KYBERNETIKA — VOLUME 23 (1987), NUMBER 4

CLASSIFICATIONS WITH RELATIONS II: ASYMPTOTIC TESTING AND ESTIMATION

JAN ŘEHÁK, BLANKA ŘEHÁKOVÁ

The asymptotic statistical theory for measures derived from the D-model provides the means for analysis of generalized categorical data with relations among categories that are specified by a matrix $\mathbf{D} = \|d_{ij}\|$. This concept was introduced in [9] and developed in [10]—[19]. Independently it appeared later in the contex of C.R. Rao's concept of diversity (see e.g. [4]—[7]). Weights (scores) d_{ij} express *dissimilarities* between categories i and j of the particular variable. They may be interpreted as values of a loss function as well. In this paper we present some asymptotic results for the coefficient of explanatory power (Section 2) and the coefficient of partial association (Section 3). The necessary definition and results are reviewed in Section 1. Further we deal with testing hypotheses of goodness-of-fit, homogeneity of independent samples and marginal homogeneity in square contingency tables for large samples (Section 4). Residual analysis in the homogeneity problem is also included (Section 4). Most of this paper is based on the results of [18].

1. DEFINITIONS AND PRELIMINARIES

A generalized categorical variable $\boldsymbol{A}$ is given by a list of its values (categories) and by a matrix of scores generating its type:

$$\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\} . \tag{1}$$

The matrix $\mathbf{D}$ is supposed to be a $K \times K$ real symmetric matrix with nonnegative elements d_{ij} ($d_{ii} = 0$ for all i). The more unlike the categories $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ are, the greater is the score d_{ij}.

A generalized variance of a distribution $\mathbf{f}$ on $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$ is defined as

$$\mathrm{Gvar}\, \mathbf{f} = \mathbf{f}'\mathbf{D}\mathbf{f} , \tag{2}$$

where $\mathbf{f} \in \mathbf{Q}_K$ and $\mathbf{Q}_K$ is the set of all possible K-dimensional probability vectors.

A distance between distributions $\mathbf{f}$ and $\mathbf{g}$ of $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$ is defined as

$$D(\mathbf{f}, \mathbf{g}) = [(\mathbf{f} - \mathbf{g})' \mathbf{D}(\mathbf{g} - \mathbf{f})]^{1/2} \tag{3}$$

whenever D is such a matrix that $(\mathbf{f} - \mathbf{g})' \mathbf{D}(\mathbf{g} - \mathbf{f}) \geqq 0$ for all $\mathbf{f}, \mathbf{g} \in \mathbf{Q}_K$. This requirement is fulfilled iff $\mathbf{D}^* = \|d_{ij}^*\|$

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

is a $(K-1) \times (K-1)$ Gramian matrix. Moreover the function $D(\cdot, \cdot)$ is a metrics (semimetrics) on $\mathbf{Q}_K$ iff the matrix $\mathbf{D}^*$ is a positive definite (positive semidefinite) matrix. Further it holds that

$$D(\mathbf{f}, \mathbf{g}) = [(\mathbf{f}^* - \mathbf{g}^*)' \mathbf{D}^* (\mathbf{f}^* - \mathbf{g}^*)]^{1/2}, \tag{5}$$

where $\mathbf{f}^* = (f_1, \ldots, f_{K-1})'$, $\mathbf{g}^* = (g_1, \ldots, g_{K-1})'$.

Let $\mathbf{f}_{(1)}, \ldots, \mathbf{f}_{(R)}$ be R distributions on $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$ for the strata that are determined by values $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R$ of a nominal variable $\boldsymbol{B}$, and let $\mathbf{w} \in \mathbf{Q}_R$ with positive components w_r. Considering a vector $\mathbf{f} = \sum w_r \mathbf{f}_{(r)}$ then the decompositional property of the generalized variance holds:

$$\begin{aligned} \operatorname{Gvar} \mathbf{f} &= \sum_{r=1}^{R} w_r \operatorname{Gvar} \mathbf{f}_{(r)} + \tfrac{1}{2} \sum_{r=1}^{R} \sum_{s=1}^{R} w_r w_s D^2(\mathbf{f}_{(r)}, \mathbf{f}_{(s)}) = \\ &= \sum_{r=1}^{R} w_r \operatorname{Gvar} \mathbf{f}_{(r)} + \sum_{r=1}^{R} w_r D^2(\mathbf{f}_{(r)}, \mathbf{f}) . \end{aligned} \tag{6}$$

A coefficient of explanatory power of decomposition $\delta_{\boldsymbol{A}|\boldsymbol{B}}$ is defined as a relative portion of the explained variability of the dependent variable $\boldsymbol{A}$ by the nominal variable $\boldsymbol{B}$ that generates the decomposition, i.e.

$$\delta_{\boldsymbol{A}|\boldsymbol{B}} = 1 - \frac{\sum_{r=1}^{R} w_r \operatorname{Gvar} \mathbf{f}_{(r)}}{\operatorname{Gvar} \mathbf{f}} . \tag{7}$$

2. ASYMPTOTIC DISTRIBUTION OF $\hat{\delta}_{\boldsymbol{A}|\boldsymbol{B}}$

The asymptotic distribution of $\hat{\delta}_{\boldsymbol{A}|\boldsymbol{B}}$ (*the sample analog* of $\delta_{\boldsymbol{A}|\boldsymbol{B}}$ that is its consistent estimate) has been investigated by Řeháková [18].

Consider a two-dimensional categorization $(\boldsymbol{B}, \boldsymbol{A})$, where $\boldsymbol{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R\}$ is a nominal variable and $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$. Let $p_{rk} = P(\boldsymbol{B} = \boldsymbol{b}_r, \boldsymbol{A} = \boldsymbol{a}_k)$, $p_{rk} > 0$ for all r, k and $\sum \sum p_{rk} = 1$. Suppose that a sample of size n is drawn from a distribution with probabilities p_{rk}. Such a sample may be described by the multinomial distribution $\mathcal{M}(n; p_{11}, \ldots, p_{RK})$ with RK cells. Let f_{rk} be observed relative frequencies.

Theorem 1. The asymptotic distribution of $\sqrt{(n)}\,(\hat{\delta}_{\boldsymbol{A}|\boldsymbol{B}} - \delta_{\boldsymbol{A}|\boldsymbol{B}})$, where

$$\delta_{\boldsymbol{A}|\boldsymbol{B}} = 1 - \frac{\sum_{r=1}^{R} p_{r+} \operatorname{Gvar} \mathbf{p}_{(r)}}{\operatorname{Gvar} \mathbf{p}} , \tag{8}$$

$\hat{\delta}_{\boldsymbol{A}|\boldsymbol{B}} = \delta_{\boldsymbol{A}|\boldsymbol{B}}(f_{11}, \ldots, f_{RK})$ is under a random sample from the multinomial distribu-

tion $\mathcal{M}(n;\, p_{11}, \ldots, p_{RK})$ a normal distribution $\mathcal{N}(0, \sigma^2)$ with

$$\sigma^2 = \frac{1}{\zeta^4} \sum_{b=1}^{R} \sum_{a=1}^{K} p_{ba}[v(2d_a^* - \zeta) - \zeta(2d_{a|b}^* - \text{Gvar } \mathbf{p}_{(b)})]^2 \tag{9}$$

under the condition that $\sigma^2 \neq 0$. Here we denote

$$v = \sum_{r=1}^{R} p_{r+} \text{ Gvar } \mathbf{p}_{(r)}\,, \quad \zeta = \text{Gvar } \mathbf{p}\,, \tag{10}$$

$$\mathbf{p}_{(r)} = (p_{r1}/p_{r+}, \ldots, p_{rK}/p_{r+})' = (p_{1|r}, \ldots, p_{K|r})'\,, \quad \mathbf{p} = (p_{+1}, \ldots, p_{+K})'\,.$$

$$d_{a|b}^* = \sum_{k=1}^{K} p_{k|b} d_{ka}\,, \quad d_a^* = \sum_{k=1}^{K} p_{+k} d_{ka}\,.$$

Further it holds that

$$\frac{\sqrt{(n)}\,(\delta_{\mathbf{A}|\mathbf{B}} - \delta_{\mathbf{A}|\mathbf{B}})}{\sigma(f_{11}, \ldots, f_{RK})}$$

has the normal distribution $\mathcal{N}(0, 1)$. The asymptotic variance σ^2 is equal to zero if and only if all $\mathbf{p}_{(b)}$ $(b = 1, \ldots, R)$ belong to the same class of equivalence generated by the $D(\cdot, \cdot)$, i.e. if and only if the coefficient $\delta_{\mathbf{A}|\mathbf{B}}$ is equal to zero.

Proof. The form of the asymptotic distribution follows from the δ method (see [3], p. 430). For the possibility of simplifying the computation of the asymptotic variance see [2]. The details are given in [18]. To prove the last part let us note that

$$\sigma^2 = 0 \Leftrightarrow v(2d_a^* - \zeta) - \zeta(2d_{a|b}^* - \text{Gvar } \mathbf{p}_{(b)}) = F_{ba} = 0 \text{ for all } a \text{ and } b.$$

1. Let $D(\mathbf{p}_{(b)}, \mathbf{p}_{(b')}) = 0$ for all choices of b and b'. Expressing $\delta_{\mathbf{A}|\mathbf{B}}$ as

$$[\tfrac{1}{2} \sum_r \sum_s p_{r+} p_{s+} D^2(\mathbf{p}_{(r)}, \mathbf{p}_{(s)})]/\text{Gvar } \mathbf{p}$$

(see (6)) we have $\delta_{\mathbf{A}|\mathbf{B}} = 0$ and therefore $v = \zeta$. From the relation

$$\zeta = v + \sum p_{r+} D^2(\mathbf{p}_{(r)}, \mathbf{p})$$

(see (6)) it follows that under the condition $v = \zeta$ it holds that $\sum p_{r+} D^2(\mathbf{p}_{(r)}, \mathbf{p}) = 0$ and therefore $D^2(\mathbf{p}_{(r)}, \mathbf{p}) = 0$ for all r. Finally

$$F_{ba} = \zeta[2(d_a^* - d_{a|b}^*) - (\text{Gvar } \mathbf{p} - \text{Gvar } \mathbf{p}_{(b)})]\,.$$

According to Corollary 1 and 2 of Theorem 4 in [15] it follows that $F_{ba} = 0$ for all a and b.

2. Conversely, let $F_{ba} = 0$ for all a and b. Then also $F_{ba_1} - F_{ba_2} = 0$. Taking the difference of this value between b_1 and b_2 we shall obtain that

$$0 = -\zeta(d_{a_1|b_1}^* - d_{a_2|b_1}^*) + \zeta(d_{a_1|b_2}^* - d_{a_2|b_2}^*)$$

and consequently

$$\sum_k p_{k|b_1}(d_{ka_1} - d_{ka_2}) = \sum_k p_{k|b_2}(d_{ka_1} - d_{ka_2})\,,$$

written in the vector form: $(\mathbf{d}_{a_1} - \mathbf{d}_{a_2})' \mathbf{v} = 0$, where

$$\mathbf{d}_j = (d_{1j}, \ldots, d_{Kj})', \quad \mathbf{v} = \mathbf{v}(b_1, b_2) = \mathbf{p}_{(b_1)} - \mathbf{p}_{(b_2)} .$$

From here it follows that $\mathbf{d}_1'\mathbf{v} = \mathbf{d}_2'\mathbf{v}, \ldots, \mathbf{d}_1'\mathbf{v} = \mathbf{d}_K'\mathbf{v} \Rightarrow \mathbf{d}_1'\mathbf{v} = 1/K(\sum_j \mathbf{d}_j'\mathbf{v}) = \mathbf{d}'\mathbf{v}$ and similarly $\mathbf{d}_2'\mathbf{v} = \mathbf{d}'\mathbf{v}, \ldots, \mathbf{d}_K'\mathbf{v} = \mathbf{d}'\mathbf{v}$, which results in $(\mathbf{d}_1 - \mathbf{d}_K)' \mathbf{v} = 0, \ldots$ $\ldots, (\mathbf{d}_{K-1} - \mathbf{d}_K)' \mathbf{v} = 0$. Written in components:

$$0 = \sum_{j=1}^{K} (d_{kj} - d_{Kj}) v_j \quad (k = 1, \ldots, K-1) .$$

If we take account of $\sum_{j=1}^{K} v_j = 0$ we shall obtain that

$$0 = \sum_{j=1}^{K-1} (d_{kj} - d_{Kj} - d_{kK}) v_j \quad (k = 1, \ldots, K-1) ,$$

which rewritten in matrix notation gives $\mathbf{D}^*(\mathbf{p}^*_{(b_1)} - \mathbf{p}^*_{(b_2)}) = \mathbf{0}$, where $\mathbf{D}^*$ is the $(K-1) \times (K-1)$ matrix with elements given by (4) and $\mathbf{p}^*_{(b_1)} = (p_{1|b_1}, \ldots, p_{K-1|b_1})'$, $\mathbf{p}^*_{(b_2)}$ is defined similarly. From here we can see that $D^2(\mathbf{p}_{(b_1)}, \mathbf{p}_{(b_2)}) = 0$ for all b_1 and b_2 and therefore $\delta_{A|B} = 0$. □

Corollary. If $\mathbf{D}$ generates a metrics on $\mathbf{Q}_K$ then every class of equivalence contains only one element and therefore under the condition that $p_{ba} > 0$ for all, b, a, the asymptotic variance σ^2 is equal to zero if and only if $\mathbf{p}_{(1)} = \ldots = \mathbf{p}_{(R)} = \mathbf{p}$, i.e. if and only if $\delta_{A|B} = 0$.

Remark. Equivalence $\sigma^2 = 0 \Leftrightarrow \delta_{A|B} = 0$ shows that $\hat{\delta}_{A|B}$ is not convenient for testing the hypothesis $\delta_{A|B} = 0$.

Now we consider product multinomial sampling with observed conditioned relative frequencies $f_{1|r}, \ldots, f_{K|r}$ $(r = 1, \ldots, R)$.

Theorem 2. The asymptotic distribution of $\sqrt{(n)}\,(\hat{\delta}_{A|B} - \delta_{A|B})$, where $\delta_{A|B}$ is given by (8), $\hat{\delta}_{A|B} = \delta_{A|B}(f_{1|1}, \ldots, f_{K|R})$ is under a random sample from the R independent multinomial distributions $\mathcal{M}(n_{r+}; \mathbf{p}_{(r)})$ $(n_{r+} = n\omega_r$ is known and positive, $\sum \omega_r = 1$, $p_{k|r} > 0$ for all r and $k)$ a normal distribution $\mathcal{N}(0, \sigma^2)$ with

$$\sigma^2 = \frac{4}{\zeta^4} \sum_{b=1}^{R} \frac{p_{b+}^2}{\omega_b} \sum_{a=1}^{K} p_{a|b} \left[v(d^*_a - \sum_{j=1}^{K} p_{j|b} d^*_j) - \zeta(d^*_{a|b} - \mathrm{Gvar}\, \mathbf{p}_{(b)}) \right]^2 , \tag{11}$$

where v, ζ, $d^*_{a|b}$, d^*_a are given by (10), under the condition that $\sigma^2 \neq 0$. Further it holds that

$$\frac{\sqrt{(n)}\,(\hat{\delta}_{A|B} - \delta_{A|B})}{\sigma(f_{1|1}, \ldots, f_{K|R})}$$

has the normal distribution $\mathcal{N}(0, 1)$. The asymptotic variance σ^2 is equal to zero if and only if all $\mathbf{p}_{(b)}$ $(b = 1, \ldots, R)$ belong to the same class of equivalence generated by the $D(\cdot, \cdot)$, i.e. if and only if the coefficient $\delta_{A|B}$ is equal to zero.

The proof is analogical to that given in Theorem 1 and therefore will be omitted. It can be found in [18].

The remainder of this section is devoted to a bias of $\hat{\delta}_{A|B}$ under multinomial and product multinomial sampling.

Theorem 3. Under the conditions of Theorem 1 the bias of $\hat{\delta}_{A|B}$ is

$$\mathsf{E}(\hat{\delta}_{A|B}) - \delta_{A|B} = \frac{1}{n}\left\{\frac{v}{\zeta} + \frac{1}{\zeta}\sum_{b=1}^{R} \operatorname{Gvar} \mathbf{p}_{(b)} + \frac{2}{\zeta^2}\sum_{b=1}^{R}\sum_{a=1}^{K}(2d^*_{a|b} - \operatorname{Gvar} \mathbf{p}_{(b)})\, d^*_a p_{ba} - \right. \tag{12}$$
$$\left. - \frac{4v}{\zeta^3}\sum_{a=1}^{K}(d^*_a)^2\, p_{+a}\right\} + O(n^{-2}).$$

Under the condition of Theorem 2 the bias of $\hat{\delta}_{A|B}$ is

$$\mathsf{E}(\hat{\delta}_{A|B}) - \delta_{A|B} = \frac{1}{n}\sum_{b=1}^{R}\frac{p^2_{b+}}{\zeta\omega_b}\left\{\operatorname{Gvar} \mathbf{p}_{(b)}\left(\frac{1}{p_{b+}} - \frac{v}{\zeta}\right) + \frac{4}{\zeta}\sum_{a=1}^{K}(d^*_{a|b} - \operatorname{Gvar} \mathbf{p}_{(b)})\, d^*_a p_{a|b} - \right. \tag{13}$$
$$\left. - \frac{4v}{\zeta^2}\left[\sum_{a=1}^{K}(d^*_a)^2\, p_{a|b} - \left(\sum_{a=1}^{K} d^*_a p_{a|b}\right)^2\right]\right\} + O(n^{-2}).$$

The proof that is based on the Taylor expansion is rather long and involves tedious algebraic manipulations. It is given in full details by Řeháková [18].

3. ASYMPTOTIC DISTRIBUTION OF $\hat{\delta}_{A|C(B)}$

Consider a three-dimensional categorization $(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{A})$, $\boldsymbol{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R\}$, $\boldsymbol{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_S\}$, $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K; \mathbf{D}\}$. The *coefficient of partial association* $\delta_{A|C(B)} = \delta_{(C\to A|B)}$, where $\boldsymbol{A}$ is a generalized categorical variable and $\boldsymbol{B}$, $\boldsymbol{C}$ are supposed to be nominal ones, has been introduced by Řeháková [17] in usual way as

$$\delta_{A|C(B)} = \frac{\delta_{A|B\times C} - \delta_{A|B}}{1 - \delta_{A|B}}. \tag{14}$$

The partial association is considered in the context of contingency tables with one dependent variable of general type and two nominal predictors. The case of a three-way contingency table does not restrict the generality since the variables $\boldsymbol{B}$ and $\boldsymbol{C}$ may be considered as a combination of several nominal predictors. The full range of possible applications and interpretations has been given in [14] and [16]. The asymptotic properties are summarized by the following theorem.

Theorem 4. The asymptotic distribution of $\sqrt{(n)}\,(\hat{\delta}_{A|C(B)} - \delta_{A|C(B)})$, where

$$\delta_{A|C(B)} = 1 - \frac{\sum_{r=1}^{R}\sum_{s=1}^{S} p_{rs+} \operatorname{Gvar} \mathbf{p}_{(rs)}}{\sum_{r=1}^{R} p_{r++} \operatorname{Gvar} \mathbf{p}_{(r)}}, \tag{15}$$

$\hat{\delta}_{A|C(B)} = \delta_{A|C(B)}(f_{111}, \ldots, f_{RSK})$ is under a random sample from the multinomial distribution $\mathcal{M}(n; p_{111}, \ldots, p_{RSK})$ ($p_{rsk} > 0$ for all r, s, k) a normal distribution $\mathcal{N}(0, \sigma^2)$ with

$$(16) \quad \sigma^2 = \frac{1}{\zeta^4} \sum_{b=1}^{R} \sum_{c=1}^{S} \sum_{a=1}^{K} p_{bca}[v(2d^*_{a|b} - \operatorname{Gvar} \mathbf{p}_{(b)}) - \zeta(2d^*_{a|bc} - \operatorname{Gvar} \mathbf{p}_{(bc)})]^2$$

assuming that $\sigma^2 \neq 0$. Here we denote

$$(17) \qquad v = \sum_{r=1}^{R} \sum_{s=1}^{S} p_{rs+} \operatorname{Gvar} \mathbf{p}_{(rs)}, \quad \zeta = \sum_{r=1}^{R} p_{r++} \operatorname{Gvar} \mathbf{p}_{(r)},$$

$$\mathbf{p}_{(rs)} = \left(\frac{p_{rs1}}{p_{rs+}}, \ldots, \frac{p_{rsK}}{p_{rs+}}\right)',$$

$$\mathbf{p}_{(r)} = \left(\frac{p_{r+1}}{p_{r++}}, \ldots, \frac{p_{r+K}}{p_{r++}}\right)'.$$

$$d^*_{a|bc} = \sum_{k=1}^{K} \frac{p_{bck}}{p_{bc+}} d_{ka}, \quad d^*_{a|b} = \sum_{k=1}^{K} \frac{p_{b+k}}{p_{b++}} d_{ka}$$

and $f_{111}, \ldots, f_{RSK}$ are observed relative frequencies. Further it holds that

$$\frac{\sqrt{(n)}\,(\hat{\delta}_{A|C(B)} - \delta_{A|C(B)})}{\sigma(f_{111}, \ldots, f_{RSK})}$$

has the normal distribution $\mathcal{N}(0, 1)$. The asymptotic variance σ^2 is equal to zero if and only if all distributions $\mathbf{p}_{(bc)}$ (with regard to c) belong to the same class of equivalence $\mathbf{p}^+_{(b)}$ generated by $D(\cdot, \cdot)$, and it holds for all b, i.e. if and only if the coefficient $\delta_{A|C(B)}$ is equal to zero.

The proof is again analogical to that given in Theorem 1 and therefore it will be omitted. It is given in [18].

4. TESTS OF GOODNESS-OF-FIT AND HOMOGENEITY FOR DISTRIBUTIONS OF A GENERALIZED CATEGORICAL VARIABLE

The hypotheses testing for D-variables in standard situations of homogeneity and goodness-of-fit has the high relevance. The nature of the problem represented by the particular matrix $\mathbf{D}$ determines a special interest in types of heterogeneity and departure from the full coincidence of the sampling distribution with the hypothetical assumption. The distance (3) gives a natural measure that considers the differences among distributions in a desired way. We begin by reminding a well known result (see e.g. [8]).

Lemma 1. If $\mathbf{X} = (X_1, \ldots, X_K)'$ has the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $R(\boldsymbol{\Sigma}) = K$ and $Y = \mathbf{X}'\mathbf{A}\mathbf{X}$ for some symmetric nonnegative definite matrix $\mathbf{A}$,

then $\mathscr{L}[Y] = \mathscr{L}[\sum_{i=1}^{K} \lambda_i Z_i^2]$ where $Z_1^2, \ldots, Z_K^2$ are independent chi-square variables with one degree of freedom each and $\lambda_1, \ldots, \lambda_K$ are the eigenvalues of $\mathbf{\Sigma A}$.

Remark. Values of the distribution function $P(Y \leqq t)$ can be computed according to Algorithm AS 106 (see [20]).

Theorem 5. (*Goodness-of-fit.*) If $\mathbf{X} = (X_1, \ldots, X_K)'$ has the multinomial distribution $\mathscr{M}_K(n, \mathbf{p})$, $\mathbf{f} = (f_1, \ldots, f_K)'$, $f_i = X_i/n$ and $\mathbf{g} \in \mathbf{Q}_K$ is a given vector with nonzero elements, then under the hypothesis $\mathbf{p} = \mathbf{g}$ it holds that

$$\mathscr{L}[nD^2(\mathbf{f}, \mathbf{g})] \to \mathscr{L}[\sum_{i=1}^{K-1} \lambda_i Z_i^2], \tag{18}$$

where Z_i^2 are independent χ_1^2 variates and λ_i are the eigenvalues of $(\mathbf{D}_{\mathbf{g}^*} - \mathbf{g}^*(\mathbf{g}^*)')\,\mathbf{D}^*$, $\mathbf{g}^* = (g_1, \ldots, g_{K-1})'$, $\mathbf{D}_{\mathbf{g}^*} = \operatorname{diag}\{g_1, \ldots, g_{K-1}\}$ and $\mathbf{D}^*$ is given by (4).

Proof. Denote $\mathbf{V}_n = \sqrt{(n)}\,(\mathbf{f}^* - \mathbf{g}^*)$. It is known that $\mathscr{L}[\mathbf{V}_n] \to \mathscr{L}[\mathbf{V}]$, where $\mathbf{V}$ has the multivariate normal distribution $\mathscr{N}(\mathbf{0}, \mathbf{D}_{\mathbf{g}^*} - \mathbf{g}^*(\mathbf{g}^*)')$. If we take into account that $R(\mathbf{D}_{\mathbf{g}^*} - \mathbf{g}^*(\mathbf{g}^*)') = K - 1$ and $nD^2(\mathbf{f}, \mathbf{g}) = \mathbf{V}_n'\mathbf{D}^*\mathbf{V}_n$ is a continuous function of $\mathbf{V}_n$ and if we apply Lemma 1, then

$$\mathscr{L}[nD^2(\mathbf{f}, \mathbf{g})] \to \mathscr{L}[\mathbf{V}'\mathbf{D}^*\mathbf{V}] = \mathscr{L}[\sum_{i=1}^{K-1} \lambda_i Z_i^2]$$

with those λ_i as stated in Theorem 5. □

Lemma 2. Let $\mathbf{X} = (X_1, \ldots, X_K)'$ and $\mathbf{Y} = (Y_1, \ldots, Y_K)'$ be independent random variables having the multinomial distributions $\mathscr{M}_K(n, \mathbf{p})$ and $\mathscr{M}_K(m, \mathbf{p})$, $\mathbf{p} = (p_1, \ldots, p_K)'$; $p_k > 0$ for all k. Denote $\mathbf{f} = n^{-1}\mathbf{X} = (f_1, \ldots, f_K)'$, $\mathbf{g} = m^{-1}\mathbf{Y} = (g_1, \ldots, g_K)'$, $\mathbf{f}^* = (f_1, \ldots, f_{K-1})'$, $\mathbf{g}^* = (g_1, \ldots, g_{K-1})'$, $\mathbf{p}^* = (p_1, \ldots, p_{K-1})'$. Then

$$\mathscr{L}[\sqrt{(n+m)}\,(\mathbf{f}^* - \mathbf{g}^*)] \to \mathscr{N}(\mathbf{0}, [\omega(1-\omega)]^{-1} (\mathbf{D}_{\mathbf{p}^*} - \mathbf{p}^*(\mathbf{p}^*)')), \tag{19}$$

where n and m tends to infinity so that $\lim m/(m+n) = \omega$, $0 < \omega < 1$.

The proof is obvious. □

Theorem 6. (*Two-sample problem.*) Under the assumptions of Lemma 2 it holds that

$$\mathscr{L}[(n+m)\,D^2(\mathbf{f}, \mathbf{g})] \to \mathscr{L}[\sum_{i=1}^{K-1} \lambda_i Z_i^2], \tag{20}$$

where Z_i^2 are independent χ_1^2 variates and λ_i are the eigenvalues of $\omega[(1-\omega)]^{-1}$ $(\mathbf{D}_{\mathbf{p}^*} - \mathbf{p}^*(\mathbf{p}^*)')\,\mathbf{D}^*$, $\mathbf{D}_{\mathbf{p}^*} = \operatorname{diag}\{p_1, \ldots, p_{K-1}\}$, $\mathbf{D}^*$ is given by (4).

Proof. $(n+m)\,D^2(\mathbf{f}, \mathbf{g}) = (n+m)\,(\mathbf{f}^* - \mathbf{g}^*)'\,\mathbf{D}^*(\mathbf{f}^* - \mathbf{g}^*) = \mathbf{H}_{n+m}'\mathbf{D}^*\mathbf{H}_{n+m}$ say. According to Lemma 2

$$\mathscr{L}[\mathbf{H}_{n+m}] \to \mathscr{L}[\mathbf{H}] = \mathscr{N}(\mathbf{0}, [\omega(1-\omega)]^{-1}\,(\mathbf{D}_{\mathbf{p}^*} - \mathbf{p}^*(\mathbf{p}^*)')).$$

Further $\mathscr{L}[\mathbf{H}'_{n+m}\mathbf{D}^*\mathbf{H}_{n+m}] \to \mathscr{L}[\mathbf{H}'\mathbf{D}^*\mathbf{H}]$, because $\mathbf{H}'_{n+m}\mathbf{D}^*\mathbf{H}_{n+m}$ is a continuous function of $\mathbf{H}_{n+m}$. From Lemma 1 it follows that $\mathscr{L}[\mathbf{H}'\mathbf{D}^*\mathbf{H}] = \mathscr{L}[\sum_{i=1}^{K-1} \lambda_i Z_i^2]$ with those λ_i as stated in Theorem 6. □

Remark. If we do not know $\mathbf{p}^*$ we may use its estimate $\hat{\mathbf{p}}^*$ with components $\hat{p}_i = (n_i + m_i)/(n + m)$. Suppose that $n_i + m_i = nf_i + mg_i > 0$ $(i = 1, \ldots, K)$. First we note that $\Sigma = \mathbf{D}_{\mathbf{p}^*} - \mathbf{p}^*(\mathbf{p}^*)'$ and $\hat{\Sigma} = \mathbf{D}_{\hat{\mathbf{p}}^*} - \hat{\mathbf{p}}^*(\hat{\mathbf{p}}^*)'$ are nonsingular matrices and that $\hat{\Sigma} \xrightarrow{P} \Sigma$. Therefore $\hat{\Sigma}^{-1/2} \xrightarrow{P} \Sigma^{-1/2}$. Denote $\mathbf{V}_{n+m} = \sqrt{(\omega(1-\omega))} \cdot \hat{\Sigma}^{-1/2}\mathbf{H}_{n+m}$. Then $\mathscr{L}[\mathbf{V}_{n+m}] \to \mathscr{L}[\mathbf{V}] = \mathscr{N}(\mathbf{0}, \mathbf{I})$ and $\mathscr{L}[\mathbf{V}'_{n+m}\mathbf{D}^*\mathbf{V}_{n+m}] \to \mathscr{L}[\mathbf{V}'\mathbf{D}^*\mathbf{V}] = \mathscr{L}[\sum_{i=1}^{K-1} \lambda_i Z_i^2]$, where λ_i are the eigenvalues of $\mathbf{D}^*$.

Lemma 3. Let $\mathbf{X}_1, \ldots, \mathbf{X}_R$, $\mathbf{X}_r = (X_{r1}, \ldots, X_{rK})'$, $r = 1, \ldots, R$, be independent random variables having the distributions $\mathscr{M}_K(n_r, \mathbf{p})$, $r = 1, \ldots, R$, $\mathbf{p} = (p_1, \ldots, p_K)'$, $p_k > 0$ for all k. Denote $\mathbf{f}_{(r)} = n_r^{-1}\mathbf{X}_r = (f_{1|r}, \ldots, f_{K|r})'$, $\mathbf{f}^*_{(r)} = (f_{1|r}, \ldots, f_{K-1|r})'$, $\mathbf{h} = \sum_{r=1}^{R} w_r \mathbf{f}_{(r)}$, $\mathbf{h}^* = \sum_{r=1}^{R} w_r \mathbf{f}^*_{(r)}$, $w_r > 0$ $(r = 1, \ldots, R)$, $\sum_{r=1}^{R} w_r = 1$. Let $n_r \to \infty$ in such a way that $n_r/n \to \omega_r$, where $n = \sum_{r=1}^{R} n_r$, $0 < \omega_r < 1$ $(r = 1, \ldots, R)$, $\sum_{r=1}^{R} \omega_r = 1$. Then

$$\mathscr{L}[\sqrt{(n)}\,(\mathbf{f}^*_{(1)} - \mathbf{h}^*, \ldots, \mathbf{f}^*_{(R-1)} - \mathbf{h}^*)] \to \mathscr{N}((\mathbf{0}, \ldots, \mathbf{0}), \mathbf{B}_{R-1} \otimes \Sigma_{K-1}), \tag{21}$$

where

$$\Sigma_{K-1} = \mathbf{D}_{\mathbf{p}^*} - \mathbf{p}^*(\mathbf{p}^*)', \quad \mathbf{D}_{\mathbf{p}^*} = \operatorname{diag}\{p_1, \ldots, p_{K-1}\} \tag{22}$$

and $\mathbf{B}_{R-1}$ is a positive definite matrix with elements

$$b_{rr} = \frac{1}{\omega_r} - \frac{2w_r}{\omega_r} + \sum_{i=1}^{R} \frac{w_i^2}{\omega_i} \quad (r = 1, \ldots, R-1) \tag{23}$$

$$b_{rs} = b_{sr} = -\frac{w_r}{\omega_r} - \frac{w_s}{\omega_s} + \sum_{i=1}^{R} \frac{w_i^2}{\omega_i} \quad (r \neq s = 1, \ldots, R-1).$$

Proof. It is easy to show that $\mathscr{L}[\sqrt{(n)}\,(\mathbf{f}^*_{(1)} - \mathbf{h}^*, \ldots, \mathbf{f}^*_{(R)} - \mathbf{h}^*)$ converges to a normal distribution with expected value vector $(\mathbf{0}, \ldots, \mathbf{0})$ and covariance matrix

$$(\mathbf{I}_{R(K-1)} - \mathbf{AC})\,\Sigma_{R(K-1)}\,(\mathbf{I}_{R(K-1)} - \mathbf{AC})', \tag{24}$$

where $\mathbf{I}_{R(K-1)}$ denotes the $R(K-1) \times R(K-1)$ unit matrix. $\mathbf{A} = (\mathbf{I}_{K-1}, \ldots, \mathbf{I}_{K-1})'$ is a $R(K-1) \times (K-1)$ matrix, $\mathbf{C} = (w_1\mathbf{I}_{K-1}, \ldots, w_R\mathbf{I}_{K-1})$ is a $(K-1) \times R(K-1)$ matrix, $\mathbf{I}_{K-1}$ denotes the $(K-1) \times (K-1)$ unit matrix and

$$\Sigma_{R(K-1)} = \left\| \begin{array}{ccc} \omega_1^{-1}\Sigma_{K-1}, & \mathbf{0}, & \ldots, \quad \mathbf{0} \\ \mathbf{0}, & \omega_2^{-1}\Sigma_{K-1}, & \ldots, \quad \mathbf{0} \\ \hline \mathbf{0}, & \mathbf{0}, & \ldots, \omega_R^{-1}\Sigma_{K-1} \end{array} \right\|$$

is a $R(K-1) \times R(K-1)$ matrix with Σ_{K-1} given by (22).

After some algebra we find that (24) is equal to $\mathbf{B}_R \otimes \Sigma_{K-1}$, where $\mathbf{B}_R$ has the elements given by (23) for $r, s = 1, \ldots, R$ and the symbol "$\otimes$" stands for the Kronecker product of matrices. And now it is easy to conclude that the statement (21) holds.

The matrix $\mathbf{B}_{R-1}$ arises from $\mathbf{B}_R$ by omitting the last row and column. $\mathbf{B}_{R-1}$ is positive semidefinite because $\mathbf{B}_{R-1} \otimes \Sigma_{K-1}$ is a covariance matrix. $\mathbf{B}_{R-1}$ is also nonsingular. It follows from the fact that $\mathbf{B}_R = \mathbf{U}_R \mathbf{U}_R'$.

$$\mathbf{U}_R = \left\| \begin{array}{c|ccc} (1 - w_1)/\sqrt{\omega_1}, & -w_2/\sqrt{\omega_2}, & \ldots, & -w_R/\sqrt{\omega_R} \\ \hline -w_1/\sqrt{\omega_1}, & -w_2/\sqrt{\omega_2}, & \ldots, & (1 - w_R)/\sqrt{\omega_R} \end{array} \right\|,$$

$$R(\mathbf{B}_R) = R(\mathbf{U}_R \mathbf{U}_R') = R(\mathbf{U}_R) = R - 1, \quad \text{since} \quad |\mathbf{U}_R| = 0,$$

$$|\mathbf{U}_{R-1}| = w_R \prod_{r=1}^{R-1} (1/\sqrt{\omega_r}) > 0.$$

The last row (column) of $\mathbf{B}_R$ is a linear combination of the preceding rows (columns). If we omit them the rank will not change. Hence $\mathbf{B}_{R-1}$ is positive definite. □

Theorem 7. (*R-sample problem, one-way ANOVA for D-variables.*) Under the assumptions of Lemma 3 it holds that

$$\mathscr{L}\Big[n \sum_{r=1}^{R} w_r D^2(\mathbf{f}_{(r)}, \mathbf{h})\Big] \to \mathscr{L}\Big[\sum_{i=1}^{(R-1)(K-1)} \lambda_i Z_i^2\Big], \tag{25}$$

where Z_i^2 are independent χ_1^2 variates and λ_i are the eigenvalues of the matrix $(\mathbf{B}_{R-1} \otimes \otimes \Sigma_{K-1})(\mathbf{W}_{R-1} \otimes \mathbf{D}^*)$, where $\mathbf{B}_{R-1}, \Sigma_{K-1}, \mathbf{W}_{R-1}$ are positive definite matrices given by (23), (22) and

$$W_{rr} = w_r(w_R + w_r)/w_R \quad (r = 1, \ldots, R - 1)$$
$$W_{rs} = W_{sr} = w_r w_s / w_R \quad (r \neq s) \tag{26}$$

$\mathbf{D}^*$ is a nonnegative definite matrix with elements given by (4).

Proof. After some algebraic manipulation we get that

$$n \sum_{r=1}^{R} w_r D^2(\mathbf{f}_{(r)}, \mathbf{h}) =$$
$$= n(\mathbf{f}_{(1)}^* - \mathbf{h}^*, \ldots, \mathbf{f}_{(R-1)}^* - \mathbf{h}^*)' (\mathbf{W}_{R-1} \otimes \mathbf{D}^*)(\mathbf{f}_{(1)}^* - \mathbf{h}^*, \ldots, \mathbf{f}_{(R-1)}^* - \mathbf{h}^*) =$$
$$= \mathbf{H}_n' \mathbf{M} \mathbf{H}_n \quad \text{(say.)}.$$

The matrix $\mathbf{W}_{R-1}$ is positive definite because its principal minors are equal to $w_1 \ldots w_i (w_1 + \ldots + w_i + w_R)/w_R$ for $i = 1, \ldots, R - 1$. The matrix $\mathbf{M}$ is positive definite (semidefinite) if $\mathbf{D}^*$ is positive definite (semidefinite). According to Lemma 3 $\mathscr{L}[\mathbf{H}_n] \to \mathscr{L}[\mathbf{H}] = \mathscr{N}((\mathbf{0}, \ldots, \mathbf{0}), \mathbf{B}_{R-1} \otimes \Sigma_{K-1})$ and according to Lemma 1 $\mathscr{L}[\mathbf{H}'\mathbf{M}\mathbf{H}] = \mathscr{L}\Big[\sum_{i=1}^{(R-1)(K-1)} \lambda_i Z_i^2\Big]$, where λ_i are the eigenvalues of $(\mathbf{B}_{R-1} \otimes \Sigma_{K-1})\mathbf{M}$.

Remark. If the vector $\mathbf{p}^*$ is unknown we may use its estimate $\hat{\mathbf{p}}^*$ with components $\hat{p}_k = \sum_{r=1}^{R} n_{rk}/n$, where n_{rk} is an observed value of X_{rk}. Suppose $\hat{p}_k > 0$ for $k = 1, \ldots, K$. $\boldsymbol{\Sigma}_{K-1}$ and its estimate $\hat{\boldsymbol{\Sigma}}_{K-1}$ are nonsingular, $\hat{\boldsymbol{\Sigma}}_{K-1} \xrightarrow{P} \boldsymbol{\Sigma}_{K-1}$, hence $\hat{\boldsymbol{\Sigma}}_{K-1}^{-1/2} \xrightarrow{P} \boldsymbol{\Sigma}_{K-1}^{-1/2}$ and

$$\mathscr{L}\big[(\mathbf{B}_{R-1} \otimes \boldsymbol{\Sigma}_{K-1})^{-1/2}\, \mathbf{H}_n\big] = \mathscr{L}[\mathbf{V}_n] \to \mathscr{L}\big[(\mathbf{B}_{R-1} \otimes \boldsymbol{\Sigma}_{K-1})^{-1/2}\, \mathbf{H}\big] =$$

$$= \mathscr{L}[\mathbf{V}] = \mathscr{N}\big((\mathbf{0}, \ldots, \mathbf{0}), \mathbf{I}_{(R-1)(K-1)}\big)$$

and $\mathscr{L}[\mathbf{V}_n'\mathbf{M}\mathbf{V}_n] \to \mathscr{L}[\mathbf{V}'\mathbf{M}\mathbf{V}] = \mathscr{L}\big[\sum_{i=1}^{(R-1)(K-1)} \lambda_i Z_i^2\big]$, where λ_i are the eigenvalues of the matrix $\mathbf{M}$.

Theorem 6 is a special case of Theorem 7. However, it has been separated for its simpler formulation and computational aspects as well as for the frequent occurrence of the two-sample case in practice. The two-sample statistic can also be used in R-sample problem *in combination with a simultaneous testing* or in the case of rejecting homogeneity hypothesis if we want to investigate the distributions *pair-wise* to find out which populations differ.

The investigation of residuals after the hypothesis of homogeneity ($R \geqq 2$) is rejected is an important part of the data analysis. For the purposes of the next theorem let us denote

$$\mathbf{f}^{(r)} = (n - n_{r+})^{-1} \sum_{u \neq r} n_{u+} \mathbf{f}_{(u)} = (f_1^{(r)}, \ldots, f_K^{(r)})',$$

$$\mathbf{p}^{(r)} = (n - n_{r+})^{-1} \sum_{u \neq r} n_{u+} \mathbf{p}_{(u)} = (p_1^{(r)}, \ldots, p_K^{(r)})',$$

$$\mathbf{q}_{(r)} = \mathbf{D}(\mathbf{f}_{(r)} - \mathbf{f}^{(r)}) = (q_{1|r}, \ldots, q_{K|r})'$$

($\mathbf{q}_{(r)}$ is a residual characteristic of the contingency table row).

Theorem 8. (*Residual analysis of the departure from homogeneity.*) Let $\mathbf{f}_{(1)}, \ldots, \mathbf{f}_{(R)}$ are distributed independently as $\mathscr{L}[n_{r+}\mathbf{f}_{(r)}] = \mathscr{M}(n_{r+}, p_{(r)})$, $p_{k|r} > 0$ for all k and r. If $n_{r+} \to \infty$ in such a way that $n_{r+}/n \to \omega_r$, $0 < \omega_r < 1$, $r = 1, \ldots, R$, $n = \sum n_{r+}$, then it holds that

$$\mathscr{L}\big[\sqrt{(n)}\,(\mathbf{q}_{(r)} - \mathbf{D}(\mathbf{p}_{(r)} - \mathbf{p}^{(r)}), \mathbf{q}_{(s)} - \mathbf{D}(\mathbf{p}_{(s)} - \mathbf{p}^{(s)}))'\big] \to$$

$$\to \mathscr{N}\big((\mathbf{0}, \mathbf{0})', \mathbf{A}\,\boldsymbol{\Sigma}\mathbf{A}'\big)$$

where

$$\mathbf{A} = \begin{Vmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{Vmatrix}, \quad \mathbf{A}\,\boldsymbol{\Sigma}\mathbf{A}' = \begin{Vmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_2 & \mathbf{A}_3 \end{Vmatrix},$$

$$(27) \qquad (A_1)_{ab} = \omega_r^{-1}\Big\{\sum_{k=1}^{K} p_{k|r}(d_{ak} - d^*_{a|r})(d_{bk} - d^*_{b|r})\Big\} +$$

$$+ (1 - \omega_r)^{-2}\Big\{\sum_{u \neq r} \omega_u\Big[\sum_{k=1}^{K} p_{k|u}(d_{ak} - d^*_{a|u})(d_{bk} - d^*_{b|u})\Big]\Big\},$$

$$(28)\qquad (A_2)_{ab} = [(1-\omega_r)(1-\omega_s)]^{-1}\{\sum_{u\neq r,s}\omega_u[\sum_{k=1}^{K} p_{k|u}(d_{ak}-d^*_{a|u})(d_{bk}-d^*_{b|u})]\} -$$
$$- (1-\omega_s)^{-1}\{\sum_{k=1}^{K} p_{k|r}(d_{ak}-d^*_{a|r})(d_{bk}-d^*_{b|r})\} -$$
$$- (1-\omega_r)^{-1}\{\sum_{k=1}^{K} p_{k|s}(d_{ak}-d^*_{a|s})(d_{bk}-d^*_{b|s})\}.$$

$$(29)\qquad (A_3)_{ab} = \omega_s^{-1}\{\sum_{k=1}^{K} p_{k|s}(d_{ak}-d^*_{a|s})(d_{bk}-d^*_{b|s})\} +$$
$$+ (1-\omega_s)^{-2}\{\sum_{v\neq s}\omega_v[\sum_{k=1}^{K} p_{k|v}(d_{ak}-d^*_{a|v})(d_{bk}-d^*_{b|v})]\},$$
$$a, b = 1, \ldots, K \text{ and } d^*_{a|r} = \sum_j d_{aj}p_{j|r}.$$

Proof. The asymptotic normality follows from the properties of frequencies. In the first step we can see that

$$\mathscr{L}[\sqrt{(n)}\,(\mathbf{f}_{(r)} - \mathbf{f}^{(r)} - \mathbf{p}_{(r)} + \mathbf{p}^{(r)}, \mathbf{f}_{(s)} - \mathbf{f}^{(s)} - \mathbf{p}_{(s)} + \mathbf{p}^{(s)})'] \to \mathscr{N}((\mathbf{0},\mathbf{0})', \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is a block matrix

$$\boldsymbol{\Sigma} = \left\| \begin{matrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_2 \\ \boldsymbol{\Sigma}_2 & \boldsymbol{\Sigma}_3 \end{matrix} \right\|,$$

$$(\Sigma_1)_{aa} = \omega_r^{-1}p_{a|r}(1-p_{a|r}) + (1-\omega_r)^{-2}\sum_{u\neq r}\omega_u p_{a|u}(1-p_{a|u}),$$

$$(\Sigma_1)_{ab} = -\omega_r^{-1}p_{a|r}p_{b|r} - (1-\omega_r)^{-2}\sum_{u\neq r}\omega_u p_{a|u}p_{b|u},$$

$$(\Sigma_2)_{aa} = [(1-\omega_r)(1-\omega_s)]^{-1}\sum_{u\neq r,s}\omega_u p_{a|u}(1-p_{a|u}) -$$
$$- (1-\omega_s)^{-1}p_{a|r}(1-p_{a|r}) - (1-\omega_r)^{-1}p_{a|s}(1-p_{a|s}),$$

$$(\Sigma_2)_{ab} = -[(1-\omega_r)(1-\omega_s)]^{-1}\sum_{u\neq r,s}\omega_u p_{a|u}p_{b|u} +$$
$$+ (1-\omega_s)^{-1}p_{a|r}p_{b|r} + (1-\omega_r)^{-1}p_{a|s}p_{b|s}.$$

The matrix $\boldsymbol{\Sigma}_3$ is the same as $\boldsymbol{\Sigma}_1$, only we must substitute r by s. The rest follows from the fact that

$$\mathbf{A}(\mathbf{f}_{(r)} - \mathbf{f}^{(r)} - \mathbf{p}_{(r)} + \mathbf{p}^{(r)}, \mathbf{f}_{(s)} - \mathbf{f}^{(s)} - \mathbf{p}_{(s)} + \mathbf{p}^{(s)})' =$$
$$= (\mathbf{q}_{(r)} - \mathbf{D}(\mathbf{p}_{(r)} - \mathbf{p}^{(r)}), \mathbf{q}_{(s)} - \mathbf{D}(\mathbf{p}_{(s)} - \mathbf{p}^{(s)}))'. \qquad \square$$

The next corollary gives the tool for examining the significance of residuals from hypothesis of homogeneity.

Corollary 1. Under the assumptions of Theorem 8 and under $\mathbf{H}_0$: $\mathbf{p}_{(r)} = \mathbf{p}$, $r = 1, \ldots, R$, it holds that

$$\mathscr{L}[\sqrt{(n)}\,q_{k|r}] \to \mathscr{N}(0, \sigma_q^2),$$

where

$$\sigma_q^2 = [\omega_r(1-\omega_r)]^{-1}\sum_{j=1}^{K} p_j(d_{jk} - \sum_{a=1}^{K} p_a d_{ak})^2$$

and

$$\mathscr{L}[\sqrt{(n)}\, q_{k|r}/\hat{\sigma}_q] \to \mathscr{N}(0, 1)\,,$$

where

$$\hat{\sigma}_q^2 = \{n^2/[n_{r+}(n - n_{r+})]\} \sum_{j=1}^{K} f_{+j}(d_{jk} - \sum_{a=1}^{K} f_{+a} d_{ak})^2$$

and $f_{+j} = n^{-1} \Sigma f_{j|s} n_{s+}$ is supposed to be positive for all j.

The statistics $\sqrt{(n)}\, q_{k|r}/\hat{\sigma}_q$ may be used for the testing $\mathbf{H}_0$ against $(\mathbf{p}_{(r)} - \mathbf{p}^{(r)})' \mathbf{D}_k \neq 0$, $\mathbf{D}_k = (d_{1k}, \ldots, d_{Kk})'$. Instead of it we may use the statistics $\sqrt{(n)}\, t_{k|r}/\hat{\sigma}_t$, where

$$t_{k|r} = \sum_{j=1}^{K} (f_{j|r} - f_{+j})\, d_{jk} = (1 - f_{r+})\, q_{k|r}\,,$$

$$\hat{\sigma}_t = (1 - f_{r+})\, \hat{\sigma}_q\,.$$

Notice that σ_q^2 and $\hat{\sigma}_q^2$ are equal to zero if and only if $\mathbf{D}_k = \mathbf{0}$. Of course such a column (and the corresponding row) of the matrix $\mathbf{D}$ would be excluded in advance, so that we can see that σ_q^2 and $\hat{\sigma}_q^2$ are always greater than zero.

Corollary 2. Under the assumptions of Theorem 8

$$\mathscr{L}[\sqrt{(n)}\, (q_{k|r} - q_{l|s} - \mathsf{E}(q_{k|r} - q_{l|s}))] \to \mathscr{N}(0, \sigma^2)$$

provided that $\sigma^2 \neq 0$, where σ^2 is given in the following way:

1. $k = l\,,\quad r \neq s \Rightarrow \sigma^2 = (A_1)_{kk} + (A_3)_{kk} - 2(A_2)_{kk}\,,$
2. $k \neq l\,,\quad r = s \Rightarrow \sigma^2 = (A_1)_{kk} + (A_1)_{ll} - 2(A_1)_{kl}\,,$
3. $k \neq l\,,\quad r \neq s \Rightarrow \sigma^2 = (A_1)_{kk} + (A_3)_{ll} - 2(A_2)_{kl}\,,$

$\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ are described in (27)–(29) and

$$\mathsf{E}(q_{k|r} - q_{l|s}) = d^*_{k|r} - (n - n_{r+})^{-1} \sum_{u \neq r} n_{u+} d^*_{k|u} - d^*_{l|s} + (n - n_{s+})^{-1} \sum_{v \neq s} n_{v+} \mathrm{d}^*_{l|v}\,.$$

The last part of this article deals with marginal homogeneity in square tables.

Lemma 4. Let $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_K; \mathbf{D}\}$, $\mathbf{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_K; \mathbf{D}\}$ be D-variables with the same matrix $\mathbf{D}$. The square table $\mathbf{A} \times \mathbf{B}$ arises as a sample from the multinomial distribution $\mathscr{M}(n; p_{11}, \ldots, p_{KK})$, $p_{ij} > 0$ for all i, j. Denote $\mathbf{p}_\mathbf{A} = (p_{1+}, \ldots, p_{K+})'$, $\mathbf{p}_\mathbf{B} = (p_{+1}, \ldots, p_{+K})'$. Let n_{ij} be the observed frequency in the cell (i, j) and denote $\mathbf{f}_\mathbf{A} = (f_{1+}, \ldots, f_{K+})'$, $\mathbf{f}_\mathbf{B} = (f_{+1}, \ldots, f_{+K})'$, $\mathbf{f}^*_\mathbf{A} = (f_{1+}, \ldots, f_{K-1,+})'$, $\mathbf{f}^*_\mathbf{B} = (f_{+1}, \ldots, f_{+,K-1})'$, $f_{i+} = n_{i+}/n$, $f_{+i} = n_{+i}/n$. Under the condition $\mathbf{p}_\mathbf{A} = \mathbf{p}_\mathbf{B}$ it holds that

$$\mathscr{L}[\sqrt{(n)}\, (\mathbf{f}^*_\mathbf{A} - \mathbf{f}^*_\mathbf{B})] \to \mathscr{N}(\mathbf{0}, \Sigma)\,, \tag{30}$$

where Σ is a $(K - 1) \times (K - 1)$ positive definite matrix with elements

$$\sigma_{ii} = p_{i+} + p_{+i} - 2p_{ii} \quad (i = 1, \ldots, K - 1)$$

$$\sigma_{ij} = -(p_{ij} + p_{ij}) \qquad (i \neq j)\,. \tag{31}$$

For a proof of this result, see e.g. [1], p. 219.

Theorem 9. (*Marginal homogeneity in square tables.*) Under the assumption of Lemma 4 the asymptotic distribution of $D^2(\mathbf{f}_A, \mathbf{f}_B)$ is given by

$$\mathscr{L}[nD^2(\mathbf{f}_A, \mathbf{f}_B)] \to \mathscr{L}[\sum_{i=1}^{K-1} \lambda_i Z_i^2] \,. \tag{32}$$

where Z_i^2 are independent χ_1^2 variables and λ_i are the eigenvalues of the matrix $\mathbf{\Sigma D}^*$, where $\mathbf{\Sigma}$ is defined as in Lemma 4 and $\mathbf{D}^*$ is given by (4).

Proof. The desired result immediately follows from Lemma 1 and 4 in the same way as in the proof of Theorem 6. □

Remark. If we do not know the probabilities p_{ij} we may use $\hat{p}_{ij} = n_{ij}/n$. Under the conditions of Lemma 4 and assuming that $\hat{p}_{ij} > 0$ $(i, j = 1, \ldots, K)$ it holds that $\mathscr{L}[\mathbf{V}_n] \to \mathscr{L}[\mathbf{V}] = \mathscr{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{V}_n = \sqrt{(n)}\, \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{f}_A^* - \mathbf{f}_B^*)$ and $\hat{\mathbf{\Sigma}}$ is an estimate of $\mathbf{\Sigma}$ with $\hat{p}_{ij}$ instead of p_{ij}. Then $\mathscr{L}[\mathbf{V}_n' \mathbf{D}^* \mathbf{V}_n] \to \mathscr{L}[\mathbf{V}' \mathbf{D}^* \mathbf{V}] = \mathscr{L}[\sum_{i=1}^{K-1} \lambda_i Z_i^2]$, where λ_i are the eigenvalues of $\mathbf{D}^*$.

CONCLUSIONS

The statistical theory of asymptotic behaviour of measures derived from the D-model provides a *unified approach* to different types of situations in which the researcher wants to specify the relations among the categories of variables. Important special cases as nominal, ordinal, cardinal or geographical variables are also included. Also analyses of special experimental or decisional data can be done this way if the elements of the matrix $\mathbf{D}$ are interpreted as the loss function values.

(Received October 27, 1986.)

REFERENCES

[1] J. Anděl: Matematická statistika (Mathematical Statistics). SNTL, Praha 1978.
[2] L. A. Goodman and W. H. Kruskal: Measures of association for cross classification IV: simplification of asymptotic variances. J. Amer. Statist. Assoc. *67* (1972), 415—421.
[3] C. R. Rao: Lineární metody statistické indukce a jejich aplikace (Linear Statistical Inference and Its Applications). Academia, Praha 1978.
[4] C. R. Rao: Diversity: its measurement, decomposition, apportionment and analysis. Sankhyā *44*, Series A (1982), 1—22.
[5] C. R. Rao: Gini-Simpson index of diversity: a characterization, generalization and applications. Utilitas Mathematica *21* B (1982), 273—282.
[6] C. R. Rao: Analysis of diversity: a unified approach. Statistical Decision Theory and Related Topics III, *2* (1982), 233—250.
[7] C. R. Rao: Diversity and dissimilarity coefficients: a unified approach. Theoretical Population Biology *21* (1982), 24—43.
[8] H. Ruben: Probability content of regions under spherical normal distributions, IV: the distribution of homogeneous and nonhomogeneous quadratic functions of normal variables. Ann. Math. Statist. *33* (1962), 542—570.

[9] 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.
[10] J. Řehák and I. Loučková: Klasické mnohorozměrné škálování (Aplikace metody DISTAN) (Classical multidimensional scaling (application of the DISTAN method)). Sociologický časopis *19* (1983), 535—554.
[11] J. Řehák and I. Loučková: Komparace podmíněných distribucí kontingenční tabulky klasickým mnohorozměrným škálováním (Comparison of conditional distributions in a contingency table by classical multidimensional scaling). In: Robust 84 (J. Antoch, J. Jurečková, eds.), JČMF, Praha 1984, 100—104.
[12] J. Řehák and B. Řeháková: Basic characteristics for finite-valued variables and a distance analysis of their distributions. 9th World Congress of Sociology, Uppsala 1978.
[13] J. Řehák and B. Řeháková: Distanční přístup k analýze kategorizovaných dat a jeho aplikace na problém shody (Distance approach to analysis of categorical data and its application to the problem of goodness-of-fit). In: Robust 82 (J. Antoch, J. Jurečková, eds.), JČMF, Praha 1982, 76—80.
[14] J. Řehák and B. Řeháková: Parciální asociační koeficienty v kontingenčních tabulkách (Partial association coefficients in contingency tables). In: Robust 84 (J. Antoch, J. Jurečková, eds.), JČMF, Praha 1984, 105—108.
[15] J. Řehák and B. Řeháková: Classifications with relations: a model for the description of distributions and their distances. Kybernetika *22* (1986), 2, 158—175.
[16] J. Řehák and B. Řeháková: Vícenásobná a parciální asociace v kontingenčních tabulkách (Multiple and partial association in contingency tables). Sociologický časopis *22* (1986), 374—394.
[17] B. Řeháková: Koeficienty parciální asociace pro obecný typ kategorizované proměnné (Partial Association Coefficients for General Type of Categorical Variable). Unpublished Ph. D. Thesis 1982.
[18] B. Řeháková: Model a metoda pro analýzu kategorizovaných dat s relacemi (A Model and Method for the Analysis of Categorical Data with Relations). Ph. D. Dissertation. Department of Probability and Mathematical Statistics, Charles University, Praha 1985.
[19] B. Řeháková: Asymptotické testy hypotéz pro rozdělení D-proměnné (Asymptotic tests of hypotheses for distributions of D-variable). In: Robust 86 (J. Antoch, J. Jurečková, eds.), JČMF, Praha 1986, 121—124.
[20] J. Sheil and I. O'Muircheartaigh: The distribution of non-negative quadratic forms in normal variables. Applied Statistics *26* (1977), 92—98.

RNDr. Jan Řehák, RNDr. Blanka Řeháková, CSc., Geografický ústav ČSAV (Geographical Institute — Czechoslovak Academy of Sciences), Wenzigova 7, 120 00 Praha 2. Czechoslovakia.