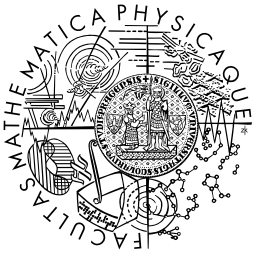


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Peter Vook

Statistické testy ve stratifikovaných čtyřpolních tabulkách

Katedra pravděpodobnosti a matematické statistiky

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

Praha 2018

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) 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

Chcel by som poďakovať vedúcemu bakalárskej práce doc. RNDr. Arnoštovi Komárkovi, Ph.D. za odborné vedenie, cenné rady pri písaní práce, ochotu a za venovaný čas.

Název práce: Statistické testy ve stratifikovaných čtyřpolních tabulkách

Autor: Peter Vook

Katedra: Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: Táto práca sa zaoberá štatistickými testami v stratifikovaných štvorpoľných tabuľkách. Je v nej odvodených viacero testov podmienenej nezávislosti. Opísaný je aj test homogénnej asociácie. Najprv sú zavedené kontingenčné tabuľky ľubovoľných rozmerov a multinomické rozdelenie. Ďalej pokračujeme opisom štvorpoľných tabuliek a ich binomickou reprezentáciou. V ďalšej časti sa zaoberáme pomerom šancí a jeho asymptotickým rozdelením. Potom nasleduje formálne zavedenie stratifikácie a súvisiacich pojmov. V ďalšej kapitole sa nachádza odvodenie testových štatistík pre testy podmienenej nezávislosti vrátane známeho Cochran-Mantel-Haenszelovho testu založeného na hypergeometrickom rozdelení. Kapitola tiež obsahuje popis Breslowovho-Dayovho testu homogénnej asociácie. Na záver je prevedená numerická simulácia vybraných testov.

Klíčová slova: stratifikované štvorpoľné tabuľky, kontingenčné tabuľky, hypergeometrické rozdelenie, Cochranov-Mantelov-Haenszelov test, Breslowov-Dayov test, pomer šancí

Title: Statistical tests in stratified fourfold tables

Author: Peter Vook

Department: Department of Probability and Mathematical Statistics

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

Abstract: This paper deals with statistical tests in stratified fourfold tables. Several tests of conditional independence are derived in it. A test of homogeneous association is also described. At first, contingency tables with arbitrary dimensions and multinomial distribution are defined. Then we continue with a description of fourfold tables and their binomial representation. In the next section we deal with an odds ratio and its asymptotic distribution. Formal definition of stratification and relevant terms follows afterwards. In the next chapter a derivation of test statistics for conditional independence tests including the well-known Cochran-Mantel-Haenszel test based on a hypergeometric distribution can be found. This chapter also includes a description of Breslow-Day test of homogeneous association. A numerical simulation of chosen tests is performed eventually.

Keywords: stratified fourfold tables, contingency tables, hypergeometric distribution, Cochran-Mantel-Haenszel test, Breslow-Day test, odds ratio

Obsah

Zoznam použitých skratiek

$\xrightarrow{d}$	konvergencia v distribúcii
$\xrightarrow{\mathsf{P}}$	konvergencia v pravdepodobnosti
$\mathsf{N}(\mu,\sigma^2)$	normálne rozdelenie so strednou hodnotou μ a rozptylom σ^2
χ^2_k	χ^2 rozdelenie s k stupňami voľnosti
$\chi^2_k(1-\alpha)$	$(1-\alpha)$-kvantil χ^2 rozdelenia s k stupňami voľnosti

Úvod

Pri pozorovaniach v praxi sa častokrát stáva, že môžeme získané dáta zatriediť do niekoľkých kategórií. Takýto popis je praktický, pretože naň potrebujeme iba konečné množstvo kategórií a schopnosť popísať príslušnosť sledovaného objektu do jednej z nich. To je dôvod, prečo sa s nimi stretneme pri rôznych odvetviach ľudskej činnosti.

Problém nastáva, keď sa rozhodujeme, aké všetky parametre potrebujeme pri meraniach brať do úvahy. Na namerané hodnoty totižto môže mať vplyv parameter, od ktorého by sme to nečakali. To nám dáva priestor na to, aby sme klasické testy prevádzané na dvoch kategoriálnych veličinách obohatili o to, že namerané hodnoty rozdelíme do niekoľko vhodne vybraných kategórií a príslušnosť do kategórie zahrnieme pri testovaní. Takýto postup nazveme stratifikácia. V tejto práci budeme skúmať stratifikované testovanie na štvorpoľných tabuľkách, čo sú tabuľky zachytávajúce vzťah dvoch kategoriálnych veličín, ktoré nadobúdajú práve 2 hodnoty (v praxi sú to napríklad odpovede áno/nie, prípadne výsledky vyliečil/nevyliečil a pod.). Počet kategórií, na ktoré pozorovania rozdelíme môže byť ľubovoľný – niekedy len 2 (pozorovania od žien a mužov), pokojne ale aj viac (pozorovania podľa krajov republiky).

V práci sa budeme venovať hlavne testovaniu toho, či sú dve sledované veličiny nezávislé pre všetky úrovne stratifikácie. Tejto vlastnosti sa hovorí podmienená nezávislosť. Krátku časť venujeme aj testu, ktorý skúma, či vplyv týchto dvoch veličín na seba (ak takýto vplyv existuje) je rovnaký pre každú úroveň. Takáto vlastnosť sa nazýva homogénna asociácia. Výskum v oblasti stratifikácie pri pozorovaniach je z veľkej časti spôsobený medicínskymi pozorovaniami. Keď sa napríklad nejaká vlastnosť objaví u jednej populácie, je prirodzené, že sa bude skúmať aj u ostatných. Preto väčšina príkladov v tejto práci aj v súvisiacej literatúre pochádza práve z medicínskeho prostredia.

V prvej kapitole si zavedieme kontingenčné tabuľky a predstavíme si ich zvyčajnú reprezentáciu pomocou multinomického rozdelenia. Potom si situáciu trocha zjednodušíme tak, že začneme uvažovať iba štvorpoľné tabuľky. Na záver kapitoly si ukážeme ďalšiu – binomickú reprezentáciu štvorpoľných tabuliek.

V druhej kapitole sa budeme venovať významnému spôsobu, ako opísať vzájomný vzťah sledovania dvoch veličín nadobúdajúcich práve dve hodnoty – pomeru šancí. Odvodíme si takisto jeho asymptotické rozdelenie.

V tretej kapitole sformalizujeme tento úvod a presne si zavedieme stratifikáciu.

Vo štvrtej kapitole si odvodíme najvýznamnejší test používaný na stratifikované tabuľky – Cochranov-Mantelov-Haenszelov test. Pozrieme sa trocha aj na históriu tohto testu. Takisto si odvodíme test založený na rozdelení pomeru šancí, ktoré sme ukázali v druhej kapitole. Načrtneme si aj Breslowov-Dayov test homogénnej asociácie.

V piatej kapitole prevedieme simuláciu v programe R. Budeme sa tým snažiť porovnať z numerického hľadiska testy odvodené vo štvrtej kapitole.

1. Kontingenčné tabuľky

V tejto kapitole uvedieme základnú teóriu kontingenčných tabuliek. Ako je uvedené v knihe autorov Fleiss, Levin a Paik (2003), kontingenčné tabuľky boli a pravdepodobne stále sú najčastejšie používaným spôsobom prezentovania štatistickej evidencie. Po zadefinovaní sa pozrieme na test nezávislosti v kontingenčných tabuľkách založený na asymptotickom rozdelení.

1.1 Zavedenie

Označme X a Y dve diskrétne rozdelené náhodné veličiny odpovedajúce kategoriálnej odozve. X má I kategórií, Y má J kategórií, I,$J \in \mathbb{N}$ (teda $X \in \{1,\ldots,I\}, Y \in \{1,\ldots,J\}$). Označme

$$p_{ij} = \mathsf{P}(X = i, Y = j), \qquad p_{i.} = \sum_j p_{ij} = \mathsf{P}(X = i), \qquad p_{.j} = \sum_i p_{ij} = \mathsf{P}(Y = j). \tag{1.1}$$

Uvažujme náhodný výber $(X_1,Y_1)^\top,\ldots,(X_n,Y_n)^\top$ s (pevným) rozsahom n z tohto rozdelenia. Označme $n_{ij} = \sum_{k=1}^{n} \mathbb{1}_{\{X_k=i,Y_k=j\}}$ počet tých prípadov, keď sa vo výbere vyskytla dvojica (i,j). Analogicky ako v prípade (1.1) označíme

$$n_{i.} = \sum_j n_{ij}, \qquad n_{.j} = \sum_i n_{ij}. \tag{1.2}$$

V zmysle definovaného značenia teda platí

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

Takto máme pripravené všetko na definíciu kontingenčnej tabuľky.

Definícia 1. *Matica $(n_{ij})_{i\in\{1,\ldots,I\},\, j\in\{1,\ldots,J\}}$ spĺňajúca podmienky z predchádzajúceho odseku sa nazýva* kontingenčná tabuľka*.

Náhodná veličina n_{ij} sa nazýva pozorovaná početnosť *pre kombináciu kategórií i a j. Čísla $n_{i.}$ a $n_{.j}$ sa nazývajú* marginálne početnosti*, hodnoty $p_{i.}$ a $p_{.j}$ sú* marginálne pravdepodobnosti*.*

Podobne ako kontingenčnú tabuľku (n_{ij}) môžeme zostaviť tabuľku pravdepodobností (p_{ij}), ktorá popisuje združené rozdelenie vektoru $(X,Y)^\top$ aj marginálne rozdelenie veličín X a Y. Príkladom na obidva druhy je Tabuľka 1.1.

Teraz sa pozrieme na rozdelenie náhodného vektoru pozorovaných početností. Na to si najprv zadefinujeme multinomické rozdelenie a potom sa pozrieme na jeho využitie pri testovaní nezávislosti v kontingenčných tabuľkách.

Multinomické rozdelenie si môžeme predstaviť na názornom príklade uvedenom v knihe Anděl (2011). Máme urnu a v nej guľôčky $k \geq 2$ farieb. Nech pravdepodobnosť vytiahnutia guľôčky s i-tou farbou je p_i; platí $\sum_{i=1}^{k} p_i = 1$. Nezávisle na sebe n-krát vyberieme po jednej guľôčke, ktorú hneď vrátime. Označme X_i počet guličiek i-tej farby, ktoré sme takto vybrali. Multinomické rozdelenie je potom združené rozdelenie veličín $X_1,\ldots,X_n$.

* anglicky *contingency table*

	Y			
X	1	$\cdots$	J	Σ
1	n_{11}	$\cdots$	n_{1J}	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	$\cdots$	n_{IJ}	$n_{I.}$
Σ	$n_{.1}$	$\cdots$	$n_{.J}$	n

(a) kontingenčná

	Y			
X	1	$\cdots$	J	Σ
1	p_{11}	$\cdots$	p_{1J}	$p_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	p_{I1}	$\cdots$	p_{IJ}	$p_{I.}$
Σ	$p_{.1}$	$\cdots$	$p_{.J}$	1

(b) pravdepodobností

Tabuľka 1.1

Definícia 2 (Multinomické rozdelenie)**.** *Nech $k \geq 2$ a n sú prirodzené čísla a $\boldsymbol{p} = (p_1, \ldots, p_k)^\top$ je vektor konštánt spĺňajúci $p_i \in (0{,}1)$ pre všetky $i \in \{1, \ldots, n\}$ a $\sum_{i=1}^{k} p_i = 1$.*

Povieme, že náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_k)^\top$ má multinomické rozdelenie s parametrami n a $\boldsymbol{p}$ (značíme $\boldsymbol{X} \sim \mathsf{Mult}_k(n,\boldsymbol{p})$) práve vtedy, keď jeho hustota k súčinovej počítacej miere na $\mathbb{Z}^k$ je

$$\mathsf{P}(X_1 = x_1, \ldots, X_k = x_k) = \begin{cases} \frac{n!}{x_1! \cdot \ldots \cdot x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} & \sum_{i=1}^{k} x_i = n, \\ & x_i \in \mathbb{N}_0 \ \forall i \in \{1, \ldots, n\} \\ 0 & \textit{inak} \end{cases}$$

1.2 Test nezávislosti

Ako uvádza Anděl (2011), najčastejšou úlohou pri rozbore dvojrozmerných kontingenčných tabuliek je prevedenie testu hypotézy, že veličiny X a Y sú na sebe nezávislé. K testu si najprv pripravíme pomocné lemma.

Lemma 1. *Veličiny X a Y sú nezávislé vtedy a len vtedy, keď platí*

$$p_{ij} = p_{i.} p_{.j} \quad \textit{pre všetky dvojice } (i,j) \ \ (1 \leq i \leq I,\ 1 \leq j \leq J)\,. \tag{1.3}$$

Dôkaz. Dve veličiny X a Y sú podľa definície nezávislé, ak

$$\mathsf{P}(Y \in A, Z \in B) = \mathsf{P}(Y \in A)\, \mathsf{P}(Z \in B) \quad A \subset \{1, \ldots, I\}, B \subset \{1, \ldots, J\} \tag{1.4}$$

Ak zvolíme za A a B jednobodové množiny, dostávame z (1.4) priamo (1.3), teda sme dokázali nutnosť podmienky (1.3).

Postačiteľnosť tejto podmienky dostaneme tak, že za jej platnosti vezmeme množiny A a B ako $A = \{i_1, \ldots, i_a\}, B = \{j_1, \ldots, j_b\}$, kde $i_1, \ldots, i_a$ sú rôzne čísla z množiny $\{1, \ldots, I\}$, čísla $j_1, \ldots, j_b$ sú rôzne z množiny $\{1, \ldots, J\}$. Za týchto predpokladov máme

$$\mathsf{P}(Y \in A, Z \in B) = \sum_{s=1}^{a} \sum_{t=1}^{b} p_{i_s j_t}, \quad \mathsf{P}(Y \in A) = \sum_{s=1}^{a} p_{i_s .}, \ \mathsf{P}(Z \in B) = \sum_{t=1}^{b} p_{.j_t}.$$

Z týchto vzorcov za platnosti 1.3 vyplýva 1.4.

□

Teraz si uvedieme tvrdenie o asymptotickom rozdelení testovej štatistiky, ktorá sa používa pri teste nezávislosti.

Tvrdenie 2. *Nech $n, n_{ij}, n_{i.}, n_{.j}$ zodpovedajú značeniu zavedenému v definícii kontingenčnej tabuľky. Potom veličina*

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}, \tag{1.5}$$

má asymptoticky rozdelenie $\chi^2_{(I-1)(J-1)}$.

Dôkaz. (Viď Anděl, 1985, Sekcia XII.1.).

□

Hypotézu H_0 o nezávislosti veličín X a Y teda zamietame, keď vyjde $\chi^2 \geq \chi^2_{(I-1)(J-1)}(1-\alpha)$. Ako je uvedené v (Anděl, 2011), obyvkle sa vyžaduje, aby všetky teoretické početnosti $n_{i.}n_{.j}/n$ boli väčšie, než 5. Ak táto podmienka nie je splnená, zvyčajne sa spájajú niektoré riadky alebo stĺpce.

1.3 Štvorpoľné tabuľky

Vo väčšine práce sa budeme zaoberať štvorpoľnými tabuľkami, čo sú kontingenčné tabuľky zachytávajúce vzťah kategoriálnych veličín nadobúdajúcich práve 2 hodnoty.

Definícia 3. *Štvorpoľná tabuľka*[†] *je kontingenčná tabuľka s $I = J = 2$.*

Štvorpoľná tabuľka sleduje prítomnosť dvoch charakteristík na nejakej populácii. Pripájam modelovú štvorpoľnú tabuľku 1.2, ako ju uvádzajú Fleiss a kol. (2003).

	Charakteristika B		
Charakteristika A	Prítomná	Neprítomná	Spolu
Prítomná	n_{11}	n_{12}	$n_{1.}$
Neprítomná	n_{21}	n_{22}	$n_{2.}$
Spolu	$n_{.1}$	$n_{.2}$	n

Tabuľka 1.2: Modelová štvorpoľná tabuľka

Vzorec (1.5) sa dá v prípade štvorpoľnej tabuľky podstatne zjednodušiť (postup uvádza Anděl (1985)).

Tvrdenie 3. *V štvorpoľnej tabuľke platí*

$$\chi^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}. \tag{1.6}$$

Veličina χ^2 tu má asymptoticky rozdelenie χ^2_1.

Dôkaz. Dostaneme podrobným rozpísaním vzťahov pre marginálne početnosti a celkovú početnosť a následnou úpravou.

□

Teraz uvediem konkrétny príklad na χ^2-test nezávislosti v štvorpoľnej tabuľke.

[†] anglicky *fourfold table*

Príklad. Bol prevedený experiment, pri ktorom sa testujú mechanické vlastnosti skrutiek. Testovalo sa 100 skrutiek, z ktorých 50 bolo vyrobených z ocele bez tepelnej úpravy a 50 z tepelne upravenej (kalenej) ocele pri zafixovaní ostatných parametrov. Pri experimente sa skrutky testovali ťahom. Niektoré z nich sa pritom mechanicky poškodili, iné nie. Údaje z experimentu sú zapísané v tabuľke 1.3. Úlohou je zistiť, či mechanická odolnosť skrutky závisí na tom, či bola vyrobená z tepelne upravenej ocele.

	Poškodená	Nepoškodená	Celkom
Tepelne upravená oceľ	8	42	50
Oceľ bez tepelnej úpravy	21	29	50
Celkom	29	71	100

Tabuľka 1.3: Údaje o skrutkách

Podľa tvrdenia 3 napočítame testovú štatistiku χ^2, ktorá vyjde 8,208. Pretože $\chi^2 \geq \chi^2_1 \geq \chi^2_1(0{,}95) = 3{,}84$, zamietame hypotézu, že materiál skrutky nemá vplyv na mechanickú odolnosť. △

Nakoniec si ukážeme trocha odlišnú reprezentáciu štvorpoľných tabuliek. Doteraz sme uvažovali prvky tabuľky generované multinomickým rozdelením. Za predpokladu pevných marginálnych početností (napr. riadkových) môžeme dáta v tabuľke reprezentovať ako realizáciu dvoch alternatívnych výberov. Potom napríklad ak marginálna početnosť v prvom riadku bude $n_{1.}$, tak dáta v prvom riadku reprezentujú výber z rozdelenia $\mathsf{Bi}(n_{1.}, p_1), p_1 \in (0{,}1)$. Dôležité je uvedomiť si, že neplatí $p_1 = p_{1.}$. Z nasledujúceho tvaru štvorpoľnej tabuľky je takáto reprezentácia dvoma nezávislými výbermi z binomického rozdelenia pekne viditeľná.

	$B = 1$	$B = 0$	Spolu
$A = 1$	X_1	$n - X_1$	n
$A = 0$	X_2	$m - X_2$	m
Spolu	$X_1 + X_2$	$n + m - X_1 - X_2$	$n + m$

Pre niektoré účely sa nám viac hodí práve takáto reprezentácia tabuľky.

2. Pomer šancí

V druhej kapitole sa zameriame na dôležitú mieru asociácie dvoch výberov z alternatívneho rozdelenia – pomer šancí a jeho asymptotické vlastnosti.

2.1 Miery asociácie

Nech teda $Y_{11}, \ldots, Y_{1n}$ je náhodný výber z alternatívneho rozdelenia $\mathsf{Alt}(p_1)$ a $Y_{21}, \ldots, Y_{2m}$ náhodný výber z $\mathsf{Alt}(p_2)$ (v zmysle binomickej reprezentácie popísanej na konci prvej kapitoly). Označme $X_1 = \sum_{i=1}^{n} Y_{1i}$ a $X_2 = \sum_{i=1}^{m} Y_{2i}$. Porovnávame teda nezávislé binomické veličiny $X_1 \sim \mathsf{Bi}(n,p_1)$ a $X_2 \sim \mathsf{Bi}(m,p_2)$. Zisťujeme či a ako sa líšia pravdepodobnosti p_1 a p_2.

Pravdepodobnosti p_1 a p_2 môžeme konzistentne odhadnúť relatívnymi početnosťami $\hat{p}_1 = X_1/n, \hat{p}_2 = X_2/m$. Pravdepodobnosti $\hat{p}_1$ a $\hat{p}_2$ spravidla porovnávame jedným z týchto spôsobov: môžeme použiť *rozdiel pravdepodobností* $d_X = p_1 - p_2$, prípadne *podiel pravdepodobností* $r_X = p_1/p_2$. Pre nás najzaujímavejší je ale pomer šancí.

Definícia 4 (Pomer šancí)**.** *Pre pravdepodobnosť úspechu p definujeme* šancu* *ako*

$$\frac{p}{1-p}.$$

Na porovnanie pravdepodobností p_1 a p_2 (definovaných ako v predchádzajúcich odsekoch) používame pomer šancí† *definovaný ako*

$$o_X = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

Využitím odhadu relatívnymi početnosťami $\hat{p}_1$ a $\hat{p}_2$ odhadujeme o_X empirickým pomerom šancí‡

$$\hat{o} = \frac{\hat{p}_1(1-\hat{p}_2)}{\hat{p}_2(1-\hat{p}_1)} = \frac{X_1(m-X_2)}{X_2(n-X_1)}.$$

Značením používaným pri kontingenčných tabuľkách dostávame

$$o_X = \frac{p_{11}p_{22}}{p_{12}p_{21}}, \qquad \hat{o} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Motivácia použitia pomeru šancí z hľadiska štvorpoľných tabuliek môže byť ilustrovaná na tomto príklade.

Príklad. Istý muž ochorel na istú chorobu. Vie, že zatiaľ na tú chorobu ochorelo 40 mužov. Niektorí z nich sa na ňu liečili, iní nie; niektorí prežili, iní zomreli. Údaje sú v tabuľke 2.1.

Tento človek by mohol uvažovať takto: ak sa bude liečiť, šanca na prežitie sa dá odhadnúť na 2:3, ak sa liečiť nebude, potom šanca na prežitie bude zhruba 3:7. Vydelením zistí, čo je výhodnejšie. Takto dostane pomer

$$\frac{2:3}{3:7} = \frac{2\cdot 7}{3\cdot 3} = \frac{14}{9}$$

väčší, než 1, a teda môže usúdiť, že výhodnejšie je dať sa liečiť. △

* anglicky *odds* † anglicky *odds ratio* ‡ anglicky *sample odds ratio*

	Prežili	Neprežili	Celkom
Liečení	12	18	30
Neliečení	3	7	10
Celkom	15	25	40

Tabuľka 2.1: Údaje o chorých mužoch

Jednou z výhod pomeru šancí je, že je z neho ľahko poznať nezávislosť pozorovaných veličín, čo dokazuje nasledujúce tvrdenie.

Tvrdenie 4. *V štvorpoľnej tabuľke platí $o_X = 1$ práve vtedy, ak je $p_{ij} = p_{i.}p_{.j}$ pre každú dvojicu (i, j).*

Pri binomickej reprezentácii ak uvažujeme i-tý riadok ako výber z rozdelenia $\mathsf{Bi}(n_i, p_i), i \in \{1, 2\}$, tak sú veličiny nezávislé práve vtedy, keď $p_1 = p_2$.

Dôkaz. Ak platí $p_{ij} = p_{i.}p_{.j}$ pre každú dvojicu (i, j), potom dosadením ihneď dostávame $o_X = 1$.

Nech naopak $o_X = 1$. Ak označíme $p_{11}/p_{12} = \lambda$, potom z toho, že $o_X = 1$ tiež plynie, že $p_{21}/p_{22} = \lambda$. Z toho

$$p_{11} = \lambda p_{12}, \qquad p_{21} = \lambda p_{22}.$$

Príslušná tabuľka pravdepodobností má teda tvar

$$\begin{array}{cc|c} \lambda p_{12} & p_{12} & (\lambda+1)p_{12} \\ \lambda p_{22} & p_{22} & (\lambda+1)p_{22} \\ \hline & & (\lambda+1)p_{.2} \end{array}$$

Vieme teda, že musí platiť $(\lambda + 1)p_{.2} = 1$, z toho $\lambda + 1 = 1/p_{.2}$. Pretože $(\lambda + 1)p_{12} = p_{1.}$ a $(\lambda + 1)p_{22} = p_{2.}$, tak platí $p_{12} = p_{1.}p_{.2}$ a $p_{22} = p_{2.}p_{.2}$. Nakoniec dostávame

$$p_{11} = p_{1.} - p_{12} = p_{1.} - p_{1.}p_{.2} = p_{1.}(1 - p_{.2}) = p_{1.}p_{.1},$$
$$p_{21} = p_{2.} - p_{22} = p_{2.} - p_{2.}p_{.2} = p_{2.}(1 - p_{.2}) = p_{2.}p_{.1}.$$

Nezávislosť veličín potom platí použitím lemmy 1.

Toto využijeme pri časti o binomickej reprezentácii. To, že pomer šancí má byť rovný 1 si vieme prepísať ako

$$p_1(1 - p_2) = p_2(1 - p_1),$$

čo je ekvivalentné s tým, že $p_1 = p_2$.

□

Uvedomme si, že takto definovaný pomer šancí $\hat{o}$ môže nadobúdať hodnoty $0 \leq \hat{o} \leq \infty$, dokonca môže byť nedefinovaný (tento prípad rovnako ako aj rovnosti na krajoch intervalu by boli spôsobené nulovými prvkami v tabuľke). Podľa Anděl (2011) práve nesymetria hodnôt $\hat{o}$ okolo bodu 1 viedla k tomu, že sa takmer výhradne začal používať *logaritmický pomer šancí* d a *teoretický logaritmický pomer šancí* δ definované ako

$$d = \log \hat{o}, \qquad \delta = \log o_X.$$

2.2 Asymptotické rozdelenie

V tejto sekcii sa pozrieme na asymptotické vlastnosti pomeru šancí, resp. jeho logaritmu. Totižto, logaritmická transformácia má aj v tomto prípade navrch; ako uvádza Agresti (2002), táto transformácia konverguje k normálnemu rozdeleniu rýchlejšie, než klasický pomer šancí. Asymptotické rozdelenie popisuje nasledujúca veta; dokazujeme ho pre multinomickú reprezentáciu.

Veta 5. *Veličina*

$$\frac{d-\delta}{\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}}$$

má asymptoticky normálne rozdelenie $\mathsf{N}(0,1)$.

Pred dôkazom tejto vety si najprv dokážeme jedno pomocné lemma.

Lemma 6. *Nech* $X_1, \ldots, X_k$ *je výber z multinomického rozdelenia* $\mathsf{Mult}_k(n,\boldsymbol{p})$, *kde* $\boldsymbol{p} = (p_i)_{i=1}^k$. *Nech* $g(\boldsymbol{p})$ *je diferencovateľná funkcia* $\boldsymbol{p}$ *s napozorovanou hodnotou* $g(\hat{\boldsymbol{p}})$. *Označíme*

$$\phi_i = \frac{\partial g(\boldsymbol{p})}{\partial p_i}, \quad i = 1, \ldots, k.$$

Potom pre $n \to \infty$

$$\sqrt{n}\Big[g(\hat{\boldsymbol{p}}) - g(\boldsymbol{p})\Big] \xrightarrow{d} \mathsf{N}\left(0, \sum_{i=1}^k p_i\phi_i^2 - \left(\sum_{i=1}^k p_i\phi_i\right)^2\right).$$

Dôkaz. Na dôkaz použijeme Δ-metódu (Veta A.4). Predpokladajme, že vektor pozorovaných početností $(n_1, \ldots, n_k)^\top$ má multinomické rozdelenie s príslušnými pravdepodobnosťami $\boldsymbol{p} = (p_1, \ldots, p_k)^\top$. Označme $n = n_1 + \ldots + n_k$ a vektor $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_k)^\top$ relatívnych početností (teda $\hat{p}_i = n_i/n$).

Vektorom $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ik})^\top$ popíšeme i-te pozorovanie ($i \in \{1, \ldots, n\}$) tak, že $Y_{ij} = 1$, ak i-te pozorovanie padlo do kategórie $j \in \{1, \ldots, k\}$, inak $Y_{ij} = 0$. Napríklad, ak vektor $\boldsymbol{Y}_3 = (0, 1, 0, \ldots, 0)^\top$, tak vieme, že tretie pozorovanie padlo do druhej skupiny. Keďže každé pozorovanie padne do práve jednej skupiny, tak vieme, že $\sum_{j=1}^k Y_{ij} = 1$ a $Y_{ij}Y_{il} = 0, j \neq l$. Ďalej $p_j = \sum_i^n Y_{ij}/n$ a

$$\mathsf{E}(Y_{ij}) = \mathsf{P}(Y_{ij} = 1) = \mathsf{E}(Y_{ij}^2), \quad \mathsf{E}(Y_{ij}Y_{il}) = 0,\ j \neq l.$$

Z toho dostaneme, že $\mathsf{E}(\boldsymbol{Y}_i) = \boldsymbol{p}$ a kovariančná matica $\boldsymbol{\Sigma} = \big(\sigma_{jk}\big)_{j,l=1}^k$, kde

$$\begin{aligned}
\sigma_{jj} &= \operatorname{var}(Y_{ij}) = \mathsf{E}(Y_{ij}^2) - \Big[\mathsf{E}(Y_{ij})\Big]^2 = p_j(1-p_j),\\
\sigma_{jl} &= \operatorname{cov}(Y_{ij}, Y_{ik}) = \mathsf{E}(Y_{ij}Y_{ik}) - \mathsf{E}(Y_{ij})\,\mathsf{E}(Y_{il}) = -p_jp_l,\ j \neq l.
\end{aligned}$$

Maticovo

$$\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top,$$

kde $\operatorname{diag}(\boldsymbol{p})$ značí diagonálnu maticu, ktorá s prvkami vektoru $\boldsymbol{p}$ na hlavnej diagonále.

Keďže $\hat{\boldsymbol{p}}$ je výberový priemer n nezávislých pozorovaní $\Big(\hat{\boldsymbol{p}} = (\sum_{i=1}^{n} \boldsymbol{Y}_i)/n\Big)$, tak platí

$$\operatorname{cov}(\hat{\boldsymbol{p}}) = \frac{\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top}{n}.$$

Použitím mnohorozmernej centrálnej limitnej vety (Veta **??**) dostávame

$$\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow{d} \mathsf{N}\Big(\mathbf{0}, \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\Big).$$

Nech ďalej $g(t_1, \ldots, t_k)$ je diferencovateľná funkcia. Označíme

$$\phi_i = \frac{\partial g}{\partial p_i} = \frac{\partial g}{\partial t_i}\Big|_{\boldsymbol{t}=\boldsymbol{p}} \qquad \boldsymbol{\phi} = (\phi_i)_{i=1}^{k}, \qquad i \in \{1, \ldots, k\}.$$

Použitím Δ-metódy (Veta A.4) dostávame

$$\sqrt{n}\Big[g(\hat{\boldsymbol{p}}) - g(\boldsymbol{p})\Big] \xrightarrow{d} \mathsf{N}\Big(0, \boldsymbol{\phi}^\top(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\boldsymbol{\phi}\Big).$$

Asymptotický rozptyl je po úprave rovný

$$\boldsymbol{\phi}^\top \operatorname{diag}\boldsymbol{\phi} - (\boldsymbol{\phi}^\top p)^2 = \sum_{i=1}^{k} p_i\phi_i^2 - \left(\sum_{i=1}^{k} p_i\phi_i\right)^2.$$

□

Dôkaz Vety 6. Vyjdeme z lemmatu 6, kde asymptotický rozptyl označíme σ^2. Tento závisí na p_i. Čo ak by sme ale použili jeho odhad $\hat{\sigma}^2$ založený na relatívnych početnostiach $\hat{p}_i$? Relatívne početnosti konvergujú v pravdepodobnosti k pravdepodobnostiam p_i vďaka zákonu veľkých čísel. Takto bude $\hat{\sigma}$ spojitá funkcia výberových početností, a teda konverguje v pravdepodobnosti k σ vďaka vete o spojitej transformácii (Veta A.1); z toho máme, že $\sigma/\hat{\sigma}$ konverguje v pravdepodobnosti k 1. Teda

$$\sqrt{n}\frac{g(\hat{\boldsymbol{p}}) - g(\boldsymbol{p})}{\hat{\sigma}} = \sqrt{n}\frac{g(\hat{\boldsymbol{p}}) - g(\boldsymbol{p})}{\sigma}\frac{\sigma}{\hat{\sigma}}.$$

Prvá časť členu na pravej strane konverguje v distribúcii k normovanému normálnemu rozdeleniu, zvyšok $(\sigma/\hat{\sigma})$ konverguje v pravdepodobnosti k 1. Použitím Cramérovej-Sluckého vety (Veta A.3) teda dostávame, že celý súčin konverguje v distribúcii k $\mathsf{N}(0,1)$.

Teraz použijeme Δ-metódu (Veta A.4) s $g(\boldsymbol{p}) = \log o_X = \log p_{11} + \log p_{22} - \log p_{12} - \log p_{21}$. Spočítame parciálne derivácie

$$\phi_{11} = \frac{\partial \log o_X}{\partial p_{11}} = \frac{1}{p_{11}}, \quad \phi_{12} = -\frac{1}{p_{12}}, \quad \phi_{21} = -\frac{1}{p_{21}}, \quad \phi_{22} = \frac{1}{p_{22}}.$$

Ďalej

$$\sum_{i=1}^{2}\sum_{j=1}^{2} p_{ij}\phi_{ij} = 0,\ \sigma^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} p_{ij}\phi_{ij}^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{p_{ij}} \text{ a } \sigma^2(d) = \frac{\sigma^2}{n} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{np_{ij}}.$$

Použitím $n\hat{p}_{ij} = n_{ij}$ dostávame tvar asymptotického rozptylu zo znenia vety.

□

3. Stratifikované tabuľky

Dôležitou súčasťou mnohých výskumov je správna voľba parametrov, ktoré sa pozorujú. Keď sledujeme vplyv jednej veličiny na druhú (napr. vplyv X na Y), mali by sme sa snažiť obmedziť (alebo aspoň kontrolovať) vplyv iných parametrov, ktoré by mohli ovplyvňovať vzťah medzi veličinami. Inak riskujeme to, že napozorovaný vplyv X na Y môže v skutočnosti ukazovať vplyv týchto iných parametrov na X a Y.

Na vysvetlenie uvediem príklad prevzatý z knihy (Agresti, 2002). Predstavme si štúdiu sledujúcu pasívne fajčenie. Sledujeme efekty spolužitia s fajčiarom na nefajčiara. To by sme mohli spraviť porovnaním výskytu ochorení dýchacieho traktu medzi nefajčiarmi, ktorí žijú s fajčiarom a tými, ktorí žijú s nefajčiarom. Pri takejto štúdii by sme ale mali takisto sledovať vek, socioekonomický stav, prípadne iné faktory, ktoré by mohli mať vplyv na fajčenie partnera alebo na výskyt ochorení. Bez tohto nemusia byť výsledky veľmi použiteľné. Napríklad ak by boli pozorovaní partneri nefajčiarov mladší, než partneri fajčiarov, tak napozorovanie toho, že medzi partnermi nefajčiarov je menší výskyt ochorení by mohlo byť spôsobené tým, že sú mladší a faktom, že u mladších ľudí je menší výskyt ochorení dýchacieho traktu.

My teda analýzu vzťahu medzi kategorickými veličinami X a Y doplníme kontrolou pre diskrétnu, potenciálne mätúcu veličinu Z.

3.1 Základné pojmy

Kontrolu Z môžeme zabezpečiť tak, že budeme vzťah X a Y sledovať pre pevné Z. Môžeme si to predstaviť tak, že veličiny X a Y sledujeme na niekoľkých vrstvách – stratách. To vysvetľuje názov *stratifikovaná tabuľka*. Predstava toho, že si jednotlivé tabuľky navrstvíme „za seba“ zas vysvetľuje použitie pojmu *trojrozmerná tabuľka**, ktorý napríklad používa aj Prášková (1985). Alternatívne pre X, Y kategoriálne veličiny nadobúdajúce (postupne) I, J hodnôt a K strat (teda hodnôt veličiny Z) ju môžeme nazvať *kontingenčná tabuľka typu* $I \times J \times K$.

My sa v práci zaoberáme *stratifikovanými štvorpoľnými tabuľkami*, teda tabuľkami typu $2 \times 2 \times K$. Nasledujúca definícia zhrňuje súvisiace pojmy.

Definícia 5. *Sledujeme vzťah X a Y na oddelených hodnotách Z.*

Jednotlivé kontingenčné tabuľky sledujúce vzťah medzi X a Y na pevnej hladine veličiny Z nazveme čiastkové tabuľky†. *Tieto tabuľky obsahujú napozorované početnosti, teda sú tvaru*

$$\left(n_{ijk}\right)_{i,\,j\in\{1,2\},\,k\in\{1,\ldots,K\}}.$$

Tabuľku, ktorá vznikne sčítaním prvkov z čiastočných tabuliek na príslušných miestach nazveme marginálnou tabuľkou‡. *Táto tabuľka má tvar*

$$\left(n_{ij.}\right)_{i,\,j\in\{1,2\}},\ kde\ n_{ij.} = \sum_{k=1}^{K} n_{ijk},\ i,j \in \{1,2\}.$$

Marginálna tabuľka neobsahuje žiadnu informáciu o veličine Z.

* anglicky *three-way contingency table* † anglicky *partial tables* ‡ anglicky *marginal table*

Analogicky predchádzajúcemu značeniu by sme mohli definovať marginálne početnosti

$$n_{i..} = \sum_{j=1}^{2}\sum_{k=1}^{K} n_{ijk}, i \in \{1,2\} \quad \text{a} \quad n_{...} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{K} n_{ijk}$$

a k nim analogické $n_{i.k}, n_{.jk}, n_{.j.}, n_{..k}$. Samozrejme platí $n_{...} = n$, kde n je počet všetkých pozorovaní.

Rovnako k tabuľkám zavedieme *podmienené pomery šancí*[§] o_{XYk} pre pevné $k \in \{1, \ldots, K\}$ a *marginálny pomer šancí* o_{XY}. Ak označíme μ_{ijk} očakávané početnosti na pozícii n_{ijk}, tak

$$o_{XYk} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}, \qquad o_{XY} = \frac{\mu_{11.}\mu_{22.}}{\mu_{12.}\mu_{21.}}.$$

Rovnako by sme zaviedli výberové alternatívy $\hat{o}_{XYk}$ a $\hat{o}_{XY}$ založené na napozorovaných početnostiach.

Na koniec tejto sekcie si uvedieme príklad toho, že niekedy nám informácie získané z čiastočných tabuliek môžu dať zdanlivo iný výsledok, než informácia z marginálnej tabuľky. Príklad je rozšírením príkladu zo sekcie 2.1.

Príklad. V tabuľke 2.1 sa nachádzajú údaje o chorých mužoch. My si k nim pridáme ešte podobnú tabuľku 3.1a o 40 chorých ženách a dáta potom sčítame do marginálnej tabuľky.

ŽENY	Prežili	Neprežili	Celkom
Liečené	8	2	10
Neliečené	21	9	30
Celkom	29	11	40

(a) Údaje o chorých ženách

$\sum$	Prežili	Neprežili	Celkom
	20	20	40
	24	16	40
	44	36	80

(b) Marginálna tabuľka

Tabuľka 3.1

Rovnako ako v sekcii 2.1 spočítame odhady pomeru šancí na prežitie. Pre ženy dostaneme 12/7, pre mužov sme spočítali 14/9. Usúdili by sme teda, že pre obidve pohlavia je výhodnejšie liečiť sa. Keď však napočítame pomer šancí na prežitie pre marginálnu tabuľku, dostaneme $2/3 < 1$. Spoločné dáta by sme teda mali interpretovať tak, že je výhodnejšie neliečiť sa.

Tento paradox sa nazýva *Simpsonov paradox*. Dobre ilustruje to, prečo častokrát potrebujeme zaviesť stratifikáciu. △

3.2 Nezávislosť v trojrozmerných tabuľkách

Kvôli pridaní kategórii musíme trocha rozšíriť aj pojem nezávislosti.

Definícia 6. *Veličiny X a Y nazveme* podmienene nezávislé pri danom Z, *ak sú nezávislé pre všetky kategórie. Presnejšie, ak je splnený vzťah (1.3) pre všetky $k \in \{1, \ldots, K\}$, $I = J = 2$.*

§ anglicky *conditional odds ratios*

Marginálna nezávislosť je potom nezávislosť plynúca z dát z marginálnej tabuľky. Nasledujúci príklad ukazuje, že podmienená nezávislosť pre všetky kategórie neznamená marginálnu nezávislosť.

Príklad. Nasledujúca tabuľka zachytáva výsledky založené na pozorovaní z 2 kliník (to sú pre nás kategórie). Na obidvoch boli pacientom s istým typom choroby podávané dva typy liečby (pre každého práve jedna) - `A` a `B`. Tabuľka zachytáva odhady pravdepodobností úspechu liečby založené na pozorovaných početnostiach.

		Výsledok	
Klinika	Liečba	Úspech	Neúspech
1	`A`	0.36	0.24
	`B`	0.24	0.16
2	`A`	0.04	0.16
	`B`	0.16	0.64
Spolu	`A`	0.2	0.2
	`B`	0.2	0.4

Tabuľka 3.2: Údaje o pacientoch

Pre charakterizáciu nezávislosti použijeme Vetu 4. Pre jednotlivé kliniky dostaneme odhady pomerov šancí $\hat{o}_{XY1} = (0.36 \cdot 0.16)/(0.24 \cdot 0.24) = 1$ a $\hat{o}_{XY2} = (0.04 \cdot 0.64)/(0.16 \cdot 0.16) = 1$. Použitá liečba a dosiahnutý výsledok sú teda podmienene nezávislé veličiny. Avšak ak napočítame odhad pomeru šancí z marginálnej tabuľky, dostaneme $\hat{o}_{XY} = (0.2 \cdot 0.4)/(0.2 \cdot 0.2) = 2$. Veličiny teda nie sú marginálne nezávislé. △

Nakoniec ešte rozšírime pojem nezávislosti aj medzi jednotlivé kategórie. Na to sa zavádza pojem homogénnej asociácie.

Definícia 7. *Veličiny X a Y z tabuľky $2 \times 2 \times K$ spĺňajú* homogénnu asociáciu*, ak*

$$o_{XY1} = o_{XY2} = \ldots = o_{XYK}.$$

Homogénna asociácia znamená, že vplyv X na Y je rovnaký pre každú kategóriu zo Z. Podmienená nezávislosť je špeciálny prípad homogénnej asociácie s $o_{XYk} = 1$ pre všetky $k \in \{1, \ldots, K\}$.

4. Testy v stratifikovaných štvorpoľných tabuľkách

V tejto kapitole sa pozrieme na testy podmienených asociácii, ktoré sme si predstavili v tretej kapitole. Zameriame sa na test podmienenej nezávislosti založený na asymptotickom rozdelení logaritmu pomeru šancí, ďalej na Cochran-Mantel-Haenszelov test a na test homogénnej asociácie (tzv. Breslow-Dayov test).

4.1 Test založený na rozdelení pomeru šancí

Vo vete 5 sme odvodili asymptotické rozdelenie logaritmu pomeru šancí. Na základe neho by sme mohli testovať hypotézu podmienenej nezávislosti.

Máme teda hypotézu a alternatívu

$$\begin{aligned} H_0: &\quad o_{XY1} = o_{XY2} = \ldots = o_{XYK} = 1, \\ H_A: &\quad o_{XY1} = o_{XY2} = \ldots = o_{XYK} \neq 1. \end{aligned} \tag{4.1}$$

Teraz určíme vhodnú testovú štatistiku a jej rozdelenie.

Veta 7. *Nech platí H_0. Potom ak $n_{i.k} \to \infty, n_{.jk} \to \infty$ pre všetky $k \in \{1, \ldots, K\}$ a $i, j \in \{1, 2\}$ platí*

$$\chi^2 = \frac{\left(\sum_{k=1}^{K} \log \hat{o}_{XYk}\right)^2}{\sum_{k=1}^{K} \left(\frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}}\right)} \xrightarrow{d} \chi^2_1.$$

Dôkaz. Z vety 5 máme, že pre všetky $k \in \{1, \ldots, K\}$ platí

$$\log \hat{o}_{XYk} \overset{as.}{\sim} \mathsf{N}\left(\log o_{XYk}, \frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}}\right).$$

V našej situácii máme K nezávislých výberov a z nich K pomerov šancí. Z nezávislosti a vlastností normálneho rozdelenia potom máme

$$\sum_{k=1}^{K} \log \hat{o}_{XYk} \overset{as.}{\sim} \mathsf{N}\left(\sum_{k=1}^{K} \log o_{XYk}, \sum_{k=1}^{K} \frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}}\right).$$

Za hypotézy je ale $\sum_{k=1}^{K} \log o_{XYk} = 0$, a teda po klasickej úprave na štandardné normálne rozdelenie máme pre $n_{i.k} \to \infty, n_{.jk} \to \infty$

$$\frac{\sum_{k=1}^{K} \log \hat{o}_{XYk}}{\sqrt{\sum_{k=1}^{K} \frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}}}} \xrightarrow{d} \mathsf{N}(0,1).$$

Z vlastností rozdelenia χ^2 teda poznáme aj rozdelenie druhej mocniny výrazu naľavo od šípky, čo je testová štatistika χ^2.

□

Hypotézu teda zamietame, ak $\chi^2 \geq \chi^2_1(1-\alpha)$.

4.2 Cochranov-Mantelov-Haenszelov test

Teraz sa pozrieme na test, ktorý v roku 1959 navrhli Nathan Mantel a William Haenszel. Opäť testuje podmienenú nezávislosť v tabuľkách $2 \times 2 \times K$.

Mantel a Haenszel zvolili trocha iný prístup. Predmetom ich záujmu bola spätná štúdia údajov o chorobách (názov publikácie je „Statistical aspects of the analysis of data from retrospective studies of disease“). Preto pri svojej práci uvažujú pevné stĺpcové marginálne súčty (tie v tomto prípade zodpovedajú výsledku liečby; pre porovnanie viď tabuľku 3.2). Ich analýza je teda založená na marginálnych súčtoch $n_{.1k}, n_{.2k}, n_{1.k}, n_{2.k}$ v jednotlivých čiastkových tabuľkách. Potom stačí študovať početnosť v ľavom hornom rohu n_{11k}, ostatné sa dajú dopočítať z tejto a marginálnych početností, čo je vidieť z nasledujúcej tabuľky.

$X \backslash Y$	1	0	Spolu
1	n_{11k}	$n_{1.k} - n_{11k}$	$n_{1.k}$
0	$n_{.1k} - n_{11k}$	$n_{.2k} - n_{1.k} + n_{11k}$	$n_{2.k}$
Spolu	$n_{.1k}$	$n_{.2k}$	$n_{..k}$

V takomto prípade má za hypotézy podmienenej nezávislosti veličina n_{11k} hypergeometrické rozdelenie. Toto si teraz zadefinujeme.

Definícia 8. *Pravdepodobnostné rozdelenie definované pravdepodobnosťou*

$$\mathsf{P}(X = k) = \begin{cases} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} & ak\ \max(0, n+K-N) \le k \le \min(K,n) \\ 0 & inak \end{cases}$$

sa nazýva hypergeometrické rozdelenie.

Toto rozdelenie popisuje pravdepodobnosť k úspechov pri n výberoch bez vracania z balíka veľkosti N, ktorý obsahuje presne K objektov, ktorých vytiahnutie považujeme za úspech.

Bez dôkazu si teraz uvedieme tvrdenie o strednej hodnote a rozptylu hypergeometrického rozdelenia.

Tvrdenie 8. *Nech náhodná veličina X má hypergeometrické rozdelenie ako v definícii 8. Potom platí*

$$\mathsf{E}X = n\frac{K}{N}, \qquad var\,X = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}.$$

Dôkaz. (Viď napr. Fleiss a kol., 2003, Sekcia 6.4).

□

V našom prípade $N = n_{..k}, n = n_{1.k}, K = n_{.1k}$.

Chceme testovať podmienenú nezávislosť, teda opäť máme hypotézu (4.1). Zadefinujeme si teda testovú štatistiku a určíme jej asymptotické rozdelenie.

Veta 9. *Nech platí hypotéza (4.1). Potom pre* $n_{..k} \to \infty$ *a* $\lim_{n\to\infty}(n_{.1k}/n_{..k}) = p \in (0,1)$ *platí*

$$CMH = \frac{\left(\sum_{k=1}^{K}(n_{11k} - \frac{n_{1.k}n_{.1k}}{n_{..k}})\right)^2}{\sum_{k=1}^{K}\frac{n_{1.k}n_{2.k}n_{.1k}n_{.2k}}{n_{..k}^2(n_{..k}-1)}} \xrightarrow{d} \chi_1^2.$$

Dôkaz. Porovnávame pozorované početnosti v ľavom hornom rohu s očakávanými početnosťami. Za hypotézy má ale početnosť hypergeometrické rozdelenie, a teda poznáme očakávanú početnosť založenú na strednej hodnote tohto rozloženia.

Podstatou dôkazu je ukázať, že hypergeometrické rozdelenie môžeme (vhodne) aproximovať normálnym rozdelením. To ukážeme tak, že ho aproximujeme vhodným binomickým rozdelením, pre ktoré už centrálnu limitnú vetu (a jeho normálnu aproximáciu poznáme).

Prepíšeme pravdepodobnosť danú hypergeometrickým rozdelením.

$$\begin{aligned}
\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} &= \frac{K!}{k!\cdot(K-k)!}\frac{(N-K)!}{(n-k)!\cdot(N-n-(K-k))!}\cdot\frac{n!\cdot(N-n)!}{N!} \\
&= \binom{n}{k}\cdot\frac{K!/(K-k)!}{N!/(N-k)!}\cdot\frac{(N-K)!\cdot(N-n)!}{(N-k)!\cdot(N-K-(n-k))!} \\
&= \binom{n}{k}\cdot\frac{K!/(K-k)!}{N!/(N-k)!}\cdot\frac{(N-K)!/(N-K-(n-k))!}{(N-n+(n-k))!/(N-n)!} \\
&= \binom{n}{k}\cdot\prod_{m=1}^{k}\frac{K-k+m}{N-k+m}\cdot\prod_{m=1}^{n-k}\frac{N-K-(n-k)+m}{N-n+m}
\end{aligned}$$

Vezmeme limitu tohto člena pre $N \to \infty$ s $\lim_{N\to\infty} K/N = p$ a pevné n a k.

$$\begin{aligned}
\lim_{N\to\infty}\frac{K-k+m}{N-k+m} &= \lim_{N\to\infty}\frac{K}{N} = p \\
\lim_{N\to\infty}\frac{N-K-(n-k)+m}{N-n+m} &= \lim_{N\to\infty}\frac{N-K}{N} = 1-p.
\end{aligned}$$

Takto sme ukázali, že je oprávnené nahradiť hypergeometrické rozdelenie limitným rozdelením $\mathsf{Bi}(n, K/N)$.

Ak by sme boli neopatrní, použili by sme teraz priamo vetu A.5, avšak dostali by sme mierne odlišný tvar. Problémom by bolo to, že na rozdiel od situácie vety A.5 sú naše náhodné veličiny síce rovnako rozdelené, ale nie nezávislé. Pri výpočte celkového rozptylu teda nesmieme zabudnúť na kovarianciu.

Rozptyl jedného pokusu je teda $\frac{K}{N}(1-\frac{K}{N})$. Kovarianciu určíme pomocou toho, že si vieme vyjadriť pravdepodobnosť vytiahnutia dvoch „správnych“ objektov v prvých dvoch ťahoch vďaka tomu, že ide o vyberanie s vracaním, ako $K/N \cdot (K-1)/(N-1)$. Potom kovariancia je

$$\begin{aligned}
\operatorname{cov}(X_1, X_2) &= \mathsf{E}(X_1, X_2) - \mathsf{E}\,X_1\,\mathsf{E}\,X_2 = \mathsf{P}(X_1 = X_2 = 1) - [\mathsf{P}(X_1 = 1)]^2 \\
&= \frac{K(K-1)}{N(N-1)} - \left(\frac{K}{N}\right)^2 = \frac{K\,N\,(K-1) - K^2\,(N-1)}{N^2(N-1)} = \frac{K(K-N)}{N^2(N-1)}
\end{aligned}$$

Celkový rozptyl je potom súčet n jednotlivých rozptylov a $n(n-1)$ kovariancií pre ostatné dvojice. Z toho máme

$$\begin{aligned}\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \operatorname{var} X_i + \sum_{i\neq j} \operatorname{cov}(X_i, X_j) \\ &= n\frac{K}{N}\frac{N-K}{N} + n(n-1)\frac{K(K-N)}{N^2(N-1)} \\ &= n\frac{K(N-K)}{N^2}\left(1 - \frac{n-1}{N-1}\right) = n\frac{K(N-K)}{N^2}\frac{N-n}{N-1}.\end{aligned}$$

Dostávame výsledok ako v tvrdení 8. Zvyšok dôkazu normality je už obdobný dôkazu vety A.5. Z toho máme, že ak náhodná veličina X má hypergeometrické rozdelenie s parametrami N, K, n (ako sme ho zaviedli), tak platí

$$X - n\frac{K}{N} \overset{as.}{\sim} \mathsf{N}\left(0,\, n\,\frac{K}{N}\,\frac{N-K}{N}\,\frac{N-n}{N-1}\right).$$

Ďalej máme K nezávislých výberov. Parametre hypergeometrického rozdelenia nahradíme príslušnými prvkami kontingenčnej tabuľky, teda $N = n_{..k}, n = n_{1.k}, K = n_{.1k}$ a X je uvažovaná náhodná veličina, teda ľavý horný roh n_{11k}. Vďaka nezávislosti a vlastnostiam normálneho rozdelenia dostávame rovnako ako vo vete 7, že

$$\frac{\sum_{k=1}^{K}\left(n_{11k} - n_{1.k}\frac{n_{.1k}}{n_{..k}}\right)}{\sqrt{\sum_{k=1}^{K}\left(\frac{n_{1.k}n_{.1k}(n_{..k}-n_{.1k})(n_{..k}-n_{1.k})}{n_{..k}^2(n_{..k}-1)}\right)}} \overset{as.}{\sim} \mathsf{N}(0,1).$$

Poznáme teda aj limitné rozdelenie druhej mocniny, čo je testová štatistika CMH. Limitným rozdelením je χ^2_1.

□

Hypotézu zamietame, ak $CMH \geq \chi^2_1(1-\alpha)$.

Päť rokov pred článkom autorov Mantela a Haenszela, teda v roku 1954 sa problematikou kombinovania výsledkov z niekoľkých kontingenčných tabuliek zaoberal aj William Cochran v článku „Some Methods for Strengthening the Common χ^2 Tests“. Zvolil iný prístup, a to že početnosti v jednotlivých riadkoch považoval za nezávislé výbery z binomických rozdelení. Tento prístup sme opísali na konci prvej kapitoly. Odvodíme si testovú štatistiku, ktorú dostal.

Veta 10. *Nech platí hypotéza (4.1). Potom pre $n_{1.k} \to \infty$ a $n_{2.k} \to \infty$ platí*

$$CMH_B = \frac{\left(\sum\limits_{k=1}^{K}\left(n_{11k} - \frac{n_{21k}n_{1.k}}{n_{2.k}}\right)\right)^2}{\sum\limits_{k=1}^{K}\frac{n_{1.k}n_{.1k}n_{.2k}}{n_{..k}n_{2.k}}} \xrightarrow{d} \chi^2_1.$$

Dôkaz. Riadky zodpovedajú dvom nezávislým výberom z binomických rozdelení $\mathsf{Bi}(n_{1.k}, p_{1k}), \mathsf{Bi}(n_{2.k}, p_{2k})$. Vďaka vete 4 vieme, že za platnosti hypotézy platí $p_{1k} = p_{2k}$. Pre prvky v prvom stĺpci tabuľky teda môžeme písať

$$n_{11k} \sim \mathsf{Bi}(n_{1.k}, p_k) \quad \text{a} \quad n_{21k} \sim \mathsf{Bi}(n_{2.k}, p_k).$$

Podľa vety A.5 o normálnej aproximácii binomického rozdelenia a vlastností normálneho rozdelenia si napíšeme, čo potom platí pre n_{11k} a n_{21k}.

$$n_{11k} - n_{1.k}p_k \overset{as.}{\sim} \mathsf{N}\Big(0, n_{1.k}p_k(1-p_k)\Big)$$
$$n_{21k} - n_{2.k}p_k \overset{as.}{\sim} \mathsf{N}\Big(0, n_{2.k}p_k(1-p_k)\Big)$$

Výrazy na ľavej strane vydelíme postupne $n_{1.k}$ a $n_{2.k}$ a využijeme to, aké rozdelenie má normálne rozdelená náhodná veličina po vynásobení konštantou. Dostávame

$$\frac{n_{11k}}{n_{1.k}} - p_k \overset{as.}{\sim} \mathsf{N}\left(0, \frac{n_{1.k}p_k(1-p_k)}{n_{1.k}^2}\right)$$
$$\frac{n_{21k}}{n_{2.k}} - p_k \overset{as.}{\sim} \mathsf{N}\left(0, \frac{n_{2.k}p_k(1-p_k)}{n_{2.k}^2}\right).$$

Výrazy na ľavej strane od seba odčítame a využijeme to, ako vyzerá rozdelenie rozdielu dvoch nezávislých normálne rozdelených náhodných veličín. Dostávame tak

$$\frac{n_{11k}}{n_{1.k}} - \frac{n_{21k}}{n_{2.k}} \overset{as.}{\sim} \mathsf{N}\left(0, \frac{p_k(1-p_k)}{n_{1.k}} + \frac{p_k(1-p_k)}{n_{2.k}}\right).$$

Ak nahradíme p_k konzistentným odhadom $\hat{p}_k$, tak podľa Cramérovej-Sluckého vety (Veta A.3) bude asymptotické rozdelenie stále platiť. Použijeme odhad $\hat{p}_k = n_{.1k}/n_{..k}$ a vlastnosti normálneho rozdelenia a dostávame

$$n_{11k} \overset{as.}{\sim} \mathsf{N}\left(\frac{n_{21k}n_{1.k}}{n_{2.k}}, n_{1.k}^2 \frac{n_{.1k}n_{.2k}}{n_{..k}^2} \frac{n_{1.k}+n_{2.k}}{n_{1.k}n_{2.k}}\right)$$

a z toho

$$n_{11k} \overset{as.}{\sim} \mathsf{N}\left(\frac{n_{21k}n_{1.k}}{n_{2.k}}, \frac{n_{1.k}n_{.1k}n_{.2k}}{n_{..k}n_{2.k}}\right).$$

Zavedenie strat ako K nezávislých výberov a následná úprava podľa vlastností normálneho rozdelenia je pre nás už rutina. Vzniknutý výraz umocníme na druhú, čím dostávame výraz

$$\frac{\left(\sum_{k=1}^{K}(n_{11k} - \frac{n_{21k}n_{1.k}}{n_{2.k}})\right)^2}{\sum_{k=1}^{K}\frac{n_{1.k}n_{.1k}n_{.2k}}{n_{..k}n_{2.k}}},$$

čo je testová štatistika CMH_B o ktorej sme takto dokázali, že jej limitným rozdelením je χ_1^2.

□

Pre podobný prístup autorov a rozšírené použitie testu v praxi sa celému výsledku dalo meno Cochranov-Mantelov-Haenszelov test. Ako uvádza Agresti (2002), hypergeometrický prístup Mantela a Haenszela je všeobecnejší, pretože nie všetky praktické pozorovania sa dajú reprezentovať ako dva nezávislé výbery z binomického rozdelenia. Pre zaujímavosť ešte spomeniem, že Mantel a Haenszel pôvodne navrhli testovú štatistiku s korekciou spojitosti, tej sa ale v tejto práci nebudeme venovať. Výhodou Cochran-Mantel-Haenszelovho testu je, že existuje rozšírenie pre tabuľky typu $I \times J \times K$. To napríklad neplatí pre Breslowov-Dayov test, ktorý si teraz predstavíme.

4.3 Breslowov-Dayov test

Na záver ešte popíšeme test homogénnej asociácie, teda test hypotézy

$$H_0 : \ o_{XY1} = \ldots = o_{XYk}$$

známy ako Breslowov-Dayov test. Tento test navrhli Breslow a Day (1980). Opäť pri ňom uvažujeme pevné marginálne súčty a analyzujeme ľavé horné políčko n_{11k}. Ak neplatí H_0, tak existuje nejaká kategória, pre ktorú je podmienený pomer šancí výrazne väčší alebo menší, než ostatné pomery šancí. V takomto prípade bude pozorovaná početnosť n_{11k} väčšia alebo menšia, než očakávaná hodnota založená na odhadnutom spoločnom pomere šancí.

Prvým návrhom bol χ^2 test, ktorý by počítal testovú štatistiku ako

$$\sum_{k=1}^{K} \frac{(n_{11k} - \mu_{11k})^2}{\operatorname{var} n_{11}};$$

μ_{11k} značíme očakávanú početnosť na pozícii (1,1). Za predpokladu dostatočného počtu pozorovaní a hypotézy by táto štatistika mala rozdelenie χ^2_{K-1}.

Autori však uvádzajú, že takýto test nie je veľmi užitočný, pretože pre veľký počet kategórií nemusí rozdelenie testovej štatistiky aproximovať rozdelenie χ^2 ani za hypotézy. Preto navrhujú alternatívny postup.

V ňom si K kategórií rozdelíme do H skupín $I_h, h \in \{1, \ldots, H\}$ takých, že podmienené pomery šancí sú rovnaké vrámci skupín, nie však medzi nimi. Testová štatistika je potom

$$\sum_{h=1}^{H} \frac{\left(\sum_{k \in I_h} n_{11k} - \mu_{11k}\right)^2}{\sum_{k \in I_h} \operatorname{var} n_{11k}}.$$

Táto má za hypotézy rozdelenie χ^2_{H-1} a podľa autorov má test založený na tejto testovej štatistike väčšiu silu, než vyššie navrhnutá varianta.

Otázkou ešte ostáva aký spoločný pomer šancí použiť pri výpočte μ_{11k}. Autori navrhujú buď odhad založený na maximálnej vierohodnosti, alebo odhad $\hat{o}_{MH}$ navrhnutý Mantelom a Haenszelom

$$\hat{o}_{MH} = \frac{\sum_{k=1}^{K} \frac{n_{11k} n_{22k}}{n_{..k}}}{\sum_{k=1}^{K} \frac{n_{12k} n_{21k}}{n_{..k}}},$$

čo je vážený priemer všetkých k pomerov šancí.

5. Simulácia

V poslednej kapitole sa pozrieme na vlastnosti testov odvodených v štvrtej kapitole z numerického hľadiska. Budeme pracovať s konečným výberom. Najprv si ukážeme, ako testové štatistiky dodržiavajú stanovenú hladinu pre rôzne rozsahy výberov a parametre rozdelenia, z ktorého vyberáme. Potom porovnáme empirické distribučné funkcie našich testových štatistík s distribučnou funkciou rozdelenia χ^2_1.

$(X,Y)^\top$ je teda opäť dvojrozmerný diskrétne rozdelený náhodný vektor s hodnotami v množine $\{0,1\}^2$; my máme niekoľko (K) náhodných výberov $(X_1,Y_1)^\top$, $\ldots$, $(X_{2n},Y_{2n})^\top$ z rozdelenia tohto vektoru (prečo je posledný index práve $2n$ bude jasné z popisu simulácie). Test založený na rozdelení pomeru šancí aj Cochranov-Mantelov-Haenszelov test, ktoré sme si odvodili v kapitole 4 testujú hypotézu podmienenej nezávislosti.

Pre simuláciu budeme uvažovať štvorpoľnú tabuľku reprezentovanú dvoma nezávislými výbermi z binomického rozdelenia. Vezmeme si pevné marginálne riadkové početnosti $n_{1.k}$ a $n_{2.k}$, tieto budeme uvažovať rovnaké a označíme ich n. Pre celú simuláciu budeme uvažovať pevný počet strat, a to $K = 5$. Pre každú z 5 tabuliek nagenerujeme pomocou softvéru prvky v prvom stĺpci n_{11} a n_{21} z binomických rozdelení $\mathsf{Bi}(n, p_{1k})$ a $\mathsf{Bi}(n, p_{2k})$ (postupne). Početnosti v druhom stĺpci určíme tak, aby bol marginálny súčet v riadku rovný n. Celkovo takto máme 5 výberov o rozsahu $2n$.

Vďaka tvrdeniu 4 vieme, že za platnosti hypotézy podmienenej nezávislosti musí (pri binomickej reprezentácii) platiť $p_{1k} = p_{2k}$ pre všetky $k \in \{1, \ldots, K\}$. My budeme pri simulácii uvažovať tieto pravdepodobnosti rovnaké pre všetky k, túto pravdepodobnosť označíme p.

Zvolíme hladinu $\alpha = 0.05$. Budeme postupne uvažovať rozsahy výberov určené $n = 10, 20, 50, 100, 50, 500$ a pravdepodobnosti $p = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$. Pre každú kombináciu n a p nagenerujeme 100 000 sád piatich tabuliek a pre každú sadu spočítame testové štatistiky (χ^2 aj CMH). V nasledujúcich dvoch tabuľkách sú zhrnuté dosiahnuté hladiny testov pre všetky kombinácie n a p. Túto hladinu sme určili ako podiel počtu prípadov, kedy test zamietol platnú hypotézu a počtu všetkých prípadov (teda 100 000).

	p					
n	0.05	0.1	0.2	0.3	0.4	0.5
10	0.0456	0.0505	0.0495	0.0496	0.0524	0.0554
20	0.048	0.0506	0.0495	0.0493	0.0518	0.0562
50	0.0493	0.0502	0.0520	0.0491	0.0498	0.0542
100	0.0496	0.0488	0.0499	0.0499	0.0491	0.0497
500	0.0515	0.0510	0.0509	0.0501	0.0498	0.0494

Tabuľka 5.1: Dosiahnutá hladina Cochran-Mantel-Haenszelovho testu

Vidíme, že test ani pre malé hodnoty p a n nie je test príliš konzervatívny.

V nasledujúcej tabuľke sú uvedené dosiahnuté hladiny pre test založený na logaritme pomeru šancí.

				p		
n	0.05	0.1	0.2	0.3	0.4	0.5
10	0	0	0.0038	0.0256	0.0464	0.0544
20	0	0.0023	0.0365	0.0522	0.0550	0.0577
50	0.0059	0.0419	0.0546	0.0514	0.0521	0.0545
100	0.0409	0.0518	0.0516	0.0510	0.0505	0.0534
500	0.0503	0.0507	0.0512	0.0502	0.0495	0.0494

Tabuľka 5.2: Dosiahnutá hladina testu založeného na $\log \hat{o}_{XYk}$

Môžeme vidieť, že tento test je pre malé hodnoty n a p nepoužiteľný. Problémom je, že pre tieto hodnoty sa často stane, že početnosť v niektorom poli tabuľky je rovná 0, čo nesmie nastať, ak uvažujeme pomer šancí (resp. jeho logaritmus). Testovú štatistiku potom nevieme vyjadriť ako číslo a zdanlivo dostaneme príliš konzervatívny test. To je spôsobené tým, že pri porovnávaní s kvantilom rozdelenia χ^2 nám naše porovnanie ($\mathsf{NaN} \overset{?}{\geq} \chi^2_1(1-\alpha)$) v takomto prípade nepovie, že by sme mali hypotézu zamietnuť. Pre dostatočné veľké hodnoty p a n sa dosiahnutá hladina blíži k požadovanej hladine, podobne ako aj pri Cochran-Mantel-Haenszelovom teste.

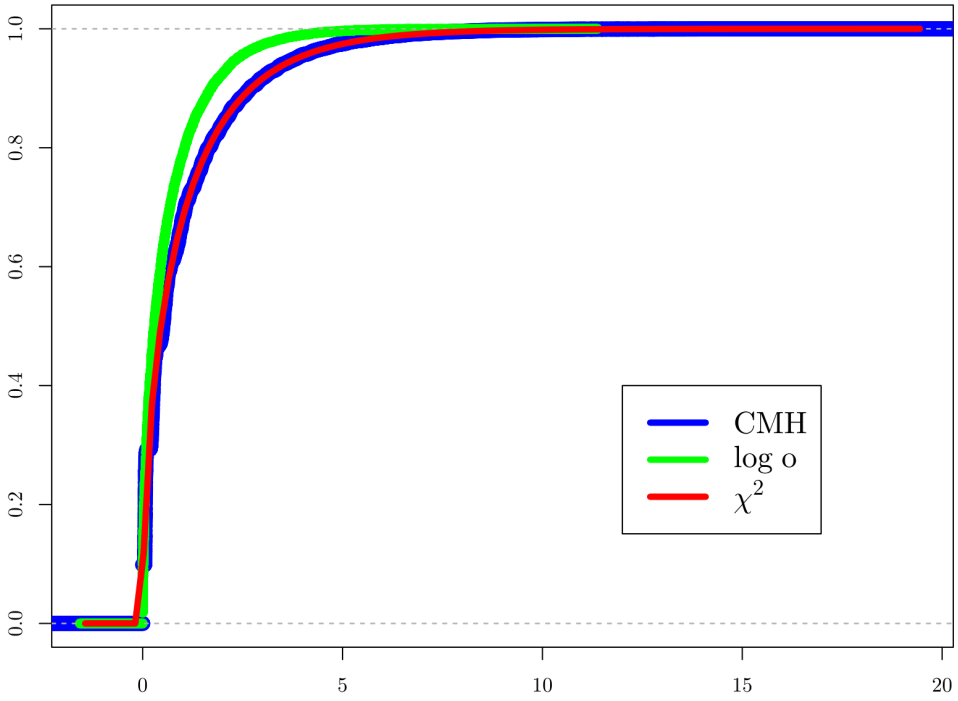
Na záver sa pozrieme na konvergenciu testových štatistík k asymptotickému rozdeleniu χ^2_1. To spravíme tak, že na niekoľkých grafoch zobrazíme pre vybrané kombinácie n a p empirické distribučné funkcie (e.d.f.) našich testových štatistík a distribučnú funkciu (d.f.) rozdelenia χ^2_1. Distribučná funkcia rozdelenia χ^2_1 je v každom grafe znázornená červenou krivkou. Všetky údaje vznikli z nagenerovania 100 000 sád piatich tabuliek.

Najprv si ukážeme, porovnanie obidvoch testov. Dostaneme podobný výsledok, ako pri hladine testu, a to že pre malé n je e.d.f. testovej štatistiky χ^2 vzdialenejšia od d.f. rozdelenia χ^2_1 než e.d.f. štatistiky CMH. Pre dostatočne veľké n tento rozdiel už nie je badateľný. Obrázok 5.1 zobrazuje empirické distribučné funkcie testových štatistík pre $p = 0.2$ a n najprv 10 (Obr. 5.1a) a potom 100 (Obr. 5.1b).

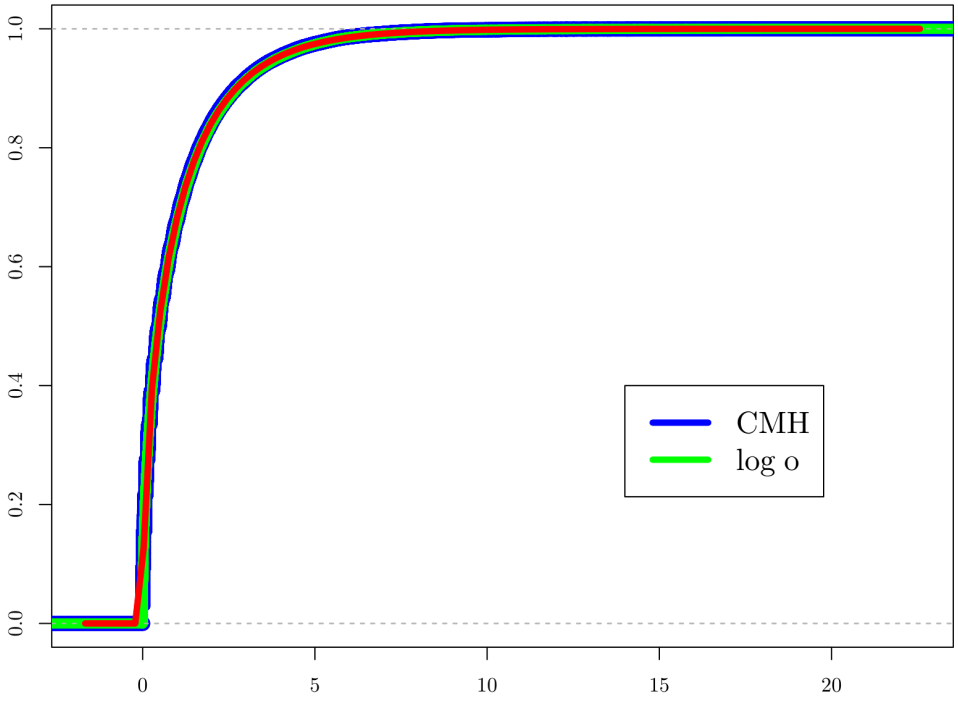
Na obrázku 5.2 porovnávame konvergenciu e.d.f. oboch štatistík pre malú veľkosť výberu ($n = 10$). Vidíme, že e.d.f. testu s testovou štatistikou χ^2 je pre malé p veľmi rozdielna od d.f. rozdelenia χ^2. CMH test zvláda dobre aj túto situáciu.

Obrázok 5.3 zobrazuje opačnú situáciu, a to že máme malé $p = 0.05$ a sledujeme vplyv zmeny veľkosti výberu. CMH test opäť situáciu zvláda aj pre malé výbery. Test s testovou štatistikou χ^2 v tomto dopadol ešte horšie, ako v predchádzajúcej časti, pretože okrem prípadu $n = 500$ je rozdiel distribučných funkcií veľký.

Úplne nakoniec nám obrázok 5.4 ukazuje, že pre dostatočne veľké p (v tomto prípade 0.3) konvergujú e.d.f. testových štatistík ku d.f. rozdelenia χ^2 spoľahlivo aj pre malé veľkosti výberu.

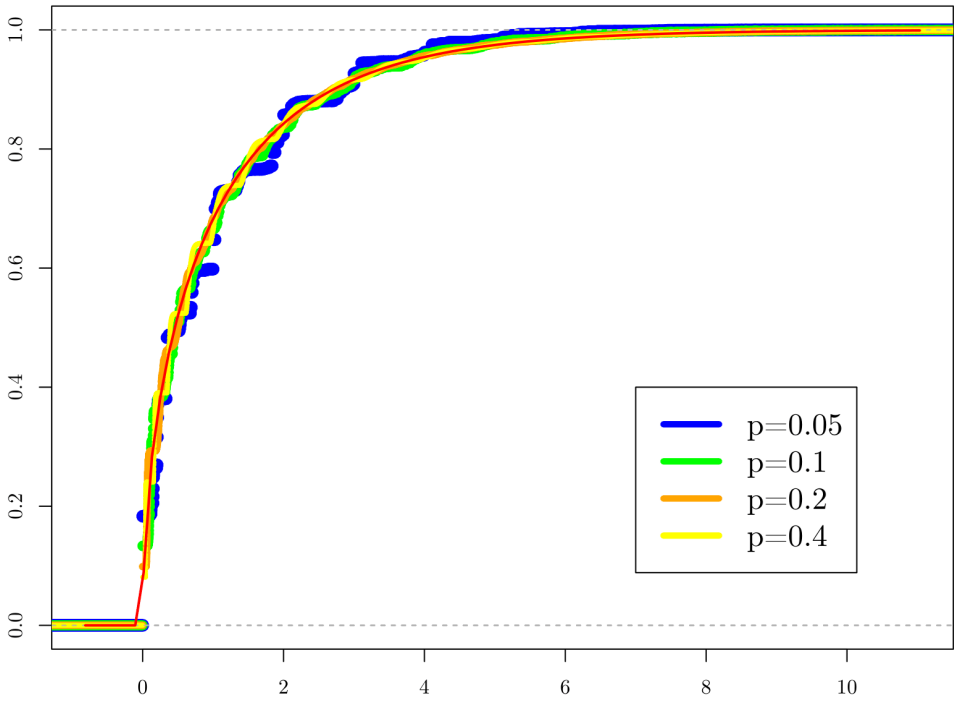


(a) $p = 0.2, n = 10$

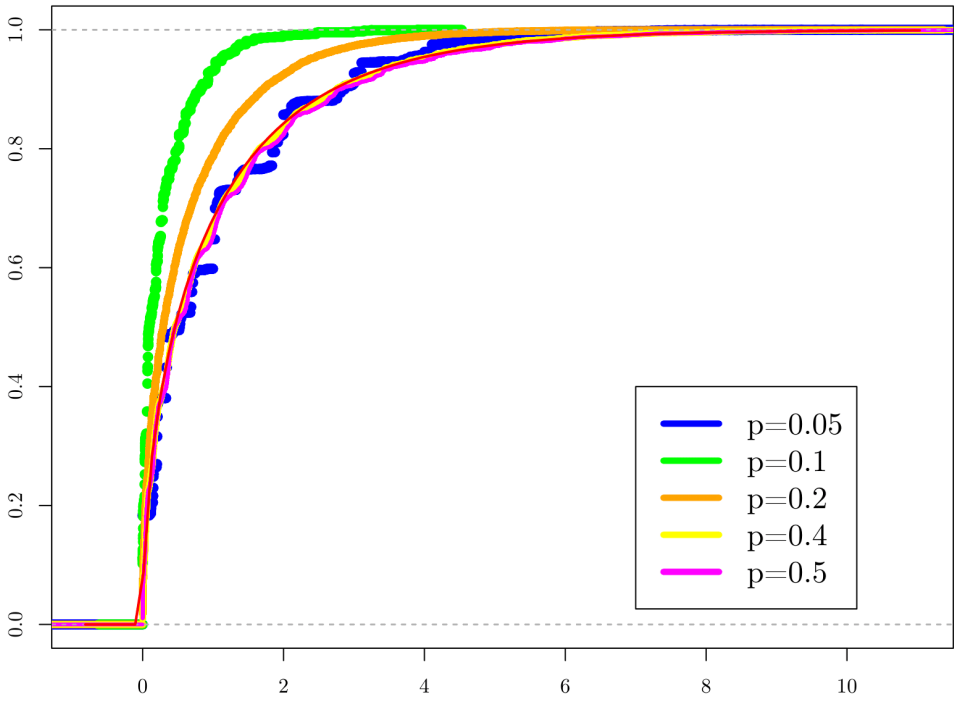


(b) $p = 0.2, n = 100$

Obr. 5.1: Porovnanie konvergencie e.d.f. obidvoch testových štatistík pre rôzne veľkosti výberov

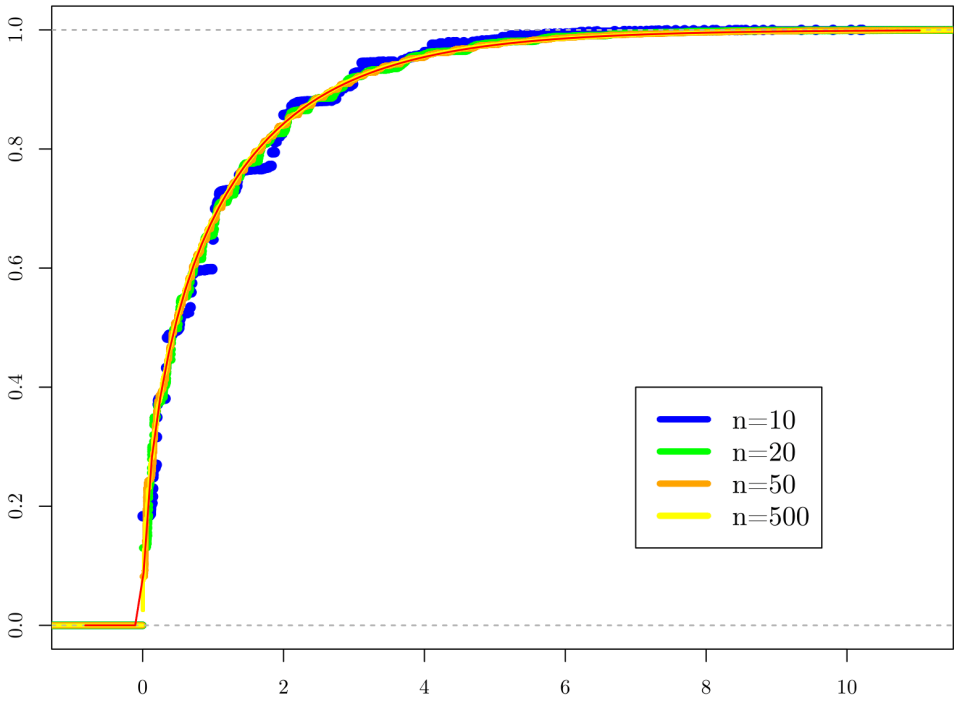


(a) $CMH, n = 10$

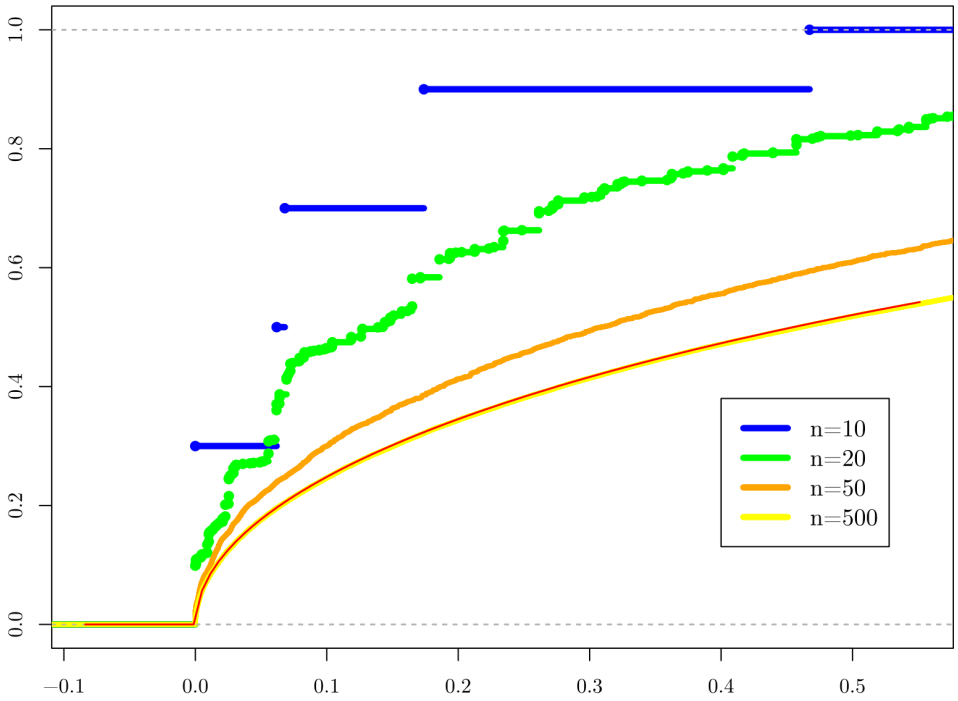


(b) $\chi^2, n = 10$

Obr. 5.2: Porovnanie konvergencie e.d.f. testových štatistík pre malé n a rôzne p

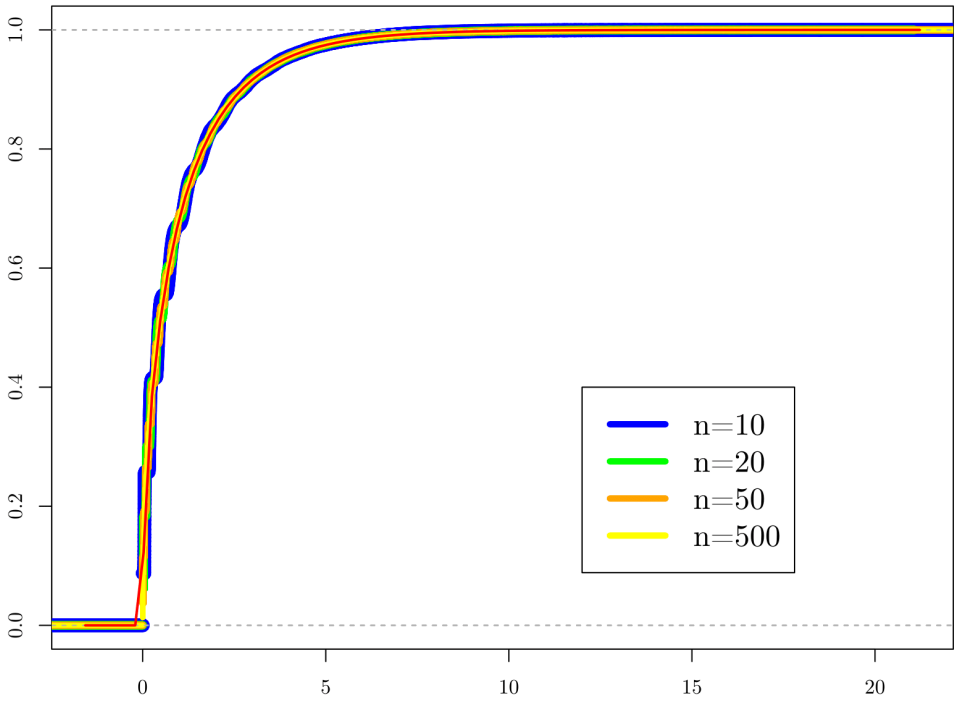


(a) $CMH, p = 0.05$

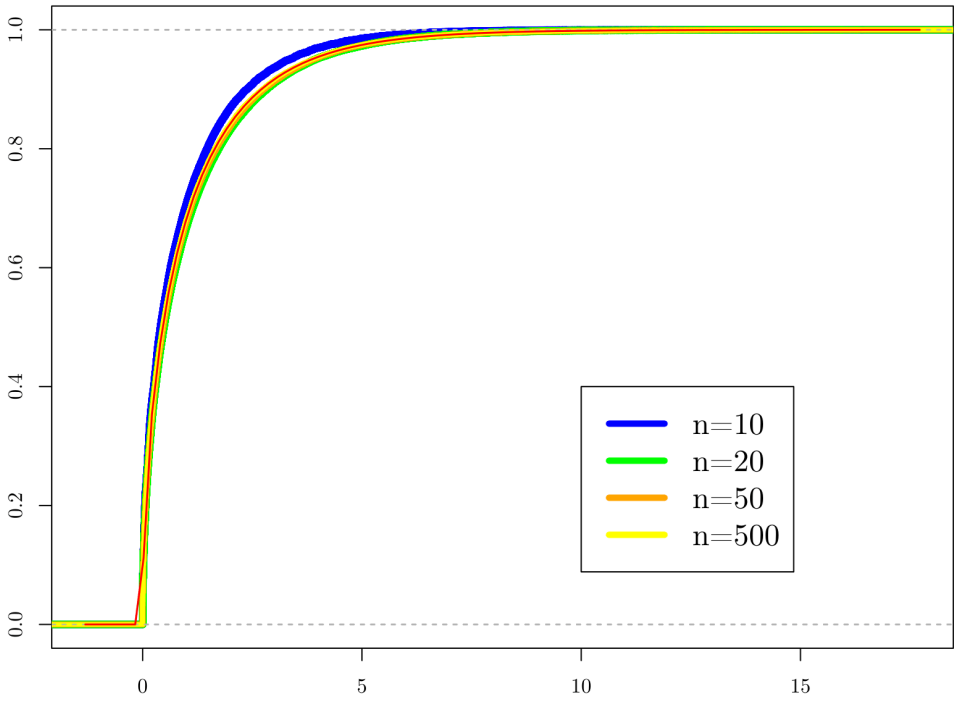


(b) $\chi^2, p = 0.05$

Obr. 5.3: Porovnanie konvergencie e.d.f. testových štatistík pre malé p a rôzne n



(a) $CMH, p = 0.3$



(b) $\chi^2, p = 0.3$

Obr. 5.4: Porovnanie konvergencie e.d.f. štatistík pre pevné p a rôzne n

Záver

V práci sme sa zaoberali testami stratifikovaných kategoriálnych dát. Pre poriadok bolo dôležité to, že sme si precízne zaviedli značenie a pojmy, na ktorých sme potom odvodzovali príslušné testy.

Dôležitou časťou bolo zoznámenie sa s pomerom šancí, ktorý nám jednak zjednodušil skúmanie vzťahu medzi sledovanými veličinami, ale takisto sme na základe jeho asymptotického rozdelenia odvodili jeden z testov.

V celej práci sme sa postupne stretli s troma reprezentáciami štvorpoľných tabuliek – najprv s multinomickou, ktorá je dôležitá pre kontingenčné tabuľky ľubovoľných rozmerov, ďalej s binomickou s pevnými riadkovými početnosťami, ktorá nám pomohla pri odvodení testovej štatistiky, s akou pracoval Cochran a s hypergeometrickou, ktorú použili Mantel a Haenszel pre odvodenie dodnes najpoužívanejšieho testu na podmienenú nezávislosť v kontingenčných tabuľkách.

Najdôležitejšou časťou práce sú teda vety 9 a 10, kde sme dokázali asymptotické rozdelenie testových štatistík. V tejto kapitole sme vychádzali predovšetkým z literatúry autorov Agresti (2002) a Fleiss a kol. (2003).

V poslednej časti sme pomocou simulácie ukázali, že odvodené testy nie sú nesprávne a fungujú spoľahlivo. Takisto sme ukázali, že Cochranov-Mantelov-Haenszelov test funguje na malé početnosti lepšie, než nami odvodený test založený na rozdelení pomeru šancí (napriek tomu, že v literatúre sa nachádzajú rôzne odporúčania aké početnosti musíme mať na to, aby bolo použitie testu oprávnené). So zväčšujúcim sa rozsahom výberu sme pre obidva testy dosiahli dobré výsledky. Konvergencia testových štatistík je jasne viditeľná z grafického porovnania ich empirických distribučných funkcií s distribučnou funkciou rozdelenia χ^2_1.

Prehľad používaných viet

V tejto kapitole uvedieme základné vety, ktoré používame pri práci, aj s odkazom na literatúru, kde sú vety dokázané.

Veta A.1 (Veta o spojitej transformácii)**.** *Ak je g spojitá reálna funkcia a $X_n \xrightarrow{\mathsf{P}} X$, potom $g(X_n) \xrightarrow{\mathsf{P}} g(X)$.*

Dôkaz. (Viď Anděl, 2011, Veta B.9).

□

Veta A.2 (Lindebergova CLV)**.** *Nech X_1,X_2,... sú nezávislé rovnako rozdelené náhodné veličiny so strednou hodnotou μ a s konečným rozptylom σ^2. Potom pre $n \to \infty$ platí*

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

Dôkaz. (Viď Anděl, 1985, Veta X.6).

□

Veta A.3 (Cramérova-Sluckého)**.** *Nech X_1,X_2,... je postupnosť náhodných veličín s distribučnými funkciami F_1,F_2,.... Nech F je distribučná funkcia a c konštanta. Nech F_n konvergujú v distribúcii k F ($F_n \xrightarrow{d} F$). Nech Y_1,Y_2,... je taká postupnosť náhodných veličín, že $Y_n \xrightarrow{\mathsf{P}} c$. Definujme*

$$R_n = X_n + Y_n, \quad S_n = X_n Y_n, \quad T_n = \frac{X_n}{Y_n}.$$

Nech F_n^R,F_n^S a F_n^T sú (v tomto poradí) distribučné funkcie veličín R_n,S_n a T_n. Potom $F_n^R(x)$ konvergujú slabo k $F(x-c)$. Ak je $c > 0$, potom $F_n^S(x)$ konvergujú v distribúcii k $F\left(\frac{x}{c}\right)$ a $F_n^T(x)$ konvergujú v distribúcii k $F(cx)$.

Dôkaz. (Viď Anděl, 2011, Veta B.10).

□

Veta A.4 (Δ-metóda)**.** *Nech postupnosť náhodných vektorov $\{\boldsymbol{T}_n\}_{n=1}^{\infty}$ spĺňa*

$$\sqrt{n}(\boldsymbol{T}_n - \mu) \xrightarrow[n\to\infty]{d} \mathsf{N}_k(0,\boldsymbol{\Sigma})$$

pre nejaký vektor konštánt $\boldsymbol{\mu} \in \mathbb{R}^k$ a maticu $\boldsymbol{\Sigma}$. Nech $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^p$ je funkcia, ktorá je spojite diferencovateľná v nejakom okolí bodu $\boldsymbol{\mu}$. Označme $\mathbb{D}(\boldsymbol{x}) = \frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}}$. Potom platí

$$\sqrt{n}\left(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\mu})\right) \xrightarrow{d} \mathsf{N}_p(\mathbf{0},\mathbb{D}(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbb{D}(\boldsymbol{\mu})^\top).$$

Dôkaz. (Viď Agresti, 2002, 14.1.5).

□

Veta A.5 (CLV pre binomické rozdelenie)**.** *Nech* $X_1, \ldots, X_n$ *je postupnosť nezávislých rovnako rozdelených náhodných veličín z binomického rozdelenia* $\mathsf{Bi}(n,p)$*,* $p \in (0,1)$*. Potom pre* $n \to \infty$ *platí*

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathsf{N}(0,1).$$

Dôkaz. (Viď Anděl, 2011, Veta B.12).

□

Zoznam použitej literatúry

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition. Wiley, Gainesville, Florida. ISBN 0-471-36093-7.

Anděl, J. (1985). *Matematická statistika*. Druhé vydání. SNTL - Nakladatelství technické literatúry, Praha.

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

Breslow, N. E. a Day, N. E. (1980). *Statistical Methods in Cancer Research*, volume 1 of *First Edition*. International Agency for Research on Cancer, Lyon. ISBN 92-832-0132-9.

Fleiss, J. L., Levin, B. a Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. Third Edition. Wiley, Columbia University, New York. ISBN 0-471-52629-0.

Prášková, Z. (1985). *Kontingenční tabulky*. 1. vydání. Univerzita Karlova, Praha.