

**Mathematical contributions to the theory of evolution. / By Karl Pearson  
[and others].**

**Contributors**

Pearson, Karl, 1857-1936.

Blakeman, John.

Drapers' Company (London, England)

London School of Hygiene and Tropical Medicine

**Publication/Creation**

London : Published by Dulau and Co., 1904-1907.

**Persistent URL**

<https://wellcomecollection.org/works/x7duh5hh>

**Provider**

London School of Hygiene and Tropical Medicine

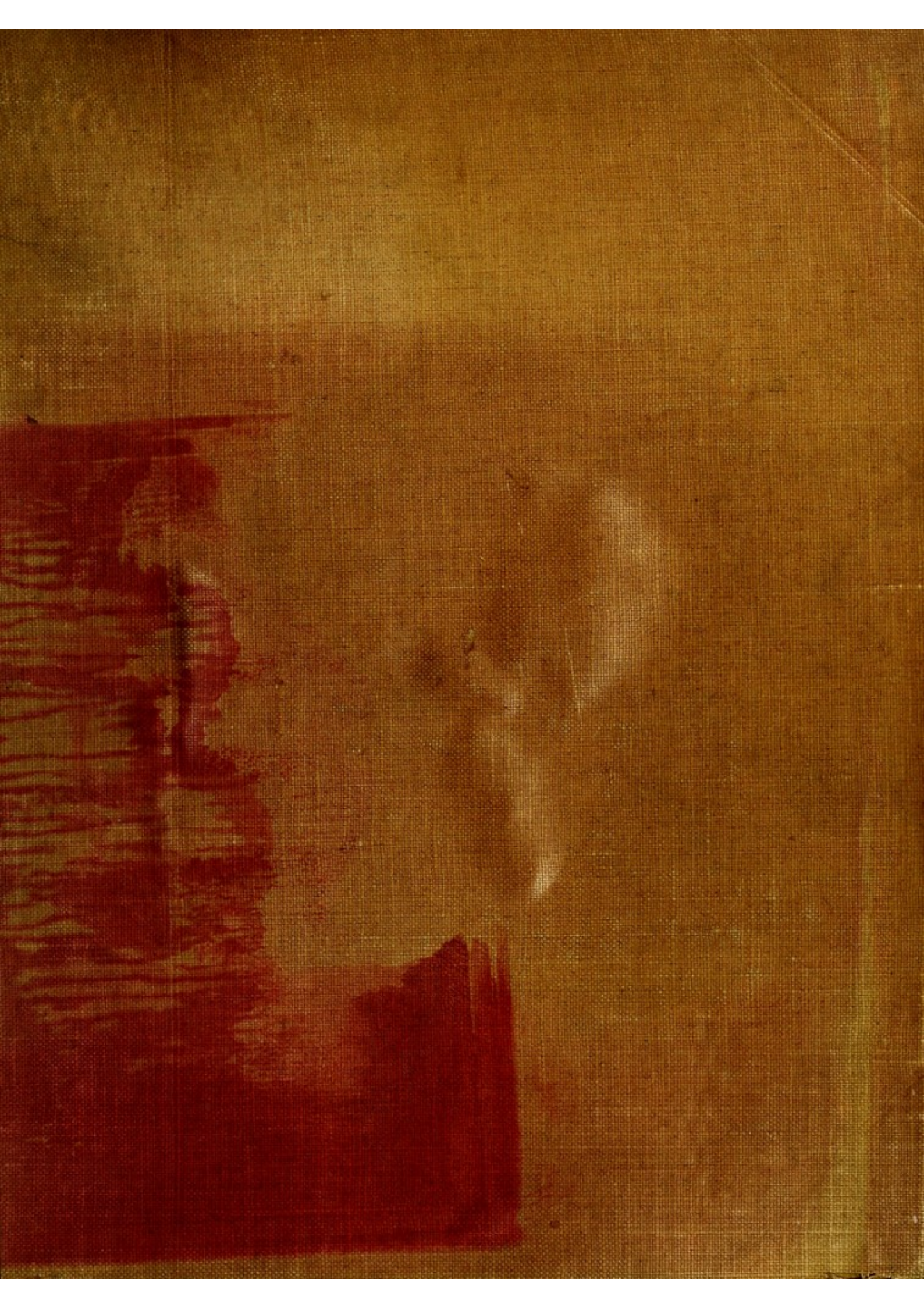
**License and attribution**

This material has been provided by This material has been provided by London School of Hygiene & Tropical Medicine Library & Archives Service. The original may be consulted at London School of Hygiene & Tropical Medicine Library & Archives Service. where the originals may be consulted. Conditions of use: it is possible this item is protected by copyright and/or related rights. You are free to use this item in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s).

**wellcome  
collection**

Wellcome Collection  
183 Euston Road  
London NW1 2BE UK  
T +44 (0)20 7611 8722  
E [library@wellcomecollection.org](mailto:library@wellcomecollection.org)  
<https://wellcomecollection.org>







12  
E

MEDICAL RESEARCH  
COUNCIL LIBRARY  
No.....




LIBRARY

Author : PEARSON (K.)			
Title : Mathematical contributions to the theory of evolution: XIII; XIV; XV: XVI.			
Acc. No.	Class Mark	Date	Volume
62244	* BEN	1904-07.	

Sold to School of Hygiene.





Digitized by the Internet Archive  
in 2015

<https://archive.org/details/b24397933>







DEPARTMENT OF APPLIED MATHEMATICS,  
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON.

---

DRAPERS' COMPANY RESEARCH  
MEMOIRS.

BIOMETRIC SERIES, I.

---

9437

95

MATHEMATICAL CONTRIBUTIONS TO THE  
THEORY OF EVOLUTION.

XIII. ON THE THEORY OF CONTINGENCY AND ITS RELATION  
TO ASSOCIATION AND NORMAL CORRELATION.

BY

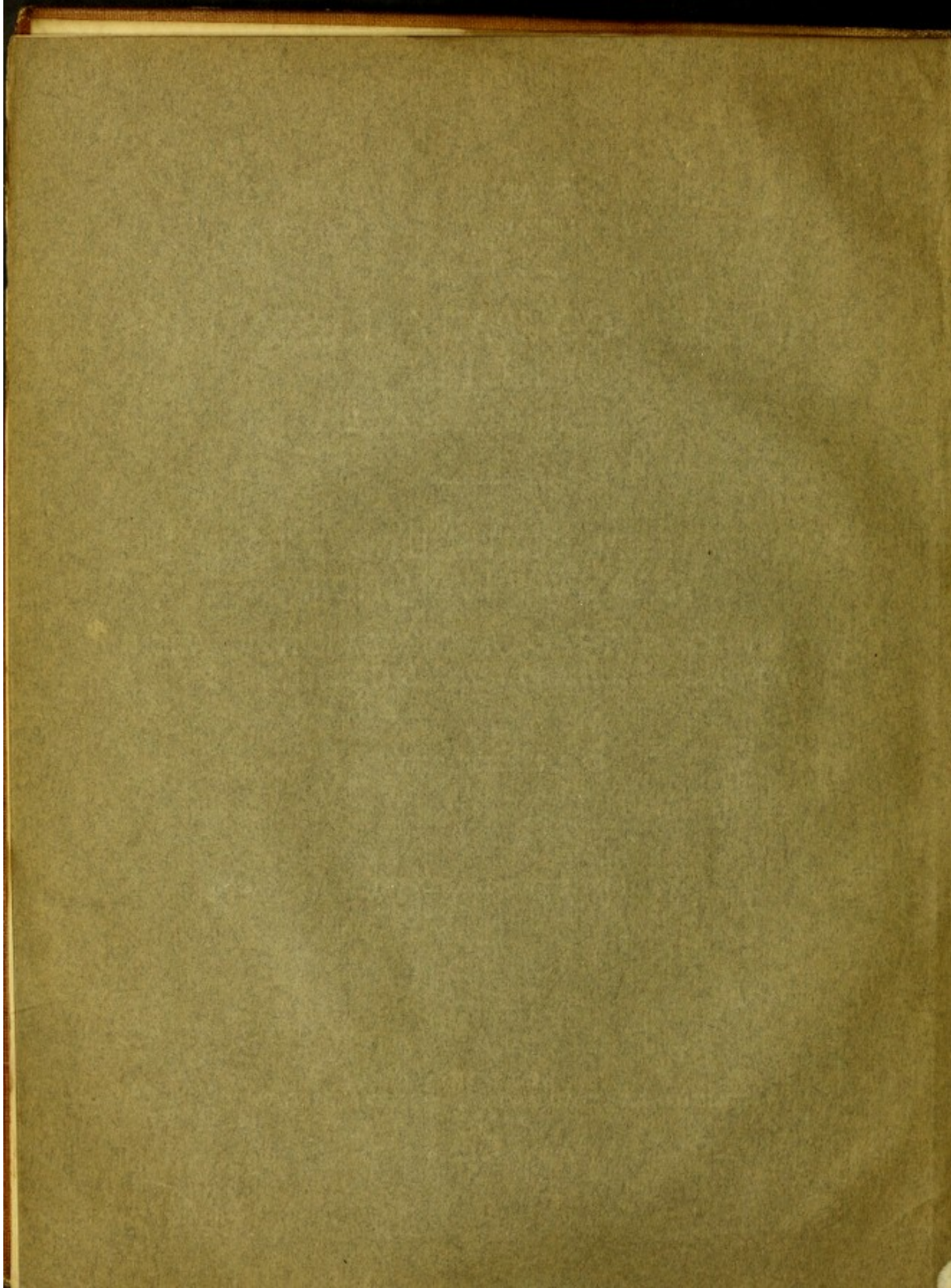
KARL PEARSON, F.R.S.

[WITH TWO DIAGRAMS.]

LONDON:  
PUBLISHED BY DULAU AND CO., 37 SOHO SQUARE, W.  
1904.

*Price Four Shillings.*







DEPARTMENT OF APPLIED MATHEMATICS,  
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON.

---

DRAPERS' COMPANY RESEARCH  
MEMOIRS.

BIOMETRIC SERIES, I.

---

MATHEMATICAL CONTRIBUTIONS TO THE  
THEORY OF EVOLUTION.

XIII. ON THE THEORY OF CONTINGENCY AND ITS RELATION  
TO ASSOCIATION AND NORMAL CORRELATION.

BY  
KARL PEARSON, F.R.S.

[WITH TWO DIAGRAMS.]

LONDON:  
PUBLISHED BY DULAU AND CO., 37 SOHO SQUARE, W.  
1904.

*Price Four Shillings.*

62244

*In March, 1903, the Worshipful Company of Drapers announced their intention of granting £1,000 to the University of London to be devoted to the furtherance of research and higher work at University College. After consultation between the University and College authorities, the Drapers' Company presented £1,000 to the University to assist the statistical work and higher teaching of the Department of Applied Mathematics. It seemed desirable to commemorate this—probably, first occasion on which a great City Company has directly endowed higher research work in mathematical science—by the issue of a special series of memoirs in the preparation of which the Department has been largely assisted by the grant. Such is the aim of the present series of "Drapers' Company Research Memoirs."*

K. P.



*On Contingency and its Relation to Association and Normal Correlation.*

By KARL PEARSON, F.R.S.

CONTENTS.

	Page
(1.) Introductory. On classification and scales . . . . .	1
(2.) On the conception of contingency . . . . .	5
(3.) On the relation between mean square contingency and normal correlation . . . . .	7
(4.) On the relation between mean contingency and normal correlation . . . . .	9
(5.) The two contingency coefficients . . . . .	15
(6.) Influence of overfine grouping . . . . .	16
(7.) Probable errors and error correlations of sub-contingencies . . . . .	17
(8.) Change of order of grouping unimportant in the general case of linear regression . . . . .	19
(9.) Relation of contingency to association . . . . .	21
(10.) Theory of multiple contingency . . . . .	22
(11.) Illustrations :— . . . . .	27
A.—Inheritance of stature in man . . . . .	27
B.—Colour inheritance in greyhounds . . . . .	30
C.—Hair colours in man . . . . .	31
D.—Occupational or professional contingency in relatives . . . . .	32
(12.) General conclusions . . . . .	34

(1.) *Introduction.*

IN dealing with the problem of the relationship of attributes, not capable of quantitative measurement, it has been usual to classify the two attributes into a number of groups,  $A_1, A_2, A_3, \dots, A_s$  and  $B_1, B_2, B_3, \dots, B_t$ . In this manner a table has been formed containing  $s$  columns and  $t$  rows, or  $s \times t$  compartments. The total frequency of the population, or of the "universe" under consideration, to use the logician's phrase, is then distributed into sub-groups corresponding to these  $s \times t$  compartments. In simple cases of association, as in that of the presence of the vaccination cicatrix and the recovery from an attack of smallpox,  $s$  and  $t$  are both equal to two, and we have a simple four-fold division of the universe. In other cases we have higher numbers, as when we classify the human eye into eight colour classes and correlate these classes with six or more classes for hair colour. We may even run up to as many as 18 to 25 classes for each attribute when we table the coat colours of thoroughbred horses or pedigree dogs in the case of pairs of blood relatives.



Hitherto, in order to obtain a measure of the degree of correlation or association, we have proceeded on the assumption that it was necessary to arrange the system of classes like  $A_1, A_2, \dots, A_n$  in some order, which corresponded to a real quantitative scale in the attribute, although we were unable to use this scale directly. Thus one arranged eye-colours in what appeared to correspond to a scale of varying amounts of orange pigment; the coat colours of horses were arranged in an order corresponding fairly to what an artist would call their "value." I even analysed hair tints by photographic processes. In all such cases the order seemed of vital importance. Once this order was settled, the methods of my memoir\* on the correlation of characters not quantitatively measurable could be applied—the actual scale corresponding to the classification could be deduced, and we were able, on the assumption of normal frequency, to actually plot the regression lines for the correlation of a variety of attributes.† The conception, however, of order in the classification was at times very hampering. Take three broad classes like those for human temper—*quick tempered, good natured, and sullen*; it is difficult to grasp the exact meaning of a quantitative scale at the basis of this classification, and it is not obvious that the right order is necessarily that with good-natured in the middle. Or, again, take the case of human hair; omitting the brown reds, we can get a practically continuous series of shades from jet black to flaxen, and from flaxen with increasing red up to the deepest reds. Only the brown reds come in and upset the system! We seem, therefore, forced to take a double scale, first one of black, and then one of red pigment. Or, again, take the coat colour of greyhounds; these are classified into as many as 40 fairly narrow groups, and we can arrange these groups in ascending order of red, or black, or other pigmentation. We have more than one possible scale.

Now in recent work on such things as temper in man, eye colour in man, and hair colour in man or other animals, I have proceeded to arrange my groups in two or three different orders, and to calculate the correlation on the basis of these different orders. The results for the different orders came out in rather striking agreement, and the first sort of conclusion that one was tempted to draw was, for example, that the inheritance of pigmentation was strikingly alike for all pigments. But the agreement was in some cases far closer than one is accustomed to find when one compares the inheritance of directly measurable characters, and I soon became convinced that owing to some important theoretical law hitherto overlooked, the order of the groups by which we classify our attributes is a matter of no importance when we are determining correlation. The group order is all important for variation, it has practically no influence on correlation. We may put sullen tempers where we please in regard to quick and good-natured; we may place the shades of red hair at either end of the hair scale or in the middle, and the inheritance coefficient will come

\* 'Phil. Trans.,' A, vol. 195, pp. 1-47.

† For example, for health and ability and for the correlation of the psychical and physical characters, see the "Fourth Annual Huxley Lecture," 'Journal of the Anthropological Institute,' vol. 33, pp. 194-195.



out nearly the same in value. Nay, we may go further, and classify finger prints like Mr. GALTON into "tents," "arches," "whorls," "croziers," &c., &c., and still be able to find a numerical value of the degree of resemblance between two blood relatives, although any arrangements of such groups into a possible quantitative scale may be inconceivable. The object of this present paper is to deal with this novel conception of what I have termed *contingency*, and to see its relation to our older notions of association and normal correlation. The great value of the idea of contingency for economic, social, and biometric statistics seems to me to lie in the fact that it frees us from the need of determining scales before classifying our attributes. I shall endeavour to illustrate the importance of this freedom in the illustrations which follow the theoretical treatment of the subject.

(2.) *On the Conception of Contingency.*

In mathematical treatises on algebra a definition is usually given of independent probability. If  $p$  be the probability of any event, and  $q$  the probability of a second event, then the two events are said to be independent, if the probability of the combined event be  $p \times q$ . Now let  $A$  be any attribute or character and let it be classified into the groups  $A_1, A_2, \dots, A_s$ , and let the total number of individuals examined be  $N$ , and let the numbers which fall into these groups be  $n_1, n_2, \dots, n_s$  respectively. Then the probability of an individual falling into one or other of these groups is given by  $n_1/N, n_2/N, \dots, n_s/N$  respectively. Now suppose the same population to be classified by any other attribute into the groups  $B_1, B_2, \dots, B_t$ , and the group frequencies of the  $N$  individuals to be  $m_1, m_2, \dots, m_t$  respectively. The probability of an individual falling into these groups will be respectively  $m_1/N, m_2/N, m_3/N, \dots, m_t/N$ . Accordingly the number of combinations of  $B_e$  with  $A_u$  to be expected on the theory of independent probability if  $N$  pairs of attributes are examined is

$$N \times \frac{n_u}{N} \times \frac{m_e}{N} = \frac{n_u \cdot m_e}{N} = \nu_{ue}, \text{ say.}$$

Let the number actually observed be  $n_{ue}$ . Then, allowing for the errors of random sampling,

$$n_{ue} - \frac{n_u m_e}{N} = n_{ue} - \nu_{ue}$$

is the deviation from independent probability in the occurrence of the groups  $A_u, B_e$ . Clearly the total deviation of the whole classification system from independent probability must be some function of the  $n_{ue} - \nu_{ue}$  quantities for the whole table. I term any measure of the total deviation of the classification from independent probability a measure of its *contingency*. Clearly the greater the contingency, the greater must be the amount of association or of correlation between the two



attributes, for such association or correlation is solely a measure from another standpoint of the degree of deviation from independence of occurrence.

Now it must be quite clear that if we make our measurement of contingency any function whatever of such quantities as  $n_{uv} - \nu_{uv}$ , its magnitude will be absolutely independent of the order of classification, *i.e.*, its value will be unchanged if we re-arrange the A's and the B's in any manner whatever. This is the fundamental gain of this new conception of contingency. But precisely as we can measure position or acceleration in a great variety of ways, so it is possible to measure contingency. We must try to select out of these ways those which: (a) bring contingency into line with the customary notions of correlation and association; and (b) permit of not too laborious calculations leading to the required measure.

We will consider these points at some length. I have shown in a paper,\* "On Deviations from the Probable in a Correlated System of Variables," that if  $m'_1, m'_2, \dots, m'_n$  be any system of observed frequencies and  $m_1, m_2, \dots, m_n$  be any system of theoretical frequencies known *à priori*, then if

$$\chi^2 = \text{Sum} \left\{ \frac{(m'_q - m_q)^2}{m_q} \right\} \text{ from } q = 0 \text{ to } n$$

be calculated, we can deduce a quantity P from  $\chi^2$  which is the probability that in any trial a system  $m''_1, m''_2, \dots, m''_n$  of observed frequencies will occur, which deviates more from  $m_1, m_2, \dots, m_n$  than the actually observed system does. Tables have been worked out by Mr. PALIN ELDERTON giving the value of P, for a considerable range of values of  $\chi^2$  and  $n$ , and have been published in 'Biometrika.'†

Now it will be obvious that if we want to measure contingency, we really want to measure the deviation of the observed results from independent probability, and therefore if we take  $m_1, m_2, \dots, m_n$  to correspond to the system  $\nu_{uv}$  and  $m'_1, m'_2, \dots, m'_n$  to correspond to the actually observed system  $n_{uv}$ ,

$$\chi^2 = S \left\{ \frac{(n_{uv} - \nu_{uv})^2}{\nu_{uv}} \right\} \dots \dots \dots (i.),$$

will be a proper quantity to calculate, and P would measure how far the observed system is or is not compatible with a basis of independent probability. If P be large the chances are in favour of the system arising from independent probability; if P be small there is certainly association between the attributes. Hence  $1 - P$  would be a proper measure of the contingency. I propose to call  $1 - P$  the *contingency grade*. Further, it is convenient to have a name for a function closely related to  $\chi^2$ . I shall call

$$\phi^2 = \chi^2/N \dots \dots \dots (ii.)$$

the *mean square contingency*.

\* 'Phil. Mag.,' July, 1900, pp. 157-175.

† Vol. I, p. 155.



It will be seen that, in the method by which we have approached the problem, we have not had to consider the question of the sign of the contingency like  $n_{uv} - v_{uv}$ , our mean square contingency is based on a summation of squares extending to all the  $s \times t$  compartments of the table. But if we treat now of quantities like  $n_{uv} - v_{uv}$  their total sum must be zero, since for the whole table

$$S(n_{uv}) = N = S(v_{uv}).$$

Let us suppose that the symbol  $\Sigma$  refers to a summation of all the *positive* contingencies, and let

$$\psi = \Sigma(n_{uv} - v_{uv})/N \dots \dots \dots (iii),$$

then  $\psi$  shall be spoken of as the *mean contingency*. Clearly any functions of either  $\phi^2$  or  $\psi$  would serve to measure the contingency. We shall be guided in our choice of such functions by considering what are the values of  $\phi^2$  and  $\psi$  in the case of normal correlation.

(3.) *On the Relation between Mean Square Contingency and Normal Correlation.*

Let  $x$  and  $y$  denote the deviations from their respective means of two characters or attributes, of which  $\sigma_x, \sigma_y$  are the standard deviations and  $r$  is the correlation. Then if we assume a normal distribution of frequency,  $z_0 \delta x \delta y$  would be the frequency of individual pairs falling between  $x$  and  $x + \delta x, y$  and  $y + \delta y$ , where

$$z_0 = \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)} \dots \dots \dots (iv),$$

on the assumption of independent probability, and  $z \delta x \delta y$ , where

$$z = \frac{N}{2\pi\sqrt{1-r^2}\sigma_x\sigma_y} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)} \dots \dots \dots (v).$$

on the assumption of contingent probability.

We then have at once

$$\phi^2 = S \left\{ \frac{(z \delta x \delta y - z_0 \delta x \delta y)^2}{N z_0 \delta x \delta y} \right\} = S \left\{ \frac{(z - z_0)^2}{N z_0} \delta x \delta y \right\},$$

and we have only to insert the values of  $z$  and  $z_0$ , given by (iv.) and (v.), and integrate all over the plane of  $x, y$ , to find the mean square contingency.

Now, if  $ac > b^2$ , we know that

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(ax^2 - 2bxy + cy^2)} dx dy = \frac{1}{\sqrt{ac - b^2}} \dots \dots \dots (vi).$$



This is all we need, for if  $x = \sigma_x x'$ ,  $y = \sigma_y y'$  :

$$\begin{aligned} \phi^2 &= \frac{1}{N} \int_{-x}^{+x} \int_{-x}^{+x} \left( \frac{z^2}{z_0} - 2z + z_0 \right) dx' dy' \\ &= \frac{1}{2\pi} \left\{ \frac{1}{1-r^2} \int_{-x}^{+x} \int_{-x}^{+x} e^{-\frac{1}{2} \left\{ x'^2 \frac{1+r^2}{1-r^2} - \frac{2rx'y'}{1-r^2} + y'^2 \frac{1+r^2}{1-r^2} \right\}} dx' dy' \right. \\ &\quad - \frac{2}{\sqrt{1-r^2}} \int_{-x}^{+x} \int_{-x}^{+x} e^{-\frac{1}{2} \left\{ x'^2 \frac{1}{1-r^2} - \frac{2rx'y'}{1-r^2} + y'^2 \frac{1}{1-r^2} \right\}} dx' dy' \\ &\quad \left. + \int_{-x}^{+x} \int_{-x}^{+x} e^{-\frac{1}{2}(x'^2+y'^2)} dx' dy' \right\} \\ &= \frac{1}{1-r^2} \frac{1}{\sqrt{\left(\frac{1+r^2}{1-r^2}\right)^2 - \frac{4r^2}{(1-r^2)^2}}} - \frac{2}{\sqrt{1-r^2}} \frac{1}{\sqrt{\frac{1}{(1-r^2)^2} - \frac{r^2}{(1-r^2)^2}}} + 1 \quad \text{(vii.)} \\ &= \frac{1}{1-r^2} - 2 + 1 = \frac{r^2}{1-r^2}. \end{aligned}$$

Thus the mean square contingency is simply  $r^2/(1-r^2)$ . Or,

$$r = \pm \sqrt{\frac{\phi^2}{1 + \phi^2}} \quad \text{(viii.)}$$

Thus the relationship between mean square contingency and correlation in the case of normal frequency is of an extremely simple character.

We see at once :—

- (i.) That since the mean square contingency is absolutely independent of the arrangement of our classes, the coefficient of correlation is also entirely independent of the arrangement of our classes on the basis of any assumed order or scale.
- (ii.) Provided our classes are sufficiently small to allow of us legitimately replacing by groupings over small areas the theoretical integrations, the coefficient of correlation can be found from the mean square contingency.

We have thus an entirely new method of finding correlation in the case of quantitatively non-measurable characters. It assumes, however, that our classification-groups are sufficiently numerous and their contents sufficiently small to justify us in supposing that the contingency has reached a definite limit. Clearly in working in the future by the contingency method, we shall have to adopt rather more numerous classes, and they should not contain too irregular proportions of individuals, but we can then afford to drop any question of scale or order of grouping.

It may be asked whether this method of deriving the correlation from the contingency cannot replace the earlier method of deducing the correlation by the fourfold division of the material. The answer is that in some cases it can do so very



advantageously, but it is very far from doing so in all. The contingency found from a fourfold table is a perfectly real and very proper measure of the deviation of its material from independent probability. But if this mean square contingency be substituted in equation (viii.), it will not give us the correlation. The proper mean square contingency to give us the correlation must be based on a sufficiently *large* number of classes. When, however, we take, say, 20 classes for each attribute, we have 400 terms to deal with in calculating  $\phi^2$ , and although the result might then possibly give a more accurate value for the correlation than that found from a fourfold division, yet the labour of determining it is far greater and may be excessive. Further, the simple classification into two or three groups may be all we are able to make at all, or all we can conveniently make. Hence the new conception of contingency, while illuminating the whole subject—especially as demonstrating that the correlation is independent of scale or grouping, does not do away with the older method of the fourfold division. I propose to call the expression

$$\sqrt{\frac{\phi^2}{1 + \phi^2}},$$

the *first coefficient of contingency*.

We note that with small enough classes the coefficient of contingency becomes the coefficient of correlation. Accordingly, with a view of lessening the number of coefficients in use, I adopt the following convention: Any expression or function of either the mean square contingency ( $\phi^2$ ) or the mean contingency ( $\psi$ ) (or indeed of any other measure of the contingency), which, when the grouping is sufficiently small, is theoretically equal to the coefficient of correlation—on the hypothesis of normal frequency—shall be termed a coefficient of contingency. All such coefficients of contingency must, on the same hypothesis, become equal on a sufficiently small grouping, and they will scarcely differ widely from each other when the frequency is not absolutely normal and the grouping is merely moderately small. These points will be illustrated later.

#### (4.) *On the Relation of Mean Contingency to Normal Correlation.*

A great deal of the labour of finding either the coefficient of contingency or the coefficient of correlation by the method of mean square contingency when the groups are numerous, depends upon the squaring of the contingencies and dividing by the frequency to be expected on the basis of independent probabilities. The whole of this labour is escaped, if we work with the mean contingency instead of the mean square contingency; further, since in this case we only sum for the positive contingencies, neglecting the negative, we have usually to deal with only, or often less than, a moiety of the terms involved in calculating  $\phi^2$ . On the other hand, there is no simple relation between the correlation and the mean contingency such as we have found between correlation and mean square contingency in equation (viii.) above.



The relation is far more complex and is only expressible in the form of integrals reducible by quadratures. Still, for practical purposes we rarely want the coefficient of contingency to more than two decimal places. Hence, if the integral be evaluated for the coefficient proceeding by equal intervals, we can plot a curve giving the value of the coefficient of contingency in terms of the mean contingency, and this will be sufficiently accurate to enable us to read off the former in terms of the latter to the required degree of accuracy. The enquiry also brings out some other points not without interest.\*

*To investigate the curve which in a normal correlation surface separates on the plane of  $xy$  areas of positive from areas of negative contingency.*

The frequency due to independent probability will be equal to that due to the actual contingent probability when

$$\frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)} = \frac{N}{2\pi\sigma_x\sigma_y} \frac{1}{\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)},$$

where  $r$  is the coefficient of correlation, or of contingency.

Clearly

$$(1-r^2) \log_e (1-r^2) = -r^2 \left\{ \frac{x^2}{\sigma_x^2} - \frac{2xy}{r\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2} \right\} \dots \dots \dots \text{(ix.)}$$

Since  $r$  is always less than unity, this curve is clearly a hyperbola, which possesses several interesting properties. We see at once that all the contingency of one sense is grouped into the space between the two branches of this hyperbola, and that the contingency of the other sense is grouped into the two separate spaces inside the two branches. Thus contingency of either sense is for normal correlation *continuous*, and abrupt changes of sign in the contingency—beyond the limits of random sampling—are not to be expected.

By testing on actual correlation tables I find this hyperbola comes out in a fairly marked manner, in fact, quite as significantly as the elliptic contours of equal frequency.

I propose to consider the properties of this zero contingency hyperbola—it forms the curve along which two really contingent events have a frequency identical with their independent probability.

Consider the two families of curves :

$$\frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2} = \alpha \dots \dots \dots \text{(x.)}$$

$$\frac{x^2}{\sigma_x^2} - \frac{2}{r} \frac{xy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2} = \beta \dots \dots \dots \text{(xi.)}$$

\* I have to heartily thank my assistant, Dr. L. N. G. FILON, for the substance of the first part of the investigation given below, down to equation (xiii). I owe the calculation and plotting of the curves  $u = e^{-x \sec \theta}$  to my assistant, Mr. J. C. M. GARNETT.



Since  $r$  is always  $< 1$ , the  $\alpha$  family form a set of concentric, similar, and similarly-situated ellipses, and the  $\beta$  family a set of concentric, similar, and similarly-situated hyperbolas. Any conic having double contact with the hyperbola  $\beta_0$ , of zero contingency defined by (ix.), at the ends of a diameter  $y = mx$ , has for its equation

$$\frac{x^2}{\sigma_x^2} - \frac{2}{r} \frac{xy}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} + \lambda (y - mx)^2 = \beta_0.$$

If this be identical with an ellipse  $\alpha$ , we have, by comparing coefficients and eliminating  $\lambda$  and  $m$ ,

$$\beta_0^2/\alpha^2 = 1/r^2.$$

Consequently  $\alpha = \pm r\beta_0$ , the sign being determined from the fact that  $\alpha$  must always be positive for real ellipses.

Now the ordinate  $z$  of the normal frequency surface is given by

$$\begin{aligned} z &= \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left\{\frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right\}}, \\ &= \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{\alpha}{2(1-r^2)}}, \end{aligned}$$

and to find the mean contingency we must determine the whole volume lying inside the *two* branches of the above hyperbola, integrating on *both* sides of the line of contact of the families of hyperbolas and ellipses.\*

We have  $\iint \frac{z}{N} dx dy$  over this area

$$= I_r = \frac{4}{2\pi} \frac{1}{\sigma_x\sigma_y} \frac{1}{\sqrt{1-r^2}} \int_{r\beta_0}^{\infty} d\alpha \int_{\beta_0}^{\alpha/r} \frac{e^{-\frac{\alpha}{2(1-r^2)}}}{J} d\beta,$$

where

$$J = \frac{\delta(\alpha, \beta)}{\delta(x, y)} = -\frac{4(1-r^2)}{r\sigma_x\sigma_y} \left( \frac{x^2}{\sigma_x^2} - \frac{y^2}{\sigma_y^2} \right)$$

from (x.) and (xi.).

But from (x.) and (xi.)

$$\left\{ \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)^2 - \frac{4x^2y^2}{\sigma_x^2\sigma_y^2} \right\} (1-r^2)^2 = (\alpha^2 - \beta^2r^2) (1-r^2).$$

Or, choosing the signs to make  $J$  positive, we have

$$J = \frac{4\sqrt{1-r^2}}{r\sigma_x\sigma_y} \sqrt{\alpha^2 - \beta^2r^2}.$$

\* The ellipses and hyperbolas have common pairs of conjugate diameters; one line of contact is one of the asymptotes of the hyperbola  $\frac{x^2}{\sigma_x^2} - \frac{y^2}{\sigma_y^2} = 1$ ; and tangents at an intersection point of any of the family of ellipses with any of the family of hyperbolas are respectively parallel to conjugate diameters of this hyperbola. These geometrical properties, however, need not detain us here.



Thus the required integral is

$$I_r = \frac{2r}{4\pi(1-r^2)} \int_{r\beta_0}^{\infty} d\alpha \int_{\beta_0}^{\alpha/r} d\beta \frac{e^{-\frac{\alpha}{2(1-r^2)}}}{\sqrt{\alpha^2 - \beta^2 r^2}},$$

$$= \frac{1}{2\pi(1-r^2)} \int_{r\beta_0}^{\infty} \cos^{-1} \frac{\beta_0 r}{\alpha} e^{-\frac{\alpha}{2(1-r^2)}} d\alpha.$$

To simplify put, using (ix.),

$$\alpha = \beta_0 r \sec \theta, \quad k = \frac{\beta_0 r}{2(1-r^2)} = -\frac{1}{2r} \log_e(1-r^2) \dots \dots \dots \text{(xii.)}$$

where  $k$  will always be positive, since  $r < 1$ .

We have

$$I_r = \frac{k}{\pi} \int_0^{\frac{\pi}{2}} e^{-k \sec \theta} \theta \sec \theta \tan \theta d\theta,$$

or, integrating by parts,

$$I_r = \frac{1}{\pi} \int_0^{\frac{\pi}{2}} e^{-k \sec \theta} d\theta \dots \dots \dots \text{(xiii.)}$$

The curves  $u = e^{-k \sec \theta}$  were then plotted with our coordinatograph for a series of values of  $k$  or  $r$  on a large scale, drawn in with a spline and integrated with a Coradi compensating planimeter. The values of  $I_r$  resulting are tabled on p. 15.

We have next to investigate what is the volume  $NQ_r$  of the surface of independent probability

$$z_0 = \frac{N}{2\pi\sigma_x\sigma_y} e^{-1(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2})},$$

which falls within the same hyperbola of contingency. We shall then have in  $Q_r - I_r$  the required value of  $\psi$ , the mean contingency on the basis of normal correlation. We have

$$Q_r = \frac{1}{2\pi\sigma_x\sigma_y} \iint e^{-1(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2})} dx dy$$

taken over the space inside the two branches of the hyperbola

$$\frac{x^2}{\sigma_x^2} - \frac{2xy}{r\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2} = \beta_0$$

Write  $x = x'\sigma_x, y = y'\sigma_y$ , and we have

$$x'^2 - 2\frac{x'y'}{r} + y'^2 = \beta_0$$

Transform to polars,  $\rho \cos \theta = x', \rho \sin \theta = y'$ ,

$$\rho^2 = \frac{r\beta_0}{r - \sin 2\theta}$$



This shows us that the axes are given by  $\theta = \frac{\pi}{4}$  and  $\frac{\pi}{2} + \frac{\pi}{4}$ , or are  $a$  and  $b$ , where

$$a^2 = r\beta_0/(1 + r), \quad b^2 = r\beta_0/(1 - r).$$

Take these axes as axes of coordinates. Then we have to integrate

$$Q_r = \frac{1}{\pi} \iint e^{-1(x^2+y^2)} dx dy,$$

over the area inside *one* branch of the hyperbola

$$(1 + r)x^2 - (1 - r)y^2 = r\beta_0 \dots \dots \dots \text{(xiv.)}$$

Let

$$\left. \begin{aligned} x^2 + y^2 &= \alpha, \\ x^2 - y^2 + r(x^2 + y^2) &= r\beta \end{aligned} \right\} \dots \dots \dots \text{(xv.)}$$

and let us transfer the integrations to  $\alpha$  and  $\beta$ .

We have

$$\begin{aligned} x^2 &= \frac{1}{2} \{ \alpha - r(\alpha - \beta) \}, \\ y^2 &= \frac{1}{2} \{ \alpha + r(\alpha - \beta) \}, \end{aligned}$$

and

$$Q_r = \frac{2}{\pi} \iint \frac{e^{-1\alpha}}{J} d\alpha d\beta,$$

over one-half one branch of the hyperbola.

$$J = \frac{d\alpha d\beta}{dy dx} - \frac{d\alpha d\beta}{dx dy} = \frac{8yx}{r} = \frac{4}{r} \sqrt{\alpha^2 - r^2(\alpha - \beta)^2}.$$

Thus we have

$$Q_r = \frac{r}{2\pi} \int_{\beta_0}^{\alpha} dae^{-1\alpha} \int_{\beta_0}^{\frac{1+r}{r}\alpha} \frac{d\beta}{\sqrt{\alpha^2 - r^2(\alpha - \beta)^2}} \dots \dots \dots \text{(xvi.)}$$

The limits are obtained from the consideration, easily seen on a figure, that for a given  $\alpha$  we must integrate from  $\beta = \beta_0$ , the given hyperbola, to  $\beta = \frac{1+r}{r}\alpha$ , the *touching* hyperbola; and then for  $\alpha$  we must take every circle from that touching  $\beta_0$ , *i.e.*,  $\alpha = r\beta_0/(1 + r)$  up to infinity.

We will first integrate with regard to  $\beta$ , and put

$$r(\alpha - \beta) = -\alpha \sin \phi.$$

This gives, when  $\beta = (1 + r)\alpha/r$ ,  $\phi = \frac{1}{2}\pi$ ; and when

$$\beta = \beta_0, \quad \phi = \sin^{-1} \frac{r(\beta_0 - \alpha)}{\alpha} = \phi_r.$$



Thus we find

$$Q_r = \frac{1}{2\pi} \int_{\frac{r\beta_0}{1+r}}^{\infty} da e^{-1/a} \int_{\phi_0}^{1/\pi} d\phi = \frac{1}{2\pi} \int_{\frac{r\beta_0}{1+r}}^{\infty} \cos^{-1} \frac{r(\beta_0 - a)}{a} e^{-1/a} da \quad \dots \quad (\text{xvii}).$$

Take

$$\cos \chi = r(\beta_0 - a)/a,$$

then

$$a = \infty, \quad \cos \chi = -r,$$

$$a = r\beta_0/(1+r), \quad \cos \chi = 1.$$

Hence

$$\begin{aligned} Q_r &= \frac{1}{2\pi} \int_0^{\cos^{-1}(-r)} \chi e^{-1/\frac{r\beta_0}{r+\cos\chi}} \frac{r\beta_0 \sin \chi}{(r+\cos\chi)^2} d\chi, \\ &= \frac{1}{\pi} \int_0^{\cos^{-1}(-r)} e^{-1/\frac{r\beta_0}{r+\cos\chi}} d\chi, \end{aligned}$$

observing that the term between the limits vanishes at both.

Take

$$\cos \theta = (r + \cos \chi)/(r + 1).$$

Then

$$\chi = 0, \quad \theta = 0,$$

$$\chi = \cos^{-1}(-r), \quad \theta = \frac{1}{2}\pi.$$

Thus we find finally, after some reductions,

$$Q_r = \frac{1}{\pi} \int_0^{1/\pi} e^{-\kappa \sec \theta} \sqrt{\frac{1 + \cos \theta}{\epsilon + \cos \theta}} d\theta \quad \dots \quad (\text{xviii}),$$

where

$$\left. \begin{aligned} \epsilon &= (1-r)/(1+r), \\ \kappa &= \frac{1}{2} \frac{r\beta_0}{1+r} = -\frac{1}{2} \frac{1-r}{r} \log_e (1-r^2) \end{aligned} \right\} \dots \quad (\text{xix}),$$

$$= (1-r)k, \text{ of the integral } I_r.$$

Tables were now formed of  $\epsilon$  and  $\kappa$  and the ordinates of the curves

$$v = e^{-\kappa \sec \theta} \sqrt{\frac{1 + \cos \theta}{\epsilon + \cos \theta}} \quad \dots \quad (\text{xx.})$$

calculated.\* These ordinates were plotted on a large scale by aid of a Coradi coordinatograph and the resulting curves integrated as before, the values of  $Q_r$  thus found are given with the values of  $I_r$  and  $\psi$  in the table below. I believe this table gives the mean contingency in terms of the correlation true to at least three places of decimals. The  $u$  and  $v$  curves are both interesting analytically and subject to rather curious changes of type. We were aided in plotting them by calculating, where

\* I owe the calculation of these ordinates to Dr. ALICE LEE.



needful,  $\frac{du}{d\theta}$  and  $\frac{dv}{d\theta}$ . Finally, the values of  $r$  were plotted by my demonstrator, Mr. L. W. ATCHERLEY, to the corresponding values of  $\psi$ . Thus a curve was obtained, which enables us to read off the correlation from the contingency correct to at least two places of decimals—sufficient for nearly all practical purposes.

TABLE I.—Table of Integrals  $I_r$ ,  $Q_r$ , and the Contingency  $\psi$  for Values of  $r$ .

$r$ .	$I_r$ .	$Q_r$ .	$\psi$ .
0.00	.5000	.5000	.0000
.05	.4620	.4762	.0142
.10	.4342	.4652	.0310
.20	.3895	.4536	.0641
.30	.3501	.4498	.0996
.40	.3162	.4547	.1385
.50	.2830	.4643	.1813
.60	.2489	.4814	.2325
.70	.2128	.5106	.2978
.80	.1700	.5524	.3824
.90	.1186	.6279	.5093
.95	.0796	.7009	.6213
1.00	.0000	1.0000	1.0000

Diagram I. at the end of this memoir will therefore serve for most purposes of interpolation, and it will be seen that now that the integrals have been evaluated and the diagram constructed, the correlation can be very easily found from mean contingency. But the method seems to me distinctly inferior to that of mean square contingency, and this for much the same reasons that mean error calculations are inferior to mean square error work in curve fitting. Further, the grade of contingency can be found at once from a knowledge of mean square contingency, and whatever be the distribution is a significant and interpretable constant. This is only true of the correlation deduced from mean contingency if the distribution be normal.

(5.) To sum up our results so far :—

We have, if

$n_{uv}$  be the actual frequency of a group in the population,  $N$  which combines the characters  $A_u$  and  $B_v$ ,  $\nu_{uv}$  be the frequency of this group on the hypothesis of independent probability, then

$n_{uv} - \nu_{uv}$  is simply a sub-contingency,

$S \left\{ \frac{(n_{uv} - \nu_{uv})^2}{\nu_{uv}} \right\} = \chi^2$  may be termed the square contingency,

$S \left\{ \frac{(n_{uv} - \nu_{uv})^2}{N\nu_{uv}} \right\} = \phi^2$  is the mean square contingency,

$\Sigma \left( \frac{n_{uv} - \nu_{uv}}{N} \right) = \psi$ , where  $\Sigma$  is the sum for positive (or negative) sub-contingencies only, is the mean contingency.



Any one of these expressions is a measure of the deviation of the system from independent probability, and therefore of the amount of association or correlation between the characters or attributes involved. But any function of these expressions is also a proper measure. Such functions are:—

(a.) The contingency grade. This is  $1 - P$ , where  $P$  is to be found from  $\chi^2$  by aid of the tables for “goodness of fit.” See ‘*Biometrika*,’ vol. 1, pp. 155, *et seq.*

(b.) The mean square contingency coefficient =  $C_1$ , where

$$C_1 = \sqrt{\frac{\phi^2}{1 + \phi^2}} \dots \dots \dots (xxi).$$

(c.) The mean contingency coefficient =  $C_2$ , where  $C_2$  is to be found from the table on p. 15 or from Diagram I. at the end of this memoir.

In the case of sufficiently small grouping and normal correlation we have

$$C_1 = C_2 = \text{coefficient of correlation.}$$

But it must not be forgotten that this is essentially a limiting, not a general case. Nevertheless the approach to equality of the two contingency coefficients will be a good measure of the normality of the distribution and the suitability as to smallness of our elements of grouping.

(6.) A little experience of actual working, however, shows that in practice it is perfectly easy to overshoot the mark in fineness of grouping. Suppose that in dealing with 1000 cattle we find a single instance of a calf inscribed as “mulberry,” say the offspring of a red cow by a dark fawn bull. Now if there be 30 dark fawn bulls, the independent probability of a dark fawn bull having a mulberry offspring is .03. Hence the sub-contingency for a ♂ parent-offspring table =  $1 - .03 = .97$ , and the corresponding contribution to the square contingency will be  $(.97)^2 / .03$ , or is upwards of 31. The fact is, that when we come to very fine groupings we get at once into difficulties owing to our having to record by *units* only. Suppose “mulberry” calves actually had no relation to any special parentage, but were rare anomalies occurring once among 1000 calves, or perhaps were merely an odd breeder’s fancy description, then a unit cannot be divided in the proportions of the colour parentage, it must fall into some one colour parentage group. The result is that a few isolated individuals will give large contributions to the mean square contingency. The above example is purely hypothetical, but similar cases have actually occurred in dealing with colour problems by the contingency method. They are exactly similar to those which occur when dealing with outlying individuals by the test for “goodness of fit.” In a frequency distribution we proceed only by units, but the theory gives fractional values of the frequency; hence in forming the value of  $\chi^2$  to measure goodness of fit, one or two unit “outliers,” although not improbable as far as the *whole* of the tail of a curve is concerned, may be exceedingly improbable if



considered from the standpoint of the actual group in which they do occur. This point must be carefully borne in mind in actual practice, for by sufficient refinement of grouping, *i.e.*, till we reduce certain groups to a single individual or two, the mean square contingency can be increased in a remarkable manner.

(7.) Of course this is merely saying that the probable errors of the sub-contingencies increase largely when we make  $n_{uv}$  very small. Unfortunately I have not yet succeeded in determining the probable errors of the contingency coefficients. If  $c_{uv}$  be the contingency, determined by

$$c_{uv} = n_{uv} - \frac{n_u n_v}{N},$$

and  $\Sigma_{c_{uv}}$  its standard deviation for random sampling, I find

$$\Sigma_{c_{uv}} = n_{uv} \left( 1 - \frac{n_{uv}}{N} \right) + \frac{n_u n_v}{N^2} \left( n_u + n_v - \frac{4n_u n_v}{N} \right) - 2 \frac{n_{uv}}{N} \left( n_u + n_v - \frac{3n_u n_v}{N} \right). \quad \text{(xxii.)}$$

so that the probable error of any individual contingency =  $\cdot 67449 \Sigma_{c_{uv}}$  is determined.

Further, if  $R_{c_{uv}, c_{u'v'}}$  be the correlation between errors due to random sampling in two contingencies  $c_{uv}$  and  $c_{u'v'}$ , *not* belonging to either the same row or column,

$$\begin{aligned} \Sigma_{c_{uv}} \Sigma_{c_{u'v'}} R_{c_{uv}, c_{u'v'}} = & - \frac{n_{uv} n_{u'v'}}{N} + 2 \frac{n_{uv} n_u n_{u'} + n_{u'v'} n_u n_v}{N^2} \\ & + \frac{n_{uv} n_v n_{u'} + n_{u'v'} n_u n_v}{N^2} - 4 \frac{n_u n_v n_u n_{u'}}{N^3} \dots \dots \dots \quad \text{(xxiii.)} \end{aligned}$$

Similarly we find for the correlation of errors of two contingencies of the same column,  $R_{c_{uv}, c_{uv'}}$ , the result

$$\begin{aligned} \Sigma_{c_{uv}} \Sigma_{c_{uv'}} R_{c_{uv}, c_{uv'}} = & - \frac{n_{uv} n_{uv'}}{N} - \frac{n_u n_{uv'}}{N} + \frac{n_v n_{uv}}{N} \left( 1 - \frac{3n_u}{N} \right) \\ & + \frac{n_u n_v n_{uv'}}{N^2} \left( 1 - \frac{4n_u}{N} \right) \dots \dots \dots \quad \text{(xxiv.)} \end{aligned}$$

and for errors of two contingencies of the same row,

$$\begin{aligned} \Sigma_{c_{uv}} \Sigma_{c_{u'v}} R_{c_{uv}, c_{u'v}} = & - \frac{n_{uv} n_{u'v}}{N} - \frac{n_u n_{u'v}}{N} + \frac{n_v n_{uv}}{N} \left( 1 - \frac{3n_v}{N} \right) \\ & + \frac{n_u n_v n_{u'v}}{N^2} \left( 1 - \frac{4n_v}{N} \right) \dots \dots \dots \quad \text{(xxv.)} \end{aligned}$$

Results (xxii.) to (xxv.) enable us to find the probable errors and the error correlations for any individual contingencies which will arise from random sampling, and are so far of value; but when we attempt to find the general expression for the probable error of either the mean or mean square contingency, it becomes so complex



that there appears little hope of deducing a simple result. Arithmetically the problem might be solved at the expense of rather troublesome numerical calculations if the number of sub-groups was not very large. A general and simple expression for the probable error of  $\psi$  or  $\phi^2$  involving  $\psi$  or  $\phi^2$  only does not appear likely to exist, and an expression involving all the sub-group frequencies would be very troublesome for computation. Practically the errors of the contingency coefficients may be fairly reasonably taken to lie between the probable errors of  $r$  as found by a fourfold division of a table and by the product method, approaching the latter more closely as the number of sub-groups is sufficiently increased. With the experience of probable errors of fourfold tables before us we may, I think, safely take the probable error of a contingency coefficient  $C$  for rough judgments to be less than

$$2 \times .67449 \frac{1 - C^2}{\sqrt{n}},$$

*i.e.*, double the probable error of a correlation coefficient found from the product moment. At the same time we must distinctly be cautious, remembering the difficulty as to isolated units referred to in the previous section.

We may look at the probable error of the contingency from another standpoint.

Taking the mean squared contingency, we have

$$1 + \phi^2 = \frac{1}{1 - r^2}.$$

Therefore

$$\delta\phi^2 = \frac{2r}{(1 - r^2)^2} \delta r,$$

and accordingly, if  $\Sigma_{\phi^2}$ ,  $\Sigma_r$  be the standard deviations in errors of  $\phi^2$  and  $r$ ,

$$\begin{aligned} \Sigma_{\phi^2} &= \frac{2r}{1 - r^2} \Sigma_r = \frac{2r}{(1 - r^2)^2} \frac{1 - r^2}{\sqrt{N}} * \\ &= \frac{2}{\sqrt{N}} \frac{r}{1 - r^2} = \frac{2}{\sqrt{N}} \sqrt{\phi^2 (1 + \phi^2)}. \end{aligned}$$

Hence if we were to determine  $\phi^2$  from  $r$ , the probable error of  $\phi^2$  would be given by

$$\text{Probable error of } \phi^2 = .67449 \frac{2}{\sqrt{N}} \sqrt{(1 + \phi^2) \phi^2}.$$

Or, we can put it into the more useful form,

$$\text{Percentage probable error of } \phi^2 = \frac{1.34898}{\sqrt{N}} \sqrt{\frac{1 + \phi^2}{\phi^2}} \dots \text{ (xxvi).}$$

Thus the percentage probable error increases rapidly as the contingency gets smaller.

\* 'Phil. Trans.,' A, vol. 191, p. 242.



Of course, the probable error of  $\phi^2$  as found from  $r$  is not necessarily the same as the probable error of  $\phi^2$  found directly, but it may serve as a guide to its approximate value.

If it were the same, the probable error of  $r$  as found from  $\phi^2$  would be  $\cdot67449/\{(1 + \phi^2)\sqrt{N}\}$ , a result, as indicated in the previous paragraph, much too small, except possibly for very successful systems of grouping.

(8.) To find under what other condition than normal correlation small changes in the order of grouping will not affect the value of the correlation.

Let us assume the unit of grouping to be very small, but not necessarily the same for all groups. Let the two characters or attributes be  $x$  and  $y$ , and suppose  $n_s$  to be the total frequency of individuals in the range  $y_s - \epsilon$  to  $y_s + \epsilon$ , and  $n_{s+1}$  to be the total frequency in the range  $y_{s+1} - \epsilon'$  to  $y_{s+1} + \epsilon'$ . Let  $y_{s+1} - y_s = \epsilon + \epsilon' = h$  be so small that its square may be neglected. Let  $\bar{x}$ ,  $\bar{y}$  be the mean values of the characters,  $N$  the total frequency. We will find the changes in the moments and constants supposing the array  $n_s$  and  $n_{s+1}$  interchanged in position.

Clearly  $\delta\bar{x} = 0$  and  $\delta\sigma_x = 0$ .

$$N(\bar{y} + \delta\bar{y}) = S(y_s n_s) + h(n_s - n_{s+1}),$$

or,

$$\delta\bar{y} = h(n_s - n_{s+1})/N.$$

$$N(\sigma_y + \delta\sigma_y)^2 = S(y_s^2 n_s) + 2h(y_s n_s - y_{s+1} n_{s+1}) - N(\bar{y} + \delta\bar{y})^2,*$$

$$2\sigma_y \delta\sigma_y = 2h(y_s n_s - y_{s+1} n_{s+1}) - 2N\bar{y}\delta\bar{y},$$

$$\frac{\delta\sigma_y}{\sigma_y} = \frac{h}{\sigma_y^2} \frac{(y_s - \bar{y})n_s - (y_{s+1} - \bar{y})n_{s+1}}{N}.$$

Next if

$$P = S(xy) - N\bar{y}\bar{x},$$

$$P + \delta P = S(xy) + h(n_s \bar{x}_s - n_{s+1} \bar{x}_{s+1}) - N\bar{y}\bar{x} - N\bar{x}\delta\bar{y},$$

or,

$$\delta P = h\{n_s(\bar{x}_s - \bar{x}) - n_{s+1}(\bar{x}_{s+1} - \bar{x})\},$$

where  $\bar{x}_s$  and  $\bar{x}_{s+1}$  are the means of the arrays  $n_s$  and  $n_{s+1}$ .

But if  $r$  be the correlation coefficient of  $x$  and  $y$  characters,

$$r = \frac{P}{N\sigma_x\sigma_y}.$$

Therefore

$$\frac{\delta r}{r} = \frac{\delta P}{P} - \frac{\delta\sigma_x}{\sigma_x} - \frac{\delta\sigma_y}{\sigma_y},$$

\* It must be noted here that the squares of the change in  $\bar{y}$  and  $\sigma_y$  are neglected. Hence the changes must not be so great that  $\delta\bar{y}$  and  $\delta\sigma_y$  are sensibly as compared with  $\bar{y}$  and  $\sigma_y$ .



and substituting the above values,

$$\frac{\delta r}{r} = h \left\{ \frac{n_s (\bar{x}_s - \bar{x}) - n_{s+1} (\bar{x}_{s+1} - \bar{x})}{P} - \frac{(y_s - \bar{y}) n_s - (y_{s+1} - \bar{y}) n_{s+1}}{N \sigma_y^2} \right\}.$$

If this is to vanish for any value of  $s$  and  $h$ , it will be sufficient, since

$$P = r \times N \sigma_x \sigma_y,$$

$$\bar{x}_s - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y_s - \bar{y}),$$

and

$$\bar{x}_{s+1} - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y_{s+1} - \bar{y}).$$

Or, if the mean  $\bar{x}_m$  of any  $y_m$ -array of individuals be determined by

$$\bar{x}_m - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y_m - \bar{y}).$$

But this is the condition for linear regression.

Hence we conclude that in any correlated system of variables, obeying the law of linear regression, we can, without sensibly modifying the correlation, interchange two adjacent  $y$ -arrays (*e.g.*, two rows of the correlation table), provided the grouping be fine. But if we can interchange any two adjacent  $y$ -arrays, we can, by a repetition of such changes, interchange any two  $y$ -arrays whatever; and a precisely similar statement must be valid for any two  $x$ -arrays (*e.g.*, two columns of the correlation table). Hence, given a sufficiently small system of grouping, we may state that in all cases of linear regression the actual order of the scales is immaterial as far as the determination of the correlation is concerned.

The practical importance of this result would appear to be great, for it frees us when dealing with scale orders from the need for supposing normal frequency; the indifference of the scale order when determining correlation is still true, provided the regression is linear; and this linearity of regression is not only found from observation to be very general—for example, in inheritance problems\*—but follows from theory itself in the case of various hypotheses.†

In actual practice, of course, the degree of fineness of the grouping is limited by many considerations, and hence it will often be better to proceed by the fourfold division method, taking that division where possible at a very distinct classification. But the general principle now demonstrated will enable us in future to pay much less

\* See "The Laws of Inheritance in Man.—I. Inheritance of the Physical Characters," 'Biometrika,' vol. 2, pp. 362-3; also "Inheritance of Mental and Moral Characters in Man," 'Huxley Memorial Lecture,' 1903. 'Journal of the Anthropological Institute,' vol. 33, pp. 185-7.

† "Contributions to the Theory of Evolution.—XII. On a Generalised Theory of Alternative Inheritance, with special reference to MENDEL'S LAWS." 'Phil. Trans.,' A, vol. 203, p. 85.



attention to the actual order chosen for the scales if we are dealing with a class of characters for which we may reasonably presume the regression to be sensibly linear.

(9.) If we take the crudest possible division of our material into only four groups, thus :—

<i>a</i>	<i>c</i>	<i>a + c</i>
<i>d</i>	<i>b</i>	<i>d + b</i>
<i>a + d</i>	<i>c + b</i>	N

corresponding to what Mr. YULE has termed the *association* of two attributes, we have at once

$$\psi = \frac{2(ab - cd)}{N^2} \dots \dots \dots (xxvii).$$

$$\phi^2 = \frac{(ab - cd)^2}{(a + d)(c + b)(a + c)(d + b)} \dots \dots \dots (xxviii).$$

Now it is clear that in this case  $\phi^2$  reduces to  $r_{hk}^2$ , where  $r_{hk}$  is the correlation between errors in the position of the means of the two characters under consideration, as determined by a fourfold table, and  $\frac{1}{2}\psi$  is in this simple case what I have defined as the transfer per unit of total frequency.\* Both are expressions intimately connected with the conception of association, and have already been discussed in relation to it.† The coefficients,  $C_1$  and  $C_2$ , of contingency—either of which might serve as a measure of the association—will not in this simple case, however, be necessarily even approximately equal to each other, still less to either the coefficient of correlation or Mr. YULE'S coefficient of association.‡

It is worth while illustrating this on a numerical example. Taking the small-pox returns for the epidemic of 1890, we have :—

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

\* 'Phil. Trans.,' A, vol. 195, pp. 12 and 14.

† *Ibid.*, p. 15 *et seq.*

‡ 'Phil. Trans.,' A, vol. 194, p. 272.



These give us  $\phi^2 = .0845$ ,  $\chi^2 = 175.76$ ,  $\psi = .0604$ . From these we find

$$C_1 = .279, \quad C_2 = .190.$$

YULE'S coefficient of association = .803.

Coefficient of correlation by fourfold division = .595.

Grade of contingency =  $1 - P$ ,\* where  $P = 718/10^{40}$ .

Now so far as numerical values go these things are all totally different.  $C_1$ ,  $C_2$ , and the coefficient of association depend very largely on where the fourfold division is taken.† It is extremely difficult to use them therefore for comparative purposes. On the other hand, the coefficient of correlation with the assumption, however, of normality is free of this restriction; it brings us into line with other things for comparative purposes. The grade of contingency is also independent in a sense of the division, *i.e.*, it has a definite physical meaning. What it tells us is this, that the deviation from independent probability in the relation between result, a case of small-pox and presence or absence of cicatrix is such that the above table could only arise 718 times in  $10^{40}$  cases if the two events were absolutely independent.

If, instead of a table like the above, we take a number of alternative possibilities for each attribute, the coefficient of association loses its uniqueness of meaning;  $C_1$  and  $C_2$  still retain their significance, and as the number of alternatives become greater, merge in the coefficient of correlation. The grade of contingency, on the other hand, retains the same perfectly definite meaning throughout. I think this statement may serve as some warning of the caution needful in using the coefficients now introduced. The degree of approach of both  $C_1$  and  $C_2$  to the correlation must be studied for each special class of cases, and only when this has been done will their use be really legitimate and effective.

(10.) *On the Relation between Multiple Contingency and Multiple Normal Correlation.*

Suppose instead of a single correlation table we have a multiple correlation system. Such a system is well illustrated by the cabinet at Scotland Yard, which contains the measurements of habitual criminals on the old system of body measurements, now discarded in favour of a finger-print index. We have in this case a division of the cabinet into 3 compartments, which mark a threefold division of long, medium, and

\* When the number of groups = 4, we have ('Phil. Mag.,' vol. 50, p. 157 *et seq.*):—

$$P = \sqrt{\frac{2}{\pi}} \int_x^\infty e^{-\frac{1}{2}x^2} dx + \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}x^2} x \\ = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}x^2} x \left\{ 1 - \frac{1}{x^2} + \frac{1}{x^4} - \frac{3}{x^6} + \frac{15}{x^8} \right\},$$

whence  $P$  is easily found if  $\chi^2$  be large.

† YULE, *loc. cit.*, p. 276.



short head lengths. Each of these vertical divisions is then sub-divided horizontally into three divisions giving the corresponding divisions for head breadth; each of these head-breadth divisions has three drawers for large, moderate, and small face breadths. Each drawer is sub-divided into three sections for three finger groups, and these again into compartments for cubit groups, and so on. If this be carried out for the seven characters dealt with, we should have ultimately  $3^7$  sub-groups forming a multiple correlation system of the 7<sup>th</sup> order.\* We may ask what is the mean square contingency of such a system and to what extent does it diverge from an independent probability system? Of course, for an ideal anthropometric index system the divergence should be very slight.

Let  $x_1, x_2 \dots x_n$  be the  $n$  variables of a multiple normal correlation surface, to which the equation is

$$z = \frac{N}{(2\pi)^n \sigma_1 \sigma_2 \dots \sigma_n \sqrt{R}} \text{expt.} - \frac{1}{2} \left\{ S_1 \left( \frac{R_{pp} x_p^2}{R \sigma_p^2} \right) + 2S_2 \left( \frac{R_{pq} x_p x_q}{R \sigma_p \sigma_q} \right) \right\}.$$

Here  $\sigma_1, \sigma_2 \dots \sigma_n$  are the standard deviations of the  $n$  variables;  $S_1$  denotes a sum of all values of  $p$  from 1 to  $n$ ,  $S_2$  a sum of all unlike values of  $p$  and  $q$  from 1 to  $n$ ; while  $R$  is the determinant

$$\begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ r_{21} & 1 & r_{23} & \dots & r_{2n} \\ r_{31} & r_{32} & 1 & \dots & r_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & r_{n3} & \dots & 1 \end{vmatrix}$$

and  $R_{st}$  is the minor corresponding to the constituent  $r_{st}$ , and the  $r$ 's are the correlation coefficients.†

Now if  $\phi^2$  be the mean square contingency, we have

$$\phi^2 = \frac{1}{N} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \frac{(z - z_0)^2}{z_0} dx_1 dx_2 \dots dx_n,$$

where  $z_0$  = value of  $z$  when all the  $r$ 's are zero, or

$$z_0 = \frac{N}{(2\pi)^n \sigma_1 \sigma_2 \dots \sigma_n} \text{expt.} - \frac{1}{2} \left\{ S_1 \left( \frac{x_p^2}{\sigma_p^2} \right) \right\}.$$

Thus we have, writing  $x_p = \sigma_p x'_p$ , etc.,

$$\phi^2 = \frac{1}{(2\pi)^n} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \left( \frac{\zeta^2}{\zeta_0} - 2\zeta + \zeta_0 \right) dx'_1 dx'_2 \dots dx'_n,$$

\* See MACDONELL, "On Criminal Anthropometry," 'Biometrika,' vol. 1, p. 205 *et seq.*

† 'Phil. Trans.,' A, vol. 187, p. 302, or *Ibid.*, A, vol. 200, pp. 3-8.







Now multiply the determinant  $\Delta'$  by the determinant R, we find, using the above relations,

$$\Delta'R = \begin{vmatrix} 1, & -r_{12}, & -r_{13}, & \dots & -r_{1n} \\ -r_{21}, & 1, & -r_{23}, & \dots & -r_{2n} \\ -r_{31}, & -r_{32}, & 1, & \dots & -r_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ -r_{n1}, & -r_{n2}, & -r_{n3}, & \dots & 1, \end{vmatrix}$$

= R', say.

Here R' is R with the sign of all the correlations *changed*. Hence it follows that

$$\phi^2 = \frac{1}{\sqrt{RR'}} - 1 \dots \dots \dots \text{(xxx.)}$$

*Special Cases.*

(i.) Simple correlation

$$R' = R = 1 - r_{12}^2, \text{ and } \phi^2 = r_{12}^2 / (1 - r_{12}^2), \text{ as before.}$$

(ii.) Triple correlation

$$R = 1 - r_{23}^2 - r_{31}^2 - r_{12}^2 + 2r_{23}r_{31}r_{12},$$

$$R' = 1 - r_{23}^2 - r_{31}^2 - r_{12}^2 - 2r_{23}r_{31}r_{12}.$$

$$\phi^2 = \frac{1}{\sqrt{(1 - r_{23}^2 - r_{31}^2 - r_{12}^2)^2 - 4r_{23}^2r_{31}^2r_{12}^2}} - 1.$$

(iii.) Quadruple correlation

$$RR' = \{1 - r_{12}^2 - r_{13}^2 - r_{14}^2 - r_{23}^2 - r_{24}^2 - r_{34}^2 + r_{12}^2r_{34}^2 + r_{23}^2r_{14}^2 + r_{13}^2r_{24}^2 - 2(r_{12}r_{14}r_{23}r_{34} + r_{14}r_{13}r_{23}r_{24} + r_{12}r_{13}r_{24}r_{34})\}^2 - 4\{r_{23}r_{24}r_{34} + r_{34}r_{14}r_{13} + r_{12}r_{14}r_{24} + r_{12}r_{13}r_{23}\}^2,$$

and so on.

Clearly a condition has to be satisfied among the correlation coefficients, or the process by which we have deduced  $\phi^2$  is not legitimate. We must have  $\Delta$  positive for equation (xxix.) to be true. Now, for *normal* correlation R must be real and positive, or the equation to the multiple correlation surfaces become imaginary. Hence it follows that  $\Delta'$  must be positive, and therefore R' must be positive. This seems to give a definite condition to be satisfied by the correlation coefficients, and in some cases rather narrow limits are enforced. For example, in the case of triple correlation we must have

$$1 - r_{23}^2 - r_{31}^2 - r_{12}^2 - 2r_{23}r_{31}r_{12}$$

positive, and this appears to reduce very considerably the possible values for the







correlationship of three characters.\* The source of this novel condition appears to lie in the integration of the term  $\zeta^2/\zeta_0$ , and this is only possible by use of equation (i.), provided the surface  $Z = \zeta^2/\zeta_0$  has "ellipsoidal" contours. If it has not, we may get the subject of integration becoming infinite with one or other of the  $x$ 's, and consequently, although both  $\zeta$  and  $\zeta_0$  vanish at  $\infty$ ,  $\zeta^2/\zeta_0$  may not do so, *i.e.*, the mean square contingency tends in certain tracks to become indefinitely large. In fact, our method of deducing multiple contingency from the normal correlation coefficients is only valid provided the system is not only a possible correlation system with the given values of the coefficients, but also when these coefficients all have their signs reversed.

(11.) *Illustrations.* A.—*Stature in Father and Son.*

Table II. gives the distribution of 1078 cases of stature in father and son.† The correlation  $r$ , as found from the product moment in the usual way, is .514.

I propose to consider the approach of  $C_1$  and  $C_2$  to  $r$  as we increase the fineness of the grouping. Clearly it would involve extreme labour to work out the contingencies—especially the mean square contingency—for the table as it stands.

To begin with I classed in three inch groups and got the following table, in which the figures in brackets are the independent probabilities.

TABLE III.—Stature of Father and Son in Inches.

		Stature of Father.						Totals.	Chances.
		58.5-61.5.	61.5-64.5.	64.5-67.5.	67.5-70.5.	70.5-73.5.	73.5-76.5.		
Stature of Son.	58.5-61.5	— (.05)	1.5 (.36)	2 (1.20)	— (1.32)	— (.50)	— (.03)	3.5	.0032
	61.5-64.5	3.5 (.84)	19 (6.50)	33 (21.75)	5.5 (23.87)	1.5 (9.02)	— (.55)	62.5	.0580
	64.5-67.5	8.5 (4.02)	53.75 (31.07)	148 (104.03)	80.5 (114.15)	8.25 (45.14)	— (2.64)	299	.2774
	67.5-70.5	2.5 (6.07)	33.25 (46.86)	149.25 (156.90)	202.25 (172.17)	60.25 (65.06)	3.5 (3.97)	451	.4184
	70.5-73.5	— (2.87)	3.5 (22.13)	39.75 (74.10)	104.25 (81.31)	62 (30.73)	3.5 (1.88)	213	.1976
	73.5-76.5	— (.56)	1 (4.31)	3 (14.44)	14.5 (15.84)	20.5 (5.99)	2.5 (.37)	41.5	.0385
	76.5-79.5	— (.10)	— (.77)	— (2.59)	4.5 (2.84)	3 (1.07)	— (.07)	7.5	.0069
Totals . .		14.5	112	375	411.5	155.5	9.5	1078	1.0000

\* For example, if .5 be the value of parental correlation, then the correlation of two brothers could not exceed .5 without making  $R'$  negative.

† See 'Biometrika,' vol 2, p. 415.



The independent probabilities were found by multiplying the "chances" of a son occurring in each group by the totals for each group of fathers. Taking the difference of the observed sub-group frequencies and the independent probability frequencies, we have  $N \times \psi = 205.62$  from the positive and  $= -205.66$  from the negative differences, a quite good agreement. Hence we find  $\psi = .1908$ .

Using Diagram I. we have

$$C_2 = .522.$$

Proceeding now to the mean square contingency obtained by squaring all the above found contingencies, dividing each by the independent probability frequency and summing, we find

$$\phi^2 = .2755,$$

whence

$$C_1 = .465.$$

The value of  $C_1$  is clearly too small. We must infer that our grouping was not fine enough. Accordingly in Table IV. I have re-arranged the matter in 2-inch groupings, and have then in the same manner proceeded to find  $\psi$  and  $\phi^2$ . In this case I found  $\psi = .2013$ , and thus

$$C_2 = .542,$$

while

$$\phi^2 = .3568,$$

and

$$C_1 = .513.$$

I thus conclude that the grouping is now fine enough to give  $C_1$  and  $C_2$  approximately equal to the correlation.\*

\* *i.e.*, within the probable error of that result.



TABLE IV.—Stature of Father and Son in Inches.

		Stature of Father.								Totals.	Chances.	
		58.5-60.5.	60.5-62.5.	62.5-64.5.	64.5-66.5.	66.5-68.5.	68.5-70.5.	70.5-72.5.	72.5-74.5.			74.5-76.5.
Stature of Son.	59.5-61.5	— (.02)	— (.08)	1.5 (.31)	1 (.77)	1 (.95)	— (.84)	— (.41)	— (.11)	— (.02)	3.5	.00325
	61.5-63.5	.5 (.14)	2.75 (.56)	5.75 (2.11)	9.5 (5.29)	5 (6.49)	.25 (5.73)	.25 (2.83)	— (.72)	— (.12)	24	.02226
	63.5-65.5	4 (.60)	7.75 (2.32)	20 (8.81)	41.5 (22.03)	17.25 (27.04)	8.25 (23.89)	1.25 (11.78)	— (3.01)	— (.51)	100	.00276
	65.5-67.5	2 (1.43)	10 (5.51)	32 (20.93)	73 (52.33)	78.75 (64.22)	33.5 (56.73)	7.25 (27.98)	1 (7.16)	— (1.21)	237.5	.22032
	67.5-69.5	— (1.95)	4.5 (7.49)	27.75 (28.46)	65.5 (71.16)	95 (87.34)	93.25 (77.15)	31.5 (38.03)	4.5 (9.74)	1 (1.65)	323	.29963
	69.5-71.5	— (1.42)	— (5.47)	6.75 (20.80)	38.25 (51.99)	61 (63.82)	77.5 (56.37)	39.5 (27.80)	11 (7.11)	2 (1.20)	236	.21892
	71.5-73.5	— (.63)	— (2.44)	.25 (9.5)	5.75 (23.13)	24.75 (28.39)	34.5 (25.08)	32.25 (12.37)	7 (3.17)	.5 (.54)	105	.09740
	73.5-75.5	— (.23)	— (.87)	1 (3.31)	3 (8.26)	6.25 (10.14)	6.75 (8.96)	13 (4.42)	5.5 (1.13)	2 (.19)	37.5	.03479
	75.5-77.5	— (.05)	— (.19)	— (.70)	— (1.76)	2.5 (2.16)	1.5 (1.91)	1.5 (.94)	2.5 (.24)	— (.04)	8	.00742
	77.5-79.5	— (.02)	— (.08)	— (.31)	— (.77)	— (.95)	2 (.84)	.5 (.41)	1 (.11)	— (.02)	3.5	.00325
	Totals.		6.5	25	95	237.5	291.5	257.5	127	32.5	5.5	1078

To show the effect of too fine a grouping, I worked out the mean contingency for the inch grouping in Table II. There resulted

$$\psi = .2309, \text{ giving } C_2 = .597.$$

I therefore conclude that with sufficiently fine grouping the new method of contingency will give contingency coefficients sensibly equal to the correlation coefficient. But that with over fine grouping, the effect of individual units scattered here and there at random over the table, becomes influential and exaggerates the value of the correlation. Hence, when a correlation table can be formed and worked in the old ways, there is little doubt that it is safer to do so, and the labour will hardly be sensibly greater, at least when compared with the method of mean square contingency. I have not faced the labour required to determine the mean square contingency of the table with 340 sub-groups. Dr. LEE has worked out the mean square contingency for a table with 400 sub-groups, and we do not think it desirable to deal with a table of more than  $10^2$  to  $15^2$  entries again. Still the mean square contingency coefficient will hardly be as great on the full table as the mean contingency coefficient.

The following table gives the results:—



## COMPARISON of Methods of Finding Correlation.

No. of groupings.	Mean contingency.	Mean square contingency.	Fourfold division.	Correlation table.
42	.522	.465	(Mean of six divisions)*	—
90	.542	.513	.550	—
340	.597	—	—	.514

Thus the first contingency method approaches the fourfold, the second, the ordinary correlation method.

Diagram II. at the end of this memoir gives the hyperbola of zero contingency for this case, calculated on the basis of the correlation coefficient being .514. The means and standard deviations are :—

Father . . . . .	67''·698,	2''·7048,
Son . . . . .	68''·661,	2''·7321,

and the equation to the hyperbola referred to the means as the origin is

$$x^2 - 3.8522yx + .9801y^2 = 6.2510.$$

The shaded squares are those of positive contingency. It will be seen that the hyperbola separates fairly well areas of positive, from areas of negative contingency. In most cases where there is an invasion across the boundary, the contingencies hardly differ from zero by amounts greater than the probable errors due to random sampling.

*Illustration B.—Data from Colour Inheritance in Greyhounds.*

In the previous example we have dealt with material in which contingency methods were directly comparable as to result with the correlation found by the "best" or product method process. In this illustration I deal with matter which can only provide a correlation to be found by the fourfold division process for comparison with the contingency coefficients. The data from which this illustration is drawn were extracted by Miss A. BARRINGTON from the 'Greyhound Studbook.' We deal with the inheritance of red and black pigments in the coat colour. I have selected six cases of the resemblance of brethren from *different* litters to compare the methods on. Tables were formed giving 16 to 25 contingency sub-groups of varying degrees of pigment, and these were worked out (a) by Miss BARRINGTON herself for the mean square contingency, (b) by myself for the mean contingency, and (c) by Dr. A. LEE

\* See 'Phil. Trans.,' A, vol. 195, p. 42. The values range from .521 to .594, or almost the same range as we obtain from the mean contingency results.



for the fourfold correlation results. The results reached are given in the accompanying table. It is desirable to state that the number dealt with was about 1000 pairs of brethren in each case.

TABLE V.—Fraternal Resemblance of Greyhounds from Different Litters.

Character.	C <sub>1</sub> , Mean Square Contingency.	C <sub>2</sub> , Mean Contingency.	r, Fourfold Table.
Red in brothers . . . . .	·478	·695	·456
"  sisters . . . . .	·528	·612	·620
"  sister and brother . . . . .	·488	·615	·450
Black in brothers . . . . .	·512	·615	·558
"  sisters . . . . .	·482	·632	·552
"  sister and brother . . . . .	·502	·622	·593
Mean . . . . .	·498	·632	·538
Mean deviation from mean . . . . .	·016	·032	·057

We see at once from this table that the method of mean square contingency gives far more uniform results than either the mean contingency method or the fourfold division method. The average given by it is close to what we have found for fraternal resemblance, *i.e.*, ·5, in other cases, and within fairly close limits, all six cases now give ·5. The mean contingency gives results more divergent among themselves, but less so than those of the fourfold division method; their average, however, diverges most from what we have found in other cases.

The lesson, I think, to be learnt from this is: That the mean square contingency coefficient, although more laborious to find, is better than the mean contingency coefficient. That even with only 16 to 25 contingency sub-groups we may deduce results comparable with those obtained by fourfold divisions. But that it is probably *always* necessary to check a series by a certain number of fourfold division workings, for such are the only test that we have not got too crude a grouping reducing the contingency coefficient below the correlation value, or too fine a grouping introducing the difficulty already referred to (see p. 16), of magnifying the contingency coefficient owing to anomalous units.

*Illustration C.—Hair Colour in Man.*

I take the subject of hair colour because it is one in which doubts have been raised as to the order of pigments in a scale.

The following table gives the resemblance of pairs of brothers in hair colour:—



TABLE VI.

		First Brother.					Totals.
		Red.	Fair.	Brown.	Dark.	Jet Black.	
Second Brother.	Red . . . . .	30·5	23	16	12	—	81·5
	Fair . . . . .	23	416	158	67·75	·25	665
	Brown . . . . .	16	158	394	98·25	8·25	674·5
	Dark . . . . .	12	67·75	98·25	328·5	19	525·5
	Jet Black . . . . .	—	·25	8·25	19	10	37·5
Totals . . . . .		81·5	665	674·5	525·5	37·5	1984

The correlation found by taking the mean of four four-fold table divisions was ·621.\*

This result is based on the above scale order. We will now see what difference will arise if we work by contingency, so that the *scale order is absolutely indifferent, e.g.,* red might follow jet black.

We find

$$\phi^2 = \cdot603896,$$

and accordingly  $C_1 = \cdot614$ , a result within the limits of the probable error identical with the value of  $r$  found from the four-fold division method.

This illustration confirms the opinion I have already expressed, *i.e.*, that if the contingency be calculated for 16 to 36 sub-groups we shall obtain by the method of mean square contingency satisfactory results, *i.e.*, values close to the coefficient of correlation as found by product moment or four-fold division methods. In this case, as in others, I find the mean contingency far inferior to the mean square contingency.

My experience seems to show that about 25 sub-groups is the distribution to be aimed at; 9 is too few. Thus I worked out the relationship of temper in sisters for three-fold division—sullen, good-tempered, quick-tempered—or for 9 sub-groups. The method of mean contingency gave ·44 and of mean squared contingency ·36. Both far too small, as I find from each of four four-fold divisions a result of about ·5.

*Illustration D.—On Occupational or Professional Correlation between Relatives.*

I take as a final illustration a case in which any idea of scale is practically inconceivable, and yet one in which it is of considerable interest to measure the deviation from independent probability. It belongs to a class of problems in which I hope this new method of contingency will be fruitful of result. In classifying men into occupational and professional groups, we clearly cannot do so on the basis of any

\* "Huxley Memorial Lecture," 'Journal of Anthropological Institute,' vol. 33, pp. 197 and 215.



scale which will put the army, church, and bar in any special order. On the other hand, it becomes of special interest to determine how far tastes and preferences for particular callings in life run in families. Miss EMILY PERRIN has undertaken a lengthy investigation of this kind, and has provided me with the pure contingency table given as Table VII. The occupations of 775 fathers and sons are here classed in broad general groups, which can be arranged purely alphabetically. More minute divisions and data for other series of relatives will be published later by Miss PERRIN, and it is not my present purpose to anticipate her conclusions, but merely to suggest the valuable applications which may be made of the novel methods to pure contingency results. What is the numerical measure of the relationship in pursuit between father and son, and how far is it removed from a mere chance relationship?

TABLE VII.—Contingency between Occupations of Fathers and Sons.

Nature of occupation.		Occupation of Son.													Totals.	
		Army.	Art.	Teacher, Clerk, Civil Servant.	Crafts.	Divinity.	Agriculture.	Landownership.	Law.	Literature.	Commerce.	Medicine.	Navy.	Politics and Court.		Scholarship and Science.
Occupation of Father.	Army . . . . .	28	—	4	—	—	—	1	3	3	—	3	1	5	2	50
	Art . . . . .	2	51	1	1	2	—	—	1	2	—	—	—	1	1	62
	Teacher, Clerk, Civil Servant	6	5	7	—	9	1	3	6	4	2	1	1	2	7	54
	Crafts . . . . .	—	12	—	6	5	—	—	1	7	1	2	—	—	10	44
	Divinity . . . . .	5	5	2	1	54	—	—	6	9	4	12	3	1	13	115
	Agriculture . . . . .	—	2	3	—	3	—	—	1	4	1	4	2	1	5	26
	Landownership . . . . .	17	1	4	—	14	—	6	11	4	1	3	3	17	7	88
	Law . . . . .	3	5	6	—	6	—	2	18	13	1	1	1	8	5	69
	Literature . . . . .	—	1	1	—	4	—	—	1	4	—	2	1	1	4	19
	Commerce . . . . .	12	16	4	1	15	—	—	5	13	11	6	1	7	15	106
	Medicine . . . . .	—	4	2	—	1	—	—	—	3	—	20	—	5	6	41
	Navy . . . . .	1	3	1	—	—	—	1	—	1	1	1	6	2	1	18
	Politics and Court . . . . .	5	—	2	—	3	—	1	8	1	2	2	3	23	1	51
	Scholarship and Science } . . . . .	5	3	—	2	6	—	1	3	1	—	—	1	1	9	32
	Totals . . . . .		84	108	37	11	122	1	15	64	69	24	57	23	74	86

Miss PERRIN has extracted this first series from the 'Dictionary of National Biography'; hence she has, as a rule, tabled the distinguished, or at least moderately distinguished, sons of less distinguished fathers. It is, for example, not easy to win any form of distinction in agriculture. For this reason the distribution of occupations



for sons differs widely from that of the occupations for fathers. There has accordingly been selection of the second generation, which undoubtedly must influence the result, *i.e.*, tend to weaken the observed relationship.

Working out the 196 contingencies, squaring, dividing by the independent probability frequencies, summing and averaging, I find for the mean square contingency

$$\phi^2 = 1.299206,$$

whence

$$\phi^2/(1 + \phi^2) = .393794,$$

and the coefficient of mean square contingency = .6275. This would correspond to the correlation in occupation between father and son. Now if occupation were settled solely by fitness or taste, and these characters were inherited as other human faculties, we should expect the correlation between father and son to be about .46.\* Or, roughly, the hereditary relationship is increased by about  $\frac{1}{3}$  in the matter of occupation. Remembering what we have noted as to selection above, the real increment is probably somewhat larger than this. Roughly, however, we may conclude from Miss PERRIN'S data that about  $\frac{3}{4}$  of the observed resemblance in occupation between father and son is due to hereditary influences, and the remaining  $\frac{1}{4}$  to environmental effect. These numbers are subject to revision when Miss PERRIN'S data are more ample and have been more fully analysed and discussed.

#### (12.) *General Conclusions.*

The general conception of contingency developed in this memoir I consider in the first place of *theoretical* importance. Its practical applications are not negligible, but are, for reasons given below, of less importance than might *à priori* be supposed.

(A.) In the first place, the conception of contingency enables us at once to generalise the notion of the association of two attributes developed by Mr. YULE. We can class individuals not into two alternate groups, but into as many groups with exclusive attributes as we please, and either the mean contingency or the mean square contingency will enable us to see the extent to which two such systems are contingent or non-contingent.

(B.) This result enables us to start from the mathematical theory of independent probability as developed in the elementary text books, and build up from it a generalised theory of association, or, as I term it, *contingency*. We reach the notion of a pure contingency table, in which the order of the sub-groups is of no importance whatever.

(C.) We then investigate the relation of contingency to normal correlation, and find that with normal frequency distributions both contingency coefficients pass with sufficiently fine grouping into the well-known correlation coefficient. Since, however,

\* 'Biometrika,' vol. 2, p. 379.



the contingency is independent of the order of grouping, we conclude that when we are dealing with alternative and exclusive sub-attributes, we need not insist on the importance of any particular order or scale for the arrangement of the sub-groups.

(D.) This conception can be extended from normal correlation to any distribution with linear regression; small changes (*i.e.*, such that the sum of their squares may be neglected as compared with the square of mean or standard deviation) may be made in the order of grouping without affecting the correlation coefficient.

(E.) The results (c) and (D) are not so fruitful for practical working as might at first sight appear, for they depend in practice on the legitimacy of replacing finite integrals by sums over a series of varying areas, where no quadrature formulâ is available. If we, to meet the difficulty, make a very great number of small classes, the calculation, especially of the mean square contingency, becomes excessively laborious. Further, since in observation individuals go by units, casual individuals, which may fairly represent the total frequency of a considerable area, will be found on some one or other isolated small area, and thus increase out of all proportion the contingency. The like difficulty occurs when we deal with outlying individuals in the case of frequency curves, only it is immensely exaggerated in the case of frequency surfaces.

(F.) It is thus not desirable in actual practice to take too many or too fine sub-groupings. It is found, under these conditions, that the correlation coefficient as determined by the product moment or fourfold division methods is approximated to more closely in the case of the contingency coefficient found from mean square contingency than in the case of that found from mean contingency. Probably 16 to 25 contingency sub-groups will give fairly good results in the case of mean square contingency, but for each particular type of investigation it appears desirable to check the number of groups proper for the purpose by comparison with the results of test fourfold division correlations. Under such conditions it appears likely that very steady and consistent results will be obtained from mean square contingency.

(G.) Finally, contingency may be applied—of course, at first tentatively and with caution—in the consideration of a whole class of problems in which no attempt at a scale or order of sub-groups is possible, in short, where alphabetical order is as good as any other. For example, it would seem to be available in a vast range of problems of exclusive and alternative inheritance.

---



The author of this work has endeavored to present a fair and accurate account of the events which have shaped the history of the United States. It is intended for the use of students and the general reader.

(1) The author has endeavored to present a fair and accurate account of the events which have shaped the history of the United States. It is intended for the use of students and the general reader.

The author has endeavored to present a fair and accurate account of the events which have shaped the history of the United States. It is intended for the use of students and the general reader.

The author has endeavored to present a fair and accurate account of the events which have shaped the history of the United States. It is intended for the use of students and the general reader.

The author has endeavored to present a fair and accurate account of the events which have shaped the history of the United States. It is intended for the use of students and the general reader.







Diagram I.

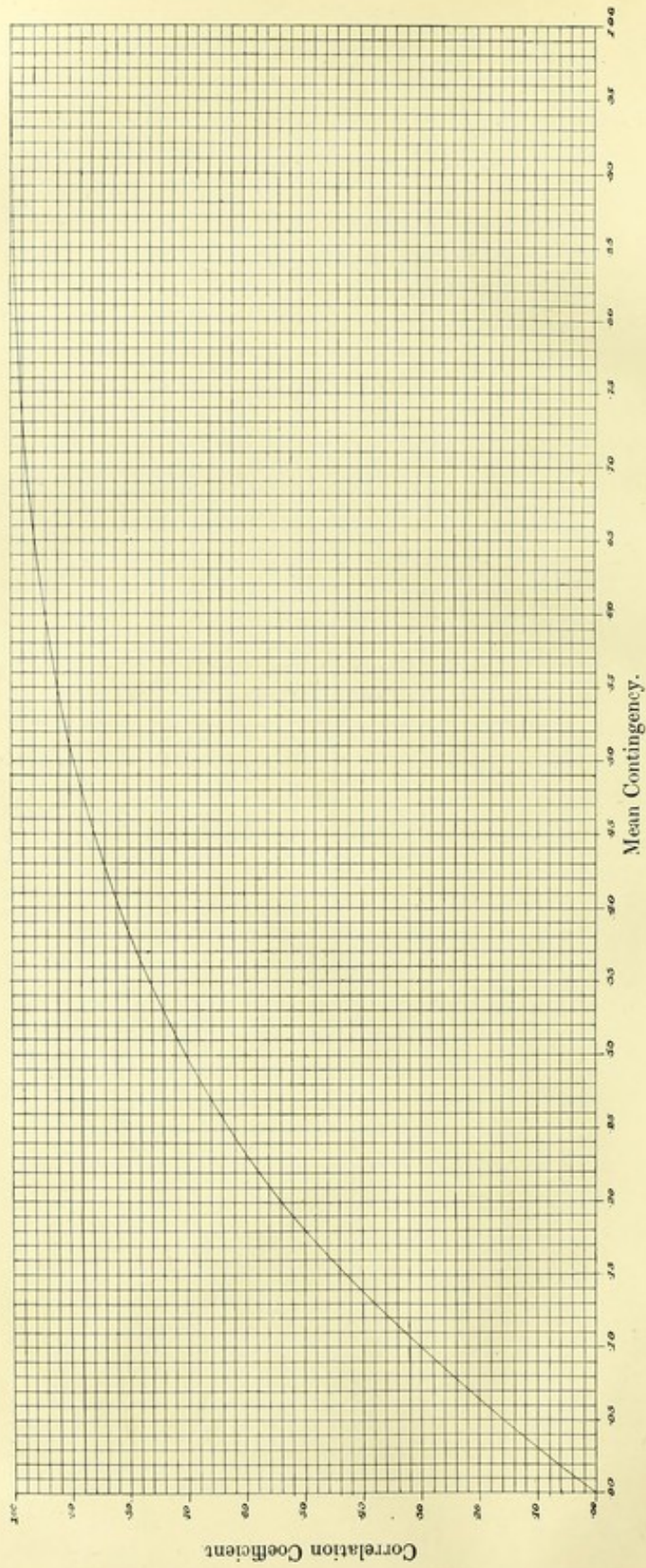
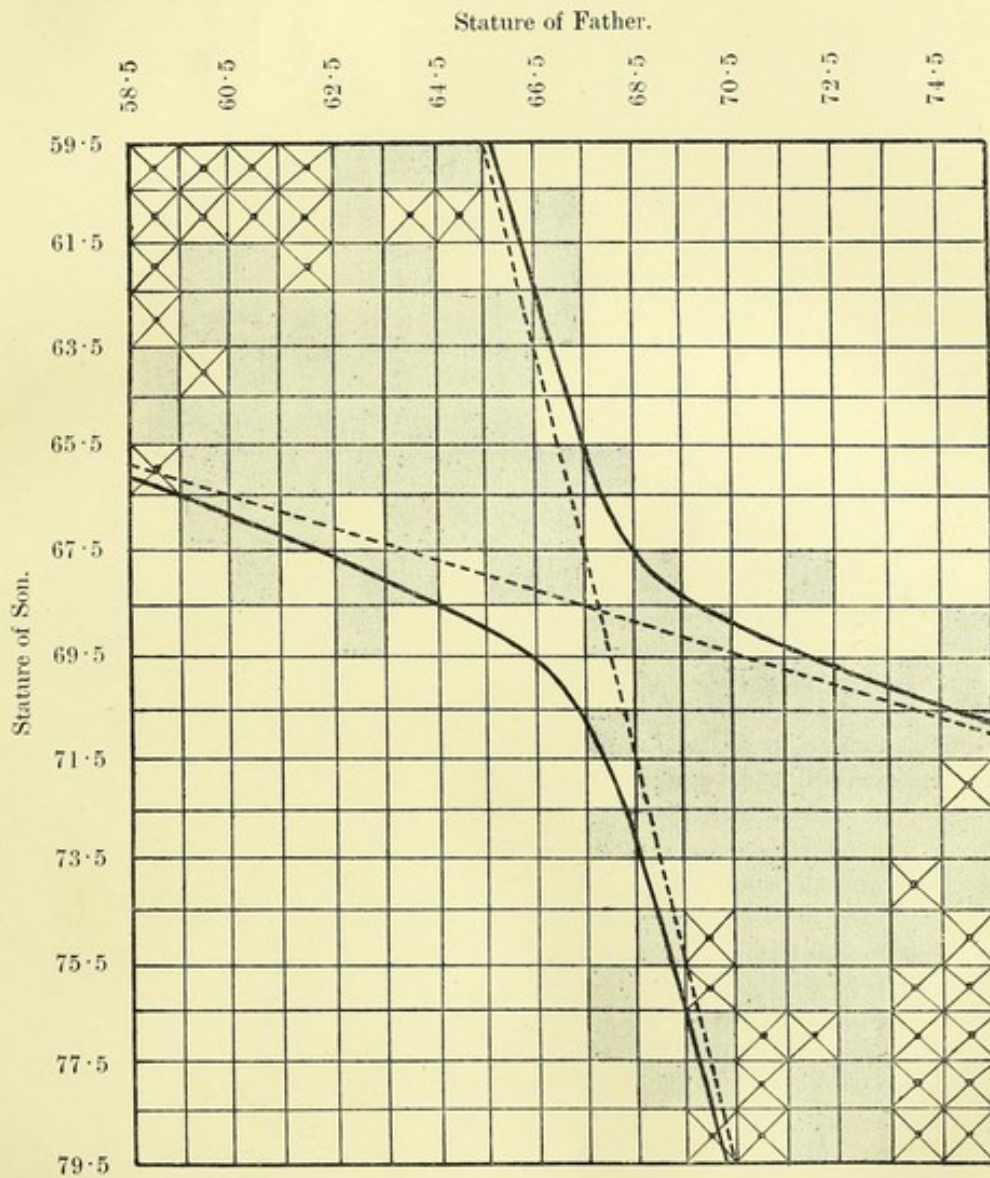




Plate II.

Diagram II. Illustrating areas of positive and negative Contingency and the Hyperbola of Zero-Contingency.



Sub-groups with plus contingency marked thus:

Sub-groups within the hyperbolic area, where there is *no* frequency in the observations, and which *must* therefore give negative contingency, marked thus: X

N.B.—Owing to an oversight on the part of the engraver, the absolute squareness of the elements in the original drawing has been disregarded in this reproduction.











## DRAPERS' COMPANY RESEARCH MEMOIRS.

DEPARTMENT OF APPLIED MATHEMATICS, UNIVERSITY COLLEGE,  
UNIVERSITY OF LONDON.

These memoirs will be issued at short intervals. The following are nearly ready and will probably appear in this series:—

### *Biometric Series I.*

- I. Mathematical Contributions to the Theory of Evolution.—XIII. On the Theory of Contingency and its Relation to Association and Normal Correlation. By KARL PEARSON, F.R.S.
- II. Mathematical Contributions to the Theory of Evolution.—XIV. On Homotypis in the Animal Kingdom. By ERNEST WARREN, D.Sc., ALICE LEE, D.Sc., EDNA LEA-SMITH, MARION RADFORD and KARL PEARSON, F.R.S.
- III. Mathematical Contributions to the Theory of Evolution.—XV. On the Theory of Skew Correlation and Non-linear Regression. By KARL PEARSON, F.R.S.

### *Technical Series II.*

- I. On some Points in the Theory of Structures:—
  - A. On Masonry Dams.
  - B. On the Relative Strength of Two-pivoted, Three-pivoted and Built-in Metal Arches. By W. L. ATCHERLEY, assisted by KARL PEARSON, F.R.S.
- II. On Crane and Coupling Hooks. By E. S. ANDREWS, B.Sc. Eng.
- III. On Torsional Vibrations in Shafting. By KARL PEARSON, F.R.S.

PUBLISHED BY DULAU AND CO.

## MATHEMATICAL CONTRIBUTIONS TO THE THEORY OF EVOLUTION.

### XI. ON THE INFLUENCE OF SELECTION ON THE VARIABILITY AND CORRELATION OF ORGANS.

By KARL PEARSON, F.R.S.

'Phil. Trans.,' vol. 200, pp. 1-56. Price 3s.

### XII. ON A GENERALISED THEORY OF ALTERNATIVE INHERITANCE, WITH SPECIAL REFERENCE TO MENDEL'S LAWS.

By KARL PEARSON, F.R.S.

'Phil. Trans.,' vol. 203, pp. 53-86. Price 1s. 6d.

PUBLISHED BY THE CAMBRIDGE UNIVERSITY PRESS.

## BIOMETRIKA.

### A JOURNAL FOR THE STATISTICAL STUDY OF BIOLOGICAL PROBLEMS.

Edited in Consultation with FRANCIS GALTON,

By W. F. R. WELDON, KARL PEARSON and C. B. DAVENPORT.

#### VOL. II, PART IV.

- I. On the Laws of Inheritance in Man. I. Inheritance of Physical Characters. (With 9 Figures.) By KARL PEARSON, F.R.S. and ALICE LEE, D.Sc.
- II. Variation in *Ophiocoma Nigra* (O. F. MÜLLER). By D. C. McINTOSH, M.A., F.R.S.E.
- III. Tables of Powers of Natural Numbers and of the Sums of Powers of the Natural Numbers from 1-100. By W. PALIN ELDERTON.
- IV. Associative Mating in Man. A Cooperative Study.
- Miscellanea. (I.) Inheritance in *Phaseolus vulgaris*. By W. F. R. WELDON and K. PEARSON.  
(II.) Addendum to "Graduation and Analysis of a Sickness Table." By W. PALIN ELDERTON.  
(III.) Chronological Notes:  
(iv.) Homogeneity and Heterogeneity in Crania. By CHARLES S. MYERS.  
Remarks on Dr. MYERS' Note. By K. PEARSON.  
(v.) On Cranial Types. By Professor AUREL VON TROSK.  
Remarks on Professor VON TROSK's Note. By K. PEARSON.

#### VOL. III, PART I.

- I. On the Result of Crossing Japanese Waltzing with Albino Mice. By A. D. DARBISHIRE.
- II. Graduation of a Sickness Table by MAKEHAM's Hypothesis. By JOHN SPENCER.
- III. On the Protective Value of Colour in *Mantis religiosa*. By A. P. DI CERNOLA.
- IV. Measurements of One Hundred and Thirty Criminals. By G. B. GRIFFITHS. With Introductory Note. By H. B. DONKIN.
- V. A First Study of the Weight, Variability and Correlation of the Human Viscera, with Special Reference to the Healthy and Diseased Heart. By M. GREENWOOD, JUN.
- VI. Sui Massimi delle Curve Dimorfiche. Dal Dr. FERNANDO HELOPERO.
- Miscellanea. (I.) On some Dangers of Extrapolation. By EMILY PERLIN.  
(II.) On Differentiation and Homotypis in the Leaves of *Fagus sylvatica*. By KARL PEARSON and MARION RADFORD.  
(III.) Albinism in Sicily and MENDEL'S LAWS. By W. F. R. WELDON.  
(IV.) A Mendelian's View of the Law of Ancestral Heredity. By K. PEARSON.

The subscription price, payable in advance, is 30s. net per volume (post free); single numbers 10s. net. Volumes I. and II. (1902-3) complete, 30s. net per volume. Bound in Buckram 34s. 6d. net per volume. Subscriptions may be sent to Messrs. C. J. Clay & Sons, Cambridge University Press Warehouse, Ave Maria Lane, London, either direct or through any bookseller.



DEPARTMENT OF APPLIED MATHEMATICS,  
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON.

---

DRAPERS' COMPANY RESEARCH  
MEMOIRS.

BIOMETRIC SERIES II.

---

MATHEMATICAL CONTRIBUTIONS TO THE  
THEORY OF EVOLUTION.

XIV. ON THE GENERAL THEORY OF SKEW CORRELATION  
AND NON-LINEAR REGRESSION.

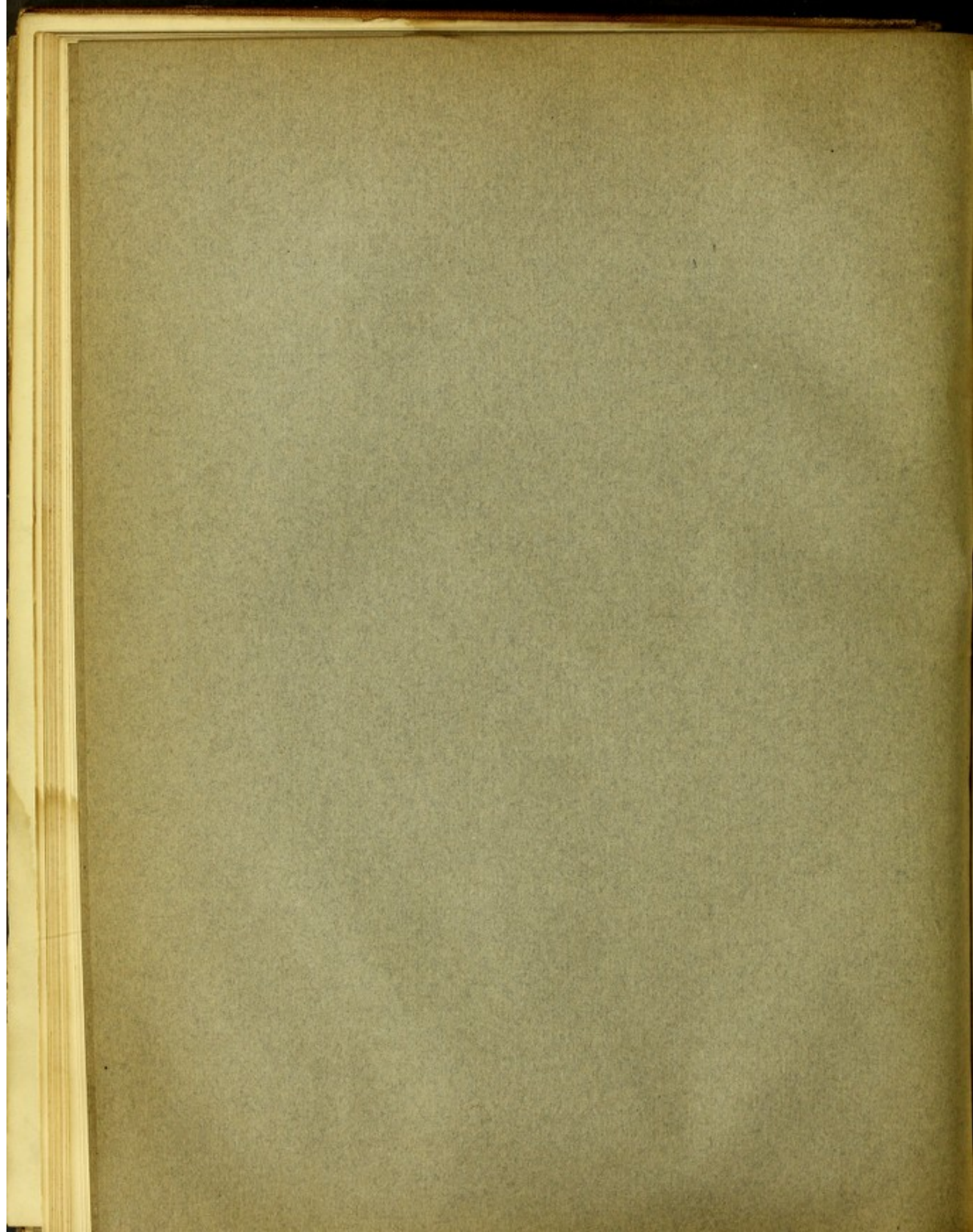
BY  
KARL PEARSON, F.R.S.

[WITH FIVE DIAGRAMS.]

LONDON:  
PUBLISHED BY DULAU AND CO., 37, SOHO SQUARE, W.  
1905.

*Price Five Shillings.*







DEPARTMENT OF APPLIED MATHEMATICS,  
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON.

---

DRAPERS' COMPANY RESEARCH  
MEMOIRS.

BIOMETRIC SERIES II.

---

MATHEMATICAL CONTRIBUTIONS TO THE  
THEORY OF EVOLUTION.

XIV. ON THE GENERAL THEORY OF SKEW CORRELATION  
AND NON-LINEAR REGRESSION.

BY

KARL PEARSON, F.R.S.

[WITH FIVE DIAGRAMS.]

LONDON:  
PUBLISHED BY DULAU AND CO., 37, SOHO SQUARE, W.  
1905.

*Price Five Shillings.*



*In March, 1903, the Worshipful Company of Drapers announced their intention of granting £1,000 to the University of London to be devoted to the furtherance of research and higher work at University College. After consultation between the University and College authorities, the Drapers' Company presented £1,000 to the University to assist the statistical work and higher teaching of the Department of Applied Mathematics. It seemed desirable to commemorate this—probably, first occasion on which a great City Company has directly endowed higher research work in mathematical science—by the issue of a special series of memoirs in the preparation of which the Department has been largely assisted by the grant. Such is the aim of the present series of "Drapers' Company Research Memoirs."*

*K. P.*



*Mathematical Contributions to the Theory of Evolution.—XIV. On the General Theory of Skew Correlation and Non-linear Regression.*

By KARL PEARSON, F.R.S.

CONTENTS.

	Page
(1.) Introductory. General conceptions as to skew variation and correlation. General theory of skew variation within the limits of practical errors of sampling. . . . .	3
(2.) Generalised idea of correlation. The correlation ratio $\eta$ and its relation to the correlation coefficient $r$ . . . . .	9
(3.) Probable errors of the correlation ratio and other constants of the arrays. Probable error of $r$ . . . . .	11
(4.) On the higher types of regression. Homoscedastic and heteroscedastic systems. Homoclitic and heteroclitic systems . . . . .	21
(5.) Cubical regression. General equations for regression of any order . . . . .	23
(6.) Parabolic regression. . . . .	28
(7.) Linear regression. . . . .	30
(8.) Illustration A.—On the skew correlation between number of branches to the whorl and position of the whorl on the spray in the case of <i>Asperula odorata</i> . . . . .	31
(9.) Illustration B.—On the skew correlation between age and head height in girls. . . . .	34
(10.) Illustration C.—On the skew correlation between size of cell and size of body in <i>Daphnia magna</i> . . . . .	38
(11.) Illustration D.—On the skew correlation between number of branches to the whorl and position of the whorl on the stem in <i>Equisetum arvense</i> . . . . .	42
(12.) Quartic regression. Necessary criteria for various types of regression . . . . .	47
(13.) Illustration E.—Calculation of quartic regression in the case of <i>Equisetum arvense</i> . . . . .	49
(14.) General conclusions. Nomenclature, clitic and scedastic curves. Difference between mere curve fitting and regression calculations. Remarks on retention of decimals . . . . .	51

(1.) *Introductory.*

IN a series of memoirs presented to the Royal Society I have endeavoured to show that the Gaussian-Laplace normal distribution is very far from being a general law of frequency distribution either for errors of observation\* or for the distribution of deviations from type such as occur in organic populations.† It is quite true that the

\* "On Errors of Judgment, &c.," 'Phil. Trans.,' A, vol. 198, pp. 235-299.

† "On Skew Variation, &c.," 'Phil. Trans.,' A, vol. 186, pp. 343-414.



normal distribution applies within certain fields with a remarkable degree of accuracy, notably in a whole series of anthropometric, particularly craniometric, observations.\* In other fields it is not even approximately correct, for example in the distribution of barometric variations,† of grades of fertility and incidence of disease.‡ For such cases I have introduced a series of skew frequency curves which serve the purpose of describing the frequency of innumerable skew distributions well within the errors of random sampling. An exact test for "goodness of fit" in the case of frequency distributions has also been now provided.§

In dealing with frequency which diverges more or less conspicuously from the normal law we require to bear in mind at least three important points:—

(i.) Any expression for frequency must be a graduation formula. It is not a disadvantage, but a fundamental requisite that it should smooth off "Scheingipfeln," so far as these are irregularities within the limits of random sampling.

Hence formulæ like those provided by THIELE|| and WUNDT's pupils,¶ which depend upon taking enough "moments" to reproduce the complete frequency, are *à priori* fallacious. Many interpolation formulæ would do this completely, but such interpolation formulæ are not graduation formulæ.

(ii.) The graduation formula must not depend upon the calculation of constants having such a high probable error that their value is practically worthless.

Now, the probable error of high moments and products increases rapidly with their dimensions; hence there is, beyond the labour of arithmetic, a practical limit to the number of moments or products which can be effectively used in a graduation formula.

(iii.) There must be a systematic method of approaching frequency distributions, which can be applied to all cases with reasonably practical ease.

Now the immense majority, if not the totality, of frequency distributions in homogeneous material show, when the frequency is indefinitely increased, a tendency to give a smooth curve characterised by the following properties:—

(i.) The frequency starts from zero, increases slowly or rapidly to a maximum, and then falls again to zero—probably at a quite different rate—as the character for which the frequency is measured is steadily increased. This is the almost universal unimodal distribution of the frequency of homogeneous series. Homogeneity may

\* 'Biometrika,' vol. I., p. 443; vol. II., p. 344; vol. III., p. 230.

† 'Phil. Trans.,' A, vol. 190, pp. 423-469.

‡ 'Phil. Trans.,' A, vol. 192, pp. 257-330; 'The Chances of Death,' vol. I., pp. 69, *et seq.*; 'Biometrika,' vol. I., p. 134 and p. 292; and for disease, 'Phil. Trans.,' A, vol. 186, pp. 390 and 407; A, vol. 197, p. 159.

§ 'Phil. Mag.,' vol. 50, 1900, pp. 157-174, and 'Biometrika,' vol. I., pp. 154-163.

|| 'Forelaesninger over Almindelig Iagttagelslaere,' Kjöbenhavn, 1889; 'Theory of Observations,' London, 1903.

¶ WUNDT, 'Philosophische Studien.' A whole series of papers, by G. F. LIPPS and others, seems to me to quite miss the point of (i.) and (ii.) above.



for practical purposes be taken to imply unimodality, although the converse is very far from true.

(ii.) In the next place there is generally contact of the frequency curve at the extremities of the range. These characteristics at once suggest the following form of frequency curve, if  $y\delta x$  measure the frequency falling between  $x$  and  $x+\delta x$  :—

$$dy/dx = \frac{y(x+a)}{F(x)} \dots \dots \dots (i).$$

For in this case we have one mode only of the frequency, *i.e.*, at  $x=-a$ , and  $dy/dx$  will vanish when  $y=0$ .

But the assumption of this form, as long as  $F(x)$  is general, is itself extremely general, and it includes cases in which  $dy/dx$  may not be zero, but take any values from 0 to  $\infty$ , when  $y=0$ .\*

Now let us assume that  $F(x)$  can be expanded by MACLAURIN'S theorem, and equals  $b_0+b_1x+b_2x^2+b_3x^3+\dots$ . Then our differential equation to the frequency will be

$$\frac{1}{y} \frac{dy}{dx} = \frac{x+a}{b_0+b_1x+b_2x^2+b_3x^3+\dots} \dots \dots \dots (ii).$$

There is now absolutely no difficulty in determining the unknown constants in terms of the moments of the system. Multiply up and also by  $x^n$ , and then integrate throughout the range of frequency, we have

$$\int x^n (b_0+b_1x+b_2x^2+b_3x^3+\dots) \frac{dy}{dx} dx = \int y(x+a) x^n dx \dots \dots (iii).$$

Or, noting that  $y=0$ , at the ends of the range we have, with the usual notation for a total frequency  $N$ , *i.e.*,

$$N\mu'_n = \int yx^n dx \dots \dots \dots (iv),$$

the result by integration by parts

$$nb_0\mu'_{n-1} + (n+1)b_1\mu'_n + (n+2)b_2\mu'_{n+1} + (n+3)b_3\mu'_{n+2} + \dots = -\mu'_{n+1} - a\mu'_n \quad (v).$$

Hence, if we write  $n=0, 1, 2, 3 \dots s$  successively, we have  $s+1$  equations to find  $a, b_0, b_1, b_2 \dots b_{s-1}$  in terms of the moments. For example, if we stop at  $b_0$  we require two moments, at  $b_1$  three moments, at  $b_2$  four moments, at  $b_3$  six moments, at  $b_4$  eight moments, and at  $b_{s-1}$ ,  $s > 2$ ,  $2s-2$  moments.

\* For example, cases in which there is a minimum frequency or antimode at  $x = -a$ , and  $dy/dx$  infinite at one or two values for which  $y=0$ , as in the frequency distributions discussed in 'Phil. Trans.,' A, vol. 186, pp. 364-5, and 'Roy. Soc. Proc.,' vol. 62, p. 287, "Cloudiness, a Novel Case of Frequency."







This equation gave Types I.-VI. of my two memoirs on skew variation,\* and provides at once the expressions

$$d = \text{distance from mode to mean} = \frac{\sigma \sqrt{\beta_1} (\beta_2 + 3)}{2 (5\beta_2 - 6\beta_1 - 9)} \dots \dots \dots \text{(xi.)}$$

$$\text{skewness} = \frac{\sqrt{\beta_1} (\beta_2 + 3)}{2 (5\beta_2 - 6\beta_1 - 9)} \dots \dots \dots \text{(xii.)}$$

where  $\sigma = \sqrt{\mu_2}$ ,  $\beta_1 = \mu_3^2/\mu_2^3$ ,  $\beta_2 = \mu_4/\mu_2^2$ , given in my memoir on the theory of errors of observation without proof.†

There is no *theoretical* limit, however, to this process; we can from (vi.) and (vii.) express the *a* and *b*'s at once in terms of determinants, and expanding obtain forms which, like the formulæ of THIELE, will fit closer and closer to the observed distribution of frequency, the more moments we take. But there are three fundamental *practical* objections to this. These are the following:—

(a.) Experience shows that the form (x.) suffices for certainly the great bulk of frequency distributions, *i.e.*, it describes them effectively within the limits of random sampling.

If the distribution be even approximately normal, the series in the denominator converges very rapidly, for the coefficients of every power of *x* vanish for moments obeying the relationships:—

$$\mu_{2s+1} = 0, \quad \mu_{2s} = (2s-1) \mu_2 \mu_{2s-2},$$

which hold for a normal series.

(b.) The labour of arithmetic and of analysis becomes very great, if we desire to keep higher moments. If we go to *b*<sub>4</sub> we should have to calculate the first eight moments of the observations about their centroid—a by no means easy task. Further, the classification of the resulting curves and the criteria for the right one to use in a special case, although not absolutely prohibitive, if we only go as far as *b*<sub>3</sub>, are for practical purposes idle in the case of taking into account *b*<sub>4</sub>.

(c.) The probable errors of the higher moments are so large that the values found for  $\mu_7$ ,  $\mu_8$ , &c., are quite untrustworthy, and even that for  $\mu_6$  is doubtful,‡ unless we have frequency series far larger than usually occur in actual observations. This is a strong argument against the utility of any descriptions of frequency, such as those suggested by THIELE or LIPPS, which depend upon moments higher than the fifth or sixth.

\* 'Phil. Trans.,' A, vol. 186, pp. 343-414, and 'Phil. Trans.,' A, vol. 197, pp. 443-459.

† 'Phil. Trans.,' A, vol. 198, p. 277.

‡ In 'Phil. Trans.,' A, vol. 185, pp. 71-110, I have given a method of breaking up a frequency distribution into two normal series. I obtained long ago the criterion for determining whether such a resolution is possible or not. But it involves moments higher than the fifth, and the probable error of the criterion is thus so great that for practical purposes it is worthless.



The question of the probable deviations of the higher moments can be illustrated as follows, by finding the standard deviation of the moment when we take a number of random samples from a general population. Let  $\Sigma_{\mu_r}$  be the standard deviation of  $\mu_r$ , then  $100 \Sigma_{\mu_r}/\mu_r$  is the percentage variability of  $\mu_r$  due to random sampling. The table below shows the increase of these percentages in the case of the moments of normal distributions, which, quite as well as any other, will illustrate the rapid increase in probable error as we use higher and higher moments. The general values of the standard deviations of some of the moments were first given by CZUBER,\* then far more completely by SHEPPARD,† and a *résumé* of all the results recently in 'Biometrika.'‡

PERCENTAGE Variability in Moments due to Random Sampling when the Series is supposed to be Normal.

Moment.	500 in series.	1000 in series.
$\mu_2$	6·3	4·5
$\mu_4$	14·6	10·3
$\mu_6$	30·1	21·3
$\mu_8$	60·6	42·9

Precisely the same rapid increase takes place when we find the variabilities of the ratios  $\mu_4/\mu_2^2$ ,  $\mu_6/\mu_2^3$ ,  $\mu_8/\mu_2^4$ , &c., which are the forms in which the moments actually occur in our coefficients. In this case we have to remember that errors in the moments are correlated, but the correlations are given in the papers cited above.§ I find in this case the following series, which is almost as suggestive as the previous table.

PERCENTAGE Variabilities in Ratio of Moments due to Random Sampling, the Series being Normal.

Ratio.	500 in series.	1000 in series.
$\mu_4/\mu_2^2$	7·3	5·2
$\mu_6/\mu_2^3$	23·3	16·5
$\mu_8/\mu_2^4$	55·1	39·0

The order of this increase of percentage variability, and therefore of probable error, is the same for skew as for normal variation, and it seems therefore, with the length

\* 'Theorie der Beobachtungsfehler,' S. 130, *et seq.*

† 'Phil. Trans.,' A, vol. 192, pp. 122, *et seq.*

‡ Vol. II., pp. 273-281.

§ *Ibid.*, p. 277.



of the series in customary use, idle to use the 7<sup>th</sup> or 8<sup>th</sup> moments; these have variabilities varying from 30 to 60 per cent. of their values, and accordingly we might easily on a random sample reach a 7<sup>th</sup> or 8<sup>th</sup> moment having half, or double the value it actually has in the general population. Constants based on these high moments will be practically idle. They may enable us to describe closely an individual random sample, but no safe argument can be drawn from this individual sample as to the general population at large, at any rate so far as the argument is based on the constants depending upon these high moments.

It seems to me accordingly obvious that, bearing in mind the object of a theory of frequency (*i.e.*, the description of the distribution in the general population by aid of a *graduated* sample, agreeing with the general population within the probable errors of random sampling), we can dismiss from practical use all theories which call upon us to use moments as high as the seventh or eighth. Any use of the general form (ii.) beyond  $b_3$ , indirectly or directly, involves such higher moments. Personally I am inclined to doubt whether the continental series using higher moments are, from the standpoint of graduation, nearly as good as my form (ii.).

Hence we seem driven to the skew curves embraced in (x.) as a practical frequency series. If we have a frequency not described by (x.) we may, perhaps, use  $\mu_5$  and  $\mu_6$ ,\* but it is difficult to see how its description can possibly be bettered by the use of still higher moments. This may seem a counsel of despair; but it is very far from being so in reality when we remember that (x.) has proved its efficiency now—I might almost say, without exception—in a wide range of economic, physical, biometric, and actuarial data.

In this memoir on skew correlation I shall accordingly confine my attention, for the most part, to constants the discovery of which does not involve the use of moments or products of higher than six dimensions, judging all above this limit to be, as a rule, disqualified for practical service by the magnitude of their probable errors.

### (2.) *Generalised Idea of Correlation.*

Given any two variables or characters A and B, we say that they are correlated when, with different values  $x$  of A, we do not find the same value  $y$  of B equally likely to be associated. In other words, certain values of B are relatively more likely to occur with the value  $x$  than others. The distribution of B's associated with a given value  $x$  of A is termed an  $x$ -array of B's. If N pairs of A and B are taken, and  $n_x$  of these have the character  $A = x$ , these  $n_x$  form the  $x$ -array of B's. This array, like any other frequency distribution, will have its mean, which we will denote by  $\bar{y}_x$ , and its

\* Referring to equation (ii.), I propose to call curves which stop at  $b_q$  skew curves of the  $q^{\text{th}}$  order. Thus the normal curve is a skew curve of zero order; curve of Type III. is a skew curve of the 1<sup>st</sup> order; Types I., II., V., and VI. are of the 2<sup>nd</sup> order. I hope shortly to publish a discussion of skew curves of the 3<sup>rd</sup> order to complete the practically legitimate range of such curves.



standard deviation, which we will denote by  $\sigma_x$ . The mean of all the B characters shall be  $\bar{y}$  and their variability given by the standard deviation  $\sigma_y$ . Similarly  $\bar{x}$ ,  $\sigma_x$  will denote the mean and standard deviation of the A's, and  $n_y$ ,  $\bar{x}_y$ , and  $\sigma_{n_y}$  the number of individuals, the mean and the standard deviation for a  $y$ -array of A's.

Now clearly a knowledge of  $\bar{y}_x$  and  $\sigma_{n_x}$  will not fix the B's which will be found associated with a given A, but it will define the limits of probable or even possible B's. The curve obtained by plotting  $\bar{y}_x$  to  $x$  is termed the regression curve of  $y$  on  $x$ . A curve in which the ratio of  $\sigma_{n_x}$  to the standard deviation  $\sigma_y$  is plotted to  $x$  may be termed a scedastic\* curve. Since the standard deviation is always a positive quantity, this curve always lies on one side of the axis; it is a horizontal line in the case of normal correlation—*i.e.*, the Gauss-Laplacian distribution of deviations—and coincides with the axis, in any case where correlation passes into causation, *i.e.*, when one value of B only is associated with each A.

The mean ordinate of this curve would clearly be a sort of general measure of the degree of correlation between A and B, but it seems for many reasons better to base our measure on the mean square of the weighted standard deviations of the arrays, or

$$\sigma_{a_x}^2 = S(n_x \sigma_{n_x}^2) / N \dots \dots \dots \text{(xiii).}$$

$\sigma_{a_x}$  will thus measure the average variability in B to be found associated with any A, its vanishing will mean that the scedastic curve as defined above will coincide with the axis. Now let a new quantity  $\eta$ , defined by

$$\sigma_{a_x}^2 = (1 - \eta^2) \sigma_y^2 \dots \dots \dots \text{(xiv).}$$

be introduced. Then clearly  $\eta$  must lie between  $\pm 1$ , because  $\sigma_{a_x}^2$  cannot be negative, being the sum of a number of positive squares. I term  $\eta$  the *correlation ratio*, to distinguish it from the *correlation coefficient* represented by  $r$ . When  $\eta = \pm 1$  the correlation is perfect or we have causation. Further we have by a well-known property of moments, if

$$\sigma_{m_x}^2 = S\{n_x (y_{n_x} - \bar{y})^2\} / N \dots \dots \dots \text{(xv).}$$

$$\sigma_y^2 = \sigma_{a_x}^2 + \sigma_{m_x}^2,$$

or

$$\eta = \sigma_{m_x} / \sigma_y \dots \dots \dots \text{(xvi).}$$

This shows us that the correlation ratio is the ratio of the variability of the means of the  $x$ -arrays to the variability of B's in general. If  $\eta = 0$ , it follows that  $\sigma_{m_x}$  is zero, or from (xv.) that every  $y_{n_x} = \bar{y}$ , *i.e.*, there is no association of B's with special A's at all, or correlation is zero. Thus the correlation ratio  $\eta$ , as defined by either (xiv.) or (xvi.), is an excellent measure of the stringency of correlation, always lying numerically between the values 0 and 1, which mark absolute independence and

\* *I.e.*, a curve which measures the "scatter" in the arrays.



complete causation respectively. Further, remembering the definition of  $r$ , the coefficient of correlation, *i.e.*,

$$\begin{aligned} N\sigma_x\sigma_y \times r &= S\{n_{xy}(x-\bar{x})(y-\bar{y})\}, \\ &= S\{n_x(x-\bar{x})(y_n-\bar{y})\} \dots \dots \dots \text{(xvii.)} \end{aligned}$$

we have, from (xv.) and (xvii.),

$$N(\eta^2 - r^2)\sigma_y^2 = S\left[n_x(y_n - \bar{y})\left\{y_n - \bar{y} - \frac{r\sigma_y}{\sigma_x}(x - \bar{x})\right\}\right].$$

Now let

$$Y = \bar{y} + \frac{r\sigma_y}{\sigma_x}(x - \bar{x}) \dots \dots \dots \text{(xviii.)}$$

then (xviii.), as is well known, gives the best fitting straight line to the series of points  $y_n$ , loaded with their respective  $n_x$ . We can now write

$$N(\eta^2 - r^2)\sigma_y^2 = S\{n_x(y_n - Y)^2\} + S\{n_x(Y - \bar{y})(y_n - Y)\}.$$

But, using (xviii.),

$$\begin{aligned} S\{n_x(Y - \bar{y})(y_n - Y)\} &= \frac{r\sigma_y}{\sigma_x} S\left[n_x(x - \bar{x})\left\{y_n - \bar{y} - \frac{r\sigma_y}{\sigma_x}(x - \bar{x})\right\}\right], \\ &= \frac{r\sigma_y}{\sigma_x} \left(Nr\sigma_x\sigma_y - \frac{r\sigma_y}{\sigma_x} N\sigma_x^2\right), \\ &= 0. \end{aligned}$$

Thus the last summation vanishes, and we have

$$N(\eta^2 - r^2)\sigma_y^2 = S\{n_x(y_n - Y)^2\} \dots \dots \dots \text{(xix.)}$$

The right-hand side must always be positive, unless  $y_n = Y$ , when it is zero. Hence we conclude that  $\eta$  is always greater than  $r$ , or the correlation ratio greater than the correlation coefficient, except in the special case when the means of the  $x$ -arrays of  $y$ 's all fall on a straight line, *i.e.*, we have linear regression, and then the two correlation constants are equal.

Thus the expression  $(\eta^2 - r^2)\sigma_y^2$  has an important physical meaning; it is the mean square deviation of the regression curve from the straight line which fits this curve most closely.\* We have now freed our treatment of correlation from any condition as to linearity of the regression, and it remains to consider the probable errors of the various quantities dealt with.

(3.) *Probable Errors of Constants of Correlation.*

We shall first prove a number of general propositions relating to the probable errors of correlation constants. We first note that if  $n$  and  $n'$  be the frequencies in

\* The properties of the correlation ratio were briefly noted in a footnote to a paper by the author in 'Roy. Soc. Proc.', vol. 71, pp. 303-4. It has been systematically used in my laboratory for some years and determined alongside  $r$  for many distributions.



any two sub-groups of a total  $N$ , for which no member of  $n$  is a member of  $n'$ , then the standard deviation of  $n$  due to random sampling is given by

$$\Sigma_n^2 = n \left( 1 - \frac{n}{N} \right) \dots \dots \dots \text{(xx.)}$$

and the correlation between deviations in  $n$  and  $n'$  due to random sampling is given by

$$R_{nn'} = - \frac{nn'}{N} \dots \dots \dots \text{(xxi.)}$$

*Problem I.—To find the correlation in deviations due to random sampling between the number  $n_x$  in the  $x_p$ -array of  $y$ 's and the number  $n_y$  in the  $y_s$ -array of  $x$ 's.*

If the symbol  $\delta n$  denote the error or deviation in  $n$ , we have with an obvious subscript notation\*

$$\delta n_{x_r} = \delta n_{x_r n_1} + \delta n_{x_r n_2} + \delta n_{x_r n_3} + \dots + \delta n_{x_r n_q}$$

if there be  $q$  groups of  $y$ 's, and again

$$\delta n_{y_s} = \delta n_{x_1 y_s} + \delta n_{x_2 y_s} + \delta n_{x_3 y_s} + \dots + \delta n_{x_i y_s}$$

if there be  $i$  groups of  $x$ 's.

Multiply the expressions for  $\delta n_{x_r}$  and  $\delta n_{y_s}$  together and we have

$$\delta n_{x_r} \delta n_{y_s} = (\delta n_{x_r y_s})^2 + S (\delta n_{x_r y_u} \delta n_{x_r y_v}),$$

where the summation is for every pair of values of  $u$  and  $v$ , differing from  $s$  and  $p$ .

Summing all such pairs of values for every random sample and dividing by the number of samples taken, we have the usual definition of correlation

$$\Sigma_{n_r} \Sigma_{n_s} R_{n_r n_s} = n_{x_r y_s} \left( 1 - \frac{n_{x_r y_s}}{N} \right) - S \left( \frac{n_{x_r y_u} n_{x_r y_v}}{N} \right);$$

or,

$$\Sigma_{n_r} \Sigma_{n_s} R_{n_r n_s} = n_{x_r y_s} - \frac{n_{x_r} n_{y_s}}{N} \dots \dots \dots \text{(xxii.)}$$

This gives  $R_{n_r n_s}$ , the required correlation, since  $\Sigma_{n_r}$  and  $\Sigma_{n_s}$  are known from (xx.).

*Problem II.—To find the correlation between deviations in the total  $n_x$  of any array and in any sub-group  $n_{x,y}$  of this array.*

We have at once

$$\delta n_x \delta n_{x,y} = (\delta n_{x,y})^2 + S (\delta n_{x,y} \delta n_{x,u})$$

where  $u$  is to be taken every value other than  $s$  in the summation term. Summing for all random samples and dividing by their number, we have, after using results like (xx.) and (xxi.),

$$R_{n_x n_{x,y}} \times \Sigma_{n_r} \Sigma_{n_s} = n_{x,y} \left( 1 - \frac{n_{x,y}}{N} \right) \dots \dots \dots \text{(xxiii.)}$$

which gives  $R_{n_x n_{x,y}}$ .

\*  $n_{xy}$  = frequency of groups with characters  $x$  and  $y$ .



*Proposition III.*—There is no correlation between deviations in the mean of an  $x$ -array  $y_x$ , and the total number in that array.

$$n_x \times y_x = S(n_{xy} y_u),$$

$$n_x \delta y_x = S(\delta n_{xy} y_u) - y_x \delta n_x,$$

$$n_x \delta y_x \delta n_x = -y_x (\delta n_x)^2 + S(\delta n_x \delta n_{xy} y_u).$$

Hence as before, using (xxiii.), &c.,

$$\begin{aligned} n_x \sum_{y_x} \sum_{n_x} R_{y_x n_x} &= -y_x n_x \left(1 - \frac{n_x}{N}\right) + S \left\{ n_x y_u \left(1 - \frac{n_x}{N}\right) y_u \right\} \\ &= -y_x n_x \left(1 - \frac{n_x}{N}\right) + \left(1 - \frac{n_x}{N}\right) n_x y_x \\ &= 0, \end{aligned}$$

which proves that  $R_{y_x n_x}$  is zero.

*Proposition IV.*—There is no correlation between deviations in the mean of an  $x$ -array and in the total number in any other array.

Proof as before.

*Proposition V.*—There is no correlation between deviations in the mean of one  $x$ -array and in the mean of a second  $x$ -array.

We have

$$n_x \delta y_x = S(\delta n_{xy} y_u) - y_x \delta n_x,$$

$$n_{x'} \delta y_{x'} = S(\delta n_{x'y} y_u) - y_{x'} \delta n_{x'}.$$

Multiply these two expressions together, sum for all random samples, and divide by the number of such samples. We find

$$\begin{aligned} n_x n_{x'} \sum_{y_x} \sum_{y_{x'}} R_{y_x y_{x'}} &= -y_x y_{x'} \frac{n_x n_{x'}}{N} \\ &\quad + y_x S(n_x n_{x'} y_u) / N \\ &\quad + y_{x'} S'(n_{x'} n_{xy} y_u) / N \\ &\quad - S(n_{xy} n_{x'y} y_u^2) / N \\ &\quad - S'(n_{xy} n_{x'y} y_u y_u) / N \\ &= -y_x y_{x'} \frac{n_x n_{x'}}{N} + y_x \frac{n_x n_{x'}}{N} y_{x'} \\ &\quad + y_{x'} \frac{n_x n_{x'}}{N} y_x - \frac{S(n_{xy} y_u) \times S(n_{x'y} y_u)}{N}. \end{aligned}$$

The last term is  $\frac{n_x y_x \times n_{x'} y_{x'}}{N}$ , and thus the right-hand side is identically zero. It thus appears that there is no correlation between errors made in finding the means of two arrays. This result is not at once obvious, although a very little consideration shows it must be true.



*Proposition VI.—To prove that the standard deviation of the mean  $y_x$  of any  $x$ -array due to random sampling equals  $\frac{\sigma_{n_x}}{\sqrt{n_x}}$ .*

We have

$$n_x \delta y_x = S'(\delta n_{x,y} y_u) - y_x \delta n_x.$$

Square, sum for all random samples, and divide by the number of such samples. We have

$$\begin{aligned} n_x^2 \Sigma y_x^2 &= y_x^2 n_x \left(1 - \frac{n_x}{N}\right) - 2y_x S \left\{ n_{x,y} \left(1 - \frac{n_x}{N}\right) y_u \right\} \\ &\quad + S \left\{ n_{x,y} \left(1 - \frac{n_x}{N}\right) y_u^2 \right\} \\ &\quad - 2S \left\{ \frac{n_{x,y} n_{x,y'}}{N} y_u y_{u'} \right\} \\ &= y_x^2 n_x \left(1 - \frac{n_x}{N}\right) - 2y_x^2 n_x \left(1 - \frac{n_x}{N}\right) \\ &\quad + S(n_{x,y} y_u^2) - \frac{S(n_{x,y} y_u) S(n_{x,y} y_{u'})}{N} \\ &= S(n_{x,y} y_u^2) - n_x y_x^2 \\ &= n_x \sigma_{n_x}^2. \end{aligned}$$

Hence

$$\Sigma y_x = \sigma_{n_x} / \sqrt{n_x} \dots \dots \dots \text{(xxiv).}$$

Thus the probable error of the mean of an array has exactly the same form as the probable error of the mean of a random sample of a *definite* number of individuals. The array may have a variable number of individuals, but we have seen in Proposition III. that there is no correlation between errors in its mean and errors in the total number of individuals contained in it.

*Problem VII.—To find the probable error of the standard deviation of any array.* By a precisely similar investigation to that of the previous proposition we find

$$\Sigma \sigma_{n_x} = \sqrt{\frac{m_4 - m_2^2}{4n_x m_2}} \dots \dots \dots \text{(xxv).}$$

where

$$m_2 = \frac{1}{n_x} S \{ (y_u - y_x)^2 n_{x,y} \}.$$

This is identical with the probable error we should have if the array were a random sample of constant size.

In many cases it will be sufficiently approximate to put  $m_4 = 3m_2^2$  and we then have

$$.67449 \Sigma \sigma_{n_x} = .67449 \frac{\sigma_{n_x}}{\sqrt{2n_x}} \dots \dots \dots \text{(xxvi).}$$



the well-known form for the probable error of the standard deviation of a normal distribution of a definite number of individuals.

*Problem VIII.*—To find the standard deviation of the standard-deviation  $\sigma_M$  of the means of the arrays due to random sampling.

Since

$$N\sigma_M^2 = S \{n_{x_r} (y_{x_r} - \bar{y})^2\}$$

$$2N\sigma_M \delta\sigma_M = S \{\delta n_{x_r} (y_{x_r} - \bar{y})^2\} + 2S \{\delta y_{x_r} n_{x_r} (y_{x_r} - \bar{y})\} - 2\delta\bar{y} S \{n_{x_r} (y_{x_r} - \bar{y})\},$$

the last term of which vanishes, since

$$N\bar{y} = S (n_{x_r} y_{x_r}).$$

Square the above relation, sum for all random samples, and divide by the number of such samples.

We find

$$\begin{aligned} 4N^2\sigma_M^2\Sigma\sigma_M^2 &= S \left\{ n_{x_r} \left( 1 - \frac{n_{x_r}}{N} \right) (y_{x_r} - \bar{y})^4 \right\} \\ &\quad - 2S \left\{ \frac{n_{x_r} n_{x_r'}}{N} (y_{x_r} - \bar{y})^2 (y_{x_r'} - \bar{y})^2 \right\} \\ &\quad + 4S \{ \Sigma_{n_{x_r}} \Sigma_{y_{x_r}} R_{n_{x_r} y_{x_r}} (y_{x_r} - \bar{y})^3 \} \\ &\quad + 4S \{ \Sigma_{n_{x_r}} \Sigma_{y_{x_r}} R_{n_{x_r} y_{x_r}} (y_{x_r} - \bar{y})^2 (y_{x_r'} - \bar{y}) \} \\ &\quad + 4S \{ \Sigma_{y_{x_r}} \Sigma_{y_{x_r'}} R_{y_{x_r} y_{x_r'}} (y_{x_r} - \bar{y}) (y_{x_r'} - \bar{y}) \} \\ &\quad + 4S \{ \Sigma y_{x_r}^2 n_{x_r}^2 (y_{x_r} - \bar{y})^2 \}. \end{aligned}$$

But  $R_{n_{x_r} y_{x_r}}$ ,  $R_{n_{x_r} y_{x_r}'}$ , and  $R_{y_{x_r} y_{x_r}'}$  vanish by Propositions III., IV., and V. Further, by VI.,  $\Sigma y_{x_r}^2 = \sigma_{n_{x_r}}^2 / n_{x_r}$ . Hence we have

$$\begin{aligned} 4N^2\sigma_M^2\Sigma\sigma_M^2 &= S \left\{ n_{x_r} \left( 1 - \frac{n_{x_r}}{N} \right) (y_{x_r} - \bar{y})^4 \right\} \\ &\quad - 2S \left\{ \frac{n_{x_r} n_{x_r'}}{N} (y_{x_r} - \bar{y})^2 (y_{x_r'} - \bar{y})^2 \right\} \\ &\quad + 4S \{ n_{x_r} \sigma_{n_{x_r}}^2 (y_{x_r} - \bar{y})^2 \} \\ &= S \{ n_{x_r} (y_{x_r} - \bar{y})^4 \} - \frac{[S \{ n_{x_r} (y_{x_r} - \bar{y}) \}]^2}{N} \\ &\quad + 4S \{ n_{x_r} \sigma_{n_{x_r}}^2 (y_{x_r} - \bar{y})^2 \}. \end{aligned}$$

Now let

$$N\lambda_q = S \{ n_{x_r} (y_{x_r} - \bar{y})^q \}$$

be the  $n^{\text{th}}$  moment of the means of the arrays about their mean. Then clearly  $\lambda_2 = \sigma_M^2$ . Further, since  $S (n_{x_r} \sigma_{n_{x_r}}^2) = N\sigma_y^2 (1 - \eta^2)$ , we can write

$$S \{ n_{x_r} \sigma_{n_{x_r}}^2 (y_{x_r} - \bar{y})^2 \} = N\sigma_y^2 (1 - \eta^2) \sigma_M^2 \times \chi_1,$$



where  $\chi_1$  is a purely numerical constant, which is equal to unity for those cases in which there is no correlation between the standard deviation of an array and the square of its mean's deviation from the mean. Thus finally we find

$$\Sigma \sigma_M^2 = \frac{\lambda_1 - \lambda_2^2}{4N\lambda_2} + \chi_1 \frac{\sigma_y^2 (1 - \eta^2)}{N} \dots \dots \dots \text{(xxvii).}$$

This enables us at once to find the probable error of the standard deviation of the means of the arrays.

*Proposition IX.*—To find the correlation between the deviations due to random sampling in the values of  $\sigma_y$  and  $\sigma_M$ .

We have

$$N\sigma_y^2 = S\{n_y (y - \bar{y})^2\},$$

$$2N\sigma_y \delta\sigma_y = S\{\delta n_y (y_s - \bar{y})^2\} - 2\delta\bar{y} S\{n_y (y_s - \bar{y})\};$$

the last term vanishes because  $S(n_y y_s) = N\bar{y}$ .

Thus

$$2N\sigma_y \delta\sigma_y = S\{\delta n_y (y_s - \bar{y})^2\}.$$

But from the previous proposition

$$2N\sigma_M \delta\sigma_M = S\{\delta n_{x_r} (y_{x_r} - \bar{y})^2\} + 2S\{\delta y_{x_r} n_{x_r} (y_{x_r} - \bar{y})\}.$$

Multiply these two expressions together, sum for all random samples and divide by the number of such samples; we find

$$4N^2 \sigma_y \sigma_M \Sigma_{\sigma_y} \Sigma_{\sigma_M} R_{\sigma_y \sigma_M} = S\{\Sigma_{n_y} \Sigma_{n_{x_r}} (y_s - \bar{y})^2 (y_{x_r} - \bar{y})^2 R_{n_y n_{x_r}}\} \\ + 2S\{n_{x_r} \Sigma_{n_y} \Sigma_{y_{x_r}} R_{n_y y_{x_r}} (y_s - \bar{y})^2 (y_{x_r} - \bar{y})\}.$$

To evaluate this, we require to find the two correlations expressed by  $R_{n_y n_{x_r}}$  and  $R_{n_y y_{x_r}}$ . We will consider the two summation terms separately.

*First Term.*  $\delta n_{x_r} = \delta n_{x_r y} + \delta n_{x_r y_2} + \dots + \delta n_{x_r y_p} + \dots$

$$\delta n_y = \delta n_{y x_1} + \delta n_{y x_2} + \dots + \delta n_{y x_p} + \dots$$

$$\delta n_{x_r} \delta n_y = (\delta n_{x_r y})^2 + S(\delta n_{x_r y} \delta n_{x_r y'}),$$

where in the summation  $p'$  and  $s'$  are not equal to  $p$  and  $s$ .

Proceeding in the usual manner we find

$$\Sigma_{n_{x_r}} \Sigma_{n_y} R_{n_{x_r} n_y} = n_{x_r y} \left(1 - \frac{n_{x_r y}}{N}\right) - S\left\{\frac{n_{x_r y} n_{x_r y'}}{N}\right\} \\ = n_{x_r y} - \frac{S(n_{x_r y'}) \times S(n_{x_r y})}{N},$$



where in the first sum  $s'$  is to take all possible values, and in the second  $p'$  is to take all possible values. Thus we have

$$\sum_{n_x} \sum_{n_y} R_{n_x n_y} = n_{x,y} - \frac{n_x n_y}{N} \dots \dots \dots \text{(xxviii.)}$$

Substituting we find

$$\begin{aligned} \text{First Term} &= S_1 \{ n_{x,y} (y_s - \bar{y})^2 (y_{x_r} - \bar{y})^2 \} \\ &\quad - S_2 \left\{ \frac{n_x n_y}{N} (y_s - \bar{y})^2 (y_{x_r} - \bar{y})^2 \right\}. \end{aligned}$$

Here both the summations are really double summations; fixing our attention on any  $x_p$ , i.e., on any array of  $y$ 's for a given value of  $x$ , we have first to sum for all  $y$ 's in this array, and then we have to sum for all arrays. This is the meaning of  $S_1$ . In  $S_2$  we are to associate every array of  $x$ 's with every array of  $y$ 's; hence this term will break up at once into two factors, i.e.,

$$\begin{aligned} &\frac{1}{N} S \{ n_x (y_{x_r} - \bar{y})^2 \} \times S \{ n_y (y_s - \bar{y})^2 \} \\ &= \sigma_y^2 \times S \{ n_x (y_{x_r} - \bar{y})^2 \} \\ &= N \sigma_y^2 \times \sigma_M^2. \end{aligned}$$

Keeping  $x_p$  constant first in  $S_1$ , we see that

$$S \{ n_{x,y} (y_s - \bar{y})^2 \}$$

is the 2<sup>nd</sup> moment of the  $y$ 's in the  $x_p$  array about the mean of the system

$$= n_x \{ \sigma_{n_x}^2 + (y_{x_r} - \bar{y})^2 \}.$$

Combining we have

$$\begin{aligned} \text{First Term} &= S \{ n_x (y_{x_r} - \bar{y})^4 \} + S \{ n_x \sigma_{n_x}^2 (y_{x_r} - \bar{y})^2 \} - N \sigma_y^2 \sigma_M^2 \\ &= N \{ \lambda_4 + \sigma_y^2 \sigma_M^2 (1 - \eta^2) \chi_1 - \sigma_y^2 \sigma_M^2 \} \dots \dots \dots \text{(xxix.)} \end{aligned}$$

We now turn to the second term which involves the discovery of  $R_{n_x y_r}$ .

$$\begin{aligned} \delta n_y \delta y_{x_r} &= (\delta n_{y,x_1} + \delta n_{y,x_2} + \dots + \delta n_{y,x_p} + \dots) \delta y_{x_r} \\ n_{x_r} \delta y_{x_r} &= -y_{x_r} \delta n_{x_r} + S (\delta n_{x,y} y_u). \end{aligned}$$

Hence

$$\begin{aligned} n_{x_r} \delta n_y \delta y_{x_r} &= -y_{x_r} (\delta n_{y,x_1} + \delta n_{y,x_2} + \dots + \delta n_{y,x_p} + \dots) \delta n_{x_r} \\ &\quad + (\delta n_{y,x_1} + \delta n_{y,x_2} + \dots + \delta n_{y,x_p} + \dots) S (\delta n_{x,y} y_u). \end{aligned}$$

Sum for all random samples and divide by the number of such samples; we have

$$\begin{aligned} n_{x_r} \sum_{n_y} \sum_{y_r} R_{n_x y_r} &= -y_{x_r} \left( n_x y_s - \frac{n_x n_y}{N} \right) \\ &\quad + n_{x,y} y_s - \frac{S (y_u n_{x,y} n_{x_r})}{N} \\ &= n_{x,y} (y_s - y_{x_r}) \dots \dots \dots \text{(xxx.)} \end{aligned}$$



Substituting we have

$$\text{Second Term} = 2S \{ n_{x,y} (y_s - y_{x_r}) (y_s - \bar{y})^2 (y_{x_r} - \bar{y}) \}.$$

Here again the summation is of a double character.

Let us first take  $x_p$  as constant and sum for every value of  $y_s$ . We may write  $y_s - \bar{y} = (y_s - y_{x_r} + y_{x_r} - \bar{y})$ , and our first summation will be

$$\begin{aligned} & 2 (y_{x_r} - \bar{y}) \times S [ n_{x,y} \{ (y_s - y_{x_r})^3 + 2 (y_s - y_{x_r})^2 (y_{x_r} - \bar{y}) + (y_s - y_{x_r}) (y_{x_r} - \bar{y})^2 \} ] \\ & = 2 (y_{x_r} - \bar{y}) n_{x_r} m_3 + 4 (y_{x_r} - \bar{y})^2 n_{x_r} m_2 + 2 (y_{x_r} - \bar{y})^3 S \{ n_{x,y} (y_s - y_{x_r}) \}, \end{aligned}$$

if

$$n_{x_r} m_2 = S \{ n_{x,y} (y_s - y_{x_r})^2 \}.$$

The last term vanishes for  $S (n_{x,y} y_s) = n_{x_r} y_{x_r}$  by the definition of the mean.

Hence

$$\text{Second Term} = 2S \{ n_{x_r} m_3 (y_{x_r} - \bar{y}) \} + 4S \{ n_{x_r} \sigma_{n_{x_r}}^2 (y_{x_r} - \bar{y})^2 \}.$$

Here  $m_3$  is the third moments of the  $x_p$  array of  $y$ 's, which will probably be very small if the arrays are nearly symmetrical and the first term clearly depends on the existence of a correlation between the skewness of the arrays and the magnitude of their means.

We may write the first term then :

$$\begin{aligned} & = 2N \sigma_y^3 \sigma_M \times \chi_2 \\ & = 2N \sigma_y^3 (1 - \eta^2)^{3/2} \sigma_M \times \chi_2, \end{aligned}$$

where  $\chi_2$  is a purely numerical quantity, which for most cases will probably be very small or even zero.

Thus we find :

$$\text{Second Term} = 2N \sigma_y^3 (1 - \eta^2)^{3/2} \sigma_M \chi_2 + 4N \sigma_y^2 \sigma_M^2 (1 - \eta^2) \chi_1 \dots \quad (\text{xxxii}).$$

We can now return to p. 16 and write down the full correlation between deviations in the values of  $\sigma_y$  and  $\sigma_M$  due to random sampling. Remembering that  $\sigma_M = \eta \sigma_y$ ,\* we find :

$$\begin{aligned} \Sigma_{\sigma_y} \Sigma_{\sigma_M} R_{\sigma_y, \sigma_M} &= \frac{1}{4N\eta} \left[ \frac{\lambda_4}{\sigma_y^2} + \eta^2 \sigma_y^2 \{ (1 - \eta^2) \chi_1 - 1 \} \right] \\ & \quad + \frac{1}{2N} \sigma_y^2 (1 - \eta^2)^{3/2} \chi_2 + \frac{\sigma_y^2}{N} \eta (1 - \eta^2) \chi_1 \\ &= \frac{\sigma_y^2}{N} \left\{ \frac{\lambda_4}{4\eta \sigma_y^4} + \frac{5}{4} \eta (1 - \eta^2) \chi_1 - \frac{1}{4} \eta + \frac{1}{2} (1 - \eta^2)^{3/2} \chi_2 \right\} \dots \quad (\text{xxxiii}). \end{aligned}$$

\* It should be remembered that this definition of  $\eta$  gives it invariably the positive sign.



*Proposition X.*—To find the standard deviation of the values of the correlation ratio  $\eta$  due to random sampling, i.e., to find the probable error of the correlation ratio  $\eta$ .

We have

$$\eta = \sigma_M / \sigma_y.$$

Hence

$$\frac{\delta\eta}{\eta} = \frac{\delta\sigma_M}{\sigma_M} - \frac{\delta\sigma_y}{\sigma_y}.$$

Squaring, summing for all random samples and dividing by the number of such samples, we have :

$$\frac{\Sigma \eta^2}{\eta^2} = \frac{\Sigma \sigma_M^2}{\sigma_M^2} + \frac{\Sigma \sigma_y^2}{\sigma_y^2} - \frac{2\Sigma \sigma_M \sigma_y R_{\sigma_M \sigma_y}}{\sigma_M \sigma_y}.$$

$\Sigma \sigma_M^2$  is given (xxvii.),  $\Sigma \sigma_M \sigma_y R_{\sigma_M \sigma_y}$  by (xxxii.) and  $\Sigma \sigma_y^2 = \frac{1}{4N} \frac{\mu_4 - \mu_2^2}{\mu_2}$  by a well-known formula.\*

Substituting, we have the complete value of  $\Sigma \eta$  given by :

$$\begin{aligned} \frac{\Sigma \eta^2}{\eta^2} &= \frac{\lambda_4 - \lambda_2^2}{4N\lambda_2^2} + \chi_1 \frac{(1-\eta^2)}{N\eta^2} + \frac{1}{4N} \frac{\mu_4 - \mu_2^2}{\mu_2} \\ &\quad - \frac{1}{2N} \frac{\lambda_4}{\lambda_2^2} \eta^2 - \frac{5}{2N} (1-\eta^2) \chi_1 + \frac{1}{2N} - \frac{(1-\eta^2)^{3/2}}{N\eta} \chi_2; \end{aligned}$$

or, after re-arranging,

$$\begin{aligned} \Sigma \eta^2 &= \frac{1}{N} \left\{ (1-\eta^2)^2 + \frac{\mu_4 - 3\mu_2^2}{4\mu_2^2} \eta^2 + \frac{\lambda_4 - 3\lambda_2^2}{4\lambda_2^2} \eta^2 (1-2\eta^2) \right. \\ &\quad \left. + (\chi_1 - 1) (1-\eta^2) (1-\frac{5}{2}\eta^2) - \chi_2 \eta (1-\eta^2)^{3/2} \right\} \dots \dots \text{(xxxiii).} \end{aligned}$$

For normal correlation,  $\mu_4 = 3\mu_2^2$ . Further

$$y_x - \bar{y} = \frac{r\sigma_y}{\sigma_x} (x_p - \bar{x}),$$

and

$$\begin{aligned} N\lambda_4 &= S\{n_x (y_x - \bar{y})^4\} = \frac{r^4 \sigma_y^4}{\sigma_x^4} S\{n_x (x_p - \bar{x})^4\} \\ &= \frac{r^4 \sigma_y^4}{\sigma_x^4} \times N3\sigma_x^4 = 3N\lambda_2^2. \end{aligned}$$

Hence the second and third terms vanish. Further  $\chi_1 = 1$  and  $\chi_2 = 0$ , while  $\eta = r$ .

Hence we have

$$\Sigma \eta^2 = \Sigma r^2 = \frac{(1-r^2)^2}{N},$$

which agrees with the special result.

\* 'Biometrika,' vol. II., p. 276.



In any other case,  $\chi_2$ ,  $\chi_1 - 1$ ,  $(\mu_4 - 3\mu_2^2)/\mu_2^2$ ,  $(\lambda_4 - 3\lambda_2^2)/\lambda_2^2$  will probably be small and thus

$$\Sigma_r^2 = \frac{1}{N} (1 - \eta^2)^2.$$

Probable error of

$$\eta = .67449 (1 - \eta^2)/\sqrt{N}, \text{ nearly } \dots \dots \dots \text{ (xxxiv.)}$$

This simple form suffices for many practical cases.

If greater exactitude is wanted, there is, however, no great labour in using (xxxiii.). We find the means and standard deviations of each array.

Then  $N\lambda_2$  and  $N\lambda_4$  are the 2<sup>nd</sup> and 4<sup>th</sup> moments of the means of these arrays about their mean.

$N\mu_2$  and  $N\mu_4$  are the 2<sup>nd</sup> and 4<sup>th</sup> moments about the mean of the  $y$ -characters, and will always be known for skew variation.

$\chi_1$  is defined by

$$\chi_1 = \frac{S\{n_x \sigma_{x_r}^2 (y_x - \bar{y})^2\}}{N \sigma_y^2 (1 - \eta^2) \sigma_M^2} \dots \dots \dots \text{ (xxxv.)}$$

and can be easily found when the means and standard deviations of each array have been found.

The most troublesome expression is  $\chi_2$  defined by

$$\chi_2 = \frac{S\{n_x m_3 (y_x - \bar{y})\}}{N \sigma_y^3 (1 - \eta^2)^2 \sigma_M^2} \dots \dots \dots \text{ (xxxvi.)}$$

But as we do not take usually more than 10 to 20 arrays, the discovery of their 3<sup>rd</sup> moments is not an extremely difficult task. As a rule, however,  $\chi_2$  is very small and may be fairly neglected, even when we must find  $\chi_1 - 1$ . All these points will be dealt with in the numerical illustrations given later in this paper. At present we note that the probable error of  $\eta$  has been determined, and that its value for the general case is not really more complex than the value of the probable error of  $r$  in the general case, which requires the determination of product moments of the 4<sup>th</sup> order.\*

\* Let  $Np_{qr} = S\{n_{xy} (x - \bar{x})^q (y - \bar{y})^r\}$ , then the probable error of  $r$  is given by

$$\Sigma_r^2 = \frac{r^2}{N} \left\{ \frac{p_{22} - 3p_{11}^2}{p_{11}^2} + \frac{p_{22} - 3p_{20}p_{02}}{2p_{20}p_{02}} + \frac{p_{40} - 3p_{20}^2}{4p_{20}^2} + \frac{p_{04} - 3p_{02}^2}{4p_{02}^2} - \frac{p_{31} - 3p_{11}p_{20}}{p_{11}p_{20}} - \frac{p_{13} - 3p_{11}p_{02}}{p_{11}p_{02}} \right\} \dots \text{ (xxxvii.)}$$

This agrees with the value given by SHEPPARD ('Phil. Trans.,' A, vol. 192, p. 128), except that the  $r^2$  factor has been dropped by a printer's error in his paper. For the special case of a normal distribution, we have easily from the equation to the normal surface

$$p_{40} = 3p_{20}^2, \quad p_{04} = 3p_{02}^2, \quad p_{31} = 3p_{11}p_{20}, \quad p_{13} = 3p_{11}p_{02}, \quad (p_{22} - 3p_{11}^2)/p_{11}^2 = (1 - r^2)/r^2$$

and

$$\frac{p_{22} - 3p_{20}p_{02}}{2p_{20}p_{02}} = r^2 - 1, \quad \text{whence } \Sigma_r = (1 - r^2)/\sqrt{N},$$

the well-known form ('Phil. Trans.,' A, vol. 191, p. 245).



(4.) *On the Higher Types of Regression.*

We have already seen how the introduction of the correlation ratio  $\eta$  enables us to drop the limitations associated with the Gauss-Laplacian form of frequency, and the Bravais correlation formulæ. The fundamental step towards this advance was undoubtedly taken by G. U. YULE in his paper in the 'Roy. Soc. Proc.' vol. 60, pp. 477 *et seq.*, wherein he shows that if the regression be linear, the Bravais type of formula applied to multiple correlation is still true, although we make no assumption as to the form of the frequency surface. It would undoubtedly be a gain to have skew frequency surfaces which would describe skew correlation for the great mass of cases as effectively as the series of skew frequency curves describe skew variation, but although a considerable amount of progress has been made in the consideration of these surfaces, their full theory has not yet been worked out owing to difficulties of analysis, and their complete discussion must still be postponed. YULE'S method of approaching the problem from the form of the regression curves is, however, available and capable of very great extension. Its chief advantage is that it makes little or no assumption as to the distribution of frequency; its chief defect lies even in this advantage of generality: it does not enable us to predict the probability of an individual with a given combination of characters. This follows at once from the fact that we make no assumption as to the form of the distribution within an array. Without some theory as to variation within the array, we are reduced to the laborious process of calculating the standard deviation, skewness, and other general characters of each array, a lengthy and troublesome process compared with a theory which would, like the Bravais theory, give these at once in terms of a few constants determined from the data as a whole.

In the great bulk of biometrical and economical enquiries, however, the regression does not diverge very markedly from the linear form. In the cases of non-linear regression that I have hitherto had to deal with, I find that parabolæ of the 2<sup>nd</sup> or 3<sup>rd</sup> order will suffice as a rule to describe the deviation from linearity. If they did not, we could, of course, use curves of higher orders, but the difficulty referred to in the first section of this paper at once arises: we then need to use in the determination moments and product-moments of such high orders that the probable errors of the constants are so high as to render valueless their calculation from such statistical data as we can hope for in most actual inquiries. In the great bulk of investigations it is practically impossible to increase our random samples from 500 to 1,000 individuals up to 50,000 to 100,000. Nor in the great bulk of statistical cases is any such increase even desirable, for a fairly wide experience shows that 2<sup>nd</sup> and 3<sup>rd</sup> order parabolæ amply suffice to describe the skewness of the regression line. I shall accordingly classify skew correlation in the following manner:—







or being asymmetrical in an equal degree about their means. I shall express this by the term *homoclitic*; generally the arrays will not be equally asymmetrical round their means, and in this case we shall speak of them as *heteroclitic*. If there were no skewness in any of the arrays, then  $m_3$  of (xxxvi.) would be zero for all of them. I term arrays of no skewness *isocurtic*, and skew arrays *allocurtic*. If we supposed that a curve of Type III. would sufficiently express the skewness of an array, we should have

$$\text{Sk.} = \frac{1}{2} m_3 / \sigma_{n_x}^3,$$

and therefore from (xxxvi.)

$$\chi_2 = \frac{2S\{n_x \sigma_{n_x}^3 (\text{Sk.})(y_x - \bar{y})\}}{N \sigma_y^3 (1 - \eta^2)^{3/2} \sigma_M} \dots \dots \dots \text{(xli.)}$$

For a homoscedastic system we have  $\sigma_{n_x} = \sigma_y \sqrt{1 - \eta^2}$ , and therefore

$$\chi_2 = \frac{2S\{n_x (\text{Sk.})(y_x - \bar{y})\}}{N \sigma_M},$$

and for a homoclitic system

$$\chi_2 = \frac{2(\text{Sk.}) S\{n_x \sigma_{n_x}^3 (y_x - \bar{y})\}}{N \sigma_y^3 (1 - \eta^2)^{3/2} \sigma_M}.$$

For a homoclitic homoscedastic system, whether isocurtic or allocurtic,

$$\chi_2 = \frac{2(\text{Sk.}) S\{n_x (y_x - \bar{y})\}}{N \sigma_M} = 0.$$

Thus  $\chi_2$  is to a certain extent a measure of both homoscedasticity and homoclisly. But as the correlation between  $\sigma_{n_x}$  and  $y_x - \bar{y}$  is in most cases extremely small, while the skewness of the array can well change its sign with arrays above or below the mean, we can fairly consider the smallness of  $\chi_2$  to be a measure of the approach to homoclisly. I am thus inclined to speak of  $\chi_1 - 1$  and  $\chi_2$  as measures of heteroscedasticity and heteroclisly. When they both vanish we have a homoscedastic homoclitic system. For such systems  $\eta$ , the correlation ratio, tells us effectively the scatter of any array, and as a rule all we want to know, in addition, is the form of the regression line.

(5.) *Cubical Regression.*

We have already used the following notation

$$N p_{qq} = S\{n_x (x - \bar{x})^q (y - \bar{y})^q\} \dots \dots \dots \text{(xlii.)}$$

We shall shorten our formulæ if we write

$$r = p_{11} / (\sigma_x \sigma_y), \quad \epsilon = p_{21} / (\sigma_x^2 \sigma_y), \quad \zeta = p_{31} / (\sigma_x^3 \sigma_y), \quad \theta = p_{41} / (\sigma_x^4 \sigma_y) \dots \text{(xliii.)}$$

We have already used  $\mu_q$  to denote  $p_{0q}$ , and we shall use  $\nu_q$  for  $p_{q0}$ . Further, we write

$$\beta_1 = \nu_3^2 / \nu_2^3, \quad \beta_2 = \nu_4 / \nu_2^2, \quad \beta_3 = \nu_5 \nu_3 / \nu_2^4, \quad \beta_4 = \nu_6 / \nu_2^3 \dots \dots \dots \text{(xliv.)}$$



$\sqrt{\beta_1} = \nu_3/\sigma_x^3$  will be of the same sign as  $\nu_3$ . These constants  $\beta$  have been previously used in the theory of skew variation.\*

We shall further put

$$\bar{\epsilon} = \epsilon - r\sqrt{\beta_1}, \quad \bar{\zeta} = \zeta - r\beta_2, \quad \bar{\theta} = \theta - r\beta_3/\sqrt{\beta_1} \dots \dots \dots \text{(xlv.)}$$

The regularity of the forms  $\bar{\epsilon}$ ,  $\bar{\zeta}$ ,  $\bar{\theta}$ , is rather screened by the above notation, which is introduced for brevity; using the  $p_{qq'}$  notation, we have

$$\bar{\epsilon} = \frac{P_{21}P_{20} - P_{11}P_{30}}{\sigma_x^4\sigma_y}, \quad \bar{\zeta} = \frac{P_{31}P_{20} - P_{11}P_{40}}{\sigma_x^5\sigma_y}, \quad \bar{\theta} = \frac{P_{41}P_{20} - P_{11}P_{50}}{\sigma_x^6\sigma_y} \dots \dots \dots \text{(xlvi.)}$$

whence the law of formation of these constants is easily seen.

The regression curve may now be conveniently put into the form

$$\frac{y_x - \bar{y}}{\sigma_y} = b_0 + b_1 \frac{x_p - \bar{x}}{\sigma_x} + b_2 \left( \frac{x_p - \bar{x}}{\sigma_x} \right)^2 + b_3 \left( \frac{x_p - \bar{x}}{\sigma_x} \right)^3 \dots \dots \dots \text{(xlvii.)}$$

Or, multiplying by  $n_x$  and summing for all arrays,

$$0 = Nb_0 + b_2N + b_3N\sqrt{\beta_1},$$

the sign of  $\sqrt{\beta_1}$  being always that of the 3<sup>rd</sup> moment. Hence, measuring from the means of the two characters, *i.e.*,  $X_p = x_p - \bar{x}$ ,  $Y_x = y_x - \bar{y}$ , we may re-write (xlvii.)

$$Y_x/\sigma_y = b_1(X_p/\sigma_x) + b_2\{(X_p/\sigma_x)^2 - 1\} + b_3\{(X_p/\sigma_x)^3 - \sqrt{\beta_1}\} \dots \dots \dots \text{(xlviii.)}$$

Now multiply by  $n_x X_p/\sigma_x$  and sum for all arrays, remembering that

$$Nr\sigma_x\sigma_y = S(n_x XY) = S(n_x X_p Y_x),$$

we find

$$r = b_1 + b_2\sqrt{\beta_1} + b_3\beta_2.$$

This enables us to get rid of  $b_1$  and write (xlviii.)

$$Y_x/\sigma_y = rX_p/\sigma_x + b_2\{(X_p/\sigma_x)^2 - \sqrt{\beta_1}(X_p/\sigma_x) - 1\} + b_3\{(X_p/\sigma_x)^3 - \beta_2(X_p/\sigma_x) - \sqrt{\beta_1}\} \dots \dots \dots \text{(xlix.)}$$

Now multiply by  $n_x(X_p/\sigma_x)^2$  and sum for all arrays. We have

$$\epsilon = r\sqrt{\beta_1} + b_2(\beta_2 - \beta_1 - 1) + b_3(\beta_3/\sqrt{\beta_1} - \beta_2\sqrt{\beta_1} - \sqrt{\beta_1}),$$

or

$$\bar{\epsilon} = b_2\phi_2 + b_3\phi_3 \dots \dots \dots \text{(l.)}$$

where

$$\left. \begin{aligned} \phi_2 &= \beta_2 - \beta_1 - 1 \\ \phi_3 &= (\beta_3 - \beta_1\beta_2 - \beta_1)/\sqrt{\beta_1} \end{aligned} \right\} \dots \dots \dots \text{(li.)}$$

\* 'Phil. Trans.,' A, vol. 186, p. 368, and A, vol. 198, p. 278.



Eliminating  $b_2$ , we can write (xlix.)

$$Y_{x_r}/\sigma_y = r(X_p/\sigma_x) + \frac{\bar{\epsilon}}{\phi_2} \{ (X_p/\sigma_x)^2 - \sqrt{\beta_1} (X_p/\sigma_x) - 1 \} \\ + b_3 \left[ (X_p/\sigma_x)^3 - \beta_2 (X_p/\sigma_x) - \sqrt{\beta_1} - \frac{\phi_3}{\phi_2} \{ (X_p/\sigma_x)^2 - \sqrt{\beta_1} (X_p/\sigma_x) - 1 \} \right]. \quad (\text{lii.})$$

Now multiply by  $n_{x_r} (X_p/\sigma_x)^3$  and sum for all arrays; we find

$$\zeta = r\beta_2 + \frac{\bar{\epsilon}}{\phi_2} \phi_3 + b_3 (\phi_4 - \phi_3^2/\phi_2),$$

or  $(\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)/(\phi_2\phi_4 - \phi_3^2) = b_3 \dots \dots \dots (\text{liii.})$

where  $\phi_4 = \beta_4 - \beta_2^2 - \beta_1 \dots \dots \dots (\text{liv.})$

It follows from (l.) that  $b_2 = (\bar{\epsilon}\phi_4 - \bar{\zeta}\phi_3)/(\phi_2\phi_4 - \phi_3^2) \dots \dots \dots (\text{lv.})$

We can thus write the cubic regression curve in either of the forms\*

\* The method is perfectly easy of extension, if we choose to use higher products and moments, to a regression curve of any order, *e.g.*,

$$Y_{x_r}/\sigma_y = b_0 + b_1(X_p/\sigma_x) + b_2(X_p/\sigma_x)^2 + \dots + b_n(X_p/\sigma_x)^n + \dots$$

For let:  $N\epsilon_{q1} = S(n_{xy}Y_{x_r}X_p^q)/(\sigma_x^q\sigma_y)$ , and  $\gamma_q = v_x/\sigma_x^q = S(n_{x_r}X_p^q)/(N\sigma_x^q)$ ,

we have:

$$0 = b_0 + 0 \times b_1 + b_2 + \gamma_3 b_3 + \dots + \gamma_n b_n + \dots$$

$$\epsilon_{11} = 0 \times b_0 + b_1 + \gamma_3 b_2 + \gamma_4 b_3 + \dots + \gamma_{n+1} b_n + \dots$$

$$\epsilon_{21} = b_0 + \gamma_3 b_1 + \gamma_4 b_2 + \gamma_5 b_3 + \dots + \gamma_{n+2} b_n + \dots$$

$$\dots \dots \dots$$

$$\epsilon_{p1} = \gamma_p b_0 + \gamma_{p+1} b_1 + \gamma_{p+2} b_2 + \gamma_{p+3} b_3 + \dots + \gamma_{n+p} b_n + \dots$$

$$\dots \dots \dots$$

Hence writing  $\epsilon_{01}$  for 0,  $\gamma_0 = 1, \gamma_1 = 0, \gamma_2 = 1$ , we have

$$b_n = (\epsilon_{01} \Delta_{0n} + \epsilon_{11} \Delta_{1n} + \epsilon_{21} \Delta_{2n} + \dots + \epsilon_{p1} \Delta_{pn} + \dots) / \Delta,$$

where  $\Delta = \begin{vmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 & \dots & \gamma_n & \dots \\ \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \dots & \gamma_{n+1} & \dots \\ \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 & \dots & \gamma_{n+2} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \gamma_p & \gamma_{p+1} & \gamma_{p+2} & \gamma_{p+3} & \dots & \gamma_{p+n} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{vmatrix}$

and  $\Delta_{qn}$  is the minor of the constituent in the  $(q+1)^{th}$  row and  $(n+1)^{th}$  column. As we have already noted, however, solutions involving anything beyond  $\gamma_6$  are hardly likely to be of practical value.

The value above for  $b_n$  is the type equation given by the method of least squares, when we strike the best fitting curve to all the entries in the correlation table. I have already pointed out that the method of moments becomes identical with that of least squares, when we fit parabolæ of any order ('Biometrika,' vol. I, p. 271). The retention of the method of moments, however, enables us, without abrupt change of method, to introduce the needful  $\eta$ , and to grasp at once the application of the proper SHEPPARD'S corrections. The extension of the method of least squares to *continua* in space has not yet, as far as I am aware, been fully considered.



$$Y_{x_p}/\sigma_y = r(X_p/\sigma_x) + \frac{\bar{\epsilon}}{\phi_2} \{(X_p/\sigma_x)^2 - \sqrt{\beta_1}(X_p/\sigma_x) - 1\} + \frac{\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3}{\phi_2\phi_4 - \phi_3^2} \left[ (X_p/\sigma_x)^3 - \beta_2(X_p/\sigma_x) - \sqrt{\beta_1} - \frac{\phi_3}{\phi_2} \{(X_p/\sigma_x)^2 - \sqrt{\beta_1}(X_p/\sigma_x) - 1\} \right]. \quad (\text{lvi.})$$

or

$$Y_{x_p}/\sigma_y = r(X_p/\sigma_x) + \frac{\bar{\epsilon}\phi_4 - \bar{\zeta}\phi_3}{\phi_2\phi_4 - \phi_3^2} \{(X_p/\sigma_x)^2 - \sqrt{\beta_1}(X_p/\sigma_x) - 1\} + \frac{\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3}{\phi_2\phi_4 - \phi_3^2} \{(X_p/\sigma_x)^3 - \beta_2(X_p/\sigma_x) - \sqrt{\beta_1}\} . . \quad (\text{lvi.}) \text{ bis.}$$

The former arrangement of the solution, while it is apparently more cumbersome, is, perhaps, the better, for it gives us at once the measure of the deviation from parabolic or 2<sup>nd</sup> order regression, *i.e.*, the approach of  $\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3$  to zero. In the case of normal correlation both  $\bar{\epsilon}$  and  $\bar{\zeta}$  vanish, and neglecting higher terms the condition for linear regression is that  $\bar{\epsilon} = 0$ , and  $\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3 = 0$ , or, again,  $\bar{\epsilon}$  and  $\bar{\zeta} = 0$ . For material in which the  $x$ -variability is isocurtic,  $\beta_1 = \beta_3 = \phi_3 = 0$ , and the regression curve takes the simple form

$$Y_{x_p}/\sigma_y = r(X_p/\sigma_x) + \frac{\bar{\epsilon}}{\phi_2} \{(X_p/\sigma_x)^2 - 1\} + \frac{\bar{\zeta}}{\phi_4} \{(X_p/\sigma_x)^3 - \beta_2(X_p/\sigma_x)\} . \quad (\text{lvi.}) \text{ ter.}$$

We now turn to express these relations in terms of the correlation ratio  $\eta$ . Multiply (lvi.) by  $n_x Y_{x_p}/\sigma_y$ , and sum for all arrays, we obtain

$$\eta^2 = r^2 + \frac{\bar{\epsilon}}{\phi_2} (\epsilon - \sqrt{\beta_1}r) + \frac{\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3}{\phi_2\phi_4 - \phi_3^2} \left\{ \zeta - \beta_2 r - \frac{\phi_3}{\phi_2} (\epsilon - \sqrt{\beta_1}r) \right\},$$

whence results

$$\phi_2 (\eta^2 - r^2) - \epsilon^2 = (\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2 / (\phi_2\phi_4 - \phi_3^2) . . . . . \quad (\text{lvii.})$$

(lvii.) is a necessary condition of cubical regression.

It is of course not a sufficient condition, as we ought to show that  $b_4, b_5, \&c.$ , all vanish, and thus any number of conditions may be found. For example, multiply by  $n_x X_p^4/\sigma_x^4$  and sum for all arrays, then

$$\theta = \frac{\bar{\epsilon}\phi_4 - \bar{\zeta}\phi_3}{\phi_2\phi_4 - \phi_3^2} (\beta_4 - \beta_3 - \beta_2) + \frac{\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3}{\phi_2\phi_4 - \phi_3^2} \frac{\beta_5 - \beta_2\beta_3 - \beta_1\beta_2}{\sqrt{\beta_1}} . . . . \quad (\text{lviii.})$$

is also a necessary condition. Here  $\beta_5 = \nu_7 \nu_3 / \sigma_x^{10}$ . But the high as well as complicated value of the probable errors of such expressions renders it idle to consider them in practice.



Substituting (lvii.) in (lvi.) we have :

$$Y_{x/y} = r(X_p/\sigma_x) + \frac{\bar{\epsilon}}{\phi_2} \{ (X_p/\sigma_x)^2 - \sqrt{\beta_1} (X_p/\sigma_x) - 1 \} \\ \pm \sqrt{\frac{\phi_2(\eta^2 - r^2) - \bar{\epsilon}^2}{\phi_2\phi_4 - \phi_3^2}} \left[ (X_p/\sigma_x)^3 - \beta_2 (X_p/\sigma_x) - \sqrt{\beta_1} \right. \\ \left. - \frac{\phi_3}{\phi_2} \{ (X_p/\sigma_x)^2 - \sqrt{\beta_1} (X_p/\sigma_x) - 1 \} \right] . \quad (\text{lix.})$$

Which sign is to be given to the root will often be visible on inspection of the observations. Otherwise the sign of the root must be the same as that of

$$\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3.$$

(lix.) will save the calculation of  $\bar{\zeta}$  if the root-sign can be found by inspection.

Finally there is a third form into which we may put the cubic. Eliminate  $\phi_2\phi_4 - \phi_3^2$  from (lix.) by aid of (lvii.) and it becomes

$$Y_{x/y} = r(X_p/\sigma_x) + \frac{\bar{\epsilon}\bar{\zeta} - \phi_3(\eta^2 - r^2)}{\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3} \{ (X_p/\sigma_x)^2 - \sqrt{\beta_1} (X_p/\sigma_x) - 1 \} \\ + \frac{\phi_2(\eta^2 - r^2) - \bar{\epsilon}^2}{\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3} \{ (X_p/\sigma_x)^3 - \beta_2 (X_p/\sigma_x) - \sqrt{\beta_1} \} . \quad (\text{lx.})$$

At first sight this might appear to be the best form of the cubic, because it does not involve the 6<sup>th</sup> moment of the variable  $x$ . But this is very far from being the case in actual practice. The reason is simply this,  $\bar{\epsilon}$ ,  $\bar{\zeta}$  and  $\eta^2 - r^2$  are in most cases very small—they vanish in normal correlation—relatively to  $\phi_2$  and  $\phi_4$ . Hence both numerators and denominators of the coefficients of the square and cubic terms are the ratio of small quantities, and accordingly subject to large probable errors. For this reason (lx.) was found in actual practice to be of no service. Of the other two forms (lvii.) and (lix.), which neither suffer from this defect,  $\phi_2\phi_4 - \phi_3^2$  being always large relative to the numerators, (lix.) while involving a 6<sup>th</sup> moment does not involve a 4<sup>th</sup> product,  $\bar{\zeta}$ , and experience shows that the former is on the whole easier to determine and more exact than the former. Hence (lix.) seems the preferable form, even if it be needful in certain cases to determine  $\bar{\zeta}$  in order to fix the sign of the radical. The cubic regression curve thus demands a knowledge of the correlation ratio  $\eta$ , of the “cubic product”  $\bar{\epsilon}$  and the sign by inspection or calculation of  $\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3$ . Besides this, we require the first six moments of the independent variable  $x$ . Of course if the regression of  $x$  on  $y$  be required, as well as that of  $y$  on  $x$ , the second correlation ratio and cubic product as well as the first six moments of  $y$  must be found. It is rare, however, that both regression curves are needed for a single enquiry.

As to the general form of (lix.), we note that there will always be a real point of inflexion given by

$$X_p/\sigma_x = \frac{1}{3} (b_3\phi_3 - \bar{\epsilon}) / (b_3\phi_2) . . . . . (\text{lx.}),$$



where

$$b_3 = \pm \sqrt{\{\phi_2(\eta^2 - r^2) - \bar{\epsilon}^2\} / (\phi_3\phi_4 - \phi_3^2)},$$

and further that there may be two points of horizontality given by a certain quadratic. Thus, in general, the regression line will tend to be part of an **S**-shaped curve. The horizontal points may be imaginary, or, if real, either they or the point of inflexion may be far beyond the portion of the curve which crosses the observed field of frequency. If we consider, however, the slope of the regression curve to measure the regression in the neighbourhood of any point, we note that the regression is a maximum at the point given by (lxi.), and grows smaller and smaller towards the two points of horizontality, *i.e.*, points of complete local independence of the two characters. These are not unfamiliar features in certain practical cases of skew correlation,\* and accordingly the cubic regression curve provides us with a ready means of describing regression phenomena, which cannot be dealt with by the simple line or the parabola.

It may of course be suggested that a quartic or quintic curve would give a better result than a cubic. The answer to this is: Possibly, but the high moments and products required render it impossible to deal even superficially with the probable errors of the constants involved. The calculation of the probable error of  $\eta$  is a sufficiently stiff task in the general case. To test the probable error of a condition like (lvii.), to say nothing of one like (lviii.), would involve an immense amount of work, since we should want the correlation of errors in  $\eta$ ,  $\bar{\epsilon}$ ,  $\zeta$ , and  $\theta$ . Speaking with some experience of practical statistical possibilities, I think, the tendency to use very high moments or product-moments must be curtailed to the minimum of actual needs. We cannot deny the existence of skew variation, nor of the sensible curvature of regression lines. We must admit their existence as the result of statistical experience. This existence involves a great widening of the old frequency notions and the need for a new means of description. But we must remember that statistics are essentially a practical study, the art of describing by a few numerical constants observational experience, and we must curtail at every turn the desire to run riot in mathematical formulæ, which cannot be generally applied in actual practice.† Still I propose later in this paper to deal with the general formulæ for quartic regression.

#### (6.) *Parabolic Regression.*

For a parabolic system  $b_3$  must vanish, or nearly vanish. Hence we have from (liii.) and (lvii.).

$$\zeta\phi_2 - \bar{\epsilon}\phi_3 = 0 \quad \dots \dots \dots \text{(lxii.)}$$

$$\phi_2(\eta^2 - r^2) - \bar{\epsilon}^2 = 0 \quad \dots \dots \dots \text{(lxiii.)}$$

\* Compare for example the regression line of age of mean age of bridegroom for actual age of bride, which gives a typical **S**-shaped curve. See 'Biometrika,' vol. II., p. 20.

† These remarks have special reference to the points dealt with on p. 6.



From these conditions we find

$$b_2 = \bar{\epsilon} / \phi_2 = \pm \sqrt{(\eta^2 - r^2) / \phi_2}.$$

These give for the form of the parabolic regression curve

$$Y_{x_p} / \sigma_y = r (X_p / \sigma_x) + \frac{\bar{\epsilon}}{\phi_2} \{ (X_p / \sigma_x)^2 - \sqrt{\beta_1} (X_p / \sigma_x) - 1 \} \quad \dots \quad (\text{lxiv.}),$$

or

$$Y_{x_p} / \sigma_y = r (Y_{p_x} / \sigma_x) \pm \sqrt{\frac{\eta^2 - r^2}{\phi_2}} \{ (X_p / \sigma_x)^2 - \sqrt{\beta_1} (X_p / \sigma_x) - 1 \} \quad \dots \quad (\text{lxv.}).$$

The latter form, besides the correlation coefficient and correlation ratio, requires only a knowledge of the skew variation constants  $\beta_1$  and  $\beta_2$ , and is therefore very easy to determine. Except for very nearly linear regression, there can be no doubt as to the sign of  $\sqrt{\eta^2 - r^2}$ , as we can tell at once whether the parabola ought to be concave or convex to the  $x$ -axis. In other cases the sign of  $\sqrt{\eta^2 - r^2}$  must be taken to coincide with that of  $\bar{\epsilon}$ , which must therefore be found. It will then be as easy to use (lxiv.) as (lxv.), although probably  $\eta$  and  $r$  can be found with less error than  $\bar{\epsilon}$ .

It is thus quite easy to allow for such curvature of the regression line as can be expressed by a parabola of the 2<sup>nd</sup> order of the type considered.

We notice at once that the regression curve does not pass through the mean of the two characters. Or, an individual with the mean of one character will most probably not have the mean of a second character. This is a rather important result, which follows at once for nearly all types of skew correlation.

It will be seen, for example, that QUETELET'S "mean man," defended by Professor EDGEWORTH as theoretically justifiable, depends entirely on human characters giving linear regression curves. Such linear curves are certainly given by many pairs of characters, *e.g.*, cranial and body measurements, but there are certainly other characters for which regression ceases to be sensibly linear, and the conception of the "mean man" in this case fails. For example, if age be considered as a character, then the regression is certainly not linear, and the individual of mean age will not necessarily have either the mean physical or psychical characters. This seems of some importance for the general conception of "type," if by type we denote the mean, for probably there are other characters than age for which regression is skew.

The regression, *i.e.*,  $dY_{x_p} / dX_p$  will be zero, for a point  $X_{(Y \text{ max.})}$  for which

$$\frac{X_{(Y \text{ max.})}}{\sigma_x} = \frac{1}{2} \left\{ \sqrt{\beta_1} - r \sqrt{\frac{\beta_2 - \beta_1 - 1}{\eta^2 - r^2}} \right\} \quad \dots \quad (\text{lxvi.})$$

the sign of the root being determined as before. Clearly, therefore, unless  $r$  be very small, or  $\eta^2$  diverges very sensibly from  $r^2$ , this point of zero regression may correspond



to a very large abscissa, and in some cases will lie entirely outside the range of observable frequency.

The parabola of regression cuts the line of regression, *i.e.*, the line of best fit to the series of regression points, or to the means of the  $x$ -arrays, in two points determined by the quadratic equation

$$\left(\frac{X_p}{\sigma_x}\right)^2 - \sqrt{\beta_1} \frac{X_p}{\sigma_x} - 1 = 0,$$

or

$$\frac{X_p}{\sigma_x} = \frac{1}{2} \{ \sqrt{\beta_1} \pm \sqrt{\beta_1 + 4} \} \quad \dots \dots \dots \quad (\text{lxvii}).$$

These points are always real, and correspond, if regression be truly parabolic, to the same values of the  $x$ -character, whatever be the  $y$ -character of which we are considering the correlation. In the case of normal variation of the  $x$ -character only; these are the points of inflexion of the  $x$ -distribution.

#### (7.) *Linear Regression.*

In this case it is necessary that both  $b_2$  and  $b_3$  vanish within the limits of random sampling, and, although these are not theoretically sufficient—for a whole series of relations between the higher product-moments could be written down\*—they are for practical purposes sufficient.

Hence we have the following conditions for linear regression:—

$$\eta^2 = r^2 \quad \dots \dots \dots \quad (\text{lxviii}).$$

or, the coefficient of correlation, without regard to sign, should be equal to the correlation ratio. Further  $\bar{\epsilon}$  should be zero, or

$$p_{21} p_{30} - p_{11} p_{30} = 0 \quad \dots \dots \dots \quad (\text{lxix}).$$

The theory of linear regression is so familiar that it need not be further discussed here. In the actual practice of statistics, the determination of the means of the  $x$ -arrays and the drawing of the regression line will often suffice to show the fairly trained eye whether the deviations from it are random or not. If they are not random, then we must proceed to the determination of  $\eta$  and of the higher product-moments.

The following are numerical examples of skew correlation, selected to illustrate the theory developed above.

\* For example, it is necessary in most cases that  $\bar{\zeta}$  should vanish. In the instance of that very special case of linear regression, the Gauss-Laplacian normal frequency, it is easy to show that the constants  $\bar{\epsilon}$ ,  $\bar{\zeta}$  both vanish as well as  $\eta^2 = r^2$ .



STATISTICAL ILLUSTRATIONS.

(8.) *Illustration A.—On the Skew Correlation between Number of Branches to the Whorl and Position of the Whorl on the Spray in the case of Asperula odorata.*

In this case the material was collected in a lane near Horsham, Sussex, at Whitsuntide, 1903, by Miss M. RADFORD. There were 150 independent sprays, the woodruff had just flowered, and the whorls were counted from the flower downwards. Being early in the season, the maximum number of whorls was five, and, in some cases, not even as many were available. The material was counted and tabled by the author, and the results are exhibited in the table below :—

TABLE I.—Correlation of Whorl-Branches and Position of Whorl.

	x.	Whorl.	Number of branches in whorl.					$n_p$ .	$y_{x_p}$ .	$\sigma_{n_p}$ .	$m_2$ .	$m_3$ .
			4.	5.	6.	7.	8.					
Position of whorl.	$x_1$	First . .	—	3	66	42	39	150	6·7800	·8553	·7316	·1535
	$x_2$	Second . .	—	3	61	47	39	150	6·8133	·8437	·7117	·0985
	$x_3$	Third . .	—	6	60	40	44	150	6·8133	·9047	·8185	·0383
	$x_4$	Fourth . .	1	12	68	39	22	142	6·4859	·8780	·7709	·1347
	$x_5$	Fifth . .	1	13	53	10	10	87	6·1724	·8605	·7404	·4049
Totals . . . .			2	37	308	178	154	679	6·6554	—	—	—

We require the regression curve giving the probable number of branches for a given whorl.

Dealing first with the skew variation in position, a purely arbitrary system depending solely on the number of whorls dealt with in each position, we find, not using SHEPPARD'S correction,\*

$$\begin{aligned} \text{Mean} &= 2\cdot802,651, & \nu_2 &= 1\cdot787,268, & \nu_5 &= 2\cdot799,638, \\ \sigma_x &= 1\cdot336,887, & \nu_3 &= \cdot311,783, & \nu_6 &= 22\cdot678,308. \\ & & \nu_4 &= 5\cdot841,682. \end{aligned}$$

Hence we determine

$$\begin{aligned} \beta_1 &= \cdot017,027, & \phi_2 &= \cdot811,740, \\ \beta_2 &= 1\cdot828,767, & \phi_3 &= \cdot286,465. \\ \beta_3 &= \cdot085,545, & \phi_4 &= \cdot610,879, \\ \beta_4 &= 3\cdot972,295, & \text{and } \sqrt{\beta_1} &= +\cdot130,487. \end{aligned}$$

\* The numbers are tabulated to six places, because we cannot be sure that the final calculations are for the data true to two places, which is all we finally retain unless this is done. Any number of figures can really be retained with perfect ease when the work is done on a calculator.



We now turn to the skew variation in the number of branches to the whorl, and get the following constants:—

$$\begin{aligned} \text{Mean} &= 6.655,375, & \mu_2 &= .806,124, \\ \sigma_y &= .897,842, & \mu_3 &= .132,090, \\ & & \mu_4 &= 1.138,410. \end{aligned}$$

The values of  $y_2$ ,  $m_2$ , and  $m_3$  are given in table above. Using them we find

$$\begin{aligned} \sigma_M &= .224,377, & \eta &= .249,911, & \sigma_{\sigma_y} &= \sigma_y \sqrt{1-\eta^2} = .869,355, \\ \lambda_2 &= \sigma_M^2 = .050,345, & \lambda_4 &= .007,474, & \chi_1 &= .990,862, \chi_2 = -.059,851. \end{aligned}$$

These give by (xxxiii.), showing the numerical contribution of each term,

$$\Sigma_s^2 = \frac{1}{N} \{ .878,991 - .010,323 - .000,888 - .007,231 + .013,578 \},$$

or the probable error of  $\eta = .0242$ .

Had we calculated the probable error of  $\eta$  from (xxxiv.), we should have found for its value .0243. It is clear that for this special case the simple formula (xxxiv.) is amply sufficient, the small terms almost cancelling.

We see that  $\chi_1$  is almost unity, and the graph of  $\sigma_{\sigma_y}/\sigma_y$  shows indeed that the system is sensibly homoscedastic.  $\chi_2$  is small, but a glance at the graph of the clitic curve on Diagram I. shows that we can hardly treat the system as homoclitic, the changes in the skewness forming a fairly uniform curve.\*

For practical purposes, we may treat the variability of the number of branches in any array as sufficiently closely given by  $\sigma_y \sqrt{1-\eta^2}$ .

We now turn to the product-moments† and find

$$\begin{aligned} p_{11} &= -.249,160, & p_{31} &= -.896,415, \\ p_{21} &= -.236,289, & p_{41} &= -1.210,225. \end{aligned}$$

\* Throughout these illustrations the clitic curve is plotted by calculating the skewness of the arrays from  $\frac{1}{2}m_3/(m_2)^{3/2}$ . See p. 23.

† In calculating these products referred to the centroid from those referred to any axes, generally corresponding to whole numbers in the table, the following reduction formulæ will be found useful. We take  $N\Pi_{xy} = S(n_{xy} x'y')$ ,  $x'$  and  $y'$  being measured from any axes, further,  $\bar{x}$ ,  $\bar{y}$  are the distances of the means from these axes, and  $v_2, v_3, v_4$  the moments of the  $x$ -character about its mean as tabled above.

$$\begin{aligned} p_{11} &= \Pi_{11} - \bar{x}'\Pi_{01}, & p_{21} &= \Pi_{21} - 2\bar{x}'\Pi_{11} + \bar{x}'^2\Pi_{01} - \bar{y}'v_2, \\ p_{31} &= \Pi_{31} - 3\bar{x}'\Pi_{21} + 3\bar{x}'^2\Pi_{11} - \bar{x}'^3\Pi_{01} - \bar{y}'v_3, \\ p_{41} &= \Pi_{41} - 4\bar{x}'\Pi_{31} + 6\bar{x}'^2\Pi_{21} + 4\bar{x}'^3\Pi_{11} + \bar{x}'^4\Pi_{01} - \bar{y}'v_4. \end{aligned}$$

The  $p$ 's should be further corrected for grouping by SHEPPARD'S corrections (given on my p. 36), provided there be high contact at the contour of the surface of frequency. SHEPPARD'S corrections have not in this



These lead to

$$r = -0.207,579, \quad \bar{\epsilon} = -0.120,164, \quad \bar{\zeta} = -0.038,241, \quad \bar{\theta} = -0.285,890.$$

Thus all the constants are determined.

We find

$$\begin{aligned} \eta^2 - r^2 &= 0.019,367, \\ \phi_2 (\eta^2 - r^2) - \bar{\epsilon}^2 &= 0.001,281, \\ \phi_2 (\eta^2 - r^2) - \bar{\epsilon}^2 - (\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2 / (\phi_2\phi_4 - \phi_3^2) &= 0.000,276. \end{aligned}$$

These should be respectively zero for linear, parabolic, and cubical regressions. It will be seen that they are satisfied with increasing closeness; we might well be satisfied even with the parabolic regression curve. The following are the regression curves determined,  $y_x$  being the actual number of branches in the whorl ( $= 6.655,375 + Y_x$ ), and  $x_p$  the actual position of the whorl:—

(a.) *Straight line* :

$$y_x = 7.046,087 - 0.139,408 x_p.$$

(b.) *Parabola* from (lxv.) :

$$y_x = 6.794,052 - 0.125,872 x_p - 0.077,592 x_p^2;$$

or,

$$y_x = 6.853,561 - 0.077,592 (x_p - 1.991,535)^2.$$

This clearly gives a maximum number of branches, 6.8536 corresponding to  $x_p = 1.9915$ , a value within the limits of observation.

(c.) *Cubic* from (lix.) :

$$y_x = 6.799,399 - 0.192,439 X_p - 0.084,230 X_p^2 + 0.020,915 X_p^3.$$

Here  $X_p$  is measured from the mean position  $= x_p - 2.802,651$ , and  $y_x$  is, as before, the total number of branches for the given position.

Condition (lvii.) is so closely satisfied that we shall here get sensibly as good results from (lix.) as from (lvi.).

In the table below and in the curves of Diagram I. the values of the mean of the arrays, as found from line, parabola, and cubic, are given and compared with observation.

case been used, as this condition is not fulfilled. The axes  $x', y'$  actually taken for woodruff were those through the third whorl and through six branches.

An obvious warning about the signs of the sums of the products may be given which may save computators some trouble. The axes being taken positive, as in the accompanying figure, then the sums of the products for  $\Pi_{11}$  and  $\Pi_{31}$  are positive in the 1<sup>st</sup> and 3<sup>rd</sup>, negative in the 2<sup>nd</sup> and 4<sup>th</sup> quadrants. For  $\Pi_{21}$  and  $\Pi_{41}$  they are positive in the 1<sup>st</sup> and 4<sup>th</sup> quadrants and negative in the 2<sup>nd</sup> and 3<sup>rd</sup> quadrants. In the figure the axes are taken so as to suit the  $x$  and  $y$ -directions of the table on p. 31. Care must, of course, be paid to this point. The products may also be found from the  $y_x$ 's in the manner indicated on p. 35, footnote. They were thus verified in this case.

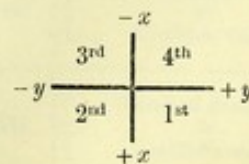




TABLE II.—Mean Branches to each Whorl.

$z_p =$	0.	1.	2.	3.	4.	5.	6.
$y_{z_p}$ from line . . . .	[7.046]	6.907	6.767	6.628	6.488	6.349	6.210
$y_{z_p}$ " parabola . . . .	[6.546]	6.777	6.854	6.775	6.541	6.151	5.607
$y_{z_p}$ " cubic . . . .	[6.117]	6.750	6.889	6.758	6.443	6.192	6.007
Observed . . . .	—	6.780	6.813	6.813	6.486	6.172	†

I think we may safely say that in the relationship of branches to position of the whorl in woodruff we have a case of homoscedastic correlation, which is effectively described by a parabolic regression curve. Thus, in a case of this kind, it is only needful, besides the moments up to the fourth of the  $x$ -character, to find the correlation coefficient  $r$  and the correlation ratio  $\eta$ .

(9.) *Illustration B.—On the Correlation between Age and Head Height in Girls.*

The data for this are taken from my School Measurement series, and involve the auricular heights of 2272 girls between the ages of 3 and 22. There was considerable paucity of material at the extreme ends of the range, and accordingly as our correlation curves are all obtained by weighting the observations, we can hardly expect good fits near 3 or 22 years of age. The actual correlation table is given as Table III. SHEPPARD'S corrections were applied throughout, and the unit of height is 2 millims.

In the first place the means, standard deviations, and 3<sup>rd</sup> moments of all the arrays of heights for different years of age were determined. These are given at the foot of Table III., but in actually calculating the constants more places of decimals were used. Then the first six moments of the frequency of the ages were found and the first four moments of the height frequencies. These are the  $x$  and  $y$ -frequencies. They give us:—











*Height Constants.*

*Age Constants.*

Mean height = 124.0467 millims.

Mean age = 12.7007

$$\left. \begin{aligned} \sigma_y &= 3.454,125 \\ \mu_2 &= 11.930,977 \\ \mu_3 &= 5.206,247 \\ \mu_4 &= 438.639,633 \end{aligned} \right\} \begin{array}{l} \text{in} \\ 2 \text{ millim.} \\ \text{units.} \end{array}$$

$$\left. \begin{aligned} \sigma_x &= 3.064,819 \\ \nu_2 &= 9.393,110 \\ \nu_3 &= 1.051,882 \\ \nu_4 &= 239.157,055 \\ \nu_5 &= 104.298,702 \\ \nu_6 &= 9536.265,059 \end{aligned} \right\} \begin{array}{l} \text{in} \\ \text{year} \\ \text{units.} \end{array}$$

$$\begin{aligned} \beta'_1 &= .015,960, \\ \beta'_2 &= 3.081,454, \end{aligned}$$

$$\begin{aligned} \beta_1 &= .001,335, \\ \beta_2 &= 2.710,593, \\ \beta_3 &= .014,093, \\ \beta_4 &= 11.506,681, \end{aligned}$$

Further

$$\begin{aligned} \Sigma_M &= 2.093,366 \text{ millims.} \\ \lambda_2 &= 4.382,181 \\ \lambda_4 &= 62.399,135 \end{aligned} \left\} \begin{array}{l} \text{in 1 millim.} \\ \text{units.} \end{array}$$

$$\begin{aligned} \sqrt{\beta_1} &= + .036,538, \\ \phi_2 &= 1.709,258, \\ \phi_3 &= .250,123. \end{aligned}$$

Hence

$$(\lambda_4 - 3\lambda_2^2)/(4\lambda_2^3) = .062,340,$$

$$\phi_4 = 4.158,032.$$

In the next place the products were worked out and referred to the means with the following results:—\*

$$\begin{aligned} p_{11} &= 3.113,712, \\ p_{21} &= - 1.957,022, \\ p_{31} &= 74.447,616, \\ p_{41} &= -108.701,559, \end{aligned}$$

$$\begin{aligned} \text{whence } r &= .294,128, \\ \bar{\epsilon} &= -.071,065, \\ \bar{\zeta} &= -.048,576, \\ \bar{\theta} &= -.470,126. \end{aligned}$$

Further, from  $\Sigma_M$ ,  $\eta = .303,024$ .

In deducing the product-moments *after they had been referred to the means*, the

\* These products were in this case (as in all other cases) verified by calculating from the means of the arrays  $y_{x_p}$ , the expressions

$$S \left\{ \frac{n_x y_{x_p} (x_p - \bar{x})}{N} \right\}, \quad S \left\{ \frac{n_x y_{x_p} (x_p - \bar{x})^2}{N} \right\}, \quad S \left\{ \frac{n_x y_{x_p} (x_p - \bar{x})^3}{N} \right\}, \quad S \left\{ \frac{n_x y_{x_p} (x_p - \bar{x})^4}{N} \right\}.$$

Of course it is easiest to calculate these products about some arbitrary origin coinciding with the abscissa of one array. If these products be then  $p'_{11}$ ,  $p'_{21}$ ,  $p'_{31}$ ,  $p'_{41}$ , and  $\bar{x}'$  be the mean, we have

$$\begin{aligned} p_{11} &= p'_{11}, \\ p_{21} &= p'_{21} - 2\bar{x}'p'_{11}, \\ p_{31} &= p'_{31} - 3\bar{x}'p'_{21} + 3\bar{x}'^2p'_{11}, \\ p_{41} &= p'_{41} - 4\bar{x}'p'_{31} + 6\bar{x}'^2p'_{21} - 4\bar{x}'^3p'_{11}, \dots \end{aligned}$$



proper SHEPPARD'S corrections were introduced. These are, if  $\{p_{11}\}$ ,  $\{p_{21}\}$ ,  $\{p_{31}\}$ ,  $\{p_{41}\}$  represent the uncorrected moments:---

$$\begin{aligned} p_{11} &= \{p_{11}\}, & p_{21} &= \{p_{21}\}, \\ p_{31} &= \{p_{31}\} - \frac{1}{4}\{p_{11}\}, & p_{41} &= \{p_{41}\} - \frac{1}{2}\{p_{21}\}, \end{aligned}$$

the units of grouping being the units throughout.

From the constants for the arrays, I found

$$\chi_1 - 1 = -\cdot000,675, \quad \chi_2 = -\cdot007,198.$$

Whence the probable error of  $\eta$  was determined by (xxxiii.). Its value was\*

$$\text{Probable error of } \eta = \cdot012,913.$$

If found from the simple formula  $\cdot67449(1-\eta^2)/N$ , the value is  $\cdot012,851$ . We accordingly are again forced to the conclusion that  $\eta$  may for practical purposes be found from this simple formula, instead of the complicated result (xxxiii.). Although both  $\chi_1 - 1$  and  $\chi_2$  are small, it is very doubtful whether we can legitimately consider the system as homoscedastic. The dotted line *ab* of Diagram II. would fairly well represent increasing variability with age. The skewness of the arrays is relatively small and changes sign so frequently, that we can certainly not attribute any law to such heteroclitic tendencies as there are. They are probably due to errors of random sampling from truly isocurtic material.

It will be seen that the height frequencies with  $\beta'_1 = \cdot0160$  and  $\beta'_2 = 3\cdot0815$  do not differ very much from a normal distribution; in fact, we can lay no stress on the heteroclysis of the system at all. But the values of the standard deviations of the arrays, or the graph of  $\sigma_x/\sigma_y$ , certainly shows increasing variation with increasing age, a phenomenon with which one is familiar in a variety of other human characters.†

This heteroscedasticity, due to increasing variation with growth, would hardly have been anticipated from a mere inspection of the smallness of  $\chi_1$ ; it is somewhat obscured by the irregular values of the standard deviations of the small arrays at the adult end of the age range. The mean value of the standard deviation of the weighted arrays is  $\sigma_y \sqrt{1-\eta^2} = 3\cdot2992$  in 2-millim. units.

We now turn to the regression curves to see how far the conditions for the different types are satisfied. We have

$$\begin{aligned} \eta^2 - r^2 &= \cdot005,312, \\ \phi_2 (\eta^2 - r^2) - \bar{\epsilon}^2 &= \cdot004,030, \\ \phi_2 (\eta^2 - r^2) - \bar{\epsilon}^2 - (\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2 / (\phi_2\phi_4 - \phi_3^2) &= \cdot000,604. \end{aligned}$$

\* The contributions of the successive terms of (xxxiii.) are in fact given by

$$\Sigma_4^2 = \frac{1}{N} \{ \cdot824,785 + \cdot001,870 + \cdot004,673 - \cdot000,472 + \cdot001,888 \}.$$

† See PEARSON: 'The Chances of Death and other Studies of Evolution,' vol. I., pp. 296, 307, 310, 314.



But the first should be zero, if the regression be linear; the second, if it be parabolic; and the third, if it be cubical.

We see increasing approximation to fulfilment of the several conditions. Referred to axes through the mean age and head height, the following are the regression curves\* :—

(a.) *Straight line* :

$$Y_{x_p} = \cdot662,979 X_p.$$

(b.) *Parabola* (from equation (lxv.)) :

$$Y_{x_p} = \cdot055,749 + \cdot667,570 X_p - \cdot041,001 X_p^2.$$

(c.) *Cubic* (from equation (lvi.)) :

$$Y_{x_p} = \cdot280,194 + \cdot722,886 X_p - \cdot029,580 X_p^2 - \cdot002,223 X_p^3.$$

(c'.) *Cubic* (from equation (lix.)) :

$$Y_{x_p} = \cdot296,076 + \cdot812,249 X_p - \cdot028,004 X_p^2 - \cdot005,740 X_p^3.$$

(c') will not give as good results as (c), for it depends on a use of the condition (lvii.) which is not absolutely fulfilled.

The following table gives the values in the case of the four curves :—

TABLE IV.— $y_{x_p}$  = Mean Auricular Height of Girl's Head at Given Age.

$x_p$ = age.	Regression line.	Regression parabola.†	Cubic (c).	Cubic (c').	Observed.
3·5	117·95	114·49	116·90	118·94	115·25
4·5	118·61	115·87	117·66	118·94	116·96
5·5	119·27	117·17	118·42	119·16	117·47
6·5	119·94	118·39	119·24	119·57	119·10
7·5	120·60	119·52	120·08	120·14	120·30
8·5	121·26	120·57	120·93	120·84	121·63
9·5	121·92	121·55	121·78	121·62	121·72
10·5	122·59	122·43	122·62	122·45	122·82
11·5	123·25	123·24	123·42	123·26	123·14
12·5	123·91	123·97	124·18	124·15	123·89
13·5	124·58	124·61	124·88	124·95	124·86
14·5	125·24	125·17	125·52	125·65	125·71
15·5	125·90	125·65	126·07	126·22	126·16
16·5	126·57	126·05	126·52	126·68	126·53
17·5	127·23	126·36	126·87	126·93	126·91
18·5	127·89	126·59	127·09	126·96	127·02
19·5	128·55	126·75	127·18	126·74	129·56
20·5	129·22	126·81	127·11	126·22	123·82
21·5	129·88	126·80	126·88	125·38	126·50
22·5	130·54	126·71	126·48	124·28	125·25

\*  $Y_{x_p}$  is here measured in millimetres and  $X_p$  in years.

† The maximum ordinate is at vertex of parabola, i.e.,  $x = 8·1409$ , or age 20·84; its magnitude = 126·82.



An examination of this table and the graphs on Diagram II. seem to show :—

- (i.) That cubic ( $c$ ) is considerably better than cubic ( $c'$ ).
- (ii.) That we do get a sensible betterment in passing from parabola to cubic, and, accordingly, that we must use in this the cubic to effectively describe the regression within the range of observation. Probably neither cubic nor parabola would effectively serve for extrapolation even close to the limits of observation.

Thus the cubic ( $c'$ ) starting at 3-4 with its point of inflection is clearly inadmissible, and the drop after 20 or 21 years of age, shown by both parabola and cubic, is, of course, only due to the anomalous character of the few girls over 18 left in the schools. Actually the shrinkage of measurements does not begin till at least 26 years, and is then far more gradual than these curves indicate.

But, as in all fitting of this kind, we obtain the best fit we can within the range, entirely at the expense of what may occur just outside the range. For this reason, as E. PERRIN\* has pointed out, a good interpolation curve is usually a bad extrapolation curve.

We might sum up our results for auricular height with age in girls by saying : That the correlation is non-linear, effectively cubic ; heteroscedastic, there being increasing variability with growth ; that while the total height frequency is not very far from normal the array frequencies are slightly heteroclitic, but so very irregular in sign, that probably we are dealing with a case of isocurtic homoclysis, to which the sparsity of data in the extreme arrays gives an appearance of anomic heteroclysis.

(10.) *Illustration C.—On the Skew Correlation between Size of Cell and Size of Body in Daphnia magna.*

Dr. E. WARREN has dealt with this point in a memoir published in 'Biometrika,' vol. II., pp. 255-9. The resulting regression curve of size of cell for given size of body is very far from linear, and it is quite clear that the correlation is skew. It has already been noted in 'Biometrika' that the relationship is considerably obscured by the irregularities produced by ecdysis. Our object at present, however, is purely theoretical, namely, to show how a certain system of constants and of curves describes the actual relationship, and for this purpose Dr. WARREN's observations form as good material for graduation as we could expect to find. The following Table V. gives the observations with the working scales attached. I must refer to Dr. WARREN's paper (p. 256) for the relation between the units of grouping on the working scales and those of the actual measurements on body and cell lengths. As far as correcting the raw moments is concerned, SHEPPARD's corrections were used for the cell sizes, but not for the body lengths, because the number of individuals in the latter case was perfectly arbitrary and there is no approach to high contact. The

\* 'Biometrika,' vol. III., p. 99.



product moments were also uncorrected. The product moments were found in both ways (see p. 35, footnote) and the results thus verified.

Table V. gives the means, standard deviations, and third moments of the arrays; the latter are all small and superficially irregular in sign. I think we may say that there is no marked and continuous heteroclisys. On the other hand, I think we may say that while the clitic curve deviates to and fro from a zero base, the scedastic curve would fit better to a parabolic curve than to the straight line which is its mean. In other words, the variability of the cells increases with size of body (*i.e.*, growth) up to a certain stage and then decreases again. This result is obscured by the fall of the variability after each ecdysis. Roughly the ecdyses produce a rhythm in all three curves, the regression curve, the scedastic curve, and the clitic curve. When the means of the arrays are above the regression cubic, then the ordinates of the scedastic curve are above their mean and those of the clitic curve show positive skewness; when they are below the regression curve, we have lessened variability and negative skewness. In other words, the ecdyses are accompanied by lessened cell variability and negative skewness of distribution. I think we may state that there is a nomic heteroscedasticity due to growth of body, giving first an increased variability with growth and afterwards a decrease with age. There is probably isocurtic homoclisys. Both of these are, however, obscured by a semi-rhythmic heteroscedasticity and heteroclisys introduced by the ecdyses.

We now turn to the constants of the cell and body length distributions, merely noting that all these constants are given in terms of the units of the working scales.

*Cell Constants.*

*Body Length Constants.*

Mean cell =	9.268,657,	Mean body length =	8.502,488,
$\sigma_y =$	2.541,734,	$\sigma_x =$	3.864,784,
$\mu_2 =$	6.460,410,	$\nu_2 =$	14.936,562,
$\mu_3 =$	2.142,362,	$\nu_3 =$	- 5.125,806,
$\mu_4 =$	123.921,496,	$\nu_4 =$	432.769,533,

$\nu_5 =$	- 425.276,682,
$\nu_6 =$	15192.5375,

$\beta_1' =$	.017,021,	$\beta_1 =$	.007,885,
$\beta_2' =$	2.969,111.	$\beta_2 =$	1.939,793,

Further

$\beta_3 =$	.043,796,	$\beta_4 =$	4.559,091,
-------------	-----------	-------------	------------

$\Sigma_M =$	1.454,600,	$\sqrt{\beta_1} =$	- .088,798,
--------------	------------	--------------------	-------------

$\lambda_2 =$	2.115,862,	$\phi_2 =$	.931,908,
---------------	------------	------------	-----------

$\lambda_4 =$	15.142,840.	$\phi_3 =$	- .232,167,
---------------	-------------	------------	-------------

Hence $(\lambda_4 - 3\lambda_2^2)/(4\lambda_2^2) =$	.095,615.	$\phi_4 =$	.788,409.
---	-----------	------------	-----------







We have next the product moments referred to the means

$$\begin{array}{ll} p_{11} = 3\cdot892,863, & \text{whence } r = \cdot394,862, \\ p_{21} = -12\cdot104,322, & \bar{\epsilon} = -\cdot281,831, \\ p_{31} = 127\cdot348,064, & \bar{\zeta} = \cdot098,578, \\ p_{41} = -541\cdot433,455, & \bar{\theta} = -\cdot759,344. \end{array}$$

Further, from  $\Sigma_M$ ,

$$\eta = \cdot572,287.$$

From the constants for the arrays I deduced

$$\chi_1 - 1 = -\cdot108,148, \quad \chi_2 = \cdot088,323.$$

These are higher values of  $\chi_1 - 1$  and  $\chi_2$  than we have found in the first two illustrations.

We now obtain, showing the contribution of each term of (xxxiii.),

$$\Sigma_v^2 = \frac{1}{N} \{ \cdot452,240 - \cdot002,528 + \cdot010,803 - \cdot013,180 - \cdot027,875 \}.$$

Whence probable error of  $\eta = \cdot67449 \Sigma_v = \cdot0097$ .

Had we calculated the probable error of  $\eta$  from (xxxiv.), we should have found it equal to  $\cdot0101$ . The difference is greater than in the two previous illustrations, but is only  $\cdot0004$ , and this would have no significance in any practical use of the probable error. We again conclude, therefore, that (xxxiv.) is sufficiently close to replace (xxxiii.) in practice.

For the mean standard deviation of the weighted arrays we have

$$\sigma_a = \sigma_y \sqrt{1 - \eta^2} = 2\cdot084,358.$$

If we now examine the criteria for the nature of the regression, we have

$$\begin{aligned} \eta^2 - r^2 &= \cdot171,596, \\ \phi_2 (\eta^2 - r^2) - \bar{\epsilon}^2 &= \cdot080,483, \\ \phi_2 (\eta^2 - r^2) - \bar{\epsilon}^2 - (\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2 / (\phi_2\phi_4 - \phi_3^2) &= \cdot079,457. \end{aligned}$$

We should conclude, therefore, that linear regression is inadmissible, but that parabolic or cubic will be moderately successful, the latter not very much better than the former. Our moderate success only in this case is, of course, due to the irregularity of the results to be graduated, the influence of the ecdyses being so disturbing that we really need a curve periodically varying from the graduated regression curve.

We have the following regression curves:—

(a.) *Straight line*:

$$Y_x = \cdot259,687 X_p$$



(b.) *Parabola* from (lxv.):

$$Y_{x_p} = 1.097,690 + .236,135 X_p - .073,490 X_p^2.$$

The maximum occurs when  $X_p = 1.6066$ , and is given by  $Y_{x_p} = 1.2874$ , thus occurring within the limits of observation.\*

(c.) *Cubic* from (lix.):

$$Y_{x_p} = .752,856 + .193,058 X_p - .049,817 X_p^2 + .001,710 X_p^3.$$

In all these cases  $Y_{x_p}$  and  $X_p$  are measured from the means of the cell and body lengths, or from 9.268,657 and 8.502,488 respectively.

Table VI. gives the calculated and observed results, and the whole system is represented in Diagram III. Either the parabola or cubic graduates quite well the results, allowing for the periodic deviation, and we may fairly describe the system as a heteroscedastic cubic regression with isocurtic homoclysis. The correlation ratio is very sensibly different from the correlation coefficient. The regression cubic does not differ widely from that given in 'Biometrika,' which was obtained without weighting the means of the arrays, and by simply striking the best cubic of the given type through the points.

TABLE VI.— $y_{x_p}$  = Mean Cell Length for Given Body Length in *Daphnia*.

$x_p$ = body length.	Regression line.	Regression parabola.	Regression cubic.	Observed.
1	7.320	4.458	5.047	5.300
2	7.580	5.724	6.190	5.833
3	7.840	6.842	7.166	7.790
4	8.099	7.813	7.986	8.050
5	8.359	8.638	8.661	9.473
6	8.619	9.315	9.200	8.436
7	8.879	9.846	9.613	8.596
8	9.138	10.229	9.912	10.267
9	9.398	10.466	10.105	10.761
10	9.658	10.555	10.205	11.027
11	9.917	10.498	10.220	10.953
12	10.177	10.293	10.161	9.100
13	10.437	9.942	10.038	9.000
14	10.696	9.443	9.861	10.036
15	10.956	8.798	9.642	10.317

(11.) *Illustration D.*—On the Skew Correlation between Number of Branches to the Whorl and Position of the Whorl on the Stem in *Equisetum arvense*.

I have selected this example not on account of any biological importance, because the material is—especially with regard to the first and last two whorls—unsatisfactory either on account of irregularity or of insufficiency of material. It has been taken

\* Actual values on working scales,  $x_p = 10.1091$  and  $y_{x_p} = 10.5560$ .



purely from its statistical interest, because it gives a series with markedly skew correlation, having a regression curve of a rough **S**-shaped character. If we omit the first and last whorls, we get, as I have already shown,\* a remarkably close fit with a cubical regression curve. My present object, however, is not to consider any law of growth, but merely a mass of statistical material, to be dealt with by the processes of the present paper.

We may anticipate that the irregularities of the series, indicated in the memoir just referred to, will make themselves manifest in a less satisfactory fitting of the regression curve than occurs when we deal with the more homogeneous group of equally weighted whorls fitted in the diagram of that paper. Table VII. gives the data, with the means, standard deviations, and third moments of each array.

The axis of  $x$  shall be taken to give the position of the whorl on the stem and that of  $y$  to denote the number of branches. We require the regression curve of  $y$  on  $x$ , or the probable number of branches on a whorl in a given position. We shall not use SHEPPARD'S corrections for the moments of either the  $x$  or  $y$ -characters, as high contact certainly does not hold for both at the low-value ends of their ranges.

We have the following constants:—

*Position Constants.*

*Branch Constants.*

Mean position =	6·403,315,	Mean number of branches =	7·216,851,
$\sigma_x =$	3·542,604,	$\sigma_y =$	3·278,499,
$\nu_2 =$	12·550,046,	$\mu_2 =$	10·748,557,
$\nu_3 =$	8·249,534,	$\mu_3 = -$	24·313,478,
$\nu_4 =$	319·515,824,	$\mu_4 =$	245·811,660,
$\nu_5 =$	644·095,176,		
$\nu_6 =$	11203·5814,		
$\beta_1 =$	·034,429,	$\beta'_1 =$	·476,044,
$\beta_2 =$	2·028,625,	$\beta'_2 =$	2·127,658.
$\beta_3 =$	·214,190,	Further	
$\beta_4 =$	5·667,884,	$\Sigma_{31} =$	2·789,949,
$\sqrt{\beta_1} =$	·185,550,	$\lambda_2 =$	7·783,815,
$\phi_2 =$	·994,196,	$\lambda_4 =$	140·441,685.
$\phi_3 =$	·592,384,	Hence	
$\phi_4 =$	1·518,136.	$(\lambda_4 - 3\lambda_2^2)/(4\lambda_2^2) =$	-·170,503.

We have next the product moments referred to the means

\* 'Proc. Roy. Soc.,' vol. 71, p. 308.



TABLE VII.

Position of Whorl.	Number of branches to the whorl.													Totals.	Mean.	Standard deviation.	Third moment.
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.				
1	2	2	3	10	9	8	13	29	22	17	11	—	—	126	7.619	2.360	- 9.437
2	—	—	—	—	1	3	5	21	37	40	16	3	—	126	9.294	1.273	- 1.111
3	—	—	—	—	—	—	9	9	35	45	23	5	—	126	9.627	1.187	- 0.650
4	—	—	—	—	—	—	6	10	33	45	28	3	1	126	9.730	1.151	- 0.464
5	—	—	—	—	—	—	8	10	35	41	30	2	—	126	9.643	1.158	- 0.780
6	—	—	—	—	2	3	6	13	35	38	24	3	—	124	9.427	1.375	- 2.171
7	—	—	1	4	2	6	12	23	28	29	17	1	—	123	8.732	1.781	- 5.013
8	—	3	7	5	5	13	21	33	24	14	4	—	—	121	7.297	2.291	- 9.727
9	8	10	13	14	9	14	19	17	10	5	—	—	—	119	5.555	2.553	- 2.693
10	18	20	13	11	17	14	11	5	1	—	—	—	—	110	3.964	2.199	+ 2.455
11	31	29	18	9	5	3	1	1	—	—	—	—	—	97	2.443	1.506	+ 4.392
12	24	34	6	2	—	—	—	—	—	—	—	—	—	67	1.866	0.960	+ 2.210
13	24	14	—	—	1	—	—	—	—	—	—	—	—	39	1.462	0.746	+ 1.132
14	8	4	—	—	—	—	—	—	—	—	—	—	—	12	1.333	0.471	+ 0.707
15	3	1	—	—	—	—	—	—	—	—	—	—	—	4	1.250	0.433	+ 0.094
16	2	—	—	—	—	—	—	—	—	—	—	—	—	2	1.000	0.000	+ 0.000
Totals . .	122	117	61	55	51	64	112	161	260	274	153	17	1	1448	7.217	3.278	- 24.313



$$\begin{array}{ll}
 p_{11} = -8.225,585, & \text{whence } r = -0.708,222, \\
 p_{21} = -21.471,321, & \bar{\epsilon} = -0.390,436, \\
 p_{31} = -205.084,042, & \bar{\zeta} = +0.029,733, \\
 p_{41} = -917.984,938, & \bar{\theta} = -0.960,212.
 \end{array}$$

Further, from  $\Sigma_M$ ,

$$\eta = 0.850,984.$$

From the constants for the arrays we deduce

$$\chi_1 - 1 = -0.356,367, \quad \chi_2 = -0.312,952.$$

We now obtain, showing the contribution of each term of (xxxiii.),

$$\Sigma_n^2 = \frac{1}{N} \{0.076,080 - 0.157,932 + 0.055,359 + 0.079,662 + 0.038,579\}.$$

Whence probable error of  $\eta = 0.67449 \Sigma_n = 0.0054$ .

Had we calculated the probable error of  $\eta$  from (xxxiv.) we should have found it equal to 0.0049. The difference 0.0005 is not of importance for practical purposes. Yet in this case it is clear that the values of  $\chi_1 - 1$  and  $\chi_2$  are very sensible. Thus we see that a very marked heteroscedastic and heteroclitic system with continuously changing standard deviation and skewness scarcely affects for practical purposes (*i.e.*, to three significant figures) the probable error of  $\eta$ . All four of our illustrations therefore confirm the conclusion that:

*For practical purposes the probable error of the correlation ratio,  $\eta$ , may be taken as  $0.67449 (1 - \eta^2)/N$ .*

Our Diagram IV. gives the values of the relative standard deviations of the arrays, or,  $\sigma_{x_i}/\sigma_y$ , the horizontal line giving  $\sqrt{1 - \eta^2} = 0.5252$ , or the mean value of the relative standard deviations of the weighted arrays. We have also the clitic curve giving  $\frac{1}{2}\sqrt{\beta_1}$ , for each array.\* The remarkable smoothness of these scedastic and clitic curves in this case indicates how far certain types of correlation surfaces diverge from pure normality of distribution, the divergence being obviously nomic.

We now turn to the regression curves and write down the conditions for the different types; the three expressions should be zero for linear, parabolic, and cubical regression respectively

$$\begin{aligned}
 \eta^2 - r^2 &= 0.222,596, \\
 \phi_2 (\eta^2 - r^2) - \bar{\epsilon}^2 &= 0.068,864, \\
 \phi_2 (\eta^2 - r^2) - \bar{\epsilon}^2 - (\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2 / (\phi_2\phi_4 - \phi_3^2) &= 0.010,127.
 \end{aligned}$$

\*  $\frac{1}{2}\sqrt{\beta_1}$  = difference between mode and mean divided by standard deviation = skewness in the case of skew-curves of Type III. ('Phil. Trans.,' A, vol. 186, p. 373), and may be taken as a reasonable measure of the skewness for those cases in which the fuller form involving  $\beta_2$  would involve too laborious calculations. If in equation (xii.) of the present memoir we put  $\beta_2 = 3 +$  a small quantity, and remember that  $\beta_1$  is itself a small quantity, we see that the more correct formula for the skewness involving  $\beta_2$  reduces, neglecting terms of 2<sup>nd</sup> order, to  $\frac{1}{2}\sqrt{\beta_1}$ .



We see at once that the straight line is inadmissible, the parabola will not be very good, and the cubic only moderately appropriate. The conditions are not nearly so closely fulfilled as in the cases of woodruff and head heights; the last two are better than in the case of *Daphnia* cells, but while the deviations in the case of *Daphnia* were irregular, there being no approximate smoothness in the scedastic or clitic curves, we shall find here more uniform deviations which would probably be partially allowed for by a quartic regression curve.

The following are the regression curves:—

(a.) *Straight line* :

$$Y_{x_p} = -\cdot655,423 X_p$$

(b.) *Parabola* from (lxv.) :

$$Y_{x_p} = 1\cdot551,307 - \cdot574,171 X_p - \cdot123,610 X_p^2$$

The maximum ordinate is at the position  $X_p = -2\cdot3225$ , or  $x_p = 4\cdot0808$ , with maximum number of branches  $y_p = 9\cdot435$ .

(c.) *Cubic* from (lvi.) :

$$Y_{x_p} = 1\cdot590,413 - \cdot987,694 X_p - \cdot137,641 X_p^2 + \cdot016,605 X_p^3$$

In all cases  $X_p$  and  $Y_{x_p}$  are measured from the mean position and the mean number of branches, *i.e.*,  $6\cdot403,315$  and  $7\cdot216,851$  respectively.

The following table contains the calculated and observed results:—

TABLE VIII.—Mean Number of Branches to each Whorl in *Equisetum*.

Position.	Regression line.	Regression parabola.	Regression cubic.	Observed.	Regression cubic without first whorl.
1	10·758	8·262	7·506	7·619	[8·207]
2	10·103	8·900	9·070	9·294	8·929
3	9·447	9·291	9·920	9·627	9·869
4	8·792	9·434	10·156	9·730	10·161
5	8·137	9·330	9·876	9·643	9·911
6	7·481	8·980	9·182	9·427	9·224
7	6·826	8·382	8·172	8·732	8·205
8	6·170	7·536	6·947	7·297	6·962
9	5·515	6·444	5·605	5·555	5·599
10	4·859	5·104	4·247	3·964	4·223
11	4·204	3·517	2·971	2·443	2·939
12	3·549	1·683	1·879	1·866	1·854
13	2·893	-0·399	1·069	1·462	1·072
14	2·238	-2·727	0·641	1·333	0·700
15	1·582	-5·303	0·694	1·250	0·844
16	0·927	-8·126	1·328	1·000	1·610

In the last column I have placed the results of re-working the whole system, omitting the first whorl as largely influenced by the ground condition at the foot of



the stem.\* The improvement of fit is not sufficiently great to justify a publication of all the constants for the distribution in this modified case. But there is improvement for the higher whorls, which are so few in number as to be wholly insignificant when compared with the weight of the first few low whorls.

It will be noticed at once that the line and the parabola (which gives at the top of the stem negative numbers!) are absolutely unsuitable for representing the facts of the case. The cubic is better and certainly gives the general trend of the observations, but in this our last illustration we have clearly reached the limit of material to which such cubical regression can be satisfactorily applied. See Diagram V.

(12.) *Quartic Regression.*

It seemed of some interest in this case of *Equisetum* to ascertain whether any real improvement in description would be reached by considering the quartic regression curve. I briefly indicate the theory in this case as developed from the general method in the footnote, p. 25. We shall now have

$$Y_{x_p}/\sigma_y = b_0 + b_1(X_p/\sigma_x) + b_2(X_p/\sigma_x)^2 + b_3(X_p/\sigma_x)^3 + b_4(X_p/\sigma_x)^4.$$

Eliminating  $b_0$  and  $b_1$ , by the processes familiar to us from the case of cubical regression, we have

$$\begin{aligned} Y_{x_p}/\sigma_y = & r(X_p/\sigma_x) + b_2\{(X_p/\sigma_x)^2 - \sqrt{\beta_1}(X_p/\sigma_x) - 1\} \\ & + b_3\{(X_p/\sigma_x)^3 - \beta_2(X_p/\sigma_x) - \sqrt{\beta_1}\} \\ & + b_4\{(X_p/\sigma_x)^4 - (\beta_3/\sqrt{\beta_1})(X_p/\sigma_x) - \beta_2\} \dots \dots \dots \text{(lxx).} \end{aligned}$$

Hence as before

$$\left. \begin{aligned} \bar{\epsilon} &= b_2\phi_2 + b_3\phi_3 + b_4\phi_5 \\ \bar{\zeta} &= b_2\phi_3 + b_3\phi_4 + b_4\phi_6 \\ \bar{\theta} &= b_2\phi_5 + b_3\phi_6 + b_4\phi_7 \end{aligned} \right\} \dots \dots \dots \text{(lxxi).}$$

where  $\phi_2, \phi_3,$  and  $\phi_4$  are given as before by (li. and liv.), while

$$\phi_5 = \beta_4 - \beta_3 - \beta_2 \dots \dots \dots \text{(lxxii).}$$

$$\phi_6 = (\beta_5 - \beta_2\beta_3 - \beta_2\beta_1)/\sqrt{\beta_1} \dots \dots \dots \text{(lxxiii).}$$

$$\phi_7 = (\beta_1\beta_6 - \beta_3^2 - \beta_1\beta_2^2)/\beta_1 \dots \dots \dots \text{(lxxiv).}$$

and

$$\beta_5 = \nu_7\nu_3/\sigma_x^{10}, \quad \beta_6 = \nu_8/\sigma_x^8 \dots \dots \dots \text{(lxxv).}$$

Solving, we have

$$b_4 = \frac{\bar{\theta}(\phi_2\phi_4 - \phi_3^2) - \bar{\epsilon}(\phi_4\phi_5 - \phi_3\phi_6) - \bar{\zeta}(\phi_2\phi_6 - \phi_3\phi_5)}{\phi_2\phi_4\phi_7 - \phi_7\phi_3^2 - \phi_4\phi_5^2 - \phi_2\phi_6^2 + 2\phi_3\phi_5\phi_6} \dots \dots \dots \text{(lxxvi).}$$

\* 'Roy. Soc. Proc.', vol. 71, pp. 308-310.



and

$$\left. \begin{aligned} b_2 &= \frac{\bar{\epsilon}\phi_4 - \bar{\zeta}\phi_3}{\phi_2\phi_4 - \phi_3^2} - b_4 \frac{\phi_4\phi_5 - \phi_3\phi_6}{\phi_2\phi_4 - \phi_3^2} \\ b_3 &= \frac{\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3}{\phi_2\phi_4 - \phi_3^2} - b_4 \frac{\phi_2\phi_6 - \phi_3\phi_5}{\phi_2\phi_4 - \phi_3^2} \end{aligned} \right\} \dots \dots \dots \text{(lxxvii).}$$

Substituting in (lxx.), the solution is completed. The advantage of this form is that we see clearly the modifications made in  $b_2$  and  $b_3$  as we pass from cubical to quartic regression. On the other hand,  $\phi_6$  and  $\phi_7$ , as shown by (lxxv.), involve the 7<sup>th</sup> and 8<sup>th</sup> moments of the  $x$ -character. These are not only very laborious to calculate, but, as we have already shown, are as a rule very untrustworthy.

If we proceed as on p. 26, equation (lvii.), we find

$$\eta^2 - r^2 = b_2\bar{\epsilon} + b_3\bar{\zeta} + b_4\bar{\theta} \dots \dots \dots \text{(lxxviii).}$$

Using this and not the third equation of (lxxi.), we replace (lxxvi.) by

$$b_4 = (\phi_2\phi_4 - \phi_3^2) \frac{\left\{ \eta^2 - r^2 - \frac{\bar{\epsilon}^2}{\phi_2} - \frac{(\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2}{\phi_2(\phi_2\phi_4 - \phi_3^2)} \right\}}{\bar{\theta}(\phi_2\phi_4 - \phi_3^2) - \bar{\epsilon}(\phi_4\phi_5 - \phi_3\phi_6) - \bar{\zeta}(\phi_2\phi_6 - \phi_3\phi_5)} \dots \text{(lxxix).}$$

This equation for  $b_4$  only involves the 7<sup>th</sup> and not the 8<sup>th</sup> moment, but like the corresponding form (lx.) suffers from being a ratio of small quantities. (lxxvii.) completes the solution as before.

(lxxvii.) and (lxxix.) in conjunction give us a necessary condition for quartic regression. We can indeed now write the whole series of conditions as follows:—

Linear regression :

$$\eta^2 - r^2 = 0.$$

Parabolic regression :

$$\eta^2 - r^2 - \bar{\epsilon}^2/\phi_2 = 0.$$

Cubical regression :

$$\eta^2 - r^2 - \bar{\epsilon}^2/\phi_2 - (\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2 / \{\phi_2(\phi_2\phi_4 - \phi_3^2)\} = 0.$$

Quartic regression :

$$\eta^2 - r^2 - \bar{\epsilon}^2/\phi_2 - \frac{(\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2}{\phi_2(\phi_2\phi_4 - \phi_3^2)} - \frac{\{\bar{\theta}(\phi_2\phi_4 - \phi_3^2) - \bar{\epsilon}(\phi_4\phi_5 - \phi_3\phi_6) - \bar{\zeta}(\phi_2\phi_6 - \phi_3\phi_5)\}^2}{(\phi_2\phi_4 - \phi_3^2)(\phi_2\phi_4\phi_7 - \phi_7\phi_3^2 - \phi_4\phi_5^2 - \phi_2\phi_6^2 + 2\phi_3\phi_5\phi_6)} = 0 \dots \dots \dots \text{(lxxx).}$$

We now have a third possibility: we can get rid of the fourth product moment  $\bar{\theta}$  from the value of  $b_4$  and write it:

$$b_4 = \pm \sqrt{\frac{\eta^2 - r^2 - \bar{\epsilon}^2/\phi_2 - (\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2 / \{\phi_2(\phi_2\phi_4 - \phi_3^2)\}}{\phi_7 - \phi_5 \frac{\phi_4\phi_5 - \phi_3\phi_6}{\phi_2\phi_4 - \phi_3^2} - \phi_6 \frac{\phi_2\phi_6 - \phi_3\phi_5}{\phi_2\phi_4 - \phi_3^2}} \dots \dots \text{(lxxxii).}$$



While this value of  $b_4$  does not suffer like (lxxix.) from being the ratio of small quantities, and would *a priori* appear to save the calculation of  $\bar{\theta}$ , yet the right sign of the root may not be obvious on inspection, so that an actual determination of  $\bar{\theta}$  to find the sign of  $b_4$  may after all be needful. If (lxxx.) were absolutely satisfied, (lxxxii.), (lxxix.) and (lxxvi.) would lead to identical results; but this will rarely be true in practice. In any of the three cases  $b_2$  and  $b_3$  will be given by (lxxviii.). On the whole, I consider that (lxxxii.) and (lxxvi.) will give the better results, and probably the former the best, but it will generally require as much arithmetic as the latter.

(13). *Illustration E.—Calculation of the Quartic Regression Curve in the Case of Equisetum arvense.*

The only new constants required are :

$$\begin{aligned} \nu_7 &= 43,207.386, & \text{whence } \beta_5 &= 1.144,882, \\ \nu_8 &= 507,649.540, & \beta_6 &= 20.463,633, \\ \text{and :} \\ \phi_5 &= 3.425,069, & \phi_6 &= 3.452,046, \\ \phi_7 &= 15.015.792. \end{aligned}$$

These lead us to :

$$\begin{aligned} \frac{\phi_4\phi_5 - \phi_3\phi_6}{\phi_2\phi_4 - \phi_3^2} &= 2.723,384, & \frac{\phi_2\phi_6 - \phi_3\phi_5}{\phi_2\phi_4 - \phi_3^2} &= 1.211,194, \\ \Delta_4 &= \begin{vmatrix} \phi_2 & \phi_3 & \phi_5 \\ \phi_3 & \phi_4 & \phi_6 \\ \phi_5 & \phi_6 & \phi_7 \end{vmatrix} &= 1.745,622. \end{aligned}$$

Our successive conditions are therefore :

$$\begin{aligned} \eta^2 - r^2 &= .222,596, \\ \eta^2 - r^2 - \bar{\epsilon}^2/\phi_2 &= .069,266, \\ \eta^2 - r^2 - \bar{\epsilon}^2/\phi_2 - (\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2/\{\phi_2(\phi_2\phi_4 - \phi_3^2)\} &= .010,186, \\ \eta^2 - r^2 - \bar{\epsilon}^2/\phi_2 - (\bar{\zeta}\phi_2 - \bar{\epsilon}\phi_3)^2/\{\phi_2(\phi_2\phi_4 - \phi_3^2)\} \\ - \frac{\{\bar{\theta}(\phi_2\phi_4 - \phi_3^2) - \bar{\epsilon}(\phi_4\phi_5 - \phi_3\phi_6) - \bar{\zeta}(\phi_2\phi_6 - \phi_3\phi_5)\}^2}{(\phi_2\phi_4 - \phi_3^2)\Delta_4} &= .007,200, \end{aligned}$$

whence we see the successive approximations to the fulfilment of the conditions. Clearly great gains arise when we pass from linear to parabolic, and from parabolic to cubic regression, but the advance is not so conspicuous when we pass to quartic regression.



We have :—

$$\text{From (lxxvi.) : } b_4 = \cdot 044,517, \text{ and } b_2 = -\cdot 648,122, \quad b_3 = \cdot 171,260,$$

$$\text{From (lxxix.) : } b_4 = \cdot 151,842, \text{ and } b_2 = -\cdot 940,410, \quad b_3 = \cdot 041,981,$$

$$\text{From (lxxxix.) : } b_4 = \cdot 025,999, \text{ and } b_2 = -\cdot 597,691, \quad b_3 = \cdot 193,688.$$

The equations to the three corresponding quartics are :

$$(a). Y_x = 1\cdot 724,611 - \cdot 913,208 X_p - \cdot 169,311 X_p^2 + \cdot 012,629 X_p^3 + \cdot 000,927 X_p^4,$$

$$(b). Y_x = 2\cdot 047,717 - \cdot 734,966 X_p - \cdot 245,667 X_p^2 + \cdot 003,096 X_p^3 + \cdot 003,161 X_p^4,$$

$$(c). Y_x = 1\cdot 668,788 - \cdot 944,192 X_p - \cdot 156,137 X_p^2 + \cdot 014,283 X_p^3 + \cdot 000,541 X_p^4.$$

The values of  $Y_x$  and  $X_p$  are as before measured from the means, or 7·216,851 and 6·403,315 respectively.

The values of the observed and calculated ordinates are given in Table IX., and the graph of the results in the lower half of Diagram V.

TABLE IX.—Mean Number of Branches to Whorl in *Equisetum* deduced from Quartic Regression.

Position.	Quartic (a).	Quartic (b).	Quartic (c).	Observed.
1	7·731	8·269	7·637	7·619
2	8·950	8·662	9·000	9·294
3	9·715	9·222	9·800	9·627
4	10·014	9·674	10·073	9·730
5	9·858	9·816	9·866	9·643
6	9·281	9·521	9·240	9·427
7	8·339	8·740	8·270	8·732
8	7·109	7·498	7·042	7·297
9	5·692	5·898	5·656	5·555
10	4·209	4·116	4·225	3·964
11	2·816	2·407	2·875	2·443
12	1·651	1·100	1·745	1·866
13	0·930	0·600	0·987	1·462
14	0·857	1·389	0·766	1·333
15	1·665	4·022	1·259	1·250
16	3·609	9·133	2·657	1·000

From these results we deduce the following conclusions :—

(i.) That the use of a quartic instead of a cubic regression curve has not very markedly bettered the fit. The failure to get a closer fit lies largely in the nature of the material. The number of plants with more than 13 whorls is very few, and their contribution allows little weight to the tail of the regression curve. Further, all our



attempts to fit a smooth regression curve show that the observed data are unduly flattened at the top. If we confine ourselves to a homogeneous series of 110 plants with ten whorls apiece, we get a remarkably good fit.\* The S-shape of the regression line as indicated in both cubic and quartic does, however, appear to be characteristic of the nature of the plant, and I take it that more ample material would allow of a closer analytical description by a simple cubic. I doubt whether for practical statistics the use of the quartic will often be requisite.

(ii.) The comparative failure of the quartic (*b*) shows us that a formula like (lxxix.) is of small service. This corresponds fully to our experience in the use of (lx.) in the case of the cubic. In both cases we get rid of a high moment by making a certain constant the ratio of two small quantities, and experience shows us that the result is unsatisfactory. It is accordingly preferable to use formulæ involving high moments of one variable in preference to those with a ratio of small quantities.

(iii.) The quartic (*c*) appears as good, if not slightly better, than quartic (*a*). In (*c*) we have got rid of a high product moment,  $\bar{\theta}$ , by supposing the quartic condition (lxxx.) rigidly fulfilled. This of course is not the case. It is clear that product moments like  $\bar{\theta}$  of the 5<sup>th</sup> order are far from advantageous, and this is the same principle which was in evidence when we found (lxv.) giving better results than (lxiv.) for parabolic regression. Hence we must further conclude that the use of third, fourth or fifth product moments is disadvantageous as compared respectively with fifth to eighth moments of one variable. Or, a moment two degrees higher is preferable to a product moment in calculating correlation values. This is, I think, consonant with our knowledge of the relative magnitude of the probable errors in the two cases.

#### (14.) *General Conclusions.*

(i.) The present paper provides us with a general method of dealing with the regression line and the variability of arrays in the case of skew correlation, without any assumption as to the analytical form of the skew correlation surface.

(ii.) It provides a nomenclature and classification of the types of array variability which may be of service.

Arrays are either *homoclitic* or *heteroclitic*, according as their skewnesses are of equal magnitude or not. Arrays are further *homoscedastic* or *heteroscedastic*, according as their standard deviations are alike or different. Skew arrays are termed *allocurtic*; if arrays are symmetrical about their mean, they are *isocurtic*.

A heteroclitic system of arrays may be *nomie* or *anomie*, according as the skewness of the arrays changes continuously or irregularly with the position of the array.

A heteroscedastic system of arrays is also either *nomie* or *anomie*, according as the standard deviation of the arrays changes continuously or irregularly with the

\* 'Roy. Soc. Proc.,' vol. 71, p. 308.



position of the arrays. Anomic heteroclisys and anomic heteroscedasticity probably only signify that our material is either heterogeneous or too sparse to free us from the large errors of random sampling in the extreme arrays. Still the terms will be found of use in describing the actual data.

The curve in which the skewness of the array is plotted to its position is termed the *clitic curve*; the curve in which the ratio of the standard deviation of the array to the standard deviation of the character in the population at large is plotted to position is termed a *scedastic curve*.

(iii.) The types of regression have been classified into *linear, parabolic, cubic* and *quartic*. For most practical purposes the first three suffice. Necessary criteria have been given for each case. But as in the case of the skew frequency of one character, an indefinite number of conditions ought theoretically to be fulfilled. Practically in dealing with frequency, no criteria are absolutely fulfilled, and the probable errors of the expressions used become unmanageable as we ascend in the scale. We must therefore be content to estimate the degree of approximation with which one or two necessary criteria are satisfied.

The fundamental test of deviation from the familiar form of linear regression is the inequality of the correlation coefficient  $r$  and the newly introduced correlation ratio  $\eta$ . The probable error of this latter is determined. It is shown that  $\sigma_y \sqrt{1 - \eta^2}$  is the mean standard deviation of a system of arrays in skew correlation. The ease with which  $\eta$  can be calculated suggests that in many cases it should accompany, if not replace the determination of the correlation coefficient.

In the determination of the constants of the regression curve we must use moments and product moments. The limitations to the order of the curve used depend: (a) on the labour of the arithmetic, (b) on the increasing probable errors of the higher moments and product moments. For these reasons it seems idle to propose going beyond the 6<sup>th</sup> to 8<sup>th</sup> moments, or the 3<sup>rd</sup> to 5<sup>th</sup> product-moments. Practical experience suggests that little is to be gained by using moments beyond the 6<sup>th</sup>, or product moments beyond the 3<sup>rd</sup>. A quartic regression curve may be useful occasionally, but it has yet to justify its necessity. As our object is not to reproduce the given data, but to provide a graduation for them, which smooths down the errors of random sampling, we believe that any legitimate and practical theory must discard the high moments and high product moments with which THIELE and LIPPS propose to deal.

(iv.) There is one point to which reference ought to be made. Some reader may enquire why the method of my paper on curving fitting\* should not be applied to these regression curves *in general*, as we have in practice once or twice already applied it. It would seem that that method is the easier, involving in the case of the quartic only quantities analogous to our  $r$ ,  $e$ ,  $\zeta$  and  $\theta$ . The answer is

\* "On the Systematic Fittings of Curves to Observations and Measurements." 'Biometrika,' vol. I., pp. 265-303, and vol. II., pp. 1-23, especially the latter, pp. 11-15.



straightforward: that process supposes every  $y_x$ , to have equal weight, or  $n_x$ , to be the same for each array. Hence the higher moments of the  $x$ -character, which are really involved, can be written down without calculation once and for all.\* The complexity of our present investigation arises from the introduction of the weighting into the calculation of the moments of the  $x$ -character, as well as into that of the product moments  $r$ ,  $e$ ,  $\zeta$ ,  $\theta$ . Our results therefore, although they might not look so good on a graph of the regression curve, would be markedly better, if due weight were given to the frequency of each array. The difference of the two conceptions is comparable to the determination of the regression on the one hand from the correlation coefficient, and on the other from merely striking a line through the plotted means of the arrays. The method of moments in the present case, if we except the use of  $\eta$ , is identical with that of fitting a curve to a *continuum* in space by the method of least squares.

(v.) No stress whatever is laid on the actual instances here selected for illustration of the methods of this paper. I have merely chosen out of available material cases in which I had come across skew regression of various types. Thus we find:—

(a.) The correlation of the number of branches and position of the whorl in *Asperula odorata* is practically parabolic, homoscedastic and of nomic heteroclisly.

(b.) The correlation between auricular height of head and age in girls is cubical, of nomic heteroscedasticity and of anomic heteroclisly. It is probably really a case of isocurtosis.

(c.) The correlation of size of cell and size of body in *Daphnia magna*, allowing for the irregularities produced by the ecdyses, is parabolic or cubic, of nomic heteroscedasticity, and probably, but for the above-mentioned irregularities, of isocurtic homoclisly.

(d.) The correlation of the number of branches and position of the whorl in *Equisetum arvense* is cubical or possibly even quartic, of markedly nomic heteroscedasticity and markedly nomic heteroclisly.

It is not impossible that slips have occurred in the lengthy arithmetic involved, but every important piece of work has been done independently twice, once by Dr. ALICE LEE, whom I have most heartily to thank for her unwearying assistance, and once by myself. To preserve uniformity of working, the constants have in each case been carried to six figures. This involves little or no additional trouble, using as we do mechanical calculators. The final results are of course of no value beyond their probable errors, which will be in the second or third place of figures. No doubt I shall be told that there is a show of accuracy in the number of decimal figures retained, which does not really exist. It does not exist (and I am as fully conscious of its non-existence as any would-be critic) so far as our results fit the actual population, of which we have but a random sample. The figures, however, are of importance, as far as testing accuracy of fit of result to *actual* sample goes. The

\* 'Biometrika,' vol. II., p. 12.



cubic or quartic curves may have coefficients insensible before the third or fourth figure of decimals, and these coefficients have to be multiplied occasionally by abscissae of the third or fourth powers of 7 to 9. Hence to get ordinates true, *as far as the sample goes*, to the second or third figure, we require to work to a fairly high number of figures. There is no magic in six figures, four or five would probably satisfy another worker, but they are easily read off the calculator we use, and if the constants had been tabled only to four or five, no reader would have been able to agree exactly, if he wished to test any of our results, even to three figures, with the final ordinates.



DIAGRAM I. SKEW CORRELATION IN ASPERULA ODORATA.

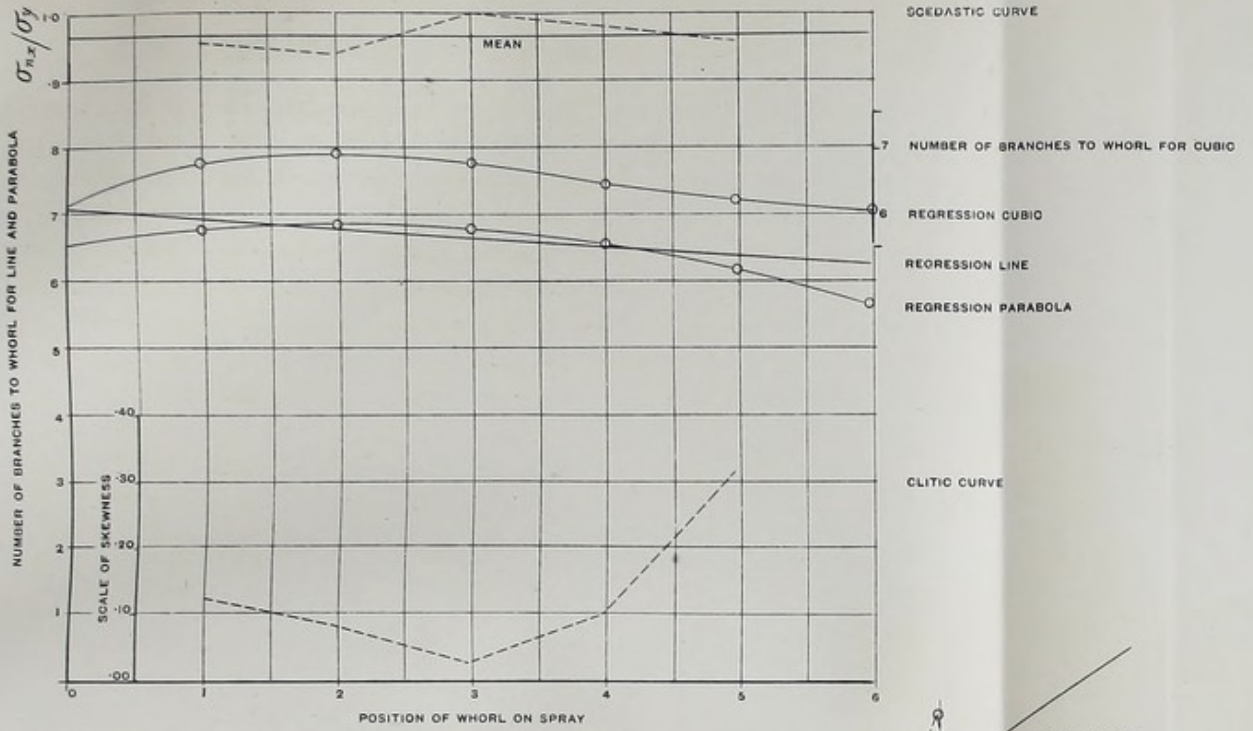


DIAGRAM II. SKEW CORRELATION, HEAD-HEIGHT AND AGE IN GIRLS.

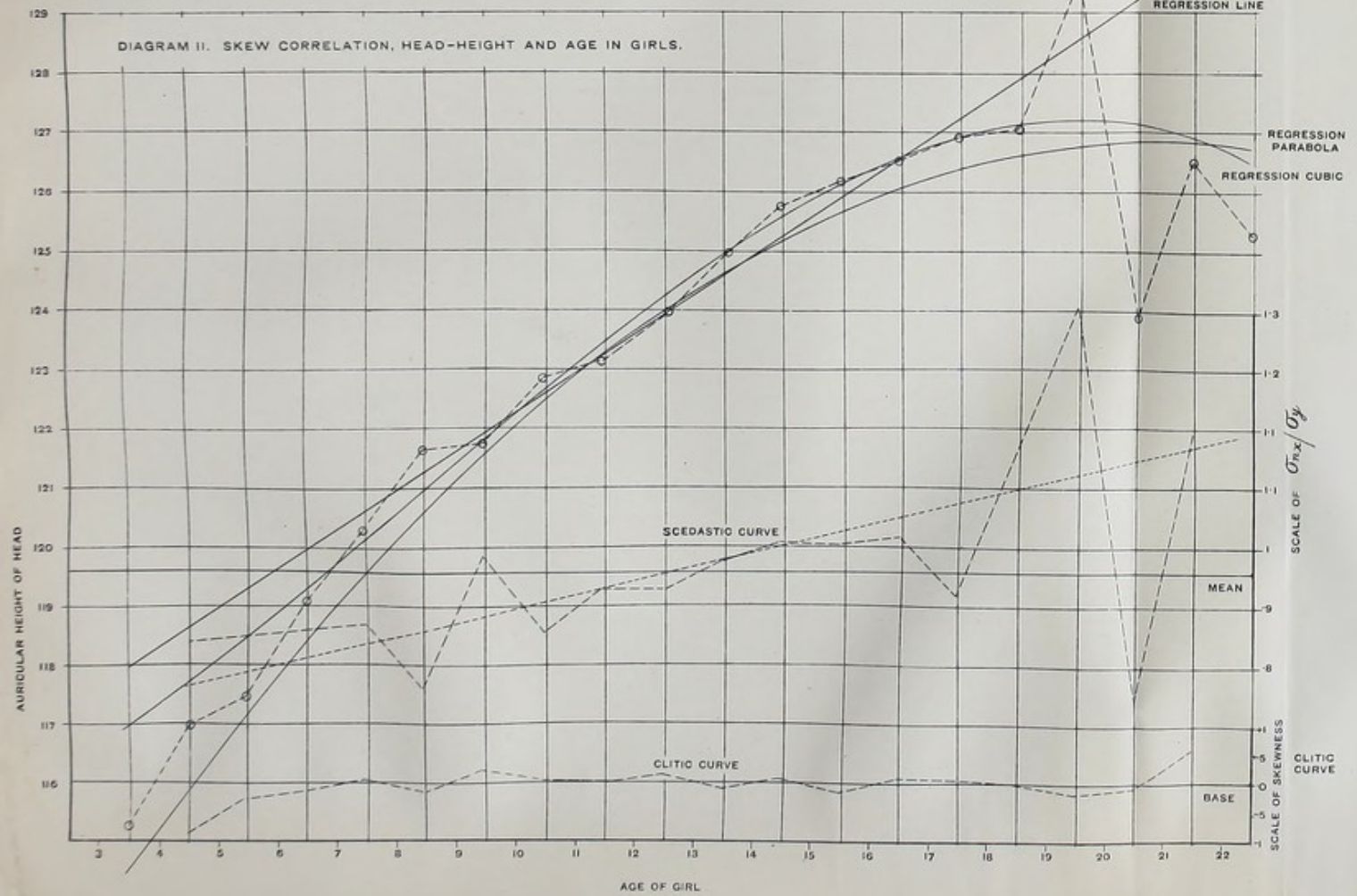








DIAGRAM III. SKEW CORRELATION BETWEEN SIZES OF CELL AND BODY IN DAPHNIA.

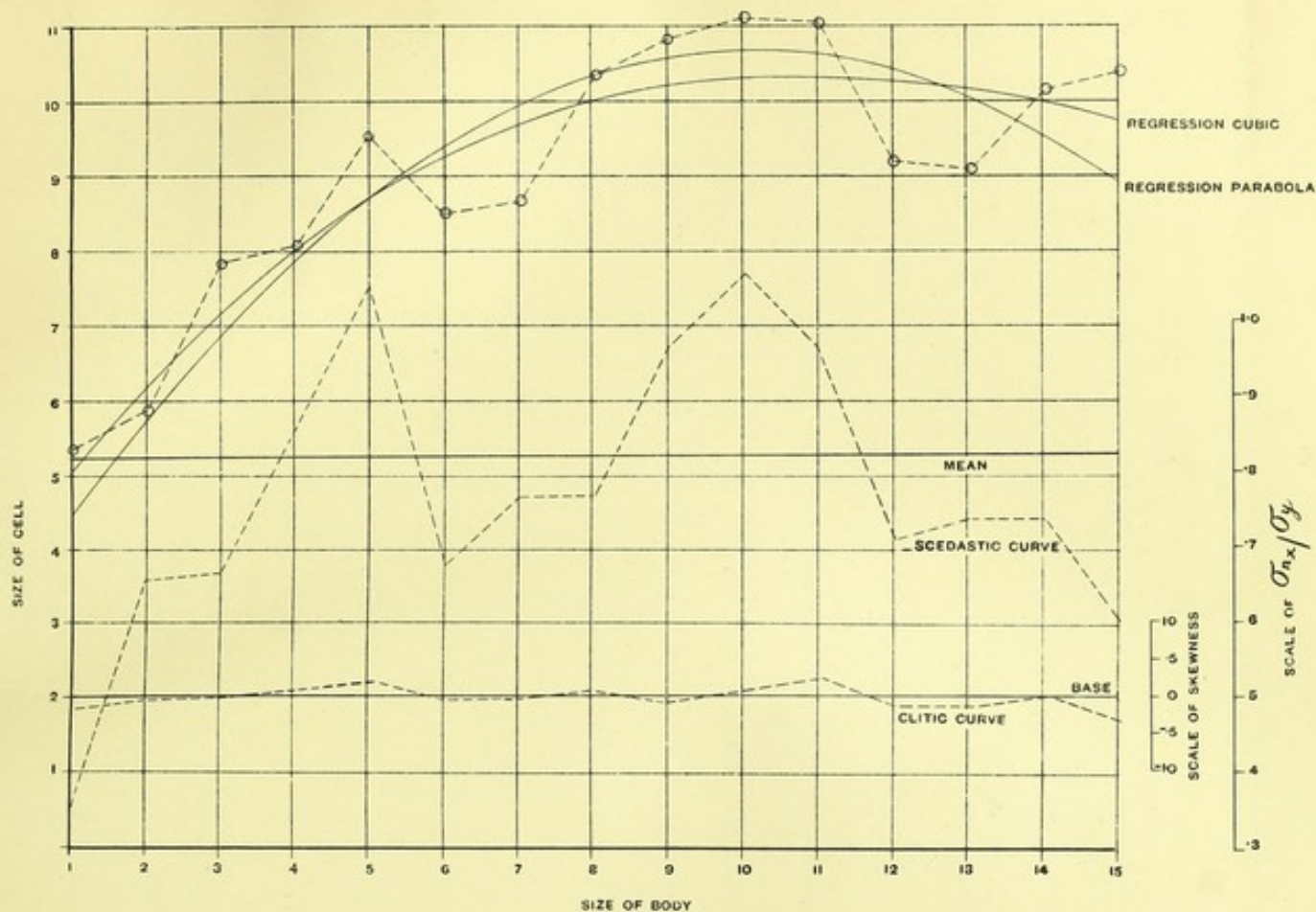


DIAGRAM IV. SKEW CORRELATION BETWEEN BRANCHES AND POSITION OF WHORL IN EQUISETUM: SCEDASTIC AND CLITIC CURVES

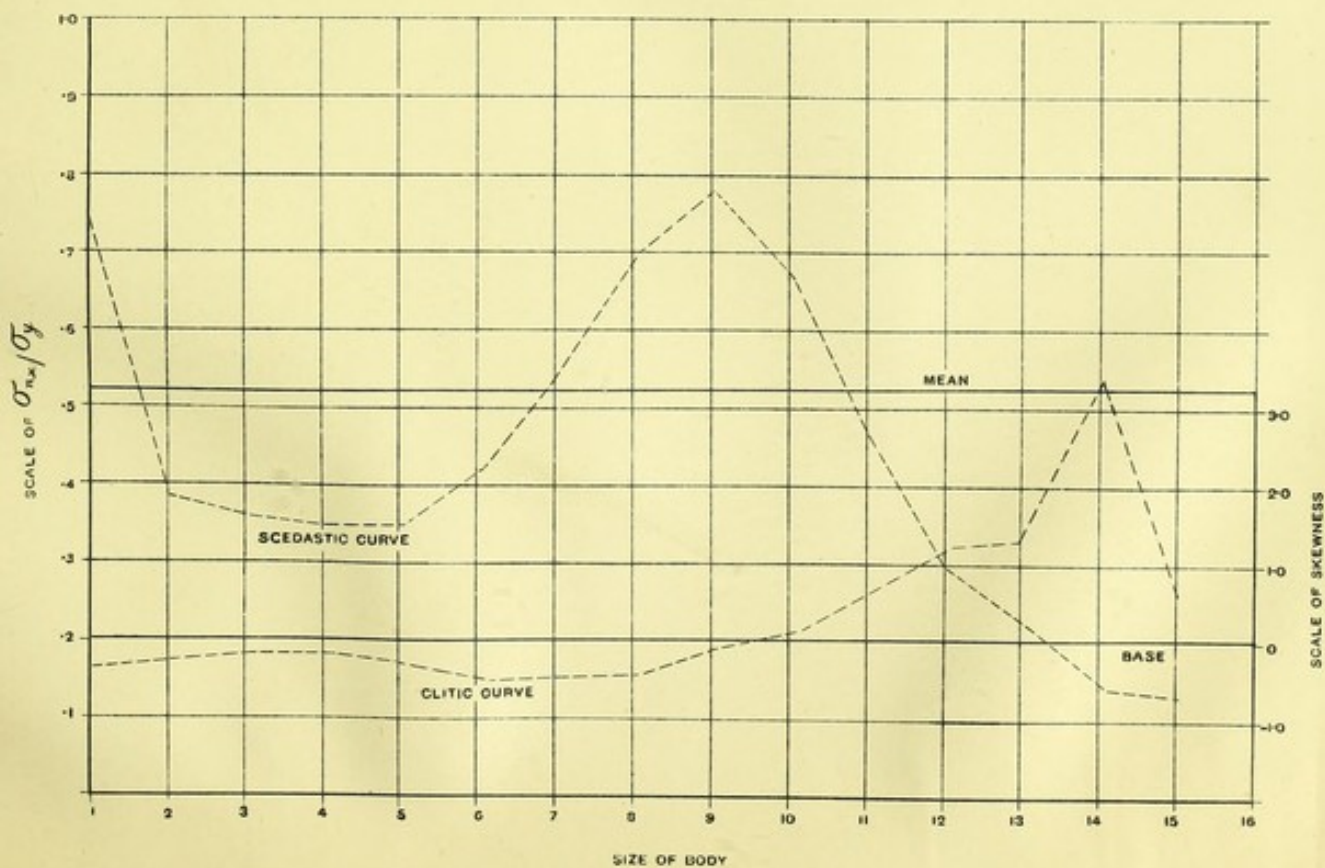
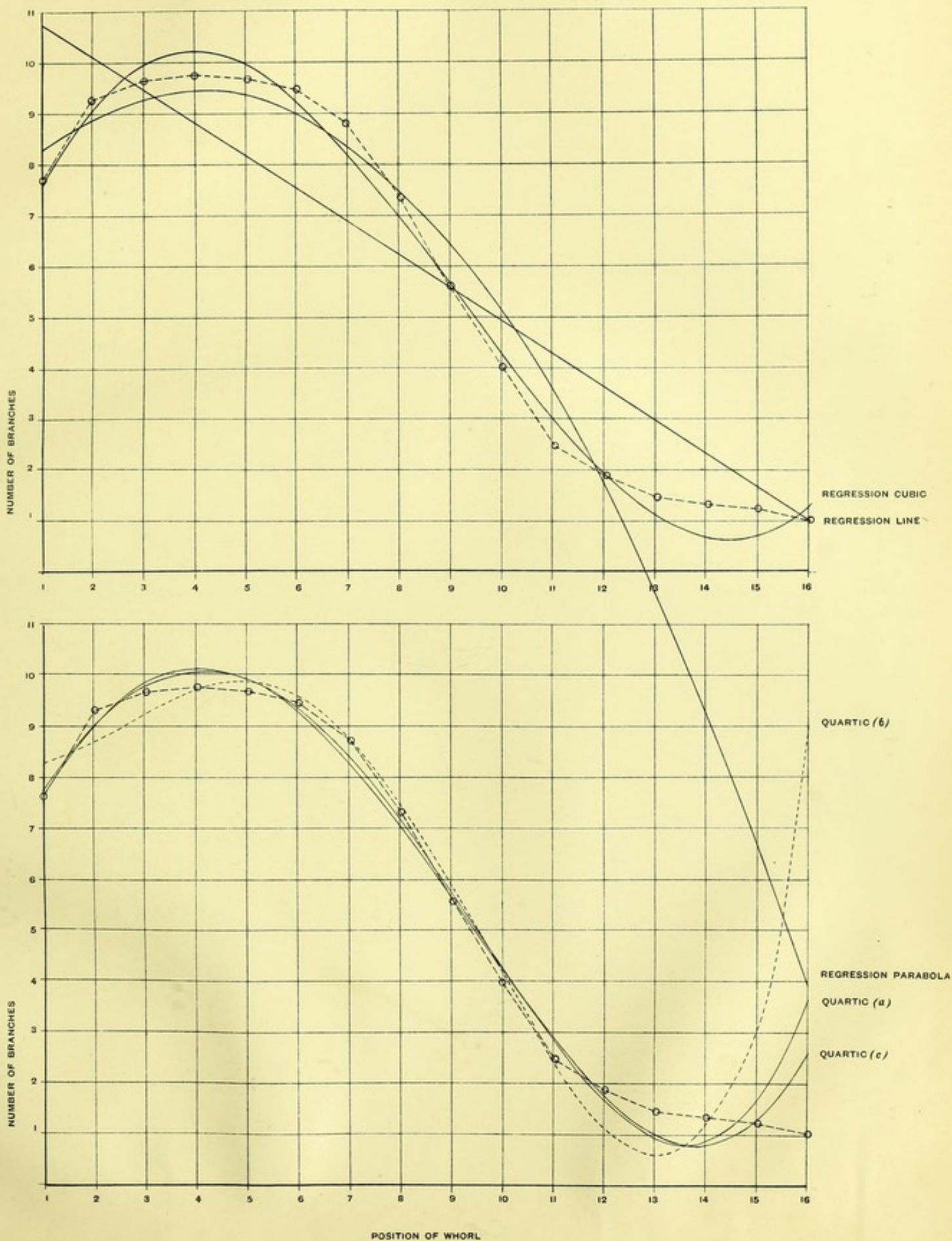








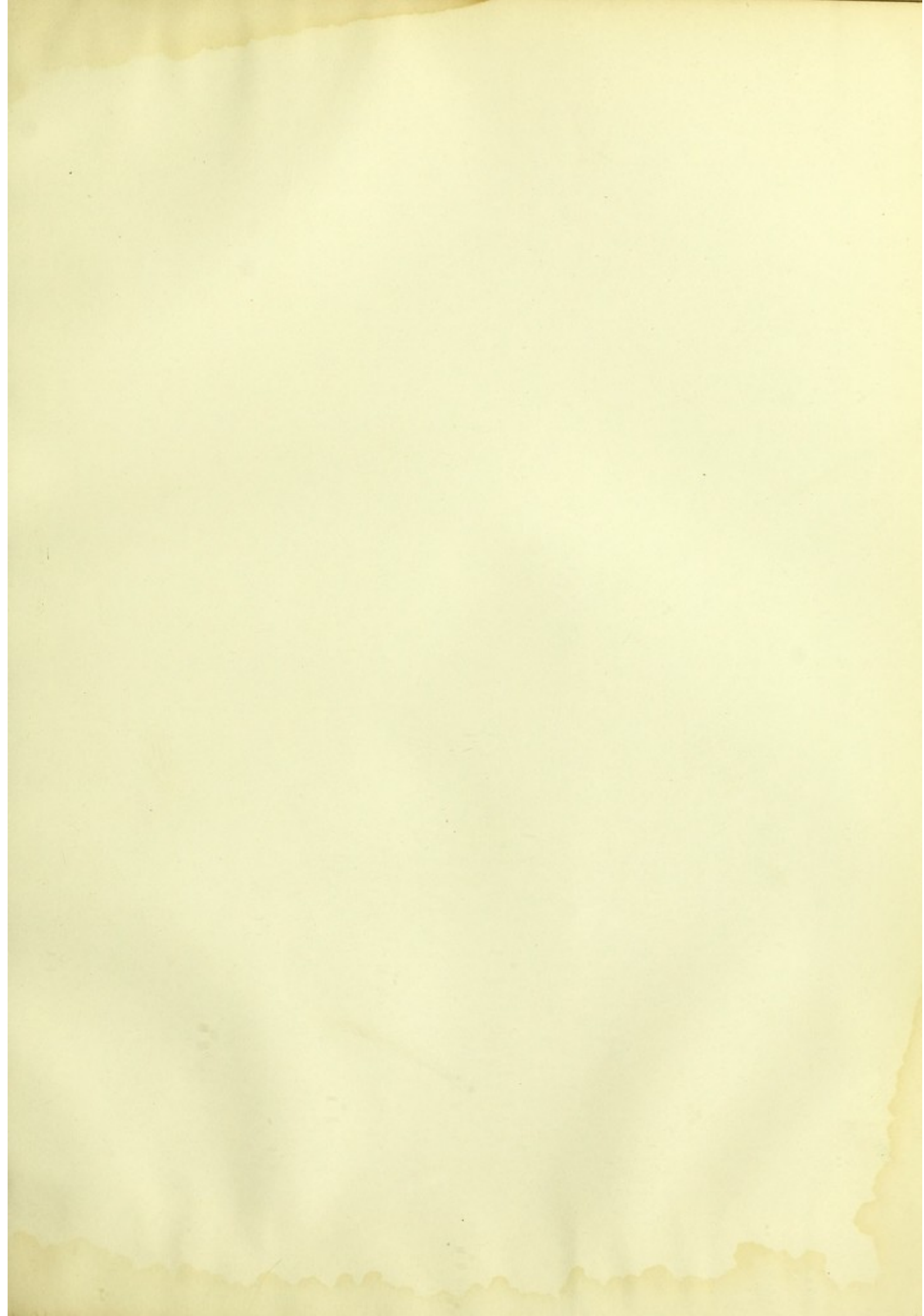
DIAGRAM V. SKEW CORRELATION BETWEEN BRANCHES AND POSITION OF WHORL IN EQUISETUM:  
REGRESSION CURVES.







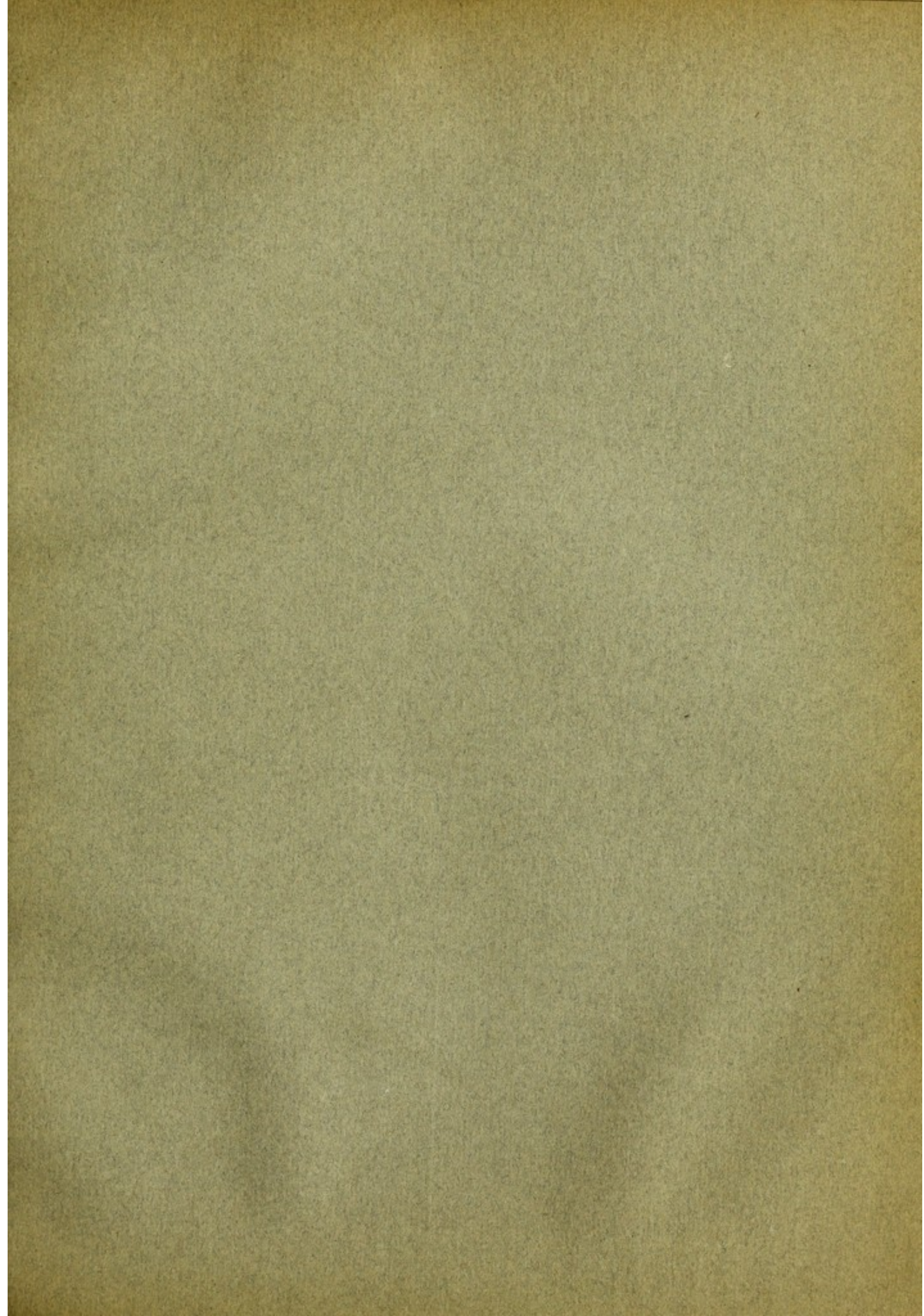














# DRAPERS' COMPANY RESEARCH MEMOIRS.

DEPARTMENT OF APPLIED MATHEMATICS, UNIVERSITY COLLEGE,  
UNIVERSITY OF LONDON.

These memoirs will be issued at short intervals. The following are ready or will probably appear later in this series:—

## *Biometric Series.*

- I. Mathematical Contributions to the Theory of Evolution.—XIII. On the Theory of Contingency and its Relation to Association and Normal Correlation. By KARL PEARSON, F.R.S. *Issued.* Price 4s.
- II. Mathematical Contributions to the Theory of Evolution.—XIV. On the Theory of Skew Correlation and Non-linear Regression. By KARL PEARSON, F.R.S. *Issued.* Price 5s.
- III. Mathematical Contributions to the Theory of Evolution.—XV. On Homotopy in the Animal Kingdom. By ERNEST WARREN, D.Sc., ALICE LEE, D.Sc., EDNA LEA-SMITH, MARION RADFORD and KARL PEARSON, F.R.S. *Shortly.*

## *Technical Series.*

- I. On a Theory of the Stresses in Crane and Coupling Hooks with Experimental Comparison with Existing Theory. By E. S. ANDREWS, B.Sc.Eng., assisted by KARL PEARSON, F.R.S. *Issued.* Price 3s.
- II. On some Disregarded Points in the Stability of Masonry Dams. By L. W. ATCHERLEY, assisted by KARL PEARSON, F.R.S. *Issued.* Price 3s. 6d.
- III. On the Graphics of Metal Arches, with Special Reference to the Relative Strength of Two-pivoted, Three-pivoted and Built-in Metal Arches. By L. W. ATCHERLEY and KARL PEARSON, F.R.S. *Issued.* Price 5s.
- IV. On Torsional Vibrations in Shafting. By KARL PEARSON, F.R.S.

PUBLISHED BY DULAU AND CO.

## MATHEMATICAL CONTRIBUTIONS TO THE THEORY OF EVOLUTION.

### XI. ON THE INFLUENCE OF SELECTION ON THE VARIABILITY AND CORRELATION OF ORGANS.

By KARL PEARSON, F.R.S.

'Phil. Trans.,' vol. 200, pp. 1-56. Price 3s.

### XII. ON A GENERALISED THEORY OF ALTERNATIVE INHERITANCE, WITH SPECIAL REFERENCE TO MENDEL'S LAWS.

By KARL PEARSON, F.R.S.

'Phil. Trans.,' vol. 203, pp. 53-86. Price 1s. 6d.

PUBLISHED BY THE CAMBRIDGE UNIVERSITY PRESS.

## BIOMETRIKA.

### A JOURNAL FOR THE STATISTICAL STUDY OF BIOLOGICAL PROBLEMS.

Edited, in Consultation with FRANCIS GALTON,

By W. F. R. WELDON, KARL PEARSON and C. B. DAVENPORT.

#### VOL. III., PARTS II. AND III.

- I. Experimental and Statistical Studies upon Lepidoptera.  
I. Variation and Elimination in *Philosamea Cynthia*.  
By HENRY EDWARD CHAMPTON.
- II. On the Laws of Inheritance in Man.—II. On the Inheritance of the Mental and Moral Characters in Man, and its Comparison with the Inheritance of the Physical Characters. By KARL PEARSON.
- III. A Study of the Variation and Correlation of the Human Skull with Special Reference to English Crania. By W. R. MACDONELL. (With 50 Plates.)
- IV. On the Inheritance of Coat-colour in the Greyhound. By AMY BARRINGTON, ALICE LEE and K. PEARSON.
- V. Note on a Race of *Clausilia itala* (Von Martens). By W. F. R. WELDON.
- Miscellaneous. On an Elementary Proof of SHEPPARD'S Corrections for Raw Moments and on some Allied Points. (Editorial.)

#### VOL. III., PART IV.

- I. Merism and Sex in *Spinax niger*. By R. C. PUNNETT.
- II. Note on Inheritance of Meristic Characters in *Spinax niger*. By K. PEARSON.
- III. On the Measurement of Internal Capacity from Cranial Circumferences. By M. A. LEWENZ and K. PEARSON. (With two Plates.)
- IV. Étude Biométrique sur les Variations de la Fleur et sur l'Hétérostyle de *Pulmonaria officinalis L.* Par EDMUND GAIN.
- Miscellaneous. (I.) On the Correlation between Hair Colour and Eye Colour in Man. By K. PEARSON.  
(II.) On the Correlation between Age and the Colour of Hair and Eyes in Man. By G. UCHIDA.  
(III.) On the Contingency between Occupations in the Case of Father and Son. By EMILY PERRIN.  
(IV.) On a Convenient Means of Drawing Curves to various Scales. By G. UDSY YULE.  
(V.) Albinism in Sicily. By W. BATESON.

The subscription price, payable in advance, is 30s. *net* per volume (post free); single numbers 10s. *net*. Volumes I. to III. (1902-4) complete, 30s. *net* per volume. Bound in Buckram 34s. 6d. *net* per volume. Subscriptions may be sent to Messrs. C. J. Clay & Sons, Cambridge University Press Warehouse, Ave Maria Lane, London, either direct or through any bookseller.



DEPARTMENT OF APPLIED MATHEMATICS,  
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON

---

DRAPERS' COMPANY RESEARCH  
MEMOIRS

BIOMETRIC SERIES. III.

---

MATHEMATICAL CONTRIBUTIONS TO THE THEORY  
OF EVOLUTION.—XV. A MATHEMATICAL  
THEORY OF RANDOM MIGRATION.

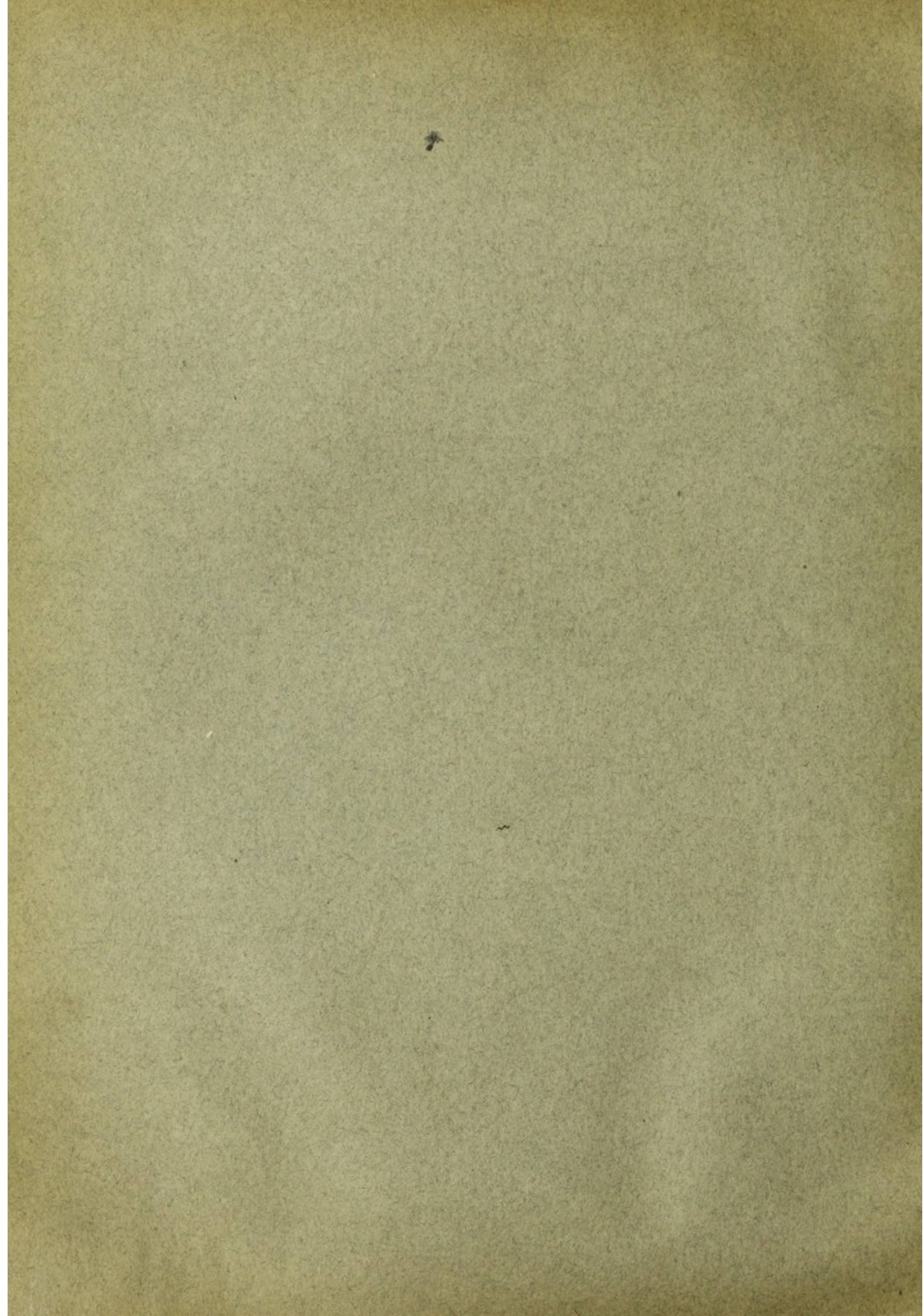
BY  
KARL PEARSON, F.R.S.  
WITH THE ASSISTANCE OF  
JOHN BLAKEMAN, M.Sc.

[With Seven Diagrams.]

LONDON:  
PUBLISHED BY DULAU AND CO., 37, SOHO SQUARE, W.  
1906

*Price Five Shillings*







DEPARTMENT OF APPLIED MATHEMATICS,  
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON

---

DRAPERS' COMPANY RESEARCH  
MEMOIRS

BIOMETRIC SERIES.

---

III. MATHEMATICAL CONTRIBUTIONS TO THE THEORY  
OF EVOLUTION.—XV. A MATHEMATICAL  
THEORY OF RANDOM MIGRATION.

BY

KARL PEARSON, F.R.S.  
WITH THE ASSISTANCE OF  
JOHN BLAKEMAN, M.Sc.

LONDON :

PUBLISHED BY DULAU AND CO., 37, SOHO SQUARE, W.  
1906



DEPARTMENT OF MATHEMATICS  
UNIVERSITY OF CALIFORNIA  
DRAHERS COMPANY MEMOIRS  
REGISTERED

III. MATHEMATICAL CONTRIBUTIONS TO THE THEORY  
OF BILINAR FORMS AND QUADRATIC  
THEORY OF BILINAR ALGEBRA

BY  
DR. H. DRÄHER  
AND  
DR. H. DRÄHER



## *Mathematical Contributions to the Theory of Evolution.*

XV. A MATHEMATICAL THEORY OF RANDOM MIGRATION. By  
KARL PEARSON, F.R.S., with the assistance of JOHN BLAKEMAN, M.Sc.

(1) *Introductory.* In dealing with any natural phenomenon,—especially one of a vital nature, with all the complexity of living organisms in type and habit,—the mathematician has to simplify the conditions until they reach the attenuated character which lies within the power of his analysis\*. The problem of migration is one which is largely statistical, but it involves at the same time a close study of geographical and geological conditions, and of food and shelter supply peculiar to each species. Some years ago the late Professor Weldon started an extensive study as to the distribution of various species and local races of land snails, but he was struck by the absence in several cases of any definite change of environment at the boundaries of the distribution of a definite race. It occurred to me in thinking over the matter that such boundaries, where they exist, may possibly not be permanent. To take a purely hypothetical illustration: A species is pushed back to a certain limit by a change of environmental conditions—say, an ice age. Does it follow that if the environment again becomes favourable, that it will *rapidly* occupy possible country? What is the rate of infiltration of a species into a possible habitat? It depends, of course, on a whole series of most complex conditions, the rate of locomotion, the channels of communication, the distribution of food areas and breeding grounds in the new country, and the connecting links between all these. Every detail must be studied by the field naturalist in relation to each species. All the mathematician can do is to make an idealised system, which may be dangerous, if applied dogmatically to any particular case, but which can hardly fail to be suggestive, if it be treated within the limits of reasonable application. The idealised system which I proposed to myself was of the following kind:

(i) Breeding grounds and food supply are supposed to have an average uniform distribution over the district under consideration. There is to be no special following of river beds or forest tracks.

\* This is of course a perfectly familiar process to every mathematical physicist, but its unfamiliarity leads the biologist to suspect or even discard mathematical reasoning, instead of testing the result as the physicist does by experiment and observation.



(ii) The species scattering from a centre is supposed to distribute itself uniformly in all directions. The average distance through which an individual of the species moves from habitat to habitat will be spoken of as a "flight," and there may be  $n$  such "flights" from locus of origin to breeding ground, or again from breeding ground to breeding ground, if the species reproduces more than once. A flight is to be distinguished from a "flitter," a mere two and fro motion associated with the quest for food or mate in the neighbourhood of the habitat.

(iii) Now taking a centre, reduced in the idealised system to a point, what would be the distribution after  $n$  random flights of  $N$  individuals departing from this centre? This is the *first* problem. I will call it the *Fundamental Problem of Random Migration*.

(iv) Supposing the first problem solved, we have now to distribute such points over an area bounded by any contour, and mark the distribution on both sides of the contour after any number of breeding seasons. The shape of the contour and the number of seasons dealt with provide a series of problems which may be spoken of as *Secondary Problems of Migration*.

A little consideration of the Fundamental Problem showed me that it presented considerable analytical difficulties, and I was by no means clear that the series of hypotheses adopted would be sufficiently close to the natural conditions of any species to repay the labour involved in the investigation. At this stage the matter rested, until last year Major Ross put before me the same problem as being of essential importance for the infiltration of mosquitoes into cleared areas, and asked me if I could not provide the statistical solution of it. He considered that we might treat a district as approximately "equi-swampous," and thus my conditions (i), (ii) above could be applied to obtain at any rate a first approximation to the solution.

Starting on the problem again I obtained the solution for the distribution after two flights, an integral expressing the distribution after three flights, which I carelessly failed to see could be at once reduced to an elliptic integral, and the general functional relation between the distribution after successive flights. At this point I failed to make further progress, and under the heading of "The Problem of the Random Walk" asked for the aid of fellow-mathematicians in *Nature*\*. The reply to my appeal was threefold. Mr Geoffrey T. Bennett sent me in terms of elliptic integrals the solution for three flights. Lord Rayleigh drew my attention to the fact that the "problem of the random walk" where the number of flights is very great becomes identical with a problem in the combination of sound amplitudes in the case of notes of the same period, which he has dealt with in several papers†. Lastly Professor J. C. Kluyver presented a paper to the Royal Academy

\* July 27th, 1905.

† *Phil. Mag.*, August, 1880, p. 75; February, 1899, p. 246.



of Sciences of Amsterdam, entitled "A local probability problem."\* Professor Kluyver obtains the general solution in terms of the integral of a product of Bessel's functions of the zero and first orders. He deduces Rayleigh's solution for  $n$  large, he shows that the Bessel function integral represents a series of different analytic functions in different intervals, and proves a number of special problems of very considerable interest. Referring to his general solution, he writes, however :

"From this result we infer that the probability sought for is of a rather intricate character. The  $n + 1$  functions  $J$  are oscillating functions, and have their signs altering in an irregular manner as the variable  $u$  increases. Hence even an approximation of the integral is not easily found, and as a solution of Pearson's problem it is little apt to meet the requirements of the proposer."†

Kluyver's solution is of extreme suggestiveness for the analytical theory of discontinuous functions. In the endeavour to express it in a form suited to my special purposes I have come across a long series of Bessel function properties, some at least of which seem to me novel, but have unfortunately no bearing on the problem of migration. If we turn to Rayleigh's solution for  $n$  large, I must confess at once to being unconvinced of the adequacy of the proofs used to deduce it, especially that in the *Theory of Sound*‡. Kluyver's proof of Rayleigh's solution§ appears to me to also require much strengthening, and in neither case do we have any practical measure of what the number of flights must be before we have in practice a reasonable accordance between the discontinuous Bessel's function integral expression and the Rayleigh solution of Gaussian frequency type.

After a good many failures I have succeeded in obtaining a solution in series of the Bessel function integral, but this not of a character to be of service for frequent arithmetical calculations. It serves, however, to test the approximation of the Rayleigh solution and the accuracy of the solutions for few flights obtained by other processes. At this stage I realised that the functional equation between the distributions for  $n$  and  $n + 1$  flights could be solved graphically, and that starting with the known distributions for  $n = 2$  or  $n = 3$ , we could by very great labour, but absolutely straightforward graphical work and the use of mechanical integrators, build up in succession the solutions for  $n = 4, 5, 6, 7$ , etc. I proposed that this process should be continued until the graphically found distribution coincided with the plotted values obtained from the above solution in series. This was achieved for  $n = 7$ . For  $n = 6$  and  $n = 7$ , the solution in series approaches to the Rayleigh solution, with which for all practical purposes it may be asserted to coincide for  $n = 10$ . We have thus reached a continuous graphical illustration of the transition of a series of discontinuous and, in many respects, remarkable analytical functions, step by step with the increase of  $n$  into a normal curve of errors. The relation-

\* *Koninklijke Akademie van Wetenschappen te Amsterdam*. Proceedings, Oct. 25, 1905, pp. 341—50.

† *loc. cit.* p. 343.

‡ 2nd Edition, § 42 a.

§ Kluyver, *loc. cit.* p. 345.



ship is a noteworthy one, and not without suggestion for other branches of investigation.

The exact method of graphical solution will be described later; the whole labour of it, involving many weeks' work, was due to my assistant, Mr John Blakeman, M.Sc.

(2) *General Analytical Solution of the Fundamental Problem.* Let the origin be taken at the centre of dispersion and  $r$  be the distance of any small elementary area  $\alpha$  from the centre of dispersion. Let  $\phi_n(r^2) \cdot \alpha$  be the frequency of individuals on  $\alpha$  after the  $n$ th flight, and  $\phi_{n+1}(r^2) \alpha$  their frequency on the same element after the  $(n+1)$ th flight. Let  $l$  be the length of the flight. Then only those individuals who were on a circle of radius  $l$  round the centre of  $\alpha$  after the  $n$ th flight can reach  $\alpha$  with the  $(n+1)$ th flight, and only those individuals of these who take their flight in one definite direction. Let  $O$  be the centre of dispersion,  $C$  the centre of  $\alpha$ ,  $P$  a point on the circle of radius  $l$  round  $C$ ,  $\angle PCO = \theta$ , then the frequency per unit area at  $P$  is  $\phi_n(r^2 + l^2 - 2rl \cos \theta)$ , and the amount which goes in directions between  $\theta$  and  $\theta + \delta\theta$  is  $d\theta/2\pi$ . Hence the frequency per unit area at  $C$  after the  $(n+1)$ th flight is given by:

$$\phi_{n+1}(r^2) = \frac{1}{2\pi} \int_0^{2\pi} \phi_n(r^2 + l^2 - 2rl \cos \theta) d\theta \dots\dots\dots (i).$$

This is the equation, which I shall speak of as the general functional relation between the densities at successive flights. Now assume:  $\phi_n(r^2) = C_n J_0(ur)$ , where  $C_n$  is any undetermined function of  $n$ ,  $l$  and  $u$ , and  $u$  is at present an undetermined variable.

Then by Neumann's Theorem\*:

$$J_0(u \sqrt{r^2 + l^2 - 2rl \cos \theta}) = J_0(ur) J_0(ul) + 2 \sum_1^{\infty} J_t(ur) J_t(ul) \cos t\theta.$$

Hence: 
$$\frac{1}{2\pi} \int_0^{2\pi} C_n J_0(u \sqrt{r^2 + l^2 - 2rl \cos \theta}) d\theta = C_n J_0(ur) J_0(ul) = C_{n+1} J_0(ur), \text{ by (i).}$$

Therefore 
$$C_{n+1} = J_0(ul) C_n.$$

It follows that  $C_n = D \{J_0(ul)\}^n$ , where  $D$  may be any function of  $l$ , but not of  $n$ .

Thus we have: 
$$\phi_n(r^2) = D J_0(ur) \{J_0(ul)\}^n,$$

where we may sum for any or all values of  $u$ .

Now when  $n = 1$ ,  $\phi_1(r^2)$  must be zero, for all values of  $r$  except  $r = l$  to  $l + \tau$ , and it then equals  $N/(2\pi l\tau)$ ,  $\tau$  being very small and  $N$  the total number issuing from the centre of dispersion. We know, however, that†:

$$\int_0^{\infty} du \int_q^p u \rho f(\rho) J_n(u\rho) J_n(ur) d\rho = f(r), \text{ if } q < r < p; \\ = 0, \text{ if } r > p \text{ or } < q.$$

\* C. Neumann, *Theorie der Besselschen Functionen*, S. 65.

† Gray and Mathews, *Treatise on Bessel's Functions*, p. 80.



Now take  $n=0, q=l, p=l+\tau$  and  $f(\rho) = \frac{N}{2\pi l\tau}$ ,

then we have:  $\int_0^\infty ul \frac{N}{2\pi l\tau} J_0(ul) J_0(ur) \tau du = \phi_1(r^2),$

or,  $\phi_1(r^2) = \frac{N}{2\pi} \int_0^\infty u J_0(ul) J_0(ur) du \dots\dots\dots(ii).$

This determines the form of  $D$  and the summation of  $u$ ; for, if we take

$$\phi_n(r^2) = \frac{N}{2\pi} \int_0^\infty u J_0(ur) \{J_0(ul)\}^n du \dots\dots\dots(iii),$$

we satisfy the general functional relation (i) and further the initial equation (ii).

Let  $P_n(r)$  be the probability that an individual after  $n$  flights will be a distance  $r$  or less from the centre of dispersion. Then clearly

$$\begin{aligned} P_n(r) &= 2\pi \int_0^r r dr \phi_{n+1}(r^2) \\ &= N \int_0^r r dr \int_0^\infty u J_0(ur) \{J_0(ul)\}^n du. \end{aligned}$$

But\*  $ur J_0(ur) = \frac{d\{J_1(ur) ur\}}{d(ur)},$

hence  $P_n(r) = N \int_0^\infty du \int_0^{ur} d(ur) \frac{d\{J_1(ur) ur\}}{d(ur)} \frac{\{J_0(ul)\}^n}{u}$   
 $= N \int_0^\infty r J_1(ur) \{J_0(ul)\}^n du,$

or if  $v=ur$ :  $= N \int_0^\infty J_1(v) J_0\left(\frac{lv}{r}\right)^n dv \dots\dots\dots(iv).$

(iv) is Kluyver's fundamental solution, which he reaches by a very different and more general analysis. (iii) is the form of it which best suits my present investigation.

(3) On an expansion in series of the expression for  $\phi_n(r^2)$ . By straightforward but somewhat laborious multiplication it can be shown that:

$$\begin{aligned} \{J_0(2\sqrt{y}) e^y\}^n &= 1 - \frac{1}{4}ny^2 - \frac{1}{8}ny^3 + \frac{(6n-11)n}{192}y^4 \\ &+ \frac{(50n-57)n}{1800}y^5 - \frac{(1892-2125n+270n^2)n}{103,680}y^6, \text{ etc.} \end{aligned}$$

\* Gray and Mathews, *loc. cit.* p. 13.



Hence putting  $2\sqrt{y} = z$ ,

$$\begin{aligned} \{J_0(z)\}^n &= e^{-\frac{1}{4}nz^2} \left\{ 1 - \frac{n}{64}z^4 - \frac{n}{576}z^6 + \frac{(6n-11)n}{192} \frac{z^8}{256} \right. \\ &\quad \left. + \frac{(50n-57)n}{1800} \frac{z^{10}}{1024} - \frac{(1892-2125n+270n^2)n}{103,680} \frac{z^{12}}{4096} - \text{etc.} \right\} \dots\dots(v) \\ &= e^{-\frac{1}{4}nz^2} \{1 - a_4z^4 - a_6z^6 - a_8z^8 - a_{10}z^{10} - a_{12}z^{12} - \text{etc.}\}, \end{aligned}$$

let us write, for brevity. The  $a$ 's are then known coefficients.

Now by (iii)

$$\phi_n(r^2) = \frac{1}{2\pi} \int_0^\infty u J_0(ur) \{J_0(ul)\}^n du.$$

But we know that\* :

$$\int_0^\infty ue^{-\frac{1}{4}nu^2} J_0(ur) du = \frac{2}{nl^2} e^{-r^2/nl^2} \dots\dots\dots(vi).$$

Write :  $\frac{1}{2}nl^2 = \sigma^2 \dots\dots\dots(vii).$

Thus :  $\int_0^\infty ue^{-\frac{1}{2}u^2\sigma^2} J_0(ur) du = \frac{1}{\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} \dots\dots\dots(viii).$

Differentiate (viii)  $s$  times with regard to  $\sigma^2$  :

$$\left(-\frac{1}{2}\right)^s \int_0^\infty u^{2s+1} e^{-\frac{1}{2}u^2\sigma^2} J_0(ur) du = \frac{d^s}{d(\sigma^2)^s} \left\{ \frac{1}{\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} \right\}.$$

Hence, if  $\beta = -2\sigma^2/r^2$ , we have :

$$\begin{aligned} I_{2s} &= \int_0^\infty u^{2s+1} e^{-\frac{1}{2}u^2\sigma^2} J_0(ur) du \\ &= -\frac{2^{2s+1}}{r^{2s+2}} \frac{d^s}{d\beta^s} \left( \frac{1}{\beta} e^{1/\beta} \right) \\ &= -\frac{2^{2s+1}}{r^{2s+2}} i_{2s}, \text{ say.} \end{aligned}$$

We have therefore :

$$\phi_n(r^2) = \frac{N}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} + \frac{N}{2\pi} \sum_{s=2}^\infty \frac{l^{2s} 2^{2s+1} a_{2s} i_{2s}}{r^{2s+2}} \dots\dots\dots(ix),$$

where it remains to evaluate  $i_{2s} = \frac{d^s}{d\beta^s} \left( \frac{1}{\beta} e^{1/\beta} \right)$ .

We find :

$$\begin{aligned} i_4 &= -\frac{1}{8} \left(\frac{r}{\sigma}\right)^6 e^{-r^2/2\sigma^2} \left( 2 - 2\frac{r^2}{\sigma^2} + \frac{1}{4}\frac{r^4}{\sigma^4} \right), \\ i_6 &= -\frac{1}{16} \left(\frac{r}{\sigma}\right)^8 e^{-r^2/2\sigma^2} \left( 6 - 9\frac{r^2}{\sigma^2} + \frac{9}{4}\frac{r^4}{\sigma^4} - \frac{1}{8}\frac{r^6}{\sigma^6} \right), \\ i_8 &= -\frac{1}{32} \left(\frac{r}{\sigma}\right)^{10} e^{-r^2/2\sigma^2} \left( 24 - 48\frac{r^2}{\sigma^2} + 18\frac{r^4}{\sigma^4} - 2\frac{r^6}{\sigma^6} + \frac{1}{16}\frac{r^8}{\sigma^8} \right), \end{aligned}$$

\* Gray and Mathews, *loc. cit.* Eqn. (162), p. 78



$$i_{10} = -\frac{1}{64} \left(\frac{r}{\sigma}\right)^{12} e^{-r^2/2\sigma^2} \left(120 - 300 \frac{r^2}{\sigma^2} + 150 \frac{r^4}{\sigma^4} - 25 \frac{r^6}{\sigma^6} + \frac{25}{16} \frac{r^8}{\sigma^8} - \frac{1}{32} \frac{r^{10}}{\sigma^{10}}\right),$$

$$i_{12} = -\frac{1}{128} \left(\frac{r}{\sigma}\right)^{14} e^{-r^2/2\sigma^2} \left(720 - 2160 \frac{r^2}{\sigma^2} + 1350 \frac{r^4}{\sigma^4} - 300 \frac{r^6}{\sigma^6} + \frac{225}{8} \frac{r^8}{\sigma^8} - \frac{9}{8} \left(\frac{r}{\sigma}\right)^{10} + \frac{1}{64} \left(\frac{r}{\sigma}\right)^{12}\right).$$

Remembering that by (vii)  $l^2 = 2\sigma^2/n$ , we have from (ix)

$$\begin{aligned} \phi_n(r^2) = & \frac{N}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} \left\{ 1 - \frac{1}{4n} \left(2 - 2 \frac{r^2}{\sigma^2} + \frac{1}{4} \frac{r^4}{\sigma^4}\right) - \frac{1}{9n^2} \left(6 - 9 \frac{r^2}{\sigma^2} + \frac{9}{4} \frac{r^4}{\sigma^4} - \frac{1}{8} \frac{r^6}{\sigma^6}\right) \right. \\ & + \frac{6n-11}{192n^3} \left(24 - 48 \frac{r^2}{\sigma^2} + 18 \frac{r^4}{\sigma^4} - 2 \frac{r^6}{\sigma^6} + \frac{1}{16} \frac{r^8}{\sigma^8}\right) \\ & + \frac{50n-57}{1800n^4} \left(120 - 300 \frac{r^2}{\sigma^2} + 150 \frac{r^4}{\sigma^4} - 25 \frac{r^6}{\sigma^6} + \frac{25}{16} \frac{r^8}{\sigma^8} - \frac{1}{32} \frac{r^{10}}{\sigma^{10}}\right) \\ & - \frac{1892 - 2125n + 270n^2}{103,680n^5} \left(720 - 2160 \frac{r^2}{\sigma^2} + 1350 \frac{r^4}{\sigma^4} - 300 \frac{r^6}{\sigma^6} + \frac{225}{8} \frac{r^8}{\sigma^8} - \frac{9}{8} \frac{r^{10}}{\sigma^{10}} + \frac{1}{64} \frac{r^{12}}{\sigma^{12}}\right) \\ & \left. - \text{etc.} \right\} \dots\dots\dots(x). \end{aligned}$$

This is the general expansion for the distribution of the individuals emerging from a centre of dispersion after  $n$  random flights. Clearly if we want to go as far as  $\frac{1}{n^2}$  we must retain terms up to  $(r^2/\sigma^2)^{2n}$ , and the convergence is small for  $n$  small. Thus for the first two or three flights, (x) as far as I have calculated the terms gives poor results, even if they are notwithstanding better than the Rayleigh solution. The arithmetical work required to calculate the ordinates is also severe. If we put  $n = \infty$ , we have

$$\phi_\infty(r^2) = \frac{N}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} \dots\dots\dots(xi),$$

Lord Rayleigh's expression. Now  $\sigma^2 = \frac{1}{2}nl^2$ , hence unless  $l$  becomes indefinitely small as  $n$  becomes indefinitely large the population becomes widely scattered. If the single flight  $l$  be very small, but the total flight  $nl$  be finite, then  $\frac{1}{2}nl^2$  tends to become vanishingly small, or the population remains close to the centre of dispersion. This is really the "flutter" as distinct from the flight.

Examining the solution found it is clear that it may be looked upon as the sum of products of two factors, one series of factors not involving  $r/\sigma$  but only  $n$  and the other not involving  $n$  but only  $r/\sigma$ . Thus we may write

$$\phi_n(r^2) = N (\nu_0 \omega_0 + \nu_2 \omega_2 + \nu_4 \omega_4 + \nu_6 \omega_6 + \dots),$$

where

$$\begin{aligned} \omega_0 &= \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2}, & \nu_0 &= 1, \\ \omega_2 &= \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} \left(1 - \frac{1}{2} \frac{r^2}{\sigma^2}\right), & \nu_2 &= 0, \end{aligned}$$



$$\begin{aligned} \omega_4 &= \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} \left( 2 - 2\frac{r^2}{\sigma^2} + \frac{1}{4}\frac{r^4}{\sigma^4} \right), & \nu_4 &= -\frac{1}{4n}, \\ \omega_6 &= \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} \left( 6 - 9\frac{r^2}{\sigma^2} + \frac{9}{4}\frac{r^4}{\sigma^4} - \frac{1}{8}\frac{r^6}{\sigma^6} \right), & \nu_6 &= -\frac{1}{9n^2}, \\ & \text{etc.} & & \dots\dots\dots(xii). \end{aligned}$$

The  $\omega$ -functions form a series of such special interest that a few of their remarkable properties will be stated in the next section.

(4) *Properties of the  $\omega$ -functions.*

Let us consider the  $p$ th moment round the origin of the 2sth  $\omega$ -function taken over all plane space. We will denote it by  $m_{p,2s}$ . Then

$$\begin{aligned} m_{p,2s} &= \int_0^{2\pi} d\theta \int_0^\infty r dr \omega_{2s} r^p \\ &= 2\pi \int_0^\infty \omega_{2s} r^{p+1} dr \dots\dots\dots(xiii). \end{aligned}$$

Now 
$$\omega_{2s} = -\frac{1}{2\pi\sigma^2} (-\beta)^{s+1} \frac{d^s}{d\beta^s} \left( \frac{1}{\beta} e^{1/\beta} \right) \dots\dots\dots(xiv),$$

and by  $\beta = -2\sigma^2/r^2$  we have 
$$d\beta = \frac{4\sigma^2}{r^3} dr.$$

Hence writing  $p = 2q$  we find

$$m_{2q,2s} = (-1)^{q+s-1} (2\sigma^2)^q \int_{-\infty}^{-0} \beta^{s-q-1} \frac{d^s}{d\beta^s} \left( \frac{1}{\beta} e^{1/\beta} \right) d\beta \dots\dots\dots(xv).$$

Integrate by parts and we have

$$m_{2q,2s} = (-1)^{q+s-1} (2\sigma^2)^q \left[ \left\{ \beta^{s-q-1} \frac{d^{s-1}}{d\beta^{s-1}} \left( \frac{1}{\beta} e^{1/\beta} \right) \right\}_{-0}^{-\infty} - (s-q-1) \int_{-\infty}^{-0} \beta^{s-q-2} \frac{d^{s-1}}{d\beta^{s-1}} \left( \frac{1}{\beta} e^{1/\beta} \right) d\beta \right].$$

The part in curled brackets vanishes at the limits and thus

$$\begin{aligned} m_{2q,2s} &= (-1)^{q+s-2} (2\sigma^2)^q (s-q-1) \int_{-\infty}^{-0} \beta^{s-q-2} \frac{d^{s-1}}{d\beta^{s-1}} \left( \frac{1}{\beta} e^{1/\beta} \right) d\beta \\ &= m_{2q,2s-2} (s-q-1). \end{aligned}$$

Repeating this process we find

$$\begin{aligned} m_{2q,2s} &= (s-1-q)(s-2-q)(s-3-q) \dots (-q) \\ &\times (-1)^{q-1} \times (2\sigma^2)^q \times \int_{-\infty}^{-0} \beta^{-q-2} e^{1/\beta} d\beta \dots\dots\dots(xvi). \end{aligned}$$

The integral is finite and known; hence if  $q$  be less than  $s$  we find for integer values

$$m_{2q,2s} = 0, \quad q < s \dots\dots\dots(xvii).$$

Now consider  $\omega_{2s}$  as made up of two parts,

$$\omega_{2s} = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} \times \chi_{2s} \dots\dots\dots(xviii).$$



Then it is clear that  $\chi_{2q}$ , if  $q$  be less than  $s$ , consists of powers of  $r^2$  less than  $s$ , and therefore

$$\int_0^\infty \omega_{2s} \chi_{2q} r dr = 0.$$

Accordingly a remarkable property holds for the  $\chi$ -function part of the  $\omega$ -function, namely, if  $\chi_{2q}$  and  $\chi_{2q'}$  be two such functions, then it follows that

$$\int_0^\infty e^{-\frac{1}{2}r^2/\sigma^2} \chi_{2q} \chi_{2q'} r dr = 0, \text{ if } q \text{ and } q' \text{ be different, } \dots\dots\dots(\text{xix}).$$

Returning to (xvi), let us put  $q = s$ , then

$$\begin{aligned} m_{2s, 2s} &= -(2\sigma^2)^s \int_{-\infty}^{\infty} \beta^{-s-2} e^{1/\beta} d\beta \\ &= \frac{(2\sigma^2)^s}{2^s} \int_0^\infty x^{2s+1} e^{-\frac{1}{2}x^2} dx, \end{aligned}$$

or, 
$$m_{2s, 2s} = (-1)^s \sigma^{2s} 2^s (|s|)^2 \dots\dots\dots(\text{xx}).$$

Let us now consider the integral over the plane

$$I = 2\pi \int_0^\infty \omega_{2s} \chi_{2s} r dr.$$

Except for the last term in  $\chi_{2s}$ , it will consist of a number of terms having for factors  $m_{2s, 2q}$  with  $q < s$  and these all vanish. The last term in  $\chi_{2s}$  is

$$(-1)^s \frac{1}{2^s} \left(\frac{r}{\sigma}\right)^{2s},$$

and accordingly

$$I = 2\pi \int_0^\infty \omega_{2s} \chi_{2s} r dr = \frac{2\pi (-1)^s}{2^s} \frac{1}{\sigma^{2s}} \int_0^\infty \omega_{2s} r^{2s+1} dr,$$

or by (xx) 
$$I = (|s|)^2 \dots\dots\dots(\text{xxi}).$$

Hence we have the following properties :

(a) The integral all over the plane of distribution of one product of a  $\chi$ -function into an  $\omega$ -function of a different order is zero.

(b) The integral all over the plane of distribution of the product of a  $\chi$ -function into an  $\omega$ -function of the same order is, if  $2s$  be the order, equal to  $(|s|)^2$ .

These properties enable us—as in the case of Bessel's functions or Legendre's functions—to expand any function symmetrical round a centre and a function only of the square of the distance from that centre in  $\omega$ -functions.

Thus let 
$$F(r^2) = \sum_{s=0}^{s=\infty} (b_{2s} \omega_{2s}),$$



multiply by  $\chi_{2s}$  and integrate all over the plane,

$$2\pi \int_0^\infty F(r^2) \chi_{2s} r dr = b_{2s} 2\pi \int_0^\infty \omega_{2s} \chi_{2s} r dr = b_{2s} \{[s]^2\}.$$

Hence 
$$b_{2s} = \frac{2\pi}{\{[s]^2\}} \int_0^\infty F(r^2) \chi_{2s} r dr \dots\dots\dots(\text{xxii}).$$

Now  $\chi_{2s}$  consists of an algebraic series in  $\left(\frac{r}{\sigma}\right)$ . Thus the discovery of the value of the integral  $\int_0^\infty F(r^2) \chi_{2s} r dr$  depends solely on the determination of the odd moments of  $F(r^2)$  between 0 and  $\infty$ . We conclude therefore that an expansion in  $\omega$ -functions involves merely the determination of moments, such as every statistician has been accustomed for years to calculate. This is not the proper occasion to deal fully with the properties of the  $\omega$ -functions, nor to generalise them for odd powers of  $r$ , and to consider the convergency of expansions in terms of  $\omega$ -functions. They will be discussed on another occasion, but the present writer believes that they will be found of not inconsiderable service, not only in statistical problems, but in certain physical problems where intensity round an axis diminishes with the distance.

(5) Two further problems are of service for the theory of dispersion. Suppose

$$F(r^2) = \sum_{s=0}^{s=\infty} (b_{2s} \omega_{2s}).$$

Integrate over the plane and remember that  $\chi_0 = 1$ ,

$$\begin{aligned} 2\pi \int_0^\infty F(r^2) r dr &= \sum_{s=0}^{s=\infty} 2\pi \int_0^\infty b_{2s} \omega_{2s} \chi_0 r dr \\ &= b_0 \dots\dots\dots(\text{xxiii}). \end{aligned}$$

Thus the first coefficient is merely the total volume of the surface  $z = F(r^2)$ , taken over the plane.

Next consider the second moment

$$2\pi \int_0^\infty r^2 F(r^2) r dr = \sum_{s=0}^{s=\infty} 2\pi \int_0^\infty b_{2s} \cdot \omega_{2s} \cdot r^2 \cdot r dr.$$

Every term of the summation vanishes except for  $s=0$  and  $s=1$ , and the left-hand side is the second moment of the function about the axis perpendicular to the plane through the centre = volume  $\times$  (swing-radius)<sup>2</sup> =  $b_0 \times K^2$ , say. Thus:

$$\begin{aligned} b_0 \times K^2 &= \frac{1}{\sigma^2} \int_0^\infty b_0 e^{-\frac{1}{2}r^2/\sigma^2} r^2 dr + \frac{1}{\sigma^2} \int_0^\infty b_2 e^{-\frac{1}{2}r^2/\sigma^2} \left(1 - \frac{1}{2} \frac{r^2}{\sigma^2}\right) r^2 dr \\ &= 2b_0 \sigma^2 + b_2 (2 - 4) \sigma^2 = 2(b_0 - b_2) \sigma^2, \end{aligned}$$

or 
$$b_2 = b_0 \left\{1 - \frac{1}{2} K^2 / \sigma^2\right\} \dots\dots\dots(\text{xxiv}).$$



Thus far no choice has been made of  $\sigma^2$ . If we take  $\sigma^2 = \frac{1}{2}K^2$ , we have  $b_2 = 0$ , or if  $\sigma^2$  be taken half the square of the swing-radius about the axis of the solid of revolution  $z = F(r^2)$ , that is if  $\sigma$  be the swing-radius of the solid about any plane through its axis, then the second term in the expansion of  $F(r^2)$  in  $\omega$ -functions disappears.

We are now able, I think, to grasp the relation of the Rayleigh solution to the complete solution of the random scatter round a centre of dispersion. If  $\phi_n(r^2)$  be the function giving the distribution after  $n$  flights, then  $\phi_n(r^2)$  can be expanded in a series of  $\omega$ -functions, *i.e.*

$$\phi_n(r^2) = b_0\omega_0 + b_2\omega_2 + b_4\omega_4 + \dots + b_{2s}\omega_{2s} + \dots$$

By choosing the  $\sigma^2$  of the  $\omega$ -functions =  $\frac{1}{2}K^2$ , this becomes, since  $b_0$  the volume =  $N$ ,

$$\phi_n(r^2) = \frac{N}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2} \{1 + b_2\chi_2 + \dots + b_{2s}\chi_{2s} + \dots\}.$$

Lord Rayleigh's solution provides the first term of this series, or is the correct solution to two terms in the expansion by  $\omega$ -functions. It possesses the properties (a) that its volume is the same as that of the complete solution, and (b) the mean square deviation from the centre of dispersion is the same, *i.e.*  $2\sigma^2$ , as for the complete solution.

The latter depends upon the fundamental property of the  $\omega$ -functions that  $\int_0^\infty \omega_{2s} r^2 dr = 0$ , if  $s$  be  $> 1$ .

The expansion in  $\omega$ -functions shows us at once that, whatever be the magnitude of  $n$ , the mean square deviation from the centre of dispersion is  $\sqrt{nl}$ , and this gives us readily a *rough* measure of the range of habitat of any species for which  $n$  and  $l$  are approximately known.

Another point may be noted here as to the Rayleigh solution. That solution is the best fitting Gaussian error surface to the distribution, *i.e.* its volume and its standard deviation are the same as those of the actual distribution, whatever  $n$  may be. If we take the section, however, of the distribution through its axis the standard deviation of this according to the Rayleigh solution is  $\sigma = \sqrt{\frac{1}{2}nl}$ , but this is not the standard deviation of the section of the actual distribution, *i.e.* the Rayleigh solution does not give the best fitting normal curve to the section. It gives only the standard deviation corresponding to  $\omega_0$ . It is of some value to note what are the standard deviations of other component  $\omega_{2s}$  terms.

To obtain this we must determine the area and even moments corresponding to any  $\omega_{2s}$  term. Let

$$A_{2s} = \int_0^\infty \omega_{2s} dr = \frac{1}{\pi\sigma} \frac{1}{2^{3/2}} (-1)^{s+\frac{1}{2}} \int_{-1}^\infty \beta^{s-\frac{1}{2}} \frac{d^s}{d\beta^s} \left( \frac{1}{\beta} e^{1/\beta} \right) d\beta,$$



whence integrating by parts :

$$\begin{aligned}
 A_{2s} &= \frac{1}{\pi\sigma} \frac{1}{2^{2s}} (-1)^{\frac{1}{2}} \left(s - \frac{1}{2}\right) \left(s - \frac{3}{2}\right) \left(s - \frac{5}{2}\right) \dots \frac{1}{2} \int_0^{-\infty} \beta^{-2s} e^{1/\beta} d\beta \\
 &= \frac{1}{2\pi\sigma} \left(s - \frac{1}{2}\right) \left(s - \frac{3}{2}\right) \left(s - \frac{5}{2}\right) \dots \frac{1}{2} \int_0^{\infty} e^{-\frac{1}{2}x^2} dx \\
 &= \frac{1}{\sqrt{2\pi\sigma}} \frac{1}{2} \left(s - \frac{1}{2}\right) \left(s - \frac{3}{2}\right) \left(s - \frac{5}{2}\right) \dots \frac{1}{2} \dots\dots\dots (xxv), \\
 & \hspace{25em} s = \text{ or } > 1.
 \end{aligned}$$

If  $s = 0$ ,  $A_0 = \frac{1}{\sqrt{2\pi\sigma}} \frac{1}{2} \dots\dots\dots (xxvi).$

I now take :  $\mu_{2p, 2s} = \int_0^{\infty} \omega_{2s} r^{2p} dr$   
 $= -\frac{1}{2\pi\sigma^2} \int_0^{\infty} r^{2p} (-\beta)^{s+1} \frac{d^s}{d\beta^s} \left(\frac{1}{\beta} e^{1/\beta}\right) dr,$

and find, reducing in the same manner,

$$\mu_{2p, 2s} = \frac{1}{\sqrt{2\pi\sigma}} \frac{\sigma^{2p}}{2} \left(s - p - \frac{1}{2}\right) \left(s - p - \frac{3}{2}\right) \dots \left(-p + \frac{1}{2}\right) \times 1 \cdot 3 \cdot 5 \dots (2p - 1) \dots (xxvii).$$

Clearly :  $m_{2p-1, 2s} = 2\pi \int_0^{\infty} \omega_{2s} r^{2p} dr$   
 $= 2\pi \mu_{2p, 2s}.$

by (xiii), hence :

Or,  $m_{2p-1, 2s} = \sqrt{2\pi} \sigma^{2p-1} \left(s - p - \frac{1}{2}\right) \left(s - p - \frac{3}{2}\right) \dots \left(-p + \frac{1}{2}\right)$   
 $\times 1 \cdot 3 \cdot 5 \dots (2p - 1) \dots\dots\dots (xxviii).$

Thus the odd moments of the  $\omega_{2s}$  functions are known\*.

For the particular case when  $p = 1$  :

$$\mu_{2, 2s} = \frac{1}{\sqrt{2\pi\sigma}} \frac{\sigma^2}{2} \left(s - \frac{3}{2}\right) \left(s - \frac{5}{2}\right) \dots \left(-\frac{1}{2}\right) \dots\dots\dots (xxxii),$$

if  $k_{2s}$  be the swing-radius round the axis of the function  $\omega_{2s}$ . Hence by (xxv)

$$k_{2s}^2 = \frac{\sigma^2 \left(-\frac{1}{2}\right)}{s - \frac{1}{2}} = -\frac{\sigma^2}{2s - 1} \dots\dots\dots (xxxiii).$$

\* If  $x = r/\sigma$  the following finite difference and differential equations are fundamental in the theory of the  $\omega$ -functions :

$$\omega_{2(s+2)} - (2s + 3 - \frac{1}{2}x^2) \omega_{2(s+1)} + (s + 1)^2 \omega_{2s} = 0 \dots\dots\dots (xxix),$$

$$\omega_{2(s+1)} = (s + 1) \omega_{2s} + \frac{1}{2}x \frac{d\omega_{2s}}{dx} \dots\dots\dots (xxx),$$

$$\frac{d^2\omega_{2s}}{dx^2} + \left(x + \frac{1}{x}\right) \frac{d\omega_{2s}}{dx} + 2(s + 1) \omega_{2s} = 0 \dots\dots\dots (xxxii).$$

But the fuller treatment of the  $\omega$ -functions must be deferred.



This is also true for  $s=0$ , as well as any integer value. It follows accordingly that while the total area of any  $\omega$ -function from 0 to  $\infty$  is positive, its  $k$  is negative for values of  $s > 1$ . In other words the negative parts of  $\omega$  are on the whole furthest from the axis. Again the absolute value of  $k_s$  decreases as  $\frac{1}{\sqrt{2s-1}}$  when  $s$  increases, or the higher the  $\omega$ -function the less it contributes relative to its area to the total mean square deviation of the curve.

Applying these results to the curve of scatter given by (x), *i.e.*

$$\phi_n(r^2) = N \left( \omega_0 - \frac{1}{4n} \omega_1 - \frac{1}{9n^2} \omega_2 + \frac{6n-11}{192n^3} \omega_3 + \frac{50n-57}{1800n^4} \omega_4 - \frac{1892-2125n+270n^2}{103,680n^5} \omega_{12} - \text{etc.} \right) \dots\dots\dots(\text{xxxiv}),$$

we have if  $A$  be the whole area and  $k$  the radius,

$$A = \frac{N}{\sqrt{2\pi\sigma}} \frac{1}{2} \left\{ 1 - \frac{3}{16} \frac{1}{n} - \frac{5}{24} \frac{1}{n^2} + \frac{35}{1024} \frac{6n-11}{n^2} + \frac{21}{1280} \frac{50n-57}{n^4} - \frac{77}{49152} \frac{1892-2125n+270n^2}{n^5} - \text{etc.} \right\} \dots\dots\dots(\text{xxxv}),$$

$$Ak^2 = \frac{N}{\sqrt{2\pi\sigma}} \frac{\sigma^2}{2} \left\{ 1 + \frac{1}{16} \frac{1}{n} + \frac{1}{24} \frac{1}{n^2} - \frac{5}{1024} \frac{6n-11}{n^2} - \frac{7}{3840} \frac{50n-57}{n^4} + \frac{7}{49152} \frac{1892-2125n+270n^2}{n^5} + \text{etc.} \right\} \dots\dots\dots(\text{xxxvi}).$$

Hence if we even neglect terms of order  $\frac{1}{n^2}$ , we see that the Rayleigh solution gives too large an area for the curve of section and too small a swing-radius; these values are

Rayleigh area,  $\frac{1}{2} \frac{N}{\sqrt{2\pi\sigma}}$ , Rayleigh swing-radius,  $\sigma$ ,

True area to  $\frac{1}{n}$ ,  $\frac{1}{2} \frac{N}{\sqrt{2\pi\sigma}} \left( 1 - \frac{3}{16} \frac{1}{n} \right)$ ; True swing-radius to  $\frac{1}{n}$ ,  $\sigma \left( 1 + \frac{1}{8n} \right)$ .

Accordingly for  $n$  small the graph of the Rayleigh solution tends to exaggerate the concentration, *i.e.* using it as an approximation we shall somewhat reduce the extreme parts of the curve at the expense of exaggerating those near the centre of dispersion.

While there is no difficulty about determining the curve of distribution when  $n$  is large from (xxxiv), beyond the great labour of dealing with hitherto untabled functions, the investigation becomes very troublesome when  $n$  is small. The functions  $\omega$  are suited in this case to represent the discontinuous functions which actually form the values of  $\phi_n(r^2)$ , but the extreme discontinuity of  $\phi_n(r^2)$  for  $n$



small, compels us to use a very great number of  $\omega$ -functions, and the convergency of (xxxiv) is then small.

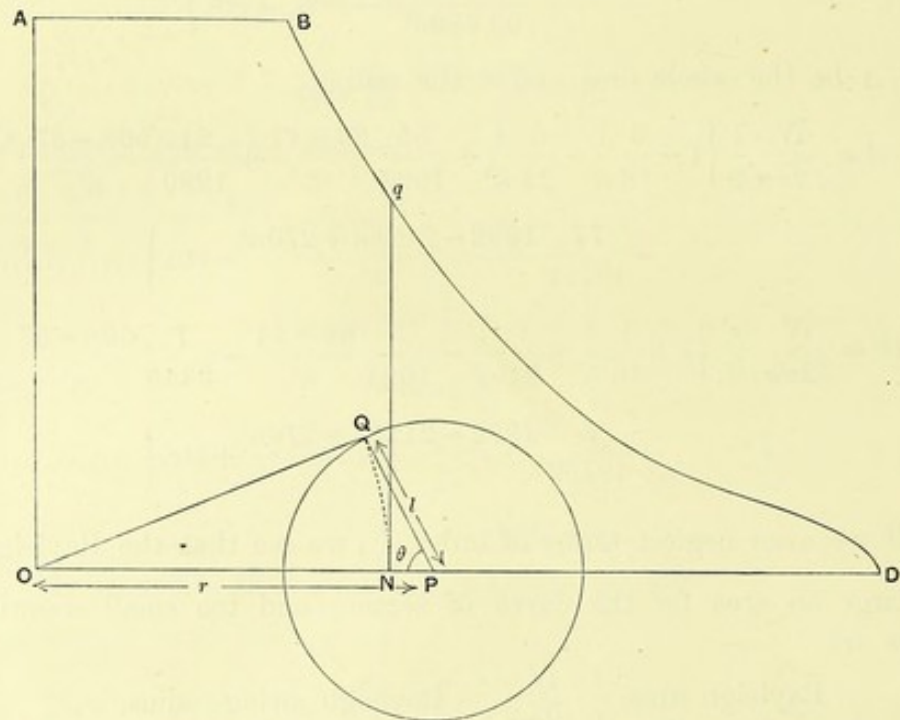
Another method of determining the distribution of the dispersed population has then to be applied to the case of  $n$  small.

(6) *Graphical Solution of the Fundamental Problem for  $n$  small.*

Let us consider the general functional relation (i)

$$\phi_{n+1}(r^2) = \frac{1}{2\pi} \int_0^{2\pi} \phi_n(r^2 + l^2 - 2rl \cos \theta) d\theta.$$

Suppose the graph of  $\phi_n$  from 0 to  $nl$  known. This may be any discontinuous function. From  $nl$  to  $\infty$ , it will be zero. Let  $ABD$  be the graph of  $\phi_n$  and  $OA$  the axis.



$OP = r$ . Round  $P$  describe a circle of radius  $l$ , take the radius  $PQ$ , so that the angle  $OPQ = \theta$ ; then clearly,  $OQ^2 = r^2 + l^2 - 2rl \cos \theta$ ; rotate  $OQ$  round  $O$  down into line  $OD$ , as  $ON$ ; draw the ordinate of the graph  $Nq$ , then we have

$$Nq = \phi_n(r^2 + l^2 - 2rl \cos \theta)$$

and

$$\phi_{n+1}(OP^2) = \frac{1}{2\pi} \int_0^{2\pi} Nq d\theta.$$

Hence if we divide the circle up into a number of equal parts, and determine the ordinates  $Nq$ , corresponding to each of them, we can plot a curve to the base  $2\pi$ , of which the mean ordinate will be  $\phi_{n+1}(OP^2)$ , or the ordinate at  $r$  of the new curve of dispersion for  $\overline{n+1}$  flights. This can be done for a series of values of  $r$  from 0 to  $\overline{n+1}l$  and thus  $\phi_{n+1}(r^2)$  will be determined as a new graph. The area



of the plotted curve which gives any new ordinate can be found mechanically. It will be seen that the process is theoretically straightforward, but very laborious. Thus for the dispersion curve after the fourth flight some 43 points had to be found, and this involved the construction of 43 subsidiary curves and their integration.

There were, of course, graphical difficulties in the construction of the subsidiary curves in the neighbourhood of the asymptotes and various devices had to be used, but at almost every point there were tests of the accuracy of the work. Some of these I shall now notice.

Case (i). The solution for two flights is:

$$\left. \begin{aligned} \phi_2(r^2) &= \frac{N}{\pi^2 r \sqrt{4l^2 - r^2}} & r < 2l \\ &= 0 & r > 2l \end{aligned} \right\} \dots\dots\dots(\text{xxxvii}).$$

The reader will find no difficulty in deducing this directly from the case of  $n=1$ , which corresponds to a narrow zone of radius  $r=l$ , the rest of the plane being unoccupied. Thus:

$$\left. \begin{aligned} \phi_1 &= \frac{N}{2\pi l \epsilon} \text{ from } r=l-\frac{1}{2}\epsilon \text{ to } r=l+\frac{1}{2}\epsilon \\ &= 0 \text{ from } r=0 \text{ to } l-\frac{1}{2}\epsilon \text{ and } r=l+\frac{1}{2}\epsilon \text{ to } \infty \end{aligned} \right\} \dots\dots (\text{xxxvii bis}),$$

$\epsilon$  being taken indefinitely small.

By distributing each element of  $\phi_1$  on the zone round a circle of radius  $l$  we obtain (xxxvii).

The result may be obtained also from (iii) by putting  $n=2$ , *i.e.*

$$\begin{aligned} \phi_2(r^2) &= \frac{N}{2\pi} \int_0^\infty u J_0(ur) \{J_0(ul)\}^2 du, \\ &= \frac{N}{2\pi} \frac{[(2l+r)r(2l-r)r]^{-\frac{1}{2}}}{\sqrt{\pi} 2^{-1} \Pi(-\frac{1}{2})} \text{ from } r=0 \text{ to } 2l, \\ &= 0 \text{ from } r=2l \text{ to } \infty, \end{aligned}$$

from a theorem of de Sonin by putting  $a=r$ ,  $b=c=l$ . Compare Gray and Mathews, p. 239, Ex. 52.

Case (ii). The solution for three flights may be obtained from that for two, by distributing analytically the density given by  $\phi_2$  round circles of radius  $l$  about each point. The resulting double integral is then expressible in elliptic integrals\*. We find:

$$\left. \begin{aligned} \phi_3(r^2) &= \frac{N}{2\pi^2 l} \frac{1}{\sqrt{rl}} \kappa F\left(\frac{\pi}{2}, \kappa\right), \\ &\quad \text{where } \kappa^2 = 16lr/\{(r+l)^2(3l-r)\}, \\ &\quad \quad \quad r > 0 \text{ and } < l; \\ &= \frac{N}{2\pi^2 l} \frac{1}{\sqrt{rl}} F\left(\frac{\pi}{2}, \kappa\right), \\ &\quad \text{where } \kappa^2 = (r+l)^2(3l-r)/(16lr), \\ &\quad \quad \quad r > l \text{ and } < 3l; \\ &= 0 & r > 3l \text{ to } r = \infty \end{aligned} \right\} \dots\dots\dots(\text{xxxviii}).$$

\* This solution, or its equivalent, was first sent me by Mr Geoffrey T. Bennett.



We have here at  $r=l$  a typical instance of the discontinuity.

In Table I. columns (i) and (ii) the calculated ordinates of  $\phi_2$  and  $\phi_3$  are given, the latter having been determined by the use of Legendre's Tables of the Elliptic Integral  $F$ . In these cases, as in the later values of the ordinates of the dispersion curves,  $N$  is taken as unity. The dispersion curves are plotted in Diagrams I. and II. The Rayleigh solution is given in broken line; it will be noticed how very far it is from representing the facts at this early stage of the number of flights. One of the most interesting features of the investigation is to mark the gradual approximation of the discontinuous series of functions to the Gaussian normal curve of errors as the value of  $n$  increases.

The first test of the graphical method of dealing with the problem was to start from the curve for  $n=2$  and construct the graph of  $\phi_r$ . The result was found to be extremely close to the elliptic integral solution obtained by analysis and calculated from Legendre, and this gave us every confidence in the correctness within reasonable limits of the graphical solution, where no such direct verification was possible. After the ordinates of any graph had been found their differences were plotted, and these difference curves submitted to most careful inspection. Larger irregularities led to a reinvestigation of the points, smaller irregularities were smoothed with the spline, and from the final smoothed difference curve the ordinates were corrected.

Another test was now possible. In every case  $2\pi \int_0^\infty \phi_n(r^2) r dr$  ought to be unity. Each ordinate was now multiplied by its  $r$  and a quadrature formula used to find the integral. The integral would usually differ very slightly from unity. Its reciprocal was then used as a factor to each ordinate and the ordinates so modified were the final corrected ordinates of the corresponding graph. The graphs were made on a large scale, and the accompanying Table I., columns (iii)—(vi), gives the ordinates of the dispersion curves from four to seven flights.

Additional tests were as follows:

Since 
$$\phi_{n+1}(r^2) = \frac{1}{2\pi} \int_{2\pi}^{\pi} \phi_n(r^2 + l^2 - 2lr \cos \theta) d\theta,$$

it follows that 
$$\phi_{n+1}(0) = \frac{1}{2\pi} \int_0^{2\pi} \phi_n(l^2) d\theta = \phi_n(l^2),$$

or: The axial ordinate of the  $\overline{n+1}$ th dispersion curve is the ordinate at a distance  $l$ , or a flight, from the centre of the  $n$ th dispersion curve. Table IV. illustrates the degree of accuracy reached here.

The ordinate at  $r=l$  of the seventh curve given by the expansion in  $\omega$ -functions is .0375, and this is precisely the value of the central ordinate of the eighth curve given by the same expansion. Thus the graphical method runs with surprising accuracy into the analytical. The Rayleigh solution gives .0398 for the central ordinate of the eighth curve as against the .0375 of the  $\omega$ -expansion, or the .0378 of the



TABLE I. Ordinates of the Dispersal Curves.

(i) Two Flights (n=2)*		(ii) Three Flights (n=3)†		(iii) Four Flights (n=4)‡		(iv) Five Flights (n=5)‡		(v) Six Flights (n=6)‡		(vi) Seven Flights (n=7)§	
r/l	z <sup>2</sup> /N	r/l	z <sup>2</sup> /N	r/l	z <sup>2</sup> /N	r/l	z <sup>2</sup> /N	r/l	z <sup>2</sup> /N	r/l	z <sup>2</sup> /N
0.00	∞	0.0	-0.585	0.0	∞	0.0	-0.537	0.0	-0.538	0.0	-0.420
0.05	1.0139	0.2	-0.593	0.1	0.1074	0.2	-0.537	0.2	-0.513	0.2	-0.419
0.07	0.7246	0.4	-0.619	0.2	0.0891	0.4	-0.537	0.4	-0.489	0.4	-0.415
0.08	0.6342	0.6	-0.674	0.3	0.0809	0.6	-0.537	0.6	-0.464	0.6	-0.407
0.09	0.5635	0.8	-0.794	0.4	0.0746	0.8	-0.537	0.8	-0.439	0.8	-0.394
0.1	0.5072	0.9	-0.933	0.5	0.0695	1.0	-0.537	1.0	-0.415	1.0	-0.378
0.2	0.2546	0.94	-1.045	0.6	0.0652	1.2	-0.474	1.2	-0.390	1.2	-0.358
0.3	0.1708	0.98	-1.285 (†)	0.7	0.0615	1.4	-0.413	1.4	-0.364	1.4	-0.335
0.4	0.1293	1.00	∞	0.8	0.0585	1.6	-0.363	1.6	-0.338	1.6	-0.310
0.5	0.1046	1.02	-1.265 (†)	0.9	0.0560	1.8	-0.322	1.8	-0.309	1.8	-0.283
0.6	0.0885	1.06	-0.990	1.0	0.0537	2.0	-0.285	2.0	-0.277	2.0	-0.256
0.7	0.0773	1.1	-0.856	1.1	0.0516	2.2	-0.250	2.2	-0.240	2.2	-0.229
0.8	0.0691	1.2	-0.672	1.2	0.0496	2.4	-0.217	2.4	-0.206	2.4	-0.202
0.9	0.0630	1.4	-0.488	1.3	0.0477	2.6	-0.185	2.6	-0.177	2.6	-0.177
1.0	0.0585	1.6	-0.386	1.4	0.0459	2.8	-0.153	2.8	-0.150	2.8	-0.153
1.1	0.0551	1.8	-0.319	1.5	0.0442	3.0	-0.121	3.0	-0.126	3.0	-0.130
1.2	0.0528	2.0	-0.270	1.6	0.0426	3.2	-0.093	3.2	-0.106	3.2	-0.110
1.3	0.0513	2.2	-0.233	1.7	0.0410	3.4	-0.070	3.4	-0.087	3.4	-0.092
1.4	0.0507	2.4	-0.205	1.8	0.0396	3.6	-0.053	3.6	-0.070	3.6	-0.076
1.5	0.0511	2.6	-0.181	1.9	0.0382	3.8	-0.039	3.8	-0.055	3.8	-0.062
1.6	0.0528	2.8	-0.162	2.0	0.0369	4.0	-0.029	4.0	-0.042	4.0	-0.050
1.7	0.0566	3.0	-0.146	2.02	0.0348	4.2	-0.021	4.2	-0.032	4.2	-0.039
1.8	0.0646			2.06	0.0308	4.4	-0.013	4.4	-0.023	4.4	-0.031
1.9	0.0854			2.1	0.0284	4.6	-0.007	4.6	-0.016	4.6	-0.024
1.94	0.1074			2.2	0.0242	4.8	-0.002	4.8	-0.011	4.8	-0.018
1.96	0.1299			2.3	0.0210	5.0	-0.000	5.0	-0.007	5.0	-0.013
1.98	0.1815			2.4	0.0185	5.2	-0.004	5.2	-0.004	5.2	-0.009
1.99	0.2552			2.5	0.0165	5.4	-0.002	5.4	-0.002	5.4	-0.006
2.00	∞			2.6	0.0148	5.6	-0.001	5.6	-0.001	5.6	-0.004
				2.7	0.0133	5.8	-0.000	5.8	-0.000	5.8	-0.003
				2.8	0.0120	6.0	-0.000	6.0	-0.000	6.0	-0.001
				2.9	0.0106			6.2	-0.000	6.2	-0.000
				3.0	0.0095			6.4	-0.000	6.4	-0.000
				3.2	0.0076			6.6	-0.000	6.6	-0.000
				3.4	0.0059			6.8	-0.000	6.8	-0.000
				3.6	0.0043			7.0	-0.000	7.0	-0.000
				3.8	0.0027						
				3.9	0.0019						
				3.92	0.0017						
				3.94	0.0015						
				3.96	0.0012						
				3.98	0.0008						
				4.00	0.0000						

\* Calculated from formula.  
 † Calculated from Tables of Elliptic Integrals.  
 ‡ Values to the left calculated graphically, values to the right found from the  $\omega$ -function expansion for comparison, the values in brackets are those given by the Rayleigh solution.  
 § Calculated graphically.



TABLE II. *Values of the  $\omega$ -functions.*

$r/\sigma$	$\sigma^2 \omega_0$	$\sigma^2 \omega_2$	$\sigma^2 \omega_4$	$\sigma^2 \omega_6$	$\sigma^2 \omega_8$	$\sigma^2 \omega_{10}$	$\sigma^2 \omega_{12}$
0	+ 159,1550	+ 159,1550	+ 318,3100	+ 954,9300	+ 3819,7200	+ 19098,6000	+ 114591,6000
1	+ 158,3611	+ 157,5693	+ 313,5589	+ 935,9497	+ 3724,9378	+ 18530,6202	+ 110620,7235
2	+ 156,0035	+ 152,8834	+ 299,5891	+ 880,4201	+ 3449,0302	+ 16885,5699	+ 99177,8011
3	+ 152,1517	+ 145,3049	+ 277,2242	+ 792,4264	+ 3016,3079	+ 14332,2150	+ 81601,7164
4	+ 146,9185	+ 135,1650	+ 247,7634	+ 678,3356	+ 2464,2124	+ 11127,4046	+ 59905,9470
5	+ 140,4537	+ 122,8970	+ 212,8751	+ 546,1784	+ 1839,0999	+ 7583,1582	+ 36489,3471
6	+ 132,9374	+ 109,0087	+ 174,4670	+ 404,8965	+ 1191,1906	+ 4027,9574	+ 13802,7339
7	+ 124,5713	+ 94,0513	+ 134,5401	+ 263,5330	+ 569,3039	+ 767,7288	+ 5975,6743
8	+ 115,5702	+ 78,5877	+ 95,0449	+ 130,4593	+ 016,0640	+ 1947,9139	+ 21205,3207
9	+ 106,1526	+ 63,1608	+ 57,7497	+ 012,7166	+ 435,8815	+ 3949,8656	+ 30951,7904
10	+ 096,5323	+ 48,2662	+ 24,1331	+ 084,4658	+ 766,2251	+ 5161,4614	+ 35039,7166
12	+ 077,4690	+ 21,6913	+ 028,0128	+ 206,6600	+ 1045,7098	+ 5351,9170	+ 28874,9613
14	+ 059,7326	+ 001,1947	+ 057,3194	+ 235,2026	+ 900,0451	+ 3455,1198	+ 12119,1731
16	+ 044,2510	+ 012,3903	+ 065,5623	+ 194,3306	+ 521,5103	+ 916,7705	+ 4126,7483
18	+ 031,4966	+ 019,5279	+ 058,4451	+ 119,4328	+ 116,5429	+ 1050,8391	+ 12770,4428
20	+ 021,5393	+ 021,5393	+ 043,0786	+ 043,0786	+ 172,3144	+ 1895,4584	+ 12751,2656
22	+ 014,1523	+ 020,0963	+ 025,8081	+ 013,8001	+ 295,4776	+ 1723,4411	+ 7400,1858
24	+ 008,9341	+ 016,7961	+ 010,9496	+ 043,9712	+ 279,7081	+ 1008,2741	+ 1194,4811
26	+ 005,4188	+ 012,8967	+ 000,5180	+ 050,7478	+ 188,3692	+ 246,6709	+ 2829,6050
28	+ 003,1578	+ 009,2208	+ 005,3253	+ 042,6344	+ 083,3863	+ 258,5489	+ 3915,1831
30	+ 001,7680	+ 006,1880	+ 007,5140	+ 028,5090	+ 003,6465	+ 439,7348	+ 2949,4384
32	+ 000,9511	+ 003,9185	+ 007,3562	+ 014,7914	+ 038,3979	+ 385,6461	+ 1308,1481
34	+ 000,4916	+ 002,3498	+ 006,0410	+ 004,6874	+ 048,6501	+ 231,6523	+ 007,0283

I owe this preliminary table of  $\omega$ -functions to the kindness of Dr Alice Lee. Much more elaborate tables will have to be calculated, if as I anticipate the  $\omega$ -functions are found valuable for other purposes. The present table suffices to indicate their general numerical character, and enables one to calculate some of the quantities needed in the present memoir.

graphical construction. The fact that the central ordinate of the sixth curve is almost identical with the ordinate of the fourth curve at  $r=l$ , seems conclusive as to the general accuracy of the process.

The above test of the general accuracy of Mr Blakeman's graphical work is only a part of the still more sufficient test that in the seventh curve the graph and the  $\omega$ -expansion practically coincide. See Diagram VI. After  $r=5l$  the two curves cannot be distinguished, and between  $r=0$  and  $3l$  the deviation is probably as much due to the neglect of higher  $\omega$ -functions as to errors in the graphical treatment.

Another method adopted by Mr Blakeman for testing the accuracy of his graphical work, especially at the end of the range, was to obtain expansions to  $\phi_n(r^2)$ , when  $r$  does not differ much from  $nl$ ,  $=nl - \xi$ , say, where  $\xi$  is supposed small. If  $f_n(\xi) = \phi_n((nl - \xi)^2)$ , then generally for  $\xi$  small :

$$f_n(\xi) = \frac{N}{l^2 (\sqrt{2})^{n+3} \pi^{n-1} \sqrt{n}} \left(\frac{\xi}{l}\right)^{(n-3)/2} I_1 I_2 I_3 \dots I_{n-3} \dots \dots \dots (\text{xxxix}),$$

where 
$$I_q = \int_0^{\pi/2} \cos^q \theta d\theta = \Gamma\left(\frac{1}{2}(n+1)\right) \Gamma\left(\frac{3}{2}\right) / \Gamma\left(\frac{1}{2}(n+2)\right).$$



TABLE III. *Table of the  $\nu$  constants.*

$m=1$	$n=6$	$n=7$	$n=8$
$\nu_4$	-041,666,667	-035,714,286	-031,250,000
$\nu_6$	-003,086,420	-002,267,574	-001,736,111
$\nu_8$	000,602,816	000,470,724	000,376,383
$\nu_{10}$	000,104,167	000,067,796	000,046,522
$\nu_{12}$	+000,001,412	-000,000,142	-000,000,639

*Table of the  $N$  constants.*

$m=5$	$n=6$	$n=7$	$n=8$
$N_4$	-008,333,333	-007,142,857	-006,250,000
$N_6$	-000,123,457	-000,090,703	-000,069,444
$N_8$	000,032,600	000,024,174	000,018,636
$N_{10}$	000,000,990	000,000,627	000,000,422
$N_{12}$	-000,000,078	-000,000,047	-000,000,033
$m=10$	$n=6$	$n=7$	$n=8$
$N_4$	-004,166,667	-003,571,429	-003,125,000
$N_6$	-000,030,864	-000,022,676	-000,017,361
$N_8$	000,008,415	000,006,211	000,004,771
$N_{10}$	000,000,126	000,000,082	000,000,053
$N_{12}$	-000,000,011	-000,000,007	-000,000,006
$m=20$	$n=6$	$n=7$	$n=8$
$N_4$	-002,083,333	-001,785,714	-001,562,500
$N_6$	-000,007,716	-000,005,669	-000,004,340
$N_8$	000,002,137	000,001,574	000,001,207
$N_{10}$	000,000,016	000,000,010	000,000,007
$N_{12}$	-000,000,0014	-000,000,0009	-000,000,0006

TABLE IV. *Central ordinates and ordinates at  $r=l$ .*

No. of Flights	Central ordinate	Ordinate at $r=l$
First .....	0	$\infty$
Second .....	$\infty$	0585
Third .....	0585	$\infty$
Fourth .....	$\infty$	0537
Fifth .....	0537	0537
Sixth .....	0538	0415
Seventh .....	0415	0378



Hence: 
$$f_n(\xi) = \frac{N(\xi/l)^{\frac{1}{2}(n-3)}}{(\sqrt{2\pi})^{n+1} l^2 \sqrt{n} 2^{n-2} \Gamma(\frac{1}{2}(n-1))} \dots\dots\dots(\text{xxxix bis}).$$

This can be proved by induction.

For most of the cases more approximate formulae still were deduced. Thus:

$$f_3(\xi) = \frac{1}{8\sqrt{3}l^2\pi^2} \left( 1 + \frac{1}{2} \frac{\xi}{l} + \frac{5}{24} \frac{\xi^2}{l^2} \right) \dots\dots\dots(\text{xl}),$$

$$f_4(\xi) = \frac{1}{16\sqrt{2}l^2\pi^3} \sqrt{\frac{\xi}{l}} \left( 1 + \frac{21}{48} \frac{\xi}{l} + \frac{557}{1536} \frac{\xi^2}{l^2} \right) \dots\dots\dots(\text{xli}),$$

$$f_5(\xi) = \frac{1}{64\sqrt{5}l^2\pi^3} \left( \frac{\xi}{l} \right) \left( 1 + \frac{2}{5} \frac{\xi}{l} + \frac{507}{2560} \frac{\xi^2}{l^2} \right) \dots\dots\dots(\text{xlii}),$$

$$f_6(\xi) = \frac{1}{96\sqrt{12}l^2\pi^4} \left( \frac{\xi}{l} \right)^{3/2} \left( 1 + \frac{9}{24} \frac{\xi}{l} + \dots \right) \dots\dots\dots(\text{xliii}),$$

after which the first term only as given by (xxxix) is sufficient. It will be observed that after  $\phi_n(r^2)$ , the curve touches at  $r=nl$  or  $\xi=0$ , and the contact becomes higher and higher as  $n$  increases. Thus, although short of  $n = \infty$ , there is no real asymptoting to the axis, still  $\phi_n(r^2)$  for  $n > 5$  not only vanishes for  $r=nl$ , but has increasingly higher contact as  $n$  increases. This explains how the Gaussian curve can fairly well represent the state of affairs towards the end of the dispersal range, if  $n$  is  $> 5$ .

Mr Blakeman found that the ends of the range for the various cases ran closely into the curves (xl) to (xliii), and they were tested and, if needful, corrected by these formulae.

Thus the whole graphical work was kept in check, and, I think, we may be confident that the true forms of the dispersal curves for  $n=4$  to  $7$  are really given by our diagrams and tables.

(7) We may note a few features of these curves.

*Dispersal Curve for Two Flights (Diagram I).*

There is no discontinuity in the solution from  $r=0$  to  $2l$ , the range within which all individuals fall. The curve asymptotes to the vertical at the axis and at  $r=2l$ . Of course, while the density becomes infinite, the number on any small area near  $r=0$  or  $r=2l$ , is finite. Thus the number between the circles of radii  $r_1$  and  $r_2$  is

$$\frac{2N}{\pi} \left( \sin^{-1} \frac{r_2}{2l} - \sin^{-1} \frac{r_1}{2l} \right).$$

If  $r_1=0$  and  $r_2=\epsilon_1$ , where  $\epsilon_1$  is small, the number  $\nu_1$  within the small circle of radius  $\epsilon_1$  at the centre of dispersion =  $N\epsilon_1/(\pi l)$ . If  $r_2=r_1+\epsilon_2$ , the number lying on the zone of breadth  $\epsilon_2$  is  $\frac{N\epsilon_2}{\pi l} \left( 1 - \frac{r_1^2}{4l^2} \right)^{-\frac{1}{2}}$ , and this if  $r_1=2l-\epsilon_2$ , is  $\nu_2 = \frac{N}{\pi} \sqrt{\frac{\epsilon_2}{l}}$ . At the position of



minimum density  $r_1 = \sqrt{2}l$ , and the number on the zone  $r_1$  to  $r_1 + \epsilon_3$  is  $\nu_3 = N\epsilon_3/(\pi l\sqrt{1/2})$ . Hence it follows that the numbers on narrow zones  $\epsilon_1, \epsilon_2, \epsilon_3$  in breadth, of equal areas  $\pi\epsilon_1^2 = \pi 4l\epsilon_2 = \pi 2\sqrt{2}l\epsilon_3$ , are given by

$$N\epsilon_1/(\pi l), \quad N\sqrt{\epsilon_2}/(\pi\sqrt{l}), \quad \text{and} \quad N\epsilon_3\sqrt{2}/(\pi l),$$

or in the ratio

$$N\epsilon_1/(\pi l) : \frac{1}{2}N\epsilon_2/(\pi l), \quad \frac{1}{2}N\epsilon_3/(\pi l) \times \frac{\epsilon_1}{l}.$$

Thus the total population on a small area at the centre of dispersion is twice that on an equal area at the periphery of the distribution, and at both indefinitely greater than on an equal belt at the distance of minimum density. The same point can be indicated in another way. From  $r=0$  to  $r=\frac{1}{2}l$  is  $\frac{1}{16}$  of the total area occupied after dispersion, it contains  $\cdot 16N$  or about  $\frac{1}{6}$  of the total population; from  $r=\frac{3}{2}l$  to  $r=2l$  is  $\frac{7}{16}$  of the total area, it contains  $\cdot 54N$ . In other words the half of the area nearest and farthest from the centre of dispersion contains  $\frac{7}{16}$  of the dispersed population; the "middle" half of the area contains only  $\frac{3}{16}$  of the population. The nature of the distribution is thus extremely different from that given by the rotation of the Gaussian curve about its axis for this small number of flights. For in the Gaussian case if the central area  $\pi\epsilon_1^2 = 2\pi r_1\epsilon_2$ , the area of the zone at distance  $r_1$ , the population on the centre patch is  $\frac{1}{2}N\epsilon_1^2/\sigma^2$  and on the zone is

$$\frac{1}{2}N\epsilon_2^2/\sigma^2 \times \epsilon^{-\frac{1}{2}r_1^2/\sigma^2},$$

which is always less and diminishes *continuously* with increase of  $r_1$ . Thus the Rayleigh solution fails in this, as in the next three cases, not only to give the form of the curve at dispersion, but to indicate that the dispersed populations on zones of equal area round the centre do not decrease uniformly in number.

*Dispersal Curve for Three Flights (Diagram II.).*

The solution is discontinuous at  $r=l$ . The density is here infinite, but has become finite at the origin. There is no discontinuity at  $r=2l$ , but at the end of the range the density drops suddenly from a finite value to zero. Thus the integral of the Bessel function product (see Eqn. (iii)) is discontinuous at two points. The Rayleigh solution is still widely divergent from the true curve of dispersal.

*Dispersal Curve for Four Flights (Diagram III.).*

By the rule already referred to (p. 18) the infinite density has returned to the origin. There are only two points of discontinuity, *i.e.*, at  $r=l$  and  $r=4l$  the end of the range, at both of which there is an abrupt change in the slope of the curve. The density at the end of the range is now zero and will remain so, but the dispersal curve rises at right angles to the axis. The true dispersal curve is bending round somewhat to the Rayleigh curve, but the latter is not even yet a rough approximation to the facts.



*Dispersal Curve for Five Flights (Diagram IV.).*

All infinite densities have now finally disappeared. The density vanishes at the end of the range, but the dispersal curve makes a finite angle with the horizontal axis. There is a marked discontinuity of slope at  $r=l$ ; a still more noteworthy feature is that from  $r=0$  to  $r=l$  the graphical construction, however carefully reinvestigated, did not permit of our considering the curve to be anything but a *straight* line. If this could be verified from the analytical expression

$$\phi_4(r^2) = \frac{N}{2\pi} \int_0^\infty u J_0(ur) \{J_0(ul)\}^4 du$$

by showing that the integral is independent of  $r$  from 0 to  $l$  it would be of much interest. Even if it be not absolutely true, it exemplifies the extraordinary power of such integrals of  $J$  products to give extremely close approximations to such simple forms as horizontal lines.

The approach of the Rayleigh curve to the result is now more noticeable.

*Dispersal Curve for Six Flights (Diagram V.).*

There is contact now of the first order at the end of the range. From  $r=0$  to  $r=l$  the curve of dispersal appears to be a sloping straight line tangential to the continuous curve from  $r=l$  to  $r=6l$ . No other discontinuity of a low order is now visible. The curve, except for the finite slope at  $r=0$ , is becoming much more of the Gaussian form. It runs fairly closely to the solution in  $\omega$ -functions up to  $\omega_{22}$ , in fact is not separable at the extreme part of the range, where the Rayleigh curve still gives finite ordinates beyond the possible range.

*Dispersal Curve for Seven Flights (Diagram VI.).*

All sign of discontinuity has gone, the curve is horizontal at the centre of dispersion and might be easily mistaken for a normal curve of errors. The expansion in  $\omega$ -functions represents the result within the limits almost of constructional error. It was not thought necessary to continue the graphical work beyond this stage. We may conclude that:

*The deviation of the Rayleigh solution for seven and more flights from the true dispersal curve is practically the same as its deviation from the solution in  $\omega$ -functions when five terms of that series are retained.*

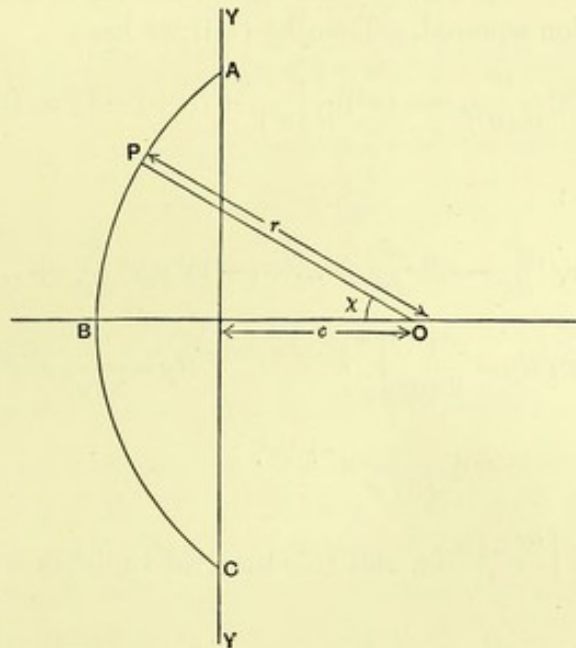
This I think completes the full solution of the fundamental problem. The dispersal curves for the cases of 2 to 7 flights are given in the Table I. of ordinates and the Diagrams I. to VI. For higher values the  $\omega$ -function series gives the solution. This solution could be applied to calculate the ordinates of the dispersal curve for fewer flights than 6 or 7, but several more  $\omega$ -functions would have to be used and the arithmetical work—especially while these functions are as yet untabled\*—then becomes somewhat severe.

\* Table II. provides a preliminary series of values of the  $\omega$  functions.



(8) *Secondary Migration Problems.*

Problem I. *On one side of a straight line there is supposed to be a uniform distribution of habitats; on the other at starting no habitats. To investigate the distribution in the unoccupied area after one migration. Each individual is supposed to take  $n$ -flights to the new habitat.*



Let  $YY$  be the straight line and  $O$  a point at distance  $c$  from it on the unoccupied side of it. Let  $N$  be the average density per unit of area on the occupied side. Then after an  $n$ -flight migration, the contribution from  $P$  (co-ordinates  $r, \chi$ ) at  $O$  will be  $Nr\delta\chi\delta r\phi_n(r^2)$ , and integrating this all round a circle of radius  $r$  from  $A$  to  $C$  within the occupied area, we have for the quantity  $F_n(c)$  at  $O$

$$F_n(c) = 2N \int_c^\infty \int_0^{\cos^{-1}c/r} \phi_n(r^2) r d\chi dr$$

$$= 2N \int_c^\infty \cos^{-1}c/r \phi_n(r^2) r dr.$$

Hence 
$$\frac{dF_n(c)}{dc} = -2N \int_c^\infty \frac{\phi_n(r^2) r dr}{\sqrt{r^2 - c^2}} = -2N \int_0^\infty \phi_n(c^2 + y^2) dy \dots\dots(xliv).$$

The evaluation of this integral needs a further consideration of the  $\omega$ -functions. By (xiv)

$$\omega_{2s} = -\frac{1}{2\pi\sigma^2} (-\beta)^{s+1} \frac{d^s}{d\beta^s} \left( \frac{1}{\beta} e^{1/\beta} \right), \text{ where } \beta = -2\sigma^2/r^2.$$



Transfer the differentiations from  $\beta$  to  $\sigma^2$  and we have:

$$\begin{aligned} \omega_{2s} &= \frac{(-1)^s}{2\pi} (\sigma^2)^s \frac{d^s}{d(\sigma^2)^s} \left( \frac{1}{\sigma^2} e^{-\frac{1}{2}\gamma^2/\sigma^2} \right) \\ &= (-1)^s (\sigma^2)^s \frac{d^s \omega_0}{d(\sigma^2)^s} \dots\dots\dots(xlv), \end{aligned}$$

or, all the  $\omega$ -functions can be found by differentiating the first  $\omega$ -function with regard to the standard-deviation squared. Then by (xii) we have

$$\phi_n(\gamma^2) = \left\{ 1 + \nu_4 (\sigma^2)^2 \frac{d^2}{d(\sigma^2)^2} - \nu_6 (\sigma^2)^3 \frac{d^3}{d(\sigma^2)^3} + \dots + (-1)^s \nu_{2s} (\sigma^2)^s \frac{d^s}{d(\sigma^2)^s} + \dots \right\} \omega_0 \dots\dots(xlvi).$$

Thus, if we put  $\sigma^2 = t$ :

$$\frac{dF_n(c)}{dc} = -2N \left( 1 + \nu_4 t^2 \frac{d^2}{dt^2} - \nu_6 t^3 \frac{d^3}{dt^3} + \dots + (-1)^s \nu_{2s} t^s \frac{d^s}{dt^s} + \dots \right) \times \int_0^\infty \omega_0(c^2 + y^2) dy.$$

But: 
$$\int_0^\infty \omega_0(c^2 + y^2) dy = \frac{1}{2\pi\sigma^2} \int_0^\infty e^{-(c^2+y^2)/2\sigma^2} dy = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}c^2/\sigma^2} \frac{1}{2} \sqrt{2\pi} \sigma$$

$$= \frac{1}{2} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2}c^2/\sigma^2}.$$

Now  $\int_x^c e^{-\frac{1}{2}c^2/\sigma^2} dc = \sigma \int_x^{c/\sigma} e^{-\frac{1}{2}x^2} dx$ , and this integral vanishes when  $c = \infty$ . Hence

$$F_n(c) = \frac{N}{\sqrt{2\pi}} \left( 1 + \nu_4 t^2 \frac{d^2}{dt^2} - \nu_6 t^3 \frac{d^3}{dt^3} + \dots + (-1)^s \nu_{2s} t^s \frac{d^s}{dt^s} + \dots \right) \int_{c/\sigma}^\infty e^{-\frac{1}{2}x^2} dx \dots\dots(xlvii).$$

Since  $F_n(c)$  clearly vanishes for  $c$  infinite, it is not needful to introduce a constant.

It remains accordingly to determine the successive differentials of the integral with regard to  $t$ . Call the integral  $i$ ; then, if  $\eta = c/\sigma = c/\sqrt{t}$ ,

$$\frac{di}{dt} = -e^{-\frac{1}{2}\eta^2} \frac{d\eta}{dt} = \frac{1}{2} \frac{c}{t^{\frac{3}{2}}} e^{-c^2/2t} = \pi \frac{c}{\sigma} \omega_0 = \pi c \omega_0 / \sqrt{t}.$$

By (xlv) we know that  $d^s \omega_0 / dt^s = (-1)^s \omega_{2s} / t^s$ . Hence differentiating  $s-1$  times we have:

$$\begin{aligned} \frac{d^s i}{dt^s} &= \frac{\pi c}{\sqrt{t}} \left( \frac{d^{s-1} \omega_0}{dt^{s-1}} - (s-1) \frac{d^{s-2} \omega_0}{dt^{s-2}} \frac{1}{2t} + \frac{(s-1)(s-2)}{2!} \frac{d^{s-3} \omega_0}{dt^{s-3}} \frac{1.3}{2.2} \frac{1}{t^2} - \text{etc.} \right) \\ &= \frac{(-1)^{s-1} \pi c}{t^{s-\frac{1}{2}}} \left( \omega_{2(s-1)} + (s-1) \omega_{2(s-2)} \frac{1}{2} + \frac{(s-1)(s-2)}{1.2} \omega_{2(s-3)} \frac{1.3}{2.2} \right. \\ &\quad \left. + \frac{(s-1)(s-2)(s-3)}{1.2.3} \omega_{2(s-4)} \frac{1.3.5}{2^3} + \dots \text{etc.} \right). \end{aligned}$$



$$\text{Thus } t^s \frac{d^s i}{dt^s} = (-1)^{s-1} \pi \frac{c}{\sigma} \left( \sigma^2 \omega_{2(s-1)} + \frac{(s-1)}{1! 2} \sigma^2 \omega_{2(s-2)} + \frac{(s-1)(s-2)}{2! 2^2} 1.3. \sigma^2 \omega_{2(s-3)} \right. \\ \left. + \frac{(s-1)(s-2)(s-3)}{3! 2^3} 1.3.5 \sigma^2 \omega_{2(s-4)} + \text{etc.} \right) \dots\dots\dots (\text{xlvi})$$

Substituting in (xlvi) we have, if  $\bar{\psi}(\eta) = \frac{1}{\sqrt{2\pi}} \int_{\eta}^{\infty} e^{-\frac{1}{2}x^2} dx \dots\dots\dots (\text{xlvii}),$

$$F_n(c) = N \left[ \bar{\psi} \left( \frac{c}{\sigma} \right) - \sqrt{\frac{\pi}{2}} \frac{c}{\sigma} \left\{ \sigma^2 \omega_0 \left( \frac{1}{2} \nu_4 + \frac{3}{4} \nu_6 + \frac{1}{8} \nu_8 + \frac{1}{16} \nu_{10} + \frac{9}{64} \nu_{12} \right) \right. \right. \\ \left. \left. + \sigma^2 \omega_2 \left( \nu_4 + \nu_6 + \frac{9}{4} \nu_8 + \frac{1}{2} \nu_{10} + \frac{5}{16} \nu_{12} \right) \right. \right. \\ \left. \left. + \sigma^2 \omega_4 \left( \nu_6 + \frac{3}{2} \nu_8 + \frac{9}{2} \nu_{10} + \frac{7}{4} \nu_{12} \right) \right. \right. \\ \left. \left. + \sigma^2 \omega_6 \left( \nu_8 + 2\nu_{10} + \frac{1}{2} \nu_{12} \right) \right. \right. \\ \left. \left. + \sigma^2 \omega_8 \left( \nu_{10} + \frac{1}{2} \nu_{12} \right) + \sigma^2 \omega_{10} \nu_{12} + \dots \right\} \right] \dots\dots\dots (1),$$

as far as coefficients of the order  $\nu_{12}$  and functions of order  $\omega_{10}$ .

This is the solution in  $\omega$ -functions. Table III., p. 21, gives the values of the  $\nu$ 's for certain values of  $n$ , and Table II., p. 20, is a preliminary table of the  $\omega$ -functions. These will enable us to readily find the values of  $F_n(c)$ . I have done this for the case of  $n=6$  and  $n=7$ , which will suffice to illustrate the character of these curves.  $\bar{\psi}(c/\sigma)$  can be found at once from Tables of the probability integral. It is drawn with a broken line in Diagram VII. and is the Rayleigh solution for this case. I term  $F_n(c)$  an "infiltration curve" of the first order.

Substituting the values of the  $\nu$ 's from Table III., we have for  $n=6$ :

$$F_6(c)/N = \bar{\psi} \left( \frac{c}{\sigma} \right) + \frac{c}{\sigma} \left\{ .026,712,414 (\sigma^2 \omega_0) + .053,325,539 (\sigma^2 \omega_2) \right. \\ \left. + .002,114,303 (\sigma^2 \omega_4) - .001,029,898 (\sigma^2 \omega_6) \right. \\ \left. - .000,134,978 (\sigma^2 \omega_8) - .000,001,770 (\sigma^2 \omega_{10}) + \dots \right\},$$

and for  $n=7$ :

$$F_7(c)/N = \bar{\psi} \left( \frac{c}{\sigma} \right) + \frac{c}{\sigma} \left\{ .022,850,925 (\sigma^2 \omega_0) + .045,644,347 (\sigma^2 \omega_2) \right. \\ \left. + .001,578,008 (\sigma^2 \omega_4) - .000,758,570 (\sigma^2 \omega_6) \right. \\ \left. - .000,084,525 (\sigma^2 \omega_8) + .000,000,178 (\sigma^2 \omega_{10}) + \dots \right\}.$$

The first term  $\bar{\psi} \left( \frac{c}{\sigma} \right)$  is the ogive curve already drawn corresponding to the Rayleigh solution. We see at once that the term  $\sigma^2 \omega_{10}$  will not affect the fourth place of decimals.



TABLE V. *Ordinates of Infiltration Curve over straight Boundary.*

$+c/\sigma$	$n=6$	$n=7$	$n=\infty$	$-c/\sigma$	$n=6$	$n=7$	$n=\infty$
0	.5000	.5000	.5000	—	—	—	—
.1	.4614	.4612	.4602	-.1	.5386	.5388	.5398
.2	.4231	.4228	.4207	-.2	.5769	.5772	.5793
.3	.3854	.3850	.3821	-.3	.6146	.6150	.6179
.4	.3488	.3483	.3446	-.4	.6512	.6517	.6554
.5	.3135	.3128	.3085	-.5	.6865	.6872	.6915
.6	.2797	.2790	.2743	-.6	.7203	.7210	.7257
.7	.2478	.2469	.2420	-.7	.7522	.7531	.7580
.8	.2177	.2169	.2119	-.8	.7823	.7831	.7881
.9	.1898	.1889	.1841	-.9	.8102	.8111	.8159
1.0	.1640	.1632	.1587	-1.0	.8360	.8368	.8413
1.2	.1193	.1186	.1151	-1.2	.8807	.8814	.8849
1.4	.0834	.0830	.0808	-1.4	.9166	.9170	.9192
1.6	.0558	.0557	.0548	-1.6	.9442	.9443	.9452
1.8	.0356	.0356	.0359	-1.8	.9644	.9644	.9641
2.0	.0215	.0214	.0228	-2.0	.9785	.9786	.9772
2.2	.0121	.0124	.0139	-2.2	.9879	.9876	.9861
2.4	.0064	.0066	.0082	-2.4	.9936	.9934	.9918
2.6	.0030	.0033	.0047	-2.6	.9970	.9967	.9953
2.8	.0015	.0013	.0026	-2.8	.9985	.9987	.9974
3.0	.00046	.00060	.00135	-3.0	.99954	.99940	.99865
3.2	.00012	.00020	.00069	-3.2	.99988	.99980	.99931
3.4	.00000	.00004	.00034	-3.4	1.00000	.99996	.99966

$n = \infty$  is used to denote the Rayleigh solution.

This table suggests some interesting points. The curves for  $n=6$  and  $n=7$  are fairly close together, but differ sensibly from the Rayleigh solution, perhaps 4 or 5 per cent., where the density is at all material. For many practical purposes this might be close enough, and we see that for infiltration as distinct from dispersal curves, the Rayleigh solution—owing to integration over an area—gives fairly close results. The greatest percentage deviations from the Rayleigh solution are to be found in the tail. Now no individual can be found beyond the range  $nl$  from the boundary, and  $\sigma = \sqrt{\frac{1}{2}nl}$ ; thus the maximum range is  $\sqrt{2n}\sigma$ , or, for  $n=6$  and  $7$ , the maximum range is  $3.46\sigma$  and  $3.74\sigma$  respectively. The  $\omega$ -function expansion brings this out well. For  $n=6$  at  $3.4\sigma$  there is not one in 100,000 individuals, while the Rayleigh solution gives 34. For  $n=7$  there are still 4 in the 100,000, because we are a little distance still from the limit of



the range. The Rayleigh solution continues to give sensible densities beyond the range, although they may be sufficiently small to be neglected in practice.

For rough purposes a first approximation to the infiltration curves may be found from the Rayleigh solution, they will err on the side of safety if we are considering the effect of a clearance at a considerable distance from the boundary. But with the aid of the tables of the  $\omega$ -functions and the  $\nu$ -coefficients, it is not difficult to obtain the actual form of the infiltration curves as I have done in the present case. Diagram VII. compares the Rayleigh approximation and the infiltration curve for  $n=7$ .

It will be seen that an infiltration curve of the first order gives not only the density of the population after a first migration into cleared or unoccupied area across a straight boundary, but also the diminution of density on the populated side of the area, when we put  $c$  negative, *i.e.* it gives both the 'depopulation' and 'repopulation.' The reduced density at the boundary is  $\frac{1}{2}N$ , and if we take the point where the infiltration curve cuts the vertical through the boundary as origin, we see that it is centrally symmetrical; or the loss of population at a given distance from the boundary is exactly equal to the gain at the same distance on the opposite side of the boundary.

If we require an infiltration curve of the second order, we must now multiply the ordinates of the curve of the first order by (i) the average fertility of the species, say  $\mu$ , and (ii) the survival rate  $\Delta$ . If the environment be the same on either side of the boundary, and neither  $\mu$  nor  $\Delta$  affected by the density of the population, then  $\mu\Delta$  may be treated as a constant and the infiltration curves of higher orders can be found with moderate ease for simple cases. We thus have the distributions after two, three or more migrations accompanied by reproduction and death. On the other hand both  $\mu$  and  $\Delta$  may be functions of the density of the population, and in this case the ordinates of the infiltration curves of the second and higher orders can only be determined when the nature of  $\mu$  and  $\Delta$  is known. On the whole it is probable that the average fertility depending on the mating frequency will be highest where the density is greatest, as mating opportunities will then be most frequent, but in such cases the survival rate  $\Delta$  may be lower, as more enemies are likely to be present and the food supply is also likely to be less, where the population is densest. Thus  $\mu\Delta$  as a whole may not be very different on the depopulated and repopulated sides of the boundary. We shall only consider in this memoir cases in which this product is (i) supposed constant throughout, or (ii) constant for each migration season; but supposing uniform environment on both sides of the boundary, it is conceivable that  $\mu\Delta$  will be correlated with the population density and this will modify the basis of the distribution from which the second and later migrations start.



(9) Problem II. *To investigate the distribution after  $m$  migrations from uniformly densely occupied space across a straight boundary into unoccupied space.*

Let the axis of  $x$  be taken perpendicular to the boundary and the axis of  $y$  be the boundary. Let us consider the density at  $x=c$ , on the originally unoccupied side of the boundary. Then the density at a distance  $x$  from the boundary is given by (xlvi), or if we write the operator as  $Q_t$ , we have

$$F_n(x) = NQ_t \frac{1}{\sqrt{2\pi}} \int_{x/\sigma}^{\infty} e^{-\frac{1}{2}x^2} dx = u_1, \text{ say, } \dots\dots\dots(\text{li}).$$

Here  $Q_t$  involves only  $n$  and  $\sigma$  and not  $x$ .

Now the distance  $r$  from the point  $x, y$  to the point  $c, 0$  at which we want the density after the next migration is given by:

$$r^2 = y^2 + (x-c)^2,$$

and  $\mu\Delta$  being the fertility-survival factor, we have for the density at  $c$ ,

$$u_2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mu\Delta u_1 \phi_n(r^2) dx dy.$$

Now 
$$\phi_n(r^2) = Q_t \omega_0 = Q_t \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}r^2/\sigma^2}.$$

To mark that this  $Q_t$  operates only on this part of the expression, write it  $Q'_t$  and suppose it to operate on  $\sigma'$  written for  $\sigma$ . After the operations are complete we can put  $\sigma'$  again =  $\sigma$ . Let

$$v_1 = \frac{1}{\sqrt{2\pi}} \int_{x/\sigma}^{+\infty} e^{-\frac{1}{2}x^2} dx.$$

Then if  $\mu\Delta$  be constant (see p. 29):

$$u_2 = \frac{\mu\Delta N}{2\pi} Q_t Q'_t \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} v_1 \frac{1}{\sigma'^2} e^{-\frac{1}{2}\{(x-c)^2 + y^2\}/\sigma'^2} dx dy.$$

Completing the integration with regard to  $y$  we have:

$$u_2 = \frac{\mu\Delta N}{\sqrt{2\pi}} Q_t Q'_t \int_{-\infty}^{+\infty} v_1 \frac{1}{\sigma'} e^{-\frac{1}{2}(x-c)^2/\sigma'^2} dx.$$

Differentiate with regard to  $c$ :

$$\frac{du_2}{dc} = \frac{\mu\Delta N}{\sqrt{2\pi}} Q_t Q'_t \int_{-\infty}^{+\infty} v_1 \frac{1}{\sigma'} \frac{d}{dx} (-e^{-\frac{1}{2}(x-c)^2/\sigma'^2}) dx.$$

Integrate by parts, and notice that the part between limits vanishes at both of them and we have:

$$\frac{du_2}{dc} = \frac{\mu\Delta N}{\sqrt{2\pi}} Q_t Q'_t \int_{-\infty}^{+\infty} \frac{dv_1}{dx} \frac{1}{\sigma'} e^{-\frac{1}{2}(x-c)^2/\sigma'^2} dx.$$

But

$$\frac{dv_1}{dx} = -\frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2}(x/\sigma)^2};$$

hence:

$$\frac{du_2}{dc} = -\frac{\mu\Delta N}{2\pi} Q_t Q'_t \frac{1}{\sigma\sigma'} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma^2} + \frac{(x-c)^2}{\sigma'^2}\right)} dx.$$



This is integrable and gives:

$$\frac{du_2}{dc} = -\mu\Delta N Q_t Q_t' \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2 + \sigma'^2}} e^{-\frac{1}{2}c^2/(\sigma^2 + \sigma'^2)}.$$

Integrate with regard to  $c$ , and remember that  $u_2 = 0$  if  $c = \infty$ ; thus:

$$\begin{aligned} u_2 &= \mu\Delta N Q_t Q_t' \frac{1}{\sqrt{2\pi}} \int_c^\infty \frac{1}{\sqrt{\sigma^2 + \sigma'^2}} e^{-\frac{1}{2}x^2/(\sigma^2 + \sigma'^2)} dx \\ &= \mu\Delta N Q_t Q_t' \frac{1}{\sqrt{2\pi}} \int_{c/\sqrt{\sigma^2 + \sigma'^2}}^\infty e^{-\frac{1}{2}x'^2} dx' \dots\dots(\text{lii}). \end{aligned}$$

Comparing this with (li) we see that  $u_2$  differs from  $u_1$  by (a) the introduction of the factors  $Q_t'$  and  $\mu\Delta$  and (b) the replacement of  $\sigma$  in the lower limit by  $\sqrt{\sigma^2 + \sigma'^2}$ . The process can therefore be repeated as often as we please, and we have for  $u_m$  the value:

$$u_m = (\mu\Delta)^{m-1} N Q_t Q_t' Q_t'' \dots \text{ to } m \text{ terms } \frac{1}{\sqrt{2\pi}} \int_{c/\Sigma}^\infty e^{-\frac{1}{2}x'^2} dx',$$

where

$$\Sigma^2 = \sigma + \sigma'^2 + \sigma''^2 + \dots \text{ to } m \text{ terms.}$$

After the operations indicated by the  $Q$ 's are completed, we are to put

$$\sigma' = \sigma'' = \sigma''' = \dots = \sigma.$$

Now it is clear that a differentiation with regard to any  $\sigma^2$  is precisely the same as one with regard to  $\Sigma^2$ . We can therefore write for all the  $Q$ 's the simple expression

$$1 + \nu_4 (\sigma^2)^2 \frac{d^2}{d(\Sigma^2)^2} - \nu_6 (\sigma^2)^3 \frac{d^3}{d(\Sigma^2)^3} + \dots + (-1)^s (\sigma^2)^s \frac{d^s}{d(\Sigma^2)^s} + \dots$$

understanding that  $d/d(\Sigma^2)$  operates only on  $\Sigma$  and that after the operation is completed we can put  $\Sigma = \sqrt{m}\sigma$ . Thus the complete solution is:

$$\begin{aligned} u_m &= N (\mu\Delta)^{m-1} \left( 1 + \nu_4 (\sigma^2)^2 \frac{d^2}{d(\Sigma^2)^2} - \nu_6 (\sigma^2)^3 \frac{d^3}{d(\Sigma^2)^3} + \dots + (-1)^s \nu_{2s} \frac{d^s}{d(\Sigma^2)^s} + \dots \right)^m \\ &\quad \frac{1}{\sqrt{2\pi}} \int_{c/\Sigma}^\infty e^{-\frac{1}{2}x'^2} dx \dots\dots(\text{liii}). \end{aligned}$$

This is true for  $c$  positive or negative, *i.e.* whether the density be considered at a point on the originally occupied or originally unoccupied side of the boundary.

Up to terms of order  $1/n^3$  we have for the operator the value

$$\begin{aligned} &1 + m (\nu_4 q^2 - \nu_6 q^3 + \nu_8 q^4 - \nu_{10} q^5 + \nu_{12} q^6) \\ &+ \frac{m(m-1)}{1.2} (\nu_4^2 q^4 - 2\nu_4 \nu_6 q^5 + 2\nu_4 \nu_8 q^6) \\ &+ \frac{m(m-1)(m-2)}{1.2.3} \nu_4^3 q^6, \text{ where } q \text{ stands for } \sigma^2 d/d(\Sigma^2). \end{aligned}$$



Now exactly as on p. 26 we may show that:

$$\begin{aligned} Q^s \frac{1}{\sqrt{2\pi}} \int_{c/\Sigma}^{\infty} e^{-\frac{1}{2}x^2} dx &= \left(\frac{\sigma}{\Sigma}\right)^{2s} \frac{1}{\sqrt{2\pi}} \frac{c}{2\Sigma} \frac{(-1)^{s-1}}{2\Sigma} e^{-\frac{1}{2}c^2/\Sigma^2} \psi_{2(s-1)}(c/\Sigma) \\ &= \frac{1}{\sqrt{2\pi}} \frac{(-1)^{s-1}}{2m^s} \eta_m e^{-\frac{1}{2}\eta_m^2} \psi_{2(s-1)}(\eta_m), \end{aligned}$$

where:  $\psi_{2(s-1)}(\eta_m) = \chi_{2(s-1)}(\eta_m) + \frac{(s-1)}{1!2} \cdot \chi_{2(s-2)}(\eta_m) + \frac{(s-1)(s-2)}{2!2^2} 1 \cdot 3 \cdot \chi_{2(s-3)}(\eta_m)$   
 $+ \frac{(s-1)(s-2)(s-3)}{3!2^3} 1 \cdot 3 \cdot 5 \cdot \chi_{2(s-4)}(\eta_m) + \dots,$

$\eta_m = c/(\sqrt{m}\sigma)$ , and  $\chi_{2s}$  is defined on p. 10, Equation (xviii).

Thus

$$\begin{aligned} u_m = N(\mu\Delta)^{m-1} \left\{ \bar{\psi}(\eta_m) - \frac{1}{2} \frac{1}{\sqrt{2\pi}} \eta_m e^{-\frac{1}{2}\eta_m^2} \left( \frac{m}{m^2} \nu_4 \psi_2(\eta_m) + \frac{m}{m^3} \nu_6 \psi_4(\eta_m) \right. \right. \\ \left. \left. + \frac{m\nu_8 + \frac{1}{2}m(m-1)\nu_4^2}{m^4} \psi_6(\eta_m) + \frac{m\nu_{10} + m(m-1)\nu_4\nu_6}{m^5} \psi_8(\eta_m) \right. \right. \\ \left. \left. + \frac{m\nu_{12} + m(m-1)\nu_4\nu_8 + \frac{1}{6}m(m-1)(m-2)\nu_4^3}{m^6} \psi_{10}(\eta_m) \right) \right\} \dots\dots(\text{liv}). \end{aligned}$$

We see that this expression converges much more rapidly than that for  $\phi_n(r^2)$ , if  $m$  be at all large.

The result (liv) might have been reached in a different manner. We might have supposed the  $(\mu\Delta)^{m-1}Na$  individuals to have started from any element  $a$  on the populated side of the boundary and taken  $mn$  flights without multiplying to their final resting-place. The effect of this would be that  $\sigma^2 = \frac{1}{2}mnl^2$ , and that in the values of the  $\nu$ 's we must write  $mn$  for  $n$ . But doing this gives us precisely the coefficients of the  $\psi$ 's in (liv). Thus (liv) is deduced directly from (xlix). The proof becomes then much shorter, but it is more artificial; the fact that we may suppose all the unborn individuals to scatter from the original centre is not so easily realised, and further it does not in the process picture what takes place until the final arrangement after the  $m$ th breeding cycle is attained. In the method I have adopted we see the exact process of each breeding multiplication, its increase of the operating factor by an additional  $Q_t$ , and its increase of the square of the standard deviation by an additional  $\sigma^2$ . Lastly the final form (liv) enables us, without recalculating the  $\nu$ 's for each breeding cycle, to see very easily the effect in the case of any  $n$ -flight species, of taking any number of breeding cycles.

So long as we keep  $\mu\Delta$  constant of course our result for  $m$  breeding cycles with  $n$  flights will be the same as for a simple scattering for  $mn$  flights of a larger number of individuals. If  $\mu\Delta$  varies, however, we must adopt the method indicated in the above proof, and work out each migration successively. The same method must be adopted if a patch be rendered permanently sterile, because



TABLE VI. Distances of given Densities from the Boundary measured into the Immigration area in terms of Boundary Density.

Number of migrations	Percentage Density	Number of Flights				Percentage Density	Number of Flights				
		6	7	8	9		10	6	7	8	9
1	50	1.17	1.26	1.35	1.43	1.51	3.39	3.67	3.92	4.16	4.38
	30	1.99	2.15	2.30	2.44	2.57	4.46	4.82	5.15	5.46	5.76
	10	2.85	3.08	3.29	3.49	3.68	5.70	6.16	6.58	6.98	7.36
2	50	1.65	1.78	1.91	2.02	2.13	4.80	5.19	5.54	5.88	6.20
	30	2.82	3.04	3.25	3.45	3.64	6.31	6.82	7.29	7.73	8.15
	10	4.03	4.35	4.65	4.93	5.20	8.06	8.71	9.31	9.87	10.41
3	50	2.02	2.19	2.34	2.48	2.61	5.88	6.35	6.79	7.20	7.59
	30	3.45	3.73	3.98	4.23	4.46	7.73	8.35	8.92	9.46	9.98
	10	4.93	5.33	5.70	6.04	6.37	9.87	10.66	11.40	12.09	12.74
4	50	2.34	2.52	2.70	2.86	3.02	6.79	7.33	7.84	8.32	8.77
	30	3.98	4.30	4.60	4.88	5.14	8.92	9.64	10.30	10.93	11.52
	10	5.70	6.16	6.58	6.98	7.36	11.40	12.31	13.16	13.96	14.72
5	50	2.61	2.82	3.02	3.20	3.37	7.59	8.20	8.77	9.30	9.80
	30	4.46	4.81	5.14	5.46	5.75	9.98	10.78	11.52	12.22	12.88
	10	6.37	6.88	7.36	7.80	8.23	12.74	13.77	14.72	15.61	16.45
10	50	3.69	3.99	4.27	4.52	4.77	10.74	11.60	12.40	13.15	13.86
	30	6.30	6.81	7.28	7.72	8.13	14.11	15.24	16.29	17.28	18.21
	10	9.01	9.73	10.40	11.03	11.63	18.02	19.47	20.81	22.07	23.27
50	50	8.26	8.92	9.54	10.11	10.66	24.00	25.93	27.72	29.40	30.99
	30	14.09	15.22	16.27	17.26	18.19	31.55	34.08	36.43	38.64	40.73
	10	20.15	21.76	23.26	24.67	26.01	40.30	43.53	46.53	49.36	52.03
100	50	11.68	12.62	13.49	14.31	15.08	33.95	36.67	39.20	41.58	43.83
	30	19.92	21.52	23.01	24.40	25.72	44.62	48.19	51.52	54.64	57.60
	10	28.49	30.78	32.90	34.90	36.78	56.99	61.56	65.81	69.80	73.58



in such a case  $\mu$  is not constant for all parts of the integrated area, and we cannot suppose the whole final population to scatter from the original centres.

If we neglect the  $\psi_2, \psi_1 \dots$  terms in (liv) we have the value which would follow from the Rayleigh solution of the fundamental problem, and this can be very readily expressed in geometrical terms. For we mark at once that  $u_1$  and  $u_m$  are in type identical curves. Take  $u_1$  and stretch it vertically in the uniform ratio of  $\mu\Delta^{m-1}$  to 1, and horizontally in the ratio of  $\sqrt{m}$  to 1, and it becomes  $u_m$ . In other words the broken line on Diagram VII. represents the approximate solution in this case after  $m$  migrations provided we read  $N(\mu\Delta)^{m-1}$  for  $N$  on the vertical scale and  $\Sigma = \sqrt{m}\sigma$  for  $\sigma$  on the horizontal scale. The Table on p. 33 gives the chief results.

The unit of this table is the length  $l$  of "flight." It will be desirable to illustrate its application. Any such application can be of course only a suggestion, and on this account the above Table has been calculated to only a few places of decimals. But such suggestions may not be without value. They will become more than suggestions when our knowledge is greater of the migratory habits of different species. At present only rough approximations can be made as to the values of  $n$  and  $l$ , and these admittedly are of small weight.

*Illustration I.* In captivity I have noted that *H. aspersa* will live for five years. For two years it does not usually lay eggs, and then it will generally, but not invariably, reproduce twice in the year. This is of course subject to claustral conditions, and while these seem in some cases unfavourable, in others they may be advantageous both in matter of longevity and—owing to the constant food supply—in number of broods. This snail, as far as my observation goes, appears to return to the same shelter after seeking its food. Leaving such "flitters" on one side, I think we might look upon thirty to forty yards as a maximum "flight" for such a snail and regard seven or eight such flights between its egg layings as on the average an exaggeration.

We might therefore take  $l = 40$  yards,  $n = 8$ , and an average during life of one brood a year as being quite possible approximations in the case of some snails.

This indicates that the progress across a boundary into unoccupied country would be such that 1 per cent. of the density at the boundary and, therefore, possibly  $\frac{1}{2}$  per cent. of the density in the fully-occupied country, would only be reached at 2061 yards from the boundary after 100 migrations. In other words, such a species would only progress a mile or two at most in a century. Such progress would hardly be noted in any studies hitherto made of distribution; the limits of a species a hundred years ago were certainly not closely defined to a mile or two, even if they have been recently. Of course there are many other ways in which a slow moving species can be transported than by its own "flights," and further no special stress is laid on the above case, but a study of Table VI. shows



that the advance of a slow scattering species\* may be comparatively small. The inference can accordingly be made that the existing boundaries of the geographical distribution of certain forms of animal and plant life which are not marked by natural barriers, and which do not correspond to obviously changing environmental conditions, need not after all be associated with subtle physical differences which have escaped the observation of the naturalist. The species may be progressing into an unoccupied area, but at a rate hardly observable in the time during which accurate distribution observations are available. If this view be correct we should expect such boundaries with no apparent environmental change in the case of species for which we might reasonably predict a small  $n$  and  $l$ .

*Illustration II.* I have endeavoured to apply the above theory to the immigration of mosquitoes into a cleared area. We will suppose in the present treatment that the area bounded by a straight line (some attempt to allow for curvature of the boundary will be considered later) has been cleared but is not kept sterile to the species. I shall speak of a district as rendered sterile to a species when it is made impossible for it to breed there, and kept sterile when the breeding possibilities are persistently destroyed. The distinction is an important one, especially in the mosquito case. For in the latter case all mosquitoes are immigrants, and in the former case we have not only immigrants, but their produce.

Major Ronald Ross, who has most kindly provided me with information as to mosquito habits, makes the following remarks:

(a) That the number of mosquitoes produced varies roughly (*ceteris paribus*) as the extent of surface breeding area.

(b) That the breeding area can be taken as consisting of numerous isolated small pools or vessels of water scattered fairly uniformly over the country.

(c) That the feeding places (houses, stables, birds, etc.) may be taken as scattered pretty uniformly between the breeding pools.

(d) That abundance or scarcity of food can scarcely influence the question much. A single man or bird will yield enough food for many mosquitoes, and if they starve it is not because the food is not there, but because they cannot reach it. They are therefore not likely to be drawn in general by special abundance of food in any special direction. Wind tends to make mosquitoes "sit tight," rather than allow themselves to be scattered.

It would thus appear that on the average an "equi-swampous" condition of the environment and random "flights" of the mosquito will not be very wide of the truth. The difficulty is to form some estimate of  $n$  and  $l$ . On these points again Major Ross came to my help, but naturally the statements he made were with great reservation.

\* Of course any more quickly moving species that depends on this for food would have the same boundary, but in its case the boundary would be environmentally defined.



( $\alpha$ ) From egg to egg (*i.e.* from laying of eggs, hatching, larval and pupal stages, to laying of eggs again) takes roughly about a fortnight in hot countries with most mosquitoes. In England, gnats may have only one generation or two in a summer, but in the tropics they may go on breeding throughout the year. In cool countries the egg to egg cycle may be prolonged to a month or two. In certain very hot and dry countries, breeding may be checked entirely except during the rainy season. I have accordingly taken 20, 10 and 5 breedings to the year to represent roughly these conditions.

( $\beta$ ) Major Ross distinguishes between "minor vicissitudes," which an insect makes when it hovers round its victim or mate, and "major vicissitudes" which it makes when it passes from feeding place to pool for egg laying. These correspond to my "flitters" and "flights." He considers that they go back to water every four or five days, so that a "major vicissitude" occurs every two days or so. We might therefore take, excluding flitters, the average number of flights to be six or seven. Of course this is the roughest approximation, but still not an unreasonable estimate of what probably takes place in the mosquito's life.

( $\gamma$ ) As to the magnitude of  $l$  we have less definite data. Mosquitoes of a rare kind have been said to have been found two or three miles from their breeding place. Major Ross thinks that *Anopheles* will exceptionally, when no houses are near, probably travel  $\frac{1}{2}$  mile for their food, or perhaps further, but he supposes the average distance scarcely to exceed  $\frac{1}{4}$  mile, and it may, as houses and suitable pools often abound not more than 50 yards apart, be not greater, perhaps, than 100 yards.

I have accordingly taken 100 yards and 500 yards as likely values for  $l$ , and considering 1 per cent. of the boundary value of the mosquitoes' density as a limit to their existence and 5 per cent. as objectionable, we have the following table:

TABLE VII. *Distances from the Boundary of a cleared but not sterile area at which 1 per cent. and 5 per cent. of the boundary density of Mosquitoes will be found in the course of a Year.*

Number of Flights.....		Supposed number of Breeding Cycles in Year					
		5		10		20	
		6	7	6	7	6	7
Density 1 per cent.	$l = 100$	998	1078	1411	1524	1995	2155
" " "	$= 500$	4990	5390	7055	7620	9975	10775
Density 5 per cent.	$= 100$	759	820	1074	1160	1518	1640
" " "	$= 500$	3745	4100	5370	5800	7590	8200



The distances are all given in yards.

Thus we see that the least of these distances for 1 per cent. is greater than half a mile, or, if an area be cleared but not rendered sterile, we might expect within a year the mosquitoes to reappear within half a mile of the boundary, and to reach an objectionable frequency even at this distance for most of the cases considered.

As far then as these rough numbers can be taken to indicate the state of affairs, it is needful not only to clear an area but to maintain it sterile. The clearance radius may be only  $\frac{1}{2}$  mile and is hardly likely to exceed a mile, and the above results only mark the progress of immigration in the course of one year after the clearance. Further the results would be accentuated if the boundary were curved or an approximately circular clearance made.

It does not appear to me that any substantial difference would be made in the main result by reducing  $n$  to 3 or 4, although some difference would occur if  $l$  were reduced to 20 or 30 yards.

(10) Problem III. *To determine the distribution after  $m$   $n$ -flight migrations starting with a centre of population  $N\alpha$ .*

The previous two problems indicate the nature of the general solutions to which I now proceed. I shall adopt the longer process of proof in this first case as being the more suggestive.

By (xii) and (xlvi), calling the operator as before  $Q_t$ , we have for the distribution at  $X, Y$  due to a centre at the origin :

$${}_1\phi_n(X, Y) = \frac{N\alpha}{2\pi} Q_t \left( \frac{1}{\sigma^2} e^{-\frac{1}{2}(X^2+Y^2)/\sigma^2} \right) \dots\dots\dots(lv).$$

Hence the distribution at  $(h, k)$  after a second migration of  $n$  flights is

$${}_2\phi_n(h, k) = \mu\Delta \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} {}_1\phi_n(X, Y) \frac{Q_t e^{-\frac{1}{2}\{(X-h)^2+(Y-k)^2\}/\sigma^2}}{\sigma^2} dXdY.$$

Call the  $Q_t$  in this  $Q_{t_2}$  and write the  $\sigma^2$  on which it operates  $\sigma_2^2$ ; call the  $Q_t$  in  ${}_1\phi_n(X, Y)$ ,  $Q_{t_1}$  and the  $\sigma^2$  on which it operates  $\sigma_1^2$ , we have :

$${}_2\phi_n(h, k) = \frac{\mu\Delta N\alpha}{(2\pi)^2} Q_{t_1} Q_{t_2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{\sigma_1^2 \sigma_2^2} e^{-\frac{1}{2}\left\{\frac{(X-h)^2+(Y-k)^2}{\sigma_2^2} + \frac{X^2+Y^2}{\sigma_1^2}\right\}} dXdY.$$

The integrations can be performed and give us

$${}_2\phi_n(h, k) = \frac{\mu\Delta N\alpha}{2\pi} Q_{t_1} Q_{t_2} \frac{1}{\sigma_1^2 + \sigma_2^2} e^{-\frac{1}{2}\{(h^2+k^2)/(\sigma_1^2+\sigma_2^2)\}} \dots\dots\dots(lvi).$$

This only differs from  ${}_1\phi_n(X, Y)$  by the introduction of  $\sigma_1^2 + \sigma_2^2$  for  $\sigma^2$  and of the factor  $\mu\Delta Q_{t_2}$ .



We can accordingly repeat the process as often as we like and we have :

$${}_m\phi_n(h, k) = \frac{(\mu\Delta)^{m-1}N\alpha}{2\pi} Q_{t_1} Q_{t_2} \dots Q_{t_m} \frac{1}{\Sigma^2} e^{-\frac{1}{2}(h^2+k^2)/\Sigma^2} \dots\dots\dots(\text{lvii}),$$

where  $\Sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \dots + \sigma_m^2,$

and after the operations have been performed we are to put all the  $\sigma$ 's equal to  $\sigma$  or  $\Sigma^2 = m\sigma^2$ . But no operator  $Q_t$  affects any  $\sigma^2$  in any other operator, and  $\frac{d}{d\sigma_i^2} = \frac{d}{d\Sigma^2}$ . Thus  $(\sigma_i^2)^s \frac{d^s}{d(\sigma_i^2)^s}$  may be put  $= (\sigma^2)^s \frac{d^s}{d(\Sigma^2)^s}$ , and this makes all the operators identical in form and we may write

$$\begin{aligned} Q_{t_1} Q_{t_2} \dots Q_{t_m} &= \left\{ 1 + \nu_4 (\sigma^2)^2 \frac{d^2}{d(\Sigma^2)^2} - \nu_6 (\sigma^2)^3 \frac{d^3}{d(\Sigma^2)^3} + \nu_8 (\sigma^2)^4 \frac{d^4}{d(\Sigma^2)^4} \right. \\ &\quad \left. + \dots + (-1)^s \nu_{2s} (\sigma^2)^s \frac{d^s}{d(\Sigma^2)^s} + \dots \right\}^m \\ &= 1 + N_4 (\sigma^2)^2 \frac{d^2}{d(\Sigma^2)^2} - N_6 (\sigma^2)^3 \frac{d^3}{d(\Sigma^2)^3} + N_8 (\sigma^2)^4 \frac{d^4}{d(\Sigma^2)^4} \\ &\quad + \dots + (-1)^s N_{2s} (\sigma^2)^s \frac{d^s}{d(\Sigma^2)^s} + \dots \end{aligned}$$

In this form of the operator we can now write at once  $\sigma^2 = \frac{1}{m}\Sigma^2$  and call the expression  $Q_t^m$ .

$$\begin{aligned} \text{Thus } Q_t^m &= 1 + N_4 (\Sigma^2)^2 \frac{d^2}{d(\Sigma^2)^2} - N_6 (\Sigma^2)^3 \frac{d^3}{d(\Sigma^2)^3} + N_8 (\Sigma^2)^4 \frac{d^4}{d(\Sigma^2)^4} \\ &\quad + \dots + (-1)^s N_{2s} (\Sigma^2)^s \frac{d^s}{d(\Sigma^2)^s} + \dots \dots\dots(\text{lviii}), \end{aligned}$$

where  $N_4 = \frac{m}{m^2} \nu_4, \quad N_6 = \frac{m}{m^3} \nu_6,$

$$N_8 = \frac{m\nu_8 + \frac{1}{2}m(m-1)\nu_4^2}{m^4}, \quad N_{10} = \frac{m\nu_{10} + m(m-1)\nu_4\nu_6}{m^5},$$

$$N_{12} = \frac{m\nu_{12} + m(m-1)\nu_4\nu_8 + \frac{1}{6}m(m-1)(m-2)\nu_4^3}{m^6}, \text{ etc. } \dots\dots\dots(\text{lix}).$$

These values of the  $N$ 's rapidly converge and their values are given in Table III. on p. 21 of this paper with those of the  $\nu$ 's for a few values of  $n$  and  $m$ . As we have seen on p. 32, they are the  $\nu$ 's obtained by using values of  $nm$  for  $n$ .

We now have the general solution of distribution from a centre :

$$\begin{aligned} {}_m\phi_n(h, k) &= (\mu\Delta)^{m-1} \frac{N\alpha}{2\pi} Q_t^m \frac{1}{\Sigma^2} e^{-\frac{1}{2}(h^2+k^2)/\Sigma^2} \dots \\ &= (\mu\Delta)^{m-1} N\alpha (\Omega_0 + N_4\Omega_4 + N_6\Omega_6 + \dots + N_{2s}\Omega_{2s} + \dots) \dots(\text{lx}). \end{aligned}$$

This is absolutely identical with (xii), except that the constants  $\nu$  are replaced



by other constants  $N$  of known value, and in every  $\omega$ -function we are to replace  $\sigma^2$  by  $m\sigma^2$  or  $\Sigma^2$ , that is to say a uniform stretch in the ratio of  $\sqrt{m}$  to  $l$  is given to any surface  $z = \omega_{zs}$  parallel to the axes of  $x$  and  $y$ . This is denoted by writing  $\Omega_{zs}$  for  $\omega_{zs}$ .

If we confine our attention to the Rayleigh part of the solution—which will be more and more nearly exact as  $m$  increases, for the  $N$ 's rapidly decrease in value—then we have

$${}_m\bar{\phi}_n(h, k) = (\mu\Delta)^{m-1} N \alpha \Omega_0 \dots \dots \dots (lxi),$$

and we see that every density gradient curve for the successive migrations is to be obtained by a stretch from the first migration density curve.

In general, however, this result is not absolutely true because the different components of the true solution are mixed in different proportions, the  $N$ 's being functions of  $m$ . We see, however, that the stretching rule becomes more and more accurate, as we increase either the number of flights or the number of migrations.

(11) Problem IV. *To find the form of the general solution for the distribution into surrounding space after  $m$  migrations of any population initially spread uniformly over any given patch with density  $N$ .*

The density at  $h, k$ , after  $m$  migrations due to a centre  $N dx dy$ , is by (lvii) above

$$= (\mu\Delta)^{m-1} \frac{N dx dy}{2\pi} Q_t^m \frac{1}{\Sigma^2} e^{-\frac{1}{2}\{(x-h)^2 + (y-k)^2\}/\Sigma^2}.$$

To give the patch let  $x$  be integrated from  $v_1$  to  $v_2$ , where  $v_1$  and  $v_2$  will usually be functions of  $y$ , and then let  $y$  be integrated from  $u_1$  to  $u_2$ . We find:

$${}_mF_n(h, k) = (\mu\Delta)^{m-1} \frac{N Q_t^m}{2\pi} \int_{u_1}^{u_2} \int_{v_1}^{v_2} \frac{1}{\Sigma^2} e^{-\frac{1}{2}\{(x-h)^2 + (y-k)^2\}/\Sigma^2} dx dy \dots \dots (lxii).$$

This is the general form of the solution when the population spreads from a uniform patch into non-sterile surrounding country.

If on the other hand we want the distribution after  $m$  migrations starting with a cleared patch, which is not kept sterile, we have

$${}_mF_n(h, k) = (\mu\Delta)^{m-1} N - {}_mF_n(h, k) \dots \dots \dots (lxiii),$$

for the whole district would have had a uniform density of  $(\mu\Delta)^{m-1} N$  had there been no clearance. Hence

$${}_mF_n(h, k) = (\mu\Delta)^{m-1} N Q_t^m \left( 1 - \frac{1}{2\pi} \int_{u_1}^{u_2} \int_{v_1}^{v_2} \frac{1}{\Sigma^2} e^{-\frac{1}{2}\{(x-h)^2 + (y-k)^2\}/\Sigma^2} dx dy \right) \dots (lxiv).$$

Now  ${}_1F_n(h, k) = N Q_t' \left( 1 - \frac{1}{2\pi} \int_{u_1}^{u_2} \int_{v_1}^{v_2} \frac{1}{\sigma^2} e^{-\frac{1}{2}\{(x-h)^2 + (y-k)^2\}/\sigma^2} dx dy \right).$



Hence the rule: If the solution can be found for a single migration, replace  $\sigma^2$  by  $m\sigma^2$ , and each  $\nu$  by the proper  $N$ , multiply by the factor  $(\mu\Delta)^{m-1}$ , and the solution for  $m$  migrations is deduced.

It will thus be clear that, if the solution can in any case be found for one migration fully, we can at once extend it to the case of any number of migrations, with constant fertility-survival factor.

(12) Problem V. *To determine the distribution after a first migration into a cleared rectangular area.*

Let the area be the rectangle  $2a \times 2b$ , and the origin be taken at its centre and axes of  $x$  and  $y$  parallel respectively to the sides  $2a$  and  $2b$ . Then the density at any point  $h, k$ , after a single migration  $F_n(h, k)$  is given by the principle of the last problem by

$$F_n(h, k) = N - F_n(h, k) \dots\dots\dots(\text{lxv}),$$

where  $F_n(h, k)$  is the distribution from a uniformly occupied rectangular area into surrounding unoccupied space.

$$\begin{aligned} \text{But} \quad F_n(h, k) &= N \int_{-a}^{+a} \int_{-b}^{+b} \phi_n \{(x-h)^2 + (y-k)^2\} dx dy \\ &= \frac{1}{2\pi} N Q_t \frac{1}{\sigma^2} \int_{-a}^{+a} \int_{-b}^{+b} e^{-\frac{1}{2} \left\{ \frac{(x-h)^2}{\sigma^2} + \frac{(y-k)^2}{\sigma^2} \right\}} dx dy \\ &= \frac{1}{2\pi} N Q_t \frac{1}{\sigma^2} \int_{-a}^{+a} e^{-\frac{1}{2}(x-h)^2/\sigma^2} dx \times \int_{-b}^{+b} e^{-\frac{1}{2}(y-k)^2/\sigma^2} dy. \end{aligned}$$

Let  $P_0(\epsilon)$  stand for the probability integral

$$\frac{1}{\sqrt{2\pi}} \int_0^\epsilon e^{-\frac{1}{2}x^2} dx.$$

$$\begin{aligned} \text{Then:} \quad \frac{1}{\sqrt{2\pi}\sigma} \int_{-a}^{+a} e^{-\frac{1}{2}(x-h)^2/\sigma^2} dx &= \frac{1}{\sqrt{2\pi}} \int_{-(a+h)/\sigma}^{(a-h)/\sigma} e^{-\frac{1}{2}x^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \left( \int_0^{(a-h)/\sigma} + \int_0^{(a+h)/\sigma} \right) e^{-\frac{1}{2}x^2} dx \\ &= P_0\left(\frac{a-h}{\sigma}\right) + P_0\left(\frac{a+h}{\sigma}\right). \end{aligned}$$

Thus:

$$F_n(h, k) = N Q_t P_0 \left\{ \left( \frac{a-h}{\sigma} \right) + P_0 \left( \frac{a+h}{\sigma} \right) \right\} \left\{ P_0 \left( \frac{b-k}{\sigma} \right) + P_0 \left( \frac{b+k}{\sigma} \right) \right\} \dots(\text{lxvi}).$$

Now consider the differentiation of  $P_0\left(\frac{u}{\sigma}\right)$  with regard to  $\sigma^2$ .

$$\frac{d}{d\sigma^2} \left\{ P_0 \left( \frac{u}{\sigma} \right) \right\} = \frac{d}{2\sigma d\sigma} \frac{1}{\sqrt{2\pi}} \int_0^{u/\sigma} e^{-\frac{1}{2}x^2} dx = -\frac{1}{2} \frac{u}{\sigma} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma^2} e^{-\frac{1}{2}(u^2/\sigma^2)}.$$



Writing  $\sigma^2 = t$  as before, we find

$$\frac{d}{dt} \left\{ P_0 \left( \frac{u}{\sigma} \right) \right\} = -\frac{1}{2} \sqrt{2\pi u} \frac{\omega_0}{\sqrt{t}} \dots\dots\dots(\text{lxvii}).$$

Hence

$$\begin{aligned} \frac{d^s}{dt^s} \left\{ P_0 \left( \frac{u}{\sigma} \right) \right\} &= -\frac{1}{2} \sqrt{2\pi u} \frac{d^{s-1}}{dt^{s-1}} \left( \frac{\omega_0}{\sqrt{t}} \right) \\ &= \frac{1}{2} \frac{\sqrt{2\pi} (-1)^s u}{t^{s-\frac{1}{2}}} \left( \omega_{2(s-1)} + (s-1) \omega_{2(s-2)} \frac{1}{2} + \frac{(s-1)(s-2)}{1 \cdot 2} \omega_{2(s-3)} \frac{1 \cdot 3}{2 \cdot 2} \right. \\ &\quad \left. + \frac{(s-1)(s-2)(s-3)}{1 \cdot 2 \cdot 3} \omega_{2(s-4)} \frac{1 \cdot 3 \cdot 5}{2 \cdot 2 \cdot 2} + \text{etc.} \right), \end{aligned}$$

$$t^s \frac{d^s}{dt^s} \left\{ P_0 \left( \frac{u}{\sigma} \right) \right\} = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{u}{\sigma} (-1)^s e^{-\frac{1}{2} u^2 / \sigma^2} \psi_{2(s-1)}(u/\sigma) \dots\dots\dots(\text{lxviii}),$$

the expression being the same as that on p. 32.

Now let us write the following for brevity where  $\eta = u/\sigma$ :

$$L_1(\eta) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \eta e^{-\frac{1}{2}\eta^2} (\nu_4 \psi_2(\eta) + \nu_6 \psi_4(\eta) + \dots + \nu_{2s} \psi_{2(s-1)}(\eta) + \dots),$$

$$L_2(\eta) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \eta e^{-\frac{1}{2}\eta^2} (2\nu_4 + 3\nu_6 \psi_2(\eta) + \dots + s\nu_{2s} \psi_{2(s-2)}(\eta) + \dots),$$

$$L_3(\eta) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \eta e^{-\frac{1}{2}\eta^2} \left( 3\nu_6 + 6\nu_8 \psi_2(\eta) + \dots + \frac{s(s-1)}{1 \cdot 2} \nu_{2s} \psi_{2(s-3)}(\eta) + \dots \right),$$

$$L_4(\eta) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \eta e^{-\frac{1}{2}\eta^2} \left( 4\nu_8 + 10\nu_{10} \psi_2(\eta) + \dots + \frac{s(s-1)(s-2)}{1 \cdot 2 \cdot 3} \nu_{2s} \psi_{2(s-4)}(\eta) + \dots \right) \dots(\text{lxix}),$$

and so on. All these functions are directly expressible in  $\omega$ -functions as on p. 27.

Further let  $P_s(\eta) = (-1)^s t^s \frac{d^s}{dt^s} P_0(\eta) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \eta e^{-\frac{1}{2}\eta^2} \psi_{2(s-1)}(\eta) \dots\dots\dots(\text{lxix}).$

Then we have, if

$$\eta_1 = (a-h)/\sigma, \quad \eta_2 = (a+h)/\sigma, \quad \epsilon_1 = (a-h)/\sigma, \quad \epsilon_2 = (a+h)/\sigma,$$

$$\begin{aligned} F_n(h, k) &= N \left[ \{P_0(\eta_1) + P_0(\eta_2)\} \{P_0(\epsilon_1) + P_0(\epsilon_2)\} + \{L_1(\eta_1) + L_1(\eta_2)\} \{P_0(\epsilon_1) + P_0(\epsilon_2)\} \right. \\ &\quad + \{P_0(\eta_1) + P_0(\eta_2)\} \{L_1(\epsilon_1) + L_1(\epsilon_2)\} + \{P_1(\epsilon_1) + P_1(\epsilon_2)\} \{L_2(\eta_1) + L_2(\eta_2)\} \\ &\quad \left. + \dots + \{P_s(\epsilon_1) + P_s(\epsilon_2)\} \{L_{s+1}(\eta_1) + L_{s+1}(\eta_2)\} + \dots \right] \dots\dots\dots(\text{lxix}). \end{aligned}$$

The  $L$ -functions involve the rapidly converging  $\nu$ -coefficients, and the first few terms will suffice to get an idea of the distribution. If we retain only the Rayleigh terms we find:

$$F_n(h, k) = N \left[ 1 - \{P_0(\eta_1) + P_0(\eta_2)\} \{P_0(\epsilon_1) + P_0(\epsilon_2)\} \right] \dots\dots\dots(\text{lxix}),$$

which can be ascertained for given values of  $a, b, h, k$  and  $\sigma$  from the ordinary tables of the probability integral.



If we make  $b$  infinite, then  $P_s(\epsilon_1)$  and  $P_s(\epsilon_2) = 0$  for  $s > 0$ , and  $L_1(\epsilon_1)$  and  $L_1(\epsilon_2) = 0$ ,  $P_0(\epsilon_1) = P_0(\epsilon_2) = \frac{1}{2}$ , and

$$F_n(h, k) = N \{1 - P_0(\eta_1) - P_0(\eta_2) - L_1(\eta_1) - L_2(\eta_2)\} \dots\dots\dots(\text{lxxiii}).$$

This could be deduced directly from (xlix) and it represents the first migration distribution into an indefinitely long cleared strip or belt. This is a result of some interest as it might approximately apply to the migration into a zone cleared by a flood or a fire of certain types of animal or vegetable life.

(13) Problem VI. *To determine the distribution after  $m$  migrations into a cleared but not sterile rectangular area.*

By the general proposition on p. 39 we have only to write  $\Sigma = m\sigma$  for  $\sigma$ , and the  $N$ 's for the  $\nu$ 's in the  $L$ 's. Let us put

$$\eta_1' = (a - h)/\Sigma = \eta_1/\sqrt{m}, \quad \eta_2' = \eta_2/\sqrt{m}, \quad \epsilon_1' = \epsilon_1/\sqrt{m}, \quad \epsilon_2' = \epsilon_2/\sqrt{m}.$$

Let

$$L_1'(\eta_1') = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \eta_1' e^{-\frac{1}{2}\eta_1'^2} \{N_4 \psi_2(\eta_1') + N_6 \psi_4(\eta_1') + \dots + N_{2s} \psi_{2(s-1)}(\eta_1') + \dots\},$$

and so forth, then we have for the full solution:

$$\begin{aligned} {}_m F_n(h, k) = (\mu\Delta)^{m-1} N \{ & \{P_0(\eta_1') + P_0(\eta_2')\} \{P_0(\epsilon_1') + P_0(\epsilon_2')\} \\ & + \{L_1'(\eta_1') + L_1'(\eta_2')\} \{P_0(\epsilon_1') + P_0(\epsilon_2')\} + \{P_0(\eta_1') + P_0(\eta_2')\} \{L_1'(\epsilon_1') + L_1'(\epsilon_2')\} \\ & + \{P_1(\epsilon_1') + P_1(\epsilon_2')\} \{L_2'(\eta_1') + L_2'(\eta_2')\} \\ & + \dots + \{P_s(\epsilon_1') + P_s(\epsilon_2')\} \{L_{s+1}'(\eta_1') + L_{s+1}'(\eta_2')\} + \dots \} \dots\dots\dots(\text{lxxiv}). \end{aligned}$$

The terms here will very rapidly converge for any fairly large value of  $m$ , so that for many purposes we may write the solution:

$${}_m F_n(h, k) = (\mu\Delta)^{m-1} N \{P_0(\eta_1') + P_0(\eta_2')\} \{P_0(\epsilon_1') + P_0(\epsilon_2')\} \dots\dots(\text{lxxv}),$$

which can be found at once from the usual tables of the probability integral.

*Illustration I.* A rectangular patch 2 miles long and 1 mile broad is cleared of mosquitoes, but not retained sterile. What would be the central density at the end of the year? Suppose 10 breeding cycles with their scatter migrations, each of 6 flights, to take place in the year. Then if we take 200 yards as a possible round value for the flight we have:

$$\begin{aligned} m = 10, \quad n = 6, \quad l = 200 \text{ yds.}, \quad \sigma^2 = \frac{1}{2}nl^2 = 120,000 \quad \text{or} \quad \sigma = 346.41 \text{ yds.} \\ a = 880 \text{ yds.}, \quad b = 1760 \text{ yds.}, \quad \eta_1 = \eta_2 = 2.540, \quad \epsilon_1 = \epsilon_2 = 5.081, \\ \Sigma = \sqrt{10}\sigma = 1095.44 \text{ yds.}, \quad \eta_1' = \eta_2' = .803, \quad \epsilon_1' = \epsilon_2' = 1.607. \end{aligned}$$

Hence  ${}_{10}F_6(0, 0) = (\mu\Delta)^9 4P_0(.803)P_0(1.607)N,$

or, using Sheppard's Tables:

$$\begin{aligned} {}_{10}F_6(0, 0) &= (\mu\Delta)^9 4 \times .2890 \times .4460N, \\ &= (\mu\Delta)^9 \times .5156N. \end{aligned}$$



$$\begin{aligned}\text{Thus } {}_{10}\mathcal{F}_c(0, 0) &= (\mu\Delta)^9 (1 - \cdot 5156) N \\ &= (\mu\Delta)^9 \cdot 48 N.\end{aligned}$$

We see accordingly that if the fertility and the death-rate were the same in the clearance and in the populated district outside, the density at the centre of the cleared patch would at the end of the year be almost 50 per cent. of that in uncleared country. It is thus obvious that clearance can be of small use, unless it is followed by permanent preservation of sterility. Even if one *annual* clearance were made it is very unlikely—if the actual values of the constants are at all near those assumed—that the mosquitoes would not by the 9th or 10th breeding cycle within the year before the annual clearance was repeated have reached a very substantial density even at the centre of the patch. We have thus an additional argument in favour of rendering a district not only sterile, but keeping it so. In such a case since  $\nu_1$  and  $\nu_2$ ,  $\psi_1$ ,  $\psi_2$  are negative we shall have a density somewhat less than :

$${}_1\mathcal{F}_c(0, 0) = N \{1 - 4P_0(2\cdot 540) P_0(5\cdot 081)\} = N(1 - \cdot 9889) \text{ about.}$$

$$\text{Thus : } {}_1\mathcal{F}_c(0, 0) = \cdot 01N \text{ approximately.}$$

It follows that in the centre of such a rectangular patch, there would roughly be only about 1 mosquito for every 100 in uncleared country.

But while this shows that such a sterile patch would be a great improvement for a denizen at the centre it is well to enquire what happens in such patches some way from the centre. I accordingly add the following illustration.

*Illustration II.* A square area of one mile side is cleared and kept permanently sterile. What will be the density at the centre and a quarter of a mile from the centre on the same assumption as before ?

$$\text{Here } a = b = 880 \text{ yds.}$$

At the centre  $\eta_1 = \eta_2 = \epsilon_1 = \epsilon_2 = 2\cdot 54$  and :

$${}_1\mathcal{F}_c(0, 0) = N[1 - 4\{P_0(2\cdot 54)\}^2] = N\{1 - (\cdot 9889)^2\} = \cdot 022N;$$

or, we find one mosquito for every fifty in uncleared country. Taking our quarter of a mile directly towards one of the boundaries, we have  $h = 440$ ,  $k = 0$ , and :

$$\eta_1 = 1\cdot 27, \quad \eta_2 = 3\cdot 81, \quad \epsilon_1 = \epsilon_2 = 2\cdot 54.$$

$$\begin{aligned}\text{Thus : } {}_1\mathcal{F}_c(440, 0) &= N[1 - \{P_0(1\cdot 27) + P_0(3\cdot 81)\} \{2P_0(2\cdot 54)\}] \\ &= N\{1 - (\cdot 3980 + \cdot 4999)(\cdot 9889)\} = \cdot 112N.\end{aligned}$$

Thus at  $\frac{1}{4}$  mile from the centre (or from the edge) of the clearance, the density is 11 per cent. of that in uncleared country. It may be doubted whether this is a sufficient reduction, and, supposing the above assumptions to be anything like roughly correct, it may be needful to render more than a square mile permanently sterile to protect a patch of one square half-mile.



On the other hand a cleared but not permanently sterile square mile would after a year have a density at the same point— $\frac{1}{4}$  mile from the centre—of:

$${}_{\infty}F_s(440, 0) = (\mu\Delta)^9 N [1 - \{P_0(.402) + P_0(1.205)\} \{2P_0(.803)\}] = (\mu\Delta)^9 \cdot 69N,$$

or of 69 per cent. of that in uncleared country.

Another point seems of some interest. What is the density at the boundary after the first migration?

At the middle point of the edge it is

$$\begin{aligned} {}_1F_s(880, 0) &= N[1 - \{P_0(0) + P_0(5.08)\} \{2P_0(2.54)\}] \\ &= N(1 - .5000 \times .9889) \\ &= .506N. \end{aligned}$$

This is almost the  $\frac{1}{2}N$  of an indefinitely long straight boundary.

At the corner it is

$${}_1F_s(880, 880) = N[1 - \{P_0(0) + P_0(5.08)\}^2] = .75N \text{ nearly,}$$

or, as we should expect, has risen much beyond the  $\frac{1}{2}N$  value.

There is no difficulty in tracing the contour lines of the population density in this case.

If we consider a cycle of 10 breedings in a non-sterile patch we have:

$$\begin{aligned} {}_{\infty}F_s(880, 0) &= (\mu\Delta)^9 N [1 - \{P_0(0) + P_0(1.607)\} \{2P_0(.808)\}] \\ &= .742N (\mu\Delta)^9, \end{aligned}$$

$$\begin{aligned} \text{and } {}_{\infty}F_s(880, 880) &= (\mu\Delta)^9 N [1 - \{P_0(0) + P_0(1.607)\}^2] \\ &= .801N (\mu\Delta)^9. \end{aligned}$$

Thus if the patch were not sterile, the effect of the clearance would at the boundary after the lapse of a year be marked by a 20 to 25 per cent. reduction. The illustrations I have given are of course dependent on the values of the constants selected. Such constants have at present been little studied, and accordingly small weight can be laid on the actual numerical results. But the theory appears to indicate useful lines of inquiry, even if its results will of course need to be controlled everywhere by local facts. In a general way there can be little doubt that a theory like the present will not only lead to a more systematic classification of local facts and to fuller observation of the habits of local species, but that this knowledge itself will in its turn test the applicability of the theory, or suggest the directions in which it may need modification.

(14) Problem VII. *To determine the distribution after a first migration into a cleared circular area.*

Let the radius of the cleared area be  $a$ . Then at distance  $c$  from the centre, inside or outside the circle of radius  $a$ , the distribution  $F_n(c)$  is given by:



$$\begin{aligned}
 F_n(c) &= N \int_a^\infty \int_0^{2\pi} \phi_n(c^2 + r^2 - 2rc \cos \theta) r d\theta dr \dots\dots\dots(lxxvi) \\
 &= \frac{N}{2\pi} Q_t \int_a^\infty \int_0^{2\pi} \frac{1}{\sigma^2} e^{-\frac{1}{2}(c^2+r^2)/\sigma^2} e^{-(rc \cos \theta)/\sigma^2} r d\theta dr \\
 &= \frac{N}{2\pi} Q_t \frac{e^{-\frac{1}{2}c^2/\sigma^2}}{\sigma^2} \int_a^\infty \int_0^{2\pi} e^{-\frac{1}{2}r^2/\sigma^2} r \sum_0^\infty \frac{1}{m!} \left(\frac{r}{\sigma}\right)^m \left(\frac{c}{\sigma}\right)^m \cos^m \theta d\theta dr.
 \end{aligned}$$

$$\begin{aligned}
 \text{Now } \int_0^{2\pi} \cos^m \theta d\theta &= 0, \text{ if } m \text{ be odd, and } = 4 \int_0^{\pi/2} \cos^{2s} \theta d\theta \\
 &= 4 \frac{(2s-1)(2s-3) \dots 1}{2s(2s-2) \dots 2} \frac{\pi}{2} = 2\pi \frac{2s!}{(2^s s!)^2},
 \end{aligned}$$

if  $m$  be even and  $= 2s$ .

Hence :

$$F_n(c) = N Q_t \frac{e^{-\frac{1}{2}c^2/\sigma^2}}{\sigma^2} \int_a^\infty e^{-\frac{1}{2}r^2/\sigma^2} r \sum_0^\infty \left\{ \frac{(r^2)^s}{(\sigma^2)^s} \frac{(c^2)^s}{(\sigma^2)^s} \frac{1}{(2^s s!)^2} \right\} dr \dots\dots\dots(lxxvii).$$

$$\begin{aligned}
 \text{Now } \int_a^\infty e^{-\frac{1}{2}r^2/\sigma^2} \frac{r^{2s+1}}{\sigma^{2s+1}} d\left(\frac{r}{\sigma}\right) &= \int_{a/\sigma}^\infty e^{-\frac{1}{2}z^2} z^{2s+1} dz \\
 &= M_{2s+1}(a/\sigma) \dots\dots\dots(lxxviii).
 \end{aligned}$$

$M_{2s+1}(a/\sigma)$  is thus the  $(2s+1)$ th moment of the 'tail' of a normal or Gaussian curve of errors (multiplied by  $\sqrt{2\pi}$ ) about its axis. Its values have been tabled for  $s=1, 2, 3$  and  $4$ .

Thus we have :

$$F_n(c) = N Q_t e^{-\frac{1}{2}c^2/\sigma^2} \sum_0^\infty \frac{(c^2)^s}{(\sigma^2)^s} \frac{M_{2s+1}(a/\sigma)}{(2^s s!)^2} \dots\dots\dots(lxxix).$$

But it is easy to see that :

$$M_{2s+1}(a/\sigma) = 2^s s! e^{-\frac{1}{2}a^2/\sigma^2} \left\{ 1 + \frac{1}{2} \frac{a^2}{\sigma^2} + \frac{1}{2 \cdot 4} \left(\frac{a^2}{\sigma^2}\right)^2 + \dots + \frac{1}{2^s s!} \left(\frac{a^2}{\sigma^2}\right)^s \right\}.$$

Accordingly :

$$F_n(c) = N Q_t e^{-\frac{1}{2}(c^2+a^2)/\sigma^2} \sum_0^\infty \frac{(c^2)^s}{(\sigma^2)^s} \left\{ 1 + \frac{1}{2} \frac{a^2}{\sigma^2} + \frac{1}{2 \cdot 4} \left(\frac{a^2}{\sigma^2}\right)^2 + \dots + \frac{1}{2^s s!} \left(\frac{a^2}{\sigma^2}\right)^s \right\} \frac{1}{2^s s!} \dots (lxxx).$$

The successive differentiations of this expression with regard to  $t = \sigma^2$ , involved in the operator  $Q_t$ , which are needful if we wish to give the corrections to the Rayleigh solution, are straightforward but extremely laborious. We can throw the solution into other forms.

Write :  $\epsilon_1 = \frac{1}{2} c^2/\sigma^2, \quad \epsilon_2 = \frac{1}{2} a^2/\sigma^2,$

then we have :

$$\begin{aligned}
 F_n(c) &= N Q_t e^{-(\epsilon_1+\epsilon_2)} \sum_0^\infty \frac{\epsilon_1^s}{s!} \left( 1 + \epsilon_2 + \frac{\epsilon_2^2}{2!} + \dots + \frac{\epsilon_2^s}{s!} \right) \\
 &= N Q_t e^{-\epsilon_1} \sum_0^\infty \frac{\epsilon_1^s}{(s!)^2} \int_{\epsilon_2}^\infty x^s e^{-x} dx \dots\dots\dots(lxxxix).
 \end{aligned}$$



Here  $\int_{\epsilon_2}^{\infty} x^s e^{-x} dx$  is the incomplete  $\Gamma$ -function for an integer value of  $s$ . This can be found fairly easily from the above series, or may be determined from tables of the incomplete  $\Gamma$ -function which it is hoped may be shortly published.

Again: 
$$J_0(2i\sqrt{z}) = \sum_0^{\infty} \frac{z^s}{(s!)^2},$$

hence we have:

$$F_n(c) = NQ_t e^{-c} \int_{\epsilon_2}^{\infty} J_0(2i\sqrt{\epsilon_1 x}) e^{-x} dx \dots\dots\dots (lxxxii),$$

a very concise form, which does not, however, simplify the calculations. Integrate by parts and we have:

$$\begin{aligned} F_n(c) &= NQ_t e^{-(c_1+c_2)} \sum_0^{\infty} \frac{d^s}{d\epsilon_2^s} \{J_0(2i\sqrt{\epsilon_1 \epsilon_2})\} \\ &= NQ_t e^{-(c_1+c_2)} \sum_0^{\infty} \epsilon_1^s \frac{d^s}{d(\epsilon_1 \epsilon_2)^s} \{J_0(2i\sqrt{\epsilon_1 \epsilon_2})\}. \end{aligned}$$

But 
$$\frac{d^s}{dz^s} J_0(2i\sqrt{z}) = \sum_0^{\infty} \frac{z^q}{q! (q+s)!} = \frac{J_s(2i\sqrt{z})}{(i\sqrt{z})^s},$$

or: 
$$\begin{aligned} F_n(c) &= NQ_t e^{-(c_1+c_2)} \sum_0^{\infty} \left\{ \epsilon_1^s \frac{J_s(2i\sqrt{\epsilon_1 \epsilon_2})}{(i\sqrt{\epsilon_1 \epsilon_2})^s} \right\} \\ &= NQ_t e^{-(c_1+c_2)} \sum_0^{\infty} \left( \sqrt{\frac{-\epsilon_1}{\epsilon_2}} \right)^s J_s(2i\sqrt{\epsilon_1 \epsilon_2}) \dots\dots\dots (lxxxiii). \end{aligned}$$

This is the solution in Bessel's functions, and inside the cleared area, where  $\epsilon_2$  is greater than  $\epsilon_1$ , would give fairly good results if tables of the higher Bessel's functions for imaginary values of the argument were available.

We can also express the solution in terms of  $\omega$ -functions as follows:

Write 
$$I_s(a) = \int_0^{\infty} \left( \frac{1}{2} \frac{r^2}{\sigma^2} \right)^s e^{-\frac{1}{2} r^2 / \sigma^2} \frac{r dr}{\sigma^2},$$

$$E_s(c) = \frac{1}{s!} e^{-\frac{1}{2} c^2 / \sigma^2} \left( \frac{1}{2} \frac{c^2}{\sigma^2} \right)^s.$$

Then 
$$F_n(c) = NQ_t \sum_0^{\infty} \frac{1}{s!} I_s(a) E_s(c).$$

Now 
$$E_s(r) = \frac{1}{s!} e^{-\frac{1}{2} r^2 / \sigma^2} \left( \frac{1}{2} \frac{r^2}{\sigma^2} \right)^s = 2\pi\sigma^2 \sum_0^{\infty} b_{sp} \omega_{sp},$$

the  $b$ 's being undetermined constants, for dividing by the exponential factor we have an integer algebraic expression in  $r^2/\sigma^2$  on both sides. Multiply both sides by  $\chi_{sp} r dr$  and integrate between 0 and  $\infty$ ,  $p$  being = or  $<$   $s$ . Then:

$$\begin{aligned} \frac{1}{s!} \int_0^{\infty} e^{-\frac{1}{2} r^2 / \sigma^2} \chi_{sp} \left( \frac{1}{2} \frac{r^2}{\sigma^2} \right)^s r dr &= 2\pi\sigma^2 b_{sp} \int_0^{\infty} \chi_{sp} \omega_{sp} r dr, \\ \frac{2\pi\sigma^2}{s!} \int_0^{\infty} \omega_{sp} \frac{r^{2s+1} dr}{(2\sigma^2)^s} &= 2\pi\sigma^2 b_{sp} (p!)^2, \text{ by (xix) and (xxi).} \end{aligned}$$



Therefore by (xvi):

$$\begin{aligned}
 b_{sp} &= \frac{1}{s!(p!)^2} (p-1-s)(p-2-s) \dots (-s) (-1)^{s-1} \int_{-0}^{-\infty} \beta^{-s-z} e^{1/\beta} d\beta \\
 &= (-1)^p \frac{s(s-1) \dots (s-p+1)}{s!(p!)^2} \int_0^{\infty} x^s e^{-x} dx, \text{ if } x = -1/\beta, \\
 &= (-1)^p \frac{s(s-1) \dots (s-p+1)}{(p!)^2}.
 \end{aligned}$$

Thus 
$$\begin{aligned}
 E_s(r) &= 2\pi\sigma^2 \left\{ \omega_0 - s\omega_2 + \frac{s(s-1)}{(2!)^2} \omega_4 - \frac{s(s-1)(s-2)}{(3!)^2} \omega_6 + \dots \right\} \dots (\text{lxxxiv}) \\
 &= 2\pi\sigma^2 U_s(r)^*, \text{ say.}
 \end{aligned}$$

Now consider:

$$\begin{aligned}
 \int_r^{\infty} \omega_{2s} \frac{rdr}{\sigma^2} &= (-1)^s (\sigma^2)^{s-1} \frac{d^s}{d(\sigma^2)^s} \int_r^{\infty} \omega_0 r dr \\
 &= (-1)^s (\sigma^2)^{s-1} \frac{d^s}{d(\sigma^2)^s} \left[ -\frac{1}{2\pi} e^{-\frac{1}{2}r^2/\sigma^2} \right]_r^{\infty} \\
 &= (-1)^s (\sigma^2)^{s-1} \frac{d^s}{d(\sigma^2)^s} (\omega_0 \sigma^2) \\
 &= \omega_{2s} - s\omega_{2s-2} \dots \dots \dots (\text{lxxxv}).
 \end{aligned}$$

We can now express  $I_s(r)$  in terms of  $\omega$ -functions.

We have:

$$\begin{aligned}
 I_s(r) &= \int_r^{\infty} s! E_s(r) \frac{rdr}{\sigma^2} \\
 &= s! 2\pi\sigma^2 \int_r^{\infty} \left\{ \omega_0 - s\omega_2 + \frac{s(s-1)}{(2!)^2} \omega_4 - \frac{s(s-1)(s-2)}{(3!)^2} \omega_6 + \dots \right\} \frac{rdr}{\sigma^2} \\
 &= s! 2\pi\sigma^2 (s+1) \left\{ \omega_0 - \frac{s\omega_2}{1!2!} + \frac{s(s-1)}{2!3!} \omega_4 - \frac{s(s-1)(s-2)}{3!4!} \omega_6 + \dots \right\} \\
 &= s! 2\pi\sigma^2 (s+1) V_s(r).
 \end{aligned}$$

Thus 
$$F_n(c) = NQ_t 4\pi^2 \sigma^4 \overset{\infty}{S}_0((s+1)U_s(c)V_s(a)) \dots \dots \dots (\text{lxxxvi}),$$

where:

$$\begin{aligned}
 U_s(r) &= \omega_0 - \frac{s}{(1!)^2} \omega_2 + \frac{s(s-1)}{(2!)^2} \omega_4 - \frac{s(s-1)(s-2)}{(3!)^2} \omega_6 + \dots, \\
 V_s(r) &= \omega_0 - \frac{s}{1!2!} \omega_2 + \frac{s(s-1)}{2!3!} \omega_4 - \frac{s(s-1)(s-2)}{3!4!} \omega_6 + \dots,
 \end{aligned}$$

a result which allows of fairly rapid determination from tables of  $\sigma^2 \omega_{2s}$ .

There is, perhaps, less difficulty in this form in allowing for the first term or two of the operator  $Q_t$ , for  $U_s(r)$  and  $V_s(r)$  can be at once differentiated with regard to  $\sigma^2$ , but even then the final result has considerable complexity.

\* This result involves the expression of any power of  $r^2$  in  $\chi$ -functions.



The Rayleigh solution value is easily found by putting  $Q_t=1$  in any of the forms of (lxxix), (lxxx), (lxxxii), (lxxxiii) or (lxxxvi).

A case of peculiar interest arises when  $c=0$ , or we take the density at the centre of the clearance. In this instance we have:

$$F_n(0) = NQ_t e^{-\frac{1}{2}a^2/\sigma^2}.$$

Now

$$Q_t = 1 + \nu_4 (\sigma^2)^2 \frac{d^2}{d(\sigma^2)^2} - \nu_6 (\sigma^2)^3 \frac{d^3}{d(\sigma^2)^3} + \dots + (-1)^s \nu_{2s} (\sigma^2)^s \frac{d^s}{d(\sigma^2)^s} + \dots$$

and

$$e^{-\frac{1}{2}a^2/\sigma^2} = 2\pi\sigma^2 \omega_0,$$

therefore 
$$\begin{aligned} (\sigma^2)^s \frac{d^s}{d(\sigma^2)^s} (e^{-\frac{1}{2}a^2/\sigma^2}) &= 2\pi\sigma^2 \left\{ \frac{d^s \omega_0}{d(\sigma^2)^s} + s \frac{d^{s-1} \omega_0}{d(\sigma^2)^{s-1}} \right\} (\sigma^2)^s \\ &= 2\pi\sigma^2 \{ (-1)^s \omega_{2s} + s (-1)^{s-1} \omega_{2(s-1)} \} \\ &= e^{-\frac{1}{2}a^2/\sigma^2} (-1)^s \{ \chi_{2s} - s \chi_{2(s-1)} \}. \end{aligned}$$

Thus

$$\begin{aligned} F_n(0) &= N e^{-\frac{1}{2}a^2/\sigma^2} \{ 1 - 2\nu_4 \chi_2 + (\nu_4 - 3\nu_6) \chi_4 + (\nu_6 - 4\nu_8) \chi_6 + (\nu_8 - 5\nu_{10}) \chi_8 + \dots \} \\ &= 2\pi\sigma^2 N (\omega_0 - 2\nu_4 \omega_2 + (\nu_4 - 3\nu_6) \omega_4 + (\nu_6 - 4\nu_8) \omega_6 + (\nu_8 - 5\nu_{10}) \omega_8 + \dots) \dots (lxxxvii). \end{aligned}$$

We are also able to consider the secondary problem:

*What is the distribution into unoccupied space surrounding a uniformly occupied circular area due to a first migration?*

Let the radius of the area be  $a$  and let the density at any distance  $c$  be  $F_n(c)$  after the first migration. Then clearly, if all space were uniformly filled, we should have uniformity after the first migration, or:

$$F_n(c) + F_n(c) = N,$$

hence: 
$$F_n(c) = N - F_n(c) \dots \dots \dots (lxxxviii).$$

The solution is thus thrown back on the solution obtained for the previous problem. In particular at the centre of the populated area we have:

$$F_n(0) = N - F_n(0) \dots \dots \dots (lxxxix).$$

We are thus able to calculate the reduced central density due to a migration from the area to the surrounding unoccupied district, i.e. the effect on population of the spread outwards of a colony.

(15) Problem VIII. *Indirect solution of the General Problem of the Random Walk.*

It may not be without interest to put on record the distribution density after  $n$  flights in the case of a cleared circular area, if it be expressed in Kluyver's manner by the integral of a Bessel's function product.



We have:

$$F_n(c) = N \int_0^a \int_0^{2\pi} \phi_n(c^2 + r^2 - 2rc \cos \theta) r d\theta dr,$$

and

$$\begin{aligned} F_n(c) &= N - F_n(c) \\ &= N \left\{ 1 - \frac{1}{2\pi} \int_0^a \int_0^{2\pi} \int_0^\infty u J_0(u\sqrt{c^2 + r^2 - 2cr \cos \theta}) J_0(ul)^n du r d\theta dr \right\}, \\ &\qquad\qquad\qquad \text{by (iii),} \\ &= N \left[ 1 - \int_0^a \int_0^\infty u J_0(ur) J_0(uc) \{J_0(ul)\}^n du r dr \right], \\ &\qquad\qquad\qquad \text{by Neumann's Theorem (see p. 6)} \\ &= N \left[ 1 - \int_0^\infty \frac{\{J_0(ul)\}^n}{u} \left\{ \int_0^a J_0(ur) ur d(ur) \right\} J_0(uc) du \right] \\ &= N \left[ 1 - \int_0^\infty \frac{\{J_0(ul)\}^n}{u} \int_0^a d\{J_1(ur) ur\} J_0(uc) du \right], \\ &\qquad\qquad\qquad \text{by the theorem cited on p. 7,} \\ &= N \left[ 1 - \int_0^\infty \frac{\{J_0(ul)\}^n}{u} \left\{ ur J_1(ur) \right\}_0^a J_0(uc) du \right] \\ &= N \left[ 1 - \int_0^\infty \{J_0(ul)\}^n a J_1(ua) J_0(uc) du \right]. \end{aligned}$$

Or, writing  $v = au$ , we have:

$$F_n(c) = N \left[ 1 - \int_0^\infty J_1(v) J_0\left(v \frac{c}{a}\right) \left\{ J_0\left(v \frac{l}{a}\right) \right\}^n dv \right] \dots\dots\dots \text{(xc).}$$

This expression is concise. The integral expresses the probability that if an individual start from the origin and take  $(n+1)$  flights, the first of magnitude  $c$  and the remaining  $n$  of magnitude  $l$ , at random, he will find himself within a distance  $a$  of his starting point. But there does not seem any convenient method of evaluating the integral. Comparing with (lxxxiii) we have the curious identity:

$$\int_0^\infty J_1(v) J_0\left(v \frac{c}{a}\right) \left\{ J_0\left(v \frac{l}{a}\right) \right\}^n dv = 1 - Q_n e^{-(a^2+c^2)/nl^2} S_0^\infty \left( i \frac{c}{a} \right)^s J_s \left( 2i \frac{ac}{nl^2} \right) \dots \text{(xci).}$$

Write  $c=l$ ,  $a=r$  and  $n-1$  for  $n$ , then

$$\int_0^\infty J_1(v) \left\{ J_0\left(v \frac{l}{r}\right) \right\}^n dv = 1 - Q_{n-1} e^{-(r^2+l^2)/((n-1)l^2)} S_0^\infty \left( i \frac{l}{r} \right)^s J_s \left( 2i \frac{r}{(n-1)l} \right),$$

where  $Q_{n-1}$  is the operator,

$$1 + \nu_1(n-1)^2 \frac{d^2}{dn^2} - \nu_2(n-1)^2 \frac{d^3}{dn^3} + \dots + (-1)^s \nu_{2s}(n-1)^s \frac{d^s}{dn^s} + \dots,$$



or, by (iv),  $P_n(r)$ , the chance that an individual taking  $n$  flights from a centre should be found within a distance  $r$  from that centre, is:

$$P_n(r) = N \left\{ 1 - Q_{n-1} e^{-\frac{(r^2+l^2)}{(n-1)l^2}} \sum_0^{\infty} \left( i \frac{l}{r} \right)^s J_s \left( 2i \frac{r}{(n-1)l} \right) \right\} \dots\dots(\text{xcii}).$$

Since 
$$\phi_n(r^2) = \frac{1}{2\pi r} \frac{d}{dr} \{P_n(r)\},$$

we have here the complete analytical solution in known functions—*i.e.* the Bessel's functions with imaginary arguments—of my original problem of the random walk. But this formal solution provides no better method for shortly determining the dispersal curves than that already indicated in these pages.

(16) Problem IX. *To find the distribution after  $m$  migrations each of  $n$  flights, there being originally a circular clearance which is not kept sterile.*

The solution is found by writing  $m\sigma^2$  for  $\sigma^2$ , putting the  $N$ 's for the  $\nu$ 's in  $Q_t$  which becomes  $Q_t^m$ , and multiplying by the factor  $(\mu\Delta)^{m-1}$  assumed to be constant. This can be done to any of the forms (lxxix)—(lxxxiii), or (lxxxvi). If we write:

$$\bar{\epsilon}_1 = \epsilon_1/m \text{ and } \bar{\epsilon}_2 = \epsilon_2/m$$

we find:

$${}_mF_n(c) = N (\mu\Delta)^{m-1} Q_t^m e^{-\bar{\epsilon}_1} \sum_0^{\infty} \left( \frac{\bar{\epsilon}_1^s}{(s!)^2} \int_{\bar{\epsilon}_1}^{\infty} x^s e^{-x} dx \right) \dots\dots\dots(\text{xciii}),$$

or: 
$${}_mF_n(c) = N (\mu\Delta)^{m-1} Q_t^m e^{-\bar{\epsilon}_1} \int_{\bar{\epsilon}_1}^{\infty} J_0(2i\sqrt{\bar{\epsilon}_1 x}) e^{-x} dx \dots\dots\dots(\text{xciv}).$$

Or again:

$${}_mF_n(c) = N (\mu\Delta)^{m-1} Q_t^m e^{-\bar{\epsilon}_1 + \bar{\epsilon}_2} \sum_0^{\infty} \left( \sqrt{-\frac{\bar{\epsilon}_1}{\bar{\epsilon}_2}} \right)^s J_s(2i\sqrt{\bar{\epsilon}_1 \bar{\epsilon}_2}) \dots\dots\dots(\text{xcv}),$$

$${}_mF_n(c) = N (\mu\Delta)^{m-1} Q_t^m 4\pi^2 m^2 \sigma^4 \sum_0^{\infty} \left( U_s \left( \frac{c}{\sqrt{m}} \right) V_s \left( \frac{a}{\sqrt{m}} \right) (s+1) \right) \dots\dots(\text{xcvi}).$$

Of these, I have found the first quite as convenient as any other to obtain numerical results from. I shall now illustrate the circular patch formulae.

*Illustration I.* A circular patch  $\frac{1}{2}$  mile radius is cleared of mosquitoes but not kept sterile. To find the density at the centre, at  $\frac{1}{4}$  mile from the centre, and at the margin after ten breeding cycles.

We shall suppose as before  $l = 200$  yards,  $n = 6$ , and therefore

$$\sigma^2 = 120,000 \text{ square yards. } \epsilon_1 = \frac{1}{2} \frac{a^2}{m\sigma^2} = .3227.$$

The second term in  $Q_t^m$  will be of the order  $\frac{1}{240}$  of the first and I shall neglect it. Accordingly the solution may be taken



$${}_m F_n(c) = e^{-(\bar{\epsilon}_1 + \bar{\epsilon}_2)} (\mu\Delta)^{m-1} N \left( 1 + \bar{\epsilon}_1 (1 + \bar{\epsilon}_2) + \frac{\bar{\epsilon}_1^2}{2!} \left( 1 + \bar{\epsilon}_2 + \frac{\bar{\epsilon}_2^2}{2!} \right) + \frac{\bar{\epsilon}_1^3}{3!} \left( 1 + \bar{\epsilon}_2 + \frac{\bar{\epsilon}_2^2}{2!} + \frac{\bar{\epsilon}_2^3}{3!} \right) + \dots \right).$$

The successive bracketted terms in  $\bar{\epsilon}_2$  are

$$1.3227, \quad 1.3748, \quad 1.3804, \quad 1.3809 \text{ and } 1.3809,$$

which is equal to  $e^{+\bar{\epsilon}_2}$  to our number of decimal places. Hence we may put

$$\begin{aligned} {}_{10} F_6(c) &= (\mu\Delta)^9 N e^{-(\bar{\epsilon}_1 + \bar{\epsilon}_2)} \left\{ 1 + \bar{\epsilon}_1 (e^{\bar{\epsilon}_2} - 0.582) + \frac{\bar{\epsilon}_1^2}{2!} (e^{\bar{\epsilon}_2} - 0.061) \right. \\ &\quad \left. + \frac{\bar{\epsilon}_1^3}{3!} (e^{\bar{\epsilon}_2} - 0.0005) + \sum_{s=4}^{\infty} \frac{\bar{\epsilon}_1^s}{s!} e^{\bar{\epsilon}_2} \right\} \\ &= (\mu\Delta)^9 N e^{-(\bar{\epsilon}_1 + \bar{\epsilon}_2)} (1 - e^{\bar{\epsilon}_2} + e^{(\bar{\epsilon}_1 + \bar{\epsilon}_2)} - 0.582\bar{\epsilon}_1 - 0.0030\bar{\epsilon}_1^2 - 0.0001\bar{\epsilon}_1^3) \\ &= (\mu\Delta)^9 N \{1 - e^{-\bar{\epsilon}_1} + e^{-(\bar{\epsilon}_1 + \bar{\epsilon}_2)} (1 - 0.582\bar{\epsilon}_1 - 0.0030\bar{\epsilon}_1^2 - 0.0001\bar{\epsilon}_1^3)\}. \end{aligned}$$

At centre

$${}_{10} F_6(0) = (\mu\Delta)^9 N \{e^{-0.3227} (1)\} = (\mu\Delta)^9 N \cdot 724.$$

We can test the accuracy of this result by using Equation (lxxvii) which, if we put  $\nu_1 = N_1$ , gives:

$${}_{10} F_6(0) = (\mu\Delta)^9 N e^{-\bar{\epsilon}_1} (1 - 2N_1\chi_2 + \dots)$$

and

$$\begin{aligned} \chi_2 &= 1 - \bar{\epsilon}_2 = (\mu\Delta)^9 N e^{-0.3227} \left( 1 + \frac{0.6773}{120} + \dots \right) \\ &= (\mu\Delta)^9 N \cdot 730. \end{aligned}$$

The agreement is accordingly good enough for practical purposes, and we may say that within a year the mosquitoes would at the centre of the patch have a density 73 per cent. of what they would have in uncleared country.

I now consider the density at a quarter of a mile from the centre,  $\bar{\epsilon}_1 = 0.807$ , and using the above formula we find:

$$\begin{aligned} {}_{10} F_6(440) &= (\mu\Delta)^9 N (1 - e^{-0.807} + e^{-0.807} \times 0.9953) \\ &= (\mu\Delta)^9 N \cdot 75, \end{aligned}$$

or, we see that at a quarter mile, midway between centre and boundary of the patch, the density is only 2 per cent. more than at the centre.

Finally, at the boundary itself,  $\bar{\epsilon}_1 = 0.3227 = \epsilon_2$ ,

$$\begin{aligned} {}_{10} F_6(880) &= (\mu\Delta)^9 N (1 - e^{-0.3227} + e^{-0.3227} \times 0.9809) \\ &= (\mu\Delta)^9 N \cdot 79. \end{aligned}$$

Thus the cleared patch would within the year have filled up with a population of mosquitoes varying in density from 73 per cent. at the centre to about 80 per cent. at the boundary, or the clearance without permanent sterility would have been quite ineffectual with the assumed values of the constants.



*Illustration II.* Let us assume precisely the same conditions as in the previous illustration, except that the area shall be supposed sterile, and we will consider what happens at the end of the first migration.

At the centre we have by Equation (lxxxvii):

$${}_1F_0(0) = Ne^{-c} \{1 - 2\nu_1\chi_2 + (\nu_1 - 3\nu_3)\chi_4 + (\nu_1 - 4\nu_5)\chi_6 + \dots\}.$$

But

$$\begin{aligned} -2\nu_1 &= \cdot083,333, & \chi_2(\epsilon_1) &= 1 - \epsilon_1 = -2\cdot227,000, \\ \nu_1 - 3\nu_3 &= -\cdot032,407, & \chi_4(\epsilon_1) &= 2 - 4\epsilon_1 + \epsilon_1^2 = -\cdot494,471, \\ \nu_1 - 4\nu_5 &= -\cdot005,498, & \chi_6(\epsilon_1) &= 6 - 18\epsilon_1 + 9\epsilon_1^2 - \epsilon_1^3 = 8\cdot031,303, \\ \nu_1 - 5\nu_{10} &= \cdot000,082, & \chi_8(\epsilon_1) &= 24 - 96\epsilon_1 + 72\epsilon_1^2 - 16\epsilon_1^3 + \epsilon_1^4, \\ & & \epsilon_1 &= 3\cdot227, & & = 34\cdot752,347. \end{aligned}$$

Hence:  ${}_1F_0(0) = Ne^{-c} (1 - \cdot185,583 + \cdot016,024 - \cdot044,156 + \cdot002,850)$   
 $= \cdot031N.$

This three per cent. of the density in uncleared area might possibly prove a trouble and on our assumptions it may be doubted whether the half-mile radius is sufficient. If we take the first term only, we find  $\cdot040N$ , or four per cent., not an important practical difference.

The introduction of even the first modifying term when  $c$  is not zero appears to lead to such complexity that I content myself with calculating the approximate value given by the Rayleigh solution for distances of  $\frac{1}{4}$  and  $\frac{1}{2}$  mile from the centre of the clearance. In this case  $\epsilon_1 = 3\cdot227$ ,  $\epsilon_2 = \cdot807$  and  $3\cdot227$  respectively half-way to and at the boundary. I proceed just as before and deduce the following approximate value for  ${}_1F_0(c)$ , *i.e.*

$$\begin{aligned} {}_1F_0(c) &= (\mu\Delta)^2 N \{1 - e^{-c} + e^{-(c_1+c_2)} (1 - 20\cdot9769\epsilon_1 - 7\cdot8851\epsilon_1^2 - \cdot2355\epsilon_1^3 \\ &\quad - \cdot0228\epsilon_1^4 - \cdot0016\epsilon_1^5 - \cdot0001\epsilon_1^6)\}. \end{aligned}$$

Hence

$${}_1F_0(440) = \cdot179 (\mu\Delta)^2 N, \text{ corresponding to } \epsilon_1 = \cdot807$$

and

$${}_1F_0(880) = \cdot709 (\mu\Delta)^2 N, \text{ corresponding to } \epsilon_1 = 3\cdot227.$$

Thus the density at  $\frac{1}{4}$  of a mile from the centre of the cleared patch would be some 18 per cent. of the density in uncleared country. In other words on our assumptions a clearance of one mile diameter, if kept sterile, would hardly suffice to keep an area of  $\frac{1}{2}$  mile diameter free of mosquitoes.

Compared with a straight boundary, where the density falls to about one half that of uncleared country at the boundary, we see that the bending of the boundary has a most marked effect in its neighbourhood, the curvature raising the boundary density from about 50 to 71 per cent. of the uncleared density. In fact the density is almost equal to the 75 per cent. in the boundary angle of a square clearance.



The differences between a square and a circular patch inscribed in it are noteworthy, indicating the marked influence of the area at the angles. Thus at the centre we have only 2 per cent. as against 3 per cent., and at  $\frac{1}{4}$  mile from the centre 11 per cent. as against 18 per cent.

As far as the above numerical investigations are to be looked upon as anything but illustrations of the nature of the calculations requisite to apply the theory of random migration to the mosquito clearance problem, they must be taken:

(i) As merely an incentive to further study of the manner in which mosquitoes scatter from the breeding ponds. It would seem possible, if difficult, to experimentally test this by in some way marking a large number of insects, and determining the nature and extent of the flight.

(ii) As indicating that permanent sterility of the protection belt is almost certainly needful. The  $\frac{1}{2}$  to 3 per cent. of mosquitoes at the centre of the clearance amounting to 6 to 18 per cent. at  $\frac{1}{4}$  mile distance may or may not be serious, but they certainly would very soon be if they were able to breed.

(iii) As showing that on the rough numbers taken, that a clearance belt of probably  $\frac{1}{2}$  mile round a settlement would be the minimum desirable sterile zone. But it is quite possible that, when the requisite constants are better known, it will be found that smaller belts will suffice. It is possibly rather an exaggerated view to suppose a mosquito to make six random flights of 200 yards between breeding spot and breeding spot. But certainly many insects I have noted will fly with great rapidity in one flight 50, 100 or 200 yards, and these flights are quite distinct from "flitters."

(17) *Conclusions.* The present memoir suffers of course from all the defects which must accompany a first attempt to develop a mathematical theory of phenomena which have hitherto not been studied with this development in view. The theory itself suggests hypotheses and constants which have never yet been considered. How far with a broad average of environment in relation to food supply, breeding places, shelter, foes, etc. is the spread of a species random? Are any of the geographical limits to plant or insect or animal life non-environmental and in course of change? If so, statistical studies of the density gradients of such species for a few miles either side of the supposed boundary would form most interesting work for biometricians. But, apart from this observational work, a good deal of experimental inquiry might be usefully attempted with regard to the constants of random scatter or flight in the cases of both seeds and insects.

On the theoretical side there are many problems left untouched. The present memoir has only opened up the outskirts of a very big field. It would be of value to investigate the number of terms in the expansion in  $\omega$ -functions requisite to practically reproduce the graphically constructed density distributions for migrations of 3, 4 or 5 flights. Our expansion to 6 terms is hardly close enough



for practical work until  $n=6$  or  $7$ . Many other shapes of populated or of cleared areas would provide problems of some interest, especially when the spread of the colony was limited in one or more directions by environmental barriers, such as sea, river or mountain range. The problem of sterile areas has been by no means exhausted, for in such cases I have only dealt with a result of the first migration, but actually there will be a second and later migrations in which not only new immigrants will appear but a portion of the first immigrants will be emigrants and again able to breed when they reach uncleared country. Our solution thus gives only a minimum limit to the percentages if the immigrants do not die at the end of the first breeding cycle. Much interest attaches also to cases in which the fertility and the death-rate are correlated with the density, *i.e.*  $\mu\Delta$  is not to be considered a constant. But in these as in other problems which suggest themselves, a further preliminary knowledge of some of the ecological constants suggested by the present enquiry would be an extremely valuable guide to the direction that research should take.

On the purely mathematical side the problem of the "random walk" may now be considered as fairly completely solved. The distribution curves have been determined until they pass into an analytical solution expressed by a new type of function. The expansion in these functions shows the limits to the accuracy of Lord Rayleigh's solution of a certain allied problem in the theory of sound. But the  $\omega$ -functions which have arisen in the enquiry have most interesting properties, and have led me to a whole series of allied functions of one and two variables which I propose to discuss on another occasion. The expansion in  $\omega$ -functions will I venture to think be found ultimately to have considerable importance for mathematical physics, especially in the evaluation of certain definite integrals which arise there. The possibility of practically carrying out such expansions depends on the determination of the successive moments (and products) of the original function, a process with which every statistician is now fairly familiar. But applied to definite mathematical functions it loses the disadvantage with which it is burdened in statistical practice—the high relative probable error of very high moments—and becomes closely allied to the process of determining the integral of the product of any function and a Legendre's coefficient (or solid harmonic). Should the generalised  $\omega$ -functions prove, as I anticipate, of some mathematical interest, it will be another illustration of how the need of the applied mathematician has thrust him, almost unawares, into the path of a novel functional development.



DIAGRAM I.

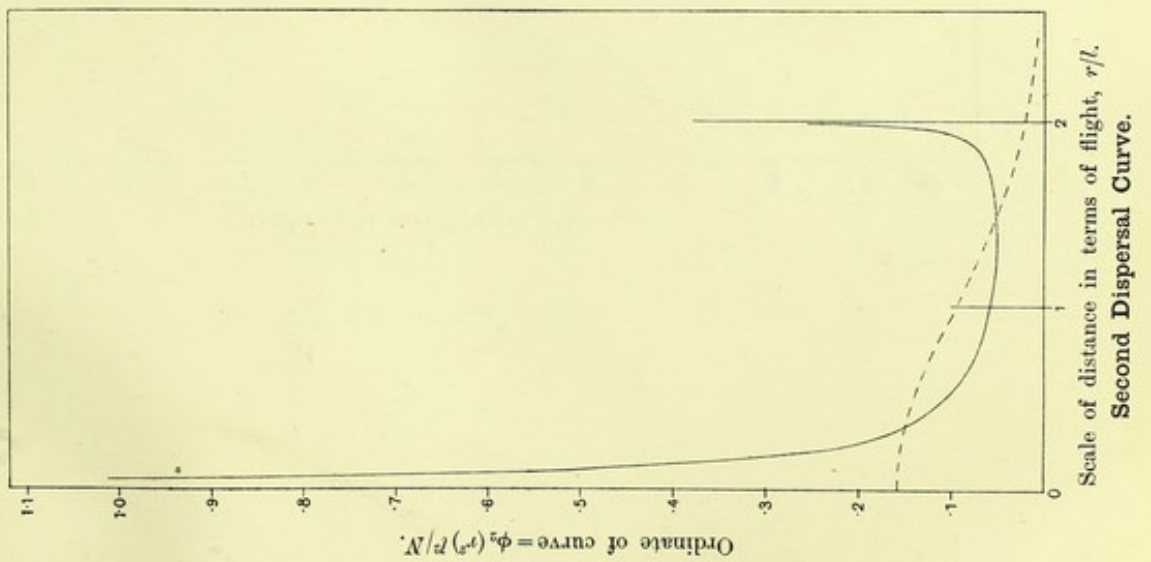


DIAGRAM II.

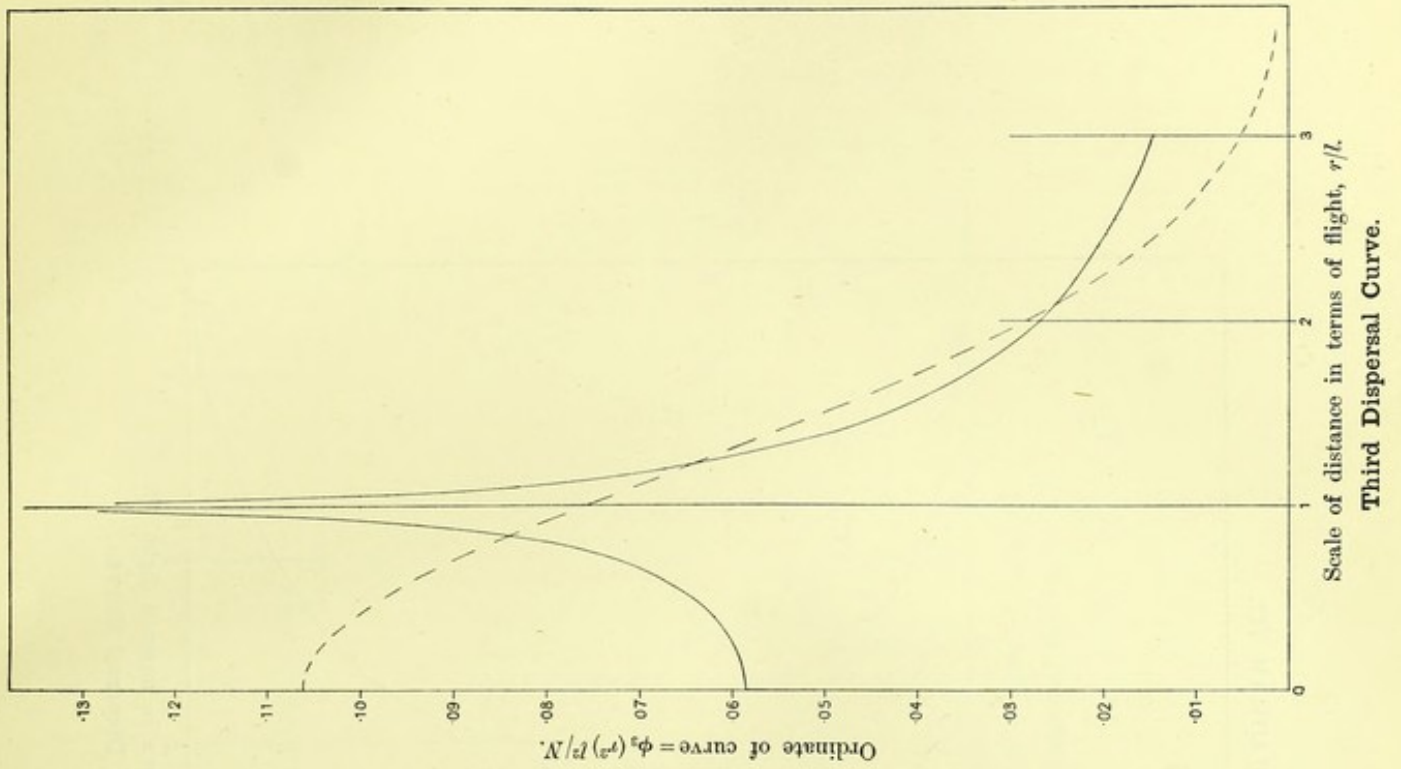




DIAGRAM III.

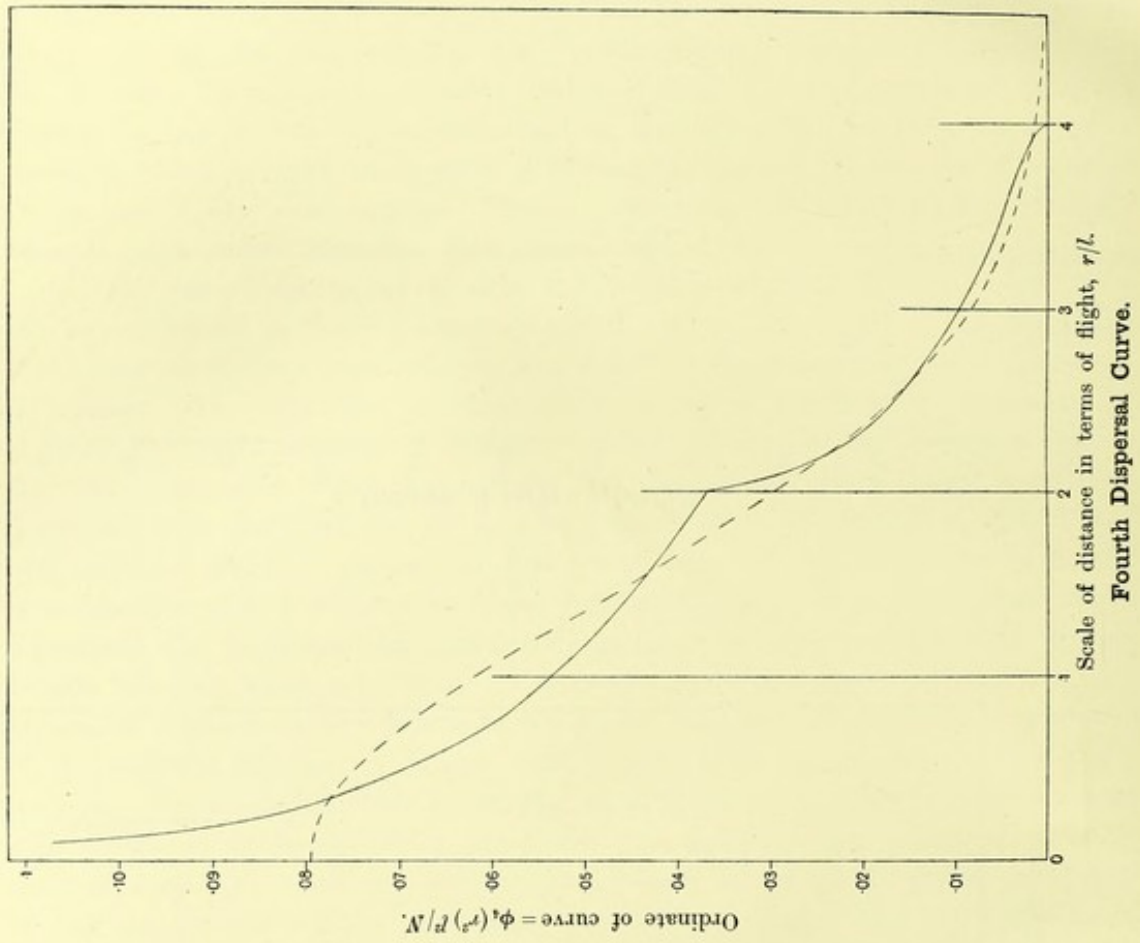




DIAGRAM IV.

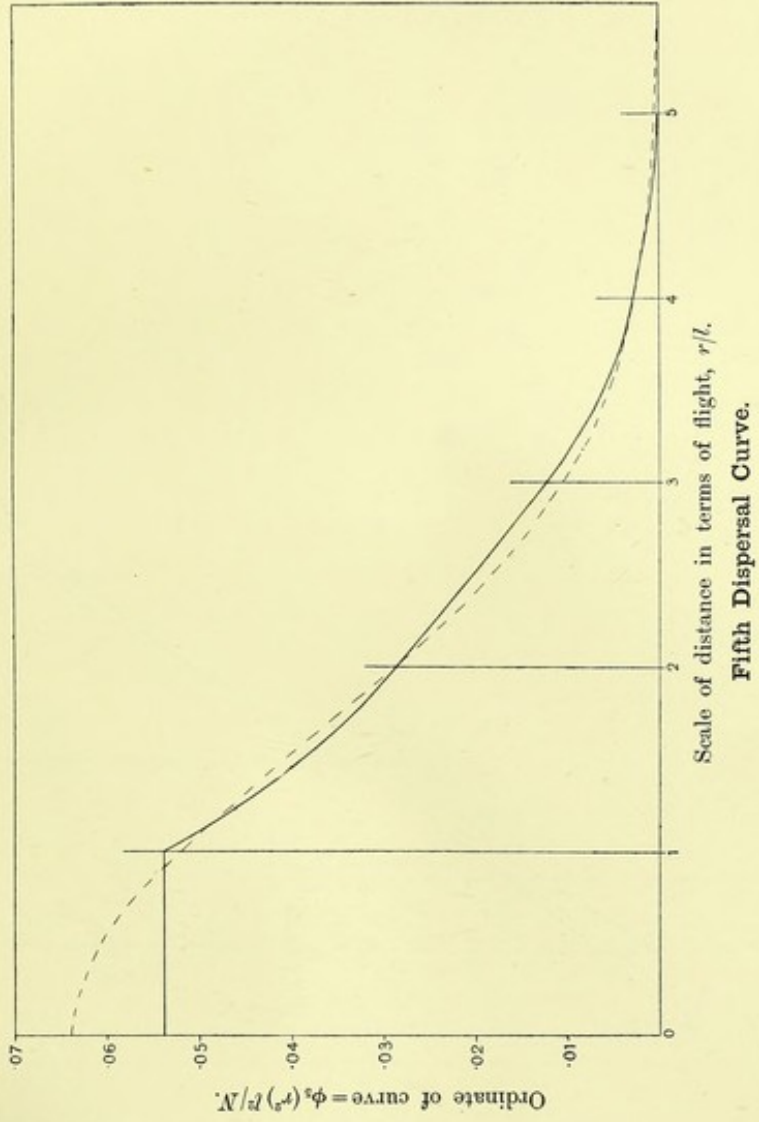




DIAGRAM V.

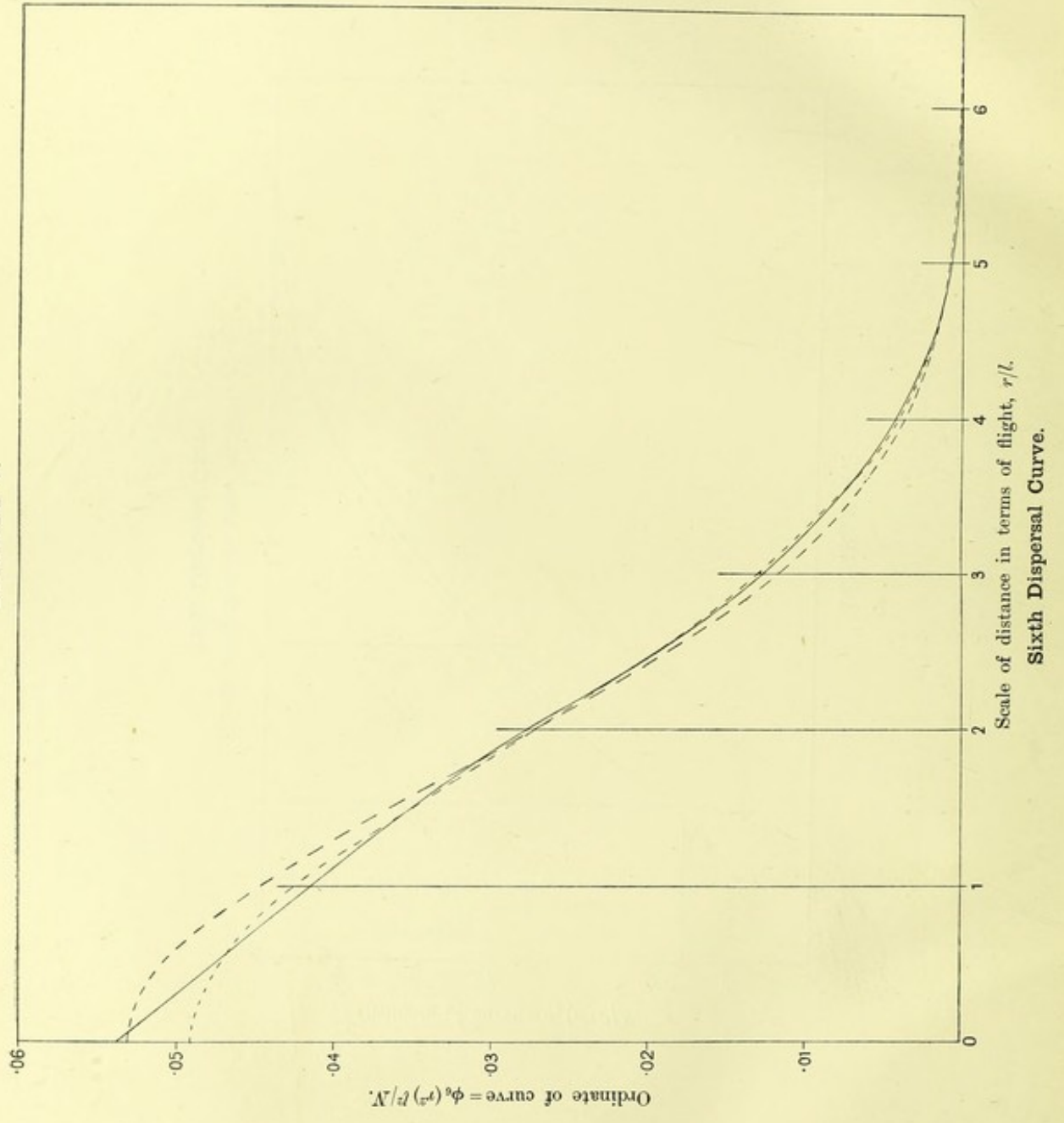
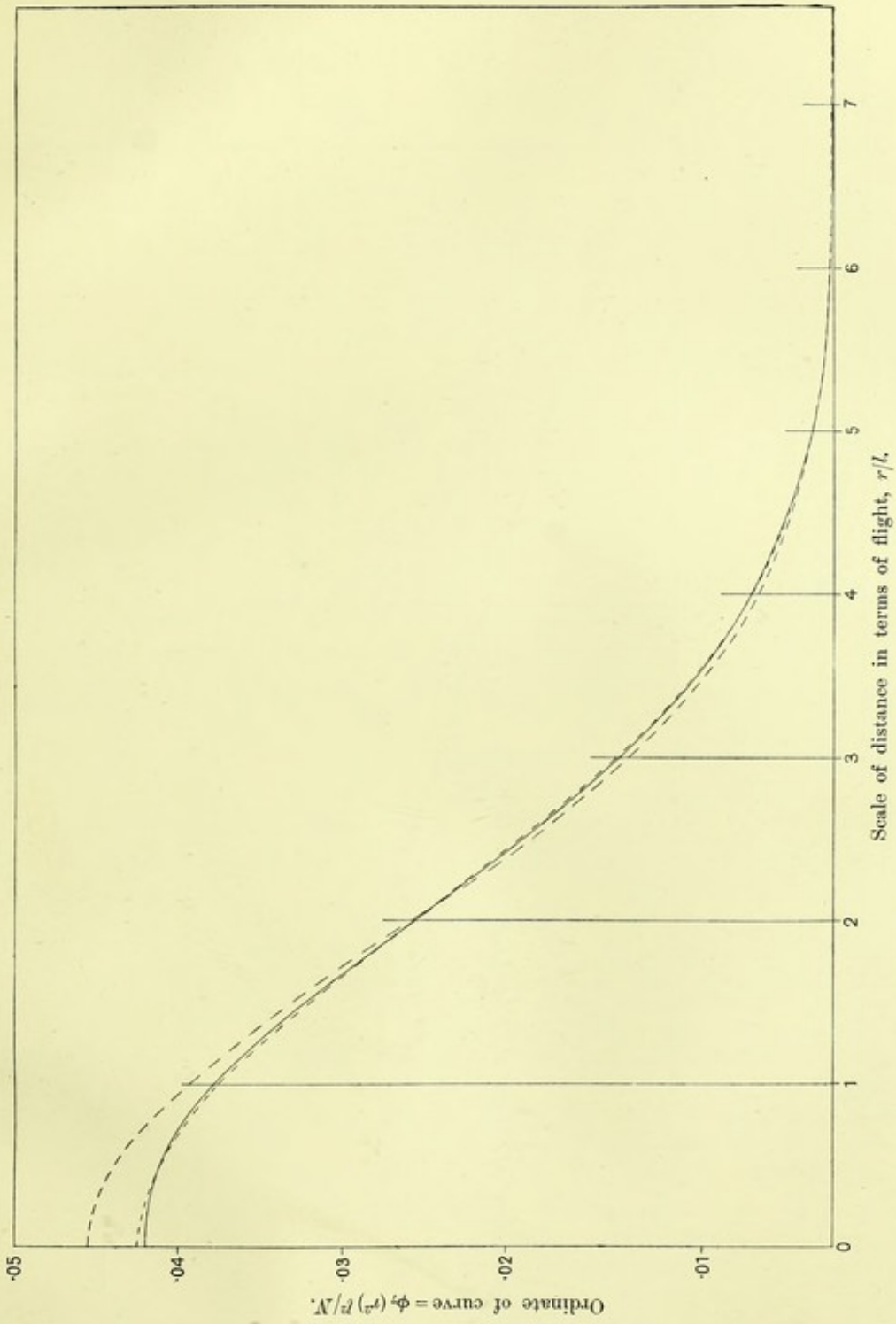




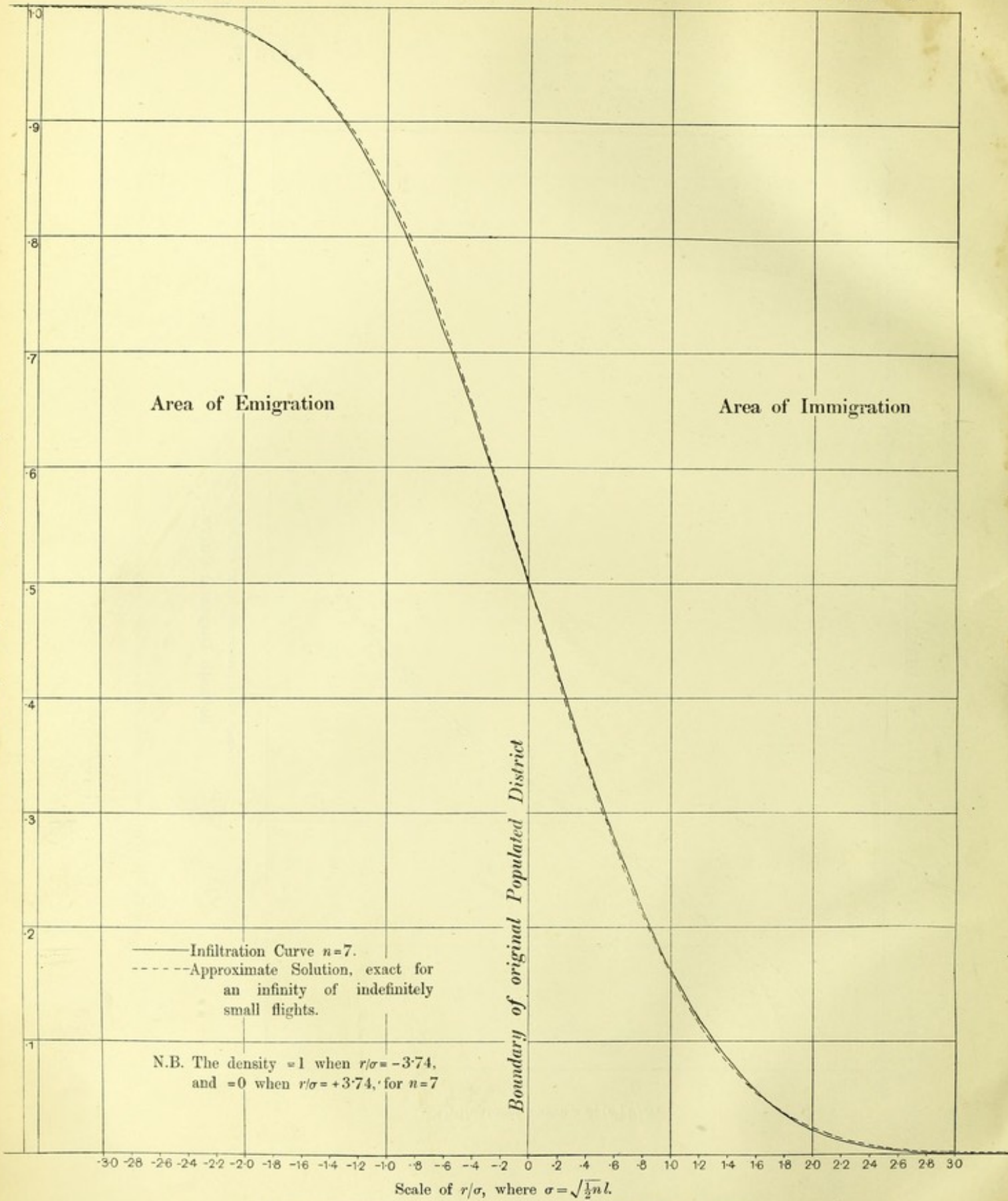
DIAGRAM VI.



Seventh Dispersal Curve.



Scale of density of Population. Original density  $N=1$ .



First order infiltration curve across a straight boundary.







PUBLISHED BY DULAU AND CO., SOHO SQUARE, LONDON, W.

**DRAPERS' COMPANY RESEARCH MEMOIRS.**  
DEPARTMENT OF APPLIED MATHEMATICS, UNIVERSITY COLLEGE,  
UNIVERSITY OF LONDON.

These memoirs will be issued at short intervals.

*Biometric Series.*

- I. Mathematical Contributions to the Theory of Evolution.—XIII. On the Theory of Contingency and its Relation to Association and Normal Correlation. By KARL PEARSON, F.R.S. *Issued.* Price 4s.
- II. Mathematical Contributions to the Theory of Evolution.—XIV. On the Theory of Skew Correlation and Non-linear Regression. By KARL PEARSON, F.R.S. *Issued.* Price 5s.
- III. Mathematical Contributions to the Theory of Evolution.—XV. On the Mathematical Theory of Random Migration. By KARL PEARSON, F.R.S., with the assistance of JOHN BLAKEMAN, M.Sc. *Issued.* Price 5s.
- IV. Mathematical Contributions to the Theory of Evolution.—XVI. On Homotypis in the Animal Kingdom. By ERNEST WARREN, D.Sc., ALICE LEE, D.Sc., EDNA LEA-SMITH, MARION RADFORD and KARL PEARSON, F.R.S. *Shortly.*

*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 3s.

*Technical Series.*

- I. On a Theory of the Stresses in Crane and Coupling Hooks with Experimental Comparison with Existing Theory. By E. S. ANDREWS, B.Sc.Eng., assisted by KARL PEARSON, F.R.S. *Issued.* Price 3s.
- II. On some Disregarded Points in the Stability of Masonry Dams. By L. W. ATCHERLEY, assisted by KARL PEARSON, F.R.S. *Issued.* Price 3s. 6d.
- III. On the Graphics of Metal Arches, with special reference to the Relative Strength of Two-pivoted, Three-pivoted and Built-in Metal Arches. By L. W. ATCHERLEY and KARL PEARSON, F.R.S. *Issued.* Price 6s.
- IV. On Torsional Vibrations in Axles and Shafting. By KARL PEARSON, F.R.S. *Issued.* Price 6s.
- V. A Further Study of the Stresses in Masonry Dams. By KARL PEARSON, F.R.S., and A. F. C. POLLARD, assisted by C. WHEEN. *Shortly.*

PUBLISHED BY THE CAMBRIDGE UNIVERSITY PRESS.

**BIOMETRIKA.**

A JOURNAL FOR THE STATISTICAL STUDY OF BIOLOGICAL PROBLEMS.

Founded by W. F. R. WELDON, FRANCIS GALTON and KARL PEARSON.

Edited by KARL PEARSON in Consultation with FRANCIS GALTON.

VOL. V., PARTS I. & II.

- |   |  |
|---|--|
| <ol style="list-style-type: none"><li>I. Walter Frank Raphael Weldon. 1860—1906. (With two Portraits and three Plates.)</li><li>II. Variation in <i>Chilomonas</i> under Favourable and Unfavourable Conditions. (With seven Figures in the text.) By RAYMOND PEARL.</li><li>III. The Non-Inheritance of Sex in Man. By FREDERICK ADAMS WOODS, M.D.</li><li>IV. On the Inheritance of the Sex-Ratio. By DAVID HERON, M.A.</li><li>V. A Second Study of the English Skull, with special Reference to the Moorfields Crania. (With superimposed Maps and 12 Plates of Crania, and four folding Tables.) By W. R. MACDONELL.</li><li>VI. On the Relationship of Intelligence to Size and Shape of Head, and to other Physical and Mental Characters. (With nine Figures in the text.) By KARL PEARSON, F.R.S.</li><li>VII. On the Relation between the Symmetry of the Egg and the Symmetry of the Embryo in the Frog (<i>Rana Temporaria</i>). (With twelve Figures in the text.) By J. W. JENKINSON, D.Sc.</li></ol> | <ol style="list-style-type: none"><li>(iii) On certain Points connected with scale Order in the Case of Correlation of two characters, which for some arrangement give a Linear Regression Line. By KARL PEARSON, F.R.S.</li><li>(iv) On the Classification of Frequency Ratios. (With one Figure in the text.) By D. M. Y. SOMMERVILLE, D.Sc.</li><li>(v) Note on the Significant or Non-Significant Character of a Sub-sample drawn from a Sample. By KARL PEARSON, F.R.S.</li><li>(vi) Professor Ziegler and Galton's Law of Ancestral Inheritance. By EDGAR SCHUSTER.</li><li>(vii) Variazione ed Omotiposi nelle infiorescenze di <i>Cichorium Intybus</i> L. (With two Figures in the text.) By Dr FERNANDO DE HELGUERO.</li><li>(viii) The Calculation of the Probable Errors of Certain Constants of the Normal Curve. By RAYMOND PEARL.</li><li>(ix) On the Probable Error of the Coefficient of Mean Square Contingency. By J. BLAKEMAN, M.Sc. and KARL PEARSON, F.R.S.</li><li>(x) On a Coefficient of Class Heterogeneity or Divergence. By KARL PEARSON, F.R.S.</li><li>(xi) Inheritance in the Female Line of Size of Litter in Poland China Sows. By G. M. ROMMEL and E. F. PHILLIPS.</li></ol> |
|---|--|

Miscellaneous.

- (i) Skew Frequency Curves. A Rejoinder to Professor Kapteyn. By KARL PEARSON, F.R.S.
- (ii) On the Curves which are most suitable for de-

scribing the Frequency of Random Samples of a Population. By KARL PEARSON, F.R.S.

The subscription price, payable in advance, is 30s. net per volume (post free); single numbers 10s. net. Volumes I., II., III. and IV. (1902-6) complete, 30s. net per volume. Bound in Buckram 34s. 6d. net per volume. Subscriptions may be sent to C. F. CLAY, Manager, Cambridge University Press Warehouse, Fetter Lane, London, E.C., either direct or through any bookseller.



DEPARTMENT OF APPLIED MATHEMATICS,  
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON

---

DRAPERS' COMPANY RESEARCH  
MEMOIRS.

BIOMETRIC SERIES. IV.

---

MATHEMATICAL CONTRIBUTIONS TO THE THEORY  
OF EVOLUTION.—XVI. ON FURTHER METHODS  
OF DETERMINING CORRELATION.

BY

KARL PEARSON, F.R.S.

LONDON:

PUBLISHED BY DULAU AND CO., 37, SOHO SQUARE, W.  
1907

*Price Four Shillings*







DEPARTMENT OF APPLIED MATHEMATICS,  
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON

---

DRAPERS' COMPANY RESEARCH  
MEMOIRS.

BIOMETRIC SERIES. IV.

---

MATHEMATICAL CONTRIBUTIONS TO THE THEORY  
OF EVOLUTION.—XVI. ON FURTHER METHODS  
OF DETERMINING CORRELATION.

BY

KARL PEARSON, F.R.S.

LONDON :

PUBLISHED BY DULAU AND CO., 37, SOHO SQUARE, W.

1907



DEPARTMENT OF APPLIED MATHEMATICS  
UNIVERSITY OF LONDON

DRABERS COMPANY RESEARCH  
MEMOIRS

Volume 10, Series IV

MATHEMATICAL CONDITIONS TO THE THEORY  
OF DIFFUSION IN A BARRIER METHOD  
OF DIFFUSION CORRELATION

1954

PRINTED BY THE UNIVERSITY OF LONDON  
PRINTERS



## *Mathematical Contributions to the Theory of Evolution.*

### XVI. ON FURTHER METHODS OF DETERMINING CORRELATION.

BY KARL PEARSON, F.R.S.

(1) *Introductory.* The object of the present paper is to give an account of some new methods of determining correlation. It is not suggested that they can with advantage replace the old processes, even when the distribution is approximately normal; to my mind the methods of determining the correlation ratio and the correlation coefficient ( $\eta$  and  $r$ ) based on moments and product moments stand foremost for the information they give and its weighable accuracy. At the same time there are series which are so short, or cases in which it is desirable to come rapidly to an approximate result or data which cannot be presented in a form suitable for product-moment working, where other methods are not only reasonable, but necessary. To such cases the present new methods apply. I have termed them *new* methods and I think this is legitimate. In the case of the first method, I have not seen any hint of it before. In the case of what I term grade methods, Dr Spearman has suggested that rank in a series should be the character correlated, but he has not taken this rank correlation as merely the stepping stone by which to reach the true correlation of the variables as dependent magnitudes, and further in the discussion he has given of the subject he has, I believe, given erroneous formulae and made quite incorrect statements as to the magnitude of probable errors.

One word must be said as to the use made of the normal distribution. I have used it here as on many other occasions as a means of suggesting fitting relations and simple formulae for correlation constants. This does not necessarily mean (i) that the constants reached may not have a perfectly definite meaning apart from normal distributions, or (ii) that the formulae obtained may not hold for all forms of distribution apart from normality. As an illustration of the first case I cite my mean square coefficient of contingency\*. This is a perfectly general measure of the deviation from independent probability in the case of an  $n \times m$  fold table, but its

\* *On the Theory of Contingency.* "Drapers' Research Memoirs, Biometric Series 1" (Dulau & Co., Soho Square, London).



actual form was selected so that it would agree with the coefficient of correlation in the case of indefinitely fine grouping and normal distribution. As an illustration of my second point I take the formulae given by me for the influence of selection on variation and correlation\*. These formulae were originally proved for normal distributions, but for a number of years past the proofs given in my lectures have been perfectly general, depending only on a more comprehensive definition of what we are to understand as correlation in the case of a complex of variables.

These points will be considered in the present treatment of correlation.

(2) *On Difference Methods of finding the Coefficient of Correlation.*

Let  $x$  and  $y$  be two correlated variables, each measured from their means  $m_x$  and  $m_y$  respectively. Then if  $v = x - y$ , and  $\sigma_x, \sigma_y, \sigma_v$  denote the three standard deviations

$$\sigma_v^2 = \sigma_x^2 + \sigma_y^2 - 2r_{xy}\sigma_x\sigma_y$$

and

$$r_{xy} = (\sigma_x^2 + \sigma_y^2 - \sigma_v^2) / (2\sigma_x\sigma_y) \dots\dots\dots(i).$$

This method of finding  $r_{xy}$  has long been in use as an alternative method to the product-moment method †.

It involves finding the mean-square difference of the values of the pairs of correlated characters. It is possible, however, to find  $r_{xy}$  from about one half these differences, if we assume the distribution to be normal.

More generally I proceed as follows. Suppose the function  $mx - ny$  formed, where  $m$  and  $n$  are at present indeterminate positive constants, and let the positive values only of this expression be taken and divided by the total frequency  $N$ . Then it will be possible to determine  $r$  from this result.

If  $z$  be the ordinate of the surface, then :

$$z = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{1}{2} \frac{1}{1-r^2} \left( \frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2} \right)} \dots\dots\dots(ii)$$

and we have the above result expressed analytically:

$$\frac{S(mx - ny)}{N} = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \iint (mx - ny) e^{-\frac{1}{2} \frac{1}{1-r^2} \left( \frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2} \right)} dydx \dots(iii).$$

The limits to  $y$  in order that  $mx - ny$  may be positive are  $y = -\infty$  to  $mx/n$ , and the limits of  $x$  will then be  $x = \infty$  to  $-\infty$ .

Put  $y' = y/\sigma_y$        $x' = x/\sigma_x$ ,  
then

$$\frac{S(mx - ny)}{N} = \frac{1}{2\pi} \frac{1}{\sqrt{1-r^2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{\frac{m\sigma_x}{n\sigma_y} x'} (m\sigma_x x' - n\sigma_y y') \times e^{-\frac{1}{2} \frac{1}{1-r^2} (x'^2 - 2rx'y' + y'^2)} dy' dx'.$$

\* *Phil. Trans. A*, Vol. 200, pp. 1-66.

† For example, *Phil. Trans. A*, Vol. 198, p. 242, and often elsewhere.



Write 
$$y'' = \frac{m\sigma_1}{n\sigma_2} x' - y'$$

and we have :

$$\frac{S(mx - ny)}{N} = \frac{1}{2\pi} \frac{n\sigma_2}{\sqrt{1-r^2}} \int_{-x}^{+\infty} \int_0^{\infty} y'' e^{-\frac{1}{2} \frac{1}{1-r^2} \left\{ x'^2 - 2rx' \left( \frac{m\sigma_1}{n\sigma_2} x' - y'' \right) + \left( \frac{m\sigma_1}{n\sigma_2} x' - y'' \right)^2 \right\}} dy'' dx''.$$

The order of integration can now be changed and we have :

$$\frac{S(mx - ny)}{N} = \frac{1}{2\pi} \frac{n\sigma_2}{\sqrt{1-r^2}} \int_0^{\infty} y'' e^{-\frac{1}{2} \frac{y''^2}{a^2}} \int_{-x}^{+\infty} e^{-\frac{1}{2} \frac{1}{\beta^2} (x' - \gamma y'')^2} dx' dy'',$$

where if  $\epsilon = m\sigma_1/(n\sigma_2)$ ,

$$\frac{1}{a^2} = \frac{1 - 2r\epsilon(1 - \epsilon) - r^2\epsilon^2}{(1 - 2r\epsilon + \epsilon^2)(1 - r^2)}, \quad \frac{1}{\beta^2} = \frac{1 - 2r\epsilon + \epsilon^2}{1 - r^2},$$

$$\gamma = \epsilon(1 - r)/(1 - 2r\epsilon + \epsilon^2).$$

But the integral with regard to  $x'$  is  $\sqrt{2\pi} \beta$ , and

$$\int_0^{\infty} y'' e^{-\frac{1}{2} \frac{y''^2}{a^2}} dy'' = a^2.$$

Hence : 
$$\frac{S(mx - ny)}{N} = \frac{1}{2\pi} \frac{n\sigma_2}{\sqrt{1-r^2}} \sqrt{2\pi} \beta a^2.$$

Or, for the positive summation

$$\frac{S(mx - ny)}{N} = \frac{n^2\sigma_2^2 (1 - r^2) \sqrt{n^2\sigma_2^2 - 2rmn\sigma_1\sigma_2 + m^2\sigma_1^2}}{\sqrt{2\pi} n^2\sigma_2^2 - 2rm\sigma_1(n\sigma_2 - m\sigma_1) - r^2m^2\sigma_1^2} \dots\dots\dots(iv).$$

This general value does not appear to be likely to be of much service. If we take  $m = n = 1$ , we obtain the result of simply summing the positive differences of paired variates. It is :

$$\frac{S(x - y)}{N} = \frac{\sigma_2^2 (1 - r^2) \sqrt{\sigma_2^2 - 2r\sigma_1\sigma_2 + \sigma_1^2}}{\sqrt{2\pi} \sigma_2^2 - 2r\sigma_1(\sigma_2 - \sigma_1) - r^2\sigma_1^2} \dots\dots\dots(v).$$

(v) leads to an equation of the 5th order to find  $r$  and again does not appear to be likely to be of any service. The variates must be reduced to a common unit before they are handled if we are to make (iv) workable. Such a unit is the standard deviation.

If we write  $m = \frac{1}{\sigma_1}$   $n = \frac{1}{\sigma_2}$ , we have at once :

$$\frac{S\left(\frac{x}{\sigma_1} - \frac{y}{\sigma_2}\right)}{N} = \sqrt{\frac{1-r}{\pi}}.$$

Thus we find :

$$r = 1 - \pi \left( \frac{S\left(\frac{x}{\sigma_1} - \frac{y}{\sigma_2}\right)}{N} \right)^2 \dots\dots\dots(vi).$$



(vi) is an extremely neat formula and might be taken as the definition of a quantity measuring correlation. But the actual determination of correlation in this way, *i.e.* the reduction of each variate to a deviation from its mean measured in terms of its s.d. as unit, would probably be as troublesome as using the product-moment method.

One special case occurs, however, in which the above formula may possibly be of good service. Suppose the two variates have the same mean =  $m$  and the same s.d. =  $\sigma$ , then :

$$r = 1 - \frac{\pi \{S(x-y)\}^2}{N^2\sigma^2} = 1 - \pi \frac{\{S(m+x-m+y)\}^2}{N^2\sigma^2} \dots\dots\dots(vii).$$

Or, the coefficient of correlation is the result of subtracting from unity  $\pi$  times the square of the mean sum of the positive differences of paired variates divided by their common standard deviation.

For cases in which both variates are the same, brothers, cousins of the same sex, homotypes, etc., and especially for some cases of short series, the method may be of value.

*Illustration I. Resemblance of Length of Little Finger in Male Cousins.* I take a short series of 68 male pairs of cousins. The average value of the length measured on the little finger was 51.02 mm. and its standard deviation 2.721 mm. There were 33 positive differences of finger length giving  $S(x-y) = 87.6$  mm. Hence we had :

$$r = 1 - \pi \frac{(87.6/68)^2}{7.4036} = .296.$$

Found by the product-moment method the answer was .287; the difference is well within the probable error of the latter value. The process of taking differences and summing was considerably shorter than finding a product moment.

*Illustration II. Assortative Mating in the case of Paramecium.* I take Dr Pearl's Table AA 3 from Vol. v. p. 295 of *Biometrika* for the lengths of conjugating Paramecia.

I choose this purposely because there was no difficulty above about the male cousins; there were only two equalities, the actual measurements of each individual being recorded. But in an ordinary correlation table owing to the method of grouping there will be a very considerable number of ties, and the problem arises how are they to be distributed. Clearly one half of them will be excesses and one half defects, if we suppose the odds against an actual tie in measuring to any degree of accuracy to be very great. Hence we may say that half the diagonal total is to be treated as in excess. But at what portion of the base unit are we to set the pair apart? If the frequency was uniformly distributed over the diagonal cells, we should take the average interval between a pair to be  $\frac{1}{2}$  the base unit. But the material is almost always clustered inside the cell, and clearly  $\frac{1}{2}$  is too much. The actual value to be taken would depend upon the value of the correlation and the size of the base unit. In fact we



can only take a rough approximation. I suggest that  $\frac{1}{3}$  will be found to work fairly well. Accordingly we take  $\frac{1}{3}$  of the contents of the diagonal cells, multiplied by the base unit. The whole process may now be written as follows :

$\frac{1}{2}$	1	2	3	4	5	6	7	8	9	10	11	12
0	1	1	1	0	1	0	-	-	-	-	-	-
2	1	0	0	0	1	0	-	-	-	-	-	-
4	3	4	1	0	0	1	-	-	-	-	-	-
4	14	7	5	1	1	0	-	-	-	-	-	-
30	25	9	5	1	1	0	-	-	-	-	-	-
22	16	5	5	0	0	0	-	-	-	-	-	-
10	16	7	1	2	0	1	-	-	-	16.3	-	-
16	4	2	0	0	0	2	-	-	-	85	-	-
4	5	3	0	0	4	-	-	-	-	76	-	-
4	0	0	0	4	-	-	-	-	-	54	-	-
2	0	0	18	-	-	-	-	-	-	16	-	-
0	0	38	-	-	-	-	-	-	-	20	-	-
0	85	-	-	-	-	-	-	-	-	12	-	-
98	-	-	-	-	-	-	-	-	-	-	-	-

$$\sigma = 19.112^*, \quad S(x - y) = 279.3 \times 10$$

$$r = 1 - \pi \left( \frac{2793}{40 \times 19.112} \right)^2 = .581.$$

The value obtained by the product-moment method is  $.588 \pm .022$ .

The correlation Table is as follows :

*Length of First Conjugant.*

	160-9	170-9	180-9	190-9	200-9	210-9	220-9	230-9	240-9	250-9	260-9	270-9	280-9	Totals
160-9	—	1	1	1	—	1	—	—	—	—	—	—	—	4
170-9	1	2	1	—	—	—	1	—	—	—	—	—	—	5
180-9	1	1	4	3	4	1	—	—	1	—	—	—	—	15
190-9	1	—	3	4	14	7	5	1	1	—	—	—	—	36
200-9	—	—	4	14	30	25	9	5	1	1	—	—	—	89
210-9	1	—	1	7	25	22	16	5	5	—	—	—	—	82
220-9	—	1	—	5	9	16	10	16	7	1	2	—	1	68
230-9	—	—	—	1	5	5	16	16	4	2	—	—	—	49
240-9	—	—	1	1	1	5	7	4	4	5	3	—	—	31
250-9	—	—	—	—	1	—	1	2	5	4	—	—	—	13
260-9	—	—	—	—	—	—	2	—	3	—	2	—	—	7
270-9	—	—	—	—	—	—	—	—	—	—	—	—	—	0
280-9	—	—	—	—	—	—	1	—	—	—	—	—	—	1
Totals	4	5	15	36	89	82	68	49	31	13	7	0	1	400

\* Pearl, *loc. cit.* p. 226, Table II.



we proceed thus: Read each column down to and including the diagonal cell, and place the total under the corresponding differences in the previous scheme. For example, take the sixth column; 1, 0, 1, 7, 25, 22, are the corresponding frequencies, and these numbers will be found, sloping from the column marked 5, *i.e.* difference  $5 \times 10$ , diagonally across the scheme. In this manner the columns of the table can be disposed in the scheme at once. The scheme columns are then added up and multiplied by the difference at the top, and, if multiplied again by the base unit, in this case 10, the total gives  $S(x-y)$ . The whole can be done with very great rapidity, and the correlation found in about 10 minutes if  $\sigma$  be known.

As other comparisons I give the homotypic results:

	Difference method	Product method
Monmouthshire Ashes (65,000)	.432	.405 $\pm$ .011
<i>Papaver Rhoeas</i> (Quantocks) (19,790)	.523	.533 $\pm$ .013
Ditto (Chilterns' Base) (25,160)	.395	.400 $\pm$ .012

These results show that there exists quite a reasonable amount of agreement between the two methods, and the difference method is much the shorter when the table contains thousands of observations as in these cases. At the same time too much reliance must not be placed upon the difference method, not only because it assumes normality of distribution but because it involves a somewhat rough method of approximation in the case of the diagonal cell.

One further point may be noted. Suppose that rank in a series was a true character which could be dealt with by a difference formula like the above then  $r$  the correlation of the ranks would be given by

$$r = 1 - \pi \left\{ \frac{S(x-y)}{N\sigma} \right\}^2.$$

Now for such ranks  $\sigma^2 = \frac{1}{12}(N^2 - 1)$ , therefore

$$r = 1 - \frac{12\pi \{S(x-y)\}^2}{N^2(N^2 - 1)} \dots\dots\dots(\text{viii}).$$

Dr Spearman has introduced a quantity  $R$  which he terms a "correlational coefficient\*" and which he defines without any special justification by:

$$R = 1 - \frac{S(x-y)}{\frac{1}{6}(N^2 - 1)} \dots\dots\dots(\text{ix}).$$

We should thus have:

$$1 - r = \frac{\pi(N^2 - 1)}{3N^2} (1 - R)^2 \dots\dots\dots(\text{x}),$$

which would give approximately:  $r = 2R - R^2$ .

\* *Journal of Psychology*, Vol. II. p. 96.



This is, of course, not true, for the distribution of ranks is not normal; the exact formula will be given later; but it suffices to indicate that the actual distribution assumed for  $x$  and  $y$  will much influence the relation between  $r$  and  $R$ . Dr Spearman from trial gives the empirical formula

$$r = \sin \left( \frac{\pi}{2} R \right) \dots\dots\dots(x_i),$$

which is also incorrect. But the above relation shows that we are not *à priori* compelled to suppose that  $r$  merely changes its sign, not its numerical value when  $R$  changes sign.

(3) *On the Correlation of Grades.* A method of representing frequency has been introduced by Francis Galton in which the extent of variation of a character is expressed by the position of the individual bearing this character in the population. This method was originally spoken of as that of percentiles but more recently as that of grades. A fundamental feature of the method is that the grade is looked upon as an index to the variate, it is not considered as in itself significant, or treated as an independent character of the individual. In order, however, to pass from the grade to the variate it is absolutely necessary to make some hypothesis as to the nature of the distribution. The hypothesis hitherto made is that the frequency follows, at least fairly closely, the normal or Gaussian law. On this assumption, tables of the probability integral enable us to pass at once from the grade to the magnitude of the variate, and *vice versa*. Quite recently, however, Dr Spearman has proposed that rank in a population for any variate should be considered as in itself the quantitative measure of the character, and he proceeds to correlate ranks as if they were quantitative measures of character, without any reference to the true value of the variate. This seems to me a retrograde step; hitherto we have dealt with grade or rank (I will distinguish between them presently) as an index to the variate, and to make rank into a unit itself cannot fail, I believe, to lead to grave misconception. Between mediocrities the unit of rank treated as a measure of a variate is practically zero, between extreme individuals, it is very large indeed. To state that two individuals differ by  $m$  ranks carries no meaning at all unless we add, (i) the size of the population dealt with, (ii) the position in the population of one or both individuals, and (iii) the nature of the frequency distribution which governs the population. I cannot therefore look upon the correlation of ranks as conveying any real idea of the correlation of variates, unless we have a means of passing from the correlation of ranks to the value of the correlation of the variates, i.e. the correlation of ranks can only be treated as a step subsidiary to determining the true variate correlation.

The correlation between variates can be made to change widely by preserving the same system of ranks, but by altering the nature of the frequency distribution. Thus consider the system:



Variates	x	-2	-1	+1	+2
	y	-2	-1	+1	+2

1	2	3	4
1	2	3	4

The correlation of variates is perfect and the correlation of ranks is also perfect. But we may also have :

Variates	x	-2	-1.9	+1.9	+2
	y	-2	-.01	+.01	+2

1	2	3	4
1	2	3	4

The correlation of variates is now .72, but the correlation of ranks remains perfect and would indicate nothing of this great difference. I think that it is safe to assert that until some assumption is made, at least as to the approximate nature of the distribution, we cannot hope to avoid misconceptions if we use the method of ranks without reference to the rank as index of the variate.

In such a case there can hardly be a doubt that the best method is first to consider to what results normal distribution will lead us, and secondly if the formulae found turn out to be of a simple character to adopt these as the basis by definition of the variate correlation constant as found from a method of ranks. This will be the course adopted in the present memoir.

(4) Let there be a population of  $N$  members and let these be under investigation for two correlated characters, means  $m_1, m_2$ , standard deviations  $\sigma_1, \sigma_2$ , correlation  $r$ . I shall suppose normality of distribution. Let  $m_1 + x, m_2 + y$  be the deviations of the two characters in any individual. Then I term :

$$\left. \begin{aligned} g_1 &= \frac{1}{2}N + \frac{N}{\sqrt{2\pi}\sigma_1} \int_0^x e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} dx \\ g_2 &= \frac{1}{2}N + \frac{N}{\sqrt{2\pi}\sigma_2} \int_0^y e^{-\frac{1}{2}\frac{y^2}{\sigma_2^2}} dy \end{aligned} \right\} \dots\dots\dots(xii),$$

the  $x$ - and  $y$ -grades of the variates for the individual. It will be obvious that  $g_1$  and  $g_2$  are mathematical functions of the variates and that accordingly the correlation between them determines that between  $x$  and  $y$ , or *vice versa*.

Obviously  $g_1$  and  $g_2$  can be found from tables of the probability integral as soon as  $x$  and  $y$ , the deviates, are known.

I term *rank* the actual position in order of an individual with regard to any variate in a given series obtained by measurement or observation. If  $v_1$  be the 'rank' of an individual for a given character this signifies that in the observed



population there are  $\nu_1 - \frac{1}{2}$  individuals with character greater than  $x$ . If therefore we were to identify this with the grade we should have

$$g_1 = \nu_1 - \frac{1}{2} \dots\dots\dots(xiii),$$

or  $g_1$  would always differ from a whole number by .5. This, of course, it does not, and the whole problem of working with ranks really centres on the degree of approximation which is made when we proceed from ranks to grades by the relation (xiii). A grade determined from a rank and not from a variate we may term a spurious grade; actually the real grade often differs by several units from the spurious grade, and the practical problem is: To what extent does this vitiate the use of ranks as a subsidiary stage to the determination of variate-correlation?

I shall first proceed to find the mean and standard deviation of a true grade; (xii) shows us at once that  $\bar{g}_1 = \bar{g}_2 = \frac{1}{2}N$  is the mean value of the grade.

The frequency of a given variate lying between  $x$  and  $x + \delta x$

$$= \frac{N}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} dx = dg_1.$$

But the frequency of the variate must also be the frequency of its grade, or:

$$\begin{aligned} N\sigma_{g_1}^2 &= \int_{-\infty}^{+\infty} (g_1 - \bar{g}_1)^2 dg_1 = \frac{1}{3} \left[ (g_1 - \bar{g}_1)^3 \right]_0^N \\ &= \frac{2}{3} \frac{N^3}{8} = \frac{N^2}{12} \times N. \end{aligned}$$

Hence we have:  $\sigma_{g_1}^2 = \sigma_{g_2}^2 = \frac{1}{12}N^2 \dots\dots\dots(xiv).$

Now whereas our grades are a continuous series, the spurious grades or ranks are discontinuous and at intervals  $h = 1$ . (xiii) shows us at once that

$$\bar{\nu}_1 = \bar{\nu}_2 = \bar{g}_1 + \frac{1}{2} = \frac{1}{2}(N + 1).$$

Further

$$\sigma_{g_1}^2 = \sigma_{\nu_1}^2 + \frac{1}{12}h^2,$$

the latter corresponding to the Sheppard's correction by which we pass from raw to adjusted moments.

Thus we have:  $\left. \begin{aligned} \sigma_{\nu_1}^2 = \sigma_{\nu_2}^2 &= \frac{1}{12}(N^2 - 1) \\ \bar{\nu}_1 = \bar{\nu}_2 &= \frac{1}{2}(N + 1) \end{aligned} \right\} \dots\dots\dots(xv),$

(xv) must be used whenever we are dealing with ranks or spurious grades.

Writing  $i_1 = g_1 - \bar{g}_1$ , and  $i_2 = g_2 - \bar{g}_2$ , I now turn to the determination of the product moment of the grades. Let us put:

$$z = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)} \dots\dots\dots(xvi),$$

then:

$$p_{g_1, g_2} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} i_1 i_2 z dx dy$$

gives the product moment of the grades.



Differentiate  $p_{g_1, g_2}$  with regard to  $r$  which is not contained in either  $i_1$  or  $i_2$ ; we have :

$$\frac{dp_{g_1, g_2}}{dr} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} i_1 i_2 \frac{dz}{dr} dx dy.$$

But I have elsewhere\* shown that :

$$\frac{dz}{dr} = \sigma_1 \sigma_2 \frac{d^2 z}{dx dy} \dots\dots\dots(xvii).$$

Accordingly : 
$$\frac{dp_{g_1, g_2}}{dr} = \sigma_1 \sigma_2 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx dy.$$

Integrating twice by parts and noting that the part between limits vanishes in both cases, we have :

$$\frac{dp_{g_1, g_2}}{dr} = \sigma_1 \sigma_2 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} z \frac{di_1}{dx} \frac{di_2}{dy} dx dy.$$

Substituting for  $di_1/dx$  and  $di_2/dy$  and writing  $x = x'\sigma_1$ ,  $y = y'\sigma_2$ , we find :

$$\begin{aligned} \frac{dp_{g_1, g_2}}{dr} &= \frac{N^3}{4\pi^2 \sqrt{1-r^2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{1}{2(1-r^2)} \{ (2-r^2)x'^2 - 2r x'y' + (2-r^2)y'^2 \}} dx' dy' \\ &= \frac{N^3}{2\pi \sqrt{1-r^2}} \frac{1}{\sqrt{\left(\frac{2-r^2}{1-r^2}\right)^2 - \frac{r^2}{(1-r^2)^2}}} = \frac{N^3}{2\pi \sqrt{4-r^2}}. \end{aligned}$$

Now if  $\rho_{12}$  be the correlation of grades, we have :

$$\rho_{12} = \frac{p_{g_1, g_2}}{N \sigma_{g_1} \sigma_{g_2}}, \text{ and } \frac{d\rho_{12}}{dr} = \frac{1}{N \sigma_{g_1} \sigma_{g_2}} \frac{dp_{g_1, g_2}}{dr}.$$

Thus remembering (xiv)

$$\frac{d\rho_{12}}{dr} = \frac{6}{\pi} \frac{1}{\sqrt{4-r^2}},$$

or, 
$$\rho_{12} = \frac{6}{\pi} \sin^{-1} \frac{1}{2} r + \text{constant.}$$

Now  $\rho_{12}$  and  $r$  must vanish together, hence the constant is zero. Accordingly we have :

$$r = 2 \sin \left( \frac{\pi}{6} \rho_{12} \right) \dots\dots\dots(xviii).$$

This remarkably simple formula enables us to determine the value of the true variate correlation from a correlation of grades on the assumption of the normal law ; or if grades may be replaced by ranks, a knowledge of the correlation of ranks will give us the correlation of the actual variates behind the order exhibited in the ranking. The important idea embodied in the above formula is the basis of the present memoir, and is as far as I am aware wholly new.

\* *Phil. Trans. A.* Vol. 195, p. 25.



It remains for us to consider methods of finding the rank or grade correlation and the probable error of such methods.

(5) A convenient method of finding the grade correlation is that of formula (i), p. 4, we have at once :

$$\rho_{12} = \frac{\sigma_1^2 + \sigma_2^2 - S(g_1 - g_2)^2/N}{2\sigma_1\sigma_2}$$

Or, 
$$\rho_{12} = 1 - \frac{6S(g_1 - g_2)^2}{N^2} \dots\dots\dots(xix),$$

if we use true grades,

but: 
$$= 1 - \frac{6S(v_1 - v_2)^2}{N(N^2 - 1)} \dots\dots\dots(xx),$$

if we use ranks  $v_1$  and  $v_2$ .

If we use ranks the discovery of  $S(v_1 - v_2)^2$  or the sum of the squares of the differences of ranks forms a very easy process of determining  $\rho_{12}$ , due regard being paid to certain points to be dealt with in the illustrations below. Then (xviii) will give the variate correlation.

The probable error of  $\rho_{12}$  and of  $r$  found in this way will be given in another section.

Since the determination of  $\rho_{12}$  by (xx) is algebraically identical with finding  $\rho_{12}$  by the product moment, and such product moment gives the least probable error in the determination of a correlation coefficient, there must be some fallacy in a statement which has been propounded among the psychologists that a difference method of determining the correlation will give  $\rho_{12}$  with about  $\frac{2}{3}$  of the probable error of the product moment method. This fallacy will be considered later.

Meanwhile it is of interest to show that the probable error\* of

$$\rho_{12} = \{S(v_1 v_2)/n - \bar{v}_1 \bar{v}_2\} / (\sigma_{v_1} \sigma_{v_2})$$

is of the form :

$$P.E. = \frac{.67449}{\sqrt{n-1}} (1 - c_2 \rho_{12}^2 + c_4 \rho_{12}^4 + c_6 \rho_{12}^6 + \dots)$$

where  $c_2, c_4, c_6, \dots$  are undetermined constants. Or, the probable error of  $\rho_{12}$  for  $\rho_{12} = 0$ , or for uncorrelated ranks is

$$.67449/\sqrt{n-1},$$

i.e. is absolutely identical with probable error of a coefficient of correlation of any two uncorrelated variables, and is not as asserted much smaller.

Since for ranks  $\sigma_{v_1}$  and  $\sigma_{v_2}$  are constant, we have † to find the value of

$$u^2 = \Sigma \left\{ \frac{S(v_1 v_2)}{n} - \left\{ \frac{1}{2}(n+1) \right\}^2 \right\}^2,$$

$v_1$  and  $v_2$  being independent, in order to reach the squared standard deviation of  $\rho_{12}$  for  $\rho_{12} = 0$ .

\*  $n$  is here put for  $N$  as more convenient for the algebraic work which follows

† I owe the following proof to the kindness of my friend "Student."



$$\text{Now: } u^2 = \Sigma \left\{ \left( \frac{S(\nu_1 \nu_2)}{n} \right)^2 - 2 \left( \frac{n+1}{2} \right)^2 \frac{S(\nu_1 \nu_2)}{n} + \left( \frac{n+1}{2} \right)^4 \right\}.$$

There being no correlation,  $n!$  arrangements of this product occur with equal frequency. Hence

$$\Sigma \left( \frac{n+1}{2} \right)^4 = \left( \frac{n+1}{2} \right)^4 n!.$$

Next any  $\nu_1 \nu_2$  occurs in  $(n-1)!$  of the arrangements, for if  $\nu_1$  be paired with  $\nu_2$ , the remaining  $n-1$  pairs may be arranged in  $(n-1)!$  ways. Thus

$$\begin{aligned} & \Sigma \left\{ 2 \left( \frac{n+1}{2} \right)^2 \frac{S(\nu_1 \nu_2)}{n} \right\} = 2(n-1)! \left( \frac{n+1}{2} \right)^2 \frac{\Sigma(\nu_1 \nu_2)}{n} \\ & = 2(n-1)! \left( \frac{n+1}{2} \right)^2 \frac{\Sigma(\nu_1) \Sigma(\nu_2)}{n} = 2(n-1)! \left( \frac{n+1}{2} \right)^2 \left\{ \frac{n(n+1)}{2} \right\}^2 / n \\ & = 2(n)! \left( \frac{n+1}{2} \right)^4. \end{aligned}$$

$$\text{Further: } \Sigma \left( \frac{S(\nu_1 \nu_2)}{n} \right)^2 = \Sigma \frac{1}{n^2} \{ S(\nu_1^2 \nu_2^2) + 2S(\nu_1 \nu_2 \nu_1' \nu_2') \},$$

where  $\nu_1', \nu_2'$  are different from  $\nu_1, \nu_2$ .

Now  $\nu_1^2 \nu_2^2$  occurs in  $(n-1)!$  arrangements; hence

$$\begin{aligned} \Sigma \left( \frac{S(\nu_1^2 \nu_2^2)}{n^2} \right) &= \frac{(n-1)!}{n^2} \Sigma(\nu_1^2) \Sigma(\nu_2^2) = \frac{(n-1)!}{n^2} \left( \frac{n(n+1)}{6} (2n+1) \right)^2 \\ &= (n-1)! \left\{ \frac{(n+1)(2n+1)}{6} \right\}^2. \end{aligned}$$

Next  $\nu_1 \nu_2 \nu_1' \nu_2'$  occurs in  $(n-2)!$  arrangements.

$$\begin{aligned} \text{Thus: } \Sigma \left( \frac{2S(\nu_1 \nu_2 \nu_1' \nu_2')}{n^2} \right) &= \frac{2(n-2)!}{n^2} \Sigma(\nu_1 \nu_2 \nu_1' \nu_2') \\ &= \frac{(n-2)!}{n^2} \Sigma(\nu_1 \nu_2) \{ \Sigma(\nu_1' \nu_2') - \nu_1 \Sigma(\nu_2') - \nu_2 \Sigma(\nu_1') + \nu_1 \nu_2 \}, \end{aligned}$$

where  $\nu_1'$  and  $\nu_2'$  may now take all values.

Thus:

$$\begin{aligned} \Sigma \left\{ \frac{2S(\nu_1 \nu_2 \nu_1' \nu_2')}{n^2} \right\} &= \frac{(n-2)!}{n^2} \left\{ \Sigma(\nu_1 \nu_2) \Sigma(\nu_1' \nu_2') - \frac{n(n+1)}{2} S(\nu_1^2 \nu_2 + \nu_2^2 \nu_1) + S(\nu_1^2 \nu_2^2) \right\} \\ &= \frac{(n-2)!}{n^2} \left\{ \left( \frac{n(n+1)}{2} \right)^2 - 2 \left( \frac{n(n+1)}{2} \right)^2 \frac{n(n+1)(2n+1)}{6} + \left( \frac{n(n+1)(2n+1)}{6} \right)^2 \right\} \\ &= \frac{(n-1)!}{144} (n+1)^2 \{ 9n^3 + 3n^2 - 8n - 4 \}. \end{aligned}$$



Collecting the various parts we find :

$$u^2 = \frac{(n-1)! (n+1)^2 (2n+1)^2}{36} + \frac{(n-1)! (n+1)^2 (9n^2 + 3n^2 - 8n - 4)}{144} - 2n! \left(\frac{n+1}{2}\right)^4 + n! \left(\frac{n+1}{2}\right)^4,$$

or, after reducing, 
$$u^2 = \frac{n! (n+1)^2 (n-1)}{144}.$$

Therefore the mean value of  $u^2$  is  $\frac{(n+1)^2 (n-1)}{144}.$

Now the probable error of  $\rho_{12}$ , for uncorrelated ranks :

$$\begin{aligned} &= .67449u/\sigma_{v_1}\sigma_{v_2} \\ &= .67449 \frac{(n+1)\sqrt{n-1}}{12} / \frac{n^2-1}{12} \\ &= .67449/\sqrt{n-1} \dots\dots\dots(xxi). \end{aligned}$$

It thus follows that if the value of  $\rho_{12}$  be not two or three times the expression (xxi), there is no significant correlation of ranks, and therefore no significant correlation of the corresponding variates.

(6) *On the Difference Method of finding the Correlation of Grades.*

Exactly as in the first section of this paper we may seek the correlation of grades by means of the sum  $S(g_1 - g_2)$  of all their positive differences. This is slightly shorter than finding  $S(g_1 - g_2)^2$ , but only very slightly so, and it may be doubted whether the increased rapidity of working at all compensates for the decreased accuracy of the process. Still the result is interesting and throws considerable light on one or two allied points.

Let  $G = S(g_1 - g_2)$ , where the sum  $S$  is for all  $x$ -grades which are greater than corresponding  $y$ -grades.

Let us put  $x = \sigma_1 x'$ ,  $y = \sigma_2 y'$ , and write

$$\begin{aligned} j_x &= \int_0^x e^{-\frac{1}{2}v^2} dv, \\ z &= \frac{1}{2\pi} \frac{1}{\sqrt{1-r^2}} \frac{1}{\sigma_1 \sigma_2} z'. \end{aligned}$$

Then :

$$\begin{aligned} G &= \frac{N}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \int_{-\infty}^{x'} (j_x - j_y) z dy dx \\ &= \frac{N^2}{(2\pi)^{\frac{3}{2}} \sqrt{1-r^2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{x'} (j_x - j_y) z' dy' dx'. \end{aligned}$$



$$\begin{aligned}
\frac{dG}{dr} &= \frac{N^2}{(2\pi)^{\frac{3}{2}}} \int_{-\infty}^{+\infty} \int_{-\infty}^{x'} (j_x - j_y) \frac{d}{dr} \left( \frac{z'}{\sqrt{1-r^2}} \right) dy' dx' \\
&= \frac{N^2}{(2\pi)^{\frac{3}{2}}} \int_{-\infty}^{+\infty} \int_{-\infty}^{x'} (j_x - j_y) \frac{d}{dx' dy'} \left( \frac{z'}{\sqrt{1-r^2}} \right) dy' dx' \\
&= \frac{N^2}{(2\pi)^{\frac{3}{2}} \sqrt{1-r^2}} \int_{-\infty}^{+\infty} \left\{ \left[ (j_x - j_y) \frac{dz'}{dx'} \right]_{-\infty}^{x'} + \int_{-\infty}^{x'} \frac{dj_y}{dy'} \frac{dz'}{dx'} dy' \right\} dx' \\
&= \frac{N^2}{(2\pi)^{\frac{3}{2}} \sqrt{1-r^2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{x'} e^{-\frac{1}{2}y'^2} \left( -\frac{(x' - ry')}{1-r^2} z' \right) dy' dx'.
\end{aligned}$$

Put  $y' = x' - y''$ ; then after rearranging :

$$\begin{aligned}
\frac{dG}{dr} &= -\frac{N^2}{(2\pi)^{\frac{3}{2}} (1-r^2)^{\frac{3}{2}}} \int_{-\infty}^{+\infty} \int_0^{\infty} \left\{ (1-r) \left( x' - \frac{2+r}{3+r} y'' \right) + \frac{2(1+r)}{3+r} y'' \right\} \\
&\quad \times e^{-\frac{1}{2} \left( \frac{3+r}{1+r} \left( x' - \frac{2+r}{3+r} y'' \right)^2 + \frac{2}{(1-r)(3+r)} y''^2 \right)} dy'' dx'.
\end{aligned}$$

The order of integration can now be transposed and if  $X$  be written for

$$x' - \frac{2+r}{3+r} y'',$$

the limits of  $X$  will also be  $-\infty$  to  $+\infty$ . Thus :

$$\frac{dG}{dr} = -\frac{N^2}{(2\pi)^{\frac{3}{2}} (1-r^2)^{\frac{3}{2}}} \int_0^{\infty} e^{-\frac{1}{2} \frac{2}{(1-r)(3+r)} y''^2} \int_{-\infty}^{+\infty} \left\{ (1-r) X + \frac{2(1+r)}{3+r} y'' \right\} e^{-\frac{1}{2} \frac{3+r}{1+r} X^2} dX dy''.$$

But if  $c$  have any value :

$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2} c^2 X^2} X dX = 0, \quad \text{and} \quad \int_{-\infty}^{+\infty} e^{-\frac{1}{2} c^2 X^2} dX = \sqrt{2\pi} \frac{1}{c}.$$

Hence :

$$\begin{aligned}
\frac{dG}{dr} &= -\frac{N^2}{(1-r^2)^{\frac{3}{2}}} \frac{1}{(2\pi)} \frac{2(1+r)}{3+r} \frac{\sqrt{1+r}}{\sqrt{3+r}} \int_0^{\infty} y'' e^{-\frac{1}{2} \frac{2}{(1-r)(3+r)} y''^2} dy'' \\
&= -\frac{N^2}{2\pi} \frac{1}{(1-r^2)^{\frac{3}{2}}} \frac{2(1+r)^{\frac{3}{2}} (1-r)(3+r)}{(3+r)^{\frac{3}{2}} \cdot 2} \\
&= -\frac{N^2}{2\pi} \frac{1}{\sqrt{(1-r)(3+r)}} = -\frac{N^2}{2\pi} \frac{1}{\sqrt{4-(1+r)^2}}.
\end{aligned}$$

Hence integrating :

$$G = \text{constant} + \frac{N^2}{2\pi} \cos^{-1} \frac{1+r}{2}.$$

But when  $r=1$ ,  $G$  must be zero ; therefore the constant is zero, or inverting :

$$r = 2 \cos 2\pi \frac{G}{N^2} - 1.$$



Or, finally\* : 
$$r = 2 \cos 2\pi \left( \frac{S(g_1 - g_2)}{N^2} \right) - 1 \dots\dots\dots(\text{xxii}).$$

This gives us the correlation of two variates from the corresponding grades by a difference method.

If  $r$  be zero, we must have  $2\pi \frac{S(g_1 - g_2)}{N^2}$  equal to  $60^\circ = \pi/3$ , or  $S(g_1 - g_2) = \frac{1}{6}N^2$  when there is no correlation of variates. This is easily proved directed, for in this case :

$$\begin{aligned} G = S(g_x - g_y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{x'} (g_x - g_y) \frac{N}{2\pi} e^{-\frac{1}{2}(x^2 + y^2)} dy dx' \\ &= \frac{1}{N} \int_0^N \int_0^{g_x} (g_x - g_y) dg_y dg_x \\ &= \frac{1}{N} \int_0^N \frac{g_x^2}{2} dg_x = \frac{N^2}{6}. \end{aligned}$$

For ranks the corresponding expression to be used is  $\frac{1}{6}(N^2 - 1)$ , or we have :

$$r = 2 \cos 2\pi \left( \frac{S(v_1 - v_2)}{N^2 - 1} \right) - 1 \dots\dots\dots(\text{xxiii}).$$

As before the truth of (xxii) depends on the approximation to normal correlation.

If we combine (xx), (xviii) and (xxiii) we have the relation between  $S(v_1 - v_2)^2$  and  $S(v_1 - v_2)$  which holds in the case of normal correlation.

Writing  $R = 1 - S(v_1 - v_2) / \frac{1}{6}(N^2 - 1)$ , we have :

$$r = 2 \sin \frac{\pi}{6} \rho_{12} = 2 \cos \frac{\pi}{3} (1 - R) - 1 \dots\dots\dots(\text{xxiv}).$$

Dr Spearman gives† the relation :

$$\rho_{12} = \sin \left( \frac{\pi}{2} R \right) \dots\dots\dots(\text{xxv})$$

(he neither connects  $\rho_{12}$ , nor  $R$ , with  $r$ ) as apparently an empirical relationship and speaks of it as "all that could be desired." It is clearly incompatible with normal

\* The relationship of (xxii) to (vii) is easily seen if we expand the cosine as far as the square of the angle. We have

$$r = 1 - \frac{4\pi^2}{N^4} \{S(g_1 - g_2)\}^2 = 1 - \frac{\pi}{3} \frac{\pi \{S(g_1 - g_2)\}^2}{N^2 \sigma^2} = 1 - 1.0472 \frac{\pi \{S(g_1 - g_2)\}^2}{N^2 \sigma^2}.$$

(vii) would have given us 1 instead of the factor 1.0472. Thus when there is high correlation, or  $S(g_1 - g_2)$  is small, we see that the difference method with grades leads us to nearly the same result, as the assumption that the grades themselves form a normal distribution. This suggests that Spearman would have got much better results for his "footrule" for measuring correlation had he taken  $R = 1 - 3 \frac{N^2}{N^2 - 1} \left( \frac{S(v_1 - v_2)}{N\sigma} \right)^2$  instead of  $1 - \frac{S(v_1 - v_2)}{2\sigma^2}$ ; for this value, i.e.  $1 - \left( \frac{S(g)}{M} \right)^2$  in his notation, would have been almost the true variate correlation  $r$ .

† *Journal of Psychology*, Vol. II. p. 102.



correlation, which at any rate is a fairly good guide for general relations of this sort in the theory of frequency. Table I. gives the values of  $r$  and  $R$  for each .05 for  $\rho_{12}$ . Table II. gives the values of  $r$  and  $\rho_{12}$  for each .05 of  $R$ , and in the last column the value of  $\rho_{12}$  which would arise if (xxv) were correct.

TABLE I. *Correlation of Variates from Mean Square Difference of Grades.*

$\rho_{12}$	$r$	$R$	$\rho_{12}$	$r$	$R$
.00	.000	.000	.50	.518	.323
.05	.052	.029	.55	.568	.361
.10	.105	.059	.60	.618	.400
.15	.157	.089	.65	.668	.442
.20	.209	.120	.70	.717	.486
.25	.261	.152	.75	.765	.533
.30	.313	.184	.80	.813	.584
.35	.364	.217	.85	.861	.644
.40	.416	.251	.90	.908	.709
.45	.467	.286	.95	.954	.796
.50	.518	.323	1.00	1.000	1.000

TABLE II. *Correlation of Variates from Difference of Grades.*

$R$	$r$	$\rho_{12}$	(xxv)	$R$	$r$	$\rho_{12}$	(xxv)
.00	.000	.000	.000	.50	.732	.716	.707
.05	.089	.085	.078	.55	.782	.767	.760
.10	.176	.168	.156	.60	.827	.814	.809
.15	.259	.248	.233	.65	.867	.856	.853
.20	.338	.324	.309	.70	.902	.894	.891
.25	.414	.398	.383	.75	.932	.926	.924
.30	.486	.469	.454	.80	.956	.952	.951
.35	.554	.536	.522	.85	.975	.973	.972
.40	.618	.600	.587	.90	.989	.988	.988
.45	.677	.660	.649	.95	.997	.997	.997
.50	.732	.716	.707	1.00	1.000	1.000	1.000

Now these Tables bring out several interesting facts. The first is the remarkable closeness between the correlation of the grades and the true correlation of the variates, if we suppose the system normal. The maximum difference as shown by Table I. is .018 and actually the maximum of  $r - \rho_{12}$  occurs when  $\rho_{12} = .5756$  and is then .0180. Thus, the difference will often be of the order of the probable error. The formula (xviii) is so simple, that we can always deduce the variate correlation at once from the grade correlation. I propose to define  $r$  as given by (xviii) as the *grade-variate correlation*. Whenever the system is normal, or approximately normal, this will agree with the true variate correlation closely. Next Table I. shows us that



equal differences of  $\rho_{12}$  give almost equal differences of  $r$ , i.e. the differences only range from .052 to .046 of  $r$  for differences of .050 of  $\rho_{12}$ . On the other hand the differences of  $r$  for equal differences .050 of  $R$  vary from .089 to .003, or second differences become of importance. Clearly for high values of  $R$ ,  $r$  will be found much more closely than for low values.

If  $E_r'$  be the error in  $r$  due to an error  $E_{\rho_{12}}$  in  $\rho_{12}$ , and  $E_r''$  be the error due to an error  $E_R$  in  $R$ , we have:

$$E_r' = \frac{\pi}{3} \cos \frac{\pi}{6} \rho_{12} \times E_{\rho_{12}},$$

$$E_r'' = \frac{2\pi}{3} \sin \frac{\pi}{3} (1 - R) \times E_R$$

if we use differentials. For the special case of  $\rho_{12} = R = 0$ , we have seen that the probable error of  $\rho_{12} = .67449/\sqrt{n-1}$ ; it will be seen later that the probable error of  $R$  is  $.4266/\sqrt{n-1}$  nearly, and if  $E_r$  be the probable error of  $r=0$ , as found in the ordinary product moment way, we have:

$$E_r : E_r' : E_r'' :: \frac{.6745}{\sqrt{n-1}} : \frac{\pi}{3} \frac{.6745}{\sqrt{n-1}} : \frac{2\pi}{3} \frac{\sqrt{3}}{2} \frac{.4266}{\sqrt{n-1}}$$

$$:: \frac{.6745}{\sqrt{n-1}} : \frac{.7063}{\sqrt{n-1}} : \frac{.7738}{\sqrt{n-1}} \dots\dots\dots(\text{xxvi}).$$

Thus we see that, contrary to what has been asserted, the accuracy of the new methods—when they are measured by the determination of the true correlation—are less than the old product moment method. In particular it requires about 30 per cent. more observations by the  $R$  method to obtain  $r$  with the same degree of certainty, when  $r=0$ .

At present we do not know the  $R$  factor term in  $E_R$ , when  $R$  differs from zero, and accordingly cannot test  $E_r$ ,  $E_r'$  and  $E_r''$  at other values of  $R$  or  $\rho_{12}$ , but I have little doubt of the general truth of the result that  $E_r$  is at all values as well as for  $r=0$ , sensibly less than  $E_r'$  and still less than  $E_r''$ .

(7) *Remarks on the Probable Error of R.*

The probable error of a quantity in which the limits of the summation vary as we make random variations in the constants is always a troublesome matter, and I have not yet succeeded in evaluating the probable error of  $S(g_1 - g_2)$  when  $g_1 > g_2$  for any value of  $r$ .

Spearman has investigated the probable error of the corresponding expression for ranks,  $S(v_1 - v_2)$ , when there is no correlation between the ranks. He finds that for  $n$  observations the probable error of  $R$  may be taken as  $.43/\sqrt{n}$ , and from this result he has drawn rather sweeping conclusions as that: "twenty cases treated in one of the ways described furnish as much certitude as 180 in another more usual way"; or that: "a probable error may at present be admitted without much hesitation up to



0.05; so that by adopting the method of calculation recommended, two to three dozen subjects would be sufficient for most purposes\*." Now these statements seem to me not without grave danger, and accordingly it is well to see where the error has crept in.

Spearman gives the value  $.4266/\sqrt{n}$ , but it should be  $.4266/\sqrt{n-1}$ †, and accordingly since we have seen that the probable error of  $\rho_{12}$  for  $\rho_{12}=0$ , is  $.6745/\sqrt{n-1}$ , the probable error of  $R$  would only be about  $\frac{2}{3}$  of the probable error of  $\rho_{12}$ , and upon this Spearman's statements are based.

Now the probable error of any quantity is conventionally  $.67449 \times$  standard deviation  $/\sqrt{n-1}$ , and accordingly for the same number of observations the probable error is less when the standard deviation is less. But there would be no meaning in asserting that the mean of 20 metacarpal bones could be found with much more exactitude than the mean of 20 humeri, because the latter being a larger bone had a greater variability. We must either measure the same quantity by different processes, or else be at any rate certain that our quantities are alike in character and function before we compare their probable errors. The probable error of  $\frac{1}{2}x$  is certainly less than that of  $x$ . Now  $\rho_{12}$  is a true correlation and ranges from +1 through 0 to -1 with a symmetrical distribution about 0, if we take the case of a random distribution of ranks. The quantity  $R$  presents nothing of this nature at all; random distribution of ranks does not give a symmetrical distribution for  $R$ , its range is not from +1 to -1, and there are certain values it can never take. In order to bring out these points I take the following table for  $R$  negative.

TABLE III. *Negative Correlation of Variates from Difference of Grades.*

$R$	$r$	$\rho_{12}$
- .05	- .092	- .088
- .10	- .187	- .178
- .15	- .283	- .271
- .20	- .382	- .367
- .25	- .482	- .465
- .30	- .584	- .566
- .35	- .687	- .670
- .40	- .791	- .777
- .45	- .895	- .886
- .50	- 1.000	- 1.000

N.B. It will be observed that when  $R$  is negative, the true variate correlation is almost double the magnitude of  $R$ , while if  $R$  be positive (Table II.)  $r$  is larger than  $R$  but not to this exaggerated extent. It will be clear that no estimate of the real correlation can be based on  $R$ , if it does not allow for this exaggeration.

\* *American Journal of Psychology*, Vol. xv. pp. 100, 101. For the proof of the probable error cited see: *British Journal of Psychology*, Vol. II. pp. 105-8.

† Spearman's result at bottom of p. 108 may be written  $\frac{.4266}{\sqrt{n-1}} \sqrt{\frac{n^2-3.5}{n^2-1}}$ , or neglecting terms in  $\frac{1}{n^2}$  not  $\frac{1}{n}$  as he does, this gives  $.4266/\sqrt{n-1}$  as we should anticipate.



Thus we see that while  $r$  and  $\rho_{12}$  run from  $-0.0$  to  $-1.0$ ,  $R$  only runs from  $-0.0$  to  $-0.50$ .

In order to obtain his probable error for  $R$  Spearman takes every random arrangement of ranks  $v_1$  and  $v_2$  for which  $v_1$  is greater than  $v_2$ . He has neglected to observe that when he does this his  $R$  will become negative, but that it will not range from  $-1.0$  to  $+1.0$ . For example, I take the following system of ranks for  $(2m + 1)$  individuals:

$v_1 =$	1	2	3	4	.	.	$2m + 1$
$v_2 =$	$2m + 1$	$2m$	$2m - 1$	$2m - 2$	.	.	1

This gives: 
$$S(v_1 - v_2)^2 = 2 \{ (2m)^2 + (2m - 2)^2 + (2m - 4)^2 + \dots + 2^2 \}$$

$$= 8m(m + 1)(2m + 1)/6,$$

$$\therefore \rho_{12} = 1 - \frac{6S(v_1 - v_2)^2}{N(N^2 - 1)} = 1 - \frac{8m(m + 1)(2m + 1)}{(2m + 1)((2m + 1)^2 - 1)} = -1.$$

But 
$$S(v_1 - v_2) = 2 + 4 + \dots + (2m - 4) + (2m - 2) + 2m = m(m + 1).$$

Therefore: 
$$R = 1 - \frac{6S(v_1 - v_2)}{N^2 - 1} = 1 - \frac{6m(m + 1)}{(2m + 1)^2 - 1} = -0.5 \dots \dots \dots (xxvii).$$

Accordingly when the correlation is negative and perfect, the number of observations being odd,  $R$  will never take the value  $-1$ , but no greater value than  $-0.5$ ; whereas if we reckon our second ranks in the negative direction  $R$  will equal  $+1$ .

Here the Spearman formula (xxv) leads to the absurd result  $\rho_{12} = -1/\sqrt{2}$ , instead of  $-1$ . On the other hand my formulae (xxiv) for  $\rho_{12} = -1$  and  $R = -0.5$  give absolutely the correct value  $r = -1$  for the variate correlation.

Again take  $N$  even  $= 2m$  and consider the system:

$v_1 =$	1	2	3	4	.	.	$2m$
$v_2 =$	$2m$	$2m - 1$	$2m - 2$	$2m - 3$	.	.	1

We find: 
$$S(v_1 - v_2)^2 = 2 \{ (2m - 1)^2 + (2m - 3)^2 + (2m - 5)^2 + \dots + 1^2 \}$$

$$= \frac{2m}{3} (4m^2 - 1),$$

and this gives 
$$\rho_{12} = 1 - \frac{6 \frac{2m}{3} (4m^2 - 1)}{2m(4m^2 - 1)} = -1.$$

Again: 
$$S(v_1 - v_2) = 1 + 3 + 5 + \dots + (2m - 5) + (2m - 3) + (2m - 1)$$

$$= m^2.$$



Hence : 
$$R = 1 - \frac{6m^2}{4m^2 - 1} = -\frac{2m^2 + 1}{4m^2 - 1} = -\cdot 5 \left\{ 1 + \frac{3}{N^2 - 1} \right\} \dots\dots(\text{xxviii}).$$

For : 
$$\begin{array}{ll} N = 4, & R = -\cdot 600; \\ N = 10, & R = -\cdot 515, \\ N = 20, & R = -\cdot 504; \\ N = 100, & R = -\cdot 500. \end{array}$$

Or, again, the limit  $-\cdot 5$  is rapidly reached as the number of observations increases. In fact solely for the simple case of *two* observations is it possible for  $R$  to reach  $-1$ .

If it be objected to (xxiv) that it would now give for values of  $R$ , greater than  $-\cdot 5$  values of the variate correlation greater than  $-1$  ( $= -1\cdot 09$  at a maximum for  $N = 4$ ), this is overlooking the point that (xxiv) is deduced from (xxii) by replacing true grades by spurious grades or ranks, and that if we retain (xxii) then

$$S(g_1 - g_2)/N^2 = \frac{m^2}{4m^2} = \frac{1}{4},$$

and  $r = -1$  as it should do.

We have now reached I think the basis of Spearman's apparent paradox. While the variation of the true rank correlation  $\rho_{12}$  lies between  $+1$  and  $-1$  and has  $\cdot 67449/\sqrt{N-1}$  for its probable error, the value of  $R$  only ranges between  $+1$  and  $-\cdot 5$ , and may well have a less value for its probable error.

Now Spearman tells us that large negative values of his  $R$  should be avoided\*. There is no necessity whatever for avoiding them if we are seeking the variate correlation by the formula given in this memoir. But if we are seeking the probable error of a zero quantity, which may vary on either side of zero (and in this case the variation is not symmetrical about zero), we cannot neglect the distribution of random variations below zero. If Spearman wishes his  $R$  to be considered always positive, then he ought to have found the probable error on the assumption that  $S(\nu_1 - \nu_2)$  should never be greater than  $\frac{1}{2}(N^2 - 1)$ . He has taken a quantity which ranges from  $+1$  to  $-\cdot 5$  and compared its random variations with one which ranges from  $+1$  to  $-1$  for the same frequency. If he had restricted his attention to variations of  $R$  between  $0$  and  $+1$  and of  $\rho_{12}$  between  $0$  and  $+1$  he would not have reached the same conclusion.

But there is a further very serious indictment to be made against Spearman's  $R$ . For values of  $N$  fairly small, which are those for which he proposes to use it,  $R$  retains a constant value for wide variations in  $\rho_{12}$ . We can show this on an exaggerated scale by writing down the possible values for Spearman's  $R$  and the true rank correlation for 4 individuals taken with random ranks. See Table on p. 23.

A little consideration will show to what much better results  $\rho_{12}$  leads us than  $R$ .  $R$  in fact remains constant and  $= -\cdot 2$  while  $\rho_{12}$  passes through the values  $0$ ,  $-\cdot 2$ ,  $-\cdot 4$  and  $-\cdot 6$ ; or  $r$  can take values from  $0$  to  $-\cdot 62$ , while its value as found from  $R$

\* *Loc. cit.*, footnote, p. 96.







remains  $-.38$ . This simple illustration of how the real rank correlation varies widely while Spearman's coefficient  $R$  remains constant shows how unsuitable the latter is, when we have to deal with small series.

Another point worth noting is that, if we take the positive values of the correlation only, the mean value of  $R$  is  $.3818$ , while the mean value of the corresponding  $\rho_{12}$ 's is  $.5454$ ; the former has a standard deviation of  $.2622$  and the latter of  $.2573$ , showing that we are not justified in asserting that  $R$  has a smaller probable error than  $\rho_{12}$  when we take comparable quantities.

Spearman appears to have an idea that  $R$  is really a coefficient comparable with  $\rho_{12}$ , and he attempts to get over some difficulties which have arisen, by telling us to reverse one series of ranks when  $R$  comes out negative. But reversing the ranks does not aid us to the right result. Thus if the ranks in the 12th and 13th column of  $\nu_2$  above be reversed, we find that  $R$  still remains negative and of the same magnitude  $-.2$ . In fact it is easy to write down a system of ranks which give a negative  $R$ , and which on reversal give a negative  $R$  six or seven times as big. The fact is simply that  $R$  is not a symmetrical function of  $\rho_{12}$  and reversal of ranks does not necessarily reverse  $\rho_{12}$  in sign.

We see accordingly (i) that the total range of  $R$  is only about  $\frac{2}{3}$  that of  $\rho_{12}$ , and that if we make the range the same by any attempt to reverse ranks, the Spearman method of calculating the probable error for  $R=0$  is erroneous. (ii) That the distribution of  $R$  for random rankings has a median which differs from zero, is very skew, and is in no ways comparable with that for  $\rho_{12}$ .

A point to be borne in mind most carefully is that for a given value of  $R$ ,  $\rho_{12}$  the true rank correlation may take a great variety of values. It is only when (i) the number of observations is fairly considerable, and (ii) we assume some distribution of associated grades such as that of normal correlation, that we are able to assert that the value of  $R$  will fix  $\rho_{12}$ , but such a relationship as that connecting  $\rho_{12}$ ,  $R$  and the variate correlation  $r$  can only be fixed, as in this memoir, by the appeal to despised mathematical analysis.

Thus the advantages claimed by Spearman for  $R$ , namely: (a) that it frees the discussion from the complexities of mathematical analysis, and (b) that it gives a less probable error than more usual ways of approaching the subject, are seen to be illusory.

The difficulty that  $\rho_{12}$  may take a whole series of values for a single value of  $R$  is only surmounted if we define the character of our frequency distribution, and there is no doubt that we shall obtain a first approximation by defining it as normal. Secondly, we cannot reverse ranks with the effect Spearman proposes, and if we could his probable error of  $R$  for  $R=0$  would be erroneous. Lastly, if we do not reverse ranks, then the probable error of one and the *same* quantity, the variate correlation, is considerably greater—for the only case yet worked out—i.e.  $R=0$ , when found by Spearman's method, than when found by the well-known method of squares of



differences, and still less than if found by the product of the variates directly. The squares of the differences of ranks can be taken so directly and quickly from a table of squares, that it does not seem to me that the slight rapidity gained in using positive differences of ranks is of any weight against its increased inaccuracy for small series, where indeed it is likely to be chiefly used.

Further no two rank correlations are in the least reliable or comparable unless we assume that the frequency distributions are of the same general character (see p. 9), and this general character will, till further advance be made in the theory of skew-correlation, be undoubtedly that provided by the hypothesis of normal distribution. On this assumption Spearman's suggestion of correlation of ranks becomes valid, but not as he supposes as a *Ding an sich*, but only as a means of passing at any rate to an approximation to the variate correlation, and this in the case of quantities where it is easier to rank individuals than to measure their attributes accurately.

For the grounds stated in this section, I propose to use as a rule  $\rho_{12}$  and not  $R$  to find  $r$ . For this reason I have spent my energies in finding the probable error of  $\rho_{12}$  instead of seeking that of  $R$ .

(8) *On the Probable Error of the Correlation of Grades.*

The following investigation is admittedly lengthy, but I have not seen my way to shorten it, and the main point is to reach by some road the expression for the probable error. The most general expression for the probable error of a correlation whatever be the distribution is to be found from  $\cdot 67449\Sigma_r$ , where\* :

$$\Sigma_r^2 = \frac{r^2}{N} \left\{ \frac{p_{22}}{p_{11}^2} + \frac{1}{2} \frac{p_{22}}{p_{20}p_{02}} + \frac{1}{4} \frac{p_{40}}{p_{20}^2} + \frac{1}{4} \frac{p_{04}}{p_{02}^2} - \frac{p_{31}}{p_{11}p_{20}} - \frac{p_{13}}{p_{11}p_{20}} \right\}$$

and 
$$p_{qs} = S \{ n_{xy} (x - \bar{x})^q (y - \bar{y})^s \} / N \dots\dots\dots(\text{xxix}).$$

Now in our case  $x$  and  $y$  are to be the grades  $g_1$  and  $g_2$  and  $r$  is to be

$$\rho_{12} = p_{11} \sqrt{p_{20} p_{02}},$$

which we will write for this investigation  $\rho$ .

We have at once :

$$N \times p_{m0} = \int_{-\infty}^{+\infty} (g_1 - \bar{g}_1)^m \frac{N}{\sqrt{2\pi\sigma_1}} e^{-\frac{1}{2}x^2/\sigma_1^2} dx,$$

and : 
$$g_1 = \bar{g}_1 + \frac{N}{\sqrt{2\pi\sigma_1}} \int_0^x e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} dx,$$

$$-dg_1 = \frac{N}{\sqrt{2\pi\sigma_1}} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} dx.$$

\* *Mathematical Contributions to the Theory of Evolution*, XIV. "Draper's Research Memoirs," Biometric Series II, p. 20. I have omitted certain terms which cancel.



Thus: 
$$p_{mo} = \frac{1}{N} \int_0^N (g_1 - \bar{g}_1)^m dg_1 = 0, \text{ if } m \text{ be odd,}$$

$$= \frac{1}{m+1} \left(\frac{1}{2}N\right)^m, \text{ if } m \text{ be even.}$$

Thus  $p_{40} = p_{04} = \frac{1}{80} N^4$ , and  $p_{20} = p_{02} = \frac{1}{12} N^2$ , as before .....(xxx).  
 We can now write (xxix) in the form :

$$\Sigma_p^2 = \frac{1}{N} \left\{ \frac{p_{22}}{p_{20}^2} (1 + \frac{1}{2}\rho^2) - 2\rho \frac{p_{31}}{p_{20}^2} + \frac{9}{16}\rho^2 \right\} \dots\dots\dots(\text{xxxix}),$$

assuming as we shall show in the sequel (p. 30) that  $p_{31} = p_{23}$ . Accordingly we have now to find  $p_{22}$  and  $p_{31}$ .

First to find: 
$$p_{22} = \frac{1}{N} \{S (g_1 - \bar{g}_1)^2 (g_2 - \bar{g}_2)^2\},$$

or if we use the notation of p. 11,

$$p_{22} = \frac{1}{N} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} i_1^2 i_2^2 z dx dy \dots\dots\dots(\text{xxxix}).$$

Now I have not succeeded in integrating this expression, although I have spent much time over it, but I have expanded it in powers of the variate correlation  $r$ .

If 
$$U = \frac{1}{\sqrt{1-r^2}} e^{-\frac{1}{2} \frac{1}{1-r^2} (x'^2 - 2rx'y' + y'^2)},$$

and  $j_x$  and  $j_y$  are the same as on p. 15, we can write :

$$p_{22} = \frac{N^4}{8\pi^3} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} j_x^2 j_y^2, U dx' dy'.$$

But\* 
$$U = S \frac{r^n}{n!} v_n w_n e^{-\frac{1}{2} (x'^2 + y'^2)},$$

or,

$$p_{22} = \frac{N^4}{8\pi^3} S \left( \frac{r^n}{n!} q_n^2 \right),$$

where 
$$q_n = \int_{-\infty}^{+\infty} j_x^2 e^{-\frac{1}{2}x'^2} v_n dx' = \int_{-\infty}^{+\infty} j_y^2 e^{-\frac{1}{2}y'^2} w_n dy' \dots\dots\dots(\text{xxxix}).$$

If  $n$  be odd,  $v_n$  and  $w_n$  have odd powers and  $q_n = 0$ , hence  $p_{22}$  contains only even powers of  $r$ .

First: 
$$q_0 = \int_{-\infty}^{+\infty} j_x^2 e^{-\frac{1}{2}x'^2} v_0 dx,$$

where we may drop the dashes from the letters now, and  $v_0 = 1$ . Therefore :

$$q_0 = \int_{-\infty}^{+\infty} j_x^2 \frac{dj_x}{dx} dx = \frac{1}{3} \left[ j_x^3 \right]_{-\sqrt{\frac{1}{2}\pi}}^{+\sqrt{\frac{1}{2}\pi}} = \sqrt{\frac{2\pi}{3}} \times \frac{\pi}{\sqrt{12}} \dots\dots\dots(\text{xxxix}).$$

\* Pearson: *Phil. Trans. A*, Vol. 195, p. 3.



Now\* :  $v_n e^{-\frac{1}{2}x^2} = -\frac{d}{dx}(v_{n-1} e^{-\frac{1}{2}x^2})$ , and  $v_{n-1} = \frac{1}{n} \frac{dv_n}{dx}$ .

Hence: 
$$q_n = -\int_{-\infty}^{+\infty} j_x^2 d(v_{n-1} e^{-\frac{1}{2}x^2}) = 2 \int_{-\infty}^{+\infty} j_x e^{-x^2} v_{n-1} dx$$

$$= \frac{2}{n} \int_{-\infty}^{+\infty} j_x e^{-x^2} dv_n = -\frac{2}{n} \int_{-\infty}^{+\infty} v_n (e^{-\frac{3}{2}x^2} - 2xe^{-x^2} j_x) dx$$

$$= -\frac{2}{n} \int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2} dx + \frac{4}{n} \int_{-\infty}^{+\infty} j_x e^{-x^2} x v_n dx.$$

But† :  $xv_n = v_{n+1} + \frac{dv_n}{dx}$ , thus :

$$q_n = -\frac{2}{n} \int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2} dx + \frac{4}{n} \int_{-\infty}^{+\infty} j_x e^{-x^2} v_{n+1} dx + \frac{4}{n} \int_{-\infty}^{+\infty} j_x e^{-x^2} \frac{dv_n}{dx} dx$$

$$= -\frac{2}{n} \int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2} dx + \frac{2}{n} q_{n+2} + 2q_n.$$

Or :  $q_{n+2} + \frac{1}{2} n q_n = \int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2} dx = \beta_n$  say, .....(xxxv).

It now remains to find  $\beta_n$ .

$$\beta_n = \int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2} dx = \int_{-\infty}^{+\infty} \{xv_{n-1} - (n-1)v_{n-2}\} e^{-\frac{3}{2}x^2} dx$$

$$= -\frac{1}{3} \int_{-\infty}^{+\infty} \frac{d}{dx} (e^{-\frac{3}{2}x^2}) v_{n-1} dx - (n-1) \int_{-\infty}^{+\infty} v_{n-2} e^{-\frac{3}{2}x^2} dx$$

$$= -\frac{n-1}{3} \int_{-\infty}^{+\infty} e^{-\frac{3}{2}x^2} v_{n-2} dx - (n-1) \int_{-\infty}^{+\infty} v_{n-2} e^{-\frac{3}{2}x^2} dx$$

$$= -\frac{2}{3} (n-1) \int_{-\infty}^{+\infty} v_{n-2} e^{-\frac{3}{2}x^2} dx = -\frac{2}{3} (n-1) \beta_{n-2}$$

$$= (-\frac{2}{3})^{\frac{1}{2}n} (n-1)(n-3)(n-5) \dots \dots 1 \times \beta_0, n \text{ being of course even.}$$

But  $\beta_0 = \int_{-\infty}^{+\infty} v_0 e^{-\frac{3}{2}x^2} dx = \sqrt{\frac{2}{3}\pi}$ .

Thus we have the reduction formula :

$$q_{n+2} + \frac{1}{2} n q_n = (-\frac{2}{3})^{\frac{1}{2}n} (n-1)(n-3)(n-5) \dots \dots 1 \times \sqrt{\frac{2\pi}{3}} \dots \dots \text{(xxxvi)}$$

\* *loc. cit.* p. 5.

† *loc. cit.* p. 4.



We can now rapidly calculate the  $q$ 's.

$$q_0 = \sqrt{\frac{2\pi}{3}} \frac{\pi}{\sqrt{12}}, \quad q_2^* = \sqrt{\frac{2\pi}{3}}, \quad q_4 = -\frac{5}{3} \sqrt{\frac{2\pi}{3}}$$

$$q_6 = \frac{14}{3} \sqrt{\frac{2\pi}{3}}, \quad q_8 = -\frac{166}{9} \sqrt{\frac{2\pi}{3}}, \quad q_{10} = \frac{2552}{27} \sqrt{\frac{2\pi}{3}}.$$

This is probably more than sufficient for most practical purposes. Evaluating the coefficients numerically we have from (xxxiii):

$$\frac{p_{22}}{p_{20}^2} = 1 + \cdot 607,9271r^2 + \cdot 140,7239r^4 + \cdot 036,7758r^6$$

$$+ \cdot 010,2587r^8 + \cdot 002,9933r^{10} \dots\dots\dots(\text{xxxvii}).$$

To test the accuracy of this result—obviously correct for  $r=0$ —consider  $r=1$ . We have:

$$(p_{22}/p_{20}^2)_{r=1} = 1.798,6788 \dots\dots\dots(\text{xxxviii}).$$

But in the case the variate correlation surface becomes a ridge and  $i_1 = i_2$ , or:

$$(p_{22})_{r=1} = \frac{1}{N} \int_{-\infty}^{+\infty} i_1^4 \frac{N}{\sqrt{2\pi\sigma_1}} e^{-\frac{1}{2} \frac{x^2}{\sigma_1^2}} dx$$

$$= \frac{1}{N} \int_{-N/2}^{+N/2} i_1^4 di_1 = \frac{N^4}{80}.$$

Thus 
$$\left(\frac{p_{22}}{p_{20}^2}\right)_{r=1} = \frac{144}{80} = 1.8 \dots\dots\dots(\text{xxxix}).$$

The difference between (xxxviii) and (xxxix) is only .001,3212 or about .07 per cent. Thus even if we omit the term in  $r^{10}$ , we shall be less than .2 per cent in error in this extreme case, when the probable error itself is zero; and for lesser values of  $r$ , where the probable error is sensible, we shall not be as much as .01 per cent in error. This is amply sufficient for statistical purposes. I now take  $p_{21}$  and find its value in a different manner.

$$p_{21} = \frac{1}{N} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} i_1^3 i_2 z dx dy,$$

$$\frac{dp_{21}}{dr} = \frac{1}{N} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} i_1^3 i_2 \frac{dz}{dr} dx dy = \frac{\sigma_1 \sigma_2}{N} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} i_1^3 i_2 \frac{d^2 z}{dx dy} dx dy.$$

This can be integrated twice by parts, and the part between limits vanishes at each integration. Writing  $x = \sigma_1 x'$ ,  $y = \sigma_2 y'$  as before, we have:

$$\frac{dp_{21}}{dr} = \frac{3N^4}{(2\pi)^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} j_x^2 e^{-\frac{1}{2}(x'^2+y'^2)} z' dx' dy'.$$

\* This is found at once from  $q_n = 2 \int_{-\infty}^{+\infty} j_x e^{-x^2} v_{n-1} dx$ , or  $q_2 = - \int_{-\infty}^{+\infty} j_x de^{-x^2}$ , since  $v_{n-1} = x$ . Thus

$$q_2 = \int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi} \frac{1}{\sqrt{3}}.$$



The integration with regard to  $y'$  can now be completed and we find :

$$\frac{dp_{31}}{dr} = \frac{3N^4}{(\sqrt{2\pi})^3} \frac{1}{\sqrt{2-r^2}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2} \frac{4-r^2}{2-r^2} x^2} j_x^2 dx \dots\dots\dots(xl),$$

where 
$$j_x = \int_0^x e^{-\frac{1}{2} x^2} dx,$$

and we have dropped the dash on  $x$  as no longer of service.

Write  $m = (4 - r^2)/(2 - r^2)$ , and we must now find :

$$I = \int_{-\infty}^{+\infty} e^{-\frac{1}{2} m x^2} j_x^2 dx \dots\dots\dots(xli).$$

Now :

$$\begin{aligned} \frac{dI}{dm} &= -\frac{1}{2} \int_{-\infty}^{+\infty} x^2 e^{-\frac{1}{2} m x^2} j_x^2 dx = \frac{1}{2m} \int_{-\infty}^{+\infty} x j_x^2 d(e^{-\frac{1}{2} m x^2}) \\ &= -\frac{1}{2m} \int_{-\infty}^{+\infty} e^{-\frac{1}{2} m x^2} d(x j_x^2) = -\frac{I}{2m} - \frac{1}{m} \int_{-\infty}^{+\infty} x j_x e^{-\frac{1}{2} (m+1) x^2} dx, \\ &= -\frac{I}{2m} + \frac{1}{m(m+1)} \int_{-\infty}^{+\infty} j_x d(e^{-\frac{1}{2} (m+1) x^2}), \end{aligned}$$

or: 
$$\frac{dI}{dm} + \frac{I}{2m} = -\frac{1}{m(m+1)} \int_{-\infty}^{+\infty} e^{-\frac{1}{2} (m+2) x^2} dx = -\frac{\sqrt{2\pi}}{m(m+1)\sqrt{m+2}},$$

thus: 
$$\frac{d}{dm} (\sqrt{m} I) = -\frac{\sqrt{2\pi}}{(m+1)\sqrt{(m+1)^2-1}}.$$

Thus: 
$$\sqrt{m} I = \text{constant} - \sqrt{2\pi} \cos^{-1} \frac{1}{m+1} \dots\dots\dots(xlii).$$

To evaluate the constant\* put  $m = 1$ , and we have :

$$\begin{aligned} \text{constant} &= I_{m=1} + \sqrt{2\pi} \cos^{-1} \frac{1}{2} \\ &= \int_{-\infty}^{+\infty} e^{-\frac{1}{2} x^2} j_x^2 dx + \sqrt{2\pi} \frac{\pi}{3} \\ &= \int_{-\sqrt{\pi/2}}^{+\sqrt{\pi/2}} j_x^2 dj_x + \sqrt{2\pi} \frac{\pi}{3} = \pi^{3/2} / \sqrt{2}. \end{aligned}$$

Or finally: 
$$\begin{aligned} I &= \frac{\sqrt{2\pi}}{\sqrt{m}} \left\{ \frac{\pi}{2} - \cos^{-1} \frac{1}{m+1} \right\} \\ &= \frac{\sqrt{2\pi}}{\sqrt{m}} \sin^{-1} \frac{1}{m+1} \dots\dots\dots(xliii). \end{aligned}$$

\* Mr L. F. Richardson has shown me that if we put  $m = 0$ , since the inverse cosine now vanishes :  
 constant = Limit  $\sigma = \infty$  of  $\sqrt{2\pi} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2} \frac{x^2}{\sigma^2}} j_x^2 dx$ , which he has evaluated with the same result.



Returning now to (xl) we can replace  $m$  by its value in terms of  $r$  and write

$$\frac{d}{dr} \left( \frac{p_{31}}{p_{20}^2} \right) = \frac{108}{\pi^2} \frac{1}{\sqrt{4-r^2}} \sin^{-1} \frac{2-r^2}{2(3-r^2)} \dots\dots\dots(xliv).$$

This expression I have not succeeded in integrating. I have therefore expanded it in  $r^2$  and then integrated. Since  $p_{31} = 0$  for  $r = 0$ , we see the constant is zero after integration; thus after some troublesome expansions I find:

$$\frac{p_{31}}{p_{20}^2} = \frac{54}{\pi^2} \{.339,8369r - .005,4820r^2 - .003,6798r^3 - .001,1836r^4\} \dots\dots(xlv).$$

The value of  $p_{31}$  is clearly the same as  $p_{13}$  for nothing would be altered if  $x$  and  $y$  were interchanged from (xl) onwards. To test the accuracy of the result, suppose  $r = 1$ . Then we have from the 'ridge':

$$\begin{aligned} (p_{31})_{r=1} &= \frac{1}{N} \int_{-\infty}^{+\infty} i_1^3 i_2 \frac{N}{\sqrt{2\pi\sigma_1}} e^{-\frac{1}{2}x^2/\sigma_1^2} dx_1 \text{ and } i_2 = i_1 \\ &= \frac{1}{N} \int_{-N/2}^{+N/2} i_1^4 di_1 = \frac{N^4}{80}, \end{aligned}$$

or:  $\left( \frac{p_{31}}{p_{20}^2} \right)_{r=1} = 1.8$ , again.

But (xlv) gives us:

$$\left( \frac{p_{31}}{p_{20}^2} \right)_{r=1} = 1.8 \times 1.00153,$$

that is a result at a maximum only .15 per cent. in error and correct enough for all statistical purposes. The next step is to determine the powers of  $r$  in terms of  $\rho$ , and substitute in the expressions just found for  $p_{22}/p_{20}^2$  and  $p_{31}/p_{20}^2$ . I find:

$$\frac{p_{22}}{p_{20}^2} = 1 + .666,6667\rho^2 + .108,3084\rho^4 + .019,7955\rho^6 + .002,7683\rho^8 \dots\dots(xlvi),$$

and:  $\frac{p_{31}}{p_{20}^2} = 1.947,1220\rho - .123,4135\rho^3 - .019,4138\rho^5 - .003,8120\rho^7 \dots\dots(xlvii).$

To verify we note that for  $\rho = 1$ , these give 1.7975 and 1.8005 instead of 1.8,—quite sufficiently close for the purpose in view.

We now substitute in (xxxi) and find as far as  $\rho^8$  that:

$$\Sigma_\rho^2 = \frac{1}{N} \left\{ 1 - 1.827,5773\rho^2 + .688,4687\rho^4 + .112,7773\rho^6 + .020,2900\rho^8 \right\} \quad (xlviii).$$

I throw this, by dividing by  $(1-\rho^2)^2$ , into the form:

$$\Sigma_\rho^2 = \frac{(1-\rho^2)^2}{N} \left\{ 1 + .086,2113\rho^2 + .012,9408\rho^4 + .002,3757\rho^6 + .000,0822\rho^8 \right\}^2,$$

or, dropping unnecessary decimals:

$$\Sigma_\rho^2 = \frac{1-\rho^2}{\sqrt{N}} \left\{ 1 + .086\rho^2 + .013\rho^4 + .002\rho^6 \right\} \dots\dots\dots(xlix).$$



Thus we see that the distribution of grades being very far from normal, the probable error  $\cdot67449\Sigma_p$  of the correlation of grades exceeds the value  $\cdot67449(1-\rho^2)/\sqrt{N}$ , which it would take on the hypothesis of normal correlation by a factor which can amount to about 10 per cent. at a maximum, but gives 0 per cent. excess when  $\rho=0$ , then agreeing with our previous result.

I propose now to find the probable error in  $r$  as determined by grade methods in terms of  $r$ . This involves expressing  $\rho$  and  $\rho^2$  in terms of  $r$ ; these are easily found from the known expansions for  $\sin^{-1}x$  and  $(\sin^{-1}x)^2$ . We have :

$$1 + \frac{1}{2}\rho^2 = 1 + \cdot455,9453r^2 + \cdot037,9954r^4 + \cdot005,0661r^6 + \cdot000,8142r^8,$$

$$2\rho = 1\cdot909,8593r + \cdot079,5775r^3 + \cdot009,2650r^5 + \cdot001,3322r^7.$$

These must be used in (xxxv), which may be written in the form :

$$\Sigma_p^2 = \frac{1}{N} \left\{ \left( \frac{p_{22}}{p_{20}^2} + 1\cdot8 \right) (1 + \frac{1}{2}\rho^2) - 2\rho \frac{p_{21}}{p_{20}} - 1\cdot8 \right\}.$$

Hence using (xxxviii) and (xlv), we deduce after some troublesome multiplications :

$$\Sigma_p^2 = \frac{1}{N} \{ 1 - 1\cdot666,5507r^2 + \cdot433,6130r^4 + \cdot161,8337r^6 + \cdot049,5042r^8 \}$$

$$= \frac{1}{N} (1 - r^2)^2 \{ 1 + \cdot333,4493r^2 + \cdot100,5116r^4 + \cdot029,4076r^6 + \cdot007,8078r^8 \}.$$

But since :  $r = 2 \sin \frac{\pi}{6} \rho,$

$$\delta r = \frac{\pi}{3} \cos \frac{\pi}{6} \rho \times \delta \rho \text{ and } \Sigma_r = \frac{\pi}{3} \sqrt{1 - \frac{r^2}{4}} \Sigma_p.$$

Thus :

$$\Sigma_r^2 = \frac{\pi^2}{9} \frac{(1 - r^2)^2}{N} \{ 1 + \cdot083,4493r^2 + \cdot017,1493r^4 + \cdot004,2797r^6 + \cdot000,4559r^8 \}.$$

Taking the square root we have :

$$\Sigma_r = 1\cdot0472 \frac{1 - r^2}{\sqrt{N}} \{ 1 + \cdot041,7246r^2 + \cdot007,7042r^4 + \cdot001,8184r^6 + \cdot000,1224r^8 \},$$

or, for all practical purposes, the probable error of  $r$  found from the grade correlation is,

$$\text{P. E. of } r = \cdot70633 \frac{1 - r^2}{\sqrt{N}} \{ 1 + \cdot042r^2 + \cdot008r^4 + \cdot002r^6 \} \dots\dots\dots(\text{i}).$$

Clearly for all values of  $r$ , this is larger than the probable error of the correlation  $r$  found by the product moment method, i.e.  $\cdot67449 \frac{1 - r^2}{\sqrt{N}}$ . The maximum difference, as  $r$  approaches unity is 10 per cent. The value can always be found from (i) without any trouble. The completer value is singularly close to

$$\frac{\cdot70633}{\sqrt{N}} \frac{1 - r^2}{\{1 - \frac{1}{3}r^2\}^{\frac{1}{2}}} \dots\dots\dots(\text{ii}),$$



but no advantage is gained in calculation by using this form, as tables of powers of  $r$  up to the 6th exist\*.

We see therefore from this section that whatever be the value of  $r$ , then for normal frequency the probable error of  $r$  found by the product moment method is less than the value found by the correlation of grades. Further there is no reason for supposing that the probable error of  $r$  found from the difference of grades ( $R$ ) is not greater than the probable error of  $r$  found from the product moment of grades.

We accordingly conclude that the new methods are less accurate than the old. But they possess some advantages,—when ranks can be easily determined,—in rapidity of calculating, and there are undoubtedly cases where they can be used effectively. In saying this I must reassert that I do not believe there is any advantage in the knowledge of rank correlation in itself; I look upon it as a mere stage to the discovery of the variate correlation. For the comparability of rank correlations depends upon the sameness of type in the frequency distributions, and this assumption is the weak step in the method. Granted approximately normal distributions, then the variate correlation flows from the rank correlation, and the whole investigation gains a rich significance.

My remaining sections will be devoted to illustration of the new methods and their comparison with the old.

(9) *Illustration III. Correlation of National Debt and Population.*

The following table is based on data for the year 1900, and raises no pretence to exactness, or financial accuracy. It is merely illustrative.

TABLE IV. *Population and Indebtedness of Various States 1900.*

State	Population in millions	Debt in million £	Population Rank	Debt Rank	$v_1 - v_2$	$(v_1 - v_2)^2$
Russia	129.20	1097.0	1	2	-1	1
United States	76.40	200.0	2	8	-6	36
German Empire†	56.34	649.4	3	4	-1	1
Austria	47.01	226.7	4	7	-3	9
Japan	43.80	51.5	5	15	-10	100
United Kingdom	41.60	705.0	6	3	+3	9
France	38.64	1242.1	7	1	+6	36
Italy	32.10	500.0	8	5	+3	9
Turkey	20.30	162.0	9	9	0	0
Spain	18.10	385.0	10	6	+4	16
Belgium	6.82	106.4	11	11	0	0
Roumania	5.50	58.0	12	14	-2	4
Sweden	5.14	18.6	13	17	-4	16
Holland	5.10	95.6	14	12	+2	4
Portugal	4.70	155.0	15	10	+5	25
Argentine	4.50	86.4	16	13	+3	9
Switzerland	3.30	3.6	17	20	-3	9
Greece	2.40	28.0	18	16	+2	4
Norway	2.20	12.7	19	18	+1	1
Denmark	2.18	11.6	20	19	+1	1
			—	—	$30 = S(v_1 - v_2)$	$290 = S(v_1 - v_2)^2$

\* See *Biometrika*, Vol. II. p. 474.

† Imperial debt and sum of state debts.



Hence:  $N^2 - 1 = 399$ , and  $\rho_{12} = 1 - 6 \times 290 / (20 \times 399) = .7820$ .

Further:  $R = 1 - 6 \times 30 / 399 = .5489$ .

These values are obtained in a few minutes, if the ranks have once been written down. If  $\rho_{12}$  only be required, we need not write down the  $\nu_1 - \nu_2$  column at all, the squares being placed down straight away from the rank columns.

Now applying equations (xxiv) we determine:

$$\begin{aligned} r &= .7962, \text{ found from } \rho_{12}, \\ &= .7810, \text{ found from } R. \end{aligned}$$

The probable error of  $r$  found from  $\rho_{12}$  as given by Equation (1) is .0596. Thus we conclude that

$$\begin{aligned} r &= .80 \pm .06, \text{ found from } \rho_{12} \\ &= .78 \pm > .063, \text{ found from } R^*. \end{aligned}$$

If we turn to the much more laborious method of moments, we find:

Mean Population = 27.26 millions; Mean Debt = 289.7 million £,

S. D. Population = 31.74 millions; S. D. Debt = 357.9 million £.

Now these results in themselves should be sufficient to warn us that both distributions are very far from normal; for the S. D.'s in both cases are greater than the means, and since in a normal distribution, we might easily have a deviation equal to the S. D. we should on that hypothesis expect to get negative debts and negative populations. The distributions are therefore very skew, or in clubbing together great and small powers, we have introduced excessive heterogeneity, completely destroying any approach to normality<sup>†</sup>. If we work out the value of  $r$  by the product moment method, we find:

$$r = .68 \pm .08.$$

We see at once that the rank method has so exaggerated the correlation that it has made the probable error of the less exact methods less than the probable error of the more exact method! The explanation of this lies simply in the fact that the system we are dealing with is not normal. If the ranks of two variables were those given in Table IV, and the distribution were normal, then the variate correlation would be .80; it actually takes the value .68, and this is a very good illustration of how much the nature of the distribution may affect a judgment from ranks.

\* The p. e. is of the form  $\frac{.7738}{\sqrt{n}} (1 - r^2) (1 + c_1 r^2 + c_2 r^4 + c_3 r^6)$ , the  $c$ 's being positive unknown constants, and this is  $> .063$ .

† If we confine our attention to the seven "great powers," Austria, France, Germany, Great Britain, Italy, Russia and the United States, we find  $\rho_{12} = -.143$ ,  $R = -.125$ , giving  $r = -.15$  and  $-.23$  with a probable error of .3; this result again emphasises the heterogeneity of the material.



Of course it is doubtful whether when we are in ignorance of the character of the distribution we could say more than

$$r = .8 \pm .1, \text{ found from } \rho_{12},$$

and  $r = .7 \pm .1, \text{ found by product-moment.}$

These might then be treated as identical for some purposes of inference. But the advantage of the longer product-moment method would be that it would have taught us that the correlation was non-Gaussian, and given us in the process the regression line. This would probably more than compensate for its greater laboriousness.

(10) *Illustration IV. Correlation between mean Size of Litter in a Generation and mean Sex Ratio in the same Generation in the case of Mice.*

The following data are taken from a paper in *Biometrika*, Vol. v., p. 439.

TABLE V.

Generation	Mean size of Litter	Mean Sex Ratio	Litter Rank	Sex Ratio Rank	$v_1 - v_2$	$(v_1 - v_2)^2$
1st	5.06	.505	5	3	+ 2	4
2nd	4.94	.491	6	4	+ 2	4
3rd	5.96	.523	1	2	- 1	1
4th	5.93	.542	2	1	+ 1	1
5th	5.53	.462	3	6	- 3	9
6th	5.23	.483	4	5	- 1	1

Thus :  $S(v_1 - v_2)^2 = 20, \quad S(v_1 - v_2) = 5,$

and  $\rho_{12} = .429, \quad R = .143.$

Whence :  $r \text{ from } \rho_{12} = .45 \pm .23,$   
 $r \text{ from } R = .25 \pm < .23.$

The actual value of  $r$  from product-moment is

$$r = .63 \pm .17.$$

This example serves to show that the correlation found from  $R$  may when the observations are few, not be definitely significant, while when we proceed in the more accurate manner it is definitely significant. The  $R$ -method is thus shown not to have special advantages, but rather peculiar disadvantages for short series. Its merit really lies in rapidity of working for assay purposes and rough treatment.

(11) *Illustration V. Resemblance of Cousins.*

(a) *Width of Hand.* The following table gives the width of the hand in 34 pairs of male adult cousins taken from my series of Cousin Measurements. These data are being used by Miss Ethel M. Elderton in a forthcoming paper on this



subject, and I have most heartily to thank her for the exhaustive manner in which she has dealt with the material in order to illustrate the whole subject of determining correlation by ranks.

TABLE VI. *Width of Hand in mm. in Pairs of Male Adult Cousins.*

1st cousin <i>x</i>	2nd cousin <i>y</i>	Rank A		Rank B		$r_1 - r_2$	$(r_1 - r_2)^2$	True Grade of A		True Grade of B	
80.7	80.0	23	17	17	23	6	36	23.51	20.74	20.74	23.51
90.0	80.0	58	17	17	58	41	1681	58.82	20.74	20.74	58.82
80.7	84.7	23	48	48	23	-25	625	23.51	40.66	40.66	23.51
90.0	84.7	58	48	48	58	10	100	58.82	40.66	40.66	58.82
80.0	84.7	17	48	48	17	-31	961	20.74	40.66	40.66	20.74
74.5	81.0	3	26	26	3	-23	529	5.52	24.74	24.74	5.52
81.0	80.0	26	17	17	26	9	81	24.74	20.74	20.74	24.74
86.0	81.0	52	26	26	52	26	676	46.00	24.74	24.74	46.00
80.7	83.7	23	43	43	23	-20	400	23.51	36.36	36.36	23.51
94.0	82.7	64	37	37	64	27	729	65.26	31.99	31.99	65.26
94.0	81.7	64	34	34	64	30	900	65.26	27.68	27.68	65.26
76.0	77.0	5	11	11	5	-6	36	8.44	10.90	10.90	8.44
76.0	79.0	5	16	16	5	-11	121	8.44	17.08	17.08	8.44
76.0	83.0	5	41	41	5	-36	1296	8.44	33.29	33.29	8.44
86.3	88.3	53	54	54	53	-1	1	47.16	54.16	54.16	47.16
92.5	85.0	60	51	51	60	9	81	63.51	41.94	41.94	63.51
83.7	81.7	43	34	34	43	9	81	36.36	27.68	27.68	36.36
83.7	83.3	43	42	42	43	1	1	36.36	34.62	34.62	36.36
83.7	78.7	43	15	15	43	28	784	36.36	16.05	16.05	36.36
82.0	81.0	36	26	26	36	10	100	28.95	24.74	24.74	28.95
80.5	80.0	22	17	17	22	5	25	22.71	20.74	20.74	22.71
75.0	76.0	4	5	5	4	-1	1	6.40	8.44	8.44	6.40
71.0	76.0	1	5	5	1	-4	16	1.70	8.44	8.44	1.70
73.0	77.0	2	11	11	2	-9	81	3.45	8.44	8.44	3.45
84.5	78.0	47	13	13	47	34	1156	39.81	13.78	13.78	39.81
76.0	78.0	5	13	13	5	-8	64	8.44	13.78	13.78	8.44
93.3	89.7	61	55	55	61	6	36	64.53	58.09	58.09	64.53
93.3	82.7	61	37	37	61	24	576	64.53	31.99	31.99	64.53
98.7	82.7	66	37	37	66	29	841	67.58	31.99	31.99	67.58
89.7	81.0	55	26	26	55	29	841	58.09	24.74	24.74	58.09
81.0	93.3	26	61	61	26	-35	1225	24.74	64.53	64.53	24.74
82.7	81.0	37	26	26	37	11	121	31.99	24.74	24.74	31.99
98.7	81.0	66	26	26	66	40	1600	67.58	24.74	24.74	67.58
98.7	89.7	66	55	55	66	11	121	67.58	58.09	58.09	67.58
Mean } Size }	= 83.16		Mean } Rank }	= 34		$S(r_1 - r_2)$ = 605	$S(r_1 - r_2)^2$ = 2 × 15923	Mean } Grade }	= 32.41		

The measurements were only read to the millimetre, but since measurements were taken two or three times in each case the fractions .3, .5 or .7 arise, when averaging. Since either cousin may be the "first" cousin, we have for a symmetrical table 68 pairs. In the third and fourth columns, we have the ranks placed, according as to which cousin is considered the "first." It will at once be obvious that many ties arise; thus no less than eight individuals tie with a width of hand 81 mm. at rank 26. It is not so clear what rank ought to be given to them. They run from 26 to 33, we may call them all 29.5. We shall speak of this as the mid-rank method. Or, we



might put them all at 26, because this would probably be the result nearest to the true grade\*. We shall speak of this as the bracket-rank method †.

The above table illustrates the work for the bracket-rank method in columns 5 and 6, the differences of ranks  $A$  and  $B$  being, however, only written down *once*, so that to find  $S(\nu_1 - \nu_2)$ , we must sum all quantities in the fifth column as if they had the same sign, and double the sum of their squares in the sixth column.

We find :  $R = \cdot 2148$  and  $\rho_{12} = \cdot 3922$ ,  
whence  $r$  from  $\rho_{12} = \cdot 408 \pm \cdot 072$ ,  
 $r$  from  $R = \cdot 361 \pm > \cdot 072$ .

If we now investigate the value of  $R$  and  $\rho_{12}$  from the mid-ranks, we find that  $S(\nu_1 - \nu_2) = 588$  and  $S(\nu_1 - \nu_2)^2 = 29812$ . Accordingly :

$R = \cdot 2369$ , and  $\rho_{12} = \cdot 4310$ .  
Whence :  $r$  from  $\rho_{12} = \cdot 448 \pm \cdot 069$ ,  
 $r$  from  $R = \cdot 396 \pm > \cdot 069$ .

Both these values for  $r$  are higher than those determined by the bracket-rank process. We must then question whether the mid-rank or the bracket-rank method is the better. Or, indeed is it not possible, that sometimes the one, and sometimes the other will be the closer according to the nature of the frequency distribution ?

To illustrate this point the actual grades on the basis of normal distribution have been calculated by Eqn. (xii). It must be remembered that  $\cdot 5$  has to be added to the grade to obtain the rank, Eqn. (xiii).

We find : Mean width of hand = 83.16 mm.  
Standard Deviation = 6.201 mm.

As illustration of the method consider the hand of width 84.7 mm., its deviation is 1.54 and the ratio of this to the S. D. =  $\cdot 248$ , this corresponds to a value of  $\frac{1}{2}(1 + \alpha)$ , in the notation of Sheppard's Tables, =  $\cdot 59793$  and multiplied by 68 gives the grade 40.66, corresponding to a rank 41.16, as against the observed rank 48 or a mid-rank 49! Thus the actual size of organ corresponding to a bracket rank may differ widely from the size really belonging to the ranked organ, or the true grade in a general population differ very considerably from the spurious grade or rank in the sample used. This point again indicates how little can be judged from ranks unless we associate the rank distribution with some frequency hypothesis.

Having found the true grades we may correlate them together to find  $\rho_{12}$ , but in using the formula

$$\rho_{12} = 1 - \frac{S(g_1 - g_2)^2}{2N \times \sigma_g^2},$$

\* That is, find  $\sigma$ , and calculate  $g_1$  and  $g_2$  from Eqn. (xii) p. 10; the true grade in this case is 24.74, and  $\nu_1 = g_1 + \cdot 5 = 25.24$  is even below 26, not above it.

† To adopt a term from the examination world, where the place number of the bracket is measured only by those above.



we may adopt either the theoretical value  $\frac{1}{12}N^2$  for  $\sigma_g^2$ , or we can actually calculate its value. Now  $\frac{1}{12}N^2 = 385\frac{1}{3}$  and  $\sigma_g^2 = 365\cdot94$ , and thus there is a very considerable deviation from normality in the series\*  $S(g_1 - g_2)^2 = 31153\cdot195$ , and thus:

$$\rho_{12} \text{ found from the true } \sigma_g^2 = \cdot3740,$$

$$\rho_{12} \text{ found from } \sigma_g^2 = \frac{1}{12}N^2 = \cdot4055.$$

Whence :

$$r \text{ from true } \sigma_g^2 = \cdot3890,$$

$$r \text{ from } \sigma_g^2 = \frac{1}{12}N^2 = \cdot4215.$$

If we might judge from this single case we should conclude that the bracket-rank method gave a closer result to the grade method than the mid-rank method. But the question now arises, how close after all are all these grade rank methods to the correlation coefficient in any short series such as the present?

Accordingly the series was worked out by product moment and the result obtained was

$$r = \cdot331 \pm \cdot073.$$

Thus we see that the actual correlation is considerably lower than that given by any of the rank or grade processes. It is perfectly true that  $\cdot33$  and  $\cdot45$  are within double the probable error, and therefore two different random samples of the real population might have given as widely divergent results. But this is really the case of two different methods applied to the *same* sample. And further the actual correlation tells us that as far as this sample is concerned the true answer is likely to lie between  $\cdot19$  and  $\cdot48$ , but the mid-rank method tells us that it is likely to lie between  $\cdot31$  and  $\cdot58$ †. Now it is clear we might for some extraneous reason hold the value likely to be  $\cdot56$ , and we should find nothing to contradict this in the mid-rank result. But the proper method of determining  $r$  would show us that such a value was itself very unlikely. Thus the latter method when it diverges less than twice the probable error from the result of the rank method may yet forbid us to interpret the results in a manner admissible on the rank method. We cannot argue in like manner from the grade or rank result because that method has assumed an hypothesis, not made in the product-moment treatment, i.e. that of normal correlation, which is here not justified by the results.

But even the amount of agreement here noted is to be considered rather exceptional. I owe to Miss Elderton the working out of three other pairs of characters in the same set of male cousins each in five different ways. I have myself done each of them in three more ways, namely by Variate Differences as in Art. 2, and by the  $R$  method. The results are given in the Table below.

\* The mean grade in fact =  $32\cdot41$  and not  $34$  also.

† Taking a range of twice the probable error on either side the means.



TABLE VII. *Comparison of Correlation Coefficients found by Various Methods. Resemblance of Hand in 68 Pairs of Male Cousins.*

Character	Product Moment	Variate Difference	Grades		Ranks			
			True $\sigma_p^2$	$\sigma_p^2 = \frac{1}{12}N^2$	Bracket-Rank		Mid-Rank	
					By $\rho_{12}$	By $R$	By $\rho_{12}$	By $R$
Width of Hand	$.33 \pm .07$	.37	.39	.42	$.41 \pm .07$	.36	$.45 \pm .07$	.40
Width of Wrist	$.17 \pm .08$	.25	.12	.22	$.07 \pm .085$	.05	$.08 \pm .085$	.03
Length of Index Finger	$.19 \pm .08$	.14	.16	.13	$.21 \pm .08$	.29	$.19 \pm .085$	.29
Length of Little Finger	$.29 \pm .075$	.26	.18	.30	$.20 \pm .08$	.19	$.24 \pm .08$	.21
Mean of Four Results	.25	.25	.21	.25	.22	.22	.24	.23
Root Mean Square Deviation from true $r$		.053	.069	.060	.079	.094	.079	.096

It will, I think, be clear from this table that for series even with as many as 68 pairs—and this is approaching the limit at which any time is gained by using rank methods—we cannot hope to ascertain the correlation of the *sample* by such methods within about .1 of its value, and as the probable error of the sample may be .07, we may well deviate .2 from the population value in our estimate. We are accordingly very unlikely to reach reliable results by rank methods for the 8 to 10 observations to which Dr Spearman proposes to apply his  $R$ -method. We see that the mean values are fairly close, although the variate difference and the second grade methods give the best results. Judged by mean square deviations from product moment results, the variate difference is easily first, then come the laborious grade methods, the rank methods by  $\rho_{12}$  about fifty per cent. worse than the variate difference, and lastly the  $R$  methods not quite 100 per cent. worse. Thus we note that when a series is not fairly long and not approximately normal, the different rank and grade methods will give very diverse results. But when a series is fairly long, say 100 or more observations, then there is no advantage in rapidity from the rank method; the formation of a grouped correlation table, and the use of the product moment is just as rapid, and further conveys a great deal more of valuable information.

(12) *Conclusions.* Three new methods of determining variate correlation have been given in this paper. The first, that of variate differences, seems likely to be of some service in the case of symmetrical tables containing large numbers, the frequency being approximately normal, homotyposis tables may be taken as illustration.

The second that of deducing variate correlation from correlation of ranks, may be of service when it is not possible to put a quantitative value on the individual character. Thus it might be easy to form a relative series of intensity of pigment, and place individuals in rank. But mere correlation of ranks is not in itself a com-



parable character, as the variate correlation may have widely different values for the same ranking. Justification for the comparability depends upon assuming a wide spread rule of frequency distribution, and this rule can hardly be other than normality. The present paper shows how to deduce variate correlation from correlation of ranks. It shows, however, that such a method of reaching variate correlation is considerably less exact than the usual product-moment method. There is no gain in accuracy, but the reverse in using such a method in the case of short series.

Thirdly, the method proposed by Spearman of deducing the correlation of ranks from the positive differences of ranks is discussed, and the error of the process by which he has deduced for it an accuracy greater than that of the more usual methods of finding correlation is indicated. A method for deducing variate correlation from positive difference of ranks is indicated. The method is very rapid for short series, say those not exceeding 20 observations, but it is less accurate than the product-moment method, and considerable changes in the final value reached will be found to arise according as we use bracket-ranks or mid-ranks in the case of ties. The comparison with true grades for a few special cases, does not enable us to say which is the better method; the deviations from normality sometimes appear to make one, sometimes the other, the closer to the true correlation.

In conclusion, I think, we may say that variate correlation found by ranks may prove to be a useful *auxiliary* method of dealing with correlation, when it is needful to give a rough answer to a problem in a brief time, or when the material itself is incapable of being accurately measured. In all such cases mean square of rank differences will be more accurate than mean positive rank difference. But both methods must be used with caution, and their easy application must not lead us to approve exaggerated statements as to their accuracy.



THE HISTORY OF THE REVOLUTIONARY WAR

The first part of the war was fought in the north, and the British were successful in capturing New York City and Philadelphia. The Continental Congress fled to Lancaster and York, and finally to Lancaster and York, and finally to Lancaster and York.

The British then moved south, and the Continental Army followed them. The Battle of the Clouds was fought on September 26, 1777, and the British were victorious. The Continental Army then moved to Lancaster and York, and finally to Lancaster and York.

The Battle of Red Bank was fought on December 19, 1777, and the British were victorious. The Continental Army then moved to Lancaster and York, and finally to Lancaster and York. The Battle of Red Bank was fought on December 19, 1777, and the British were victorious.







PUBLISHED BY DULAU AND Co., SOHO SQUARE, LONDON, W.

## DRAPERS' COMPANY RESEARCH MEMOIRS.

DEPARTMENT OF APPLIED MATHEMATICS, UNIVERSITY COLLEGE,  
UNIVERSITY OF LONDON.

These memoirs will be issued at short intervals.

### Technical Series.

- I. On a Theory of the Stresses in Crane and Coupling Hooks with Experimental Comparison with Existing Theory. By E. S. ANDREWS, B.Sc.Eng., assisted by KARL PEARSON, F.R.S. *Issued.* Price 3s.
- II. On some Disregarded Points in the Stability of Masonry Dams. By L. W. ATCHERLEY, assisted by KARL PEARSON, F.R.S. *Issued.* Price 3s. 6d.
- III. On the Graphics of Masonry Arches, with special reference to the Relative Strength of Two-pivoted, Three-pivoted and Built-in Metal Arches. By L. W. ATCHERLEY and KARL PEARSON, F.R.S. *Issued.* Price 6s.
- IV. On Torsional Vibrations in Axles and Shafting. By KARL PEARSON, F.R.S. *Issued.* Price 6s.
- V. An Experimental Study of the Stresses in Masonry Dams. By KARL PEARSON, F.R.S., and A. F. CAMPBELL POLLARD, assisted by C. W. WHEEN, B.Sc.Eng., and L. F. RICHARDSON, B.A. *Issued.* Price 7s.
- VI. On the Graphics of Masonry Structures. By N. G. DUNBAR, Chadwick Scholar, and KARL PEARSON, F.R.S. *Shortly.*

### Biometric Series.

- I. Mathematical Contributions to the Theory of Evolution.—XIII. On the Theory of Contingency and its Relation to Association and Normal Correlation. By KARL PEARSON, F.R.S. *Issued.* Price 4s.
- II. Mathematical Contributions to the Theory of Evolution.—XIV. On the Theory of Skew Correlation and Non-linear Regression. By KARL PEARSON, F.R.S. *Issued.* Price 5s.
- III. Mathematical Contributions to the Theory of Evolution.—XV. On the Mathematical Theory of Random Migration. By KARL PEARSON, F.R.S., with the assistance of JOHN BLAKEMAN, M.Sc. *Issued.* Price 5s.
- IV. Mathematical Contributions to the Theory of Evolution.—XVI. On Further Methods of Determining Correlation. By KARL PEARSON, F.R.S. *Issued.* Price 4s.
- V. Mathematical Contributions to the Theory of Evolution.—XVII. On Homotyposis in the Animal Kingdom. By E. WARREN, D.Sc., A. LEE, D.Sc., EDNA LEA-SMITH, MARION RADFORD and KARL PEARSON, F.R.S. *Shortly.*

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

## PUBLICATIONS OF THE FRANCIS GALTON LABORATORY FOR NATIONAL EUGENICS, UNIVERSITY OF LONDON. (Published by Dulau and Co.)

- 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., First Francis Galton Research Fellow in National Eugenics, and E. M. ELDERTON, Galton Research Scholar in National Eugenics. *Issued.* Price 4s.
- II. A First Study of the Statistics of Insanity and the Inheritance of the Insane Diathesis. By DAVID HERON, M.A., Second Galton Research Fellow. *At Press.*
- 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., First Galton Research Fellow in National Eugenics. *Issued.* Price 2s. 6d.
- IV. On the Degree of Resemblance of First Cousins. By ETHEL M. ELDERTON, Galton Research Scholar. *At Press.*

PUBLISHED BY THE CAMBRIDGE UNIVERSITY PRESS.

### BIOMETRIKA.

A JOURNAL FOR THE STATISTICAL STUDY OF BIOLOGICAL PROBLEMS.

Founded by W. F. R. WELDON, FRANCIS GALTON and KARL PEARSON.

Edited by KARL PEARSON in Consultation with FRANCIS GALTON.

VOL. V., PART III.

- I. A Biometrical Study of Conjugation in *Paramecium*. (With eleven Diagrams in the text.) By RAYMOND PEARL, Ph.D.
- II. The Anthropometric Characteristics of the Inmates of Asylums in Scotland. (With eleven Diagrams and eight Maps in the text and three Plates of Maps.) By J. F. TOCHER.
- III. On the Error of Counting with a Haemocytometer. By Student. (With two Diagrams in the text.)
- Miscellanea: (i) On the Distribution of Severity of Attack in Cases of Smallpox. By F. M. TURNER, M.D. (ii) Remarks on Dr Turner's Note. By KARL PEARSON, F.R.S.
- Supplement to Vol. V. Anthropometric Survey of the Inmates of Asylums in Scotland. (With Maps.) By J. F. TOCHER. Issued by permission of the Henderson Trustees.

VOL. V., PART IV.

- I. Statistical Observations on Wasps and Bees. By F. Y. EDGEWORTH.
- II. Natural Selection in *Helix Arbutorum*. By A. P. DI CESNOLA.
- III. Grades and Deviates. By FRANCIS GALTON, with a Table of Deviates by W. F. SHEPPARD.
- IV. A Cooperative Study of Queens, Drones and Workers in *Vespa Vulgaris*. By A. WRIGHT, A. LEE and KARL PEARSON, F.R.S.
- V. Statistical Studies in Immunity. A Discussion of the Means of estimating the severity of cases of Acute Disease. By JOHN BROWNLEE, M.D., D.Sc.
- VI. On Heredity in Mice from the Records of the late W. F. R. WELDON. Part I. On the Inheritance of the Sex-Ratio and of the Size of Litter.
- VII. The Calculation of the Moments of a Frequency-Distribution. By W. F. SHEPPARD.
- Miscellanea: (i) On the Inheritance of Psychological Characters. Being further Statistical Treatment of Material Collected and Analysed by Messrs Heymans and Wiersma. By EDGAR SCHUSTER, M.A., and ETHEL M. ELDERTON. (ii) Reply to certain criticisms of Mr G. U. Yule. (With coloured Plate XXIII.) By KARL PEARSON, F.R.S.

Notices and Bibliography.

The subscription price, payable in advance, is 30s. net per volume (post free); single numbers 10s. net. Volumes I, II, III, IV. and V. (1902-7) complete, 30s. net per volume. Bound in Buckram 34s. 6d. net per volume. Subscriptions may be sent to C. F. CLAY, Manager, Cambridge University Press Warehouse, Fetter Lane, London, E.C., either direct or through any bookseller.















