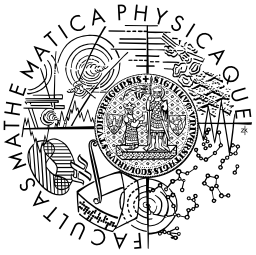


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Petr Míchal

Testy pro párová kategoriální data

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Arnošt Komárek, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2017

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

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

Rád bych poděkoval vedoucímu práce doc. RNDr. Arnoštu Komárkovi, Ph.D. za jeho vstřícnost, cenné rady, návrhy na zlepšení a věnovaný čas.

Název práce: Testy pro párová kategoriální data

Autor: Petr Míchal

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá testy pro párová kategoriální data. Testovanými vlastnostmi jsou shoda marginálních rozdělení a symetrie příslušné tabulky pravděpodobností. Nejprve je zavedeno a popsáno multinomické rozdělení a kontingenční tabulky. V další části se zabýváme dichotomickými párovými kategoriálními daty, odvodíme McNemarův test a popíšeme test pro malé rozsahy výběrů. Dále uvádíme testy pro obecná párová kategoriální data, nejprve Stuartův test, dále Bhapkarův test. Na testování symetrie tabulky pravděpodobností ukážeme test, který odvodil Bowker. Na závěr provedeme simulace McNemarova testu v programu R.

Klíčová slova: kategoriální data, kontingenční tabulky, McNemarův test, shoda marginálních rozdělení

Title: Tests for Paired Categorical Data

Author: Petr Míchal

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this paper we deal with paired categorical data. We will test marginal homogeneity and symmetry of corresponding probability table. At first, we describe multinomial distribution and contingency tables. In the next section, we deal with dichotomic paired categorical data, we derive McNemar's test and describe test for small sample sizes. Further, we state tests for general paired categorical data, Stuart's and Bhapkar's test are described. We then state test derived by Bowker, which is used for testing symmetry of probability table. In the last section, we show simulations of McNemar's test in software R.

Keywords: categorical data, contingency tables, McNemar test, marginal homogeneity

Obsah

Seznam použitých zkratek

$\xrightarrow{d}$	konvergence v distribuci
$\xrightarrow{P}$	konvergence v pravděpodobnosti
$\boldsymbol{I}_k$	jednotková matice o rozměrech $k \times k$
$\mathrm{Var}(\boldsymbol{X})$	varianční matice náhodného vektoru $\boldsymbol{X}$
$\mathrm{rank}(A)$	hodnost matice A
$\mathrm{tr}(A)$	stopa matice A
$\chi^2_k(\alpha)$	α-kvantil rozdělení χ^2 s k stupni volnosti
u_α	α-kvantil rozdělení $\mathrm{N}(0,1)$

Úvod

Kategoriální data vznikají v různorodých odvětvích lidské činnosti. Náhodné veličiny reprezentující tato data popisují příslušnost sledovaného objektu do některé z možných kategorií. Proto se u kategoriálních dat zpravidla testují odlišné hypotézy než u dat kvantitativních, která mají většinou jasnou numerickou interpretaci.
Párová kategoriální data se potom objevují např. při opakovaných měřeních nebo dotazníkových šetřeních, kde zastupují odpovědi dotazovaných osob na dvě otázky se stejnými možnými odpověďmi. Naším cílem je testovat hypotézy o vlastnostech rozdělení odpovědí na jednotlivé otázky. V této práci se budeme zabývat testováním shody marginálních rozdělení diskrétně rozděleného náhodného vektoru reprezentujícího kategoriální data. Druhou sledovanou vlastností bude symetrie příslušné tabulky pravděpodobností. V práci ukážeme, jaký je mezi těmito vlastnostmi vztah, a uvedeme příslušné testy. Pro jeden z těchto testů také na závěr provedeme simulace pomocí softwaru R.
V první kapitole zavedeme multinomické rozdělení, ukážeme některé jeho vlastnosti a asymptotické rozdělení. To následně využijeme při testech dobré shody. Ukážeme variantu těchto testů při neznámých parametrech, kterou budeme dále v práci využívat. Dále v této kapitole zavedeme kontingenční tabulky a k nim potřebné značení.
V druhé kapitole se budeme zabývat dichotomickými párovými kategoriálními daty, pro něž obě sledované vlastnosti splývají, jak ukážeme. Dále v této kapitole odvodíme McNemarův test a test pro malé rozsahy výběrů.
Ve třetí kapitole budeme uvažovat obecná párová kategoriální data. Nejprve se budeme zabývat testováním shody marginálních rozdělení, zde se seznámíme se Stuartovým testem a uvedeme ještě jiný test, který odvodil Bhapkar. Dále budeme testovat symetrii tabulky pravděpodobností. Ukážeme si test, který odvodil Bowker, založený na testech dobré shody.
Ve čtvrté kapitole se budeme opět zabývat McNemarovým testem. Provedeme simulace v programu R a spočteme dosaženou hladinu tohoto testu a konvergenci testové statistiky k limitnímu rozdělení při různých pravděpodobnostech a rozsazích výběru.
V poslední části uvádíme věty, které v práci používáme, s odkazem na příslušný důkaz.

1. Multinomické rozdělení

V úvodní kapitole zavedeme pojmy a označení, se kterými budeme pracovat v následujících kapitolách.

1.1 Zavedení

S multinomickým rozdělením budeme pracovat v celé práci, proto ho v této části zavedeme a ukážeme některé jeho vlastnosti a asymptotické chování.

Definice 1. *Buďte $k \geq 2$, n přirozená čísla, vektor $\boldsymbol{p} = (p_1, \ldots, p_k)^\top$ takový, že $p_i \in (0,1)$ pro každé $i = 1, \ldots, n$ a $p_1 + \cdots + p_k = 1$. Řekneme, že náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_k)^\top$ s nezápornými celočíselnými hodnotami má multinomické rozdělení s parametry $\boldsymbol{p}$ a n, značíme $\boldsymbol{X} \sim \mathrm{Mult}_k(n, \boldsymbol{p})$, jestliže*

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k},$$

pro $x_i \in \{0, 1, \ldots, n\}, i = 1, \ldots, k$ a $x_1 + \cdots + x_k = n$.

Multinomické rozdělení můžeme chápat jako rozdělení počtu kuliček v k přihrádkách. Představme si, že máme $k \geq 2$ přihrádek, a n kol. V každém kole nezávisle na ostatních umístíme do jedné z k přihrádek jednu kuličku, přičemž pravděpodobnosti přidělení kuličky do jednotlivých přihrádek jsou dány vektorem $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)^\top$, kde $p_1 + \cdots + p_n = 1$. Potom vektor $\boldsymbol{X} = (X_1, \ldots, X_k)^\top$ udávající výsledné počty kuliček v jednotlivých přihrádkách má multinomické rozdělení $\mathrm{Mult}_k(n, \boldsymbol{p})$.

Příklad. Mějme spravedlivou šestistěnnou kostku, tj. pravděpodobnosti padnutí jednotlivých hodnot jsou všechny $\frac{1}{6}$. Po n hodech má vektor $\boldsymbol{X}$ počtu padnutí jednotlivých čísel multinomické rozdělení $\mathrm{Mult}_6(n, \boldsymbol{p})$, kde $\boldsymbol{p} = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)^\top$.

Uveďme si některé základní vlastnosti multinomického rozdělení.

Věta 1. *Buď $\boldsymbol{X} \sim \mathrm{Mult}_k(n, \boldsymbol{p})$ a $1 \leq l < k$, pak pro marginální rozdělení $X_1, \ldots, X_l$ platí*

$$P(X_1 = x_1, \ldots, X_l = x_l) = \\ = \frac{n!}{x_1! \ldots x_l! (n - x_1 - \cdots - x_l)!} p_1^{x_1} \ldots p_l^{x_l} (1 - p_1 - \cdots p_l)^{n - x_1 - \cdots - x_l}$$

a pro podmíněné rozdělení $X_1, \ldots, X_{l-1}$ za podmínky $X_l = x_l, \ldots, X_k = x_k$ platí

$$P(X_1 = x_1, \ldots, X_{l-1} = x_{l-1} | X_l = x_l, \ldots, X_k = x_k) = \\ = \frac{(n - x_l - \cdots - x_k)!}{x_1! \ldots x_{l-1}!} \prod_{i=1}^{l-1} \left(\frac{p_i}{1 - p_l - \cdots - p_k} \right)^{x_i}$$

pro $x_1 + \cdots + x_{l-1} = n - x_l - \cdots - x_k$ *a* $x_i \in \{0, 1, \ldots, n - x_l - \cdots - x_k\}$, $i = 1, \ldots, l-1$.
Dále platí

$$\begin{aligned} &\mathsf{E}\, X_i = np_i, \qquad \mathrm{var} X_i = np_i\left(1-p_i\right), \quad 1 \le i \le n, \\ &\mathrm{cov}\left(X_i, X_j\right) = -np_ip_j, \quad 1 \le i \ne j \le n. \end{aligned}$$

Důkaz. Interpretujme hodnotu multinomicky rozděleného náhodného vektoru jako počet kuliček v k přihrádkách po n kolech rozdělování. Pří počítání marginálního rozdělení se vlastně zajímáme o počet kuliček umístěných v prvních l přihrádkách, všechny ostatní přihrádky zahrneme do jedné, ve které bude zbylých $n - x_1 - \cdots - x_l$ kuliček. Odtud plyne vzorec pro marginální pravděpodobnosti. Tvrzení o podmíněném rozdělení ověříme výpočtem, z definice podmíněné pravděpodobnosti je

$$\begin{aligned} &\mathsf{P}(X_1 = x_1, \ldots, X_{l-1} = x_{l-1} | X_l = x_l, \ldots, X_k = x_k) = \\ &\qquad = \mathsf{P}(X_1 = x_1, \ldots, X_k = x_k) / \mathsf{P}(X_l = x_l, \ldots, X_k = x_k). \quad (1.1) \end{aligned}$$

Z definice a již dokázané části máme

$$\begin{aligned} &\mathsf{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}, \\ &\mathsf{P}(X_l = x_l, \ldots, X_k = x_k) = \\ &\qquad = \frac{n!}{x_l! \ldots x_k! (n - x_l - \ldots - x_k)!} p_l^{x_l} \ldots p_k^{x_k} (1 - p_l - \ldots p_k)^{n - x_l - \ldots - x_k}, \end{aligned}$$

platí $n - x_l - \cdots - x_k = x_1 + \cdots + x_{l-1}$, po dosazení do (1.1) a úpravě dostáváme vzorec ve větě.
Dále z již dokázané první části si všimněme, že $X_i \sim \mathrm{Bi}\left(n, p_i\right), i = 1, \ldots, k$. Odtud plynou vzorce pro rozptyl a střední hodnotu. Pro ověření kovariance uvažujme náhodné veličiny $Y_{i,j} \in \{0,1\}$ takové, že $Y_{i,j} = 1$, pokud jsme v j-tém kole přiřadili kuličku do i-té přihrádky, jinak $Y_{i,j} = 0,\ i = 1, \ldots, n$. Potom $\mathsf{E}\, Y_{i,j} = p_i$, pro $j \ne l$ jsou $Y_{i,j}$ a $Y_{m,l}$ nezávislé a platí

$$X_i = \sum_{j=1}^{n} Y_{i,j}.$$

Můžeme tedy psát

$$\begin{aligned} \mathrm{cov}(X_i, X_j) &= \mathrm{cov}\left(\sum_{m=1}^{n} Y_{i,m}, \sum_{r=1}^{n} Y_{j,r}\right) = \sum_{m=1}^{n}\sum_{r=1}^{n} \mathrm{cov}\left(Y_{i,m}, Y_{j,r}\right) \\ &= \sum_{m=1}^{n} \mathrm{cov}\left(Y_{i,m}, Y_{j,m}\right) = \sum_{m=1}^{n} \left(\mathsf{E}\, Y_{i,m}Y_{j,m} - \mathsf{E}\, Y_{i,m}\, \mathsf{E}\, Y_{j,m}\right) \\ &= \sum_{m=1}^{n} \left(\mathsf{E}\, Y_{i,m}Y_{j,m} - p_ip_j\right) = -np_ip_j, \end{aligned}$$

kde jsme využili nezávislosti jednotlivých Y_i a dále v poslední rovnosti faktu, že v každém kole umisťujeme právě jednu kuličku, a tedy pro $i \neq j$ musí být jedna z veličin $Y_{i,m}, Y_{j,m}$ nulová, takže je nulová i střední hodnota jejich součinu.

□

Podívejme se nyní na varianční matici vektoru $\boldsymbol{X}$. Pro sloupcový vektor $\boldsymbol{a} \in \mathbb{R}^n$ budeme značit $\mathrm{diag}(\boldsymbol{a})$ matici, která má na diagonále vektor $\boldsymbol{a}$ a jinde 0, dále $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a} \cdot \boldsymbol{a}^\top$. Při tomto značení můžeme psát

$$\begin{aligned}\mathrm{Var}(\boldsymbol{X}) &= n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \ddots & & \vdots \\ -p_kp_1 & -p_kp_2 & \cdots & p_k(1-p_k) \end{pmatrix} = n\left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}\right) \\ &= n\left(\mathrm{diag}(\sqrt{\boldsymbol{p}})\boldsymbol{I}_k\mathrm{diag}(\sqrt{\boldsymbol{p}}) - \mathrm{diag}(\sqrt{\boldsymbol{p}})(\sqrt{\boldsymbol{p}})^{\otimes 2}\mathrm{diag}(\sqrt{\boldsymbol{p}})\right) \\ &= n\,\mathrm{diag}(\sqrt{\boldsymbol{p}})\left(\boldsymbol{I}_k - (\sqrt{\boldsymbol{p}})^{\otimes 2}\right)\mathrm{diag}(\sqrt{\boldsymbol{p}}).\end{aligned}$$

Ukážeme si ještě asymptotické vlastnosti multinomického rozdělení.

Věta 2. *Buď* $\boldsymbol{X} \sim \mathrm{Mult}_k(n, \boldsymbol{p}), \boldsymbol{p} = (p_1, \ldots, p_k)^\top$. *Pak pro* $n \to \infty$ *platí*

i) $\boldsymbol{Y}_n = \frac{1}{\sqrt{n}}(diag(\sqrt{\boldsymbol{p}}))^{-1}(\boldsymbol{X} - n\boldsymbol{p}) \xrightarrow{d} N_k(0, \boldsymbol{I}_k - (\sqrt{\boldsymbol{p}})^{\otimes 2})$,

ii) $\boldsymbol{Y}_n^\top \boldsymbol{Y}_n = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} \xrightarrow{d} \chi^2_{k-1}$.

Důkaz. (i) Můžeme psát $\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{Z}_i$, kde $\boldsymbol{Z}_i$ jsou nezávislé stejně rozdělené náhodné vektory z rozdělení $\mathrm{Mult}_k(1, \boldsymbol{p})$. Potom z centrální limitní věty (věta A.1) máme

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (\boldsymbol{Z}_i - \boldsymbol{p}) = \frac{1}{\sqrt{n}}(\boldsymbol{X} - n\boldsymbol{p}) \xrightarrow{d} \mathrm{N}_k(0, \mathrm{Var}(\boldsymbol{Z}_1)), \quad n \to \infty,$$

kde $\mathrm{Var}(\boldsymbol{Z}_1) = \mathrm{diag}(\sqrt{\boldsymbol{p}})\left(\boldsymbol{I}_k - (\sqrt{\boldsymbol{p}})^{\otimes 2}\right)\mathrm{diag}(\sqrt{\boldsymbol{p}})$. Matice $\mathrm{diag}(\sqrt{\boldsymbol{p}})$ je regulární, po přenásobení maticí $(\mathrm{diag}(\sqrt{\boldsymbol{p}}))^{-1}$ dostáváme (i).

(ii) Označme $\boldsymbol{\Sigma} = \boldsymbol{I}_k - (\sqrt{\boldsymbol{p}})^{\otimes 2}$, z *(i)* víme, že $\boldsymbol{Y}_n \xrightarrow{d} N_k(0, \boldsymbol{\Sigma})$. Ukážeme, že matice Σ je idempotentní, potom z věty A.5 je $\mathrm{rank}(\Sigma) = \mathrm{tr}(\Sigma)$, a tedy dostáváme $\mathrm{rank}(\Sigma) = k - \mathrm{tr}(\sqrt{\boldsymbol{p}} \cdot \sqrt{\boldsymbol{p}}^\top) = k - tr(\sqrt{\boldsymbol{p}}^\top \cdot \sqrt{\boldsymbol{p}}) = k - 1$. Platí

$$\begin{aligned}\boldsymbol{\Sigma}\boldsymbol{\Sigma} &= \left(\boldsymbol{I}_k - (\sqrt{\boldsymbol{p}})^{\otimes 2}\right) \cdot \left(\boldsymbol{I}_k - (\sqrt{\boldsymbol{p}})^{\otimes 2}\right) = \boldsymbol{I}_k - 2(\sqrt{\boldsymbol{p}})^{\otimes 2} + \sqrt{\boldsymbol{p}}(\sqrt{\boldsymbol{p}})^\top\sqrt{\boldsymbol{p}}(\sqrt{\boldsymbol{p}})^\top \\ &= \boldsymbol{I}_k - 2(\sqrt{\boldsymbol{p}})^{\otimes 2} + \sqrt{\boldsymbol{p}}(\sqrt{\boldsymbol{p}})^\top = \boldsymbol{\Sigma}.\end{aligned}$$

Z věty A.8 celkově dostáváme $\boldsymbol{Y}_n^\top \boldsymbol{I}_k \boldsymbol{Y}_n \xrightarrow{d} \chi^2_{k-1},\ n \to \infty$.

□

1.2 Testy dobré shody

Ve větě 2 jsme ukázali, že náhodný vektor $\boldsymbol{X} \sim \text{Mult}_k(n, \boldsymbol{p}),\ \boldsymbol{p} = (p_1, \ldots, p_k)^\top$ má asymptoticky normální rozdělení a pro příslušnou kvadratickou formu $\boldsymbol{Y}_n^\top \boldsymbol{Y}_n$, kde $\boldsymbol{Y}_n$ je z věty 2, platí

$$\chi^2 = \boldsymbol{Y}_n^\top \boldsymbol{Y}_n = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \xrightarrow{d} \chi^2_{k-1}.$$

Na tomto poznatku jsou založeny tzv. χ^2 testy dobré shody, pomocí nichž můžeme testovat platnost $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$, proti $H_1 : \boldsymbol{p} \neq \boldsymbol{p}_0$ pro nějaký vektor konstant $\boldsymbol{p}_0$. Tento test používá testovou statistiku χ^2, která za platnosti H_0 konverguje v distribuci k rozdělení χ^2_{k-1} a hypotézu zamítá, je-li napozorovaná hodnota $\chi^2 \geq \chi^2_{k-1}(1-\alpha)$, kde $\chi^2_{k-1}(1-\alpha)$ je $(1-\alpha)$-kvantil rozdělení χ^2 s $k-1$ stupni volnosti.

Často ale potřebujeme testovat, jestli je náš neznámý vektor $\boldsymbol{p}$ roven vektoru pravděpodobností $\boldsymbol{p}(\boldsymbol{a}) = (p_1(\boldsymbol{a}), \ldots, p_k(\boldsymbol{a}))^\top$, které závisí na neznámém parametru $\boldsymbol{a} \in \mathbb{R}^m$, tj. jestli $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{a})$. Pro tyto testy se používá tzv. metoda minimálního χ^2 resp. modifikovaná metoda minimálního χ^2. Nejprve upravíme statistiku χ^2, jež je nyní funkcí parametru $\boldsymbol{a}$:

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{X_i^2}{np_i} - 2\sum_{i=1}^{k} X_i + n\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} \frac{X_i^2}{np_i} - n.$$

Nyní budeme chtít χ^2 minimalizovat v proměnné $\boldsymbol{a}$, můžeme například řešit soustavu rovnic

$$\frac{\partial \chi^2(\boldsymbol{a})}{\partial a_j} = -\sum_{i=1}^{k} \frac{X_i^2}{np_i(\boldsymbol{a})^2} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1, \ldots, m.$$

Můžeme ale také vyjít z původního vyjádření χ^2, pro $j = 1, \ldots, m$ dostaneme rovnice

$$\frac{\partial \chi^2(\boldsymbol{a})}{\partial a_j} = -\sum_{i=1}^{k} \left(\frac{2(X_i - np_i(\boldsymbol{a}))}{p_i(\boldsymbol{a})} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} + \frac{(X_i - np_i(\boldsymbol{a}))^2}{np_i(\boldsymbol{a})^2} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} \right) = 0.$$

S rostoucím n klesá vliv druhého členu, proto se v praxi řeší jen soustava

$$\sum_{i=1}^{k} \frac{(X_i - np_i(\boldsymbol{a}))}{p_i(\boldsymbol{a})} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1, \ldots, m. \tag{1.2}$$

Dále pro každé $\boldsymbol{a}$ musí platit $p_1(\boldsymbol{a}) + \cdots + p_k(\boldsymbol{a}) = 1$, derivováním dostáváme vztah

$$\frac{\partial p_1(\boldsymbol{a})}{\partial a_j} + \cdots + \frac{\partial p_k(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1, \ldots, m.$$

S využitím této rovnosti dostáváme z (1.2) soustavu

$$\sum_{i=1}^{k} \frac{X_i}{p_i(\boldsymbol{a})} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1, \ldots, m, \tag{1.3}$$

neboť $n\sum_{i=1}^{k}\frac{\partial p_i(\boldsymbol{a})}{\partial a_j}=0$. Vyřešením této soustavy dostaneme vektor $\widehat{\boldsymbol{a}}_n$, který se nazývá odhad parametru $\boldsymbol{a}$ modifikovanou metodou minimálního χ^2. Použijeme-li tento získaný odhad $\widehat{\boldsymbol{a}}_n$ jako odhad parametru $\boldsymbol{a}$ ve statistice $\chi^2(\boldsymbol{a})$, pak tato statistika má asymptoticky χ^2 rozdělení s $k-m-1$ stupni volnosti, jak tvrdí následující věta.

Věta 3. *Buď $m<k-1$, $A\subset\mathbb{R}^m$ nedegenerovaný omezený uzavřený interval takový, že pro všechny body $\boldsymbol{a}\in A$ platí:*

i) $p_1(\boldsymbol{a})+\cdots+p_k(\boldsymbol{a})=1$,

ii) Existuje $c>0$ takové, že $p_i(\boldsymbol{a})>c^2$ pro $i=1,\ldots,k$,

iii) Pro každé $i=1,\ldots,k$ má funkce $p_i(\boldsymbol{a})$ spojité derivace $\frac{\partial p_i(\boldsymbol{a})}{\partial a_j}$, $\frac{\partial^2 p_i(\boldsymbol{a})}{\partial a_j\partial a_s}$ pro $j,s=1,\ldots,m$,

iv) Matice typu $k\times m$ s prvky $\left(\frac{\partial p_i(\boldsymbol{a})}{\partial a_j}\right)$ má hodnost m.

Nechť $\boldsymbol{a^0}$ je vnitřním bodem A. Označme $p_i^0=p_i(\boldsymbol{a^0})$. Předpodkládejme, že $\boldsymbol{X}=(X_1,\ldots,X_k)^\top\sim Mult_k(n;p_1^0,\ldots,p_k^0)$. Pak exstijí takové posloupnosti kladných čísel $\epsilon_n\to 0$ a $\delta_n\to 0$ při $n\to\infty$, že soustava (1.3) má s pravděpodobností alespoň $1-\epsilon_n$ právě jeden kořen $\widehat{\boldsymbol{a}}_n$ takový, že $\|\widehat{\boldsymbol{a}}_n-\boldsymbol{a}_0\|<\delta_n$. Existuje-li $\widehat{\boldsymbol{a}}_n$ pro všechna dostatečně velká n, má statistika $\chi^2(\widehat{\boldsymbol{a}}_n)$ asypmtoticky χ^2_{k-m-1}rozdělení při $n\to\infty$.

Důkaz. (viz Anděl, 1978, Věta XI.6).

□

Všimněme si ještě, že oproti χ^2 testu se známými parametry se v tomto případě počet stupňů volnosti sníží právě o dimenzi neznámého parametru.

1.3 Kontingenční tabulky

Mějme dvourozměrný diskrétně rozdělený náhodný vektor $(X,Y)^\top$ s hodnotami v $\{1,2,\ldots,I\}\times\{1,2,\ldots,J\}$ reprezentující kategoriální data a označme

$$\mathsf{P}\,(X=i,Y=j)=p_{i,j},\quad \mathsf{P}\,(X=i)=p_{i,+},\quad \mathsf{P}\,(Y=j)=p_{+,j}.$$

Potom platí

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

Takto označené pravděpodobnosti se obvykle zapisují do tabulky 1.1$\boldsymbol{a}$.

Dále předpokládejme, že máme náhodný výběr $(X_1,Y_1)^\top,\ldots,(X_n,Y_n)^\top$ z rozdělení vektoru $(X,Y)^\top$, kde X a Y jsou jako výše. Označme $N_{i,j}$ počet pozorování z našeho výběru, kde je $X_m=i$ a $Y_m=j$, $m=1,\ldots,n$ ($N_{i,j}$ se nazývají pozorované nebo napozorované četnosti). Dále označme $N_{i,+}=\sum_{j=1}^{J}N_{i,j}$ a $N_{+,j}=\sum_{i=1}^{I}N_{i,j}$. Zřejmě platí $n=\sum_{i=1}^{I}N_{i,+}=\sum_{j=1}^{J}N_{+,j}$. Pozorované četnosti $N_{i,j}$ se také obvykle zapisují do tabulek 1.1$\boldsymbol{b}$, ty se nazývají kontingenční tabulky.

Častá úloha je testování nezávislosti X a Y. V této práci se budeme ale zabývat tzv. párovými kategoriálními daty a úlohami specifickými pro tento typ dat. Za

Tabulka 1.1: Tabulky pravděpodobností a napozorovaných četností

a) Teoretické pravděpodobnosti

	Y			
X	1	...	J	Σ
1	$p_{1,1}$	...	$p_{1,J}$	$p_{1,+}$
⋮	⋮	⋱	⋮	⋮
I	$p_{I,1}$	...	$p_{I,J}$	$p_{I,+}$
Σ	$p_{+,1}$	...	$p_{+,J}$	1

b) Napozorované četnosti

	Y			
X	1	...	J	Σ
1	$N_{1,1}$	...	$N_{1,J}$	$N_{1,+}$
⋮	⋮	⋱	⋮	⋮
I	$N_{I,1}$	...	$N_{I,J}$	$N_{I,+}$
Σ	$N_{+,1}$	...	$N_{+,J}$	n

párová kategoriální data budeme považovat situaci, kdy X i Y mají hodnoty ve stejné množině $\{1,\ldots,I\}$, typicky reprezentující hodnoty jedné kategoriální proměnné zjišťované za dvou odlišných podmínek. Příslušné tabulky budou tedy čtvercové.
Budeme se zabývat následujími vlastnostmi:

(*i*) Symetrie tabulky pravděpodobností, tj. jestli $p_{i,j} = p_{j,i}$ pro všechna $i,j \in \{1,2,\ldots,I\}$;

(*ii*) Shoda marginálních rozdělení X a Y, tj. jestli $\mathsf{P}(X = i) = \mathsf{P}(Y = i)$, neboli $p_{i,+} = p_{+,i}$ pro všechna $i \in \{1,2,\ldots,I\}$.

Mezi těmito vlastnostmi je následující vztah.

Lemma 4. *Buď $(X,Y)^\top$ dvourozměrný diskrétně rozdělený náhodný vektor s hodnotami v množině $\{1,\ldots,I\}^2$. Je-li příslušná tabulka pravděpodobností symetrická, pak obě jeho složky mají shodné marginální rozdělení. Je-li $I = 2$, pak jsou obě vlastnosti ekvivalentní.*

Důkaz. Nechť je tabulka pravděpodobností X a Y symetrická, potom máme pro $i \in \{1,\ldots,I\}$ rovnosti $p_{i,+} = \sum_{j=1}^{I} p_{i,j} = \sum_{j=1}^{I} p_{j,i} = p_{+,i}$. Dále pro $I = 2$ máme $p_{1,+} = p_{+,1} \Leftrightarrow p_{1,1} + p_{1,2} = p_{1,1} + p_{2,1} \Leftrightarrow p_{1,2} = p_{2,1} \Leftrightarrow p_{2,+} = p_{+,2}$.

□

2. Dichotomická párová kategoriální data

V této časti se budeme zabývat případem, kdy složky náhodného vektoru $(X,Y)^\top$ mohou nabývat pouze dvou různých hodnot, budeme tedy pracovat s tabulkami 2×2. Z lemmatu 4 je v tomto případě symetrie tabulky pravděpodobností a shoda marginálních rozdělení ekvivalentní.

Nechť v celé kapitole je $(X,Y)^\top$ dvourozměrný diskrétně rozdělený náhodný vektor s hodnotami v množině $\{1,2\}^2$ a $(X_1,Y_1)^\top,\dots,(X_n,Y_n)^\top$ náhodný výběr z rozdělení vektoru $(X,Y)^\top$.

2.1 McNemarův test

Nyní odvodíme McNemarův test, který se používá na testování shody marginálních rozdělení, v případě uvažovaném v této kapitole tedy též symetrie sdruženého rozdělení.

Označme $\delta = p_{+,1}-p_{1,+} = p_{2,1}-p_{1,2}$, budeme testovat hypotézu

$$H_0:\delta=0,\quad H_1:\delta\neq 0,\quad \text{ekvivalentně}\quad H_0:p_{+,1}=p_{1,+},\quad H_1:p_{+,1}\neq p_{1,+}.$$

Pro δ uvažme přímočarý bodový odhad daný jako rozdíl příslušných relativních četností, tj: $D_n = (\widehat{p}_{+,1}-\widehat{p}_{1,+}) = \frac{1}{n}\left(N_{+,1}-N_{1,+}\right)$. Nyní ukážeme některé jeho vlastnosti.

Tvrzení 5. *Pro $D_n = (\widehat{p}_{+,1}-\widehat{p}_{1,+})$ platí*

i) D_n je je nestranný a konzistentní odhad δ,

ii) $\operatorname{var} D_n = \frac{1}{n}\left[p_{2,1}+p_{1,2}-(p_{2,1}-p_{1,2})^2\right]$,

iii) $\frac{D_n-\delta}{\sqrt{\operatorname{var} D_n}} \xrightarrow{d} \mathrm{N}(0,1),\quad n\to\infty.$

Důkaz. Pro $i=1,\dots,n$ definujme náhodnou veličinu Z_i následovně

$$Z_i = \begin{cases} 1, & \text{je-li } (X_i,Y_i)=(2,1),\\ 0, & \text{je-li } (X_i,Y_i)=(1,1) \text{ nebo } (2,2),\\ -1, & \text{je-li } (X_i,Y_i)=(1,2).\end{cases}$$

Veličiny $Z_i, i=1,\dots,n$ jsou nezávislé a stejně rozdělené díky nezávislosti jednotlivých (X_i,Y_i) a $D_n = \frac{1}{n}\sum_{i=1}^n Z_i = \overline{Z}_n$.
Dále s využitím věty 1 máme $\mathsf{E}\ D_n = \frac{1}{n}\left(\mathsf{E}\ N_{+,1}-\mathsf{E}\ N_{1,+}\right) = \frac{1}{n}(np_{+,1}-np_{1,+}) = \delta$

a D_n je tedy nestranný odhad δ. Pro rozptyl D_n máme

$$\begin{aligned}\operatorname{var} D_n &= \frac{1}{n^2}\operatorname{var}\left(N_{+,1} - N_{1,+}\right) = \frac{1}{n^2}\operatorname{var}\left(N_{2,1} - N_{1,2}\right)\\ &= \frac{1}{n^2}\left(\operatorname{var} N_{2,1} + \operatorname{var} N_{1,2} - 2\operatorname{cov}\left(N_{2,1}, N_{1,2}\right)\right)\\ &= \frac{1}{n^2} n\left(p_{2,1}\left(1-p_{2,1}\right) + p_{1,2}\left(1-p_{1,2}\right) + 2p_{2,1}p_{1,2}\right)\\ &= \frac{1}{n}\left(p_{2,1} + p_{1,2} - \left(p_{2,1} - p_{1,2}\right)^2\right).\end{aligned}$$

Odtud plyne také konzistence D_n neboť $\mathsf{E}\ D_n = \delta$ a $\operatorname{var} D_n \to 0$ pro $n \to \infty$.
Z centrální limitní věty (věta A.1) máme pro nezávislé stejně rozdělené veličiny Z_i

$$\sqrt{n}\frac{\overline{Z}_n - \mathsf{E}\ Z_1}{\sqrt{\operatorname{var} Z_1}} \xrightarrow{d} \mathrm{N}(0,1), \quad n \to \infty.$$

K ověření (iii) výraz upravíme a dostaneme

$$\sqrt{n}\frac{\overline{Z}_n - \mathsf{E}\ Z_1}{\sqrt{\operatorname{var} Z_1}} = \frac{D_n - \delta}{\sqrt{\frac{1}{n}\operatorname{var} Z_1}} = \frac{D_n - \delta}{\sqrt{\operatorname{var} D_n}} \xrightarrow{d} \mathrm{N}(0,1), \quad n \to \infty. \tag{2.1}$$

□

Pravděpodobnosti $p_{i,j}$ jsou ale neznámé, proto uvažujme jejich příslušné odhady $\widehat{p}_{i,j} = \frac{1}{n}N_{i,j}$. Tyto odhady jsou konzistentní, neboť $\mathsf{E}\ \widehat{p}_{i,j} = \frac{1}{n}\mathsf{E}\ N_{i,j} = p_{i,j}$ a

$$\operatorname{var}\widehat{p}_{i,j} = \frac{1}{n^2}\operatorname{var} N_{i,j} = \frac{1}{n}p_{i,j} \to 0, \quad n \to \infty.$$

Označíme-li $\sigma_n^2 = \operatorname{var} D_n$ a $\widehat{\sigma}_n^2 = \frac{1}{n}\left(\widehat{p}_{2,1} + \widehat{p}_{1,2} - \left(\widehat{p}_{2,1} - \widehat{p}_{1,2}\right)^2\right)$, pak

$$\frac{D_n - \delta}{\sqrt{\operatorname{var} D_n}} = \frac{D_n - \delta}{\sigma_n}\cdot\frac{\widehat{\sigma}_n}{\widehat{\sigma}_n} = \frac{D_n - \delta}{\widehat{\sigma}_n}\cdot\frac{\widehat{\sigma}_n}{\sigma_n}.$$

Odhady $\widehat{p}_{i,j}$ pro jednotlivá i, j jsou konzistentní, tedy $\widehat{p}_{i,j} \xrightarrow{P} p_{i,j}$, z věty A.3 o spojité transformaci také $\widehat{\sigma}_n^2 \xrightarrow{P} \sigma_n^2$, a tedy $\frac{\widehat{\sigma}_n}{\sigma_n} \xrightarrow{P} 1$. Z Cramér-Sluckého věty (věta A.4) dostáváme

$$\frac{D_n - \delta}{\widehat{\sigma}_n} \xrightarrow{d} \mathrm{N}(0,1), \quad n \to \infty.$$

Odtud mimo jiné můžeme odvodit intervalový odhad pro δ o asymptotickém pokrytí $1-\alpha$. Označíme-li u_α α-kvantil rozdělení N(0,1), pak daný interval má zřejmě tvar

$$\left(D_n - u_{1-\alpha/2}\widehat{\sigma}_n,\ D_n + u_{1-\alpha/2}\widehat{\sigma}_n\right).$$

Za platnosti testované hypotézy $\delta = 0$ platí následující věta.

Věta 6. *Nechť platí* $H_0 : p_{+,1} = p_{1,+}$. *Potom*

$$M_n^2 = \frac{(N_{2,1} - N_{1,2})^2}{N_{2,1} + N_{1,2}} \xrightarrow{d} \chi_1^2, \quad n \to \infty. \tag{2.2}$$

Důkaz. Z věty 5 víme

$$\frac{D_n - \delta}{\sqrt{\sigma_n^2}} \xrightarrow{d} \mathrm{N}(0,1), \quad n \to \infty.$$

Za platnosti H_0 je navíc $\delta = 0$ a rozptyl D_n je roven $\sigma_n^2 = \frac{1}{n}\left(p_{2,1} + p_{1,2}\right)$, označme $\widehat{\sigma}_n^2 = \frac{1}{n}\left(\widehat{p}_{2,1} + \widehat{p}_{1,2}\right)$, pak

$$\frac{D_n - \delta}{\sigma_n} = \frac{D_n - \delta}{\widehat{\sigma}_n^2} \cdot \frac{\widehat{\sigma}_n^2}{\sigma_n}.$$

Díky konzistenci jednotlivých odhadů $\widehat{p}_{i,j}$ platí $\frac{\widehat{\sigma}_n^2}{\sigma_n^2} \xrightarrow{P} 1$ a z Cramér-Sluckého věty (věta A.4) za H_0 platí

$$\frac{D_n - \delta}{\widehat{\sigma}_n^2} \xrightarrow{d} \mathrm{N}(0,1), \quad n \to \infty.$$

Po úpravě

$$\frac{D_n - \delta}{\widehat{\sigma}_n^2} = \frac{\frac{1}{n}(N_{+,1} - N_{1,+})}{\sqrt{\frac{1}{n}\left(\widehat{p}_{2,1} + \widehat{p}_{1,2}\right)}} = \frac{\frac{N_{2,1} - N_{1,2}}{n}}{\sqrt{\frac{N_{2,1} + N_{1,2}}{n^2}}} = \frac{N_{2,1} - N_{1,2}}{\sqrt{N_{2,1} + N_{1,2}}} \xrightarrow{d} \mathrm{N}(0,1), \quad n \to \infty.$$

Známe tedy i asymptotické rozdělení druhé mocniny

$$M_n^2 = \frac{(N_{2,1} - N_{1,2})^2}{N_{2,1} + N_{1,2}} \xrightarrow{d} \chi_1^2, \quad n \to \infty.$$

□

Test, který testuje $H_0 : p_{+,1} = p_{1,+}$ s testovou statistikou M_n^2 se nazývá McNemarův test. Test na hladině α hypotézu zamítá, je-li $M_n^2 \geq \chi_1^2(1-\alpha)$.

2.2 Přesné rozdělení

V předešlé části jsme odvodili McNemarův test založený na asymptotickém rozdělení napozorovaných četností. Pro malé rozsahy umíme určit i přesné rozdělení a na jeho základě sestavit přesný test hypotézy $H_0 : p_{+,1} = p_{1,+}$. Lze z nich také odvodit testovou statistiku McNemarova testu uvedenou v (2.2).

Jak již bylo řečeno, napozorované četnosti $(N_{1,1}, N_{1,2}, N_{2,1}, N_{2,2})^\top$ mají multinomické rozdělení $\mathrm{Mult}_4(n, \boldsymbol{p})$, kde $\boldsymbol{p} = (p_{1,1}, p_{1,2}, p_{2,1}, p_{2,2})^\top$. Uvažujme nyní podmíněné rozdělení $N_{1,2}, N_{2,1}$ při pevných $N_{1,1}, N_{2,2}$ a n. Potom z věty 1 máme

$$P(N_{1,2} = n_{1,2}, N_{2,1} = n_{2,1} | N_{1,1} = n_{1,1}, N_{2,2} = n_{2,2}) =$$
$$= \frac{(n - n_{1,1} - n_{2,2})!}{n_{1,2}!\, n_{2,1}!} \left(\frac{p_{1,2}}{1 - p_{1,1} - p_{2,2}}\right)^{n_{1,2}} \left(\frac{p_{2,1}}{1 - p_{1,1} - p_{2,2}}\right)^{n_{2,1}}.$$

Za platnosti $H_0 : p_{+,1} = p_{1,+}$ je $p_{1,2} = p_{2,1}$ a výraz lze upravit

$$P(N_{1,2} = n_{1,2}, N_{2,1} = n_{2,1} | N_{1,1} = n_{1,1}, N_{2,2} = n_{2,2}) =$$
$$= \frac{(n_{1,2} + n_{2,1})!}{n_{1,2}!\, n_{2,1}!} \left(\frac{p_{1,2}}{2p_{1,2}}\right)^{n_{1,2}} \left(\frac{p_{1,2}}{2p_{1,2}}\right)^{n_{2,1}} = \binom{n_{1,2} + n_{2,1}}{n_{2,1}} \left(\frac{1}{2}\right)^{n_{1,2}+n_{2,1}}.$$

Při pevném $N_{1,1}, N_{2,2}$ a n je i $N_{1,2} + N_{2,1} = n - N_{1,1} - N_{2,2}$ pevné, a proto označíme-li $N = N_{1,2} + N_{2,1}$, máme při pevném $N_{1,1}, N_{2,2}$ a n

$$N_{2,1} \sim Bi\left(N, \frac{1}{2}.\right).$$

Na tomto výsledku je založen přesný test naší hypotézy $p_{+,1} = p_{1,+}$. Tu zamítáme, je-li $N_{2,1} \leq k_1$ nebo $N_{2,1} \geq k_2$, kde

$$k_1 \text{ je největší celé číslo splňující } \sum_{k=0}^{k_1} \binom{N}{k} \left(\frac{1}{2}\right)^N \leq \frac{\alpha}{2},$$
$$k_2 \text{ je nejmenší celé číslo splňující } \sum_{k=k_2}^{N} \binom{N}{k} \left(\frac{1}{2}\right)^N \leq \frac{\alpha}{2}.$$

Tento test má zřejmě hladinu nejvýše α.
Je-li $N_{1,1}, N_{2,2}$ a n pevné, pak je pevné i $N = N_{1,2} + N_{2,1} = n - N_{1,1} - N_{2,2}$, proto z centrální limitní věty (věta A.1) víme, že pro $N \to \infty$ je

$$\frac{N_{2,1} - \mathsf{E}\ N_{2,1}}{\sqrt{\text{var}\, N_{2,1}}} = \frac{N_{2,1} - \frac{1}{2}N}{\sqrt{\frac{1}{4}N}} = \frac{2N_{2,1} - (N_{1,2} + N_{2,1})}{\sqrt{N_{1,2} + N_{2,1}}} = \frac{N_{2,1} - N_{1,2}}{\sqrt{N_{1,2} + N_{2,1}}} \xrightarrow{d} \text{N(0,1)}.$$

Druhá mocnina tohoto výrazu je tedy rovna statistice M_n^2 z McNemarova testu. Hodnoty N, pro které je asympotické rozdělení dostatečně přesné se u různých autorů liší, např. v knize Agresti (2002) se uvádí $N > 10$.

3. Obecná párová kategoriální data

V této kapitole se budeme zabývat případem, kdy je odpovídající tabulka čtvercová typu $I \times I$. Pro tyto tabulky už neplatí ekvivalence obou sledovaných vlastností jako v případě $I = 2$. Proto odvodíme testy na tyto vlastnosti zvlášť.

Nechť v celé kapitole je $(X, Y)^\top$ dvourozměrný diskrétně rozdělený náhodný vektor s hodnotami v množině $\{1, \ldots, I\}^2$ a $(X_1, Y_1)^\top, \ldots, (X_n, Y_n)^\top$ náhodný výběr z rozdělení vektoru $(X, Y)^\top$.
Z lemmatu 4 plyne, že symetrie tabulky pravděpodobností implikuje shodu marginálních rozdělení, ale opačná implikace obecně neplatí.

3.1 Shoda marginálních rozdělení

V této části odvodíme Stuartův test na testování shody marginálních rozdělení X a Y. Uvedeme také Bhapkarův test, který se někdy používá na testování téhož. Zmíněné testy byly poprvé navrženy v pracech Stuart (1955), resp. Bhapkar (1966).

Testujeme

$$H_0 : p_{+,i} = p_{i,+}, \quad \forall\, i = 1, \ldots, I, \qquad H_1 : \exists\, i \in \{1, \ldots, I\} : p_{+,i} \neq p_{i,+}. \tag{3.1}$$

V předchozí kapitole jsme používali testovou statistiku založenou na rozdílu napozorovaných četností $N_{+1} - N_{1+}$. Podobný princip uplatníme i zde.
Položme

$$D_{n,i} = N_{i,+} - N_{+,i},\ i = 1, \ldots, I \quad \text{a dále} \quad \boldsymbol{D}_n = (D_{n,1}, \ldots, D_{n,I})^\top. \tag{3.2}$$

Některé vlastnosti náhodného vektoru $\boldsymbol{D}_n$ popisuje následující tvrzení.

Tvrzení 7. *Pro náhodný vektor $\boldsymbol{D}_n$ definovaný v (3.2) a $i, j \in \{1, \ldots, I\}$ platí:*

i) $\mathsf{E}\, D_{n,i} = n(p_{i,+} - p_{+,i})$;

ii) $\operatorname{var} D_{n,i} = n\left[p_{i,+} + p_{+,i} - 2p_{i,i} - (p_{i,+} - p_{+,i})^2\right]$;

iii) $\operatorname{cov}(D_{n,i}, D_{n,j}) = -n\left[p_{i,j} + p_{j,i} + (p_{i,+} - p_{+,i})(p_{j,+} - p_{+,j})\right]$, *je-li* $i \neq j$.

Důkaz. S vlastnostmi multinomického rozdělení z věty 1 dostáváme
$\mathsf{E}\, D_{n,i} = \mathsf{E}\, N_{i,+} - \mathsf{E}\, N_{+,i} = n(p_{i,+} - p_{+,i})$.
Dále počítejme rozptyl

$$\operatorname{var} D_{n,i} = \operatorname{var} N_{i,+} + \operatorname{var} N_{+,i} - 2\operatorname{cov}(N_{i,+}, N_{+,i})$$

Pro kovarianci $N_{i,+}, N_{+,j}$ platí

$$\begin{aligned}\operatorname{cov}(N_{i,+}, N_{+,j}) &= \operatorname{cov}\left(\sum_{a=1}^{I} N_{i,a}, \sum_{b=1}^{I} N_{b,j}\right) = \operatorname{var} N_{i,j} + \sum_{(i,a)\neq(b,j)}\sum \operatorname{cov}\left(N_{i,a}, N_{b,j}\right) \\ &= np_{i,j}(1-p_{i,j}) + \sum_{(i,a)\neq(b,j)}\sum (-np_{i,a}p_{b,j}) \\ &= np_{i,j}(1-p_{i,j}) - n(p_{i,+}p_{+,j} - p_{i,j}^2) \\ &= n(p_{i,j} - p_{i,+}p_{+,j}).\end{aligned}$$

Celkově tedy pro rozptyl máme

$$\begin{aligned}\operatorname{var} D_{n,i} &= np_{i,+}(1-p_{i,+}) + np_{+,i}(1-p_{+,i}) - 2n(p_{i,i} - p_{i,+}p_{+,i}) \\ &= n\left[p_{i,+} + p_{+,i} - 2p_{i,i} - \left(p_{i,+}^2 - 2p_{i,+}p_{+,i} + p_{+,i}^2\right)\right] \\ &= n\left[p_{i,+} + p_{+,i} - 2p_{i,i} - \left(p_{i,+} - p_{+,i}\right)^2\right]\end{aligned}$$

Pro ověření vzorce pro kovarianci $D_{n,i}, D_{n,j}$, je-li $i \neq j$, počítejme s využitím vlstností kovariance a odvozeného vyjádření $\operatorname{cov}(N_{i,+}, N_{+,j})$,

$$\begin{aligned}\operatorname{cov}(D_{n,i}, D_{n,j}) &= \operatorname{cov}(N_{i,+} - N_{+,i}, N_{j,+} - N_{+,j}) \\ &= -np_{i,+}p_{j,+} - \operatorname{cov}(N_{i,+}, N_{+,j}) - \operatorname{cov}(N_{+,i}, N_{j,+}) - np_{+,i}p_{+,j} \\ &= -np_{i,+}p_{j,+} - n(p_{i,j} - p_{i,+}p_{+,j}) - n(p_{j,i} - p_{+,i}p_{j,+}) - np_{+,i}p_{+,j} \\ &= -n(p_{i,j} + p_{j,i} + p_{i,+}p_{j,+} - p_{i,+}p_{+,j} - p_{+,i}p_{j,+} + p_{+,i}p_{+,j}) \\ &= -n\left[p_{i,j} + p_{j,i} + (p_{i,+} - p_{+,i})(p_{j,+} - p_{+,j})\right].\end{aligned}$$

□

Z tvrzení 7 vidíme, že za platnosti H_0 je $\mathsf{E}\,\boldsymbol{D}_n = \mathbf{0}$ a rozptyl i kovariance se trochu zjednoduší. Toho využijeme při výběru testové statistiky. Označme

$$\begin{aligned}w_{i,j} &= -(p_{i,j} + p_{j,i}), \text{ je-li } i \neq j, \\ w_{i,i} &= p_{i,+} + p_{+,i} - 2p_{i,i}, \\ \widetilde{\boldsymbol{D}}_n &= (D_{n,1}, \ldots, D_{n,I-1})^\top, \\ \widetilde{\boldsymbol{W}} &= (w_{i,j})_{i,j=1}^{I-1}, \\ \widetilde{\boldsymbol{W}}_n &= n \cdot \widetilde{\boldsymbol{W}}.\end{aligned}$$

Pro takto definované objekty dokážeme tvrzení o jejich asymptotickém rozdělení, nejprve ale ukážeme pomocné lemma.

Lemma 8. *Za H_0 platí* $\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{D}}_n \xrightarrow{d} N_{I-1}(\mathbf{0}, \widetilde{\boldsymbol{W}}), \quad n \to \infty.$

Důkaz. Zvolme i takové, že $1 \leq i \leq I-1$, ukážeme, že $\frac{1}{\sqrt{n}} D_{n,i}$ má za H_0 asymptoticky normální rozdělení. Pro $k = 1, \ldots, n$ definujme náhodnu veličiny U_k předpisem

$$U_k = \begin{cases} 1, & \text{je-li } X_k = i, Y_k \neq i, \\ 0, & \text{je-li } X_k = i, Y_k = i \text{ nebo } X_k \neq i, Y_k \neq i, \\ -1, & \text{je-li } X_k \neq i, Y_k = i. \end{cases}$$

Veličiny U_k, $k = 1, \ldots, n$ jsou nezávislé a stejně rozdělené z nezávislosti jednotlivých (X_k, Y_k) a platí $D_{n,i} = N_{i,+} - N_{+,i} = \sum_{k=1}^{n} U_k = n\overline{U}_n$.
Z centrální limitní věty (věta A.1) máme

$$\sqrt{n}\frac{\overline{U}_n - \mathsf{E}\, U_1}{\sqrt{\operatorname{var} U_1}} \xrightarrow{d} \mathrm{N}(0,1), \quad n \to \infty.$$

Úpravou dostáváme pro

$$\sqrt{n}\frac{\overline{U}_n - \mathsf{E}\, U_1}{\sqrt{\operatorname{var} U_1}} = \frac{\sqrt{n}}{n} \cdot \frac{\sum_{k=1}^{n} U_k - n\,\mathsf{E}\, U_1}{\sqrt{\operatorname{var} U_1}} = \frac{D_{n,i} - n\,\mathsf{E}\, U_1}{\sqrt{n \operatorname{var} U_1}}.$$

Dále $\mathsf{E}\, U_1 = 1 \cdot (p_{i,+} - p_{i,i}) - 1 \cdot (p_{+,i} - p_{i,i}) = p_{i,+} - p_{+,i}$, za platnosti H_0 je tedy $\mathsf{E}\, U_1 = 0$ a dále $\operatorname{var} D_{n,i} = n \operatorname{var} U_1$ neboli za H_0 z věty 7 dostáváme $n \operatorname{var} U_1 = n(p_{i,+} + p_{+,i} - 2p_{i,i})$.
Celkově tedy za H_0

$$\frac{D_{n,i}}{\sqrt{n}} \xrightarrow{d} \mathrm{N}(0, p_{i,+} + p_{+,i} - 2p_{i,i}), \quad n \to \infty.$$

Každá lineární kombinace jednotlivých $D_{n,i}/\sqrt{n}$ má asymptoticky normální rozdělení s vhodnými parametry, proto $\widetilde{\boldsymbol{D}}_n/\sqrt{n}$ má za H_0 asymptoticky normální rozdělení s nulovým vektorem středních hodnot a varianční maticí $\widetilde{\boldsymbol{W}}$, neboli

$$\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{D}}_n \xrightarrow{d} \mathrm{N}_{I-1}(\boldsymbol{0}, \widetilde{\boldsymbol{W}}), \quad n \to \infty.$$

□

Tvrzení 9. *Za platnosti H_0 platí*

$$Q = \widetilde{\boldsymbol{D}}_n^{\top} \widetilde{\boldsymbol{W}}_n^{-1} \widetilde{\boldsymbol{D}}_n \xrightarrow{d} \chi^2_{I-1}, \quad n \to \infty. \tag{3.3}$$

Důkaz. Z předešlého lemmatu víme, že

$$\frac{\widetilde{\boldsymbol{D}}_n}{\sqrt{n}} \xrightarrow{d} \mathrm{N}_{I-1}(\boldsymbol{0}, \widetilde{\boldsymbol{W}}), \quad n \to \infty,$$

Dále $\widetilde{\boldsymbol{W}}$ je regulární: kdyby nebyla, tak by z věty A.6 musela existovat nějaká netriviální lineární kombinace složek vektoru $n^{-1/2}\widetilde{\boldsymbol{D}}_n$, která by byla skoro jistě rovna nějaké konstantě. Složky vektoru $n^{-1/2}\widetilde{\boldsymbol{D}}_n$ mezi sebou ale nijak nezávisí, taková lineární kombinace tedy nemůže existovat (pro celý vektor $\boldsymbol{D}_n$ bychom

toto tvrdit nemohli, protože $D_{n,i}$ bychom mohli vyjádřit pomocí předchozích složek a n). Tedy $\widetilde{\boldsymbol{W}}$ je regulární a z věty $A.7$

$$(n^{-1/2}\widetilde{\boldsymbol{D}}_n)^\top \widetilde{\boldsymbol{W}}^{-1}(n^{-1/2}\widetilde{\boldsymbol{D}}_n) \xrightarrow{d} \chi^2_{I-1}, \quad n \to \infty.$$

Úpravou dostáváme pro $n \to \infty$

$$(n^{-1/2}\widetilde{\boldsymbol{D}}_n)^\top \widetilde{\boldsymbol{W}}^{-1}(n^{-1/2}\widetilde{\boldsymbol{D}}_n) = \widetilde{\boldsymbol{D}}_n^\top (n\widetilde{\boldsymbol{W}})^{-1}\widetilde{\boldsymbol{D}}_n = \widetilde{\boldsymbol{D}}_n^\top \widetilde{\boldsymbol{W}_n}^{-1}\widetilde{\boldsymbol{D}}_n = Q \xrightarrow{d} \chi^2_{I-1}.$$

□

Nyní už známe asymptotické rozdělení statistiky Q, v matici $\widetilde{\boldsymbol{W}_n}$ jsou ale neznámé pravděpodobnosti. Dále budeme pracovat s jejich odhady N_{ij}/n a ukážeme, že asymptotické rozdělení se tímto nezmění.

Věta 10 (Stuartova)**.** *Pro* $1 \le i, j \le I-1$ *označme*

$$V_{i,i} = N_{i,+} + N_{+,i} - 2N_{i,i},$$
$$V_{i,j} = -(N_{i,j} - N_{j,i}),\ \textit{je-li}\ i \neq j,$$
$$\boldsymbol{V} = (V_{i,j})_{i,j=1}^{I-1}.$$

Pak za platnosti H_0 *platí* $Q_S = \widetilde{\boldsymbol{D}}_n^\top \boldsymbol{V}^{-1}\widetilde{\boldsymbol{D}}_n \xrightarrow{d} \chi^2_{I-1}, \quad n \to \infty.$

Důkaz. Po úpravě můžeme psát

$$Q_S = \widetilde{\boldsymbol{D}}_n^\top \boldsymbol{V}^{-1}\widetilde{\boldsymbol{D}}_n = (n^{-1/2}\widetilde{\boldsymbol{D}}_n)^\top (n^{-1}\boldsymbol{V})^{-1}(n^{-1/2}\widetilde{\boldsymbol{D}}_n).$$

Každý prvek matice $n^{-1}\boldsymbol{V}$ konverguje v pravděpodobnosti k prvku příslušnému prvku matice $\widetilde{\boldsymbol{W}}$, neboť odhad $N_{i,j}/n$ konzistentním odhadem p_{ij}, jak jsme ukázali v předchozí kapitole. V předchozí větě jsme ukázali, že matice $\widetilde{\boldsymbol{W}}$ je regulární, proto také každý prvek matice $(n^{-1}\boldsymbol{V})^{-1}$ konverguje v pravděpodobnosti k příslušnému prvku matice $\widetilde{\boldsymbol{W}}^{-1}$.
Z Cramér-Sluckeho věty (věta A.4) víme, že náhodné vektory $\widetilde{\boldsymbol{W}}^{-1}(n^{-1/2}\widetilde{\boldsymbol{D}}_n)$ a $(n^{-1}\boldsymbol{V})^{-1}(n^{-1/2}\widetilde{\boldsymbol{D}}_n)$ mají stejné asymptotické rozdělení.
Z věty o spojité transformaci (věta A.3) proto také náhodná veličina

$$(n^{-1/2}\widetilde{\boldsymbol{D}}_n)^\top (n^{-1}\boldsymbol{V})^{-1}(n^{-1/2}\widetilde{\boldsymbol{D}}_n) = Q_S$$

má asymptoticky stejné rozdělení jako náhodná veličina

$$(n^{-1/2}\widetilde{\boldsymbol{D}}_n)^\top \widetilde{\boldsymbol{W}}^{-1}(n^{-1/2}\widetilde{\boldsymbol{D}}_n).$$

Dostáváme tedy

$$Q_S \xrightarrow{d} \chi^2_{I-1}, \quad n \to \infty.$$

□

Testová statistika Q_S se používá ve Stuartově testu, který testuje hypotézu (3.1).

Tento test má zřejmě asymptotickou hladinu α, pokud hypotézu zamítneme, jestliže $Q_S \geq \chi^2_{I-1}(1-\alpha)$.

Stuartův test je zobecněním McNemarova testu z předchozí kapitoly, pro $I = 2$ dostáváme

$$Q_s = (N_{1,+} - N_{+,1})(N_{1,+} + N_{+,1} - 2N_{1,1})^{-1}(N_{1,+} - N_{+,1}) = \frac{(N_{1,2} - N_{2,1})^2}{N_{1,2} + N_{2,1}} = M_n^2.$$

Na testování (3.1) existuje ještě jiný přístup, založený na odhadování celého rozptylu a kovariance složek $D_{n,i}$ (oproti odhadnutí jejich zjednodušených verzí ve Stuartově testu).

Věta 11 (Bhapkarova)**.** *Pro $1 \leq i, j \leq I-1$ označme*

$$\begin{aligned} K_{i,i} &= N_{i,+} + N_{+,i} - 2N_{i,i} - \frac{(N_{i,+} - N_{+,i})^2}{n}, \\ K_{i,j} &= -(N_{i,j} - N_{j,i}) - \frac{(N_{i,+} - N_{+,i})(N_{j,+} - N_{+,j})}{n}, \text{ je-li } i \neq j, \\ \boldsymbol{K} &= (K_{i,j})_{i,j=1}^{I-1}. \end{aligned}$$

Pak za platnosti H_0 platí $Q_B = \widetilde{\boldsymbol{D}}_n^\top \boldsymbol{K}^{-1} \widetilde{\boldsymbol{D}}_n \xrightarrow{d} \chi^2_{I-1}, \quad n \to \infty.$

Důkaz lze nalézt v článku Bhapkar (1966).

3.2 Testování symetrie

Nyní budeme testovat druhou ze sledovaných vlastností, symetrii tabulky pravděpodobností, tj. hypotézu

$$H_0 : p_{i,j} = p_{j,i} \quad \forall\, i,j \in \{1, \ldots, I\}, \quad H_1 : \exists\, i,j \in \{1, \ldots, I\} : p_{i,j} \neq p_{j,i}.$$

Z lemmatu 4 víme, že za platnosti H_0 nastává i shoda marginálních rozdělení, ale pro $I \neq 2$ opačná implikace obecně neplatí. Při odvozování testu vyjdeme z testové statistiky

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{(N_{i,j} - np_{i,j})^2}{np_{i,j}} = \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{N_{i,j}^2}{np_{i,j}} - n.$$

V tomto případě použijeme variantu χ^2 testu s neznámými parametry. Těmi jsou pro nás všechny pravděpodobnosti p_{ij}. Za H_0 pro ně však máme vyjádření ve tvaru

$$\begin{aligned} p_{i,j} &= p_{j,i}, \quad i > j \\ p_{I,I} &= 1 - 2\sum_{1 \leq i < j \leq I} p_{i,j} - \sum_{i<I} p_{i,i}, \end{aligned} \tag{3.4}$$

neboť prvky v matici pravděpodobností jsou díky H_0 symetrické podle hlavní diagonály a prvek na místě (I, I) můžeme určit jako doplňek ostatních do jedničky.

Proto díky vztahům z (3.4) stačí pracovat s vektorem pravděpodobností $(p_{1,1}, p_{1,2}, \ldots, p_{1,I}, p_{2,2}, \ldots, p_{2,I}, \ldots, p_{I-1,I-1}, p_{I-1,I})^\top = \boldsymbol{a}$, ačkoliv neznámé jsou všechny $p_{i,j}$. Ve značení z části 1.2 je tedy náš vektor neznámých parametrů roven $\boldsymbol{a}$. Ostatní parametry můžeme vyjádřit jako funkci složek $\boldsymbol{a}$. Označme m počet neznámých parametrů za H_0, v našem případě $m = \frac{I(I+1)}{2} - 1$.

Pro určení odhadu $\boldsymbol{a}$ metodou modifikovaného minimálního χ^2 řešíme soutavu rovnic

$$\sum_{i=1}^{I}\sum_{j=1}^{I} \frac{N_{i,j}}{p_{i,j}(\boldsymbol{a})} \frac{\partial p_{i,j}(\boldsymbol{a})}{\partial a_k} = 0, \qquad k = 1, \ldots, m. \tag{3.5}$$

Na určení parciálních derivací použijeme vyjádření v (3.4) a dostáváme

$$\frac{\partial p_{i,j}(\boldsymbol{a})}{\partial a_k} = \frac{\partial p_{i,j}(\boldsymbol{a})}{\partial p_{i,i}} = -1, \text{ odpovídá-li } k\text{tá souřadnice vektoru } \boldsymbol{a} \text{ prvku } p_{i,i},$$

$$\frac{\partial p_{i,j}(\boldsymbol{a})}{\partial a_k} = \frac{\partial p_{i,j}(\boldsymbol{a})}{\partial p_{i,j}} = -2, \text{ odpovídá-li } k\text{tá souřadnice vektoru } \boldsymbol{a} \text{ prvku } p_{i,j},$$

pro $1 \leq i < I$ resp. $1 \leq i < j \leq I$.
Nyní můžeme pomocí těchto vyjádření upravit soustavu (3.5), dostáváme

$$\frac{N_{i,i}}{p_{i,i}} \cdot 1 + \frac{N_{I,I}}{p_{I,I}} \cdot (-1) = 0, \quad i \leq I, \tag{3.6}$$

$$\frac{N_{i,j} + N_{j,i}}{p_{i,j}} \cdot 1 + \frac{N_{I,I}}{p_{I,I}} \cdot (-2) = 0, \quad i < j, \tag{3.7}$$

kde v (3.6) jsme přidali rovnici pro $i = I$ (ačkoli neplyne z vyjádření derivací výše, platí triviálně). Odtud plyne vyjádření

$$p_{ii} = N_{i,i}\frac{p_{I,I}}{N_{I,I}}, \qquad p_{i,j} = \frac{N_{i,j} + N_{j,i}}{2} \cdot \frac{p_{I,I}}{N_{I,I}}. \tag{3.8}$$

Víme, že součet všech N_{ij} je roven celkovému počtu pozorování n, z rovnic (3.6), (3.7) máme vyjádření

$$N_{i,i} = \frac{N_{I,I}}{p_{I,I}} p_{i,i}, \qquad N_{i,j} + N_{j,i} = \frac{N_{I,I}}{p_{I,I}} 2p_{i,j},$$

sečtením dostáváme

$$n = \sum_{i=1}^{I} N_{i,i} + \sum_{1 \leq i < j \leq I} (N_{i,j} + N_{j,i}) = \frac{N_{I,I}}{p_{I,I}} \left(\sum_{i=1}^{I} p_{i,i} + \sum_{1 \leq i < j \leq I} 2p_{i,j} \right) = \frac{N_{I,I}}{p_{I,I}},$$

neboť součet všech pravděpodobností $p_{i,j}$ dává 1 a za H_0 je $p_{i,j} = p_{j,i}$.
Dosadíme do (3.8) a dostáváme odhady pravděpodobností $p_{i,j}$ za H_0 ve tvaru

$$\widehat{p}_{i,i} = \frac{N_{i,i}}{n}, \qquad \widehat{p}_{i,j} = \frac{N_{i,j} + N_{j,i}}{2n}. \tag{3.9}$$

Tyto odhady dosadíme do statistiky χ^2 a obdržíme

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{(N_{i,j} - n\widehat{p}_{i,j})^2}{n\widehat{p}_{i,j}} = \sum_{i=1}^{I} \frac{(N_{i,i} - n\frac{N_{i,i}}{n})^2}{n\frac{N_{i,i}}{n}} + \sum\sum_{i\neq j} \frac{(N_{i,j} - n\frac{N_{i,j}+N_{j,i}}{2n})^2}{n\frac{N_{i,j}+N_{j,i}}{2n}} \\
&= \frac{1}{2}\sum\sum_{i\neq j} \frac{(N_{i,j} - N_{j,i})^2}{N_{i,j} + N_{j,i}} \\
&= \frac{1}{2}\left(\sum\sum_{i>j} \frac{(N_{i,j} - N_{j,i})^2}{N_{i,j} + N_{j,i}} + \sum\sum_{i<j} \frac{(N_{i,j} - N_{j,i})^2}{N_{i,j} + N_{j,i}}\right) \\
&= \sum\sum_{i>j} \frac{(N_{i,j} - N_{j,i})^2}{N_{i,j} + N_{j,i}},
\end{aligned}$$

neboť $(N_{i,j} - N_{j,i})^2 = (N_{j,i} - N_{i,j})^2$ a obě dvojité sumy se tedy rovnají. Označíme-li

$$B = \sum\sum_{i>j} \frac{(N_{i,j} - N_{j,i})^2}{N_{i,j} + N_{j,i}},$$

potom z věty 3 plyne, že za platnosti H_0 platí

$$B \xrightarrow{D} \chi^2_{\frac{I(I-1)}{2}}, \quad n \to \infty.$$

neboť $I^2 - (\frac{I(I+1)}{2} - 1) - 1 = \frac{I(I-1)}{2}$. Hypotézu budeme tedy zamítat, naměříme-li $B \geq \chi^2_{\frac{I(I-1)}{2}}(1-\alpha)$.

Tento test odvodil Bowker v článku Bowker (1948). Všimněme si, že jde opět o zobecnění McNemarova testu: pro $I = 2$ dostáváme $B = \frac{(N_{2,1}-N_{1,2})^2}{N_{2,1}+N_{1,2}}$, což je právě testová statistika M_n^2 v McNemarově testu.

4. Simulace

Vraťme se nyní k McNemarovu testu. V této části ukážeme, jak testová statistika dodržuje stanovenou hladinu pro různé rozsahy výběrů n a různé hodnoty p_{1+}. Dále pro tyto marginální pravděpodobnosti na vygenerovaných datech znázorníme konvergenci rozdělení testové statistiky k rozdělení χ^2_1.

Nechť $(X, Y)^\top$ je dvourozměrný diskrétně rozdělený náhodný vektor s hodnotami v množině $\{1, 2\}^2$ a $(X_1, Y_1)^\top, \ldots, (X_n, Y_n)^\top$ náhodný výběr z rozdělení vektoru $(X, Y)^\top$. McNemarův test, který jsme odvodili v části 2.1 testuje hypotézu

$$H_0 : p_{1,+} = p_{+,1}, \quad \text{proti} \quad H_1 : p_{1,+} \neq p_{+,1}.$$

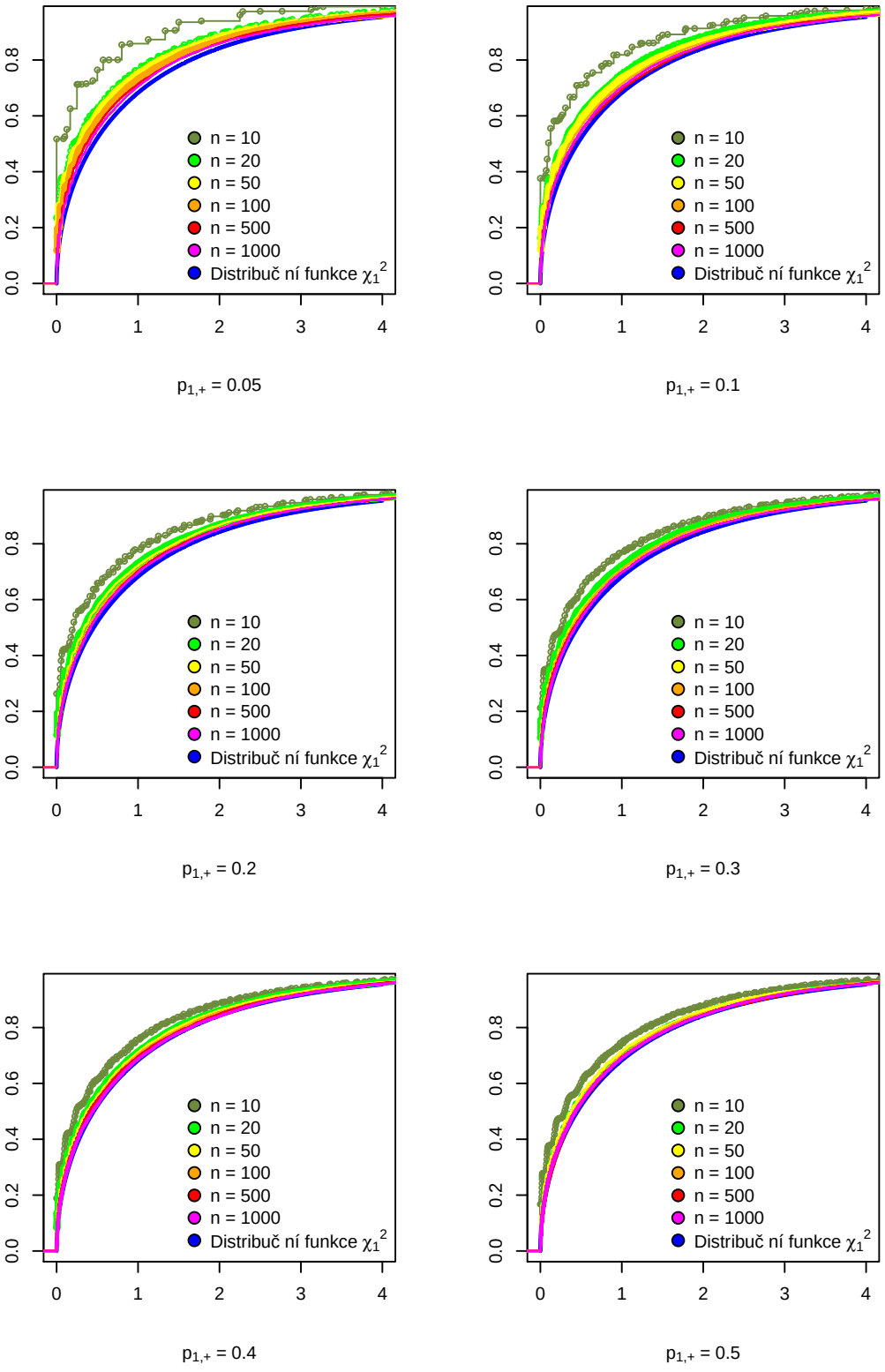
Nechť H_0 platí, uvažujme marginální pravděpodobnosti $p_{1,+}$ postupně 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 a rozsahy výběrů $n = 10, 20, 50, 100, 500, 1000$.
Při dané pravděpodobnosti $p_{1,+}$ určíme zbylé pravděpodobnosti následovně:
$p_{1,2} = p_{2,1} = p_{1,1} = p_{1,+}/2$ a $p_{2,2}$ se určí jako doplněk do jedničky.

Zvolme hladinu $\alpha = 0.05$. Pro každou kombinaci n a $p_{1,+}$ postupně vygenerujeme 100 000 náhodných výběrů z multinomického rozdělení o rozsahu n a vektorem pravděpodobností $(p_{1,1}, p_{1,2}, p_{2,1}, p_{2,2})^\top$, u každého spočteme a uložíme testovou statistiku M_n^2.
Následující tabulka ukazuje dosaženou hladinu, spočtenou jako podíl případů, kdy jsme platnou hypotézu zamítli, a celkového počtu provedených simulací. Vidíme, že pro malé pravděpodobnosti $p_{1,+}$ se zvolené hladiny α pro menší n nedosahuje, při větších hodnotách $p_{1,+}$ a rostoucím n se spočtená hladina postupně blíží ke stanovené hodnotě $\alpha = 0.05$.

Nyní se pro konkrétní pravděpodobnost p_{1+} zaměříme na konvergenci rozdělení testové statistiky k asymptotickému rozdělení χ^2_1. Na obrázku jsou pro jednotlivé pravděpodobnosti p_{1+} znázorněné distribuční funkce rozdělení χ^2_1 a empirické distribuční funkce testové statistiky spočtené ze 100 000 simulací pro n postupně $10, 20, 50, 100, 500$ a 1000. Při malém $p_{1,+}$ je pro malá n spočtená empirická distribuční funkce dost vzdálená od distribuční funkce limitního rozdělení, s rostoucím n se k ní blíží. Pro větší $p_{1,+}$ je už i při menším n aproximace přesnější, i konvergence k limitnímu rozdělení je rychlejší.

Tabulka 4.1: Dosažená hladina McNemarova testu

	$\mathbf{p_{1,+}}$					
n	0.05	0.1	0.2	0.3	0.4	0.5
10	0.0001	0.0013	0.0127	0.0332	0.0485	0.0522
20	0.0018	0.0130	0.0438	0.0479	0.0433	0.0420
50	0.0223	0.0471	0.0433	0.0670	0.0532	0.0537
100	0.0457	0.0437	0.0513	0.0524	0.0492	0.0491
500	0.0523	0.0497	0.0487	0.0501	0.0505	0.0504
1000	0.0484	0.0504	0.0497	0.0498	0.0506	0.0499



Obrázek 4.1: Konvergence k rozdělení χ_1^2

Závěr

V práci jsme se zabývali testy pro párová kategoriální data. Nejprve jsme zavedli potřebné značení, pojmy a formulovali jsme zkoumané vlastnosti. V dalších částech jsme v tomto značení shrnuli testy, které se na testování těchto vlastností používají.
Napřed jsme se zabývali dichotomickými párovými kategoriálními daty, ukázali jsme, že obě zkoumané vlastnosti jsou pro tato data ekvivalentní. V této části jsme odvodili McNemarův test, vycházeli jsme přitom z publikace Agresti (2002). Samostatně jsme odvodili detaily v důkazu tvrzení 5 a větě 6. Dále jsme v této kapitole odvodili přesný test pro malé rozsahy výběrů, zde jsme vycházeli především z publikací Anděl (1978) a Agresti (2007).
Dále jsme pracovali s obecnými párovými kategoriálními daty. Zde jsme vycházeli nejvíce z prací Stuart (1955), Bowker (1948) a dále z knihy Anděl (2007). Postupně jsme se zabývali oběma sledovanými vlastnostmi, neboť zde obecně nejsou ekvivalentní, jak ukazuje lemma 4. Na testování shody marginálních rozdělení jsme samostatně ukázali lemma 8, s jeho pomocí jsme poté odvodili Stuartův test, uvedli jsme také Bhapkarův test. Dále jsme v části 3.2 ukázali odvození testu na testování symetrie tabulky pravděpodobností, který poprvé odvodil Bowker.
V další části jsme samostatně provedli simulace McNemarova testu. Ověřovali jsme, jaké dosahuje hladiny při různých pravděpodobnostech a rozsazích výběrů. Zjistili jsme, že i při malé marginální pravděpodobnosti už pro náhodný výběr o rozsahu 100 dodržuje stanovenou hladinu. Také jsme znázornili konvergenci rozdělení testové statistiky tohoto testu k limitnímu rozdělení χ^2. Zjistili jsme, že pro malé marginální pravděpodobnosti se pro malé rozsahy výběrů může podstatně odlišovat, postupně se ale se zvětšujícím se rozsahem výběru blíží k asymptotickému rozdělení a při větší marginální pravděpodobnosti je už i při malém rozsahu aproximace přesnější.

Používané věty

V této části uvedeme znění vět, které v práci používáme, s příslušnými odkazy na důkazy.

Věta A.1 (Lindebergova CLV)**.** *Nechť $X_1, X_2, \ldots$ jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotou μ a s konečným rozptylem σ^2. Pak pro $n \to \infty$ platí*

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}} \xrightarrow{d} N(0, \sigma^2).$$

Důkaz. (viz Anděl, 1978, Věta X.6).

□

Věta A.2 (Lindebergova mnohorozměrná CLV)**.** *Nechť $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ jsou nezávislé stejně rozdělené náhodné vektory s vektorem středních hodnot $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{V}$ s konečnými prvky. Pak pro $n \to \infty$ platí*

$$\frac{1}{\sqrt{n}}(\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n - k\boldsymbol{\mu}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{V}).$$

Důkaz. (viz Anděl, 1978, Věta X.7).

□

Věta A.3 (Věta o spojité tranformaci)**.** *Je-li g spojitá reálná funkce a $X_n \xrightarrow{d} X$, pak $g(X_n) \xrightarrow{d} g(X)$.*

Důkaz. (viz Anděl, 2007, Věta B.8).

□

Věta A.4 (Cramér-Sluckého)**.** *Nechť $X_1, X_2, \ldots$ je posloupnost náhodných veličin s distribučními funkcemi $F_1, F_2, \ldots$. Nechť F je distribuční funkce a c konstanta. Nechť F_n konvergují slabě k F. Nechť $Y_1, Y_2, \ldots$ je taková posloupnost náhodných veličin, že $Y_n \xrightarrow{P} c$. Definujme*

$$S_n = X_n \cdot Y_n, \qquad T_n = X_n / Y_n.$$

Nechť F_n^S, F_n^T jsou po řadě distribuční funkce veličin S_n, T_n. Je-li $c > 0$, pak $F_n^S(x)$ konvergují slabě k $F(x/c)$ a $F_n^T(x)$ konvergují slabě k $F(cx)$.

Důkaz. (viz Anděl, 2007, Věta B.10).

□

Věta A.5. *Hodnost idempotentní matice se rovná její stopě.*

Důkaz. (viz Anděl, 2007, Věta A.12).

□

Věta A.6. *Nechť* $\boldsymbol{X} = (X_1, \ldots, X_r)^\top$ *je náhodný vektor s konečnými druhými momenty, vektorem středních hodnot* $\boldsymbol{\mu}$ *a varianční maticí* $\boldsymbol{V}$ *a* $\boldsymbol{c} \in \mathbb{R}^r$. *Pak vztah* $\boldsymbol{V}\boldsymbol{c} = \boldsymbol{0}$ *platí právě když je* $\boldsymbol{c}^\top(\boldsymbol{X} - \boldsymbol{\mu}) = 0$ *skoro jistě.*

Důkaz. (viz Anděl, 2007, Věta A.26).

□

Věta A.7. *Nechť* $\boldsymbol{X} \sim N_r(\boldsymbol{\mu},\boldsymbol{V})$ *a matice* $\boldsymbol{V}$ *je regulární. Pak náhodná veličina* $Y = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ *má rozdělení* χ^2_r.

Důkaz. Věta se speciálním případem věty 4.15 v knize (Anděl, 2007).

□

Věta A.8. *Nechť* $\boldsymbol{X} \sim N_r(\boldsymbol{0}, \boldsymbol{V})$ *a nechť* $\boldsymbol{A}$ *je symetrická pozitivně definitní matice typu* $r \times r$. *Je-li matice* $\boldsymbol{A}\boldsymbol{V}$ *nenulová a idempotentní, pak náhodná veličina* $\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}$ *má rozdělení* χ^2 *s počtem stupňů volnosti rovným* $tr(\boldsymbol{A}\boldsymbol{V})$.

Důkaz. (viz Anděl, 2007, Věta 4.16).

□

Literatura

Agresti, A. (2002). *Categorical Data Analysis*. Druhé vydání. Wiley Series in Probability and Statistics, Gainesville, Florida. ISBN 0-471-36093-7.

Agresti, A. (2007). *An introduction to Categorical Data Analysis*. Druhé vydání. Wiley Series in Probability and Statistics, Gainesville, Florida. ISBN 978-0-471-22618-5.

Anděl, J. (1978). *Matematická statistika*. První vydání. SNTL - Nakladatelství technické literatury, Praha. ISBN 80-85863-27-8.

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

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypothesis in categorical data. *J. Amer. Statist. Assoc.*, **61**, 228–235.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, **43**(244), 572–574.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412–416.