Milan Terek
Peter Kročitý

ANALÝZA ASOCIÁCIE MEDZI NOMINÁLNYMI PREMENNÝMI V ŠTATISTICKÝCH PRIESKUMOCH[1]

Abstract: *The paper deals with the possibilities of analysis of the association between two nominal variables in statistical surveys. The procedure of realization of chi-square test of homogeneity is described. Then, residual analysis, concretely the possibilities of using adjusted standardized residuals in the analysis of character of association between variables, is presented. Finally, statistic chi-square and measures of strength of association – difference of proportions and odds ratio are described. The application of all described methods and tools is illustrated on the examples.*

Keywords: *nominal variables, association, chi-square test of homogeneity, adjusted standardized residuals, difference of proportions, odds ratio*

JEL: C 46

Úvod

V posledných rokoch sa výrazne rozšírilo používanie výberových skúmaní na báze štatistických prieskumov na získavanie informácií potrebných pri rozhodovaní. Štatistický prieskum možno charakterizovať ako proces zhromažďovania dát prostredníctvom zisťovania odpovedí respondentov na otázky. Väčšinou sa vytvorí zoznam otázok, ktoré sa zhromaždia v dotazníku. Potom sa získavajú odpovede respondentov na otázky. Dáta možno získavať osobným interview, telefonicky, e-mailom a podobne. V dotazníkoch sa často vyskytujú otázky, ktoré možno z formálno-matematickej stránky charakterizovať ako nominálne premenné. Všimneme si metódy detekcie a opisu spojenia medzi dvoma kvalitatívnymi, nominálnymi premennými.

Všeobecne, keď určité hodnoty jednej premennej majú tendenciu meniť sa s určitými hodnotami druhej premennej, hovoríme, že medzi premennými je asociácia alebo spojenie (*association*).

[1] Článok vznikol s prispením grantovej agentúry VEGA v rámci projektu č. 1/0761/12: Alternatívne prístupy k meraniu sociálno-ekonomického rozvoja (v kontexte Stratégie 2020 a poučení z globálnej finančnej krízy).

Kvalitatívne premenné nadobúdajú hodnoty (kategórie, obmeny, úrovne), ktoré umožňujú identifikovať znak každej jednotky. Ide napríklad o pohlavie, národnosť a podobne. Napríklad kvalitatívna premenná pohlavie nadobúda dve hodnoty – muž a žena. Jednotka môže mať znak – muž, alebo znak – žena.

Keď hodnoty kvalitatívnej premennej umožňujú identifikovať znak každej jednotky, stupnica merania kvalitatívnej premennej je nominálna.[2] Napríklad stupnica merania pohlavia osôb je nominálna.

Kvalitatívne premenné sa merajú v ordinálnej stupnici, keď ich hodnoty majú vlastnosti nominálnych dát a ich usporiadanie má zmysel.[3] Napríklad zákazník by mohol pri kúpe automobilu hodnotiť jeho farbu pomocou subjektívnej stupnice s kategóriami: veľmi pekná, pekná, nepekná. Veľmi pekná je preferovaná pred peknou, pekná je preferovaná pred nepeknou. Hodnoty premennej možno usporiadať podľa preferencie.

Keď sú hodnoty kvalitatívnej premennej vyjadrené slovne, ide o meranie v nominálnej alebo ordinálnej slovnej stupnici, keď hodnotám priradíme číselné kódy, ide o meranie v nominálnej alebo ordinálnej číselnej (numerickej) stupnici.

Dáta pre analýzu asociácie dvoch kvalitatívnych premenných sa sústreďujú v kontingenčných tabuľkách. V kontingenčnej tabuľke obyčajne každý riadok korešponduje s jednou hodnotou jednej premennej, každý stĺpec korešponduje s jednou hodnotou druhej premennej. V políčkach tabuľky sú počty jednotiek s príslušnou kombináciou hodnôt premenných.

Základné otázky, ktoré si analytik kladie pri analýze kontingenčnej tabuľky, sú najčastejšie takéto:

– Existuje asociácia medzi premennými? Odpoveď možno nájsť pomocou testu chí-kvadrát. Čím menšia je *p*-hodnota, tým viac to svedčí v prospech asociácie medzi premennými.
– Ako sa dáta líšia od situácie, v ktorej premenné nie sú spojené? Korigované normované rezíduá identifikujú políčka, viac alebo menej podobné stavu, ktorý by sa očakával pri neexistencii asociácie.
– Aká silná je asociácia? Na charakterizáciu sily asociácie sa používajú napríklad štatistiky rozdiel podielov (*difference of proportions*) a pomer šancí (*odds ratio*).

Aplikácia uvedených metód bude ilustrovaná na príklade štúdia asociácie niektorých nominálnych premenných, ktoré sa analyzovali pri štatistickom prieskume zameranom na štúdium informačných tokov a manažmentu znalostí o akademickej etike, v širšom kontexte znalostného manažmentu, na Vysokej škole manažmentu v Trenčíne, City University of Seattle, v jej pobočkách v Trenčíne a v Bratislave. Náhodne vybraným študentom bol poslaný dotazník, ktorý obsahoval 8 otázok. Jeho súčasťou bola aj otázka č. 2: „Kedy ste po prvý raz počuli o pravidlách akademickej etiky na VŠM/CU?“. Odpovede na túto otázku v súvislosti s pobočkou, na ktorej respondent študuje, a s formou jeho štúdia si všimneme podrobnejšie.

[2] Vtedy hovoríme o nominálnych dátach.
[3] Vtedy hovoríme o ordinálnych dátach.

1 Chí-kvadrát test homogenity

Všeobecne v štatistike sa hovorí o homogenite, keď sú štatistické vlastnosti nejakej časti súboru dát rovnaké ako vlastnosti nejakej jeho inej časti.

Často treba rozhodnúť, či pozorované rozdiely medzi hodnotami výberových podielov sú významné, alebo vyplývajú len z náhodnosti vyberania. Všimneme si testy o rozdieloch medzi viacerými podielmi. Budeme analyzovať kontingenčnú tabuľku typu r x c ($c > 2$) a uvažovať o náhodných výberoch z r základných súborov s multinomickým rozdelením.[4] V každom pokuse môže nastať jeden z c možných výsledkov. V kontingenčnej tabuľke sú rozsahy výberov v poslednom stĺpci fixované, stĺpcové súčty sú ovplyvnené náhodnosťou vyberania.

Nech π_{ij} je pravdepodobnosť j-teho výstupu pre i-ty základný súbor. Testujeme:

$$H_0: \pi_{1j} = \pi_{2j} = \ldots = \pi_{rj} \qquad \text{pre } j = 1, 2, \ldots c$$

t. j., že náhodné výbery sú z r základných súborov s rovnakým multinomickým rozdelením, oproti alternatíve
H_1: $\pi_{1j}, \pi_{2j}, \ldots \pi_{rj}$ sa všetky nerovnajú aspoň pre jednu hodnotu j

Nech n_{ij} je pozorovaná početnosť v i-tom riadku a j-tom stĺpci, n_i je súčet hodnôt n_{ij} v i-tom riadku, n_j je súčet hodnôt n_{ij} v j-tom stĺpci. Súčet všetkých hodnôt n_{ij} označíme n.

Keď vyjdeme z predpokladu, že H_0 je správna, potom teoretické početnosti sa vypočítajú takto:

$$o_{ij} = \frac{n_i \cdot n_j}{n} \tag{1}$$

Hodnota testovacej štatistiky sa vypočíta podľa vzťahu

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - o_{ij})^2}{o_{ij}}$$

Kritická oblasť testu na hladine významnosti α je: $\chi^2 \geq \chi^2_{1-\alpha}((r-1)(c-1))$

Pretože uvažovaná testovacia štatistika má len približne rozdelenie chí-kvadrát, odporúča sa používať tento test len vtedy,[5] keď žiadna hodnota o_{ij} nie je menšia ako 5. Tieto testy sa často nazývajú testy homogenity.[6]

Príklad 1. Zaujíma nás, či sa spôsob získavania informácií o akademickej etike významne líši v pobočkách Trenčín (TN) a Bratislava (BA). Náhodným vyberaním s opakovaním sa získali odpovede na otázku č. 2: „Kedy ste po prvý raz počuli o pra-

[4] Multinomické rozdelenie pozri napríklad v ([8], s. 86 – 87).

[5] Tamtiež.

[6] Viac o chí-kvadrát testoch homogenity (o rozdieloch medzi viacerými podielmi) a nezávislosti možno nájsť v ([8], s. 197 – 207).

vidlách akademickej etiky na VŠM/CU?“ od 62 respondentov z pobočky Bratislava a od 45 respondentov z pobočky Trenčín. Výsledky sú v tabuľke č. 1. V zátvorkách sú teoretické početnosti vypočítané podľa vzťahu (1).

Tab. č. 1

Rozdelenie odpovedí na otázku č. 2 v pobočkách

Odpovede na otázku č. 2 Pobočka	Pred podaním prihlášky na štúdium 1	Na začiatku štúdia – počas zhromaždenia nových študentov 2	Počas štúdia v prvom trimestri 3	Počas štúdia v prvom ročníku (neskôr ako 1. trim.) 4	Počas druhého roku štúdia 5	Počas tretieho roku štúdia 6	Neskôr ako počas tretieho roku štúdia 7	Dozvedel som sa inak 8	Nikdy som o nich nepočul 9	n_i
BA	5 (5)	36 (41)	14 (12)	2 (1)	0 (0)	0 (0)	0 (0)	4 (2)	1 (1)	62
TN	4 (4)	34 (30)	7 (9)	0 (1)	0 (0)	0 (0)	0 (0)	0 (2)	0 (0)	45
n_j	9	70	21	2	0	0	0	4	1	107

Stĺpce s teoretickými početnosťami menšími ako 5 treba zlúčiť. Výsledné rozdelenie je v tabuľke č. 2. V tabuľke č. 2 sú v zátvorkách teoretické početnosti vypočítané podľa vzťahu (1).

Tab. č. 2

Rozdelenie odpovedí na otázku č. 2 v pobočkách, po zlúčení

Odpovede na otázku č. 2 Pobočka	Pred podaním prihlášky na štúdium alebo v 1. ročníku (neskôr ako 1 trim.), 2. alebo 3. ročníku, neskôr ako v 3. ročníku, dozvedel som sa inak alebo som o nich nikdy nepočul. 1+4+5+6+7+8+9	Na začiatku štúdia – počas zhromaždenia nových študentov 2	Počas štúdia v prvom trimestri 3	n_i
BA	12 (9)	36 (41)	14 (12)	62
TN	4 (7)	34 (29)	7 (9)	45
n_j	16	70	21	107

Všeobecne, v kontingenčnej tabuľke sa vo všetkých riadkoch riadkový súčet pozorovaných početností rovná riadkovému súčtu teoretických početností. Podobne vo všetkých stĺpcoch sa stĺpcový súčet pozorovaných početností rovná stĺpcovému súčtu teoretických početností.

Test chí-kvadrát bol realizovaný pomocou štatistickej funkcie CHISQ. TEST v Exceli.[7] V teste vyšla p-hodnota = 0,103554. To znamená, že na hladine významnosti napríklad 0,05 nezamietame predpoklad, že výbery sú z rovnakého multinomického rozdelenia, alebo, čo je ekvivalentné, že náhodný výber zo základného súboru – Študenti z Bratislavy a náhodný výber zo základného súboru – Študenti z Trenčína, sú z rovnakého rozdelenia pravdepodobnosti náhodnej premennej – Odpovede na otázku č. 2. Na hladine významnosti 0,05 teda nezamietame predpoklad, že spôsob získavania informácií o pravidlách

[7] Podrobnejšie v [9].

akademickej etiky sa v jednotlivých pobočkách nelíši. Nezamietame predpoklad, že neexistuje asociácia medzi spôsobom získavania informácií o akademickej etike a pobočkou.

Príklad 2. Preskúmame spojenie medzi spôsobom získavania informácií o pravidlách akademickej etiky a formou štúdia – externí alebo denní študenti. Výsledné rozdelenie je v nasledujúcej tabuľke. V tabuľke sú v zátvorkách teoretické početnosti.

Tab. č. 3

Rozdelenie odpovedí na otázku č. 2 externých a denných študentov, po zlúčení

Odpovede na otázku č. 2 Forma štúdia	Pred podaním prihlášky na štúdium alebo v 1. ročníku (neskôr ako 1 trim.), 2. alebo 3. ročníku, neskôr ako v 3. ročníku, dozvedel som sa inak alebo som o nich nikdy nepočul. 1+4+5+6+7+8+9	Na začiatku štúdia – počas zhromaždenia nových študentov 2	Počas štúdia v prvom trimestri 3	Spolu n_i
Externí	10 (10,89)	22 (27,78)	35 (28,33)	67
Denní	10 (9,11)	29 (23,22)	17 (23,67)	56
Spolu n_j	20	51	52	123

V teste vyšla *p*-hodnota = 0,043915. To znamená, že na hladine významnosti 0,05 zamietame predpoklad, že výbery zo základných súborov – Externí študenti a Denní študenti sú z rovnakého rozdelenia pravdepodobnosti náhodnej premennej – Odpovede na otázku č. 2, a prijímame alternatívnu hypotézu, že nie sú z rovnakého rozdelenia. Teda spôsob získavania informácií o pravidlách akademickej etiky sa významne líši u externých a denných študentov. Prijímame predpoklad, že existuje asociácia medzi spôsobom získavania informácií o akademickej etike a formou štúdia.

Samotný test ale nič nehovorí o charaktere alebo sile asociácie. Test nenaznačuje, či sa všetky políčka významne odchyľujú od homogenity, prípadne len jedno alebo dve políčka. To umožňuje analýza rezíduí.

2 Analýza rezíduí

Porovnanie pozorovaných a teoretických početností umožňuje analyzovať charakter spojenia medzi premennými. Rozdiel (n_{ij}-o_{ij}) sa nazýva rezíduum. Možno definovať korigované (upravené) normované rezíduá (*adjusted standardized residuals*):

$$r_{ij} = \frac{n_{ij} - o_{ij}}{\sqrt{o_{ij}\left(1 - \frac{n_i}{n}\right)\left(1 - \frac{n_j}{n}\right)}} \qquad pre\ i = 1,2,...r, \quad j = 1,2...c \qquad (2)$$

kde $\frac{n_i}{n}$ je odhadnutá marginálna pravdepodobnosť v riadku *i*,

$\frac{n_{.j}}{n}$ – odhadnutá marginálna pravdepodobnosť v stĺpci *j*.

Menovateľ vo vzťahu (2) je smerodajná chyba náhodnej premennej (n_{ij}-o_{ij}), keď je H_0 správna ([1], s. 230).

Korigované normované rezíduá r_{ij} majú asymptoticky normované normálne rozdelenie. Možno ich použiť neformálnym spôsobom na opis obrazu spojenia medzi políčkami tabuľky. Veľká hodnota korigovaného normovaného rezídua indikuje odklon od homogenity v políčku. Keď je H_0 správna, potom je približne len 5 % šanca, že korigované normované rezíduum presiahne v absolútnej hodnote 2. Hodnoty v absolútnej hodnote väčšie ako 3 už jasne naznačujú spojenie v políčku.

Príklad 2 – pokračovanie 1. Vypočítame hodnoty korigovaných normovaných rezíduí r_{ij} podľa vzťahu (2).

Tab. č. 4

Hodnoty korigovaných normovaných rezíduí r_{ij}

Odpovede na otázku č. 2 Forma štúdia	Pred podaním prihlášky na štúdium alebo v 1. ročníku (neskôr ako 1 trim.), 2. alebo 3. ročníku, neskôr ako v 3. ročníku, dozvedel som sa inak alebo som o nich nikdy nepočul. 1+4+5+6+7+8+9	Na začiatku štúdia – počas zhromaždenia nových študentov 2	Počas štúdia v prvom trimestri 3
Externí	−0,44	−2,12	2,44
Denní	0,44	2,12	−2,44

V tabuľke sú pomerne veľké kladné hodnoty rezíduí pre denných študentov, ktorí získali informácie na začiatku štúdia – počas zhromaždenia nových študentov, a pre externých študentov, ktorí získali informácie počas štúdia v prvom trimestri. To znamená, že je viac denných študentov, ktorí získali informácie na začiatku štúdia – počas zhromaždenia nových študentov, a viac externých študentov, ktorí získali informácie počas štúdia v prvom trimestri, ako predpokladá hypotéza o homogenite.

Podobne v tabuľke sú pomerne veľké záporné hodnoty rezíduí pre externých študentov, ktorí získali informácie na začiatku štúdia – počas zhromaždenia nových študentov, a pre denných študentov, ktorí získali informácie počas štúdia v prvom trimestri. To znamená, že je menej externých študentov, ktorí získali informácie na začiatku štúdia – počas zhromaždenia nových študentov, a menej denných študentov, ktorí získali informácie počas štúdia v prvom trimestri ako predpokladá hypotéza o homogenite.

To znamená, že denní študenti získali informácie skôr na začiatku štúdia – počas zhromaždenia nových študentov, externí študenti získali informácie skôr počas štúdia v prvom trimestri.

3 Charakteristiky sily asociácie

Charakteristika sily asociácie je štatistika alebo parameter, ktorý charakterizuje silu asociácie medzi dvoma premennými ([1], s. 233).

3.1 Rozdiel podielov

Všimnime si najprv niektoré rozdelenia početností. Podobne ako je definované združené, marginálne[8] a podmienené[9] rozdelenie pravdepodobnosti, možno definovať združené, marginálne a podmienené rozdelenie početností. V rozdeleniach početností sa namiesto pravdepodobností uvažuje o početnostiach. Napríklad v tabuľke č. 3, v druhom a treťom riadku, v druhom, treťom a štvrtom stĺpci (čísla mimo zátvoriek), je združené rozdelenie absolútnych početností, v poslednom stĺpci, v druhom a treťom riadku je marginálne rozdelenie absolútnych početností premennej Forma štúdia a v poslednom riadku, v druhom, treťom a štvrtom stĺpci je marginálne rozdelenie absolútnych početností premennej Odpovede na otázku č. 2. V tabuľke č. 5 je združené rozdelenie relatívnych početností a marginálne rozdelenia relatívnych početností premennej Forma štúdia a premennej Odpovede na otázku č. 2.

Tab. č. 5

Rozdelenie relatívnych početností odpovedí na otázku č. 2 externých a denných študentov

Odpovede na otázku č. 2 Forma štúdia	Pred podaním prihlášky na štúdium alebo v 1. ročníku (neskôr ako 1 trim.), 2. alebo 3. ročníku, neskôr ako v 3. ročníku, dozvedel som sa inak alebo som o nich nikdy nepočul. 1+4+5+6+7+8+9	Na začiatku štúdia – počas zhromaždenia nových študentov 2	Počas štúdia v prvom trimestri 3	Spolu n_i
Externí	10/123	22/123	35/123	67/123
Denní	10/123	29/123	17/123	56/123
Spolu n_j	20/123	51/123	52/123	1

V tabuľke č. 6 sú podmienené rozdelenia početností premennej Odpovede na otázku č. 2.[10] Ide o rozdelenia podmienené formou štúdia. Napríklad (0,149, 0,328, 0,522) je podmienené rozdelenie premennej Odpovede na otázku č. 2 pre externú formu štúdia a (0,179, 0,518, 0,304) je podmienené rozdelenie premennej Odpovede na otázku č. 2 pre dennú formu štúdia.

[8] Pozri napríklad v [8].

[9] Pozri napríklad v [6].

[10] Pri ich výpočte možno vychádzať z absolútnych alebo relatívnych početností.

Tab. č. 6

Podmienené rozdelenia Odpovede na otázku č. 2

Odpovede na otázku č. 2 Forma štúdia	Pred podaním prihlášky na štúdium alebo v 1. ročníku (neskôr ako 1 trim.), 2. alebo 3. ročníku, neskôr ako v 3. ročníku, dozvedel som sa inak alebo som o nich nikdy nepočul. 1+4+5+6+7+8+9	Na začiatku štúdia – počas zhromaždenia nových študentov 2	Počas štúdia v prvom trimestri 3	n_i
Externí	10/67 ≈ 0,149	22/67 ≈ 0,328	35/67 ≈ 0,522	67
Denní	10/56 ≈ 0,179	29/56 ≈ 0,518	17/56 ≈ 0,304	56
n_j	20	51	52	123

Keď každá z dvoch premenných nadobúda len dve hodnoty, príslušná kontingenčná tabuľka je typu 2 x 2. Vtedy je vhodnou charakteristikou asociácie rozdiel podielov. V tejto analýze je užitočné rozlišovať závisle a nezávisle premennú. Ktorá z dvoch premenných sa považuje za závislú, závisí od cieľa skúmania.

Hodnoty tejto charakteristiky sa počítajú ako rozdiely medzi hodnotami podmieneného rozdelenia závisle premennej pre danú hodnotu závisle premennej.

Príklad 2 – pokračovanie 2. Predpokladajme, že by sme v prieskume uvažovali len o externej a dennej forme štúdia a o dvoch možných odpovediach na otázku č. 2, označených v tabuľke ako 2 a 3. Ďalej predpokladajme hypotetické výsledky, ktoré sú uvedené v tabuľkách č. 7 a č. 8. Závisle premennou nech je premenná Odpovede na otázku č. 2.

Tab. č. 7

Hypotetické rozdelenie 1 Odpovede na otázku č. 2

Odpovede na otázku č. 2 Forma štúdia	Na začiatku štúdia – počas zhromaždenia nových študentov 2	Počas štúdia v prvom trimestri 3	n_i
Externí	0	60	60
Denní	40	0	40

Tab. č. 8

Hypotetické rozdelenie 2 Odpovede na otázku č. 2

Odpovede na otázku č. 2 Forma štúdia	Na začiatku štúdia – počas zhromaždenia nových študentov 2	Počas štúdia v prvom trimestri 3	n_i
Externí	36	24	60
Denní	24	16	40

Vypočítame početnosti podmieneného rozdelenia početností premennej Odpovede na otázku č. 2 pre hypotetické rozdelenie 1[11], pre externých študentov:

$$(\frac{0}{60}, \frac{60}{60}) = (0, 1)$$

a pre denných študentov:

$$(\frac{40}{40}, \frac{0}{40}) = (1, 0)$$

Rozdiel hodnôt podmieneného rozdelenia pre odpoveď 2 je:
0 – 1 = – 1
a rozdiel hodnôt podmieneného rozdelenia pre odpoveď 3 je:
1 – 0 = 1

Každý z výsledkov charakterizuje najsilnejšiu možnú asociáciu. Externí študenti totiž volili len odpoveď 3, denní študenti len odpoveď 2.

Teraz vypočítame početnosti podmieneného rozdelenia početností premennej Odpovede na otázku č. 2 pre hypotetické rozdelenie 2[12], pre externých študentov:

$$(\frac{36}{60}, \frac{24}{60}) = (0{,}6;\ 0{,}4)$$

a pre denných študentov:

$$(\frac{24}{40}, \frac{16}{40}) = (0{,}6;\ 0{,}4)$$

Rozdiel hodnôt podmieneného rozdelenia pre odpoveď 2 je:
0,6 – 0,6 = 0
a rozdiel hodnôt podmieneného rozdelenia pre odpoveď 3 je:
0,4 – 0,4 = 0

Každý z výsledkov charakterizuje neexistenciu asociácie. Podmienené rozdelenia sú totiž rovnaké, čo platí v prípade, keď medzi premennými nie je asociácia. V príklade rovnaký podiel externých a denných študentov volil odpoveď 2 a rovnaký podiel externých a denných študentov volil odpoveď 3. Odpoveď na otázku č. 2 nie je spojená s formou štúdia.

Všeobecne – čím je asociácia silnejšia, tým väčšia je absolútna hodnota rozdielu hodnôt podmieneného rozdelenia.

[11] V tabuľke č. 7.
[12] V tabuľke č. 8.

Ešte raz pripomíname, že túto charakteristiku možno použiť len keď má každá z dvoch premenných dve hodnoty. V uvádzanej aplikácii to nebolo možné.

3.2 Štatistika chí-kvadrát nie je charakteristikou sily asociácie

Veľká hodnota testovacej štatistiky chí-kvadrát naznačuje, že premenné sú spojené. Nič ale nehovorí o sile tohto spojenia.

Príklad 3. Majme hypotetické dáta v tabuľkách č. 9 a č. 10. Na hladine významnosti 0,01 budeme testovať homogenitu rozdelení.

Tab. č. 9

Hypotetické výsledky 3

Y X	y_1	y_2	n_i
x_1	490	510	1000
x_2	510	490	1000
n_j	1000	1000	2000

Hodnota testovacej štatistiky je $x^2 = 0,8$ a p-hodnota je 0,371093. Na hladine významnosti 0,01 nezamietame H_0, že výbery sú z rovnakého rozdelenia.

Tab. č. 10

Hypotetické výsledky 4

Y X	y_1	y_2	n_i
x_1	4900	5100	10000
x_2	5100	4900	10000
n_j	10000	10000	20000

Hodnota testovacej štatistiky je $\chi^2 = 8$ a p-hodnota je 0,004678. Na hladine významnosti 0,01 zamietame H_0 a prijímame H_1, že výbery nie sú z rovnakého rozdelenia.

V oboch prípadoch je rozdiel podielov rovnaký – 0,02, ide o veľmi slabú asociáciu.

Všeobecne aj v prípade slabej asociácie, keď je rozsah výberu veľký, p-hodnota môže byť malá.

3.3 Pomer šancí

Pre závisle premennú s dvoma hodnotami je šanca úspechu definovaná takto:

$$\text{Šanca} = \frac{\text{Pravdepodobnosť úspechu}}{\text{Pravdepodobnosť neúspechu}}$$

Všeobecne, – odhadnutá šanca[13] pre závisle premennú s dvoma hodnotami sa rovná počtu úspechov delenému počtom neúspechov. Pomer šancí θ v kontingenčných tabuľkách typu 2 x 2 sa rovná pomeru šance v prvom riadku a šance v druhom riadku.

Príklad 2 – pokračovanie 3. V tabuľke č. 3 z príkladu 2 si všimneme len stĺpce 2 a 3, v ktorých rezíduá indikujú významnú asociáciu. Za úspech budeme považovať možnosť 2, za neúspech možnosť 3. Dáta sú v tabuľke č. 11. Vypočítame odhadnuté šance a pomer šancí.

Tab. č. 11

Možnosti 2 a 3 z tabuľky č. 3

Odpovede na otázku č. 2 / Forma štúdia	Na začiatku štúdia – počas zhromaždenia nových študentov 2	Počas štúdia v prvom trimestri 3	Spolu n_i
Externí	22	35	57
Denní	29	17	46
Spolu n_j	51	52	103

Odhadnutá šanca pre externých študentov $= \dfrac{\frac{22}{57}}{\frac{35}{57}} = \dfrac{22}{35} \approx 0{,}629$

Pre externých študentov pripadá približne 0,629 študenta, ktorý sa dozvedel o etických pravidlách prostredníctvom možnosti 2, na jedného študenta, ktorý sa to dozvedel prostredníctvom možnosti 3.

Odhadnutá šanca pre denných študentov $= \dfrac{\frac{29}{46}}{\frac{17}{46}} = \dfrac{29}{17} \approx 1{,}706$

Pre denných študentov pripadá približne 1,706 študenta, ktorý sa dozvedel o etických pravidlách prostredníctvom možnosti 2, na jedného študenta, ktorý sa to dozvedel prostredníctvom možnosti 3.

Vypočítame pomer šancí pre denných študentov: $\theta = \dfrac{1{,}706}{0{,}629} \approx 2{,}712$

Denný študent má 2,712-krát väčšiu šancu, že sa dozvie o pravidlách etiky prostredníctvom možnosti 2, ako externý študent.

[13] Vypočítanou hodnotou sa len odhaduje skutočná, neznáma šanca v základnom súbore, preto sa hovorí o odhadnutej šanci.

V kontingenčných tabuľkách typu *r* x *c* možno pomer šancí počítať pre dáta v ľubovoľnej subtabuľke typu 2 x 2. To sme realizovali v príklade 2 – pokračovanie 3.

3.4 Sumárne charakteristiky sily asociácie pre tabuľky typu *r* x *c*

Namiesto štúdia asociácie v subtabuľkách 2 x 2 možno charakterizovať silu asociácie v celej tabuľke jediným číslom. Na analýzu nominálnych dát sa používa napríklad charakteristika lambda.[14] V práci [1] sa na s. 239 uvádza: „Niektoré charakteristiky sily asociácie, napríklad koeficient kontingencie (*contingency coefficient*) alebo Cramerovo V (*Cramer` s V*)[15], sú ťažko interpretovateľné a podľa nášho názoru nie sú veľmi užitočné“. Na tej istej strane sa uvádza aj toto: “Veríme, že lepšia predstava o asociácii sa dá získať porovnaním podielov, na základe korigovaných normovaných rezíduí a na základe pomerov šancí v subtabuľkách typu 2 x 2. Tieto metódy sa stávajú vo viacrozmernej analýze oveľa viac preferovanými pred sumárnymi charakteristikami sily asociácie.“

Záver

V článku sme uviedli možnosti analýzy spojenia medzi dvoma nominálnymi premennými. Na príkladoch sme ilustrovali postup pri získavaní podkladu pre rozhodnutie, či existuje asociácia medzi premennými, pomocou testu chí-kvadrát. Na realizáciu testu nie je nevyhnutné používať špecializovaný štatistický softvér, postačí Excel. Konkrétne sme použili štatistickú funkciu CHISQ.TEST.

Keď je identifikovaná asociácia medzi premennými, má zmysel skúmať, ktoré kombinácie hodnôt premenných ju hlavne spôsobujú. Ukázali sme možnosti využitia korigovaných normovaných rezíduí, ktoré umožňujú identifikovať v kontingenčnej tabuľke políčka, viac alebo menej podobné stavu, ktorý by sa očakával pri neexistencii asociácie.

Napokon je užitočné hľadať odpovede na otázku: „Aká silná je asociácia?“ Na charakterizáciu sily asociácie boli použité štatistiky rozdiel podielov a pomer šancí. Tieto štatistiky sú dobre interpretovateľné a poskytujú jasnú predstavu o sile asociácie. Súhlasíme s názorom uvedeným v práci [1], že poskytujú lepšiu predstavu o asociácii ako sumárne charakteristiky.

V článku sme uvažovali len o nominálnych premenných. Keď ide o ordinálne premenné, možno použiť „silnejšie štatistické metódy“, určené pre vyššiu úroveň merania.

Literatúra

[1] AGRESTI, A. – FINLAY, B. (2014): *Statistical Methods for the Social Sciences*. Essex: Pearson. ISBN 978-1-29202-166-9.

[2] ANDERSON, C. J. (dátum neznámy): *Two-Way Tables: Chi-Square Tests Edpsy/Psych/Soc 589.* Dostupné na < http://courses.education.illinois.edu/EdPsy589/lectures/2way_chi-ha-online.pdf >

[14] Podrobnejšie v ([7], s. 34 – 36).

[15] Podrobnejšie v práci ([7], s. 33 – 34).

[3] DIPANKAR BANDYOPADHYAY (dátum neznámy): *Lecture 10: Partitioning Chi Squares and Residual Analysis.* Dostupné na < http://www.biostat.umn.edu/~dipankar/bmtry711.11/lecture_10.pdf >

[4] JOHNSON, N. L. – KEMP, A. W. – KOTZ, S. (2005): *Univariate Discrete Distributions*. Third Edition. USA: J. Wiley and Sons. ISBN 0-471-27246-9.

[5] LEVINE, D. M. – STEPHAN, D. F. – KREHBIEL, T. C. – BERENSON, M. L. (2011): *Statistics for Managers.* Boston: Pearson. ISBN 0-13-611349-4.

[6] MILLER, I. – MILLER, M. (2004): *John E. Freund` s Mathematical Statistics with Applications*. Pearson Prentice Hall. ISBN 0-13-124646-1.

[7] RUBLÍKOVÁ, E. – LABUDOVÁ, V. – SANDTNEROVÁ, S. (2009): *Analýza kategoriálnych údajov*. Bratislava: Ekonóm. ISBN 978-80-2710-1.

[8] TEREK, M. (2014 – 1): *Interpretácia štatistiky a dát. Tretie, doplnené vydanie.* Košice: Equilibria. ISBN 978-80-8143-139-5.

[9] TEREK, M. (2014 – 2): *Interpretácia štatistiky a dát. Podporný učebný materiál. Tretie, doplnené vydanie.* Košice: Equilibria. ISBN 978-80-8143-138-8.

[10] *Analysing Tables. Part V. Interpreting Chi-Square* (dátum neznámy). Dostupné na < http://www.helsinki.fi/~komulain/Tilastokirjat/09.%20Ristiintaulukko.pdf >

[11] *Chi-square Test of Independence* (dátum neznámy). Dostupné na < http://www.geneseo.edu/~bearden/socl211/chisquareweb/chisquare.html >