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.