Přehled metod odhadu statistické chyby ve výběrových šetřeních

Martin Anděl, Rostislav Černý, Pavel Charamza, Jan Neustadt

22. července 2005

1 Úvod

V článku se budeme zabývat problematikou velikosti chyby odhadů vznikajících v oblasti výzkumů veřejného mínění. Jak již naznačuje název článku, budeme se zabývat chybou statistickou, tj. pomineme diskuse okolo chyb systematických vznikajících například odmítáním odpovědí specifickými částmi populace apod. Metody odhadu statistické chyby popisované v článku jsou pochopitelně použitelné i v jiných praktických situacích, než jsou výše uvedené výzkumy veřejného mínění. Článek vznikl na základě problémů, se kterými se setkávali jeho autoři při hledání správných cest pro řešení uvedené problematiky. Článek neobsahuje nové matematické poznatky, je spíše inspirován negativními zkušenostmi, se kterými se autoři setkali při snaze aplikovat známé postupy a teorii do praxe. Negativními v tom smyslu, že obtížně hledali všeobecná doporučení, která by šla použít v praxi velmi častých situacích. Proto se v závěru článku autoři pokoušejí tato doporučení stanovit, i když si jsou vědomi jisté ošidnosti obecných doporučení pro praxi, která může být mírně odlišná. Rovněž si nekladou nárok na úplné pokrytí problematiky, protože některé problémy, například odhady chyb při závislých pozorováních v šetřeních jednotlivců na základě odpovědí celých domácností nebo podrobnější studium vážených odhadů jsou velmi komplikované a teoreticky pravděpodobně neodvoditelné.

V kapitole 2 jsou uvedeny varianty pro postup v nejjednodušších případech, kdy odhadujeme relativní či absolutní četnosti výskytu nějakého jevu v konečné nebo aproximativně nekonečné populaci. V kapitole 3 problematiku zobecňujeme na porovnání dvou odhadů, ať již vzájemně závislých nebo nezávislých. V kapitole 4 se zabýváme odhadem chyby v případě obecnějších odhadů typu průměr, úhrn či poměr úhrnů. V kapitole 5 potom připomínáme teorii skupinkových výběrů a z nich plynoucích odhadů chyb jako možnou motivaci pro vylepšení stávajících postupů. V posledních dvou kapitolách potom uvádíme postupy použitelné v případech, kdy uspořádání výběru, resp. jiná omezení neumožňují použít postupy předchozích kapitol. V závěru pak velmi stručně shrnujeme, které ze všech uvedených vzorců použít v praktických situacích.

Hned v úvodu bychom rádi poděkovali recenzentům za detailní úsilí, které

věnovali přečtení tohoto článku a za jejich komentáře a poznámky, které pomohly k doplnění, zpřesnění a opravě původní verze.

2 Relativní a absolutní četnost v neváženém vzorku

2.1 Modelování hypergeometrickým rozdělením

V tomto odstavci budeme předpokládat, že chceme zjistit množství prvků s určitou vlastností v daném základním (konečném) souboru. Velikost tohoto konečného souboru označíme N a neznámý počet prvků s hledanou vlastností v tomto souboru M. Odhad tohoto neznámého počtu provedeme na základě prostého náhodného výběru, jehož rozsah označíme n. Hodnoty pozorování budeme značit jako X_i, $i = 1, \ldots, n$. Veličiny X_i, $i = 1, \ldots, n$ nabývají hodnot 1, nebo 0, podle toho, zda vybraný prvek má, nebo nemá hledanou vlastnost. Počet vybraných prvků s hledanou vlastností, tj. $\sum_{i=1}^{n} X_i$ má hypergeometrické rozdělení s parametry N, M a n, tj.

$$p_k = \mathrm{P}[\sum_{i=1}^{n} X_i = k] = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},\ k = \max\left(0, M-(N-n)\right), \ldots, \min(M, n). \tag{1}$$

Pro zkrácení zápisu budeme dále označovat $X = \sum_{i=1}^{n} X_i$.

Předpokládáme, že parametry N a n jsou známy. Odhadujeme obvykle dva neznámé parametry základního souboru:

- relativní četnost prvků s danou vlastností, tj. parametr $p = \frac{M}{N}$,
- absolutní četnost prvků s danou vlastností, tj. parametr M.

Odhad parametru p založíme na statistice

$$\widehat{p} = \frac{X}{n}, \tag{2}$$

odhad absolutní četnosti na statistice

$$\widehat{M} = \frac{N}{n} X. \tag{3}$$

2.1.1 Vlastnosti odhadů

1. Střední hodnota náhodné veličiny X je dána vzorcem

$$\mathrm{E}\, X = n\frac{M}{N}, \tag{4}$$

její rozptyl má hodnotou

$$\mathrm{var}\, X = \frac{N-n}{N-1} n \frac{M}{N}\left(1 - \frac{M}{N}\right). \tag{5}$$

Z uvedených vlastností vyplývá, že $\mathrm{E}\,\widehat{p} = \frac{M}{N}$ a $\mathrm{var}\,\widehat{p} = \frac{N-n}{N-1}\frac{\frac{M}{N}\left(1-\frac{M}{N}\right)}{n}$.

2. Odhad $\widehat{p}$ je nestranným odhadem parametru p a jeho rozptyl se blíží k nule, když n se blíží k N.

3. Z rovnic (4) a (5) plyne, že odhad $\widehat{M}$ je nestranným odhadem parametru M a jeho rozptyl je roven

$$\operatorname{var}\widehat{M} = N^2 \frac{N-n}{N-1} \frac{\frac{M}{N}\left(1-\frac{M}{N}\right)}{n}.$$

2.1.2 Konstrukce intervalu spolehlivosti pro parametry p a M

Hledáme interval $[\underline{p}, \overline{p}]$ takový, že pokrývá skutečnou hodnotu parametru p s pravděpodobností alespoň $1-\alpha$, tj. že $\mathrm{P}_p\left[\underline{p} \le p \le \overline{p}\right] \ge 1-\alpha$, kde α je vhodně zvolené malé číslo (obvykle 0,05, méně často 0,01). Můžeme postupovat například následujícím způsobem, který je analogií postupu z [7]:

Označme $F(x,p,n) = \sum_{k=0}^{x} p_k$ a podobně $\overline{F}(x,p,n) = \sum_{k=x}^{n} p_k$. Závislost F, resp. $\overline{F}$ na parametrech p, n je dána prostřednictvím p_k.

Stanovíme pro dané $x \in \{1,\ldots,n-1\}$ dolní hranici intervalu spolehlivosti pro parametr p jako největší hodnotu $\underline{p}$ z množiny $\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$ takovou, že

$$\overline{F}(x,\underline{p},n) \le \frac{\alpha}{2} < \overline{F}(x-1,\underline{p},n),$$

a horní hranici intervalu spolehlivosti jako nejmenší hodnotu $\overline{p}$ ze stejné množiny, pro kterou

$$F(x,\overline{p},n) \le \frac{\alpha}{2} < F(x+1,\overline{p},n).$$

Interval $(\underline{p}, \overline{p})$ je potom intervalem spolehlivosti pokrývající skutečnou hodnotu p s pravděpodobností alespoň $1-\alpha$. Interval spolehlivosti pro parametr M lze stanovit jako

$$\left[N\underline{p}, N\overline{p}\right]. \tag{6}$$

V případě, že $x=0$, resp. $x=n$, lze zkonstruovat podobně intervaly spolehlivosti $[0,\overline{p})$ resp. $(\underline{p},1]$, kde pravděpodobnost pokrytí správné hodnoty je alespoň $1-\frac{\alpha}{2}$.

Poznamenejme, že pro stanovení hodnot $\underline{p}$ a $\overline{p}$ lze využít vztahy

$$\underline{p} = \max\{p; \overline{F}(x,p,n) \le \frac{\alpha}{2}\}, \tag{7}$$

$$\overline{p} = \min\{p; F(x,p,n) \le \frac{\alpha}{2}\}. \tag{8}$$

To vyplývá z vlastnosti hypergeometrického rozdělení, že funkce F je pro libovolné pevné x nerostoucí funkcí p. Tuto vlastnost mají i další rozdělení popisovaná v tomto článku, která se používají pro aproximaci intervalu spolehlivosti. Kompletní rozbor tohoto problému lze nalézt například v [27].

Na grafu 2.1 jsou zobrazeny některé speciální případy intervalů spolehlivosti na základě popsané metody. Jsou zde uvedeny intervaly spolehlivosti pro hypotetickou populaci o velikosti $N = 10000$ jednotek s různou velikostí rozsahu

výběru n. Intervaly spolehlivosti se pochopitelně zužují s rozsahem výběru. Interval s nulovou šířkou (tedy stoprocentně přesný odhad) získáváme pro $n = N$. Je důležité poznamenat, že intervaly spolehlivosti se dále výrazně nemění s rostoucí hodnotou N. Platí zde proto aproximace, že pro dostatečně velkou populaci (stačí i méně než 10x větší než velikost výběru) jsou intervaly spolehlivosti v podstatě identické. Proto se používá v praxi místo hypergeometrického rozdělení modelování rozdělením binomickým, jak popisuje následující kapitola 2.2.

Na úplný závěr tohoto odstavce poznamenejme, že skutečné pravděpodobnosti pokrytí správné hodnoty parametru p jsou vzhledem k diskrétnosti hypergeometrického rozdělení většinou větší než hodnota $1 - \alpha$, tedy že konstruované intervaly spolehlivosti jsou konzervativnější. Tento jev se týká i aproximací, o kterých budeme mluvit v následujících kapitolách. O problematice skutečné pravděpodobnosti v případě normální aproximace potom hovoří ilustrace v odstavci 2.5.

2.2 Modelování binomickým rozdělením

V případě, že N je dostatečně veliké a přitom n nepříliš veliké ve srovnání s N, můžeme předchozí úvahy převést snadno pomocí limitního přechodu $N \to \infty$ na následující vyjádření:

$$p_k = \mathrm{P}[\sum_{i=1}^{n} X_i = k] = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n \tag{9}$$

$$\mathrm{E}\, X = np, \quad \mathrm{var}\, X = np(1-p) \tag{10}$$

$$\mathrm{E}\, \widehat{p} = p, \quad \mathrm{var}\, \widehat{p} = \frac{p\,(1-p)}{n}. \tag{11}$$

Interval spolehlivosti lze potom stanovit stejným postupem jako v předchozí kapitole s tím, že p_k ze vztahu (1) nahradíme hodnotou ze vztahu (9). V případě binomického modelu lze navíc s výhodou použít pro výpočet vztahu

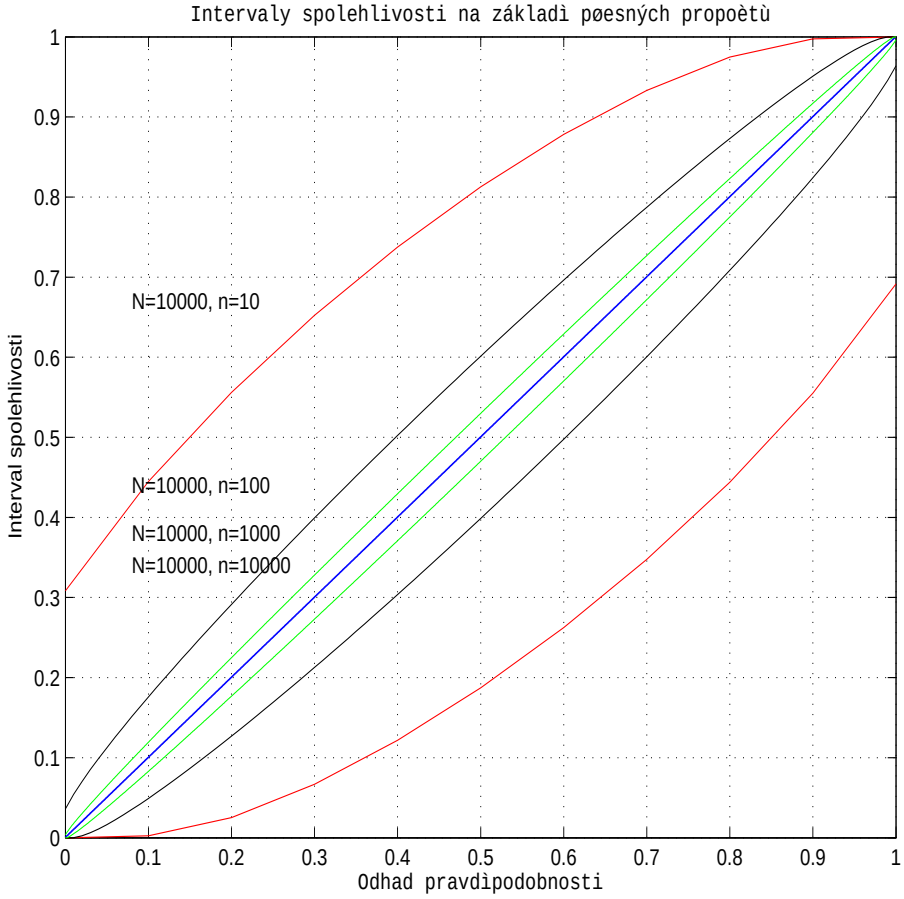
$$B(k, p, n) = F(k, p, n) = F^*_{2(n-k),2(k+1)} \left(\frac{(k+1)\,(1-p)}{p\,(n-k)} \right), \tag{12}$$

kde $F^*_{m,n}(x)$ je hodnota distribuční funkce Fisherova–Snedecorova rozdělení o m a n stupních volnosti. Uvedený vzorec lze nalézt například v [2]. V této knize lze najít další citace vztahující se k tomuto problému a některá další doporučení. Na základě tohoto vztahu lze odvodit, že

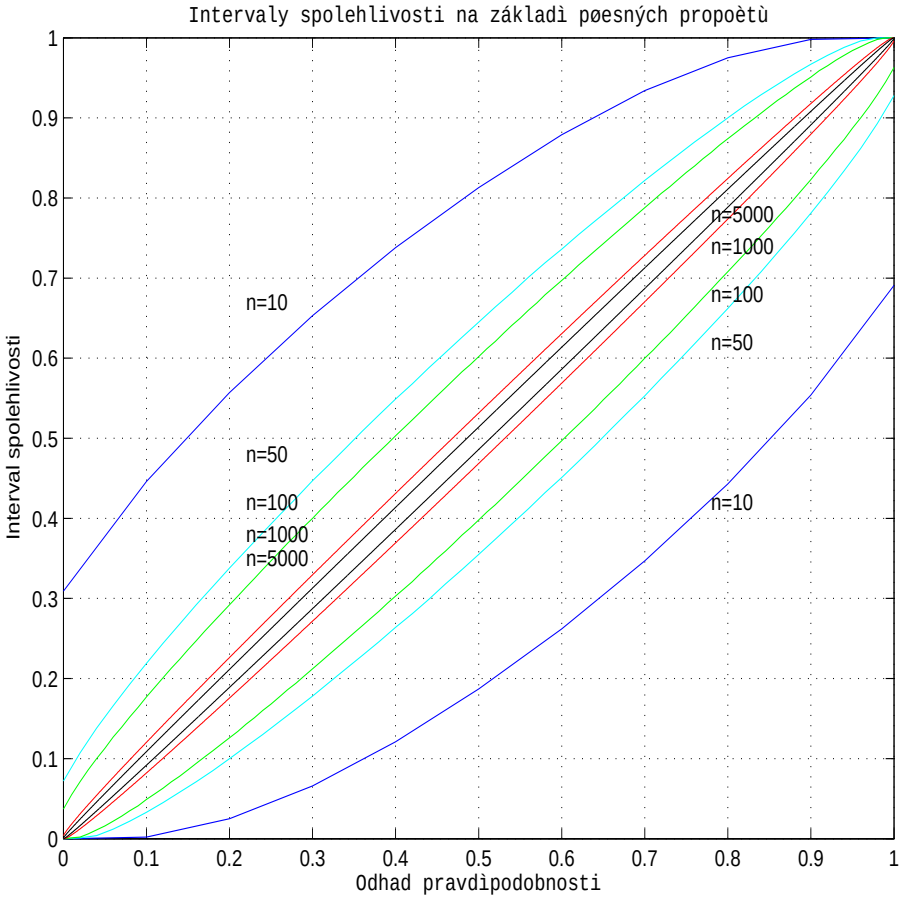
$$\underline{p} = \frac{x}{x + (n - x + 1) \cdot F^{*-1}_{2(n-x+1),2x}(1 - \frac{\alpha}{2})}$$

$$\overline{p} = \frac{(x+1) \cdot F^{*-1}_{2(x+1),2(n-x)}(1 - \frac{\alpha}{2})}{n - x + (x+1) \cdot F^{*-1}_{2(x+1),2(n-x)}(1 - \frac{\alpha}{2})},$$

kde $F^{*-1}_{m,n}(\gamma)$ je $100\gamma\%$ kvantil F rozdělení o m, n stupních volnosti.



Graf 2.1: Intervaly spolehlivosti pro různé velikosti výběru a populace na základě hypergeometrického modelu



Graf 2.2: Intervaly spolehlivosti pro binomický model pro různé velikosti výběru

Ilustraci šířky intervalu spolehlivosti na základě binomické aproximace udává graf 2.2.

Na základě laskavého upozornění jednoho z recenzentů uvádíme ještě podobnou aproximaci, která má výhodu v přesnějším posouzení hodnot distribuční funkce hypergeometrického rozdělení. Při této aproximaci (viz [24], [27]) je

$$\mathrm{P}[\sum_{i=1}^{n} X_i \leq x] \approx \sum_{k=0}^{x} \binom{n}{k} p^{*k} \left(1-p^*\right)^{n-k}, \tag{13}$$

kde

$$p^* = \frac{M - \frac{x}{2}}{N - \frac{n-1}{2}}. \tag{14}$$

Interval spolehlivosti lze potom zkonstruovat podle následujícího postupu:

1. Nalezněme řešení $\overline{p^*}$, resp. $\underline{p^*}$ následujících rovnic vzhledem k proměnné p.

$$B(x,p,n) = \frac{\alpha}{2}, \quad \text{resp.} \quad 1 - B(x-1,p,n) = \frac{\alpha}{2}. \tag{15}$$

2. Vypočítejme $\underline{M}$, $\overline{M}$ z rovnic

$$\underline{M} = \frac{x}{2} + \underline{p^*}\left(N - \frac{n-1}{2}\right),$$

$$\overline{M} = \frac{x}{2} + \overline{p^*}\left(N - \frac{n-1}{2}\right).$$

3. Stanovme hodnoty $\underline{p} = \frac{\underline{M}}{N}$ a $\overline{p} = \frac{\overline{M}}{N}$.

Poznamenejme, že pro řešení rovnic (15) lze opět s výhodou použít vzorce (12).

2.3 Modelování Poissonovým rozdělením

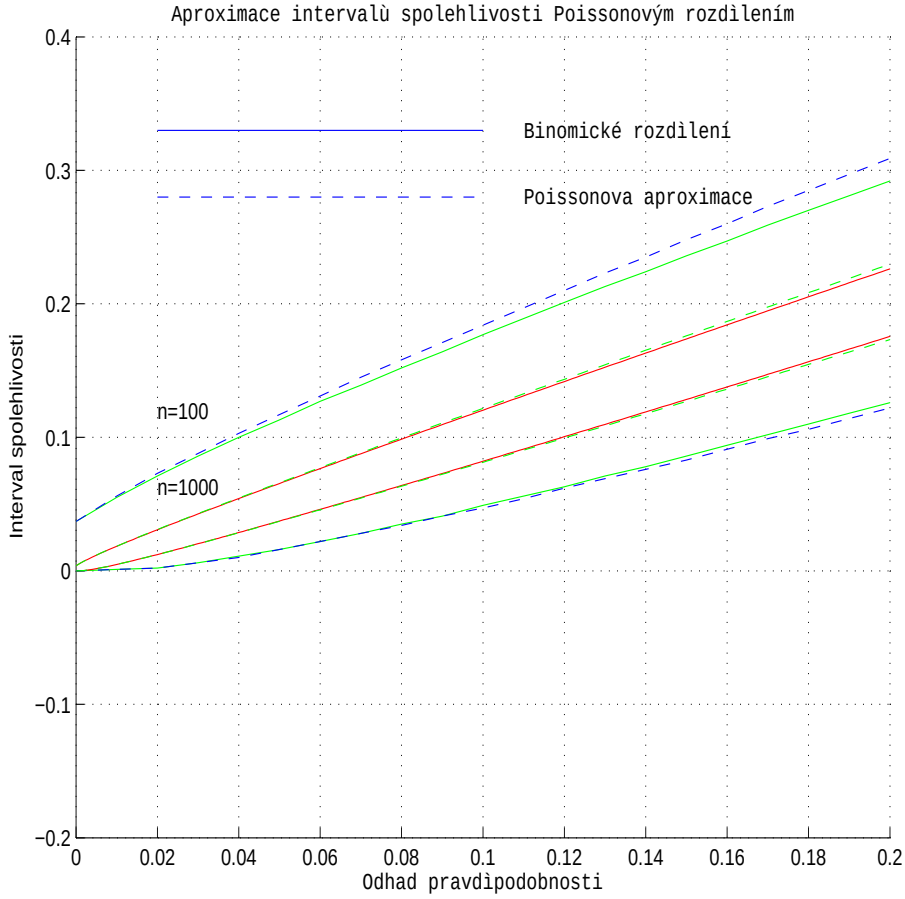
V případě, že n je dostatečně veliké ($n \geq 50,\ n \leq 0,1N$) a naopak odhady parametru p vycházejí malé ($\widehat{p} \leq 0,1$), lze dále ještě aproximovat následujícím způsobem:

$$p_k = \mathrm{P}[\sum_{i=1}^{n} X_i = k] \approx \frac{(np)^k}{k!} \exp(-np), \tag{16}$$

$$\mathrm{E}\, X \approx np, \quad \mathrm{var}\, X \approx np, \tag{17}$$

$$\mathrm{E}\, \widehat{p} = p, \quad \mathrm{var}\, \widehat{p} = \frac{p}{n}. \tag{18}$$

Interval spolehlivosti lze potom stanovit stejným postupem jako v předchozích kapitolách s tím, že p_k ze vztahu (1) nahradíme hodnotou ze vztahu (16). Ilustraci šířky intervalu spolehlivosti na základě aproximace Poissonovým rozdělením je možno spatřit na grafu 2.3. Při stanovení intervalů spolehlivosti se



Graf 2.3: Intervaly spolehlivosti pro porovnání binomického modelu a Poissonovy aproximace

potom dá využít vztahu

$$F(k,p,n) = 1 - \chi^2_{2(k+1)}\left(2np\right), \tag{19}$$

kde $\chi^2_n(x)$ je hodnota distribuční funkce rozdělení χ^2 o n stupních volnosti v bodě x (viz např. [16]). Je proto

$$\underline{p} = \frac{1}{2n}\left(\chi^2\right)^{-1}_{2(x+1)}\left(\frac{\alpha}{2}\right)$$

$$\overline{p} = \frac{1}{2n}\left(\chi^2\right)^{-1}_{2(x+1)}\left(1-\frac{\alpha}{2}\right),$$

kde $\left(\chi^2\right)^{-1}_f(\gamma)$ je $100\gamma\%$ kvantil χ^2 rozdělení o f stupních volnosti.

2.4 Aproximace normálním rozdělením

V případě, že n je dostatečně velké, lze na základě centrální limitní věty použít další aproximaci rozdělení náhodné veličiny $X = \sum_{i=1}^{n} X_i$. Obvykle se uvažuje vhodnost této aproximace v případě, kdy $np(1-p) \geq 10$. K tomuto problému viz též 4.2.1. Díky centrální limitní větě platí

$$\mathrm{P}\left[\frac{X - \mathrm{E}\,X}{\sqrt{\mathrm{var}\,X}} \leq x\right] \approx \Phi(x),$$

kde Φ je distribuční funkce normálního rozdělení se střední hodnotou 0 a rozptylem 1. Je tedy vzhledem k odvozeným hodnotám střední hodnoty a rozptylu odhadu $\widehat{p}$

$$\lim_{n\to\infty} \mathrm{P}\left[|\widehat{p} - p| \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right] = 1-\alpha, \tag{20}$$

kde $z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ je $100(1-\frac{\alpha}{2})\%$ kvantil normálního rozdělení $N(0,1)$. Odtud dostáváme velmi často používaný přibližný tvar intervalu spolehlivosti $[\underline{p}, \overline{p}]$, kde

$$\begin{aligned} \underline{p} &= \widehat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \\ \overline{p} &= \widehat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}. \end{aligned} \tag{21}$$

Při volbě $\alpha = 0,05$ je hodnota $z_{1-\frac{\alpha}{2}}$ přibližně rovna 1,96, místo toho se často nahrazuje méně přesnou hodnotou 2. Vzorec (20) lze zpřesnit použitím přesného rozptylu $\widehat{p}$ na základě hypergeometrického rozdělení. Ve vzorcích (20), resp. (21) se nahradí výraz $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ výrazem $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\sqrt{\frac{N-n}{N-1}}$. Toto zpřesnění se nazývá konečnostním zpřesnění. Dále je potřeba poznamenat, že nestranným odhadem hodnoty $p(1-p)$ není hodnota $\widehat{p}(1-\widehat{p})$, ale hodnota $\frac{n}{n-1}\widehat{p}(1-\widehat{p})$, je proto správnější použít ve vzorcích (21) ve jmenovatelích hodnotu $n-1$. Pravda je,

že při rozsazích výběrů, jakých se zpravidla používá je výsledné zpřesnění bezvýznamné.

Zajímavou variantu pro odhad intervalu spolehlivosti je možno najít např. v [31]. Vzorec vychází opět z aproximace normálním rozdělením, tentokrát však ve tvaru

$$\lim_{n\to\infty} \mathrm{P}\left[|\widehat{p}-p| \le z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right] = 1-\alpha. \tag{22}$$

Označme pro zkrácení zápisu $z := z_{1-\frac{\alpha}{2}}$. Ekvivalentními úpravami (22) lze dosáhnout tvaru $\lim_{n\to\infty} \mathrm{P}\left[p \in [\underline{p},\, \overline{p}]\right] = 1-\alpha$, kde

$$\begin{aligned} \underline{p} &= \frac{2n\widehat{p} + z^2 - z\sqrt{4n\widehat{p}\,(1-\widehat{p}) + z^2}}{2\,(n+z^2)}, \\ \overline{p} &= \frac{2n\widehat{p} + z^2 + z\sqrt{4n\widehat{p}\,(1-\widehat{p}) + z^2}}{2\,(n+z^2)}. \end{aligned} \tag{23}$$

Blyth a Still v [4] uvádějí alternativní aproximaci s tzv. spojitostní korekcí, kde se interval spolehlivosti bere jako množina p, pro kterou platí

$$\lim_{n\to\infty} \mathrm{P}\left[|\widehat{p}-p| - \frac{1}{2n} \le z\,\sqrt{\frac{p(1-p)}{n}}\right] = 1-\alpha, \tag{24}$$

čemuž odpovídá interval spolehlivosti s mezemi

$$\begin{aligned} \underline{p} &= \begin{cases} \dfrac{2n\widehat{p} + z^2 - 1 - z\sqrt{z^2 - 2 - \frac{1}{n} + 4\widehat{p}\,(n\,(1-\widehat{p}) + 1)}}{2\,(n+z^2)}, & \widehat{p} > 0, \\ 0, & \widehat{p} = 0, \end{cases} \\ \overline{p} &= \begin{cases} \dfrac{2n\widehat{p} + z^2 + 1 + z\sqrt{z^2 + 2 - \frac{1}{n} + 4\widehat{p}\,(n\,(1-\widehat{p}) - 1)}}{2\,(n+z^2)}, & \widehat{p} < 1, \\ 1, & \widehat{p} = 1. \end{cases} \end{aligned} \tag{25}$$

V článku [23] je uvedeno srovnání jednotlivých metod pro odhad relativních četností, ze kterého vyplývá preference intervalových odhadů (24), (25). Hodnotu M odhadujeme potom stejně jako v kapitole 2.1 pomocí vztahu (3) a interval spolehlivosti podle (6).

Poznamenejme, že v případě, kdy je rozsah výběru srovnatelný s velikostí populace, by šlo dosáhnout zpřesnění tohoto odhadu (podobně jako u klasické normální aproximace) nahrazením výrazu $\frac{p(1-p)}{n}$ v (24) výrazem $\frac{N-n}{N-1}\frac{p(1-p)}{n}$, tj. rozptylem $\widehat{p}$ v hypergeometrickém modelu – viz 2.1.1. Meze intervalu spolehlivosti pak dostaneme nahrazením hodnoty z v (25) hodnotou

$$z^* = z\sqrt{\frac{N-n}{N-1}}. \tag{26}$$

Tuto aproximaci budeme v tomto článku nazývat konečnostním zpřesněním Blythovy–Stillovy aproximace.

Na závěr této kapitoly ukážeme grafickou reprezentaci vybraných metod odhadu. Jako referenční model uvažujeme přesný propočet intervalu spolehlivosti na základě kapitoly 2.2. Ten je na grafech vyznačen vždy plnou čarou. Přerušovanou čarou jsou vyznačeny aproximativní modely. Na grafu 2.4 lze porovnat přesný propočet s normální aproximaci založenou na vzorci (21). Z grafu je vidět rostoucí přesnost aproximace pro rostoucí n. Na druhou stranu, pro malé hodnoty parametru p (resp. hodnoty parametru p blízké 1) může i zde docházet k nepřesnostem, jak ukazuje detail na grafu 2.5.

Lepší vlastnosti aproximace (25) ilustruje graf 2.6 (srovnej s 2.4). I zde však dochází k určitým nepřesnostem pro hodnoty p blízké 0 (resp. 1), které jsou však výrazně menší než v případě klasické normální aproximace, jak ukazuje graf 2.7 (srovnej s 2.5).

Na posledním grafu 2.8 ukážeme porovnání přesného hypergeometrického modelu s modelem binomickým, Blythovou–Stillovou aproximací a jejím konečnostním zpřesněním dle (26).

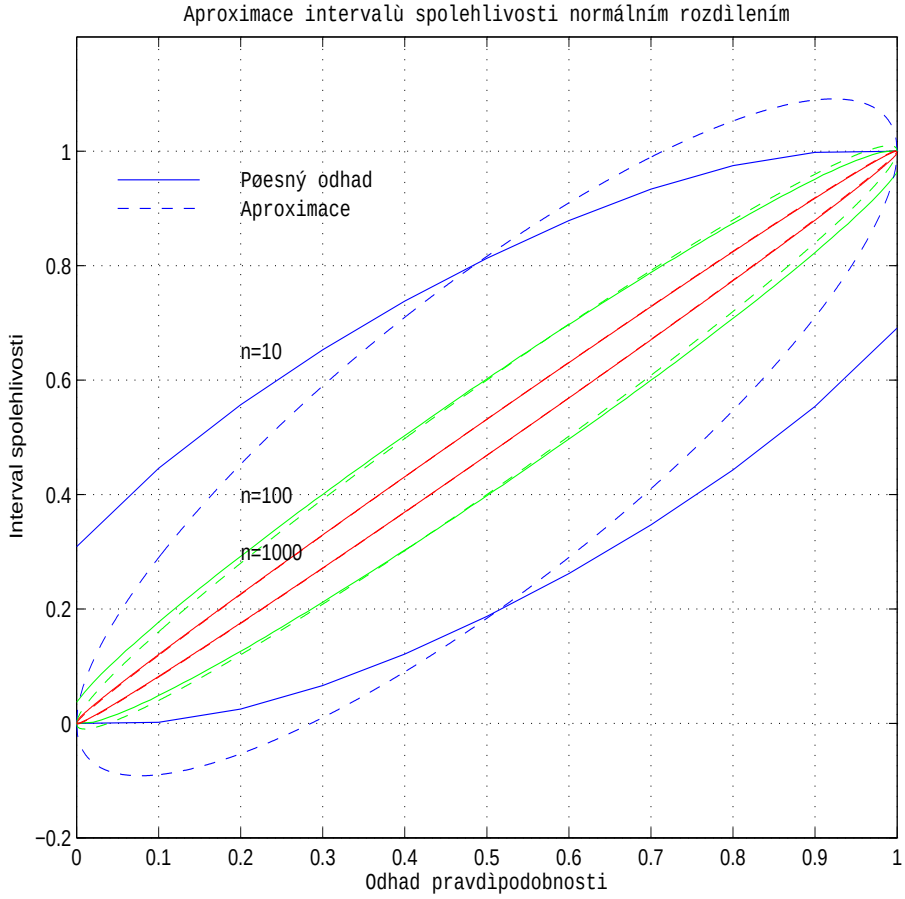
2.5 Pravděpodobnost pokrytí skutečné hodnoty parametru při aproximaci normálním rozdělením

Klasická aproximace normálním rozdělením (21) se doporučuje pro dostatečně velký rozsah výběru (velikost vzorku) n, a není-li parametr p příliš blízko 0 nebo 1. Očekávali bychom, že při rostoucím n se skutečná pravděpodobnost zahrnutí správné hodnoty $100(1-\alpha)\%$ intervalem spolehlivosti pro parametr p bude blížit hodnotě $100(1-\alpha)$ %. O tom, že tato konvergence není rovnoměrná, ale má oscilující charakter, se můžeme přesvědčit na grafu 2.9. Zajímavé je i chování skutečné pravděpodobnosti zahrnutí pro pevné n a pro různé hodnoty p (viz graf 2.10). Podrobnější rozbor lze nalézt v [6].

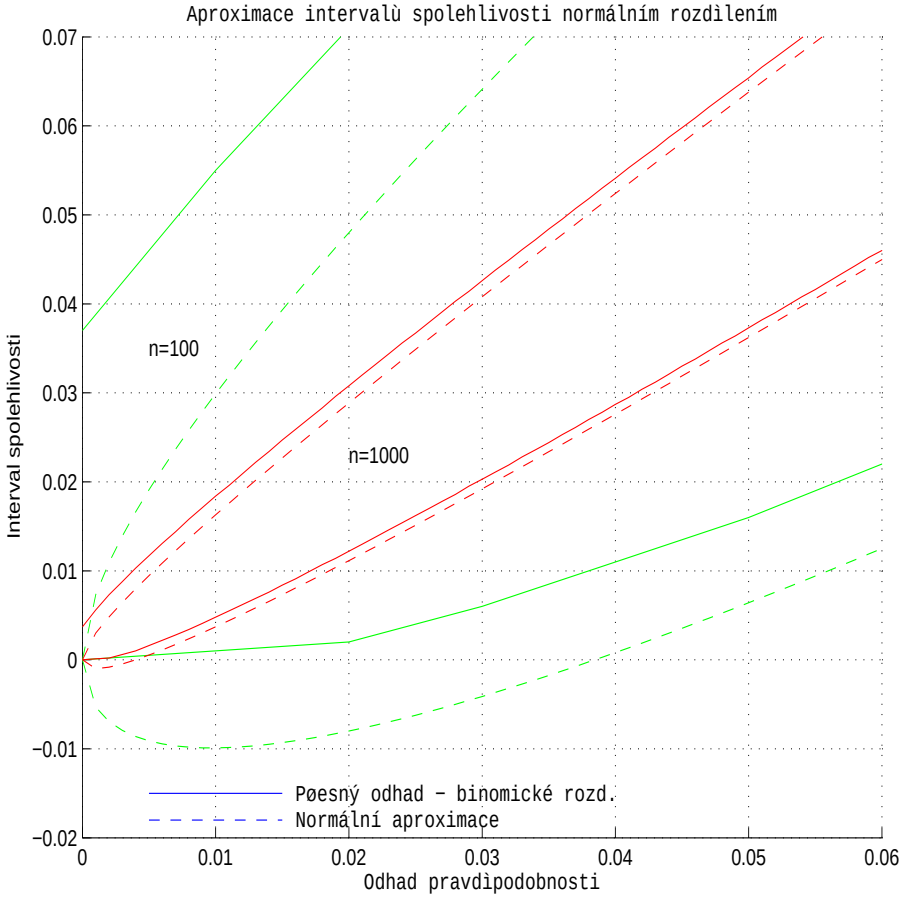
2.6 Odhady četností v rámci cílové skupiny

Cílovou skupinou zde rozumíme v „nestatistické obci“ ustálený pojem pro určitou část (podmnožinu) populace, která je předmětem zájmu. Cílovou skupinou mohou být různé věkové kategorie (například *děti ve věku 4–14 let*), lidé s různým vzděláním (například *vysokoškoláci*) apod. Pro danou cílovou skupinu potom opět chceme odhadnout relativní nebo absolutní četnost nějakého jevu (například relativní počet posluchačů dané rozhlasové stanice, absolutní počet notorických alkoholiků apod.). Podobně jako při odhadech v rámci celé populace budeme odhadovat parametr p, který značí poměr výskytu zkoumaného jevu v rámci cílové skupiny.

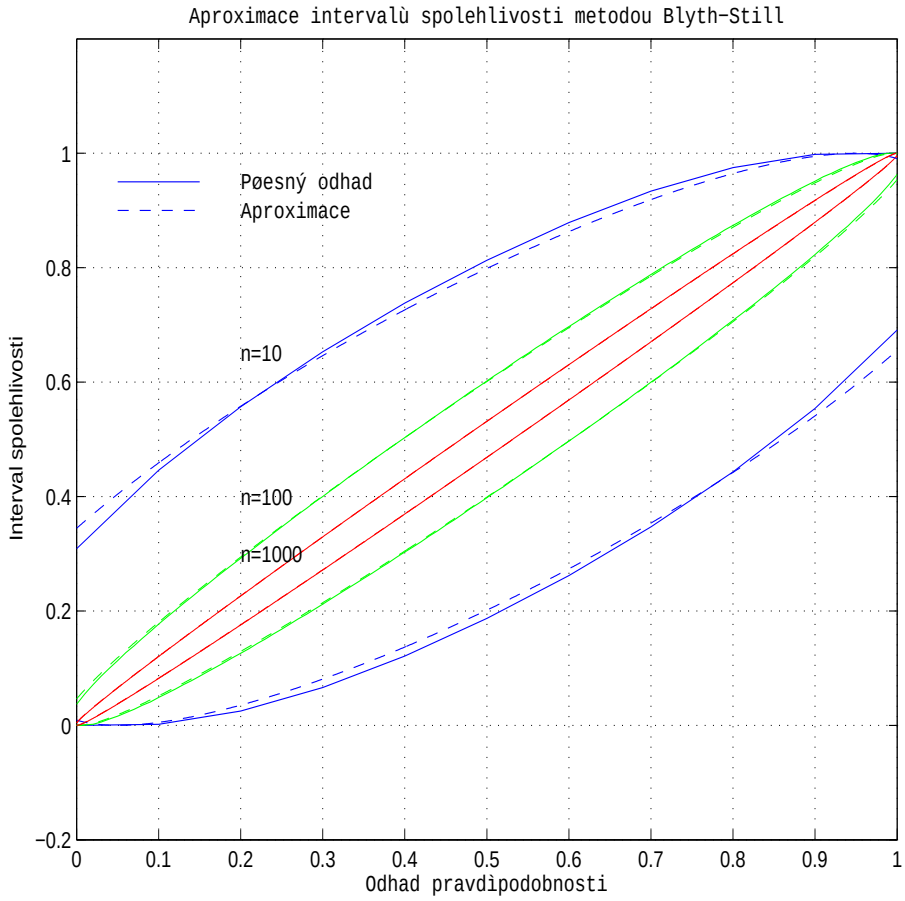
Při detailnějším pohledu zde vzniká i další úloha, jaký je v rámci celé populace výskyt jedinců, kteří odpovídají zkoumanému jevu a současně jsou členy zkoumané cílové skupiny. Tato úloha je však úlohou, kdy zkoumáme výskyt určitého jevu v rámci celé populace (například počet dětí, které jsou posluchači



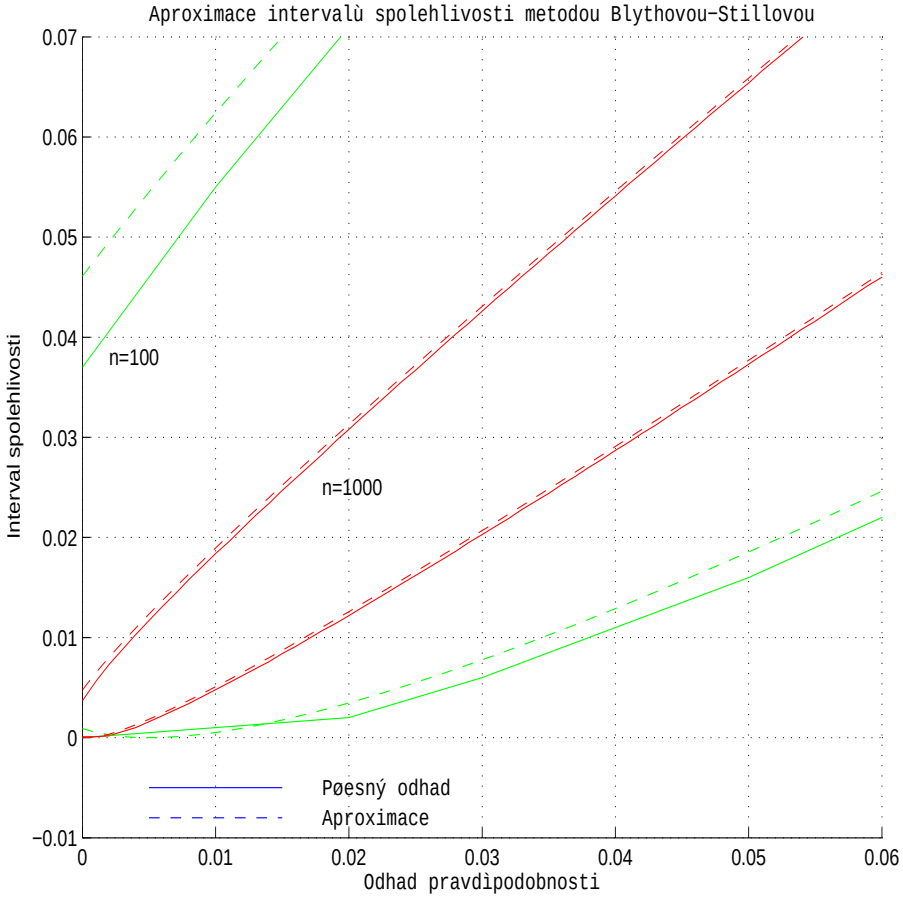
Graf 2.4: Porovnání intervalu spolehlivosti pro normální aproximaci s binomickým modelem



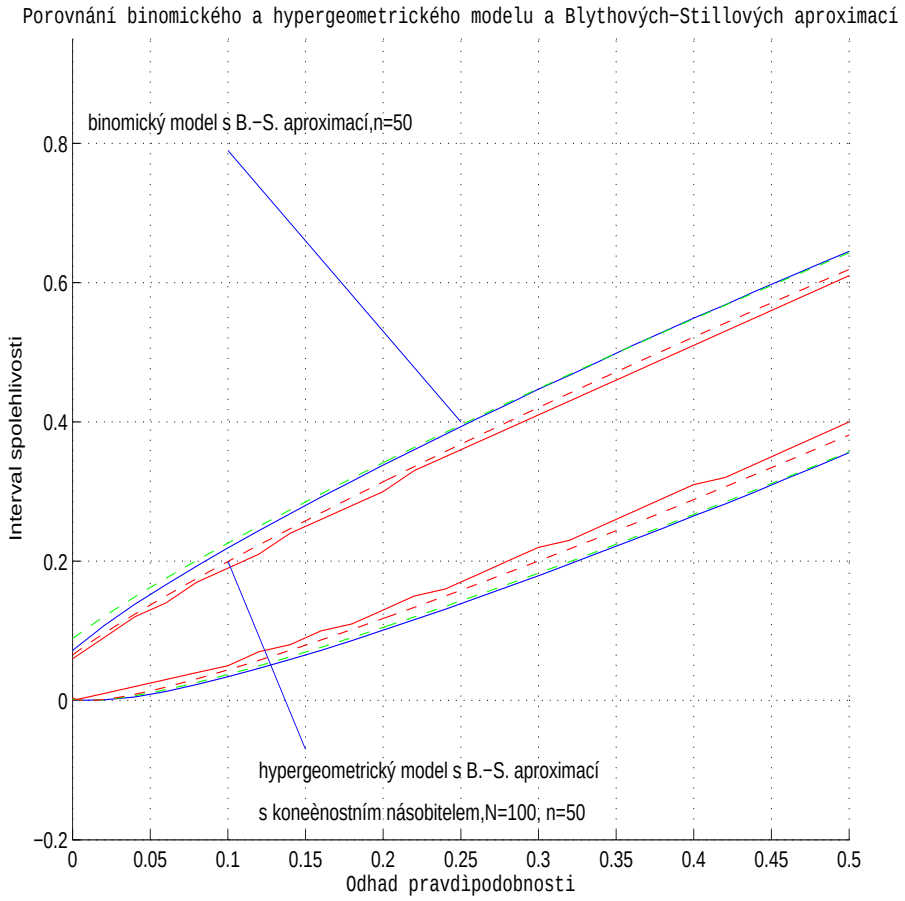
Graf 2.5: Porovnání intervalu spolehlivosti binomického modelu s normální aproximací v detailu



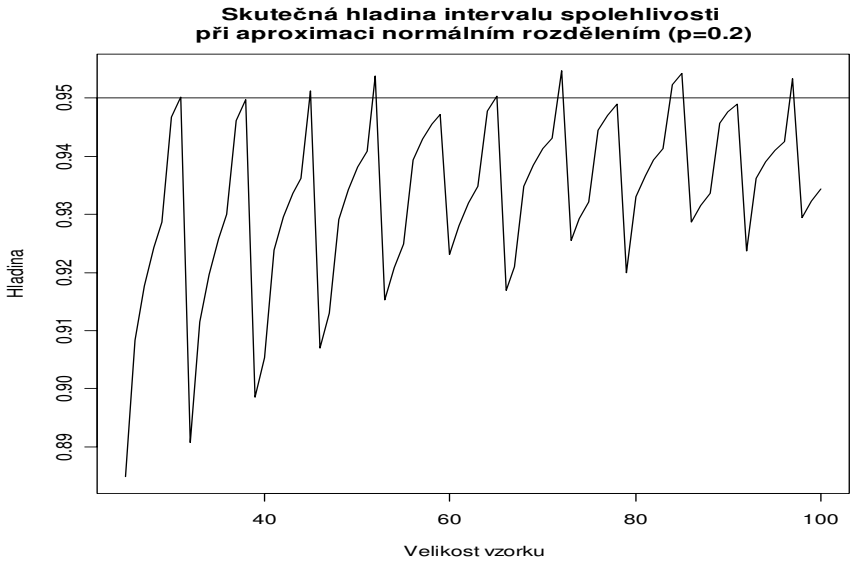
Graf 2.6: Porovnání intervalů spolehlivosti na základě binomického modelu a Blythovy–Stillovy aproximace



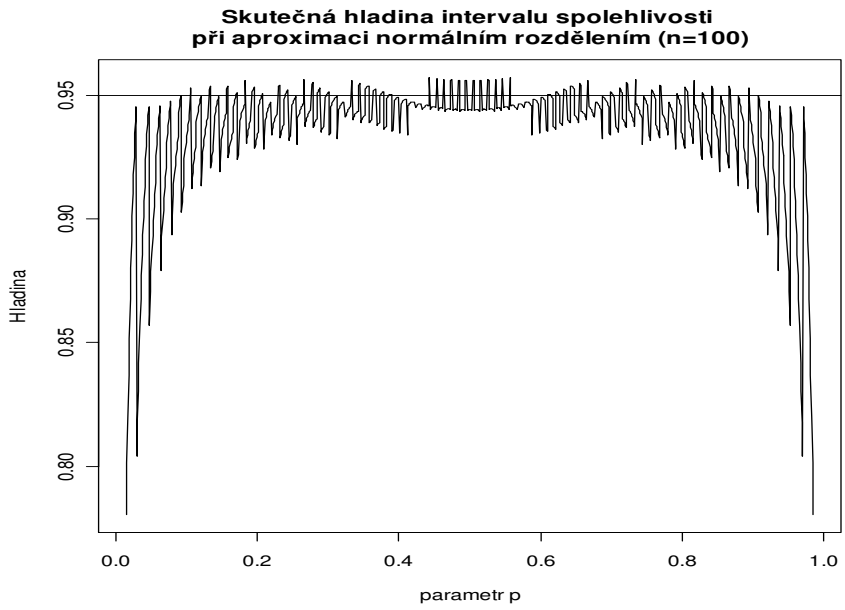
Graf 2.7: Porovnání intervalu spolehlivosti pro Blythovu–Stillovu aproximaci s binomickým modelem v detailu



Graf 2.8: Porovnání intervalu spolehlivosti hypergeometrického a binomického modelu s Blythovou–Stillovou aproximací a jejím konečnostním zpřesněním



Graf 2.9: Pravděpodobnost pokrytí skutečné hodnoty 95% intervalem spolehlivosti při aproximaci normálním rozdělením pro parametr $p = 0{,}2$ při různých hodnotách velikosti vzorku n



Graf 2.10: Pravděpodobnost pokrytí skutečné hodnoty 95% intervalem spolehlivosti při aproximaci normálním rozdělením pro rozsah výběru (velikost vzorku) $n = 100$ při různých hodnotách parametru p

dané rozhlasové stanice, počet notorických alkoholiků, kteří jsou současně vysokoškoláky apod.).

V případě, že odhadujeme četnosti (ať již absolutní nebo relativní) v rámci cílové skupiny, je úloha analogická úloze z předchozích kapitol. První rozdíl spočívá v tom, že musíme zmenšit výchozí hodnotu N, která nyní odpovídá pouze části populace, přesněji velikosti cílové skupiny. Ve většině případů je cílová skupina dostatečně velká (i když třeba její velikost, tj. N, neznáme), takže můžeme použít přímo postupy kapitoly 2.2. I v případě, že tomu tak není, je lepší použít postupy založené přímo na hypergemetrickém rozdělení nebo postupy zmiňované v kapitole 6, resp. 7 (v případě, že neznáme hodnotu N). Druhý rozdíl spočívá v tom, že hodnota rozsahu výběru n (to jest počet vybraných prvků v rámci dané cílové skupiny) je obvykle před vlastním provedením výběru neznámá a tedy náhodná. Uvedené intervaly spolehlivosti jsou proto podmíněny tím, že počet respondentů z dané cílové skupiny je právě n. Tento postup je pro většinu praktických situací naprosto postačující.

Rozbor dalších situací (např. kdybychom chtěli stanovit interval spolehlivosti před vlastním provedením výběru pro stanovení jeho vhodného rozsahu) nebudeme v tomto článku pro jeho komplikovanost rozebírat. Bylo by například možné na základě způsobu výběru určit pravděpodobnosti realizace jednotlivých velikostí cílové skupiny n a provést propočet nepodmíněných rozptylů a odpovídajících intervalů spolehlivosti. Zde by opět mohla najít uplatnění metoda bootstrapu, i když simulace rozsahu výběru může být komplikovaná. Dalším možným postupem by byl přístup na základě poměrových odhadů (viz 4.3).

2.7 Problematika odhadu chyby pro relativní a absolutní četnost ve váženém vzorku

Při téměř každém výstupu z výzkumu veřejného mínění dochází k tzv. převažování výsledků. Jedná se o to, že i sebelépe zorganizovaný náhodný výběr je poškozen odmítnutím jedinců zúčastnit se výzkumu. Pomineme-li úmyslné falšování odpovědí ze strany respondentů, které se objevuje zejména ve výzkumech politických preferencí, v otázkách na velikost osobních nebo rodinných příjmů, případně v otázkách na dosažené vzdělání apod., je problematika návratnosti a odmítnutých rozhovorů jedním z nejvážnějších problémů odhadů správných hodnot. Vážení jako takové může částečně napravit alespoň jednu nesrovnalost, a to případné vychýlení způsobené převahou určité skupiny obyvatel ve výsledném výběrovém vzorku oproti skutečnému poměru v celé populaci, pokud tato vychýlená skupina projevuje v měřené charakteristice odlišnost od zbytku populace. Na druhou stranu vážení nemůže nahradit chybějící a potenciálně odlišnou informaci o respondentech odmítajících odpověď.

Vážení znamená přiřazení vah jednotlivcům ve výsledném vzorku, tj. hodnot w_i, $i = 1, \ldots, n$, kterými jsou výsledné hodnoty odpovědí převažovány. Nechceme zde diskutovat filozofii vážení jako takového, ani postupy pro stanovení vah w_i. Z metod výpočtu vah citujme pouze především postupy založené na práci [11], ve které Deville a Särndal navrhli třídu kalibračních odhadů. Tyto kalibrační odhady jsou asymptoticky ekvivalentí zobecněnému regresnímu od-

hadu (GREG), viz například [28]. Speciálním případem kalibračních odhadů je i iterační metoda nazývaná „raking“, „rim weighting“ nebo též „proportional fitting“, viz [10], její porovnání s metodou založenou na minimalizaci χ^2 lze nalézt například v [21]. Další metody výpočtu vah lze najít například v [15] nebo [25]. Porovnání některých metod odhadu rozptylu zobecněného regresního odhadu je uvedeno například v [29]. Postup pro výpočet vah, které odpovídají tzv. „kosmeticky kalibrovaným odhadům“ lze nalézt v [5]. Dále uvedeme pouze nejčastěji používaný přístup pro odhad chyby vážených odhadů.

Ve všech předchozích postupech pro odhad relativní četnosti, resp. četnosti absolutní byly použity vzorce, ve kterých kromě známých deterministických hodnot byla použita náhodná veličina X, označující počet výskytů sledovaného jevu v n pozorováních – viz zavedení značení za vzorcem (1). Při vážených odhadech a odhadech jejich přesnosti se tato veličina nahradí jejím váženým protějškem. Myšlenka je v zásadě taková, že místo jednoho jedince jsme dostali odpovědi od w_i jedinců. Odhady relativních a absolutních četností proto založíme na stejných vzorcích jako v předchozích kapitolách s tím, že veličinu X nahradíme odhadem

$$X = \frac{n}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i X_i. \tag{27}$$

Protože i odhady velikosti rozptylů a intervalů spolehlivosti uvedené prozatím v tomto článku lze zkonstruovat na základě znalosti veličiny X, je možno je použít s tím, že tuto veličinu nahradíme váženým součtem (27).

Poznamenejme pouze na okraj slabé místo tohoto přístupu. Pokud totiž vážíme, upravujeme výběrový vzorek na výběr dosažený stratifikovaným náhodným výběrem, kde vážení vlastně stanovuje jednotlivé proměnné, vzhledem k nimž má být výběr reprezentativní. Výslednou hodnotu X potom nelze považovat za veličinu s hypergeometrickým, resp. binomickým rozdělením. Správnější by bylo propočítat odhady přes jednotlivá „strata“ a pro každý takovýto odhad vypočítat vlastní velikost chyby. Výslednou chybu pak propočítat jako váženou kombinaci chyb odhadů v jednotlivých „stratech“ výběru. Vzhledem k tomu, že vážení často probíhá přes kategorie, které nejsou navzájem disjunktní, je však takovýto přístup kromě velké komplikovanosti v podstatě nemožný. Rovněž stanovení intervalu spolehlivosti na základě kombinace intervalů spolehlivosti odhadů v jednotlivých „stratech“ je obtížné, ne-li nemožné. Proto se v praktických situacích většinou spokojujeme s postupem uvedeným v předchozím odstavci. Tento přístup je tím korektnější, čím jsou hodnoty w_i blíže k jedné.

3 Srovnávání výběrových četností

V tomto odstavci probereme situaci, kdy měříme výskyt dvou jevů v populaci. Jako zásadní budeme uvažovat dva modely:

- Naše měření odpovídají jiným jedincům populace a porovnáváme mezi sebou odhady výskytu jevů na základě těchto dvou nezávislých výběrů.

- Naše měření probíhají na stejných jedincích a měříme změnu výskytu jevu při změněných podmínkách.

3.1 Porovnání dvou nezávislých výběrových četností

V tomto odstavci budeme předpokládat, že zkoumáme výskyt jednoho nebo dvou znaků na základě měření, která jsou všechna vzájemně nezávislá. Počet měření výskytu prvního znaku označíme n_1 a příslušná měření jako X_i, $i = 1, \ldots, n_1$. Podobně počet měření druhého znaku (resp. stejného znaku v populaci, ale na jiných jednotkách) označíme jako n_2 a příslušná měření jako Y_i, $i = 1, \ldots, n_2$. Dále označíme jako p_1, resp. p_2 teoretickou pravděpodobnost výskytu prvního znaku, resp. druhého znaku. Budeme zkoumat intervalové odhady parametru $\theta = p_1 - p_2$. Označíme dále $\psi = \frac{p_1+p_2}{2}$. Vzhledem k tomu, že velikost populace, na které jsou prováděna měření, je obvykle výrazně větší než velikost výběrových vzorků, budeme v tomto odstavci předpokládat, že X_i, $i = 1, \ldots, n_1$ a Y_i, $i = 1, \ldots, n_2$ jsou nezávislé náhodné veličiny.

Při konstrukci intervalových odhadů parametru θ lze postupovat analogicky jako v případě jednoho výběru. Vycházíme zde zejména z přehledného článku [22]. Jako nejjednodušší se jeví využít asymptotické normality maximálně věrohodného odhadu $\widehat{\theta} = \frac{\sum_{i=1}^{n_1} X_i}{n_1} - \frac{\sum_{i=1}^{n_2} Y_i}{n_2}$. Rozptyl tohoto odhadu, který je roven $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$, se odhadne nahrazením teoretických pravděpodobností p_1, p_2 jejich výběrovými ekvivalenty $\widehat{p_1} = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$, $\widehat{p_2} = \frac{\sum_{i=1}^{n_2} Y_i}{n_2}$. Intervalový odhad potom dostaneme jako

$$\widehat{\theta} \pm z \sqrt{\frac{\widehat{p_1}\,(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}\,(1-\widehat{p_2})}{n_2}}, \tag{28}$$

kde z je příslušný kvantil normálního rozdělení.

Podobně jako při odhadu v jednorozměrném případě se používá princip spojitostní korekce (viz např. [12]), který dává intervalový odhad ve tvaru

$$\widehat{\theta} \pm \left(z \sqrt{\frac{\widehat{p_1}\,(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}\,(1-\widehat{p_2})}{n_2}} + \frac{\frac{1}{n_1} + \frac{1}{n_2}}{2} \right). \tag{29}$$

Variant pro výpočet chyby při porovnání dvou výběrových četností je celá řada. Velmi pěkný přehledný článek porovnávající jednotlivé přístupy lze najít v [22]. Vyjměme z něj například tam citovanou (pod číslem 10) i doporučovanou metodu. Interval spolehlivosti pro rozdíl θ se udává ve tvaru $[\widehat{\theta} - \delta, \widehat{\theta} + \epsilon]$, kde

$$\delta = z\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}} \tag{30}$$

$$\epsilon = z\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}}. \tag{31}$$

l_1, u_1 jsou kořeny rovnice (vzhledem k π)

$$|\widehat{p_1} - \pi| = z\sqrt{\frac{\pi(1-\pi)}{n_1}}$$

a l_2, u_2 jsou kořeny rovnice (opět vzhledem k π)

$$|\widehat{p_2} - \pi| = z\sqrt{\frac{\pi(1-\pi)}{n_2}}.$$

Tyto kořeny se dají vyjádřit ve tvaru, který je analogický výrazu (23).

3.2 Porovnání dvou nebo více relativních četností na stejném výběrovém vzorku

V tomto případě rozšíříme naše značení následujícím způsobem. Budeme předpokládat, že na jednom jedinci máme k dispozici k měření. Veličina X_{i1} bude určovat výskyt znaku při prvním zkoumání na jedinci i. Veličina X_{i2} potom výskyt znaku při druhém zkoumání, teoreticky může jít i o zkoumání výskytu jiného znaku. Takto budeme postupovat dále, až konečně označíme X_{ik} veličinu, která nabude hodnoty 1 při výskytu znaku na i-tém jedinci při k-tém zkoumání (měření). Jako příklad zde může sloužit například kontinuální výzkum výskytu nějakého jevu v panelu jednotlivců v určitých časových obdobích.

Označme počet výskytů jevu při j-tém zkoumání $T_j = \sum_{i=1}^{n} X_{ij}$. Vhodnou statistikou pro ověření hypotézy, že mezi teoretickými relativními četnostmi neexistuje významný rozdíl, je

$$\sum_{j=1}^{k}(T_j - \overline{T})^2,$$

kde $\overline{T} = \frac{\sum_{j=1}^{k} T_j}{k}$. V článku [8] je uvedeno asymptotické rozdělení pro tuto statistiku ve tvaru

$$\frac{k(k-1)\sum\limits_{j=1}^{k}(T_j - \overline{T})^2}{k\sum\limits_{i} u_i - \sum\limits_{i} u_i^2} \tag{32}$$

jako χ^2 s $k-1$ stupni volnosti, přičemž u_i značí počet výskytů jevu u i-tého jedince, tedy

$$u_i = \sum_{j=1}^{k} X_{ij},\ \forall i.$$

V případě $k = 2$ přechází uvedený test do známého McNemarova testu (viz např. [3, 19]).

4 Odhady dalších charakteristik populace a jejich chyba

4.1 Terminologie a značení při výběrech z konečných populací

V literatuře zabývající se výběry z konečných populací se používá výrazně odlišné značení než v teorii pravděpodobnosti. Velká písmena jsou zde vyhrazena pro populační, tj. pevné (nenáhodné) charakteristiky, a naopak malými písmeny se značí charakteristiky výběrové, které jsou ovšem náhodnými veličinami. (Při označování hodnot zkoumaných znaků na jednotkách populace značení kolísá – srov. např. [9] a [14].) V následujících dvou kapitolách proto budeme v souladu s literaturou používat odlišné značení oproti předchozímu textu.

Není-li řečeno jinak, rozumí se prostým náhodným výběrem výběr bez vracení se stejnou pravděpodobností zahrnutí pro každou jednotku. Výběr s vracením se v praxi patrně příliš nepoužívá, je však výhodný z teoretického hlediska – většina odvození je v tomto modelu mnohem jednodušší, neboť jednotlivé pokusy jsou na rozdíl od výběru bez vracení nezávislé. V obvyklých případech, kdy je rozsah výběru mnohem menší než velikost populace, je navíc zpřesnění odhadů odvozených za předpokladu výběru bez vracení prakticky zanedbatelné.

Předpokládejme, že sledujeme dva znaky, které u jednotek v populaci S nabývají hodnot $x_1, \ldots, x_N$, resp. $y_1, \ldots, y_N$. Označme populační (základní) úhrny

$$X = \sum_{i=1}^{N} x_i, \qquad Y = \sum_{i=1}^{N} y_i$$

a jejich výběrové protějšky

$$x = \sum_{i \in s} x_i, \qquad y = \sum_{i \in s} y_i,$$

kde s je prostý náhodný výběr o rozsahu n z populace S. Analogicky označíme populační průměry

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i, \qquad \overline{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

a výběrové průměry

$$\overline{x} = \frac{1}{n} \sum_{i \in s} x_i, \qquad \overline{y} = \frac{1}{n} \sum_{i \in s} y_i.$$

Dále zavedeme populační, resp. výběrový rozptyl

$$\sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{Y})^2, \quad s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \overline{y})^2.$$

Místo výrazu σ_y^2 se někdy používá též

$$\sigma_y^{*2} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{Y})^2 = \frac{N-1}{N}\sigma_y^2.$$

(Vzorce pro σ_x^2, s_x^2 apod. zde již neuvádíme, neboť jsou zcela analogické.)

4.2 Odhad průměru a úhrnu

Při prostém náhodném výběru platí mj. následující vztahy (viz např. [14]):

$$\mathrm{E}\,\overline{y} = \overline{Y}, \tag{33}$$

$$\mathrm{var}\,\overline{y} = \left(1 - \frac{n}{N}\right)\frac{1}{n}\sigma_y^2, \tag{34}$$

$$\mathrm{E}\,s_y^2 = \sigma_y^2. \tag{35}$$

Při výběru s vracením platí stále (33), ale (34)–(35) se změní (viz např. [9]):

$$\mathrm{var}\,\overline{y} = \frac{1}{n}\sigma_y^{*2}, \tag{36}$$

$$\mathrm{E}\,s_y^2 = \sigma_y^{*2}. \tag{37}$$

Jak vyplývá z rovnice (33), $\overline{y}$ je nestranným odhadem průměru $\overline{Y}$ a

$$\widehat{Y} = N\overline{y} = \frac{N}{n}y$$

je nestranným odhadem úhrnu Y. Rozptyl odhadu $\widehat{Y}$ je N^2násobkem rozptylu $\overline{y}$, tj. při výběru bez vracení

$$\mathrm{var}\,\widehat{Y} = N^2\left(1 - \frac{n}{N}\right)\frac{1}{n}\sigma_y^2. \tag{38}$$

Intervalové odhady průměru, resp. úhrnu jsou pak tvaru

$$\overline{y} \pm z_{1-\frac{\alpha}{2}}\sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n}s_y^2}, \tag{39}$$

$$\widehat{Y} \pm N z_{1-\frac{\alpha}{2}}\sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n}s_y^2}, \tag{40}$$

kde $z_{1-\frac{\alpha}{2}}$ je kvantil normálního rozdělení, je-li $n \geq 30$. V opačném případě se za $z_{1-\frac{\alpha}{2}}$ dosadí příslušný kvantil Studentova t rozdělení s $n-1$ stupni volnosti. Vynecháním konečnostního násobitele $(1-\frac{n}{N})$ dostaneme odhady, které bychom odvodili z výběru s vracením.

Uvedený postup je oprávněný pouze v případě, že rozdělení hodnot y_i v populaci je alespoň přibližně normální. Dle Ing. Machka přitom nejvíce vadí zejména výrazná asymetrie (šikmost) tohoto rozdělení. V takovém případě rozdělení $\overline{y}$

konverguje (při rostoucím n) k normálnímu pomaleji, a je proto potřeba vyšší rozsah výběru na to, aby intervaly spolehlivosti (39) a (40) opravdu pokrývaly příslušné populační charakteristiky s požadovanou pravděpodobností. Problematické je dle [9] rovněž použití těchto intervalových odhadů pro vysoké koeficienty spolehlivosti (99 % a více, tj. $\alpha \leq 0{,}01$), neboť rozdělení výběrových průměrů je na obou koncích velmi citlivé na stupeň odchýlení rozdělení y_i od normálního a také na přítomnost extrémních hodnot.

Určitou možnost kvantifikace, nakolik je rozdělení hodnot y_i blízké normálnímu, lze nalézt v [14], včetně v praxi bohužel nepříliš použitelného postupu pro případy, kdy toto rozdělení není blízké normálnímu. Lepší řešení, které by patrně spočívalo v konstrukci asymetrického intervalu spolehlivosti (podobně jako v předchozích kapitolách popsané intervalové odhady relativní četnosti), nám ale zatím není známo.

4.2.1 Odhad četnosti a relativní četnosti jako speciální případ úhrnu a průměru

Speciálním případem je 0-1 znak, kdy úhrn nazýváme četností a průměr relativní (nebo poměrnou) četností. Označme populační, resp. výběrovou četnost symboly N_y, resp. n_y a populační, resp. výběrovou relativní četnost symboly P_y, resp. p_y. Výběrová četnost n_y je při prostém náhodném výběru bez vracení náhodnou veličinou s hypergeometrickým rozdělením – viz 2.1.

Dle [14] lze v tomto případě použít odhady (39) a (40), je-li $n_y \geq 50$, přičemž

$$s_y^2 = \frac{np_y(1-p_y)}{n-1}. \tag{41}$$

Jak vyplývá z (35), s_y^2 je nestranným odhadem rozptylu σ_y^2, pro který v případě 0-1 znaku platí

$$\sigma_y^2 = \frac{N}{N-1}P_y(1-P_y) \doteq P_y(1-P_y) = \sigma_y^{*2}.$$

Pro $10 \leq n_y < 50$ doporučuje Hájek (viz[14]) korigovat nespojitost rozdělení výběrové četnosti n_y přidáním členu $\frac{1}{2n}$, resp. $\frac{N}{2n}$, tj.

$$p_y \pm \frac{1}{2n} \pm z_{1-\frac{\alpha}{2}}\sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n}s_y^2}, \tag{42}$$

$$Np_y \pm \frac{N}{2n} \pm Nz_{1-\frac{\alpha}{2}}\sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n}s_y^2}. \tag{43}$$

Pro $n_y < 10$ je dle Hájka (viz[14]) lepší použít odhady založené na aproximaci Poissonovým rozdělením se střední hodnotou nP_y – viz 2.3. Tyto odhady už ale nebudou symetrické kolem bodového odhadu (podobně jako nejsou symetrické ani přesné intervaly spolehlivosti odvozené na základě hypergeometrického rozdělení).

V případě prostého náhodného výběru s vracením je výběrová četnost n_y náhodnou veličinou s binomickým rozdělením – viz 2.2. Při použití normální

aproximace dostaneme analogické vzorce pro intervalové odhady, které se od (39) a (40) budou lišit pouze absencí konečnostního násobitele $(1 - \frac{n}{N})$ a budou (až na odlišné značení) stejné jako ty ve vztahu (21) po nahrazení jmenovatele n výrazem $n - 1$. Znovu poznamenejme, že jak vyplývá z (37), nestranným odhadem rozptylu $\sigma_y^{*2} = P_y(1 - P_y)$ není výraz $p_y(1 - p_y)$, ale s_y^2.

4.3 Odhad poměru dvou úhrnů

Poměr úhrnů Y/X (který je samozřejmě roven i poměru průměrů $\overline{Y}/\overline{X}$) odhadujeme výběrovým poměrem y/x. Intervalový odhad je odvozen v [14] zavedením pomocných hodnot

$$z_i = y_i - \frac{Y}{X}x_i,$$

pro jejichž úhrn platí, že

$$Z = \sum_{i=1}^{N} z_i \equiv 0.$$

Intervalový odhad poměru dvou úhrnů Y/X je tvaru

$$\frac{y}{x} \pm \frac{z_{1-\frac{\alpha}{2}}}{x}\sqrt{\left(1 - \frac{n}{N}\right)\frac{n}{n-1}\sum_{i\in s}\left(y_i - \frac{y}{x}x_i\right)^2}, \tag{44}$$

resp. při zanedbání konečnostního násobitele (je-li N mnohem větší než n)

$$\frac{y}{x} \pm \frac{z_{1-\frac{\alpha}{2}}}{x}\sqrt{\frac{n}{n-1}\sum_{i\in s}\left(y_i - \frac{y}{x}x_i\right)^2}. \tag{45}$$

Odhad (44) je tím uspokojivější, čím menší je poměrná výběrová chyba výběrového úhrnu x

$$\gamma_x = V_x\sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n}},$$

kde

$$V_x = \frac{\sigma_x}{\overline{X}}$$

je variační koeficient. Hájek v [14] uvádí, že odhad (44) lze pokládat za uspokojivý, pokud je výběrový odhad poměrné výběrové chyby $g_x < 0{,}2$, kde

$$g_x = v_x\sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n}},$$
$$v_x = \frac{s_x}{\overline{x}}.$$

Nabývají-li x_i hodnot pouze 0 nebo 1, lze variační koeficient V_x vyjádřit jako

$$V_x = \frac{1}{P_x}\sqrt{\frac{N}{N-1}P_x(1 - P_x)} = \sqrt{\frac{N}{N-1}\left(\frac{1}{P_x} - 1\right)} \doteq \sqrt{\frac{1}{P_x} - 1}. \tag{46}$$

V tomto případě je tedy (při zanedbání konečnostního násobitele)

$$g_x \doteq \sqrt{\frac{1}{n}\left(\frac{1}{p_x}-1\right)} = \sqrt{\frac{1}{n_x}-\frac{1}{n}}.$$

Nerovnost $g_x < 0{,}2$ je pak splněna pro

$$n_x > \frac{25n}{n+25},$$

tedy např. pro $n_x > 25$.

Speciálním využitím odhadu poměrů dvou úhrnů je odhad průměru (či relativní četnosti) nějakého jevu na cílové skupině, jejíž velikost v populaci neznáme. V tomto případě udávají hodnoty x_i příslušnost k dané cílové skupině ($x = n_x$ je její velikost ve výběru a $X = N_x$ neznámá velikost v celé populaci) a y_i hodnoty zkoumaného jevu pro jedince z dané cílové skupiny, přičemž $y_i = 0$, pokud $x_i = 0$.

4.4 Odhady ve váženém vzorku

V případě váženého vzorku se všechny odhady odvozené pro nevážený vzorek upraví dále popsaným způsobem.

Výběrový úhrn y, průměr $\overline{y}$ a rozptyl s_y^2 se nahradí váženými protějšky:

$$\begin{aligned} y_w &= \frac{n}{\sum\limits_{i\in s} w_i} \sum_{i\in s} w_i y_i, \\ \overline{y}_w &= \frac{\sum\limits_{i\in s} w_i y_i}{\sum\limits_{i\in s} w_i}, \\ s_{wy}^2 &= \frac{1}{n-1}\frac{n}{\sum\limits_{i\in s} w_i} \sum_{i\in s} w_i (y_i - \overline{y}_w)^2. \end{aligned}$$

Poznamenejme, že pro 0-1 znak odpovídá s_{wy}^2 výrazu $\frac{n}{n-1} p_{wy}(1-p_{wy})$, kde p_{wy} je vážená relativní četnost znaku y.

Výraz $\sum_{i\in s}\left(y_i - \frac{y}{x}x_i\right)^2$ z odhadů (44) a (45) se nahradí výrazem

$$\frac{n}{\sum\limits_{i\in s} w_i} \sum_{i\in s} w_i \left(y_i - \frac{y_w}{x_w}x_i\right)^2,$$

kde x_w je vážený úhrn znaku x definovaný analogicky jako u znaku y.

5 Skupinkové výběry

V některých případech je ekonomičtější a mnohdy i logičtější nevybírat přímo zkoumané elementy, ale jejich určitá uskupení, a šetření provádět na všech jednotkách vybrané skupinky. Příkladem jsou třeba výběry domácností namísto

několika jednotlivců v nich žijících či celých školních tříd namísto jednotlivých žáků. V takových případech však nemůžeme počítat s nezávislostí vybraných jednotek z téže skupinky, a dosud odvozené vztahy proto mohou být mírně zkreslující.

Pro definici modelu popisujícího vlastnosti odhadů této modifikace náhodného výběru se opřeme o moderní teorii pravděpodobnostního výběru, jak je popsána například v [13] či [26].

5.1 Náhodný výběr z konečné populace

Než přistoupíme k teorii skupinkových výběrů, popišme standardní teorii náhodných pravděpodobnostních výběrů a tu pak zobecněme na studovaný problém.

Jako výše předpokládáme populaci N prvků

$$S = \{1, \ldots, N\},$$

na které provádíme výběr $s \subseteq S$, o n prvcích. Velikost výběru n bývá většinou předem stanovena, existují však i postupy, při kterých počet vybraných jednotek je náhodnou veličinou.

Náhodným výběrem budeme v této kapitole rozumět pravděpodobnostní rozdělení $\{\mathrm{P}(s)\}_{s\subseteq S}$ na množině všech možných podmnožin S. Budeme předpokládat, že každá jednotka populace může být vybrána s kladnou pravděpodobností. Matematicky lze tedy náhodný výběr popsat pomocí systému pravděpodobností

$$\begin{gathered} 0 \le \mathrm{P}(s) \le 1 \quad \forall s \subseteq S, \\ \sum_{s\subseteq S} \mathrm{P}(s) = 1, \end{gathered}$$

přičemž

$$\forall i \in S \ \exists s \subseteq S : \ i \in s, \ P(s) > 0.$$

Přiřazením nulových pravděpodobností celým třídám podmnožin S můžeme docílit požadovaných vlastností výběru. Pokud například požadujeme pevnou velikost výběru n, kladnou pravděpodobnost přiřadíme pouze takovým podmnožinám $s \subseteq S$, které mají právě n prvků.

Pro studium vlastností odhadů jsou velmi důležité dva ukazatele: pravděpodobnost, že do výběru bude zahrnuta jednotka i, a pravděpodobnost, že do jednoho výběru budou zahrnuty jednotky i a j zároveň. Označíme tyto pravděpodobnosti Π_i a $\Pi_{i,j}$, tj.

$$\begin{aligned} \Pi_i &= \mathrm{P}(i \in s), \\ \Pi_{i,j} &= \mathrm{P}(i, j \in s). \end{aligned}$$

Pomocí takto zavedeného aparátu samozřejmě můžeme popsat všechny známé druhy pravděpodobnostních výběrů.

Příklad 5.1 *Je–li např.*

$$\mathrm{P}(s) = \frac{1}{\binom{N}{n}}$$

pro všechny podmnožiny S, které mají právě n prvků, jde o prostý náhodný výběr (PNV). Pravděpodobnosti zahrnutí jsou tedy rovny

$$\begin{aligned}\Pi_i &= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},\\ \Pi_{i,j} &= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}.\end{aligned}$$

□

Jako v předchozích kapitolách předpokládejme, že pro každý prvek $i \in S$ lze jednoznačně určit hodnotu y_i nějaké veličiny $\mathbb{Y}$, přičemž připouštíme i diskrétní a spojité rozdělení $\mathbb{Y}$ (nikoliv jen hodnoty 0 a 1).

Budeme se zde zabývat pouze odhadem $\widehat{Y}$ populačního úhrnu

$$Y = \sum_{i \in S} y_i,$$

přičemž v případě průměru (u alternativní veličiny odpovídá relativní četnosti – viz kapitolu 2) lze odhad jednoduše zkonstruovat jako

$$\widehat{\overline{Y}} = \frac{\widehat{Y}}{N}.$$

Rozptyl tohoto odhadu pak dostaneme podělením druhou mocninou N, tj.

$$\operatorname{var} \widehat{\overline{Y}} = \frac{\operatorname{var} \widehat{Y}}{N^2}.$$

Nadále budeme předpokládat, že odhad $\widehat{Y}$ úhrnu Y bude konstruován jako tzv. Horvitzův–Thompsonův, tj. má tvar

$$\widehat{Y} = \sum_{i \in s} \frac{y_i}{\Pi_i}. \tag{47}$$

Jak je známo (viz např. [26]), odhad (47) je nestranný a jeho rozptyl je

$$\operatorname{var} \widehat{Y} = \sum_{i=1}^{N} \frac{y_i^2}{\Pi_i^2} \Pi_i (1 - \Pi_i) + \sum_{i \neq j} \frac{y_i}{\Pi_i} \frac{y_j}{\Pi_j} (\Pi_{i,j} - \Pi_i \Pi_j).$$

Odhad tohoto rozptylu je roven

$$\widehat{\operatorname{var} \widehat{Y}} = \sum_{i \in s} \frac{y_i^2}{\Pi_i} \left(\frac{1}{\Pi_i} - 1 \right) + \sum_{i \neq j \in s} \frac{y_i y_j}{\Pi_{i,j}} \left(\frac{\Pi_{i,j}}{\Pi_i \Pi_j} - 1 \right), \tag{48}$$

což lze dále upravit pro případ výběru o pevné velikosti na

$$\widehat{\operatorname{var} \widehat{Y}} = \frac{1}{2} \sum_{i \neq j \in s} \left(\frac{y_i}{\Pi_i} - \frac{y_j}{\Pi_j} \right)^2 \frac{\Pi_i \Pi_j - \Pi_{i,j}}{\Pi_{i,j}}. \tag{49}$$

Příklad 5.2 *Vraťme se opět k PNV.*

$$\widehat{Y} = \sum_{i \in s} \frac{N y_i}{n} = \frac{N}{n} \sum_{i \in s} y_i, \tag{50}$$

tedy H–T odhad odpovídá běžně používanému odhadu průměrem vybraných jednotek vynásobeným velikostí populace, srov. s 4.2. Dosazením do (49) a úpravami pak dostaneme stejný odhad rozptylu, jako je použit v (40). □

5.2 Odhad úhrnu a jeho rozptyl při skupinkových výběrech

Předpokládejme, že máme populaci $S = \{1, \ldots, N\}$, jejíž elementy jsou seskupeny do skupinek $S_1, \ldots, S_M$. Označme

$$S_G = \{S_1, \ldots, S_M\}$$

množinu všech skupinek populace S. Podobně jako S můžeme S_G jednoznačně ztotožnit s množinou $\{1, \ldots, M\}$. Nechť N_i značí velikost skupinky S_i, $i = 1, \ldots, M$.

Skupinkovým výběrem nazveme každý náhodný výběr $s \subseteq S$, pro který platí

$$i \in s,\ i \in S_k \Rightarrow S_k \subseteq s. \tag{51}$$

Je zřejmé, že každý skupinkový výběr s z S lze jednoznačně ztotožnit s nějakým náhodným výběrem s_G z S_G.

Nechť $\Pi'_k = \mathrm{P}(S_k \in s_G),\ \Pi'_{k,l} = \mathrm{P}(S_k, S_l \in s_G)$. Pak platí

$$\begin{array}{rcccl} \Pi_i & = & \mathrm{P}(i \in s) & = & \Pi'_k \qquad \text{pro } i \in S_k, \\ \Pi_{i,j} & = & \mathrm{P}(i, j \in s) & = & \Pi'_{k,l} \qquad \text{pro } i \in S_k,\ j \in S_l, \end{array}$$

přičemž k, l z poslední rovnosti nemusejí být nutně různá.

Situace se velmi zjednoduší, pokud budeme na skupinkové výběry nahlížet jako na výběry skupinek a zjišťované hodnoty budou tvořit celkové součty za jednotlivé skupinky. Takto motivovaný odhad a odhad (47) jsou pak totožné

$$\widehat{Y} = \sum_{i \in s} \frac{y_i}{\Pi_i} = \sum_{S_k \subseteq s} \sum_{i \in S_k} \frac{y_i}{\Pi'_k} = \sum_{S_k \subseteq s} \frac{\sum\limits_{i \in S_k} y_i}{\Pi'_k}.$$

Z výše uvedeného důvodu pak můžeme na tento odhad jednoduše aplikovat vzorce (48) a (49). Obdržíme tak odhady

$$\widehat{\mathrm{var}\, \widehat{Y}} = \sum_{S_k \subseteq s} \frac{\left(\sum\limits_{i \in S_k} y_i \right)^2}{\Pi'_k} \left(\frac{1}{\Pi'_k} - 1 \right) + \sum_{\substack{S_k, S_l \subseteq s, \\ k \neq l}} \frac{\sum\limits_{i \in S_k} y_i \sum\limits_{j \in S_l} y_j}{\Pi'_{k,l}} \left(\frac{\Pi'_{k,l}}{\Pi'_k \Pi'_l} - 1 \right) \tag{52}$$

a pro patrně používanější výběr o pevném počtu skupinek

$$\widehat{\operatorname{var}\,\widehat{Y}} = \frac{1}{2} \sum_{\substack{S_k, S_l \subseteq s, \\ k \neq l}} \left(\frac{\sum\limits_{i \in S_k} y_i}{\Pi'_k} - \frac{\sum\limits_{j \in S_l} y_j}{\Pi'_l} \right)^2 \frac{\Pi'_k \Pi'_l - \Pi'_{k,l}}{\Pi'_{k,l}}. \tag{53}$$

Odhadem rozptylu samozřejmě ještě nedostáváme intervalový odhad pro Y. Ten lze ovšem jednoduše získat například pomocí aproximace normálním rozdělením popsané v odstavci 2.4.

Příklad 5.3 *Podívejme se opět na případ prostého náhodného výběru, tentokrát ve smyslu výběru skupinek.*

Nechť $s_G \subset S_G$ je prostý náhodný výběr m skupinek. Platí

$$\begin{aligned} \Pi'_k &= \frac{m}{M}, \\ \Pi'_{k,l} &= \frac{m(m-1)}{M(M-1)}. \end{aligned}$$

Dosazením do (53) dostáváme

$$\widehat{\operatorname{var}\,\widehat{Y}}_{PNV} = \sum_{S_k \neq S_l \in s_G} \frac{M(M-m)}{m^2(m-1)} \left(\sum_{i \in S_k} y_i^2 + \sum_{i \neq j \in S_k} y_i y_j - \sum_{i \in S_k} \sum_{j \in S_l} y_i y_j \right), \tag{54}$$

což je ekvivalentní vzorci uvedenému např. v [14]. □

V knize [26] se autoři věnují kvalitativnímu porovnání skupinkových výběrů s obyčejnými. Ze vzorce (53) je patrné, že výběry skupinek budou velice efektivní v případě, že pravděpodobnosti zahrnutí do výběru budou úměrné celkovému úhrnu skupinky. To lze alespoň částečně zařídit úměrností vůči velikosti skupinky (pokud předpokládáme, že průměrné hodnoty skupinek se příliš neliší).

U prostého výběru skupinek se ukazuje, že výběr je neefektivní (má větší rozptyl odhadu než běžný výběr) v případě, že v jistém smyslu průměrná variace uvnitř skupinek je malá v porovnání s celkovou variací zkoumané veličiny. Toto je ovšem v praxi velmi obvyklé (např. členové stejné domácnosti se chovají podobně). Z toho plyne, že použitím běžných odhadů pro skupinkový výběr můžeme značně podhodnocovat rozptyl, a tudíž vyvozovat zkreslené závěry o chybě, které jsme se dopustili odhadem z výběru.

5.3 Aplikace odvozených vztahů v praxi

V předchozí kapitole jsme se v příkladech zaměřili především na prostý náhodný výběr. Je důležité připomenout, že v praxi se výběr jen zřídkakdy realizuje jako prostý. Na druhou stranu vzorce odvozené pro PNV jsou velmi oblíbené, jen výjimečně se lze setkat s jinými odhady.

Tím, že jsem ukázali, že moderní aparát pravděpodobnostních výběrů se pro prostý náhodný výběr shoduje s odhady v praxi běžně užívanými, jsme chtěli především v čtenáři vzbudit důvěru k těmto řídce používaným postupům. Cílem je, aby v případě výběrů již z principu prováděných jinak, než je prováděn PNV, byly používány přesnější metody a odhady.

Následující příklad demonstruje rozdíl výsledků při používání odhadů pro běžný výběr u výběru, který byl prováděn jako skupinkový.

Příklad 5.4 *Použitím dat získaných při výzkumu realizovaném na jednotlivcích rodin jsme zjišťovali rozdíl v odhadech chyby počítané pro skupinkový výběr od odhadu počítaného pro obyčejný výběr jednotlivců. Bylo vytipováno celkem 10 alternativně rozdělených veličin tak, aby bodové odhady jejich průměrů přibližně odpovídaly postupně hodnotám 2, 4, 7, 10, 15, 20, 25, 30, 40 a 50 %, a dostatečně tak pokrývaly interval* $\left(0, \frac{1}{2}\right]$. *Mezilehlé hodnoty pak byly odhadnuty pouze lineární interpolací.*

Pravděpodobnosti zahrnutí byly počítány jako pro oblastní výběr, v tomto případě se jednalo o proporcionální alokaci v jednotlivých krajích ČR, kdy se v každém kraji uvažuje prostý výběr o rozsahu úměrnému počtu domácností v kraji a výběry v různých krajích jsou vzájemně nezávislé. Pravděpodobnosti zahrnutí byly vypočítány zvlášť pro výběr jednotlivců a pro výběr domácností.

Velikost výběru lehce přesahovala 1000 domácností s více než 2500 jednotlivci. Zkonstruované intervalové odhady odpovídají odhadům pro celou populaci ČR.

Výsledné odhady jsou zachyceny na grafu 5.1, kde na vodorovné ose jsou vyneseny bodové odhady jednotlivých veličin a svislá osa určuje hodnoty pro dolní a horní mez 95% intervalu spolehlivosti. Plná čára pak odpovídá lineární interpolaci pro odhad konstruovaný jako pro výběr skupinek, zatímco čerchovaná odpovídá odhadu pro výběr jednotlivců.

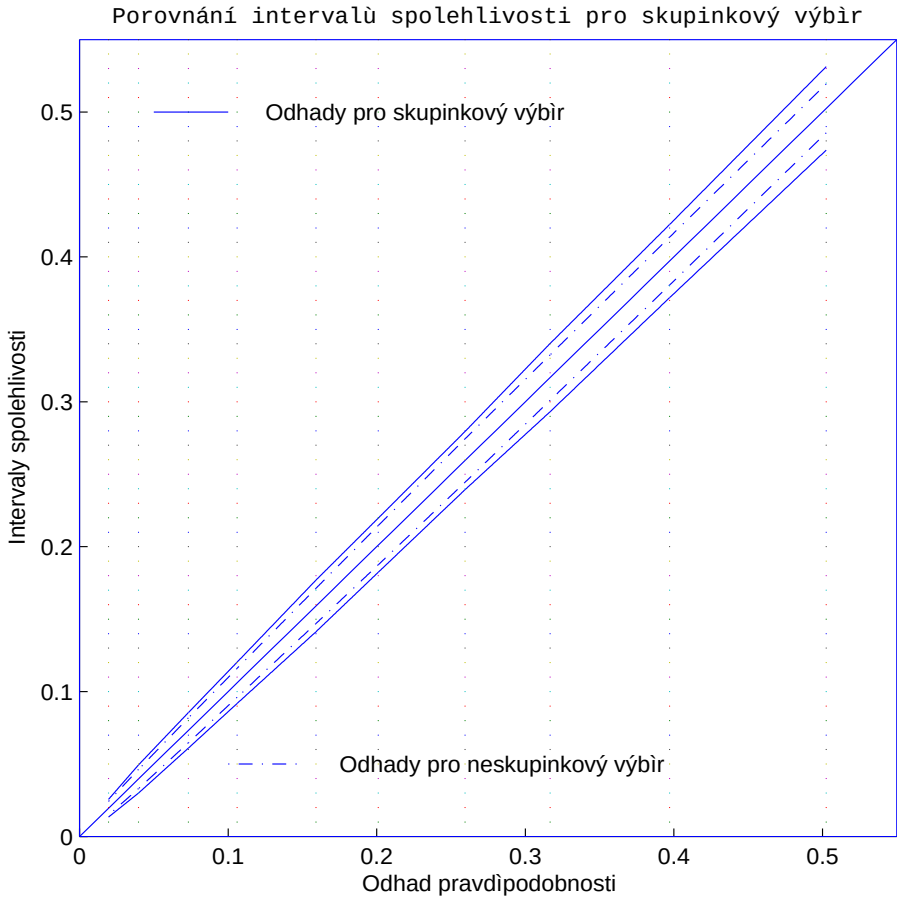
Pro relativní četnost blízkou polovině je konfidenční interval počítaný běžnou metodou skoro dvakrát užší, než by správně měl být. □

6 Bootstrap

Chceme-li zjistit statistické vlastnosti nějakého odhadu a nemáme-li k dispozici podrobnější informace o rozdělení odhadu, můžeme tyto vlastnosti odhadnout např. pomocí metody bootstrap. Metodu bootstrap navrhl Efron (1979). Bootstrap může být např. použit k odhadu rozptylu, vychýlení nebo ke stanovení intervalu spolehlivosti zkoumaného odhadu. Dále uvedený popis bootstrapu je založen na textu [30].

Předpokládejme, že odhadujeme nějaký populační parametr θ a z populace jsme vybrali n jednotek tvořící výběr s. Na vybraných jednotkách mějme k dispozici hodnoty $x_1, \ldots, x_n$ sledované veličiny, obecně však může jít o vektory hodnot sledovaných veličin. Na základě vybraných jednotek vytvoříme odhad

$$\hat{\theta}_0 = f(x_1, \ldots, x_n) \tag{55}$$



Graf 5.1: Porovnání intervalů spolehlivosti pro skupinkový výběr

neznámého parametru θ. Přejeme si nyní například zjistit, jaké je vychýlení odhadu $\hat{\theta_0}$ a stanovit jeho směrodatnou odchylku, případně stanovit 95% interval spolehlivosti. Za tímto účelem budeme z výběru s generovat B náhodných výběrů s vracením, každý o rozsahu n. V každém z generovaných výběrů se tak jednotka i z výběru s může vyskytovat i vícekrát, nebo naopak ani jednou. Označme nyní $\hat{\theta}_j$ odhad populačního parametru θ zkonstruovaný na základě j. bootstrapového výběru. Dále označme

$$\hat{\theta}_{\mathrm{B}} = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j. \tag{56}$$

Protože $\hat{\theta}_{\mathrm{B}}$ je bootstrapovým odhadem střední hodnoty $\mathrm{E}\,\hat{\theta}_0$, můžeme vychýlení odhadu $\hat{\theta}_0$ odhadnout jako

$$\widehat{\mathrm{bias}}(\hat{\theta}_0) = \hat{\theta}_{\mathrm{B}} - \hat{\theta}_0. \tag{57}$$

Odečteme-li odhad vychýlení $\widehat{\mathrm{bias}}(\hat{\theta}_0)$ od původního odhadu $\hat{\theta}_0$, získáme nový odhad opravený vzhledem k vychýlení

$$\hat{\theta}_{\mathrm{unbiased}} = 2\hat{\theta}_0 - \hat{\theta}_{\mathrm{B}}. \tag{58}$$

Pro odhad intervalu spolehlivosti pomocí bootstrapu bylo navrženo několik metod. Předpokládejme nejprve, že odhad $\hat{\theta}_0$ má normální rozdělení $N(\mu, \sigma^2)$. Označme

$$S_{\mathrm{B}}^2 = \frac{1}{n-1} \sum_{j=1}^{B} (\hat{\theta}_j - \hat{\theta}_{\mathrm{B}})^2 \tag{59}$$

výběrový rozptyl bootstrapových odhadů, který můžeme vzít za odhad parametru σ^2. $100(1-\alpha)\%$ interval spolehlivosti je pak dán vztahem

$$\hat{\theta}_0 \pm z_{1-\frac{\alpha}{2}} S_{\mathrm{B}} = \hat{\theta}_0 \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n-1} \sum_{j=1}^{B} (\hat{\theta}_j - \hat{\theta}_{\mathrm{B}})^2}, \tag{60}$$

kde $z_{1-\frac{\alpha}{2}}$ je kvantil rozdělení $N(0,1)$ (např. $z_{1-0{,}025} = 1{,}96$ pro 95% interval spolehlivosti). Použijeme-li odhad opravený o vychýlení, pak dostaneme interval spolehlivosti

$$2\hat{\theta}_0 - \hat{\theta}_B \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n-1} \sum_{j=1}^{B} (\hat{\theta}_j - \hat{\theta}_{\mathrm{B}})^2}. \tag{61}$$

Pro stanovení směrodatné odchylky by mělo dle [30] stačit provést 100 až 200 bootstrapových výběrů.

Přímější cestou ke stanovení $100(1-\alpha)\%$ intervalu spolehlivosti je využití rozdělení bootstrapových odhadů a stanovení dolního a horního $\alpha/2$ kvantilu přímo z tohoto rozdělení. Označme $\hat{\theta}_{L,\alpha/2}$ hodnotu, pod kterou padne $100\frac{\alpha}{2}$ % hodnot z bootstrapových odhadů $\hat{\theta}_j$ a $\hat{\theta}_{U,\alpha/2}$ hodnotu, nad kterou padne $100\frac{\alpha}{2}$ %

těchto bootstrapových odhadů. $100(1-\alpha)\%$ interval spolehlivosti je aproximován jako

$$[\hat{\theta}_{L,\alpha/2},\ \hat{\theta}_{U,\alpha/2}]. \tag{62}$$

Ke stanovení 95% intervalu spolehlivosti se podle [30] doporučuje provést alespoň 1000, pro stanovení 99% intervalu spolehlivosti alespoň 5000 bootstrapových výběrů. Při malém rozsahu výběru n pak s rostoucím počtem bootstrapových výběrů B dochází k jejich opakování a bootstrapový odhad již dále není zpřesňován.

Porovnání přesného intervalu spolehlivosti pro hodnoty $p \in [0,1]$ podílu výskytu sledovaného 0-1 znaku v populaci o velikosti $N = 10000$ a rozsahu výběru $n = 200$ s bootstrapovými odhady pro hodnoty $B = 10$ a $B = 100$ ukazuje graf 6.1.

7 Jackknife

Motivací pro odhady jackknife je redukce vychýlení. Metodu Jackknife navrhl Quenouille (1949). Jackknife může být použit nejen k redukci vychýlení odhadu, ale i k odhadu rozptylu. Jde především o situace, kdy neznáme vzorec pro odhad rozptylu nebo máme k dispozici jen vzorec přibližný. Příkladem může být odhad rozptylu odhadu poměru dvou úhrnů. Základní postup metody jackknife je popsán např. v [30]. Nechť

$$\hat{\theta}_0 = f(x_1, \ldots, x_n) \tag{63}$$

je odhad populačního parametru založeného na výběru s, který obsahuje n jednotek. Vytvořme nyní n výběrů, které vzniknou z výběru s odstraněním jedné náhodně vybrané jednotky. Může se stát, že některá jednotka byla odstraněna ve více výběrech, jiná nemusela být odstraněna nikdy. Označme

$$\hat{\theta}_{-j} = f(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_n) \tag{64}$$

odhad založený na výběru s, ze kterého byla odstraněna jednotka j. Definujme

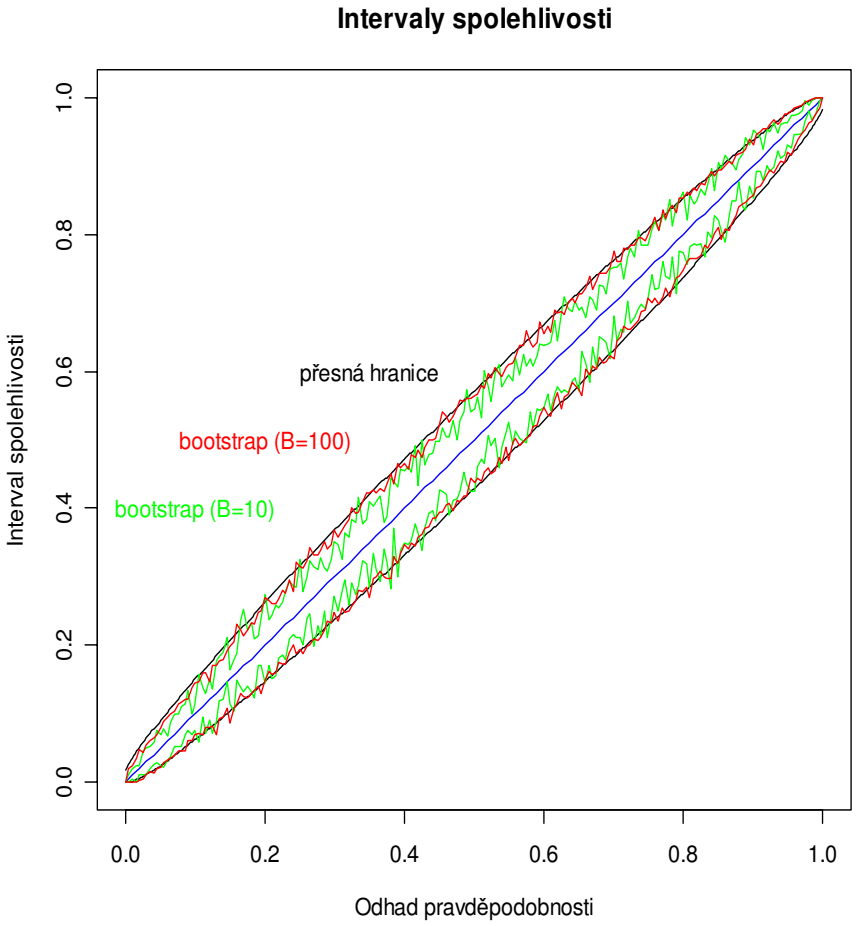
$$\hat{\theta}^*_j = n\hat{\theta}_0 - (n-1)\hat{\theta}_{-j}. \tag{65}$$

Jackknife odhadem $\hat{\theta}^*$ rozumíme odhad

$$\hat{\theta}^* = \frac{1}{n}\sum_{k=1}^{n}\hat{\theta}^*_k. \tag{66}$$

Výběrovou chybu odhadu $\hat{\theta}^*$ můžeme přibližně odhadnout

$$\widehat{\mathrm{var}}\left(\hat{\theta}^*\right) = \frac{\mathrm{var}\left(\hat{\theta}^*_j\right)}{n} = \frac{\sum_{k=1}^{n}\left(\hat{\theta}^*_k - \hat{\theta}^*\right)^2}{n(n-1)}. \tag{67}$$



Graf 6.1: Porovnání přesného intervalu spolehlivosti a bootstrapového odhadu

Pokud původní odhad $\hat{\theta}_0$ měl vychýlení řádu $1/n$

$$\mathrm{E}\,\hat{\theta}_0 = \theta\left(1 + \tfrac{C}{n}\right), \tag{68}$$

pro odhad $\hat{\theta}_{-j}$ platí

$$\mathrm{E}\,\hat{\theta}_{-j} = \theta\left(1 + \tfrac{C}{n-1}\right). \tag{69}$$

Pro hodnoty $\hat{\theta}_j^*$ máme

$$\mathrm{E}\,\hat{\theta}_j^* = n\,\mathrm{E}\,\hat{\theta}_0 - (n-1)\,\mathrm{E}\,\hat{\theta}_{-j} = \theta\left(n\left(1 + \tfrac{C}{n}\right) - (n-1)\left(1 + \tfrac{C}{n-1}\right)\right) = \theta. \tag{70}$$

Odtud je vidět, že odhad metodou jackknife odstraňuje vychýlení řádu $1/n$. Obecnější metody jackknife odstraňují z výběrového souboru m hodnot, čímž mohou dosáhnout odstranění vychýlení řádu $1/n^m$.

V obecnější situaci, kdy odstraňujeme z výběrového souboru m hodnot, můžeme postupovat např. následovně (viz [26]). Nejprve rozdělíme náhodným způsobem výběr s na A skupin stejné velikosti m, $m = n/A$ (pro jednoduchost se předpokládá, že takové rozdělení je možné). Pro každou skupinu a, $a = 1, \ldots A$, spočítáme odhad $\hat{\theta}_{-a}$, který počítáme stejným funkčním tvarem jako odhad $\hat{\theta}_0$, ale vynecháním jednotek ze skupiny a. Dále definujme pseudohodnoty

$$\hat{\theta}_a^* = A\hat{\theta}_0 - (A-1)\hat{\theta}_{-a}. \tag{71}$$

Jackknife odhadem $\hat{\theta}_{\mathrm{JK}}^*$ rozumíme odhad

$$\hat{\theta}_{\mathrm{JK}}^* = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a^*. \tag{72}$$

Jackknife odhad rozptylu odhadu $\hat{\theta}_{\mathrm{JK}}^*$ je

$$\widehat{\mathrm{var}}_{\mathrm{JK}\,1}\left(\hat{\theta}_{\mathrm{JK}}^*\right) = \frac{\sum_{a=1}^{A}\left(\hat{\theta}_a^* - \hat{\theta}_{\mathrm{JK}}^*\right)^2}{A(A-1)}. \tag{73}$$

Odhad rozptylu je používán nejen pro odhad $\hat{\theta}_{\mathrm{JK}}^*$, ale i pro původní odhad $\hat{\theta}_0$. Někdy se pro odhad rozptylu používá i

$$\widehat{\mathrm{var}}_{\mathrm{JK}\,2}\left(\hat{\theta}_{\mathrm{JK}}^*\right) = \frac{\sum_{a=1}^{A}\left(\hat{\theta}_a^* - \hat{\theta}_0^*\right)^2}{A(A-1)}, \tag{74}$$

přičemž platí $\widehat{\mathrm{var}}_{\mathrm{JK}\,2}\left(\hat{\theta}_{\mathrm{JK}}^*\right) \geq \widehat{\mathrm{var}}_{\mathrm{JK}\,1}\left(\hat{\theta}_{\mathrm{JK}}^*\right)$.

Postup pro případ stratifikovaného nebo vícestupňového výběru lze najít např. v [26]. Při použití metody jackknife musíme stanovit počet skupin A. Častou volbou bývá $A = n$, tedy $m = 1$. Z výpočetních důvodů se ale může preferovat kompromisní volba mezi hodnotami $A = 2$ a $A = n$.

8 Závěr

Výpočty statistické chyby v praxi výzkumných agentur zejména v oblasti výběrových šetření se omezují většinou na situace, kdy je zkoumanou veličinou výskyt určitého jevu v populaci. Vzhledem k potřebě rychlého a snadno dostupného odhadu chyby potom nutí klienti nepřímo výzkumníky k používání velmi aproximativního vzorce (21). Tento přístup má pochopitelně svá úskalí, zejména pro malé pravděpodobnosti, resp. pravděpodobnosti blízké jedné.

Vzhledem k výraznému vylepšení přesnosti v oblastech hodnot p blízkých nule nebo jedné při použití vzorců (23), resp. (25) by je bylo vhodné používat více, než je v současné praxi obvyklé. Vzhledem ke komplikovanější formě těchto vzorců je potom důležité, aby byly klientům „nestatistikům“ předávány ve formě tabulek nebo ještě lépe jako součást softwarů, které tito klienti používají pro potřeby svých analýz. Je zde nutné podotknout, že většina významných softwarů v oblasti výpočetní matematiky (Matlab, Mathematica) ale dnes již i softwary popularizační (MS Excel) umožňují provádět odhady intervalů spolehlivosti na základě přesných propočtů v přijatelném čase.

Je proto téměř více rozhodující neumdlévat ve statistické osvětě a rozšířit nebo spíše začít s popularizačními články v nestatistických časopisech, protože povědomí o statistické chybě sice existuje, ale povědomí o její velikost je velmi zamlženo a většinou ve svém důsledku je statistická chyba hrubě podceňována. Lze se setkat i s reakcemi typu „jestli je ta statistická chyba tak velká, potom máte špatně výzkum“, což vede k tomu, že statistici v praxi problém chyby někdy musí spíše skrývat, aby se nedostali do podezření, že „celý výzkum je špatně“.

Na druhou stranu v komplikovanějších situacích, kdy se nejedná o měření prostého výskytu jevu, nejsou statistické chyby zveřejňovány v naprosté většině případů vůbec. Vzorce (39), (40) jsou sice absolventům statistických oborů pochopitelně známy, ale jejich použití v praxi je velmi řídké. Navíc, protože situace v praxi je komplikovanější o problematiku vážení, je dobré používat analogii těchto vzorců (viz 4.4), která již ani předmětem základních kurzů není, a proto i její použití lze považovat za velmi řídké. Použití vzorců pro odhad chyby poměru dvou úhrnů (44), (45) je potom v praxi výzkumů v ČR rovněž velmi sporadické, přestože odhady tohoto typu (například podíly na trhu apod.) se běžně vyskytují. Autoři článku doporučují navíc použití jejich analogií na základě 4.4 v případě vážených odhadů.

Dalším obvyklým problémem, který rovněž není v praxi řešen zcela korektně, je porovnání výsledků dvou výzkumů. S tímto problémem se můžeme setkat téměř denně na stránkách denního tisku bez serióznějšího komentáře ohledně významnosti rozdílu. Jedná se o problematiku porovnání odhadů poskytovaných více agenturami k témuž časovému období, stejný problém však nastává i při kontinuálních výzkumech určitého jevu na jiných jedincích populace. Postupy uváděné zde v kapitole 3 by rovněž měly být přehlednou formou tabelizovány, případně by měly být dostupné v běžně používaných softwarech. Je potřeba poznamenat, že vzorce uváděné v [22], zde reprezentované postupem (30), nelze najít ani v běžně dostupných statistických softwarech.

Situace výběrových postupů a modelů a tedy i postupů pro výpočet chyby bývá však v praxi obvykle komplikovanější, než by odpovídalo přímému použití výše uvedených vzorců. Zde lze říci, že používané výběrové postupy jsou korektnější než v případě výpočtu statistické chyby, zejména proto, že kvalita výběru již přímo souvisí s kvalitou výzkumu jako takového, a proto je věcí profesionální cti i ekonomického zájmu výzkumných agentur, aby postupy byly promyšlené do větších detailů. V případě průzkumů „face-to-face“ se používají nejčastěji výběry „kvótní“, které aproximují stratifikované náhodné výběry, kde stratifikace probíhá většinou na regionální úrovni, reprezentativitě vůči pohlaví a věku, obvykle také vzdělání. Pro výpočet statistické chyby lze pak doporučit obecné postupy – viz (48), ale ani aproximace vzorci pro prosté náhodné výběry z kapitol 2, resp. 4 asi nepřinese hrubší chybu.

Postupy kapitoly 5, tzv. skupinkové výběry mohou hrát větší význam při specifických druzích výzkumů, kdy výběrovou jednotkou bývá domácnost, ale měřenými jednotkami potom všichni jednotlivci, resp. jednotlivci vybrané domácnosti s určitou vlastností (věková omezení, ekonomická omezení apod.). Tyto tzv. panelové výzkumy navíc obvykle mají tu specifičnost, že odpovědi jednotlivců se opakují s určitou periodou, obvykle denní (klasickým příkladem jsou panely zkoumající spotřebu produktů, panely pro elektronická měření sledovanosti televize, případně deníčkové panely pro poslechovost rozhlasu). Vlastní odhady používané v praxi bývají téměř výhradně odhady odvozené pro prostý náhodný výběr, stejně tak jako postupy pro odhad statistické chyby s tím, že problematika závislosti odpovědí jednotlivců v rámci panelu se obvykle zanedbává. Korektní statistické vyhodnocení pozorování ve dvou různých časových snímcích na stejném vzorku se potom většinou neprovádí, jedná se většinou pouze o „expertní“ posouzení velikosti odchylky.

Pro základní porovnání odhadů na stejném vzorku respondentů doporučují autoři postupy zmiňované v kapitole 3.2. Metody vedoucí k ještě korektnějšímu posouzení vývoje a vlivu různých částí populace na odhadované veličiny ať již založené na lineární regresní analýze nebo regresi loglineární přesahují rámec tohoto článku, bohužel však ve většině případů i možnosti skutečného pochopení a racionálního uplatnění klientskou veřejností.

Stále větší komplikovanost vzorců pro odhady chyb, která se začala projevovat již v kapitole 5, lze do jisté míry nahradit metodami zmiňovanými v posledních dvou kapitolách. Tyto metody jsou možná laické veřejnosti lépe objasnitelné, a tudíž i obhajitelné než zdánlivě komplikované vzorce. Myšlenka opakování výběru se stejnými vlastnostmi, resp. myšlenka porovnání odhadu v případě vynechaných pozorování jsou snadno vysvětlitelné, avšak problematické použití spočívá „pouze“ v nutnosti existence softwaru, který by na uživatelské úrovni statistické chyby počítal, případně nutnost přesvědčit klienta, že uvedené odhady by měly být součástí výsledné zprávy výzkumné agentury. Zde se však klient nechá přesvědčit poměrně snadno pouze v případě, kdy za tuto speciální statistickou službu nebude muset platit.

Na úplný závěr by rádi autoři vyjádřili záměr pokračovat v diskusi o statistické chybě, zejména s laickou veřejností. Na tento článek by rádi navázali řadou článků popularizačních, které by formou srozumitelnější širší veřejnosti

měly vést ke korektnějšímu používání odhadů statistické chyby v konkrétních praktických situacích. Na druhou stranu praxe přináší velmi zajímavé podněty i pro specializovaná statistická pracoviště ve formě poměrně komplikovaných zadání, jejichž správná řešení by stačila na několiměsíční či možná i roční vědeckou práci.

Reference

[1] Agresti, A., Coull, B. A. (1998). *Approximate is better than "exact" for interval estimation with binomial proportions*, American Statistician, 52, 119–126.

[2] Anděl, J. (1993). *Statistické metody*, Matfyzpress, Vydavatelství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze.

[3] Anděl, J. (2002). *Základy matematické statistiky*, Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Preprint.

[4] Blyth, C. R., Still, H. A. (1983). *Binomial confidence intervals*, Journal of the American Statistical Association, 78, 108–116.

[5] Brewer, K. (2002). *Combined Survey Sampling Inference*, London: Arnold.

[6] Brown, L. D., Cai, T. T., DasGupta, A. (—). *Interval Estimation for a Binomial Proportion*, http://ljsavage.wharton.upenn.edu/∼tcai/paper/binomial-statsci.pdf.

[7] Clopper, C. J. and Pearson, E. S. (1934). *The use of confidence interval or fiducial limits illustrated in the case of the binomial*, Biometrika, 26, 404–413.

[8] Cochran, W. G. (1950). *The comparison of percentages in matched samples*, Biometrika, 37, 256–266.

[9] Čermák, V. (1968). *Statistika II. díl*, SNTL a SVTL, Praha a Bratislava.

[10] Deming, W. E., and Stephan, F. F. (1940). *On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known*, The Annals of Mathematical Statistics, 11, 427–444.

[11] Deville, J. C., and Särndal, C.-E. (1992). *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association, 87, 376–382.

[12] Fleiss, J. L., (1981). *Statistical Methods for Rates and Proportions*, 2nd edn. Wiley, New York.

[13] Hájek, J. (1981). *Sampling from a finite populations*, Marcel Dekker, inc., New York.

[14] Hájek, J. (1960). *Teorie pravděpodobnostního výběru s aplikacemi na výběrová šetření*, Nakladatelství ČSAV, Praha.

[15] Chen, J., Sitter, R. R., and Wu, C. (2002). *Using Empirical Likelihood Methods to Obtain Range Restricted Weights in Regression Estimators for Surveys*, Biometrika, to appear.

[16] Janko, J. (1958). *Statistické tabulky*, Praha, Nakladatelství Československé akademie věd.

[17] Kott, P. S., Anderson, P. G., Nerman, O. (—). *Two-sided coverage intervals for small proportions based on survey data*, http://www.fcsm.gov/01papers/Kott.pdf.

[18] Likeš J., Machek, J. (1983). *Matematická statistika*, Praha, SNTL.

[19] McNemar, Q. (1949). *Psychological Statistics*, New York, John Wiley and Sons.

[20] Miettinen, O. S. and Nurminen, N. (1985). *Comparative analysis of two rates*, Statistics in Medicine, 4, 213–226.

[21] Neustadt, J. (1997). *Problematika vah ve výběrových šetřeních*, diplomová práce MFF UK.

[22] Newcombe, R. G. (1998). *Interval estimation for the difference between independent proportions: Comparison of eleven methods*, Statistics in Medicine, 17, 873–890.

[23] Newcombe, R. G. (1998). *Two-sided confidence intervals for the single proportion: Comparison of seven methods*, Statistics in Medicine, 857–872.

[24] Ord, J. K. (1968). *Approximations to Distribution Functions which are Hypergeometric Series*, Biometrika, 55, 1968.

[25] Rao, J. N. K., and Singh, A. C. (1997) *A Ridge-Shrinkage method for Range-Restricted Weight Calibration in Survey Sampling*, http://www.amstat.org/sections/srms/Proceedings/papers/1997_009.pdf.

[26] Sarndal, C. E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling*, New York, Springer-Verlag.

[27] Šmejcová, M. (1993). *Hypergeometrické rozdělení ve statistice*, diplomová práce MFF UK.

[28] Valliant, R., Dorfmann, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.*, New York: John Wiley & Sons.

[29] Valliant, R. (2002). *Variance Estimation for the General Regression Estimator*, Survey Methodology, 28, 103–114.

[30] Walsh, B. (2000). *Resampling Methods: Randomization. Tests, Jackknife and Bootstrap Estimators*, Lecture Notes for EEB, 596z.

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