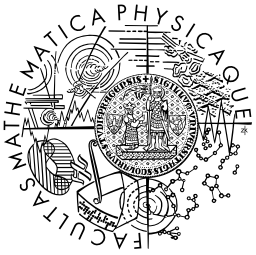

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Miroslava Gažová

Testy nezávislosti v kontingenčných tabuľkách

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D.

Studijní program: Obecná matematika

Praha 2022

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

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

V ..... dne ....

Podpis autora

Touto cestou by som sa chcela poďakovať doc. Ing. Marekovi Omelkovi, Ph.D. za trpezlivosť, odbornú pomoc, rýchlu komunikáciu a predovšetkým za nadšenie, ktoré vkladá do svojej práce. Ďalej by som rada poďakovala rodičom a priateľovi Dominikovi, ktorý mi poskytol maximálnu podporu pri písaní práce.
Ďakujem Bohu za vedenie a príležitosti, ktoré mi dáva.

Název práce: Testy nezávislosti v kontingenčných tabuľkách

Autor: Miroslava Gažová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Táto práca sa zaoberá problémom testovania nezávislosti dvoch diskrétnych náhodných veličín. Najprv definujeme kontingenčnú tabuľku a základné značenia v kontexte testov nezávislosti. Popíšeme najčastejšie používané testy v tejto oblasti. Následne predstavíme USP test nezávislosti, ktorý bol prvý krát predstavený autormi T.B.Berrett a R.J.Samworth (2021). V ďalšej kapitole sa podrobnejšie zameriame na štvorpoľné kontingenčné tabuľky a s nimi súvisiaci problém testovania zhody parametrov z dvoch nezávislých binomických rozdelení. Na konci aplikujeme testy na reálne dáta s využitím prostredia R.

Klíčová slova: chi-kvadrát test nezávisosti, kontingenčné tabuľky, štvorpoľné tabuľky, testy nezávislosti, USP test

Title: Tests of independence in contingency tables

Author: Miroslava Gažová

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis deals with the problem of independence testing between two discrete random variables. At first, we define contingency table and the basic notations in the context of independence tests. We describe the most commonly used tests in this field. Next, we present the U-statistics Permutation test of independence (USP), which was first presented by authors T.B.Berrett and R.J.Samworth (2021). In the next section, we focus in better detail on fourfold contingency tables and corresponding problem of testing for the equality of parameters from two independent binomial distributions. In the end, we apply the tests on the real data using the R enviroment.

Keywords: contingency tables, fourfold tables, chi-squared test, tests of independence, USP test

Obsah

Úvod

Testovanie nezávislosti patrí medzi základné štatistické otázky. V práci sa omezíme na testovanie nezávislosti dvoch diskrétnych náhodných veličín, ktoré sú kategoriálne. Testy nezávislosti nachádzajú využitie v mnohých oblastiach, ako sú napríklad vedecké a lekárske štúdie alebo finančné analýzy. V práci podrobne popíšeme USP test nezávislosti v kontingenej tabuľke, ktorého autormi sú T.B.Berrett a R.J.Samworth (2021). USP test budeme porovnávať najmä s Pearsonovým χ^2-testom, ktorý sa v situácii testovania nezávislosti používa najčastejšie. V práci budeme čerpať hlavne z článkov Berrett a Samworth (2021) a Berrett a kol. (2021b).

V prvej kapitole predstavíme Pearsonov χ^2-test nezávislosti a test pomerom vierohodností (G-test). Uvedieme ich základné charakteristiky spolu s jednoduchými príkladmi, na ktorých ilustrujeme ich použitie. Testy dáme do súvislosti s χ^2-*divergenciou*, resp. *Kullbackovou-Leiblerovou divergencou*. Následne odvodíme vzťah testových statistík týchto tvoch testov. Potom predstavíme permutačný test nezávislosti v kontingenčných tabuľkách - USP test. V tejto kapitole uvedieme jeho základné charakteristiky spolu s popisom realizácie testu.

V druhej kapitole sa zameriame na štvorpoľné kontingenčné tabuľky, ktoré vznikli ako špeciálny prípad kontingenčnej tabuľky veľkosti 2×2. V tejto situácii budeme kontingenčnú tabuľku interpretovať ako výber dvoch nezávislých binomických rozdelení $Bi(n,p_1), Bi(m,p_2)$ a následne sa budeme zaoberať problematikou testovania zhody parametrov p_1 a p_2. Testovú statistiku χ^2-testu a USP testu prepíšeme pomocou odhadov parametrov p_1 a p_2 a nájdeme ich asymptotické rozdelenia za nulovej hypotézy. Na záver kapitoly predstavíme Fisherov faktoriálový test pre štvorpoľné kontingenčné tabuľky, ktorý spolu s ostatnými testami aplikujeme na reálnych dátach, pochádzajúcich z určitej lekárskej štúdie.

1. Testy nezávislosti

1.1 Kontingenčné tabuľky

V mnohých situáciách sa stretneme s otázkou či sú na seba dva javy závislé. Môže nás zaujímať či závisí výskyt cukrovky od množstva cukru v strave, alebo či závisí výška platu od dosiahnutého vzdelania. Napozorované údaje z nášho náhodného výberu môžeme zostaviť do tabuľky, ktorú nazývame *kontingenčná tabuľka.*

Uvažujme náhodný vektor $(X,Y)^T$ s diskrétnym rozdelením, pričom X nadobúda hodnoty i z množiny $\{1,\dots,I\}$ a Y nadobúda hodnoty j z množiny $\{1,\dots,J\}$. Obe náhodné veličiny X, Y sú kategoriálne. Kategoriálne náhodné veličiny nadobúdajú hodnoty z daných kategórií, ktoré nemusia byť nutne číselné. Napríklad pri pozorovaní farby očí môžeme pozorované dáta rozdeliť do kategórií *„modrá“, „zelená“, „hnedá“*. Ďalej uvažujme náhodný výber $(X_1,Y_1)^T,\dots,(X_N,Y_N)^T$ o rozsahu N, kde $(X_k,Y_k)^T$ sú rovnako rozdelené ako $(X,Y)^T$ pre všetky $k \in \{1,\dots,N\}$.

Častokrát pracujeme s dátami, ktoré boli dopredu rozdelené do kategórií aj keď pochádzajú zo spojitého rozdelenia. Ako príklad uvedieme náhodnú veličinu udávajúcu výšku IQ v populácii. Napozorované hodnoty môžu byť zaradené do kategórií *„<60“,„61–80“ , „81–100“, „101–120“, „121–140“, „141>“*.
Označme počet pozorovaní zaradených do i-tej kategórie náhodnej veličiny X a j-tej kategórie náhodnej veličiny Y ako o_{ij}, tj.

$$o_{ij} = \sum_{k=1}^{N} 1\{X_k = i, Y_k = j\}. \tag{1.1}$$

Hodnoty o_{ij} budeme nazývať *pozorované četnosti.* Môžeme ich zoradiť do tabuľky, ktorú budeme nazývať *kontingenčná tabuľka* (viz Tabuľka 1.1).

	$Y=1$	...	$Y=J$	$\sum$
$X=1$	o_{11}	...	o_{1J}	o_{1+}
$X=2$	o_{21}	...	o_{2J}	o_{2+}
...	...	...	...	...
$X=I$	o_{I1}	...	o_{IJ}	o_{I+}
$\sum$	o_{+1}	...	o_{+J}	N

Tabulka 1.1: Kontingenčná tabuľka

Hodnoty o_{i+} v kontingenčnej tabuľke označujú počet pozorovaní v i-tom riadku, hodnoty o_{+j} počet pozorovaní v j-tom stĺpci. V literatúre sú často nazývané aj ako *marginálne četnosti* a sú definované ako:

$$o_{i+} = \sum_{j=1}^{J} o_{ij}, \qquad o_{+j} = \sum_{i=1}^{I} o_{ij},$$

pričom

$$\sum_{i=1}^{I} o_{i+} = \sum_{j=1}^{J} o_{+j} = \sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij} = N.$$

Ďalej označme pravdepodobnosti:

$$\mathsf{P}(X = i,Y = j) = p_{ij}, \qquad \mathsf{P}(X = i) = q_i, \qquad \mathsf{P}(Y = j) = r_j. \tag{1.2}$$

Teda platí:

$$q_i = \sum_{j=1}^{J} p_{ij}, \qquad r_j = \sum_{i=1}^{I} p_{ij}, \qquad \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1.$$

Pravdepodobnosti p_{ij} určujú združené rozdelenie náhodných veličín X a Y. Hodnoty q_i,r_j označujú *marginálne pravdepodobnosti*. Označme nulovú hypotézu H_0 a alternatívu H_1 nasledovne:

$$H_0 : \{ \text{ X a Y sú nezávislé } \}$$
$$H_1 : \{ \text{ X a Y nie sú nezávislé } \}.$$

Náhodné veličiny X a Y sú nezávislé práve vtedy, keď platí $p_{ij} = q_i r_j$ pre všetky dvojice (i,j), $i = 1,\ldots,I, j = 1,\ldots,J$. Nulovú hypotézu môžeme teda ekvivalentne prepísať ako:

$$H_0 : p_{ij} = q_i r_j, \qquad \forall i \in \{1,\ldots,I\}, \forall j \in \{1,\ldots,J\}. \tag{1.3}$$

Ako príklad uvedieme pozorovanie z 360 dospelých ľudí, u ktorých pozorujeme krvnú skupinu a index telesnej hmotnosti (BMI), ktorý bol rozdelený do piatich kategórií. Četnosti zapíšeme do kontingenčnej tabuľky 1.2. Údaje sú iba ilustračné, nezakladajú sa na žiadnej reálnej štúdii. Tiež je nutné poznamenať, že pri testoch, ktorými sa budeme zaoberať, nezáleží na usporiadaní jednotlivých kategórií.

	A	B	AB	0
<18,5	3	5	3	8
18.5–24.9	18	37	35	21
25–29.9	45	32	36	18
30–34.9	10	20	17	22
35<	8	6	9	7

Tabulka 1.2: Pozorované četnosti

Hladáme odpoveď na otázku či závisí BMI v populácii od toho, akú majú ľudia krvnú skupinu. Môžeme použiť rôzne štatistické testy, ako napríklad Pearsonov χ^2-test, G-test alebo Fisherov faktoriálový test. V tejto kapitole sa budeme zaoberať najmä prvými dvoma testami.

1.2 Pearsonov χ^2-test nezávislosti

Pearsonov χ^2-test je jeden z najpoužívanejších testov na zistenie nezávislosti dvoch náhodných veličín, ktoré sú kategoriálne. χ^2-test je možné použiť aj ako test dobrej shody, kedy sa snažíme zistiť či pozorované dáta pochádzajú z dopredu daného rozdelenia. V tejto práci sa budeme zaoberať iba testovaním hypotéz nezávislosti. Dve náhodné veličiny sú navzájom nezávislé, ak hodnoty jednej náhodnej veličiny neovplyvnujú, akých hodnôt nadobúda druhá náhodná veličina.

Statistika Pearsonova χ^2-testu má tvar:

$$\chi^2_N = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}.$$

Rozdelenie testovej statistiky za H_0, viz (Anděl, 2007, str.283):

$$\chi^2_N \xrightarrow[N\to\infty]{D} \chi^2_{(I-1)(J-1)} \tag{1.4}$$

Hodnota $e_{ij} = \frac{o_{i+}o_{+j}}{N}$ udáva očakávanú četnosť za nulovej hypotézy. χ^2-test patrí medzi asymptotické testy. Nulovú hypotézu zamietame na základe pravidla:

$$H_0 \text{ zamietni } \iff \chi^2_N \geq \chi^2_{(I-1)(J-1)}(1-\alpha),$$

kde $\chi^2_{(I-1)(J-1)}(1-\alpha)$ označuje $(1-\alpha)$-tý kvantil χ^2-rozdelenia s $(I-1)(J-1)$ stupňami voľnosti a α udáva hladinu testu.

Súvislosť s χ^2-divergenciou

Jedna z možností, ako môžeme nahliadnuť statistiku χ^2-testu je pomocou χ^2-divergencie. Táto divergencia je špeciálny prípad tzv. f-divergencie. Definíciu f-divergencie spolu s príkladmi môžeme nájsť v (Dragomir a kol., 2001, str.3). Uvažujme dve diskrétne pravdepodobnostné rozdelenia $P = (p_{ij})$ a $Q = (q_{ij})$ s rovnakými nosičmi. χ^2-divergencia P od Q má tvar:

$$D_{\chi^2}(P,Q) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-q_{ij})^2}{q_{ij}}.$$

χ^2-divergencia nie je symetrická ($D_{\chi^2}(P,Q) \neq D_{\chi^2}(Q,P)$), preto ju nemôžeme označiť ako vzdialenosť. Pri testovaní hypotézy nezávislosti nás zaujíma či platí vzťah $p_{ij} = q_i r_j$ pre všetky $i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}$. Táto rovnosť platí práve vtedy keď je vzdialenosť (v zmysle χ^2-divergencie) medzi rozdeleniami $P = (p_{ij})$ a $Q = (q_i r_j)$ nulová. Pravdepodobnosti p_{ij},q_i,r_j nepoznáme, preto ich musíme odhadnúť.

Pre $k = 1,\ldots,N$ a pre dané $i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}$ označme

$$Z_{ij}^{(k)} = 1\{X_k = i, Y_k = j\}. \tag{1.5}$$

Pozorované četnosti o_{ij} môžeme písať ako v (1.1). Potom platí

$$\frac{o_{ij}}{N} = \frac{1}{N}\sum_{k=1}^{N} Z_{ij}^{(k)} := \overline{Z_N}. \tag{1.6}$$

Náhodné veličiny $Z_{ij}^{(k)}$ sú nezávislé a rovnako rozdelené a stredná hodnota $\mathbb{E}Z_{ij}^{(k)} = \mathbb{E}(1\{X_k = i, Y_k = j\}) = \mathsf{P}(X_k = i, Y_k = j) = p_{ij}$ je konečná. Preto zo zákona veľkých čísel dostávame

$$\overline{Z_N} \xrightarrow[N\to\infty]{s.j.} \mathbb{E}Z_{ij}^{(k)} = p_{ij}, \tag{1.7}$$

a teda

$$\frac{o_{ij}}{N} \xrightarrow[N\to\infty]{s.j.} p_{ij}. \tag{1.8}$$

Podobne zkonstruujeme odhad pre hodnoty q_i,r_j. Z vety o spojitej transformácii dostávame

$$\frac{o_{i+}}{N} = \sum_{j=1}^{J} \frac{o_{ij}}{N} \xrightarrow[N\to\infty]{s.j.} \sum_{j=1}^{J} p_{ij} = q_i, \qquad \frac{o_{+j}}{N} = \sum_{i=1}^{I} \frac{o_{ij}}{N} \xrightarrow[N\to\infty]{s.j.} \sum_{i=1}^{I} p_{ij} = r_j,$$

$$\frac{e_{ij}}{N} = \frac{o_{i+}}{N} \cdot \frac{o_{+j}}{N} \xrightarrow[N\to\infty]{s.j.} q_i r_j. \tag{1.9}$$

Následne môžeme zkonstruovať odhad hodnoty $D_{\chi^2}(P,Q)$

$$\hat{D}_{\chi^2}(P,Q) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\frac{o_{ij}}{N} - \frac{e_{ij}}{N})^2}{\frac{e_{ij}}{N}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{N e_{ij}} = \frac{\chi^2_N}{N}.$$

Odhad $\hat{D}_{\chi^2}(P,Q)$ konverguje v pravdepodobnosti k $D_{\chi^2}(P,Q)$, čo môžeme nahliadnuť z nasledujúceho výpočtu.
Označme vektory $\mathbf{x} = (x_{11},x_{12},...,x_{IJ})$, $\mathbf{y} = (y_{11},y_{12},...,y_{IJ})$ a definujme funkciu

$$g(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(x_{ij} - y_{ij})^2}{y_{ij}},$$

ktorá je spojitá na $(0,\infty)^{(2IJ)}$. Ďalej označme náhodné vektory

$$\mathbf{o} = (o_{11},o_{12},\ldots,o_{IJ}) \qquad \mathbf{e} = (e_{11},e_{12},\ldots,e_{IJ}) \tag{1.10}$$

$$\mathbf{p} = (p_{11},\ldots,p_{IJ}) \qquad \mathbf{qr} = (q_1r_1,\ldots,q_Ir_J). \tag{1.11}$$

Z vety o spojitej transformácii dostávame:

$$g\Big(\frac{\mathbf{o}}{N}, \frac{\mathbf{e}}{N}\Big) \xrightarrow[N\to\infty]{s.j.} g(\mathbf{p}, \mathbf{qr})^T,$$

a teda

$$\hat{D}_{\chi^2}(P,Q) \xrightarrow[N\to\infty]{s.j.} \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - q_ir_j)^2}{q_ir_j} = D_{\chi^2}(P,Q).$$

Získavame tak vzťah

$$\frac{\chi^2_N}{N} \xrightarrow[N\to\infty]{s.j.} D_{\chi^2}(P,Q).$$

1.3 Test pomerom vierohodností - G-test

Ďalšou možnosťou testovania nezávislosti dvoch náhodných veličín je *G-test*, častokrát nazývaný aj ako *test pomerom vierohodností*. G-test môžeme použiť namiesto Pearsonova χ^2 testu. G-test je daný testovou statistikou

$$G = 2\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij} \log \frac{o_{ij}}{e_{ij}}. \tag{1.12}$$

1.3.1 Odvodenie testovej statistiky

V tejto kapitole popíšeme odvodenie testovej statistiky (1.12) pomocou *pomeru vierohodností*.

Na kontingenčnú tabuľku môžeme nahliadnuť ako na realizáciu multinomického rozdelenia. Vektor $(o_{11},o_{12},...,o_{IJ})$ má multinomické rozdelenie $Mult_{IJ}(N,\mathbf{p})$, kde $\mathbf{p}$ má tvar $\mathbf{p} = (p_{11},p_{12},...,p_{IJ})^T$ a p_{ij} je definovaná ako v (1.2). Zo znalosti multinomického rozdelenia dostávame:

$$\mathsf{P}(O_{11} = x_{11},...,O_{IJ} = x_{IJ}) = \frac{N!}{x_{11}!x_{12}!\cdots x_{IJ}!} p_{11}^{x_{11}} p_{12}^{x_{12}} \cdots p_{IJ}^{x_{IJ}}.$$

Maximálne vierohodný odhad vektoru $\mathbf{p}$ označíme ako $\hat{\mathbf{p}} = (\hat{p}_{11}, \hat{p}_{12},...,\hat{p}_{IJ})^T$, kde

$$\hat{p}_{ij} = \frac{o_{ij}}{N}.$$

Označme ako $\tilde{\mathbf{p}}$ maximálne vierohodný odhad vektora $\mathbf{p}$ za nulovej hypotézy. Nulovú hypotézu môžeme ekvivalentne prepísať ako (1.3). Označme teda $\tilde{p}_{ij}$ odhad súčinu marginálnych pravdepodobností $q_i r_j$. Potom odhad $\tilde{\mathbf{p}}$ má tvar $\tilde{\mathbf{p}} = (\tilde{p}_{11}, \tilde{p}_{12},...,\tilde{p}_{IJ})^T$, kde

$$\tilde{p}_{ij} = \frac{e_{ij}}{N} = \frac{o_{i+}o_{+j}}{N^2},$$

viz (Omelka, 2022, str.148-149). Definujme vierohodnostný pomer ako

$$V_N = \frac{L(\hat{\mathbf{p}})}{L(\tilde{\mathbf{p}})},$$

kde funkcia $L(p)$ označuje vierohodnostnú funkciu. Za platnosti H_0 dostávame:

$$2\log V_N \xrightarrow[N\to\infty]{D} \chi^2_{(I-1)(J-1).}$$

Testová statistika založená na pomeru vierohodností má nasledujúci tvar (Anděl, 2007, str.177).

$$\begin{aligned}
G &= 2\log V_N = 2(\log L(\hat{\mathbf{p}}) - \log L(\tilde{\mathbf{p}})) \\
&= 2\cdot \log\left(\frac{\hat{p}_{11}^{o_{11}}\hat{p}_{12}^{o_{12}}\cdots\hat{p}_{IJ}^{o_{IJ}}}{\tilde{p}_{11}^{o_{11}}\tilde{p}_{12}^{o_{12}}\cdots\tilde{p}_{IJ}^{o_{IJ}}}\right) \\
&= 2\cdot \log\prod_{i=1}^{I}\prod_{j=1}^{J}\left(\frac{\hat{p}_{ij}}{\tilde{p}_{ij}}\right)^{o_{ij}} = 2\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij}\cdot\log\left(\frac{\frac{o_{ij}}{N}}{\frac{e_{ij}}{N}}\right) \\
&= 2\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij}\cdot\log\frac{o_{ij}}{e_{ij}}.
\end{aligned}$$

Hodnotu $G/2N$ môžeme nahliadnuť ako odhad tzv. *Kullbackovej–Leiblerovej divergencie*. Uvažujme znovu dve diskrétne pravdepodobnostné rozdelenia $P = (p_{ij})$, $Q = (q_i r_j)$ s rovnakými nosičmi. Kullbackova–Leiblerova divergencia má tvar

$$D_{KL}(P,Q) = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} \log\left(\frac{p_{ij}}{q_i r_j}\right).$$

Uvažujme odhady (1.6), (1.9) a vektory (1.10). Definujme funkciu

$$g(\mathbf{x},\mathbf{y}) = 2\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij} \log\left(\frac{x_{ij}}{y_{ij}}\right),$$

ktorá je spojitá na $(0,\infty)^{(2IJ)}$. S využitím vety o spojitej transformácii dostávame

$$2\frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij} \log\left(\frac{\frac{o_{ij}}{N}}{\frac{e_{ij}}{N}}\right) = g\Big(\frac{\mathbf{o}}{N},\frac{\mathbf{e}}{N}\Big) \xrightarrow[N\to\infty]{P} 2D_{KL}(P,Q),$$

a teda

$$\frac{G}{N} \xrightarrow[N\to\infty]{P} 2D_{KL}(P,Q).$$

Za platnosti nulovej hypotézy nezávislosti má testová statistika G asymptoticky χ^2-rozdelenie s $(I-1)(J-1)$ stupňami voľnosti (viz (Anděl, 2007, str.184). Hypotézu H_0 zamietame pre $G \geq \chi^2_{(I-1)(J-1)}(1-\alpha)$, kde α predstavuje hladinu testu a $\chi^2_{(I-1)(J-1)}(1-\alpha)$ je $(1-\alpha)$-tý kvantil χ^2 rozdelenia s $(I-1)(J-1)$ stupňami voľnosti.

1.3.2 Vzťah testovej statistiky χ^2-testu a G-testu

Testovú statistiku G-testu je možné aproximovať pomocou testovej statistiky χ^2-testu, s využitím Taylorovho rozvoja logaritmu. Nasledujúca veta vyjadruje vzťah medzi testovou statistikou χ^2-testu a G-testu.

Veta 1. *Za platnosti hypotézy nezávislosti platí:* $G_N - \chi^2_N \xrightarrow[N\to\infty]{P} 0$.

Dôkaz. Rozdiel pozorovanej četnosti o_{ij} a očakávanej četnosti e_{ij} označíme ako δ_{ij}, teda $\delta_{ij} = o_{ij} - e_{ij}$. Súčet odchyliek v kontingenčnej tabulke je 0, čo môžeme nahliadnuť v nasledujúcom výpočte:

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{J}\delta_{ij} &= \sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij} - e_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{o_{i+}o_{+j}}{N} \\
&= N - \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\Big(\sum_{j=1}^{J} o_{ij}\sum_{i=1}^{I} o_{ij}\Big) = N - \frac{1}{N}\Big(\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij}\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij}\Big) \\
&= N - \frac{1}{N}N^2 = 0.
\end{aligned}$$

Po substitúcii $o_{ij} = e_{ij} + \delta_{ij}$ dostávame

$$G = 2\sum_{i=1}^{I}\sum_{j=1}^{J}(e_{ij}+\delta_{ij})\log\frac{e_{ij}+\delta_{ij}}{e_{ij}} = 2\sum_{i=1}^{I}\sum_{j=1}^{J}(e_{ij}+\delta_{ij})\log\left(1+\frac{\delta_{ij}}{e_{ij}}\right).$$

Následne prepíšeme výraz $\log(1+\frac{\delta_{ij}}{e_{ij}})$ pomocou Taylorovho rozvoja logaritmu $\log(1+x) = x - \frac{1}{2}x^2 + \mathcal{O}(x^3)$:

$$\log(1+\tfrac{\delta_{ij}}{e_{ij}}) = \tfrac{\delta_{ij}}{e_{ij}} - \tfrac{1}{2}\tfrac{\delta_{ij}^2}{e_{ij}^2} + \mathcal{O}((\tfrac{\delta_{ij}}{e_{ij}})^3).$$

Výrazy v zátvorkách roznásobíme a využijeme rovnosť $\sum_{i=1}^{I}\sum_{j=1}^{J}\delta_{ij} = 0$.

$$\begin{aligned}
G &= 2\sum_{i=1}^{I}\sum_{j=1}^{J}(e_{ij}+\delta_{ij})\left(\frac{\delta_{ij}}{e_{ij}} - \frac{1}{2}\frac{\delta_{ij}^2}{e_{ij}^2} + \mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big)\right)\\
&= 2\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\delta_{ij} - \frac{1}{2}\frac{\delta_{ij}^2}{e_{ij}} + \frac{\delta_{ij}^2}{e_{ij}} - \frac{1}{2}\frac{\delta_{ij}^3}{e_{ij}^2} + e_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big) + \delta_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big)\right)\\
&= 2\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\delta_{ij} + \frac{1}{2}\frac{\delta_{ij}^2}{e_{ij}} - \frac{1}{2}\frac{\delta_{ij}^3}{e_{ij}^2} + e_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big) + \delta_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big)\right)\\
&= 2\sum_{i=1}^{I}\sum_{j=1}^{J}\delta_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\delta_{ij}^2}{e_{ij}} - \frac{1}{2}\frac{\delta_{ij}^3}{e_{ij}^2} + e_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big) + \delta_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big)\\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} - \frac{1}{2}\frac{\delta_{ij}^3}{e_{ij}^2} + e_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big) + \delta_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big).
\end{aligned}$$

Zostáva ukázať

$$\left|-\frac{1}{2}\frac{\delta_{ij}^3}{e_{ij}^2} + e_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big) + \delta_{ij}\mathcal{O}\Big(\frac{\delta_{ij}^3}{e_{ij}^3}\Big)\right| \xrightarrow[N\to\infty]{P} 0. \tag{1.13}$$

Označme $\mathbf{Z} = (Z_{11}^{(k)}, Z_{12}^{(k)}, ..., Z_{IJ}^{(k)})^T$, kde $Z_{ij}^{(k)}$ je ako v (1.5) .Platí, že stredná hodnota $\mathbb{E}\mathbf{Z} = (p_{11},...,p_{IJ}) := \boldsymbol{\mu}$ a rozptylová matica $\mathbb{V} = \mathsf{var}\mathbf{Z}$ je konečná, a teda z (1.6) a centrálnej limitnej vety dostávame

$$\sqrt{N}\left[\begin{pmatrix}\frac{o_{11}}{N}\\ \vdots \\ \frac{o_{IJ}}{N}\end{pmatrix} - \begin{pmatrix}p_{11}\\ \vdots \\ p_{IJ}\end{pmatrix}\right] \xrightarrow[N\to\infty]{D} \mathbf{N}_{IJ}(\mathbf{0}_{IJ},\mathbb{V}).$$

Definujme funkciu

$$g(x_{11},\ldots,x_{IJ}) = x_{ij} - \sum_{j=1}^{J}x_{ij}\sum_{i=1}^{I}x_{ij}.$$

Tá má zrejme spojité parciálne derivácie na nejakom okolí bodu $\boldsymbol{\mu}$. Z delta-vety plynie

$$\sqrt{N}\Big(\frac{o_{ij}}{N} - \frac{e_{ij}}{N}\Big) = \sqrt{N}\Big(g\Big(\frac{o_{11}}{N},...,\frac{o_{IJ}}{N}\Big) - g(\boldsymbol{\mu})\Big) \xrightarrow[N\to\infty]{D} \mathbf{N}(0,v^2),$$

kde

$$v^2 = \nabla g(\boldsymbol{\mu}) \mathbb{V}[\nabla g(\boldsymbol{\mu})]^T,$$

pričom za nulovej hypotézy platí vzťah

$$g(\boldsymbol{\mu}) = p_{ij} - \sum_{j=1}^{J} p_{ij} \sum_{i=1}^{I} p_{ij} = p_{ij} - p_{i+}p_{+j} = 0.$$

S využitím (1.9) a vety o spojitej transformácii dostávame

$$\frac{\delta_{ij}^3}{e_{ij}^2} = \frac{\frac{1}{\sqrt{N}}\left(\frac{1}{\sqrt{N}}\delta_{ij}\right)^3}{\left(\frac{e_{ij}}{N}\right)^2} = \frac{\frac{1}{\sqrt{N}}\left(\sqrt{N}\left(\frac{o_{ij}}{N} - \frac{e_{ij}}{n}\right)\right)^3}{\left(\frac{e_{ij}}{N}\right)^2} \xrightarrow[N\to\infty]{P} 0,$$

$$\frac{\delta_{ij}^4}{e_{ij}^3} = \frac{\frac{1}{N}\left(\frac{1}{\sqrt{N}}\delta_{ij}\right)^4}{\left(\frac{e_{ij}}{N}\right)^3} = \frac{\frac{1}{N}\left(\sqrt{N}\left(\frac{o_{ij}}{N} - \frac{e_{ij}}{N}\right)\right)^4}{\left(\frac{e_{ij}}{N}\right)^3} \xrightarrow[N\to\infty]{P} 0,$$

pre $\forall i \in \{1,...,I\}, \forall j \in \{1,...,J\}$, z čoho plynie vzťah (1.13), a teda

$$G_N - \chi_N^2 \xrightarrow[N\to\infty]{P} 0.$$

□

χ^2-test a G-test patria medzi najpoužívanejšie testy, no pre malé četnosti v kontingenčnej tabuľke nemusia dodržiavať hladinu α. Často sa stretneme s pravidlom, že všetky pozorovné četnosti o_{ij} musia byť ≥ 5.
Oba testy môžeme ilustrovať na príklade s krvnými skupinami, daný tabuľkou 1.2. P-hodnota χ^2-testu je 0.0006 a teda nulovú hypotézu môžeme zamietnuť na hladine $\alpha = 0{,}05$. G-test dáva p-hodnotu 0.0015, teda H_0 môžeme rovnako zamietnuť na tejto hladine.

1.4 USP test

V tejto kapitole sa budeme venovať USP testu (U-statistic Permutation test), ktorý bol vytvorený na testovanie hypotéz nezávislosti. V celej kapitole budeme čerpať najmä z článku Berrett a Samworth (2021).
Definujeme novú mieru vzdialenosti dvoch diskrétnych pravdepodobnostných rozdelení $P = (p_{ij}), Q = (q_i r_j)$ s rovnakými nosičmi ako

$$D := D(P,Q) = \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - q_i r_j)^2.$$

Všimnime si, že $D(P,Q) = D(Q,P)$, a teda táto nová miera vzdialenosti je symetrická (na rozdiel od χ^2-divergence alebo Kullbackovej–Leiblerovej divergencie).

Rovnako ako pre χ^2-divergenci platí, že za platnosti nulovej hypotézy nezávislosti je hodnota D nulová. Pravdepodobnosti p_{ij},q_i,r_j nepoznáme, preto musíme mieru D odhadnúť. K tomu využijeme teóriu o U-statistikách, ktorá nám umožní nájsť nestranný odhad hodnoty D. Na základe tohoto odhadu potom zostrojíme testovú statistiku USP testu.

1.4.1 Testová statistika USP testu

Testová statistika USP testu má tvar

$$U = \frac{1}{N(N-3)} \sum_{i=1}^{I} \sum_{j=1}^{J} (o_{ij} - e_{ij})^2 - \frac{4}{N(N-2)(N-3)} \sum_{i=1}^{I} \sum_{j=1}^{J} o_{ij} e_{ij}. \tag{1.14}$$

K tejto štatistike sa autori článku (Berrett a Samworth, 2021) dostali pomocou odhadu hodnoty D, ktorý má tvar

$$\hat{D} = \frac{1}{4! \binom{N}{4}} \sum_{(i_1,\ldots,i_4) \in P_{4,N}} h((X_{i_1} Y_{i_1}),\ldots,(X_{i_m} Y_{i_m})),$$

kde $P_{4,N}$ označuje množinu všetkých usporiadaných štvoríc z $\{1,\ldots,N\}$. Môžeme si všimnúť, že $\hat{D}$ je U-statistika štvrtého rádu s jadrom (2.18), kde

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

Definíciu U-statistiky nájdeme v Appendixu. Podľa článku (Berrett a Samworth, 2021) môžeme $\hat{D}$ ďalej prepísať ako

$$\begin{aligned} \hat{D} = \frac{1}{N(N-3)} \sum_{i=1}^{I} \sum_{j=1}^{J} (o_{ij} - e_{ij})^2 - \frac{4}{N(N-2)(N-3)} \sum_{i=1}^{I} \sum_{j=1}^{J} o_{ij} e_{ij} \\ + \frac{\sum_{i=1}^{I} o_{i+}^2 + \sum_{j=1}^{J} o_{+j}^2}{N(N-1)(N-3)} + \frac{(3N-2)(\sum_{i=1}^{I} o_{i+}^2)(\sum_{j=1}^{J} o_{+j}^2)}{N^3(N-1)(N-2)(N-3)} \\ - \frac{N}{(N-1)(N-3)}. \end{aligned}$$

Všimnime si, že tretí a štvrtý sčítanec v tejto sume závisí iba na marginálnych četnostiach o_{i+},o_{+j}. Tieto marginálne četnosti sa nezmenia, ani keď pracujeme so spermutovanými dátami

$$(X_1, Y_{\pi(1)})^T,\ldots,(X_N, Y_{\pi(N)})^T, \tag{1.15}$$

kde $\Pi = (\pi(1),\ldots,\pi(N))$ je permutácia množiny $\{1,\ldots,N\}$. USP test je permutačný test (definíciu permutačného testu môžeme nájsť v publikácii (Salmaso a Pesarin, 2010, kapitola 1.2), preto pri konstrukcii jeho testovej statistiky môžeme tretí a štvrtý sčítanec uplne zanedbať. Rovnako môžeme zanedbať aj posledný konstantný člen. Na základe tejto úvahy získame testovú statistiku USP testu (1.14). $\hat{D}$ je nestranným odhadom D, čo môžeme nahliadnuť z nasledujúceho výpočtu:

$$\begin{aligned} \mathbb{E}h((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)) = \\ \mathbb{E}\Big(\sum_{i=1}^{I} \sum_{j=1}^{J} (1_{\{X_1=i,Y_1=j\}} 1_{\{X_2=i,Y_2=j\}} - 2 \cdot 1_{\{X_1=i,Y_2=j\}} 1_{\{X_2=i\}} 1_{\{Y_3=j\}} \end{aligned}$$

$$+1_{\{X_1=i\}}1_{\{Y_2=j\}}1_{\{X_3=i\}}1_{\{X_4=j\}})\Big)$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J}(\mathsf{P}(X_1 = i,Y_1 = j)P(X_2 = i,Y_2 = j) -$$

$$-2\mathsf{P}(X_1 = i,Y_1 = j)\mathsf{P}(X_2 = i)\mathsf{P}(Y_3 = j) +$$

$$+\mathsf{P}(X_1 = i)\mathsf{P}(Y_2 = j)\mathsf{P}(X_3 = i)\mathsf{P}(X_4 = j))$$

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

Naviac, ak vytvoríme ľubovoľnú permutáciu $\Pi = (\pi(1),\ldots,\pi(N))$ z množiny $\{1,\ldots,N\}$, ktorá je nezávislá s napozorovanými dátami, tak pre strednú hodnotu jadra h, aplikovaného na spermutovaný náhodný výber (1.15), platí:

$$\mathbb{E}h((X_1,Y_{\pi(1)}),(X_2,Y_{\pi(2)}),(X_3,Y_{\pi(3)}),(X_4,Y_{\pi(4)})) = 0.$$

Toto tvrdenie nájdeme v článku (Berrett a kol., 2021b, str.7). Tento výsledok sa dá interpretovať nasledovne: Ak sa vrátime k príkladu v kapitole 1, tak za platnosti nulovej hypotézy nemá krvná skupina ľudí žiadny vplyv na ich telesnú hmotnosť. Teda za platnosti H_0 je rovnako pravdepodobné, že napozorujeme dáta $(X_1,Y_1),\ldots,(X_N,Y_N)$ alebo spermutované dáta (1.15). Hodnota $\hat{D}$ odhaduje mieru vzdialenosti D, ktorá je nulová práve vtedy, keď sú X a Y nezávislé.

1.4.2 Konstrukce USP testu

USP test nezávislosti môžeme previesť v nasledujúcich krokoch:

1. Testovú statistiku U aplikujeme na pôvodné dáta, ktoré označíme

   $$S_N = \{(X_1,Y_1)^T,\ldots,(X_N,Y_N)^T\}.$$

   Príslušnú hodnotu testovej statistiky označíme ako $U = U(S_N)$.

2. Zvolíme číslo B, ktoré bude udávať počet permutácií z náhodného výberu S_N. Najčastejšie sa volí $B = 999$.

3. Vytvoríme nezávislú permutáciu $\Pi_b = (\pi_b(1),\ldots,\pi_b(N))$ množiny $\{1,\ldots,N\}$ pre všetky $b = 1,\ldots,B$. Uvažujme spermutovaný náhodný výber $S_N^{(b)}$, ktorý vznikol pomocou permutácie Π_b:

   $$S_N^{(b)} = \left\{\begin{pmatrix}X_1\\Y_{\pi_b(1)}\end{pmatrix},\ldots,\begin{pmatrix}X_N\\Y_{\pi_b(N)}\end{pmatrix}\right\}.$$

4. Statistiku U aplikujeme na množinu $S_N^{(b)}$ pre všetky $b = 1,\ldots,B$. Príslušné hodnoty testovej statistiky U na jednotlivých množinách $S_N^{(b)}$ označíme ako

   $$U^{(b)} = U(S_N^{(b)}).$$

5. Týmto spôsobom skonstruujeme $B+1$ hodnôt U,$U^{(1)},\ldots,U^{(B)}$.

6. **kritický obor**
   Postupnosť U,$U^{(1)},\ldots,U^{(B)}$ zoradíme zostupne a poradie hodnoty U v tejto postupnosti označíme ako l. Nulovú hypotézu zamietame na základe pravidla:

$$l \leq \alpha(B+1) : H_0 \text{ zamietame},$$
$$l > \alpha(B+1) : H_0 \text{ nezamietame},$$

   kde α označuje hladinu testu.

Poznámka. P-hodnota testu predstavuje najmenšiu hladinu α, na ktorej by sme ešte mohli zamietnuť nulovú hypotézu. Hypotézu H_0 zamietame, ak l splnuje $l \leq \alpha(B+1)$. Z tejto nerovnosti odvodíme vzťah pre α:

$$\alpha \geq \frac{l}{B+1}. \tag{1.16}$$

Najmenšia hladina α, ktorá splnuje vzťah (1.16) je

$$\alpha = \frac{l}{B+1} = \frac{1+\sum_{b=1}^{B} 1\{U^{(b)} \geq U\}}{B+1},$$

ktorá predstavuje p-hodnotu testu.

USP test aplikujeme na príklad s krvnými skupinami z kapitoly 1. Na realizáciu testu použijeme prostredie R Core Team (2021) a funkciu USP.test , ktorá je implementovaná v balíku Berrett a kol. (2021a). Výsledná p-hodnota je 0.008 (pri voľbe $B = 999$), teda nulovú hypotézu môžeme zamietnuť na hladine $\alpha = 0.05$.

V nasledujúcej vete ukážeme, že $\hat{D}$ je najlepším nestranným odhadom miery D. V dôkaze tejto vety využijeme druhú Lehmannovu-Scheffého vetu (Anděl, 2007, str.136).

Veta 2. *$\hat{D}$ je najlepší nestranným odhadom D, a to jediný.*

Dôkaz. Na kontingenčnú tabuľku sa môžeme pozerať ako na realizáciu multinomického rozdelenia. Nech $d = IJ$. Označme vektory

$$\mathbf{x} = (x_1,x_2,...,x_d) := (x_{11},x_{12},...,x_{IJ})^T$$

$$\mathbf{p} = (p_1,p_2,...,p_d) := (p_{11},p_{12},...,p_{IJ})^T,$$

pre ktoré platí:

$$\sum_{i=1}^{d} p_i = 1, \qquad \sum_{i=1}^{d} x_i = N, \qquad p_i \geq 0.$$

Označme $P_D := \mathsf{P}(O_1 = x_1, O_2 = x_2, ..., O_d = x_d)$. Zo znalosti multinomického rozdelenia dostávame:

$$P_D = \frac{N!}{\prod_{i=1}^{I}\prod_{j=1}^{J} x_{ij}!} p_1^{x_1} p_2^{x_2} \cdots p_d^{x_d} = \frac{N!}{\prod_{i=1}^{I}\prod_{j=1}^{J} x_{ij}!} \exp\left\{ \sum_{i=1}^{d} x_i \log(p_i) \right\}$$

$$= \frac{N!}{\prod_{i=1}^{I}\prod_{j=1}^{J} x_{ij}!} \exp\left\{ \sum_{i=1}^{d-1} x_i \log(p_i) + \left(N - \sum_{i=1}^{d-1} x_i\right) \log\left(1 - \sum_{i=1}^{d-1} p_i\right) \right\}$$

$$= \frac{N!}{\prod_{i=1}^{I}\prod_{j=1}^{J} x_{ij}!} \exp\left\{ \sum_{i=1}^{d-1} \log\left(\frac{p_i}{1 - \sum_{i=1}^{d-1} p_i}\right) x_i \right\} \exp\left\{ N \cdot \log\left(1 - \sum_{i=1}^{d-1} p_i\right) \right\}.$$

Množina

$$\left\{ \log\left(\frac{p_1}{1 - \sum_{i=1}^{d-1} p_i}\right), \log\left(\frac{p_2}{1 - \sum_{i=1}^{d-1} p_i}\right), \ldots, \log\left(\frac{p_{d-1}}{1 - \sum_{i=1}^{d-1} p_i}\right) \right\},$$

kde $\sum_{i=1}^{d} p_i = 1, p_i \geq 0$, obsahuje nedegenerovaný $(d-1)$-dimenzionálny interval, a teda z vety o exponenciálnom systéme (Anděl, 2007, str.133) dostávame, že statistika $S_1 = (o_1, ..., o_{d-1})$ je úplná postačujúca. Statistika $S = (o_1, ..., o_d)$ je funkciou statistiky S_1, čo plynie zo vzťahu $o_d = N - \sum_{i=1}^{d-1} o_i$. Ukázali sme teda, že aj statistika S je úplná postačujúca a $\hat{D}$ je funkciou úplnej postačujúcej statistiky. V sekcii 1.4.1 sme ukázali, že $\hat{D}$ je nestranný odhad miery D. Z druhej Lehmannovej-Scheffého vety (Anděl, 2007, str.136) dostávame, že $\hat{D}$ je najlepší nestranný odhad D.

□

Poznámka. Uvedieme alternatívny dôkaz vety 2, ktorého skrátenú verziu môžeme nájsť aj v článku Berrett a Samworth (2021).

Dôkaz. V predchádzajúcom dôkaze sme odvodili vzťah

$$\mathsf{P}(O_1 = x_1, O_2 = x_2, ..., O_d = x_d) = \frac{N!}{\prod_{i=1}^{I}\prod_{j=1}^{J} x_{ij}!} \exp\left\{ \sum_{i=1}^{d} x_i \log(p_i) \right\}.$$

Z Neumannova faktorizačného kritéria (Anděl, 2007, str.125) dostávame, že statistika $S = (o_1, ..., o_d)$ je postačujúca. V ďalšom kroku ukážeme, že statistika S je úplná a to priamo z definície.

Označme N_d množinu všetkých d-dimenzionálnych vektorov, ktoré majú súčet prvkov rovný N.

$$N_d = \{(x_1, x_2, ..., x_d) : x_i \in \mathbb{N}_0 \text{ pre } \forall i \in \{1, ..., d\}, \sum_{i=1}^{d} x_i = N.\}$$

Bez ujmy na obecnosti môžeme predpokladať, že $p_d > 0$. V opačnom prípade stačí zvoliť index i taký, že $p_i > 0$. Nech existuje funkcia $g : \mathbb{R}^d \to \mathbb{R}$ taká, že pre všetky $p_1, \ldots, p_d$ platí

$$\begin{aligned}
0 &= \sum_{(x_1,\ldots,x_d)\in N_d} g(x_1,\ldots,x_d)\mathsf{P}(O_1 = x_1, O_2 = x_2, \ldots, O_d = x_d) \\
&= \sum_{(x_1,\ldots,x_d)\in N_d} g(x_1,\ldots,x_d) N! \prod_{i=1}^{d} \frac{{p_i}^{x_i}}{x_i!} \\
&= N! \sum_{(x_1,\ldots,x_d)\in N_d} g(x_1,\ldots,x_d) \frac{{p_d}^{x_d}}{x_d!} \prod_{i=1}^{d-1} \frac{{p_i}^{x_i}}{x_i!} \\
&= N! \sum_{(x_1,\ldots,x_d)\in N_d} g(x_1,\ldots,x_d) \frac{(1-\sum_{i=1}^{d-1} p_i)^{x_d}}{x_d!} \prod_{i=1}^{d-1} \frac{{p_i}^{x_i}}{x_i!} \\
&= N! \sum_{(x_1,\ldots,x_d)\in N_d} g(x_1,\ldots,x_d) \frac{1}{x_d!} \frac{(1-\sum_{i=1}^{d-1} p_i)^{N}}{(1-\sum_{i=1}^{d-1} p_i)^{\sum_{i=1}^{d-1} x_i}} \prod_{i=1}^{d-1} \frac{{p_i}^{x_i}}{x_i!} \\
&= \frac{N!}{x_d!}(1-\sum_{i=1}^{d-1} p_i)^N \sum_{(x_1,\ldots,x_d)\in N_d} g(x_1,\ldots,x_d) \prod_{i=1}^{d-1} \left(\frac{p_i}{1-\sum_{i=1}^{d-1} p_i}\right)^{x_i} \frac{1}{x_i!}.
\end{aligned}$$

V štvrtej a piatej rovnosti sme využili nasledujúce vzťahy

$$p_d = 1 - \sum_{i=1}^{d-1} p_i, \qquad x_d = N - \sum_{i=1}^{d-1} x_i, \qquad p_i \geq 0,$$

v poslednej rovnosti sme potom využili vzťah

$$\prod_{i=1}^{d-1} \left(1 - \sum_{i=1}^{d-1} p_i\right)^{x_i} = \left(1 - \sum_{i=1}^{d-1} p_i\right)^{\sum_{i=1}^{d-1} x_i}.$$

Označme

$$z_i = \frac{p_i}{1 - \sum_{i=1}^{d-1} p_i} \quad i = 1, \ldots, d-1.$$

Všimneme si, že z_i nadobúda iba nezáporné hodnoty, pretože $\sum_{i=1}^{d-1} p_i < 1$. Úpravou dostávame

$$\sum_{(x_1,\ldots,x_d)\in N_d} g(x_1,\ldots,x_d) \prod_{i=1}^{d-1} \frac{z_i}{x_i!} = 0.$$

Tento výraz môžeme interpretovať ako polynóm v premenných $z_1,\ldots,z_{d-1}$ s koeficientami $g(x_1,\ldots,x_d)$, ktorý je rovný nule na $\mathbb{R}_+^{d-1}$. Ak je polynóm rovný nule na $\mathbb{R}_+^{d-1}$, tak musia byť všetky jeho koeficienty nulové na $\mathbb{R}_+^{d-1}$, preto platí

$$g(x_1,\ldots,x_d) = 0 \quad \forall (x_1,\ldots,x_d) \in N_d,$$

čo dokazuje úplnosť statistiky S. Podobne ako v prvom dôkaze plynie z druhej Lehmannovej-Scheffého vety, že $\hat{D}$ je najlepší nestranný odhad D.

□

2. Štvorpoľné kontingenčné tabuľky

Uvažujme špeciálny prípad kontingenčnej tabuľky, pre ktorú $I = 2$ a $J = 2$, teda obe náhodné veličiny X a Y nadobúdajú práve dvoch hodnôt. Túto tabuľku budeme nazývať *štvorpoľná kontingenčná tabuľka* (viz tabuľka 2.1).

	$Y = 1$	$Y = 2$	$\sum$
$X = 1$	o_{11}	o_{12}	o_{1+}
$X = 2$	o_{21}	o_{22}	o_{2+}
$\sum$	o_{+1}	o_{+2}	N

Tabulka 2.1: Štvorpoľná kontingenčná tabuľka

Testová statistika χ^2-testu pre štvorpoľnú kontingenčnú tabuľku sa dá zjednodušiť nasledujúcim spôsobom (viz (Anděl, 2007)):

$$\chi^2_{2\times 2} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} = N\frac{(o_{11}o_{22} - o_{12}o_{21})^2}{o_{1+}o_{2+} + o_{+1}o_{+2}}.$$

Za platnosti nulovej hypotézy má testová statistika $\chi^2_{2\times 2}$ asymptoticky χ^2_1 rozdelenie. Testová statistika G-testu má tvar:

$$G = 2\sum_{i=1}^{2}\sum_{j=1}^{2} o_{ij}\log\frac{o_{ij}}{e_{ij}}.$$

Štvorpoľné kontingenčné tabuľky sa často používajú pri medicínskych výzkumoch. Môžeme testovať napríklad účinnosť nového lieku, kedy je jednej skupine ľudí podávaný nový liek a druhá skupina dostáva placebo (alebo je na ňu použitá štandartná liečba). U oboch skupín ľudí pozorujeme či sa ich zdravotný stav zlepšil alebo zhoršil. Pri klinických štúdiách často porovnávame dva rôzne liečebné zákroky alebo dva rôzne lieky. Štúdie sa vykonávajú na dvoch skupinách ľudí, pričom na každú skupinu je aplikovaný jeden z dvoch liekov (alebo zákrokov). Výsledky týchto pozorovaní nadobúdajú dvoch hodnôt, ako napríklad „vyliečený“ a „nevyliečený“ alebo „prítomnosť vedlajších účinkov“ a „bez prítomnosti vedlajších účinkov“.
Ako príklad uvedieme štúdiu, ktorá skúmala súvislosť pravidelného príjmu aspirinu s výskytom infarktu. Účastníci boli rozdelení do dvoch skupín, pričom jednej skupine bol podávaný aspirin a druhej placebo. Výsledky pozorovania môžeme zapísať do tabuľky 2.2. Údaje v tabuľke sú iba ilustračné.

	Infarkt	Bez infarktu	Spolu
Placebo	3	5	8
Aspirin	7	9	16
Spolu	10	14	24

Tabulka 2.2: Výsledky štúdie závislosti užívania aspirinu a infarktu

2.1 Test homogenity dvoch binomických rozdelení

Doposiaľ sme na kontingenčnú tabuľku nahliadali ako na realizáciu multinomického rozdelenia. Častokrát býva vhodné interpretovať štvorpoľnú kontingenčnú tabuľku po stĺpcoch (resp. po riadkoch) ako realizáciu dvoch binomických rozdelení.
Uvažujme nezávislé náhodné výbery X_1,...,X_n o rozsahu n z alternatívneho rozdelenia $Alt(p_1)$ a Y_1,...,Y_m o rozsahu m z alternatívneho rozdelenia $Alt(p_2)$, pričom $p_1 \in (0,1)$, $p_2 \in (0,1)$. Náhodná veličina $X'_n = \sum_{i=1}^{n} X_i$ má binomické rozdelenie $Bi(n,p_1)$ a náhodná veličina $Y'_m = \sum_{i=1}^{m} Y_i$ má binomické rozdelenie $Bi(m,p_2)$.
Na základe tohoto značenia môžeme prepísať tabuľku 2.1 nasledovne:

	$Y=1$	$Y=2$	$\sum$
$X=1$	X'_n	Y'_m	$X'_n + Y'_m$
$X=2$	$n - X'_n$	$m - Y'_m$	$n + m - X'_n - Y'_m$
$\sum$	n	m	$n+m$

Tabulka 2.3: Realizácia dvoch binomických rozdelení

Nezávislosť X a Y je v tomto prípade ekvivalentná s rovnosťou $p_1 = p_2$. Vráťme s sa znovu k tabuľke 2.2, udávajúcu výsledky určitej lekárskej štúdie. Túto tabuľku môžeme interpretovať po riadkoch ako realizáciu dvoch nezávislých binomických rozdelení $Bi(8,p_1), Bi(16,p_2)$, kde p_1 udáva pravdepodobnosť že pacient, ktorému bolo podávané placebo, dostane infarkt. Analogicky p_2 označuje pravdepodobnosť infarktu pri podávaní Aspirinu. Rovnosť $p_1 = p_2$ implikuje, že riziko infarktu nezáviselo na konzumácii Aspirinu, a teda prevencia nebola účinná.

Veta 3. *Označme $\hat{p}_1 = \frac{X'_n}{n}$ ako odhad pravdepodobnosti úspechu p_1, $\hat{p}_2 = \frac{Y'_m}{m}$ ako odhad pravdepodobnosti úspechu p_2 a $\tilde{p} = \frac{X'_n + Y'_m}{n+m}$ ako odhad spoločnej pravdepodobnosti úspechu za nulovej hypotézy. Nech $\tilde{p} \neq 0$, $\tilde{p} \neq 1$. Testovú statistiku $\chi^2_{2\times 2}$ môžeme prepísať ako*

$$\chi^2_{2x2} = \left(\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\tilde{p}(1-\tilde{p})(\frac{1}{n} + \frac{1}{m})}} \right)^2 . \tag{2.1}$$

Dôkaz. Pomocou značenia zavedeného v tabuľke 2.3 prepíšeme:

$$e_{11} = \frac{(X'_n + Y'_n)n}{n+m} = n\tilde{p} \qquad e_{21} = \frac{n(n+m-(X'_n+Y'_n))}{n+m} = n(1-\tilde{p})$$

$$e_{12} = \frac{(X'_n + Y'_n)m}{n+m} = m\tilde{p} \qquad e_{22} = \frac{m(n+m-(X'_n+Y'_n))}{n+m} = m(1-\tilde{p})$$

$$o_{11} = n\hat{p}_1 \qquad o_{12} = m\hat{p}_2 \qquad o_{21} = n(1-\hat{p}_1) \qquad o_{22} = m(1-\hat{p}_2).$$

Počítajme:

$$\frac{(o_{11}-e_{11})^2}{e_{11}} = \frac{n^2(\hat{p}_1-\tilde{p})^2}{n\tilde{p}} := A_{11} \quad \frac{(o_{12}-e_{12})^2}{e_{12}} = \frac{m^2(\hat{p}_2-\tilde{p})^2}{m\tilde{p}} := A_{12}$$

$$\frac{(o_{21}-e_{21})^2}{e_{21}} = \frac{(n(1-\hat{p}_1-1+\tilde{p}))^2}{n(1-\tilde{p})} = \frac{n^2(\hat{p}_1-\tilde{p})^2}{n(1-\tilde{p})} := A_{21}$$

$$\frac{(o_{22}-e_{22})^2}{e_{22}} = \frac{(m(1-\hat{p}_2-1+\tilde{p}))^2}{m(1-\tilde{p})} = \frac{m^2(\hat{p}_2-\tilde{p})^2}{m(1-\tilde{p})} := A_{22}$$

Testovú statistiku χ^2-testu pre štvorpoľné kontingenčné tabuľky môžeme prepísať ako

$$\begin{aligned}\chi^2_{2\times 2} &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = A_{11}+A_{12}+A_{21}+A_{22} = \\ &= \frac{mn^2(\hat{p}_1-\tilde{p})^2(\tilde{p}+1-\tilde{p}) + nm^2(\hat{p}_2-\tilde{p})^2(\tilde{p}+1-\tilde{p})}{nm\tilde{p}(1-\tilde{p})} = \\ &= \frac{n(\hat{p}_1-\tilde{p})^2+m(\hat{p}_2-\tilde{p})^2}{\tilde{p}(1-\tilde{p})} \end{aligned} \tag{2.2}$$

S využitím vzťahu $\tilde{p} = \frac{n\hat{p}_1+m\hat{p}_2}{n+m}$ dostávame:

$$\begin{aligned} n(\hat{p}_1-\tilde{p})^2+m(\hat{p}_2-\tilde{p})^2 &= n\Big(\hat{p}_1-\frac{n\hat{p}_1+m\hat{p}_2}{n+m}\Big)^2+m\Big(\hat{p}_2-\frac{n\hat{p}_1+m\hat{p}_2}{n+m}\Big)^2 = \\ &= n\Big(\frac{m(\hat{p}_1-\hat{p}_2)}{n+m}\Big)^2+m\Big(\frac{n(\hat{p}_2-\hat{p}_1)}{n+m}\Big)^2 = \\ &= \frac{nm(\hat{p}_1-\hat{p}_2)^2(n+m)}{(n+m)^2} = \frac{nm(\hat{p}_1-\hat{p}_2)^2}{(n+m)}, \end{aligned}$$

čo spoločne s (2.2) dáva

$$\chi^2_{2\times 2} = \frac{(\hat{p}_1-\hat{p}_2)^2}{\tilde{p}(1-\tilde{p})}\frac{nm}{n+m} = \frac{(\hat{p}_1-\hat{p}_2)^2}{\tilde{p}(1-\tilde{p})(\frac{1}{n}+\frac{1}{m})} = \left(\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\tilde{p}(1-\tilde{p})(\frac{1}{n}+\frac{1}{m})}}\right)^2.$$

□

Veta 4. *Za platnosti nulovej hypotézy má testová statistika z vety 3 asymptoticky* χ^2_1 *rozdelenie, teda*

$$\left(\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\tilde{p}(1-\tilde{p})(\frac{1}{n}+\frac{1}{m})}}\right)^2 \xrightarrow{D} \chi^2_1, \quad m,n\to\infty, \frac{n}{m}\to\lambda\in(0,\infty).$$

Dôkaz. Dôkaz je modifikáciou vety 6.2 z (Omelka, 2022). Predpokladajme, že platí nulová hypotéza. Prepíšeme

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\tilde{p}(1-\tilde{p})(\frac{1}{n}+\frac{1}{m})}} = \frac{\sqrt{m}(\hat{p}_1-\hat{p}_2)}{\sqrt{\tilde{p}(1-\tilde{p})\frac{m}{n}+\tilde{p}(1-\tilde{p})}} := T.$$

Za platnosti nulovej hypotézy je $p_1 = p_2 := p$. Stredné hodnoty $\mathbb{E}X_1 = p_1$, $\mathbb{E}Y_1 = p_2$ sú konečné a teda zo zákona veľkých čísel dostávame:

$$\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{s.j.} p \qquad \hat{p}_2 = \frac{1}{m}\sum_{i=1}^{m} Y_i \xrightarrow{s.j.} p$$

Vďaka vete o spojitej transformácii platí vzťah:

$$\sqrt{\tilde{p}(1-\tilde{p})\frac{m}{n}+\tilde{p}(1-\tilde{p})} \xrightarrow{s.j.} \sqrt{\frac{p(1-p)}{\lambda}+p(1-p)} \tag{2.3}$$

Z konečnosti rozptylov $\mathsf{var}X_1 = p_1(1-p_1)$, $\mathsf{var}Y_1 = p_2(1-p_2)$ a s využitém centrálnej limitnej vety

$$\sqrt{n}(\hat{p}_1-p) \xrightarrow{d} N(0,p(1-p)) \quad \sqrt{m}(\hat{p}_2-p) \xrightarrow{d} N(0,p(1-p)),$$

a teda

$$\sqrt{m}(\hat{p}_1-p) = \sqrt{\frac{m}{n}}\sqrt{n}(\hat{p}_1-p) \xrightarrow{d} N(0,p(1-p)/\lambda).$$

Odhady $\hat{p}_1$ a $\hat{p}_2$ sú výberové priemery z náhodného výberu $X_1,\ldots,X_n$ resp. $Y_1,\ldots,Y_m$, a teda sú nezávislé. Preto dostávame:

$$\sqrt{m}\begin{pmatrix}\hat{p}_1-p\\ \hat{p}_2-p\end{pmatrix} \xrightarrow{D} N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\frac{p(1-p)}{\lambda} & 0\\ 0 & p(1-p)\end{pmatrix}\right).$$

Z delta-metódy (Omelka, 2022, str.14) dostávame, že pre všetky $\mathbf{c}$ z $\mathbb{R}^2$ platí

$$\mathbf{c}^T\sqrt{m}\begin{pmatrix}\hat{p}_1-p\\ \hat{p}_2-p\end{pmatrix} \xrightarrow{D} N(0,\mathbf{c}^T\Sigma\mathbf{c}), \quad \text{kde} \quad \Sigma = \begin{pmatrix}\frac{p(1-p)}{\lambda} & 0\\ 0 & p(1-p)\end{pmatrix}.$$

Voľnou $\mathbf{c} = (1,-1)^T$ dostávame

$$\sqrt{m}(\hat{p}_1-\hat{p}_2) \xrightarrow{D} N\left(0,\frac{p(1-p)}{\lambda}+p(1-p)\right). \tag{2.4}$$

Nakoniec využijeme vzťahy (2.3), (2.4) a Cramérovu-Sluckého vetu, pomocou ktorej odvodíme vzťah

$$T = \frac{\sqrt{\frac{p(1-p)}{\lambda}+p(1-p)}}{\sqrt{\tilde{p}(1-\tilde{p})\frac{m}{n}+\tilde{p}(1-\tilde{p})}} \frac{\sqrt{m}(\hat{p}_1-\hat{p}_2)}{\sqrt{\frac{p(1-p)}{\lambda}+p(1-p)}} \xrightarrow{D} N(0,1).$$

Z definície χ^2-rozdelenia dostávame, že za platnosti nulovej hypotézy testová statistika z vety 4 konverguje v distribuci ku χ^2_1- rozdeleniu.

□

Poznámka. Všimnime si, že z vety 3 a 4 plynie aj vzťah 2.4 pre $I = 2, J = 2$.

2.1.1 USP test pre štvorpoľnú tabuľku

Podobným spôsobom ako vo vete 3 môžeme prepísať aj statistiku USP testu U (viz (1.14)) pre štvorpoľnú kontingenčnú tabuľku. Označme ju

$$U_{2\times 2} = U_1 - U_2, \tag{2.5}$$

kde

$$U_1 := \frac{1}{(n+m)(n+m-3)} \sum_{i=1}^{2}\sum_{j=1}^{2}(o_{ij}-e_{ij})^2,$$

$$U_2 := \frac{4}{(n+m)(n+m-2)(n+m-3)} \sum_{i=1}^{2}\sum_{j=1}^{2} o_{ij}e_{ij}.$$

Pomocou značenia zavedeného v kapitole 2.1 prepíšeme

$$(o_{11}-e_{11})^2 = n^2(\hat{p}_1-\tilde{p})^2 \qquad (o_{12}-e_{12})^2 = m^2(\hat{p}_2-\tilde{p})^2$$

$$(o_{21}-e_{21})^2 = n^2(\hat{p}_1-\tilde{p})^2 \qquad (o_{22}-e_{22})^2 = m^2(\hat{p}_2-\tilde{p})^2.$$

Počítajme

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J}(o_{ij}-e_{ij})^2 &= 2n^2(\hat{p}_1-\tilde{p})^2 + 2m^2(\hat{p}_2-\tilde{p})^2 = \\ &= 2n^2\Big(\hat{p}_1 - \frac{n\hat{p}_1+m\hat{p}_2}{n+m}\Big)^2 + 2m^2\Big(\hat{p}_2 - \frac{n\hat{p}_1+m\hat{p}_2}{n+m}\Big)^2 \\ &= \frac{2n^2m^2(\hat{p}_1-\hat{p}_2)^2}{(n+m)^2} + \frac{2m^2n^2(\hat{p}_1-\hat{p}_2)^2}{(n+m)^2} = \frac{4m^2n^2(\hat{p}_1-\hat{p}_2)^2}{(n+m)^2}.\end{aligned}$$

Z toho odvodíme vzťah pre U_1 ako

$$\begin{aligned}U_1 &= \frac{1}{(n+m)(n+m-3)}\frac{4m^2n^2(\hat{p}_1-\hat{p}_2)^2}{(n+m)^2} = (\hat{p}_1-\hat{p}_2)^2\frac{4m^2n^2}{(n+m)^3(n+m-3)} = \\ &= \left(\frac{(\hat{p}_1-\hat{p}_2)2mn}{(n+m)^{3/2}(n+m-3)^{1/2}}\right)^2.\end{aligned}$$

Veta 5. *Označme* $\sigma^2_{p\lambda} = \frac{2p(1-p)}{(1+\lambda)}$, $\omega_{p\lambda} = \frac{4(\lambda^2+1)(2p^2-2p+1)}{(\lambda+1)^3}$. *Za platnosti nulovej hypotézy platí*

$$mU_{2\times 2} \xrightarrow{D} \sigma^2_{p\lambda}\chi^2_1 - \omega_{p\lambda} \quad m,n\to\infty, \frac{n}{m}\to\lambda\in(0,\infty).$$

Dôkaz. Ako prvé ukážeme, že platí

$$mU_1 \xrightarrow{D} {\sigma_{p\lambda}}^2\chi^2_1.$$

Označme

$$Z_{n,m} := \sqrt{m}\frac{(\hat{p}_1-\hat{p}_2)2mn}{(n+m)^{3/2}(n+m-3)^{1/2}}.$$

Prepíšeme

$$Z_{n,m} = \frac{\sqrt{m}(\hat{p}_1 - \hat{p}_2)2mn}{m^{3/2}(\frac{n}{m}+1)^{3/2}m^{1/2}(\frac{n}{m}+1-\frac{3}{m})} = \frac{\sqrt{m}(\hat{p}_1 - \hat{p}_2)2\frac{n}{m}}{(\frac{n}{m}+1)^{3/2}(\frac{n}{m}+1-\frac{3}{m})}.$$

S využitím Cramérovej–Sluckého vety a vzťahu (2.4) dostávame

$$Z_{n,m} \xrightarrow{D} N\Big(0, \frac{2\lambda}{(1+\lambda)^2}\Big(\frac{p(1-p)}{\lambda} + p(1-p)\Big)\Big) = N(0,{\sigma_{p\lambda}}^2).$$

Potom platí

$$\frac{Z_{n,m}}{\sigma_{p\lambda}} \xrightarrow{D} N(0,1).$$

Z definície χ^2 rozdelenia ďalej platí

$$\Big(\frac{Z_{n,m}}{\sigma_{p\lambda}}\Big)^2 \xrightarrow{D} \chi^2_1,$$

z čoho plynie dokazovaný vzťah

$$(Z_{n,m})^2 = m\left(\frac{(\hat{p}_1 - \hat{p}_2)2mn}{(n+m)^{3/2}(n+m-3)^{1/2}}\right)^2 = mU_1 \xrightarrow{D} {\sigma_{p\lambda}}^2\chi^2_1.$$

Ďalej ukážeme platnosť vzťahu

$$mU_2 \xrightarrow{D} \omega_{p\lambda}. \tag{2.6}$$

S využitím (1.7), (1.9), značenia zavedeného v kapitole 2.1 a za predpokladu platnosti nulovej hypotézy odvodíme vzťahy

$$\frac{o_{11}}{n} \xrightarrow{s.j.} p, \qquad \frac{o_{12}}{m} \xrightarrow{s.j.} p, \qquad \frac{o_{21}}{n} \xrightarrow{s.j.} (1-p), \qquad \frac{o_{22}}{m} \xrightarrow{s.j.} (1-p), \tag{2.7}$$

$$\frac{e_{11}}{n} \xrightarrow{s.j.} p, \qquad \frac{e_{12}}{m} \xrightarrow{s.j.} p, \qquad \frac{e_{21}}{n} \xrightarrow{s.j.} (1-p), \qquad \frac{e_{22}}{m} \xrightarrow{s.j.} (1-p). \tag{2.8}$$

Označme

$$B_{11} := \frac{o_{11}e_{11}}{n^2}, \qquad B_{12} := \frac{o_{12}e_{12}}{m^2}, \qquad B_{21} := \frac{o_{21}e_{21}}{n^2}, \qquad B_{22} := \frac{o_{22}e_{22}}{m^2}.$$

S využitím vzťahov (2.7), (2.8) a vety o spojitej transformácii dostávame

$$B_{11} \xrightarrow{s.j.} p^2, \qquad B_{12} \xrightarrow{s.j.} p^2, \qquad B_{21} \xrightarrow{s.j.} (1-p)^2, \qquad B_{21} \xrightarrow{s.j.} (1-p)^2. \tag{2.9}$$

Prepíšeme

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij}e_{ij} &= o_{11}e_{11} + o_{12}e_{12} + o_{21}e_{21} + o_{22}e_{22} \\ &= n^2B_{11} + m^2B_{12} + n^2B_{21} + m^2B_{22} \\ &= n^2(B_{11} + B_{21}) + m^2(B_{12} + B_{22}) \\ &= m^2\left(\Big(\frac{n}{m}\Big)^2(B_{11} + B_{21}) + B_{12} + B_{22}\right) := m^2C \end{aligned} \tag{2.10}$$

Pomocou (2.9) a vety o spojitej transformácii máme

$$C \xrightarrow{s.j.} \lambda^2(p^2 + (1-p)^2) + p^2 + (1-p)^2 = (\lambda^2+1)(2p^2 - 2p + 1). \tag{2.11}$$

S využitím (2.10) a (2.11) platí

$$\begin{aligned} mU_2 &= \frac{4m^3 C}{(n+m)(n+m-2)(n+m-3)} \\ &= \frac{4m^3 C}{m(\frac{n}{m}+1)m(\frac{n}{m}+1-\frac{2}{m})m(\frac{n}{m}+1-\frac{3}{m})} \xrightarrow{s.j.} \frac{4(\lambda^2+1)(2p^2-2p+1)}{(\lambda+1)^3}. \end{aligned}$$

Konvergencia skoro isto implikuje konvergenciu v pravdepodobnosti, a preto platí požadovaný vzťah (2.6).

□

Uvažujme znovu spermutovaný náhodný výber 1.15. V testovej statistike U_1, aplikovanej na 1.15, môžeme zanedbať konstantu závislú na m a n. Získame tak výraz

$$U_{1,Perm} = (\hat{p}_1 - \hat{p}_2)^2. \tag{2.12}$$

Vidíme, že permutačný test založený na statistike U_1 by bol ekvivalentný permutačnej verzii χ^2-testu. Jednoduchým výpočtom je možné overiť, že hodnota U_2 sa pri aplikácii na spermutovaný náhodný výbere zmení. Z tohoto dôvodu nie je permutačný χ^2-test ekvivalentný s USP testom s testovou statistikou (2.5).

2.2 Fisherov faktoriálový test

Fisherov faktoriálový test je možné použiť na testovanie hypotézy nezávislosti pre obecnú kontingenčnú tabuľku veľkosti $I \times J$, no v tejto práci uvedieme len jeho verziu pre tabuľky veľkosti 2×2. Tento test patrí medzi presné testy, pretože dodržiava požadovanú hladinu α presne. Typicky ale býva konzervatívy. Fisherov faktoriálový test úzko súvisí s USP testom, ako neskôr nahliadneme.

Uvažujme tabuľku 2.2. Ďale predpokladajme, že sú dané marginálne četnosti. Za platnosti tejto podmienky určuje četnosť v prvej bunke tabuľky o_{11} četnosti v zvyšných troch bunkách. Označme ako P_{Fish} podmienenú pravdepodobnosť, že vznikne tabuľka s četnostami o_{11},o_{12},o_{21},o_{22} za podmienky, že sú marginálne četnosti o_{1+},o_{2+},o_{+1},o_{+2} dané. Za nulovej hypotézy platí nasledujúci vzťah (Agresti, 2018, str.47)

$$\mathsf{P}_{Fish} = \frac{\binom{o_{1+}}{o_{11}}\binom{o_{2+}}{o_{+1}-o_{11}}}{\binom{N}{o_{+1}}}.$$

Kritický obor Fisherova faktoriálového testu odvodíme nasledovne, viz (Anděl, 2007). Označme

$$\beta = \frac{p_{11}p_{22}}{p_{12}p_{21}}, \qquad b = \frac{o_{11}o_{22}}{o_{12}o_{21}},$$

kde β budeme nazývať pomer šancí a b odhad teoretického pomeru šancí. Ďalej definujme logaritmickú interakciu $d = \ln(b)$ a teoretickú logaritmickú interakciu $\delta = \ln(\beta)$.

Pri testu hypotézy nezávislosti sčítame pravdepodobnosti P_{Fish} pre tie tabuľky, ktoré majú rovnaké marginálne četnosti ako pôvodná tabuľka a pre ktoré zároveň platí, že absolútne hodnoty ich logaritmických interakcií sú väčšie alebo rovné ako absolútna hodnota z logaritmickej interakcie pôvodnej tabuľky. Ak je tento súčet menší alebo rovný ako α, tak zamietame hypotézu H_0.

Príklad

Vráťme sa k tabuľke 2.2, zachytávajúcu výsledky štúdie, ktorá mala za úlohu zistiť súvislosť medzi užívaním aspirinu a rizikom infarktu. Prevedieme Fisherov faktoriálový test. Ako prvé vypíšeme všetky tabuľky, ktoré majú rovnaké marginálne četnosti ako tabuľka 2.2. Pre každú tabuľku spočítame logaritmickú interakciu d a hodnotu P_{Fish} (v tabuľkách ju budeme označovať pre prehľadnosť len ako P).

0	8
10	6

$d = -\infty$
$\mathsf{P} \doteq 0.00656$

1	7
9	7

$d \doteq$-2.197
$\mathsf{P} \doteq 0.04666$

2	6
8	8

$d \doteq$-1.098
$\mathsf{P} \doteq 0.11432$

3	5
7	9

$d \doteq$ -0.26
$\mathsf{P} \doteq$ 0.12472

4	4
6	10

$d \doteq 0.511$
$\mathsf{P} \doteq 0.06496$

5	3
5	11

$d \doteq 1.299$
$\mathsf{P} \doteq 0.1599$

6	2
4	12

$d \doteq 2.197$
$\mathsf{P} \doteq 0.00171$

7	1
3	13

$d \doteq$ 3.412
$\mathsf{P} \doteq 0.00653$

8	0
2	14

$d = \infty$
$\mathsf{P} \doteq 0.0051$

Následne sčítame všetky pravdepodobnosti P, u ktorých je absolútna hodnota logaritmickej interakcie d väčšia alebo rovná ako 0.26 (číslo 0.26 predstavuje absolútnu hodnotu logaritmickej interakcie pôvodnej tabuľky 2.2). V našom prípade spočítame hodnoty P všetkých tabuliek. Obdržíme tak číslo 0.53046, ktoré je väčšie ako hladina $\alpha = 0.05$. Hypotézu nezávislosti teda nemôžeme zamietnuť na tejto hladine.

Súvislosť Fisherova Faktoriálového testu a USP testu

Všimnime si, že tabuľky, ktoré majú rovnaké marginálne četnosti ako pôvodná tabuľka, na ktorú aplikujeme Fisherov faktoriálový test, odpovedajú kontingenčným tabuľkám, ktoré vznikli na základne spermutovaných dát 1.15. Pri USP teste aplikujeme testovú statistiku USP testu na každú kontingenčnú tabuľku, ktorá vznikla zo spermutovaných dát. P-hodnotu testu určíme na základe poradia hodnoty realizácie testovej statistiky na pôvodných dátach, ako je vysvetlené v 1.4.2. Naopak, pri Fisherovom faktoriálovom teste nás zaujíma poradie logaritmickej interakcie pôvodnej tabuľky medzi logaritmickými interagciami zvyšných tabuliek a zároveň ich príslušné hodnoty P_{Fish}.

2.2.1 Riadková a stĺpcová intepretácia

Tabuľku 2.3 je možné interpretovať ako realizáciu dvoch nezávislých binomických rozdelení po riadkoch alebo po stĺpcoch. Môžeme sa zaoberať otázkou či pri testovaní rovnosti parametrov p_1 a p_2 záleží, ktorú zo zmienených interpretácií zvolíme. Jednoduchým výpočtom môžeme overiť, že pri použití χ^2-testu nezávislosti na interpretácii nezáleží. Rovnako nezáleží na riadkovej a stĺpcovej interpretácii pri použití Fisherova faktoriálového testu a testu pomerom vierohodností. Pre testovú statistiku USP testu $U_{2\times 2}$ na zvolenej interpretácii záleží.

Pre štvorpoľnú kontingenčnú tabuľku môžeme uvažovať aj statistiky založené na rozdielu pravdepodobností, podielu pravdepodobností a relatívnom riziku, ktoré uvedieme v nasledujúcom odstavci.

Označme

$$d = p_1 - p_2, \qquad r = \frac{p_1}{p_2}, \qquad o = \frac{p_1(1-p_2)}{p_2(1-p_1)},$$

kde d,r a o nazývame postupne *rozdiel pravdepodobností*, *podiel pravdepodobností* a *pomer šancí*. Rovnosť $p_1 = p_2$ nastáva práve vtedy keď je splnený aspoň jeden zo vzťahov

$$d = 0, \qquad r = 1, \qquad o = 1. \tag{2.13}$$

Uvažujme postupne testové statistiky založené na rozdielu pravdepodobností, podielu pravdepodobností a pomeru šancí (viz (Omelka, 2022, str. 134-137))

$$T_d = \frac{\hat{d}}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}}, \tag{2.14}$$

$$T_r = \frac{\log \hat{r}}{\sqrt{\frac{1-\hat{p}_1}{n\hat{p}_1} + \frac{1-\hat{p}_2}{m\hat{p}_2}}}, \tag{2.15}$$

$$T_o = \frac{\log \hat{o}}{\sqrt{\hat{V}_0}}. \tag{2.16}$$

Znovu môžeme jednoduchým výpočtom overiť, že pri použití testovej statistiky (2.16) na interpretácii nezáleží. Naopak, pri teste založenom na statistike (2.14) alebo (2.15) na zvolenej interpretácii záleží.
Pre prehľadnosť zhrnieme výsledky do tabuľky 2.4.

	χ^2-test	G-test	Fish. test	USP test	T_d	T_r	T_o
záleží na interpretácii	NIE	NIE	NIE	ANO	ANO	ANO	NIE

Tabulka 2.4: riadková a stĺpcová interpretácia testov

2.3 Aplikácia testov na reálnych dátach

V tejto kapitole predvedieme testovanie hypotézy nezávislosti na reálnych dátach s využitím testov, ktoré sme predstavili v práci. Jednotlivé testy budeme realizovať pomocou softwaru R Core Team (2021). Testované dáta pochádzajú zo známej harvardskej štúdie, ktorá skúmala vplyv aspirinu na výskyt infarktu. Testy prevedieme na hladine $\alpha = 0.05$.

Physicians' Health Study

Dáta pochádzajú z randomizovanej štúdie s názvom *Physicians' Health Study* (Agresti, 2018, str.30), ktorá testovala či môže pravidelná konzumácia aspirinu zabrániť vzniku infarktu myokardu. Štúdia prebiehala 5 rokov a pozorovala účinky lieku na 22 071 pacientoch. Účastníci boli náhodne rozdelení do dvoch skupín. Jednej skupine bol podávaný aspirin a druhá skupina dostávala placebo. Výsledky pozorovania môžeme nájsť v tabuľke 2.5.

	Infarkt	Bez infarktu	Spolu
Placebo	189	10 845	11 034
Aspirin	104	10 933	11 037

Tabulka 2.5: výsledky štúdie závislosti užívania aspirinu a infarktu

Označíme hypotézu a alternatívu

$$H_0 : \{ \text{ výskyt infarktu nezávisí na konzumácii aspirinu } \},$$
$$H_1 : \{ \text{ výskyt infarktu závisí na konzumácii aspirinu } \}.$$

Na testovanie hypotézy H_0 proti alternatíve H_1 využijeme USP test (počet permutácií $B = 999$), χ^2-test, G-test (test pomerom vierohodností) a Fisherov faktoriálový test. Výsledné p-hodnoty zapíšeme do tabuľky 2.6.

test:	USP	χ^2-test	G-test	Fisherov test
p-hodnota:	0.001	7.7×10^{-7}	4.7×10^{-7}	5.03×10^{-7}

Tabulka 2.6: p-hodnoty testov

P-hodnoty v tabuľke sú menšie ako hladina $\alpha = 0.05$, a teda nulovú hypotézu môžeme zamietnuť na tejto hladine. Ukázali sme, že je štatisticky preukázateľná závislosť medzi konzumáciou aspirinu a výskytom infarktu.

Appendix

U-statistiky

Mnoho testových statistík je založených na *U-statistikách*. Medzi najznámejšie patria výberový priemer $\bar{X}_n$ a výberový rozptyl ${S_n}^2$. Definície spolu s príkladmi v celej kapitole sú čerpané z publikácie (Serfling, 2009, kapitola 5).

Definícia U-statistiky

Definícia 1 (U-statistika)**.** *Nech $X_1,...,X_N$ je náhodný výber s rozdelením F o rozsahu N. Pre symetrickú funkciu $h = h(x_1,...,x_m)$, kde $m \leq N$, definujeme U-statistiku ako*

$$U_n = U(X_1,...,X_n) = \frac{1}{\binom{N}{m}} \sum_{P_{m,n}} h(X_{i_1},...,X_{i_m}), \tag{2.17}$$

kde sčítavame cez množinu $P_{m,n}$ všetkých m-prvkových podmnožín $(i_1,...,i_m)$ z množiny $\{1,...,N\}$. Funkciu h nazývame jadro U-statistiky.

Uvažujme náhodný výber $X_1,...,X_n$ s rozdelením daným distribučnou funkciou F. Ďalej uvažujme parameter $\theta = \theta(F)$, pre ktorý existuje nestranný odhad. Inými slovami, parameter $\theta(F)$ môžeme písať ako

$$\theta(F) = E_F\{h(X_1,...,X_m)\}, \qquad m \leq N$$

kde funkciu $h = h(x_1,...,x_m)$ budeme nazývať *jadro* parametru $\theta(F)$. Všimnime si, že h nemusí byť funkciou celého náhodného výberu, ale iba jeho prvých m pozorovaní, kde $m \leq N$. Z tohoto dôvodu nie je funkcia h dobrým odhadom parametru θ. Definujme nový odhad ako priemer hodnôt funkcie $h(x_1,...,x_m)$ cez všetky m-prvkové permutácie z nášho náhodného výberu. Takto definovaný odhad nazývame *U-statistika*.

Poznámka. V definícii U-statistiky uvažujeme jadro h symetrické, teda také $h(x_{i1},...,x_{im})$, ktoré má rovnakú hodnotu pre všetky permutácie z množiny $P_{m,n}$. V opačnom prípade môže byť nahradené symetrickým jadrom

$$\frac{1}{m!} \sum_{P_m} h(x_{i_1},...,x_{i_m}), \tag{2.18}$$

kde P_m označuje množinu všetkých m-prvkových permutácií postupnosti $x_1,...,x_m$.

Záver

Práca sa zaoberá testami nezávislosti pre dáta, ktoré pochádzajú z dvojrozmerného diskrétneho rozdelenia. Cieľom práce bolo predstaviť základné testy nezávislosti a skúmať vzťahy medzi nimi. V práci dopĺňame dôležité kroky niektorých dôkazov (najmä v dôkazu vety 1) a uviedli sme alternatívny dôkazy vety 2. Podrobne sme odvodili súvislosť Pearsonova χ^2-testu s χ^2-divergenciou a súvislosť testu pomerom vierohodností (G-testu) s Kullbackovou-Leiblerovou divergenciou. Toto odvodenie poskytuje priamočiarejšie nahliadnutie testovej statistiky USP testu, ktorá je uvedená v kapitole 1.4.1.

V práci sme sa podrobne zamerali na problematiku testovania zhody dvoch nezávislých binomických rozdelení pomocou štvorpoľnej kontingenčnej tabuľky. Medzi hlavné prínosy v tejto kapitole patrí prepis testovej statistiky USP testu pomocou odhadov parametrov p_1 a p_2 a následné odvodenie jeho asymptotického rozdelenia. Prácu sme ilustrovali názornými príkladmi spolu s ukážkami praktického použitia testov. Na záver sme pridali ukážku realizácie testov pomocou softwaru R Core Team (2021).

V ďalšom výskume by bolo zaujimavé porovnať USP test s permutačnou verziou Pearsonova χ^2-testu. Zaujimavým pozorovaním by bolo aj porovnanie jednotlivých testov pomocou simulačnej štúdie.

Zoznam použitej literatúry

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons, Hoboken, New Jersey.

Anděl, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Berrett, T. B. a Samworth, R. J. (2021). USP: an independence test that improves on Pearson's chi-squared and the G-test. *Proceedings of the Royal Society A*, **477**(2256), 20210549.

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

Berrett, T. B., Kontoyiannis, I. a Samworth, R. J. (2021b). Optimal rates for independence testing via u-statistic permutation tests. *The Annals of Statistics*, **49**(5), 2457–2490.

Dragomir, S. S., Gluscevic, V. a Pearce, C. E. M. (2001). Csiszár f-divergence, Ostrowski's inequality and mutual information. *Nonlinear Analysis-Theory Methods and Applications*, **47**(4), 2375–2386.

Omelka, M. (2022). Matematická statistika 1, Poznámky k přednášce. URL `https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmsa331/ms1.pdf`. dátum prístupu: 1.5.2022, dátum poslednej úpravy: 25.1.2022.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Salmaso, L. a Pesarin, F. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, Chippenham, Wiltshire, United Kingdom.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*. John Wiley & Sons, New York.