Milan Terek
Peter Kročitý

ANALÝZA ASOCIÁCIE MEDZI ORDINÁLNYMI PREMENNÝMI V ŠTATISTICKÝCH PRIESKUMOCH[1]

Abstract: *The paper deals with the possibilities of analysis of the association between two ordinal variables in statistical surveys. The procedure of calculating ordinal measure gamma, based on the number of concordant and discordant pairs of observations is described. Then, the possibility of testing the independence of ordinal variables, based on ordinal measure is presented. It is followed by the comparison of the chi-square test of independence and a test based on some ordinal measure. Described procedures and conclusions are illustrated with examples. Finally, there are described the possibilities of analysis of contingence tables with one nominal variable and one ordinal variable.*

Keywords: *ordinal variables, association, chi-square test of independence, concordance and discordance, gamma, test of independence using ordinal measures*

JEL: C 46

Úvod

V posledných rokoch sa na získavanie informácií potrebných pri rozhodovaní výrazne rozšírilo používanie výberových skúmaní na báze štatistických prieskumov. Štatistický prieskum možno charakterizovať ako proces zhromažďovania dát prostredníctvom zisťovania odpovedí respondentov na otázky. Vytvorí sa zoznam otázok, ktoré sa zhromaždia v dotazníku, a získajú sa odpovede respondentov na otázky. V dotazníkoch sa často vyskytujú otázky s odpoveďami, ktoré možno z formálno-matematickej stránky charakterizovať ako ordinálne premenné.

Keď hodnoty kvalitatívnej premennej umožňujú identifikovať znak každej jednotky, stupnica merania kvalitatívnej premennej je nominálna.[2] Kvalitatívne premenné sa merajú v ordinálnej stupnici, keď ich hodnoty majú vlastnosti nominálnych dát a ich usporiadanie má zmysel.[3] Hodnotu ordinálnej premennej, ktorá je v stupnici merania umiestnená bližšie k hornému koncu stupnice, budeme nazývať

[1] Príspevok vznikol s prispením grantovej agentúry VEGA v rámci projektu č. 1/0092/15: *Moderné prístupy k navrhovaniu komplexných štatistických prieskumov.*

[2] Vtedy hovoríme o nominálnych dátach.

[3] Vtedy hovoríme o ordinálnych dátach.

jednoducho vyššou v porovnaní s hodnotou, ktorá je v porovnaní s ňou umiestnená ďalej od horného konca stupnice. Napríklad hodnota „Veľmi dôležitá súčasť štúdia" premennej „Dôležitosť uplatňovania pravidiel akademickej etiky" je vyššia ako jej hodnota „Dôležitá súčasť štúdia". Hodnota „Druhý ročník Bc. štúdia" premennej „Ročník Bc. štúdia" je vyššia ako jej hodnota „Prvý ročník Bc. štúdia".

Všimneme si metódy detekcie a opisu spojenia medzi dvoma kvalitatívnymi, ordinálnymi 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*).

Dáta na 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. Všimneme si analýzu kontingenčnej tabuľky, v ktorej sú obe premenné ordinálne.

1 Asociácia medzi ordinálnymi premennými

Štatistická analýza ordinálnych dát berie do úvahy usporiadanie hodnôt premenných. Uvedieme, ako možno rozhodnúť, či medzi dvoma ordinálnymi premennými existuje asociácia a možnosti merania sily tejto asociácie. Postupy budeme ilustrovať na príklade štúdia asociácie medzi dvoma ordinálnymi premennými, ktoré sa analyzovali pri štatistickom prieskume zameranom na manažment znalostí o akademickej etike na Vysokej škole manažmentu v Trenčíne/City University of Seattle. Náhodne vybratým študentom bol poslaný dotazník, ktorý obsahoval 8 otázok. Jeho súčasťou bola aj otázka č. 6: "Posúďte dôležitosť uplatňovania pravidiel akademickej etiky na VŠM/CU". Analýzu odpovedí na túto otázku v členení podľa ročníkov štúdia si všimneme podrobnejšie.

Príklad 1. Skúma sa názor študentov VŠM/CU na dôležitosť uplatňovania pravidiel akademickej etiky na ich vysokej škole. Výsledky prieskumu sú v tabuľke č. 1. V tabuľke je združené rozdelenie absolútnych početností, v zátvorkách sú hodnoty podmieneného rozdelenia početností premennej „Dôležitosť uplatňovania pravidiel akademickej etiky" v percentách.

Tab. č. 1

Názor študentov VŠM/CU na dôležitosť uplatňovania pravidiel akademickej etiky

Dôležitosť Ročník štúdia	Menej dôležitá, resp. zbytočná súčasť štúdia, resp. neviem posúdiť	Dôležitá súčasť štúdia	Veľmi dôležitá súčasť štúdia	Spolu
Prvý ročník Bc. štúdia	13 (19,40)	35 (52,24)	19 (28,36)	67
Druhý ročník Bc. štúdia	16 (30,19)	24 (45,28)	13 (24,53)	53
Tretí ročník Bc. štúdia	4 (6,56)	30 (49,18)	27 (44,26)	61
Spolu	33	89	59	181

Podmienené rozdelenie v percentách (19,40; 52,24; 28,36) ukazuje dôležitosť uplatňovania pravidiel akademickej etiky u vybratých študentov prvého ročníka, podmienené rozdelenie (30,19; 45,28; 24,53) u vybratých študentov druhého ročníka a (6,56; 49,18; 44,26) u vybratých študentov tretieho ročníka. Vo výberovom súbore druháci považujú uplatňovanie pravidiel akademickej etiky za menej dôležité ako prváci, naopak, výrazná väčšina tretiakov ich považuje za dôležitú až veľmi dôležitú súčasť štúdia.

V ordinálnych dátach sa môžu vyskytnúť hlavne dva typy asociácie medzi dvoma premennými – pozitívna a negatívna. Pri pozitívnej asociácii má jednotka s vyššou hodnotou jednej premennej tendenciu mať súčasne vyššiu hodnotu druhej premennej a jednotka s nižšou hodnotou jednej premennej má tendenciu mať aj nižšiu hodnotu druhej premennej. Napríklad, keby študenti v nižších ročníkoch prikladali menšiu dôležitosť uplatňovaniu pravidiel akademickej etiky a študenti vo vyšších ročníkoch by jej prikladali väčšiu dôležitosť, išlo by o pozitívnu asociáciu medzi ročníkom štúdia a názorom na uplatňovanie pravidiel akademickej etiky.

Pri negatívnej asociácii má jednotka s vyššou hodnotou jednej premennej tendenciu mať súčasne nižšiu hodnotu druhej premennej a jednotka s nižšou hodnotou jednej premennej má tendenciu mať vyššiu hodnotu druhej premennej. Keby študenti v nižších ročníkoch prikladali väčšiu dôležitosť uplatňovaniu pravidiel akademickej etiky a študenti vo vyšších ročníkoch menšiu, išlo by o negatívnu asociáciu medzi ročníkom štúdia a názorom na dôležitosť uplatňovania pravidiel akademickej etiky.

1 Zhoda a nezhoda

Mnoho charakteristík asociácie je založených na informácii o spojeniami medzi všetkými dvojicami pozorovaní.

Dvojica pozorovaní pre dve jednotky je zhodná (*concordant),* keď jednotka, ktorá má vyššiu hodnotu jednej premennej, má vyššiu aj hodnotu druhej premennej.

Dvojica pozorovaní pre dve jednotky je nezhodná (*discordant*), keď jednotka, ktorá má vyššiu hodnotu jednej premennej, má nižšiu hodnotu druhej premennej.

V tabuľke č. 1 je hodnota „Menej dôležitá, resp. zbytočná súčasť štúdia, resp. neviem posúdiť“ na dolnom konci a hodnota „Veľmi dôležitá súčasť štúdia“ na hor-

nom konci stupnice merania premennej „Dôležitosť uplatňovania pravidiel akademickej etiky“. Hodnota „Prvý ročník Bc. štúdia“ je na dolnom konci a hodnota „Tretí ročník Bc. štúdia“ je na hornom konci stupnice merania premennej „Ročník štúdia“. Kontingenčná tabuľka pre ordinálne premenné sa obyčajne konštruuje tak, že najnižšia hodnota premennej s hodnotami v prvom stĺpci je v prvom riadku a najnižšia hodnota premennej s hodnotami v prvom riadku je v prvom stĺpci. Tak je to aj v tabuľke č. 1.

Uvažujme teraz o jednotke s hodnotami premenných (Menej dôležitá, resp. zbytočná súčasť štúdia, resp. neviem posúdiť; Prvý ročník Bc. štúdia) a o inej jednotke s hodnotami premenných (Dôležitá súčasť štúdia; Druhý ročník Bc. štúdia). Prvá jednotka je jedna z 13 jednotiek v ľavom hornom políčku tabuľky č. 1. Druhá jednotka je jedna z 24 jednotiek v strednom políčku tabuľky č. 1. Dvojica pozorovaní pre tieto dve jednotky je zhodná, pretože druhá jednotka má vyššiu hodnotu prvej aj druhej premennej. Každá zhodná dvojica svedčí v prospech pozitívnej asociácie. Každá z 13 jednotiek s hodnotami premenných (Menej dôležitá, resp. zbytočná súčasť štúdia, resp. neviem posúdiť; Prvý ročník Bc. štúdia) môže tvoriť dvojicu s 24 jednotkami s hodnotami premenných (Dôležitá súčasť štúdia; Druhý ročník Bc. štúdia). Potom máme $(13 \cdot 24) = 312$ zhodných párov pozorovaní v týchto dvoch políčkach tabuľky 1.

Naopak, každá z 35 jednotiek s hodnotami premenných (Dôležitá súčasť štúdia; Prvý ročník Bc. štúdia) tvorí nezhodnú dvojicu s každou zo 16 jednotiek s hodnotami premenných (Menej dôležitá, resp. zbytočná súčasť štúdia, resp. neviem posúdiť; Druhý ročník Bc. štúdia). 35 jednotiek má vyššiu hodnotu premennej „Dôležitosť uplatňovania pravidiel akademickej etiky“, ale nižšiu hodnotu premennej „Ročník štúdia“. V týchto dvoch políčkach tabuľky je $(35 \cdot 16) = 560$ nezhodných párov pozorovaní. Každá nezhodná dvojica svedčí v prospech negatívnej asociácie.

Symbolom C označíme celkový počet zhodných párov pozorovaní a symbolom D celkový počet nezhodných párov pozorovaní v kontingenčnej tabuľke. Hodnoty C a D možno vypočítať podľa vzťahov (Agresti, 2010 [1], s. 186):

$$C = \sum_{i<k} \sum \sum_{j<l} \sum n_{ij} n_{kl} \quad \text{a} \qquad D = \sum_{i<k} \sum \sum_{j>l} \sum n_{ij} n_{kl} \tag{1}$$

kde n_{ij} (i, k sú indexy hodnôt prvej premennej; j, l sú indexy hodnôt druhej premennej) sú absolútne početnosti v kontingenčnej tabuľke (hodnoty združeného rozdelenia absolútnych početností).

Príklad 1 − pokračovanie 1 V príklade 1 vypočítame hodnoty C a D.

Počítame podľa vzťahu (1). Výpočet C začneme v ľavom hornom (severozápadnom) políčku tabuľky č. 1 (v druhom riadku a v druhom stĺpci). Podobne pre všetky ostatné políčka tabuľky, treba príslušnú početnosť vynásobiť so súčtom početností v políčkach s vyššími hodnotami oboch premenných:

$$C = 13(24 + 13 + 30 + 27) + 35(13 + 27) + 16(30 + 27) + 24 \cdot 27 = 4182$$

Výpočet D sa začína v pravom hornom (severovýchodnom) políčku tabuľky s najvyššou hodnotou prvej premennej a s najnižšou hodnotou druhej premennej. Početnosť v tomto políčku vynásobíme so súčtom početností vo všetkých políčkach s vyššou hodnotou jednej premennej a nižšou hodnotou druhej premennej. Podobne pre všetky ostatné políčka tabuľky treba príslušnú početnosť vynásobiť so súčtom početností v políčkach, s vyššou hodnotou jednej premennej a nižšou hodnotou druhej premennej:

$$D = 19(16 + 24 + 4 + 30) + 13(4 + 30) + 35(16 + 4) + 24 \cdot 4 = 2644$$

Viac dvojíc svedčí o pozitívnej ako o negatívnej asociácii.

2 Charakteristika asociácie gama

Z ($C + D$) dvojíc pozorovaní, ktoré sú zhodné alebo nezhodné, je podiel $C/(C + D)$ zhodných a podiel $D/(C + D)$ nezhodných. Rozdiel medzi týmito podielmi sa nazýva gama[4] :

$$\hat{\gamma} = \frac{C - D}{C + D} \tag{2}$$

Hodnota $\hat{\gamma} \in [-1, 1]$, kladná hodnota, naznačuje pozitívnu asociáciu, a záporná hodnota negatívnu asociáciu. Čím väčšia je absolútna hodnota gama, tým je asociácia silnejšia.

Príklad 1 − pokračovanie 2 V príklade 1 vypočítame hodnotu gama podľa (2).

$$\hat{\gamma} = \frac{C - D}{C + D} = \frac{4182 - 2644}{4182 + 2644} \approx 0{,}225$$

Hodnota $\hat{\gamma}$ indikuje pomerne slabú pozitívnu asociáciu medzi ročníkom štúdia a názorom na dôležitosť uplatňovania pravidiel akademickej etiky.

V základnom súbore je gama definovaná takto:

$$\gamma = \frac{\pi_C - \pi_D}{\pi_C + \pi_D}$$

kde π_C je pravdepodobnosť zhody pre náhodne vybratú dvojicu pozorovaní a π_D je pravdepodobnosť nezhody pre náhodne vybratú dvojicu pozorovaní. Hodnotu charakteristiky γ odhadujeme hodnotou $\hat{\gamma}$.

Okrem charakteristiky gama sú známe aj iné, podobné charakteristiky – Kendallovo tau-b a tau-c, Spearmanovo ró-b a ró-c a Somersovo d.[5]

[4] Táto charakteristika bola navrhnutá Goodmanom a Kruskalom v roku 1954.

[5] Podrobnejšie o týchto charakteristikách pozri v Agresti, 2010 ([1], s. 188 – 190).

3 Testovanie nezávislosti na báze *C* a *D*

Všimnime si teraz vzťah medzi premennými v základnom súbore. Budeme testovať nulovú hypotézu H_0: premenné sú štatisticky nezávislé oproti jednej z alternatívnych hypotéz: $H_1: \pi_C - \pi_D \neq 0$, $H_1: \pi_C - \pi_D > 0$ alebo $H_1: \pi_C - \pi_D < 0$. Hodnota ordinálnej charakteristiky $\pi_C - \pi_D$ je záporná, nulová alebo kladná, podobne ako gama.

Pre veľké náhodné výbery[6] má náhodná premenná $(C - D)$ približne normálne rozdelenie. Test nezávislosti možno založiť na testovacej štatistike s normovaným normálnym rozdelením, ktorej hodnota sa vypočíta podľa vzťahu:

$$z = \frac{C - D}{\sigma_{C-D}} \tag{3}$$

kde σ_{C-D} je smerodajná odchýlka náhodnej premennej $(C - D)$, keď je H_0 správna. Príslušný rozptyl možno vypočítať podľa vzťahu (Agresti, 2010 [1], s. 196):

$$\sigma^2_{C-D} = \frac{n(n-1)(2n+5) - \sum_i n_i(n_i-1)(2n_i+5)}{18} - \frac{\sum_j n_j(n_j-1)(2n_j+5)}{18} + \tag{4}$$

$$+ \frac{\left[\sum_i n_i(n_i-1)(n_i-2)\right]\left[\sum_j n_j(n_j-1)(n_j-2)\right]}{9n(n-1)(n-2)} + \frac{\left[\sum_i n_i(n_i-1)\right]\left[\sum_j n_j(n_j-1)\right]}{2n(n-1)}$$

kde *n* je súčet všetkých absolútnych početností n_{ij} v kontingenčnej tabuľke,

n_i je súčet všetkých hodnôt n_{ij} v *i*-tom riadku,

n_j je súčet všetkých hodnôt n_{ij} v *j*-tom stĺpci.

Alternatívne možno použiť Waldov test, ktorý používa pri výpočte testovacej štatistiky $\hat{\gamma}$. V Agresti, 2010 ([1], s. 197) sa odporúča preferovať testovaciu štatistiku, ktorej hodnota sa vypočíta podľa (3).

Test nezávislosti založený na jednej z ordinálnych charakteristík sa obyčajne preferuje pred testom nezávislosti chí-kvadrát, keď sú obe premenné ordinálne. Testovacia štatistika χ^2 totiž ignoruje usporiadanie hodnôt ordinálnych premenných. Keď je prítomná pozitívna alebo negatívna asociácia, ordinálne charakteristiky sú na jeho detekciu obyčajne silnejšie.

[6] Odporúča sa, aby *C* aj *D* bolo väčšie ako 50 (Agresti, Finlay, 2014 [2], s. 243).

Niekedy však môže byť test nezávislosti χ^2 silnejší aj pre ordinálne dáta. Totiž, štatistická nezávislosť premenných implikuje $\pi_C - \pi_D = 0$, ale opačná implikácia neplatí. Môže byť $\pi_C - \pi_D = 0$ a premenné nemusia byť štatisticky nezávislé. Trendom asociácie nazveme tendenciu vývoja asociácie medzi jednou a druhou premennou. Možno ju charakterizovať napríklad množinou hodnôt $\hat{\gamma}$, ktoré hodnoty sa počítajú pre všetky postupnosti dvoch susedných riadkov a pre všetky postupnosti dvoch susedných stĺpcov v kontingenčnej tabuľke. Keď bude charakter asociácie rovnaký (všetky hodnoty $\hat{\gamma}$ budú kladné alebo budú všetky záporné), pôjde o jednoduchý trend asociácie (pozitívny alebo negatívny), v opačnom prípade pôjde o komplexný trend asociácie. V prípade komplexného trendu môže byť test nezávislosti χ^2 silnejší aj pre ordinálne dáta (Agresti, Finlay, 2014 [2], s. 245). Všimnime si hypotetické dáta v tabuľke č. 2.

Pre dáta v tabuľke č. 2 je $(C - D) = 0$, aj $\hat{\gamma} = 0$. Uvažujme teraz v tabuľke č. 2 len o prvých dvoch dátových riadkoch. Pre tieto dáta vyjde $\hat{\gamma} = 1$. Ide o najsilnejšiu pozitívnu asociáciu. Keď rastie *x* , rastie aj *y.* Keď v tabuľke č. 2 uvažujeme len o posledných dvoch dátových riadkoch, vyjde $\hat{\gamma} = -1$. Tu ide o najsilnejšiu negatívnu asociáciu. Keď rastie *x*, klesá *y.*

Tab. č. 2.

Hypotetické dáta

Premenná *y* Premenná *x*	Nízky	Vysoký
Veľmi nízky	100	0
Nízky	0	100
Vysoký	0	100
Veľmi vysoký	100	0

Takáto situácia však neindikuje štatistickú nezávislosť príslušných náhodných premenných. Všeobecne môže byť $\pi_C - \pi_D = 0$, alebo iná ordinálna charakteristika asociácie sa môže rovnať nule, keď sú náhodné premenné štatisticky závislé, ale trend asociácie je komplexný. Vtedy môže byť test nezávislosti χ^2 lepší ako test založený na ordinálnej charakteristike.

Príklad 1 − pokračovanie 3 V príklade 1 vykonáme na hladine významnosti 0,01 test nezávislosti χ^2 a potom test nezávislosti na báze *C* a *D* s alternatívnou hypotézou $H_1: \pi_C - \pi_D \neq 0$.

Test nezávislosti χ^2 sme vykonali pomocou štatistickej funkcie CHISQ.TEST v Exceli.[7] Vyšla *p*-hodnota[8] rovná 0,011119. Na hladine významnosti 0,01 nezamietame predpoklad o štatistickej nezávislosti medzi ročníkom štúdia a názorom študentov na dôležitosť uplatňovania pravidiel akademickej etiky.

Na hladine významnosti 0,01 teraz vykonáme test nezávislosti na báze *C* a *D*. Najprv vypočítame σ^2_{C-D} podľa (4).

$$\sigma^2_{C-D} = \frac{181\cdot 180\cdot 367 - (67\cdot 66\cdot 139 + 53\cdot 52\cdot 111 + 61\cdot 60\cdot 127)}{18} -$$

$$- \frac{33\cdot 32\cdot 71 + 89\cdot 88\cdot 183 + 59\cdot 58\cdot 123}{18} +$$

$$+ \frac{(67\cdot 66\cdot 65 + 53\cdot 52\cdot 51 + 61\cdot 60\cdot 59)(33\cdot 32\cdot 31 + 89\cdot 88\cdot 87 + 59\cdot 58\cdot 57)}{9\cdot 181\cdot 180\cdot 179} +$$

$$+ \frac{(67\cdot 66 + 53\cdot 52 + 61\cdot 60)(33\cdot 32 + 89\cdot 88 + 59\cdot 58)}{2\cdot 181\cdot 180} \approx 493\ 330{,}9904$$

Potom σ_{C-D} = 702,375249.

Hodnotu testovacej štatistiky vypočítame podľa (3):

$$z = \frac{4182 - 2644}{702{,}375249} \approx 2{,}19$$

P-hodnota[9] je približne 0,028524. Na hladine významnosti 0,01 nezamietame predpoklad o štatistickej nezávislosti medzi ročníkom štúdia a názorom študentov na dôležitosť uplatňovania pravidiel akademickej etiky. Dospeli sme k rovnakému záveru ako pri aplikácii testu χ^2.

Príklad 1 − pokračovanie 4 Preskúmame v kontingenčnej tabuľke z príkladu 1 trend asociácie, pomocou charakteristiky $\hat{\gamma}$.

Všimnime si len prvé dva riadky v tabuľke **č.** 1. Vyjde $\hat{\gamma} \approx -0{,}1702$. Výsledok indikuje negatívnu asociáciu. Dáta z výberu naznačujú, že druháci považujú uplatňovanie pravidiel akademickej etiky za menej dôležité ako prváci. Keď zvlášť analyzujeme len druhý a tretí riadok v tabuľke 1, dostaneme $\hat{\gamma} \approx 0{,}4871$. Výsledok

[7] Podrobný postup možno nájsť v Terek, 2014 - 2 [5].
[8] Podrobnejšie o *p*-hodnote pozri v Terek, 2014 - 1 [4], s. 155 – 156.
[9] Spôsob výpočtu *p*-hodnoty v Exceli možno nájsť v Terek, 2014 - 2 [5].

indikuje pomerne silnú pozitívnu asociáciu. Dáta z výberu naznačujú, že tretiaci považujú uplatňovanie pravidiel akademickej etiky za dôležitejšie ako druháci, pričom rozdiel medzi nimi je pomerne významný. Toto je určite užitočná informácia pre manažment znalostí o akademickej etike na VŠM/CU.

V príklade 1 – pokračovanie 4 sme ukázali, že existujú silné indície, že vzťah medzi uvažovanými premennými nemá jednoduchý trend. Vtedy môže byť test nezávislosti χ^2 lepší ako test založený na ordinálnej charakteristike. To sa aj stalo, pretože v teste nezávislosti χ^2 aj v teste založenom na C a D sme síce dospeli k rovnakému záveru – na hladine významnosti 0,01 nezamietame nulovú hypotézu o štatistickej nezávislosti premenných, ale v teste χ^2 je p-hodnota len 0,011119 a vyjadruje menšiu podporu nulovej hypotéze ako v teste založenom na C a D, v ktorom sa p-hodnota rovná 0,028524. Keďže máme silné indície o závislosti premenných, je v tomto prípade test χ^2 skutočne lepší.

V príklade 1 – pokračovanie 5 budeme uvažovať o hypotetickej situácii, v ktorej sa v tabuľke 1 vymenia dáta v prvom a druhom riadku. Možno sa ľahko presvedčiť, že po takejto modifikácii dáta z výberu naznačujú jednoduchý pozitívny trend asociácie.

Príklad 1 − pokračovanie 5 V tabuľke č. 3 sú hypotetické dáta, ktoré vznikli tak, že sa v tabuľke č. 1 vymenili dáta v prvom a druhom riadku. V kontingenčnej tabuľke preskúmame trend asociácie. Na hladine významnosti 0,01 vykonáme test nezávislosti χ^2 a potom test nezávislosti na báze *C* a *D* s alternatívnou hypotézou $H_1: \pi_C - \pi_D \neq 0$.

Tab. č. 3

Hypotetické dáta na základe dát v tabuľke č. 1

Dôležitosť Ročník štúdia	Menej dôležitá, resp. zbytočná súčasť štúdia, resp. neviem posúdiť	Dôležitá súčasť štúdia	Veľmi dôležitá súčasť štúdia	Spolu
Prvý ročník Bc. štúdia	16 (30,19)	24 (45,28)	13 (24,53)	53
Druhý ročník Bc. štúdia	13 (19,40)	35 (52,24)	19 (28,36)	67
Tretí ročník Bc. štúdia	4 (6,56)	30 (49,18)	27 (44,26)	61
Spolu	33	89	59	181

Všimnime si najprv len prvé dva riadky v tabuľke **č.** 3. Vyjde $\hat{\gamma} \approx 0{,}1702$. Výsledok indikuje pozitívnu asociáciu. Keď zvlášť analyzujeme len druhý a tretí riadok v tabuľke č. 3, dostaneme $\hat{\gamma} \approx 0{,}3641$. Podobne, keď zvlášť analyzujeme prvý a druhý stĺpec a potom druhý a tretí stĺpec, dostaneme kladné hodnoty $\hat{\gamma}$. Dáta z výberu naznačujú jednoduchý pozitívny trend. V takomto prípade by mal byť test na základe *C* **a** *D* lepší.

V teste χ^2 *p*-hodnota vyšla rovnako ako predtým – 0,011119; tento test nerozlišuje usporiadanie hodnôt premenných. Na hladine významnosti 0,01 nezamietame H_0.

V teste založenom na C a D vypočítame hodnotu testovacej štatistiky:

$$z = \frac{4566 - 2260}{702{,}375249} \approx 3{,}28$$

P-hodnota vyjde 0,001038. Na hladine významnosti 0,01 zamietame H_0 a prijímame alternatívnu hypotézu H_1, ktorá hovorí o štatistickej závislosti premenných.

Keď budeme uvažovať o alternatívnej hypotéze $H_1 : \pi_C - \pi_D > 0$, vyjde p-hodnota 0,000519. Na hladine významnosti 0,01 zamietame H_0 a prijímame alternatívnu hypotézu H_1 ktorá hovorí o štatistickej závislosti premenných, s pozitívnou asociáciou.

V príklade má asociácia jednoduchý pozitívny trend. V takýchto prípadoch je použitie testu založeného na C a D alebo na inej ordinálnej charakteristike lepšie ako použitie testu nezávislosti χ^2.

Alternatívny prístup k analýze trendu asociácie v kontingenčnej tabuľke s ordinálnymi premennými spočíva v priradení skóre hodnotám premenných. Potom možno premenné analyzovať ako kvantitatívne a v rámci korelačnej analýzy napríklad testovať štatistickú významnosť lineárnej regresie prostredníctvom testu o koeficiente korelácie s využitím testovacej štatistiky Z.

Vždy, keď je to možné, je lepšie uvažovať o čo najväčšom počte hodnôt ordinálnych premenných. Je lepšie napríklad uvažovať o štyroch alebo piatich hodnotách premennej, ako o dvoch. Smerodajné chyby charakteristík majú totiž tendenciu byť pri viacerých hodnotách premenných, pri rovnakom rozsahu výberu, menšie. Okrem toho „jemnejšie meranie" je výhodnejšie vtedy, keď sa dáta analyzujú ako kvantitatívne a používajú sa „silnejšie" metódy (Agresti, Finlay, 2014 [2], s. 245).

4 Analýza kontingenčných tabuliek s nominálnou a ordinálnou premennou

Pre analýzu kontingenčnej tabuľky s ordinálnou premennou a s nominálnou premennou, ktorá má dve hodnoty, sú postupy, ktoré sme uviedli, platné. V tomto prípade znamienko charakteristiky ukazuje, ktorá hodnota nominálnej premennej je spojená s vyššou hodnotou ordinálnej premennej. Predpokladajme napríklad, že by nás zaujímalo spojenie medzi pohlavím študenta (žena, muž) a dôležitosťou, ktorú študent prisudzuje dodržiavaniu pravidiel akademickej etiky (menej dôležité, dôležité, veľmi dôležité). Predpokladajme napríklad, že vyjde $\hat{\gamma} = -0{,}2$. Keďže výsledok je záporné číslo, „vyššia" hodnota pohlavia, t. j. muž, má tendenciu objavovať sa s menšou dôležitosťou dodržiavania pravidiel. Výsledok indikuje negatívnu asociáciu.

Keď má nominálna premenná viac ako dve hodnoty, nie je vhodné použiť ordinálnu charakteristiku, napríklad gama. V takýchto prípadoch možno uplatniť postup, v ktorom sa hodnotám ordinálnej premennej priradia skóre a považuje sa za kvantitatívnu premennú. Potom možno napríklad pomocou analýzy rozptylu[10] porovnávať

[10] Viac o analýze rozptylu, pozri napríklad v Anderson et al., 2007 [3].

stredné hodnoty skupín, definovaných hodnotami nominálnej premennej. Niekedy je vhodné použiť logistickú regresiu[11], v ktorej je závisle premenná nominálna. Vtedy hodnotám ordinálnej premennej netreba priraďovať skóre.

Záver

V článku boli ukázané možnosti analýzy asociácie medzi dvoma ordinálnymi premennými. Boli definované dva typy asociácie medzi dvoma premennými – pozitívna a negatívna a tiež zhodné a nezhodné dvojice pozorovaní. Na základe počtu zhodných a nezhodných počtov pozorovaní možno formulovať napríklad charakteristiku asociácie gama.

Testovanie nezávislosti na báze *C* a *D* umožňuje prijať rozhodnutie o štatistickej nezávislosti dvoch ordinálnych premenných. Test na báze *C* a *D* je ale lepší ako test nezávislosti χ^2 len vtedy, keď je medzi premennými jednoduchý trend asociácie. Keď medzi premennými nie je jednoduchý trend asociácie, môže byť test nezávislosti χ^2 lepší ako test založený na niektorej z ordinálnych charakteristík. Platnosť uvedených záverov bola ilustrovaná na príkladoch.

Alternatívny prístup k analýze trendu asociácie **v** kontingenčnej tabuľke s ordinálnymi premennými spočíva v priradení skóre hodnotám premenných. Potom možno premenné analyzovať ako kvantitatívne.

Nakoniec boli uvedené možnosti analýzy kontingenčných tabuliek s jednou nominálnou a jednou ordinálnou premennou.

Literatúra

[1] AGRESTI, A. (2010): *Analysis of Ordinal Categorical Data*. New York: Wiley and Sons. ISBN 978-0-470-08289-8.

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

[3] ANDERSON, D. R. – SWEENEY, D. J. – WILLIAMS, T. A. – FREEMAN, J. – SHOESMITH, E. (2007): *Statistics for Business and Economics*. USA: Thomson Learning, 2007. ISBN 978-1-84480-313-2.

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

[5] TEREK, M. – HORNÍKOVÁ, A. – LABUDOVÁ, V. (2010): *Hĺbková analýza údajov.* Bratislava: IURA EDITION, 2010. ISBN 978-80-8078-336-5.

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

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

[8] TEREK, M. – KROČITÝ, P. (2014): Analýza asociácie medzi nominálnymi premennými v štatistických prieskumoch. In: *Ekonomické rozhľady* 4/2014.

[11] Viac o logistickej regresii pozri napríklad v Terek et al., 2010 [5] alebo v Rublíková et al., 2009 [4].