ESTIMATION OF THE CORRELATION COEFFICIENT IN CONTINGENCY TABLES WITH POSSIBLY NONMETRICAL CHARACTERS

H. O. Lancaster and M. A. Hamdan

University of Sydney

A method of estimating the product moment correlation from the polychoric series is developed. This method is shown to be a generalization of the method which uses the tetrachoric series to obtain the tetrachoric correlation. Although this new method involves more computational labor, it is shown to be superior to older methods for data grouped into a small number of classes.

1. Introduction

Pearson [8] suggested that any deviation of the contingency table from independence might be called "a measure of its contingency"; the usual $\chi^2$ might be termed the square contingency and $\chi^2/N = \phi^2$ the mean square contingency. $\phi^2$ might be expected to converge to some invariant function of the bivariate distribution as the sample size $N$ became indefinitely large. $\phi^2$ was independent of the ordering of the marginal classes. For the bivariate normal, $\phi^2$ was an estimate of $\rho^2(1 - \rho^2)^{-1}$, so that, inverting this relation,

$$R = \phi/(1 + \phi^2)^{1/2}$$

would be an estimate of the coefficient of correlation $\rho$ of the underlying joint normal distribution. Pearson [8] assumed that $N$ was large and that the marginal classes were fine.

In later discussions, Pearson [9] and Pearson and Heron [10] suggested that a correction for the random sampling error should be made and that $\phi^2$ should be calculated by

$$\phi^2 = \{\chi^2 - (r - 1)(c - 1)\}/N.$$ 

In other words, the "degrees of freedom" or expected value of $\chi^2$ should be subtracted. The use of this correction by Pearson is worthy of more extended comment elsewhere. We have followed Pearson in this paper, using $\phi^2$ for both theoretical and observed values. Tschuprow [15] modified the Pearson procedure by defining

$$T^2 = \phi^2\{(r - 1)(c - 1)\}^{-1/2},$$

which attains its maximum value of unity only when $r = c$ and the de-
dependence is complete. Little has come out of this theory in the non-normal case.

It has never been clear how the broadness of the classes, that is, the fineness of the partition of the marginal distributions, affects the estimate of $\rho$ by the $\phi^2$ method. In Section 2 of the present paper, we illustrate by operations on a table of Pearson and Lee [11] how the Pearson $\phi^2$ method fails when the partition is not fine and when symmetric pooling of classes has been carried out. In Section 3, the theory of orthonormal functions is used to estimate $\rho$ in the general $r \times c$ contingency tables. In Section 4, the new method is shown to produce accurate estimates of $\rho$ for widely grouped data. Briefly, the unmodified method is effective if each of the marginal variables can be closely approximated by linear functions of the indicator variables of the marginal classes. If, however, symmetrical grouping is applied to either marginal distribution or if the partition is not sufficiently fine, the marginal variable cannot be approximated closely by a linear function of the new indicator variables.

2. The Bias of the Pearson $\phi^2$ Method

The statures of the 1376 pairs of fathers and daughters are given in Table 1, which has been obtained from Table XXV of Pearson and Lee [11] by freeing it of fractional frequencies and reducing the numbers of classes. Using this table, the estimate of $\rho$ by the product moment method is 0.5157. Obviously, the assumption of normal correlation is plausible here; this has been tested by Lancaster [3], and hence the Pearson $\phi^2$ method is relevant. It is found that $\chi^2_{299} = 763.02$ so that the corrected $\phi^2$ of (2) is 0.3445, yielding an estimate of $R = 0.5062$. The Pearson $\phi^2$ method has its more important applications when the marginal variables are not metrical. The example has been chosen, however, because the value of $\rho$ obtained by the product moment method can be compared with those given by other methods of estimation and the effects of pooling of the classes of the marginal variables can be studied.

Now let the rows and columns of Table 1 be numbered serially 1, 2, $\cdots$, 18 and let $i, j, k, \cdots, l$; denote that the rows (or columns) $i, j, k, \cdots, l$ are grouped into one row (or column). Let us name the grouping of the classes as follows.

- **A**: 1, 2; 3, 4; $\cdots$; 13, 14; 15 to 18
- **B**: 1 to 3; 4, 5; 6, 7; $\cdots$; 14, 15; 16 to 18
- **C**: 1, 2, 17, 18; 3, 16; 4, 15; $\cdots$; 8, 11; 9, 10
- **D**: 1 to 6; 7 to 10; 11 to 18
- **E**: 1 to 7; 8 to 11; 12 to 18
- **F**: 1 to 8; 9 to 18