Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Kateřina Poláchová

Poměr šancí v kontingenčních tabulkách

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Petr Klášterecký
Studijní program: Matematika, obecná matematika

2007

Poděkování:
Na tomto místě bych chtěla poděkovat svému vedoucímu Mgr. Petrovi Klášterecképu za odborné vedení, trpělivost a cenné rady k danému tématu.

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

V Praze dne 28.5.2007 Kateřina Poláchová

Obsah

Název práce: Poměr šancí v kontingenčních tabulkách
Autor: Kateřina Poláchová
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Petr Klášterecký
e-mail vedoucího: petr.klasterecky@matfyz.cz

Abstrakt: Poměr šancí a jeho logaritmus jsou v praxi velmi často používané statistiky pro měření závislostí mezi kategoriálními veličinami. V této práci se zabýváme teoretickými vlastnostmi poměru šancí v kontingenčních tabulkách, odhadováním poměru šancí z náhodného výběru a v neposlední řadě jeho využitím v testování hypotézy o nezávislosti. Okrajově jsou v práci zmíněny další míry závislosti v kontingenčních tabulkách a jejich vztah k poměru šancí.

Klíčová slova: Kategoriální veličina, kontingenční tabulka, poměr šancí.

Title: Odds ratio in contingency tables
Author: Kateřina Poláchová
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Petr Klášterecký
Supervisor's e-mail address: petr.klasterecky@matfyz.cz

Abstract: The odds ratio and his logarithm are ones of the most used statistics for measuring of dependencies between categorical variables. The core of this work are theoretical properties of the odds ratio in contingency tables, estimators of odds ratio from random sample, and it's use in independence hypothesis testing. The work also roughly sketches other dependency measures in contingency tables and their relation to the odds ratio.

Keywords: Categorical variable, contingency table, odds ratio.

Kapitola 1

Úvod

V této práci se budeme zabývat kategoriálními veličinami a zkoumáním a testováním asociací mezi nimi. Ve druhé kapitole popíšeme tabulky, pomocí kterých se zobrazují vztahy mezi kategoriálními veličinami. Dále se budeme věnovat mírám závislostí, převážně poměru šancí a jeho vlastnostem v dvoj- a trojrozměrných tabulkách. Ve třetí kapitole se zaměříme na odhady poměru šancí a na odvození intervalového odhadu pomocí delta metody. Samotná delta metoda je uvedena v Dodatku. Čtvrtá kapitola je věnována testování hypotéz o nezávilosti v dvojrozměrných kontingenčních tabulkách.

1.1 Kategoriální veličiny

Kategoriální veličina je diskrétní náhodná veličina, která může nabývat jen konečně mnoha hodnot, obvykle malého počtu. Tyto hodnoty se nazývají kategorie veličiny. Příkladem kategoriální veličiny mohou být povahy lidí s kategoriemi cholerik, sangvinik, melancholik nebo flegmatik.

Rozlišujeme dva základní typy kategoriálních veličin. Veličina, jejíž hodnoty nemají přirozené uspořádání, se nazývá *nominální*. To znamená, že nezávisí na pořadí, ve kterém volíme jednotlivé kategorie. Je to například náboženské vyznání s prvky křesťanství, islám, budhismus a jiné. O dvou hodnotách nominální proměnné můžeme říct, že jsou buď stejné nebo různé.

Naopak veličina, pro kterou platí, že její kategorie mají jednoznačně určené pořadí, se nazývá *ordinální*. Taková veličina je například známka ze zkoušky na vysoké škole s hodnotami výborně, velmi dobře, dobře a neprospěl(a).

Podle toho, jak na veličinu pohlížíme, určujeme její klasifikaci. Mějme kategoriální veličinu vzdělání. Pokud máme kategorie státní nebo soukromá škola, jedná se o nominální veličinu. V případě, že nás zajímá stupeň vzdělání, tedy základní, střední, vyšší a vysoké, veličina je ordinální.

Kapitola 2

Měření závislostí v kontingenčních tabulkách

2.1 Kontingenční tabulky

Definice 2.1. *Nechť $(X_1, X_2, \ldots, X_n)$ je náhodný vektor, který má diskrétní rozdělení. Nechť X_j jsou kategoriální veličiny, které mají I_j kategorií, $j = 1, 2, \ldots, n$. Tím získáme $I_1 I_2 \ldots I_n$ možných kombinací hodnot všech veličin, které můžeme umístit do n-rozměrné tabulky. Tabulka takového typu, ve které jsou v jednotlivých polích obsaženy počty možných výstupů, se nazývá kontingenční tabulka.*

V následujícím textu se omezíme na tabulky, které sledují dvě nebo tři veličiny, neboli na *dvojrozměrné* a *trojrozměrné* kontingenční tabulky. Nejprve popíšeme pravděpodobnostní strukturu dvojrozměrných tabulek. Mějme tedy veličiny X a Y s počtem kategorií I a J. Tabulka má nyní tvar matice typu $I \times J$.

Nechť π_{ij} označuje pravděpodobnost, že se vektor (X, Y) nachází v poli tabulky v řádku i a ve sloupci j, tedy

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

Pravděpodobnostní rozdělení $\{\pi_{ij}\}$ je *sdružené rozdělení* X a Y. Označme dále

$$\pi_{i.} = \sum_j \pi_{ij},$$

$$\pi_{.j} = \sum_i \pi_{ij}.$$

Čísla $\pi_{i.}$ a $\pi_{.j}$ se nazývají *marginální pravděpodobnosti*. Tedy, $\{\pi_{i.}\}$ je *marginální rozdělení* veličiny X a $\{\pi_{.j}\}$ je *marginální rozdělení* veličiny Y. Zřejmě platí

$$\sum_i \pi_{i.} = \sum_j \pi_{.j} = \sum_i \sum_j \pi_{ij} = 1.$$

Tabulka 2.1: Pravděpodobnostní struktura kontingenční tabulky typu 2×2

	Y		
X	1	2	$\sum$
1	π_{11}	π_{12}	$\pi_{1.}$
	$(\pi_{1\|1})$	$(\pi_{2\|1})$	(1.0)
2	π_{21}	π_{22}	$\pi_{2.}$
	$(\pi_{1\|2})$	$(\pi_{2\|2})$	(1.0)
$\sum$	$\pi_{.1}$	$\pi_{.2}$	1.0

Pravděpodobnostní struktura kontingenční tabulky typu 2×2 je uvedena v tabulce 2.1.

V některých kontingenčních tabulkách je jedna veličina vysvětlovaná, nechť je to Y, a druhá veličina (X) je vysvětlující. V tomto případě se nabízí určit samostatné pravděpodobnostní rozdělení veličiny Y na každé úrovni veličiny X. Neboli, zajímá nás, jak se mění rozdělení Y při změnách kategorií X. Mějme objekt, který se nachází v řádku i, tedy v *i-té* kategorii veličiny X. Pravděpodobnost, že se objekt nachází ve sloupci j, tedy v *j-té* kategorii veličiny Y, se označuje $\pi_{j|i}$, $j = 1, ..., J$. Pravděpodobnosti $\{\pi_{1|i}, ..., \pi_{J|i}\}$ tvoří *podmíněné rozdělení* Y v *i-té* kategorii X.

2.2 Vztahy mezi kategoriálními veličinami

Závislost mezi kategoriálními veličinami můžeme popsat buď pomocí sdruženého rozdělení nebo pomocí podmíněného rozdělení Y podle X nebo X podle Y. Vztah mezi sdruženým rozdělením a podmíněným rozdělením veličiny Y podle X se dá podle definice podmíněné pravděpodobnosti vyjádřit pomocí vzorce

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i.}} \qquad \text{pro všechna } i, j.$$

Definice 2.2. *Řekneme, že dvě kategoriální veličiny jsou nezávislé, právě když platí*

$$\pi_{ij} = \pi_{i.}\pi_{.j} \qquad \textit{pro všechna } i = 1, \ldots, I,\ j = 1, \ldots, J. \tag{2.1}$$

Poznámka 2.1. *Jestliže jsou veličiny X a Y nezávislé, pak*

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i.}} = \frac{(\pi_{i.}\pi_{.j})}{\pi_{i.}} = \pi_{.j}, \qquad \textit{pro všechna } i = 1, \ldots, I.$$

To znamená, že podmíněné pravděpodobnosti veličiny Y jsou rovny marginálním pravděpodobnostem Y.

Obdobou nezávislosti je tzv. *homogenita* rozdělení, kterou můžeme popsat jako

$$\{\pi_{j|1} = \cdots = \pi_{j|I},\ \textit{pro } j = 1, \ldots, J\}.$$

Homogenita nastává, pokud pravděpodobnost, že se objekt nachází v *j-tém* sloupci, je ve všech řádcích tabulky stejná, neboli pravděpodobnosti nezávisejí na řádkovém indexu i. To znamená, že všechny řádky tabulky jsou stejné.

V další části kapitoly se zaměříme na binární veličiny, neboli veličiny, které mohou nabývat pouze dvou hodnot. Předpokládejme, že Y je veličina s kategoriemi úspěch a neúspěch. Mějme dále dvě skupiny objektů, které tvoří kategorie veličiny X. Data můžeme zobrazit v kontingenční tabulce typu 2×2, přičemž skupiny objektů tvoří řádky tabulky a sloupce jsou kategorie veličiny Y.

2.3 Rozdíl pravděpodobností

Vezměme objekt nacházející se v řádku i. Pravděpodobnost, že výsledek bude ležet v kategorii úspěch, je $\pi_{1|i}$. Jedná se tedy o pravděpodobnost úspěchu. Naopak pravděpodobnost neúspěchu $\pi_{2|i}$ má tvar $\pi_{2|i} = 1 - \pi_{1|i}$. Pro jednoduchost zavedeme značení π_i místo $\pi_{1|i}$. Nejjednodušší porovnání pravděpodobností úspěchů ve dvou řádcích tabulky je *rozdíl pravděpodobností*

$$\pi_1 - \pi_2.$$

Porovnání neúspěchů je ekvivalentní s porovnáním úspěchů:

$$(1 - \pi_1) - (1 - \pi_2) = \pi_2 - \pi_1.$$

Rozdíl pravděpodobností nabývá hodnot z intervalu $(-1, 1)$, přičemž nule je roven v případě, že $\pi_1 = \pi_2$.

2.4 Relativní riziko

Rozdíl pravděpodobností není vhodné používat v případě, že se obě hodnoty π_i nachází blízko 0 nebo 1. Například, pokud nás zajímá výsledek léčby (použití určitého léku) na podílu pacientů, kteří po této léčbě zemřeli, rozdíl mezi 0,01 a 0,001 může být zajímavější než rozdíl mezi 0,41 a 0,401, přestože oba tyto rozdíly jsou rovny 0,009. V takových případech je lepší pracovat s podílem pravděpodobností. Proto se zavádí *relativní riziko*, které definujeme jako podíl

$$\frac{\pi_1}{\pi_2}. \tag{2.2}$$

Relativní riziko může nabývat libovolných nezáporných hodnot. Pokud je podíl roven jedné, veličiny X a Y jsou nezávislé.

Jestliže se vrátíme k výše zmíněnému příkladu, je relativní riziko v prvním případě rovno $0,01/0,001 = 10$ a ve druhém případě $0,41/0,401 = 1,02$. To znamená, že v prvním případě je pravděpodobnost nepříznivé reakce pacientů na jeden lék desetkrát větší než na druhý.

2.5 Poměr šancí

Nechť π je pravděpodobnost úspěchu, potom definujeme *šance* jako

$$\Omega = \frac{\pi}{(1-\pi)}.$$

Šance mohou nabývat nezáporných hodnot. Jestliže je $\Omega > 1$, potom platí, že úspěch nastává častěji než neúspěch. Například, je-li $\pi = 0.8$, potom $\Omega = (0.8)/(0.2) = 4$ znamená, že na každý jeden neúspěch očekáváme čtyři úspěchy. Na druhou stranu, pokud známe šance Ω, potom pro pravděpodobnost úspěchu π platí

$$\pi = \frac{\Omega}{(\Omega+1)}.$$

Šance na úspěch v dvojrozměrné kontingenční tabulce typu 2×2 v řádku i jsou definovány jako

$$\Omega_i = \frac{\pi_i}{(1-\pi_i)}.$$

Definice 2.3. *Podíl šancí Ω_1 a Ω_2 ve dvou řádcích tabulky typu 2×2,*

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}, \tag{2.3}$$

se nazývá poměr šancí.

Nechť $\{\pi_{ij}\}$ jsou pravděpodobnosti jednotlivých polí tabulky. Potom pro šance v řádku i platí

$$\Omega_i = \frac{\pi_{i1}}{\pi_{i2}}.$$

Dosazením Ω_i do (2.3) získáme ekvivalentní definici poměru šancí,

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \tag{2.4}$$

2.5.1 Vlastnosti poměru šancí

Poměr šancí může být roven libovolnému nezápornému číslu. Jestliže $1 < \theta < \infty$, potom pravděpodobnost na úspěch je větší pro objekty v řádku 1 než pro objekty v řádku 2. To znamená, že $\pi_1 > \pi_2$. Například pro $\theta = 5$, šance na úspěch v řádku 1 jsou pětkrát větší než v řádku 2. Jestliže $0 < \theta < 1$, potom $\pi_1 < \pi_2$. Pokud některé pole tabulky obsahuje nulovou pravděpodobnost, poměr šancí θ je roven 0 nebo ∞.

Pomocí poměru šancí můžeme posuzovat závislost veličin X a Y. Předpokládejme, že všechny pravděpodobnosti jednotlivých polí tabulky jsou kladné. Pokud platí

$$\Omega_1 = \Omega_2,$$

neboli $\theta = 1$, potom jsou veličiny X a Y nezávislé. Protože v kontingenční tabulce jsou veličiny X a Y nezávislé právě tehdy, když $\pi_{ij} = \pi_{i.}\pi_{.j}$ pro každou dvojici (i, j), platí následující věta:

Věta 2.1. *V dvojrozměrné kontingenční tabulce typu* 2×2 *je* $\theta = 1$ *právě tehdy, když pro každou dvojici* (i, j) *platí*

$$\pi_{ij} = \pi_{i.}\pi_{.j}.$$

Důkaz. Je-li nejprve $\pi_{ij} = \pi_{i.}\pi_{.j}$ pro všechny dvojice (i, j), dosazením do (2.4) okamžitě dostaneme, že $\theta = 1$.

Nechť nyní platí $\theta = 1$. Označme $\pi_{11}/\pi_{12} = \lambda$. Pak z $\theta = 1$ plyne, že i $\pi_{21}/\pi_{22} = \lambda$. Odtud dostáváme

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

Příslušnou tabulku pravděpodobností nyní můžeme psát ve tvaru

$\lambda\pi_{12}$	π_{12}	$(\lambda+1)\pi_{12}$
$\lambda\pi_{22}$	π_{22}	$(\lambda+1)\pi_{22}$
		$(\lambda+1)\pi_{.2}$

Protože $(\lambda + 1)\pi_{.2} = 1$, dostáváme, že $\lambda + 1 = 1/\pi_{.2}$. Ze vztahu $(\lambda + 1)\pi_{12} = \pi_{1.}$ plyne, že $\pi_{12} = \pi_{1.}\pi_{.2}$. Podobně z $(\lambda + 1)\pi_{22} = \pi_{2.}$ vyplývá $\pi_{22} = \pi_{2.}\pi_{.2}$. Nakonec dostaneme

$$\pi_{11} = \pi_{1.} - \pi_{12} = \pi_{1.} - \pi_{1.}\pi_{.2} = \pi_{1.}(1 - \pi_{.2}) = \pi_{1.}\pi_{.1}$$

Podobně

$$\pi_{21} = \pi_{2.} - \pi_{22} = \pi_{2.} - \pi_{2.}\pi_{.2} = \pi_{2.}(1 - \pi_{.2}) = \pi_{2.}\pi_{.1}.$$

□

Čím více je hodnota θ vzdálená od jedné, tím silnější je závislost sledovaných znaků. Pokud pro dvě hodnoty θ platí, že jedna je převrácenou hodnotou druhé, potom obě hodnoty definují stejný vztah mezi objekty, ale v opačném směru. Tedy, je-li například $\theta = 0.25$, šance na úspěch v řádku 1 jsou čtyřikrát menší než v řádku 2. Naopak, pokud $\theta = 4$, šance na úspěch v řádku 1 jsou čtyřikrát větší než v řádku 2. Jestliže změníme pořadí řádků nebo sloupců, nový poměr šancí je roven převrácené hodnotě původního. Ze vzorce (2.4) plyne, že při změně orientace tabulky, neboli v tabulce transponované, se hodnota poměru šancí nezmění.

Poměr šancí můžeme také dobře definovat pomocí opačných podmíněných pravděpodobností. Pokud máme sdružené rozdělení X a Y, podmíněné rozdělení existuje a pro poměr šancí platí

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{P(Y = 1|X = 1)/P(Y = 2|X = 1)}{P(Y = 1|X = 2)/P(Y = 2|X = 2)}$$

$$= \frac{P(X = 1|Y = 1)/P(X = 2|Y = 1)}{P(X = 1|Y = 2)/P(X = 2|Y = 2)}.$$

Podle definice poměru šancí (2.3) a relativního rizika (2.2) můžeme vztah mezi nimi vyjádřit jako

$$\text{poměr šancí} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \text{relativní riziko} \times \left(\frac{1-\pi_2}{1-\pi_1}\right).$$

Poměr šancí a relativní riziko mají podobný význam, kdykoli jsou pravděpodobnosti π_i blízko nule. Tato vlastnost se ve statistice často využívá v případech, kdy relativní riziko nelze z dostupných dat odhadnout, např. v retrospektivních studiích typu case-control. Tento typ studií se používá např. v epidemiologii, kdy ke skupině nemocných (cases) náhodně z populace vybereme kontrolní skupinu (controls) a zpětně zjišťujeme, zda sledovaní jedinci byli nebo nebyli vystaveni nějakému rizikovému faktoru.

2.5.2 Logaritmický poměr šancí

Protože hodnoty poměru šancí θ jsou rozloženy nesymetricky kolem jedné, zavádíme *logaritmický poměr šancí* $\log\theta$. Veličiny X a Y jsou potom nezávislé, pokud

$$\log\theta = 0.$$

Logaritmický poměr šancí je symetrický kolem 0. Při přehození řádků nebo sloupců dojde pouze ke změně znaménka. Dvě hodnoty $\log\theta$, které jsou stejné až na znaménko, určují stejnou sílu závislosti.

Logaritmický poměr šancí využijeme později především k odvození intervalu spolehlivosti pro poměr šancí θ.

2.5.3 Poměr šancí v trojrozměrných tabulkách

Mějme trojrozměrné kontingenční tabulky, které obsahují kombinace hodnot veličin *X, Y* a *Z*. Cílem je analyzovat vztah mezi X a Y, buď na pevně daných úrovních veličiny Z nebo při sloučení tabulky přes hodnoty veličiny Z. V prvním případě používáme *parciální tabulky*, což jsou dvojrozměrné průřezy trojrozměrnou tabulkou. Ve druhém případě pracujeme s dvojrozměrnými kontingenčními tabulkami, které získáme kombinací parciálních tabulek. Takové tabulky se nazývají *XY marginální tabulky*. Hodnota v určitém poli marginální tabulky je rovna součtu hodnot ze stejných polí jednotlivých parciálních tabulek.

Předpokládejme, že máme tabulku typu $2 \times 2 \times K$, kde K značí celkový počet kategorií veličiny Z. Nechť $\{\mu_{ijk}\}$ označují očekávané četnosti v polích tabulky. V rámci kategorie k veličiny Z definujeme *poměr šancí* jako

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}.$$

Poměr šancí popisuje podmíněný vztah mezi X a Y v *k-té* parciální tabulce. Poměry šancí pro všech K parciálních tabulek se nazývají *XY podmíněné poměry šancí*.

Nechť dále $\{\mu_{ij.} = \sum_k \mu_{ijk}\}$ jsou očekávané četnosti v XY marginálních tabulkách. XY *marginální poměr šancí* je roven

$$\theta_{XY} = \frac{\mu_{11.}\mu_{22.}}{\mu_{12.}\mu_{21.}}.$$

Důležitou vlastností trojrozměrných tabulek typu $2 \times 2 \times K$ je *homogenní XY asociace*. Tabulka má tuto vlastnost, pokud

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}.$$

To znamená, že vliv X na Y je stejný v každé úrovni Z. Speciálním případem homogenní asociace je *podmíněná nezávislost*, kde $\theta_{XY(k)} = 1$ pro všechna k. Neboli X a Y jsou nezávislé v každé parciální tabulce.

Problematikou trojrozměrných tabulek se v této práci nebudeme podrobně zabývat, informace lze nalézt např. v knize [2].

Kapitola 3

Odhady poměru šancí

Mějme opět náhodný vektor (X, Y) a příslušnou tabulku jako v kapitole 2.1. Předpokládejme nyní, že se uskutečnil výběr z tohoto rozdělení o rozsahu n. Nechť n_{ij} jsou počty případů, kdy se ve výběru vyskytla dvojice (i, j). Hodnoty n_{ij} se nazývají *četnosti*. Odhadem teoretických pravděpodobností π_{ij} jsou výrazy $\hat{\pi}_{ij}$, pro které platí

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

Podobně podmíněné pravděpodobnosti $\pi_{j|i}$ odhadujeme pomocí

$$\hat{\pi}_{j|i} = \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i.}} = \frac{n_{ij}}{n_{i.}}.$$

Platí $n_{i.} = n\hat{\pi}_{i.} = \sum_j n_{ij}$. Čísla $n_{i.}$ a $n_{.j}$ se nazývají *marginální četnosti*. Vztah mezi četnostmi a celkovým rozsahem výběru n můžeme vyjádřit jako

$$n = \sum_i n_{i.} = \sum_j n_{.j} = \sum_i \sum_j n_{ij}. \tag{3.1}$$

.

3.1 Bodový odhad poměru šancí

Poměr šancí θ je ve vzorci (2.4) definován pomocí pravděpodobností $\{\pi_{ij}\}$. Pokud k výpočtu použijeme četnosti $\{n_{ij}\}$, potom získáme *empirický poměr šancí*

$$\hat{\theta} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Tato hodnota se nemění, pokud obě četnosti v libovolném řádku nebo sloupci násobíme nenulovou konstantou. Podle předchozího víme, že

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

je odhadem π_{ij}. Z toho plyne, že empirický poměr šancí je odhadem poměru šancí.

Jestliže nějaká četnost n_{ij} je nulová, potom $\hat{\theta}$ je rovno 0 nebo ∞. Pokud obě četnosti v řádku nebo sloupci jsou nulové s kladnou pravděpodobností, $\hat{\theta}$ není definováno. V takovém případě se doporučuje ke každé četnosti v poli tabulky přičíst např. hodnotu 0.5. Tím získáme odhad

$$\tilde{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)}.$$

3.2 Intervalový odhad poměru šancí

Odhad logaritmického poměru šancí je roven $\log \hat{\theta}$. K odvození intervalového odhadu poměru šancí použijeme právě $\log \hat{\theta}$.

Tvrzení 3.1. *Přibližná směrodatná odchylka logaritmického poměru šancí* $\log \hat{\theta}$ *je rovna*

$$\hat{\sigma}(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{3.2}$$

Důkaz. Směrodatná odchylka $\hat{\sigma}(\log \hat{\theta})$ je výsledkem vícerozměrné delta metody, která je podrobně popsána v Dodatku. Položme $g(\boldsymbol{\pi}) = \log \theta = \log \pi_{11} + \log \pi_{22} - \log \pi_{12} - \log \pi_{21}$. Dále spočteme

$$\phi_{11} = \frac{\partial(\log \theta)}{\partial \pi_{11}} = \frac{1}{\pi_{11}}.$$

Podobným způsobem získáme $\phi_{12} = -1/\pi_{12}$, $\phi_{21} = -1/\pi_{21}$ a $\phi_{22} = 1/\pi_{22}$. Po dosazení předchozích hodnot do vzorce (A.4) dostaneme, že $\sum_i \sum_j \pi_{ij}\phi_{ij} = 0$ a

$$\sigma^2 = \sum_i \sum_j \pi_{ij}\phi_{ij}^2 = \sum_i \sum_j \left(\frac{1}{\pi_{ij}}\right).$$

Asymptotická směrodatná odchylka $\log \hat{\theta}$ při výběru z multinomického rozdělení s četnostmi $\{n_{ij}\}$ je rovna

$$\sigma(\log \hat{\theta}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\sum_i \sum_j \frac{1}{n\pi_{ij}}}.$$

Z odhadu pravděpodobnosti π_{ij} plyne, že $n\hat{\pi}_{ij} = n_{ij}$. Dosazením za $n\pi_{ij}$ dostaneme směrodatnou odchylku $\hat{\sigma}(\log \hat{\theta})$ ze vzorce (3.2). □

Čím více pozorování máme k dispozici, tím přesnější odhad dostaneme. Neboli hodnota $\hat{\sigma}$ roste, pokud četnosti $\{n_{ij};\ i = 1, 2,\ j = 1, 2\}$ klesají. Podle [2] má $\log \hat{\theta}$ při velkém počtu pozorování asymptoticky normální rozdělení se střední hodnotou $\log \theta$. Asymptotický interval spolehlivosti pro $\log \theta$ má proto tvar

$$\log \hat{\theta} \pm u\left(\frac{\alpha}{2}\right) \hat{\sigma}(\log \hat{\theta}),$$

kde $u(\alpha/2)$ je kvantil normálního rozdělení. Pokud na krajní body tohoto intervalu použijeme exponenciální funkci, získáme asymptotický interval spolehlivosti pro poměr šancí

$$\hat{\theta} \pm \exp\{u\left(\frac{\alpha}{2}\right)\hat{\sigma}(\log\hat{\theta})\}.$$

Kapitola 4

Testování hypotéz

V této kapitole se budeme zabývat testováním nezávislosti dvou kategoriálních veličin v dvojrozměrných kontingenčních tabulkách. Mějme výběr z multinomického rozdělení s parametry n a $\{\pi_{ij}\}$. Podle vzorce (2.1) víme, že dvě veličiny jsou nezávislé, pokud $\pi_{ij} = \pi_{i.}\pi_{.j}$ pro všechny dvojice (i, j). Obecně proto nulovou hypotézu nezávislosti můžeme psát ve tvaru

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \qquad pro\ \forall i = 1, \ldots, I;\ j = 1, \ldots, J. \tag{4.1}$$

Zároveň pro nezávislost platí $\theta = 1$ a $\log\theta = 0$. Proto lze (4.1) ekvivalentně psát jako

$$H_0 : \log\theta = 0. \tag{4.2}$$

4.1 Test nezávislosti pomocí poměru šancí

V kontingenčních tabulkách typu 2×2 můžeme nezávislost testovat pomocí poměru šancí, přesněji pomocí jeho logaritmu. Víme, že veličiny jsou nezávislé, pokud $\log\theta = 0$. Budeme tedy testovat nulovou hypotézu nezávislosti jako v (4.2),

$$H_0 : \log\theta = 0.$$

Nejprve dokážeme pomocnou větu, kterou dále použijeme k odvození testu.

Věta 4.1. *Nechť $\alpha_1, \ldots, \alpha_k$ jsou taková reálná čísla, že $\alpha_1 + \ldots + \alpha_k = 0$, $|\alpha_1| + \ldots + |\alpha_k| > 0$. Nechť $\mathbf{X} = (X_1, \ldots, X_k)$ má multinomické rozdělení s parametry n a $\pi_1, \ldots, \pi_k$. Položme*

$$\delta = \sum_{i=1}^{k} \alpha_i \log\pi_i, \qquad d = \sum_{i=1}^{k} \alpha_i \log X_i, \qquad \sigma = \sqrt{\sum_{i=1}^{k} \frac{\alpha_i^2}{\pi_i}}.$$

Pak pro $n \to \infty$ platí

$$\frac{d - \delta}{\sigma} \xrightarrow{d} N(0, 1).$$

Důkaz. Předpokládejme, že $\mathbf{T}_n = (T_{n_1}, \ldots, T_{n_k})$ je náhodný vektor, který má asymptoticky normální rozdělení $N(0, \mathbf{\Sigma})$ a $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ je vektor parametrů. Označme $\sigma_{ij} = \sigma_{ij}(\boldsymbol{\theta})$ prvky matice $\mathbf{\Sigma}$. Položme $\theta_i = \pi_i$, $T_{n_i} = X_i/n$ pro $i = 1, \ldots, k$. Potom $\boldsymbol{\theta} = \boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$. Víme, že $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\pi}) \xrightarrow{d} N(0, \mathbf{\Sigma})$, kde $\sigma_{ii} = \pi_i(1 - \pi_i)$, $\sigma_{ij} = -\pi_i\pi_j$ pro $i \neq j$. Definujme $g(\boldsymbol{\pi}) = \sum_{i=1}^{k} \alpha_i \log \pi_i$, kde g je měřitelná funkce k proměnných, která má totální diferenciál v bodě $\boldsymbol{\pi}$, který je skutečnou hodnotou parametru. Pak $\sqrt{n}(d - \delta) = \sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\pi}))$. Označme

$$v(\boldsymbol{\pi}) = \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma_{ij} \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_i} \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_j}.$$

Dosazením do předchozího vzorce dostaneme

$$v(\boldsymbol{\pi}) = \sum_{i=1}^{k} \pi_i(1 - \pi_i) \left(\frac{\alpha_i}{\pi_i}\right)^2 + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} (-\pi_i\pi_j) \frac{\alpha_i}{\pi_i} \frac{\alpha_j}{\pi_j} = \sum_{i=1}^{k} \frac{\alpha_i^2}{\pi_i}.$$

Jestliže je $v(\boldsymbol{\pi}) \neq 0$, potom

$$\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\pi})) \xrightarrow{d} N[0, v(\boldsymbol{\pi})].$$

Z toho plyne, že

$$\frac{\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\pi}))}{\sqrt{v(\boldsymbol{\pi})}} \xrightarrow{d} N(0, 1).$$

□

Položme

$$d = \log \hat{\theta}, \qquad \delta = \log \theta, \qquad \sigma = \hat{\sigma}(\log \hat{\theta}).$$

Náhodná veličina

$$D = \frac{\log \hat{\theta} - \log \theta}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

má podle věty 4.1. asymptoticky rozdělení $N(0, 1)$. Při testování hypotézy $H_0 : \log \theta = 0$ do veličiny D za $\log \theta$ dosadíme nulu a pokud platí

$$|D| \geq u\left(\frac{\alpha}{2}\right),$$

nulovou hypotézu zamítneme na hladině α.

Protože se jedná o asymptotický test, můžeme ho používat jen při dostatečně velkých četnostech.

4.2 Pearsonův χ^2 test

Pearsonův χ^2 test umožňuje testovat nezávislost v obecnějších tabulkách než v tabulkách typu 2×2. V takovém případě testujeme hypotézu

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \qquad pro\ \forall i = 1, \ldots, I;\ j = 1, \ldots, J.$$

Pearsonova χ^2 statistika je definována jako

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}. \tag{4.3}$$

Při $n \to \infty$ má náhodná veličina χ^2 asymptoticky rozdělení χ^2 s $IJ-1$ stupni volnosti. Své minimální hodnoty nabývá v případě, že $n_{ij} = \mu_{ij}$. Hodnoty $\{\mu_{ij} = n\pi_{ij}\}$ jsou očekávané četnosti. Za platnosti hypotézy $H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}$ je $\mu_{ij} = n\pi_{i.}\pi_{.j}$ a platí, že $\mu_{ij} = E(n_{ij})$. Parametry $\pi_{i.}$ a $\pi_{.j}$ většinou nejsou známé, proto k testování použijeme jejich odhady $\hat{\pi}_{i.} = n_{i.}/n$ a $\hat{\pi}_{.j} = n_{.j}/n$. Podrobné odvození těchto odhadů je uvedeno např. v knize [3]. Tím získáme odhady očekávaných četností

$$\hat{\mu}_{ij} = n\hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{n_{i.}n_{.j}}{n}.$$

Dosazením $\{\hat{\mu}_{ij}\}$ do (4.3) získáme nový tvar veličiny χ^2:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}. \tag{4.4}$$

Nahrazením $\{\hat{\mu}_{ij}\}$ za $\{\mu_{ij}\}$ dojde ke změně stupňů volnosti. Náhodná veličina χ^2 definovaná ve vzorci (4.4) má také asymptoticky rozdělení χ^2 s počtem stupňů volnosti

$$df = (IJ - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1),$$

kde $(I - 1)$ je počet parametrů $\{\pi_{i.}\}$ a $(J - 1)$ je počet $\{\pi_{.j}\}$, což vyplývá z podmínky $\sum_i \pi_{i.} = \sum_j \pi_{.j} = 1$. Nulovou hypotézu H_0 o nezávislosti veličin zamítneme na hladině α, jestliže

$$\chi^2 \geq \chi^2_{(I-1)(J-1)}(\alpha).$$

Nyní ještě odvodíme χ^2 statistiku speciálně pro tabulky typu 2×2. Pokud do vzorce (4.4) dosadíme za očekávané četnosti $\{\hat{\mu_{ij}}\}$, dostaneme

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

Jestliže rozepíšeme druhý součet přes indexy j, dosadíme $n_{i2} = n_{i.} - n_{i1}$ a $n_{.2} = n - n_{.1}$ a použijeme vztah (3.1), postupně dostaneme

$$\begin{aligned}
\chi^2 &= n \sum_{i=1}^{2} \left[\frac{\left(n_{i1} - \frac{n_{i.}n_{.1}}{n}\right)^2}{n_{i.}n_{.1}} + \frac{\left(n_{i2} - \frac{n_{i.}n_{.2}}{n}\right)^2}{n_{i.}n_{.2}} \right] \\
&= \frac{1}{n} \sum_{i=1}^{2} (nn_{i1} - n_{i.}n_{.1})^2 \left[\frac{1}{n_{i.}n_{.1}} + \frac{1}{n_{i.}n_{.2}} \right] \\
&= \frac{1}{n} \sum_{i=1}^{2} n^2 \left(n_{i1} - \frac{n_{i.}n_{.1}}{n} \right)^2 \left(\frac{n_{.2} + n_{.1}}{n_{i.}n_{.1}n_{.2}} \right) \\
&= \frac{n^2}{n_{.1}n_{.2}} \sum_{i=1}^{2} \frac{1}{n_{i.}} \left(n_{i1} - \frac{n_{i.}n_{.1}}{n} \right)^2 \\
&= \frac{n^2}{n_{.1}n_{.2}} \sum_{i=1}^{2} n_{i.} \left(\frac{n_{i1}}{n_{i.}} - \frac{n_{.1}}{n} \right)^2 \\
&= \frac{n^2}{n_{.1}n_{.2}} \left[n_{1.} \left(\frac{n_{11}}{n_{1.}} - \frac{n_{.1}}{n} \right)^2 + n_{2.} \left(\frac{n_{21}}{n_{2.}} - \frac{n_{.1}}{n} \right)^2 \right] \qquad (4.5)
\end{aligned}$$

Nejprve vypočteme

$$\left(\frac{n_{11}}{n_{1.}} - \frac{n_{.1}}{n} \right)^2 = \left(\frac{n_{11}n - n_{.1}n_{1.}}{n_{1.}n} \right)^2 = \left[\frac{n_{11}(n_{11} + n_{12} + n_{21} + n_{22}) - (n_{11} + n_{21})(n_{11} + n_{12})}{n_{1.}n} \right].$$

Odtud snadnou úpravou dostaneme, že

$$\left(\frac{n_{11}}{n_{1.}} - \frac{n_{.1}}{n} \right)^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}^2 n^2}.$$

Podobně

$$\left(\frac{n_{21}}{n_{2.}} - \frac{n_{.1}}{n} \right)^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}^2 n^2}.$$

Předchozí výrazy dosadíme do vzorce (4.5), následně vytkneme stejné hodnoty a použijeme vztah $n_{1.} + n_{2.} = n$. Nakonec dostaneme

$$\chi^2 = n \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}. \qquad (4.6)$$

Pokud testujeme hypotézu nezávislosti $H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}$, porovnáváme hodnotu χ^2 s kritickou hodnotou $\chi_1^2(\alpha)$. Platí-li

$$\chi^2 > \chi_1^2(\alpha),$$

zamítáme nulovou hypotézu nezávislosti v tabulkách 2×2 na hladině α.

4.3 Poměr věrohodností

Alternativní statistická metoda pro testování hypotézy $H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}$ v tabulkách typu $I \times J$ vychází z metody maximální věrohodnosti. Věrohodnostní funkce multinomického rozdělení je

$$\prod_i \prod_j \pi_{ij}^{n_{ij}}, \quad \text{kde všechna } \pi_{ij} \geq 0 \quad \text{a} \quad \sum_i \sum_j \pi_{ij} = 1.$$

Za platnosti hypotézy H_0 je $\hat{\pi}_{ij} = \hat{\pi}_{i.}\hat{\pi}_{.j} = n_{i.}n_{.j}/n^2$. Obecně je $\hat{\pi}_{ij} = n_{ij}/n$. Poměr příslušných věrohodnostních funkcí je roven

$$\Lambda = \frac{\prod_i \prod_j (n_{i.}n_{.j})^{n_{ij}}}{n^n \prod_i \prod_j n_{ij}^{n_{ij}}}.$$

Maximálně věrohodná χ^2 statistika je definována výrazem

$$-2 \log \Lambda.$$

Tato statistika se značí G^2 a je rovna

$$G^2 = -2 \log \Lambda = 2 \sum_i \sum_j n_{ij} \log \left(\frac{n_{ij}}{\hat{\mu}_{ij}} \right), \tag{4.7}$$

kde $\{\hat{\mu}_{ij} = n_{i.}n_{.j}/n\}$. Obecně je parametrický prostor tvořen pravděpodobnostmi $\{\pi_{ij}\}$. Jeho dimenze je rovna $IJ - 1$, neboť $\sum_i \sum_j \pi_{ij} = 1$. Za platnosti hypotézy H_0 jsou $\{\pi_{ij}\}$ vypočteny pomocí $\{\pi_{i.}\}$ a $\{\pi_{.j}\}$, tedy dimenze je rovna $(I-1)+(J-1)$. Rozdíl mezi těmito dimenzemi je $(IJ-1)-(I-1)-(J-1) = (I-1)(J-1)$. Pro velká n má náhodná veličina G^2 asymptoticky rozdělení $\chi^2_{(I-1)(J-1)}$. Stejně jako v předchozí metodě zamítáme nulovou hypotézu na hladině α, je-li

$$G^2 \geq \chi^2_{(I-1)(J-1)}(\alpha).$$

Veličiny X^2 i G^2 při $n \to \infty$ konvergují ke stejnému rozdělení χ^2, tedy jejich rozdíl $X^2 - G^2$ konverguje v pravděpodobnosti k nule. Právě probrané metody můžeme používat v případě, že rozsah výběru n je dostatečně velký. S rostoucím n roste i $\{\mu_{ij} = n\pi_{ij}\}$ a podle [2] se rozdělení X^2 a G^2 se blíží k rozdělení χ^2, přičemž X^2 k němu konverguje rychleji než G^2. K použití statistiky G^2 by mělo platit, že $n/IJ \geq 5$. Pokud je hodnota I nebo J hodně velká, pro použití X^2 by některé očekávané četnosti měly být přibližně 1, ale většina by měla překročit hodnotu 5.

Pro výběry malého rozsahu je vhodnější použít některý z exaktních diskrétních testů, kterými se v této práci nebudeme zabývat. Jejich přehled a odvození lze nalézt např. v [2].

4.4 Rezidua

Předchozí testy v obecných kontingenčních tabulkách a jejich p-hodnoty vypovídají pouze o tom, zda nulovou hypotézu o nezávislosti zamítáme nebo naopak. P-hodnota je nejnižší hladina, na které zamítáme hypotézu. Porovnáváním pozorovaných a odhadovaných

očekávaných četností v jednotlivých polích tabulky můžeme zjistit, které dvojice indexů se nejvýrazněji podílejí na zamítnutí hypotézy H_0.

Pro pole tabulky s většími hodnotami μ_{ij} jsou však i rozdíly mezi n_{ij} a $\hat{\mu}_{ij}$ větší, proto obyčejný rozdíl $(n_{ij} - \hat{\mu}_{ij})$ není pro takovou analýzu postačující. Výhodnější je používat *Pearsonovo reziduum* definované

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}.$$

Vztah mezi Pearsonovo reziduem a Pearsonovou statistikou se dá vyjádřit jako

$$\sum_i \sum_j e_{ij}^2 = X^2.$$

Za platnosti hypotézy H_0 mají $\{e_{ij}\}$ podle [2] asymptoticky normální rozdělení se střední hodnotou 0. Jejich asymptotický rozptyl je menší než 1. Jako konzervativní test o porušení nezávislosti dvojicí indexů (i, j) můžeme použít následující pravidlo. Je-li

$$e_{ij} > u\left(\frac{\alpha}{2}\right),$$

potom pro dvojici indexů (i, j) je $\pi_{ij} \neq \pi_{i.}\pi_{.j}$.

Pokud vydělíme Pearsonovo reziduum jeho směrodatnou odchylkou, dostaneme

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - \hat{\pi}_{i.})(1 - \hat{\pi}_{.j})}}. \tag{4.8}$$

Podíl (4.8) se nazývá *standardizované Pearsonovo reziduum*. Toto reziduum má asymptoticky standardní normální rozdělení. Hodnota kvantilu normálního rozdělení $u(\alpha/2)$ na hladině $\alpha = 5$ je rovna 1.96. Jestliže standardizované Pearsonovo reziduum v absolutní hodnotě překročí hodnotu 2 (případně 3, pokud testujeme hypotézu na hladině $\alpha < 1$), potom je hypotéza nezávislosti H_0 porušena v odpovídajícím poli tabulky.

Dodatek A

Delta metoda

Delta metoda se často používá k odvozování směrodatných odchylek pro velké výběry z dat. Nechť θ je neznámý parametr. Nechť T_n je statistika, která má asymptotické normální rozdělení. Pomocí T_n chceme odhadnout parametr θ. Nechť dále $g(T_n)$ je odhadem θ. Pro dostatečně velké n má $g(T_n)$ normální rozdělení. Směrodatná odchylka závisí na tom, jak rychle se $g(t)$ v závislosti na t mění v blízkosti θ.

Předpokládejme, že T_n má normální rozdělení se střední hodnotou θ a směrodatnou odchylkou $\sigma/\sqrt{n}$. Při $n \to \infty$ platí

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2). \tag{A.1}$$

Tvrzení A.1. *Pokud je g alespoň dvakrát diferencovatelná v θ, potom pro velká n je*

$$\sqrt{n}[g(T_n) - g(\theta)] \approx \sqrt{n}(T_n - \theta)g'(\theta), \tag{A.2}$$

kde $g'(\theta) = \partial g/\partial\theta$.

Víme, že pokud náhodná veličina $Y \backsim N(0, \sigma^2)$, potom veličina $cY \backsim N(0, c^2\sigma^2)$. Z tohoto tvrzení a ze vzorců (A.1) a (A.2) plyne, že

$$\sqrt{n}[g(T_n) - g(\theta)] \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2). \tag{A.3}$$

Výsledek (A.3) se nazývá *delta metoda*.

Nyní ještě popíšeme vícerozměrnou delta metodu. Nechť $g(T_n)$ je v tomto případě diferencovatelná funkce $g(\boldsymbol{\pi})$ proměnných $\{\pi_i\}$. Mějme výběr z multinomického rozdělení s parametry $(n, \{\pi_i\}),\ i = 1, \ldots, c$. Výběrová hodnota $g(\boldsymbol{\pi})$ je $g(\hat{\boldsymbol{\pi}})$. Nechť dále

$$\phi_i = \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_i}, \qquad i = 1, \ldots, c.$$

Pokud $n \to \infty$, potom rozdělení veličiny $\sqrt{n}[g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})]/\sigma$ konverguje k normálnímu rozdělení, kde

$$\sigma^2 = \sum \pi_i\phi_i^2 - (\sum \pi_i\phi_i)^2. \tag{A.4}$$

Nahrazením $\{\pi_i\}$ a $\{\phi_i\}$ ve vzorci (A.4) jejich výběrovými hodnotami získáme $\hat{\sigma}^2$, což je odhad σ^2. Směrodatná odchylka funkce $g(\hat{\boldsymbol{\pi}})$ je potom rovna $\hat{\sigma}/\sqrt{n}$ a interval spolehlivosti pro $g(\boldsymbol{\pi})$ je

$$g(\hat{\boldsymbol{\pi}}) \pm u\left(\frac{\alpha}{2}\right)\hat{\sigma}/\sqrt{n}.$$

Poznámka A.1. *Pokud ve vzorci (A.4) nahradíme σ hodnotou $\hat{\sigma}$, limitní rozdělení je stále standardní normální, ale konvergence je pomalejší. Podle slabého zákona velkých čísel výběrové pravděpodobnosti konvergují v pravděpodobnosti k $\{\pi_i\}$. Odtud $\hat{\sigma}$ konverguje v pravděpodobnosti k σ, tedy $\sigma/\hat{\sigma}$ konverguje v pravděpodobnosti k jedné. Pak*

$$\sqrt{n}\frac{g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})}{\hat{\sigma}} = \sqrt{n}\frac{g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})}{\sigma}\frac{\sigma}{\hat{\sigma}}. \tag{A.5}$$

Vezměme pravou stranu vzorce A.5. První výraz konverguje v distribuci ke standardnímu normánímu rozdělení a druhý výraz konverguje v pravděpodobnosti k jedné. To znamená, že jejich součin má limitní normální rozdělení.

Literatura

[1] Agresti, A.: *An Introduction to Categorical Data Analysis*, Wiley-Interscience, 1996

[2] Agresti, A.: *Categorical Data Analysis*, Wiley-Interscience, 2002

[3] Anděl, J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2005.