# REGRESSION AND CORRELATION EVALUATED BY A METHOD OF PARTIAL SUMS

By Felix Bernstein

> "To be sure, Laplace viewed the matter in a similar way but he selected the absolute value of the error as a measure of loss. But if we mistake not, this position is certainly not less arbitrary than our own; that is to say, whether the double error is to be considered just as tolerable as, or worse than, the simple error twice repeated and whether it is thus more fitting to ascribe to the double error only a double weight, or a greater one, is a question which is neither in itself clear nor determinable by mathematical proof but has to be left entirely to individual discretion.
>
> "Furthermore, it cannot be denied that the assumption under discussion violates the principle of continuity and precisely for this reason the procedure based on it strongly defies analytic treatment while the results to which our principle leads have the advantage of simplicity as well as of generality."—*F. G. Gauss: Theoria combinationis observationum, pars prior, art. 6.*

Since the "Theoria Combinationis" of C. F. Gauss appeared in the year 1821 a century of Mathematical Statistics has been dominated by the ideas of this classical treatise—ideas whose fertility does not seem to be exhausted even today.

The germ of most modern contributions to mathematical statistics—in fact also those of Karl Pearson and his school—go back decidedly to this paper. Though the immediate achievements of Gauss are so conspicuous as not to need any comment, a true critical appreciation of the work can be gained only by comparing it with the previous methods of Laplace, superseded by those of Gauss.

For such critical appreciation, C. F. Gauss himself has prepared the ground in the lines quoted at the beginning of this article. To Gauss the standard deviation is a measure of uncertainty or risk of a game in which the errors of observation are considered as causing only losses. In this he follows the lead of his great predecessor. The difference between them is that Gauss adopts the square of the error as a measure of the loss while Laplace adopts its absolute value for this purpose, Either choice frees the error from its sign so that the loss is the same regardless of the sign of the error.

Gauss considers this choice of the measure of the loss as purely conventional. Therefore he feels justified in adopting the square of the error because in adopting the square instead of the absolute value of the error, the mathematics he uses remains in the easily accessible domain of analytical processes. This creates for these methods a superiority in elegance, simplicity, and generality.

The modern developments of mathematical statistics, based on the principles of Gauss, have confirmed the correctness of this viewpoint. This has proved true particularly in the theory of analysis of variance developed by R. A. Fisher and in the more general theory of semi-invariants, first defined by N. H. Thiele.

The inadequacy of the Gaussian method seriously impairing its value for statistical use has come to light through the investigations of Karl Pearson of distributions of one and two variables. Since the moments of higher order involve standard deviations of increasing magnitude the characterization of the distributions by means of the moments, in line with the Gauss-Thiele concepts, becomes practically impossible. Therefore it was of the greatest interest that Lindeberg was able to derive an expression for the standard deviation of a measure of skewness constructed not on Gaussian but on Laplacian lines, namely based exclusively upon the sign of the error. The mathematical difficulties surmounted by Lindeberg by a very involved and difficult analysis—with some clearly indicated gaps in the proofs—are precisely of the character of those that Gauss wished to avoid. Encouraged by the success of Lindeberg, I have developed in two papers[1] the standard deviations of more general moments and the correlations between them of which the mean deviation of Laplace and Lindeberg's measure of skewness are special cases. The proofs have been arrived at by a rather simple and rigorous procedure. These new moments, together with the old ones, form a new system of statistical characteristics by which a distribution in one or two variables can be described by expressions of lower order and therefore of greater precision. This method makes unnecessary the use of moments of higher order than the third.

But another point of interest is still involved. It has been assumed that the Gaussian characteristics give a greater amount of information than those of Laplace. This is proved, however, only for the case of the normal distribution $\frac{h}{\sqrt{\pi}} e^{-h^2x^2}$ This was recognized by Gauss himself in his paper of April, 1816, that appeared five years earlier than the Theoria Combinationis Observationum. In article 6 of his paper, he says, that the constant $h$ of a normal distribution obtained from one hundred observations by the use of the standard error is as exact as that obtained from one hundred fourteen observations in which the mean deviation is used. Hence with a given number of observations only the equivalent of 88% of the total are used by the second method. This does not hold true for all distributions. The following theorem can easily be proved: The amount of information as defined above, furnished by the use of the mean deviation is greater, equal to, or less than that furnished by the standard deviation, depending respectively upon whether

[1] Felix Bernstein: "Die mittleren Fehlerquadrate und Korrelationen der Potenzmomente und ihre Anwendung auf Funktionen der Potenzmomente," Metron, Vol. X, N. 3 (Nov. 1932).

Felix Bernstein: "Uber den mittleren Fehler der Potenzmomente." Zeitschr. f. d. ges. Vers.-Wissenschaft, Bard 30, Heft 3, March 1930.

$$(\beta_2 - 1) \gtreqless 4(\beta_0 - 1)$$

where

$$\beta_0 = \frac{\mu_2}{\vartheta^2}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$\mu_k$ the $k$-th moment and $\vartheta$ = the mean deviation.

For example, in the distribution $\frac{h}{2} e^{-h|x|}$, the mean deviaition furnishes a greater amount of information than the standard deviation.[2]

In the present paper, we shall discuss the practical use of expressions for correlation and regression in which the new type of statistics formed along Laplacian lines will be used. These new expressions are of a linear form and can be computed therefore more easily than those of Karl Pearson. The amount of information given by these expressions is less than that given by the expressions of Pearson if the normal law, in two variables, is fulfilled. For other distributions, however, this is not generally true. The determination of the standard deviations of these new expressions is given in Metron.[3]

The application of the new expressions of regression and correlation to grouped data is set forth here for the first time. The method is strongly recommended for all cases in which the data lose reliability with increasing deviations from the mean. Deviations in the new method enter the expressions only in the first degree and not in the second as in the case of Pearson's. It is obvious that the influence of the doubtful extreme readings is, therefore, considerably lessened. Since our expressions are linear, no adjustments for grouping (Sheppard's corrections) are necessary.

It ought to be mentioned here that linear expressions for the measurement of correlation have been set up before.

K. Pearson (Biometrika) and Egon Pearson (Biometrika) have derived an expression called "linear correlation ratio" which in case of linear regression is identical with the correlation coefficient.

K. Pearson also discusses the linear correlation coefficient

$$r = \frac{1}{2}\left(S\frac{ysgx}{xsgx} + S\frac{xsgy}{ysgy}\right),$$

[2] To this second type of distribution curves also belongs $y = \psi(x)$ where $x(x)$ is the mean of two Gaussian curves with the same origin, i.e. $\psi(x) = \frac{1}{2}\left(\frac{h}{\sqrt{\pi}} e^{-h^2x^2} + \frac{kh}{\sqrt{\pi}} e^{-h^2h^2x^2}\right)$ $1.6 < k < 3.4$.

I owe this remark and some other valuable suggestions regarding the subject of this paper to Mr. Myron Fuchs.

[3] *Op. cit.*

suggested by Lenz and various other linear expressions, all similar to our expression (1). He finds that they are all equal to his quadratic correlation coefficient in the case of a Gaussian distribution.

However, their expressions were not recommended by those authors for the determination of correlation between quantitative variables, because—

1. No easy and practicable methods were given for their evaluation in the case of grouped data.
2. Their standard deviations were not determined.

We now proceed to define the new formulas and to describe the methods for their evaluation. The proofs are furnished in the Appendix to this paper.

Let $r_1$ and $r_2$ denote the regression coefficients of $x$ on $y$ and $y$ on $x$ respectively, and $r$, as usual, the coefficient of correlation, and by $\bar{x}$ and $\bar{y}$ the arithmetic means of the $x$'s and $y$'s. Let us take $\bar{x}$, $\bar{y}$ as the origin, so that $x$, $y$ are the deviations from the mean. We have

$$(1) \qquad \begin{gathered} r_1 = \frac{\underset{+y}{Sx}}{\underset{+y}{Sy}} \quad \text{or} \quad r_1 = \frac{\underset{-y}{Sx}}{\underset{-y}{Sy}} \\ r_2 \quad \frac{\underset{+x}{Sy}}{\underset{+x}{Sx}} \quad \text{or} \quad r_2 \quad \frac{\underset{-x}{Sy}}{\underset{-x}{Sx}} \\ r = \sqrt{r_1 \times r_2} \end{gathered}$$

$\underset{+y}{Sx}$ denotes a partial sum of the $x$'s, this sum being extended over all the $x$'s of the observations whose $y$ is positive and the other sums have a corresponding meaning.

It should be noted though that if data occur whose $y$-deviation is 0 (practically never in a grouped table) one-half of the sum of these $x$'s should be added to $\underset{+y}{Sx}$. In the $\underset{+x}{S}$ a similar addition should be made in case observations occur in which $x$ is zero. (See Table IV.)

The formulas (1) and all following ones will be proved in the appendix to this article.[4]

[4] Using $r_1$ and $r_2$ of (1) the regression lines are $y = \mathrm{r}_2x$ and $x = r_1y$. They are those straight lines which fit the data best according to the method of least squares, if the weight of the deviations is taken inversely proportional to the absolute value of the variable. Taking $x$ for instance as the independent variable, $r_2$ is the value of $m$ which minimizes $S\frac{1}{|x|}(y - mx)^2$ (the sum extended over all data $x\ y$).

The standard deviations of $r_1$ and $r_2$ are

$$\sigma^2_{r_1} = \frac{r_1^{2\pi}}{2N}(1 + m(m - 2r)) \qquad \text{where } m = \frac{\underset{+x}{Sx}}{\underset{+y}{Sx}}$$

(2)

$$\sigma^2_{r_2} = \frac{r_2^{2\pi}}{2N}(1 + n(n - 2r)) \qquad \text{where } n = \frac{\underset{+y}{Sy}}{\underset{+x}{Sy}}$$

We are now going to illustrate the computation of $r$ and for this purpose we shall use a table of Pearson's which gives the correlation between the heights of fathers and daughters.

The totals at the right and lower end of the table are first computed and the bracketed numbers are the sums of the numbers that precede. The means are

$$\bar{x} = \frac{1659.5 - 1179}{1376} = +\frac{480.5}{1376}$$

and

$$\bar{y} = \frac{1650.9 - 1390}{1376} = +\frac{260.5}{1376}$$

whose signs determine on which side of the working mean to "quarter" the table. This quartering is done in Table 1 by the lines *vv* and *hh*. Then the totals above the heavy horizontal separating line *hh* and those to the left of the vertical separating line *vv* are found, e.g. 2, 4.5, 7.25, ⋯ and .5, .5, 0, ⋯ . Multiplying these totals by the respective class marks, we find the outside lines: 18, 36, 50.75, ⋯ and 5.5, 5, 0, ⋯ .

$\underset{-y}{Sx}$ is now = 1107.5 − 420.5 = 687, and an adjustment for the fact that a working mean has been used has yet to be made. This adjustment is $\bar{x}N_{-y}$ where $N_{-y}$ is the number of negative $y$'s. ($N_{-y} = 728$.)

We have therefore for the adjusted values

$$\underset{-y}{Sx} = 1107.5 - 420.5 + \frac{260.5}{1376}\cdot 728 = 825.07$$

$$\underset{-y}{Sy} = 1179 + \frac{480.5}{1376}\cdot 728 = 1433.21$$

$$r_1 = .5757 \qquad r_2 = .5170$$

$$r = .546$$

The standard deviations, according to the formulas (2) are

$$\sigma_{r_1} = .031 \qquad \sigma_{r_2} = .027$$

# TABLE 1

*Correlation between Heights of Fathers and Daughters*

x → Height of Fathers    y ↓ Height of Daughters

In Inches

v

| | | | 18 | 36 | 50.75 | 66 | 206.25 | 183 | 202.5 | 212.5 | 132.5 | (1107.5) | 86.25 | 116.5 | 115.5 | 69 | 23.75 | 7.5 | 1.75 | | (420.25) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Totals above line | | 2 | 4.5 | 7.25 | 11 | 41.25 | 45.75 | 67.5 | 106.25 | 132.5 | | 86.25 | 58.25 | 38.5 | 17.25 | 4.75 | 1.25 | .25 | | Totals | |
| | Totals left of line | | −9 | −8 | −7 | −6 | −5 | −4 | −3 | −2 | −1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | (728) | |
| 5.5 | .5 | −11 | | | | | .25 | .25 | | | | | | | | | | | | | .5 | 5.5 |
| 5 | .5 | −10 | | | | | .25 | .25 | | | | | | | | | | | | | .5 | 5 |
| | | −9 | | | | | | | | | | | | | | | | | | | | |
| 8 | 1 | −8 | | | | | | | | 1 | | | | | | | | | | | 1 | 8 |
| 31.5 | 4.5 | −7 | .25 | .25 | | .25 | 1.25 | .5 | | 1 | .5 | .5 | | | | | | | | | 4.5 | 31.5 |
| 81 | 13.5 | −6 | .25 | .25 | .5 | 1.5 | 4.5 | 1 | 1.5 | 1.5 | 2.5 | | .5 | .5 | | | | | | | 14.5 | 88 |
| 73.75 | 14.75 | −5 | .25 | .75 | .5 | .75 | .75 | 1 | 1.75 | 1.25 | 5 | 2.75 | .5 | .25 | | | | | | | 15.5 | 77.5 |
| 163 | 40.75 | −4 | .5 | 1 | 2 | | 6 | 4.75 | 5 | 6.25 | 11.75 | 3.5 | 3.5 | 2 | 1.75 | .5 | | | | | 48.5 | 194 |
| 240.75 | 80.25 | −3 | 75 | .75 | | 2.5 | 8 | 6.25 | 12.5 | 18.25 | 20.25 | 11 | 9 | 4.75 | 2.5 | 1.25 | 1.25 | | | | 99 | 297 |
| 212.5 | 106.25 | −2 | | .5 | 1.75 | 2 | 9.75 | 11.5 | 13 | 23.75 | 23.75 | 20.25 | 16.5 | 10.25 | 4.25 | 3 | 1.25 | | | | 141.5 | 283 |
| 131.75 | 131.75 | −1 | | 1 | 2.25 | 2 | 4.5 | 12 | 22.75 | 26 | 33 | 28.25 | 24.75 | 14.25 | 13.75 | 4.75 | .75 | .5 | | | 190.5 | 190.5 |
| (952.75) | | 0 | | | .25 | 2 | 6 | 8.25 | 11 | 22.75 | 35.75 | 37.25 | 31.5 | 26.25 | 16.25 | 7.75 | 1.5 | .75 | .25 | | 212 | (1179) |
| 87.25 | 87.25 | 1 | | | .25 | 2.5 | 1.75 | 3.25 | 9.25 | 23 | 18.75 | 28.5 | 33 | 34.25 | 24.5 | 11.75 | 5.5 | 1 | .25 | 1 | 198.5 | 198.5 |
| 108.5 | 54.25 | 2 | | | | .5 | 1 | .5 | 11 | 12.25 | 9.25 | 19.75 | 30 | 26.5 | 22.25 | 15 | 4.75 | 3.75 | 2 | 1 | 159.5 | 319 |
| 113.25 | 37.75 | 3 | | | | .5 | .5 | 1.5 | 3.25 | 7.25 | 8.75 | 16 | 26.25 | 26.75 | 20.5 | 18.5 | 7.75 | 4.25 | .25 | .5 | 142.5 | 427.5 |
| 71 | 17.75 | 4 | | | | | | | 1 | 5.75 | 7 | 4 | 14.25 | 13.25 | 12 | 11.25 | 4.5 | 3.75 | .75 | | 77.5 | 310 |
| 22.5 | 5.5 | 5 | | | | | .25 | .25 | .25 | .25 | 1.5 | 3 | 5.5 | 4.25 | 5.75 | 5.25 | 3.75 | 2.5 | 1.5 | 2 | 36 | 180 |
| 9 | 1.5 | 6 | | | | | .25 | .25 | .25 | .25 | .25 | .25 | 1 | 2.5 | 6.5 | 2.25 | 2.75 | 2 | 1 | | 19.5 | 117 |
| | | 7 | | | | | | | | | | | 1.75 | .25 | 4.5 | .75 | 1.25 | .75 | .25 | | 9.5 | 66.5 |
| | | 8 | | | | | | | | | | | .5 | | .5 | .5 | 1.5 | .75 | .25 | | 4 | 32 |
| | | 9 | | | | | | | | | | | 1 | | | | | | | | 1 | 9 |
| (416.5) | Totals | (728.5) | 2 | 4.5 | 7.5 | 14.5 | 45 | 51.5 | 92.5 | 155 | 178 | 175 | 199.5 | 166 | 135 | 82.5 | 36.5 | 20 | 6.5 | 4.5 | (1376) | (1659.5) |
| | | | 18 | 36 | 52.5 | 87 | 225 | 206 | 277.5 | 310 | 178 | (1390) | 199.5 | 332 | 405 | 330 | 182.5 | 120 | 45.5 | 36 | (1650.5) | |

v

Working Mean    x = 67.5    y = 63.5

Class width 1 Inch

The standard deviation of $r^2 = r_1 \times r_2$ has to be estimated by using the general formula for the standard deviation of the product $c$ of two variables $a$ and $b$;

$$\frac{\sigma_c^2}{c^2} = \frac{\sigma_a^2}{a^2} + \frac{\sigma_b^2}{b^2} + \frac{2R\sigma_a\sigma_b}{ab}$$

R being the correlation coefficient between $a$ and $b$. Since $-1 < R < +1$, substitution of these limits for $R$ leads to the inequalities

$$\left(\frac{\sigma_a}{a} - \frac{\sigma_b}{b}\right)^2 < \frac{\sigma_c^2}{c} < \left(\frac{\sigma_a}{a} + \frac{\sigma_b}{b}\right)^2$$

putting $a = r_1$, $b = r_2$, $c = r^2$ we have

$$\frac{\sigma_{r_1}}{r_1} - \frac{\sigma_{r_2}}{r_2} < \frac{\sigma_{r^2}}{r} < \frac{\sigma_{r_1}}{r_1} + \frac{\sigma_{r_2}}{r_2}$$

Considering the relation $\sigma_r = \frac{\sigma_{r_2}}{2r}$
we have $2r(\sigma_{r_1} r_2 - \sigma_{r_2} r_1) < \sigma_r < 2r(\sigma_{r_1} r_2 + \sigma_{r_2} r_1)$
from which we derive with sufficient approximation

$$\sigma_r < \cdot 030$$

A slightly different arrangement for computing $r$ has been made in the following table.

TABLE II

*Correlation between diameter of the stem and length of the lonest flower petal of Trientalis europaea**

| | PS | 3 | 15 | 34 | 45 | 30 | 6 | 2 | 0 | 0 | 0 | 0 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PS | | −4 | −3 | −2 | −1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| 1 | −4 | 1 | | | | | | | | | | | 1 |
| 7 | −3 | 1 | 4 | 1 | 1 | | | | | | | | 7 |
| 29 | −2 | 1 | 9 | 16 | 3 | 1 | | | | | | | 30 |
| 33 | −1 | | 2 | 9 | 22 | 9 | 2 | 1 | | | | | 45 |
| 27 | 0 | | | 8 | 19 | 20 | 4 | 1 | | | | | 52 |
| 8 | 1 | 1 | | | 7 | 18 | 12 | 6 | 4 | | | | 48 |
| 1 | 2 | | | | 1 | 8 | 9 | 3 | 2 | 1 | | | 24 |
| | 3 | | | | | | 3 | 6 | 4 | 1 | | | 14 |
| | 4 | | | | | | | 2 | 2 | 1 | 2 | | 7 |
| | 5 | | | | | | | | | 1 | 3 | | 4 |
| | 6 | | | | | | | | | 1 | | 1 | 2 |
| | Total | 4 | 15 | 34 | 53 | 56 | 30 | 19 | 12 | 5 | 5 | 1 | 234 |

* E. Czuber: Die statistischen Forschungsmethoden, Wien, 1921.

## TABLE III

$x$ = Diameter of the stem.
$y$ = Length of the longest flower petal in millimeters.
Working mean, $x_m = .825$, $y_m = 34.5$.
Class width of $x = .4$ mm. of $y = 6$ mm.

| $x$ | Total times $x$ | P.S. times $x$ | $y$ | Total times $y$ | P.S. times $y$ |
|---|---|---|---|---|---|
| −4 | 16 | 12 | −4 | 4 | 4 |
| −3 | 45 | 45 | −3 | 21 | 21 |
| −2 | 68 | 68 | −2 | 60 | 58 |
| −1 | 53 | 45 | −1 | 45 | 33 |
| 0 | (182) | (170) | 0 | (130) | (116) |
| 1 | 30 | 6 | 1 | 48 | 8 |
| 2 | 38 | 4 | 2 | 48 | 2 |
| 3 | 36 | 0 | 3 | 42 | |
| 4 | 20 | 0 | 4 | 28 | |
| 5 | 25 | 0 | 5 | 20 | |
| 6 | 6 | 0 | 6 | 12 | |
| | (155) | (10) | | (198) | (10) |
| Mean | −27 | | | +68 | |

The P.S. columns are the partial sums as explained in the previous table. The work of multiplying the totals by the class marks and of adding them has been separated here from the table.

We obtain $N = 234$, $N_{-x} = 106$, $N_{-y} = 135$

$$r_1 = \frac{170 - 10 - \frac{27}{234} \times 135}{130 + \frac{68}{234} \times 135} = .805$$

$$r_2 = \frac{116 - 10 + \frac{68}{234} \times 106}{182 - \frac{27}{234} \times 106} = .834$$

$$r = .82$$

Pearson's coefficient for this table is $r = .83$.

Finally we illustrate by a small non-grouped table where the partial sums can be written down immediately.

TABLE IV

*Correlation between Ages of Husband and Wife*

| Age of Husband | Age of Wife | Deviation Husband | Deviation Wife |
|---|---|---|---|
| 22 | 18 | −8 | −8 |
| 24 | 20 | −6 | −6 |
| 26 | 20 | −4 | −6 |
| 26 | 24 | −4 | −2 |
| 27 | 22 | −3 | −4 |
| 27 | 24 | −3 | −2 |
| 28 | 27 | −2 | +1 |
| 28 | 24 | −2 | −2 |
| 29 | 21 | −1 | −5 |
| 30 | 25 | 0 | −1 |
| 30 | 29 | 0 | +3 |
| 30 | 32 | 0 | +6 |
| 31 | 27 | +1 | +1 |
| 32 | 27 | +2 | +1 |
| 33 | 30 | +3 | +4 |
| 34 | 27 | +4 | +1 |
| 35 | 30 | +5 | +4 |
| 35 | 31 | +5 | +5 |
| 36 | 30 | +6 | +4 |
| 37 | 32 | +7 | +6 |
| Ave 30 | 26 | | |

Here 0-deviations occur in the third column. Hence[5]

$$\underset{+x}{Sy} = 26 + \tfrac{1}{2} \times 8 = 30, \qquad \underset{+x}{Sx} = 33, \qquad \underset{+y}{Sx} = 31, \qquad \underset{+y}{Sy} = 36,$$

$$r_1 = .86, \quad r_2 = .91, \quad r = .88 \text{ (Pearson's } r = .86)$$

**Appendix**

Proof of formula (1), page 1. The following notations will be used:

$(f(x))^0$ = probable value of $f(x)$

$(f(y))^0_x$ = probable value of $f(y)$ for a fixed $x$.

$$sgx = \text{sign of } x = \frac{x}{|x|} \text{ for } x \neq 0. \quad sgx = \begin{matrix} +1 \\ 0 \\ -1 \end{matrix} \text{ if } x \gtreqless 0.$$

[5] See page 7.

The assumption of linear regression means that

$$y_x^0 - y^0 = r_{y:x}(x - x^0) \tag{4}$$

We multiply both sides of (4) by some arbitrary function $\phi(x)$ of $x$ and get

$$(y_x^0 - y^0)\phi(x) = r_{y:x}(x - x^0)\phi(x).$$

Both sides are functions of $x$. We shall take their probable values for all $x$'s.

Now, for a fixed $x$, $y_x^0\phi(x) = (y\phi(x))_x^0$ and the probable value of $(y\phi(x))_x^0$ for all $x$'s is equal to the total probable value $(y\phi(x))^0$. So we have

$$(y\phi(x))^0 - (y^0\phi(x))^0 = r_{y:x}((x - x^0)\phi(x))^0$$

$$r_{y:x} = \frac{((y - y^0)\phi(x))^0}{((x - x^0)\phi(x))^0} \tag{5}$$

If now we take $x^0y^0$ as the origin, we get

$$r_{y:x} = \frac{(y\phi(x))^0}{(x\phi(x))^0}$$

and similarly

$$r_{x:y} = \frac{(x\phi_1(y))^0}{(y\phi_1(y))^0}$$

where $\phi_1$ is another arbitrary function.

Replacing the probable values by the respective arithmetic means we get

$$r_{y:x} = \frac{Sy\phi(x)}{Sx\phi(x)} \quad \text{and} \quad r_{x:y} = \frac{Sx\phi_1(y)}{Sy\phi_1(y)} \tag{6}$$

with $\bar{x}$, $\bar{y}$ as the origin.

By a suitable choice of the still arbitrary functions $\phi$ and $\phi_1$, we may derive all the various expressions for regression coefficients. Taking, for instance, $\phi(x) = x$, $\phi_1(y) = y$, we get Pearson's expressions. Taking $\phi(x) = sg(x - \alpha_1)$, $\phi_1(y) = sg(y - \alpha_2)$, $\alpha_1$ and $\alpha_2$ being constants, we have

$$r_{y:x} = \frac{Sy\ sg(x - \alpha_1)}{Sx\ sg(x - \alpha_1)}, \quad r_{x:y} = \frac{Sx\ sg(y - \alpha_2)}{Sy\ sg(y - \alpha_2)} \tag{7}$$

and if we make $\alpha_1 = \alpha_2 = 0$

$$r_{y:x} = \frac{Sy\ sg\ x}{Sx\ sg\ x}, \quad r_{x:y} = \frac{Sx\ sg\ y}{Sy\ sg\ y} \tag{8}$$

Since $Sx = Sy = 0$, we can add $Sy$ or $Sx$ to the numerators and denominators. Adding $Sy$ to the numerator, $Sx$ to the denominator and multiplying both sides of the fraction by $\frac{1}{2}$ we get

$$r_{y:x} = \frac{\frac{1}{2}Sy(sg(x - \alpha_1) + 1)}{\frac{1}{2}Sx(sg(x - \alpha_1) + 1)} \tag{9}$$

Instead of (9) we can write

$$(10) \qquad r_{y:x} = \frac{\underset{x>\alpha_1}{S}\, y + \frac{1}{2}\underset{x=\alpha_1}{S}\, y}{\underset{x>\alpha_1}{S}\, x + \frac{1}{2}\underset{x=\alpha_1}{S}\, x}$$

since the operations of (9) multiply the $y$ ordinates by 0, $\frac{1}{2}$, 1 according as the $x$'s are $\lesseqgtr \alpha_1$.

The expression (10), with a suitable choice of $\alpha_1$ should be used for the purpose of numerical calculation of $r$. For instance, when calculating $r$ from the data of Table IV, we took $\alpha_1 = \alpha_2 = 0$ and had

$$r_{y:x} = \frac{\underset{+x}{S}y + \frac{1}{2}\underset{x=0}{S}\, y}{\underset{+x}{S}x}$$

When dealing with data which are arranged in a grouped table (Tables I and II) we take $\alpha_1$ equal to the $x$-ordinate of that classline which is nearest to the mean. $\left(\text{In Table I } \alpha_1 = .5 - \frac{480.5}{1376}\right)$.[6] With that choice of $\alpha_1$ the sums $\underset{x=\alpha_1}{S}$ disappear and the sums $\underset{x>\alpha_1}{S}$ are equivalent to the corresponding sums $\underset{+x}{S}$. Hence we have

$$(11) \qquad r_{y:x} = \frac{\underset{+x}{S}y}{\underset{+x}{S}x} \quad \text{and similarly} \quad r_{x:y} = \frac{\underset{+y}{S}x}{\underset{+y}{S}y}$$

Instead of (9) we can also write

$$(9a) \qquad r_{y:x} = \frac{\frac{1}{2}Sy(sg(x-\alpha_1)-1)}{\frac{1}{2}Sx(sg(x-\alpha_1)-1)}$$

This leads to

$$(11a) \qquad r_{y:x} = \frac{\underset{-x}{S}y}{\underset{-x}{S}x} \quad \text{and} \quad r_{x:y} = \frac{\underset{-y}{S}x}{\underset{-y}{S}y}$$

[6] It is desirable to chose the absolute values of the $\alpha$'s small so that the maximum number of data enter into the calculation of $r$. However, to take $\alpha_1 = \alpha_2 = 0$ would necessitate a division of the middle arrays of a grouped table, a laborious process. Hence the choice of the $\alpha$'s as described above.

Proof of the standard deviations of Formula (2).

In my article on standard deviations and correlations of moments[7] the standard deviations of the expressions used in this article have been derived.

In the following, the notation of the Metron article just referred to will be used. We use the symbols:

$$\begin{aligned} P_{m,n} &= \sum x^m \cdot y^n \\ P_{/m,n} &= \sum x^m sgxy^n \\ P_{m,/n} &= \sum x^m y^n sgy \\ P_{/m,/n} &= \sum x^m sgxy^n sgy \end{aligned}$$

The summations indicated extend over all observations. The true or probable values of the same expressions are indicated by using $p$ instead of $P$.

$$r_{x:y} = r_1 = \frac{P_{1/0}}{P_{0/1}}$$

We derive the standard deviations by defining the deviations as first variations.

$$\log r_1 = \log P_{1/0} - \log P_{0/1}$$

$$\frac{\delta r_1}{r_1'} = \frac{\delta P_{1/0}}{p_{1/0}} - \frac{\delta P_{0/1}}{p_{0/1}}$$

$$\sigma r_1^2 = [(\delta r_1)^2]^0 = (r_1')^2 \left[\left(\frac{\delta P_{1/0}}{p_{1/0}} - \frac{\delta P_{0/1}}{p_{0/1}}\right)^2\right]^0 \tag{12}$$

The probable values of the terms on the right hand side of the last equation are derived on pages 17–19 and listed on pages 32–33 of the Metron article referred to. The proofs which imply essentially a process of variation of Stieltje's integrals will not be given here. From pages 32–33 we take

$$\begin{gathered} [(\delta P_{1/0})^2]^0 = \frac{p_{20} - p_{1/0}^2}{N}, \qquad [(\delta P_{0/1})^2]^0 = \frac{p_{02} - p_{0/1}^2}{N} \\ [(P_{1/0}\delta P_{0/1}]^0 = \frac{p_{11} - p_{1/0}p_{0/1}}{N} \end{gathered} \tag{13}$$

so that

$$\sigma_{r_1}^2 = \frac{1}{N}(r_1')^2\left[\frac{p_{20}}{p_{1/0}^2} + \frac{p_{02}}{p_{0/1}^2} - \frac{2p_{11}}{p_{1/0}p_{0/1}}\right] \tag{14}$$

Assuming Gaussian distribution, we can put

$$p_{20} = \frac{\pi}{2}p_{/10}^2 \qquad p_{02} = \frac{\pi}{2}p_{0/1}^2 \qquad p_{11} = r\sqrt{p_{02}p_{20}} = r\frac{\pi}{2}p_{/10}p_{0/1}$$

[7] Felix Bernstein: "Die mittleren Fehlerquadrate und Korrelationen der Potenzmomente und ihre Anwendung auf Funktionen der Potenzmomente," Metron, Vol. X, N. 3 Nov. 1932).

Hence

$$\sigma_{r_1}^2 = \frac{1}{N}\cdot\frac{\pi}{2}(r_1')^2\left(1 + \frac{p_{/10}^2}{p_{1/0}^2} - 2r\frac{p_{/10}}{p_{1/0}}\right) \tag{15}$$

Replacing the theoretical values by their corresponding empirical values, we have

$$\sigma_{r_1}^2 = \frac{\pi r_1^2}{2N}(1 + m^2 - 2rm) \qquad \text{where } m = \frac{Sx\,sg\,x}{Sx\,sg\,y} \tag{16}$$

The formula for $\sigma_{r_1}^2$ has been derived here for the value of $r_1$ as given by (8) i.e. $r_1 = \dfrac{Sx\,sg\,y}{Sy\,sg\,y}$. In fact, we used $r_1 = \dfrac{Sx\,sg\,(y - \alpha)}{Sy\,sg\,(y - \alpha)}$ in the examples in the article, and $\alpha$ had some value absolutely smaller than .5. To use equation (16) for the standard deviation of $r_1$ is within the limits of the required degree of accuracy; hence we shall disregard the difference. In a later paper the standard deviation of $r_1$ for any $\alpha$ will be derived by using the method described in the Metron article, for a different purpose.

To prove the statement in the footnote to page 7

To find the value of $r_2$ that makes

$$Sf(x)\,(y - r_2x)^2 \text{ a minimum.}$$

By differentiating we get

$$Sf(x)(y - r_2x)x = 0$$

$$r_2 = \frac{Sxf(x)y}{Sxf(x)x}$$

If $f(x) = 1$ we get Pearson's coefficient.

If $f(x) = \dfrac{1}{|x|}$ $(x \neq 0)$ we get

$$r_2 = \frac{S\dfrac{x}{|x|}y}{S\dfrac{x}{|x|}x} = \frac{Sy\,sg\,x}{Sx\,sg\,x}$$

New York University,
Departments of Anatomy of the Graduate School and the College of Dentistry.