Základní charakteristiky proměnných s konečným počtem hodnot a distanční analýza jejich rozložení*

JAN ŘEHÁK
Ústav pro filozofii a sociologii ČSAV, Praha

BLANKA ŘEHÁKOVÁ
Ústav pro výzkum veřejného mínění, Praha

1. Úvod

Práce s daty je založena na používání smysluplných statistických charakteristik a indexů popisujících různé aspekty rozložení dat a na teorii statistické inference. V odborné literatuře jak statistické, tak meritorních věd byl už mnohokrát vznesen požadavek jednoduchých operacionálních definic pojmů, které by umožnily vhodnou interpretaci číselných naplnění zavedených statistických měr. Příkladem toho je PRE přístup ke konstrukci koeficientů asociace, který byl od důležité práce L. Goodmana a W. Kruskala [1954] velmi rychle rozvíjen.

Tato práce se pokouší o konstrukci modelu, který by umožnil sjednocený pohled a základ a tím i sjednocenou interpretaci měr variability, komparativní analýzy a měr asociace pro distribuce proměnných s konečným počtem hodnot. Uvedený model lze aplikovat na tři základní typy proměnných, které se vyskytují ve společenských vědách: nominální, ordinální a kardinální.

Práce byla především motivována snahou zavést do analýzy dat smysluplné míry vlastností ordinálních dat a odlišit explanační a predikční koeficienty (tj. měření asociace a měření prediktivní síly), které byly zavedeny v práci Řehák [1976]. Zobecnění metody pak dává možnost pracovat podobně s velice širokou třídou proměnných, včetně areálních (které vyjadřují např. geografické vztahy) a vícerozměrných. Model také ukazuje na možnosti aplikace schématu ANOVA pro různé typy dat, jak to již bylo provedeno (model CATANOVA v práci Light, Margolin [1971]) nebo navrženo v práci Řehák [1976].

Vycházíme z matice skóre generujících typ proměnné a zavádíme míry variability, střední polohy a koncentrace. Dále ukazujeme, jak lze pro daný typ proměnné získat metriku na množině distribucí dané proměnné a jak současně platí rozkladová vlastnost variance. V souvislosti s metrikou zavádíme kovarianci distribucí a korelační koeficient distribucí. Z odvozených vlastností pak plyne přirozená definice koeficientu explanační síly rozkladu a koeficientu predikční síly rozkladové proměnné, který je založen na PRE principu.

2. Matice skóre generujících typ proměnné

V této práci předpokládáme, že proměnná A je dána soupisem svých hodnot: $A = \{a_1, a_2, \ldots, a_K\}$. Typ proměnné je dán soustavou relací na množině hodnot $\{a_i\}$. Relace mohou být vyjádřeny skóre, která generují typ proměnné. Skóre, která budeme dále značit $d(a_i, a_j) = d(i, j) = d_{ij}$ pro $i, j = 1, 2, \ldots, K$, určují míru nepodobnosti, resp. vzdálenosti (nikoli v matematickém smyslu slova) mezi každými dvěma hodnotami proměnné A. Skóre d_{ij} lze také chápat jako míru nepodobnosti mezi dvěma jedinci, z nichž jeden má hodnotu a_i a druhý a_j. Pro predikční úlohy má d_{ij} též interpretaci ztráty při predikci hodnoty a_j místo správné hodnoty a_i.

Poznámka redakce: Tento článek je součástí příspěvků, které byly prezentovány na IX. světovém sociologickém kongresu v Uppsale ve dnech 14.–19. VIII. 1978. Jejich větší část byla uveřejněna v Sociologickém časopise č. 1/1979.

Definice 1. Matici D, řádu K, $D = \|d_{ij}\|$, nazveme *maticí skóre generujících typ proměnné* $A = \{a_1, a_2, \ldots, a_K\}$, jsou-li splněny tyto požadavky:

(1)
1. *identičnost*: $d(i, i) = 0$ pro $i = 1, 2, \ldots, K$
2. *symetričnost*: $d(i, j) = d(j, i)$ pro $i, j = 1, 2, \ldots, K$
3. *nezápornost*: $d(i, j) \geqq 0$ pro $i, j = 1, 2, \ldots, K$
   $d(i, j) > 0$ alespoň pro jednu dvojici (i, j)
4. *interpretabilita*: funkce $d(i, j)$ zachovává relaci „nepodobnosti", která je definována na množině hodnot proměnné A a pro daný typ proměnné; tj. je-li (a_i, a_j) „nepodobnější" než (a_k, a_m), pak $d(i, j) \geqq d(k, m)$.

Základní typy proměnných, které se běžně vyskytují ve společenských vědách, mají modely určené definicí 2.

Definice 2. Nechť $D = \|d_{ij}\|$ je matice skóre generujících typ proměnné. Řekneme, že proměnná $A = \{a_1, a_2, \ldots, a_k\}$ je

a) *nominální* (resp. *klasifikace*), jestliže

(2) $\quad d_{ij} = 1$ pro $i \neq j$, $d_{ii} = 0$ pro všechna i, j .

b) *diskrétní ordinální* (resp. *uspořádaná klasifikace*), jestliže

(3) $\quad d_{ij} = |i - j|$ pro všechna i, j .

c) *diskrétní kardinální* (resp. *číselná*), jestliže její hodnoty jsou čísla x_i, tj. $A = \{x_1, x_2, \ldots, x_K\}$ a

(4) $\quad d_{ij} = (x_i - x_j)^2$ pro všechna i, j .

Význam skóre je dán intuitivně tak, že u nominálních proměnných mají všechny hodnoty stejnocenný nominální charakter, všechny jsou od sebe stejně „vzdálené". U diskrétního ordinálního znaku je vzdálenost dvou kategorií počet kategorií, který leží mezi nimi. Jde tedy o „neparametrická" skóre a všechny výsledky je pak nutno takto interpretovat. Skóre pro diskrétní kardinální znak odpovídají představám o eukleidovském prostoru; je možno též zavést skóre $d_{ij} = |x_i - x_j|$, která spíše odpovídají diskrétnímu charakteru zkoumané veličiny. Použitá skóre intuitivně odpovídají spíše spojitému případu (vzdušné vzdálenosti), např. jsou-li x_i středy intervalů.

3. Základní míry pro vlastnosti rozložení četností

Základní charakteristikou rozložení četností je míra variability. Stupeň vnitřní diferenciace prvků v souboru podle hlediska dané proměnné je základním kamenem statistické analýzy, a to ať již pro explanační účely, či pro přesnost predikcí, nebo stupeň typičnosti některé reprezentující hodnoty. Proto od pojmu variability vycházíme.

A. Značení

V dalším textu budeme používat jednak značení sumačního, jednak maticového, v němž $f, g, h, \ldots$ budou sloupcové vektory ze simplexu $Q_K = \{f' = (f_1, \ldots, f_K), \Sigma f_i = 1, f_i \geqq 0, i = 1, 2, \ldots, K\}$ a $D = \|d_{ij}\|$ matice skóre generujících typ proměnné. V případě, kdy budeme uvažovat soubor R populací, budeme značit vektory rozložení jako $f'_{(r)} = (f_{1/r}, f_{2/r}, \ldots, f_{K/r})$, $r = 1, 2, \ldots, R$. Rozložení f odpovídá souboru dat, který popisujeme, má tedy charakter populační charakteristiky a v dalším proto nebudeme rozlišovat pojem relativní četnosti a pravděpodobnosti. Je pochopitelně možné model použít i na výběrové soubory.

B. Míra variability: Gvar

Míra variability Gvar (resp. Genvar) vzniká přirozeně jako průměrná (očekávaná) nepodobnost mezi dvojicemi jedinců v populaci:

$$(5)^* \qquad \begin{aligned} \mathrm{Gvar} &= \underset{i}{E}\underset{j}{E}\, d(i,j) \\ &= \sum_i \sum_j f_i f_j d(i,j) \\ &= f' D f \end{aligned}$$

Čím větší je hodnota Gvar, tím větší je variabilita v dané populaci. Proto je Gvar operacionální definicí stupně variability vzhledem k relacím určeným maticí D. Vlastnosti míry variability Gvar:

a) Gvar $= 0$ právě tehdy, když je populace stoprocentně homogenní (tj. existuje-li $f_i = 1$) v případě, že matice D je taková, že $d_{ij} > 0$ pro všechna $i \neq j$. Tuto vlastnost má většina běžně používaných proměnných.

b) Pro obecnou matici $D = \|d_{ij}\|$ platí, že Gvar $= 0$ právě když $\sum_W f_i = 1$ pro všechny množiny W, což jsou množiny indexů, pro něž $d_{ij} = 0$ pro všechna $i, j \in W$. Vlastnost a) je speciálním případem, v němž všechny takové množiny W jsou jednobodové.

c) Nechť $\bar{d} = \max_{(i,j)} d_{ij}$, t je velikost množiny indexů T, pro niž platí, že příslušná submatice D_T řádu $t \times t$ má prvky $d_{ij} = (1 - \delta_{ij})\,\bar{d}$, kde δ_{ij} je Kroneckerovo delta a neexistuje submatice většího řádu s touto vlastností. Pak Gvar nabývá svého maxima

$$\max_{Q_K} \mathrm{Gvar} = \frac{t-1}{t}\,\bar{d}$$

pro rozložení $f_{\max}$ s hodnotami $f_i = \frac{1}{t}$ pro $i \in T$ a $f_i = 0$ pro $i \notin T$. Častý speciální případ:

$$T = \{r, s\}\,, \quad \bar{d} = d_{rs}\,, \quad f_r = f_s = \tfrac{1}{2}, \quad \max \mathrm{Gvar} = \frac{d_{rs}}{2}\,.$$

Důkaz vlastností míry Gvar:

Vlastnost a) je zřejmá z toho, že pro jakékoli rozložení s $f_i = 1$ je Gvar $= 0$ a pro jakékoli rozložení, v němž jsou alespoň dvě složky nenulové, je Gvar kladný. Vlastnost b): Nechť Gvar $=$ $= 0$ a nechť existují $f_i \neq 0$, $f_j \neq 0$ takové, že $d_{ij} \neq 0$. Pak Gvar $\geqq 2 f_i f_j d_{ij} > 0$, což je spor. Pro Gvar $= 0$ musí tedy platit, že f má nenulové složky jen na takové podmnožině hodnot W, pro kterou je $d_{ij} = 0$ pro $i, j \in W$. Opačně platí implikace okamžitě. Vlastnost c): Nechť $\bar{d} =$ $= \max_{i,j} d_{ij}$, T je množina indexů velikosti t pro kterou platí: $r, s \in T \Rightarrow d_{rs} = (1 - \delta_{rs})\,\bar{d}$. Pro f_r takové, že $r \in T$, $\sum_T f_r = \bar{f} \leqq 1$ platí, že maximum výrazu

$$\sum_T \sum_T f_r f_s \bar{d} = \bar{d}\left[\left(\sum_T f_r\right)^2 - \sum_T f_r^2\right] = \bar{d}\left(\bar{f}^2 - \sum_T f_r^2\right)$$

je dosaženo pro $f_r = \frac{\bar{f}}{t}$ a má hodnotu

$$\bar{d}\left(\bar{f}^2 - \frac{\bar{f}^2}{t}\right) = \bar{d}\bar{f}^2\,\frac{t-1}{t}\,.$$

Jelikož $1-\bar{f}$ je rozloženo na $\bar{T}$, pak

$$\text{Gvar} = \sum\sum f_i f_j d_{ij} \leqq \bar{d}\bar{f}^2 \frac{t-1}{t} + 2 \sum_{i\in T' j\in\bar{T}} \sum f_i f_j d_{ij} + \sum_{i\in\bar{T} j\in\bar{T}} \sum f_i f_j d_{ij} \leqq$$

$$\leqq \bar{d}\bar{f}^2 \frac{t-1}{t} + 2 \sum_{i\in\bar{T} j\in\bar{T}} \sum f_i f_j \bar{d} + \sum_{i\in\bar{T} j\in\bar{T}} \sum f_i f_j \bar{d} = \bar{d}\bar{f}^2 \frac{t-1}{t} + 2\bar{f}(1-\bar{f})\bar{d} + (1-\bar{f})^2 \bar{d} =$$

$$= \bar{d}\left[1 - \frac{\bar{f}^2}{t}\right] = H(\bar{f})$$

$\max_{Q_K} \text{Gvar} \leqq H(\bar{f})$ a je dosaženo pro $\bar{f} = 1$ při $t > 1$, tj. $\max_{Q_K} \text{Gvar} = H(1) = \bar{d}\frac{t-1}{t}$.

Kdyby existovalo $T' \supset T$, $t' > t$, pak $\bar{d}\frac{t'-1}{t'} > \bar{d}\frac{t-1}{t}$. Proto je maxima dosaženo pro takovou množinu T s danou vlastností, která obsahuje největší možný počet indexů.

Q. E. D.

Míra variability Gvar je mírou vnitřní diferenciovanosti souboru a mírou stupně variability. Je základem pro zkoumání různorodosti souboru a pro zkoumání vlivů, které na tuto různorodost působí. V dalším půjde o popsání vlastností variability daného rozložení.

C. Míra koncentrace kolem hodnoty a: d_a^*

Mírou koncentrace kolem hodnoty a nazveme průměrné (očekávané) skóre vzdálenosti od této hodnoty.

(6) $$d_a^* = E_i d(a_i, a) = \sum_i f_i d(a_i, a)$$

Hodnotu znaku A, pro kterou je d_j^* minimální, nazveme *centrem* c (středem) *rozložení* f.

(7) $$c = a_i \quad \text{takové, že} \quad d_i^* = \min_j d_j^*$$

Platí, že a)

(8) $$\text{Gvar} = E d_j^* = \sum f_j d_j^*$$

b) rozložení může mít více středů,

c) středem může být neobsazená hodnota znaku, tj. $f_c = 0$.

D. Míra koncentrace kolem středu: d^*

Jako obdobu vlastní definice rozptylu pro číselné veličiny můžeme zavést míru koncentrace kolem středu jako očekávanou odchylku všech pozorování od středu c

(9) $$d^* = E d(a_i, c) = d_c^* = \sum f_i d(a_i, c)$$

Čím větší je hodnota d^*, tím vyšší je odchýlenost populace od centra rozložení.

Lze říci, že d^* je též stupněm atypičnosti středu pro dané rozložení. Čím menší je d^*, tím typičtějším představitelem populace centrum je. V predikčním modelu pak hodnota d^* charakterizuje kvalitu optimální predikce.

Vlastnosti d^*:

a) $d^* = 0$ je ekvivalentní s tím, že $f_i \neq 0$ pouze pro takové hodnoty, pro které $d(a_i, c) = 0$

b) $d^* \leqq$ Gvar.

4. Komparace distribucí — (polo)metrika generovaná maticí D

Z testování hypotéz známe řadu statistik, které umožňují komparaci dvou či více rozložení. Distribuce můžeme komparovat též přímo pomocí některé metriky, viz např. Wishart, Leach [1970], Charvát [1977]. Testovací statistiky jsou většinou též založeny, ať už přímo či intuitivně, na metrických vlastnostech vzdáleností a vznikají jako standardizace metriky na veličiny mající vhodné statistické vlastnosti. Na pojmu vzdálenosti či podobnosti jsou založeny celé moderní metodologie zpracování dat, jako například seskupovací analýza, multidimenzionální škálování, různé neparametrické přístupy k podobnostní analýze, metody AID.

Zde zavedeme důležitou třídu metrik na množině Q_K (Q_K je jak už uvedeno výše K-rozměrný simplex: $Q_K \equiv \{f : f' = (f_1, \ldots, f_K), f_i \geqq 0, i = 1, \ldots, K, \Sigma f_i = 1\}$). Ukážeme, že Q_K tvoří metrický prostor s metrikou určenou prvky matice D. Nejprve vyslovíme definici pojmu, o nějž se budeme v celé další části opírat.

Definice 3. Funkci (reálnou) $D(f, g)$ na $Q_K \times Q_K$ nazveme *polometrikou* (*pseudometrikou*), jestliže platí:

(10)
1. $f = g \Rightarrow D(f, g) = 0$
2. $D(f, g) = D(g, f)$
3. $D(f, g) + D(g, h) \geqq D(f, h)$ pro všechna $f, g, h \in Q_K$.

Funkci $D(f, g)$ nazveme *metrikou*, jestliže navíc platí

1'. $D(f, g) = 0 \Rightarrow f = g$.

Přechod od pseudometriky k metrice lze provést tak, že zavedeme množinu Q_K^+ jako množinu ekvivalentních tříd distribucí vzhledem k $D(f, g) = 0$. Označíme-li tyto ekvivalentní třídy f^+ a položíme-li $D^+(f^+, g^+) = D(f, g)$, kde $f \in f^+$ a $g \in g^+$, pak je-li $D(f, g)$ polometrika na Q_K, je $D^+(f^+, g^+)$ metrika na Q_K^+.

V dalším ukážeme, jak lze určit polometriku, resp. metriku na základě matice skóre generujících typ proměnné. Taková polometrika bude mít interpretaci vzhledem k matici D a její věcný význam v analýze dat bude záviset na významu skóre $d(i, j)$.

Budeme potřebovat ještě tuto definici:

Kvadratickou formu $X'AX$ nazveme pozitivně definitní na množině S, jestliže

(11) $X'AX \geqq 0$ pro $X \in S$ a

(12) $X'AX = 0 \Leftrightarrow X = 0$, je-li $0 \in S$.

Matici A pak nazveme pozitivně definitní maticí vzhledem k S. Obdobně definujeme pozitivně semidefinitní, negativně definitní a negativně semidefinitní formy a matice vzhledem k S (pro semidefinitnost se nevyžaduje vlastnost (12)).

Věta 1. Buď $D = \|d_{ij}\|$, matice skóre generujících typ proměnné řádu K. Pak platí následující tři tvrzení.

1. Funkce

$$D(f, g) = \sqrt{(f - g)' D(g - f)} \tag{13}$$

je polometrika právě tehdy, když matice $D^* = \|d^*_{ij}\|$ řádu $K - 1$, kde

$$d^*_{ij} = d_{iK} + d_{Kj} - d_{ij} \tag{14}$$

je pozitivně semidefinitní matice. $D(f, g)$ nazveme D-polometrikou na Q_K. Je-li $D(f, g)$ metrika, nazveme ji D-metrikou na Q_K.

2. Je-li $D(f, g)$ D-polometrika, pak každá z následujících podmínek je postačující k tomu, aby $D(f, g)$ byla D-metrika.

a) D je regulární a $\Sigma\Sigma d^{(-1)}_{ij} \neq 0$, kde $D^{-1} = \|d^{(-1)}_{ij}\|$.

b) Blokova matice $G = \begin{Vmatrix} D & J \\ J' & 0 \end{Vmatrix}$ je regulární; $J' = (1, 1, \ldots, 1)$.

c) D^* je pozitivně definitní matice.

3. Je-li $D(f, g)$ D-polometrika, pak ke každému $f \in Q_K$ existuje ekvivalentní třída f^+, do níž patří všechny distribuce z Q_K splňující podmínku

$$g = f + \tfrac{1}{2}\lambda D^- J + (D^- D - I) Z \tag{15}$$

pro reálná λ, kde D^- je zobecněná inverze matice D, I je jednotková matice, Z je libovolný vektor. Množina Q^+_K všech f^+ tvoří metrický prostor s D-metrikou určenou

$$D^+(f^+, g^+) = D(f, g)\ ; \quad f \in f^+,\ g \in g^+ . \tag{16}$$

Důkaz.

1a) Forma $X'DX$ je negativně semidefinitní (definitní) vzhledem k $S = \{X : \Sigma x_i = 0\}$ právě když forma $Y'D^*Y$ je pozitivně semidefinitní (definitní). Položme $Y' = (x_1, x_2, \ldots, x_{K-1})$.

Na S platí
$$\begin{aligned} X'DX &= \sum^{K}\sum^{K} x_i x_j d_{ij} = \sum^{K-1}\sum^{K-1} x_i x_j d_{ij} \\ &+ \sum^{K-1} x_i x_K d_{iK} + \sum^{K-1} x_K x_j d_{Kj} = \\ &= \sum^{K-1}\sum^{K-1} x_i x_j (d_{ij} - d_{iK} - d_{Kj}) = -Y'D^*Y \end{aligned}$$

Odtud plyne, že $X'DX$ je negativně semidefinitní na $S \Leftrightarrow Y'D^*Y$ je pozitivně semidefinitní. Na S je také $Y = 0 \Leftrightarrow X = 0$, a proto $X'DX$ je negativně definitní na $S \Leftrightarrow Y'D^*Y$ je pozitivně definitní.
Je-li $D(f, g)$ polometrika, pak D je negativně semidefinitní na Q_K, a tedy D^* je pozitivně semidefinitní.

1b) Nechť D^* je pozitivně semidefinitní reálná symetrická matice. Existuje tedy rozklad $P'D^*P = \Lambda$, kde P je ortogonální a Λ diagonální matice, λ_i jsou charakteristická čísla. Proto je D^* Gramova matice a pro $Y'D^*Z$ platí Cauchy-Schwartz-Buňakovského nerovnost. Dále platí, že $D^2(f, g) = (f - g)' D(g - f) = (\bar{f} - \bar{g})' D^*(\bar{f} - \bar{g})$, kde $\bar{f}' = (f_1, f_2, \ldots, f_{K-1})$, $\bar{g}' = (g_1, g_2, \ldots, g_{K-1})$ a obdobně i $(f - g)' D(g - h) = (\bar{g} - \bar{f})' D^*(\bar{g} - \bar{h})$. Vzhledem k nezápornosti členů je vztah

$$D(f, g) + D(g, h) \geqq D(f, h) \tag{17}$$

ekvivalentní vztahu

$$D^2(f, g) + D^2(g, h) + 2D(f, g)\, D(g, h) \geqq D^2(f, h)$$

a ten je vzhledem k rovnosti

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

ekvivalentní vztahu

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

který však je identicky splněn vzhledem k tomu, že D^* je Gramova matice, pro kterou platí

$$-(\bar{g} - \bar{f})'\, D^*(\bar{g} - \bar{h}) \leqq |(\bar{g} - \bar{f})'\, D^*(\bar{g} - \bar{h})| \leqq D(f, g)\,.\, D(g, h)\,.$$

Trojúhelníková nerovnost (17) tedy platí, a proto části 1 a 2c věty 1 jsou dokázány, neboť ostatní vlastnosti plynou okamžitě.

2. Je-li $D(f, g)$ polometrika, pak forma $X'DX$ dosahuje na $S = \{X : \Sigma x_i = 0\}$ maxima nuly, je-li splněna podmínka (a) nebo (b). To plyne z věty o Lagrangeových multiplikátorech. Zaveďme funkci

$$F(X) = X'DX - \lambda\, \Sigma x_j$$

Platí

$$\frac{\partial F(X)}{\partial X} = 2DX - \lambda J = 0$$

Řešení rovnice $DX = \frac{1}{2}\lambda J$ je (viz Rao [1965], str. 24)

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

kde Z je libovolný vektor.

Rovnice má právě jedno řešení v případě že D je regulární.

Z podmínky

$$0 = J'X = \tfrac{1}{2}\lambda J' D^{-1} J = \tfrac{1}{2}\lambda\, \Sigma\, \Sigma d_{ij}^{(-1)}$$

vidíme, že je-li $\Sigma\, \Sigma d_{ij}^{(-1)} \neq 0$, pak nutně $\lambda = 0$ a $X = 0$ je jediným řešením rovnice $DX = \frac{1}{2}\lambda J$. Je-li $\lambda \neq 0$, pak existuje nenulové řešení X.

Poznamenejme, že

$$\det G = (-1)^K \det D\,.\,\det (J' D^{-1} J) = (-1)^K \det D\,.\,\Sigma\, \Sigma d_{ij}^{(-1)}.$$

f^+ tvoří ekvivalentní třídu: reflexivnost a symetričnost jsou zřejmé, neboť $D(f, g) = 0 \Rightarrow$ $\Rightarrow D(g, f) = 0$ a $D(f, f) = 0$. Tranzitivita je dána trojúhelníkovou nerovností.

$$D(f, g) = D(g, h) = 0 \Rightarrow 0 \leqq D(f, h) \leqq D(f, g) + D(g, h) = 0$$

Q. E. D.

Poznámky k větě 1.

1. Matice D^* vznikne tak, že od všech řádků odečteme poslední řádek a od všech sloupců odečteme poslední sloupec a redukujeme řád o K-tý řádek a K-tý sloupec. Věta ovšem platí i tehdy, jestliže tuto proceduru provedeme s kterýmkoli jiným indexem, tj. odečteme a vypustíme s-tý řádek a sloupec ($s = 1, 2, \ldots, K$). Ověřování podmínky pro konkrétní případ tedy není omezeno na manipulaci s posledním řádkem a sloupcem.

2. D-polometrika vytváří ekvivalentní třídy distribucí, které se v analýze ztotožňují. Dále uvidíme, že všechny distribuce číselné proměnné se stejným prům-

rem patří do jedné ekvivalentní třídy. Proto určení a charakterizace těchto ekvivalentních tříd je pro práci s daty velice důležitá.

Pro ověřování možnosti aplikace modelu je vhodná věta 2.

Věta 2. Nechť D je matice skóre vytvářejících typ proměnné. Pak platí:

1. Pro $K = 2$, $d_{12} \neq 0$, je $D(f, g)$ násobek Eukleidovské metriky

$$D(f,g) = \sqrt{2d_{12}}\,|f_1 - g_1| = \sqrt{2d_{12}}\,|f_2 - g_2| \tag{18}$$

2. Pro $K = 3$ je $D(f, g)$ D-polometrika právě tehdy, když

$$d^2 \leqq 4d_{ij}d_{jk} \tag{19}$$

pro všechny permutace indexů (1, 2, 3) a $D(f, g)$ je D-metrika, když

$$\bar{d}^2 < 4d_{ij}d_{jk} \tag{20}$$

pro všechny permutace indexů (1, 2, 3).

3. Pro $K > 3$ je platnost vztahu

$$\bar{d}^2 < 4d_{ij}d_{jk} \tag{21}$$

nutnou podmínkou proto, aby $D(f, g)$ byla D-polometrika pro všechny trojice indexů i, j, k.

Přitom

$$\bar{d} = d_{ik} - d_{ij} - d_{kj}\,.$$

Důkaz.

1. Pro $K = 2$ je

$$D(f,g) = \sqrt{2d_{12}(f_1 - g_1)(g_2 - f_2)} = \sqrt{d_{12}[(f_1 - g_1)^2 + (f_2 - g_2)^2]} = \sqrt{2d_{12}}\,|f_1 - g_1|$$

2. Pro $K = 3$ provedeme diskusi formy vzhledem k $S = \{X : \Sigma x_j = 0\}$.
$x_i d_{ij} x_j + x_j d_{jk} x_k + x_i d_{ik} x_k = \frac{1}{2} X'DX = F$

Položme $x_j = -x_i - x_k$.

Pak $$F = -x_i^2 d_{ij} - x_k^2 d_{kj} + 2x_i x_k \left(\frac{\bar{d}}{2}\right).$$

Diskriminant δ formy F je roven

$$\delta = \begin{vmatrix} -d_{ij} & \frac{1}{2}\bar{d} \\ \frac{1}{2}\bar{d} & -d_{kj} \end{vmatrix} = d_{ij}d_{jk} - \tfrac{1}{4}\bar{d}^2$$

a) Je-li $\delta > 0$, je F eliptická s negativními prvky na diagonále, je tedy negativně definitní.

b) Je-li $\delta = 0$, je F parabolická, existuje netriviální nulové řešení, forma je negativně semi-

definitní. Pro $d_{ij} \neq 0$ je nenulové řešení dáno: $x_i = \frac{1}{2}\frac{\overline{d}}{d_{ij}}x_k$. Pro $d_{ij} = 0$, $d_{jk} \neq 0$ je $x_k = 0$, $x_i = -x_j$.

c) Je-li $\delta < 0$, je F hyperbolická, mění znaménko, a proto nemůže odpovídat metrice. Ostatní plyne podobně jako v důkazu věty 1.

3. Vztah (21) je nutný proto, aby $D(f, g)$ byla D-polometrika v obecném případě $K > 3$. Není-li totiž splněna pro některou trojici indexů i, j, k, pak lze nalézt trojici čísel x_i, x_j, x_k různých od nuly tak, že $x_i + x_j + x_k = 0$ a $x_ix_jd_{ij} + x_ix_kd_{ik} + x_jx_kd_{jk} > 0$. Položíme-li ostatní složky vektoru X rovny nule, máme spor s předpokladem negativní semidefinitnosti matice D vzhledem k $S = \{X : \Sigma x_i = 0\}$.

Q. E. D.

Věta 2. ukazuje, že pro $K = 2$ existuje jednoduché zobrazení Q_K na $\langle 0, 1\rangle$, resp. na $\langle 0, \sqrt{2d_{12}}\rangle$. Proto nemá význam pomocí složitých technik a koeficientů podobnosti provádět multidimenzionální škálovací techniky na rozložení dichotomických dat, neboť škálování do přímky je jednoznačně a přirozeně určeno. To ovšem platí i pro K-rozměrný znak. Nemá význam provádět škálování zobrazením do více rozměrů eukleidovského prostoru, než je hodnost matice D^*. Škálováním například dichotomických proměnných do vícerozměrného prostoru tedy dostáváme pouze artefakty, které jsou způsobeny neadekvátností použité míry podobnosti distribucí.

5. Dekompozice zobecněné variance

Zobecněná variance Genvar může být rozložena na dvě složky, rozkládáme-li populaci na několik částí. V dalším budeme předpokládat, že je dán rozkladový znak $B = (b_1, b_2, \ldots, b_R)$, který určuje rozklad na zkoumané populaci, a že mu odpovídá rozložení $p' = (p_1, \ldots, p_R)$, $p \in Q_R$. Označme $f'_{(r)} = (f_{1/r}, f_{2/r}, \ldots, f_{K/r})$ rozložení znaku A na jednotlivých podmnožinách zkoumané populace a h rozložení celé populace, $h = \Sigma p_r f_{(r)}$. Věta 3 dává námět na zobecněnou analýzu rozptylu (GANOVA).

Věta 3. Buď $D^2(f, g) = (f - g)' D(g - f)$ pozitivně semidefinitní forma vzhledem k $S = \{f - g, f \in Q_K, g \in Q_K\}$. Pak pro soubor distribucí $f_{(r)} \in Q_K$, $r = 1, 2, \ldots, R$ a jejich směs $h = \Sigma p_r f_{(r)}$, $p_r \in Q_R$ platí

$$\text{(22)} \qquad \begin{aligned} \operatorname{Gvar} h &= \sum_r p_r \operatorname{Gvar} f_{(r)} + \tfrac{1}{2}\sum_r\sum_s p_r p_s D^2(f_{(r)}, f_{(s)}) \\ &= \sum_r p_r \operatorname{Gvar} f_{(r)} + \sum_r p_r D^2(f_{(r)}, h) \end{aligned}$$

Důkaz.

Jednoduchým rozpisem se přesvědčíme, že

$$\begin{aligned} \operatorname{Gvar} h &= h'Dh = (\Sigma p_r f_{(r)})' D(\Sigma p_r f_{(r)}) \\ &= \Sigma\Sigma p_r p_s f'_{(r)} D f_{(s)} \\ &= \tfrac{1}{2}\Sigma\Sigma p_r p_s f'_{(r)} D f_{(s)} + \tfrac{1}{2}\Sigma\Sigma p_r p_s f'_{(s)} D f_{(r)} - \tfrac{1}{2}\Sigma\Sigma p_r p_s f'_{(r)} D f_{(r)} - \\ &\quad - \tfrac{1}{2}\Sigma\Sigma p_r p_s f'_{(s)} D f_{(s)} + \Sigma p_r f'_{(r)} D f_{(r)} \\ &= \Sigma p_r \operatorname{Gvar} f_{(r)} + \tfrac{1}{2}\Sigma\Sigma p_r p_s D^2(f_{(r)}, f_{(s)}) \end{aligned}$$

Obdobně

$$\begin{aligned} \operatorname{Gvar} h &= h'Dh = (\Sigma p_r f_{(r)})' Dh = \Sigma p_r f'_{(r)} Dh \\ &= \Sigma p_r f'_{(r)} Dh - \Sigma p_r h' Dh + \Sigma p_r f'_{(r)} Dh \end{aligned}$$

$$+ \Sigma p_r f'_{(r)} Df_{(r)} - \Sigma p_r f'_{(r)} Df_{(r)}$$
$$= \Sigma p_r (f_{(r)} - h)' D(h - f_{(r)}) + \Sigma p_r f'_{(r)} Df_{(r)}$$
$$= \Sigma p_r \operatorname{Gvar} f_{(r)} + \Sigma p_r D^2(f_{(r)}, h)$$

Q. E. D.

I když lze uvedenou vlastnost formulovat obecněji, zde se omezujeme na negativně semidefinitní matice vzhledem k S, tak jak to odpovídá úloze analýzy dat. Negativní semidefinitnost matice D vzhledem k S je ovšem ekvivalentní tomu, že $D(f, g)$ je D-polometrika. Tento rozklad umožňuje odvodit neparametrickou analýzu rozptylu pro různé typy kategorizovaných dat, alespoň v její asymptotické teorii, vycházeje z asymptotické normality četností $f_{(r)}$. Pro nominální znaky je toto uděláno v práci Light, Margolin [1971]. Rozklad však také umožní tvorbu smysluplných koeficientů asociace mezi rozkladovými a závislými znaky.

6. Skalární součin dvou populací a jejich korelační koeficient

Pro Gvar a $D^2(f, g)$ lze napsat analogický vztah jako pro vzdálenost, normy a skalární součin ve vektorových prostorech:

$$\text{(23)} \qquad D^2(f, g) = \operatorname{Gvar} f + \operatorname{Gvar} g - 2s(f, g)$$

Odtud dostaneme výraz pro $s(f, g)$, který nazveme *D-kovariancí* distribucí f a g:

$$\text{(24)} \qquad \begin{aligned} s(f, g) &= f'Df + g'Dg - f'Dg \\ &= f'Dg - D^2(f, g) \\ &= \tfrac{1}{2}[f'Df + g'Dg - D^2(f, g)] \end{aligned}$$

D-kovariance je omezena dosažitelnými hranicemi

$$\text{(25)} \qquad -f'Dg \leqq s(f, g) \leqq f'Dg\,,$$

přičemž levá nerovnost se změní v rovnost právě tehdy, když $\operatorname{Gvar} f = \operatorname{Gvar} g = 0$ a pravá nerovnost se změní v rovnost právě tehdy, když $D(f, g) = 0$. Dále platí, že

$$\text{(26)} \qquad f'Dg = \tfrac{1}{2}(f'Df + g'Dg + D^2(f, g))$$

Protože $0 \leqq D^2(f, g) \leqq 2f'Dg$ a horní hranice je dosažena pouze pro nerozptýlené rozložení ($\operatorname{Gvar} f = \operatorname{Gvar} g = 0$), můžeme pro $D(f, g) \neq 0$ normalizovat $s(f, g)$ a definovat *korelační koeficient dvou rozložení vzhledem k D*,

$$\text{(27)} \qquad \operatorname{corr}(f, g) = \frac{s(f, g)}{f'Dg} = 1 - \frac{D^2(f, g)}{f'Dg}\,,$$

který nabývá hodnot z intervalu $(-1, +1)$, přičemž jeho krajní hodnoty jsou dosahovány při maximální a minimální vzdálenosti obou rozložení.

Definice 4. Dvě distribuce $f, g \in Q_K$, pro které $D(f, g) \neq 0$ nazveme *ortogonální vzhledem k D*, jestliže $s(f, g) = 0$.

Věta 4. Dvě distribuce $f, g \in Q_K$ jsou ortogonální vzhledem k D právě když platí jedna z podmínek (a), (b):

$$\text{(28)} \qquad \text{(a)} \quad f'Dg = D^2(f, g)$$

$$\text{(29)} \qquad \text{(b)} \quad \operatorname{Gvar} f + \operatorname{Gvar} g = D^2(f, g) \quad \text{(Pythagorova věta)}$$

Důkaz věty 4 plyne okamžitě ze vztahů (24).

Pojem kovariance, korelace a ortogonality dvou distribucí vzhledem k D může být užitečný v analýze dat. Založíme-li komparativní analýzu na corr (f, g), dostáváme výsledky invariantní k násobení matice D konstantou a normalizované v intervalu $(-1, +1)$. To znamená, že tato míra umožňuje komparovat populace z hlediska různých proměnných A_i i jejich různých typů. Je to tedy krok ke sjednocení komparativní analýzy, která je založena na stejném principu pro různé typy proměnných.

7. Míra explanační a predikční síly rozkladu

Rozklad uvedený větou 3 umožňuje zavést třídu koeficientů asociace mezi nominálním znakem a jakýmkoli znakem charakterizovaným maticí D. Na dekompozici jsou založeny korelační poměr η^2, a Wallisovo τ (viz Goodman, Kruskal [1954]), a β (viz Řehák [1976]), stejně jako i další koeficienty odvozené z PRE principu. Na principu rozkladu variability tedy můžeme podle stejného schématu zavést viz Řehák [1976])

(30) celková variabilita = variabilita uvnitř oblastí + + variabilita mezi oblastmi

což je umožněno právě větou 3.

Je-li tedy dán rozkladový znak, který má R-tříd, můžeme definovat *koeficient explanační síly rozkladu* jako relativní podíl vysvětlené variance

$$\delta = 1 - \frac{\Sigma p_r \operatorname{Gvar} f_{(r)}}{\operatorname{Gvar} h} = \frac{\frac{1}{2} \Sigma \Sigma p_r p_s D^2(f_{(r)}, f_{(s)})}{\operatorname{Gvar} h}, \tag{31}$$

použijeme-li značení věty 3.

Vlastnosti koeficientu plynou z věty 3.

1. $0 \leqq \delta \leqq 1$
2. $\delta = 0$ právě když $f_{(r)}$, $r = 1, \ldots, R$ patří do jedné třídy ekvivalence f^+ vzhledem k $D(f, g) = 0$.
3. $\delta = 1$ právě když zobecněná variance ve všech oblastech vymizí, tj. $\operatorname{Gvar} f_{(r)} = 0$ pro všechna $r = 1, \ldots, R$.
4. Koeficient je definován pouze pro $\operatorname{Gvar} h \neq 0$.

I když tento koeficient je možné formulovat v termínech predikčního přístupu PRE, navrhli jsme nový název, neboť situace predikčnímu přístupu neodpovídá. PRE model vychází z interpretace:

a) $d(i, j)$ = ztráta daná predikcí i, má-li proměnná pro daný objekt hodnotu j,
b) predikci provádíme proporcionálně k pravděpodobnostem v jednotlivých třídách,
c) δ je poměr chyby predikce při znalosti hodnoty stratifikačního znaku a chyby predikce bez znalosti hodnoty stratifikačního znaku.

Proporcionální predikce není optimální, a proto v predikční situaci bude lépe použít jiné míry — *koeficientu predikční síly rozkladu*. Ten je založen na stejných předpokladech a), c), ale predikční pravidlo b) je optimální: proveď predikci, která má nejmenší očekávanou chybu. Tomuto modelu odpovídá predikce centrem rozložení a mírou kvality predikce je míra koncentrace kolem této hodnoty, tj. d^*. Koeficient predikční síly rozkladu je dán vzorcem

(32) $$\delta^* = 1 - \frac{\Sigma p_r d^*_{(r)}}{d^*},$$

kde d^* je míra koncentrace kolem středu pro celkové rozložení a $d^*_{(r)}$ je míra koncentrace kolem středu pro r-tou oblast.

Vlastnosti koeficientu δ^*:

1. $0 \leqq \delta^* \leqq 1$
2. Je-li $f_{(r)} = h$ pro $r = 1, \ldots, R$ (případ statistické nezávislosti rozložení v oblastech), pak $\delta^* = 0$. Opačně platí pouze: je-li $\delta^* = 0$, pak střed c rozložení h je zároveň středem každého rozložení $f_{(r)}$.
3. $\delta^* = 1$ právě tehdy, když vymizí všechny odchylky od centra v oblastech, tj. všechna $d^*_{(r)} = 0$.
4. Koeficient je definován pouze pro případy $d^* \neq 0$.

Důkaz vlastností koeficientu δ^*

1. $d^* \geqq 0$ a $\Sigma p_r d^*_{(r)} \geqq 0$, neboť $D = \|d_{ij}\|$ má nezáporné prvky. Označme c_r a c centra rozložení $f_{(r)}$ a h. Pak platí:

$$d^* = d^*_c = \Sigma h_j d_{jc} = \Sigma \Sigma p_r f_{(r)i} d_{ic} = \Sigma p_r \Sigma f_{(r)j} d_{jc}$$

$$\geqq \Sigma p_r d^*_{(r)c} = \Sigma p_r d^*_{(r)}$$

Odtud $0 \leqq \Sigma p_r d^*_{(r)}/d^* \leqq 1$, a tudíž $0 \leqq \delta^* \leqq 1$.

2. Je-li $f_{(r)} = h$ pro všechna r, platí $d^*_{(r)} = d^*$, a tudíž $\delta^* = 0$. Je-li $\delta^* = 0$, pak $d^* = \Sigma p_r d^*_{(r)}$. Předpokládejme dále, že $p_r > 0$ a že alespoň pro jedno s platí, že c není středem rozložení $f_{(s)}$, tj. $d^*_{(s)} < d^*_{(s)c}$. Pak $d^* = \Sigma p_r d^*_{(r)c} > \Sigma p_r d^*_{(r)}$, což je spor. To znamená, že pro $\delta^* = 0$ musí c patřit ke středům u každého z rozložení $f_{(r)}$.

3. $\delta^* = 1 \Leftrightarrow \Sigma p_r d^*_{(r)} = 0 \Rightarrow d^*_{(r)} = 0$ pro všechna r.

4. Vlastnost je zřejmá.

Q. E. D.

Wilson [1971] zadal tři požadavky na predikční pravidlo, které má být základem PRE koeficientů:

1. *Úplná prediktabilita* (totally predictive rule): pravidlo predikce závisí jen na charakteristikách společného rozložení závislé a stratifikační proměnné a nevychází z žádné informace o hodnotách predikované proměnné. To je u δ^* splněno, neboť predikční pravidlo vychází pouze z hodnot d^* a $d^*_{(r)}$.
2. *Přesnost v průměru* (rule accurate in average): pravidlo má tu vlastnost, že očekávaná hodnota predikce z modelu koinciduje s průměrným výsledkem predikcí z celého souboru. Tento požadavek je splněn vzhledem k procentu správné predikce, neboť každému jedinci přiřazujeme hodnotu středu rozložení, jejíž pravděpodobnost je stejná jako relativní četnost v souboru.
3. *Minimalizace chyby* (minimum — error rule): pravidlo minimalizuje chybu. Na tomto základě je koeficient δ^* konstruován. Právě třetí Wilsonův požadavek, s nímž autoři plně souhlasí, neboť patří k požadavkům predikce jakožto praktické úlohy, vede na rozlišení dvou typů koeficientů a k navržení jiného názvu pro koeficienty typu δ (ty splňují požadavky 1 a 2, ale ne 3).

Naproti tomu koeficient δ splňuje důležitou vlastnost, kterou δ^* postrádá: *rozložitelnost míry variability*, tj. existence dvou aditivních složek vztahu (29), které jsou zaručeny větou 3. Tato vlastnost pak zajišťuje nejen uvedené vlastnosti koeficientu δ, ale také jeho přirozenou interpretaci.

(33) $100\delta\ \%$ = procento variability, které u závislé proměnné vysvětluje rozklad nezávislou proměnnou.

V dalších částech uvidíme, které známé a používané koeficienty splňují tuto vlastnost jako speciální případy koeficientu δ.

8. Nominální znak (klasifikace)

Vztahem (2) jsme definovali nominální typ znaku, u něhož jsou všechny dvojice hodnot stejně nepodobné, takže d_{ij} pouze indikují, že každé dvě hodnoty jsou z hlediska typu rozdílné. Vlastnosti matice D a D^* je možno ověřit přímo, stejně jako zavedení vzdálenosti a dalších měr. Výsledky, které jsou známé, lze shrnout ve větě 5.

Věta 5. (Vlastnosti nominálních znaků.) Nechť matice skóre vytvářející typ proměnné $D = \|d_{ij}\|$ je dána skóre $d_{ij} = 1 - \delta_{ij}$, kde δ_{ij} je Kroneckerovo delta a $A = \{a_1, \ldots, a_K\}$ je proměnná s distribucemi $f_{(r)}$, $r = 1, \ldots, R$. Nechť $f,\ g \in Q_K$, $p \in Q_R$, $f = \Sigma p_r f_{(r)}$. Pak

1.

$$\text{(34)} \qquad \text{Genvar}\, f = \text{nomvar}\, f = 1 - \sum_{i=1}^{K} f_i^2$$

$$\text{(35)} \qquad \begin{aligned} d_i^* &= 1 - f_i \quad \text{pro všechna} \quad i = 1, \ldots, K \\ d^* &= 1 - f_M\,, \end{aligned}$$

kde M je index středu rozložení daného vztahem $f_M = \max_i f_i$, středem rozložení je jeho modus M.

2. Nomvar nabývá svého maxima $\dfrac{K-1}{K}$ pro $f_i = \dfrac{1}{K}$, $i = 1, \ldots, K$.

3. Matice D^* je pozitivně definitní a vytváří Eukleidovskou metriku.

$$\text{(36)} \qquad D(f, g) = \sqrt{\sum_{i=1}^{K} (f_i - g_i)^2}$$

4. Skalární součin dvou distribucí f, g je určen

$$\text{(37)} \qquad \begin{aligned} s(f, g) &= 1 - \Sigma f_i^2 - \Sigma g_i^2 + \Sigma f_i g_i = \\ &= 1 - \tfrac{1}{2}(\Sigma f_i^2 + \Sigma g_i^2) - \tfrac{1}{2}\Sigma (f_i - g_i)^2 \end{aligned}$$

a má hranici

$$\text{(38)} \qquad f' D g = 1 - \Sigma f_i g_i$$

5. Korelační koeficient corr (f, g) je dán pro $\Sigma f_i g_i \neq 1$

$$\text{corr}\,(f, g) = \frac{1 - \Sigma f_i^2 - \Sigma g_i^2 + \Sigma f_i g_i}{1 - \Sigma f_i g_i} = 1 - \frac{\Sigma (f_i - g_i)^2}{1 - \Sigma f_i g_i} \tag{39}$$

6. Koeficient explanační síly rozkladu je Wallisovo τ pro $\Sigma f_i^2 \neq 1$

$$\delta = \tau = \frac{\Sigma p_r \Sigma f_{(r)i}^2 - \Sigma f_i^2}{1 - \Sigma f_i^2} \tag{40}$$

7. Koeficient predičkční síly rozkladu je Guttmanovo λ pro $f_{\max} \neq 1$

$$\delta^* = \lambda = \frac{\Sigma p_r f_{(r)\max} - f_{\max}}{1 - f_{\max}}, \tag{41}$$

kde $f_{\max}$ a $f_{(r)\max}$ jsou maximální hodnoty vektorů příslušných rozložení.

Věta 5 shrnuje výsledky modelu pro nominální znak. Důkaz neuvádíme, neboť jej lze provést přímým ověřením. Míra nomvar (34) byla zavedena C. Ginim a použita v práci Light, Margolin [1971] jako základ pro model analýzy rozptylu pro nominální znaky (CATANOVA). Věta 3 je zobecněním základů tohoto modelu. Koeficienty (40), (41) jsou dobře známé z práce Goodman, Kruskal [1954], v níž byly odvozeny v rámci úvah predikčních modelů (optimálního a proporcionálního predikčního pravidla). Míra variability nomvar a míra koncentrace d^* mají v tomto kontextu pravděpodobnostní interpretaci pravděpodobnosti predikčních chyb.

Speciálním případem proměnné je též dichotomický znak ($K = 2$). Pro dichotomický znak platí, že je-li $d_{12} > 0$, pak až na násobek touto konstantou je D matice skóre vytvářejících nominální proměnnou. Věta 5 platí tedy pro libovolný dichotomický znak s tím, že u vzorců (34), (35), (37), (38) musíme pravou stranu násobit hodnotou d_{12} a u vzorce (36) odmocninou této hodnoty ($\sqrt{d_{12}}$). Vzorce (39), (40), (41) zůstanou beze změny. Formule ovšem mohou být dále zjednodušeny vzhledem ke vztahu $f_2 = 1 - f_1$.

9. Diskrétní ordinální znak (uspořádaná klasifikace)

Diskrétní ordinální znak či uspořádanou klasifikaci jsme definovali pomocí vztahu (3). Matice $D = \|d_{ij}\|$ zde má význam $d_{ij} = |i - j|$ = počet kategorií, které leží mezi kategoriemi a_i, a_j. Toto neparametrické skóre odpovídá diskrétnímu uspořádání hodnot jako „žebříčku", jehož příčky je třeba překonat, aby daná statistická jednotka (jedinec) se dostala z hodnoty a_i do a_j. Diskrétnost zde zdůrazňujeme proto, že každá kategorie se počítá i v případě, že je neobsazena ($f_i = 0$). Výsledky pro tento typ znaku jsou uvedeny v práci Řehák [1976] a zde je opět shrneme ve formě věty.

Věta 6. (Diskrétní ordinální proměnná.) Nechť matice skóre $D = \|d_{ij}\|$ je dána $d_{ij} = |i - j|$ pro $i, j = 1, 2, \ldots, K$. Nechť $A = \{a_1, \ldots, a_K\}$ je proměnná s distribucemi $f_{(r)}, f, g \in Q_K$, $r = 1, \ldots, R$, $p' = (p_r)$, $\Sigma f_{(r)} p_r = f$, $F_i = \sum_{j=1}^{i} f_j$, $F_{i/r} = \sum_{j=1}^{i} f_{j/r}$.

Pak

1.

(42) $$\text{Genvar}\, f = \text{Dorvar}\, f = 2\ \Sigma F_i(1 - F_i)$$

(43) $$d_i^* = \Sigma f_j |j - i|$$
$$d^* = \Sigma f_j |j - Me|\,,$$

kde Me je mediánová kategorie určená vztahem $F_{Me-1} < 0{,}5$, $F_{Me} \geqq 0{,}5$, $f_{Me} \neq 0$.

2. Dorvar nabývá svého maxima $\dfrac{K-1}{2}$ pro $f_1 = \frac{1}{2} = f_K$.

3. Matice D^* je pozitivně definitní a vytváří Eukleidovskou metriku na množině distribučních funkcí

$$C_K = \{F : F = (F_1, F_2, \ldots, F_K),\ 0 \leqq F_i \leqq F_j,\ i < j,\ F_K = 1\}$$

(44) $$D(f, g) = \sqrt{2\ \Sigma(F_i - G_i)^2}$$

4. Skalární součin dvou distribucí f, g je určen

(45) $$s(f, g) = \Sigma F_i(1 - F_i) + \Sigma G_i(1 - G_i) - \Sigma(F_i - G_i)^2$$

a má horní hranici

(46) $$\begin{aligned} f'Dg &= \Sigma G_i(1 - F_i) + \Sigma F_i(1 - G_i) \\ &= \Sigma G_i + \Sigma F_i - 2\ \Sigma F_i G_i \\ &= \Sigma F_i(1 - F_i) + \Sigma G_i(1 - G_i) + \Sigma(F_i - G_i)^2 \end{aligned}$$

5. Korelační koeficient pro rozložení f, g je pro $\Sigma f_i g_i \neq 1$

(47) $$\text{corr}\,(f, g) = \frac{\Sigma F_i(1 - F_i) + \Sigma G_i(1 - G_i) - \Sigma(F_i - G_i)^2}{\Sigma F_i(1 - F_i) + \Sigma G_i(1 - G_i) + \Sigma(F_i - G_i)^2}$$

6. Koeficient explanační síly rozkladu je dán pro Dorvar $\neq 0$

(48) $$\delta = \beta = 1 - \frac{\Sigma p_r\ \Sigma F_{i/r}(1 - F_{i/r})}{\Sigma F_i(1 - F_i)}$$

7. Koeficient predikční síly rozkladu je dán pro $d^* \neq 0$

(49) $$\delta^* = \beta^* = 1 - \frac{\Sigma p_r\ \Sigma f_{i/r} |i - Me_{(r)}|}{\Sigma f_i |i - Me|}\,,$$

kde Me mediánová kategorie celého souboru a $Me_{(r)}$ jsou mediánové kategorie distribucí na rozkladových množinách.

Dorvar je tedy Giniho průměrná diference a d^* je Giniho průměrná odchylka od mediánové kategorie. Zde však mají obě míry „neparametrický" význam a jsou nezávislé na případné kvantifikaci kategorií uspořádané klasifikace (jsou invariantní k jakékoli rostoucí transformaci hodnot $x_i = f(a_i)$. Koeficienty β a β^* byly uvedeny

v práci Řehák [1976], kde byla též ukázána dekompoziční vlastnost Dorvar (tj. platnost věty 3 pro tento případ).

Diskrétní ordinální znaky hrají důležitou roli v analýze dat společenských i jiných věd. Sjednocený obecný model pro různé typy proměnných, tak jak jej prezentujeme v této práci, byl především motivován snahou vyvinout smysluplné a jednoduché míry charakterizující tento typ, které by zároveň byly založeny na srovnatelném principu s nominálními a kardinálními proměnnými. Práce nad ordinálními daty pak umožnila vyvinout model daleko obecnější, který zahrnuje mnohem více typů proměnných, nejen tři základní.

10. Diskrétní kardinální znak

Předpokládejme, že zkoumaná proměnná má konečný počet hodnot x_i, $i = 1, 2, \ldots, K$. Zde si ukážeme, jak známé a běžně používané míry odpovídají našemu modelu.

Věta 7. (Diskrétní kardinální proměnná). Nechť matice skóre $D = \|d_{ij}\|$ je dána prvky $d_{ij} = (x_i - x_j)^2$, kde $x_1, x_2, \ldots, x_K$ je K různých reálných hodnot proměnné X. Pak při značení distribucí jako ve větě 5 platí:

1.

$$\text{Genvar } f = 2 \text{ var } X = 2\,\Sigma f_i (x_i - \bar{X})^2 = 2\sigma^2 , \tag{50}$$

kde $\bar{X} = \Sigma f_i x_i$. Středem rozložení X_c je ta hodnota x_i, pro kterou platí, že

$$|X_c - \bar{X}| = \min_i |x_i - \bar{X}| \tag{51}$$

$$d^* = \sigma^2 + (\bar{X} - X_c)^2$$

2. Maximum $\text{Genvar } f = \dfrac{(x_{\max} - x_{\min})^2}{2}$ je dosaženo pro rozložení, které má váhy $\frac{1}{2}$ u maximální a minimální hodnoty proměnné X.

3. $D(f, g)$ je D-polometrika a je dána rozdílem průměrů

$$D(f, g) = \sqrt{2}\,|\bar{X}_f - \bar{X}_g| , \tag{52}$$

kde $\bar{X}_f = \Sigma f_i x_i$, $\bar{X}_g = \Sigma g_i x_i$

Ekvivalentní třídy $\{f^+\}$ jsou tvořeny všemi distribucemi majícími stejný aritmetický průměr.

4. Skalární součin dvou distribucí je

$$s(f, g) = \text{var } (X/f) + \text{var } (X/g) - (\bar{X}_f - \bar{X}_g)^2 \tag{53}$$

$$= \Sigma f_i x_i^2 + \Sigma g_i x_i^2 - 2(\bar{X}_f^2 + \bar{X}_g^2 - \bar{X}_f \bar{X}_g)$$

s horní hranicí

$$f' D g = \Sigma f_i x_i^2 + \Sigma g_i x_i^2 - 2\bar{X}_f \bar{X}_g \tag{54}$$

5. Koeficient korelace corr (f, g) dvojice rozložení f, g je dán pro $\mu_2(f) + \mu_2(g) \neq \neq 2\mu_1(f)\,\mu_1(g)$, kde μ_i je i-tý necentrální moment

(55) $$\operatorname{corr}(f, g) = \frac{\operatorname{var}(X/f) + \operatorname{var}(X/g) - (\bar{X}_f - \bar{X}_g)^2}{\operatorname{var}(X/f) + \operatorname{var}(X/g) + (\bar{X}_f - \bar{X}_g)^2}$$

6. Koeficient explanační síly rozkladu je pro $\operatorname{var}(X/f) \neq 0$

(56) $$\delta = \eta^2 = 1 - \frac{\Sigma p_r \operatorname{var}(X/f_{(r)})}{\operatorname{var}(X/f)}$$

7. Koeficient predikční síly rozkladu je pro $\operatorname{var}(X/f) \neq 0$

(57) $$\delta^* = \eta^{*2} = 1 - \frac{\Sigma p_r \operatorname{var}(X/f_{(r)}) + \Sigma p_r |\bar{X}_{f(r)} - X_{c_r}|^2}{\operatorname{var}(X/f) + |X_f - X_c|^2}$$

Věta 7 je ověřitelná přímými výpočty. Ukazuje, jak známé míry pro číselná data odpovídají obecnému modelu. Je zajímavé si všimnout, že pro diskrétně chápaná číselná data je rozdíl mezi σ^2 a d^* a mezi η^2 a η^{*2}. I zde tedy má význam odlišit explanační a predikční úlohu. Ze vzorců je též vidět, že d^* se může od σ^2 značně lišit. Rozdíl mezi η^2 a η^{*2} bývá obvykle malý a pro praktické účely zanedbatelný. V další části si ukážeme, že model platí pro vícerozměrné zobecnění kardinálního znaku.

11. Areální a M-rozměrné metrické znaky

Důležitý typ proměnné vzniká tehdy, jestliže relace na množině jejich hodnot odpovídá metrickým vlastnostem (M-rozměrná Eukleidovská vzdálenost). Taková situace vzniká například tehdy, když závislá proměnná je souborem M číselných proměnných (M-rozměrný vektor proměnných). Vzniká však také tam, kde vůbec neznáme souřadnice a pouze víme, jak jsou jednotlivé hodnoty od sebe vzdáleny. Tento případ nastává buď při přímém zjišťování vzdáleností z mapy či plánu (odtud název areální proměnná), nebo metodami analýzy — seskupovacími algoritmy a multidimenzionálním škálováním. Multidimenzionální škálování (viz např. Lingoes [1973]) umožňuje metrizovat relace objektů, jejichž podobnosti jsou pouze nemetricky určeny, a proto má multidimenzování škálování novou důležitou roli v souvislosti s naším modelem rozkladu zobecněné variance. Jsou-li totiž skóre d_{ij} modelu D odvozeny z Eukleidovské M-rozměrné metriky, jsou splněny podmínky pro rozkladové vlastnosti Genvar a $D(f, g)$ je D-polometrika, jak zajišťuje věta 8 (viz dále). To znamená, že u proměnných s obecnými relacemi d_{ij} můžeme postupovat tak, že je nejlépe „vyhladíme" např. Guttman-Lingoesovými algoritmy (MINISSA a další) a poté na nich provedeme analýzu pomocí dekompozičních vlastností či komparaci pomocí D-vzdáleností. Znamená to také, že metoda automatické detekce interakcí navržená Morganem a Sonquistem může být tímto postupem použita univerzálně nejen na číselné, uspořádané a nominální proměnné, ale na všechny typy proměnných, u nichž výsledek multidimenzionálního škálování znamená dostatečně přesný obraz vstupní relace. Věta 8 pak také umožňuje důležitou analýzu typologií vzniklých seskupovacími algoritmy. Jsou-li vzniklé typy považovány za hodnoty proměnné A, pak relace na těchto hodnotách je určena vzdáleností (M-rozměrnou) mezi těžišti vzniklých skupin. Model umožňuje analy-

zovat buď predikční, nebo explanační úlohy vztahu k různým rozkladům populace. Zde naznačená strategie (vyzkoušená v praxi analýzy dat) ukazuje důležitost modelu a věty 8.

Definice 5. M-rozměrnou metrickou proměnnou nazveme takovou proměnnou, pro kterou platí: existuje K vektorů o M složkách $X_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$, $i = 1, 2, \ldots, K$ tak, že platí

$$d_{ij} = \sum_{a=1}^{M} (x_{ia} - x_{ja})^2 \tag{58}$$

$$X_i \neq X_j \quad \text{pro} \quad i \neq j$$

Areální proměnnou nazveme dvourozměrný metrický znak.

Věta 8. Buď $A = \{a_1, a_2, \ldots, a_K\}$ M-rozměrný metrický znak f, $g \in Q_K$. Pak $D(f, g) = \sqrt{(f - g)' D(g - f)}$ je D-polometrika na Q_K.

Důkaz.

Existuje K vektorů o M složkách X_i, $i = 1, \ldots, K$ tak, že d_{ij} splňuje podmínku (58). Označme $\bar{X}_{fm} = \Sigma f_j x_{mj}$, $D_m = \|d_{ij}^{(m)}\|$, $d_{ij}^{(m)} = (x_{im} - x_{jm})^2$. Platí, že

$$D(f, g) = \sqrt{(f-g)' D(g-f)} = \sqrt{\sum_{m=1}^{M} (f-g)' D_m (g-f)} = \sqrt{2 \sum_{m=1}^{M} (\bar{X}_{fm} - \bar{X}_{gm})^2},$$

což je metrika pro $\bar{X}_f = (\bar{X}_{f1}, \ldots, \bar{X}_{fM} Y)$ a polometrika pro $f \in Q_K$.

Q. E. D.

Důkaz věty 8 zároveň ukazuje, čemu se rovná $D(f, g)$ a jak je lze vyjádřit, jsou-li vektory X_i známé. Vzdálenost dvou distribucí je v tomto případě vzdálenost dvou těžišť určených body X_i a vahami f_i, resp. g_i na nich.

Praktická aplikace, jejíž strategie už byla popsána výše, může nalézt širší a okamžité uplatnění především v sociální ekologii, v sociologii městských celků či aglomerací, v analýze sociální struktury, v sociologii etnika, v antropologii, v lingvistice (dialektologii), ve sledování koncentrace veřejného mínění, ve hledání příčin sociální distance a jiných vícerozměrných pojmů, v diagnostice, v terapeutických modelech apod. Ve všech těchto případech můžeme sledovat areální či vícerozměrnou variabilitu, rozšířenost, koncentrovanost, váhu jednotlivých bodů vzhledem ke koncentrovanosti, můžeme smysluplně měřit kvalitu predikce i vytvářet smysluplné míry asociace.

Věta 8, ač jednoduchý důsledek modelu, je společně s větou 6 nejdůležitějším výsledkem, který znamená obohacení možností analýzy dat.

12. Závěr

Cílem modelu je sjednotit pohled na různé a různě aplikované míry variability, střední polohy i statistické asociace. Z rozboru by měl také vyplynout rozdíl mezi

koeficientem asociace používaným v explanační a predikční analýze. Důležitost modelu však především leží v přirozené možnosti zobecnění známého postupu vytváření koeficientů na daleko širší třídu závislých znaků, z nichž jsme uvedli pouze areální znak a jeho M-rozměrné rozšíření na znaky vznikající z multidimenzionálního škálování, resp. seskupovací analýzy. Podobně však může být uvedená teorie specifikována například na pojem vzdálenosti uspořádání (viz Kemeny, Snell [1962]), a tak mohou být pomocí zde uvedených postupů analyzovány příčiny uspořádání a variability posuzovatelů, resp. respondentů.

Koeficienty asociace byly formulovány pro rozkladový znak. Pojem vzdálenosti populací je však důležitý hlavně v komparativní metodě, tj. tam, kde chceme komparovat rozložení na složitých či vícerozměrných znacích. K výše uvedeným případům můžeme ještě přidat rozložení na grafech.

Celá práce byla motivována třemi různými úlohami: komparací, explanací a predikcí. Tyto tři typy úloh mají mnoho společného, ale jejich ztotožnění by bylo podle našeho názoru chybou. Dalším podnětem byla potřeba zavést vhodné charakteristiky pro ordinální data, a to na bázi komparovatelné s ostatními typy znaků.

Literatura

Goodman, L. A. — Kruskal, W. H.: *Measures of Association for Cross-Classifications.* Journal of American Statistical Association 1954, No. 49, s. 732—764.

Charvát, F.: *O jednom možném přístupu k analýze sociálních rozdílů (diferenciací).* (Analysis of the Structure of Social Differences — One of the Possible Approaches). Sociologický časopis 1977, č. 4, s. 403—415.

Kemeny, J. G. — Snell, J. L.: *Mathematical Models in the Social Sciences.* New York, Blaisdell Publ. Comp. 1962.

Light, R. I. — Margolin, B. H.: *An Analysis of Variance for Categorical Data.* Journal of American Statistical Association 1971, No. 66, s. 534—544.

Lingoes, J. C.: *The Guttman-Lingoes Nonmetric Program Series.* Ann Arbor, Mathesis Press 1973.

Margolin, B. H. — Light, R. I.: *An Analysis of Variance for Categorical Data. II.: Small Sample Comparisons with Chi Square and other Competitors.* Journal of American Statistical Association 1974, No. 69, s. 755—764.

Řehák, J.: *Základní deskriptivní míry pro rozložení ordinálních dat.* (Basic Descriptive Measures for the Distribution of Ordinal Data). Sociologický časopis 1976, č. 4, s. 416—431.

Wilson, T. P.: *Critique of Ordinal Variables.* Social Forces 1971, No. 49, s. 432—444.

Wishart, D. — Leach, S. V.: *A Multivariate Analysis of Platonic Prose Rhytm.* Computer Studies in the Humanities and Verbal Behaviour 1970, No. 2.

Резюме

Я. Ржегак - Б. Ржегакова: Основные характеристики переменных с конечным числом значений и анализ расстояний их распределений

Статья содержит основную модель, которая даст разные меры описывающие свойства распределений дискретных знаков различных типов. Из модели можно таким способом получить дисперсию, среднюю арифметическую, моду, категорию медияны, среднюю разность и среднее отклонение Джини и так дальше. Модель основана на понятии несходства двух значений переменной, так называемой D-матрицы. Спецификация ее элементов приводит к специальным типам переменных и следовательно к подходящим характеристикам. В статье дана спецификация модели на номинальные, ординальные и кардинальные знаки; тоже введен новый тип: ареальная и m-размерная метрическая переменная (знак), который оказывается в анализе очень полезным.

Кроме самых основных характеристик для описания статистических расспределений, в статье даны дальнейшие важные следствия вытекающие из основной модели: семи-метрика между двумя распределениями, их коэффициент корреляции, меры экспланационной и предикционной силы.

Результаты оказываются полезными в практике анализа данных: в анализе статистических ассоциаций, в сравнительном анализе расспределений, в многоразмерном шкалировании, в группировке (распознавании образов) и в возможности ввести различные и нестандартные типы переменных.

Summary

J. Řehák – B. Řeháková: Basic Characteristics for Finite-Valued Variables and a Distance Analysis of Their Distribution

The paper develops a general approach to data description of discrete – valued variables and to a distance analysis of their distributions. The general model includes various well-known measures as variance, arithmetic mean, Gini's mean difference and mean deviation, mode and median category etc. The model is based on the notion of dissimilarity of two values of the variable, so called D-matrix. The specification of its elements gives the type of the variable and consequently appropriate characteristics. Three basic types are covered: nominal, ordinal and cardinal variables. Also new types, areal and m-dimensional metrical variables are treated as special cases of the model.

Chapter four shows how the semi-metrics can be obtained for general type variable. The decomposition property is derived in chapter five. A notion scalar product of two distributions and the correlation coeficient is introduced in chapter six. The distance and the decomposition property is then used for a natural definition of measures of explanatory power and measures of predictional power that contain as special cases various associational coeficients known in the analytical work with the data. The model gives results that are relevant in connection with multidimensional scaling, clustering and associational analysis as well as for detailed analysis of contingency tables and comparative method applied for general type of data.