Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Kamila Bárnetová

Intervaly spolehlivosti pro parametry multinomického rozdělení

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	prof. RNDr. Jiří Anděl, DrSc.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2014

Ráda bych poděkovala prof. RNDr. Jiřímu Andělovi, DrSc. za odbornou pomoc při vypracování bakalářské práce.

Prohlašuji, že jsem tuto bakalářskou práci vypracovala 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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V .... dne .... Podpis autora

Název práce: Intervaly spolehlivosti pro parametry multinomického rozdělení

Autor: Kamila Bárnetová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: prof. RNDr. Jiří Anděl, DrSc., katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této práci jsou popsány intervaly spolehlivosti pro parametry binomického a multinomického rozdělení. Tyto intervaly se dají v praxi použít např. pro předvolební odhady. První dvě kapitoly popisují odvození těchto intervalů. Poslední kapitola je věnována simulacím a porovnání několika vybraných metod. Na základě provedených simulací považujeme za vhodné volit pro výpočet intervalu spolehlivosti pro parametry multinomického rozdělení intervaly založené na Bonferroniho nerovnosti, případně jejich modifikace. Tyto intervaly se dají snadno spočítat a zároveň jejich pravděpodobnost pokrytí je alespoň 0.89.

Klíčová slova: interval spolehlivosti, multinomické rozdělení, binomické rozdělení, Bonferroniho nerovnost

Title: Confidence intervals for parameters of multinomial distribution

Author: Kamila Bárnetová

Department: Department of Probability and Mathematical Statistics

Supervisor: prof. RNDr. Jiří Anděl, DrSc., Department of Probability and Mathematical Statistics

Abstract: Confidence intervals for parameters for binomial and multinomial distribution are described in this thesis. These intervals can be used in practice, for exemple- pre-election estimates. The first two chapter are devoted to derivation of these intervals. Simulations and comparison of several selected methods can be found in the last chapter. Based on the simulations, we consider it appropriate, to choose to calculate confidence intervals for parameters of multinomial distribution intervals based on Bonferroniho inequality, or their modifications. These intervals are easy to calculate, while their coverage probability is at least 0.89.

Keywords: confidence interval, multinomial distribution, binomial distribution, Bonferroni inequality

Obsah

Úvod

V praxi se často setkáváme s úlohou hledání intervalů spolehlivosti pro multinomické rozdělení. Typickým příkladem jsou volební modely, hojně užívané především pro předvolební odhady výsledků. Uvažujme tento problém: Mějme k kandidátů ve volbách. Dále z dat získaných při předvolebním průzkumu známe rozsah výběru n, tj. počet respondentů, a četnosti X_i, představují počet hlasů pro jednotlivé kandidáty.

Kapitola 1 této práce se věnuje intervalům spolehlivosti pro parametr binomického rozdělení. Jsou v ní popsány a odvozeny běžně užívané intervaly i ty, které jsou méně známé.

Dva různé způsoby konstrukce intervalů pro parametry multinomického rozdělení jsou pak popsány v kapitole 2. První z nich je založen na *Pearsonově statistice*, je odvozen v článku [6] a dále upravován v článku [5]. Druhý je založený na *Bonferroniho nerovnosti* a popsán v článku [4].

Kapitola 3 je věnována simulacím multinomického rozdělení a jejich využití pro porovnání vybraných intervalů spolehlivosti z předchozích dvou kapitol.

Ještě poznamenejme, že v celé práci je používána desetinná tečka namísto desetinné čárky. Z jazykového hlediska by byla správnější desetinná čárka, na druhou stranu většina programů používá desetinnou tečku.

Kapitola 1

Intervaly spolehlivosti pro parametr binomického rozdělení

1.1 Definice a základní vlastnosti binomického rozdělení

Definice 1. *Nechť $n \in \mathbb{N}$ a $p \in (0,1)$. Předpokládejme, že veličina X nabývá pouze hodnot $0,1,\dots,n$, a to s pravděpodobnostmi $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$, $k=0,1,\dots,n$. Pak říkáme, že X má binomické rozdělení. Značíme $X \sim Bi(n,p)$.*

Nyní uvedeme dvě známé věty, jejichž důkazy lze snadno najít v učebnicích statistiky.

Věta 1. *Nechť $X \sim Bi(n,p)$, pak $\mathsf{E}\,X = p$ a $var X = p(1-p)$.*

Věta 2. *Nechť $X \sim Bi(n,p)$, kde $p \in (0,1)$. Pak pro $n \to \infty$ platí*

$$\tfrac{X-np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1).$$

Označme $z = u(\alpha/2)$ kritickou hodnotu normálního rozdělení $N(0,1)$ na hladině $\alpha/2$.

1.2 Waldův interval

Nechť $X \neq 0, X \neq n$, pak $\hat{p} = \frac{X}{n}$ je maximálně věrohodný odhad parametru p. Dále označme $\hat{q} = 1-\hat{p}$. Podle věty 2 platí, že

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} N(0,1). \tag{1.1}$$

Tedy

$$P\left(\frac{|\hat{p}-p|}{\sqrt{\frac{p(1-p)}{n}}} \leq z\right) \longrightarrow 1-\alpha. \tag{1.2}$$

Odtud

$$P\left[\hat{p} - z\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z\sqrt{\frac{p(1-p)}{n}}\right] \longrightarrow 1-\alpha. \tag{1.3}$$

Tím získáme interval

$$I = \left[\hat{p} - z\sqrt{\frac{p(1-p)}{n}}, \hat{p} + z\sqrt{\frac{p(1-p)}{n}}\right]. \tag{1.4}$$

Protože krajní body tohoto intervalu závisí na parametru p, jehož hodnotu neznáme, nahradíme p maximálně věrohodným odhadem $\hat{p}$ a dostaneme

$$I_s = \left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]. \tag{1.5}$$

Tím jsme získali *Waldův interval spolehlivosti*. Jedná se o jeden z nejčastěji v literatuře uváděných intervalů spolehlivosti pro binomické rozdělení. Někdy ho můžeme najít i pod názvem *standardní interval spolehlivosti*.

1.3 Wilsonův skórový interval

Alternativou k *Waldovu intervalu* je interval spolehlivosti založený na úpravách následující nerovnosti:

$$|\hat{p} - p| \leq z\sqrt{\frac{p(1-p)}{n}}. \tag{1.6}$$

Postupně dostaneme

$$\begin{aligned} (\hat{p}-p)^2 &\leq z^2 p(1-p)/n, \\ n(\hat{p}-p)^2 &\leq z^2 p(1-p), \\ n(\hat{p}^2 - 2p\hat{p} + p^2) &\leq z^2 p - z^2 p^2, \\ p^2(n+z^2) - p(2\hat{p}n + z^2) + n\hat{p}^2 &\leq 0 \end{aligned} \tag{1.7}$$

kvadratickou nerovnici pro p, kterou budeme dále řešit.

Rovnice

$$p^2(n+z^2) - p(2\hat{p}n + z^2) + n\hat{p}^2 = 0 \tag{1.8}$$

má kořeny

$$\begin{aligned} p_{1,2} &= \frac{(2\hat{p}n + z^2) \pm \sqrt{(2\hat{p}n+z^2)^2 - 4(n+z^2)n\hat{p}^2}}{2(n+z^2)} \\ &= \frac{(2\hat{p}n + z^2) \pm \sqrt{4\hat{p}nz^2 + z^4 - 4z^2n\hat{p}^2}}{2(n+z^2)} \\ &= \frac{\hat{p}n + \frac{z^2}{2}}{n+z^2} \pm \frac{z\sqrt{n}}{n+z^2}\sqrt{\hat{p}(1-\hat{p}) + \frac{z^2}{4n}}. \end{aligned} \tag{1.9}$$

Koeficient u p^2 v řešené nerovnici je kladný, proto interval (p_1,p_2) je námi hledaný interval spolehlivosti. Použijeme-li navíc značení $\hat{q} = 1 - \hat{p}$, získáme

$$I_W = \frac{\hat{p}n + \frac{z^2}{2}}{n + z^2} \pm \frac{z\sqrt{n}}{n + z^2}\sqrt{\hat{p}\hat{q} + \frac{z^2}{4n}}. \tag{1.10}$$

Tento interval byl zřejmě poprvé uveden v práci [8], proto ho nazýváme *Wilsonův interval spolehlivosti*, nebo také *skórový* či *q-interval.*

1.4 Agresti-Coullův interval

Ke zkonstruování tohoto intervalu spolehlivosti použijeme střed *Wilsonova* intervalu spolehlivosti z vyjádření v (1.10). Označme $\tilde{X} = X + \frac{z^2}{2}$ a $\tilde{n} = n + z^2$. Nechť $\tilde{p} = \frac{\tilde{X}}{\tilde{n}}$ a $\tilde{q} = 1 - \tilde{p}$. Dosadíme do vztahu (1.4) místo parametrů p a $\hat{p}$ odhad $\tilde{p}$. Získáme *Agresti-Coullův interval* s koncovými body

$$I_{AC} = \tilde{p} \pm z\sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}}. \tag{1.11}$$

Oba intervaly, *Wilsonův* i *Agresti-Coullův*, mají střed ve stejném bodě $\tilde{p}$. Podle [3] *Agresti-Coullův interval* nikdy není kratší než *Wilsonův.* Je-li $\alpha = 0.05$ a dosadíme-li za z hodnotu 2 místo 1.96, je tento interval „přidání dvou úspěchů a dvou neúspěchů“ intervalem popsaným v [1]. Z tohoto důvodu ho nazýváme *Agresti-Coullův.*

1.5 Bayesův interval

Beta rozdělení je konjugováno s apriorní hustotou binomického rozdělení a je běžné používat apriorní hustotu rozdělení beta k usuzování o parametru p. Jedná se o tzv. *beta-binomický model.*

Nechť $X \sim Bi(n,p)$ a nechť p má apriorní rozdělení *Beta*(a_1, a_2). Nechť dále nastane x úspěchů v n nezávislých pokusech. Pak aposteriorní hustota parametru p je

$$\frac{1}{B(a_1 + x, a_2 + n - x)} p^{a_1+x-1}(1-p)^{a_2+n-x-1}. \tag{1.12}$$

Tato hustota je hustotou rozdělení *Beta*$(X + a_1, n - X + a_2)$.

Dále budeme hledat interval $I_B = [L,U]$ tak, aby platilo

$$\int_0^L \frac{1}{B(a_1 + x, a_2 + n - x)} p^{a_1+x-1}(1-p)^{a_2+n-x-1}\,\mathrm{d}x = \frac{\alpha}{2}. \tag{1.13}$$

a

$$\int_U^1 \frac{1}{B(a_1+x,a_2+n-x)} p^{a_1+x-1}(1-p)^{a_2+n-x-1}\,\mathrm{d}x = \frac{\alpha}{2}. \qquad (1.14)$$

Tedy chceme, aby plocha pod grafem aposteriorní hustoty byla pro $x \in (0,L)$ rovna $\alpha/2$, podobně pro $x \in (U,1)$. Protože pak obsah plochy pod grafem bude pro $x \in (L,U)$ roven $1-\alpha$.

Integrál v (1.13) je roven $P(X \leq L)$. Z toho vyplývá, že L je $(\alpha/2)$-kvantil rozdělení $Beta(X+a_1, n-X+a_2)$, který budeme značit
$Beta(\alpha/2; X+a_1, n-X+a_2)$. Rovnice v (1.14) je ekvivalentní s

$$\int_0^U \frac{1}{B(a_1+x,a_2+n-x)} p^{a_1+x-1}(1-p)^{a_2+n-x-1}\,\mathrm{d}x = 1-\frac{\alpha}{2}. \qquad (1.15)$$

Podobně jako v předchozím případě z (1.15) vyplývá, že U je $(1-\alpha/2)$-kvantil rozdělení $Beta(X+a_1, n-X+a_2)$.
Proto *Bayesův interval spolehlivosti* je dán vztahem

$$I_B = [B(\alpha/2; X+a_1, n-X+a_2), B(1-\alpha/2; X+a_1, n-X+a_2)]. \qquad (1.16)$$

1.6 Jeffreysův interval

Speciálním případem *Bayesova intervalu spolehlivosti* je interval *Jeffreysův*. Jeho konstrukce je založena *Jeffreysově apriorním rozdělení*, které definujeme takto: $\pi_j(p) \propto \sqrt{I(p)}$, kde $I(p) = -\mathsf{E}_p \dfrac{\mathrm{d}^2 \ln P(X|p)}{\mathrm{d}p^2}$ je Fisherova informace.

Nechť $X \sim Bi(n,p), \quad 0 \leq p \leq 1$ a

$$P(X|p) = \tbinom{n}{X} p^X (1-p)^{n-X}.$$

Postupně tuto rovnici zlogaritmujeme a dvakrát zderivujeme:

$$\begin{aligned}
\ln P(X|p) &= \ln \binom{n}{X} + X \ln p + (n-X)\ln(1-p),\\
\frac{\mathrm{d}}{\mathrm{d}p} \ln P(X|p) &= \frac{X}{p} - \frac{n-X}{1-p}, \qquad (1.17)\\
\frac{\mathrm{d}^2}{\mathrm{d}p^2} \ln P(X|p) &= \frac{-X}{p^2} - \frac{n-X}{(1-p)^2}.
\end{aligned}$$

Protože $\mathsf{E}_p X = np$, je

$$
\begin{aligned}
I(p) &= \left[\frac{-X}{p^2} - \frac{n-X}{(1-p)^2}\right] = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} \\
&= \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}.
\end{aligned}
\tag{1.18}
$$

Tedy $\pi_j(p) = \sqrt{I(p)} \propto p^{-1/2}(1-p)^{-1/2}$ je hustota rozdělení $B(1/2{,}1/2)$.

Nyní využijeme poznatky o aposteriorním rozdělení a intervalu spolehlivosti získané v předchozí sekci. V tomto případě se $B(a_1, a_2) = B(1/2, 1/2)$. Po pozorování x úspěchů v n pokusech je aposteriorní rozdělení $B(x + 1/2, n - x + 1/2)$. Je-li $x \neq 0$ a $x \neq n$, je interval

$$
I_J = [B(\alpha/2; x + 1/2{,}n - x + 1/2), B(1 - \alpha/2; x + 1/2, n - x + 1/2)] \tag{1.19}
$$

Jeffreysův interval spolehlivosti.

Příslušné kvantily *Beta rozdělení* je třeba dopočítat numericky, to je ale snadné s využitím programů jako např. Mathematica.

1.7 Clopper-Pearsonův interval

Nechť x je počet úspěchů v n pokusech. Pak hledaný interval je

$$
I_{CP} = [L_{CP}(x){,}U_{CP}(x)],
$$

kde $L_{CP}(x)$ je ta hodnota p, pro kterou platí rovnice

$$
P_p(X \geq x) = \alpha/2,
$$

podobně hodnotu $U_{CP}(x)$ získáme z rovnice

$$
P_p(X \leq x) = \alpha/2.
$$

K řešení těchto rovnic použijeme vztah mezi binomickým a beta rozdělením.

Tvrzení 3. *Nechť $X \sim Bi(n{,}p)$, pak*

$$
P(X \geq x) = \sum_{j=x}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \int_0^p \frac{t^{x-1}(1-t)^{n-x}}{B(x, n-x+1)} \mathrm{d}t. \tag{1.20}
$$

Důkaz. Nejprve připomeňme známé vztahy

$$
B(a_1{,}a_2) = \frac{\Gamma(a_1)\Gamma(a_2)}{\Gamma(a_1 + a_2)}, \qquad \Gamma(a) = (a-1)!.
$$

Proto

$$\int_0^p \frac{t^{x-1}(1-t)^{n-x}}{B(x,n-x+1)}\mathrm{d}t = \frac{n(n-1)\dots(n-x+1)}{(x-1)!}\int_0^p t^{x-1}(1-t)^{n-x}\mathrm{d}t.$$

Tento integrál vypočítáme metodou *per partes*

$$\begin{aligned}
&\frac{n(n-1)\dots(n-x+1)}{(x-1)!}\int_0^p t^{x-1}(1-t)^{n-x}\mathrm{d}t\\
&= \frac{n(n-1)\dots(n-x+1)}{(x-1)!}\frac{p^x}{x}(1-p)^{n-x}\\
&\quad + \frac{n(n-1)\dots(n-x+1)}{(x-1)!}\int_0^p \frac{t^x}{x}(n-x)(1-t)^{n-x-1}\mathrm{d}t\\
&= \binom{n}{x}p^x(1-p)^{n-x} + \frac{n(n-1)\dots(n-x+1)(n-x)}{(x-1)!}\frac{p^{x+1}}{x+1}(1-p)^{n-x+1}\\
&\quad + \frac{n(n-1)\dots(n-x+1)}{(x-1)!}\int_0^p \frac{t^{x+1}}{x+1}(n-x-1)(1-t)^{n-x-2}\mathrm{d}t\\
&= \binom{n}{x}p^x(1-p)^{n-x} + \binom{n}{x+1}p^{x+1}(1-p)^{n-x-1} + \dots\\
&= \sum_{j=x}^{n}\binom{n}{j}p^j(1-p)^{n-j}.
\end{aligned}$$

□

Z předchozího tvrzení vyplývá, že

$$P_p(X \geq x) = \int_0^p \frac{t^{x-1}(1-t)^{n-x}}{B(x,n-x+1)}\mathrm{d}t = \frac{\alpha}{2}, \tag{1.21}$$

tedy p je $\alpha/2 - kvantil\, Beta(x, n-x+1)$. Podobně z platnosti

$$P_p(X \leq x) = 1 - P_p(X > x) = \frac{\alpha}{2}$$

a z předchozího tvrzení plyne, že

$$P_p(X > x) = \int_0^p \frac{t^x(1-t)^{n-x-1}}{B(x+1,n-x)}\mathrm{d}t = 1 - \frac{\alpha}{2}, \tag{1.22}$$

a tedy p je $(1-\alpha/2) - kvantil\, Beta(x+1, n-x)$.
Proto $L_{CP}(x) = B(\alpha/2; x, n-x+1)$ a $U_{CP}(x) = B(1-\alpha/2; x+1, n-x)$.
Tedy

$$I_{CP} = [B(\alpha/2; x, n-x+1), B(1-\alpha/2, x+1, n+x)] \tag{1.23}$$

je *Clopper-Pearsonův interval spolehlivosti.*

1.8 Arkussinový interval

Konstrukce tohoto intervalu je založená na hojně využívané transformaci stabilizující rozptyl pro binomické rozdělení.

Nechť X je náhodná veličina, $\mathsf{E}\, X = \theta, \quad \operatorname{var} X = \sigma^2(\theta)$. Hledáme hladkou funkci g tak, aby rozptyl $\operatorname{var} g(X)$ nezávisel na θ. Z Taylorova rozvoje máme

$$g(X) \doteq g(\theta) + (X-\theta)g'(\theta),$$

a tedy $\mathsf{E}\, g(x) \doteq g(\theta)$ a $\operatorname{var} g(x) \doteq \sigma^2(\theta)[g'(\theta)]^2$.

Rozptyl $\operatorname{var} g(X)$ nebude záviset na θ , bude-li $\sigma(\theta)[g'(\theta)] = c$, kde c je nějaká konstanta. Z toho vyplývá, že $g'(\theta) = \frac{c}{\sigma(\theta)}$. Tedy $g(\theta) = c \int \frac{\mathrm{d}\theta}{\sigma(\theta)}$.

Nechť $\xi \sim Bi(n,p)$. Pak pro $X = \frac{\xi}{n}$ je $\mathsf{E}\, X = p, \operatorname{var} X = \frac{p(1-p)}{n}$.
Počítejme

$$\begin{aligned} g(p) = c \int \frac{\mathrm{d}p}{\sqrt{\frac{p(1-p)}{n}}} &= 2c\sqrt{n} \int \frac{\mathrm{d}p}{2\sqrt{p}\sqrt{1-p}} \\ &= 2c\sqrt{n} \int \frac{\mathrm{d}t}{\sqrt{1-t^2}} = 2c\sqrt{n} \arcsin t = 2c\sqrt{n} \arcsin \sqrt{p}. \end{aligned} \tag{1.24}$$

Volbou $c = \frac{1}{2\sqrt{n}}$, dostáváme $g(p) = \arcsin \sqrt{p}$. Odtud $\mathsf{E}\, g(X) \doteq g(p)$ a $\operatorname{var} g(X) \doteq \frac{1}{4n}$.

Dá se dokázat, že

$$\frac{\arcsin \sqrt{\hat{p}} - \arcsin \sqrt{p}}{\sqrt{\frac{1}{4n}}} \to N(0,1). \tag{1.25}$$

Pak

$$P\left(\frac{|\arcsin \sqrt{\hat{p}} - \arcsin \sqrt{p}|}{\sqrt{\frac{1}{4n}}} \le z \right) \longrightarrow 1-\alpha. \tag{1.26}$$

Postupně upravujeme výše uvedenou nerovnost:

$$
\begin{aligned}
\frac{|\arcsin\sqrt{\hat{p}} - \arcsin\sqrt{p}|}{\sqrt{\dfrac{1}{4n}}} &\leq z, \\
|2\sqrt{n}(\arcsin\sqrt{\hat{p}} - \arcsin\sqrt{p})| &\leq z, \\
\arcsin\sqrt{\hat{p}} - z\frac{1}{2\sqrt{n}} \leq \arcsin\sqrt{p} &\leq \arcsin\sqrt{\hat{p}} + z\frac{1}{2\sqrt{n}}, \qquad (1.27) \\
\sin\left(\arcsin\sqrt{\hat{p}} - z\frac{1}{2\sqrt{n}}\right) \leq \sqrt{p} &\leq \sin\left(\arcsin\sqrt{\hat{p}} + z\frac{1}{2\sqrt{n}}\right), \\
\sin^2\left(\arcsin\sqrt{\hat{p}} - z\frac{1}{2\sqrt{n}}\right) \leq p &\leq \sin^2\left(\arcsin\sqrt{\hat{p}} + z\frac{1}{2\sqrt{n}}\right).
\end{aligned}
$$

Tím jsme získali hledaný interval spolehlivosti:

$$
I_{Arc} = \left[\sin^2\left(\arcsin\sqrt{\hat{p}} - \frac{z}{2\sqrt{n}}\right), \sin^2\left(\arcsin\sqrt{\hat{p}} + \frac{z}{2\sqrt{n}}\right)\right]. \qquad (1.28)
$$

V knize [7] na str. 472 je uvedeno, že nahrazením maximálně věrohodného odhadu parametru $\hat{p}$ odhadem $\check{p} = (X+3/8)/(n+3/4)$ dostaneme lepší stabilizaci rozptylu.

1.9 Interval spolehlivosti poměrem věrohodností

Nechť $X \sim Bi(n,p)$. Nechť dále $\theta = p$. Maximálně věrohodný odhad $\hat{\theta} = \hat{p}$, kde $\hat{p} = \dfrac{X}{n}$ pro $X \neq 0, X \neq n$. Dále mějme hypotézu $H_0 : p_i = p_i^0, \quad i = 1{,}2$, kde p_i^0 jsou kladná čísla splňující podmínku $\sum_{i=1}^{2} p_i^0 = 1$. Pak pro logaritmické věrohodnostní funkce platí

$$
\begin{aligned}
L(\theta) &= \ln\frac{n!}{X!(n-X)!} + \sum_{i=1}^{2} X_i \ln p_i^0, \\
L(\hat{\theta}) &= \ln\frac{n!}{X!(n-X)!} + \sum_{i=1}^{2} X_i \ln\frac{X_i}{n}.
\end{aligned}
$$

Proto

$$
2[L(\hat{\theta}) - L(\theta)] = \sum_{i=1}^{2} X_i \ln\frac{X_i}{np_i^0}
$$

má za platnosti hypotézy H_0 asymptoticky rozdělení χ_1^2. Tedy

$$
2[L(\hat{\theta}) - L(\theta)] \leq \chi_1^2(1-\alpha).
$$

Označme $q = \chi_1^2(1-\alpha)$ a upravujme předchozí nerovnost:

$$
\begin{aligned}
2\left[X\ln\frac{X}{n} + (n-X)\ln\frac{n-x}{n} - X\ln p - (n-X)\ln(1-p)\right] &\le q,\\
X\ln\frac{X}{n} + (n-X)\ln\frac{n-X}{n} - [X\ln p + (n-X)\ln(1-p)] &\le \frac{q}{2},
\end{aligned}
$$

$$
\begin{aligned}
\ln p^X + \ln(1-p)^{n-X} &\ge X\ln\frac{X}{n} + (n-X)\ln\frac{n-X}{n} - \frac{q}{2},\\
\ln p^X(1-p)^{n-X} &\ge X\ln\frac{X}{n} + (n-X)\ln\frac{n-X}{n} - \frac{q}{2},\\
p^X(1-p)^{n-X} &\ge \exp\left\{\left(X\ln\frac{X}{n} + (n-X)\ln\frac{n-X}{n} - \frac{q}{2}\right)\right\}.
\end{aligned}
$$

Pro známé hodnoty n, X a α je na pravé straně výsledné nerovnice konstanta. Tuto nerovnici s neznámým parametrem p lze snadno vyřešit numericky např. pomocí programu Mathematica. Připomeňme, že hledáme řešení pro $p \in (0{,}1)$. Tím dostaneme hledaný interval spolehlivosti.

Kapitola 2

Intervaly spolehlivosti pro parametry multinomického rozdělení

Definice 2. *Nechť* $n, k \in \mathbb{N}$ *a* $p_1, \ldots, p_k$ *jsou nezáporné parametry, pro které platí* $\sum_{j=1}^{k} p_j = 1$. *Nechť dále pro náhodné veličiny* $X_1, \ldots, X_k$ *platí* $\sum_{i=1}^{k} X_i = n$. *Pak*

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \cdots p_k^{x_k}, \tag{2.1}$$

pro nezáporná celá čísla $x_1, \ldots, x_k$, *jejichž součet je* n. *Pravděpodobnostní rozdělení dané vzorcem* (2.1) *se nazývá multinomické. Značíme* $X_1, X_2, \ldots, X_k \sim M(n, p_1, p_2, \ldots, p_k)$.

Chceme-li zkonstruovat intervaly spolehlivosti pro parametry multinomického rozdělení, znamená to, že potřebujeme najít takovou množinu intervalů $\{I_i\}$ $i = 1, \ldots, k$, pro kterou platí, že

$$P\left\{\bigcap_{i=1}^{k} (p_i \in I_i)\right\} \geq 1 - \alpha, \tag{2.2}$$

pro danou hodnotu α. Tedy chceme, aby pravděpodobnost, že každý interval I_i obsahuje příslušnou hodnotu parametru p_i, byla nejméně $1 - \alpha$.

Stejně jako v minulé kapitole bude z označovat kritickou hodnotu normálního rozdělení $N(0,1)$ na hladině $\alpha/2$.

2.1 Intervaly založené na Pearsonově statistice

Nyní se budeme věnovat odvození intervalů spolehlivosti na základě *Pearsonovy statistiky*, které je popsané v [6].

Věta 4. *Je-li* $(X_1, X_2, \ldots, X_k) \sim M(n, p_1, p_2, \ldots, p_k)$, *pak* Pearsonova statistika

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{2.3}$$

má při $n \to \infty$ asymptoticky rozdělení χ^2_{k-1}.

Důkaz. Viz ([2], věta 12.5)

$\square$

Předpokládejme že n je„dostatečně velké". Tedy takové, že i pro nejmenší hodnotu p_i $(i = 1, \ldots, k)$ je np_i alespoň 5. Pak lze tuto aproximaci použít. Proto

$$P\left(\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \leq \chi^2_{\alpha,k-1}\right) \doteq 1 - \alpha, \tag{2.4}$$

kde $\chi^2_{\alpha,k-1}$ značí α-kvantil rozdělení χ^2_{k-1}.
Dosadíme maximálně věrohodný odhad parametru p_i, $\hat{p_i} = \dfrac{X_i}{n}$ a upravujeme

$$\begin{aligned} P\left(\sum_{i=1}^{k} \frac{n\hat{p_i}^2}{p_i} - n \leq \chi^2_{\alpha,k-1}\right) &\doteq 1 - \alpha, \\ P\left(n\sum_{i=1}^{k} \frac{\hat{p_i}^2}{p_i} \leq \chi^2_{\alpha,k-1} + n\right) &\doteq 1 - \alpha. \end{aligned} \tag{2.5}$$

Tím získáme nerovnici

$$\sum_{i=1}^{k} \frac{\hat{p_i}^2}{p_i} \leq \frac{\chi^2_{\alpha,k-1}}{n} + 1, \tag{2.6}$$

která vymezuje pro pevné α, k a n oblast v parametrickém prostoru, jejíž hranice získáme řešením příslušné rovnice

$$\sum_{i=1}^{k} \frac{\hat{p_i}^2}{p_i} = C,$$

kde $C = \dfrac{\chi^2_{\alpha,k-1}}{n} + 1$. Vyjádříme

$$p_i = \frac{\hat{p_i}^2}{C - \sum_{j \neq i} \frac{\hat{p_j}^2}{p_j}}, \quad i = 1,2, \ldots, k$$

Pro nalezení maxima a minima této funkce více proměnných $(p_1, p_2, \ldots, p_k)$ s vazbou

$$\sum_{i=1}^{k} p_i = 1. \tag{2.7}$$

použijeme metodu řešení pomocí *Lagrangeových multiplikátorů*. Proto hledáme maximum a minimum funkce

$$Q = (1 - \lambda)\frac{\hat{p_i}^2}{C - \sum_{j \neq i} \frac{\hat{p_j}^2}{p_j}} - \lambda\left(\sum_{i=1}^{k} p_i - 1\right),$$

kde λ je *Lagrangeův multiplikátor*. Tuto funkci parciálně zderivujeme podle všech proměnných a získané výrazy položíme rovny 0. Vyřešením těchto k rovnic a rovnice (2.7), dostáváme

$$p_i^+, p_i^- = \frac{C + 2\hat{p}_i - 1 \pm \sqrt{(C + 2p_i - 1)^2 - 4C\hat{p}_i{}^2}}{2C},$$

po dosazení za C

$$p_i^+, p_i^- = \frac{\chi^2_{k-1,\alpha} + 2X_i \pm \sqrt{\chi^2_{k-1,\alpha}[\chi^2_{k-1,\alpha} + 4X_i(n - X_i)/n]}}{2(n + \chi^2_{k-1,\alpha})}. \tag{2.8}$$

Proto $I_i = [p_i^-, p_i^+]$ $\quad i = 1,2,\ldots,k$ jsou hledané intervaly spolehlivosti.
Poznamenejme, že hodnoty p_i^-, p_i^+ jsme mohli podle [5] získat i snadnějším způsobem, a to vyřešením kvadratické rovnice pro neznámou p_i

$$(\hat{p}_i - p_i)^2 = \frac{\chi^2_{k-1,\alpha} p_i(1 - p_i)}{n}, \quad i = 1,2,\ldots,k. \tag{2.9}$$

Jestliže $n \to \infty$, je pravděpodobnost, že všechny intervaly spolehlivosti z (2.8) obsahují skutečnou hodnotu parametru, alespoň $1 - \alpha$.

Dále se budeme věnovat modifikacím intervalů z (2.8), které jsou popsané v [5]. Naším cílem bude tyto intervaly zjednodušit a zkrátit (zejména pro běžně používané hodnoty $\alpha = 0.01$, 0.05 a 0.1). Zároveň ale pořád musí platit, že všechny intervaly současně pokrývají příslušnou hodnotu parametru s pravděpodobností alespoň $1 - \alpha$. Pro $k = 2$ a $n \to \infty$, je tato pravděpodobnost rovna $1 - \alpha$, ale pro $k > 2$ je pravděpodobnost vyšší než $1 - \alpha$. Je-li $k = 2$, jedná se o binomické rozdělení. Proto můžeme předpokládat, že $k > 2$. Spolehlivost intervalů z (2.8) zlepšíme, nahradíme-li $\chi^2_k(\alpha)$ kvantilem $\chi^2_1(\alpha/k)$. Pro běžné hodnoty $\alpha = 0.01, 0.05, 0.1$ platí $F < E$, kde $E = \chi^2_k(\alpha)$ a $F = \chi^2_1(\alpha/k)$.

Nyní popíšeme jiný způsob, jak intervaly (2.8) zjednodušit. U binomického rozdělení se často používá nahrazení $p_i(1-p_i)$ výrazem $\hat{p}_i(1-\hat{p}_i)$, pro dostatečně velká n. Provedeme-li nyní tu záměnu v (2.9), pak dostaneme kvadratickou rovnici

$$(\hat{p}_i - p_i)^2 = \frac{\chi^2_{k-1,\alpha} \hat{p}_i(1 - \hat{p}_i)}{N}, \quad i = 1,2,\ldots,k.$$

Její řešení jsou

$$\begin{aligned} \overline{p_i^-} &= \hat{p}_i - \sqrt{\frac{\chi^2_{k-1,\alpha} \hat{p}_i(1 - \hat{p}_i)}{N}}, \\ \overline{p_i^+} &= \hat{p}_i + \sqrt{\frac{\chi^2_{k-1,\alpha} \hat{p}_i(1 - \hat{p}_i)}{N}}, \end{aligned} \tag{2.10}$$

kde $i = 1,2,\ldots,k$. I v tomto případě lze nahradit E hodnotou F. Porovnáním těchto intervalů (2.10), kdy v jednom případě $\chi^2_k(\alpha) = E$ a v druhém případě je $F = \chi_1^2(\alpha/k)$, snadno získáme podíl délek $\sqrt{\dfrac{F}{E}}$. Protože intervaly z (2.8) se asymptoticky rovnají (2.10), jejich poměr je pro $n \to \infty$ také roven $\sqrt{\dfrac{F}{E}}$.

2.2 Intervaly založené na Bonfferoniho nerovnosti

Na začátek připomeňme, že ze vzorce $(\cap A_i)^c = \cup A_i^c$ vyplývá

$$P(\cap A_i) = 1 - P(\cup A_i^c) \geq 1 - \sum P(A_i^c) = 1 - \sum[1 - P(A_i)].$$

Tato nerovnost se nazývá *Bonferroniho*

Nechť $(X_1,X_2,\dots,X_k) \sim M(n,p_1,p_2,\dots,p_k)$. Maximálně věrohodný odhad parametru p_i je $\hat{p}_i = \dfrac{x_i}{n}$, pro $x_i \neq 0, x_i \neq n$.
Označme

$$A_i = \left\{ |\hat{p}_i - p_i| \leq \frac{z}{2\sqrt{n}} \right\}, \qquad B_i = \left\{ |\hat{p}_i - p_i| \leq z\sqrt{\frac{p_i(1-p_i)}{n}} \right\}.$$

Protože platí $p_i(1-p_i) \leq \dfrac{1}{4}$, máme $B_i \subset A_i$, a tedy

$$P(B_i) \leq P(A_i).$$

Počítejme

$$\begin{aligned}
P(A_i) &= P\left\{ \frac{|\hat{p}_i - p_i|}{\sqrt{\frac{p_i(1-p_i)}{n}}} \leq \frac{z}{2\sqrt{n}}\sqrt{\frac{n}{p_i(1-p_i)}} \right\} \\
&= P\left\{ \frac{|\hat{p}_i - p_i|}{\sqrt{\frac{p_i(1-p_i)}{n}}} \leq \frac{z}{2\sqrt{p_i(1-p_i)}} \right\} \\
&\approx \Phi\left(\frac{z}{2\sqrt{p_i(1-p_i)}}\right) - \Phi\left(-\frac{z}{2\sqrt{p_i(1-p_i}}\right) \\
&= 2\Phi\left(\frac{z}{2\sqrt{p_i(1-p_i)}}\right) - 1.
\end{aligned}$$

Proto

$$\begin{aligned}P(\cap A_i) &= 1 - \sum_{i=1}^{k}[1 - P(A_i)] \\ &\approx 1 - k + \sum_{i=1}^{k} 2\Phi\left(\frac{z}{2\sqrt{p_i(1-p_i)}}\right) - k \\ &= 1 - 2k + 2\sum_{i=1}^{k} 2\Phi\left(\frac{z}{2\sqrt{p_i(1-p_i)}}\right) - k \\ &= F(k,\mathbf{p},z).\end{aligned} \tag{2.11}$$

Označme

$$\Pi(k,\mathbf{p},\alpha) = \lim_{n\to\infty} P\left(\bigcap_{i=1}^{k} A_i(\alpha)\right).$$

Z *Bonferroniho nerovnosti* vyplývá, že

$$\begin{aligned}\Pi(k,\mathbf{p},\alpha) &\geq \lim_{n\to\infty} P\left[1 - \sum_{i=1}^{k}[1 - P(A_i)]\right] \\ &= F(k,\mathbf{p},z).\end{aligned}$$

Podle tvrzení, které je dokázané v článku [4], nabývá funkce $F(k,\mathbf{p},z)$ svého minima pro $p_i = 1/k \quad i = 1,2,\ldots,k$.

Položme $p_i = 1/k$ pro $i = 1,\ldots,k$ a označme

$$g(k,\alpha) = 1 - 2k + 2k\Phi\left(\frac{kz}{2\sqrt{k-1}}\right).$$

Pak tedy

$$1 - \sum_{i=1}^{k}[1 - P(A_i)] \approx g(k,\alpha).$$

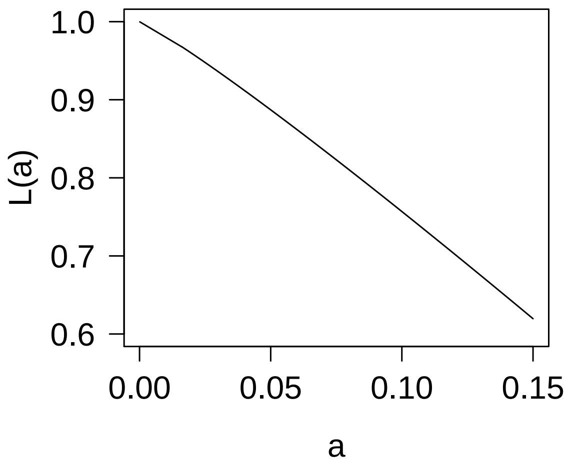
V článku [4] je dokázáno že pro funkci $\Pi(k,\mathbf{p},\alpha)$ a $k = 2,3\ldots$ platí nerovnost

$$\Pi(k,\mathbf{p},\alpha) \geq L(\alpha),$$

kde

$$Ł(\alpha) = \begin{cases} 1 - 2\alpha, & \alpha \leq\ 0.016, \\ 6\Phi\left(\frac{3z(\alpha/2)}{\sqrt{8}}\right) - 5, & 0.016 \leq \alpha \leq 0.150. \end{cases} \tag{2.12}$$

Graf této funkce můžeme vidět na obrázku 2.1, kde a je rovno α.



Obrázek 2.1: Obrázek 1: Funkce L(a)

Tedy 99% intervaly spolehlivosti $p_i \pm \frac{z(0.01/2)}{2\sqrt{n}} \doteq p_i \pm \frac{2.58}{2\sqrt{n}}$ mají pravděpodobnost pokrytí alespoň 0.98 a 95% intervaly $p_i \pm \frac{z(0.05/2)}{2\sqrt{n}} \doteq p_i \pm \frac{1.96}{2\sqrt{n}}$ alespoň 0.8871. Z toho vyplývá, že často používané intervaly spolehlivosti $p_i \pm \frac{1}{\sqrt{n}}$ mají pravděpodobnost pokrytí alespoň 0.89.

Kapitola 3

Simulační studie

V této kapitole budeme pomocí programu R simulovat náhodné výběry z multinomického rozdělení a budeme pro vybrané intervaly spolehlivosti zjišťovat, jaká je celková pravděpodobnost pokrytí, tedy jestli pro každé $i = 1,2,\dots,k$ leží hodnota p_i v příslušném intervalu.

Výsledky těchto simulací budeme zaznamenávat do tabulek, kde n značí rozsah výběru, k je počet tříd multinomického rozdělení, vektor $\mathbf{p} = (p_1,p_2,\dots,p_k)'$ udává skutečné hodnoty parametru. Také si označíme dva vektory. Nejprve

$$\begin{aligned}\triangle &= \mathbf{p}\\ &= (0.2045,0.1865,0.1461,0.1199,0.0772,0.0688,0.0678,0.0319,0.0266,0.0246,0.0151)'\end{aligned}$$

Tento vektor jsme nezvolili náhodně, jedná se o výsledek parlamentních voleb z roku 2013 a jsou zde zaznamenány pouze výsledky stran, které získaly ve volbách více než jedno procento hlasů. Dále vektor
$\clubsuit = \mathbf{p} = (0.2421,0.2340,0.1635,0.1612,0.0684,0.0495,0.0323,0.0246,0.0239)'$ znázorňuje výsledky prvního kola prezidentských voleb z roku 2013.

3.1 Simulace pro intervaly založené na Bonferroniho nerovnosti

Nejprve uveďme kód podle kterého budeme simulace provádět.

```
z<-qnorm(1-a/2)
o<-10000
set.seed(78)
m<-0
for (i in 1:o)
  {x=rmultinomial(n,p)

      h<-1
     for (j in 1:k){if (abs((x[1,j]/n)-p[1,j])>(z/(2*sqrt(n)))){h=h+1}}
       if (h==1){m=m+1}
        }
m
```

Generovali jsme 10 000 náhodných výběrů z multinomického rozdělení každý o rozsahu n pozorování. Pro každý tento výběr a pro každou z k-tříd multinomického rozdělení jsme pomocí vzorce $\frac{X_i}{n} \pm \frac{z}{2\sqrt{n}}$ spočítali interval spolehlivosti a zjišťovali, jestli v něm leží skutečná hodnota parametru p_i, pro $i = 1, \ldots, k$. Tedy ověřovali jsme platnost nerovnosti $|X_i/n - p_i| \leq \frac{z}{2\sqrt{n}}$. Byla-li tato nerovnost splněna ve všech k-případech, označili jsme tento výběr za „úspěšný". Podíl počtu „úspěšných" výběrů ku všem výběrům vyjadřuje pravděpodobnost pokrytí. Výsledky jsou zaznamenány v tabulce, kde je navíc v posledním sloupci uvedena i hodnota funkce $L(\alpha)$, kterou jsme popisovali v sekci 2.2.

α	n	k	$\mathbf{p} = (p_1,p_2,\ldots,p_k)'$	**Pravděpodobnost pokrytí**	$L(\alpha)$
0.01	100	3	(1/3,1/3,1/3)	0.9859	0.98
0.01	100	11	△	0.9958	0.98
0.01	100	9	♣	0.9948	0.98
0.05	100	3	(1/3,1/3,1/3)	0.9161	0.8871
0.05	100	3	(2/3,1/6,1/6)	0.9604	0.8871
0.05	100	4	(1/4,1/4,1/4,1/4)	0.9073	0.8871
0.05	100	5	(1/5,1/5,1/5,1/5,1/5)	0.9207	0.8871
0.05	100	11	△	0.9552	0.8871
0.05	100	9	♣	0.9502	0.8871
0.05	500	3	(1/3,1/3,1/3)	0.9054	0.8871
0.05	500	3	(2/3,1/6,1/6)	0.9537	0.8871
0.05	500	4	(1/4,1/4,1/4,1/4)	0.9100	0.8871
0.05	500	5	(1/5,1/5,1/5,1/5,1/5)	0.9259	0.8871
0.05	500	11	△	0.9417	0.8871
0.05	500	9	♣	0.9442	0.8871
0.1	100	3	(1/3,1/3,1/3)	0.8003	0.7568
0.1	100	3	(2/3,1/6,1/6)	0.8753	0.7568
0.1	100	4	(1/4,1/4,1/4,1/4)	0.8451	0.7568
0.1	100	5	(1/5,1/5,1/5,1/5,1/5)	0.8528	0.7568
0.1	100	11	△	0.8611	0.7568
0.1	100	9	♣	0.8569	0.7568
0.1	500	3	(1/3,1/3,1/3)	0.8163	0.7568
0.1	500	3	(2/3,1/6,1/6)	0.8935	0.7568
0.1	500	4	(1/4,1/4,1/4,1/4)	0.8212	0.7568
0.1	500	5	(1/5,1/5,1/5,1/5,1/5)	0.8340	0.7568
0.1	500	11	△	0.8486	0.7568
0.1	500	9	♣	0.8634	0.7568
0.15	100	3	(1/3,1/3,1/3)	0.7034	0.6196
0.15	100	4	(1/4,1/4,1/4,1/4)	0.7471	0.6196
0.15	500	3	(1/3,1/3,1/3)	0.7194	0.6196
0.15	500	4	(1/4,1/4,1/4,1/4)	0.7315	0.6196

Tabulka 3.1: Pravděpodobnost pokrytí pro intervaly spolehlivosti

Z tabulky můžeme vidět, že pravděpodobnost pokrytí byla vždy vyšší než hodnota $L(\alpha)$. Nejblíže příslušné hodnotě $L(\alpha)$ byla hodnota pravděpodobnosti pokrytí u trinomického rozdělení při výběru o rozsahu $n = 100$ a pro $\alpha = 0.01$. Dále vidíme, že nejnižší hodnota pravděpodobnosti pokrytí byla vždy u již zníměného trinomického rozdělení. Naopak nejvyšší hodnotu jsme zaznamenali u rozdělení daných vektory $(2/3,1/6,1/6)'$, $\triangle$, $\clubsuit$.

3.2 Simulace pro Waldův interval

Pro simulace použijeme následující kód:

```
z<-qnorm(1-a/2)
o<-10000
set.seed(78)
m<-0
for (i in 1:o)
{x=rmultinomial(n,p)

 h<-1
 for (j in 1:k){if (abs((x[1,j]/n)-p[1,j])>
            (z*(sqrt((x[1,j]/n*(1-(x[1,j])/n))/n))))
 {h=h+1}}
 if (h==1){m=m+1}
}
m
```

Jako v minulém případě jsme generovali jsme 10 000 náhodných výběrů z multinomického rozdělení každý o rozsahu, v tomto případě $n = 500$ pozorování. Pro každý tento výběr a pro každou z k-tříd multinomického rozdělení jsme počítali pomocí vzorce pro Waldův interval spolehlivosti, viz (1.5). Pro zjištění, zda v něm leží skutečná hodnota parametru p_i, pro $i = 1, \ldots, k$, jsme ověřovali platnost nerovnosti $|X_i/n - p_i| \leq z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Byla-li tato nerovnost splněna ve všech k-případech, označili jsme tento výběr za „úspěšný“ a dopočítali jsme pravděpodobnost pokrytí. Výsledky najdeme v tabulce 3.2.

α	k	$\mathbf{p} = (p_1,p_2,\ldots,p_k)'$	**Pravděpodobnost pokrytí**
0.01	3	(1/3,1/3,1/3)	0.9726
0.01	3	(2/3,1/6,1/6)	0.9667
0.01	11	$\triangle$	0.8268
0.01	9	$\clubsuit$	0.8601
0.05	3	(1/3,1/3,1/3)	0.8735
0.05	3	(2/3,1/6,1/6)	0.8656
0.05	11	$\triangle$	0.5502
0.05	9	$\clubsuit$	0.5925
0.1	3	(1/3,1/3,1/3)	0.7804
0.1	3	(2/3,1/6,1/6)	0.7700
0.1	11	$\triangle$	0.2808
0.1	9	$\clubsuit$	0.3973

Tabulka 3.2: Pravděpodobnost pokrytí pro Waldův interval

Z tabulky vidíme, že pro vektory (1/3,1/3,1/3), $(2/3, 1/6, 1/6)$ je hodnota pravděpodobnosti pokrytí vysoká, ale o málo nižší než v předchozím případě.

Naopak pro vektory $\triangle$, $\clubsuit$ je tato hodnota velmi malá, a to hlavně v případě $\alpha = 0.05$ a 0.1.

3.3 Simulace pro Agresti-Coullův interval

Opět uvedeme kód, který použijeme:

```
z<-qnorm(1-a/2)
o<-10000
set.seed(78)
m<-0
for (i in 1:o)
{x=rmultinomial(n,p)

 h<-1
 for (j in 1:k){if (abs(((x[1,j]+y)/(n+2*y))-p[1,j])>
               (z*sqrt(( ((x[1,j]+y)/(n+2*y))
            *(1-((x[1,j]+y)/(n+2*y))))/(n+2*y))))
 {h=h+1}}
 if (h==1){m=m+1}
}
m
```

Postupovali jsme naprosto obdobně jako v předchozím případě. Ověřovali jsme platnost nerovnosti $|(X_i + y)/(n + 2y) - p_i| \leq z\sqrt{\dfrac{\dfrac{X_i + y}{n + 2y}\left(1 - \dfrac{X_i + y}{n + 2y}\right)}{n + 2y}}$, kde $y = z^2$ vycházející z (1.11).

α	k	$\mathbf{p} = (p_1,p_2,\ldots,p_k)'$	**Pravděpodobnost pokrytí**
0.01	3	(1/3,1/3,1/3)	0.9725
0.01	3	(2/3,1/6,1/6)	0.9740
0.01	11	$\triangle$	0.8733
0.01	9	$\clubsuit$	0.9298
0.05	3	(1/3,1/3,1/3)	0.8835
0.05	3	(2/3,1/6,1/6)	0.8858
0.05	11	$\triangle$	0.5235
0.05	9	$\clubsuit$	0.6559
0.1	3	(1/3,1/3,1/3)	0.7804
0.1	3	(2/3,1/6,1/6)	0.7864
0.1	11	$\triangle$	0.2833
0.1	9	$\clubsuit$	0.4196

Tabulka 3.3: Pravděpodobnost pokrytí pro Agresti-Coullův interval

Pohledem do tabulky zjistíme, že jsem získali podobné výsledky jako u *Waldova intervalu.* Například pro vektor $\clubsuit$ je pravděpodobnost pokrytí vždy vyšší, ale pořád je pro hodnoty $\alpha = 0.05$ a 0.1 výrazně nižší než u intervalů ze sekce 3.1.

Závěr

V této práci jsem popsala a odvodila intervaly spolehlivosti pro parametry binomického a multinomického rozdělení. V kapitole 3 jsem se věnovala simulacím multinomického rozdělení a následnému výpočtu pravděpodobnosti pokrytí pro konkrétní intervaly spolehlivosti.

Podrobně jsem tuto pravděpodobnost počítala pro interval spolehlivosti založený na *Bonferroniho nerovnosti*, interval $\frac{X_i}{n} \pm \frac{z}{2\sqrt{n}}$, jehož odvození jsem se věnovala v sekci 2.2. Pravděpodobnost pokrytí vycházela „poměrně“ vysoká, např. pro běžně používanou hladinu $\alpha = 0.05$ byla její hodnota vždy vyšší než 0.9, a vždy byla vyšší než hodnota funkce $L(\alpha)$. Proto je vhodné pro výpočet intervalů spolehlivosti pro parametry multinomického rozdělení používat intervaly $\frac{X_i}{n} \pm \frac{1}{\sqrt{n}}$. Tyto intervaly lze snadno spočítat a zároveň jejich pravděpodobnost pokrytí je alespoň 0.89.

Dále jsem zjišťovala, jaká je pravděpodobnost pokrytí, konstruujeme-li intervaly spolehlivosti pro parametry multinomického rozdělení individuálně pro každou pravděpodobnost. V tomto případě jsem vybrala *Waldův interval* a *Agresti-Coullův interval*. U obou vycházely pravděpodobnosti pokrytí podobně. V případě trinomického rozdělení byla její hodnota podobná jako u výše popsaného intervalu. Naproti tomu u příkladu z praxe - výsledků minulých voleb - byla pravděpodobnost pokrytí velmi nízká, pro $\alpha = 0.05$ se její hodnota pohybovala jen mezi 0.52 a 0.65.

Literatura

[1] A. Agresti and B. A. Coull. Aproximation is better than "exact"for interval estimation of binomial proportion. *The American Statistician*, 52:119–126, 1998.

[2] J. Anděl. *Základy matematické statistiky.* Druhé opravené vydání. Matfyzpress, Praha, 2007.

[3] L.D. Brown, T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16:101–133, 2001.

[4] S. Fitzpatrick and A. Scott. Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82(399):875–878, 1978.

[5] L. A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7(2):247–254, 1965.

[6] C. P. Quesenberry and D. C. Hurst. Large sample simultaneous confidence intervals. *Technometrics*, 6(2):191–195, 1964.

[7] Radhakrishna C. Rao. *Lineární metody statistické indukce a jejich aplikace.* 1. vydání překladu. Academia, Praha, 1978.

[8] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212, 1927.