Měření statistické závislosti nominálních znaků*

JAN ŘEHÁK
Ústav pro filosofii a sociologii ČSAV, Praha
BLANKA ŘEHÁKOVÁ
Ústav hematologie a krevní transfuse, Praha

1. Úvodní poznámky

Statistická závislost je jedním z centrálních pojmů analýzy sociologických dat; je to pojem velice komplexní, zahrnuje mnoho různých aspektů a v souvislosti s ním je třeba vyrovnat se s celou řadou statistických metodologických úkolů, jako jsou:

a) vztah statistické a skutečné závislosti;

b) závislost dvou a více proměnných;

c) závislost dvou proměnných na celé populaci a podmíněné závislosti na dílčích souborech; parciální závislost;

d) zobecňování výsledků měření závislostí z výběrového souboru na soubor základní — testování hypotéz a konfidenční intervaly;

e) odlišení různých úrovní závislosti: nominální, ordinální, regresní (lineární, nelineární apod.);

f) odlišení asymetrických a symetrických závislostí;

g) sledování závislostí více proměnných po dvojicích, resp. současné vyhodnocení celé závislosti sítě — závislostní modely;

h) volba přesného popisu, definice, modelu statistické závislosti, a tím určitého koeficientu.

Tyto základní úlohy se vyskytují v každém sociologickém výzkumu, ve kterém se pracuje s pojmy statistická závislost, asociace, souvislost; vždy je třeba se v uvedených bodech rozhodnout pro konkrétní alternativu.

V článku se chceme zaměřit na problém dvojice nominálních znaků, vztah asymetrický i symetrický, na modely, ze kterých vycházíme při konstrukci koeficientů a jež jsou základem pro interpretaci numerických výsledků.

Pro volbu takto zaměřeného tématu je několik důvodů:

1. Především je to potřeba praxe sociologického výzkumu (výzkumní pracovníci často žádají o konzultaci v tomto směru).

2. Chceme seznámit čtenáře se základy, na nichž jednotlivé koeficienty stojí; tato informace má posloužit adekvátnější volbě koeficientů a seznámit naši sociologickou veřejnost s technikami, které jsou prověřeny matematickostatistickou teorií i výzkumnou praxí a mohou být tedy doporučeny k obecnému použití. Znalost modelů a východisek také omezí triviální chyby, např. srovnávání numerických hodnot různých typů koeficientů, což může vést ke zcela absurdním výsledkům.

3. Chceme, aby základní vzorce byly čtenářům snadno dostupné; v praxi je často nutné spočítat některý koeficient pro několik málo tabulek — v tom případě samozřejmě nebudeme používat počítače, ale rozhodneme se pro ruční výpočty; často také potřebujeme spočítat některé míry pro rychlou orientaci v datech.

Rozeznáváme dva typy vzorců:

a) *definitorické* — vyjadřují smysl míry; většinou je možné tyto vzorce dobře přeložit do běžného jazyka, nejsou však vhodné pro výpočet, protože dosazování je velice pracné a velký počet aritmetických operací vede

* Cílem této stati je přispět praxi sociologického výzkumu informací o jednom z důležitých témat oboru metod a technik statistické analýzy dat. Příspěvek však může být současně chápán jako ilustrativní příloha k diskusi o měření. Měříme zde statistickou závislost v dvourozměrných kontingenčních tabulkách — empirickým systémem jsou tedy kontingenční tabulky se dvěma vstupy. Tuto závislost měříme numericky — abstraktním systémem je část reálné přímky. Abychom však mohli měřit, musíme přesně specifikovat, co rozumíme statistickou závislostí, musíme vytvořit model, jehož výsledkem je výpočetní formule, která je pravidlem, podle něhož (dosazením) provádíme zobrazení mezi objekty (tabulkami) a měřící stupnicí (částí reálné přímky). Výpočetní formule je zde zobrazením z množiny kontingenčních tabulek na interval $\langle 0,1 \rangle$. Tím dostáváme měření, které není specificky sociologické. Sociologickým měřením se stane tehdy, když tento model použijeme jako mezistupeň pro zjišťování vztahů (měření závislosti znaků) proměnných pro určitý daný soubor. V souvislosti s problematikou měření chceme upozornit na záměrnost názvu „Statistická závislost znaků". Složitost problematiky sociologického měření je totiž dále komplikována tím, že pro určitou vlastnost můžeme konstruovat více znaků a pro různé znaky téže vlastnosti dostaneme jiné hodnoty míry statistické závislosti. I z toho je tedy vidět složitost vztahu „vlastnost — znak" a nutnost gnozeologicko-metodologické diskuse problému. Současně jsou patrny i meze aplikace statistických měr pro odhalování skutečných vztahů.

obvykle ke zvýšení numerických chyb. Definitorický vzorec, který odráží obsah pojmu, obsahuje způsob interpretace výsledků, je překladem racionálních základních požadavků na míru či zprecizovaným vyjádřením výchozí intuice;

b) *výpočetní* — jsou vzorce formálně zcela ekvivalentní definitorickým, není z nich ovšem vždy vidět smysl a význam (i když v některých případech mohou ukázat další vlastnosti a možnosti dalších interpretací míry); jejich význam spočívá v tom, že zkracují výpočetní čas a námahu a též minimalizují numerické chyby (obsahují většinou menší počet operací dělení i operací vůbec). Výpočetní vzorec tedy obsahuje optimální algoritmus ručního výpočtu.

V této přehledové stati nám jde o problém, který hraje roli v metodách a technikách výzkumu, ve statistické analýze dat. Pojednáme o *statistické závislosti*, tj. o empirických jevech, vztazích, o tom, jaké relace lze nalézt mezi daty. *Jelikož náš technickomatematický úkol je dosti rozsáhlý, nemůžeme se věnovat vztahu příčinnosti a statistické závislosti.* Chceme tu však alespoň upozornit na to, že *zaměňování statistické závislosti se skutečnou a reálnou kauzální závislostí je chybné a že je nejhrubším projevem empirismu.* (Tento problém by jistě zasluhoval samostatnou stať.) Koeficienty statistické závislosti *můžeme používat jen v rámci teoretických vztahových modelů nebo při vyhledávání a formulacích teoretických hypotéz — v obou případech je to pomocný aparát.*

Jde tedy o problém statistický a závislost (resp. nezávislost) bude v celém dalším textu chápána v tomto smyslu.

Informace o statistickém vztahu je uložena ve dvourozměrném rozložení četností zkoumaných proměnných a konstrukce jakéhokoli koeficientu znamená redukci informace [(r — 1) (s — 1) údajů převádíme na jeden údaj]. Aby tato *redukce byla účelná*, aby *měla smysl* a byla pro praktické aplikace *přeložitelná*, musí odpovídat našim požadavkům a *musí být založena na modelu*, který interpretaci a přeložitelnost redukované informace do jednoho koeficientu umožňuje. V článku uvádíme míry, které takový rozumný model mají; u klasických měr (vzniklých normalizací chí-kvadrátu) tento model zcela chybí, a i když samy splňují požadavky na měření statistické závislosti, uvedené v druhé části, absence modelu znamená z hlediska moderního matematickostatistického nazírání absenci základního požadavku, a proto tyto míry nemohou být doporučeny.

Rozhodně však nelze přeceňovat koeficienty zde uváděné; informace kondenzovaná a redukovaná v koeficientech je výhodná pro analytickou práci s velkým počtem statistických vztahů. Při hlubší analýze se však *neobejdeme bez studia a zhodnocení celé struktury kontingenční tabulky* (dvojrozměrného rozložení četností), a to přímo analýzou relativních četností, různým seskupováním kategorií (úpravami původního znaku) nebo pomocí tzv. znaménkového schématu, které právě v podobných situacích je vhodným metodologickým aparátem.

Koeficienty mají tedy svou roli v určitém typu analýzy dat. Z modelu plynou možnosti i meze koeficientů. Jeho absence může vést k libovolnosti nasazování a interpretace.

Při analýze dat musíme ovšem odlišit dva typy úloh, které nás mohou zajímat:

a) odhalení existence závislosti, její signifikantní prokázání — tuto úlohu řešíme testováním hypotéz (statistických);

b) měření síly závislosti.

Obě úlohy nelze směšovat, jejich role v postupu poznání založeném na analýze dat je různá; nám jde o problém druhý.

V těchto úvodních poznámkách ještě chceme upozornit na to, že formálně řešíme úlohu měřenír síly závislosti stejně jako úlohu heterogenity (nepodobnosti) souborů.[1,2] Koeficienty zde uvedené lze tedy použít také v komparačních výzkumech jako míry podobnosti či rozdílnosti populací (souborů dat).

---

[1] Termín homogenní, resp. heterogenní má ve statistice dva významy:
a) Vztahuje se k podobnosti, resp. rozdílnosti souborů dat — jde o podobnost statistických rozložení jednoho znaku v různých populacích, resp. oblastech; v tomto významu je termín použit zde. Homogenní je ve statistice použito ve významu „směsitelný" s jinými populacemi, aniž by se výsledné rozložení relativních četností, resp. pravděpodobností, změnilo. Termín vychází z homogennosti (stejnorodosti) skupin dat.
b) Vztahuje se k podobnosti údajů v jedné skupině dat — jde o homogenitu souboru dat, tj. podobnost jedinců v souboru vzhledem k určitému danému znaku; v tomto smyslu termín v článku nebudeme používat a nahradíme jej jiným termínem „neurčitost", který odpovídá heterogenitě v souboru.

[2] Ekvivalence úlohy nezávislosti (resp. závislosti) a homogenity (resp. heterogenity) souborů je vidět z toho, že určení dat do jednotlivých souborů si můžeme představit jako realizaci určitého znaku, jehož hodnoty jsou názvy souborů. Homogenita je pak totéž jako prohlášení, že zkoumaný znak je na takto zavedeném znaku nezávislý, tj. rozložení hodnot znaků je ve všech podsouborech stejné (jsou stejná podmíněná rozložení).

2. Měření závislosti asymetrického typu — principy

Předpokládejme, že u dvou proměnných (resp. znaků) X, Y můžeme určit, která z nich je nezávisle (např. X) a která závisle proměnná (např. Y), tzn., že můžeme přijmout závislostní model (schéma)

$$X \rightarrow Y$$

(X ovlivňuje Y, X časově předchází Y, X je nezávisle a Y závisle proměnná atp.).

Úkolem měření statistické závislosti je především odvodit nějaké smysluplné míry, které by charakterizovaly těsnost statistického vztahu obou proměnných.

Intuitivní požadavky na tyto míry mohou být vysloveny obecně ještě dříve, než zavedeme přesnou definici statistické závislosti pouze s intuitivním chápáním tohoto termínu.

Obecné požadavky na míru statistické závislosti α:

a) $0 \leq \alpha \leq 1$ (požadavek normalizace měřicí stupnice, který je motivován zvykem a tím, že se nám s takovou stupnicí dobře pracuje);

b) $\alpha = 0$, nastane právě když obě proměnné (znaky), které uvažujeme, jsou *statisticky nezávislé*;

$\alpha = 1$ nastane právě když obě proměnné jsou jednoznačně vázány, tj. jsou *funkčně závislé*.[3] (Tyto dva požadavky vymezují krajní hodnoty a určují orientaci stupnice.)

c) $\alpha \doteq 0$, nastává nepatrný odklon od nezávislosti; $\alpha \doteq 1$, nastává nepatrný odklon od závislosti (upřesnění významu krajních hodnot);

d) α je invariantní k uspořádání sloupců a řádků tabulky, [požadavek určuje nominální úroveň — u nominálních znaků je uspořádání hodnot (kategorií) zcela libovolné, proto míra souvislosti musí být stejná, ať už volíme uspořádání jakékoli];

e) čím větší je α, tím větší je závislost — u tohoto intuitivního požadavku je však třeba specifikovat jednoznačně jeho význam tím, že určíme konkrétní model.

Statistická data z určitého souboru uspořádaná v tabulce, tj. dvojrozměrné sdružené statistické rozložení četností u hodnot obou znaků, marginální rozložení každého znaku zvlášť a podmíněná rozložení nám přinášejí jednak informaci o každé proměnné zvlášť, jednak o vztahu obou proměnných v daném souboru dat. Princip, na kterém je většina zde uvedených měr založena, lze verbálně popsat takto: Předpokládáme, že platí model $X \rightarrow Y$. Proměnná Y má dané statistické rozložení četností pro zkoumaný soubor; toto rozložení je buď více, nebo méně neurčité v tom smyslu, že nám přináší méně nebo více informace o náhodně zvoleném jedinci — jestliže jsou všichni jedinci v jedné kategorii znaku Y, pak můžeme samozřejmě u každého z nich jednoznačně určit, do které kategorie patří, aniž bychom to zkoumali složitými metodami; jsou-li naproti tomu jednotky rozmístěny stejnoměrně v kategoriích, pak u náhodně zvolené jednotky bude naše predikce správné hodnoty riskantní (v pravděpodobnostním smyslu). Pro každé rozložení četností znaku Y máme tedy jistý stupeň neurčitosti. Z hlediska znaku X se daný soubor dat rozpadá na části (subpopulace, dílčí soubory, oblasti, strata apod.). V každé této části můžeme zjistit podmíněné rozložení četností znaku Y. To znamená, že nalezneme relativní četnosti, jimiž jsou hodnoty znaku Y zastoupeny v jednotlivých dílčích souborech. Ve většině případů se budou tato podmíněná rozložení lišit; budou se tedy lišit i predikce (určování hodnot) a neurčitost predikcí hodnot Y u náhodně zvolených jednotek vybraných z jednotlivých dílčích populací. Určíme-li jakékoli predikční pravidlo vycházející z rozložení četností, bude jistě lépe vycházet z dílčího podmíněného rozložení četností, které znamená upřesnění informace o jednotce. Budeme tedy očekávat, že specifikace X nám umožní lépe „uhodnout" Y, tj., že X přináší nějakou informaci (empirickou) o Y, že znalost X redukuje neurčitost Y.

Jestliže redukce neurčitosti v datech je velká, pak závislost $X \rightarrow Y$ je silná (těsná), jestliže je redukce zanedbatelná, pak nemůžeme o závislosti mluvit.

[3] Někdy se tento požadavek formuluje tak, že koeficient nabývá hodnoty 1 právě tehdy, když četnostní (resp. pravděpodobnostní) rozložení obou proměnných je *singulární*. V sociologické analýze se zdá být rozumnější požadavek uvedený v textu; znamená, že pro každou hodnotu nezávisle proměnné můžeme pro daný soubor dat jednoznačně určit hodnotu závislé proměnné. Podrobnější diskuse těchto dvou přístupů by připadala v úvahu v případech symetrického vztahu. Pojem funkční závislosti je vzat z matematiky a je chápán jako jednoznačné zobrazení z množiny hodnot X do množiny hodnot Y, zprostředkované kontingenční tabulkou; takto chápaná funkční závislost je v našem případě tedy krajním případem statistické závislosti.

Úkol[4] odvození míry tedy zní:

— nalézt nějakou míru neurčitosti,

— vyjádřit matematicky redukci neurčitosti plynoucí ze znalosti hodnoty X a položit ji do relace k původní neurčitosti proměnné Y.

Před upřesněním modelů zaveďme ještě značení, které je obvyklé ve statistických příručkách a které bude jednotně používáno i v tomto textu (viz tab. 1).

Tabulka 1.

X (Kategorie znaku X)	Y: Kategorie znaku Y — 1	2 ...	j ...	s	Součet
1	n_{11}	n_{12} ...	n_{1j} ...	n_{1s}	$n_{1.}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
i	n_{i1}	n_{i2} ...	n_{ij} ...	n_{is}	$n_{i.}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
r	n_{r1}	n_{r2} ...	n_{rj} ...	n_{rs}	$n_{r.}$
Součet	$n_{.1}$	$n_{.2}$ ...	$n_{.j}$ ...	$n_{.s}$	n

n_{ij} je počet statistických jednotek v souboru, které mají současně hodnotu i-té kategorie znaku X a j-té kategorie znaku Y;

$n_{i.}$ je řádkový součet — počet jednotek, které mají hodnotu i-té kategorie znaku X;

$n_{.j}$ je sloupcový součet — počet jednotek, které mají hodnotu j-té kategorie znaku Y;

n je celkový počet jednotek uvažovaného souboru vstupujících do uvedené kontingenční tabulky.

Hodnoty $n_{i.}$ určují marginální rozložení znaku X, hodnoty $n_{.j}$ určují marginální rozložení znaku Y.

Písmeny f_{ij}, $f_{i.}$, $f_{.j}$ budeme značit relativní četnosti z celého souboru[5] v téže tabulce, tzn., že to jsou postupně $\frac{n_{ij}}{n}$, $\frac{n_{i.}}{n}$, $\frac{n_{.j}}{n}$; součet hodnot f_{ij} je jednička.

Nakonec ještě musíme vyslovit definici statistické nezávislosti. Ta je zcela jednoznačná, zatímco pojem stupně závislosti dvou proměnných je specifikován u každého modelu trochu jinak, jak uvidíme později.

Znaky X a Y jsou statisticky vzájemně nezávislé, jestliže platí pro všechny relativní četnosti v tabulce vztah:

$$f_{ij} = f_{i.}\, f_{.j}$$

(tzn., že výskyty všech dvojic hodnot X a Y jsou statisticky nezávislé jevy).

Pro absolutní četnosti lze tento vztah vyjádřit tak, že tzv. „očekávané četnosti" jsou rovny skutečným zjištěným četnostem:

$$n_{ij} = \frac{n_{i.}\, n_{.j}}{n}$$

Podrobnější diskuse tohoto pojmu a motivaci definice nalezne čtenář v kterékoli učebnici počtu pravděpodobnosti.

Lze k ní přejít z definice nezávislosti jedné proměnné na druhé: Y je nezávislá na X, jestliže rozložení četností Y ve všech dílčích podsouborech určených hodnotami X jsou stejná, tj. jestliže rozložení Y postupně podmíněná jednotlivými hodnotami X jsou stejná — a tedy stejná jako nepodmíněné rozložení Y (specifikace podle hodnoty X nemění statistickou informaci o Y, uloženou v rozložení četností).

Označíme-li $(f_{1/1}, f_{2/i}, \ldots, f_{s/1})$ podmíněné relativní četnosti Y v subpopulaci, která odpovídá i-té kategorii znaku X, pak v případě nezávislosti Y na X jsou všechna tato rozložení stejná mezi sebou a též stejná jako hodnoty $(f_{1.}, f_{2.}, \ldots, f_{s.})$. Poznamenejme, že $f_{j/i} = n_{ij}/n_{i.}$ a že $\sum_j f_{j/i} = 1$ pro všechna $i = 1, 2, \ldots, r$.

K definici vzájemné nezávislosti přejdeme odtud velice snadno jednoduchými algebraickými úpravami a konstatováním, že „Y je statisticky nezávislé na X $\Leftrightarrow$ X je statisticky nezávislé na Y". Tato věta říká, že *statistická nezávislost je symetrická relace na množině znaků (proměnných)*. U asymetrického případu vycházíme z nezávislosti Y na X (tj. jeden krajní případ) a pojem statistické závislosti budeme specifikovat třemi různými možnostmi vyjádření neurčitosti (tři modely, tři východiska), a tedy tři různé míry. Vy-

[4] Čtenář, který je obeznámen s korelačním poměrem, jistě okamžitě postřehl analogii — korelační poměr je založen na přesně stejném principu. Rozptýlení (neurčitost) dat je zde však charakterizováno součtem čtverců hodnot od průměru — to ovšem je možné jen u kardinálních znaků. Redukce neurčitosti je provedena v dílčích souborech a výsledná formule je podílem této redukce a původní hodnoty součtu čtverců.

[5] Čísla f_{ij} považujeme za relativní četnosti celého souboru; můžeme je také považovat za pravděpodobnosti výskytu pro neomezené populace; v tom případě výpočet koeficientů provádíme ze vzorců obsahujících f_{ij} (nikoliv absolutní četnosti) nebo při dosazení výběrových absolutních četností n_{ij} dostáváme výběrové hodnoty koeficientu jako konzistentní odhady koeficientů populačních.

jádření redukce neurčitosti je u všech těchto případů obdobné.

3. Míry asymetrické statistické závislosti

A. Guttmanův koeficient prediktability

Koeficient vychází z následujícího modelu: Máme-li nějaké statistické rozložení[6]

$$f_1^*, f_2^*, \ldots, f_j^*, \ldots, f_s^*; \sum f_i^* = 1$$

a vezmeme-li jednu náhodně zvolenou statistickou jednotku, zvolíme toto predikční pravidlo pro neznámou (pro nás) hodnotu znaku jednotky: *přiřadíme této jednotce modální kategorii* (tj. kategorii nejvíce četnou; jestliže máme více modálních kategorií, volíme z nich náhodně).

Z modální predikce plyne jednoduchá míra neurčitosti: pravděpodobnost, že určíme kategorii respondenta chybně. (Čím menší je tato pravděpodobnost, tím menší je neurčitost v datech a tím větší šance na správnou predikci; přístup je tedy až potud jistě rozumný.)

Při znalosti informace o X nebudeme určovat Y hůře. V krajním případě se nám dokonce může stát, že budeme určovat hodnotu Y zcela jednoznačně pomocí X — a to bude situace, kdy v jednotlivých dílčích souborech (podle X) už nemáme žádnou neurčitost, tj. predikujeme zcela bez chyb.

Například u souboru o 100 lidech víme, že 70 osob sleduje sportovní pořad v televizi a 30 nesleduje. Rozdělíme-li soubor podle pohlaví, můžeme zjistit, že nastane tato možná situace: ze 70 sledovaných mužů všichni sledují pořad, ze 30 žen žádná pořad nesleduje. Z těchto dat (z této informace) a z informace, že osoba je muž, plyne jednoznačně závěr o hodnotě znaku sledování sportovního pořadu.

Označíme-li $f_{.m}$ četnost modální kategorie v marginálním rozložení znaku Y, pak chyba modální predikce má pravděpodobnost:

P (chyba predikce hodnoty znaku Y) = $1 - f_{.m}$

(Chyby se dopustíme vždy, když prvek souboru nepatří do modální kategorie.)

Jestliže obdobně označíme ($f_{im} = \max f_{ij}$) a $f_{m/i}$ je největší (modální) četnost pro podmíněné rozložení znaku Y v j-té kategorii znaku X, dostáváme postupně:
P (chybné určení Y za předpokladu, že jednotka patří do i-té kategorie znaku X) = $= 1 - f_{m/i}$, pro $i = 1, 2, \ldots, r$.

Podle vzorce úplné pravděpodobnosti pak P (chybné určení hodnoty Y při znalosti kategorie znaku X) = $1 - \sum_i f_{im}$

a redukce chyby je tedy:

$$(1 - f_{.m}) - (1 - \sum f_{im}) = \sum f_{im} - f_{.m}$$

Můžeme tedy zavést koeficient modální predikce jako podíl redukované a původní neurčitosti:

$$(1) \qquad \lambda_{Y/X} = \frac{\sum f_{im} - f_{.m}}{1 - f_{.m}}$$

$$(2) \qquad = \frac{\sum n_{im} - n_{.m}}{n - n_{.m}}$$

$n_{.m}$, n_{im} jsou postupně maxima marginálního rozložení a řádkových rozložení absolutních četností, obdobně jako u relativních četností.

Vlastnosti koeficientu:

1. Koeficient nemá věcný smysl, je-li zastoupena pouze jedna kategorie znaku X; jeho hodnota je 0; v interpretaci však nelze chápat tuto hodnotu jako nezávislost, protože X není variabilní, její hodnota tudíž nemůže přispět k predikci hodnoty Y.

2. Koeficient není definován, jsou-li všechny údaje ze souboru dat seskupeny v jednom sloupci — tj. nemáme žádnou chybu v určování hodnoty Y z marginálního rozložení, X proto nemůže už přinést žádné další zlepšení predikce.

3. Koeficient má hodnoty mezi nulou a jedničkou včetně, přitom

a) $\lambda_{Y/X} = 0$, právě když znalost hodnoty X nepřináší žádnou informaci pro předikci Y; všechna řádková maxima jsou ve stejném sloupci jako marginální maximum; podmíněná rozložení v řádcích mají stejné modální kategorie; (existuje j_0 tak, že $f_{ij_0} = f_{im}$ pro všechna i).

b) $\lambda_{Y/X} = 1$ nastává, právě když znalost hodnoty X umožňuje jednoznačnou predikci hodnoty Y, tzn., že v každém řádku tabulky je právě jedno nenulové pole;

4. Koeficient se nemění, změníme-li pořadí řádků nebo sloupců: je invariantní k permutaci hodnot obou proměnných. To odpovídá vlastnostem závislosti nominálních znaků.

Vlastnost 3a) nezaručuje ekvivalenci nulové hodnoty a případu nezávislosti a to znamená, že z *nulové hodnoty koeficientu ne-*

[6] Značení f* je použito proto, aby nedošlo ke konfúzi se značením použitým výše.

můžeme usuzovat na nezávislost obou proměnných; jelikož taková ekvivalence je pro nás důležitá, koeficient není příliš používán. (Stejný nedostatek mají některé tradiční míry.) Dalším nedostatkem je, že v případě závislosti je citlivý pouze na maxima, a ne na celkovou strukturu tabulky.

Výhody koeficientu:

a) Vzhledem k snadno pochopitelnému modelu a vlastnostem můžeme bez potíží určit, zda se nám jeho použití hodí, nebo ne.

b) Snadná spočitatelnost; algoritmus spočívá v zatržení řádkových maxim, jejich součtu a jednom úkolu dělení.

c) Jsou známy různé testy hypotéz pro tento koeficient.

Pro obecnou aplikaci v sociologických výzkumech jej nedoporučujeme. Uvádíme jej zde jednak proto, že jeho model je jednoduchý a snadno pochopitelný a může tedy sloužit jako úvod do problému pro nematematiky, a hlavně proto, že jej velice často *používáme pro rychlé ruční výpočty, které slouží k první orientaci v datech.*

Obecná formule pro tento asymetrický koeficient:

$$\frac{P\left(\begin{array}{c}\text{chyba při predikci Y}\\ \text{z marginálního}\\ \text{rozložení}\end{array}\right) - P\left(\begin{array}{c}\text{chyba při predik-}\\ \text{ci Y při znalosti}\\ \text{kat. X}\end{array}\right)}{P\ (\text{chyba při predikci Y z marginálního rozložení})} \qquad (3)$$

Tato formule, navržená Guttmanem, byla specifikována tím, že byl přesně navržen způsob predikce pro náhodně zvolenou statistickou jednotku. Z překladu této formule lze heuristicky rychle odvodit smysluplnost míry: je to relativní úbytek pravděpodobnosti chyby predikce při znalosti hodnoty jiné proměnné; je to míra informačního přínosu jedné proměnné pro predikci proměnné druhé.

B. Wallisův koeficient proporcionální predikce

Wallisův koeficient predikce je založen na stejném obecném modelu, vyjádřeném ve vzorci (3), pouze způsob predikce je jiný: známe-li rozložení četností v kategoriích znaku, pak predikujeme hodnotu znaku pomocí náhodného mechanismu (např. tabulky náhodných čísel) tak, že *přiřazujeme vybrané jednotce hodnoty znaku s pravděpodobnostmi úměrnými zastoupení jednotlivých hodnot v populaci*; tzn., že realizujeme rozložení pravděpodobností (nebo četností) a výsledek přiřazujeme k dané jednotce. Tento postup opět uplatňujeme pro Y buď bez znalosti, nebo se znalostí X, spočítáme příslušné pravděpodobnosti chyb predikce a po dosazení do obecného vzorce dostáváme koficient:

$$\tau_{Y/X} = \frac{\sum\sum \frac{f_{ij}^2}{f_{i.}} - \sum f_{.j}^2}{1 - \sum f_{.j}^2} \qquad (4)$$

$$= \frac{\sum\sum \frac{1}{f_{i.}} (f_{ij} - f_{i.} f_{.j})^2}{1 - \sum f_{.j}^2} \qquad (5)$$

$$= \frac{n \sum\sum \frac{n_{ij}^2}{n_{i.}} - \sum n_{.j}^2}{n^2 - \sum n_{.j}^2} \qquad (6)$$

Vlastnosti koeficientu:

1. V případě, že data jsou seskupena do jednoho řádku, nemá věcný smysl koeficient počítat, neboť X nemá žádnou variabilitu a nemůže tedy přinášet žádnou informaci.

2. $\tau_{Y/X}$ není definován pro případ, že se všechna data seskupí do jednoho sloupce tabulky (v takovém případě nemá smysl o závislosti mluvit).

3. $\tau_{Y/X}$ nabývá hodnot mezi nulou a jedničkou včetně; čím vyšší je hodnota koeficientu, tím vyšší je prediktabilita Y z X, a tedy tím vyšší je i statistická závislost Y na X; přitom

a) $\tau_{Y/X} = 0$, *právě když* X a Y jsou *statisticky nezávislé*, tzn., že $f_{ij} = f_{i.}\, f_{.j}$ pro všechna pole tabulky;
b) $\tau_{Y/X} = 1$, právě když znalost hodnoty X *jednoznačně umožňuje určení hodnoty znaku* Y.

4. Koeficient se nezmění permutací řádků nebo sloupců tabulky — to odpovídá práci s nominálními znaky a obecnému typu statistické závislosti.

K bodu 2. lze ještě dodat: koeficient nemá smysl, je-li jeden řádek nulový. V praxi postupujeme tak, že tento řádek z tabulky vynecháváme a redukujeme tabulku jen na ty řádky, které obsahují některé údaje — při analýze je ovšem třeba vzít takovou redukci hodnot znaku v úvahu. Při závislostních úvahách pracujeme s neúplnými znaky a naše informace o závislosti tedy není dokonalá.

Z uváděných vzorců používáme pro vý-

počty většinou vzorce (6), vycházejícího z absolutních četností.[7]

Vlastnost 3a) je velice důležitá oproti obdobné vlastnosti koeficientu $\lambda_{Y/X}$, protože zde máme ekvivalenci nabývání nuly a statistické nezávislosti, tj., je tu splněna podmínka b), kterou jsme kladli na obecné míry. Tím je také dána preference tohoto koeficientu vzhledem ke koeficientu Guttmanovu. Výpočty vedoucí k tomuto koeficientu jsou však zdlouhavější, a proto se tento koeficient nehodí pro rychlou analýzu a pro orientaci v datech v případě ručních výpočtů. Při použití počítače mu však dáváme přednost.

C. Informační míra asymetrické závislosti

Tato míra je založena na modelu teorie informace. V předchozích případech jsme dosazovali do obecného vzorce Guttmanova, v němž se vyskytovaly pravděpodobnosti chyb. Nyní vezmeme zcela analogický obecný vzorec, v němž místo pravděpodobností chyby predikce budeme uvažovat obecnou míru neurčitosti (variability) statistického rozložení. Zobecníme-li takto základní vzorec, dostáváme:

$$(7)\qquad \frac{\begin{pmatrix}\text{neurčitost v}\\ \text{rozložení Y}\end{pmatrix} - \begin{pmatrix}\text{neurčitost v rozložení Y}\\ \text{při znalosti X}\end{pmatrix}}{(\text{neurčitost v rozložení Y})}$$

Místo redukce pravděpodobnosti chyby zde máme relativní úbytek obecné míry neurčitosti v distribuci proměnné Y při znalosti hodnoty X, a to v poměru k původní hodnotě neurčitosti.

Teorie informace zavádí jako míru neurčitosti tzv. entropii[8] a informaci jako míru redukce entropie.

Po dosazení těchto pojmů do obecného vzorce dostáváme (viz Nikl, Perez (8]):

$$\zeta_{Y/X} = \frac{H(Y) - H(Y/X)}{H(Y)} \qquad (8)$$

$$= \frac{I(X, Y)}{H(Y)} \qquad (9)$$

$$= 1 - \frac{\sum f_{i.} \log f_{i.} - \sum\sum f_{ij} \log f_{ij}}{-\sum f_{.j} \log f_{.j}} \qquad (10)$$

$$= 1 - \frac{\sum n_{i.} \log n_{i.} - \sum\sum n_{ij} \log n_{ij}}{n \log n - \sum n_{.j} \log n_{.j}} \qquad (11)$$

Vlastnosti tohoto koeficientu jsou stejné jako vlastnosti koeficientu $\tau_{Y/X}$ (viz vlastnosti Wallisova koeficientu). Proto je jeho aplikace obdobná.

Navíc má tento koeficient — oproti předcházejícímu — jednu velkou výhodu — lze jej velice snadno spočítat, máme-li k dispozici tabulky hodnot $k \log k$ (pro přirozená čísla k) nebo $-p \log p$ (pro hodnoty p mezi nulou a jedničkou). Jsou-li tabulky takových hodnot dostupné,[9] pak je výpočet velice rychlý — znamená pouze sčítání a odčítání tabulkových hodnot a jedno dělení.

4. Symetrická statistická závislost

V mnoha případech nemůžeme určit kauzální nebo časovou posloupnost proměnných, tj. roli nezávisle a závisle proměnné; v tom případě nemá asymetrické měření statistické závislosti žádný smysl a je navíc chybné a zavádějící. Proto musíme volit jiný přístup k problému a jiné východisko modelu.

Rozumný model symetrizace, který zde popisujeme, je v matematické statistice obecně považován za přijatelný a je také často používán. Pojem symetrické závislosti používáme tam, kde nevíme, která z obou zkoumaných vlastností předchází druhé, naše poznání neumožňuje určit mezi vlastnostmi vztah nezávislé a závislé proměnné, žádná z nich nemůže sloužit jako explanační faktor pro druhou apod.; zatímco mezi věkem a postojovými otázkami je směr možného statistického vztahu jednoznačný, u dvou postojových otázek to již tak snadné není. Symetričnost navíc může znamenat také vzájemné současné ovlivňování obou vlastností. Model je založen na této úvaze: neznáme směr asymetrie, proto uplatníme predikci oběma směry a výsledky vážíme

[7] Druhý tvar vzorce ukazuje souvislost s tzv. chí-kvadrátovými nebo klasickými měrami, ve kterých vystupují stejné členy: čtverce rozdílů mezi zjištěnými a očekávanými četnostmi v polích, $(f_{ij} - f_{i.}\ f_{.j})^2$: tyto členy jsou v součtu násobeny vahami $(1 - \Sigma f^2_{.j}) / f_{i.}$; tím vzniká určitý způsob normalizace základních členů. $f_{i.}$ ve jmenovateli jsou váhy, které zrovnoprávňují řádky s nižším zastoupením se řádkami četnějšími.

[8] S teorií informace zde seznamovat nemůžeme, čtenáře odkazujeme např. na práci Jasin (5). Uvádíme zde jen základní pojmy a značení: H(X) — entropie znaku X, H(Y) — entropie znaku Y, H(X, Y) — entropie dvourozměrného rozložení vzniklého kombinací obou znaků X, Y. Informace X o Y (stejná jako Y o X) je I (X, Y) = H(X) + + H(Y) — H(X, Y) = H(Y) — H(Y/X) = „neurčitost v Y" — „neurčitost v Y při znalosti X". Dále definujeme $H(Y) = -\Sigma f_{.j} \log f_{.j}$ atd.

[9] Tabulky jsou publikovány v knize Kullback [6], jejíž ruský překlad z roku 1967 je u nás dostupný. Při výpočtech nerozhoduje základ logaritmu, protože se faktor převodu při dělení zkrátí. Používáme-li však mezivýsledky pro jiné metody (např. testování hypotéz), pracujeme s přirozeným logaritmem.

stejnou vahou (rovnou jedné polovině), tj. vezmeme náhodně jeden prvek, rozhodneme se s pravděpodobností 1/2, který směr predikce budeme uvažovat, a pro ten provedeme predikci.

Místo pravděpodobnosti chyb predikce zde dostaneme očekávané hodnoty pravděpodobností příslušný k oběma směrům.

P (chybná predikce jedné proměnné bez znalosti druhé) = 1 — 1/2 [P (správně pro X) + P (správně pro Y)]

Obdobné platí pro predikci jedné proměnné při znalosti proměnné druhé. Tyto hodnoty dosadíme do obecného Guttmanova schématu.

A. Symetrizovaný Guttmanův koeficient predikce

Vycházíme z modelu modální predikce a dostáváme koeficient, který charakterizuje míru symetrické statistické závislosti:

$$\lambda = \frac{\frac{1}{2}[\Sigma f_{im} + \Sigma f_{mj} - f_{.m} - f_{m.}]}{1 - \frac{1}{2}(f_{.m} + f_{m.})} \quad (12)$$

f_{im} značí relativní četnost (pro celé dvojrozměrné rozložení), která je maximální v i-tém řádku;
f_{mj} je maximální relativní četnost v j-tém sloupci;
$f_{m.}$ je maximum v marginálním rozložení veličiny X;
$f_{.m}$ je maximum v marginálním rozložení veličiny Y.

Obdobné značení přijmeme pro absolutní četnosti; po úpravách dostaneme vzorec vhodný k výpočtům:

$$\lambda = \frac{\sum n_{im} + \sum n_{mj} - n_{.m} - n_{m.}}{2n - (n_{.m} + n_{m.})} \quad (13)$$

Vlastnosti koeficientu:

1. Nemá interpretační smysl, jsou-li data soustředěna v jednom řádku nebo v jednom sloupci.

2. Není definován pro případ, že data jsou soustředěna do jednoho pole tabulky.

3. Leží v uzavřeném intervalu [0, 1], přitom:

a) $\lambda = 0$, když X a Y *jsou nezávislé, ale neplatí opak* (může mít nulovou hodnotu i v případech statistické závislosti);

b) $\lambda = 1$, právě když hodnoty obou znaků jsou pro danou populaci *jednoznačně párovány*, tj. každý řádek a každý sloupec v tabulce obsahuje právě jedno nenulové pole.

4. Není citlivý na libovolné permutace řádků a sloupců; zároveň nezávisí na změně role sloupcové a řádkové proměnné (toto je vlastnost symetrizace; u asymetrických koeficientů neplatí).

5. Leží mezi $\lambda_{Y/X}$ a $\lambda_{X/Y}$ včetně.

Jeho výhodou je opět výpočetní jednoduchost, a *proto se používá hlavně pro rychlé analýzy dat* s ručními výpočty. Jeho nevýhodou je vlastnost 3a), ve které opět chybí ekvivalence mezi nabýváním nulové hodnoty a případem statistické nezávislosti obou znaků.

B. Symetrizovaný Wallisův koeficient proporcionální predikce

Zcela stejným postupem symetrizace, ale za použití modelu proporcionální predikce dostaneme vzorec

$$\tau = \frac{\frac{1}{2}\sum\sum\left[(f_{ij} - f_{i.}f_{.j})^2 \cdot \left(\frac{f_{i.} + f_{.j}}{f_{i.}f_{.j}}\right)\right]}{1 - \frac{1}{2}\sum f_{i.}^2 - \frac{1}{2}\sum f_{.j}^2} \quad (14)$$

Tento vzorec opět má tu vlastnost, že čtverce rozdílu očekávaných četností a skutečných četností v polích jsou váženy — váhy jsou nyní symetrické vzhledem k i a j (tj. k řádkům a sloupcům). Za výpočetní formuli může být vzat například některý z uvedených výrazů:

$$\tau = \frac{n\sum\sum\left[(nn_{ij} - n_{i.}n_{.j})^2\left(\frac{n_{i.} + n_{.j}}{n_{i.}n_{.j}}\right)\right]}{2n^2 - \sum n_{i.}^2 - \sum n_{.j}^2} \quad (15)$$

$$= \frac{\sum\sum\frac{f_{ij}^2}{f_{.j}} + \sum\sum\frac{f_{ij}^2}{f_{.i}} - \sum f_{i.}^2 - \sum f_{.j}^2}{2 - \sum f_{i.}^2 - \sum f_{.j}^2} \quad (16)$$

$$= \frac{n\left(\sum\sum\frac{n_{ij}^2}{n_{i.}} + \sum\sum\frac{n_{ij}^2}{n_{.j}}\right) - \sum n_{i.}^2 - \sum n_{.j}^2}{2n^2 - \sum n_{i.}^2 - \sum n_{.j}^2} \quad (17)$$

Vlastnosti jsou stejné jako u λ, jen s tím rozdílem, že místo 3a) máme *požadovanou ekvivalenci*:

$\tau = 0$, právě když X a Y jsou statisticky nezávislé, tj. $f_{ij} = f_{i.}f_{.j}$ pro všechny dvojice i, j (pole tabulky).

Při výpočtu vynecháváme prázdné řádky a prázdné sloupce.

C. Symetrizovaný informační koeficient

Symetrizační princip můžeme použít i pro zobecněný vzorec (7) při dosazení entropií. Navrhujeme zde koeficient, který z tohoto přístupu vyplývá; jeho tvar udává vzorec:

$$\zeta = \frac{I(X, Y)}{\frac{1}{2}[H(X) + H(Y)]} \tag{18}$$

$$= 2\left[1 - \frac{H(X, Y)}{H(X) + H(Y)}\right] \tag{19}$$

$$= 2\left[1 - \frac{-\sum\sum f_{ij} \log f_{ij}}{-\sum f_{i.} \log f_{i.} - \sum f_{.j} \log f_{.j}}\right] \tag{20}$$

$$= 2\left[1 - \frac{n \log n - \sum\sum n_{ij} \log n_{ij}}{2n \log n - \sum n_{i.} \log n_{i.} - \sum n_{.j} \log n_{.j}}\right] \tag{21}$$

Vlastnosti jsou stejné jako u koeficientu τ (viz předcházející oddíl). Výpočet koeficientu je však daleko snadnější, máme-li k dispozici tabulky *k log k*, resp. *— p log p*.

D. Koeficienty založené na statistice chí — kvadrát

Klasické koeficienty jsou zavedeny pro symetrické vztahy a jsou založeny na zcela jiné myšlence než koeficienty predikční.

Vycházíme ze stavu nezávislosti a hledáme nějaké charakteristiky, které by porovnávaly stav daný v tabulce vzhledem ke stavu nezávislosti. K. Pearson navrhl „míru vzdálenosti", která je takovou charakteristikou, odrážející „vzdálenost" dvourozměrného rozložení četností v tabulce od rozložení, které by odpovídalo případu nezávislosti při stejných marginálních četnostech:

$$\chi^2 = n \sum\sum \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}} \tag{22}$$

$$= \sum\sum \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}} \tag{23}$$

$$= n \sum\sum \frac{n_{ij}^2}{n_{i.} n_{.j}} - n \tag{24}$$

$$= n\left(\sum\sum \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1\right) \tag{25}$$

Za charakteristiku vzdálenosti u jednoho pole byl vzat výraz $(f_{ij} - f_{i.} f_{.j})^2$; (v literatuře se ovšem vyskytly i jiné typy takovýchto vzdáleností).

Poslední dva vzorce používáme pro výpočet.

Veličinu chí-kvadrát musíme normovat tak, aby ležela mezi nulou a jedničkou včetně a aby tak vyhovovala požadavkům specifikovaným v § 2. Z mnoha pokusů o normalizaci je nejlepší koeficient Cramérův:

$$Cr = \frac{\chi^2}{n \cdot \min(r - 1, s - 1)} \tag{26}$$

Při výpočtu vynecháváme nulové řádky a sloupce a tím také snižujeme příslušné hodnoty r a s ve vzorci.

Vlastnosti koeficientu:

1. Nemá věcný smysl, jsou-li data soustředěna pouze v jednom sloupci nebo pouze v jednom řádku.

2. Není definován pro nulové sloupce nebo řádky — ty musíme při výpočtu vynechávat; s takovou podmínkou je definován vždy, když min $(r - 1, s - 1)$ není nula, to znamená, že data nejsou soustředěna v jednom řádku nebo sloupci.

3. Leží mezi nulou a jedničkou včetně pro libovolné rozměry tabulky:
a) $Cr = 0$, právě když X a Y jsou nezávislé,
b) $Cr = 1$, právě když je závislost kompletní.

4. Je invariantní vzhledem k libovolné permutaci sloupců nebo řádků, je symetrický vzhledem k přehození řádkové a sloupcové proměnné.

Tento typ koeficientů nelze doporučit, protože *nejsou založeny na dobře interpretovatelném modelu*. Je to názor, který se v literatuře vyskytl již v základní práci Goodman-Kruskalově [2]. Chí-kvadrát je statistika, která je jednoznačně přijata pro testování nezávislosti a jako taková se dobře osvědčila (v poslední době se však i od ní upouští na úkor statistik teorie informace). Testování nezávislosti a měření závislosti jsou však

dvě zcela různé úlohy, které vycházejí ze zcela různých vstupů do výzkumu a z jiných modelů analýzy dat.

Vzhledem k tomu, že volba základní charakteristiky odpovídá nulové hypotéze nezávislosti, ale neodpovídá modelu měření statistické závislosti, je cesta adaptace charakteristiky χ^2 na koeficient statistické závislosti dosti libovolná. Výsledkem různých adaptací je dnes celá řada různých normalizací této charakteristiky (Pearsonův normalizovaný koeficient, Čuprovův koeficient, výše uvedený Cramérův koeficient a další). Libovolnost zavedení konečného tvaru a absence základního modelu vede k neurčené interpretaci. Zatímco ostatní v článku uvedené míry mají interpretaci přirozenou a jejich numerické hodnoty mají jednoznačný obsah, chí-kvadrátové míry sice slouží k orientaci v datech, ale obsah jim přiřazovat nelze. Proto je také nesprávné nasazovat je v dalších složitějších modelech, do nichž musí vstupovat míry s přesně určeným obsahem (aby tak byly zajištěny například návaznosti měr v kauzálních sítích apod.).

Nedostatek klasických měr je patrný také z toho, že neodpovídají požadavkům asymetričnosti a symetričnosti a že tyto dva případy nerozlišují.

Chí-kvadrátové míry jsou dnešním vývojem matematické statistiky překonány; jistě splnily svoji roli ve vývoji statistického myšlení, v dnešní době by ale bylo vhodné nahradit je měrami, které odpovídají moderním požadavkům metodologie práce s daty a jsou prověřeny jak matematickostatistickou teorií, tak praxí výzkumů.

Důvodem pro jejich zachovávání nemůže být to, že jsou (nebo byly) běžně v našich výzkumech používány a že je třeba je udržet pro srovnatelnost — v našich výzkumech sice používány byly, ale v tak různých nejednotných variantách, že srovnatelnost hodnot stejně není možná.

Ani argument, že vystupují ve většině současných programů, neobstojí; je totiž velmi snadné doprogramovat nebo připrogramovat vzorec, neboť východisko (kontingenční tabulka) zůstává stejné.

5. Numerické příklady[10]

Při sociologickém šetření postojů (viz poznámka 10) byly položeny (mimo jiné) například otázky, které vyústily v následující znaky:

A. Podle názoru respondenta je modernizace a přestavba historické části města
   1 — nezbytná,
   2 — mohla by počkat,
   3 — respondent neví, nerozumí těmto věcem.

B. Podle názoru respondenta by měla přestavba probíhat tak,
   1 — aby se většina obyvatel mohla vrátit do této čtvrti,
   2 — že by se většina obyvatel měla stěhovat jinam,
   3 — respondentovi na tom nezáleží,
   4 — respondent neví.

C. Respondent říká za sebe a svou rodinu, že po skončení restaurací bytového fondu:
   1 — by se chtěli v každém případě vrátit do svého domu,
   2 — by chtěli bydlet v téže čtvrti (bez ohledu na dům),
   3 — by dali přednost jiné čtvrti,
   4 — by se odstěhovali mimo Tábor,
   5 — je pro jiné řešení (např. dům důchodců),
   6 — neví, je mu to jedno.

D. Nejvyšší dosažené vzdělání respondenta:
   1 — nedokončené základní,
   2 — základní,
   3 — základní a vyučen,
   4 — střední bez maturity,
   5 — střední všeobecné s maturitou,
   6 — střední odborné s maturitou,
   7 — vysokoškolské.

Především si uvědomíme, že koeficienty spočítané pro uvedené soubory respondentů (viz níže), jsou výběrové koeficienty pro celou populaci, tzn., že bychom měli s koeficientem počítat i jeho konfidenční interval, abychom mohli výsledek zobecnit. Tento problém jsme však neřešili a nebudeme se jím zabývat ani zde.

V tabulkách jsou sloučeny některé kategorie; zajímá nás například vztah k historické části města (taková úprava kategorií se dělá často: data se získávají pro jemnější znak — podrobnější dělení kategorií — který se pak účelově zúží, někdy i několika způsoby). Respondenti, kteří neodpověděli ale-

[10] Oba příklady jsou vzaty z *výzkumů postojů obyvatel historického jádra města Tábor k otázkám rekonstrukce, jehož autory jsou J. Linhart a M. Matějů.* J. Linhartovi zde chceme vyjádřit poděkování za laskavé poskytnutí dat, pomoc při jejich výběru a průběžný zájem o tento text. Slučování kategorií bylo provedeno autory.

spoň na jednu otázku, nebyli do tabulky zařazeni.

Tabulka 2 je příkladem *asymetrického vztahu* „vzdělání — názor na modernizaci města".

Tabulka 2.

Nutnost přestavby

D	A			Celkem
	1	2	3	
(1, 2)	/133	32	24	189
(3, 4)	/129	26	13	168
(5, 6, 7)	/58	6	0	64
Celkem	/320	64	37	421

Vzdělání

A. Výpočet $\lambda_{A/D}$:

1. Zatrhneme maxima v řádcích

$$n_{1m} = 133 \qquad n_{2m} = 129 \qquad n_{3m} = 58$$
$$n_{.m} = 320 \qquad n = 421$$

2. Spočteme

$$\Sigma n_{im} = 133 + 129 + 58 = 320$$

3. Dosadíme do vzorce (2)

$$\lambda_{A/D} = \frac{\sum n_{im} - n_{.m}}{n - n_{.m}} = \frac{320 - 320}{421 - 320} = 0$$

B. Výpočet $\tau_{A/D}$:

1. V každém řádku nalezneme součet čtverců absolutních četností, $\Sigma\Sigma n_{ij}^2$; tato čísla dělíme příslušnými řádkovým součty:

$$\frac{1}{n_1} \sum n_{1j}^2 = (133^2 + 32^2 + 24^2) : 189 =$$
$$= 102{,}058$$

$$\frac{1}{n_2} \sum n_{2j}^2 = (129^2 + 26^2 + 13^2) : 168 =$$
$$= 104{,}083$$

$$\frac{1}{n_3} \sum n_{3j}^2 = (58^2 + 6^2 + 0^2) : 64 = 53{,}125$$

2. Součet těchto čísel (259,266) násobíme n:

$$259{,}266 \times 421 = 109\,150{,}986 \doteq 109\,151$$

3. Součet čtverců marginálního řádku:

$$\Sigma n_{.j}^2 = 320^2 + 64^2 + 37^2 = 107\,865$$

a

$$n^2 = 421^2 = 177\,241$$

4. Dosadíme do vzorce (6):

$$\tau_{A/D} = \frac{109\,151 - 107\,865}{177\,241 - 107\,865} = \frac{1\,286}{69\,376} = 0{,}0185$$

C. Výpočet informační míry:

1. Nejprve získáme příslušné hodnoty *n log n* pro dosazení do vzorce (11); zde je pro názornost vypisujeme do tabulky (ve výpočtech pak tato čísla přímo načítáme na kalkulačce):

2. Z těchto čísel získáme $\Sigma n_{ij} \log n_{ij}$ jako součet hodnot ve vnitřní části tabulky. Tato hodnota se rovná 1828,8201.

$\Sigma n_i \log n_{i.}$ = součet v posledním sloupci (bez pravého dolního rohového pole tabulky, která odpovídá hodnotě n) = 2117,6846.

$\Sigma n_{.j} \log n_{.j}$ = součet posledního řádku (opět bez pravého dolního rohového pole tabulky) = 2245,6352.

n log n je v pravém dolním rohovém poli: 2543,9484

3. Dosadíme do vzorce (11)

$$\zeta_{D/A} = 1 - \frac{2117{,}6846 - 1828{,}8201}{2543{,}9484 - 2245{,}6352} =$$
$$= 1 - 0{,}9683 = 0{,}0317$$

Tabulka 3.

650,4164	110,9035	76,2733	990,6902
626,9158	84,7105	76,2733	721,6264
235,5057	10,7506	0,0000	266,1685
1 845,8627	266,1685	133,6040	2 543,9484

Tabulka 4.

B-Názor na návrat ostatních obyvatel	C — Přání nastěhovat se zpět				Součet
	1	2	3	(4, 5, 6)	
1	/198×	31×	20	15	264×
2	12	8	/21	6	47
(3, 4)	/38	25	24×	20×	107
Součet	/248	64	65	41	418

Všechny míry mají nízké hodnoty. Názory na přestavbu jsou na vzdělání statisticky velmi málo závislé; při další interpretaci zjistíme modální volbu společnou všem skupinám — tato kategorie je dominantní.

Příklad symetrické asociace (viz tabulku 4)

A. Výpočet λ:

1. Zatrhneme řádková maxima v tab. č. 4 (|_)

$$n_{1m} = 198 \quad n_{2m} = 21 \quad n_{3m} = 38$$
$$\Sigma n_{im} = 257$$

2. Křížkem označíme sloupcová maxima (×)

$$n_{m1} = 198 \quad n_{m2} = 31 \quad n_{m3} = 24$$
$$n_{m4} = 20 \quad \Sigma n_{mj} = 273$$

3. Zjistíme maxima u marginálních rozložení

$$n_{m.} = 264 \quad n_{.m} = 248$$

4. Dosadíme do vzorce (13)

$$\lambda = \frac{257 + 273 - 264 - 248}{2 \times 418 - 264 - 248} = \frac{18}{324} =$$
$$= 0{,}05555 \doteq 0{,}0556$$

B. Výpočet τ:

Postup asymetrického případu aplikujeme dvakrát, a to jednou ve směru řádků a poté ve směru sloupců:

1. a) Spočítáme výrazy:

$$\frac{1}{n_{1.}} \sum n_{1j}^2 = 40\,790 : 264 = 154{,}51$$

$$\frac{1}{n_{2.}} \sum n_{2j}^2 = 685 : 47 = 14{,}57$$

$$\frac{1}{n_{.3}} \sum n_{3j}^2 = 3\,045 : 107 = 28{,}46$$

b) Obdobně:

$$\frac{1}{n_{.1}} \sum n_{i1}^2 = 164{,}48 \qquad \frac{1}{n_{.2}} \sum n_{i2}^2 = 25{,}78$$

$$\frac{1}{n_{.3}} \sum n_{i3}^2 = 29{,}49 \qquad \frac{1}{n_{.4}} \sum n_{i4}^2 = 16{,}12$$

c) Součet výrazů ad a) a ad b) je roven $197{,}54 + 235{,}87 = 433{,}41$

2. Nalezneme součet čtverců marginálních rozložení

a) u součtového řádku (veličiny Y):

$$\Sigma n_{.j} = 71\,506$$

b) u součtového sloupce (veličiny X):

$$\Sigma n_{i.} = 83\,354$$

c) dohromady součet marginálních čtverců:

154 860

3. Výsledek dosadíme do vzorce (17):

$$\tau = \frac{418 \times 433{,}41 - 154\,860}{2 \times 174\,724 - 154\,860} =$$
$$= \frac{26\,305{,}38}{194\,588} = 0{,}1352$$

C. Výpočet ζ:

1. K výpočtu informační míry musíme získat nejprve příslušné hodnoty n log n; opět je pro názornost vypíšeme do tabulky č. 5.

2. Z tabulky 5 zjistíme $\Sigma\Sigma n_{ij} \log n_{ij} = 1730{,}0940$

Tabulka 5.

1047,0759	106,4536	59,9146	40,6208	1472,0506
29,8189	16,6355	63,9350	10,7506	180,9569
138,2283	80,4719	76,2733	59,9146	499,9927
1367,3303	266,1685	271,3351	152,2565	2522,8312

$\Sigma n_{.j} \log n_{.j} = 2057{,}0904$
$\Sigma n_{i.} \log n_{i.} = 2153{,}0002$
$n \log n = 2522{,}8312$

3) Uvedené dílčí výsledky dosadíme do vzorce (21):

$$\xi = 2\left[1 - \frac{2522{,}8312 - 1730{,}0940}{2 \times 2522{,}8312 - (2153{,}0002 + 2057{,}0904)}\right] = \underline{0{,}1028}$$

D. Cramérův koeficient Cr:

1. Nejprve spočítáme chí-kvadrát podle vzorce (25); přitom ovšem nám stačí získat hodnotu v závorce, neboť n se ve vzorci (26) zkrátí: $\chi^2 = n\,(1{,}2161 - 1)$

2. $Cr = \dfrac{1{,}2161 - 1}{2} = \underline{0{,}1080}$

6. Závěry

V článku jsme se nezabývali problematikou konfidenčních intervalů, ani jsme neuváděli testy hypotéz; vzorce a výpočty jsou pro tyto úlohy dost obtížné a jen zřídka je budeme počítat ručně. Zájemce odkazujeme na uvedenou statistickou literaturu (Goodman, Kruskal [4], Zvárová [9]) a na odbornou statistickou konzultaci.

Pro nedostatek místa jsme se také nezabývali případem čtyřpolních tabulek, u nichž se vzorce pro jednotlivé míry značně zjednodušují a v různých případech i ztotožňují. Problém čtyřpolních tabulek však zasluhuje samostatné zhodnocení, protože používáme různé modely, odpovídající různým pozadím dichotomizace znaků.

Při aplikaci zásadně odlišujeme symetrické a asymetrické vztahy. Doporučujeme nepoužívat chí-kvadrátové normované míry. Podle našeho názoru a zkušeností je pro rychlou analýzu a orientaci v datech nejlepší koeficient λ, pro výpočet na počítači však vždy volíme τ nebo ζ. Výběr mezi těmito koeficienty je vcelku libovolný. Jak predikční model, tak model teorie informace je heuristicky snadno pochopitelný a oba vyhovují našim požadavkům. Koeficienty vycházející z teorie informace mají však dříve uvedenou výhodu, totiž že při možnosti použití tabulek jsou snadno spočitatelné. To je důležité obzvláště v případech, kdy provádíme významové úpravy tabulek (slučujeme kategorie), přepočítáváme jednotlivé míry spočítané pro původní tabulku počítačem a potřebujeme zaručit srovnatelnost s jinými tabulkami či s tabulkou původní. U těchto koeficientů jsou zajímavé také mezivýpočty, které lze použít i pro jiné úlohy analýzy dat.

Asymetrické a symetrické koeficienty stejného typu jsou srovnatelné, protože jsou srovnatelné i jejich modely, z nichž plyne věcný smysl koeficientů. Je však samozřejmé, že nelze srovnávat τ u jedné tabulky a ζ u druhé tabulky, ať jde o stav symetrický nebo asymetrický.

V praxi většinou při výpočtech vynecháváme kategorie „odmítnutí odpovědi", „chybějící informace" atd. a počet jednotek vstupujících do tabulky tak snižujeme.

Všechny asymetrické míry, které jsme zde uváděli, mohou sloužit také pro účely komparace (v úloze jednoduché stratifikace), tj. srovnávání dílčích souborů. Malé hodnoty měr značí vysokou podobnost statistických rozložení znaků v dílčích souborech, vysoké hodnoty velkou nepodobnost (heterogenitu). Koeficient λ může sloužit také jako charakteristika pro testování hypotézy, která říká, že všechny dílčí soubory mají stejnou modální kategorii.

Literatura

[1] Fano, R. M.: *Transmission of Information.* New York, Wiley 1961.
[2] Goodman, L. A. - Kruskal, W. H.: *Measures of Association for Cross Classifications.* Journal of American Statistical Association 49, 1954, s. 732—764.
[3] 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.
[4] Goodman, L. A. - Kruskal, W. H.: *Measures of Association for Cross Classifications III: Approximate Sampling Theory.* Journal of American Statistical Association 54, 1963, s. 310—364.
[5] Jasin, Je. G.: *Těorija informacii i ekonomičeskije issledovanija.* Moskva, Statistika 1970.
[6] Kullback, S.: *Information Theory and Statistics.* New York, Wiley 1958. Ruský překlad: *Těorija informacii i statistika.* Moskva, Nauka 1967.
[7] Linfoot, E. H.: *An Informational Measure of Correlation.* Information and Control 1, 1957, s. 85—89.
[8] Nikl, J., Perez, A.: *Aplikace metod teorie informace při studiu závislosti v biologických systémech.* Československá fyziologie 10, 1961, s. 454—460.
[9] Zvárová, J.: *Asymptotická distribuce výběrové informační míry závislosti.* Kybernetika 5, 1969, s. 50—59.

Резюме

Ржегак Я. - Ржегакова Б.: Измерение статистической зависимости номинальных признаков

В анализе социологических данных существенную роль играет понятие статистической зависимости между отдельными переменными. Статья вводит в проблематику измерения статистической зависимости номинальных признаков (ассоциация качественных переменных). Не дискутируется отношение статистической зависимости и каузальной реальной зависимости; авторы сосредоточивают свое внимание на моделях, которые приводят к измерению статистической зависимости.

Во второй части высказываются общие требования к мерам статистической зависимости и описывается принцип, который может служить основой для отдельных мер. Третья часть вводит модель для измерения ассиметрического статистического отношения. Этот принцип ввел Гуттман как относительное уменьшение вероятности ошибки предикции переменной Y при знании величины переменной X в сравнении с вероятностью ошибки предикции Y без знания X .Это отношение, примененное к конкретной контингентной таблице, является мерой статистической зависимости ассиметрического типа. Если мы в эту модель поставим модальное предикционное правило, мы получим формулы (1), (2); λ Гуттмана является хорошей мерой для быстрых анализов, оно легко поддается подсчитанию, однако у него тот недостаток, что нулевая величина коэффициента не эквивалентна состоянию статистической независимости, но только состоянию тождества модальных категорий отдельных субпопуляций. Пропорциональное предикционное правило, примененное в общей формуле (3), приводит к формулам (4), (5), (6). Коэффициент Валлиса τ уже является превосходной мерой, которая выполняет все априори поставленные требования.

Если обобщить принцип (3) в выраженные другим образом неопределенности в данных, то мы получим общую формулу (7). При поставлении значений энтропии (понятие теории информации) мы получим формулы (8)—(11) для ζ. Этот коэффициент выполняет одинаково хорошо все требования к статистическим мерам зависимости номинальных признаков. Сверх того он обладает тем преимуществом, что довольно просто его можно подсчитать, если в нашем распоряжении есть таблицы значений к log к и для естественных к или p log p для p из интервала (0,1).

Четвертая часть вводит принцип симметризации предшествующих моделей на случай, что мы не в состоянии определить направление ассиметрического отношения. Этот принцип приводит к симметризированному λ [формулы (12), (13)] и к симметризированному τ [формулы (14)—(17)]; авторы предлагают также аналогичную симметризацию для коэффициента ζ [формулы (18)—(21)]. Преимущество последнего коэффициента заключается опять в простой возможности подсчитания при наличии доступных соответствующих таблиц. Преимущества и недостатки симметризированных мер аналогичны как и у мер ассиметрических.

Классицческие chi-квадратные меры в статье представлены только коэффициентом Крамера (26). Но они для исследовательских целей не рекомендуются, так как они не имеют никакой основной модели, которая давала бы возможность их толковой, одинаковой и также простой интерпретации.

Пятая часть содержит и численные примеры для ассиметрического и симметрического отношения переменных. Ни тесты гипотез о коэффициентах ни их доверительные и нтервалы не приводятся.

Summary

Řehák J. - Řeháková B.: Measurement of Statistical Dependence Among Nominal Variables

The notion of statistical dependence (association) among individual variables plays a substantial part in the analysis of sociological data. The paper introduces into problems of measuring the statistical dependence of nominal variables (association of qualitative variables). The relation between statistical dependence and causal real dependence is not discussed; the authors pay attention to models leading to the measurement of statistical dependence.

Part 2 presents general demands on the measures of statistical dependence and de-

scribes the principle that may form the basis of the separate measures.

Part 3 introduces a model for measuring the asymmetrical statistical relation. This principle was introduced by Guttman as the relative decrease of the probability of error in predicting the variable Y if the value of the variable X is known, relative to the probability of error in predicting Y if X is not known. This relation applied to the concrete contingency table is the measure of statistical dependence of the asymmetric type. If the modal prediction rule is introduced into this model, we get the formulas (1), (2); Guttman's λ is a good measure for quick analyses, it is easily computable; its disadvantage, however, lies in the fact that the zero value of the coefficient is not equivalent to the statistical independence but only to the case in which the modal categories of individual subpopulations coincide. The proportional prediction rule applied in the general formula (3) leads to the formulas (4), (5), (6). Wallis' coefficient τ then represents a perfect measure fulfilling all a priori demands.

If principle (3) is generalized to differently expressed uncertainties in the data, the general formula (7) is obtained. By substituting the values of entropy (a concept of the theory of information), we get the formulas (8) — (11) for ζ. This coefficient equally well fulfils all the demands on the statistical measures of dependence of nominal variables. Moreover, it has the advantage that it may be obtained by rather simple computation if the tables of values k log k (for natural k) or p log p (for p from the interval (0,11 are available.

Part 4 introduces the principle of the symmetrization of the preceding models for the case of our inability to determine the direction of the asymmetrical relation. This principle leads to a symmetrized λ [formulas (12), (13)] and to a symmetrized τ [formulas (14)—(17)]; the authors also suggest a similar symmetrization for the coefficient ζ [formulas (18), (21)]. Again, the advantage of the last coefficient lies in a simple computability if the respective tables are available. The advantages and disadvantages of symmetrized measures are analogous to those of asymmetrical measures.

In the present paper, the classic χ^2 measures are represented only by Cramér's coefficient (26). They are, however, not recommended for research purposes, since they have no basic model facilitating their reasonable, unique and simple interpretation.

Part 5 contains numerical examples illustrating asymmetrical and symmetrical relations of the variables. Tests of hypotheses on coefficients and their confidence intervals are not mentioned.