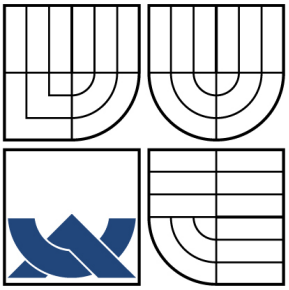
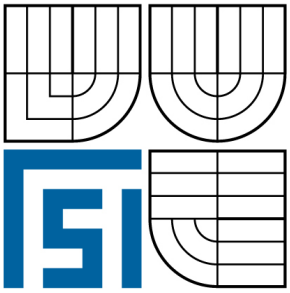


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

METODA BOOTSTRAP A JEJÍ APLIKACE

BOOTSTRAP METHOD AND ITS APPLICATION

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — LUCIE PAVLÍČKOVÁ
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. ZDENĚK KARPÍŠEK, CSc.
SUPERVISOR

BRNO 2009

Vysoké učení technické v Brně, Fakulta strojního inženýrství

Ústav matematiky
Akademický rok: 2009/2010

ZADÁNÍ DIPLOMOVÉ PRÁCE

student(ka): Lucie Pavlíčková

který/která studuje v **magisterském studijním programu**

obor: **Matematické inženýrství (3901T021)**

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Metoda bootstrap a její aplikace

v anglickém jazyce:

Bootstrap Method and its Application

Stručná charakteristika problematiky úkolu:

Popis metody bootstrap, jejích vlastností a užití na reálných datových souborech.

Cíle diplomové práce:

Princip a vlastnosti metody bootstrap. Odhady momentových a kvantilových charakteristik rozdělení pravděpodobnosti. PC realizace metody ve statistických softwarech. Aplikace bootstrapu na reálných datových souborech.

Seznam odborné literatury:

1. Chernick, M. R. Bootstrap Methods: A Practitioner's Guide. New York: Wiley, 1999.
2. Davison, A. C. and Hinkley, D. V. Bootstrap Methods and Their Application. Cambridge, England: Cambridge University Press, 1997.
3. Efron, B. and Tibshirani, R. J. An Introduction to the Bootstrap. Boca Raton, FL: CRC Press, 1994.
4. Mooney, C. Z. and Duval, R. D. Bootstrapping: A Nonparametric Approach to Statistical Inference. Sage, 1993.
5. Odborné články dle pokynů vedoucího diplomové práce.

Vedoucí diplomové práce: doc. RNDr. Zdeněk Karpíšek, CSc.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2009/2010.

V Brně, dne

L.S.

prof. RNDr. Josef Šlapal, CSc. Ředitel ústavu	doc. RNDr. Miroslav Doupovec, CSc. Děkan fakulty

Abstrakt

Diplomová práce popisuje metodu bootstrap a její použití pro určení přesnosti odhadu, tvorbu konfidenčních intervalů a testování statistických hypotéz. Dále je předložena metoda odhadu diskrétního rozdělení pravděpodobnosti kategoriální veličiny využívající gradient kvazinormy tohoto rozdělení. Metoda bootstrap je v konkrétních příkladech aplikována k získání konfidenčního intervalu pravděpodobnostní funkce kategoriální veličiny.

Diplomová práce je součástí řešení projektu MŠMT České republiky čís. 1M06047 „Centrum pro jakost a spolehlivost výroby“, projektu Grantové agentury České republiky reg. čís. 103/08/1658 „Pokročilá optimalizace návrhu složených betonových konstrukcí“ a výzkumného záměru MŠMT České republiky čís. MSM0021630519 „Progresivní spolehlivé a trvanlivé nosné stavební konstrukce“.

Summary

The diploma thesis describes the bootstrap method and its applications in the estimate accuracy statement, in the confidence intervals generation and in the testing of statistical hypotheses. Further the method of the discrete probability estimation of the categorical quantity is presented, making use the gradient of the quasi-norm hereof distribution. On concrete examples the bootstrap method is applied in the confidence intervals forming of the categorical quantity probability function.

The diploma thesis was supported by the project of MŠMT of the Czech Republic no. 1M06047 “Centre for Quality and Reliability of Production”, by the grant of Grant Agency of the Czech Republic (Czech Science Foundation) reg. no. 103/08/1658 “Advanced optimum design of composed concrete structures” and by the research plan of MŠMT of the Czech Republic no. MSM0021630519 “Progressive reliable and durable structures”.

Klíčová slova

bootstrap, odhad parametru, přesnost odhadu, konfidenční interval, BCA, test statistické hypotézy, f-divergence, kvazinorma, diskrétní rozdělení pravděpodobnosti, gradientní odhad

Keywords

bootstrap, parameter estimate, accuracy of the estimate, confidence interval, BCA, statistical hypothesis testing, f-divergence, quasi-norm, discrete probability distribution, gradient estimate

PAVLÍČKOVÁ, L. *Metoda bootstrap a její aplikace.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2009. 64 s. Vedoucí diplomové práce doc. RNDr. Zdeněk Karpíšek, CSc.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně a že jsem uvedla všechny využité prameny a literaturu.

Děkuji vedoucímu práce za jeho trpělivost, ochotu, vstřícnost a pomocnou ruku. Děkuji svému otci za cenné rady při psaní této práce a za její důkladné přečtení. A konečně děkuji svému manželovi za jeho obrovskou podporu, pomoc a lásku, která mi dodávala sílu v probdělých nocích.

Obsah

Kapitola 1

Úvod

Statistická teorie se pokouší odpovědět na tři základní otázky: jak získat data, jak tato data analyzovat a shrnout a jak ověřit jejich přesnost. Bootstrap je metoda, která přináší snadno pochopitelnou a proveditelnou odpověď na třetí otázku. Jeho princip poprvé popsal Bradly Efron, profesor Stanfordské univerzity, v roce 1979. Jednalo se tehdy o jednu z prvních metod, která ve statistice nahrazovala tradiční algebraické výpočty počítačovými simulacemi na pozorovaných datech. Bootstrap přinesl možnost odhadnout přesnost libovolného odhadu libovolného parametru. Přitom spočívá v prosté myšlence mnohonásobného opakování jednoduchého algoritmu. Navíc není závislý na centrální limitní větě, a proto ho lze s úspěchem použít i pro výběry s malým rozsahem. S rozvojem a zrychlováním počítačů se otevřely dveře pro další aplikace bootstrapu, zejména pro konstruování konfidenčních intervalů, testování statistických hypotéz a v oblasti regresní analýzy. Dnes je bootstrap stále častěji používanou metodou s širokým využitím, bylo o něm napsáno množství knih a má své pevné místo mezi matematickým softwarem.

V této diplomové práci popíšeme, co je to bootstrap a jak ho použít pro výpočet střední kvadratické chyby odhadu, rozptylu nebo vychýlení. Předložíme několik přístupů, kterými lze získat bootstrapový konfidenční interval, a ukážeme, jak s jeho pomocí testovat statistické hypotézy. Načrtneme také, jak aplikovat bootstrap na časové řady, vícerozměrné náhodné výběry a jak ho lze použít v regresní analýze.

V deváté kapitole seznámíme čtenáře s novou metodou odhadu pravděpodobnostní funkce kategoriální veličiny pomocí gradientu kvazinormy jejího rozdělení. V desáté kapitole se nakonec budeme věnovat aplikaci bootstrapu na tuto metodu a sestrojíme intervalové odhady pravděpodobnostních funkcí konkrétních kategoriálních veličin. Otestujeme při tom nový a stále se vyvíjející software Shine bootstrap.

Kapitola 2

Základní pojmy

Na úvod připomeneme některé základní pojmy z teorie pravděpodobnosti a statistiky, jako jsou náhodná veličina a náhodný vektor a jejich charakteristiky, náhodný výběr, rozdělení pravděpodobnosti, odhady parametrů, testování hypotéz nebo regresní analýza. Viz také [5].

2.1. Pravděpodobnost, náhodná veličina, náhodný vektor a jejich charakteristiky

Pokusem rozumíme realizaci určitého systému podmínek, které jsou opakovatelné a neměnné. Výsledkem pokusu je náhodný jev. Jednotlivým možným výsledkům pokusu odpovídají *elementární náhodné jevy*, které vyjadřujeme pomocí jednoprvkových množin $\{\omega\}$. Všechny výsledky pokusu tvoří množinu Ω, kterou nazýváme *základní prostor*. Platí $\omega \in \Omega$. *Náhodným jevem* A pak rozumíme podmnožinu základního prostoru, tedy $A \subseteq \Omega$. *Opačný náhodný jev* k jevu A je jev $\overline{A}$, který nastane právě tehdy, když nenastane jev A; tj. $\overline{A}$ je doplněk A v Ω. *Jevové pole* Σ na Ω je množina náhodných jevů s vlastnostmi:

1. $\Omega \in \Sigma$,
2. pro každé $A \in \Sigma$ je $\overline{A} \in \Sigma$,
3. pro každou posloupnost náhodných jevů $A_i \in \Sigma$, $i = 1, 2, \ldots,$ je

$$\bigcap_{i=1}^{\infty} A_i \in \Sigma.$$

Pravděpodobnost $P(A)$ náhodného jevu $A \in \Sigma$ je reálná funkce definovaná na jevovém poli Σ s vlastnostmi:

1. $P(A) \geqq 0$ pro všechny náhodné jevy $A \in \Sigma$,
2. $P(\Omega) = 1$,
3. pro každou posloupnost disjunktních náhodných jevů $A_i \in \Sigma$, $i = 1, 2, \ldots,$ je

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Uspořádaná trojice (Ω, Σ, P) se nazývá *pravděpodobnostní prostor*.

Náhodná veličina (náhodná proměnná) X je funkce, která nabývá náhodně reálných číselných hodnot x. Formálně jde tedy o zobrazení $X\colon \Omega \to \mathbb{R}$. Přitom požadujeme, aby pro každé $c \in \mathbb{R}$ množina $\{X < c\} = \{\omega \in \Omega : X(\omega) \in (-\infty, c)\}$ byla jevem, tj. $\{X < c\} \in \Sigma$ (tzv. *borelovsky měřitelné* zobrazení). Množina všech hodnot náhodné veličiny X se nazývá *základní soubor* nebo také *populace*. *Distribuční funkce* náhodné veličiny X je reálná funkce

$$F(x) = P(X < x) = P\big(X \in (-\infty; x)\big),$$

definovaná pro všechna $x \in (-\infty; \infty)$. Distribuční funkcí je náhodná veličina plně popsána a říkáme, že je dáno její *rozdělení pravděpodobnosti*.

Náhodná veličina X je *diskrétní* (má *diskrétní rozdělení pravděpodobnosti*), jestliže nabývá s nenulovou pravděpodobností nejvýše spočetně mnoha hodnot $x_1, x_2, \ldots$. Její *pravděpodobnostní funkce* je posloupnost

$$p(x) = P(X = x) > 0 \text{ pro } x = x_1, x_2, \ldots.$$

Náhodná veličina X je *spojitá* (má *spojité rozdělení pravděpodobnosti*), jestliže má tzv. *absolutně spojitou* distribuční funkci $F(x)$; tzn. že existuje nezáporná funkce $f(x)$ taková, že pro každé $x \in (-\infty; \infty)$ je

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

Funkci $f(x)$ nazýváme její *hustotou pravděpodobnosti*.

Číselné charakteristiky náhodné veličiny X jsou reálná čísla, která vyjadřují její důležité vlastnosti. Nejprve uvedeme *momentové charakteristiky*.

Polohu rozdělení pravděpodobnosti náhodné veličiny X charakterizuje její *střední hodnota*

$$E(X) = \sum_{x} x p(x)$$

pro diskrétní náhodnou veličinu X (pokud řada konverguje absolutně),

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

pro spojitou náhodnou veličinu X (pokud integrál konverguje absolutně).

Míru kolísání hodnot náhodné veličiny X kolem její střední hodnoty $E(X)$ vyjadřuje její *rozptyl (variance)*

$$D(X) = E\big([X - E(X)]^2\big).$$

Odmocninu z rozptylu pak nazýváme *směrodatnou odchylkou* náhodné veličiny X a značíme $\sigma(X) = \sqrt{D(X)}$. Odtud pak plyne také alternativní označení pro rozptyl $\sigma^2(X)$.

Míru asymetrie rozdělení náhodné veličiny X vzhledem k její střední hodnotě vyjadřuje *koeficient šikmosti (asymetrie)*

$$A(X) = \frac{E\big([X - E(X)]^3\big)}{[\sigma(X)]^3}.$$

Koeficient špičatosti (excesu)

$$\gamma_2(X) = \frac{E[X - E(X)]^4}{\big[D(x)\big]^2} - 3$$

porovnává rozdělení pravděpodobnosti náhodné veličiny X s normálním rozdělením.

Nejdůležitější *kvantilovou číselnou charakteristikou* náhodné veličiny X je *p-kvantil* náhodné veličiny $x_p = \inf\{x; F(x) \geqq p\}$, kde $0 < p < 1$. Pro spojitou náhodnou veličinu X s rostoucí distribuční funkcí je $F(x_p) = p$. Kvantil $x_{0,5}$ je *medián* náhodné veličiny X a charakterizuje polohu jejího rozdělení pravděpodobnosti.

Modus $\hat{x}$ náhodné veličiny X je taková hodnota, v níž nabývá její pravděpodobnostní funkce nebo hustota pravděpodobnosti maximum, popř. supremum.

Uspořádanou n-tici náhodných veličin $X_1, \ldots, X_n$ nazýváme *n-rozměrným náhodným vektorem* $(X_1, \ldots, X_n)$. Dále *simultánní (sdružená) distribuční funkce* náhodného vektoru $(X_1, \ldots, X_n)$ je reálná funkce

$$F(x_1, \ldots, x_n) = P(X_1 < x_1, \ldots, X_n < x_n) = P\big((X_1, \ldots, X_n) \in (-\infty; x_1) \times \ldots \times (-\infty; x_n)\big)$$

definovaná pro všechny n-tice $(x_1, \ldots, x_n) \in (-\infty; \infty)^n$. Simultánní distribuční funkcí je náhodný vektor $(X_1, \ldots, X_n)$ jednoznačně popsán a říkáme, že je dáno jeho *simultánní rozdělení pravděpodobnosti*.

Náhodný vektor je *diskrétní*, jestliže všechny jeho složky jsou diskrétní. Náhodný vektor je *spojitý*, jestliže všechny jeho složky jsou nezávislé a spojité. Jestliže u náhodného vektoru vynecháme některé jeho složky, dostaneme *marginální rozdělení pravděpodobnosti*. *Podmíněné rozdělení pravděpodobnosti* dostaneme, pokud budeme uvažovat podmínku, že některé ze složek náhodného vektoru nabývají libovolné pevné hodnoty.

Mezi nejdůležitější číselné charakteristiky náhodného vektoru patří *střed (centrum)*, což je uspořádaná n-tice $(E(X_1), \ldots, E(X_n))$, jejímiž složkami jsou střední hodnoty složek náhodného vektoru.

Vzájemný vztah dvou složek X_i, X_j náhodného vektoru $(X_1, \ldots, X_n)$ vyjadřuje jejich *kovariance*

$$\operatorname{cov}(X_i, X_j) = E\big([x_i - E(X_i)][x_j - E(X_j)]\big) = E(X_i X_j) - E(X_i)E(X_j).$$

Z kovariancí se sestavuje *kovarianční matice*

$$\mathbf{cov}(X_1, \ldots, X_n) = \begin{pmatrix} D(X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & D(X_2) & \cdots & \operatorname{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \cdots & D(X_n) \end{pmatrix}.$$

Mírou lineární závislosti dvou složek X_i, X_j náhodného vektoru $(X_1, \ldots, X_n)$ je jejich *koeficient korelace*

$$\rho(X_i, X_j) = \operatorname{cov}\left(\frac{X_i - E(X_i)}{\sigma(X_i)}, \frac{X_j - E(X_j)}{\sigma(X_j)}\right) = \frac{\operatorname{cov}(X_i, X_j)}{\sigma(X_i)\sigma(X_j)}.$$

Z koeficientů korelace se sestavuje *korelační matice* náhodného vektoru $(X_1, \ldots, X_n)$

$$\boldsymbol{\rho}(X_1, \ldots, X_n) = \begin{pmatrix} 1 & \rho(X_1, X_2) & \cdots & \rho(X_1, X_n) \\ \rho(X_2, X_1) & 1 & \cdots & \rho(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(X_n, X_1) & \rho(X_n, X_2) & \cdots & 1 \end{pmatrix}.$$

2.2. Náhodný výběr a jeho charakteristiky

Opakujeme-li n-krát nezávisle pokus, jehož výsledkem je hodnota náhodné veličiny X s distribuční funkcí $F(x, \theta)$, kde θ je reálný parametr (případně vektor parametrů nebo jejich funkce) daného rozdělení pravděpodobnosti, pozorujeme vlastně náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)$ a předpokládáme, že jeho složky jsou nezávislé náhodné veličiny X_i se stejnou distribuční funkcí, jako má pozorovaná náhodná veličina X. Náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)$ pak nazýváme *náhodným výběrem* z náhodné veličiny X nebo z jejího rozdělení pravděpodobnosti a číslo n *rozsahem náhodného výběru*. Množinu všech uvažovaných hodnot parametru θ nazýváme *parametrickým prostorem*.

Číselný vektor $\boldsymbol{x} = (x_1, \ldots, x_n)$, který získáme při realizaci náhodného výběru, kde x_i je pozorovaná hodnota složky X_i, $i = 1, \ldots, n$, nazýváme *pozorovanou hodnotou náhodného výběru* $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Funkci náhodného výběru $T(X_1, \ldots, X_n)$ nazýváme *výběrovou charakteristikou* nebo *statistikou*, její hodnotu na pozorované hodnotě náhodného výběru $t = T(x_1, \ldots, x_n)$ nazýváme *empirickou charakteristikou* nebo *pozorovanou hodnotou statistiky* T.

Nejvýznamnějšími výběrovými charakteristikami jsou:

1. *výběrový průměr*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

2. *výběrový rozptyl*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

3. *výběrová směrodatná odchylka*

$$S = \sqrt{S^2},$$

4. *výběrový koeficient korelace*

$$R = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{S(X)S(Y)}$$

pro náhodný výběr z náhodného vektoru (X, Y), kde $S(X)$, $S(Y)$ jsou výběrové směrodatné odchylky náhodných veličin X, Y.

Základní vlastnosti výběrového průměru a výběrového rozptylu jsou:

1. Jestliže pozorovaná náhodná veličina X má střední hodnotu $E(X)$, pak

$$E(\overline{X}) = E(X).$$

2. Jestliže pozorovaná náhodná veličina X má rozptyl $D(X)$, pak

$$D(\overline{X}) = \frac{D(X)}{n}, \quad \sigma(\overline{X}) = \frac{\sigma(X)}{\sqrt{n}}, \quad E(S^2) = D(X).$$

Hodnoty empirických charakteristik jsou náhodné, při opakovaných realizacích náhodného výběru se náhodně mění. Z předcházejícího však plyne, že např. pro $n \to \infty$ rozptyl výběrového průměru $D(\overline{X}) \to 0$, takže pro dostatečně velké n je „takřka jistě" aritmetický průměr $\overline{x}$ blízký neznámé střední hodnotě rozdělení $E(X)$.

2.3. Rozdělení pravděpodobnosti pro aplikace

Uvedeme si některá známá rozdělení pravděpodobnosti.

Normální (Gaussovo) rozdělení $N(\mu, \sigma^2)$, kde $\mu, \sigma^2 \in \mathbb{R}$, $\sigma > 0$, má hustotu pravděpodobnosti

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad x \in (-\infty, \infty),$$

střední hodnotu $E(X) = \mu$, rozptyl $D(X) = \sigma^2$ a koeficient šikmosti $A(X) = 0$.

Transformací náhodné veličiny X s normálním rozdělením $N(\mu, \sigma^2)$ na náhodnou veličinu

$$Z = \frac{X - \mu}{\sigma}$$

dostaneme *normované normální rozdělení* $N(0, 1)$. Pro jeho kvantily platí $z_{1-\alpha} = -z_\alpha$, $\alpha \in (0, 1)$.

Pearsonovo rozdělení $\chi^2(k)$ s k stupni volnosti, kde $k \in \mathbb{N}$, má hustotu pravděpodobnosti

$$f(x) = \begin{cases} \frac{\mathrm{e}^{-x/2} x^{k/2-1}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} & \text{pro} \quad x \in \langle 0, \infty), \\ 0 & \text{pro} \quad x \in (-\infty, 0), \end{cases}$$

kde

$$\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}\,\mathrm{d}t, \quad z > 0,$$

je tzv. *gama funkce*, střední hodnotu $E(X) = k$, rozptyl $D(X) = 2k$ a koeficient šikmosti $A(X) = 4/\sqrt{2k}$.

Studentovo rozdělení $S(k)$ s k stupni volnosti, kde $k \in \mathbb{N}$, má hustotu pravděpodobnosti

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad x \in (-\infty, \infty),$$

střední hodnotu $E(X) = 0$ pro $k > 1$, rozptyl $D(X) = k/(k-2)$ pro $k > 2$ a koeficient šikmosti $A(X) = 0$ pro $k > 3$. Pro jeho kvantily platí $t_{1-\alpha} = -t_\alpha$, $\alpha \in (0, 1)$.

Fisherovo-Snedecorovo rozdělení $F(k_1, k_2)$ s k_1, k_2 stupni volnosti, kde $k_1, k_2 \in \mathbb{N}$, má hustotu pravděpodobnosti

$$f(x) = \begin{cases} \dfrac{\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} x^{\frac{k_1}{2}-1}\left(1+\frac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}}}{B\left(\frac{k_1}{2},\frac{k_2}{2}\right)} & \text{pro} \quad x \in \langle 0, \infty), \\ 0 & \text{pro} \quad x \in (-\infty, 0), \end{cases}$$

kde

$$B(z_1, z_2) = \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\mathrm{d}t = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}, \quad z_1 > 0, z_2 > 0,$$

je tzv. *beta funkce*, střední hodnotu $E(X) = \frac{k_2}{k_2-2}$ pro $k_2 > 2$ a rozptyl $D(X) = \frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}$ pro $k_2 > 4$. Pro jeho kvantily platí $F_{1-\alpha}(k_1, k_2) = 1/F_\alpha(k_2, k_1)$, $\alpha \in (0, 1)$.

Multinomické rozdělení $M(n, p_1, \ldots, p_m)$, kde $n \in \mathbb{N}$, $p_1, \ldots, p_m > 0$, $\sum_{i=1}^{m} p_i < 1$, má sdruženou pravděpodobnostní funkci

$$p(x_1, x_2, ..., x_m) = \begin{cases} \dfrac{n! p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}\left(1-\sum_{i=1}^{m} p_i\right)^{n-\sum_{i=1}^{m} x_i}}{x_1! x_2! \cdots x_m!\left(n-\sum_{i=1}^{m} x_i\right)!} & \text{pro } x_1, \ldots, x_m = 1, \ldots, n, \ \sum_{i=1}^{m} x_i \leqq n \\ 0 & \text{jinak,} \end{cases}$$

střed $(E(X_1), \ldots, E(X_m)) = (np_1, \ldots, np_m)$ a kovarianční matici

$$\mathbf{cov}(X_1, \ldots, X_m) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_m \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_2p_m \\ \vdots & & \ddots & \vdots \\ -np_mp_1 & -np_mp_2 & \cdots & np_m(1-p_m) \end{pmatrix}.$$

Pro jediné $m = 1$ se $M(n, p)$ nazývá *binomické rozdělení* $Bi(n, p)$.

Weibullovo rozdělení $W(b, c, \delta)$, kde $b, \delta > 0$, $c \in \mathbb{R}$, má hustotu pravděpodobnosti

$$f(x) = \frac{b}{\delta}\left(\frac{x-c}{\delta}\right)^{b-1} \exp\left[-\left(\frac{x-c}{\delta}\right)^{b}\right], \quad x \in \langle c, \infty),$$

střední hodnotu $E(X) = c + \delta\Gamma\left(1/b+1\right)$, rozptyl $D(X) = \delta^2\left[\Gamma\left(2/b+1\right) - \Gamma^2\left(1/b+1\right)\right]$ a koeficient šikmosti

$$A(X) = \frac{\Gamma\left(3/b+1\right) - 3\Gamma\left(1/b+1\right)\Gamma\left(2/b+1\right) + 2\Gamma^3\left(1/b+1\right)}{\left[\Gamma\left(2/b+1\right) - \Gamma^2\left(1/b+1\right)\right]^{3/2}}.$$

Přitom b je *parametr tvaru*, c je *prahový parametr* a δ je *parametr měřítka*. Pro $b \approx 3{,}6$ je Weibullovo rozdělení blízké normálnímu rozdělení.

Pro $b = 1$ se $W(1, c, \delta)$ nazývá *exponenciální rozdělení* $E(\lambda, c)$.

2.4. Odhady parametrů a testování hypotéz

Odhadem T parametru θ je statistika $T(X_1, \ldots, X_n)$, která na celém parametrickém prostoru nabývá hodnot blízkých parametru θ. Odhad T parametru θ je *nestranný (nevychýlený)*, jestliže $E(T) = \theta$. V opačném případě je odhad *stranný (vychýlený)*.

Bodovým odhadem parametru θ je pozorovaná hodnota $t = T(x_1, \ldots, x_n)$ odhadu T. *Interval spolehlivosti* nebo také *konfidenční interval* pro parametr θ se spolehlivostí $1 - \alpha$, kde $\alpha \in \langle 0; 1\rangle$, je dvojice takových statistik T_1, T_2, že $P(T_1 \leqq \theta \leqq T_2) = 1 - \alpha$ pro libovolnou hodnotu parametru θ. *Intervalový odhad parametru θ* se spolehlivostí $1 - \alpha$ je pak interval $\langle t_1; t_2\rangle$, kde t_1, t_2 jsou pozorované hodnoty statistik T_1, T_2. Intervalové odhady dělíme na *oboustranné* a *jednostranné* podle toho, zda je ohraničujeme oboustranně nebo jednostranně.

Statistická hypotéza H je tvrzení o vlastnostech rozdělení pravděpodobnosti pozorované náhodné veličiny X s distribuční funkcí $F(x, \theta)$. Platnost hypotézy ověřujeme postupem, který se nazývá *test statistické hypotézy*. Proti testované (nulové) hypotéze H stavíme tzv. *alternativní hypotézu* $\overline{H}$, kterou volíme dle požadavků úlohy. Hypotéza je *jednoduchá*, jestliže uvažujeme jedinou hypotetickou hodnotu, v opačném případě je hypotéza *složená*. Složená hypotéza může být *jednostranná* (např. $H: \theta < \theta_0$) nebo *oboustranná* (např. $H: \theta \neq \theta_0$). *Parametrická hypotéza* je tvrzení o parametrech pozorované náhodné veličiny, *neparametrická hypotéza* je tvrzení o kvalitativních vlastnostech pozorované náhodné veličiny.

Testové kritérium je vhodná statistika $T(X_1, \ldots, X_n)$ zkonstruovaná pro test dané hypotézy H proti dané alternativní hypotéze $\overline{H}$. Obor hodnot testového kritéria T se za předpokladu platnosti hypotézy H rozdělí na dvě disjunktní podmnožiny, a to *kritický obor* W_α a jeho doplněk $\overline{W}_\alpha$. Kritický obor W_α se vzhledem k alternativní hypotéze stanoví tak, aby pravděpodobnost, že testové kritérium T nabyde hodnotu z kritického oboru, byla nejvýše α. Číslo $\alpha > 0$ nazýváme *hladina významnosti testu*. O hypotéze rozhodujeme na základě *pozorované hodnoty testového kritéria* $t = T(x_1, \ldots, x_n)$. Jestliže $t \in W_\alpha$, zamítáme hypotézu H a současně nezamítáme alternativní hypotézu $\overline{H}$ na hladině významnosti α. Jestliže naopak $t \in \overline{W}_\alpha$, nezamítáme hypotézu H a současně zamítáme alternativní hypotézu $\overline{H}$ na hladině významnosti α.

Chyba prvního druhu nastane, jestliže zamítneme platnou hypotézu. Její pravděpodobnost je hladina významnosti testu α. *Chyba druhého druhu nastane*, jestliže nezamítneme neplatnou hypotézu. Pravděpodobnost této chyby značíme β a číslo $1 - \beta$ nazýváme *silou testu*.

K testování statistických hypotéz lze rovněž použít přímo intervalové odhady. Při testování na hladině významnosti α pak místo testového kritéria zvolíme vhodný intervalový odhad se spolehlivostí $1 - \alpha$.

Při testování statistických hypotéz lze také místo kritického oboru použít tzv. *p-hodnotu*. Pro pozorovanou hodnotu t testového kritéria T je p-hodnotou číslo $1 - P(-t \leqq T \leqq t)$. Jestliže $p < \alpha$, zamítáme hypotézu H a současně nezamítáme alternativní hypotézu $\overline{H}$ na hladině významnosti α. Jestliže naopak $p \geqq \alpha$, nezamítáme hypotézu H a současně zamítáme alternativní hypotézu $\overline{H}$ na hladině významnosti α.

2.5. Lineární regresní analýza

Regresní analýza zkoumá závislost závisle proměnné náhodné veličiny Y na nezávisle proměnném náhodném vektoru $\boldsymbol{X} = (X_1, \ldots, X_k)$. *Základní lineární regresní model* uvažujeme ve tvaru

$$Y_i = \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \quad i = 1, \ldots, n,$$

vektorově zapisujeme

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} X_{11} & \cdots & X_{k1} \\ \vdots & \ddots & \vdots \\ X_{1n} & \cdots & X_{kn} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{pmatrix},$$

přičemž $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ a $\mathbf{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$, kde $\sigma^2 > 0$, takže složky vektoru $\boldsymbol{\varepsilon}$ jsou nekorelované a mají stejný rozptyl σ^2. Platí

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}, \quad \mathbf{cov}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}.$$

Nejlepší nestranný odhad vektoru *regresních koeficientů* $\boldsymbol{\beta}$ je vektor $\boldsymbol{b}$, který získáme *metodou nejmenších čtverců*, tj. minimalizací *reziduálního součtu čtverců*

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Jestliže matice $\boldsymbol{X}$ má *plnou hodnost* $k < n$, existuje jediný vektor $\boldsymbol{b}$, který je řešením *soustavy normálních rovnic*

$$\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T \boldsymbol{Y},$$

kde matice $\boldsymbol{X}^T \boldsymbol{X}$ je regulární, a to

$$\boldsymbol{b} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

Platí

$$E(\boldsymbol{b}) = \boldsymbol{\beta}, \quad \mathbf{cov}(\boldsymbol{b}) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}.$$

Vektor bodových odhadů $\widehat{Y}_i$ hodnot (středních hodnot) Y_i, $i = 1, \ldots, n$, je

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

Bodový odhad rozptylu σ^2 je

$$s^2 = \frac{\boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{b}^T \boldsymbol{X}^T \boldsymbol{Y}}{n-k} = \frac{\boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{Y}^T \boldsymbol{X}\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}}{n-k} = \frac{\boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{Y}^T \widehat{\boldsymbol{Y}}}{n-k}. \tag{2.1}$$

Kapitola 3

Základní metoda bootstrap — přesnost odhadu

3.1. Střední kvadratická chyba a tolerance chyby odhadu

Nechť X je náhodná veličina a nechť θ je parametr jejího rozdělení pravděpodobnosti. Realizujeme náhodný výběr $(X_1, \ldots, X_n)$ z této náhodné veličiny o rozsahu n. Na základě pozorovaných hodnot vypočítáme odhad parametru θ a označíme jej $\widehat{\theta}$. Pak *střední kvadratickou chybou odhadu* $\widehat{\theta}$ rozumíme hodnotu

$$\mathrm{MSE} = E(\widehat{\theta} - \theta)^2.$$

Důležitost střední kvadratické chyby vyplývá z následující *Čebyševovy-Markovovy nerovnosti*:

$$P\left(|\widehat{\theta} - \theta| \leqq k\sqrt{\mathrm{MSE}}\right) \geqq 1 - \frac{1}{k^2} \text{ pro libovolné } k.$$

Pokud např. zvolíme $k = 2$, dostáváme

$$P\left(\theta - 2\sqrt{\mathrm{MSE}} \leqq \widehat{\theta} \leqq \theta + 2\sqrt{\mathrm{MSE}}\right) \geqq 0{,}75.$$

Číslo $2\sqrt{\mathrm{MSE}}$ nazýváme *tolerancí chyby odhadu* a používáme ho jako hrubou míru přesnosti odhadu v případech, kdy žádnou vhodnější míru nemáme k dispozici.

Vychýlením odhadu $\widehat{\theta}$ parametru θ rozumíme $\mathrm{Bias}\left(\widehat{\theta}\right) = E\left(\widehat{\theta}\right) - \theta$. Snadno se ukáže, že platí $\mathrm{MSE} = D\left(\widehat{\theta}\right) + \left(\mathrm{Bias}\left(\widehat{\theta}\right)\right)^2$.

Pokud například odhadujeme střední hodnotu rozdělení $E(X)$ výběrovým průměrem $\overline{X}$, můžeme využít faktu, že rozptyl výběrového průměru je

$$D(\overline{X}) = \frac{D(X)}{n}.$$

Odhadneme-li dále rozptyl rozdělení $D(X)$ výběrovým rozptylem S^2, pak s využitím centrální limitní věty dostáváme asymptotický konfidenční interval se spolehlivostí 0,95:

$$E(X) \in \left(\overline{X} - 2\frac{S}{\sqrt{n}}; \overline{X} + 2\frac{S}{\sqrt{n}}\right).$$

Pak $2Sn^{-1/2}$ je v tomto případě tolerancí chyby odhadu $\overline{X}$.

V mnoha případech ovšem neznáme analytickou metodu výpočtu střední kvadratické chyby a tolerance chyby odhadu. Představme si ale, že bychom mohli mnohokrát opakovat realizaci náhodného výběru o rozsahu n z náhodné veličiny X. Označme $\widehat{\theta}_i$ odhad parametru θ vypočítaný z pozorovaných hodnot náhodného výběru při i-tém opakování a MSE_i střední kvadratickou chybu tohoto odhadu. Pak při dostatečně velkém počtu opakování můžeme střední kvadratickou chybu odhadu $\widehat{\theta}$ odhadnout

$$\widehat{\mathrm{MSE}} = \frac{1}{B}\sum_{i=1}^{B}\mathrm{MSE}_i = \frac{1}{B}\sum_{i=1}^{B}\left(\widehat{\theta}_i - \theta\right)^2,$$

kde číslo B označuje počet provedených opakování. Takovéto opakování obvykle není ve skutečnosti možné, a proto musíme přistoupit k další aproximaci výpočtu odhadu MSE.

3.2. Bootstrapový odhad střední kvadratické chyby, rozptylu, směrodatné odchylky a vychýlení

Pokud neznáme rozdělení pravděpodobnosti pozorované náhodné veličiny X nebo není k dispozici intervalový odhad jejího parametru θ, nahradíme soubor pozorovaných hodnot $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ novým souborem získaným z $(x_1, \ldots, x_n)$ náhodným výběrem s opakováním (s vracením). Takto získaný náhodný výběr nazýváme *bootstrapovým výběrem*.

Při odhadu střední kvadratické chyby, rozptylu, směrodatné odchylky a vychýlení odhadu $\widehat{\theta}$ parametru θ rozdělení pravděpodobnosti náhodné veličiny X postupujeme následovně:

1. Z pozorovaných hodnot $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ vypočítáme odhad $\widehat{\theta}$ parametru θ.
2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n z pozorovaných hodnot $(x_1, \ldots, x_n)$. Obvykle volíme $B \geqq 1000$.
3. Pro každý bootstrapový výběr vypočítáme odhad parametru θ a označíme $\widehat{\theta}_{b,i}$, kde $i = = 1, 2, \ldots, B$.
4. *Bootstrapovým odhadem parametru* θ rozumíme obvykle aritmetický průměr

$$\widehat{\theta}_b = \frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_{b,i}.$$

5. *Bootstrapovým odhadem střední kvadratické chyby* MSE odhadu $\widehat{\theta}$ je

$$\widehat{\mathrm{MSE}}_b = \frac{1}{B}\sum_{i=1}^{B}\left(\widehat{\theta}_{b,i} - \widehat{\theta}\right)^2.$$

6. *Bootstrapovým odhadem rozptylu* $D\left(\widehat{\theta}\right)$ je

$$\widehat{D}\left(\widehat{\theta}\right)_b = \frac{1}{B-1}\sum_{i=1}^{B}\left(\widehat{\theta}_{b,i} - \frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_{b,i}\right)^2.$$

7. *Bootstrapovým odhadem směrodatné odchylky* $\sigma\left(\widehat{\theta}\right)$ je

$$\widehat{\sigma}\left(\widehat{\theta}\right)_b = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\widehat{\theta}_{b,i} - \frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_{b,i}\right)^2}. \tag{3.1}$$

8. *Bootstrapovým odhadem vychýlení* $\mathrm{Bias}\left(\widehat{\theta}\right)$ odhadu $\widehat{\theta}$ je

$$\widehat{\mathrm{B}}\left(\widehat{\theta}\right)_b = \frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_{b,i} - \widehat{\theta}.$$

3.3. Parametrický bootstrap

Na metodu bootstrap lze také pohlížet *parametricky*. V tom případě předpokládáme, že známe rozdělení pravděpodobnosti pozorované náhodné veličiny X. Z pozorovaných hodnot $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ odhadneme všechny potřebné parametry její distribuční funkce a bootstrapové náhodné výběry nyní realizujeme jako náhodné výběry z takto odhadnuté distribuční funkce. Všechny následné výpočty se pak už neliší od těch popsaných výše.

Kapitola 4

Intervalové odhady parametrů rozdělení pravděpodobnosti

4.1. Pivotové odhady

Nechť Z je spojitá náhodná veličina se střední hodnotou $E(Z) = 0$, rozptylem $D(Z) = 1$ a hustotou pravděpodobnosti $f(z)$. Nechť X je spojitá náhodná veličina daná vztahem

$$X = \mu + \sigma Z, \quad \text{kde} \quad \sigma > 0,$$

tedy má hustotu pravděpodobnosti

$$g(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right).$$

Potom střední hodnota $E(X) = \mu$, rozptyl $D(X) = \sigma^2$ a směrodatná odchylka $\sigma(X) = \sigma$.

V této podkapitole ukážeme, jak získat metodou bootstrap odhad konfidenčního intervalu pro odhady střední hodnoty μ, rozptylu σ^2 a směrodatné odchylky σ. Přitom budeme používat následující odhady parametrů: μ odhadneme výběrovým průměrem $\overline{X}$ a σ odhadneme výběrovou směrodatnou odchylkou S. Více viz [3]

V následujícím budeme ukazovat konfidenční intervaly se spolehlivostí $1 - 2\alpha$ získané pomocí α-kvantilů a $(1 - \alpha)$-kvantilů příslušných rozdělení pravděpodobnosti. Pochopitelně není nutné volit právě tyto kvantily. Konfidenční interval se spolehlivostí ξ získáme pomocí libovolné dvojice ζ-kvantil, η-kvantil, kde $\eta - \zeta = \xi$ a $\xi, \zeta, \eta \in \langle 0; 1 \rangle$.

4.1.1. Intervalový odhad střední hodnoty

Pokud Z má normované normální rozdělení pravděpodobnosti $N(0; 1)$, pak statistika

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

má Studentovo rozdělení pravděpodobnosti s $n - 1$ stupni volnosti a platí

$$P\left(-t_{1-\alpha} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{1-\alpha}\right) = 1 - 2\alpha,$$

kde $t_{1-\alpha}$ je $(1-\alpha)$-kvantil Studentova rozdělení s $n-1$ stupni volnosti. Odtud odvodíme konfidenční interval se spolehlivostí $1-2\alpha$

$$\mu \in \left(\overline{X} - t_{1-\alpha} \frac{S}{\sqrt{n}}; \overline{X} + t_{1-\alpha} \frac{S}{\sqrt{n}} \right).$$

Pokud Z nemá normální rozdělení pravděpodobnosti, rozdělení pravděpodobnosti statistiky t je stále nezávislé na μ i σ, ale už se nejedná o Studentovo rozdělení. Nicméně pokud bychom mohli nějakým způsobem zjistit hodnoty kvantilů tohoto neznámého rozdělení, stále by platilo

$$P\left(t_\alpha < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{1-\alpha} \right) = 1 - 2\alpha$$

a konfidenční interval se spolehlivostí $1-2\alpha$ by měl tvar

$$\mu \in \left(\overline{X} - t_{1-\alpha} \frac{S}{\sqrt{n}}; \overline{X} - t_\alpha \frac{S}{\sqrt{n}} \right).$$

Hodnoty kvantilů rozdělení pravděpodobnosti statistiky t odhadneme pomocí metody bootstrap.

Postup pro získání konfidenčního intervalu pro $\mu = E(X)$ bude tedy následující:

1. Z pozorovaných hodnot $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ vypočítáme pozorované hodnoty výběrového průměru $\overline{X}$ a výběrové směrodatné odchylky S.
2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n z pozorovaných hodnot $(x_1, \ldots, x_n)$. Obvykle volíme $B \geqq 1000$.
3. Pro každý bootstrapový výběr vypočítáme pozorovanou hodnotu výběrového průměru $\overline{X}_{b,i}$ a výběrové směrodatné odchylky $S_{b,i}$ a hodnotu statistiky t

$$t_{b,i} = \frac{\overline{X}_{b,i} - \overline{X}}{S_{b,i}/\sqrt{n}},$$

   kde $i = 1, 2, \ldots, B$.
4. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky t_b odhadneme hodnotami $t_{b,\alpha}$ a $t_{b,1-\alpha}$ splňujícími co nejpřesněji

$$\left| \{ t_{b,i};\, t_{b,i} \leqq t_{b,\alpha} \} \right| / B \doteq \alpha, \quad \left| \{ t_{b,i};\, t_{b,i} \leqq t_{b,1-\alpha} \} \right| / B \doteq 1 - \alpha,$$

   kde $\left| \{\ldots\} \right|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.
5. *Bootstrapovým konfidenčním intervalem se spolehlivostí* $1-2\alpha$ *pro střední hodnotu* $E(X)$ je

$$\mu \in \left(\overline{X} - t_{b,1-\alpha} \frac{S}{\sqrt{n}}; \overline{X} - t_{b,\alpha} \frac{S}{\sqrt{n}} \right).$$

4.1.2. Intervalový odhad rozptylu a směrodatné odchylky

Pokud Z má normované normální rozdělení pravděpodobnosti $N(0; 1)$, pak statistika

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

má Pearsonovo rozdělení pravděpodobnosti s $n-1$ stupni volnosti a platí

$$P\left(\chi^2_\alpha < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{1-\alpha}\right) = 1-2\alpha,$$

kde χ^2_α a $\chi^2_{1-\alpha}$ jsou α-kvantil a $(1-\alpha)$-kvantil Pearsonova rozdělení s $n-1$ stupni volnosti. Odtud odvodíme konfidenční interval se spolehlivostí $1-2\alpha$

$$\sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{1-\alpha}}; \frac{(n-1)S^2}{\chi^2_\alpha}\right).$$

Pokud Z nemá normální rozdělení pravděpodobnosti, rozdělení pravděpodobnosti statistiky χ^2 je stále nezávislé na μ i σ, ale už se nejedná o Pearsonovo rozdělení. Nicméně pokud bychom mohli nějakým způsobem zjistit hodnoty kvantilů tohoto neznámého rozdělení, výše uvedené rovnosti by stále platily. Hodnoty kvantilů rozdělení pravděpodobnosti statistiky χ^2 odhadneme pomocí metody bootstrap.

Postup pro získání konfidenčního intervalu pro $\sigma^2 = D(X)$ a $\sigma = \sigma(X)$ bude tedy následující:

1. Z pozorovaných hodnot $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ vypočítáme pozorovanou hodnotu výběrového rozptylu S^2.
2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n z pozorovaných hodnot $(x_1, \ldots, x_n)$. Obvykle volíme $B \geqq 1000$.
3. Pro každý bootstrapový výběr vypočítáme pozorovanou hodnotu výběrového rozptylu $S^2_{b,i}$ a hodnotu statistiky χ^2
$$\chi^2_{b,i} = \frac{(n-1)S^2_{b,i}}{S^2},$$
kde $i = 1, 2, \ldots, B$.
4. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky χ^2_b odhadneme hodnotami $\chi^2_{b,\alpha}$ a $\chi^2_{b,1-\alpha}$ splňujícími co nejpřesněji
$$\left|\left\{\chi^2_{b,i};\, \chi^2_{b,i} \leqq \chi^2_{b,\alpha}\right\}\right| / B \doteq \alpha, \quad \left|\left\{\chi^2_{b,i};\, \chi^2_{b,i} \leqq \chi^2_{b,1-\alpha}\right\}\right| / B \doteq 1-\alpha,$$
kde $\left|\{\ldots\}\right|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.
5. *Bootstrapovým konfidenčním intervalem se spolehlivostí* $1-2\alpha$ *pro rozptyl* $D(X)$ je
$$\sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{b,1-\alpha}}; \frac{(n-1)S^2}{\chi^2_{b,\alpha}}\right).$$
6. *Bootstrapovým konfidenčním intervalem se spolehlivostí* $1-2\alpha$ *pro směrodatnou odchylku* $\sigma(X)$ je
$$\sigma \in \left(\sqrt{\frac{(n-1)S^2}{\chi^2_{b,1-\alpha}}}; \sqrt{\frac{(n-1)S^2}{\chi^2_{b,\alpha}}}\right).$$

4.1.3. Obecný pivotový konfidenční interval

Výše uvedené konfidenční intervaly jsou konkrétními příklady obecnějších formulí. Zcela obecné konfidenční intervaly pro libovolné parametry rozdělení pravděpodobnosti odhadované libovolnými statistikami uvádí [2].

Nechť θ je libovolný parametr (popř. parametrická funkce) rozdělení pravděpodobnosti náhodné veličiny X a nechť $\widehat{\theta}$ je nějakým jeho odhadem. Nechť $\widehat{\sigma}\left(\widehat{\theta}\right)_b$ je bootstrapovým odhadem směrodatné odchylky odhadu $\widehat{\theta}$ podle (3.1). Pokud u náhodné veličiny $\widehat{\theta}_b$, tj. odhadu parametru θ na základě bootstrapových náhodných výběrů, můžeme předpokládat normální rozdělení pravděpodobnosti, pak statistika

$$t = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}\left(\widehat{\theta}\right)_b}$$

má Studentovo rozdělení pravděpodobnosti s $n - 1$ stupni volnosti a platí

$$P\left(-t_{1-\alpha} < \frac{\widehat{\theta} - \theta}{\widehat{\sigma}\left(\widehat{\theta}\right)_b} < t_{1-\alpha}\right) = 1 - 2\alpha,$$

kde $t_{1-\alpha}$ je $(1-\alpha)$-kvantil Studentova rozdělení pravděpodobnosti s $n - 1$ stupni volnosti. Odtud snadno odvodíme bootstrapový konfidenční interval se spolehlivostí $1 - 2\alpha$ pro odhad parametru θ

$$\theta \in \left(\widehat{\theta} - t_{1-\alpha}\widehat{\sigma}\left(\widehat{\theta}\right)_b; \widehat{\theta} + t_{1-\alpha}\widehat{\sigma}\left(\widehat{\theta}\right)_b\right).$$

V případě, kdy u náhodné veličiny $\widehat{\theta}_b$ nelze předpokládat normální rozdělení, musíme hodnoty kvantilů rozdělení pravděpodobnosti statistiky t opět odhadnout metodou bootstrap.

Postup pro získání konfidenčního intervalu pro parametr θ bude tedy následující:

1. Z pozorovaných hodnot $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ vypočítáme odhad $\widehat{\theta}$ parametru θ.
2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n z pozorovaných hodnot $(x_1, \ldots, x_n)$. Obvykle volíme $B \geqq 1000$.
3. Pro každý bootstrapový výběr vypočítáme odhad $\widehat{\theta}_{b,i}$ parametru θ a jeho směrodatnou odchylku $\sigma(\widehat{\theta}_{b,i})$. Pokud pro ni neznáme žádné analytické vyjádření, odhadneme ji vnořenou metodou bootstrap podle (3.1). Obvykle volíme $B \geqq 100$.
4. Pro každý bootstrapový výběr dále vypočítáme hodnotu statistiky

$$t_{b,i} = \frac{\widehat{\theta}_{b,i} - \widehat{\theta}}{\sigma\left(\widehat{\theta}_{b,i}\right)},$$

   kde $i = 1, 2, \ldots, B$.
5. Podle (3.1) spočítáme odhad směrodatné odchylky $\widehat{\sigma}\left(\widehat{\theta}\right)_b$ odhadu $\widehat{\theta}$.
6. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky t_b odhadneme hodnotami $t_{b,\alpha}$ a $t_{b,1-\alpha}$ splňujícími co nejpřesněji

$$\left|\{t_{b,i};\, t_{b,i} \leqq t_{b,\alpha}\}\right| / B \doteq \alpha, \quad \left|\{t_{b,i};\, t_{b,i} \leqq t_{b,1-\alpha}\}\right| / B \doteq 1 - \alpha,$$

   kde $\left|\{\ldots\}\right|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.
7. *Bootstrapovým konfidenčním intervalem se spolehlivostí* $1 - 2\alpha$ *pro parametr* θ je

$$\theta \in \left(\widehat{\theta} - t_{b,1-\alpha}\widehat{\sigma}\left(\widehat{\theta}\right)_b; \widehat{\theta} - t_{b,\alpha}\widehat{\sigma}(\widehat{\theta})_b\right).$$

4.2. Kvantilové odhady

V této podkapitole se seznámíme s intervalovými odhady, které vycházejí přímo z rozdělení pravděpodobnosti bodových odhadů. Jsou proto zcela obecné, použitelné pro libovolný parametr, příp. parametrickou funkci, a pro libovolný jeho odhad.

4.2.1. Jednoduchý kvantilový konfidenční interval

Postup pro získání jednoduchého kvantilového konfidenčního intervalu je následující:

1. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n z pozorovaných hodnot $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$. Obvykle volíme $B \geqq 1000$.
2. Pro každý bootstrapový výběr vypočítáme odhad $\widehat{\theta}_{b,i}$ parametru θ.
3. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky $\widehat{\theta}_b$ odhadneme hodnotami $\widehat{\theta}_{b,\alpha}$ a $\widehat{\theta}_{b,1-\alpha}$ splňujícími co nejpřesněji
$$\left|\left\{\widehat{\theta}_{b,i};\, \widehat{\theta}_{b,i} \leqq \widehat{\theta}_{b,\alpha}\right\}\right| / B \doteq \alpha, \quad \left|\left\{\widehat{\theta}_{b,i};\, \widehat{\theta}_{b,i} \leqq \widehat{\theta}_{b,1-\alpha}\right\}\right| / B \doteq 1-\alpha,$$
kde $\left|\{\ldots\}\right|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.
4. *Bootstrapovým jednoduchým kvantilovým konfidenčním intervalem se spolehlivostí* $1-2\alpha$ *pro parametr* θ je
$$\theta \in \left\langle \widehat{\theta}_{b,\alpha};\, \widehat{\theta}_{b,1-\alpha} \right\rangle.$$

4.2.2. Reziduový kvantilový konfidenční interval

Reziduem rozumíme $\varepsilon = \widehat{\theta} - \theta$. Označíme α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti náhodné veličiny ε jako ε_α a $\varepsilon_{1-\alpha}$. Pak platí
$$P\left(\varepsilon_\alpha < \widehat{\theta} - \theta \leqq \varepsilon_{1-\alpha}\right) = 1 - 2\alpha.$$
Odtud odvodíme konfidenční interval se spolehlivostí $1-2\alpha$
$$\theta \in \left\langle \widehat{\theta} - \varepsilon_{1-\alpha};\, \widehat{\theta} - \varepsilon_\alpha \right).$$
Kvantily rozdělení pravděpodobnosti rezidua ovšem neznáme a odhadneme je metodou bootstrap.

Postup pro získání reziduového kvantilového konfidenčního intervalu je následující:

1. Z pozorovaných hodnot $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ vypočítáme odhad $\widehat{\theta}$ parametru θ.
2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n z pozorovaných hodnot $(x_1, \ldots, x_n)$. Obvykle volíme $B \geqq 1000$.
3. Pro každý bootstrapový výběr vypočítáme odhad $\widehat{\theta}_{b,i}$ parametru θ a reziduum $e_{b,i} = \widehat{\theta}_{b,i} - \widehat{\theta}$.
4. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti reziduí e_b odhadneme hodnotami $e_{b,\alpha}$ a $e_{b,1-\alpha}$ splňujícími co nejpřesněji
$$\left|\{e_{b,i};\, e_{b,i} \leqq e_{b,\alpha}\}\right| / B \doteq \alpha, \quad \left|\{e_{b,i};\, e_{b,i} \leqq e_{b,1-\alpha}\}\right| / B \doteq 1-\alpha,$$
kde $\left|\{\ldots\}\right|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.

5. *Bootstrapovým reziduovým kvantilovým konfidenčním intervalem se spolehlivostí* $1-2\alpha$ *pro parametr* θ je
$$\theta \in \langle \widehat{\theta} - e_{b,1-\alpha}; \widehat{\theta} - e_{b,\alpha} \rangle .$$

4.2.3. BCA kvantilový konfidenční interval

Může se stát, že jednoduché a reziduové kvantilové konfidenční intervaly jsou vychýlené nebo příliš široké oproti hodnotám pozorovaným v praxi. K odstranění těchto nedostatků proto konstruujeme tzv. *BCA konfidenční intervaly* (z anglického bias corrected and accelerated), které jsou sice také omezeny dvěma kvantily rozdělení pravděpodobnosti bootstrapového odhadu $\widehat{\theta}_b$, ale na rozdíl od předchozích metod se již nemusí nutně jednat o α-kvantil a $(1-\alpha)$-kvantil pro spolehlivost $1-2\alpha$.

V této metodě vycházíme z předpokladu, že existuje nějaká transformace parametru θ, jejíž rozdělení pravděpodobnosti je normální a jejíž střední hodnota a rozptyl závisejí na θ. Konfidenční interval pak zkonstruujeme pro transformovaný parametr a inverzní transformací jeho mezí získáme konfidenční interval pro θ. Elegance metody spočívá v tom, že předpokládanou transformaci vůbec nepotřebujeme znát v explicitním vyjádření, realizujeme ji metodou bootstrap.

Předpokládejme, že existuje rostoucí transformační zobrazení T takové, že $T(\widehat{\theta})$ má normální rozdělení pravděpodobnosti se střední hodnotou

$$E\left[T\left(\widehat{\theta}\right)\right] = T(\theta) - z_0\left[1 + aT(\theta)\right]$$

a směrodatnou odchylkou

$$\sigma\left[T\left(\widehat{\theta}\right)\right] = 1 + aT(\theta).$$

Nechť $z_{1-\alpha}$ je $(1-\alpha)$-kvantil normovaného normálního rozdělení pravděpodobnosti. Pak

$$P\left(-z_{1-\alpha} < \frac{T\left(\widehat{\theta}\right) - T(\theta)}{1 + aT(\theta)} + z_0 < z_{1-\alpha}\right) = 1 - 2\alpha,$$

odkud snadno odvodíme konfidenční interval se spolehlivostí $1-2\alpha$

$$T(\theta) \in \left(\frac{T\left(\widehat{\theta}\right) + z_0 - z_{1-\alpha}}{1 - a(z_0 - z_{1-\alpha})}; \frac{T\left(\widehat{\theta}\right) + z_0 + z_{1-\alpha}}{1 - a(z_0 + z_{1-\alpha})}\right).$$

Vzhledem k tomu, že náhodné veličiny

$$\frac{T\left(\widehat{\theta}_b\right) - T\left(\widehat{\theta}\right)}{1 + aT\left(\widehat{\theta}\right)} + z_0 \quad \text{a} \quad \frac{T\left(\widehat{\theta}\right) - T(\theta)}{1 + aT(\theta)} + z_0$$

mají stejné rozdělení pravděpodobnosti (podle předpokladu normované normální), platí

$$P\left(T\left(\widehat{\theta}_b\right) < \frac{T\left(\widehat{\theta}\right) + z_0 + z_{1-\alpha}}{1 - a(z_0 + z_{1-\alpha})}\right) = P\left(\frac{T\left(\widehat{\theta}_b\right) - T\left(\widehat{\theta}\right)}{1 + aT\left(\widehat{\theta}\right)} + z_0 < \frac{z_0 + z_{1-\alpha}}{1 - a(z_0 + z_{1-\alpha})} + z_0\right) =$$

$$= P\left(Z < \frac{z_0 + z_{1-\alpha}}{1 - a(z_0 + z_{1-\alpha})} + z_0\right),$$

kde Z má normované normální rozdělení pravděpodobnosti. Odtud vyplývá, že horní mez konfidenčního intervalu pro $T(\theta)$ je

$$z_H = \frac{z_0 + z_{1-\alpha}}{1 - a(z_0 + z_{1-\alpha})} + z_0.$$

Obdobně se odvodí dolní mez

$$z_D = \frac{z_0 - z_{1-\alpha}}{1 - a(z_0 - z_{1-\alpha})} + z_0.$$

Zbývá odhadnout hodnoty z_0 a a. Nechť p_0 je podíl pozorovaných hodnot rozdělení pravděpodobnosti náhodné veličiny $\widehat{\theta}_b$ s vlastností $\widehat{\theta}_b \leqq \widehat{\theta}$ ku všem pozorovaným hodnotám. Pak z_0 má takovou hodnotu, že platí $P(Z \leqq z_0) = p_0$, kde Z má normované normální rozdělení pravděpodobnosti. Z toho vyplývá, že z_0 *koriguje vychýlení mediánu* bootstrapového odhadu.

Akcelerace a měří rychlost změny $\widehat{\sigma}\left(\widehat{\theta}\right)_b$ (viz rovnice (3.1)) v závislosti na změně skutečné hodnoty parametru θ. Uvádí se více různých odhadů pro a, nejčastěji však následující založený na míře špičatosti rozdělení náhodné veličiny X.

Označme $\widehat{\theta}_{-i}$ odhad parametru θ spočítaný s vynecháním X_i, tj. z náhodného výběru $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Označme

$$\widehat{\theta}_{(\ )} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{-i}.$$

Pak

$$a = \frac{\sum_{i=1}^{n} \left(\widehat{\theta}_{(\ )} - \widehat{\theta}_{-i}\right)^3}{\sigma \left[\sum_{i=1}^{n} \left(\widehat{\theta}_{(\ )} - \widehat{\theta}_{-i}\right)^2\right]^{3/2}}. \tag{4.1}$$

Postup pro získání BCA konfidenčního intervalu pro parametr θ je tedy následující:

1. Z pozorovaných hodnot $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ vypočítáme odhad $\widehat{\theta}$ parametru θ.
2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n z pozorovaných hodnot $(x_1, \ldots, x_n)$. Obvykle volíme $B \geqq 1000$.
3. Pro každý bootstrapový výběr vypočítáme odhad $\widehat{\theta}_{b,i}$ parametru θ.
4. Spočítáme korekci vychýlení mediánu

$$z_0 = \Phi^{-1}\left(\frac{\left|\left\{\widehat{\theta}_{b,i}; \widehat{\theta}_{b,i} < \widehat{\theta}\right\}\right|}{B}\right),$$

   kde Φ je distribuční funkce normovaného normálního rozdělení pravděpodobnosti a $\left|\{\ldots\}\right|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.
5. Spočítáme akceleraci a podle vzorce (4.1).
6. Spočítáme

$$\alpha_1 = \Phi\left(\frac{z_0 - z_{1-\alpha}}{1 - a(z_0 - z_{1-\alpha})} + z_0\right) \quad \text{a} \quad \alpha_2 = \Phi\left(\frac{z_0 + z_{1-\alpha}}{1 - a(z_0 + z_{1-\alpha})} + z_0\right).$$

7. α_1-kvantil a $(1-\alpha_2)$-kvantil rozdělení pravděpodobnosti statistiky $\widehat{\theta}_b$ odhadneme hodnotami $\widehat{\theta}_{b,\alpha_1}$ a $\widehat{\theta}_{b,1-\alpha_2}$ splňujícími co nejpřesněji

$$\left|\left\{\widehat{\theta}_{b,i};\, \widehat{\theta}_{b,i} \leqq \widehat{\theta}_{b,\alpha_1}\right\}\right| \Big/ B \doteq \alpha_1, \quad \left|\left\{\widehat{\theta}_{b,i};\, \widehat{\theta}_{b,i} \leqq \widehat{\theta}_{b,1-\alpha_2}\right\}\right| \Big/ B \doteq 1-\alpha_2.$$

8. *BCA konfidenčním intervalem se spolehlivostí* $1-2\alpha$ *pro parametr* θ je

$$\theta \in \left(\widehat{\theta}_{b,\alpha_1};\, \widehat{\theta}_{b,1-\alpha_2}\right).$$

Jednodušší verzí BCA intervalů jsou tzv. *BC konfidenční intervaly*, u kterých se volí $a = 0$.

4.3. Testování hypotéz

Bootstrapové konfidenční intervaly mají zásadní využití také v oblasti testování hypotéz. Chceme-li testovat hypotézu $H\colon \theta = \theta_0$ na hladině významnosti 2α, sestrojíme jednoduše konfidenční interval pro θ se spolehlivostí $1-2\alpha$. Hypotézu na hladině významnosti 2α nezamítneme, pokud θ_0 bude prvkem intervalu, a zamítneme, pokud θ_0 nebude prvkem intervalu.

Kvantily ke konstrukci konfidenčního intervalu přitom zvolíme intuitivně v závislosti na tvaru alternativy. Testujeme-li např. proti alternativě $\overline{H}\colon \theta \neq \theta_0$, budeme při konstrukci konfidenčního intervalu libovolného typu volit α-kvantil a $(1-\alpha)$-kvantil. Testujeme-li proti alternativě $\overline{H}\colon \theta < \theta_0$, budeme při konstrukci reziduového kvantilového a pivotových konfidenčních intervalů volit 2α-kvantil a 1-kvantil, zatímco při konstrukci ostatních kvantilových konfidenčních intervalů $(1-2\alpha)$-kvantil a 0-kvantil. Přesně naopak pak při testování proti alternativě $\overline{H}\colon \theta > \theta_0$, tj. při konstrukci reziduového kvantilového a pivotových konfidenčních intervalů $(1-2\alpha)$-kvantil a 0-kvantil, zatímco při konstrukci ostatních kvantilových konfidenčních intervalů 2α-kvantil a 1-kvantil.

Kapitola 5

Bootstrapový výběr z časové řady

Nechť (Ω, Σ, P) je pravděpodobnostní prostor a $T \subset \mathbb{R}$. *Náhodným (stochastickým) procesem* nazveme reálnou funkci $X(\omega, t)$ definovanou na množině $\Omega \times T$ takovou, že při každém pevném $t_0 \in T$ je $X(\omega, t_0)$ náhodnou veličinou. Stručně značíme náhodný proces pouze $\{X(t)\}_{t\in T}$. Jestliže množina parametrů T je diskrétní (spočetná), je $\{X(t)\}_{t\in T}$ *náhodný proces s diskrétním časem*. Nazýváme ho též *náhodnou posloupností* nebo *časovou řadou*. Je-li množina parametrů T nespočetná, je $\{X(t)\}_{t\in T}$ *náhodný proces se spojitým časem*. Nazýváme ho též *náhodnou funkcí*.

Náhodný proces $\{X(t)\}_{t\in T}$ nazýváme *Markovův (markovský)*, jestliže pro libovolná reálná čísla a, b a pro libovolná $t_1 < \cdots < t_n < t$, kde $t_1, \ldots, t_n, t \in T$ platí tzv. *markovská vlastnost*

$$P\big(a \leqq X(t) < b \big| X(t_1) = x_1, \ldots, X(t_n) = x_n\big) = P\big(a \leqq X(t) < b \big| X(t_n) = x_n\big).$$

Chceme-li metodu bootstrap aplikovat na Markovův proces, je třeba pro bootstrapový výběr zachovat markovskou vlastnost. Toho nejlépe dosáhneme vhodným rekurentním vyjádřením, odhadem odchylek a realizací bootstrapového výběru z odchylek.

Nechť $(X_1, \ldots, X_n)$ je markovská časová řada, pro kterou platí $X_i = h(X_{i-1}) + \varepsilon_i$ pro každé $i = 2, \ldots, n$, kde odchylky $\varepsilon_2, \ldots, \varepsilon_n$ tvoří náhodný výběr z neznámé distribuční funkce s nulovou střední hodnotou. Pak bootstrapovou časovou řadu získáme takto:

1. Odhadneme funkci $h(X)$ odhadem $\widehat{h}(X)$ a pro každé $i = 2, \ldots, n$ vypočítáme odhad odchylky ε_i jako $e_i = X_i - \widehat{h}(X_{i-1})$.
2. Realizujeme bootstrapový náhodný výběr o rozsahu $n-1$ z hodnot $e_2, \ldots, e_n$ a označíme ho $e_{b,2}, \ldots, e_{b,n}$.
3. *Bootstrapová časová řada* se tvoří rekurzivně: $X_{b,1} = X_1$, $X_{b,i} = \widehat{h}(X_{i-1}) + e_{b,i}$ pro $i = 2, \ldots, n$.
4. Dále vygenerujeme B takových bootstrapových časových řad, které použijeme k získání potřebných odhadů podle postupů uvedených dříve.

V [2] se uvádí další metoda získání bootstrapové časové řady z časové řady $(X_1, \ldots, X_n)$:

1. Zvolíme vhodnou *délku bloku d* tak, že každá dvě pozorování z původní časové řady vzdálená od sebe o víc než d kroků jsou již nezávislá. Uvažujeme pak všechny *bloky* délky d, tj. $(d+1)$-tice $X_i, \ldots, X_{i+d}$. (Např. pro markovskou časovou řadu bude $d = 1$.)
2. Bootstrapovou časovou řadu získáme náhodným výběrem s opakováním z těchto bloků o rozsahu k, kde $n \doteq (d+1)k$.

Kapitola 6

Vícerozměrný bootstrapový výběr

Je známo více metod, kterými lze získat vícerozměrný bootstrapový náhodný výběr. Při jeho realizaci obvykle nemůžeme postupovat u každé složky samostatně, protože chceme zachovat případnou závislost (kovarianci, korelaci), která mezi složkami může existovat. Pouze pokud předpokládáme nezávislost složek nebo případná závislost nemá vliv na statistiky, které nás dále zajímají, můžeme si dovolit samostatný bootstrapový výběr pro každou složku. Uvedeme si dva základní přístupy k realizaci vícerozměrného bootstrapového výběru.

6.1. Vícerozměrný bootstrapový výběr z *k*-tic

Nechť $(X_{1i}, \ldots, X_{ki})$, kde $i = 1, 2, \ldots, n$, je náhodný výběr o rozsahu n z k-rozměrného náhodného vektoru $(X_1, \ldots, X_k)$. Potom *k-rozměrným bootstrapovým výběrem z k-tic* rozumíme náhodný výběr s opakováním z k-tic $(X_{1i}, \ldots, X_{ki})$, tj. $\big((X_{1b_1}, \ldots, X_{kb_1}), \ldots, (X_{1b_n}, \ldots, X_{kb_n})\big)$, kde $b_1, \ldots, b_n$ je náhodný výběr s opakováním z čísel $1, \ldots, n$. Takto získaný k-rozměrný bootstrapový náhodný výběr se používá zejména pro $k = 2$ k výpočtům konfidenčních intervalů dvojvýběrových charakteristik, jako jsou kovariance, koeficient korelace nebo poměr středních hodnot. Je také vhodný pro lineární a zejména nelineární regresní analýzu.

Uvedeme si např. postup pro získání konfidenčního intervalu pro koeficient korelace:

1. Realizujeme B bootstrapových náhodných výběrů z pozorovaných hodnot náhodného vektoru (X, Y). Obvykle volíme $B \geqq 1000$.
2. Pro každý bootstrapový výběr spočítáme pozorovanou hodnotu výběrového koeficientu korelace $R_{b,i}$, $i = 1, \ldots, B$.
3. Pro intervalový odhad koeficientu korelace $\rho(X, Y)$ použijeme jednoduchý kvantilový konfidenční interval nebo lépe BCA konfidenční interval.

Reziduový kvantilový konfidenční interval se nedoporučuje, protože v některých případech dává meze mimo interval $\langle -1, 1\rangle$. Použití obecného pivotového konfidenčního intervalu je také možné, ale v tomto případě výrazně výpočtově náročnější.

6.2. Vícerozměrný bootstrapový výběr z odchylek

Předpokládáme, že závislost náhodných veličin $X_1, \ldots, X_k, Y$ můžeme vyjádřit následujícím vztahem:

$$Y = h(X_1, \ldots, X_k) + \varepsilon, \tag{6.1}$$

kde $h(X_1, \ldots, X_k)$ je nějaká funkce náhodných veličin $X_1, \ldots, X_k$ a *odchylka* ε je na nich nezávislá náhodná proměnná se střední hodnotou 0 a směrodatnou odchylkou σ. Myšlenkou této metody je, že zafixujeme hodnoty náhodných veličin $X_1, \ldots, X_k$, odhadneme odchylku ε_i pro každé $i = 1, \ldots, n$ a realizujeme bootstrapový výběr z odchylek. Dosazením do vztahu (6.1) dostaneme pro každou původní hodnotu náhodných veličin $X_1, \ldots, X_k$ novou hodnotu náhodné veličiny Y.

Postup pro získání dvojrozměrného bootstrapového výběru je tedy následující:

1. Vypočítáme odhad $\widehat{h}(X_1, \ldots, X_k)$ funkce $h(X_1, \ldots, X_k)$.
2. Pro každé $i = 1, \ldots, n$ vypočítáme odhad odchylky $e_i = Y_i - \widehat{h}(X_{1i}, \ldots, X_{ki})$.
3. Realizujeme náhodný výběr s opakováním z hodnot $e_1, \ldots, e_n$ a označíme $e_{b,1}, \ldots, e_{b,n}$.
4. Pro každé $i = 1, \ldots, n$ vypočítáme $Y_{b,i} = \widehat{h}(X_{1i}, \ldots, X_{ki}) + e_{b,i}$.
5. $(k+1)$-rozměrný bootstrapový výběr pak dostáváme ve tvaru $\big((X_{11}, \ldots, X_{k1}, Y_{b,1})$ až $(X_{1n}, \ldots, X_{kn}, Y_{b,n})\big)$.

Tento přístup používáme samozřejmě tehdy, když hodnoty náhodných veličin $X_1, \ldots, X_k$ byly v praxi skutečně zafixované, např. se měřily hodnoty náhodné veličiny Y v pravidelných časových a prostorových odstupech. Je také vhodná v případech regresní analýzy, pokud X_1 až X_k chápeme jako nezávisle proměnné a Y jako závisle proměnnou.

Odhady odchylek $e_1, \ldots, e_n$ mají ale rozptyly závislé na náhodných veličinách X_1 až X_k. Označme

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & \cdots & X_{k1} \\ \vdots & \ddots & \vdots \\ X_{1n} & \cdots & X_{kn} \end{pmatrix} \quad \text{a} \quad \boldsymbol{H} = \boldsymbol{X}\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T.$$

Definujme pro každé $i = 1, \ldots, n$ i-tý diagonální prvek matice $\boldsymbol{H}$ jako *vliv i-tého pozorování* Y_i a označme h_i. Pak rozptyl e_i je $\sigma^2(1 - h_i)$, kde σ^2 je rozptyl odchylky ε. Zavedeme pro každé $i = 1, \ldots, n$

$$r_i = \frac{e_i}{\sqrt{1 - h_i}}.$$

Pak *korigovanou odchylkou* rozumíme pro každé $i = 1, \ldots, n$ veličinu

$$e_i^* = r_i - \frac{1}{n}\sum_{i=1}^{n} r_i.$$

Korigované odchylky mají konstantní rozptyl σ^2 stejně jako ε a jejich součet je nulový jako součet pozorovaných hodnot $e_1, \ldots, e_n$. Proto se někdy doporučuje realizovat bootstrapový výběr nikoliv přímo z pozorovaných hodnot $e_1, \ldots, e_n$, ale z korigovaných odchylek e_1^* až e_n^*. Zejména je tento postup vhodný, pokud rozsah náhodného výběru n je velmi malý nebo pokud jsou mezi pozorovanými hodnotami extrémně odlehlá pozorování.

Kapitola 7

Testy hypotéz o středních hodnotách

7.1. Náhodný výběr z dvojrozměrného náhodného vektoru

Uvažujme nyní dvě náhodné veličiny Y_1, Y_2 a náhodné výběry z těchto náhodných veličin $(Y_{11}, \ldots, Y_{1n_1})$, $(Y_{21}, \ldots, Y_{2n_2})$, kde Y_{ij} označuje j-té pozorování náhodné veličiny Y_i, $j = 1, \ldots, n_i$, $i = 1, 2$. Označme dále μ_i střední hodnotu a $F_i(y)$ distribuční funkci náhodné veličiny Y_i. Odchylky definujeme jako $\varepsilon_{ij} = Y_{ij} - \mu_i$.

Budeme se zabývat problémem, jak vytvořit konfidenční interval pro rozdíl středních hodnot $\mu_1 - \mu_2$. Pomocí tohoto konfidenčního intervalu pak také můžeme testovat hypotézu $H: \mu_1 - \mu_2 = \Delta$. Jejím speciálním případem pro $\Delta = 0$ je hypotéza $H: \mu_1 = \mu_2$. Budeme přitom uvažovat dva případy.

V prvním případě budeme předpokládat, že rozdělení pravděpodobnosti odchylek jsou pro obě náhodné veličiny shodná. Pak lze ukázat, že rozdělení pravděpodobnosti obou náhodných veličin Y_1, Y_2 jsou stejného typu a jsou pouze navzájem posunutá o $\Delta = \mu_1 - \mu_2$, tj. $F_1(y) = F_2(y - \Delta)$. Pokud odchylky mají normální rozdělení pravděpodobnosti, nastává tento případ právě tehdy, když rozptyly náhodných veličin Y_1, Y_2 jsou shodné, tj. $D(Y_1) = D(Y_2)$.

V druhém případě předpokládáme, že rozdělení pravděpodobnosti odchylek jsou různá. Pokud se přitom jedná o dvě různá normální rozdělení, nastává tento případ právě tehdy, když rozptyly náhodných veličin Y_1, Y_2 jsou různé, tj. $D(Y_1) \neq D(Y_2)$.

7.1.1. Shodná rozdělení pravděpodobnosti odchylek

Předpokládáme shodná rozdělení pravděpodobnosti odchylek pro obě náhodné veličiny Y_1, Y_2, tj. existuje jediná náhodná veličina ε s distribuční funkcí $F(\varepsilon)$. Označme

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}},$$

kde $\overline{Y}_1$, $\overline{Y}_2$ jsou příslušné výběrové průměry a

$$S_p = \sqrt{\sum_{i=1}^{2} \sum_{j=1}^{n_i} \frac{\left(Y_{ij} - \overline{Y}_i\right)^2}{n_1 + n_2 - 2}}.$$

Rozdělení pravděpodobnosti statistiky t závisí pouze na $F(\varepsilon)$, což se ukáže snadno, když dosadíme $\mu_i + \varepsilon_{ij}$ za Y_{ij}. Po úpravě dostaneme

$$t_\varepsilon = \frac{\overline{\varepsilon}_1 - \overline{\varepsilon}_2}{S_{\varepsilon p}\sqrt{1/n_1 + 1/n_2}},$$

kde $\overline{\varepsilon}_1$, $\overline{\varepsilon}_2$ jsou výběrové průměry odchylek pro náhodné veličiny Y_1, Y_2 a

$$S_{\varepsilon p} = \sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{(\varepsilon_{ij} - \overline{\varepsilon}_i)^2}{n_1 + n_2 - 2}}.$$

Pokud ε má normální rozdělení pravděpodobnosti, pak statistiky t i t_ε mají Studentovo rozdělení pravděpodobnosti s $n_1 + n_2 - 2$ stupni volnosti a platí

$$P\left(-t_{1-\alpha} < \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} < t_{1-\alpha}\right) = 1 - 2\alpha,$$

kde $t_{1-\alpha}$ je $(1-\alpha)$-kvantil Studentova rozdělení. Odtud konfidenční interval se spolehlivostí $1 - 2\alpha$ pro rozdíl $\mu_1 - \mu_2$ dostaneme ve tvaru

$$\mu_1 - \mu_2 \in \left(\overline{Y}_1 - \overline{Y}_2 - t_{1-\alpha}S_p\sqrt{1/n_1 + 1/n_2}; \overline{Y}_1 - \overline{Y}_2 + t_{1-\alpha}S_p\sqrt{1/n_1 + 1/n_2}\right).$$

Pokud ε nemá normální rozdělení pravděpodobnosti, odhadneme hodnoty kvantilů rozdělení pravděpodobnosti statistiky t_ε metodou bootstrap. Bootstrapový konfidenční interval pro rozdíl $\mu_1 - \mu_2$ získáme následujícím postupem:

1. Z pozorovaných hodnot náhodných výběrů z náhodných veličin Y_1, Y_2 vypočítáme pozorované hodnoty výběrových průměrů $\overline{Y}_1$, $\overline{Y}_2$ a odchylek $e_{ij} = Y_{ij} - \overline{Y}_i$.
2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n_1+n_2 z pozorovaných hodnot odchylek $e_{11}, \ldots, e_{1n_1}, e_{21}, \ldots, e_{2n_2}$ a označíme $e_{b,11}, \ldots, e_{b,1n_1}$, $e_{b,21}, \ldots, e_{b,2n_2}$. Obvykle volíme $B \geqq 1000$.
3. Pro každý bootstrapový výběr vypočítáme pozorované hodnoty výběrových průměrů $\overline{e}_{b,i}$ a hodnotu statistiky $t_{e,b}$

$$t_{e,b,k} = \frac{\overline{e}_{b,1} - \overline{e}_{b,2}}{S_{b,ep}\sqrt{1/n_1 + 1/n_2}}, \quad \text{kde} \quad S_{b,ep} = \sqrt{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{(e_{b,ij} - \overline{e}_{b,i})^2}{n_1 + n_2 - 2}}.$$

   kde $k = 1, \ldots, B$.
4. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky $t_{e,b}$ odhadneme hodnotami $t_{b,\alpha}$ a $t_{b,1-\alpha}$ splňujícími co nejpřesněji

$$\left|\{t_{e,b,k};\, t_{e,b,k} \leqq t_{b,\alpha}\}\right| / B \doteq \alpha, \quad \left|\{t_{e,b,k};\, t_{e,b,k} \leqq t_{b,1-\alpha}\}\right| / B \doteq 1 - \alpha,$$

   kde $\left|\{\ldots\}\right|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.
5. *Bootstrapovým konfidenčním intervalem se spolehlivostí* $1 - 2\alpha$ *pro rozdíl středních hodnot* $\mu_1 - \mu_2$ je

$$\mu_1 - \mu_2 \in \left(\overline{Y}_1 - \overline{Y}_2 - t_{b,1-\alpha}S_p\sqrt{1/n_1 + 1/n_2}; \overline{Y}_1 - \overline{Y}_2 - t_{b,\alpha}S_p\sqrt{1/n_1 + 1/n_2}\right).$$

7.1.2. Různá rozdělení pravděpodobnosti odchylek

Pokud rozdělení pravděpodobnosti odchylek pro náhodné veličiny Y_1, Y_2 jsou různá, musíme změnit statistiku i způsob realizace bootstrapových výběrů z odchylek. Označme

$$z = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}},$$

kde S_1^2, S_2^2 jsou výběrové rozptyly náhodných výběrů z Y_1, Y_2. Rozdělení pravděpodobnosti statistiky z závisí pouze na rozdělení pravděpodobnosti odchylek, což se ukáže snadno, dosadíme-li $\mu_i + \varepsilon_{ij}$ za Y_{ij}. Po úpravě dostaneme

$$z_\varepsilon = \frac{\overline{\varepsilon}_1 - \overline{\varepsilon}_2}{\sqrt{S_{\varepsilon 1}^2/n_1 + S_{\varepsilon 2}^2/n_2}},$$

kde

$$S_{\varepsilon i}^2 = \sum_{j=1}^{n_i} \frac{(\varepsilon_{ij} - \overline{\varepsilon}_i)^2}{n_i - 1}, \quad i = 1, 2.$$

Hodnoty kvantilů rozdělení pravděpodobnosti statistiky z_ε odhadneme metodou bootstrap. Bootstrapový konfidenční interval pro rozdíl $\mu_1 - \mu_2$ získáme následujícím postupem:

1. Z pozorovaných hodnot náhodných výběrů z náhodných veličin Y_1, Y_2 vypočítáme pozorované hodnoty výběrových průměrů $\overline{Y}_1$, $\overline{Y}_2$ a odchylek $e_{ij} = Y_{ij} - \overline{Y}_i$.
2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n_1 z pozorovaných hodnot odchylek $e_{11}, \ldots, e_{1n_1}$ a o rozsahu n_2 z pozorovaných hodnot odchylek $e_{21}, \ldots, e_{2n_2}$ a označíme $e_{b,11}, \ldots, e_{b,1n_1}$ a $e_{b,21}, \ldots, e_{b,2n_2}$. Obvykle volíme $B \geqq 1000$.
3. Pro každý bootstrapový výběr vypočítáme pozorované hodnoty výběrových průměrů $\overline{e_{b,i}}$ a hodnotu statistiky $z_{e,b}$

$$z_{e,b,k} = \frac{\overline{e}_{b,1} - \overline{e}_{b,2}}{\sqrt{S_{b,e1}^2/n_1 + S_{b,e2}^2/n_2}}, \quad \text{kde} \quad S_{b,ei}^2 = \sum_{j=1}^{n_i} \frac{(e_{b,ij} - \overline{e}_{b,i})^2}{n_i - 1}, \quad i = 1, 2,$$

   pro $k = 1, 2, \ldots, B$.
4. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky $z_{e,b}$ odhadneme hodnotami $z_{b,\alpha}$ a $z_{b,1-\alpha}$ splňujícími co nejpřesněji

$$\big|\{z_{e,b,k};\, z_{e,b,k} \leqq z_{b,\alpha}\}\big| / B \doteq \alpha, \quad \big|\{z_{e,b,k};\, z_{e,b,k} \leqq z_{b,1-\alpha}\}\big| / B \doteq 1 - \alpha,$$

   kde $\big|\{\ldots\}\big|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.
5. *Bootstrapovým konfidenčním intervalem se spolehlivostí* $1 - 2\alpha$ *pro rozdíl středních hodnot* $\mu_1 - \mu_2$ je

$$\mu_1 - \mu_2 \in \left(\overline{Y}_1 - \overline{Y}_2 - z_{b,1-\alpha}\sqrt{S_1^2/n_1 + S_2^2/n_2};\, \overline{Y}_1 - \overline{Y}_2 - z_{b,\alpha}\sqrt{S_1^2/n_1 + S_2^2/n_2}\right).$$

Další metody

Je nepříjemné, že ani v případě normálních rozdělení pravděpodobnosti odchylek nemají statistiky z a z_ε Studentovo rozdělení pravděpodobnosti a hodnoty jejich kvantilů je třeba odhadovat metodou bootstrap. To lze obejít tzv. *Satterthwaitovou aproximací*, podle které má statistika z Studentovo rozdělení s K stupni volnosti, kde

$$K \doteq \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}.$$

Namísto statistiky z můžeme také použít přímo statistiku $\overline{Y}_1 - \overline{Y}_2$ a pro ni sestrojit některý z dříve uvedených kvantilových konfidenčních intervalů. V tom případě realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n_1 přímo z pozorovaných hodnot $y_{11}, \ldots, y_{1n_1}$ a o rozsahu n_2 z pozorovaných hodnot $y_{21}, \ldots, y_{2n_2}$.

7.1.3. Testování hypotéz

Popsali jsme konstrukci konfidenčních intervalů pivotového typu pro statistiku $\mu_1 - \mu_2$, využili jsme přitom α-kvantily a $(1-\alpha)$-kvantily rozdělení pravděpodobnosti odchylek. Jiné kvantily těchto rozdělení se získají analogicky.

Testujeme-li hypotézu $H\colon \mu_1 - \mu_2 = \Delta$ proti alternativní hypotéze $\overline{H}\colon \mu_1 - \mu_2 \neq \Delta$, využijeme příslušný konfidenční interval sestrojený pomocí α-kvantilu a $(1-\alpha)$-kvantilu. Testujeme-li ale proti alternativní hypotéze $\overline{H}\colon \mu_1 - \mu_2 > \Delta$, využijeme příslušný konfidenční interval sestrojený pomocí $(1-2\alpha)$-kvantilu a 0-kvantilu. A konečně testujeme-li proti alternativní hypotéze $\overline{H}\colon \mu_1 - \mu_2 < \Delta$, využijeme příslušný konfidenční interval sestrojený pomocí 2α-kvantilu a 1-kvantilu.

Pokud pozorovaná hodnota parametru Δ je prvkem intervalového odhadu, pak na hladině významnosti 2α nezamítáme hypotézu H a zároveň zamítáme alternativní hypotézu $\overline{H}$. Pokud pozorovaná hodnota parametru Δ není prvkem intervalového odhadu, pak zamítáme hypotézu H a zároveň nezamítáme alternativní hypotézu $\overline{H}$.

7.2. Náhodný výběr z k-rozměrného náhodného vektoru

Rozšiřme nyní úvahy z minulé části na obecnější případ. Uvažujme k náhodných veličin Y_1 až Y_k a náhodné výběry z těchto náhodných veličin $(Y_{11}, \ldots, Y_{1n_1}), \ldots, (Y_{k1}, \ldots, Y_{kn_k})$, kde Y_{ij} označuje j-té pozorování náhodné veličiny Y_i, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$. Označme dále μ_i střední hodnotu náhodné veličiny Y_i. Odchylky definujeme jako $\varepsilon_{ij} = Y_{ij} - \mu_i$.

Budeme se zabývat problémem, jak testovat hypotézu $H\colon \mu_1 = \cdots = \mu_k = 0$ proti alternativní hypotéze $\overline{H}\colon \exists i\ (\mu_i \neq 0)$. Budeme přitom opět uvažovat dva případy. V prvním předpokládáme, že rozdělení pravděpodobnosti odchylek jsou pro všechny náhodné veličiny $Y_1, \ldots, Y_k$ shodná, ve druhém naopak předpokládáme, že některá jsou různá.

7.2.1. Shodná rozdělení pravděpodobnosti odchylek

Předpokládáme, že všechny odchylky ε_{ij} jsou náhodnými výběry z jediné náhodné veličiny ε. Pro testování rovnosti středních hodnot zavedeme statistiku

$$F = \frac{\sum_{i=1}^{k} n_i \left(\overline{Y}_i - \overline{Y}\right)^2 \Big/ (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_i\right)^2 \Big/ \left(\sum_{i=1}^{k} n_i - k\right)},$$

kde

$$\overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \qquad \text{a} \qquad \overline{Y} = \frac{\sum_{i=1}^{k} n_i \overline{Y}_i}{\sum_{i=1}^{k} n_i}.$$

Po dosazení $\mu_i + \varepsilon_{ij}$ za Y_{ij} a při uvažování nulové hypotézy $\mu_1 = \cdots = \mu_k = 0$ dostáváme

$$F_\varepsilon = \frac{\sum_{i=1}^{k} n_i (\overline{\varepsilon}_i - \overline{\varepsilon})^2 \Big/ (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\varepsilon_{ij} - \overline{\varepsilon}_i)^2 \Big/ \left(\sum_{i=1}^{k} n_i - k\right)},$$

kde

$$\overline{\varepsilon}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \varepsilon_{ij} \qquad \text{a} \qquad \overline{\varepsilon} = \frac{\sum_{i=1}^{k} n_i \overline{\varepsilon}_i}{\sum_{i=1}^{k} n_i}.$$

Pokud náhodná veličina ε má normální rozdělení pravděpodobnosti, statistika F_ε má Fisherovo-Snedecorovo rozdělení pravděpodobnosti s $k-1$ a $\sum_{i=1}^{k} n_i - k$ stupni volnosti a platí:

$$P\left(F_\alpha < \frac{\sum_{i=1}^{k} n_i (\overline{\varepsilon}_i - \overline{\varepsilon})^2 \Big/ (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\varepsilon_{ij} - \overline{\varepsilon}_i)^2 \Big/ \left(\sum_{i=1}^{k} n_i - k\right)} < F_{1-\alpha} \right) = 1 - 2\alpha,$$

kde F_α, $F_{1-\alpha}$ jsou α-kvantil a $(1-\alpha)$-kvantil Fisherova-Snedecorova rozdělení pravděpodobnosti s $k-1$ a $\sum_{i=1}^{k} n_i - k$ stupni volnosti.

Pokud nemůžeme u ε předpokládat normální rozdělení, musíme hodnoty kvantilů rozdělení pravděpodobnosti statistiky F odhadnout metodou bootstrap. Postup je následující:

1. Z pozorovaných hodnot náhodných výběrů z náhodných veličin $Y_1, \ldots, Y_k$ vypočítáme pozorované hodnoty výběrových průměrů $\overline{Y}_1, \ldots, \overline{Y}_k$ a odchylek $e_{ij} = Y_{ij} - \overline{Y}_i$.

2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu $\sum_{i=1}^{k} n_i$ z pozorovaných hodnot odchylek $e_{11}, \ldots, e_{kn_k}$ a označíme $e_{b,11}, \ldots, e_{b,kn_k}$. Obvykle volíme $B \geqq 1000$.
3. Pro každý bootstrapový výběr vypočítáme pozorované hodnoty výběrových průměrů $\overline{e}_{b,i}$ a $\overline{e}_b$ a hodnotu statistiky $F_{e,b}$

$$F_{e,b,l} = \frac{\sum_{i=1}^{k} n_i (\overline{e}_{b,i} - \overline{e}_b)^2 \Big/ (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (e_{b,ij} - \overline{e}_{b,i})^2 \Big/ \Big(\sum_{i=1}^{k} n_i - k\Big)},$$

kde $l = 1, \ldots, B$.
4. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky $F_{e,b}$ odhadneme hodnotami $F_{b,\alpha}$ a $F_{b,1-\alpha}$ splňujícími co nejpřesněji

$$\big|\{F_{e,b,l};\, F_{e,b,l} \leqq F_{b,\alpha}\}\big| / B \doteq \alpha, \quad \big|\{F_{e,b,l};\, F_{e,b,l} \leqq F_{b,1-\alpha}\}\big| / B \doteq 1-\alpha,$$

kde $\big|\{\ldots\}\big|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.

Pokud tedy pozorovaná hodnota statistiky F bude ležet v intervalu

$$(F_\alpha, F_{1-\alpha}), \quad \text{resp.} \quad (F_{b,\alpha}, F_{b,1-\alpha}),$$

na hladině významnosti 2α nezamítneme hypotézu $H: \mu_1 = \cdots = \mu_k = 0$ a zamítneme alternativní hypotézu $\overline{H}$. Pokud pozorovaná hodnota statistiky F bude ležet mimo tento interval, hypotézu H naopak zamítneme a nezamítneme alternativní hypotézu $\overline{H}$.

7.2.2. Různá rozdělení pravděpodobnosti odchylek

V případě různých rozdělení pravděpodobnosti odchylek lze uvažovat více možností, jak vhodně upravit statistiku F. Někteří autoři uvádějí různé modifikace statistiky nebo stupňů volnosti jejího rozdělení. Mnoho příkladů z praxe ale ukazuje, že je možné i zde použít statistiku F s Fisherovým-Snedecorovým rozdělením pravděpodobnosti, jako by odchylky měly stejné normální rozdělení pravděpodobnosti a náhodné veličiny $Y_1, \ldots, Y_k$ měly stejné rozptyly. To je možné díky tomu, že statistika F je robustní a její rozdělení pravděpodobnosti je jen málo ovlivněno mírným porušením předpokladů o normalitě rozdělení odchylek nebo o shodnosti rozptylů náhodných veličin $Y_1, \ldots, Y_k$.

Vhodným kompromisem se zdá být použití statistiky F v nezměněném tvaru, avšak s použitím metody bootstrap k odhadu hodnot kvantilů jejího rozdělení pravděpodobnosti. Postupujeme tedy takto:

1. Z pozorovaných hodnot náhodných výběrů z náhodných veličin $Y_1, \ldots, Y_k$ vypočítáme pozorované hodnoty výběrových průměrů $\overline{Y}_1, \ldots, \overline{Y}_k$ a odchylek $e_{ij} = Y_{ij} - \overline{Y}_i$.
2. Realizujeme B náhodných bootstrapových výběrů (s opakováním) o rozsahu n_i z pozorovaných hodnot odchylek $e_{i1}, \ldots, e_{in_i}$ pro $i = 1, \ldots, k$ a označíme je $e_{b,i1}, \ldots, e_{b,in_i}$. Obvykle volíme $B \geqq 1000$.

3. Pro každý bootstrapový výběr vypočítáme pozorované hodnoty výběrových průměrů $\overline{e}_{b,i}$ a $\overline{e}_b$ a hodnotu statistiky $F_{e,b}$

$$F_{e,b,l} = \frac{\sum_{i=1}^{k} n_i(\overline{e}_{b,i} - \overline{e}_b)^2 \Big/ (k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(e_{b,ij} - \overline{e}_{b,i})^2 \Big/ \Big(\sum_{i=1}^{k} n_i - k\Big)},$$

kde $l = 1, \ldots, B$.

4. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky $F_{e,b}$ odhadneme hodnotami $F_{b,\alpha}$ a $F_{b,1-\alpha}$ splňujícími co nejpřesněji

$$\big|\{F_{e,b,l};\, F_{e,b,l} \leqq F_{b,\alpha}\}\big| / B \doteq \alpha, \quad \big|\{F_{e,b,l};\, F_{e,b,l} \leqq F_{b,1-\alpha}\}\big| / B \doteq 1-\alpha,$$

kde $\big|\{\ldots\}\big|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.

Pokud tedy pozorovaná hodnota statistiky F bude ležet v intervalu

$$(F_{b,\alpha},\, F_{b,1-\alpha}),$$

na hladině významnosti 2α nezamítneme hypotézu $H\colon \mu_1 = \cdots = \mu_k = 0$ a zamítneme alternativní hypotézu $\overline{H}$. Pokud pozorovaná hodnota statistiky F bude ležet mimo tento interval, hypotézu H naopak zamítneme a nezamítneme alternativní hypotézu $\overline{H}$.

Kapitola 8

Testy hypotéz o parametrech lineárního regresního modelu

Uvažujme lineární regresní model tak, jak byl zaveden v první kapitole, tj. $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ a bodovým odhadem vektoru parametrů $\boldsymbol{\beta}$ je $\boldsymbol{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

8.1. Konfidenční interval pro β_j

Konfidenční interval pro jeden libovolný koeficient β_j sestrojíme pomocí statistiky

$$t = \frac{b_j - \beta_j}{s\sqrt{\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{jj}}},$$

kde $\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{jj}$ je j-tý diagonální prvek matice $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ a s je bodový odhad směrodatné odchylky podle (2.1). Pokud odchylky $\boldsymbol{\varepsilon}$ mají normální rozdělení pravděpodobnosti, má t Studentovo rozdělení pravděpodobnosti s $n - k$ stupni volnosti a platí

$$P\left(-t_{1-\alpha} \leqq \frac{b_j - \beta_j}{s\sqrt{\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{jj}}} \leqq t_{1-\alpha}\right) = 1 - 2\alpha,$$

odkud dostáváme konfidenční interval pro β_j se spolehlivostí $1 - 2\alpha$

$$\beta_j \in \left\langle b_j - t_{1-\alpha}s\sqrt{\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{jj}}; b_j + t_{1-\alpha}s\sqrt{\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{jj}}\right\rangle,$$

kde $t_{1-\alpha}$ je $(1-\alpha)$-kvantil Studentova rozdělení pravděpodobnosti s $n - k$ stupni volnosti.

Pokud ale u odchylek $\boldsymbol{\varepsilon}$ nemůžeme normální rozdělení předpokládat, musíme hodnoty kvantilů rozdělení pravděpodobnosti statistiky t odhadnout metodou bootstrap. Vícerozměrný bootstrapový výběr přitom budeme realizovat jako výběr z odchylek.

1. Z pozorovaných hodnot pro $\boldsymbol{X}$ a $\boldsymbol{Y}$ vypočítáme metodou nejmenších čtverců odhad $\boldsymbol{b}$ a pozorované hodnoty odchylek $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}$.
2. Realizujeme B bootstrapových výběrů (s opakováním) o rozsahu n z $e_1, \ldots, e_n$ a označíme $(e_{b,1,i}, \ldots, e_{b,n,i})$, $i = 1, \ldots, B$. Obvykle volíme $B \geqq 1000$.

3. Pro každé $i = 1, \ldots, B$ vypočítáme $\boldsymbol{Y}_{b,i} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}_{b,i}$. Dosazením $\boldsymbol{Y}_{b,i}$ za $\boldsymbol{Y}$ dále spočítáme odhad parametrů $\boldsymbol{b}_{b,i}$, odhad rozptylu $s_{b,i}$ a hodnotu statistiky t_b

$$t_{b,i} = \frac{b_{b,i,j} - b_j}{s_{b,i}\sqrt{\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{jj}}}.$$

4. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky t_b odhadneme hodnotami $t_{b,\alpha}$, $t_{b,1-\alpha}$ splňujícími co nejpřesněji

$$\left|\{t_{b,i};\, t_{b,i} \leqq t_{b,\alpha}\}\right| / B \doteq \alpha, \quad \left|\{t_{b,i};\, t_{b,i} \leqq t_{b,1-\alpha}\}\right| / B \doteq 1 - \alpha,$$

kde $\left|\{\ldots\}\right|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.

5. *Bootstrapovým konfidenčním intervalem se spolehlivostí* $1 - 2\alpha$ *pro* j*-tý regresní parametr* β_j je

$$\beta_j \in \left\langle b_j - t_{b,1-\alpha}s\sqrt{\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{jj}};\, b_j - t_{b,\alpha}s\sqrt{\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{jj}}\right\rangle.$$

8.2. Test hypotézy $C\beta = 0$

V této podkapitole se zaměříme na problém, jak lze metodou bootstrap testovat hypotézu, která může být vyjádřena ve tvaru $H\colon \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$, kde $\boldsymbol{C}$ je matice typu $q \times k$ s plnou hodností q. Např. pro hypotézu $H\colon \beta_1 = \cdots = \beta_k = 0$ by $\boldsymbol{C}$ byla matice typu $k \times k$

$$\boldsymbol{C} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

nebo pro hypotézu $H\colon \beta_1 = \beta_2$ by $\boldsymbol{C}$ byla matice typu $1 \times k$

$$\boldsymbol{C} = (\ 1 \quad -1 \quad 0 \quad \cdots \quad 0\ ).$$

Testovat budeme proti alternativní hypotéze $\overline{H}\colon \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{0}$.

Hypotézu H testujeme pomocí statistiky

$$F = \frac{(\boldsymbol{C}\boldsymbol{b})^T\left[\boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T\right]^{-1}\boldsymbol{C}\boldsymbol{b}}{q\mathrm{MSE}},$$

kde

$$\mathrm{MSE} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})}{n-k}. \tag{8.1}$$

Pokud odchylky $\boldsymbol{\varepsilon}$ mají normální rozdělení pravděpodobnosti a platí hypotéza H, pak statistika F má Fisherovo-Snedecorovo rozdělení pravděpodobnosti s q a $n-k$ stupni volnosti. Pokud tedy pozorovaná hodnota statistiky F bude prvkem intervalu $(F_\alpha, F_{1-\alpha})$, na hladině významnosti 2α nezamítneme hypotézu H a zamítneme alternativní hypotézu $\overline{H}$. Pokud pozorovaná hodnota F nebude prvkem intervalu, pak naopak zamítneme hypotézu H a nezamítneme alternativní hypotézu $\overline{H}$.

V případě, kdy $\boldsymbol{C}$ je řádkový vektor, tj. $q = 1$, můžeme také pro test hypotézy $H: \boldsymbol{C\beta} = \Delta$ proti alternativní hypotéze $\overline{H}: \boldsymbol{C\beta} \neq \Delta$ použít statistiku

$$t = \frac{\boldsymbol{Cb} - \boldsymbol{C\beta}}{\sqrt{\text{MSE } \boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T}}.$$

Pokud odchylky $\boldsymbol{\varepsilon}$ mají normální rozdělení pravděpodobnosti, pak t má Studentovo rozdělení pravděpodobnosti s $n - k$ stupni volnosti a platí

$$P\left(-t_{1-\alpha} < \frac{\boldsymbol{Cb} - \boldsymbol{C\beta}}{\sqrt{\text{MSE } \boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T}} < t_{1-\alpha}\right) = 1 - 2\alpha,$$

kde $t_{1-\alpha}$ je $(1-\alpha)$-kvantil Studentova rozdělení pravděpodobnosti s $n - k$ stupni volnosti. Odtud dostáváme konfidenční interval se spolehlivostí $1 - 2\alpha$ pro $\boldsymbol{C\beta}$

$$\boldsymbol{C\beta} \in \left(\boldsymbol{Cb} - t_{1-\alpha}\sqrt{\text{MSE } \boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T}; \boldsymbol{Cb} + t_{1-\alpha}\sqrt{\text{MSE } \boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T}\right).$$

Pokud Δ bude prvkem tohoto konfidenčního intervalu, na hladině významnosti 2α nezamítneme hypotézu H a zamítneme alternativní hypotézu $\overline{H}$. Pokud Δ nebude prvkem tohoto intervalu, pak naopak zamítneme hypotézu H a nezamítneme alternativní hypotézu $\overline{H}$. Tato statistika se používá zejména k testování hypotéz typu $H: \beta_j = \beta_{j0}$. Můžeme pochopitelně testovat i proti jednostranné alternativní hypotéze. V tom případě zvolíme odpovídající kvantily Studentova rozdělení.

Pokud odchylky $\boldsymbol{\varepsilon}$ nemají normální rozdělení, musíme najít vhodný tvar statistiky F, pro který můžeme odhadnout hodnoty kvantilů jejího rozdělení pravděpodobnosti metodou bootstrap. Dosaďme $\boldsymbol{\varepsilon}$ za $\boldsymbol{Y}$ a označme F_ε statistiku získanou takto ze statistiky F, dále $\boldsymbol{b}_\varepsilon$ odhad $\boldsymbol{\beta}$ metodou nejmenších čtverců a MSE_ε střední kvadratickou chybu podle vzorce (8.1). Pak platí

$$\boldsymbol{b} - \boldsymbol{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon} = \boldsymbol{b}_\varepsilon,$$

a

$$(n-k)\text{MSE} = (\boldsymbol{Y} - \boldsymbol{Xb})^T(\boldsymbol{Y} - \boldsymbol{Xb}) = (\boldsymbol{X\beta} + \boldsymbol{\varepsilon} - \boldsymbol{Xb})^T(\boldsymbol{X\beta} + \boldsymbol{\varepsilon} - \boldsymbol{Xb}) =$$
$$= (\boldsymbol{\varepsilon} - \boldsymbol{Xb}_\varepsilon)^T(\boldsymbol{\varepsilon} - \boldsymbol{Xb}_\varepsilon) = (n-k)\text{MSE}_\varepsilon.$$

S užitím tohoto poznatku a za předpokladu platnosti hypotézy $H: \boldsymbol{C\beta} = \boldsymbol{0}$ dostáváme

$$F = \frac{(\boldsymbol{Cb})^T\left[\boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T\right]^{-1}\boldsymbol{Cb}}{q\text{MSE}} = \frac{\left[\boldsymbol{C}(\boldsymbol{b} - \boldsymbol{\beta})\right]^T\left[\boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T\right]^{-1}\boldsymbol{C}(\boldsymbol{b} - \boldsymbol{\beta})}{q\text{MSE}} =$$

$$= \frac{(\boldsymbol{Cb}_\varepsilon)^T\left[\boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T\right]^{-1}\boldsymbol{Cb}_\varepsilon}{q\text{MSE}_\varepsilon} = F_\varepsilon.$$

Statistiky F a F_ε tedy mají stejné rozdělení pravděpodobnosti, čehož využijeme v následujícím postupu pro test hypotézy H metodou bootstrap. Vícerozměrný bootstrapový výběr přitom opět realizujeme výběrem z odchylek.

1. Z pozorovaných hodnot pro $\boldsymbol{X}$ a $\boldsymbol{Y}$ vypočítáme metodou nejmenších čtverců odhad $\boldsymbol{b}$ a pozorované hodnoty odchylek $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}$.
2. Realizujeme B bootstrapových výběrů (s opakováním) o rozsahu n z $e_1, \ldots, e_n$ a označíme $(e_{b,1,i}, \ldots, e_{b,n,i})$, $i = 1, \ldots, B$. Obvykle volíme $B \geqq 1000$.
3. Pro každé $i = 1, \ldots, B$ dosazením $\boldsymbol{e}_{b,i}$ za $\boldsymbol{Y}$ spočítáme odhad parametrů $\boldsymbol{b}_{b,e,i}$, střední kvadratickou chybu $\mathrm{MSE}_{b,e,i}$ a hodnotu statistiky $F_{b,e}$

$$F_{b,e,i} = \frac{(\boldsymbol{C}\boldsymbol{b}_{b,e,i})^T \left[\boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T\right]^{-1}\boldsymbol{C}\boldsymbol{b}_{b,e,i}}{q\mathrm{MSE}_{b,e,i}}$$

nebo v případě $q = 1$ hodnotu statistiky $t_{b,e}$

$$t_{b,e,i} = \frac{\boldsymbol{C}\boldsymbol{b}_{b,e,i}}{\sqrt{\mathrm{MSE}_{b,e,i}\ \boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T}}.$$

4. α-kvantil a $(1-\alpha)$-kvantil rozdělení pravděpodobnosti statistiky $F_{b,e}$ nebo $t_{b,e}$ odhadneme hodnotami $F_{b,\alpha}$, $F_{b,1-\alpha}$ nebo $t_{b,\alpha}$, $t_{b,1-\alpha}$ splňujícími co nejpřesněji

$$\left|\{F_{b,e,i};\, F_{b,e,i} \leqq F_{b,\alpha}\}\right|/B \doteq \alpha, \quad \left|\{F_{b,e,i};\, F_{b,e,i} \leqq F_{b,1-\alpha}\}\right|/B \doteq 1-\alpha,$$

$$\left|\{t_{b,e,i};\, t_{b,e,i} \leqq t_{b,\alpha}\}\right|/B \doteq \alpha, \quad \left|\{t_{b,e,i};\, t_{b,e,i} \leqq t_{b,1-\alpha}\}\right|/B \doteq 1-\alpha,$$

kde $\left|\{\ldots\}\right|$ značí velikost množiny, tj. v tomto případě počet jejích prvků.
5. Pokud pozorovaná hodnota statistiky F bude ležet v intervalu $(F_{b,\alpha}, F_{b,1-\alpha})$, na hladině významnosti 2α nezamítáme hypotézu $H\colon \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$ a zamítáme alternativní hypotézu $\overline{H}\colon \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{0}$. Pokud pozorovaná hodnota statistiky F nebude ležet v intervalu $(F_{b,\alpha}, F_{b,1-\alpha})$, pak naopak zamítáme hypotézu H a nezamítáme alternativní hypotézu $\overline{H}$.
6. V případě $q = 1$ je konfidenční interval se spolehlivostí $1 - 2\alpha$ pro $\boldsymbol{C}\boldsymbol{\beta}$

$$\boldsymbol{C}\boldsymbol{\beta} \in \left(\boldsymbol{C}\boldsymbol{b} - t_{b,1-\alpha}\sqrt{\mathrm{MSE}\ \boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T};\, \boldsymbol{C}\boldsymbol{b} - t_{b,\alpha}\sqrt{\mathrm{MSE}\ \boldsymbol{C}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{C}^T}\right).$$

Pokud Δ leží v tomto konfidenčním intervalu, nezamítáme na hladině významnosti 2α hypotézu $H\colon \boldsymbol{C}\boldsymbol{\beta} = \Delta$ a zamítáme alternativní hypotézu $\overline{H}\colon \boldsymbol{C}\boldsymbol{\beta} \neq \Delta$. Pokud Δ neleží v konfidenčním intervalu, zamítáme naopak hypotézu H a nezamítáme alternativní hypotézu $\overline{H}$.

8.3. Další metody

Pro testy hypotéz o regresních parametrech nebo konstruování konfidenčních intervalů pro regresní parametry můžeme využít také vícerozměrné bootstrapové výběry z $(k+1)$-tic a kvantilové konfidenční intervaly. Doporučuje se kombinovat pivotové konfidenční intervaly s vícerozměrnými bootstrapovými výběry z odchylek (jak jsme ukázali výše v této kapitole) a kvantilové konfidenční intervaly s bootstrapovými výběry z $(k+1)$-tic. Pak postupujeme takto:

1. Realizujeme B bootstrapových výběrů z $(k+1)$-tic pozorovaných hodnot náhodných veličin $(X_1, \ldots, X_k, Y)$.

2. Pro každý takový bootstrapový výběr spočítáme odhady regresních parametrů nebo jejich funkcí, které nás zajímají.
3. Odhadneme kvantily rozdělení pravděpodobnosti těchto odhadů regresních parametrů nebo jejich funkcí.
4. Sestrojíme některý z typů kvantilových konfidenčních intervalů, nejlépe BCA konfidenční interval.
5. Pomocí tohoto konfidenčního intervalu můžeme dále testovat nějakou hypotézu o regresních parametrech nebo jejich funkcích proti jednostranné nebo oboustranné alternativní hypotéze.

Kapitola 9

Odhad diskrétního rozdělení pravděpodobnosti kategoriální veličiny pomocí gradientu kvazinormy

Nechť X je kategoriální veličina, která nabývá náhodně konečně mnoha různých slovních hodnot x_j^*, $j = 1, \ldots, m$, kde $m \geqq 2$. Pozorováním veličiny X získáme statistický soubor $(x_1, \ldots, x_n)$, roztříděním získáme roztříděný statistický soubor $\big((x_1^*, f_1/n), \ldots, (x_m^*, f_m/n)\big)$, kde $f_j/n \neq 0$ je relativní četnost pozorované hodnoty x_j^*, $j = 1, \ldots, m$. Neznámé rozdělení pravděpodobnosti kategoriální veličiny X označíme $\boldsymbol{p} = (p_1, \ldots, p_m)$, kde $p_j = P(X = x_j^*)$. Odhad pravděpodobností $\boldsymbol{p}$ je vlastně odhadem parametrů multinomického rozdělení pravděpodobnosti $M(n, p_1, \ldots, p_n)$. Nestranných odhadem $\boldsymbol{p}$ je $\widehat{\boldsymbol{p}} = (f_1/n, \ldots, f_m/n)$. V [7] je předložen následující pesimistický odhad $\boldsymbol{p}$ založený na gradientu kvazinormy rozdělení $\boldsymbol{p}$.

Nechť funkce $f : (0, \infty) \to \mathbb{R}^*$, kde $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, \infty\}$, je konvexní na $(0, \infty)$, striktně konvexní v bodě $u = 1$ a nabývá zde hodnoty $f(1) = 0$. Jestliže $\boldsymbol{p} = (p_1, \ldots, p_m)$, resp. $\boldsymbol{q} = (q_1, \ldots, q_m)$ je diskrétní rozdělení pravděpodobnosti z pravděpodobnostního prostoru (Ω, Σ, P), resp. $(\Omega, \Sigma, \mathrm{Q})$, pak *$f$-divergencí* těchto rozdělení rozumíme funkcionál

$$D_f(\boldsymbol{p}, \boldsymbol{q}) = \sum_{j=1}^{m} q_j f\left(\frac{p_j}{q_j}\right).$$

Pojem f-divergence má význam vzdálenosti daných rozdělení. Platí

1. $\boldsymbol{p} = \boldsymbol{q} \Leftrightarrow D_f(\boldsymbol{p}, \boldsymbol{q}) = 0$,
2. $D_f(\boldsymbol{p}, \boldsymbol{q})$ nabývá v $\mathbb{R}^*$ svého maxima $\Leftrightarrow$ $\boldsymbol{p}$ a $\boldsymbol{q}$ jsou ortogonální, tj. existují takové disjunktní množiny $E, F \subset \Omega$, že

$$\sum_{x_j^* \in E} p_j = 1 \quad \text{a} \quad \sum_{x_j^* \in F} q_j = 1.$$

Nechť $S = \left\{\boldsymbol{p} \in \mathbb{R}^m : \forall p_j \geqq 0, \sum_{j=1}^{m} p_j = 1\right\}$ je množina všech diskrétních rozdělení pravděpodobnosti na Ω. *Kvazinormou* rozdělení $\boldsymbol{p} \in S$ rozumíme f-divergenci $D_f(\boldsymbol{p}, \boldsymbol{p}_0)$, kde $\boldsymbol{p}_0 = (1/m, \ldots, 1/m)$, a o funkci f říkáme, že *generuje* kvazinormu $D_f(\boldsymbol{p}, \boldsymbol{p}_0)$ na S. Platí

1. $D_f(\boldsymbol{p}, \boldsymbol{p}_0) = \frac{1}{m} \sum_{j=1}^{m} f(mp_j)$,
2. $D_f(\boldsymbol{p}, \boldsymbol{p}_0)$ je nezáporná konvexní funkce na S symetrická vzhledem k proměnným p_j, $j = 1, \ldots, m$.
3. $\boldsymbol{p}_0$ minimalizuje integrál všech f-divergencí $D_f(\boldsymbol{p}, \boldsymbol{q})$ na S a má maximální entropii.

Budeme hledat takové rozdělení pravděpodobnosti v S, které je nejblíže $\boldsymbol{p}_0$ a k němuž se dostaneme od empirického rozdělení co nejrychleji. Tomu odpovídá minimalizace kvazinormy $D_f(\boldsymbol{p}, \boldsymbol{p}_0)$ a hledání rozdělení na křivce největšího spádu v S.

Nechť $D_f(\boldsymbol{p}, \boldsymbol{p}_0)$ je kvazinorma na S. *Gradientním odhadem* rozdělení pravděpodobnosti $\boldsymbol{p} \in S$ z empirického rozdělení $(f_1/n, \ldots, f_m/n)$ rozumíme takové rozdělení pravděpodobnosti $\boldsymbol{p}(t) \in S$, že

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p}(t) = -\operatorname{grad} D_f(\boldsymbol{p}(t), \boldsymbol{p}_0) \quad \forall t \in \langle 0, \infty) \quad \text{a} \quad \boldsymbol{p}(0) = \boldsymbol{f}/n = (f_1/n, \ldots, f_m/n).$$

Jestliže funkce $f(u)$ generuje kvazinormu $D_f(\boldsymbol{p}, \boldsymbol{p}_0)$ na S, má výše uvedené vlastnosti a má spojitou derivaci $f'(u)$ pro každé $u \in (0, \infty)$, pak existuje jediný gradientní odhad $\boldsymbol{p}(t) = (p_1(t), \ldots, p_m(t))$ rozdělení pravděpodobnosti $\boldsymbol{p} \in S$. Jeho složky $p_1(t), \ldots, p_{m-1}(t)$ jsou $\forall t \in \langle 0, \infty)$ partikulárním řešením soustavy obyčejných diferenciálních rovnic prvního řádu

$$\begin{aligned}
p_1'(t) &= \quad -f'(mp_1(t)) \quad + f'\Big(m\Big[1 - \sum_{j=1}^{m-1} p_j(t)\Big]\Big), \\
&\vdots \\
p_{m-1}'(t) &= -f'(mp_{m-1}(t)) \quad + f'\Big(m\Big[1 - \sum_{j=1}^{m-1} p_j(t)\Big]\Big)
\end{aligned}$$

s počátečními podmínkami

$$p_1(0) = f_1/n, \ldots, p_{m-1}(0) = f_{m-1}/n$$

a složka $p_m(t) = 1 - \sum_{j=1}^{m-1} p_j(t) \quad \forall t \in \langle 0, \infty)$.

Vhodnou hodnotu parametru $t_0 \in \langle 0, \infty)$ nalezneme pomocí testu dobré shody jako takovou hodnotu t, kdy ještě nezamítáme hypotézu o vhodnosti rozdělení $\boldsymbol{p}(t)$ na hladině významnosti α. Gradientní odhad $\boldsymbol{p}(t)$ se pro rostoucí parametr t vzdaluje po křivce největšího spádu v S od empirického rozdělení k $\boldsymbol{p}_0$. Odhad $\boldsymbol{p}(t_0)$ je nejhorším z odhadů, které splňují zvolené testové kritérium na hladině významnosti alespoň α, proto ho nazýváme *pesimistickým gradientním odhadem*.

Nechť $f(u) = (u - 1)^2$. Pak $D_f(\boldsymbol{p}, \boldsymbol{p}_0) = \frac{1}{m} \sum_{j=1}^{m} (mp_j - 1)^2$ je tzv. *kvadratická kvazinorma*. Potom složky gradientního odhadu $\boldsymbol{p}(t) = (p_1(t), \ldots, p_m(t))$ z empirického rozdělení $(f_1/n, \ldots, f_m/n)$ jsou $\forall t \in \langle 0, \infty)$ partikulárním řešením nehomogenní lineární soustavy obyčejných diferenciálních rovnic prvního řádu s konstantními koeficienty a pravými stranami

$$\begin{aligned}
p_1'(t) &= -4mp_1(t) - 2mp_2(t) - \cdots - 2mp_{m-1}(t) + 2m, \\
p_2'(t) &= -2mp_1(t) - 4mp_2(t) - \cdots - 2mp_{m-1}(t) + 2m, \\
&\vdots \\
p_{m-1}'(t) &= -2mp_1(t) - 2mp_2(t) - \cdots - 4mp_{m-1}(t) + 2m
\end{aligned}$$

s počátečními podmínkami

$$p_1(0) = f_1/n, \ldots, p_{m-1}(0) = f_{m-1}/n$$

a složka $p_m(t) = 1 - \sum_{j=1}^{m-1} p_j(t) \quad \forall t \in \langle 0, \infty)$.

Řešením soustavy získáváme složky gradientního odhadu $\boldsymbol{p}(t)$

$$\begin{aligned}
p_1(t) &= c_1\,\mathrm{e}^{-2m^2t} + c_2\,\mathrm{e}^{-2mt} + 1/m,\\
p_2(t) &= c_1\,\mathrm{e}^{-2m^2t} + c_3\,\mathrm{e}^{-2mt} + 1/m,\\
&\vdots\\
p_{m-2}(t) &= c_1\,\mathrm{e}^{-2m^2t} + c_{m-1}\,\mathrm{e}^{-2mt} + 1/m,\\
p_{m-1}(t) &= c_1\,\mathrm{e}^{-2m^2t} - c_2\,\mathrm{e}^{-2mt} \cdots - c_{m-1}\,\mathrm{e}^{-2mt} + 1/m,\\
p_m(t) &= -(m-1)\,c_1\,\mathrm{e}^{-2m^2t} + 1/m,
\end{aligned}$$

kde

$$\begin{aligned}
c_1 &= \frac{f_1/n + f_2/n + \cdots + f_{m-1}/n}{m-1} - \frac{1}{m},\\
c_2 &= \frac{(m-2)\,f_1/n - f_2/n - \cdots - f_{m-1}/n}{m-1},\\
c_3 &= \frac{-f_1/n + (m-2)\,f_2/n - \cdots - f_{m-1}/n}{m-1},\\
&\vdots\\
c_{m-1} &= \frac{-f_1/n - f_2/n - \cdots + (m-2)\,f_{m-2}/n - f_{m-1}/n}{m-1}.
\end{aligned}$$

Složky takto získaného gradientního odhadu z empirického rozdělení jsou asymptoticky nestrannými odhady složek pozorovaného rozdělení pravděpodobnosti $\boldsymbol{p}$.

Výpočet hodnoty parametru t_0 je dosti citlivý na „strmost“ zvolené kvazinormy, související numerické problémy s řešením odpovídajících nelineárních diferenciálních rovnic lze alespoň částečně zmenšit vhodným kladným násobkem kvazinormy. Kvadratická kvazinorma je přitom až na tento násobek jediná, která vede na lineární soustavu diferenciálních rovnic s konstantními koeficienty.

Kapitola 10

Intervalové odhady diskrétního rozdělení pravděpodobnosti kategoriální veličiny

V minulé kapitole jsme uvedli dva odhady diskrétního rozdělení pravděpodobnosti kategoriální veličiny, a to odhad relativními četnostmi a pesimistický gradientní odhad pomocí kvadratické kvazinormy (dále zkráceně jen pesimistický gradientní odhad). V obou případech se ale jedná pouze o bodové odhady. Jednou z možností, jak získat intervalové odhady, je použití metody bootstrap. V této kapitole uvedeme dva příklady kategoriálních veličin, u nichž předvedeme získání intervalových odhadů jejich rozdělení pravděpodobnosti metodou bootstrap. Dále provedeme srovnání intervalových odhadů získaných pomocí obou výše uvedených bodových odhadů a také ukážeme vliv rozsahu pozorovaného náhodného výběru, stejně jako vliv počtu bootstrapových výběrů B.

K simulacím a výpočtům uvedeným v této kapitole jsme použili matematický software Shine bootstrap a Matlab R2008a. Shine bootstrap byl vytvořen speciálně pro generování náhodných výběrů z diskrétních rozdělení pravděpodobnosti a pro výpočty bodových odhadů pravděpodobnostních funkcí těchto rozdělení. Zejména je v něm implementován výpočet pesimistického gradientního odhadu podle postupu předloženého v předchozí kapitole. Matlab R2008a byl použit zejména k výpočtu kvantilů a generování histogramů.

10.1. Falešná kostka

V prvním příkladě budeme uvažovat hrací kostku o šesti stranách, které si označíme čísly $1, \ldots, 6$. Házíme-li kostkou, pozorujeme diskrétní náhodnou veličinu X — číslo, které padne. Základní prostor je tedy tvořen šesti elementárními náhodnými jevy, které odpovídají číslům $1, \ldots, 6$. V případě, kdy kostka není falešná, je pravděpodobnostní funkce této náhodné veličiny $\boldsymbol{p} = (p_1, \ldots, p_6) = (1/6, \ldots, 1/6) \doteq (0{,}1667; \ldots; 0{,}1667)$. My ale budeme uvažovat kostku falešnou, např. takovou, která má poněkud těžší stranu s číslem 6. Její pravděpodobnostní funkci zvolíme $\boldsymbol{p} = (p_1, \ldots, p_6) = (0{,}10; 0{,}15; 0{,}15; 0{,}15; 0{,}15; 0{,}30)$. Pozorování náhodné veličiny X realizujeme na počítači jako náhodný výběr z diskrétního rozdělení se zadanou pravděpodobnostní funkcí. Protože chceme mimo jiné zkoumat vliv rozsahu n pozorovaného náhodného výběru, realizujeme tři pozorování a získáme tři náhodné výběry z veličiny X, a to o rozsazích 60, 100 a 400. Pro všechny tři náhodné výběry vypočítáme bodové

	strana 1	strana 2	strana 3	strana 4	strana 5	strana 6	rozsah n
skutečné $\boldsymbol{p}$	0,10	0,15	0,15	0,15	0,15	0,30	
odhad	0,1167	0,1333	0,1333	0,0833	0,1667	0,3667	60
relativními	0,1600	0,1400	0,1800	0,1100	0,1000	0,3100	100
četnostmi	0,1200	0,1425	0,1525	0,1350	0,1500	0,3000	400
pesimistický	0,1235	0,1397	0,1397	0,0911	0,1720	0,3341	60
gradientní	0,1642	0,1448	0,1835	0,1157	0,1060	0,2859	100
odhad	0,1228	0,1449	0,1548	0,1375	0,1523	0,2877	400

Tab. 10.1: Bodové odhady pravděpodobnostní funkce falešné kostky

odhady pravděpodobnostní funkce $\boldsymbol{p}$, a to oběma výše uvedenými způsoby, tj. bodový odhad relativními četnostmi i bodový pesimistický gradientní odhad. Výsledky jsou zaznamenány v tabulce 10.1.

Vidíme, že pro strany $1, \ldots, 5$ je vždy bodový pesimistický gradientní odhad vyšší než příslušný bodový odhad relativními četnostmi, zatímco pro stranu 6 je vždy nižší. Pesimistický gradientní odhad má tedy tendenci snižovat rozdíly mezi jednotlivými pravděpodobnostmi oproti pozorovaným relativním četnostem. Pro strany $2, \ldots, 6$ jsme obdrželi nejpřesnější bodové odhady jednotlivých pravděpodobností pro náhodný výběr o rozsahu 400, u všech stran $1, \ldots, 6$ jsou odhady jednotlivých pravděpodobností pro náhodný výběr o rozsahu 400 blíže skutečným hodnotám než odhady pro náhodný výběr o rozsahu 100. Z toho lze odvodit, že přesnost bodových odhadů stoupá s rozsahem pozorovaného náhodného výběru. Srovnáním přesnosti všech 18 bodových odhadů jednotlivých pravděpodobností získaných jako odhad relativními četnostmi a jako pesimistický gradientní odhad zjišťujeme, že v 9 případech byl přesnější odhad relativními četnostmi a v 9 případech byl přesnější pesimistický gradientní odhad. Bodové odhady získané oběma metodami jsou tedy zhruba stejně přesné, žádná z metod se nejeví být lepší. Bude nás zajímat, zda bude pozorovatelný nějaký rozdíl přesnosti obou metod u intervalových odhadů.

Nejjednodušší a výpočtově nejméně náročný způsob zisku intervalového odhadu je v tomto případě jednoduchý kvantilový bootstrapový konfidenční interval, který nevyžaduje žádné dodatečné výpočty směrodatných odchylek, vychýlení mediánu, akcelerace ani reziduí. Jednoduše realizujeme B bootstrapových výběrů z pozorovaného náhodného výběru o rozsahu n a pro každý z nich vypočítáme bodový odhad pravděpodobnostní funkce. Pro každou stranu kostky (každý elementární náhodný jev náhodné veličiny X) tak dostaneme B hodnot (bodových odhadů). Jejich α-kvantil a $(1-\alpha)$-kvantil pak jsou krajními body intervalového odhadu se spolehlivostí $1-2\alpha$. My přitom celý postup zopakujeme pro všechny tři náhodné výběry o rozsazích 60, 100 a 400 a také pro oba odhady pravděpodobnostní funkce, tj. pro odhad relativními četnostmi i pro pesimistický gradientní odhad. Protože chceme dále zkoumat i vliv počtu bootstrapových výběrů, zvolíme také dvě různé hodnoty B, a to 1000 a 5000. Výsledky jsou zaznamenány v tabulce 10.2, volili jsme spolehlivost 0,95.

Celkem je tedy v tabulce 72 intervalových odhadů. Každý z těchto intervalových odhadů přitom obsahuje příslušnou skutečnou hodnotu pravděpodobnosti daného náhodného jevu. Dále je ve všech případech splněno, že šířka intervalu klesá se stoupajícím rozsahem pozorovaného náhodného výběru. Ve 36 případech můžeme srovnat šířku intervalu získaného pomocí

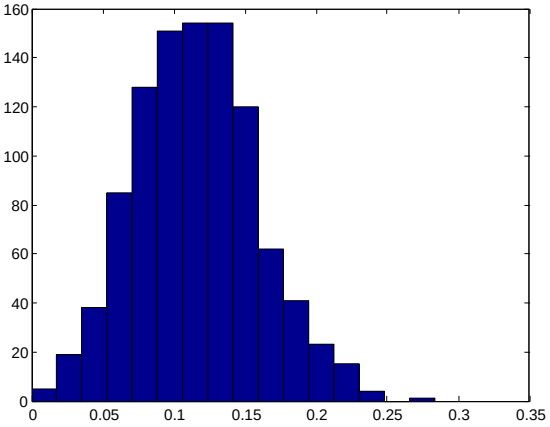
	strana 1		strana 2		strana 3		rozsah n	počet B
skutečné $\boldsymbol{p}$	0,10		0,15		0,15			
	0,0500	0,2000	0,0500	0,2167	0,0500	0,2167	60	1000
odhad	0,0500	0,2000	0,0500	0,2167	0,0500	0,2167	60	5000
relativními	0,0900	0,2400	0,0800	0,2100	0,1200	0,2700	100	1000
četnostmi	0,0900	0,2300	0,0800	0,2100	0,1200	0,2700	100	5000
	0,0900	0,1525	0,1100	0,1800	0,1187	0,1900	400	1000
	0,0900	0,1525	0,1075	0,1775	0,1175	0,1900	400	5000
	0,0573	0,2042	0,0585	0,2205	0,0580	0,2354	60	1000
pesimistický	0,0572	0,2043	0,0583	0,2206	0,0582	0,2206	60	5000
gradientní	0,0959	0,2411	0,0769	0,2128	0,1155	0,2609	100	1000
odhad	0,0960	0,2318	0,0859	0,2127	0,1154	0,2612	100	5000
	0,0931	0,1548	0,1129	0,1792	0,1202	0,1917	400	1000
	0,0931	0,1548	0,1128	0,1793	0,1203	0,1892	400	5000

	strana 4		strana 5		strana 6		rozsah n	počet B
skutečné $\boldsymbol{p}$	0,15		0,15		0,30			
	0,0167	0,1667	0,0833	0,2667	0,2500	0,4833	60	1000
odhad	0,0167	0,1667	0,0833	0,2667	0,2500	0,4833	60	5000
relativními	0,0500	0,1700	0,0400	0,1600	0,2200	0,4100	100	1000
četnostmi	0,0500	0,1700	0,0400	0,1600	0,2200	0,4000	100	5000
	0,1025	0,1725	0,1150	0,1875	0,2550	0,3425	400	1000
	0,1025	0,1700	0,1175	0,1875	0,2575	0,3450	400	5000
	0,0241	0,1558	0,0751	0,2674	0,2230	0,4670	60	1000
pesimistický	0,0242	0,1559	0,0901	0,2676	0,2219	0,4662	60	5000
gradientní	0,0567	0,1830	0,0561	0,1641	0,2005	0,3803	100	1000
odhad	0,0566	0,1800	0,0475	0,1641	0,2019	0,3750	100	5000
	0,1055	0,1719	0,1178	0,1843	0,2422	0,3347	400	1000
	0,1078	0,1719	0,1178	0,1868	0,2434	0,3323	400	5000

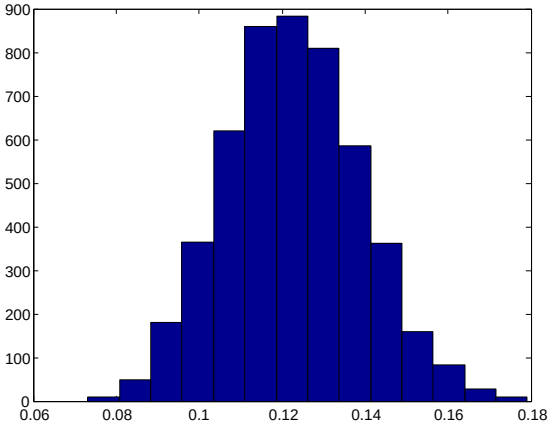
Tab. 10.2: Bootstrapové intervalové odhady pravděpodobnostní funkce falešné kostky

1000 bootstrapových výběrů a pomocí 5000 bootstrapových výběrů. Pouze v 8 případech jsme pro $B = 5000$ obdrželi širší intervalový odhad než pro $B = 1000$, přičemž největší rozšíření intervalu bylo o 0,0086. V 15 případech byly oba intervaly stejně široké, v ostatních 13 případech se interval s rostoucím B zúžil. Ve 36 případech můžeme také srovnat šířku intervalového odhadu relativními četnostmi a intervalového pesimistického gradientního odhadu. Ve 26 z nich byl intervalový pesimistický gradientní odhad užší.

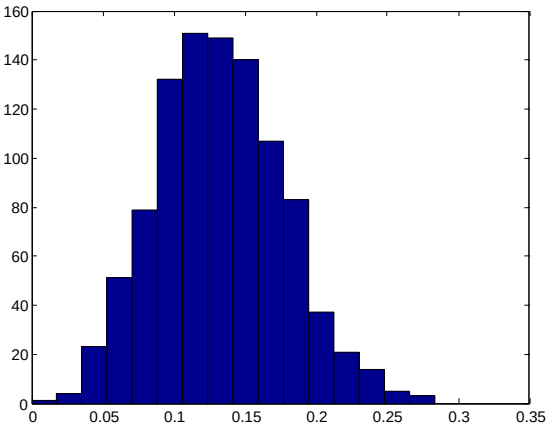
Na následujících stranách uvádíme 12 histogramů B bodových odhadů pro jednotlivé strany kostky (jednotlivé náhodné jevy). Pro každou ze stran 1, ..., 6 jsou to vždy případy $n = 60$ a $n = 400$, kterým odpovídají nejširší a nejužší ze všech intervalových odhadů pro tu kterou stranu kostky. U každého histogramu je uvedena hodnota B, a zda se v daném případě jedná o intervalový odhad relativními četnostmi nebo intervalový pesimistický gradientní odhad.



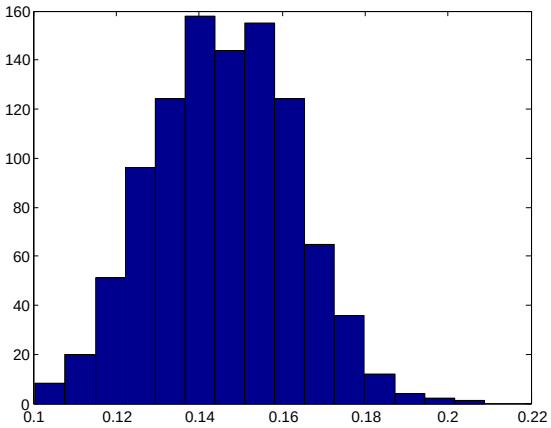
Strana 1, $n = 60$, $B = 1000$,
odhad relativními četnostmi



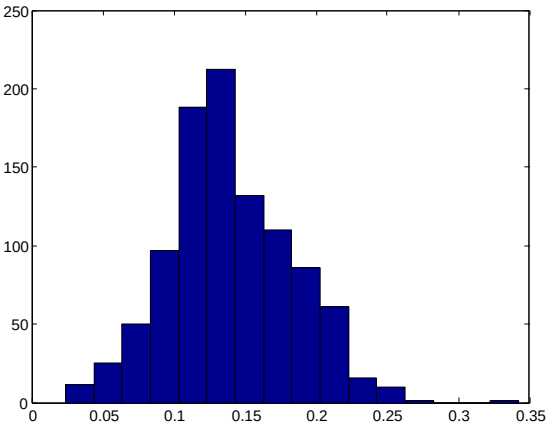
Strana 1, $n = 400$, $B = 5000$,
pesimistický gradientní odhad



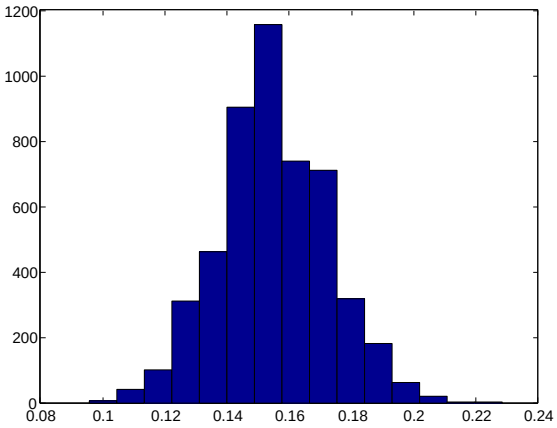
Strana 2, $n = 60$, $B = 1000$,
odhad relativními četnostmi



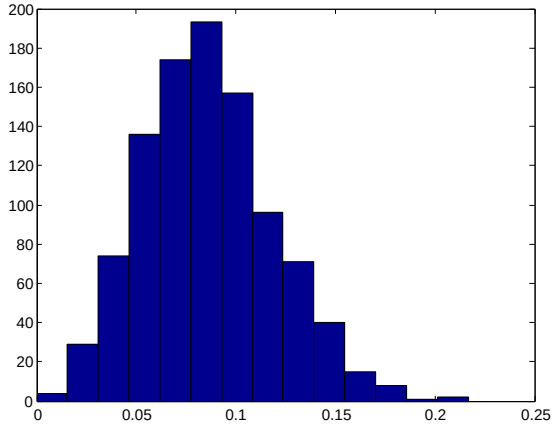
Strana 2, $n = 400$, $B = 1000$,
pesimistický gradientní odhad



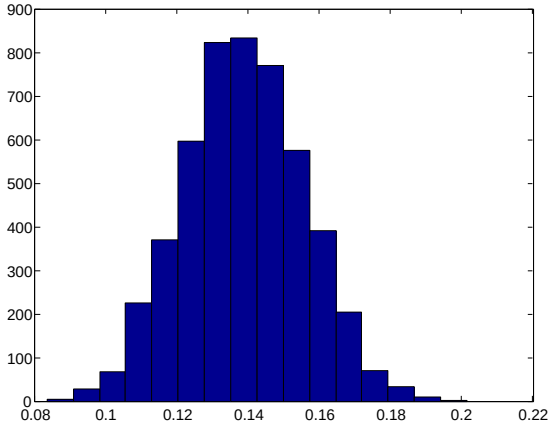
Strana 3, $n = 60$, $B = 1000$,
pesimistický gradientní odhad



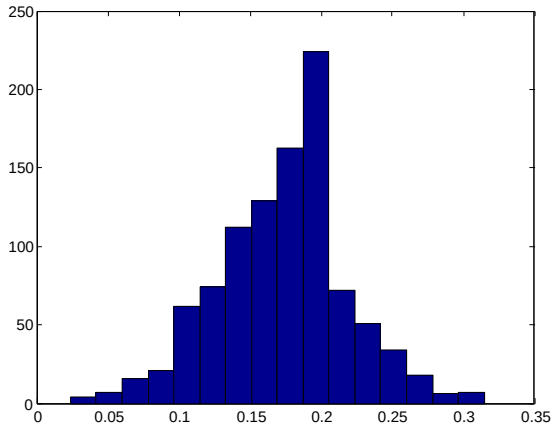
Strana 3, $n = 400$, $B = 5000$,
pesimistický gradientní odhad



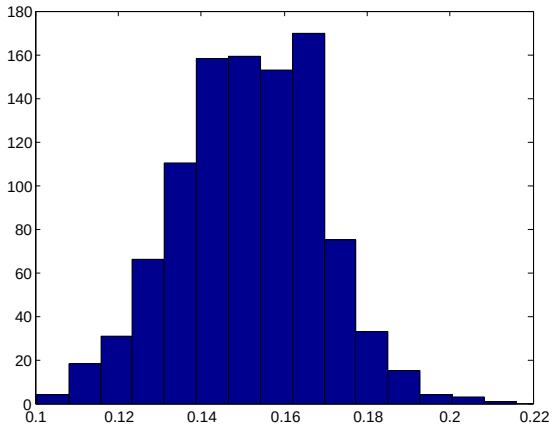
Strana 4, $n = 60$, $B = 1000$,
odhad relativními četnostmi



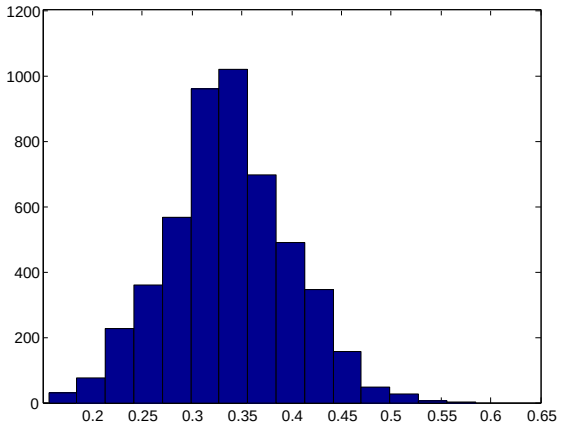
Strana 4, $n = 400$, $B = 5000$,
pesimistický gradientní odhad



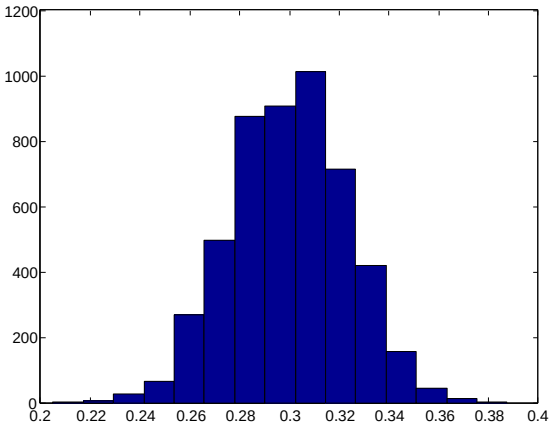
Strana 5, $n = 60$, $B = 1000$,
pesimistický gradientní odhad



Strana 5, $n = 400$, $B = 1000$,
pesimistický gradientní odhad



Strana 6, $n = 60$, $B = 5000$,
pesimistický gradientní odhad



Strana 6, $n = 400$, $B = 5000$,
odhad relativními četnostmi

Podstatnou otázkou také je, zda jsou získané intervalové odhady pravděpodobnostní funkce dostatečně úzké, abychom na základě nich byli schopni rozhodnout, zda kostka je nebo není falešná. Testujme hypotézu $H\colon p_6 = 1/6$ proti alternativní hypotéze $\overline{H}\colon p_6 \neq 1/6$. Z tabulky je patrné, že již pro náhodný výběr o rozsahu 60 (a dále i pro oba větší rozsahy) žádný ze čtyř intervalových odhadů pro stranu s číslem 6 neobsahuje hodnotu 0,1667. Tedy na hladině významnosti 0,05 zamítáme hypotézu H a nezamítáme alternativní hypotézu $\overline{H}$. Z toho ovšem vyplývá, že kostka je falešná. Vzhledem k tomu, že pro $\alpha < 0{,}5$ platí, že 0-kvantil $\leqq \alpha$-kvantil $\leqq 2\alpha$-kvantil $\leqq (1-2\alpha)$-kvantil $\leqq (1-\alpha)$-kvantil $\leqq$ 1-kvantil, můžeme bez dalších výpočtů říci, že také při testu hypotézy H proti alternativní hypotéze $\overline{H}\colon p_6 \geqq 1/6$ zamítáme H a nezamítáme $\overline{H}$. Strana s číslem 6 padá na této kostce příliš často.

Dále pro náhodný výběr o rozsahu 100 vidíme, že ani žádný ze čtyř intervalových odhadů pro stranu s číslem 5 neobsahuje hodnotu 0,1667, pro náhodný výběr o rozsahu 400 žádný ze čtyř intervalových odhadů pro stranu s číslem 1 neobsahuje hodnotu 0,1667. Podobně jako v předchozím odstavci dojdeme k závěru, že strany s čísly 1 a 5 naopak padají příliš málo.

Závěrem lze říci, že metodu bootstrap lze pro zisk intervalových odhadů pravděpodobnostní funkce této kategoriální veličiny doporučit, získané intervalové odhady jsou dostatečně přesné a úzké. Dalšího zlepšení lze dosáhnout větším rozsahem pozorovaného náhodného výběru, větším počtem bootstrapových výběrů a použitím metody pesimistického gradientního odhadu.

10.2. Volební model

Ve druhém příkladu se budeme věnovat předvolebnímu průzkumu veřejného mínění. Ve dnech 17. až 18. října 2008 se konaly volby do krajských zastupitelstev. Hlasy voličů se v tomto případě sčítají pro každý kraj samostatně, zaměříme se proto např. na Jihomoravský kraj. Ve dnech 5. až 7. října 2008 realizovala společnost CS&C pro Českou televizi tzv. bleskový předvolební průzkum veřejného mínění (viz [10]) občanů Jihomoravského kraje. Respondentů bylo 633, z nich pouze 306 bylo zahrnuto do tzv. volebního modelu. Do volebního modelu se zahrnují pouze odpovědi těch respondentů, kteří chtějí jít volit a zároveň už vědí, koho budou volit. K volbám se nakonec v Jihomoravském kraji dostavilo 377 706 voličů (jejichž hlasy byly platné — viz [9]). Kandidovalo celkem 18 politických stran a hnutí. Ve volebním modelu je uvedeno 8 z nich, které měly nejvyšší preference, všechny ostatní tvoří devátou položku modelu.

Volební model je tedy kategoriální veličina X, která může nabývat devíti slovních hodnot (elementárních náhodných jevů), a to ODS, ČSSD, KSČM, KDU-ČSL, SZ, SNK-ED, NEZÁVISLÍ, Moravané a ostatní. Pozorováním kategoriální veličiny X byl získán náhodný výběr o rozsahu 306. V závěrečné zprávě průzkumu jsou uvedeny absolutní četnosti pozorovaných hodnot a na jejich základě vytvořený odhad procentuálního zisku jednotlivých politických stran v nadcházejících volbách. Je uveden bodový odhad relativními četnostmi a dále i intervalový odhad se spolehlivostí 0,95, není ale zveřejněno, jakou metodou byl získán. K jeho výpočtu nicméně byla použita asymptotická metoda, která předpokládá absolutní četnost pozorování každého elementárního náhodného jevu alespoň 5. Tento předpoklad ale nebyl splněn.

My odhadneme složky pravděpodobnostní funkce $\boldsymbol{p}$ kategoriální veličiny X oběma výše uvedenými bodovými odhady, tj. realizujeme odhad relativními četnostmi a pesimistický gradientní odhad. Pro zisk intervalových odhadů použijeme opět jednoduchý kvantilový bootstra-

pový konfidenční interval. Realizujeme B bootstrapových výběrů z pozorovaného náhodného výběru a pro každý z nich vypočítáme odhad pravděpodobnostní funkce. Pro každý elementární náhodný jev tak dostaneme B hodnot (bodových odhadů). Jejich α-kvantil a $(1-\alpha)$--kvantil pak jsou krajními body intervalového odhadu se spolehlivostí $1-2\alpha$. Celý postup přitom zopakujeme pro oba bodové odhady, tj. odhad relativními četnostmi i pesimistický gradientní odhad. Zajímá nás také vliv počtu bootstrapových výběrů, proto budeme volit dvě různé hodnoty B, a to 1000 a 5000. Spolehlivost volíme 0,95, abychom mohli provést relevantní srovnání s intervalovými odhady uvedenými v [10]. Pozorovaný náhodný výběr, odhady převzaté z [10], výsledky provedených výpočtů i skutečný výsledek voleb jsou uvedeny v tabulce 10.3.

Vidíme, že pro prvních pět elementárních náhodných jevů, jejichž relativní četnosti přesahují 0,05, je bodový pesimistický gradientní odhad nižší než bodový odhad relativními četnostmi, zatímco pro následující čtyři elementární jevy, jejichž relativní četnosti nedosahují ani 0,03, je bodový pesimistický gradientní odhad vyšší než bodový odhad relativními četnostmi. Opět se tedy potvrzuje, že pesimistický gradientní odhad má tendenci snižovat rozdíly mezi jednotlivými pravděpodobnostmi oproti pozorovaným relativním četnostem.

V tabulce je zaznamenáno celkem 45 intervalových odhadů, z toho 9 jsme převzali z [10] a 36 jsme vypočítali metodou bootstrap, a to 18 pomocí odhadu relativními četnostmi a 18 pomocí pesimistického gradientního odhadu. Můžeme tedy právě v 18 případech srovnat intervalový odhad relativními četnostmi s intervalovým pesimistickým gradientním odhadem. Ve 12 případech je intervalový pesimistický gradientní odhad užší než příslušný intervalový odhad relativními četnostmi, v 1 případě byly oba intervaly stejně široké a pouze v 5 případech je intervalový pesimistický gradientní odhad širší než příslušný intervalový odhad relativními četnostmi. Opět tedy vidíme, že intervalové pesimistické gradientní odhady jsou přesnější. Dále je patrné, že pro prvních 5 elementárních náhodných jevů jsou intervalové pesimistické gradientní odhady oproti intervalovým odhadům relativními četnostmi posunuté doleva, zatímco pro zbylé 4 elementární jevy naopak doprava, čímž se opět potvrzuje, že pesimistický gradientní odhad má tendenci snižovat rozdíly mezi jednotlivými pravděpodobnostmi oproti pozorovaným relativním četnostem.

V 18 případech také můžeme srovnat intervaly získané pomocí $B = 1000$ a $B = 5000$ bootstrapových výběrů. V 8 případech se interval s rostoucím B zúžil, v 6 případech jsme získali zcela stejné (tudíž i stejně široké) intervaly a ve 4 případech se interval s rostoucím B rozšířil, a to nejvýše o 0,0049. Vyšší počet bootstrapových výběrů má tedy i v tomto příkladu kategoriální veličiny pozitivní vliv na intervalové odhady pravděpodobnostní funkce.

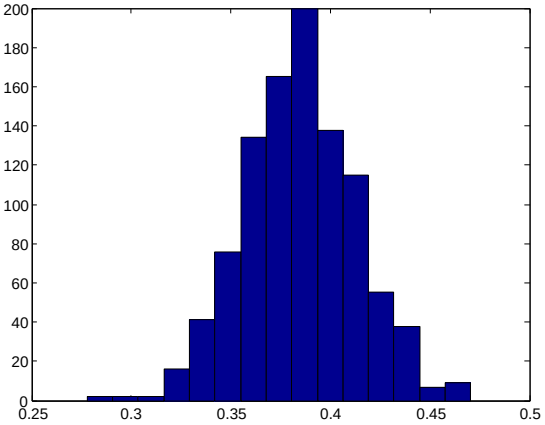
Na následujících stranách uvádíme 18 histogramů B bodových odhadů pro jednotlivé elementární náhodné jevy. Pro každý z jevů 2 histogramy, které odpovídají nejširšímu a nejužšímu intervalovému odhadu pro daný jev. U každého histogramu je uvedeno, zda se v daném případě jedná o intervalový odhad relativními četnostmi, nebo intervalový pesimistický gradientní odhad, a pomocí kolika bootstrapových výběrů byl získán.

Dále je třeba srovnat převzaté intervalové odhady s těmi, které jsme vypočítali metodou bootstrap. Krajní meze převzatých intervalových odhadů byly autory zaokrouhleny na celá procenta, tj. na dvě desetinná místa. Pro prvních šest elementárních náhodných jevů lze říci, že pokud bychom zaokrouhlili hodnoty krajních mezí bootstrapových intervalových odhadů na dvě desetinná místa, obdrželi bychom takřka vždy, konkrétně ve 45 případech z 48, odhady převzaté z [10], ve zbylých 3 případech hodnotu o 0,01 vyšší. U zbylých tří elementárních jevů

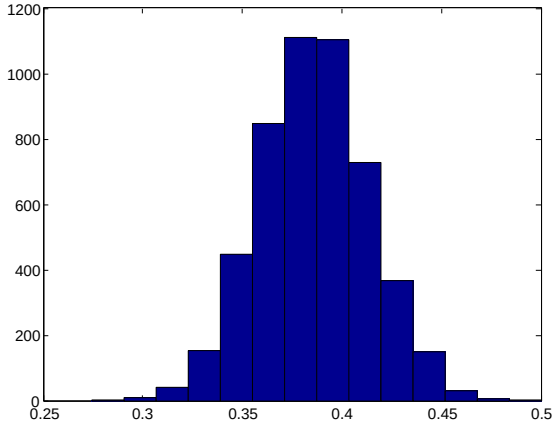
	ČSSD		ODS		KSČM		KDU-ČSL		SZ		počet B
absolutní četnosti	118		71		35		52		17		
odhad SC&C	0,33	0,44	0,18	0,28	0,8	0,15	0,13	0,21	0,3	0,8	
odhad relativními četnostmi	0,3856		0,2320		0,1144		0,1699		0,0556		
pesimistický gradientní odhad	0,3815		0,2296		0,1133		0,1682		0,0551		
intervalový odhad	0,3301	0,4444	0,1863	0,2810	0,0784	0,1503	0,1291	0,2124	0,0294	0,0817	1000
relativními četnostmi	0,3333	0,4412	0,1863	0,2810	0,0784	0,1503	0,1275	0,2157	0,0327	0,0817	5000
intervalový pesimistický	0,3270	0,4389	0,1846	0,2788	0,0778	0,1490	0,1277	0,2113	0,0293	0,0812	1000
gradientní odhad	0,3279	0,4368	0,1842	0,2776	0,0780	0,1494	0,1265	0,2128	0,0325	0,0810	5000
výsledky voleb	0,3484		0,1588		0,1441		0,2389		0,0364		
	NEZÁVISLÍ		Moravané		SNK-ED		ostatní				počet B
absolutní četnosti	3		4		2		4				
odhad SC&C	0,00	0,02	0,00	0,02	0,00	0,01	0,00	0,02			
odhad relativními četnostmi	0,0098		0,0131		0,0065		0,0131				
pesimistický gradientní odhad	0,0099		0,0132		0,0067		0,0225				
intervalový odhad	0,0000	0,0229	0,0033	0,0294	0,0000	0,0163	0,0033	0,0261			1000
relativními četnostmi	0,0000	0,0229	0,0030	0,0261	0,0000	0,0163	0,0033	0,0261			5000
intervalový pesimistický	0,0002	0,0228	0,0033	0,0293	0,0001	0,0165	0,0093	0,0382			1000
gradientní odhad	0,0002	0,0228	0,0033	0,0261	0,0001	0,0164	0,0093	0,0383			5000
výsledky voleb	0,0189		0,0089		0,0139		0,0417				

Tab. 10.3: Volební model a výsledky voleb do krajských zastupitelstev 2008

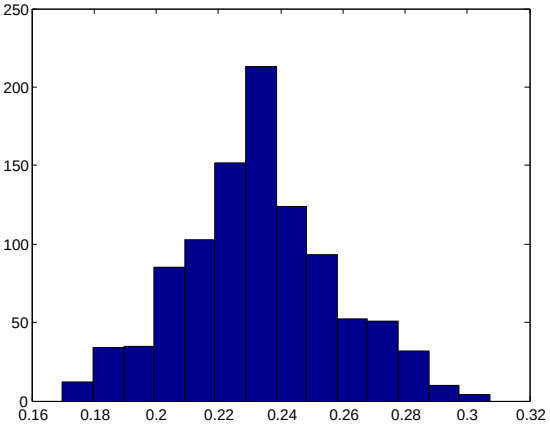
se po zaokrouhlení shodují pouze dolní meze intervalů, a to v 10 případech z 12, horní meze



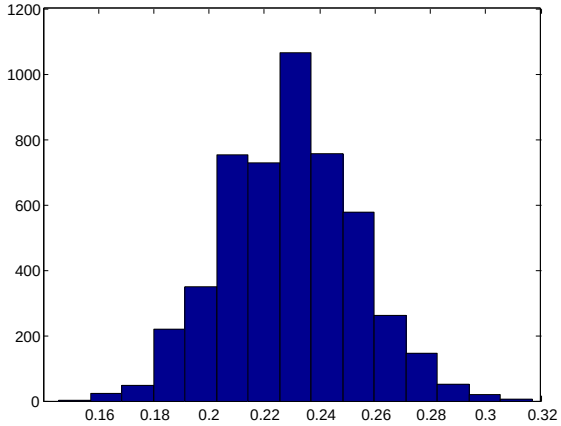
ČSSD, $B = 1000$,
odhad relativními četnostmi



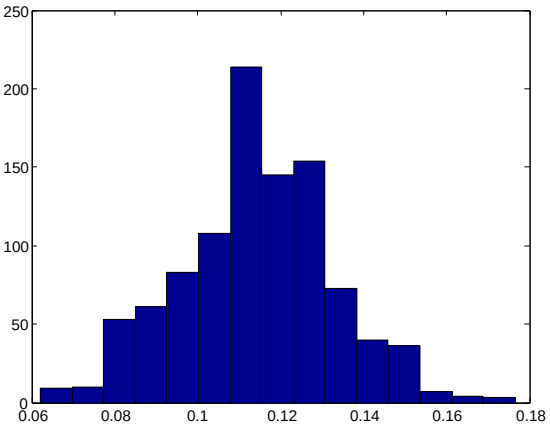
ČSSD, $B = 5000$,
odhad relativními četnostmi



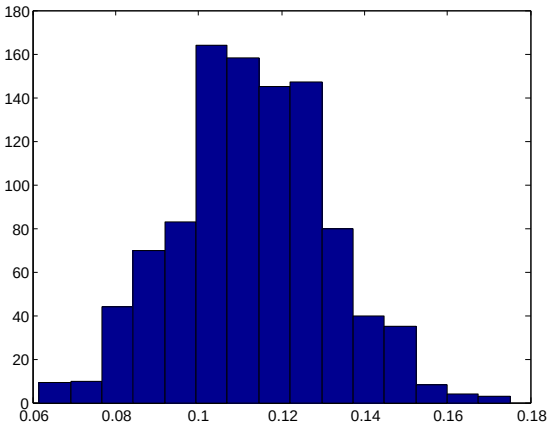
ODS, $B = 1000$,
odhad relativními četnostmi



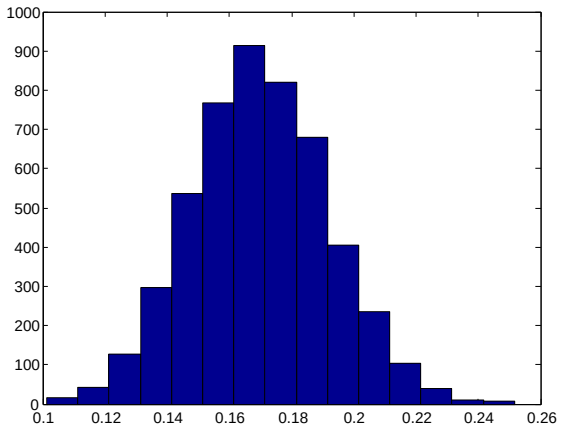
ODS, $B = 5000$,
pesimistický gradientní odhad



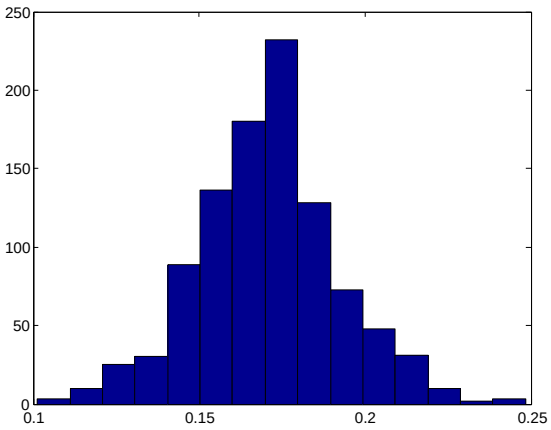
KSČM, $B = 1000$,
odhad relativními četnostmi



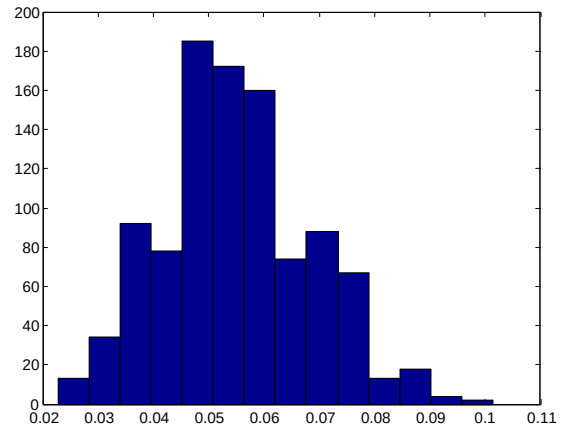
KSČM, $B = 1000$,
pesimistický gradientní odhad



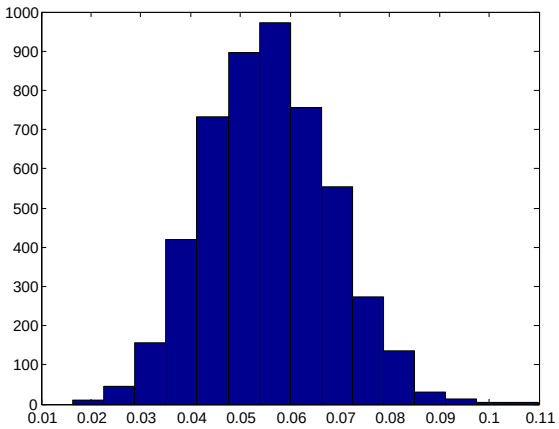
KDU-ČSL, $B = 5000$,
odhad relativními četnostmi



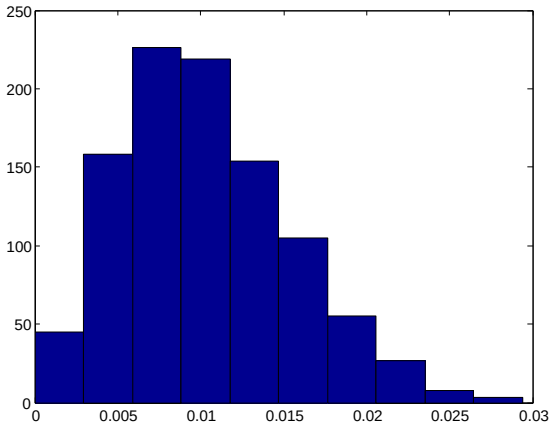
KDU-ČSL, $B = 1000$,
odhad relativními četnostmi



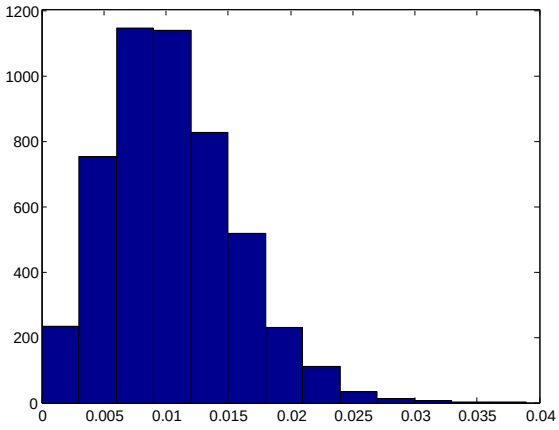
SZ, $B = 1000$,
odhad relativními četnostmi



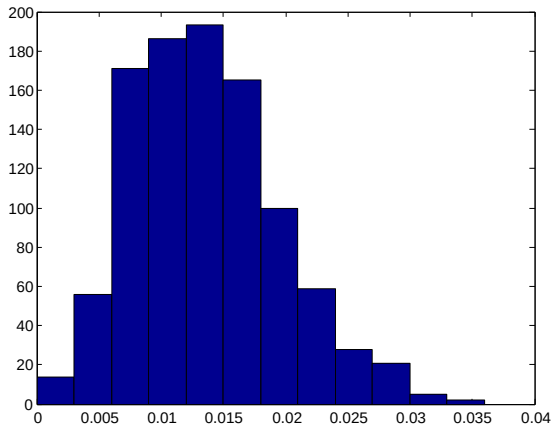
SZ, $B = 5000$,
pesimistický gradientní odhad



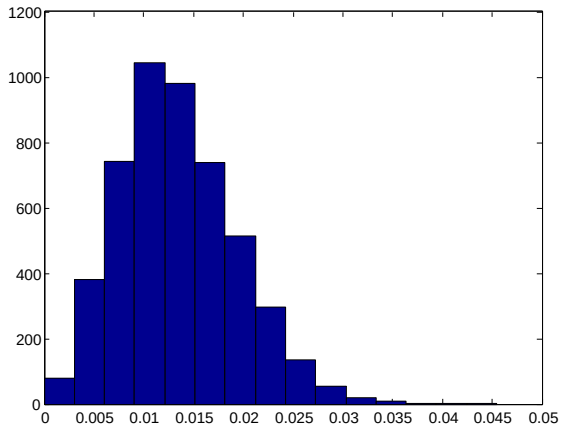
NEZÁVISLÍ, $B = 1000$,
odhad relativními četnostmi



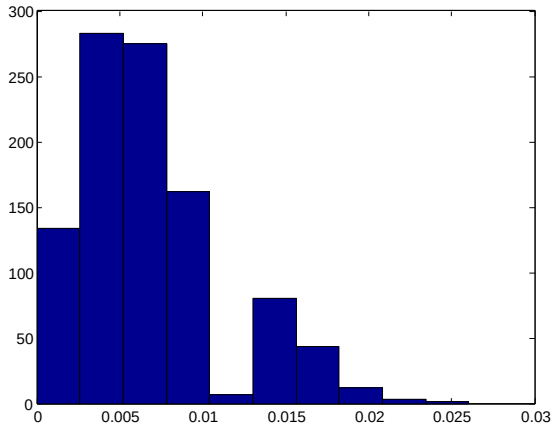
NEZÁVISLÍ, $B = 5000$,
pesimistický gradientní odhad



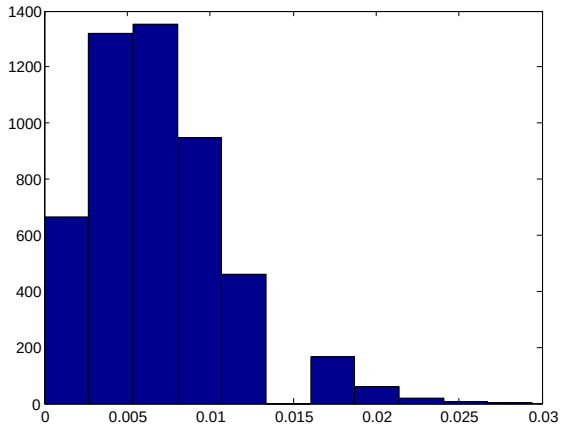
Moravané, $B = 1000$,
odhad relativními četnostmi



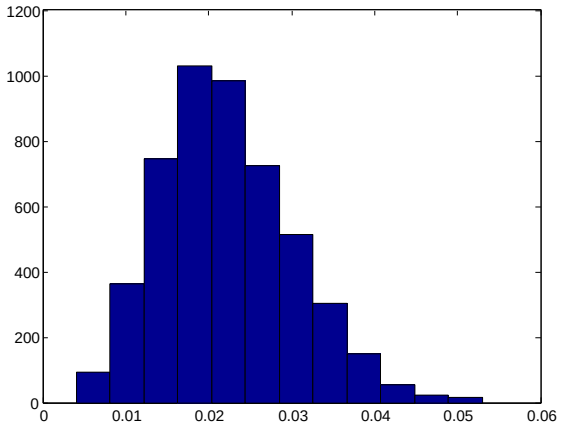
Moravané, $B = 5000$,
pesimistický gradientní odhad



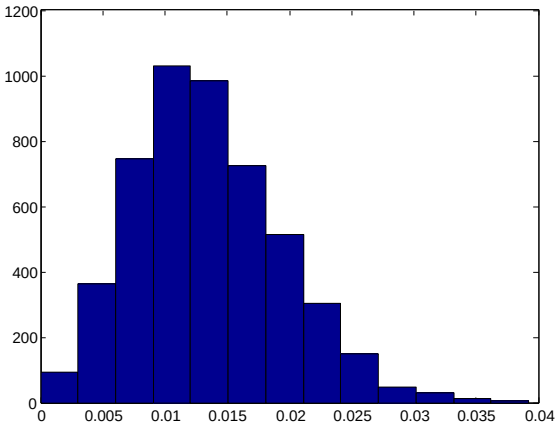
SNK-ED, $B = 1000$,
pesimistický gradientní odhad



SNK-ED, $B = 5000$,
odhad relativními četnostmi



ostatní, $B = 5000$,
pesimistický gradientní odhad



ostatní, $B = 5000$,
odhad relativními četnostmi

všech bootstrapových intervalových odhadů jsou vyšší než horní meze převzatých intervalů, a to o 0,0061 až 0,0183. Jedná se o ty elementární náhodné jevy, u nichž byly pozorované absolutní četnosti nižší než 5, a tedy nebyly splněny předpoklady asymptotické metody použité pro výpočet převzatých intervalových odhadů. Lze se tedy domnívat, že metoda bootstrap v těchto případech přináší „spolehlivější“ intervalové odhady.

Srovnáme-li převzaté i vypočítané intervalové odhady se skutečným výsledkem voleb, zjistíme, že pouze 5 z 9 převzatých intervalových odhadů obsahuje skutečnou hodnotu volebního výsledku, u elementárního náhodného jevu ODS byl volební výsledek nižší, naopak u jevů KDU-ČSL, SNK-ED a ostatní byl vyšší. U jevů ODS, KDU-ČSL a ostatní není skutečný volební výsledek ani prvkem žádného z příslušných bootstrapových intervalových odhadů, zatímco u jevu SNK-ED je prvkem všech čtyř příslušných bootstrapových intervalových odhadů. Dokázali jsme tedy odhadnout volební výsledek úspěšněji než metoda použitá v praxi, i tak ale pouze u šesti elementárních náhodných jevů z devíti. Příčinu musíme hledat zejména v nízkém rozsahu pozorovaného náhodného výběru, 306 respondentů volebního modelu se ukazuje jako zcela nedostatečné množství.

Závěrem lze říci, že metoda bootstrap se ukázala pro odhad pravděpodobnostní funkce volebního modelu jako srovnatelná nebo mírně lepší než v praxi používaná metoda a lze ji doporučit. Zlepšení intervalových odhadů lze dosáhnout zvýšením rozsahu pozorovaného náhodného výběru, zvýšením počtu bootstrapových výběrů a použitím pesimistického gradientního odhadu.

Kapitola 11

Závěr

Připomeneme největší výhody metody bootstrap, jak je uvádí [2]. Umožňuje odhadnout přesnost odhadu, získaného třeba i velmi složitým postupem, užitím výkonu a rychlosti počítače. Její použití zbavuje nutnosti studovat do hloubky teorii a složitě odvozovat přesné vztahy, navíc přináší řešení i v případech, kdy analytické odvození neznáme. Může být použita parametricky i neparametricky. Odhady odvozené neparametrickým bootstrapem jsou pro dostatečně rozsáhlé výběry přesné bez ohledu na pozorované rozdělení pravděpodobnosti. Přitom často jsou dostatečně přesné již pro velmi malé rozsahy. Nicméně základní bootstrap je jen hrubým odhadem přesnosti, který by měl být použit pouze tehdy, když není možné realizovat rozsáhlejší výpočty. Přednost dáváme vždy zkonstruování některého z konfidenčních intervalů pro daný odhad.

Jak se uvádí v [7], metoda pesimistického gradientního odhadu pravděpodobnostní funkce diskrétního rozdělení pravděpodobnosti kategoriální veličiny již na řadě úloh ukázala svoji použitelnost v kategoriální analýze. Až dosud byla ale používána pouze pro bodové odhady. My jsme nyní na příkladech demonstrovali, že ve spojení s metodou bootstrap umožňuje také získat uspokojivé intervalové odhady pravděpodobnostní funkce.

Naším úkolem bylo také otestovat možnosti využití softwaru Shine bootstrap. Jeho použití k výpočtu pesimistického gradientního odhadu pomocí kvadratické kvazinormy lze rozhodně doporučit, bohužel však neumožňuje výpočet kvantilů, a tak je při konstruování intervalových odhadů třeba používat i další software.

Shine bootstrap také umožňuje výpočet maximálně věrohodných bodových odhadů všech tří parametrů tříparametrického Weibullova rozdělení pravděpodobnosti. Je známo, že je-li jeho prahový parametr nenulový, je jeho hodnotu velmi obtížné odhadnout. Shine bootstrap má implementován řešič NOMAD, který funguje na principu ortogonálního prohledávání bez využití derivací. Bohužel se však ukázalo, že spojením takto získaných bodových odhadů parametrů Weibullova rozdělení s metodou bootstrap získáme příliš široké a v případě prahového parametru také příliš vychýlené intervalové odhady, proto tento postup nelze doporučit. Pro získání uspokojivých intervalových odhadů bude zřejmě třeba vyzkoušet implementaci jiných metod výpočtu bodových odhadů parametrů Weibullova rozdělení.

Použité zdroje

[1] DAVISON, A. C., HINKLEY, D. V. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press, 2003. ISBN 0-521-57471-4.

[2] EFRON, B., TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. [New York]: Chapmann & Hall, 1993. 436 s. ISBN 0-412-04231-2.

[3] HIGGINS, J. J. *An Introduction to Modern Nonparametric Statistics*. [USA]: Brooks/Cole, 2004. 366 s. ISBN 0-534-38775-6.

[4] CHERNICK, M. R. *Bootstrap methods: a guide for practitioners and researchers*. New Jersey: John Wiley & Sons, 2008. ISBN 0-471-75621-0.

[5] KARPÍŠEK, Z. *Matematika IV: Statistika a pravděpodobnost*. 2. doplněné vydání. Brno: Akademické nakladatelství CERM, 2003. 170 s. Skriptum. ISBN 80-214-2522-9.

[6] KARPÍŠEK, Z., NERADOVÁ, V. *Estimation of Categorical Variable Probability Distribution (Odhad rozdělení pravděpodobnosti kategoriální veličiny)*. Bratislava: 7th International Conference APLIMAT 2008, 5. – 8. 2. 2008. Book of abstracts, s. 101. ISBN 978-80-89313-02-0. Sborník, s. 1145–1154. ISBN 978-80-89313-03-7.

[7] KARPÍŠEK, Z., NERADOVÁ, V., ŽAMPACHOVÁ, E. *A Contribution to the Estimation of Discrete Probability Distribution*. Brno: 14th International Conference on Soft Computing MENDEL 2008, 18. – 20. 6. 2008. Sborník, s. 287–292. ISBN 978-80-214-3675-6.

[8] NERADOVÁ, V. *Progresivní metody odhadů neznámých rozdělení pravděpodobnosti*. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2007. Vedoucí diplomové práce doc. RNDr. Zdeněk Karpíšek, CSc.

[9] *Volby do zastupitelstev krajů konané dne 17. – 18. 10. 2008. Výsledky hlasování za krajské zastupitelstvo. Jihomoravský kraj* [online]. Praha: Český statistický úřad, 2008. URL: <http://www.volby.cz/pls/kz2008/kz21?xjazyk=CZ&xdatum=20081017&xkraj=10>

[10] *Výzkum veřejného mínění. Krajské volby 2008. Jihomoravský kraj* [online]. Praha: SC&C spol. s r. o., 2008-10-8. 47 s. URL: <http://img1.ct24.cz/multimedia/documents/5/455/45447.doc>

[11] *Wikipedia* [online]. URL: <http://www.wikipedia.com>