Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra statistiky a teorie pravděpodobnosti

Studijní program: Statistika a ekonometrie
Obor: Kvantitativní metody v ekonomii

Použití metody bootstrap ve statistice

BAKALÁŘSKÁ PRÁCE

Student : Jiří Procházka
Vedoucí : Mgr. Milan Bašta, Ph.D.
Oponent :

2013

Prohlášení:

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně a že jsem uvedl všechny použité prameny a literaturu, ze které jsem čerpal.

V Praze dne 20. 05. 2013 . . . . . .

Jiří Procházka

Poděkování

Rád bych poděkoval mému vedoucímu práce, Mgr. Milanu Baštovi, Ph.D., za věnovaný čas a podnětné rady, jimiž mi do značné míry pomohl při vypracování této bakalářské práce.

Abstrakt

Cílem této bakalářské práce je představit základní využití metody bootstrap ve statistice. V následujících kapitolách bude vysvětleno fungování metody bootstrap, aplikace této metody pro odhadování směrodatné odchylky, vychýlení odhadu a pro tvorbu intervalů spolehlivosti. Taktéž v této práci bude prezentována možnost využití metody bootstrap pro testování statistických hypotéz. Následně budou tyto metody aplikovány na reálná či simulovaná data. Bude-li to možné, metoda bootstrap bude porovnána s analytickým přístupem. Práce bude členěna do pěti hlavních kapitol, vyjma úvodu. První z těchto pěti kapitol představí některé světové autory zabývající se bootstrapem a jejich knihy. Následující kapitola bude jakýmsi úvodem do problematiky. Třetí část se bude zabývat aplikací metody bootstrap pro hodnocení kvality bodových odhadů. V čtvrté části budou představeny aplikace některých metod bootstrapu pro tvorbu intervalů spolehlivosti. V poslední části této práce se budu zabývat testováním statistických hypotéz za pomoci metody bootstrap.

Klíčová slova

bootstrap, jackknife, statistická indukce, interval spolehlivosti, testování statistických hypotéz

Abstract

The main aim of this bachelor thesis is to introduce the basic utilization of bootstrap in statistics. In following chapters will be explained the functionality of bootstrap, applications of this method on estimating of standard errors of estimators, biases of estimators and creating of interval estimates. In this bachelor thesis will be also presented the possibility of utilization bootstrap for statistical hypothesis testing. Afterwards these methods will be used on real or simulated datasets. If possible bootstrap methods will be compared with the analytic approach. Excluding the preface the whole thesis is divided into five parts. The first part introduces several authors (as well as their books) who contributed largely to the theory and application of bootstrap. The next chapter serves as an introduction to the topic of bootstrap. The third part will deal with the application of bootstrap in evaluating the quality of estimators. In the fourth several bootstrap methods for constructing interval estimates will be introduced. The last chapter of this thesis is devoted to statistical hypothesis testing using the bootstrap approach.

Keywords

bootstrap, jackknife, statistical inference, interval estimates, statistical hypothesis testing

Obsah

1 Úvod 1

2 Bootstrap ve světě 2

2.1 Autoři 2

2.2 Knihy 3

3 Úvod do problematiky 4

3.1 Symbolika a terminologie 4

3.2 Základní myšlenka bootstrapu 5

3.3 Empirická distribuční funkce 7

3.4 Jackknife 9

3.5 Bootstrap 12

4 Aplikace metody bootstrap pro určení přesnosti odhadu 13

4.1 Bootstrapový odhad směrodatné odchylky odhadu 13

4.2 Bootstrapový odhad vychýlení 15

4.3 Počet bootstrapových replikací 15

4.3.1 Chyba bootstrapového odhadu vychýlení 16

4.3.2 Chyba bootstrapového odhadu směrodatné odchylky odhadu 17

4.3.3 Jackknife after bootstrap 18

4.4 Parametrický bootstrapový odhad 18

5 Aplikace metody bootstrap pro tvorbu intervalů spolehlivosti 21

5.1 Bootstrapový-t interval 21

5.2 Kvantilový bootstrapový interval 26

5.3 BC_a intervaly spolehlivosti 27

5.4 Implementace tvorby bootstrapových intervalů spolehlivosti 32

5.4.1 Parametrický bootstrapový odhad intervalů spolehlivosti 32

5.4.2 Snížení počtu nutných bootstrapových replikací 32

5.4.3 ABC intervaly spolehlivosti 32

6 Testování statistických hypotéz užitím metody bootstrap 33

6.1 Testování bootstrapovými intervaly spolehlivosti 33

6.2 Permutační test 34

6.3 Bootstrapové testování hypotéz 38

6.3.1 Dvouvýběrový bootstrapový test o střední hodnotě 39

6.3.2 Jednovýběrový bootstrapový test o rovnosti střední hodnoty 40

7 Závěr 42

1 Úvod

V podstatě každý den se člověk setkává se situací, kdy je nucen vyhodnocovat určitou entitu na základě dostupných dat. Jen málo kdy při onom vyhodnocování má dostupná veškerá relevantní data, a tedy se dá říci, že je prováděn jistý typ úsudku o celkové entitě na základě výběrových dat. Jako příklad si lze představit řidiče, který na základě zkušeností uvažuje, zdali bude v centru horší dopravní situace než na okruhu. Tyto situace je možné připodobnit k statistické indukci, kdy na základě analýzy výběrových dat je prováděn úsudek o populaci, respektive o základním souboru. Stejně jako v realitě, i v statistice je důležitá schopnost vyhodnocení přesnosti těchto úsudků. V každodenním životě je jen málo kdy využíváno sofistikovaných metod k určení přesnosti úsudků, naopak ve statistice existuje mnoho analytických nástrojů pro vyhodnocení přesnosti odhadů. Bohužel většina z nich funguje pouze za určitých podmínek a jen pro specifické typy úsudků. Metoda bootstrap vnáší jisté zobecnění do odhadování přesnosti úsudků za pomocí statistické indukce. Jedná se o metodu hojně využívající počítačových simulací. Význam této metody rostl spolu s růstem výkonnosti procesorů. Oproti sedmdesátým létům, kdy byla metoda bootstrap poprvé prezentována, se zvýšil počet tranzistorů zapojených v procesoru zhruba stotisíci násobně.

Tato bakalářská práce popisuje co je metoda bootstrap, jaké jsou její základní principy, k čemu lze tuto metodu využít. V práci je zkoumána možnost praktické aplikace metody při odhadování směrodatné odchylky odhadu, vychýlení odhadu, tvorbě intervalů spolehlivosti a testování statistických hypotéz. V případech, v kterých lze využít i analytický přístup bude provedeno porovnání metody bootstrap s analytickým přístupem. Veškeré grafy, výpočty a analýzy jsou tvořeny a prováděny v programu R.

2 Bootstrap ve světě

Bootstrap je relativně mladou metodou statistiky. I když byl objeven již v sedmdesátých letech, jeho plný rozvoj byl umožněn až spolu s rozvojem výpočetní techniky. Tato kapitola je věnována především odkazům a referencím. Čtenář zde nalezne hrubý výčet světových autorů, reference na knihy a články.

2.1 Autoři

Bradley Efron

Bradle Efron je profesorem na Standfordské universitě. Je označován za autora metody bootstrap. Z jeho hojné publikační činnosti bych uvedl monografii *An Introduction to the Bootstrap*. Tato kniha je jedním z hlavních zdrojů mé práce.

Jun Shao

Jun Shao, profesor Wisconsinské university, napsal spolu s profesorem Dongshen Tu knihu *The Jackknife and Bootstrap* vydanou nakladatelstvím Springers.

Peter Hall

Peter Gavin Hall je australským profesorem habilitujícím na Australian Nation University, nyní vyučující na University of Melbourne. Mezi jeho přínosy patří například práce na téma bootstrapových Monte Carlo aproximací a jejich zefektivnění. Z Hallových monografií bych rád uvedl knihy *On Efficient Bootstrap Simulation* a *Antithetic Resampling for the Bootstrap.*

John Hartigan

John Hartigen se mimo jiné věnoval výzkumu užití nejrůznějších resamplovacích metod ve statistice. Rád bych zde uvedl jeho knihu, *Using subsample values as typical values*, kterou napsal v roce 1969, a v které se zabývá zobecněním metody jackknife.

John Tukey

John Tukey je osoba, která položila jeden ze základních kamenů, na který navázali výše zmínění autoři. Mezi jeho četnými monografiemi bych rád vypíchnul knihu *Bias and confidence in not quite large samples*.

2.2 Knihy

An Introdutcion to the Bootstrap
Bradley Efron, Robert J. Tibshirani

Tato kniha je dělená do dvou celků. V první části knihy je uvedena nezbytná základní statistická teorie spolu s technikami metody bootstrap. Samotná teorie a odvození metody bootstrap je včetně některých odvození zapsána až v druhé části knihy. Tato kniha je velice přínosná i pro čtenáře, kteří nejsou studenty statistiky.

The Jackknife and Bootstrap
Jun Shao, Dongsheng Tu

Na rozdíl od výše zmíněné knihy, tuto monografii profesorů Tua a Shaa bych doporučil především studentům doktorského nebo alespoň navazujícího stupně statistiky. Jsou zde veškeré metody popsány velice sofistikovaně a k většině problémů, zde řešených, je napsána i alternativa, existuje-li, a příslušná reference na monografii, popřípadě článek, zabývající se do hloubky daným problémem.

Exploring the Limits of Bootstrap
Raoul lePage, Lynne Billard (editováno)

Tato kniha je sbírkou článků mnoha autorů. Čtenář se zde seznámí s hrubým nástinem různorodých směrů využití bootstrapu. Chce-li se čtenář dozvědět více o specifické problematice, nalezne v této knize příslušné reference.

3 Úvod do problematiky

3.1 Symbolika a terminologie

V této práci se budu snažit dodržovat určitou formu definovanou v této kapitole. Pojmem populace je rozuměn soubor všech statistických jednotek uvažovaných v dané statistické úloze. Pojem základní soubor bude v této práci užíván jako synonymum termínu populace. Statistická jednotka je nejnižší možná jednotka populace, u níž lze měřit určité statistické vlastnosti. Pod pojmem výběr bude myšlena určitá libovolně velká podmnožina populace. Pojmem náhodný výběr bude myšlen výběr z nekonečné populace, respektive výběr z konečné populace s vracením, kdy každá statistická jednotka má stejnou pravděpodobnost vybrání. Veličina, ať už náhodná či nenáhodná, bude značit určitou statistickou vlastnost různých jednotek populace. Náhodnou veličinu budu v následujícím textu značit vždy velkým písmenem. Náhodným vektorem bude myšlen vektor náhodných veličin. Náhodný vektor se bude značit malým písmenem. Například má-li sledovaná náhodná veličina X určité rozdělení, náhodný výběr o velikosti n může být před provedením výběru uvažován jako n náhodných veličin $X_1, X_2, \dots, X_n$ se stejným rozdělením jako má X. Poté náhodný vektor **x** bude značen malým písmenem a bude se zapisovat takto:

$$\boldsymbol{x} = (X_1, X_2, \dots, X_n) \qquad 3.1.1$$

Každá funkce náhodné veličiny je opět náhodná veličina. Například výběrový průměr náhodné veličiny X budu značit jako $\bar{X}$. Výše zmiňované n bude v této práci vždy značit počet prvků v konkrétní množině. Po provedení náhodného výběru budu značit konkrétní hodnoty malým písmenem, tudíž náhodný vektor **x** nabývá hodnot $x_1, x_2, \dots, x_n$ a výběrový průměr konkrétní realizace náhodného výběru budu značit $\bar{x}$. (Hebák, Bílková, & Svobodová, 2009)

Rozdělení pravděpodobnosti, zkráceně rozdělení, náhodné veličiny X bude dáno distribuční funkcí

$$F(x) = P(X \leq x); -\infty < x < \infty \qquad 3.1.2$$

Tedy pojmem neznámé rozdělení náhodné veličiny X a náhodná veličina X s neznámou distribuční funkcí bude myšleno to samé. Je-li náhodná veličina spojitá, bývá tato veličina

popisována hustotou pravděpodobnosti. Hustota pravděpodobnosti je definována jako první derivace distribuční funkce $F(x)$ podle x a značí se pro odlišení malým písmenem $f(x)$.

$$f(x) = dF(x) \qquad 3.1.3$$

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad 3.1.4$$

V tomto vzorci je substituce *t* za *x* pouze formální zápis. Zbylé značení bude definováno v průběhu textu vždy při prvním výskytu.

3.2 Základní myšlenka bootstrapu

V realitě je chování náhodných veličin značně odlišné. Zachytit a popsat tyto vlastnosti je v obecné rovině velice náročné, občas až nemožné. Proto se teorie pravděpodobnosti zabývá jen popisem nejčastějších pravděpodobnostních rozdělení. Častokrát je možné se setkat se situací, kdy získaný výběr z populace, o které nejsou dostupné žádné další informace, vykazuje prvky nenormality. Za předpokladů dostatečně velkého výběru je pochopitelně možné aplikovat centrální limitní větu, nicméně ani centrální limitní věta není všemocná a je ji možné aplikovat pouze v případě limitního rozdělení výběrového úhrnu, respektive výběrového průměru. Je-li výběr $X_1, \ldots, X_n$, nabývající hodnot $x_1, \ldots, x_n$, vybrán náhodně z populace, s neznámou distribuční funkcí F a neznámým parametrem θ, první prací statistika bývá vhodně odhadnout onen neznámý parametr, respektive funkci tohoto parametru. Toho je možno docílit vhodně zvoleným odhadem $t(\boldsymbol{x})$. Například je-li hledaným parametr střední hodnota μ a jako vhodný odhad je volen výběrový průměr $t(\boldsymbol{x}) = \bar{X}$, pak

$$\mu = \int x * dF(x) \qquad 3.2.1$$

$$\hat{\mu} = \bar{X} \qquad 3.2.2$$

Další otázkou je, jak přesné jsou úsudky o daném populačním parametru. Mírou variability bodového odhadu je střední čtvercová chyba, která udává průměrné čtvercové odchýlení střední hodnoty odhadu od skutečné hodnoty populačního parametru.

$$E[t(\boldsymbol{x}) - \theta]^2 = D[t(\boldsymbol{x})] + b_\theta^2[t(\boldsymbol{x})] \qquad 3.2.3$$

Ve výše uvedeném vzorci $D[t(\boldsymbol{x})]$ značí rozptyl odhadu a $b_\theta^2[t(\boldsymbol{x})]$ značí vychýlení odhadu na druhou. V této práci budu $D[t(\boldsymbol{x})]$ značit taktéž pro zkrácení zápisu jako $D(\hat{\theta})$. Rozptyl odhadu je absolutní mírou variability odhadu. Je definován jako průměrné čtvercové odchýlení odhadu od své střední hodnoty. V mé práci bude $D(X)$ vždy značit rozptyl náhodné veličiny X a $D(x)$ jako rozptyl konkrétní realizace náhodné veličiny. Vychýlení odhadu značí rozdíl mezi hodnotou odhadovaného parametru a střední hodnotou odhadu. V mé práci budu vždy značit vychýlení odhadu jako $b_\theta[t(\boldsymbol{x})]$, nebo zkráceně jen $b(\hat{\theta})$.

Je-li odhad nevychýlený, poté je střední čtvercová chyba rovna rozptylu odhadu, tedy absolutní mírou variability pro všechny nevychýlené odhady je rozptyl, respektive druhou odmocninu z něj, směrodatná odchylka odhadu, též někdy v literatuře označovaná jako směrodatná chyba odhadu.

$$se[t(\boldsymbol{x})] = \sqrt{D[t(\boldsymbol{x})]} \qquad 3.2.4$$

Označím-li σ^2 rozptyl neznámé populace s distribuční funkcí *F*, pak σ^2 je roven

$$\sigma^2 = (\int x - [\int x * dF(x)])^2 dF(x) \qquad 3.1.5$$

a mezi rozptylem σ^2 a směrodatnou odchylkou odhadu $\bar{X}$ platí jednoduchý vztah:

$$se(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} \qquad 3.2.6$$

Tento tvar je vcelku bezcenný, jelikož není známa distribuční funkce *F(x)*, tudíž i σ^2 je neznámý. Nicméně existuje nevychýlený odhad σ^2

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \qquad 3.2.7$$

Po dosazení nevychýleného odhadu do rovnice č. *3.2.6* je získán odhad směrodatné odchylky odhadu

$$se(\bar{X}) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n*(n-1)}} \qquad 3.2.8$$

Bohužel analogické vyjádření vztahu rovnice č. *3.2.6* není k dispozici pro složitější odhady. Po generace se statistici snažili najít nějakou obecnější formuli, avšak tyto snahy ztroskotaly na odhadech složitějších, jako je například výběrový medián, výběrový korelační koeficient a jím podobné. (kol. autorů, 1992)

3.3 Empirická distribuční funkce

Stále nevyřešenou otázkou je, jak obecně co nejlépe odhadnout neznámý parametr a určit, jak přesný náš odhad je. Jedná-li se o parametrické odhady, kdy náhodná veličina X má známou distribuční funkci $F(x)$, s neznámým parametrem, či neznámým vektorem parametrů, existuje mnoho postupů pro konstrukci vhodného odhadu. Nejznámější skupinu odhadů tvoří tak zvané M-estimátory, které lze považovat v obecné rovině pouze jako zobecnění maximálně věrohodných odhadů. Je-li $F(x)$ regulární rozdělení, je možné za pomocí metody maximální věrohodnosti zkonstruovat nevychýlený, konzistentní a vydatný odhad neznámého parametru daného rozdělení. Vhodnou aproximací variability tohoto odhadu je pak Fischerova míra informace. Avšak odhady parametrů získaných metodou maximální věrohodnosti mají sklon k tomu být koncentrovanější okolo střední hodnoty. Nechť $X_1, X_2, \ldots, X_n$ jsou náhodně vybrané nezávislé náhodné veličiny pocházející ze stejného rozdělení s distribuční funkcí $F(x)$, pak neparametrickým maximálně věrohodným odhadem $F(x)$ je empirická distribuční funkce $\hat{F}(x)$. Empirická distribuční funkce je diskrétní distribuční funkcí přiřazující pravděpodobnost vybrání i-tého pozorování jako $\frac{1}{n}$

$$P_{\hat{F}}(\mathrm{X} = \mathrm{x}) = \frac{1}{n} \qquad 3.2.1$$

$$\hat{F}(x) = P_{\hat{F}}(\mathrm{X} \leq \mathrm{x}) = \frac{\#(X_i \leq \mathrm{x})}{n} \qquad 3.2.2$$

V této práci používám dolní index s dosazenou distribuční funkcí pouze v rámci porovnávání dvou entit lišících se pouze dosazenou distribuční funkcí a v případě kdy není jasné, jaká distribuční funkce byla pro daný případ dosazena. Je-li v dolním indexu $\hat{F}$, znamená to, že pro tento případ byla dosazena za neznámou distribuční funkci empirická distribuční funkce. Je-li v dolním indexu pouze F, znamená to, že pro daný případ byla ponechána neznámá distribuční funkce. Funkce $\#(X_i \leq \mathrm{x})$ vyjadřuje počet náhodných veličin výběru menších nebo rovných x. V dalším textu budu tuto funkci často používat. Argument této funkce je určitá podmínka. Funkce kumuluje jedničku pokaždé, je-li tato podmínka splněna. To znamená, že je-li podmínka splněná třikrát, funkce vrátí hodnotu tři.

Empirická distribuční funkce poskytuje nevychýlený odhad neznámé distribuční funkce,

$$E\left[\hat{F}(x)\right] = F(x) \qquad 3.2.3$$

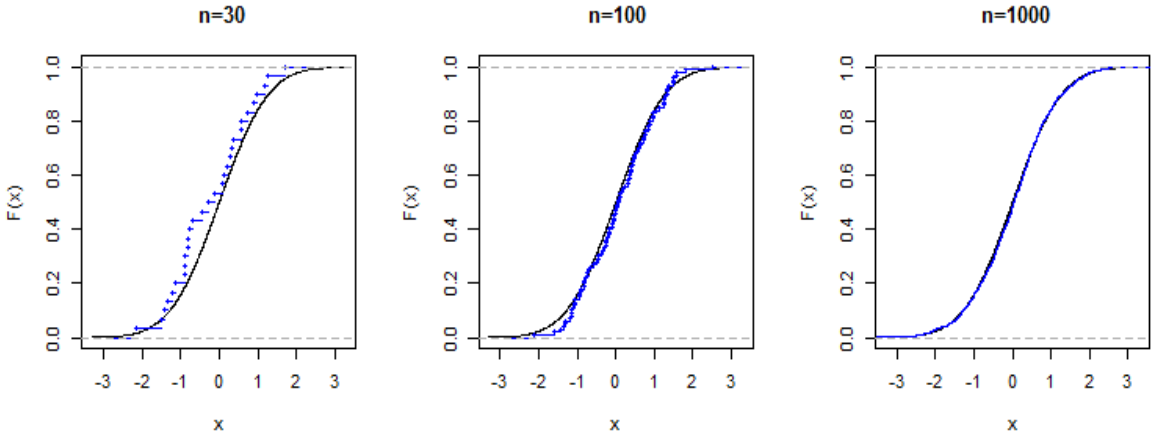
Nicméně jak již bylo řečeno rozdělení odhadnuté metodou maximální věrohodnosti má vyšší koncentraci okolo střední hodnoty, tudíž systematicky podhodnocuje rozptyl. Pro výpočet nevychýleného odhadu rozptylu je nutno výpočet doplnit následujícím vztahem

$$E[D_{\hat{F}}(X)] = \frac{n-1}{n} D_F(X) \qquad 3.2.4$$

Proto, jak je nám známo ze základních kurzů statistiky, je doporučeno raději používat jako odhad populačního rozptylu upravený výběrový rozptyl, než samotný neupravený odhad za pomocí empirické distribuční funkce.

$$\frac{n}{n-1} D_{\hat{F}}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = s^2 \qquad 3.2.5$$

Bude-li v následujícím textu zmíněn pojem neparametrický odhad, bude tím myšlen odhad neznámé distribuční funkce za pomocí neparametrického maximálně věrohodného odhadu, s využitím empirické distribuční funkce. (kol. autorů, 1992)



Graf číslo 1 – Konvergence empirické distribuční funkce
zdroj: vlastní zpracování na základě simulovaných dat

Graf číslo 1 znázorňuje konvergenci empirické distribuční funkce k distribuční funkci normovaného normálního rozdělení. V příloze *A* je vložen kód v jazyce R pro tvorbu tohoto grafu.

3.4 Jackknife

Roku 1958 představil John Wilder Tukey techniku pro odhadování chyb odhadů zvanou „Jackknife“, navazující na Quenouillevu práci v oblasti odhadu vychýlení odhadů. John Tukey získal doktorát z matematiky na Princetonské universitě. Mimo jeho nesmírný přínos na poli aplikované matematiky, je možné se s jeho jménem taktéž setkat v historii výpočetní techniky. Údajně je autorem definice pojmu „bit“ a taktéž použil ve své publikaci jako první slovo „software“. Ve statistice můžeme narazit na jeho jméno například v souvislosti s ANOVOU, kde Tukeyova metoda mnohonásobného porovnávání určuje jaká dvojice $\mu_i \neq \mu_j$ způsobila zamítnutí nulové hypotézy o shodě středních hodnot.

Princip metody jackknife je následující. Za předpokladů, že výběr dat **x** byl vybrán nezávisle, ze stejného neznámého rozdělení, což mohu zaznačit následovně.[1]

$$F \xrightarrow{i.i.d.} (x_1, x_2, \dots, x_n) = \mathbf{x} \qquad 3.3.1$$

$$\hat{\theta} = t(\boldsymbol{x}) \qquad 3.3.2$$

Je vytvořen nový jackknifový výběr $\boldsymbol{x}_{(i)}$ tak, že z původního vzorku je vždy vymazáno i-té pozorování. Je jasné, že tímto krokem lze dosáhnout *n* nových vzorků o velikosti *n-1*.

$$\boldsymbol{x}_{(i)} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \qquad 3.3.2$$

Nechť $\hat{\theta}_{(i)}=t(\boldsymbol{x}_{(i)})$ je replikací $\hat{\theta}$ s vyřazeným i-tým pozorováním. Pak jackknifový odhad směrodatné odchylky odhadu (*3.3.3*) a vychýlení odhadu (*3.3.4*) je:

$$\widehat{se}_{jack}(\hat{\theta}) = \sqrt{\frac{(n-1)}{n} * \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \bar{\theta}_{(.)})^2} \qquad 3.3.3$$

$$\hat{b}_{jack}(\hat{\theta}) = \frac{(\bar{\theta}_{(.)} - \hat{\theta})}{n-1} \qquad 3.3.4$$

$$\bar{\theta}_{(.)} = \frac{\sum_{i=1}^{n} \hat{\theta}_i}{n} \qquad 3.3.5$$

[1] V následujícím textu bude vždy pojmem náhodný výběr myšleno, že byl výběr proveden nezávisle ze stejného rozdělení, nebude-li výslovně řečeno něco jiného.

$\hat{b}_{jack}(\hat{\theta})$ je v tomto případě jackknifový odhad $b_{t(x)}(\theta)$ a $\widehat{se}_{jack}(\hat{\theta})$ je jackknifovým odhadem $se(\hat{\theta})$

Hlavní deviza této metody spočívá, že lze aplikovat i na mnohem složitější odhady než je výběrový průměr. Nevýhodou je, že tato metoda využívá pouze n nových výběrů, čili takto získaná informace o výběru je omezena. Bohužel v době, kdy byla tato metoda představena, byla výpočetní technologie v pomyslných plenkách, výpočty se převážně prováděly ručně, a proto je toto omezení vcelku obhajitelné. Lze dokázat, že jackknife je velice kvalitní aproximací bootsrapu, nic méně jen tehdy, je-li odhad hladký, což v tomto případě znamená, že malá změna ve výběru způsobí pouze malou změnu ve vypočítaném odhadu. Klasickým příkladem nehladkého odhadu je například výběrový medián, výběrová maximální hodnota nebo výběrový průměr čísel s lichými indexy. (Effron & Timbishirani, 1993) Například je-li vybrán výběr o devíti pozorováních

$\boldsymbol{x} = (\,11, 12, 13, 14, 15, 16, 17, 18, 19)$,

pak jsou výše zmíněné výběrové statistiky rovny:

$\tilde{x} = 15$

$max(x) = 19$

$liché_i(\bar{X}) = 15$

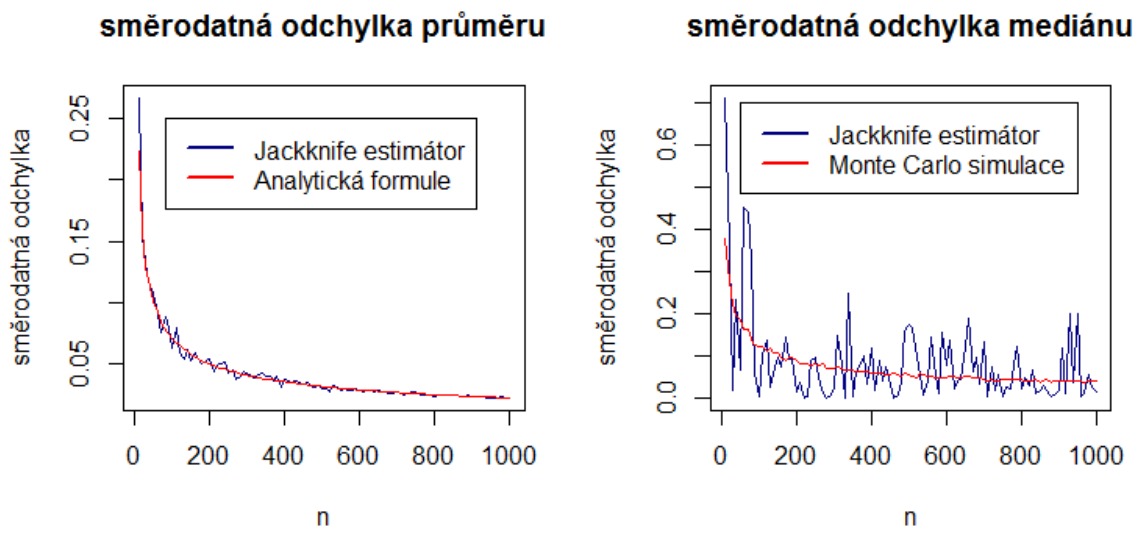
Změní-li se hodnota druhého pozorování o dvě, výběrový medián, výběrové maximum ani výběrový průměr čísel s lichými indexy se nezmění. Bereme-li v úvahu medián.

$\hat{\theta} = \tilde{X}$

$\hat{\theta}_{(i)} = (15.5, 15.5, 15.5, 15.5, 15, 14.5, 14.5, 14.5, 14.5)$

$\widehat{se}_{jack}(\tilde{x}) = 1,\overline{3}$

Je zřejmé, že jackknifová replikace mediánu nabývá maximálně tří hodnot. Lze dokázat, že je pro tento případ porušena konzistence.



Graf číslo 2 – Jackknifový odhad směrodatné odchylky odhadu
Zdroj: vlastní zpracování ze simulovaných dat

Graf zobrazuje průběh jackknife odhadu v závislosti na *n*. Byl proveden náhodný výběr o velikosti *n*. První graf zachycuje odhad směrodatné odchylky výběrového průměru. Modrá křivka znázorňuje hodnoty vypočtené metodou jackknife. Červená křivka znázorňuje průběh analytické funkce směrodatné odchylky výběrového průměru dle vzorce č. *3.2.6*. Druhý graf zachycuje odhad směrodatné odchylky výběrového mediánu. Bohužel tato výběrová statistika nemá triviální vyjádření analytické funkce rozptylu, proto byla metoda jackknife komparována se simulací Monte Carlo. V Monte Carlo simulaci bylo realizováno deset tisíc náhodných výběrů o velikosti *n* ze známého rozdělení. Následně byl pro každý náhodný výběr vypočten výběrový medián. V posledním kroku byla vypočtena výběrová směrodatná odchylka těchto „Monte Carlo mediánů“. Z grafu je patrná nekonzistentnost metody pro nehladké odhady.

Dalším a asi hlavním selháním metody jackknife je tak zvaná studentizace při tvorbě intervalů spolehlivosti. Nechť je odhad parametru θ aproximovatelný k normálnímu rozdělení.

$$\frac{(\hat{\theta}-\theta)}{\mathrm{se}(\hat{\theta})} \sim \mathrm{u} \qquad 3.3.6$$

Zde u značí normované normální rozdělení. Je-li nahrazena neznámá statistika $\mathrm{se}(\hat{\theta})$ jackknifovým odhadem, je možné zapsat,

$$\hat{\theta} \pm u^{\alpha/2} * \widehat{se}_{jack}(\hat{\theta}) \qquad 3.3.7$$

V této rovnici $u^{\alpha/2}$ značí $\frac{\alpha}{2}$ kvantil normovaného normálního rozdělení. Je známo, že pro neznámý rozptyl je normální rozdělení pouze limitní aproximací studentova t-rozdělení a tudíž pro menší vzorky lze tento interval spolehlivosti vylepšit aproximací t-rozdělením o *n-1* stupni volnosti.

$$\hat{\theta} \pm t_{n-1}^{(1-\alpha/2)} * \widehat{se}_{jack}(\hat{\theta}) \qquad 3.3.8$$

V této rovnici $t_{n-1}^{(1-\alpha/2)}$ značí $\frac{\alpha}{2}$ kvantil t-studentova rozdělení s *n-1* stupni volnosti. Bohužel je dokázáno, že jackknifový interval spolehlivosti s aproximací parametru k t-rozdělení neposkytuje statisticky významně lepší intervalový odhad θ než s aproximací k studentovu t-rozdělení. (kol. autorů, 1992)

3.5 Bootstrap

Poprvé se myšlenka bootstrapu objevila již roku 1969. Autorem této myšlenky byl australský statistik John Hartigan. Hartigan tvrdil, že každý náhodný výběr o rozsahu *n* má $2^n - 1$ neprázdných podmnožin. Jackknife využívá pouze *n* z nich. Možností jak jackknife vylepšit, by bylo použít metodu založenou na více než *n* podmnožinách původního výběru. Bohužel v této době byly počítače stále ještě velice pomalé a nevhodné pro rozsáhlé simulace. Proto Hartigan zůstal pouze u teoretických úvah. O deset let později, roku 1979, byla Bradem Efronem, profesorem Stanfordské university, představena metoda nazvaná „bootstrap“. Myšlenka bootstrapu vznikla při zkoumání vlastností odhadů směrodatné chyby odhadu metodou jackknife. Principem metody bootstrap je, že z náhodného výběru o rozsahu *n* je vybráno aplikací určitého pravděpodobnostního modelu *n* pozorování s vracením. Každou tuto realizaci výběru s vracení z původního výběru nazvěme bootstrapový výběr. Z toho vyplývá, že je možno vytvořit více než *n* nových výběru, a tedy je potenciál celé metody mnohem vyšší než u jackknifu. Metodou bootstrap lze získat přesně $\binom{2n-1}{n}$ různých výběrů. K tomuto kombinačnímu číslu se dojde aplikací kombinace s opakováním *n* prvků na *k-tici* rovnou *n*, z čehož plyne, že jednotlivé bootstrapové výběry se od sebe liší hodnotami, nikoli pořadím (Shao & Tu, 1995).

4 Aplikace metody bootstrap pro určení přesnosti odhadu

Tuto kapitolu mé práce věnuji aplikacím metody bootstrap pro hodnocení kvality odhadu. Jak již bylo řečeno pro určení kvality odhadu je klíčová střední čtvercová chyba. Za pomocí této statistiky lze porovnávat kvalitu jednotlivých odhadů. Jak již bylo napsáno v kapitole třetí, střední čtvercová chyba je rovna součtu rozptylu a druhé mocnině vychýlení odhadu.

4.1 Bootstrapový odhad směrodatné odchylky odhadu

V předchozí kapitole byl popsán vztah mezi střední čtvercovou chybou, vychýlením a rozptylem odhadu, z něhož plyne že je-li odhad nevychýlený, jeho jedinou relevantní měrou variability je rozptyl, potažmo směrodatná odchylka odhadu. Bootstrapový odhad směrodatné odchylky odhadu probíhá ve dvou krocích:

1. Nechť **x** je náhodně vybraný výběr z populace a neparametrický odhad θ, $\hat{\theta} = t(\mathbf{x})$. Je realizováno B bootstrapových náhodných výběrů z původního výběru $\mathbf{x}=(x_1, x_2, \ldots, x_n)$ o velikosti n. B je obvykle voleno okolo dvou set.

$$\hat{F} \rightarrow (x_1^*, x_2^*, \ldots, x_n^*) = \boldsymbol{x}_b^* \quad b \in \{1, 2, \ldots, B\} \tag{4.1.1}$$

$$\overline{\overline{\boldsymbol{x}}}^* = (\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_B^*)^T \tag{4.1.2}$$

2. Pro všechny bootstrapové výběry je vypočten bootstrapový odhad parametru θ. Bootstrapový odhad parametru θ budu značit jako $\hat{\theta}^*(b)$, kde b značí číslo bootstrapového výběru, pro který byl tento odhad dopočítán. Vektor s bootstrapovými odhady budu značit jako $\hat{\boldsymbol{\theta}}^*$.

$$\hat{\theta}^*(b) = T(\boldsymbol{x}_b^*) \quad b \in \{1, 2, \ldots, B\} \tag{4.1.3}$$

$$\hat{\boldsymbol{\theta}}^* = \{\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)\} \tag{4.1.4}$$

3. V posledním kroku je vypočten bootstrapový odhad směrodatné odchylky odhadu parametru θ, $\widehat{se}_B(\hat{\theta})$.

$$\widehat{se}_B(\hat{\theta}) = \sqrt{\frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2}{B-1}} \tag{4.1.5}$$

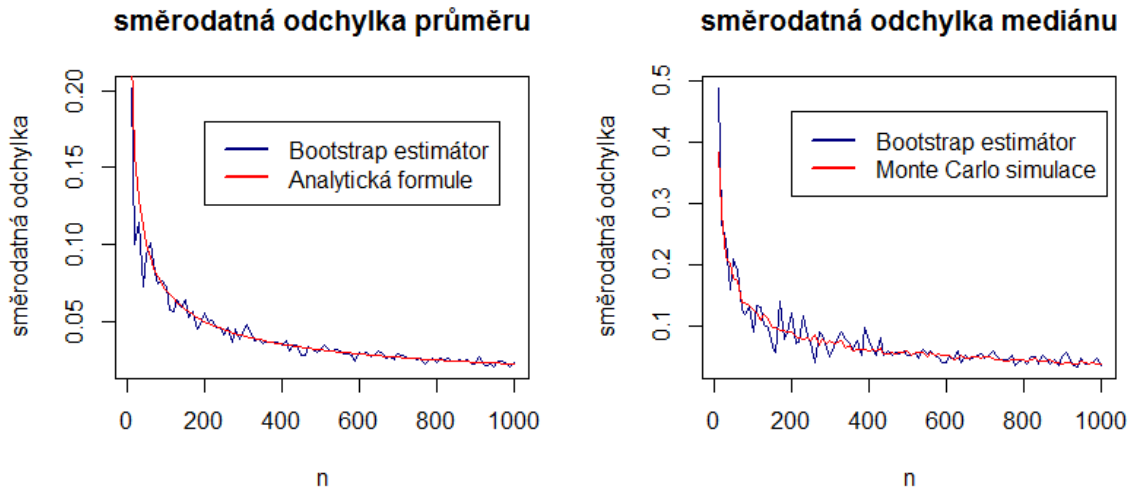
$$\hat{\theta}^*(.) = \frac{\sum_{b=1}^{B}\hat{\theta}^*(b)}{B} \tag{4.1.6}$$

Limitní tvar vzorce č. *4.1.5*, kde *B* je rovno nekonečnu, se taktéž nazývá ideální bootstrapový odhad směrodatné odchylky. Ideální bootstrap je roven přesné hodnotě směrodatné odchylky odhadu za situace, kdy je neznámá distribuční funkce nahrazena empirickou distribuční funkcí. Ideální bootstrap konverguje ke skutečné hodnotě směrodatné odchylky odhadu, s neznámou distribuční funkci, s *n* rostoucím do nekonečna.

$$\lim_{B\to\infty} \widehat{se}_B(\hat{\theta}) = \widehat{se}_\infty(\hat{\theta}) = \widehat{se}_{\hat{F}}(\hat{\theta}) \qquad 4.1.7$$

$$\lim_{n\to\infty} \widehat{se}_{\hat{F}}(\hat{\theta}) = \widehat{se}_F(\hat{\theta}) \qquad 4.1.8$$

Této metodě bootstrapového odhadu se taktéž říká neparametrický bootstrap, protože je zde v procesu odhadu substituována neznámá distribuční funkce empirickou distribuční funkcí, viz kapitolu č. *3.3*. (Effron & Timbishirani, 1993)



Graf číslo 3 – Bootstrapový odhad směrodatné odchylky odhadu
Zdroj: vlastní zpracování simulovaných dat

Na grafu je znázorněn průběh bootstrapového odhadu v závislosti na *n*. Byla zde použita stejná metoda simulace jako v kapitole č. *3.4*. Na rozdíl od odhadu metodou jackknife, je vidět že bootstrap funguje skvěle i pro nehladké odhady.

4.2 Bootstrapový odhad vychýlení

Další složkou střední čtvercové chyby je vychýlení. Vychýlení je definováno jako rozdíl mezi střední hodnotou odhadu a skutečnou hodnotou parametru. Kromě k posouzení kvality je vychýlení používáno ke korekcím odhadů na nevychýlené. Asi nejznámější příklad korekce je výběrový rozptyl, kde je rozptyl opraven na nevychýlený za pomocí násobku členem $\frac{n-1}{n}$. Bootstrapový odhad vychýlení odhadu opět probíhá ve třech krocích: (Effron & Timbishirani, 1993)

1. Z náhodného výběru **x** je vypočten neparametrický odhad θ, $\hat{\theta} = t(\mathbf{x})$ a je provedeno *B* bootstrapových výběrů s vracením. Zde je ovšem *B* voleno mnohem vyšší, než u odhadu směrodatné odchylky. Tomuto tématu bude věnováno pár řádků v kapitole *4.3*.
2. Pro každý Bootstrapový výběr je vypočítán bootstrapový odhad dle vzorce č. *4.1.3*
3. Je vypočítán bootstrapový odhad vychýlení odhadu $\hat{\theta}$.

$$\hat{b}_B(\hat{\theta}) = \hat{\theta}^*(.) - \hat{\theta} \qquad 4.2.1$$

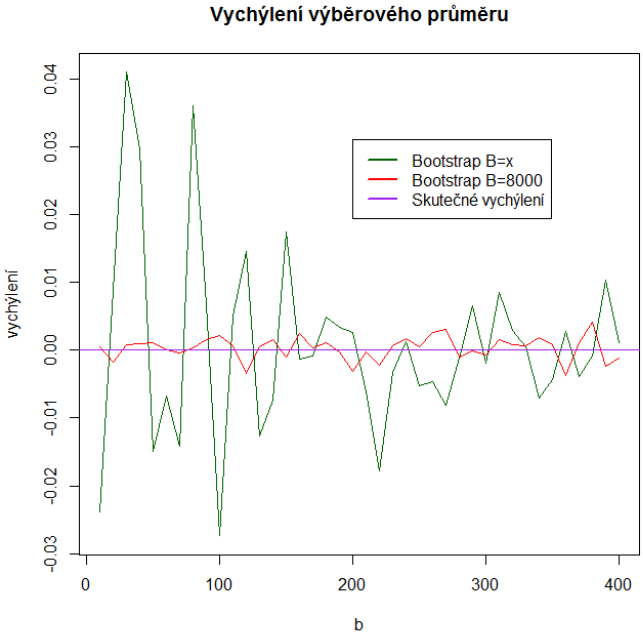
$$\hat{\theta}^*(.) = \frac{\sum_{i=1}^{B} \hat{\theta}^*(b)}{B} \qquad 4.2.2$$

4.3 Počet bootstrapových replikací

V předchozích dvou kapitolách byl počet bootstrapových odhadů pouze hrubě a orientačně stanoven. Ideální bootstrapový odhad je takový bootstrapový odhad, pro něhož je počet replikací roven nekonečnu. I bootstrap sám o sobě má určitou variabilitu. Tato variabilita vzniká díky dvěma hlavním složkám. První složkou je výběrová variabilita vzniklá díky tomu, že rozsah výběru je příliš malý pro kvalitní aproximaci neznámé populační distribuční funkce empirickou distribuční funkcí, viz kapitolu č. *3.3*. Druhá složka variability bootstrapu je způsobena odlišností ideálního bootstrapu od bootstrapového odhadu pro konečné *B*. Aproximaci bootstrapového odhadu pro konečné *B* k ideálnímu bootstrapovému odhadu budu taktéž nazývat Monte Carlo aproximaci bootstrapového odhadu. (Effron & Timbishirani, 1993)

4.3.1 Chyba bootstrapového odhadu vychýlení

Konvergence odhadu bootstrapového vychýlení k ideálnímu bootstrapovému odhadu vychýlení je mnohem pomalejší, než je tomu u odhadu směrodatné odchylky. Proto v kapitole č. *4.2* je *B* voleno mnohem vyšší, než je tomu u bootstrapového odhadu směrodatné odchylky.



Graf číslo 4 – Bootstrapový odhad vychýlení výběrového průměru pro různá *B*
Zdroj vlastní zpracování simulovaných dat

Na grafu je znázorněn bootstrapový odhad vychýlení výběrového průměru. Zelená křivka znázorňuje průběh odhadu závislý na *B* (počet bootstrapových replikací), červená křivka znázorňuje průběh odhadu pro *B* fixované na hodnotě osm tisíc a fialová přímka zobrazuje hodnotu skutečného vychýlení výběrového průměru. Z grafu je patrné, že i když je skutečné vychýlení výběrového průměru rovno nule, odhad vychýlení metodou bootstrap, pro konečná *B* kolem této hodnoty osciluje. Hodnota této fluktuace lze díky centrální limitní větě aproximativně zachytit následujícím vztahem: (Effron & Timbishirani, 1993)

$$\hat{\theta}^*(.) \sim N\left(E(\hat{\theta}^*), \frac{\hat{D}_B(\hat{\theta})}{B}\right) \qquad 4.3.1$$

$$P\left(\left|\hat{b}_B(\hat{\theta}) - \hat{b}_\infty(\hat{\theta})\right| < 2\frac{\widehat{se}_B(\hat{\theta})}{\sqrt{B}}\right) = 0{,}95 \qquad 4.3.2$$

Kde $\widehat{D}_B(\hat{\theta})$ je bootstrapový odhad rozptylu $\hat{\theta}$ a 0,95 je zaokrouhlená pravděpodobnost, že absolutní hodnota normálně rozdělené náhodné veličiny překročí dvojnásobek směrodatné odchylky.

4.3.2 Chyba bootstrapového odhadu směrodatné odchylky odhadu

Rozkladem rozptylu bootstrapového odhadu směrodatné odchylky odhadu lze dojít k následujícímu obecnému tvaru rozptylu:

$$D[\widehat{se}_B(\hat{\theta})] = \frac{c_1}{n^2} + \frac{c_2}{n*B} \qquad 4.3.3$$

Ve vzorci č. *4.3.3* jsou obě konstanty c_1 i c_2 nezávislé na velikosti *n* či *B*. Velikost obou konstant určuje tvar neznámého populačního rozdělení. První podíl vyjadřuje variabilitu vzniklou díky nedokonalé aproximaci empirické distribuční funkce. Druhý podíl vyjadřuje pro fixované *n* variabilitu vzniklou díky Monte Carlo aproximaci bootstrapového odhadu pro konečné *B* k ideálnímu bootstrapovému odhadu.

Relativní mírou variability, sloužící k určení počtu bootstrapových simulaci je variační koeficient. Variační koeficient je definován jako podíl směrodatné odchylky a odmocniny ze střední hodnoty na druhou.

$$V(\widehat{se}_B) = \frac{\sqrt{D[\widehat{se}_B(\hat{\theta})]}}{E[\widehat{se}_B(\hat{\theta})]} \qquad 4.3.4$$

$$V(\widehat{se}_B) = \frac{\sqrt{D[\widehat{se}_B(\hat{\theta})]}}{E[\widehat{se}_B(\hat{\theta})]} * 100\% \qquad 4.3.5$$

Je očividné, že je-li dělitel střední hodnotou směrodatné odchylky, nemůže být nikdy menší než nula a proto je zbytečné vyjadřovat jej v absolutní hodnotě. Vzorec *4.3.5* je pouze procentuální vyjádření variačního koeficientu a udává, z kolika procent se podílí variabilita bootstrapového odhadu na bootstrapovém odhadu směrodatné odchylky odhadu. Matematickými operacemi lze vyjádřit variační koeficient bootstrapového odhadu ve tvaru

$$V(\widehat{se}_B) = \sqrt{V(\widehat{se}_\infty) + \frac{E(\Delta)+2}{4*B}} \qquad 4.3.6$$

První sčítanec reprezentuje variační koeficient ideálního bootstrapového odhadu. Druhý sčítanec ve tvaru podílu reprezentuje odlišnost mezi ideálním bootstrapovým odhadem a

bootstrapovým odhadem pro konečné *B*. Střední hodnota delty v dělenci nabývá hodnotu dle špičatosti populačního rozdělení. Pro rozdělení s nejvyšší možnou koncentrací hodnot okolo střední hodnoty nabývá delta hodnoty -2. Čím je rozdělení plošší, tím hodnota delty roste. Pro normální rozdělení nabývá delta hodnoty nula. Z této skutečnosti vyplývá, že je třeba pro plošší rozdělení volit konstantu *B* vyšší, aby se dosáhlo stejné přesnosti jako u špičatého rozdělení. (Effron & Timbishirani, 1993)

4.3.3 Jackknife after bootstrap

Pro určení rozptylu bootstrapového odhadu směrodatné odchylky odhadu se používá metoda zvané Jackknife after bootstrap. Metodu jackknife jsem již stručně představil v kapitole tři. Výpočet rozptylu bootstrapového odhadu směrodatné odchylky odhadu je prováděn v následujících krocích: (Effron & Timbishirani, 1993)

1. Nechť $\mathbf{x}=(x_1, x_2, \ldots, x_n)$ je náhodně vybraný výběr z populace. Je vybráno *B* bootstrapových výběrů. Pro každý tento výběr je spočten bootstrapový odhad parametru θ, viz vzorec č. *4.1.3*.
2. Pro i = 1, 2, … n vynecháme takový bootstrapový výběr, ve kterém se vyskytuje i-té pozorování vektoru **x**.
3. Pro všechny *i* je bootstrapový odhad směrodatné odchylky odhadu dle vzorce č. *4.1.3* a č. *4.1.5*. Nesmí být opomenuto snížit *B* ve vzorci č. *4.1.5* o počet vyškrtnutých bootstrapových výběrů. Takto vypočtený bootstrapový odhad směrodatné odchylky odhadu označím jako $\widehat{se}_{B(i)}$.
4. Vypočte se odhad rozptylu bootstrapového odhadu směrodatné odchylky.

$$\widehat{se}_{jack}[\widehat{se}_B(\hat{\theta})] = \sqrt{\frac{(n-1)}{n} * \sum_{i=1}^{n}(\widehat{se}_{B(i)} - \widehat{se}_{B(.)})^2} \qquad 4.3.5$$

$$\widehat{se}_{B(.)} = \frac{\sum_{i=1}^{n}\widehat{se}_{B(i)}}{n} \qquad 4.3.6$$

4.4 Parametrický bootstrapový odhad

Je-li známa populační distribuční funkce *F*, může se jevit nesmyslné používat bootstrapový odhad například pro odhad směrodatné odchylky odhadu, zvláště existuje-li, například pro odhady získané metodou maximální věrohodnosti, kvalitní aproximace směrodatné od-

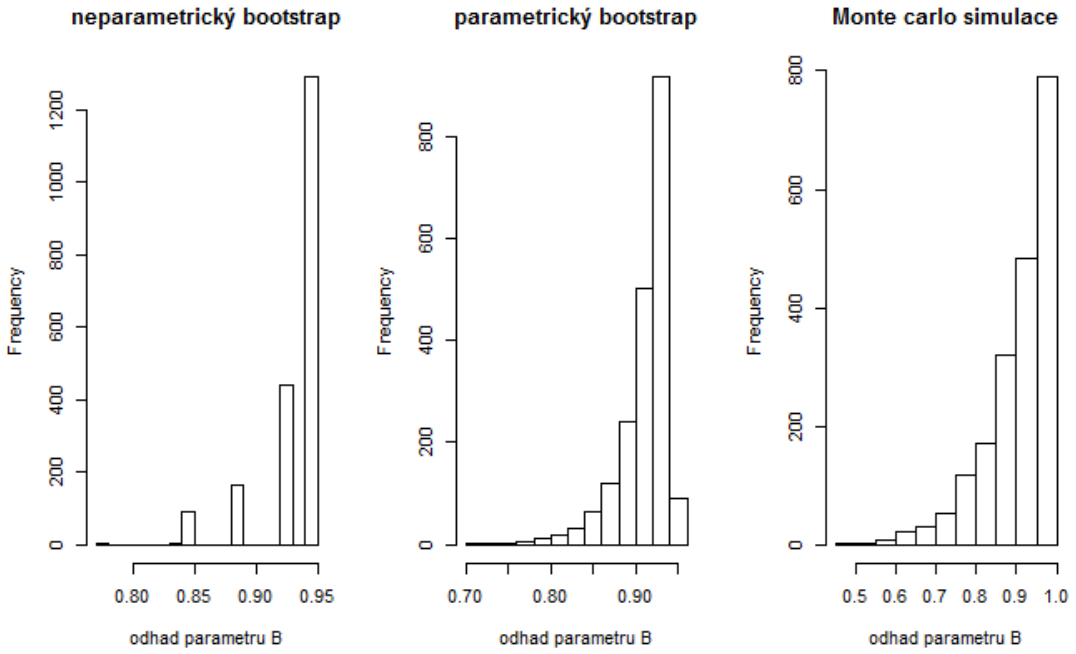
chylky odhadu díky Fischerově míře informace. Nicméně parametrický bootstrap poskytuje exaktní hodnotu variability daného odhadu. Parametrický bootstrap využívá pro výběr vypočtený parametrický pravděpodobnostní model. Algoritmus realizace parametrického bootstrapu je velice obdobný tomu neparametrickému. Liší se jen volbě onoho parametrického modelu bootstrapového výběru, který je v tomto případě roven známé populační distribuční funkci. Je-li θ jediným neznámým parametrem distribuční funkce F a $\mathbf{x}=(x_1, x_2, \ldots, x_n)$, je realizací náhodného výběru z tohoto rozdělení, vypočte se odhad parametru θ. Na místo neznámého parametru distribuční funkce F, je do pravděpodobnostního modelu bootstrapového výběru dosazen jeho odhad. Je realizováno B bootstrapových výběrů s aplikací stanoveného pravděpodobnostního modelu (*4.4.1*).

$$\hat{F}_{par} \xrightarrow{i.i.d.} (x_1^*, x_2^*, \ldots, x_n^*) = \boldsymbol{x}^* \qquad 4.4.1$$

A dále je postup stejný jako v neparametrickém bootstrapu. Tedy například, je-li realizován náhodný výběr z normálního rozdělení s neznámou střední hodnotou, ale známým parametrem σ=1, $\mathbf{x}=(x_1, x_2, \ldots, x_n)$, je vypočten výběrový průměr jako maximálně věrohodný odhad parametru μ. Následně je bootstrapový výběr realizován náhodným výběrem z normálního rozdělení, s parametry $\mu = \bar{x}$ a σ=1. Takto lze postupovat i v případě, že není znám ani jeden parametr. V takové situaci jsou všechny neznámé parametry odhadnuty a bootstrapový výběr je realizován ze známého rozdělení s odhadnutými parametry.

V případech, kdy neparametrický bootstrap selhává, parametrický bootstrap často funguje bez problémů. Klasickým příkladem této situace je odhad parametru b rovnoměrného rozdělení. V takovém případě je pravděpodobnost, že bude vybrána maximální hodnota vektoru $\mathbf{x}$ v libovolném bootstrapovém výběru, a tedy že odhad parametru θ bude roven bootstrapovému odhadu parametru θ (Effron & Timbishirani, 1993)

$$P[\hat{\theta}^*(b) = \hat{\theta}] = 1 - \frac{n-1}{n}^n \qquad 4.4.2$$



Graf číslo 5 – Různé metody odhadu parametru *b* rovnoměrného rozdělení
Zdroj: vlastní zpracování simulovaných dat

V této studii bylo generováno čtyřicet hodnot z rovnoměrného rozdělení. Z prvního histogramu je patrné, že podíl počtu situací, kdy bootstrapový odhad θ je roven analytickému odhadu θ, ku dvěma tisícům, což je celkový počet generovaných hodnot, je zhruba roven pravděpodobnosti vypočtené dle vzorce č. *4.4.2*. Dále je možné zaregistrovat, že první histogram se jeví jako velice chabá aproximace třetího histogramu, který je generován ze skutečného, populačního rozdělení. Druhý histogram zachycuje rozdělení četností odhadu parametru rovnoměrného rozdělení užitím parametrického bootstrapu. Je očividné, že v tomto případě dokáže parametrický bootstrap kvalitně zastoupit neparametrický bootstrap, který jak vidno v této dané situaci selhává.

Toto ovšem neznamená, že je-li odhadem neznámého parametru s neznámou populační distribuční funkcí, maximální hodnota vzorku, nemůže být neparametrický bootstrap použit. V takové situaci je ovšem třeba upravit algoritmus neparametrického bootstrapu. Upravení spočívá v tom, že v každém bootstrapovém výběru je náhodně taženo s vracením *n-k*, namísto *n*, pozorování z vektoru **x**. (Politis & Romano, 1993)

5 Aplikace metody bootstrap pro tvorbu intervalů spolehlivosti

V předchozích sekcích byly probrány pouze bodové odhady. Bodové odhady jsou základem matematické statistiky. Nicméně pravděpodobnost, že bodový odhad určí přesnou hodnotu parametru, se obvykle blíží nule. Z tohoto důvodu se v matematické statistice zavádí pojem intervalový odhad. Intervalový odhad vymezuje interval, ve kterém leží odhadovaný parametr s pravděpodobností $1 - \alpha$, kde α je libovolně volená konstanta ležící v otevřeném intervalu od nuly do jedničky. Číslo $1 - \alpha$ se taktéž nazývá koeficient spolehlivosti. Čím vyšší je koeficient spolehlivosti, čím bližší je číslu jedna, tím širší je konfidenční interval. Hodnotu $100*\alpha$ lze prezentovat taktéž jako očekávaný počet intervalů spolehlivosti, nepokrývajících odhadovaný parametr, vytvořených ze sta náhodných výběrů. Je-li očekávaný počet intervalů spolehlivosti nepokrývajících parametr θ na sto náhodných výběrů roven $100*\alpha$, jedná se o takzvaně přesný interval spolehlivosti. Klasické intervaly spolehlivosti se konstruují za pomocí výběrových rozdělení, z čehož plyne, že v klasickém přístupu je nutné splnit určitá kritéria pro tvorbu přesných intervalů spolehlivosti.

Za pomocí bootstrapu lze tvorbu intervalů spolehlivosti zobecnit i na případech, na které nelze aplikovat centrální limitní věty a jiné aproximace k známým rozdělením. V této kapitole demonstruji fungování dvou nejzákladnějších metod tvorby bootstrapových intervalů spolehlivosti. Uvedu zde pouze obecné formy intervalových odhadů. Specifickými případy se budu zabývat pouze u aplikace na reálná či simulovaná data.

5.1 Bootstrapový-t interval

Bootstrapový interval spolehlivosti je robustní alternativou klasického odhadu. Je doporučeno jej využít především v případě, že nelze využít klasický přístup při tvorbě intervalů spolehlivosti, a to například je-li vzorek příliš malý k využití asymptotických vlastností, nebo je-li populační rozdělení příliš složité. Bootstrapový interval spolehlivosti lze použít univerzálně pro všechny druhy populačních parametrů. Kvalita se bohužel dle rozdělení a typu parametru, jehož interval spolehlivosti je odhadován, liší.

Princip bootstrapového-t intervalu je jednoduchý. Je provedeno alespoň tisíc bootstrapových výběrů. Pro každý z těchto výběrů je dopočten bootstrapový odhad parametru θ. Je-li $\widehat{\boldsymbol{\theta}}^*$, viz vzorec č. *4.1.3*, blízký normálnímu rozdělení, pak lze předpokládat následující vztah:

$$Z = \frac{\hat{\theta}-\theta}{\widehat{se}_B(\hat{\theta})} \sim t_{n-1} \qquad 5.1.1$$

Bootstrapový-t interval poté bude mít následující formu:

$$P[\hat{\theta} - t_{n-1}^{1-\frac{\alpha}{2}} * \widehat{se}_B(\hat{\theta}) < \theta < \hat{\theta} + t_{n-1}^{1-\alpha/2} * \widehat{se}_B(\hat{\theta})] = 1-\propto \qquad 5.1.2$$

Ověřit normalitu $\widehat{\boldsymbol{\theta}}^*$ je možné mnoha způsoby. Nejpraktičtější se mi osobně jeví Q-Q plot, který porovnává kvantily ověřovaného rozdělení a graficky je porovnává s kvantily normálního rozdělení, popřípadě Shapiro-Wilkův test fungující na podobném principu. (Bartoszyński & Niewiadomska-Bugaj, 2008)

Avšak normalita není pro $\widehat{\boldsymbol{\theta}}^*$ ničím garantována. Naopak jen pro určité typy odhadů populačních parametrů vyjde rozdělení bootstrapového odhadu θ normální. Nicméně i v případě nesplnění normality je možné využít bootstrapového-t intervalu, i když existuje jednodušší metoda, které se budu věnovat v příští kapitole. Tedy nesplňuje-li $\widehat{\boldsymbol{\theta}}^*$ předpoklad normality, statistika Z již není aproximovatelná k studentovu rozdělení. Nicméně můžeme říci, že s pravděpodobností $1-\alpha$ leží tato statistika v intervalu $(t^{\alpha}, t^{1-\alpha})$, kde jednotlivé meze intervalu jsou příslušné kvantily neznámého rozdělení.

$$P\left(t^{\alpha} < \frac{\hat{\theta}-\theta}{\widehat{se}(\hat{\theta})} < t^{1-\alpha}\right) = 1 - \frac{\alpha}{2} \qquad 5.1.3$$

V případě, že existuje podobné analytické vyjádření, jakým je například vzorec č. *3.2.6* pro odhad směrodatné odchylky odhadu parametru θ, a tedy existuje analytická forma výpočtu směrodatné odchylky odhadu, konstrukce bootstrapového-t intervalu spolehlivosti probíhá v následujících krocích:

1. Je realizováno *B* bootstrapových výběrů z náhodného výběru **x**. Tentokrát je *B* voleno vyšší než tisíc.
2. Pro každý z výběrů se vypočte bootstrapový odhad $\hat{\theta}^*(b)$ a dle analytického vzorce a je dopočtena jeho směrodatná odchylka $\widehat{se}[\hat{\theta}^*(b)]$. Výpočet směrodatné odchylky probíhá tak, že se do vzorce pro výpočet směrodatné odchylky odhadu dosadí místo původního výběru **x** bootstrapový výběr $\boldsymbol{x}_b^*$.

3. Vypočte se pro každý bootstrapový odhad statistika

$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{se}[\hat{\theta}^*(b)]} \quad b \in \{1, 2, \ldots, B\} \qquad 5.1.4$$

$$\boldsymbol{Z}^* = Z^*(1), Z^*(2), \ldots, Z^*(B) \qquad 5.1.5$$

4. Odhadne se α a $1 - \alpha$ kvantilu $\boldsymbol{Z}^*$ rozdělení. Příslušný kvantil neznámého rozdělení $\hat{t}^{\alpha}$ se odhadne tak, že je stanovena hodnota vektoru $\boldsymbol{Z}^*$ dopočítaných statistik v kroku tři taková, že počet hodnot v tomto vektoru menších nebo rovných $\hat{t}^{\alpha}$ vydělený celkovým počtem bootstrapových výběrů B je roven α.

$$\frac{\#[\boldsymbol{Z}^* \leq \hat{t}^{\alpha}]}{B} = \alpha \qquad 5.1.5$$

5. Dopočítá se bootstrapový-t interval

$$P[\hat{\theta} - \hat{t}^{1-\alpha/2} * \widehat{se}(\hat{\theta}) < \theta < \hat{\theta} - \hat{t}^{\alpha/2} * \widehat{se}(\hat{\theta})] = 1 - \alpha \qquad 5.1.6$$

Směrodatná odchylka odhadu je zde vypočítána dostupnou analytickou formulí. Neexistuje-li tato analytická formule, je třeba výpočet modifikovat. Namísto dosazení do analytického vyjádření $\widehat{se}[\hat{\theta}^*(b)]$, je potřeba pro každý bootstrapový výběr odhadnout, dalším vnořeným bootstrapovým odhadem směrodatné odchylky $\widehat{se}_{B_2}[\hat{\theta}^*(b)]$, onu směrodatnou odchylku. Namísto $\widehat{se}(\hat{\theta})$ je v kroku pět dosazen bootstrapový odhad směrodatné odchylky $\widehat{se}_{B_1}(\hat{\theta})$. Je jasné, že počet celkových simulací je $B_1 * B_2$, kde B_1 reprezentuje počet bootstrapových replikací nutných ke kvalitnímu odhadu kvantilu $\hat{t}^{\alpha}$ a B_2 indikuje počet vnořených bootstrapových replikací nutných k určení kvalitního odhadu směrodatné odchylky $\widehat{se}^*_{B_2}(\hat{\theta})$. (Effron & Timbishirani, 1993)

Příklad č. 1 – Bootstrapový-t intervalový odhad střední hodnoty

Do experimentu bylo vybráno náhodně deset studentů. Byl zkoumán účinek léku na spaní. Data zachycují průměrný denní přírůstek spánku v hodinách.

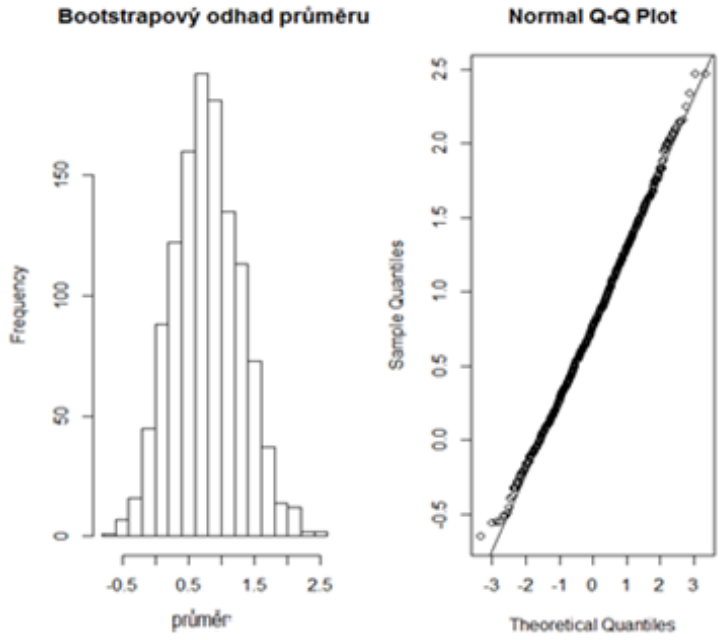
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0,7	-1,6	-0,2	-1,2	-0,1	3,4	3,7	0,8	0,0	2,0

$\bar{x} = 0{,}75$

$\alpha = 0{,}05$

$n = 10$

Abychom vycházeli z reprezentativního testu normálního rozdělení $\widehat{\boldsymbol{\theta}}^*$, je nutno generovat alespoň tisíc bootstrapových replikací. Nižší počet replikací by mohl vést k chybnému vyhodnocení normality.



Graf číslo 6 – Rozdělení $\widehat{\theta}^*$ pro výběrový průměr
***Zdroj: vlastní zpracování dat* (R Core Team, 2012)**

Na základě Shapirova-Wilkova testu nezamítám na hladině významnosti hypotézu o normálním rozdělení bootstrapového odhadu průměrného hodinového přírůstku spánku za den. Pro porovnání intervalů spolehlivosti budu tedy aplikovat Bootstrapový-t interval založený na normalitě bootstrapového odhadu parametru θ, obecný bootstrapový-t interval s aplikací analytické formule pro výpočet rozptylu odhadu a analytický t-studentův interval spolehlivosti.

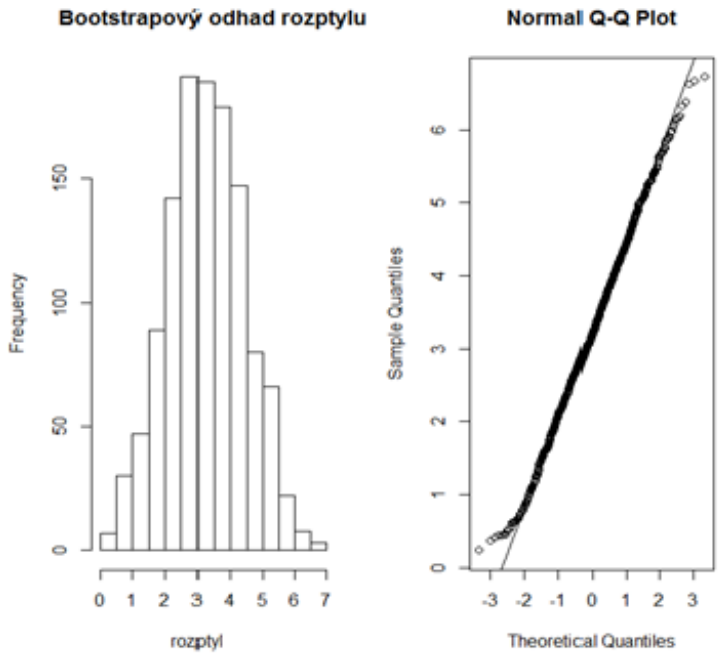
Užitá metoda	dolní mez	Střed intervalu	horní mez
Bootstrap-t (obecný)	-0,291	0,685	1,660
Bootstrap-t (normalita)	-0,397	0,75	1,897
Analytický interval	-0,599	0,75	2,099

Výstup číslo 1 – Intervaly spolehlivosti pro odhad střední hodnoty
***Zdroj vlastní zpracování na základě dat* (R Core Team, 2012)**

Z výstupu je patrné, že všechny odhady vyšly velice podobně. Nejužším intervalem je obecný bootstrapový-t interval a nejširším je analytický interval spolehlivosti. Je možné zaznamenat dvě menší postřehnutí. Prvním je, že Bootstrapový-t interval založený na normalitě $\widehat{\boldsymbol{\theta}}^*$ a analytický interval se v tomto případě liší pouze dosazením bootstrapového odhadu směrodatné odchylky, do konečného intervalu spolehlivosti, namísto analytické formule pro odhad směrodatné odchylky výběrového průměru. Druhým postřehnutím je, že obecný bootstrapový-t interval spolehlivosti není symetrický okolo odhadu parametru θ.

Příklad č. 2 – Bootstrapový-t intervalový odhad rozptylu

Pro data z příkladu č. *1* určím intervalový odhad rozptylu.



Graf číslo 7 - Rozdělení $\widehat{\boldsymbol{\theta}}^*$ pro výběrový rozptyl
***Zdroj vlastní zpracování na základě dat* (R Core Team, 2012)**

Opět na základě Shapirova-Wilkova testu nevyvracím hypotézu o normalitě na hladině významnosti pěti procent.

Užitá metoda	Dolní mez	Střed intervalu	Horní mez
Bootstrap-t (normalita)	0.861	3,556	6.252
Analytický interval	1.682	6,767	11.8522

Výstup číslo 2 – Intervaly spolehlivosti pro odhad rozptylu
***Zdroj vlastní zpracování na základě dat* (R Core Team, 2012)**

Z vypočtených hodnot je vidět, menší odlišnost mezi oběma intervaly. Analytický interval, předpokládající normální rozdělení populace, je mnohem širší a střed jeho intervalu leží až za hranicí horní meze obecného bootstrapového-t intervalu. Taktéž je možné si všimnout, že analytický interval není symetrický okolo odhadu parametru θ. Tato skutečnost je dána nesymetrií chí-kvadrát rozdělení. Pro určení intervalu spolehlivosti rozptylu metodou bootstrap existují vhodnější „pivotové" bootstrapové intervaly spolehlivosti, vycházející z analytické statistiky aproximované k chí-kvadrát rozdělení.

5.2 Kvantilový bootstrapový interval

Kvantilové intervaly spolehlivosti vycházejí přímo z rozdělení $\widehat{\boldsymbol{\theta}}^*$. Toto rozdělení nemusí být nutně normální. Pochopitelně se naskýtá otázka, jaká z těchto dvou metod konstrukce bootstrapového intervalu spolehlivosti je lepší. Je-li situace taková, že $\widehat{\boldsymbol{\theta}}^*$ vykazuje statisticky významné odchylky od normality, je alespoň dle mého názoru lepší využít kvantilový přístup. V obecné rovině se počet replikací pro bootstrapový-t interval v takovém případě pohybuje řádově v desítkách tisíc. V případě, je-li $\widehat{\boldsymbol{\theta}}^*$ normální, je nejintuitivnější přístup zkonstruovat oba intervaly a vzájemně je porovnat. Liší-li se nějakým významným způsobem, či se dokonce nepřekrývají, nastala někde chyba a neměl by být použit ani jeden z nich. Obecně ale platí, že je-li bootstrapový odhad vychýlení nízký, doporučuje se užít kvantilový bootstrapový interval. To samé platí i v situaci, kdy je výběr příliš malý a většina variability bootstrapového odhadu je způsobena špatnou aproximací empirické distribuční funkce, viz kapitolu č. *4.3*. (Bartoszyński & Niewiadomska-Bugaj, 2008)

Kvantilový bootstrapový interval lze odhadnout následujícím algoritmem:

1. Ze vzorku **x**, se vybere *B* bootstrapových výběrů. *B* by mělo být opět voleno vyšší než tisíc.

2. Vypočítáme $\hat{\theta}^*(b)$ pro každý bootstrapový výběr.

3. Nechť $\hat{\theta}_B^{*(\alpha/2)}$ a $\hat{\theta}_B^{*(1-\alpha/2)}$ jsou $\frac{\alpha}{2}$ kvantil a $\left(1-\frac{\alpha}{2}\right)$ kvantil $\widehat{\boldsymbol{\theta}}^*$, pak

$$P(\hat{\theta}^*_{\%\,Dolní} < \theta < \hat{\theta}^*_{\%\,Horní}) = 1-\alpha \qquad 5.2.1$$

$$[\hat{\theta}^*_{\%\,Dolní}, \hat{\theta}^*_{\%\,Horní}] \sim [\hat{\theta}_B^{*(\alpha/2)}, \hat{\theta}_B^{*(1-\alpha/2)}] \qquad 5.2.2$$

$$P(\hat{\theta}_B^{*(\alpha/2)} < \theta < \hat{\theta}_B^{*(1-\alpha/2)}) = 1 - \alpha \qquad 5.2.3$$

Lze si všimnout, že tento interval je velice nenáročný, na rozdíl od bootstrapového-t intervalu, na počet nutných replikací i pro složité odhady. (Effron & Timbishirani, 1993)

5.3 BC_a intervaly spolehlivosti

Pod třemi písmeny BC_a se skrývá zkratka z anglického *Bias-corrected and Accelerated*. Jedná se o vylepšení metody kvantilových odhadů. V případě, je-li vychýlení odhadu příliš velké, aby mohla být aplikována kvantilová metoda bootstrapového odhadu intervalu spolehlivosti, není nutné uchylovat se k bootstrapovému-t intervalu. Zvláště nastává-li výše popsaná situace, kdy je třeba u bootstrapového-t intervalu generovat $B_1 * B_2$ replikací, je výhodné zvážit užití BC_a intervalu, který nabízí mnohem kvalitnější odhad intervalu spolehlivosti. Od kvantilového odhadu se postup liší pouze posledním, třetím krokem algoritmu.

$$BC_a: P\left(\hat{\theta}_B^{*(\alpha 1)} < \theta < \hat{\theta}_B^{*(\alpha 2)}\right) = 1 - \alpha \qquad 5.3.1$$

$$\alpha 1 = \Phi[\hat{u}_0 + \frac{\hat{u}_0 + u^{\alpha/2}}{1 - \hat{a}(\hat{u}_o + u^{\alpha/2})}] \qquad 5.3.2$$

$$\alpha 2 = \Phi[\hat{u}_0 + \frac{\hat{u}_0 + u^{1-\alpha/2}}{1 - \hat{a}(\hat{u}_o + u^{1-\alpha/2})}] \qquad 5.3.3$$

$\hat{u}_0$	koeficient opravy na nevychýlenost
$\hat{a}$	koeficient zrychlení
u^{α}	kvantil normovaného normálního rozdělení

Oba koeficienty se dají odhadnout následujícími vzorci.

$$\hat{u}_0 = \Phi^{-1}[\frac{\#[\hat{\theta}^*(b) < \hat{\theta}]}{B}] \qquad 5.3.4$$

Tedy $\hat{u}_0$ se vypočítá inverzní funkcí kvantilu normovaného normálního rozdělení, a argument této funkce je roven počtu čísel ve vektoru $\hat{\boldsymbol{\theta}}^*$ menších než je odhad parametru θ, $\hat{\theta}$, vydělený počtem bootstrapových replikací. Koeficient opravy na nevychýlenost je odhadem vychýlení mediánu rozdělení $\hat{\theta}$.

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^3}{6 * \sqrt{[\sum_{i=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2]^3}} \qquad 5.3.5$$

Odhad koeficient zrychlení $\hat{a}$ se spočte za pomocí za pomocí metody jackknife, viz kapitolu č. *3.3* pro bližší vysvětlení. Vyjadřuje rychlost změny směrodatné odchylky odhadu. (Effron & Timbishirani, 1993)

Příklad č. 3 – Porovnání metod tvorby intervalu spolehlivosti pro střední hodnotu

V tomto příkladu bude proveden simulační test, na kterém porovnám vhodnost metod užití konstrukce jednotlivých intervalů spolehlivosti. Cílem bude vytvořit intervalovou předpověď střední hodnoty délky života jednotlivých pacientů nakažených nemocí AIDS od dne diagnózy. Zúženým základním souborem bude celý balíček dat, ze kterého bude proveden náhodný výběr. Během samotné statistické indukce budu přistupovat k vyhodnocování, jako kdyby daná populace nebyla dostupná, pouze α_{simul} bude vypočten za pomocí znalostí populace.

Bylo vybráno třicet hodnot bez vracení[2] ze základního souboru.

$\bar{x} = 382{,}83$

$\alpha = 0{,}05$

$n = 30$

$B = 1200$

Užitá metoda	**dolní mez**	**střed intervalu**	**horní mez**	α_{simul}
Analytický interval	253,98	382,83	511,69	0,058
Bootstrap-t (obecný)	251,18	373,33	495,32	0,040
Kvantilový odhad	270,34	392,42	514,49	0,092
BC$_\alpha$ odhad	269,59	391,34	513,09	0,066

Výstup číslo 3 – Intervalové odhady střední hodnoty
Zdroj vlastní zpracování dat (R Core Team, 2012)

[2] Výběrem bez vracení z konečné populace nebyl porušen náhodný výběr, neboť výběrový podíl je menší než 0,05 (konkrétně je roven přibližně 0,01). (Hebák, Bílková, & Svobodová, 2009)

$\boldsymbol{\alpha_{simul}}$ je Monte Carlo odhad přesnosti intervalu. Bylo vytvořeno náhodným výběrem bez vracení tisíc intervalů danou zkoumanou metodou a následně byl počet intervalů ležících mimo skutečný odhadovaný parametr vydělen tisíci.

Z výstupu je patrné, že v tomto případě vyšel nejlépe analytický odhad, nicméně jen díky tomu, že vzorek byl již dostatečně velký k aplikaci centrální limitní věty. Nejhůř dopadnul obecný kvantilový interval. Naopak obecný bootstrapový-t interval vyšel dobře, z výsledků vyplývá, že mírně podhodnocuje koeficient α. Taktéž BC_α interval vyšel velice dobře. Tato studie koresponduje s doporučením, že tvorba intervalů spolehlivosti za pomocí metody bootstrap se doporučuje především tehdy, nejsou-li splněny podmínky pro analytický přístup.

Příklad č. 4 – Porovnání tvorby intervalu spolehlivosti pro mezikvantilový průměr

Nyní na ta samá data jako v příkladu č. *3* aplikuji vychýlenější odhad, mezikvantilový průměr. Mezikvantilový průměr je robustním odhadem střední hodnoty pro symetrická rozdělení. Vypouští pozorování obsažená za hranicí voleného β a *1-* β kvantilu výběru.

$\bar{x}^{0.05} = 364{,}79$

$\alpha = 0{,}05$

$B = 1200$

Vychýlení bylo odhadnuto za pomocí metody bootstrap:

$\widehat{(b)}_{500}(\bar{x}^{0.05}) = -1{,}13$

Užitá metoda	**dolní mez**	**střed intervalu**	**horní mez**	$\boldsymbol{\alpha_{real}}$
Bootstrap-t (obecný)	248,99	388,96	528,93	—
Kvantilový odhad	253.62	376,00	494.66	0,046
BC_α odhad	253.64	379,14	494.93	0,048

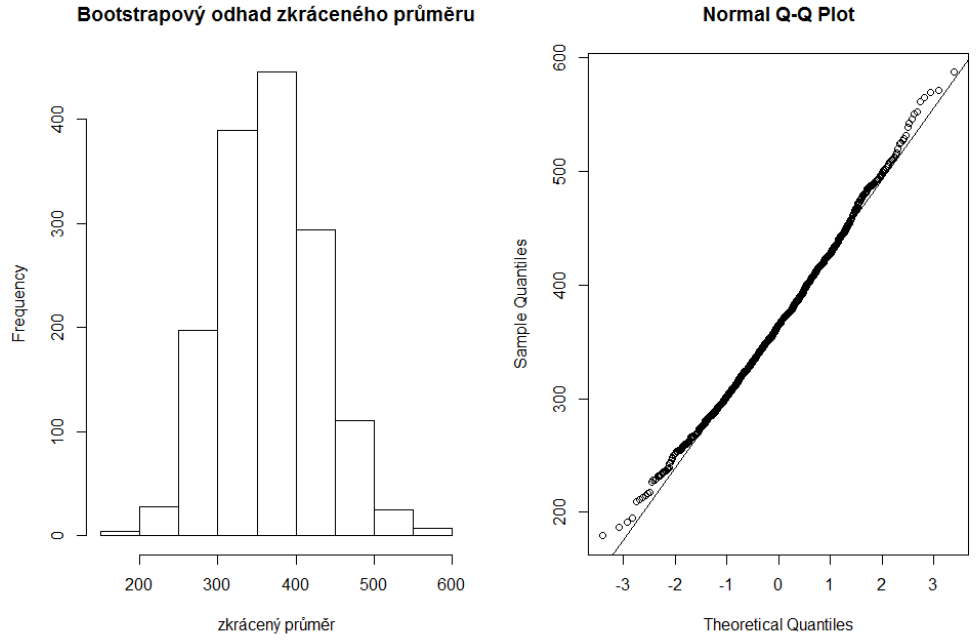
Výstup číslo 4 – Robustní intervalový odhad střední hodnoty
Zdroj vlastní zpracování dat (Ripley & Venables, 2002)

$\hat{u}_0 = 0{,}0017$

Koeficient opravy na nevychýlenost se blíží nule, což znamená, že vychýlení není nikterak markantní. V podstatě bootstrapové replikace zkráceného průměru jsou téměř symetrické kolem analytického odhadu.

$\hat{a} = 0{,}031$

Odhad koeficientu zrychlení je taktéž velice nízký. Je vidět, že BC_a odhad opět vylepšil přesnost kvantilového odhadu. V tomto případě nebyla přesnost počítána pro bootstrapový-t interval z důvodu velké náročnosti simulace. Pro dosažení výsledku v rozumném čase by bylo třeba zmenšit některý z parametrů simulace a tím by se porušila porovnatelnost s ostatními simulacemi.



Graf číslo 8 - Rozdělení $\widehat{\theta}^*$ pro mezikvantilového průměru
***Zdroj vlastní zpracování dat* (Ripley & Venables, 2002)**

Příklad č. 5 - Porovnání tvorby intervalu spolehlivosti pro medián

Nyní porovnám různé metody konstrukce na nehladkém odhadu, mediánu, jehož bootstrapová replikace nebude vycházet blízké normálnímu rozdělení.

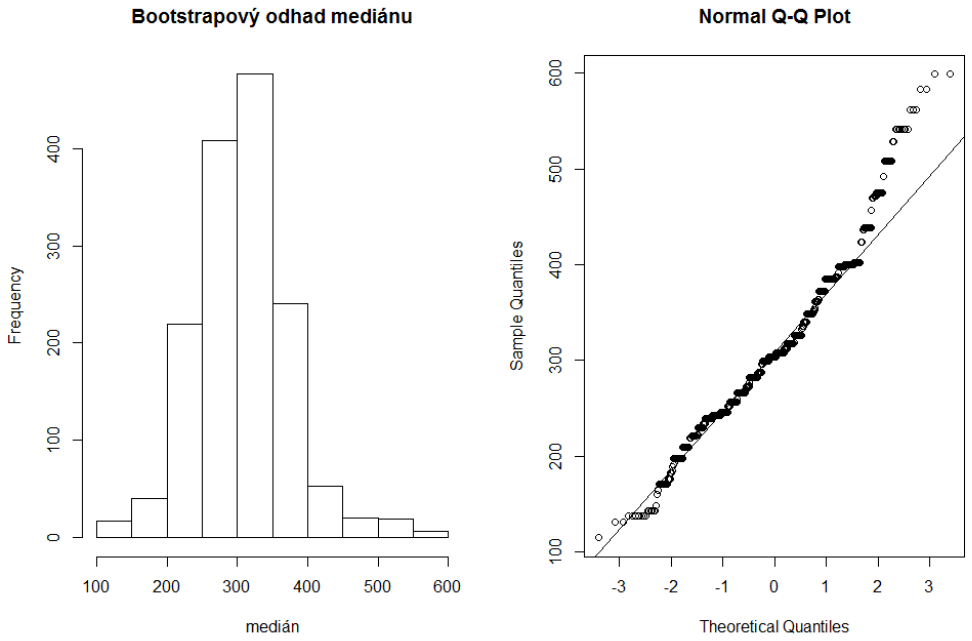
$\tilde{x} = 303{,}5$

$\hat{b}_{500}(\tilde{x}) = -2{,}409$

$B = 1200$

Užitá metoda	dolní mez	střed intervalu	horní mez	α_{real}
Bootstrap-t (obecný)	193,77	310,19	426,60	—
Kvantilový odhad	282,96	435,8	588,65	0,044
BC$_\alpha$ odhad	293,92	449	604,09	0,046

Výstup č. 5 – Interval spolehlivosti pro odhad mediánu
Zdroj vlastní zpracování dat (Ripley & Venables, 2002)



Graf číslo 9 - Rozdělení $\widehat{\theta}^*$ pro býběrový medián
***Zdroj vlastní zpracování dat* (Ripley & Venables, 2002)**

Z výstupu lze usoudit, že přesnost není vůbec špatná. Navíc je vidět, že BC$_\alpha$ dokázal přesnost odhadu vylepšit vzhledem k volenému koeficientu α. Bohužel u BC$_\alpha$ odhadu je problém v tom, že koeficient zrychlení zde vyjde nulový, což nemusí do jisté míry vypovídat nic o koeficientu samotném jako o faktu, že metoda jackknife, jak jsem již uvedl v kapitole č. *3.3*, není vhodnou aplikací pro nehladké odhady. Existuje mnoho dalších metod odhadů koeficientu zrychlení, nicméně v tomto případě je lepší rovnou zvolit za $\hat{a}$ nulu, než dojít ke zkresleným závěrům, a tak získám tak zvaný BC odhad, čili pouze „*Biass corrected*“.

5.4 Implementace tvorby bootstrapových intervalů spolehlivosti

5.4.1 Parametrický bootstrapový odhad intervalů spolehlivosti

Stejně jako u parametrického bootstrapového odhadu, známe-li rozdělení základního souboru, neparametrickým přístupem není možno dosáhnout lepšího výsledku než parametrickým bootstrapovým odhadem. Stejně tak i u tvorby intervalů spolehlivosti, je-li známé pravděpodobnostní rozdělení základního souboru, dosáhneme přesnějšího intervalu spolehlivosti aplikací parametrického bootstrapu. Postup je identický, liší se pouze v metodě generování bootstrapových výběrů, viz kapitolu č. *4.4*.

5.4.2 Snížení počtu nutných bootstrapových replikací

Počet nutných bootstrapových replikací lze obecně snížit tehdy, sníží-li se rozptyl Monte Carlo simulace užité při generování bootstrapových výběrů. Snížení rozptylu ovšem nesmí jít na úkor porušení konzistence Monte Carlo aproximace. I dnes, když má téměř každý k dispozici superrychlé procesory, jsou tyto metody používané. Speciálně při tvorbě intervalů spolehlivosti, kde například v metodě bootstrapového-t intervalu, kdy je generováno $B_1 * B_2$ bootstrapových replikací, dokáží tyto algoritmy snížit obě B o konstantu. (Hall, Antithetic Resampling for the Bootstrap, 1989), (Hall, On efficient bootstrap simulation, 1989)

5.4.3 ABC intervaly spolehlivosti

ABC je aproximací metody BC_{α}, kdy meze bootstrapového intervalu spolehlivosti jsou v tomto případě odvozeny analyticky. Monte Carlo simulace je zde nahrazena aproximací Taylorovým rozvojem. ABC metoda funguje jako kvalitní odhad intervalu spolehlivosti pouze tehdy, je-li výběrová statistika $t(\boldsymbol{x})$ hladkým odhadem. (Effron & Timbishirani, 1993)

6 Testování statistických hypotéz užitím metody bootstrap

Testování statistických hypotéz je určitým typem statistické indukce. Test sestává ze dvou základních entit, z nulové hypotézy a alternativy. Cílem testu je rozhodnout o platnosti nulové hypotézy, ergo výsledkem testu je buď zamítnutí nulové hypotézy ve prospěch alternativy, nebo nezamítnutí nulové hypotézy. Hypotéza a alternativa musí být dva disjunktní jevy. Za výše zmíněných předpokladů, je možné docílit čtyř možných situací:

	Hypotéza platí	alternativa platí
Zamítáme hypotézu	Chyba I. typu (α)	Síla testu (1-β)
Nezamítáme hypotézu	Správně (1-α)	Chyba II. typu (β)

Výstup číslo 6 – Testová rozhodnutí a jejich podmíněné pravděpodobnosti
Zdroj vlastní zpracování

Zpravidla existují tři metody vyhodnocení testu statistických hypotéz: testování skrz intervaly spolehlivosti, kritickou hodnotu a p-hodnotou.

Otázkou je, kdy je výhodné využít metodu bootstrap pro testování statistických hypotéz. Nejčastější důvod vedoucí k užití metody bootstrap pro testování statistických hypotéz je porušení podmínek, znemožňující užití klasického parametrického přístupu. Jako příklad bych uvedl typicky neznámé rozdělení a příliš malý vzorek znemožňující užití asymptotických vlastností. Další otázkou je, proč nevyužít jiných neparametrických metod? Bootstrap je na rozdíl od ostatních neparametrických metod univerzální. Lze jej s určitými modifikacemi použít na téměř libovolné typy testů.

6.1 Testování bootstrapovými intervaly spolehlivosti

Nejjednodušší metodou testování statistických hypotéz je prosté dosazení do intervalu spolehlivosti. Je-li testována hypotéza o rovnosti populačního parametru zvolené konstantě na hladině významnosti α, zkonstruuje se pouze příslušný interval spolehlivosti a zjistí se, leží-li testovaná konstanta uvnitř či vně intervalu. Leží-li vně, zamítáme hypotézu na hla-

dině významnosti α, leží-li uvnitř, na dané hladině významnosti hypotézu nezamítáme. Je potřebné dávat pozor při dosazování hladiny významnosti do rovnice. U jednostranné hypotézy musí být za α dosazen dvojnásobek testované hladiny (pouze v případě zápisu P(…)=1-α), čímž je test proveden na dvojnásobné hladině významnosti, a konstanta je porovnána pouze s příslušnou stranou intervalu spolehlivosti. Naopak je-li testována dvoustranná hypotéza, je za α dosazena požadovaná hladina významnosti a konstanta je porovnána s celým intervalem.

6.2 Permutační test

Jak již bylo v této kapitole nastíněno, pro vyhodnocení nulové hypotézy je potřeba nějakým způsobem určit rozdělení testového kritéria. Bohužel ne vždy je toto umožněno, respektive častokrát bývá toto rozdělení asymptotické. Avšak je-li velikost výběru malá, je potřeba aplikovat nějaký exaktní test nezávislý na asymptotických vlastnostech. Jsou-li dva výběry na sobě nezávislé, je možné testovat například, zda-li pocházejí ze stejného rozdělení.

$$F \rightarrow \boldsymbol{y} = \{y_1, y_2, \dots, y_n\} \tag{6.2.1}$$

$$G \rightarrow \boldsymbol{z} = \{z_1, z_2, \dots, z_m\} \tag{6.2.2}$$

$$H_0: F = G \tag{6.2.3}$$

Platí-li nulová hypotéza, pak jsou obě rozdělení stejná a pravděpodobnost vybrání libovolné podmnožiny A z jednoho rozdělení, je rovna pravděpodobnosti vybrání té samé podmnožiny A z druhého rozdělení.

$$P(X = A) = P(Z = A) \tag{6.2.4}$$

Je vytvořen sloučený vektor **v**, čímž je myšlen vektor sdružující vektor **y** a vektor **z**, a indexový vektor **g**, určující, ke kterému ze dvou testovaných výběrů daná hodnota náleží, nula přiřazuje hodnotu k vektoru **y** a jedna přiřazuje hodnotu k vektoru **z**. Následně vypočteno příslušné testové kritérium

$$\boldsymbol{v} = \{y_1, y_2, \dots, y_n, z_1, z_2, \dots, z_m\} \tag{6.2.5}$$

$$\boldsymbol{g} = \{0,0, \dots,0,1,1, \dots,1\} \tag{6.2.6}$$

$$\hat{\theta} = t(\boldsymbol{g}, \boldsymbol{v}) \qquad 6.2.7$$

Algoritmus permutačního testu je následující:

1) Je vybráno *B* náhodných permutací vektoru **g**. Sloučený vektor **v** je po celou dobu testu fixován.

$$\boldsymbol{g}^*(b) \quad b \in 1, 2, \ldots, B \qquad 6.2.8$$

2) Pro každý z těchto vektorů je vypočtena replikace příslušného kritéria

$$\widehat{\boldsymbol{\theta}}^*(b) = t[\mathbf{g}^*(b), v] \qquad 6.2.9$$

$$\widehat{\boldsymbol{\theta}}^* = \{\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)\} \qquad 6.2.10$$

3) Je aproximována p-hodnota permutačního testu pro pravostrannou alternativu

$$\hat{p} = \frac{\#\{\widehat{\boldsymbol{\theta}}^* \geq \widehat{\theta}\}}{B} \qquad 6.2.11$$

Tvar aproximace p-hodnoty záleží na tvaru alternativy. Ve vzorci č. *6.2.*11 je popsán test pro pravostrannou alternativu. Je-li alternativa dvoustranná, pak se dělenec ve vzorci mění na $\#\{|\widehat{\boldsymbol{\theta}}^*| > |\hat{\theta}|\}$ a na závěr, jedná-li se o levostrannou alternativu, pak se dělenec mění v $\#\{\widehat{\boldsymbol{\theta}}^* \leq \hat{\theta}\}$. Tuto analogii budu dodržovat i v následujících kapitolách.

Otázkou je, kolik permutací je požadováno. Záleží, jak přesný odhad p-hodnoty je požadován. Je logické, že bude-li zvoleno *B* rovno stu, nejnižší možná p-hodnota bude 0,01. Standardně se volí *B* vyšší tisíci. Počet replikací lze určit i díky znalosti rozdělení $\hat{p}$, které má alternativní rozdělení. Tedy lze vyjádřit následující.

$$B * \hat{p} \sim Bi(B, p) \qquad 6.2.12$$

$$E(\hat{p}) = p \qquad 6.2.13$$

$$D(\hat{p}) = \frac{p*(1-p)}{B} \qquad 6.2.14$$

$$V(\hat{p}) = \sqrt{\frac{p*(1-p)}{p*B}} \qquad 6.2.15$$

Testuji-li hypotézu na hladině významnosti 5%, zamítám hypotézu, bude-li $\hat{p}$ menší nebo rovno 0,05. Z formule *6.2.14* plyne, že skutečná p-hodnota permutačního testu se může od aproximace lišit při *B* rovno tisíci zhruba o 0,016. Tedy leží-li aproximovaná p-hodnota $\hat{p}$

v mezích $< \alpha - se(\hat{p}), \alpha + se(\hat{p}) >$, je třeba zpřísnit test a zvolit o řád vyšší *B*. Například může nastat situace, kdy $\hat{p}$ bude rovno 0,049 a zamítnutím hypotézy bychom se dopustil chyby, nýbrž přesnější p-hodnota, ke které bychom se dopracovali po zvýšení *B* o desetinásobek, by vyšla třeba 0,052.

Je-li vzorek malý, je možné provést permutační test bez Monte Carlo simulace. Je-li vektor ***v*** složen z *n* pozorování, vybraných z prvního rozdělení *F*, a z *m* pozorování, vybraných z rozdělení *G*, je možné docílit přesně $\binom{n+m}{n}$ různých odhadů dle formule *6.2.9*. Tento počet kombinací je odvozen z faktu, že výsledek získaný dle této formule je určen prvními *n* čísly ve vektoru **g**. Je-li fixováno prvních *n* čísel v tomto vektoru, pak pro libovolnou permutaci zbylých *m* čísel je výsledek formule *6.2.9* stejný. (Rice & Thomas, 2008) (Effron & Timbishirani, 1993)

Příklad č. 6 – Test rovnosti středních hodnot dvou jinak stejných rozdělení

V tomto příkladu porovnám za pomocí permutačního testu rozdíl středních hodnot spánkového přírůstku po aplikaci prvního typu a druhého typu prášku.

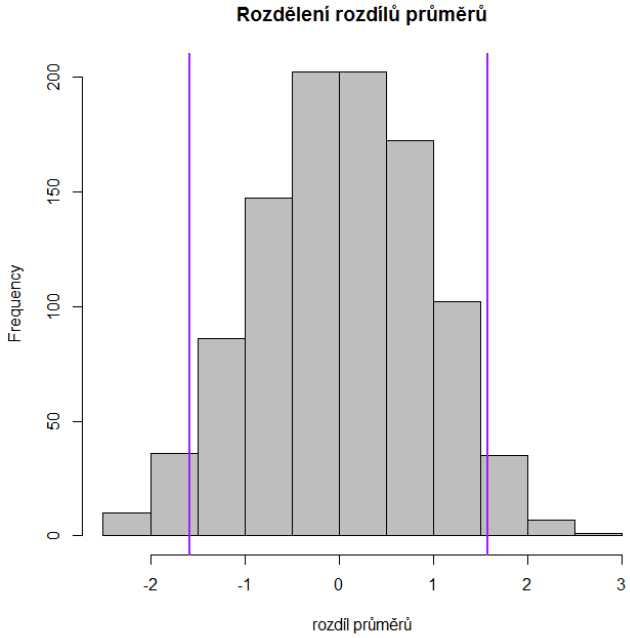
$H_0: \mu_y - \mu_z = 0$

$H_1: \mu_y - \mu_z \neq 0$

$$\hat{\theta} = \frac{\sum_{g_i=0} v_i}{n} - \frac{\sum_{g_i=1} v_i}{m} \qquad 6.2.15$$

$\hat{\theta} = -1{,}58$

Permutačním testem ověřím, je-li tento rozdíl natolik signifikantní, abych byl schopen na rozumné hladině významnosti zamítnout hypotézu o rovnosti středních hodnot obou rozdělení. Test provedu v dvoustranné alternativě.



Graf číslo 10 - Rozdělení $\widehat{\theta}^*$ pro rozdíl průměrů
Zdroj vlastní zpracování dat **(R Core Team, 2012)**

Aproximace p-hodnoty dvoustranného permutačního testu byla rovna 0,07. To znamená, že na základě toho testu jsem nebyl schopen označit rozdíl mezi průměry obou rozdělení jako signifikantní. Je nutné si povšimnout, že rozdělení četností rozdílů průměru není vzhledem k aproximaci symetrické, tudíž pro jednostranný test nestačí aproximovanou p-hodnotu vydělit dvěma, respektive následně odečíst od jedničky.

Příklad č. 7 – Srovnání t-testu a permutačního testu

Byl simulován náhodný výběr ze dvou normálních rozdělení, lišících se od sebe pouze střední hodnotou.

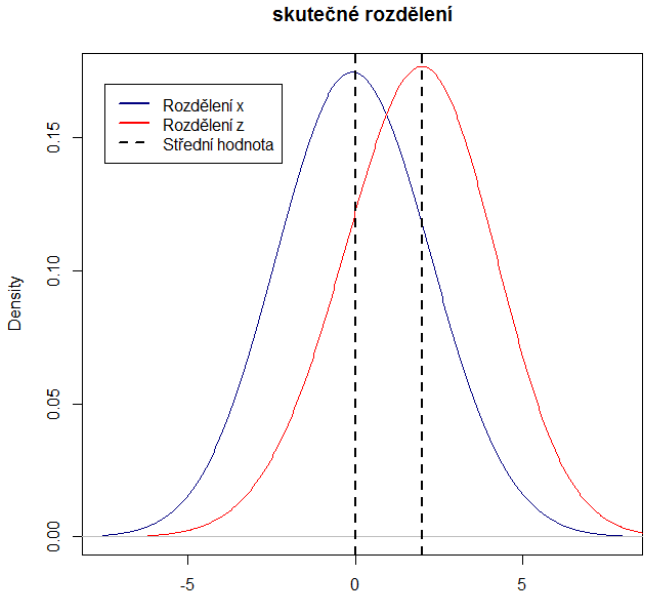
$\boldsymbol{x} \sim N(0,1)$

$\boldsymbol{z} \sim N(2,1)$

$H_0: \mu_x = \mu_z$

$H_1: \mu_x \neq \mu_z$

Z prvního rozdělení bylo náhodně generováno deset hodnot, z druhého pouze pět.



Graf číslo 11 – Graf dvou normálních rozdělení s různou střední hodnotou
Zdroj vlastní zpracování simulovaných dat

U oboustranného t-testu, který může být v tomto případě normálních populačních rozdělení použit i pro takto malé vzorky, vyšla p-hodnota rovna 0,014.

Permutačním testem vychází aproximativní p-hodnota 0,009. Zvýším-li počet permutací o desetinásobek, vyjde již tato aproximace 0,0114. Je možné se v literatuře setkat s tím, že mezi permutace je zahrnut i původní indexový vektor **g**. U jednostranné alternativy se pak zvýší při výpočtu p-hodnoty čitatel i jmenovatel o jedničku.

6.3 Bootstrapové testování hypotéz

Bootstrapové testování hypotéz je do jisté míry zobecněním permutačních testů. Permutační test vyžaduje symetrii uspořádaného vektoru ***(v,g)*** za předpokladu platnosti nulové hypotézy. Tato symetrie je splněna, když pochází oba výběry sloučeného vektoru **v** ze stejného rozdělení. Metoda bootstrap tuto symetrii nevyžaduje a tedy mohou jí být testovány obecnější problémy. V následujícím textu budu značit sloučený vektor více výběrů **x** namísto **v**. Nicméně se principiálně jedná o ten samý vektor jako v kapitole č. *6.2*.

Pro obecné vysvětlení opět předpokládám hypotézu s rovností dvou rozdělení a testovou statistiku jako rozdíl výběrových průměrů. V návaznosti na vzorce č. 6.2.1, 6.2.2 a 6.2.3

$\boldsymbol{x} = \{y_1, y_2, \dots, y_n, z_1, z_2, \dots, z_m\}$

$t(\boldsymbol{x}) = f(\boldsymbol{y}, \boldsymbol{z})$

Algoritmus testu je následující.

1. Je taženo *B* bootstrapových výběrů z **x.** prvních *n* pozorováních tohoto výběru je označeno jako $\boldsymbol{y^{*b}}$ a zbylých *m* je označeno jako $\boldsymbol{z^{*b}}$.
2. Je určena bootstrapová testová statistika.

$$\hat{\theta}^*(b) = t(\boldsymbol{x}_b^*) = f(\boldsymbol{y_b^*}, \boldsymbol{z_b^*}) \quad b \in \{1, 2, \dots, B\} \qquad 6.3.1$$

$$\hat{\boldsymbol{\theta}}^* = \{\hat{\theta}^*(1), \hat{\theta}^*(2), \dots, \hat{\theta}^*(B)\} \qquad 6.3.2$$

3. Je aproximována p-hodnota.

$$\hat{p}_{boot} = \frac{\#[\hat{\boldsymbol{\theta}}^* \geq t(\boldsymbol{x})]}{B} \qquad 6.3.3$$

Nicméně výše uvedený algoritmus opět zahrnuje do hypotézy, že rozdělení obou vzorků jsou stejná, což znamená, že například je-li testována rovnost dvou středních hodnot, může být hypotéza zamítnuta i za předpokladů rovnosti obou středních hodnot, kvůli rozdílným rozptylům. Nicméně jak jsem uvedl v úvodu, metoda bootstrap tuto symetrii nevyžaduje. (Effron & Timbishirani, 1993)

6.3.1 Dvouvýběrový bootstrapový test o střední hodnotě

Za platnosti výše uvedených vztahů a hypotézy rovnosti středních hodnot obou výběrů, oproti dvoustranné alternativě,

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

Jsou vypočteny hodnoty $\tilde{z}_i$ a $\tilde{y}_i$, kde

$$\tilde{y}_i = y_i - (\bar{y} - \bar{x}) \qquad 6.3.4$$

$$\tilde{z}_i = z_i - (\bar{z} - \bar{x}) \qquad 6.3.5$$

$\bar{y}$ ve vzorci č. *6.3.4* a $\bar{z}$ v 6.3.5 jsou skupinové průměry vektorů **y** a **z**, $\bar{x}$ vyjadřuje průměr sloučeného vektoru **x**. Bootstrapový test rovnosti dvou skupinových průměrů pak probíhá ve třech krocích. (Effron & Timbishirani, 1993)

1. Je provedeno *B* replikací dvou na sobě nezávislých bootstrapových výběrů. V prvním výběru je taženo *n* pozorování s vracením z vektoru $\tilde{\mathbf{y}}$, v druhém je taženo *m* pozorování s vracením z vektoru $\tilde{\mathbf{z}}$.

$$\hat{F} \rightarrow (\tilde{y}_1^*, \tilde{y}_2^*, \ldots, \tilde{y}_n^*) = \mathbf{y}_b^* \quad b \in \{1, 2, \ldots, B\} \qquad 6.3.6$$

$$\hat{G} \rightarrow (\tilde{z}_1^*, \tilde{z}_2^*, \ldots, \tilde{z}_m^*) = \mathbf{z}_b^* \quad b \in \{1, 2, \ldots, B\} \qquad 6.3.7$$

2. Je vypočtena testová statistika založená na bootstrapových výběrech z kroku 1. Zde $\bar{y}_b^*$ značí průměr vektoru $\mathbf{y}_b^*$

$$\hat{\theta}^*(b) = t(\boldsymbol{x}_b^*) = \frac{\bar{y}_b^* - \bar{z}_b^*}{\sqrt{\frac{\sigma_{b1}^{2*}}{n} + \frac{\sigma_{b2}^{2*}}{m}}} \quad b \in \{1, 2, \ldots, B\} \qquad 6.3.8$$

$$\hat{\boldsymbol{\theta}}^* = \{\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)\} \qquad 6.3.9$$

$$\bar{y}_b^* = \frac{\sum_{i=1}^{n} \tilde{y}_{bi}^*}{n} \quad b \in \{1, 2, \ldots, B\} \qquad 6.3.10$$

$$\bar{z}_b^* = \frac{\sum_{i=1}^{n} \tilde{z}_{bi}^*}{m} \quad b \in \{1, 2, \ldots, B\} \qquad 6.3.11$$

$$\sigma_{b1}^{2*} = \frac{\sum_{i=1}^{n} (\tilde{y}_{bi}^* - \bar{y}_b^*)^2}{n-1} \quad b \in \{1, 2, \ldots, B\} \qquad 6.3.12$$

$$\sigma_{b2}^{2*} = \frac{\sum_{i=1}^{m} (z_{bi}^* - \bar{z}_b^*)^2}{m-1} \quad b \in \{1, 2, \ldots, B\} \qquad 6.3.13$$

3. V posledním kroku je aproximována p-hodnota vzhledem k pravostranné alternativě,

$$\hat{p} = \frac{\#[\hat{\boldsymbol{\theta}}^* \geq t(\boldsymbol{x})]}{B} \qquad 6.3.14$$

6.3.2 Jednovýběrový bootstrapový test o rovnosti střední hodnoty

Jednovýběrový bootstrapový test funguje na podobném principu jako dvouvýběrový test. Testuje se rovnost střední hodnoty, neznámé populac, s neznámou distribuční funkcí *F* konstantě μ_0, oproti dvoustranné alternativě.

$H_0: \mu_1 - \mu_0 = 0$

$H_1: \mu_1 - \mu_0 \neq 0$

Nechť je vzorek $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ náhodně vybrán z neznámé populace. Vypočte se vektor $\tilde{\boldsymbol{x}}$ dle vzorce č. *6.3.3*. Jednovýběrový bootstrapový test pak probíhá v následujících krocích. (Effron & Timbishirani, 1993)

1. Je vybráno *B* bootstrapových výběrů vektoru $\tilde{\boldsymbol{x}}$
2. Je vypočtena bootstrapová testovací statistika. $\overline{x_b^*}$ a σ_b^{2*} se vypočítají analogicky dle vzorců č. 6.3.10 a 6.3.12.

$$\hat{\theta}^*(b) = t(\boldsymbol{x}_b^*) = \frac{\overline{x_b^*} - \mu_0}{\sqrt{\frac{\sigma_b^{2*}}{n}}} \qquad 6.3.15$$

$$\hat{\boldsymbol{\theta}}^* = \{\hat{\theta}^*(1), \hat{\theta}^*(2), \dots, \hat{\theta}^*(B)\} \qquad 6.3.16$$

3. V posledním kroku je opět odhadnuta p-hodnota analogicky dle vzorce č. 6.3.14. Pouze je tento vzorec opraven o oboustrannou alternativu.

$$\hat{p} = \frac{\#[|\hat{\boldsymbol{\theta}}^*| > |t(\boldsymbol{x})|]}{B} \qquad 6.3.17$$

7 Závěr

John Tukey kdysi řekl, že přibližná odpověď na přesnou otázku jest lepší, než přesná odpověď na přibližnou otázku. Celý život lidé hledají co nejpřesnější odpovědi a nezeptají se sami sebe, není-li jejich otázka pouhou aproximací skutečně řešeného problému. Stejně tak i ve statistice existuje celá škála možností, jak řešit určitý problém. Každá z těchto metod má určitou přesnost. Stejně jako v životě není cílem hledat nejpřesnější, nýbrž nejsprávnější řešení. V situacích kdy není vhodné užít klasický přístup, neboť jsou porušeny podmínky jeho užití, je bootstrap přibližnou odpovědí na přesnější otázku.

V této práci jsem shrnul základní užití metody bootstrap. Uvedl jsem v jakých případech by tato metoda měla být použita, a porovnal jsem její vlastnosti s klasickým přístupem. Tato práce je pouhým nástinem mnohem složitější problematiky. Celkový potencionál bootstrapových odhadů se neomezuje pouze na případy zde uvedené. Metodu bootstrap lze aplikovat na lineární, nelineární, parametrické i neparametrické modely, časové řady. Jsou známy i její aplikace v oblasti bayesovské statistiky. Popularita této metody velmi roste. Bootstrap je v dnešní době součástí většiny statistických softwarů jako jsou SPSS, R, SAS a Eviews. Stále je mnoho otázek v této oblasti nezodpovězených. Jejich výčet by zahrnoval mnohem hlubší a detailnější studii daného problému. V neposlední řadě bych chtěl podotknout, že nejen vývoj rychlosti procesorů měl vliv na rozvoj různých vědních oborů, ale především internet má velkou zásluhu na tempu vývoje různých vědních disciplín díky rychlosti a lehkosti, s kterou dokáží být v dnešní době díky němu informace šířeny.

Seznam Grafů

Graf číslo 1 – Konvergence empirické distribuční funkce 8
Graf číslo 2 – Jackknifový odhad směrodatné odchylky odhadu 11
Graf číslo 3 – Bootstrapový odhad směrodatné odchylky odhadu 14
Graf číslo 4 – Bootstrapový odhad vychýlení výběrového průměru pro různá *B* 16
Graf číslo 5 – Různé metody odhadu parametru *b* rovnoměrného rozdělení 20
Graf číslo 6 – Rozdělení $\boldsymbol{\theta}*$ pro výběrový průměr 24
Graf číslo 7 - Rozdělení $\boldsymbol{\theta}*$ pro výběrový rozptyl 25
Graf číslo 8 - Rozdělení $\boldsymbol{\theta}*$ pro mezikvantilového průměru 30
Graf číslo 9 - Rozdělení $\boldsymbol{\theta}*$ pro býběrový medián 31
Graf číslo 10 - Rozdělení $\boldsymbol{\theta}*$ pro rozdíl průměrů 37
Graf číslo 11 – Graf dvou normálních rozdělení s různou střední hodnotou 38

Seznam výstupů

Výstup číslo 1 – Intervaly spolehlivosti pro odhad střední hodnoty _____ 24
Výstup číslo 2 – Intervaly spolehlivosti pro odhad rozptylu _____ 25
Výstup číslo 3 – Intervalové odhady střední hodnoty _____ 28
Výstup číslo 4 – Robustní intervalový odhad střední hodnoty _____ 29
Výstup číslo 5 – Interval spolehlivosti pro odhad mediánu _____ 31
Výstup číslo 6 – Testová rozhodnutí a jejich podmíněné pravděpodobnosti _____ 33

Bibliografie

- Bartoszyński, R., & Niewiadomska-Bugaj, M. (2008). *Probability and Statistical Inference.* New Jersey: John Wiley & Sons Inc.
- Effron, B., & Timbishirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.
- Hall, P. (1989). *Antithetic Resampling for the Bootstrap.* Biometrika.
- Hall, P. (1989). *On efficient bootstrap simulation.* Biometrika.
- Hartigan, J. (1969). *Using subsample values as typical values.* Journal of the American Statistical Association.
- Hebák, P., Bílková, D., & Svobodová, A. (2009). *Praktikum k výuce matematické statistiky II.* Praha: Nakladatelství Oeconomica.
- kol. autorů. (1992). *Exploring the limits of bootstrap.* New York: Wiley-Interscience Publication.
- Politis, D. N., & Romano, J. P. (1993). *Estimating the distribution of a studentized statistic by subsampling.* Bulletine of international statistical institution.
- R Core Team. (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria.
- Rice, K., & Thomas, L. (2008). Permutation Test. *UW Biostatistics*, Seatle.
- Ripley, B. D., & Venables, W. N. (2002). Modern Applied Statistics with S. New York, USA: Springers.
- Shao, J., & Tu, D. (1995). *The Jackknife and Bootstrap.* New York: Springers.
- Tukey, J. (1959). *Bias and confidence in not quite large samples.* Anals of Mathematical Statistics.

Příloha A: Kód k vytvoření grafu číslo 1

```
> par(mfrow=c(1,3))
> x=rnorm(3000)
> x=sort(x)
> n30=rnorm(30)
> n100=rnorm(100)
> n1000=rnorm(1000)
> plot(x,pnorm(x),type="l",ylab="F(x)",xlab="x",main=" n=30")
> lines(ecdf(n30),col="blue")
> plot(x,pnorm(x),type="l",ylab="F(x)",xlab="x",main=" n=100")
> lines(ecdf(n100),col="blue")
> plot(x,pnorm(x),type="l",ylab="F(x)",xlab="x",main=" n=1000")
> lines(ecdf(n1000),col="blue")
```

Příloha B: Kód k vytvoření grafu číslo 3

```
> par(mfrow=c(1,2))
> n=seq(10,1000,by=20)
> b1=numeric(length(n))
> b2=numeric(length(n))
> mc=numeric(length(n))
> for (i in 1:length(n))
> {
> x=rnorm(n[i])
> b1[i]=bootstrap_se(x,2,200)$std.err
> b2[i]=bootstrap_se(x,1,200)$std.err
> mc[i]=sqrt((n[i]/(n[i]-1))*var(replicate(10000,median(rnorm(n[i])))))
> }
> plot(b1~n,type="l",ylab="směrodatná odchylka",main="směrodatná odchylka
  průměru",col="blue")
> curve(sqrt(1/x),add=T,col="red")
> plot(b2~n,type="l",ylab="směrodatná odchylka",main="směrodatná odchylka
  mediánu",col="blue")
> lines(mc~n,type="l",col="red")
```

Zde použitá funkce *bootstrap_se* je mnou vytvořený skript sloužící k výpočtu bootstrapového odhadu směrodatné odchylky odhadu. Vektor nazvaný *mc* zde ukládá hodnoty rozptylu odhadu vypočítaných za pomocí Monte Carlo simulace. Argument funkce *curve* zde představuje analytickou funkci směrodatné odchylky výběrového průměru s tím, že v tomto případě je nám rozptyl a je roven jedné.

Příloha C: Kód k simulaci v příkladu číslo 5

```
> kvantilovy=0
> bc=0
> n=30
> for (i in 1:1000)
> {
> x=sample(dny,30)
> b=bootstrap_t_median(x,1200)
> u_nula=qnorm(sum(b$boot.gen<median(x))/1200)
> alfa1=pnorm(u_nula+(u_nula+qnorm(0.025)))
> alfa2=pnorm(u_nula+(u_nula+qnorm(0.975)))
> dk=quantile(b$boot.gen,0.025)
> hk=quantile(b$boot.gen,0.975)
> dbc=quantile(b$boot.gen,alfa1)
> hbc=quantile(b$boot.gen,alfa2)
> if (median(dny)<dk | median(dny)>hk)
> {
> kvantilovy=kvantilovy+1
> }
> if (median(dny)<dbc | median(dny)>hbc)
> {
> bc=bc+1
> }
> }
```

Je zde užit mnou napsaný skript *bootstrap_t_median* , který počítá bootstrapový-t intervalový odhad i kvantilový intervalový odhad. Tato simulace nebyla zaměřena na bootstrapový-t interval, proto jsem v rámci urychlení simulace tuto funkci uvnitř skriptu zakázal. Vektor "b$boot.gen" odkazuje na simulací vytvořený vektor $\widehat{\boldsymbol{\theta}}^*$. V každém cyklu je vypočten kvantilový intervalový odhad a na základě těch samých dat i BC_a intervalový odhad. Následně je s oběma intervaly porovnána skutečná hodnotou mediánu základního souboru. Leží-li tato hodnota mimo tento interval, je přičtena do příslušné proměnné jednička.

Příloha D: Kód k vytvoření příkladu číslo 6

```
> y=sleep[sleep$group==1,1]
> z=sleep[sleep$group==2,1]
> n=length(y)
> m=length(z)
> g=rep(c(0,1),c(n,m))
> rozdil.prumeru=mean(z)-mean(y)
> permutation.test=function(g,v)
> {
> permutace=sample(g)
> mean(v[permutace==1])-mean(v[permutace==0])
> }
> replikace=replicate(1000,permutation.test(g,v))
> p.hodnota=mean(abs(replikace)>abs(rozdil.prumeru))
```