Biometrika Trust

On the Measurement of the Influence of "Broad Categories" On Correlation

Author(s): Karl Pearson

Source: *Biometrika*, Vol. 9, No. 1/2 (Mar., 1913), pp. 116-139

Published by: [Biometrika Trust](#)

Stable URL: <http://www.jstor.org/stable/2331803>

Accessed: 17/06/2014 23:25

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
<http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Biometrika Trust is collaborating with JSTOR to digitize, preserve and extend access to *Biometrika*.

<http://www.jstor.org>

ON THE MEASUREMENT OF THE INFLUENCE OF “BROAD CATEGORIES” ON CORRELATION.

BY KARL PEARSON, F.R.S.

(1) By a “broad category” I understand one of a finite small number of groups into which we class a variable. For example: we may divide General Health into the categories Very Robust, Robust, Normally Healthy, Rather Delicate, Delicate and Very Delicate. These categories may be verbally defined or have their boundaries determined by quantitative limits as when we state that the limits of the Delicate coincide with so many weeks of sickness or of absence from work in the year. Again we may put into four to six classes the competitors in an examination, and the boundaries to these classes may be really percentages of marks gained. Such broad categories are very common not only in social investigations, but also in psychological records, and quite recently Dr G. A. Jaederholm, a Swedish psychologist, wrote to me asking what was the correlation between the true quantitative value of a variate in any individual and that individual’s category or class-mark. The answer is an obvious one, but I do not know that I have seen it stated or any discussion of it given. It of course assumes that at the back of the categorical classification a true quantitative value lies.

Suppose a population of N individuals divided into p classes, and that C_s is the class-mark of an individual in the s th category, whose true variate is x . The problem is what is the correlation of x and the class-mark. Let $\bar{x}_s$ be the mean variate of the group of n_s individuals who fall into the s th class. Then it is just as reasonable to call $\bar{x}_s$ the class-mark as C_s , for given one the other is fixed. We really want then the correlation of x and $\bar{x}_s$.

We may either find this directly* or indirectly, and the latter is the easier course. Clearly if $\tilde{x}$ be the mean value of x for a given class, then

$$\tilde{x} = \bar{x}_s,$$

or

$$\tilde{x} - \bar{x} = \bar{x}_s - \bar{x},$$

* Let x within the class = $\bar{x}_s + x_s'$, S = sum for classes and Σ = sum within class. Then

$$S\Sigma (n_x x \bar{x}_s) = S\Sigma \{n_x \bar{x}_s (\bar{x}_s + x_s')\} = S (n_s \bar{x}_s^2) + S \{\bar{x}_s \Sigma (n_x x_s')\},$$

but

$$\Sigma (n_x x_s') = 0,$$

$$S\Sigma (n_x x \bar{x}_s) = S (n_s \bar{x}_s^2),$$

$$r_x \bar{x}_s = \frac{S\Sigma (n_x x \bar{x}_s)}{N \cdot \sigma_{\bar{x}_s} \times \sigma_x} = \frac{S (n_s \bar{x}_s^2)}{N \sigma_{\bar{x}_s} \sigma_x} = \frac{\sigma_{\bar{x}_s}^2}{\sigma_{\bar{x}_s} \sigma_x} \\ = \sigma_{\bar{x}_s} / \sigma_x \text{ as before.}$$

is the equation to the regression line; therefore

$$\frac{\tilde{x} - \bar{x}}{\sigma_x} = \frac{\sigma_{\tilde{x}_s}}{\sigma_x} \times \frac{\bar{x}_s - \bar{x}}{\sigma_{\tilde{x}_s}}.$$

Whence it follows that

$$r_{xC_s} = \frac{\sigma_{\tilde{x}_s}}{\sigma_x} \dots\dots\dots (i),$$

or the correlation of a variate with its class-mark is the ratio of the standard deviation of the means of class-marks to the standard deviation of the variate.

It follows from this that in classifying a variable into broad categories the test of their efficiency, as far as number and arrangement are concerned, lies in the standard deviation of their means not differing widely from the standard deviation of the variate itself.

$$(2) \quad \text{Clearly} \quad \sigma_{\tilde{x}_s}^2 = S(n_s \bar{x}_s^2)/N \dots\dots\dots (ii),$$

where S is a sum involving all classes.

If before classifying into broad categories we have made a quantitative determination of these classes on a sample frequency, the values of $\bar{x}_s$'s can be determined. If this has not been done, or cannot be done, we are bound to assume some form of frequency distribution. Suppose we take a normal distribution, then we know that

$$n_s \bar{x}_s = \frac{N}{\sqrt{2\pi}\sigma_x} \int_{x_{s_1}}^{x_{s_2}} x e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}} dx = N\sigma_x(z_{s_1} - z_{s_2}),$$

where
$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x/\sigma_x)^2},$$

and can be found from Sheppard's Tables of the ordinates of the normal curve as soon as n_s , etc. are known, for the z 's are the ordinates at start and finish of the s th class, reduced by the factor N/σ_x .

Hence

$$\sigma_{\tilde{x}_s}^2 = S \left\{ \frac{N}{n_s} (z_{s_1} - z_{s_2})^2 \times \sigma_x^2 \right\},$$

or

$$r_{xC_s} = \sqrt{S \left\{ \frac{N}{n_s} (z_{s_1} - z_{s_2})^2 \right\}} \dots\dots\dots (iii).$$

Thus r_{xC_s} can be found at once from Sheppard's Tables, when the totals of the broad classes are known.

(3) Let us now suppose a second variate y and assume that the correlation of x and y for practical purposes is linear. Then clearly since a given x will have a constant class-mark, the correlation of y and C_x for a constant x is zero; that is to say that the partial correlation coefficient

$$r_{yC_x} = \frac{r_{yC_x} - r_{xy} r_{xC_x}}{\sqrt{1 - r_{xy}^2} \sqrt{1 - r_{xC_x}^2}} = 0.$$

Hence

$$r_{yC_x} - r_{xy} r_{xC_x} = 0,$$

or

$$r_{xy} = \frac{r_{yC_x}}{r_{xC_x}} \dots\dots\dots (iv).$$

In other words, to find the true correlation of x and any other variate y , divide the correlation of y and the class-mark of x , by the correlation of x with its class-mark.

From (iv) we can deduce the correction for number of arrays when we find η_{xy} the correlation ratio of x and y on the supposition that r_{xy} equals sufficiently closely η_{xy} , i.e. when we "find r by η methods." Let H_{yC_x} equal the value of η found when y is finely classified and the mean of an array can be determined for y , but the arrays of x are broad classes*. Then

$$\eta_{xy} = \frac{H_{yC_x}}{r_{xC_x}} = \frac{H_{yC_x}}{\sqrt{S(n_s \bar{x}_s^2)/(N\sigma_x^2)}} \dots\dots\dots (v),$$

$$= \frac{H_{yC_x}}{\sqrt{S \frac{N}{n_s} (z_{s_1} - z_{s_2})^2}} \dots\dots\dots (vi),$$

if we suppose the distribution normal.

(4) Now let us consider two variables x and y given by their class-marks C_x and C_y . If we correlate y with any variable u we have at once by (iv)

$$r_{uy} = \frac{r_{C_y u}}{r_{yC_y}} \dots\dots\dots (vii).$$

Now as u is quite arbitrary put it equal to the class-mark of x or C_x , then

$$r_{yC_x} = \frac{r_{C_y C_x}}{r_{yC_y}} \dots\dots\dots (viii).$$

Substitute this in (iv) and we have†

$$r_{xy} = \frac{r_{C_y C_x}}{r_{xC_x} r_{yC_y}} \dots\dots\dots (ix).$$

* The primary establishment of equation (v) is due to "Student." His paper published in this number reached me, as I was writing this paper. I had obtained (iv) and (ix) and used them to correct contingency, but not to correct η .

† A somewhat different proof of this formula may be obtained as follows: the partial correlation $xy \cdot r_{C_x C_y}$ is clearly zero, for when x and y are constant C_x and C_y do not vary. Hence

$$r_{C_x C_y} (1 - r_{xy}^2) - r_{xC_x} r_{yC_y} - r_{yC_y} r_{yC_x} + r_{xy} (r_{xC_x} r_{yC_y} + r_{xC_y} r_{yC_x}) = 0.$$

But

$$r_{xy} = r_{yC_x} / r_{xC_x} = r_{xC_y} / r_{yC_y},$$

and if we substitute for r_{yC_x} and r_{xC_y} we have

$$r_{C_x C_y} (1 - r_{xy}^2) - (r_{xy} - r_{xy}^2) r_{yC_y} r_{xC_x} = 0,$$

or

$$r_{xy} = \frac{r_{C_x C_y}}{r_{yC_y} r_{xC_x}}, \text{ as before.}$$

Thus $1/(r_{xC_x} r_{yC_y})$ appears to be the corrective factor when we group the variates x and y both into broad categories, the ranges of which are of any nature.

We have

$$r_{C_y C_x} = \frac{S(n_{st} \bar{x}_{st} \bar{y}_{st})/N}{\sqrt{S(n_s \bar{x}_s^2)/N} \sqrt{S(n_t \bar{y}_t^2)/N}} \dots\dots\dots(\text{x}),$$

and

$$r_{xy} = \frac{S\left(\frac{n_{st} \bar{x}_{st} \bar{y}_{st}}{N \sigma_x \sigma_y}\right)}{S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) \times S\left(\frac{n_t \bar{y}_t^2}{N \sigma_y^2}\right)} \dots\dots\dots(\text{xi}),$$

where n_{st} is the frequency of the cell in the s th category of x and t th category of y .

It is obvious that r_{xy} takes its usual value when the classes of x and y become very numerous, for both factors in the denominator are then unity.

In the particular case of normal distributions, on the assumption that the product $\bar{x}_{st} \bar{y}_{st}$ is sufficiently close to $\bar{x}_s \bar{y}_t$ to be replaced by it:

$$r_{xy} = \frac{S\left\{\frac{n_{st} N}{n_s n_t} (z_{s_1} - z_{s_2})(z_{t_1} - z_{t_2})\right\}}{S\left\{\frac{N}{n_s} (z_{s_1} - z_{s_2})^2\right\} S\left\{\frac{N}{n_t} (z_{t_1} - z_{t_2})^2\right\}} \dots\dots\dots(\text{xii}),$$

which admits of fairly ready determination from Sheppard's Tables.

(5) An approximate value of r_{xC_x} can be reached in the following manner, if we suppose the range of the broad classes are all equal and given by h .

Assume a parabola $y = a + bx + cx^2$

to give by its area the three class frequencies n_{s-1} , n_s , n_{s+1} , the origin being at the start of the n_{s-1} class. Its equation will be found to be

$$y = \frac{2n_{s+1} - 7n_s + 11n_{s-1}}{6h} - \frac{n_{s+1} - 3n_s + 2n_{s-1}}{h} \left(\frac{x}{h}\right) + \frac{n_{s+1} - 2n_s + n_{s-1}}{2h} \left(\frac{x}{h}\right)^2 \dots\dots(\text{xiii}).$$

But
$$n_s \bar{x}_s' = \int_h^{2h} yx dx = \frac{h}{24} (n_{s+1} + 36n_s - n_{s-1}),$$

or
$$\bar{x}_s' = \frac{3}{2}h + \frac{h}{24} \frac{n_{s+1} - n_{s-1}}{n_s}.$$

Thus, if x_s' be the distance from the origin to the *middle* of the s th class,

$$\bar{x}_s' - x_s' = \frac{1}{24} h \frac{n_{s+1} - n_{s-1}}{n_s},$$

and if $\bar{x}_s$ and x_s be measured from the mean of the whole population

$$\bar{x}_s = x_s - \frac{1}{24} h \frac{n_{s-1} - n_{s+1}}{n_s},$$

and

$$\begin{aligned} n_s \bar{x}_s^2 &= n_s x_s^2 - \frac{n_s}{12} h x_s \frac{n_{s-1} - n_{s+1}}{n_s} + \frac{h^2}{576} \frac{(n_{s-1} - n_{s+1})^2}{n_s} \\ &= n_s x_s^2 - \frac{h}{12} (x_{s-1} n_{s-1} - x_{s+1} n_{s+1}) - \frac{h^2}{12} n_{s-1} - \frac{h^2}{12} n_{s+1} + \frac{h^2}{576} \frac{(n_{s-1} - n_{s+1})^2}{n_s}. \end{aligned}$$

Summing for all classes and dividing by $N\sigma_x^2$

$$S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) = \frac{S(n_s x_s^2)}{N \sigma_x^2} - \frac{1}{12} \frac{h^2}{\sigma_x^2} - \frac{1}{12} \frac{h^2}{\sigma_x^2} + \frac{h^2}{576 \sigma_x^2} S\left(\frac{(n_{s-1} - n_{s+1})^2}{N n_s}\right).$$

But by Sheppard's Theorem, with contact at the tails of the order we are supposing :

$$\sigma_x^2 = \frac{S(n_s x_s^2)}{N} - \frac{1}{12} h^2.$$

Further, if we suppose y_{s-1} , y_s , y_{s+1} , y_{s+2} to be the bounding ordinates of the classes n_{s-1} , n_s , n_{s+1} of the frequency curve,

$$\alpha_{s-1} = n_{s-1} - y_s h = n_{s-1} - \frac{(-n_{s+1} + 5n_s + 2n_{s-1})}{6} = \frac{1}{6} (4n_{s-1} - 5n_s + n_{s+1}).$$

Similarly $\alpha_{s+1} = n_{s+1} - y_{s+2} h = \frac{1}{6} (n_{s-1} - 5n_s + 4n_{s+1}).$

Here α_{s-1} and α_{s+1} are the excesses of n_1 and n_3 above rectangles ; and we have

$$\alpha_{s-1} - \alpha_{s+1} = \frac{1}{2} (n_{s-1} - n_{s+1}).$$

Thus finally

$$S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) = 1 - \frac{1}{12} \frac{h^2}{\sigma_x^2} + \frac{h^2}{144 \sigma_x^2} S\left(\frac{(\alpha_{s-1} - \alpha_{s+1})^2}{N n_s}\right) \dots\dots\dots(\text{xiv}).$$

Consider the last term ; it may be written

$$S\left\{\frac{\frac{1}{2}(\alpha_s - \alpha_{s-1} + \alpha_{s+1} - \alpha_s)}{n_s} \times \frac{\frac{1}{2}(\alpha_s - \alpha_{s-1} + \alpha_{s+1} - \alpha_s)}{N}\right\} \times \frac{h^2}{36 \sigma_x^2}.$$

Now the numerators are mean differences of what will be, if h be at all small, small areas, and these mean differences are divided in the one case by n_s and the other by N , much larger quantities. Hence the last term is as a rule very much smaller than the second, and we have approximately

$$S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) = 1 - \frac{1}{12} \frac{h^2}{\sigma_x^2} \dots\dots\dots(\text{xv}).$$

Thus we see how rapidly, if h be small with regard to σ_x , the value of the quantity

$$S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right),$$

and therefore of r_{xC_x} , its square root, approaches unity.

(6) It is desirable to illustrate the approach of r_{xC_x} to unity as we increase the number of groups. I have therefore worked out this approach for three and more *symmetrical* groups when (a) we use the approximate formula (xv), (b) we still suppose the classes to have their ranges equal but the frequencies to be given by a normal curve, (c) we assume the frequencies in the classes and not the class ranges to be equal, the frequency being supposed normal, (d) we suppose the frequencies to increase by 50 % at each stage, i.e. to be as 1, 3/2, 9/4, 27/8, etc.

These cases will be sufficient to indicate what sort of frequencies we should take for few classes in order to get the highest correlation between variate and its class-mark. I shall after these theoretical investigations consider a few actual cases of "broad" categories.

Values of r_{xC_x} for Various Groupings.*

No. of Classes	(a) Equal Ranges, Any Frequency. Formula (xv)	(b) Equal Ranges, Normal Frequency	(c) Equal Subfrequencies, Normal Frequency	(d) Increasing Subfrequencies, Normal Frequency
3	·817 ($h=2\sigma$)	·859	·891	·876
4	·901 ($h=1.5\sigma$)	·915	·928	·912
5	·938 ($h=1.2\sigma$)	·943	·947	·928
6	·957 ($h=\sigma$)	·960	·959	·939
8	·978 ($h=.75\sigma$)	·977	·972	—
10	·985 ($h=.6\sigma$)	·985	·979	—
12	·990 ($h=.5\sigma$)	·989	·984	—
14	·992 ($h=.4\sigma$)	·992	·987	—
16	·994 ($h=.3\sigma$)	·994	·990	—
20	·996 ($h=.3\sigma$)	·996	·992	—

It is clear from the fourth column that nothing whatever is gained by exaggerating the frequencies of the extreme or tail sections. Further:

(a) Equal frequencies are better than equal ranges up to about six classes. After six classes it is better to take equal ranges.

(b) After six classes the approximate formula (xv) is amply sufficient in the case of equal ranges to obtain the value of r_{xC_x} .

Thus we have the general rule that up to six broad categories it is desirable to make the frequencies of those classes approximately equal, but beyond this greater

* The number of classes given in (a) and (b) is based upon the range of 6σ covering for such data as we usually deal with, i.e. 500 to 1000 cases, the total frequency.

exactness will be reached by equal ranges. The differences, however, between equal ranges and equal frequencies are after six classes so small—within the value of the usual probable error of random sampling—that either method will give practically quite good results.

(7) It is interesting to note the relation of the present method to the usual correction for grouping in the value of the correlation. In that method we take the groups at their mid-points, and we do not correct the product, but only the standard deviations of the two variates by the usual Sheppard's correction, $\frac{1}{12}$ th of the sub-range squared being subtracted from the raw second moment coefficient. Now we have

$$r_{xy} = S(n_{st} x_s y_t) / \sigma_x \sigma_y,$$

and we have to replace x_s, y_t by $\bar{x}_s$ and $\bar{y}_t$.

But we have seen that approximately with equal ranges

$$\bar{x}_s = x_s - \frac{1}{24} h \frac{n_{s-1} - n_{s+1}}{n_s},$$

$$\bar{y}_t = y_t - \frac{1}{24} k \frac{n_{t-1} - n_{t+1}}{n_t},$$

and
$$S\left\{\frac{n_s}{N} \left(\frac{\bar{x}_s}{\sigma_x}\right)^2\right\} = 1 - \frac{1}{12} \frac{h^2}{\sigma_x^2}, \quad S\left\{\frac{n_t}{N} \left(\frac{\bar{y}_t}{\sigma_y}\right)^2\right\} = 1 - \frac{1}{12} \frac{k^2}{\sigma_y^2}.$$

Now

$$\begin{aligned} S(n_{st} \bar{x}_s \bar{y}_t) &= S(n_{st} x_s y_t) - \frac{1}{24} h S\left\{\frac{n_{s-1} - n_{s+1}}{n_s} n_{st} y_t\right\} \\ &\quad - \frac{1}{24} k S\left\{\frac{n_{t-1} - n_{t+1}}{n_t} n_{st} x_s\right\} + \frac{1}{576} h k S\left\{n_{st} \frac{n_{s-1} - n_{s+1}}{n_s} \frac{n_{t-1} - n_{t+1}}{n_t}\right\} \dots\dots (xvi). \end{aligned}$$

Consider first the last term*, it contains not only the product of hk , but also the product of differences, and is of the form, when we divide by $N\sigma_x\sigma_y$,

$$\frac{1}{36} \frac{h}{\sigma_x} \frac{k}{\sigma_y} S\left\{\frac{n_{st}}{N} \times \frac{\frac{1}{2}(\alpha_s - \alpha_{s-1}) + \frac{1}{2}(\alpha_{s+1} - \alpha_s)}{n_s} \times \frac{\frac{1}{2}(\alpha'_t - \alpha'_{t-1}) + \frac{1}{2}(\alpha'_{t+1} - \alpha'_t)}{n_t}\right\},$$

which will be of the fourth order of small quantities if h/σ_x and k/σ_y be small and may be neglected as compared to the square of small quantities.

* As a matter of fact for Gaussian frequency

$$S\left(n_{st} \frac{n_{s-1} - n_{s+1}}{n_s} \frac{n_{t-1} - n_{t+1}}{n_t}\right) = 4hk S(n_{st} \bar{x}_s \bar{y}_t)$$

to our degree of approximation. Thus the fourth term may be written $\frac{1}{144} h^2 k^2 S(n_{st} x_s y_t)$, which gives

$$\begin{aligned} S(n_{st} \bar{x}_s \bar{y}_t) &= S(n_{st} x_s y_t) \left(1 - \frac{1}{12} h^2 - \frac{1}{12} k^2 + \frac{1}{144} h^2 k^2\right) \\ &= S(n_{st} x_s y_t) \left(1 - \frac{1}{12} h^2\right) \left(1 - \frac{1}{12} k^2\right). \end{aligned}$$

Next, considering the second summation and summing first $n_{st}y_t$ for all values of t , it equals $n_s\bar{y}_s$, where $\bar{y}_s$ is mean of all y 's in array corresponding to x_s . But if the regression be linear $\bar{y}_s = \frac{r\sigma_y}{\sigma_x}x_s$. Thus the second sum equals

$$\begin{aligned} & \frac{1}{24} r \sigma_y \sigma_x \frac{h}{\sigma_x^2} S \{ (n_{s-1} - n_{s+1}) x_s \} \\ &= \frac{1}{24} r \sigma_y \sigma_x \frac{h}{\sigma_x^2} \{ S(x_{s-1} n_{s-1} - x_{s+1} n_{s+1}) + h S(n_{s-1}) + h S(n_{s+1}) \} \\ &= \frac{1}{24} r \sigma_y \sigma_x \frac{h^2}{\sigma_x^2} 2N \end{aligned}$$

as before. But in a term of this order we may put

$$Nr\sigma_y\sigma_x = S(n_{st}x_sy_t).$$

Similarly the third summation

$$= \frac{1}{12} S(n_{st}x_sy_t) \times \frac{k^2}{\sigma_y^2}.$$

Or, finally, $S(n_{st}\bar{x}_s\bar{y}_t) = S(n_{st}x_sy_t) \left(1 - \frac{h^2}{12\sigma_x^2} - \frac{k^2}{12\sigma_y^2}\right)$ (xvi) *bis*.

Thus

$$\frac{S\left(\frac{n_{st}\bar{x}_s\bar{y}_t}{N\sigma_x\sigma_y}\right)}{S\left(\frac{n_s\bar{x}_s^2}{N\sigma_x^2}\right)S\left(\frac{n_t\bar{y}_t^2}{N\sigma_y^2}\right)} = \frac{\frac{1}{\sigma_x\sigma_y}S\left(\frac{n_{st}x_sy_t}{N}\right)\left(1 - \frac{h^2}{12\sigma_x^2} - \frac{k^2}{12\sigma_y^2}\right)}{\left(1 - \frac{1}{12}\frac{h^2}{\sigma_x^2}\right)\left(1 - \frac{1}{12}\frac{k^2}{\sigma_y^2}\right)},$$

or, since we neglect terms of fourth order,

$$\frac{S\left(\frac{n_{st}\bar{x}_s\bar{y}_t}{N\sigma_x\sigma_y}\right)}{S\left(\frac{n_s\bar{x}_s^2}{N\sigma_x^2}\right)S\left(\frac{n_t\bar{y}_t^2}{N\sigma_y^2}\right)} = \frac{S\left(\frac{n_{st}x_sy_t}{N}\right)}{\sigma_x\sigma_y} = r_{xy} \text{(xvii),}$$

the usual value, σ_x and σ_y having of course to be corrected by Sheppard.

We see by (xv) that

$$\sigma_x^2 = S\left(\frac{n_s\bar{x}_s^2}{N}\right) \left(1 + \frac{1}{12}\frac{h^2}{\sigma_x^2}\right) \text{(xvii) } bis,$$

but by Sheppard

$$\sigma_x^2 = S\left(\frac{n_s\bar{x}_s^2}{N}\right) \left(1 - \frac{1}{12}\frac{h^2}{\sigma_x^2}\right).$$

Accordingly

$$S\left(\frac{n_s\bar{x}_s^2}{N}\right) = S\left(\frac{n_s\bar{x}_s^2}{N}\right) \left(1 - \frac{1}{6}\frac{h^2}{\sigma_x^2}\right) \text{(xviii),}$$

which enables us with equal small subranges to use the standard deviation of means or mid-abcissae at our pleasure.

(8) Of course it is absurd in practice to push our results to the extreme* of two categories only, but theoretically it is not without interest to note the results which flow from such an assumption.

* We have seen that $\bar{x}_{st}\bar{y}_{st}$ can only be replaced by $\bar{x}_s\bar{y}_t$ in (xi), or as a special case, (xii) used, provided the subranges are equal and fairly small.

Let the division be into two categories containing $N \times \frac{1}{2}(1 + \alpha)$ and $N \times \frac{1}{2}(1 - \alpha)$ individuals, and let no assumption be made as to the nature of the frequency. Then

$$S\left(\frac{n_s}{N} \bar{x}_s^2\right) = \frac{1}{2}(1 + \alpha) \bar{x}_1^2 + \frac{1}{2}(1 - \alpha) \bar{x}_2^2,$$

but

$$\frac{1}{2}(1 + \alpha) \bar{x}_1 + \frac{1}{2}(1 - \alpha) \bar{x}_2 = 0,$$

and therefore

$$S\left(\frac{n_s}{N} \bar{x}_s^2\right) = \bar{x}_2^2 \frac{\frac{1}{2}(1 - \alpha)}{\frac{1}{2}(1 + \alpha)}.$$

Thus

$$r_{xC_x} = \frac{\bar{x}_2}{\sigma_x} \sqrt{\frac{\frac{1}{2}(1 - \alpha)}{\frac{1}{2}(1 + \alpha)}} \dots\dots\dots(\text{xix}).$$

We note therefore that unless we suppose all the frequency in each category concentrated at one point—a very rare occurrence— r_{xC_x} will always be less than unity, and therefore in correlating class indices a correction will always have to be made. We will consider two cases: (i) when the frequency distribution is a rectangle of length l . In this case $\bar{x}_2$ is $= l \{ \frac{1}{2}(1 + \alpha) \}$ and $\sigma_x = l/\sqrt{3}$, we have then

$$r_{xC_x} = \sqrt{3} \times \frac{1}{2}(1 + \alpha) \times \frac{1}{2}(1 - \alpha) \dots\dots\dots(\text{xx});$$

(ii) the distribution of frequency is supposed to be Gaussian. In this case

$$\bar{x}_2 = \frac{z_\alpha \times \sigma_x}{\frac{1}{2}(1 - \alpha)},$$

and

$$r_{xC_x} = \frac{z_\alpha}{\sqrt{\frac{1}{2}(1 - \alpha) \times \frac{1}{2}(1 + \alpha)}} \dots\dots\dots(\text{xxi}),$$

in other words, r_{xC_x} is the reciprocal of the χ_α already tabled in this number of *Biometrika*, p. 27.

We have the following results:

Value of $\frac{1}{2}(1 + \alpha)$ as percentage of total	Value of r_{xC_x}	
	Gaussian Frequency	Rectangular Frequency
50 %	·798	·866
60 %	·789	·849
70 %	·759	·794
80 %	·700	·693
90 %	·585	·520
95 %	·473	·377
99 %	·268	·172

It will be seen at once how rapidly the corrective factor $1/r_{xC_x}$ rises as the division of the categories becomes more and more unequal. But in both cases very sensible correction is needful and its nature depends on the particular frequency.

Let us now pass to the consideration of the correlation r_{xy} as obtained by correction from the class index correlation, i.e.

$$r_{xy} = \frac{r_{C_x C_y}}{r_x C_x \times r_y C_y}.$$

The difficulty is to know how to determine the product $S(n_{st}\bar{x}_{st}\bar{y}_{st})$ for such a slender division as a fourfold table. If we assume it still to be equal to

$$S(n_{st}\bar{x}_s\bar{y}_t)$$

we can then find r_{xy} .

$$\text{We have} \quad S(n_{st}\bar{x}_s\bar{y}_t) = n_1\bar{x}_1\bar{y}_1 + n_2\bar{x}_2\bar{y}_2 + n_3\bar{x}_3\bar{y}_3 + n_4\bar{x}_4\bar{y}_4,$$

where we take for our fourfold table:

n_1	n_2	$n_1 + n_2$
n_4	n_3	$n_3 + n_4$
$n_1 + n_4$	$n_2 + n_3$	N

But clearly on the above assumption

$$\bar{x}_1 = \bar{x}_4, \quad \bar{x}_2 = \bar{x}_3, \quad \bar{y}_1 = \bar{y}_2, \quad \bar{y}_3 = \bar{y}_4.$$

Thus

$$S(n_{st}\bar{x}_s\bar{y}_t) = n_1\bar{x}_1\bar{y}_1 + n_2\bar{x}_2\bar{y}_1 + n_3\bar{x}_2\bar{y}_4 + n_4\bar{x}_1\bar{y}_4,$$

then since

$$(n_1 + n_4)\bar{x}_1 + (n_2 + n_3)\bar{x}_2 = 0,$$

$$(n_1 + n_2)\bar{y}_1 + (n_3 + n_4)\bar{y}_4 = 0,$$

we easily deduce

$$S(n_{st}\bar{x}_s\bar{y}_t) = \frac{\bar{x}_2\bar{y}_4(n_1n_3 - n_3n_1)}{(n_1 + n_4)(n_1 + n_2)} \dots\dots\dots(\text{xxii}).$$

Similarly we find

$$\left. \begin{aligned} \sigma_{\bar{x}_s}^2 &= \frac{n_2 + n_3}{n_1 + n_4} \bar{x}_2^2 \\ \sigma_{\bar{y}_t}^2 &= \frac{n_3 + n_4}{n_1 + n_2} \bar{y}_4^2 \end{aligned} \right\} \dots\dots\dots(\text{xxiii}).$$

And accordingly

$$\begin{aligned} r_{C_x C_y} &= \frac{S(n_{st}\bar{x}_s\bar{y}_t)}{\sigma_{\bar{x}}\sigma_{\bar{y}}} \\ &= \frac{n_1n_3 - n_2n_4}{\sqrt{(n_1 + n_2)(n_1 + n_4)(n_2 + n_3)(n_3 + n_4)}} \dots\dots\dots(\text{xxiv}). \end{aligned}$$

This is not really the true value of the class index correlation unless we assume $\bar{x}_s\bar{y}_t = \bar{x}_{st}\bar{y}_{st}$, which is not generally true. It is really the correlation of the means used as class-marks, and as shown long ago by me it is the correlation of errors in the means of the two variates, and also equals ϕ the basis of the mean square contingency. If we correct to obtain r_{xy} by the factors

$$r_x C_x = \sigma_{\bar{x}_s}/\sigma_x \quad \text{and} \quad r_y C_y = \sigma_{\bar{y}_t}/\sigma_y,$$

we find

$$r_{xy} = \frac{n_1 n_3 - n_2 n_4}{(n_2 + n_3)(n_3 + n_4) \times \frac{\bar{x}_2}{\sigma_x} \times \frac{\bar{y}_4}{\sigma_y}} \dots\dots\dots(\text{xxv}).$$

Mr Yule has termed (xxiv) the "theoretical value of the correlation*." It will be quite obvious from the present discussion that it is only an approximate value even to the class index correlation, and cannot represent r_{xy} at all, unless the frequency distributions each consist of two isolated points. Generally

$$\begin{aligned} r_{xy} &= \frac{n_1 n_3 - n_2 n_4}{\sqrt{(n_1 + n_2)(n_1 + n_4)(n_2 + n_3)(n_2 + n_4)}} \times \sqrt{\frac{\sigma_x}{\bar{x}_1} \frac{\sigma_x}{\bar{x}_2} \frac{\sigma_y}{\bar{y}_1} \frac{\sigma_y}{\bar{y}_4}} \\ &= r_{C_x C_y} \times \sqrt{\frac{\sigma_x}{\bar{x}_1} \cdot \frac{\sigma_x}{\bar{x}_2} \cdot \frac{\sigma_y}{\bar{y}_1} \cdot \frac{\sigma_y}{\bar{y}_4}} \dots\dots\dots(\text{xxvi}). \end{aligned}$$

This corrective value on $r_{C_x C_y}$ is always true, although the above value of $r_{C_x C_y}$ is only approximate.

This is brought out at once by considering what (xxv) reduces to in the case of Gaussian frequency. Here

$$\begin{aligned} (n_2 + n_3) \bar{x}_2 &= N z_2 \sigma_x = N \sigma_x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \frac{h^2}{\sigma_x^2}}, \\ (n_3 + n_4) \bar{y}_4 &= N z_4 \sigma_y = N \sigma_y \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \frac{k^2}{\sigma_y^2}}, \end{aligned}$$

where h and k are the distances of the dividing lines from the mean. Hence if H and K have their usual meanings (xxv) gives us

$$r_{xy} = \frac{n_1 n_3 - n_2 n_4}{N^2 H K} = \epsilon \dots\dots\dots(\text{xxvii}).$$

ϵ is the expression used in my memoir on the correlation of characters not quantitatively measurable†. Thus the physical meaning of ϵ is $r_{C_x C_y}$ corrected for the use of class indices, but not for the assumption that $\bar{x}_{st} \times \bar{y}_{st}$ may be replaced by $\bar{x}_s \times \bar{y}_t$. We may therefore anticipate that ϵ or, failing Gaussian frequency, (xxv), will give a better approximation than $r_{C_x C_y}$ to the true value of the correlation of x and y .

I shall discuss elsewhere the possibility of any absolute identity in the members of a class; the only thing in which I have personally come across it is in theoretical investigations—never in practical—on Mendelian units. When we are absolutely ignorant of the nature of the frequency, then I feel sure that in dealing with either physical or mental characteristics in living forms, the Gaussian distribution will in the long run be found closer than any single type of distribution hitherto discussed, and that accordingly even ϵ —to say nothing

* *Introduction to the Theory of Statistics*, p. 212.

† *Phil. Trans.* Vol. 195, A, p. 7.

of r by fourfold—would be far safer to use than $r_{C_x C_y}$, or any arbitrary coefficient of association, which wholly neglects the question of the type of the frequency distribution. In the rare cases in which the frequency is collected into points—absolute homogeneity of category contents—then $r_{C_x C_y}$ is the right coefficient to use and not Yule's coefficient of association.

(9) I propose in this section to consider the relations between

$$S\left(\frac{n_{st}}{N} \cdot \bar{x}_{st} \cdot \bar{y}_{st}\right), \quad S\left(\frac{n_{st}}{N} \cdot \bar{x}_s \cdot \bar{y}_s\right) \quad \text{and} \quad S\left(\frac{n_{st}}{N} \cdot x_s \cdot y_s\right)$$

when the subranges for each character are small and equal. Let those for x be h , and for y be k . Then consider the surface

$$z = a + bx + cy + dx^2 + ey^2,$$

and choose a, b, c, d, e so that it gives by its volumes correctly the five class frequencies

$$n_{s-1, t}, \quad n_{s, t}, \quad n_{s+1, t}, \quad n_{s, t-1}, \quad n_{s, t+1}.$$

Then if we take the origin so that it lies at the mid-point of the $n_{s, t}$ group, we find

$$\begin{aligned} z = \frac{1}{hk} & \left\{ n_{s, t} - \frac{1}{12} (n_{s-1, t} + n_{s, t-1} - 2n_{s, t}) - \frac{1}{24} (n_{s+1, t} - n_{s-1, t} + n_{s, t+1} - n_{s, t-1}) \right\} \\ & + \frac{1}{2} \frac{x}{h} \frac{n_{s+1, t} - n_{s-1, t}}{hk} + \frac{1}{2} \frac{y}{k} \frac{n_{s, t+1} - n_{s, t-1}}{hk} \\ & + \frac{x^2}{h^2} \left(\frac{n_{s-1, t} - n_{s, t}}{hk} + \frac{n_{s+1, t} - n_{s, t}}{2hk} \right) + \frac{y^2}{k^2} \left(\frac{n_{s, t-1} - n_{s, t}}{hk} + \frac{n_{s, t+1} - n_{s, t}}{2hk} \right). \end{aligned}$$

Further

$$n_{st} (\bar{x}_{st} - x_s) = \int_{-\frac{1}{2}h}^{+\frac{1}{2}h} \int_{-\frac{1}{2}k}^{+\frac{1}{2}k} xz \, dx = \frac{1}{12} b h^3 k = \frac{h^3 k}{24} \left(\frac{n_{s+1, t} - n_{s-1, t}}{h^2 k} \right).$$

Or

$$\bar{x}_{st} = x_s + \frac{1}{24} h \frac{n_{s+1, t} - n_{s-1, t}}{n_{st}} \dots\dots\dots(\text{xxviii}).$$

Similarly

$$\bar{y}_{st} = y_t + \frac{1}{24} k \frac{n_{s, t+1} - n_{s, t-1}}{n_{st}} \dots\dots\dots(\text{xxix}).$$

Now take the product of these, multiply by n_{st} and sum, we have

$$\begin{aligned} S\left(\frac{n_{st}}{N} \bar{x}_{st} \bar{y}_{st}\right) &= S\left(\frac{n_{st}}{N} x_s y_t\right) + \frac{1}{24} h S\left\{\frac{y_t}{N} (n_{s+1, t} - n_{s-1, t})\right\} \\ &+ \frac{1}{24} k S\left\{\frac{x_s}{N} (n_{s, t+1} - n_{s, t-1})\right\} \\ &+ \frac{1}{144} h k S\left(\frac{n_{s, t+1} - n_{s, t} + n_{s, t} - n_{s, t-1}}{2N}\right) \times \left(\frac{n_{s+1, t} - n_{s, t} + n_{s, t} - n_{s-1, t}}{2n_{st}}\right). \end{aligned}$$

But if h and k be small the last summation is, precisely as that dealt with on p. 122, of the fourth order, and we have

$$S\left(\frac{n_{st}}{N} \bar{x}_{st} \bar{y}_{st}\right) = S\left(\frac{n_{st}}{N} x_s y_t\right) \dots\dots\dots(\text{xxx}),$$

for the second and third sums obviously vanish if we suppose zero frequencies at the boundaries of each row or column. Hence by (xvi) *bis* we have

$$S\left(\frac{n_{st}}{N} \bar{x}_{st} \bar{y}_{st}\right) = S(n_{st} \bar{x}_s \bar{y}_t) \times \left(1 + \frac{h^2}{12\sigma_x^2} + \frac{k^2}{12\sigma_y^2}\right) \dots\dots\dots(\text{xxxix}).$$

Thus we cannot replace $S(n_{st} \bar{x}_{st} \bar{y}_{st})$ by $S(n_{st} \bar{x}_s \bar{y}_t)$ unless h and k are very small relative to σ_x and σ_y respectively.

But now consider $r_{xy} = \frac{S(n_{xy} xy)}{N \sigma_x \sigma_y}$, the accurate value.

We may replace $S(n_{xy} xy)$ by $S(n_{st} x_s y_s)$, there being no Sheppard's correction for the product moment. Hence

$$\begin{aligned} r_{xy} &= \frac{S\left(\frac{n_{st}}{N} x_s y_s\right)}{\sigma_x \sigma_y} = \frac{S\left(\frac{n_{st}}{N} \bar{x}_{st} \bar{y}_{st}\right)}{\sigma_x \sigma_y} \text{ by (xxx)} \\ &= S\left(\frac{n_{st}}{N} \frac{\bar{x}_s}{\sigma_x} \frac{\bar{y}_t}{\sigma_y}\right) \left(1 + \frac{h^2}{12\sigma_x^2} + \frac{k^2}{12\sigma_y^2}\right) \text{ by (xxxix)} \\ &= \frac{S\left(\frac{n_{st}}{N} \frac{\bar{x}_s}{\sigma_x} \frac{\bar{y}_t}{\sigma_y}\right)}{\left(1 - \frac{h^2}{12\sigma_x^2}\right) \left(1 - \frac{k^2}{12\sigma_y^2}\right)} \text{ to the same degree of approximation} \\ &= \frac{S\left(\frac{n_{st}}{N} \frac{\bar{x}_s}{\sigma_x} \frac{\bar{y}_t}{\sigma_y}\right)}{S\left(\frac{n_s}{N} \frac{\bar{x}_s^2}{\sigma_x^2}\right) \times S\left(\frac{n_t}{N} \frac{\bar{y}_t^2}{\sigma_y^2}\right)} \text{ by (xvii) bis } \dots\dots\dots(\text{xxxii}). \end{aligned}$$

This formula is free from all Sheppard's corrections, and *in form* it does not involve the equality of the subranges. Hence it seems likely to give moderately good results even for unequal subranges, provided they do not differ very widely, and there are not too few of them. It leads as before in the case of normal frequency to

$$r_{xy} = \frac{S\left\{\frac{N n_{st}}{n_s n_t} (z_{s_1} - z_{s_2})(z_{t_1} - z_{t_2})\right\}}{S\left\{\frac{N}{n_s} (z_{s_1} - z_{s_2})^2\right\} S\left\{\frac{N}{n_t} (z_{t_1} - z_{t_2})^2\right\}},$$

but it is clear that this cannot be pressed so far as to make only a very few, and those very unequal, divisions for each variate.

(10) The present method of correction seems likely to give good results in the case of the method of mean square contingency*, when we may reasonably assume the distribution not widely divergent from the Gaussian.

* A much fuller discussion of all the corrections for contingency has been some years in progress and will shortly appear.

Illustration I. Thus I take the table for correlation of stature in father and son, the subranges being each two inches.

Stature of Father.

Stature of Son.	Stature of Father.									Totals
	58·5—60·5	60·5—62·5	62·5—64·5	64·5—66·5	66·5—68·5	68·5—70·5	70·5—72·5	72·5—74·5	74·5—76·5	
59·5—61·5	—	—	1·5	1	1	—	—	—	—	3·5
61·5—63·5	0·5	2·75	5·75	9·5	5	0·25	0·25	—	—	24
63·5—65·5	4	7·75	20	41·5	17·25	8·25	1·25	—	—	100
65·5—67·5	2	10	32	73	78·75	33·5	7·25	1	—	237·5
67·5—69·5	—	4·5	27·75	65·5	95	93·25	31·5	4·5	1	323
69·5—71·5	—	—	6·75	38·25	61	77·5	39·5	11	2	236
71·5—73·5	—	—	0·25	5·75	24·75	34·5	32·25	7	0·5	105
73·5—75·5	—	—	1	3	6·25	6·75	13	5·5	2	37·5
75·5—77·5	—	—	—	—	2·5	1·5	1·5	2·5	—	8
77·5—79·5	—	—	—	—	—	2	0·5	1	—	3·5
Totals	6·5	25	95	237·5	291·5	257·5	127	32·5	5·5	1078

Worked out by the product-moment method and using the Sheppard correction for the S.D.'s, $r_{xy} = .5259$.

I then proceed to make it into two further tables (i) a 5×5 , and (ii) a 3×3 celled table. These are as follows:

Stature of Father.

Stature of Son.	Stature of Father.					Totals
	58·5—64·5	64·5—66·5	66·5—68·5	68·5—70·5	70·5—76·5	
59·5—65·5	42·25	52	23·25	8·5	1·5	127·5
65·5—67·5	44	73	78·75	33·5	8·25	237·5
67·5—69·5	32·25	65·5	95	93·25	37	323
69·5—71·5	6·75	38·25	61	77·5	52·5	236
71·5—79·5	1·25	8·75	33·5	44·75	65·75	154
Totals	126·5	237·5	291·5	257·5	165	1078

and

Stature of Father.

Stature of Son.	Stature of Father.			Totals
	58·5—66·5	66·5—68·5	68·5—76·5	
59·5—67·5	211·25	102	51·75	365
67·5—69·5	97·75	95	130·25	323
69·5—79·5	55	94·5	240·5	390
Totals	364	291·5	422·5	1078

In obtaining these tables I was generally guided by an endeavour to roughly equalise the frequencies in the totals.

If x be the stature of father and y of son, then for the 5×5 -fold table we have, reading from left to right and from top to bottom :

$$\begin{aligned}\bar{x}_s &= -1.6780, & -.7650, & -.0694, & +.6191, & +1.5440, \\ \bar{y}_t &= -1.6741, & -.7617, & -.0298, & +.6811, & +1.5795.\end{aligned}$$

These give

$$\begin{aligned}S\left(\frac{n_s}{N} \frac{\bar{x}_s^2}{\sigma_x^2}\right) &= .917,091, & S\left(\frac{n_t}{N} \frac{\bar{y}_t^2}{\sigma_y^2}\right) &= .863,6599, \\ r_{xC_x} &= .9576, & r_{yC_y} &= .9293.\end{aligned}$$

I now calculated ϕ^2 and found for raw $N\phi^2$ the value 314.7514; deducting the correction for 5×5 cells, this equals 298.7514, whence $\phi^2 = .277,1349$,

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}} = .4659.$$

Taking this to represent the correlation of C_x and C_y we should have

$$\begin{aligned}r_{xy} &= .4659 / (.9576 \times .9293) \\ &= .5235,\end{aligned}$$

which is excellently in keeping with the value found from the 9×10 table by product moment, and not very different from the value .5140 found from the original 17×20 table in inches by the same process.

I now turn to the 3×3 -fold table and find the raw $N\phi^2 = 219.2194$ or less the correction for 3×3 cells = 215.2194, whence $\phi^2 = .199,6469$. Thus $C = .40795$.

For this table reading in same directions as before :

$$\begin{aligned}\bar{x}_s &= -1.0823, & -.0694, & +.9803, \\ \bar{y}_t &= -1.0804, & -.0298, & +1.0358,\end{aligned}$$

whence

$$\begin{aligned}S\left(\frac{n_s}{N} \frac{\bar{x}_s^2}{\sigma_x^2}\right) &= .773,470, & S\left(\frac{n_t}{N} \frac{\bar{y}_t^2}{\sigma_y^2}\right) &= .783,6384, \\ r_{xC_x} &= .8795, & r_{yC_y} &= .8852.\end{aligned}$$

Thus taking as before $r_{xy} = C / (r_{xC_x} r_{yC_y})$,

we find $r_{xy} = .5240$.

Again in excellent agreement with the 5×5 contingency result and also the product-moment result. The evaluations of $\bar{x}_s$ and $\bar{y}_t$ are of course based on the assumption of a Gaussian distribution for those variates.

Illustration II. The following table is given by Mr W. H. Gilby in a paper in *Biometrika* (Vol. VIII, p. 106) on the "Teacher's Appreciation of General Intelligence." It correlates class of intelligence with character of the clothing in boys. By treating the grades of intelligence and clothing as Gaussian variates an almost linear regression line was obtained in that paper. Hence it seems a favourable case to examine the present corrections upon. Intelligence was

measured on my scale (see *Biometrika*, Vol. VIII, p. 93) and clothing was divided into five categories, IV and V being ultimately classed together. The better intelligence has the higher letter, the worse clothing the higher number. The table runs:

		Intelligence.						
Clothing.		<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	Totals
	I	33	48	113	209	194	39	636
	II	41	100	202	255	138	15	751
	III	39	58	70	61	33	4	265
	IV and V	17	13	22	10	10	1	73
	Totals	130	219	407	535	375	59	1725

I shall use x for Intelligence and y for Clothing. Mr Gilby found raw $\phi^2 = \cdot 1014$, corrected for 4×6 cells $\phi^2 = \cdot 0927$ and $C = \cdot 291$. I find from left to right:

$$\bar{x}_s = -1\cdot8859, \quad -1\cdot1009, \quad -\cdot4758, \quad +\cdot2428, \quad +1\cdot1179, \quad +2\cdot2167,$$

$$\bar{y}_t = -1\cdot0066, \quad +\cdot2173, \quad +1\cdot2130, \quad +2\cdot1317.$$

Hence

$$S\left(\frac{n_s}{N} \frac{\bar{x}_s^2}{\sigma_x^2}\right) = \cdot 933,393, \quad S\left(\frac{n_t}{N} \frac{\bar{y}_t^2}{\sigma_y^2}\right) = \cdot 812,475,$$

$$r_{xC_x} = \cdot 9661, \quad r_{yC_y} = \cdot 9014.$$

If we now correct Mr Gilby's contingency for the class-index correlations of x and y we have

$$r_{xy} = C/(r_{xC_x} r_{yC_y}) = \cdot 3342.$$

This is very near the values, $\cdot 343$ and $\cdot 340$, found by converting the table into a three-rowed table, classes III, IV and V of clothing being grouped together and a bi-serial η method used (*Biometrika*, Vol. VIII. p. 98).

I then proceeded to work out on the full table given above the value of r as determined by the approximate formula (xvii). This can be conveniently arranged thus:

	-1·8859	-1·1009	-.4758	+.2428	+1·1179	+2·2167
-1·0066	33	48	113	209	194	39
+·2173	41	100	202	255	138	15
+1·2130	39	58	70	61	33	4
+2·1317	17	13	22	10	10	1
	+33·2178	+48·3168	+113·7458	-210·3794	-195·2804	-39·2574
	-8·9093	-21·7300	-43·8946	+55·4115	+29·9874	+3·2595
	-47·3070	-70·3540	-84·9100	+73·9930	+40·0290	+4·8520
	-36·2389	-27·7121	-46·8974	+21·3170	+21·3170	+2·1317
	-59·2374	-71·4793	-61·9562	-59·6579	-103·9470	-29·0142

Therefore

$$S\left(\frac{n_{st}\bar{x}_s\bar{y}_t}{N\sigma_x\sigma_y}\right) = -\cdot240,51548,$$

$$r_{C_x C_y} = -\frac{\cdot240,51548}{\cdot9661 \times \cdot9014} = -\cdot2762,$$

and

$$r_{C_x C_y} / (r_{x C_x} r_{y C_y}) = -\cdot3171*.$$

In the above table each row is first multiplied by the y_t on the left and the *final* sign at once given to the product; this forms the lower half of the table. The columns are then added up with the results given at the foot. The column sums are then multiplied by their respective $\bar{x}_s$'s, shown at the top of the first half of the table, regardless of the sign of $\bar{x}_s$, and the sum divided by 1725 as a continuous operation on the calculator, which shows finally $\cdot240,51548$.

The value ultimately obtained, $\cdot32$, is somewhat less than the $\cdot33$ of the contingency result but of the same order for all practical purposes.

I next reduce my table to a 4×4 table as follows:

Intelligence.

Clothing.		$B+C$	D	E	$F+G$	Totals
	I	81	113	209	233	636
	II	141	202	255	153	751
	III	97	70	61	37	265
	IV and V	30	22	10	11	73
	Totals	349	407	535	434	1725

We have

$$\bar{x}_s = -1\cdot3934, \quad -\cdot4758, \quad +\cdot2428, \quad +1\cdot2674,$$

$$\bar{y}_t = -1\cdot0066, \quad +\cdot2173, \quad +1\cdot2130, \quad +2\cdot1317,$$

$$S\left(\frac{n_s\bar{x}_s^2}{N\sigma_x^2}\right) = \cdot868,606, \quad S\left(\frac{n_t\bar{y}_t^2}{N\sigma_y^2}\right) = \cdot812,475,$$

$$r_{x C_x} = \cdot9320, \quad r_{y C_y} = \cdot9014.$$

Proceeding as before by formula (xvii) we find

$$r_{xy} = \frac{-\cdot228,76485}{\cdot868,606 \times \cdot812,475} = -\cdot3242,$$

a result which is again in excellent agreement with the previous ones.

Lastly I convert my table into a 3×3 table:

Intelligence.

Clothing.		$B+C+D$	E	$F+G$	Totals
	I	194	209	233	636
	II	343	255	153	751
	III+IV+V	219	71	48	338
	Totals	756	535	434	1725

* The negative sign shows that worse intelligence is associated with the poorer clothing.

I found $\bar{x}_s = -\cdot8995, +\cdot2428, +\cdot1\cdot2674,$
 $\bar{y}_t = -\cdot1\cdot0066, +\cdot2173, +\cdot1\cdot4110,$

and $S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) = \cdot776,941, \quad S\left(\frac{n_t \bar{y}_t^2}{N \sigma_y^2}\right) = \cdot784,241,$
 $r_{xC_x} = \cdot8814, \quad r_{yC_y} = \cdot8856.$

Proceeding exactly as before by formula (xvii) we have

$$r_{xy} = \frac{-\cdot204,01275}{\cdot784,241 \times \cdot776,941} = -\cdot3348.$$

For comparison I worked out the above 3×3 table by contingency. The contingency corrected for 3×3 cells is $\phi^2 = \cdot076,0779$, and therefore $C = \cdot2659$, hence

$$r_{xy} = \cdot2659 / (\cdot8814 \times \cdot8856) \\ = \cdot3406, \text{ regardless of sign.}$$

To sum up our results for this case we have :

	Bi-Serial η	Corrected Contingency	Formula (xvii)
6×2 Table	$\cdot340$ and $\cdot343$	—	—
6×4 Table	—	$\cdot3342$	$\cdot3171$
4×4 Table	—	—	$\cdot3242$
3×3 Table	—	$\cdot3406$	$\cdot3348$

The results are very satisfactory and show that for correlations of this magnitude our corrective factors work excellently when we use contingency to find $r_{C_x C_y}$. Formula (xvii) also gives fairly consistent results, but this is partially due to the relative smallness of r . For the smaller r the more nearly $\bar{x}_{st} \times \bar{y}_{st}$ may be replaced by $\bar{x}_s \times \bar{y}_t$ as is well shown in the case of the simple 2×2 table.

On my scale of *Intelligence* seven categories are used ; there were in Mr Gilby's schools none of the *A* or Mentally Defective class, so that he only used six, *B—G*. We see that there is a high degree of correlation between the actual intelligence of a child and its class index. We must have

For seven classes	$r > \cdot97,$
„ six „	$r = \cdot97,$
„ four „	$r = \cdot93,$
„ three „	$r = \cdot88.$

It is of interest to put beside these the frequencies of my General Health Scale and the resulting class index correlations. The scale had six classes :

Very Robust, Robust, Normally Healthy, Rather Delicate, Delicate and Very Delicate. Applied to 2037 schoolgirls (*Biometrika*, Vol. III. p. 166) it gave:

We have : $\frac{\bar{x}_s}{\sigma_x} =$

V. R.	R.	N. H.	R. D. + D.	V. D.	Total
109	578·5	803·5	478·5	67·5	2037
-2·0342	-·9033	+·0936	+1·0837	+2·2293	

leading to $S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) = \cdot 897,1616$, and $r_{xC_x} = \cdot 9472$.

If we divide into four classes, thus:

We have : $\frac{\bar{x}_s}{\sigma_x} =$

V. R.	R.	N. H.	R. D. + D. + V. D.	Total
109	578·5	803·5	546	2037
-2·0342	-·9033	+·0936	+1·2253	

leading to $S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) = \cdot 858,9714$, and $r_{xC_x} = \cdot 9268$.

And lastly if we divide into three classes, thus:

We have : $\frac{\bar{x}_s}{\sigma_x} =$

V. R. + R.	N. H.	R. D. + D. + V. D.	Total
687·5	803·5	546	2037
-1·0826	+·0936	+1·2253	

leading to $S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) = \cdot 801,3866$, and $r_{xC_x} = \cdot 8952$.

Now if we put these results with those on p. 121 together we find:

Number of Groups	Gaussian Frequency Equal Subfrequencies	Pearson's Scale General Intelligence	Pearson's Scale General Health
3	·891	·881	·895
4	·928	·932	·927
5	·947	—	·947
6	·959	·966	—

It will be noticed that the frequencies differ considerably, although not extremely, from equality. Accordingly we may find it sufficient in many cases,

although the frequencies are not symmetrical or equal, to state the value of r_{xC_x} for the same number of groups with equal frequencies from the table on p. 121, and save the trouble of calculating r_{xC_x} .

Illustration III. I took the 3×3 table for correlation of health in pairs of sisters, i.e.

First Sister.

Second Sister.		V. R. + R.	N. H.	R. D. + D. + V. D.	Totals
	V. R. + R. ...	428	172	87·5	687·5
	N. H.	172	411	220·5	803·5
	R. D. + D. + V. D.	87·5	220·5	238	546
	Totals	687·5	803·5	546	2037

and I determined the contingency. The raw value of $N\phi^2 = 424·9285$, and after correction for number of cells = $420·9285$. Hence

$$\phi^2 = .206,6414,$$

$$C = .4138.$$

Therefore by the value just found for r_{xC_x} for three health classes

$$r_{xy} = \frac{.4138}{.801,3866} = .5164.$$

I then worked out r_{xy} by aid of the approximate formula

$$r_{xy} = \frac{S\left(\frac{n_{st}}{N} \frac{\bar{x}_s}{\sigma_x} \frac{\bar{y}_t}{\sigma_y}\right)}{S\left\{\frac{n_s}{N} \left(\frac{\bar{x}_s}{\sigma_x}\right)^2\right\} S\left\{\frac{n_t}{N} \left(\frac{\bar{y}_t}{\sigma_y}\right)^2\right\}},$$

which gives in this case

$$r_{xy} = .4940.$$

The value as given in my Huxley Lecture* for the mean of two fourfold tables with divisions first between Robust and Healthy and then between Delicate and Healthy was .51. The agreement between the three methods appears reasonably good. I compared the last value with the same formula applied to the original 5×5 table (see *Biometrika*, Vol. III., p. 166). I found

$$S\left(\frac{n_{st}}{N} \frac{\bar{x}_s}{\sigma_x} \frac{\bar{y}_t}{\sigma_y}\right) = .392,7573*,$$

and

$$r_{xy} = .4880.$$

It will be clear that for the type of argument which is generally based on such numbers the accordance is very satisfactory.

* In the case of the 3×3 table this product was .317,2398, which measures the amount of correction made by denominator.

Illustration IV. I take now an extreme case, namely the Table :

A.

	a_1	a_2	a_3	Totals
b_1	4983	761	49	5793
b_2	1166	1661	572	3399
b_3	30	248	530	808
Totals	6179	2670	1151	10,000

I call this an extreme case because the table is only 3×3 , and the frequencies and the subranges are very unequal.

We have $\bar{x}_s/\sigma_x = -\cdot6172, \quad +\cdot7011, \quad +\cdot16871,$

$\bar{y}_t/\sigma_y = -\cdot6750, \quad +\cdot7100, \quad +\cdot18531.$

Whence I find $S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) = \cdot694,222, \quad r_x C_x = \cdot8332,$

$S\left(\frac{n_t \bar{y}_t^2}{N \sigma_y^2}\right) = \cdot712,753, \quad r_y C_y = \cdot8442.$

I then worked out the mean square contingency and found for crude numbers

$$N\phi^2 = 5146\cdot6.$$

Corrected for number of cells $\phi^2 = \cdot51426.$

Whence $C = \cdot5827.$

Therefore $r_{xy} = \cdot5827/(\cdot8332 \times \cdot8442) = \cdot8284.$

The actual table has been constructed *in round numbers* as the distribution of a Gaussian frequency surface for $r = \cdot80$, the divisions being at $x/\sigma_x = +\cdot3$ and $+1\cdot2$ and $y/\sigma_y = +\cdot2$ and $+1\cdot4$, and the correlation is hardly likely to differ by a unit from $\cdot80$. Considering the marked inequality of the subfrequencies the corrected contingency approaches closely to the correlation—at least the approach is sufficient for any argument likely to be drawn from the data in a 3×3 -fold table.

Illustration V. It is profitable to show the amount of error which will be introduced by taking a marked case of non-Gaussian frequency. When looking out for such cases many years ago, I found one of the most representative instances in the case of barometric heights. I take the following table from the memoir by Dr Lee and myself*, and I have arranged the 3×3 -fold table so as to give very unequal frequencies. We have thus in this case non-Gaussian frequency and very unequal frequencies. We find

$$\bar{x}_s = -1\cdot1479, \quad -\cdot1581, \quad +\cdot9161,$$

$$\bar{y}_t = -1\cdot2797, \quad -\cdot2940, \quad +\cdot8424,$$

$$S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) = \cdot768,494, \quad r_x C_x = \cdot8823,$$

$$S\left(\frac{n_t \bar{y}_t^2}{N \sigma_y^2}\right) = \cdot763,093, \quad r_y C_y = \cdot8736.$$

* *Phil. Trans.* Vol. 190, A, p. 453.

Southampton.

Laudale.		Over 30·15	30·15—29·95	Under 29·95	Totals
	Over 30·15 ...	545	148·5	26·5	720
	30·15—29·85 ...	263·25	340·75	217·5	821·5
	Under 29·85 ...	83·75	288·75	1008	1380·5
	Totals ...	892	778	1252	2922

We have raw $N\phi^2 = 1449·52$.

ϕ^2 corrected for number of cells = ·494,7023, $C = ·5753$.

Thus :
$$r_{xy} = \frac{·5753}{·8823 \times ·8736} = ·7464.$$

Found by the product-moment method*

Without Sheppard : $r_{xy} = ·7752 \pm ·0050$, with Sheppard : $r_{xy} = ·7802 \pm ·0049$.

Thus the difference is more than four times the probable error, but is not of a nature upon which any sweeping conclusions would be drawn. Indeed some might prefer the value drawn from the corrected 3×3 -fold table, owing to the distrust of isolated outlying observations which often widely modify the constants of a distribution calculated by product-moment methods.

To still further test the pliability of the method, I rearranged the data in a 3×3 -fold table with markedly unequal frequencies but nearly equal ranges, thus :

Southampton.

Laudale.		31·05—30·15	30·15—29·25	29·25—28·35	Totals
	30·85—29·85 ...	808·25	733·25	0	1541·5
	29·85—28·85 ...	83·75	1223·25	45·5	1352·5
	28·85—27·85 ...	0	14	14	28
	Totals ...	892	1970·5	59·5	2922

A more unpromising series of totals is hardly likely to be met with! We have

$$\bar{x}_s = -1·1479, \quad +·4467, \quad +2·4133.$$

$$S \left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2} \right) = ·635,4498, \quad r_{xC_x} = ·7972.$$

$$\bar{y}_t = -·7544, \quad +·8044, \quad +2·6797,$$

$$S \left(\frac{n_t \bar{y}_t^2}{N \sigma_y^2} \right) = ·568,5512, \quad r_{yC_y} = ·7540.$$

Thus we reach some of the lowest values we have come across for r_{xC_x} and r_{yC_y} in the case of threefold divisions.

* In *Phil. Trans.* Vol. 190 A, p. 453, the value ·7572 was printed in error for ·7752.

Raw $N\phi^2 = 766.887$, corrected $\phi^2 = .761,0838$ and $C = .4550$,

$$r_{xy} = .4550 / (.7972 \times .7540) = .7570,$$

a result differing only by .02 from that of the product-moment method.

It will be noticed that I have arranged the material so that the subranges are *equal*. If we work the 3×3 -fold table out, using Sheppard's correction, we find by product-moment

$$r_{xy} = .7746,$$

which is considerably nearer the mark than the corrected contingency value.

In this case a poor result is given by the formula:

$$S\left(\frac{n_{st} \bar{x}_s \bar{y}_t}{N \sigma_x \sigma_y}\right) / S\left(\frac{n_s x_s^2}{N \sigma_x^2}\right) \times S\left(\frac{n_t y_t^2}{N \sigma_y^2}\right),$$

the assumption that $S(n_{st} \bar{x}_s \bar{y}_t) = S(n_{st} \bar{x}_s \bar{y}_t)$ being by no means satisfactory for (i) very non-Gaussian distributions, or (ii) high correlations, when the subranges if equal are very large. For *equal* ranges with a moderate approach to the Gaussian distribution and a correlation not much exceeding 0.5, it gives quite good results. Thus if we take the table for stature in Father and Son and arrange it in practically equal subranges thus:

Stature of Father.					
Stature of Son.		58.5—64.5	64.5—70.5	70.5—76.5	Totals
	Under 67.5 ...	86.25	269	9.75	365
	67.5—73.5 ...	39.25	495.5	129.25	664
	73.5—79.5 ...	1	22	26	49
	Totals ...	126.5	786.5	165	1078

we find:

$$\begin{aligned} \bar{x}_s &= -1.6780, & - .0540, & + 1.5439, \\ \bar{y}_t &= -1.0804, & + .4387, & + 2.1022, \\ S\left(\frac{n_s \bar{x}_s^2}{N \sigma_x^2}\right) &= .697,738, & r_{x C_x} &= .8353, \\ S\left(\frac{n_t \bar{y}_t^2}{N \sigma_y^2}\right) &= .714,644, & r_{y C_y} &= .8454, \\ S\left(\frac{n_{st} \bar{x}_s \bar{y}_t}{N \sigma_x \sigma_y}\right) &= .260,7392, \\ \therefore r_{xy} &= .5229, \end{aligned}$$

against .5259 by the product-moment methods.

The same table yields by 3×3 -fold contingency

$$N\phi^2 = 163.909, \quad \text{corrected } \phi^2 = .148,3386, \quad C = .3594,$$

whence $r_{xy} = .5089$.

Thus the corrected contingency for roughly equal frequencies and equal sub-ranges gives respectively $\cdot 52$ and $\cdot 51$, but the formula (xvii) gives for equal ranges $\cdot 52$ and for roughly equal frequencies $\cdot 57$. It seems to me therefore that corrected contingency can be more safely applied than formula (xvii).

Conclusions.

A further discussion of the corrections needful when using the method of mean square contingency will shortly be published, but the present paper seems to indicate that it can in the bulk of cases for 3×3 -fold, 4×3 -fold or 4×4 -fold tables be used effectively. I am not aware that any other effective method has been proposed for such tables. The assumption of the Gaussian distribution to determine the correlation of the variate with its class index need not be made, if the material thrown into broad classes has had a sample quantitatively determined. But we have seen in this paper that the Gaussian assumption to fix the means of the broad classes is in many cases amply sufficient to give good results even in non-Gaussian frequencies. Of course a control series to determine r_{xC_x} , where practicable, may be advantageously sought*, but until some better method can be suggested, the present seems to me the best available for dealing with the correlation of variates classed in a few broad categories.

With contingency tables of 5×5 or 6×6 , the correcting factors will generally be small, but corrections must certainly be made for 4×4 and 3×3 tables.

* For example, I took Southampton frequency, and calculated from the original table the means of the three groups 892, 1970·5, 59·5, corresponding to the table on p. 137; they were $-1\cdot1006$, $+4123$, and $+2\cdot1746$, leading to $r_{xC_x} = \cdot 8056$, as against the Gaussian value $\cdot 7972$. This is for a sensibly non-Gaussian frequency, and a much worse agreement is permissible in actual practice, where the nature of the argument rarely turns on correlation differences of less than $\cdot 05$.