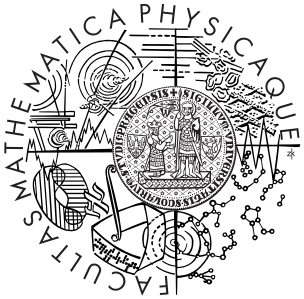
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Jaroslav Pazdera

Markovovy řetězce a kategoriální data

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Doc. RNDr. Zuzana Prášková, CSc.

Studijní program: matematika

2006

Na tomto místě bych chtěl poděkovat především vedoucí mojí bakalářské práce, Doc. RNDr. Zuzaně Práškové, CSc., za volbu zajímavého tématu, za cenné připomínky a rady, za ochotu k častým konzultacím a za pomoc s problémy a otázkami, které při psaní této práce vznikaly. Také děkuji za zapůjčení potřebné literatury.

Dále bych chtěl poděkovat mým rodičům a prarodičům, protože bez jejich podpory po celou dobu mého studia by tato práce nemohla vzniknout.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne Jaroslav Pazdera

Obsah

Název práce: Markovovy řetězce a kategoriální data
Autor: Jaroslav Pazdera
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Doc. RNDr. Zuzana Prášková, CSc.
e-mail vedoucího: praskova@karlin.mff.cuni.cz

Abstrakt: V předložené práci se zabýváme základními testy v kontingenčních tabulkách (testy nezávislosti, symetrie, homogenity) a paralelními testy pro pravděpodobnosti přechodu v Markovových řetězcích s diskrétním časem a konečnou množinou stavů (testy homogenity, symetrie, nezávislosti pravděpodobností přechodu na čase). Jsou odvozeny testové statistiky a testy jsou v závěrečné kapitole aplikovány na reálná data.

Klíčová slova: kontingenční tabulka, LR-test, multinomické rozdělení, Markovský řetězec, χ^2 test

Title: Markov Chains and Categorial Data
Author: Jaroslav Pazdera
Department: Department of Probability and Mathematical Statistics
Supervisor: Doc. RNDr. Zuzana Prášková, CSc.
Supervisor's e-mail address: praskova@karlin.mff.cuni.cz

Abstract: In the present thesis we study basic tests for two dimensional categorical data (tests of independence, homogeneity, symmetry) and their parallel tests for transition probabilities in Markov chains with discrete time and with finite set of states (test of homogeneity, symmetry, stacionarity). Test statistics are developed and tests are applied to some real data in the final chapter.

Keywords: categorial data, LR-test, multinomial distribution, Markov chain, χ^2 test.

Úvod

V následující práci definujeme multinomické rozdělení a jeho základní vlastnosti. Dále si popíšeme dvě metody odhadu parametrů a ukážeme jejich použití v testech. Následně upozorníme na souvislost mezi nimi. Zavedeme pojem kontingenční tabulky a vysvětlíme použití jednotlivých testů. V teorii náhodných procesů popíšeme markovovskou vlastnost a ukážeme použití jednotlivých testů na matice pravděpodobností přechodů. V závěrečné kapitole budeme ilustrovat použití testů na datech, která zveřejňuje pravidelně Český statistický úřad.

Kapitola 1

Multinomické rozdělení

1.1 Základní vlastnosti

V této kapitole se budeme zabývat základními vlastnostmi multinomického rozdělení a metodami odhadu parametrů.

Definice 1. *Mějme náhodný vektor $X = (X_1, \dots, X_k)$, který nabývá hodnot $(x_1, \dots, x_k)$ s pravděpodobnostmi*

$$P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \ , \tag{1.1}$$

$$kde \ \sum_{i=1}^{k} x_i = n \quad \sum_{i=1}^{k} p_i = 1 \quad x_i = 0, 1, \dots, n \ pro \ i = 1, \dots, k; \tag{1.2}$$

pak říkáme, že X má multinomické rozdělení s parametry $(n, p_1, \dots, p_k)$.

Věta 1. *Nechť $X = (X_1, \dots, X_k)$ má multinomické rozdělení s parametry $(n, p_1, \dots, p_k)$. Nechť $1 < r \leq k$. Pak marginální rozdělení veličin $X_r, \dots X_k$, je*

$$P(X_r = x_r, \dots, X_k = x_k) =$$
$$= \frac{n!}{x_r! \dots x_k! (n - x_r - \dots - x_k)!} p_r^{x_r} \dots p_k^{x_k} (1 - p_r - \dots - p_k)^{n - x_r - \dots - x_k}.$$

Dále platí

$$EX_i = np_i, \qquad varX_i = np_i(1 - p_i).$$

Důkaz. Viz [1, str. 268] □

Všimněme si, že každé jednorozměrné marginální rozdělení je binomické s parametry (n, p_i).

Věta 2. *Nechť $X = (X_1, ..X_k)$ je náhodný vektor s multinomickým rozdělením o parametrech $(n, p_1, .., p_k)$, pak*

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{1.3}$$

má asymptoticky pro $n \to \infty$ χ^2 rozdělení o $k-1$ stupních volnosti.

Důkaz. [1, str.270] □

1.2 Metoda maximální věrohodnosti

Označme $x = (x_1, \ldots, x_k)$ náhodný výběr z X a $p = (p_1, \ldots, p_k)$. Nechť vektor $X = (X_1, \ldots, X_k)$ má multinomické rozdělení s parametry $(n, p_1, \ldots, p_k)$. Nechť pro $P(X = x)$ platí (1.1) a (1.2). Cílem je $\max_p P(X = x)$.

Zlogaritmujeme výraz $P(X = x)$ a pokračujme standardním způsobem pro hledání extrémů. Extrémy budeme hledat na otevřené množině; předpokládejme proto, že $p_i > 0$ a $x_i > 0$ pro $i = 1, \ldots, k$.

$$\begin{aligned} L(p) := \ln P(X = x) &= \ln \frac{n!}{x_1! \ldots x_k!} + \ln p_1^{x_1} + \ldots + \ln p_k^{x_k} \\ &= c + \sum_{i=1}^{k} x_i \ln p_i \end{aligned}$$

Logaritmus je funkce konkávní a prostá, proto můžeme úlohu přeformulovat jako hledání

$$\max_{p=(p_1,\ldots,p_k)} \sum_{i=1}^{k} x_i \ln p_i$$

za podmínek

$$\sum_{i=1}^{k} p_i = 1.$$

Úlohu hledání lokálních extrémů vyřešíme pomocí Lagrangeových multiplikátorů. Lagrangeova funkce a její derivace v tomto případě jsou

$$L(p,\lambda) = \sum_{i=1}^{k} x_i \ln p_i + \lambda \left(\sum_{i=1}^{k} p_i - 1 \right), \tag{1.4}$$

$$\frac{\partial L(p,\lambda)}{\partial p_i} = \frac{x_i}{p_i} + \lambda = 0 \quad \text{pro } i = 1, \ldots, k, \tag{1.5}$$

$$\frac{\partial L(p,\lambda)}{\partial \lambda} = \sum_{i=1}^{k} p_i - 1 = 0. \tag{1.6}$$

Vyřešením (1.5) a (1.6) dostaneme

$$x_i = -\lambda p_i \quad \text{pro } i = 1, \ldots, k. \tag{1.7}$$

Sečtením přes všechna i získáme rovnost $n = -\lambda$. Dosazením zpět do (1.7) dostáváme odhad

$$\hat{p}_i = \frac{x_i}{n} \quad \text{pro } i = 1, \ldots, k.$$

Nyní ještě zbývá oddiskutovat mezní případy, jako jsou extrémy na hranici a body nespojitosti derivace. Dle předpokladu, že $x_i > 0$ pro všechna $i = 1, \ldots, k$, můžeme uvažovat takto. Pro posloupnost p_i^{β}, která pro libovolné i při $\beta \to \infty$ konverguje k nule, platí, že $P(X = x)$ vypočtená podle (1.1) také konverguje k nule. Proto existuje-li kladný extrém uvnitř množiny $\{p_i > 0, \text{ pro } i = 1, \ldots k; \sum_i p_i = 1\}$, je zároveň maximem. Protože $P(X = x)$ podle našich odhadnutých parametrů $\hat{p}_i$ bude kladná, je to maximum.

Výsledek: $\hat{p}_i = \frac{x_i}{n}$ je hodnota parametrů p_i maximalizující $P(X = x)$ pro dané x.

1.3 Metoda minimálního χ^2

Zůstaňme u multinomického rozdělení. Mějme pravděpodobnosti $p_1, ..., p_k$ a nechť tyto pravděpodobnosti závisí na m-rozměrném parametru $a = (a_1, .., a_m)$. Stále však musí platit $\sum_{i=1}^{k} p_i(a) = 1$. Za předpokladu spojitosti $p_i(a)$ pro všechna a zderivujme předcházející rovnost. Dostaneme

$$\sum_{i=1}^{k} \frac{\partial p_i(a)}{\partial a_j} = 0 \quad \text{pro všechny } j = 1, ..., m. \tag{1.8}$$

Upravme vztah (1.3) a uvažujme p_i závislá na parametru a:

$$\chi^2(a) = \sum_{i=1}^{k} \frac{(x_i - np_i(a))^2}{np_i(a)} = \sum_{i=1}^{k} \frac{x_i^2}{np_i(a)} - n. \tag{1.9}$$

Pak lze označit np_i jako teoretické četnosti a x_i jako četnosti skutečně pozorované.

Nyní použijeme podobnou myšlenku jako v metodě maximální věrohodnosti. Nebudeme chtít maximalizovat pravděpodobnost, ale minimalizovat $\chi^2(a)$. Tato metoda odhadu a se nazývá metoda minimálního χ^2. Odhad a získáme z (1.9) jako řešení rovnic

$$\frac{\partial \chi^2(a)}{\partial a_j} = -\frac{1}{n} \sum_{i=1}^{k} \frac{x_i^2}{p_i^2(a)} \frac{\partial p_i(a)}{\partial a_j} = 0, \quad \text{pro } j = 1, \ldots, m. \tag{1.10}$$

Tato soustava se však těžko řeší, proto se používá jiná metoda, která využívá derivací výrazu

$$\chi^2(a) = \sum_{i=1}^{k} \frac{(x_i - np_i(a))^2}{np_i(a)}.$$

V tomto případě dostaneme rovnice

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

Zanedbáme-li vliv druhého členu v součtu, získáme jednodušší rovnici

$$\sum_{i=1}^{k} \frac{x_i - np_i(a)}{p_i(a)} \frac{\partial p_i(a)}{\partial a_j} = 0.$$

Použijeme-li ještě vlastnost (1.8), získáme soustavu

$$\sum_{i=1}^{k} \frac{x_i}{p_i(a)} \frac{\partial p_i(a)}{\partial a_j} = 0, \qquad j = 1, \ldots, m. \tag{1.11}$$

Nechť $\hat{a}$ je řešení této zjednodušené soustavy, pak $\hat{a}$ minimalizuje $\chi^2(a)$, nazveme jej odhad parametru a modifikovanou metodou minimálního χ^2.

Věta 3. *Předpokládejme, že* $X = (X_1, \ldots, X_k)$ *je výběr z* $M(n, p_1^0, \ldots, p_k^0)$ *rozdělení. Nechť* $p_i^0 = p_i(a^0)$. *Nechť parametr* a^0 *je* m*-rozměrný, kde* $m < k - 1$. *Nechť pro všechny vektory* $a = (a_1, \ldots, a_m)$ *z nedegenerovaného omezeného uzavřeného intervalu* $A \subset \mathbb{R}_m$ *platí:*

(1) $p_1(a) + \cdot + p_k(a) = 1$.
(2) Existuje takové $c > 0$, *že* $p_i(a) > c^2$ *pro* $i = 1, \ldots, k$.
(3) Každá funkce $p_i(a)$ *má spojité parciální derivace* $\frac{\partial p_i(a)}{\partial a_j}$ *a* $\frac{\partial^2 p_i(a)}{\partial a_j \partial a_s}$ *pro j=1,...,m s=1,...,m.*
(4) Matice derivací $\left(\frac{\partial p_i(a)}{\partial a_j}\right)_{i,j}$, *která je typu* $k \times m$, *má hodnost* m.

Nechť a^0 *je vnitřním bodem* A, *pak existují takové posloupnosti kladných čísel* $\varepsilon_n \to 0$ *a* $\delta_n \to 0$, *že soustava* (1.11) *má s pravděpodobností alespoň* $1 - \varepsilon_n$ *právě jeden kořen* $\hat{a}_n$ *takový, že* $\|\hat{a}_n - a^0\| < \delta_n$. *Existuje-li* $\hat{a}_n$ *pro všechna dostatečně velká* n, *má veličina* $\chi^2(\hat{a}_n)$ *vypočtená podle* (1.9) *při* $n \to \infty$ *asymptoticky* χ^2_{k-m-1} *rozdělení.*

Důkaz. Odkaz na důkaz je možné najít v [1, str. 273-274]. □

Testujeme-li tedy hypotézu $H_0 : \ p_i(a) = p_i(a^0)$ pro všechna i proti alternativě $H_1 : \ p_i(a) \neq p_i(a^0)$ pro alespoň jedno i, tak hypotézu H_0 zamítáme, pokud $\chi^2(\hat{a})$ vypočtené podle (1.9) je větší než příslušná kritická hodnota χ^2_{k-m-1} rozdělení.

1.4 LR-test pro multinomické rozdělení

Nejdříve si uveďme 2 lemmata.

Lemma 1. *Čtvrtý centrální moment binomického rozdělení s parametry* (n, p) *je roven*

$$E[X - EX]^4 = 3(npq)^2 + npq(1 - 6pq), \quad kde \quad q = 1 - p. \tag{1.12}$$

Důkaz. Lze dokázat přímým výpočtem nebo lze nalézt například v [3, str. 51]. □

Lemma 2. *Nechť existuje* $E|X|^4 < \infty$, *pak platí*

$$\left(E|X|^3\right)^{\frac{1}{3}} \leq \left(E|X|^4\right)^{\frac{1}{4}}. \tag{1.13}$$

Důkaz. Využijeme poznatku z teorie prostorů L_p. Zde platí pro $x \in L_p$

$$\|x\|_1 \leq \|x\|_p \leq \|x\|_q \leq \|x\|_\infty \quad \text{pro } 1 < p < q < \infty.$$

Pak si stačí uvědomit, že pro $t \in (1, \infty)$ a $d > t$

$$\left(E|X|^t\right)^{\frac{1}{t}} = \left(\int_\Omega |X|^t dP\right)^{\frac{1}{t}} = \|X\|_d \leq \|X\|_t = \left(E|X|^d\right)^{\frac{1}{d}}.$$

□

Podle věty 2 víme, že

$$\sum_{i=1}^{k} np_i \Lambda_i^2 \ , \quad \text{kde } \Lambda_i = \frac{x_i - np_i}{np_i} \ , \tag{1.14}$$

má asymptoticky pro $n \to \infty$ χ^2 rozdělení o $k-1$ stupních volnosti, kde $x = (x_1, \ldots, x_k)$ je náhodný výběr z multinomického rozdělení. Testujme hypotézu, zda p_i mohou být předem dané hodnoty p_i^0 ,

$$H_0 : p_i = p_i^0$$

proti hypotéze $H_1 : p_i \neq p_i^0$. Předpokládejme, že máme náhodný výběr $x = (x_1, \ldots, x_k)$. Označme nyní $\theta_0 := (p_1^0, \ldots, p_k^0)$ jako testovaný parametr a $\hat{\theta}_n$ jako odhad parametru $(p_1, \ldots, p_k)$ pomocí metody maximální věrohodnosti. Pak tedy pravděpodobnost, že nastal náš náhodný výběr x, je za platnosti hypotézy H_0

$$P(\theta_0) = \frac{n!}{x_1! \ldots x_k!} (p_1^0)^{x_1} \ldots (p_k^0)^{x_k}.$$

Zlogaritmujme tuto funkci a logaritmus označme $L(\theta_0)$, dostaneme

$$L(\theta_0) := \ln \frac{n!}{x_1! \ldots x_k!} + \sum_{i=1}^{k} x_i \ln p_i^0. \tag{1.15}$$

Použijme stejnou úvahu na náš odhad parametru $\hat{\theta}_n$. Dostáváme

$$L(\hat{\theta}_n) = \ln \frac{n!}{x_1! \ldots x_k!} + \sum_{i=1}^{k} x_i \ln \frac{x_i}{n}, \tag{1.16}$$

neboť $\hat{p}_i = \frac{x_i}{n}$. Z (1.15) a (1.16) přímo plyne, že

$$LR := 2\left(L(\hat{\theta}_n) - L(\theta_0)\right) = 2 \sum_{i=1}^{k} x_i \ln \frac{x_i}{np_i^0}. \tag{1.17}$$

Poznámka. *Předchozí statistiku jsme nazvali LR podle "likehood ratio", neboli poměr věrohodností.*

Věta 4 (o LR statistice)**.** *Nechť* $(x_1, \ldots, x_k)$ *je výběr z multinomického rozdělení s parametry* $(n, p_1, \ldots, p_k)$*, pak při platnosti hypotézy*

$$H_0 : p_i = p_i^0 \quad \textit{pro } i = 1, \ldots, k$$

má (1.17) *asymptoticky* χ^2 *rozdělení o* $k-1$ *stupních volnosti.*

Důkaz. Větu dokážeme za zjednodušujícího předpokladu $\Lambda_i \in (-1, 1)$ pro všechna i, kde Λ_i je zavedeno v (1.14). Při značení z (1.14) je

$$x_i = np_i^0 \left(1 + \frac{x_i - np_i^0}{np_i^0}\right) = np_i^0(1 + \Lambda_i).$$

Dosaďme do (1.17) a použijme Taylorův rozvoj logaritmu:

$$\begin{aligned}
2\Big(L(\hat{\theta}_n) - L(\theta_0)\Big) &= 2\sum_{i=1}^{k} x_i \ln \frac{x_i}{np_i^0} \\
&= 2\sum_{i=1}^{k} np_i^0(1 + \Lambda_i)\ln(1 + \Lambda_i) \\
&= 2\sum_{i=1}^{k} np_i^0(1 + \Lambda_i)\left(\Lambda_i - \frac{1}{2}\Lambda_i^2 + \frac{1}{3}\zeta_i^3\right) \\
&= 2\sum_{i=1}^{k} np_i^0\left(\Lambda_i + \frac{1}{2}\Lambda_i^2 + \frac{1}{2}\Lambda_i^3 + \frac{1}{3}\zeta_i^3 + \frac{1}{3}\Lambda_i\zeta_i^3\right) \\
&= 2\sum_{i=1}^{k} np_i^0\Lambda_i + \sum_{i=1}^{k} np_i^0\Lambda_i^2 + \sum_{i=1}^{k} np_i^0\left(\Lambda_i^3 + \frac{2}{3}\zeta_i^3 + \frac{2}{3}\Lambda_i\zeta_i^3\right),
\end{aligned}$$

kde $\zeta_i \in (0; \Lambda_i)$. O $\sum_{i=1}^{k} np_i\Lambda_i^2$ víme, že má asymptoticky χ^2 rozdělení, dle věty 2. Budeme dále chtít dokázat, že zbylé dva sčítance konvergují k nule v pravděpodobnosti. První sčítanec je roven nule, protože z vlastností multinomického rozdělení plyne

$$2\sum_{i=1}^{k} np_i\Lambda_i = 2\sum_{i=1}^{k}(X_i - np_i) = 2\sum_{i=1}^{k} X_i - 2\sum_{i=1}^{k} np_i = 2n - 2n = 0.$$

Druhý sčítanec odhadněme následovně:

$$\begin{aligned}\left|\sum_{i=1}^{k} np_i(\Lambda_i^3+\frac{2}{3}\zeta_i^3+\frac{2}{3}\Lambda_i\zeta_i^3)\right| \leq & \sum_{i=1}^{k} np_i(3\left|\Lambda_i^3\right|+2\left|\zeta_i^3\right|+2\left|\Lambda_i\right|\left|\zeta_i^3\right|)\\ \leq & 5\sum_{i=1}^{k} np_i\left|\Lambda_i^3\right|+2\sum_{i=1}^{k} np_i\left|\Lambda_i^4\right| \qquad (1.18)\\ \leq & 7\sum_{i=1}^{k} np_i\left|\Lambda_i^3\right|.\end{aligned}$$

Aproximujme sumu v (1.18), tj.

$$\sum_{i=1}^{k} np_i\left|\Lambda_i^3\right|=\sum_{i=1}^{k}\frac{\left|X_i-np_i\right|^3}{(np_i)^2}.$$

Ze zobecněné Čebyševovy nerovnosti plyne:

$$\begin{aligned}P\left(\left|\sum_{i=1}^{k}\frac{\left|X_i-np_i\right|^3}{(np_i)^2}\right|>\varepsilon\right) \leq & \frac{1}{\varepsilon}E\left|\sum_{i=1}^{k}\frac{\left|X_i-np_i\right|^3}{(np_i)^2}\right|\\ = & \frac{1}{\varepsilon}\sum_{i=1}^{k}\frac{1}{n^2p_i^2}E\left|X_i-np_i\right|^3.\end{aligned}$$

Nyní využijeme toho, že náhodné veličiny X_i mají binomické rozdělení. Pokud použijeme lemma 2 a lemma 1, získáme

$$\frac{1}{\varepsilon}\sum_{i=1}^{k}\frac{E\left|X_i-np_i\right|^3}{n^2p_i^2}\leq\frac{1}{\varepsilon}\sum_{i=1}^{k}\frac{(3(np_iq_i)^2+np_iq_i(1-6p_iq_i))^{\frac{3}{4}}}{n^2p_i^2}.$$

Pravá strana poslední nerovnosti konverguje k 0 pro $n\to\infty$.

Podle Cramérovy-Sluckého věty, např. [1, věta B.10], jsme dokázali, že náhodná veličina $2\Big(L(\hat{\theta}_n)-L(\theta_0)\Big)$ má asymptoticky χ^2_{k-1} rozdělení. □

Hypotézu $H_0:\theta=\theta_0$ zamítáme ve prospěch $H_1:\theta\neq\theta_0$, jestliže $2\Big(L(\hat{\theta}_n)-L(\theta_0)\Big)$ překročí příslušnou kritickou hodnotu χ^2_{k-1} rozdělení.

Uvažujme nyní $p_i=p_i(a)$, pro neznámý parametr $a=(a_1,\ldots,a_m)$. Pokud bychom nyní chtěli maximalizovat pravděpodobnostní funkci

$$P(a)=\frac{n!}{x_1!\ldots x_k!}\prod_{i=1}^{k}p_i^{x_i}(a),$$

můžeme maximalizovat její logaritmus

$$\ln P(a) = \ln\left(\frac{n!}{x_1! \dots x_k!}\right) + \sum_{i=1}^{k} \ln p_i(a).$$

Věrohodnostní rovnice jsou:

$$\frac{\partial}{\partial a_j} \sum_{i=1}^{k} x_i \left(\ln p_i(a)\right) = \sum_{i=1}^{k} x_i \frac{1}{p_i(a)} \frac{\partial p_i(a)}{\partial a_j} = 0 \quad \text{pro } j = 1, \dots, m \ .$$

Tyto rovnice jsou totožné s rovnicemi (1.11), které jsme odvodili z metody minimálního χ^2. Potom za velmi podobných předpokladů jako ve větě 3 platí následující věta:

Věta 5. *Nechť $x = (x_1, \dots, x_k)$ je náhodný výběr z multinomického rozdělení o parametrech $(n, p_1, \dots, p_k)$. Nechť všechny pravděpodobnosti p_i závisí na m-rozměrném parametru $a = (a_1, \dots, a_m)$, kde $a \in A$. Nechť dále platí:*

(1) $\ln p_i(a)$ *má spojité deriavace prvního a druhého řádu pro všechna i.*
(2) $\sum_{i=1}^{k} x_i \frac{\partial \ln p_i(a)}{\partial a_h} = 0$ *má jediné řešení.*
(3) $p_i(a') \neq p_i(a'')$ *pro alespoň jedno i, když $a' \neq a''$.*
(4) Matice s prvky $\{m_{hl}(a)\}$ *má hodnost m, kde* $\{m_{hl}(a)\} = \sum_{i=1}^{k} \frac{1}{p_i(a)} \frac{\partial p_i(a)}{\partial a_h} \frac{\partial p_i(a)}{\partial a_l}$.

Nechť $\hat{a}$ je odhad parametrů metodou maximální věrohodnosti, pak statistika $2\sum_{i=1}^{k} x_i \left(\ln x_i - \ln(np_i(\hat{a}))\right)$ konverguje v distribuci k χ^2_{k-m-1} .

Důkaz. Důkaz je možné najít v [2, str. 101]. □

Tímto jsme ukázali souvislost mezi χ^2 statistikou a LR statistikou, která v tomto případě má tvar $2\sum_{i=1}^{k} x_i \left(\ln x_i - \ln(np_i(\hat{a}))\right)$.

Kapitola 2

Kontingenční tabulky

2.1 Úvod

Mějme náhodné veličiny X, Y s diskrétním rozdělením, kde X nabývá hodnot $1, \ldots, r$ a Y nabývá hodnot $1, \ldots, c$. Definujme náhodný vektor $Z = (X, Y)$. Výběr o rozsahu n z rozdělení, kterým se řídí vektor Z, můžeme uspořádat do matice $\{n_{ij}\}_{i=1,j=1}^{r,c}$, která je typu $r \times c$. Prvek n_{ij} v matici reprezentuje četnost dvojce $(X = i, Y = j)$ v tomto výběru. Označme

$$p_{ij} = P(X = i, Y = j), \quad p_{i\cdot} = \sum_{j=1}^{c} p_{ij}, \quad p_{\cdot j} = \sum_{i=1}^{r} p_{ij}.$$

Analogické značení zaveďme pro pozorované realizace

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{c} n_{ij}, \quad n = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j} = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}.$$

Matici $\{n_{ij}\}$, která obsahuje empirické četnosti, nazveme kontingenční tabulkou. Čísla $p_{i\cdot}$ a $p_{\cdot j}$ nazveme marginálními pravděpodobnostmi a analogicky $n_{i\cdot}$ a $n_{\cdot j}$ marginálními četnostmi, viz tabulka 2.1.

Náhodný výběr $\{n_{ij}\}_{i=1,j=1}^{r,c}$, který jsme zavedli výše, má multinomické rozdělení s parametry $\left(n, \{p_{ij}\}_{i=1,j=1}^{r,c}\right)$. Nadále tuto vlastnost budeme značit jako

$$\{n_{ij}\}_{i=1,j=1}^{r,c} \sim M\left(n, p_{1,1}, \ldots, p_{r,c}\right)$$

Do kontingenční tabulky o velikosti 3×2 můžeme například zapsat naše pozorování týkající se barvy očí a barvy vlasů u 25-ti jedinců způsobem,

Tabulka 2.1: Tabulka pravděpodobností

$Z = (X, Y)$	Y	$\sum$
X	$p_{11} \dots p_{1c}$ $\vdots \quad \vdots$ $p_{r1} \dots p_{rc}$	$p_{1.}$ $\vdots$ $p_{r.}$
$\sum$	$p_{.1} \dots p_{.c}$	1

Tabulka 2.2: Barva očí a vlasů

Barva očí	světlá	tmavá	celkem
modrá	3	2	5
zelená	2	4	6
tmavá	6	8	14
celkem	11	14	25

jakým je uvedeno v tabulce 2.2, kde barva očí je psána v řádcích a barva vlasů v sloupcích.

Poznámka. *V praxi nemusíme vždy mít zcela přesně diskrétně rozdělená data, například pokud budeme porovnávat míru hluku, který vydávaji 3 typy strojů, je třeba míru hluku vhodným způsobem kategorizovat ("zdiskrétnit").*

2.2 Test nezávislosti

V praxi můžeme chtít testovat, zda jsou X a Y nezávislé. Mějme jejich náhodný výběr, reprezentovaný jejich četnostmi v kontingenční tabulce.

Věta 6. *Náhodné veličiny X a Y s diskrétním rozdělením jsou nezávislé právě tehdy, když $p_{ij} = p_{i.}p_{.j}$ pro všechny (i,j).*

Důkaz. Víme, že veličiny s diskrétním rozdělením jsou nezávislé právě tehdy, když P(X=i,Y=j) = P(X=i)P(Y=j) pro všechny možné i a j. Stačí pak, že $\sum_j$ P(X=i,Y=j) = P(X=i) = $p_{i.}$.

$$p_{ij} = P(X = i, Y = j) = P(X = i)P(Y = j) = p_{i.}p_{.j}$$

□

Naše hypotézy tedy jsou

$$H_0 \;:\; p_{ij} = p_{i.}p_{.j} \quad \text{pro} \quad i = 1, \ldots, r \quad j = 1, \ldots, c \tag{2.1}$$
$$H_1 \;:\; p_{ij} \neq p_{i.}p_{.j} \quad \text{pro nějakou dvojici } (i, j). \tag{2.2}$$

Kontigenční tabulka je zaznamenání realizací n pokusů z rozdělení $Z = (X, Y)$, jehož distribuci můžeme popsat maticí pavděpodobností $\{p_{ij}\}$, kde musí platit

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1.$$

Náhodný vektor Z má multinomické rozdělení s parametry $(n, p_{11}, ..., p_{rc})$. Za předpokladu hypotézy H_0 jsou všechny pravděpodobnosti p_{ij}, jejichž počet je rc, určeny pravděpodobnostmi $p_{1.}, \ldots, p_{r.}$ a $p_{.1}, \ldots, p_{.c}$, jejichž počet je $r + c$. Parametry $p_{1.}, ..., p_{r.}$ nejsou lineárně nezávislé, neboť $\sum_i p_{i.} = 1$, $p_{r.}$ lze vyjádřit z předchozího jako

$$1 - \sum_{i=1}^{r-1} p_{i.} = p_{r.}.$$

Stejně tak můžeme vyjádřit $p_{.c}$ jako součet $c - 1$ nezávislých parametrů. Dohromady máme tedy

$$m = r - 1 + c - 1 = r + c - 2$$

neznámých parametrů. Nyní dosaďme do (1.11), kde X_i jsou naše n_{ij}, vektor a je m-rozměrný. Po dosazení máme soustavu rovnic

$$\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}}{p_{ij}} \frac{\partial p_{ij}}{\partial a_k} = 0 \qquad \text{pro } k = 1, \ldots, m,$$

kde a_k je prvek vektoru lineárně nezávislých parametrů $a = (a_1, .., a_m)$, tedy v našem případě $a = (p_{1.}, \ldots, p_{(r-1)\cdot}, p_{.1}, \ldots, p_{.(c-1)})$. Za platnosti hypotézy $p_{ij} = p_{i.}p_{.j}$ jde o soustavu $r + c - 2$ rovnic

$$\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}}{p_{i.}p_{.j}} \frac{\partial p_{i.}p_{.j}}{\partial a_k} = 0.$$

Máme tedy $$\frac{\partial p_{i.}p_{.j}}{\partial p_{k.}} = p_{.j} \quad \text{pro } i = k$$
$$= \frac{\partial(1 - (p_{1.} + .. + p_{(r-1).})p_{.j})}{\partial p_{k.}} = -p_{.j} \quad \text{pro } i = r$$
$$= 0 \quad \text{jinak.}$$

Budeme-li psát sčítance do tabulky podle n_{ij}, bude výsledná matice pro $a_k = p_{1.}$

$$\begin{pmatrix} \frac{n_{11}}{p_{1.}} & \frac{n_{12}}{p_{1.}} & \dots \\ 0 & 0 & \dots \\ \dots & \dots & \dots \\ -\frac{n_{r1}}{p_{r.}} & -\frac{n_{r2}}{p_{r.}} & \dots \end{pmatrix}.$$

Sečtením všech členů matice získáme vzorec

$$\sum_{j=1}^{c} \left(\frac{n_{ij}}{p_{i.}} - \frac{n_{rj}}{p_{r.}} \right) = 0 \quad \text{pro } i = 1, \dots, r-1 \ .$$

Triviálně platí však i pro případ $i = r$. Sečtením přes index $j = 1, \dots, c$ upravíme na tvar:

$$\begin{aligned} \frac{n_{i.}}{p_{i.}} - \frac{n_{r.}}{p_{r.}} &= 0 \\ n_{i.} &= \frac{n_{r.}}{p_{r.}} p_{i.} \ . \end{aligned}$$

Sečtením tohoto přes všechna i, při znalosti $n = \sum_{i=1}^{r} n_{i.}$, dostaneme řešení pro $p_{r.} = \frac{n_{r.}}{n}$. Když dosadíme toto řešení do každého sčítance, získáme odhad

$$p_{i.} = \frac{n_{i.}}{n} \quad \text{pro } i = 1, \dots, r \ . \tag{2.3}$$

Obdobnou úvahou dostaneme rovnice pro druhou část parametrů:

$$p_{.j} = \frac{n_{.j}}{n} \quad \text{pro } j = 1, \dots, c \ . \tag{2.4}$$

Tímto jsme dostali odhady parametrů modifikovanou metodou minimálního χ^2. Označme je $\hat{p}_{i.}$ a $\hat{p}_{.j}$. Nyní využijme toho, že podle věty 3 má veličina

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - n\frac{n_{i.}}{n}\frac{n_{.j}}{n}\right)^2}{n\frac{n_{i.}}{n}\frac{n_{.j}}{n}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \tag{2.5}$$

pro $n \to \infty$ χ^2 rozdělení. Počet stupňů volnosti je

$$rc - m - 1 = rc - (r + c - 2) - 1 = (r - 1)(c - 1).$$

Upravme (2.5)

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} \left(n\frac{n_{ij}^2}{n_{i.}n_{.j}} - 2n_{ij} + \frac{n_{i.}n_{.j}}{n} \right) \\
&= n\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} + \sum_{i=1}^{r} n_{i.} \sum_{j=1}^{c} \frac{n_{.j}}{n}
\end{aligned}$$

do výsledné podoby

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i.}n_{.j}} - n \ . \tag{2.6}$$

Hypotézu H_0 zamítáme na hladině α, pokud χ^2 je větší než kritická hodnota $\chi^2_{(r-1)(c-1)}$ pro hladinu α.

2.3 Test homogenity

Někdy je třeba testovat, zda je určitá vlastnost stejně zastoupena u zkoumaných předmětů. X_j nechť jsou testované vlastnosti a Y_i jsou předměty, na nichž dané vlastnosti zkoumáme. Testujeme-li jednou předmět Y_i, pak výsledkem může být pouze jedna vlastnost X_j. Pravděpodobnost, že se objeví právě vlastnost X_j, je p_j. U takovéhoto modelu víme, kolikrát testujeme daný předmět Y_i, což je při našem označení znalost $n_{i.}$. Naše hypotéza H_0 je, že každý řádek kontingenční tabulky má stejné multinomické rozdělení s parametry $(n_{i.}, p_1 \dots p_c)$. Jinak zapsáno:

$$H_0 : \ p_{i1}, \dots, p_{ic} \quad \text{jsou stejné pro všechna } i = 1, \dots, r$$

Podle [1, str. 283] lze dokázat, že za platnosti H_0 má veličina χ^2 počítaná podle (2.6) také asymptoticky $\chi^2_{(r-1)(c-1)}$ rozdělení. Hypotézu H_0 tedy zamítáme, překročí-li naše χ^2 kritickou hodnotu $\chi^2_{(r-1)(c-1)}$ pro hladinu α.

2.4 Test symetrie

Mějme čtvercovou kontingenční tabulku typu $c \times c$. Budeme chtít testovat, zda odpovídá realizacím symetrické matice pravděpodobností. Testujeme hypotézu symetrie

$$H_0 \ : p_{ij} = p_{ji} \quad \text{pro } i = 1, \ldots, c \quad j = 1, \ldots, c \ .$$

Neznámé parametry a_k jsou v celé matici pravděpodobností pouze

$$\begin{array}{ccccc} p_{11} & p_{12} & \cdots & p_{1,c-1} & p_{1,c} \\ & p_{22} & \cdots & p_{2,c-1} & p_{2,c} \\ & & \ddots & \vdots & \vdots \\ & & & p_{c-1,c-1} & p_{c-1,c} \ . \end{array}$$

Ostatní plynou z platnosti hypotézy H_0 a triviální rovnosti $\sum_{ij} p_{ij} = 1$. Počet nezávislých parametrů je

$$m := c + (c-1) + \ldots + 2 = \frac{c(c+1)}{2} - 1 \ .$$

Před dosazením do rovnice (1.11) je třeba si rozmyslet hodnotu $\frac{\partial p_i(a)}{\partial a_k}$. Uvažujme a_k na diagonále, tedy $k \in \{(i,i) | i = 1, \ldots, c-1\}$, pak zbudou v rovnici 2 nenulové členy:

$$\begin{aligned} \frac{\partial p_{ij}(a)}{\partial a_k} &= 1 \quad \text{pro } (i,j) = (i,i) = k \\ &= -1 \quad \text{pro } (i,j) = (c,c) \\ &= 0 \quad \text{jinak,} \end{aligned}$$

protože $p_{cc}(a) = 1 - \sum_{(i,j) \neq (c,c)} p_{ij}$. Nechť nyní a_k není na diagonále, tedy

$$k \in \{(i,j) |\ j > i;\ i = 1, \ldots, c-1;\ j = 2, \ldots, c\} \, .$$

Pak nám v součtu zbudou také 2 nenulové členy:

$$\begin{aligned} \frac{\partial p_{ij}(a)}{\partial a_k} &= 1 \quad \text{pro } k = (i,j) \text{ nebo } k = (j,i) \\ &= -2 \quad \text{pro } (i,j) = (c,c) \\ &= 0 \quad \text{jinak,} \end{aligned}$$

protože $p_{cc} = (1 - \sum_i p_{ii} - 2\sum_{j>i} p_{ij})$. Dosadíme-li do vzorce (1.11), dostáváme soustavu rovnic

$$\begin{aligned} \frac{n_{ii}}{p_{ii}} - \frac{n_{cc}}{p_{cc}} &= 0 \quad \text{pro } 1 \le i < c \qquad (2.7) \\ \frac{n_{ij}}{p_{ij}} + \frac{n_{ji}}{p_{ij}} - 2\frac{n_{cc}}{p_{cc}} &= 0 \quad \text{pro } 1 \le i < j \le c\ . \qquad (2.8) \end{aligned}$$

Upravme rovnice následovně:

$$\begin{aligned} n_{ii} &= \frac{n_{cc}}{p_{cc}} p_{ii} \quad \text{pro } 1 \le i < c \\ n_{ij} + n_{ji} &= 2\frac{n_{cc}}{p_{cc}} p_{ij} \quad \text{pro } 1 \le i < j \le c\ . \end{aligned}$$

První rovnost lze také triviálně rozšířit pro $i = c$. Prvně sečteme tyto dvě rovnice a poté sečtením přes všechny přípustné dvojce (i, j) získáme

$$\begin{aligned} n_{ij} + n_{ji} + n_{ii} &= (2p_{ij} + p_{ii})\frac{n_{cc}}{p_{cc}} \quad \text{pro } 1 \le i < j \le c \\ n &= \frac{n_{cc}}{p_{cc}}. \end{aligned}$$

Dosazením do (2.7) získáme odhady modifikovanou metodou minimálního χ^2.

$$\hat{p}_{ii} = \frac{n_{ii}}{n}, \quad \hat{p}_{ij} = \frac{n_{ij} + n_{ji}}{2n} \qquad (2.9)$$

Dosadíme nyní do věty 3, kde za $p_i(a)$ použijme odhady $\hat{p}_{ij}(a)$, dostáváme

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{c} \frac{\left[n_{ii} - n\frac{n_{ii}}{n}\right]^2}{n\frac{n_{ii}}{n}} + \sum_{i \ne j}^{c} \frac{\left[n_{ij} - n\frac{n_{ij}+n_{ji}}{2n}\right]^2}{n\frac{n_{ij}+n_{ji}}{2n}} = \\ &= \sum_{i<j}^{c} \frac{\left[n_{ij} - n_{ji}\right]^2}{n_{ij} + n_{ji}}. \end{aligned} \qquad (2.10)$$

Dle věty 3 a za platnosti H_0 má (2.10) pro $n \to \infty$ χ^2-rozdělení o

$$c^2 - m - 1 = c^2 - \left(\frac{c(c-1)}{2} - 1\right) - 1 = \frac{c(c-1)}{2} \qquad (2.11)$$

stupních volnosti. Proto hypotézu symetrie H_0 zamítáme, pokud χ^2 je větší než kritická hodnota $\chi^2_{\frac{c(c-1)}{2}}$ na hladině α.

Kapitola 3

Markovovy řetězce

3.1 Úvod

Definice 2. *Nechť (Ω, A, P) je pravděpodobnostní prostor. Mějme $T \subset \mathbb{R}$. Soustava reálných náhodných veličin $\{X_t, t \in T\}$ definovaných na tomto pravděpodobnostním prostoru se nazývá náhodný proces. Nechť $\{X_t, t \in T\}$ nabývá hodnot z množiny S. Potom S se nazývá množina stavů procesu $\{X_t, t \in T\}$.*

Definice 3. *Pokud $T = \mathbb{Z}$ nebo $T = \mathbb{N}^0$, pak $\{X_t, t \in T\}$ nazveme náhodný proces s diskrétním časem.*

Nadále předpokládejme, že máme náhodný proces s diskrétním časem, kde $T = \mathbb{N}^0$, s množinou stavů S, kde $S = \{1, \ldots, m\}$. Budeme se dále zajímat pouze o procesy, které mají speciální vlastnost. Stav, do kterého se proces dostane v čase $t+1$, závisí pouze na tom, ve kterém stavu byl v čase t. Nezávisí na tom, jakými stavy prošel do času t. Řetězce s touto vlastností nazveme Markovovými řetězci.

Definice 4. *Náhodný proces s diskrétním časem a množinou stavů $S = (1, \ldots, m)$ a s vlastností*

$$P\left[X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \ldots, X_0 = i_0\right] = P\left[X_{t+1} = j | X_t = i\right], \quad (3.1)$$

kde i_n je stav, ve kterém se nacházel proces v čase n pro $n = t-1, \ldots, 1, 0$, nazveme Markovovým řetězcem.

Rovnost (3.1) se nazývá markovská vlastnost. Dále budeme uvažovat pouze procesy s touto vlastností.

Definice 5. *Pokud podmíněné pravděpodobnosti*

$$P(X_{t+1} = j | X_t = i) = p_{ij}(t, t+1)$$

existují, nazveme je pravděpodobnostmi přechodu ze stavu i v čase t do stavu j v čase $t+1$.

Z předchozí definice přímo plyne rovnost

$$\sum_{j=1}^{m} p_{ij}(t, t+1) = 1 \quad \text{pro všechna } t \in T.$$

Definice 6. *Pravděpodobnosti přechodu $p_{ij}(t, t+1)$ pro čas t můžeme zapsat do matice typu $m \times m$. Tuto matici pak nazveme maticí pravděpodobností přechodu v čase t.*

Definice 7. *Označme*

$$P(X_0 = i) = p_i \quad \textit{pro } i = 1, \ldots, m,$$

pak $\{p_i, i = 1, \ldots, m\}$ se nazývá počátečním rozdělením Markovova řetězce.

Definice 8. *Řekneme že Markovův řetězec je homogenní, když pravděpodobnosti přechodu nezávisí na čase, tzn. když*

$$p_{ij}(t, t+1) = p_{ij} \quad \textit{pro } i, j \in S, \quad t \in T.$$

K definování homogenního Markovova řetězce stačí jediná matice pravděpodobností přechodu:

$$\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}$$

V praxi často nemáme zadané teoretické pravděpodobnosti, ale máme možnost sledovat jednotlivé řetězce. Jako v předchozím předpokládejme, že máme homogenní Markovův řetězec s diskrétním časem a s konečnou množinou stavů $S = \{1, \ldots, m\}$ s počátečním rozdělením $(p_1, \ldots, p_m)$ a s maticí pravděpodobností přechodu $\{p_{ij}\}$. Chceme pozorovat k nezávislých

náhodných procesů začínajících v libovolných stavech $i_0, \ldots, i_k$. Označme y_i jako počet procesů začínajících ve stavu i. Pak zřejmě platí

$$\sum_{i=1}^{m} y_i = k.$$

Označme ještě $x_{ij}(t, t+1)$ počet přechodů ze stavu i v čase t do stavu j v čase $t+1$; dostaneme

$$\sum_{j=1}^{m} \sum_{i=1}^{m} x_{ij}(t, t+1) = k \quad \text{pro všechna } t \in T.$$

Pozorujme proces v čase n. Za předpokladu nezávislosti k pozorovaných procesů můžeme vypočítat pravděpodobnost jedné určité trajektorie

$$f(y_1, \ldots, y_m, x_{11}(0,1), \ldots, x_{mm}(n-1, n)) = \prod_{i=1}^{m} p_i^{y_i} \prod_{i=1}^{m} \prod_{j=1}^{m} \prod_{t=0}^{n-1} p_{ij}^{x_{ij}(t,t+1)}. \tag{3.2}$$

Označme $x_{ij} = \sum_{t=0}^{n-1} x_{ij}(t, t+1)$, tím zjednodušíme (3.2). Přeuspořádáním pak vznikne

$$f(y_1, \ldots, y_m, x_{11}(0,1), \ldots, x_{mm}(n-1, n)) = \\ = \left\{ \prod_{i=1}^{m} p_i^{y_i} \right\} \left\{ \prod_{j} p_{1j}^{x_{1j}} \right\} \cdots \left\{ \prod_{j} p_{mj}^{x_{mj}} \right\}. \tag{3.3}$$

Každý součinitel ve složených závorkách odpovídá pravděpodobnostem multinomického rozdělení, až na přenásobení konstantou, neboť $\sum_{j=1}^{m} p_{ij} = 1$ pro $i = 1, \ldots, m$ a $\sum_{i=1}^{m} p_i = 1$. Funkci f tedy maximalizujeme, pokud budeme maximalizovat všech $m+1$ součinitelů. Odhady parametrů metodou maximální věrohodnosti jsme již odvodili v 1. kapitole:

$$\hat{p}_i = \frac{y_i}{k}, \qquad \hat{p}_{ij} = \frac{x_{ij}}{n_i}, \quad \text{kde } n_i = \sum_{j=1}^{m} x_{ij} \ .$$

Jak vyplývá z definice funkce f, můžeme na naše pozorování nahlížet, jako kdybychom měli m nezávislých pozorování z multinomických rozdělení s parametry $(n_i, p_{i1}, \ldots, p_{im})$,

$$\begin{array}{cc} \text{1. pozorování} & \sim M(n_1, p_{11}, \ldots, p_{1m}) \\ \text{2. pozorování} & \sim M(n_2, p_{21}, \ldots, p_{2m}) \\ \vdots & \vdots \\ m\text{. pozorování} & \sim M(n_m, p_{m1}, \ldots, p_{mm}) \ . \end{array}$$

Matici četností $\{x_{ij}\}_{i=1,j=1}^{m,m}$ můžeme považovat za typ kontingenční tabulky.

3.2 Test $H : p_{ij} = p_{ij}^0$

Předpokládejme, že budeme chtít testovat hypotézy

$$\begin{array}{rcll} H_0 & : & p_{ij} = p_{ij}^0 & \text{pro } i = 1, \ldots, m \quad j = 1, \ldots, m \\ H_1 & : & p_{ij} \neq p_{ij}^0 & \text{pro nějakou dvojici } (i, j) \end{array}$$

Jak z formulace hypotézy H_0 vyplývá, uvažujme diskrétní homogenní Markovův řetězec s konečnou množinou stavů S. Použijeme-li LR-test, má veličina

$$LR = 2 \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ij} \ln \frac{x_{ij}}{n_i p_{ij}^0} \tag{3.4}$$

asymptoticky $\chi^2_{m(m-1)}$ rozdělení, neboť máme m multinomických rozdělení a v každém $m - 1$ nezávislých proměnných. Připomeňme, že například p_{im} lze vypočítat ze vztahu $\sum_j p_{ij} = 1$.

3.3 Test symetrie

Nyní chtějme testovat hypotézu symetrie, tedy že matice pravděpodobností přechodu je symetrická.

$$\begin{array}{rcll} H_0 & : & p_{ij} = p_{ji} & \text{pro } i = 1, \ldots, m \quad j = 1, \ldots, m \\ H_1 & : & p_{ij} \neq p_{ji} & \text{pro nějakou dvojici } (i, j) \end{array}$$

Opět použijeme LR-test a uplatníme také předchozí odhad parametru p_{ij} modifikovanou metodou minimálního χ^2 dle (2.9).

$$LR = 2 \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ij} \left(\ln x_{ij} - \ln \frac{1}{2}(x_{ij} + x_{ji}) \right) = 2 \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ij} \left(\ln \frac{2x_{ij}}{x_{ij} + x_{ji}} \right);$$

LR má $\chi^2_{\frac{m(m-1)}{2}}$ rozdělení. Předchozí výraz je možno dále zjednodušit, neboť pro diagonální prvky je logaritmus roven nule. Dostaneme

$$LR = 2\sum_{i\neq j} x_{ij}\left(\ln\frac{2x_{ij}}{x_{ij}+x_{ji}}\right) \ .$$

Počet stupňů volnosti jsme získali stejně jako v testu symetrie (2.11) v kapitole 2.

3.4 Test homogenity řádků matice přechodu

Zde budeme testovat, zda máme speciální typ matice pravděpodobností se shodnými řádky:

$$\begin{pmatrix} p_1 & p_2 & \dots & p_m \\ p_1 & p_2 & \dots & p_m \\ \vdots & \vdots & \vdots & \vdots \\ p_1 & p_2 & \dots & p_m \end{pmatrix}$$

Testujeme hypotézu

$$\begin{aligned} H_0 &: p_{ij} = p_j \quad \text{pro } i = 1,\dots,m, \quad j = 1,\dots,m \\ H_1 &: p_{ij} \neq p_j \quad \text{pro nějakou dvojici} (i,j) \ . \end{aligned}$$

Použijme LR-test. Pak z (2.3),(2.4) a (2.1) plyne, že

$$LR = 2\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}\left(\ln x_{ij} - \ln\frac{x_{i.}x_{.j}}{n}\right) \tag{3.5}$$

má $\chi^2_{(m-1)^2}$ rozdělení, kde $x_{i.}$ a $x_{.j}$ jsou marginální četnosti definované dříve.

3.5 Test nezávislosti na t

Nyní přestaňme předpokládat, že námi pozorovaný Markovský řetězec je homogenní. Jinými slovy, nechť p_{ij} se s časem t mění. Maximální věrohodnostní odhady parametru p_{ij} v čase $t = 0,\dots,n-1$ jsou

$$\hat{p}_{ij}(t,t+1) = \frac{x_{ij}(t,t+1)}{x_{i.}(t,t+1)}, \quad \text{kde } x_{i.}(t,t+1) = \sum_{j=1}^{m} x_{ij}(t,t+1) \ .$$

Testujme hypotézu, zda p_{ij} opravdu závisí na t.

$$\begin{aligned} H_0 &: p_{ij}(t,t+1) = p_{ij} \quad \text{pro } t = 0,\ldots,n-1 \\ H_1 &: p_{ij}(t,t+1) \neq p_{ij} \quad \text{pro nějaké } t. \end{aligned}$$

Za platnosti H_0 je maximální věrohodnostní odhad p_{ij} roven

$$\hat{p}_{ij} = \frac{x_{ij}}{n_{i.}}, \quad \text{kde } x_{ij} = \sum_{t=0}^{n-1} x_{ij}(t,t+1), \quad n_{i.} = \sum_{t=0}^{n-1} x_{i.}(t,t+1) \ .$$

Použijeme-li LR-test pro H_0, dostaneme

$$LR = 2\sum_{t=}^{n-1}\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}(t,t+1)\left[\ln x_{ij}(t,t+1) - \ln\left(x_{i.}(t,t+1)\frac{x_{ij}}{n_{i.}}\right)\right] \ . \tag{3.6}$$

LR má $\chi^2_{(n-1)(m-1)m}$ rozdělení. Máme celkem $m(m-1)$ proměných v n tabulkách, za H_0 jsou závislé na $m(m-1)$ parametrech. Počet stupňů volnosti je tedy $m(m-1)(n-1)$. Hypotézu H_0 zamítáme na hladině α, pokud $LR > \chi^2_{(n-1)(m-1)m}(\alpha)$.

Kapitola 4

Příklady

Ukažme si nyní použití některých uvedených testů na příkladech. Data jsou ze Statistické ročenky České republiky pro rok 2002, 2003 a 2004 (viz [5]).

4.1 Příklady ke kontingenčním tabulkám

Test nezávislosti

Vycházejme z tabulky 4.1. V řádcích je uvedeno maximální dosažené vzdělání ženicha v době svatby a ve sloupcích vzdělání nevěsty. Například počet sňatků za rok 2002 uzavřených mezi muži základního vzdělání a ženami, které dokončily vysokou školu, je 66.

Budeme testovat hypotézu, že míra dosaženého vzdělání neovlivňuje volbu manžela nebo manželky. Podle (2.6) je $\chi^2 = 29068$. Počet stupňů volnosti je $(n-1)(n-1) = 9$. Kritická hodnota $\chi^2_9(0,99) = 21,67$, proto na hladině

Tabulka 4.1: Počet sňatků podle vzdělání ženicha a nevěsty za rok 2002

ženich ↓	základní	vyučená	maturita	vysoká š.	$\sum$
základní	2278	1011	664	66	4019
vyučen	2106	11310	7700	644	21760
maturita	817	3000	12896	2390	19103
vysoká š.	139	334	3597	3780	7850
$\sum$	5340	15655	24857	6880	52732

$0,01$ hypotézu nezávislosti zamítáme. Uveďme ještě pro doplnění hodnotu, počítanou pomocí LR-testu, $LR = 23715$. I tento test zamítá nezávislost.

Test homogenity

Pro příklad testu homogenity použijme tabulku 4.2. Předpokládejme, že se v následujícím roce (t+1) vyloví tolik ryb určitého druhu, jako byla poptávka po tomto druhu ve sledovaném roce (t). Testujeme, zda se mění poptávka po 3 následujících letech. Uvažme druhy ryb, které se chovají převážně v sádkách. Předpokládejme proto, že chov, tedy celkový počet ryb, ovlivňuje jeden faktor. Nechť tento faktor ovlivňuje všechny vybrané druhy ryb stejně, např. počasí.

Testujeme homogenitu sloupců. Počítáme-li podle (2.6), získáme $\chi^2 = 0,5264$. Počet stupňů volnosti je $(r-1)(c-1) = 4$. Kritická hodnota $\chi^2_4(0,01) = 13,2767$. Hypotézu homogenity nezamítáme na hladině $0,01$. Opět pro doplnění je $LR = 0,5244$, což je také nevýznamné.

Test symetrie

Nyní nás v tabulce (4.1) bude zajímat, zda je symetický vztah mužů a žen vůči vzdělání partnera. Testujeme, zda muži se vzděláním A preferují nevěstu se vzděláním B stejnou měrou, jako ženy se vzděláním A upřednostňují ženicha se vzděláním B. Podle (2.10) je $\chi^2 = 2832,6$. Kritická hodnota $\chi^2_{n(n-1)/2=6}(0,01) = 16,814$ je menší než χ^2. Proto hypotézu symetrie zamítáme. Pro zajímavost $LR = 2917$ je také větší než příslušná kritická hodnota, proto i zde zamítáme hypotézu symetrie.

Tabulka 4.2: Množství vylovených ryb v jednotlivých letech (v tunách)

Druhy ryb	2002	2003	2004	$\sum$
Pstruh	50	40	37	127
Úhoř	29	27	25	81
Okoun	54	52	44	150
$\sum$	133	119	106	358

Tabulka 4.3: Počet sňatků podle vzdělání snoubenců za rok 2003

ženich ↓	základní	vyučená	maturita	vysoká š.	$\sum$
základní	1830	790	582	70	3272
vyučen	1720	9824	6977	639	19160
maturita	749	2830	12413	2371	18363
vysoká š.	112	367	3635	4034	8148
$\sum$	4411	13811	23607	7114	48943

4.2 Příklad k Markovovým řetězcům

Test nezávislosti na t

Testujme, zda se nemění matice přechodu s časem t. Tedy, zda muž se vzděláním A si vezme ženu se vzděláním B v roce 2002 se stejnou pravděpodobností jako jiný pár stejného vzdělání v roce 2003.

Použijme data z tabulek 4.1 a 4.3. Dosadíme-li do (3.6), získáme $LR = 28,87$. Počet stupňů volnosti je $m(m-1)(n-1) = 12$. Porovnáním s kritickou hodnotou $\chi^2_{12}(0,01) = 26,22$ musíme hypotézu nezávislosti na t zamítnout na hladině 0,01. Nyní uveďme pro zajímavost hodnotu χ^2 testu: $\chi^2 = 28,88$. I v tomto případě jsou testové statistiky velmi podobné

Závěr

V této práci jsme studovali základní testy v kontingenčních tabulkách. Ukázali jsme jejich podobnost s testy matic pravděpodobností přechodu pro Markovské řetězce s konečnou množinou stavů. Ukázali jsme také souvislost mezi χ^2 testem a LR testem a také mezi odhadem parametrů metodou minimálního χ^2 a metodou maximální věrohodnosti. Na závěr jsme našli několik příkladů pro použití těchto testů. Testy v této práci uvedené jsou poměrně nestabilní pro odlehlá data, která je možné v praxi získat (třeba jako chybu v měření). Proto je potřeba zvážit jejich použití v rozsáhlých souborech dat.

Literatura

[1] Anděl J.: *Základy matematické statistiky*, Matfyzpress, Praha, 1985.

[2] Anderson E. B.: *Discrete Statistical Models with Social Science Aplications*, Noth-Holland, 1980.

[3] Johnson L. N., Kotz S.: *Discrete Distributions*, Wiley, New York, 1969.

[4] Prášková Z., Lachout P.: *Základy náhodných procesů*, Karolinum, Praha, 2005.

[5] Český statistický úřad: *Statistická ročenka České republiky, 2002, 2003, 2004*, Scientia, Praha, 2002, 2003, 2004.