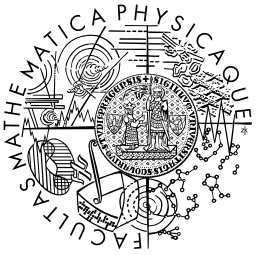


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

BAKALÁRSKA PRÁCA

Jakub Krett

Intervalové odhady pre pomery

Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: doc. Ing. Marek Omelka, Ph.D.
Študijný program: Matematika
Študijný odbor: Obecná matematika

Praha 2021

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov. Táto práca nebola využitá k získaniu iného alebo rovnakého titulu.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, najmä skutočnosť, že Univerzita Karlova má právo na uzavrenie licenčnej zmluvy o použití tejto práce ako školského diela podľa §60 odst. 1 autorského zákona.

V ..... dňa ....

Podpis autora

Na tomto mieste by som sa chcel poďakovať môjmu školiteľovi, doc. Ing. Marekovi Omelkovi, Ph.D., za ochotu, čas a odbornú pomoc. Za podporu počas štúdia ďakujem svojej rodine a priateľke.

Názov práce: Intervalové odhady pre pomery

Autor: Jakub Krett

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: doc. Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Táto práca sa venuje odvodeniu rôznych typov intervalových odhadov pre podiel stredných hodnôt. Zmyslom práce je aplikácia získaných teoretických poznatkov do problematiky triedenia odpadu, napríklad odhad hmotnosti nevytriedenej zložky odpadu vzhľadom k celkovej hmotnosti zmesového odpadu. Najprv sa v práci predstavia intervaly spoľahlivosti odvodené na základe štandardnej asymptotickej inferencie ako štandardný asymptotický interval spoľahlivosti a intervalový odhad odvodený s využitím logitovej transformácie. Ďalej sa vysvetlí metóda bootstrap, ktorá vedie k odvodeniu základného, percentilového a študentizovaného bootstrapového intervalu spoľahlivosti. V závere práce sa skúmajú vlastnosti uvedených intervalových odhadov pomocou dvoch simulačných modelov.

Kľúčové slová: asymptotický interval spoľahlivosti, logitová transformácia, bootstrap.

Title: Confidence intervals for ratios

Author: Jakub Krett

Department: Department of Probability and Mathematical Statistics

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

Abstract: This thesis is devoted to the derivation of various types of confidence intervals for ratios of mean values. The inspiration for this work is applying the acquired theoretical knowledge to the problem of waste sorting, such as estimating the weight of the unsorted waste components concerning the total weight of the mixture. Firstly, the confidence intervals based on the standard asymptotic inference are derived, such as the standard asymptotic confidence interval and the interval estimate derived using the logit transformation. Furthermore, the thesis introduces the bootstrap method, which leads to the derivation of the basic, percentile, and studentized bootstrap confidence interval. Finally, the end of the thesis explores the properties of these interval estimates using two simulation models.

Keywords: asymptotic confidence interval, logit transformation, bootstrap.

Obsah

Úvod

Separácia odpadu sa v posledných rokoch stáva bežnou súčasťou našich životov. Bohužiaľ, časť odpadu, ktorú je možné vytriediť, sa nachádza v košoch určených na zmesový odpad. Naskytuje sa prirodzená otázka na hmotnosť nevytriedeného odpadu vzhľadom k celkovej hmotnosti odpadu. V bakalárskej práci sa preto budeme venovať odvodeniu rôznych typov intervalových odhadov.

V prvej kapitole predstavíme štandardný asymptotický interval spoľahlivosti. Na jej konci uvedieme modifikáciu odhadu rozptylu, ktorá bude neskôr použitá pri simuláciách.

Následne v druhej kapitole odvodíme interval spoľahlivosti s využitím logitovej transformácie. Dôvodom zavedenia tejto transformácie je domnienka, že takto skonštruované intervalové odhady budú vykazovať lepšie výsledky ako štandardné intervaly z prvej kapitoly.

Tretia kapitola je venovaná neparametrickému bootstrapu, ktorý predstavuje alternatívu ku štandardnej asymptotickej inferencii. Po motivácii tejto metódy budú podrobne odvodené bootstrapové konfidenčné intervaly - základný, percentilový a študentizovaný. Celá kapitola je sprevádzaná ilustračným príkladom, ktorý slúži na lepšie pochopenie výpočtov bootstrapových intervalov spoľahlivosti.

Nakoniec budeme v simulačnej štúdii skúmať vlastnosti všetkých odvodených intervalových odhadov v dvoch rôznych modeloch.

1. Štandardný asymptotický interval spoľahlivosti

Majme náhodný výber (nezávislé, rovnako rozdelené náhodné vektory)

$$\begin{pmatrix} Y_1 \\ T_1 \end{pmatrix}, \begin{pmatrix} Y_2 \\ T_2 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ T_n \end{pmatrix},\ n \in \mathbb{N}.$$

Pre jednoduchosť značenia označme $\begin{pmatrix} Y \\ T \end{pmatrix}$ náhodný vektor, ktorý má rovnaké rozdelenie ako náhodný vektor $\begin{pmatrix} Y_i \\ T_i \end{pmatrix}$. Našim cieľom bude odvodiť bodový a intervalový odhad pre podiel stredných hodnôt

$$\theta = \frac{\mathsf{E}\, Y}{\mathsf{E}\, T}.$$

Aby mal výraz θ zmysel, budeme v práci predpokladať, že $\mathsf{E}\, T \neq 0$. Ako bodový odhad parametru θ použijeme podiel výberových priemerov

$$\hat{\theta}_n = \frac{\overline{Y}_n}{\overline{T}_n}.$$

Odhad $\hat{\theta}_n$ je konzistentný, pretože zo silného zákonu veľkých čísel (SZVČ) aplikovaného po zložkách (viď Dupač a Hušková, 2009, Věta 4.8.) platí

$$(\overline{Y}_n, \overline{T}_n)^\top \xrightarrow[n\to\infty]{s.j.} (\mathsf{E}\, Y, \mathsf{E}\, T)^\top.$$

Ďalej použijeme Vetu o spojitej transformácii (VoST) (viď van der Vaart, 2000, Theorem 2.3) pre spojitú funkciu $g(y,t) = y/t$ na $\mathbb{R} \times \mathbb{R} \setminus \{0\}$ a dostávame

$$\hat{\theta}_n = \frac{\overline{Y}_n}{\overline{T}_n} = g(\overline{Y}_n, \overline{T}_n) \xrightarrow[n\to\infty]{s.j.} g\left(\mathsf{E}\, Y, \mathsf{E}\, T\right) = \frac{\mathsf{E}\, Y}{\mathsf{E}\, T} = \theta, \tag{1.1}$$

kde sme využili predpoklad $\mathsf{E}\, T \neq 0$.

Náhodný výber môže v praxi slúžiť ako model pre triedenie odpadu. Nech Y_i značí hmotnosť nevytriedeného odpadu (papieru, skla, bioodpadu, ...) v i-tom komunálnom odpadkovom koši a T_i predstavuje celkovú hmotnosť odpadu v i-tom odpadkovom koši. Nasledujúca veta nám poskytuje informáciu o asymptotickom rozdelení $\hat{\theta}_n$. Predtým než ju sformulujeme, označme $\operatorname{var} Y = \sigma_Y^2$, $\operatorname{var} T = \sigma_T^2$ a $\operatorname{cov}(Y,T) = \operatorname{cov}(T,Y) = \sigma_{YT}$.

Veta 1. Nech $\begin{pmatrix} Y_1 \\ T_1 \end{pmatrix}, \begin{pmatrix} Y_2 \\ T_2 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ T_n \end{pmatrix}$ je náhodný výber s konečnou rozptylovou maticou Σ spĺňajúci $\mathsf{E}\, T \neq 0$. Potom platí

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{D} \mathsf{N}(0, \sigma^2), \tag{1.2}$$

kde $\sigma^2 = \frac{\sigma_Y^2}{(\mathsf{E}\, T)^2} - 2\frac{\sigma_{YT}\, \mathsf{E}\, Y}{(\mathsf{E}\, T)^3} + \frac{\sigma_T^2 (\mathsf{E}\, Y)^2}{(\mathsf{E}\, T)^4} \in (0, \infty)$.

Dôkaz. Z Centrálnej limitnej vety pre nezávislé, rovnako rozdelené vektory (CLV) (viď van der Vaart, 2000, Proposition 2.27) vyplýva

$$\sqrt{n}\left(\begin{pmatrix}\overline{Y}_n\\ \overline{T}_n\end{pmatrix} - \begin{pmatrix}\mathsf{E}\ Y\\ \mathsf{E}\ T\end{pmatrix}\right) \xrightarrow[n\to\infty]{D} \mathsf{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_Y^2 & \sigma_{YT}\\ \sigma_{YT} & \sigma_T^2\end{pmatrix}\right).$$

Ďalej označme $\mathbb{D}(y,t) = \left(\frac{\partial g(y,t)}{\partial y},\ \frac{\partial g(y,t)}{\partial t}\right)$.
Použitím Δ-metódy (viď van der Vaart, 2000, Theorem 3.1) pre funkciu

$$g(y,t) = y/t,\ g:\mathbb{R}^2 \to \mathbb{R},$$

ktorá je spojite diferencovateľná na intervale $\mathbb{R}\times\mathbb{R}\setminus\{0\}$, dostávame

$$\begin{aligned}&\mathbb{D}(y,t) = \left(1/t,\ \ -y/t^2\right) \implies \mathbb{D}(\mathsf{E}\ Y, \mathsf{E}\ T) = \left(1/\mathsf{E}\ T,\ \ -\mathsf{E}\ Y/(\mathsf{E}\ T)^2\right),\\ &\sqrt{n}\left(g\left(\begin{pmatrix}\overline{Y}_n\\ \overline{T}_n\end{pmatrix}\right) - g\left(\begin{pmatrix}\mathsf{E}\ Y\\ \mathsf{E}\ T\end{pmatrix}\right)\right) \xrightarrow[n\to\infty]{D} \mathsf{N}(0,\sigma^2),\end{aligned} \tag{1.3}$$

kde

$$\begin{aligned}\sigma^2 &= \mathbb{D}(\mathsf{E}\ Y, \mathsf{E}\ T)\,\Sigma\,\mathbb{D}(\mathsf{E}\ Y, \mathsf{E}\ T)^\top = \left(\tfrac{1}{\mathsf{E}\ T}, -\tfrac{\mathsf{E}\ Y}{(\mathsf{E}\ T)^2}\right)\begin{pmatrix}\sigma_Y^2 & \sigma_{YT}\\ \sigma_{YT} & \sigma_T^2\end{pmatrix}\left(\tfrac{1}{\mathsf{E}\ T}, -\tfrac{\mathsf{E}\ Y}{(\mathsf{E}\ T)^2}\right)^\top\\ &= \frac{\sigma_Y^2}{(\mathsf{E}\ T)^2} - 2\frac{\sigma_{YT}\,\mathsf{E}\ Y}{(\mathsf{E}\ T)^3} + \frac{\sigma_T^2(\mathsf{E}\ Y)^2}{(\mathsf{E}\ T)^4}.\end{aligned} \tag{1.4}$$

Po dosadení do (1.3) dostávame $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{D} \mathsf{N}(0,\sigma^2)$.

□

Poznámka. Rozptyl σ^2 z Vety 1 vieme vyjadriť ako

$$\sigma^2 = \frac{1}{(\mathsf{E}\ T)^2}\mathrm{var}\,(Y - \theta T). \tag{1.5}$$

To sa ukáže pomocou platnosti vzťahu

$$\mathsf{E}\,(Y - \theta T) = \mathsf{E}\ Y - \theta\,\mathsf{E}\ T = \mathsf{E}\ Y - \frac{\mathsf{E}\ Y}{\mathsf{E}\ T}\,\mathsf{E}\ T = 0,$$

ktorý použijeme v prvých dvoch rovnostiach nasledujúceho výpočtu. Počítajme

$$\begin{aligned}\sigma^2 &= \frac{1}{(\mathsf{E}\ T)^2}\left[\mathsf{E}\,(Y-\theta T)^2\right] = \frac{1}{(\mathsf{E}\ T)^2}\left[\mathsf{E}\,(Y - \mathsf{E}\ Y + \theta\,\mathsf{E}\ T - \theta T)^2\right]\\ &= \frac{1}{(\mathsf{E}\ T)^2}\left[\mathsf{E}\,(Y-\mathsf{E}\ Y)^2 + 2\theta\,\mathsf{E}\,[(Y-\mathsf{E}\ Y)(\mathsf{E}\ T - T)] + \theta^2\,\mathsf{E}\,(\mathsf{E}\ T - T)^2\right]\\ &= \frac{1}{(\mathsf{E}\ T)^2}\left[\sigma_Y^2 - 2\frac{\mathsf{E}\ Y}{\mathsf{E}\ T}\,\mathsf{E}\,[(Y-\mathsf{E}\ Y)(T-\mathsf{E}\ T)] + \frac{\sigma_T^2(\mathsf{E}\ Y)^2}{(\mathsf{E}\ T)^2}\right]\\ &= \frac{1}{(\mathsf{E}\ T)^2}\left[\sigma_Y^2 - 2\frac{\sigma_{YT}\,\mathsf{E}\ Y}{\mathsf{E}\ T} + \frac{\sigma_T^2(\mathsf{E}\ Y)^2}{(\mathsf{E}\ T)^2}\right].\end{aligned}$$

1.1 Odhad rozptylu

Aby sme vedeli skonštruovať interval spoľahlivosti, potrebujeme poznať (alebo odhadnúť) každý jeho člen. Preto musíme nájsť konzistentný odhad $\hat{\sigma}_n^2$ parametru σ^2 z (1.2). Nazvime veličinu

$$\hat{\sigma}_{YT} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)(T_i - \overline{T}_n)$$

výberová kovariancia a veličiny

$$\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2, \quad \hat{\sigma}_T^2 = \frac{1}{n-1}\sum_{i=1}^{n}(T_i - \overline{T}_n)^2$$

výberové rozptyly náhodných výberov Y a T.
Uvedomme si, že $\overline{Y}_n$, $\overline{T}_n$, $\hat{\sigma}_Y^2$, $\hat{\sigma}_T^2$, $\hat{\sigma}_{YT}$ sú odhady $\mathsf{E}\ Y$, $\mathsf{E}\ T$, σ_Y^2, σ_T^2, σ_{YT}.

Poznámka. Výberový priemer a výberový rozptyl sú konzistentné odhady pre príslušné parametre. To platí aj pre výberovú kovarianciu, pretože

$$\hat{\sigma}_{YT} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_iT_i - Y_i\overline{T}_n - \overline{Y}_nT_i + \overline{Y}_n\overline{T}_n) = \frac{1}{n-1}\sum_{i=1}^{n}Y_iT_i - \frac{n}{n-1}\overline{Y}_n\overline{T}_n.$$

Zo SZVČ máme $\overline{Y}_n \xrightarrow[n\to\infty]{s.j.} \mathsf{E}\ Y$, $\overline{T}_n \xrightarrow[n\to\infty]{s.j.} \mathsf{E}\ T$, $\frac{1}{n-1}\sum_{i=1}^{n}Y_iT_i \xrightarrow[n\to\infty]{s.j.} \mathsf{E}\ YT$.

Použitím VoST pre spojitú funkciu $g(x,y,z) = x - yz$ dostávame

$$\hat{\sigma}_{YT} = g\left(\frac{\sum_{i=1}^{n}Y_iT_i}{n-1}, \frac{n\overline{Y}_n}{n-1}, \overline{T}_n\right) \xrightarrow[n\to\infty]{s.j.} g(\mathsf{E}\ YT, \mathsf{E}\ Y, \mathsf{E}\ T) = \mathsf{E}\ YT - \mathsf{E}\ Y\ \mathsf{E}\ T = \operatorname{cov}(Y, T).$$

Po dosadení konzistentných odhadov z predošlej poznámky do (1.4) získame odhad

$$\hat{\sigma}_n^2 = \frac{\hat{\sigma}_Y^2}{(\overline{T}_n)^2} - 2\frac{\hat{\sigma}_{YT}\overline{Y}_n}{(\overline{T}_n)^3} + \frac{\hat{\sigma}_T^2(\overline{Y}_n)^2}{(\overline{T}_n)^4}$$

parametru σ^2, o ktorom chceme dokázať, že je taktiež konzistentný.
Preto použijeme VoST pre funkciu $g(a,b,c,d,e) = \frac{a}{b^2} - 2\frac{cd}{b^3} + \frac{ed^2}{b^4}$, ktorá je spojitá ak $b \neq 0$ a dostávame

$$\hat{\sigma}_n^2 = g(\hat{\sigma}_Y^2, \overline{T}_n, \hat{\sigma}_{YT}, \overline{Y}_n, \hat{\sigma}_T^2) \xrightarrow[n\to\infty]{s.j.} g(\sigma_Y^2, \mathsf{E}\ T, \sigma_{YT}, \mathsf{E}\ Y, \sigma_T^2) = \sigma^2. \tag{1.6}$$

Poznámka. Použitím podobných úprav ako pri odvodzovaní výrazu (1.5) vieme previesť odhad rozptylu na tvar

$$\hat{\sigma}_n^2 = \frac{1}{(\overline{T}_n)^2}\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \hat{\theta}_nT_i\right)^2. \tag{1.7}$$

V prvej rovnosti nasledujúceho výpočtu používame vzťah $\hat{\theta}_n \overline{T}_n - \overline{Y}_n = 0$.

$$\begin{aligned}\hat{\sigma}_n^2 &= \frac{1}{(\overline{T}_n)^2}\frac{1}{n-1}\sum_{i=1}^{n}\left[Y_i - \overline{Y}_n - \hat{\theta}_n(T_i - \overline{T}_n)\right]^2 \\ &= \frac{1}{(\overline{T}_n)^2}\frac{1}{n-1}\sum_{i=1}^{n}\left[(Y_i - \overline{Y}_n)^2 - 2\hat{\theta}_n(Y_i - \overline{Y}_n)(T_i - \overline{T}_n) + \hat{\theta}_n^2(T_i - \overline{T}_n)^2\right] \\ &= \frac{1}{(\overline{T}_n)^2}\left[\hat{\sigma}_Y^2 - 2\hat{\sigma}_{YT}\hat{\theta}_n + \hat{\sigma}_T^2\hat{\theta}_n^2\right] = \frac{1}{(\overline{T}_n)^2}\left[\hat{\sigma}_Y^2 - 2\frac{\hat{\sigma}_{YT}\overline{Y}_n}{\overline{T}_n} + \frac{\hat{\sigma}_T^2(\overline{Y}_n)^2}{(\overline{T}_n)^2}\right].\end{aligned}$$

1.2 Interval spoľahlivosti

V predošlej sekcii sme našli konzistentný odhad $\hat{\sigma}_n^2$ parametru σ^2, ktorý zohráva kľúčovú úlohu pri konštrukcii intervalu spoľahlivosti. Pripomeňme, že sme z Vety 1 ukázali platnosť vyjadrenia

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\sigma^2}} \xrightarrow[n\to\infty]{D} \mathsf{N}(0,1),$$

pričom predpokladajme, že $\sigma^2 > 0$. Keďže funkcia $g(x) = \sqrt{x}$ je spojitá na intervale $(0,\infty)$, obdržíme pomocou VoST konvergenciu $\sqrt{\hat{\sigma}_n^2} \xrightarrow[n\to\infty]{s.j.} \sqrt{\sigma^2}$. Použitím Cramér-Sluckého vety (viď Gut, 2005, Chapter 5, Theorem 11.4.) dostávame

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\hat{\sigma}_n^2}} \xrightarrow[n\to\infty]{D} \mathsf{N}(0,1),$$

čo vieme prepísať ako

$$\mathsf{P}\left(u_{\alpha/2} < \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\hat{\sigma}_n^2}} < u_{1-\alpha/2}\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

Využitím symetrie kvantilovej funkcie, t. j. $u_{\alpha/2} = -u_{1-\alpha/2}$, vieme jednoduchými úpravami odvodiť interval o asymptotickej spoľahlivosti $(1-\alpha)100\,\%$ pre podiel $\theta = \frac{\mathsf{E}\,Y}{\mathsf{E}\,T}$, ktorý má nasledujúci tvar

$$\left(\hat{\theta}_n - u_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}}, \hat{\theta}_n + u_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}}\right). \tag{1.8}$$

Výhodou intervalu spoľahlivosti (1.8) je jeho pomerne jednoduché odvodenie. Naopak, nevýhodou je skutočnosť, že môže obsahovať hodnoty mimo intervalu [0,1] a to v prípade, že platí $\hat{\theta}_n < \hat{\sigma}_n u_{1-\alpha/2}/\sqrt{n}$ alebo $\hat{\theta}_n > 1 - \hat{\sigma}_n u_{1-\alpha/2}/\sqrt{n}$. Túto nepríjemnosť vieme odstrániť pomocou prieniku konfidenčného intervalu (1.8) s množinou [0,1]. Ďalším problémom je, že odvodené výpočty fungujú iba asymptoticky, avšak rozdelenie náhodnej veličiny $\frac{\sqrt{n}(\hat{\theta}_n-\theta)}{\hat{\sigma}_n}$ môže byť vzdialené od rozdelenia $\mathsf{N}(0,1)$ pre malé rozsahy náhodného výberu. V nasledujúcich kapitolách predstavíme metódy, ktoré vykazujú lepšie výsledky pre malé rozsahy výberov. Viac v kapitole 4.

1.3 Modifikácia odhadu rozptylu

Vo Vete 1 sme ukázali, že

$$\operatorname{avar}(\hat{\theta}_n) = \sigma^2/n.$$

Teda odhad asymptotického rozptylu $\hat{\theta}_n$ je

$$\widehat{\operatorname{avar}}(\hat{\theta}_n) = \hat{\sigma}_n^2/n.$$

Teraz predpokladajme, že navyše platí

$$\mathsf{E}[Y_i \mid T_i] = \theta\, T_i, \;\; \operatorname{var}(Y_i \mid T_i) = \gamma^2\, T_i. \tag{1.9}$$

Z definície podmieneného rozptylu a podmienenej kovariancie máme vzťahy

$$\operatorname{var}(Y \mid T) = \mathsf{E}[Y^2 \mid T] - [\mathsf{E}(Y \mid T)]^2, \tag{1.10}$$
$$\operatorname{cov}[(X,Y) \mid T] = \mathsf{E}[XY \mid T] - \mathsf{E}[X \mid T]\,\mathsf{E}[Y \mid T]. \tag{1.11}$$

Sformulujme vetu, pomocou ktorej získame modifikáciu odhadu $\hat{\sigma}_n^2/n$.

Veta 2. Nech platia predpoklady (1.9) a označme $\boldsymbol{T}_n = (T_1, \ldots, T_n)^\top$. Potom platí

(*i*) $\operatorname{var}(\hat{\theta}_n \mid \boldsymbol{T}_n) = \frac{\gamma^2}{\sum_{i=1}^n T_i}$,

(*ii*) $\mathsf{E}\left[\frac{\hat{\sigma}_n^2}{n} \mid \boldsymbol{T}_n\right] = \frac{\gamma^2}{\sum_{i=1}^n T_i} \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^n T_i^2}{(\sum_{i=1}^n T_i)^2}\right]$.

Dôkaz. V rámci celého dôkazu používame vlastnosti podmienenej strednej hodnoty, ktoré sú uvedené vo Vete D1 a vzťahy (1.10) a (1.11) pre výpočet podmieneného rozptylu a podmienenej kovariancie.

Bod (*i*): V tretej rovnosti nasledujúceho výpočtu využívame nezávislosť náhodného výberu $\binom{Y_1}{T_1}, \binom{Y_2}{T_2}, \ldots, \binom{Y_n}{T_n}$ a v štvrtej rovnosti druhý predpoklad z (1.9).

$$\begin{aligned}\operatorname{var}(\hat{\theta}_n \mid \boldsymbol{T}_n) = \operatorname{var}\left(\frac{\overline{Y}_n}{\overline{T}_n} \mid \boldsymbol{T}_n\right) &= \frac{\operatorname{var}(\sum_{i=1}^n Y_i \mid \boldsymbol{T}_n)}{n^2(\overline{T}_n)^2} = \frac{\sum_{i=1}^n \operatorname{var}(Y_i \mid T_i)}{n^2(\overline{T}_n)^2} \\ &= \frac{\gamma^2 \sum_{i=1}^n T_i}{n^2(\overline{T}_n)^2} = \frac{\gamma^2}{n\overline{T}_n} = \frac{\gamma^2}{n\frac{1}{n}\sum_{i=1}^n T_i} = \frac{\gamma^2}{\sum_{i=1}^n T_i}.\end{aligned}$$

Bod (*ii*): S využitím rovnosti (1.7) počítajme:

$$\begin{aligned}\mathsf{E}\left[\frac{\hat{\sigma}_n^2}{n} \mid \boldsymbol{T}_n\right] &\overset{(1.7)}{=} \mathsf{E}\left[\frac{1}{(\overline{T}_n)^2}\frac{1}{(n-1)n}\sum_{i=1}^n \left(Y_i - \theta T_i + \theta T_i - \hat{\theta}_n T_i\right)^2 \mid \boldsymbol{T}_n\right] \\ &= \frac{\sum_{i=1}^n \mathsf{E}[(Y_i - \theta T_i)^2 \mid \boldsymbol{T}_n]}{(\overline{T}_n)^2(n-1)n} + \frac{2\sum_{i=1}^n \mathsf{E}\left[(Y_i - \theta T_i)(\theta T_i - \hat{\theta}_n T_i) \mid \boldsymbol{T}_n\right]}{(\overline{T}_n)^2(n-1)n} \\ &+ \frac{\sum_{i=1}^n \mathsf{E}\left[(\theta T_i - \hat{\theta}_n T_i)^2 \mid \boldsymbol{T}_n\right]}{(\overline{T}_n)^2(n-1)n} = A_n + B_n + C_n,\end{aligned} \tag{1.12}$$

kde A_n, B_n a C_n sme postupne označili jednotlivé členy na pravej strane (1.12). Najprv začneme upravovať člen A_n.

$$A_n = \frac{\sum_{i=1}^n \mathsf{E}\left[(Y_i - \theta T_i)^2 \mid \boldsymbol{T}_n\right]}{(\overline{T}_n)^2(n-1)n} \overset{(1.10)}{=} \frac{\sum_{i=1}^n \operatorname{var}\left[(Y_i - \theta T_i) \mid \boldsymbol{T}_n\right]}{(\overline{T}_n)^2(n-1)n} = \frac{\sum_{i=1}^n \operatorname{var}\left[Y_i \mid T_i\right]}{(\overline{T}_n)^2(n-1)n}$$
$$\overset{(1.9)}{=} \frac{\gamma^2 \sum_{i=1}^n T_i}{(\overline{T}_n)^2(n-1)n} = \frac{\gamma^2}{\overline{T}_n(n-1)} = \frac{\gamma^2 n}{\sum_{i=1}^n T_i(n-1)}. \tag{1.13}$$

Teraz pre prehľadnosť upravíme nasledujúce výrazy, ktoré využijeme vo výpočte členu B_n.

$$\begin{aligned}
\mathsf{E}\left[Y_i\hat{\theta}_n \mid \boldsymbol{T}_n\right] &= \frac{\mathsf{E}\left[Y_i \sum_{j=1}^n Y_j \mid \boldsymbol{T}_n\right]}{n\overline{T}_n} = \frac{\mathsf{E}\left[Y_i^2 \mid T_i\right]}{n\overline{T}_n} + \frac{\sum_{j=1, j\neq i}^n \mathsf{E}\left[Y_i Y_j \mid T_i, T_j\right]}{n\overline{T}_n} \\
&= \frac{\left[\operatorname{var}\left(Y_i \mid T_i\right) + (\mathsf{E}\left[Y_i \mid T_i\right])^2\right]}{n\overline{T}_n} + \frac{\sum_{j=1, j\neq i}^n \mathsf{E}\left[Y_i \mid T_i\right] \mathsf{E}\left[Y_j \mid T_j\right]}{n\overline{T}_n} \\
&= \frac{(\gamma^2 T_i + \theta^2 T_i^2)}{n\overline{T}_n} + \frac{\theta^2 T_i(\sum_{j=1}^n T_j - T_i)}{n\overline{T}_n} \\
&= \frac{\gamma^2 T_i}{n\overline{T}_n} + \frac{\theta^2 T_i^2}{n\overline{T}_n} + \frac{\theta^2 T_i \sum_{j=1}^n T_j}{n\overline{T}_n} - \frac{\theta^2 T_i^2}{n\overline{T}_n} = \frac{\gamma^2 T_i}{n\overline{T}_n} + \theta^2 T_i.
\end{aligned} \tag{1.14}$$

$$\mathsf{E}\left[\hat{\theta}_n \mid \boldsymbol{T}_n\right] = \frac{\mathsf{E}\left[\sum_{i=1}^n Y_i \mid T_i\right]}{n\overline{T}_n} = \frac{\theta \sum_{i=1}^n T_i}{n\overline{T}_n} = \theta. \tag{1.15}$$

S využitím (1.14) a (1.15) vo štvrtej rovnosti počítajme B_n.

$$\begin{aligned}
B_n &= \frac{2\sum_{i=1}^n T_i \operatorname{cov}\left[(Y_i - \theta T_i,\, \theta - \hat{\theta}_n) \mid \boldsymbol{T}_n\right]}{(\overline{T}_n)^2(n-1)n} = -\frac{2\sum_{i=1}^n T_i \operatorname{cov}\left[(Y_i, \hat{\theta}_n) \mid \boldsymbol{T}_n\right]}{(\overline{T}_n)^2(n-1)n} \\
&\overset{(1.11)}{=} -\frac{2\sum_{i=1}^n T_i \mathsf{E}\left[Y_i\hat{\theta}_n \mid \boldsymbol{T}_n\right]}{(\overline{T}_n)^2(n-1)n} + \frac{2\sum_{i=1}^n T_i \mathsf{E}\left[Y_i \mid T_i\right] \mathsf{E}\left[\hat{\theta}_n \mid \boldsymbol{T}_n\right]}{(\overline{T}_n)^2(n-1)n} \\
&= -\frac{2\sum_{i=1}^n T_i(\frac{\gamma^2 T_i}{n\overline{T}_n} + \theta^2 T_i)}{(\overline{T}_n)^2(n-1)n} + \frac{2\sum_{i=1}^n T_i \mathsf{E}\left[Y_i \mid T_i\right]\theta}{(\overline{T}_n)^2(n-1)n} \\
&= -\frac{2\gamma^2 \sum_{i=1}^n T_i^2}{(\overline{T}_n)^3(n-1)n^2} - \frac{2\theta^2 \sum_{i=1}^n T_i^2}{(\overline{T}_n)^2(n-1)n} + \frac{2\theta^2 \sum_{i=1}^n T_i^2}{(\overline{T}_n)^2(n-1)n} \\
&= -\frac{2\gamma^2 n \sum_{i=1}^n T_i^2}{(n-1)(\sum_{i=1}^n T_i)^3}.
\end{aligned} \tag{1.16}$$

Nakoniec ostáva spočítať C_n.

$$\begin{aligned}
C_n &\overset{(1.10)}{=} \frac{\sum_{i=1}^n T_i^2 \operatorname{var}\left[\hat{\theta}_n \mid \boldsymbol{T}_n\right]}{(\overline{T}_n)^2(n-1)n} + \frac{\sum_{i=1}^n T_i^2 \left(\mathsf{E}\left[(\theta - \hat{\theta}_n) \mid \boldsymbol{T}_n\right]\right)^2}{(\overline{T}_n)^2(n-1)n} \\
&\overset{(i)}{=} \frac{\gamma^2 \sum_{i=1}^n T_i^2}{(\overline{T}_n)^3(n-1)n^2} + \frac{\sum_{i=1}^n T_i^2 \left(\theta - \frac{\theta \sum_{j=1}^n T_j}{n\overline{T}_n}\right)^2}{(\overline{T}_n)^2(n-1)n} = \frac{\gamma^2 n \sum_{i=1}^n T_i^2}{(n-1)(\sum_{i=1}^n T_i)^3}.
\end{aligned} \tag{1.17}$$

Na záver s využitím (1.12), (1.13), (1.16) a (1.17) dostávame

$$\mathsf{E}\left[\frac{\hat{\sigma}_n^2}{n} \mid \boldsymbol{T}_n\right] = A_n + B_n + C_n = \frac{\gamma^2 n}{\sum_{i=1}^n T_i(n-1)} - \frac{\gamma^2 n \sum_{i=1}^n T_i^2}{(n-1)(\sum_{i=1}^n T_i)^3}$$
$$= \frac{\gamma^2}{\sum_{i=1}^n T_i} \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^n T_i^2}{(\sum_{i=1}^n T_i)^2}\right],$$

čo sme chceli dokázať.

□

Definujme

$$\tilde{\sigma}_n^2 = \hat{\sigma}_n^2 \frac{n-1}{n} \frac{1}{\left[1 - \frac{\sum_{i=1}^n T_i^2}{(\sum_{i=1}^n T_i)^2}\right]}. \tag{1.18}$$

Potom je $\tilde{\sigma}_n^2/n$ podľa vety 2 nestranný odhad $\mathrm{var}\,(\hat{\theta}_n \mid \boldsymbol{T}_n)$ v modeli (1.9).

Poznámka. Ukážeme platnosť vzťahu $\tilde{\sigma}_n^2 \geq \hat{\sigma}_n^2$ pre $n > 1, T_i > 0\ \forall i \in \{1, \ldots, n\}$. Nech $x_i > 0\ \forall i \in \{1, \ldots, n\}$. Označme $\boldsymbol{x}_n = (x_1, \ldots, x_n)^\top$ a

$$f(\boldsymbol{x}_n) = \frac{1}{\left[1 - \frac{\sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i)^2}\right]}.$$

Potom pomocou Lagrangeovej vety o multiplikátoroch (viď Pick a kol., 2020, 11.5.8. Věta) ukážeme, že funkcia $f(\boldsymbol{x}_n)$ dosahuje globálne minimum v bode $\boldsymbol{x}_n = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)^\top$ a nadobúda v ňom hodnotu $n/(n-1)$. Pre jednoduchosť výpočtov stačí hľadať globálne minimum funkcie

$$f_1(\boldsymbol{x}_n) = \frac{\sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i)^2}.$$

Definujme bez ujmy na všeobecnosti väzobnú podmienku

$$g(\boldsymbol{x}_n) = \sum_{i=1}^n x_i - 1 = 0,$$

pretože inak

$$f_1(\boldsymbol{x}_n) = \sum_{i=1}^n (x_i^*)^2,$$

kde $x_i^* = x_i / \sum_{j=1}^n x_j$. Tým dostávame Lagrangeovu funkciu pre $\lambda \in \mathbb{R}$

$$\mathcal{L}(\boldsymbol{x}_n, \lambda) = \sum_{i=1}^n x_i^2 - \lambda\left(\sum_{i=1}^n x_i - 1\right).$$

Derivovaním by sa zistilo, že funkcia $f_1(\boldsymbol{x}_n)$ nadobúda globálne minimum v bode $\boldsymbol{x}_n = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)^\top$ a nadobúda v ňom hodnotu $1/n$. Potom funkcia $f(\boldsymbol{x}_n)$ nadobúda v rovnakom bode globálne minimum veľkosti $n/(n-1)$. S využitím (1.18) počítajme

$$\tilde{\sigma}_n^2 = \hat{\sigma}_n^2 \frac{n-1}{n} f(\boldsymbol{T}_n) \geq \hat{\sigma}_n^2 \frac{n-1}{n} \frac{n}{n-1} = \hat{\sigma}_n^2,$$

pričom rovnosť nastáva pre $T_1 = T_2 = \cdots = T_n$, čo v kontexte aplikácie práce znamená, že všetky pozorované odpadkové koše majú rovnakú hmotnosť.

Poznámka. Použitím modifikovaného rozptylu (1.18) vieme upraviť asymptotický intervalový odhad (1.8) pre parameter θ na tvar

$$\left(\hat{\theta}_n - u_{1-\alpha/2}\sqrt{\frac{\tilde{\sigma}_n^2}{n}}, \hat{\theta}_n + u_{1-\alpha/2}\sqrt{\frac{\tilde{\sigma}_n^2}{n}}\right). \tag{1.19}$$

V 4. kapitole budeme skúmať, či je takto modifikovaný štandardný interval spoľahlivosti (1.19) lepší ako (1.8).

2. Logitová transformácia

V tejto kapitole predstavíme logitovú transformáciu, ktorá vykazuje lepšie výsledky (viď kapitola 4) ako štandardný asymptotický interval spoľahlivosti (1.8) na str. 6. Nech platí

$$T_i > 0 \text{ a } 0 \leq Y_i \leq T_i,\ i = 1, \ldots, n.$$

Tieto predpoklady sú v súlade s aplikáciou triedenia odpadu, pretože chceme pracovať so zmesovým odpadom kladnej hmotnosti a zároveň s nezápornou hmotnosťou nevytriedenej zložky odpadu, ktorej je nanajvýš rovnako ako celkového odpadu. Ďalej nazveme funkciu

$$\text{logit}(x) = \log \frac{x}{1-x}, x \in (0,1)$$

logitová funkcia premennej x. Táto funkcia zobrazuje hodnoty z intervalu (0,1) do $\mathbb{R}$. Inverzná funkcia k logitovej sa nazýva expitová funkcia a má tvar

$$\text{expit}(x) = \text{logit}^{-1}(x) = \frac{e^x}{1+e^x}, x \in \mathbb{R}.$$

Expitová funkcia zobrazuje ľubovoľný bod reálnej osi do intervalu (0,1).

2.1 Odhad rozptylu

Označme $\lambda = \text{logit}(\theta)$ a $\hat{\lambda}_n = \text{logit}(\hat{\theta}_n)$ logitovú transformáciu parametru θ, respektíve odhadu $\hat{\theta}_n$. V prípade, že platí $Y_i = 0\ \forall i$ alebo $Y_i = T_i\ \forall i$, tak funkcia logit nie je definovaná. Zároveň však nie je v tejto situácii zaujímavé počítať žiaden interval spoľahlivosti, preto nebudeme takto degenerované prípady uvažovať.

Asymptotický rozptyl náhodnej veličiny $\hat{\lambda}_n$ odvodíme pomocou Δ-metódy (van der Vaart, 2000, Theorem 3.1). Funkcia $g(t) = \text{logit}(t)$ je spojitá pre $t \in (0,1)$. Derivujme ako rozdiel logaritmov

$$g'(t) = \frac{1}{t} + \frac{1}{1-t} = \frac{1}{t(1-t)}.$$

Použitím (1.2) z Vety 1 na str. 3 a Δ-metódy môžeme napísať

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow[n\to\infty]{D} \mathsf{N}(0, \nu^2),$$

kde $\nu^2 = \frac{\sigma^2}{\theta^2(1-\theta)^2}$. Asymptotický rozptyl ν^2 odhadneme ako

$$\hat{\nu}_n^2 = \frac{\hat{\sigma}_n^2}{\hat{\theta}_n^2(1-\hat{\theta}_n)^2}.$$

Tento odhad je konzistentný, pretože spojením (1.1) a (1.6) platí konvergencia

$$(\hat{\sigma}_n^2, \hat{\theta}_n)^\top \xrightarrow[n\to\infty]{s.j.} (\sigma^2, \theta)^\top$$

a následne použijeme VoST pre funkciu $g(x,y) = \frac{x}{y^2(1-y)^2}$, ktorá je spojitá na $\mathbb{R} \times \mathbb{R} \setminus \{0,1\}$ a dostávame

$$\hat{\nu}_n^2 = g(\hat{\sigma}_n^2, \hat{\theta}_n) \xrightarrow[n\to\infty]{s.j.} g(\sigma^2, \theta) = \nu^2. \tag{2.1}$$

2.2 Interval spoľahlivosti

V tejto časti sa budeme venovať odvodeniu intervalu spoľahlivosti pomocou logitovej transformácie. Budeme postupovať v dvoch krokoch. Najprv odvodíme konfidenčný interval pre parameter λ a následne intervalový odhad pre parameter θ pomocou vhodne zvolenej transformácie. V sekcii 2.1 sme ukázali platnosť vzťahu

$$\frac{\sqrt{n}(\hat{\lambda}_n - \lambda)}{\sqrt{\nu^2}} \xrightarrow[n\to\infty]{D} \mathsf{N}(0,1)\,.$$

Využitím (2.1), VoST a Cramér-Sluckého vety dostávame

$$\frac{\sqrt{n}(\hat{\lambda}_n - \lambda)}{\sqrt{\hat{\nu}_n^2}} \xrightarrow[n\to\infty]{D} \mathsf{N}(0,1)\,.$$

Jednoduchými úpravami obdržíme asymptotický intervalový odhad o spoľahlivosti $(1-\alpha)100\,\%$ pre parameter λ

$$\left(\hat{\lambda}_n - u_{1-\alpha/2}\sqrt{\frac{\hat{\nu}_n^2}{n}}, \hat{\lambda}_n + u_{1-\alpha/2}\sqrt{\frac{\hat{\nu}_n^2}{n}}\right). \tag{2.2}$$

V našom záujme je ale odvodiť interval spoľahlivosti pre parameter θ. Ako transformačnú funkciu použijeme $g(t) = \operatorname{expit}(t)$, ktorá je monotónna na $\mathbb{R}$. Aplikovaním funkcie expit na interval (2.2) získame asymptotický intervalový odhad pre parameter θ o spoľahlivosti $(1-\alpha)100\,\%$

$$\left(\frac{\exp\left(\hat{\lambda}_n - u_{1-\alpha/2}\sqrt{\frac{\hat{\nu}_n^2}{n}}\right)}{1+\exp\left(\hat{\lambda}_n - u_{1-\alpha/2}\sqrt{\frac{\hat{\nu}_n^2}{n}}\right)}, \frac{\exp\left(\hat{\lambda}_n + u_{1-\alpha/2}\sqrt{\frac{\hat{\nu}_n^2}{n}}\right)}{1+\exp\left(\hat{\lambda}_n + u_{1-\alpha/2}\sqrt{\frac{\hat{\nu}_n^2}{n}}\right)}\right). \tag{2.3}$$

Poznámka. Použitím modifikácie rozptylu (1.18) označme $\tilde{\nu}_n^2 = \frac{\tilde{\sigma}_n^2}{\hat{\theta}_n^2(1-\hat{\theta}_n)^2}$. Potom vieme upraviť logitový intervalový odhad pre parameter θ na tvar

$$\left(\frac{\exp\left(\hat{\lambda}_n - u_{1-\alpha/2}\sqrt{\frac{\tilde{\nu}_n^2}{n}}\right)}{1+\exp\left(\hat{\lambda}_n - u_{1-\alpha/2}\sqrt{\frac{\tilde{\nu}_n^2}{n}}\right)}, \frac{\exp\left(\hat{\lambda}_n + u_{1-\alpha/2}\sqrt{\frac{\tilde{\nu}_n^2}{n}}\right)}{1+\exp\left(\hat{\lambda}_n + u_{1-\alpha/2}\sqrt{\frac{\tilde{\nu}_n^2}{n}}\right)}\right), \tag{2.4}$$

ktorý budeme využívať v 4. kapitole.

3. Bootstrap

3.1 Motivácia

Neparametrický bootstrap je štatistická metóda, ktorú v roku 1979 publikoval prof. Bradley Efron v práci Efron (1979). Principiálne ide o jednoduchú metódu, ktorú však nejde zrealizovať bez použitia počítača. Hlavná myšlienka spočíva v prevedení mnohých náhodných výberov z pozorovaných dát.

Označme F distribučnú funkciu náhodného vektoru $\binom{Y}{T}$ a pripomeňme, že $\theta = \frac{\mathsf{E}\, Y}{\mathsf{E}\, T}$. Ako bodový odhad neznámeho parametru θ môžeme stále použiť podiel výberových priemerov $\hat{\theta}_n = \frac{\overline{Y}_n}{\overline{T}_n}$. Avšak, keďže nepoznáme distribučnú funkciu F, nepoznáme ani rozdelenie náhodnej veličiny

$$\sqrt{n}(\hat{\theta}_n - \theta)$$

a teda nevieme skonštruovať interval spoľahlivosti pre parameter θ.

3.2 Princíp bootstrapu

Majme náhodný výber $\binom{Y_1}{T_1}, \binom{Y_2}{T_2}, \ldots, \binom{Y_n}{T_n}$, $n \in \mathbb{N}$, z neznámeho rozdelenia F a spočítajme z tohto výberu potrebné odhady. Pre naše účely budeme potrebovať podiel výberových priemerov, značíme $\hat{\theta}_n$, ako konzistentný odhad podielu stredných hodnôt, značíme θ. Následne uvažujme prostý náhodný výber

$$\binom{Y_1^*}{T_1^*}, \binom{Y_2^*}{T_2^*}, \ldots, \binom{Y_n^*}{T_n^*}$$

z realizácie dát $\binom{Y_1}{T_1}, \binom{Y_2}{T_2}, \ldots, \binom{Y_n}{T_n}$ s **vracaním** veľkosti n. Hodnoty realizácie dát sa môžu v novom výbere opakovať. Tento krok zopakujeme B-krát a tým získame B bootstrapových výberov. Formálnejšie, obdržíme

$$\binom{Y_{1,b}^*}{T_{1,b}^*}, \binom{Y_{2,b}^*}{T_{2,b}^*}, \ldots, \binom{Y_{n,b}^*}{T_{n,b}^*}, \; b = 1, \ldots, B. \tag{3.1}$$

Ďalej spočítajme odhady

$$\hat{\theta}_{n,b}^* = \frac{\overline{Y}_{n,b}^*}{\overline{T}_{n,b}^*},$$

kde $\overline{Y}_{n,b}^* = \frac{1}{n}\sum_{i=1}^n Y_{i,b}^*$ a $\overline{T}_{n,b}^* = \frac{1}{n}\sum_{i=1}^n T_{i,b}^*$. Celkovo tak získame B odhadov podielu stredných hodnôt

$$\hat{\theta}_{n,1}^*, \ldots, \hat{\theta}_{n,B}^*.$$

Zvyšovaním B získame spoľahlivejšie odhady bootstrapových kvantilov, ale dáta s ktorými pracujeme ostávajú stále rovnaké, pretože zakaždým generujeme výbery z počiatočného náhodného výberu. Pre 95% intervalové odhady zvyčajne volíme $B = 999$ viď Davison a Hinkley (1997), kapitola 5.2.

Poznámka. Veľkosť každého z B výberov musí byť rovná veľkosti počiatočného náhodného výberu, v našom prípade n. Dôvodom je fakt, že rozdelenie náhodnej veličiny $\sqrt{n}(\hat{\theta}_n - \theta)$ výberu $\binom{Y_1}{T_1}, \binom{Y_2}{T_2}, \ldots, \binom{Y_n}{T_n}$ závisí na jeho veľkosti. Preto, ak ju chceme aproximovať, potrebujeme konštruovať výbery rovnakej veľkosti.

3.3 Bootstrapové intervaly spoľahlivosti

V ďalších sekciách sa budeme venovať trom alternatívam bootstrapových intervalov spoľahlivosti. Informácie sú čerpané z knihy Davison a Hinkley (1997), kapitoly 5.1–5.3 a zo skript Omelka (2021), kapitola 8.3.

Uvažujme funkciu $g(y,t) = y/t$, ktorá je spojite diferencovateľná pre $t \neq 0$, čo nám zaručuje funkčnosť bootstrapu (viď Omelka, 2021, Theorem 15). Uvažujme neznáme rozdelenie

$$R_n = \sqrt{n}(\hat{\theta}_n - \theta).$$

Z vety 1 vieme, že

$$R_n \xrightarrow[n\to\infty]{D} R, \tag{3.2}$$

kde R je náhodná veličina so spojitou distribučnou funkciou. Nech $r_n(\alpha)$ a $r(\alpha)$ značia α-kvantil rozdelenia R_n, respektíve R. Potom z van der Vaart (2000), Lemma 21.2 platí

$$r_n(\alpha) \xrightarrow[n\to\infty]{} r(\alpha). \tag{3.3}$$

V našom záujme bude odvodenie intervalu spoľahlivosti pre parameter θ.

Ak by sme poznali rozdelenie R_n, vedeli by sme určiť aj jeho kvantily $r_n(\alpha/2)$ a $r_n(1-\alpha/2)$. Potom by sme dostali vyjadrenie

$$\begin{aligned}
&\mathsf{P}\left[r_n(\alpha/2) < \sqrt{n}(\hat{\theta}_n - \theta) < r_n(1-\alpha/2)\right] = 1-\alpha \text{ a ekvivalentne} \\
&\mathsf{P}\left[\hat{\theta}_n - \frac{r_n(1-\alpha/2)}{\sqrt{n}} < \theta < \hat{\theta}_n - \frac{r_n(\alpha/2)}{\sqrt{n}}\right] = 1-\alpha,
\end{aligned}$$

z ktorého vieme odvodiť $(1-\alpha)100\%$ interval spoľahlivosti pre parameter θ

$$\left(\hat{\theta}_n - \frac{r_n(1-\alpha/2)}{\sqrt{n}},\, \hat{\theta}_n - \frac{r_n(\alpha/2)}{\sqrt{n}}\right).$$

Problém je, že rozdelenie R_n nepoznáme, a teda nevieme spočítať interval spoľahlivosti.

3.3.1 Základný bootstrapový interval spoľahlivosti

Označme

$$R^*_{n,b} = \sqrt{n}\left(\hat{\theta}^*_{n,b} - \hat{\theta}_n\right),\ b = 1, \ldots, B$$

a nech $r_n^*(\alpha)$ značí α-kvantil rozdelenia náhodnej veličiny $R_{n,b}^*$. Potom z (3.2) a z Omelka (2021, Theorem 13 a Theorem 15) platí

$$r_n^*(\alpha) \xrightarrow[n\to\infty]{P} r(\alpha). \tag{3.4}$$

Rozdielom (3.3) a (3.4) dostávame konvergenciu

$$r_n(\alpha) - r_n^*(\alpha) \xrightarrow[n\to\infty]{P} r(\alpha) - r(\alpha) = 0,$$

z ktorej môžeme usúdiť, že bootstrapové kvantily $r_n^*(\alpha)$ sú limitne „dobrou“ aproximáciou teoretických kvantilov $r_n(\alpha)$. Ďalej máme

$$\mathsf{P}\left[r_n^*(\alpha/2) < \sqrt{n}(\hat{\theta}_n - \theta) < r_n^*(1-\alpha/2)\right] \xrightarrow[n\to\infty]{} 1-\alpha.$$

Z tohto tvaru vieme odvodiť asymptotický interval spoľahlivosti pre parameter θ

$$\left(\hat{\theta}_n - \frac{r_{n,B}^*(1-\alpha/2)}{\sqrt{n}}, \hat{\theta}_n - \frac{r_{n,B}^*(\alpha/2)}{\sqrt{n}}\right), \tag{3.5}$$

kde $r_{n,B}^*(\alpha)$ je odhad α-kvantilu $r_n^*(\alpha)$ spočítaný ako výberový α-kvantil z hodnôt $R_{n,1}^*, \ldots, R_{n,B}^*$.

Poznámka. Interval spoľahlivosti (3.5) vieme ekvivalentne upraviť na tvar

$$\left(2\,\hat{\theta}_n - q_{n,B}^*(1-\alpha/2), 2\,\hat{\theta}_n - q_{n,B}^*(\alpha/2)\right),$$

pričom sme využili platnosť rovnosti

$$r_{n,B}^*(\alpha) = \sqrt{n}(q_{n,B}^*(\alpha) - \hat{\theta}_n), \tag{3.6}$$

kde $q_{n,B}^*(\alpha)$ je výberový α-kvantil spočítaný z hodnôt $\hat{\theta}_{n,1}^*, \ldots, \hat{\theta}_{n,B}^*$.

Príklad. Ilustrujme na príklade výpočet intervalových odhadov s pravdepodobnosťou pokrytia 95 %, t. j. $\alpha = 0{,}05$. Začneme so základným bootstrapovým intervalom spoľahlivosti. Uvažujme 10 pozorovaní

$$\binom{68}{72}, \binom{45}{52}, \binom{42}{96}, \binom{17}{73}, \binom{60}{77}, \binom{55}{55}, \binom{57}{80}, \binom{91}{97}, \binom{16}{80}, \binom{5}{46}.$$

Podiel výberových priemerov dát je $\hat{\theta}_n = 0{,}626$. Pre konštrukciu intervalu spoľahlivosti potrebujeme vedieť, ako veľmi sa $\hat{\theta}_n$ líši od θ. Matematicky, zaujíma nás rozdelenie

$$R_n = \sqrt{n}(\hat{\theta}_n - \theta).$$

Pomôžeme si bootstrapom, na základe ktorého vieme aproximovať neznáme rozdelenie R_n rozdelením náhodnej veličiny

$$R_{n,b}^* = \sqrt{n}\left(\hat{\theta}_{n,b}^* - \hat{\theta}_n\right),\ b = 1, \ldots, B,$$

kde $\hat{\theta}^*_{n,\,b}$ predstavuje podiel výberových priemerov b-tého bootstrapového výberu. Ďalej sme v softvéri R vygenerovali $B = 999$ bootstrapových výberov z pôvodného výberu, každý veľkosti $n = 10$. Zobrazenie týchto výberov v samostatnej tabuľke by bolo neprehľadné, z toho dôvodu spočítame hodnoty $R^*_{n,\,b}$, $b = 1, \ldots, B$ a zobrazíme ich v histograme na obrázku 3.1. Ďalej zoradíme spočítané hodnoty vzostupne. Dostávame postupnosť 999 hodnôt, z ktorých pre prehľadnosť vypíšeme iba prvé a posledné:

$$-0{,}921, \quad -0{,}862, \quad -0{,}808, \quad \ldots, \quad 0{,}805, \quad 0{,}817, \quad 0{,}848. \tag{3.7}$$

Teraz môžeme aproximovať neznáme hodnoty

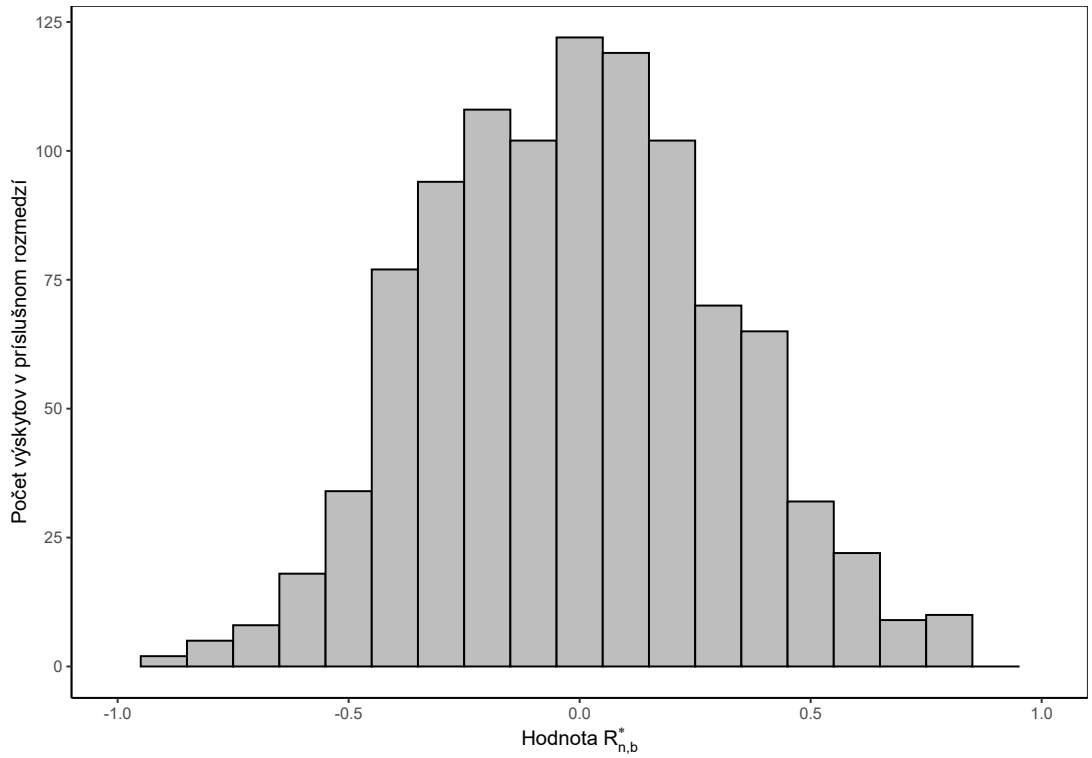
$$r_n(\alpha/2) = r_n(0{,}025) \text{ a } r_n(1-\alpha/2) = r_n(0{,}975)$$

hodnotami $r^*_{n,\,B}(0{,}025)$ a $r^*_{n,\,B}(0{,}975)$, ktoré predstavujú 2,5. a 97,5. výberový percentil. Volíme preto 25. a 975. hodnotu zo zoradenej postupnosti čísel (3.7) a obdržíme odhady

$$r^*_{n,\,B}(0{,}025) = -0{,}589 \text{ a } r^*_{n,\,B}(0{,}975) = 0{,}601.$$

Výsledný konfidenčný interval so spoľahlivosťou 95 % pre parameter θ má tvar

$$\left(\hat{\theta}_n - \frac{r^*_{n,\,B}(0{,}975)}{\sqrt{n}},\, \hat{\theta}_n - \frac{r^*_{n,\,B}(0{,}025)}{\sqrt{n}}\right) = \left(0{,}626 - \frac{0{,}601}{\sqrt{10}},\, 0{,}626 + \frac{0{,}589}{\sqrt{10}}\right) = (0{,}436,\, 0{,}813)\,.$$



Obr. 3.1: Histogram zobrazujúci hodnoty $R^*_{n,\,b}$, $b = 1, \ldots, 999$.

3.3.2 Percentilový bootstrapový interval spoľahlivosti

Druhou alternatívou je percentilový interval spoľahlivosti, ktorý má tvar

$$\left(q^*_{n,\,B}(\alpha/2), q^*_{n,\,B}(1-\alpha/2)\right),$$

kde $q^*_{n,\,B}(\alpha)$ je výberový α-kvantil spočítaný z hodnôt $\hat{\theta}^*_{n,\,1}, \ldots, \hat{\theta}^*_{n,\,B}$. Využitím rovnosti (3.6) môžeme percentilový intervalový odhad upraviť a dostaneme

$$\left(\hat{\theta}_n + \frac{r^*_{n,\,B}(\alpha/2)}{\sqrt{n}}, \hat{\theta}_n + \frac{r^*_{n,\,B}(1-\alpha/2)}{\sqrt{n}}\right). \tag{3.8}$$

Poznámka. Z vyjadrení (3.5) a (3.8) si môžeme všimnúť, že základné a percentilové bootstrapové intervaly spoľahlivosti sú rovnako široké.

Príklad. Našou úlohou bude spočítať percentilový 95% konfidenčný interval použitím dát z predošlého príkladu. Pre získanie požadovaného intervalu spoľahlivosti stačí dosadiť už zistené odhady do (3.8) a obdržíme

$$\left(0{,}626 - \frac{0{,}589}{\sqrt{10}},\ 0{,}626 + \frac{0{,}601}{\sqrt{10}}\right) = (0{,}440,\ 0{,}817)\,.$$

3.3.3 Študentizovaný bootstrapový interval spoľahlivosti

Pre odvodenie posledného typu bootstrapového intervalu spoľahlivosti pripomeňme, že z vety 1 platí

$$R_n = \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{D} \mathsf{N}(0, \sigma^2).$$

Označme

$$\widehat{R}_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\hat{\sigma}^2_n}},$$

kde $\hat{\sigma}^2_n$ je konzistentný odhad parametru σ^2. Nech $\hat{r}^*_n(\alpha)$ je α-kvantil rozdelenia

$$\widehat{R}^*_{n,\,b} = \frac{\sqrt{n}(\hat{\theta}^*_{n,\,b} - \hat{\theta}_n)}{\sqrt{\hat{\sigma}^{2*}_{n,\,b}}},\ b = 1, \ldots, B, \tag{3.9}$$

kde $\hat{\sigma}^{2*}_{n,\,b}$ sú odhady rozptylu σ^2 skonštruované z bootstrapových výberov. Tým dostávame študentizovaný intervalový odhad

$$\left(\hat{\theta}_n - \frac{\hat{r}^*_{n,\,B}(1-\alpha/2)\sqrt{\hat{\sigma}^2_n}}{\sqrt{n}},\ \hat{\theta}_n - \frac{\hat{r}^*_{n,\,B}(\alpha/2)\sqrt{\hat{\sigma}^2_n}}{\sqrt{n}}\right), \tag{3.10}$$

kde $\hat{r}^*_{n,\,B}(\alpha)$ je odhad α-kvantilu $\hat{r}^*_n(\alpha)$ spočítaný ako výberový α-kvantil z hodnôt $\widehat{R}^*_{n,\,1}, \ldots, \widehat{R}^*_{n,\,B}$.

Príklad. Nakoniec ukážeme na našich dátach konštrukciu 95% študentizovaného bootstrapového intervalu spoľahlivosti. Pre konzistentný odhad parametru σ^2 použijeme odhad (1.7) zo str. 5 založený na pôvodných dátach. Dostávame $\hat{\sigma}_n^2 = 0{,}107$. Ďalej spočítame $\widehat{R}^*_{n,\,b}$ z bootstrapových výberov a obdržíme postupnosť 999 čísel zoradených vzostupne

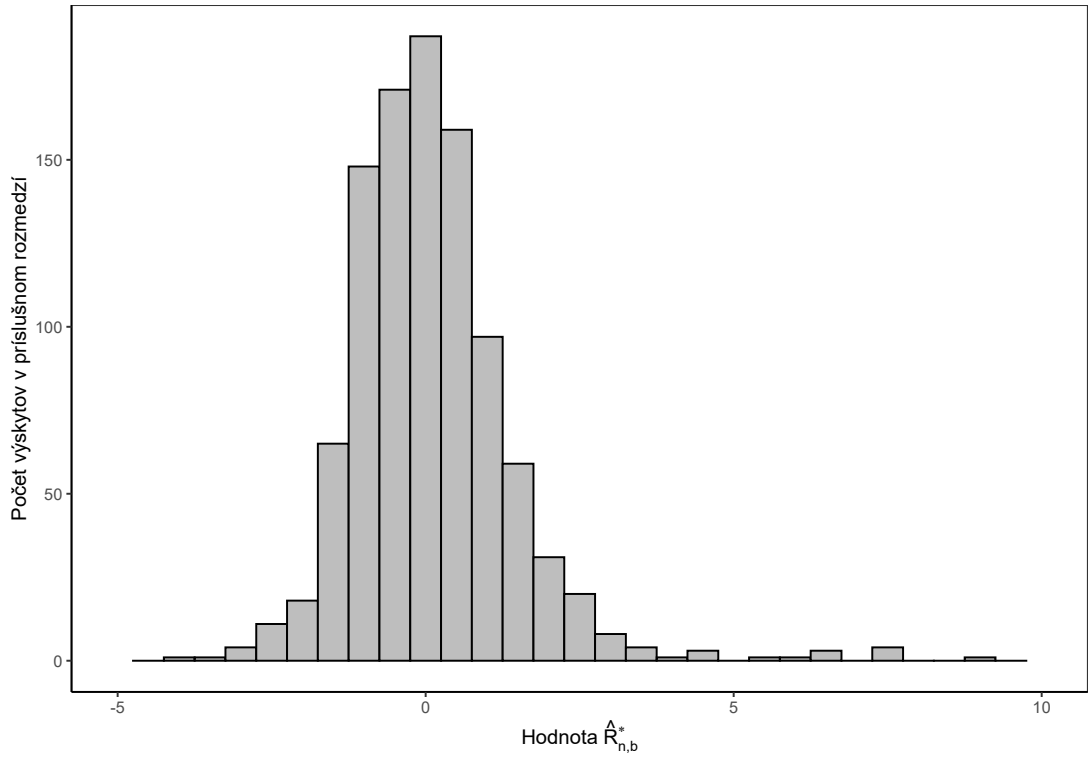
$$-4{,}188, \quad -3{,}729, \quad -3{,}161, \quad \ldots, \quad 7{,}619, \quad 9{,}066, \quad 9{,}839. \tag{3.11}$$

Pre lepšiu predstavu sú tieto hodnoty zobrazené v histograme na obrázku 3.2. Ako odhady neznámych hodnôt $r_n(0{,}025)$ a $r_n(0{,}975)$ volíme 25. a 975. hodnotu zo zoradenej postupnosti čísel (3.11) a obdržíme odhady

$$\hat{r}^*_{n,\,B}(0{,}025) = \ -2{,}022 \text{ a } \hat{r}^*_{n,\,B}(0{,}975) = \ 2{,}880.$$

Výsledný interval spoľahlivosti má po dosadení do (3.10) tvar

$$\left(0{,}626 - \frac{2{,}880\sqrt{0{,}107}}{\sqrt{10}},\ 0{,}626 + \frac{2{,}022\sqrt{0{,}107}}{\sqrt{10}}\right) = (0{,}329,\ 0{,}835)\,.$$



Obr. 3.2: Histogram zobrazujúci hodnoty $\widehat{R}^*_{n,\,b}$, $b = 1, \ldots, 999$.

Pre úplnosť, spočítajme ešte štandardný asymptotický interval spoľahlivosti (1.8) zo str. 6 a interval spoľahlivosti (2.3) zo str. 12 odvodený s použitím logitovej transformácie.

Z predchádzajúcich výpočtov sme spočítali hodnoty $\hat{\theta}_n = 0{,}626$ a $\hat{\sigma}_n^2 = 0{,}107$. Ďalej máme $n = 10$ pozorovaní a príslušný kvantil štandardného normálneho roz-

delenia je $u_{0,975} = 1{,}96$. Dosadením do (1.8) dostávame 95% štandardný asymptotický interval spoľahlivosti pre parameter θ

$$\left(0{,}626 - 1{,}96\sqrt{\frac{0{,}107}{10}},\ 0{,}626 + 1{,}96\sqrt{\frac{0{,}107}{10}}\right) = (0{,}424,\ 0{,}829)\,.$$

Pre logitovú transformáciu spočítame $\hat{\lambda}_n = 0{,}517$ a $\hat{\nu}_n^2 = 1{,}947$, z čoho po dosadení do (2.3) dostávame

$$\left(\frac{\exp\left(0{,}517 - 1{,}96\sqrt{\frac{1{,}947}{10}}\right)}{1 + \exp\left(0{,}517 - 1{,}96\sqrt{\frac{1{,}947}{10}}\right)},\ \frac{\exp\left(0{,}517 + 1{,}96\sqrt{\frac{1{,}947}{10}}\right)}{1 + \exp\left(0{,}517 + 1{,}96\sqrt{\frac{1{,}947}{10}}\right)}\right) = (0{,}414,\ 0{,}799)\,.$$

Zhrňme spočítané výsledky do tabuľky:

Metóda konštrukcie	Dĺžka intervalu	Interval spoľahlivosti
Základný bootstrap	0,377	(0,436; 0,813)
Percentilový bootstrap	0,377	(0,440; 0,817)
Študentizovaný bootstrap	0,506	(0,329; 0,835)
Štandardný asymptotický	0,405	(0,424; 0,829)
Logitová transformácia	0,385	(0,414; 0,799)

Tabuľka 3.1: Intervalové odhady pre parameter θ so spoľahlivosťou 95 %.

Z tabuľky 3.1 si môžeme všimnúť, že základný a percentilový bootstrap sú naozaj rovnako široké. Líšia sa iba posunutím. Zároveň sú tieto intervaly najužšie. Naopak, najširší interval pre naše dáta vykazuje študentizovaný bootstrap. Štandardný asymptotický a logitový interval sú šírkou podobné a nachádzajú sa v strede spomedzi spočítaných intervalov spoľahlivosti. Treba podotknúť, že všetky uvedené intervaly fungujú iba asymptoticky, teda pre veľké rozsahy výberu, pričom v príklade pracujeme iba s $n = 10$.

4. Simulácie

Na záver práce si ukážeme porovnanie intervalov spoľahlivosti podľa odvodenej teórie v predošlých kapitolách. Budeme konštruovať:

- Štandardný asymptotický interval spoľahlivosti (1.8) s rozptylom (1.7) - Štandardný asym.
- Štandardný asymptotický interval spoľahlivosti s modifikovaným rozptylom (1.19) - Štan. asym. modif.
- Logitový interval spoľahlivosti (2.3) s rozptylom (1.7) - Logitový
- Logitový interval spoľahlivosti s modifikovaným rozptylom (2.4) - Logitový modif.
- Základný bootstrapový interval spoľahlivosti (3.5) - Základný b.
- Percentilový bootstrapový interval spoľahlivosti (3.8) - Percentilový b.
- Študentizovaný bootstrapový interval spoľahlivosti (3.10) s rozptylom (1.7) - Študentizovaný b.

Simulácie budeme prirovnávať k problematike (ne)separovania odpadu načrtnutej v kapitole 1. Majme náhodný výber $\binom{Y_1}{T_1}, \binom{Y_2}{T_2}, \ldots, \binom{Y_n}{T_n}$, $n \in \mathbb{N}$. Pripomeňme, že Y_i značí hmotnosť nevytriedeného odpadu v i-tom komunálnom odpadkovom koši a T_i celkovú hmotnosť odpadu v i-tom odpadkovom koši.

Z povahy aplikácie chceme, aby sa výsledné intervaly spoľahlivosti nachádzali v intervale [0,1]. Preto budeme podobne ako v kapitole 2 predpokladať, že platí

$$T_i > 0 \text{ a } 0 \leq Y_i \leq T_i,\ i = 1, \ldots, n.$$

Občas sa vo výpočtoch stane, že výsledný intervalový odhad nie je podmnožinou intervalu [0,1]. V takých prípadoch budeme uvažovať prienik intervalu spoľahlivosti s intervalom [0,1]. Je to najmä z dôvodu, aby nám záporné spodné medze alebo horné medze väčšie ako 1 neskresľovali informácie o priemernej dĺžke jednotlivých intervalov spoľahlivosti.

4.1 Prvý simulačný model

Ako prvý typ simulácií uvažujme vzájomne nezávislé náhodné veličiny

$$\begin{aligned} Y_i &\sim \mathsf{Exp}(\lambda_1), \text{ s hustotou } f(x) = \lambda_1 e^{-\lambda_1 x}\, \mathbb{I}_{(0,\infty)}(x), \lambda_1 > 0,\ i = 1, \ldots, n, \\ Z_i &\sim \mathsf{Exp}(\lambda_2), \text{ s hustotou } f(x) = \lambda_2 e^{-\lambda_2 x}\, \mathbb{I}_{(0,\infty)}(x), \lambda_2 > 0,\ i = 1, \ldots, n. \end{aligned}$$

Potom pre náhodnú veličinu $T_i = Y_i + Z_i$ platí

$$T_i \sim \mathsf{Hypo}\left(\lambda_1, \lambda_2\right), \lambda_1 \neq \lambda_2.$$

Z vlastností exponenciálneho rozdelenia vieme, že

$$\mathsf{E}\ Y = 1/\lambda_1,\ \mathsf{E}\ Z = 1/\lambda_2 \text{ a teda } \mathsf{E}\ T = \mathsf{E}\ Y + \mathsf{E}\ Z = 1/\lambda_1 + 1/\lambda_2.$$

Teda parameter θ má tvar

$$\theta = \frac{\mathsf{E}\ Y}{\mathsf{E}\ T} = \frac{1/\lambda_1}{1/\lambda_1 + 1/\lambda_2} \in (0, 1)\,.$$

Pre simulovanie volíme 10 000 náhodných výberov Y_i a Z_i dĺžky $n = 20$ a raz pre porovnanie zvolíme výbery dĺžky 50. Bootstrapové intervaly sú generované pre $B = 999$ bootstrapových výberov. Všetky intervaly sú uvažované so spoľahlivosťou 95 %, t. j. $u_{0,975} = 1{,}96$. Rozoberme podrobne tri rôzne voľby parametru $\theta \in \{0{,}05,\ 0{,}25,\ 0{,}5\}$.

4.1.1 Parameter $\theta = 0{,}05$

Nech $\lambda_1 = 19$, $\lambda_2 = 1$. Potom $\theta = 0{,}05$ predstavuje 5 % nevytriedeného odpadu v komunálnom odpadkovom koši.

Metóda konštrukcie	Pokrytie θ	Priemerná dĺžka	Priemerné medze	Min. dolná a max. horná medza
Štandardný asym.	91,23 %	0,059	(0,023; 0,082)	(0,000; 0,296)
Štan. asym. modif.	91,56 %	0,060	(0,022; 0,082)	(0,000; 0,298)
Logitový	92,17 %	0,061	(0,030; 0,091)	(0,008; 0,317)
Logitový modif.	92,82 %	0,063	(0,029; 0,092)	(0,008; 0,320)
Základný b.	87,36 %	0,061	(0,014; 0,075)	(0,000; 0,272)
Percentilový b.	92,04 %	0,061	(0,030; 0,091)	(0,008; 0,313)
Študentizovaný b.	94,83 %	0,073	(0,026; 0,098)	(0,000; 0,314)

Tabuľka 4.1: Intervaly spoľahlivosti prvého modelu pre parameter $\theta = 0{,}05$ s pravdepodobnosťou pokrytia 95 %.

Pokrytia parametru θ z tabuľky 4.1 neaproximujú veľmi dobre požadovanú pravdepodobnosť s výnimkou študentizovaného bootstrapu, ktorý dopadol v prvom modeli pre $\theta = 0{,}05$ najlepšie. Naopak, najhoršie pokrytie ukázal základný bootstrap. Skutočnosť, že intervaly nespĺňajú presne predpísanú pravdepodobnosť pokrytia je spôsobená tým, že pracujeme s náhodnými výbermi dĺžky 20 a uvedené intervaly fungujú iba asymptoticky. Dôvod, prečo volíme $n = 20$ je ten, že v praxi nemáme k dispozícii veľa informácií o hmotnosti nevytriedeného odpadu v komunálnych odpadoch.

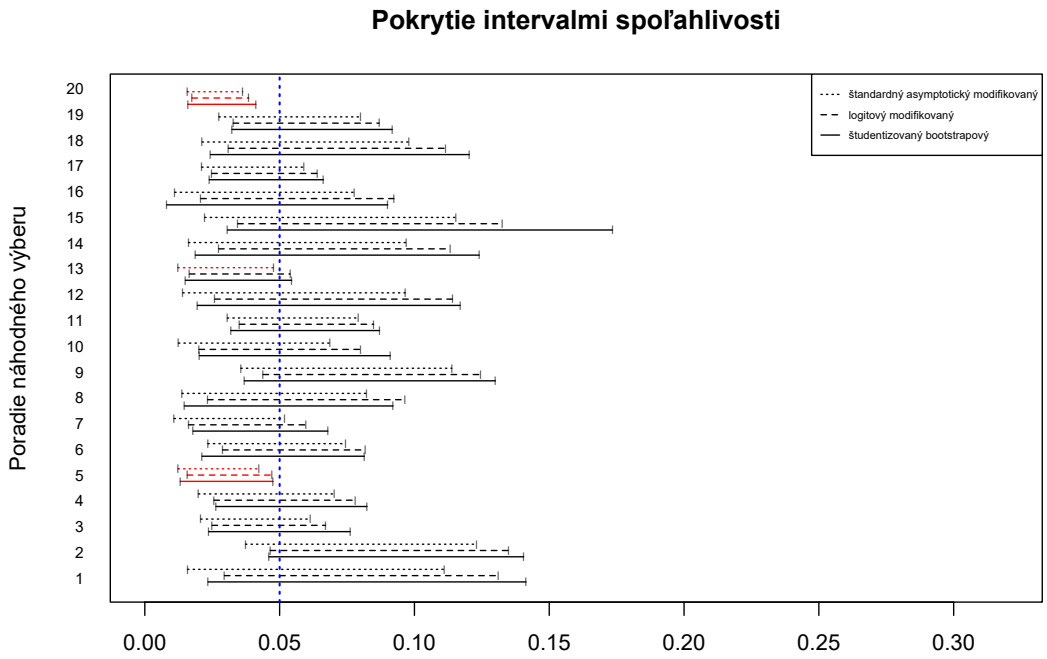
Za povšimnutie stojí stĺpec priemernej dĺžky intervalov, ktorá je veľmi krátka, čo je však prirodzené vzhľadom k tomu, že sme zvolili $\theta = 0{,}05$. To naznačuje, že v prípade malého množstva nevytriedeného odpadu môžeme očakávať dobrý odhad skutočnosti na základe náhodných výberov dĺžky 20 za predpokladu, že sme zvolili vhodný model.

Ďalej si môžeme všimnúť, že štandardný asymptotický, štandardný asymptotický modifikovaný, základný bootstrap a študentizovaný bootstrap majú minimálnu spodnú medzu rovnú 0 a to z dôvodu, že sme ňou nahradili zápornú dolnú medzu. Počet, koľkokrát vykazovali intervaly záporné spodné medze pre uvedené metódy je po poradí $2, 8, 961, 64$ z celkového počtu 10 000 opakovaní. Teda aj v tejto charakteristike dopadol základný bootstrap najhoršie.

Ďalšou zaujímavosťou je, že percentilový bootstrap sa javí byť vhodnejším kandidátom ako základný bootstrap, pričom oba majú vždy presne rovnakú dĺžku a líšia sa iba posunutím.

Modifikácia rozptylu (1.18) pomohla pre oba konfidenčné intervaly. Zlepšila pokrytie a zanedbateľne zvýšila priemernú dĺžku intervalov.

Celkovo hodnotíme najlepšie interval konštruovaný na základe logitovej transformácie s modifikovaným rozptylom. Taktiež musíme spomenúť aj študentizovaný bootstrap, ktorý má vynikajúce pokrytie, následkom čoho má ale najväčšiu priemernú dĺžku. Taktiež sme mu museli 64-krát orezávať spodnú medzu. Najhoršie hodnotíme základný bootstrap.



Obr. 4.1: Ilustrácia intervalov spoľahlivosti pre 20 náhodných výberov prvej simulačnej štúdie. Modrá vertikálna čiara symbolizuje skutočnú hodnotu parametru $\theta = 0{,}05$ a červená farba intervalu znamená, že nepokryl skutočnú hodnotu θ.

Na obrázku 4.1 vidíme tri typy intervalových odhadov. Bodkovaná čiara predstavuje štandardný asymptotický interval spoľahlivosti s modifikovaným rozptylom, prerušovaná logitový s taktiež modifikovaným rozptylom a plná študenzitovaný bootstrapový. Môžeme si všimnúť, že študentizovaný bootstrap vychádza najdlhší spomedzi zobrazených metód na obrázku.

V tabuľke 4.2 nájdeme výsledky prvého modelu pre $\theta = 0{,}05$ s rozsahom výberu $n = 50$. V porovnaní s tabuľkou 4.1 si môžeme všimnúť, že pre zväčšený rozsah výberu vyšli všetky výsledky lepšie, čo je v súlade s očakávaním. Podobne by sa výsledky zlepšili aj keby sme použili zväčšený rozsah náhodného výberu pre model použitý v sekcii 4.2.

Metóda konštrukcie	Pokrytie θ	Priemerná dĺžka	Priemerné medze	Min. dolná a max. horná medza
Štandardný asym.	93,21 %	0,037	(0,032; 0,069)	(0,013; 0,145)
Štan. asym. modif.	93,44 %	0,037	(0,032; 0,069)	(0,013; 0,145)
Logitový	93,91 %	0,038	(0,035; 0,073)	(0,015; 0,155)
Logitový modif.	94,14 %	0,038	(0,035; 0,073)	(0,015; 0,155)
Základný b.	90,99 %	0,038	(0,029; 0,066)	(0,010; 0,136)
Percentilový b.	93,93 %	0,038	(0,035; 0,073)	(0,015; 0,151)
Študentizovaný b.	95,00 %	0,040	(0,034; 0,074)	(0,014; 0,158)

Tabuľka 4.2: Intervaly spoľahlivosti prvého modelu pre parameter $\theta = 0{,}05$ s pravdepodobnosťou pokrytia 95 % pre náhodné výbery dĺžky 50.

4.1.2 Parameter $\theta = 0{,}25$

Voľme $\lambda_1 = 3$, $\lambda_2 = 1$. Potom $\theta = 0{,}25$ znázorňuje štvrtinu nevytriedeného odpadu v komunálnom odpade.

Metóda konštrukcie	Pokrytie θ	Priemerná dĺžka	Priemerné medze	Min. dolná a max. horná medza
Štandardný asym.	90,97 %	0,221	(0,145; 0,366)	(0,025; 0,761)
Štan. asym. modif.	91,37 %	0,225	(0,143; 0,368)	(0,018; 0,763)
Logitový	92,17 %	0,219	(0,161; 0,380)	(0,050; 0,746)
Logitový modif.	92,62 %	0,222	(0,160; 0,382)	(0,050; 0,747)
Základný b.	87,88 %	0,219	(0,130; 0,350)	(0,000; 0,755)
Percentilový b.	92,04 %	0,219	(0,161; 0,380)	(0,049; 0,743)
Študentizovaný b.	95,35 %	0,274	(0,137; 0,411)	(0,000; 1,000)

Tabuľka 4.3: Intervaly spoľahlivosti prvého modelu pre parameter $\theta = 0{,}25$ s pravdepodobnosťou pokrytia 95 %.

Z tabuľky 4.3 vidíme, že pokrytia θ sa v porovnaní s tabuľkou 4.1 zmenili minimálne. Prirodzene sa ale zmenila priemerná dĺžka intervalov, ktorá sa zvýšila.

Pre $\theta = 0{,}25$ zašiel do záporných hodnôt základný bootstrap 24-krát a študentizovaný bootstrap 54-krát. Zároveň bola raz horná medza pri študentizovanom bootstrape väčšia ako 1.

Opäť hodnotíme najlepšie logitový interval spoľahlivosti s modifikovaným rozptylom spolu so študentizovaným bootstrapom a naopak najhoršie základný bootstrap.

4.1.3 Parameter $\theta = 0{,}5$

Nakoniec, nech $\lambda_1 = 1$ a $\lambda_2 = 1$. Potom $\theta = 0{,}5$. Toto pokrytie si môžeme predstaviť tak, že polovica komunálneho odpadu obsahuje odpad, ktorý je možné vytriediť.

Metóda konštrukcie	Pokrytie θ	Priemerná dĺžka	Priemerné medze	Min. dolná a max. horná medza
Štandardný asym.	90,67 %	0,289	(0,356; 0,645)	(0,106; 0,918)
Štan. asym. modif.	91,07 %	0,293	(0,354; 0,647)	(0,103; 0,919)
Logitový	92,17 %	0,281	(0,360; 0,641)	(0,137; 0,898)
Logitový modif.	92,54 %	0,284	(0,358; 0,643)	(0,136; 0,899)
Základný b.	87,89 %	0,281	(0,360; 0,641)	(0,069; 0,937)
Percentilový b.	92,04 %	0,281	(0,360; 0,641)	(0,134; 0,896)
Študentizovaný b.	95,41 %	0,358	(0,322; 0,680)	(0,000; 1,000)

Tabuľka 4.4: Intervaly spoľahlivosti prvého modelu pre parameter $\theta = 0{,}5$ s pravdepodobnosťou pokrytia 95 %.

Výsledky z tabuľky 4.4 vieme interpretovať analogicky ako výsledky v tabuľke 4.3. Medze študentizovaného bootstrapu sa pre prípad $\theta = 0{,}5$ dostali do záporných hodnôt 13-krát a takisto 13-krát prekročili hodnotu 1.

4.2 Druhý simulačný model

Druhý model založíme na predpokladoch (1.9). Voľme rozdelenia náhodných výberov Y_i a T_i nasledovne

$$T_i \sim \mathsf{R}(1, 20) \text{ a } Y_i \mid T_i \sim \mathsf{Po}(\theta\, T_i),\ \theta \in (0, 1)\,,\ i = 1, \ldots, n.$$

Z vlastností Poissonového rozdelenia platí

$$\mathsf{E}\,[Y_i \mid T_i] = \theta\, T_i,\ \text{var}\,[Y_i \mid T_i] = \theta\, T_i.$$

Opäť generujeme 10 000 náhodných výberov T_i a Y_i dĺžky 20 a 999 bootstrapových výberov. Hodnoty parametru θ volíme presne ako v prvom modeli.

4.2.1 Parameter $\theta = 0{,}05$

Interpretácia parametru θ je rovnaká ako pre prvý model.

Metóda konštrukcie	Pokrytie θ	Priemerná dĺžka	Priemerné medze	Min. dolná a max. horná medza
Štandardný asym.	91,14 %	0,059	(0,021; 0,079)	(0,000; 0,177)
Štan. asym. modif.	91,34 %	0,059	(0,020; 0,080)	(0,000; 0,178)
Logitový	94,12 %	0,062	(0,028; 0,090)	(0,001; 0,200)
Logitový modif.	94,31 %	0,062	(0,028; 0,090)	(0,001; 0,202)
Základný b.	89,67 %	0,058	(0,020; 0,077)	(0,000; 0,171)
Percentilový b.	91,42 %	0,058	(0,023; 0,081)	(0,000; 0,184)
Študentizovaný b.	96,89 %	0,092	(0,022; 0,114)	(0,000; 1,000)

Tabuľka 4.5: Intervaly spoľahlivosti druhého modelu pre parameter $\theta = 0{,}05$ s pravdepodobnosťou pokrytia 95 %.

Z tabuľky 4.5 vidíme, že všetky metódy okrem logitových majú minimálnu spodnú medzu rovnú 0 čo znamená, že sa dolné medze dostali do záporných hodnôt s výnimkou percentilového bootstrapového intervalového odhadu, ktorý mal 8-krát dolnú medzu presne rovnú nule. V poradí zobrazenom v tabuľke mali intervaly zápornú dolnú medzu celkovo 261-, 272-, 0-, 0-, 505-, 0-, 59-krát a horná medza prevýšila hodnotu 1 pri študentizovanom bootstrape 210-krát.

Dôvod, prečo bola horná medza nahradená 1 pri študentizovanom bootstrape pre θ iba 0,05 je ten, že výberový kvantil $\hat{r}^*_{n,B}(0{,}025) = -\infty$, čo sa prejavilo na hornej medzi, ktorá vyšla ∞. To nastáva v prípade, keď je bootstrapový výber (3.1) nulový pre všetky zložky $Y^*_{1,b}, \ldots, Y^*_{n,b}$, čo má za následok delenie nulou vo výraze (3.9). Keby nám vyšiel takýto prípad v praxi, tak by sme študentizovaný bootstrap nepoužili.

Pri porovnaní tabuľky 4.5 s tabuľkou 4.1 si môžeme všimnúť, že štandardné asymptotické intervaly s oboma rozptylmi dopadli takmer identicky. Výraznejší rozdiel môžeme pozorovať iba pri maximálnej hornej medzi, ktorá vyšla v druhom modeli približne o 0,12 nižšie ako v prípade prvého modelu.

Oba logitové intervaly majú v druhom modeli výrazne lepšie pokrytie v porovnaní s prvým modelom, pričom priemerná dĺžka ostala bez viditeľnej zmeny. Minimálna spodná aj maximálna horná medza sa znížili, čo mohlo mať v konečnom dôsledku vplyv na zlepšenie pokrytia, keďže priemerné medze ostali takmer rovnaké.

Základný bootstrap zvýšil pravdepodobnosť pokrytia o viac než 2 %, avšak pokrytie 89,67 % stále nie je ani zďaleka uspokojivé.

Pri percentilovom bootstrape pozorujeme naopak nižšiu pravdepodobnosť pokrytia, obe priemerné medze sa znížili a takisto sa znížila aj minimálna spodná a maximálna horná medza. Z týchto pozorovaní môžeme urobiť záver, že sa intervaly konštruované percentilovým bootstrapom v druhom modeli posunuli mierne doľava.

Celkovo hodnotíme výsledky druhého modelu s $\theta = 0{,}05$ veľmi pozitívne pre obe logitové transformácie a aj pre základný bootstrap, ktorý však ostáva aj napriek zlepšenému pokrytiu stále najhorší. Naopak, negatívne výsledky ukázal druhý model najmä v prípade percentilového bootstrapu.

4.2.2 Parameter $\theta = 0{,}25$

Metóda konštrukcie	Pokrytie θ	Priemerná dĺžka	Priemerné medze	Min. dolná a max. horná medza
Štandardný asym.	92,82 %	0,132	(0,184; 0,316)	(0,070; 0,487)
Štan. asym. modif.	93,07 %	0,133	(0,183; 0,317)	(0,069; 0,487)
Logitový	93,48 %	0,132	(0,190; 0,322)	(0,084; 0,491)
Logitový modif.	93,56 %	0,133	(0,190; 0,322)	(0,084; 0,492)
Základný b.	92,19 %	0,131	(0,183; 0,313)	(0,069; 0,491)
Percentilový b.	92,67 %	0,131	(0,186; 0,317)	(0,076; 0,486)
Študentizovaný b.	95,12 %	0,151	(0,182; 0,332)	(0,073; 0,587)

Tabuľka 4.6: Intervaly spoľahlivosti druhého modelu pre parameter $\theta = 0{,}25$ s pravdepodobnosťou pokrytia 95 %.

Pri porovnaní tabuliek 4.6 a 4.3 si môžeme všimnúť, že pokrytie parametru θ je lepšie v druhom modeli pre všetky metódy s výnimkou študentizovaného bootstrapu, ktorý ale aj naďalej spĺňa predpísané pokrytie.

Priemerná dĺžka intervalov je v druhom modeli nižšia približne o 0,09 a v prípade študentizovaného bootstrapu až o 0,12. Z týchto dvoch kritérií by sme mohli usúdiť, že lepšie výsledky intervalov spoľahlivosti pre hodnotu parametra $\theta = 0{,}25$ generoval druhý model. Túto domnienku potvrdzujú aj fakty, že priemerná dolná medza je vyššia a priemerná horná nižšia v druhom modeli pre všetky použité metódy.

Zároveň z výsledkov prvého modelu vieme, že minimálna spodná medza pri študentizovanom bootstrape zašla do záporných čísel a to 13-krát. Horná medza prekročila hodnotu 1 tiež 13-krát, pričom druhý model sa obom týmto problémom vyhol.

Za povšimnutie stojí úprava pokrytia pri základnom bootstrape, ktoré sa zvýšilo až o 4,31 %.

4.2.3 Parameter $\theta = 0{,}5$

Porovnanie oboch modelov pre parameter $\theta = 0{,}5$ z tabuliek 4.7 a 4.4 vieme diskutovať analogicky ako pre $\theta = 0{,}25$. Dospejeme k záveru, že aj pre parameter $\theta = 0{,}5$ vykazujú intervaly spoľahlivosti lepšie výsledky v druhom modeli.

Metóda konštrukcie	Pokrytie θ	Priemerná dĺžka	Priemerné medze	Min. dolná a max. horná medza
Štandardný asym.	92,86 %	0,187	(0,406; 0,594)	(0,238; 0,817)
Štan. asym. modif.	93,13 %	0,189	(0,405; 0,594)	(0,237; 0,819)
Logitový	93,47 %	0,185	(0,407; 0,592)	(0,250; 0,795)
Logitový modif.	93,69 %	0,186	(0,407; 0,593)	(0,248; 0,796)
Základný b.	92,39 %	0,185	(0,406; 0,591)	(0,221; 0,823)
Percentilový b.	92,56 %	0,185	(0,409; 0,594)	(0,236; 0,835)
Študentizovaný b.	95,05 %	0,212	(0,401; 0,613)	(0,233; 0,945)

Tabuľka 4.7: Intervaly spoľahlivosti druhého modelu pre parameter $\theta = 0{,}5$ s pravdepodobnosťou pokrytia 95 %.

Záver

V tejto práci sme sa venovali odvodeniu rôznych typov intervalových odhadov pre parameter $\theta = \frac{\mathsf{E}\, Y}{\mathsf{E}\, T}$.

Odvodili sme štandardný asymptotický interval spoľahlivosti, logitovú transformáciu štandardného asymptotického intervalu spoľahlivosti a bootstrapové konfidenčné intervaly (základný, percentilový a študentizovaný).

Následne sme pomocou simulácií skúmali vlastnosti všetkých odvodených typov intervalových odhadov v dvoch modeloch. Zistilo sa, že najlepšie pokrytie vykazoval študentizovaný bootstrapový interval spoľahlivosti. Avšak, má najširšie rozpätie, je potrebné občas orezávať medze a je najnáročnejší na konštrukciu. Z tohto dôvodu radšej odporúčame zvoliť logitovú transformáciu s modifikovaným rozptylom. Túto metódu je možné jednoduchšie spočítať a to aj bez použitia počítača. Zároveň je jej pravdepodobnosť pokrytia veľmi blízka tej predpísanej a výsledok vždy vedie k intervalovému odhadu nachádzajúcemu sa v intervale $(0, 1)$.

Na záver simulačnej štúdie nahradíme v modeli z kapitoly 4.1 pre $\theta = 0{,}05$ použité kvantily $u_{1-\alpha/2}$ kvantilmi $t_{n-1}(1-\alpha/2)$ Studentovho t-rozdelenia s $n-1$ stupňami voľnosti, ktoré sú väčšie ako kvantily štandardného normálneho rozdelenia. Bootstrapové intervalové odhady vychádzajú bez zmeny, pretože sa v nich použitie iných kvantilov neprejaví. Výsledky tejto simulácie popisuje tabuľka 4.8. Vidíme, že sa výmena kvantilov pozitívne prejavila na zlepšenom pokrytí v porovnaní s pokrytím v tabuľke 4.1.

Predmetom ďalšieho skúmania by preto mohlo byť odvodenie pravidla pre voľbu stupňov voľnosti, ktoré by reflektovalo, že veličiny $T_1, \ldots, T_n$ sú rôzne.

Metóda konštrukcie	Pokrytie θ	Priemerná dĺžka	Priemerné medze	Min. dolná a max. horná medza
Štandardný asym.	92,60 %	0,063	(0,021; 0,084)	(0,000; 0,303)
Štan. asym. modif.	92,98 %	0,064	(0,020; 0,084)	(0,000; 0,306)
Logitový	93,94 %	0,066	(0,029; 0,094)	(0,008; 0,327)
Logitový modif.	94,39 %	0,067	(0,028; 0,096)	(0,008; 0,330)
Základný b.	87,36 %	0,061	(0,014; 0,075)	(0,000; 0,272)
Percentilový b.	92,04 %	0,061	(0,030; 0,091)	(0,008; 0,313)
Študentizovaný b.	94,83 %	0,073	(0,026; 0,098)	(0,000; 0,314)

Tabuľka 4.8: Intervaly spoľahlivosti prvého modelu pre parameter $\theta = 0{,}05$ s pravdepodobnosťou pokrytia 95 % s použitím kvantilov Studentovho rozdelenia s $n-1$ stupňami voľnosti.

Dodatky

Veta D1. Majme náhodné veličiny X, Y, Z, konštanty a, $b \in \mathbb{R}$ a $g : \mathbb{R} \to \mathbb{R}$. Prepokladajme, že podmienené stredné hodnoty existujú. Potom platí

(*i*) $\mathsf{E}[a \mid Y] = a$,

(*ii*) $\mathsf{E}[aX + bZ \mid Y] = a\,\mathsf{E}[X \mid Y] + b\,\mathsf{E}[Z \mid Y]$,

(*iii*) $\mathsf{E}[X \mid Y] = \mathsf{E}[X]$ pre X a Y nezávislé,

(*iv*) $\mathsf{E}[Xg(Y) \mid Y] = g(Y)\,\mathsf{E}[X \mid Y]$.

Dôkaz. Viď Lachout (2004), kapitola 7. □

Zoznam použitej literatúry

Davison, A. C. a Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. ISBN 0 521 57391 2.

Dupač, V. a Hušková, M. (2009). *Pravděpodobnost a matematická statistika*. Prvé vydání. Nakladatelství Karolinum, Praha. ISBN 978-80-246-009-3.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**(1), 1 – 26. doi: 10.1214/aos/1176344552.

Gut, A. (2005). *Probability: A Graduate Course*. Springer Texts in Statistics. Springer. ISBN 978-0-387-27332-7.

Lachout, P. (2004). *Teorie pravděpodobnosti*. Druhé vydání. Nakladatelství Karolinum, Praha. ISBN 80-246-0872-3.

Omelka, M. (2021). Modern statistical methods. `https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmst434/nmst434_course-notes_2020.pdf`. Posledný prístup: 20-02-2021.

Pick, L., Hencl, S., Spurný, J. a Zelený, M. (2020). Matematická analýza 1. `https://www2.karlin.mff.cuni.cz/~pick/analyza.pdf`. Posledný prístup: 12-10-2020.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. ISBN 0 521 49603 9.