ON THE PROBABLE ERROR OF THE CORRELATION COEFFICIENT TO A SECOND APPROXIMATION*.

By H. E. SOPER, M.A.

(1) It is very important in determining whether the coefficient of correlation as found by any particular method differs significantly from the calculated value to know not only its standard deviation but also to have some idea of the nature of the frequency distribution. When the numbers dealt with are large, then, provided r be not nearly ± 1, we may quite legitimately assume a normal distribution and calculate the frequency of r on this basis. But if n be small, or if r have a value near either end of the range, then the usual values for the S.D. of r are not applicable and what is more in the latter case the frequency of r is of a markedly skew character and differs widely from a Gaussian curve. In such case the value of r found from a single sample will most probably be neither the true r of the material nor the mean value of r as deduced from a large number of samples of the same size, but the modal value of r in the given frequency distribution of r for samples of this size. In this paper the following notation will be used:

$\rho =$ correlation coefficient of the material from which the sample is drawn;

$\bar{r} =$ mean value of correlation coefficient for N samples of size n;

$\check{r} =$ modal value of the correlation in the distribution of the values of r as found from N samples of size n;

$r =$ correlation coefficient of any arbitrary sample of size n.

The first question we have to answer is what is likely to be the distribution of the r's. Clearly, when ρ differs from unity, it must be a skew distribution of limited range lying between $+1$ and -1. The general skew curves discussed in *Phil. Trans.*, Vol. 186 A, pp. 343—414, have proved themselves so capable of describing all sorts of types of frequency that one naturally turns to them in the *first* place in the present problem. There appears very little chance of successfully determining—at least for a product-moment table—the distribution of r. We must start with the assumption of a reasonable frequency distribution and justify

* The frequency-distribution of the correlation coefficient in small samples was first discussed by "Student" in his paper in *Biometrika*, Vol. VI. pp. 302–10; he invited further mathematical investigation and to a large extent supplied the impulse and direction to the present paper.

it *a posteriori* by means of experimental samples for given ρ and given n. Now the only type among the skew curves mentioned applicable in the present case is of the form:

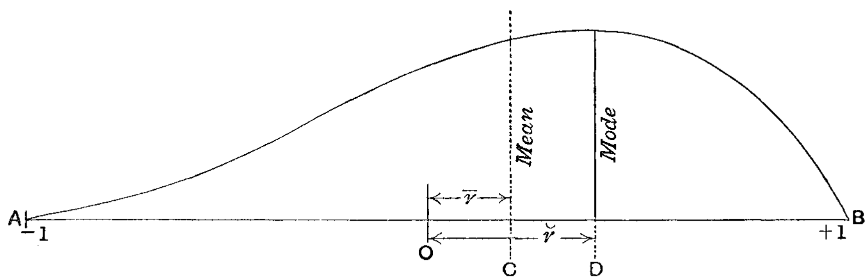
$$y = y_0 \left(1 - \frac{x}{a_1}\right)^{m_1} \left(1 + \frac{x}{a_2}\right)^{m_2} \quad \text{..........................(i)},$$

where, if the origin be at the mode, we must have

$$m_1/a_1 = m_2/a_2 \quad \text{..................................(ii)}$$

Now if we suppose ρ to be positive, we clearly have

$$a_1 = 1 - \check{r}, \qquad a_2 = 1 + \check{r}.$$



Hence from (i) and (ii)

$$y = y_0 \left(1 - \frac{x}{1 - \check{r}}\right)^{m_1} \left(1 + \frac{x}{1 + \check{r}}\right)^{m_2} \quad \text{.....................(iii)},$$

$$\check{r} = (m_2 - m_1)/(m_2 + m_1) \quad \text{..........................(iv)}$$

Now let σ_r denote the standard-deviation of the distribution. Then we easily find (*Phil. Trans., loc. cit.* p. 368)

$$AC = \frac{2(m_2 + 1)}{m_1 + m_2 + 2} = 1 + \bar{r} \quad \text{..........................(v)},$$

$$\sigma_r^2 = \frac{4(m_1 + 1)(m_2 + 1)}{(m_1 + m_2 + 2)^2 (m_1 + m_2 + 3)} \quad \text{.....................(vi)}.$$

Thus
$$\bar{r} = (m_2 - m_1)/(m_1 + m_2 + 2) \quad \text{..........................(vii)},$$

$$1 - \bar{r}^2 = 4(m_1 + 1)(m_2 + 1)/(m_1 + m_2 + 2)^2,$$

and
$$\sigma_r^2 = (1 - \bar{r}^2)/(m_1 + m_2 + 3) \quad \text{..............................(viii)}.$$

It follows that

$$m_1 + m_2 + 3 = \frac{1 - \bar{r}^2}{\sigma_r^2} = \lambda, \text{ say},$$

$$m_2 - m_1 = \bar{r} \left\{\frac{1 - \bar{r}^2}{\sigma_r^2} - 1\right\} = \bar{r}(\lambda - 1).$$

Accordingly
$$\left.\begin{aligned} m_1 &= \tfrac{1}{2}(\lambda - 1)(1 - \bar{r}) - 1 \\ m_2 &= \tfrac{1}{2}(\lambda - 1)(1 + \bar{r}) - 1 \end{aligned}\right\} \quad \text{........................(ix)}.$$

Substituting in (iv) we have

$$\left.\begin{aligned} \check{r} &= \bar{r}(\lambda - 1)/(\lambda - 3) \\ d &= \check{r} - \bar{r} = 2\bar{r}/(\lambda - 3) \end{aligned}\right\} \quad \text{..........................(x)}.$$

and

Since $\sigma_r^2 = (1-\bar{r}^2)/\lambda$, and must grow very small as the number in the sample grows large, i.e. λ grows large, we see that $\breve{r}$ and $\bar{r}$ rapidly become equal as the sample increases or the distribution becomes symmetrical.

The value of y_0 can be found from (*Phil. Trans., loc. cit.* p. 369)

$$y_0 = \tfrac{1}{2} N \frac{m_1^{m_1} m_2^{m_2}}{(m_1+m_2)^{m_1+m_2}} \frac{(m_1+m_2+1)\,\Gamma(m_1+m_2+1)}{\Gamma(m_1+1)\,\Gamma(m_2+1)} \text{...........(xi)}.$$

The problem of the distribution of r would thus be completely solved, if we knew:

(*a*) $\breve{r}$ in terms of ρ,

(*b*) σ_r^2 in terms of ρ.

Using Stirling's Theorem we can reduce the expression of y_0 to

$$\begin{aligned} y_0 &= \frac{N}{2\sqrt{2\pi}} \frac{(m_1+m_2+1)\sqrt{(m_1+m_2)}}{\sqrt{m_1 m_2}} e^{\frac{1}{12}\left(\frac{1}{m_1+m_2}-\frac{1}{m_1}-\frac{1}{m_2}\right)} \\ &= \frac{N}{\sqrt{2\pi}\,\sigma_r} \sqrt{\left(1+\frac{1}{m_1}\right)\left(1+\frac{1}{m_2}\right)} \Big/ \sqrt{\left(1+\frac{3}{m_1+m_2}\right)} \left(1+\frac{1}{m_1+m_2+1}\right) \\ &\qquad \times e^{\frac{1}{12}\left(\frac{1}{m_1+m_2}-\frac{1}{m_1}-\frac{1}{m_2}\right)} \\ &= \frac{N}{\sqrt{2\pi}\,\sigma_r}\left\{1+\frac{1}{12}\left(\frac{5}{m_1}+\frac{5}{m_2}-\frac{29}{m_1+m_2}\right)\right\} \text{.....................................(xii)}. \end{aligned}$$

This approaches rapidly to the Gaussian value $N/(\sqrt{2\pi}\,\sigma_r)$, if σ_r be at all small and therefore λ and m_1 and m_2 large. For most purposes it is sufficient to take

$$y = \frac{N}{\sqrt{2\pi}\,\sigma_r}\left(1-\frac{x}{1-\breve{r}}\right)^{\frac{1}{2}(\lambda-1)(1-\bar{r})-1}\left(1+\frac{x}{1+\breve{r}}\right)^{\frac{1}{2}(\lambda-1)(1+\bar{r})-1} \text{......(xiii)},$$

where the relation between $\breve{r}$ and $\bar{r}$ is given by (x) and it only remains to determine $\bar{r}$ and σ_r in terms of ρ.

(2) Now the product moment value of the coefficient of correlation, ρ, between two measured characters in any population is defined by

$$\rho = \frac{p_{11} - p_{10}p_{01}}{\sqrt{(p_{20}-p_{10}^2)} \times \sqrt{(p_{02}-p_{01}^2)}} \text{ ....(xiv)},$$

p_{10}, p_{20} being the first and second moments in respect of the first character, p_{01}, p_{02} those in respect of the second character, and p_{11} being the first product moment, all derived from measures of individuals taken from some arbitrary origins of measurement in the two characters.

If samples of number n are selected at random the moments will have different values p_{10}', p_{20}' etc., and in consequence the coefficient of correlation a different value, r, in any sample, and

$$r = \frac{p_{11}' - p_{10}'p_{01}'}{\sqrt{(p_{20}'-p_{10}'^2)} \times \sqrt{(p_{02}'-p_{01}'^2)}} \text{ ....(xv)}.$$

The mean values of p_{10}', p_{20}' etc. in all samples are p_{10}, p_{20} etc., since the moments are crude and simple averages of individual values. Let dp_{10}, dp_{20}, dp_{01}, dp_{02}, dp_{11} be the deviations of p_{10}', p_{20}', p_{01}', p_{02}', p_{11}' from their means p_{10}, p_{20}, p_{01}, p_{02}, p_{11}. The mean value of r we have called $\bar{r}$. Let dr be the deviation of r from its mean $\bar{r}$, then (xv) becomes

$$\bar{r} + dr = \frac{p_{11} + dp_{11} - (p_{10} + dp_{10})(p_{01} + dp_{01})}{\sqrt{\{p_{20} + dp_{20} - (p_{10} + dp_{10})^2\}} \times \sqrt{\{p_{02} + dp_{02} - (p_{01} + dp_{01})^2\}}} \quad \ldots\text{(xvi)}.$$

Choose the fixed origin of measurement of each character to be the mean of that character in the whole population, then $p_{10} = p_{01} = 0$, and (xvi) becomes

$$\bar{r} + dr = \frac{p_{11} + dp_{11} - dp_{10}dp_{01}}{\sqrt{\{p_{20} + dp_{20} - (dp_{10})^2\}} \times \sqrt{\{p_{02} + dp_{02} - (dp_{01})^2\}}} \quad \ldots\ldots\text{(xvii)}.$$

If the distributions and correlations of the deviations of the moments in samples of n are known this is the equation for determining the distribution of the values of the correlation coefficient. The average value of the right-hand side of (xvii) will be $\bar{r}$. The average values of the square, cube etc. of the right-hand expression will give the crude second, third etc. moments of r from which the moments of deviations from mean value of the correlation coefficient can be derived.

Now if (xvii) be expanded in powers and products of the deviations it may be anticipated that the average values of terms of higher order in the deviations are of higher order in $1/n$, and that there is a limit to the number of terms needed to give a required approximation. The approximation sought in the crude moments of r is to terms in $1/n^2$ only, in order that the moments from the mean may be to terms in $1/n^2$, and so that σ_r^2 for instance, which is known* to have the value $(1-\rho^2)^2/n$ to the first approximation for normal frequency, may be further carried to a term in $1/n^2$.

Thus the process for determining $\bar{r}$ is to expand (xvii) and to find and insert in it the average values in samples of n of the various powers and products of the deviations of the moments involved, carrying the process on as far as is necessary to gather in all significant terms as defined above: and a similar process applied to the squares and cubes etc. of (xvii) determines the higher moments of r.

Were the samples sufficiently large these deviations would approximate, as has been shown, to normal distributions, and the known properties of such distributions could be utilised in evaluating the complicated mean, but we are dealing with small samples where the deviations are not so distributed, and it is necessary, in the first place, to evaluate these moments of deviations in terms of the higher moments of the whole distribution without making any assumptions or any approximations within the limits assigned. After this is done the distribution of the two characters will be assumed normal and the results expressed in terms of ρ, the coefficient of correlation of the material examined, and n, the number in the sample, only.

* *Biometrika*, Vol. IX. p. 5 (if $\beta_2 = \beta_2' = 3$).

The method adopted in this paper is that of grade groups. It is well known that if in an indefinitely large population the fraction f fall into a certain grade of a character or combination of grades of two characters, then in taking random samples of n the numbers of this group to be found in such samples follow the binomial distribution of frequency

$$f^n + nf^{n-1}(1-f) + \frac{n(n-1)}{1.2} f^{n-2}(1-f)^2 + \ldots + (1-f)^n,$$

and that the mean number is nf and that the deviations from this mean number have moments

$$\text{mean } (df)^2 = \frac{1}{n} f(1-f),$$

$$\text{,, } (df)^3 = \frac{1}{n^2} f(1-f)(1-2f),$$

$$\text{,, } (df)^4 = \frac{3}{n^2} f^2(1-f)^2 + \frac{1}{n^3} f(1-f)(1-6f+6f^2),$$

$$\text{,, } (df)^5 = \frac{10}{n^3} f^2(1-f)^2(1-2f) + \frac{1}{n^4} f(1-f)(1-2f)(1-12f+12f^2),$$

etc., the fourth moment being the last which gives terms in $1/n^2$ [see Pearson, *Phil. Trans.* Vol. 186, p. 347 and *Phil. Mag.* 1899, pp. 240, 241]. Here df is the deviation from mean value f of the frequency of the group in a sample of n.

Moreover if $f_1, f_2, f_3 \ldots$ are the totality of frequencies of the various detached groups into which the population is divided by the graduation (which in our case is a double one) of character the various product moments of the deviations in samples of n may be deduced. These and the above, as far as our approximation needs, are put in one table as follows:

$$\text{mean } (df_1)^2 = \frac{1}{n} f_1(1-f_1) \qquad \text{etc.} \quad \ldots\text{(xviii)},$$

$$\text{,, } df_1 . df_2 = -\frac{1}{n} f_1 f_2 \qquad \text{,,} \quad \ldots\ldots\text{(xix)},$$

$$\text{,, } (df_1)^3 = \frac{1}{n^2} f_1(1-f_1)(1-2f_1) \qquad \text{,,} \quad \ldots\ldots\ldots\text{(xx)},$$

$$\text{,, } (df_1)^2 . df_2 = -\frac{1}{n^2} f_1 f_2 (1-2f_1) \qquad \text{,,} \quad \ldots\ldots\text{(xxi)},$$

$$\text{,, } df_1 . df_2 . df_3 = \frac{2}{n^2} f_1 f_2 f_3 \qquad \text{,,} \quad \ldots\ldots\text{(xxii)},$$

$$\text{,, } (df_1)^4 = \frac{3}{n^2} f_1^2 (1-f_1)^2 \qquad \text{,,} \quad \ldots\ldots\text{(xxiii)},$$

$$\text{,, } (df_1)^3 . df_2 = -\frac{3}{n^2} f_1^2 f_2 (1-f_1) \qquad \text{,,} \quad \ldots\ldots\text{(xxiv)},$$

$$\text{,, } (df_1)^2 . (df_2)^2 = \frac{1}{n^2} f_1 f_2 (1-f_1-f_2+3f_1 f_2) \qquad \text{,,} \quad \ldots\ldots\text{(xxv)},$$

$$\text{,, } (df_1)^2 . df_2 . df_3 = -\frac{1}{n^2} f_1 f_2 f_3 (1-3f_1) \qquad \text{,,} \quad \ldots\ldots\text{(xxvi)},$$

$$\text{,, } df_1 . df_2 . df_3 . df_4 = \frac{3}{n^2} f_1 f_2 f_3 f_4 \qquad \text{,,} \quad \ldots\text{(xxvii)},$$

the last five values being approximate and wanting terms in $1/n^3$ to render them exact.

The method of derivation of the product from the power moments is illustrated in the following example.

$$\text{Mean } (df_1)^2 . df_2 . df_3 = \text{mean } \{(df_1)^2 \times \text{mean } df_2 df_3 \text{ for constant } df_1\}.$$

Now in samples of constant df_1 the number of 1's is $nf_1 + ndf_1$ and of not 1's $n - nf_1 - ndf_1$, amongst which latter restricted population in the whole community the frequency of 2's will be $\frac{f_2}{1-f_1}$ and of 3's $\frac{f_3}{1-f_1}$. Hence the mean number of 2's in such samples will be $(n - nf_1 - ndf_1)\frac{f_2}{1-f_1}$, and of 3's

$$(n - nf_1 - ndf_1)\frac{f_3}{1-f_1},$$

differing from the mean numbers in all samples, nf_2, nf_3, by $-ndf_1\frac{f_2}{1-f_1}$ and $-ndf_1\frac{f_3}{1-f_1}$, and the mean product of the deviations from such means in the restricted samples will be

$$-(n - nf_1 - ndf_1) . \frac{f_2}{1-f_1} . \frac{f_3}{1-f_1}$$

by (xix). It follows that the mean product of the deviations ndf_2, ndf_3, which are measured from the means of all samples, will be

$$-(n - nf_1 - ndf_1) . \frac{f_2}{1-f_1} . \frac{f_3}{1-f_1} + \left(-ndf_1\frac{f_2}{1-f_1}\right)\left(-ndf_1\frac{f_3}{1-f_1}\right),$$

i.e.
$$-n\frac{f_2 f_3}{1-f_1} + ndf_1\frac{f_2 f_3}{(1-f_1)^2} + n^2 (df_1)^2 \frac{f_2 f_3}{(1-f_1)^2},$$

in samples of constant df_1. Dividing by n^2 we get the value of mean $df_2 . df_3$ for constant df_1 and so obtain finally

$$\text{mean } (df_1)^2 df_2 df_3 = \text{mean } -\frac{(df_1)^2}{n} . \frac{f_2 f_3}{1-f_1} + \frac{(df_1)^3}{n} . \frac{f_2 f_3}{(1-f_1)^2} + (df_1)^4 . \frac{f_2 f_3}{(1-f_1)^2},$$

which by (xviii), (xx) and (xxiii)

$$= -\frac{1}{n^2} f_1 f_2 f_3 + \frac{1}{n^3}\frac{f_1 f_2 f_3 (1 - 2f_1)}{1-f_1} + \frac{3}{n^2} f_1^2 f_2 f_3$$

$$= -\frac{1}{n^2} f_1 f_2 f_3 (1 - 3f_1),$$

to our approximation.

The other formulae were arrived at in like manner but the process is lengthy and these formulae and the general formulae that follow have been verified by a shorter process, which however being less direct in method is not introduced in this paper.

There is no necessity to take the products of the deviations more than four together, for these do not give terms in $1/n^2$. Did any products, five together for instance, give terms in $1/n^2$, then mean $(df_1 + df_2 + df_3 + df_4 + df_5)^5$ would have such terms, which is contrary to the formula arrived at above for mean $(df)^5$.

Having obtained the various mean products of deviations of group frequencies shown in equations (xviii)—(xxvii) the mean products of deviations of moments, formed by associating with such group frequencies their grade values, follow.

Let a_1, a_2 ... be the values to be assigned to the grades 1, 2 ... in the formation of the moment p (these values will in the present case be the product of one power of one grade of one character with another or the same power of one grade of the second character). Then

$$p = a_1 f_1 + a_2 f_2 + a_3 f_3 + \ldots$$

and if a_1', a_2' ... are the values proper to a second moment, p', in like manner

$$p' = a_1' f_1 + a_2' f_2 + a_3' f_3 + \ldots,$$

and if in random samples of n deviations df_1, df_2 ... in the frequencies lead to deviations dp, dp', in the moments, all deviations being taken from the above universal values which are also the mean values in samples, then

$$dp = a_1 df_1 + a_2 df_2 + a_3 df_3 + \ldots$$
$$dp' = a_1' df_1 + a_2' df_2 + a_3' df_3 + \ldots$$

and so

$$\begin{aligned}\text{mean } dp \,.\, dp' &= \text{mean}\,[a_1 a_1' (df_1)^2 + a_2 a_2' (df_2)^2 + \ldots + a_1 a_2' df_1 df_2 + a_1' a_2 df_1 df_2 + \ldots] \\ &= \frac{1}{n}[a_1 a_1' f_1 (1 - f_1) + a_2 a_2' f_2 (1 - f_2) + \ldots - a_1 a_2' f_1 f_2 - a_1' a_2 f_1 f_2 - \ldots] \\ &= \frac{1}{n}[a_1 a_1' f_1 + a_2 a_2' f_2 + \ldots - (a_1 f_1 + a_2 f_2 + \ldots)(a_1' f_1 + a_2' f_2 + \ldots)].\end{aligned}$$

If then p is the u, v moment defined by

$$p_{uv} = a_1^u b_1^v f_{11} + a_1^u b_2^v f_{12} + a_2^u b_2^v f_{22} + \ldots$$

obtained by summing the products of the group frequencies f by the uth power of the grade value a of the first character and the vth power of the grade value b of the second character in that group; and p' is the u', v' moment defined by

$$p_{u'v'} = a_1^{u'} b_1^{v'} f_{11} + a_1^{u'} b_2^{v'} f_{12} + a_2^{u'} b_2^{v'} f_{22} + \ldots,$$

it follows that the first term in the above square brackets is

$$a_1^{u+u'} b_1^{v+v'} f_{11} + a_1^{u+u'} b_2^{v+v'} f_{12} + \ldots$$

or $p_{u+u'\,v+v'}$, and the general formula for the mean products two together of deviations of grade moments is*

$$\text{mean } dp_{uv} \,.\, dp_{u'v'} = \frac{1}{n}[p_{u+u'\,v+v'} - p_{uv}\, p_{u'v'}] \quad \ldots\ldots\ldots\ldots(\text{xxviii}).$$

* See W. F. Sheppard, "On the application of the theory of error to cases of normal distribution and correlation," *Phil. Trans.* 1899 (192 A), in which paper (p. 127) are given formulae for the mean products, two together, of errors of moments calculated from the means of samples. In the present paper, it should be noted, the moments of the samples are crude, being calculated, not from the means of the samples, but from the mean values of the measured characters in the whole population; and dp is the deviation in the value of the crude moment in any particular sample from its mean value in all samples, which is mean $a_1(f_1 + df_1) + a_2(f_2 + df_2) + \ldots = a_1 f_1 + a_2 f_2 + \ldots = p$ or the moment in the whole population. This latter is a true moment, the general means having been taken as the origin of measurement.

It will be observed that nothing in the proof prevents u', v' from having the same values as u, v and the formula is true for any second order moment whether power or product.

In like manner if p, p', p'' are any three moments of the material sampled we have the equations of deviations

$$dp = a_1 df_1 + a_2 df_2 + a_3 df_3 + \dots$$
$$dp' = a_1' df_1 + a_2' df_2 + a_3' df_3 + \dots$$
$$dp'' = a_1'' df_1 + a_2'' df_2 + a_3'' df_3 + \dots$$

giving

$$\begin{aligned}\text{mean } dp \,.\, dp' \,.\, dp'' = \text{mean } [&a_1 a_1' a_2'' (df_1)^3 + \dots \text{ all grades} \\ &+ (a_1 a_1' a_2'' + a_1' a_1'' a_2 + a_1 a_1'' a_2')(df_1)^2 df_2 \\ &\quad + (a_1 a_2' a_2'' + a_1' a_2 a_2'' + a_1'' a_2 a_2') df_1 \,.\, (df_2)^2 + \dots \text{ all pairs} \\ &+ (a_1 a_2' a_3'' + a_1 a_2'' a_3' + a_1' a_2 a_3'' \\ &\quad + a_1' a_2'' a_3 + a_1'' a_2 a_3' + a_1'' a_2' a_3) df_1 \,.\, df_2 \,.\, df_3 + \dots \text{all triads}],\end{aligned}$$

and inserting the values from equations (xx), (xxi), (xxii),

$$\begin{aligned}\text{mean } dp \,.\, dp' \,.\, dp'' = \frac{1}{n^2} [&a_1 a_1' a_1'' f_1 (1 - f_1)(1 - 2f_1) + \dots \text{ all grades} \\ &- (a_1 a_1' a_2'' + a_1' a_1'' a_2 + a_1 a_1'' a_2') f_1 f_2 (1 - 2f_1) \\ &- (a_1 a_2' a_2'' + a_1' a_2 a_2'' + a_1'' a_2 a_2') f_1 f_2 (1 - 2f_2) - \dots \text{ all pairs} \\ &+ (a_1 a_2' a_3'' + a_1 a_2'' a_3' + a_1' a_2 a_3'' \\ &\quad + a_1' a_2'' a_3 + a_1'' a_2 a_3' + a_1'' a_2' a_3) \, 2f_1 f_2 f_3 + \dots \text{ all triads}],\end{aligned}$$

and collecting terms of first, second and third degree in f's and suitably commuting the f's and a's this is seen to be

$$\begin{aligned}&= \frac{1}{n^2} [(a_1 a_1' a_1'' f_1 + \dots) \\ &- (a_1 a_1' f_1 + \dots)(a_1'' f_1 + \dots) - (a_1 a_1'' f_1 + \dots)(a_1' f_1 + \dots) - (a_1' a_1'' f_1 + \dots)(a_1 f_1 + \dots) \\ &\quad + 2 (a_1 f_1 + \dots)(a_1' f_1 + \dots)(a_1'' f_1 + \dots)],\end{aligned}$$

the sums being for all grades.

If then p, p' have the double grade values p_{uv}, $p_{u'v'}$ previously assigned and p'' stands in the same way for $p_{u''v''}$ where

$$p_{u''v''} = a_1^{u''} b_1^{v''} f_{11} + a_1^{u''} b_2^{v''} f_{12} + a_2^{u''} b_2^{v''} f_{22} + \dots$$

there results the general formula for the mean products three together of deviations in moments as follows

$$\begin{aligned}\text{mean } dp_{uv} \,.\, dp_{u'v'} \,.\, dp_{u''v''} = \frac{1}{n^2} [&p_{u+u'+u''\, v+v'+v''} - p_{u+u'\, v+v'} \, p_{u''v''} \\ &- p_{u+u''\, v+v''} \, p_{u'v'} - p_{u'+u''\, v'+v''} \, p_{uv} + 2 p_{uv} p_{u'v'} p_{u''v''}] \quad \dots\dots\text{(xxix)},\end{aligned}$$

where, as before, the values of the suffixes may be any the same and the formula gives power moments equally well with the product moments of the deviations.

Precisely the same process evaluates the mean products four together. We shall have, putting in representative terms only of each series, and using equations (xxiii), (xxiv), (xxv), (xxvi), (xxvii),

$$\begin{aligned}\text{mean } dp\,.\,dp'\,.\,dp''\,.\,dp''' = \frac{1}{n^2}\,[&a_1a_1'a_1''a_1'''3f_1^2(1-f_1)^2+\dots\\ &-a_1a_1'a_1''a_2'''3f_1^2(1-f_1)f_2+\dots\\ &+a_1a_1'a_2''a_2'''f_1f_2(1-f_1-f_2+3f_1f_2)+\dots\\ &-a_1a_1'a_2''a_3'''f_1f_2f_3(1-3f_1)-\dots\\ &+a_1a_2'a_3''a_4'''f_1f_2f_3f_4+\dots]\end{aligned}$$

$$\begin{aligned}=\frac{1}{n^2}[&(a_1a_1'f_1+\dots)(a_1''a_1'''f_1+\dots)\\ &+(a_1a_1''f_1+\dots)(a_1'a_1'''f_1+\dots)\\ &+(a_1a_1'''f_1+\dots)(a_1'a_1''f_1+\dots)\\ &-(a_1a_1'f_1+\dots)(a_1''f_1+\dots)(a_1'''f_1+\dots)\\ &-(a_1a_1''f_1+\dots)(a_1'f_1+\dots)(a_1'''f_1+\dots)\\ &-(a_1a_1'''f_1+\dots)(a_1'f_1+\dots)(a_1''f_1+\dots)\\ &-(a_1'a_1''f_1+\dots)(a_1f_1+\dots)(a_1'''f_1+\dots)\\ &-(a_1'a_1'''f_1+\dots)(a_1f_1+\dots)(a_1''f_1+\dots)\\ &-(a_1''a_1'''f_1+\dots)(a_1f_1+\dots)(a_1'f_1+\dots)\\ &+3(a_1f_1+\dots)(a_1'f_1+\dots)(a_1''f_1+\dots)(a_1'''f_1+\dots)]\end{aligned}$$

on collecting terms and rearranging the associations of f's and a's as before. And putting into factors this

$$\begin{aligned}=\frac{1}{n^2}[&\{(a_1a_1'f_1+\dots)-(a_1f_1+\dots)(a_1'f_1+\dots)\}\{(a_1''a_1'''f_1+\dots)-(a_1''f_1+\dots)(a_1'''f_1+\dots)\}\\ &+\{(a_1a_1''f_1+\dots)-(a_1f_1+\dots)(a_1''f_1+\dots)\}\{(a_1'a_1'''f_1+\dots)-(a_1'f_1+\dots)(a_1'''f_1+\dots)\}\\ &+\{(a_1a_1'''f_1+\dots)-(a_1f_1+\dots)(a_1'''f_1+\dots)\}\{(a_1'a_1''f_1+\dots)-(a_1'f_1+\dots)(a_1''f_1+\dots)\}].\end{aligned}$$

And so again if the material is double graded and p_{uv}, $p_{u'v'}$, $p_{u''v''}$, $p_{u'''v'''}$ are any four moments involving products of powers of both grades, the general formula for the mean products four together of the deviations of such moments in samples of n becomes

$$\begin{aligned}&\text{mean } dp_{uv}\,.\,dp_{u'v'}\,.\,dp_{u''v''}\,.\,dp_{u'''v'''}\\ &\quad=\frac{1}{n^2}[(p_{u+u'\,v+v'}-p_{uv}\,p_{u'v'})(p_{u''+u'''\,v''+v'''}-p_{u''v''}\,p_{u'''v'''})\\ &\qquad+(p_{u+u''\,v+v''}-p_{uv}p_{u''v''})(p_{u'+u'''\,v'+v'''}-p_{u'v'}p_{u'''v'''})\\ &\qquad+(p_{u+u'''\,v+v'''}-p_{uv}p_{u'''v'''})(p_{u'+u''\,v'+v''}-p_{u'v'}\,p_{u''v''})]\ \dots\dots\dots\text{(xxx)},\end{aligned}$$

and it is to be recalled that this formula omits terms in $\frac{1}{n^3}$, not wanted to the degree of approximation laid down.

Comparing (xxx) and (xxviii) it appears that the mean values four together are equal, within our degree of approximation, to the sum of the products of the mean values two together of the complementary pairs making the four, the division being possible in three ways.

$$\begin{aligned}\text{mean } dp_{uv}dp_{u'v'}dp_{u''v''}dp_{u'''v'''} &= \text{mean } dp_{uv}dp_{u'v'} \times \text{mean } dp_{u''v''}dp_{u'''v'''}\\ &+ \text{mean } dp_{uv}dp_{u''v''} \times \text{mean } dp_{u'v'}dp_{u'''v'''}\\ &+ \text{mean } dp_{uv}dp_{u'''v'''} \times \text{mean } dp_{u'v'}dp_{u''v''} \quad \ldots\text{(xxxi)}.\end{aligned}$$

It is unnecessary to find the general formula for the mean products of deviations five together, which by p. 96 will contribute nothing within our approximation, and formulae (xxviii), (xxix) and (xxx) applied to the expansions of (xvii) and its powers are sufficient to evaluate the general formulae for the mean and moments of deviations of r as far as terms in $1/n^2$.

It is not proposed to exhibit these general formulae for moments of deviations of r in terms of the higher moments of the given distribution at length, but to proceed at once to the simpler case of a normal distribution in the two correlated characters and reduce the higher moments to second moments and the coefficient of correlation, ρ, as such distributions, it is well known, admit. In order to reduce in this way the values of the various mean products at the same time that they are evaluated by the formulae (xxviii), (xxix), (xxx), the necessary formulae of reduction are next obtained.

The expression for dr involves dp_{10}, dp_{20}, dp_{01}, dp_{02} and dp_{11}, and (xxix) shows that we shall require to reduce $p_{60}, p_{51} \ldots p_{50} \ldots p_{40} \ldots p_{30} \ldots$ in the above way.

Now it is well known that in normal distributions, following the Gaussian law of frequency, the odd moments from the mean in either character vanish and the even moments are derived from the second moment by a simple formula of reduction, from which there results that

$$p_{50} = p_{30} = p_{10} = 0, \qquad p_{05} = p_{03} = p_{01} = 0,$$
$$p_{40} = 3p_{20}^2, \quad p_{60} = 5 \,.\, 3p_{20}^3, \quad p_{04} = 3p_{02}^2, \quad p_{06} = 5 \,.\, 3p_{02}^3.$$

And, utilising these results, the higher product moments of normal distributions in two characters may be derived from the first product moment and the second moments by two well-known properties of the Gaussian surface.

If x, y are deviations from their mean value of two normally correlated characters the mean value of y for a given x is $\frac{p_{11}}{p_{20}}x$, and if y' is the deviation of y from its mean value in the array the distribution of y' is normal; its second moment is $p_{02}-\frac{p_{11}^2}{p_{20}}$; and its higher moments follow the same laws of reduction as above. Hence

$$\begin{aligned}
p_{uv} &= \text{mean } x^u y^v \\
&= \text{mean } \{x^u \times \text{mean } y^v \text{ for given } x\} \\
&= \text{mean } \left\{x^u \times \text{mean } \left(\frac{p_{11}}{p_{20}}x + y'\right)^v \text{ for given } x\right\} \\
&= \text{mean } \left\{x^u \times \text{mean } \left(\frac{p_{11}^v}{p_{20}^v}x^v + v\,.\,\frac{p_{11}^{v-1}}{p_{20}^{v-1}}x^{v-1}y' \right.\right. \\
&\qquad \left.\left. + \frac{v\,.\,(v-1)}{1\,.\,2}\,.\,\frac{p_{11}^{v-2}}{p_{20}^{v-2}}x^{v-2}y'^2 + \dots\right) \text{ for given } x\right\},
\end{aligned}$$

and so remembering that mean y', mean y'^3, etc. vanish, and mean y'^4, mean y'^6, etc. reduce by the above formulae we have

$$\begin{aligned}
p_{uv} &= \left(\frac{p_{11}}{p_{20}}\right)^v p_{u+v\,0} + \frac{v\,(v-1)}{1\,.\,2}\left(\frac{p_{11}}{p_{20}}\right)^{v-2} p_{u+v-2\,0} \\
&\quad + \frac{v\,(v-1)\,(v-2)\,(v-3)}{1\,.\,2\,.\,3\,.\,4}\left(\frac{p_{11}}{p_{20}}\right)^{v-4} p_{u+v-4\,0}\left(p_{02}-\frac{p_{11}^2}{p_{20}}\right) \\
&\quad + \frac{v!}{6!\,(v-6)!}\left(\frac{p_{11}}{p_{20}}\right)^{v-6} p_{u+v-6\,0}\,.\,3\left(p_{02}-\frac{p_{11}^2}{p_{20}}\right)^2 + \dots \qquad \text{.........(xxxii).}
\end{aligned}$$

It follows that if $u+v$ is odd p_{uv} is zero, that is

$$p_{30} = p_{21} = p_{12} = p_{03} = 0,$$
$$p_{50} = p_{41} = p_{32} = p_{23} = p_{14} = p_{05} = 0.$$

If $u+v$ is even it is convenient to divide by suitable powers of p_{20} and p_{02}, and putting ρ for $\frac{p_{11}}{\sqrt{p_{20}}\sqrt{p_{02}}}$ exhibit all the reduction formulae together as follows:

$$\begin{aligned}
&p_{40}/p_{20}^2 = p_{04}/p_{02}^2 = 3, \quad p_{31}/p_{20}^{\frac{3}{2}}p_{02}^{\frac{1}{2}} = p_{13}/p_{20}^{\frac{1}{2}}p_{02}^{\frac{3}{2}} = 3\rho, \quad p_{22}/p_{20}p_{02} = 1+2\rho^2, \\
&p_{60}/p_{20}^3 = p_{06}/p_{02}^3 = 15, \quad p_{51}/p_{20}^{\frac{5}{2}}p_{02}^{\frac{1}{2}} = p_{15}/p_{20}^{\frac{1}{2}}p_{02}^{\frac{5}{2}} = 15\rho, \\
&p_{42}/p_{20}^2p_{02} = p_{24}/p_{20}p_{02}^2 = 3+12\rho^2, \quad p_{33}/p_{20}^{\frac{3}{2}}p_{02}^{\frac{3}{2}} = 9\rho+6\rho^3 \quad \text{...............(xxxiii).}
\end{aligned}$$

If in like manner the numerator and denominator of the expression (xvii) for the deviation in r in terms of the deviations in the moments be divided by $\sqrt{p_{20}}\,.\sqrt{p_{02}}$ and we write

$$\alpha_1 = \frac{dp_{10}}{\sqrt{p_{20}}}, \quad \alpha_2 = \frac{dp_{20}}{p_{20}}, \quad \beta_1 = \frac{dp_{01}}{\sqrt{p_{02}}}, \quad \beta_2 = \frac{dp_{02}}{p_{02}}, \quad \gamma = \frac{dp_{11}}{\sqrt{p_{20}}\sqrt{p_{02}}}$$

it becomes
$$\bar{r} + dr = \frac{\rho + \gamma - \alpha_1\beta_1}{\sqrt{(1+\alpha_2-\alpha_1^2)} \times \sqrt{(1+\beta_2-\beta_1^2)}} \quad \text{...........(xxxiv).}$$

When this is expanded, the mean values of α_1, β_1, α_2, β_2, and γ are of course zero. In the following tables the general formulae (xxviii), (xxix) and (xxxi) for the mean products two, three and four together of deviations of moments in samples of n are given at the head, the suffixes used and their composition in the several formulae are shown in the initial columns (omitting repetitions of dp and p to abbreviate the printed matter), and the resulting formulae for the mean products of the α's, β's and γ's required in the expansion of (xxxiv) are given in the last column, the reductions of the higher moments having been made by (xxxiii) to suit the case we are investigating, that of a normal distribution of the two characters in the material sampled. Since the first and second moments only of r are at present sought it is unnecessary to take products involving higher powers of γ than the second. The four additional formulae to γ^4 are inserted, however, to complete the formulae for the third and fourth moments when required.

As an illustration, if in equation (xxix) we put

$$u=0, \quad v=2, \quad u'=1, \quad v'=1, \quad u''=1, \quad v''=1,$$

we get
$$\text{mean } dp_{02}dp_{11}dp_{11}=\frac{1}{n^2}[p_{24}-p_{13}p_{11}-p_{13}p_{11}-p_{22}p_{02}+2p_{02}p_{11}p_{11}].$$

Dividing by $p_{20}p_{02}{}^2$ and using (xxxiii) it follows that

$$\text{mean } \beta_2\gamma^2=\frac{1}{n^2}[(3+12\rho^2)-3\rho^2-3\rho^2-(1+2\rho^2)+2\rho^2]$$
$$=\frac{1}{n^2}[2+6\rho^2],$$

the penultimate formula in table (xxxvi). It will be seen that the values to be attributed to the terms signified by the suffixes between the double rules are the right-hand sides of (xxxiii), the composition of the terms being shown in the last column of the formula. For 24 put $3+12\rho^2$, for 13 put 3ρ, for 11 put ρ, for 22 put $1+2\rho^2$, for 02 put 1, and the formula for mean $\beta_2\gamma^2$ may be written down without any division being necessary.

The formulae (xxxvii) are derived from those in (xxxv) and the suffixes between the double rules are for reference to the first columns of the latter table. Thus in the penultimate formula look up 02 11 in (xxxv) and find 2ρ. Look up 11 11 and find $1+\rho^2$. $2\rho\times(1+\rho^2)=2\rho+2\rho^3$, the first of the three component terms added together in the last column of the formula.

$$\text{mean } dp_{uv} \,.\, dp_{u'v'} = \frac{1}{n}\left[p_{u+u'\,v+v'} - p_{uv}\,p_{u'v'}\right]$$

			−		Mean value of	$=\frac{1}{n}\times$
uv	$u'v'$	$u+u'$ $v+v'$	uv	$u'v'$		
10	10	20	10	10	a_1^2	1
01	01	02	01	01	β_1^2	1
10	20	30	10	20	a_1a_2	0
01	02	03	01	02	$\beta_1\beta_2$	0
20	20	40	20	20	a_2^2	2
02	02	04	02	02	β_2^2	2
10	01	11	10	01	$a_1\beta_1$	ρ
20	02	22	20	02	$a_2\beta_2$	$2\rho^2$
10	02	12	10	02	$a_1\beta_2$	0
20	01	21	20	01	$a_2\beta_1$	0
10	11	21	10	11	$a_1\gamma$	0
01	11	12	01	11	$\beta_1\gamma$	0
20	11	31	20	11	$a_2\gamma$	2ρ
02	11	13	02	11	$\beta_2\gamma$	2ρ
11	11	22	11	11	γ^2	$1+\rho^2$

......(xxxv).

$$\text{mean } dp_{uv}.dp_{u'v'}.dp_{u''v''}$$

$$= \frac{1}{n^2}\left[p_{u+u'+u''\,v+v'+v''} - p_{u+u'\,v+v'}\,p_{u''v''} - p_{u+u''\,v+v''}\,p_{u'v'} - p_{u'+u''\,v'+v''}\,p_{uv} + 2p_{uv}\,p_{u'v'}\,p_{u''v''}\right]$$

				−		−		−		$+2\times$			Mean value of	$=\frac{1}{n^2}\times$
uv	$u'v'$	$u''v''$	$u+u'+u''$ $v+v'+v''$	$u+u'$ $v+v'$	$u''v''$	$u+u''$ $v+v''$	$u'v'$	$u'+u''$ $v'+v''$	uv	uv	$u'v'$	$u''v''$		
10	10	20	40	20	20	30	10	30	10	10	10	20	$a_1^2a_2$	2
01	01	02	04	02	02	03	01	03	01	01	01	02	$\beta_1^2\beta_2$	2
20	20	20	60	40	20	40	20	40	20	20	20	20	a_2^3	8
02	02	02	06	04	02	04	02	04	02	02	02	02	β_2^3	8
10	20	01	31	30	01	11	20	21	10	10	20	01	$a_1a_2\beta_1$	2ρ
10	01	02	13	11	02	12	01	03	10	10	01	02	$a_1\beta_1\beta_2$	2ρ
10	10	02	22	20	02	12	10	12	10	10	10	02	$a_1^2\beta_2$	$2\rho^2$
20	01	01	22	21	01	21	01	02	20	20	01	01	$a_2\beta_1^2$	$2\rho^2$
20	20	02	42	40	02	22	20	22	20	20	20	02	$a_2^2\beta_2$	$8\rho^2$
20	02	02	24	22	02	22	02	04	20	20	02	02	$a_2\beta_2^2$	$8\rho^2$
10	10	11	31	20	11	21	10	21	10	10	10	11	$a_1^2\gamma$	2ρ
01	01	11	13	02	11	12	01	12	01	01	01	11	$\beta_1^2\gamma$	2ρ
10	01	11	22	11	11	21	01	12	10	10	01	11	$a_1\beta_1\gamma$	$1+\rho^2$
20	02	11	33	22	11	31	02	13	20	20	02	11	$a_2\beta_2\gamma$	$4\rho+4\rho^3$
20	20	11	51	40	11	31	20	31	20	20	20	11	$a_2^2\gamma$	8ρ
02	02	11	15	04	11	13	02	13	02	02	02	11	$\beta_2^2\gamma$	8ρ
20	11	11	42	31	11	31	11	22	20	20	11	11	$a_2\gamma^2$	$2+6\rho^2$
02	11	11	24	13	11	13	11	22	02	02	11	11	$\beta_2\gamma^2$	$2+6\rho^2$
11	11	11	33	22	11	22	11	22	11	11	11	11	γ^3	$6\rho+2\rho^3$

...........(xxxvi).

$$\begin{aligned}\text{mean } dp_{uv}.dp_{u'v'}.dp_{u''v''}.dp_{u'''v'''} &= \text{mean } dp_{uv}.dp_{u'v'} \times \text{mean } dp_{u''v''}.dp_{u'''v'''}\\ &\quad + \text{mean } dp_{uv}.dp_{u''v''} \times \text{mean } dp_{u'v'}.dp_{u'''v'''}\\ &\quad\quad + \text{mean } dp_{uv}.dp_{u'''v'''} \times \text{mean } dp_{u'v'}.dp_{u''v''}.\end{aligned}$$

						×		+		×		+		×		Mean value of	$=\frac{1}{n^2}\times$
uv	$u'v'$	$u''v''$	$u'''v'''$	uv	$u'v'$	$u''v''$	$u'''v'''$	uv	$u''v''$	$u'v'$	$u'''v'''$	uv	$u'''v'''$	$u'v'$	$u''v''$		
10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	$\alpha_1{}^4$	3
01	01	01	01	01	01	01	01	01	01	01	01	01	01	01	01	$\beta_1{}^4$	3
10	10	20	20	10	10	20	20	10	20	10	20	10	20	10	20	$\alpha_1{}^2\alpha_2{}^2$	2
01	01	02	02	01	01	02	02	01	02	01	02	01	02	01	02	$\beta_1{}^2\beta_2{}^2$	2
20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	$\alpha_2{}^4$	12
02	02	02	02	02	02	02	02	02	02	02	02	02	02	02	02	$\beta_2{}^4$	12
10	10	10	01	10	10	10	01	10	10	10	01	10	01	10	10	$\alpha_1{}^3\beta_1$	3ρ
10	01	01	01	10	01	01	01	10	01	01	01	10	01	01	01	$\alpha_1\beta_1{}^3$	3ρ
10	10	01	01	10	10	01	01	10	01	10	01	10	01	10	01	$\alpha_1{}^2\beta_1{}^2$	$1+2\rho^2$
10	10	02	02	10	10	02	02	10	02	10	02	10	02	10	02	$\alpha_1{}^2\beta_2{}^2$	2
20	20	01	01	20	20	01	01	20	01	20	01	20	01	20	01	$\alpha_2{}^2\beta_1{}^2$	2
20	20	20	02	20	20	20	02	20	20	20	02	20	02	20	20	$\alpha_2{}^3\beta_2$	$12\rho^2$
20	02	02	02	20	02	02	02	20	02	02	02	20	02	02	02	$\alpha_2\beta_2{}^3$	$12\rho^2$
20	20	02	02	20	20	02	02	20	02	20	02	20	02	20	02	$\alpha_2{}^2\beta_2{}^2$	$4+8\rho^4$
10	20	20	01	10	20	20	01	10	20	20	01	10	01	20	20	$\alpha_1\alpha_2{}^2\beta_1$	2ρ
10	01	02	02	10	01	02	02	10	02	01	02	10	02	01	02	$\alpha_1\beta_1\beta_2{}^2$	2ρ
10	10	20	02	10	10	20	02	10	20	10	02	10	02	10	20	$\alpha_1{}^2\alpha_2\beta_2$	$2\rho^2$
20	01	01	02	20	01	01	02	20	01	01	02	20	02	01	01	$\alpha_2\beta_1{}^2\beta_2$	$2\rho^2$
10	01	20	02	10	01	20	02	10	20	01	02	10	02	01	20	$\alpha_1\beta_1\alpha_2\beta_2$	$2\rho^3$
10	10	20	11	10	10	20	11	10	20	10	11	10	11	10	20	$\alpha_1{}^2\alpha_2\gamma$	2ρ
01	01	02	11	01	01	02	11	01	02	01	11	01	11	01	02	$\beta_1{}^2\beta_2\gamma$	2ρ
20	20	20	11	20	20	20	11	20	20	20	11	20	11	20	20	$\alpha_2{}^3\gamma$	12ρ
02	02	02	11	02	02	02	11	02	02	02	11	02	11	02	02	$\beta_2{}^3\gamma$	12ρ
10	20	01	11	10	20	01	11	10	01	20	11	10	11	20	01	$\alpha_1\alpha_2\beta_1\gamma$	$2\rho^2$
10	01	02	11	10	01	02	11	10	02	01	11	10	11	01	02	$\alpha_1\beta_1\beta_2\gamma$	$2\rho^2$
10	10	02	11	10	10	02	11	10	02	10	11	10	11	10	02	$\alpha_1{}^2\beta_2\gamma$	2ρ
20	01	01	11	20	01	01	11	20	01	01	11	20	11	01	01	$\alpha_2\beta_1{}^2\gamma$	2ρ
20	20	02	11	20	20	02	11	20	02	20	11	20	11	20	02	$\alpha_2{}^2\beta_2\gamma$	$4\rho+8\rho^3$
20	02	02	11	20	02	02	11	20	02	02	11	20	11	02	02	$\alpha_2\beta_2{}^2\gamma$	$4\rho+8\rho^3$
10	10	11	11	10	10	11	11	10	11	10	11	10	11	10	11	$\alpha_1{}^2\gamma^2$	$1+\rho^2$
01	01	11	11	01	01	11	11	01	11	01	11	01	11	01	11	$\beta_1{}^2\gamma^2$	$1+\rho^2$
10	01	11	11	10	01	11	11	10	11	01	11	10	11	01	11	$\alpha_1\beta_1\gamma^2$	$\rho+\rho^3$
20	02	11	11	20	02	11	11	20	11	02	11	20	11	02	11	$\alpha_2\beta_2\gamma^2$	$10\rho^2+2\rho^4$
20	20	11	11	20	20	11	11	20	11	20	11	20	11	20	11	$\alpha_2{}^2\gamma^2$	$2+10\rho^2$
02	02	11	11	02	02	11	11	02	11	02	11	02	11	02	11	$\beta_2{}^2\gamma^2$	$2+10\rho^2$
20	11	11	11	20	11	11	11	20	11	11	11	20	11	11	11	$\alpha_2\gamma^3$	$6\rho+6\rho^3$
02	11	11	11	02	11	11	11	02	11	11	11	02	11	11	11	$\beta_2\gamma^3$	$6\rho+6\rho^3$
11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	γ^4	$2+4\rho^2+2\rho^4$

..................(xxxvii).

We are now in a position to expand (xxxiv) to terms of the fourth degree in α, β and γ, take mean values of the α, β, γ products as found for samples of normal distributions and so obtain $\bar{r}$ the mean value of the correlation coefficient in samples of n of such distributions correct to terms in $1/n^2$.

$$\begin{aligned}
&\bar{r} + dr \\
&= (\rho + \gamma - \alpha_1\beta_1) \\
&\quad \times (1 - \tfrac{1}{2}\alpha_2 + \tfrac{1}{2}\alpha_1^2 + \tfrac{3}{8}\alpha_2^2 - \tfrac{3}{4}\alpha_1^2\alpha_2 - \tfrac{5}{16}\alpha_2^3 + \tfrac{3}{8}\alpha_1^4 + \tfrac{15}{16}\alpha_1^2\alpha_2^2 + \tfrac{35}{128}\alpha_2^4) \\
&\quad \times (1 - \tfrac{1}{2}\beta_2 + \tfrac{1}{2}\beta_1^2 + \tfrac{3}{8}\beta_2^2 - \tfrac{3}{4}\beta_1^2\beta_2 - \tfrac{5}{16}\beta_2^3 + \tfrac{3}{8}\beta_1^4 + \tfrac{15}{16}\beta_1^2\beta_2^2 + \tfrac{35}{128}\beta_2^4) \\
&= \rho - \tfrac{1}{2}\{\rho(\alpha_2 + \beta_2) - 2\gamma\} \\
&\quad + \tfrac{1}{8}\{4\rho(\alpha_1^2 + \beta_1^2) + 3\rho(\alpha_2^2 + \beta_2^2) + 2\rho\alpha_2\beta_2 - 4(\alpha_2\gamma + \beta_2\gamma) - 8\alpha_1\beta_1\} \\
&\quad + \tfrac{1}{16}\{-12\rho(\alpha_1^2\alpha_2 + \beta_1^2\beta_2) - 5\rho(\alpha_2^3 + \beta_2^3) - 4\rho(\alpha_1^2\beta_2 + \alpha_2\beta_1^2) - 3\rho(\alpha_2^2\beta_2 + \alpha_2\beta_2^2) \\
&\qquad + 8(\alpha_1^2\gamma + \beta_1^2\gamma) + 6(\alpha_2^2\gamma + \beta_2^2\gamma) + 4\alpha_2\beta_2\gamma + 8(\alpha_1\alpha_2\beta_1 + \alpha_1\beta_1\beta_2)\} \\
&\quad + \tfrac{1}{128}\{48\rho(\alpha_1^4 + \beta_1^4) + 120\rho(\alpha_1^2\alpha_2^2 + \beta_1^2\beta_2^2) + 35\rho(\alpha_2^4 + \beta_2^4) \\
&\quad + 48\rho(\alpha_1^2\alpha_2\beta_2 + \alpha_2\beta_1^2\beta_2) + 20\rho(\alpha_2^3\beta_2 + \alpha_2\beta_2^3) + 32\rho\alpha_1^2\beta_1^2 + 24\rho(\alpha_1^2\beta_2^2 + \alpha_2^2\beta_1^2) \\
&\qquad\qquad + 18\rho\alpha_2^2\beta_2^2 \\
&\quad - 96(\alpha_1^2\alpha_2\gamma + \beta_1^2\beta_2\gamma) - 40(\alpha_2^3\gamma + \beta_2^3\gamma) - 32(\alpha_1^2\beta_2\gamma + \alpha_2\beta_1^2\gamma) - 24(\alpha_2^2\beta_2\gamma + \alpha_2\beta_2^2\gamma) \\
&\quad - 64(\alpha_1^3\beta_1 + \alpha_1\beta_1^3) - 48(\alpha_1\alpha_2^2\beta_1 + \alpha_1\beta_1\beta_2^2) - 32\alpha_1\beta_1\alpha_2\beta_2\} \quad \ldots\ldots\ldots\ldots\ldots\text{(xxxviii)}.
\end{aligned}$$

Whence taking mean values in samples of n of a normal distribution,

$$\begin{aligned}
\bar{r} = \rho &+ \frac{1}{8n}\{8\rho + 12\rho + 4\rho^3 - 16\rho - 8\rho\} \\
&+ \frac{1}{16n^2}\{-48\rho - 80\rho - 16\rho^3 - 48\rho^3 \\
&\qquad + 32\rho + 96\rho + 16\rho + 16\rho^3 + 32\rho\} \\
&+ \frac{1}{128n^2}\{288\rho + 480\rho + 840\rho \\
&\qquad + 192\rho^3 + 480\rho^3 + 32\rho + 64\rho^3 + 96\rho + 72\rho + 144\rho^5 \\
&\qquad - 384\rho - 960\rho - 128\rho - 192\rho - 384\rho^3 \\
&\qquad - 384\rho - 192\rho - 64\rho^3\} \\
= \rho &- \frac{1}{2n}\rho(1 - \rho^2) - \frac{3}{8n^2}\rho(1 - \rho^2)(1 + 3\rho^2) \ldots\ldots\ldots\ldots\ldots\ldots\text{(xxxix)}.
\end{aligned}$$

Or, expressing the result in terms of $n - 1$ (by changing n into $n' + 1$ and expanding) we may write to the same degree of approximation

$$\bar{r} = \rho\left[1 - \frac{1 - \rho^2}{2(n-1)}\left\{1 - \frac{1 - 9\rho^2}{4(n-1)}\right\}\right] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xl)},$$

from which follows that

$$1 - \bar{r}^2 = (1 - \rho^2)\left[1 + \frac{\rho^2}{n-1}\left\{1 - \frac{1 - 5\rho^2}{2(n-1)}\right\}\right] \ldots\ldots\ldots\ldots\ldots\text{(xli)}.$$

And, again, by squaring (xxxiv) and expanding we obtain in the same way

$$\begin{aligned}
&\bar{r}^2 + 2\bar{r}\,dr + (dr)^2 \\
&= (\rho^2 + 2\rho\gamma + \gamma^2 - 2\rho\alpha_1\beta_1 - 2\alpha_1\beta_1\gamma + \alpha_1^2\beta_1^2) \\
&\quad \times (1 - \alpha_2 + \alpha_1^2 + \alpha_2^2 - 2\alpha_1^2\alpha_2 - \alpha_2^3 + \alpha_1^4 + 3\alpha_1^2\alpha_2^2 + \alpha_2^4) \\
&\quad \times (1 - \beta_2 + \beta_1^2 + \beta_2^2 - 2\beta_1^2\beta_2 - \beta_2^3 + \beta_1^4 + 3\beta_1^2\beta_2^2 + \beta_2^4) \\
&= \rho^2 + \{2\rho\gamma - \rho^2\alpha_2 - \rho^2\beta_2\} \\
&\quad + \{\gamma^2 - 2\rho(\alpha_1\beta_1 + \alpha_2\gamma + \beta_2\gamma) + \rho^2(\alpha_1^2 + \beta_1^2 + \alpha_2^2 + \beta_2^2 + \alpha_2\beta_2)\} \\
&\quad + \{-2\alpha_1\beta_1\gamma - \alpha_2\gamma^2 - \beta_2\gamma^2 + 2\rho(\alpha_1\alpha_2\beta_1 + \alpha_1\beta_1\beta_2 + \alpha_1^2\gamma + \beta_1^2\gamma + \alpha_2^2\gamma + \beta_2^2\gamma + \alpha_2\beta_2\gamma) \\
&\qquad + \rho^2(-2\alpha_1^2\alpha_2 - 2\beta_1^2\beta_2 - \alpha_2^3 - \beta_2^3 - \alpha_1^2\beta_2 - \alpha_2\beta_1^2 - \alpha_2^2\beta_2 - \alpha_2\beta_2^2)\} \\
&\quad + \{\alpha_1^2\beta_1^2 + 2\alpha_1\alpha_2\beta_1\gamma + 2\alpha_1\beta_1\beta_2\gamma + \alpha_1^2\gamma^2 + \beta_1^2\gamma^2 + \alpha_2^2\gamma^2 + \beta_2^2\gamma^2 + \alpha_2\beta_2\gamma^2 \\
&\qquad + 2\rho(-\alpha_1^3\beta_1 - \alpha_1\beta_1^3 - \alpha_1\alpha_2^2\beta_1 - \alpha_1\beta_1\beta_2^2 - \alpha_1\beta_1\alpha_2\beta_2 \\
&\qquad - 2\alpha_1^2\alpha_2\gamma - 2\beta_1^2\beta_2\gamma - \alpha_2^3\gamma - \beta_2^3\gamma - \alpha_1^2\beta_2\gamma - \alpha_2\beta_1^2\gamma - \alpha_2^2\beta_2\gamma - \alpha_2\beta_2^2\gamma) \\
&\qquad + \rho^2(\alpha_1^4 + \beta_1^4 + 3\alpha_1^2\alpha_2^2 + 3\beta_1^2\beta_2^2 + \alpha_2^4 + \beta_2^4 + 2\alpha_1^2\alpha_2\beta_2 + 2\alpha_2\beta_1^2\beta_2 \\
&\qquad + \alpha_2^3\beta_2 + \alpha_2\beta_2^3 + \alpha_1^2\beta_1^2 + \alpha_2^2\beta_2^2 + \alpha_1^2\beta_2^2 + \alpha_2^2\beta_1^2)\} \quad \text{..........................(xlii).}
\end{aligned}$$

And taking mean values in samples of n of a normal distribution,

$$\begin{aligned}
&\bar{r}^2 + \text{mean}\,(dr)^2 \\
&= \rho^2 + \frac{1}{n}\{1 + \rho^2 - 2\rho(\rho + 4\rho) + \rho^2(2 + 4 + 2\rho^2)\} \\
&\quad + \frac{1}{n^2}\{-2 - 2\rho^2 - 4 - 12\rho^2 + 2\rho(4\rho + 4\rho + 16\rho + 4\rho + 4\rho^3) \\
&\qquad + \rho^2(-8 - 16 - 4\rho^2 - 16\rho^2)\} \\
&\quad + \frac{1}{n^2}\{1 + 2\rho^2 + 8\rho^2 + 2 + 2\rho^2 + 4 + 20\rho^2 + 10\rho^2 + 2\rho^4 \\
&\qquad + 2\rho(-6\rho - 4\rho - 2\rho^3 - 8\rho - 24\rho - 4\rho - 8\rho - 16\rho^3) \\
&\qquad + \rho^2(6 + 12 + 24 + 8\rho^2 + 24\rho^2 + 1 + 2\rho^2 + 4 + 8\rho^4 + 4)\} \\
&= \rho^2 + \frac{1}{n}\{1 - 3\rho^2 + 2\rho^4\} \\
&\quad + \frac{1}{n^2}\{-6 + 18\rho^2 - 12\rho^4\} \\
&\quad + \frac{1}{n^2}\{7 - 15\rho^2 + 8\rho^6\} \\
&= \rho^2 + \frac{1}{n}(1 - \rho^2)(1 - 2\rho^2) + \frac{1}{n^2}(1 - \rho^2)(1 + 4\rho^2 - 8\rho^4).
\end{aligned}$$

And by squaring (xxxix),

$$\bar{r}^2 = \rho^2 - \frac{1}{n}\rho^2(1 - \rho^2) - \frac{1}{2n^2}\rho^2(1 - \rho^2)(1 + 5\rho^2).$$

Hence by subtraction,

$$\text{mean } (dr)^2 = \frac{1}{n}(1-\rho^2)^2 + \frac{1}{n^2}(1-\rho^2)^2(1+5\tfrac{1}{2}\rho^2).$$

Or taking the square root

$$\sigma_r = \frac{1-\rho^2}{\sqrt{n}}\left(1 + \frac{1+5\frac{1}{2}\rho^2}{2n}\right),$$

which may be expressed in like manner with $\bar{r}$ in the form

$$\sigma_r = \frac{1-\rho^2}{\sqrt{n-1}}\left[1 + \frac{11\rho^2}{4(n-1)}\right] \quad \text{..........................(xliii)},$$

to the same degree of approximation.

It appears then from the above results that if the coefficient of correlation existing between two measured characters in a large aggregate of individuals be computed from the product moment values in small samples, these values are subject to errors from a mean value, the standard deviation of which errors may be very approximately represented by the formula

$$\frac{1-\rho^2}{\sqrt{n-1}},$$

and with greater degree of accuracy by the formula

$$\frac{1-\rho^2}{\sqrt{n-1}}\left(1 + \frac{11\rho^2}{4n}\right),$$

ρ being the coefficient of correlation between the characters in the material sampled and n being the number in the sample.

Moreover the mean value of the correlation coefficients obtained from such small samples will be less than the true coefficient of the aggregate and will be approximately represented by the formula

$$\rho\left(1 - \frac{1-\rho^2}{2n}\right),$$

the defect being very small when ρ is large, and when ρ is small being of the order 5 °/。 in samples of 10 and ·5 °/。 in samples of 100.

On the other hand the modal value of the correlation coefficients, or the most likely value in a single sample, will be greater than the true correlation coefficient (that is to say numerically greater: the correlation being supposed to be measured positively).

We have, by definition $\qquad 1/\lambda = \dfrac{\sigma_r^2}{1-\bar{r}^2},$

and so from equations (xli), (xliii) putting $n-1=n'$

$$1/\lambda = \frac{1-\rho^2}{n'}\left(1+\frac{11\rho^2}{4n'}\right)\left(1+\frac{\rho^2}{n'}\right)^{-1}$$
$$= \frac{1-\rho^2}{n'}\left(1+\frac{7\rho^2}{4n'}\right),$$

using second approximations.

And hence from (x) going now to third approximations,

$$\begin{aligned}\breve{r} &= \bar{r}\,\frac{1-1/\lambda}{1-3/\lambda}\\ &= \rho\left\{1-\frac{1-\rho^2}{2n'}+\frac{(1-\rho^2)(1-9\rho^2)}{8n'^2}\right\}\\ &\quad\times\left\{1-\frac{1-\rho^2}{n'}-\frac{(1-\rho^2)\,9\rho^2}{2n'^2}\right\}\\ &\quad\times\left\{1-\frac{3(1-\rho^2)}{n'}-\frac{(1-\rho^2)\,27\rho^2}{2n'^2}\right\}^{-1}\\ &= \rho\left\{1+\frac{3(1-\rho^2)}{2n'}+\frac{(41+23\rho^2)(1-\rho^2)}{8n'^2}\right\}.\end{aligned}$$

The excess of $\breve{r}$ over true ρ is zero if $\rho=0$ and if $\rho=1$, but if n is small and ρ fairly large the excess may be such as to make the modal value unity or greater than unity. If for instance n is so small as 4, n' being thus 3, the above approximate equation gives

$$\breve{r}=\rho+\tfrac{1}{2}\rho(1-\rho^2)+(\tfrac{41}{72}\rho+\tfrac{23}{72}\rho^3)(1-\rho^2)$$

or

$$\begin{aligned}\breve{r} &= 2{\cdot}069\rho-{\cdot}750\rho^3-{\cdot}319\rho^5\\ &= {\cdot}93 \text{ when } \rho={\cdot}5\\ &= 1{\cdot}05 \quad ,, \quad \rho={\cdot}6\\ &= 1{\cdot}14 \quad ,, \quad \rho={\cdot}7.\end{aligned}$$

The frequency distribution in the last two cases is of the J type, there being no mode within the range. The greatest frequency is at the extremity of the range, or at value unity. The interpretation of this result is clearly that such small samples as 3, 4 or 5, as might be expected, fail altogether to give by the product moment formula an approximation to the correlation coefficient. Under some circumstances the points which graphically represent the observed measures are more likely to be in a line than to have a configuration represented by any specified fractional correlation coefficient. This will happen if the correlation in the material has a larger coefficient than ·6 (approx.) when samples of four are drawn: or a larger coefficient than ·3 (approx.) when samples of three are drawn. If samples of two are drawn the coefficient of correlation is necessarily unity in the sample whatever it may be in the material*. All the distribution is concentrated at value unity and $\breve{r}$ should in this case be infinite for all values of ρ. Our approximation, neglecting terms in $1/n'^3$ etc., cannot of course show this if $n'=1$. It gives $\breve{r}$ greater than unity and so a J type, but fails to show the complete concentration at unity.

It appears from (x) that $\breve{r}$ will be infinite when $\lambda=3$ and $\bar{r}$ any value other than zero, whilst $\breve{r}$ will be zero if $\bar{r}$ is zero and λ other than 3. If $\bar{r}$, and therefore

* Supposing the material ungrouped. If it is grouped some values will be indeterminate in small samples, viz. when all observations fall into the same group.

ρ, is zero and $\lambda=3$, $\check{r}$ is indeterminate. This case is a little singular and (ix) shows that m_1 and m_2 are zero and that the frequency is therefore constant or that all values of r are equally likely, from -1 to $+1$. Remembering that $\lambda=\frac{1-\bar{r}^2}{\sigma_r^2}$ it will happen when $\sigma_r^2=\frac{1}{3}$. But when $\rho=0$ the value we have found for σ_r^2 is $\frac{1}{n-1}$ which will be $\frac{1}{3}$ when $n=4$. If therefore from material possessing zero correlation samples of four are drawn, the values of the correlation coefficients in these samples should, if the above formulae and assumed type of frequency distribution are correct, be equally distributed in value throughout the whole range.

Complete experimental confirmation of the distributions found in this paper for the product moment expression of the coefficient of correlation in small samples is difficult to obtain. The second order differences in the values of the mean and standard deviation of the distribution are necessarily small in large samples and comparable with the errors of sampling of such samples unless a very large *number* of samples is taken. On the other hand if small samples are taken the theory is not so exact and in addition the distribution tends to concentrate in certain grades if the original material is coarsely grouped, whilst fine grouping adds very much to the labour of computation.

In a paper published in *Biometrika*, Vol. VI., p. 302, "Student" gives the results of a very painstaking investigation to determine experimentally these distributions. Material based on W. R. Macdonell's measures of finger length and stature in 3000 criminals [*Biometrika*, Vol. I., p. 219] was formed having correlation with coefficient* ·66 and samples of 4, 8 and 30 were drawn. Material with correlation zero was also constructed and samples of 4 and 8 drawn.

The comparison of the actual means and standard deviations found in these experiments with those computed from the formulae in this paper is shown in the table that follows.

	1	*2*	*3*	*4*	*5*
Number in sample, n	4	8	4	8	30
Correlation coefficient in material, ρ	0	0	·66	·66	·66
Mean correlation coefficient in samples, observed ...	—	—	·5609	·6139	·661
" " " " calculated, $\bar{r}$ ...	0	0	·5933	·6317	·6542
Standard deviation of coefficient in samples, observed ...	·5512	·3731	·4680	·2684	·1001
" " " " calculated, σ_r	·5773	·3780	·4234	·2453	·1088
Number of samples taken	745	750	745	750	100
Probable error of mean	·0143	·0093	·0073	·0105	·0060
" " standard deviation	·0101	·0066	·0052	·0074	·0043

* The exact figure was ·6608.

It will be seen that the differences of observed and calculated values are for the most part several times the probable errors. Estimated in this way case (3) $\rho = \cdot 66$, $n = 4$, is the worst fit, the difference of the mean being five times and that of the standard deviation nine times the probable errors.

From the values of $\bar{r}$ and σ_r found above are calculated

	1	*2*	*3*	*4*	*5*
$\lambda = \dfrac{1 - \bar{r}^2}{\sigma_r^2} =$	3	7	3·6140	9·9905	48·325
$\check{r} = \dfrac{\lambda - 1}{\lambda - 3} \cdot \bar{r} =$	$\frac{0}{0}$	0	2·5259	·8124	·6831

The frequency distributions of Type I to fit the numbers of samples taken in the experiments and the values of $\bar{r}$ and σ_r calculated will be of the form

$$y = y_0 (1 - x)^{m_1} (1 + x)^{m_2}$$

when referred to the absolute origin and unit of measurement of r, and the constants* will be

	1	*2*	*3*	*4*	*5*
$m_1 = \frac{1}{2}(\lambda - 1)(1 - \check{r}) - 1 \quad =$	0	2	$-$·46844	·6557	7·1825
$m_2 = \frac{1}{2}(\lambda - 1)(1 + \check{r}) - 1 \quad =$	0	2	1·08243	6·3348	38·1425
$y_0 = \dfrac{N}{2^{m_1 + m_2 + 1}} \cdot \dfrac{\Gamma(m_1 + m_2 + 2)}{\Gamma(m_1 + 1) \cdot \Gamma(m_2 + 1)} =$	372·5	703·12	202·25	95·131	·0033889

When the above frequency curves are plotted they appear as shown on the diagrams pp. 112, 113 and are seen to be in fair consonance with the frequencies observed in the experiments and shown by the rectangles upon the same diagrams. They are perhaps as good an expression of these frequencies as could be found amongst the type of frequency curve assumed. The case of $\rho = 0$, $n = 4$, for which theory prescribes a horizontal straight line is seen to be very nearly so in the experiment, apart from individual fluctuations. In $\rho = 0$, $n = 8$, the curve well fits the deviations from zero correlation observed. In $\rho = \cdot 66$, $n = 4$, the asymptotic nature of the distribution towards the value unity which the fitted curve foreshadows is borne out in the samples of four drawn. With larger samples from the same class of material the displacement of the mode and the skewness of the distribution resulting from the assumed types are corroborated in the tests.

At the same time it must be admitted that there are considerable differences

* For the special case of $\rho = 0$, eqns. (ix), (xl) and (xliii) show that the curve is $y = y_0 (1 - x^2)^{\frac{1}{2}(n-4)}$, the form suggested for this case by "Student," *Biometrika*, Vol. VI. p. 306.

needing to be accounted for. The individual irregularities in the observations are more than would be expected in random samples of homogeneous material and it is probable that these jumps which make a good fit of any continuous curve whatever an impossibility are partly due to grouping. If the grouping of the original material were too coarse there would be a tendency in small samples for statistical constants to centre round certain values. Another possible source of error is in the mixing and in the drawing out of samples. Although a great deal of trouble was undoubtedly taken in these experiments, yet there always seems room for a little involuntary order in repetitions intended to go solely by chance.

The curves were planimetered and the following tables show the comparison of theoretical with observed frequencies of the grade values. It will be seen that the differences are not systematic but that the + and − errors are fairly mixed. On calculating the square contingency, χ^2, and deducing from this and the number of groups, n', the probability, P, of these differences being purely that of sampling* such probability comes out in most cases very small. It seems legitimate in this instance to see to what extent grouping will smooth down the irregularities and yet show the general resemblance and, with this in view, the differences in the columns headed e' are calculated and the grouping therein shown may be taken to indicate what is necessary to bring the probabilities within reasonable distance of expectation.

	$\rho=0,\quad n=4.$				$\rho=0,\quad n=8.$			
r	Calculated frequency m	Observed frequency	Difference e	$\frac{e^2}{m}$	Calculated frequency m	Observed frequency	Difference e	$\frac{e^2}{m}$
·925—1·0	27·94	22·5	− 5·44	1·06	·6	—	− ·6	·60
·825—	37·25	31·5	− 5·75	·89	4·1	3	− 1·1	·30
·725—	37·25	24·0	− 13·25	4·71	11·6	12	+ ·4	·01
·625—	37·25	35·0	− 2·25	·14	20·6	11·5	− 9·1	4·02
·525—	,,	34·0	− 3·25	·28	31·9	28·5	− 3·4	·36
·425—	,,	47·0	+ 9·75	2·55	42·3	46	+ 3·7	·32
·325—	,,	30·5	− 6·75	1·22	51·6	47·5	− 4·1	·33
·225—	,,	46·5	+ 9·25	2·30	59·9	70	+ 10·1	1·70
·125—	,,	44·0	+ 6·75	1·22	65·8	57·5	− 8·3	1·05
·025—	,,	32·0	− 5·25	·74	69·4	70	+ ·6	·01
$\bar{1}$·925—	,,	45·0	+ 7·75	1·61	70·3	60·5	− 9·8	1·37
$\bar{1}$·825—	,,	43·0	+ 5·75	·89	67·9	71·5	+ 3·6	·19
$\bar{1}$·725—	,,	41·0	+ 3·75	·38	63·1	76	+ 12·9	2·64
$\bar{1}$·625—	,,	37·0	− ·25	·00	56·3	63	+ 6·7	·80
$\bar{1}$·525—	,,	44·0	+ 6·75	1·22	47·1	42	− 5·1	·55
$\bar{1}$·425—	,,	40·0	+ 2·75	·20	36·9	33	− 3·9	·41
$\bar{1}$·325—	,,	38·5	+ 1·25	·04	26·1	29	+ 2·9	·32
$\bar{1}$·225—	,,	32·5	− 4·75	·61	15·3	20	+ 4·7	1·44
$\bar{1}$·125—	,,	36·0	− 1·25	·04	7·4	8	+ ·6	·05
$\bar{1}$—$\bar{1}$·125	46·56	41·0	− 5·56	·83	1·8	1	− ·8	·36
		745		20·93		750		16·83

$n'=20,\quad \chi^2=20·93,\quad P=·341.$ $\qquad n'=20,\quad \chi^2=16·83,\quad P=·601.$

* Tables for testing the goodness of fit, W. Palin Elderton, *Biometrika*, Vol. I. p. 155.

FIG. 1.

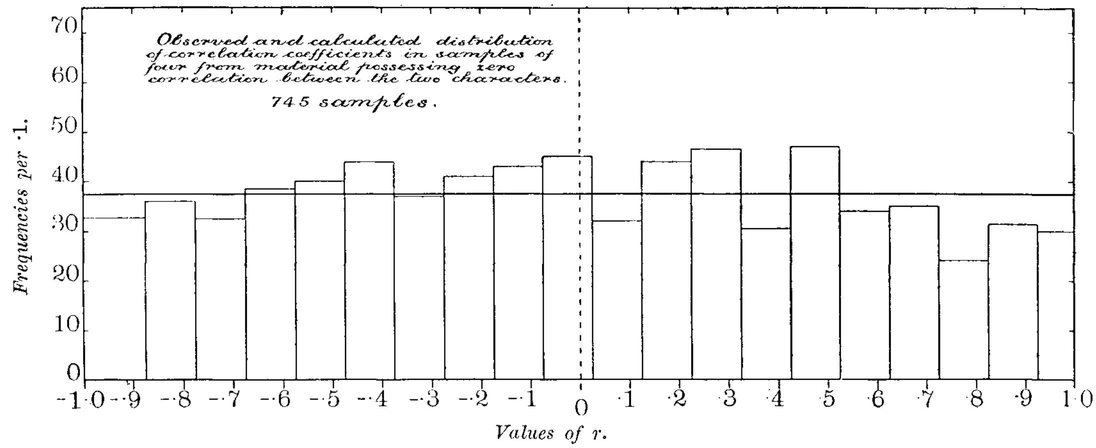

Observed and calculated distribution of correlation coefficients in samples of four from material possessing zero correlation between the two characters.
745 samples.
Frequencies per ·1.
Values of r.

FIG. 3.

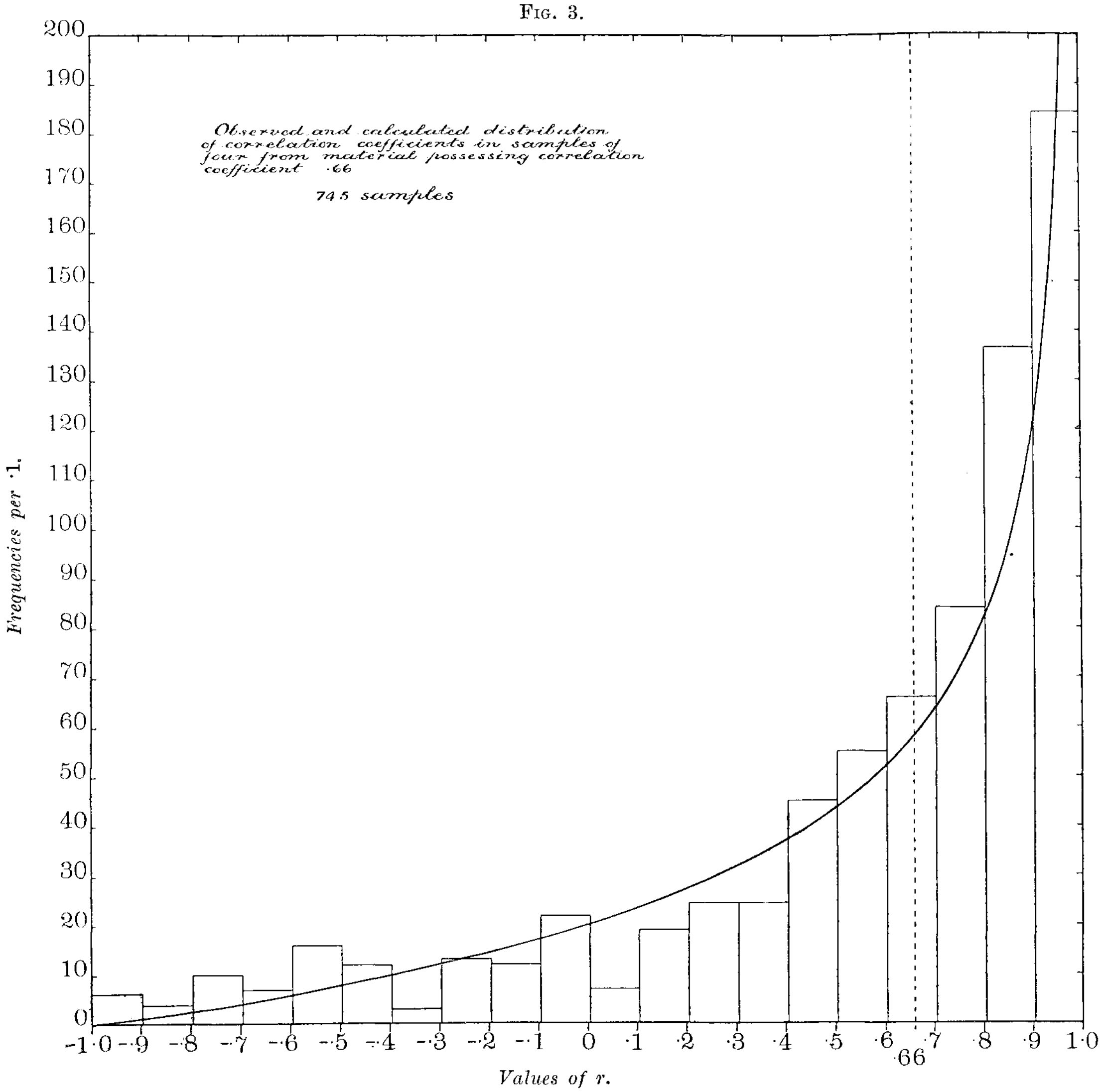
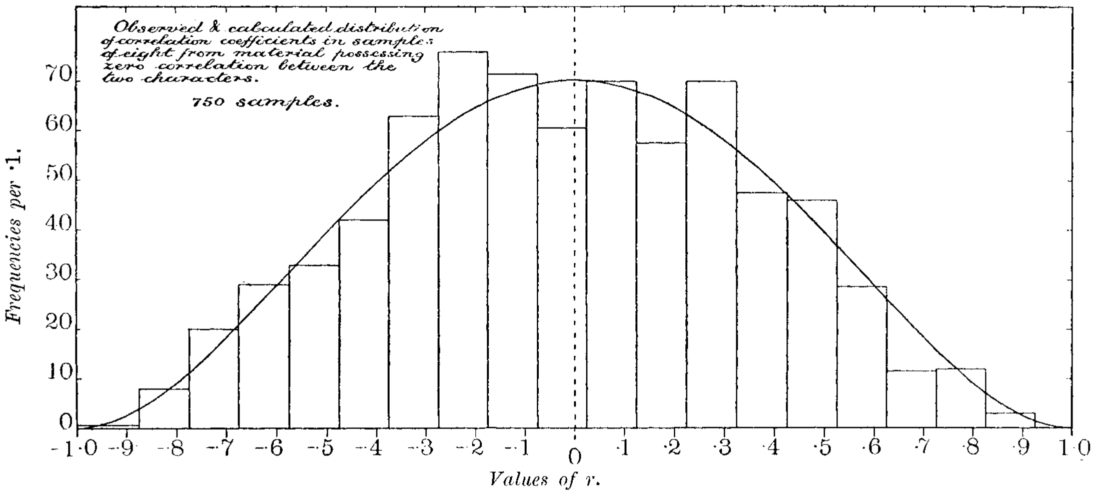
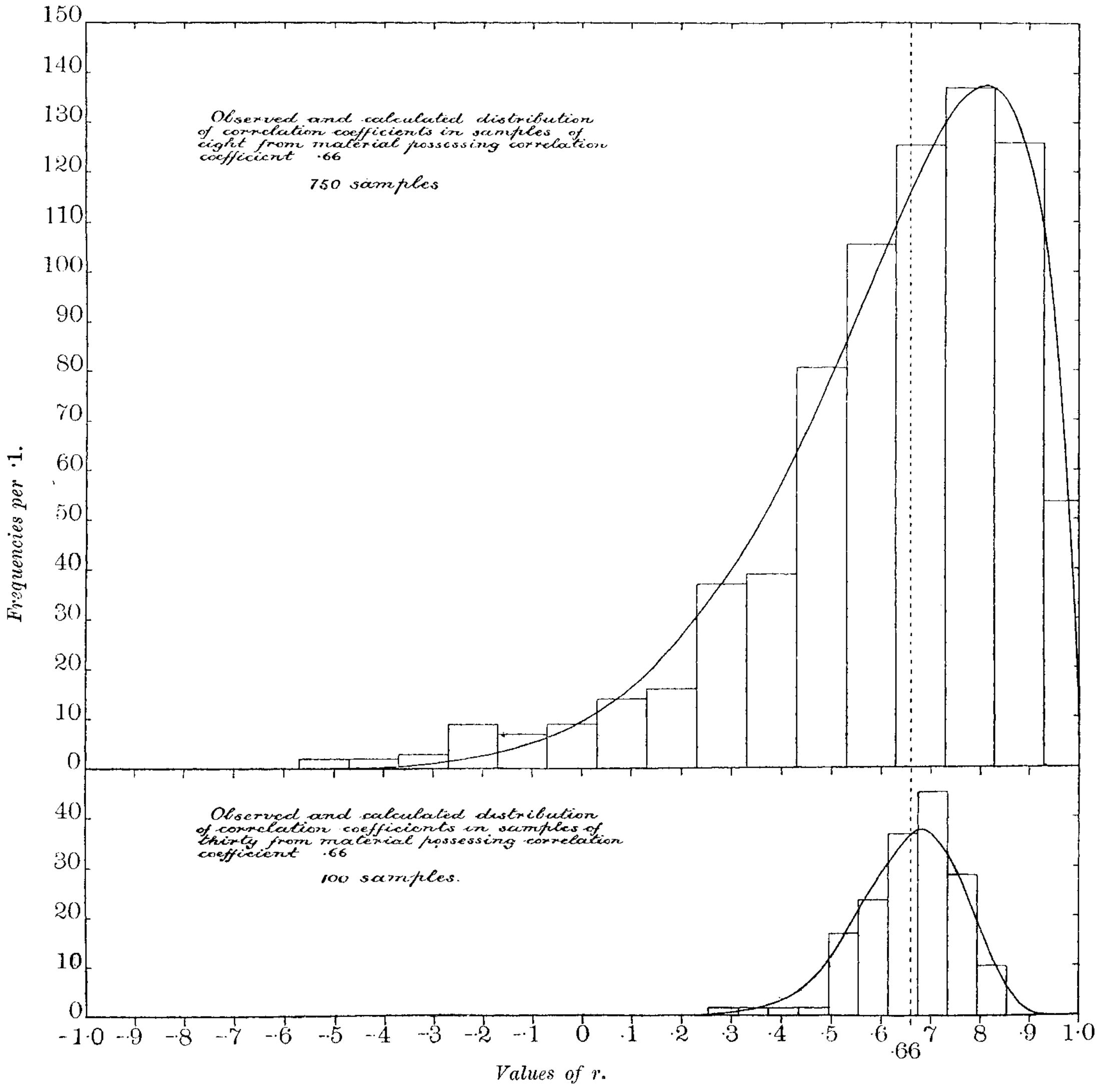

Observed and calculated distribution of correlation coefficients in samples of four from material possessing correlation coefficient ·66
745 samples
Frequencies per ·1.
Values of r.
·66

Fig. 2.



Figs. 4 and 5.



$\rho = \cdot 66, \quad n = 4.$

r	Calculated frequency m	Observed frequency	Difference e	$\frac{e^2}{m}$	Difference e'	$\frac{e'^2}{m}$
·905—1·0	230·3	175·5	− 17·2	·90	− 17·2	·90
·805—	98·9	136·5				
·705—	72·1	84	+ 20·3	3·18	+ 20·3	3·18
·605—	57·6	66				
·505—	48·0	55	+ 11·8	1·58	+ 2·0	·03
·405—	40·2	45				
·305—	34·3	24·5	− 15·0	3·52		
·205—	29·7	24·5				
·105—	25·6	19	− 21·6	9·80	− 5·1	·16
·005—	22·0	7				
$\bar{1}$·905—	18·8	22	− ·8	·02		
$\bar{1}$·805—	16·0	12				
$\bar{1}$·705—	13·5	13	− 8·7	3·06		
$\bar{1}$·605—	11·2	3				
$\bar{1}$·505—	9·0	12	+ 12·1	9·21		
$\bar{1}$·405—	6·9	16				
$\bar{1}$·305—	5·1	7	+ 8·6	8·80		
$\bar{1}$·205—	3·3	10				
$\bar{1}$·105—	1·9	4	+ 10·5	44·10		
$\bar{1}$—$\bar{1}$·105	·6	9				
	745			84·17		4·27

$n' = 10, \quad \chi^2 = 84\cdot 17, \quad P$ very small.

$n' = 4, \quad \chi^2 = 4\cdot 27, \quad P = \cdot 237.$

$\rho = \cdot 66, \quad n = 8.$

r	Calculated frequency m	Observed frequency	Difference e	$\frac{e^2}{m}$	Difference e'	$\frac{e'^2}{m}$
·925—1·0	48·9	37·5	− 11·4	2·67	− 13·2	·99
·825—	127·6	126	− 1·8	·02		
·725—	135·25	137	+ 1·6	·02	+ 1·6	·02
·625—	120·0	125·5	+ 5·4	·24	+ 5·4	·24
·525—	97·25	105·5	+ 8·1	·68	+ 8·1	·68
·425—	73·9	80·5	+ 6·5	·58	− 8·0	·51
·325—	53·5	39	− 14·6	3·96		
·225—	37·0	37	0·0	·00	+ 6·2	·41
·125—	24·25	16	− 8·3	2·82		
·025—	14·75	14	− ·8	·04		
$\bar{1}$·925—	8·5	9	+ ·5	·03		
$\bar{1}$·825—	4·75	7	+ 2·2	1·05		
$\bar{1}$·725—	2·25	9	+ 6·75	20·25		
$\bar{1}$·625—	·9	3	+ 2·1	5·22		
$\bar{1}$·525—	·25	2	+ 1·75	12·24		
$\bar{1}$·425—	·10	2	+ 1·9	36·10		
$\bar{1}$·325—						
$\bar{1}$·225—						
$\bar{1}$·125—						
$\bar{1}$—$\bar{1}$·125						
	750			85·92		2·85

$n' = 16, \quad \chi^2 = 85\cdot 92, \quad P$ very small.

$n' = 6, \quad \chi^2 = 2\cdot 85, \quad P = \cdot 722.$

$\rho = \cdot 66, \quad n = 30.$

r	Calculated frequency m	Observed frequency	Difference e	$\frac{e^2}{m}$
·795—	8·5	6	−2·5	·74
·735—	15·8	17	+1·2	·09
·675—	21·6	27	+5·4	1·35
·615—	20·9	22	+1·1	·06
·555—	15·8	14	−1·8	·21
·495—	9·8	10	+ ·2	·04
$\bar{1}$—·495	7·6	4	−3·6	1·71
	100			4·20

$n' = 7, \quad \chi^2 = 4{\cdot}20, \quad P = \cdot 650.$

It is hoped that further experiments may be shortly carried out which will have regard to the points raised and show definitely whether the distributions theoretically arrived at in this paper are good presentations of fact and whether the application of the standard types of frequency curves to the distributions of statistical constants in small samples is justified.

I am indebted to Professor Pearson for drafting the lines of this investigation and for critical supervision.