Analýza kontingenčních tabulek: rozlišení dvou základních typů úloh a znaménkové schéma

JAN ŘEHÁK
Ústav pro filozofii a sociologii ČSAV, Praha
BLANKA ŘEHÁKOVÁ
Ústav pro výzkum veřejného mínění při FSÚ, Praha

Kontingenční tabulka patří k nejzákladnější metodologické výbavě každého sociologa pracujícího s daty. Představuje také nejčastější způsob uspořádání dat, který dostáváme z počítače a který prezentujeme ve výzkumných zprávách. O kontingenčních tabulkách či o *vyhodnocení třídění druhého stupně* bylo napsáno už mnoho, přesto při práci s nimi často chybujeme. Účelem této stati je upozornit na dva základní typy tabulky, především pak na znaménkové schéma.
Znaménkové schéma v kontingenčních tabulkách je jedním z velmi vhodných grafických prostředků pro analýzu typu vztahů v kontingenční tabulce a patří k nejpraktičtějším inovacím naší sociologické metodologie. Bylo zavedeno v práci Linhart, Šafář [1967] a je založeno na intuitivním testování shody pozorované a očekávané četnosti v každém poli kontingenční tabulky. Na základě tohoto testování pak tiskne počítač schéma znamének (+++, ++, +, 0, −, −−, −−−). Znaménka + znamenají, že pozorované četnosti významně převyšují očekávané četnosti, znaménka −, že pozorované četnosti jsou nižší než očekávané a 0 značí nevýznamný rozdíl. Počet znamének je odstupňován podle významnosti rozdílu, např. „+" značí významnost na 0,05, „++" na 0,01 a „+++" na 0,001 (obdobně „−"). Tyto hranice jsou zvoleny podle přání uživatele (resp. zadavatele programu), jejich volba je určitou konvencí a zvyklostí. Autoři schématu [Linhart, Šafář 1967] zavedli pro testování intuitivní statistiku, o které předpokládají, že má asymptoticky normální rozložení a tak má i známé vlastnosti tzv. *z*-skóre.

Označme n_{ij} absolutní četnost v tabulce o rozměrech $R \times C$, n_{i+} je řádkový součet, n_{+j} je sloupcový součet, n je celkový počet pozorování[1]. Linhart, Šafář [1967] navrhují tuto statistiku

$$(1) \qquad z_{ij} = \frac{n_{ij} - \frac{n_{i+}n_{+j}}{n}}{\sqrt{n}\sqrt{\frac{n_{i+}}{n}\frac{n_{+j}}{n}\left(1 - \frac{n_{i+}n_{+j}}{n^2}\right)}}$$

$$(2) \qquad = \sqrt{n}\,\frac{nn_{ij} - n_{i+}n_{+j}}{\sqrt{n_{i+}n_{+j}(n^2 - n_{i+}n_{+j})}}$$

Výhody grafického přehledu, rychlá orientace a možnost pohodlně hodnotit kvalitativní strukturu vztahu dvou proměnných pomocí systému znamének, našly velkou řadu uživatelů. Myšlenka se rychle rozšířila do různých programů s různými modifikacemi a chybami. Účelem této stati je zavést správné vzorce pro tuto me-

[1]) Toto je běžné značení v kontingenčních tabulkách. Platí: $n_{i+} = \sum_{j=1}^{C} n_{ij}$, $n_{+j} = \sum_{i=1}^{R} n_{ij}$, $n = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}$

todu a tak ji revidovat i současně sjednotit různé existující přístupy. Výsledky zakládáme na statistické teorii asymptotických testů, tj. testů pro velké výběry (viz Rao [1968], zvláště kapitola 6). Uvádíme zde nejprve dva nejčastější typy kontingenčních tabulek. Ukazujeme, že pro oba diskutované typy úloh je formální výpočet znaménkového schématu i testovacích charakteristik stejný, i když testy základních hypotéz a znaménkové schéma mají různé významy. Matematická argumentace a technické aspekty jsou obsahem poslední části stati a některých poznámek, které jsou určeny pro specialisty, zatímco text je určen pro práci při vlastní analýze dat.

Uveďme ještě další značení:

$f_{j/i} = \frac{n_{ij}}{n_{i+}}$ = řádková relativní četnost

$f_{i+} = \frac{n_{i+}}{n}$ = relativní četnost pro zastoupení řádkové proměnné v celém souboru dat (marginální distribuce odpovídající součtovému sloupci tabulky)

$f_{+j} = \frac{n_{+j}}{n}$ = relativní četnost pro zastoupení sloupcové proměnné v celém souboru dat (marginální distribuce odvozená ze součtového řádku tabulky)

$f_{ij} = \frac{n_{ij}}{n}$ = relativní četnost obsazení pole vzhledem k celému souboru dat

$p_{j/i}$, p_{i+}, p_{+j}, p_{ij} mají obdobný význam, ale odpovídají populačním rozložením, zatímco f odpovídají výběrovým hodnotám.

1. Úloha srovnání několika souborů — typ A

Data uspořádaná v kontingenční tabulce reprezentují nezávislé základní soubory, které chceme komparovat z hlediska nominálního znaku. Je dáno R takových souborů, nominální znak má C hodnot. Každý řádek odpovídá jednomu souboru a obsahuje rozložení četností, které vznikly prostým náhodným výběrem[2] z daného souboru.

Příklady:

a) soubory: země, dělníci vybraných závodů; znak: typ vztahu dělníka k práci;
b) soubory: kraje, dospělé obyvatelstvo; znak: typ čtenářské aktivity vzhledem k týdeníkům;
c) soubory: mládež v Praze, ostatní mládež v ČSR; znak: typ trávení volného času.

Tento případ je velmi častý. Jeho *charakterizace* spočívá v tom, že:

a) data v řádcích vznikají nezávisle,
b) řádkový součet je předem dán velikostí výběru v daném souboru, nevzniká tedy náhodně jako součást náhodných mechanismů výběru;
c) velikost výběrových souborů nemusí být proporcionální k velikostem základních souborů;
d) společné rozložení četností není specifikováno.

Hypotéza: Všechny základní soubory mají stejná rozložení, odchylky ve výběrových souborech můžeme vysvětlit náhodným působením výběru či náhodnými vlivy při genezi jednotlivých údajů (chyba měření nebo proces vzniku souboru).

[2]) Tento předpoklad bývá většinou v sociologii porušen. Proto pravděpodobnostní implikace metody nemohou být chápány doslova, ale slouží spíše jako vodítko pro intepretaci. Metodu lze u obou typů A i B ovšem použít i za předpokladu, že šetříme celé základní soubory, ale předpokládáme, že tyto soubory vznikly náhodným procesem (např. výběr lidí do dílen určitého typu).

Formálně můžeme tedy hypotézu formulovat jako:

(3) $p_{j/i} = p_{+j}$ pro všechna i a j

(existuje společná distribuce u všech R základních souborů reprezentovaných tabulkou); čísla p_{+j} nejsou specifikována.

Alternativní hypotéza: Alespoň jeden soubor se liší svou distribucí od ostatních, a to alespoň u relativní četnosti jedné z hodnot znaku; alespoň jeden řádek reprezentuje populační rozložení četností rozdílné od ostatních.
Alternativní hypotéza tedy zahrnuje všechny případy lišící se od případu vzájemné shody rozložení. Jde tedy o *omnibusovou* (*obecnou*) alternativní hypotézu.

Příklad 1.

Srovnávací tabulka rozložení typů trávení volného času mládeže (typy *A*, *B*, *C*, *D*) pro dva soubory: I. mládež ČSR (mimo Prahu) II. mládež s pobytem v Praze.

Tabulka 1. Absolutní četnosti a řádková procenta (v závorce)[3]

	Typ				Velikost souboru	Poměr zastoupení souborů
	A	B	C	D		
Soubor I	236 (20,8 %)	370 (32,7 %)	384 (33,9 %)	143 (12,6 %)	1133	56,8 %
Soubor II	159 (18,4 %)	320 (37,1 %)	232 (26,9 %)	151 (17,5 %)	862	43,2 %
Celkem dva soubory	395 (19,8 %)	690 (34,6 %)	616 (30,9 %)	294 (14,7 %)	1995	

2. Úloha testování nezávislosti dvou nominálních znaků — typ B

Data uspořádaná v kontingenční tabulce pocházejí z jednoho prostého náhodného výběru, řádky a sloupce odpovídají dvěma různým nominálním znakům, jejichž statistickou nezávislost chceme testovat. Tento případ ještě může být specifikován analyticky tak, že nás buď zajímá vzájemná statistická souvislost (symetrický statistický vztah), nebo asymetrická závislost jednoho znaku na druhém. Na rozdíl od zjišťování koeficientů souvislostí, které se počítají v obou případech jinak (viz např. [Řehák, Řeháková 1973]), testy nezávislosti se počítají zcela stejným způsobem. Rozlišení obou případů je interpretační, věcné, nikoli statistické. Příklady známe z běžné výzkumné praxe.

Příklady :

a) znaky: pohlaví, typ prožívání volného času; populace: mládež ČSR;
b) znaky: forma účasti na řízení, obsahové zaměření účasti na řízení; populace: dělníci potravinářského průmyslu;
c) znaky: typ organizace práce, typ interpersonálních vztahů; populace: výzkumné kolektivy.

[3]) Pro ilustraci můžeme ukázat značení, např. absolutní četnosti: $n_{11} = 236$, $n_{1+} = 1133$, $n_{+1} = 395$, $n = 1995$; procenta: $100 . f_{2/1}\ \% = 32{,}7\ \%$, $100 . f_{4/2}\ \% = 17{,}5\ \%$, $100 . f_{+3}\ \% = 30{,}1\ \%$, $100 . f_{1+}\ \% = 56{,}8\ \%$, $100 . f_{2+}\ \% = 43{,}2\ \%$.

Charakteristika případu:

a) Data celé tabulky vznikají jako realizace jednoho náhodného výběru[4] z jedné populace.
b) Celkový součet n je předem dán, ani řádkové, ani sloupcové součty nejsou ve výběru určeny.
c) Očekáváme, že řádkové i sloupcové absolutní četnosti budou v rámci povolené statistické tolerance proporčně odpovídat situaci v základním souboru.
d) Ani řádkové, ani sloupcové marginální rozložení není specifikováno.

Hypotéza: Oba znaky spolu v dané populaci nesouvisí, proto i jejich empirické výskyty odpovídají statistické nezávislosti. Formálně můžeme tuto hypotézu formulovat tak, že

(4) $p_{ij} = p_{i+} \cdot p_{+j}$ pro všechna i, j
p_{i+}, p_{+j} nejsou specifikovány.

Alternativní omnibusová hypotéza: Alespoň u jedné dvojice hodnot (i, j) je statistická nezávislost porušena.

(5) $p_{ij} \neq p_{i+} \cdot p_{+j}$

Zahrnuje tedy všechny možné případy nejrůznějších typů závislostí.

Příklad 2

Pro populaci mládeže ČSR (mimo Prahu) zkoumáme souvislost výskytu dvou typologií (trávení volného času: typy A, B, C, D a zájmové orientace: typy a, b, c). Společné (sdružené) rozložení absolutních a relativních četností je dáno v tabulce 2.

Tabulka 2. Sdružené absolutní a relativní rozložení dvou znaků [5]

		Typ				Celkem
		A	B	C	D	
Typ	a	100 (9,1 %)	32 (2,9 %)	151 (13,8 %)	124 (11,3 %)	407 (37,1 %)
	b	51 (4,6 %)	16 (1,5 %)	103 (9,4 %)	40 (3,6 %)	210 (19,1 %)
	c	118 (10,8 %)	73 (6,7 %)	159 (14,5 %)	130 (11,9 %)	480 (43,8 %)
Celkem		269 (24,5 %)	121 (11,0 %)	413 (37,6 %)	294 (26,8 %)	1097 (100 %)

3. Testování nulové hypotézy pro oba případy (typ A i B)

I když se úlohy typu **A** a **B** liší svým analytickým cílem a zasazením do sociologického kontextu interpretace dat, testování obou nulových hypotéz provádíme for-

[4] I zde platí požadavek prostého náhodného výběru a poznámka 2.
[5] Jako u předchozí tabulky ilustrujeme značení: $n_{23} = 103$, $n_{2+} = 210$, $n_{+3} = 413$, $n = 1097$, $100 \cdot f_{23}$ % $= 9{,}4$ %, $100 \cdot f_{2+}$ % $= 19{,}1$ %, $100 \cdot f_{+3}$ % $= 37{,}6$ %.

málně zcela stejně tak, že spočítáme známou[6] statistiku *chí*-kvadrát (X^2). Zde pro přehlednost i praktické cíle uvádíme celou řadu ekvivalentních vzorců[7]:

$$(6) \qquad X^2 = \sum\sum \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}$$

$$(7) \qquad = \frac{1}{n} \sum\sum \frac{(nn_{ij} - n_{i+}n_{+j})^2}{n_{i+}n_{+j}}$$

$$(8) \qquad = n \left(\sum\sum \frac{{n_{ij}}^2}{n_{i+}n_{+j}} - 1\right)$$

$$(9) \qquad = n \left(\sum\sum \frac{n_{ij}f_{j/i}}{n_{+j}} - 1\right)$$

$$(10) \qquad = n \left(\sum\sum \frac{{f_{ij}}^2}{f_{i+}f_{+j}} - 1\right)$$

(11) Počet stupňů volnosti $\text{d.f.} = (R - 1)(C - 1)$

Postup testu:

1. Určíme nulovou hypotézu H_0 a hladinu významnosti α.
2. Spočítáme X^2 a určíme d.f.
3. Nalezneme kritickou hodnotu $\chi^2_{\alpha;\text{d.f.}}$ ve statistických tabulkách.
4. Provedeme rozhodnutí.

(12) Zamítneme H_0, jestliže $X^2 \geqq \chi^2_{\alpha;\text{d.f.}}$

V opačném případě nemáme důvod k zamítnutí. Alternativně můžeme postupovat také tak, že k hodnotě X^2 a k d.f. nalezneme její hladinu významnosti α^* a podle její výše určíme, zda hypotézu přijmeme či odmítneme, tj. je-li $\alpha^* \leqq \alpha$ (předem zvolené), nulovou hypotézu zamítneme. Tento postup je v poslední době běžnější u počítačových výstupů (je jednodušší jej naprogramovat a výstup slouží všem uživatelům bez rozdílu jejich volby α).

---

[6] Test je založen na standardizovaném porovnání pozorovaných a očekávaných četností v jednotlivých polích podle Pearsonova principu:

$$X^2 = \Sigma \frac{(\textit{pozorovaná četnost} - \textit{očekávaná četnost})^2}{\textit{očekávaná četnost}}$$

Vypočtenou statistiku značíme X^2 a porovnáváme ji s teoretickou distribucí *chí*-kvadrátu pro daný počet stupňů volnosti.

[7] Uvádíme různé vzorce, které jsou ovšem matematicky ekvivalentní. Jejich výběr záleží na několika faktorech: a) na typu dat, které máme k dispozici z počítače nebo z nějaké publikace (absolutní četnosti, relativní četnosti, či různé kombinace); b) na typu programování, resp. typu kapesní kalkulačky a její logiky; c) na osobní preferenci uživatele. Uvedené vzorce nereprezentují všechny úpravy. Někdy je výhodné si vzorce upravit podle toho, k čemu bude třeba dále využít mezivýsledků apod. Při výpočtech se řídíme zásadami: a) Bez ohledu na to, že výsledky publikujeme zaokrouhleně a zjednodušeně, udržujeme mezivýsledky na nejvyšší možné numerické přesnosti. b) Vybíráme postupy, kde provádíme co nejméně zaokrouhlování během výpočtů (týká se především dělení). c) Výpočet provádíme pokud možno kontinuálně, s tím, že mezivýsledky si uchovávámе v pamětích. d) Zásadně všechny výpočty provádíme dvakrát, pokud možno dvěma různými osobami a nezávisle.

Omezení při použití tohoto testu:[8]

1. Žádná očekávaná hodnota by neměla být menší než 1; je-li třetina nebo čtvrtina očekávaných hodnot mezi 1 až 5, lze test použít (Lancaster [1969]).
2. Tabulky $2 \times C$ (resp. $R \times 2$) lze testovat tímto testem, pokud jsou všechny očekávané hodnoty větší nebo rovné 1; i toto pravidlo je však konzervativní, hranici lze snížit až na 0,5 (Lewontin, Felsenstein [1965]).

Příklad 2 (pokračování)

Statistiku X^2 spočteme například podle vzorce (3) [9]

$$X^2 = 1097\left[\frac{100^2}{407 \cdot 269} + \frac{32^2}{407 \cdot 121} + \ldots + \frac{130^2}{480 \cdot 294} - 1\right]$$
$$= 1097[1.027507345 - 1]$$
$$= 30.18$$
$$R = 3, \quad C = 4, \quad \mathrm{df} = (R-1)(C-1) = 6$$

Významnost můžeme zjistit například v Janko [1958: tabulka 4, 5]. X^2 je vysoce významné, x je menší než 0.0001.

4. Znaménkové schéma[10]

V případě, že zamítáme nulovou hypotézu, zajímá nás struktura závislosti (typ **B**) či rozdílnosti souborů (typ **A**). K tomu můžeme použít postup testování odchylky pozorovaných a očekávaných četností v jednotlivých polích tabulky. I když používáme postup testování hypotéz, nejde nám o žádnou specifickou hypotézu — jde nám o to, jak specifikovat přijatou obecnou alternativu existence rozdílnosti či závislosti. Proto je následující postup explorativní.

Přesto, že oba typy **A** i **B** odpovídají různým meritorním i statistickým situacím, znaménkové schéma (stejně jako test) budeme určovat formálně v obou případech

[8]) Názory na aplikabilitu testu se značně liší. Ve srovnání se základní prací (Cochran [1954]) se požadavky značně uvolnily. Konečné slovo zatím vysloveno nebylo. Závěrečné hodnocení bude zřejmě založeno na dalších extenzívních simulacích kontingenčních tabulek metodou Monte Carlo a na zkoumání takto získaných empirických distribucí testových charakteristik. Stejně tak se liší i názory na modifikaci testu při malých výběrech použitím tzv. korekcí pro spojitost pro tabulky 2×2. Pirie, Hamdan [1972] uvádějí zlepšení korekce pro spojitost pro typ A i B kontingenčních tabulek, Grizzle [1967] doporučuje korekci pro spojitost nepoužívat vzhledem k tomu, že přispívá ke konzervativnosti testu, a to obzvlášť u malých výběrových souborů. Pro malé četnosti v tabulkách 2×2 se používá přesný Fisherův test, pro který existují podrobné tabulky (není třeba nic počítat, vyhodnocení nalezneme v tabulkách přímo z rozložení četností). Pro velké tabulky (d.f. > 30) a malé očekávané četnosti viz Cochran [1954]. Vzhledem k těmto nejednotnostem a absenci konečného zhodnocení problému doporučujeme zachovávat omezení uvedená v textu. Jsme však přesvědčeni, že je možné uvolnit první požadavek tak, aby byl analogický požadavku druhému pro $2 \times C$ tabulky. Ve sporných případech doporučujeme odbornou konzultaci. Při řešení opakovaných úloh v tabulkách speciálních typů doporučujeme aplikaci metody Monte Carlo pro zjištění aplikability či případné nutné modifikace testu.

[9]) Na většině kalkulátorů je tento postup nejvýhodnější vzhledem k tomu, že je jednoduché načítat součiny. I numericky je tento postup vhodný.

[10]) Uvedení správného postupu pro znaménkové schéma je vlastním cílem této statě. Moderní uživatelské výpočetní systémy tento postup nepoužívají. Vzhledem k tomu, že se postup osvědčil, bylo by škoda, aby tento dobrý nápad československých metodologů zapadl. Postup výpočtu z-skóre můžeme použít nejen pro konstrukci celého schématu znamének, ale především jako test odchylky v určitém předem zvoleném poli. Pak má výsledek testování svoji původní a správnou pravděpodobnostní interpretaci. Podle povahy úlohy zvolíme dvoustranný nebo jednostranný test. Pro dvoustranný test jsou kritické hodnoty pro $|z_{ij}| < z_\alpha$ uvedeny v poznámce 11. Pro jednostranný test (předem specifikovanou alternativu co do znaménka rozdílu) máme kritické hodnoty pro $\alpha = 0{,}1$; 0,05; 0,01; 0,001 postupně: $z_\alpha = 1.2816$, 1.6449, 2.3263, 3.0902. Tyto hodnoty platí také v případě, že znaménkové schéma chceme vytvořit pomocí jednostranných testů (což však vnáší další chyby do pravděpodobnostních závěrů, a proto to nedoporučujeme).

stejně. Způsob odvození však vychází ze dvou různých modelů a provádí se, stejně jako odvození χ^2-testu, různě (viz část 6).

Postup se skládá z několika kroků:

1. Určíme si tři hladiny významnosti (např. 0,05; 0,01; 0,001, nebo 0,1; 0,05; 0,01) a nalezneme k nim příslušné kritické hodnoty normálního rozložení (podle dohody vezmeme hodnoty buď pro jednostranný, nebo dvoustranný test; doporučujeme dvoustranný test)[11].

2. Spočítáme z-skóre pro každé pole tabulky.

$$z_{ij} = \frac{n_{ij} - nf_{i+}f_{+j}}{\sqrt{nf_{i+}(1 - f_{i+})\,f_{+j}(1 - f_{+j})}} \tag{13}$$

$$= n\,\frac{n_{ij} - \dfrac{n_{i+}n_{+j}}{n}}{\sqrt{\dfrac{n_{i+}n_{+j}}{n}\,(n - n_{+j})\,(n - n_{i+})}} \tag{14}$$

$$= \sqrt{n}\,\frac{nn_{ij} - n_{i+}n_{+j}}{\sqrt{n_{i+}n_{+j}(n - n_{+j})\,(n - n_{i+})}} \tag{15}$$

$$= \sqrt{n}\,\frac{n_{ij}(n - n_{i+} - n_{+j} + n_{ij}) - (n_{i+} - n_{ij})\,(n_{+j} - n_{ij})}{\sqrt{n_{i+}n_{+j}(n - n_{+j})(n - n_{i+})}} \tag{16}$$

3. Výsledky tiskneme[12] podle pravidla znázorněného v tabulce A.

Tabulka A

tiskni	*je-li z_{ij} v intervalu*
+ + +	$\langle z_3, \infty)$
+ +	$\langle z_2, z_3)$
+	$\langle z_1, z_2)$
0	$(-z_1, z_1\rangle$
−	$(-z_2, -z_1\rangle$
− −	$(-z_3, -z_2\rangle$
− − −	$(-\infty, -z_3\rangle$

Tento postup má ovšem smysl pouze tehdy, když jsme přijali hypotézu o rozdílnosti souborů, resp. hypotézu o závislosti znaků, a chceme specifikovat, kde se tyto rozdílnosti, resp. závislosti projevují.

[11] Dvoustrannému testovému kritériu s volbou hladin významnosti $\alpha_1 = 0{,}05$, $\alpha_2 = 0.01$, $\alpha_3 = 0.001$ odpovídají tyto hodnoty z: $z_1 = 1.9600$, $z_2 = 2.5758$, $z_3 = 3.2905$. V praxi by bylo možné zjednodušit rozhodovací prahy na 2.0; 2.5; 3.0. Tím se nedopustíme velké chyby. Toto zjednodušení však komplikuje situaci tam, kde v programu chceme určit schéma pomocí přesného binomického rozložení (tj. v případě $p_{i+}p_{+j} < 5$). Kdybychom chtěli použít $\alpha = 0.1$, pak příslušné $z = 1.6449$.

[12] Pochopitelně můžeme přijmout jiné grafické zásady pro tisk znaménkového schématu. Je možné omezit tisk znamének na (+ +, +, 0, −, − −) nebo naopak rozšířit jej na čtyři stupně v každém směru (nebo i více): (+ + + +, + + +, + +, +, 0, −, − −, − − −, − − − −). Domníváme se však, že návrh z práce Linhart, Šafář [1967], je pro analytickou práci optimální. Zároveň doporučujeme, aby byla uchována praxe některých programů spočívající v tisku skóre z_{ij}.

Poznámky k metodě.

1. Test platí pro každé pole zvlášť. Schéma jako celek, tj. struktura všech výsledků testování v jednotlivých polích, má daleko nižší stupeň statistické přesvědčivosti než každé jednotlivé pole, protože předpokládáme současnou platnost všech jednotlivých výsledků, které jsou všechny zatíženy chybou. Proto nemůžeme chápat znaménkové schéma jako absolutní, ale pouze jako grafickou pomůcku pro analytickou práci a neustále mít na paměti, že jde o výsledky zatížené větší pravděpodobnostní chybou, než na jakou jsme zvyklí.

2. Postup je ekvivalentní testu χ^2 s jedním stupněm volnosti, když ho použijeme na čtyřpolní tabulku vytvořenou z původní tabulky tak, že jedním jejím polem je dané pole a zbývající pole vzniknou spojením zbylých řádků a sloupců. Znaménko však musíme určit zvlášť.

Příklad 2 (pokračování)

Určíme znaménkové schéma pro tabulku závislosti dvou typologií z příkladu 2. Spočítáme nejprve z-skóre pro jednotlivá pole tabulky podle vzorce (15). Například pro z_{11} dostaneme:[13]

$$z_{11} = \sqrt{\frac{1097}{407(1097 - 407)}} \cdot \frac{[1097 \cdot 100 - 407 \cdot 269]}{\sqrt{269(1097 - 269)}} = .03$$

Tabulka z-skóre pro tabulku 2 je:

	A	B	C	D
a	.03	−2.57	− .29	2.11
b	−.09	−1.75	+3.79	−2.82
c	.04	3.90	−2.73	.19

Tabulka znamének pro $z_1 = 1.96$ $z_2 = 2.58$ $z_3 = 3.29$ ($\alpha_1 = 0.05$ $\alpha_2 = 0.01$ $\alpha_3 = 0.001$)

	A	B	C	D
a	0	− −	0	+
b	0	0	+ + +	− −
c	0	+ + +	− −	0

Typ A se tedy na souvislosti nepodílí. Spolu korelované jsou (b, C), (c, B) a (a, D), které mají vyšší výskyt a (a, B), (b, D), (c, C), které se spolu vyskytují málo.

3. Omezení na výpočet z-skóre plyne z teorie nahrazení binomického rozložení normálním.

Abychom mohli tuto aproximaci použít, je nutné, aby očekávaná četnost (tj. $n_{i+}p_{+j}$ pro případ **A** a $np_{i+}p_{+j}$ pro případ **B**) byla rovna alespoň pěti. Pro menší hodnoty bychom měli spočítat přesný binomický test a určit jeho hladinu významnosti. V praxi nahrazujeme ovšem neznámé parametry p_{i+}, p_{+j} jejich výběrovými odhady f_{i+}, f_{+j}. V případě **A** pak testujeme hypotézu, že binomický výběr o velikosti n_{i+} odpovídá pravděpodobnosti $P = p_{+j}$. V případě **B** pak testujeme hypotézu, že binomický výběr o velikosti n odpovídá pravděpodobnosti $P = p_{i+}p_{+j}$. Počítáme jednostranné nebo dvoustranné testy podle toho, jaké jsme volili hladiny u normální aproximace, aby hodnoty α byly srovnatelné s jednostrannými nebo dvoustrannými hranicemi u z-skóre.

5. Zjednodušení výpočtů v případě 2×C (resp. R×2)

Pro tabulku $2 \times C$ (resp. $R \times 2$ po transformaci) můžeme uvést zjednodušené vzorce pro výpočet *chi*-kvadrátu. Tento případ je častý, zahrnuje porovnání dvou souborů a porovnání R binomických distribucí. Kromě uvedených vzorců (6)—(10) můžeme použít některý z následujících. Označíme si
(pro tabulku $2 \times C$)

[13] Vzorec (15) je pro výpočet poněkud upraven. Pro kalkulačky, které mají paměť, je totiž výhodné spočítat výraz první odmocniny následujícího výrazu pro z_{11}, uchovat jej v paměti a použít pro výpočet z-skóre všech polí prvního řádku. Jestliže je počet řádků větší než počet sloupců, uchováváme obdobný výraz pro sloupec.

$$g_j = \frac{n_{1j}}{n_{+j}}, \quad g = \frac{n_{1+}}{n}$$

$$X^2 = \frac{1}{g(1-g)} \Sigma n_{+j}(g_j - g)^2 \tag{16}$$

$$= \frac{1}{g(1-g)} (\Sigma n_{+j} g_j^2 - n g^2) \tag{17}$$

$$= \frac{1}{g(1-g)} (\Sigma n_{1j} g_j - n_{1+} g) \tag{18}$$

Zároveň se zjednoduší i výpočet znaménkového schématu, které stačí určit jen pro první řádek, protože z-skóre druhého řádku budou opačná čísla.

$$z_{1j} = -z_{2j} \tag{19}$$

Příklad 1 (pokračování)

K výpočtu X^2 použijeme upravený vzorec (18):

$$X^2 = \frac{1}{g\,(1-g)} (\Sigma n_{1j}^2/n_{+j} - n_{1+}^2/n) \tag{18'}$$

$$= \frac{1995^2}{1133 \,.\, 862} \left(\frac{236^2}{395} + \frac{370^2}{690} + \frac{384^2}{616} + \frac{143^2}{294} - \frac{1133^2}{1995} \right)$$

$$= 19.967$$

$$\mathrm{df} = 3$$

V tabulce 4 či 5 Statistických tabulek (Janko [1958]) nalezneme, že $\alpha = 0.003$, tedy X^2 je vysoce významný, přijímáme alternativní hypotézu o rozdílnosti obou souborů. Znaménkové schéma nám může ukázat, u kterých typů se obě populace významně liší. Stačí nám určit z-skóre pro první řádek (druhý řádek obsahuje čísla opačná):

	A	B	C	D
Soubor I	1.32	−2.08	3.34	−3.06
Soubor II	−1.32	+2.08	−3.34	+3.06

Znaménkové schéma:

	A	B	C	D
Soubor I	0	−	+++	−−−
Soubor II	0	+	−−−	+++

Největší rozdíl můžeme hledat u typu C a D, menší, ale významný též u B.

Ještě více se zjednoduší výpočet X^2 i znaménkového schématu pro čtyřpolní tabulku.

$$X^2 = n \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} \tag{20}$$

$$\mathrm{d.f.} = 1 \tag{21}$$

$$z_{11} = \sqrt{n} \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}} \tag{22}$$

$$z_{11} = z_{22} = -z_{12} = -z_{21} \tag{23}$$

$$z_{ij}^2 = X^2 \tag{24}$$

Významnost *chí*-kvadrátu tedy zároveň určuje znaménkové schéma, oba postupy jsou ekvivalentní. Přitom stačí určit jedno znaménko a ostatní plynou okamžitě.

6. Odvození vzorce pro test odchylky v jednotlivém poli tabulky

Bez újmy na obecnosti stačí ukázat odvození pro pole (1,1). Musíme však respektovat modely pro typy **A** a **B**.

Typ A

Předpokládáme, že rozložení v řádcích jsou nezávislá a že platí nulová hypotéza $p_{j/i} = p_{+j}$, p_{+j} není známo, $p_{i+} = \frac{n_{i+}}{n}$ je předem určeno. Test můžeme založit na dvou přístupech:

Přístup 1

Pro danou hodnotu znaku porovnáme relativní četnost ve zkoumaném řádku a relativní četnost v celém zbylém souboru, tj.

$$w = \frac{n_{11}}{n_{1+}} - \frac{n_{+1} - n_{11}}{n - n_{1+}} \tag{25}$$

Obě veličiny mají asymptoticky normální rozložení, w má tedy také asymptoticky normální rozložení. Spočteme jeho varianci

$$\begin{aligned} \operatorname{var} w &= \frac{1}{n_{1+}^2} \operatorname{var} n_{11} + \frac{1}{(n - n_{1+})^2} \operatorname{var} (n_{+1} - n_{11}) \\ &= p_{+1}(1 - p_{+1}) \left[\frac{1}{n_{1+}} + \frac{1}{n - n_{1+}} \right] \\ &= \frac{1}{n} \frac{p_{+1}(1 - p_{+1})}{p_{1+}(1 - p_{1+})} \end{aligned} \tag{26}$$

Statistika $z = \frac{w}{\sqrt{\operatorname{var} w}}$ má tedy standardní normální rozložení s průměrem nula a rozptylem jedna.

Dosazením $p_{+1} = f_{+1}$ a snadnou úpravou dostaneme

$$z = \sqrt{n} \frac{n n_{11} - n_{1+} n_{+1}}{\sqrt{n_{1+} n_{+1} (n - n_{1+}) (n - n_{+1})}} \tag{27}$$

což je vzorec (15).

Přístup 2

Porovnáme relativní četnost jednoho souboru a relativní četnost téhož sloupce pro celý soubor dat

$$w' = \frac{n_{11}}{n_{1+}} - \frac{n_{+1}}{n} \tag{28}$$

w' má opět asymptoticky normální rozložení. Výpočet variance se nepatrně komplikuje, neboť oba členy z (28) nejsou statisticky nezávislé. Proto nejprve upravíme součet na dva nezávislé členy.

$$\begin{aligned} \operatorname{var} w' &= \operatorname{var} \left(\frac{n_{11}}{n_{1+}} - \frac{n_{+1}}{n} \right) \\ &= \operatorname{var} \left(\frac{n_{11}}{n_{1+}} - \frac{n_{11} + n_{+1} - n_{11}}{n} \right) \\ &= \operatorname{var} \left[\left(\frac{n_{11}}{n_{1+}} - \frac{n_{11}}{n} \right) + \left(\frac{n_{+1} - n_{11}}{n} \right) \right] \end{aligned} \tag{29}$$

$$= (1 - p_{1+})^2 \left[\operatorname{var} \frac{n_{11}}{n_{1+}} + \operatorname{var} \frac{n_{+1} - n_{11}}{n - n_{1+}} \right]$$

$$= \frac{1}{np_{1+}} (1 - p_{1+}) p_{+1}(1 - p_{+1})$$

Pro $z' = \dfrac{w'}{\sqrt{\operatorname{var} w'}}$ dostaneme po úpravách stejný výraz jako (27). Oba návrhy vedou tedy na stejnou statistickou techniku.

Typ B

Odvození z-skóre provedeme pomocí asymptotické teorie rozložení statistik, které jsou funkcí četností (viz Rao [1968], část 6.b. 3, „Test pro odchylku v jednom poli"). Model nulové hypotézy odpovídá multinomickému rozložení s $R \times C$ třídami, jimž odpovídají pravděpodobnosti p_{ij}, dané pomocí $R + C - 2$ parametrů p_{i+} $(i = 1, \ldots, R - 1)$, p_{+j} $(j = 1, \ldots, C - 1)$ jako $p_{ij} = p_{i+}p_{+j}$. Formulujeme-li tento model, pak zbytek je už aplikace Raovy teorie. Nebudeme zde ukazovat detaily, neboť výsledky plynou ze zdlouhavých algebraických úprav; uvedeme pouze výsledky nejdůležitějších kroků. Zájemce odkazujeme na Raovu knihu (Rao [1968]), kde lze nalézt definice, terminologii a podrobnosti teorie. Označíme $p_{R+} = 1 - \sum_{i=1}^{R-1} p_{i+}$, $p_{+C} = 1 - \sum_{j=1}^{C-1} p_{+j}$, δ_{ij} = Kroneckerova delta.

Krok 1:

Výpočet informační matice **I** pro parametry p_{i+}, p_{+j}. **I** je čtvercová řádu $(R + C - 2)$ rozdělená na bloky:

$$\mathbf{I} = \left\| \begin{matrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{matrix} \right|$$

Matice **B** a **C** obsahují samé nuly.
Prvky matice **A**, která je čtvercová řádu $(R - 1)$

$$a_{rs} = \delta_{rs} \cdot \frac{1}{p_{r+}} + \frac{1}{p_{R+}} \tag{30}$$

Prvky matice **D**, která je čtvercová řádu $(C - 1)$:

$$d_{rs} = \delta_{rs} \cdot \frac{1}{p_{+r}} + \frac{1}{p_{+C}} \tag{31}$$

Krok 2:

Inverze $\mathbf{I}^{-1}$ informační matice **I**.

$$\mathbf{I}^{-1} = \left\| \begin{matrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{matrix} \right\|$$

Prvky matice $\mathbf{A}^{-1}$:

$$a^{rs} = \delta_{rs}p_{r+} - p_{r+}p_{s+} \tag{32}$$

Prvky matice $\mathbf{D}^{-1}$:

$$d^{rs} = \delta_{rs}p_{+r} - p_{+r}p_{+s} \tag{33}$$

Krok 3:

Výpočet hodnoty v_{ij} (Rao [1968], vztahu 6.b.3.2.)

(34) $$v_{ij} = (1 - p_{i+})(1 - p_{+j})$$

Krok 4:

Výpočet z-skóre (Rao [1968], vztah 6.b.3.3)

(35) $$z_{ij} = \frac{n_{ij} - np_{i+}p_{+j}}{\sqrt{np_{i+}p_{+j}(1 - p_{i+})(1 - p_{+j})}}$$

což po dosazení $p_{i+} = f_{i+}$, $p_{+j} = f_{+j}$ dává vztahy (13)—(16).

Q.E.D

Oba vztahy (27) i (35) vedou na stejnou statistiku z_{ij}. Testování provádíme proto stejně bez ohledu na to, ze kterého modelu statistiky z_{ij} vznikly. Obě statistiky mají ovšem zcela stejné vlastnosti pouze asymptoticky. V případě **A** máme pouze $(C - 1)$ neznámých parametrů, které odhadujeme pomocí empirických dat, v případě typu **B** je jich $(R + C - 2)$. Proto odhady a testy u typu **A** budou pro stejný rozměr tabulky a stejný počet pozorování zřejmě přesnější než v případě **B**.

7. Závěr

V této stati jsme chtěli uvést a dokázat správný postup pro určení znaménkového schématu pro strukturu vztahů či rozdílností v kontingenční tabulce. Motivace byla hlavně dvojí: a) podnítit znovu zájem o znaménkové schéma, které se ukázalo jako velmi užitečné u uživatelů (tj. pro praktickou práci sociologa analyzujícího data); b) uvést správný postup pro konstrukci schématu. Bylo by škoda, kdyby dobrý nápad konstrukce znaménkového schématu zapadl na druhé straně však je nutné, aby programování postupu bylo založeno na teoreticky správných základech. Vzhledem k tomu, že autorům jsou známy různé verze tohoto postupu, které si vzájemně neodpovídají, bylo by vhodné prověřit, resp. opravit, stávající programy obsahující tuto metodu (programátorsky jde o nepatrné a časově zanedbatelné změny).Vycházeli jsme z návrhu práce Linhart, Šafář [1967]. Znaménkové schéma však lze určovat také na základě zcela jiných statistických principů. Nakonec bychom chtěli znovu zdůraznit, že tato vhodná metoda má svá omezení a výsledky schématu nelze v žádném případě absolutizovat.[14]

(14) Během doby, kdy byl článek v tisku, autoři zjistili, že vzorce pro reziduální odchylky (bez odvození) jsou uvedeny v článku S. J. Haberman: The Analysis of Residuals in Cross-Classified Tables. Biometrics 1973, s. 205—220. a to v kontextu grafické metody pro analyzu reziduí v modelech kontingenčních tabulek.

Literatura

Cochran, W. G.: *Some Methods for Strengthening the Common χ^2 Tests.* Biometrics 10, 1954, s. 417—451.
Grizzle, J. E.: *Continuity Correction in the χ^2 — test for 2 × 2 Tables.* The American Statistician 1967, s. 28—32.
Janko, J.: *Statistické tabulky.* Praha, Nakladatelství ČSAV 1958.
Kelley, T. L.: *Statističeskije tablicy.* Moskva, Vyčislitělnyj centr AN SSSR 1966.
Lancaster, H. O.: *The chi-squared Distribution.* New York, John Wiley &Sons 1969.
Lewontin, R. C.; Felsenstein, J.: *The Robustness of Homogeneity Tests in 2 × N Tables.* Biometrics 21, 1965, s. 19—33.
Linhart, J., Šafář, Z.: *Programování třídění a statistických výpočtů pomocí samočinných počítačů.* Sociologický časopis 1967, č. 4, s. 435—443.
Pirie, W. R., Hamdan, M. A.: *Some Revised Continuity Corrections for Discrete Distributions.* Biometrics 28, 1972, s. 693—701.
Rao, C. R.: *Linějnyje statističeskije metody i ich primeněnija.* Moskva, Nauka 1968.
Rao, C. R.: *Lineární metody statistické indukce a jejich aplikace.* Praha, Academia 1978.
Řehák, J., Řeháková, B.: *Měření statistické závislosti nominálních znaků.* Sociologický časopis 1973, č. 4, s. 404—418.

Резюме

Ржегак Я. - Ржегакова, Б.: Анализ таблиц сравнения параллельных выборок и таблиц сопряженности признаков при помощи схемы знаков

Статья вводит две основные аналитические задачи ведущие к широко известному критерию хи-квадрат:

а) таблицы сравнения параллельных выборок (критерий однородности)

б) таблицы сопряженности признаков (критерий независимости)

Знаковая схема образуется путем применения к каждому полю таблицы критерия для проверки отклонения между установленным наблюдением и ожидаемой частотой в каждом поле таблицы. По формулам (13)—(16) получаются величины критерия, которые сопоставляются с данными критическими значениями нормального распределения.

В статье приводятся два нумерических примера и способ, как вести доказательство согласно теории асимптотических приближений по Рао [1968].

Summary

Řehák J. — Řeháková B.: Analysis of Contingency Tables: Two Basic Types of Analytical Problems and Sign Scheme

The paper introduces two basic and most frequent types of analytical problems, both leading to a common chi-square test technique, namely comparison of samples with respect to one nominal variable and investigating independence of two nominal variables with respect to one sample. The main task of this paper is to show a revised method of calculating the sign scheme. It might be used in a case of the rejection of null hypothesis of homogeneity or independence for specifying a type of differences or dependence that occurs in the data.

The sign scheme is based on a test for deviation of observed and expected frequencies in a single cell. This test is applied to each cell in a table. The form of the test is a result of Rao asymptotic theory [Rao 1968], and the test statistics is given by formulas (13) — (16). The table of sings is a result of a rule applying a two-sided normal test for z-scores for three given significance levels.

The sign scheme is a good graphical device for analytical work. The purpose of this paper is to show the correct formulas for this method (and present its abridged mathematical documentation).