number of spikelets and the number of sterile spikelets at the base of the head in varieties of wheat. Let t = total spikelets and s = sterile spikelets. Then from their two tables I find, without applying SHEPPARD'S correction,

Bearded varieties,

$$r_{ts} = .635 \pm .039, r_{tzs} = +.164 \pm .063$$

Beardless varieties,

$$r_{ts} = .542 \pm .054, r_{tzs} = +.232 \pm .072$$

For both groups of varieties there is a positive correlation between total spikelets and sterile spikelets and between total spikelets and the deviation of the number of sterile spikelets from their probable value. Sterility is, therefore, not merely absolutely but relatively more frequent in varieties with larger numbers of spikelets.

Illustration 10. Viability of dominants and recessives in F_2 generation of Mendelian hybrids

F_1 D(R) $\times$ D(R), according to Mendelian theory, gives in F_2 a population with somas 3 D: R. Let there be N individual families in this population and the number of individuals per family be variable. This variation may be due solely to fluctuations in the number of zygotes produced, or it may be in part attributable to the failure of some of the zygotes to reach maturity. This failure may be random or differential with respect to the alternative characters involved. If differential, one might expect a correlation between the actual number of individuals per family and the deviation from the probable number of either of the alternative types of individuals which it contains.

In current Mendelian literature discrepancies between observed and theoretical ratios are often explained as due to selective fertilization or to a lower viability of particular zygotes.

If the latter explanation be correct r_{lz} , the correlation between the number in the litter and the deviation from their probable frequency of the number which shows a particular character, should have a statistically significant value. As an illustration of the kind of data to which this test may be applied, I take YULE'S (1914) table of DARBISHIRE'S (1904) results for albinos in mice families. If the fertilized ovum which is to produce an albino be less capable of development than that which is to develop into a normal individual r_{la} should take a substantial positive value, i.e., the relative numbers of albino young should be larger in the large litters, since the small litters are assumed to be small *in part* because albino-producing fertilized ova are supposed to be less viable.

The actual results are:

$$\begin{aligned}\bar{l} &= 4.5868 & \bar{a} &= 1.1322 \\ \sigma_l &= 1.7518 & \sigma_a &= 1.0752 \\ V_l &= 38.193 & V_a &= 94.963 \\ r_{la} &= .3756 \pm .0527 \\ r_{lz} &= -.0287 \pm .0612\end{aligned}$$

where l = number in litter, a = number of albinos in litter, and z = deviation of number of albinos from their probable value.

The constant of critical significance, r_{lz} , is sensibly zero, with due regard to its probable error. Hence there is no evidence of a greater intra-uterine mortality of albino zygotes.

RECAPITULATION

In many instances the biologist has to consider the relationship between a measurement and some of its logical subdivisions or components. The first has been called an independent and the second, which is always some fraction of the first, a dependent variable.

For the analysis of such relationships two coefficients are required, the correlation between the variable and the dependent variable and the correlation between the variable and the deviation of the dependent variable from its probable value on the assumption that the relative magnitude of the dependent variable is independent of the magnitude of the variable.

This paper gives (*a*) the supplementary formulae which are required in certain cases in which the correlation between a variable and the deviation of a dependent variable from its probable value is to be computed, and (*b*) a series of illustrations of the applicability of this coefficient, drawn from a wide range of biological phenomena.

The method has been most extensively applied to the problems of the physiology of seed production in plants. The relationship of the number of ovaries which develop to maturity to the total number of ovaries formed, and the relationship of the number of seeds which mature to the total number of ovules laid down, have been determined in a considerable number of plant forms. That it may be useful in the study of the relationship of seed viability to the number of seed formed is shown by analysis of meagre data for carnation crosses.

The formulae have been applied to the problem of sex by an investigation of the relative proportion of macrosporophylls and microsporophylls in data for *Arisarum*, *Ficaria* and *Homogyne*, and for the relative numbers of males and females in litters of swine.

Under certain conditions the method may be of use in testing the assumption of the existence of a differential viability in Mendelian dominants and recessives.

The usefulness of the method in morphology has been illustrated by its application to the problem of the relationship between total length and anterior length in *Paramecium*, to that of the relative size of the head in developing trout, to that of the relative frequency of abnormal pedicels in the inflorescence of *Spiraea*, and to that of the relationship of the numbers of cotyledons and primordial leaves to the total number of leaf homologs in a highly variable race of *Phaseolus*. Such studies have not merely an independent morphological value, but have their bearing

upon such biological theories as DRIESCH's annunciation that in a differentiated system the proportionality of the parts is absolutely independent of the size of the system.

In dealing with materials of economic importance, the method has been applied to the problem of the relative proportion of sterile spikelets in varieties of wheat with varying numbers of total spikelets per head, to the relationship between total solids and sucrose content in the juice of sugar beets, and has been found of the greatest value in analyzing the relationship between the egg production of the individual months and that of the entire year in the domestic fowl.

These illustrations are perhaps sufficient to show that the coefficient has great analytical value, and should have the widest usefulness.

Finally it must be understood that the conclusions drawn from the illustrations given in this paper are not in all cases to be extended beyond the specific series of data to which the formulae have been applied. The data employed are, *in some instances*, intended to illustrate the type of biological problem to which the formulae may be applied when more extensive data are available. No attempt has been made to discuss in detail the biological significance of results of so diverse kinds of illustrations. Such discussion must be left to the specialist in the particular field in which the observations fall.

LITERATURE CITED

- DARBISHIRE, A. D., 1904 On the result of crossing Japanese waltzing with albino mice. *Biometrika* **3**: 1-51.
- GRANTHAM, A. E., and GROFF, F., 1916 Occurrence of sterile spikelets in wheat. *Jour. Agr. Res.* **6**: 235-250.
- HARRIS, J. ARTHUR, 1909 a The correlation between a variable and the deviation of a dependent variable from its probable value. *Biometrika* **6**: 436-443.
- 1909 b Correlation in the inflorescence of *Celastrus scandens*. *Ann. Rep. Mo. Bot. Gard.* **20**: 116-122.
- 1910 a A quantitative study of the morphology of the blood root, *Sanguinaria canadense*. *Biometrika* **7**: 305-351.
- 1910 b The arithmetic of the product moment method of calculating the coefficient of correlation. *Amer. Nat.* **44**: 693-699.
- 1912 Biometric data on the inflorescence and fruit of *Crinum longifolium*. *Ann. Rep. Mo. Bot. Gard.* **23**: 75-99.
- 1913 On the relationship between the number of ovules formed and the capacity of the ovary for maturing its ovules into seeds. *Bull. Torr. Bot. Club* **40**: 447-455.
- 1914 a On the relationship between the number of ovules formed and the number of seeds developing in *Cercis*. *Bull. Torr. Bot. Club.* **41**: 243-256.
- 1914 b Further observations on the relationship between the number of ovules formed and the number of seeds developing in *Cercis*. *Bull. Torr. Bot. Club* **41**: 533-549.

- HARRIS, J. ARTHUR, 1916 a On the distribution and correlation of the sexes (staminate and pistillate flowers) in the inflorescence of the aroids *Arisarum vulgare* and *Arisarum proboscideum*. Bull. Torr. Bot. Club **42**: 663-673.
- 1916 b A tetracotyledonous race of *Phaseolus vulgaris*. Mem. N. Y. Bot. Gard. **6**: 229-244, *figs. 1-3*.
- 1917 a Supplementary determinations of the relationship between the number of ovules per pod and fertility in *Phaseolus*. Genetics **2**: 282-290.
- 1917 b Biometric studies on the somatic and genetic physiology of the sugar beet. Amer. Nat. **51**: 507-512.
- 1917 c The correlation between a component, and between the sum of two or more components, and the sum of the remaining components of a variable. Quart. Publ. Amer. Stat. Ass. **15**: 854-859. *pls. 30, 31*.
- 1918 a The interrelationship of the number of stamens and pistils in the flowers of *Ficaria*. Biol. Bull. **34**: 7-17, *figs. 1-5*.
- 1918 b On the distribution of abnormalities in the inflorescence of *Spiraea Van-houttei*. Amer. Jour. Bot. **4**: 624-636, *diagrams 1-4, pls. 30, 31*.
- HARRIS, J. ARTHUR, and BLAKESLEE, A. F., with the coöperation of KIRKPATRICK, WM. F., 1917 Inter-periodic correlation in the egg production of the domestic fowl. Proc. Nat. Acad. Sci. **3**: 565-569, *figs. 1, 2*.
- HARRIS, J. ARTHUR, and GORTNER, R. A., 1913 On the relationship between the weight of the sugar beet and the composition of its juice. Jour. Ind. Eng. Chem. **5**: 192-195.
- 1918 The correlation between egg production during the various periods of the year in the domestic fowl. Genetics **3**: 27-72, *figs. 1-11*.
- JENKINSON, J. W., 1912 Growth, variability and correlation in young trout. Biometrika **8**: 444-455.
- JENNINGS, H. S., 1911 Assortative mating, variability and inheritance of size in the conjugation of *Paramecium*. Jour. Exp. Zool. **11**: 1-134.
- LUDWIG, F., 1901 Variationsstatistische Probleme und Materialien. Biometrika **1**: 11-29.
- PARKER, G. H., and BULLARD, C., 1913 On the size of litters and the numbers of nipples in swine. Proc. Amer. Acad. Arts and Sci. **49**: 399-426.
- PEARL, R., 1909 The frequency constants of a variable $z = f(x_1, x_2)$. Biometrika **6**: 437-438.
- STUART, W., 1912 Mendelian inheritance in the carnation. Bull. Vt. Agr. Exp. Sta. 163.
- YULE, G. U., 1914 Fluctuations of sampling in Mendelian ratios. Proc. Cambridge Phil. Soc. **17**: 425-432.