Null Correlation for Proportions

John N. Darroch

INTRODUCTION

Consider a set \{X_1, X_2, \ldots, X_{t-1}, s-X_1-X_2-\ldots-X_{t-1}\} of \( t \) nonnegative random variables whose sum \( s \) is a fixed number. For example, a randomly selected volume \( s \) of rock or blood is analyzed according to its component minerals or sera yielding, in each situation, \( t \) nonnegative, real-valued random variables. Alternatively, a random sample of \( s \) moths or pollen grains is classified according to its \( t \) types yielding \( t \) non-negative, integer-valued random variables.

We are concerned in this paper with the problem of describing “null correlation” among such variables, a problem of particular importance in geology in the study of paragenesis. Precisely, we wish to find the value of the correlation coefficient \( \rho(X_1, X_2) \), which corresponds to zero correlation in the usual situation of unconstrained variables. The correlation coefficient of \( X_1, X_2 \) is, of course, the same as the correlation coefficient of the proportions \( Y_1 = \frac{X_1}{s}, Y_2 = \frac{X_2}{s} \).

A more general version of this problem occurs when \( X_i = \sum_{i=1}^{t} X_i = S \) is a random variable, reflecting the experimenter’s sampling procedure. What is at question now is the null value of \( \rho(Y_1, Y_2) \), where \( Y_1 = \frac{X_1}{S}, Y_2 = \frac{X_2}{S} \). This correlation coefficient is, in general, different from \( \rho(X_1, X_2) \).

The relevant literature is found mainly in geological journals. A null value of \( \rho(Y_1, Y_2) \) is derived from some simple assumptions, and the corresponding multivariate results are given. We refer to the null value of a correlation coefficient as “the correlation due to the constraint.”

The results presented here are related closely to the author’s work on null dependence, which will be published separately.

REVIEW OF LITERATURE

Let \( Y_1, Y_2, \ldots, Y_t \) denote the proportions \( \frac{X_1}{S}, \frac{X_2}{S}, \ldots, \frac{X_t}{S} \).

Chayes (1960) showed that, if \( Y_1, Y_2, \ldots, Y_t \) all have equal variances, then the
average value of $\rho(Y_i, Y_j)$ over all pairs $(i, j)$, $i \neq j$, is $-(t-1)^{-1}$. The result follows from the expansion of the left side of the identity $\text{var} \left( \sum_{i=1}^{t} Y_i \right) = 0$.

Sarmanov (1961), independently of Chayes, made essentially the same observation. Namely, that, if $X_1, X_2, \ldots, X_t$ are distributed symmetrically (exchangeable), then so are $Y_1, Y_2, \ldots, Y_t$ and $\rho(Y_i, Y_j) = -(t-1)^{-1}$, $i \neq j$. He showed further that, if

$$X_i = \sum_{a=1}^{n_i} X_{ia}, \quad i = 1, 2, \ldots, t$$

where the $n_1 + n_2 + \ldots + n_t$ variables $\{X_{ia}\}$ are symmetrically distributed, then

$$\rho(Y_i, Y_j) = -\frac{E[Y_i]E[Y_j]}{\sqrt{(1-E[Y_i])(1-E[Y_j])}}, \quad i \neq j$$

This is the null value that we shall derive. The difficulty with Sarmanov’s derivation of eq (1) is that seemingly there is no connection between the properties “the correlation of $Y_i$, $Y_j$ is due only to the constraint” and “$X_i$ and $X_j$ are sums of symmetrically distributed random variables.”

Mosimann (1962) was concerned with statistical dependence due to the constraint as well as correlation due to the constraint. He noted, independently of Sarmanov, that when $X_1, X_2, \ldots, X_t; \sum_{i=1}^{t} X_i = s$, have the Dirichlet (multivariate beta) distribution or the multinomial distribution or certain compounds of the multinomial distribution (including the Dirichlet compound) the correlation coefficient of $Y_i, Y_j$ is given by eq (1). He suggested that eq (1) is probably of wide validity.

Chayes and Kruskal (1966) proposed the following model for $Y_1, Y_2, \ldots, Y_t$. Let $U_1, U_2, \ldots, U_t$ be uncorrelated (in the usual sense) random variables and let $V_1, V_2, \ldots, V_t$ denote the proportions obtained from them, that is,

$$V_i = U_i/\sum_{j=1}^{t} U_j, \quad i = 1, 2, \ldots, t$$

Now regard the joint distribution of $V_1, V_2, \ldots, V_t$ as a model for the joint distribution of $Y_1, Y_2, \ldots, Y_t$. The $U$’s are called open variables and the $V$’s closed variables. (This approach is similar to Mosimann’s approach to dependence due to the constraint. He takes the $U$’s to be independent.) Chayes and Kruskal neglect the third-order and higher moments of the $U$’s to obtain the approximate formula

$$\rho(V_i, V_j) = \gamma_{ij}/\sqrt{\gamma_{ii}\gamma_{jj}}, \quad i \neq j$$

where

$$\gamma_{ij} = E[V_i]E[V_j] \sum_{i=1}^{t} \text{var}[U_i] - E[V_i]\text{var}[U_j] - E[V_j]\text{var}[U_i] + \delta_{ij}\text{var}(U)$$

If $\text{var}(U_i) = \alpha E[V_i], i = 1, 2, \ldots, t$, then formula (2) is equivalent to eq (1).

Two related papers by Sarmanov and Vistelius (1959, 1961) are not so much