DEPARTMENT OF APPLIED STATISTICS
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON

DRAPERS' COMPANY RESEARCH MEMOIRS

BIOMETRIC SERIES. VIII

MATHEMATICAL CONTRIBUTIONS TO THE THEORY OF EVOLUTION. XVIII.

ON A NOVEL METHOD OF REGARDING THE ASSOCIATION OF TWO VARIATES CLASSED SOLELY IN ALTERNATE CATEGORIES

BY
KARL PEARSON, F.R.S.

WITH TWO ABACS CONSTRUCTED BY G. H. SOPER, M.A.

CAMBRIDGE UNIVERSITY PRESS
LONDON: FETTER LANE, E.C. 4
ALSO
H. K. LEWIS & CO., LTD., 136, Gower Street, London, W.C. 1
WHELDON & WESLEY, LTD., 2-4, Arthur Street, New Oxford Street, London, W.C. 2
Bombay, Calcutta, Madras: Macmillan & Co., Limited
Tokyo: The Maruzen-Kabushiki-Kaisha

1912

Price [illegible] *net*

Price 5s. 0d. Net. C.U.P.

UNIVERSITY OF LONDON.

The Francis Galton Laboratory for National Eugenics.

This Laboratory was founded by Sir FRANCIS GALTON, and is under the direction of Professor KARL PEARSON, F.R.S.

Assistants: DAVID HERON, M.A., D.Sc., ETHEL M. ELDERTON, AMY BARRINGTON, KATHLEEN T. RYLEY. Hon. Sec.: H. GERTRUDE JONES.

National Eugenics is the study of agencies under social control, that may improve or impair the racial qualities of future generations, either physically or mentally.

It was the intention of the Founder, that the Laboratory should serve (i) as a storehouse of statistical material bearing on the mental and physical conditions in man, and the relation of these conditions to inheritance and environment; (ii) as a centre for the publication or other form of distribution of information concerning National Eugenics; (iii) as a school for training and assisting research workers in the special problems of Eugenics.

Short courses are provided for those who are engaged in social, medical, or anthropometric work.

EUGENICS LABORATORY LECTURE SERIES.

I. The Scope and Importance to the State of the Science of National Eugenics. By KARL PEARSON, F.R.S. *Issued.* Third Edition. Price 1*s. net.*

II. The Groundwork of Eugenics. By KARL PEARSON, F.R.S. *Issued.* Price 1*s. net.*

III. The Relative Strength of Nurture and Nature. By ETHEL M. ELDERTON. *Issued.* Price 1*s. net.*

IV. On the Marriage of First Cousins. By ETHEL M. ELDERTON. *Issued.* Price 1*s. net.*

V. The Problem of Practical Eugenics. By KARL PEARSON, F.R.S. *Issued.* Second Edition. Price 1*s. net.*

VI. Nature and Nurture, the Problem of the Future. By KARL PEARSON, F.R.S. *Issued.* Price 1*s. net.*

VII. The Academic Aspect of the Science of National Eugenics. By KARL PEARSON, F.R.S. *Issued.* Price 1*s. net.*

VIII. Tuberculosis, Heredity and Environment. By KARL PEARSON, F.R.S. *Issued.* Price 1*s. net.*

UNIVERSITY COLLEGE, LONDON.

The Biometric Laboratory.

This Laboratory is intended to forward the statistical study of Biological Problems. Director: KARL PEARSON, F.R.S.

Assistants: JULIA BELL, M.A., HERBERT G. SOPER, M.A., EVELINE Y. THOMSON. Benington Studentship in Craniometry: Vacant.

Until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science.—FRANCIS GALTON.

The Laboratory is assisted by a grant from the Worshipful Company of Drapers. It provides a complete training in statistical method and assists research workers engaged on biometric problems.

All communications for *both* Laboratories should be addressed to University College, London, W.C.

QUESTIONS OF THE DAY AND OF THE FRAY.

I. The Influence of Parental Alcoholism on the Physique and Ability of the Offspring. A Reply to the Cambridge Economists. By KARL PEARSON, F.R.S. *Issued.* Price 1*s. net.*

II. Mental Defect, Mal-Nutrition, and the Teacher's Appreciation of Intelligence. A Reply to Criticisms of the Memoir on "The Influence of Defective Physique and Unfavourable Home Environment on the Intelligence of School Children." By DAVID HERON, D.Sc. *Issued.* Price 1*s. net.*

III. An Attempt to correct some of the Misstatements made by Sir VICTOR HORSLEY, F.R.S., F.R.C.S., and MARY D. STURGE, M.D., in their Criticisms of the Galton Laboratory Memoir: "A First Study of the Influence of Parental Alcoholism," &c. By KARL PEARSON, F.R.S. *Issued.* Price 1*s. net.*

IV. The Fight against Tuberculosis and the Death-rate from Phthisis. By KARL PEARSON, F.R.S. *Issued.* Price 1*s. net.*

V. Social Problems, Their Treatment, Past, Present and Future. By KARL PEARSON, F.R.S. *Issued.* Price 1*s. net.*

DEPARTMENT OF APPLIED STATISTICS, UNIVERSITY COLLEGE, LONDON.

DRAPERS' COMPANY RESEARCH MEMOIRS

C. SERIES. *Studies in National Deterioration.*

I. On the Relation of Fertility in Man to Social Status, and on the changes in this Relation that have taken place in the last 50 years. By DAVID HERON, M.A. *Issued.* Price 3*s. net.*

II. A First Study of the Statistics of Pulmonary Tuberculosis: Inheritance. By KARL PEARSON, F.R.S. *Issued.* Price 3*s. net.*

III. A Second Study of the Statistics of Pulmonary Tuberculosis: Marital Infection. By the late E. G. POPE. Edited and revised by KARL PEARSON, F.R.S. with an Appendix on Assortive Mating from Data reduced by ETHEL M. ELDERTON. *Issued.* Price 3*s. net.*

IV. The Health of the School-Child in relation to its Mental Characters. By KARL PEARSON, F.R.S. *Shortly.*

V. On the Inheritance of the Diatheses of Phthisis and Insanity. A Statistical Study based upon the Family History of 1500 Criminals. By CHARLES GORING, M.D., B.Sc. *Issued.* Price 3*s. net.*

VI. A Third Study of the Statistics of Pulmonary Tuberculosis: The Mortality of the Tuberculous and Sanatorium Treatment. By W. PALIN ELDERTON, F.I.A. and S. J. PERRY, A.I.A. *Issued.* Price 3*s. net.*

VII. On the Intensity of Natural Selection in Man (On the relation of Darwinism to the Infantile Death-rate). By E. C. SNOW, M.A. *Issued.* Price 3*s. net.*

Published by the Cambridge University Press, Fetter Lane, E.C. 4

DEPARTMENT OF APPLIED STATISTICS
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON

DRAPERS' COMPANY RESEARCH MEMOIRS

BIOMETRIC SERIES. VIII.

MATHEMATICAL CONTRIBUTIONS TO THE THEORY OF EVOLUTION. XVIII.

ON A NOVEL METHOD OF REGARDING THE ASSOCIATION OF TWO VARIATES CLASSED SOLELY IN ALTERNATE CATEGORIES

BY
KARL PEARSON, F.R.S.

WITH TWO ABACS CONSTRUCTED BY G. H. SOPER, M.A.

Published by the Cambridge University Press, Fetter Lane, E.C. 4
1912

ON A NOVEL METHOD OF REGARDING THE ASSOCIATION OF TWO VARIATES CLASSED SOLELY IN ALTERNATIVE CATEGORIES. By Karl Pearson, F.R.S.

In a memoir published twelve years ago in the *Phil. Trans.* I have shewn that in the case of the fourfold table for the correlation of two variates, *i.e.*

	A_1	A_2	Totals
B_1	a	b	$a+b$
B_2	c	d	$c+d$
Totals	$a+c$	$b+d$	N

the correlation between the means of the two variates, when each is measured in terms of its standard deviation, is*

$$r_{hk}=\frac{ad-bc}{\sqrt{(b+d)(a+c)(c+d)(a+b)}} \quad \text{.....(i).}$$

This correlation naturally vanishes with the transfer, *i.e.* $\epsilon=(ad-bc)/N$, or if the two variates are absolutely independent. Further if r_{xy} be the correlation between the two variates x and y concerned, r_{hk} must of course vanish with r_{xy}, but it is very far from equal to it, or proportional to it, as has been apparently assumed by certain recent writers on correlation, the multiplying factor varying with the values of h and k, *i.e.* with the positions where the dividing classifications are made.

The above statements depend upon the assumption that the distribution of frequency is normal or Gaussian in character.

Quite apart from any assumption as to the nature of the distribution, I have shewn that the mean square contingency † of a fourfold table is

$$\phi^2=\frac{(ab-cd)^2}{(a+d)(c+b)(a+c)(d+b)}=r_{hk}^{\ 2} \quad \text{.....(ii),}$$

and that, if we take

$$\chi^2=N\phi^2,$$

we can from the general theory of the deviations from the probable in a correlated system of variables‡ reach a quantity P giving the probability that the system is

* *Phil. Trans.* Vol. 195, A, p. 12, 1900. It is well to take as our standard arrangement of the table one in which $a+b>c+d$ and $a+c>b+d$.

† "On the Theory of Contingency and its relation to Association and Normal Correlation," *Drapers' Company Research Memoirs*, I. p. 21.

‡ *Phil. Mag.* July 1900, pp. 157—75.

really a random sample from material in which the two variates are independent. Tables for finding P from χ^2 have been calculated by Mr Palin Elderton, the well-known actuary*.

By determining the value of P, we are always in a position to ascertain the improbability of independence or $1-P$ is a proper measure of the grade of relationship. Unfortunately we do not think in millions, and to say that $P=718/10^{40}$ gives us a very poor mental estimate of the interrelationship of two quantities, compared with the simple statement that their coefficient of correlation is ·60. We do not think on such an extended scale of figures as the improbability scale provides us with, and we are bound to ask ourselves whether it is not possible to translate it into the simpler ideas of correlation. We might ask: what is the probability P' that in a population of N individuals an observed correlation r has arisen not from real association but from random sampling? We should then reduce our correlation to a probability scale. There is no difficulty about such a process at all, it depends solely on the distribution of frequency of r in random sampling. By simply equating the above value of P to P', we could then determine on a correlation scale—that is on a scale readily appreciable, the improbability of a given deviation being due to random sampling and not to true association. We should say it is as unreasonable (or as reasonable) to suppose this contingency has arisen by random sampling in a population of N individuals as to suppose that a correlation coefficient of magnitude r could arise solely from random sampling. Thus r would not be used in any way to represent features of linear or other regression lines, but solely as an artifice for transferring to an adequate mental scale improbabilities often sensible only in the 30th or 40th decimal place.

Now the improbability of r, arising from a random sampling of material having its variates unassociated, depends on the size of the standard deviation of r, and this size depends on the method by which r is determined. It is not the same when found from (i) a product moment table assumed to represent a Gaussian frequency†, or (ii) from a fourfold table representing the same frequency divided at its means‡, or again (iii) from a fourfold table of Gaussian frequency divided very far from its means§, or lastly (iv) from a product moment table for a frequency which is very far from Gaussian‖. Hence to obtain a scale of correlations by which to represent contingency improbabilities, we must select the nature of the method by which r is supposed to be reached as well as the size of the population. It will not do to say that $\cdot 67449\,(1-r^2)/\sqrt{N}$ or, for zero real association $\cdot 67449/\sqrt{N}$, is the probable error of r, because this probable error depends on the determination of r by a method which is never applicable to a fourfold table. It seems needful to select our correlation scale to be such: (i) that the standard deviation of our correlation will vary

* *Biometrika*, Vol. I. p. 155.

† Pearson and Filur, *Phil. Trans.* Vol. 191, A, p. 242, 1898.

‡ Sheppard, *Phil. Trans.* Vol. 192, A, p. 148.

§ Pearson, *Phil. Trans.* Vol. 195, A, p. 14.

‖ Sheppard, *Phil. Trans.* Vol. 192, A, p. 128. See also Pearson, *Drapers' Company Research Memoirs*, II. p. 20.

with the relative frequency of the two pairs of groups into which our categories divide the variates. This has no relation at all to the assumption of a Gaussian frequency. If a group contain n_1 individuals, its probable error is

$$\cdot 67449\sqrt{n_1(1-n_1/N)},$$

whether the variate be Gaussian or not, and the ratio of probable error to the number in the group contains the factor $1/\sqrt{n_1}$, and increases rapidly as n_1 becomes small. Any categories which contain only small percentages of the total in a fourfold division, even if the variates be purely categorical and not quantitatively measurable (*e.g.* divorced women and married women), involve increased probable error of our conclusions, and no scale of r will be satisfactory which does not recognise this; (ii) beyond this when we suppose our fourfold table to represent Gaussian material the value of r obtained by our selected scale of correlation deduced from "equality of improbability" ought to be reasonably close to the r given by the usual process on a Gaussian fourfold table.

The probable error of the coefficient of correlation for a fourfold table on the supposition that it is a random sample from uncorrelated material of Gaussian distribution is

$$\frac{\cdot 67449}{\sqrt{N}HK}\sqrt{\frac{(a+b)(a+c)(d+b)(d+c)}{N^4}} \quad \text{.....(iii)},$$

where $$H=\frac{1}{\sqrt{2\pi}}c^{-\frac{1}{2}h^2}, \qquad K=\frac{1}{\sqrt{2\pi}}c^{-\frac{1}{2}k^2},$$

h and k corresponding to the ratios of the distances of the means from the dividing lines of the categories to their respective standard deviations.

It is clear that this value increases rapidly with h or k, since

$$\frac{1}{H}\sqrt{\frac{(a+c)}{N}\times\frac{(b+d)}{N}} \text{ or } \frac{1}{K}\sqrt{\frac{(a+b)}{N}\times\frac{(c+d)}{N}}$$

increase to infinite values with h and k.

Now the value of r from a Gaussian fourfold table has to be determined from an equation of the form

$$\frac{ad-bc}{N^2HK}=r\left\{1+\underset{1}{\overset{\infty}{S}}\frac{(r^{n-1}\bar{v}_{n-1}\bar{w}_{n-1})}{\underline{|n}}\right\} \quad \text{.....(iv)},$$

where $\bar{v}_{n-1}$ and $\bar{w}_{n-1}$ are known converging factors in h and k respectively*.

It follows that the ratio of the standard deviation of r on the supposition that it is truly zero to its observed value is *approximately*

$${}_0\sigma_r/r=\frac{\sqrt{(a+b)(a+c)(d+b)(d+c)}}{\sqrt{N}(ad-bc)} \quad \text{.....(v)},$$

or approximately:

$$\chi^2=\frac{N(ad-bc)^2}{(a+b)(a+c)(d+b)(d+c)}=\frac{r^2}{{}_0\sigma_r^2} \quad \text{.....(vi)}.$$

* *Phil. Trans.* Vol. 195, p. 6.

Now let us determine the probable error of r_{hk}, for uncorrelated material,

$$r_{hk}=\frac{\gamma}{\sqrt{(b+d)(a+c)(c+d)(a+b)}}=\frac{\gamma}{\lambda},$$

where γ is zero for uncorrelated material. We have, using differentials,

$$\delta r_{hk}=\frac{\delta\gamma}{\lambda}-\frac{\gamma}{\lambda^2}\delta\lambda.$$

But for uncorrelated material we may put $\gamma=0$ after the variation due to random sampling has been allowed for, *i.e.* $\gamma=0$, but not $\delta\gamma=0$. Hence

$$(\delta r_{hk})^2=(\delta\gamma)^2\frac{1}{\lambda^2},$$

or, summing and dividing by the number of random samples,

$${}_0\sigma_{r_{hk}}=\frac{1}{\lambda}\sigma_\gamma.$$

To find $\sigma_{r_{hk}}$ therefore for uncorrelated material, we require to find σ_γ.

Now $$\gamma=ad-bc;$$

$$\therefore\ \delta\gamma=a\,\delta d+d\,\delta a-b\,\delta c-c\,\delta b.$$

Square, sum for all random samples and remember that

$$\sigma_a^2=a\left(1-\frac{a}{N}\right),$$

and $$\sigma_a\sigma_b r_{ab}=-\frac{ab}{N},$$

we find

$$\sigma_\gamma^2=a^2d\left(1-\frac{d}{N}\right)+d^2a\left(1-\frac{a}{N}\right)+b^2c\left(1-\frac{c}{N}\right)+c^2b\left(1-\frac{b}{N}\right)$$
$$-\frac{2a^2d^2}{N}-\frac{2b^2c^2}{N}+\frac{2abcd}{N}+\frac{2abcd}{N}+\frac{2abcd}{N}+\frac{2abcd}{N}$$

$$=ad\,(a+d)+bc\,(b+c)-\frac{4\,(a^2d^2-2adbc+b^2c^2)}{N}$$

$$=ad\,(a+d)+bc\,(b+c)-\frac{4\,(ad-bc)^2}{N}$$

$$=(ad-bc)(a+d)+Nbc-\frac{4\,(ad-bc)^2}{N}$$

$=Nbc$, for truly uncorrelated material.

But if $ad=bc$,

$$b=\frac{b\,(a+b+c+d)}{N}=\frac{(b+d)(a+b)}{N}$$

and $$c=\frac{c\,(a+b+c+d)}{N}=\frac{(a+c)(c+d)}{N}$$

Hence
$$\sigma_\gamma^2 = \frac{(a+b)(a+c)(d+b)(d+c)}{N} \quad \text{.....(vii)},$$
and as a result
$$_0\sigma_{r_{hk}} = \frac{1}{\lambda}\,\sigma_\gamma = \frac{1}{\sqrt{N}} \quad \text{.....(viii)}.$$

Thus the probable error of r_{hk} for material with zero association takes the simple form of ·67449 $\frac{1}{\sqrt{N}}$, precisely the value of the probable error of a zero r found from product moment formula, and quite independent of the division between the categories.

We have accordingly
$$\frac{r^2_{hk}}{_0\sigma^2_{r_{hk}}} = \frac{N\,(ad-bc)^2}{(a+b)(a+c)(d+b)(d+c)} \quad \text{.....(ix)}.$$
Thus χ which gives the probability of the observed mean square contingency is actually equal to the ratio $r_{hk}/_0\sigma_{r_{hk}}$ and approximately equal to $r/_0\sigma_r$; in other words, for the simple case of a fourfold table χ which determines the probability that the system is a random sample from unassociated material is really the ratio of an observed correlation on the Gaussian hypothesis of distribution to its probable error on the assumption of unassociated variates*.

It will be clear, however, that $\chi = r/_0\sigma_r$ will only give a very rough approximation to the value of r, since all terms but the first in the series for r in (iv) must be

* If ϕ^2 be the mean square contingency, then it is easy to show that $_0\sigma_\phi$, i.e. the standard deviation of ϕ, if there were no association of variates $= 1/\sqrt{N}$. Thus we have $\phi/_0\sigma_\phi = \chi$, or the ratio of ϕ to its standard deviation is also χ. In the same manner the standard deviation of a Yule "coefficient of association," $Q = (ad-bc)/(ad+bc)$ is
$$\sigma_Q = \frac{(1-Q^2)}{2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}},$$
and
$$_0\sigma_Q = \frac{1}{2}\sqrt{\frac{d}{bc}+\frac{1}{b}+\frac{1}{c}+\frac{a}{bc}} = \frac{1}{2}\sqrt{\frac{N}{bc}},$$
but as before $b = (b+d)(a+b)/N$, $c = (a+c)(c+d)/N$, hence
$$_0\sigma_Q = \frac{1}{2}\sqrt{\frac{N^3}{(a+c)(c+d)(b+d)(a+b)}}$$
$$= \frac{1}{2}\,\frac{N^{\frac{3}{2}}\sqrt{(a+c)(c+d)(b+d)(a+b)}}{(a+c)(c+d)(b+d)(a+b)}$$
$$= \frac{1}{2}\,\frac{1}{\sqrt{N}}\,\frac{\sqrt{(a+c)(c+d)(b+d)(a+b)}}{bc}$$
$$= \frac{1}{\sqrt{N}}\,\frac{\sqrt{(a+c)(c+d)(b+d)(a+b)}}{ad+bc}.$$
Hence if we use the constants of the observed material
$$\frac{Q}{_0\sigma_Q} = \frac{\sqrt{N}\,(ad-bc)}{\sqrt{(a+c)(c+d)(b+d)(a+b)}} = \chi, \text{ as before} \quad \text{.....(x)}.$$
Or, the improbability of the sample as a sample of uncorrelated material is the same when found from Q as from r_{hk} or from contingency, all are expressible in terms of χ.

neglected in order to obtain it. But the relation pointed out between the correlation of the variates and the correlation of their means to χ suggests a possible and not unreasonable scale of reducing contingency probabilities of independence to correlation probabilities of independence. In order to think on a readily apprehensible scale—which that of excessive improbabilities is not—find the correlation which would give the same improbability as the contingency does on the hypothesis that the two variates are unassociated. The correlation is thus used—not as a measure of regression, of which we know nothing in the case of a purely categorical fourfold division, but as a standard of relative improbability.

It will be at once obvious that if we are to carry out this suggestion of a correlation measure of improbability we want a very considerable extension of existing tables. We need

(i) A large extension of Palin Elderton's table for goodness of fit, shewing the values of P, or better log P, for such high values of χ^2 as occur in contingency tables with populations running from a few hundred to two or three thousand.

(ii) A nearly new table indicating the probability that in uncorrelated material a value of r will be reached which is many times its standard deviation; at present we only know these improbabilities up to about 5 or 6, on the hypothesis that the values of r follow a Gaussian curve, and this is inadmissible.

(iii) A table giving on some reasonable hypothesis the values of ${}_0\sigma_r$ for various relative frequencies of the variates in the fourfold classification.

Before we consider such tables it is well to note some points which arise in dealing with the epistemological side of contingency. If we are dealing with a table

Contingency Table.

	A_1	A_2	.....	A_m	Totals
B_1	n_{11}	—	.....	—	n_{11}
B_2	—	n_{22}	.....	—	n_{22}
⋮	—	—	.....	—	—
⋮	—	—	.....	—	—
B_m	—	—	.....	n_{mm}	n_{mm}
Totals	n_{11}	n_{22}	.....	n_{mm}	N

of the following type, we find that, the frequency occurring only in the diagonal column and there being $m \times m$ cells, the coefficient, C_2, of mean-square contingency

$$= \sqrt{\frac{m-1}{m}}.$$

For the case of the fourfold table for example

	A_1	A_2	Totals
B_1	a	—	a
B_2	—	d	d
Totals	a	d	N

we find that $C_2 = \cdot 707$.

Mr G. U. Yule in his recent *Introduction to the Theory of Statistics*, p. 66, suggests that the coefficient of contingency should not be used when m is less than 5 or so. But such advice could only result from wholly overlooking the essential nature of a contingency coefficient. The true position is that it is not comparable with a coefficient of correlation for like tables assumed to be Gaussian unless we use a 5×5 or finer classification. It actually measures the probability that the observed results arise from independent material, whatever be the classification.

A very little consideration will show that the table

$$\begin{array}{c|c} a & 0 \\ \hline 0 & d \end{array}$$

ought not, in the absence of knowledge as to the finer classification of a and d, to give a unity coefficient. In my conception of contingency $C_2 = 1$ marks absolute dependence, *i.e.* every individual A is associated with its own individual B. But a table of the above kind gives us no information with regard to the distribution of the a group of A_1's associated with B_1's, or of the d group of A_2's associated with B_2's. Such a coefficient as Mr Yule's coefficient of association seems to me absolutely misleading on this very account, because it gives such tables a complete association of the variates, or a unity value for their coefficient. It is perfectly true that if we *assume* the Gaussian distribution of frequency, then a wasp-waist distribution like $\begin{array}{c|c} a & 0 \\ \hline 0 & d \end{array}$ is impossible unless the correlation is perfect. But the very essence of any theory of contingency or association is to proceed from purely logical grounds and avoid any assumption that our distribution is quantitative much less that it is Gaussian. Hence when the coefficient of contingency gives for a fourfold table with only entries in the diagonal cells ·707 and for a 10×10 fold table with only entries in the diagonal cells ·949, while a Yule association coefficient gives 1·0 in both cases, this is not as Mr Yule seems to argue a disadvantage of the contingency method, but one of its chief merits. It indicates how we pass to closer and closer relationship as we classify more finely, and any coefficient which neglects this essential is to that extent, I hold, defective. There ought to be no attempt to modify the coefficient of contingency so as to raise $m \times m$ tables when m is small up to correlation values, for such values assume some further knowledge not

conveyed in the table, *e.g.* that the distribution is Gaussian, or that wasp-waist distributions are impossible.

Consider for example the following table. Actually it is purely hypothetical and small frequencies would occur where I have put empty cells, but I have done this to emphasise the principle involved, which would have acted equally, but have been obscured had I given a real table:

Mothers—before birth of child.

Mothers—after birth of child.		Employed			Unemployed	Totals
		Factory	Charing	Work taken at home		
Employed	Factory	540	—	—	—	540
	Charing	58	20	10	—	88
	Work taken at home ...	70	30	22	—	122
	Unemployed ...	—	—	—	250	250
	Totals	668	50	32	250	1000

Now, if the division here is between "employed" and "unemployed" mothers only, the coefficient of correlation on the assumption of a Gaussian frequency is unity, and the coefficient of association is also unity. Against these we find that the coefficient of contingency is only ·707. There cannot be I think a doubt that this is the better estimate. It leaves ·293 over in reserve until we know something of the sub-classifications of mothers' employment before and after the birth of the children. In the example we see that the degree of employment changes after the birth and that a number of the factory workers tend to take in work or to go charing. The coefficient of contingency is now $C_2 = \cdot 755$. It will be clear from such an illustration that we should have got a poorer result had we tried to correct contingency on the basis of tables with diagonal cells only occupied being representable by a coefficient of value unity. Contingency very properly allows for the extent of our ignorance, the coefficient of association does not. The correlation coefficient assumes a knowledge of the exact character of the distribution, and even if a Gaussian distribution be adopted, still no trained statistician would dream of applying it to a table of the form $\begin{array}{c|c} a & 0 \\ \hline 0 & d \end{array}$ and argue that the correlation was therefore perfect.

According to the view taken here we should anticipate that when a fourfold table really represents approximately Gaussian material, then the value of $r/_0\sigma_r$ will give a probability fairly closely approaching that deduced from the contingency; on the other hand when the material has no approach to a Gaussian distribution the value of r found by equality of improbabilites will be higher than that deduced by a simple fourfold correlation table.

To sum up then, the present paper proposes to deal with tables of few cells by using the probability P determined from the square contingency χ^2. I can see

absolutely no valid theoretical objection to this method of reckoning the relationship of two characters. Practically it suffers from the transcendent difficulty of mentally appreciating the relative differences of indefinitely large improbabilities. In order to surmount this difficulty I propose to think in a scale of correlations; I ask what r would have equal improbability if it arose from a random sampling of Gaussian material at the same dividing lines. In a paper now in type I have given tables from which the probable error of r can be readily found for a given division. I use these tables to find ${}_0\sigma_r$. From an extended Elderton's Table for "Goodness of Fit," I find log P. I then determine on what appears to be a reasonable hypothesis a value of the correlation coefficient which would be equally improbable; and thus reduce my improbability of independence to a mentally apprehensible scale. Thus the coefficient of correlation is merely used as a standard of improbability, and we pledge ourselves to no hypothesis as to frequency distribution, of which in many cases we know nothing. Still as we wish to approach fairly closely to the actual value of the correlation when the distribution is Gaussian, we select by preference a standard scale of correlation improbabilities, which will not contradict Gaussian results, when the fourfold table is of that character. We should not anticipate absolute agreement, for the reasons already stated, and it would be a sufficient justification of our method, if the results obtained by it, when the material is truly Gaussian, lie within a range limited by twice the probable error taken on either side of the Gaussian value.

We have next to consider how the frequency of correlation coefficients for an actual value zero is to be distributed in large random samples. We cannot use a normal curve of standard deviation ${}_0\sigma_r$; for it is quite obvious that the tails of this will extend beyond the limits -1 to $+1$, and although such a curve is quite legitimate for ordinary probabilities in the neighbourhood of $r=0$, it is wholly inadequate when we have to ask for example what is the probability that r will equal 0·8, when its actual value is zero, for say a sample of 1000. The curve of distribution of r must be symmetrical about $r=0$, and vanish for $r=\pm 1$. The only one of my generalised frequency curves* which satisfies these conditions is Type II, *i.e.*

$$y = y_0 (1-x^2)^m \quad \text{.....(xi).}$$

If N be the total frequency, $N = y_0 \int_{-1}^{+1} (1-x^2)^m dx$, and

$$N\mu_2 = N\sigma^2 = y_0 \int_{-1}^{+1} (1-x^2)^m x^2 dx$$

$$= -y_0 \int_{-1}^{+1} \frac{x d\,(1-x^2)^{m+1}}{2\,(m+1)}$$

$$= y_0 \int_{-1}^{+1} \frac{(1-x^2)^{m+1} dx}{2\,(m+1)} = \frac{1}{2\,(m+1)} (N - N\mu_2);$$

hence $$\sigma^2 = \frac{1}{2m+3}, \text{ or } m = \frac{1}{2}\left(\frac{1}{\sigma^2} - 3\right) \quad \text{.....(xii).}$$

* *Phil. Trans.* Vol. 186, A, p. 372.

We now want y_0. Remembering that the range is 2, we have*

$$y_0 = N\frac{\Gamma(m+1\cdot5)}{\sqrt{\pi}\,\Gamma(m+1)} \quad\text{.....(xiii).}$$

But σ is small, *i.e.* ·07 would be large for σ and therefore σ^2 large if equal to ·005. Thus $1/\sigma^2$ will be small if it is 200 and m small if only of order 100. We may therefore safely use Stirling's theorem to obtain the Γ-functions. Thus

$$\begin{aligned}\frac{\Gamma(m+1\cdot5)}{\Gamma(m+1)} &= \frac{\sqrt{2\pi(m+1\cdot5)}}{\sqrt{2\pi(m+1)}}\,\frac{e^{-(m+1\cdot5)}}{e^{-(m+1)}}\,\frac{(m+1\cdot5)^{m+1\cdot5}}{(m+1)^{m+1}}\\ &= e^{-\cdot5}\times\frac{m+1\cdot5}{\sqrt{m+1}}\times\left(\frac{m+1\cdot5}{m+1}\right)^{m+1}\\ &= e^{-\cdot5}\times\frac{m+1\cdot5}{\sqrt{m+1}}\left(1+\frac{\cdot5}{m+1}\right)^{m+1}\\ &= \frac{m+1\cdot5}{\sqrt{m+1}} = \sqrt{m+1}+\frac{\cdot5}{\sqrt{m+1}}\\ &= \sqrt{m+1}\left\{1+\frac{1}{2(m+1)}\right\}\\ &= \frac{1}{\sqrt{2}\sigma}\text{ nearly.}\end{aligned}$$

Hence we can take for our frequency curve for r

$$y = \frac{N}{\sqrt{2\pi}\sigma}(1-x^2)^{\frac{1}{2}\left(\frac{1}{\sigma^2}-3\right)} \quad\text{.....(xiv).}$$

When the sample is very small we must retain the full value

$$y = \frac{N}{\sqrt{\pi}}\,\frac{\Gamma\left(\frac{1}{2\sigma^2}\right)}{\Gamma\left(\frac{1}{2}\left(\frac{1}{\sigma^2}-1\right)\right)}(1-x^2)^{\frac{1}{2}\left(\frac{1}{\sigma^2}-3\right)} \quad\text{.....(xv).}$$

This agrees in the special case for product-moment r's when $\sigma^2 = \frac{1}{n-1}$ with the form proposed by "Student" in *Biometrika*, Vol. VI. p. 306 and experimentally justified by him.

We have next to find the area of the tails of this curve beyond a given value r†, to measure the improbability that with no correlation a random sample could give a value r. We clearly have

$$\begin{aligned}P &= 2\int_r^1 \frac{1}{\sqrt{2\pi}\sigma}(1-x^2)^{\frac{1}{2}\left(\frac{1}{\sigma^2}-3\right)}dx = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}\int_r^1(1-x^2)^m\,dx\\ &= -\sqrt{\frac{2}{\pi}}\frac{1}{\sigma}\frac{1}{2(m+1)}\int_r^1\frac{1}{x}\,d(1-x^2)^{m+1}\\ &= \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}\left[\frac{(1-r^2)^{m+1}}{r}\frac{1}{2(m+1)} - \frac{1}{2(m+1)}\int_r^1\frac{1}{x^2}(1-x^2)^{m+1}\,dx\right],\end{aligned}$$

* *Phil. Trans.* Vol. 186, A, p. 372.

† r is to be treated as a quantity without sign, a mere numerical quantity and therefore both tails of the frequency distribution are taken—this is the origin of the first factor 2 in the value of P.

and continuing the integration in this way by parts we have

$$P=\sqrt{\frac{2}{\pi}}\frac{1}{\sigma}\frac{(1-r^2)^{m+1}}{r}\frac{1}{2(m+1)}\left\{1-\frac{1}{2(m+2)}\frac{1-r^2}{r^2}+\frac{1.3}{2(m+2)\,2(m+3)}\left(\frac{1-r^2}{r^2}\right)^2\right.$$
$$\left.-\frac{1.3.5}{2(m+2)\,2(m+3)\,2(m+4)}\left(\frac{1-r^2}{r^2}\right)^3+\dots\right\}.$$

This series converges with considerable rapidity if σ be not greater than ·07 and r moderately large. If $1/\sigma^2=s$, $2m=s-3$, and if we write $\lambda=(1-r^2)/r^2$, we have

$$P=\sqrt{\frac{2}{\pi}}\sqrt{s}\frac{1}{s-1}\frac{(1-r^2)^{\frac{1}{2}(s-1)}}{r}\left\{1-\frac{1.\lambda}{s+1}+\frac{1.3.\lambda^2}{(s+1)(s+3)}-\frac{1.3.5.\lambda^3}{(s+1)(s+3)(s+5)}+\dots\right\}$$
$$\dots\dots\dots\text{(xvi)},$$

whence P may be fairly easily calculated.

The series is a semi-converging one, which is satisfactory enough until r gets small and therefore λ large. When r is small or σ very large (xvi) fails to give the result closely enough*. In these cases the value of the integral $\int_x^1(1-x^2)^m\,dx$ must be found from other considerations. Thus

$$\int_x^1(1-x^2)^m\,dx=\int_x^1 e^{m\log(1-x^2)}\,dx=\int_x^1 e^{-mx^2}\times e^{-m\left(\frac{x^4}{2}+\frac{x^6}{3}+\dots\right)}\,dx$$
$$=\int_x^1 e^{-mx^2}\left(1-m\frac{x^4}{2}-m\frac{x^6}{3}+m^2\frac{x^8}{8}+\dots\right)dx.$$

Let $mx^2=\frac{1}{2}z^2$, then

$$\int_x^1(1-x^2)^m\,dz=\frac{1}{\sqrt{2m}}\int_{\sqrt{2m}r}^{\sqrt{2m}}e^{-\frac{1}{2}z^2}\left(1-\frac{1}{8m}z^4-\frac{1}{24m^2}z^6+\frac{1}{128m^2}z^8+\dots\right)dz.$$

Now the integrals

$$\frac{1}{\sqrt{2\pi}}\int_z^\infty e^{-\frac{1}{2}z^2}\,dz \quad\text{and}\quad \frac{1}{\sqrt{2\pi}}\int_z^\infty e^{-\frac{1}{2}z^2}z^n\,dz$$

are tabled integrals, the first being the usual probability integral (*Biometrika*, Vol. II. p. 182), and the second being the incomplete normal moment function which has been calculated for $n=1$ to 10 (*Biometrika*, Vol. VI. p. 66). Thus

$$P=\frac{2}{\sigma}\frac{1}{\sqrt{1/\sigma^2-3}}\{\mu_0(\sqrt{2m})-\mu_0(\sqrt{2m}r)\}-\frac{1}{8m}\{\mu_4(\sqrt{2m})-\mu_4(\sqrt{2m}r)\}$$
$$-\frac{1}{24m^2}\{\mu_6(\sqrt{2m})-\mu_6(\sqrt{2m}r)\}+\frac{1}{128m^2}\{\mu_8(\sqrt{2m})-\mu_8(\sqrt{2m}r)\}-\text{etc.}\dots\text{(xvii)},$$

where
$$\mu_n(z)=\frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{1}{2}z^2}z^n\,dz.$$

* For most values of σ for $r=$ ·02 the two formulae coincide for practical purposes, so that the formula to be given below may be used for values of $r=$ ·02 or under. For $\sigma=$ ·08, we cannot use (xvi) for $r=0$·3.

As a matter of fact $\sqrt{2m}$ may be treated as infinite compared with $\sqrt{2mr}$ for values of $r = 0{\cdot}2$ and less, in which case

$$\mu_{2n}(\sqrt{2m}) = (2n-1)(2n-3)(2n-5)\ldots\ldots 1 \times {\cdot}5.$$

It will be found as a rule unnecessary to go beyond the terms in μ_4*.

From (xvi) and (xvii) Table I has been constructed; this gives the value of $-\log P$ for each value of r and ${}_0\sigma_r$. In other words it expresses the probability that with a given probable error of a zero coefficient (*i.e.* ${\cdot}67449\,{}_0\sigma_r$) a given value of the correlation will arise from this uncorrelated material in a random sample of definite size. The size of the sample and the nature of the process by which r is obtained are indifferent, provided regard is paid to them in determining ${}_0\sigma_r$.

Table II is the required extension of Palin Elderton's Table of Goodness of Fit (see *Biometrika*, Vol. I. p. 159). It gives the value of $-\log P$ for $n' = 4$, *i.e.* for four groups, from $\chi^2 = 1$ to $\chi^2 = 25{,}000$. This enables us to ascertain the improbability of a given χ^2, even when that improbability is only significant in the 5000th place of figures.

Table III gives the χ^2, which would have the same improbability as the r of Table I, and is obtained by simple interpolation from Table II.

Table IV replaces χ^2 by $\log \chi^2$ and forms a reasonable working table. Given $\log \chi^2$ from the fourfold table and ${}_0\sigma_r$, we can find the value of r which expresses the same improbability.

Thus far we have not even selected our scale of correlation, which is wholly determined by the choice of ${}_0\sigma_r$. We might take ${}_0\sigma_r$ simply equal to $1/\sqrt{n}$, but this would not be a really satisfactory scale of correlation improbabilities. The reason is obvious; it supposes a knowledge never conveyed by a fourfold table, *i.e.* the knowledge involved in our having the material in a large number of equal-ranged cells. Very naturally, therefore, we avoid this scale, for it certainly would not give at all comparable values of r for those cases of fourfold table where the material is known, or may legitimately be supposed, to be Gaussian. Accordingly we adopt for ${}_0\sigma_r$ the value as given by a fourfold Gaussian table, *i.e.*

$${}_0\sigma_r = \frac{1}{\sqrt{N}} \times \chi_{a_1} \times \chi_{a_2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xviii)},$$

where χ_{a_1} and χ_{a_2} are respectively

$$\frac{\sqrt{\frac{1}{2}(1+a_1)\frac{1}{2}(1-a_1)}}{H} \quad \text{and} \quad \frac{\sqrt{\frac{1}{2}(1+a_2)\frac{1}{2}(1-a_2)}}{K},$$

a table of which function has been recently published by me, and is reproduced here as Table V. It enables us at once to determine ${}_0\sigma_r$.

Finally, Table IV has been converted into an "abac," upon which the value of r—the "equal improbability correlation"—can be read off as soon as $\log \chi^2$ and ${}_0\sigma_r$

* For example for $r = 0{\cdot}1$ and ${}_0\sigma_r = {\cdot}03$, $-\log P = 3{\cdot}072$ by (xvi), it equals $3{\cdot}076$ from (xvii) and $3{\cdot}066$ from the Gaussian, or probability integral. The most troublesome values were those for ${}_0\sigma_r = {\cdot}08$, $r = 0{\cdot}2$ and $0{\cdot}3$. They were finally determined as $1{\cdot}924$ and $3{\cdot}903$, but they cannot be guaranteed to a unit in the last figure.

are determined to two figures. This abac was constructed in the following manner: Accurate curves were drawn of the values of each ${}_0\sigma_r$ of $\log \chi^2$. From these curves the values of $\log \chi^2$ were read off for each value of r proceeding by ·01, from ·05 to ·95. It was then possible to plot the family of curves which for each r give the relationship of $\log \chi^2$ and ${}_0\sigma_r$. I owe this excellent diagram to Mr G. H. Soper. The bulk of the laborious calculating work on the tables has been carried out by Miss Julia Bell.

In the course of our investigations two additional tables have been calculated. In the first place it was needful to extend Sheppard's Tables far beyond the limits of published work on the probability integral in order to compare how far it was possible to trust the Gaussian to give the distribution of frequency of correlation coefficients obtained by sampling independent material. As we have seen, the Gaussian is of no service for this purpose except for very low values of r (*e.g.* 0·1 or less). But the table has independently considerable value, and most statisticians will remember cases when they have had laboriously to calculate

$$F = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{1}{2}x^2}\, dx$$

beyond the usual limit of $x = 5$. I reproduce this table here as Table VI. It was calculated by aid of Schölmilch's formula*, *i.e.*

$$F = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{1}{2}x^2}\, dx = \frac{1}{\sqrt{2\pi}}\, x e^{-\frac{1}{2}x^2} \left\{ \frac{1}{x^2} - \frac{1}{x^2(x^2+2)} + \frac{1}{x^2(x^2+2)(x^2+4)} \right.$$
$$- \frac{5}{x^2(x^2+2)(x^2+4)(x^2+6)} + \frac{9}{x^2(x^2+2)(x^2+4)(x^2+6)(x^2+8)}$$
$$\left. - \frac{129}{x^2(x^2+2)(x^2+4)(x^2+6)(x^2+8)(x^2+10)} - \dots \right\} \dots\dots\dots$$

and the table gives $-\log F$, for ease of interpolation. It was calculated to seven decimal places, but only five are retained, as the seventh figure was not trustworthy and occasionally the sixth is doubtful.

It remains to examine some of the correlations found for fourfold tables by this novel process, and to compare the value as found by the assumption of Gaussian frequency.

Illustration I. The following table is given by me for good and bad temper in pairs of brothers (*Phil. Trans.* Vol. 195 A, p. 147).

First Brother

Second Brother	Good temper	Bad temper	Totals
Good temper	330	255	585
Bad temper	255	454	709
Totals	585	709	1294

* *Compendium der höheren Analysis*, Bd. II. S. 270, Braunschweig, 1879.

Here $$\log \chi^2 = 1{\cdot}9614,$$
$$\tfrac{1}{2}(1+a_1) = \tfrac{1}{2}(1+a_2) = {\cdot}5479.$$
Hence, by Table V, $$\chi_{a_1} = \chi_{a_2} = 1{\cdot}2566,$$
and $$_0\sigma_r = \frac{1}{\sqrt{1294}} \times \chi_{a_1} \times \chi_{a_2} = {\cdot}0439.$$

Interpolating from Table IV we have for $_0\sigma_r = {\cdot}0439$,
$$r = 0{\cdot}3, \quad \log \chi^2 = 1{\cdot}7596,$$
$$r = 0{\cdot}4, \quad \log \chi^2 = 2{\cdot}0025.$$
Hence $$\log \chi^2 = 1{\cdot}9614, \quad r = {\cdot}38.$$

Had we treated "Temper" as a continuous variate of Gaussian distribution, we find
$$r = {\cdot}32 \pm {\cdot}03.$$
Mr Soper's abac gives us at once $r = {\cdot}38$, and saves the labour of the second interpolation.

Illustration II. The following table gives the relation between deaths or recoveries from small-pox and the presence of a vaccination cicatrix (*Phil. Trans.* Vol. 195 A, p. 43).

Small-pox

Cicatrix	Recoveries	Deaths	Totals
Present	1562	42	1604
Absent	383	94	477
Totals.....	1945	136	2081

Here $$\log \chi^2 = 2{\cdot}2549,$$
$$\tfrac{1}{2}(1+a_1) = {\cdot}9346, \quad \tfrac{1}{2}(1+a_2) = {\cdot}7708.$$
Hence, by Table V, $$\chi_{a_1} = 1{\cdot}9437, \quad \chi_{a_2} = 1{\cdot}3869,$$
and $$_0\sigma_r = \frac{1}{\sqrt{2081}} \times \chi_{a_1} \times \chi_{a_2} = {\cdot}0590.$$

Interpolating from Table IV we have for $_0\sigma_r = {\cdot}0590$,
$$r = 0{\cdot}6, \quad \log \chi^2 = 2{\cdot}1090,$$
$$r = 0{\cdot}7, \quad \log \chi^2 = 2{\cdot}3088.$$
Hence for $$\log \chi^2 = 2{\cdot}2549, \quad r = {\cdot}67,$$
precisely the value read off beforehand from the abac.

Had we treated Cicatrix and Recovery or Death as Gaussian variates, the correlation would be
$$r = {\cdot}60 \pm {\cdot}03.$$

Illustration III. The following fourfold table is given by Macdonell (*Biometrika*, Vol. I. p. 193) as connecting stature and head breadth in animals.

Head breadth \ Stature	5′ $4\frac{9}{16}$″ and under	Over 5′ $4\frac{9}{16}$″	Totals
14·8 cm. and under ...	455	622	1077
Over 14·8 cm.	599	1324	1923
Totals.....	1054	1946	3000

Here
$$\log \chi^2 = 1{\cdot}5718,$$
$$\tfrac{1}{2}(1+a_1) = {\cdot}6487, \quad \tfrac{1}{2}(1+a_2) = {\cdot}6410.$$
Hence, by Table V,
$$\chi_{a_1} = 1{\cdot}2871, \quad \chi_{a_2} = 1{\cdot}2791,$$
and
$$_0\sigma_r = \frac{1}{\sqrt{3000}} \times \chi_{a_1} \times \chi_{a_2} = {\cdot}0301.$$

Either by interpolation from tables or the abac we reach ·165 (actually ·1654), so that $r = {\cdot}17$ is the nearest second place figure. Macdonell, assuming Gaussian frequencies, finds
$$r = {\cdot}18 \pm {\cdot}02.$$

Illustration IV. The following table is taken from my memoir on the relation of intelligence to other mental characters (*Biometrika*, Vol. V. p. 146). It gives the relationship of self-consciousness to intelligence in 2054 boys. The intelligent group cover the quick intelligent and intelligent, the other group the slow intelligent to very dull.

Consciousness \ Intellectual Grade	Intelligent	Slow Intelligent to Dull	Totals
Self-conscious	447·5	544	991·5
Unself-conscious	438·5	624	1062·5
Totals	886	1168	2054

Here
$$\log \chi^2 = 0{\cdot}4942,$$
$$\tfrac{1}{2}(1+a_1) = {\cdot}5686, \quad \tfrac{1}{2}(1+a_2) = {\cdot}5173.$$
Hence, by Table V,
$$\chi_{a_1} = 1{\cdot}2601, \quad \chi_{a_2} = 1{\cdot}2538,$$
and
$$_0\sigma_r = \frac{1}{\sqrt{2054}} \times \chi_{a_1} \times \chi_{a_2} = {\cdot}0349.$$

Interpolating from Table IV we have, for ${}_0\sigma_r = \cdot 0349$, $r = \cdot 05$, $\log \chi^2 = 0\cdot 7303$ Thus the correlation is less than ·05, and an inspection of the abac and rough extrapolation indicates that it must be about ·03. To test this, remember that for such low values of r, the Gaussian curve gives the area closely. Now $\log \chi^2 = 0\cdot 4942$ corresponds to $\chi^2 = 3\cdot 12$, and this from Table II to $-\log P = 0\cdot 428$, or $\log P = \bar{1}\cdot 572$, *i.e.* $P = \cdot 3733$, which gives a *single* tail of ·1866, or we must enter the probability integral table with $\frac{1}{2}(1+\alpha) = \cdot 8134$. This gives $x = \cdot 89 = r/{}_0\sigma_r = r/(\cdot 0349)$, or $r = \cdot 031$, agreeing excellently with the value read from the abac by extrapolation. Using Everitt's Tables we have for a Gaussian distribution

$$\cdot 009{,}631 = \cdot 156{,}650r + \cdot 000{,}621r^2 + \cdot 025{,}271r^3 + \cdot 000{,}461r^4,$$

which gives $r = \cdot 061 \pm \cdot 024$, a result within the limits of random sampling of the r obtained by the previous method, *i.e.* ·031.

Illustration V. I take for this illustration absolutely Gaussian material for a population of 1000 destined to give ·80 correlation.

	A	Not-*A*	Totals
B	704	160	864
Not-*B*	22	114	136
Totals	726	274	1000

Such a table is easily constructed from Everitt's Supplementary Tables of the Tetrachoric Functions (*Biometrika*, Vol. VIII. p. 385).

We find $$\log \chi^2 = 2\cdot 4013,$$

$$\tfrac{1}{2}(1+\alpha_1) = \cdot 726, \quad \tfrac{1}{2}(1+\alpha_2) = \cdot 864.$$

Hence, by Table V, $$\chi_{\alpha_1} = 1\cdot 3391, \quad \chi_{\alpha_2} = 1\cdot 5713,$$

and $${}_0\sigma_r = \frac{1}{\sqrt{1000}} \times \chi_{\alpha_1} \times \chi_{\alpha_2} = \cdot 06654.$$

Interpolating from Table IV, we have

$$r = \cdot 8, \quad \log \chi^2 = 2\cdot 3828, \quad r = \cdot 9, \quad \log \chi^2 = 2\cdot 5873,$$

whence we find for $\log \chi^2 = 2\cdot 4013$, $r = \cdot 809$. From the abac $r = \cdot 81$, as against the $r = \cdot 800 \pm \cdot 022$ actual value.

Illustration VI. I take another illustration of truly Gaussian material, namely 1000 cases distributed so as to give $r = \cdot 50$.

Here $$\log \chi^2 = 1\cdot 9452.$$

We have $$\tfrac{1}{2}(1+\alpha_1) = \cdot 709, \quad \tfrac{1}{2}(1+\alpha_2) = \cdot 813,$$

giving $$\chi_{\alpha_1} = 1\cdot 3248, \quad \chi_{\alpha_2} = 1\cdot 4512,$$

and $${}_0\sigma_r = \cdot 06080.$$

	A	Not-*A*	Totals
B	629	184	813
Not-*B*	80	107	187
Totals	709	291	1000

From Table IV, by interpolation for ${}_0\sigma_r = \cdot 0608$,

$$r = 0\cdot 5, \quad \log \chi^2 = 1\cdot 9366,$$
$$r = 0\cdot 6, \quad \log \chi^2 = 2\cdot 1138,$$

whence $r = \cdot 506$, when $\chi^2 = 1\cdot 9452$. The abac gives us $\cdot 51$. We have to set these values against

$$r = \cdot 500 \pm \cdot 033,$$

and we see that the difference is again not significant.

Illustration VII. As a last Gaussian material case, I take the following table, namely 2000 distributed so as to give $r = \cdot 25$.

	A	Not-*A*	Totals
B	1115	85	1200
Not-*B*	685	115	800
Totals	1800	200	2000

Here

$$\log \chi^2 = 1\cdot 4526,$$
$$\tfrac{1}{2}(1 + a_1) = \cdot 9, \quad \tfrac{1}{2}(1 + a_2) = \cdot 6.$$

These give by Table V

$${}_0\sigma_r = \frac{1}{\sqrt{2000}} \times 1\cdot 7094 \times 1\cdot 2680 = \cdot 048{,}467.$$

Whence by interpolation we find $r = \cdot 226$ and by the abac $\cdot 23$. We have then to compare the Gaussian

$$r = \cdot 25 \pm \cdot 03,$$

with $r = \cdot 23$, and we see that the difference is again less than the probable error of random sampling.

Illustrations V, VI and VII seem to show that, for truly Gaussian material, the two methods lead to closely similar results.

Illustration VIII. The following table represents in a fourfold form the correlation between the length of left foot and left middle finger in 3000 criminals. It is taken from Macdonell's paper (*Biometrika*, Vol. I. p. 226).

Left Foot

Left Middle Finger	25·5 cms. and under	Over 25·5 cms.	Totals
11·5 cms. and under...	1103	411	1514
Over 11·5 cms.	274	1212	1486
Totals	1377	1623	3000

Here
$$\log \chi^2 = 2{\cdot}9514,$$
$$\tfrac{1}{2}(1+\alpha_1) = {\cdot}5410, \quad \tfrac{1}{2}(1+\alpha_2) = {\cdot}5047.$$
Hence
$$\chi_{\alpha_1} = 1{\cdot}2557, \quad \chi_{\alpha_2} = 1{\cdot}2534,$$
and
$$_0\sigma_r = \frac{1}{\sqrt{3000}} \times \chi_{\alpha_1} \times \chi_{\alpha_2} = {\cdot}028735.$$

Interpolation and the abac give us $r = {\cdot}72$ and Macdonell has ·76. The probable error is only ·01. These last two numbers on the assumption of a Gaussian distribution.

Illustration IX. I consider lastly a table which would by many be considered to represent perfect correlation.

	A	Not-*A*	Totals
B	800	0	800
Not-*B*	0	200	200
Totals	800	200	1000

Here
$$\log \chi^2 = 3,$$
$$\tfrac{1}{2}(1+\alpha_1) = \tfrac{1}{2}(1+\alpha_2) = {\cdot}8,$$
$$\chi_{\alpha_1} = \chi_{\alpha_2} = 1{\cdot}4288\,;$$
thus
$$_0\sigma_r = {\cdot}064{,}557.$$

The point on the abac is outside the contour $r = {\cdot}95$, and some might be prepared on this account to consider the correlation as perfect. We must however proceed, as the value lies outside Table IV, by a slightly different method. By Table II, $\chi^2 = 1000$ gives us
$$-\log P = 215{\cdot}745,$$
or
$$\log P = \overline{216}{\cdot}255.$$

We now turn back to equation (xvi) and note that
$$s = \frac{1}{\sigma^2} = 239{\cdot}946.$$

Suppose
$$r = \cdot 995, \quad r^2 = \cdot 990025,$$
$$1 - r^2 = \cdot 009975 \text{ and } \lambda = \cdot 01007,$$
$$P = \sqrt{\frac{2}{\pi}} \frac{\sqrt{239 \cdot 946}}{238 \cdot 946} \frac{(\cdot 009975)^{119 \cdot 473}}{\cdot 995} \left(1 - \frac{\cdot 01}{240 \cdot 946} + \frac{\cdot 0003}{241 \times 242} + \ldots\right).$$

It is clear that for such high values we may treat the series factor as unity. We find
$$-\log P = 240 \cdot 360.$$

Now putting $r = \cdot 99$, we have
$$-\log P = 204 \cdot 525,$$
and thence by interpolation
$$-\log P = 215 \cdot 745,$$
where $r = \cdot 992$.

The correlation is thus very high, but not perfect, and this seems reasonable because we are not really making the assumption of a Gaussian frequency, and the value of P if very small is still not zero.

Conclusions. Without laying too much stress on a short series of numerical illustrations, which were merely taken at random for various values of the correlation and for various divisions of the categories*, we may, I think, conclude that our correlation scale of the improbability of independence in variates gives quite reasonable results when tested on fourfold tables treated as or really representing Gaussian distributions. In these cases we shall rarely get a divergence amounting to twice the probable error, and usually a result well within the probable error. In the following table I have put together the chief results for the present series of illustrations. In the first column we have the value of χ^2; in the second column the resulting probability of independence, P; in the third column we have ϕ^2, the mean square contingency; in the fourth column C_2, the coefficient of mean square contingency; in the fifth column $r_{hk} = \phi$; in the sixth column Yule's coefficient of association, Q_2; in the seventh column the coefficient of correlation, r_G, as found by a fourfold table on the assumption of Gaussian distribution of frequency; and lastly, in the eighth column, the value of the coefficient of correlation, r_P, on the equal improbability scale discussed in the present memoir.

Several interesting results are at once manifest:

(i) The reader will at once appreciate the difficulty of mental apprehension of the relative probabilities involved in the column P.

(ii) The order of the probabilities is not the same as that of the coefficients of correlation r_G found by assuming a Gaussian frequency. Nor should we anticipate that they would be; for by increasing the total population, and still distributing it

* I have worked out numerous other examples since the illustrations here given, and on the basis of them might have very reasonably supposed the divergences in I and II to be due to arithmetical errors. But I have not found such errors. For example take, to compare with II, the Table given by Macdonell (*Biometrika*, I. p. 222) for stature and left m. finger length (Table XII of his paper): it gives ·68 as against Macdonell's $\cdot 663 \pm \cdot 013$, as found by Gaussian methods.

among the four categories in the same proportions, we increase in the same ratio χ^2, and also decrease P, but we do not modify r_G. Our scale therefore of appreciation ought to allow for this factor in P, and this is done when we reckon r_P by considering the relation of r to ${}_0\sigma_r$, which varies with the size of the population.

Analysis of Results of Illustrations.

Illustration	χ^2	P	ϕ^2	C_2	r_{hk}	Q_2	r_G	r_P
IV (2054)	3·12	$37/10^2$	·0015	·04	·04	·08	·06 ± ·02	·03
III (3000)	37·31	$40/10^9$	·0124	·11	·11	·24	·18 ± ·02	·17
VII (2000)	28·35	$31/10^7$	·0142	·12	·12	·38	·25 ± ·03	·23
I (1294)	91·50	$10/10^{20}$	·0707	·26	·27	·39	·32 ± ·03	·38
VI (1000)	88·15	$55/10^{20}$	·0882	·28	·30	·64	·50 ± ·03	·51
II (2081)	179·85	$95/10^{40}$	·0864	·28	·29	·80	·60 ± ·03	·67
VIII (3000)	894·13	$38/10^{194}$	·2780	·47	·53	·84	·76 ± ·01	·72
V (1000)	251·94	$25/10^{54}$	·2519	·45	·50	·92	·80 ± ·02	·81
IX (1000)	1000·00	$18/10^{216}$	1·000	·71	1·00	1·00	1·00 ± ·00	·992
X (1000)	0·1423	$99/10^2$	·000142	·012	·012	1·00	*·31 ± ·26	·008

(iii) C_2 and r_{hk} are of course free from this objection, but they are absolutely incomparable with true coefficients of correlation; the former because the coefficient of contingency must be based at least on a 4×4, and better a 4×5 or 5×5, table, before it approaches r, and the latter because $r_{hk} = \phi$ is never equal to r, by its very definition and nature.

I pointed this out many years ago when first dealing with r_{hk}†. Quite recently Mr G. U. Yule has reintroduced r_{hk} under the novel name of a "theoretical value" for the correlation coefficient of a fourfold table. I am unable to see why it should be a "theoretical value," as it seems so far as I can follow Mr Yule's deduction to involve, when deduced by his method, a very arbitrary relationship between the standard deviation and the position of the mean of each subrange in the case of both frequencies. Like C_2, r_{hk} may even displace the true order of relationship in the series. I do not think that r_{hk} can be used, as Mr Yule suggests, as a measure of association, at any rate it is a measure wholly incomparable with true correlation, and it is quite possible—out of the indefinitely large number of measures of association—to select one practically as easy of determination and which does approximate to the true correlation‡.

* This value is insignificant as compared to its probable error.

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

‡ Given $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$ as our fourfold table, the correlation is not necessarily perfect in actual practice, if either b alone or c alone be zero. This is quite clear if the distribution be Gaussian, and the dividing lines of the classification be taken so as to meet on the elliptic contour of the frequency surface which contains inside itself the whole volume of the population. Thus in practice it is quite possible to obtain $Q_2 = 1$, where the correlation is small or even zero.

(iv) The coefficient of association comes out badly from these tables—it gives a difference of mean square $= \cdot 109$ against that of $r_P = \cdot 036$. But its value here is by no means as bad as it can be. Its chief evil is that it gives wildly different values according to the position of the dividing lines, and when for Gaussian material $r = 0$, Q_2 may take any value from 0 to 1 according to the position of the dividing lines, *i.e.* the percentages of the two variates in their sub-categories. When we know this is so, for material the distribution of which we can measure, what confidence can we have that the result has any significance when we know nothing at all of the frequency distribution? This may be exemplified as follows:

Illustration X.

	A	Not-*A*	Totals
B	23	971	994
Not-*B*	0	6	6
Totals	23	977	1000

Here $$\chi^2 = \cdot 14225 \text{ and } P = \cdot 986^*,$$

$$\tfrac{1}{2}(1+\alpha_1) = \cdot 977, \quad \tfrac{1}{2}(1+\alpha_2) = \cdot 994,$$

which give $$\chi_{\alpha_1} = 2\cdot 7377, \quad \chi_{\alpha_2} = 4\cdot 5419,$$

and thus $$_0\sigma_r = \frac{1}{\sqrt{1000}} \times \chi_{\alpha_1} \times \chi_{\alpha_2} = \cdot 3932,$$

and this gives r for equal probability $= \cdot 008$, *i.e.* sensibly zero. Actually the table was obtained from material having *zero* correlation. The same material divided at the mean gave

250	250	500
250	250	500
500	500	1000

,

for which the correlation is absolutely zero. In the above table, however, the association coefficient of Mr Yule is unity, in this second table it is zero!—Clearly such a coefficient when it is liable to swing over from zero to unity can be of no real service for accurate work, such as the determination of the relationships between

* Calculated from

$$P = \sqrt{\frac{2}{\pi}} \int_\chi^\infty e^{-\frac{1}{2}\chi^2} d\chi + \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\chi^2} \chi:$$

see Pearson, *Phil. Mag.* Vol. L. pp. 157—75, or *Biometrika*, Vol. I. p. 156.

deformities and in other cases to which Mr Yule, and—I regret to say—continental anthropologists and economists on his authority are now applying it*.

(v) The value of r_P seems to me based on a scientific conception. We agree to measure the closeness of association by the improbability that the material could have arisen from a random sample of unassociated variates. But this probability offers no easily apprehensible mental scale. Accordingly we determine to replace our improbabilities by correlations which would have been equally unlikely to arise from a random sample of uncorrelated material. The choice then to be made is one of a correlation scale. The probability of any r can be determined in terms of the standard deviation of r for uncorrelated material. But what standard deviation shall we select? In order that our results shall agree fairly closely with the results for Gaussian distributions we select our arbitrary standard deviation, and so our scale, to be that of a zero correlation for a fourfold Gaussian table with its variates divided in the same proportions as in the actual material. If we estimate our probability of independence on this correlation scale, we see that the values of r_P, the probability correlation, never differ very widely from those which would be obtained by supposing the fourfold table to represent a Gaussian frequency distribution. In other words, even when a table is non-Gaussian, or cannot be thought of as representing continuously varying material at all, so that r_G ceases to have any meaning as connected with regression or array variation, still its value has a perfectly definite and new significance: it measures reasonably closely the improbability that the sample could have arisen from non-associated material; it is a measure of association on a probability scale.

(vi) By aid of Table V and of Table IV, or the accompanying abac, r_P, or approximately r_G, this measure of the improbability of independence on a standard correlation scale can be found for any fourfold table in a few minutes. The extension of the fundamental idea of this paper to 3×3 tables suggests itself, and I hope shortly to publish a supplementary paper on that point.

* I am sorry to animadvert thus strongly on the work of an old pupil and colleague, but I consider that the association coefficient never had more than formal logical interest, and that to try to resuscitate it in practical statistics is to check the advance of modern scientific methods. Since this paper was printed, I have seen a memoir by Mr Yule in type, which is shortly to be issued in the *Journal of the Royal Statistical Society*. In that paper he defends, on what appear to me to be wholly inadequate grounds, the use of his Coefficient of Association and introduces what he terms a "Colligation Coefficient" —a very old friend with a new name. A reply, at length, to that memoir will appear in the forthcoming number of *Biometrika*.

TABLE I.

Values of $(-\log P)$, *entering with* r *and* ${}_0\sigma_r$.

Values of ${}_0\sigma_r$.

Values of r.	·01	·02	·03	·04	·05	·06	·07	·08
0·05	6·248	1·907	1·020	0·675	0·498	0·392	0·322	0·273
0·075	13·228	3·760	1·908	1·217	0·874	0·674	0·545	0·456
0·1	22·924	6·267	3·076	1·910	1·343	1·019	0·814	0·675
0·15	50·687	13·329	6·298	3·784	2·586	1·916	1·498	1·218
0·2	90·035	23·254	10·771	6·343	4·259	3·100	2·384	1·924
0·3	206·348	52·453	23·836	13·758	9·057	6·478	4·906	3·903
0·4	380·266	96·013	43·254	24·726	16·112	11·407	8·552	6·686
0·5	626·428	157·607	70·669	40·177	26·025	18·312	13·642	10·597
0·6	970·879	243·753	108·980	61·747	39·845	27·922	20·713	16·020
0·7	1463·946	367·033	163·781	92·579	59·584	41·634	30·792	23·740
0·8	2220·267	556·100	247·801	139·832	89·819	62·625	46·209	35·539
0·9	3607·924	902·949	401·907	226·479	145·241	101·085	74·442	57·134
0·95	5056·547	1265·013	562·757	316·904	203·069	141·207	103·886	79·671

TABLE II.

Values of $(-\log P)$ *corresponding to given values of* χ^2 *in a* 2×2 *table.*

(Extension of Palin Elderton's Table for $n' = 4$.)

χ^2	$-\log P$	χ^2	$-\log P$	χ^2	$-\log P$	χ^2	$-\log P$	χ^2	$-\log P$	χ^2	$-\log P$
1	0·096	26	5·021	50	10·097	1100	237·439	2600	562·973	13500	2929·521
2	0·242	27	5·230	60	12·231	1150	248·287	2700	584·680	14000	3038·086
3	0·407	28	5·440	70	14·370	1200	259·135	2800	606·387	14500	3146·652
4	0·583	29	5·650	80	16·513	1250	269·983	2900	628·094	15000	3255·219
5	0·765	30	5·860	90	18·659	1300	280·832	3000	649·801	15500	3363·785
6	0·952	31	6·071	100	20·809	1350	291·681	3500	758·341	16000	3472·352
7	1·143	32	6·281	150	31·579	1400	302·531	4000	866·886	16500	3580·919
8	1·337	33	6·492	200	42·375	1450	313·381	4500	975·434	17000	3689·486
9	1·533	34	6·703	250	53·184	1500	324·231	5000	1083·995	17500	3798·053
10	1·731	35	6·914	300	64·002	1550	335·081	5500	1192·538	18000	3906·621
11	1·931	36	7·126	350	74·826	1600	345·931	6000	1301·092	18500	4015·188
12	2·132	37	7·337	400	85·655	1650	356·782	6500	1409·649	19000	4123·756
13	2·334	38	7·549	450	96·487	1700	367·633	7000	1518·206	19500	4232·324
14	2·537	39	7·761	500	107·321	1750	378·484	7500	1626·765	20000	4340·892
15	2·741	40	7·972	550	118·158	1800	389·335	8000	1735·324	20500	4449·461
16	2·945	41	8·184	600	128·997	1850	400·187	8500	1843·885	21000	4558·029
17	3·151	42	8·397	650	139·837	1900	411·038	9000	1952·446	21500	4666·597
18	3·357	43	8·609	700	150·678	1950	421·890	9500	2061·008	22000	4775·166
19	3·564	44	8·821	750	161·520	2000	432·742	10000	2169·570	22500	4883·735
20	3·770	45	9·034	800	172·364	2050	443·594	10500	2278·133	23000	4992·304
21	3·978	46	9·246	850	183·208	2100	454·446	11000	2386·697	23500	5100·873
22	4·186	47	9·459	900	194·053	2200	476·151	11500	2495·261	24000	5209·442
23	4·394	48	9·672	950	204·899	2300	497·856	12000	2603·825	24500	5318·011
24	4·602	49	9·885	1000	215·745	2400	519·561	12500	2712·390	25000	5426·580
25	4·811	50	10·097	1050	226·592	2500	541·267	13000	2820·955		
26	5·021			1100	237·439	2600	562·973	13500	2929·521		

TABLE III.

Values of χ^2 corresponding to the values of $(-\log P)$ in Table I.

Values of ${}_0\sigma_r$.

Values of r.	·01	·02	·03	·04	·05	·06	·07	·08
0·05	31·84	10·88	6·36	4·51	3·52	2·91	2·48	2·19
0·075	64·66	19·95	10·89	7·38	5·58	4·51	3·78	3·28
0·1	109·82	31·93	16·64	10·90	8·03	6·35	5·26	4·51
0·15	238·45	65·13	32·08	20·07	14·24	10·93	8·82	7·39
0·2	422·29	111·35	53·16	32·29	22·35	16·75	13·25	10·97
0·3	956·68	246·62	114·05	67·14	45·11	32·93	25·45	20·64
0·4	1758·21	447·81	204·07	118·18	78·13	56·14	42·73	33·92
0·5	2892·33	731·95	330·80	189·82	124·22	88·38	66·60	52·34
0·6	4479·02	1129·10	507·65	289·58	188·28	133·02	99·55	77·70
0·7	6750·09	1697·24	760·43	431·96	279·58	196·57	146·35	113·61
0·8	10233·49	2568·34	1147·76	649·98	419·22	293·64	217·74	168·34
0·9	16624·37	4166·12	1857·93	1049·48	674·92	471·22	348·23	268·26
0·95	23295·86	5833·82	2599·00	1466·24	941·56	656·32	484·15	372·37

TABLE IV.

Values of $\log \chi^2$ corresponding to values of r and ${}_0\sigma_r$ in Tables I and II.

Values of ${}_0\sigma_r$.

Values of r.	·01	·02	·03	·04	·05	·06	·07	·08
0·05	1·5030	1·0366	0·8035	0·6542	0·5465	0·4639	0·3945	0·3404
0·075	1·8106	1·2999	1·0370	0·8681	0·7466	0·6542	0·5775	0·5159
0·1	2·0407	1·5042	1·2212	1·0374	0·9047	0·8028	0·7210	0·6522
0·15	2·3774	1·8138	1·5062	1·3025	1·1535	1·0386	0·9455	0·8686
0·2	2·6256	2·0467	1·7256	1·5091	1·3493	1·2240	1·1222	1·0400
0·3	2·9808	2·3920	2·0571	1·8270	1·6543	1·5176	1·4057	1·3148
0·4	3·2451	2·6511	2·3098	2·0725	1·8928	1·7493	1·6307	1·5305
0·5	3·4612	2·8645	2·5196	2·2783	2·0942	1·9464	1·8235	1·7188
0·6	3·6512	3·0527	2·7056	2·4618	2·2748	2·1239	1·9980	1·8904
0·7	3·8293	3·2297	2·8811	2·6354	2·4465	2·2935	2·1654	2·0554
0·8	4·0100	3·4097	3·0598	2·8129	2·6224	2·4678	2·3379	2·2262
0·9	4·2207	3·6197	3·2690	3·0210	2·8293	2·6732	2·5419	2·4286
0·95	4·3673	3·7660	3·4148	3·1662	2·9738	2·8171	2·6850	2·5710

TABLE V.*

Values of χ_a for values of $\frac{1}{2}(1+a)$.

$\frac{1}{2}(1+a)$	χ_a	$\frac{1}{2}(1+a)$	χ_a	$\frac{1}{2}(1+a)$	χ_a	$\frac{1}{2}(1+a)$	χ_a
·50	1·2533	·65	1·2877	·80	1·4288	·95	2·1132
·51	1·2535	·66	1·2928	·81	1·4457	·96	2·2740
·52	1·2539	·67	1·2984	·82	1·4641	·97	2·5071
·53	1·2546	·68	1·3044	·83	1·4844	·98	2·8915
·54	1·2556	·69	1·3109	·84	1·5067	·985	3·2097
·55	1·2569	·70	1·3180	·85	1·5315	·990	3·7333
·56	1·2585	·71	1·3256	·86	1·5590	·991	3·8854
·57	1·2604	·72	1·3338	·87	1·5897	·992	4·0639
·58	1·2626	·73	1·3427	·88	1·6245	·993	4·2784
·59	1·2652	·74	1·3523	·89	1·6640	·994	4·5419
·60	1·2680	·75	1·3626	·90	1·7094	·995	4·8779
·61	1·2712	·76	1·3738	·91	1·7623	·996	5·3278
·62	1·2748	·77	1·3859	·92	1·8249	·997	5·9776
·63	1·2787	·78	1·3990	·93	1·9003	·998	7·0465
·64	1·2830	·79	1·4133	·94	1·9937	·999	9·3870

* Reprinted from *Biometrika*, Vol. IX.

TABLE VI.

Extension of Sheppard's Table of the Probability Integral

$$F=\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{1}{2}x^2}\,dx, \text{ giving } (-\log F) \text{ for } x.$$

x	$-\log F$	x	$-\log F$	x	$-\log F$
5	6·54265	30	197·30921	50	544·96634
6	9·00586	31	210·56940	60	783·90743
7	11·89285	32	224·26344	70	1066·26576
8	15·20614	33	238·39135	80	1392·04459
9	18·94746	34	252·95315	90	1761·24604
10	23·11805	35	267·94888	100	2173·87154
11	27·71882	36	283·37855	150	4888·38812
12	32·75044	37	299·24218	200	8688·58977
13	38·21345	38	315·53979	250	13574·49960
14	44·10827	39	332·27139	300	19546·12790
15	50·43522	40	349·43701	350	26603·48018
16	57·19458	41	367·03664	400	34746·55970
17	64·38658	42	385·07032	450	43975·36860
18	72·01140	43	403·53804	500	54289·40830
19	80·06919	44	422·43983		
20	88·56010	45	441·77568		N.B. To obtain anything but a rough appreciation after $x=50$, the table would require much extension, but for many practical problems it suffices to take after $x=50$: $F=\frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-\frac{1}{2}x^2}$.
21	97·48422	46	461·54561		
22	106·84167	47	481·74964		
23	116·63253	48	502·38776		
24	126·85686	49	523·45999		
25	137·51475	50	544·96634		
26	148·60624				
27	160·13139				
28	172·09024				
29	184·48283				
30	197·30921				

To each of the values in this table ·30103 must be added, if we wish to obtain the probability that the value is greater than x, without regard to sign.

DESCRIPTION OF PLATES.

This memoir is accompanied by two *abacs* for which I have heartily to thank my assistant Mr G. H. Soper.

The first abac Plate I gives ${}_0\sigma_r$. It has not been used in the text because interpolation from the Tables gave slightly more accurate values, but its readings are quite sufficient for most practical purposes. The maximum difference we found in determining ${}_0\sigma_r$ for the ten illustrations of this paper was ·0006, or a unit in the value of ${}_0\sigma_r$ read to two significant figures.

The method of using it is as follows: Run along the horizontal giving the size of the population (left-hand scale), until you meet the vertical giving the value $\frac{1}{2}(1+\alpha_1)$ (bottom scale); then follow the 45° line through that point till you reach the left-hand scale, take the horizontal through this point and follow it, till you meet the vertical through $\frac{1}{2}(1+\alpha_2)$ (bottom scale), then again follow the 45° line to the left-hand side, and from the point reached traverse the horizontal to the right-hand side of the diagram where the scale gives the proper value of ${}_0\sigma_r$. If, in traversing the 45° line we meet the *top of the diagram* instead of the left-hand side, we follow the usual rule of such abacs, *i.e.* drop by a vertical through the point to the bottom scale and run up the 45° line through that point to the left-hand scale and continue as before; the final value read for ${}_0\sigma_r$ on the right-hand scale has for each such drop to be multiplied by 10.

The second abac Plate II is entered by the value of $\log \chi^2$ on the left-hand scale and the value of ${}_0\sigma_r$ on the bottom scale, the meet of the horizontal and vertical lines through these points determine a contour the value of which in "probability correlation" r_P is recorded on the right-hand vertical scale.

CAMBRIDGE: PRINTED BY JOHN CLAY, M.A. AT THE UNIVERSITY PRESS

NOTE

In the course of the present memoir it is shewn that

$$\chi = \frac{r_{hk}}{{}_0\sigma_{hk}} = \frac{Q}{{}_0\sigma_Q} = \frac{\phi}{{}_0\sigma_\phi}.$$

It may then be asked why not measure the improbability of ϕ exceeding ${}_0\sigma_\phi$ by the ordinary theory of the probability integral? We know that ϕ^2 is by its essence positive and we should probably have to take ϕ positive also. The distribution of ϕ^2 for samples from a population in which ϕ^2 is zero, has not been studied; we know the mean value, $\bar{\phi}^2$, for such a population, but we do not know the frequency of ϕ in terms of ${}_0\sigma_\phi$ and we have no reason to suppose that it can be expressed by aid of a Gaussian distribution in terms of the constant ${}_0\sigma_\phi$. Similar remarks apply to $Q/{}_0\sigma_Q$ and $r_{hk}/{}_0\sigma_{hk}$; the latter will clearly be a limited range frequency, and a normal curve distribution especially for large values of r_{hk} will certainly be inadmissible. As a matter of fact the probability that a sample occurs with a χ over a given value is

$$P = \sqrt{\frac{2}{\pi}} \int_\chi^\infty e^{-\frac{1}{2}x^2}\, d\chi + \chi \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}x^2},$$

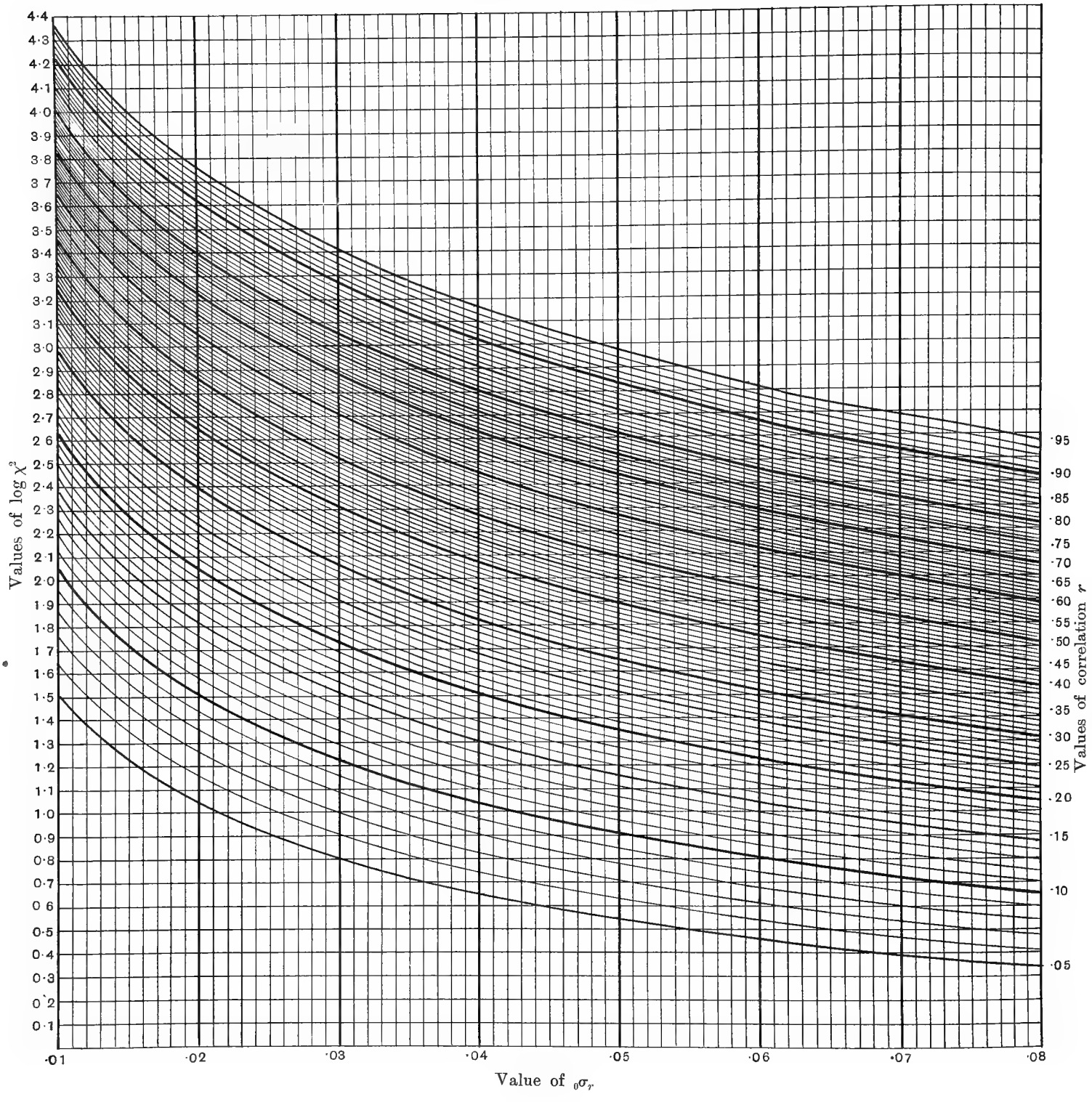
which connotes a frequency distribution

$$f = \sqrt{\frac{2}{\pi}}\, \chi^2 e^{-\frac{1}{2}x^2},$$

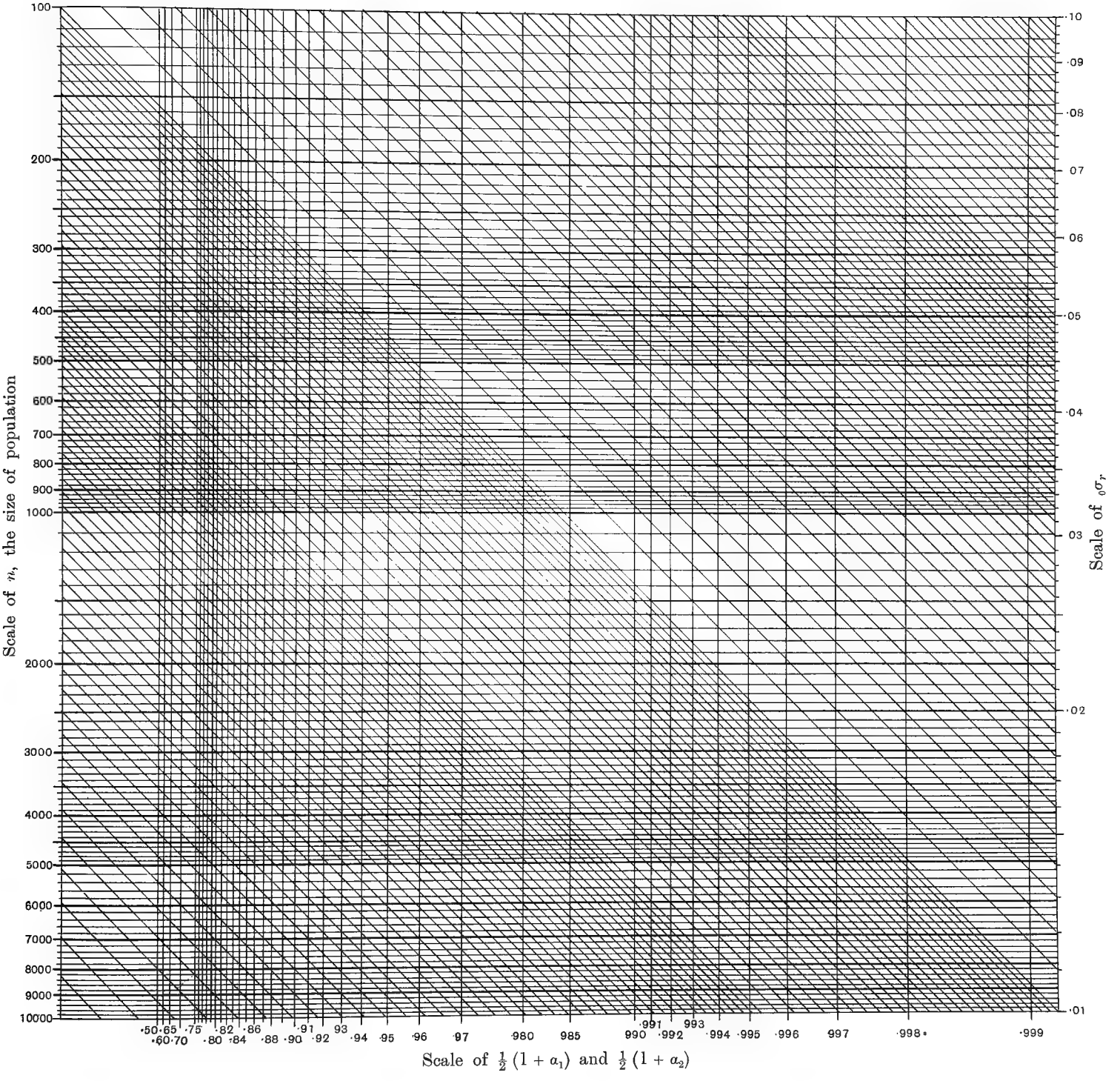
and this is not a normal distribution*. The above relations between χ, r_{hk}, Q, ϕ and their standard deviations are interesting, but they do not provide us, by aid of a Gaussian probability table, with the requisite "equality in probability," which we are seeking in this memoir. There is very grave danger when, having found in some case the value of the standard deviation of a statistical quantity, we then assume that for this case the Gaussian distribution must apply, and deduce thereby a measure of the significance of the quantity, *e.g.* estimate the significance of Q from a knowledge of ${}_0\sigma_Q$.

* It also is only a close approximation; we have actually assumed the Gaussian may be used to describe the frequency binomials, but it is a far closer approximation than using a Gaussian to describe the frequency of χ, or of $Q/{}_0\sigma_Q$.

Abac to determine r_p



Abac to determine ${}_0\sigma_r$



EUGENICS LABORATORY PUBLICATIONS.

Published by the Cambridge University Press, Fetter Lane, E.C. 4

Memoir Series.

I. **The Inheritance of Ability.** Being a statistical Examination of the Oxford Class Lists from the year 1800 onwards, and of the School Lists of Harrow and Charterhouse. By EDGAR SCHUSTER, M.A., Formerly Galton Research Fellow in National Eugenics, and E. M. ELDERTON, Galton Research Scholar in National Eugenics. *Issued.* Price 4*s. net.*

II. **A First Study of the Statistics of Insanity and the Inheritance of the Insane** Diathesis. By DAVID HERON, M.A., Galton Research Fellow. *Issued.* Price 3*s. net.*

III. **The Promise of Youth and the Performance of Manhood.** Being a statistical Examination into the Relation existing between Success in the Examinations for the B.A. Degree at Oxford and subsequent Success in professional Life. (The professions considered are the Bar and the Church.) By EDGAR SCHUSTER, M.A., D.Sc., Formerly Galton Research Fellow in National Eugenics. *Issued.* Price 2*s.* 6*d. net.*

IV. **On the Measure of the Resemblance of First Cousins.** By ETHEL M. ELDERTON, Galton Research Scholar, assisted by KARL PEARSON, F.R.S. *Issued.* Price 3*s.* 6*d. net.*

V. **A First Study of the Inheritance of Vision and of the Relative Influence** of Heredity and Environment on Sight. By AMY BARRINGTON and KARL PEARSON, F.R.S. *Issued.* Price 4*s. net.*

VI. **Treasury of Human Inheritance** (Pedigrees of physical, psychical, and pathological Characters in Man). Parts I. and II (double part). (Diabetes insipidus, Split-Foot, Polydactylism, Brachydactylism, Tuberculosis, Deaf-Mutism, and Legal Ability.) Issued by the Galton Laboratory. Price 14*s. net.*

VII. **The Influence of Parental Occupation and Home Conditions on the** Physique of the Offspring. By ETHEL M. ELDERTON, Galton Research Scholar. *Shortly.*

VIII. **The Influence of Unfavourable Home Environment and Defective Physique** on the Intelligence of School Children. By DAVID HERON, M.A., Galton Research Fellow. *Issued.* Price 4*s. net.*

IX. **The Treasury of Human Inheritance** (Pedigrees of physical, psychical, and pathological Characters in Man). Part III. (Angioneurotic Oedema, Hermaphroditism, Deaf-mutism, Insanity, Commercial Ability.) *Issued.* Price 6*s. net.*

X. **A First Study of the Influence of Parental Alcoholism on the Physique** and Intelligence of the Offspring. By ETHEL M. ELDERTON, Galton Research Scholar, assisted by KARL PEARSON, F.R.S. *Issued. Second Edition.* Price 4*s. net.*

XI. **The Treasury of Human Inheritance** (Pedigrees of physical, psychical, and pathological Characters in Man). Part IV. (Cleft Palate, Hare-Lip, Deaf-mutism, and Congenital Cataract.) *Issued.* Price 10*s. net.*

XII. **The Treasury of Human Inheritance** (Pedigrees of physical, psychical, and pathological Characters in Man). Parts V and VI. (Haemophilia.) *Issued.* Price 15*s. net.*

XIII. **A Second Study of the Influence of Parental Alcoholism on the Physique** and Intelligence of the Offspring. A Reply to certain Medical Critics and an Examination of the rebutting Evidence cited by them. By KARL PEARSON, F.R.S., and ETHEL M. ELDERTON. *Issued.* Price 4*s. net.*

XIV. **A Preliminary Study of Extreme Alcoholism in Adults.** By AMY BARRINGTON and KARL PEARSON, F.R.S., assisted by DAVID HERON, D.Sc. *Issued.* Price 4*s. net.*

XV. **The Treasury of Human Inheritance** (Pedigrees of physical, psychical, and pathological Characters in Man). Parts VII and VIII. (Dwarfism.) With 49 Plates of Illustrations and 8 Plates of Pedigrees. *Issued.* Price 15*s. net.*

XVI. **The Treasury of Human Inheritance.** Prefatory Matter and complete Name and Subject Indices to Vol. I. With Frontispiece Portraits of Sir Francis Galton and Ancestry. *Issued.* Price 3*s. net.*

XVII. **A Second Study of Extreme Alcoholism in Adults.** By DAVID HERON, D.Sc. *Immediately.*

Buckram covers for binding Volume I. of the *Treasury of Human Inheritance* with impress of the bust of Sir FRANCIS GALTON by Sir GEORGE FRAMPTON can be obtained from the Eugenics Laboratory by sending a postal order for 2*s.* 9*d.* to the Hon. Secretary.

A large photograph (11″ × 13″) of Sir FRANCIS GALTON by the late Mr DEW SMITH can also be obtained from the Laboratory by sending a postal order for 10*s.* 6*d.* to the Hon. Secretary.