Analýza dat v neurologii

XXI. Kontingenční tabulky: test nezávislosti kategoriálních znaků

L. Dušek, T. Pavlík,
I. Jarkovský, J. Koptíková

Institut biostatistiky a analýz
Masarykova univerzita, Brno

✉
doc. RNDr. Ladislav Dušek, Dr.
Institut biostatistiky a analýz
Masarykova univerzita, Brno
e-mail: dusek@cba.muni.cz

Úvod: co je kontingenční tabulka

V tomto díle seriálu budeme pokračovat ve výkladu testů pro kategoriální data; na řadě je velmi často používané hodnocení vztahu mezi dvěma kategoriálními znaky. Jde např. o zkoumání vztahu mezi charakteristikami pacienta a výskytem komplikací při léčbě, sledování souvislostí ve výskytu znaku u rodičů a dětí nebo testy vzájemné nezávislosti prognostických faktorů. Výklad nám usnadní předchozí díl seriálu, kde jsme vysvětlili tzv. test dobré shody. Tento test hodnotí pomocí χ^2 statistiky rozdíly mezi očekávanými a pozorovanými četnostmi kategorií znaků.

Hodnotíme-li vzájemnou souvislost dvou znaků (např. pohlaví pacienta a komplikace při léčbě), musíme oba znaky sledovat u náhodného výběru N pacientů. Výskyt obou znaků považujeme za náhodný. Získaná data zapisujeme do tabulky, která staví výskyt kategorií obou znaků proti sobě a obsahuje pozorované četnosti jednotlivých kategorií. Řádky tabulky odpovídají možným hodnotám (kategoriím) prvního znaku, sloupce pak možným hodnotám (kategoriím) druhého znaku. Jde o tzv. **kontingenční tabulku** *(contingency table)*. Nejjednodušší verzi tabulky (dva znaky, každý jen o dvou kategoriích) označujeme jako tabulku 2 × 2 (příklad 1), při více kategoriích pak obecně hovoříme o tabulce $R \times C$, kde R je počet řádků a C počet sloupců tabulky (příklad 2). Toto označení je používáno i mezinárodně. Nejjednodušší verze tabulky 2 × 2 bývá někdy označována jako čtyřpolní tabulka četností (sloužící ke srovnání dvou binárních neboli dichotomických znaků).

Znaky zpracovávané v kontingenční tabulce musí nabývat diskrétních hodnot,

	X+	X-	
Y+	a	b	a + b
Y-	c	d	c + d
	a + c	b + d	N

Poznámky:

Alternativní formou výpočtu v bodě 4 je následující vztah:

$$\chi^2_{\nu=1} = n\,\frac{(ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$

Výpočet statistiky χ^2 s využitím tzv. Yatesovy korekce (korekce používaná pokud pozorovaná četnost v kterémkoli poli tabulky klesne pod 5; viz též text článku):

$$\chi^2_{\nu=1} = n\,\frac{\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+b)(a+c)(c+d)(b+d)}$$

1. Je dána kontingenční tabulka dvou proměnných X a Y, každá z nich nabývá dvou možných hodnot "+" a "–", jde tedy o 2 × 2 kontingenční tabulku.
2. Celkově čtyři možné kombinace hodnot dvou binárních proměnných v tabulce: a až d značí četnosti (počty) případů, které nastaly v jednotlivých kombinacích. Platí, že $a + b + c + d = N$.
3. Cílem je testovat nezávislost proměnných X a Y pomocí testu dobré shody, kdy testujeme rozdíl mezi četnostmi pozorovanými v tabulce a tzv. očekávanými četnostmi. Očekávané četnosti kalkulujeme tak, aby odpovídaly teoretické situaci, kdy jsou proměnné X a Y zcela nezávislé.
4. Testová statistika je pro kontingenční tabulku 2 × 2 dána rovnicí

$$\chi^2_{\nu=1} = \frac{(f_a - F_a)^2}{F_a} + \frac{(f_b - F_b)^2}{F_b} + \frac{(f_c - F_c)^2}{F_c} + \frac{(f_d - F_d)^2}{F_d}$$

   kde $\boldsymbol{f_{a-d}}$ jsou pozorované četnosti v buňkách tabulky, $\boldsymbol{F_{a-d}}$ jsou očekávané četnosti v buňkách tabulky.
5. Očekávané četnosti jsou vypočteny následovně:

$$F_a = \frac{(a+b)(a+c)}{N} \qquad F_c = \frac{(a+c)(d+c)}{N}$$

$$F_b = \frac{(a+b)(b+d)}{N} \qquad F_d = \frac{(b+d)(c+d)}{N}$$

6. Po dosazení do vztahu v bodě 4 je vypočtena hodnota testové statistiky χ^2 pro ν stupňů volnosti; ν = (počet řádků – 1) * (počet sloupců – 1) = 1.
7. Vypočtená hodnota testové statistiky je porovnána s kritickou hodnotou, která je například pro $\alpha = 0{,}05$ $\chi^2_{(0{,}95;\,\nu=1)} = 3{,}84$.

Příklad 1. Vzorový výpočet testu nezávislosti znaků pro 2 × 2 kontingenční tabulku.

	Y_1	Y_2	Y_3	Y_4
X_1	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$
X_2	$f_{2,1}$	..	..	..
X_3	$f_{3,1}$	..	..	$f_{i,j}$

1. Je dána kontingenční tabulka dvou proměnných *X* a *Y*, každá z nich nabývá více možných hodnot (kategorií) $X_{1..R}$ a $Y_{1..C}$
2. Kombinace proměnných jsou označeny jako $f_{x,y}$, kde *f* je pozorovaná četnost kombinace a *i*, *j* jsou indexy pro daný řádek a sloupec tabulky
3. Cílem je testovat nezávislost proměnných *X* a *Y* pomocí testu dobré shody, kdy testujeme rozdíl mezi pozorovanými četnostmi v tabulce a tzv. očekávanými četnostmi. Očekávané četnosti kalkulujeme tak, aby odpovídaly teoretické situaci, kdy jsou proměnné *X* a *Y* zcela nezávislé.
4. Testová statistika je pro kontingenční tabulku *R* × *C* dána vztahem

$$\chi^2_{(R-1)*(C-1)} = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(f_{i,j} - F_{i,j})^2}{F_{i,j}}$$

kde $\boldsymbol{f}_{i,j}$ jsou pozorované četnosti v buňkách tabulky, $\boldsymbol{F}_{i,j}$ jsou očekávané četnosti v buňkách tabulky.

5. Očekávané četnosti jsou kalkulovány podle vztahu

$$F_{i,j} = \frac{\sum_{z=1}^{C} f_{i,z} * \sum_{z=1}^{R} f_{z,j}}{N}$$

6. Po dosazení do vztahu v bodě 4 je vypočtena hodnota testové statistiky χ^2 pro ν stupňů volnosti; ν = (počet řádků – 1) * (počet sloupců – 1) stupňů volnosti.
7. Vypočtená hodnota testové statistiky je porovnána s kritickou hodnotou testu.

Příklad 2. Vzorový výpočet testu nezávislosti znaků pro R × C kontingenční tabulku.

Příklad A	muž	žena	
alergie ano	16	6	22
alergie ne	3	24	27
	19	30	49

Příklad B	muž	žena	
alergie ano	7	15	22
alergie ne	12	15	27
	19	30	49

1. Analyzujeme vztah mezi pohlavím a výskytem alergie ve dvou skupinách pacientů – příklady A a B. Nulovou hypotézou je nezávislost výskytu alergie na pohlaví.
2. Hypotézu testujeme pomocí testu dobré shody pro 2 × 2 kontingenční tabulku dle postupu, který je obecně popsán v příkladě 1.
3. Po dosazení do vztahu je vypočtena hodnota testové statistiky pro příklady A a B:

$$\text{A}: \chi^2_{\nu=1} = \frac{(16-8.5)^2}{8.5} + \frac{(6-10.5)^2}{10.5} + \frac{(3-13.5)^2}{13.5} + \frac{(24-16.5)^2}{16.5} = 19.4$$

$$\text{B}: \chi^2_{\nu=1} = \frac{(7-8.5)^2}{8.5} + \frac{(15-10.5)^2}{10.5} + \frac{(12-13.5)^2}{13.5} + \frac{(15-16.5)^2}{16.5} = 0.8$$

4. Vypočtená hodnota testové statistiky je porovnána s kritickou hodnotou testu, která je pro $\alpha = 0{,}05$ $\chi^2_{(0{,}95;\ \nu=1)} = 3{,}84$.
5. Výsledné hodnoty statistické významnosti jsou:
   Příklad A: $p < 0{,}001$ – zamítáme nulovou hypotézu o nezávislosti znaků
   Příklad B: $p = 0{,}367$ – nezamítáme nulovou hypotézu o nezávislosti znaků

Příklad 3. Ukázka výpočtu testu nezávislosti dvou znaků v 2 × 2 kontingenční tabulce.

musí tedy jít o data nominální nebo ordinální. Spojité (kvantitativní) znaky v tabulce četností zpracovat nelze, je ale možné je rozdělit do tříd a tyto jako kategorie následně do tabulky zadat.

Test nezávislosti kategoriálních znaků

Jak vidno, sestavení kontingenční tabulky není složité, v podstatě tak zpřehledňujeme výskyt všech kombinací kategorií dvou znaků mezi *N* subjekty. Obdobně jednoduché je i statistické hodnocení kontingenční tabulky. Nejčastěji testovanou **hypotézou je nezávislost výskytu sledovaných znaků**. K hodnocení se zde

Příklad A	Komplikace onemocnění		
	I	II	III
muž	16	6	22
žena	3	24	27

Příklad B	Komplikace onemocnění		
	I	II	III
muž	7	15	22
žena	12	15	27

1. Analyzujeme vztah mezi výskytem komplikací daného primárního onemocnění a pohlavím pacienta. Nulovou hypotézou je nezávislost výskytu obou znaků.

2. Test dobré shody pro kontingenční tabulku $R \times C$ vypočteme podle vztahů obecně popsaných v příkladě 2. Počet stupňů volnosti pro tabulku 2 × 3 je roven 2. Po dosazení do vztahu je vypočtena hodnota testové statistiky pro příklad A a B:

$$\mathrm{A}: \chi^2_{\nu=2} = 18.3$$

$$\mathrm{B}: \chi^2_{\nu=2} = 0.8$$

3. Vypočtená hodnota testové statistiky je porovnána s kritickou hodnotou testu pro $\alpha = 0{,}05$ $\chi^2_{(0{,}95;\ \nu=2)} = 5{,}99$.

4. Výsledné hodnoty statistické významnosti jsou:
Příklad A: $p < 0{,}001$ – zamítáme nulovou hypotézu o nezávislosti znaků
Příklad B: $p = 0{,}665$ – nezamítáme nulovou hypotézu o nezávislosti znaků

Příklad 4. Provedení testu nezávislosti dvou znaků v *R* × *C* kontingenční tabulce.

standardně používá nám již známý test dobré shody s testovou statistikou, která má χ^2 rozdělení. Výpočet stručně shrneme v následujících bodech, blíže jej přibližují příklady 1–4:

- K pozorovaným četnostem v tabulce musíme vypočítat četnosti očekávané, které jsou kalkulovány pro teoretickou situaci naprosté nezávislosti výskytu sledovaných znaků.
- Testovou statistiku χ^2 počítáme jako součet vážených čtverců rozdílů pozorovaných a očekávaných četností přes všechna políčka kontingenční tabulky. Jde tedy o klasický výpočet testu dobré shody, pouze počet stupňů volnosti je jiný než při porovnávání očekávaných a pozorovaných četností u kategorií jednoho znaku (díl XX seriálu). Test nezávislosti dvou znaků v kontingenční tabulce má počet stupňů volnosti roven μ = (počet řádků – 1) × (počet sloupců – 1).
- Vypočítanou hodnotu testové statistiky srovnáváme s hodnotou kvantilu rozdělení χ^2, a pokud tuto hranici odpovídající zvolené hladině chyby 1. druhu (α) překročí, pak zamítáme nulovou hypotézu o nezávislosti sledovaných znaků. Zamítnutím takto postavené hypotézy prokazujeme závislost, tedy existující vztah (vazbu, asociaci) ve výskytu kategorií sledovaných znaků. Naopak nezamítnutí nulové hypotézy znamená, že oba znaky jsou ve svém výskytu nezávislé a hodnota jednoho znaku neovlivňuje podmíněné rozdělení znaku druhého a naopak.

Test dobré shody pro hodnocení nezávislosti dvou znaků je neparametrický, nicméně i on má jistá omezení daná především výrazně heterogenním nebo nízkým výskytem kategorií v kontingenční tabulce. Pro takovou situaci lze nabídnout dva alternativní postupy výpočtu:

- **Yatesova korekce** je doporučena, pokud v jakémkoli poli kontingenční tabulky klesne pozorovaná četnost pod 5. Za této situace je klasický výpočet χ^2 testu nevhodný. Yatesova korekce je využitelná pro tabulky četností 2 × 2, kde má výsledná statistika χ^2 jeden stupeň volnosti. Ze vztahu pro výpočet korigovaného testu (viz příklad 1) je patrné, že oproti nekorigovanému výpočtu statistiky χ^2 odečítáme v čitateli hodnotu N/2, a tím číselnou hodnotu statistiky χ^2 snižujeme. Budeme tedy hůře zamítat nulovou hypotézu o nezávislosti znaků (říkáme, že test je více konzervativní), neboť nižší hodnota testové statistiky pochopitelně méně pravděpodobně překročí kritickou mez.
- **Fisherův exaktní test** je permutační neparametrický test vyvinutý k hodnocení nezávislosti dvou binárních znaků; tedy dat, která lze vložit do tabulky četností 2 × 2. Výhodou je použitelnost i pro velmi malé náhodné výběry. Test byl podrobně vysvětlen ve XIV. díle seriálu.

Korekce testu dobré shody při malé velikosti vzorku jsou v literatuře doporučovány, ale i zatracovány. Snížením hodnoty χ^2 snižujeme pravděpodobnost chyby 1. druhu (tedy zamítnutí platné hypotézy), ale na druhou stranu klesá naše šance rozpoznat skutečně neplatnou hypotézu (roste pravděpodobnost chyby 2. druhu). Nadto publikovaná doporučení, kdy přistoupit ke korekci, nejsou zcela jednoznačná. Těmto diskuzím se lze vyhnout následovně:

- plánovitým výběrem hodnocených subjektů o dostatečném *N,*
- při malém *N* v některých polích tabulky slučováním řádků nebo sloupců tabulky (vzácné kategorie mohou být spojeny); vždy je však možno sloučit pouze kategorie, kde to dovoluje logika nebo podstata problému,
- pokud není možné dosáhnout dostatečného *N,* pak je vysoce doporučenou alternativou Fisherův exaktní test, který jsme detailně rozebrali v díle XIV našeho seriálu.

Další metodou provedení testu nezávislosti kategoriálních znaků je tzv. *G* test. *G* test lze použít pro výpočet testu dobré shody, kdykoli platí pro jakoukoli buňku

	Y_1	Y_2	Y_3	Y_4
X_1	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$
X_2	$f_{2,1}$	..	..	..
X_3	$f_{3,1}$	..	..	$f_{i,j}$

1. Je dána kontingenční tabulka dvou proměnných *X* a *Y*, každá z nich nabývá více možných hodnot (kategorií) $X_{1..R}$ a $Y_{1..C}$.
2. Kombinace hodnot proměnných jsou označeny jako $f_{i,j}$ kde *f* je pozorovaná četnost kombinace a *i*, *j* jsou indexy pro daný řádek a sloupec tabulky.
3. Cílem je testovat nezávislost proměnných *X* a *Y* pomocí *G*-testu, který je alternativou ke klasickému χ^2 testu dobré shody.
4. Testová statistika *G*-testu je pro kontingenční tabulku *R* × *C* dána vztahem

$$G = 2 * \sum_{i=1}^{R} \sum_{j=1}^{C} f_{i,j} * \ln\left(\frac{f_{i,j}}{F_{i,j}}\right)$$

   kde $\boldsymbol{f_{i,j}}$ jsou pozorované četnosti v buňkách tabulky, $\boldsymbol{F_{i,j}}$ jsou očekávané hodnoty v buňkách tabulky; ve výpočtu je použit je přirozený logaritmus a součet je počítán pro nenulové buňky tabulky četností

5. Očekávané četnosti jsou vypočteny dle vztahu

$$F_{i,j} = \frac{\sum_{z=1}^{C} f_{i,z} * \sum_{z=1}^{R} f_{z,j}}{N}$$

6. Rozdělení testové statistiky *G* je přibližně shodné s rozdělením χ^2 pro obdobný χ^2 test dobré shody, tedy χ^2 rozdělení s ν = (počet řádků – 1) * (počet sloupců – 1) stupňů volnosti.
7. χ^2 test dobré shody a *G*-test vedou většinou k velmi obdobným výsledkům, nicméně *G*-test je oproti χ^2 testu preferován ve všech případech, kdy pro jakoukoli buňku tabulky platí, že $|\boldsymbol{f_{i,j} - F_{i,j}}| > \boldsymbol{F_{i,j}}$.

Příklad 5. Vzorový výpočet testu nezávislosti znaků pro *R* × *C* kontingenční tabulku: G-test.

Nominální data ⟶ Hodnocení asociace

	muž	žena
alergie ano	16	6
alergie ne	3	24

Koeficient	Hodnota	*p*
Koeficient *phi*	0,629	< 0,001
Cramérovo *V*	0,629	< 0,001
Koeficient kontingence *C*	0,532	< 0,001

Ordinální data ⟶ Hodnocení asociace

Anémie	Tíže onemocnění		
	I	II	III
bez anémie	16	6	8
lehká anémie	3	24	27
těžká anémie	3	17	40

Koeficient	Hodnota	*p*
Kendallův korelační koeficient τ	0,348	< 0,001
Goodman-Kruskalův koeficient (γ)	0,522	< 0,001

Příklad 6. Ukázka výpočtu dalších ukazatelů nezávislosti dvou nominálních a ordinálních znaků.

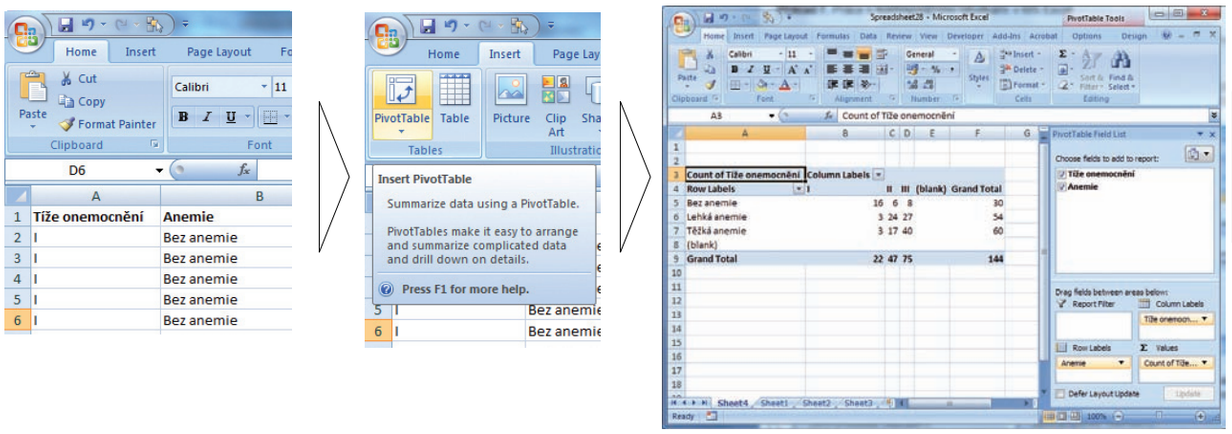
tabulky, že $|f_{i,j} - F_{i,j}| > F_{i,j}$ (viz též příklad 5). Opět tedy jde o situace, kdy je rozdělení pozorovaných četností v tabulce heterogenní a kdy je v některých polích nedostatečná četnost. *G* test ale není jen jakousi alternativou testu dobré shody, naopak sám χ^2 test je vlastně aproximací *G* testu. G test je založen na principu maximální věrohodnosti (*MLE, maximum likelihood estimation)*, což je obecná statistická metoda pro získávání odhadů. Výpočet touto metodou byl před nástu-

1. Kontingenční tabulky v MS Excel jsou nástrojem pro sumarizaci kategoriálních dat ve formě tabulek četností
2. Kontingenční tabulky v MS Excel neumožňují výpočet statistických testů

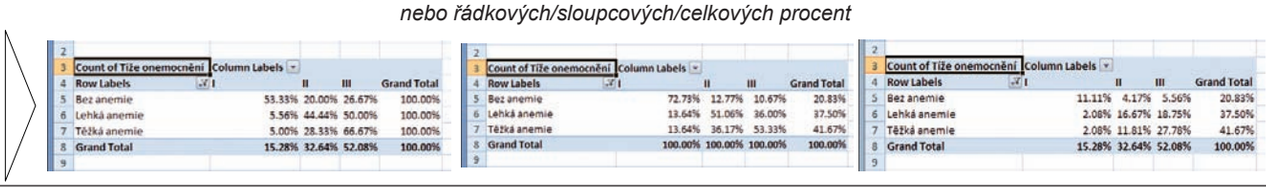
A) Vstupní data ve formě databázové tabulky

B) Vytvoření kontingenční tabulky v menu MS Excel

C) Definice kontingenční tabulky prokřížením primárních proměnných



D) Kontingenční tabulky umožňují zobrazit data z pohledu absolutních počtů nebo řádkových/sloupcových/celkových procent



Příklad 7. Práce s kontingenčními tabulkami v MS Excel.

Tab. 1. Hodnocení asociace dvou nominálních znaků – ukázky výpočtu dalších ukazatelů.

Φ (phi koeficient)

zdrojová data	kontingenční tabulka 2 × 2
výpočet – vztah	$C = \sqrt{\frac{\chi^2}{N}}$, kde χ^2 je hodnota χ^2 testové statistiky z klasického testu dobré shody, N je celkový počet hodnot
význam – komentář	Jde o míru vztahu dvou binárních proměnných interpretačně obdobnou Pearsonovu korelačnímu koeficientu pro spojitá data. Maximální hodnota rozsahu phi koeficientu je dána poměrem četností kategorií analyzovaných proměnných, v případě stejného zastoupení kategorií u obou proměnných je maximum 1, v jiných případech nižší.

Cramérovo *V*

zdrojová data	kontingenční tabulka $R \times C$
výpočet – vztah	$V = \sqrt{\frac{\chi^2}{N(k-1)}}$, kde χ^2 je hodnota testové statistiky z klasického testu dobré shody, N je celkový počet hodnot a k je menší hodnota z počtu řádků nebo počtu sloupců tabulky
význam – komentář	Míra asociace mezi dvěma nominálními proměnnými, hodnota 0 znamená nulovou asociaci proměnných, čím blíže je hodnotě 1, tím silnější je jejich asociace.

Koeficient kontingence *C* dle Pearsona

zdrojová data	kontingenční tabulka $R \times C$ – symetrická (stejný počet řádků a sloupců)
výpočet – vztah	$C = \sqrt{\frac{\chi^2}{N+\chi^2}}$, kde χ^2 je hodnota testové statistiky z klasického testu dobré shody, N je celkový počet hodnot
význam – komentář	Míra asociace mezi dvěma nominálními proměnnými, hodnota 0 znamená nulovou asociaci proměnných, čím vyšší hodnota, tím silnější je jejich asociace. Maximální hodnota se liší podle velikosti tabulky, pro 2 × 2 tabulky je to hodnota 0,707; s velikostí tabulky roste.

Tab. 2. Hodnocení asociace dvou ordinálních znaků – ukázky výpočtu dalších ukazatelů.

Kendallův korelační koeficient τ

zdrojová data	kontingenční tabulka $R \times C$ ordinálních proměnných
výpočet – vztah	$\tau_A = \frac{n_s - n_d}{0{,}5n\,(n-1)}$, kde n_s je počet souhlasných dvojic subjektů (se stejným pořadím hodnot v obou znacích), n_d je počet nesouhlasných dvojic subjektů (s rozdílným pořadím hodnot v obou znacích) a n celkový počet dvojic subjektů.
význam – komentář	míra asociace dvou ordinálních proměnných s rozsahem –1 (negativní asociace) až +1 (pozitivní asociace)

Goodmanův-Kruskalův koeficient (γ)

zdrojová data	kontingenční tabulka $R \times C$ ordinálních proměnných
výpočet – vztah	$\tau_A = \frac{n_s - n_d}{n_s - n_d}$, kde n_s je počet souhlasných dvojic subjektů (se stejným pořadím hodnot v obou znacích) a n_d je počet nesouhlasných dvojic subjektů (s rozdílným pořadím hodnot v obou znacích).
význam – komentář	míra asociace dvou ordinálních proměnných s rozsahem –1 (negativní asociace) až +1 (pozitivní asociace)

pem výkonných osobních počítačů velmi náročný, a proto Karl Pearson odvodil na počátku minulého století aproximaci v podobě testu dobré shody. Nicméně dnes již aplikaci obecnějšího výpočtu nic nebrání, a v mnoha statistických programech tak naleznete vedle „obyčejného" testu χ^2 také *ML* χ^2 (tedy *maximum likelihood* χ^2).

Stručný přehled dalších ukazatelů asociace kategoriálních znaků

Kromě testu nezávislosti založeného na χ^2 testu se v literatuře často objevují i jiné ukazatele vztahu dvou kategoriálních znaků. Jde o různé koeficienty, které aspirují i na kvantifikaci síly vztahu. Avšak ne vždy bývají tyto ukazatele přehledně vysvětleny a uživatelé si pak relativně často pletou jejich význam a někdy i názvy, které jsou dost podobné. Z tohoto důvodu jsme připravili komentovaný přehled nejvýznamnějších ukazatelů v tab. 1 a 2, jednoduchou ukázku výpočtu obsahuje příklad 6.

Je patrné, že hlavní přidaná hodnota těchto koeficientů je v měření síly vztahu, kterou vyjadřuje již sama hodnota příslušného koeficientu. Na rozdíl od hodnoty testu χ^2 nabývají tyto ukazatele hodnot v známém rozsahu, například od 0 (žádný vztah) do 1 (maximální vztah). Známý rozsah možných hodnot koeficientu usnadňuje posouzení míry vztahu dvou znaků nebo srovnání výsledků analýzy v různých náhodných výběrech.

Zvláštní situaci představují kontingenční tabulky, které dávají do vztahu dvě ordinální proměnné. Zde již můžeme uvažovat i o jistém trendu v jejich případném vztahu, neboť ordinalita znamená, že kategorie znaků umíme seřadit od nejnižší po nejvyšší. Kromě obecného průkazu závislosti, který nabízí test dobré shody, zde můžeme sledovat, zda jde o závislost kladnou (oba znaky spolu v kategoriích rostou), nebo naopak zápornou. Tab. 2 krátce shrnuje dva tzv. neparametrické koeficienty korelace, které umí tento trend podchytit; škála jejich hodnot je od –1 do +1.

Průkaz a kvantifikace závislosti ordinálních znaků jsou velmi významné i pro biomedicínské obory. Mnoho klinických a diagnostických znaků je vyjadřováno na ordinální škále (stupeň toxicity léčby, škály dosahované léčebné odpovědi, hodnocení semikvantitativních metod tzv. křížkováním apod.). Zkoumání vzájemného vztahu těchto znaků je tedy velmi důležité a budeme mu věnovat samostatně jeden z následujících dílů seriálu.

Kontingenční tabulky jako nástroj zviditelnění a konzolidace dat

Ačkoli jsou kontingenční tabulky spjaty se statistikou, celá řada oborů s nimi pracuje, aniž je využívá pro statistické testování. V ekonomice nebo v manažerském rozhodování jsou kontingenční tabulky využívány čistě jako nástroj pro zviditelnění, popis a zpřehlednění dat. Pojem kontingenční tabulka tak nalezete v manuálu mnoha softwarových produktů zaměřených na práci s daty, ale často bez výše popsané statistické nadstavby. Například nástroje pro prohlížení dat nad datovými sklady nabízejí kontingenční tabulky pro interaktivní křížení parametrů. Numerické výstupy pak pracují s procentickými přepočty původní tabulky (řádková, sloupcová nebo celková procenta) anebo graficky znázorňují marginální součty řádků a sloupců. Příklad 7 stručně ukazuje práci s kontingenční tabulkou v MS Excel. Chceme-li ale provádět testy nezávislosti znaků a jiná statistická hodnocení, musíme sáhnout po statistickém programu (např. Statistica for Windows, SPSS, R…).

Závěrem: není test jako test

Doufáme, že jsme relativně jednoduché téma příliš nezkomplikovali. Základ hodnocení nezávislosti kategoriálních jevů je a zůstává stejný již více než sto let: sestavíme kontingenční tabulku a použijeme test dobré shody s adekvátním počtem stupňů volnosti. Avšak způsobů výpočtu a využitelných ukazatelů je mnohem více a v dnešním světě se i početně náročné postupy stávají lehce dostupnými laickému uživateli. Statistické programy nabízejí pro analýzu kontingenčních tabulek často několik ukazatelů současně, a pokud se v nich uživatel neorientuje, je vlastně obětí redundantní nabídky. Jistě jste i vy již někdy slyšeli větu: „Nabídlo mi to na stejná data několik výpočtů a pokaždé to vyšlo jinak, tak jaká je v tom věda?" Snad je po pročtení našeho textu jasné, jaký výpočet bychom si měli v jaké situaci vybrat. To je zcela zásadní, protože zamítnutá hypotéza nezávislosti dvou znaků znamená jejich faktickou závislost nebo vazbu mezi nimi. Odtud je již jen krůček k závažným interpretacím o příčině a následku. Jistě nikdo z nás nechce publikovat tak vážné závěry na základě špatných výpočtů.

Literatura

Plackett RL. Karl Pearson and the Chi-Squared Test. International Statistical Review 1983; 51(1): 59–72.
Sokal RR, Rohlf FJ. Biometry: the Principles and Practice of Statistics in Biological Research. 3rd ed. New York: Freeman 1994.