CORRELATION COEFFICIENTS FOR MORE THAN ONE SCALE TYPE: AN ALTERNATIVE TO THE JANSON AND VELELIUS APPROACH

FRITS E. ZEGERS AND JOS M. F. TEN BERGE
UNIVERSITY OF GRONINGEN

Janson and Vegelius have recently suggested a family of correlations for variables of mixed scale types, including nominal scales. The resulting correlations are E-coefficients, which means that they are unity if the variables involved are identical up to permissible transformations, and that they can be considered as inner products in a Euclidian space. Some of the coefficients of the correlation family suggested by Janson and Vegelius are generalized squared product-moment correlations and some are not. In the present paper, a family of correlations for variables of mixed scale types is advocated all members of which are generalized squared product-moment correlations. Some practical advantages of the latter family are explained.

Key words: mixed levels of measurement; association coefficients.

The problem how to measure similarity between variables has received attention for many years. Quite a number of coefficients have been suggested which preserve certain characteristics of variables while ignoring others, thus being appropriate for certain levels of measurement of "scale types." For instance, Kendall's tau (Kendall, 1955) preserves only ordinal information of the variables, hence it is a measure of similarity for ordinal scales.

A vast number of coefficients satisfy the additional requirements that the similarity between a variable and itself be 1, and that, for an arbitrary set of k variables of a given scale type, the k x k matrix of similarity coefficients between these variables be Gramian (symmetric and positive semi-definite). Coefficients which satisfy these additional requirements have been christened E-coefficients by Vegelius (1978), because they can be interpreted as scalar products in a Euclidian space. E-coefficients have several attractive features (Janson & Vegelius, 1982a, p. 272), including the possibility of being subjected to principal components analysis (Janson & Vegelius, 1978a).

Whereas the literature on coefficients for variables of the same scale type is quite abundant, very little work seems to have been done on the problem of assessing similarity between variables of different scale types. However, a major contribution in this area has been made by Janson and Vegelius (1982a). They constructed a "family of E-correlations" between variables of various scale types, including the nominal scale type. If this family of coefficients is applied to an arbitrary set of k variables of various scale types, the resulting k x k matrix of coefficients is Gramian and has unit diagonal elements.

It should be noted that the family of E-correlations does not rely on optimal scaling, as is the case in nonlinear principal components analysis (Gifi, 1981). As a result, the family of E-correlations provides a method for handling variables of nominal or mixed measurement levels which does not capitalize on chance.

The purpose of the present paper is to advocate an alternative E-correlation family for the case where nominal scales are involved. In fact, this alternative family was anticipated in Janson and Vegelius (1978b) but was ignored in Janson and Vegelius (1982a).
Both the $E$-correlation family adopted by Janson and Vegelius (1982a) and the one advocated here rely on general principles that can conveniently be explained after first considering the case of variables of the same scale type.

**$E$-Coefficients for Variables of the Same Scale Type**

Consider an arbitrary pair of variables $x$ and $y$ of the same scale type. Let the responses of $N$ subjects on $x$ and $y$ be quantified by $N \times m$ matrices $X$ and $Y$, respectively (examples will be given shortly). Then an $E$-coefficient is given, regardless of how $X$ and $Y$ are defined, by

\[
E(x, y) = \text{tr } X'Y(\text{tr } X'X)^{-1/2}(\text{tr } Y'Y)^{-1/2},
\]

(1) (Janson & Vegelius, 1982a). Clearly, $E(x, y)$ can be interpreted as the inner product between two vectors in $Nm$-space, which have as coordinates the elements of $X$ and $Y$, respectively, scaled to unit sums of squares. In fact (1) gives Tucker's congruence (Tucker, 1951) between these vectors, which reduces to the product-moment correlation (PMC) when $X$ and $Y$ have zero means.

The matrix $X$ shall be referred to as the "primary quantification for variable $x$" in the sequel. Specific primary quantifications generate specific $E$-coefficients. For interval scales, the primary quantification can conveniently be taken as the $N \times 1$ vector of deviation scores, thus rendering $E(x, y)$ equal to the PMC. For ordinal scales, one may conveniently adopt the $N \times 1$ vectors of ranks, expressed in deviations from their means, as primary quantifications, which renders $E(x, y)$ equal to Spearman's rho. An alternative quantification for ordinal scales, which generates Kendall's tau, has been suggested by Daniels (1944).

The problem how to quantify responses to nominal scales is crucial in the present context. Janson and Vegelius (1979) have emphasized two approaches, both of which are of particular interest. First, one may define $X$ for a nominal scale $x$ as the $N \times m$ matrix with elements $X_{ij} = m - 1$ if the response of subject $i$ is in category $j$, and $X_{ij} = -1$ otherwise, $i = 1, \ldots, N; j = 1, \ldots, m$. This primary quantification will be referred to as the "row-centered quantification," for obvious reasons. The resulting $E(x, y)$ is the $C$-coefficient, introduced by Janson and Vegelius (1979) as a generalized G-index of agreement (Holley & Guilford, 1964).

Second, one may define $X$ for a nominal scale $x$ as the $N \times m$ matrix with elements $X_{ij} = N - f_j$, where $f_j$ is the response frequency of category $j$ of $x$, if the response of subject $i$ is in that category, and $X_{ij} = -f_j$, otherwise, $i = 1, \ldots, N; j = 1, \ldots, m$. This quantification will be referred to as the "double-centered quantification," for obvious reasons. The resulting $E(x, y)$ is the $S$-coefficient, introduced by Janson and Vegelius (1979) as a generalization of the phi coefficient (the PMC between dichotomous variables).

Clearly, other quantifications than the row-centered and the double-centered quantification could be considered, but are beyond the scope of the present paper.

Expressing the various coefficients as special cases of (1) applied to primary quantification matrices is by no means attractive from a computational point of view. Computational short-cut formulas are available but are of no concern at this point.

**$E$-Correlation Families for Variables of Different Scale Types**

The coefficients $E(x, y)$ defined in (1) do not apply to variables of different scale types, for two reasons. First, $\text{tr } X'Y$ is only defined if $X$ and $Y$ have the same order, a condition which is not generally met when $x$ and $y$ have different scale types. Second, $E(x, y)$ must be insensitive to transformations that are permissible for the scale types involved. For nominal scales, the primary quantifications discussed above are sensitive to permutations.