The Poisson Correlation Function

By J. T. Campbell, Edinburgh University.

(*Received* 18*th July*, 1933. *Read* 4*th November*, 1933.)

§ 1. *Introduction.*

In this study, the use of factorial moments and factorial moment generating functions as applied (2) to the Poisson frequency function and Charlier's Type B function is further extended towards developing a theory of these distributions for the case of two or more correlated variables.

The results obtained bear a close analogy to the known ones for the correlated normal function and Type A function, though there does not appear to be a simple closed form for the multiple Poisson function of correlated variables compared with that of the normal frequency function of several linearly correlated variables.

In the last section a numerical example is given to verify the theoretical formulae of double Poisson correlation. As no statistical records of correlation of two variables, each following a Poisson law, could be discovered, an artificial experiment, suggested by Darbishire's experiments[1] in throws of dice, was devised to provide the necessary data.

§ 2. *The Poisson Correlation Function.*

Consider first the case of the "fourfold table." Let the following fourfold table display the frequency in a population of two characters A and B which are not independent:

	$\bar{A}$	A	
$\bar{B}$	a	b	$\bar{p}_2$
B	c	d	p_2
	$\bar{p}_1$	p_2	1

[1] Whittaker and Robinson, *Calculus of Observations*, p. 330.

https://doi.org/10.1017/S0013091500024135 Published online by Cambridge University Press

Here p_1 and p_2 are the relative frequency (probability) of A and B respectively, $\bar{p}_1$ and $\bar{p}_2$ that of the absence of thes echaracters. Since A and B are not independent the relative frequency d of both together will be different from $p_1 p_2$. The frequency generating function (5) for this population will be

$$a + bA + cB + dAB = 1 + p_1(A - 1) + p_2(B - 1) + d(A - 1)(B - 1).$$

Putting $A = 1 + \alpha$, $B = 1 + \beta$, we have the factorial moment generating function (m. g. f.) of the population,

$$1 + p_1 \alpha + p_2 \beta + d\alpha\beta,$$

and so that of samples of N drawn from it, with replacement of individuals if the population is finite, is the N^{th} power of the above.

The case of interest to us is when p_1, p_2 and d are all $O(N^{-1})$. Wc then have

$$(1 + p_1\alpha + p_2\beta + d\alpha\beta)^N = [(1 + p_1\alpha)(1 + p_2\beta)\{1 + \overline{d - p_1 p_2}\alpha\beta + O(N^{-2})\}]^N$$

which tends to

$$e^{m_1\alpha + m_2\beta + \bar{m}\alpha\beta}, \tag{2.1}$$

where $\qquad Np_1 = m_1, \; Np_2 = m_2, \; N(d - p_1 p_2) = \bar{m}.$

To find the corresponding frequency function $\phi(x, y)$ we have to solve the sum equation

$$e^{m_1\alpha + m_2\beta + \bar{m}\alpha\beta} = \sum_x \sum_y \phi(x, y)(1 + \alpha)^x (1 + \beta)^y.$$

Now

$$m_1\alpha + m_2\beta + \bar{m}\alpha\beta = (m_1 - \bar{m})(\alpha + 1) + (m_2 - \bar{m})(\beta + 1) \\ + \bar{m}(\alpha + 1)(\beta + 1) + \bar{m} - (m_1 + m_2),$$

and so, picking out the coefficient of $(1 + \alpha)^x (1 + \beta)^y$ on the left hand side of the m. g. f. just found, we have the frequency function

$$\phi(x, y) = e^{-(m_1 + m_2 - \bar{m})} \sum_{s=0} \frac{(m_1 - \bar{m})^{x-s}}{(x - s)!} \frac{(m_2 - \bar{m})^{y-s}}{(y - s)!} \frac{\bar{m}^s}{s!}, \tag{2.2}$$

where the upper limit of the summation is the lesser of x and y.

This series is identical with that obtained by A. G. McKendrick (4) by a quite different approach, the assumptions of which were that the x and y series were distributed according to the Poisson law, so that the variances were equal to the means.

https://doi.org/10.1017/S0013091500024135 Published online by Cambridge University Press

§ 3. *A Series Form in Products of Orthogonal Polynomials.*

The form (2.2) of the frequency function is not very convenient for most purposes, and so we proceed to derive another, analogous to that form of the normal correlation function which involves a power series in r, the correlation coefficient.

If (2.1) be expanded as a power series in $\bar{m}$, we have

$$\sum_{s=0}^{\infty} \alpha^s e^{m_1 \alpha} . \beta^s e^{m_2 \beta} . \bar{m}^s / s!,$$

which (2) can be written

$$\sum_{s=0}^{\infty} \{\sum_{x} (-)^s K_s(x) \psi(x) (1+\alpha)^x\}\{\sum_{y} (-)^s K_s(y) \psi(y) (1+\beta)^y\} \frac{\bar{m}^s}{s!},$$

where $\psi(x) = e^{-m_1} m_1^x / x!$, $\psi(y) = e^{-m_2} m_2^y / y!$, and $K_s(x)$, $K_s(y)$ are the orthogonal polynomials (1) appropriate to the Poisson frequency function. The theorem used here is (2, § 1) that multiplication of a factorial m. g. f. by α^r is equivalent to the operation $(-\nabla)^r$ on the frequency function, where $\nabla f(x) = f(x) - f(x-1)$.

Thus finally, assuming that the order of summation may be interchanged, we have

$$\phi(x, y) = \frac{e^{-m_1} m_1^x}{x!} . \frac{e^{-m_2} m_2^y}{y!} \left\{1 + K_1(x) K_1(y) \frac{\bar{m}}{1!} + K_2(x) K_2(y) \frac{\bar{m}^2}{2!} + \, . \, . \right\}, \qquad (3.1)$$

which could also be written in operational form

$$\phi(x, y) = e^{\bar{m} \nabla_x \nabla_y} \{e^{-m_1} m_1^x . e^{-m_2} m_2^y / (x! \, y!)\}.$$

§ 4. *The Factorial Moments of the Distribution.*

Let $m_{(r,s)}$ be the factorial moment of order r in x, s in y, defined by

$$m_{(r,s)} = \sum_x \sum_y \phi(x, y) \, x(x-1) \ldots . (x-r+1) \, y(y-1) \ldots . (y-s+1).$$

Since it is the coefficient of $\alpha^r \beta^s / (r! \, s!)$ in (2.1), we have

$$m_{(r,s)} = r! \, s! \sum_{t=0} \frac{m_1^{r-t}}{(r-t)!} . \frac{m_2^{s-t}}{(s-t)!} . \frac{\bar{m}^{-t}}{t!}, \qquad (4.1)$$

where the upper limit for t is the lesser of r and s. In particular

$$m_{(1,0)} = m_1, \quad m_{(0,1)} = m_2, \quad m_{(1,1)} = \bar{m}.$$

These three moments can be computed directly from any set of data, and so the three parameters of the frequency function are completely determined. Also, further factorial moments of the fitted distribution can be obtained by summing the necessary terms of (4.1).

It is easy to prove, by summing (2.2) over all values of y, that the distribution of x alone is an ordinary Poisson distribution with mean m_1, with a similar result for the distribution of y. Thus the totals of the x and y arrays in the correlation table are distributed according to the Poisson exponential law.

§ 5. *Regression Lines.*

To find the mean x corresponding to a fixed $y = k$, which we shall denote by $\widehat{x}_k$, we have, by (2.1),

$$\sum_x x\phi(x, k) = e^{-m_2}\left\{\frac{(m_1-\bar{m})(m_2-\bar{m})^k}{k!} + \frac{(m_1-\bar{m}+1)(m_2-\bar{m})^{k-1}}{(k-1)!}\cdot\frac{\bar{m}}{1!} + \ldots. + \frac{(m_1-\bar{m}+k)\bar{m}^k}{k!}\right\}$$

$$= e^{-m_2}\left\{\frac{(m_1-\bar{m})\,m_2^k}{k!} + \frac{\bar{m}\,.\,m_2^{k-1}}{(k-1)!}\right\},$$

and $\qquad \Sigma\,\phi(x, k) = e^{-m_2}m_2^k/k!,$

so that $\qquad \widehat{x}_k = \Sigma\, x\phi(x, k)/\Sigma\,\phi(x, k) = m_1 - \bar{m} + \dfrac{\bar{m}}{m_2}k.$

Hence the locus of the means of the x's corresponding to any y is

$$x - m_1 = \frac{\bar{m}}{m_2}(y - m_2),$$

with a similar expression for the locus of the means of the y's corresponding to any x. These straight lines may be regarded as the "regression lines" of the distribution, though they do not, as in the case of ordinary normal correlation, give the most probable value of one variable corresponding to any value of another; they give the *mean* value.

We may further extend the analogy and define a coefficient of correlation r by

$$r = \sqrt{\left(\frac{\bar{m}}{m_1}\cdot\frac{\bar{m}}{m_2}\right)} = \frac{\bar{m}}{\sqrt{(m_1 m_2)}} = \frac{m_{(1,\,1)} - m_{(0,\,1)}\,m_{(1,\,0)}}{\sqrt{(m_{(0,\,1)}\,m_{(1,\,0)})}}.$$

https://doi.org/10.1017/S0013091500024135 Published online by Cambridge University Press

Since the totals of the x and y arrays are Poisson distributions, the variances of these are equal to their means; so that the above result may be written

$$r = \frac{\mu_{1,\,1}}{\sqrt{(\mu_{0,\,2}\,\mu_{2,\,0})}},$$

where $\mu_{1,1}$ is the ordinary product moment, $\mu_{2,0}$ and $\mu_{0,2}$ the variances of the x and y arrays, all three being calculated from the means. This is simply the ordinary Pearsonian coefficient of correlation.

The definition may be justified in another way. In the case of Poisson correlation the factor expressing the correlation involves orthogonal polynomials $K_s(x)$, $K_s(y)$, which contain the parameters m_1 and m_2. In the corresponding series for normal correlation, as usually given, the variables are *normalized*, so that r is given absolutely, free from scale units. The true analogue of our series will therefore be the normal series, when the Hermite polynomials that occur in it are defined by *non-normalized* variables. In such a case we must have

$$H_s(x,\,\sigma_x) = e^{\frac{1}{2}x^2/\sigma_x^2}\,(-)^s\left(\frac{d}{dx}\right)^s e^{-\frac{1}{2}x^2/\sigma_x^2}$$

$$= \sigma_x^{-s}\,H_s(x/\sigma_x),$$

and the series for the normal correlation function then takes the shape

$$1 + r\sigma_x\sigma_y H_1(x,\,\sigma_x)\,H_1(y,\,\sigma_y) + \frac{r^2\sigma_x^2\,\sigma_y^2}{2!}\,H_2(x,\,\sigma_x)\,H_2(y,\,\sigma_y) + \ldots.,$$

with which our form for the Poisson correlation function,

$$1 + \bar{m}\,K_1(x,\,m_1)\,K_2(x,\,m_2) + \frac{\bar{m}^2}{2!}\,K_2(x,\,m_1)\,K_2(x,\,m_2) + \ldots.,$$

is now in satisfactory correspondence.

From the nature of the Poisson function and its orthogonal polynomials, we are compelled to leave the parameters m_1 and m_2 implicit, being unable to remove them by normalizing the variables, since not powers but factorials in x and y are involved. But we may normalize $\bar{m}$ in terms of m_1 and m_2, thus arriving at a proper analogy and at the definition of the correlation coefficient given above. It is of course known that the Pearsonian coefficient can be applied to distributions other than the normal; what is not always known in such cases is its sampling distribution.

§ 6. *The Correlation Function of Type B.*

The distribution given by (2.2) can be derived under more general conditions. We shall now consider the samples of N to be formed by taking one individual from each of N different populations constituted similarly to the population taken in § 2, that is, consisting of two correlated characters. If the parameters of the populations are p_1, p'_1, d_1; p_2, p'_2, d_2; .... ; p_N, p'_N, d_N, then the factorial m. g. f. of the samples of N, since the selections are from independent populations, is

$$\prod_{i}^{N} (1 + p_i \alpha + p'_i \beta + d_i \alpha\beta) = \Pi [(1 + p_i \alpha)(1 + p'_i \beta) \{1 + d_i - p_i p'_i \alpha\beta + O(N^{-2})\}]$$
$$= e^{m_1\alpha + m_2\beta + \bar{m}\alpha\beta}\{1 + B_{2,0}\,\alpha^2 + B_{0,2}\,\beta^2 + B_{3,0}\,\alpha^3 + B_{2,1}\,\alpha^2\beta + B_{1,2}\alpha\beta^2 + B_{0,3}\beta^3 + ..\}, \quad (6.1)$$

where $m_1 = \Sigma p_i$, $m_2 = \Sigma p'_i$, $\bar{m} = \Sigma (d_i - p_i p'_i)$.

Here $B_{2,0}$ and $B_{0,2}$ are $O(N^{-1})$, later terms being of smaller order. Hence in the limit we have the same factorial m. g. f. as in § 2, and therefore the same kind of frequency function $\phi(x, y)$.

If however we do not proceed to the limit but retain the terms of higher order, then by the operational theorem mentioned in § 3 we have, from (6.1), the more general correlation function,

$$f(x, y) = \phi(x, y) + B_{2,0}\,\nabla_x^2 \phi(x, y) + B_{0,2}\,\nabla_y^2 \phi(x, y) + \ldots., \quad (6.2)$$

where $\phi(x, y)$ has the form given in (2.2). The function $f(x, y)$ may be called the correlation function of Type B, and may be compared with the correlation function (3) of Type A,

$$\phi(x, y) + \sum_{k, r} A_{k, r} \frac{\partial^{k+r}}{\partial x^k \, \partial y^r} \phi(x, y),$$

where $k + r \geqq 3$, and $\phi(x, y)$ is the normal correlation function for two variables.

The parameters m_1, m_2 and $\bar{m}$ and the B's can be expressed directly (2) in terms of the factorial moments of the distribution.

Expansions similar to those for more than two normally correlated variables can be developed by the method of the present paper, but the analogies are then less marked.

§ 7. *Experimental Tests and Numerical Results.*

The data for these were obtained in the following way. A stock of 420 marbles, of which 21 differed from the rest in colour only, was subjected to thorough mixing and then distributed at random in 20

https://doi.org/10.1017/S0013091500024135 Published online by Cambridge University Press

grooves capable of holding 20 marbles each exactly. (An excess of 20 over the 400 was taken in order that the mean number of coloured marbles in a groove should not be *forced* to be unity.) The number of coloured marbles in each groove was then noted. The 420 marbles were then mixed again and redistributed at random. This was carried out 50 times in all, and so a frequency distribution of 1000 observations was obtained.

The successive frequencies recorded being denoted by u_1, u_2, u_3, ..., the sequence $(u_1 + u_2)$, $(u_2 + u_3)$, $(u_3 + u_4)$, ... was then constructed, and the correlated pairs $(u_1 + u_2, u_2 + u_3)$, $(u_3 + u_4{:}\ u_4 + u_5)$, ... were taken to be the values (x_i, y_i) of the double distribution.

The single frequencies u_1, u_2, ... are uncorrelated, and can be regarded as distributed, to a sufficient order of approximation, according to the Poisson law, with mean unity and so with variance unity. (Actually the distribution, obtained by sampling without replacement, is of hypergeometric type, but since the probability of success is small, $p = 0{\cdot}05$, the Poisson law is a close approximation.) Hence the variance of the x-series alone is $\sigma_x^2 = 2$, since the series is composed of sums of two independent terms each of variance unity; similarly $\sigma_y^2 = 2$. Hence the expected value of the correlation coefficient r, as given by

$$\begin{aligned} r\sigma_x \sigma_y &= \Sigma\,(u_i + u_j)\,(u_j + u_k)/N \\ &= \Sigma\, u_j^2/N = 1, \end{aligned}$$

(since u_i, u_j, u_k are uncorrelated), is $r = 0{\cdot}5$, as might be expected.

The experimental results based on these pairs of terms gave the following correlation table.

	y 0	1	2	3	4	5	6	7	f_x
x 0	24	19	12	7	2				64
1	28	39	45	17	2				131
2	12	41	47	32	3	3	2		140
3	6	23	29	21	11	6		1	97
4		1	12	14	13	4	2	2	48
5	1		3	1	1	2	4		12
6			1		2	2			5
7					1	1	1		3
f_y	71	123	149	92	35	18	9	3	500

https://doi.org/10.1017/S0013091500024135 Published online by Cambridge University Press

In the usual way we obtain from this the values of the required parameters:

$$m_1 = m_x = 2{\cdot}01, \qquad \sigma_x = 1{\cdot}37,$$

$$m_2 = m_y = 2{\cdot}00, \qquad \sigma_y = 1{\cdot}43, \qquad \bar{m} = 1{\cdot}02.$$

Hence $r = 0{\cdot}51$.

The next table shows the theoretical *uncorrelated* distribution of 500 observations having the totals of rows and columns distributed according to a Poisson law with $m_x = 2$, $m_y = 2$, the frequency in class (x, y) being thus given by

$$500\, e^{-2}\, 2^x \,.\, e^{-2}\, 2^y / (x!\, y!).$$

	y									
	0	1	2	3	4	5	6	7	8	
x 0	9·2	18·3	18·3	12·2	6·1	2·5	0·8	0·2	0·1	67·7
1	18·3	36·6	36·6	24·4	12·2	4·9	1·6	0·5	0·1	135·2
2	18·3	36·6	36·6	24·4	12·2	4·9	1·6	0·5	0·1	135·2
3	12·2	24·4	24·4	16·3	8·1	3·3	1·1	0·3	0·1	90·2
4	6·1	12·2	12·2	8·1	4·1	1·6	0·5	0·2		45·0
5	2·5	4·9	4·9	3·3	1·6	0·7	0·2	0·1		18·2
6	0·8	1·6	1·6	1·1	0·5	0·2	0·1			5·9
7	0·2	0·5	0·5	0·3	0·2	0·1				1·8
8	0·1	0·1	0·1	0·1						0·4
										499·6

In order to obtain the theoretical correlation distribution, the frequencies in the cells of the above table must be multiplied by the proper value of the series in orthogonal polynomials of (3 . 1), with $\bar{m} = 1$. For this purpose we require the values of these polynomials for $m = 2$, as far as the 6th order. Tables of the polynomials have been constructed (2), depending on the recurrence relation (1)

$$K_r(x+1) - K_r(x) = \frac{r}{m} K_{r-1}(x),$$

a partial control on the computations being provided by the relation

$$K_r(r+1) = -K_{r+1}(r).$$

In this way we obtain the theoretical correlated frequencies of the following table, which is to be compared with that first given:

	y 0	1	2	3	4	5	6	7	8	
x 0	24·9	24·9	12·4	4·2	1·0	0·2				67·6
1	24·9	49·7	37·4	16·5	5·2	1·3	0·2	0·1		135·3
2	12·4	37·4	43·3	27·1	10·9	3·2	0·9	0·1		135·2
3	4·2	16·5	27·1	23·4	12·6	4·7	1·3	0·3	0·1	90·2
4	1·0	5·2	10·9	12·6	9·0	4·4	1·5	0·4	0·1	45·1
5	0·2	1·3	3·2	4·7	4·4	2·6	1·2	0·4	0·1	18·1
6		0·2	0·9	1·3	1·5	1·2	0·6	0·2		5·9
7		0·1	0·1	0·3	0·4	0·4	0·6	0·2		1·7
8				0·1	0·1	0·1				0·3
										499·4

As a measure of the agreement Pearson's χ^2-test of goodness of fit may be applied, but since the test is not valid for too fine a grouping, it was decided to pool the frequencies for x and $y \geqq 3$. When this is done we have 16 frequency classes with 15 degrees of freedom, since the total frequencies of the tables have been made to agree by the process of fitting. For the correlation table of the experimental results and the theoretical correlation table just given χ^2 proves to be 10·32, for which the value of P, the probability of obtaining as great or greater values of χ^2, is 0·80. This can be considered a reasonable value. On the other hand, applying a similar test to the theoretical uncorrelated table, we find $\chi^2 = 105{\cdot}6$, for which the value of P is of extreme smallness.

We conclude from the experiment that correlation of the type discussed in this paper is present, and that the mathematical representation of it by the bracket factor involving the orthogonal polynomials of Type B is an adequate one.

REFERENCES.

1. A. C. Aitken. *Proc. Roy. Soc. Edin.*, **52** (1932), 174-182.

2. J. T. Campbell. *Proc. Edin. Math. Soc.* (2), **3** (1932), 99-106. Also, for tables of the orthogonal polynomials, and other details, a Thesis, deposited in the Library of the University of Edinburgh, 1932.

3. C. V. L. Charlier. *Arkiv för Mat., Astr. och Fys.* **9** (1914), 26.

4. A. G. M^c^Kendrick. *Proc. Edin. Math. Soc.*, **44** (1926), 106.

5. H. E. Soper. *Frequency Arrays*, Cambridge, 1922.