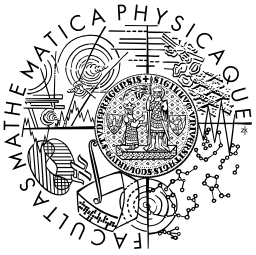


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Lukáš Pavlík

Testy nezávislosti v kontingenční tabulce

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Matúš Maciak, Ph.D.
Studijní program: Finanční matematika
Studijní obor: Finanční matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Zde bych chtěl poděkovat především RNDr. Matúši Maciakovi, Ph.D. za cenné rady, připomínky a vstřícnost a také za čas, který věnoval průběžnému opravování a konzultacím při vedení této práce. Také děkuji rodině a přátelům za poskytnuté zázemí a podporu během vypracování.

Název práce: Testy nezávislosti v kontingenční tabulce

Autor: Lukáš Pavlík

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Matúš Maciak, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této bakalářské práci se zabýváme různými metodami testování nezávislosti v dvourozměrných kontingenčních tabulkách. Metody vysvětlujeme, diskutujeme jejich výhody a nedostatky a ilustrujeme na příkladu. Dále je porovnáváme na simulovaných datech prostřednictvím statistického softwaru R. Na základě výsledků simulací se snažíme rozhodnout, jaký test je pro danou situaci nejlepší. Zvláštní pozornost věnujeme nové metodě, USP testu, který je založen na tzv. U-statistikách, s nimiž čtenáře seznámíme. Ukážeme, že USP test v určitých případech poskytuje výrazné zlepšení výsledků, v jiných se mu naopak významně nedaří. Tyto případy specifikujeme a učiníme závěr, kdy je výhodné test použít a kdy nikoli.

Klíčová slova: Kontingenční tabulky, Testy nezávislosti, Pearsonův χ^2 test, Permutační testy, U-statistiky

Title: Tests of independence in contingency tables

Author: Lukáš Pavlík

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Matúš Maciak, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this thesis we investigate various methods for testing independence in two-way contingency tables. The methods are explained, their advantages and drawbacks are discussed, and we also illustrate the methods on an example. Further, we compare the tests on simulated data using R statistic programming language. Based on simulation results we try to decide which test is the best choice for a situation. In particular, we investigate a new method, USP test, which is based on the theory of so called U-statistics. We therefore describe these, too. It is shown that USP test performs much better than other tests in particular cases, but fails in some others. These cases are specified and guidelines are made about when the test is advantageous to use and when it is not.

Keywords: Contingency tables, Independence tests, Pearson's χ^2 Test, Permutation tests, U-statistics

Obsah

Úvod

Data jsou v současné době velmi cenné zboží. Ještě cennější jsou však metody jejich analýzy, tedy proces sběru, transformace, získávání a vyhodnocení informací obsažených v datech a jejich vizualizace. Analýza dat nám umožňuje odhalit v datech vzorce, vztahy a trendy. Tato činnost má zásadní význam pro optimální rozhodování o budoucnosti, protože poskytuje podklady pro informovaná rozhodnutí na základě faktů a důkazů. Analyzováním dat získáváme přehled o současné situaci, schopnost identifikovat klíčové problémy a objevujeme příležitosti pro vylepšení procesů, produktů nebo služeb. Díky analýze dat mohou firmy, organizace i jednotlivci lépe porozumět svým zákazníkům, efektivněji řídit své aktivity a dosahovat svých cílů.

V této práci se budeme zabývat analýzou kategoriálních dat, která lze reprezentovat dvourozměrnou kontingenční tabulkou. Kategoriální data jsou taková data, která většinou nemají žádnou přirozenou číselnou interpretaci, neboť vyjadřují příslušnost statistické jednotky do určité skupiny (kategorie) nebo přítomnost nějakého znaku. Proto se jejich analýza zaměřuje především na pravděpodobnosti jednotlivých kategorií. Kategoriální data se dále dělí podle typu stupnice kategorií na ordinální a nominální. Ordinální data se dají přirozeně uspořádat podle velikosti, příkladem může být nejvyšší dosažené vzdělání. Nominální data jsou naopak neporovnatelná, nelze je seřadit. Jako příklad uvedeme pohlaví nebo krevní skupinu.

V rámci analýzy kategoriálních dat nás bude konkrétně zajímat nezávislost dvou veličin. Například můžeme chtít zkoumat na základě souboru určitého počtu jedinců, zda dosažené vzdělání a pohlaví na sobě závisí nebo jestli nezaměstnanost a kraj, ve kterém jedinec bydlí, spolu nějak souvisí. Čemu se naopak věnovat nebudeme, a tedy o ní nebudeme činit závěry, je kauzalita, tedy příčinná souvislost mezi veličinami (tj. zda jedna veličina může být příčinou druhé, která je následkem).

K formálnímu rozhodnutí o nezávislosti dvou veličin slouží statistické testy. Stručně zde připomeneme základní charakteristiky statistického testu. Hladina testu je předem stanovené číslo $\alpha \in (0, 1)$ vyjadřující pravděpodobnost zamítnutí platné hypotézy. Síla testu proti dané alternativě je pravděpodobnost zamítnutí neplatné hypotézy při dané konkrétní alternativě. Pravděpodobnost napozorování náhodného výběru za platnosti hypotézy, který je ve stejném nebo větším rozporu s hypotézou než realizovaný náhodný výběr, vyjadřuje p-hodnota. Chyba I. druhu znamená zamítnutí platné hypotézy, její pravděpodobnost je u správně fungujícího testu omezena hladinou. Chyba II. druhu vyjadřuje nezamítnutí neplatné hypotézy při dané konkrétní alternativě, její pravděpodobnost není pod kontrolou.

Na následujících řádcích si shrneme a ilustrujeme na příkladu nejpoužívanější testy nezávislosti a představíme jeden poměrně nový. Testy porovnáme jak teoreticky, tak empiricky pomocí simulační studie ve statistickém softwaru R (R Core Team, 2022). Na základě simulací rozhodneme o praktickém použití testů v různých situacích.

1. Kontingenční tabulky

Začneme tím, že se seznámíme se základními pojmy a notací. Uvažujme náhodný vektor $(X, Y)^\top$ s diskrétním rozdělením, jehož složky jsou dvě kategoriální veličiny $X \in \{x_1, \ldots, x_I\}$ a $Y \in \{y_1, \ldots, y_J\}$, kde x_i pro $i = 1, \ldots, I$ je i-tá kategorie náhodné veličiny X a y_j pro $j = 1, \ldots, J$ je j-tá kategorie náhodné veličiny Y. Pro jednodušší zápis budeme někdy používat pro označení kategorií pouze i a j namísto x_i a y_j. Předpokládejme, že se uskutečnil náhodný výběr $(X_1, Y_1)^\top, \ldots, (X_n, Y_n)^\top$ z rozdělení náhodného vektoru $(X, Y)^\top$ o daném rozsahu $n \in \mathbb{N}$. Tento výběr zobrazíme v podobě dvourozměrné kontingenční tabulky o I řádcích a J sloupcích (viz Tabulka 1.1), kde náhodná veličina n_{ij} udává pozorovanou četnost vektorů rovných $(x_i, y_j)^\top$, tedy počet pozorování spadajících do i-té kategorie náhodné veličiny X a do j-té kategorie náhodné veličiny Y. Dále n_{i+} označuje počet pozorování spadajících do i-té kategorie veličiny X (i-tého řádku) a n_{+j} počet pozorování v j-té kategorii veličiny Y (j-tém sloupci). Tyto veličiny se nazývají marginální četnosti a splňují

$$n_{i+} = \sum_{j=1}^{J} n_{ij}, \quad n_{+j} = \sum_{i=1}^{I} n_{ij}.$$

	$Y = y_1$	$\ldots$	$Y = y_J$	$\sum$
$X = x_1$	n_{11}	$\ldots$	n_{1J}	n_{1+}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$X = x_I$	n_{I1}	$\ldots$	n_{IJ}	n_{I+}
$\sum$	n_{+1}	$\ldots$	n_{+J}	n

Tabulka 1.1: Kontingenční tabulka $I \times J$.

Zajímají nás také pravděpodobnosti jednotlivých kategorií, které v praxi však neznáme (viz Tabulka 1.2). Sdružené rozdělení náhodného vektoru $(X, Y)^\top$ je multinomické a je určeno pravděpodobnostmi $p_{ij} = \mathsf{P}[X = x_i, Y = y_j]$. Marginální rozdělení náhodných veličin X, resp. Y jsou také multinomická a určují je pravděpodobnosti $p_{i+} = \mathsf{P}[X = x_i]$, resp. $p_{+j} = \mathsf{P}[Y = y_j]$.

	$Y = y_1$	$\ldots$	$Y = y_J$	$\sum$
$X = x_1$	p_{11}	$\ldots$	p_{1J}	p_{1+}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$X = x_I$	p_{I1}	$\ldots$	p_{IJ}	p_{I+}
$\sum$	p_{+1}	$\ldots$	p_{+J}	1

Tabulka 1.2: Pravděpodobnostní rozdělení pro kontingenční tabulku $I \times J$.

Zřejmě tedy platí následující vztahy:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathsf{Mult}_{I\,J}\left(1, (p_{11}, p_{12}, \ldots, p_{IJ})^\top\right),$$

$$X \sim \mathsf{Mult}_I\left(1, (p_{1+}, \ldots, p_{I+})^\top\right), \quad Y \sim \mathsf{Mult}_J\left(1, (p_{+1}, \ldots, p_{+J})^\top\right).$$

V praxi kontingenční tabulka vzniká tak, že pozorujeme na statistických jednotkách dva znaky (veličiny) a určíme četnosti kombinací jejich diskrétních hodnot. Samozřejmě nezáleží na tom, kterou veličinu zapíšeme do řádků a kterou do sloupců tabulky. Pro nominální veličinu nezáleží ani na uspořádání jejích kategorií v rámci řádků (sloupců), zatímco u ordinální veličiny je nutné zachovat řazení kategorií (sestupně nebo vzestupně) v rámci uspořádání v tabulce, pokud chceme použít metody statistické analýzy určené pro ordinální veličiny. Metody určené pro analýzu nominálních veličin lze samozřejmě použít i na ordinální veličiny, pak ale typicky mají menší sílu než metody určené specificky pro ordinální veličiny, protože nevyužívají informaci o uspořádání. Ordinální data lze zkonstruovat i z dat kvantitativních (spojitých) tak, že je převedeme na diskrétní data vytvořením vhodně zvolených kategorií, obvykle rozdělením množiny přípustných hodnot na podintervaly. Například v případě zkoumání hmotnosti lidí bychom mohli vytvořit intervaly s délkou odpovídající deseti kilogramům s tím, že první a poslední interval by byly zvoleny delší tak, aby do nich padl dostatečný počet pozorování.

1.1 Testování nezávislosti

Často nás při analýze kontingenčních tabulek zajímá otázka nezávislosti dvou marginálních veličin, tedy zda četnosti v jednotlivých kategoriích veličiny X závisí na hodnotě veličiny Y, či nikoliv. Je důležité poznamenat, že v případě nalezení závislosti obecně nemůžeme bez dalších znalostí vyvozovat závěry o kauzalitě.

Naším cílem je tedy vyšetřit, zda jsou veličiny X a Y nezávislé. Budeme testovat hypotézu

$$H_0 : X \text{ a } Y \text{ jsou nezávislé,}$$

což je, matematicky vyjádřeno, ekvivalentní hypotéze

$$H_0' : p_{ij} = p_{i+}\, p_{+j} \quad \forall\, i \in \{1, \dots, I\}, \forall\, j \in \{1, \dots, J\}, \tag{1.1}$$

proti alternativě

$$H_1 : X \text{ a } Y \text{ nejsou nezávislé,}$$

což je ekvivalentní s

$$H_1' : \exists\, i \in \{1, \dots, I\}\ \exists\, j \in \{1, \dots, J\} :\ p_{ij} \neq p_{i+}\, p_{+j}. \tag{1.2}$$

Definujeme si ještě některé pojmy, které budeme dále používat. Nechť $\widehat{p}_{i+}$ a $\widehat{p}_{+j}$ značí odhady marginálních pravděpodobností metodou maximální věrohodnosti na základě náhodného výběru $(X_1, Y_1)^\top, \dots, (X_n, Y_n)^\top$. Víme, že pro ně platí

$$\widehat{p}_{i+} = \frac{n_{i+}}{n}, \quad \widehat{p}_{+j} = \frac{n_{+j}}{n}.$$

Dále označme

$$e_{ij} = n\, \widehat{p}_{i+}\, \widehat{p}_{+j} = \frac{n_{i+}\, n_{+j}}{n} \tag{1.3}$$

odhadnuté očekávané četnosti pozorování v (i, j)-té kategorii (buňce) *za platnosti hypotézy* H_0.

Existuje mnoho různých metod pro testování hypotézy H_0 proti alternativě H_1, některé mají široké uplatnění, jiné užší. My v této práci budeme porovnávat testy nezávislosti použitelné ve zcela obecném případě, tj. pro nominální veličiny (tedy i ordinální) s libovolnými počty kategorií $I \geq 2$, $J \geq 2$. Specifickým statistickým metodám určeným pouze pro tabulky s ordinálními veličinami nebo výlučně pro čtyřpolní tabulky (tj. $I = 2$, $J = 2$) se věnovat nebudeme.

2. Testy nezávislosti

V dalším textu představíme různé přístupy a metody pro testování nezávislosti. Budeme vycházet především z knih Agresti (2007), Anděl (2011) a článku Berrett a Samworth (2021). Další zdroje uvedeme u konkrétních pasáží. Nejdříve popíšeme Pearsonův χ^2 test a G-test, což jsou asymptotické testy, které využívají známé limitní rozdělení svých testových statistik. Následně přejdeme k tzv. permutačním testům, jejichž speciálními případy jsou Fisherův přesný (nebo přibližný) test a USP test, které také vysvětlíme. Každý z testů pro ilustraci aplikujeme na simulovaný příklad, jenž si nyní zformulujeme.

Příklad 1. *U 1000 náhodně vybraných obyvatel ČR jsme zjistili, v jakém typu sídla (podle velikosti) bydlí a jaký je jejich preferovaný druh dovolené. Sídlo představuje kategoriální veličinu se třemi kategoriemi: velkoměsto (nad 100 000 obyvatel), město (alespoň 3000 obyvatel) a obec (méně než 3000 obyvatel). Veličina dovolená zahrnuje čtyři kategorie: rekreační (pobytová), poznávací, aktivní (sportovní nebo turistická) a jiná dovolená (například trávená doma aj.). Pozorované (resp. nasimulované) četnosti jsme zapsali do kontingenční tabulky (Tabulka 2.1). Jsou vygenerované z pravděpodobnostního rozdělení, které mírně porušuje hypotézu nezávislosti* (1.1)*, přesný popis použitého scénáře uvedeme v Kapitole 4. Zajímá nás, zda preferovaný druh dovolené nějak závisí na typu sídla jedince.*

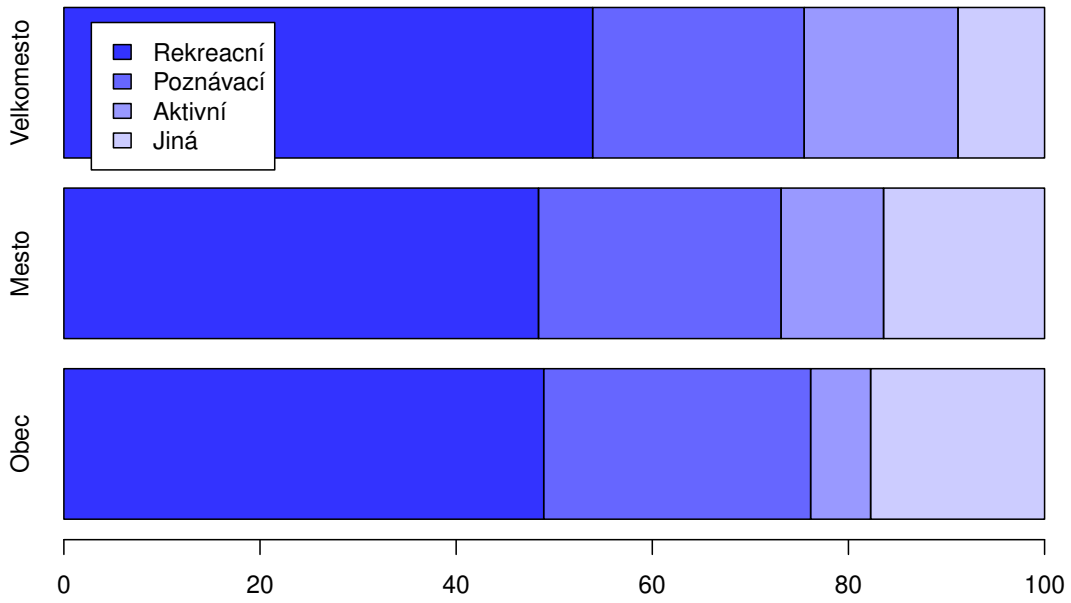
	Rekreační	Poznávací	Aktivní	Jiná
Velkoměsto	110	44	32	18
Město	227	116	49	77
Obec	160	89	20	58

Tabulka 2.1: Kontingenční tabulka shrnující absolutní četnosti preferovaného druhu dovolené v různých typech sídel.

V rámci explorační analýzy spočítáme řádkové relativní četnosti preferovaného druhu dovolené v daném sídle (viz Tabulka 2.2). Vidíme, že mezi sídly jsou ve veličině dovolená určité odlišnosti. Například, zatímco asi 16 % obyvatel velkoměst preferuje aktivní dovolenou, tak u lidí z obcí je to pouze 6 %. Podobné větší či menší rozdíly lze pozorovat i pro ostatní kategorie dovolené. Vizualizovat si je můžeme pomocí sloupcového diagramu (viz Obrázek 2.1). Zda jsou tyto rozdíly statisticky signifikantní a nikoli pouze dílem náhody, rozhodneme provedením vhodně zvoleného statistického testu.

	Rekreační	Poznávací	Aktivní	Jiná
Velkoměsto	53,92	21,57	15,69	8,82
Město	48,40	24,73	10,45	16,42
Obec	48,93	27,22	6,11	17,74

Tabulka 2.2: Tabulka řádkových relativních četností (vyjádřených v procentech) druhů dovolené pro jednotlivé typy sídel.



Obrázek 2.1: Sloupcový graf s relativními četnostmi druhů dovolené pro dané sídlo (na svislé ose) v procentech (na vodorovné ose).

Nyní přejdeme k popisu různých testů nezávislosti, které na tento příklad můžeme aplikovat. Zajímají nás výsledky testu pro stanovenou hladinu významnosti $\alpha = 0{,}05$. Žádoucí výsledek je zamítnutí hypotézy nezávislosti, která pro uvažované rozdělení neplatí.

2.1 Pearsonův χ^2 test nezávislosti

Velmi často používaným a asi nejznámějším testem k testování nezávislosti je Pearsonův χ^2 test nezávislosti (viz např. Agresti, 2007, str. 34). Navrhl jej již v roce 1900 britský statistik Karl Pearson. Jeho testová statistika

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \tag{2.1}$$

kde n_{ij} jsou pozorované četnosti a e_{ij} jsou odhadnuté očekávané četnosti definované v (1.3), zřejmě nabývá pouze nezáporných hodnot, neboť je součtem nezáporných „příspěvků“. Testová statistika (2.1) má za platnosti nulové hypotézy (1.1) asymptoticky χ^2-rozdělení s $(I-1)(J-1)$ stupni volnosti (dále budeme užívat značení $\chi^2_{(I-1)(J-1)}$). Počet stupňů volnosti se získá odečtením počtu odhadnutých parametrů metodou maximální věrohodnosti od počtu všech testovaných parametrů. Podrobněji to popíšeme v rámci odvození G-testu v následující sekci.

Pearsonův χ^2 test vlastně přímo vychází z χ^2 testu dobré shody pro multinomické rozdělení s odhadnutými parametry. Za platnosti hypotézy bude pozorovaná hodnota testové statistiky (2.1) malá, protože hodnoty n_{ij} budou blízké hodnotám e_{ij}. Naopak velké hodnoty testové statistiky nasvědčují platnosti alternativy (1.2), tedy že existuje souvislost mezi dvěma veličinami.

Ze známého limitního rozdělení testové statistiky můžeme ihned zkonstruovat kritický obor, tedy pro předem stanovenou hladinu testu $\alpha \in (0, 1)$ hypotézu

nezávislosti zamítáme pro hodnoty $\chi^2 \geq \chi^2_{(I-1)(J-1)}(1-\alpha)$, kde pravá strana nerovnosti značí $(1-\alpha)$-kvantil příslušného rozdělení.

Vzhledem k jednostrannému kritickému oboru dostaneme vyjádření p-hodnoty jako $p = 1 - F_\infty(\chi^2)$, kde F_∞ značí distribuční funkci limitního rozdělení.

Provedeme-li Pearsonův χ^2 test na kontingenční tabulce z Příkladu 1, dostaneme p-hodnotu 0,0016. Tedy na hladině významnosti 0,05 s velkou rezervou zamítáme nezávislost preferovaného druhu dovolené a typu sídla. Prokázali jsme tak, že mezi veličinami je nějaká závislost, ale nevíme jaká. Existují různé způsoby, jak lze tuto závislost kvantifikovat (např. Cramerovo V nebo Goodman a Kruskalova lambda), které ale nejsou předmětem této práce. My se spokojíme s interpretací pomocí provedené explorační analýzy. Z Tabulky 2.2 vidíme, že ve velkoměstech je častější zájem o rekreační nebo aktivní dovolenou, zatímco ve městech a obcích lidé častěji preferují poznávací nebo jinou dovolenou.

2.2 G-test nezávislosti

Dalším známým testem pro testování hypotézy nezávislosti je takzvaný G-test, který vznikl aplikací testu založeného na věrohodnostním poměru na multinomické rozdělení (odvození viz dále). Jeho testová statistika

$$G = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log \frac{n_{ij}}{e_{ij}}, \tag{2.2}$$

kde log značí přirozený logaritmus, má za platnosti nulové hypotézy (1.1) asymptotické rozdělení $\chi^2_{(I-1)(J-1)}$, tedy stejné jako testová statistika (2.1). Za platnosti nulové hypotézy bude hodnota G blízká nule, protože podíly pozorovaných četností n_{ij} a odhadnutých očekávaných četností e_{ij} budou blízké 1. Naopak velké hodnoty testové statistiky G nasvědčují platnosti alternativy (1.2), tedy že veličiny nejsou nezávislé.

Hypotézu nezávislosti (1.1) zamítáme pro hodnoty $G \geq \chi^2_{(I-1)(J-1)}(1-\alpha)$. Pro p-hodnotu platí $p = 1 - F_\infty(G)$, kde F_∞ značí distribuční funkci limitního rozdělení.

Testovou statistiku zde odvodíme, vycházíme z knihy Anděl (2011, příklad 8.23). Pro přehlednější výpočty si nejdříve zavedeme upravené značení. Nechť

$$\begin{aligned}\boldsymbol{U} &= (U_1, U_2, \ldots, U_K)^\top = (n_{11}, n_{12}, \ldots, n_{IJ})^\top,\\ \boldsymbol{p} &= (\pi_1, \pi_2, \ldots, \pi_K)^\top = (p_{11}, p_{12}, \ldots, p_{IJ})^\top,\end{aligned}$$

kde $\boldsymbol{U}$ je vektor pozorovaných četností splňující $\sum_{k=1}^{K} U_k = n$, $\boldsymbol{p} \in (0,1)^K$ je vektor pravděpodobností kategorií splňující $\sum_{k=1}^{K} \pi_k = 1$, $K = IJ$. Pak $\boldsymbol{U}$ má multinomické rozdělení $\mathsf{Mult}_K(n, \boldsymbol{p})$ a dá se vyjádřit jako součet n nezávislých stejně rozdělených náhodných vektorů $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$, kde

$$\boldsymbol{V}_i = (V_{i1}, \ldots, V_{iK})^\top,\ i \in \{1, \ldots, n\}$$

má multinomické rozdělení $\mathsf{Mult}_K(1, \boldsymbol{p})$, takových, že vektor $\boldsymbol{V}_i$ udává, do které kategorie padlo i-té pozorování. Pro jeho složky proto platí $V_{ik} = 1$ pro k rovné kategorii, do které i-té pozorování padlo, a $V_{ik} = 0$ jinak, $k \in \{1, \ldots, K\}$. Platí tedy $\boldsymbol{U} = \sum_{i=1}^{n} \boldsymbol{V}_i$. Testovaný parametr bude vektor $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{K-1})^\top$, což je vektor

$\boldsymbol{p}$ bez poslední složky π_K, kterou testovat nepotřebujeme, neboť je jednoznačně určena ostatními složkami vektoru pravděpodobností ($\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$). Je známo, že maximálně věrohodný odhad vektoru $\boldsymbol{\theta}$ je $\widehat{\boldsymbol{\theta}}_n = (U_1/n, \ldots, U_{K-1}/n)^\top$. Nejdříve použijeme test poměrem věrohodností na jednoduchou hypotézu $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, kde $\boldsymbol{\theta}_0 = (\pi_1^0, \ldots, \pi_{K-1}^0)^\top$ je předem zvolený vektor kladných čísel splňující $\sum_{k=1}^{K-1} \pi_k^0 < 1$. Hodnota pro poslední ($K$-tou) kategorii se pak dopočítá z $\pi_K^0 = 1 - \sum_{k=1}^{K-1} \pi_k^0$. Víme, že testová statistika testu věrohodnostním poměrem

$$2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}_0)),$$

kde $\ell_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta})$ je logaritmická věrohodnost ($L_n(\boldsymbol{\theta})$ značí věrohodnostní funkci), má za platnosti hypotézy asymptotické rozdělení χ^2_{K-1}. Máme

$$\begin{aligned} L_n(\boldsymbol{\theta}_0) &= \prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k^0)^{V_{ik}} = \prod_{k=1}^{K} (\pi_k^0)^{U_k}, \\ \ell_n(\boldsymbol{\theta}_0) &= \sum_{k=1}^{K} U_k \log \pi_k^0, \\ \ell_n(\widehat{\boldsymbol{\theta}}_n) &= \sum_{k=1}^{K} U_k \log \frac{U_k}{n}. \end{aligned}$$

Tudíž

$$2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}_0)) = 2 \sum_{k=1}^{K} U_k \log \frac{U_k}{n\pi_k^0}.$$

G-test ale testuje hypotézu nezávislosti, tedy hypotézu složenou. V tom případě je potřeba určit $(I-1) + (J-1)$ maximálně věrohodných odhadů parametrů $p_{i+}, i \in \{1, \ldots, I-1\}$ a $p_{+j}, j \in \{1, \ldots, J-1\}$. Vrátíme-li se k původnímu značení, dostaneme vyjádření (2.2) pro statistiku G, která má za platnosti hypotézy nezávislosti (1.1) asymptoticky rozdělení $\chi^2_{(I-1)(J-1)}$, jehož počet stupňů volnosti jsme získali odečtením jednoho stupně za každý odhadnutý parametr od původního počtu stupňů ($IJ - 1$).

Lze ukázat, že Pearsonův χ^2 test je aproximací G-testu. Oba testy dávají přibližně stejné výsledky, jsou asymptoticky ekvivalentní, avšak konvergence k limitnímu rozdělení je u G-testu o trochu rychlejší. Statistik John H. McDonald napsal: „*Vyberte si jeden z nich a držte se ho po zbytek svého života.*“ (viz McDonald, 2014, str. 60).

Pro Příklad 1 dává G-test nepřekvapivě velmi podobný výsledek jako Pearsonův χ^2 test. Vychází p-hodnota 0,0011. Tedy na hladině významnosti 0,05 opět s velkou rezervou zamítáme nezávislost dvou uvažovaných veličin.

Nedostatky χ^2 testu a G-testu

Ve zbylé části této sekce si popíšeme nedostatky χ^2 testu a G-testu. Zaprvé oba testy obecně nedodržují stanovenou hladinu (pravděpodobnost chyby I. druhu může výrazně překročit hladinu významnosti), což je většinou způsobeno tím, že se jedná o asymptotické testy. Skutečné rozdělení testové statistiky totiž může být v případech, kdy rozsah výběru $n \in \mathbb{N}$ není dostatečně velký v poměru k počtu buněk $I\,J$, dost odlišné od limitního, nebo dokonce i skutečné limitní rozdělení (pro $n \to \infty$) může být pro některé modely jiné než teoreticky odvozené rozdělení

$\chi^2_{(I-1)(J-1)}$, jak ukázali Berrett a Samworth (2021), . Zadruhé testové statistiky (2.1) a (2.2) nejsou definovány (pro dělení nulou) v případě, že v nějakém řádku nebo sloupci tabulky nejsou žádná pozorování. Třetím problémem je neznámá síla testů proti různým alternativám.

První dva nedostatky souvisí s malými četnostmi v buňkách či nedostatečným rozsahem výběru. Z toho důvodu se v odborné literatuře uvádí různá doporučení (tzv. „rules of thumb"), kdy je vhodné testy použít, která však nejsou jednotná. Agresti tvrdí, že pro uspokojivé výsledky je dostatečné, když všechny odhadnuté očekávané četnosti e_{ij} jsou rovny alespoň 5 (Agresti, 2007, str. 35). Jiné zdroje však uvádí některá méně přísná pravidla, takže není úplně zřejmé, v jakých případech testy lze použít a v jakých nikoli.

Tabulka 2.1 z Příkladu 1 předpoklady o odhadnutých očekávaných četnostech e_{ij} uvedených výše splňuje (dokonce by splňovala i o dost přísnější), navíc 1000 pozorování je pro naše počty kategorií dostatečně velký rozsah. Tedy použití asymptotických testů bylo korektní.

Problém s nulovými řádky či sloupci se typicky řeší vynecháním příslušné kategorie nebo sjednocením s jinou kategorií (tzv. „pooling"). Odstranění řádku či sloupce však lze provést pouze za předpokladu, že má nulovou pravděpodobnost. V případě kladné pravděpodobnosti by došlo ke změně testované hypotézy. Předpokládejme, například, že I-tý řádek neobsahuje žádná pozorování, ale $p_{I+} > 0$, pak po jeho odstranění testujeme hypotézu $H_0 : p_{ij} = p_{i+}\, p_{+j}$ pro $i = 1, \ldots, I-1$ a $j = 1, \ldots, J$. Ta však není ekvivalentní s hypotézou nezávislosti (1.1).

Další nevýhodou je problém se sílou testů. Obecně, pro jakýkoli statistický test, bychom chtěli maximalizovat sílu, tedy zamítat neplatnou hypotézu s co největší pravděpodobností při různých alternativách. To ale většinou není možné pro všechny druhy alternativ. Je však důležité prokázat u daného testu, že má dostatečnou sílu aspoň proti některým druhům alternativ a ty specifikovat. O síle χ^2-testu a G-testu se však z odborné literatury mnoho nedozvíme.

Thomas B. Berrett a Richard J. Samworth se ve svém článku (Berrett a Samworth, 2021) zamysleli nad tím, jak výše zmíněné nedostatky odstranit. Problém nedodržení stanovené hladiny $\alpha \in (0, 1)$ lze vyřešit tak, že kritický obor spočítáme pomocí permutačního testu, který zaručí, že pravděpodobnost chyby I. druhu nepřekročí hladinu pro libovolný rozsah výběru $n \in \mathbb{N}$. Test tedy budeme moct použít i v případě malého počtu pozorování. Permutačním testům se proto budeme věnovat v další sekci.

2.3 Permutační test nezávislosti

Permutační testy (viz např. Pesarin a Salmaso, 2010) jsou přesné nebo přibližné testy, které k výpočtu kritického oboru používají pouze dostupná data, tedy není nutné znát přesné ani asymptotické rozdělení jejich testové statistiky. Nejčastěji se používají při řešení vícevýběrových problémů nebo pro testování nezávislosti, což je naším zájmem.

Připomeneme si problém, který řešíme. Předpokládáme, že máme náhodný výběr $\boldsymbol{X} = (X_1, Y_1)^\top, \ldots, (X_n, Y_n)^\top$ (dále označovaný také jako „data") z rozdělení náhodného vektoru $(X, Y)^\top$ o rozsahu $n \in \mathbb{N}$, kde X a Y jsou dvě kategoriální náhodné veličiny. Chceme testovat nulovou hypotézu (1.1), že veličiny X a Y jsou nezávislé. Zvolíme nějakou vhodnou testovou statistiku T_n citlivou na porušení

hypotézy nezávislosti. Za T_n můžeme vzít například testovou statistiku Pearsonova χ^2 testu (2.1) nebo G-testu (2.2). Pokud nebude řečeno jinak, permutačním testem budeme implicitně označovat jeho přibližnou verzi, která se používá ve většině případů. Kritický obor permutačního testu se určuje na základě tzv. permutačního rozdělení. Jeho konstrukci popisuje Algoritmus 1.

Algoritmus 1 Permutační rozdělení pro test nezávislosti.

Vstup: Náhodný výběr vektorů $\boldsymbol{X} = (X_1, Y_1)^\top, \ldots, (X_n, Y_n)^\top$ o rozsahu $n \in \mathbb{N}$.
Výstup: Permutační rozdělení $\{T_0, T_1, \ldots, T_B\}$ testové statistiky T_n za platnosti nulové hypotézy nezávislosti (1.1).

1: vytvoř kontingenční tabulku O_0 z náhodného výběru $\boldsymbol{X}$
2: spočítej $T_0 \leftarrow T_n(O_0)$
3: zvol $B \in \mathbb{N}$
4: **for** $b = 1$ to B **do**
5: náhodně vyber permutaci $\Pi_b = (\alpha_b(1), \ldots, \alpha_b(n))$ z $(1, \ldots, n)$ bez opakování ze všech $n!$ možných permutací (s diskrétním rovnoměrným rozdělením)
6: **end for**
7: **for** $b = 1$ to B **do**
8: $\boldsymbol{X}_b = (X_1, Y_{\alpha_b(1)})^\top, \ldots, (X_n, Y_{\alpha_b(n)})^\top$
9: vytvoř kontingenční tabulku O_b z $\boldsymbol{X}_b$
10: spočítej $T_b \leftarrow T_n(O_b)$
11: **end for**

Poznamenejme, že je třeba zvolit $B \in \mathbb{N}$ dostatečně velké (v praxi minimálně $B = 1000$). Permutované výběry $\boldsymbol{X}_b$ vzniknou z původního náhodného výběru $\boldsymbol{X}$ permutací druhých složek náhodných vektorů $(X_i, Y_i)^\top$ při zachování jejich prvních složek. Výsledek permutačního testu nezávislosti určíme porovnáním pozorované hodnoty T_0 testové statistiky T_n s teoretickými hodnotami T_b, vypočítanými z B vygenerovaných tabulek. Formálně, určíme si hladinu testu $\alpha \in (0, 1)$. Předpokládejme, že T_n za platnosti H_0 nabývá malých hodnot. Hypotézu nezávislosti zamítneme, právě když $T_0 > c_\alpha$, kde c_α je kritická hodnota definovaná jako

$$c_\alpha = \inf \left\{ r \in \mathbb{R} : \frac{1}{B+1} \sum_{b=0}^{B} \mathbb{1}\{T_b \geq r\} \leq \alpha \right\},$$

což je vlastně $(1 - \alpha)$-kvantil permutačního rozdělení $\{T_0, T_1, \ldots, T_B\}$. Odhad p-hodnoty pak získáme jako

$$\widehat{p}_B = \frac{\sum_{b=0}^{B} \mathbb{1}\{T_b \geq T_0\}}{B+1}.$$

Pokud počet všech permutací $n!$ (tj. počet všech možných tabulek, které lze z původních dat generovat) není příliš velký (pro velmi malá n), a tedy výpočetně ne moc náročný, lze za B vzít jejich celkový počet a p-hodnotu spočítat přesně. Pak se jedná o tzv. přesný permutační test. Ve většině případů však volíme (nebo musíme zvolit kvůli neproveditelnosti přesného testu) menší počet náhodných permutací, typicky $B = 2000$, a spokojíme se s odhadem p-hodnoty. Tomuto způsobu výpočtu se říká Monte Carlo metoda.

Je důležité si uvědomit, že pro správné fungování testu musí permutovaná data $\boldsymbol{X}_b$ použitá pro výpočet teoretických hodnot T_b testové statistiky T_n splňovat nulovou hypotézu, abychom na tyto hodnoty mohli nahlížet jako na odhad rozdělení T_n za nulové hypotézy. Fakt, že tabulky O_b vzniklé z permutovaných dat splňují nulovou hypotézu, plyne z toho, že náhodné vektory v původním výběru $\boldsymbol{X}$ jsou nezávislé, a tedy první složka X_k k-tého vektoru a druhá složka Y_l l-tého vektoru jsou nezávislé pro všechna $k, l \in \{1, \ldots, n\}$, $k \neq l$.

Také je dobré si všimnout, že řádkové a sloupcové součty pozorování v kontingenční tabulce n_{i+} a n_{+j} jsou identické pro původní tabulku O_0 i všechny vytvořené tabulky O_b z náhodně permutovaných dat. Konstrukce permutačního testu nezávislosti se tak dá zjednodušit pouze na náhodné generování kontingenčních tabulek O_b z původní tabulky O_0, sestrojené z pozorovaných dat, při zachování marginálních četností.

Příhodnou vlastností permutačních testů je, že i v případě testování složené hypotézy (hypotéza nezávislosti) je pravděpodobnost chyby I. druhu omezená shora stanovenou hladinou $\alpha \in (0, 1)$ pro všechny rozsahy výběru $n \in \mathbb{N}$, pro které je test definován. Tento fakt je dokázán v článku Berrett a Samworth (2017, lemma 3). Test je tedy konzervativní, z čehož plyne i možná nevýhoda, že skutečná hladina testu bývá menší než tolerovaná pravděpodobnost α. Je to způsobeno tím, že permutační rozdělení je diskrétní, a tak ve výpočtu kritické hodnoty c_α často nelze dosáhnout přesně α. Hlavní nevýhodou permutačních testů je jejich větší výpočetní náročnost oproti asymptotickým testům.

Pro aplikaci na Příklad 1 uvažujme permutační verze Pearsonova χ^2 testu a G-testu s odhadem p-hodnoty metodou Monte Carlo na základě $B = 2000$ vytvořených tabulek. Příslušné p-hodnoty vyjdou popořadě 0,0015 a 0,0010, tedy téměř shodně jako pro asymptotické verze testů. Opět zamítáme hypotézu nezávislosti.

Jak jsme již zmínili, permutační test nezávislosti lze zkonstruovat pro libovolnou vhodně zvolenou testovou statistiku. V další části si představíme specifický permutační test, který klasickou testovou statistiku nemá, funguje na jiném principu.

2.4 Fisherův přesný test

Fisherův přesný test, nazývaný také Fisherův exaktní nebo faktoriálový test, je neparametrická metoda původně navržená v roce 1934 britským statistikem R. A. Fisherem pro kontingenční tabulky typu 2×2 (viz McDonald, 2014, od str. 77). Později, s rozvojem počítačů, byla zobecněna pro libovolné rozměry I, J. Jedná se o speciální typ přesného permutačního testu, který se používá k testování nezávislosti zejména v případě menšího počtu pozorování n a menších rozměrů I, J, tedy v situacích, kdy není vhodné používat asymptotické testy (χ^2 test nebo G-test).

Fisherův přesný test nemá klasickou testovou statistiku. Jeho princip spočívá ve výpočtu podmíněné pravděpodobnosti, že při daných marginálních četnostech n_{i+} a n_{+j} vznikne tabulka s četnostmi n_{ij}. Tato pravděpodobnost se spočítá pro všechny možné realizace tabulek s fixovanými marginálními četnostmi. Označme O_0 původní tabulku sestavenou z pozorovaných dat. Určíme všechny *různé* tabulky $O_1, \ldots, O_B$, které lze sestavit při pevných marginálních četnostech n_{i+} a n_{+j}

původní tabulky O_0. Dále nechť $\mathsf{P}(O_b)$, $b = 0, 1, \ldots, B$, značí příslušné podmíněné pravděpodobnosti jejich nastání. Dá se ukázat (viz např. Anděl, 2011, str. 290), že pro čtyřpolní tabulku tyto pravděpodobnosti tvoří hypergeometrické rozdělení a pro obecnou tabulku s libovolnými rozměry dostaneme mnohorozměrné hypergeometrické rozdělení.

Test rozhoduje o platnosti hypotézy nezávislosti (1.1) na základě p-hodnoty, která se spočítá jako součet pravděpodobností těch tabulek, které představují stejné nebo větší odchýlení od hypotézy nezávislosti než původní tabulka, jsou tedy stejně nebo více „extrémní“. Jedná se o takové tabulky O_b, které splňují $\mathsf{P}(O_b) \leq \mathsf{P}(O_0)$. Pro přesnou p-hodnotu tak platí

$$p = \sum_{b=0}^{B} \mathsf{P}(O_b)\, \mathbb{1}\left\{\mathsf{P}(O_b) \leq \mathsf{P}(O_0)\right\}.$$

Pro tabulky s velkými rozsahy výběru a většími rozměry může být obtížné či nemožné Fisherův test provést přesně kvůli výpočetní náročnosti (i přes postupné zefektivňování algoritmů). V tom případě je nutné použít jiný test nebo využít možnosti přibližného permutačního testu spočítat odhad p-hodnoty

$$\widehat{p}_B = \frac{\sum_{b=0}^{B} \mathbb{1}\{\mathsf{P}(O_b) \leq \mathsf{P}(O_0)\}}{B+1}$$

metodou Monte Carlo na základě menšího počtu $B \in \mathbb{N}$ náhodně generovaných tabulek pomocí permutace původních dat. Pak provedený test pojmenujeme jako *Fisherův přibližný test*.

Jakožto permutační test je Fisherův test konzervativní v důsledku diskrétního pravděpodobnostního rozdělení tabulek. Pro tabulky 2×2 existuje několik testů vycházejících z Fisherova, které sice poskytují o trochu větší sílu, ale jejich výpočetní náročnost je ještě větší než pro Fisherův test. Navíc tyto testy nemusí být (i odborným) čtenářům známé. Příkladem je Barnardův test, který pravděpodobnosti tabulek nepodmiňuje na součtech řádků, ale pouze na součtech sloupců, čímž má větší sílu.

Fisherův přesný test nelze provést na Příkladu 1 kvůli velkému rozsahu výběru ($n = 1000$). Můžeme však použít Fisherův přibližný test s $B = 2000$ vytvořenými tabulkami. Dostaneme pak p-hodnotu 0,0020, na jejímž základě vyvrátíme hypotézu nezávislosti veličin dovolená a sídlo. Vidíme, že i tento test poskytuje silný důkaz proti neplatné hypotéze.

Výčet statistických testů uzavřeme novějším příspěvkem k metodám pro testování nezávislosti.

2.5 USP test nezávislosti

USP test byl představen v roce 2021 v článku Berrett a Samworth (2021). Zkratka USP vychází z tzv. „U-statistic permutation test“, což bychom mohli přeložit jako permutační test založený na U-statistice. Zjednodušeně řečeno, U-statistika je v podstatě aritmetický průměr nějakého nestranného odhadu parametru přes všechny jeho možné hodnoty, spočítané z náhodného výběru. Tomuto nestrannému odhadu se říká jádro. Podrobněji se U-statistikám budeme věnovat v Kapitole 3. Opět nejdříve připomeňme problém, který řešíme. Máme náhodný

výběr $\boldsymbol{X} = (X_1, Y_1)^\top, \ldots, (X_n, Y_n)^\top$ z rozdělení náhodného vektoru $(X, Y)^\top$ o rozsahu $n \in \mathbb{N}$, kde X a Y jsou dvě kategoriální náhodné veličiny. Chceme testovat nulovou hypotézu (1.1), že veličiny X a Y jsou nezávislé. Data uspořádáme do kontingenční tabulky (Tabulka 1.1), n_{ij} jsou pozorované četnosti a e_{ij} (definované v (1.3)) jsou odhadnuté očekávané četnosti za platnosti nulové hypotézy. Testová statistika USP testu je definována pro $n \geq 4$ jako

$$\widehat{U} = \frac{1}{n(n-3)} \sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - e_{ij})^2 - \frac{4}{n(n-2)(n-3)} \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}\, e_{ij}. \tag{2.3}$$

Její první člen kvantifikuje odchýlení od hypotézy nezávislosti, zatímco druhý člen vyjadřuje jakousi korekci zohledňující fakt, že náhodné veličiny n_{ij} a e_{ij} nejsou nezávislé, neboť je počítáme ze stejné tabulky. USP test, stejně jako Fisherův test, netrpí většinou nedostatků, které mají asymptotické testy χ^2 test a G-test, protože jde o permutační test a jeho testová statistika neobsahuje jmenovatel e_{ij}.

Kritický obor a p-hodnotu USP testu určíme na základě permutačního rozdělení vypočteného podle Algoritmu 1. Nechť $\widehat{U}_0$ značí pozorovanou hodnotu testové statistiky (2.3) pro původní tabulku a nechť $\widehat{U}_b$ značí hodnoty testové statistiky vypočítané z B náhodně generovaných tabulek pro $b \in \{1, \ldots, B\}$. Hypotézu nezávislosti zamítneme na hladině $\alpha \in (0, 1)$, právě když $\widehat{U}_0 > c_\alpha$, kde c_α je kritická hodnota definovaná jako

$$c_\alpha = \inf \left\{ r \in \mathbb{R} : \frac{1}{B+1} \sum_{b=0}^{B} \mathbb{1}\{\widehat{U}_b \geq r\} \leq \alpha \right\},$$

což je vlastně $(1-\alpha)$-kvantil permutačního rozdělení $\{\widehat{U}_0, \widehat{U}_1, \ldots, \widehat{U}_B\}$. Odhad p-hodnoty pak získáme jako

$$\widehat{p}_B = \frac{\sum_{b=0}^{B} \mathbb{1}\{\widehat{U}_b \geq \widehat{U}_0\}}{B+1}.$$

Odvození testové statistiky vychází z následující úvahy. Mnoho nedostatků χ^2 testu a G-testu plyne z přítomnosti členů e_{ij} ve jmenovateli jejich testových statistik (2.1) a (2.2). Proto se nabízí posuzovat rozdílnost dvou pravděpodobnostních rozdělení P, P' pomocí klasické eukleidovské metriky, respektive jejího čtverce

$$D(P, P') = \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - p'_{ij})^2.$$

V kontextu testování nezávislosti nás zajímá případ, kdy je rozdělení P' určeno součinem marginálních rozdělení veličin X a Y, tedy $P' = (p'_{ij}) = (p_{i+}\, p_{+j})$. Dosazením do předchozího vztahu dostáváme definici míry závislosti v kontingenční tabulce jako

$$D = \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - p_{i+}\, p_{+j})^2.$$

Nulová hypotéza nezávislosti (1.1) platí, právě když $D = 0$. Čím více je nezávislost porušena, tím se zvětšuje hodnota D.

Míra závislosti D však v praxi nelze spočítat, protože skutečné hodnoty parametrů p_{ij}, p_{i+} a p_{+j} jsou neznámé. Je tedy nasnadě pokusit se D vhodně odhadnout. K odhadu využijeme teorii U-statistik. Uvažujme funkci

$$h\left((x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\Big(\mathbb{1}_{\{x_1=i,\,y_1=j\}}\,\mathbb{1}_{\{x_2=i,\,y_2=j\}} - \\ - 2\,\mathbb{1}_{\{x_1=i,\,y_1=j\}}\mathbb{1}_{\{x_2=i\}}\,\mathbb{1}_{\{y_3=j\}} + \mathbb{1}_{\{x_1=i\}}\,\mathbb{1}_{\{y_2=j\}}\,\mathbb{1}_{\{x_3=i\}}\,\mathbb{1}_{\{y_4=j\}}\Big).$$

Ukážeme, že funkce h aplikovaná na náš náhodný výběr je (nesymetrické, viz Definice 1) jádro parametrické funkce D, tj. že $h\left((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)\right)$ je nestranný odhad D. Ve výpočtu se využívá nezávislost náhodných vektorů a vlastnosti střední hodnoty.

$$\begin{aligned}
&\mathsf{E}\,h((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)) = \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\Big(\mathsf{P}[X_1=i,Y_1=j]\,\mathsf{P}[X_2=i,Y_2=j] - \\
&- 2\;\mathsf{P}[X_1=i,Y_1=j]\,\mathsf{P}[X_2=i]\,\mathsf{P}[Y_3=j] + \\
&+ \mathsf{P}[X_1=i]\,\mathsf{P}[Y_2=j]\,\mathsf{P}[X_3=i]\,\mathsf{P}[Y_4=j]\Big) = \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}(p_{ij}^2 - 2\,p_{ij}\,p_{i+}\,p_{+j} + p_{i+}^2\,p_{+j}^2) = \sum_{i=1}^{I}\sum_{j=1}^{J}(p_{ij} - p_{i+}\,p_{+j})^2 = D.
\end{aligned}$$

Dosazením jádra h do Definice 3.3 dostaneme pro rozsah výběru $n \geq 4$ odhad $\widehat{D}$ hodnoty parametrické funkce D jako U-statistiku čtvrtého řádu (viz Definice 2)

$$\widehat{D} = \frac{1}{n(n-1)(n-2)(n-3)}\sum_{V} h((X_{\alpha_1},Y_{\alpha_1}),(X_{\alpha_2},Y_{\alpha_2}),(X_{\alpha_3},Y_{\alpha_3}),(X_{\alpha_4},Y_{\alpha_4})),$$

kde v sumě sčítáme přes všechny čtyřprvkové variace $V = (\alpha_1,\alpha_2,\alpha_3,\alpha_4)$ z $(1,\ldots,n)$. Tento výraz lze zjednodušit na tvar

$$\begin{aligned}
\widehat{D} =& \frac{1}{n(n-3)}\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-e_{ij})^2 - \frac{4}{n(n-2)(n-3)}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\,e_{ij} + \\
&+ \frac{\sum_{i=1}^{I} n_{i+}^2 + \sum_{j=1}^{J} n_{+j}^2}{n(n-1)(n-3)} + \frac{(3n-2)(\sum_{i=1}^{I} n_{i+}^2)(\sum_{j=1}^{J} n_{+j}^2)}{n^3(n-1)(n-2)(n-3)} - \frac{n}{(n-1)(n-3)}.
\end{aligned}$$

Úprava, kterou se k vyjádření výše dospěje, je nad rámec této práce (viz Berrett, Kontoyiannis a Samworth, 2020). I toto vyjádření je dost složité. Protože však konstruujeme permutační test nezávislosti, ke správnému fungování testu stačí vzít jako testovou statistiku pouze první dva členy odhadu $\widehat{D}$. Je to kvůli tomu, že zbývající členy závisí pouze na hodnotách n, n_{i+} a n_{+j}, které nemění pořadí pozorované hodnoty odhadu $\widehat{D}$ pro náš datový soubor v rámci permutačního rozdělení při konstrukci kritického oboru. Dostaneme tak vzorec testové statistiky USP testu (2.3).

Statistika $\widehat{D}$ je navíc jediný nestranný odhad D s nejmenším rozptylem. Tuto důležitou optimální vlastnost, která zdůvodňuje volbu U-statistiky $\widehat{D}$ pro odhad míry závislosti D, dokázali Berrett a Samworth (2021). Z toho plyne i vhodnost

testové statistiky $\widehat{U}$, neboť permutační test s testovou statistikou $\widehat{D}$ je ekvivalentní permutačnímu testu s testovou statistikou $\widehat{U}$.

Použijme nyní USP test na náš modelový Příklad 1. Jako u předchozích permutačních testů i zde zvolíme počet náhodně generovaných tabulek $B = 2000$. Vyjde nám p-hodnota 0,0920, na základě které nemůžeme zamítnout hypotézu nezávislosti dvou náhodných veličin na hladině 0,05. Jinak řečeno, test nám neposkytl dostatečně silný důkaz k vyvrácení nezávislosti.

Abychom to shrnuli, při vědomí, že v Příkladě 1 neplatí nezávislost, vidíme, že správně (a s podobnými výsledky) rozhodly všechny uvažované testy až na USP test, který v tomto případě selhal. Proč tomu tak je, nahlédneme v rámci simulační studie v Kapitole 4.

3. U-statistiky

Na tomto místě dodatečně vysvětlíme a ilustrujeme na příkladech pojem U-statistika, který je stěžejním nástrojem pro konstrukci USP testu. Budeme vycházet z poznatků v knize R. J. Serflinga (Serfling, 1980, kapitola 5) a částečně také z článku W. Hoeffdinga (Hoeffding, 1948), který s tímto pojmem přišel a položil základy teorie U-statistik.

Definice 1 (Jádro)**.** *Nechť $n, m \in \mathbb{N}$. Nechť $X_1, \ldots, X_n$ je náhodný výběr reálných veličin z rozdělení s distribuční funkcí F. Uvažujme parametr $\theta \in \Theta$, kde $\Theta \subset \mathbb{R}$ je parametrický prostor, jako funkcionál $\theta = \theta(F)$, pro nějž existuje nestranný odhad. Pak $\theta(F)$ nazýváme* regulární funkcionál*. Tedy existuje $m \leq n$ a nějaká měřitelná funkce h splňující*

$$\theta(F) = \mathsf{E}\left[h(X_1, \ldots, X_m)\right] = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} h(x_1, \ldots, x_m)\, \mathrm{d}F(x_1) \cdots \mathrm{d}F(x_m).$$

Funkci $h(x_1, \ldots, x_m)$ nazýváme jádro *(kernel) regulárního funkcionálu $\theta(F)$. Bez újmy na obecnosti budeme předpokládat, že jádro h je symetrické v $x_1, \ldots, x_m$. Když nebude, nahradíme ho symetrickým jádrem*

$$\frac{1}{m!} \sum_{\Pi} h(x_{\alpha_1}, \ldots, x_{\alpha_m}), \tag{3.1}$$

kde v sumě sčítáme přes všechny permutace $\Pi = (\alpha_1, \ldots, \alpha_m)$ z $(1, \ldots, m)$.

Definice 2 (U-statistika)**.** *Uvažujme náhodný výběr veličin $X_1, \ldots, X_n$ z rozdělení s distribuční funkcí F a parametr θ jako regulární funkcionál $\theta = \theta(F)$. Nechť $h(x_1, \ldots, x_m)$ je jádro $\theta(F)$, $m \leq n$. Pak* U-statistiku (m-tého řádu) *definujeme jako funkci výběru*

$$U_n = U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{m}} \sum_{K} h(X_{\alpha_1}, \ldots, X_{\alpha_m}), \tag{3.2}$$

kde v sumě sčítáme přes všechny m-prvkové kombinace $K = (\alpha_1, \ldots, \alpha_m)$ z $(1, \ldots, n)$. Funkci h nazveme jádro statistiky U*.*

V Definici 2 implicitně předpokládáme, že jádro h je symetrické. V případě, že symetrické není, dostaneme dosazením (3.1) do (3.2) alternativní definici U-statistiky pro obecné jádro h

$$U_n = U(X_1, \ldots, X_n) = \frac{1}{n(n-1)\cdots(n-m+1)} \sum_{V} h(X_{\alpha_1}, \ldots, X_{\alpha_m}), \tag{3.3}$$

kde v sumě sčítáme přes všechny m-prvkové variace $V = (\alpha_1, \ldots, \alpha_m)$ z $(1, \ldots, n)$.

Statistika U_n je zřejmě symetrická v $X_1, \ldots, X_n$ a je nestranným odhadem parametru θ (odtud U jako „unbiased“). Je to vlastně aritmetický průměr přes všechny možné hodnoty jádra h, tedy pro výpočet odhadu parametru θ využíváme maximum informací, které máme k dispozici z dat. Jádro h vyčíslené v prvních m pozorováních je sice také nestranný odhad θ, ale méně přesný než U_n (h má větší rozptyl). Intuitivně to plyne z toho, že při výpočtu h nevyužíváme všechnu informaci z pozorovaných dat. Analyticky si to rozebereme níže.

Na následujícím příkladu si ukážeme, že U-statistiky jsou v jistém smyslu zobecněním výběrového průměru. Mnoho známých statistik jsou ve skutečnosti U-statistiky.

Příklad 2.

(i) Chceme určit U-statistiku pro střední hodnotu $\mathsf{E}\,X$ náhodné veličiny X na základě náhodného výběru $X_1, \ldots, X_n$. Máme $\theta = \mathsf{E}\,X = \int_{\mathbb{R}} x \,\mathrm{d}F(x)$. Pro symetrické jádro $h(x_1) = x_1$ dostaneme

$$U(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}_n,$$

což je výběrový průměr.

(ii) Chceme určit U-statistiku pro rozptyl $\mathsf{var}\,X$ náhodné veličiny X na základě náhodného výběru $X_1, \ldots, X_n$. Máme $\theta = \mathsf{var}\,X = \int_{\mathbb{R}} (x-\mu)^2 \,\mathrm{d}F(x)$. Z rovnosti $\mathsf{var}\,X = \mathsf{E}\,X^2 - (\mathsf{E}\,X)^2$ plyne, že můžeme vzít jádro $h_0(x_1, x_2) = x_1^2 - x_1\,x_2$, které však není symetrické. Aplikací výrazu (3.1) na h_0 dostaneme symetrické jádro

$$h(x_1, x_2) = \frac{x_1^2 - 2\,x_1\,x_2 + x_2^2}{2} = \frac{1}{2}(x_1 - x_2)^2,$$

jemuž odpovídající U-statistika je

$$\begin{aligned} U(X_1, \ldots, X_n) &= \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h(X_i, X_j) = \\ &= \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} (X_i - X_j)^2 = \\ &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{X_i^2 + X_j^2 - 2\,X_i\,X_j}{2} = \\ &= \frac{1}{n(n-1)} \left(n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right) = \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\,\overline{X}_n^2 \right) = S_n^2, \end{aligned}$$

tedy výběrový rozptyl.

(iii) Na základě náhodného výběru $X_1, \ldots, X_n$ chceme určit U-statistiku pro distribuční funkci, tedy odhadovaným parametrem bude celá distribuční funkce $F(u)$ pro $u \in \mathbb{R}$. Máme $\theta = F(u) = \int_{-\infty}^{u} \mathrm{d}F(x) = \mathsf{P}[X \le u]$. Pro symetrický kernel $h(x_1) = \mathbb{1}_{(-\infty,\,u\rangle}(x_1)$ dostaneme

$$U(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,\,u\rangle}(X_i) = \widehat{F}_n(u),$$

což je empirická (výběrová) distribuční funkce.

Dále si uvedeme a dokážeme zmíněnou vlastnost U-statistik, totiž že jsou lepšími nestrannými odhady ve smyslu menšího rozptylu než jiné nestranné odhady, což motivuje volbu statistiky $\widehat{D}$ při konstrukci USP testu. Nejprve však odvodíme vyjádření U-statistiky jako podmíněné střední hodnoty potřebné k důkazu této vlastnosti.

Lemma 1. *Nechť $X_1, \dots, X_n$ je náhodný výběr veličin z rozdělení s distribuční funkcí F a mějme jádro $h(x_1, \dots, x_m)$, kde $m \leq n$. Označme $\boldsymbol{X}_{(n)}$ vektor pořádkových statistik $\boldsymbol{X}_{(n)} = (X_{(1)}, \dots, X_{(n)})^\top$. Pak platí*

$$U_n = \mathsf{E}_{\widehat{F}_n}[h(X_1, \dots, X_m) \mid \boldsymbol{X}_{(n)}]. \tag{3.4}$$

Důkaz. [Vlastní] Ve výpočtu střední hodnoty integrujeme podle mnohorozměrné empirické distribuční funkce

$$\widehat{F}_n\big((y_1, \dots, y_m)^\top \mid \boldsymbol{X}_{(n)} = \boldsymbol{x}_{(n)}\big),$$

která určuje podmíněné diskrétní rozdělení náhodného vektoru $(Y_1, \dots, Y_m)^\top$. Jeho nosičem je množina vektorů

$$\left\{(X_{\alpha_1}, \dots, X_{\alpha_m})^\top : (\alpha_1, \dots, \alpha_m) \text{ je } m\text{-prvková variace z } (1, \dots, n)\right\},$$

jejichž složky tvoří m-prvkové variace ze složek vektoru $\boldsymbol{X}_{(n)}$, s pravděpodobnostmi

$$\mathsf{P}[Y_1 = x_{\alpha_1}, \dots, Y_m = x_{\alpha_m} \mid \boldsymbol{X}_{(n)} = \boldsymbol{x}_{(n)}] = \frac{1}{\binom{n}{m}\, m!},$$

kde x_{α_i} jsou realizace náhodných veličin X_{α_i} pro $i = 1, \dots, m$. Bez újmy na obecnosti předpokládáme, že vektor $\boldsymbol{X}_{(n)}$ neobsahuje shodné složky. Z definice podmíněné střední hodnoty pro diskrétně rozdělený náhodný vektor dostaneme vyjádření pravé strany (3.4) jako

$$\sum_V h(X_{\alpha_1}, \dots, X_{\alpha_m})\, \frac{1}{\binom{n}{m}\, m!} = \frac{1}{\binom{n}{m}} \sum_K h(X_{\alpha_1}, \dots, X_{\alpha_m}) = U_n.$$

V první rovnosti jsme využili předpoklad symetrie funkce h. □

Věta 2. *Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení s distribuční funkcí F. Nechť $S = S(X_1, \dots, X_m)$, kde $m \leq n$, je nestranný odhad regulárního funkcionálu $\theta(F)$ s odpovídající U-statistikou U_n. Pak U_n je také nestranným odhadem a platí*

$$\mathsf{var}\, U_n \leq \mathsf{var}\, S, \tag{3.5}$$

přičemž rovnost nastane právě, když $\mathsf{P}[U_n = S] = 1$.

Důkaz. Díky vztahu (3.4) máme $U_n = \mathsf{E}_{\widehat{F}_n}[S \mid \boldsymbol{X}_{(n)}]$, tedy

$$\mathsf{E}\, U_n = \mathsf{E}\,[\mathsf{E}_{\widehat{F}_n}[S \mid \boldsymbol{X}_{(n)}]] = \mathsf{E}\, S = \theta,$$

takže U_n je nestranný. Tudíž k důkazu nerovnosti (3.5) stačí ukázat, že $\mathsf{E}\, U_n^2 \leq \mathsf{E}\, S^2$. S využitím Jensenovy nerovnosti dostaneme

$$\mathsf{E}\, U_n^2 = \mathsf{E}\,[\mathsf{E}_{\widehat{F}_n}[S \mid \boldsymbol{X}_{(n)}]]^2 \leq \mathsf{E}\,[\mathsf{E}_{\widehat{F}_n}[S^2 \mid \boldsymbol{X}_{(n)}]] = \mathsf{E}\, S^2,$$

kde rovnost je právě, když $\mathsf{E}_{\widehat{F}_n}[S \mid \boldsymbol{X}_{(n)}] = S$ skoro jistě. □

Tedy každý nestranný odhad S lze vylepšit (zpřesnit) příslušnou U-statistikou ve smyslu zmenšení rozptylu.

4. Simulační studie

V této kapitole empiricky porovnáme probrané testy nezávislosti pomocí simulovaných dat. Navážeme na simulační studie provedené ve výše zmíněném článku Berrett a Samworth (2021) a doplníme je o nové závěry. Autoři se rozhodli do porovnání testů nezávislosti v kontingenční tabulce nezahrnout Fisherův test. My ho do našich simulací zahrneme, protože se také jedná o poměrně rozšířený test, který je často zmiňován ve statistické literatuře. Budeme však uvažovat pouze jeho přibližnou verzi, protože přesný test nelze na naše simulační scénáře aplikovat kvůli výpočetní náročnosti.

V softwaru R pro statistické výpočty, ve kterém simulace provádíme, je zabudována permutační varianta Pearsonova χ^2 testu (s odhadem p-hodnoty metodou Monte Carlo), ale nikoliv G-testu. Modifikací existujících příkazů jsme si proto vytvořili vlastní funkci `G.test`, jejíž kód poskytujeme v Příloze A. Ostatní zkoumané testy provádíme pomocí příkazů `chisq.test`, `fisher.test` z balíčku `stats` a `USP.test` z balíčku `USP` (Berrett a kol., 2021).

Pro porovnání testů budeme zkoumat jejich *empirickou hladinu*, což je podíl případů zamítnutí platné hypotézy, a *empirickou sílu*, což je podíl případů zamítnutí neplatné hypotézy. Na základě vhodně zvolených simulačních scénářů učiníme závěry a doporučení, v jakých případech je vhodné použít jaký test. Za účelem přehledných definic těchto scénářů si zavedeme následující značení a pojmy. Označme popořadě

$$\boldsymbol{p_{i+}} = (p_{1+}, \ldots, p_{I+})^\top, \quad \boldsymbol{p_{+j}} = (p_{+1}, \ldots, p_{+J})^\top$$

vektory řádkových a sloupcových marginálních pravděpodobností určujících marginální rozdělení dvou kategoriálních veličin. Dále nechť $\mathbf{P}$ značí matici sdružených pravděpodobností $(p_{ij})_{I\times J}$ pro příslušnou kontingenční tabulku $I \times J$. Pak *za platnosti nulové hypotézy* H_0' (1.1) platí

$$\mathbf{P} = \boldsymbol{p_{i+}}\, \boldsymbol{p_{+j}}^\top, \tag{4.1}$$

kde pravá strana vyjadřuje maticový součin příslušných vektorů. Výsledkem je tedy matice typu $I \times J$. Abychom mohli porovnávat sílu testů, je potřeba nějakým způsobem narušit rovnost (4.1) při zachování marginálních pravděpodobností, čímž porušíme nezávislost dvou veličin a vytvoříme alternativu. Matice $\mathbf{P}$ se pak *za platnosti alternativy* H_1' (1.2) dá vyjádřit jako

$$\mathbf{P} = \boldsymbol{p_{i+}}\, \boldsymbol{p_{+j}}^\top + \mathcal{E}, \tag{4.2}$$

kde $\mathcal{E}$ je *matice perturbací* (poruch) nulové hypotézy s nulovými řádkovými a sloupcovými součty. Prvky matice $\mathcal{E}$ lze volit mnoha způsoby. Nabízí se však možnost rozlišit je na dva případy. *Řídké* perturbace se týkají pouze malého počtu prvků matice $\mathbf{P}$, typicky čtyř, což je nejmenší počet prvků, který lze perturbovat. Na druhé straně *husté* perturbace porušují hypotézu ve velkém počtu prvků matice $\mathbf{P}$. Maximální počet perturbovaných prvků zřejmě závisí na tom, zda jsou rozměry I, J matice $\mathbf{P}$ sudé nebo liché. Ve všech simulacích budeme používat nominální hladinu $\alpha = 0{,}05$ a výpočty budeme provádět na 10000 kontingenčních tabulkách generovaných z rozdělení daného maticí $\mathbf{P}$. Pro testy využívající permutační rozdělení bereme $B = 2000$.

Začneme simulacemi, pomocí kterých budeme zkoumat dodržování hladiny testů a jejich sílu proti konkrétní alternativě pro různé rozsahy výběru n. Zvolme $I = 3$ a $J = 4$. Definujme první simulační scénář

$$\boldsymbol{p_{i+}} = \begin{pmatrix} 0{,}22 \\ 0{,}45 \\ 0{,}33 \end{pmatrix}, \quad \boldsymbol{p_{+j}} = \begin{pmatrix} 0{,}50 \\ 0{,}25 \\ 0{,}10 \\ 0{,}15 \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \epsilon & -\epsilon & \epsilon & -\epsilon \\ -\epsilon & \epsilon & 0 & 0 \\ 0 & 0 & -\epsilon & \epsilon \end{pmatrix}, \tag{4.3}$$

kde $\mathcal{E}$ je matice hustých perturbací $\epsilon \geq 0$, které musí být zvoleny dostatečně malé, aby porušené pravděpodobnosti p_{ij} podle (4.2) ležely stále mezi 0 a 1. Zřejmě platí čím větší ϵ, tím větší odchýlení od nezávislosti dvou veličin.

Pro porovnání empirické hladiny testů budeme tabulky generovat z rozdělení určeného v (4.1) s nulovou odchylkou od hypotézy s dosazenými hodnotami definovanými v (4.3). Výsledky simulace, kde χ^2 test a G-test jsou provedeny v klasické asymptotické verzi, pro různé rozsahy výběru n jsou uvedeny v Tabulce 4.1. Jejich grafické znázornění poskytuje Obrázek 4.1 vlevo. Výsledky uka-

Rozsah výběru	**Pearsonův χ^2 test**[a]	**G-test**[a]	**USP test**	**Fisherův test**
20	0,039	0,061	0,051	0,045
50	0,047	0,083	0,052	0,051
100	0,048	0,071	0,049	0,048
200	0,052	0,061	0,053	0,053
300	0,049	0,055	0,050	0,049
500	0,051	0,053	0,050	0,051
700	0,048	0,050	0,050	0,048
1000	0,048	0,050	0,049	0,048
1300	0,048	0,049	0,052	0,050
1600	0,052	0,052	0,052	0,052

Pozn: [a] Asymptotické verze testů.

Tabulka 4.1: Hodnoty empirické hladiny testů pro různé rozsahy výběru podle prvního simulačního scénáře. Pearsonův χ^2 test a G-test byly provedeny v klasické asymptotické verzi.

zují, že pro malé rozsahy výběru ($n < 500$) skutečná hladina (pravděpodobnost chyby I. druhu) asymptotického G-testu výrazně překračuje stanovenou (nominální) hladinu $\alpha = 0{,}05$. Například pro $n = 50$ test zamítal platnou hypotézu v 8,3 % případů namísto požadovaných 5 %. Naopak asymptotický χ^2 test je pro velmi malé rozsahy dost konzervativní, jeho skutečná hladina je menší než nominální. Například pro $n = 20$ test zamítl platnou hypotézu v méně než 4 % případů. Nedodržování stanovené hladiny u asymptotických testů (pro menší rozsahy výběru) se dá předejít použitím permutačních verzí testů, které tímto nedostatkem netrpí, viz Tabulka 4.2 a Obrázek 4.1 vpravo. Empirické hladiny všech testů pro libovolný rozsah n kolísají blízko okolo nominální hladiny $\alpha = 0{,}05$. Odpovídá to teoretickému očekávání. Fisherův a USP test (jakožto permutační testy) hladinu zřejmě také dodržují pro libovolný rozsah n. Simulace jejich hladiny byla pro lepší porovnání provedena v obou případech.

Pro porovnání síly testů proti konkrétní alternativě zvolíme $\epsilon = 0{,}0079$ a definujeme rozdělení splňující alternativu pomocí $\mathbf{P} = \boldsymbol{p_{i+}}\,\boldsymbol{p_{+j}}^{\top} + \mathcal{E}$, kam dosadíme

Rozsah výběru	**Pearsonův χ^2 test**[a]	**G-test**[a]	**USP test**	**Fisherův test**
20	0,049	0,046	0,051	0,045
50	0,050	0,051	0,051	0,050
100	0,054	0,053	0,051	0,055
200	0,050	0,050	0,050	0,050
300	0,055	0,053	0,053	0,054
500	0,050	0,048	0,053	0,048
700	0,048	0,049	0,046	0,048
1000	0,049	0,049	0,049	0,049
1300	0,048	0,047	0,046	0,048
1600	0,052	0,052	0,048	0,052

Pozn: [a] Permutační verze testů.

Tabulka 4.2: Hodnoty empirické hladiny testů pro různé rozsahy výběru podle prvního simulačního scénáře. Pearsonův χ^2 test a G-test byly provedeny v permutační verzi. Výsledky pro USP a Fisherův test se mírně liší od těch v Tabulce 4.1 z důvodu jiné posloupnosti pseudonáhodných tabulek.

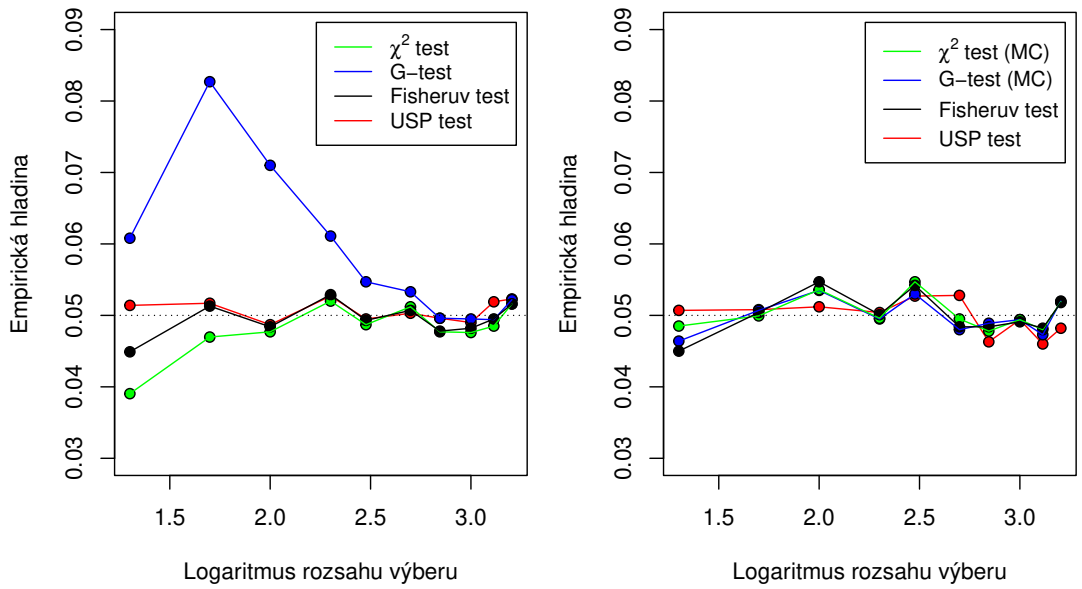
hodnoty z (4.3). Ze stejného rozdělení jsme také vygenerovali Tabulku 2.1 v Příkladu 1. Provedeme simulaci empirické síly testů v závislosti na rozsahu výběru n.

Rozsah výběru	**Pearsonův χ^2 test**[a]	**G-test**[a]	**USP test**	**Fisherův test**
100	0,093	0,095	0,071	0,092
200	0,150	0,147	0,100	0,148
300	0,201	0,200	0,122	0,201
500	0,344	0,346	0,193	0,344
700	0,482	0,482	0,266	0,481
1000	0,664	0,664	0,404	0,663
1300	0,797	0,800	0,536	0,797
1600	0,883	0,884	0,650	0,883
2000	0,950	0,951	0,784	0,951
2500	0,983	0,983	0,884	0,983
3000	0,995	0,996	0,948	0,996

Pozn: [a] Permutační verze testů.

Tabulka 4.3: Hodnoty empirické síly testů pro různé rozsahy výběru pro perturbaci $\epsilon = 0{,}0079$ podle prvního simulačního scénáře. Pearsonův χ^2 test a G-test byly provedeny v permutační verzi.

Obdržené výsledky uvádíme v Tabulce 4.3. Jejich grafické znázornění poskytuje Obrázek 4.2. Použili jsme pouze permutační verze χ^2 testu a G-testu, protože, jak jsme ukázali, pro malá n jsou asymptotické testy nevhodné, a navíc pro velká n je jejich síla v podstatě stejná jako u permutačních. Vidíme, že s rostoucím n se síla testů zvětšuje (podle očekávání), ale pro USP test v tomto případě síla roste o dost pomaleji než pro ostatní testy, které vykazují téměř shodné výsledky. Například pro $n = 1000$ jako v Příkladě 1 je empirická síla USP testu pouze 0,40

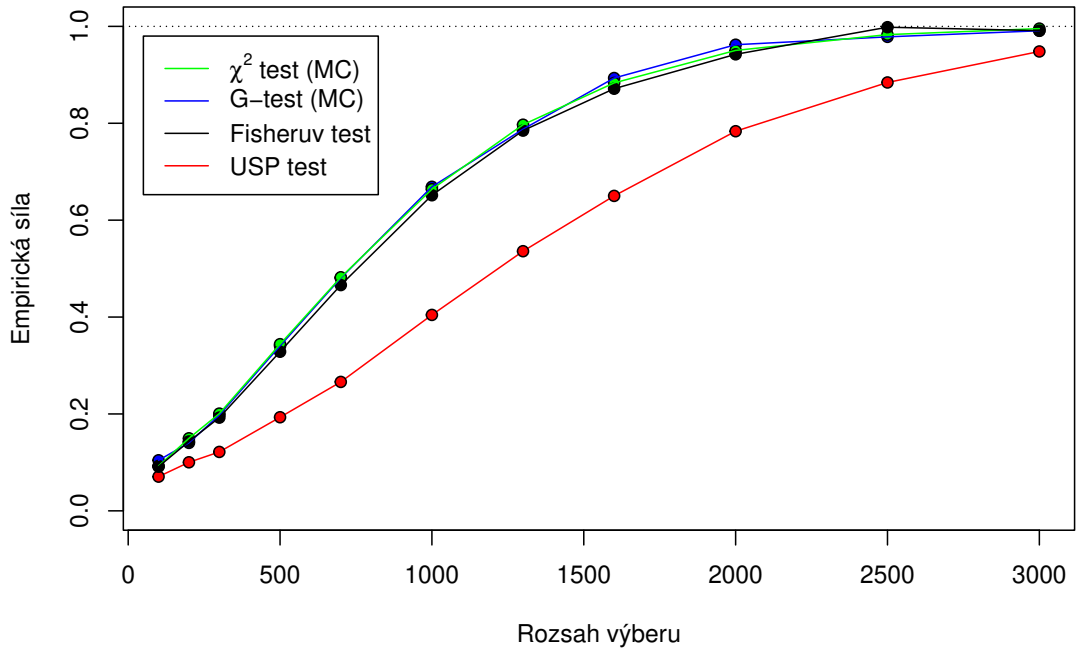


Obrázek 4.1: Grafy empirické hladiny v závislosti na (dekadickém) logaritmu rozsahu výběru n pro první simulační scénář. Vlevo pro asymptotické verze Pearsonova χ^2 testu a G-testu, vpravo pro jejich permutační verze (v legendě označeny MC jako Monte Carlo).

(tedy neplatnou hypotézu zamítal pouze ve 40 % případů), zatímco pro ostatní testy se rovná přibližně 0,66. To nám poskytuje vysvětlení, proč v Příkladu 1 USP test jako jediný chybně nezamítl neplatnou hypotézu nezávislosti.

Že se USP testu pro scénář (4.3) nedaří v porovnání s ostatními testy, nahlédneme ještě pomocí simulace empirické síly v závislosti na různých hodnotách perturbace ϵ pro pevný počet pozorování $n = 1000$ (viz Tabulka 4.4 a Obrázek 4.3). Tentokrát uvažujeme pouze asymptotické varianty χ^2 testu a G-testu, neboť n je velké. Permutační verze však dávají analogické výsledky. Je zřejmé, že USP test má menší sílu než ostatní testy, které dávají lepší a v podstatě stejné výsledky. Například pro $\epsilon = 0{,}009$ USP test zamítal hypotézu pouze v 54 % případů, zatímco ostatní testy neplatnost hypotézy detekovaly v 80 % experimentů.

USP test zřejmě tedy v některých případech selhává. Berrett a Samworth (2021) však předložili pouze příklady a simulační scénáře, ve kterých si USP test vede nejlépe, případně velmi podobně jako ostatní testy. Neukázali však situace, kdy USP test výrazně selhává oproti jiným testům. Jedním z těchto případů je scénář (4.3). Autoři článku ukázali, že USP test má poměrně velkou sílu i proti velmi malým odchylkám od nulové hypotézy nezávislosti, ale pouze v případě, kdy jsou odchylky přítomné zejména v buňkách s vysokými pravděpodobnostmi. Z toho plyne, že USP test je žádoucí použít v situaci, kdy (v praxi neznámé) pravděpodobnostní rozdělení pro pozorovanou tabulku obsahuje řídké perturbace a pouze v kategoriích s vyššími pravděpodobnostmi p_{ij}. Pokud jsou perturbovány také (nebo jenom) nízké pravděpodobnosti p_{ij}, pak USP test ztrácí na síle a může být o dost horší než jiné testy. To je případ scénáře (4.3), snadno totiž nahlédneme, že nezávislost je porušena i v buňkách s malými pravděpodobnostmi (p_{13} a p_{14}). Zformulovaná tvrzení o síle USP testu demonstrujeme na následující simulaci.



Obrázek 4.2: Graf empirické síly v závislosti na rozsahu výběru n pro perturbaci $\epsilon = 0{,}0079$ pro první simulační scénář. Pearsonův χ^2 test a G-test byly provedeny v permutační verzi. Kvůli lepšímu zobrazení byl k hodnotám prvních tří testů v legendě přičten náhodný šum (jitter).

Zvolíme opět rozsah $n = 1000$ a rozměry $I = J = 4$. Definujme

$$\boldsymbol{p}_{i+} = \begin{pmatrix} 0{,}50 \\ 0{,}30 \\ 0{,}12 \\ 0{,}08 \end{pmatrix}, \quad \boldsymbol{p}_{+j} = \begin{pmatrix} 0{,}40 \\ 0{,}35 \\ 0{,}15 \\ 0{,}10 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \epsilon & -\epsilon \\ -\epsilon & \epsilon \end{pmatrix}, \quad \mathcal{E}_1 = \left(\begin{array}{c|c} \boldsymbol{\varepsilon} & 0_{2\times 2} \\ \hline 0_{2\times 2} & 0_{2\times 2} \end{array}\right), \tag{4.4}$$

$$\mathcal{E}_2 = \left(\begin{array}{c|c} \boldsymbol{\varepsilon} & 0_{2\times 2} \\ \hline \boldsymbol{\varepsilon} & 0_{2\times 2} \end{array}\right), \quad \mathcal{E}_3 = \left(\begin{array}{c|c} 0_{1\times 2} & 0_{1\times 2} \\ \hline \boldsymbol{\varepsilon} & \boldsymbol{\varepsilon} \\ \hline 0_{1\times 2} & 0_{1\times 2} \end{array}\right), \quad \mathcal{E}_4 = \left(\begin{array}{c|c} \boldsymbol{\varepsilon} & 0_{2\times 2} \\ \hline 0_{2\times 2} & \boldsymbol{\varepsilon} \end{array}\right),$$

kde $\mathcal{E}_k$, $k = 1, 2, 3, 4$ jsou blokové matice typu 4×4 různě uspořádaných bloků $\boldsymbol{\varepsilon}$ s perturbacemi $\epsilon \geq 0$ hypotézy nezávislosti. Matice $\mathcal{E}_k$ jsou zvoleny tak, aby s rostoucím indexem k byly odchylky ϵ přítomny ve větším počtu buněk s menšími pravděpodobnostmi p_{ij}. Budeme tedy uvažovat čtyři různé matice pravděpodobností $\mathbf{P}_k$ za platné alternativy, splňující $\mathbf{P}_k = \boldsymbol{p}_{i+}\,\boldsymbol{p}_{+j}^{\top} + \mathcal{E}_k$, kde $k \in \{1, 2, 3, 4\}$. Zajímá nás síla testů pro čtyři definované scénáře a pro různé velikosti odchylek ϵ. Do simulací opět zahrneme pouze asymptotické verze χ^2 testu a G-testu, neboť n je dostatečně velké. V prvním případě, kdy perturbujeme pouze relativně vysoké pravděpodobnosti (scénář $\mathbf{P}_1$) si USP test vede podle očekávání nejlépe, má největší sílu, zatímco ostatní testy si vedou o dost hůře (a vzájemně v podstatě stejně), viz Tabulka 4.5 a Obrázek 4.4 vlevo. Například pro $\epsilon = 0{,}018$ USP test zamítal hypotézu v 81 % případů, zatímco ostatní testy neplatnost hypotézy detekovaly pouze v cca 52 % experimentů.

V druhé simulaci podle matice $\mathbf{P}_2$ se už projevuje přítomnost odchylek ϵ i v menších pravděpodobnostech p_{ij} ztrátou síly USP testu (viz Tabulka 4.6 a Obrázek 4.4 vpravo). Rozdíl v síle mezi ním a ostatními testy ale zatím není propastný (pro $\epsilon = 0{,}0102$ je síla USP testu 0,66, u ostatních asi 0,77).

Velikost perturbace	**Pearsonův χ^2 test**[a]	**G-test**[a]	**USP test**	**Fisherův test**
0,0000	0,052	0,054	0,054	0,053
0,0015	0,065	0,068	0,056	0,066
0,0030	0,116	0,121	0,084	0,118
0,0045	0,230	0,234	0,137	0,230
0,0060	0,414	0,420	0,234	0,412
0,0075	0,610	0,617	0,364	0,609
0,0090	0,800	0,804	0,539	0,799
0,0105	0,915	0,917	0,713	0,915
0,0120	0,980	0,980	0,861	0,979
0,0135	0,995	0,996	0,947	0,996
0,0150	0,999	0,999	0,984	0,999

Pozn: [a] Asymptotické verze testů.

Tabulka 4.4: Hodnoty empirické síly testů pro různé velikosti perturbace ϵ při rozsahu výběru $n = 1000$ podle prvního simulačního scénáře. Pearsonův χ^2 test a G-test byly provedeny v klasické asymptotické verzi.

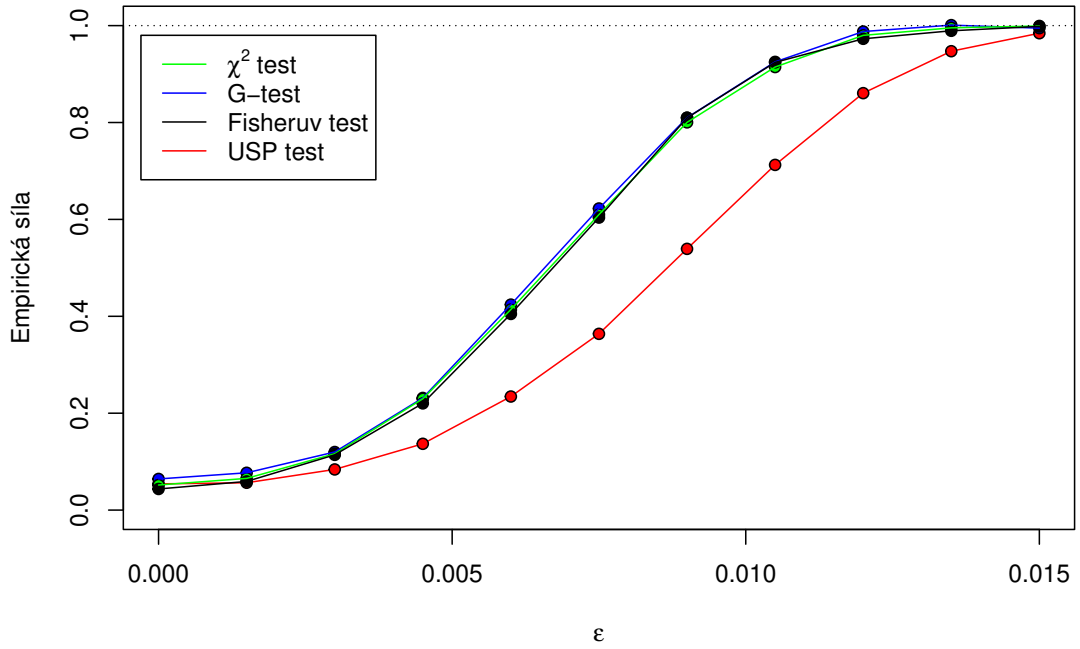
Pro třetí scénář $\mathbf{P}_3$ už je síla USP testu chabá v porovnání s jinými testy (viz Tabulka 4.7 a Obrázek 4.5 vlevo). Perturbujeme totiž spíše malé pravděpodobnosti, naopak ty velké nikoli. Markantní rozdíl v síle ilustrujeme pro $\epsilon = 0{,}0075$. USP test zamítal neplatnou hypotézu pouze v 37 % případů, zatímco ostatní testy ji detekovaly v 75 % experimentů, což dělá rozdíl 38 procentních bodů.

V posledním případě, kdy generujeme tabulky z rozdělení $\mathbf{P}_4$, už USP test takřka zcela selhává a není vhodné ho používat (viz Tabulka 4.8 a Obrázek 4.5 vpravo). Je to způsobeno tím, že odchylky ϵ jsou přítomny i v buňkách s nejmenšími pravděpodobnostmi. Ostatní testy mají v tomto případě znatelně větší sílu, nejlépe si vede G-test. Jeho síla pro $\epsilon = 0{,}0072$ je 0,917, zatímco pro USP test pouze 0,325, což činí rozdíl téměř 0,6.

Pokud bychom ještě uvažovali případ, kdy v matici $\mathcal{E}_4$ v (4.4) vyměníme levý horní blok za nulovou matici $0_{2\times 2}$ a ostatní bloky necháme, čímž bychom perturbovali jenom ty nejnižší pravděpodobnosti p_{ij}, dostali bychom ještě horší výsledky USP testu než pro poslední uvažovaný scénář. To znovu potvrzuje nevhodnost použití USP testu v těchto případech.

Na předešlých simulacích jsme si ukázali, kdy je (respektive není) vhodné USP test použít v případě různě velkých pravděpodobností buněk kontingenční tabulky. Pro tabulky, jejichž všechny pravděpodobnosti p_{ij} jsou před porušením hypotézy nezávislosti (jak řídkými, tak hustými perturbacemi) přibližně stejně velké, mají všechny testy podobnou sílu, takže můžeme použít libovolný z nich, respektive se rozhodovat na základě jiných kritérií (např. rozsahu výběru). Tento závěr plyne i z druhé simulace v uvedeném článku (sekce 3.2).

Poznamenejme ještě, že v našich simulacích jsme pro jednoduchost uvažovali pouze konstantní perturbaci ϵ, která se v absolutní hodnotě nesnižovala s klesajícími pravděpodobnostmi p_{ij}. Jeden takový případ, kde tomu je naopak, ukázali Berrett a Samworth (2021) v sekci 5.5. Plyne z něj, že pro velikosti poruch klesající úměrně pravděpodobnostem má USP test stále trochu větší sílu.

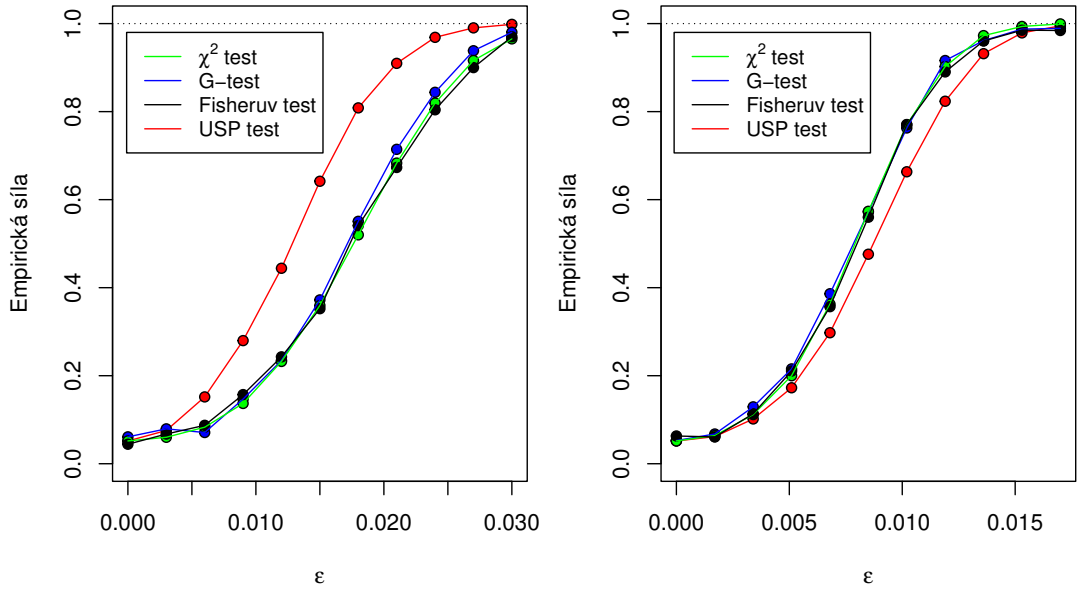


Obrázek 4.3: Graf empirické síly v závislosti na velikosti perturbace ϵ pro rozsah výběru $n = 1000$ pro první simulační scénář. Pearsonův χ^2 test a G-test byly provedeny v klasické asymptotické verzi. Kvůli lepšímu zobrazení byl k hodnotám prvních tří testů v legendě přičten náhodný šum (jitter).

Velikost perturbace	**Pearsonův χ^2 test**[a]	**G-test**[a]	**USP test**	**Fisherův test**
0,000	0,052	0,057	0,052	0,054
0,003	0,060	0,066	0,075	0,060
0,006	0,083	0,088	0,152	0,084
0,009	0,137	0,141	0,280	0,139
0,012	0,232	0,237	0,444	0,233
0,015	0,359	0,365	0,642	0,361
0,018	0,520	0,526	0,808	0,525
0,021	0,683	0,687	0,910	0,686
0,024	0,819	0,822	0,969	0,822
0,027	0,916	0,918	0,990	0,918
0,030	0,965	0,966	0,998	0,966

Pozn: [a] Asymptotické verze testů.

Tabulka 4.5: Hodnoty empirické síly testů pro různé velikosti perturbace ϵ při rozsahu výběru $n = 1000$ podle simulačního scénáře určeného maticí $\mathbf{P}_1$. Pearsonův χ^2 test a G-test byly provedeny v klasické asymptotické verzi.



Obrázek 4.4: Grafy empirické síly v závislosti na velikosti perturbace ϵ pro rozsah výběru $n = 1000$. Vlevo jsou výsledky pro scénář $\mathbf{P}_1$, vpravo pro $\mathbf{P}_2$. Pearsonův χ^2 test a G-test byly provedeny v klasické asymptotické verzi. Kvůli lepšímu zobrazení byl v obou grafech k hodnotám prvních tří testů v legendě přičten náhodný šum (jitter).

Velikost perturbace	**Pearsonův χ^2 test**[a]	**G-test**[a]	**USP test**	**Fisherův test**
0,0000	0,052	0,057	0,052	0,054
0,0017	0,064	0,068	0,062	0,066
0,0034	0,112	0,118	0,102	0,112
0,0051	0,200	0,209	0,173	0,203
0,0068	0,362	0,373	0,298	0,365
0,0085	0,574	0,581	0,476	0,576
0,0102	0,770	0,776	0,663	0,772
0,0119	0,902	0,906	0,824	0,904
0,0136	0,973	0,974	0,932	0,973
0,0153	0,994	0,994	0,979	0,994
0,0170	0,999	1,000	0,995	1,000

Pozn: [a] Asymptotické verze testů.

Tabulka 4.6: Hodnoty empirické síly testů pro různé velikosti perturbace ϵ při rozsahu výběru $n = 1000$ podle simulačního scénáře určeného maticí $\mathbf{P}_2$. Pearsonův χ^2 test a G-test byly provedeny v klasické asymptotické verzi.

Velikost perturbace	**Pearsonův χ^2 test**[a]	**G-test**[a]	**USP test**	**Fisherův test**
0,0000	0,052	0,057	0,052	0,054
0,0015	0,064	0,068	0,053	0,065
0,0030	0,133	0,134	0,076	0,132
0,0045	0,289	0,289	0,131	0,286
0,0060	0,524	0,527	0,220	0,521
0,0075	0,751	0,754	0,371	0,751
0,0090	0,917	0,923	0,576	0,921
0,0105	0,983	0,984	0,768	0,983
0,0120	0,997	0,997	0,905	0,997
0,0135	1,000	1,000	0,975	1,000
0,0150	1,000	1,000	0,995	1,000

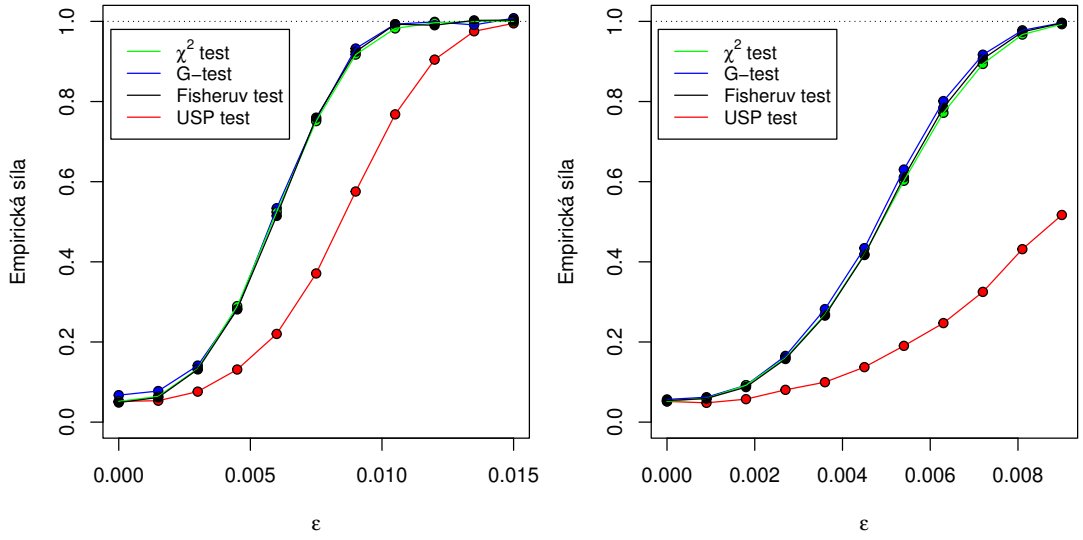
Pozn: [a] Asymptotické verze testů.

Tabulka 4.7: Hodnoty empirické síly testů pro různé velikosti perturbace ϵ při rozsahu výběru $n = 1000$ podle simulačního scénáře určeného maticí $\mathbf{P}_3$. Pearsonův χ^2 test a G-test byly provedeny v klasické asymptotické verzi.

Velikost perturbace	**Pearsonův χ^2 test**[a]	**G-test**[a]	**USP test**	**Fisherův test**
0,0000	0,052	0,057	0,052	0,054
0,0009	0,059	0,062	0,048	0,059
0,0018	0,092	0,092	0,057	0,087
0,0027	0,160	0,165	0,080	0,158
0,0036	0,268	0,282	0,100	0,266
0,0045	0,419	0,434	0,137	0,417
0,0054	0,602	0,630	0,190	0,611
0,0063	0,772	0,801	0,247	0,785
0,0072	0,894	0,917	0,325	0,906
0,0081	0,967	0,978	0,432	0,974
0,0090	0,993	0,996	0,517	0,996

Pozn: [a] Asymptotické verze testů.

Tabulka 4.8: Hodnoty empirické síly testů pro různé velikosti perturbace ϵ při rozsahu výběru $n = 1000$ podle simulačního scénáře určeného maticí $\mathbf{P}_4$. Pearsonův χ^2 test a G-test byly provedeny v klasické asymptotické verzi.



Obrázek 4.5: Grafy empirické síly v závislosti na velikosti perturbace ϵ pro rozsah výběru $n = 1000$. Vlevo jsou výsledky pro scénář $\mathbf{P}_3$, vpravo pro $\mathbf{P}_4$. Pearsonův χ^2 test a G-test byly provedeny v klasické asymptotické verzi. Kvůli lepšímu zobrazení byl v grafu vlevo (vpravo nikoli) k hodnotám prvních tří testů v legendě přičten náhodný šum (jitter).

Závěr

Představili jsme a popsali nejpoužívanější testy nezávislosti pro dvourozměrné kontingenční tabulky, k nimž jsme přidali jednoho nováčka – USP test. Diskutovali jsme jejich předpoklady, výhody a nedostatky. Seznámili jsme čtenáře se základy teorie U-statistik, na nichž je USP test postaven. Na základě empirických poznatků ze simulačních studií jsme učinili závěry o tom, v jakých případech je daný test vhodný použít a v jakých nikoli.

Kvůli nedodržování nominální hladiny $\alpha \in (0, 1)$ pro malé rozsahy výběru $n \in \mathbb{N}$ u klasických asymptotických verzí Pearsonova χ^2 testu a G-testu doporučujeme pro malá n používat jejich permutační verze nebo Fisherův přesný test, u nichž je nepřekročení hladiny zaručeno pro libovolná n. K rozhodnutí o tom, co je malé n, může orientačně posloužit asi nejpřísnější požadavek z literatury na odhadnuté očekávané četnosti e_{ij}, tedy $e_{ij} \geq 5$ pro všechna i, j. Nicméně v současné době není pro počítače problém spočítat permutační testy i pro velká n, takže když si člověk není jistý, zda je n dostatečně velké, s permutační verzí neudělá chybu. Pro velká n asymptotické verze většinou hladinu dodržují, takže klidně můžeme zůstat u nich, výpočty budou rychlejší.

Zajímavých výsledků v našich simulacích dosahoval Fisherův přibližný test, který byl stejně silný jako asymptotický χ^2 test a G-test, ale na rozdíl od nich se dá použít i pro malá n. Ale protože neposkytl žádnou sílu navíc, pro velká n je lepší zůstat u zmíněných dvou testů.

Největší pozornost jsme věnovali USP testu, který se dá použít pro libovolný rozsah n a který v určitých případech poskytuje extra sílu v porovnání se všemi ostatními testy. Ukázali jsme, že se jedná o případy, kdy je nezávislost dvou kategoriálních veličin porušena především v buňkách s relativně vysokými sdruženými pravděpodobnostmi p_{ij} (za předpokladu, že sdružené rozdělení zahrnuje různě velké pravděpodobnosti). Pak je USP test nejlepší. Pokud je však výrazně porušena nezávislost i (nebo jenom) v méně pravděpodobných kategoriích, síla USP testu se výrazně zhoršuje proti ostatním testům, až může úplně selhat. V případě přibližně stejně velkých pravděpodobností kategorií má USP test víceméně stejně velkou sílu jako ostatní testy.

S USP testem je tedy potřeba zacházet obezřetně a jeho volbu podložit dostatečnou explorační analýzou (například porovnáním pozorovaných četností n_{ij} s odhadnutými očekávanými četnostmi e_{ij} nebo vhodnými grafickými nástroji), na základě které přijmeme předpoklad, že nezávislost je porušena nejvíce a především v kategorích s velkými četnostmi (s vysokou pravděpodobností výskytu).

Seznam použité literatury

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Second edition. Wiley-Interscience, New Jersey. ISBN 978-0471226185.

Anděl, J. (2011). *Základy matematické statistiky*. Třetí vydání. Matfyzpress, Praha. ISBN 978-80-7378-162-0.

Berrett, T. B. a Samworth, R. J. (2017). Nonparametric independence testing via mutual information. *arXiv e-prints*. URL `https://arxiv.org/abs/1711.06642`.

Berrett, T. B. a Samworth, R. J. (2021). USP: an independence test that improves on Pearson's chi-squared and the G-test. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **477**(2256). URL `https://doi.org/10.1098%2Frspa.2021.0549`.

Berrett, T. B., Kontoyiannis, I. a Samworth, R. J. (2020). Optimal rates for independence testing via U-statistic permutation tests. *arXiv e-prints*. URL `https://arxiv.org/abs/2001.05513`.

Berrett, T. B., Kontoyiannis, I. a Samworth, R. J. (2021). *USP: U-Statistic Permutation Tests of Independence for all Data Types*. URL `https://CRAN.R-project.org/package=USP`. R package version 0.1.2.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, **19**, 293–325. ISSN 0003-4851. doi: 10.1214/aoms/1177730196.

McDonald, J. H. (2014). *Handbook of Biological Statistics*. Third edition. Sparky House Publishing, Baltimore.

Pesarin, F. a Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*. First edition. Wiley-Interscience, New Jersey. ISBN 978-0-470-51641-6.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley-Interscience, New York. ISBN 0-471-21927-4.

A. Příloha

Elektronická příloha obsahující zdrojový kód G-testu s možností simulované p-hodnoty metodou Monte Carlo v programu R.