THE STRUCTURE OF BIVARIATE DISTRIBUTIONS

By. H. O. Lancaster

School of Public Health and Tropical Medicine, Sydney, Australia

1. Introduction. K. Pearson [18] in his study on the association between two chance variables defined a measure, the mean square contingency, $\phi^2 = \chi^2/N$, where χ^2 is that, usually calculated in a contingency table with fixed marginal totals, and N is the size of the sample. In a bivariate joint normal distribution with coefficient of correlation, ρ, Pearson showed that ϕ^2 would have a limiting value if the sample size became indefinitely large, while the subdivisions of the marginal distributions were made increasingly fine. In effect, he was considering a property of the parent joint normal distribution, rather than of a sample drawn from it. He noted that this limiting ϕ^2 was independent of the scale of the marginal variables and was invariant under any bi-unique transformations of the marginal variables of the form, $x \to x'(x)$, $y \to y'(y)$. If the distribution was the bivariate joint normal, he showed that $\rho^2 = \phi^2/(1 + \phi^2)$. In some distributions, jointly normal with appropriate choice of the marginal variable, but not so with the variables actually chosen, he took the value of ρ^2 still to have the meaning that an appropriate transformation would yield the variables of the underlying joint normal distribution.

Hirshfeld [8], considering contigency tables with a finite number of discrete values of the variables, sought for transformations of the marginal variables that would yield linear least squares regression lines. He found that these variables maximised the coefficients of correlation.

Fisher [3] defined a set of variables on each of the marginal distributions of an $m \times n$ contigency table, such that $x_j = 1$ for an observation falling into the jth class and $x_j = 0$ elsewhere for $j = 1, 2 \cdots m - 1$, and similarly for y_j with $j = 1, 2 \cdots (n - 1)$. His problem was to find a linear form in the x_j, which would have maximum correlation with any linear form in the y_j. For convenience, these linear forms were considered without loss of generality as being normalised. Fisher referred to such a variable and the corresponding correlation as canonical and thus identified them with the canonical variables and correlation of Hotelling [10]. Fisher's theory was amplified by Maung [13] and Williams [25], who considered observational data in the form of a contingency table. We shall see later that in this case, the problem of finding the canonical correlations is equivalent to the determination of the canonical form of a rectangular matrix under pre- and post-multiplication by orthogonal matrices.

It is of interest to extend this type of analysis to the theoretical parent population and to more general classes of bivariate distributions. Lancaster [12] applied the methods of the theory of integral equations to find the canonical correlations and variables in the joint normal distribution and this work leads to a generalisa-

Received September 30, 1957; revised December 10, 1957.

tion of the canonical correlation theory. If the correlation is to have meaning, the canonical variables must have a finite variance, so that each canonical variable can be expressed as an orthonormal linear form in a complete set of orthogonal functions defined on the marginal distribution. The problem is now one in eigenvalue theory. Indeed, it is shown that the canonical correlations are the eigenvalues and the canonical variables on each marginal distribution form a subset, perhaps improper, of a complete set; the canonical variables are, moreover, the eigenfunctions except for a factor. This analysis holds provided the limiting value of Pearson's ϕ^2 is finite. If ϕ^2 is finite, it is further shown that the bivariate distribution can be expanded in an eigenfunction expansion. ϕ^2 is then the sum of the squares of the canonical correlations. The contingency table is then shown to be a special case of the general theory.

Once the canonical form of a bivariate population, that is, the eigenfunction expansion, has been obtained, some further applications of the theory can be made. First, the regressions take a particularly simple form and are confirmed to be the solution of Hirschfeld's problem. Second, given the marginal distributions it is possible to obtain bivariate distributions with prescribed correlations. Third, a goodness of fit test can be devised for the bivariate joint normal distribution, which displays as components of χ^2, the contributions of the regressions of the ith Hermite-Chebishev polynomial in x on the jth polynomial in y. The test is made of the total contributions from those pairs for which $i \neq j$.

2. Pearson's ϕ^2 as the Sum of Squares of the Correlation Coefficients. K. Pearson [18] introduced ϕ^2 as the "mean square contigency" for a bivariate distribution in order to derive a measure of association independent of the sample size, N. He wrote $\phi^2 = \chi^2/N$. Pearson saw that χ^2 (or rather ϕ^2) had a use as a descriptive measure, whereas it is usually thought of as a criterion of goodness of fit, e.g., as in the test due to Pearson [16]. It is convenient to modify Pearson's definition by using the integral sign in the sense of Lebesgue-Stieltjes and adopting the notation of Hellinger [7], which has been justified by Hobson [9].

DEFINITION.

$$\phi^2 = \iint_{-\infty}^{+\infty} [dF(x, y)]^2/[dG(x)\, dH(y)] - 1 \tag{1A}$$

$$= \iint_{-\infty}^{\infty} \Omega^2(x, y)\, dG(x)\, dH(y) - 1 \tag{1B}$$

where

$$\Omega(x, y) = dF(x, y)/[dG(x)\, dH(y)]. \tag{2}$$

$\Omega(x, y)$, and so the integrand of (1A), is to be taken as zero, if the point (x, y) does not correspond to points of increase of both $G(x)$ and $H(y)$. ϕ^2 can evidently be regarded as the limit of the sum $\sum_{i,j} f_{ij}^2/(f_{i.}f_{.j}) - 1$, where f_{ij} is the weight of the bivariate distribution corresponding to marginal sets, A_i and B_j, and where $f_{i.}$ and $f_{.j}$ are the weights of the marginal distributions corresponding to the same sets.

Examples of bounded -ϕ^2 distributions are provided by the joint distribution of independent stochastic variables, in which case ϕ^2 is zero, and by the bivariate normal distribution with the absolute value of the correlation less than unity. All discrete distributions with finitely many points of increase in both variables will also have a finite ϕ^2. A case of special interest is provided by the bivariate joint normal distribution. In this distribution we may write $g(x)\,dx$ and $h(y)\,dy$ in place of $dG(x)$ and $dH(y)$ respectively and $f(x, y)\,dx\,dy$ in place of $dF(x, y)$. Pearson derived the relation,

$$\phi^2 = \iint f^2(x, y)/[g(x)h(y)]\,dx\,dy - 1 = \rho^2/(1 - \rho^2), \tag{3}$$

where $|\rho| < 1$. This result has been discussed by Lancaster [12]. However, if $|\rho| = 1$ and so the bivariate normal distribution is singular, ϕ^2 is unbounded. Indeed, ϕ^2 is unbounded for any bivariate distribution distributed along a straight line, with infinitely many points of increase.

It follows from the definition by an analysis similar to that used to justify the Riemann integral that ϕ^2 is uniquely determined by the passage to the limit if it is bounded.

DEFINITION. Let $\{x^{(i)}\}$ and $\{y^{(i)}\}$ be complete sets of orthonormal functions defined on the marginal distributions, $G(x)$ and $H(y)$, respectively by

$$\int x^{(i)}x^{(j)}\,dG(x) = \int y^{(i)}y^{(j)}\,dH(y) = \delta_{ij}. \tag{4}$$

Let ρ_{ij} be the correlation coefficients,

$$\rho_{ij} = \iint x^{(i)}y^{(j)}\,dF(x, y). \tag{5}$$

By the Schwarz inequality ρ_{ij} always exists and is not greater than unity in absolute value. Further,

$$\rho_{00} = 1, \qquad \rho_{0k} = \rho_{k0} = 0 \qquad k \neq 0. \tag{6}$$

The following discussion gives a statistical content to some well known analysis. The steps taken can be justified by the theory of integral equations as set out in Courant and Hilbert [2] or Riesz and Szent-Nagy [22].

THEOREM 1. *If $F(x, y)$ is a ϕ^2-bounded distribution and if*

$$S_{mn} = S_{mn}(x, y) = \sum_{i=0}^{m} \sum_{j=0}^{n} \lambda_{ij}\, x^{(i)}y^{(j)}, \tag{7}$$

then

$$Q_{mn} = \iint (\Omega - S_{mn})^2\,dG(x)\,dH(y) \quad . \tag{8}$$

is minimised by taking

$$\lambda_{ij} = \rho_{ij}, \qquad i = 0, 1, 2, \cdots m; j = 0, 1, 2, \cdots n. \tag{9}$$

Writing S for S_{mn} as $m \to \infty$ and $n \to \infty$,

(10) $$\Omega(x, y) = S(x, y), \quad \textit{almost everywhere}$$

and

(11) $$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \rho_{ij}^2 = \phi^2.$$

PROOF. The set $\{x^{(i)}\} \times \{y^{(i)}\}$ is complete over the distribution $G(x) \times H(y)$, and $\Omega(x, y)$, as defined in (2), is square summable by (1B) and the hypothesis of the theorem. The result (9) follows by differentiating (7) with regard to λ_{ij} for $i = 0, 1, 2, \cdots m; j = 0, 1, 2, \cdots n$. For any finite m and n, the sum $\sum_{i,j} \rho_{ij}^2 \leqq \phi^2$, so that $\sum_{i,j} \rho_{ij}^2$ converges. The completeness assures the truth of (10) and of (11), which is the Parseval equality.

It is our aim now to redefine the sets $\{x^{(i)}\}$ and $\{y^{(j)}\}$ so that the correlation matrix,

(12) $$R = (\rho_{ij}), \qquad i = 1, 2, \cdots, j = 1, 2, \cdots,$$

assumes as simple a form as possible. The theorems of the next section show that R is diagonal if we choose, for the sets $\{x^{(i)}\}$ and $\{y^{(i)}\}$, the canonical variables in the sense of Fisher. The chief difficulty lies in the need to prove that the canonical variables form subsets of complete sets of orthonormal functions. We have, therefore, to proceed indirectly.

3. The Canonical Variables. The canonical variables have been defined on discrete distributions with finitely many points of increase. They are usually thought of as "scores to be assigned" but may also be thought of as functions of the marginal variables. Often no marginal variable has been explicitly defined; then, we may take the row or column position as the variable. The following definition may be regarded as the appropriate extension of Fisher's definition.

DEFINITION. The canonical variables (or functions) are two sets of orthonormal functions defined on the marginal distributions in a recursive manner such that the correlation between corresponding members of the two sets is maximal. Unity may be considered as a member of zero order of each set of variables. Symbolically, the orthogonal and normalising conditions are

(13) $$\begin{cases} \xi^{(i)} = \xi^{(i)}(x), \eta^{(i)} = \eta^{(i)}(y), & \\ \int \xi^{(i)}\, dG(x) = \int \eta^{(i)}\, dH(y) = 0, & i = 1, 2, \ldots, \\ \int \xi^{(i)2}\, dG(x) = \int \eta^{(i)2}\, dH(y) = 1, & i = 1, 2, \ldots, \\ \int \xi^{(i)}\xi^{(j)}\, dG(x) = \int \eta^{(i)}\eta^{(j)}\, dH(y) = 0 & \text{for } i \neq j, \end{cases}$$

and the maximisation conditions are that

$$\rho_i = \text{corr}\,(\xi^{(i)}, \eta^{(i)}) = \iint \xi^{(i)}\eta^{(i)}\, dF(x, y) \tag{14}$$

should be maximal for each i, given the preceding canonical variables. The ρ_i are the canonical correlations and can by convention be taken always to be positive.

THEOREM 2. *The canonical variables obey a second set of orthogonal conditions,*

$$E(\xi^{(i)}\eta^{(j)}) = \iint \xi^{(i)}\eta^{(j)}\, dF(x, y) = 0, \qquad \textit{if } i \neq j. \tag{15}$$

PROOF. For definiteness, let $j > i$. By hypothesis $E(\xi^{(i)}\eta^{(i)})$ is maximal in the sense of the definition above and is equal to ρ_i, say. Suppose that $E(\xi^{(i)}\eta^{(j)})$ is not zero but equal to $\rho_i \tan\theta$. Now $\eta^{(j)}$ has been defined according to (13) and so the function, $\text{Cos}\,\theta\eta^{(i)} + \text{Sin}\,\theta\eta^{(j)}$, obeys all the necessary orthogonal and normalising conditions, and its correlation with $\xi^{(i)}$ is easily found to be $\rho_i \sec\theta$ and this is greater than ρ_i, a contradiction results and so the theorem is proved.

As has been already noted, the canonical functions are necessarily square summable and so can be written as linear forms in any complete set of orthonormal functions, defined on the marginal distributions. Thus we can write

$$\begin{cases} \xi^{(i)} = \sum_{k=1}^{\infty} a_{ik} x^{(k)}, & \sum_k a_{ik}^2 = 1, \\ \eta^{(i)} = \sum_{k=1}^{\infty} b_{ik} y^{(k)}, & \sum_k b_{ik}^2 = 1. \end{cases} \tag{16}$$

Now let us determine $\xi^{(1)}$ and $\eta^{(1)}$ in terms of the $\{x^{(i)}\}$ and $\{y^{(i)}\}$ respectively.

$$\begin{aligned} \text{Corr}\,(\xi^{(1)}, \eta^{(1)}) &= \text{corr}\,(\textstyle\sum_i a_i x^{(i)}, \sum_k b_k x^{(k)}) \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_i b_j \rho_{ij}. \end{aligned} \tag{17}$$

Now $\sum_{i,j} \rho_{ij}^2$ is convergent and so the bilinear form on the right of (17) can be treated by the theory of quadratic forms in infinitely many variables. The normalising conditions (13) assure us that $\sum_i a_i^2 = 1$ and $\sum_j b_j^2 = 1$ and that neither $\xi^{(1)}$ nor $\eta^{(1)}$ contains any constant term. The bilinear form will have an attained maximum value for variations in the a_i and b_j. We take the coefficients of one such maximum to define a new set of variables

$$\begin{cases} x^{*(1)} = \xi^{(1)} = \sum_i a_i x^{(i)}, \\ x^{*(2)} = a_{21}x^{(1)} + a_{22}x^{(2)}, \\ x^{*(3)} = a_{31}x^{(1)} + a_{32}x^{(2)} + a_{33}x^{(3)}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{cases} \tag{18}$$

where the a_{2j}, a_{3j}, $\cdots$ are chosen to satisfy the orthogonal and normalising conditions. A similar transformation is applied to the $y^{(i)}$:

$$(19)\quad \begin{cases} y^{*(1)} = \eta^{(1)} = \sum_j b_j y^{(j)}, \\ y^{*(2)} = b_{21}y^{(1)} + b_{22}y^{(2)}, \\ y^{*(3)} = b_{31}y^{(1)} + b_{32}y^{(2)} + b_{33}y^{(3)}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{cases}$$

But now the correlation matrix, $R = (\rho_{ij})$, in the new variables is simpler in that, because of Theorem 2,

$$(20)\quad \rho_{i1} = \rho_{1i} = 0 \qquad i \neq 1.$$

We can proceed similarly to find $\xi^{(2)}$ and $\eta^{(2)}$ in terms of the $\{x^{(i)}\}$ and $\{y^{(i)}\}$ respectively. Since $\xi^{(2)}$ is orthogonal to $\xi^{(1)}$

$$(21)\quad \xi^{(2)} = \sum_2^\infty a_i^* x^{*(i)},$$

and similarly,

$$(22)\quad \eta^{(2)} = \sum_2^\infty b_j^* y^{*(j)}$$

with $\sum_i a_i^{*2} = \sum_i b_i^{*2} = 1$. Now to find $\xi^{(2)}$ and $\eta^{(2)}$ we shall have to maximise $\sum_{i=2}^\infty \sum_{j=2}^\infty a_i^* b_j^* \rho_{ij}^*$. This again has an attained maximum and we take again a new set of variables

$$(23)\quad \begin{aligned} x^{+(1)} &= x^{*(1)} = \xi^{(1)}, \\ x^{+(2)} &= \textstyle\sum_2^\infty a_i^* x^{*(1)} = \xi^{(2)}, \\ x^{+(3)} &= a_{32}^* x^{*(2)} + a_{33}^* x^{*(3)}, \\ x^{+(4)} &= a_{42}^* x^{*(2)} + a_{43}^* x^{*(3)} + a_{44}^* x^{*(4)}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots, \end{aligned}$$

and similarly define $y^{+(1)}$, $y^{+(2)}$, $y^{+(3)}$ $\cdots$ in terms of the $y^{(1)}$. The correlation matrix is simplified again for now

$$(24)\quad \begin{cases} \rho^+_{1i} = \rho^+_{i1} = 0 & \text{for } i \neq 1, \\ \rho^+_{2i} = \rho^+_{i2} = 0 & \text{for } i \neq 2. \end{cases}$$

This process may be continued a denumerable infinity of times or until all ρ_i are zero for $i > r$ or $j > r$ for some value of r. We may follow Williams [25] and refer to r as the rank of the departure from independence. r may be infinite. At each step, since the transformation is orthogonal, a complete set is transformed into a complete set. It is evident that we may pass from the sets $\{x^{(i)}\}$ and $\{y^{(i)}\}$ by a series of orthogonal transformations to complete sets of orthonormal functions, of which the sets $\{\xi^{(i)}\}$ and $\{\eta^{(i)}\}$ are subsets and conversely. We can sum up these results in

THEOREM 3. *If $F(x, y)$ is a ϕ^2-bounded bivariate distribution with marginal distribution, $G(x)$ and $H(y)$, then complete sets of orthonormal functions can be defined on the marginal distributions such that each member of a set of canonical variables appears as a member of the complete set of orthonormal functions. The element of frequency can be expressed in terms of the marginal distributions,*

$$dF(x, y) = \left\{1 + \sum_1^\infty \rho_i x^{(i)} y^{(i)}\right\} dG(x)\, dH(y), \quad \text{a.e.,} \tag{25}$$

and

$$\phi^2 = \sum_{i=1}^\infty \rho_i^2. \tag{26}$$

PROOF. We have just proved the first statement. To prove the second we write, in the same way as in Theorem 1,

$$Q = \iint \{\Omega(x, y) - S_{mn}(x, y)\}^2\, dG(x)\, dH(y) \tag{27}$$

and take the partial differentials of Q with respect to λ_{ij}. Owing to the simplified form of the correlation matrix, ρ_{ij} is now zero for $i \neq j$ and ρ_{ii} is ρ_i. Since $\{x^{(i)}\} \times \{y^{(i)}\}$ is a complete set on $G(x) \times H(y)$, it follows that the minimised Q tends to zero as $m \to \infty$ and $n \to \infty$, and (26) which is the Parseval equality follows.

It may be proved that the choice of orthonormal functions is unique except for a convention as to sign if the ρ_i form a pair-wise different set. It is assumed throughout that, once $x^{(i)}$ is chosen, $y^{(i)}$ is defined so as to give the expectation of $x^{(i)}y^{(i)}$ a positive value. If, however, $\rho_{j+1}, \rho_{j+2}, \cdots \rho_{j+k}$ are of equal magnitude and $x^{(j+1)}, x^{(j+2)}, \cdots x^{(j+k)}$ is one solution for the corresponding canonical variables, then every other solution is given by an arbitrary orthogonal transformation on these $x^{(j+1)} \cdots x^{(j+k)}$ and the same transformation on the $y^{(j+1)} \cdots y^{(j+k)}$. A converse of Theorem 3 holds.

THEOREM 4. *If a bivariate distribution can be written in the form* (25) *with $\{x^{(i)}\}$ and $\{y^{(i)}\}$ forming complete sets on the marginal distribution and if $\sum_i \rho_i^2$ is finite, then the ρ_i are the canonical correlations, $x^{(i)}$ and $y^{(i)}$ are the canonical variables and $\sum_i \rho_i^2 = \phi^2$.*

PROOF. The proof is by induction. We suppose first that the ρ_i are pairwise different. Then if ξ and η are the first pair of canonical variables

$$\begin{aligned} \text{corr}(\xi, \eta) &= \text{corr}\left(\sum_i a_i x^{(i)}, \sum_j b_j y^{(j)}\right) \\ &= \sum_i a_i h_i \rho_i. \end{aligned} \tag{28}$$

Now $\sum_i a_i^2 = \sum_i b_i^2 = 1$ and Cauchy's inequality shows that the sum on the right of (28) is maximised by taking $a_1 = b_1 = 1$ and all other coefficients zero. Similarly, if $\rho_1 = \rho_2 \cdots = \rho_k$, Cauchy's inequality shows that the correlation of ξ and η is ρ_i if $\sum_1^k a_i^2 = 1$ and $a_i = b_i$ and that this is the maximum. Clearly however in this case too we can take $a_1 = b_1 = 1$, and once again $x^{(1)}$ and $y^{(1)}$ are the pair of first canonical variables or functions. We can proceed by induction to prove the main statement of the theorem. Defining $\Omega(x, y)$ as in (2) and writing out its value by the use of (25), we derive

$$\sum_1^\infty \rho_i^2 = \phi^2.$$

This is a generalisation of a result of Hirschfeld [8] and Maung [13] in the finite case. Further, we may note that Theorem 3 is a generalisation of the Mehler identity; for, using the notation of (3), we define complete sets of orthogonal functions $\{x^{(i)}\} = \{\psi_i(x)\}$ and $\{y^{(i)}\} = \{\psi_i(y)\}$ on the marginal distributions where $\psi_i(x)$ is a polynomial of precise degree i standardised by the formula

$$\text{(29)} \qquad \begin{cases} \int \psi_i(x)\psi_j(x)g(x)\,dx = \delta_{ij}, \\ \int \psi_i(y)\psi_j(y)h(y)\,dy = \delta_{ij}. \end{cases}$$

$g(x)$ and $h(y)$ have the same functional form in this case. By considering the expectation of $\exp\{tx - \frac{1}{2}t^2 + uy - \frac{1}{2}u^2\}$, namely $\exp \rho ut$, we find that

$$\text{(30)} \qquad E\, x^{(i)}y^{(j)} = \delta_{ij}\rho^i$$

and Mehler's identity (Mehler [14]; Watson [24]) follows after Theorem 3 and continuity considerations. Conversely, given Mehler's identity, Theorem 4 shows that $|\rho|^i$ are the canonical correlations in this special case and the standardised Hermite-Chebishev polynomials, the canonical variables. Pearson [17] showed the great value of the Mehler identity in discussing normal correlation, although he and his collaborator, Bramley-Moore, failed to note that the tetrachoric expansion is indeed the Mehler identity. The Mehler identity is the special case when $f(x)$ and $g(y)$ are standardised normal distributions and $h(x, y)$ is the bivariate normal distribution with coefficient or correlation, ρ. This identity is given in Szegö's textbook [27] on page 371, where Szegö has $x\sqrt{2}$ and $y\sqrt{2}$ corresponding to our x and y and w for our ρ. Our $\psi_i(x)$ is $H_i(2^{-\frac{1}{2}}x)/\sqrt{i!}$ in his notation.

Dr. G. S. Watson (personal communication) has pointed out that the usual eigenfunction and kernel theory might be applied. The analogy is quite easy to establish in purely discrete or purely continuous distributions. In the continuous case we should define a kernel

$$\text{(31)} \qquad K(x, y) = f(x, y)\,\{g(x)\,h(y)\}^{-\frac{1}{2}}$$

where $g(x) > 0$, $h(y) > 0$, with the convention that $K(x, y) = 0$ if $g(x)\,h(y) = 0$. $K(x, y)$ would in general be unsymmetric. It would follow that

$$\text{(32)} \qquad \begin{cases} \rho_j y^{(j)}\sqrt{h(y)} = \int K(x, y)x^{(j)}\sqrt{g(x)}\,dx, \\ \rho_j x^{(j)}\sqrt{g(x)} = \int K(x, y)y^{(j)}\sqrt{h(y)}\,dy, \end{cases}$$

in precisely the same way as in equation (26) and (27) of Schmidt [23], noting the different definitions for the eigenvalues. (32) is proved by the application of Theorem 3. In the finite discrete case, where the frequencies are f_{ij}, the kernel $K(x, y)$ is replaced by $f_{ij}f_{i\cdot}^{-\frac{1}{2}}f_{\cdot j}^{-\frac{1}{2}} = b_{ij}$ and this is discussed in the next section. (32) is simplified if the marginal distributions are rectangular with $g(x) = h(y) = 1$.

4. The Finite Case. The discussion above is a generalisation of a procedure, alternative to that of Fisher [3] and Maung [13], which may be used in the finite discrete case of an m by n contigency table with proportions f_{ij} in the cell of the ith row and jth column, with $f_{i\cdot} = \sum_j f_{ij} > 0, f_{\cdot j} = \sum_i f_{ij} > 0$, and for definiteness, $m \leqq n$. It follows from Theorem 3 that if we construct matrices, X and Y, with the $(k + 1)$th column consisting of the values of the kth canonical variable, then $X'FY$ will have a canonical form with non-zero elements everywhere except along the leading diagonal. It is found simpler to deal with a matrix B derived from F and then the problem is reduced to determining a canonical form for a rectangular matrix under pre- and post-multiplication by orthogonal matrices, which we consider by an adaptation of the argument of Murnaghan [15] on his pages, 26 and 27. The defining conditions for the matrices X and Y may be written

$$\begin{aligned} x_{i1} &= 1 & i &= 1, 2, \cdots m, \\ y_{i1} &= 1 & i &= 1, 2, \cdots n, \\ x_{ij} &= \xi_{(i)}^{(j-1)} = \xi_i^{(j-1)}, & j &= 2, 3, \cdots, m, \\ y_{ij} &= \eta_{(i)}^{(j-1)} = \eta_i^{(j-1)}, & j &= 2, 3, \cdots, n. \end{aligned} \tag{33}$$

(13) now becomes

$$\begin{cases} X' \operatorname{diag} f_{i\cdot} X = 1_m, \\ Y' \operatorname{diag} f_{\cdot j} Y = 1_n, \end{cases} \tag{34}$$

and the elements of the leading diagonal of $X'FY$ are to be maximised. Theorem 4 ensures that it is sufficient and Theorem 2 that it is necessary for $X'FY$ to be in canonical form. We therefore state without completing the proof

THEOREM 5. *Given an $m \times n$ contigency table with proportions f_{ij} in the cell of the* ith *and* jth *column, let an $m \times n$ matrix, B, be defined by*

$$b_{ij} = f_{ij}f_{i\cdot}^{-\frac{1}{2}}f_{\cdot j}^{-\frac{1}{2}}. \tag{35}$$

Then orthogonal matrices M and N exist with elements of the first column $\sqrt{f_{i\cdot}}$ and $\sqrt{f_{\cdot j}}$ respectively such that $M'BN$ is in canonical form, namely

$$M'BN = C = [\operatorname{diag}(1, \rho_1 \cdots \rho_{m-1}), 0_{m,n-m}]. \tag{36}$$

It is evident further by a consideration of the forms of $(M'BN)$ $(M'BN)'$ and $(M'BN)'$ $(M'BN)$ that M and N are the orthogonal matrices that reduce BB' and $B'B$ respectively to canonical form. Conversely, it can be shown that if N

transforms $B'B$ to canonical form with unity in the leading position and k other non-zero diagonal elements, then an M, having for its first $(k+1)$ columns the first $(k+1)$ columns of BN normalised, can be constructed so that $M'BN$ is in the required form. In fact, the first $(k+1)$ columns of BN are mutually orthogonal because $(NB)'$ (BN) is diagonal. Maung [13], obtains the latent roots of BB' or $B'B$ by solving the determinantal equation, $| BB' - \lambda 1 | = 0$, in the usual manner. An alternative is to use the iteration method of Frazer, Duncan and Collar ([6], page 133). We note further that M and N must be of the form,

$$\begin{cases} M = M_1(1 \dotplus M_2), \\ N = N_1(1 \dotplus N_2), \end{cases} \tag{37}$$

where M_1 and N_1 are of the Helmert type with first columns having elements $f_{i.}^{\frac{1}{2}}$ and $f_{.j}^{\frac{1}{2}}$ respectively. Now the elements of

$$M_1'BB'M_1 = 1 \dotplus W \tag{38}$$

can be computed readily. Using the observed number, a_{ij} in the contigency table,

$$W_{kk'} = \frac{a_{..}\left(a_{k+1,.}\sum_{i=1}^{k} a_{ij} - a_{k+1,j}\sum_{i=1}^{k} a_{i.}\right)\left(a_{k'+1,.}\sum_{i=1}^{k'} a_{ij} - a_{k'+1,j}\sum_{i=1}^{k'} a_{i.}\right)}{a_{.j}\left\{a_{k.}\, a_{k'}\, a_{k+1,.}\, a_{k'+1,.}\sum_{1}^{k} a_{i.}\sum_{1}^{k'} a_{i.}\right\}^{\frac{1}{2}}} \tag{39}$$

The trace of W is χ^2. It does not take much more time to compute W than χ^2 if m is not too large. A computing routine is to form a matrix with elements in the first row, $(a_{1j}a_{2.} - a_{1.}a_{2j})$, elements in the second row $(a_{1j} + a_{2j})a_{3.} - (a_{1.} + a_{2.})a_{3j}$ and so on. For each row, a standardising factor is computed,

$$\left\{a_{k.}\, a_{k+1,.}\sum_{1}^{k} a_{i.}\right\}^{\frac{1}{2}}.$$

The elements of W are then simply computed by formula (39). The Helmert matrix can be looked upon as generating sets of orthonormal functions, which take a simple form. The values for the canonical variables are then calculated by an orthogonal transformation

$$\begin{aligned} X &= \operatorname{diag} f_{i.}^{-\frac{1}{2}} M \\ &= f_{i.}^{-\frac{1}{2}} M_1(1 \dotplus M_2) \end{aligned} \tag{40}$$

where M_1 is the Helmert Matrix and $M_2'WM_2$ is diagonal, M_2 being obtained by iteration and similarly Y can be written in terms of N_1 and N_2.

A NUMERICAL EXAMPLE. Maung [13] has given the following example of a classification of Aberdeen schoolchildren by hair and eye colours (see Table I).

A matrix of elements, U, with $u_{kj} = (a_{k+1,.}\sum_{i=1}^{k} a_{ij} - a_{k+1,j}\sum_{i=1}^{k} a_{i.})$ is given by

$$\begin{bmatrix} 1{,}487{,}190 & -273{,}082 & -1{,}077{,}957 & -110{,}090 & -26{,}061 \\ 16{,}182{,}645 & 773{,}584 & -8{,}895{,}366 & -7{,}831{,}720 & -229{,}143 \\ 19{,}806{,}181 & 1{,}123{,}770 & 7{,}415{,}022 & -26{,}653{,}016 & -1{,}691{,}957 \end{bmatrix}.$$

TABLE I

Eye colour	Hair colour					Total
	Fair	Red	Medium	Dark	Black	
Blue	1368	170	1041	398	1	2978
Light	2577	474	2703	932	11	6697
Medium	1390	420	3826	1842	33	7511
Dark	454	255	1848	2506	112	5175
Total	5789	1319	9418	5678	157	22,361

The elements of this matrix are now divided by the corresponding column totals of the contigency table to give a matrix (v_{ij}). Divisors appropriate to each row of U are now computed, $\{a_{k.}a_{k+1,.} \sum_1^k a_{i.}\}^{\frac{1}{2}} = d_{k.}$. Then w_{ij} is $\sum_k u_{ik}v_{jk}/\{d_i d_j\}$ or $\sum_k v_{ik}u_{jk}/\{d_i d_j\}$. We thus obtain the matrix, W, of (38).

$$\begin{bmatrix} 65.8744811 & 237.1027158 & 173.4280109 \\ 237.1027158 & 1167.9147643 & 1252.2082711 \\ 173.4280109 & 1252.2082711 & 2450.0865906 \end{bmatrix}.$$

The trace of W is 3683.875836 agreeing with Maung's value for χ^2.

The orthogonal matrix, M_2, of (37) is then derived from W by an iteration process and is

$$\begin{bmatrix} 0.085413 & 0.272546 & 0.958344 \\ 0.522636 & 0.806650 & -0.275985 \\ 0.848266 & -0.524438 & 0.073545 \end{bmatrix}.$$

The values of the complete set of orthonormal variables associated with the Helmert matrix, M_1, may be displayed as a matrix,

$$\begin{bmatrix} 1 & 2.279806 & 1.005036 & 0.548741 \\ 1 & -1.013777 & 1.005036 & 0.548741 \\ 1 & 0 & -1.294598 & 0.548741 \\ 1 & 0 & 0 & -1.822352 \end{bmatrix}.$$

In the jth column, all elements above the diagonal are equal to $\{p_{j.}/(\sum_1^{j-1} p_{k.}\sum_1^j p_{k.})\}^{\frac{1}{2}}$, the diagonal element is $-\{\sum_1^{j-1} p_{k.}/(p_{j.}\sum_1^j p_{k.}\}^{\frac{1}{2}}$ and element below the diagonal are zero. Post-multiplication of this matrix by $(1 \dotplus M_2)$ yields the sets of canonical variables in the form of a 4×4 matrix, X, of Equation (40)

$$\begin{bmatrix} 1 & +1.1855 & +1.1443 & +1.9478 \\ 1 & +0.9042 & +0.2466 & -1.2086 \\ 1 & -0.2111 & -1.3321 & +0.3976 \\ 1 & -1.5458 & +0.9557 & -0.1340 \end{bmatrix}.$$

The values of the elements agree with those given by Maung.

The canonical variables in y can now be obtained by using Fisher's algorithm as in (45), below, and we may write the first four columns of the matrix, Y, as

$$\begin{bmatrix} 1 & +1.3419 & +0.9713 & +0.3288 \\ 1 & +0.2933 & -0.0236 & -3.7389 \\ 1 & +0.0038 & -1.1224 & +0.1666 \\ 1 & -1.3643 & +0.7922 & +0.3625 \\ 1 & -2.8278 & +3.0607 & -3.8177 \end{bmatrix}.$$

Programs, similar to the computational process used above, are now available on electronic computers.

Interpreting the findings, the first set of canonical variables arranges both hair colour and eye colour in the same order as was suggested by biological considerations. If there is an underlying bivariate distribution the first set of canonical variables gives the best values to be assigned to the marginal variables.

5. Identifications of the Finite and the General Cases. We now state some corollaries deducible from the theorems above in such a way as to bring out the identity of the theory of canonical correlation as a special case of the more general theory; where appropriate, we have numbered these "a" for the finite case, "b" for the more general.

COROLLARIES.

(ia). ρ_i^2 are the non-zero latent roots of the matrices BB' and BB'; ρ_i are the "roots" of B under transformation by pre- and post-multiplication of B by orthogonal matrices.

(ib). ρ_i^2 are the eigenvalues of certain symmetric kernels and ρ_i are the eigenvalues of a certain, possibly asymmetric, kernel.

(iia). The identity of Fisher [4]

$$f_{ij} = f_{i.}f_{.j}\left\{1 + \sum_{k=1}^{m-1} \rho_k x^{(k)} y^{(k)}\right\} \tag{41}$$

is a special case of our Theorem 3. It is also proved by noting that

$$X'AY = M'BN = C, \tag{42}$$

and the inverse of X' is diag $f_{i.}X$ and the inverse of Y is Y' diag $f_{.j}$ by (34).

(iib). The generalisation of Fisher's identity is given by Theorem 3.

(iiia) and (iiib). If $\mathbf{m}_k$ and $\mathbf{n}_k$ are the kth column vectors of M and N respectively

$$\begin{cases} \rho_k \mathbf{n}_k = B'\mathbf{m}_k\,, \\ \rho_k \mathbf{m}_k = B\mathbf{n}_k\,, \end{cases} \tag{43}$$

or alternatively after (36)

$$\begin{cases} BN = MC, \\ B'M = NC', \end{cases} \tag{44}$$

or

$$\begin{cases} AY = \operatorname{diag} f_{i.}XC, \\ A'X = \operatorname{diag} f_{.j}YC', \end{cases} \tag{45}$$

(45) corresponds exactly with equations (26) and (27) of Schmidt [23] as modified in our (32). The equation (45) is the basis of Fisher's [3] algorithm for the computation of the canonical correlations, which we give as a corollary.

(iv). The canonical variables can be obtained by iteration if $\rho_{j+1} > \rho_{j+2}$. From (45) it follows that

$$\text{diag} f_{1.}^{-1} A \ \text{diag} f_{.1}^{-1} A' X = XCC', \tag{46}$$

and so

$$(\text{diag} f_{i.}^{-1} A \ \text{diag} f_{.i}^{-1} A')^p X \, \mathbf{n}_0 = X(CC')^p \, \mathbf{x}_0 . \tag{47}$$

Therefore if any vector $\mathbf{x}_0$ is taken orthogonal to the first j columns of X but not orthogonal to the $(j + 1)$th column, the iteration of the form (45) will yield a vector proportional to the $(j + 1)$th column of X. This is a special case of iterating using Schmidt's (26) and (27), which we could rewrite as (ivb).

(v). In Yates [26], arises the problem to find values for y such that y will have maximum correlation with an x, which has prescribed values.

We may write

$$x = \sum_{i=1}^{m-1} a_i x^{(i)}, \qquad \sum_i a_i^2 = 1. \tag{48}$$

Then from the canonical form of Theorem 3 and the use of the Cauchy inequality, we find that

$$y = \sum_{i=1}^{m-1} a_i \rho_i y^{(i)}, \tag{49}$$

is such that the correlation of x and y is maximal and

$$\text{corr}\,(x, y) = \left(\sum_{i=1}^{m-1} a_i^2 \rho_i^2 \right)^{\frac{1}{2}}. \tag{50}$$

(vi). In either finite or infinite cases, it can be proved that the existence of k canonical correlations of unity means the distribution consists of $(k+1)$ disjunct pieces. The case of one canonical correlation of unity has been treated by Richter [21].

6. Regression in the Bivariate Distribution. If the bivariate surface can be described in the canonical form (25), then regression takes a particularly simple form.

THEOREM 6. *The regressions of the canonical variables are given by the lines,*

$$\begin{aligned} x^{(i)} &= \rho_i y^{(i)}, \\ y^{(i)} &= \rho_i x^{(i)}. \end{aligned} \tag{51}$$

For $i \neq j$ the regression of $x^{(i)}$ on $y^{(j)}$ and $y^{(i)}$ on $x^{(j)}$ are zero.

PROOF. This follows in the usual way by minimising

$$\iint (x^{(i)} - \lambda y^{(j)})^2 \, dF(x, y).$$

Incidentally, we have proved that the regression of $x^{(i)}$ on $y^{(i)}$ is linear since any square summable function of $y^{(i)}$ orthogonal to $y^{(i)}$ can be expanded in terms of the other orthonormal functions.

7. Generalization of the Notion of Correlation. Many attempts have been made to find some way of obtaining bivariate distributions which would generalize the normal case. Pretorius [20] has given many references to such attempts. Fisher's theory of canonical correlation gives an alternative approach. Suppose we are given marginal variables with distribution functions $G(x)$ and $H(y)$, then a bivariate distribution can be formed using (25) provided that the series $[1 + \sum_1^\infty \rho_i x^{(i)} y^{(i)}]$ is non-negative at points corresponding to increase in both $G(x)$ and $H(y)$. We may take one of the simplest possible pairs of distributions for the margins, namely the rectangular over the range $-\frac{1}{2}$ to $\frac{1}{2}$ and set up three different bivariate distributions.

EXAMPLE 1. We take as our orthonormal sets of functions the normalised Legendre polynomials, in particular

$$\begin{cases} x^{(1)} = x\sqrt{12}, \\ x^{(2)} = 6\sqrt{5}\,(x^2 - \tfrac{1}{12}). \end{cases} \tag{52}$$

We can now assign correlations ρ_1 and ρ_2 subject to the condition that the density becomes nowhere negative

$$dF(x, y) = \{1 + 12\rho_1 xy + 180\rho_2(x^2 - \tfrac{1}{12})(y^2 - \tfrac{1}{12})\}\, dx\, dy. \tag{53}$$

But the maximum absolute value of $x^{(1)}y^{(1)}$ is 3 and that of $x^{(2)}y^{(2)}$ is 5, so the expression in (53) will be positive if

$$3 \mid \rho_1 \mid + 5 \mid \rho_2 \mid < 1. \tag{54}$$

EXAMPLE 2. We choose the cosine series as the orthonormal sets,

$$\begin{cases} x^{(1)} = \sqrt{2}\,\mathrm{Cos}(2\pi x), \\ x^{(2)} = \sqrt{2}\,\mathrm{Cos}(4\pi x), \end{cases} \tag{55}$$

and similarly define $y^{(1)}$ and $y^{(2)}$

$$dF(x, y) = \{1 + 2\rho_1\,\mathrm{Cos}\,(2\pi x)\,\mathrm{Cos}\,(2\pi y) + 2\rho_2\,\mathrm{Cos}\,(4\pi x)\,\mathrm{Cos}\,(4\pi y)\}\, dx\, dY. \tag{56}$$

This is non-negative if the absolute value of ρ_1 and ρ_2 are both less than $\frac{1}{2}$.

EXAMPLE 3. A further possibility results from forming arbitrary bivariate distributions, e.g., we might divide the square with corners at $(+\frac{1}{2}, +\frac{1}{2})$ into four quarters and add $+\rho_1$ to the density in the first and third quadrants and subtract ρ_1 from the density in the second and fourth quadrants. We could also subdivide the original square into 16 parts and add ρ_2 to the four corner subdivisions and to the four central subdivisions and subtract ρ_2 from the remainder.

The resulting distribution can be described with the aid of step-functions

$$dF(x, y) = \{1 + \rho_1 x^{(1)} y^{(1)} + \rho_2 x^{(2)} y^{(2)}\}\, dx\, dy, \tag{57}$$

where

$$\begin{cases} x^{(1)} = +1 \text{ for } x \leqq 0, \text{ for } x < 0, \\ x^{(2)} = -1 \text{ for } \tfrac{1}{4} \leqq x < \tfrac{1}{4} \text{ and } +1 \text{ for } x \text{ elsewhere.} \end{cases} \tag{58}$$

To obtain a complete set of orthogonal functions defined on $[-\frac{1}{2}, \frac{1}{2}]$ we divide this interval into four subintervals of equal length. On each complete sets of orthonormal functions may be defined. For example, we may choose the Legendre polynomials as our set, standardized so as to be orthonormal on the uniform distribution $[-\frac{1}{2}, \frac{1}{2}]$. Corresponding to the first interval we define a set of orthogonal polynomials which have the values $1 = P^{(0)}(X)$, $P^{(i)}X$, $i = 1, 2 \cdots$ where $X + \frac{1}{2}$ is the fractional part of $4(x + 1)$, on the first interval and zero elsewhere and similar sets on the other subintervals. The four sets of functions may be displayed as the elements of a four rowed matrix, P, of infinitely many columns. The rows of this matrix are obviously mutually orthogonal since no two elements of the same column can be simultaneously non-zero. Let us now define $Q = AP$, where A is the matrix

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 \end{bmatrix}$$

The elements of Q are now an orthonormal set on the whole interval. $q_{H} = x^{(0)}$, a term constant on $[\frac{1}{2}, \frac{1}{2}]$. $q_{21} = x^{(1)}$, $q_{31} = x^{(2)}$. q_{41} is necessary for completeness. It is constant on any subinterval but changes sign being -1 on the odd intervals. Every other function q_{ij} of the form $\pm P^{(i)}(X)$. The $y^{(i)}$ may be similarly defined.

It is clear from the examples that the same correlations can arise in a great many different ways. In the next section, we show how the methods can be used as a test of normality.

These three examples show how bivariate distributions can be formed with arbitrarily prescribed correlation coefficients. Barrett and Lampard [1] give two other examples where such bivariate distributions arise naturally out of a physical problem.

8. A Canonical Partition of χ^2. In testing whether a bivariate distribution is normal, the marginal distributions can be tested in the usual way by an overall χ^2 or the individual degrees of freedom can be displayed as previously suggested by Lancaster [11] by the aid of orthogonal polynomials. Moreover, according to the analysis of the present paper and that of Lancaster [11] the regressions of the orthogonal polynomials in x and y on one another should be zero except for polynomials of the same degree. We therefore may compute the regressions and display them in the form of a matrix, which we explain with the aid of a well-known example, the correlation table of Pearson and Lee (*Biometrika* 2,257), easily accessible in ([5], paragraph 30). After estimating the mean and variance of both variables, the regressions of the theoretical Hermite-Chebishev in one variable on those of the other may be computed and set out as suggested by Lancaster [11]. The mean and standard deviations have been computed using n as a divisor. The table of Pearson and Lee has been modified to 8 columns representing classifications of daughters' heights. The $\psi_i(x)\psi_j(y)$ sums of products of polynomials of the form, $\psi_i(x)\psi_j(y)f_{ij}$, have been computed and divided by 1376 the number of observations to give component χ's of a partition of χ^2. The leading 4 × 4 submatrix is as follows—

$$\begin{bmatrix} \cdot & \cdot & \cdot & -1.006 \\ \cdot & 19.238 & -0.053 & -1.834 \\ \cdot & 0.398 & 8.325 & -0.460 \\ -0.328 & -0.578 & -0.350 & 2.390 \end{bmatrix}$$

The term 19.238 corresponds to the regression of the first polynomial in the fathers' heights on first polynomial in the daughters' heights and to a correlation of 0.5186, which is slightly different from that given by Fisher [5] as the grouping is different. It may be noted also that the squares of the 3 × 3 submatrix excluding the marginal terms accounts for over 446 of a χ^2 of 504.23 if the table is analysed by the usual χ^2 with fixed marginal totals, so that all the significant departure from independence is shown to be accounted for by the first three not identically zero diagonal terms, the sum of whose squares is 445.

Pearson [19] gave a rule which substantially states that the number of degrees of freedom must be subtracted from the χ^2 of the test of homogeneity when computing ϕ^2. We have

$$\begin{aligned} \phi^2 &= (504.234 - 98)/1376 \\ &= 0.295228, \\ \rho^2 &= 0.295228/1.295228 \\ &= 0.227935, \\ \rho &= 0.477, \end{aligned}$$

which gives a correlation approximately equal to that calculated here, 0.5186.

An alternative canonical partition is given by estimating the means and variances and computing the marginal frequencies on the assumption of normality. A partition of χ^2 is obtained as shown in Table II.

It is clear that the distribution of Pearson and Lee is fitted very well by the assumption that it is a sample of a bivariate normal distribution. The residual χ^2 of 101.04 with 95 degrees of freedom represents the sums of squares due to all other regressions than the first three regressions of the form $\psi_i(x)$ on $\psi_i(y)$. The assumption of normality of the marginal distributions and a non-zero correlation are sufficient to account for the total χ^2, for the residual χ^2 is little greater than the corresponding degrees of freedom.

TABLE II

Source of χ^2	Degrees of Freedom	χ^2
Difference of distribution of father's heights from theoretical	5	7.20
Difference of distribution of daughter's heights from theoretical	12	12.77
Regression of $\psi_1(y)$ on $\psi_1(x)$	1	370.10
Regression of $\psi_2(y)$ on $\psi_2(x)$	1	69.31
Regression of $\psi_3(y)$ on $\psi_3(x)$	1	5.71
Residual	95	101.04
Total	115	566.13

9. Summary. The problems of Hirschfeld [8] and of the description of a contingency table by means of the canonical variables and correlations have been generalised to distributions limited only by the condition that the Pearson ϕ^2 is finite. Any theoretical or observed distribution subject to this condition can be described by the canonical variables (that is, subsets of complete sets of orthogonal functions in the variables of the two marginal distributions, which obey the second orthogonality condition that $Ex^{(i)}y^{(j)}$ is zero for $i \neq j$, and the canonical correlations. The theory of Fisher [3], Maung [13] and Williams [25] has been related to the eigenfunction theory.

Mehler's identity, or in statistical language, the expansion of the bivariate normal frequency in tetrachoric functions, has been generalised. The approach of Maung [13] has been modified to allow for an extension of the canonical theory to continuous marginal distributions.

The methods used give a new test of goodness of fit for the bivariate normal distribution and enable populations to be constructed with arbitrary marginal distributions and correlations.

Acknowledgments. This paper is published with the permission of the Director-General of Health, Canberra.

Dr. G. S. Watson has offered valuable help in its preparation.

REFERENCES

[1] J. F. Barrett, and D. G. Lampard, "An expansion for some second-order probability distributions and its application to noise problems," IRE *Transactions of the Professional Group on Information Theory*, IT 1 (1955), pp. 10–15.

[2] R. Courant, and D. Hilbert, *Methods of Mathematical Physics*, Interscience Publishers, Inc., New York, 1953.

[3] R. A. Fisher, "The precision of discriminant functions," *Ann. Eugenics*, Vol. 10, (1940), pp. 422–429.

[4] R. A. Fisher, cited by Maung [13].

[5] R. A. Fisher, *Statistical Methods for Research Workers*, 10th ed., Oliver and Boyd, Edinburgh, 1944.

[6] R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices*, Cambridge University Press, 1946.

[7] E. Hellinger, "Neue Begruendung der Theorie quadratischer Formen von unendlichvielen Veraenderlichen", *J. Reine Angew. Math.* (1909), Vol. 136, pp. 210–271.

[8] H. O. Hirschfeld, "A Connection between Correlation and Contingency", *Proc. Cambridge Philos. Soc.* Vol. 31 (1935), 520–524.

[9] E. W. Hobson, "On Hellinger's Integrals", *Proc. London Math. Soc.* (2), Vol. 18 (1919), pp. 249–265.

[10] H. Hotelling, "Relations between two Sets of Variates", *Biometrika*, Vol. 28 (1936), pp. 321–377.

[11] H. O. Lancaster, "A Reconciliation of χ^2 Considered from Metrical and Enumerative Aspects", *Sankhyā*, Vol. 13 (1953), pp. 1–10.

[12] H. O. Lancaster, "Some properties of the bivariate normal distribution considered in the form of a contingency table", *Biometrika*, Vol. 44 (1957), pp. 289–292.

[13] K. Maung, "Measurement of Association in a Contingency Table with Special Reference to the Pigmentation of Hair and Eye Colours of Scottish School Children", *Ann. Eugenics*, Vol. 11 (1941), pp. 189–223.

[14] F. G. Mehler, "Ueber die Entwicklung einer Funktion von beliebig vielen Variablen nach Laplaceschen Funktionen hoeherer Ordnung", *J. Reine Angew. Math.*, Vol. 66 (1866), pp. 161–176.

[15] F. D. Murnaghan, *The Theory of Group Representations*, The Johns Hopkins Press, Baltimore, 1938.

[16] K. Pearson, "On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling", *Philos. Mag.* (5), Vol. 50 (1900), pp. 157–175.

[17] K. Pearson, "Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable", *Philos. Trans. Roy. Soc. (Lond.)* Vol. 195A (1900), pp. 1–47.

[18] K. Pearson, "Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation", *Drapers Company Research Memoirs Biometric Series, No. 1* (1904)

[19] K. Pearson, "On the Measurement of the Influence of Broad Categories on Correlation", *Biometrika*, Vol. 9 (1913), pp. 116–139.

[20] S. J. Pretorius, "Skew bivariate frequency surfaces examined in the light of numerical illustrations", *Biometrika*, Vol. 22 (1930), pp. 109–223.

[21] H. Richter, "Zur Maximalkorrelation", *Z. angew. Math. Mech.* Vol. 29 (1949), pp. 127–128.

[22] F. Riesz, and B. Szent-Nagy, *Functional Analysis*. Translated from the 2d French Edition, 1956.

[23] E. Schmidt, "Zur Theorie der linearen und nichtlinearen Integralgleichungen", *Math. Ann.*, Vol. 63 (1907), pp. 433–476.

[24] G. N. Watson, "Notes on generating functions of polynomials: Hermite polynomials", *J. London Math. Soc.*, Vol. 8 (1933), pp. 194–199.

[25] E. J. Williams, "Use of scores for the analysis of association in contingency tables", *Biometrika*, Vol. 39 (1952) pp. 274–289.

[26] F. Yates, "The analysis of contingency tables with groupings based on quantitative characters", *Biometrika*, Vol. 35 (1948), pp. 176–181.

[27] G. Szegö, (1939), *Orthogonal Polynomials*, Colloquium Publ. No. 23, Amer. Math. Soc.