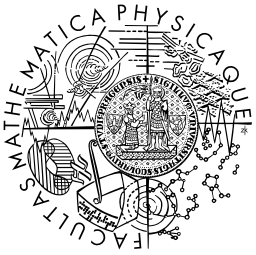


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Petr Klička

Kalibrační odhady ve výběrových šetřeních

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	Ing. Marek Omelka, Ph.D.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie

Praha 2018

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 19. července 2018 Petr Klička

Na tomto místě bych rád poděkoval svému vedoucímu, panu Ing. Marku Omelkovi, Ph.D., za jeho trpělivost, cenné rady, podněty a připomínky, které stály za vznikem této práce. Dále děkuji svým nejbližším, kteří mě podporovali během celého studia.

Název práce: Kalibrační odhady ve výběrových šetřeních

Autor: Petr Klička

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této práci se zabýváme odhady populačního úhrnu s využitím pomocných informací. V práci je popsán obecný regresní odhad a předpoklady, za kterých je splněna asymptotická normalita tohoto odhadu. Dále jsou zde popsány kalibrační odhady a zmínka o jejich asymptotické ekvivalenci s obecným regresním odhadem. Odvozené závěry aplikujeme na data z RADIOPROJEKTu a porovnáme je s výsledky získanými společnostmi, které tento projekt realizovali. Na závěr pomocí simulací porovnáme skutečné pravděpodobnosti pokrytí intervalů spolehlivosti pro populační úhrn spočítané na základě teorie uvedené v této práci a na základě metod společností realizujících RADIOPROJEKT.

Klíčová slova: výběrová šetření, populační úhrn, obecný regresní odhad, kalibrační odhad, kalibrační rovnice

Title: Calibration Estimators in Survey Sampling

Author: Petr Klička

Department: Department of Probability and Mathematical Statistics

Supervisor: Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this thesis we deal with the estimation of population totals with the use of auxiliary information. We describe the generalized regression estimator and assumptions of its asymptotic normality. We also describe calibration estimators and mention their asymptotic equivalence to the generalized regression estimator. We apply our obtained conclusions to data from the RADIOPROJEKT and we compare our results with results obtained by companies that worked on this project. Finally, we compare observed coverage probabilities of confidence intervals for population total based on the theory described in this thesis and also on methods used by the companies.

Keywords: survey sampling, population total, generalized regression estimator, calibration estimator, calibration equation

Obsah

Úvod

V této práci se budeme zabývat kalibračními odhady ve výběrových šetřeních, přesněji řečeno kalibračními odhady populačního úhrnu. Populační úhrn je základní ukazatel zkoumané vlastnosti v celé populaci, jedná se o součet hodnot zkoumané vlastnosti přes všechny jednotky v populaci. Pokud zkoumaná vlastnost nabývá pouze hodnot 0 a 1 (tj. například 1 značí, že jedinec z populace poslouchal rádio R), potom populační úhrn nám udává počet jednotek z populace, u nichž je zkoumaná vlastnost rovna 1 (tj. počet jedinců z populaci, kteří poslouchali rádio R). Tento populační úhrn se následně snažíme odhadnout za pomoci výběrových šetřeních. Zjednodušeně se dá říci, že výběrová šetření jsou postupy, které nám říkají, jak vybrat nějaký vzorek (výběr nebo také výběrový soubor) z populace, abychom získali odhad dané vlastnosti v populaci. V této práci se budeme zabývat odhadem poslechovosti rádií v České republice, dalším typickým příkladem jsou předvolební průzkumy a další.

V některých případech známe o daném výběru (případně celé populaci) i jiné informace než pouze tu o dané vlastnosti, jež je hlavním předmětem našeho zkoumání. Tyto pomocné informace, které lze získat například z dotazníků, z údajů ze statistického úřadu, případně z jiných zdrojů, chceme následně využít při hledání odhadu dané populační vlastnosti. Existuje několik možností, jak tyto pomocné informace využít, například regresní odhady nebo právě kalibrační odhady a mnohé další.

V první kapitole této práce jsou shrnuty základní pojmy používané ve výběrových šetřeních. Dále je zde zavedeno značení, které je používáno v celé práci, a také je zde uveden Horvitzův-Thompsonův odhad populačního úhrnu, který je jedním ze základních odhadů populačního úhrnu.

Druhá kapitola je věnována prostému náhodnému výběru bez vracení, který je následně využit při aplikaci v poslední části práce. Zásadní částí této kapitoly je asymptotická normalita a další asymptotické vlastností odhadu populačního úhrnu v případě prostého náhodného výběru.

Ve třetí kapitole jsou již při odhadování populačního úhrnu využity pomocné informace, které o výběru, případně o populaci, známe. Nejdříve je však uveden obecný regresní odhad založený na lineárním modelu a jeho asymptotické vlastnosti. Následně je ukázáno, jakým způsobem lze tento obecný regresní odhad vyjádřit jako kalibrační odhad. Dále jsou odvozeny obecné kalibrační odhady a také konkrétní případy těchto odhadů.

Ve čtvrté kapitole jsou představeny dvě možnosti, jak se vypořádat s kalibračními odhady, pokud známe četnosti v kontingenčních tabulkách. V prvním případě se jedná o tzv. *úplnou post-stratifikaci*, u které známe četnosti v každé buňce kontingenční tabulky. V druhém případě se věnujeme *neúplné post-stratifikaci*, u které místo četností v každé buňce známe pouze jednotlivé marginální četnosti.

Teorie odvozená v předchozích kapitolách je v páté kapitole aplikována na data o poslechovosti rádií osob trvale bydlících v České republice ve věku 12 – 79 let. Takto získané výsledky jsou dále porovnány s výsledky RADIOPROJEKTu, který realizují společnosti MEDIAN a STEM/MARK.

Šestá kapitola navazuje na předchozí kapitolu, je zde provedena simulační studie, ve které porovnáváme vlastnosti odhadů populačního úhrnu získané různými přístupy.

1. Základní pojmy

V úvodní části nejprve zavedeme základní pojmy, které se budou v práci dále vyskytovat. Populace U je soubor jednotek očíslovaných $1, \ldots, N$, tedy $U = \{1, \ldots, k, \ldots, N\}$, kde N je konečné. Výběrem s rozumíme podmnožinu populace U ($s \subseteq U$). U výběru s neuvažujeme uspořádání jednotek, ani jejich opakování (viz Vorlíčková, 1985, strana 7), celkem tedy existuje 2^N různých výběrů. Pravděpodobnost, že vybereme konkrétní výběr s označíme $p(s)$. Výběrovým plánem nazveme množinu rozdělení pravděpodobností $\{p(s), s \subseteq U\}$ splňující podmínky $\sum_{s \subseteq U} p(s) = 1$ a $p(s) \geq 0 \; \forall s \subseteq U$. Dále pro $k = 1, \ldots, N$ definujeme indikátory $I_k(s)$ zahrnutí jednotky k do výběru s, tj.

$$I_k(s) = \mathbb{I}(k \in s) = \begin{cases} 1, & \text{když } k \in s, \\ 0, & \text{když } k \notin s. \end{cases}$$

Počet jednotek ve výběru s nazveme rozsahem výběru a značíme jej $K(s)$, platí

$$K(s) = \sum_{k \in U} I_k(s).$$

Rozsah výběru $K(s)$ může být náhodná veličina (například u poissonovského výběru, kde $K(s)$ může nabývat hodnot $0, \ldots, N$, viz Vorlíčková, 1985, strana 14), ale u některých výběrů to může být stanovená konstanta (například u prostého náhodného výběru, viz Vorlíčková, 1985, strana 12).

Pravděpodobnosti zahrnutí jednotek do výběru s definujeme následovně:

$$\pi_k = \mathsf{P}(k \in s) = \mathsf{P}(I_k(s) = 1) = \sum_{s \ni k} p(s),$$
$$\pi_{kl} = \mathsf{P}(k, l \in s) = \mathsf{P}(I_k(s) I_l(s) = 1) = \sum_{s \ni k, l} p(s).$$

Z této definice vyplývá, že

$$\pi_{kl} = \pi_{lk} \text{ a dále}$$
$$\pi_{kk} = \mathsf{P}\left(I_k^2(s) = 1\right) = \mathsf{P}(I_k(s) = 1) = \pi_k.$$

Předpokládáme, že π_k a π_{kl} jsou kladné.

Sledovaný znak v populaci označíme y, tedy y_k je hodnota sledovaného znaku pro k-tou jednotku z populace U. Parametr populace U označíme $\theta(y_1, \ldots, y_N)$, neboť se jedná o funkci hodnot sledovaných znaků y všech jednotek v populaci U. Odhad parametru $\theta(y_1, \ldots, y_N)$ označíme $\hat{\theta}(s)$, jedná se tedy o odhad, který založíme na výběru s, jelikož hodnoty y_k, $k \in s$, známe, lze odhad $\hat{\theta}(s)$ spočítat. Základním parametrem, kterým se zabýváme v této práci, je populační úhrn

$$Y = \sum_{k \in U} y_k.$$

Dalším parametrem je populační průměr $\bar{Y}$, tj.

$$\bar{Y} = \frac{1}{N} \sum_{k \in U} y_k = \frac{Y}{N}.$$

Populační rozptyl definujeme následovně:

$$\sigma_y^2 = \frac{1}{N-1} \sum_{k \in U} \left(y_k - \bar{Y} \right)^2 .$$

Není-li uvedeno jinak, tak se v celé práci jedná o tzv. *design-based approach*, tedy náhoda spočívá v tom, jaké jednotky z populace U vybereme (hodnoty těchto jednotek jsou předem dané).

Definice 1. *Řekneme, že odhad $\hat{\theta}(s) = \hat{\theta}$ je designově nestranný (D-nestranný) odhad parametru $\theta = \theta(y_1, \ldots, y_N)$, jestliže pro všechny vektory $(y_1, \ldots, y_N)^\top \in \mathbb{R}^N$ platí*

$$\mathsf{E}\, \hat{\theta} = \sum_{s \subseteq U} p(s)\, \hat{\theta}(s) = \theta.$$

Poznámka 1. Z definice 1 vidíme, že střední hodnota je definována jako vážený průměr možných hodnot $\hat{\theta}(s)$ odhadu $\hat{\theta}$ s váhami $p(s)$.

△

Obdobně definujeme designový rozptyl odhadu $\hat{\theta}(s) = \hat{\theta}$ parametru $\theta = \theta(y_1, \ldots, y_N)$ (viz Särdnal a kol., 1992, strana 40):

$$\operatorname{var}\left(\hat{\theta}\right) = \sum_{s \subseteq U} p(s) \left(\hat{\theta}(s) - \mathsf{E}\left(\hat{\theta}\right) \right)^2 .$$

Designové vychýlení a designová střední čtvercová chyba odhadu jsou definovány standardním způsobem, tj. designové vychýlení odhadu $\hat{\theta}$:

$$B\left(\hat{\theta}\right) = \mathsf{E}\left(\hat{\theta}\right) - \theta$$

a designová střední čtvercová chyba odhadu $\hat{\theta}$:

$$\operatorname{MSE}\left(\hat{\theta}\right) = \mathsf{E}\left(\hat{\theta} - \theta\right)^2 = \sum_{s \subseteq U} p(s) \left(\hat{\theta}(s) - \theta \right)^2 .$$

1.1 Horvitzův-Thompsonův odhad

Jedním ze základních odhadů populačního úhrnu Y je Horvitzův-Thompsonův odhad

$$\hat{Y}_{HT} = \sum_{k \in s} d_k y_k, \tag{1.1}$$

kde $d_k = 1/\pi_k > 0$, neboť předpokládáme, že $\pi_k > 0$. Tento odhad je speciální případ lineárního odhadu, který lze zapsat ve tvaru $\hat{Y} = \sum_{k \in s} w_k y_k$, kde w_k, $k \in s$, jsou zvolené váhy (pro Horvitzův-Thompsonův odhad tedy platí, že $w_k = d_k$). Horvitzův-Thompsonův odhad je designově nestranný, tj.

$$\mathsf{E}\, \hat{Y}_{HT} = Y, \tag{1.2}$$

s designovým rozptylem

$$\mathrm{var}(\hat{Y}_{HT}) = \sum_{k\in U}\sum_{l\in U}\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) y_k y_l, \tag{1.3}$$

který lze odhadnout pomocí

$$\widehat{\mathrm{var}(\hat{Y}_{HT})} = \sum_{k\in s}\sum_{l\in s}\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right)\frac{1}{\pi_{kl}} y_k y_l = \sum_{k\in s}\sum_{l\in s}\left(\frac{1}{\pi_k\pi_l} - \frac{1}{\pi_{kl}}\right) y_k y_l, \tag{1.4}$$

(viz Särdnal a kol., 1992, Result 2.8.1).

Pro výběry s pevným rozsahem výběru $K(s)$ lze použít pro designový rozptyl Horvitzova-Thompsonova odhadu Yatesovu-Grundyho formuli

$$\mathrm{var}(\hat{Y}_{HT}) = -\frac{1}{2}\sum_{\substack{k\in U\\ k\neq l}}\sum_{l\in U}\left(\pi_{kl} - \pi_k\pi_l\right)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \tag{1.5}$$

Rozptyl Horvitzova-Thompsonova odhadu ve tvaru (1.5) můžeme odhadnout pomocí vztahu

$$\widehat{\mathrm{var}(\hat{Y}_{HT})} = -\frac{1}{2}\sum_{\substack{k\in s\\ k\neq l}}\sum_{l\in s}\left(1 - \frac{\pi_k\pi_l}{\pi_{kl}}\right)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2, \tag{1.6}$$

(viz Särdnal a kol., 1992, Result 2.8.2, Remark 2.8.3).

2. Prostý náhodný výběr

Jednou ze základních možností výběrů je prostý náhodný výběr bez vracení. Jedná se o výběr s pevným rozsahem n a je definován výběrovým plánem:

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}}, & \text{když } K(s) = n, \\ 0, & \text{když } K(s) \neq n. \end{cases}$$

Tedy všechny výběry o rozsahu n mají stejnou pravděpodobnost (viz Vorlíčková, 1985, strana 12).

V tomto případě máme následující vztahy pro pravděpodobnosti zahrnutí jednotek do výběru:

$$\begin{aligned} \pi_k &= \sum_{s \ni k} p(s) = \sum_{s \ni k} \frac{1}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, \\ \pi_{kl} &= \sum_{s \ni k,\, l} p(s) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}. \end{aligned} \tag{2.1}$$

Nyní zavedeme následující pojmy, výběrový průměr $\bar{y}$:

$$\bar{y} = \frac{1}{n} \sum_{k \in s} y_k,$$

odhad úhrnu $\hat{Y}$:

$$\hat{Y} = N\bar{y} = \frac{N}{n} \sum_{k \in s} y_k, \tag{2.2}$$

a také výběrový rozptyl s_y^2:

$$s_y^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y})^2 .$$

Nyní uvedeme několik základních vlastností týkajících se těchto pojmů, které platí u prostého náhodného výběru. Výběrový průměr $\bar{y}$ je D-nestranný odhad populačního průměru $\bar{Y}$. Designový rozptyl tohoto odhadu $\bar{y}$ je

$$\operatorname{var}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sigma_y^2 \tag{2.3}$$

a D-nestranný odhad tohoto rozptylu je

$$\widehat{\operatorname{var}(\bar{y})} = \left(1 - \frac{n}{N}\right) \frac{1}{n} s_y^2$$

(viz Särdnal a kol., 1992, Result 3.3.2).

Pro populační úhrn Y máme obdobné výsledky. Odhad $N\bar{y}$ je D-nestranným odhadem populačního úhrnu Y s designovým rozptylem

$$\operatorname{var}(N\bar{y}) = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \sigma_y^2 \tag{2.4}$$

a jeho D-nestranným odhadem

$$\widehat{\text{var}\,(N\bar{y})} = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} s_y^2 \tag{2.5}$$

(viz Särdnal a kol., 1992, Result 3.3.1).

Horvitzův-Thompsonův odhad pro prostý náhodný výběr získáme po dosazení výše uvedeného vzorce π_k do rovnice (1.1) a má tvar

$$\hat{Y}_{HT} = \sum_{k \in s} \frac{1}{\frac{n}{N}} y_k = \frac{N}{n} \sum_{k \in s} y_k = N\bar{y}. \tag{2.6}$$

Vzhledem k tomu, že prostý náhodný výběr má pevný rozsah výběru, lze designový rozptyl vypočítat také z Yatesovy-Grundyho formule (1.5). Můžeme si všimnout, že pro rozptyl odhadu $N\bar{y}$ populačního úhrnu Y máme nyní dvě vyjádření — pomocí vzorce (2.4), a také pomocí Yatesovy-Grundyho formule (1.5), která pro prostý náhodný výběr má tvar

$$-\frac{1}{2} \sum_{\substack{k \in U \\ k \neq l}} \sum_{l \in U} \left(\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \right) \frac{N^2}{n^2} \left(y_k - y_l\right)^2 .$$

Nyní ukážeme, že se jedná o shodná vyjádření. Začneme upravovat Yatesovu-Grundyho formuli, kterou upravíme do podoby vzorce (2.4).

$$\begin{aligned}
&-\frac{1}{2} \sum_{\substack{k \in U \\ k \neq l}} \sum_{l \in U} \left(\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \right) \frac{N^2}{n^2} \left(y_k - y_l\right)^2 \\
&\qquad = -\frac{1}{2} \left(\frac{N(n-1)}{n(N-1)} - 1 \right) \sum_{\substack{k \in U \\ k \neq l}} \sum_{l \in U} \left(y_k - y_l\right)^2 \\
&\qquad = -\frac{1}{2} \frac{n-N}{n(N-1)} \sum_{k \in U} \sum_{l \in U} \left(y_k^2 - 2 y_k y_l + y_l^2 \right) \\
&\qquad = -\frac{1}{2} \frac{n-N}{n(N-1)} \left(2N \sum_{k \in U} y_k^2 - 2 \left(Y\right)^2 \right) \\
&\qquad = -N \frac{n-N}{n(N-1)} \left(\sum_{k \in U} y_k^2 - N \left(\bar{Y}\right)^2 \right) \\
&\qquad = \left(\frac{N^2}{n} - N \right) \frac{1}{N-1} \sum_{k \in U} \left(y_k - \bar{Y} \right)^2 \\
&\qquad = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \sigma_y^2.
\end{aligned}$$

Vidíme, že oba vzorce se rovnají, tedy rozptyl odhadu $N\bar{y}$ populačního úhrnu Y lze vyjádřit dvěma způsoby, které jsou shodné.

2.1 Asymptotická normalita odhadu populačního úhrnu

Díky Horvitzově-Thompsonově odhadu pro prostý náhodný výběr máme odhad populačního úhrnu Y (2.6). Spolu s bodovým odhadem je pro úplnost vhodné

uvádět i interval spolehlivosti. Abychom mohli asymptotický interval spolehlivosti vypočítat v případě prostého náhodného výběru (bez vracení), uvedeme nyní obdobu centrální limitní věty pro tento případ.

U prostého náhodného výběru jednotky ve výběru s nejsou nezávislé, tedy nemůžeme použít centrální limitní větu pro nezávislé stejně rozdělené veličiny. Ovšem pro prostý náhodný výběr z konečné populace existuje varianta centrální limitní věty (viz Thompson, 2012, Kapitola 3.2.). Tuto centrální limitní větu si nyní popíšeme.

Uvažujme posloupnost konečných populací $U_1, \ldots, U_N, \ldots$, kde U_N obsahuje N jednotek, konkrétně označíme-li sledovaný znak v populaci y, pak $U_N = \{y_{1N}, \ldots, y_{NN}\}$. Z populace U_N děláme prostý náhodný výběr s_N s rozsahem n_N. Populační průměr této populace označme $\bar{Y}_N = \frac{1}{N}\sum_{k\in U_N} y_{kN}$ a výběrový průměr $\bar{y}_N$, tj. $\bar{y}_N = \frac{1}{n_N}\sum_{k\in s_N} y_{kN}$. Populační úhrn této populace označme $Y_N = \sum_{k\in U_N} y_{kN}$, Horvitzův-Thompsonův odhad tohoto úhrnu označme $\hat{Y}_N$, tj. $\hat{Y}_N = N\bar{y}_N$. Dále pro každé $\epsilon > 0$ označme $A_N(\epsilon)$ množinu jednotek z populace U_N, které jsou „příliš vzdálené“ od populačního průměru $\bar{Y}_N$, tj.

$$A_N(\epsilon) = \left\{ k \in U_N : \left| y_{kN} - \bar{Y}_N \right| > \epsilon \sqrt{\frac{n_N}{N}\left(1 - \frac{n_N}{N}\right) \sum_{l\in U_N} \left(y_{lN} - \bar{Y}_N\right)^2} \right\}. \quad (2.7)$$

Za platnosti předpokladů a značení výše platí:

$$\frac{\hat{Y}_N - Y_N}{\sqrt{\mathrm{var}\left(\hat{Y}_N\right)}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1) \quad \Leftrightarrow \quad \lim_{N\to\infty} \frac{\sum_{k\in A_N(\epsilon)} \left(y_{kN} - \bar{Y}_N\right)^2}{\sum_{k\in U_N} \left(y_{kN} - \bar{Y}_N\right)^2} = 0 \quad \forall \epsilon > 0. \quad (2.8)$$

Nyní se budeme zabývat otázkou, za jakých předpokladů je splněna pravá strana ekvivalence (2.8). Nejprve se zaměříme na množinu $A_N(\epsilon)$ definovanou vztahem (2.7), uvedenou nerovnost lze zapsat ve tvaru:

$$\left| y_{kN} - \bar{Y}_N \right| > \epsilon\sqrt{N} \sqrt{\frac{n_N}{N}\left(1 - \frac{n_N}{N}\right) \frac{1}{N} \sum_{l\in U_N} \left(y_{lN} - \bar{Y}_N\right)^2},$$

z čehož pro každé $\delta > 0$ a každé $k \in A_N(\epsilon)$ vyplývá nerovnost

$$\left(\frac{\left| y_{kN} - \bar{Y}_N \right|}{\epsilon\sqrt{N}\sqrt{\frac{n_N}{N}\left(1 - \frac{n_N}{N}\right) \frac{1}{N} \sum_{l\in U_N} \left(y_{lN} - \bar{Y}_N\right)^2}} \right)^{\delta} > 1. \quad (2.9)$$

Pro zbývající část kapitoly 2 předpokládejme, že

(a) existuje $\sigma^2 \in (0, \infty)$ takové, že $\lim_{N\to\infty} \sigma^2_{yN} = \sigma^2$, kde σ^2_{yN} je populační rozptyl populace U_N, tj. $\sigma^2_{yN} = \frac{1}{N-1}\sum_{k\in U_N} \left(y_{kN} - \bar{Y}_N\right)^2$

(b) $\lim_{N\to\infty} n_N = \infty$,

(c) $\limsup_{N\to\infty} \frac{n_N}{N} < 1$.

Pokud navíc předpokládáme, že

- existuje $\delta > 0$ takové, že $\limsup_{N\to\infty} \frac{1}{N-1} \sum_{k\in U_N} \left|y_{kN} - \bar{Y}_N\right|^{2+\delta} < \infty$,

tak můžeme ukázat, že pravá strana ekvivalence (2.8) je splněna, a tedy platí, že $\frac{\hat{Y}_N - Y_N}{\sqrt{\mathrm{var}\left(\hat{Y}_N\right)}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1)$. To nyní ukážeme:

$$
\begin{aligned}
&\lim_{N\to\infty} \frac{\sum_{k\in A_N(\epsilon)} \left(y_{kN} - \bar{Y}_N\right)^2}{\sum_{k\in U_N} \left(y_{kN} - \bar{Y}_N\right)^2} \\
&\quad = \lim_{N\to\infty} \frac{\sum_{k\in A_N(\epsilon)} \left(y_{kN} - \bar{Y}_N\right)^2}{\sum_{k\in U_N} \left(y_{kN} - \bar{Y}_N\right)^2} \cdot \frac{\frac{1}{N-1}}{\frac{1}{N-1}} \\
&\quad \leq \lim_{N\to\infty} \frac{\frac{1}{N-1} \sum_{k\in A_N(\epsilon)} \left(y_{kN} - \bar{Y}_N\right)^2}{\sigma^2} \\
&\quad \leq \frac{1}{\sigma^2} \lim_{N\to\infty} \frac{\frac{1}{N-1} \sum_{k\in A_N(\epsilon)} \left|y_{kN} - \bar{Y}_N\right|^{2+\delta}}{\left(\epsilon\sqrt{N}\sqrt{\frac{n_N}{N}\left(1 - \frac{n_N}{N}\right)\frac{1}{N}\sum_{l\in U_N}\left(y_{lN} - \bar{Y}_N\right)^2}\right)^{\delta}} \qquad (2.10) \\
&\quad \leq \frac{1}{\epsilon^{\delta}\sigma^{2+\delta}} \lim_{N\to\infty} \frac{\frac{1}{N-1} \sum_{k\in U_N} \left|y_{kN} - \bar{Y}_N\right|^{2+\delta}}{\left(\sqrt{N}\right)^{\delta}\left(\sqrt{\frac{n_N}{N}\left(1 - \frac{n_N}{N}\right)}\right)^{\delta}} = 0,
\end{aligned}
$$

kde nerovnost (2.10) je splněna díky tomu, že pro každé $k \in A_N(\epsilon)$ je $k-$tý člen sumy v čitateli přenásoben výrazem $\frac{\left|y_{kN} - \bar{Y}_N\right|^{\delta}}{\left(\epsilon\sqrt{N}\sqrt{\frac{n_N}{N}\left(1-\frac{n_N}{N}\right)\frac{1}{N}\sum_{l\in U_N}\left(y_{lN} - \bar{Y}_N\right)^2}\right)^{\delta}} > 1$, kde uvedená nerovnost platí díky vzorci (2.9).

2.1.1 Designově konzistentní odhad

Výše jsme ukázali, za jakých předpokladů platí asymptotická normalita populačního průměru. Nyní chceme ve zlomku $(\hat{Y}_N - Y_N)/\sqrt{\mathrm{var}(\hat{Y}_N)}$ nahradit rozptyl Horvitzova-Thompsonova odhadu $\hat{Y}_N$ za jeho odhad a zároveň požadujeme, aby byla normalita zachována. Za tímto účelem nejdříve uvedeme definici designově konzistentního odhadu a pomocné lemma a následně ukážeme, za jakých předpokladů můžeme dříve zmíněný rozptyl nahradit za jeho odhad.

Definice 2. *Řekneme, že odhad $\hat{\theta}(s) = \hat{\theta}_{n_N}$ je designově konzistentní (D-konzistentní) odhad parametru $\theta_N = \theta(y_{1N}, \ldots, y_{NN})$, jestliže*

$$\forall \epsilon > 0: \quad \mathsf{P}\left(\left|\hat{\theta}_{n_N} - \theta_N\right| \geq \epsilon\right) \xrightarrow[N\to\infty]{} 0.$$

Dříve než uvedeme pomocné lemma, se zaměříme na populační průměry. Konkrétně v celé práci budeme předpokládat, že sledované znaky v_{kN}, $k \in U_N$, jsou konečné a navíc pro $N \to \infty$ platí, že

$$\lim_{N\to\infty} \bar{V}_N \in (-\infty, \infty), \qquad (2.11)$$

kde $\bar{V}_N = \frac{1}{N}\sum_{k\in U_N} v_{kN}$.

Lemma 1. *Uvažujme posloupnost konečných populací* $U_1, \ldots, U_N, \ldots$, *kde populace* U_N *obsahuje* N *jednotek. Pro* $k-$*tou jednotku populace* U_N *uvažujme znaky* $\{v_{kN}, k \in U_N\}$, *kde* $v_{kN} \in \mathbb{R}$. *Z populace* U_N *děláme prostý náhodný výběr* s_N *s rozsahem* n_N. *Dále předpokládejme, že platí*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k\in U_N} v_{kN}^2 < \infty. \tag{2.12}$$

Označíme-li populační průměr $\bar{V}_N = \frac{1}{N}\sum_{k\in U_N} v_{kN}$ *a výběrový průměr* $\bar{v}_N = \frac{1}{n_N}\sum_{k\in s_N} v_{kN}$, *potom platí*

$$\bar{v}_N - \bar{V}_N \xrightarrow[N\to\infty]{\mathcal{P}} 0.$$

Poznámka 2. Pokud bychom nepředpokládali, že sledované znaky v_{kN}, $k \in U_N$, nabývají konečných hodnot (viz krátká diskuze před lemmatem 1), tak bychom předpoklad (2.12) v lemmatu 1 museli nahradit za předpoklad

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k\in U_N} \left(v_{kN} - \bar{V}_N\right)^2 < \infty.$$

△

Důkaz. Pro důkaz lemmatu aplikujeme Čebyševovu nerovnost, dostaneme tedy

$$\forall\, \epsilon > 0 :\ \mathsf{P}\left(\left|\bar{v}_N - \bar{V}_N\right| > \epsilon\right) \le \frac{\operatorname{var}(\bar{v}_N)}{\epsilon^2} \xrightarrow[N\to\infty]{} 0,$$

neboť díky vztahu (2.3) máme, že

$$\operatorname{var}(\bar{v}_N) = \frac{1}{n_N}\left(1 - \frac{n_N}{N}\right)\frac{1}{N-1}\sum_{k\in U_N}\left(v_{kN} - \bar{V}_N\right)^2 \xrightarrow[N\to\infty]{} 0.$$

❑

Poznámka 3. Čebyševovu nerovnost v předchozím důkazu můžeme použít díky tomu, že z populace U_N děláme prostý náhodný výběr s_N a tedy platí, že výběrový průměr $\bar{v}_N$ je designově nestranný odhad populačního průměru $\bar{V}_N$.

△

Předpokládáme-li navíc, že

(A) $\lim_{N\to\infty} \frac{1}{N} \sum_{k\in U_N} y_{kN}^4 < \infty$,

lze ukázat, že potom platí $\frac{\widehat{\operatorname{var}\left(\hat{Y}_N\right)}}{\operatorname{var}\left(\hat{Y}_N\right)} = \frac{N^2\widehat{\operatorname{var}(\bar{y}_N)}}{N^2\operatorname{var}(\bar{y}_N)} = \frac{\widehat{\operatorname{var}(\bar{y}_N)}}{\operatorname{var}(\bar{y}_N)} \xrightarrow[N\to\infty]{\mathcal{P}} 1$. Nejprve ukážeme, že $\bar{y}_N$ je D-konzistentní odhad $\bar{Y}_N$. Díky faktu, že $\bar{y}_N$ je D-nestranný odhad $\bar{Y}_N$ a předpokladu (a), máme splněny předpoklady lemmatu 1, kde $v_{kN} = y_k$, tedy jsme ukázali, že $\bar{y}_N$ je D-konzistentní odhad $\bar{Y}_N$. Nyní za pomoci předpokladů (A)

a faktu (2.11) ukážeme, že s^2_{yN} je D-konzistentní odhad σ^2_{yN}, kde s^2_{yN} je výběrový rozptyl populace U_N, tj. $s^2_{yN} = \frac{1}{n_N-1}\sum_{k\in s_N}(y_{kN}-\bar{y}_N)^2$. Platí:

$$\begin{aligned}
s^2_{yN} - \sigma^2_{yN} &= \frac{1}{n_N-1}\sum_{k\in s_N}(y_{kN}-\bar{y}_N)^2 - \frac{1}{N-1}\sum_{k\in U_N}\left(y_{kN}-\bar{Y}_N\right)^2 \\
&= \left(\frac{1}{n_N-1}\sum_{k\in s_N} y^2_{kN} - \frac{n_N}{n_N-1}(\bar{y}_N)^2\right) - \left(\frac{1}{N-1}\sum_{k\in U_N} y^2_{kN} - \frac{N}{N-1}\left(\bar{Y}_N\right)^2\right) \\
&= \left(\frac{n_N}{n_N-1}\sum_{k\in s_N}\frac{y^2_{kN}}{n_N} - \frac{N}{N-1}\sum_{k\in U_N}\frac{y^2_{kN}}{N}\right) - \left(\frac{n_N}{n_N-1}(\bar{y}_N)^2 - \frac{N}{N-1}\left(\bar{Y}_N\right)^2\right) \\
&= C_N - D_N. \qquad (2.13)
\end{aligned}$$

Nejprve se zaměříme na člen C_N z výše uvedeného vztahu (2.13). Díky předpokladu (A) máme splněny předpoklady lemmatu 1 pro $v_{kN} = y^2_k$ a dále víme, že $\frac{n_N}{n_N-1} \xrightarrow[N\to\infty]{} 1$, $\frac{N}{N-1} \xrightarrow[N\to\infty]{} 1$. Tedy celkem dostáváme, že

$$C_N \xrightarrow[N\to\infty]{\mathcal{P}} 0.$$

Nyní se zaměříme na druhý člen D_N vztahu (2.13). Díky vztahům $\frac{n_N}{n_N-1} \xrightarrow[N\to\infty]{} 1$, $\frac{N}{N-1} \xrightarrow[N\to\infty]{} 1$, dále díky tomu, že máme splněny předpoklady lemmatu 1, tentokrát pro $v_{kN} = y_k$ (s využitím předpokladu (a)) a díky vztahu (2.11) můžeme použít větu o spojité transformace (viz poznámka 4) a celkem dostáváme, že

$$D_N \xrightarrow[N\to\infty]{\mathcal{P}} 0.$$

Poznámka 4. Věta o spojité transformaci je aplikována následovně.

- Z předchozí části díky lemmatu 1 pro $v_{kN} = y_k$ víme, že $\bar{y}_N - \bar{Y}_N \xrightarrow[N\to\infty]{\mathcal{P}} 0$.
- Víme, že $\bar{Y}_N \xrightarrow[N\to\infty]{} C \in (-\infty, \infty)$ (viz vztah (2.11)).
- Celkem tedy dostáváme, že $\bar{y}_N \xrightarrow[N\to\infty]{\mathcal{P}} C$.

Chceme ukázat, že $g(\bar{y}_N) - g\left(\bar{Y}_N\right) \xrightarrow[N\to\infty]{} 0$, pro $g(\cdot)$ spojitou funkci.

$$g(\bar{y}_N) - g\left(\bar{Y}_N\right) = g(\bar{y}_N) - g(C) + g(C) - g\left(\bar{Y}_N\right) \xrightarrow[N\to\infty]{\mathcal{P}} 0,$$

neboť $g(\bar{y}_N) - g(C) \xrightarrow[N\to\infty]{\mathcal{P}} 0$ z věty o spojité transformaci pro g spojitou v C a $g(C) - g\left(\bar{Y}_N\right) \xrightarrow[N\to\infty]{} 0$.

△

Ze vztahu (2.13) tedy celkem dostáváme, že

$$s^2_{yN} - \sigma^2_{yN} \xrightarrow[N\to\infty]{\mathcal{P}} 0 + 0 = 0,$$

tedy s^2_{yN} je D-konzistentní odhad σ^2_{yN}.

Použijeme-li předpoklad (a), dostaneme

$$\frac{\widehat{\text{var}\left(\hat{Y}_N\right)}}{\text{var}\left(\hat{Y}_N\right)} = \frac{\widehat{\text{var}\left(\bar{y}_N\right)}}{\text{var}\left(\bar{y}_N\right)} = \frac{\left(1-\frac{n_N}{N}\right)\frac{1}{n_N}s_{yN}^2}{\left(1-\frac{n_N}{N}\right)\frac{1}{n_N}\sigma_{yN}^2} \xrightarrow[N\to\infty]{\mathcal{P}} 1. \tag{2.14}$$

Tedy za platnosti výše uvedených předpokladů a díky Cramérově-Slutského větě získáváme vztah:

$$\frac{\hat{Y}_N - Y_N}{\sqrt{\widehat{\text{var}\left(\hat{Y}_N\right)}}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1). \tag{2.15}$$

3. Odhad populačního úhrnu Y při využití pomocné informace

V některých situacích v praxi máme k dispozici pomocné informace o populaci U (například údaje ze statistického úřadu). V takovém případě se snažíme tyto informace využít při odhadu populačního úhrnu Y.

Pro k-tou jednotku populace U uvažujeme vektor pomocných hodnot $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^\top$. Rozlišujeme dva případy (viz Särndal, 2010, strana 102):

- vektor $\mathbf{x}_k$ známe pro každé $k \in U$ (úplná pomocná informace),
- vektor $\mathbf{x}_k$ známe pro každé $k \in s$, a dále známe populační úhrn $\mathbf{X}$ hodnot $\mathbf{x}_k$, tj. $\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$.

V této práci se zabýváme druhým případem, tedy k dispozici máme data (y_k, $\mathbf{x}_k$) pro $k \in s$, známe populační úhrn $\mathbf{X}$ a naším hlavním cílem je odhadnout populační úhrn Y.

V této kapitole se budeme zabývat odhady populačního úhrnu Y, které využívají pomocné informace obsažené ve vektorech $\mathbf{x}_k$, $k \in U$, ale v našem případě jsou známy pouze pro jedince z populace, kteří jsou ve výběru s. Naším cílem nyní je pomocí těchto hodnot vylepšit původní výběrové váhy z Horvitzova-Thompsonova odhadu $d_k = 1/\pi_k$ tak, abychom získali přesnější odhad populačního úhrnu Y.

3.1 Obecný regresní odhad

Jedna z možností, jak při odhadu úhrnu Y zohlednit pomocné informace, je za pomoci obecného regresního odhadu, který lze zapsat ve tvaru lineárního odhadu

$$\hat{Y} = \sum_{k \in s} w_k y_k. \tag{3.1}$$

Obecný regresní odhad je založený na lineárním modelu

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + e_k, \quad k \in 1, \ldots, N, \tag{3.2}$$

kde $e_1, \ldots, e_N$ jsou nezávislé náhodné veličiny, $\mathsf{E}_M e_k = 0$, $k \in U$, $\mathrm{var}_M e_k = \sigma_k^2$, $k \in U$, kde E_M a var_M jsou střední hodnota a rozptyl vzhledem k modelu (3.2), tj. vzhledem k rozdělení náhodných veličin $e_1, \ldots, e_N$ a $\sigma_k^2, k \in U$, známe a $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^\top$ je regresní parametr (např. viz Särndal, 2010, strana 103). Odhad parametru $\boldsymbol{\beta}$ můžeme klasicky získat metodou vážených nejmenších čtverců jako řešení normálních rovnic:

$$\left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2} \right) \boldsymbol{\beta}_N = \sum_{k \in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2}, \tag{3.3}$$

tedy pokud lze matici $\left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2} \right)$ invertovat, lze tento odhad zapsat ve tvaru

$$\boldsymbol{\beta}_N = \left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2}. \tag{3.4}$$

Odhad $\boldsymbol{\beta}_N$ je nejlepší lineární nestranný odhad parametru $\boldsymbol{\beta}$, ovšem odpovídá hypotetickému populačnímu odhadu, který neznáme, neboť máme pouze data $(y_k, \mathbf{x}_k), k \in s$. Vzhledem k tomu, že populační odhad $\boldsymbol{\beta}_N$ neznáme, tak se v praxi využívá odhad $\hat{\boldsymbol{\beta}}_s$ založený na konkrétním výběru s (na základě Horvitzova-Thompsonova odhadu, viz Särdnal a kol., 1992, strana 227), tedy $\hat{\boldsymbol{\beta}}_s$ nalezneme jako řešení normálních rovnic založených na konkrétním výběru s:

$$\left(\sum_{k\in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2}\right) \hat{\boldsymbol{\beta}}_s = \sum_{k\in s} \frac{d_k \mathbf{x}_k y_k}{\sigma_k^2}. \tag{3.5}$$

Pokud lze matici $\left(\sum_{k\in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2}\right)$ invertovat, tak tento odhad lze zapsat ve tvaru:

$$\hat{\boldsymbol{\beta}}_s = \left(\sum_{k\in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2}\right)^{-1} \sum_{k\in s} \frac{d_k \mathbf{x}_k y_k}{\sigma_k^2}. \tag{3.6}$$

Hlavní myšlenka odhadu je založena na vyrovnaných hodnotách $\hat{y}_k, k \in U$, na základě modelu, v našem případě $\hat{y}_k = \hat{\boldsymbol{\beta}}_s^\top \mathbf{x}_k$ (viz Särndal, 2010, strana 103) a také na faktu, že populační úhrn Y můžeme napsat následovně:

$$Y = \sum_{k\in U} y_k = \sum_{k\in U} \hat{y}_k + \sum_{k\in U} (y_k - \hat{y}_k). \tag{3.7}$$

Druhý člen na pravé straně rovnice (3.7) odhadneme pomocí Horvitzova-Thompsonova odhadu a za pomoci vyrovnaných hodnot $\hat{y}_k$ sestavíme regresní odhad úhrnu Y následovně:

$$\begin{aligned}\hat{Y}_{reg} &= \sum_{k\in U} \hat{y}_k + \sum_{k\in s} d_k(y_k - \hat{y}_k) = \sum_{k\in s} d_k y_k + \left(\sum_{k\in U} \hat{y}_k - \sum_{k\in s} d_k \hat{y}_k\right) \\ &= \hat{Y}_{HT} + \hat{\boldsymbol{\beta}}_s^\top \mathbf{X} - \hat{\boldsymbol{\beta}}_s^\top \hat{\mathbf{X}}_{HT},\end{aligned} \tag{3.8}$$

kde $\hat{\mathbf{X}}_{HT} = \sum_{k\in s} d_k x_k$, je Horvitzův-Thompsonův odhad populačního úhrnu $\mathbf{X}$. Předtím než se zaměříme na rozptyl regresního odhadu $\hat{Y}_{reg}$, zapíšeme tento odhad ve tvaru (3.1), kde váhy w_k získáme postupnými úpravami (3.8) po dosazení $\hat{\boldsymbol{\beta}}_s^\top$, které získáme ze vzorce (3.6):

$$\begin{aligned}\hat{Y}_{reg} &= \hat{Y}_{HT} + \hat{\boldsymbol{\beta}}_s^\top \mathbf{X} - \hat{\boldsymbol{\beta}}_s^\top \hat{\mathbf{X}}_{HT} \\ &= \sum_{k\in s} d_k y_k + \sum_{k\in s} \frac{d_k \mathbf{x}_k^\top y_k}{\sigma_k^2} \left(\sum_{k\in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2}\right)^{-1} \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right) \\ &= \sum_{k\in s} \left[d_k + \frac{d_k \mathbf{x}_k^\top}{\sigma_k^2} \left(\sum_{k\in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2}\right)^{-1} \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right)\right] y_k.\end{aligned} \tag{3.9}$$

Tedy ze vztahu (3.9) dostáváme následující vztah pro váhy w_k:

$$w_k = d_k \left(1 + \frac{\mathbf{x}_k^\top}{\sigma_k^2} \left(\sum_{k\in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2}\right)^{-1} \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right)\right) = d_k g_k, k \in s, \tag{3.10}$$

kde $g_k = 1 + \frac{\mathbf{x}_k^\top}{\sigma_k^2} \left(\sum_{k\in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2}\right)^{-1} \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right)$.

3.1.1 Asymptotická normalita obecného regresního odhadu

Nyní se zaměříme na asymptotickou normalitu regresního odhadu $\hat{Y}_{reg}$ z rovnice (3.8), především se budeme věnovat asymptotickému rozptylu tohoto odhadu. Vzhledem k tomu, že se v této části budeme zabývat asymptotickými vlastnostmi, musíme stejně jako v části 2.1 uvažovat posloupnost konečných populací $U_1, \ldots, U_N, \ldots$, kde U_N obsahuje N jednotek, konkrétně $U_N = \{y_{1N}, \ldots, y_{NN}\}$. Z populace U_N děláme výběr s_N s rozsahem n_N (pokud se jedná o výběr s pevným rozsahem výběru, platí, že $K(s) = n_N$, v případě výběrů, kde rozsah výběru je náhodná veličina, platí, že $\mathsf{E}\, K(s) = n_N$). Pro rozsah výběru n_N předpokládáme, že platí $\lim_{N\to\infty} n_N = \infty$ a $\limsup_{N\to\infty} \frac{n_N}{N} < \infty$. Obdobně jako v části 2.1 budeme spodním indexem N značit jednotlivé znaky, parametry a další charakteristiky vypočítané z populace U_N, například $\hat{Y}_{HT,N} = \sum_{k\in s_N} d_{kN} y_{kN}$ je Horvitzův-Thompsonův odhad populačního úhrnu $Y_N = \sum_{k\in U_N} y_{kN}$, $d_{kN} = 1/\pi_{kN}$, $\pi_{kN} = \mathsf{P}\,(k \in s_N)$.

Pomocí značení zmíněného výše můžeme regresní odhad $\hat{Y}_{reg,N}$ z rovnice (3.8) rozepsat následovně:

$$\begin{aligned}\hat{Y}_{reg,N} &= \hat{Y}_{HT,N} + \hat{\boldsymbol{\beta}}_{s_N}^{\top}\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right) \\ &= \hat{Y}_{HT,N} + \boldsymbol{\beta}_N^{\top}\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right) + \left(\hat{\boldsymbol{\beta}}_{s_N} - \boldsymbol{\beta}_N\right)^{\top}\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right), \quad (3.11)\end{aligned}$$

kde $\boldsymbol{\beta}_N = \left(\sum_{k\in U_N} \frac{\mathbf{x}_{kN}\mathbf{x}_{kN}^{\top}}{\sigma_k^2}\right)^{-1} \sum_{k\in U_N} \frac{\mathbf{x}_{kN} y_{kN}}{\sigma_k^2}$

a $\hat{\boldsymbol{\beta}}_{s_N} = \left(\sum_{k\in s_N} \frac{d_{kN}\mathbf{x}_{kN}\mathbf{x}_{kN}^{\top}}{\sigma_k^2}\right)^{-1} \sum_{k\in s_N} \frac{d_{kN}\mathbf{x}_{kN} y_{kN}}{\sigma_k^2}$.

Nejprve označme $R_N = \left(\hat{\boldsymbol{\beta}}_{s_N} - \boldsymbol{\beta}_N\right)^{\top}\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right)$, pro jednotky z populace U_N označme $z_{kN} = y_{kN} - \boldsymbol{\beta}_N^{\top}\mathbf{x}_{kN}$, $k \in U_N$ a dále označme $\hat{Z}_{HT,N}$ Horvitzův-Thompsonův odhad populačního úhrnu $Z_N = \sum_{k\in U_N} z_{kN}$, tj.

$$\hat{Z}_{HT,N} = \sum_{k\in s_N} d_{kN} z_{kN} = \sum_{k\in s_N} d_{kN} y_{kN} - \sum_{k\in s_N} d_{kN}\boldsymbol{\beta}_N^{\top}\mathbf{x}_{kN} = \hat{Y}_{HT,N} - \boldsymbol{\beta}_N^{\top}\hat{\mathbf{X}}_{HT,N}.$$

Díky vlastnostem Horvitzova-Thompsonova odhadu platí (viz (1.2) a (1.3)):

$$\begin{aligned}\mathsf{E}\;\hat{Z}_{HT,N} &= Y_N - \boldsymbol{\beta}_N^{\top}\mathbf{X}_N = Z_N, \\ \operatorname{var}\hat{Z}_{HT,N} &= \sum_{k\in U_N}\sum_{l\in U_N}\left(\frac{\pi_{klN}}{\pi_{kN}\pi_{lN}} - 1\right) z_{kN} z_{lN}.\end{aligned}$$

Za pomoci výše uvedených vztahů a značení lze rovnice (3.11) upravit do tvaru:

$$\hat{Y}_{reg,N} = \hat{Z}_{HT,N} + \boldsymbol{\beta}_N^{\top}\mathbf{X}_N + R_N = \hat{Z}_{HT,N} + Y_N - \mathsf{E}\;\hat{Z}_{HT,N} + R_N.$$

Odečteme-li od obou stran rovnice populační úhrn Y_N a podělíme je odmocninou z rozptylu Horvitzova-Thompsonova odhadu $\hat{Z}_{HT,N}$, získáme vztah:

$$\frac{\hat{Y}_{reg,N} - Y_N}{\sqrt{\operatorname{var}\hat{Z}_{HT,N}}} = \frac{\hat{Z}_{HT,N} - \mathsf{E}\;\hat{Z}_{HT,N}}{\sqrt{\operatorname{var}\hat{Z}_{HT,N}}} + \frac{R_N}{\sqrt{\operatorname{var}\hat{Z}_{HT,N}}}. \quad (3.12)$$

Podíváme-li se nyní blíže na rovnost (3.12), vidíme, že pokud bychom mohli člen $\frac{R_N}{\sqrt{\operatorname{var} \hat{Z}_{HT,\,N}}}$ zanedbat, tak za platnosti vztahu $\frac{\hat{Z}_{HT,\,N} - Z_N}{\sqrt{\operatorname{var} \hat{Z}_{HT,\,N}}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1)$ bychom již získali požadovanou asymptotickou normalitu odhadu $\hat{Y}_{reg,\,N}$. Abychom však mohli člen $\frac{R_N}{\sqrt{\operatorname{var} \hat{Z}_{HT,\,N}}}$ zanedbat a dosáhli požadované normality regresního odhadu $\hat{Y}_{reg,\,N}$, je třeba dále předpokládat (I) – (IV):

(I) $\frac{\hat{Z}_{HT,\,N} - Z_N}{\sqrt{\operatorname{var} \hat{Z}_{HT,\,N}}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1)$,

(II) $\hat{\boldsymbol{\beta}}_{s_N} - \boldsymbol{\beta}_N \xrightarrow[N\to\infty]{\mathcal{P}} \mathbf{0}$,

(III) $\frac{\hat{X}^{(j)}_{HT,\,N} - X^{(j)}_N}{\sqrt{\operatorname{var} \hat{X}^{(j)}_{HT,\,N}}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1)$, pro $j = 1, \ldots, J$,
kde $\hat{X}^{(j)}_{HT,\,N}$ je $j-$tá složka vektoru $\hat{\mathbf{X}}_{HT,\,N}$, tedy $\hat{X}^{(j)}_{HT,\,N} = \sum_{k\in s_N} d_{kN} x_{kjN}$ a $X^{(j)}_N$ je $j-$tá složka vektoru $\mathbf{X}_N$, tedy $X^{(j)}_N = \sum_{k\in U_N} x_{kjN}$,

(IV) $\liminf_{N\to\infty} \frac{\operatorname{var} \hat{Z}_{HT,\,N}}{\max_{j=1,\,\ldots,\,J} \operatorname{var} \hat{X}^{(j)}_{HT,\,N}} > 0$.

Za těchto předpokladů můžeme člen $\frac{R_N}{\sqrt{\operatorname{var} \hat{Z}_{HT,\,N}}}$ z rovnice (3.12) odhadnout následovně:

$$\begin{aligned}
\frac{|R_N|}{\sqrt{\operatorname{var} \hat{Z}_{HT,\,N}}} &= \frac{\left|\left(\hat{\boldsymbol{\beta}}_{s_N} - \boldsymbol{\beta}_N\right)^\top \left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,\,N}\right)\right|}{\sqrt{\operatorname{var} \hat{Z}_{HT,\,N}}} \\
&\leq \left|\left(\hat{\boldsymbol{\beta}}_{s_N} - \boldsymbol{\beta}_N\right)^\top\right| \sum_{j=1}^{J} \frac{\left|\hat{X}^{(j)}_{HT,\,N} - X^j_N\right|}{\sqrt{\operatorname{var} \hat{X}^{(j)}_{HT,\,N}}} \sqrt{\frac{\operatorname{var} \hat{X}^{(j)}_{HT,\,N}}{\operatorname{var} \hat{Z}_{HT,\,N}}} \xrightarrow[N\to\infty]{\mathcal{P}} 0,
\end{aligned}$$

neboť $\left(\hat{\boldsymbol{\beta}}_{s_N} - \boldsymbol{\beta}_N\right)^\top \overset{\text{(II)}}{=} o_{\mathcal{P}}(1)$, $\frac{\left|\hat{X}^{(j)}_{HT,\,N} - X^{(j)}_N\right|}{\sqrt{\operatorname{var} \hat{X}^{(j)}_{HT,\,N}}} \overset{\text{(III)}}{=} O_{\mathcal{P}}(1)$ a $\sqrt{\frac{\operatorname{var} \hat{X}^{(j)}_{HT,\,N}}{\operatorname{var} \hat{Z}_{HT,\,N}}} \overset{\text{(I)},\,\text{(IV)}}{=} O(1)$, kde symboly o, O, $o_{\mathcal{P}}$ a $O_{\mathcal{P}}$, které jsou používané i v dalších částech práce, jsou zavedeny například v knize Jiang (2010, kapitola 3).

Celkem tedy ze vztahu (3.12), výše uvedených předpokladů a Cramérovy-Slutského věty dostáváme:

$$\frac{\hat{Y}_{reg,\,N} - Y_N}{\sqrt{\operatorname{var} \hat{Z}_{HT,\,N}}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1). \tag{3.13}$$

Z tohoto výsledku zároveň dostáváme asymptotický rozptyl obecného regresního odhadu $\hat{Y}_{reg,\,N}$ (3.8), pro který platí:

$$\operatorname{avar}\left(\hat{Y}_{reg,\,N}\right) = \operatorname{var} \hat{Z}_{HT,\,N} = \sum_{k\in U_N} \sum_{l\in U_N} \left(\frac{\pi_{klN}}{\pi_{kN}\pi_{lN}} - 1\right) z_{kN} z_{lN}, \tag{3.14}$$

kde $z_{kN} = y_{kN} - \boldsymbol{\beta}_N^\top \mathbf{x}_{kN}$, $k \in U_N$.

Nyní se zaměříme na odhad tohoto asymptotického rozptylu. Z předchozího víme, že avar $\left(\hat{Y}_{reg,\,N}\right) = \operatorname{var}\hat{Z}_{HT,\,N}$, tedy můžeme aplikovat teorii Horvitzova-Thompsonova odhadu. Pokud aplikujeme pouze teorii Horvitzova-Thomposonova odhadu, tak díky vzorci (1.4) dostáváme vztah:

$$\sum_{k\in s_N}\sum_{l\in s_N}\left(\frac{1}{\pi_{kN}\pi_{lN}}-\frac{1}{\pi_{klN}}\right)z_{kN}z_{lN},$$

kde $z_{kN} = y_{kN} - \boldsymbol{\beta}_N^\top \mathbf{x}_{kN}$, $k \in U_N$. Je zřejmé, že pro výběry s pevným rozsahem výběru $K(s_N)$ lze použít odhad rozptylu Horvitzova-Thompsonova odhadu, který je založený na Yatesově-Grundyho formuli (viz vzorec (1.6)):

$$-\frac{1}{2}\sum_{\substack{k\in s_N\\ k\neq l}}\sum_{l\in s_N}\left(1-\frac{\pi_{kN}\pi_{lN}}{\pi_{klN}}\right)\left(\frac{z_{kN}}{\pi_{kN}}-\frac{z_{lN}}{\pi_{lN}}\right)^2.$$

V praxi ovšem tyto vztahy nemůžeme použít, neboť $z_{kN} = y_{kN} - \boldsymbol{\beta}_N^\top \mathbf{x}_{kN}$ závisí na hypotetickém populačním odhadu $\boldsymbol{\beta}_N$, který je neznámý, tedy se nejedná o odhad ve smyslu, že odhad závisí pouze na známých datech. Z toho důvodu se pro odhady asymptotického rozptylu nahrazuje hypotetický populační odhad $\boldsymbol{\beta}_N$ za odhad $\hat{\boldsymbol{\beta}}_{s_N}$, který je založený na konkrétním výběru s_N. Dostáváme tedy

$$\widehat{\operatorname{avar}\left(\hat{Y}_{reg,\,N}\right)} = \sum_{k\in s_N}\sum_{l\in s_N}\left(\frac{1}{\pi_{kN}\pi_{lN}}-\frac{1}{\pi_{klN}}\right)\hat{z}_{kN}\hat{z}_{lN}, \tag{3.15}$$

případně pro výběry s pevným rozsahem výběru $K(s_N)$ dostáváme také druhý tvar pro odhad asymptotického rozptylu $\hat{Y}_{reg,\,N}$:

$$\widehat{\operatorname{avar}(\hat{Y}_{reg,\,N})} = -\frac{1}{2}\sum_{\substack{k\in s_N\\ k\neq l}}\sum_{l\in s_N}\left(1-\frac{\pi_{kN}\pi_{lN}}{\pi_{klN}}\right)\left(\frac{\hat{z}_{kN}}{\pi_{kN}}-\frac{\hat{z}_{lN}}{\pi_{lN}}\right)^2, \tag{3.16}$$

kde $\hat{z}_{kN} = y_{kN} - \hat{\boldsymbol{\beta}}_{s_N}^\top \mathbf{x}_{kN}$, $k \in s_N$.

Poznámka 5. Někdy se doporučuje v odhadu asymptotického rozptylu (3.15), respektive (3.16) pro výběry s pevným rozsahem výběru $K(s_N)$, nahradit $\hat{z}_{kN}$ za $\tilde{z}_{kN} = g_{kN}\hat{z}_{kN}$, kde

$$g_{kN} = 1 + \frac{\mathbf{x}_{kN}^\top}{\sigma_k^2}\left(\sum_{k\in s_N}\frac{d_{kN}\mathbf{x}_{kN}\mathbf{x}_{kN}^\top}{\sigma_k^2}\right)^{-1}\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,\,N}\right) = \frac{w_{kN}}{d_{kN}}, k \in s_N,$$

kde w_{kN} jsou určeny rovnicí (3.10) (viz Särdnal a kol., 1992, strana 235). Toto nahrazení je motivováno následující úvahou. Obecný regresní odhad (3.8) můžeme za pomoci vyjádření $y_{kN} = z_{kN} + \boldsymbol{\beta}_N^\top \mathbf{x}_{kN}$ zapsat ve tvaru

$$\hat{Y}_{reg,\,N} = \sum_{k\in s_N} d_{kN}g_{kN}y_{kN} = \sum_{k\in s_N} d_{kN}g_{kN}z_{kN} + \sum_{k\in s_N} d_{kN}g_{kN}\boldsymbol{\beta}_N^\top \mathbf{x}_{kN}.$$

Druhou sumu na pravé straně rovnice lze upravit následovně:

$$\begin{aligned}\sum_{k\in s_N} d_{kN} g_{kN} \boldsymbol{\beta}_N^\top \mathbf{x}_{kN} &= \boldsymbol{\beta}_N^\top \sum_{k\in s_N} d_{kN}\mathbf{x}_{kN} g_{kN} \\ &= \boldsymbol{\beta}_N^\top \sum_{k\in s_N} d_{kN}\mathbf{x}_{kN} \left(1 + \frac{\mathbf{x}_{kN}^\top}{\sigma_k^2}\left(\sum_{k\in s_N}\frac{d_{kN}\mathbf{x}_{kN}\mathbf{x}_{kN}^\top}{\sigma_k^2}\right)^{-1}\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right)\right) \\ &= \boldsymbol{\beta}_N^\top \left(\sum_{k\in s_N} d_{kN}\mathbf{x}_{kN} + \mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right) = \boldsymbol{\beta}_N^\top \left(\hat{\mathbf{X}}_{HT,N} + \mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right) \\ &= \boldsymbol{\beta}_N^\top \mathbf{X}_N.\end{aligned}$$

Celkem tedy máme

$$\hat{Y}_{reg,N} = \sum_{k\in s_N} d_{kN} g_{kN} z_{kN} + \boldsymbol{\beta}_N^\top \mathbf{X}_N. \tag{3.17}$$

Díky tomu, že lze obecný regresní odhad vyjádřit ve tvaru (3.17), můžeme rozptyl tohoto odhadu vyjádřit ve tvaru

$$\operatorname{var}\left(\hat{Y}_{reg,N}\right) = \operatorname{var}\left(\sum_{k\in s_N} d_{kN} g_{kN} z_{kN}\right).$$

Můžeme si všimnout, že na pravé straně rovnice počítáme rozptyl z výrazu, který připomíná Horvitzův-Thompsonův odhad. Aplikujeme-li tedy teorii Horvitzova-Thompsonova odhadu, nahradíme-li hypotetická z_{kN} za výběrová $\hat{z}_{kN}$ a navíc váhy g_{kN} uvažujeme jako nenáhodné (i když nejsou, neboť váhy g_{kN} závisí na konkrétním výběru s), dostaneme odhad asymptotického rozptylu tvaru

$$\widehat{\operatorname{avar}\left(\hat{Y}_{reg,N}\right)} = \sum_{k\in s_N}\sum_{l\in s_N}\left(\frac{1}{\pi_{kN}\pi_{lN}} - \frac{1}{\pi_{klN}}\right) g_{kN} g_{lN} \hat{z}_{kN}\hat{z}_{lN}, \tag{3.18}$$

případně pro výběry s pevným rozsahem výběru $K(s_N)$ lze využít vzorec

$$\widehat{\operatorname{avar}(\hat{Y}_{reg,N})} = -\frac{1}{2}\sum_{\substack{k\in s_N \\ k\neq l}}\sum_{l\in s_N}\left(1 - \frac{\pi_{kN}\pi_{lN}}{\pi_{klN}}\right)\left(\frac{g_{kN}\hat{z}_{kN}}{\pi_{kN}} - \frac{g_{lN}\hat{z}_{lN}}{\pi_{lN}}\right)^2, \tag{3.19}$$

kde $\hat{z}_{kN} = y_{kN} - \hat{\boldsymbol{\beta}}_{s_N}^\top \mathbf{x}_{kN}$, $k \in s_N$.

△

3.1.2 Aplikace pro prostý náhodný výběr

Obecný regresní odhad

Pro prostý náhodný výběr jsme odvodili, že $\pi_k = \frac{n}{N}$ (viz rovnost (2.1)), tedy $d_k = \frac{1}{\pi_k} = \frac{N}{n}$. Jelikož v praxi zůstávají hodnoty σ_k^2 neznámé, budeme pro zjednodušení dále předpokládat, že σ_k^2 nezávisí na k, tedy $\sigma_k^2 = \sigma^2$, $k \in U$. Jak si můžeme všimnout, ve všech výrazech, ve kterých se σ^2 vyskytuje ((3.4), (3.6), (3.10)), se tyto hodnoty pokrátí, tedy lze předpokládat, že $\sigma_k^2 = 1\,\forall k \in U$. Dosadíme-li tyto

vztahy do normálních rovnic založených na konkrétním výběru s, pomocí kterých hledáme odhad $\hat{\boldsymbol{\beta}}_s$ (vztah (3.5)), získáme

$$\frac{N}{n}\left(\sum_{k\in s}\mathbf{x}_k\mathbf{x}_k^\top\right)\hat{\boldsymbol{\beta}}_s = \frac{N}{n}\sum_{k\in s}\mathbf{x}_k y_k,$$
$$\left(\sum_{k\in s}\mathbf{x}_k\mathbf{x}_k^\top\right)\hat{\boldsymbol{\beta}}_s = \sum_{k\in s}\mathbf{x}_k y_k.$$

Tedy pokud lze matici $\left(\sum_{k\in s}\mathbf{x}_k\mathbf{x}_k^\top\right)$ invertovat, můžeme pro prostý náhodný výběr odhad $\hat{\boldsymbol{\beta}}_s$ zapsat ve tvaru

$$\hat{\boldsymbol{\beta}}_s = \left(\sum_{k\in s}\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{k\in s}\mathbf{x}_k y_k.$$

Z normálních rovnic pro hledání hypotetického populačního odhadu $\boldsymbol{\beta}_N$ (viz vztah (3.3)) lze nahlédnout, že v případě prostého náhodného výběru pro $\sigma_k^2 = 1\,\forall k \in U$ platí

$$\boldsymbol{\beta}_N = \left(\sum_{k\in U}\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{k\in U}\mathbf{x}_k y_k.$$

Tedy odhady $\hat{\boldsymbol{\beta}}_s$ a $\boldsymbol{\beta}_N$ mají stejný tvar, jen v případě hypotetického populačního odhadu $\boldsymbol{\beta}_N$ se sčítá přes jednotky z celé populace U a v případě odhadu $\hat{\boldsymbol{\beta}}_s$ založeném na konkrétním výběru s se sčítá přes jednotky z výběru s.

Zapíšeme-li obecný regresní odhad $\hat{Y}_{reg}$ ve tvaru $\hat{Y}_{reg} = \sum_{k\in s} w_k y_k$, tak v případě prostého náhodného výběru mají váhy w_k na základě vztahu (3.10) tvar

$$\begin{aligned} w_k = d_k g_k &= \frac{N}{n}\left(1 + \mathbf{x}_k^\top\left(\frac{N}{n}\sum_{k\in s}\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right)\right) \\ &= \frac{N}{n} + \mathbf{x}_k^\top\left(\sum_{k\in s}\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right), k \in s. \end{aligned} \tag{3.20}$$

Tedy obecný regresní odhad $\hat{Y}_{reg}$ lze v případě prostého náhodného výběru zapsat ve tvaru

$$\begin{aligned} \hat{Y}_{reg} = \sum_{k\in s} w_k y_k &\overset{(3.20)}{=} \frac{N}{n}\sum_{k\in s} y_k + \left(\sum_{k\in s}\mathbf{x}_k^\top y_k\right)\left(\sum_{k\in s}\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right) \\ &= N\bar{y} + \hat{\boldsymbol{\beta}}_s^\top\left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right) = N\bar{y} + \hat{\boldsymbol{\beta}}_s^\top N\left(\bar{\mathbf{X}} - \bar{\mathbf{x}}\right) = N\left(\bar{y} + \hat{\boldsymbol{\beta}}_s^\top\left(\bar{\mathbf{X}} - \bar{\mathbf{x}}\right)\right), \end{aligned} \tag{3.21}$$

kde $\bar{\mathbf{X}} = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k$ a $\bar{\mathbf{x}} = \frac{1}{n}\sum_{k\in s}\mathbf{x}_k$.

Asymptotická normalita obecného regresního odhadu

Nyní se zaměříme na asymptotickou normalitu a především asymptotický rozptyl tohoto odhadu v případě prostého náhodného výběru a za předpokladu, že

$\sigma_k^2 = 1$, $k \in U$. K tomu jsme v sekci 3.1.1 zavedli předpoklady (I) – (IV), kterými se nyní budeme zabývat. Tedy uvažujeme posloupnost konečných populací $U_1, \ldots, U_N, \ldots$, kde U_N obsahuje N jednotek. Jestliže sledovaný znak označíme y, potom $U_N = \{y_{1N}, \ldots, y_{NN}\}$. Z populace U_N děláme prostý náhodný výběr s_N s rozsahem n_N, pro který platí, že $\lim_{N\to\infty} n_N = \infty$ a $\limsup_{N\to\infty} \frac{n_N}{N} < 1$. Splnění předpokladů (I) a (III) můžeme zaručit splněním obdobné podmínky jako na pravé straně ekvivalence (2.8), tentokrát ovšem pro sledované znaky $\{z_{kN}, k \in U_N\}$, kde kde $z_{kN} = y_{kN} - \boldsymbol{\beta}_N^\top \mathbf{x}_{kN}$, respektive $\{x_{kjN}, k \in U_N\}$ $\forall\, j = 1, \ldots, J$, kde x_{kjN} je $j-$tá složka vektoru pomocných hodnot $\mathbf{x}_{kN}$ z populace U_N. Pro splnění těchto podmínek můžeme obdobně jako v části 2.1 předpokládat, že

- $\lim_{N\to\infty} \frac{1}{N-1} \sum_{k\in U_N} \left(z_{kN} - \bar{Z}_N\right)^2 \in (0, \infty)$, kde $\bar{Z}_N = \frac{1}{N} \sum_{k\in U_N} z_{kN}$,
- existuje $\delta > 0$ takové, že $\limsup_{N\to\infty} \frac{1}{N-1} \sum_{k\in U_N} \left|z_{kN} - \bar{Z}_N\right|^{2+\delta} < \infty$,

respektive

- $\lim_{N\to\infty} \frac{1}{N-1} \sum_{k\in U_N} \left(x_{kjN} - \bar{X}_N^{(j)}\right)^2 \in (0, \infty)$, kde $\bar{X}_N^{(j)} = \frac{1}{N} \sum_{k\in U_N} x_{kjN}$,
- existuje $\delta > 0$ takové, že $\limsup_{N\to\infty} \frac{1}{N-1} \sum_{k\in U_N} \left|x_{kjN} - \bar{X}_N^{(j)}\right|^{2+\delta} < \infty$.

Z výše uvedeného si můžeme všimnout, že pro splnění předpokladů (I) a (III) je třeba předpokládat existence $(2+\delta)-$tého absolutního momentu u daného sledovaného znaku z populace U_N, kde $\delta > 0$.

Nyní se budeme zabývat předpokladem (II). Můžeme vidět, že odhady $\hat{\boldsymbol{\beta}}_{s_N}$ a $\boldsymbol{\beta}_N$ lze zapsat následovně

$$\hat{\boldsymbol{\beta}}_{s_N} = \left(\frac{1}{n_N} \sum_{k\in s_N} \mathbf{x}_{kN}\mathbf{x}_{kN}^\top\right)^{-1} \left(\frac{1}{n_N} \sum_{k\in s_N} \mathbf{x}_{kN} y_{kN}\right),$$
$$\boldsymbol{\beta}_N = \left(\frac{1}{N} \sum_{k\in U_N} \mathbf{x}_{kN}\mathbf{x}_{kN}^\top\right)^{-1} \left(\frac{1}{N} \sum_{k\in U_N} \mathbf{x}_{kN} y_{kN}\right).$$

Předpokládáme-li, že

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k\in U_N} x_{kiN}^4 < \infty \quad \forall i = 1, \ldots, J, \tag{3.22}$$

pak již můžeme aplikovat lemma 1 s $v_{kN} = x_{kiN} x_{kjN}^\top$, kde $i, j = 1, \ldots, J$, na každou složku matice $\mathbf{x}_{kN}\mathbf{x}_{kN}^\top$, neboť díky (3.22) a Cauchyho-Schwarzově nerovnosti máme splněny všechny předpoklady tohoto lemmatu a tedy dostaneme, že

$$\frac{1}{n_N} \sum_{k\in s_N} \mathbf{x}_{kN}\mathbf{x}_{kN}^\top - \frac{1}{N} \sum_{k\in U_N} \mathbf{x}_{kN}\mathbf{x}_{kN}^\top \xrightarrow[N\to\infty]{\mathcal{P}} 0.$$

Předpokládáme-li, že

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k\in U_N} (x_{kjN} y_{kN})^2 < \infty \quad \forall j = 1, \ldots, J,$$

můžeme lemma 1 aplikovat na každou složku vektoru $\mathbf{x}_{kN}y_{kN}$ a dostáváme, že

$$\frac{1}{n_N}\sum_{k\in s_N}\mathbf{x}_{kN}y_{kN}-\frac{1}{N}\sum_{k\in U_N}\mathbf{x}_{kN}y_{kN}\xrightarrow[N\to\infty]{\mathcal{P}}0.$$

Celkem tedy máme

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{s_N}-\boldsymbol{\beta}_N&=\left(\frac{1}{n_N}\sum_{k\in s_N}\mathbf{x}_{kN}\mathbf{x}_{kN}^\top\right)^{-1}\left(\frac{1}{n_N}\sum_{k\in s_N}\mathbf{x}_{kN}y_{kN}\right)\\&\quad-\left(\frac{1}{N}\sum_{k\in U_N}\mathbf{x}_{kN}\mathbf{x}_{kN}^\top\right)^{-1}\left(\frac{1}{N}\sum_{k\in U_N}\mathbf{x}_{kN}y_{kN}\right)\\&=\left[\left(\frac{1}{n_N}\sum_{k\in s_N}\mathbf{x}_{kN}\mathbf{x}_{kN}^\top\right)^{-1}-\left(\frac{1}{N}\sum_{k\in U_N}\mathbf{x}_{kN}\mathbf{x}_{kN}^\top\right)^{-1}\right]\left(\frac{1}{n_N}\sum_{k\in s_N}\mathbf{x}_{kN}y_{kN}\right)\\&\quad+\left(\frac{1}{N}\sum_{k\in U_N}\mathbf{x}_{kN}\mathbf{x}_{kN}^\top\right)^{-1}\left[\left(\frac{1}{n_N}\sum_{k\in s_N}\mathbf{x}_{kN}y_{kN}\right)-\left(\frac{1}{N}\sum_{k\in U_N}\mathbf{x}_{kN}y_{kN}\right)\right].\end{aligned}\tag{3.23}$$

Předpokládáme-li navíc, že matice $\left(\frac{1}{N}\sum_{k\in U_N}\mathbf{x}_{kN}\mathbf{x}_{kN}^\top\right)$ konverguje pro $N\to\infty$ k regulární matici, tak díky výše uvedeným předpokladům a větě o spojité transformaci, kterou aplikujeme obdobně jako je uvedeno v poznámce 4 na první část vztahu (3.23), získáváme platnost předpokladu (II), tedy $\hat{\boldsymbol{\beta}}_{s_N}-\boldsymbol{\beta}_N\xrightarrow[N\to\infty]{\mathcal{P}}\mathbf{0}$.

Nyní se zaměříme na poslední předpoklad (IV). Tento předpoklad je technického charakteru a říká nám, že maximální hodnota rozptylu přes jednotlivé složky Horvitzova-Thompsonova odhadu $\hat{\mathbf{X}}_{HT,N}$ populačního průměru $\mathbf{X}_N$ není mnohonásobně větší než hodnota rozptylu Horvitzova-Thompsonova odhadu $\hat{Z}_{HT,N}$ populačního průměru Z_N. V případě prostého náhodného výběru platí

$$\begin{aligned}\frac{\operatorname{var}\hat{Z}_{HT,N}}{\max_{j=1,\dots,J}\operatorname{var}\hat{X}^{(j)}_{HT,N}}&=\frac{-\frac{1}{2}\left(1-\frac{N}{n_N}\right)\frac{1}{N-1}\sum_{k\in U_N}\sum_{\substack{l\in U_N\\k\neq l}}\left(z_{kN}-z_{lN}\right)^2}{-\frac{1}{2}\left(1-\frac{N}{n_N}\right)\frac{1}{N-1}\sum_{k\in U_N}\sum_{\substack{l\in U_N\\k\neq l}}\left(x_{kjN}-x_{ljN}\right)^2}\\&=\frac{\frac{1}{N-1}\sum_{k\in U_N}\left(z_{kN}-\bar{Z}_N\right)^2}{\frac{1}{N-1}\sum_{k\in U_N}\left(x_{kjN}-\bar{X}^{(j)}_N\right)^2}=\frac{\sigma^2_{zN}}{\sigma^2_{x^{(j)}N}},\end{aligned}$$

kde $\bar{Z}_N=\frac{1}{N}\sum_{k\in U_N}z_{kN}$ a $\bar{X}^{(j)}_N=\frac{1}{N}\sum_{k\in U_N}x_{kjN}$. Celkem tedy lze předpoklad (IV) zapsat následovně:

$$\liminf_{N\to\infty}\frac{\operatorname{var}\hat{Z}_{HT,N}}{\max_{j=1,\dots,J}\operatorname{var}\hat{X}^{(j)}_{HT,N}}=\liminf_{N\to\infty}\frac{\sigma^2_{zN}}{\sigma^2_{x^{(j)}N}},$$

kde $\sigma^2_{zN}=\frac{1}{N-1}\sum_{k\in U_N}\left(z_{kN}-\bar{Z}_N\right)^2$ a $\sigma^2_{x^{(j)}N}=\frac{1}{N-1}\sum_{k\in U_N}\left(x_{kjN}-\bar{X}^{(j)}_N\right)^2$. Díky předpokladu (3.22) dostáváme, že $\sigma^2_{x^{(j)}N}<\infty$, tedy jediný případ, kdy by mohl být porušen předpoklad (IV) je, pokud $\sigma^2_{zN}=0$. To by ovšem mohlo nastat pouze v případě, že pro každou jednotku k z populace U_N by platilo, že $z_{kN}=\bar{Z}_N$,

$k \in U_N$. Jinými slovy by muselo platit, že všechny jednotky v populaci U_N mají shodné z_{kN}, což není možné, tedy jsme ověřili, že předpoklad (IV) je splněn.

Výše jsme specifikovali předpoklady (I) – (IV) pro případ prostého náhodného výběru a volbu $\sigma_k^2 = 1$, $k \in U$. Tedy ze vztahu (3.13) víme, že

$$\frac{\hat{Y}_{reg,\,N} - Y_N}{\sqrt{\operatorname{var} \hat{Z}_{HT,\,N}}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1)\,.$$

Nyní bychom v tomto vztahu chtěli nahradit rozptyl odhadu $\hat{Z}_{HT,\,N}$ za jeho odhad. Díky tomu, že prostý náhodný výběr je výběr s pevným rozsahem výběru $K(s_N)$, můžeme pro odhad rozptylu využít rovnici (3.16) a s využitím rovností (2.1) získáme odhad asymptotického rozptylu pro obecný regresní odhad $\hat{Y}_{reg}$ (3.8), který v případě prostého náhodného výběru a pro $\sigma_k^2 = 1$, $k \in U$, má tvar:

$$\widehat{\operatorname{var}\, \hat{Z}_{HT,\,N}} = \widehat{\operatorname{avar}\left(\hat{Y}_{reg,\,N}\right)} = -\frac{1}{2} \sum_{\substack{k \in s_N \\ k \neq l}} \sum_{l \in s_N} \left(1 - \frac{n_N(N-1)}{N(n_N - 1)}\right) \frac{N^2}{n_N^2} \left(\hat{z}_{kN} - \hat{z}_{lN}\right)^2,$$

což lze obdobnými úpravami jako při úpravě výrazů před částí 2.1 upravit do tvaru

$$\begin{aligned}\widehat{\operatorname{avar}\left(\hat{Y}_{reg,\,N}\right)} &= N^2 \left(1 - \frac{n_N}{N}\right) \frac{1}{n_N} \frac{1}{n_N - 1} \sum_{k \in s_N} \left(\hat{z}_{kN} - \bar{\hat{z}}_N\right)^2 \\ &= N^2 \left(1 - \frac{n_N}{N}\right) \frac{1}{n_N} s_{\hat{z}N}^2,\end{aligned}$$

kde $\hat{z}_{kN} = y_{kN} - \hat{\boldsymbol{\beta}}_{s_N}^\top \mathbf{x}_{kN} = y_{kN} - \left(\sum_{k \in s_N} \mathbf{x}_{kN}^\top y_{kN}\right) \left(\sum_{k \in s_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top\right)^{-1} \mathbf{x}_{kN}$, $k \in s_N$, $\bar{\hat{z}}_N = \frac{1}{n_N} \sum_{k \in s_N} \hat{z}_{kN}$ a $s_{\hat{z}N}^2 = \frac{1}{n_N - 1} \sum_{k \in s_N} \left(\hat{z}_{kN} - \bar{\hat{z}}_N\right)^2$.

Abychom ovšem získali požadovaný vztah

$$\frac{\hat{Y}_{reg,\,N} - Y_N}{\sqrt{\widehat{\operatorname{var}\, \hat{Z}_{HT,\,N}}}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1)\,,$$

je třeba využít faktu, že pro $\sqrt{\widehat{\operatorname{var}\, \hat{Z}_{HT,\,N}}}$ platí

$$\frac{\sqrt{\widehat{\operatorname{var}\, \hat{Z}_{HT,\,N}}}}{\sqrt{\operatorname{var}\, \hat{Z}_{HT,\,N}}} \xrightarrow[N\to\infty]{\mathcal{P}} 1.$$

(viz Kott, 1990). Díky výše uvedenému faktu a Cramérově-Slutského větě tedy dostáváme, že

$$\frac{\hat{Y}_{reg,\,N} - Y_N}{\sqrt{\widehat{\operatorname{var}\, \hat{Z}_{HT,\,N}}}} = \frac{\frac{N}{n_N} \sum_{k \in s_N} g_{kN} y_{kN} - \sum_{k \in U_N} y_{kN}}{\sqrt{N^2 \left(1 - \frac{n_N}{N}\right) \frac{1}{n_N} s_{\hat{z}N}^2}} \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1)\,, \tag{3.24}$$

kde $s_{\hat{z}N}^2 = \frac{1}{n_N - 1} \sum_{k \in s_N} \left(\hat{z}_{kN} - \bar{\hat{z}}_N\right)^2$, $\bar{\hat{z}}_N = \frac{1}{n_N} \sum_{k \in s_N} \hat{z}_{kN}$, $\hat{z}_{kN} = y_{kN} - \hat{\boldsymbol{\beta}}_{sN}^\top \mathbf{x}_{kN} = y_{kN} - \left(\sum_{k \in s_N} \mathbf{x}_{kN}^\top y_{kN}\right) \left(\sum_{k \in s_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top\right)^{-1} \mathbf{x}_{kN}$, $k \in s_N$
a $g_{kN} = 1 + \mathbf{x}_{kN}^\top \left(\frac{N}{n_N} \sum_{k \in s_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top\right)^{-1} \left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,\,N}\right)$, $k \in s_N$.

Zaměříme-li se nyní podrobněji na váhy g_{kN}, konkrétně na limitní chování těchto vah pro $N \to \infty$ (z hlediska designového přístupu), tak ze vztahu (3.20) máme, že v případě prostého náhodného výběru pro $\sigma_k^2 = 1$, $k \in U$, platí

$$g_{kN} = 1 + \mathbf{x}_{kN}^\top \left(\frac{N}{n_N} \sum_{k \in s_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1} \left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N} \right), k \in s_N. \tag{3.25}$$

Nejprve se zaměříme na část $\left(\frac{N}{n_N} \sum_{k \in s_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1}$, pro kterou platí

$$\left(\frac{N}{n_N} \sum_{k \in s_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1} = \frac{1}{N} \left[\left(\frac{1}{n_N} \sum_{k \in s_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1} - \left(\frac{1}{N} \sum_{k \in U_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1} \right] \tag{3.26}$$

$$+ \frac{1}{N} \left(\frac{1}{N} \sum_{k \in U_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1}. \tag{3.27}$$

Dále víme, že platí-li (3.22) a předpoklad, že matice $\left(\frac{1}{N} \sum_{k \in U_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)$ konverguje pro $N \to \infty$ k regulární matici, pak obdobně jako ve vztahu (3.23) platí pro část (3.26), že

$$\frac{1}{N} \left[\left(\frac{1}{n_N} \sum_{k \in s_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1} - \left(\frac{1}{N} \sum_{k \in U_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1} \right] = \frac{1}{N} o_{\mathcal{P}}(1) = o_{\mathcal{P}} \left(\frac{1}{N} \right).$$

Z předpokladu, že matice $\left(\frac{1}{N} \sum_{k \in U_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)$ konverguje pro $N \to \infty$ k regulární matici, dále pro část (3.27) dostáváme, že

$$\frac{1}{N} \left(\frac{1}{N} \sum_{k \in U_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1} = \frac{1}{N} O(1) = O \left(\frac{1}{N} \right).$$

Celkem tedy máme

$$\left(\frac{N}{n_N} \sum_{k \in s_N} \mathbf{x}_{kN} \mathbf{x}_{kN}^\top \right)^{-1} = o_{\mathcal{P}} \left(\frac{1}{N} \right) + O \left(\frac{1}{N} \right) = O_{\mathcal{P}} \left(\frac{1}{N} \right).$$

Nyní se blíže podíváme na část $\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N} \right)$ ze vztahu (3.25), platí:

$$\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N} \right) = \frac{\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}}{\sqrt{\operatorname{var} \left(\hat{\mathbf{X}}_{HT,N} \right)}} \sqrt{\operatorname{var} \left(\hat{\mathbf{X}}_{HT,N} \right)}.$$

Díky předpokladu (III) pro zlomek $\frac{\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}}{\sqrt{\operatorname{var} \left(\hat{\mathbf{X}}_{HT,N} \right)}}$ platí, že

$$\frac{\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}}{\sqrt{\operatorname{var} \left(\hat{\mathbf{X}}_{HT,N} \right)}} = O_{\mathcal{P}}(1),$$

a dále pro rozptyl Horvitzova-Thompsonova odhadu $\hat{\mathbf{X}}_{HT,N}$ populačního úhrnu $\mathbf{X}_N$ za předpokladu (3.22), platí, že

$$\sqrt{\mathrm{var}\left(\hat{\mathbf{X}}_{HT,N}\right)} = \sqrt{\frac{N^2}{n_N}} O_{\mathcal{P}}(1) = O_{\mathcal{P}}\left(\frac{N}{\sqrt{n_N}}\right).$$

Pro výraz $\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right)$ tedy platí, že

$$\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right) = O_{\mathcal{P}}(1) \cdot O_{\mathcal{P}}\left(\frac{N}{\sqrt{n_N}}\right) = O_{\mathcal{P}}\left(\frac{N}{\sqrt{n_N}}\right).$$

Pro váhy g_{kN} (3.25), $k \in s_N$, tedy celkem dostáváme, že

$$g_{kN} = 1 + O_{\mathcal{P}}\left(\frac{1}{N}\right) O_{\mathcal{P}}\left(\frac{N}{\sqrt{n_N}}\right) = 1 + O_{\mathcal{P}}\left(\frac{1}{\sqrt{n_N}}\right). \tag{3.28}$$

Poznámka 6. Vlastnost vah g_{kN} (3.28), kterou jsme ukázali v případě prostého náhodného výběru pro $\sigma_k^2 = 1$, $k \in s_N$, se standardně předpokládá i v ostatních případech, tedy nejen u prostého náhodného výběru bez vracení.

△

Díky vlastnosti (3.28) tedy platí, že i odhad rozptylu Horvitzova-Thompsonova odhadu $\hat{Z}_{HT,N}$ ve tvaru (3.18) je konzistentním odhadem rozptylu Horvitzova-Thompsonova odhadu $\hat{Z}_{HT,N}$ (viz Särndal a kol., 1989), tedy platí, že

$$\begin{aligned}\frac{\hat{Y}_{reg,N} - Y_N}{\sqrt{\widehat{\mathrm{var}}\,\hat{Z}_{HT,N}}} &= \frac{\frac{N}{n_N}\sum_{k\in s_N} g_{kN} y_{kN} - \sum_{k\in U_N} y_{kN}}{\sqrt{N^2\left(1-\frac{n_N}{N}\right)\frac{1}{n_N}\frac{1}{n_N-1}\sum_{k\in s_N}\left(g_{kN}\hat{z}_{kN} - \frac{1}{n_N}\sum_{k\in s_N} g_{kN}\hat{z}_{kN}\right)^2}} \\ &\xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{N}(0,1),\end{aligned} \tag{3.29}$$

kde $\hat{z}_{kN} = y_{kN} - \hat{\boldsymbol{\beta}}_{s_N}^{\top}\mathbf{x}_{kN} = y_{kN} - \left(\sum_{k\in s_N}\mathbf{x}_{kN}^{\top} y_{kN}\right)\left(\sum_{k\in s_N}\mathbf{x}_{kN}\mathbf{x}_{kN}^{\top}\right)^{-1}\mathbf{x}_{kN}$, $k \in s_N$ a $g_{kN} = 1 + \mathbf{x}_{kN}^{\top}\left(\frac{N}{n_N}\sum_{k\in s_N}\mathbf{x}_{kN}\mathbf{x}_{kN}^{\top}\right)^{-1}\left(\mathbf{X}_N - \hat{\mathbf{X}}_{HT,N}\right)$, $k \in s_N$.

3.2 Kalibrační odhady

Kalibrační odhady jsou odhady ve tvaru (3.1), které pro odhad populačního úhrnu Y také zohledňují pomocné informace obsažené ve vektorech $\mathbf{x}_k$, $k \in U$. Nové váhy w_k hledáme tak, aby byly vzhledem k námi zvolené funkci vzdálenosti co nejblíže původním vahám d_k a zároveň byly splněny tzv. kalibrační rovnice

$$\sum_{k\in s} w_k \mathbf{x}_k = \mathbf{X}, \tag{3.30}$$

(viz Deville a Särndal, 1992 a Deville a kol., 1993). Tento fakt je motivován tím, že pokud srovnáme vychýlení kalibračního odhadu (3.1) s Horvitzovým-Thompsonovým odhadem (1.1), o kterém víme, že je designově nestranný, tak dostaneme

$$\mathsf{E}\left(\sum_{k\in s} w_k y_k\right) - Y = \mathsf{E}\left(\sum_{k\in s} w_k y_k\right) - \mathsf{E}\,\hat{Y}_{HT} = \mathsf{E}\left(\sum_{k\in s}(w_k - d_k)\, y_k\right),$$

z čehož vyplývá, že pokud hledáme odhad, který je „téměř“ designově nestranný (tj. má malé designové vychýlení, $\mathsf{E}\left[\sum_{k\in s} w_k y_k\right] \approx Y$), tak nové váhy w_k musí být co nejblíže původním vahám d_k (viz Särndal, 2010, strana 105).

Poznámka 7. V případě prostého náhodného výběru platí, že $d_k = \frac{N}{n}$, $k \in s$. Tato volba původních vah d_k dává smysl, neboť všem jednotkám ve výběru s dáváme stejnou váhu, tedy žádnou z nich neupřednostňujeme před ostatními. Tedy při hledání nových vah w_k, $k \in s$, chceme, aby tyto váhy byly co nejblíže původním vahám $\frac{N}{n}$, tj. aby byly co „nejpodobnější“.

△

Váhy w_k budeme hledat minimalizací dané funkce vzdálenosti za podmínky, že platí kalibrační rovnice (3.30), což povede na úlohu Lagrangeových multiplikátorů.

Příklad. (viz Deville a Särndal, 1992, strana 378) Zaměříme se na kalibrační odhad $\hat{Y}_{kal} = \sum_{k\in s} w_k y_k$, kde w_k splňují kalibrační rovnice (3.30). Nechť existuje vektor konstant $\boldsymbol{\alpha}$ takový, že platí $y_k = \mathbf{x}_k^\top \boldsymbol{\alpha}$ pro každou jednotku k z populace U. Potom platí

$$\begin{aligned}\hat{Y}_{kal} &= \sum_{k\in s} w_k y_k = \sum_{k\in s} w_k \mathbf{x}_k^\top \boldsymbol{\alpha} = \left(\sum_{k\in s} w_k \mathbf{x}_k^\top\right) \boldsymbol{\alpha} \overset{(3.30)}{=} \mathbf{X}^\top \boldsymbol{\alpha} \\ &= \sum_{k\in U} \mathbf{x}_k^\top \boldsymbol{\alpha} = \sum_{k\in U} y_k = Y.\end{aligned}$$

Z rovnice výše vidíme, že pokud hodnoty sledovaných znaků y_k, $k \in U$, jsou pouze nějaké lineární kombinace pomocných hodnot $\mathbf{x}_k$, $k \in U$, tak potom kalibrační odhad $\hat{Y}_{kal}$ je díky kalibračním rovnicím (3.30) roven populačnímu úhrnu Y. Tedy získáme přesný odhad, což však v praxi nastane zřídkakdy, neboť ve většině případů neexistuje vektor konstant $\boldsymbol{\alpha}$ takový, že $y_k = \mathbf{x}_k^\top \boldsymbol{\alpha}$. Často platí, že rozdíly $y_k - \mathbf{x}_k^\top \boldsymbol{\alpha}$ jsou malé pro většinu jednotek z populace U.

3.2.1 Obecný regresní odhad jako kalibrační odhad

V této části si ukážeme, jak lze obecný regresní odhad (3.8) odvodit pomocí teorie kalibračních odhadů. Za tímto účelem nejprve ukážeme, jak zvolit funkci vzdálenosti tak, abychom dostali odhad shodný s obecným regresním odhadem $\hat{Y}_{reg}$ (3.8). Funkci vzdálenosti zvolíme tak, aby připomínala chí-kvadrát statistiku, tedy je tvaru $\sum_{k\in s} \frac{(w_k - d_k)^2}{d_k}$. Tuto funkci, kterou měříme vzdálenost nových vah w_k od původních vah d_k, musíme ovšem uvažovat v obecnější verzi, ve které každý člen součtu navíc podělíme hodnotou $2q_k > 0$, která nijak nesouvisí s d_k (pro zobecnění by stačilo podělit každý člen součtu hodnotou q_k, ale s hodnotu $2q_k$ se následně lépe pracuje). Toto zobecnění uvažujeme, abychom umožnili různá $\{\sigma_k^2, k \in U\}$ ve vzorci obecného regresního odhadu (3.8). Kdybychom ho neuvažovali, tak bychom nedostali shodný odhad s obecným regresním odhadem (3.8). Minimalizujeme tedy

$$\sum_{k\in s} \frac{(w_k - d_k)^2}{2 d_k q_k}$$

za podmínky, že platí kalibrační rovnice (3.30), což vede na úlohu Lagrangeových multiplikátorů. Lagrangeova funkce má tvar

$$L\left(\mathbf{w}, \boldsymbol{\lambda}\right) = \sum_{k \in s} \frac{\left(w_k - d_k\right)^2}{2 d_k q_k} - \boldsymbol{\lambda}^\top \left(\sum_{k \in s} w_k \mathbf{x}_k - \sum_{k \in U} \mathbf{x}_k \right), \tag{3.31}$$

kde jsme pro přehlednost označili $\mathbf{w} = \{w_k\}_{k \in s}$. Zderivováním Lagrangeovy funkce (3.31) podle proměnné w_k a následném položením rovno 0, získáme vztah:

$$\frac{w_k - d_k}{d_k q_k} - \boldsymbol{\lambda}^\top \mathbf{x}_k = 0, k \in s.$$

Upravením této rovnice získáme vztah pro nové váhy w_k, které jsou tvaru

$$w_k = d_k \left(1 + q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right), k \in s, \tag{3.32}$$

kde $\boldsymbol{\lambda}$ je vektor Lagrangeových multiplikátorů, který dopočítáme z kalibračních rovnic (3.30), tedy

$$\begin{aligned} \sum_{k \in s} \mathbf{x}_k d_k \left(1 + q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right) &= \mathbf{X}, \\ \sum_{k \in s} d_k \mathbf{x}_k + \left(\sum_{k \in s} d_k q_k \mathbf{x}_k \mathbf{x}_k^\top \right) \boldsymbol{\lambda} &= \mathbf{X}, \\ \left(\sum_{k \in s} d_k q_k \mathbf{x}_k \mathbf{x}_k^\top \right) \boldsymbol{\lambda} &= \mathbf{X} - \hat{\mathbf{X}}_{HT}. \end{aligned}$$

Pokud lze matici $\left(\sum_{k \in s} d_k q_k \mathbf{x}_k \mathbf{x}_k^\top\right)$ invertovat, tak pro $\boldsymbol{\lambda}$ platí:

$$\boldsymbol{\lambda} = \left(\sum_{k \in s} q_k d_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \left(\mathbf{X} - \hat{\mathbf{X}}_{HT} \right). \tag{3.33}$$

Po dosazení vypočítané hodnoty $\boldsymbol{\lambda}$ (3.33) do vzorce (3.32), při volbě $q_k = 1/\sigma_k^2$, získáme pro váhy w_k následující tvar:

$$w_k = d_k \left(1 + \frac{\mathbf{x}_k^\top}{\sigma_k^2} \left(\sum_{k \in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{\sigma_k^2} \right)^{-1} \left(\mathbf{X} - \hat{\mathbf{X}}_{HT} \right) \right), k \in s,$$

který je shodný se vzorcem (3.10), tedy získáváme shodný odhad s obecným regresním odhadem (3.8).

3.2.2 Obecné kalibrační odhady

V této části si odvodíme kalibrační odhady populačního úhrnu Y s obecnou funkcí vzdálenosti $G_k(w, d)$ (viz Särndal, 2010 a Deville a Särndal, 1992). Tedy se budeme zabývat úlohou, kde budeme minimalizovat

$$\sum_{k \in s} G_k(w_k, d_k)$$

za podmínky, že platí kalibrační rovnice (3.30). O funkci vzdálenosti G_k předpokládáme, že:

- pro každé pevné $d > 0$ je $G_k(w, d)$ nezáporná, ryze konvexní, diferencovatelná podle proměnné w a má spojitou derivací $g_k(w, d) = \frac{\partial G_k(w,d)}{\partial w}$,
- $G_k(d, d) = g_k(d, d) = 0$.

Dále předpokládejme, že g_k lze zapsat ve tvaru $g_k(w, d) = \frac{g(w/d)}{q_k}$, kde $q_k > 0$ nijak nesouvisí s d_k. Funkce $g(\cdot)$ je spojitá a ryze rostoucí (neboť G_k je ryze konvexní). O této funkci dále předpokládáme, že $g(1) = 0$ a $g'(1) = 1$, kde g' značí derivaci funkce g. První z těchto předpokladů nám říká, že pro $w = d$, tedy $w/d = 1$ (při aplikaci, že za nové váhy w_k uvažujeme původní váhy d_k), funkce $g(\cdot)$ nabývá svého minima, tj. pokud uvažujeme nové váhy w_k různé od původních vah d_k, tak jsou od původních vah d_k vzdálenější než samotné původní váhy d_k. Druhý předpoklad říká, že pro $w = d$ jsou původní funkce G_k na okolí tohoto bodu stejně zakřivené, tj. funkce $G_k(w, d)$ mají v bodě (d, d) stejné první 3 členy Taylorovy řady (členy do druhé derivace).

Při hledání nových vah w_k minimalizujeme zvolenou funkci vzdálenosti za podmínky, že platí kalibrační rovnice (3.30), což vede na následující úlohu:

$$\begin{aligned} &\min_{\mathbf{w}} \sum_{k \in s} G_k\left(w_k, d_k\right) \\ \text{za podmínky } &\sum_{k \in s} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k, \end{aligned}$$

kde $\mathbf{w} = \{w_k\}_{k \in s}$, kterou vyřešíme metodou Lagrangeových multiplikátorů, tj. Lagrangeova funkce má tvar

$$L\left(\mathbf{w}, \boldsymbol{\lambda}\right) = \sum_{k \in s} G_k\left(w_k, d_k\right) - \boldsymbol{\lambda}^\top \left(\sum_{k \in s} w_k \mathbf{x}_k - \sum_{k \in U} \mathbf{x}_k \right), \tag{3.34}$$

kde $\mathbf{w} = \{w_k\}_{k \in s}$ a $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)^\top$ je vektor Lagrangeových multiplikátorů. Zderivováním $L\left(\mathbf{w}, \boldsymbol{\lambda}\right)$ (3.34) podle proměnné w_k a položením rovno 0, získáme vztah

$$\frac{g\left(w_k/d_k\right)}{q_k} - \mathbf{x}_k^\top \boldsymbol{\lambda} = 0, k \in s.$$

Označíme-li F inverzní funkci k funkci g, tj. $F(u) = g^{-1}(u)$ (existence inverzní funkce F vyplývá z předpokladu, že G_k je ryze konvexní), získáme pro váhy w_k vztah:

$$w_k = d_k F\left(q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right), k \in s, \tag{3.35}$$

kde $\boldsymbol{\lambda}$ dopočítáme z kalibračních rovnic (3.30), tedy

$$\sum_{k \in s} d_k F\left(q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k. \tag{3.36}$$

Takto jsme odvodili kalibrační odhad populačního úhrnu Y a tento odhad je tvaru:

$$\hat{Y}_{kal} = \sum_{k \in s} d_k F\left(q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k, \tag{3.37}$$

kde $\boldsymbol{\lambda}$ dopočítáme ze vztahu (3.36).

Nyní se vrátíme k sekci 3.2.1, abychom ověřili předpoklady pro funkci vzdálenosti vah w_k a d_k, kterou jsme zavedli tak, abychom získali shodný odhad s obecným regresním odhadem (3.8). V tomto případě máme podle zavedeného značení následující: $G_k(w,d) = \frac{(w-d)^2}{2dq_k}$, zřejmě se tedy jedná o nezápornou funkci se spojitou derivací $g_k(w,d) = \frac{\partial G_k(w,d)}{\partial w} = \frac{(w-d)}{dq_k}$. Funkce $G_k(w,d)$ je ryze konvexní, což můžeme ověřit pomocí druhé derivace podle, která je rovna $\frac{1}{dq_k}$, tedy je kladná, neboť $d > 0$ a $q_k > 0$, z čehož vyplývá, že funkce $G_k(w,d)$ je ryze konvexní. Derivaci $g_k(w,d)$ lze zapsat ve tvaru $\frac{g(w/d)}{q_k}$, kde $g(w/d) = w/d - 1$, platí tedy, že $g(1) = 1 - 1 = 0$ a $g'(1) = 1$. Inverzní funkce F k funkci g je tvaru $F(u) = 1 + u$. Tedy vidíme, že funkce vzdálenosti $\frac{(w-d)^2}{2dq_k}$, splňuje všechny předpoklady, které jsme výše zavedli a podle (3.37) kalibrační odhad pro takto zvolenou funkci vzdálenosti má tvar

$$\hat{Y}_{kal} = d_k \left(1 + q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k,$$

kde $\boldsymbol{\lambda}$ lze vyjádřit ze vztahu (3.36), který je v tomto případě

$$\sum_{k \in s} d_k \left(1 + q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k.$$

Volby funkce vzdálenosti

V této části se budeme zabývat možnostmi, jak zvolit funkci vzdálenosti G_k, kterou měříme vzdálenosti nových vah w_k od původních vah d_k. Pro funkci G_k jsme zavedli několik předpokladů, které chceme, aby byly splněny. Názvy metod, jak zvolit funkci vzdálenosti G_k, jsou převzaty ze článků Deville a Särndal (1992) a Deville, Särndal a Sautory (1993).

Lineární metoda. V části 3.2.1 jsme již uvedli příklad takové funkce, jednalo se o funkci

$$G_k(w,d) = \frac{(w-d)^2}{2dq_k}.$$

Zderivováním této funkce podle proměnné w jsme získali, že

$$g_k(w,d) = \frac{w-d}{dq_k} = \frac{1}{q_k}\left(\frac{w}{d} - 1\right) = \frac{1}{q_k} g\left(\frac{w}{d}\right),$$

a dále jsme výše ukázali, že předpoklady funkcí G_k a g_k jsou splněny. Inverzní funkce F k funkci g je tedy tvaru $F(u) = 1 + u$.

V části 3.2.1 jsme pro tuto volbu funkce vzdálenosti odvodili nové váhy w_k (viz rovnost (3.32)). V tomto případě ovšem není zaručeno, že váhy jsou nezáporné. Tuto vlastnost bychom však od nových vah w_k očekávali. Tento problém můžeme vyřešit jinými volbami funkce vzdálenosti G_k, které nyní uvedeme (například viz Deville a Särndal, 1992).

Metoda *rakingu* (logitová metoda). První metodou, u které máme zaručeno, že váhy budou nezáporné je tzv. *raking*. V tomto případě je funkce vzdálenosti G_k definována následovně:

$$G_k(w,d) = \frac{1}{q_k}\left(w \log\left(\frac{w}{d}\right) - w + d\right),$$

tedy aby byla funkce logaritmus definovaná, tak musí platit, že $w > 0$. Zderivujeme-li tuto funkci podle proměnné w, dostaneme

$$g_k(w,d) = \frac{1}{q_k}\left(\log\left(\frac{w}{d}\right) + w\frac{d}{w}\frac{1}{d} - 1\right) = \frac{1}{q_k}\log\left(\frac{w}{d}\right).$$

Vidíme, že předpoklady o funkci vzdálenosti G_k, respektive g_k, jsou splněny: G_k je funkce se spojitou derivací, která lze zapsat ve tvaru $\frac{g(w/d)}{q_k}$, kde $g(x) = \log(x)$, $G_k(d,d) = \frac{1}{q_k}(-d+d) = 0$, $g_k(d,d) = 0$ a $g'(x)|_{x=1} = \frac{1}{x}|_{x=1} = 1$. Dále platí, že G_k je ryze konvexní, což můžeme vidět z druhé derivace této funkce podle proměnné w, pro kterou platí $\frac{\partial^2 G_k(w,d)}{\partial w^2} = \frac{1}{q_k w}$. Druhá derivace je kladná, neboť $w > 0$, tedy G_k je ryze konvexní. Inverzní funkce F k funkci g je tvaru $F(u) = \exp(u)$, tedy nové váhy w_k jsou podle rovnice (3.35) rovny

$$w_k = d_k \exp\left(q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right), k \in s, \tag{3.38}$$

zřejmě tedy platí, že $w_k > 0, k \in s$. Dříve zavedené předpoklady jsou tedy splněny a podle rovnice (3.37) dostaneme kalibrační odhad tvaru

$$\hat{Y}_{kal} = \sum_{k\in s} d_k \exp\left(q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k,$$

kde $\boldsymbol{\lambda}$ dopočítáme z kalibračních rovnic (3.36), které v tomto případě mají tvar

$$\sum_{k\in s} d_k \exp\left(q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right) \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k.$$

Můžeme si všimnout, že v případě *rakingu* mohou nové váhy w_k nabývat hodnot z intervalu $(0,\infty)$, neboť platí rovnice (3.38). Ovšem v některých případech se chceme vyvarovat extrémně velkým váhám, tedy je chceme nějak omezit, neboť pro nějaké výběry s bychom mohli v případě extrémních vah získat nesmyslný odhad.

Logitová (L,U) metoda. Abychom nové váhy w_k omezili, definujeme konstanty L a U takové, že $L < 1 < U$, a dále definujeme $A = \frac{U-L}{(1-L)(U-1)}$. Tímto omezením je motivována volba funkce vzdálenosti ve tvaru

$$G_k(w,d) = \frac{d}{q_k A}\left[\left(\frac{w}{d} - L\right)\log\frac{\frac{w}{d} - L}{1-L} + \left(U - \frac{w}{d}\right)\log\frac{U - \frac{w}{d}}{U-1}\right].$$

Derivací této funkce podle proměnné w získáme

$$\begin{aligned} q_k(w,d) &= \frac{d}{q_k A}\left[\frac{1}{d}\log\frac{\frac{w}{d} - L}{1-L} + \frac{1}{d} - \frac{1}{d}\log\frac{U - \frac{w}{d}}{U-1} - \frac{1}{d}\right] \\ &= \frac{1}{q_k}\left[\frac{1}{A}\left(\log\frac{\frac{w}{d} - L}{1-L} - \log\frac{U - \frac{w}{d}}{U-1}\right)\right]. \end{aligned}$$

Nyní opět ověříme předpoklady, funkce G_k má spojitou derivaci g_k, která lze zapsat ve tvaru $\frac{g(w/d)}{q_k}$, kde $g(x) = \frac{1}{A}\left(\log\frac{x-L}{1-L} - \log\frac{U-x}{U-1}\right)$, $G_k(d,d) = g_k(d,d) = 0$. Nyní ověříme, že $g'(1) = 1$. Díky tvaru konstanty A máme

$$g'(x) = \frac{1}{A}\frac{U-L}{(x-L)(U-x)} = \frac{(1-L)(U-1)}{U-L}\frac{U-L}{(x-L)(U-x)} = \frac{(1-L)(U-1)}{(x-L)(U-x)},$$

tedy $g'(1) = \frac{(1-L)(U-1)}{(1-L)(U-1)} = 1$. Inverzní funkce F k funkci g je tvaru

$$F(u) = \frac{L(U-1) + (1-L)U\exp(Au)}{U-1+(1-L)\exp(Au)}, \tag{3.39}$$

tedy nové váhy jsou podle rovnice (3.35) rovny

$$w_k = d_k \frac{L(U-1) + (1-L)U\exp(Aq_k\mathbf{x}_k^\top \boldsymbol{\lambda})}{U-1+(1-L)\exp(Aq_k\mathbf{x}_k^\top \boldsymbol{\lambda})}, k \in s.$$

Podíváme-li se podrobněji na inverzní funkci F (3.39), vidíme, že platí následující:

- $\lim_{u\to-\infty} F(u) = L$,
- $F(0) = 1$,
- $\lim_{u\to\infty} F(u) = U$.

Díky vztahu (3.35) pro nové váhy w_k vyplývá, že nové váhy w_k jsou v tomto případě omezeny následovně

$$Ld_k < w_k < Ud_k.$$

Lineární (L, U) metoda. Obdobně se lze vyhnout záporným vahám $w_k, k \in s$, při volbě funkce vzdálenosti $G_k(w, d) = \frac{(w-d)^2}{2dq_k}$ (viz část 3.2.1). Při volbě funkce vzdálenosti ve tvaru

$$G_k(w,d) = \begin{cases} \frac{(w-d)^2}{2dq_k}, & \text{když } L < \frac{w}{d} < U \\ \infty, & \text{jinak,} \end{cases}$$

kde $L < 1 < U$ a volbě $L > 0$ získáme nezáporné váhy $w_k, k \in s$.

Hellingerova metoda. Další možností, jak zvolit funkci vzdálenosti G_k, je volba

$$G_k(w,d) = \frac{2}{q_k}\left(\sqrt{w} - \sqrt{d}\right)^2.$$

Derivací této funkce podle proměnné w získáme

$$g_k(w,d) = \frac{1}{q_k} 2\left(1 - \sqrt{\frac{d}{w}}\right).$$

Snadno můžeme vidět, že jsou splněny všechny předpoklady o funkci G_k, respektive g_k. Inverzní funkce F je tvaru $F(u) = \left(1 - \frac{u}{2}\right)^{-2}$, tedy podle rovnice (3.37) získáme odhad ve tvaru

$$\hat{Y}_{kal} = \sum_{k\in s} d_k \left(1 - \frac{q_k\mathbf{x}_k^\top \boldsymbol{\lambda}}{2}\right)^{-2} y_k,$$

kde $\boldsymbol{\lambda}$ dopočítáme z kalibračních rovnic (3.36), které v tomto případě mají tvar

$$\sum_{k\in s} d_k \left(1 - \frac{q_k\mathbf{x}_k^\top \boldsymbol{\lambda}}{2}\right)^{-2} \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k.$$

Metoda minimalizace entropie. Dále ještě uvedeme funkci vzdálenosti ve tvaru

$$G_k(w,d) = \frac{1}{q_k}\left(-d\log\frac{w}{d} + w - d\right),$$

která má derivaci podle proměnné w tvaru

$$g_k(w,d) = \frac{1}{q_k}\left(1 - \frac{d}{w}\right).$$

O těchto funkcích opět snadno ověříme všechny předpoklady a dále dopočítáme, že pro inverzní funkci F platí $F(u) = (1-u)^{-1}$. Tedy dle rovnice (3.37) máme kalibrační odhad tvaru

$$\hat{Y}_{kal} = \sum_{k\in s} d_k \left(1 - q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right)^{-1},$$

kde $\boldsymbol{\lambda}$ dopočítáme z kalibračních rovnic (3.36), které v tomto případě mají tvar

$$\sum_{k\in s} d_k \left(1 - q_k \mathbf{x}_k^\top \boldsymbol{\lambda}\right)^{-1} \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k.$$

3.2.3 Rozptyl kalibračního odhadu

V této části se zaměříme na rozptyl obecného kalibračního odhadu $\hat{Y}_{kal}$ definovaného vztahem (3.37). Následující výsledek je převzat ze článku Deville a Särndal (1992, strana 379). Předpokládám-li navíc, že

- $\max_{k\in U_N} \|\mathbf{x}_{kN}\| = M < \infty$,
- $F''(0) = \tilde{M} < \infty$,

tak potom platí, že obecné kalibrační odhady $\hat{Y}_{kal}$ (3.37) jsou asymptoticky ekvivalentní s obecným regresním odhadem $\hat{Y}_{reg}$ (3.9), kde F jsou inverzní funkce uvedené v předchozí části 3.2.2, tedy příslušné funkce G_k, respektive g_k, splňují dříve uvedené předpoklady.

Poznámka 8. Asymptotická ekvivalence odhadů (3.37) a (3.9) je myšlena ve smyslu, že platí $\frac{1}{N}\left(\hat{Y}_{kal} - \hat{Y}_{reg}\right) = O_{\mathcal{P}}\left(\frac{1}{n_N}\right)$ (viz Deville a Särndal, 1992). △

Vzhledem k výsledku výše dále platí, že obecné kalibrační odhady $\hat{Y}_{kal}$ (3.37) a obecný regresní odhad $\hat{Y}_{reg}$ (3.9) mají stejný asymptotický rozptyl, který jsme odvodili v části 3.1.1. Tedy pro kalibrační odhady $\hat{Y}_{kal}$ (3.37) platí, že asymptotický rozptyl je tvaru (3.14).

4. Kalibrační odhady v kontingenčních tabulkách

V předchozí části této práce jsme se zabývali odvozením obecných kalibračních odhadů, které nyní budeme chtít aplikovat na příklady z praxe, kde často můžeme narazit na situace, kdy chceme dělat kalibrační odhady na základě známých četností v kontingenčních tabulkách. Rozlišujeme dva případy (viz Deville a kol., 1993):

- *úplná post-stratifikace* — známe četnosti jednotlivých buněk v kontingenční tabulce,
- *neúplná post-stratifikace* — známe marginální četnosti v kontingenční tabulce.

Větší pozornost budeme věnovat druhému případu, se kterým se v reálných situacích setkáváme častěji. Můžeme se s ním setkat například v případech, kdy známe data o populaci ze dvou zdrojů, kde z jednoho známe například věkové strukturu populace a ze druhého známe nejvyšší dosažené vzdělání této populace, ale neznáme překřížení těchto informací, tj. neznáme nejvyšší dosažené vzdělání podle různých věkových skupin populace.

Níže uvedené závěry platí pro kontingenční tabulky různých rozměrů a dimenzí, ale pro jednoduchost se zde budeme zabývat dvourozměrnými kontingenčními tabulkami o rozměrech $r \times c$. Máme tedy kontingenční tabulku s r řádky a c sloupci, která nám rozděluje populaci U, která obsahuje N jednotek, do $r \times c$ jednotlivých buněk. Pro buňku (i, j), $i = 1, \ldots, r$, $j = 1, \ldots, c$, máme populaci U_{ij}, která obsahuje N_{ij} jednotek. Platí tedy, že

$$U = \bigcup_{i=1}^{r} \bigcup_{j=1}^{c} U_{ij},$$
$$N = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij}.$$

4.1 Úplná post-stratifikace

Jak bylo uvedeno výše, při úplné post-stratifikaci známe četnosti buněk v kontingenční tabulce, tedy pro $i = 1, \ldots, r$ a $j = 1, \ldots, c$ známe počty N_{ij} jednotek z populace v jednotlivých buňkách. Znalost těchto počtů jednotek nyní chceme využít při hledání kalibračních odhadů (viz Deville a kol., 1993). Za tímto účelem definujeme vektor pomocných hodnot $\mathbf{x}_k$ tak, abychom věděli do jaké buňky $k-$tý jedinec z populace patří. Nejjednodušší tedy je definovat pomocný vektor $\mathbf{x}_k$ tak, že obsahuje $(rc - 1)$-krát hodnotu 0 a dále obsahuje jednu hodnotu 1. Hodnota 1 nám tedy říká, do jaké buňky $k-$tý jedinec z populace patří. Podíváme-li se nyní na kalibrační rovnice (3.30), tj. $\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X}$, vidíme, že v tomto případě platí, že

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k = (N_{11}, \ldots, N_{rc})^\top .$$

Tedy populační úhrn $\mathbf{X}$ pomocných hodnot $\mathbf{x}_k, k \in U$, obsahuje všechny známé počty jednotek N_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, v jednotlivých buňkách. Pokud nyní uvažujeme kalibrační odhady s obecnou funkcí vzdálenosti $G_k(w, d)$, která společně se svou derivací $g_k(w, d) = \frac{g(w/d)}{q_k}$ a také s inverzní funkcí F k funkci g splňují předpoklady z části 3.2.2, a dále pokud předpokládáme, že $q_k = 1 \forall k \in U$, tak kalibrační rovnice (3.30) můžeme podle vztahu (3.36) zapsat ve tvaru

$$\sum_{k \in s} d_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) \mathbf{x}_k = (N_{11}, \ldots, N_{rc})^\top . \tag{4.1}$$

Nyní si uvědomíme, že pro pro všechny jednotky k z buňky (i, j) je výraz $\mathbf{x}_k^\top \boldsymbol{\lambda}$ konstantní a označíme jej λ_{ij}, tj. $\lambda_{ij} = \mathbf{x}_k^\top \boldsymbol{\lambda}$ pro jednotky k patřící do buňky (i, j). Označíme-li s_{ij} množinu jednotek k z výběru s patřící zároveň do části populace U_{ij}, tj. $s_{ij} = s \cap U_{ij}$, potom můžeme kalibrační rovnice (4.1) přepsat do tvaru $\sum_{k \in s_{ij}} d_k F(\lambda_{ij}) = N_{ij}$, z čehož získáme vztah

$$F(\lambda_{ij}) = \frac{N_{ij}}{\sum_{k \in s_{ij}} d_k} \quad \forall i = 1, \ldots, r \ \forall j = 1, \ldots, c.$$

Dosadíme-li tento výsledek do vztahu pro nové váhy w_k (3.35), získáme, že

$$w_k = d_k \frac{N_{ij}}{\sum_{k \in s_{ij}} d_k}, k \in s_{ij}, \ i = 1, \ldots, r, j = 1, \ldots, c.$$

Pokud nyní vyjádříme hledaný kalibrační odhad populačního úhrnu Y (3.37), dostáváme

$$\hat{Y}_{kal} = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k \in s_{ij}} d_k \frac{N_{ij}}{\sum_{k \in s_{ij}} d_k} y_k = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \frac{\sum_{k \in s_{ij}} d_k y_k}{\sum_{k \in s_{ij}} d_k}. \tag{4.2}$$

4.2 Neúplná post-stratifikace

Jedním z možných důvodů, proč využít neúplnou post-stratifikaci, je, pokud neznáme četnosti ve všech buňkách kontingenční tabulky, ale známe pouze marginální četnosti v této kontingenční tabulce, tj. pro $i = 1, \ldots, r$ známe hodnoty $N_{i+} = \sum_{j=1}^{c} N_{ij}$ a pro $j = 1, \ldots, c$ známe hodnoty $N_{+j} = \sum_{i=1}^{r} N_{ij}$ (viz Deville a Särndal, 1992 a Deville a kol., 1993). Dalším důvodem, proč upřednostňujeme neúplnou post-stratifikaci před úplnou, je, pokud v některých buňkách kontingenční tabulky máme malý nebo nulový počet jednotek. Pokud bychom v tomto případě použili úplnou post-stratifikaci, mohla by nastat situace, že kalibrační odhad (4.2) nebude definován, neboť $\sum_{k \in s_{ij}} d_k = 0$.

Při neúplné post-stratifikaci se tedy snažíme využít známé hodnoty marginálních četností $N_{i+} = \sum_{j=1}^{c} N_{ij}$, $i = 1, \ldots, r$, a $N_{+j} = \sum_{i=1}^{r} N_{ij}$, $j = 1, \ldots, c$, při hledání kalibračních odhadů. Za tímto účelem definujeme vektory pomocných hodnot $\mathbf{x}_k$ následovně

$$\mathbf{x}_k = (\delta_{1.k}, \ldots, \delta_{r.k}, \delta_{.1k}, \ldots, \delta_{.ck})^\top ,$$

kde $\delta_{i.k} = 1$, jestliže $k-$tý jedinec patří do $i-$tého řádku a $\delta_{i.k} = 0$ v ostatních případech. Obdobně je definováno i $\delta_{.jk}$. Pomocné vektory $\mathbf{x}_k$ obsahují tedy $r + c$

hodnot, z toho dvakrát obsahují hodnotu 1, které nám udávají do jakého řádku a sloupce $k-$tý jedinec z populace patří a ostatní hodnoty jsou 0 (tj. $(r+c-2)$-krát obsahují hodnotu 0). Díky této definici vektorů pomocných hodnot $\mathbf{x}_k$ dostáváme, že

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k = (N_{1+}, \ldots, N_{r+}, N_{+1}, \ldots, N_{+c})^\top .$$

Pokud opět předpokládáme, že funkce $G_k(w, d)$, $g_k(w, d)$ a $F(u)$ splňují předpoklady zavedené v části 3.2.2 a pokud navíc platí, že $q_k = 1 \, \forall k \in U$, tak můžeme kalibrační rovnice (3.30) zapsat ve tvaru

$$\sum_{k \in s} d_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) \mathbf{x}_k = (N_{1+}, \ldots, N_{r+}, N_{+1}, \ldots, N_{+c})^\top . \tag{4.3}$$

Všimněme si, že pro všechny jedince k patřící do $i-$tého řádku a $j-$tého sloupce platí, že $\mathbf{x}_k^\top \boldsymbol{\lambda}$ je konstantní, pokud označíme $\boldsymbol{\lambda} = (u_1, \ldots, u_r, v_1, \ldots, v_c)^\top$, tak potom platí, že $\mathbf{x}_k^\top \boldsymbol{\lambda} = u_i + v_j$, pokud $k-$tý jedinec patří do buňky (i, j), pro $i = 1, \ldots, r$ a $j = 1, \ldots, c$. Podobnou úvahou jako u úplné post-stratifikace dostáváme, že kalibrační rovnice (3.36) můžeme zapsat ve tvaru

$$\begin{aligned} \sum_{j=1}^{c} F(u_i + v_j) \sum_{k \in s_{ij}} d_k &= N_{i+}, \quad i = 1, \ldots, r, \\ \sum_{i=1}^{r} F(u_i + v_j) \sum_{k \in s_{ij}} d_k &= N_{+j}, \quad j = 1, \ldots, c. \end{aligned} \tag{4.4}$$

Máme tedy $r + c$ rovnic o $r + c$ neznámých. Jedna z rovnic (4.4) je ovšem přebytečná, neboť lze vyjádřit jako lineární kombinace zbývajících $r+c-1$ rovnic. Tedy jednu z neznámých můžeme zvolit jako parametr, konkrétně zvolíme, že $v_c = 0$, a poté řešíme rovnice (4.4) pouze pro $i = 1, \ldots, r$ a $j = 1, \ldots, c - 1$.

Poznámka 9. Parametr v_c můžeme zvolit rovný 0, neboť v rovnicích (4.4) pracujeme vždy s argumentem funkce $F(\cdot)$ ve tvaru $u_i + v_j$ pro $i = 1, \ldots, r$ a $j = 1, \ldots, c$, tedy pokud bychom zvolili $v_c = v$, byly by ostatní parametry o v posunuté, ale součet $u_i + v_j$ by byl stejný.

△

Poznámka 10. Jak je uvedeno ve článku Deville a Särndal (1992), k vyřešení rovnic (4.4) je často potřeba použít iterativní řešení.

△

Jestliže z rovnic (4.4) máme vypočítány jednotlivé složky vektoru $\boldsymbol{\lambda}$, tj. známe u_i, $i = 1, \ldots, r$, a v_j, $j = 1, \ldots, c$, potom lze odhadnout počty jednotek N_{ij} v jednotlivých buňkách na základě rovnic (4.4) pomocí vztahu

$$\hat{N}_{ij}^w = F(u_i + v_j) \sum_{k \in s_{ij}} d_k, \; i = 1, \ldots, r, \; j = 1, \ldots, c.$$

Vyjádříme-li z výše uvedeného vztahu předpis pro funkci $F(\cdot)$, dostaneme, že

$$F(u_i + v_j) = \frac{\hat{N}_{ij}^w}{\sum_{k \in s_{ij}} d_k} \;\; i = 1, \ldots, r, j = 1, \ldots, c. \tag{4.5}$$

Pokud dosadíme rovnost (4.5) do vztahu pro nové váhy w_k (3.35), získáme, že

$$w_k = d_k \frac{\hat{N}_{ij}^w}{\sum_{k \in s_{ij}} d_k}, k \in s_{ij},\ i = 1, \ldots, r, j = 1, \ldots, c.$$

Tedy hledaný kalibrační odhad $\hat{Y}_{kal}$ populačního úhrnu Y je podle vzorce (3.37) tvaru

$$\hat{Y}_{kal} = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k \in s_{ij}} d_k \frac{\hat{N}_{ij}^w}{\sum_{k \in s_{ij}} d_k} y_k = \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{N}_{ij}^w \frac{\sum_{k \in s_{ij}} d_k y_k}{\sum_{k \in s_{ij}} d_k}. \tag{4.6}$$

Porovnáme-li nyní kalibrační odhady populačního úhrnu Y získané při úplné a neúplné post-stratifikaci (odhady (4.2), respektive (4.6)), vidíme, že mají stejný tvar a liší se pouze v tom, že u odhadu při úplné post-stratifikaci pracujeme ve vzorci (4.2) se známými počty jednotek $N_{ij}, i = 1, \ldots, r,\ j = 1, \ldots, c$ a u odhadu při neúplné post-stratifikaci (4.6) pracujeme s odhadnutými počty jednotek $\hat{N}_{ij}^w, i = 1, \ldots, r,\ j = 1, \ldots, c$.

4.2.1 Aplikace pro lineární metodu a metodu *rakingu*

V této části práce uvedeme kalibrační odhad $\hat{Y}_{kal}$ populačního úhrnu Y při neúplné post-stratifikaci (odhad ve tvaru (4.6)) pro konkrétní volby funkce vzdálenosti $G_k(\cdot, \cdot)$, tedy i pro příslušné funkce $F(\cdot)$, které ve vztahu (4.6) využijeme. Některé možnosti, jak zvolit tyto funkce, jsme uvedli v podkapitole 3.2.2.

První možností, která byla výše uvedena, byla lineární metoda, pro kterou platí, že $G_k(w, d) = \frac{(w-d)^2}{2dq_k}$ a příslušná inverzní funkce F je tvaru $F(u) = 1 + u$. Tedy s využitím vztahu (4.5) je kalibrační odhad (4.6) tvaru

$$\hat{Y}_{kal} = \sum_{i=1}^{r} \sum_{j=1}^{c} (1 + u_i + v_j) \sum_{k \in s_{ij}} d_k y_k,$$

kde $u_i, i = 1, \ldots, r$ a $v_j, j = 1, \ldots, c$ vypočítáme z rovnic (4.4), které v případě lineární metody mají tvar

$$\begin{aligned} \sum_{j=1}^{c} (1 + u_i + v_j) \sum_{k \in s_{ij}} d_k &= N_{i+},\ \ i = 1, \ldots, r, \\ \sum_{i=1}^{r} (1 + u_i + v_j) \sum_{k \in s_{ij}} d_k &= N_{+j},\ \ j = 1, \ldots, c. \end{aligned} \tag{4.7}$$

Pokud označíme n_{ij} počet jednotek v podvýběru s_{ij}, tj. počet jednotek které patří do výběru s a zároveň do části populace U_{ij}, tak v případě prostého náhodného výběru bez vracení, pro který platí, že $d_k = \frac{1}{\pi_k} = \frac{N}{n}$ (viz (2.1)), lze rovnice (4.7) pro $i = 1, \ldots, r$ upravit, stejně jak je uvedeno v článku Deming a Stephan (1940), do tvaru:

$$\sum_{j=1}^{c}(1+u_i+v_j)\sum_{k\in s_{ij}} d_k = N_{i+},\ \ i=1,\ldots,r,$$

$$\frac{N}{n}\left[\sum_{j=1}^{c}\sum_{k\in s_{ij}} 1 + u_i\sum_{j=1}^{c}\sum_{k\in s_{ij}} 1 + \sum_{j=1}^{c} v_j \sum_{k\in s_{ij}} 1\right] = N_{i+},\ \ i=1,\ldots,r,$$

$$\sum_{j=1}^{c} n_{ij} + u_i\sum_{j=1}^{c} n_{ij} + \sum_{j=1}^{c} v_j n_{ij} = N_{i+}\frac{n}{N},\ \ i=1,\ldots,r,$$

$$n_{i+} + u_i n_{i+} + \sum_{j=1}^{c} v_j n_{ij} = m_{i+},\ \ i=1,\ldots,r,$$

kde $n_{i+} = \sum_{j=1}^{c} n_{ij}$ a $m_{i+} = N_{i+}\frac{n}{N}$. Obdobně můžeme upravit rovnice pro $j = 1,\ldots,c$ a celkem rovnice (4.7) můžeme přepsat do tvaru

$$u_i n_{i+} + \sum_{j=1}^{c} v_j n_{ij} = m_{i+} - n_{i+},\ \ i=1,\ldots,r,$$

$$v_j n_{+j} + \sum_{i=1}^{r} u_i n_{ij} = m_{+j} - n_{+j},\ \ j=1,\ldots,c,$$

kde $n_{+j} = \sum_{i=1}^{r} n_{ij}$ a $m_{+j} = N_{+j}\frac{n}{N}$. Způsob, jak vyřešit tyto rovnice, je uveden v článku Deming a Stephan (1940).

Další možností volby funkce vzdálenosti byla metoda *rakingu*, tj. $G_k(w,d) = \frac{1}{q_k}\left(w\log\left(\frac{w}{d}\right) - w + d\right)$, která byla motivována tím, abychom se vyvarovali záporným vahám, které mohou vycházet u lineární metody. Jak je uvedeno v části 3.2.2, příslušná inverzní funkce F k derivaci funkce G_k je tvaru $F(u) = \exp(u)$, tedy opět s využitím vztahu (4.5) je kalibrační odhad (4.6) tvaru

$$\hat{Y}_{kal} = \sum_{i=1}^{r}\sum_{j=1}^{c}\exp(u_i+v_j)\sum_{k\in s_{ij}} d_k y_k,$$

kde $u_i, i=1,\ldots,r$ a $v_j, j=1,\ldots,c$ vypočítáme z rovnic (4.4), které v případě metody *rakingu* mají tvar

$$\sum_{j=1}^{c}\exp(u_i+v_j)\sum_{k\in s_{ij}} d_k = N_{i+},\ \ i=1,\ldots,r,$$

$$\sum_{i=1}^{r}\exp(u_i+v_j)\sum_{k\in s_{ij}} d_k = N_{+j},\ \ j=1,\ldots,c.$$

V případě prostého náhodného výběru, $d_k = \frac{N}{n}$, a při značení, které bylo zavedeno výše, lze tyto rovnice zapsat ve tvaru

$$\exp(u_i)\sum_{j=1}^{c}\exp(v_j)n_{ij} = m_{i+},\ \ i=1,\ldots,r,$$

$$\exp(v_j)\sum_{i=1}^{r}\exp(u_i)n_{ij} = m_{+j},\ \ j=1,\ldots,c.$$

Jak je uvedeno v článku Deville a Särndal (1992), řešení těchto rovnic lze získat pomocí *IPF* algoritmu (*iterative proportional fitting*), který je pro lineární metodu uveden v článku Deming a Stephan (1940).

5. Aplikace

V této části práce aplikujeme dříve uvedené teoretické závěry na reálná data, konkrétně se budeme zabývat poslechovostí rádií obyvatel České republiky — tzv. RADIOPROJEKTem. Tento projekt je realizován společnostmi MEDIAN a STEM/MARK. Od těchto společností máme k dispozici data a také soubor, ve kterém je popsáno, jak tento výzkum vznikal, a ze kterého tato část práce vychází. Výzkum je prováděn pomocí telefonních rozhovorů a k dispozici máme data ze čtvrtého čtvrtletí roku 2015 a z prvního čtvrtletí roku 2016 (tj. z období od 1.10.2015 – 31.3.2016), kdy bylo dotázáno 15245 respondentů. Respondenti pro výzkum jsou vybíráni pomocí generování náhodných telefonních čísel a databáze telefonních čísel domácností ČR (pevné linky). Dotazovány jsou pouze osoby ve věku 12 – 79 let trvale bydlící na území České republiky.

Vzhledem k tomu, že od roku 2007 je normovaná velikost vzorku určena čtvrtletně na 7500 respondentů, byla velikost výběru převážena na 15000 respondentů, neboť máme data za dvě čtvrtletí. K převážení dat byl použit algoritmus *iterative proportional fitting* (Deming a Stephan, 1940). Data byla převážena podle vybraných základních údajů a také pomocí jejich vybraných dvojných kombinací tak, aby převážená data odpovídala předepsaným četnostem v celé populaci obyvatel České republiky ve věku 12 – 79 let. Vzhledem k tomu, že z dostupných dat známe pouze marginální četnosti obyvatel v České republice (například počet mužů, počet obyvatel Pardubického kraje, ...), případně jejich dvojné překřížení (například počet mužů v Pardubickém kraji a další), jedná se o neúplnou post-stratifikaci (viz sekce 4.2). Při převažování dále společnosti kontrolují hodnoty vypočítaných vah a to tím způsobem, že tyto váhy musí být v intervalu od 0.3 do 3. Pokud některá z vah vyšla mimo tento interval, byla její hodnota poté nastavena na bližší z krajních hodnot intervalu. Na tomto místě je třeba zdůraznit fakt, že z dostupných informací není zřejmé, zda hodnota vah mimo interval byla měněna v průběhu algoritmu, nebo až na jeho konci. Tento fakt budeme muset brát v potaz při simulační studii, která následuje v další kapitole této práce.

5.1 Odhady poslechovosti rádií na základě metod společností MEDIAN a STEM/MARK

Jak bylo zmíněno výše, zabýváme se poslechovostí radií v České republice. Tato poslechovost se uvádí v tisících obyvatel. Populačním úhrnem poslechovosti rádia rozumíme počet všech občanů České republiky ve věku 12 – 79 let, kteří během včerejšího dne poslouchali dané rádio. Pokud populační úhrn poslechovosti odhadneme stejným způsobem, jakým to dělají společnosti MEDIAN a STEM/MARK, tak odhad populačního úhrnu poslechovosti daného rádia je určen součtem vah všech respondentů, kteří v telefonním rozhovoru uvedli, že během včerejšího dne dané rádio poslouchali. Tento součet vah je následně vhodně převážen na celou populaci. Označíme-li Y_R počet lidí z celé populace, kteří včera poslouchali konkrétní rádio R, tak odhad populačního úhrnu poslechovosti $\hat{Y}_R$

pro konkrétní rádio R vypočítáme podle vzorce

$$\hat{Y}_R = N\frac{\sum_{k\in s} y_k w_k}{\sum_{k\in s} w_k}, \tag{5.1}$$

kde N je velikost celé populace U, s je výběr z této populace, w_k je váha pro $k-$tého respondenta, $y_k = 1$, pokud $k-$tý respondent včera poslouchal rádio R a $y_k = 0$, pokud rádio R včera neposlouchal. Označíme-li $w = \sum_{k\in s} w_k$ a $\bar{y}_w = \frac{1}{w}\sum_{k\in s} y_k w_k$, potom máme, že $\hat{Y}_R = N\bar{y}_w$. Pro data, která máme k dispozici, platí, že $w = 15005.1$, což není přesně 15000, jak uvádějí společnosti. Tento rozdíl je způsoben jednak tím, že váhy mimo interval 0.3 – 3 byly změněny na bližší z krajních hodnot tohoto intervalu a také tím, že váhy jsou zaokrouhlené na jedno desetinné místo. Vzhledem k tvaru odhadu populačního úhrnu poslechovosti rádia R (5.1) a vzhledem k tomu, že váhy w_k udávané společnostmi byli vypočítány tak, aby převážená data odpovídala teoretickým četnostem v populaci České republiky, se jedná také o kalibrační odhad.

Nyní se zaměříme na asymptotický interval spolehlivosti pro populační úhrn poslechovosti Y_R. Na základě souboru, který poskytly společnosti STEM/MARK a MEDIAN, víme, že k výpočtu asymptotického intervalu spolehlivosti s pravděpodobností pokrytí 0.95 používají vzorec

$$\left(\hat{Y}_R - u_{0.975}\sqrt{\bar{y}_w\left(1-\bar{y}_w\right)w}\frac{N}{w}, \hat{Y}_R + u_{0.975}\sqrt{\bar{y}_w\left(1-\bar{y}_w\right)w}\frac{N}{w}\right), \tag{5.2}$$

kde u_α značí α-kvantil rozdělení $\mathsf{N}(0,1)$. Tvar tohoto asymptotického intervalu spolehlivosti je získán na základě tvaru asymptotického intervalu spolehlivosti v případě prostého náhodného výběru, který je založen na asymptotické normalitě odhadu populačního úhrnu (viz vztah (2.15)). Z tohoto vztahu je asymptotický interval spolehlivosti pro populační úhrn Y_N s pravděpodobností pokrytí $(1-\alpha)$ tvaru:

$$\left(\hat{Y}_N - u_{1-\alpha/2}\sqrt{\widehat{\operatorname{var}\left(\hat{Y}_N\right)}}, \hat{Y}_N + u_{1-\alpha/2}\sqrt{\widehat{\operatorname{var}\left(\hat{Y}_N\right)}}\right), \tag{5.3}$$

kde $\hat{Y}_N = \frac{N}{n}\sum_{k\in s} y_k = N\bar{y}$. V případě odhadu rozptylu odhadu populačního úhrnu za předpokladu prostého náhodného výběru dostáváme díky vztahu (2.5), že

$$\begin{aligned}
\widehat{\operatorname{var}\left(\hat{Y}_N\right)} &= \widehat{\operatorname{var}\left(N\bar{y}\right)} = N^2\left(1-\frac{n}{N}\right)\frac{1}{n}s_y^2 = N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{n-1}\sum_{k\in s}\left(y_k-\bar{y}\right)^2 \\
&= N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{n-1}\left(\sum_{k\in s} y_k^2 - n\bar{y}^2\right) \\
&= \frac{n}{n-1}\left(1-\frac{n}{N}\right)\frac{N^2}{n}\left(\frac{1}{n}\sum_{k\in s} y_k^2 - \bar{y}^2\right),
\end{aligned}$$

což za předpokladu, že $y_k \in \{0,1\}, k \in U$, lze zapsat ve tvaru

$$\widehat{\operatorname{var}\left(\hat{Y}_N\right)} = \frac{n}{n-1}\left(1-\frac{n}{N}\right)\frac{N^2}{n}\left(\bar{y}-\bar{y}^2\right), \tag{5.4}$$

neboť $\frac{1}{n}\sum_{k\in s} y_k^2 = \frac{1}{n}\sum_{k\in s} y_k = \bar{y}$ (2.2). Tedy asymptotický interval spolehlivosti pro populační úhrn Y_N s pravděpodobností pokrytí $(1-\alpha)$ (5.3) můžeme v případě prostého náhodného výběru a za předpokladu, že $y_k \in \{0,1\}, k \in U$, zapsat ve tvaru

$$\left(\hat{Y}_N - u_{1-\alpha/2}\sqrt{\frac{n}{n-1}\left(1-\frac{n}{N}\right)\frac{N^2}{n}(\bar{y}-\bar{y}^2)},\right.$$
$$\left.\hat{Y}_N + u_{1-\alpha/2}\sqrt{\frac{n}{n-1}\left(1-\frac{n}{N}\right)\frac{N^2}{n}(\bar{y}-\bar{y}^2)}\right), \quad (5.5)$$

Interval spolehlivosti (5.2) má obdobný předpis jako interval spolehlivosti (5.5) na základě předpokladu prostého náhodného výběru. Za odhad populačního úhrnu Y_N uvažují společnosti odhad $\hat{Y}_R$ (5.1), obdobně to je i u odhadu rozptylu, který je navíc značně zjednodušený. Z tvaru intervalu spolehlivosti (5.2) máme, že platí:

$$\widehat{\text{var}}\left(\hat{Y}_R\right) = \left(\sqrt{\bar{y}_w\left(1-\bar{y}_w\right)w}\frac{N}{w}\right)^2 = \bar{y}_w(1-\bar{y}_w)w\frac{N^2}{w^2} = \frac{N^2}{w}\left(\bar{y}_w - \bar{y}_w^2\right). \quad (5.6)$$

Porovnáme-li nyní vzorce pro odhad rozptylu (5.4) a (5.6) a uvážíme-li, že společnosti místo n, respektive $\bar{y}$, do předpisu pro odhad rozptylu (5.4) dosazují w, respektive $\bar{y}_w$, můžeme si všimnout, že odhad rozptylu (5.4) je přenásoben výrazem $\frac{n}{n-1}\left(1-\frac{n}{N}\right)$. Pokud předpokládáme stejně jako v kapitole 2, že rozsah výběru n závisí na velikosti populace N a platí $\lim_{N\to\infty} n_N = \infty$ a dále, že $\limsup_{N\to\infty} \frac{n_N}{N} < 1$ (předpoklady (b) a (c) z části 2.1), tak potom platí

$$\frac{n_N}{n_N-1}\left(1-\frac{n_N}{N}\right) \xrightarrow[N\to\infty]{} 1\cdot(1-c),$$

kde $c \in (0,1)$ je konstanta. Za předpokladu, že číslo c bude blízko 0 (tj. $N >> n_N$), dostáváme, že odhady rozptylu (5.4) a (5.6) mají podobný tvar (pro naše data platí, že $c = \frac{15245}{8795000} \doteq 0.002$).

Výše uvedené závěry pro odhad poslechovosti Y_R rádia R (5.1) a také asymptotický interval spolehlivosti pro poslechovost rádia R (5.2) následně porovnáme společně s odhady založenými na kalibračních odhadech, kterými se budeme zabývat nyní.

5.2 Odhady poslechovosti rádií na základě teorie kalibračních odhadů

Jak bylo zmíněno v kapitole 3, u kalibračních odhadů využíváme pomocné informace, které o jednotkách z populace (výběru) známe. Tyto pomocné informace následně mají vliv jak na bodový, tak na intervalový odhad populačního úhrnu. Při výpočtu těchto odhadů budeme uvažovat stejné proměnné, které využívaly společnosti MEDIAN a STEM/MARK pro výpočet vah. Jedná se o kategoriální proměnné (například věk respondenta, pohlaví respondenta, ...) a vybrané jejich dvojné kombinace (pohlaví-věk, pohlaví-kraj, ...). Celkem se jedná o devět proměnných a 13 dvojných kombinací.

Poznámka 11. V případě výpočtu vah společnostmi MEDIAN a STEM/MARK se jednalo o deset kategoriálních proměnných, zatímco jak je zmíněno výše, my

jsme využili pouze devět kategoriálních proměnných, neboť z dostupných zdrojů o datech, které máme k dispozici, nebylo možné zbývající proměnnou získat. Tato vynechaná proměnná při převažování nefigurovala v žádné dvojné kombinaci.

△

Vzhledem k tomu, že při výpočtu odhadů poslechovosti rádií podle společností MEDIAN a STEM/MARK byl asymptotický interval spolehlivosti založený na prostém náhodném výběru, tak pro odhad populačního úhrnu poslechovosti rádia R využijeme vzorec (3.21). Z tohoto vztahu vidíme, že musíme znát populační úhrn X, tj. znát počty lidí z celé populace v jednotlivých kategoriích u všech proměnných a všech dvojných kombinací, které byly použity k výpočtu vah. Z přiloženého souboru od společností jsme schopni získat ovšem pouze počty v jednotlivých kategoriích všech proměnných, ale už nejsme schopni získat všechny počty u dvojných kombinací těchto proměnných. Tyto hodnoty jsme tedy museli dopočítat (viz následující poznámka 12). Výpočet jsme založili na příslušných vahách, které společnosti používají. Vzhledem k tomuto faktu ovšem musíme brát výsledky s rezervou, neboť kvůli tomu, že váhy mimo interval 0.3 – 3 byly měněny na bližší z krajních hodnot tohoto intervalu, tak dopočítané počty u jednotlivých dvojných kombinací jsou tímto faktem ovlivněny a nejsou tedy natolik přesné.

Poznámka 12. V této poznámce detailně popíšeme, jak jsme vypočítali požadované počty u dvojných kombinací vybraných proměnných. Uvažujme dvojnou kombinaci kategoriálních proměnných A a B, kde proměnná A má p_1 kategorií — $A_1, \ldots, A_{p_1}$, obsahující po řadě $N_{A_1}, \ldots, N_{A_{p_1}}$ jednotek z populace, a proměnná B má p_2 kategorií — $B_1, \ldots, B_{p_2}$, obsahující po řadě $N_{B_1}, \ldots, N_{B_{p_2}}$ jednotek z populace, kde $N_i, i = A_1, \ldots, A_{p_1}, B_1, \ldots, B_{p_2}$, známe. Zřejmě platí, že $N_{A_1}+\cdots+N_{A_{p_1}} = N_{B_1}+\cdots+N_{B_{p_2}} = N$. Pro $i = 1, \ldots, p_1$ a $j = 1, \ldots, p_2$ chceme znát počet N_{ij} jednotek z populace v buňce (i, j), tj. počet jednotek spadajících do kategorie A_i a B_j. K tomu využijeme znalosti vah jednotek z výběru, které spadají do kategorie A_i a B_j, neboť tyto váhy jsou zvoleny tak, aby převážená data odpovídala předepsaným četnostem této dvojné interakce v celé populaci. Sečteme tedy jejich váhy a podělíme celkovým součtem vah w všech jednotek ve výběru. Takto získáme poměr zastoupení jednotek ve výběru patřící do buňky (i, j). Pokud tento poměr přenásobíme velikostí populace N, získáme odhadnutý počet $\hat{N}_{ij}$, který následně využijeme při počítání odhadu (3.21) a dalších charakteristik závislých na populačním úhrnu X.

△

Asymptotický interval spolehlivosti pro populační úhrn Y_R můžeme založit na vztahu (3.24), případně, jak je ukázáno v části 3.1.2, na vztahu (3.29). Tyto asymptotické intervaly spolehlivosti s pravděpodobností pokrytí $1 - \alpha$ mají tvar:

$$\left(\hat{Y}_R \;-\; u_{1-\alpha/2}\sqrt{N^2\left(1-\frac{n}{N}\right)\frac{1}{n}s^2_{\hat{z}}},\;\; \hat{Y}_R \;-\; u_{1-\alpha/2}\sqrt{N^2\left(1-\frac{n}{N}\right)\frac{1}{n}s^2_{\hat{z}}}\right), \quad (5.7)$$

případně

$$\left(\hat{Y}_R - u_{1-\alpha/2} \sqrt{N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{n-1} \sum_{k \in s} \left(g_k \hat{z}_k - \frac{1}{n} \sum_{k \in s} g_k \hat{z}_k \right)^2}, \right.$$

$$\left. \hat{Y}_R - u_{1-\alpha/2} \sqrt{N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{n-1} \sum_{k \in s} \left(g_k \hat{z}_k - \frac{1}{n} \sum_{k \in s} g_k \hat{z}_k \right)^2} \right), \quad (5.8)$$

kde $\hat{z}_k = y_k - \hat{\boldsymbol{\beta}}_s^\top \mathbf{x}_k = y_k - \left(\sum_{k \in s} \mathbf{x}_k^\top y_k \right) \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \mathbf{x}_k$, $k \in s$, $\bar{\hat{z}} = \frac{1}{n} \sum_{k \in s} \hat{z}_k$, $s_{\hat{z}}^2 = \frac{1}{n-1} \sum_{k \in s} \left(\hat{z}_k - \bar{\hat{z}} \right)^2$ a $g_k = 1 + \mathbf{x}_k^\top \left(\frac{N}{n} \sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \left(\mathbf{X} - \hat{\mathbf{X}}_{HT} \right)$, $k \in s$.

5.3 Porovnání přístupů

Nyní na základě dat z RADIOPROJEKTu porovnáme odhady poslechovosti rádií založených na metodách společností MEDIAN a STEM/MARK s odhady založenými na teorii kalibračních odhadů, která byla popsána v této práci. Pro srovnání navíc porovnáme odhady populačních úhrnů s neváženým odhadem populačního úhrnu (2.2), tj. s úhrnem $\hat{Y}_R = N\bar{y}$, kde $\bar{y} = \frac{1}{n} \sum_{k \in s} y_k$ a $y_k = 1$, pokud $k-$tý jedinec z výběru s poslouchal rádio R a $y_k = 0$ jinak. Tyto odhady označíme písmenem P a jejich hodnoty uvedeme v tabulce 5.1, ve které jsou uvedeny odhady populačních úhrnů poslechovosti deseti nejposlouchanějších rádií v České republice v tisících obyvatel a jejich směrodatné odchylky. Písmeno M označuje přístup založený na metodách společností MEDIAN a STEM/MARK,

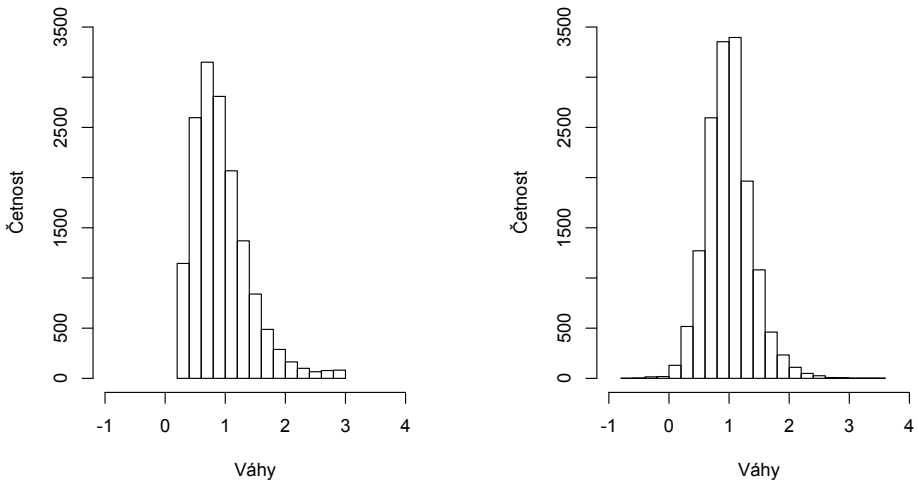
Tabulka 5.1: Odhady populačních úhrnů poslechovosti deseti nejposlouchanějších rádií v České republice ve čtvrtém čtvrtletí roku 2015 a prvním čtvrtletí roku 2016 uvedených v tisících obyvatel spolu s jejich směrodatnými odchylkami. Písmeno P značí nevážený populační úhrn, písmeno M značí výpočet odhadu a jeho směrodatné odchylky podle metod společností MEDIAN a STEM/MARK, zkratka DP značí výpočet odhadu populačního úhrnu a jeho směrodatné odchylky pomocí teorie kalibračních odhadů. Výpočet směrodatné odchylky DPg je motivován poznámkou 5.

Rádio	Odhad úhrnu			Směrodatná odchylka		
	P	M	DP	M	DP	DPg
Rádio Impuls	958.2	962.1	958.5	22.41	21.49	22.84
Evropa 2	889.6	894.9	872.8	21.71	20.33	21.33
Frekvence 1	863.1	892.7	857.7	21.68	20.68	21.82
ČRo Radiožurnál	822.7	825.2	796.9	20.94	19.91	20.59
Rádio Blaník	613.8	641.5	625.7	18.67	17.33	18.66
ČRo Dvojka (Praha)	419.4	393.3	409.1	14.84	14.29	15.05
Rádio Beat	255.0	254.2	244.8	12.03	11.61	11.84
Country radio	225.0	239.1	235.0	11.68	10.96	12.15
Hitrádio Orion	152.9	156.7	151.9	9.50	8.71	9.03
Radio Čas	147.7	152.0	143.6	9.36	8.80	9.07

tj. odhady populačního úhrnu vypočítané podle vzorce (5.1) a směrodatné odchylky použité ve vzorci (5.2), a zkratka DP, respektive DPg, označuje přístup založený na teorii kalibračních odhadů, tj. odhady populačního úhrnu vypočítané podle vzorce (3.21) a směrodatné odchylky použité ve vzorci (5.7), respektive ve vzorci (5.8). Výpočty využité k získání následující tabulky 5.1 byly provedeny ve statistickém softwaru R (R Core Team, 2016).

Vidíme, že se liší jak odhady populačních úhrnů poslechovosti rádií počítané podle přístupu M a podle přístupu DP, tak směrodatné odchylky těchto odhadů použité k výpočtů intervalů spolehlivosti. Můžeme si všimnout, že nevážené odhady P populačního úhrnu poslechovosti se od populačních úhrnů vypočítaných na základě zbývajících přístupů liší, ovšem nelze obecně říci, zda jsou vyšší či nižší. Porovnáme-li odhady populačního úhrnu vypočítané podle přístupů M a podle přístupu DP, vidíme, že odhady populačního úhrnu spočítané podle přístupu M vycházejí vyšší, až na výjimku v podobě rádia ČRo Dvojka (Praha), u kterého je odhad populačního úhrnu DP větší o 3.9%. U Radia Čas je odhad populačního úhrnu M větší o 5.9%, zatímco u Rádia Impuls je to pouze o 0.4%, ostatní rádia se pohybují v tomto rozmezí. Porovnáme-li směrodatné odchylky spočítané na základě různých přístupů, vidíme, že směrodatné odchylky vypočítané podle přístupu DPg, tedy bez využití vah g_k, $k \in s$, vycházejí nejnižší. Pokud dále srovnáme zbylé dva přístupy, můžeme vidět, že tyto směrodatné odchylky vycházejí podobně a nelze obecně říci, že jedny jsou vyšší než druhé.

Rozdíly mezi jednotlivými odhady populačních úhrnů a jejich směrodatných odchylek jsou způsobeny odlišným výpočtem vah. Při výpočtu odhadů a směrodatných odchylek na základě přístupu M se využívají váhy w_k, které jsou uvedeny v datech, ovšem neznáme přesný postup jejich výpočtu. U odhadů a směrodatných odchylek počítaných podle přístupu DP, respektive DPg, využíváme váhy g_k vypočítané podle vzorce (3.25). U těchto vah ovšem musíme pro uvažované dvojné kombinace proměnných využívat odhadnuté hodnoty jejich populačního úhrnu, neboť tyto údaje nelze z informací od společností získat. Na obrázku 5.1 můžeme vidět porovnání vah w_k a g_k pomocí histogramů. Můžeme si všimnout, že rozdíl je především způsobený tím, že váhy w_k používané společnostmi jsou z intervalu 0.3 – 3, jinak se histogramy příliš neliší.



(a) Váhy w_k, $k \in s$.

(b) Váhy g_k, $k \in s$.

Obrázek 5.1: Srovnání histogramů vah w_k, $k \in s$, které jsou použity při výpočtech odhadů a směrodatných odchylek na základě metod společností MEDIAN a STEM/MARK, a histogramů vah g_k, $k \in s$, které jsou použity při výpočtech na základě teorie kalibračních odhadů.

6. Simulace

V této kapitole pomocí simulačních studií zhodnotíme, jak jednotlivé asymptotické intervaly spolehlivosti populačního úhrnu poslechovosti rádií uvedené v předchozí kapitole 5 dodržují pravděpodobnost pokrytí. Budeme porovnávat pravděpodobnost pokrytí asymptotického intervalu spolehlivosti používaného společnostmi MEDIAN a STEM/MARK (5.2) a asymptotických intervalů spolehlivosti založených na teorii kalibračních odhadů (5.7) a (5.8).

Abychom mohli toto porovnání provést, potřebujeme znát skutečné hodnoty populačních úhrnů poslechovosti rádií v České republice. Tyto hodnoty jsou ovšem neznámé. Budeme tedy předpokládat, že se rovnají hodnotám, které následně vypočítáme na základě informací, které máme k dispozici. Nyní popíšeme, jak budeme postupovat pro dané rádio R.

Pro každého jedince z výběru s známe hodnotu jeho váhy vypočítanou společnostmi MEDIAN a STEM/MARK. Tyto váhy byly vypočítány tak, aby po převážení data odpovídala předepsaným četnostem v celé populaci U obyvatel České republiky. Tedy pokud w označuje součet vah všech jedinců ve výběru, tj. $w = \sum_{k \in s} w_k$, potom se lze na jednotlivé váhy w_k dívat tím způsobem, že jedinec k z výběru s zastupuje $\left(\frac{w_k}{w} \cdot N\right)$ jedinců z populace, kde N je velikost populace U. Vzhledem k tomu, že číslo $\left(\frac{w_k}{w} \cdot N\right)$ není přirozené, zaokrouhlíme ho na jednotky. Dostaneme tedy $\hat{N}_k$ jedinců z populace, které ve výběru zastupuje jedinec k. Navíc pro každého jedince k z výběru s známe jeho odpověď, zda včera rádio R poslouchal ($y_k = 1$), či nikoliv ($y_k = 0$), a také jeho vektor pomocných hodnot $\mathbf{x}_k$. V našem případě se jedná o vektor délky 433, který obsahuje pouze hodnoty 0 a 1, podle toho do jaké skupiny tento jedinec patří, vytvořený na základě devíti kategoriálních proměnných a vybraných dvojných kombinací těchto proměnných — viz neúplná post-stratifikace (sekce 4.2).

Na základě jedinců z výběru s o rozsahu $n = 15245$ získáme za pomoci logistického modelu odhad $\hat{p}_k$, $k \in s$, pravděpodobnosti, že z $\hat{N}_k$ jedinců z populace, které jedinec k ve výběru zastupuje, byl do výběru vybrán takový jedinec, který včera dané rádio R poslouchal. Pokud nezávisle vygenerujeme $\hat{N}_k$ hodnot z alternativního rozdělení s parametrem $\hat{p}_k$ a budeme je brát jako odpovědi $\hat{N}_k$ jedinců, zda včera poslouchali rádio R, potom je splněno, že z těchto $\hat{N}_k$ jedinců jsme do výběru s vybrali jedince, který včera rádio R poslouchal, s pravděpodobností $\hat{p}_k$. Tyto odpovědi budeme dále uvažovat jako odpovědi $\hat{N}_k$ jedinců ze skupiny, kterou ve výběru s zastupuje jedinec k, tj. jedinci v této skupině mají shodný vektor pomocných hodnot $\mathbf{x}_k$.

0. Vygenerování populace.

   Pro každého jedince k ve výběru s vygenerujeme odpovědi $\hat{N}_k$ jedinců ze skupiny, které tento jedinec zastupuje (viz výše). Součet odpovědí všech $\hat{N}_k$ jedinců využijeme k získání populačního úhrnu poslechovosti rádia R. Definujme

   $$Y_R = \sum_{k \in s} Y_k,$$

   kde $Y_k = \sum_{i=1}^{\hat{N}_k} Z_i$, $Z_i \sim \mathrm{Alt}(\hat{p}_k)$ a $\hat{N}_k$ je $\left(\frac{w_k}{w} \cdot N\right)$ zaokrouhleno na jednotky.

Budeme tedy předpokládat, že skutečný úhrn poslechovosti rádia R v České republice je roven Y_R.

Pro $b = 1, \ldots, 10000$ budeme provádět následující kroky.

1. Vytvoření výběru s_b.

   Pro každé $k = 1, \ldots, n$ vybereme ze skupiny, která je ve výběru s zastoupena jedincem k a obsahuje $\hat{N}_k$ jedinců z populace, náhodně jedince, kterého zařadíme do výběru s_b. Tímto získáme nový výběr s_b o rozsahu n.

2. Výpočet charakteristik na základě výběru s_b.

   Za pomoci výběru s_b následně odhadneme jednotlivé populační úhrny na základě metod společností MEDIAN a STEM/MARK (přístup M) a na základě teorie kalibračních odhadů (přístup DP, respektive DPg) a pro doplnění také nevážený populační úhrn poslechovosti rádia R.

Poznámka 13. Musíme brát ovšem ještě v potaz, že v kroku 0 kvůli zaokrouhlování čísel $\left(\frac{w_k}{w} \cdot N\right)$, $k \in s$, k výpočtu skutečného úhrnu Y_R využíváme populaci větší o 388 jedinců, než je velikost populace udávaná společnostmi MEDIAN a STEM/MARK.

△

Z dat získaných ze simulací provedených podle popisu výše následně vypočítáme průměrné odhady populačního úhrnu poslechovosti (nevážené (2.2), podle přístupu M (5.1) i DP (3.21)), dále průměrné délky a skutečné pravděpodobnosti pokrytí příslušných asymptotických intervalů spolehlivosti.

6.1 Výsledky simulací

V této části uvedeme skutečné pravděpodobnosti pokrytí asymptotických intervalů spolehlivosti s pravděpodobností pokrytí 0.95 pro pět nejposlouchanějších rádií v České republice. Pro větší přehlednost budeme pro odlišení jednotlivých přístupů používat stejné značení jako v části 5.3. Všechny výpočty byly provedeny ve statistickém softwaru R (R Core Team, 2016).

Jak můžeme vidět v tabulce 6.1, nevážené odhady P populačního úhrnu poslechovosti jsou u všech rádií vyšší než hodnota skutečného úhrnu. Dále vidíme, že průměrné odhady úhrnu poslechovosti pro jednotlivá rádia vypočítané podle přístupu M jsou až nápadně blízko hodnotám skutečného úhrnu, což je zapříčiněno tím, že jsme „skutečné“ hodnoty úhrnu poslechovosti vypočítali za pomoci vah w_k, $k \in s$, které tyto společnosti využívají při výpočty odhadů. Průměrné odhady úhrnu vypočítané na základě teorie kalibračních odhadů se již mírně liší od skutečných hodnot úhrnu. Výraznější rozdíly můžeme vidět u rádia Evropa 2 a u rádia ČRo Radiožurnál.

Nyní se zaměříme na průměrné délky příslušných asymptotických intervalů spolehlivosti (viz tabulka 6.2). Můžeme si všimnout, že délky intervalů vypočítané na základě přístupu DP, jsou výrazně menší než délky intervalů vypočítané na základě ostatních přístupů. S tím souvisí i skutečná pravděpodobnost pokrytí asymptotických intervalů spolehlivosti sestrojených na základě přístupu DP, která

Tabulka 6.1: Průměrná hodnota odhadů populačních úhrnů pěti nejposlouchanějších rádií v České republice uvedené v tisících obyvatel získané na základě simulací. Písmeno P značí průměrný neváženy úhrn, písmeno M značí průměrný úhrn podle metod společností MEDIAN a STEM/MARK, zkratka DP značí průměrný úhrn vypočítaný pomocí teorie kalibračních odhadů.

Rádio	Skutečný úhrn	Průměrný odhad úhrnu		
		P	M	DP
Rádio Impuls	955.3	957.5	955.1	957.4
Evropa 2	883.9	890.3	883.9	873.6
Frekvence 1	856.8	864.1	857.0	858.5
ČRo Radiožurnál	789.2	822.0	789.4	796.0
Rádio Blaník	626.9	614.3	626.8	626.4

se pohybuje v rozmezí 0.91 – 0.93. Nyní porovnáme zbývající dva přístupy, jak z pohledu průměrné délky asymptotických intervalů spolehlivosti, tak z pohledu skutečné pravděpodobnosti pokrytí těchto intervalů. Můžeme si všimnout, že nejvýraznější rozdíly jsou opět u rádií Evropa 2 a ČRo Radiožurnál. Z pohledu průměrné délky intervalu spolehlivosti je v případě rádia Evropa 2, respektive rádia ČRo Radiožurnál, tato průměrná délka větší o 3.2, respektive o 2.9, u intervalu sestrojeného podle přístupu M. Z pohledu skutečné pravděpodobnosti pokrytí je v případě rádia Evropa 2, respektive rádia ČRo Radiožurnál, skutečná pravděpodobnost pokrytí u intervalu sestrojeného podle přístupu DPg rovna 0.92, respektive 0.94, zatímco u intervalu sestrojeného podle přístupu M jsou tyto hodnoty rovny 0.95. U zbývajících rádií si můžeme všimnout, že skutečná pravděpodobnost pokrytí je vždy vyšší (a velmi blízko hodnotě 0.95) u intervalu sestrojeného podle přístupu DPg, zatímco průměrná délka intervalu je srovnatelná u obou přístupů.

Tabulka 6.2: Průměrné délky a skutečné pravděpodobnosti pokrytí asymptotických intervalů spolehlivosti pro populační úhrnů pěti nejposlouchanějších rádií v České republice získané na základě simulací. Písmeno M značí výpočet daných charakteristik podle metod společností MEDIAN a STEM/MARK, zkratka DP značí výpočet daných charakteristik pomocí teorie kalibračních odhadů. Výpočet charakteristik DPg je motivován poznámkou 5.

Rádio	Průměrná délka intervalu spolehlivosti			Skutečná pravděpodobnost pokrytí		
	M	DP	DPg	M	DP	DPg
Rádio Impuls	88.3	83.0	88.0	0.940	0.931	0.946
Evropa 2	85.5	78.6	82.3	0.949	0.905	0.919
Frekvence 1	84.4	80.0	84.5	0.939	0.933	0.947
ČRo Radiožurnál	82.5	77.0	79.6	0.952	0.928	0.938
Rádio Blaník	72.3	67.0	72.0	0.941	0.934	0.950

Ačkoliv výsledky získané na základě těchto simulací nejsou jednoznačné, můžeme si všimnout, že v případě rádií, kdy máme u přístupů M a DPg srovnatelné hodnoty průměrných odhadů úhrnu a srovnatelné průměrné délky příslušných asymptotických intervalů spolehlivosti, je přístup DPg k získání odhadu populačního úhrnu lepší. Bohužel toto nelze říct u přístupu DP bez využití vah g_k, $k \in s$.

6.2 Možná vylepšení simulací

V této části zhodnotíme výše uvedené simulace a pokusíme se navrhnout jejich možná vylepšení. Při rozhodování, jakým způsobem simulace provést, jsme naráželi na problém, že pokud chceme porovnat přístupy DP a DPg popsané v této diplomové práci s přístupem M společností MEDIAN a STEM/MARK, je třeba pracovat přesně s výběrem z datového souboru od těchto společností. V opačném případě bychom totiž nebyli schopni přesně dodržet metody z přístupu M, neboť společnosti neuvádějí přesný postup pro výpočet vah w_k. Jak již bylo zmíněné výše, při takto zvoleném přístupu neznáme skutečnou hodnotu populačního úhrnu. I přes veškerou snahu přiblížit se skutečné hodnotě populačního úhrnu může být dopočítaná „skutečná“ hodnota úhrnu daleko od skutečnosti. Toto může být důvodem nejednoznačných výsledků simulací.

Dalším důvodem, který mohl způsobit nejednoznačnost výsledků simulací, je způsob sběru dat společnostmi MEDIAN a STEM/MARK. Zásadním problémem je neochota některých náhodně vybraných obyvatel České republiky odpovídat na dotazník. Ovšem vyskytují se zde i další problémy, například výběr telefonních čísel. Z tohoto důvodu se nemusí jednat o prostý náhodný výběr, které tyto společnosti předpokládají.

Nyní se budeme zabývat možnými vylepšeními simulací v případě, že bychom byli schopni dopočítat váhy w_k, $k \in s$, používané v přístupu M pro libovolný výběr s. Jednou z možností by bylo považovat jedince z výběru s z datového souboru společností za celou populaci U a z této populace následně vybírat nezávisle výběry s menším rozsahem. Na základě těchto výběrů bychom následně mohli odhadovat populační úhrn, jehož skutečná hodnota by byla již známá, byla by rovna počtu jedinců z původního výběru, kteří poslouchali dané rádio.

Další variantou by mohlo být generování vlastní populace U. Na základě nezávislých výběrů s z této populace U bychom následně odhadli pomocí všech přístupů populační úhrn a následně porovnali tyto přístupy obdobně jako v části 6.1.

Závěr

V této práci jsme představili několik možností, jak odhadnout populační úhrn. Po zavedení základních pojmů a značení spojených s výběrovými šetřeními jsme v kapitole 2 v případě prostého náhodného výběru uvedli asymptotickou normalitu odhadu populačního úhrnu. Byla popsána varianta centrální limitní věty vhodná pro tento případ a také jsme se podrobně zabývali tím, za jakých předpokladů je splněna podmínka z této centrální limitní věty. K tomu jsme mimo jiné formulovali lemma o designové konzistenci výběrových průměrů.

Dále jsme již pro odhad populačního úhrnu využívali pomocné informace. Popsali jsme obecný regresní odhad populačního úhrnu, u kterého jsme největší pozornost věnovali jeho asymptotické normalitě. Byly uvedeny předpoklady, za kterých je tato normalita zaručena. Tuto část jsme následně aplikovali pro prostý náhodný výběr a navíc jsme se ještě zabývali limitním chováním vah g_k. Následně jsme představili kalibrační odhady populačního úhrnu, jak obecné odvození, tak některé konkrétní příklady těchto odhadů. Na závěr byly zmíněny předpoklady, za jakých je asymptotický rozptyl kalibračních odhadů shodný s asymptotickým rozptylem obecného regresního odhadu, který je speciálním případem kalibračního odhadu.

V následující části byly představeny teoretické závěry o kalibračních odhadech v kontingenčních tabulkách, ve kterých rozlišujeme, zda se jedná o úplnou, či neúplnou post-stratifikaci. Neúplná post-stratifikace bývá často využívána na konkrétních příkladech z praxe a stejně tomu bylo v této práci. Odvozené teoretické závěry jsme aplikovali na data z RADIOPROJEKTu, který se zabývá poslechovostí radií obyvatel České republiky. Výsledky založené na teorii představené v této práci jsme následně porovnali s výsledky založenými na metodě společností MEDIAN a STEM/MARK. Dospěli jsme k závěru, že výsledky z těchto přístupů se liší, což bylo částečně způsobeno tím, že jsme od společností neměli přesné informace o datech a z toho důvodu byly odhady založené na této práci počítány bez jedné kategoriální proměnné a také z odhadnutých teoretických četností u jednotlivých dvojných kombinací vybraných proměnných namísto skutečných teoretických četností. Porovnáním intervalů spolehlivosti založených na teorii uvedené v této práci jsme zjistili, že pro data z RADIOPROJEKTu vychází intervaly spolehlivosti využívající váhy g_k širší.

V poslední části práce jsme se pokusili porovnat přístupy uvedené v této práci a přístup používaný společnostmi MEDIAN a STEM/MARK. Bohužel kvůli neznáme hodnotě skutečného úhrnu jsme nedospěli k jednoznačným závěrům. Následně jsme navrhli možná vylepšení těchto simulací.

Seznam použité literatury

Deming, W. E. a 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**(4), 427–444.

Deville, J.-C. a Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**(418), 376–382.

Deville, J.-C., Särndal, C.-E. a Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, **88**(423), 1013–1020.

Jiang, J. (2010). *Large sample techniques for statistics*. Springer Science & Business Media, New York.

Kott, P. S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, **24**(3), 287–296.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Särdnal, C.-E., Swensson, B. a Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag.

Särndal, C.-E. (2010). The calibration approach in survey theory and practice. *Survey Methodology*, **33**(2), 99–119.

Särndal, C.-E., Swensson, B. a Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, **76**(3), 527–537.

Thompson, S. K. (2012). *Sampling*. Wiley, New York, Third Edition.

Vorlíčková, D. (1985). *Výběry z konečných souborů*. Univerzita Karlova, Praha.