O jednom přístupu k analýze prvotní sociologické informace

B. G. MIRKIN
Novosibirsk

1. Práce rozvíjí nový přístup k matematické analýze a zpracování prvotní sociologické informace. Tento přístup se vyznačuje těmito zvláštnostmi:

— je aplikovatelný na zpracování kvalitativních dat;
— je založen na přísně axiomatickém přístupu;
— umožňuje řešit širokou škálu úloh pro zpracování sociologických dat standardními metodami;
— umožňuje ruční výpočty.

Některé z popsaných metod byly již aplikovány v praxi s dobrými výsledky.

2. Prověřované sociologické údaje je možné znázornit tímto způsobem:

Mějme konečnou množinu objektů $\mathfrak{N} = \{e_1, e_2, \ldots, e_n\}$, přičemž každý objekt $e_K \in \mathfrak{N}$ je charakterizován hodnotami znaků (proměnných) $P_1, P_2, \ldots, P_n$. Znaky mohou být např. výška, pohlaví, profese, spokojenost s vykonávanou prací atd.

Předpokládáme, že každý znak P_i, může nabývat m_i hodnot:

$$p_i^1, p_i^2, \ldots, p_i^{m_i} \quad (i=1, \ldots, n).$$

Každý objekt e_K je tedy charakterizován uspořádaným souborem $p(e_K) = \langle p_1(e_K), \ldots, p_n(e_K) \rangle$, kde $p_i(e_K)$ jsou hodnoty znaku P_i ($i = 1, 2, \ldots, n$), odpovídající objektu e_K.

Každý znak P_i generuje rozklad množiny $\mathfrak{N}$ na m_i tříd, z nichž do jedné patří objekty e_K, pro něž $p_i(e_K) = p_i^1$ atd.; poslední třída je složena z takových objektů e_K, pro něž $p_i(e_K) = p_i^{m_i}$. Rozklad tohoto typu budeme označovat P_i.

Je užitečné rozlišovat tyto typy znaků:

P nazveme kvantitativním znakem, jsou-li jeho hodnotami $p^1, p^2, \ldots, p^m$ čísla. Takovými znaky jsou např. výška, mzda, apod. P nazveme pořadovým znakem, jestliže jeho hodnoty $p^1, \ldots, p^m$ nejsou čísla, ale charakterizují nám rozdílný stupeň projevu tohoto znaku, takže mezi nimi existuje přirozené uspořádání dané stupněm projevu znaku. Pořadovými znaky jsou např. kvalifikace (od „nízké" do „vysoké"), spokojenost s vykonávanou prací (od „velmi spokojen" do „velmi nespokojen"). P je klasifikačním znakem, jestliže jeho hodnoty nejsou čísly a nejsou vázány přirozeným uspořádáním. Do klasifikačních znaků je možné zahrnout pohlaví, profesi, příčiny opuštění daného města atd.

Existující metody zpracování údajů sociologického zkoumání jsou v podstatě spjaty s problematikou zpracování kvantitativních znaků. Pořadové znaky se při zpracování obvykle převádějí na kvantitativní (ne vždy evidentně) bodovým ohodnocením a pak se aplikují statistické metody.

Statistické zpracování takových obodovaných znaků (např. výpočet střední hodnoty, disperze apod.) vzbuzuje silné námitky, protože obecně řečeno přiřazené body charakterizují pouze uspořádání hodnot znaku a aplikace aritmetických operací na ně vyžaduje zdůvodnění pro každý konkrétní případ. Studium otázek spjatých s takovým zdůvodněním si vyžádalo přezkoumat a výrazně prohloubit koncepci měření. Výsledky získané v r. 1963 byly shrnuty ve sborníku [2], ze kterého je dobře zřejmé, že rozpracování takového zdůvodnění je zatím v počátečním stadiu.

Ještě nejasnější jsou otázky související se zpracováním klasifikačních znaků. Někdy se hodnoty klasifikačního znaku předběžně uspořádají (s přiřazením bodů) v souvislosti s některou obsahovou hypotézou nebo míněním expertů. V některých případech se použijí statistické ukazatele, které mají smysl i pro klasifikační znaky (např. modus nebo entropie rozdělení). Jindy zavádíme speciální ukazatele zdůvodněné heuristicky. Např. nejvíce upotřebitelnými ukazateli závislosti klasifikačních znaků jsou koeficienty kontigence Pearsona a Čuprova [1], které v podstatě ohodnocují statistický význam odlišnosti empirického rozdělení od rovnoměrného teore-

tického rozdělení. Takových heuristických ukazatelů je možné zavádět mnoho, ovšem kterému dát přednost — to jasné není.

Existující metody zpracování údajů jsou z těchto důvodů v podstatě zaměřeny na kvantitativní znaky. Badatele to nutí využít pro komplexní zpracování údajů především kvantitativní nebo pořadové, „bodové" znaky.

A zatím klasifikační znaky hrají významnou úlohu při charakterizování sociologických objektů. Proto tedy je rozpracování speciálního aparátu pro hodnocení klasifikačních znaků přitažlivé. Aparát takového typu by měl z jednotícího hlediska řešit široký okruh úloh.

Poznamenejme, že při zpracování dat je v terminologii klasifikačních znaků v řadě případů potřebné ke sjednocení výsledků využít kvantitativní a pořadové znaky jako klasifikační (analogicky je tomu tak, když převádíme znaky jiných typů na kvantitativní). Přitom ztrácíme část informace. Zpravidla je však počet hodnot znaků sociologických objektů nesrovnatelně menší než úhrn N sledovaných objektů, a proto základní informace o znacích je obsažena v jim odpovídajících rozkladech a ztráta informace je malá. Současně existující metody zpracování dat, které převádějí znaky na čísla, nesou s sebou dodatečnou informaci o kvantitativních znacích, jejíž spolehlivost je většinou velmi málo opodstatněna.

Možný způsob vytvoření aparátu pro zpracování klasifikačních znaků, který dává jednotné metody řešení širokého okruhu úloh, budeme uvádět v dalším.

3. Informace o klasifikačním znaku P, získaná zkoumáním množiny $\mathfrak{N}$, je uchovávána rozkladem P, na třídy objektů odpovídajících jedněm a týmž hodnotám znaků. Proto se problém studia znaků převádí na studium odpovídajících rozkladů.

Za základní nástroj zpracování klasifikačních znaků budeme považovat kvantitativní míru těsnosti rozkladů dané množiny $\mathfrak{N}$. Míru těsnosti rozkladů P, Q označíme d (P, Q) a požadujeme, aby splňovala jisté podmínky.

Jednou z těchto podmínek je požadavek, aby d (P, Q) měla základní vlastnosti geometrické délky. Požadavek tohoto typu je zdůvodněn potřebami další práce s veličinou d (P, Q) vzhledem k tomu, že použití matematického aparátu je nejefektivnější a přirozené pro míry mající vlastnosti geometrické délky. Je možné namítat, že taková podmínka nesouvisí s vnitřními vlastnostmi znaků nebo objektů, a je dokonce možné, že v jistých případech je s nimi nekonzistentní. A priori však není možné takovou nekonzistenci objevit a její eventuální projev, zvláště při praktické aplikaci „geometrických" měr, je sám o sobě důležitým výsledkem. Přednosti pohodlného zpracování jsou však zatím rozhodující. Uvedeme přesnou formulaci prvé podmínky.

Axióm 1. Míra d (P, Q) má tyto vlastnosti geometrické délky:

a) $d(P, Q) \geq 0$ a $d(P, Q) = 0$ tehdy a jenom tehdy, jestliže $P = Q$ (tzn. třídy P jsou identické s třídami Q);

b) $d(P, Q) = d(Q, P)$;

c) pro libovolné rozklady P, Q, R $d(P, Q) \leq d(P, R) + d(R, Q)$, přičemž přesné rovnosti je dosaženo tehdy, jestliže rozklad R leží mezi rozklady P, Q.[1] [3]

Následující axióm je diktován nutností zajistit rovnoprávnost všech objektů $e_K \in \mathfrak{N}$ vzhledem k míře d (P, Q), a to bez ohledu na jejich vnitřní specifiku.

Axióm 2. Získáme-li rozklad P' z rozkladu P permutací některých objektů a rozklad Q' z Q toutéž permutací, pak $d(P' Q') = d(P, Q)$. Nyní potřebujeme, aby při dílčí totožnosti rozkladů P, Q bylo možné pro výpočet vzdálenosti P, Q použít těch tříd, které nejsou totožné.

Axióm 3. Jsou-li rozklady P, Q všude totožné s výjimkou množiny $E \in \mathfrak{N}$, která je jejich segmentem,[2] pak d (P, Q) počítáme tak, že bereme v úvahu pouze rozklad na množně E.

Axióm 4. Maximální vzdálenost mezi rozklady množiny $\mathfrak{N}$ je rovna $\frac{N(N-1)}{2}$ [3].

[1] To znamená, že pro libovolné dva objekty $e_k, e_l \in \mathfrak{N}$:
a) patří-li do jedné třídy v rozkladu P a v rozkladu Q, pak patří do jedné třídy rozkladu R;
b) nalézají-li se v různých třídách jak v rozkladu P, tak i v rozkladu Q, pak se nalézají v různých třídách rozkladu R [3].

[2] Segmentem rozkladu nazýváme množinu, která je sjednocením jistých jeho tříd.

[3] Bylo by možné brát libovolné jiné číslo; použité číslo vede k výhodnému tvaru vzorce pro d (P, Q).

Platí následující výsledek [3]:

Míra d (P, Q) těsnosti rozkladů splňující axiómy 1—4 existuje a je jednoznačně definována.

Můžeme se přesvědčit, že axiómy **1—4** jsou splněny pro následující míru

$$d(P,Q) = \frac{1}{2} \sum_{k,l=1}^{N} |p_{kl} - q_{kl}| \qquad (*)$$

kde $p_{kl} = \begin{cases} 1, \text{ jestliže } e_k, e_e \text{ jsou v jedné třídě P;} \\ 0, \text{ jestliže } e_e, e_k \text{ jsou v různých třídách P} \end{cases}$

q_{kl} je definováno analogicky pro Q.

Pokládáme-li tedy axiómy 1—4 za přirozené, pak d (P, Q) vypočítáme ze vzorce (*): v důsledku platnosti věty žádná jiná míra nesplňuje všechny axiómy 1—4 současně.

Máme-li kombinované seskupování podle znaků P, Q, takže $N_{i.}$ označuje počet osob odpovídajících i—té hodnotě znaku P, $N_{.j}$ počet osob odpovídajících j—té hodnotě znaku Q a N_{ij} počet osob odpovídajících i—té hodnotě znaku P a j—té hodnotě znaku Q současně, pak míru d (P, Q) můžeme počítat z následujícího vzorce ekvivalentního vzorci (*):

$$d(P,Q) = \frac{1}{2}\left(\sum_i N_{i.}^2 + \sum_j N_{.j}^2 - 2\sum_{ij} N_{ij}^2\right) \qquad (**)$$

kde součet jde přes všechny hodnoty i, j, znaků P, Q.

Užití vzorce (**) umožňuje ruční výpočet veličiny d (P, Q) pro množiny složené z tisíce objektů.

4. Nyní vyšetřme v terminologii míry d (P, Q) některé úlohy a metody jejich řešení, se kterými se setkáváme při zpracování údajů sociologického výzkumu.

1⁰ *Klasifikace*

Pod termínem klasifikace (taxonomie) rozumíme obvykle rozklad zkoumaného souboru na skupiny objektů „podobných si" vzhledem k vybranému systému znaků. V naší terminologii klasifikací rozumíme rozklad množiny $\mathfrak{N}$ na třídy objektů $e_K \in \mathfrak{N}$ „blízkých si" vzhledem k systému znaků $P_1, \ldots, P_n$.

Jestliže systém sestává z jednoho znaku P_1, pak je klasifikací příslušný rozklad P_1, jestliže existují dva znaky P_1 a P_2, je klasifikací rozklad, který je v určitém smyslu průměrem rozkladů P_1 a P_2.

Pro větší přesnost vezmeme rozklad Π ležící mezi (ve smyslu poznámky 1) rozklady P_1 a P_2, takový, aby absolutní hodnota diference mezi d (Π, P_1) a d (Π, P_2) byla minimální. Nulové hodnoty nemůže nabývat, protože prostor rozkladu je diskrétní.

Jestliže tento postup zobecníme, pak v analogii se zjišťováním těžiště v eukleidovském prostoru můžeme říci, že těžištěm je rozklad Π ležící mezi P a Q; je-li přiřazena rozkladu P váha p a rozkladu Q váha q, je pak $\left|\frac{d(P,\Pi)}{d(\Pi,Q)} - \frac{p}{q}\right|$ minimální.

Těžiště systému složeného ze tří rozkladů P_1, P_2, P_3, je potom možné určit jako těžiště rozkladu P_1 s vahou 1 a rozkladu Π s vahou 2, kde Π je těžištěm rozkladů P_2 a P_3 (s totožnými vahami). Podobně je možné určit těžiště systému složeného z n rozkladů.

V obecném případě n znaku P_1, P_2, ..., P_n to umožňuje pro klasifikaci použít těžiště množiny rozkladů P_1, P_2, ..., P_n.

Poznamenejme, že hledání těžiště je v důsledku platnosti vzorce (**) převedeno na jistou manipulaci s kombinačními skupinami znaků P_1, ..., P_n. Na rozdíl od dnes existujících způsobů [4] nevyžaduje daná metoda odhad těsnosti mezi zkoumanými objekty $e_K \in \mathfrak{N}$ V současných metodách je naopak zjištění vzdálenosti mezi $e_K \in \mathfrak{N}$ základní obtíží vzhledem k tomu, že je pro ně nezbytná souměřitelnost různých znaků.

2⁰ *Hodnocení závislosti znaků*

Obsažnější úvahy ukazují, že blízké znaky jsou charakterizovány blízkými rozklady. Vezměme proto míru těsnosti znaků δ (P, Q) úměrnou míře těsnosti odpovídajících rozkladů d (P, Q). Pro větší snadnost vybereme koeficient úměrnosti tak, aby maximální vzdálenost mezi znaky byla rovna 1. Pak v souhlase se vzorcem (*):

$$\delta(P,Q) = \frac{2d(P,Q)}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i,j=1}^{N} |p_{ij} - q_{ij}|$$

Skutečnost δ P, Q) $\approx$ O znamená, že znaky P, Q se vzájemně dublují;

δ (P, Q) $\approx$ 1 znamená, že znaky P, Q jsou nezávislé.

Poznamenejme, že δ (P, Q) je v podstatě statistický odhad pravděpodobnosti δ^* (P, Q) toho, že v základním souboru se libovolné 2 objekty nalézají v jedné třídě jednoho z rozkladů P, Q a v různých třídách jiného rozkladu.

Analogicky k výsledkům stati [5] můžeme dokázat, že diference $\sqrt{N}\,|\delta - \delta^*|$ konverguje k normálnímu rozdělení s nulovou matematickou nadějí a konečnou disperzí (A. V. Bekker).

3^0 *Hodnocení významnosti znaků*

Mějme na množině $\mathfrak{N}$ dán jistý rozklad R. Je pochopitelné, že význam znaku P pro rozklad R se projeví tím více, čím blíže si budou rozklady P, R. Z toho plyne, že významnost znaku P vzhledem k rozkladu R je nepřímo úměrná vzdálenosti d (P, R).

Absolutní významnost znaku P je jeho významnost vztažená ke klasifikaci popsané v 1^0. Jak je zvykem, chápeme zde významnost znaku ve smyslu jeho informativnosti. V konkrétních výzkumech se však často taková významnost traktuje ve smyslu „síly vlivu" znaku P na faktor vytvářející rozklad R. Že takový přístup není neopodstatněný, potvrzuje se v konkrétních situacích tím, že neodporuje utvořeným představám.

4^0 *Faktorová analýza*

Studujme úlohu měření relativně nezávislých faktorů v daném systému závislých znaků. Úloha tohoto typu se pro kvantitativní znaky řeší obvykle metodami faktorové analýzy: k daným znakům P_i vytváříme lineární kombinace R_j (faktory), mezi nimiž je korelace nulová [6].

V případě klasifikačních znaků budeme postupovat následovně: Sestavíme matici δ (P_i, P_j) vzdáleností mezi znaky a rozdělíme znaky P_i na skupiny (je možné i s neprázdným průnikem), uvnitř kterých jsou si znaky maximálně blízké, a na další, kde jsou si maximálně vzdáleny. To je možné provést např. známými metodami taxonomie [4]. Získané skupiny jsou hledanými faktory. Jsou sice pouze relativně nezávislé, ale ve faktorové analýze platí: rovná-li se korelace nule, jde pouze o nutnou, a nikoliv postačující podmínku nezávislosti. Rozklady odpovídající příslušným faktorům získáme jako těžiště množiny znaků vytvářejících faktory. Známe-li tyto rozklady, můžeme ohodnotit významnost znaků ve faktorech (to zn. „faktorové zatížení" znaků) v souhlase s 3^0.

Literatura:

[1] Yule G. and M. G. Kendall, *An Introduction to the Theory of Statistics*, 14 th ed; Ch. Griffin and Co, London 1950.

[2] *Hanbook Mathematical Psychology*, ed R. Luce, B. Bush, E. Galanter, vol. 1. New York—London, J. Wiley and Sons, 1963.

(3) Миркин Б. Г., Черный Л. Б: „Об измерения расстояния между разбиениями конечного множества". — В кн.: *Математические методы моделирования и решения экономических задач*. Новосибирск, „Наука", 1968.

(4) *Распознавание образов в социальных исследованиях*. Новосибирск, „Наука", 1968.

[5] L. Goodman, W. Kruskal, *Measures Association for Cross Classifications, III: Approximate Sampling Theory Journ.* Amer. Stat., 1963, N 2, p. 310—365.

(6) Лоули Д., Максвелл А. *Факторный анализ как статистический метод*. М., 1967.