Základní deskriptivní míry pro rozložení ordinálních dat

JAN ŘEHÁK
Ústav pro filozofii a sociologii ČSAV, Praha

Základní typy znaků v práci sociologa jsou tři: nominální, ordinální a kardinální. Pro nominální a kardinální znaky máme k dispozici řadu statistických charakteristik, které lze dobře využít pro analytickou práci s daty, ale pro ordinální data máme analytických prostředků daleko méně. Ordinální data, to jest realizace proměnných, jejichž stavy jsou nekvantifikovatelné, ale uspořádané, jsou velmi častá právě v sociologii a ve společenských vědách vůbec. Proto statistická metodologie práce s těmito daty je rozvíjena převážně právě ve společenských vědách nebo pod jejich tlakem.

Úkolem této stati je především uvést jednoduchý model pro práci s ordinálními daty, jenž umožňuje zavést smysluplně interpretovatelné míry pro tato data a zároveň umožňuje hlubší pohled na tyto míry uvedením jejich paralel pro data nominální a kardinální. V článku je popsán obecný postup při zavádění měr pro různé typy dat a je zde ukázáno, že dobře známé míry pro nominální data (opakovací poměr, variační poměr, modus, koeficienty λ, τ) a míry pro číselná data (průměr, rozptyl, korelační poměr) plynou logicky z tohoto modelu. O podobném modelu uvažovali (i když s důležitou rozdílností) Light a Margolin [1971], když zaváděli svou metodu CATANOVA pro testování shody rozložení nominálního znaku v různých populacích. V této stati není řešen problém vzájemné, či jednostranné (symetrické, či asymetrické) statistické závislosti dvou ordinálních znaků; k tomu viz [Řehák - Řeháková 1975]. Přístup použitý zde pro ordinální znaky naznačuje také možnost vyvinout ordinální analýzu rozptylu paralelně s ANOVA (analýza rozptylu pro normálně rozložené číselné veličiny) a s CATANOVA [Light - Margolin 1971], resp. s analýzou založenou na rozkladu entropií [Kullback 1967].

1. Úvod

Tři základní typy dat jsou dobře známy každému výzkumníkovi.

Nominální znak je model vlastnosti, jejíž stavy tvoří třídy ekvivalence; u znaku neuvažujeme žádné relace na množině jeho hodnot. Nominální úroveň nepředpokládá žádné uspořádání, tím méně pak jakoukoli kvantifikaci svých hodnot. Všechny míry, které pro tento znak odvozujeme, musí být invariantní k libovolným permutacím jeho hodnot, nesmějí se měnit při jakémkoli označení či popisu pořadí hodnot.

Kardinální znak je model vlastnosti, jejíž stavy jsou uspořádány, a navíc nám tato vlastnost a naše znalosti umožňují přiřazovat jednotlivým stavům interpretovatelná čísla. Zobrazení převádějící stavy vlastnosti do číselné množiny umožňuje také číselné zpracování měr vycházejících z těchto dat a citlivých na změny číselných hodnot znaku.[1]

Ordinální znak je model, který odráží uspořádání množiny stavů vlastnosti. Hodnoty znaku jsou úplně (lineárně) uspořádány. Proto míry pro tento typ by měly být citlivé k permutacím hodnot znaku, tj. ke změnám jejich uspořádání. Kdybychom měli hodnoty znaku označeny číselně, pak míry založené na ordinalitě jsou na číselné změny necitlivé, pokud je zachováno uspořádání; míry vycházející z ordinálního modelu jsou pro číselná

[1] Požadavek citlivosti na různé typy transformací je dán členěním kardinálních proměnných vzhledem k jednoznačnosti určení číselných hodnot znaku. Různé míry jsou pak různě citlivé na změny: rozptyl se nemění s posunutím počátku číselné stupnice, korelační poměr se nemění s libovolnou lineární transformací, korelační koeficient se nemění s libovolnými kladnými lineárními transformacemi obou znaků a v případě, že obě lineární transformace jsou záporné.

data invariantní vzhledem k libovolné rostoucí transformaci.

Toto základní dělení pochází od Stevense [1959]. Podrobnější typologii uvádí Řehák [1972].

Ordinální znaky jsou předmětem dlouhodobé diskuse a zájmu ve společenských vědách. Je to zřejmě vyvoláno tím, že jejich výskyt je zcela běžný v každém sociologickém výzkumu.

V současné době lze nalézt v literatuře tři přístupy k ordinálním datům:

1. Kritika a argumentace proti používání ordinálního modelu; rozbor nepoužitelnosti, nevhodnosti a neinformativnosti těchto dat ve společenskovědním výzkumu (např. Wilson [1969; 1971]).

2. Zkoumání vlastností pseudokvantifikace, resp. jednoduchého skórování pořadím kategorie a následného použití technik analýzy pro kardinální znaky. Jde tu o zkoumání robustnosti ordinálního modelu směrem k libovolným rostoucím transformacím zavádějícím číselné modely. Důležité výsledky tohoto typu uvádí Labowitz [1967]. Pro tyto úvahy je důležitý také výsledek z práce Goodovy [1972].

3. Snaha vypracovat speciální míry pro ordinální data, jmenovitě pro ordinální korelaci a asociaci (odkazy na literaturu k tomuto tématu lze nalézt v práci Řeháka a Řehákové [1975]), a pro testování hypotéz (literatura k tomuto oboru je velice bohatá).

Všechny tři přístupy jsou velice důležité pro rozvoj metodologie sociologické práce. Tato stať má být příspěvkem ke třetímu přístupu, a to především systematickým zavedením měr variability a analogie ke korelačnímu poměru.

Kritika ordinálních měr neobstojí, pokud bude vedena z pohledu dat číselných nebo dat kvalitativních. Ordinální znaky tvoří velice širokou samostatnou (a též uvnitř diferencovanou) kategorii znaků. Model ordinálního znaku nelze prostě vyškrtnout z výzbroje společenského výzkumu.[2] Rozhodnutí pro typ znaku či pro konečné skórování je vždy závislé na zkušenosti s danou vlastností, na cílech analýzy a na hlediscích minimalizace chyb i výzkumných nákladů. Domnívám se, že ač můžeme často bez velkého nebezpečí používat skórovací metody pro ordinální data [Labowitz 1967], má smysl a je nejen užitečné, ale i nutné rozpracovávat ordinální model pro data jako samostatný, na kvantifikaci nezávislý typ.

2. Obecný model pro zavedení statistických charakteristik

A. Základním pojmem v analýze dat je pojem variability rozložení dat — rozložení prvků populace vzhledem k hodnotám daného znaku. Heterogenita populace π vzhledem k danému znaku A je měřena podle toho, jak jsou si prvky populace nepodobné.

Označíme hodnoty znaku $A = \{a\}$ a zavedeme funkci $d(a, b)$ na množině všech dvojic hodnot znaku. Tato funkce bude charakterizovat vzdálenost,[3] resp. nepodobnost, rozdílnost těchto dvou hodnot.[4] Je jasné, že zavedení funkce $d(a, b)$ závisí na typu znaku, který zkoumáme, že nepodobnost je různá u nominálních, u ordinálních a u kardinálních dat. A právě vhodné určení funkce $d(a, b)$ vede k vhodné charakterizaci typu znaku a k vhodným mírám. Zde uvedeme pouze velice slabá omezení a obecné požadavky na tuto funkci:

1. $d(a, a) = 0$ pro *všechna* a a *všechny* znaky A;
2. $d(a, b) = d(b, a)$ pro *všechny* dvojice (a, b) *všech* znaků A;
3. $d(a, b) \geqq 0$ pro *všechny* dvojice $a, b \in A$;
4. funkce $d(a, b)$ zachovává relaci „nepodobnosti", tj., je-li definována taková relace nepodobnosti na množině všech dvojic, pak jsou-li si (a, b) podobnější než (a', b'), $d(a, b) \leqq d(a', b')$.

[2] Nelze např. přijmout tvrzení, že ordinální znak je méně informativní než kardinální, aniž prostudujeme vlastnost, jíž znak reprezentuje, aniž rozebereme obsah znaku. Kardinální znak je informativnější pro kvantifikovatelné vlastnosti, u nichž je adekvátní kvantifikace známa. Je tedy nutno brát v úvahu strukturu vlastností (vztahy mezi jejich stavy), ale též naše možnosti kvantitativní strukturu odhalit. Jestliže vlastnost sama je pouze ordinální, pak její kvantifikace je pouze aproximací, a tudíž znepřesněním informace; kardinální znak v takovém případě může přinášet falešnou (i když zdánlivě bohatou) informaci.

[3] V tomto článku nemáme na mysli matematický pojem vzdálenosti, ale rozdílnost znaku nebo dvou jedinců vzhledem ke škále znaku (k oboru hodnot znaku).

[4] Tato funkce je zároveň mírou nepodobnosti, rozdílnosti či vzdálenosti dvou prvků z populace π, z nichž u prvního se vyskytuje hodnota a a u druhého hodnota b.

Tyto heuristické požadavky mají jednoduchou a přirozenou interpretaci. Dále si průběžně s rozvíjením postupu ukážeme na příkladech, jak lze tyto pojmy nalézt u zavedených měr pro znaky nominální a kardinální.

Příklad 1

$A = \{a_1, a_2, \ldots, a_K\}$ je nominální znak o K kategoriích.
Zavedeme

(1) $d(a_i, a_i) = 0$ pro všechna i
$d(a_i, a_j) = 1$ pro všechna $i \neq j$

Je-li A například znak *kouření*, mající dvě hodnoty (a_1 = kuřák, a_2 = nekuřák), pak pro každé dva jedince zavedeme $d = 0$ (jsou-li oba kuřáci či oba nekuřáci) a $d = 1$ (je-li jeden kuřák a jeden nekuřák). Toto přirozené skóre nepodobnosti je tedy indikátorovou funkcí různých hodnot na množině dvojic.

Příklad 2

$X = \{x\}$ je kardinální znak.
Položíme

(2) $d(x, y) = (x - y)^2$ pro všechna x, y

Jestliže je např. X = {cena ropy} a π je soubor producentů ropy, pak $(x - y)^2$ je mírou podobnosti dvojice. Takto zavedená funkce je velice výhodná (i když možností je více) a vede k metodě nejmenších čtverců a k rozkladům analýzy rozptylu.

B. Přirozenou *mírou rozptýlení* pro populaci π je očekávané skóre d pro všechny dvojice[5] z populace.[6] Je-li $\{f_i\}$ rozložení pravděpodobnosti na hodnotách znaku $A = \{a_i\}$, pak výběr dvojice hodnot (a_i, a_j) z populace π, který odpovídá výběru s opakováním, resp. cestě nezávislého opakování realizace znaku A, má pravděpodobnost

$$P\,[\text{vybereme } (a_i, a_j)] = f_i \,.\, f_j \text{ pro lib. } i, j$$

Očekávaná hodnota $d(a, b)$ je tedy dána

(3) $$d = \sum_i \sum_j f_i f_j d(a_i, a_j)$$

Tato míra má velice jednoduchou interpretaci: je to průměrná nepodobnost náhodně vybrané dvojice prvků z populace.

Základní vlastnosti:

a) Je-li $d = 0$, pak jsou všechna $f_i f_j d(a_i, a_j) = 0$, což dále značí, že pro nenulové skóre $d(a_i, a_j)$ musí být váhy f_i, f_j nulové, a tudíž všechny prvky populace musí být soustředěny tak, že mají nulová $d(a, b)$. Ve většině případů bude $d(a, b) = 0$ ekvivalentní s $a = b$; v těchto případech pujde o stoprocentní statistickou homogennost — *všechny prvky populace musí mít stejné hodnoty znaku.*

b) d je tím větší, čím větší jsou f_i u vzdálenějších prvků populace.

Příklad 1 (pokr.)

Zavedeme-li pro Khodnotový znak A rozložení četností $\{f_1, f_2, \ldots, f_K\}$, pak aplikace definice (1) ve vzorci (3) dává

(4) $$d = 1 - \Sigma f_i^2$$

Dostali jsme známý *opakovací poměr* jako míru variability pro nominální znaky. Tato míra byla zavedena C. Ginim už v roce 1938 (odkaz: [Light - Margolin 1971]).

d je relativní počet dvojic populace, které nepatří do stejné kategorie; pro $d = 0$ musí být $\Sigma f_i^2 = 1$ a to není možné jinak, než že pro jedno i je $f_i = 1$, což znamená, že populace je soustředěna v jedné kategorii znaku.

Příklad 2 (pokr.)

Pro kardinální znak, u něhož se vyskytují v populaci hodnoty $X = \{x_i\}$, zavedeme (3) pomocí (2):

(5) $$\begin{aligned} d &= \sum_i \sum_j f_i f_j (x_i - x_j)^2 \\ &= 2 \sum_i f_i (x_i - \bar{X})^2 \\ &= 2\, var\, X \\ &= 2\sigma^2 \end{aligned}$$

$\bar{X}$ zde značí aritmetický průměr proměnné X a σ^2 je běžné označení pro rozptyl (např. [Rodriguez 1945]).

Nulová hodnota d vzniká právě tehdy, když všechny nenulové rozdíly $(x_i - x_j)$ mají nulové váhy f_i, f_j, tedy právě tehdy, když pro jednu z hodnot x_i máme $f_i = 1$ a žádná jiná hodnota se v populaci už nevyskytuje.

C. Pro každou hodnotu a znaku $A = \{a_i\}$ možno zavést průměrnou vzdálenost všech prvků populace od této hodnoty:

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

[5] Model může předpokládat buď všechny různé dvojice prvků konečné populace π, nebo všechny dvojice prvků (včetně opakování). Dále se musíme rozhodnout, zda budeme brát všechny dvojice i vzhledem k uspořádání, nebo pouze dvojice bez ohledu na uspořádání. Rozdíly v postupu u různých přístupů nejsou nijak podstatné a je pouze nutné se pro jeden z nich rozhodnout a konzistentně se ho držet. V dalším textu předpokládáme přístup výběru s opakováním, takže dvojice lišící se uspořádáním jsou považovány za různé. Úvahy v této stati se týkají konečných znaků. Rozšíření na spojité znaky a nekonečné znaky diskrétní je přímočaré. Důvodem pro naše dobrovolné omezení je jednak to, že takový případ je ve výzkumu nejčastější, a jednak snaha vyhnout se značení pomocí integrálů, se kterými není většina sociologů dobře obeznámena.

[6] Jinak vyjádřeno: *průměrná nepodobnost, rozdílnost, pro všechny možné dvojice, které model připouští.* Též můžeme říci, že se jedná o *míru rozptýlení daného statistického rozložení.*

Tento ukazatel je *mírou centrality* a pro populaci π. Čím menší je d_a^*, tím menší je průměrná vzdálenost všech prvků od této hodnoty. To a, pro něž je d_a^* nejmenší, proto nazveme *středem* (*centrem*) rozložení znaku A pro populaci π.

Je vidět, že d můžeme vyjádřit

$$d = \sum f_i d_{a_i}^* \tag{7}$$

Příklad 1 (pokr.)

Pro nominální znak je

$$d_a^* = 1 - f_a \tag{8}$$

a minima je dosaženo pro $\bar{a}$, pro něž je $f_{\bar{a}}$ maximální, to jest pro *modus*. Dostáváme tak modus jako střed rozložení pro nominální znaky.

Příklad 2 (pokr.)

Pro kardinální znak je

$$d_x^* = \sum f_i (x_i - x)^2$$

a tato hodnota je minimalizována pro *aritmetický průměr*

$$\bar{X} = \sum f_i x_i$$

který je tak středem rozložení. Toto běžné zavedení však je přesné pouze pro případ, že X chápeme jako spojitý znak, u něhož je $\bar{X}$ přípustnou hodnotou. Jde-li o diskrétní hodnoty x_i (například celočíselné proměnné), pak za střed rozložení je nutno vzít *celé číslo nejbližší k aritmetickému průměru* (zaokrouhlené $\bar{X}$, jsou-li x_i celočíselná).[7] Pro toto číslo X^* platí, že jeho míra centrality $d_{x^*}^*$ je nejmenší mezi všemi hodnotami $d_{x_i}^*$.

D. Minimální hodnotu d^*, připadající středu rozložení, můžeme vzít též za míru rozptýlení; tato míra měří *koncentrovanost kolem středu* $\bar{a}$.

$$\begin{aligned} d^* &= min\, d_a^* \\ &= d_{\bar{a}}^* \\ &= \sum f_i d(a_i, \bar{a}) \end{aligned} \tag{9}$$

Míra vyjadřuje průměrnou nepodobnost, vzdálenost všech prvků od středu.

Její vlastnosti:

a) je nulová právě tehdy, když všechny znaky mají nulovou vzdálenost od $\bar{a}$;

b) je tím větší, čím větší četnosti mají hodnoty vzdálenější od středu rozložení.

Příklad 1 (pokr.)

Pro nominální znak je

$$d^* = 1 - f_{Mod} \tag{10}$$

kde f_{Mod} je četnost modální kategorie v rozložení $\{f_i\}$; d^* je známý *variační poměr*.

Příklad 2 (pokr.)

Pro kardinální spojitý znak X je

$$d^* = \sigma^2 = \sum f_i (x_i - \bar{X})^2 \tag{11}$$

kde $\bar{X}$ je aritmetický průměr.

Pro diskrétní $X = \{x_i\}$, u něhož nepřipouštíme žádné hodnoty mimo $\{x_i\}$, je

$$\begin{aligned} d^* &= \sum f_i (x_i - X^*)^2 \\ &= \sum f_i (x_i - \bar{X})^2 + (\bar{X} - X^*)^2 \\ &= \sigma^2 + (\bar{X} - X^*)^2 \end{aligned} \tag{12}$$

kde X^* je celé číslo nejbližší k $\bar{X}$.

E. Dále se budeme zabývat rozkladem populace na oblasti a explanační silou rozkladu. Předcházející míry měly především deskriptivní roli. Variabilita však je základní pojem pro statistické zkoumání empiricky se projevivších kauzálních vztahů a pro explanační cíle analýzy. Běžný postup spočívá v tom, že se spočítá zvlášť variabilita v každé oblasti, poté se spočítá průměrná variabilita uvnitř těchto oblastí a nakonec se porovná s celkovou variabilitou v populaci. Tento postup je základem pro zavedení indexu asociace rozkladové proměnné a daného znaku A. Tento index můžeme nazvat obecně *indexem explanativní síly rozkladu*. Běžně se konstruuje jako relativní redukce variability:

$$\delta = \frac{\begin{pmatrix}\text{variabilita} \\ \text{v celé populaci}\end{pmatrix} - \begin{pmatrix}\text{průměrná variabi-} \\ \text{lita v oblastech}\end{pmatrix}}{\text{variabilita v celé populaci}} \tag{13}$$

Obecné vlastnosti indexu:[8]

a) Index je definován jen pro případy, kdy existuje nenulová variabilita v celé populaci.

b) Jestliže variabilita v oblastech vymizí, pak index je roven jedné.

c) Jestliže variabilita je stejná ve všech oblastech a je stejná jako variabilita v celé

[7] Tím ovšem nemá být řečeno, že aritmetický průměr je špatná míra pro diskrétní proměnné. Také ho v této souvislosti běžně používáme, a i když jeho použití nemusí být vždy zcela „čisté", máme pro ně řadu dobrých důvodů. Rozdíl mezi aritmetickým průměrem a zaokrouhleným aritmetickým průměrem ovšem vyvstane jasně při predikční interpretaci tohoto modelu.

[8] Vhodnost indexu (13) zde nebudeme hodnotit obecně; jeho vlastnosti nutno posoudit zvlášť pro každé $d(a, b)$. Záleží na tom, zda průměrná variabilita je nutně menší než celková variabilita, na tom, jaká je intepretace $d(a, b)$, použité míry variability apod.

populaci, pak je index roven nule. Index je v praxi aplikován jak pro d, tak pro d^*.

Zaveďme pro další úvahy jednotná označení:

$A = \{a_i\}$ je znak s hodnotami a_i.
f_i je pravděpodobnost výskytu hodnoty a_i v populaci π.
$R = \{R_i\}$ je rozkladový znak, který dělí populaci na oblasti π_i.

Další označení:

$f_{i/r}$ je podmíněná pravděpodobnost výskytu hodnoty a_i v $r$té oblasti π_r rozkladu R.
g_r je rozložení na znaku R, tj. relativní četnost příslušná k hodnotě R_r rozkladového znaku.

Index (13) pak lze přepsat jako

$$\delta' = \frac{d(A) - \Sigma g_r d(A/R_i = r)}{d(A)} \tag{13a}$$

nebo

$$\delta'' = \frac{d^*(A) - \Sigma g_r d^*(A/R_i = r)}{d^*(A)} \tag{13b}$$

kde $d(A)$ odkazuje k π a $d(A/R_i = r)$ odkazuje k π_r.

Příklad 1 (pokr.)

Je-li A nominální a R rozklad, pak aplikace míry (4) na (13), tj. určení (13a), vede k *Wallisovu koeficientu* $\tau_{A/R}$; aplikace míry (8), tj. použití vzorce (13b), vede k $\lambda_{A/R}$ (k tomu viz podrobně [Řehák - Řeháková 1973], kde jsou též reference na původní práce).

Příklad 2 (pokr.)

Pro kardinální znak X, který považujeme za spojitý, vede aplikace (13a) i (13b) ke *korelačnímu poměru* $\eta^2_{X,R}$; při diskrétním chápání číselného znaku bychom museli provést korekce všech používaných rozptylů v duchu vzorce (12).[9]

F. V úvahu přichází také tento *predikční model*:

1. Určujeme hodnotu znaku A náhodně vybraného prvku z populace π podle pravidla založeného na $\{f_i\}_{i=1}^{K}$, tj. na pravděpodobnostním rozložení na hodnotách znaku A.

2. Je-li dáno R a u náhodně vybraného prvku známe jeho třídu R_r (oblast π_r), pak určujeme hodnotu znaku pro tento prvek pomocí rozložení $\{f_{i/r}\}_{i=1}^{K}$.

3. Kvalitu predikce (přiřazené hodnoty) vyjadřuje pro vybraný prvek funkce $d(a, b)$, kde a je hodnota přiřazená podle predikčního pravidla a b je skutečná hodnota.

4. Očekávaná chyba predikce (průměrná chyba) je mírou variability populace. Čím neurčitější je predikce, tím rozptýlenější je populace.

5. Relativní úbytek chyby predikce při znalosti a bez znalosti hodnoty R_r je mírou informačního přínosu znaku R pro znak A a mírou predikční síly znaku R pro znak A vzhledem k danému statistickému rozložení. Zároveň může být tato míra chápána jako míra asymetrické statistické závislosti A na R. Takto vzniklé míry se nazývají PRE-míry asociace (proportional reduction in error).

Predikční pravidla můžeme uvést dvě:

I. *proporcionální predikce*: hodnotu přiřadíme náhodně, s pravděpodobnostmi stejnými, jaké jsou pravděpodobnosti rozložení znaku;

II. *optimální predikce*: přiřazenou hodnotou je optimální prediktor, tj. hodnota, která minimalizuje chybu.

Predikční chyba $d(a, b)$, kterou můžeme chápat také jako ztrátovou funkci predikčního rozhodnutí, je určena podle typu znaku a cíle predikčního rozhodování. Pro statistické rozhodování můžeme zavést funkci $d(a, b)$ obdobně jako ve výše popsaném modelu.

Hodnota d ve vzorci (3) ukazuje kvalitu pravidla I, d^* ve vzorci (9) ukazuje kvalitu pravidla II. Střed rozložení je optimálním prediktorem, index (13a) je redukcí predikční chyby pro d a (13b) pro d^* při znalosti R.

Příklad 1 a 2 (pokr.)

Výše zavedená skóre nepodobnosti $d(a, b)$ pro nominální a kardinální znaky lze použít stejně dobře jako ukazatele kvality predikce. Všechny odvozené míry pak mají predikční interpretace a koeficienty statistické závislosti mají PRE-interpretace.

3. Ordinální znaky

Ordinální znak vzniká jako přiřazení hodnot z úplně uspořádané množiny A ke stavům vlastnosti A, tj. zobrazení A na A. Ve vý-

[9] Zde rozebíráme modely, a proto je třeba rozlišovat model spojitý a diskrétní. V praxi výzkumu však zřejmě můžeme takovou míru korekce u diskrétních znaků zanedbat, neboť numerická hodnota se mění nepatrně.

zkumné praxi ordinální znaky většinou analyzujeme tak, že používáme techniky pro znaky nominální (chí-kvadrát-test), nebo tak, že používáme skórování kategorií a techniky pro číselné proměnné. Výjimkou z tohoto postupu jsou koeficienty ordinální asociace.

Dříve než navrhneme vlastní popisné míry pro ordinální data, všimneme si některých základních vlastností a tím i vnitřní diferenciace ordinálních znaků.

Ordinální vlastnosti (vlastnosti s uspořádanými stavy) mohou být chápány jako *diskrétní* (*rozpojité*), nebo *kontinuální* (*spojité*).

Diskrétní ordinální vlastnosti jsou např.:

— hodnost v armádě
— funkční zařazení v organizaci
— stupeň ukončeného školního vzdělání
— akademická a vědecká hodnost
— postavení v kastovním systému
— Výsledky Guttmanova škálovacího postupu

Spojité vlastnosti jsou např.:

— spokojenost s prací
— schopnost řídit organizaci
— stupeň přizpůsobivosti
— inteligence
— sportovní výkonnost

Rozpojitost, či spojitost chápání ordinální vlastnosti ovšem ještě neznamená, že také její znak musí být rozpojitý, či spojitý. Mnohdy je dost obtížné určit (či zavést předpoklad), zda znak (proměnná) je spojitý, či diskrétní (např. stadia vývoje, generační rozvrstvení, velikostní charakteristiky, statistická asociace dvou proměnných pro konečnou populaci). V takovém případě se rozhodujeme pro model na základě výzkumné zkušenosti, intuice a dostupnosti statistických technik.

Data o spojitých vlastnostech tvoří seskupení, neboť jednotlivce většinou není možno našimi měřicími prostředky rozlišit a není to ani nutné, protože už metodologií dotazu spojitou vlastnost diskretizujeme. Tak dospíváme k situaci, kdy získaná uspořádaná kategorizace je podložena kontinuem.

Na druhé straně i jasně diskrétní vlastnost (např. hodnost v armádě či funkce v organizaci) může být indikací jiné, spojité vlastnosti (např. odborných schopností nebo doby služby). V takovém případě opět můžeme hodnoty diskrétní vlastnosti považovat za skupiny vzniklé na určitém kontinuu.

V modelovém přístupu je nutno nakonec brát v úvahu i způsob, jak seskupování stavů spojité vlastnosti vzniká. Buď může být způsobeno skutečnou nerozlišitelností, nebo k němu může docházet na základě jistých kritérií; tak může vzniknout ordinální znak např. i z číselných měření (věkové hodnoty můžeme kategorizovat na děti, mládež, mladé plnoleté, střední věk, skupinu před důchodem, důchodce) nebo z grafických zjištění (metoda sémantického diferenciálu). Na druhé straně může být kategorizace předem dána (uzavřené otázky týkající se spokojenosti, hodnocení, školní klasifikace).

O dále uvedených dvou modelech, tj. diskrétním a spojitém znaku, proto vůbec není jednoduché rozhodnout, který z nich je pro danou situaci vhodnější. Pro autora této stati je hlavním kritériem to, zda ke kategorizaci dochází s apriorním záměrem, či na základě aposteriorního (po měření) rozdělení kontinua.

4. Charakteristiky dat diskrétního ordinálního znaku

Postup zde aplikovaný sleduje obecný přístup k zavedení měr, vypracovaný v části 2 této stati. To jednak umožní posouzení vhodnosti modelu srovnáním s jinými typy znaků a jednak ukáže na paralelnost jednotlivých měr.

Ještě před uvedením modelu samého si zavedeme další označení pro *distribuční funkce*:

$$F_i = \sum_{j=1}^{i} f_j \quad \text{pro} \quad i = 1, 2, \ldots, K$$

$$F_{i/r} = \sum_{j=1}^{i} f_{j/r} \quad \text{pro} \quad i = 1, 2, \ldots, K$$

a pro všechna r

Ordinální znak $B = \{b_i\}$ bude mít hodnoty $\{b_1, b_2, \ldots, b_K\}$, uspořádané podle rostoucích hodnot indexu.[10]

A. Určení charakteristiky nepodobnosti, rozdílnosti či vzdálenosti dvou jedinců popu-

[10] V zájmu jednoduchosti předpokládáme, že znak B je konečný. Veškeré úvahy ovšem lze snadno rozšířit na spočetné znaky, a tak všechny výsledky lze např. aplikovat na diskrétní celočíselná rozdělení, jako je Poissonovo, geometrické, negativní binomické, logaritmické apod.

lace vzhledem ke znaku B musí vycházet z heuristických požadavků a musí charakterizovat diskrétní ordinalitu znaku. Budeme zde vycházet z funkce

(14) $d(b_i, b_j) = |i - j|$ pro všechna i, j

To znamená, že rozdíl mezi dvěma prvky je dán počtem kroků, které musí učinit jeden, aby se dostal na pozici (úroveň) druhého. Např., jsou-li b_3 a b_4 třetí a čtvrtý stupeň armádních hodností, pak rozdíl mezi dvěma osobami s hodnostmi b_3 a b_4 je jeden stupeň; osoba s b_3 musí být jednou povýšena, aby se dostala na stupeň b_4. Je-li B znak spokojenosti, pak dva respondenti s hodnotami b_2 a b_3 se liší o jeden stupeň spokojenosti.

Tato neparametrická funkce je nezávislá na přiřazení skóre ke kategoriím b_i; je tedy (a též vše, co je z ní odvozeno) invariantní k libovolné rostoucí transformaci hodnot $\{b_i\}$, jsou-li tyto hodnoty vyjádřeny číselně.

Model se nehodí pro případ spojitých dat seskupených pro svou nerozlišitelnost. Hodí se tedy pouze tam, kde je předem dána kategorizace, ať už diskrétní podstatou dané vlastnosti, či vzniklá metodologickou nebo teoretickou apriorní diskretizací spojité vlastnosti, např. pomocí uzavřené otázky či seskupovacího pravidla. Diskrétní ordinální model je tedy vhodný jako *uspořádaná klasifikace*, a to buď s podloženou latentní spojitou proměnnou, nebo bez ní.[11]

B. Míra variability pro diskrétní ordinální znak může být proto zavedena jako veličina

(15) $$d = dorvar\ B = \sum_i \sum_j f_i f_j |i - j|$$

(16) $$= 2 \sum_i F_i (1 - F_j)$$

Tato *diskrétní ordinální variance* značí průměrný počet kategorií, o něž se liší dvojice prvků populace π.[12]

Její vlastnosti jsou obdobné jako u obecné míry:

a) $d = 0$, právě když všichni jedinci populace patří do jedné (libovolné) kategorie;

b) d je tím větší, čím větší hodnoty f_i jsou umístěny u okrajových kategorií znaku B, to jest, vyšší váhy máme u velkých rozdílů $|i - j|$.

C. Mírou centrality hodnoty b_i je $d_i^* = \sum_j f_j |j - i|$. Jak je možno ukázat podle [Gini 1970], tato míra nabývá pro b_i minima, které nazveme *mediánovou kategorií*,[13] označíme b_M a určíme takto:

(17) b_M je hodnota znaku, pro niž $F_{M-1} < 0{,}5$, $F_M \geqq 0{,}5$ a $f_M > 0$

Mediánová kategorie je tedy centrem rozložení diskrétního ordinálního znaku.

D. Jako další míru variability můžeme vzít kocentraci dat kolem mediánové kategorie, tedy

(18) $$d^* = Dorvar\ B = \Sigma f_i |i - M|$$

kde M je mediánová kategorie.

Tato míra má stejné vlastnosti jako d dané vzorcem (16). Míra $d^* = Dorvar$ je průměrný počet kategorií, o které se liší prvky populace od centra rozložení.

K oběma mírám variability je nutno poznamenat, že to jsou míry spjaté s danou kategorizací. Jestliže zjemníme kategorizaci, tj. určíme více tříd, pak se míra změní. Pokud je znak odrazem rozkladu kontinua, pak velice záleží na tom, na kolik kategorií je rozložíme. Provedeme-li takový rozklad a po zjištění dat pro danou populaci se ukáže, že jedna z tříd je prázdná, pak ji *nemůžeme* prostě vynechat anebo sloučit se sousední kategorií, *aniž* provedeme *obsahové přehodnocení znaku* a tím míry. V této vlastnosti

[11] Nutno poznamenat, že většina pořadových metod neparametrické statistiky, např. mediánový test, Spearmanův a Kendallův koeficient pořadové korelace nebo Wilcoxonův test, odpovídají nikoli diskrétnímu, ale spojitému chápání proměnné, která je po měření buď rozdělena do tříd, nebo jsou u ní konstatována spojená pořadí. Rozložení takových statistických testů, koeficientů apod. ve většině diskrétních případů není známo a stejné testové statistiky pro uvedené dva modely se mohou navzájem lišit rozložením a kritickými hodnotami.

[12] Přijetí, či zamítnutí modelu padá s přijetím, či zamítnutím funkce $d(a, b)$. Vše ostatní je jednoznačné, tak jako u jiných typů znaků. Domnívám se, že zavedení $d(a, b)$ tímto způsobem je logické a že plně odpovídá chápání modelu. Tak jako u nominálních a kardinálních znaků, i zde je funkce $d(a, b)$ zavedena nejjednodušším možným způsobem. — Zkratka *dorvar* značí *diskrétní ordinální varianci*.

[13] Mediánová kategorie je taková, do níž by vstoupil prostřední člen populace uspořádané podle hodnot znaku B.

a požadavku je vidět základní interpretační rozdíl mezi spojitými a diskrétními daty.

E. F. Aplikací vztahů (13a) a (13b) a predikční interpretace dostaneme dvě asymetrické asociace mezi nominálním znakem (rozkladovým znakem R) a daným diskrétním ordinálním znakem B.

Koeficient proporcionální predikce:

$$(19)\quad \beta_{B/R} = 1 - \frac{\sum_{r=1}^{R} g_r \sum_{i=1}^{K} F_{i/r}(1 - F_{i/r})}{\sum_{i=1}^{K} F_i(1 - F_i)}$$

Koeficient optimální predikce:

$$(20)\quad \beta^*_{B/R} = 1 - \frac{\sum_{r=1}^{R} g_r \sum_{i=1}^{K} f_{i/r}|i - M_r|}{\sum f_i|i - M|}$$

kde M je pořadí mediánové kategorie b_M v populaci π a M_r je pořadí mediánové kategorie b_M v $r$té oblasti rozkladu R.

Diskrétní ordinální variance *dorvar* [viz (15) a (16)] má podobné rozkladové vlastnosti jako ANOVA pro kardinální znaky a CATANOVA, která je těsně spjata s Wallisovým $\tau_{A/R}$ [Light - Margolin 1971; Margolin - Light 1974].

Můžeme psát

$$(21)\quad \sum_i F_i(1 - F_i) = \sum_r g_r \sum_i F_{i/r}(1 - F_{i/r}) + \sum_i \sum_r g_r(F_{i/r} - F_i)^2$$

Tento výraz je analogický známému $TSS = WSS + BSS$ z analýzy rozptylu (celkový součet čtverců je roven součtu čtverců uvnitř oblastí plus součtu čtverců mezi oblastmi). Druhý člen je vážený součet čtverců eukleidovských vzdáleností distribučních funkcí jednotlivých rozložení v oblastech a distribuční funkce rozložení v celé populaci. Je možné jej vyjádřit též takto:

$$(22)\quad \sum_i^K \sum_r^R g_r(F_{i\,r} - F_i)^2 = \frac{1}{2} \sum_r^R \sum_s^R g_r g_s \sum_i^K (F_{i/r} - F_{i/s})^2$$

Obě strany vyjadřují základní interpretace variability mezi oblastmi, tedy analog *BSS*; pravá strana ukazuje, že tento člen je váženým součtem eukleidovských vzdáleností všech dvojic podmíněných distribucí v oblastech rozkladu R. Lze tedy očekávat, že bude možno vypracovat neparametrickou metodu DORANOVA, paralelní s analýzami rozptylu pro kardinální a nominální znaky.

Tyto vlastnosti ukazují, že k obecným vlastnostem koeficientu (13) můžeme pro β připojit ještě další důležité vlastnosti:

d) $0 \leqq \beta \leqq 1$.

e) $\beta = 0$ má za následek, že všechna podmíněná rozložení jsou stejná, tj. že B je na R statisticky nezávislé.

f) $\beta = 1$ má za následek, že veškerá variabilita ve všech $\{\pi_r\}$ vymizí. To znamená, že predikce B za pomoci R je jednoznačná; statistický vztah je nahrazen úplnou jednoznačnou funkční závislostí $R \to B$.

5. Spojitý ordinální znak

Spojitý ordinální model předpokládá, že vlastnost je kontinuem a že seskupování je samovolné; je způsobeno buď nedostatečně citlivým měřicím aparátem, nebo aposteriorními úvahami výzkumníka. Typickým příkladem je rozdělování sportovců do skupin podle výkonu (výkon je spojitá proměnná); skupiny se vytvářejí tak, že se seskupí sportovci přibližně stejného výkonu a sportovních schopností. Jiným častým postupem je seskupování nezávislých objektů (výběrových souborů), u kterých se při testování statistické hypotézy o rovnosti, či různosti některé míry střední hodnoty neukázaly signifikantně rozdílné výsledky. Nakonec uveďme seskupování pomocí seskupovací analýzy[14] nebo pomocí obsahové analýzy (obojí je postup aposteriorní — po získání dat).

Na tento typ znaku lze aplikovat všechny pořadové metody neparametrické statistiky s jejich modifikacemi pro realizovaná spojení. Jejich modely odpovídají spojité veličině

[14] *Seskupovací analýza* je název vhodný pro metody, které se v anglické odborné literatuře nazývají *cluster analysis, clustering method* nebo *mathematical taxonomy*. Autor jej navrhuje po diskusi s pracovníkem Ústavu pro jazyk český ČSAV.

a náhodnému vytváření skupin nerozlišitelných či a posteriori sloučených pozorování.

Dále, stejně jako v předcházející části, budeme postupně sledovat obecný model z části 2. Pro všechny závěry je ovšem nejdůležitější opět oddíl A, ve kterém se zavádí míra nepodobnosti dvou prvků populace.

A. Vzdálenost či nepodobnost dvou jedinců populace *nelze zde založit na hodnotách znaku samých*, neboť jednak je známa pouze informace o uspořádání a seskupování prvků a jednak jakékoliv míry musí být zcela invariantní k libovolnému roztažení či libovolné rostoucí transformaci škály. Proto zavedeme *vzdálenost dvou jedinců jako očekávaný počet jedinců, kteří jsou mezi nimi*. Je-li nyní A spojitý ordinální znak, jenž vytvořil seskupení, označená a_i a uspořádaná podle rostoucí hodnoty indexu i ($A = \{a_i\}$), pak je-li f_i relativní četnost hodnoty znaku a_i (resp. pravděpodobnost, že náhodně vybraný prvek je ze skupiny a_i), definujeme:

$$(23)\quad d(a_i, a_j) = \tfrac{1}{2} f_i + \sum_{k=i+1}^{j-1} f_k + \tfrac{1}{2} f_j \quad \text{pro } j > i$$
$$d(a_i, a_j) = d(a_j, a_i) \text{ pro } i > j$$
$$d(a_i, a_i) = 0 \text{ pro všechna } i$$

První výraz můžeme přepsat též jako

$$(24)\quad d(a_i, a_j) = \tfrac{1}{2} f_i + (F_{j-1} - F_i) + \tfrac{1}{2} f_j = \tfrac{1}{2}(F_{j-1} + F_j) - \tfrac{1}{2}(F_{i-1} + F_i)$$

B. Pro zavedení míry variability (3) použijeme vztahu (7). To znamená, že nejprve určíme d_m^* ze (6):

$$(25)\quad d_m^* = (F_m - \tfrac{1}{2})(F_{m-1} - \tfrac{1}{2}) + \tfrac{1}{4}$$

Použitím (7) dostaneme „míru variability" spojitého ordinálního znaku:

$$(26)\quad d = corvar = \sum_{m=1}^{K} f_m (F_m - \tfrac{1}{2})(F_{m-1} - \tfrac{1}{2}) + \tfrac{1}{4}$$

C. Lze ukázat, že výraz (25) nabývá minima pro index M, který je číslem *mediánové skupiny* (*kategorie*) znaku A.

D. Proto druhá „míra variability" spojitého ordinálního znaku je

$$(27)\quad d_m^* = Corvar = (F_M - \tfrac{1}{2})(F_{M-1} - \tfrac{1}{2}) + \tfrac{1}{4}$$

O variabilitě či rozptýlenosti dat vzhledem k hodnotám znaku však zde musíme mluvit pouze velice opatrně, vzhledem k libovolnosti škál vlastností. Proto zde můžeme mluvit spíše o míře kvality predikce mediánovou kategorií (27), popř. o míře seskupení dat (26) a proporcionální prediktabilitě. Hodnota d totiž značí průměrné procento osob z populace (až na faktor sto), které se vyskytnou na škále mezi dvěma náhodně vybranými osobami. Čím je toto číslo menší, tím seskupenější jsou data ve skupinách (tím méně je tu kategorií). Proto je d ukazatelem neurčitosti rozložení, resp. jeho variability, pouze v uvedeném smyslu. Je ovšem obtížné najít jiný obsah variability pro data tohoto typu.[15]

Vlastnosti míry *corvar* (26):

a) je rovna nule, jestliže všechna data spadají do jedné skupiny;

b) je tím větší, čím „rozptýlenější" jsou data, tj. čím méně seskupená data dostáváme;

c) na rozdíl od *dorvar* — variance pro diskrétní ordinální data — je tato míra necitlivá na vložení, či vynechání libovolného počtu prázdných kategorií mezi dvěma skupinami.

Z vlastnosti c) plyne, že kdybychom tuto míru aplikovali na uspořádanou kategorizaci, pak bychom prázdné kategorie mohli vynechat (vzorec je vynechává automaticky).[16]

Druhá míra — *Corvar* (27) — má tyto vlastnosti:

a) je rovna nule, jestliže se všechna data soustředí do jedné skupiny (to značí perfektní prediktabilitu);

b) je tím větší, čím menší je obsazení mediánové kategorie;

c) je necitlivá na obsazení nemediánových kategorií.

Vlastnost a) má obdobu ve vlastnosti variačního poměru ($1 - f_{Mod}$), použitelného pro nominální znaky. *Corvar* i variační po-

[15] Maximální varianci *corvar* bychom dostali pro populaci absolutně seřazenou do pořadí. Proto výraz d/d_{max} je vlastně indexem seskupenosti dat.

[16] Tuto vlastnost má např. Spearmanův koeficient pro kontingenční tabulky a též všechny míry založené na rozdílu mezi počtem shod a počtem neshod (obvykle označeném $P - Q$), jako je Goodman-Kruskalovo gama, Kendallovo τ, Sommersovo d. Všechny tyto míry také mají v podtextu spojitý model, na nějž jsou aplikovatelné (i když se při odvozování na spojitý případ neodvolávají).

Tabulka 1: Rozložení četnosti a distribuční kumulativní funkce rozložení návštěvnosti koncertů „pop-music", džezové a taneční hudby (v %, distribuční funkce v závorkách)

Oblast	Název a číslo kategorie odpovědí							Součet	Zastoupení oblasti v populaci
	Nenavštěvuje (1)	1× za sezónu (2)	2–3× za sezónu (3)	4–6× za sezónu (4)	1× za měsíc (5)	1× za 14 dní (6)	1× za týden (7)		
Osoby mladší než 30 let (π_1)	33 (33)	21 (54)	24 (78)	10 (88)	6 (94)	6 (100)	— (100)	100	50
Osoby ve věku 30 a více let (π_2)	81 (81)	7 (88)	12 (100)	— (100)	— (100)	— (100)	— (100)	100	50
Celá populace (π)	57 (57)	14 (71)	18 (89)	5 (94)	3 (97)	3 (100)	— (100)	100	100

měr se hodí jako charakteristiky variability pouze částečně; v praxi je možno je použít pro první informaci.

E. F. Nakonec nás zajímají koeficienty souvislosti s rozkladovým znakem R. Aplikací vzorce (13a) dostaneme koeficient proporcionální predikce:

$$(28) \qquad \alpha_{B/R} = 1 - \frac{[\sum_r^R g_r \sum_i f_{i/r}(F_{i/r} - \frac{1}{2})\,(F_{i-1/r} - \frac{1}{2})] + \frac{1}{4}}{[\sum_i f_i(F_i - \frac{1}{2})\,(F_{i-1} - \frac{1}{2})] + \frac{1}{4}}$$

Obdobně, aplikací vzorce (13b), dostaneme

$$(29) \qquad \alpha^*_{B/R} = 1 - \frac{[\sum_r^R g_r(F_{M_r/r} - \frac{1}{2})\,(F_{M_{r-1}/r} - \frac{1}{2})] + \frac{1}{4}}{(F_M - \frac{1}{2})\,(F_{M-1} - \frac{1}{2}) + \frac{1}{4}}$$

kde M je pořadí mediánové skupiny v celé populaci π a M_r je pořadí mediánové skupiny v oblasti π_r, tj. číslo, které slouží pouze k identifikaci mediánové skupiny a tím příslušných hodnot funkce $F_{i/r}$; přitom není důležité, zda se číslování vztahuje k číslování skupin v π, či zda se vynechávají ty skupiny, které v π_r mizí.

Vlastnosti koeficientu $\alpha_{B/R}$ pro asymetrickou statistickou závislost mezi nominálním rozkladovým znakem R a spojitým ordinálním znakem B:

a) index je aplikovatelný pouze tehdy, jestliže variabilita v B je nenulová, tj. jestliže existují alespoň dvě obsazené nenulové skupiny (kategorie);

b) jestliže v jednotlivých oblastech R_r se vyskytuje vždy pouze jedna skupina, pak je predikce na základě znalosti R_r jednoznačná a $\alpha_{B/R} = 1$;

c) $\alpha_{B/R} = 0$, právě když B je nezávislé na R.

Vlastnosti koeficientu $\alpha^*_{B/R}$:

platí a) i b);

c) jestliže B je nezávislé na R, pak $\alpha^*_{B/R} = 0$.

Oba koeficienty jsou v mezích mezi nulou a jedničkou.

6. Numerický příklad

Na příkladu si ukážeme postup výpočtů pro všechny charakteristiky ordinálních znaků zavedené v této stati. Na jedné kontingenční tabulce si ukážeme výpočty pro oba modely: spojitý i diskrétní. Z instruktivních důvodů jsou uvedeny i mezivýpočty a různé kontroly.

Výchozí je tabulka 1.[17]

[17] Data tabulky jsou vzata ze sociologického výzkumu, rozložení však jsou získána jako náhodný výběr z původních, rozsáhlých dat. U znaku, který je uveden v tabulce, je těžké se rozhodnout pro jakoukoli kvantifikaci, neboť kategorie odpovědí nejsou dobře kvantifikovatelné (vzhledem k neurčité delce sezóny).

Tabulka 2: Výrazy $F(1-F)$ pro jednotlivá pole tabulky 1

Oblast	Kategorie							Součet ($\frac{1}{2}$ *dorvar*)	g_r
	(1)	(2)	(3)	(4)	(5)	(6)	(7)		
π_1	.2211	.2484	.1716	.1056	.0564	.0000	.0000	.8031	.50
π_2	.1539	.1056	.0000	.0000	.0000	0000.	.0000	.2595	.50
π	.2451	.2059	.0979	.0564	.0291	.0000	.0000	.6344	.5313

1. Výpočet koeficientu $\beta_{B/R}$ zahajujeme sestavením tabulky 2.[18] Je zbytečné výrazy $F(1-F)$ násobit dvěma, neboť při výpočtu koeficientu asociace se tento faktor zruší. V posledním řádku a posledním sloupci uvedené tabulky je hodnota

$$\Sigma g_r \Sigma F_{i/r}(1-F_{i/r}) = \\ = 0{,}50 \times 0{,}8031 + 0{,}50 \times 0{,}2595 = 0{,}53130$$

Dosazením do vzorce (19) dostaneme

$$\beta = \frac{0{,}6344 - 0{,}5313}{0{,}6344} = 0{,}1625$$

Proveďme ještě kontrolu rozkladu diskrétní ordinální variance pomocí vzorců (21) a (22). Nejprve vypočítáme výraz

$$\tfrac{1}{2} \Sigma \Sigma g_r g_s \Sigma(F_{i/r} - F_{i/s})^2 = 0{,}5 \times 0{,}5 \times \\ \times 0{,}5 \times 2[(0{,}33 - 0{,}81)^2 + (0{,}54 - 0{,}88)^2 + \\ + (0{,}78 - 1{,}00)^2 + (0{,}88 - 1{,}00)^2 + \\ + 0{,}94 - 1{,}00)^2 + (1{,}00 - 1{,}00)^2 + \\ + (1{,}00 - 1{,}00)^2] = 0{,}25 \times 0{,}4124 = 0{,}1031$$

A nyní musí platit součtová vlastnost

$$\begin{array}{ll} \Sigma g_r \Sigma F_{i/r}(1-F_{i/r}) & = 0{,}5313 \\ \tfrac{1}{2} \Sigma \Sigma g_r g_s \Sigma(F_{i/r} - F_{i/s})^2 & = 0{,}1031 \\ \hline \Sigma F_i(1-F_i) & = 0{,}6344 \end{array}$$

Poslední, součtový řádek se rovná $\frac{1}{2}$ *dorvar B*.

2. Výsledky výpočtů měr centrality a měr *Dorvar* jsou uvedeny v tabulce 3.[19] Míry centrality pro všechna pole jsou uvedeny jen pro ilustraci. Nebyly k dalším výpočtům pro

Tabulka 3: Míry centrality $\Sigma f_j|j - i|$ pro každé pole tabulky 1

Oblast	Kategorie							*Dorvar*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
π_1	1.53	1.19	1.27	1.83	2.59	3.47	4.47	1.19
π_2	.31	.93	1.69	2.69	3.69	4.69	5.69	.31
π	.92	1.06	1.48	2.26	3.14	4.08	5.08	.92

[18] Hodnoty v polích tabulky se vypočítají dosazením do příslušného součinu $F(1-F)$. Např. u oblasti osob mladších než 30 let bude pro první pole (nenavštěvuje)

$$F_{1/1}(1 - F_{1/1}) = 0{,}33(1 - 0{,}33) = 0{,}2211$$

Pro druhé pole je

$$F_{2/1}(1 - F_{2/1}) = 0{,}54(1 - 0{,}54) = 0{,}2484$$

Pro celkové rozložení a čtvrté pole (4—6krát za sezónu) je hodnota

$$F_4(1 - F_4) = 0{,}94(1 - 0{,}94) = 0{,}0564$$

[19] Výpočty hodnot v tabulce můžeme ilustrovat např. pro čtvrté pole první oblasti:

$$\Sigma f_{j/1}|j - 4| = 0{,}33|1 - 4| + 0{,}21|2 - 4| + \\ + 0{,}24|3 - 4| + 0{,}10|4 - 4| + 0{,}06|5 - 4| + \\ + 0{,}06|6 - 4| = 3 \times 0{,}33 + 2 \times 0{,}21 + 0{,}24 + \\ + 0{,}06 + 2 \times 0{,}06 = 1{,}83$$

Pro první pole celého rozložení dostaneme

$$\Sigma f_j|j - 1| = 0{,}57|1 - 1| + 0{,}14|2 - 1| + \\ + 0{,}18|3 - 1| + 0{,}05|4 - 1| + 0{,}03|5 - 1| + \\ + 0{,}03|6 - 1| = 0{,}92$$

Tabulka 4: Hodnoty $(F_m - \frac{1}{2})(F_{m-1} - \frac{1}{2}) + \frac{1}{4}$ *pro tabulku 1*

Oblast	Kategorie							*corvar*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
π_1	.3350 (.33)	.2432 (.21)	.2612 (.24)	.3564 (.10)	.4172 (.06)	.4700 (.06)	.5000 (0.0)	0.313182
π_2	.0950 (.81)	.3678 (.07)	.4400 (.12)	.5000 (0.0)	.5000 (0.0)	.5000 (0.0)	.5000 (0.0)	0.155496
π	.2150 (.57)	.2647 (.14)	.3319 (.18)	.4216 (.05)	.4568 (.03)	.4850 (.03)	.5000 (0.0)	0.268684

diskrétní model zapotřebí; stačilo spočítat pro každý řádek nejmenší z nich, která odpovídá mediánové kategorii.

Z tabulky 3 plyne také kontrola:

$$\sum f_i(\sum_j f_j|i-j|) =$$
$$= 2\sum F(1-F) \text{ pro každý řádek}$$

Např. pro druhý řádek

$$0{,}81 \times 0{,}31 + 0{,}07 \times 0{,}93 +$$
$$+ 0{,}12 \times 1{,}69 = 0{,}519 \text{ (levá strana)}$$
$$2 \times 0{,}2595 \quad = 0{,}519 \text{ (pravá strana)}$$

Dosadíme do vzorce (20):

$$\beta^*_{B/R} = 1 - \frac{0{,}50 \times 1{,}19 + 0{,}50 \times 0{,}31}{0{,}92}$$

$$\beta^*_{B/R} = 1 - \frac{0{,}75}{0{,}92} = 0{,}1848$$

Koeficienty β a β^*, které vyjadřují redukci variability rozkladem na dvě věkové skupiny, jsou značně vysoké a ukazují, že věkové skupiny vysvětlují 16,25 %, popř. 18,48 % chování vyjádřeného znakem *návštěvnost koncertů* ..., což je hodně.

3. Míry pro spojitý model vypočteme nejlépe tak, že nejprve sestavíme tabulku hodnot d^*_m z výrazu (25); tato tabulka 4 je obdobná k tabulce 3. Výpočty zde budou vycházet z distribučních funkcí.[20] Poslední sloupec vznikne podle vzorce (26), tj. vážením hodnot řádků příslušnými relativními četnostmi, které jsou pro názornost připojeny v závorkách. Vážené hodnoty v posledním sloupci tabulky jsou „míry variability" *corvar*.

Minimální hodnoty d_m určují v každém řádku mediánovou kategorii. Tyto mediánové hodnoty jsou současně příslušné „míry variability" *Corvar*.

Míry asociace vypočteme podle vzorců (28) a (29).

Koeficient proporcionální predikce:

$$\alpha_{B/R} =$$
$$= 1 - \frac{0{,}5 \times 0{,}3132 + 0{,}5 \times 0{,}1555}{0{,}2687} = 0{,}1278$$

Koeficient optimální predikce:

$$\alpha^*_{B/R} =$$
$$= 1 - \frac{0{,}5 \times 0{,}2432 + 0{,}5 \times 0{,}0950}{0{,}2150} = 0{,}2135$$

4. Všechny výsledky shrneme do přehledné tabulky 5. Nejde o to komentovat tyto výsledky z hlediska interpretace; je zřejmé, že číselné výsledky v tabulce odpovídají situaci a že jsme podobné číselné relace očekávali. Zde je nutno zdůraznit, že *řádkově nejsou hodnoty srovnatelné*, neboť každý sloupec má

[20] Jako příklad výpočtů hodnot pro tabulku 4 uvedeme nejdříve výpočet hodnoty ve třetím poli prvního řádku:

$$(F_{3/1} - \tfrac{1}{2})(F_{2/1} - \tfrac{1}{2}) + \tfrac{1}{4} =$$
$$= (0{,}79 - 0{,}50)(0{,}54 - 0{,}50) + 0{,}25 = 0{,}2612$$

Hodnotu v prvním poli posledního řádku dostaneme takto:

$$(F_1 - \tfrac{1}{2})(F_0 - \tfrac{1}{2}) + \tfrac{1}{4} =$$
$$= (0{,}57 - 0{,}50)(0{,}00 - 0{,}50) + 0{,}25 = 0{,}2150$$

Pro výpočet hodnot v prvním poli každého řádku potřebujeme znát F_0, popř. $F_{0/1}$ nebo $F_{0/2}$. Tyto hodnoty jsou vždy rovny nule.

Tabulka 5: Numerické hodnoty charakteristik ordinálního znaku z tabulky 1

Oblast	Mediánová kategorie (17)	Diskrétní model		Spojitý model	
		dorvar (16)	*Dorvar* (18)	*corvar* (26)	*Corvar* (27)
π_1	1× za sezónu	.8031	1.19	.3132	.2432
π_2	nenavštěvuje	.2595	.31	.1555	.0950
π	nenavštěvuje	.6344	.92	.2687	.2150
		$\beta = .1625$	$\beta^* = .1848$	$\alpha = .1278$	$\alpha^* = .2135$

jinou operacionální definici, a tedy *jiný význam*. Vyšší hodnoty koeficientů optimální predikce (obzvláště α^*) ve srovnání s koeficienty proporcionální predikce jsou způsobeny zřejmě tím, že koeficienty β^* a α^* (především α^*) jsou citlivější na obsazení mediánových kategorií. V našich datech jsou velice vysoké četnosti mediánových kategorií, což zvyšuje přesnost optimální predikce mediánovou kategorií a tím i koeficienty β^* a α^*.

7. Závěry

V práci byly zavedeny deskriptivní charakteristiky pro analýzu ordinálních dat. Byly předloženy dva modely, diskrétní a spojitý, které mají podstatně rozdílné základy. Rozhodnutí pro některý z obou modelů není jednoduché, neboť záleží na určení typu vlastnosti, kterou znak reprezentuje, a na tom, která funkce $d(a, b)$ danou vlastnost lépe charakterizuje.

Zde uvedené metody analýzy dat ovšem nejsou jedinými přístupy; v současné statistické literatuře lze nalézt několik zcela odlišných přístupů k testování hypotéz v kontingenčních tabulkách s ordinálními vstupy, tj. s uspořádanými kategoriemi. Přednosti obou předložených modelů spočívají v jejich prostotě a srovnatelnosti s přístupy k datům jiných typů.

Domnívám se, že míry d jsou pro oba případy (spojitý i diskrétní znak) užitečnější než d^* (stejně jako v případě nominálního znaku), neboť hodnotí četnostní rozložení jako celek a nepřeceňují význam jedné hodnoty (jíž je u d^* střed rozložení). Též koeficienty β, α mají z tohoto důvodu přednost před β^*, α^* (podobně jako τ má přednost před λ u nominálních znaků). Koeficient β je výpočetně obdobou koeficientu τ, jen místo relativních četností se zde dosazují hodnoty distribučních funkcí.

Z obou modelů má větší důležitost model diskrétní, neboť podle mého názoru je pro sociologická šetření daleko použitelnější. U tohoto modelu bude velice nadějné zkoumat rozklady analýzy rozptylu, jejich statistická rozložení a jejich použitelnost pro testování statistických hypotéz.

Zde naznačená teorie může být zobecněna a matematicky dále zkoumána. V této stati byly složitější matematické úvahy a formulace pokud možno vynechávány. Cílem tohoto článku bylo zavedení měr a jejich modelové operacionalizace, nikoliv zkoumání statistických vlastností. Proto zde nejsou uvedeny asymptotické konfidenční intervaly a další vlastnosti, které budou publikovány zvlášť.

Literatura

Gini, C.: *Srednije veličiny. Moskva,* Statistika 1970.

Good, J. I.: *Correlation for Power Functions.* Biometrics, 1972, s. 1127—1129.

Kullback, S.: *Teorija informacii i statistika.* Moskva, Nauka 1967 (překlad z angličtiny).

Labowitz, S.: *Some Observations on Measurement and Statistics.* Social Forces, *46,* December 1967, s. 151—160.

Light, R. J. - Margolin, B. H.: *An Analysis of Variance for Categorical Data.* Journal

of the American Statistical Association, *66*, September 1971, s. 534–544.
Margolin, B. H. - Light, R. J.: *An Analysis of Variance for Categorical Data. II: Small Sample Comparisons with Chi Square and Other Competitors.* Journal of the American Statistical Association, *69*, September 1974, s. 755–764.
Rodriguez, Milton da Silva: *On an Extension of the Concept of Moment with Applications to Measures of Variability, General Similarity and Overlapping.* Annals of Mathematical Statistics, *16*, 1945, s. 74–84.
Řehák, J.: *K pojmu znak v sociologii.* Sociologický časopis, 1972, č. 6, s. 615–625.
Řehák, J. - Řeháková, B.: *Měření statistické závislost nominálních znaků.* Sociologický časopis, 1973, č. 4, s. 404–418.
Řehák, J. - Řeháková, B.: *Míry statistické závislosti pro ordinální znaky.* Sociologický časopis, 1975, č. 1, s. 75–92.
Stevens, S. S.: *Measurement, Psychophysics and Utility.* In: Churchman, C. W. - Ratoosh, P. (eds): Measurement: Definitions and Theories. New York, John Wiley 1959.
Wilson, T. P.: *A Proportional-Reduction-in--Error Interpretation for Kendall's TAU - B.* Social Forces, *47*, March 1969, s. 340–342.
Wilson, T. P.: *Critique of Ordinal Variables.* Social Forces, *49*, March 1971, s. 432–444.

Резюме

Ржегак Я.: Основные дескриптивные меры для дистрибуции ординальных данных

В статье обсуждается исходная модель для разработки дескриптивных (описательных) характеристик. Указывается на то, в какой степени принятые для номинальных и числовых данных меры соответствуют этой модели и вместе с тем модель прилагается к ординальным данным.

Общая модель развивается в следующих шагах:

А. *Дается определение отношения несхожести, разброса, расстояния* $d(a, b)$ для двух элементов популяции, а именно по отношению к шкале изучаемого признака.

Б. Ожидаемое отношение $d(a, b)$ служит *мерой разброса* (3).

В. *Центр дистрибуции* определяется как значение признака, у которого предполагается наименьшее числовое значение несхожества (6).

Г. Ожидаемое значение отношения расстояния от центра дистрибуции является следующей мерой разброса, которая характеризует *концентрированность вокруг центра* (9).

Д. Связь классифицирующего признака и изучаемого признака мы измеряем индексом, который можно назвать *индексом объясняющей силы разбивки* и который возникает в качестве относительной редукции вариабильности по отношению к разбивке [формулы (13), (13a), (13b)].

Е. Предшествующие шаги просто интерпретировать при помощи *предикционной модели.* Оптимальным предиктором дистрибуции является его центр по отношению к ошибке предикции, вводящийся с помощью выражения несхожести $d(a, b)$. Для мер ассоциации (13) затем существует известная PRE-интерпретация (proportional reduction in error) для ассиметрических случаев.

Приложение этого метода к номинальным признакам (см. пример 1) дает при определении отношения $d(a, b)$ с помощью (1) известные меры вроде отношения повторимости (4), модуса (8) и отношения вариации (10); применение формулы (13a) дает $\tau_{A/R}$ Валлиса, а применение формулы (13b) — $\lambda_{A/R}$ Гуттмана. Для кардинального признака мы постепенно получаем дисперсию, среднюю и отношение корреляции (см. пример 2). Обе меры разбивки в данном случае эквивалентны (за исключением фактора 2) для непрерывного понимания признака. Для дискретного цифрового знака приложение формул (6) и (9) приводит к исправлению дисперсии в обычном понимании членом $(\bar{X} - X^*)^2$, где $\bar{X}$ есть средняя а X^* есть ближайшее к ней цифровое значение признака.

Приложение этого метода к ординальным признакам приводит к выделению двух моделей:

1. Дискретная ординальная модель предполагает упорядоченную категоризацию. Значение несхожества (14) выражает число категорий, которыми отличаются два элемента популяции. Его ожидаемое значение для всех пар обозначается как *dorvar* (дискретная ординальная вариация) и определяется формулой (15) или ее простой версией (16). Центром дистрибуции дискретного ординального признака является категория медианы (17). Мерой концентрации вокруг категории медианы является следующая характеристика разброса, означаемая как *Dorvar* (18). Статистическую зависимость дискретного ординального признака B от номинального (классифицирующего) признака R мы можем измерять с помощью коэффициента $\beta_{B|R}$ (19), у которого существует PRE-интерпретация для модели оптимальной предикции и может одновременно служить и как мера статистической зависимости между номинальным и ординальным признаками. Коэффициент $\beta^*_{B|R}$ служит как мера владеющая свойствами PRE-модели для оптимальной предикции и тоже может быть использована как мера статистической ассиметрической ассоциации между номинальным и ординальным признаками. Характеристика разброса *dorvar* обладает дистрибутивными свойствами (21) и (22), аналогичными тем, которые мы можем найти у CATANOVA и ANOVA; поэтому будет уместно взять ее в основу DORANOVA, т. е. непараметрического анализа дисперсии разброса для дискретных ординальных признаков (упорядоченой категоризации).

2. Непрерывная ординальная модель исходит из того, что на непрерывной шкале нельзя класть в основу несхожести никаких значений, ибо они произвольно монотонно изменяемы. Поэтому соответствующее отношение определяется количеством элементов ((23), (24)), встречающихся между избранными элементами

популяции. Аналогично определяются меры *corvar* (26) и *Corvar* (27) как меры разброса данных и категория медианы или, точнее, медианова группа в качестве центра дистрибуции. Далее приложением формул (13a) и (13b) мы получаем два коэффициента ассиметрической статистической зависимости между непрерывным ординальным признаком $\alpha_{B/R}$ (28), основанным на пропорциональной предикции, и $\alpha^*_{B/R}$ (29), основанным на оптимальной предикции.

Для аналитической работы с данными удобнее меры разброса и статистических зависимостей основанные на пропорциональной предикции, ибо они отражают поведение дистрибуции как целого. Меры, основанные на оптимальной предикционной модели, более удобны в качестве специальных показателей в реальных ситуациях предикции.

В 6 части приведен пример таблицы контингентов с двумя входами: строчная переменная определяет две равновеликие возрастные группы, а столбцовая переменная является ординальной переменной посещаемости концертов «поп-музыки». Для этой таблицы затем вычислены все характеристики ординальных данных, приведенные в статье.

Summary

Řehák J.: Basic Descriptive Measures for the Distribution of Ordinal Data

The present paper discusses the basic model for working out descriptive characteristics. Measures usual for nominal and cardinal (numerical) data are shown to correspond to this model. Furthermore, the model is applied to ordinal data.

The general model involves the following steps:

A. *The score of dissimilarity, dispersion, distance d(a, b)* is defined for two elements of the population with a view to the scale of the variable under examination.

B. The expected score *d(a, b)* serves as *a measure of dispersion* (3).

C. *The centre of distribution* is defined as a value of the variable having the lowest expected score of dissimilarity (6).

D. The expected value of the score of distance from the centre of distribution is another measure of dispersion characterizing *the concentration around the centre* (9).

E. The relation between the decompositional variable and the variable under examination is measured by an index which can be designated as the *index of the explanatory power of decomposition* and arises as a relative reduction of variability in relation to the given decomposition [(13), (13a), (13b)].

F. The preceding procedure has also a simple interpretation with the help of the *prediction model.* The optimum distribution predictor is its mean in relation to the prediction error introduced by means of the score of dissimilarity *d(a, b).* Measures of association (13) then have the well-known PRE-interpretation (i.e. proportional reduction in error) for the asymmetric case.

In ascertaining the score *d(a, b)* by means of (1), the application of this procedure to nominal variables (see example 1) yields the known measures, as repetition ratio (4), modus (8) and variation ratio (10); by applying formula (13a), we get Wallis' τ, and by applying formula (13 b), we get Guttman's λ. For the cardinal variable, the dispersion, the average and the correlational ratio are successively obtained (see example 2). In this case, both the dispersion measures are equivalent (with the exception of factor 2) for the continuous conception of the variable. As regards the discrete numerical variable, the application of formulae (6) and (9) leads to the correction of the currently conceived dispersion by the member $(\overline{X} - X^*)^2$, where $\overline{X}$ is the average and X^* is the numerical value of the variable nearest to it.

The application of the procedure to ordinal variables leads to the differentiation of two models:

1. The discrete ordinal model presupposes an ordered categorization. The score of dissimilarity (14) expresses the number of categories by which two population elements differ. The expected value of this score for all the pairs is designated as *dorvar* (i.e. discrete ordinal variance), and is given by formula (15) or by its simple version (16). The centre of distribution follows as the median category (17). Formula (18) yields another measure of variability, a measure of concentration around the centre called *Dorvar.* The statistical dependence of the discrete ordinal variable B on the nominal (decompositional) variable R can be measured by means of the coefficient $\beta_{B/R}$ (19), having the PRE-interpretation for the model of proportional prediction. The coefficient $\beta^*_{B/R}$ (20) then has the PRE-interpretation for the optimum prediction model, and can also serve as the measure of statistical dependence between the nominal and the ordinal variables. The dispersion characteristic *dorvar* has decompositional qualities (21) and (22), analogical to those that can be found in CATANOVA and ANOVA; consequently, it will be adequate to make it the basis of DORANOVA, i.e. of the nonparametric analysis of variance for discrete ordinal variables (ordered categorizations).

2. The continuous ordinal model is based on the fact that, on the continuous scale, no values can be considered as the basis of dissimilarity, since they may be arbitrarily monotonously transformed. This is why the number of elements occurring among the selected population elements is taken as the dissimilarity score (23), (24). Analogically, the measures *corvar* (26) and *Corvar* (27) are determined as measures of data dispersion, and the *median category* or, better said, the

median group as the centre of distribution. By the application of formulae (13a) and (13b), two coefficients of the asymmetrical statistical dependence between the nominal and the continuous ordinal variable are further obtained: $\alpha_{B/R}$ (28), based on proportional prediction, and $\alpha^*_{B/R}$ (29), based on optimum prediction.

Measures of dispersion and statistical dependences based on proportional prediction are adequate for analytical work with data, due to the fact that they reflect the decomposition of the dispersion as a whole. Measures based on the optimum prediction are more convenient as special indicators of prediction quality.

In part 6, an example of a two-way contingency table is introduced: the row variable defines two age groups of equal size, and the column variable is the ordinal variable of attendance at „pop-music" concerts. For this table, all the characteristics of ordinal data mentioned in the present paper are successively presented.