A CONNECTION BETWEEN CORRELATION AND CONTINGENCY

BY H. O. HIRSCHFELD, Fitzwilliam House

[Communicated by MR J. WISHART]

[*Received* 7 May, *read* 28 October 1935]

INTRODUCTION

Let us consider a discontinuous bivariate distribution. That is, let us consider $N \times Q$ non-negative values $p_{\nu q}$ ($\nu = 1, 2, \ldots, N$; $q = 1, 2, \ldots, Q$), being the theoretical probabilities of the νth value of a variate X_μ ($\mu = 1, 2, \ldots, N$) concurring with the qth value of a second variate Y_s ($s = 1, 2, \ldots, Q$).

It is well known that the correlation theory for such a distribution gives much better results, if both regressions are linear, and that these regressions are transformed by a change of the variates X_ν, Y_q. On the other hand the original scales or better values assigned to the X_ν and Y_q are often chosen in a conventional or artificial way and, if a distribution of characteristics is treated, they are not known at all. Thus the following question naturally arises: Given a discontinuous distribution $p_{\nu q}$, is it always possible to introduce (instead of the original variates, if there are any) new values for the variates x_ν, y_q, such that *both* regressions are linear?

In this note we show that there are indeed at least $[\max(N, Q) - 1]$ different possibilities of introducing variates of the above kind*. We are concerned with an "infinite population" only, and, in this note, shall not enter into the problem of how these variates x_ν, y_q are affected if we replace the probabilities $p_{\nu q}$ by suitable estimations from a finite sample drawn from the population. On the other hand, we point out an interesting relation between population *parameters*. It concerns the correlation coefficients belonging to the fixed frequency distribution $p_{\nu q}$ and to each of the different sets of above variates x_ν, y_q respectively. We show that the square-mean of these correlation coefficients is equal to Pearson's Mean Square Contingency.

Without loss of generality we assume

$$\sum_{q=1}^{Q} p_{\nu q} > 0 \text{ for all } \nu = 1, 2, \ldots, N; \quad \sum_{\nu=1}^{N} p_{\nu q} > 0 \text{ for all } q = 1, 2, \ldots, Q;$$

$$\sum_{\nu=1}^{N} \sum_{q=1}^{Q} p_{\nu q} = 1; \quad 2 \leqslant Q,\ 2 \leqslant N.$$

* However, at most $[\min(N, Q) - 1]$ and at least one possibility is of practical use.

(i) Suppose that we know already a set of variates x_ν, y_q, which together with the given distribution $p_{\nu q}$ yield linear regressions. For these we obtain, by introducing the means

$$(a)\quad m_x = \sum_{\nu=1}^{N}\sum_{q=1}^{Q} x_\nu p_{\nu q}, \qquad (b)\quad m_y = \sum_{\nu=1}^{N}\sum_{q=1}^{Q} y_q p_{\nu q}, \tag{1}$$

the following equations:

$$\left.\begin{aligned} (a)\quad & \lambda_1 (x_\nu - m_x) = \sum_{q=1}^{Q} p_{\nu q}(y_q - m_y) \Big/ \sum_{s=1}^{Q} p_{\nu s}; \quad \nu = 1, 2, \ldots, N, \\ (b)\quad & \lambda_2 (y_q - m_y) = \sum_{\tau=1}^{N} p_{\tau q}(x_\tau - m_x) \Big/ \sum_{\alpha=1}^{N} p_{\alpha q}; \quad q = 1, 2, \ldots, Q. \end{aligned}\right\} \tag{2}$$

Now, if one of the values λ_1 and λ_2 (say λ_2) does not vanish, then by substituting $(2b)$ in $(2a)$ and putting $\rho^2 = \lambda_1\lambda_2$ we have

$$\rho^2 (x_\nu - m_x) = \sum_{q=1}^{Q} p_{\nu q} \left\{ \frac{\sum\limits_{\tau=1}^{N} p_{\tau q}(x_\tau - m_x)}{\sum\limits_{\alpha=1}^{N} p_{\alpha q}} \right\} \frac{1}{\sum\limits_{s=1}^{Q} p_{\nu s}}; \quad \nu = 1, 2, \ldots, N. \tag{3}$$

This may be written as*

$$\begin{aligned} & \rho^2 (x_\nu - m_x) \sqrt{\sum_{s=1}^{Q} p_{\nu s}} \\ & = \sum_{\tau=1}^{N} \left\{ \sum_{q=1}^{Q} \left[\frac{p_{\nu q} p_{\tau q}}{\sum\limits_{\alpha=1}^{N} p_{\alpha q}} \right] \frac{1}{\sqrt{\sum\limits_{s=1}^{Q} p_{\nu s}} \sqrt{\sum\limits_{s=1}^{Q} p_{\tau s}}} \right\} (x_\tau - m_x) \sqrt{\sum_{s=1}^{Q} p_{\tau s}}. \end{aligned} \tag{4}$$

(ii) Conversely, if a set of N values x_ν together with an m_x and a ρ satisfies (4) and $(1a)$, it gives rise to a set x_ν, y_q, λ_1, $\lambda_2 > 0$ satisfying (1) and (2), i.e. having linear regressions†. For from (4) we infer (3) and, defining Q values y_q by $(2b)$, putting $m_y = 0$ and $\lambda_2 = 1$‡, we infer $(2a)$ with $\lambda_1 = \rho^2$. To show that $(1b)$ is true, we observe that by $(2b)$ and $(1a)$ (if $m_y = 0$, $\lambda_2 = 1$)

$$\begin{aligned} \sum_{\nu=1}^{N}\sum_{q=1}^{Q} p_{\nu q} y_q &= \sum_{q=1}^{Q} \left(\sum_{\nu=1}^{N} p_{\nu q} \sum_{\tau=1}^{N} p_{\tau q}(x_\tau - m_x) \Big/ \sum_{\alpha=1}^{N} p_{\alpha q} \right) \\ &= \sum_{q=1}^{Q}\sum_{\tau=1}^{N} p_{\tau q}(x_\tau - m_x) = \sum_{q=1}^{Q}\sum_{\tau=1}^{N} p_{\tau q} x_\tau - m_x = 0. \end{aligned}$$

* If $\lambda_1 \neq 0$ *and* $\lambda_2 \neq 0$, then in addition to equation (4) a corresponding equation (4′), obtained by interchanging in (4) $x_\nu - m_x$ and $y_q - m_y$, N and Q, and Greek and Latin indices, is found from (2) by eliminating the $x_\nu - m_x$ instead of the $y_q - m_y$.

† It will be obvious by an argument parallel to the following one, that equations $(1b)$ and (4′) [see footnote * above] would be sufficient conditions for (1) and (2) as well. Thus we may assume from now on that $N \geqslant Q$.

‡ Thus, if $\rho^2 = 0$, then all $y_q = 0 = m_y$. But we shall show later that we can always find at least one solution of (4) and $(1a)$ with $\rho^2 > 0$ except in the case of absolute non-correlation, i.e. if $p_{\nu q} = P_\nu \tau_q$. In this case, however, the problem is trivial.

(iii) Thus our task is reduced to solving (1a) and (4). We first leave out condition (1a) and consider only condition (4) for arbitrary x_ν, m_x and ρ^2. Introducing

$$\xi_\nu = (x_\nu - m_x) \sqrt{\sum_{s=1}^{Q} p_{\nu s}}$$

and writing $A_{\nu\tau}$ for the term inside the { } in (4), we see that this condition is a characteristic-value problem

$$\rho^2 \xi_\nu = \sum_{\tau=1}^{N} A_{\nu\tau} \xi_\tau$$

for the symmetrical matrix $(A_{\nu\tau})$. Now it is well known that the greatest of the N characteristic-values of $(A_{\nu\tau})$ may be obtained by maximizing the quadric

$$K(\xi, \xi) = \sum_{\nu,\tau=1}^{N} A_{\nu\tau} \xi_\nu \xi_\tau$$

among normalized sets, i.e. sets satisfying

$$\sum_{\nu=1}^{N} \xi_\nu^2 = 1, \tag{5}$$

by a set $\xi_\nu^{(1)}$ ($\nu = 1, 2, \ldots, N$) say, called the first characteristic set. It is furthermore known, that then the second (ith $\{i = 2, 3, \ldots, N\}$) characteristic-value and corresponding set $\xi_\nu^{(2)}$ ($\xi_\nu^{(i)}$) may be obtained by maximizing $K(\xi, \xi)$ among all ξ_ν satisfying (5) and orthogonal to $\xi_\nu^{(1)}$ ($\xi_\nu^{(1)}, \xi_\nu^{(2)}, \ldots, \xi_\nu^{(i-1)}$), i.e. for which

$$\sum_{\nu=1}^{N} \xi_\nu^{(1)} \xi_\nu = 0 \left(\sum_{\nu=1}^{N} \xi_\nu^{(1)} \xi_\nu = 0, \ldots, \sum_{\nu=1}^{N} \xi_\nu^{(i-1)} \xi_\nu = 0 \right) \tag{6}$$

is true in addition to (5).

If we apply this process to our special matrix $(A_{\nu\tau})$, condition (5) becomes

$$\sum_{\nu=1}^{N} (x_\nu - m_x)^2 \sum_{s=1}^{Q} p_{\nu s} = \sum_{\nu=1}^{N} \sum_{s=1}^{Q} p_{\nu s} (x_\nu - m_x)^2 = 1. \tag{7}$$

To estimate the quadric $K(\xi, \xi)$ we write

$$K(\xi, \xi) = \sum_{\nu,\tau=1}^{N} \sum_{q=1}^{Q} \frac{p_{\tau q} p_{\nu q} \sqrt{\sum_{s=1}^{Q} p_{\nu s}} \sqrt{\sum_{s=1}^{Q} p_{\tau s}}}{\sum_{\alpha=1}^{N} p_{\alpha q} \sqrt{\sum_{s=1}^{Q} p_{\nu s}} \sqrt{\sum_{s=1}^{Q} p_{\tau s}}} (x_\nu - m_x)(x_\tau - m_x),$$

$$K(\xi, \xi) = \sum_{q=1}^{Q} \left(\sum_{\nu=1}^{N} p_{\nu q} (x_\nu - m_x) \right)^2 \Bigg/ \sum_{\alpha=1}^{N} p_{\alpha q}. \tag{8}$$

Hence $K(\xi, \xi) \geqslant 0$, but, applying Schwarz's inequality to $(\sum_\nu)^2$ in (8) (for each q separately), we obtain by (7)

$$K(\xi, \xi) \leqslant \sum_{q=1}^{Q} \frac{\sum_{\nu=1}^{N} p_{\nu q} \sum_{\nu=1}^{N} p_{\nu q} (x_\nu - m_x)^2}{\sum_{\alpha} p_{\alpha q}} \leqslant \sum_{q=1}^{Q} \sum_{\nu=1}^{N} p_{\nu q} (x_\nu - m_x)^2 = 1, \tag{9}$$

where the sign of equality is true only if $(x_\nu - m_x)\sqrt{p_{\nu q}} = c(q)\sqrt{p_{\nu q}}$, i.e. if

$$x_\nu - m_x = c(q). \tag{10}$$

Thus we see that, under condition (5),

$$0 \leqslant K(\xi, \xi) \leqslant 1 \tag{11}$$

and that $K(\xi, \xi) = 1$ if and only if

$$\left.\begin{array}{l} (a)\ \text{Either } x_\nu - m_x = 1 \text{ for all } \nu = 1, 2, \ldots, N, \\ (b)\ \text{Or if the variate } x_\nu - m_x \text{ can be made a unique function } c(q) \\ \qquad \text{of the index } q. \end{array}\right\} \tag{12}$$

Thus by (11) and (12*a*) 1 is in any case the greatest characteristic-value and

$$\xi_\nu^{(1)} = \sqrt{\sum_{q=1}^{Q} p_{\nu q}} \tag{13}$$

is a characteristic set belonging to it. Now for all remaining characteristic sets $\xi_\nu^{(i)}$ $(i = 2, \ldots, N)$ orthogonality to the first set $\xi_\nu^{(1)}$ (at least) is necessary. Writing down this condition, we have, by (13),

$$0 = \sum_{\nu=1}^{N} \xi_\nu \sqrt{\sum_{q=1}^{Q} p_{\nu q}} = \sum_{\nu=1}^{N} \sum_{q=1}^{Q} (x_\nu - m_x) p_{\nu q} = \sum_{\nu=1}^{N} \sum_{q=1}^{Q} x_\nu p_{\nu q} - m_x. \tag{14}$$

But (14) turns out to be our equation (1*a*). Therefore from now m_x has to be the mean of the x_ν with regard to the distribution $p_{\nu q}$, if the set $(x_\nu - m_x)\sqrt{\sum_{s=1}^{Q} p_{\nu s}}$ is to be admitted to the competition in maximizing $K(\xi, \xi)$ and to become one of the characteristic sets $\xi_\nu^{(i)}$ $(i = 2, 3, \ldots, N)$. Thus we now define, with the help of these $\xi_\nu^{(i)}$ and an arbitrary m_x, the $N - 1$ sets

$$x_\nu^{(i)} = \left(\xi_\nu^{(i)} \Big/ \sqrt{\sum_{s=1}^{Q} p_{\nu s}}\right) + m_x; \quad \nu = 1, 2, \ldots, N; \quad i = 2, \ldots, N.$$

These sets will fulfil conditions (1*a*) as well as (4), if we take here for ρ^2 the ith characteristic-value $(\rho^{(i)})^2 = K(\xi^{(i)}, \xi^{(i)})$, for which, by (11),

$$0 \leqslant (\rho^{(i)})^2 \leqslant (\rho^{(2)})^2 \leqslant 1.$$

They are the $N - 1$ solutions* desired and may be constructed by the maximizing principle described above or by finding the roots of the characteristic polynomial $|A_{\nu\tau} - \rho^2 E| = 0$ and solving (4).

(iv) For our statistical purposes we are interested in the quadric $K(\xi, \xi)$ under the condition (14) only, i.e. if m_x is the distribution mean of the x_ν. But then $K(\xi, \xi)$ has a proper meaning too, for we infer from (8) that

$$K(\xi, \xi) = \sum_{q=1}^{Q} \sum_{a=1}^{N} p_{\alpha q} \left\{\left(\sum_{\nu=1}^{N} p_{\nu q} x_\nu \Big/ \sum_{a=1}^{N} p_{\alpha q}\right) - m_x\right\}^2$$

* See p. 521, note †. If $N > Q$, then at least $N - Q$ of the $(\rho^{(i)})^2$ must vanish. For if $(\rho^i)^2 > 0$, then by (ii) $\lambda_2 = 1$, $\lambda_1 = (\rho^{(i)})^2 > 0$, but there are at most Q (linearly independent) characteristic sets of (4′) (see p. 521, note *). Furthermore it may occur that for some $i = 2, \ldots, N$ two coordinates $\xi_\nu^{(i)}$ and $\xi_\mu^{(i)}$ are equal. However, this property is of no statistical relevance, since it depends on the exact values of the $p_{\nu q}$. The same applies if two characteristic-values are equal.

and thus that K is equal to the square of the correlation ratio of our distribution*. The $N-1$ correlation ratios $(\rho^{(i)}) = \sqrt{K}\,(\xi^{(i)}, \xi^{(i)})$, $i=2, \ldots, N$, are at the same time correlation coefficients corresponding to the $p_{\nu q}$, $x_\nu^{(i)}$ and $y_q^{(i)}$ (defined by (2b)) in virtue of the linear regressions. In particular $(\rho^{(2)})$ is the greatest correlation coefficient that can be constructed with the fixed distribution $p_{\nu q}$. By (12b) we see that

$$\rho^{(2)} = 1 \tag{15}$$

is true if and only if there is perfect dependence in the $p_{\nu q}$ distribution, since for (12a) the condition (14) does not hold. We now show that

$$\rho^{(2)} = 0$$

is true if and only if there is perfect non-correlation. It is easily verified that $K(\xi, \xi) = 0$ for any distribution of the form $p_{\nu q} = P_\nu \tau_q$ and variates x_ν satisfying conditions (7) and (14). Hence $\rho^{(2)} = 0$. Conversely, if $\rho^{(2)} = 0$, i.e. if $K = 0$ for all $x_\nu - m_x$ satisfying (7) and (14), then, by (8),

$$\sum_{\nu=1}^{N} p_{\nu q}(x_\nu - m_x) = 0 \tag{16}$$

has to be true for all $q = 1, 2, \ldots, Q$ and all $x_\nu - m_x$ for which (14) holds. Hence (16) must have rank not greater than 1, i.e. $p_{\nu q} = P_\nu \tau_q$.

$(\rho^{(2)})^2$ would be a very general contingency and correlation measure common to statistics of characteristics as well as to ordinary bivariate distributions having the advantages of linear correlation theory. It could be defined for continuous distributions as well. However, the complexity of its construction makes its practical use nearly impossible. Searching for simpler parameters we naturally examine the symmetrical functions of the set

$$1, (\rho^{(2)})^2, (\rho^{(3)})^2, \ldots, (\rho^{(N)})^2, \tag{17}$$

since these can be expressed in terms of the $A_{\nu\tau}$ and thus of the $p_{\nu q}$.

In particular we examine the arithmetic mean of our $N-1$ squared correlation-coefficients

$$f^2 = 1/(N-1)\{(\rho^{(2)})^2 + \ldots + (\rho^{(N)})^2\}.$$

Now this is equal to Pearson's Mean Square Contingency. For, since $(\rho^{(1)})^2 = 1$, we have†

$$\begin{aligned}(N-1)f^2 &= \sum_{i=1}^{N} (\rho^{(i)})^2 - 1 = \left(\sum_{\nu=1}^{N} A_{\nu\nu}\right) - 1 = \sum_{\nu=1}^{N}\sum_{q=1}^{Q} \left\{p_{\nu q}^2 \Big/ \sum_{\alpha=1}^{N} p_{\alpha q} \sum_{s=1}^{Q} p_{\nu s}\right\} - 1 \\ &= \sum_{\nu=1}^{N}\sum_{q=1}^{Q} \left\{\left[p_{\nu q} - \sum_{\alpha=1}^{N} p_{\alpha q} \sum_{s=1}^{Q} p_{\nu s}\right]^2 \Big/ \sum_{\alpha=1}^{N} p_{\alpha q} \sum_{s=1}^{Q} p_{\nu s}\right\}. \end{aligned} \tag{18}$$

* Provided that the variance (7) is 1, which is true for the $x_\nu^{(i)}$.

† Usually Pearson's Mean Square Contingency is only defined for $N = Q$. But if $N > Q$ (say) there are at least two possibilities of defining a measure having its properties: one is given by (18), another would have $Q-1$ instead of $N-1$ (as the corresponding argument for the y_q would show).