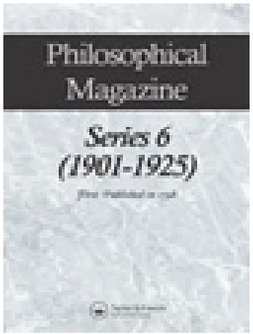

Taylor & Francis
Taylor & Francis Group

Philosophical Magazine Series 6

ISSN: 1941-5982 (Print) 1941-5990 (Online) Journal homepage: http://www.tandfonline.com/loi/tphm17

XXXIX. The application of solid hypergeometrical series to frequency distributions in space

S.D. Wicksell Dr. Phil.

To cite this article: S.D. Wicksell Dr. Phil. (1917) XXXIX. The application of solid hypergeometrical series to frequency distributions in space , Philosophical Magazine Series 6, 33:197, 389-394, DOI: 10.1080/14786440508635654

To link to this article: http://dx.doi.org/10.1080/14786440508635654

For the present purpose we need only to introduce $\delta\beta$, and with sufficient accuracy we may take

$$\delta(U^2 - 2gy) = 2\delta\beta \cos 2x. \quad . \quad . \quad . \quad . \quad (26)$$

We suppose $\delta\beta = -\cdot 000{,}000{,}2$, so that the new value of β is $-\cdot 000{,}052{,}6$. Introducing corrections according to (26) and writing only the last two figures, we obtain column 5 of Table I., in which the greatest discrepancy is reduced from 10 to 4—almost as far as the arithmetic allows—and becomes but one-millionth of the statical difference between crest and trough. This is the degree of accuracy attained when we take simply

$$\psi = y - \alpha e^{-y} \cos x - \beta e^{-2y} \cos 2x - \gamma e^{-3y} \cos 3x, \quad . \quad (27)$$

with $\alpha = \frac{1}{10}$, g and γ determined by Stokes' method, and β determined so as to give the best agreement.

XXXIX. *The Application of Solid Hypergeometrical Series to Frequency Distributions in Space.* *By* S. D. Wicksell, *Dr. Phil., Lund, Sweden* *.

In the number of this Journal issued in September 1914, Dr. L. Isserlis, under the above title, published a paper on the fitting of hypergeometrical series to correlation surfaces. The problem to describe curves of variation by aid of hypergeometrical series was treated as long ago as 1895 by Prof. Pearson in his classical Memoir: "Skew Variation in Homogeneous Material," Phil. Trans. vol. clxxxvi. Later, in 1899, Prof. Pearson gave a fuller discussion of the hypergeometrical series (Phil. Mag. vol. xlvii.). It is this paper that is the starting-point and chief place of reference of Dr. Isserlis. On the whole, the hypergeometrical series and its special case for $n = \infty$, the binomial series, play a dominating part in Prof. Pearson's celebrated theory of variation of one variate. As a consequence hereof, it was natural that the attempt should be made to employ solid hypergeometrical series as a means to describe also surfaces of correlation. Hereby, however, a fact has evidently been overlooked that greatly limits the range of applicability of any hypergeometrical or multinomial types of correlation functions. Of course, there must be some identical relations between the moments that should be more or less fulfilled in all cases of application. Dr. Isserlis also produces several such relations. But it is evident that he has not ascribed too much importance to these limitations. In the theory of variation of one variate there are similar conditions, but they

* Communicated by the Author.

have proved to be of only little harm to the generality of the frequency-curves generated. Now, the writer of these lines has reason to think that the limitations in case of the application to correlation surfaces is of far greater importance. The fact which is the ground for the opinion of the writer is the following: *All surfaces of correlation described by aid of the multinomial or hypergeometrical series must necessarily have linear regression.* With regard to the multinomial correlation function, which is the coefficient of $x^s y^{s'}$ in the development of $(p_1xy+p_2x+p_3y+p_4)^r$, the fact has already been demonstrated in my paper in *Svenska Aktuarieföreningens Tidskrift*, Nr. 4–5, 1916 (see also *Meddelanden från Lunds Astronomiska Observatorium*, 1916).

In order to prove our proposition also in case of the solid hypergeometrical series, we must first recall the formulation of the corresponding chance problem. In Dr. Isserlis' own words it is: a bag contains n balls of which np are white and nq are black; r balls are drawn and not replaced; a second draw of r' balls is made. This is repeated N times. If N is a large number, the theoretical frequency of s black balls in the first draw and s' in the second is (in a somewhat different notation)

$$\frac{\mathrm{N}\,.\,r\,!\,r'\,!\,(qn)\,!\,(pn)\,!\,(n-r-r')\,!}{s\,!\,s'\,!\,(r-s)\,!\,(r'-s')\,!\,(qn-s-s')\,!\,(pn-r-r'+s+s')\,!\,n\,!}.$$

Calling this function $z(s, s')$, the moments p'_{ij} may be deduced either, as does Dr. Isserlis, with the aid of a system of differential equations or more directly by performing the summations

$$p'_{ij}=\sum_{s,\,s'} z(s,\,s')\,s^i s'^j,$$

by which method recursion formulæ for the moments are easily derived. Especially it will readily be found that the mean values of s and s' are

$$p'_{10}=rq\,;\ p'_{01}=r'q.$$

Thus it is seen that the means of the number of "lucky" events are equal to the product of the number of trials and the probability of a "lucky" event at the *beginning of the drawings*. Now, it is an easy task to show that the regression is linear. Indeed, if from the N sets of drawings we pick out all that have given a certain number, say $s=\mathrm{S}$ black balls in the first r trials, they will all have that in common, that the second set of r' drawings has been extracted from a bag that contained $n-r$ balls, of which only $qn-\mathrm{S}$ were black. Hence, for these samples, if N be a large enough

number, the mean of the number of black balls in the second sets of trials will be

$$s'_S = \frac{r'(nq-S)}{n-r} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

Thus the mean value of s' for a given value of s is a linear function of s, by which the linearity of regression is proved.

It will be of some interest to see how the moments of the hypergeometrical series, as found by Dr. Isserlis, are consistent with the general conditions for linear regression, and especially how these conditions are even contained in the identical relations between the moments as far as such have been deduced. First, we must, however, find the relations between the moments that are the necessary conditions of linearity of regression. Denoting by ξ and η the deviations from the means $s-rq$ and $s'-r'q$, we must (as from the above regression formula it is obvious that when $S=rq$ we have $s'_S=r'q$) give to the equation of the regression line of, for instance, s' on s, the form

$$\eta_\xi = b\xi. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Denoting the coefficient of correlation by ρ and the standard deviations of s and s' by σ and σ', we necessarily have

$$b = \rho \frac{\sigma'}{\sigma} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

The truth of formula (3) is well known and may be demonstrated in the following way. Multiplying (2) with

$$\xi \sum_\eta z(\xi+rq, \eta+r'q)$$

and summing for all values of the variable ξ, we obtain

$$\sum_\xi \xi\eta_\xi \sum_\eta z(\xi+rq, \eta+r'q) = b\sum_\xi \xi^2 \sum_\eta z(\xi+rq, \eta+r'q). \quad (4)$$

According to the signification of η_ξ as a mean of η for constant ξ, this may be written

$$\sum_\xi \sum_\eta \xi\eta z(\xi+rq, \eta+r'q) = b\sum_\xi \sum_\eta \xi^2 z(\xi+rq, \eta+r'q)$$

or, denoting by p_{ij} the moments about the mean of $z(s, s')$,

$$p_{11} = bp_{20}.$$

As $\rho = \dfrac{p_{11}}{\sqrt{p_{20}p_{02}}}$, we have thus proved the truth of (3).

The equation of the regression line is hence

$$\eta_\xi = \rho\frac{\sigma'}{\sigma}\xi. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

In order to derive the general relations between the moments that are a consequence of the linearity of regression we proceed thus: Multiplying (5) by

$$\xi^2 \sum_{\eta} z(\xi + rq,\ \eta + r'q),$$

we find, summing for all values of ξ,

$$p_{21} = \rho \frac{\sigma'}{\sigma} p_{30} \quad \text{or} \quad \rho_{21}\rho_{20} = p_{30}p_{11}. \quad . \quad . \quad . \quad . \quad (6)$$

This formula has been deduced by Pearson in his well-known memoir on the skew regression.

Multiplying further by

$$\xi^3 \sum_{\eta} z(\xi + rq,\ \eta + r'q)$$

and summing, we obtain

$$p_{31}p_{20} = p_{10}p_{11}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

and similarly proceeding for higher powers of ξ and having recourse also to the equation of the other regression line, we have as conditions of linear regression,

$$\begin{aligned} p_{\alpha,1}p_{20} &= p_{\alpha+1,0}p_{11}, \\ p_{1,\beta}p_{12} &= p_{0,\beta+1}p_{11}. \end{aligned} \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

Dr. Isserlis has not deduced the moments p_{31} and p_{13}, so we are not in a position to test his formulæ on linearity of regression otherwise than in case of the moments of the third order. In case of these moments Dr. Isserlis has found the identical relations

$$\begin{aligned} p_{21}\,p_{20}\,p_{03} &\equiv p_{12}\,p_{02}\,p_{30}, \\ p_{02}\,p_{20}\,p_{21}\,p_{12} &\equiv p_{11}^2\,p_{03}\,p_{30}. \end{aligned}$$

Dividing these relations, we find

$$\begin{aligned} p_{21}^2\,p_{20}^2 &= p_{30}^2\,p_{11}^2, \\ p_{12}^2\,p_{02}^2 &= p_{03}^2\,p_{11}^2. \end{aligned}$$

Taking regard of the fact that according to the results of Dr. Isserlis, the moments p_{20} and p_{11} as well as the moments p_{30} and p_{21} have inverse signs, we see that the identities contain in them the conditions of linear regression

$$\begin{aligned} p_{21}\,p_{20} &= p_{30}\,p_{11}, \\ p_{12}\,p_{02} &= p_{03}\,p_{11}. \end{aligned}$$

Obviously, as the regression has already been shown to be strictly linear. it should be found on deducing the moments of higher order than the third that the hypergeometrical series is subject to the general identities

$$p_{\alpha,1}p_{20}=p_{\alpha+1,0}p_{11},$$
$$p_{1,\beta}p_{02}=p_{0,\beta+1}p_{11}.$$

Of course, these are not the only relations possible to find, also relations between the pure marginal moments are at hand.

The application of solid hypergeometrical series to correlation surfaces must, as we have shown, be confined to cases of strictly linear regression. That this has not been observed by Isserlis is evident, as otherwise he would have mentioned it, or, at least, he would not have attempted to apply the series to a case of decidedly curvilinear regression, as in the example of the correlation of the ages of bachelors and spinsters at the epoch of marriage. As regards the example of the numbers of trumps in whist, the regression is linear, but there is an error in the computation of the correlation coefficient, which is $-0{\cdot}3305$, not $-0{\cdot}2559$.

Note I.—The chance problem that gives rise to the above-mentioned multinomial series is the following: A bag contains n balls. The balls are either white or black, besides being marked by either an even or an odd number. Of the balls np_1 are black and even, np_2 are black and odd, np_3 are white and even, and np_4 are white and odd ; r balls are drawn and *each ball is replaced after drawing.* This is repeated N times. If N is a large number, the theoretical frequency of sets with s black balls and s' even balls is the coefficient of $x^s y^{s'}$ in the development of

$$(p_1xy+p_2x+p_3y+p_4)r.$$

The moments of this series are deduced in my memoir in the *Meddelanden från Lunds Astronomiska Observatorium* cited above. The regression is strictly linear. If the balls are *not* replaced there arises a series in which the terms are certain sums of the terms of a hypergeometrical series in three dimensions. Hereby the regression will still be strictly linear, as the even balls in samples of $s=\mathrm{S}$ black balls come forth as if they had been drawn in S trials from a bag containing all the black balls and in $r-\mathrm{S}$ trials from a bag containing all the white balls of the initial bag. The mean of the number of even balls in samples of S black balls will

then be the sum of two quantities, of which the one is proportional to S and the other to $r-S$, thus being a linear function of S.

Note II.—We have shown above that the condition for linear regression is that the following relations between the moments are valid :

$$p_{\alpha,1}p_{20}=p_{\alpha+1,0}p_{11},$$

$$p_{1,\beta}p_{02}=p_{0,\beta+1}p_{11}.$$

Introducing the notations

$$\Sigma_{ij}=\frac{p_{ij}}{\sigma^i\sigma^{ij}},$$

we may write the conditions thus

$$\alpha!\,\rho_{\alpha+1,0}=\Sigma_{\alpha,1}-\rho\,\Sigma_{\alpha+1,0}=0,$$

$$\beta!\,\rho_{0,\beta+1}=\Sigma_{1,\beta}-\rho\,\Sigma_{0,\beta+1}=0.$$

When the regression is not linear the quantities $\rho_{i,0}$ and $\rho_{0,j}$, or some of them, will not disappear. In another place I shall soon demonstrate that the equations to the curves of regression, when the correlation is only moderately skew, may be expressed in a very convenient form with the aid of the coefficients $\rho_{i,0}$ and $\rho_{0,j}$.

By Pearson's definition there is no correlation when the regression is linear and parallel to the axes. Though this definition seems to me to be not quite sufficient, as it does not necessarily coincide with the definition required from the standpoint of the theory of probability, *i. e.* that the variates should be separated in the correlation function, it is any way the best one to have recourse to when we have no adequate correlation function available. In the sense of Pearson's definition the variates will be independent of each other if all the coefficients ρ, ρ_{30}, ρ_{03}, ρ_{40}, ρ_{04}, &c. are zero. ρ is the usual coefficient of correlation; as the quantities ρ_{30}, ρ_{03}, ρ_{40}, ρ_{04}, &c., are abstract numbers, independent of any units, I propose that they be called the coefficients of correlation of higher order. The numerical factors are inserted for purposes of which I hope soon to give the explanation.