KYBERNETIKA — VOLUME 22 (1986), NUMBER 2

CLASSIFICATIONS WITH RELATIONS: A MODEL FOR THE DESCRIPTION OF DISTRIBUTIONS AND THEIR DISTANCES

JAN ŘEHÁK, BLANKA ŘEHÁKOVÁ

Distributions on classifications are met wherever we work with categorical variables. A vast investigation has been done in developing methods for statistical analysis of nominal variables (e.g. variables with simple classification), partially are solved also problems for ordered classifications and classifications with assigned numbers. In this paper we propose a general model which enables us to develop descriptive measures for distributions on various types of classifications with relations. The simple, ordered and quantitative classifications will be special cases of this general model. In this way, it is possible to handle with general decision and predictive models as well as with the analysis of generalized categorical variables.

1. BASIC DEFINITIONS: A GENERALIZED CATEGORICAL VARIABLE

The presented model (we call it the *D-model*) includes classifications with binary relations on the set of categories. A set of categories $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K\}$ is said to be a classification if the categories form a mutually exclusive system of events whose union covers all possibilities of the classification process. The distribution $\mathbf{f}$ on a classification $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K\}$ is a column vector from the simplex $\mathbf{Q}_K = \{\mathbf{p}: \mathbf{p} = (p_1, \ldots, p_K)', p_k \geqq 0 \ (k = 1, \ldots, K), \sum_{k=1}^{K} p_k = 1\}$. We restrict ourselves to the class of relations which can be expressed by numerical values. This class has been found meaningful with regard to the data analysis.

Definition 1. Let $\mathbf{D} = \|d_{ij}\| = \|d(\boldsymbol{a}_i, \boldsymbol{a}_j)\| = \|d(i, j)\|$ be a real, square matrix of order K. Matrix $\mathbf{D}$ is said to be the *matrix of scores generating the type* of the variable $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K\}$, if the following conditions hold:

a) *identity*: $d_{ii} = 0$ for $i = 1, \ldots, K$;

b) *symmetry*: $d_{ij} = d_{ji}$ for $i, j = 1, \ldots, K$;

c) *nonnegativity*: $d_{ij} \geqq 0$ for $i, j = 1, \ldots, K$ and $d_{ij} > 0$ for at least one pair (i, j);

d) *interpretability*: the more unlike categories $\boldsymbol{a}_i$, $\boldsymbol{a}_j$ are, the greater is the score $d(\boldsymbol{a}_i, \boldsymbol{a}_j)$ characterizing their dissimilarity.

Elements of $\mathbf{D}$ are called the scores of distances or dissimilarity scores of the categories $(\boldsymbol{a}_i, \boldsymbol{a}_j)$. Properties a)–d) correspond to meaningful practical requirements on d_{ij}.

A general categorical variable (*a classification with numerical relations, D-generalized categorical variable*) is given by a list of its values and by a matrix of scores generating its type:

$$\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\} .$$

The basic types of variables as they occur in practice have the models stated by Definition 2.

Definition 2. We say that $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$ is

a) *a nominal variable* (*a simple classification*) if $d_{ij} = 1$ for all $i \neq j$; we denote it also $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K\}$;

b) *a discrete ordinal variable* (*an ordered classification*) if $d_{ij} = |i - j|$ for all i, j; we denote it also $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}_0\}$;

c) *a discrete cardinal variable* (*a numerical or quantified classification*) if its values are numbers $x_1, \ldots, x_K$ assigned to categories by the mapping $x_i = x(\boldsymbol{a}_i)$, and $d_{ij} = (x_i - x_j)^2$ for all i, j; we denote it also $\boldsymbol{X} = \{x_1, \ldots, x_K\}$ or $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}_{\boldsymbol{X}}\}$.

2. BASIC CHARACTERISTICS OF THE DISTRIBUTION OF A GENERALIZED CATEGORICAL VARIABLE

The definition of a variability is based on C. Gini's idea: the variance is taken as the expected dissimilarity among pairs of independently repeated random events.

Definition 3. We define *the generalized variance* of a distribution $\mathbf{f} \in \boldsymbol{Q}_K$ of $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$ as

$$\text{(1)} \qquad \operatorname{Gvar}_{\mathbf{D}} \mathbf{f} = \operatorname{Gvar} \mathbf{f} = \sum_{i=1}^{K} \sum_{j=1}^{K} f_i f_j d_{ij} = \mathbf{f}'\mathbf{D}\mathbf{f} = \mathop{\mathsf{EE}}_{i\,j}(d_{ij}) .$$

Properties of the generalized variance are summed up in Theorem 1.

Theorem 1. (*Properties of the generalized variance.*)

a) Let $d_{ij} > 0$ for all $i \neq j$. $\operatorname{Gvar} \mathbf{f} = 0$ if, and only if, there exists $f_i = 1$ $(1 \leqq i \leqq K)$.

b) Let $\mathbf{D}$ be a general matrix. Then $\operatorname{Gvar} \mathbf{f} = 0$, if, and only if $\sum_{i \in \boldsymbol{W}} f_i = 1$ for some set $\boldsymbol{W} = \{i: d(\boldsymbol{a}_i, \boldsymbol{a}_j) = 0;\ i \in \boldsymbol{W},\ j \in \boldsymbol{W}\}$. (Property a) is a special case in which all sets $\boldsymbol{W}$ are singletons.)

c) Define d as $\max_{(i,j)} d_{ij}$ and let $t > 1$ be the size of an index set $\boldsymbol{T}$ for which it holds that the respective submatrix $\mathbf{D}_{\boldsymbol{T}}$ of order t has the elements $d_{ij} = (1 - \delta_{ij})\, d$, where δ_{ij} is the Kronecker delta and let there exists no submatrix of a higher order with this property. Let $d_{iy} < d$ when $(i, j) \notin \boldsymbol{T} \times \boldsymbol{T}, i \neq j$. Then the generalized variance attains its maximal value

$$\max_{\mathbf{f} \in \mathbf{Q}_K} \operatorname{Gvar} \mathbf{f} = \operatorname{Gvar} \mathbf{f}_{\max} = \frac{t-1}{t}\, d$$

for the distribution $\mathbf{f}_{\max}$ with the components $f_i = 1/t$ for $i \in \boldsymbol{T}$ and $f_i = 0$ for $i \notin \boldsymbol{T}$.

Proof. The property a) is obvious.

b) Let $\operatorname{Gvar} \mathbf{f} = 0$ and let there exist positive f_i, f_j and d_{ij}. Then $\operatorname{Gvar} \mathbf{f} \geqq 2 f_i f_j d_{ij} > 0$ and the contradiction follows. Therefore under the assumption that $\operatorname{Gvar} \mathbf{f} = 0$ it must be true that $\mathbf{f}$ has nonzero components only on such a subset of indices $\boldsymbol{W}$ for which $d_{ij} = 0$ for all pairs $i, j \in \boldsymbol{W}$. Let $\boldsymbol{W}$ be a subset of $\{1, 2, \ldots, K\}$ with $d_{ij} = 0$ for all pairs $(i, j) \in \boldsymbol{W}$ and $\sum_{k \in \boldsymbol{W}} f_k = 1$. Then

$$\operatorname{Gvar} \mathbf{f} = 2 \sum_{\substack{i<j \\ \{(i,j):\, d_{ij} \neq 0\}}} \sum f_i f_j d_{ij}\,.$$

Under the assumption $d_{ij} \neq 0$ there are three possibilities: either $i \in \boldsymbol{W}$ and $j \notin \boldsymbol{W}$, or $i \notin \boldsymbol{W}$ and $j \in \boldsymbol{W}$, or $i \notin \boldsymbol{W}$ and $j \notin \boldsymbol{W}$. It follows that $f_j = 0$, $f_i = 0$, $f_i = f_j = 0$ respectively. Therefore $\operatorname{Gvar} \mathbf{f} = 0$.

c)
$$\operatorname{Gvar} \mathbf{f} = \sum_{i=1}^{K} \sum_{j=1}^{K} f_i f_j d_{ij} = \sum_{i \in \boldsymbol{T}} \sum_{j \in \boldsymbol{T}} f_i f_j d_{ij} + 2 \sum_{i \in \boldsymbol{T}} \sum_{j \in \boldsymbol{T}^c} f_i f_j d_{ij} +$$
$$+ \sum_{i \in \boldsymbol{T}^c} \sum_{j \in \boldsymbol{T}^c} f_i f_j d_{ij} = \sum_{\substack{i \in \boldsymbol{T} \\ i \neq j}} \sum_{j \in \boldsymbol{T}} f_i f_j d + 2 \sum_{\substack{i \in \boldsymbol{T} \\ i \neq j}} \sum_{j \in \boldsymbol{T}^c} f_i f_j (d - g_{ij}) +$$
$$+ \sum_{\substack{i \in \boldsymbol{T}^c \\ i \neq j}} \sum_{j \in \boldsymbol{T}^c} f_i f_j (d - g_{ij}) = d \sum_{\substack{i=1 \\ i \neq j}}^{K} \sum_{j=1}^{K} f_i f_j - 2 \sum_{\substack{i \in \boldsymbol{T} \\ i \neq j}} \sum_{j \in \boldsymbol{T}^c} f_i f_j g_{ij} -$$
$$- \sum_{\substack{i \in \boldsymbol{T}^c \\ i \neq j}} \sum_{j \in \boldsymbol{T}^c} f_i f_j g_{ij}\,,$$

where $g_{ij} = d - d_{ij} \geqq 0$, $\boldsymbol{T}^c$ is the complement of $\boldsymbol{T}$. Therefore it is necessary to look for the distribution which maximizes the generalized variance among those distributions $\mathbf{f}$ for which $\sum_{i \in \boldsymbol{T}} f_i = f_{+\boldsymbol{T}} = 1$. Then

$$\operatorname{Gvar} \mathbf{f} = d \sum_{\substack{i \in \boldsymbol{T} \\ i \neq j}} \sum_{j \in \boldsymbol{T}} f_i f_j = d\left[f_{+\boldsymbol{T}}^2 - \sum_{i \in \boldsymbol{T}} f_i^2\right] = d\left[1 - \sum_{i \in \boldsymbol{T}} f_i^2\right].$$

The maximum of this expression is equal to $d(t-1)/t$ and it is attained if $f_i = 1/t$ for all $i \in \boldsymbol{T}$. If there exists a set $\boldsymbol{T}' \supset \boldsymbol{T}$, $t' > t$, then $d(t'-1)/t' > d(t-1)/t$. Therefore, the maximum is obtained for the set $\boldsymbol{T}$ having the property given above and containing the maximal possible number of indices. □

Remark. The normalized measure of variability can be introduced as

$$\operatorname{norm} \operatorname{Gvar} \mathbf{f} = \frac{\operatorname{Gvar} \mathbf{f}}{\operatorname{Gvar} \mathbf{f}_{\max}} .$$

We introduce characterisation of each classificatory value with regard to the distribution $\mathbf{f}$ in the following manner.

Definition 4. Consider a variable $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$. We define the *measure of concentration* about the value $\boldsymbol{a}_k$ as the expected score of the distance from the value $\boldsymbol{a}_k$

$$d_k^* = d_{\boldsymbol{a}_k}^* = \mathop{\mathsf{E}}_{i}[d(\boldsymbol{a}_i, \boldsymbol{a}_k)] = \sum_{i=1}^{K} f_i \, d(\boldsymbol{a}_i, \boldsymbol{a}_k) . \tag{2}$$

Any value $\mathbf{c}$ of the variable $\boldsymbol{A}$ for which

$$d^* = \min_{j} d_{\boldsymbol{a}_j}^* = \sum_{i=1}^{K} f_i \, d(\boldsymbol{a}_i, \mathbf{c})$$

is called *a centre of the distribution* $\mathbf{f}$.

It is obvious that

a) $\operatorname{Gvar} \mathbf{f} = \mathsf{E} d_{\boldsymbol{a}_k}^*$,

b) the distribution can have more centres,

c) $d^* = 0$, only if $f_i \neq 0$ for those i's satisfying $d(\boldsymbol{a}_i, \mathbf{c}) = 0$,

d) $d^* \leqq \operatorname{Gvar} \mathbf{f}$.

Let us denote $\mathbf{C}_f = \mathbf{C}$ the vector of concentrations $(d_1^*, \ldots, d_K^*)'$. We see that $\mathbf{C} = \mathbf{D}\mathbf{f}$ and for a nonsingular matrix $\mathbf{D}$ it holds that $\mathbf{f} = \mathbf{D}^{-1}\mathbf{C}$. In this case we can characterize a set of distributions $\{\mathbf{f}\}$ by means of $\{\mathbf{D}\mathbf{f}\}$ and this can be a useful transformation in the data analysis.

3. METRICS GENERATED BY THE MATRIX $\mathbf{D}$

Consider a real function

$$D(\mathbf{f}, \mathbf{g}) = \sqrt{((\mathbf{f} - \mathbf{g})' \, \mathbf{D}(\mathbf{g} - \mathbf{f}))} \tag{3}$$

on $\mathbf{Q}_K \times \mathbf{Q}_K$. An important class of variables $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$ is determined by Theorem 2 which shows when $D(\mathbf{f}, \mathbf{g})$ can be regarded as a measure of dissimilarity of two distributions $\mathbf{f}$, $\mathbf{g}$ and moreover possesses the properties of metrics.

Theorem 2. (*Existence of metrics.*) The function $D(\mathbf{f}, \mathbf{g})$ is a metrics (semimetrics) on $\mathbf{Q}_K \times \mathbf{Q}_K$ if and only if the matrix $\mathbf{D}^* = \|d_{ij}^*\|$ of order $K - 1$, where

$$d_{ij}^* = d_{iK} + d_{Kj} - d_{ij} \tag{4}$$

is a positive (positive semi-definite) matrix.

Proof. a) Let $\mathbf{S} = \{\mathbf{u} = (u_1, \ldots, u_K)' : \sum_{i=1}^{K} u_i = 0\}$. Let us denote $\mathbf{v} = (u_1, \ldots, u_{K-1})'$. Then

$$\mathbf{u}'\mathbf{D}\mathbf{u} = \sum_{i=1}^{K-1}\sum_{j=1}^{K-1} u_i u_j d_{ij} + u_K \sum_{j=1}^{K-1} u_j d_{Kj} + u_K \sum_{i=1}^{K-1} u_i d_{iK} =$$

$$= \sum_{i=1}^{K-1}\sum_{j=1}^{K-1} u_i u_j (d_{ij} - d_{Kj} - d_{iK}) = -\sum_{i=1}^{K-1}\sum_{j=1}^{K-1} u_i u_j d_{ij}^* = -\mathbf{v}'\mathbf{D}^*\mathbf{v}\,.$$

It is seen that $\mathbf{u}'\mathbf{D}\mathbf{u}$ is a negative semi-definite form on $\mathbf{S}$ if, and only if, $\mathbf{v}'\mathbf{D}^*\mathbf{v}$ is a positive semi-definite form. (A quadratic form $\mathbf{u}'\mathbf{D}\mathbf{u}$ is said to be a positive definite on a set $\mathbf{S}$ if $\mathbf{u}'\mathbf{D}\mathbf{u} \geqq 0$ for all $\mathbf{u} \in \mathbf{S}$ and $\mathbf{u}'\mathbf{D}\mathbf{u} = 0 \Leftrightarrow \mathbf{u} = \mathbf{0}$, if $\mathbf{0} \in \mathbf{S}$. A quadratic form $\mathbf{u}'\mathbf{D}\mathbf{u}$ is said to be a positive semi-definite on a set $\mathbf{S}$ if $\mathbf{u}'\mathbf{D}\mathbf{u} \geqq 0$ for all $\mathbf{u} \in \mathbf{S}$. Negative definite and negative semi-definite forms on $\mathbf{S}$ are defined similarly.) It follows that $\mathbf{u} = (u_1, \ldots, u_K)' \in \mathbf{S}$ is the zero vector iff $\mathbf{v} = (u_1, \ldots, u_{K-1})'$ is the zero vector and therefore $\mathbf{u}'\mathbf{D}\mathbf{u}$ is a negative definite form on $\mathbf{S}$ if, and only if, $\mathbf{v}'\mathbf{D}^*\mathbf{v}$ is a positive definite form. Consequently $(\mathbf{f} - \mathbf{g})'\,\mathbf{D}(\mathbf{g} - \mathbf{f})$ is a positive definite (positive semi-definite) form on $\mathbf{Q}_K$ if, and only if, $(\mathbf{f}^* - \mathbf{g}^*)'\,\mathbf{D}^*(\mathbf{f}^* - \mathbf{g}^*)$ is a positive definite (positive semi-definite) form and $\mathbf{f}^* = (f_1, \ldots, f_{K-1})'$, $\mathbf{g}^* = (g_1, \ldots, g_{K-1})'$.

b) Let $\mathbf{D}^*$ be a positive definite or a positive semi-definite matrix (i.e. a Gramian matrix). With regard to non-negativity of all members the relation

$$D(\mathbf{f}, \mathbf{g}) + D(\mathbf{g}, \mathbf{h}) \geqq D(\mathbf{f}, \mathbf{h})$$

holds for all $\mathbf{f}, \mathbf{g}, \mathbf{h} \in \mathbf{Q}_K$ if and only if

$$D^2(\mathbf{f}, \mathbf{g}) + 2D(\mathbf{f}, \mathbf{g})\,D(\mathbf{g}, \mathbf{h}) + D^2(\mathbf{g}, \mathbf{h}) \geqq D^2(\mathbf{f}, \mathbf{h})\,. \tag{5}$$

With regard to the identity

$$D^2(\mathbf{f}, \mathbf{g}) + D^2(\mathbf{g}, \mathbf{h}) - D^2(\mathbf{f}, \mathbf{h}) = 2(\mathbf{g} - \mathbf{f})'\,\mathbf{D}(\mathbf{h} - \mathbf{g}) = 2(\mathbf{g}^* - \mathbf{f}^*)'\,\mathbf{D}^*(\mathbf{g}^* - \mathbf{h}^*)$$

(5) holds if and only if

$$(\mathbf{g}^* - \mathbf{f}^*)'\,\mathbf{D}^*(\mathbf{g}^* - \mathbf{h}^*) + D(\mathbf{f}, \mathbf{g})\,D(\mathbf{g}, \mathbf{h}) \geqq 0\,.$$

The statement of the theorem follows from here and from the Schwarz inequality. □

Direct examination of existence of metric can be based on the following properties of the matrix $\mathbf{D}$.

Theorem 3. If $D(\mathbf{f}, \mathbf{g})$ is a semimetrics on $\mathbf{Q}_K \times \mathbf{Q}_K$, then each of the following conditions is sufficient for $D(\mathbf{f}, \mathbf{g})$ being a metrics on $\mathbf{Q}_K \times \mathbf{Q}_K$:

a) $\mathbf{D}$ is a nonsingular matrix and $\sum_{i=1}^{K}\sum_{j=1}^{K} d_{ij}^{(-1)} \neq 0$, where $\mathbf{D}^{-1} = \|d_{ij}^{(-1)}\|$.

b) Both $\mathbf{D}$ and the block matrix

$$\mathbf{G} = \begin{Vmatrix} \mathbf{D} & \mathbf{J} \\ \mathbf{J}' & \mathbf{0} \end{Vmatrix}$$

where $\mathbf{J} = (1, \ldots, 1)'$, are nonsingular matrices.

Proof. a) If $D(\mathbf{f}, \mathbf{g})$ is a semimetrics on $\mathbf{Q}_K \times \mathbf{Q}_K$ then $\mathbf{Y}'\mathbf{D}\mathbf{Y} \leqq 0$ for all vectors $\mathbf{Y} = (y_1, \ldots, y_K)'$ such that $\sum_{i=1}^{K} y_i = 0$. Let us denote $F(\mathbf{Y}) = \mathbf{Y}'\mathbf{D}\mathbf{Y} - \lambda \sum_{i=1}^{K} y_i$, where λ is the Lagrangian multiplier. Then

$$\frac{\partial F(\mathbf{Y})}{\partial \mathbf{Y}} = 2\mathbf{D}\mathbf{Y} - \lambda \mathbf{J}, \quad \mathbf{J} = (1, \ldots, 1)'.$$

Let $\mathbf{D}\mathbf{Y} = \frac{1}{2}\lambda\mathbf{J}$ be a consistent system and $\mathbf{D}^-$ be a generalized inverse matrix to $\mathbf{D}$. Then $\mathbf{Y}_0 = \frac{1}{2}\lambda\mathbf{D}^-\mathbf{J}$ is a solution of the given system and all solutions are expressed as $\mathbf{Y} = \mathbf{Y}_0 + (\mathbf{D}^-\mathbf{D} - \mathbf{I})\,\mathbf{Z}$, where $\mathbf{Z}$ is an arbitrary vector. From consistency it also follows that $\frac{1}{2}\lambda\mathbf{D}\mathbf{D}^-\mathbf{J} = \frac{1}{2}\lambda\mathbf{J}$. If $\mathbf{D}$ is a nonsingular matrix, then $\mathbf{Y} = \frac{1}{2}\lambda\mathbf{D}^{-1}\mathbf{J}$ is the solution of $\mathbf{D}\mathbf{Y} = \frac{1}{2}\lambda\mathbf{J}$. It is easily seen that

$$0 = \sum_{i=1}^{K} y_i = \mathbf{J}'\mathbf{Y} = \tfrac{1}{2}\lambda\mathbf{J}'\mathbf{D}^{-1}\mathbf{J} = \tfrac{1}{2}\lambda \sum_{i=1}^{K}\sum_{j=1}^{K} d_{ij}^{(-1)}.$$

If $\sum_{i=1}^{K}\sum_{j=1}^{K} d_{ij}^{(-1)} \neq 0$, then $\lambda = 0$ and consequently $\mathbf{Y} = \mathbf{0}$ is the only solution of the equation $\mathbf{D}\mathbf{Y} = \frac{1}{2}\lambda\mathbf{J}$ and $\mathbf{Y}'\mathbf{D}\mathbf{Y} = 0$ only for $\mathbf{Y} = \mathbf{0}$.

b) If $\mathbf{D}$ is a nonsingular matrix, then

$$|\mathbf{G}| = \begin{vmatrix} \mathbf{D} & \mathbf{J} \\ \mathbf{J}' & \mathbf{0} \end{vmatrix} = |\mathbf{D}|\,|\mathbf{0} - \mathbf{J}'\mathbf{D}^{-1}\mathbf{J}| = |\mathbf{D}|\,|-\mathbf{J}'\mathbf{D}^{-1}\mathbf{J}| = -|\mathbf{D}| \sum_{i=1}^{K}\sum_{j=1}^{K} d_{ij}^{(-1)}. \quad \square$$

The following theorem shows how equivalence classes are generated by a semimetrics.

Theorem 4. (*Classes of equivalent distributions*). If $D(\mathbf{f}, \mathbf{g})$ is a semimetrics on $\mathbf{Q}_K \times \mathbf{Q}_K$, then for each $\mathbf{f} \in \mathbf{Q}_K$ there exists an equivalence class $\mathbf{f}^+$ which contains all distributions $\mathbf{g}$ from $\mathbf{Q}_K$ given by

$$\mathbf{g} = \mathbf{f} + \tfrac{1}{2}\lambda\mathbf{D}^-\mathbf{J} + (\mathbf{D}^-\mathbf{D} - \mathbf{I})\,\mathbf{Z}$$

where $\lambda = (2/K)\,\mathbf{J}'\mathbf{D}(\mathbf{g} - \mathbf{f}) = 2\mathbf{D}_k(\mathbf{g} - \mathbf{f})$ for arbitrary $k = 1, \ldots, K$, $\mathbf{D}^-$ is a generalized inverse matrix to $\mathbf{D}$, $\mathbf{I}$ is the identity matrix and $\mathbf{Z}$ is an arbitrary vector, $\mathbf{D}_k$ is the kth row of the matrix $\mathbf{D}$. The set $\mathbf{Q}_K^+$ of all equivalence classes with $D(\mathbf{f}^+, \mathbf{h}^+) = D(\mathbf{f}, \mathbf{h})$, where $\mathbf{f} \in \mathbf{f}^+$, $\mathbf{h} \in \mathbf{h}^+$ is a metric space.

Proof. Let us write $\mathbf{f} \sim \mathbf{g} \Leftrightarrow D(\mathbf{f}, \mathbf{g}) = 0$. It is obvious that this relation is equivalence on $\mathbf{Q}_K$ by which $\mathbf{Q}_K$ decomposes into equivalence classes (reflexivity and symmetry are obvious, since from $D(\mathbf{f}, \mathbf{g}) = 0$ it follows that $D(\mathbf{g}, \mathbf{f}) = 0$ and $D(\mathbf{f}, \mathbf{f}) = 0$; transitivity follows from the triangular inequality: $0 \leqq D(\mathbf{f}, \mathbf{h}) \leqq D(\mathbf{f}, \mathbf{g}) + D(\mathbf{g}, \mathbf{h}) = 0$). Further, we can see that $D(\mathbf{f}, \mathbf{g}) = 0 \Leftrightarrow \mathbf{g} - \mathbf{f} = \frac{1}{2}\lambda\mathbf{D}^-\mathbf{J} + (\mathbf{D}^-\mathbf{D} - \mathbf{I})\,\mathbf{Z}$. In fact if $D(\mathbf{f}, \mathbf{g}) = 0$, then $\mathbf{g} - \mathbf{f}$ must be a solution of the system $\mathbf{D}\mathbf{Y} = \frac{1}{2}\lambda\mathbf{J}$ and consequently $\mathbf{g} - \mathbf{f} = \frac{1}{2}\lambda\mathbf{D}^-\mathbf{J} + (\mathbf{D}^-\mathbf{D} - \mathbf{I})\,\mathbf{Z}$ (see the proof of Theorem 3). If $\mathbf{g} - \mathbf{f} = \frac{1}{2}\lambda\mathbf{D}^-\mathbf{J} + (\mathbf{D}^-\mathbf{D} - \mathbf{I})\,\mathbf{Z}$, then $D(\mathbf{g} - \mathbf{f}) = \frac{1}{2}\lambda\mathbf{D}\mathbf{D}^-\mathbf{J} =$

$= \frac{1}{2}\lambda\mathbf{J}$ (it follows from consistency of the system $\mathbf{DY} = \frac{1}{2}\lambda\mathbf{J}$) and thus $(\mathbf{f} - \mathbf{g})' \mathbf{D}(\mathbf{g} - \mathbf{f}) = \frac{1}{2}\lambda(\mathbf{f} - \mathbf{g})' \mathbf{J} = 0$.

If $\mathbf{f}^+, \mathbf{h}^+$ are two equivalence classes, let us choose $\mathbf{f} \in \mathbf{f}^+$, $\mathbf{h} \in \mathbf{h}^+$ and define $D(\mathbf{f}^+, \mathbf{h}^+) = D(\mathbf{f}, \mathbf{h})$. If $\mathbf{f} \sim \mathbf{u}$, $\mathbf{h} \sim \mathbf{v}$, then $D(\mathbf{f}, \mathbf{h}) = D(\mathbf{u}, \mathbf{v})$, so that $D(\mathbf{f}^+, \mathbf{h}^+)$ is well defined. In fact if $D(\mathbf{f}, \mathbf{u}) = D(\mathbf{h}, \mathbf{v}) = 0$, then $\mathbf{D}(\mathbf{f} - \mathbf{u}) = \frac{1}{2}\lambda_1\mathbf{J}$, $\mathbf{D}(\mathbf{h} - \mathbf{v}) = \frac{1}{2}\lambda_2\mathbf{J}$ (see proof of Theorem 3). From here $\mathbf{D}(\mathbf{h} - \mathbf{f}) = \mathbf{D}(\mathbf{v} - \mathbf{u}) + \frac{1}{2}(\lambda_2 - \lambda_1)\mathbf{J}$ and consequently $D^2(\mathbf{f}, \mathbf{h}) = (\mathbf{f} - \mathbf{h})' \mathbf{D}(\mathbf{h} - \mathbf{f}) = (\mathbf{f} - \mathbf{h})' \mathbf{D}(\mathbf{v} - \mathbf{u}) + \frac{1}{2}(\lambda_2 - \lambda_1) \cdot (\mathbf{f} - \mathbf{h})' \mathbf{J} = (\mathbf{f} - \mathbf{h})' \mathbf{D}(\mathbf{v} - \mathbf{u})$. Further we can see that $D^2(\mathbf{u}, \mathbf{v}) = (\mathbf{u} - \mathbf{v})' \cdot \mathbf{D}(\mathbf{v} - \mathbf{u}) = (\mathbf{u} - \mathbf{v})' \mathbf{D}(\mathbf{h} - \mathbf{f}) - \frac{1}{2}(\lambda_2 - \lambda_1)(\mathbf{u} - \mathbf{v})' \mathbf{J} = (\mathbf{u} - \mathbf{v})' \mathbf{D}(\mathbf{h} - \mathbf{f}) = (\mathbf{f} - \mathbf{h})' \mathbf{D}(\mathbf{v} - \mathbf{u}) \Rightarrow D^2(\mathbf{f}, \mathbf{h}) = D^2(\mathbf{u}, \mathbf{v})$.

From the proof of Theorem 3 it follows that λ belonging to the matrix $\mathbf{D}$ and the vector $\mathbf{Y}(\sum y_i = 0)$ is $\lambda = (2/K)\mathbf{J}'\mathbf{DY} = 2\mathbf{D}_k'\mathbf{Y}(\mathbf{DY} = \frac{1}{2}\lambda\mathbf{J} \Rightarrow \mathbf{J}'\mathbf{DY} = \frac{1}{2}\lambda\mathbf{J}'\mathbf{J} = (K/2)\lambda)$. □

Corollary 1. Let $\mathbf{D}$ generates a semi-metrics on $\mathbf{Q}_K \times \mathbf{Q}_K$. Then $\mathbf{f}, \mathbf{g} \in \mathbf{Q}_K$ belong to the same equivalence class if and only if $\mathbf{D}(\mathbf{g} - \mathbf{f}) = \frac{1}{2}\lambda\mathbf{J}$, where λ is a constant which is dependent on $\mathbf{D}$ and $\mathbf{g} - \mathbf{f}$, $\mathbf{J} = (1, \ldots, 1)'$.

Proof. If $\mathbf{D}(\mathbf{g} - \mathbf{f}) = \frac{1}{2}\lambda\mathbf{J}$, then $D^2(\mathbf{f}, \mathbf{g}) = (\mathbf{f} - \mathbf{g})' \mathbf{D}(\mathbf{g} - \mathbf{f}) = \frac{1}{2}\lambda(\mathbf{f} - \mathbf{g})' \mathbf{J} = 0$. Let $D^2(\mathbf{f}, \mathbf{g}) = 0$. The difference $\mathbf{g} - \mathbf{f}$ must fulfil the stationarity condition for the function $F(\mathbf{Y}) = \mathbf{Y}'\mathbf{DY} - \lambda \sum_{k=1}^{K} y_k$ (see the proof of Theorem 3) because $\mathbf{g} - \mathbf{f}$ belongs to the vectors $\mathbf{Y}$ which maximize the form $\mathbf{Y}'\mathbf{DY}$ on the set $\boldsymbol{S} = \{\mathbf{Y}: \sum_{k=1}^{K} y_k = 0\}$ and $\max \mathbf{Y}'\mathbf{DY} = 0$. Hence $\mathbf{D}(\mathbf{g} - \mathbf{f}) = \frac{1}{2}\lambda\mathbf{J}$. □

Corrolary 2. Let $\mathbf{D}$ generates a semi-metrics on $\mathbf{Q}_K \times \mathbf{Q}_K$ and $\mathbf{f}, \mathbf{g} \in \mathbf{Q}_K$ belong to the same equivalence class. Then

$$\operatorname{Gvar} \mathbf{g} - \operatorname{Gvar} \mathbf{f} = \lambda \mathbf{f}'\mathbf{DD}^-\mathbf{J} = \lambda ,$$

where $\mathbf{D}^-$ is a generalized inverse matrix to $\mathbf{D}$, $\mathbf{J} = (1. \ldots, 1)'$, $\lambda = (2/K)\mathbf{J}'\mathbf{D}(\mathbf{g} - \mathbf{f}) = 2\mathbf{D}_k'(\mathbf{g} - \mathbf{f})$, $\mathbf{D}_k$ is the kth row of $\mathbf{D}$, $k = 1, \ldots, K$.

Proof. $\operatorname{Gvar} \mathbf{g} = \operatorname{Gvar}[\mathbf{f} + \frac{1}{2}\lambda\mathbf{D}^-\mathbf{J} + (\mathbf{D}^-\mathbf{D} - \mathbf{I})\mathbf{Z}] = (\mathbf{f} + \frac{1}{2}\lambda\mathbf{D}^-\mathbf{J} + (\mathbf{D}^-\mathbf{D} - \mathbf{I})\mathbf{Z})' \mathbf{D}(\mathbf{f} + \frac{1}{2}\lambda\mathbf{D}^-\mathbf{J} + (\mathbf{D}^-\mathbf{D} - \mathbf{I})\mathbf{Z}) = \mathbf{f}'\mathbf{Df} + \frac{1}{4}\lambda\mathbf{f}'\mathbf{DD}^-\mathbf{J} + \mathbf{f}'\mathbf{D} \cdot (\mathbf{D}^-\mathbf{D} - \mathbf{I})\mathbf{Z} + \frac{1}{2}\lambda(\mathbf{D}^-\mathbf{J})' \mathbf{Df} + \frac{1}{4}\lambda^2(\mathbf{D}^-\mathbf{J})' \mathbf{DD}^-\mathbf{J} + \frac{1}{2}\lambda(\mathbf{D}^-\mathbf{J})' \mathbf{D}(\mathbf{D}^-\mathbf{D} - \mathbf{I})\mathbf{Z} + \mathbf{Z}'(\mathbf{D}^-\mathbf{D} - \mathbf{I})' \mathbf{Df} + \frac{1}{2}\lambda\mathbf{Z}'(\mathbf{D}^-\mathbf{D} - \mathbf{I})' \mathbf{DD}^-\mathbf{J} + \mathbf{Z}'(\mathbf{D}^-\mathbf{D} - \mathbf{I})' \mathbf{D}(\mathbf{D}^-\mathbf{D} - \mathbf{I})\mathbf{Z} = \operatorname{Gvar} \mathbf{f} + \lambda\mathbf{f}'\mathbf{DD}^-\mathbf{J} + \frac{1}{4}\lambda^2(\mathbf{D}^-\mathbf{J})' \mathbf{DD}^-\mathbf{J}$.

The same method can be applied to the condition $(\mathbf{f} - \mathbf{g})' \mathbf{D}(\mathbf{g} - \mathbf{f}) = 0$. We obtain that $\frac{1}{4}\lambda^2(\mathbf{D}^-\mathbf{J})' \mathbf{DD}^-\mathbf{J} = 0$. If we take into consideration the fact that $\frac{1}{2}\lambda\mathbf{DD}^-\mathbf{J} = \frac{1}{2}\lambda\mathbf{J}$ then we have $\operatorname{Gvar} \mathbf{g} = \operatorname{Gvar} \mathbf{f} + \lambda\mathbf{f}'\mathbf{J} = \operatorname{Gvar} \mathbf{f} + \lambda$. □

Theorem 5. Let $D(\mathbf{f}, \mathbf{g})$ be a metrics on $\mathbf{Q}_K \times \mathbf{Q}_K$, then the metrics $D(\mathbf{f}, \mathbf{g})$ and the metrics for the nominal case $\varrho(\mathbf{f}, \mathbf{g}) = \sqrt{((\mathbf{f} - \mathbf{g})' (\mathbf{f} - \mathbf{g}))}$ are almost equal, i.e.

there exist finite positive numbers a, b so that it holds:

$$a \leqq \frac{D(\mathbf{f}, \mathbf{g})}{\varrho(\mathbf{f}, \mathbf{g})} \leqq b$$

for all $\mathbf{f} \in \mathbf{Q}_K$, $\mathbf{g} \in \mathbf{Q}_K$, $\mathbf{f} \neq \mathbf{g}$.

Proof.

$$\frac{D^2(\mathbf{f}, \mathbf{g})}{\varrho^2(\mathbf{f}, \mathbf{g})} = \frac{(\mathbf{f} - \mathbf{g})' \mathbf{D}(\mathbf{g} - \mathbf{f})}{(\mathbf{f} - \mathbf{g})' (\mathbf{f} - \mathbf{g})} = \frac{(\mathbf{f}^* - \mathbf{g}^*)' \mathbf{D}^*(\mathbf{f}^* - \mathbf{g}^*)}{(\mathbf{f}^* - \mathbf{g}^*)' \mathbf{C}(\mathbf{f}^* - \mathbf{g}^*)}$$

where $\mathbf{f}^* = (f_1, \ldots, f_{K-1})'$, $\mathbf{g}^* = (g_1, \ldots, g_{K-1})'$, $\mathbf{D}^*$ (see Theorem 2) and $\mathbf{C} = \|c_{ij}\|$, $c_{ii} = 2$ $(i = 1, \ldots, K-1)$, $c_{ij} = 1$ $(i \neq j)$ are positive definite matrices and hence the following relation holds (see [1], pp. 287, 289, 295).

$$0 < \lambda_1 \leqq \frac{D^2(\mathbf{f}, \mathbf{g})}{\varrho^2(\mathbf{f}, \mathbf{g})} \leqq \lambda_{K-1},$$

where $0 < \lambda_1 \leqq \lambda_2 \leqq \ldots \leqq \lambda_{K-1}$ are the roots of the equation $|\mathbf{D}^* - \lambda \mathbf{C}| = 0$, i.e. the eigenvalues of the matrix $\mathbf{D}^* \mathbf{C}^{-1}$. We see that $a = \sqrt{(\lambda_1)} > 0$, $b = \sqrt{(\lambda_{K-1})} > 0$.

Corollary 3. If the semimetrics $D(\mathbf{f}, \mathbf{g})$ does not have the property of a metrics on $\mathbf{Q}_K \times \mathbf{Q}_K$ then

$$0 \leqq \frac{D^2(\mathbf{f}, \mathbf{g})}{\varrho^2(\mathbf{f}, \mathbf{g})} \leqq \lambda_{K-1}.$$

For the proof see [1], pp. 287, 289, 295.

Remark. The simplex $\mathbf{Q}_K$ is not the space with the inner product. In spite of that it is appropriate to introduce an analogical concept that we obtain from the expression

$$D^2(\mathbf{f}, \mathbf{g}) = \operatorname{Gvar} \mathbf{f} + \operatorname{Gvar} \mathbf{g} - 2s(\mathbf{f}, \mathbf{g}).$$

It is an analogy of the relation among distances, norms and inner products in vector spaces. From here we can obtain $s(\mathbf{f}, \mathbf{g})$ that will be called *covariance of distribution* $\mathbf{f}$ *and* $\mathbf{g}$:

$$s(\mathbf{f}, \mathbf{g}) = \mathbf{f}'\mathbf{D}\mathbf{f} + \mathbf{g}'\mathbf{D}\mathbf{g} - \mathbf{f}'\mathbf{D}\mathbf{g} = \mathbf{f}'\mathbf{D}\mathbf{g} - D^2(\mathbf{f}, \mathbf{g}) =$$

$$= \tfrac{1}{2}(\mathbf{f}'\mathbf{D}\mathbf{f} + \mathbf{g}'\mathbf{D}\mathbf{g} - D^2(\mathbf{f}, \mathbf{g})). \tag{6}$$

It holds that

$$-\mathbf{f}'\mathbf{D}\mathbf{g} \leqq s(\mathbf{f}, \mathbf{g}) \leqq \mathbf{f}'\mathbf{D}\mathbf{g}.$$

The lower bound is attained iff $\operatorname{Gvar} \mathbf{f} = \operatorname{Gvar} \mathbf{g} = 0$ and the upper bound is attained iff $D(\mathbf{f}, \mathbf{g}) = 0$. Since $0 \leqq D^2(\mathbf{f}, \mathbf{g}) \leqq 2\mathbf{f}'\mathbf{D}\mathbf{g}$ and the upper bound in this inequality is attained only when $\operatorname{Gvar} \mathbf{f} = \operatorname{Gvar} \mathbf{g} = 0$, we can define the *correlation coefficient of two distributions* by means of normalization of $s(\mathbf{f}, \mathbf{g})$ under the condition that

$\mathbf{f}'\mathbf{D}\mathbf{g} \neq 0$, namely

$$\text{corr}(\mathbf{f}, \mathbf{g}) = \frac{s(\mathbf{f}, \mathbf{g})}{\mathbf{f}'\mathbf{D}\mathbf{g}} = 1 - \frac{D^2(\mathbf{f}, \mathbf{g})}{\mathbf{f}'\mathbf{D}\mathbf{g}}. \tag{7}$$

This coefficient is a *measure of similarity* between two distributions and it has the following properties.

1. It is defined only for $\mathbf{f}'\mathbf{D}\mathbf{g} \neq 0$ (i.e. there exists $f_i g_j d_{ij} \neq 0$) and it takes values from the interval $\langle -1, 1 \rangle$.

2. $\text{corr}(\mathbf{f}, \mathbf{g}) = -1$ iff $\text{Gvar}\,\mathbf{f} = \text{Gvar}\,\mathbf{g} = 0$ and at the same time $D(\mathbf{f}, \mathbf{g}) \neq 0$, $\text{corr}(\mathbf{f}, \mathbf{g}) = 1$ iff $D(\mathbf{f}, \mathbf{g}) = 0$ and at the same time either $\text{Gvar}\,\mathbf{f} \neq 0$ or $\text{Gvar}\,\mathbf{g} \neq 0$.

Two more remarks.

1. The function $D(\mathbf{f}, \mathbf{g})$ can be expressed also by means of the vectors of concentrations $\mathbf{C}_f = \mathbf{D}\mathbf{f}$, $\mathbf{C}_g = \mathbf{D}\mathbf{g}$ as

$$D(\mathbf{f}, \mathbf{g}) = \sqrt{((\mathbf{C}_f - \mathbf{C}_g)' \mathbf{D}^{-1}(\mathbf{C}_g - \mathbf{C}_f))}$$

under the condition that $\mathbf{D}$ is a nonsingular matrix.

2. Measurement of dissimilarity of distributions with respect to $\mathbf{D}$ can be based also on the Euclidean distance of vectors of concentrations, namely

$$D^*(\mathbf{f}, \mathbf{g}) = \sqrt{(\mathbf{C}_f - \mathbf{C}_g)'(\mathbf{C}_f - \mathbf{C}_g)} = \sqrt{(\mathbf{D}\mathbf{f} - \mathbf{D}\mathbf{g})'(\mathbf{D}\mathbf{f} - \mathbf{D}\mathbf{g})} =$$
$$= \sqrt{(\mathbf{f} - \mathbf{g})' \mathbf{D}^2(\mathbf{f} - \mathbf{g})}.$$

4. DECOMPOSITION OF THE GENERALIZED VARIANCE

Now we will investigate the relation of a set of distributions $\{\mathbf{f}_{(1)}, \ldots, \mathbf{f}_{(R)}\}$ to their convex (probabilistic) mixture. Its relevance in data analysis is obvious. The result of Theorem 6 is fundamental for data analysis because it enables us to formulate the problem of the analysis of variance for a generalized categorical variable and to introduce meaningful measures of association.

Theorem 6. (*Decomposition of the generalized variance.*) Let $\{\mathbf{f}_{(1)}, \ldots, \mathbf{f}_{(R)}\}$ be R distributions of the variable $\boldsymbol{A} = \{a_1, \ldots, a_K; \mathbf{D}\}$ and let $\mathbf{w} = (w_1, \ldots, w_R)'$ be a vector from $\mathbf{Q}_R$. Let us consider the vector $\mathbf{f} = \sum_{r=1}^{R} w_r \mathbf{f}_{(r)}$, where $\mathbf{f}_{(r)} = (f_{1|r}, \ldots, f_{K|r})'$. Then

$$\text{Gvar}\,\mathbf{f} = \sum_{r=1}^{R} w_r \,\text{Gvar}\,\mathbf{f}_{(r)} + \tfrac{1}{2}\sum_{r=1}^{R}\sum_{s=1}^{R} w_r w_s D^2(\mathbf{f}_{(r)}, \mathbf{f}_{(s)}) =$$
$$= \sum_{r=1}^{R} w_r \,\text{Gvar}\,\mathbf{f}_{(r)} + \sum_{r=1}^{R} w_r D^2(\mathbf{f}_{(r)}, \mathbf{f}). \tag{8}$$

Proof.

$$\text{Gvar}\,\mathbf{f} = \mathbf{f}'\mathbf{D}\mathbf{f} = \Big(\sum_{r=1}^{R} w_r \mathbf{f}_{(r)}\Big)' \mathbf{D}\Big(\sum_{r=1}^{R} w_r \mathbf{f}_{(r)}\Big) = \sum_{r=1}^{R}\sum_{s=1}^{R} w_r w_s \mathbf{f}'_{(r)} \mathbf{D}\mathbf{f}_{(s)} -$$

$$- \tfrac{1}{2} \sum_{r=1}^{R} \sum_{s=1}^{R} w_r w_s \mathbf{f}'_{(r)} \mathbf{D} \mathbf{f}_{(r)} - \tfrac{1}{2} \sum_{r=1}^{R} \sum_{s=1}^{R} w_r w_s \mathbf{f}'_{(s)} \mathbf{D} \mathbf{f}_{(s)} + \sum_{r=1}^{R} w_r \mathbf{f}'_{(r)} \mathbf{D} \mathbf{f}_{(r)} =$$

$$= \tfrac{1}{2} \sum_{r=1}^{R} \sum_{s=1}^{R} w_r w_s \, D^2(\mathbf{f}_{(r)}, \mathbf{f}_{(s)}) + \sum_{r=1}^{R} w_r \operatorname{Gvar} \mathbf{f}_{(r)} .$$

Similarly

$$\operatorname{Gvar} \mathbf{f} = \mathbf{f}' \mathbf{D} \mathbf{f} = \Big(\sum_{r=1}^{R} w_r \mathbf{f}_{(r)} \Big)' \mathbf{D} \mathbf{f} = \sum_{r=1}^{R} w_r \mathbf{f}'_{(r)} \mathbf{D} \mathbf{f} = \sum_{r=1}^{R} w_r \mathbf{f}'_{(r)} \mathbf{D} \mathbf{f} - \sum_{r=1}^{R} w_r \mathbf{f}' \mathbf{D} \mathbf{f} +$$

$$+ \sum_{r=1}^{R} w_r \mathbf{f}'_{(r)} \mathbf{D} \mathbf{f} + \sum_{r=1}^{R} w_r \mathbf{f}'_{(r)} \mathbf{D} \mathbf{f}_{(r)} - \sum_{r=1}^{R} w_r \mathbf{f}'_{(r)} \mathbf{D} \mathbf{f}_{(r)} =$$

$$= \sum_{r=1}^{R} w_r (\mathbf{f}_{(r)} - \mathbf{f})' \, \mathbf{D} (\mathbf{f} - \mathbf{f}_{(r)}) + \sum_{r=1}^{R} w_r \mathbf{f}'_{(r)} \mathbf{D} \mathbf{f}_{(r)} =$$

$$= \sum_{r=1}^{R} w_r \, D^2(\mathbf{f}_{(r)}, \mathbf{f}) + \sum_{r=1}^{R} w_r \operatorname{Gvar} \mathbf{f}_{(r)} . \qquad \square$$

Theorem 6 can easily be extended to the case in which the distributions $\mathbf{f}_{(r)}$ are again convex mixtures of the distributions $\mathbf{f}_{(r,s)}$ etc. Straightforward application is possible for the analysis of contingency tables in which distributions of rows are conditioned by values of simple classification $\boldsymbol{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R\}$ and $\mathbf{w}$ is the marginal distribution of $\boldsymbol{B}$. Decomposition of population by $\boldsymbol{B}$ results in R strata, each stratum determined by individual value $\boldsymbol{b}_r$, $r = 1, \ldots, R$. Then the result of Theorem 6 for the contingency table can be described as:

total variability of a distribution $\mathbf{f}$

= *variability within strata* + *variability between strata*

= *mean variability of conditioned distributions* + *mean distance between conditioned distributions.*

The analysis of variance for the nominal case based on this decomposition was developed by Light and Margolin [3]. Theorem 6 enables us to solve analogical problems for generalized categorical variables.

5. MEASURES OF EXPLANATORY AND PREDICTIVE POWER OF DECOMPOSITION

Theorem 6 and results of the previous parts enable us to introduce meaningful measures of association between an independent nominal variable $\boldsymbol{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R\}$ and a dependent generalized variable $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$.

Definition 5. *The coefficient of explanatory power of decomposition* is defined as a relative portion of the variability of the dependent variable $\boldsymbol{A}$ explained by the

nominal variable $\boldsymbol{B}$ that generates the decomposition (8)

$$\delta = \delta_{\mathbf{A}/\mathbf{B}} = 1 - \frac{\sum_{r=1}^{R} w_r \operatorname{Gvar} \mathbf{f}_{(r)}}{\operatorname{Gvar} \mathbf{f}} = \frac{\frac{1}{2}\sum_{r=1}^{R}\sum_{s=1}^{R} w_r w_s D^2(\mathbf{f}_{(r)}, \mathbf{f}_{(s)})}{\operatorname{Gvar} \mathbf{f}}. \tag{9}$$

Properties of the coefficient δ follow immediately from Theorem 6.

1. The coefficient is defined whenever $\operatorname{Gvar} \mathbf{f} \neq 0$.
2. $0 \leqq \delta \leqq 1$.
3. $\delta = 0$ iff $\mathbf{f}_{(r)}$ belongs to the same equivalence class with respect to $D(\mathbf{f}, \mathbf{g}) = 0$ for all r for which $w_r > 0$.
4. $\delta = 1$ iff $\operatorname{Gvar} \mathbf{f}_{(r)} = 0$ for all r for which $w_r > 0$.

The coefficient of predictive power of decomposition is based on *proportional-reduction-in error principle* (PRE principle, see [8]). The optimal prediction is considered with respect to the matrix $\mathbf{D}$, a predictive value being a centre of distribution which minimizes the expected loss expressed by d'_{ij}'s (expected dissimilarity scores between true and predicted values). The coefficient is given as a ratio of reduction in expected predictional error in $\boldsymbol{A}$ that provides knowledge about the value of $\boldsymbol{B}$.

Definition 6. *The coefficient of predictive power of decomposition is* defined as

$$\delta^* = 1 - \frac{\sum_{r=1}^{R} w_r d^*_{(r)}}{d^*}, \tag{10}$$

where d^* and $d^*_{(r)}$ are the measures of concentration around the centre for the distribution $\mathbf{f}$ and $\mathbf{f}_{(r)}$ respectively.

Theorem 7. (*Properties of the coefficient* δ^*.)

1. The coefficient is defined only if $d^* \neq 0$.
2. $0 \leqq \delta^* \leqq 1$.
3. If $\mathbf{f}_{(r)} = \mathbf{f}$ for all $r = 1, \ldots, R$ (the case of statistical independence), then $\delta^* = 0$. Conversely it holds only: if $\delta^* = 0$ then any centre of the distribution $\mathbf{f}$ is also a centre of every distribution $\mathbf{f}_{(r)}$ for all such r that $w_r > 0$.
4. $\delta^* = 1$ iff $d^*_{(r)} = 0$ for all $r = 1, \ldots, R$ for which $w_r > 0$.

Proof. The first property is obvious.

2. Denote $\mathbf{c}$ and $\mathbf{c}_{(r)}$ centres of the distribution $\mathbf{f}$ and $\mathbf{f}_{(r)}$ $(r = 1, \ldots, R)$ respectively. The result follows from the inequality

$$d^* = \sum_{i=1}^{K} f_i d(\boldsymbol{a}_i, \mathbf{c}) = \sum_{i=1}^{K}\sum_{r=1}^{R} w_r f_{i/r} d(\boldsymbol{a}_i, \mathbf{c}) = \sum_{r=1}^{R} w_r \sum_{i=1}^{K} f_{i/r} d(\boldsymbol{a}_i, \mathbf{c}) \geqq$$

$$\geqq \sum_{r=1}^{R} w_r \sum_{i=1}^{K} f_{i/r} d(\boldsymbol{a}_i, \mathbf{c}_{(r)}) = \sum_{r=1}^{R} w_r d^*_{(r)} \geqq 0 .$$

3. If $\mathbf{f}_{(r)} = \mathbf{f}$ for all r, then $d^*_{(r)} = d^*$ for all r and hence $\delta^* = 0$. If $\delta^* = 0$, then $\sum_{r=1}^{R} w_r d^*_{(r)} = d^*$. Let us suppose that $w_r > 0$ for all $r = 1, \ldots, R$ and that at least for one s holds that a centre $\mathbf{c}$ of $\mathbf{f}$ is not a centre of the distribution $\mathbf{f}_{(s)}$, i.e. $d^*_{(s)} = \sum_{i=1}^{K} f_{i/s} d(\mathbf{a}_i, \mathbf{c}_{(s)}) < \sum_{i=1}^{K} f_{i/s} d(\mathbf{a}_i, \mathbf{c})$. Then $d^* = \sum_{r=1}^{R} w_r \sum_{i=1}^{K} f_{i/r} d(\mathbf{a}_i, \mathbf{c}) > \sum_{r=1}^{R} w_r d^*_{(r)}$ and it is the contradiction. Thus if $\delta^* = 0$ then $\mathbf{c}$ is a centre of all distribution $\mathbf{f}_{(r)}$ $(r = 1, \ldots, R)$.

4. $\delta^* = 1 \Leftrightarrow \sum_{r=1}^{R} w_r d^*_{(r)} = 0 \Leftrightarrow d^*_{(r)} = 0$ for all r for which $w_r > 0$. □

Remark. The coefficient of explanatory power of decomposition has also PRE interpretation, where prediction is done proportionaly to the distribution of a given variable.

6. RESULTS FOR SIMPLE, ORDERED AND QUANTITATIVE CLASSIFICATION

In this part we present the previous general results for three most frequent and important cases of dependent variables. Theorems 8, 9, 10 are straightforwards consequences of the general properties that were established in the preceding sections.

Theorem 8. (*Simple classification, nominal variable.*) Let $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_K; \mathbf{D}\}$ be a nominal variable $(d_{ij} = (1 - \delta_{ij}))$, $\mathbf{f}, \mathbf{g}, \mathbf{f}_{(1)}, \ldots, \mathbf{f}_{(R)} \in \mathbf{Q}_K$. Then

1. $\operatorname{Gvar} \mathbf{f} = \operatorname{nomvar} \mathbf{f} = 1 - \sum_{i=1}^{K} f_i^2$.

2. $\max_{\mathbf{f} \in \mathbf{Q}_K} \operatorname{nomvar} \mathbf{f} = (K - 1)/K$; the maximum is attained for the distribution $\mathbf{f} = (f_1, \ldots, f_K)$, where $f_i = 1/K$ for all $i = 1, \ldots, K$.

3. $d^*_{\mathbf{a}_i} = 1 - f_i$ for all $i = 1, \ldots, K$.

4. The centre of a distribution $\mathbf{f}$ is its *modal* category.

5. $d^* = 1 - f_M$, where M is an index of the centre of the distribution, i.e. $f_M = \max_i f_i$.

6. $D(\mathbf{f}, \mathbf{g}) = \sqrt{(\sum_{i=1}^{K} (f_i - g_i)^2)}$ and it is an Euclidean metrics.

7. The covariance of distribution $\mathbf{f}$ and $\mathbf{g}$ is

$$s(\mathbf{f}, \mathbf{g}) = 1 - \sum_{i=1}^{K} f_i^2 - \sum_{i=1}^{K} g_i^2 + \sum_{i=1}^{K} f_i g_i =$$

$$= 1 - \tfrac{1}{2}\left(\sum_{i=1}^{K} f_i^2 + \sum_{i=1}^{K} g_i^2\right) - \tfrac{1}{2} \sum_{i=1}^{K} (f_i - g_i)^2$$

and it holds that

$$\sum_{i=1}^{K} f_i g_i - 1 \leqq s(\mathbf{f}, \mathbf{g}) \leqq 1 - \sum_{i=1}^{K} f_i g_i .$$

8. The correlation coefficient of distributions $\mathbf{f}$, $\mathbf{g}$ under the condition $\sum_{i=1}^{K} f_i g_i \neq 1$ is:

$$\operatorname{corr}(\mathbf{f}, \mathbf{g}) = \frac{1 - \sum_{i=1}^{K} f_i^2 - \sum_{i=1}^{K} g_i^2 + \sum_{i=1}^{K} f_i g_i}{1 - \sum_{i=1}^{K} f_i g_i} = 1 - \frac{\sum_{i=1}^{K} (f_i - g_i)^2}{1 - \sum_{i=1}^{K} f_i g_i} .$$

9. The coefficient of explanatory power of decomposition is (for $\sum_{i=1}^{K} f_i^2 \neq 1$) Wallis'tau (cf. [2]):

$$\delta = \tau = 1 - \frac{1 - \sum_{r=1}^{R} w_r \sum_{i=1}^{K} f_{i/r}^2}{1 - \sum_{i=1}^{K} f_i^2} = \frac{\sum_{r=1}^{R} w_r \sum_{i=1}^{K} f_{i/r}^2 - \sum_{i=1}^{K} f_i^2}{1 - \sum_{i=1}^{K} f_i^2} .$$

10. The coefficient of predictive power of decomposition is (for $\max_i f_i \neq 1$) Guttman's lambda (cf. [2]):

$$\delta^* = \lambda = 1 - \frac{1 - \sum_{r=1}^{R} w_r \max_i f_{i/r}}{1 - \max_i f_i} = \frac{\sum_{r=1}^{R} w_r \max_i f_{i/r} - \max_i f_i}{1 - \max_i f_i} .$$

Theorem 9. (*Ordered classification, discrete ordinal variable.*) Let $\mathbf{A} = \{a_1, \ldots, a_K; \mathbf{D}\}$ be a discrete ordinal variable ($d_{ij} = |i - j|$), $\mathbf{f}, \mathbf{f}_{(1)}, \ldots, \mathbf{f}_{(R)} \in \mathbf{Q}_K$. Then

1. $\operatorname{Gvar} \mathbf{f} = \operatorname{dorvar} \mathbf{f} = 2 \sum_{i=1}^{K-1} F_i(1 - F_i)$, $F_i = \sum_{j=1}^{i} f_j$.

2. $\max_{\mathbf{f} \in \mathbf{Q}_K} \operatorname{dorvar} \mathbf{f} = (K - 1)/2$; the maximum is attained for the distribution $\mathbf{f}$ in which $f_1 = f_K = 0{,}5$.

3. $d^*_{a_i} = \sum_{j=1}^{K} f_j |j - i|$ for all $i = 1, \ldots, K$.

4. The centre of a distribution is its *median category*, i.e. the category defined by the relation $F_{Me-1} < 0{,}5$, $F_{Me} \geqq 0{,}5$.

5. $d^* = \sum_{j=1}^{K} f_j |j - Me|$.

6. $D(\mathbf{f}, \mathbf{g}) = \sqrt{[2 \sum_{i=1}^{K-1} (F_i - G_i)^2]}$ and it is a metrics.

7. The covariance of distributions $\mathbf{f}$ and $\mathbf{g}$ is

$$s(\mathbf{f}, \mathbf{g}) = \sum_{i=1}^{K-1} F_i(1 - F_i) + \sum_{i=1}^{K-1} G_i(1 - G_i) - \sum_{i=1}^{K-1} (F_i - G_i)^2 ,$$

$$- \mathbf{f}'\mathbf{D}\mathbf{g} \leqq s(\mathbf{f}, \mathbf{g}) \leqq \mathbf{f}'\mathbf{D}\mathbf{g} ,$$

$$\mathbf{f}'\mathbf{D}\mathbf{g} = \sum_{i=1}^{K-1} F_i(1 - G_i) + \sum_{i=1}^{K-1} G_i(1 - F_i) = \sum_{i=1}^{K-1} (F_i + G_i - 2F_iG_i) =$$

$$= \sum_{i=1}^{K-1} [F_i(1 - F_i) + G_i(1 - G_i) + (F_i - G_i)^2] .$$

8. The correlation coefficient of distributions $\mathbf{f}$, $\mathbf{g}$ under $\sum_{i=1}^{K} f_i g_i \neq 1$ is:

$$\operatorname{corr}(\mathbf{f}, \mathbf{g}) = \frac{\sum_{i=1}^{K-1} [F_i(1 - F_i) + G_i(1 - G_i) - (F_i - G_i)^2]}{\sum_{i=1}^{K-1} [F_i(1 - F_i) + G_i(1 - G_i) + (F_i - G_i)^2]} .$$

9. The coefficient of explanatory power of decomposition under $\sum_{i=1}^{K} F_i(1 - F_i) \neq 0$ is the coefficient β (cf. [4]):

$$\delta = \beta = 1 - \frac{\sum_{r=1}^{R} w_r \sum_{i=1}^{K} F_{i/r}(1 - F_{i/r})}{\sum_{i=1}^{K} F_i(1 - F_i)} .$$

10. The coefficient of predictive power of decomposition under $\sum_{i=1}^{K} f_i |i\text{-}Me| \neq 0$ is the coefficient β^* (cf. [4]):

$$\delta^* = \beta^* = 1 - \frac{\sum_{r=1}^{R} w_r \sum_{i=1}^{K} f_{i/r} |i - Me_{(r)}|}{\sum_{i=1}^{K} f_i |i - Me|} ,$$

where $\boldsymbol{a}_{Me}$ and $\boldsymbol{a}_{Me_{(r)}}$ are median categories with respect to the distribution $\mathbf{f}$ and $\mathbf{f}_{(r)}$ $(r = 1, \ldots, R)$ respectively.

Theorem 10. (*Numerical classification, discrete cardinal variable.*) Let $\boldsymbol{X} = \{x_1, \ldots, x_K\}$ be a discrete cardinal variable $(d_{ij} = (x_i - x_j)^2)$, $\mathbf{f}, \mathbf{f}_{(1)}, \ldots, \mathbf{f}_{(R)} \in \mathbf{Q}_K$. Then

1. $\operatorname{Gvar} \mathbf{f} = 2 \operatorname{var} \boldsymbol{X} = 2 \sum_{i=1}^{K} f_i (x_i - \bar{X})^2$, $\bar{X} = \sum_{i=1}^{K} f_i x_i$.

2. $\max_{\mathbf{f} \in \mathbf{Q}_K} \operatorname{Gvar} \mathbf{f} = \frac{(x_{\max} - x_{\min})^2}{2}$, $x_{\max} = \max(x_1, \ldots, x_K)$, $x_{\min} = \min(x_1, \ldots, x_K)$; the maximum is attained under the distribution $\mathbf{f}$ which has on the places corresponding to the values $x_{\min}$, $x_{\max}$ the values 0,5.

3. $d^*_{x_i} = \sum_{j=1}^{K} f_j(x_j - x_i)^2 = \sum_{j=1}^{K} f_j(x_j - \bar{X})^2 + (x_i - \bar{X})^2 = \operatorname{var} \boldsymbol{X} + (x_i - \bar{X})^2$.

4. The centre X_c of a distribution is the value of the variable $\boldsymbol{X}$ for which it holds that $|X_c - \bar{X}| = \min_i |x_i - \bar{X}|$.

5. $d^* = \operatorname{var} \boldsymbol{X} + (X_c - \bar{X})^2$.

6. $D(\mathbf{f}, \mathbf{g}) = \sqrt{(2)}\,|\bar{X}_f - \bar{X}_g|$, $\bar{X}_f = \sum_{i=1}^{K} f_i x_i$, $\bar{X}_g = \sum_{i=1}^{K} g_i x_i$ and it is a semimetrics. All distributions having the same arithmetic mean belongs to the same equivalence class.

7. The covariance of distributions $\mathbf{f}$ and $\mathbf{g}$ is

$$s(\mathbf{f}, \mathbf{g}) = \sum_{i=1}^{K} f_i x_i^2 + \sum_{i=1}^{K} g_i x_i^2 - 2(\bar{X}_f^2 + \bar{X}_g^2 - \bar{X}_f \bar{X}_g) =$$

$$= (\sum_{i=1}^{K} f_i x_i^2 - \bar{X}_f^2) + (\sum_{i=1}^{K} g_i x_i^2 - \bar{X}_g^2) - (\bar{X}_f - \bar{X}_g)^2 =$$

$$= \operatorname{var}(\boldsymbol{X} \mid \mathbf{f}) + \operatorname{var}(\boldsymbol{X} \mid \mathbf{g}) - (\bar{X}_f - \bar{X}_g)^2 .$$

It holds that $|s(\mathbf{f}, \mathbf{g})| \leqq \mathbf{f}'\mathbf{D}\mathbf{g}$, where

$$\mathbf{f}'\mathbf{D}\mathbf{g} = \sum_{i=1}^{K} f_i x_i^2 + \sum_{i=1}^{K} g_i x_i^2 - 2\bar{X}_f \bar{X}_g =$$

$$= \operatorname{var}(\boldsymbol{X} \mid \mathbf{f}) + \operatorname{var}(\boldsymbol{X} \mid \mathbf{g}) + (\bar{X}_f - \bar{X}_g)^2 .$$

8. The correlation coefficient of distributions $\mathbf{f}$, $\mathbf{g}$ under $\mathbf{f}'\mathbf{D}\mathbf{g} \neq 0$ is

$$\operatorname{corr}(\mathbf{f}, \mathbf{g}) = \frac{\operatorname{var}(\boldsymbol{X} \mid \mathbf{f}) + \operatorname{var}(\boldsymbol{X} \mid \mathbf{g}) - (\bar{X}_f - \bar{X}_g)^2}{\operatorname{var}(\boldsymbol{X} \mid \mathbf{f}) + \operatorname{var}(\boldsymbol{X} \mid \mathbf{g}) + (\bar{X}_f - \bar{X}_g)^2} .$$

9. The coefficient of explanatory power of decomposition under $\operatorname{var}(\boldsymbol{X} \mid \mathbf{f}) \neq 0$ is the correlation ratio:

$$\delta = \eta^2 = 1 - \frac{\sum_{r=1}^{R} w_r \sum_{i=1}^{K} f_{i/r}(x_i - \bar{X}_{(r)})^2}{\sum_{i=1}^{K} f_i (x_i - \bar{X})^2} = 1 - \frac{\sum_{r=1}^{R} w_r \operatorname{var}(\boldsymbol{X} \mid \mathbf{f}_{(r)})}{\operatorname{var}(\boldsymbol{X} \mid \mathbf{f})} ,$$

$$\bar{X}_{(r)} = \sum_{i=1}^{K} f_{i/r} x_i , \quad \bar{X} = \sum_{i=1}^{K} f_i x_i .$$

10. The coefficient of predictive power of decomposition under $\operatorname{var}(\boldsymbol{X} \mid \mathbf{f}) \neq 0$ is

$$\delta^* = 1 - \frac{\sum_{r=1}^{R} w_r [\sum_{i=1}^{K} f_{i/r}(x_i - \bar{X}_{(r)})^2 + (X_{c_{(r)}} - \bar{X}_{(r)})^2]}{\sum_{i=1}^{K} f_i (x_i - \bar{X})^2 + (X_c - \bar{X})^2} =$$

$$= 1 - \frac{\sum_{r=1}^{R} w_r \left[\operatorname{var}(\boldsymbol{X} \mid \mathbf{f}_{(r)}) + (X_{c_{(r)}} - \bar{X}_{(r)})^2\right]}{\operatorname{var}(\boldsymbol{X} \mid \mathbf{f}) + (X_c - \bar{X})^2},$$

where X_c and $X_{c_{(r)}}$ are centres of distributions $\mathbf{f}$, $\mathbf{f}_{(r)}$ $(r = 1, \ldots, R)$ respectively.

Further important special cases concern matrices $\mathbf{D}$, elements of which come from geographic or M-dimensional distances.

Theorem 11. Let $\mathbf{f}^* = (\mathbf{f}_{(1)}, \mathbf{f}_{(2)}, \ldots, \mathbf{f}_{(H)})$, $\mathbf{g}^* = (\mathbf{g}_{(1)}, \mathbf{g}_{(2)}, \ldots, \mathbf{g}_{(H)})$ be two sets of distributions that belong to variables $\boldsymbol{A}_{(1)}, \boldsymbol{A}_{(2)}, \ldots, \boldsymbol{A}_{(H)}$, let $D_{(i)}(\mathbf{f}_{(i)}, \mathbf{g}_{(i)})$ be a metrics (semimetrics) for $i = 1, \ldots, H$ and $W_i > 0$ $(i = 1, 2, \ldots, H)$. Then the functions

$$\bar{D}(\mathbf{f}^*, \mathbf{g}^*) = \sum_{i=1}^{H} W_i D_{(i)}(\mathbf{f}_{(i)}, \mathbf{g}_{(i)}),$$

$$D(\mathbf{f}^*, \mathbf{g}^*) = \sqrt{\left[\sum_{i=1}^{H} W_i D_{(i)}^2(\mathbf{f}_{(i)}, \mathbf{g}_{(i)})\right]}$$

are the metrics (semimetrics) for sets of distributions from $\boldsymbol{Q}^* = \boldsymbol{Q}_{K_1} \times \boldsymbol{Q}_{K_2} \times \ldots \times \boldsymbol{Q}_{K_H}$.

Proof. Both functions are symmetric, non-negative and are equal to zero if, and only if, all $D_{(i)}(\mathbf{f}_{(i)}, \mathbf{g}_{(i)})$ are equal to zero. The triangular inequality for $\bar{D}(\mathbf{f}^*, \mathbf{g}^*)$ is a straightforward consequence of the triangular inequalities for $D_{(i)}(\mathbf{f}_{(i)}, \mathbf{g}_{(i)})$. For $D(\mathbf{f}^*, \mathbf{g}^*)$ it follows:

$$D^2(\mathbf{f}^*, \mathbf{g}^*) = \sum_i W_i D_{(i)}^2(\mathbf{f}_{(i)}, \mathbf{g}_{(i)}) \leqq \sum_i W_i D_{(i)}^2(\mathbf{f}_{(i)}, \mathbf{h}_{(i)}) +$$

$$+ \sum_i W_i D_{(i)}^2(\mathbf{h}_{(i)}, \mathbf{g}_{(i)}) + 2 \sum_i W_i D_{(i)}(\mathbf{f}_{(i)}, \mathbf{h}_{(i)}) D_{(i)}(\mathbf{h}_{(i)}, \mathbf{g}_{(i)}) \leqq$$

$$\leqq \sum_i W_i D_{(i)}^2(\mathbf{f}_{(i)}, \mathbf{h}_{(i)}) + \sum_i W_i D_{(i)}^2(\mathbf{h}_{(i)}, \mathbf{g}_{(i)}) +$$

$$+ 2 \sqrt{\left[\sum_i W_i D_{(i)}^2(\mathbf{f}_{(i)}, \mathbf{h}_{(i)})\right]} \sqrt{\left[\sum_i W_i D_{(i)}^2(\mathbf{h}_{(i)}, \mathbf{g}_{(i)})\right]} =$$

$$= \left\{\sqrt{\left[\sum_i W_i D_{(i)}^2(\mathbf{f}_{(i)}, \mathbf{h}_{(i)})\right]} + \sqrt{\left[\sum_i W_i D_{(i)}^2(\mathbf{h}_{(i)}, \mathbf{g}_{(i)})\right]}\right\}^2 =$$

$$= (D(\mathbf{f}^*, \mathbf{h}^*) + D(\mathbf{h}^*, \mathbf{g}^*))^2 . \qquad \square$$

Theorem 11 permits the multidimensional analysis of distances. Moreover, if we set $\mathbf{f}_{(i)} = \mathbf{f}$ and $\boldsymbol{A}_{(i)} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}_{(i)}\}$, we can introduce two important types of variables, the matrix of distances of which generates a metrics or semimetrics.

Definition 7. If there exist K vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$, $\boldsymbol{X}_k = (x_{k1}, \ldots, x_{kM})$, $k = 1, \ldots, K$ so that $d_{kj} = \sum_{m=1}^{M} (x_{km} - x_{jm})^2$, $\boldsymbol{X}_k \neq \boldsymbol{X}_j$ for $k \neq j$, we call the variable $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$ where $\mathbf{D} = \|d_{kj}\|$ the *M-dimensional metrical variable.* If $M = 2$ the variable $\boldsymbol{A}$ is called the *areal or geographical variable.*

Definition 8. If there exist M different vector $\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_M$, $\boldsymbol{R}_m = (r_{1m}, \ldots, r_{Km})$, $m = 1, \ldots, M$ whose components are permutations of the numbers $(1, 2, \ldots, K)$ so that $d_{kj} = \sum_m |r_{km} - r_{jm}|$, we call the variable $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$ where $\mathbf{D} = \|d_{kj}\|$ the *M-dimensional ordinal variable.*

Corollary 4. a) The matrix of distances of an M-dimensional metrical variable generates a semimetrics on $\boldsymbol{Q}_K \times \boldsymbol{Q}_K$.

b) The matrix of distances of an M-dimensional ordinal variable generates a metrics on $\boldsymbol{Q}_K \times \boldsymbol{Q}_K$.

Proof. a) Let us set in Theorem 11 $\mathbf{f}_{(m)} = \mathbf{f}$, $\mathbf{g}_{(m)} = \mathbf{g}$, $\mathbf{D}_{(m)} = \|d_{kj}^{(m)}\|$, where $d_{kj}^{(m)} = (x_{km} - x_{jm})^2$, $m = 1, \ldots, M$. We see that $D_{(m)}(\mathbf{f}, \mathbf{g})$ are semimetrics on $\boldsymbol{Q}_K \times \boldsymbol{Q}_K$ and that

$$D(\mathbf{f}, \mathbf{g}) = \sqrt{[(\mathbf{f} - \mathbf{g})' \, \mathbf{D}(\mathbf{g} - \mathbf{f})]} = \sqrt{[(\mathbf{f} - \mathbf{g})' \, (\sum_m \mathbf{D}_{(m)}) \, (\mathbf{g} - \mathbf{f})]} =$$

$$= \sqrt{[\sum_{m=1}^{M} (\mathbf{f} - \mathbf{g})' \, \mathbf{D}_{(m)}(\mathbf{g} - \mathbf{f})]} = \sqrt{[\sum_{m=1}^{M} D_{(m)}^2(\mathbf{f}, \mathbf{g})]} = D(\mathbf{f}^*, \mathbf{g}^*) .$$

b) The proof is analogical but $D_{(m)}(\mathbf{f}, \mathbf{g})$, where $\mathbf{D}_{(m)} = \| \, |r_{km} - r_{jm}| \, \|$ are the metrics for $m = 1, \ldots, M$.

CONCLUSIONS

The presented general model aims towards a unifying approach to and a look at various measures of variability, central tendency, predictive and explanatory coefficients. The paper deals with the discrete data that are the most frequent in the social sciences. The model generalizes well known statistical characteristics and enables to fill the existing gap for analogical measures for ordinal data. The important generalization goes to the M-dimensional metric or ordinal variable.

(Received October 30, 1984.)

REFERENCES

[1] F. R. Gantmakher: Teoriya matric (The Theory of Matrices). Second edition. Nauka, Moskva 1966.

[2] L. A. Goodman and W. H. Kruskal: Measures of association for cross-classifications. J. Amer. Statist. Assoc. *49* (1954), 732—764.

[3] R. I. Light and B. H. Margolin: An analysis of variance for categorical data. J. Amer. Statist. Assoc. *66* (1971), 534—544.

[4] J. Řehák: Základní deskriptivní míry pro rozložení ordinálních dat (Basic descriptive measures for the distribution of ordinal data). Sociologický časopis *12* (1976), 4, 416—431.

[5] J. Řehák and B. Řeháková: Základní charakteristiky proměnných s konečným počtem hodnot a distanční analýza jejich rozložení (Basic characteristics for finite-values variables and a distance analysis of their distributions). Sociologický časopis *15* (1979), 2, 214—233.

[6] J. Řehák and B. Řeháková: Distanční přístup k analýze kategorizovaných dat a jeho aplikace

na problém shody (Distance approach to the analysis of the categorical data and its application to the problem of the goodness-of-fit). In: Robust 1982 (J. Antoch, J. Jurečková, eds.). JČMF, Praha 1982, 76—80.

[7] J. Řehák and B. Řeháková: Klassifikacija s numeričeskimi sootnošenijami — *D*-model dlja analiza respredělenij (Classification with numerical relations — *D*-model for an analysis of the distributions). 2-nd Soviet-Czechoslovak seminar "Analysis and Modelling of the Social and Economic Development of the Regions", Kemerovo 1984.

[8] T. P. Wilson: Critique of ordinal variables. Social Forces *49* (1971), 432—444.

RNDr. Jan Řehák, Ústav pro filozofii a sociologii ČSAV (Institute for Philosophy and Sociology — Czechoslovak Academy of Sciences), Jilská 1, 110 00 Praha 1. Czechoslovakia.

RNDr. Blanka Řeháková, Ústav pro výzkum veřejného mínění při Federálním statistickém úřadu (Institute for Public Opinion Research at the Federal Statistical Office), Sokolovská 142, Praha 8. Czechoslovakia.