Míry statistické závislosti pro ordinální znaky

JAN ŘEHÁK
Ústav pro filozofii a sociologii ČSAV, Praha
BLANKA ŘEHÁKOVÁ
Ústav hematologie a krevní transfúze, Praha

V sociologickém výzkumu rozlišujeme různé typy znaků (viz např. [13], z nichž při statistickém zpracování sociologických dat většinou používáme tři základní: nominální, ordinální a kardinální. Ostatní typy se používají ve speciálních situacích: data o malých skupinách, zpracování výsledků škálovacích postupů, zkoumání relací u vlastností, tvorbu a zavádění sociologických proměnných a podobně.

V této práci se zabýváme technickými aspekty metodologické práce s ordinálními znaky. Navazujeme tak na stať [14], kde jsme shrnuli míry statistické závislosti pro nominální znaky. I u ordinálních měr je důležité vědět, jaká je logika a racionální základy každé míry. Chceme a musíme znát statisticko pravděpodobnostní model, z něhož je ta která míra odvozena. Obdobně jako v [14], ani zde se nezabýváme statistickou inferencí, tj. konfidenčními intervaly a testováním hypotéz. Vede nás k tomu nedostatek místa a vědomí, že hlavním úkolem sociologické metodologie v tomto oboru je vybrat, popřípadě navrhnout vhodné a interpretovatelné míry pro populaci. Proces zobecňování z výběrových souborů na cílové populace je záležitost jiná, pro sociologa stejně důležitá, avšak ještě techničtější (i když výběr inferenční strategie plyne z úvah filozofických a z obecné statistické metodologie). Čtenáře, který se zajímá o inferenci spojenou s uváděnými koeficienty, odkazujeme na uvedenou literaturu. Dále chceme zdůraznit, že se zabýváme (stejně jako v [14]) měřením *statistické závislosti určitého typu*, že se pohybujeme na poli technické části metodologie, na poli matematických abstrakcí odvozovaných z různých konkrétních věd.[1] Nejdeme a ani nemůžeme jít za hranice statistické interpretace, a to ani do problematiky filozofického rozboru pojmu závislosti, ani do interpretací konkrétních sociologických dat.[2]

Značení, které používáme, je obvyklé ve statistické literatuře i v učebnicích, nebudeme ho proto rozvádět v tabulce (odkazujeme na podrobnější ilustraci v [14]). Kontingenční tabulka, kterou tvoří kombinace dvou znaků, má rozměry $r \times s$, kde

r = počet kategorií řádkové proměnné (počet řádků tabulky)

s = počet kategorií sloupcové proměnné (počet sloupců tabulky)[3]

n_{ij} = absolutní četnost v poli i-tého řádku a j-tého sloupce tabulky

$n_{i.}$ = součet absolutních četností v i-tém řádku

$n_{.j}$ = součet absolutních četností v j-tém sloupci

n = celkový počet případů v tabulce

$p_{ij}, p_{i.}, p_{.j}$ jsou obdobně relativní četnosti vzhledem k celkovému počtu n případů

$p_{j/i}$ = relativní četnost j-té kategorie vzhledem k případům v i-tém řádku ($p_{j/i} = n_{ij}/n_{i.}$)

Kategorie proměnných jsou uspořádány podle vzrůstajících hodnot obou proměnných. Jedná-li se o asymetrický vztah, řádková proměnná je nezávislá (příčina, prediktor, časově předchází), sloupcová proměnná je závislá (důsledek, predikant, časově následující). Nezávisle proměnnou značíme X a její hodnoty $X_1, X_2, \ldots, X_r$, závislou proměnnou značíme Y a její hodnoty $Y_1, Y_2, \ldots, Y_s$ (indexy u hodnot značí pořadí v jejich uspořádání u znaku).

Ordinální a nominální statistická závislost

Mluvíme-li o ordinální asociaci či ordinální statistické závislosti, máme vždy na mysli

[1] Je zajímavé, že právě na utváření ordinálních měr se sociologie podílela více než na tvorbě jiných statistických postupů. Plyne to z potřeby zpracování dat, která jsou velmi často právě ordinálního charakteru.

[2] Interpretace konkrétních dat vždy závisí na tom kterém případě a na posouzení používaného modelu závislosti a souvislosti proměnných jako celku.

[3] V literatuře se také značí rozměry $r \times c$, což pochází z angličtiny a mnemotechnicky naznačuje řádky (rows) a sloupce (columns).

vztah dvou ordinálních znaků, tj. kategorizací s uspořádanými kategoriemi (hodnotami). Statistická nezávislost se definuje stejně jako v nominálním případě — je to situace, ve které neexistuje žádná asociace, v níž nelze nalézt žádnou informaci, žádnou predikční schopnost jedné proměnné o druhé. Matematicky lze nezávislost vyjádřit dobře známou terminologií: očekávané četnosti jsou rovny skutečným empirickým četnostem

$$(1) \quad n_{ij} = \frac{n_{i.} \cdot n_{.j}}{n} \text{ resp. } p_{ij} = p_{i.} \cdot p_{.j}$$

Tento vztah plyne ze vztahu (a je mu i ekvivalentní), který lépe odpovídá významu pojmu

(2) $p_{j/i} = p_{.j}$ pro všechna j a pro všechna i (řádkové relativní četnosti jsou stejné).

Pojem statistické závislosti je velice bohatý na různé případy. Podle typu znaků můžeme rozlišovat i různé typy závislostí. Zatímco u kardinálních znaků můžeme zkoumat velmi jemné závislosti odpovídající různým funkcím (lineární, exponenciální, mocniny, polynomy, logické S-křivky), u nominálních znaků charakterizujeme pouze odchýlení od vztahu (1) nebo (2). U nominálních znaků (vzhledem k jejich obsahu a smyslu) nesmí být míry statistické závislosti citlivé na permutování sloupců nebo řádků — musí mít tedy stejnou hodnotu při libovolném uspořádání kategorií. Kardinální míry naproti tomu závisí mnohdy i na nepatrných transformacích hodnot.[4]

Ordinální znak stojí někde uprostřed mezi nominálním a kardinálním. Jelikož zde uspořádání hraje velkou roli, nemůžeme libovolně permutovat a odtud plyne požadavek, aby ordinální míry byly citlivé na uspořádání hodnot znaku. Naprosto však nezáleží na tom, jaké hodnoty přisoudíme jednotlivým kategoriím, zda je například očíslujeme vzestupně počínaje jedničkou, či zda jim přiřadíme kód nějakého číselníku.

Ordinální znak reprezentuje vlastnost, jejíž stavy jsou odstupňovány. Podle této vlastnosti můžeme statistické jednotky úplně nebo částečně uspořádat (pomocí hodnot ordinálního znaku je můžeme uspořádat pouze částečně). Má tedy smysl hovořit o vyšším či nižším stupni vlastnosti a paralelně o vyšší či nižší hodnotě znaku. Při zkoumání statistické závislosti ordinálních znaků nás proto zajímají nejen libovolné odchylky od stavu nezávislosti, ale i typ závislosti. Jakmile zavedeme nižší a vyšší hodnoty znaku, zajímá nás (a má smysl o tom uvažovat), zda se v populaci vyskytují současně vyšší (a tedy na druhé straně nižší) hodnoty u obou znaků, či zda se společně vyskytují vyšší hodnoty jednoho znaku s nižšími hodnotami druhého. U asymetrických situací s příčinnou interpretací je pak zajímavé, zda nižší hodnoty nezávisle proměnné způsobují nižší či vyšší hodnoty závisle proměnné. V tomto se již projevuje *ordinální statistická závislost*, kterou později u jednotlivých měr operacionálně upřesníme. Intuitivní charakter pojmu je však tím dán. Z toho, co bylo řečeno výše, rovněž plyne, že má smysl hovořit o přímé a nepřímé statistické ordinální závislosti. O *přímé závislosti* mluvíme tehdy, párují-li se vysoké hodnoty X s vysokými hodnotami Y a nízké s nízkými. *Nepřímá závislost* vzniká tehdy, párují-li se nízké s vysokými. Míry ordinální závislosti jsou konstruovány tak, aby odrážely i tuto vlastnost (znaménko „+" vyznačí přímou závislost, „—" nepřímou).

Míry, které v další části uvedeme, jsou tedy konstruovány proto, aby charakterizovaly ordinalitu vztahu. Je zapotřebí zdůraznit, že nula či nízké hodnoty koeficientů neznamenají statistickou nezávislost či situaci jí blízkou. Ukazují pouze nepřítomnost ordinální závislosti (což neznamená pochopitelně nezávislost) nebo její nízký stupeň.[5] Proto také zkoumáme-li v sociologickém výzkumu závislosti u ordinálních znaků, počítáme nejen míry ordinální, ale i nominální asociace. Případ současného výskytu nízké ordinální závislosti a vysoké nominální závislosti vyžaduje podrobnější analýzu rozložení četností v tabulce[6] (někdy může rovněž indikovat nedostatky při konstrukci znaků).

Znaménko u měr a jeho interpretace závisí na orientaci uspořádání kategorií znaku. Například znak vlastnosti „důležitost určitého jevu pro respondenta" může mít hodnoty orientovány (gradovány) od „nedůležitý" až

[4]) Závisí ovšem na typu koeficientu, např. koeficient lineární korelace r je necitlivý na libovolné lineární transformace obou proměnných (s jediným omezením, že součin směrnic musí být kladný).

[5]) Chybná interpretace nízkých hodnot ordinálních měr je velmi častá; může vést a často vede k nepříjemným chybám v závěrech z dat.

[6]) Zde také možno vhodně využít znaménkových schémat, která pomáhají odhalit typ závislosti v tabulce se vyskytující pomocí obrazce plusových a minusových znamének.

po „velmi důležitý", ale může to být i naopak. Formálně matematicky je zcela libovolné, kterou z obou možností použijeme při uspořádání kategorií v tabulce. Obsahově ji ovšem musíme přesně určit a hlavně na ni nesmíme zapomenout při interpretaci. Změníme-li totiž orientaci jedné proměnné, změní se znaménko koeficientu (při zachování absolutní hodnoty). Toto je logické, neboť struktura a vzájemné vztahy četností v tabulce se nezmění, změníme-li například pořadí sloupců v opačné; četnosti jsou pouze zrcadlově obráceny a to se projeví na znaménku.

Ordinální znaky byly ve vývoji statistiky značně zanedbávány. Zájem se soustředil na zpracování číselných výsledků a souběžně (i když daleko méně) na zpracování kontingenčních tabulek s nominálními vstupy (systematické rozpracování a doplňování statistické teorie pro analýzu nominálních znaků začalo však poměrně nedávno a zdaleka ještě není ukončeno). Ordinální znaky byly studovány v oblasti neparametrických technik, ordinální asociace v problematice neparametrických korelací. To ovšem odpovídá situaci prostých ordinálních znaků (uvažovala se data odpovídající úplnému uspořádání souboru). Tento případ je pro sociologický dotazník netypický[7], a proto tyto metody nejsou pro analýzu dat sociologického dotazníku vhodné.

První míry ordinální asociace vznikly úpravou neparametrických korelačních koeficientů (Spearmanova a Kendallova) na situaci tzv. *spojení* (dvě nebo více jednotek má stejné pořadí). Stejné znamenalo *původně nerozlišitelné*, avšak v analýze kontingenčních tabulek byl tento význam přeměněn na *nerozlišované*. To bylo provedeno proto, že takto upravené korelační míry měly vhodné matematické vlastnosti a splňovaly požadavky na míry ordinální asociace. Jejich nedostatkem bylo, že model vycházel z úplných uspořádání a že přechod od významu nerozlišitelné k významu nerozlišované spojení byl proveden poněkud lehce.

Přelom v používání měr statistické závislosti nastal po publikaci velice zásadních statistických článků [4], [5] a [6], které daly podnět k požadavku pravděpodobnostní interpretace a modelování racionálních základů pro každou míru. Naší snahou je zavádět koeficienty, u nichž numerické hodnoty mají jednoduchou a pochopitelnou interpretaci a míry mají operacionalizací daný pevný obsah. Používání měr tohoto typu je už dnes běžné. Stupeň složitosti modelu je ovšem různý. U některých měr (Spearmanův koeficient ϱ) byl model nalezen a posteriori a je tak nejasný, že se prakticky nedá využít (viz [10]).

Nejjednodušší princip modelování měr je PRE-princip.[8] Míry vznikají jako relativní úbytek chyby predikace závislé proměnné bez použití a při použití nezávisle proměnné v predikčním pravidle (viz např. [1], [4], [14]). Koeficient PRE je dán jako

$$\text{PRE} = \frac{P(I) - P(II)}{P(I)}, \tag{3}$$

kde $P(I)$ je pravděpodobnost chyby predikce Y bez znalosti X a $P(II)$ je pravděpodobnost chyby predikce Y s využitím znalosti X.

Volba predikčního pravidla je různá podle problému. U symetrického případu randomizujeme směr predikce — počítáme průměrné chyby predikce v obou směrech a používáme jejich průměr.

Shrňme intuitivní požadavky na míry ordinální statistické závislosti a doplňme je ještě dalšími, snadno pochopitelnými.

1. Míry ordinální statistické závislosti používáme v situaci, kdy hodnotíme vztah dvou ordinálních znaků, tj. vyhodnocujeme kontingenční tabulky, jejichž vstupy jsou dvě uspořádané kategorizace.

2. Nepředpokládáme spojitý charakter znaků, tj. nepředpokládáme žádné kontinuum, které je kategorizaci podloženo; kategorie nemusí vznikat jako výseky na něm.

3. Rozeznáváme asymetrický a symetrický případ podle toho, zda známe vztah příčinnosti či následnosti, a podle typu závislostního modelu, který používáme (testujeme, zkoumáme).

4. Míra ordinální statistické závislosti α nemá smysl, když jsou všechna data koncentrována v jednom řádku či sloupci.

---

[7]) To ovšem neznamená, že neparametrické techniky jsou pro sociologa neužitečné; situací, kdy jsou aplikovány, je mnoho — analýza malých skupin, analýza souhrnných populačních charakteristik, vyhodnocování různých speciálních situací, komparace atd.

[8]) PRE = Proportional-Reduction-in-Error; je to princip vyjádřený Guttmanem při konstrukci λ koeficientu predikce.

5. $-1 \leqq \alpha \leqq 1$ odráží požadavek normalizace měřicí stupnice, který je motivován především zvykem a vhodností (hranice pro α je ovšem libovolná a mohla by být zvolena jakkoli[9].

6. $\alpha = 1$, když nastává perfektní přímá ordinální závislost ve smyslu operacionalizace dané zvoleným modelem.

7. $\alpha = -1$, když nastává perfektní nepřímá závislost.

8. $\alpha = 0$ znamená absenci ordinální závislosti, neznamená nezávislost; $\alpha = 0$ je ekvivalentní statistické nezávislosti pouze pro tabulky 2×2 (dvouhodnotové znaky).

9. α je závislá na uspořádání řádků i sloupců, je nezávislá na číselném ohodnocení kategorií.

10. α změní znaménko, když změníme pořadí kategorií u jednoho ze znaků na pořadí právě opačné (inverze pořadí hodnot jednoho znaku);

α se nezmění, změníme-li pořadí hodnot u obou znaků v pořadí právě opačná.

11. U symetrických měr nezáleží na tom, která z proměnných je řádková a která sloupcová. U asymetrických měr záměna proměnných vede k jiné hodnotě koeficientu.

Rozlišení asymetrického a symetrického případu je obzvláště důležité při vyhodnocování závislostí ve složitějších schématech, kde se objevují souběžně oba případy a kde potřebujeme pro oba srovnatelné koeficienty.

Důležitý případ vyhodnocení závislosti je sledování konzistence indikátorů určité proměnné. Mezi nimi nejsou přímé vazby, jejich vztah je zprostředkován vlivem proměnné, jejímiž projevy jsou. Jejich vazba je tedy symetrická, je to korelace hodnot, která neodráží kauzální vztahy, ale tu skutečnost, že jde o manifestní projevy téže proměnné, která sama není přímo zjistitelná.

Přestože mezi specialisty panuje názor, že každý vhodný a použitelný koeficient statistické závislosti byl už někdy nalezen a používán, domníváme se, že na poli ordinálních znaků tomu tak není a že zbývá ještě mnoho práce, než budeme v praxi analýzy dat plně uspokojeni.

Přístupy k měření asymetrické ordinální závislosti

Právě u asymetrických vztahů bychom očekávali PRE koeficienty, a proto absenci tohoto nebo jiného vhodného modelu zde tíživě pociťujeme. Známé PRE interpretace mají Goodman-Kruskalovo γ, Kendallovo τ_c a Somersovo d.

Podobné interpretace dále mohou mít koeficienty pro kardinální znaky, které využívají skórování znaků, nebo koeficienty, které byly odvozeny pro symetrický případ. Většina koeficientů je založena na paralele k linearitě, z čehož plyne, že pro symetrický a asymetrický případ nabývají stejné hodnoty. Asymetričnost se pak musí odlišit interpretačně.[10]

Přístupy k měření symetrické ordinální závislosti

Symetrické míry by měly vznikat symetrizací PRE asymetrických koeficientů (stejně jako v případě nominálních znaků, viz [14]). Většina měr je založena na pojmu *shody* a *neshody* (concordance a discordance) dvojic a jejich poměru k různým jmenovatelům: shodou u dvojice jednotek A, B nazýváme případ, kdy relace pro obě jednotky jsou stejné jak vzhledem k X, tak vzhledem k Y; neshodou, jsou-li opačné. Na míry asociace vede jednoduchý princip: čím více je shod mezi všemi dvojicemi, tím vyšší je přímá závislost; čím více je neshod mezi všemi dvo-

[9]) K požadavku normalizace bylo již řečeno mnoho; někteří autoři vyžadují, aby horní hranice (jednička), resp. dolní hranice (minus jedna) byly za všech případů dosažitelné. Obdobně u nominálních měr, byly chí-kvadrátové koeficienty normovány různými způsoby, aby mohly vždy dosáhnout jedné. Takový požadavek je však dosti umělý a není zcela opodstatněn. Je motivován především snahou po vytvoření absolutního kardinálního znaku, který by umožňoval srovnatelnost ruzných tabulek (o různých velikostech souborů a různých rozměrech) a zároveň umožňoval posouzení jedné samostatné tabulky porovnáním číselné hodnoty k oběma krajním hodnotám. Jednička má v našem případě znamenat úplnou ordinální závislost. To ovšem neodpovídá obdélníkovým tabulkám, kde je úplná závislost nedosažitelná. Znak také není absolutní podstatou, ale konvencí, je to většinou znak pomocný. Daleko vhodnějším přístupem je operacionalizace pravděpodobnostním modelem (např. PRE), který umožňuje obsahovou interpretaci výsledného čísla bez ohledu na to, je-li krajní hranice dosažitelná, či ne. Požadavek normalizace ovšem odpovídá zvyklostem a nelze proti němu nic namítat (naopak); poznámka směřuje pouze k problému *dosažitelnosti* krajních hodnot $+1$, -1, které jsou ovšem voleny zcela přirozeně.

[10]) U nominálních závislostí máme jednoduché modely, které rozlišují asymetrický a symetrický vztah. U ordinálních závislostí už, bohužel, takové jednoduché zázemí nemáme. Costner [1] říká, že vzhledem k určitým vlastnostem koeficientu γ není třeba konstruovat asymetrické míry a že pro takové případy můžeme též použít γ. S tím nelze souhlasit, protože, ač je γ koeficient dobrý a má symetrickou i asymetrickou PRE interpretaci, není zdaleka ideální. PRE interpretace koeficientu je prováděna vzhledem k jistému predikčnímu pravidlu, které má určité nedostatky.

jicemi, tím vyšší je nepřímá závislost. Problémem zůstává, co se spojeními v jedné nebo v obou proměnných. Podle způsobů, jimiž je zahrneme do modelu, dostáváme různé koeficienty. Jednotlivé pojmy tohoto přístupu rozebereme v dalším oddílu společně s uvedením výpočetních algoritmů.

Druhý přístup pro symetrické asociace je velice příbuzný. Spočívá v aplikaci obecného koeficientu korelace, který navrhl Daniels ([7], [8]).

$$I' = \frac{\sum\sum a_{ij}\, b_{ij}}{\sqrt{\sum\sum a_{ij}^2 \; \sum\sum b_{ij}^2}} \tag{4}$$

Dvojitá suma značí součet přes všechna i a všechna j; a_{ij}, b_{ij} jsou skóre, která charakterizují vztah dvojice (i-tého a j-tého prvku souboru) vzhledem k X a Y. Tento koeficient byl původně určen pro případ různých pozorování, tj. pro případ, kdy se v datech nevyskytují žádná spojení. Lze z něho odvodit Spearmanův a Kendallův koeficient pořadové korelace i Pearsonův koeficient lineární korelace pro kardiální znaky[11]. I když dáváme přednost měrám, které jsou modelem nějakého obsahu, chceme zdůraznit, že aplikace vzorce (4) je zcela legitimní a že je vhodná především pro případ, kdy je vztah zprostředkován nějakou další proměnnou, tj. kdy jde (obsahově) o zprostředkovanou korelaci a nikoli o vzájemnou závislost (např. výše uvedený případ indikátorů jedné proměnné).

Počet shod a příbuzné pojmy

Nyní zavedeme základní pojmy, z nichž počítáme dnes nejpoužívanější koeficienty γ, d, τ_c, a ukážeme jejich algoritmy. Vzorce koeficientů jsou výpočetně značně komplikované, a proto vedle nich uvádíme i výpočetní postupy tak, jak je v praxi provádíme. Numerické příklady budou uvedeny na konci článku. Značení dodržujeme podle [15].

a) *Počet dvojic.* V celém souboru předpokládáme n jednotek, které tvoří $\binom{n}{2} = \frac{n(n-1)}{2}$ dvojic.

b) *Shody.* Shoda u dvojice (A, B) nastává tehdy, když nastává současně

$$X_A < X_B\,, \quad Y_A < Y_B$$

nebo

$$X_A > X_B\,, \quad Y_A > Y_B$$

Páry, u nichž jsou relace takto souběžné, nazveme *konkordantní*. Jejich počet určíme podle vzorce

$$P = \sum_i \sum_j n_{ij} \sum_{p>i} \sum_{q>j} n_{pq} \tag{5}$$

Výpočet provádíme tak, že ke každému poli tabulky (s četností n_{ij}) přiřadíme plochu polí, která jsou současně pod i-tým řádkem (*níže*) a za j-tým sloupcem (*napravo*) a nalezneme součet všech četností v této ploše (tak dostaneme $\sum_{p>i} \sum_{q>j} n_{pq}$). Tento součet násobíme četností n_{ij}. Pro každé pole tabulky dostaneme jeden takový součin. Nakonec všechny tyto součiny sečteme. Součty, které odpovídají polím posledního sloupce a posledního řádku, jsou nuly.

c) *Neshody.* Neshoda u dvojice jednotek A, B nastává tehdy, když platí současně:

$$X_A > X_B\,, \quad Y_A < Y_B$$

nebo

$$X_A < X_B\,, \quad Y_A > Y_B$$

Páry, u nichž se vyskytují neshody, se nazývají *diskordantní*. Jejich počet Q je dán formulí:

$$Q = \sum_i \sum_j n_{ij} \sum_{p>i} \sum_{q>j} n_{pq} \tag{6}$$

Výpočet provádíme podobně jako u shod. Ke každému poli (i, j) nalezneme všechna pole, která leží *níže* a *nalevo* od tohoto pole. Zjistíme součet četností celé takové plochy a násobíme ho četností n_{ij}. Sečtením všech těchto součinů dostaneme Q. Součty odpovídající polím prvního sloupce a posledního řádku jsou nuly.

d) *Spojení, která nastávají pouze u X.* Je to případ dvojice A, B, u níž platí současně

$$X_A = X_B\,, \quad Y_A > Y_B$$

nebo

$$X_A = X_B\,, \quad Y_A < Y_B$$

[11]) Pearsonovo r dostaneme tak, že dosadíme $a_{ij} = X_i - X_j$, $b_{ij} = Y_i - Y_j$; tím dostaneme v čitateli dvojnásobek kovariance proměnných X a Y a ve jmenovateli geometrický průměr dvojnásobků variancí X a Y (X a Y jsou v tomto případě ovšem číselné proměnné).

Počet X_0 takových dvojic spočteme podle formule:

$$X_0 = \sum_i \sum_j n_{ij} \sum_{q>j} n_{iq} \tag{7}$$

Vlastní výpočet provádíme takto: ke každému poli (i, j) vezmeme všechna pole *napravo ve stejném řádku* a v takto vzniklém pásu sečteme četnosti ($\sum_{q>j} n_{iq}$); získanou hodnotu násobíme číslem n_{ij}; součiny pro všechna pole tabulky sečteme a výsledek je X_0. Součiny příslušné polím v posledním sloupci jsou nuly.

e) *Spojení, která nastávají pouze u Y*, vznikají tehdy, platí-li pro A, B, současně:

$$X_A > X_B\,, \quad Y_A = Y_B$$

nebo

$$X_A < X_B\,, \quad Y_A = Y_B$$

Součet těchto dvojic označíme Y_0, přičemž

$$Y_0 = \sum_i \sum_j n_{ij} \sum_{p>i} n_{pj} \tag{8}$$

Výpočet provádíme takto: ke každému poli (i, j) vezmeme všechna pole směrem *dolů ve stejném sloupci* a v takto vzniklém pásu sečteme četnosti ($\sum_{p>i} n_{pj}$); další postup je stejný jako v *d*).

f) *Současná spojení u X a Y* nastávají tehdy, když platí pro A, B současně:

$$X_A = X_B\,, \quad Y_A = Y_B$$

Jejich počet Z je dán vzorcem:

$$Z = \tfrac{1}{2} \sum_i \sum_j n_{ij}(n_{ij} - 1) \tag{9}$$

g) Počet párů, u kterých *není spojení vzhledem k X*, označíme X_u, přičemž

$$X_u = P + Q + Y_0 \tag{10}$$

$$= \sum_{i<k} \sum n_{i.} n_{k.} \tag{11}$$

$$= \tfrac{1}{2}(n^2 - \sum_i n_{i.}^2) \tag{12}$$

Výpočetně nejjednodušší je vzorec (12), nepotřebujeme-li současně znát Y_0.

h) Počet párů, u nichž *není spojení vzhledem k Y*, vypočteme podle vzorců

$$Y_u = P + Q + X_0 \tag{13}$$

$$= \sum_{j<k} \sum n_{.j} n_{.k} \tag{14}$$

$$= \tfrac{1}{2}(n^2 - \sum_j n_{.j}^2) \tag{15}$$

Nepotřebujeme-li znát X_0, pak výpočetně nejjednodušší je vzorec (15).

Hodnoty a)—h) vystupují u většiny koeficientů. Platí identita[12]

$$P + Q + X_0 + Y_0 + Z = \text{počet všech párů} = \tfrac{1}{2}n(n-1) \tag{16}$$

A) Goodman-Kruskalovo gama

V práci [4] navrhli autoři koeficient, který vychází z jednoduché úvahy, že čím více nastane shod, tím vyšší bude pozitivní (přímá) asociace, a čím více neshod, tím vyšší bude negativní (nepřímá) ordinální asociace. Koeficient je založen jen na těch dvojicích, u nichž nenastává žádné spojení. Všechna spojení jsou jak z úvah, tak z výpočtů eliminována[13]. Koeficient je zadán velice jednoduše jako rozdíl pravděpodobností dvou jevů: 1. u náhodně zvolené dvojice jednotek nastane shoda, 2. u náhodně zvolené dvojice nastane neshoda; obojí za podmínky, že ani u X, ani u Y nenastane spojení.[14] Tedy

$$\gamma = P(\text{shoda/není spojení}) - P(\text{neshoda/není spojení}) \tag{17}$$

$$= \frac{P(\text{shoda}) - P(\text{neshoda})}{1 - P(\text{spojení})} \tag{18}$$

Výpočetní vzorec a zároveň nejjednodušší vyjádření vztahu (17) je

$$\gamma = \frac{P - Q}{P + Q} \tag{19}$$

K výpočtu γ tedy stačí znát počet shod P a počet neshod Q. Základní vlastnosti koeficientu γ:

a) není definován, jsou-li data soustředěna pouze v jednom řádku nebo sloupci;

b) $-1 \leqq \gamma \leqq 1$;

c) $\gamma = 1$, právě když existují kromě spojení samé shody, tj. $Q = 0$. Příklady na

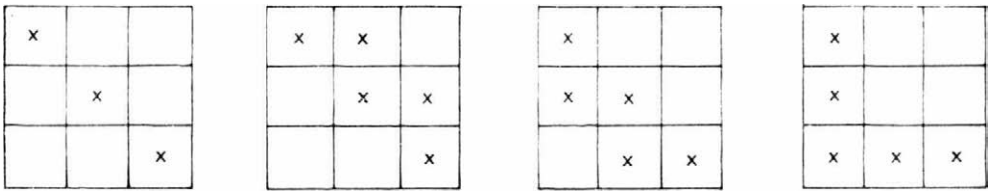
[12]) Vztah (16) používáme při ručních výpočtech *vždy* jako kontrolu.

[13]) Tento fakt je nepříjemným aspektem koeficientu. Spojení jsou důležitou složkou relačního komplexu, který z hlediska uvažovaných znaků vzniká na dané populaci. Jejich zanedbávání je proto ochuzením — projeví se to u vlastnosti koeficientu zvláště u nejednoznačnosti interpretace krajních hodnot, která zahrnuje širokou třídu možných případů.

[14]) To znamená, že náhodný výběr dvojice provádíme jen v tom souboru, u něhož se vyskytují jen shody nebo neshody, všechna spojení jsou eliminována. V takovém podmíněném souboru je $P + Q$ dvojic a rozdíl pravděpodobností proto vede na vzorec (19). Ekvivalentně je to postup: vybereme dvojici a zjistíme, zda u ní existuje spojení (pravděpodobnost takového jevu je P (spojení)). Jestliže se vyskytuje spojení, dvojici neuvažujeme — neodpovídá podmínce a podmíněným pravděpodobnostem v (17). Odtud pak plyne (18).

$\gamma = 1$ (viz [4]) naznačíme v tabulkách obr. 1, kde křížek značí obsazené pole; nezakřížkované pole znamená nulovou četnost.

Obr. 1

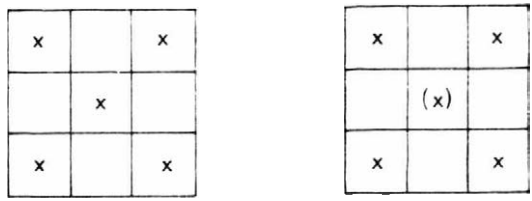


d) $\gamma = -1$, právě když v tabulce jsou kromě spojení samé neshody, tj. $P = 0$. Za příklady mohou sloužit zrcadlově převrácené případy z obr. 1.

e) $\gamma = 0$, právě když počet shod je stejný jako počet neshod ($P = Q$); $\gamma = 0$ při nezávislosti, opačná implikace však obecně neplatí, platí pouze u tabulek 2×2.

Obr. 2 ukazuje případ, kdy je $\gamma = 0$ a přitom nejde o statistickou nezávislost (křížky značí stejné nenulové četnosti).

Obr. 2



f) γ závisí na uspořádání řádků a sloupců;

g) γ změní znaménko, když převrátíme pořadí kategorií u jednoho znaku; γ se nezmění, převrátíme-li pořadí u obou znaků;

h) γ se nezmění při záměně sloupcové proměnné za řádkovou a naopak.

Aplikujeme-li γ na úplné uspořádání jedinců (oba znaky mají tolik hodnot, kolik je jedinců v populaci, jde tedy o prosté ordinální znaky), je γ totožné s Kendallovým koeficientem pořadové korelace τ. Má-li tabulka rozměry 2×2, pak je γ totožné s nulovým koeficientem asociace Q

$$\gamma = Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} \tag{20}$$

Z obrázků a z úvah o shodách a neshodách v kontingenčních tabulkách lze získat cit pro význam pojmu ordinální statistická závislost ve smyslu koeficientu γ. Pro úplnost uvádíme také PRE-interpretaci modelu. (Autoři koeficientu ho nazývají koeficientem založeným na optimální predikci pořadí.) Uvažujme situaci, v níž u náhodně vybrané dvojice A, B chceme určit relaci k Y, tj. zda $Y_A < Y_B$ nebo $Y_A > Y_B$. Vzhledem k náhodnosti výběru se můžeme rozhodnout pro jeden ze vztahů s pravděpodobností $\frac{1}{2}$, nebo (což je totéž) prostě prohlásit $Y_A > Y_B$ (predikční pravidlo I.). Chyba takové predikce (za předpokladu vynechání všech spojení) je

0,5. Při znalosti relace vzhledem k X zvolíme jednoduché pravidlo II pro predikci vztahu v Y (opět vynecháme všechna spojení): je-li $X_A > X_B$, předvídáme $Y_A > Y_B$ při $\gamma > 0$ a $Y_A < Y_B$ při $\gamma < 0$ [15] a obdobně pro opačnou relaci v X. Pravděpodobnost chyby[16] při predikčním pravidle II je $0{,}5 - \left|\frac{\gamma}{2}\right|$, a tedy

$$\mathrm{PRE} = \frac{P(I) - P(II)}{P(I)} = \frac{0{,}5 - \left(0{,}5 - \left|\frac{\gamma}{2}\right|\right)}{0{,}5} = |\gamma|.$$

Z tohoto modelu plyne možnost *asymetrického použití* koeficientu. Symetrický model relativní redukce pravděpodobnosti chyby je stejný jen s tím rozdílem, že při predikci se nejprve s pravděpodobností 0,5 rozhodujeme o směru predikce, tj. o tom, která z proměnných bude nezávislá a která závislá. Protože pojmy shody a neshody jsou oba symetrické ve své podstatě, je koeficient po uvedené symetrizaci zcela stejný jako pro asymetrický případ, a tedy i hodnoty numerické budou pro danou tabulku stejné, ať už má symetrickou či asymetrickou interpretaci. I tento malý příklad ukazuje, jak opatrně je třeba přistupovat k číselným výsledkům. Shodnost asymetrické a symetrické formy koeficientu je paralelní k Pearsonovu koeficientu lineární korelace[17].

Koeficient γ se dobře hodí pro výzkumnou práci, a to i přes nejednoznačnost situací charakterizovaných krajními hodnotami ± 1, které nelze jednoznačně interpretovat v termínech struktury tabulky.

B) Kendallův koeficient pořadové korelace — znaménková korelace

Kendallův koeficient pořadové korelace τ byl původně zaveden pro případ úplného uspořádání objektů vzhledem k oběma proměnným. Vychází (podobně jako γ) ze souběžnosti relací u všech párů souboru. Název „koeficient znaménkové korelace" pochází z definice, která je založena na počtu shod, neshod a Danielsově koeficientu (4), ve kterém jsou zavedeny skóre a, b jako znaménka. Skóre pro X jsou a_{AB}, přičemž

$$a_{AB} = +1, \quad \text{když} \quad X_A > X_B$$
$$a_{AB} = -1, \quad \text{když} \quad X_A < X_B$$

Obdobně jsou definovány skóre b_{AB} pro proměnnou Y. Shoda nastane právě tehdy, když $a_{AB} \cdot b_{AB} = +1$, neshoda právě tehdy, když $a_{AB} \cdot b_{AB} = -1$.

M. Kendall zavedl veličinu

$$2S = \sum_A \sum_B a_{AB} \cdot b_{AB}, \tag{21}$$

což v našem značení není nic jiného než[18]

$$S = P - Q = \text{počet shod} - \text{počet neshod} \tag{22}$$

Koeficient τ vzniká normalizací S na počet všech dvojic

$$\tau = \frac{S}{\frac{n(n-1)}{2}} = \frac{2(P-Q)}{n(n-1)} \tag{23}$$

Pro *kontingenční tabulky* tento koeficient modifikujeme tak, že uvažujeme spojení u každé dvojice, která padne do jedné kategorie. Máme tři možnosti:

a) ponecháme původní výraz s vědomím, že vzhledem ke spojením nemůže koeficient nabývat hodnoty ± 1; tak dostaneme koeficient

$$\tau_a = \frac{P - Q}{\frac{1}{2}n(n-1)} \tag{24}$$

Pravděpodobnostní vyjádření (v pojmech modelu náhodného výběru dvojic jako u γ):

[15]) Je-li $\gamma = 0$, pak predikci pro relaci I provádíme s pravděpodobností 0,5 pro jednu možnost a se stejnou pravděpodobností pro druhou možnost.

[16]) Pro kladné i záporné gama nalezneme: $|\gamma| = P$ (správná predikce) $- P$ (chybná predikce)|, z čehož plyne, že $|\gamma| = |1 - 2P$ (chybná predikce)| a dále, že $|P$ (chybná predikce) $| = 0{,}5 - \left|\frac{\gamma}{2}\right|$.

[17]) r vzniká jako speciální případ koeficientu (4); současně je to parametr dvourozměrného normálního pravděpodobnostního modelu, ve kterém také odpovídá vzájemné vazbě proměnných; zároveň je r^2 relativním úbytkem chyby při predikci jedné proměnné pomocí druhé proměnné za použití lineárního modelu (asymetrická PRE interpretace).

[18]) Ve (21) vystupuje dvojnásobek hodnoty S proto, že při sčítání počítáme každou dvojici dvakrát (jako A, B) a (B, A), přičemž v obou případech nastane buď shoda, nebo neshoda.

P (shoda) — P (neshoda). Zde, na rozdíl od podobného výrazu pro γ, vystupují nepodmíněné pravděpodobnosti. Proto ani jedna z pravděpodobností, a tedy ani koeficient $|\tau_a|$ nemohou být rovny $+1$, existují-li spojení.

b) Upravíme koeficient dosazením modifikovaných skóre a, b do Danielsova obecného vzorce (4). K výše uvedené konvenci ještě přidáme

$$a_{AB} = 0\,, \quad \text{když} \quad X_A = X_B$$
$$b_{AB} = 0\,, \quad \text{když} \quad Y_A = Y_B$$

Vzniklý koeficient se v literatuře značí jako

$$\tau_b = \frac{P - Q}{\frac{1}{2}\sqrt{(n^2 - \sum n_{i.}^2)\,(n^2 - \sum n_{.j}^2)}} \tag{25}$$

$$= \frac{P - Q}{\sqrt{(P + Q + X_0)\,(P + Q + Y_0)}} \tag{26}$$

$$= \frac{P - Q}{\sqrt{Y_u \,.\, X_u}} \tag{27}$$

Koeficient τ_b dosahuje hodnot ± 1 pouze v případech čtvercových matic [19], a to tehdy, když data jsou soustředěna na diagonále jdoucí od levého horního rohu k pravému dolnímu rohu ($\tau_b = +1$) nebo na diagonále jdoucí od pravého horního rohu k levému dolnímu ($\tau_b = -1$). Z toho je vidět, že se τ_b hodí pro zkoumání reliability ordinálních znaků při metodě opakovaného testování (test — retest) a při zkoumání shod výpovědí s objektivními stavy.

c) Stuart [18] navrhl modifikaci koeficientu τ tak, aby mohl dosahovat krajních hodnot ± 1 při všech tvarech tabulky. Koeficient se značí τ_c a je dán výrazem

$$\tau_c = \frac{2m(P - Q)}{n^2(m - 1)} \tag{28}$$

$$= \frac{n - 1}{n}\,\frac{m}{m - 1}\,\tau_a$$

kde $m = \min(r, s)$. Ve výzkumné práci se τ_c téměř nevyskytuje.

Ze tří variant Kendallova koeficientu používáme nejvíce τ_b. Kromě uvedené specifikace Danielsova Γ má koeficient PRE interpretaci, o které referuje Wilson [19].

Vlastnosti koeficientu τ_b:

a) není definován, když data jsou soustředěna v jednom řádku nebo v jednom sloupci tabulky;

b) $-1 \leqq \tau_b \leqq 1$, přičemž krajních hodnot dosahuje jen u čtvercových tabulek (po vynechání prázdných sloupců a řádků);

c) $\tau_b = 1$ právě když jsou všechna data v hlavní diagonále přímé závislosti;

d) $\tau_b = -1$ právě když jsou všechna data v hlavní diagonále nepřímé závislosti;

e) $\tau_b = 0$ právě když $P = Q$; nejde tedy o nezávislost, ale o ordinální nekorelovanost ve smyslu stejného počtu shod a neshod; $\tau_b = 0$ je ekvivalentní se statistickou nezávislostí pouze u čtyřpolních tabulek;

f) τ_b je závislé na uspořádání řádků a sloupců;

g) τ_b změní znaménko, když změníme pořadí kategorií u jednoho znaku v opačné; τ_b se nezmění, když provedeme tuto změnu u obou proměnných.

Pro tabulky 2×2 je koeficient roven korelačnímu koeficientu r pro čtyřpolní tabulku, u níž jsou vyšší hodnoty znaků kvantifikovány jedničkou a nižší nulou (indikátorové znaky vyšších hodnot příslušných vlastností)[20].

$$\tau_b = \frac{n_{11}n_{22} - n_{21}n_{12}}{\sqrt{n_{1.}\,n_{2.}\,n_{.1}\,n_{.2}}} \tag{29}$$

$$= \sqrt{\frac{\chi^2}{n}} \tag{30}$$

Koeficient τ je odvozen pro symetrickou korelaci proměnných. Výše uvedený odkaz [19] uvádí asymetrickou PRE interpretaci, která může být ovšem snadno symetrizována; výsledek je týž.

C) Somersovo d

Somers [15] vyšel z analogie s lineárním modelem pro kardinální znaky, v němž platí

$$r^2_{XY} = b_{XY} \,.\, b_{YX}$$

(korelační koeficient je geometrickým průměrem obou regresních koeficientů), a poku-

[19]) To plyne ze vzorce (4) a ze známé Cauchyovy nerovnosti: rovnost jmenovatele a čitatele (v absolutní hodnotě) nastane právě tehdy, když $a_{ij} = k \,.\, b_{ij}$, (k může být buď $+1$ nebo -1), a z toho plyne, že všechna data musí být soustředěna na jedné z hlavních diagonál (ovšem po vynechání všech neobsazených sloupců a řádků.)

[20]) Kvantifikace nulou a jedničkou je nepodstatná, protože koeficient korelace je invariantní k lineárním transformacím s kladným součinem směrnic a každá dvojice stejně orientovaných hodnot je převoditelná na čísla 0, 1; u dvouhodnotových znaků je tedy z technického hlediska intervalový znak totéž co ordinální znak.

sil se o asymetrizaci založenou na stejném principu. Jelikož platí:

$$\tau^2_b = \frac{(P-Q)^2}{X_u \cdot Y_u} = \frac{P-Q}{X_u} \cdot \frac{P-Q}{Y_u},$$

můžeme vybrat souběžně k regresním koeficientům oba činitele jako míry asymetrické ordinální statistické závislosti; označme je

$$d_{Y/X} = \frac{P-Q}{X_u} \tag{31}$$

$$= \frac{P-Q}{P+Q+Y_0} \tag{32}$$

$$= \frac{2(P-Q)}{n^2 - \sum n_{i.}^2} \tag{33}$$

$$d_{X/Y} = \frac{P-Q}{Y_u} \tag{34}$$

$$= \frac{P-Q}{P+Q+X_0} \tag{35}$$

$$= \frac{2(P-Q)}{n^2 - \sum n_{j}^2} \tag{36}$$

$d_{Y/X}$ je mírou vztahu $X \rightarrow Y$, v němž X je nezávislá a Y závislá proměnná; u $d_{X/Y}$ je tomu naopak.

Koeficient můžeme též zavést pravděpodobnostně (pravděpodobnosti se vztahují k náhodně zvolené dvojici jednotek):

$d_{Y/X}$ = P (shoda/u X nenastalo spojení) — — P (neshoda/u X nenastalo spojení).

Bez ohledu na PRE-interpretaci, pravděpodobnostní zavedení je rozumné — je to počet shod u těch dvojic, u nichž můžeme predikovat pořadí v Y, tj. u nichž není spojení v nezávislé proměnné X.

Vlastnosti koeficientu:

a) není definován, jsou-li data umístěna pouze v jednom řádku nebo v jednom sloupci tabulky;

b) $-1 \leqq d_{Y/X} \leqq 1$;

c) $d_{Y/X} = 1$ právě když $Q = 0$ (nevyskytují se neshody) a zároveň $X_0 = 0$ (neexistují dvojice spojené v Y a nespojené v X); to může však nastat jen tehdy, jsou-li data soustředěna v hlavní diagonále přímé závislosti (po vynechání neobsazených řádků a sloupců);

d) $d_{Y/X} = -1$ právě když $P = 0$ a $X_0 = 0$, což může nastat jen tehdy, když jsou data soustředěna v hlavní diagonále nepřímé závislosti;

e) $d_{Y/X} = 0$ právě když $P = Q$; tento případ reprezentuje ordinální nezávislost, nikoli statistickou; s tou splývá pouze u čtyřpolních tabulek;

f) $d_{Y/X}$ závisí na uspořádání řádků a sloupců;

g) při změně pořadí kategorií u jedné proměnné v opačné se změní znaménko koeficientu; při této změně u obou znaků se koeficient nezmění;

h) koeficient se změní při záměně sloupcové proměnné za řádkovou a naopak (asymetričnost);

i) oba koeficienty $d_{Y/X}$ a $d_{X/Y}$ mají stejné znaménko.

U tabulek 2×2 je $d_{Y/X}$ totožné s rozdílem relativních četností $R_{Y/X}$, který je mírou asociace pro čtyřpolní tabulky v asymetrických případech.

$$d_{Y/X} = R_{Y/X} = \frac{n_{11}}{n_{1.}} - \frac{n_{21}}{n_{2.}} = \tag{37}$$

$$= \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{1.}n_{2.}} \tag{38}$$

Somers se v práci [15] snaží ukázat, že i ve větších tabulkách odpovídá $d_{Y/X}$ rozdílům četností.

Somersův koeficient má velmi dobré vlastnosti, má jasně definované polohy (± 1), má pravděpodobnostní zázemí, které se zdá být rozumné. Existuje u něj analogie k lineárnímu modelu, u tabulek 2×2 je totožný s nejjednodušší a nejpoužívanější mírou asymetrického vztahu $R_{Y/X}$. Jeho aplikaci je možno v asymetrických případech doporučit. PRE interpretaci lze nalézt v [17].

D) Symetrizovaný Somersův koeficient d

Koeficienty $d_{Y/X}$ a $d_{X/Y}$ vznikly asymetrizací koeficientu τ_b, který lze proto považovat za symetrickou variantu Somersova koeficientu.

V [11] uvádějí autoři programu SPSS jinou symetrizaci. Je provedena stejně jako v PRE modelech: randomizací směru predikce. Vzniklá míra má velmi dobré vlastnosti a rozumný obsah numerických hodnot; s koeficientem můžeme dobře pracovat.

$$d = \frac{P-Q}{\frac{1}{2}(X_u + Y_u)} \tag{39}$$

$$= \frac{2[P-Q]}{\frac{1}{2}[(n^2 - \sum n_{i.}^2) + (n^2 - \sum n_{j}^2)]} \tag{40}$$

Vlastnosti jsou tytéž jako u asymetrického koeficientu, až na to, že d se nezmění, zaměníme-li sloupcovou a řádkovou proměnnou.

Pro čtyřpolní tabulky dostáváme jednoduchý tvar

$$(41)_2' \qquad d = \frac{n_{11}n_{22} - n_{12}n_{21}}{\frac{1}{2}(n_{1.}n_{2.} + n_{.1}n_{.2})}$$

Zatímco τ_b je geometrický průměr veličin $d_{Y/X}$ a $d_{X/Y}$, je d jejich harmonický průměr.

E) Spearmanův koeficient pořadové korelace pro kontingenční tabulky

Spearmanův koeficient ϱ vzniká tak, že do obecného Danielsova koeficientu (4) dosadíme skóre, která jsou definována jako rozdíly pořadí u jednotlivých proměnných:

a_{AB} = rozdíl pořadí dvou objektů A a B vzhledem k uspořádání X,

b_{AB} = rozdíl pořadí A a B vzhledem k uspořádání Y.

Tento předpis vede na obvyklý vzorec pro ϱ v situaci prostých ordinálních znaků. Pro kontingenční tabulky se koeficient ϱ adaptuje na spojená pořadí, vznikající uvnitř kategorií znaků a značí se ϱ_b. Vzorec opět plyne z Danielsova koeficientu (4). Výpočet je založen na jiných charakteristikách než předcházející koeficienty. Uvádíme ho v postupných krocích.

1) Určení pořadí (spojených) pro jednotlivé jednotky. Pro znak X máme v za sebou jdoucích kategoriích postupně četnosti $n_{1.}, n_{2.}, n_{3.}, \ldots, n_{r.}$. Kdyby byly všechny jednotky rozlišitelné, měly by pořadí $1, 2, 3, \ldots, n$. Nyní však je spojeno $n_{1.}$ jednotek, které by stály na místech $1, 2, 3, \ldots, n_{1.}$. Proto jim všem přiřadíme průměrné pořadí $\frac{n_{1.}+1}{2}$. Dalších $n_{2.}$ jednotek by mělo při úplném uspořádání pořadí $n_{1.}+1, n_{1.}+2, \ldots, n_{1.}+n_{2.}$, nyní jim opět přiřadíme průměrné pořadí $\frac{2n_{1.}+n_{2.}+1}{2}$. Stejně postupujeme dále. Pro znak X dostaneme tedy pro jednotlivé kategorie průměrná pořadí

$$\frac{n_{1.}+1}{2}, \frac{2n_{1.}+n_{2.}+1}{2}, \frac{2n_{1.}+2n_{2.}+n_{3.}+1}{2}, \ldots, \frac{2n_{1.}+2n_{2.}+\ldots+2n_{r-1.}+n_{r.}+1}{2}.$$

Obdobně pro znak Y dostaneme pro jednotlivé kategorie průměrná pořadí

$$\frac{n_{.1}+1}{2}, \frac{2n_{.1}+n_{.2}+1}{2}, \frac{2n_{.1}+2n_{.2}+n_{.3}+1}{2}, \ldots, \frac{2n_{.1}+2n_{.2}+\ldots+2n_{.s-1}+n_{.s}+1}{2}.$$

2) Výpočet veličiny $S(d^2) = \sum_A (X_A - Y_A)^2$; X_A je pořadí jednotky A v proměnné X a Y_A je pořadí v Y (jde o pořadí zjištěná v kroku 1). Pro každé další pole tabulky (i, j) zjistíme rozdíl skóre $X(i) - Y(j)$, umocníme ho na druhou a vynásobíme četností pole. Tak dostaneme:

$$(42) \qquad S(d^2) = \sum\sum n_{ij}[X(i) - Y(j)]^2 ;$$

$X(i)$ je pořadí přiřazené (v kroku 1) i-tému řádku tabulky,

$Y(j)$ je pořadí přiřazené (v kroku 1) j-tému sloupci tabulky.

3) Výpočet veličin T, U

$$(43) \qquad T = \frac{1}{12}\sum_i (n_{i.}^3 - n_{i.})$$

$$(44) \qquad U = \frac{1}{12}\sum_j (n_{.j}^3 - n_{.j})$$

4) Dosazení do vzorce pro ϱ_b

$$(45) \qquad \varrho_b = \frac{\frac{1}{6}(n^3 - n) - S(d^2) - T - U}{\sqrt{[\frac{1}{6}(n^3-n)-2T] \cdot [\frac{1}{6}(n^3-n)-2U]}}$$

Po určení průměrných pořadí pro jednotlivé kategorie X_i, Y_j, můžeme hodnoty průměrných pořadí $X(i), Y(j)$ považovat za skóre a dosadit je do vzorce pro výpočet Pearsonova lineárního korelačního koeficientu. Výsledek bude stejný jako výsledek získaný výše uvedeným algoritmem (42)—(45).

V literatuře ([10]) lze najít dosti složitý pravděpodobnostní model pro koeficient ϱ, jehož interpretace je však pro praktické použití značně obtížná.

Koeficient ϱ_b je symetrický a výrazně vyjadřuje korelovanost hodnot pořadí (člen $S(d^2)$). Můžeme jej interpretovat také jako lineární korelaci mezi pořadími. Proto může mít i asymetrickou interpretaci v modelu, v němž hledáme lineární predikci pořadí znaku Y pomocí pořadí znaku X u daného objektu. Pak má i jednoduchou PRE interpretaci: predikci pořadí pomocí přímky.

Vlastnosti:

a) ϱ_b není definován, jsou-li hodnoty soustředěny v jednom řádku nebo v jednom sloupci tabulky;

b) $-1 \leqq \varrho_b \leqq 1$; krajní hodnoty dosahuje pouze v případě, že pro všechny dvojice

A, B platí $X_A - X_B = k(Y_A - Y_B)$, (k je konstanta), což znamená, že krajních hodnot lze docílit pouze ve čtvercové tabulce (po vynechání prázdných sloupců a řádků), a to jen tehdy, když jsou data rozložena v jedné z hlavních diagonál;

c) $\varrho_b = 1$ právě když jsou data rozložena v diagonále jdoucí od levého horního do pravého dolního rohu čtvercové tabulky;

d) $\varrho_b = -1$ právě když jsou data rozložena v diagonále jdoucí od pravého horního k levému dolnímu rohu čtvercové tabulky;

e) $\varrho_b = 0$ značí nekorelovanost, situace v tabulce není ovšem jednoznačná;

f) ϱ_b je závislé na uspořádání řádků i sloupců a současně je závislé na rozložení marginálií, které určují skóre spojeného pořadí;

g) při změně pořadí kategorií u jedné z obou proměnných se změní znaménko u ϱ_b; při změně pořadí u obou se hodnota ϱ_b nemění;

h) koeficient se nezmění při záměně sloupcové a řádkové proměnné.

U čtyřpolních tabulek je ϱ_b stejný jako koeficient korelace (počítaný z (0, 1)-veličin)

$$\varrho_b = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}} \tag{46}$$

F) Skórovací metody

Nejjednodušší přístup k pořadovým korelacím je skórovat kategorie a pro daná skóre spočítat lineární korelační koeficient

$$r = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var} X \,.\, \operatorname{var} Y}}. \tag{47}$$

Za skórovací postup mohl být považován i Spearmanův koeficient pořadové korelace. Skórovací metody mohou být odvozeny z Danielsova vzorce (4) při vhodné volbě skóre.

Přirozené je volit za skóre čísla pořadí kategorií, tzn. přiřadit kategorii X_i číslo i a kategorii Y_j číslo j. Tak dostáváme pseudočíselné proměnné, na které aplikujeme lineární koeficient r (resp. Danielsův koeficient Γ). Tento postup je zcela legitimní, pokud nezapomeneme při interpretaci na naše východisko a na to, co korelujeme. K tomuto přístupu existuje predikční model s PRE interpretací: předpovídáme *číslo kategorie* proměnné Y bez znalosti čísla kategorie X a poté se znalostí čísla kategorie X. Jako predikční pravidlo slouží lineární regresní funkce, která charakterizuje vztah čísel kategorií. Uvedená metoda se aplikuje dosti často. Poskytuje jak symetrické, tak asymetrické interpretace koeficientu. Pro výpočty přípustně transformujeme skóre tak, abychom si ulehčili výpočetní námahu, zkrátili numerické kroky a snížili počet chyb. Obvykle volíme skóre nula u nejčetnější kategorie znaku nebo u prostřední a pak volíme hodnoty $1, 2, \ldots$ směrem napravo a $-1, -2, \ldots$ směrem nalevo.

Můžeme-li uvažovat nějaký typ rozložení na kontinuu podloženém kategorizací (např. normální standardizované rozložení $N(0, 1)$), zavádíme skóre odpovídající četnostním hodnotám a tomuto rozložení. I když je tato metoda uváděna u problematiky uspořádaných kategorizací, není vhodná pro měření ordinální statistické závislosti, neboť její výsledky nejsou invariantní k monotónním transformacím původního kontinua. Je to tedy metoda zpracování kardinálních znaků.

Málo častý je případ hledání skóre tak, aby korelace mezi proměnnými X, Y byla maximální. Tato maximální hodnota se pak používá jako charakteristika vazby (viz [8], kap. 33). Pro nás tento postup vycházející z kanonické analýzy nemá většinou praktické použití, i když lze pro něj nalézt velmi dobré metodologické důvody.

Numerické příklady[21]

Příklad **1**

Tento vztah může být v určitých kontextech považován za asymetrický a v určitých kontextech za symetrický. Proto zde předvedeme výpočet obou typů měr. Víme, že většinu koeficientů lze spočíst z několika základních hodnot; nejprve určíme proto pomocné veličiny:

$$\begin{aligned} P &= 144(600 + 363 + 136 + 166) + \\ &+ 550(363 + 166) + 114(136 + 166) + \\ &+ 600 \,.\, 166 = \mathbf{607\,138} \quad \text{(aplikace vzorce (5))} \end{aligned}$$

[21]) Za příklady ze sociologické praxe děkujeme H. Ječábkovi, který vyhledal a poskytl data tab. 1, a J. Linhartovi, který vyhledal a poskytl data tab. 2 a 3 a velmi aktivně podněcoval autory k napsání této stati. Původní data byla z důvodů zjednodušení upravena sjednocením kategorií. Příklad 1 pochází z výzkumu dlouhodobé spotřeby obyvatel provedeného Výzkumným ústavem obchodu (J. Bérová, H. Ječábek), příklady 2 a 3 pocházejí z výzkumu postojů obyvatel historického jádra města Tábora (J. Linhart, M. Matějů).

Tabulka 1: Vztah příjmu přednosty domácnosti a stupně vybavenosti domácnosti

Příjem přednosty domácnosti (X)	Stupeň vybavenosti domácnosti (Y)			
	nižší	standard	vyšší	Celkem
do 1500	144	550	89	783
1501—2500	114	600	363	1077
2501 a víc	15	136	166	317
Celkem	273	1286	618	2177

$$Q = 550(114 + 15) + 89(114 + 600 + 15 + 136) + 600 \cdot 15 + 363(15 + 136) = \mathbf{211\,748}$$
(aplikace vzorce (6))

$$X_0 = 144(550 + 89) + 550 \cdot 89 + 114(600 + 363) + 600 \cdot 363 + 15(136 + 166) + 136 \cdot 166 = \mathbf{495\,654}$$
(aplikace vzorce (7))

$$Y_0 = 144(114 + 15) + 114 \cdot 15 + 550(600 + 136) + 600 \cdot 136 + 89(363 + 166) + 363 \cdot 166 = \mathbf{614\,025}$$
(aplikace (8))

$$Z = \tfrac{1}{2}(144 \cdot 143 + 550 \cdot 549 + 89 \cdot 88 + 114 \cdot 113 + 600 \cdot 599 + 363 \cdot 362 + 15 \cdot 14 + 136 \cdot 135 + 166 \cdot 165) = \mathbf{440\,011}$$
(aplikace (9))

$$X_u = 607\,138 + 211\,748 + 614\,025 = \mathbf{1\,432\,911}$$
(aplikace (10))

nebo

$$X_u = \tfrac{1}{2}(2177^2 - 783^2 - 1077^2 - 317^2) = \mathbf{1\,432\,911}$$
(aplikace (12))

$$Y_u = 607\,138 + 211\,748 + 495\,654 = \mathbf{1\,314\,540}$$
(aplikace (13))

nebo

$$Y_u = \tfrac{1}{2}(2177^2 - 273^2 - 1286^2 - 618^2) = \mathbf{1\,314\,540}$$
(aplikace (15))

Nakonec vždy provádíme numerickou kontrolu podle (16):

$$P + Q + X_0 + Y_0 + Z = 607\,138 + 211\,748 + 495\,654 + 614\,025 + 440\,011 = \mathbf{2\,368\,576}$$

$$\tfrac{1}{2}n(n-1) = \tfrac{1}{2}2177 \cdot 2176 = \mathbf{2\,368\,576}.$$

Pomocí těchto hodnot spočteme

$$\gamma = \frac{P - Q}{P + Q} = \frac{607\,138 - 211\,748}{607\,138 + 211\,748} = \frac{395\,390}{818\,886} = \mathbf{0{,}48}$$
(viz (19))

$$\tau_a = \frac{P - Q}{\frac{1}{2}n(n-1)} = \frac{395\,390}{2\,368\,576} = \mathbf{0{,}11}$$
(viz (24))

$$\tau_b = \frac{P - Q}{\sqrt{Y_u X_u}} = \frac{395\,390}{(1\,314\,540 \cdot 1\,432\,911)^{\frac{1}{2}}} = \mathbf{0{,}29}$$
(viz (27))

$$\tau_c = \frac{2m(P - Q)}{n^2(m-1)} = \frac{2 \cdot 3 \cdot 395\,390}{2177^2 \cdot 2} = \mathbf{0{,}25}$$
(viz (28))

$$d_{Y/X} = \frac{P - Q}{X_u} = \frac{395\,390}{1\,432\,911} = \mathbf{0{,}28}$$
(viz 31))

$$d = \frac{P - Q}{\frac{1}{2}(X_u + Y_u)} = \frac{2 \cdot 395\,390}{1\,432\,911 + 1\,314\,540} = \mathbf{0{,}29}$$
(viz 39))

Pro výpočet Spearmanova koeficientu potřebujeme nejprve znát skóre pro jednotlivé kategorie obou znaků. Ty vznikají jako průměrná pořadí v těchto kategoriích:

$$X(1) = \frac{783 + 1}{2} = 392$$

$$X(2) = \frac{2 \cdot 783 + 1077 + 1}{2} = 1322$$

$$X(3) = \frac{2 \cdot 783 + 2 \cdot 1077 + 317 + 1}{2} = 2019$$

$$Y(1) = \frac{273 + 1}{2} = 137$$

$$Y(2) = \frac{2 \cdot 273 + 1286 + 1}{2} = 916{,}5$$

$$Y(3) = \frac{2 \cdot 273 + 21\,286 + 618 + 1}{2} = 1868{,}5$$

Nyní se můžeme rozhodnout pro dva postupy — buď aplikujeme vzorec (45), nebo

Tabulka 2: Vztah vzdělání přednosty domácnosti a spokojenosti s bydlením u části výběrového souboru z historické části města Tábora

Stupeň vzdělání	Spokojenost s bydlením					
	vysoká				nízká	Celkem
	1	2	3	4	5	
Neúplné základní	14	16	6	2	3	41
Základní	1	16	23	19	18	77
Vyučen	1	14	26	7	4	52
Nižší střední	0	2	3	9	0	14
Maturita	0	1	6	3	1	11
Vysokoškolské	1	1	0	0	3	5
Celkem	17	50	64	40	29	200

spočítáme lineární koeficient r pro takto získané skóre. Zde naznačíme aplikaci vzorce (45). Snadno nahlédneme, že Spearmanovo ϱ_b se nehodí nikde tam, kde jsme závislí na ručních výpočtech, neboť postup je velice zdlouhavý. Výraz $S(d^2)$ spočteme podle vzorce (42):

$$S(d^2) = \sum\sum n_{ij}(X(i) - Y(j)^2 =$$
$$= (392 - 137)^2 \,.\, 144 + (392 - 916{,}5)^2 \,.$$
$$\,.\, 550 + (392 - 1868{,}5)^2 \,.\, 89 + \ldots +$$
$$+ (2019 - 916{,}5)^2 \,.\, 136 +$$
$$+ (2019 - 1868{,}5)^2 \,.\, 166 =$$
$$= \mathbf{944\,045\,236}.$$

Dále aplikujeme vzorce (43), (44), pro výpočet T a U:

$$T = \tfrac{1}{12} \sum n_i.(n_i^2 - 1) =$$
$$= \tfrac{1}{12} [783\,(783^2 - 1) + 1077\,(1077^2 - 1) +$$
$$+ 317(317^2 - 1)] = \mathbf{146\,762\,088}.$$

Obdobně u $U = \mathbf{198\,596\,244}$.

Dosazením do (45) dostaneme $\varrho_b = \mathbf{0{,}31}$.

Poslední možnost je aplikace vzorce lineárního Pearsonova koeficientu korelace pro kategorie skórované pořadími, tj.

$$X(1) = 1,\ X(2) = 2,\ X(3) = 3,$$
$$Y(1) = 1,\ Y(2) = 2,\ Y(3) = 3.$$

Toto heuristické skórování ukazuje smysl nasazení míry. Pro výpočet je pohodlnější skórování

$$X(1) = Y(1) = -1,\ X(2) = Y(2) = 0,$$
$$X(3) = Y(3) = 1.$$

Numerický výsledek je ovšem stejný: $r^2 = 0{,}0936$, $r = \mathbf{0{,}31}$.

Srovnávat hodnoty jednotlivých koeficientů nemá smysl, protože každý z nich je jinak definován, měří proto poněkud jiné aspekty rozložení v tabulce a má svou vlastní významovou škálu.

Příklad **2**

Tabulka 2 reprezentuje zajímavý případ, v němž se projevuje „křížení" dvou závislostí, které jdou po hlavních diagonálách. Diagonála přímé závislosti je však slaběji početně obsazena, proto koeficienty ordinální závislosti budou především odrážet závislost nepřímou, ale budou oslabeny závislostí opačného směru. (Poznamenejme, že přímá závislost mezi spokojeností a vzděláním je ovšem nepřímou statistickou závislostí v tomto uspořádání tabulky.) Vztah považujeme za asymetrický. Tabulka ukazuje, že zkoumaná populace 200 osob může být rozdělena na dvě části (pravděpodobně různě velké), v nichž se projevují opačné závislosti. (Třídění dat vyššího stupně by tuto hypotézu potvrdilo.) Numerické zpracování dává tyto výsledky:

$P = 7353$, $Q = 4187$, $X_0 = 3797$, $Y_0 = 3132$,
$Z = 1431$, $X_u = 14\,672$, $Y_u = 15\,337$;
$\gamma = 0{,}27$; $\tau_b = 0{,}21$; $d_{Y/X} = 0{,}22$.

Pro zajímavost uvádíme asymetrické nominální míry statistické závislosti:

$\tau_{Y/X} = 0{,}10$ (Wallisův koeficient proporcionální predikce)
$\xi_{Y/X} = 0{,}15$ (Informační míra)
(vzorce viz [14]).

Tabulka 3: Vztah současného stavu placení nájemného a ochoty zaplatit za lepší byt u výběrového souboru domácností z asanační oblasti města Tábora

V současné době platí nájemné	Je ochoten platit nájemné za předpokladu lepšího bytu							
	−50	−100	−150	−200	−250	−300	více	Celkem
do 50	10	16	15	5	1	1	2	50
51−100	2	20	18	17	4	4	5	70
101−150	1	2	31	10	11	4	2	61
151−200	1	1	8	16	15	10	9	60
201−250	1	0	2	4	7	10	5	29
251−300	0	1	0	2	8	4	4	19
301 a více	0	0	0	0	0	0	1	1
Celkem	15	40	74	54	46	33	28	290

Příklad 3

Oba vstupy jsou kardinální, a proto především uvažujeme o použití lineárního korelačního koeficientu r. Vztah v tabulce však lineárně nevypadá a v takových případech je lépe volit míry ordinální asociace, obzvláště u tabulek větších rozměrů. Stupeň ordinální asociace libovolného typu bude charakterizován koeficientem γ. Současně s ordinální závislostí nás však zajímá v tomto speciálním případě stupeň neshody, tj. jaká je shoda (či neshoda) mezi stavem a ochotou. Pro tento účel je vhodný τ_b nebo $d_{Y/X}$ (neboť jde o asymetrický vztah). Všimněme si, že tento koeficient může být vzat za jeden z indikátorů spokojenosti šetřené populace s bydlením; (čím vyšší hodnota τ_b či $d_{Y/X}$, tím vyšší shoda stavu a ochoty a tím nižší potřeba měnit i za jisté oběti).

Přestože uvádíme tuto tabulku, domníváme se, že pro interpretaci závislosti v ní a po podrobnější inferenci z ní jsou míry závislosti málo zajímavé, protože tento typ tabulek (a podobně např. mobilitní tabulky, které mívají často podobnou strukturu) vyžaduje speciální typ analýzy (zejména studium posunutí marginálních rozložení a převodní mechanismy).

Numerické výpočty:

$P = 22\,539$, $Q = 6405$, $X_0 = 6003$, $Y_0 = 5144$, $Z = 1814$, $X_u = 34\,088$, $Y_u = 34\,947$; $\gamma = 0{,}56$; $\tau_b = 0{,}47$; $d_{Y/X} = 0{,}47$, $r = 0{,}54$.

Závěr

Snažili jsme se uvést pokud možno úplný přehled užitečných a dnes více nebo méně používaných měr ordinální statistické závislosti dvou znaků. Zatímco nominální závislost byla charakterizována spojením kategorií, ordinální závislost je charakterizována souběžností relací u dvojic v souboru.

Výběr jedné z měr je záležitostí výzkumného kontextu. Každá míra odpovídá poněkud jinému typu problému, a tudíž její nasazení je možné pouze vzhledem k příslušnému výzkumnému cíli.

Používání mnoha měr je nevhodné, neboť to snižuje komparabilitu dat různých výzkumů, ztěžuje to práci v sekundární analýze apod. Doporučit konečný výběr je však velmi obtížné. Jemné rozdíly mezi koeficienty jsou těžko postřehnutelné a závisí na citu, zkušenosti, někdy na tradici či módnosti, který koeficient je ve výzkumu použit (poslední dva důvody by však měly být pokud možno eliminovány).

Naše kritéria pro výběr jsou:

1) jednoduchá a smysluplná interpretace založená na vhodném pravděpodobnostním principu,
2) jednoduchý a rychlý výpočet.

Druhý aspekt je důležitý i v době, kdy máme k dispozici počítače. Nezřídka se stává, že potřebujeme dopočítat nějaké hodnoty pro další tabulky nebo části tabulek pro detailní analýzu. V naší praxi používáme většinou skórovacích postupů (skórování číslem kategorie), a to především pro nedostupnost vhodných a levných programů s širším výběrem měr. Tento postup je pouze náhražkou v situaci, kdy nelze použít koeficienty adekvátnější. V žádném případě jej nedopo-

ručujeme čtenářům; neodpovídá ani jednomu z obou požadavků právě vyjádřených. Dále pak Goodman-Kruskalovo gama. V symetrické a asymetrické interpretaci se velmi nadějné zdá Somersovo *d*. Jeho použití je omezeno tím, že nebývá běžnou součástí programů na počítače. Dáváme přednost γ před τ_b hlavně tam, kde široká třída krajních případů dobře odpovídá pojmu ordinální závislosti. Zároveň tu přistupuje technický aspekt — z několika zkušeností se zdá, že gama má menší směrodatnou odchylku. Ke zkoumání reliability metodou „test-retest" bychom však spíše měli používat τ_b než γ. Spearmanův koeficient může být použit v nejrůznějších kontextech, nejlépe se však hodí na situace, kde jde o skutečné spojení dat ve smyslu shluků na kontinuu. Nepříjemný je jeho dost zdlouhavý výpočet.

V rukou zkušených výzkumníků, kteří nikdy neabsolutizují čísla a hodnoty koeficientů, je i při neoptimální volbě koeficientů pouze malé nebezpečí, že se dopustí chyby. Znovu zdůrazňujeme, že *hodnoty koeficientů jsou jen vodítkem pro interpretaci.*

Uvedli jsme zde přehled, ale se stavem spokojeni nejsme. Domníváme se, že ordinální míry a obzvláště jejich zapojení do dalších následných úloh byly studovány poměrně velmi málo. Vývoj těchto měr by měl být především záležitostí sociologicko-metodologických úvah, protože ordinální znaky jsou nejtypičtější právě pro sociologii a jí příbuzné vědy.

Při zpracování tabulek s ordinálními vstupy zjišťujeme nejen míry ordinální, ale i nominální závislosti, což doporučujeme i čtenářům. Pro tabulky s nominálními vstupy však ordinální míry nepočítáme (ač to zní neuvěřitelně, ve výzkumu lze najít příklady interpretací Spearmanova koeficientu pro čistě nominální vstupy, a to i se znaménkem).

Zároveň chceme upozornit na běžnou, ale velmi hrubou chybu při výpočtu koeficientů: při použití počítačů zahrneme též kódy pro reziduální kategorie, například „neodpověděl", „odmítl odpovědět", „neví", „není si jist" apod., do uspořádané stupnice. Kódy pro tyto hodnoty nesmí být zahrnuty do výpočtů, jinak dostaneme (a mnohdy, bohužel dostáváme) zcela zkreslené a nesmyslné výsledky (které ve výzkumných zprávách figurují jako validní).

Ve stati jsme neuvedli ani metody intervalového odhadu (asymptotickou normalitu a její parametry), ani metody testování hypotéz. Domníváme se, že v sociologii na prvním místě stojí obsahová stránka populačních měr, i když samozřejmě testování a odhad jsou nutnou součástí analýzy dat. O inferenčních problémech se čtenáři mohou poučit v citované literatuře, rozsah této práce nám neumožnil výsledky na tomto poli shrnout současně s definicemi koeficientů.

Literatura

[1] Costner H. L.: *Criteria for Measures of Association.* American Sociological Review 30, 1965, s. 341—353.

[2] Costner H. L.: *Reply to Somers.* American Sociological Review 33, 1968, s. 292.

[3] Četyrkin E. M.: Předmluva in *Statističeskoje izměrenije kačestvennych charakteristik.* Moskva. Statistika 1972, s. 3—7.

[4] Goodman L. A., Kruskal W. H.: *Measures of Association for Cross Classifications.* Journal of American Statistical Association 49/1954, s. 732—764.

[5] Goodman L. A., Kruskal W. H.: *Measures of Association for Cross Classifications II: Further Discussion and References.* Journal of American Statistical Association 54/1959, s. 123—163.

[6] Goodman L. A., Kruskal W. H.: *Measures of Association for Cross Classifications III: Approximate Sampling Theory.* Journal of American Statistical Association 58/1963, s. 310—364.

[7] Kendall M. G.: *Rank Correlation Methods.* London, Ch. Griffin, 1962.

[8] Kendall M. G., Stuart A.: *The Advanced Theory of Statistics.* Vol. 2: Inference and Relationship. London 1967.

[9] *Statističeskije vyvody i svjazy.* Moskva, Nauka 1973.

[10] Kruskal W. H.: *Ordinal Measures of Association.* Journal of American Statistical Association 53, 1958, s. 814—816.

[11] Nie H. N., Bent D. H., Hull C. H.: *Statistical Package for the Social Sciences.* Mc Graw Hill, 1970.

[12] Rosenthal J.: *Distribution of the Sample Version of the Measure of Association — Gamma.* Journal of American Statistical Association, 1966, s. 440—453.

[13] Řehák J.: *K pojmu znak v sociologii.* Sociologický časopis č. 6, 1972, s. 615—625.

[14] Řehák J., Řeháková B.: *Měření statistické závislosti nominálních znaků.* Sociologický časopis č. 4/1973, s. 404—418.

[15] Somers R. H.: *A New Asymetric Measure of Association for Ordinal Variables.* American Sociological Review 27/1962, s. 799—811.

[16] Somers R. H.: *A Similarity between*

Goodman's and Kruskal's Tau and Kendall's Tau, with a Partial Interpretation of the Latter. Journal of American Association 57/1962, s. 804—812.
[17] Somers R. H.: *On the Measurement of Association.* American Sociological Review 33/1968, s. 291—292.
[18] Stuart A.: *The Estimation and Comparison of Strengths of Association in Contingency Tables.* Biometrika 40/1953, s. 105—110.
[19] Wilson T. P.: *A Proportional Reduction — Error Interpretation for Kendall's Tau — b.* Social Forces 47, March 1969, s. 340—342.
[20] Wilson T. P.: *Critigue of Ordinal Variables.* Social Forces 49, March 1971, s. 432—444.
[21] Vencelová, E. S.: *Těorija verojatnostěj.* Moskva, Nauka 1964.

Резюме

Ржегак Я., Ржегакова Б.: Меры статистической зависимости для ординальных переменных

У пары ординальных переменных мы можем измерять или номинальную или ординальную зависимость. Статья разбирает меры ординальной зависимости. Перечень приведенных коэффициентов учитывает отчасти симметрическое и отчасти ассиметрическое отношение и отчасти разные подходы к измерению ординальной зависимости. Пока применя емые в литературе два основных принципа есть следующие:

а) PRE (proportional-reduction-in-error) — принцип, исходящий из того, что возможность предвидеть одну переменную при помощи информации о второй переменной дана статистической связью и предиктивность, таким образом, является индикатором статистической зависимости. Для определения статистической меры этого типа в таком случае мы используем уравнение (3), которое отвечает ассимметрическому отношению. Выборочностью направления предикции мы тогда можем дойти до симметризации меры.

б) Применение общего коэффициента Даниельса корреляции (уравнение (4)), который отвечает симметрическому отношению: в обеих ординальных спецификациях он может быть использован также ассимметрически.

Большинство мер высчитывается из нескольких основных данных; поэтому вводятся эти понятия и приводятся их алгоритмы: число сходств (5), число несходств (6), число разных типов связей (7)—(9) и разные другие отношения (10)—(16).

В перечне содержатся следующие меры:

А) *Гамма Гудмэн-Крускала* (формулы (17—(19)) — приводятся его ассимметрические и симметрические интерпретации, PRE-модель и свойства.

Б) *Тау Кендалла и его модификации* (формулы (20)—(30)) — приводим его как результат общего коэффициента Даниельса, приводим свойства чаще всего употребляемого Γ_b и ссылаемся на литературу на PRE-модель и на возможность использовать его как для симметрических, так и для ассимметрических отношений.

В) $d_{y/x}$ *Сомерса* возникает как ассимметризаци коэффициента тау-b и пригодно для измерения ассиметрического отношения. Приводятся его свойства и ссылка на PRE-модель.

Г) *Симметризированное d* возникает выборочностью направления предикции в PRE-модели; следовательно, симметризация коэффициента другая чем у тау-b Кендалла.

Д) *Коэффициент корреляции Спирмэна для контингенционных таблиц* (формулы (42)—(45)) возникает путем подходящего отбора счетов в ведении Даниельса (формула (4)). Генезис общей примененной версии коэффициента для контингенционных таблиц с упорядоченными категориями отвечает ситуации, когда ординальный признак возникает из сплошного континуума и категории (связавный порядок) отвечают неразличимым наблюдениям. Его возможно понимать также как коэффициент линейной корреляции порядка и его можно выводить как PRE-коэффициент, возникающий при линейной предикции ранга одной переменной при помощи ранга второй переменной. У коэффициента аналогичные свойства, как тау-b Кендалла, его большой недостаток заключается в затруднительности и медлительности вычислений. Представлены два вычислительных приема.

Е) *Балловые методы.* Самым простым балловым методом является использование ранга категории как значения признака. К таким счетам можно в таком случае приложить корреляционный коэффициент Пирсона линейной зависимости кардинальных признаков, который можно интерпретировать в рамках PRE-модели, в которой мы предсказываем число категории одного признака при помощи числа категории второго признака и линейного регрессивного уравнения. Упомянуты также балловые методы исходящие из предпосылки обоснованной вероятностной модели, распределение частот и модель максимальной корреляции.

Для наглядности и лучшее понимание текста и особенно для иллюстрации алгоритмов включены три нумерических примера, исходящие из контингенционных таблиц социологических исследований.

В заключение меры коротко комментируются.

Summary

Řehák J. — Řeháková B.: Measures of Statistical Dependence for Ordinal Variables

In a pair of ordinal variables, either the nominal or the ordinal statistical dependence can be measured. The present paper deals with measures of ordinal dependence. The

survey of introduced coefficients respects the symmetrical and the asymmetrical relation on the one hand, and various approaches to the measurement of ordinal dependence on the other. The two fundamental principles that have hitherto been applied in literature are the following:

a) The PRE (Proportional-Reduction-in-Error) principle based on the fact that the possibility of predicting one variable with the aid of information on the second variable is given by statistical dependence, and thus predictability represents the indicator of statistical dependence. Equation (3), corresponding to the asymmetrical relation, is then used for defining the statistical measure of this type. By randomizing the direction of the prediction, a symmetrization of the measure can be achieved.

b) The application of Daniels' general correlation coefficient (equation [4]) which corresponds to the symmetrical relation; in both ordinal specifications, it can be applied also asymetrically. Most of the measures are calculated from several basic data; this is why these concepts have been introduced and their algorisms indicated: number of congruences (5), number of incongruences (6), number of various types of combination (7)—(9), and various further relations (10)—(16).

The survey includes the following measures:

A. *Goodman-Kruskal's gamma* (formulas [17]—[19]) — its asymmetrical and symmetrical interpretations, PRE model and attributes are given.

B. *Kendall's tau and its modifications* (formulas [21]—[30]) — it is introduced as the consequence of Daniel's general coefficient; attributes of the most frequently used τ_b are mentioned. We refer to literature for the PRE model, as well as for the possibility of applying it both in symmetrical and in asymmetrical relations.

C. *Somers'* $d_{y/x}$ arises as the asymmetrization of the coefficient tau-b and is suitable for measuring the asymmetrical relation. Its attributes are mentioned and reference is made to the PRE model.

D. *Symmetrized d* arises by randomizing the direction of prediction in the PRE model; it is, therefore, a symmetrization of the coefficient that is different from Kendall's tau b.

E. *Spearman's correlation coefficient for contingency tables* (formula [42]—[45]) arises from an adequate choice of scores in Daniels' application (formula [4]). The genesis of the general applied version of the coefficient for contingency tables with ordered categories corresponds to the situation where the ordinal variable arises from the joint continuum and the categories (joint order) correspond to undistinguishable observation. It may also be conceived as the linear-correlation coefficient of order and derived as the PRE coefficient arising in the linear prediction of the order of one variable with the aid of the order of the second variable. The coefficient has similar attributes as Kendall's tau-b; its great disadvantage consists in the difficulty and lengthiness of calculations. Two calculation procedures are presented.

F. *Scoring methods.* The simplest scoring method is the application of the category order as the value of the variable. To such scores, pearson's correlation coefficient of linear dependence of cardinal variables may be applied. This coefficient may be interpreted within the framework of the PRE model wherein we predict the number of the category of one variable with the aid of the number of the category of the second variable and the linear regressive equation. Also scoring methods based on the assumption of the underlying probability model of frequency distribution and the model of maximum correlation are mentioned.

For the sake of elucidation and a better comprehensibility of the text, and particularly for the sake of illustrating the algorisms, three numerical examples proceeding from contingency tables of sociological researches are presented.

In conclusion, the measures are briefly commented upon.