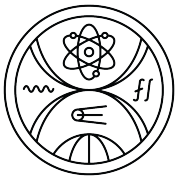
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



Testovanie nezávislosti vo svete ordinálnych dát

DIPLOMOVÁ PRÁCA

2026 Bc. Veronika GAŠPAROVIČOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Testovanie nezávislosti vo svete ordinálnych dát

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Ján Somorčík, PhD.

Bratislava 2026 **Bc. Veronika GAŠPAROVIČOVÁ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Veronika Gašparovičová
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Testovanie nezávislosti vo svete ordinálnych dát
Independence testing in the world of ordinal data

Anotácia: Ordinálna premenná nadobúda hodnoty, ktoré síce sú nečíselné, no možno ich usporiadať. Aj pre takéto premenné boli vymyslené číselné ukazovatele merajúce ich vzájomnú závislosť. V drvivej väčšine prípadov sú však následné testy nezávislosti len asymptotické, teda na správne fungovanie vyžadujú veľký počet dát. Oveľa menej známe sú testy nezávislosti 'vyšľachtené' práve pre malé sady dát.

Vedúci: Mgr. Ján Somorčík, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: doc. Mgr. Igor Melicherčík, PhD.

Dátum zadania: 09.01.2025

Dátum schválenia: 15.01.2025 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Čestné vyhlásenie

Čestne vyhlasujem, že celú diplomovú prácu na tému „Testovanie nezávislosti vo svete ordinálnych dát“, vrátane všetkých jej príloh a obrázkov, som vypracovala samostatne, a to s použitím literatúry uvedenej v priloženom zozname a nástrojov umelej inteligencie. Vyhlasujem, že nástroje umelej inteligencie som používala v súlade s príslušnými právnymi predpismi, akademickými právami a slobodami, etickými a morálnymi zásadami za súčasného dodržiavania akademickej integrity a že ich použitie je v práci vhodným spôsobom označené.

Miesto a dátum: ..... Titul, meno a priezvisko: .....

Abstrakt v štátnom jazyku

GAŠPAROVIČOVÁ, Veronika: *Testovanie nezávislosti vo svete ordinálnych dát* [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Somorčík, PhD., Bratislava, 2026, 81 s.

Diplomová práca sa zaoberá problematikou testovania nezávislosti ordinálnych premenných. V práci sú predstavené vybrané miery asociácie, konkrétne Kendallovo τ_b, Goodman-Kruskalova γ, Pearsonovo ρ a Kendallovo ρ_b založené na ridit skóre. Pre tieto miery sú odvodené najmä ich asymptotické variancie pri rôznych spôsoboch generovania dát. Hlavným cieľom práce je analyzovať správanie testov nezávislosti pri ordinálnych dátach. Teoretické poznatky sú doplnené o simulačné štúdie, ktoré skúmajú vplyv rôznych pravdepodobnostných štruktúr kontingenčných tabuliek na pravdepodobnosť chyby I. druhu a silu testu. Prínos práce spočíva v prepojení teoretických odvodení so simulačnými experimentmi, identifikácii chyby v použitej literatúre a v detailnej analýze správania testov nezávislosti za rôznych podmienok.

Kľúčové slová: ordinálne dáta, miera asociácie, test nezávislosti, chyba I. druhu, simulácie

Abstract

GAŠPAROVIČOVÁ, Veronika: *Independence testing in the world of ordinal data* [Master's Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; supervisor: Mgr. Ján Somorčík, PhD., Bratislava, 2026, 81 p.

This master's thesis deals with the problem of testing independence between ordinal variables. The thesis presents selected measures of association, namely Kendall's τ_b, Goodman–Kruskal's γ, Pearson's ρ, and Kendall's ρ_b based on ridit scores. For these measures, their asymptotic variances are derived, in particular, under different data-generating mechanisms. The main objective of the thesis is to analyze the behavior of independence tests for ordinal data. The theoretical results are complemented by simulation studies that investigate the effect of different probability structures of contingency tables on the type I error rate and the power of the test. The contribution of the thesis lies in linking theoretical derivations with simulation experiments, identifying an error in the referenced literature, and providing a detailed analysis of the behavior of independence tests under various conditions.

Keywords: ordinal data, measure of association, test of independence, type I error, simulations

Obsah

Úvod **9**

1 Popis dôležitých pojmov **11**

2 Kendallovo τ_b **18**

2.1 Odvodenie variancie pre Kendallovo τ_b pre jeden multinomický výber . . . 18

2.2 Odvodenie variancie pre Kendallovo τ_b pre nezávislé multinomické vzorky . 24

2.3 Dôkaz o rovnosti variancií 26

2.4 Chyba v Agrestiho vzorci 30

2.4.1 Relatívna chyba pre populačnú disperziu 31

2.4.2 Relatívna chyba pre výberovú disperziu 33

2.4.3 Vplyv chyby na správanie testov 37

2.4.4 Funkcia z balíka DescTools a komunikácia s profesorom Agrestim . 41

2.5 Simulačné štúdie 42

2.5.1 Simulácie za platnosti H_0 42

2.5.2 Simulácie za platnosti H_1 46

2.6 Vplyv zlučovania kategórií 48

2.6.1 Za platnosti H_0 48

2.6.2 Za platnosti H_1 50

3 Goodman-Kruskalova γ **52**

3.1 Odvodenie variancie pre Goodman–Kruskalovu γ pre nezávislé multinomické vzorky 53

3.2 Dôkaz o rovnosti sigiem 54

3.3 Simulačné štúdie 55

3.3.1 Simulácie za platnosti H_0 55

3.3.2 Simulačné porovnanie pravdepodobnosti chyby I. druhu pre Goodman–Kruskalovu γ a jej zeta-transformáciu 58

3.3.3 Simulácie za platnosti H_1 61

3.4 Vplyv zlučovania kategórií 63

3.4.1 Za platnosti H_0 63

3.4.2 Za platnosti H_1 65

4 Korelačné koeficienty $\hat{\rho}$ a $\hat{\rho}_b$ **67**

4.1 Základné vlastnosti ridit skóre 68

4.2 Simulácie za platnosti H_0 69

4.3 Vzťah medzi Kendallovým τ_b a Spearmanovou, resp. Pearsonovou koreláciou pre 2×2 tabuľku 71

5 Správanie testov podľa jednotlivých mier asociácie **75**

5.1 Za platnosti H_0 75

5.2 Za platnosti H_1 77

Záver **79**

Zoznam použitej literatúry **81**

Úvod

Vzťah medzi ordinálnymi premennými patrí medzi často skúmané problémy v analýze dát, najmä v oblastiach, kde sú pozorovania usporiadané do prirodzeného poradia, avšak bez presne definovaných vzdialeností medzi jednotlivými kategóriami. Na opis a testovanie takýchto vzťahov sa používajú špeciálne miery asociácie, ktoré vychádzajú z porovnávania počtov konkordantných a diskordantných párov pozorovaní.

Cieľom tejto práce je analyzovať vybrané testy nezávislosti založené na ordinálnych mierach asociácie a preskúmať ich vlastnosti z hľadiska empirickej pravdepodobnosti chyby I. druhu a sily.

Prvá kapitola práce je zameraná na zavedenie základných pojmov a označenia, na ktoré sa v ďalších častiach práce budeme odvolávať. Čerpali sme najmä z knihy od Alana Agrestiho [1]. Sú v nej definované pojmy konkordantnosti a diskordantnosti, asymptotické vlastnosti používaných mier, formulácia testov nezávislosti založených na konkordantných a diskordantných pároch a kapitola taktiež opisuje spôsoby konštrukcie pravdepodobnostných tabuliek, ktoré tvoria základ následnej simulačnej štúdie.

Druhá kapitola sa venuje Kendallovmu τ_b. Najskôr je odvodená jeho asymptotická variancia pre jeden multinomický výber aj pre nezávislé multinomické výbery, následne je analyzovaná chyba vo vzorci v knihe [1] a jej vplyv na výsledky testovania, ktorý je ilustrovaný pomocou simulačných štúdií. Pri odvodení asymptotickej variancie sme sa inšpirovali článkom [7]. V kapitole prezentujeme aj výsledky simulačných štúdií zameraných na pravdepodobnosť chyby I. druhu a silu testu pri použití správneho, resp. nesprávneho odhadu štandardnej odchýlky vzhľadom na spôsob generovania dát. Kapitolu uzatvárame simuláciami ilustrujúcimi vplyv zlučovania kategórií v kontingenčnej tabuľke na pravdepodobnosť chyby I. druhu a silu testu.

Tretia kapitola sa zaoberá Goodman–Kruskalovou mierou asociácie γ. Najskôr odvodzujeme jej asymptotickú varianciu pri nezávislých multinomických výberoch, ktorú následne využívame v simulačných štúdiách na porovnanie výsledkov testov o nezávislosti pri použití správneho, resp. nesprávneho odhadu štandardnej odchýlky vzhľadom na spôsob generovania dát. Ďalej prostredníctvom simulácií porovnávame vlastnosti testov založených na pôvodnej miere γ a na jej zeta-transformácii, ktorá vykazuje priaznivejšie asymptotické vlastnosti. Kapitolu uzatvárame simuláciami ilustrujúcimi vplyv zlučovania

kategórií na pravdepodobnosť chyby I. druhu a silu testu o nezávislosti.

V štvrtej kapitole porovnávame správania testov nezávislosti založených na Spearmanovom korelačnom koeficiente ρ a Kendallovmu ρ_b pomocou ridit skóre. Pomocou simulačnej štúdie skúmame pravdepodobnosť chyby I. druhu a hodnotíme vplyv veľkosti výberu na dodržanie nominálnej hladiny významnosti. Simulačné štúdie sme doplnili dôkazom, ktorý bližšie vysvetľuje vlastnosti ridit skóre a tým aj správanie Kendallovho ρ_b pomocou ridit skóre.

V poslednej kapitole zhrnieme všetky výsledky dokopy a prevedieme poslednú sadu simulačných štúdií, kde porovnáme správanie jednotlivých testov založených na spomínaných mierach asociácie ordinálnych dát.

Implementácia použitá v tejto práci je dostupná na GitHub-e[1].

[1] https://github.com/veronikagasparovicova13/DP_Nezavislost_a_ordinalne_data

1 Popis dôležitých pojmov

V prvej kapitole sa budeme venovať definovaniu základných pojmov, na ktoré sa budeme v ďalšom texte odvolávať bez potreby ich opätovného vysvetľovania. Vychádzame pritom najmä z publikácie Alana Agrestiho [1].

Multinomické rozdelenie. Pri generovaní kontingenčných tabuliek vychádzame z multinomického rozdelenia. Uvažujme dve diskrétne náhodné premenné X a Y, ktorých hodnoty zisťujeme na n jedincoch a ktoré nadobúdajú hodnoty $X_1, ..., X_r$ a $Y_1, ..., Y_c$. Predpokladáme, že tieto hodnoty sú usporiadané (ordinálne), t. j. platí

$$X_1 < X_2 < \cdots < X_r, \qquad Y_1 < Y_2 < \cdots < Y_c.$$

Počty pozorovaní usporiadame do kontingenčnej tabuľky, kde bunky n_{ij} udávajú počty jedincov v kategóriách X_i a Y_j, pričom celkový počet pozorovaní je n.

	Y_1	Y_2	$\cdots$	Y_c	
X_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
X_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	n

Tomu zodpovedá pravdepodobnostná tabuľka, ktorej prvky π_{ij} predstavujú pravdepodobnosť, že náhodne vybraný jedinec patrí do bunky (i, j). Platí pritom, že súčet pravdepodobností je 1.

	Y_1	Y_2	$\cdots$	Y_c	
X_1	π_{11}	π_{12}	$\cdots$	π_{1c}	$\pi_{1\cdot}$
X_2	π_{21}	π_{22}	$\cdots$	π_{2c}	$\pi_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	π_{r1}	π_{r2}	$\cdots$	π_{rc}	$\pi_{r\cdot}$
	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	$\cdots$	$\pi_{\cdot c}$	1

Na základe zvolenej pravdepodobnostnej tabuľky následne generujeme kontingenčné tabuľky jedincov. Pre daný počet jedincov n tak získavame náhodné tabuľky, kde hodnoty

n_{ij} predstavujú počty jedincov a zodpovedajú multinomickému výberu s pravdepodobnosťami π_{ij}. Kontingenčnú tabuľku pritom chápeme ako vektor, ktorý vznikne postupným zoradením jej prvkov po riadkoch, t. j. v tvare $(n_{11}, n_{12}, ..., n_{1c}, n_{21}, ..., n_{rc})$. Pravdepodobnosť konkrétneho rozdelenia počtov v tabuľke je potom daná vzťahom

$$P(n_{11}, \ldots, n_{rc}) = \frac{n!}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \prod_{i=1}^{r} \prod_{j=1}^{c} \pi_{ij}^{n_{ij}},$$

pričom platí $\sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} = n$.

Konkordantnosť a diskordantnosť. Nech X a Y sú ordinálne premenné, kde indexy $i = 1, \ldots, r$ označujú riadky a $j = 1, \ldots, c$ označujú stĺpce kontingenčnej tabuľky uvedenej vyššie. Označme n_{ij} počty jedincov v jednotlivých bunkách tejto tabuľky, pričom celkový počet jedincov je $n = \sum_{i,j} n_{ij}$. Dvojicu buniek (i, j) a (k, l) nazývame *konkordantnou*, ak sa poradie v oboch premenných zhoduje, t. j. $i < k$ a $j < l$ alebo $i > k$ a $j > l$. Naopak, dvojica buniek je *diskordantná*, ak sú poradia opačné, teda $i < k$ a $j > l$ alebo $i > k$ a $j < l$. Počty konkordantných a diskordantných párov potom definujeme ako

$$C = \sum_{i<k} \sum_{j<\ell} n_{ij} n_{k\ell}, \qquad D = \sum_{i<k} \sum_{j>\ell} n_{ij} n_{k\ell}. \tag{1}$$

Štandardná chyba pre ordinálne miery asociácie. Vychádzajúc z pravdepodobnostnej tabuľky $\{\pi_{ij}\}$ zavedenej vyššie, môžeme odvodiť vzorec pre disperziu výberovej miery asociácie. Nech teoretická miera asociácie nadobúda tvar $\upsilon = \frac{\nu}{\delta}$, kde ν predstavuje čitateľ a δ menovateľ. Obe premenné ν a δ sú funkcie bunkových pravdepodobností $\{\pi_{ij}\}$. Podľa [1] definujeme

$$\phi_{ij} = \delta \left(\frac{\partial \nu}{\partial \pi_{ij}} \right) - \nu \left(\frac{\partial \delta}{\partial \pi_{ij}} \right). \tag{2}$$

Následne označme $\hat{\upsilon}$ odhad miery asociácie υ pre jeden multinomický výber. Tento odhad získame dosadením odhadnutých pravdepodobností $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$ namiesto π_{ij} do definičného vzťahu pre υ. Ďalej definujme marginálne pravdepodobnosti

$$\pi_{i\cdot} = \sum_{j=1}^{c} \pi_{ij}, \qquad \pi_{\cdot j} = \sum_{i=1}^{r} \pi_{ij},$$

ktoré odhadujeme ako

$$\hat{\pi}_{i\cdot} = \sum_{j=1}^{c} \hat{\pi}_{ij}, \qquad \hat{\pi}_{\cdot j} = \sum_{i=1}^{r} \hat{\pi}_{ij}.$$

Tieto odhady následne analogicky dosadzujeme do všetkých výrazov, v ktorých vystupujú marginálne pravdepodobnosti. Potom pre $n \to \infty$ konverguje $\sqrt{n}\,(\hat{v} - v)$ k normálnemu rozdeleniu so strednou hodnotou 0 a varianciou

$$\sigma^2 = \frac{\sum_i \sum_j \pi_{ij}\, \phi_{ij}^2 - \left(\sum_i \sum_j \pi_{ij}\, \phi_{ij}\right)^2}{\delta^4}. \tag{3}$$

Keďže pravdepodobnosti π_{ij} nie sú v praxi známe, v asymptotickej variancii σ^2 ich nahrádzame výberovými odhadmi $\frac{n_{ij}}{n}$. Následne vypočítame aj odhad disperzie $\hat{\sigma}^2$ dosadením premenných $\hat{\pi}_{ij}$, $\hat{\pi}_{i\cdot}$ a $\hat{\pi}_{\cdot j}$. Odhadnutá štandardná odchýlka odhadu $\hat{v}$ je potom daná vzťahom $SE = \frac{\hat{\sigma}}{\sqrt{n}}$.

V [4], [5] a [6] možno nájsť vzorce pre σ^2 v podstatne zložitejších tvaroch, ktoré nevyužívajú prístup založený na matici ϕ.

Štandardná chyba pri nezávislých multinomických výberoch. V tejto časti sa namiesto jednej multinomickej vzorky zameriavame na viacero navzájom nezávislých multinomických výberov. Tento prístup je vhodný najmä v situáciách, keď napríklad premennú X nemožno vnímať stochasticky, ale pre jej rôzne hodnoty $X_1, ..., X_r$ získavame náhodné výbery premennej Y. Na lepšie pochopenie rozdielu medzi jedným multinomickým výberom a viacerými nezávislými multinomickými výbermi uvažujme jednoduchý príklad, kde skúmame vzťah medzi pohlavím a preferovaným typom produktu.

Pri jednom multinomickom výbere uvažujeme, že všetkých n jedincov pochádza z jednej populácie a celú kontingenčnú tabuľku generujeme naraz ako jeden výber s pravdepodobnosťami π_{ij}. Naopak, pri viacerých nezávislých multinomických výberoch jednotlivé riadky tabuľky zodpovedajú rôznym skupinám (napr. mužom a ženám) a pre každú skupinu generujeme údaje samostatne. V takomto prípade najprv určujeme, aká časť všetkých jedincov pripadá na jednotlivé riadky tabuľky. Označme ω_i podiel jedincov patriacich do i-teho riadku, teda podiel danej skupiny na celkovom počte n, pričom platí $\sum_i \omega_i = 1$. Následne v rámci každého riadku generujeme z multinomického rozdelenia hodnoty samostatne podľa podmienených pravdepodobností $\pi_{j|i}$, pričom jednotlivé výbery sú navzájom nezávislé.

Podmienené pravdepodobnosti $\pi_{j|i}$ vyjadrujú pravdepodobnosť, že jedinec patrí do kategórie Y_j za predpokladu, že patrí do skupiny X_i. Pre každý riadok i tak platí $\sum_j \pi_{j|i} = 1$. Každý riadok tejto tabuľky teda predstavuje samostatné rozdelenie pravdepodobnosti

pre danú skupinu X_i. Podmienená pravdepodobnostná tabuľka potom vyzerá nasledovne:

	Y_1	Y_2	$\cdots$	Y_c
X_1	$\pi_{1\|1}$	$\pi_{2\|1}$	$\cdots$	$\pi_{c\|1}$
X_2	$\pi_{1\|2}$	$\pi_{2\|2}$	$\cdots$	$\pi_{c\|2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_r	$\pi_{1\|r}$	$\pi_{2\|r}$	$\cdots$	$\pi_{c\|r}$

V tejto práci sa zaoberáme kontingenčnými tabuľkami veľkosti 3×4 a ω_i odhadujeme ako $\hat{\omega}_i = \frac{n_{i\cdot}}{n}$ pre $\forall i$.

Pre teoretickú mieru asociácie tvaru $\upsilon = \frac{\nu}{\delta}$, vyjadrenú pomocou podmienených pravdepodobností $\pi_{j|i}$, označme podľa [1]

$$\phi_{j|i} = \delta \left(\frac{\partial \nu}{\partial \pi_{j|i}} \right) - \nu \left(\frac{\partial \delta}{\partial \pi_{j|i}} \right), \tag{4}$$

a asymptotickú varianciu miery asociácie υ vieme podľa [1] vypočítať ako

$$\sigma^2 \;=\; \frac{1}{\delta^4} \sum_i \frac{1}{\omega_i} \left(\sum_j \pi_{j|i}\, \phi_{j|i}^2 - \left(\sum_j \pi_{j|i}\, \phi_{j|i} \right)^2 \right). \tag{5}$$

V praxi nepoznáme pravdepodobnosti $\pi_{j|i}$, preto ich nahrádzame výberovými odhadmi $\frac{n_{ij}}{n_{i\cdot}}$. Odhad miery asociácie $\hat{\upsilon}$ tak získame dosadením týchto podielov do definície υ. Následne rovnako vypočítame aj odhad disperzie $\hat{\sigma}^2$. Odhadnutá štandardná odchýlka odhadu $\hat{\upsilon}$ je potom daná vzťahom $SE = \frac{\hat{\sigma}}{\sqrt{n}}$.

Testovanie nezávislosti využitím konkordantných a diskordantných párov. Pri štatistických testoch založených na konkordantných a diskordantných párach uvažujeme nulovú hypotézu, že ordinálne premenné X a Y sú nezávislé. Ak sú nezávislé, tak miera asociácie $\upsilon = \frac{\nu}{\delta} = 0$. Čitateľ pre miery asociácie Kendallovo τ_b a Goodman-Kruskalova γ sú rovnaké a rovnajú sa $\Pi_C - \Pi_D$, kde

$$\Pi_C = 2 \sum_{i<k} \sum_{j<\ell} \pi_{ij} \pi_{k\ell} \quad \text{a} \quad \Pi_D = 2 \sum_{i<k} \sum_{j>\ell} \pi_{ij} \pi_{k\ell} \tag{6}$$

označujú pravdepodobnosti konkordantných, resp. diskordantných párov pozorovaní.

Pravdepodobnosti Π_C a Π_D si vysvetlíme nasledovne. Náhodne vyberieme dvoch jedincov a pozrieme sa, do ktorých buniek kontingenčnej tabuľky patria. Podľa pozície buniek jednej voči druhej vieme určiť, či tvoria konkordantný alebo diskordantný pár. Ak prvý

jedinec patrí do bunky (i, j) a druhý do bunky (k, l), pravdepodobnosť takejto dvojice je daná súčinom $\pi_{ij}\pi_{k\ell}$, keďže ide o nezávislé výbery jedincov. Ak dvojica spĺňa podmienky pre konkordantný pár, teda $i < k, j < l$, pripočítame túto pravdepodobnosť k Π_C. Ak spĺňa podmienky pre diskordantný pár, teda $i < k, j > l$, pripočítame ju k Π_D. Koeficient 2 vo vzťahoch pre Π_C a Π_D zohľadňuje skutočnosť, že pri výbere dvoch jedincov nezáleží na poradí výberu. Každý pár buniek (i, j) a (k, ℓ), ktorý spĺňa podmienky konkordantnosti (resp. diskordantnosti), môže vzniknúť v dvoch usporiadaniach: buď prvý jedinec patrí do bunky (i, j) a druhý do (k, ℓ), alebo naopak. Keďže súčet vo vzťahoch (6) zahŕňa len prípady $i < k$, každú takúto dvojicu započítavame iba raz, a preto ju násobíme dvomi, aby sme zohľadnili obe možné poradia výberu.

Preto testovaná hypotéza pre tieto dve miery je

$$H_0 : \Pi_C = \Pi_D \qquad H_1 : \Pi_C \neq \Pi_D,$$

pretože z nezávislosti vyplýva $\Pi_C = \Pi_D$. Testová štatistika má tvar $z = \frac{\hat{v}}{SE}$, kde $\hat{v}$ je odhad miery asociácie v a SE je odhadnutá štandardná odchýlka danej miery.

Tvorenie pravdepodobnostných tabuliek. Pravdepodobnostné tabuľky tvoríme za platnosti nulovej hypotézy aj za jej neplatnosti. Tabuľky sú veľkosti 3×4, z ktorých následne generovaním z multinomického rozdelenia tvoríme kontingenčné tabuľky jedincov veľkosti 3×4. Z týchto tabuliek jedincov ďalej vypočítavame výberové miery asociácie a vyhodnocujeme testové štatistiky. Pravdepodobnostné tabuľky za platnosti nulovej hypotézy tvoríme dvoma spôsobmi:

1. Rovnomerný spôsob: Pravdepodobnostnú tabuľku konštruujeme tak, že všetkým bunkám priradíme rovnakú pravdepodobnosť. V prípade tabuľky rozmeru 3×4 má každá bunka pravdepodobnosť $\pi_{ij} = \frac{1}{12}$. Takto definovaná tabuľka spĺňa nulovú hypotézu nezávislosti, keďže riadkové aj stĺpcové marginálne pravdepodobnosti sú rovnomerne rozdelené.

2. Nerovnomerný spôsob: Nech $\boldsymbol{p} = (0.6, 0.3, 0.1)$ označuje vektor riadkových pravdepodobností a $\boldsymbol{q} = (0.16, 0.28, 0.3, 0.26)$ vektor stĺpcových pravdepodobností. Pravdepodobnosti buniek potom definujeme ako

$$\pi_{ij} = p_i q_j,$$

čím je zabezpečená platnosť nulovej hypotézy nezávislosti.

Vektor $\boldsymbol{p}$ bol zvolený zámerne tak, aby jeho zložky boli násobkami $\frac{1}{30}$, t. j. aby platilo $p_i \cdot n \in \mathbb{N}$ pre uvažované veľkosti výberu $n = 30$ a $n = 300$.

Toto obmedzenie súvisí so spôsobom generovania kontingenčných tabuliek v simuláciách. Okrem generovania z jedného multinomického rozdelenia uvažujeme aj generovanie pomocou troch podmienených multinomických rozdelení, kde v každom riadku generujeme počty jedincov samostatne s pevne daným počtom pozorovaní $n_{i\cdot} = np_i$.

Aby boli oba prístupy porovnateľné, je potrebné, aby hodnoty $n_{i\cdot}$ boli celé čísla. V opačnom prípade by bolo nutné tieto hodnoty zaokrúhľovať, čo by viedlo k tomu, že tabuľky generované z jedného multinomického rozdelenia by neboli plne konzistentné s tabuľkami generovanými z podmienených multinomických rozdelení.

Z tohto dôvodu volíme $\boldsymbol{p}$ tak, aby boli riadkové marginálne pravdepodobnosti kompatibilné s veľkosťou výberu a aby sa predišlo dodatočným aproximačným chybám spôsobeným zaokrúhľovaním.

Za neplatnosti nulovej hypotézy konštruujeme pravdepodobnostné tabuľky tak, aby boli zachované riadkové pravdepodobnosti dané vektorom $\boldsymbol{p} = (0.6, 0.3, 0.1)$, teda aby súčty pravdepodobností v jednotlivých riadkoch zodpovedali týmto hodnotám. V rámci jednotlivých riadkov však pravdepodobnosti rozdeľujeme nerovnomerne medzi stĺpce tak, aby nebolo možné zapísať pravdepodobnosti buniek v tvare súčinu $\pi_{ij} = p_i q_j$. Týmto spôsobom dochádza k porušeniu podmienky nezávislosti medzi premennými X a Y.

Simulácie. V rámci simulačnej štúdie za platnosti nulovej hypotézy uvažujeme dva typy pravdepodobnostných kontingenčných tabuliek: tabuľky s rovnomerne rozdelenými pravdepodobnosťami buniek a tabuľky s nerovnomernými pravdepodobnosťami, konštruované za platnosti nulovej hypotézy nezávislosti. Pre každý typ tabuľky generujeme dáta dvoma rôznymi spôsobmi, a to pomocou jedného multinomického rozdelenia a pomocou troch nezávislých multinomických rozdelení, zodpovedajúcich podmienenej konštrukcii tabuľky.

Pre uvažované miery asociácie sú k dispozícii dva spôsoby odhadu asymptotickej variancie, zodpovedajúce jednotlivým spôsobom generovania dát. V simuláciách preto kombinujeme odhady mier asociácie s oboma typmi variancií, pričom zámerne používame aj varianciu, ktorá nezodpovedá skutočnému spôsobu generovania dát.

Na základe takto skonštruovaných testových štatistík realizujeme testy nezávislosti a sledujeme empirickú "zamietavosť" testu, konkrétne výskyt chyby I. druhu.

Simulácie sme realizovali aj v situáciách, keď nulová hypotéza neplatí. V takto vytvorených tabuľkách dochádza k porušeniu podmienky nezávislosti medzi premennými. Tento prístup nám umožňuje analyzovať správanie testu za prítomnosti závislosti a hodnotiť silu testu.

2 Kendallovo τ_b

V tejto kapitole sa zameriame na odvodenie asymptotickej variancie Kendallovho $\hat{\tau}_b$, ktorá je nevyhnutná pre konštrukciu testu nezávislosti založeného na tejto miere. Odvodený tvar asymptotickej variancie závisí od spôsobu, akým je zostavená pravdepodobnostná kontingenčná tabuľka, z ktorej sú následne generované výbery pomocou multinomického rozdelenia. Z tohto dôvodu sa odvodenie člení na dve časti: odvodenie variancie pri jednom multinomickom výbere a odvodenie pri viacerých multinomických výberoch. V praktickej časti práce sa následne porovnávajú simulačné štúdie založené na jednom multinomickom výbere a na troch nezávislých multinomických výberoch.

Po odvodení variancie pre jeden multinomický výber sme identifikovali chybu v jednom zo vzorcov uvedených v [1]. Uvádzame aj komunikáciu s autorom knihy a analyzujeme vplyv tejto chyby na test nezávislosti.

Následne sa venujeme simulačným štúdiám, ktoré dopĺňame dôkazom o rovnosti asymptotických variancií za platnosti nezávislosti, čo nám umožňuje lepšie interpretovať výsledky simulácií za platnosti H_0. Kapitolu uzatvárame simuláciami zameranými na zlučovanie kategórií.

2.1 Odvodenie variancie pre Kendallovo τ_b pre jeden multinomický výber

Ako prvé uvedieme definíciu τ_b a príslušné veličiny. Potom odvodíme potrebné premenné, ktoré dosadíme do vzorca pre asymptotickú varianciu (3).

Na základe pojmov zavedených v predchádzajúcej kapitole, najmä pravdepodobností konkordantných a diskordantných párov Π_C a Π_D (6), môžeme definovať Kendallovo τ_b ako mieru asociácie medzi dvoma ordinálnymi premennými [8].

Definícia 1 (Kendallovo τ_b)**.** Kendallovo τ_b je definované vzťahom

$$\tau_b \;=\; \frac{\Pi_C - \Pi_D}{\sqrt{\left(1 - \sum_i \pi_{i\cdot}^2\right)\left(1 - \sum_j \pi_{\cdot j}^2\right)}},$$

kde $\pi_{i\cdot} = \sum_j \pi_{ij}$ a $\pi_{\cdot j} = \sum_i \pi_{ij}$ sú marginálne pravdepodobnosti. Odhad Kendallovho τ_b

používaný v implementácií má tvar

$$\hat{\tau}_b \;=\; \frac{\hat{\Pi}_C - \hat{\Pi}_D}{\sqrt{\Big(1-\sum_i \hat{\pi}_{i\cdot}^2\Big)\Big(1-\sum_j \hat{\pi}_{\cdot j}^2\Big)}},$$

kde $\hat{\Pi}_C = \frac{C}{\binom{n}{2}}$, $\hat{\Pi}_D = \frac{D}{\binom{n}{2}}$ a $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$. Premenné C a D označujú počty konkordantných a diskordantných párov (1).

Označme

$$\nu := \Pi_C - \Pi_D, \quad \delta := \sqrt{\Big(1-\sum_i \pi_{i\cdot}^2\Big)\Big(1-\sum_j \pi_{\cdot j}^2\Big)}. \tag{7}$$

Ako prvé potrebujeme odvodiť premenné ϕ_{ij} podľa vzorca (2). Preto potrebujeme derivácie premenných ν a δ uvedené v (7). Začneme s označením istých premenných pre neskoršiu prehľadnosť.

Kvadrantové súčty vzhľadom na bunku (i,j). Pre ľubovoľnú bunku (i,j) definujme štyri kvadrantové sumy:

$$\Pi_{I;ij} := \sum_{k>i}\sum_{\ell>j} \pi_{k\ell}, \qquad \Pi_{II;ij} := \sum_{k<i}\sum_{\ell>j} \pi_{k\ell},$$
$$\Pi_{III;ij} := \sum_{k<i}\sum_{\ell<j} \pi_{k\ell}, \qquad \Pi_{IV;ij} := \sum_{k>i}\sum_{\ell<j} \pi_{k\ell}.$$

Potom

$$\begin{aligned} R_{ij}^C &:= \Pi_{I;ij} + \Pi_{III;ij} && \text{je súhlasný kvadrantový súčet,} \\ R_{ij}^D &:= \Pi_{II;ij} + \Pi_{IV;ij} && \text{je nesúhlasný kvadrantový súčet.} \end{aligned} \tag{8}$$

Pre každú bunku (i,j) tieto súčty predstavujú pravdepodobnosti toho, že náhodne vybraný jedinec z populácie sa nachádza v bunke, ktorá je vzhľadom na bunku (i,j) konkordantná, resp. diskordantná.

Derivácia čitateľa ν. Začneme s deriváciou čitateľa ν uvedeného (7). Pripomenieme si definíciu pravdepodobnosti konkordantných párov Π_C:

$$\Pi_C = 2\sum_{i<k}\sum_{j<\ell} \pi_{ij}\pi_{k\ell}.$$

Derivácia podľa π_{ab}:

$$\frac{\partial \Pi_C}{\partial \pi_{ab}} \;=\; 2\sum_{i<k}\sum_{j<\ell} \frac{\partial}{\partial \pi_{ab}}\Big(\pi_{ij}\,\pi_{k\ell}\Big).$$

V danej dvojitej sume sa π_{ab} môže objaviť *iba* v dvoch rolách:

(A) ako prvý činiteľ: $(i,j)=(a,b)$. Indexové podmienky $i<k$, $j<\ell$ sa menia na $a<k$ a $b<\ell$. Pre tieto členy platí

$$\frac{\partial}{\partial\pi_{ab}}\Big(\pi_{ab}\,\pi_{k\ell}\Big) \;=\; \pi_{k\ell},$$

a príspevok do derivácie je

$$2\sum_{a<k}\sum_{b<\ell}\pi_{k\ell} \;=\; 2\,\Pi_{I;ab}.$$

(B) ako druhý činiteľ: $(k,\ell)=(a,b)$. Z podmienok $i<k$, $j<\ell$ dostaneme $i<a$ a $j<b$. Pre tieto členy platí

$$\frac{\partial}{\partial\pi_{ab}}\Big(\pi_{ij}\,\pi_{ab}\Big) \;=\; \pi_{ij},$$

a príspevok do derivácie je

$$2\sum_{i<a}\sum_{j<b}\pi_{ij} \;=\; 2\,\Pi_{III;ab}.$$

Množiny indexov v (A) a (B) sú disjunktné, pretože v rámci podmienok $i<k$, $j<\ell$ nemôže nastať naraz $(i,j)=(a,b)$ a zároveň $(k,\ell)=(a,b)$. Preto

$$\frac{\partial\Pi_C}{\partial\pi_{ab}} \;=\; 2\sum_{a<k}\sum_{b<\ell}\pi_{k\ell} \;+\; 2\sum_{i<a}\sum_{j<b}\pi_{ij} \;=\; 2\Big(\Pi_{I;ab}+\Pi_{III;ab}\Big)=2R^C_{ab}.$$

Analogicky vyjdime z definície pravdepodobnosti diskordantných párov, kde

$$\Pi_D = 2\sum_{i<k}\sum_{j>\ell}\pi_{ij}\pi_{k\ell}.$$

Derivácia podľa π_{ab}:

$$\frac{\partial\Pi_D}{\partial\pi_{ab}} \;=\; 2\sum_{i<k}\sum_{j>\ell}\frac{\partial}{\partial\pi_{ab}}\Big(\pi_{ij}\,\pi_{k\ell}\Big).$$

V danej dvojitej sume sa π_{ab} môže opäť objaviť iba v dvoch rolách:

(C) ako prvý činiteľ: $(i,j)=(a,b)$. Indexové podmienky $i<k$, $j>\ell$ sa menia na $a<k$ a $b>\ell$. Pre tieto členy platí

$$\frac{\partial}{\partial\pi_{ab}}\Big(\pi_{ab}\,\pi_{k\ell}\Big) \;=\; \pi_{k\ell},$$

a príspevok do derivácie je

$$2\sum_{a<k}\sum_{b>\ell}\pi_{k\ell} \;=\; 2\,\Pi_{IV;ab}.$$

(D) ako druhý činiteľ: $(k,\ell) = (a,b)$. Z podmienok $i<k$, $j>\ell$ dostaneme $i<a$ a $j>b$.

Pre tieto členy platí

$$\frac{\partial}{\partial \pi_{ab}}\Big(\pi_{ij}\,\pi_{ab}\Big) \;=\; \pi_{ij},$$

a príspevok do derivácie je

$$2\sum_{i<a}\sum_{j>b}\pi_{ij} \;=\; 2\,\Pi_{II;ab}.$$

Analogicky, keďže množiny indexov sú disjunktné, tak platí

$$\frac{\partial \Pi_D}{\partial \pi_{ab}} \;=\; 2\sum_{k>a}\sum_{\ell<b}\pi_{k\ell} \;+\; 2\sum_{i<a}\sum_{j>b}\pi_{ij} \;=\; 2\Big(\Pi_{II;ab}+\Pi_{IV;ab}\Big) = 2R^D_{ab}.$$

Spojením predchádzajúcich výsledkov dostávame deriváciu čitateľa

$$\nu'_{ab} := \frac{\partial \nu}{\partial \pi_{ab}} = \frac{\partial \Pi_C}{\partial \pi_{ab}} - \frac{\partial \Pi_D}{\partial \pi_{ab}} = 2\Big(R^C_{ab} - R^D_{ab}\Big). \tag{9}$$

Derivácia menovateľa δ. Definujme

$$\delta \;=\; \sqrt{\Big(1-\sum_i \pi^2_{i\cdot}\Big)\Big(1-\sum_j \pi^2_{\cdot j}\Big)} \;=\; \sqrt{A\,B}, \qquad A := 1-\sum_i \pi^2_{i\cdot},\;\; B := 1-\sum_j \pi^2_{\cdot j}.$$

Potom platí

$$\frac{\partial \delta}{\partial \pi_{ab}} = \frac{1}{2\sqrt{AB}}\left(B\,\frac{\partial A}{\partial \pi_{ab}} + A\,\frac{\partial B}{\partial \pi_{ab}}\right) = \frac{1}{2\delta}\left(B\,\frac{\partial A}{\partial \pi_{ab}} + A\,\frac{\partial B}{\partial \pi_{ab}}\right).$$

Najprv zderivujeme premenné A a B. Keďže $\pi_{i\cdot} = \sum_j \pi_{ij}$ a $\pi_{\cdot j} = \sum_i \pi_{ij}$, potom

$$\frac{\partial \pi_{i\cdot}}{\partial \pi_{ab}} = \begin{cases} 1, & i = a, \\ 0, & i \neq a, \end{cases} \qquad \frac{\partial \pi_{\cdot j}}{\partial \pi_{ab}} = \begin{cases} 1, & j = b, \\ 0, & j \neq b. \end{cases}$$

Teda

$$\frac{\partial A}{\partial \pi_{ab}} = -\,\frac{\partial}{\partial \pi_{ab}}\sum_i \pi^2_{i\cdot} = -\sum_i 2\pi_{i\cdot}\frac{\partial \pi_{i\cdot}}{\partial \pi_{ab}} = -\,2\pi_{a\cdot}, \qquad \frac{\partial B}{\partial \pi_{ab}} = -\,\frac{\partial}{\partial \pi_{ab}}\sum_j \pi^2_{\cdot j} = -\,2\pi_{\cdot b}.$$

Dosadením $\frac{\partial A}{\partial \pi_{ab}}$ a $\frac{\partial B}{\partial \pi_{ab}}$ dostávame

$$\begin{aligned}
\frac{\partial \delta}{\partial \pi_{ab}} &= \frac{1}{2\delta}\Big(B\cdot(-2\pi_{a\cdot}) + A\cdot(-2\pi_{\cdot b})\Big) \\
&= -\,\frac{1}{\delta}\Big(B\,\pi_{a\cdot} + A\,\pi_{\cdot b}\Big) \\
&= -\Big(\frac{B}{\delta}\pi_{a\cdot} + \frac{A}{\delta}\pi_{\cdot b}\Big) \\
&= -\frac{1}{\delta}\Big[\pi_{a\cdot}\Big(1-\sum_j \pi^2_{\cdot j}\Big) + \pi_{\cdot b}\Big(1-\sum_i \pi^2_{i\cdot}\Big)\Big] =: \widetilde{\delta'_{ab}}.
\end{aligned}$$

Posledný tvar derivácie menovateľa, označené $\widetilde{\delta'_{ab}}$, sa ukáže ako vhodný pri dosadení do výrazu (2) pre ϕ_{ij}, keďže umožňuje následné porovnanie s formuláciou uvedenou v práci [1], v ktorej sa objavuje odlišný tvar tejto veličiny. Avšak pre vyjadrenie variancie je vhodnejšie urobiť v tomto bode menší technický krok a upraviť predposledný tvar $\frac{\partial\delta}{\partial\pi_{ab}}$ tak, aby sme si ušetrili pridlhé numerické úpravy neskôr, keď budeme dosádzať do variancie. Vieme, že $\delta^2 = AB$. Z toho vyplýva, že

$$\frac{\delta}{A} = \frac{B}{\delta} \qquad \text{a} \qquad \frac{\delta}{B} = \frac{A}{\delta}.$$

Potom vieme predposledný tvar $\frac{\partial\delta}{\partial\pi_{ab}}$ napísať ako

$$\delta'_{ab} := -\Big(\frac{\delta}{A}\pi_{a\cdot} + \frac{\delta}{B}\pi_{\cdot b}\Big) = -\delta\Big(\frac{\pi_{a\cdot}}{A} + \frac{\pi_{\cdot b}}{B}\Big).$$

Po dosadení premenných A a B dostávame

$$\delta'_{ab} = -\delta\left(\frac{\pi_{a\cdot}}{1-\sum_i \pi_{i\cdot}^2} + \frac{\pi_{\cdot b}}{1-\sum_j \pi_{\cdot j}^2}\right).$$

Matica ϕ. Z predošlých krokov sme dostali

$$\begin{aligned}
\nu'_{ab} &= 2\Big(R^C_{ab} - R^D_{ab}\Big),\\
\delta'_{ab} &= -\delta\left(\frac{\pi_{a\cdot}}{1-\sum_i \pi_{i\cdot}^2} + \frac{\pi_{\cdot b}}{1-\sum_j \pi_{\cdot j}^2}\right),\\
\widetilde{\delta'_{ab}} &= -\frac{1}{\delta}\Big[\pi_{a\cdot}\Big(1-\sum_j \pi_{\cdot j}^2\Big) + \pi_{\cdot b}\Big(1-\sum_i \pi_{i\cdot}^2\Big)\Big].
\end{aligned}$$

kde δ'_{ab} a $\widetilde{\delta'_{ab}}$ sú dve odlišné vyjadrenia tej istej hodnoty.
Po dosadení ν'_{ab} a δ'_{ab} do ϕ_{ab}, čiže do (2), dostávame

$$\phi_{ab} = \nu'_{ab}\,\delta - \nu\,\delta'_{ab} = \delta\Big[2\Big(R^C_{ab} - R^D_{ab}\Big) + \nu\left(\frac{\pi_{a\cdot}}{1-\sum_i \pi_{i\cdot}^2} + \frac{\pi_{\cdot b}}{1-\sum_j \pi_{\cdot j}^2}\right)\Big]. \tag{10}$$

Označme $T_{ab} := 2\Big(R^C_{ab} - R^D_{ab}\Big) + \nu\left(\frac{\pi_{a\cdot}}{1-\sum_i \pi_{i\cdot}^2} + \frac{\pi_{\cdot b}}{1-\sum_j \pi_{\cdot j}^2}\right)$, ktoré neskôr využijeme pri zápise variancie.
Po dosadení ν'_{ab} a $\widetilde{\delta'_{ab}}$ do ϕ_{ab}, čiže do (2), dostávame iný tvar pre ϕ_{ab}, ktorý označíme $\widetilde{\phi_{ab}}$:

$$\widetilde{\phi_{ab}} = \nu'_{ab}\,\delta - \nu\,\widetilde{\delta'_{ab}} = 2\delta\Big(R^C_{ab} - R^D_{ab}\Big) + \frac{\nu}{\delta}\Big[\pi_{a\cdot}\Big(1-\sum_j \pi_{\cdot j}^2\Big) + \pi_{\cdot j}\Big(1-\sum_i \pi_{i\cdot}^2\Big)\Big]. \tag{11}$$

Tvar $\widetilde{\phi_{ab}}$ využijeme neskôr pri porovnávaní s Agrestiho tvarom premennej ϕ_{ab} v knihe [1].

Variancia $\hat{\tau}_b$. Vzorec pre asymptotickú varianciu (3) sa dá podľa článku [7] prepísať na

$$\sigma^2 = \frac{1}{\delta^4} \sum_{ij} \pi_{ij} (\phi_{ij} - \bar{\phi})^2, \tag{12}$$

kde $\bar{\phi} = \sum_{ij} \pi_{ij} \phi_{ij}$. Najprv si vyjadríme, čo predstavuje premenná $\bar{\phi}$. Dosadíme

$$\bar{\phi} = \sum_{ij} \pi_{ij} \phi_{ij} = \delta \Big[2 \sum_{ij} \pi_{ij} (R^C_{ij} - R^D_{ij}) + \nu \Big(\frac{1}{1 - \sum_i \pi^2_{i\cdot}} \sum_{ij} \pi_{ij} \pi_{i\cdot} + \frac{1}{1 - \sum_j \pi^2_{\cdot j}} \sum_{ij} \pi_{ij} \pi_{\cdot j} \Big) \Big].$$

Pozrieme sa na jednotlivé členy samostatne. Člen $\sum_{ij} \pi_{ij} R^C_{ij}$ vieme upraviť nasledovne.

$$\begin{aligned} \sum_{ij} \pi_{ij} R^C_{ij} &= \sum_{ij} \pi_{ij} \big(\Pi_{I;ij} + \Pi_{III;ij} \big) = \sum_{ij} \pi_{ij} \left(\sum_{k>i} \sum_{\ell>j} \pi_{k\ell} + \sum_{k<i} \sum_{\ell<j} \pi_{k\ell} \right) \\ &= \sum_{i<k} \sum_{j<\ell} \pi_{ij}\, \pi_{k\ell} \;+\; \sum_{k<i} \sum_{\ell<j} \pi_{ij}\, \pi_{k\ell}. \end{aligned}$$

Tieto dva členy sú rovnaké, ide len o premenovanie názvov indexov v druhej sume. Preto podľa (6)

$$\sum_{ij} \pi_{ij} R^C_{ij} = 2 \sum_{i<k} \sum_{j<\ell} \pi_{ij}\, \pi_{k\ell} = \Pi^C. \tag{13}$$

Analogicky

$$\sum_{ij} \pi_{ij} R^D_{ij} = \Pi^D. \tag{14}$$

Ďalej, člen $\sum_{ij} \pi_{ij} \pi_{i\cdot}$ upravíme ako

$$\sum_{ij} \pi_{ij} \pi_{i\cdot} = \sum_i \pi_{i\cdot} \sum_j \pi_{ij} = \sum_i \pi_{i\cdot} \pi_{i\cdot} = \sum_i \pi^2_{i\cdot}. \tag{15}$$

Analogicky dostávame

$$\sum_{ij} \pi_{ij} \pi_{\cdot j} = \sum_j \pi^2_{\cdot j}. \tag{16}$$

Teraz vieme $\bar{\phi}$ podľa (13), (14), (15), (16) a vzťahu $\nu = \Pi_C - \Pi_D$ upraviť nasledovne:

$$\bar{\phi} = \delta \Big[2(\Pi^C - \Pi^D) + \nu \Big(\frac{\sum_i \pi^2_{i\cdot}}{1 - \sum_i \pi^2_{i\cdot}} + \frac{\sum_j \pi^2_{\cdot j}}{1 - \sum_j \pi^2_{\cdot j}} \Big) \Big] = \delta \nu \Big(2 + \frac{\sum_i \pi^2_{i\cdot}}{1 - \sum_i \pi^2_{i\cdot}} + \frac{\sum_j \pi^2_{\cdot j}}{1 - \sum_j \pi^2_{\cdot j}} \Big).$$

Označme $K := \nu \Big(2 + \frac{\sum_i \pi^2_{i\cdot}}{1 - \sum_i \pi^2_{i\cdot}} + \frac{\sum_j \pi^2_{\cdot j}}{1 - \sum_j \pi^2_{\cdot j}} \Big)$.
Nakoniec vieme vzorec (12) upraviť ako

$$\sigma^2 = \frac{1}{\delta^4} \sum_{ij} \pi_{ij} (\delta T_{ij} - \delta K)^2 = \frac{1}{\delta^2} \sum_{ij} \pi_{ij} (T_{ij} - K)^2, \tag{17}$$

kde $T_{ab} = 2\big(R^C_{ab} - R^D_{ab} \big) + \nu \left(\frac{\pi_{a\cdot}}{1 - \sum_i \pi^2_{i\cdot}} + \frac{\pi_{\cdot b}}{1 - \sum_j \pi^2_{\cdot j}} \right)$ a $K = \nu \Big(2 + \frac{\sum_i \pi^2_{i\cdot}}{1 - \sum_i \pi^2_{i\cdot}} + \frac{\sum_j \pi^2_{\cdot j}}{1 - \sum_j \pi^2_{\cdot j}} \Big)$.

2.2 Odvodenie variancie pre Kendallovo τ_b pre nezávislé multinomické vzorky

Budeme vychádzať z parametrizácie pomocou podmienenej pravdepodobnosti:

$$\pi_{ij} = \pi_{i\cdot}\,\pi_{j|i} = \omega_i\,\pi_{j|i}, \qquad \text{kde} \qquad \omega_i := \pi_{i\cdot}.$$

Potom

$$A = 1 - \sum_i \pi_{i\cdot}^2 \;=\; 1 - \sum_i \omega_i^2, \qquad B = 1 - \sum_j \pi_{\cdot j}^2.$$

Pripomenieme, že Kendallovo τ_b definujeme ako

$$\tau_b = \frac{\nu}{\delta}, \qquad \nu = \Pi_C - \Pi_D \qquad \delta = \sqrt{AB}.$$

Derivácia čitateľa ν podľa $\pi_{b|a}$. Začneme s pripomenutím definície

$$\Pi_C \;=\; 2\sum_{i<k}\sum_{j<\ell} \pi_{ij}\,\pi_{k\ell}, \qquad \Pi_D \;=\; 2\sum_{i<k}\sum_{j>\ell} \pi_{ij}\,\pi_{k\ell}.$$

Platí $\pi_{ij} = \omega_i\,\pi_{j|i}$, kde $\omega_i = \pi_{i\cdot}$ sú *pevné*, teda nezávislé od $\pi_{b|a}$. Preto

$$\frac{\partial \pi_{ij}}{\partial \pi_{b|a}} = \begin{cases} \omega_a, & \text{ak } (i,j) = (a,b), \\ 0, & \text{inak.} \end{cases}$$

Derivácia $\partial\Pi_C/\partial\pi_{b|a}$. Derivujeme z definície:

$$\frac{\partial \Pi_C}{\partial \pi_{b|a}} = 2\sum_{i<k}\sum_{j<\ell} \frac{\partial}{\partial \pi_{b|a}}\Big(\pi_{ij}\,\pi_{k\ell}\Big) = 2\sum_{i<k}\sum_{j<\ell}\Big(\underbrace{\frac{\partial \pi_{ij}}{\partial \pi_{b|a}}}_{\text{(A)}}\,\pi_{k\ell} \;+\; \pi_{ij}\,\underbrace{\frac{\partial \pi_{k\ell}}{\partial \pi_{b|a}}}_{\text{(B)}}\Big).$$

V dvojnásobnej sume sa $\pi_{b|a}$ môže objaviť iba v dvoch rolách:

(A) ako prvý činiteľ: $(i,j) = (a,b)$. Z podmienok $i<k,\ j<\ell$ dostaneme $a<k,\ b<\ell$. Potom

$$\frac{\partial \pi_{ij}}{\partial \pi_{b|a}} = \omega_a, \qquad \Rightarrow \qquad \text{príspevok (A)} \;=\; 2\sum_{a<k}\sum_{b<\ell} \omega_a\,\pi_{k\ell} \;=\; 2\,\omega_a\,\Pi_{I;ab}.$$

(B) ako druhý činiteľ: $(k,\ell) = (a,b)$. Z podmienok $i<k,\ j<\ell$ dostaneme $i<a,\ j<b$. Potom

$$\frac{\partial \pi_{k\ell}}{\partial \pi_{b|a}} = \omega_a, \qquad \Rightarrow \qquad \text{príspevok (B)} \;=\; 2\sum_{i<a}\sum_{j<b} \pi_{ij}\,\omega_a \;=\; 2\,\omega_a\,\Pi_{III;ab}.$$

Sú to disjunktné indexové množiny, preto

$$\frac{\partial \Pi_C}{\partial \pi_{ab}} = 2\,\omega_a\Big(\Pi_{I;ab} + \Pi_{III;ab}\Big) = 2\,\omega_a\,R^C_{ab},$$

kde $R^C_{ab} := \Pi_{I;ab} + \Pi_{III;ab}$ je súhlasný kvadrantový súčet.

Derivácia $\partial\Pi_D/\partial\pi_{b|a}$. Postup je analogický, iba s indexovými obmedzeniami pre nesúhlasné páry:

$$\frac{\partial \Pi_D}{\partial \pi_{b|a}} = 2\sum_{i<k}\sum_{j>\ell}\frac{\partial}{\partial \pi_{b|a}}\Big(\pi_{ij}\,\pi_{k\ell}\Big).$$

(A) ako prvý činiteľ: $(i,j) = (a,b) \Rightarrow a < k,\ b > \ell$,

$$\text{príspevok (A)} = 2\sum_{a<k}\sum_{b>\ell}\omega_a\,\pi_{k\ell} = 2\,\omega_a\,\Pi_{IV;ab}.$$

(B) ako druhý činiteľ: $(k,\ell) = (a,b) \Rightarrow i < a,\ j > b$,

$$\text{príspevok (B)} = 2\sum_{i<a}\sum_{j>b}\pi_{ij}\,\omega_a = 2\,\omega_a\,\Pi_{II;ab}.$$

Potom

$$\frac{\partial \Pi_D}{\partial \pi_{b|a}} = 2\,\omega_a\Big(\Pi_{II;ab} + \Pi_{IV;ab}\Big) = 2\,\omega_a\,R^D_{ab},$$

kde $R^D_{ab} := \Pi_{II;ab} + \Pi_{IV;ab}$ je nesúhlasný kvadrantový súčet.

Derivácia čitateľa je potom

$$\frac{\partial \nu}{\partial \pi_{b|a}} = \frac{\partial \Pi_C}{\partial \pi_{b|a}} - \frac{\partial \Pi_D}{\partial \pi_{b|a}} = 2\,\omega_a\Big(R^C_{ab} - R^D_{ab}\Big). \tag{18}$$

Derivácia menovateľa δ podľa $\pi_{b|a}$. Pripomeňme si

$$\delta = \sqrt{AB}, \qquad A = 1 - \sum_i \pi^2_{i\cdot} = 1 - \sum_i \omega_i^2, \qquad B = 1 - \sum_j \pi^2_{\cdot j}.$$

Premenné ω_i sú pevné a A je konštantné vzhľadom na $\pi_{b|a}$. Od $\pi_{b|a}$ závisí iba B, konkrétne v ňom vystupujúce $\pi_{\cdot b}$:

$$\pi_{\cdot b} = \sum_i \pi_{ib} = \sum_i \omega_i \pi_{b|i}.$$

Pre fixné (a,b) dostávame

$$\frac{\partial \pi_{\cdot b}}{\partial \pi_{b|a}} = \omega_a.$$

Odtiaľ plynie

$$\frac{\partial B}{\partial \pi_{b|a}} = -2\,\pi_{\cdot b}\,\frac{\partial \pi_{\cdot b}}{\partial \pi_{b|a}} = -2\,\pi_{\cdot b}\,\omega_a.$$

Deriváciou menovateľa δ teda dostávame

$$\frac{\partial \delta}{\partial \pi_{b|a}} = \frac{1}{2\delta}\Big(A \cdot \frac{\partial B}{\partial \pi_{b|a}}\Big) = -\frac{A\,\omega_a\,\pi_{\cdot b}}{\delta}. \tag{19}$$

Výpočet $\phi_{b|a}$. Po dosadení derivácií premenných $\nu_{b|a}$ a $\delta_{b|a}$ do vzorca pre $\phi_{b|a}$ uvedeného v (4) dostávame

$$\phi_{b|a} = \delta \cdot 2\omega_a(R^C_{ab} - R^D_{ab}) - \nu\left(-\frac{A\,\omega_a\,\pi_{\cdot b}}{\delta}\right) \tag{20}$$

$$= 2\delta\,\omega_a\Big(R^C_{ab} - R^D_{ab}\Big) + \frac{\nu}{\delta} A\omega_a\pi_{\cdot b}. \tag{21}$$

Pre stručnosť si označme

$$T_{b|a} := \frac{\phi_{b|a}}{\omega_a} = 2\delta(R^C_{ab} - R^D_{ab}) + \frac{\nu}{\delta}\,A\,\pi_{\cdot b}.$$

Variancia $\hat{\tau}_b$. Dosadíme do vzorca (5) a variancia nadobúda tvar

$$\begin{aligned}
\sigma^2_{\hat{\tau}_b} &= \frac{1}{\delta^4}\sum_i \frac{1}{\omega_i}\left(\sum_j \pi_{j|i}\,\phi^2_{j|i} - \left(\sum_j \pi_{j|i}\,\phi_{j|i}\right)^2\right) \\
&= \frac{1}{\delta^4}\sum_i \frac{1}{\omega_i}\left(\sum_j \pi_{j|i}\,\omega_i^2\frac{\phi^2_{j|i}}{\omega_i^2} - \left(\sum_j \pi_{j|i}\,\omega_i\frac{\phi_{j|i}}{\omega_i}\right)^2\right) \\
&= \frac{1}{\delta^4}\sum_i \frac{1}{\omega_i}\omega_i^2\left(\sum_j \pi_{j|i}\frac{\phi^2_{j|i}}{\omega_i^2} - \left(\sum_j \pi_{j|i}\frac{\phi_{j|i}}{\omega_i}\right)^2\right) \\
&= \frac{1}{\delta^4}\sum_i \omega_i\left(\sum_j \pi_{j|i}\,T^2_{j|i} - \left(\sum_j \pi_{j|i}\,T_{j|i}\right)^2\right).
\end{aligned} \tag{22}$$

2.3 Dôkaz o rovnosti variancií

Ukážeme, že za platnosti H_0 o nezávislosti sa σ^2 pre jedno multinomické rozdelenie, označíme σ_1^2, a σ^2 pre viacero multinomických rozdelení, označíme σ_M^2, pre koeficient τ_b rovnajú.

Pripomenieme si značenie $\omega_i = P(X = i) = \pi_{i.}$, ktoré sme predstavili v odvodení variancie pre Kendallovo τ_b pre nezávislé multinomické vzorky. Pre podmienené pravdepodobnosti potom platí

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i.}} = \frac{\pi_{ij}}{\omega_i}, \tag{23}$$

odkiaľ dostávame vzťah

$$\omega_i \pi_{j|i} = \pi_{ij}. \tag{24}$$

Za platnosti nezávislosti vieme vzťah (23) upraviť ako

$$\pi_{j|i} = \frac{\pi_{i.}\pi_{.j}}{\pi_{i.}} = \pi_{.j}, \tag{25}$$

čo znamená, že podmienené pravdepodobnosti nezávisia od indexu i.

Pre Kendallovo τ_b sa premenná ϕ_{ij} (10) za platnosti H_0 zjednoduší, keďže $\nu = 0$, na tvar

$$\phi_{ij} = 2\delta \left(R_{ij}^C - R_{ij}^D \right). \tag{26}$$

Analogicky premenná $\phi_{j|i}$ pre viacnásobný multinomický výber (21) sa zjednoduší na

$$\phi_{j|i} = 2\omega_i \delta \left(R_{ij}^C - R_{ij}^D \right). \tag{27}$$

Z (26) a (27) vyplýva

$$\phi_{j|i} = \omega_i \phi_{ij}. \tag{28}$$

Obe variancie σ_1^2 a σ_M^2 upravíme na rovnaký tvar za platnosti H_0 o nezávislosti. Začneme s σ_M^2. Do výrazu pre varianciu pre viac multinomických rozdelení (5) dosadíme výraz (28):

$$\sigma_M^2 = \frac{1}{\delta^4} \sum_i \frac{1}{\omega_i} \left(\sum_j \pi_{j|i} \omega_i^2 \phi_{ij}^2 - \left(\sum_j \pi_{j|i} \omega_i \phi_{ij} \right)^2 \right) \tag{29}$$

$$= \frac{1}{\delta^4} \sum_i \left(\sum_j \omega_i \pi_{j|i} \phi_{ij}^2 - \omega_i \left(\sum_j \pi_{j|i} \phi_{ij} \right)^2 \right). \tag{30}$$

Keďže platí (24), dostávame

$$\sigma_M^2 = \frac{1}{\delta^4} \sum_i \left(\sum_j \pi_{ij} \phi_{ij}^2 - \omega_i \left(\sum_j \pi_{j|i} \phi_{ij} \right)^2 \right). \tag{31}$$

Stačí ukázať, že druhý člen $\sum_j \pi_{j|i}\phi_{ij}$ je rovný 0. Keďže platí (25), tento druhý člen vieme prepísať ako

$$\sum_j \pi_{j|i}\phi_{ij} = \sum_j \pi_{\cdot j}\phi_{ij}. \tag{32}$$

Dosadíme premennú ϕ_{ij} za platnosti H_0 (26):

$$\sum_j \pi_{\cdot j}\phi_{ij} = 2\delta \sum_j \pi_{\cdot j}\left(R^C_{ij} - R^D_{ij}\right). \tag{33}$$

Výraz $R^C_{ij} - R^D_{ij}$ môžeme pri nezávislosti prepísať ako súčin dvoch členov závislých len od marginálnych pravdepodobností. Platí podľa (8)

$$R^C_{ij} - R^D_{ij} = \Pi_{I;i,j} + \Pi_{III;i,j} - \Pi_{II;i,j} - \Pi_{IV;i,j}.$$

Po rozpísaní jednotlivých členov dostávame

$$R^C_{ij} - R^D_{ij} = \sum_{k>i}\sum_{\ell>j}\pi_{k\ell} + \sum_{k<i}\sum_{\ell<j}\pi_{k\ell} - \sum_{k<i}\sum_{\ell>j}\pi_{k\ell} - \sum_{k>i}\sum_{\ell<j}\pi_{k\ell}.$$

Keďže za platnosti H_0 platí $\pi_{k\ell} = \pi_{k\cdot}\pi_{\cdot\ell}$, môžeme písať

$$\begin{aligned} R^C_{ij} - R^D_{ij} = &\left(\sum_{k>i}\pi_{k\cdot}\right)\left(\sum_{\ell>j}\pi_{\cdot\ell}\right) + \left(\sum_{k<i}\pi_{k\cdot}\right)\left(\sum_{\ell<j}\pi_{\cdot\ell}\right) \\ &- \left(\sum_{k<i}\pi_{k\cdot}\right)\left(\sum_{\ell>j}\pi_{\cdot\ell}\right) - \left(\sum_{k>i}\pi_{k\cdot}\right)\left(\sum_{\ell<j}\pi_{\cdot\ell}\right). \end{aligned}$$

Označme si pre jednoduchosť

$$a = \sum_{k>i}\pi_{k\cdot}, \qquad c = \sum_{k<i}\pi_{k\cdot}, \qquad b = \sum_{\ell>j}\pi_{\cdot\ell}, \qquad d = \sum_{\ell<j}\pi_{\cdot\ell}.$$

Potom

$$R^C_{ij} - R^D_{ij} = ab + cd - cb - ad.$$

Úpravou dostávame

$$ab + cd - cb - ad = a(b-d) + c(d-b) = a(b-d) - c(b-d) = (a-c)(b-d).$$

Teda

$$R^C_{ij} - R^D_{ij} = \left(\sum_{k>i}\pi_{k\cdot} - \sum_{k<i}\pi_{k\cdot}\right)\left(\sum_{\ell>j}\pi_{\cdot\ell} - \sum_{\ell<j}\pi_{\cdot\ell}\right).$$

Vidíme teda, že za platnosti H_0 sa rozdiel $R_{ij}^C - R_{ij}^D$ dá prepísať ako súčin člena, ktorý je závislí iba od indexu i a člena, ktorý je závislý iba od indexu j. Preto

$$\sum_j \pi_{\cdot j}(R_{ij}^C - R_{ij}^D) = \left(\sum_{k>i} \pi_{k\cdot} - \sum_{k<i} \pi_{k\cdot}\right) \sum_j \pi_{\cdot j} \left(\sum_{\ell>j} \pi_{\cdot \ell} - \sum_{\ell<j} \pi_{\cdot \ell}\right). \tag{34}$$

Zostáva ukázať, že

$$\sum_j \pi_{\cdot j} \left(\sum_{\ell>j} \pi_{\cdot \ell} - \sum_{\ell<j} \pi_{\cdot \ell}\right) = 0.$$

Roznásobme zátvorku a pozrime sa na jednotlivé členy. Potom

$$\sum_j \pi_{\cdot j} \sum_{\ell>j} \pi_{\cdot \ell} = \sum_j \sum_{\ell>j} \pi_{\cdot j}\pi_{\cdot \ell}, \qquad \sum_j \pi_{\cdot j} \sum_{\ell<j} \pi_{\cdot \ell} = \sum_j \sum_{\ell<j} \pi_{\cdot j}\pi_{\cdot \ell}.$$

Obe dvojité sumy sú rovnaké, pretože v oboch prípadoch sčítavame tie isté súčiny $\pi_{\cdot j}\pi_{\cdot \ell}$ pre všetky dvojice $j \neq \ell$, len raz s podmienkou $\ell > j$ a raz s podmienkou $\ell < j$. Preto

$$\sum_j \pi_{\cdot j} \left(\sum_{\ell>j} \pi_{\cdot \ell} - \sum_{\ell<j} \pi_{\cdot \ell}\right) = 0,$$

a teda podľa (8)

$$\sum_j \pi_{\cdot j}(R_{ij}^C - R_{ij}^D) = 0. \tag{35}$$

Z toho podľa (32) a (33) vyplýva, že

$$\sum_j \pi_{j|i}\phi_{ij} = \sum_j \pi_{\cdot j}\phi_{ij} = 0, \tag{36}$$

a teda $\sum_j \pi_{j|i}\phi_{ij}$ vo výraze (31) mizne. Dostávame

$$\sigma_M^2 = \frac{1}{\delta^4} \sum_i \sum_j \pi_{ij}\phi_{ij}^2.$$

Na druhej strane pre varianciu pre jeden multinomický výber platí (3). Keďže sme ukázali, že platí (27), (28) a (36), člen $\sum_j \pi_{j|i}\phi_{ij}$ vo výraze (3) mizne:

$$\sum_i \sum_j \pi_{ij}\,\phi_{ij} = \sum_i \sum_j \omega_i\,\pi_{j|i}\,\phi_{ij} = \sum_i \sum_j \omega_i\,\pi_{\cdot j}\,\phi_{ij} = \sum_i \omega_i \left(\sum_j \pi_{j|i}\,\phi_{ij}\right) = 0.$$

Potom vieme σ_1 upraviť na

$$\sigma_1^2 = \frac{1}{\delta^4} \sum_i \sum_j \pi_{ij}\phi_{ij}^2.$$

Tým sme ukázali, že $\sigma_1^2 = \sigma_M^2$, teda aj že $\sigma_1 = \sigma_M$, pre Kendallovo τ_b za platnosti H_0.

2.4 Chyba v Agrestiho vzorci

Pri porovnaní nášho odvodenia premennej $\widetilde{\phi_{ij}}$ v (11) s publikovaným vzorcom v [1] na strane 217 sme našli nesúlad v tvare tejto premennej. Agresti uvádza výsledok v tvare

$$\phi_{ij}^{\text{Agresti}} = 2\delta\left(R_{ij}^{C} - R_{ij}^{D}\right) + \nu\left[\pi_{i\cdot}\sqrt{\frac{1-\sum_b \pi_{\cdot b}^2}{1-\sum_a \pi_{a\cdot}^2}} + \pi_{\cdot j}\sqrt{\left(1-\sum_a \pi_{a\cdot}^2\right)\left(1-\sum_b \pi_{\cdot b}^2\right)}\right].$$

Pripomenieme si náš výsledok z rovnosti (11)

$$\widetilde{\phi_{ij}} = 2\delta\Big(R_{ij}^{C} - R_{ij}^{D}\Big) + \frac{\nu}{\delta}\left[\pi_{i\cdot}\left(1-\sum_b \pi_{\cdot b}^2\right) + \ \pi_{\cdot j}\left(1-\sum_a \pi_{a\cdot}^2\right)\right].$$

Vidíme, že prvý člen Agrestiho vzorca je rovnaký ako v našom odvodení. Teraz sa pokúsime prepísať náš tvar premennej $\widetilde{\phi_{ij}}$ na Agrestiho tvar premennej $\phi_{ij}^{\text{Agresti}}$.
Keďže $\delta = \sqrt{\Big(1-\sum_a \pi_{a\cdot}^2\Big)\Big(1-\sum_b \pi_{\cdot b}^2\Big)}$, potom:

$$\begin{aligned}\widetilde{\phi_{ij}} &= 2\delta\Big(R_{ij}^{C} - R_{ij}^{D}\Big) + \nu\left[\pi_{i\cdot}\frac{1-\sum_b \pi_{\cdot b}^2}{\sqrt{\Big(1-\sum_a \pi_{a\cdot}^2\Big)\Big(1-\sum_b \pi_{\cdot b}^2\Big)}} + \pi_{\cdot j}\frac{1-\sum_a \pi_{\cdot a}^2}{\sqrt{\Big(1-\sum_a \pi_{a\cdot}^2\Big)\Big(1-\sum_b \pi_{\cdot b}^2\Big)}}\right] \\ &= 2\delta\Big(R_{ij}^{C} - R_{ij}^{D}\Big) + \nu\left[\pi_{i\cdot}\sqrt{\frac{1-\sum_b \pi_{\cdot b}^2}{1-\sum_a \pi_{a\cdot}^2}} + \pi_{\cdot j}\sqrt{\frac{1-\sum_a \pi_{\cdot a}^2}{1-\sum_b \pi_{b\cdot}^2}}\right].\end{aligned}$$

Vidíme, že pri prepisovaní nášho vzorca na Agrestiho tvar vyzerá posledný člen odlišne. V Agrestiho zápise sa objavuje výraz

$$\nu\pi_{\cdot j}\sqrt{(1-\sum_a \pi_{a\cdot}^2)(1-\sum_b \pi_{\cdot b}^2)},$$

ktorý v našom odvodení nenachádzame. Náš postup ukazuje, že správny tvar obsahuje namiesto toho výraz

$$\nu\pi_{\cdot j}\sqrt{\frac{1-\sum_a \pi_{\cdot a}^2}{1-\sum_b \pi_{b\cdot}^2}},$$

teda v tvare podielu a nie v tvare súčinu.

Rozdiel medzi oboma zápismi spočíva práve v tomto poslednom sčítanci. Je veľmi pravdepodobné, že u Agrestiho ide o preklep alebo chybu v sadzbe, ktorá sa preniesla do publikovaného vzorca. Náš výsledok je konzistentný so štruktúrou predchádzajúcich krokov, preto považujeme našu verziu za korektnú. Tento detail je dôležitý, pretože nesprávny tvar by sa priamo premietol do výpočtu asymptotickej variancie a tým aj do štatistických záverov.

2.4.1 Relatívna chyba pre populačnú disperziu

V predchádzajúcej podkapitole sme ukázali chybu v jednej z premenných, ktoré vstupujú do výpočtu asymptotickej disperzie odhadu $\hat{\tau}_b$ pre jeden multinomický výber. V tejto časti sa na túto chybu pozrieme podrobnejšie, a to pomocou relatívnej chyby štandardnej odchýlky. Zaujíma nás najmä to, aká veľká táto chyba môže byť, a to v populačnom aj vo výberovom prípade.

Relatívna chyba vyjadruje, aká veľká je chyba vzhľadom na skutočnú (správnu) hodnotu. Definujeme ju nasledovne:

$$\xi = \frac{\sigma_{wrong} - \sigma_{ok}}{\sigma_{ok}}.$$

V ďalšom texte ju budeme uvádzať v percentách.

Veľkosť tejto chyby závisí predovšetkým od zvolenej pravdepodobnostnej tabuľky s hodnotami π_{ij}. V populačnom prípade sa z nej priamo vypočítajú premenné ϕ_{ij}, ktoré sú za danú chybu zodpovedné. Vo výberovom prípade sa najprv pomocou multinomického výberu vygeneruje kontingenčná tabuľka, z ktorej sa následne pomocou výberovej verzie premenných ϕ_{ij} vypočíta výberová variancia a výberová štandardná odchýlka. Preto je výber pravdepodobnostnej tabuľky veľmi dôležitý.

Najprv sa zameriame na populačnú štandardnú odchýlku a vypočítame relatívnu chybu v populačnom prípade. Ak pravdepodobnostná tabuľka spĺňa nulovú hypotézu o nezávislosti, člen ν, ktorým je chybný člen násobený, je nulový a chyba sa tak vôbec neprejaví. V prípade, že nulová hypotéza neplatí, sme vyskúšali množstvo pravdepodobnostných tabuliek s hodnotami π_{ij}. Na začiatok sme zvolili diagonálne pravdepodobnostné tabuľky A_1, A_2 a A_3.

$$A_1 = \begin{pmatrix} 0.6 & 0 & 0 & 0 \\ 0 & 0.1 & 0.2 & 0 \\ 0 & 0 & 0 & 0.1 \end{pmatrix} \quad A_2 = \begin{pmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0.2 & 0.1 \end{pmatrix} \quad A_3 = \begin{pmatrix} 0 & 0 & 0 & 0.1 \\ 0 & 0.1 & 0.2 & 0 \\ 0.6 & 0 & 0 & 0 \end{pmatrix} \tag{37}$$

Tabuľka A_1 je východisková. V tabuľke A_2 sme zmenili poradie riadkov a tabuľka A_3 predstavuje prípad s opačne otočenou diagonálou oproti tabuľke A_1. Prvé dve tabuľky majú populačné $\tau_b = 0{,}96$, zatiaľ čo tretia tabuľka s opačnou diagonálou má $\tau_b = -0{,}96$. Všetky tri tabuľky vykazujú rovnakú relatívnu chybu vo výške $\xi = 57{,}63\%$. Vidíme teda,

že chyba vo vzorci pre disperziu vedie v populačnom prípade k vysokej chybovosti. Tá je spôsobená práve extrémnou závislosťou medzi premennými, ktorá vedie k výraznej nenulovej hodnote člena ν, a tým aj k nárastu relatívnej chyby. Ďalšie príklady ukazujú, že vysoká relatívna chyba sa nevyskytuje len pri extrémnych diagonálnych usporiadaniach pravdepodobností π_{ij}, ale aj pri menej pravidelných rozdeleniach pravdepodobností. Napríklad pre tabuľku

$$A_4 = \begin{pmatrix} 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 \\ 0 & 0.1 & 0.2 & 0 \end{pmatrix} \tag{38}$$

dostávame relatívnu chybu $\xi = 34{,}69\%$ pri hodnote $\tau_b = 0{,}75$. Podobne aj pre tabuľku

$$A_5 = \begin{pmatrix} 0 & 0.05 & 0 & 0.05 \\ 0 & 0.1 & 0.1 & 0.1 \\ 0.6 & 0 & 0 & 0 \end{pmatrix}$$

je relatívna chyba $\xi = 39{,}47\%$, pričom $\tau_b = -0{,}85$. V oboch prípadoch ide stále o pomerne silnú závislosť medzi premennými, čo sa prejavuje vyššou chybovosťou. Na druhej strane, ak sa rozdelenie pravdepodobností viac „rozptýli“ mimo dominantnej diagonály, relatívna chyba výrazne klesá. Napríklad pre tabuľku

$$A_6 = \begin{pmatrix} 0.025 & 0.025 & 0.025 & 0.025 \\ 0.05 & 0.05 & 0.1 & 0.1 \\ 0.6 & 0 & 0 & 0 \end{pmatrix} \tag{39}$$

pozorujeme pokles, kde relatívna chyba klesá na $\xi = 5{,}51\%$ a $\tau_b = -0{,}73$.

Tieto výsledky naznačujú, že veľkosť relatívnej chyby nie je daná len hodnotou τ_b, ale aj konkrétnou štruktúrou rozdelenia pravdepodobností π_{ij} v tabuľke. Najväčšie chyby vznikajú pri silne koncentrovaných rozdeleniach, najmä blízko jednej z diagonál, keď je závislosť medzi premennými veľmi výrazná. Naopak, rovnomernejšie rozdelenie pravdepodobností π_{ij} vedie k zníženiu vplyvu člena ν, a tým aj k menšej relatívnej chybe.

2.4.2 Relatívna chyba pre výberovú disperziu

Ďalej sa zameriame na relatívnu chybu v prípade výberovej štandardnej odchýlky. Najprv uvedieme všeobecný vzorec pre výberovú štandardnú odchýlku:

$$\hat{\sigma} = \sqrt{\frac{\sum_i \sum_j \hat{\pi}_{ij}\, \hat{\phi}_{ij}^2 - \left(\sum_i \sum_j \hat{\pi}_{ij}\, \hat{\phi}_{ij}\right)^2}{\hat{\delta}^4}},$$

kde $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$ a

$$\hat{\delta} = \sqrt{\left(1 - \sum_i \hat{\pi}_{i\cdot}^2\right)\left(1 - \sum_j \hat{\pi}_{\cdot j}^2\right)}.$$

V ďalšom rozlišujeme dve verzie podľa použitého výrazu pre $\hat{\phi}_{ij}$.

Pre odhad $\hat{\sigma}_{\text{wrong}}$ dosádzame Agrestiho tvar

$$\hat{\phi}_{ij}^{\text{Agresti}} = 2\hat{\delta}(\hat{R}_{ij}^C - \hat{R}_{ij}^D) + \hat{\nu}\hat{\pi}_{\cdot j}\sqrt{\left(1 - \sum_a \hat{\pi}_{a\cdot}^2\right)\left(1 - \sum_b \hat{\pi}_{\cdot b}^2\right)} + \hat{\nu}\hat{\pi}_{i\cdot}\sqrt{\frac{1 - \sum_b \hat{\pi}_{\cdot b}^2}{1 - \sum_a \hat{\pi}_{a\cdot}^2}}.$$

Pre odhad $\hat{\sigma}_{\text{ok}}$ použijeme náš upravený tvar

$$\hat{\phi}_{ij} = 2\hat{\delta}(\hat{R}_{ij}^C - \hat{R}_{ij}^D) + \hat{\nu}\hat{\pi}_{\cdot j}\sqrt{\frac{1 - \sum_a \hat{\pi}_{a\cdot}^2}{1 - \sum_b \hat{\pi}_{\cdot b}^2}} + \hat{\nu}\hat{\pi}_{i\cdot}\sqrt{\frac{1 - \sum_b \hat{\pi}_{\cdot b}^2}{1 - \sum_a \hat{\pi}_{a\cdot}^2}}.$$

kde $\hat{R}_{ij}^C = \sum_{a<i}\sum_{b<j}\hat{\pi}_{ab} + \sum_{a>i}\sum_{b>j}\hat{\pi}_{ab}$, $\hat{R}_{ij}^D = \sum_{a<i}\sum_{b>j}\hat{\pi}_{ab} + \sum_{a>i}\sum_{b<j}\hat{\pi}_{ab}$ a $\hat{\nu} = \frac{C-D}{\binom{n}{2}}$.

Keďže štandardné odchýlky budú počítané z odhadov, výberová relatívna chyba je definovaná ako

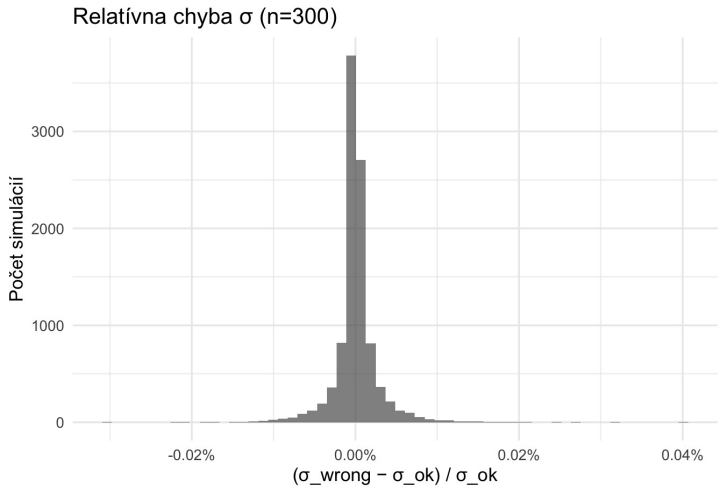
$$\hat{\xi} = \frac{\hat{\sigma}_{wrong} - \hat{\sigma}_{ok}}{\hat{\sigma}_{ok}}.$$

V tomto prípade má zmysel analyzovať aj situáciu, keď je nulová hypotéza splnená, keďže premenná ν sa vtedy odhaduje pomocou počtu konkordantných a diskordantných párov označených C a D, viď (1), a vtedy $\hat{\nu}$ nadobúda hodnoty oscilujúce okolo nuly, no nie nutne nulu. Očakávame, že výsledná chyba bude v takom prípade menej výrazná ako pri porušení predpokladu nezávislosti. Pozrieme sa na túto skutočnosť pomocou príkladov.

Začnime prípadom pravdepodobnostnej tabuľky s hodnotami π_{ij}, ktorá má rovnomerné rozdelenie pravdepodobností:

$$\begin{pmatrix} \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \end{pmatrix}. \tag{40}$$

Na základe tejto tabuľky sme realizovali 10 000 simulácií s veľkosťou výberu $n = 300$. Pre každú simuláciu bola vypočítaná výberová relatívna chyba a následne bol zostrojený jej histogram. Z obrázka 1 vidíme, že rozdelenie relatívnej chyby je sústredené okolo nuly.



Obr. 1: Rovnomerná pravdepodobnostná tabuľka

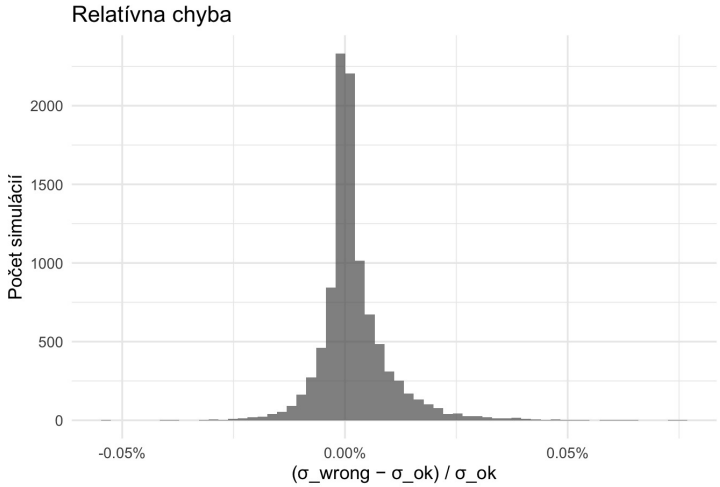
Tento výsledok je v súlade s tým, čo sme očakávali, keďže pri splnení nulovej hypotézy o nezávislosti nadobúda premenná $\hat{\nu}$ hodnoty oscilujúce okolo nuly. V dôsledku toho je vplyv chybného člena vo vzorci zanedbateľný, a teda chyba je minimálna.

Ako druhý príklad, v ktorom je splnená nulová hypotéza, uvažujme pravdepodobnostnú tabuľku, kde pravdepodobnosti nie sú rozdelené rovnomerne:

$$\begin{pmatrix} 0.096 & 0.168 & 0.18 & 0.156 \\ 0.048 & 0.084 & 0.09 & 0.078 \\ 0.016 & 0.028 & 0.03 & 0.026 \end{pmatrix}. \tag{41}$$

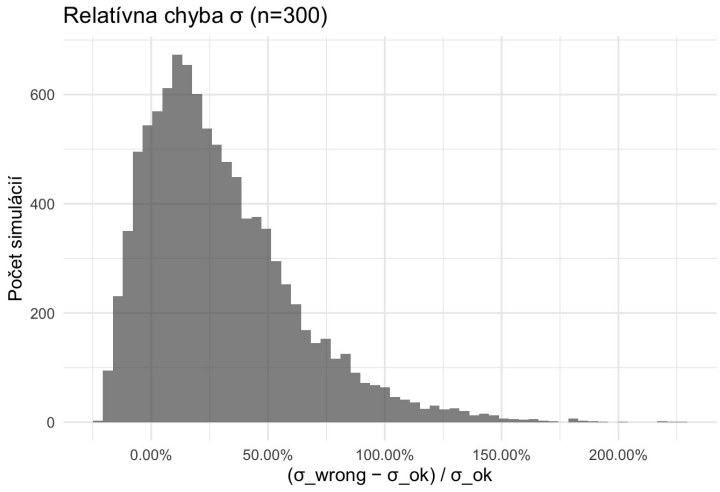
Rovnako ako v predchádzajúcom prípade sme zrealizovali 10 000 simulácií s veľkosťou výberu $n = 300$ a vykreslili histogram relatívnej chyby. Z obrázka 2 vidíme, že relatívna chyba aj v tomto prípade osciluje okolo nuly, čo je v súlade so splnením nulovej hypotézy o nezávislosti. Na rozdiel od predchádzajúceho príkladu však vidíme mierny posun rozdelenia doprava, čo naznačuje, že hodnota odhadu $\hat{\sigma}_{wrong}$ má tendenciu byť mierne vyššia ako hodnota $\hat{\sigma}_{ok}$. Tento efekt je však veľmi slabý.

Teraz sa zameriame na prípady, keď nulová hypotéza nie je splnená, a budeme sledovať správanie relatívnej chyby. Začnime pravdepodobnostnou tabuľkou A_3, v ktorej sú nenulové hodnoty výlučne na opačnej diagonále, uvedenou v predchádzajúcej podkapitole



Obr. 2: Nerovnomerná pravdepodobnostná tabuľka

(pozri (37)). Parametre simulácií ostávajú rovnaké ako v predchádzajúcich príkladoch. Výsledný histogram relatívnej chyby je zobrazený na obrázku 3.

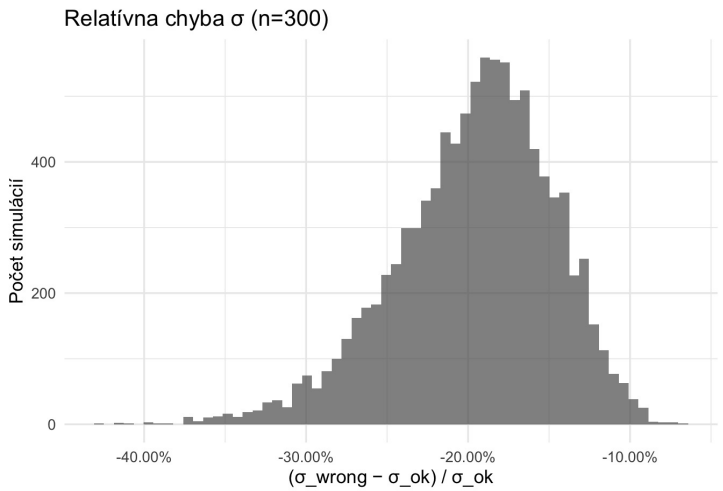


Obr. 3: Diagonálna pravdepodobnostná tabuľka A_3

Z histogramu 3 vidíme, že pri takto zostavenej tabuľke môže byť nesprávne určená hodnota $\hat{\sigma}$ často výrazne väčšia než správna hodnota. V niektorých prípadoch relatívna chyba presahuje 100%, čo znamená, že $\hat{\sigma}_{wrong}$ je viac než dvojnásobná oproti $\hat{\sigma}_{ok}$. Ojedinele sa objavujú aj hodnoty relatívnej chyby presahujúce 200%.

Teraz ukážeme, že voľba pravdepodobnostnej tabuľky má výrazný vplyv na veľkosť relatívnej chyby. Uvažujme tabuľku A_4 ako v predchádzajúcom príklade (pozri (38)). Pri tejto voľbe môže relatívna chyba nadobúdať výrazne odlišné hodnoty v porovnaní s predchádzajúcim prípadom. Postup výpočtu bol rovnaký ako v predošlom príklade,

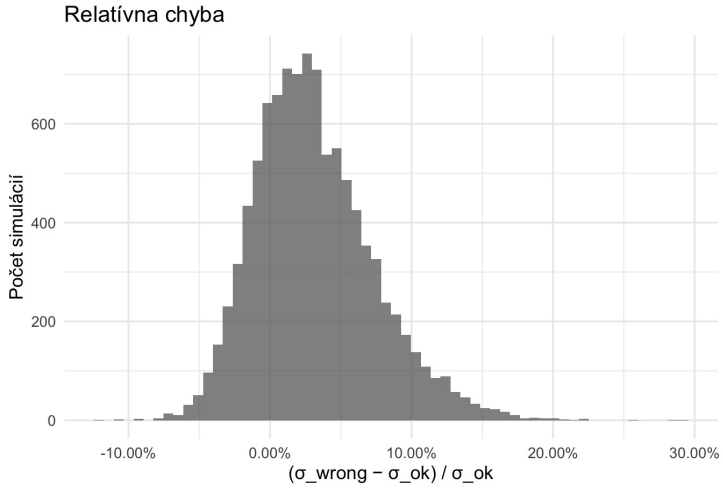
pričom výsledky simulácie sú znázornené na obr. 4.



Obr. 4: Pravdepodobnostná tabuľka A_4

Z histogramu na obr. 4 vidíme, že relatívna chyba nadobúda prevažne záporné hodnoty. To znamená, že nesprávne určená hodnota $\hat{\sigma}_{wrong}$ podhodnocuje správnu hodnotu $\hat{\sigma}_{ok}$. Najviac hodnôt relatívnej chyby sa nachádza približne v intervale od -30% do -15%. To naznačuje, že odhad štandardnej odchýlky môže byť podhodnotený približne o 15 až 30%, v extrémnejších prípadoch až o 40%.

Ako posledný príklad uvažujme tabuľku A_6 uvedenú v (39). Parametre simulácie zostávajú nezmenené a histogram relatívnej chyby je zobrazený na obrázku 5.



Obr. 5: Pravdepodobnostná tabuľka A_6

V prípade uvažovanej tabuľky sú pravdepodobnosti rozložené nerovnomerne, pričom výrazná časť je sústredená mimo hlavnej diagonály. To vedie k prevahe diskordantných

párov nad konkordantnými, čo sa prejavuje záporným ν, a teda aj záporným τ_b. Histogram relatívnej chyby ukazuje, že rozdelenie je posunuté doprava, teda nad nulu. To znamená, že vo väčšine simulácií platí $\hat{\sigma}_{wrong} > \hat{\sigma}_{ok}$.

Z uvedených histogramov vyplýva, že správanie relatívnej chyby je výrazne ovplyvnené štruktúrou pravdepodobnostnej tabuľky s hodnotami π_{ij} a mierou závislosti medzi premennými. V prípadoch, keď je splnená nulová hypotéza o nezávislosti, sa relatívna chyba sústreďuje okolo nuly a nesprávny vzorec v priemere ani nenadhodnocuje, ani nepodhodnocuje štandardnú odchýlku. Naopak, pri porušení nezávislosti dochádza k skresleniu, ktoré sa prejaví posunom rozdelenia relatívnej chyby buď do kladných, alebo záporných hodnôt.

2.4.3 Vplyv chyby na správanie testov

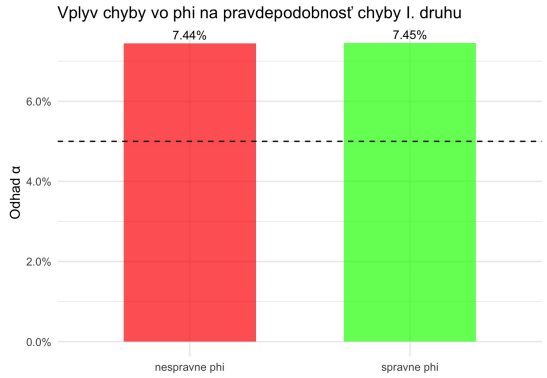
V tejto časti sa zameriame na vplyv Agrestiho chyby vo vzorci pre disperziu na výsledky správanie testov, konkrétne na pravdepodobnosti chyby I. druhu za platnosti nulovej hypotézy, resp. na silu testu.

Najskôr sa pozrieme na situáciu, keď je nulová hypotéza H_0 splnená, a to pre prípady rovnomerne aj nerovnomerne rozdelenej pravdepodobnostnej tabuľky. Zároveň budeme meniť veľkosť výberu: uvažujeme malý rozsah $n = 30$, pri ktorom sa rozdiely medzi výškou pravdepodobnosti chyby I. druhu medzi dvoma uvažovanými spôsobmi môžu prejaviť výraznejšie, a väčší rozsah $n = 300$, kde sa už test asymptoticky približuje normálnemu rozdeleniu. Za platnosti nezávislosti však platí, že $\nu = 0$, pričom práve tento člen je súčasťou chybného výrazu vo vzorci. Preto očakávame, že rozdiel medzi použitiami správneho a nesprávneho vzorca bude v tomto prípade zanedbateľný.

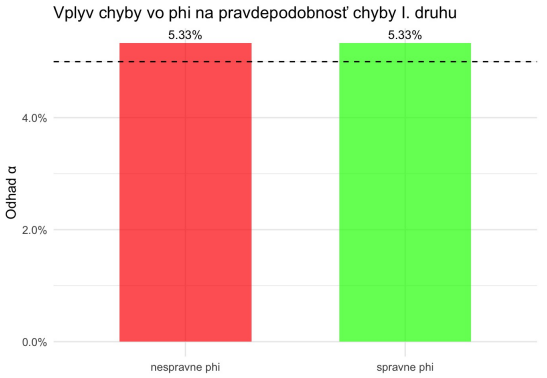
Najprv uvažujme pravdepodobnostnú tabuľku s rovnomerne rozdelenými pravdepodobnosťami π_{ij} uvedenú v (40), ktorá spĺňa hypotézu nezávislosti. Výsledky simulácií pre malé aj veľké n sú zobrazené na obrázku 6.

Z obrázka 6 pozorujeme, že v prípade menšieho rozsahu $n = 30$ je odhadnutá pravdepodobnosť chyby I. druhu v oboch prípadoch prakticky rovnaká, približne 7,44% až 7,45%, a prevyšuje nominálnu hladinu $\alpha = 5\%$. Pri väčšom rozsahu $n = 300$ sa odhad pravdepodobnosti chyby I. druhu približuje k nominálnej hodnote, približne 5,33%.

Uvažujeme aj pravdepodobnostnú tabuľku, kde rozdelenie pravdepodobností π_{ij} je ne-



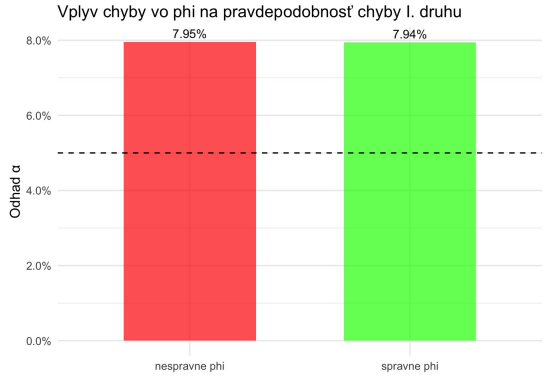
(a) $n = 30$



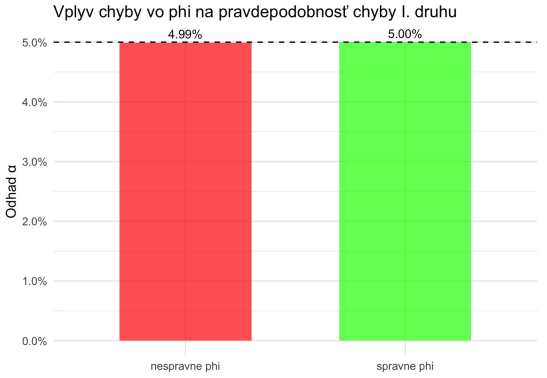
(b) $n = 300$

Obr. 6: Rovnomerná pravdepodobnostná tabuľka

rovnomerné, teda tabuľku uvedenú v (41), ktorá taktiež spĺňa nulovú hypotézu. Výsledky simulácií pre malý aj veľký výber sú uvedené na obrázku 7.



(a) $n = 30$



(b) $n = 300$

Obr. 7: Nerovnomerná pravdepodobnostná tabuľka

Z obrázka 7 vidíme, že v prípade menšieho rozsahu $n = 30$ dosahuje odhad pravdepodobnosti chyby I. druhu približne 7,94% až 7,95%, čo síce predstavuje mierne nadhodnotenie oproti nominálnej hladine $\alpha = 5\%$, rozdiel medzi oboma prístupmi je však veľmi malý. Pri väčšom rozsahu $n = 300$ sa odhad pravdepodobnosti chyby I. druhu výrazne približuje k nominálnej hodnote, približne 5%, pričom hodnoty pre správny aj nesprávny vzorec sú prakticky totožné.

Tieto výsledky, v prípade rovnomernej aj nerovnomernej pravdepodobnosti, potvrdzujú, že za platnosti nulovej hypotézy nemá Agrestiho chyba vo ϕ_{ij} významný vplyv na výsledky testovania. Keďže platí $\hat{\nu} \doteq 0$, chybný člen vo vzorci sa neprejaví temer vôbec a

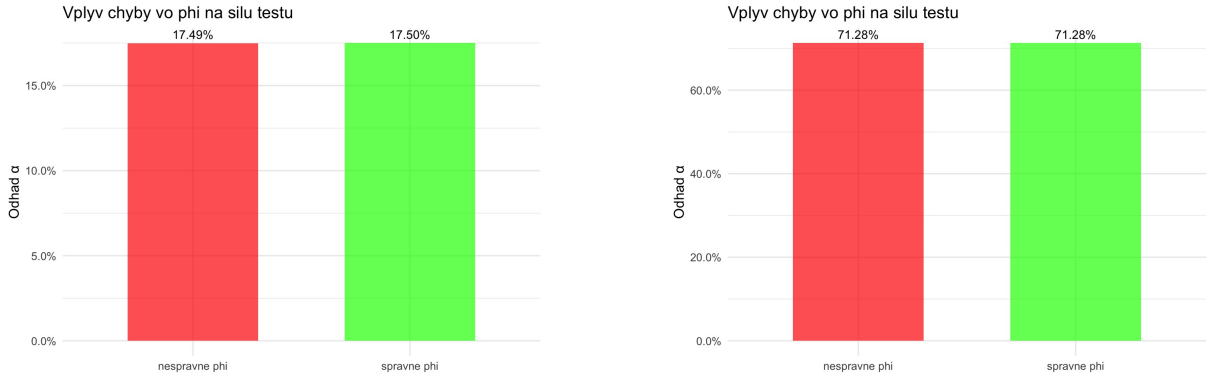
rozdiel medzi správnym a nesprávnym prístupom je zanedbateľný. S rastúcou veľkosťou výberu sa navyše odhad pravdepodobnosti chyby I. druhu stabilizuje okolo nominálnej hodnoty 5%.

Teraz sa pozrieme na pravdepodobnostné tabuľky, kde je H_0 porušená. Budeme sledovať, aký vplyv má Agrestiho chyba na silu testu. Opäť meníme veľkosť výberu, kde uvažujeme $n = 30$ a $n = 300$. Keďže nezávislosť je porušená, premenná ν bude nenulová, s čím by sa aj Agrestiho chyba mohla prejaviť podstatne viac.

Uvažujme tabuľku

$$A_7 = \begin{pmatrix} 0.14 & 0.18 & 0.16 & 0.12 \\ 0.05 & 0.09 & 0.08 & 0.08 \\ 0.01 & 0.01 & 0.06 & 0.02 \end{pmatrix},$$

kde je porušená H_0 a $\tau_b = 0{,}12$. Teda nezávislosť je porušená, ale iba trochu. Opäť prevedieme rovnakú simuláciu s oboma veľkosťami rozsahu a porovnáme. Výsledky sú uvedené na obrázku 8.



(a) $n = 30$ (b) $n = 300$

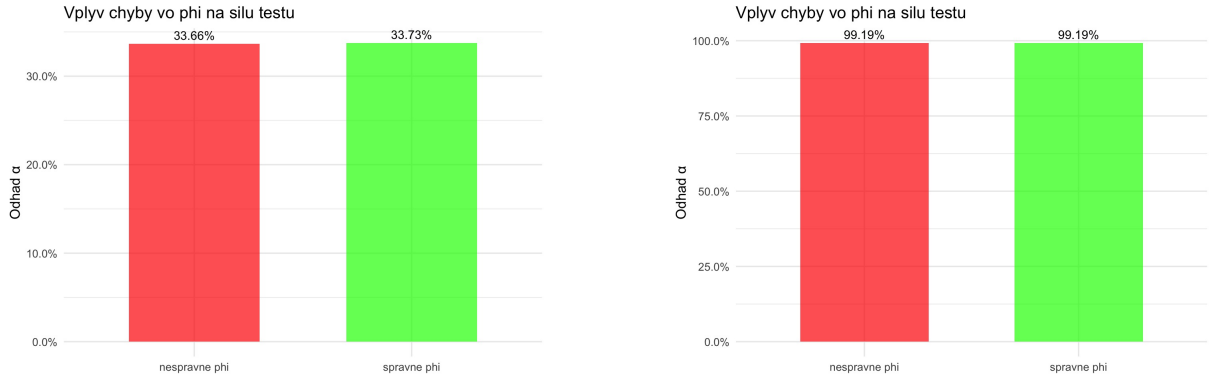
Obr. 8: Pravdepodobnostná tabuľka A_7

Z obrázka 8 vidíme porovnanie sily testu pre nesprávne a správne ϕ_{ij} pri rôznych veľkostiach výberu. V prípade menšieho rozsahu $n = 30$ dosahuje sila testu približne 17,5%. Rozdiel medzi správnym a nesprávnym vzorcom je opäť zanedbateľný. Pri väčšom rozsahu $n = 300$ sa sila testu výrazne zvyšuje na približne 71,3%. Teda, test dokáže odhaliť závislosť so zvyšujúcim sa počtom pozorovaní lepšie ako v prípade s nízkym počtom pozorovaní. Aj v tomto prípade sú výsledky pre správny a nesprávny vzorec prakticky rovnaké.

Teraz uvažujme pravdepodobnostnú tabuľku

$$A_8 = \begin{pmatrix} 0.18 & 0.12 & 0.06 & 0.04 \\ 0.10 & 0.14 & 0.04 & 0.02 \\ 0.06 & 0.08 & 0.10 & 0.06 \end{pmatrix},$$

kde H_0 je porušená a $\tau_b = 0.23$, teda závislosť je vyššia ako v príklade predtým. Výsledky simulácií sú zobrazené na obrázku 9.



(a) $n = 30$ (b) $n = 300$

Obr. 9: Pravdepodobnostná tabuľka A_8

Z obrázka 9 vidíme, že v prípade menšieho rozsahu $n = 30$ dosahuje sila testu približne 34%. Rozdiel medzi správnym a nesprávnym vzorcom je opäť zanedbateľný. Pri väčšom rozsahu $n = 300$ sa sila testu zvyšuje a dosahuje takmer 100%. Test takmer vždy správne zamieta nulovú hypotézu, čo zodpovedá pomerne silnej závislosti medzi premennými. Aj v tomto prípade sú výsledky pre správny aj nesprávny vzorec prakticky rovnaké.

Z uvedených simulácií vidíme, že Agrestiho chyba nemá v žiadnom z uvažovaných scenárov výrazný vplyv na výsledky testovania. Či už je nulová hypotéza splnená alebo nie, rozdiel medzi správnym a nesprávnym vzorcom je prakticky zanedbateľný. Naopak, oveľa dôležitejším faktorom je veľkosť výberu. Pri menšom počte jedincov je sila testu nižšia, zatiaľ čo pri väčšom počte pozorovaní sila testu výrazne rastie.

Poznamenajme, že veľkosť relatívnej chyby úzko súvisí s hodnotou miery asociácie. Pri vysokých hodnotách Kendallovho τ_b (napr. nad 0,75) dochádza k tomu, že sila testu dosahuje hodnoty blízke 100% už aj pri malých výberoch (napr. $n = 30$), a to bez ohľadu na použitý odhad variability. Tento jav sa prejavil aj pri pravdepodobnostných tabuľkách typu A_3, A_4 a A_6 uvedených v (37), (38) a (39), kde v dôsledku veľmi silnej závislosti test

vždy zamietol nulovú hypotézu. V takom prípade sa rozdiel medzi použitím rôznych odhadov štandardnej odchýlky prakticky neprejaví, keďže výsledok testu je v oboch prípadoch rovnaký.

Naopak, pri nižších hodnotách τ_b je síce sila testu menšia, avšak relatívna chyba medzi jednotlivými odhadmi variability je taktiež nižšia, čo vedie k veľmi malým rozdielom v praktických výsledkoch simulácií.

Z toho vyplýva, že výrazná relatívna chyba sa typicky vyskytuje práve v situáciách so silnou závislosťou, kde však jej vplyv na rozhodovanie testu nie je významný. Tento efekt by sa výraznejšie prejavil napríklad pri konštrukcii intervalov spoľahlivosti pre τ_b, kde by väčšia chyba odhadu variability viedla k širším intervalom.

2.4.4 Funkcia z balíka DescTools a komunikácia s profesorom Agrestim

V tejto časti opisujeme našu snahu o nápravu chyby v balíku `DescTools` v prostredí `R` a následnú komunikáciu s profesorom Alanom Agrestim.

Balík `DescTools` obsahuje funkciu `KendallTauB`, ktorá slúži na výpočet Kendallovho τ_b pre kontingenčné tabuľky alebo dvojice ordinálnych premenných. Výpočet τ_b je založený na počte konkordantných a diskordantných párov. Funkcia okrem bodového odhadu zahŕňa aj výpočet intervalov spoľahlivosti pre τ_b. Pri výpočte intervalu spoľahlivosti funkcia využíva odhad disperzie koeficientu τ_b, ktorý je implementovaný priamo v zdrojovom kóde balíka. Tento výpočet vychádza zo vzorca uvedeného v literatúre [1]. Túto skutočnosť sme overili analýzou zdrojového kódu funkcie. Po objavení chyby v implementácii sme kontaktovali správcu balíka, ktorý sa však na e-mail nevyjadril.

Následne sme kontaktovali pána Alana Agrestiho, ktorého literatúra je stavebným kameňom tejto diplomovej práce a uviedli sme do pozornosti chybu vo vzorci pre premennú ϕ_{ij} na strane 217 v knihe [1]. Argumentovali sme dôkazom z časti 2.1 a numerickým porovnaním vzorca pre varianciu τ_b z autorovho článku [3]. Vo svojej odpovedi uviedol, že sa danej problematike už dlhšie nevenuje, avšak predpokladá správnosť nášho zistenia. Zároveň informáciu o možnej chybe doplnil do zoznamu opráv k svojej publikácii, ktorý je dostupný na jeho oficiálnej webovej stránke [2].

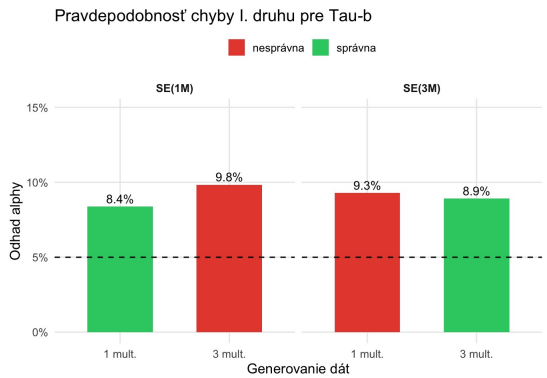
2.5 Simulačné štúdie

V tejto časti sa venujeme simulačným štúdiám za platnosti nulovej hypotézy, ak vytvoríme kontingenčnú tabuľku jedným alebo tromi multinomickými výbermi. Ak vytvoríme kontingenčnú tabuľku jedným multinomickým výberom, použijeme varianciu uvedenú v (17). Ak vytvoríme kontingenčnú tabuľku tromi nezávislými multinomickými výbermi, použijeme varianciu (22). Tieto variancie vymeníme a sledujeme pravdepodobnosť chyby I. druhu aj silu testu. Nakoniec uvedieme do súvisu dôkaz z časti 2.3, ktorý nám neskôr pomôže lepšie vysvetliť výsledky simulácií za platnosti H_0.

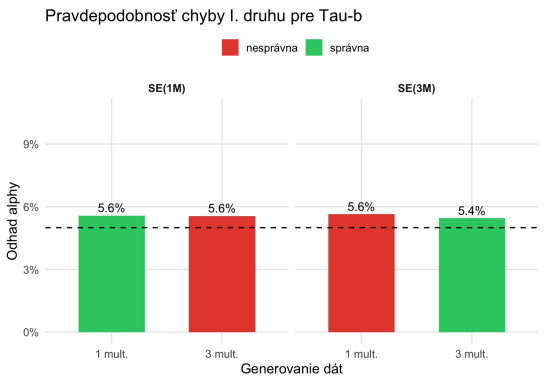
2.5.1 Simulácie za platnosti H_0

V rámci simulácií sme uvažovali dva typy pravdepodobnostných kontingenčných tabuliek: tabuľky s rovnomerne rozdelenými pravdepodobnosťami buniek a tabuľky s nerovnomerným rozdelením pravdepodobností. Pre oba typy tabuliek bola realizovaná rovnaká simulačná štúdia opísaná v časti 1. Hlavným cieľom bolo analyzovať "zamietavosť" testu, konkrétne pravdepodobnosť chyby I. druhu, a porovnať jej správanie pri použití správnej a nesprávnej asymptotickej variancie vzhľadom na generovanie dát.

Simulačná štúdia bola vykonaná na 10 000 replikáciách, pričom najskôr bola analyzovaná situácia pre výberový rozsah $n = 30$ a následne pre $n = 300$. Výsledky pre rovnomerné rozdelenie pravdepodobností sú znázornené na obrázku 10, zatiaľ čo výsledky pre nerovnomerné rozdelenie pravdepodobností sú uvedené na obrázku 11.

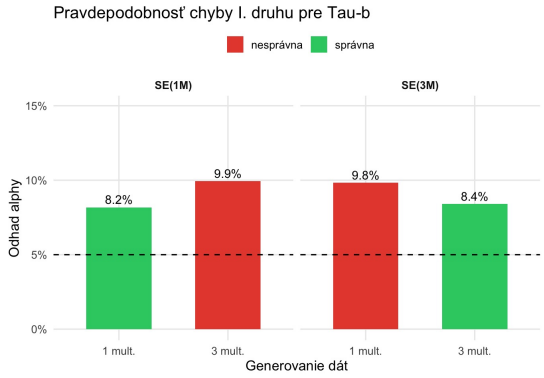


(a) $n = 30$

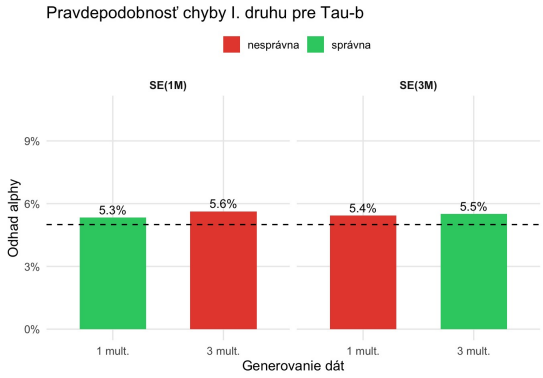


(b) $n = 300$

Obr. 10: Pravdepodobnosť chyby I. druhu pre τ_b - rovnomerná pravdepodobnosť



(a) $n = 30$



(b) $n = 300$

Obr. 11: Pravdepodobnosť chyby I. druhu pre τ_b - nerovnomerná pravdepodobnosť

Zelené stĺpce v grafoch reprezentujú prípady, keď bola použitá asymptotická variancia zodpovedajúca skutočnému spôsobu generovania dát. Konkrétne, pri generovaní tabuľky z jedného multinomického rozdelenia bola použitá variancia odvodená pre jedno multinomické rozdelenie, zatiaľ čo pri generovaní tabuľky z viacerých multinomických rozdelení bola použitá variancia zodpovedajúca tomuto spôsobu. Naopak, červené stĺpce znázorňujú situácie, v ktorých bola použitá nesprávna variancia, teda variancia nezodpovedajúca skutočnému spôsobu generovania dát.

Z výsledkov je zrejmé, že so zvyšujúcim sa počtom pozorovaní dochádza k približovaniu "zamietavosti" testu k hodnote 5%. Pri menších výberoch má odhadnutá pravdepodobnosť chyby I. druhu tendenciu sa od tejto hodnoty odchýliť, čo je dôsledkom nedostatočnej presnosti asymptotickej aproximácie.

Porovnanie výsledkov pri použití správnej a nesprávnej variancie ukazuje, že nesprávne použitie variancie vedie k zvýšeniu pravdepodobnosti chyby I. druhu. Tento efekt sa prejavuje vo všetkých uvažovaných scenároch generovania dát, pričom je výraznejší pri menšom počte výberov.

Pri nerovnomernom rozdelení pravdepodobností sú rozdiely medzi jednotlivými postupmi výraznejšie. Dôvodom je, že pri nerovnomernom rozdelení niektoré bunky tabuľky majú podstatne vyššiu pravdepodobnosť než iné, teda niektoré bunky tabuľky jedincov obsahujú podstatne viac jedincov než iné. To spôsobuje, že informácia o nezávislosti je „nerovnomerne rozložená" v tabuľke, a tým sa zhoršuje presnosť asymptotickej aproximácie, na ktorej je test založený. Preto sa odhadovaná „zamietavosť" testu viac odchyľuje od

nominálnej hodnoty. S rastúcim počtom pozorovaní sa však tento efekt postupne zmenšuje a pri dostatočne veľkých výberoch (napr. $n = 300$) sa rozdiely stávajú zanedbateľnými.

Pre obe pravdepodobnostné tabuľky sme sa pozreli aj na relatívnu chybu medzi správnou a nesprávnou sigmou vzhľadom na generovanie dát. Teda, ak sme generovali dáta pomocou jedného multinomického rozdelenia, populačná relatívna chyba vyzerala nasledovne:

$$\kappa_1 = \frac{\sigma_3 - \sigma_1}{\sigma_1}, \tag{42}$$

kde σ_3 je sigma, ktorú by sme použili, ak by dáta boli generované z troch multinomických rozdelení, zatiaľ čo σ_1 je správna sigma, teda sigma, ktorú používame, ak sú dáta generované z jedného multinomického rozdelenia. V simuláciách sme používali odhady týchto sigiem, teda odhad relatívnej chyby je

$$\hat{\kappa}_1 = \frac{\hat{\sigma}_3 - \hat{\sigma}_1}{\hat{\sigma}_1}. \tag{43}$$

V prípade, že generujeme dáta z troch multinomických rozdelení, je populačná relatívna chyba

$$\kappa_2 = \frac{\sigma_1 - \sigma_3}{\sigma_3},$$

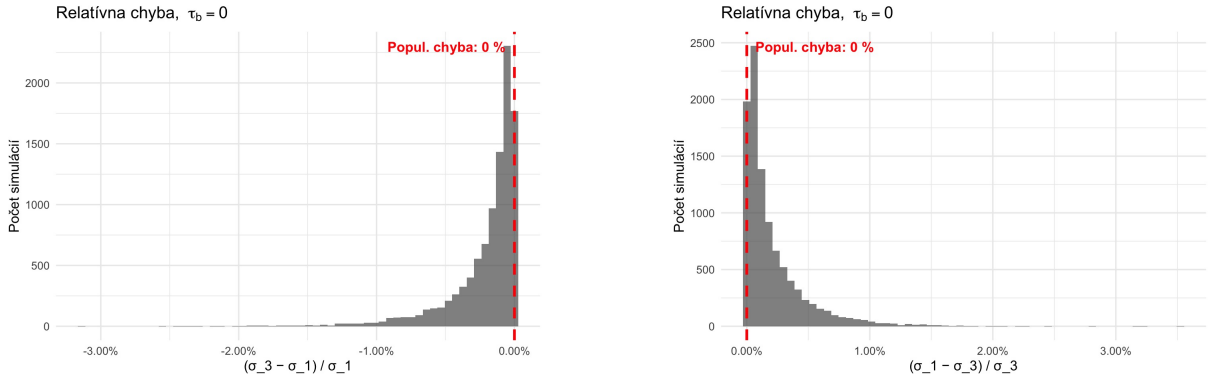
a jej odhad je

$$\hat{\kappa}_2 = \frac{\hat{\sigma}_1 - \hat{\sigma}_3}{\hat{\sigma}_3}.$$

V tejto podkapitole sa venujeme pravdepodobnostným tabuľkám, ktoré spĺňajú nulovú hypotézu, takže populačná relatívna chyba pre rovnomernú aj nerovnomernú pravdepodobnostnú tabuľku je rovná 0. To práve vyplýva z dôkazu v kapitole 2.3, kde sme ukázali, že ak je H_0 splnená, štandardné odchýlky pre rôzne generovanie dát sa rovnajú.

Na histogramoch 12 vidíme výsledky simulácií pre $n = 300$ pre rovnomerne rozdelenú tabuľku pravdepodobností. Ľavý histogram zobrazuje situáciu, kedy sú dáta generované pomocou jedného multinomického rozdelenia a pravý histogram situáciu, kedy sú dáta generované pomocou troch mutinomických rozdelení.

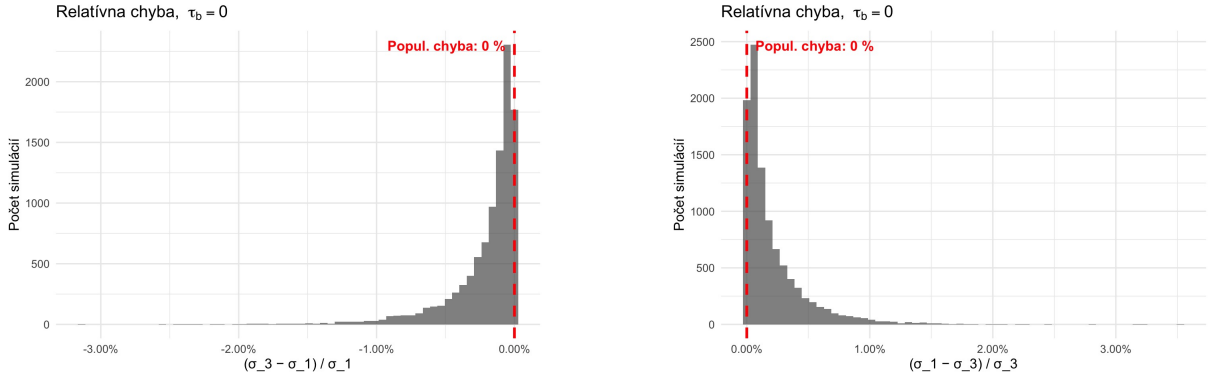
Populačná sigma je znázornená na histogramoch červenou prerušovanou čiarou. Z histogramov vidíme, že väčšina hodnôt relatívnej chyby sa nachádza v blízkosti 0, čo potvrdzuje, že oba odhady štandardnej odchýlky sú si vo väčšine prípadov veľmi podobné. Rozdelenie je výrazne asymetrické a nadobúda prevažne hodnoty jedného znamienka. Tento jav je spôsobený tým, že odhad $\hat{\sigma}_3$ podhodnocuje skutočnú variabilitu, ak sú dáta



Obr. 12: Histogramy pre $\hat{\kappa}_1$ a $\hat{\kappa}_2$ - rovnomerná pravdepodobnostná tabuľka

generované cez 1 multinom (obr. 12 vľavo). Na druhej strane, $\hat{\sigma}_1$ nadhodnocuje, ak sú dáta generované z 3 multinomických rozdelení (obr. 12 vpravo).

Na histogramoch 13 vidíme vyzobrazené výsledky pre nerovnomerne rozdelené tabuľku za platnosti H_0. Ľavý histogram zobrazuje situáciu, kedy sú dáta generované pomocou jedného multinomického rozdelenia a pravý histogram situáciu, kedy sú dáta generované pomocou troch mutinomických rozdelení.



Obr. 13: Histogramy pre $\hat{\kappa}_1$ a $\hat{\kappa}_2$ - nerovnomerná pravdepodobnostná tabuľka

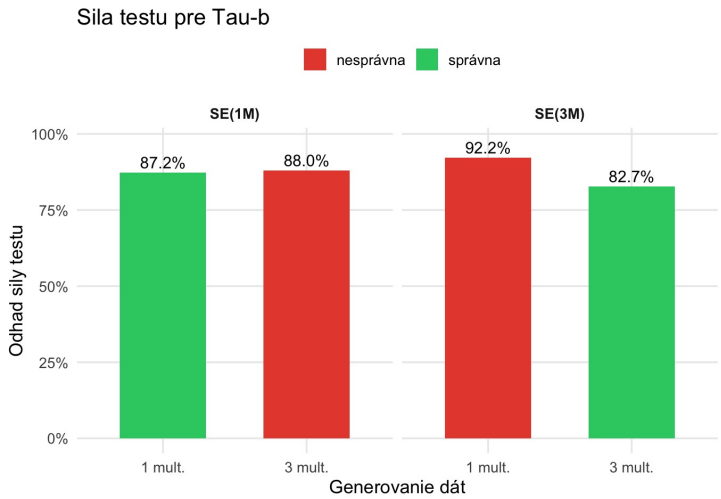
Na obrázku 13 vidíme veľmi podobný výsledok ako v prípade rovnomernej pravdepodobnosti. Aj pri nerovnomernom rozdelení pravdepodobností platí, že odhad $\hat{\sigma}_3$ podhodnocuje skutočnú variabilitu, ak sú dáta generované cez 1 multinom (obr. 13 vľavo). Na druhej strane, $\hat{\sigma}_1$ nadhodnocuje, ak sú dáta generované z 3 multinomických rozdelení (obr. 13 vpravo).

2.5.2 Simulácie za platnosti H_1

Ak si zvolíme pravdepodobnostnú tabuľku tak, aby H_0 o nezávislosti prestala platiť, začneme sledovať namiesto pravdepodobnosti chyby I. druhu silu testu. Zvolili sme túto pravdepodobnostnú tabuľku, ktorá má $\tau_b = 0.5$, teda s pomerne silnou mierou závislosti:

$$\begin{pmatrix} 0.03 & 0.35 & 0.20 & 0.02 \\ 0.02 & 0.06 & 0.13 & 0.09 \\ 0.00 & 0.00 & 0.01 & 0.09 \end{pmatrix}. \tag{44}$$

Previedli sme simulačnú štúdiu iba pre $n = 30$, keďže kvôli tejto silnej závislosti pri $n = 300$ je sila testu 100% a nevidíme rozdiel medzi týmito dvoma prístupmi. Výsledky sú zobrazené na obrázku 14.



Obr. 14: Sila testu pre τ_b - H_0 neplatí

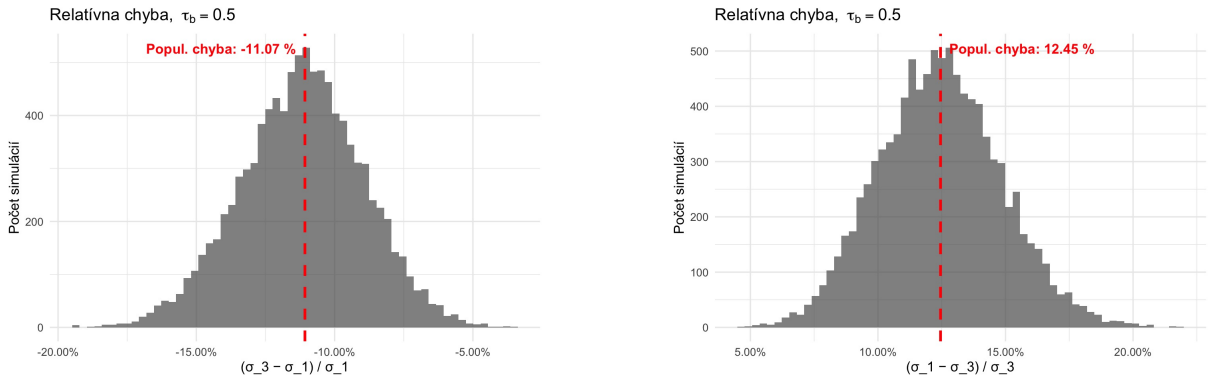
Ako prvé si na obrázku 14 všimneme, že už pri $n = 30$ je sila testu vysoká pre všetky uvažované scenáre. Ak generujeme dáta jedným multinomickým rozdelením a použijeme správnu štandardnú odchýlku, test správne zamietne nulovú hypotézu o nezávislosti približne v 87,2% prípadov. Pri použití nesprávnej štandardnej odchýlky sa tento podiel zvyšuje na 88%. V prípade generovania dát cez tri multinomické rozdelenia je sila testu pri použití štandardnej odchýlky pre jeden multinom až 92,2% a pri použití štandardnej odchýlky pre viacero multinomických rozdelení 82,7%.

Tento jav súvisí so skutočnosťou, že nesprávne zvolená štandardná odchýlka spôsobuje skreslenie testovej štatistiky. Konkrétne, ak je štandardná odchýlka podhodnotená

oproti skutočnému spôsobu generovania dát, dochádza k „nafúknutiu“ hodnoty testovej štatistiky, čo vedie k častejšiemu prekročeniu kritických hodnôt a tým aj k vyššej sile testu. Tento efekt však nie je možné interpretovať ako zlepšenie kvality testu, lebo test má tendenciu častejšie zamietať nulovú hypotézu, čo síce vedie k vyššej sile, ale zároveň aj k nárastu pravdepodobnosti chyby I. druhu, ako sme videli v predchádzajúcej podkapitole. Ide teda o situáciu, keď je vyššia sila dosiahnutá na úkor správnej kontroly hladiny významnosti.

Na danej pravdepodobnostnej tabuľke sme sa pozreli aj na relatívne chyby κ_1, κ_2 a ich odhady. Relatívna populačná chyba $\kappa_1 \doteq -11{,}07\%$ a $\kappa_2 \doteq 12{,}45\%$.

Histogramy pre $\hat{\kappa}_1$ a $\hat{\kappa}_2$ pre $n = 300$ sú zobrazené na obrázku 15.



Obr. 15: Histogramy $\hat{\kappa}_1$ a $\hat{\kappa}_2$

Z obrázka 15 vidíme, že väčšina hodnôt sa sústreďuje okolo populačnej relatívnej chyby, ktorá je vyznačená červenou prerušovanou čiarou. Pri $\hat{\kappa}_1$ sa hodnoty pohybujú v záporných číslach, zatiaľ čo pri $\hat{\kappa}_2$ sú naopak kladné. To súvisí s tým, že tieto dve veličiny majú opačné znamienko, hoci vychádzajú z toho istého rozdielu medzi σ_1 a σ_3.

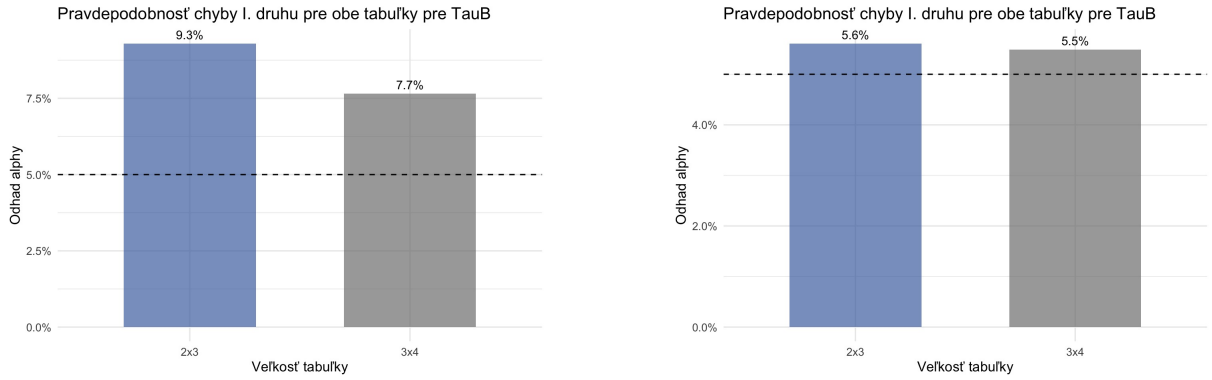
Hoci relatívna chyba odhadu štandardnej odchýlky dosahuje približne $11 - 12\%$, jej vplyv na silu testu je pomerne malý. Dôvodom je, že pri hodnote $\tau_b = 0{,}5$ je závislosť medzi premennými už dostatočne výrazná, takže veľká časť testových štatistík prekračuje kritickú hodnotu bez ohľadu na to, ktorý odhad štandardnej odchýlky použijeme. Chyba v odhade variability teda síce mení hodnotu testovej štatistiky, ale vo väčšine simulácií nezmení samotné rozhodnutie testu. Výraznejší rozdiel by sa prejavil najmä v situáciách, keď by sa hodnoty testovej štatistiky nachádzali v blízkosti kritickej hranice.

2.6 Vplyv zlučovania kategórií

V tejto časti sa pozrieme na to, ako vplýva zlučovanie kategórií na simulačné štúdie. Teda, ak máme kontingenčnú tabuľku výberov veľkosti napr. 3×4, po zlúčení kategórií (budeme zlučovať prvý riadok s druhým a prvý stĺpec s druhým stĺpcom), dostaneme tabuľku veľkosti 2×3. Potom sa pozrieme na to, ako toto zlúčenie ovplyvní test nezávislosti. Urobíme tak aj pre pravdepodobnostné tabuľky, kde je H_0 splnená, aj pre pravdepodobnostnú tabuľku, kde nie je splnená. V tejto podkapitole budeme generovať dátovú tabuľku len z jedného multinomického rozdelenia.

2.6.1 Za platnosti H_0

Výsledky pre rovnomernú pravdepodobnosť sú zobrazené na obrázku 16.

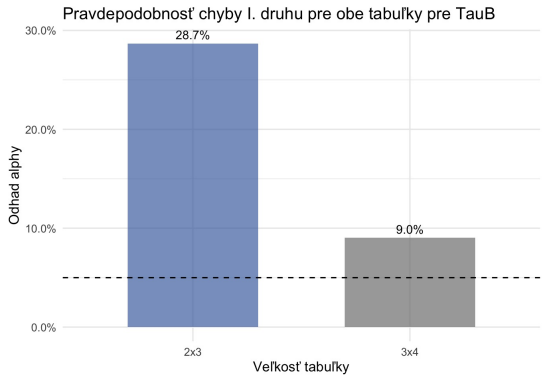


(a) $n = 30$ (b) $n = 300$

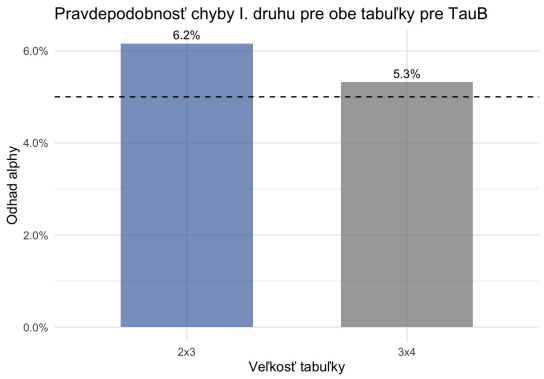
Obr. 16: Pravdepodobnosť chyby I. druhu pre τ_b - rovnomerná pravdepodobnosť

Z grafov opäť vidíme, že rozdiely sú badateľné, ak použijeme nižší počet jedincov. Je to kvôli asymptotickým vlastnostiam tejto miery. Ak sa pozrieme na rozdiel medzi pravdepodobnosťami chýb I. druhu pre $n = 30$, vidíme, že pravdepodobnosť chyby I. druhu je po zlúčení na tabuľku veľkosti 2×3 vyššia ako pred zlúčením kategórií. Treba však upozorniť, že to môže byť spôsobené práve tým, že po zlúčení už H_0 o nezávislosti platiť nemusí. V tomto prípade s rovnomernou tabuľkou pravdepodobnosti veľkosti 3×4 po zlúčení na tabuľku pravdepodobnosti 2×3 už pravdepodobnosti síce nebudú rovnomerne rozdelené, koeficient τ_b však zostáva rovný 0. Takže nezávislosť v tabuľke pravdepodobností zostáva neporušená. Ak zväčšíme rozsah výberu na 300 jedincov, pravdepodobnosť chyby I. druhu po zlúčení riadkov a stĺpcov na veľkosť 2×3 je mierne vyššia, asymptoticky

sa však približuje k 5%.



(a) $n = 30$



(b) $n = 300$

Obr. 17: Pravdepodobnosť chyby I. druhu pre τ_b - nerovnomerná pravdepodobnosť

Na obrázku 17 je znázornený vplyv zlúčenia prvého riadku s druhým a prvého stĺpca s druhým stĺpcom v kontingenčnej tabuľke na výsledky simulačných štúdií, konkrétne na pravdepodobnosť chyby I. druhu za platnosti nulovej hypotézy o nezávislosti, pričom dáta sú generované z multinomického rozdelenia, ktorého vektor pravdepodobností má nerovnaké zložky. Pravdepodobnostnú tabuľku pred a po zlúčení uvádzame tu:

$$\mathbf{Q_1} = \begin{pmatrix} 0.096 & 0.168 & 0.180 & 0.156 \\ 0.048 & 0.084 & 0.090 & 0.078 \\ 0.016 & 0.028 & 0.030 & 0.026 \end{pmatrix} \Rightarrow \begin{pmatrix} 0.396 & 0.270 & 0.234 \\ 0.044 & 0.030 & 0.026 \end{pmatrix}.$$

V porovnaní s prípadom rovnomerného rozdelenia vidíme, že pri menšom výberovom rozsahu sú rozdiely výraznejšie. Koeficient τ_b je aj po zlúčení rovný 0 napriek tomu je pravdepodobnosť chyby pre nízke n po zlúčení až 28,7%. Pri nerovnomerných pravdepodobnostiach má zlúčenie buniek výraznejší dopad na rozdelenie pravdepodobností v tabuľke, čo sa následne prejavuje zvýšením pravdepodobnosti chyby I. druhu. Keďže pravdepodobnostná tabuľka veľkosti 3×4 obsahuje 60% jedincov v prvom riadku a 30% jedincov v druhom riadku, ako bolo vysvetlené v kapitole 1, po zlúčení na tabuľku 2×3 už nová tabuľka jedincov bude obsahovať v prvom riadku približne 90% jedincov a v druhom riadku iba 10% jedincov. To spôsobuje veľkú nevyrovnanosť informácie v tabuľke o závislosti, resp. nezávislosti premenných. To sa prejaví práve vysokou pravdepodobnosťou chyby I. druhu, pretože na fungovanie asymptotiky nestačí, aby celkové n bolo veľké - všetky n_{ij} by mali byť veľké.

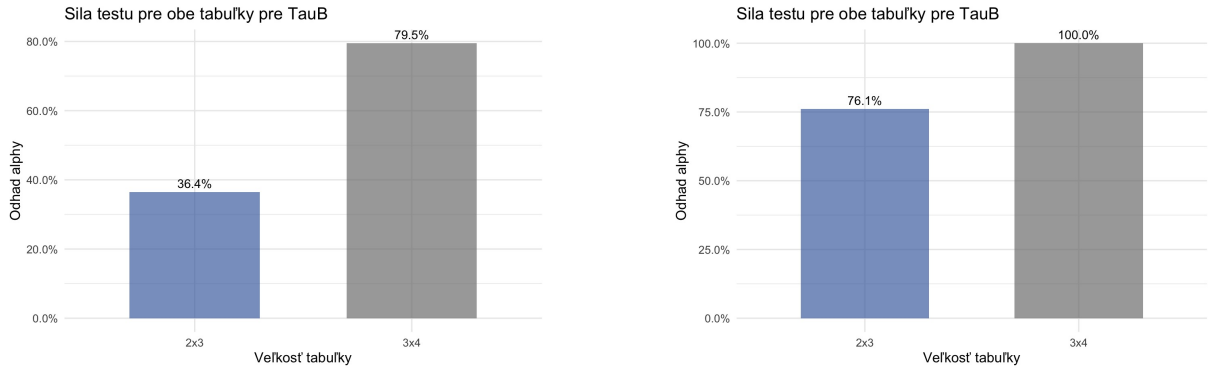
Naopak, pri dostatočne veľkom výberovom rozsahu sa tieto rozdiely postupne zmenšujú a výsledky sa stabilizujú na 5%.

2.6.2 Za platnosti H_1

Nakoniec simulačnú štúdiu prevedieme aj v prípade, kedy dáta generujeme z pravdepodobnostnej tabuľky, kde H_0 neplatí:

$$\mathbf{Q_2} = \begin{pmatrix} 0.30 & 0.20 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.10 & 0.10 \\ 0.01 & 0.03 & 0.03 & 0.03 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} 0.60 & 0.15 & 0.15 \\ 0.04 & 0.03 & 0.03 \end{pmatrix}.$$

Opäť zlúčime prvý a druhý riadok a prvý a druhý stĺpec. V tomto prípade už nesledujeme pravdepodobnosť chyby I. druhu, ale silu testu.



(a) $n = 30$ (b) $n = 300$

Obr. 18: Sila testu pre τ_b - neplatnosť H_0

Na obrázku 18 sú zobrazené výsledky z tejto simulačnej štúdie. Pri menšom výberovom rozsahu pozorujeme výrazný rozdiel v sile testu medzi tabuľkami. Pre tabuľku 2×3 dosahuje sila testu približne 36,4%, zatiaľ čo pre tabuľku 3×4 je výrazne vyššia, približne 79,5% pre $n = 30$. To naznačuje, že pri menších výberoch má väčšia tabuľka vyššiu schopnosť zachytiť závislosť medzi premennými.

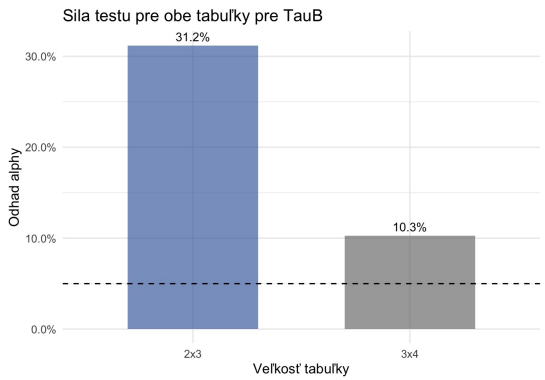
Pri väčšom výberovom rozsahu sa sila testu zvyšuje v oboch prípadoch. Pre tabuľku 2×3 dosahuje približne 76,1%, zatiaľ čo pre tabuľku 3×4 dosahuje maximálnu hodnotu 100%. Rozdiel medzi tabuľkami síce pretrváva, avšak v oboch prípadoch je test výrazne silnejší než pri menšom výbere.

Celkovo možno konštatovať, že pri tejto tabuľke pravdepodobností sila testu rastie so zvyšujúcim sa výberovým rozsahom a zároveň je vyššia pre tabuľku s väčším počtom kategórií.

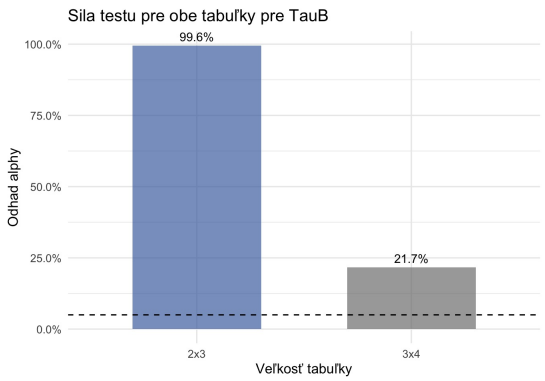
Nakoniec sa pozrieme ešte na jednu pravdepodobnostnú tabuľku P, ktorá taktiež nespĺňa nulovú hypotézu o nezávislosti, a zlúčime ju na veľkosť 2×3:

$$P = \begin{pmatrix} \frac{4}{74} & \frac{3}{74} & \frac{1}{74} & \frac{6}{74} \\ \frac{4}{74} & \frac{20}{74} & \frac{6}{74} & \frac{1}{74} \\ \frac{7}{74} & \frac{3}{74} & \frac{16}{74} & \frac{3}{74} \end{pmatrix} \quad \Longrightarrow \quad P^* = \begin{pmatrix} \frac{31}{74} & \frac{7}{74} & \frac{7}{74} \\ \frac{10}{74} & \frac{16}{74} & \frac{3}{74} \end{pmatrix}.$$

Koeficient τ_b pre maticu P je približne 0,073 a pre maticu P^* je približne 0,249. Teda, ak zlúčime tabuľku z veľkosti 3×4 na veľkosť 2×3, závislosť v tabuľke sa nám podstatne zvýši. Spustili sme simuláciu pre túto tabuľku a sledujeme silu testu. Výsledky sú zobrazené na obrázku 19.



(a) $n = 30$



(b) $n = 300$

Obr. 19: Sila testu pre τ_b - pravdepodobnostné tabuľky P a P^*

Na obrázku 19 vidíme, že sila testu po zlúčení kategórií je podstatne vyššia ako pred zlúčením, a to pre oba rozsahy výberu. Toto zistenie je v súlade s koeficientom τ_b, ktorý je vyšší pre zlúčenú tabuľku, čo znamená silnejšiu závislosť medzi premennými. Zlúčením kategórií sa teda v tomto prípade zvýraznila táto závislosť, čo vedie k vyššej „zamietavosti" testu. Môžeme tvrdiť teda, že nie iba počet kategórií je podstatným faktorom ovplyvňujúcim silu testu, ale aj samotná výška závislosti pred a po zlúčení kategórií.

3 Goodman-Kruskalova γ

V tretej kapitole odvodíme v prípade viacerých multinomických rozdelení vzorec pre varianciu výberovej Goodman-Kruskalovej $\hat{\gamma}$, ktorého odhad následne využijeme pri implementácii simulačných štúdií. V podkapitole venovanej simuláciám analogicky ako pri τ_b ukážeme, že štandardná odchýlka pre jeden aj pre viacero multinomických výberov sa za platnosti nezávislosti zhoduje. Tento výsledok nám opäť pomôže vysvetliť niektoré výsledky zo simulácií.

V simuláciách budeme porovnávať vplyv použitia správneho a nesprávneho odhadu štandardnej odchýlky vzhľadom na spôsob generovania dát, a to sledovaním relatívnej chyby, pravdepodobnosti chyby I. druhu a sily testu. Uvedieme aj tzv. zeta-transformáciu, ktorá zlepšuje asymptotické vlastnosti testu. Nakoniec sa, rovnako ako pri τ_b, zameriame aj na vplyv zlučovania kategórií na výsledky simulácií.

Začnime definíciou Goodman-Kruskalovej γ.

Definícia 2 (Goodman-Kruskalova γ)**.** Goodman–Kruskalova miera asociácie je definovaná ako

$$\gamma = \frac{\Pi_C - \Pi_D}{\Pi_C + \Pi_D},$$

kde Π_C, resp. Π_D sú uvedené v prvej kapitole, viď (6). Odhad Goodman-Kruskalovej γ potom je

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

V druhej kapitole bol odvodený vzorec pre varianciu koeficientu $\hat{\tau}_b$ v prípade jedného multinomického výberu pomocou (2) a (3). V prípade variancie pre $\hat{\gamma}$ odvodenie neuvádzame, keďže je analogické, s tým rozdielom, že menovateľ má jednoduchší tvar $\delta = \Pi_C - \Pi_D$. Pripomeňme si pre každú bunku (i, j) kvadrantové súčty

$$R_{ij}^C = \sum_{u<i}\sum_{v<j}\pi_{uv} + \sum_{u>i}\sum_{v>j}\pi_{uv}, \quad R_{ij}^D = \sum_{u<i}\sum_{v>j}\pi_{uv} + \sum_{u>i}\sum_{v<j}\pi_{uv}.$$

Potom sa dá ukázať, že

$$\phi_{ij} = 4\Big(\Pi_D\, R_{ij}^C - \Pi_C\, R_{ij}^D\Big).$$

Variancia pre Goodman–Kruskalovu γ je

$$\sigma_{\hat{\gamma}}^2 \;=\; \frac{\sum_i \sum_j \pi_{ij}\, \phi_{ij}^2}{(\Pi_C + \Pi_D)^4}.$$

Odhad variancie je

$$\hat{\sigma}^2_{\hat{\gamma}} \;=\; \frac{\sum_i \sum_j \hat{\pi}_{ij}\, \hat{\phi}^2_{ij}}{(\hat{\Pi}_C + \hat{\Pi}_D)^4}.$$

3.1 Odvodenie variancie pre Goodman–Kruskalovu γ pre nezávislé multinomické vzorky

Rovnako ako pri odvádzaní variancie pre nezávislé multinomické vzorky koeficientu τ_b, začneme s pripomenutím definícií:

$$\gamma \;=\; \frac{\nu}{\delta}, \qquad \nu = \Pi_C - \Pi_D, \qquad \delta = \Pi_C + \Pi_D.$$

Derivácie ν a δ podľa $\pi_{b|a}$. V časti 2.2 sme v (18) sme odvodili podmienenú deriváciu čitateľa pre τ_b. Keďže čitatele pre τ_b a γ sú rovnaké, prevezmeme výsledok tejto derivácie z tejto časti. Preto:

$$\nu'_{b|a} \;= \frac{\partial \nu}{\partial \pi_{b|a}} = 2\omega_a\Big(R^C_{ab} - R^D_{ab}\Big)$$

$$\frac{\partial \Pi_C}{\partial \pi_{b|a}} = 2\,\omega_a\,R^C_{ab} \qquad \frac{\partial \Pi_D}{\partial \pi_{b|a}} = 2\,\omega_a\,R^D_{ab}.$$

Potom

$$\delta'_{b|a} = \frac{\partial \delta}{\partial \pi_{b|a}} = \frac{\partial \Pi_C}{\partial \pi_{b|a}} + \frac{\partial \Pi_D}{\partial \pi_{b|a}} = 2\omega_a\Big(R^C_{ab} + R^D_{ab}\Big).$$

Matica ϕ. Všeobecný vzorec pre podmienenú premennú $\phi_{b|a}$ je v (4). Po dosadení premenných $\nu'_{b|a}$ a $\delta'_{b|a}$ dostávame

$$\begin{aligned}\phi_{b|a} &= \delta \cdot 2\omega_a\Big(R^C_{ab} - R^D_{ab}\Big) - \nu \cdot 2\omega_a\Big(R^C_{ab} + R^D_{ab}\Big)\\ &= 2\omega_a\left[\delta\Big(R^C_{ab} - R^D_{ab}\Big) - \nu\Big(R^C_{ab} + R^D_{ab}\Big)\right].\end{aligned}$$

Zaveďme si označenie

$$F_{b|a} := \delta\Big(R^C_{ab} - R^D_{ab}\Big) - \nu\Big(R^C_{ab} + R^D_{ab}\Big).$$

Variancia $\hat{\gamma}$. Z predošlých častí dostávame

$$\phi_{b|a} = 2\,\omega_a\,F_{b|a}, \qquad F_{b|a} = \delta\Big(R^C_{ab} - R^D_{ab}\Big) - \nu\Big(R^C_{ab} + R^D_{ab}\Big).$$

Po dosadení do (5) dostávame

$$\begin{aligned}\sigma^2_{\hat{\gamma}} &= \frac{1}{\delta^4}\sum_i \frac{1}{\omega_i}\left[\sum_j \pi_{j|i}\, 4\omega_i^2 F_{j|i}^2 - \left(\sum_j \pi_{j|i}\, 2\omega_i F_{j|i}\right)^2\right] \\ &= \frac{1}{\delta^4}\sum_i \frac{4\omega_i^2}{\omega_i}\left[\sum_j \pi_{j|i} F_{j|i}^2 - \left(\sum_j \pi_{j|i} F_{j|i}\right)^2\right] \\ &= \frac{4}{\delta^4}\sum_i \omega_i \left[\sum_j \pi_{j|i}\, F_{j|i}^2 - \left(\sum_j \pi_{j|i}\, F_{j|i}\right)^2\right].\end{aligned}$$

3.2 Dôkaz o rovnosti sigiem

Analogický postup, ako pri dôkaze o rovnosti variancii pre Kendallovo $\hat{\tau}_b$, môžeme použiť aj pre dôkaz o rovnosti variancií pre koeficient $\hat{\gamma}$. Pripomeňme, že

$$\gamma = \frac{\Pi_C - \Pi_D}{\Pi_C + \Pi_D} = \frac{\nu}{\delta},$$

kde $\nu = \Pi_C - \Pi_D$ a $\delta = \Pi_C + \Pi_D$.

Za platnosti H_0 o nezávislosti opäť dostávame $\nu = 0$ a preto sa výraz pre ϕ_{ij} mení na

$$\phi_{ij} = \delta \frac{\partial \nu}{\partial \pi_{ij}} = 2\delta\left(R_{ij}^C - R_{ij}^D\right),$$

lebo $\frac{\partial \nu}{\partial \pi_{ij}}$ sme už odvodili v dôkaze pre varianciu pre Kendallovo τ_b pre jeden multinomický výber (9). Keďže za platnosti H_0 platí $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ a (35), rovnako ako pri $\hat{\tau}_b$, potom

$$\sum_i \sum_j \pi_{ij}\phi_{ij} = 2\delta \sum_i \pi_{i\cdot} \sum_j \pi_{\cdot j}\left(R_{ij}^C - R_{ij}^D\right) = 0.$$

Pre varianciu pri jednom multinomickom výbere teda dostávame

$$\sigma_1^2 = \frac{1}{\delta^4}\left(\sum_i\sum_j \pi_{ij}\phi_{ij}^2 - \left(\sum_i\sum_j \pi_{ij}\phi_{ij}\right)^2\right) = \frac{1}{\delta^4}\sum_i\sum_j \pi_{ij}\phi_{ij}^2.$$

Pre varianciu pri viacnásobnom multinomickom výbere platí (5). Rovnako ako pri τ_b platí vzťah (28). Po dosadení dostávame

$$\sigma_M^2 = \frac{1}{\delta^4}\sum_i \left(\sum_j \pi_{ij}\phi_{ij}^2 - \omega_i\left(\sum_j \pi_{j|i}\phi_{ij}\right)^2\right).$$

Za platnosti H_0 platí $\pi_{j|i} = \pi_{\cdot j}$ (25) a (35), preto

$$\sum_j \pi_{j|i}\phi_{ij} = \sum_j \pi_{\cdot j}\phi_{ij} = 2\delta\sum_j \pi_{\cdot j}\left(R_{ij}^C - R_{ij}^D\right) = 0.$$

Tým pádom druhý člen opäť vypadne a dostávame

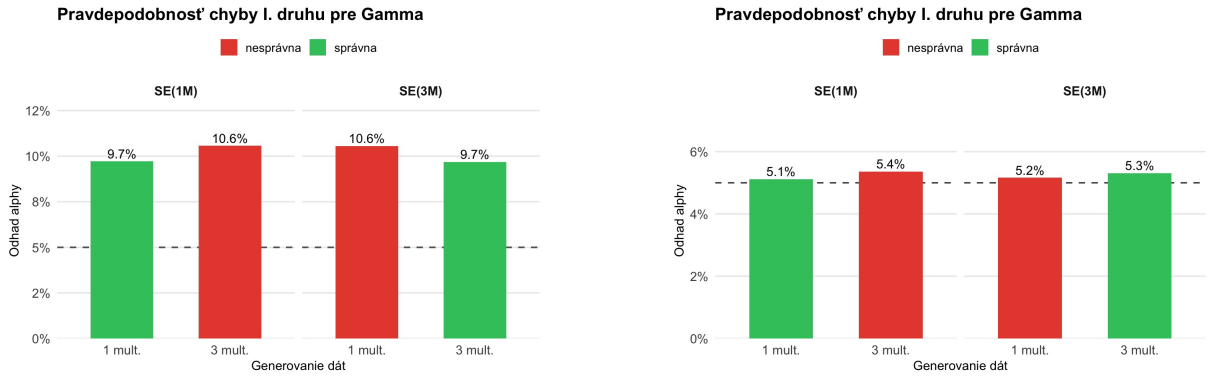
$$\sigma_M^2 = \frac{1}{\delta^4} \sum_i \sum_j \pi_{ij} \phi_{ij}^2.$$

Teda aj pre koeficient γ taktiež platí $\sigma_1^2 = \sigma_M^2$ a $\sigma_1 = \sigma_M$.

3.3 Simulačné štúdie

3.3.1 Simulácie za platnosti H_0

Opätovne sme realizovali simulačné štúdie na dvoch typoch pravdepodobnostných kontingenčných tabuliek, analogicky ako pri τ_b, teda pre tabuľky s rovnomerným a nerovnomerným rozdelením pravdepodobností. Cieľom bolo preskúmať správanie pravdepodobnosti chyby I. druhu a analyzovať jej hodnoty pri použití správnej a nesprávnej asymptotickej variancie. Výsledky sú uvedené na obrázku 20.



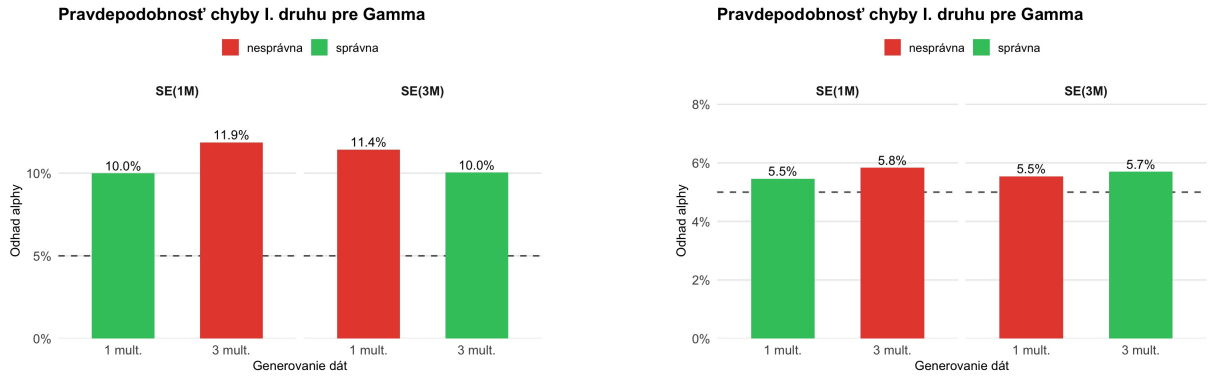
(a) $n = 30$ (b) $n = 300$

Obr. 20: Pravdepodobnosť chyby I. druhu pre γ - rovnomerná pravdepodobnosť

Z obrázka 20 je zrejmé, že pri menšom rozsahu výberu ($n = 30$) dosahuje pravdepodobnosť chyby I. druhu vyššie hodnoty ako pri väčšom rozsahu výberu ($n = 300$). Tento rozdiel je spôsobený nižšou presnosťou asymptotickej aproximácie pri menších výberoch, čo vedie k odchýlkam „zamietavosti" testu od nominálnej hodnoty 5%.

Zároveň, podobne ako v prípade τ_b, pozorujeme, že použitie nesprávnej variancie vzhľadom na spôsob generovania dát vedie k zvýšeniu pravdepodobnosti chyby I. druhu. Tento efekt sa prejavuje v oboch uvažovaných scenároch, pričom je výraznejší pri menších výberových rozsahoch. Pri väčších výberoch sa rozdiely medzi jednotlivými prístupmi zmenšujú.

Rovnakú simulačnú štúdiu sme zrealizovali aj v prípade nerovnomerného rozdelenia pravdepodobností. Výsledky sú zobrazené na obrázku 21.



(a) $n = 30$ (b) $n = 300$

Obr. 21: Pravdepodobnosť chyby I. druhu pre γ - nerovnomerná pravdepodobnosť

Pozorujeme podobné trendy ako v prípade rovnomerného rozdelenia, avšak s väčšími odchýlkami od nominálnej hladiny testu. Pri menšom výbere ($n = 30$) dosahuje pravdepodobnosť chyby I. druhu hodnoty približne medzi 10% až 12%, pričom pri použití nesprávnej variancie sú tieto hodnoty vyššie ako pri použití správnej. Rozdiel medzi použitím správnej a nesprávnej variancie je v tomto prípade zreteľnejší ako pri rovnomernom rozdelení pravdepodobností.

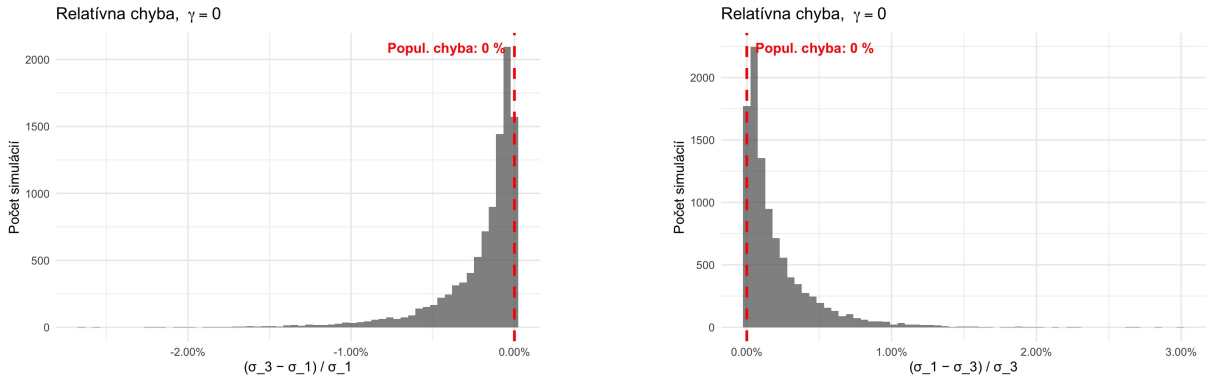
Pri väčšom výbere ($n = 300$) sa pravdepodobnosť chyby I. druhu znižuje a pohybuje sa približne okolo 5,5%, pričom rozdiely medzi jednotlivými scenármi sú už minimálne. Napriek tomu aj v tomto prípade zostáva mierne vyššia pri nesprávnej variancii.

V porovnaní s rovnomerným rozdelením pravdepodobností je zrejmé, že nerovnomerné rozdelenie vedie k vyšším hodnotám pravdepodobnosti chyby I. druhu, najmä pri menších výberoch. To možno pripísať väčšej nevyváženosti kontingenčných tabuliek, ktorá spôsobuje horšiu asymptotickú aproximáciu testovej štatistiky.

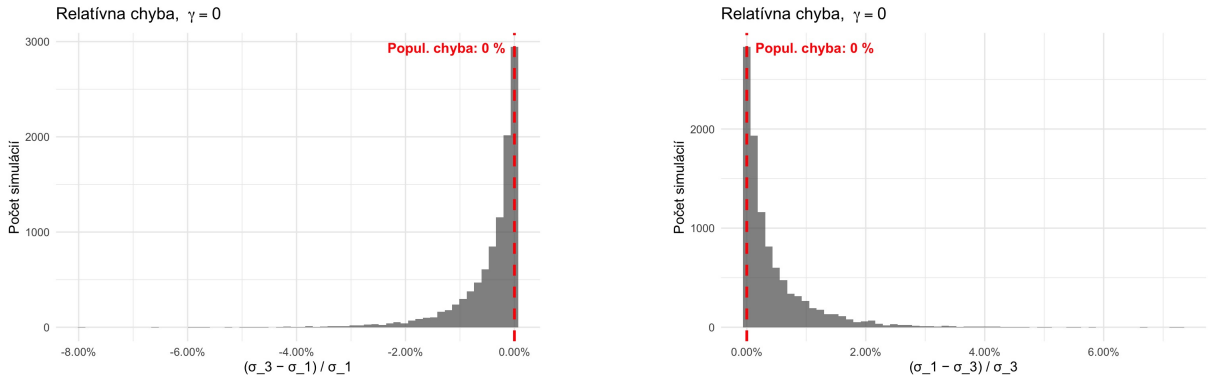
Analogicky ako pri simuláciách ku koeficientu τ_b sa pozrieme aj na relatívnu chybu κ_1 (42) a κ_2 (43) a ich odhady.

Rovnako ako pre τ_b sme dokázali, že aj v prípade koeficientu γ platí, že za platnosti nezávislosti sa štandardné odchýlky pre oba spôsoby generovania dát rovnajú, viď kapitolu 3.2. Preto pre nasledujúce dve pravdepodobnostné tabuľky bude populačná relatívna chyba κ_1 a κ_2 rovná 0.

Pozrime sa teraz na odhad relatívnych chýb κ_1 a κ_2. Výsledky pre rovnomerne rozdelenú tabuľku pravdepodobností sú zobrazené na obrázku 22 a pre nerovnomerne na obrázku 23.



Obr. 22: Histogramy pre $\hat{\kappa}_1$ a $\hat{\kappa}_2$ - rovnomerná - dáta generované z multinomického rozdelenia, ktorého vektor pravdepodobností má rovnaké zložky



Obr. 23: Histogramy pre $\hat{\kappa}_1$ a $\hat{\kappa}_2$ - dáta generované z multinomického rozdelenia, ktorého vektor pravdepodobností má nerovnaké zložky

Správanie relatívnej chyby na obrázku 22 a obrázku 23 je prakticky identické ako v prípade τ_b. Správanie sa relatívnej chyby zodpovedá predchádzajúcim odvodeniam, podľa ktorých sa za platnosti hypotézy H_0 obe miery líšia len v menovateli δ vystupujúcom vo variancii a ďalej v štandardnej odchýlke. Z tohto dôvodu sa relatívna chyba pre obe miery správa prakticky rovnako ako v prípade koeficientu τ_b.

3.3.2 Simulačné porovnanie pravdepodobnosti chyby I. druhu pre Goodman–Kruskalovu γ a jej zeta-transformáciu

V tejto časti sa venujeme porovnaniu konvergencie pravdepodobnosti chyby I. druhu k 5% medzi Goodman-Kruskalovou γ a jej zeta-transformáciou. Začneme s vysvetlením Fisherovej zeta-transformácie [1].

Keďže odhad γ, označený ako $\hat{\gamma}$, sa normalite približuje pomaly, použijeme Fisherovu známu zeta-transformáciu, ktorá preukázateľne zlepšuje konvergenciu. Fisherova zeta-transformácia je definovaná ako

$$\hat{\zeta} = \frac{1}{2}\log\left(\frac{1+\hat{\gamma}}{1-\hat{\gamma}}\right), \qquad \text{čiže} \qquad \hat{\gamma} = \frac{e^{2\hat{\zeta}}-1}{e^{2\hat{\zeta}}+1}$$

a odhad asymptorickej variancie je

$$\hat{\sigma}^2_{\hat{\zeta}} \approx \frac{\hat{\sigma}^2_{\hat{\gamma}}}{(1-\hat{\gamma}^2)^2}.$$

Transformácia vedie k lepšej normalite testovej štatistiky, čo znamená, že $\sqrt{n}(\hat{\zeta}-\zeta)$ konverguje ku $N(0,\sigma^2_{\hat{\zeta}})$ rýchlejšie než $\sqrt{n}(\hat{\gamma}-\gamma)$ ku $N(0,\sigma^2_{\hat{\gamma}})$ aj pre malé počty výberov.

V simulácii testujeme hypotézu $H_0 : \gamma = 0$, čo je ekvivalentné s $H_0 : \zeta = 0$ na hladine $\alpha = 0.05$, kde

$$\zeta = \frac{1}{2}\log\left(\frac{1+\gamma}{1-\gamma}\right).$$

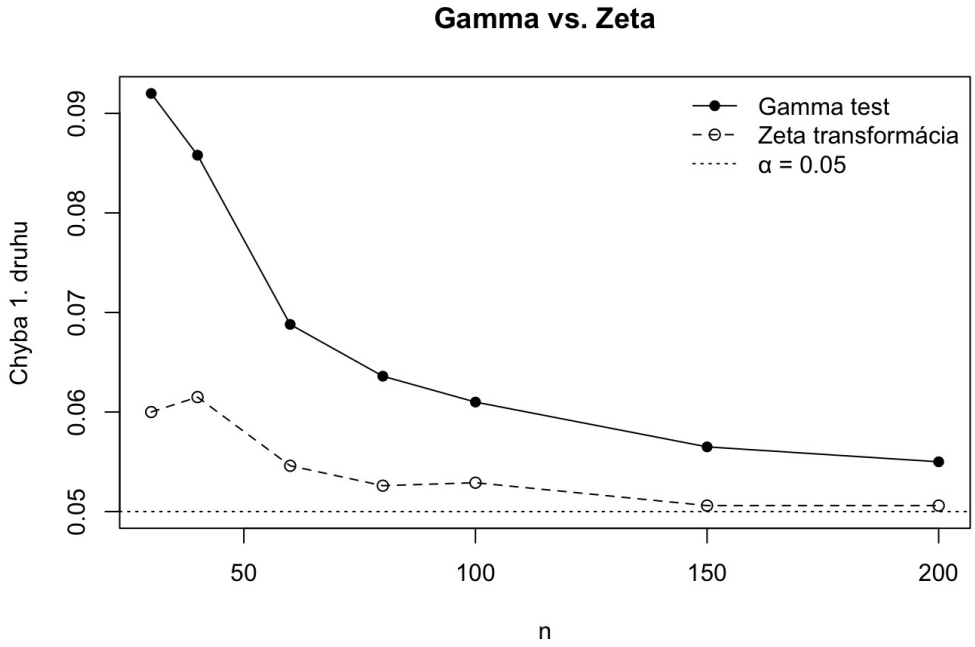
Waldove štatistiky sú

$$Z_\gamma = \frac{\hat{\gamma}}{\sqrt{\frac{\hat{\sigma}^2_{\hat{\gamma}}}{n}}}, \qquad Z_\zeta = \frac{\hat{\zeta}}{\sqrt{\frac{\hat{\sigma}^2_{\hat{\zeta}}}{n}}}$$

a za platnosti H_0 platí aproximácia $Z_\gamma, Z_\zeta \dot{\sim} \mathcal{N}(0,1)$. Hypotézu H_0 zamietame, ak $|Z| > z_{1-\frac{\alpha}{2}}$.

Pre každý rozsah výberu $n \in \{30, 40, 60, 80, 100, 150, 200\}$ urobíme 10 000 simulácií. V každej simulácii vypočítame odhady $\hat{\gamma}$ a $\hat{\zeta}$ spolu s ich odhadmi variancií a následne príslušné testové štatistiky. Na základe nich rozhodneme o zamietnutí nulovej hypotézy. Pravdepodobnosť chyby I. druhu potom odhadujeme ako podiel zamietnutí nulovej hypotézy zo všetkých simulácií, osobitne pre γ aj pre jeho zeta-transformáciu.

Simulácie sme spustili pre dva rôzne scenáre: rovnomerný a nerovnomerný, na základe použitej pravdepodobnostnej tabuľky.



Obr. 24: Pravdepodobnosť chyby I. druhu pre test γ vs. ζ (rovnomerný scenár)

Obrázok 24 zobrazuje odhad pravdepodobnosti chyby I. druhu za platnosti nulovej hypotézy pre test založený na odhade $\hat{\gamma}$ (plná čiara) a pre test využívajúci zeta-transformáciu (čiarkovaná čiara). Vodorovná bodkovaná čiara znázorňuje hladinu významnosti $\alpha = 0{,}05$. V tabuľke 1 sú uvedené konkrétne hodnoty k tomuto príkladu.

n	$\hat{\alpha}_\gamma$	$\hat{\alpha}_\zeta$
30	0.092	0.060
40	0.086	0.062
60	0.069	0.055
80	0.064	0.053
100	0.061	0.053
150	0.056	0.051
200	0.055	0.051

Tabuľka 1: Pravdepodobnosť chyba I. druhu pre test pomocou γ a test pomocou jej zeta-transformácie (rovnomerný scenár)

Pre malé rozsahy výberu sa test založený na $\hat{\gamma}$ správa liberálne. Napríklad pri $n = 30$

dosahuje empirická hladina hodnotu $\hat{\alpha}_\gamma = 9{,}2\,\%$, čo znamená prekročenie nominálnej hladiny o 4,2%. S rastúcim rozsahom výberu n sa pravdepodobnosť chyby I. druhu postupne znižuje, avšak ani pri $n = 200$ ešte úplne nekodonvergovala k nominálnej hodnote (5,5 %, teda prekročenie o 0,5%).

Test založený na zeta-transformácii má priaznivejšie vlastnosti z hľadiska dodržania nominálnej hladiny významnosti. Pri $n = 30$ dosahuje empirická hladina hodnotu $\hat{\alpha}_\zeta = 6\,\%$, čo predstavuje prekročenie o 1%. S rastúcim rozsahom výberu sa tento rozdiel rýchlo zmenšuje a približne od $n = 150$ sa empirická hladina zhruba zhoduje s nominálnou hodnotou $\alpha = 5\,\%$.

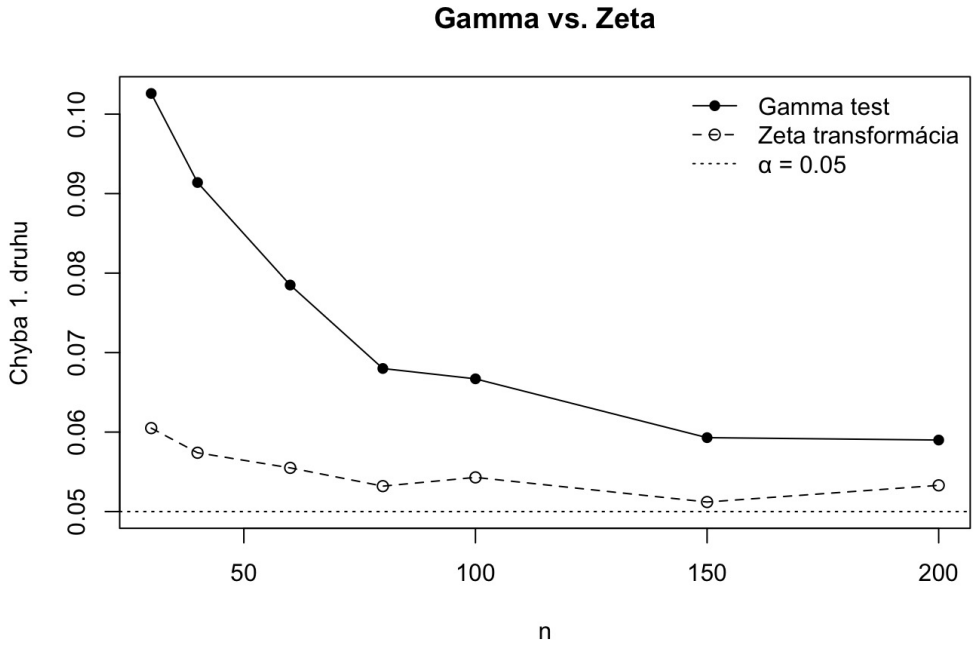
Test využívajúci zeta-transformáciu sa s rastúcim n k nominálnej hladine významnosti približuje podstatne rýchlejšie než test založený na $\hat{\gamma}$ a pri malých až stredných veľkostiach výberu má menšie odchýlky od nominálnej hladiny. Rozdiely medzi oboma testami sa so zvyšujúcim sa rozsahom výberu postupne zmenšujú.

V druhom scenári sme ponechali nulovú hypotézu nezávislosti, no zvolili sme nerovnomernú pravdepodobnosť. Pravdepodobnostná tabuľka s nerovnomernými pravdepodobnosťami bola zostavená rovnakým spôsobom, ako bolo vysvetlené v prvej kapitole. Výsledky simulácií sú zobrazené na obrázku 25 a konkrétne hodnoty k tomuto príkladu sú v tabuľke 2.

n	$\hat{\alpha}_\gamma$	$\hat{\alpha}_\zeta$
30	0.103	0.060
40	0.091	0.057
60	0.078	0.056
80	0.068	0.053
100	0.067	0.054
150	0.059	0.051
200	0.059	0.053

Tabuľka 2: Pravdepodobnosť chyby I. druhu pre γ a jeho zeta-transformáciu (nerovnomerný scenár)

Z obrázka 25 vidíme rovnaký trend ako pri rovnomernej pravdepodobnosti, a síce, že s narastajúcim n klesá odhad pravdepodobnosti chyby I. druhu, pričom pre zeta-



Obr. 25: Pravdepodobnosť chyby I. druhu pre test γ vs. ζ (nerovnomerný scenár)

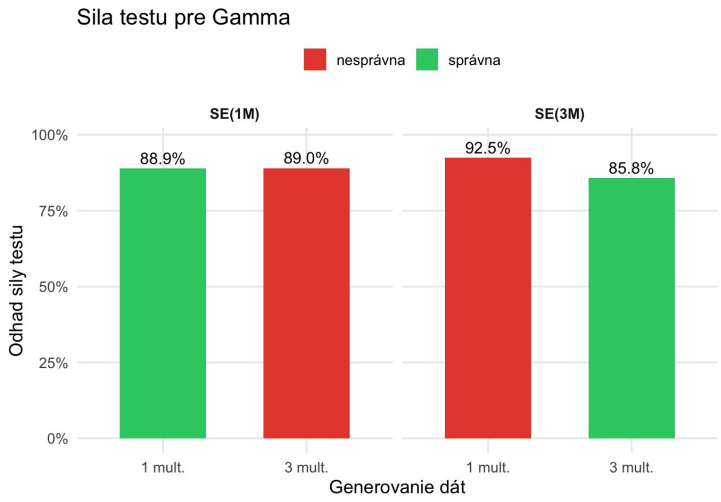
transformáciu rýchlejšie. Pre $n = 30$ jedincov je pre γ 10,3% odhad pravdepodobnosti chyby I. druhu, čo je o 5,3% vyššie ako nominálna hodnota. Aj pri $n = 200$ jedincoch test nedosahuje nominálnu hodnotu 5% a presahuje ju o 0,9%.

Zeta transformácia je hlavne v nižších rozsahoch stabilnejšia ako test využívajúci koeficient γ. Už pri $n = 80$ vieme spoľahlivo tvrdiť, že test má pravdepodobnosť chyby I. druhu približne 5%.

3.3.3 Simulácie za platnosti H_1

V prípade, že H_0 o nezávislosti prestane platiť, teda premenné X a Y už nie sú nezávislé, vieme skúmať silu testu. Aj pre mieru asociácie Goodman-Kruskalova γ sme teda urobili simulačné štúdie ako pri Kendallovom τ_b. Sme použili pravdepodobnostnú tabuľku (44), ktorá má koeficient $\gamma = 0.73$. Simuláciu sme spustili pre $n = 30$, pretože kvôli silnej závislosti je pre $n = 300$ sila testu pre všetky uvažované scenáre 100%. Výsledky simulácie sú zobrazené na obrázku 26.

Z obrázka 26 je zrejmé, že pri použití nesprávnej variancie vzhľadom na spôsob generovania dát dosahuje test mierne vyššiu silu v porovnaní so správnou varianciou. Tento rozdiel je viditeľný najmä pri generovaní dát z troch multinomických rozdelení, kde sa

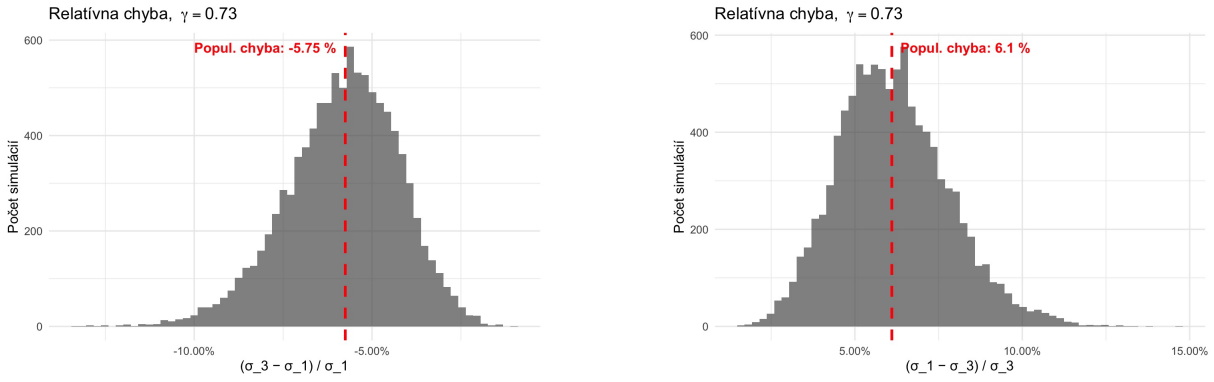


Obr. 26: Sila testu pre γ

sila testu pohybuje približne medzi 85,8% až 92,5%. Pri generovaní dát z jedného multinomického rozdelenia je rozdiel menej viditeľný.

Je však potrebné zdôrazniť, že vyššia sila testu pri použití nesprávnej variancie neznamená zlepšenie jeho vlastností. Ide o dôsledok nesprávneho použitia variancie, ktoré vedie k nadhodnoteniu testovej štatistiky a tým k častejšiemu zamietaniu nulovej hypotézy. Tento jav sa prejavuje nielen zvýšenou silou testu, ale zároveň aj zvýšenou pravdepodobnosťou chyby I. druhu, ako bolo ukázané v predchádzajúcej časti.

Nakoniec použijeme rovnakú pravdepodobnostnú tabuľku, ktorá nespĺňa H_0, ako v prípade τ_b na zobrazenie odhadov relatívnych chýb κ_1 a κ_2. Populačné $\kappa_1 \doteq -5{,}75\%$ a $\kappa_2 \doteq 6{,}1\%$. Histogramy pre $n = 300$ sú zobrazené na obrázku 27.



Obr. 27: Histogramy pre $\hat{\kappa}_1$ a $\hat{\kappa}_2$

Z obrázka 27 vidíme, že aj v prípade, keď hypotéza H_0 neplatí, majú histogramy

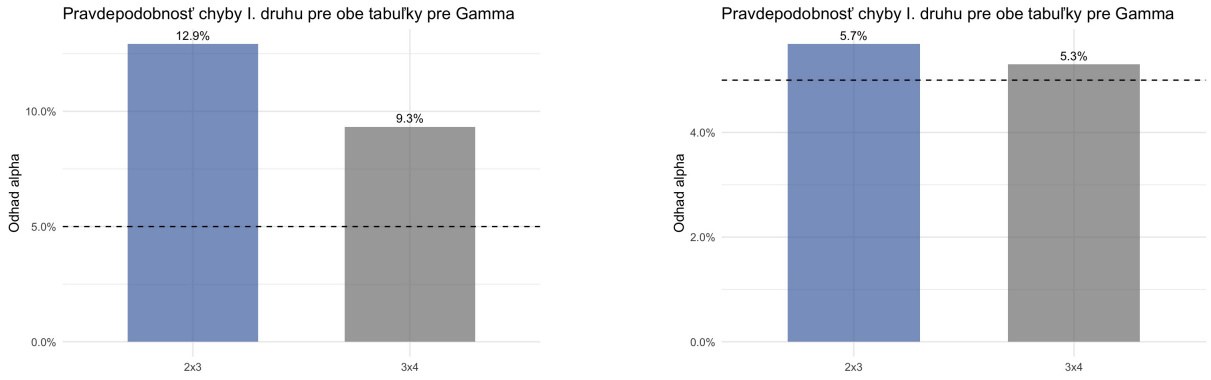
relatívnej chyby pre $\hat{\kappa}_1$ a $\hat{\kappa}_2$ rovnaký charakter ako v prípade τ_b. Pre $\hat{\kappa}_1$ sa hodnoty pohybujú v záporných číslach a pre $\hat{\kappa}_2$ naopak v kladných. To súvisí s tým, že $\hat{\sigma}_3 < \hat{\sigma}_1$, teda že $\hat{\sigma}_3$ je systematicky podhodnotená. To môže byť spôsobené tým, že odhad $\hat{\sigma}_3$ vychádza z viacnásobného multinomického výberu, ktorý nezohľadňuje celkovú variabilitu dát, ale iba variabilitu v rámci jednotlivých riadkov tabuľky.

3.4 Vplyv zlučovania kategórií

V tejto časti diplomovej práce sa pozrieme na vplyv zlučovania riadkov a stĺpcov na pravdepodobnosť chyby I. druhu, resp. silu testu. Zlučovať budeme prvý riadok s druhým riadkom a prvý stĺpec s druhým, rovnako ako tomu bolo pri τ_b v časti 2.6 o zlučovaní kategórií. Teda z pôvodnej tabuľky výberov veľkosti 3×4 dostaneme tabuľku 2×3. Dáta budú v tejto časti opäť generované iba pomocou jedného multinomického rozdelenia. Simulácie spustíme za platnosti H_0 o nezávislosti, ako aj za platnosti H_1, kde nezávislosť porušíme.

3.4.1 Za platnosti H_0

Najprv použijeme pravdepodobnostnú tabuľku s rovnomerným rozdelením pravdepodobností. Po vygenerovaní dátovej tabuľky z multinomického rozdelenia zlúčime kategórie a dostaneme tabuľku veľkosti 2×3. Potom na oboch tabuľkách sledujeme pravdepodobnosť chyby I. druhu. Výsledky sú zobrazené na obrázku 28.



(a) $n = 30$ (b) $n = 300$

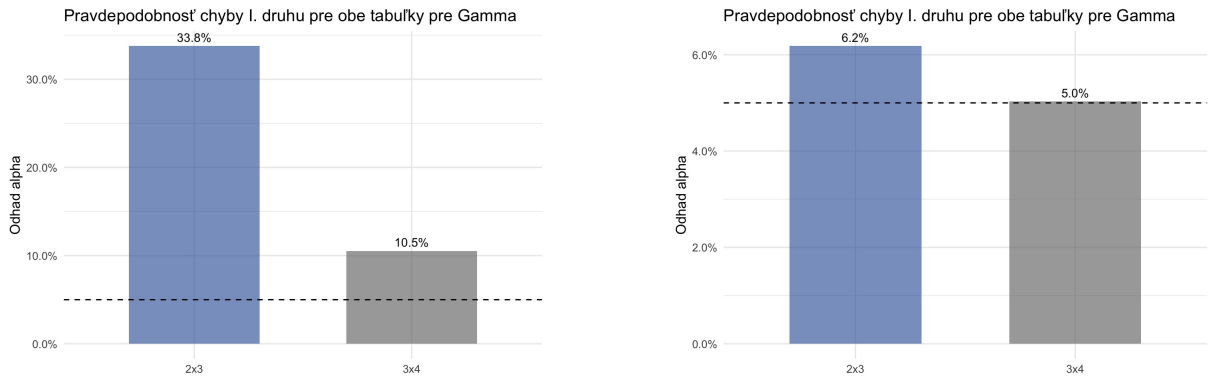
Obr. 28: Zlúčenie kategórií - rovnomerne rozdelené pravdepodobnosti

Na obrázku 28 vidíme dva grafy zobrazujúce odhad pravdepodobnosti chyby I. druhu

pre $n = 30$ a $n = 300$. Pri malom rozsahu výberu $n = 30$ vidíme väčší rozdiel medzi tabuľkami z pohľadu pravdepodobnosti chyby I. druhu. Po zlúčení na menšiu tabuľku chyba stúpa z 9,3% na 12,9%, síce $\gamma = 0$ aj po zlúčení kategórií. Je to možno spôsobené tým, že zlúčením buniek sa znižuje počet buniek v tabuľke a tým aj množstvo informácie o vzťahoch medzi premennými. To vedie k ešte menej presnej aproximácií jeho rozdelenia normálnym pri malých výberoch. V dôsledku toho má test tendenciu častejšie zamietať nulovú hypotézu, než by zodpovedalo nominálnej hladine významnosti 5%.

Pre vyšší rozsah výberu $n = 300$ sa rozdiel medzi tabuľkami výrazne zmenšuje, čo je spôsobené väčším počtom pozorovaní. Testová štatistika sa v tomto prípade lepšie približuje k svojmu asymptotickému normálnemu rozdeleniu, v dôsledku čoho sa pravdepodobnosť chyby I. druhu približuje k nominálnej hladine 5%.

Teraz namiesto tabuľky s rovnomernou pravdepodobnosťou zvolíme tabuľku s nerovnomernou pravdepodobnosťou, pričom nezávislosť v tabuľke naďalej platí. Výsledky simulácie pre oba rozsahy výberu sú zobrazené na obrázku 29.



(a) $n = 30$ (b) $n = 300$

Obr. 29: Zlúčenie kategórií - nerovnomerne rozdelené pravdepodobnosti

Z obrázka 29 pozorujeme, že pre $n = 30$ pravdepodobnosť chyby I. druhu rapídne stúpa v porovnaní s rovnomerne rozdelenou tabuľkou pravdepodobností. Nárast je spôsobený tým, že po zlúčení vzniká tabuľka s veľkou nevyrovnanosťou počtu jedincov - v prvom riadku je približne 90% jedincov a v druhom riadku je iba 10% jedincov. Táto nevyváženosť spôsobuje vyššiu variabilitu testovej štatistiky a zhoršuje presnosť asymptotickej aproximácie testovej štatistiky normálnym rozdelením. Naopak, pre tabuľku veľkosti 3×4 je pravdepodobnosť rozdelená medzi väčší počet buniek, čím je nevyváženosť v tabuľke

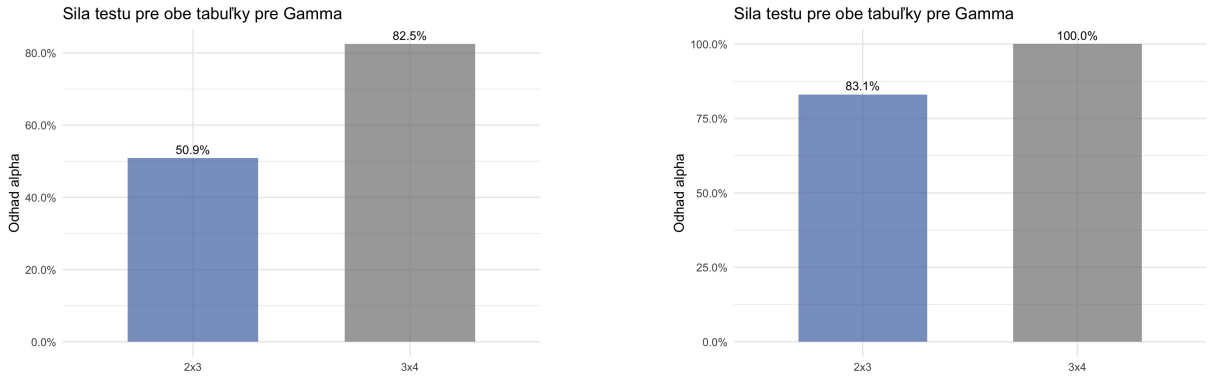
miernejšia a testová štatistika vykazuje normalitu viacej.

Relatívne výrazný rozdiel v odhade pravdepodobnosti chyby I. druhu vidíme aj v prípade vysokého počtu jedincov $n = 300$. To znamená, že po zlúčení kategórií vzniká natoľko nerovnomerné rozdelenie pravdepodobností, že ani pri tomto rozsahu výberu ešte nie je asymptotická aproximácia testovej štatistiky dostatočne presná. Na jej ustálenie by bol potrebný výrazne vyšší počet pozorovaní.

3.4.2 Za platnosti H_1

Pozrime sa na pravdepodobnostné tabuľky, kde H_0 o nezávislosti prestane platiť a budeme sledovať silu testu. Majme pravdepodobnostnú tabuľku:

$$\begin{pmatrix} 0.30 & 0.20 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.10 & 0.10 \\ 0.01 & 0.03 & 0.03 & 0.03 \end{pmatrix}.$$



(a) $n = 30$ (b) $n = 300$

Obr. 30: Zlúčenie kategórií - H_0 neplatí

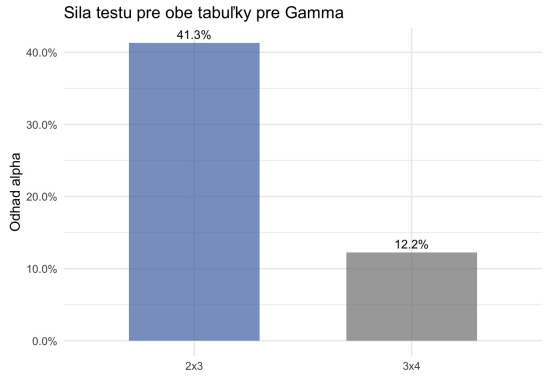
Na obrázku 30 je zobrazený výsledok simulácií. Tabuľka 3×4 má koeficient $\gamma = 0{,}591$. Po zlúčení má tabuľka veľkosti 2×3 koeficient γ hodnotu 0,421. Teda, závislosť medzi premennými jemne klesá, stále však ide o relatívne silnú pozitívnu závislosť.

Pre malý počet výberov $n = 30$ vidíme, že tabuľka s vyšším počtom kategórií má silu testu 82,5%, zatiaľ čo po zlúčení na tabuľku veľkosti 2×3 sila testu klesá na 50,9%. To je celkom veľký pokles, ktorý je spôsobený stratou informácie o závislosti medzi premennými. Test pracuje s menej výraznou štruktúrou tabuľky, čo vedie k poklesu sily testu.

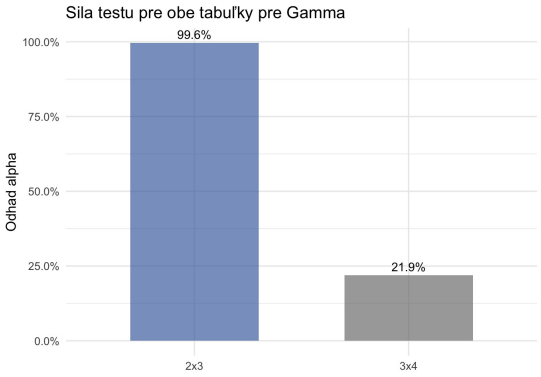
Pre vyšší počet jedincov $n = 300$ sila testu narastá pre prípad tabuľky veľkosti 3×4 až na 100% a pre tabuľku veľkosti 2×3 na 83,1%. Teda aj pri väčších výberoch zostáva sila testu po zlúčení kategórií nižšia, čo potvrdzuje stratu informácie spôsobenú zjednodušením tabuľky. Na druhej strane však aj zlúčená tabuľka dosahuje pomerne vysokú silu testu, čo hovorí o tom, že pri dostatočne veľkom rozsahu výberu je test stále schopný spoľahlivo zachytiť existujúcu závislosť medzi premennými.

Na druhej strane zvoľme tabuľku, ktorá po zlúčení bude mať závislosť vyššiu ako predtým, teda že koeficient γ bude vyšší. Zoberme tabuľku veľkosti 3×4, ktorá má koeficient $\gamma = 0.097$ a po zlúčení koeficient $\gamma = 0.415$:

$$\frac{1}{74}\begin{pmatrix} 4 & 3 & 1 & 6 \\ 4 & 20 & 6 & 1 \\ 7 & 3 & 16 & 3 \end{pmatrix}.$$



(a) $n = 30$



(b) $n = 300$

Obr. 31: Zlúčenie kategórií - H_0 neplatí

Výsledky na obrázku 31 ukazujú, že pre $n = 30$ aj tabuľka s nižším počtom kategórií, ale s podstatne silnejšou závislosťou, môže mať vyššiu hodnotu sily testu ako tabuľka s vyšším počtom kategórií, a teda aj väčšou šancou zachytiť závislosť. Pre výber $n = 300$ sa tento efekt ešte zvýrazňuje, keďže sila testu pre tabuľku veľkosti 2×3 dosahuje takmer 100 %, zatiaľ čo pre tabuľku 3×4 zostáva výrazne nižšia.

To naznačuje, že niekedy môže byť efekt zlúčenia kontraintuitívny. Zlúčením kategórií síce dochádza k strate informácie, avšak zároveň môže dôjsť k zosilneniu zachytenej závislosti, čo vedie k zvýšeniu sily testu.

4 Korelačné koeficienty $\hat{\rho}$ a $\hat{\rho}_b$

V tejto časti sa venujeme sledovaniu pravdepodobnosti chyby I. druhu pre dve miery asociácie, ktoré už nie sú založené na konkordantných a diskordantných pároch ako Kendallovo τ_b alebo Goodman-Kruskalova γ. Koeficienty korelácie Spearmanovo ρ a Kendallovo ρ_b pomocou ridit skóre označujú koreláciu, ktorá vznikne priradením pevných alebo poradových skóre hodnotám v riadkoch a stĺpcoch. Teda kategórie sa nahradia číslami a medzi nimi sa určí korelácia. Spearmanov korelačný koeficient bol pôvodne zavedený v práci [10] a riditové skóre bolo zavedené v [9]. Taktiež dokážeme prepojenie Spearmanovho korelačného koeficientu ρ s koeficientom Kendallovo τ_b a taktiež dokážeme jednu z vlastností riditových skóre. Pre túto kapitolu sme čerpali najmä z [1].

Koeficient $\hat{\rho}$ vychádza z Pearsonovho korelačného koeficientu, pričom ordinálnym kategóriám riadkov a stĺpcov v kontingenčnej tabuľke sú priradené číselné hodnoty (skóre), ktoré vyjadrujú ich poradie. Následne sa medzi týmito hodnotami vypočíta hodnota podobná Pearsonovmu korelačnému koeficientu.

Definícia 3. (Spearmanov korelačný koeficient $\hat{\rho}$) Nech $u_1 \leq u_2 \leq \cdots \leq u_r$ a $v_1 \leq v_2 \leq \cdots \leq v_c$ sú číselné skóre priradené hodnotám v riadkoch a stĺpcoch podľa ich poradia a $\hat{\pi}_{ij}$ označujú odhady pravdepodobností kontingenčnej tabuľky veľkosti $r \times c$. Potom miera asociácie Spearmanovo $\hat{\rho}$ je definovaná ako

$$\hat{\rho} = \frac{\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})\hat{\pi}_{ij}}{\sqrt{\sum_i (u_i - \bar{u})^2 \hat{\pi}_{i\cdot}} \sqrt{\sum_j (v_j - \bar{v})^2 \hat{\pi}_{\cdot j}}}, \tag{45}$$

kde $\bar{u} = \sum_i u_i \hat{\pi}_{i\cdot}$ a $\bar{v} = \sum_j v_j \hat{\pi}_{\cdot j}$ označujú priemerné hodnoty priradených skóre pre riadky a stĺpce.

Ak Spearmanovo $\hat{\rho}$ označuje výberovú koreláciu, ktorá vznikne pevným priradením skóre riadkom a stĺpcom, potom za platnosti H_0 o nezávislosti má testová štatistika $(n-1)\hat{\rho}^2$ asymptotické χ^2_1 rozdelenie.

Špeciálnym prípadom Spearmanovho koeficientu $\hat{\rho}$ je koeficient asociácie $\hat{\rho}_b$, ktorý využíva takzvané *riditové skóre*.

Definícia 4. (Kendallovo $\hat{\rho}_b$ pomocou ridit skóre) Nech $\hat{\pi}_{ij}$ označuje odhady pravdepodobností kontingenčnej tabuľky veľkosti $r \times c$. Riditové skóre pre premenné X a Y potom

sú

$$a_i^X = \sum_{k=1}^{i-1} \hat{\pi}_{k\cdot} + \frac{\hat{\pi}_{i\cdot}}{2}, \qquad a_j^Y = \sum_{k=1}^{j-1} \hat{\pi}_{\cdot k} + \frac{\hat{\pi}_{\cdot j}}{2},$$

kde $i = 1, ..., r$ a $j = 1, ..., c$. Následne sa $\hat{\rho}_b$ počíta ako

$$\hat{\rho}_b = \frac{\sum_i \sum_j (a_i^X - 0{,}50)(a_j^Y - 0,50)\,\hat{\pi}_{ij}}{\sqrt{\left[\sum_i (a_i^X - 0{,}50)^2 \hat{\pi}_{i+}\right]\left[\sum_j (a_j^Y - 0{,}50)^2 \hat{\pi}_{+j}\right]}}.$$

Za platnosti H_0 o nezávislosti má testová štatistika $(n-1)\hat{\rho}_b^2$ taktiež asymptotické χ_1^2 rozdelenie.

Vzorec pre $\hat{\rho}_b$ dostaneme zo vzorca (45), keďže vážené priemery $\overline{a^X}$ a $\overline{a^Y}$ sú presne 0,5.

4.1 Základné vlastnosti ridit skóre

Ukážeme, že pre každé marginálne rozdelenie platí, že vážený priemer riditových skóre je rovný 0,5.

Nech $\hat{\pi}_{i\cdot}$ sú odhadnuté riadkové pravdepodobnosti. Ridit skóre pre i-ty riadok definujeme ako

$$a_i^X = \sum_{k=1}^{i-1} \hat{\pi}_{k\cdot} + \frac{1}{2}\hat{\pi}_{i\cdot}, \quad i = 1, \ldots, r.$$

Vážený priemer ridit skóre potom má tvar

$$\bar{a}^X = \sum_{i=1}^{r} a_i^X \hat{\pi}_{i\cdot}.$$

Vážený priemer ridit skóre má tento tvar, pretože každé skóre a_i^X sa berie s váhou zodpovedajúcou odhadnutej pravdepodobnosti $\hat{\pi}_{i+}$, s akou sa daný riadok vyskytuje. Dosadením dostávame

$$\bar{a}^X = \sum_{i=1}^{r} \left(\sum_{k=1}^{i-1} \hat{\pi}_{k\cdot} + \frac{1}{2}\hat{\pi}_{i\cdot} \right) \hat{\pi}_{i\cdot}.$$

Rozdelíme sumu na dve časti:

$$\bar{a}^X = \sum_{i=1}^{r} \sum_{k=1}^{i-1} \hat{\pi}_{k\cdot}\hat{\pi}_{i\cdot} + \frac{1}{2} \sum_{i=1}^{r} \hat{\pi}_{i\cdot}^2. \tag{46}$$

Všimnime si, že

$$\sum_{i=1}^{r} \sum_{k=1}^{i-1} \hat{\pi}_{k\cdot}\hat{\pi}_{i\cdot} = \sum_{i>k} \hat{\pi}_{i\cdot}\hat{\pi}_{k\cdot}.$$

Zároveň platí

$$\sum_{i=1}^{r} \sum_{k=1}^{r} \hat{\pi}_{i\cdot}\hat{\pi}_{k\cdot} = \left(\sum_{i=1}^{r} \hat{\pi}_{i\cdot} \right)^2 = 1^2 = 1.$$

Následne túto dvojitú sumu rozdelíme na diagonálnu a mimo-diagonálnu časť:

$$1 = \sum_{i=1}^{r}\sum_{k=1}^{r} \hat{\pi}_{i\cdot}\hat{\pi}_{k\cdot} = \sum_{i=1}^{r} \hat{\pi}_{i\cdot}^2 + \sum_{i\neq k} \hat{\pi}_{i\cdot}\hat{\pi}_{k\cdot}. \tag{47}$$

Keďže platí symetria, tak

$$\sum_{i\neq k} \hat{\pi}_{i\cdot}\hat{\pi}_{k\cdot} = 2\sum_{i>k} \hat{\pi}_{i\cdot}\hat{\pi}_{k\cdot},$$

dostávame z (47):

$$\sum_{i=1}^{r} \hat{\pi}_{i\cdot}^2 + 2\sum_{i>k} \hat{\pi}_{i\cdot}\hat{\pi}_{k\cdot} = 1.$$

Odtiaľ dostávame

$$\sum_{i>k} \hat{\pi}_{i\cdot}\hat{\pi}_{k\cdot} = \frac{1 - \sum_{i=1}^{r} \hat{\pi}_{i\cdot}^2}{2}.$$

Dosadením späť do výrazu (46) pre $\bar{a}^X$:

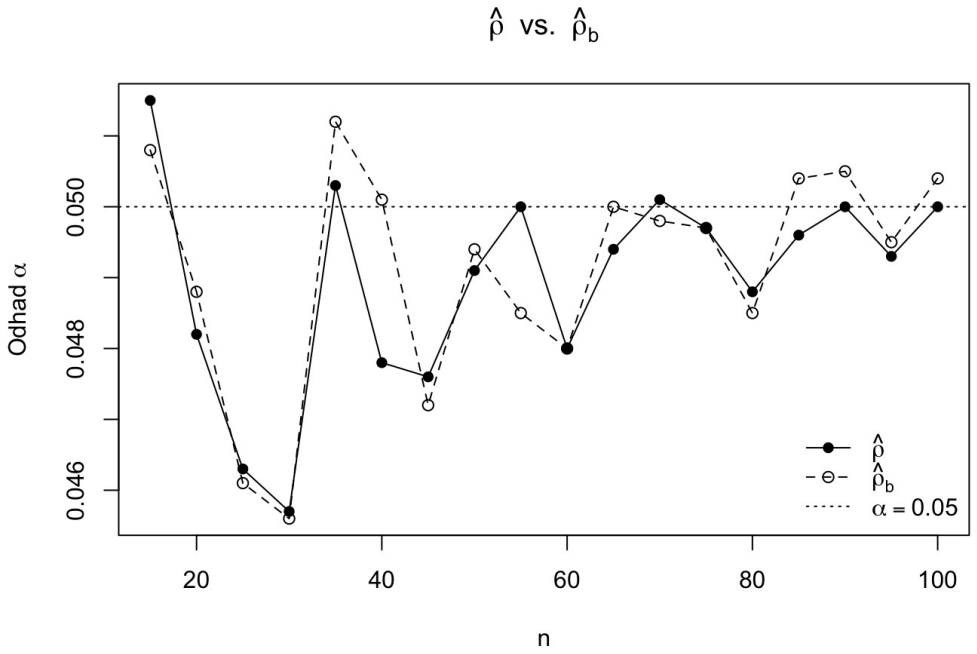
$$\bar{a}^X = \frac{1 - \sum_{i=1}^{r} p_{i+}^2}{2} + \frac{1}{2}\sum_{i=1}^{r} p_{i+}^2 = \frac{1}{2}.$$

Teda platí, že priemer ridit skóre pre marginálne rozdelenie je rovný 0,5.

4.2 Simulácie za platnosti H_0

Pre tieto dve miery asociácie sme sledovali pomocou simulácie pravdepodobnosť chyby I. druhu a ako sa odlišuje od nominálnej hodnoty 5%. Pre Pearsonov koeficient $\hat{\rho}$ sme zvolili $u_1 \leq u_2 \leq \cdots \leq u_r$ a $v_1 \leq v_2 \leq \cdots \leq v_c$ ako $1 < 2 < 3$ a $1 < 2 < 3 < 4$, čím sme dostali Spearmanov korelačný koeficient. Simulácie sme spustili 10 000-krát. Počas každej simulácie sa vytvorí tabuľka dát pomocou multinomického rozdelenia a rovnomerne rozdelenej pravdepodobnostnej tabuľky veľkosti 3×4, ktorá spĺňa nulovú hypotézu o nezávislosti. Následne sa vypočíta a vyhodnotí testová štatistika pre obe miery $\hat{\rho}$ aj $\hat{\rho}_b$. Počet dát n sme zvolili postupne z množiny $\{10, 15, 20, ..., 100\}$. Výsledky sú zobrazené na obrázku 32 a konkrétne hodnoty sú uvedené v tabuľke 3.

Na obrázku 32 vidíme výsledky testov založených na $\hat{\rho}$ (plná čiara) a výsledky testov založených na $\hat{\rho}_b$ (čiarkovaná čiara). Z obrázka vidíme, že odhady pravdepodobnosti chyby I. druhu pre obe miery oscilujú okolo nominálnej hladiny významnosti $\alpha = 5\%$, pričom rozdiely medzi nimi sú minimálne. Riditová verzia $\hat{\rho}_b$ sa od klasickej $\hat{\rho}$ líši len mierne a nedodáva výrazné zlepšenie z hľadiska dodržania nominálnej hladiny významnosti ani pri malom počte dát n.



Obr. 32: Pravdepodobnosť chyby I. druhu pre test asociácie pomocou $\hat{\rho}$ a riditovej verzie $\hat{\rho}_b$

n	$\hat{\alpha}_{\hat{\rho}}$	$\hat{\alpha}_{\hat{\rho}_b}$	n	$\hat{\alpha}_{\hat{\rho}}$	$\hat{\alpha}_{\hat{\rho}_b}$
15	0.052	0.051	60	0.048	0.048
20	0.048	0.049	65	0.049	0.050
25	0.046	0.046	70	0.050	0.050
30	0.046	0.046	75	0.050	0.050
35	0.050	0.051	80	0.049	0.048
40	0.048	0.050	85	0.050	0.050
45	0.048	0.047	90	0.050	0.051
50	0.049	0.049	95	0.049	0.050
55	0.050	0.048	100	0.050	0.050

Tabuľka 3: Pravdepodobnosť chyby I. druhu pre $\hat{\rho}$ a $\hat{\rho}_b$ - hodnoty

4.3 Vzťah medzi Kendallovým τ_b a Spearmanovou, resp. Pearsonovou koreláciou pre 2×2 tabuľku

Ukážeme, že pre 2×2 kontingenčnú tabuľku je Kendallov koeficient $\hat{\tau}_b$ identický s Pearsonovým korelačným koeficientom, ktorý možno vypočítať, ak sú hodnoty $X_1 < X_2$ a $Y_1 < Y_2$ číselné.

Uvažujme kontingenčnú tabuľku, kde namiesto značenia n_{11}, n_{12}, n_{21}, n_{22} z kapitoly 1 používame značenie a, b, c, d.

	$Y_1 = 0$	$Y_2 = 1$	súčet
$X_1 = 0$	a	b	$a + b$
$X_2 = 1$	c	d	$c + d$
súčet	$a + c$	$b + d$	n

Hodnoty možno bez újmy na všeobecnosti nastaviť na 0 a 1, lebo výberový Pearsonov korelačný koeficient je invariantný na lineárne transformácie, čiže odčítanie konštanty a predelenie konštantou. Počet konkordantných a diskordantných párov je daný vzťahmi

$$C = ad, \qquad D = bc.$$

Pearsonov korelačný koeficient

Pre binárne premenné $X, Y \in \{0, 1\}$ platí, že ich priemery môžeme vyjadriť pomocou počtu jedincov v kontingenčnej tabuľke.

Premenná X nadobúda hodnotu 1 práve vtedy, keď pozorovanie patrí do druhého riadku tabuľky. Preto

$$\bar{X} = \frac{a \cdot 0 + b \cdot 0 + c \cdot 1 + d \cdot 1}{n} = \frac{c + d}{n}.$$

Podobne premenná Y nadobúda hodnotu 1, ak pozorovanie patrí do druhého stĺpca, teda

$$\bar{Y} = \frac{a \cdot 0 + b \cdot 1 + c \cdot 0 + d \cdot 1}{n} = \frac{b + d}{n}.$$

Nakoniec súčin XY nadobúda hodnotu 1 iba v prípade, že súčasne platí $X = 1$ a $Y = 1$, čo zodpovedá bunke $(1, 1)$ s početnosťou d. Preto

$$\overline{XY} = \frac{a \cdot 0 \cdot 0 + b \cdot 0 \cdot 1 + c \cdot 1 \cdot 0 + d \cdot 1 \cdot 1}{n} = \frac{d}{n}.$$

Výberová kovariancia:

$$\widehat{\mathrm{Cov}}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \overline{XY} - \bar{X}\,\bar{Y} = \frac{d}{n} - \frac{(b+d)(c+d)}{n^2}.$$

Výberové rozptyly:

$$\widehat{\mathrm{Var}}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Keďže $X = 0$ alebo $X = 1$ platí $X = X^2$. Preto

$$\widehat{\mathrm{Var}}(X) = \bar{X} - (\bar{X})^2 = \bar{X}(1 - \bar{X}) = \frac{c+d}{n}\left(1 - \frac{c+d}{n}\right).$$

Úpravou výrazu v zátvorke:

$$1 - \frac{c+d}{n} = \frac{n-(c+d)}{n} = \frac{a+b}{n},$$

preto

$$\widehat{\mathrm{Var}}(X) = \frac{c+d}{n}\cdot\frac{a+b}{n} = \frac{(a+b)(c+d)}{n^2}.$$

Analogicky pre premennú Y:

$$\widehat{\mathrm{Var}}(Y) = \bar{Y}(1 - \bar{Y}) = \frac{(a+c)(b+d)}{n^2}.$$

Pearsonov korelačný koeficient má tvar

$$\hat{\rho} = \frac{\widehat{\mathrm{Cov}}(X,Y)}{\sqrt{\widehat{\mathrm{Var}}(X)\widehat{\mathrm{Var}}(Y)}}.$$

Dosadením do vzorca dostávame:

$$\hat{\rho} = \frac{\frac{dn-(c+d)(b+d)}{n^2}}{\sqrt{\frac{(a+b)(c+d)(a+c)(b+d)}{n^2}}} = \frac{dn-(c+d)(b+d)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Napokon upravíme čitateľ:

$$dn - (c+d)(b+d) = d(a+b+c+d) - (bc+bd+cd+d^2) = ad - bc.$$

Teda výsledný tvar:

$$\hat{\rho} = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Tento tvar Pearsonovho $\hat{\rho}$ sa zhoduje so Spearmanovým $\hat{\rho}$ uvedeným vo vzťahu (45) (pri voľbe $u_1 = 1$, $u_2 = 2$, $v_1 = 1$, $v_2 = 2$), ktoré by sme použili v prípade, že premenné

$X_1 < X_2$ a $Y_1 < Y_2$ nie sú číselné. Táto rovnosť vyplýva z toho, že hodnoty u_1, u_2 a v_1, v_2 možno bez ujmy na všeobecnosti zvoliť ako $u_1 = 0$, $u_2 = 1$ a $v_1 = 0$, $v_2 = 1$.

Kendallov koeficient $\hat{\tau}_b$ Pre Kendallov koeficient $\hat{\tau}_b$ z definície 1 sa dá prepísať ako

$$\hat{\tau}_b = \frac{C - D}{\sqrt{\left(\binom{n}{2} - T_X\right)\left(\binom{n}{2} - T_Y\right)}},$$

kde C je počet konkordantných párov, D počet diskordantných párov, T_X a T_Y počty viazaných párov podľa premenných X a Y, čiže

$$T_X = \sum_{i=1}^{r} \frac{n_{i\cdot}(n_{i\cdot} - 1)}{2} \quad \text{a} \quad T_Y = \sum_{j=1}^{c} \frac{n_{\cdot j}(n_{\cdot j} - 1)}{2}.$$

V prípade tabuľky veľkosti 2×2 na začiatku tejto podkapitoly 4.3 platí

$$T_X = \frac{n_{1\cdot}(n_{1\cdot} - 1)}{2} + \frac{n_{2\cdot}(n_{2\cdot} - 1)}{2} = \frac{(a+b)(a+b-1)}{2} + \frac{(c+d)(c+d-1)}{2}.$$

Analogicky

$$T_Y = \frac{(a+c)(a+c-1)}{2} + \frac{(b+d)(b+d-1)}{2}.$$

Najprv upravíme výraz $\binom{n}{2} - T_X$:

$$\binom{n}{2} - T_X = \frac{n(n-1)}{2} - \frac{(a+b)(a+b-1)}{2} - \frac{(c+d)(c+d-1)}{2}.$$

Keďže $n = (a+b) + (c+d)$, označme

$$M := a + b, \quad N := c + d.$$

Potom

$$\binom{n}{2} = \frac{[(a+b) + (c+d)][(a+b) + (c+d) - 1]}{2} = \frac{(M+N)(M+N-1)}{2}.$$

Dosadením dostávame

$$\binom{n}{2} - T_X = \frac{(M+N)(M+N-1) - M(M-1) - N(N-1)}{2}.$$

Po roznásobení a upravení výrazu dostávame

$$\binom{n}{2} - T_X = MN = (a+b)(c+d).$$

Analogicky

$$\binom{n}{2} - T_Y = (a+c)(b+d).$$

Keďže pre tabuľky veľkosti 2×2 platí $C = ab$ a $D = cd$, po dosadení do vzorca pre $\hat{\tau}_b$ dostávame

$$\hat{\tau}_b = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Táto vlastnosť ukazuje, že pre binárne premenné sú $\hat{\tau}_b$ a Pearsonovo $\hat{\rho}$ ekvivalentné a vedú k rovnakému hodnoteniu závislosti.

5 Správanie testov podľa jednotlivých mier asociácie

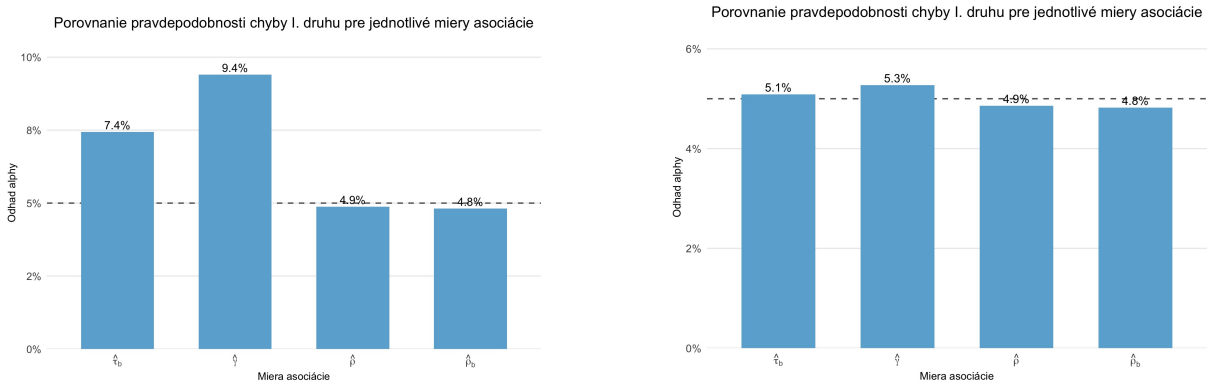
V poslednej kapitole sa venujeme porovnaniu pravdepodobnosti chyby I. druhu testov podľa všetkých doteraz rozoberaných mier asociácie. Pre každú mieru sme realizovali 10 000 simulácií, v ktorých sme sledovali odhad α pri menšom aj väčšom rozsahu výberu, ako aj pre rôzne pravdepodobnostné tabuľky v závislosti od platnosti hypotézy H_0.

Tabuľka výberov bola generovaná ako jeden multinomický výber, čomu zodpovedá aj použitý odhad variancie a štandardnej odchýlky pre miery $\hat{\tau}_b$ a $\hat{\gamma}$.

V každej simulácii sa následne počítali testové štatistiky: pre $\hat{\tau}_b$ štatistika $\frac{\hat{\tau}_b}{SE_{\hat{\tau}_b}}$, pre $\hat{\gamma}$ štatistika $\frac{\hat{\gamma}}{SE_{\hat{\gamma}}}$, pre Spearmanovo $\hat{\rho}$ štatistika $(n-1)\hat{\rho}^2$ a pre $\hat{\rho}_b$ štatistika $(n-1)\hat{\rho}_b^2$.

5.1 Za platnosti H_0

Najprv sme použili rovnomernú tabuľku pravdepodobností, pre ktorú platí hypotéza H_0 o nezávislosti. Obrázok 33 zobrazuje výsledky simulácií a zachytáva, ako sa líši pravdepodobnosť chyby I. druhu pre jednotlivé miery asociácie.



(a) $n = 30$ (b) $n = 300$

Obr. 33: Rovnomerne rozdelená pravdepodobnostná tabuľka

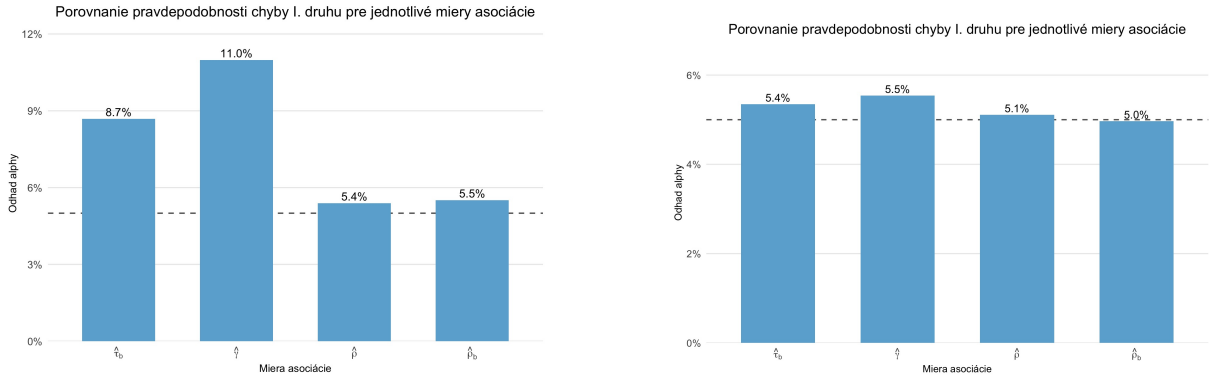
Z obrázka 33 vidíme, že výraznejšie rozdiely sa prejavujú pri menšom rozsahu výberu. Najvyššiu pravdepodobnosť chyby I. druhu dosahuje test založený na koeficiente $\hat{\gamma}$, približne 9,4 %. Druhú najvyššiu hodnotu má test pre koeficient $\hat{\tau}_b$ s pravdepodobnosťou približne 7,4 %. Naopak, testy založené na koeficientoch Spearmanovo $\hat{\rho}$ a $\hat{\rho}_b$ sa aj pri menších rozsahoch výberu pohybujú blízko nominálnej hladiny významnosti 5 %.

Rozdiely medzi jednotlivými mierami pri menších rozsahoch výberu sú spôsobené

najmä odlišnými vlastnosťami ich asymptotických aproximácií. Koeficient $\hat{\gamma}$ ignoruje viazané páry, čo vedie k horšej presnosti normálnej aproximácie, a tým aj k zvýšenej pravdepodobnosti chyby I. druhu. Koeficient $\hat{\tau}_b$ síce viazané páry zohľadňuje, avšak jeho rozdelenie sa pri malých výberoch ešte dostatočne nepribližuje asymptotickému.

Naopak, testy založené na koeficientoch Spearmanovo $\hat{\rho}$ a $\hat{\rho}_b$ využívajú chí-kvadrát aproximáciu, ktorá sa ukazuje byť stabilnejšia aj pri menších rozsahoch výberu, čo vedie k lepšiemu dodržaniu nominálnej hladiny významnosti. S rastúcim rozsahom výberu sa rozdiely medzi jednotlivými mierami postupne zmenšujú.

Vezmime o niečo netriviálnejšiu tabuľku pravdepodobností, a to pravdepodobnostnú tabuľku s nerovnomernou pravdepodobnosťou, zatiaľ čo hypotéza H_0 stále platí. Parametre simulácie sa nemenia a výsledky sú zobrazené na obrázku 34.



(a) $n = 30$ (b) $n = 300$

Obr. 34: Nerovnomerne rozdelená pravdepodobnostná tabuľka

V porovnaní s predchádzajúcou simuláciou si môžeme na obrázku 34 všimnúť navýšenie pravdepodobnosti chyby I. druhu, najmä pre koeficienty $\hat{\tau}_b$ a $\hat{\gamma}$. Zvýšenie pravdepodobnosti chyby I. druhu pri nerovnomerne rozdelenej pravdepodobnostnej tabuľke je spôsobené zhoršenou presnosťou asymptotických aproximácií testových štatistík. Aj keď hypotéza H_0 o nezávislosti platí, nerovnomerné rozdelenie vedie k nízkemu počtu jedincov v niektorých bunkách, čo zhoršuje aproximáciu testovej štatistiky normálnym rozdelením.

Tento efekt sa prejavuje najmä pri koeficientoch $\hat{\gamma}$ a $\hat{\tau}_b$, ktoré sú citlivejšie na nízke početnosti v tabuľke výberov a nerovnomernosť pravdepodobnostnej tabuľky. Naopak, testy založené na koeficientoch Spearmanovo $\hat{\rho}$ a $\hat{\rho}_b$ sa ukazujú byť robustnejšie a aj pri menších rozsahoch výberu lepšie dodržiavajú nominálnu hladinu významnosti. S rastúcim

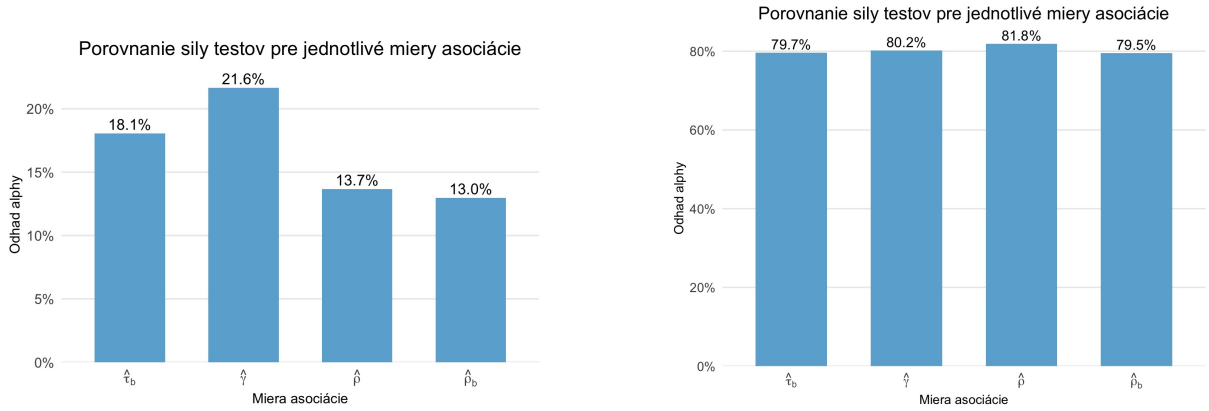
rozsahom výberu sa rozdiely medzi jednotlivými mierami postupne zmenšujú.

5.2 Za platnosti H_1

Nakoniec sme zvolili pravdepodobnostnú tabuľku

$$\begin{pmatrix} 0.04 & 0.028 & 0.018 & 0.014 \\ 0.075 & 0.095 & 0.075 & 0.055 \\ 0.12 & 0.147 & 0.177 & 0.156 \end{pmatrix},$$

ktorá porušuje hypotézu H_0 o nezávislosti, a pomocou simulácií sme sledovali silu testov pre všetky uvažované miery asociácie $\hat{\tau}_b$, $\hat{\gamma}$, $\hat{\rho}$ a $\hat{\rho}_b$. Výsledky pre túto pravdepodobnostnú tabuľku sú zobrazené na obrázku 35.



(a) $n = 30$ (b) $n = 300$

Obr. 35: Hypotéza H_0 neplatí

Z obrázka 35 môžeme pozorovať, že pri porušení hypotézy H_0 dosahujú všetky testy pomerne vysokú silu. Pri menšom rozsahu výberu ($n = 30$) sú však medzi jednotlivými mierami výraznejšie rozdiely. Najvyššiu silu dosahuje test založený na koeficiente $\hat{\gamma}$, zatiaľ čo najnižšiu silu majú testy založené na koeficientoch Spearmanovo $\hat{\rho}$ a $\hat{\rho}_b$.

S rastúcim rozsahom výberu ($n = 300$) sa sila všetkých testov výrazne zvyšuje a rozdiely medzi jednotlivými mierami sa zmenšujú. V tomto prípade už všetky testy dosahujú silu približne 80 %, pričom najvyššiu hodnotu dosahuje test založený na Spearmanovom koeficiente $\hat{\rho}$. Treba však dodať, že testy založené na $\hat{\gamma}$ (aj $\hat{\tau}_b$) majú aj najvyššiu pravdepodobnosť chyby I. druhu, čiže sa celkovo správajú liberálnejšie.

Tieto výsledky naznačujú, že pri menších rozsahoch výberu môže byť test založený na koeficiente $\hat{\gamma}$ citlivejší na zachytenie závislosti medzi premennými. Dôvodom je najmä to, že koeficient γ je definovaný iba na základe konkordantných a diskordantných párov a ignoruje viazané páry.

Testy založené na koeficientoch Spearmanovo $\hat{\rho}$ a $\hat{\rho}_b$ vychádzajú z Pearsonovej korelácie medzi číselnými skóre priradenými ordinálnym kategóriám a využívajú chí-kvadrát aproximáciu testovej štatistiky. Tieto testy sa ukazujú byť robustnejšie, avšak pri menších rozsahoch výberu môžu mať nižšiu silu v porovnaní s testom založeným na $\hat{\gamma}$.

S rastúcim rozsahom výberu sa rozdiely medzi jednotlivými mierami postupne zmenšujú všetky testy dosahujú podobnú silu.

Záver

V tejto diplomovej práci sme sa zaoberali mierami asociácie pre ordinálne dáta, a to Kendallovým $\hat{\tau}_b$, Goodman–Kruskalovou $\hat{\gamma}$ a korelačnými koeficientmi $\hat{\rho}$ a Kendallovo $\hat{\rho}_b$ pomocou ridit skóre. Cieľom práce bolo preskúmať správanie pravdepodobnosti chyby I. druhu za platnosti nezávislosti, ako aj silu testov založených na normálnom a chí-kvadrát rozdelení.

V práci sme odvodili varianciu pre jeden multinomický výber pre koeficienty $\hat{\tau}_b$ a $\hat{\gamma}$. Pri odvádzaní variancie pre $\hat{\tau}_b$ sme narazili na chybu vo vzorci uvedenom v [1], konkrétne v kritickej premennej vstupujúcej do výpočtu variancie. Po následnej komunikácii s autorom bola táto chyba zohľadnená v korekcii knihy, čo predstavuje jeden z hlavných prínosov tejto práce.

Simulácie sme realizovali pre všetky štyri miery asociácie pri dvoch veľkostiach výberu, $n = 30$ a $n = 300$, aby sme mohli porovnať správanie testov pri malých a väčších výberoch a lepšie vyhodnotiť platnosť asymptotiky. Pre $\hat{\tau}_b$ a $\hat{\gamma}$ sme navyše skúmali vplyv použitia nesprávnej smerodajnej odchýlky vzhľadom na povahu dát a analyzovali sme relatívnu chybu v súvislosti s výsledkami testov. Okrem toho sme sledovali aj vplyv zlučovania kategórií sledovaných premenných, teda situáciu, keď sa mení rozdelenie informácie o závislosti medzi premennými.

Na základe simulácií sme ukázali, že použitie nesprávnej smerodajnej odchýlky vedie k vyššej sile testu, avšak za cenu zvýšenej pravdepodobnosti chyby I. druhu. Tento efekt je výraznejší pri menších výberoch, kde asymptotické vlastnosti testov ešte neplatia dostatočne dobre. Výsledky tak poukazujú na to, že nesprávna špecifikácia variancie vedie k systematickému skresleniu testovej štatistiky a môže viesť k nesprávnym záverom.

Ďalej sme ukázali, že zlučovanie kategórií sledovaných premenných ovplyvňuje pravdepodobnosť chyby I. druhu, pričom tento vplyv závisí od toho, či sa týmto krokom závislosť medzi premennými zosilní alebo naopak oslabí. Zlučovanie zároveň vedie k strate informácie, čo sa prejavuje aj na stabilite odhadov.

Prínos práce spočíva najmä v prepojení teoretických odvodení so simuláciami, v odhalení chyby v existujúcej literatúre a v detailnej analýze správania testov pri rôznych podmienkach, najmä pri malých výberoch, ktoré sú v praxi často problematické.

Výsledky práce zdôrazňujú potrebu opatrnosti pri aplikácii asymptotických testov na

malé výbery a poukazujú na význam správnej špecifikácie modelových predpokladov pri testovaní nezávislosti v ordinálnych dátach.

Zoznam použitej literatúry

[1] Agresti, A.: *Analysis of Ordinal Categorical Data*, Second Edition, John Wiley & Sons, New Jersey, 2010.

[2] Agresti, A.: *Ordinal Data Methods*, dostupné na: https://alanagresti.com/ordinal/ord.html (prístup 2.5.2026).

[3] Agresti, A.: *The Effect of Category Choice on Some Ordinal Measures of Association*, Journal of the American Statistical Association, 71 (353), 1976, 49–55.

[4] Goodman, L. A., Kruskal, W. H.: *Measures of Association for Cross Classifications, I*, Journal of the American Statistical Association, 49, 1954, 732–764.

[5] Goodman, L. A., Kruskal, W. H.: *Measures of Association for Cross Classifications, II: Further Discussion and References*, Journal of the American Statistical Association, 54, 1959, 123–163.

[6] Goodman, L. A., Kruskal, W. H.: *Measures of Association for Cross Classifications, III: Approximate Sampling Theory*, Journal of the American Statistical Association, 58, 1963, 310–364.

[7] Goodman, L. A., Kruskal, W. H.: *Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances*, Journal of the American Statistical Association, 67 (338), 1972, 415-421.

[8] Kendall, M. G.: *The Treatment of Ties in Rank Problems*, Biometrika, 33, 1945, 239–251.

[9] Kendall, M. G.: *Rank Correlation Methods*, 4th Edition, Charles Griffin, London, 1970.

[10] Spearman, C.: *The Proof and Measurement of Association Between Two Things*, American Journal of Psychology, 15, 1904, 72–101.