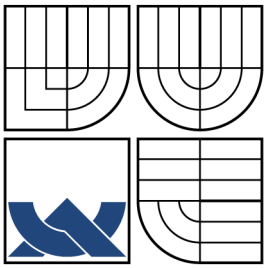
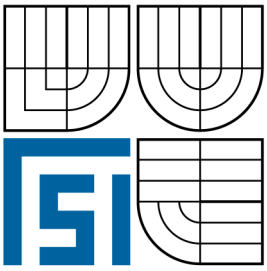


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
DEPARTMENT OF MATHEMATICS

FITOVÁNÍ ROZDĚLENÍ PRAVDĚPODOBNOSTI PRO APLIKACE

FITTING OF PROBABILITY DISTRIBUTIONS FOR APPLICATIONS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. LENKA PAVLÍČKOVÁ
AUTHOR

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

BRNO 2012

Vysoké učení technické v Brně, Fakulta strojního inženýrství

Ústav matematiky
Akademický rok: 2011/2012

ZADÁNÍ DIPLOMOVÉ PRÁCE

student(ka): Bc. Lenka Pavlíčková

který/která studuje v **magisterském navazujícím 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:

Fitování rozdělení pravděpodobnosti pro aplikace

v anglickém jazyce:

Fitting of Probability Distributions for Applications

Stručná charakteristika problematiky úkolu:

Studium moderních efektivních metod odhadů parametrů a rozdělení pravděpodobnosti pomocí bootstrapu z pozorovaných hodnot náhodných veličin, náhodných vektorů a kategoriálních veličin s ohledem na aspekty jejich aplikací v technických a dalších oborech.

Cíle diplomové práce:

Popis, zhodnocení a rozvoj současných efektivních statistických metod odhadů parametrů a rozdělení pravděpodobnosti pomocí bootstrapu respektujících omezení a neurčitost dat, jejich realizace na PC a aplikace na konkrétních datových souborech.

Seznam odborné literatury:

1. Montgomery, D. C., Renger, G.: Probability and Statistics. New York: John Wiley & Sons, 1996.
2. Anděl, J.: Statistické metody. Praha: MATFYZPRESS, 2003.
3. Anděl, J.: Základy matematické statistiky. Praha: MATFYZPRESS, 2002.
4. Silverman, B. W.: Density Estimation for Statistics and Data Analysis. London: Chapman and Hall,1985.
5. Vajda, I.: Theory of Statistical Inference and Information. London: Kluwer Academic Press, 1989.
6. Scott, D.W.: Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley, 1992.
7. Články a materiály z odborných časopisů, sborníků konferencí a Internetu dle pokynů vedoucího diplomové práce.

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

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2011/2012.

V Brně, dne 18.11.2010

L.S.

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

ABSTRAKT

Diplomová práce popisuje metodu bootstrap a její použití pro tvorbu konfidenčních intervalů, při testování statistických hypotéz a v regresní analýze. Představujeme konfidenční interval pro individuální hodnotu. Dále se zaobíráme metodou odhadu diskrétního rozdělení pravděpodobnosti kategoriální veličiny pomocí gradientního a přímkového odhadu.

KLÍČOVÁ SLOVA

bootstrap, odhad parametru, konfidenční interval, test statistické hypotézy, regresní analýza, individuální hodnota, f-divergence, kvazinorma, diskrétní rozdělení pravděpodobnosti, gradientní odhad, přímkový odhad

ABSTRACT

The diploma thesis describes the bootstrap method and its applications in the confidence intervals generation, in the testing of statistical hypotheses and in the regression analysis. We present the confidence interval for individual value. Further the method of discrete probability estimation of the categorical quantity is presented, making use the gradient and the line estimate.

KEYWORDS

bootstrap, parameter estimate, confidence interval, statistical hypothesis testing, individual value, f-divergence, quasi-norm, discrete probability distribution, gradient estimate, line estimate

PAVLÍČKOVÁ, Lenka *Fitování rozdělení pravděpodobnosti pro aplikace*: diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, Ústav matematiky, 2012. 72 s. Vedoucí práce byl doc. RNDr. Zdeněk Karpíšek, CSc.

Prohlašuji, že svou diplomovou práci na téma „Fitování rozdělení pravděpodobnosti pro aplikace“ jsem vypracovala samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Děkuji vedoucímu mé diplomové práce panu doc. RNDr. Zdeňku Karpíškovi CSc.
za pomoc a cenné rady, které mi pomohly při práci na daném tématu.

OBSAH

1 ÚVOD

Výraz bootstrap v doslovném překladu znamená poutko u bot. Název pochází z legendy o baronovi Münchhausenovi od autora Ericha Raspeho, která vypráví, že se jednou baron pomalu topil v blátě a zachránil se zatažením za šňůrky u svých bot, což by žádný z tonoucích ke své záchraně neudělal.

Základní principy metody bootstrap poprvé popsal Brad Efron roku 1979 v článku *Bootstrap Methods: Another look at the jackknife*. Článek vzbudil velký ohlas a metoda dokázala, že svou přesností předčí i klasickou aproximaci rozdělením. Metoda bootstrap přinesla možnost odhadnout přesnost libovolného odhadu libovolného parametru. Princip metody spočívá v jednoduché myšlence mnohonásobného opakování jednoduchého algoritmu. Metoda bootstrap je použitelná pro výběry s malým rozsahem, protože není závislá na centrální limitní větě.

Diplomová práce je rozdělena do šesti celků. V této diplomové práci popisujeme, jak metoda bootstrap pracuje. Ukážeme si, jak můžeme s přispěním metody bootstrap zkonstruovat konfidenční intervaly, uvádíme více přístupů a obohatíme to i praktickou aplikací hledání konfidenčních intervalů. Předvedeme si, jak lze testovat pomocí metody bootstrap statistické hypotézy. Podíváme se i na regresní analýzu a spojení s metodou bootstrap, kde se seznámíme i s praktickou aplikací metody bootstrap a regresní analýzy, v této praktické části jsme více experimentovali a snažili se i najít rozdělení pravděpodobnosti, které fituje danou veličinu. Popíšeme si, jak budeme přistupovat ke konstrukci konfidenčního intervalu pro individuální hodnotu a v poslední kapitole diplomové práce jsme se zabývali kategoriální veličinou ve spojení s metodou bootstrap. Otestovali jsme pesimistický přímkový odhad ve spojení s metodou bootstrap na reálných datech.

Praktické aplikace jsme prováděli v MS Excel a ve statistickém softwaru Statgraphic Centurion.

2 BOOTSTRAP

Výsledky v této kapitole jsou podloženy [8], [9], [10].

Základem metody bootstrap je opakovaná realizace výběru z naměřených dat nebo odhadnutého modelu.

Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé stejně rozdělené *(iid) náhodné veličiny* a nechť F je distribuční funkce, která je blíže nespecifikovaná. Nechť $\theta = \theta(F)$ je nezávislý parametr rozdělení pravděpodobnosti náhodné veličiny X, který má být odhadnut na základě realizace náhodného výběru. Parametr θ může být střední hodnota, variance nebo jiné charakteristiky rozdělení pravděpodobnosti F. Pak realizujeme *náhodný výběr*

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)$$

z náhodné veličiny X o rozsahu n, nechť

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

je realizace (soubor pozorovaných hodnot) náhodného výběru $\mathbf{X}$. Na základě realizace náhodného výběru $\mathbf{X}$ vypočítáme odhad parametru θ. Odhad θ označme $\widehat{\theta}$,

$$\widehat{\theta} = \widehat{\theta}(x_1, x_2, \ldots, x_n).$$

Nechť

$$T = T(X_1, X_2, \ldots, X_n)$$

je statistika pro odhad parametru θ a

$$R = R(X_1, X_2, \ldots, X_n)$$

je její vhodně standardizovaná verze.
Nechť

$$H(x) = P[R(X_1, X_2, \ldots, X_n, F) \leq x]$$

vyjadřuje distribuční funkci statistiky R.
Výpočet rozdělení H může být komplikované a to i v případě známé distribuční funkce F. Pokud známe distribuční funkci F, tak lze použít metodu Monte Carlo:

- generovat dlouhou sérii nezávislých náhodných výběrů z rozdělení s danou distribuční funkcí,
- spočítat pro každé opakování hodnotu příslušné charakteristiky,
- skutečné rozdělení charakteristiky aproximovat empirickým rozdělením získaným z řady uměle získaných hodnot.

Pokud distribuční funkci F neznáme, což je častější situace, tak H aproximujeme asymptotickým rozdělením (odvozené na základě limitních vět teorie pravděpodobnosti). Přesnost aproximace při neznáme distribuční funkci F je ovlivněna a omezena počtem pozorování.
Metoda bootstrap kombinuje *substituční princip* a *metodu Monte Carlo*.

Substituční princip

Nechť $F(x)$ je nějaký odhad distribuční funkce, nejčastěji empirická distribuční funkce založená na náhodném výběru $X_1, X_2, \ldots, X_n$, tj.

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} I[X_i \leq x],$$

kde $I[A]$ označuje indikátor množiny A. Při daných hodnotách $X_1, X_2, \ldots, X_n$ je F známá funkce.
Nechť $\mathbf{X}^* = ({X_1}^*, X_2^*, \ldots, X_n^*)$ je nezávislý náhodný výběr z F, tj. při daných pozorováních x_i jsou X_i^* iid náhodné veličiny a každá nabývá hodnot x_i s pravděpodobností $p = \frac{1}{n}$.
Bootstrapový výběr je soubor

$$\mathbf{X}^* = ({X_1}^*, X_2^*, \ldots, X_n^*).$$

Dále se původní výběr $\mathbf{X}$ nahradí bootstrapovým výběrem $\mathbf{X}^*$, neznámou distribuční funkci F nahradíme známou distribuční funkcí F. Pak dostaneme parametr

$$\theta^* = \theta(F)$$

a statistiky

$$T^* = T({X_1}^*, X_2^*, \ldots, X_n^*)$$

a standardizované verze statistik

$$R^* = R({X_1}^*, X_2^*, \ldots, X_n^*, F).$$

Pak můžeme definovat *teoretické charakteristiky*

$$E^*T^* = \int T(x_1, x_2, \ldots, x_n)\, d(F(x_1), F(x_2), \ldots, F(x_n)),$$

$$var^*T^* = \int [T(x_1, x_2, \ldots, x_n) - E^*T^*]^2\, d(F(x_1), F(x_2), \ldots, F(x_n))$$

a *teoretickou distribuční funkci*

$$H^*(x) = P^*(R({X_1}^*, X_2^*, \ldots, X_n^*, F) \leq x) = \\ = P(R({X_1}^*, X_2^*, \ldots, X_n^*, F) \leq x | X_1, X_2, \ldots, X_n).$$

Tyto teoretické charakteristiky a distribuční funkce jsou získané metodou bootstrap a v praxi se využijí, pokud jsou explicitními funkcemi pozorování $X_1, X_2, \ldots, X_n$. Pokud bychom chtěli přesně určit bootstrapové rozdělení, tak by se provedlo všech n^n výběrů s vracením z populace pozorovaných hodnot $x_1, x_2, \ldots, x_n$. Toto je možné provést jen pro malé n.
Proto se na bootstrapový výběr $\mathbf{X}^*$ a známou distribuční funkci F nejčastěji aplikuje metoda Monte Carlo.

Metoda Monte Carlo

Jedná se o metodu, kdy se generuje mnohokrát (B-krát) nezávislý náhodný výběr o rozsahu n z rozdělení F. Pravděpodobnost, že vybereme prvek metodou Monte Carlo je $\frac{1}{n}$. V literatuře studující statistiku se tento termín označuje jako výběr z původní množiny s opakováním. Prvek x_i z původní množiny se může ve výběru generovaném metodou Monte Carlo objevit jedenkrát, dvakrát, ale i n-krát. Pravděpodobnost rozdělení, že prvek x_i se vyskytne n-krát ve výběru, je blízké Poissonovu rozdělení se střední hodnotou rovnou jedné. Předpokládejme, že x_i se v generovaném výběru metodou Monte Carlo vyskytne n_i-krát. Pak každá bootstrapová množina obsahuje přesně n prvků,

$$\sum_{i=1}^{n} n_i = n.$$

Podívejme se na případ, kdy budeme uvažovat jednu proměnnou x_i. Pravděpodobnost výskytu x_i v jednom bootstrapovém výběru je $\frac{1}{n}$, tuto pravděpodobnost označíme p. Pak pravděpodobnost, že x_i se vyskytne n_i-krát, je

$$P(n_i) = \frac{n!}{n_i!(n-n_i)!} p^{n_i}(1-p)^{n-n_i}.$$

Pokud máme vygenerované náhodné výběry, tak pro každou realizaci náhodného výběru spočítáme hodnoty T^* a R^* a z nich se pak stanoví aritmetický průměr. Tak se získají *bootstrapové odhady* původního rozdělení a původních charakteristik. Tedy bootstrapový odhad rozptylu T získáme tak, že se B-krát opakuje nezávislý náhodný výběr z F a vždy se spočte hodnota statistiky T^*.Postupně se získají hodnoty

$$T_1^*, T_2^*, \ldots, T_B^*,$$

ze kterých se spočte

$$\widehat{var}^* T^* = \frac{1}{B} \sum_{b=1}^{B} \left(T_b^* - \frac{1}{B} \sum_{k=1}^{B} T_k^* \right)^2 .$$

Obdobně odhadneme distribuční funkci statistiky R,

$$\widehat{H}^*(x) = \frac{1}{B} \sum_{b=1}^{B} I \left\{ R \left(X_{1,b}^*, X_{2,b}^*, \ldots, X_{n,b}^*, F \right) \leq x \right\},$$

kde $\left\{ X_{1,b}^*, X_{2,b}^*, \ldots, X_{n,b}^* \right\}, b = 1, 2, \ldots, B$ jsou nezávislé výběry z F.

3 KONFIDENČNÍ INTERVALY

Teoretické výsledky jsou podloženy [8], [9], [10], [12].

Mezi základní úlohy matematické statistiky patří úloha stanovení hodnot parametrů rozdělení, ze kterého máme k dispozici náhodný výběr. Nejčastěji se zabýváme dvěma druhy odhadů:

- *bodový odhad*, který je odhadem parametru pomocí statistiky (funkce náhodného výběru), jejíž hodnotu pro datový soubor považujeme za hledanou hodnotu neznámého parametru rozdělení (nebo jeho funkce),
- *intervalový odhad* (*konfidenční interval*) je interval, ve kterém se hodnota neznámého parametru vyskytuje s požadovanou pravděpodobností.

Podívejme se na hledání intervalového odhadu. Uvažujme, že θ je neznámý parametr zkoumaného rozdělení a $\tau(\theta)$ je funkce parametru θ, kterou odhadujeme, pak hledáme statistiky T_D a T_H takové, že pro koeficient $(1-\alpha)$ platí

$$P(T_D \leq \tau(\theta) \leq T_H) = 1 - \alpha$$

a navíc vyžadujeme

$$P(\tau(\theta) < T_D) = P(\tau(\theta) > T_H) = \frac{\alpha}{2}.$$

Pak intervalovým odhadem funkce $\tau(\theta)$ je interval (T_D, T_H). V tomto případě mluvíme o *oboustranném odhadu*.

Někdy ovšem potřebujeme jen *jednostranné odhady*. Pak dostaneme

$$\tau(\theta) \in (T_D, \infty), \text{ kde } P(\tau(\theta) \geq T_D) = 1 - \alpha \text{ a } P(\tau(\theta) < T_D) = \frac{\alpha}{2};$$
$$\tau(\theta) \in (-\infty, T_H), \text{ kde } P(\tau(\theta) \leq T_H) = 1 - \alpha \text{ a } P(\tau(\theta) > T_H) = \frac{\alpha}{2}.$$

α obvykle volíme

$$\alpha = 0,1;\ 0,05;\ 0,01.$$

Spolehlivost odhadu je pak

$$1 - \alpha = 0,9;\ 0,95;\ 0,99.$$

Tedy v

$$90\%\ ; 95\%\ ; 99\%$$

je náš odhad pro parametr správný.

3.1 Intervalové odhady parametrů normálního rozdělení

Odhad parametru μ rozdělení $N(\mu, \sigma^2)$ při známém rozptylu σ^2

Použijeme statistiku $\overline{X}$ (výběrový průměr) jako jeho odhad. Víme, že náhodná veličina

$$U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$$

má normované normální rozdělení $N(0, 1)$. Pak

$$P(|U| \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha \Leftrightarrow -u_{1-\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \leq u_{1-\frac{\alpha}{2}}.$$

Symbolem $u_{1-\frac{\alpha}{2}}$, $0 < u_{1-\frac{\alpha}{2}} < 1$ označujeme $(1-\frac{\alpha}{2})$-kvantil normovaného normálního rozdělení $N(0, 1)$. Odtud pak dostaneme, že

$$T_D = \overline{X} - \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}} = T_H.$$

Jednostranné odhady jsou pak levostranný interval

$$\mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha} = T_H$$

a pravostranný interval

$$\mu \geq \overline{X}\frac{\sigma}{\sqrt{n}}u_{1-\alpha} = T_D.$$

Odhad parametru σ^2 při známé střední hodnotě μ

V tomto případě využijeme skutečnosti, že náhodná veličina

$$U_i = \frac{X_i - \mu}{\sigma}$$

má normované normální rozdělení $N(0, 1)$. Pak náhodná veličina

$$V = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

má rozdělení $\chi^2(n)$. Pak

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sigma^2}{n}V$$

a statistika $V = \frac{ns^2}{\sigma^2}$ má rozdělení $\chi^2(n)$. Pro oboustranný odhad dostaneme

$$P(v_1 \leq V \leq v_2) = 1 - \alpha \Rightarrow v_1 = \chi^2_{\frac{\alpha}{2}}(n) \text{ a } v_2 = \chi^2_{1-\frac{\alpha}{2}}(n).$$

Symbolem $\chi^2_{\frac{\alpha}{2}}(n)$ označujeme $\frac{\alpha}{2}$-kvantil rozdělení $\chi^2(n)$ a $\chi^2_{1-\frac{\alpha}{2}}(n)$ označuje $(1-\frac{\alpha}{2})$-kvantil rozdělení $\chi^2(n)$. Odtud odvodíme odhad pro σ^2

$$\chi^2_{\frac{\alpha}{2}}(n) \leq \frac{ns^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}(n)$$
$$\frac{ns^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{\frac{\alpha}{2}}(n)}$$

Obdobně dostaneme jednostranné odhady, levostranný interval

$$\frac{ns^2}{\chi^2_{1-\alpha}(n)} \geq \sigma^2,$$

pravostranný interval

$$\sigma^2 \leq \frac{ns^2}{\chi^2_{\alpha}(n)}.$$

Odhad střední hodnoty μ za podmínky, že rozptyl σ^2 uvažovaného rozdělení není znám

Pro určení intervalu spolehlivosti použijeme statistiku

$$T = \frac{\overline{X} - \mu}{S}\sqrt{n},$$

o které víme, že má Studentovo t-rozdělení $t(n-1)$ o $(n-1)$ stupních volnosti.

$$T = \frac{\frac{\overline{X}-\mu}{\sigma}\sqrt{n}}{\frac{S}{\sigma}}$$

a

$$U = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \chi^2(n-1),$$

neboť $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$. Dále

$$Z = (n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{\overline{X} - X_i}{\sigma}\right)^2 \sim \chi^2(n-1),$$

a

$$T = \frac{U}{\sqrt{\frac{Z}{n-1}}}$$

má Studentovo rozdělení $t(n-1)$.

Interval spolehlivosti určíme z podmínky

$$P(|T| \leq t_{1-\frac{\alpha}{2}}(n-1)) = 1-\alpha.$$

Odtud

$$-t_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{S}\sqrt{n} \leq t_{1-\frac{\alpha}{2}}$$
$$\overline{X} - \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}$$

je oboustranný interval spolehlivosti pro parametru μ.

Obdobně dostaneme jednostranné intervaly, levostranný interval

$$\mu \geq \overline{X} - \frac{S}{\sqrt{n}}t_{1-\alpha}$$

a pravostranný interval

$$\mu \leq \overline{X} + \frac{S}{\sqrt{n}}t_{1-\alpha},$$

kde symbolem $t_{1-\alpha}$ označujeme $(1-\alpha)$-kvantil uvažovaného rozdělení.

Odhad parametru σ^2 při neznámé střední hodnotě μ

Pro určení intervalu spolehlivosti použijeme statistiku

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

která má rozdělení $\chi^2(n-1)$ a dále vycházíme ze skutečnosti, že pro statistiku S^2 je $E(S^2) = \sigma^2$ a může tedy sloužit jako vhodný odhad parametru σ^2. Oboustranný interval spolehlivosti dostaneme z podmínky

$$P(v_1 \leq Y \leq v_2) = 1-\alpha \Rightarrow v_1 = \chi^2_{\frac{\alpha}{2}}(n-1) \text{ a } v_2 = \chi^2_{1-\frac{\alpha}{2}}(n-1).$$

Odtud plyne pro oboustranný interval spolehlivosti

$$v_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq v_2 \Rightarrow \frac{n-1}{v_2}S^2 \leq \sigma^2 \leq \frac{(n-1)}{v_1}S^2.$$

Jednoduchou úpravou získáme jednostranné intervaly spolehlivosti, levostranný interval

$$\frac{(n-1)}{v_2}S^2 \leq \sigma^2$$

a pravostranný interval

$$\sigma^2 \leq \frac{(n-1)}{v_1}S^2,$$

kde v_1 je $(\chi^2_\alpha(n-1))$-kvantil a v_2 je $(\chi^2_{1-\alpha}(n-1))$-kvantil rozdělení $\chi^2(n-1)$ o $n-1$ stupních volnosti.

3.2 Konfidenční intervaly a metoda bootstrap

Určování konfidenčních intervalů metodou bootstrap je založeno na myšlence, že pokud ze základního souboru, který obsahuje hodnotu θ_0 zjišťovaného parametru θ, získáme (měřením, pokusy, ...) náhodný výběr $x_1, \ldots, x_n$, pro který bude mít vypočtený parametr hodnotu θ, tak tento parametr θ je od parametru θ_0 vzdálen o hodnotu $\Delta\theta = \theta - \theta_0$. Naopak pokud by soubor $x_1, \ldots, x_n$ představoval základní soubor daného parametru θ o střední hodnotě μ_{θ^*}, tak některý z náhodných výběrů, vytvořený z tohoto souboru bude mít parametr θ^* vzdálen od μ_{θ^*} také o $\Delta\theta$. Rozdělení bootstrap parametru θ^* bude s jistou pravděpodobností obsahovat také hodnotu skutečného parametru θ_0.

Dál si ukážeme, jak se konfidenční intervaly konstruují.

Předpokládejme, že jsme získali n nezávislých hodnot $x_1, \ldots, x_n$, ze kterých spočítáme neznámý parametr θ, $\theta = \theta(F)$, kde F je neznámé rozdělení pravděpodobnosti. Parametr θ může být výběrový průměr, směrodatná odchylka, hodnota nějakého kvantilu, ... Nechť $\widehat{\theta} = \theta(\widehat{F})$ je odhad parametru θ, dále nechť $\widehat{se}$ je odhad směrodatné chyby $\widehat{\theta}$. S rostoucím n se rozdělení odhadu $\widehat{\theta}$ stále více přibližuje normálnímu rozdělení se střední hodnotou blízkou θ a s rozptylem blízkým $\widehat{se}^2$. Píšeme

$$\widehat{\theta} \dot{\sim} N(\theta, \widehat{se}^2) \text{ neboli } \frac{\widehat{\theta} - \theta}{\widehat{\mathrm{se}}} \dot{\sim} \mathrm{N}(0,1).$$

Bootstrapový výběr získáme z množiny $x_1, \ldots, x_n$ vygenerováním (výběrem s opakováním) opět n prvků, $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$, který nazveme bootstrapovým výběrem. Počet všech různých bootstrapových výběrů rozsahu n je $\binom{2n-1}{n}$. Pro každý bootstrapový výběr vypočítáme příslušný parametr θ^*. Pokud celý tento proces zopakujeme B-krát, dostaneme $\theta_1^*, \ldots, \theta_B^*$, které představují bootstrap populaci parametru θ^*. Obvyklým způsobem se pak dá spočítat aritmetický průměr a směrodatná odchylka a pro velké B pak můžeme sestrojit histogram, který odpovídá rozdělení parametru θ^*.

Pivotové odhady

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

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

s hustotou pravděpodobnosti

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

Pak X je náhodná spojitá veličina se střední hodnotou $E(X) = \mu$, rozptylem $D(X) = \sigma^2$ a směrodatnou odchylkou $\sigma(X) = \sigma$.

O pár řádků níž si ukážeme, jak metodou bootstrap můžeme získat odhad konfidenčního intervalu pro odhady střední hodnoty μ, rozptylu σ^2 a směrodatné odchylky σ. Budeme odhadovat μ výběrovým průměrem $\overline{X}$ a σ výběrovou směrodatnou odchylkou S.

Intervalový odhad střední hodnoty

Pokud U 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

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

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

$$\mu \in \left(\overline{X} - t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}; \overline{X} + t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right).$$

Pokud U nemá normované normální rozdělení pravděpodobnosti a rozdělení pravděpodobnosti statistiky t je stále nezávislé na μ, σ, tak už nejde o Studentovo t - rozdělení. I přesto, pokud bychom chtěli zjistit hodnoty kvantilů tohoto neznámého rozdělení, tak by stále platilo

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

Metodou bootstrap odhadneme hodnoty kvantilů rozdělení pravděpodobnosti statistiky t.

Na následujícím postupu si ukážeme, jak můžeme získat konfidenční interval pro $\mu = E(X)$.

- Pro pozorování $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ spočteme hodnoty výběrového průměru $\overline{X}$ a výběrové směrodatné odchylky S,
- vygenerujeme B náhodných bootstrapových výběrů $\mathbf{X}_i^*$, $i = 1, \ldots, B$, s opakováním o rozsahu n z původního souboru $\mathbf{X}$,
- pro každý takto vygenerovaný bootstrapový výběr se spočte pozorovaná hodnota výběrového průměru $\overline{X}_i^*$ a hodnota výběrové směrodatné odchylky S_i^* a hodnota statistiky t,

$$t_i^* = \frac{\overline{X}_i^* - \overline{X}}{S_i^*/\sqrt{n}},$$

kde $i = 1, \ldots, B$,

- odhadneme $\frac{\alpha}{2}$-kvantil a $(1 - \frac{\alpha}{2})$-kvantil rozdělení pravděpodobnosti statistiky t^* pomocí hodnot $t^*_{\frac{\alpha}{2}}$ a $t^*_{1-\frac{\alpha}{2}}$ tak, že

$$\frac{\left|\left\{t_i^*; t_i^* \leq t^*_{\frac{\alpha}{2}}\right\}\right|}{B} \doteq \frac{\alpha}{2},$$
$$\frac{\left|\left\{t_i^*; t_i^* \leq t^*_{1-\frac{\alpha}{2}}\right\}\right|}{B} \doteq 1 - \frac{\alpha}{2},$$

- pak bootstrapovým konfidenčním intervalem se spolehlivostí $1-\alpha$ pro střední hodnotu $E(X)$ je

$$\left(\overline{X} - t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}; \overline{X} - t^*_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right).$$

Intervalový odhad rozptylu a směrodatné odchylky

Nechť U 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

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

kde $\chi^2_{\frac{\alpha}{2}}$ a $\chi^2_{1-\frac{\alpha}{2}}$ jsou $\frac{\alpha}{2}$-kvantil a $(1 - \frac{\alpha}{2})$-kvantil Pearsonova rozdělení s $n - 1$ stupni volnosti. Konfidenční interval se spolehlivostí $1 - \alpha$ je

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

Pokud U nemá normální rozdělení pravděpodobnosti a rozdělení pravděpodobnosti statistiky χ^2 je nezávislé na μ a σ, tak už nejde o Pearsonovo rozdělení. I přesto pokud bychom chtěli zjistit hodnoty kvantilů tohoto neznámého rozdělení, tak

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

by stále platilo.

Metodou bootstrap odhadneme hodnoty kvantilů rozdělení pravděpodobnosti statistiky χ^2.

Na následujícím postupu si ukážeme, jak můžeme získat konfidenční interval pro $\sigma^2 = D(X)$ a $\sigma = \sigma(X)$.

- Pro pozorování $(x_1, \ldots, x_n)$ náhodného výběru $(X_1, \ldots, X_n)$ spočteme hodnoty výběrového rozptylu S^2,

- vygenerujeme B náhodných bootstrapových výběrů $\mathbf{X}_i^*$, $i = 1, \ldots, B$, s opakováním o rozsahu n z původního souboru $\mathbf{X}$,
- pro každý takto vygenerovaný bootstrapový výběr se spočte pozorovaná hodnota výběrového rozptylu S_i^{2*} a hodnota statistiky χ^2,

$$\chi_i^{2*} = \frac{(n-1)S_i^{2*}}{S^2},$$

kde $i = 1, \ldots, B$,

- odhadneme $\frac{\alpha}{2}$-kvantil a $(1 - \frac{\alpha}{2})$-kvantil rozdělení pravděpodobnosti statistiky χ^{2*} pomocí hodnot $\chi^{2*}_{\frac{\alpha}{2}}$ a $\chi^{2*}_{1-\frac{\alpha}{2}}$ tak, že

$$\frac{\left|\left\{\chi_i^{2*}; \chi_i^{2*} \leq \chi^{2*}_{\frac{\alpha}{2}}\right\}\right|}{B} \doteq \frac{\alpha}{2},$$
$$\frac{\left|\left\{\chi_i^{2*}; \chi_i^{2*} \leq \chi^{*}_{1-\frac{\alpha}{2}}\right\}\right|}{B} \doteq 1 - \frac{\alpha}{2},$$

- pak bootstrapovým konfidenčním intervalem se spolehlivostí $1 - \alpha$ pro rozptyl $D(X)$ je

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

- a bootstrapovým konfidenčním intervalem se spolehlivostí $1 - \alpha$ pro směrodatnou odchylku $\sigma(X)$ je

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

Kvantilové odhady

V této části se podíváme na intervalové odhady vycházející přímo z rozdělení pravděpodobnosti bodových odhadů. Jedná se o zcela obecné postupy, proto je můžeme použít pro libovolné parametry, případně parametrickou funkci, a pro libovolný jeho odhad.

Jednoduchý kvantilový konfidenční interval

Nechť $\widehat{\theta}$ je odhad parametru θ a $\widehat{\sigma}$ je odhad jeho směrodatné odchylky. Standardní normální konfidenční interval je tvaru

$$\left(\widehat{\theta} - z_{1-\frac{\alpha}{2}}\widehat{\sigma}; \widehat{\theta} - z_{\frac{\alpha}{2}}\widehat{\sigma}\right).$$

Nechť $\widehat{\theta}^*$ označuje náhodnou proměnnou s normálním rozdělením pravděpodobnosti $N\left(\widehat{\theta}, \widehat{\sigma}^2\right)$. Pak

$$\widehat{\theta} - z_{1-\frac{\alpha}{2}}\widehat{\sigma} = \widehat{\theta}^*_{1-\frac{\alpha}{2}} = 100 \cdot (1 - \frac{\alpha}{2}) \text{ percentil rozdělení } \widehat{\theta}^*,$$

$$\widehat{\theta} - z_{\frac{\alpha}{2}}\widehat{\sigma} = \widehat{\theta}^*_{\frac{\alpha}{2}} = \widehat{\theta}^*_{1-\frac{\alpha}{2}} = 100 \cdot (\frac{\alpha}{2}) \text{ percentil rozdělení } \widehat{\theta}^*.$$

Následující postup nám demonstruje, jak lze sestrojit jednoduchý kvantilový konfidenční interval.

- Vygenerujeme B náhodných bootstrapových výběrů $\mathbf{X}^*_i$, $i = 1, \ldots, B$, s opakováním o rozsahu n z původního souboru $\mathbf{X}$,
- pro každý takto vygenerovaný bootstrapový výběr se spočte odhad $\widehat{\theta}^*_i$ parametru θ,
- odhadneme kvantily $(1 - \frac{\alpha}{2})$ a $\frac{\alpha}{2}$-kvantil rozdělení pravděpodobnosti bootstrapové statistky $\widehat{\theta}^*$ pomocí hodnot $\widehat{\theta}^*_{\frac{\alpha}{2}}, \widehat{\theta}^*_{1-\frac{\alpha}{2}}$ tak, že

$$\frac{\left|\left\{\widehat{\theta}^*_i; \widehat{\theta}^*_i \leq \widehat{\theta}^*_{\frac{\alpha}{2}}\right\}\right|}{B} \doteq \frac{\alpha}{2},$$
$$\frac{\left|\left\{\widehat{\theta}^*_i; \widehat{\theta}^*_i \leq \widehat{\theta}^*_{1-\frac{\alpha}{2}}\right\}\right|}{B} \doteq 1 - \frac{\alpha}{2}$$

- bootstrapovým jednoduchým kvantilovým konfidenčním intervalem se spolehlivostí $1 - \alpha$ je

$$\left(\widehat{\theta}^*_{\frac{\alpha}{2}}; \widehat{\theta}^*_{1-\frac{\alpha}{2}}\right).$$

Reziduový kvantilový konfidenční interval

Výrazem $e = \widehat{\theta} - \theta$ chápeme *reziduum*. $\frac{\alpha}{2}$-kvantil a $(1 - \frac{\alpha}{2})$-kvantil rozdělení pravděpodobnosti náhodné veličiny e označíme $e_{\frac{\alpha}{2}}$ a $e_{1-\frac{\alpha}{2}}$. Pak platí

$$P\left(e_{\frac{\alpha}{2}} < \widehat{\theta} - \theta \leq e_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Na základě toho odvodíme konfidenční interval se spolehlivostí $1 - \alpha$

$$\left(\widehat{\theta} - e_{1-\frac{\alpha}{2}}; \widehat{\theta} - e_{\frac{\alpha}{2}}\right).$$

Protože kvantily rozdělení pravděpodobnosti rezidua neznáme, musíme je odhadnout metodou bootstrap.

Nyní si ukážeme postup pro získání reziduového konfidenčního intervalu.

- Nejprve z pozorovaných hodnot $\mathbf{x}$ náhodného výběru $\mathbf{X}$ spočteme odhad $\widehat{\theta}$ parametru θ,

- vygenerujeme B bootstrapových výběrů $\mathbf{X}_i^*$, $i = 1, \ldots, B$, s opakováním o rozsahu n z původního souboru pozorovaných hodnot $\mathbf{x}$,
- pro každý takto vygenerovaný bootstrapový výběr se spočte odhad $\widehat{\theta}_i^*$ parametru θ a reziduum $e_i^* = \widehat{\theta}_i^* - \widehat{\theta}$,
- v předposledním kroku odhadneme $\frac{\alpha}{2}$-kvantil a $(1 - \frac{\alpha}{2})$-kvantil rozdělení pravděpodobnosti reziduí e^* pomocí $e^*_{\frac{\alpha}{2}}$ a $e^*_{1-\frac{\alpha}{2}}$ tak, že

$$\frac{\left|\left\{e_i^*; e_i^* \leq e^*_{\frac{\alpha}{2}}\right\}\right|}{B} \doteq \frac{\alpha}{2},$$
$$\frac{\left|\left\{e_i^*; e_i^* \leq e^*_{1-\frac{\alpha}{2}}\right\}\right|}{B} \doteq 1 - \frac{\alpha}{2}$$

- a nakonec bootstrapovým reziduovým konfidenčním intervalem se spolehlivostí $1 - \alpha$ je

$$\left(\widehat{\theta} - e^*_{1-\frac{\alpha}{2}}; \widehat{\theta} - e^*_{\frac{\alpha}{2}}\right).$$

Možností, jak zkonstruovat konfidenční interval metodou bootstrap je více:

- a) studentizovaný bootstrap interval,
- b) BC_a interval (bias-corrected a accelerated),
- c) ABC interval (approximate bootstrap konfidence interval),
- d) prepivoting bootstrap interval.

Studentizované intervaly spolehlivosti

Nechť statistika $T = \frac{\widehat{\theta}-\theta}{\widehat{\sigma}}$ má normální rozdělení pravděpodobnosti $N(0, 1)$ a nechť u_α značí α-kvantil normálního rozdělení $N(0, 1)$. Pak obecně lez zapsat interval spolehlivosti jako

$$\left(\widehat{\theta} - u_{1-\frac{\alpha}{2}}\widehat{\sigma}; \widehat{\theta} - u_{\frac{\alpha}{2}}\widehat{\sigma}\right)$$

a tento výraz se nazývá *standardní interval spolehlivosti* parametru θ se spolehlivostí $1 - \alpha$, kde $\widehat{\sigma}$ je odhad směrodatné odchylky a $u_{1-\frac{\alpha}{2}}$ je $(1 - \frac{\alpha}{2})$-kvantil normálního rozdělení (tj. $1,96$ pro 95% konfidenční interval, kde $\alpha = 0,05$).

Pro interval spolehlivosti $\left(\widehat{\theta} - u_{1-\frac{\alpha}{2}}\widehat{\sigma}; \widehat{\theta} - u_{\frac{\alpha}{2}}\widehat{\sigma}\right)$ vycházíme z předpokladu $T = \frac{\widehat{\theta}-\theta}{\widehat{\sigma}} \dot{\sim} N(0, 1)$. Pokud ale můžeme předpokládat, že $T = \frac{\widehat{\theta}-\theta}{\widehat{\sigma}} \dot{\sim} t(n-1)$, kde $t(n-1)$ reprezentuje Studentovo t-rozdělení pravděpodobnosti s $n-1$ stupni volnosti. A platí

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

kde $t_{1-\frac{\alpha}{2}}$ je $(1 - \frac{\alpha}{2})$-kvantil Studentova rozdělení pravděpodobnosti s $n - 1$ stupni volnosti. Pak užitím této t aproximace dostaneme interval spolehlivosti pro parametr θ,

$$\theta \in \left(\widehat{\theta} - t_{1-\frac{\alpha}{2}}\widehat{\sigma}; \widehat{\theta} + t_{1-\frac{\alpha}{2}}\widehat{\sigma}\right).$$

Teď použijme metodu bootstrap ke zjištění konfidenčního intevalu bez nutnosti předpokladu normality. Postup pro zjištění bootstrapového intervalu je následující,

- nejprve z pozorovaných hodnot $\mathbf{x}$ náhodného výběru $\mathbf{X}$ spočteme odhad $\widehat{\theta}$ parametru θ,
- vygenerujeme B bootstrapových výběrů $\mathbf{X}_i^*$, $i = 1, \ldots, B$, s opakováním o rozsahu n z původního souboru $\mathbf{X}$,
- pro každý takto vygenerovaný bootstrapový výběr se spočte odhad $\widehat{\theta}_i^*$ parametru θ a jeho směrodatná odchylka σ_i^*,
- dále pro každý bootstrapový výběr spočteme hodnotu statistiky T,

$$t_i^* = \frac{\widehat{\theta}_i^* - \widehat{\theta}}{\sigma_i^*},$$

  kde $i = 1, \ldots, B$,
- spočteme odhad výběrové směrodatné odchylky $\widehat{\sigma}_i^*$ odhadu $\widehat{\theta}$,
- v předposledním kroku se spočte odhad hodnoty $(1 - \frac{\alpha}{2})$- kvantilu rozdělení pravděpodobnosti bootstrapové statistky T^* pomocí hodnot $t^*_{1-\frac{\alpha}{2}}$ tak, že

$$\frac{\left|\left\{t_i^*; t_i^* \leq t^*_{1-\frac{\alpha}{2}}\right\}\right|}{B} \doteq 1 - \frac{\alpha}{2},$$
$$\frac{\left|\left\{t_i^*; t_i^* \leq t^*_{\frac{\alpha}{2}}\right\}\right|}{B} \doteq \frac{\alpha}{2},$$

- a nakonec bootstrapovým t-konfidenčním intervalem se spolehlivostí $1 - \alpha$ je

$$\left(\widehat{\theta} - t^*_{\frac{\alpha}{2}}\widehat{\sigma}^*; \widehat{\theta} + t^*_{1-\frac{\alpha}{2}}\widehat{\sigma}^*\right).$$

Výhodou studentizovaného bootstrapu je, že je to přístup intuitivní a jednoduchý na pochopení. Ale nevýhodou je, že tato metoda není "automaticky spočítatelná", protože závisí na existenci věrohodného odhadu směrodatné odchylky $\widehat{\sigma}(x_1, x_2, \ldots, x_n)$. V praxi tato metoda může dávat zavádějící výsledky a může být silně ovlivněna odlehlými pozorováními. Proto metody založené na percentilu jsou více spolehlivé.

BC_a interval (bias-corrected and accelerated)

Rozdělení pravděpodobnosti $\widehat{\theta}^*$ je obvykle nesymetrické a zešikmené na jednu stranu. Pokud jednoduché a reziduové kvantilové konfidenční intervaly jsou vychýlené nebo příliš široké oproti hodnotám z praktického pozorování, tak tyto nedostatky můžeme odstranit pomocí BC_a konfidenčních intervalů. Tyto intervaly jsou také omezeny

dvěma kvantily rozdělení pravděpododbnosti bootstrapového odhadu $\widehat{\theta}^*$, ale nemusí se nutně jednat o $\frac{\alpha}{2}$-kvantil a $(1-\frac{\alpha}{2})$-kvantil se spolehlivostí $1-\alpha$ jako v předchozích metodách.

Pro tuto metodu je důležitý předpoklad, že existuje nějaká transformace parametru θ s normálním rozdělením pravděpodobnosti a se střední hodnotou a rozptylem závisejícím na θ. Konfidenční interval se pak zkonstruuje pro transformovaný parametr a pomocí inverzní transformace mezí získáme konfidenční interval pro θ. Transformaci z předpokladu nemusíme znát v explicitním tvaru, stačí na ni použít metodu bootstrap.

BC_a metoda závisí na dvou numerických parametrech: bias-korekce z_0, zrychlení a.

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

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

a směrodatnou odchylkou

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

Konfidenční interval se spolehlivostí $1 - \alpha$ se odvodí z

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

jako

$$\left(\frac{T(\widehat{\theta}) + z_0 - u_{1-\frac{\alpha}{2}}}{1 - a(z_0 - u_{1-\frac{\alpha}{2}})}; \frac{T(\widehat{\theta}) + z_0 + u_{1-\frac{\alpha}{2}}}{1 - a(z_0 + u_{1-\frac{\alpha}{2}})}\right).$$

Protože náhodné veličiny

$$\frac{T(\widehat{\theta}^*) - T(\widehat{\theta})}{1 + aT(\widehat{\theta})} + z_0,$$
$$\frac{T(\widehat{\theta}) - T(\theta)}{1 + aT(\theta)} + z_0$$

mají stejné rozdělení pravděpodobnosti (dle předpokladu se jedná o normované normální rozdělení pravděpodobnosti) a platí

$$P\left(T(\widehat{\theta}^*) < \frac{T(\widehat{\theta}) + z_0 - u_{1-\frac{\alpha}{2}}}{1 - a(z_0 + u_{1-\frac{\alpha}{2}})}\right) = P\left(\frac{T(\widehat{\theta}^*) - T(\widehat{\theta})}{1 + aT(\widehat{\theta})} + z_0 < \frac{z_0 + u_{1-\frac{\alpha}{2}}}{1 - a(z_0 + u_{1-\frac{\alpha}{2}})} + z_0\right)$$
$$= P\left(U < \frac{z_0 + u_{1-\frac{\alpha}{2}}}{1 - a(z_0 + u_{1-\frac{\alpha}{2}})} + z_0\right),$$

kde U má normované normální rozdělení pravděpodobnosti. Pak horní mez konfidenčního intervalu pro $T(\theta)$ je

$$u_H = \frac{z_0 + u_{1-\frac{\alpha}{2}}}{1 - a(z_0 + u_{1-\frac{\alpha}{2}})} + z_0.$$

Obdobně dostaneme i dolní mez konfidenčního intervalu

$$u_D = \frac{z_0 - u_{1-\frac{\alpha}{2}}}{1 - a(z_0 - u_{1-\frac{\alpha}{2}})} + z_0.$$

Nyní odhadneme hodnoty parametrů z_0, a.
Nechť $\widehat{\theta}_{-i}$ je odhad parametru θ, který dostaneme vynecháním i-tého pozorování z náhodného výběru, tj. náhodný výběr se zmodifikuje jako

$$(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Dále označíme

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

Pak spočítáme

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

Nyní si předvedeme postup pro získání BC_a konfidenčního intervalu pro parametr θ.

- Nejprve z pozorovaných hodnot $\mathbf{x}$ náhodného výběru $\mathbf{X}$ spočteme odhad $\widehat{\theta}$ parametru θ,
- vygenerujeme B bootstrapových výběrů $\mathbf{X}_i^*$, $i = 1, \ldots, B$, s opakováním o rozsahu n z původního souboru pozorovaných hodnot $\mathbf{x}$,
- pro každý takto vygenerovaný bootstrapový výběr se spočte odhad $\widehat{\theta}_i^*$ parametru θ,
- spočteme korekci vychýlení mediánu

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

kde Φ^{-1} je inverzní distribuční funkce normovaného normálního rozdělení pravděpodobnosti,

- dále spočteme parametr akcelerace a,

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

- z předchozích výpočtů spočteme

$$\alpha_1 = \Phi\left(\frac{z_0 - u_{1-\frac{\alpha}{2}}}{1 - a(z_0 - u_{1-\frac{\alpha}{2}})} + z_0\right),$$
$$\alpha_2 = \Phi\left(\frac{z_0 + u_{1-\frac{\alpha}{2}}}{1 - a(z_0 + u_{1-\frac{\alpha}{2}})} + z_0\right),$$

- v předposledním kroku odhadneme α_1-kvantil a $(1-\alpha_2)$-kvantil rozdělení pravděpodobnosti statistiky $\widehat{\theta}^*$ pomocí $\widehat{\theta}^*_{\alpha_1}$, $\widehat{\theta}^*_{1-\alpha_2}$ tak, že

$$\frac{\left|\left\{\widehat{\theta}^*_i; \widehat{\theta}^*_i \leq \widehat{\theta}^*_{\alpha_1}\right\}\right|}{B} \doteq \alpha_1,$$
$$\frac{\left|\left\{\widehat{\theta}^*_i; \widehat{\theta}^*_i \leq \widehat{\theta}^*_{1-\alpha_2}\right\}\right|}{B} \doteq 1 - \alpha_2,$$

- BC_a konfidenční interval se spolehlivostí $1 - \alpha$ pro parametr θ je

$$\left(\widehat{\theta}^*_{\alpha_1}; \widehat{\theta}^*_{1-\alpha_2}\right).$$

Parametr zrychlení a se nedá přímo odhadnout z bootstrapových dat. Protože BC_a intervaly závisí na parametru a, který nejde odhadnout z bootstrapových dat, tak se tato metoda stává méně intuitivní. Pokud zvolíme $a = 0$, tak dostaneme jednodušší verzi BC_a konfidenčních intervalů, tzv. *BC konfidenční interval*.

ABC metoda (approximate bootstrap konfidence interval)

Nyní opustíme oblast , kdy jsme měli jen jeden parametr a přejdeme ke složitější situaci. V mnoha takových případech je možné aproximovat koncové body BC_a intervalu analyticky. Hlavní nevýhodou BC_a intervalů je velký počet bootstrapových výběrů. *ABC* metoda konstrukce konfidenčních intervalů je metoda aproximující koncové body BC_a konfidenčních intervalů analyticky. Je možné tento přístup aplikovat také na neparametrický problém.

Koncové body BC_a intervalu závisí na distribuční funkci $\widehat{G}$ a na odhadech parametrů a, z_0. ABC přístup vyžaduje navíc jeden odhad nelineárního parametru c_q, ale to nijak nekomplikuje výpočet distribuční funkce $\widehat{G}$. Standardní intervaly závisí pouze na dvou veličinách, $(\widehat{\theta}, \widehat{\sigma})$. ABC intervaly závisí na pěti veličinách, $(\widehat{\theta}, \widehat{\sigma}, \widehat{a}, \widehat{z_0}, \widehat{c_q})$.

Místo odhadu z_0 se použije bootstrapové rozdělení pravděpodobnosti jako v BC_a metodě.

Prepivoting metoda (kalibrace bootstrap)

Kalibrace je bootstrapová technika na získání konfidenčního intervalu s vyšším řádem přesnosti. Předpokládejme, že $\widehat{\theta}(\alpha)$ je horní mez jednostranné α-aproximace konfidenčního intervalu pro parametr θ. Nechť $\gamma(\alpha) = P(\theta < \widehat{\theta}(\alpha))$ je *kalibrační křivka*. Pokud aproximace je přesná, pak $\gamma(\alpha) = \alpha$ pro nějaké dané α. Nebo také můžeme použít kalibrační křivku pro aproximaci konfidenčního intervalu.
Např.: pro $\gamma(0,03) = 0,025$, $\gamma(0,96) = 0,975$ lze použít $(\widehat{\theta}[0,03], \widehat{\theta}[0,96])$ jako aproximaci $0,95$ konfidenčního intervalu.
V aplikacích obvykle neznáme kalibrační křivku $\gamma(\alpha)$, ale můžeme použít bootstrap metodu na odhad $\gamma(\alpha)$:

$$\widehat{\gamma}(\alpha) = P^*(\widehat{\theta} < \widehat{\theta}^*(\alpha)),$$

kde P^* udává bootstrapová data a $\widehat{\theta}^*(\alpha)$ je horní mez α-intervalu.
Tato metoda se dá aplikovat na všechny předchozí metody, např.: k získání třetího řádu přesnosti konfidenčních intervalů ze studentizovaných bootstrap intervalů.
Bootstrap kalibrace řeší více výpočtů, v praxi obvykle velikost testovaného souboru nebývá tak velká, tak lze použít jednu bootstrap kalibraci.

V mnoha literaturách o metodě bootstrap se řeší problém, jak zkonstruovat konfidenční intervaly vyšších řádů přesnosti než prvního. Pro tyto případy existují právě předchozí 4 metody:

- a) studentizovaný bootstrap interval,
- b) BC_a interval (bias-corrected a accelerated),
- c) ABC interval (approximate bootstrap konfidence interval),
- d) prepivoting bootstrap interval.

3.3 Praktická aplikace

Uvažujme situaci, kdy jsme vygenerovali soubor $\mathbf{X}$ o n pozorování, pro který spočteme příslušné charakteristiky, střední hodnotu, směrodatnou odchylku, rozptyl a špičatost.

$\mathbf{x}$

-2,214	0,206	0,537	-0,254	2,247	-1,284	1,542	2,780	-0,531	1,065
0,315	-1,997	-1,095	-1,186	1,466	0,169	1,544	1,802	1,791	0,680

$$E(X) = 0,379$$
$$\sigma(X) = 1,393$$
$$\sigma^2(X) = D(X) = 1,94$$
$$\gamma_2 = -0,254$$

Na tento soubor $\mathbf{X}$ aplikujeme metodu bootstrap. Tedy provedeme B-krát výběr s opakováním a dostaneme boostrapový výběr $\mathbf{X}^*$. Každý soubor X_i^* obsahuje hodnoty $x_1, \ldots, x_n$ z původního výběru $\mathbf{X}$ s pravděpodobností $\frac{1}{n}$.

$\mathbf{x}_1^*$

1,542	0,537	1,466	1,542	-1,186	-1,997	0,680	-2,214	-1,284	-0,531
1,802	1,466	-0,531	-1,095	-2,214	-0,254	-1,095	1,544	-1,284	-2,214

$\mathbf{x}_2^*$

2,780	2,247	-1,997	-0,531	0,169	-1,095	-1,997	1,791	2,780	-1,284
1,802	1,544	1,065	1,791	-0,254	1,065	1,065	0,169	-0,531	-1,186

$\mathbf{x}_3^*$

-1,186	-2,214	1,802	-0,254	-2,214	-1,186	-0,254	-2,214	-2,214	0,680
1,065	-2,214	2,780	0,537	0,206	2,247	-1,997	-0,254	-2,214	1,065

$\mathbf{x}_4^*$

2,780	0,206	1,791	-1,186	1,466	1,466	0,537	1,802	-1,997	0,169
-1,186	-1,997	0,169	-2,214	1,542	-2,214	1,065	1,065	0,537	1,544

$\mathbf{x}_5^*$

1,466	1,065	0,315	-1,095	0,169	0,537	-2,214	-0,254	1,791	2,780
1,791	0,680	2,247	1,542	1,802	-1,997	2,247	2,247	0,680	-1,997

$\vdots$

Pro jednotlivá X_i^* spočteme příslušné charakteristiky, tím dostaneme B hodnot charakteristik a tedy i statistický soubor o B prvcích, pro který budeme počítat jednotlivé konfidenční intervaly.
Spočtěme střední hodnotu, směrodatnou odchylku, rozptyl a špičatost.

$E(X^*)$	$\sigma(X^*)$	$D(X^*)$	γ_2^*
-0,266	1,459	2,128	0,130
0,470	1,530	2,340	-0,116
-0,402	1,658	2,750	0,377
0,267	1,545	2,387	-0,448
0,690	1,524	2,324	-0,733
0,299	1,399	1,957	-0,018
0,576	1,457	2,124	-0,188
0,298	1,416	2,006	-0,367
-0,066	1,302	1,695	-0,095
0,883	1,195	1,427	-0,240
⋮	⋮	⋮	⋮

Nyní máme potřebné podklady k tomu, abychom se mohli zabývat konfidenčními intervaly pro jednotlivé charakteristiky. Spočítáme 90% konfidenční interval střední hodnoty, směrodatné odchylky, rozptylu a špičatosti.

$$\begin{aligned} E(X^*) &\in \langle -1,978; 1,887\rangle \\ \sigma(X^*) &\in \langle 3,395; 4,985\rangle \\ D(X^*) &\in \langle 11,526; 24,852\rangle \\ \gamma_2^* &\in \langle -0,574; 0,596\rangle, \end{aligned}$$

kde $E(X^*) \in \left(\overline{X} - t^*_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}; \overline{X} - t^*_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)$, $D(X^*) \in \left(\frac{(n-1)S^2}{\chi^{2*}_{1-\frac{\alpha}{2}}}; \frac{(n-1)S^2}{\chi^{2*}_{\frac{\alpha}{2}}}\right)$ a bootstrapovým konfidenčním intervalem se spolehlivostí $1-\alpha$ pro směrodatnou odchylku $\sigma(X)$ je $\sigma(X^*) \in \left(\sqrt{\frac{(n-1)S^2}{\chi^{2*}_{1-\frac{\alpha}{2}}}}; \sqrt{\frac{(n-1)S^2}{\chi^{2*}_{\frac{\alpha}{2}}}}\right)$. Bootstrapovým reziduovým konfidenčním intervalem špičatosti γ_2 se spolehlivostí $1-\alpha$ je $\gamma_2^* \in \left(\widehat{\theta} - e^*_{1-\frac{\alpha}{2}}; \widehat{\theta} - e^*_{\frac{\alpha}{2}}\right)$.

Se znalostí konfidenčních intervalů můžeme testovat statistické hypotézy, např, zda střední hodnota bootstrapového souboru je nula, $H_0 : E(X^*) = 0$, proti alternativě, že střední hodnota je různá od nuly, $H_1 : E(X^*) \neq 0$. Pro naši situaci vidíme, že nula je prvkem konfidenčního intervalu pro $E(X^*)$, tak hypotézu H_0 nezamítáme na hladině významnosti α, $\alpha = 0,1$.

4 TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

Výsledky v této části diplomové práce jsou podloženy zdrojem [11].

Sledujeme-li náhodné veličiny a náhodné vektory, může se stát, že okolnosti nás přimějí ověřit určité předpoklady či domněnky o jejich vlastnostech pomocí jejich pozorovaných hodnot. Tato tvrzení se nazývají statistické hypotézy a neexistuje matematický postup, který by prokázal, že daná statistická hypotéza platí. Pouze rozhodne, zda tuto hypotézu zamítáme a dopustíme se chyby s pravděpodobností menší než α, nebo hypotézu nezamítáme, ale neznamená to, že hypotéza musí platit, můžeme mít jen nedostatek informací, abychom hypotézu zamítli.

Statistická hypotéza H_0 je tvrzení o vlastnostech rozdělení pravděpodobnosti pozorované náhodné veličiny X s distribuční funkcí $F(x, \theta)$ nebo náhodného vektoru (X, Y) se simultánní distribuční funkcí $F(x, y, \theta)$.

Postup, kterým ověřujeme danou statistickou hypotézu, se nazývá *test statistické hypotézy.* Hypotézu H_0 také nazýváme nulovou hypotézou a testujeme ji proti hypotéze H_1, která se nazývá *alternativní hypotéza*, která se volí dle požadavků úlohy.

Řekněme, že hypotéza H_0 je tvrzení, které říká, že parametr θ má hodnotu θ_0, pak píšeme $H_0 : \theta = \theta_0$. Podle tvaru hypotézy H_1 mohou nastat dva případy. Pokud H_1 je tvaru $H_1 : \theta \neq \theta_0$, tak se jedná o *dvoustrannou alternativní hypotézu.* Je-li H_1 tvaru $H_1 : \theta > \theta_0$, resp. $H_1 : \theta < \theta_0$, jedná se o *jednostrannou alternativní hypotézu.*

Pokud testujeme hypotézu $H_0 : \theta = \theta_0$ proti nějaké zvolené alternativě H_1, tak zkonstruujeme vhodnou statistiku $T(X_1, \ldots, X_n)$ a tuto statistiku T nazýváme *testové kritérium.*

Předpokládejme, že platí hypotéza $H_0 : \theta = \theta_0$, pak obor hodnot testového kritéria $T(X_1, \ldots, X_n)$ se rozdělí na dvě disjunktní podmnožiny, *kritický obor* W_α a jeho doplněk $\overline{W}_\alpha$. Kritický obor W_α se volí tak, aby pravděpodobnost, že $T(X_1, \ldots, X_n)$ nabude hodnotu z kritického oboru W_α, byla α (přesněji pro diskrétní náhodnou veličinu T nejvýše α). $\overline{W}_\alpha$ se nazývá *obor nezamítnutí.*

Číslo α se nazývá *hladina významnosti* testu a volíme ji blízkou nule (obvykle $0,05$ nebo $0,01$).

Rozhodnutí o hypotéze H_0 se provede podle konvence, pokud *pozorovaná hodnota testového kritéria* $t = T(x_1, \ldots, x_n)$ na statistickém souboru $(x_1, \ldots, x_n)$ padne do kritického oboru W_α, neboli $t \in W_\alpha$, hypotézu H_0 zamítáme a současně hypotézu H_1 nezamítáme na hladině významnosti α. Pokud nastane opačná situace, tedy t nepadne do kritického oboru W_α, neboli $t \in \overline{W}_\alpha$, tak nezamítáme hypotézu H_0 a současně zamítáme hypotézu H_1 na hladině významnosti α.

Při testování hypotézy H_0 mohou nastat čtyři možnosti.

Chyba prvního druhu nastane, jestliže hypotéza H_0 platí, avšak $t \in W_\alpha$, takže hypotézu H_0 zamítáme. Pravděpodobnost této chyby je $\alpha = P(T \in W_\alpha | H_0)$.

H_0	PLATÍ	NEPLATÍ
ZAMÍTÁME	CHYBA 1. DRUHU (α)	$- - -$
NEZAMÍTÁME	$- - -$	CHYBA 2. DRUHU (β)

Tab. 4.1: Skutečnost versus rozhodnutí

Chyba druhého druhu nastane, jestliže hypotéza H_0 neplatí, avšak $t \notin W_\alpha$, takže hypotézu H_0 nezamítáme. Pravděpodobnost této chyby je $\beta = P(T \notin W_\alpha | H_1)$ a pravděpodobnost $1 - \beta = P(T \in W_\alpha | H_1)$ se nazývá *síla testu*.

Hladina významnosti α, tedy pravděpodobnost chyby prvního druhu, má statistický význam, pokud mnohokrát opakujeme experiment za stejných podmínek a současně platí hypotéza H_0, tak se přibližně v $100\alpha\%$ testech této hypotézy zmýlíme, tudíž zamítneme platnou hypotézu. Obdobně, pokud hypotéza H_0 neplatí, tak se přibližně v $100\beta\%$ testech této hypotézy zmýlíme a nezamítneme ji.

Snížíme-li hladinu významnosti α a nezměníme rozsah statistického souboru n, zvýší se β a naopak, tudíž pro zvolenou hladinu významnosti α zajišťujeme snížení β zvýšením rozsahu n.

Pokud testujeme statistické hypotézy na počítačích, tak se místo kritické ho oboru $\overline{W}_\alpha$ používá tzv. *P-hodnota*. Testujeme-li hypotézu $H_0 : \mu = \mu_0$ proti dvoustranné alternativě $H_1 : \mu \neq \mu_0$, pak pro pozorovanou hodnotu t testového kritéria T je P-hodnota číslo

$$1 - P(-t \leq T \leq t).$$

Při dané konvenci rozhodnutí pomocí kritického oboru odpovídá postup, pokud $P < \alpha$, pak zamítáme hypotézu H_0 a současně nezamítáme H_1 na hladině významnosti α. Pokud $P \geq \alpha$, pak nezamítáme hypotézu H_0 a současně zamítáme hypotézu H_1 na hladině významnosti α.

Použití metody bootstrap pro testování hypotéz je více či méně samozřejmé, protože můžeme použít poznatků z konfidenčních intervalů, kdy můžeme testovat rovnost parametru nějaké specifické hodnoty, obvykle se testuje nulovost parametru. Tedy pokud budeme testovat hypotézu $H_0 : \theta = \theta_0$ na hladině významnosti α, pak sestrojíme konfidenční interval pro parametr θ se spolehlivostí $1 - \alpha$.
Hypotézu H_0 nezamítáme na hladině významnosti α, pokud θ_0 bude prvkem příslušného konfidečního intervalu. V opačném případě hypotézu H_0 zamítáme.

Testujeme-li hypotézu $H_0 : \theta = \theta_0$ proti alternativní hypotéze $H_1 : \theta \neq \theta_0$, a pokud $t \in \overline{W}_\alpha = \left\langle \widehat{t}_{\frac{\alpha}{2}}; \widehat{t}_{1-\frac{\alpha}{2}} \right\rangle$, pak hypotézu H_0 nezamítáme a současně hypotézu H_1 zamítáme na hladině významnosti α.

Testujeme-li hypotézu $H_0 : \theta = \theta_0$ proti jednostranné hypotéze $H_1 : \theta > \theta_0$, a pokud $t \in \overline{W}_\alpha = \left(-\infty; \widehat{t}_{1-\alpha}\right\rangle$, pak hypotézu H_0 nezamítáme a současně hypotézu H_1 zamítáme na hladině významnosti α. Nebo pokud testujeme hypotézu $H_0 : \theta = \theta_0$

proti jednostranné hypotéze $H_1 : \theta < \theta_0$, a pokud $t \in \overline{W}_\alpha = \langle \widehat{t}_\alpha; \infty \rangle$, pak hypotézu H_0 nezamítáme a současně hypotézu H_1 zamítáme na hladině významnosti α.

5 REGRESNÍ ANALÝZA

Teoretické výsledky této kapitoly jsou podloženy zdroji [11], [17], [19].

Ve statistice důležitou roli hraje hledání, zkoumání a hodnocení závislostí proměnných, jejichž hodnoty získáme při realizaci experimentu. Podle charakteru proměnných dostáváme náhodný vektor $\mathbf{X}$ nezávisle proměnných (regresorů) $X_1, \ldots, X_k$ a závisle proměnné (regresanty, responze) $Y_1, \ldots, Y_n$. Náhodný vektor $\mathbf{X}$ může být i nenáhodný (časté v aplikacích) nebo rozptyly všech složek $X_1, \ldots, X_k$ jsou zanedbatelné vůči rozptylu náhodné veličiny Y. Nástrojem pro popis a vyšetřování závislosti Y na $\mathbf{X}$ je *regresní analýza*. Tuto závislost lze vyjádřit ve tvaru

$$Y = f(X_1, \ldots, X_k) + e.$$

Člen e v modelu zastupuje *náhodnou chybu* reprezentující odchylku od aproximace. Funkce f se nazývá *regresní funkce*, která může nabývat mnoha podob. Podle typu regresní funkce rozeznáváme dva typy regresních modelů, *lineární regresní model* (lineární vzhledem k parametrům):

$$\begin{aligned} y &= a + bx \\ y &= a + bx + cx^2 \\ y &= a + \left(\frac{b}{x}\right) \end{aligned}$$

a *nelineární regresní model* (nelineární postavení parametrů):

$$\begin{aligned} y &= a \cdot x^b \\ y &= a \cdot e^{bx} \\ y &= a \cdot e^{\frac{k}{x}}. \end{aligned}$$

Nejčastější vztah mezi proměnnými je lineární a model je pak tvaru

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + e.$$

$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)^T$ značí vektor parametrů, tzv. *regresních koeficientů*. Tyto koeficienty je potřeba z dostupných dat aproximovat.

Regresní analýza se skládá z několika kroků:

1. Specifikovat úlohu, neboli určit, jakou máme úlohu k řešení.
2. Vybrat proměnné, které by mohly mít vliv na závislou proměnnou, neboli vybrat nezávisle proměnné.

3. Shromáždit data a vytvořit *matici plánu* $\mathbf{X}$ o n řádcích a k sloupcích.
4. Specifikování modelu patří k nejdůležitější části regresní analýzy, protože nevhodně zvolený model může vést k zavádějícím výsledkům.
5. K odhadnutí regresních koeficientů se nejčastěji používá *metoda nejmenších čtverců*, která minimalizuje součet čtverců vzdáleností n bodů v $k+1$ rozměrném prostoru od přímky proložené tímto prostorem a tato proložená přímka reprezentuje výslednou regresní rovnici. Součet čtverců

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

   kde $\mathbf{e}$ je vektor reziduí, $\boldsymbol{\beta}$ je vektor regresních parametrů, $\mathbf{X}$ je matice nezávislých proměnných a $\mathbf{Y}$ je vektor závislé proměnné. Pro aplikaci metody nejmenších čtverců by měly být splněny některé požadavky. Mezi nejdůležitější podmínky patří:
   - regresní parametry mohou nabývat libovolných hodnot,
   - náhodné chyby mají normální rozdělení $N(0, \sigma^2)$, pokud není splněna podmínka nulovosti střední hodnoty, tak se absolutní člen posune; rozptyl by měl být konečný a konstantní,
   - náhodné chyby jsou vzájemně nekorelované.
6. Pokud zjistíme přesnou podobu regresní rovnice, tak můžeme pro každé pozorování vyjádřit *reziduum*, rozdíl skutečné hodnoty závisle proměnné a výsledku vypočteného modelu. Za pomoci reziduí a *reziduálního součtu čtverců (RSS)* můžeme odhadnout, jak je model správně sestaven a porovnat ho s jinými modely.
7. Shledáme-li model za dostatečně dobrý, pak jej můžeme použít k řešení úlohy.

Podstatou řešení regresní analýzy je stanovit nejlepší regresní model (zjistit matematickou rovnici, která popisuje závislost $\mathbf{Y}$ na $\mathbf{X}$), parametry modelu (určit nejlepší odhady parametrů β), statistickou významnost modelu (rozhodnout, zda nalezený model přispěje ke zlepšení odhadu závisle proměnné proti použití průměru) a výsledky dané modelem interpretovat z hlediska zadání.

5.1 Lineární regresní model

Mějme náhodné veličiny $Y_1, \ldots, Y_n$ a matici daných čísel $\mathbf{X} = (x_{ij})$ typu $n \times k$, kde $k < n$. Mějme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, o kterém předpokládáme, že platí

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

kde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ je vektor neznámých parametrů, tzv. *regresních koeficientů* a $\mathbf{e} = (e_1, \ldots, e_n)^T$ značí vektor náhodných chyb, který splňuje podmínky

$$\mathrm{E}\mathbf{e} = \mathbf{0}, \ \mathrm{var}\mathbf{e} = \sigma^2\mathbf{I}.$$

Dále $\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ představuje matici závisle proměnné a matice

$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$ představuje matici nezávisle proměnné nebo-li *matici plánu.*

Dále musí být splněny některé předpoklady:

1. $\mathbf{X}_{n\times k}$ je matice reálných čísel,
2. $h(\mathbf{X}) \leq k$, tj. $n \geq k$,
   - Pokud $h(\mathbf{X}) = k$, pak říkáme, že se jedn8 o model plné hodnosti,
   - Pokud $h(\mathbf{X}) < k$, pak říkáme, že model není plné hodnosti.
3. $\mathrm{E}\mathbf{e} = \mathbf{0}$, tj. $\mathrm{E}e_i = 0, \ i = 1, \ldots, n$, tedy náhodné chyby jsou systematické. Pak $\mathrm{E}\mathbf{Y} = \mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \mathbf{X}\boldsymbol{\beta} + \mathrm{E}\mathbf{e} = \mathbf{X}(\beta)$,
4. $\mathrm{var}\mathbf{e} = \sigma^2\mathbf{I}$, tj. náhodné chyby e_i, e_j jsou nekorelované pro $i \neq j$, $\mathrm{D}e_i = \sigma^2$, $i = 1, \ldots, n$. Tedy jedná se o nekorelované chyby s homogenním rozptylem. Rozptyl σ^2 je neznámý parametr, $\sigma^2 > 0$. Zřejmě $\mathrm{var}\mathbf{e} = \mathrm{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Uvedený model nazveme *lineární regresní model* a označíme $(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Dále budeme pracovat s lineárním regresním modelem plné hodnosti, tedy $h(\mathbf{X}) = k \leq n$. Navíc předpokládejme, že $k < n$, tj. matice $\mathbf{X}_{n\times k}$ má hodnost k, tedy má nezávislé sloupce.

Uvědomme si, že počet sloupečků matice $\mathbf{X}$ se musí rovnat počtu řádků matice $\boldsymbol{\beta}$. Pokud se požadují např. dva parametry β_0, β_1 a pokud máme změřena jen data typu (x_i, y_i), pak se matice $\mathbf{X}$ zkonstruuje tak, že se uměle vloží jeden sloupec se samými jedničkami a dostaneme

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}.$$

Úkolem lineární regrese je najít odhad vektoru regresních koeficientů $\boldsymbol{\beta}$, ozn. $\mathbf{b}$. Pro tyto účely byly vedle metody nejmenších čtverců vypracovány i jiné metody, např. metoda maximální věrohodnosti, minimalizace absolutní odchylky, minimalizace maximální chyby.

Pohlédneme-li na *metodu nejmenších čtverců* jako na minimalizaci účelové funkce, tak tento pohled můžeme řešit analyticky nebo algebraicky. Z algebraického hlediska

minimalizujeme výraz

$$S^2(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \to min,$$

čímž získáme vektor $\mathbf{b}$, který nazveme odhadem vektoru $\boldsymbol{\beta}$ metodou nejmenších čtverců. (Nejlepší nestranný odhad vektoru regresních koeficientů $\boldsymbol{\beta}$ je vektor $\mathbf{b}$ získaný *metodou nejmenších čtverců*, tedy minimalizací *reziduálního součtu čtverců*). Tedy

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

minimalizuje $S^2(\boldsymbol{\beta})$, zřejmě $\mathbf{b}$ je určeno jednoznačně v lineárním regresním modelu plné hodnosti. Odhad parametru $\boldsymbol{\beta}$ v lineárním regresním modelu plné hodnosti je dán řešením *normálních rovnic*

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}.$$

Výrazu

$$S_e = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T(\mathbf{Y} - \mathbf{X}\mathbf{b})$$

se říká *reziduální součet čtverců*, pomocí kterého můžeme vyčíslit odhad σ^2 neboli *reziduální rozptyl* s^2

$$s^2 = \frac{S_e}{n-k},$$

kde k je počet regresních koeficientů. A druhé odmocnině z reziduálního rozptylu se říká *reziduální směrodatná odchylka.*

Při vyhodnocování lineárních regresních modelů se můžeme setkat s pojmem *koeficient determinace* R^2 a *koeficient vícenásobné korelace* r. Koeficient determinace je definován jako

$$R^2 = 1 - \frac{S_e}{S_t} = 1 - \frac{S_e}{\sum(Y_i - \overline{Y})^2},$$

kde S_e je reziduální součet čtverců a S_t je *celkový součet čtverců* odchylek Y_i od $\overline{Y}$. Koeficient determinace numericky souvisí s výběrovým korelačním koeficientem $r^2_{X,Y}$, který je spočtený z dvojic x_i, Y_i,

$$R^2 = r^2_{X,Y}.$$

Koeficient determinace se často uvádí v procentech $100 \cdot R^2$ a udává procento variability.

V lineární regresní analýze se nejčastěji testuje, zda se některý z regresních koeficientů nerovná nějaké známe konstantě, např $\beta = 0$.

Pro testování shody regresního koeficientu s konstantou se za testovací statistiku při nulové hypotéze $H_0 : \beta = 0$ volí statistika T s rozdělením $t(n-k)$,

$$T = \frac{b_1}{\sqrt{var(b_1)}},$$

pokud $|T| \geq t_{1-\frac{\alpha}{2}}(n-k)$ se H_0 zamítá na úrovni významnosti α.

5.2 Základní regresní modely

V této části se podíváme na některé základní regresní modely.

Přímka procházející počátkem

Uvažujme model

$$Y_i = \boldsymbol{\beta} x_i + e_i,\ i = 1, \ldots, n,$$

kde předpokládáme, že $e_i \sim N(0, \sigma^2)$. Pro případ přímky procházející počátkem se vektor $\boldsymbol{\beta}$ skládá z jediného prvku β_1 a matice $\mathbf{X}$ je velikosti $n \times 1$,

$$\mathbf{X} = (x_1, \ldots, x_n)^T.$$

Ze vztahu $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ určíme odhad β_1,

$$b_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

Podobně ze vztahu $s^2 = \frac{S_e}{n-k}$ vypočteme odhad reziduálního rozptylu

$$s^2 = \frac{S_e}{n-k} = \frac{(\mathbf{Y} - \mathbf{Xb})^T(\mathbf{Y}\text{-}\mathbf{Xb})}{n-k} = \frac{\sum_{i=1}^n Y_i^2 - \beta_1 \sum_{i=1}^n x_i Y_i}{n-1}.$$

Budeme-li chtít testovat hypotézu o hodnotě parametru β_i, tedy hypotézu $H_0 : \beta_i = a$, zejména $a = 0$, pak použijeme statistiku

$$T = \frac{b-a}{s\sqrt{\sum x_i^2}},$$

která má rozdělení $t(n-1)$. Pak hypotézu H_0 zamítáme na hladině α v případě, že $|T| \geq t_{1-\frac{\alpha}{2}}(n-1)$.

Obecná přímka

Jedná se o obecnější model

$$Y_i = \beta_0 + \beta_1 x_i + e_i,\ i = 1, \ldots, n,$$

pro tento model lze psát

$$\boldsymbol{\beta} = (\beta_0, \beta_1)^T$$

a matici $\mathbf{X}$ pak píšeme ve tvaru

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Označme výběrové průměry

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Pak

$$\mathbf{X}\mathbf{X}^T = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{pmatrix},\ \mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}$$

a dostaneme odhad $\mathbf{b}$ vektoru $\boldsymbol{\beta}$

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \overline{Y} - b_1\overline{x} \\ \frac{\sum x_i Y_i - n\overline{x}\overline{Y}}{\sum x_i^2 - n\overline{x}^2} \end{pmatrix}$$

a reziduální rozptyl

$$s^2 = \frac{\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i}{n-2}.$$

Zamyslíme-li se nad významem koeficientu b_1, tak zjistíme, že se jedná o vážený průměr směrnic všech přímek, které prochází pozorovanými body (x_i, Y_i) a těžištěm bodů $(\overline{x}, \overline{Y})$, přičemž váha každého bodu roste se zvětšující se vzdáleností $|x_i - \overline{x}|$. Díky tomu zjistíme, že odlehlé body mohou velmi hrubě zatížit odhad regresního parametru.

Stejně jako v předchozím případě můžeme testovat, jestli závislá veličina Y závisí na nezávislé veličině X, nebo-li jestli $\beta_1 = 0$. Pro testování nulové hypotézy $H_0 : \beta_1 = 0$ se používá statistika T ve tvaru

$$T_1 = \frac{b_1 - \beta_1}{s}\sqrt{\sum x_i^2 - n\overline{x}^2}$$

neboli

$$T_1 = \frac{b_1}{s}\sqrt{\sum x_i^2 - n\overline{x}^2} \sim t(n-2).$$

Nulovou hypotézu zamítáme na hladině významnosti α, pokud $|T_1| \geq t_{1-\frac{\alpha}{2}}(n-2)$.

Interval spolehlivosti pro závislou veličinu Y se zkonstruuje jako

$$b_0 + b_1 x \pm t_{1-\frac{\alpha}{2}}(n-2)s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum x_i^2 - n\overline{x}^2}},$$

který s pravděpodobností $1-\alpha$ překrývá hodnotu $\beta_0 + \beta_1 x$. Protože dopředu nevíme, pro které hodnoty se má interval spolehlivosti vyčíslit, tak se počítají hodnoty pro všechna $x \in [\min x_i, \max x_i]$. Pokud x probíhá daný interval, tak vypočtené hodnoty vytváří kolem regresní přímky dvě větve hyperboly, mezi nimiž leží *pás spolehlivosti* pro predikovanou závisle proměnnou. Pás spolehlivosti zaručí překrytí jedné hodnoty $\beta_0 + \beta_1 x$ s pravděpodobností $1-\alpha$. Lze odvodit i *pás spolehlivosti pro regresní přímku*, který překrývá celou přímku s danou pravděpodobností. Tento pás je obecně širší, i když rozdíly nejsou velké.

5.3 Mnohonásobná lineární regrese

Pro případ mnohonásobné lineární regrese můžeme vycházet z rovnice

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

a výpočet regresních parametrů se zredukuje na maticové operace.

5.4 Testování hypotéz

Můžeme testovat hypotézu o shodě vektoru regresních koeficientů (kromě absolutního členu) se známým vektorem,

$$H_0 : \mathbf{b} = \boldsymbol{\beta}$$

oproti alternativě, že H_0 alespoň pro jednu složku neplatí. Mezi nejčastější testy hypotéz patří testování významnosti parametru β, kdy se známý vektor položí roven nule, $\boldsymbol{\beta} = \mathbf{0}$. Tento test je shodný s testem nezávislosti lineárního regresního modelu, $H_0 : R^2 = 0$ oproti alternativě $H_1 : R^2 > 0$. Testovací statistika F_e se testuje proti hodnotě $F_{p-1,n-p}(\alpha)$, kde p zastupuje počet regresních parametrů a F_e je definována jako

$$F_e = \frac{(n-p)R^2}{(1-R^2)(p-1)}.$$

Pomocí t-testu testujeme jednotlivé parametry, $H_0 : b_i = \beta_i$ proti $H_1 : b_i \neq \beta_i$. Často parametry β_i testujeme na významnost, tedy $\beta_i = 0$. Testujeme testovací kritérium tvaru

$$t_i = \frac{|b_i - \beta_i|}{\sqrt{s^2(\mathbf{X}^T\mathbf{X})^{-1}}}$$

proti kritické hodnotě $t_{1-\frac{\alpha}{2}}(n-p)$. Pokud vyčíslíme a vyhodnotíme tyto testy, tak mohou nastat tyto případy:

- F-test vyjde nevýznamný společně se všemi t-testy. Pak se model považuje za nevhodný, neboť nevystihuje variabilitu $\mathbf{Y}$,
- F-test a všechny t-testy vyjdou významné. Pak se model považuje za vhodný, ale nezaručí, že je model přijatelný a správný,
- F-test vyjde významný, ale t-testy vycházejí nevýznamné u několika regresních parametrů. Pak se model považuje za vhodný a pokud je to nutné, tak se provede vypuštění nevýznamných parametrů ve vazbě na výsledky multikolinearity,
- F-test vyjde významný, ale všechny t-testy jsou nevýznamné, model sice vyhovuje, ale žádný regresní parametr není významný, což bývá důsledkem kolinearity.

5.5 Bootstrap regresní model

Také metodou bootstrap můžeme odhadnout regresní koeficienty.
Uvažujme, že pro každé pozorování máme hodnoty závislých a nezávislých proměnných uloženy ve vektoru $\mathbf{z}_i$,

$$\mathbf{z}_i = (Y_i, X_{i1}, \ldots, X_{ik}),\ i = 1, \ldots, n.$$

Tak dostaneme n pozorování $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$, ze kterých výběrem s opakováním vytvoříme bootstrapový výběr $\mathbf{z}^*$, který má B množin $\mathbf{z}_b^*$ o velikosti n,

$$\mathbf{z}_b^* = (\mathbf{z}_{b1}^*, \mathbf{z}_{b2}^*, \ldots, \mathbf{z}_{bn}^*),\ b = 1, \ldots, B.$$

Tedy dostali jsme B bootstrapových množin $\mathbf{z}_{b1}^*, \mathbf{z}_{b2}^*, \ldots, \mathbf{z}_{bn}^*$ a pro každé pozorování $\mathbf{z}_b^*$ spočítáme odhad regresních koeficientů, tedy

$$\mathbf{b}_b^* = [A_b^*, B_{b1}^*, \ldots, B_{bk}^*]^T,\ b = 1, \ldots, B.$$

Tento způsob metody regrese můžeme aplikovat pro výpočet směrodatné odchylky nebo pro výpočet konfidenčních intervalů pro regresní odhady.

Výběr s opakováním $\mathbf{z}_i^{'}$ implicitně považuje regresory $X_1, \ldots, X_k$ za náhodné více než za závislé. Pokud bychom chtěli uvažovat X jako vázané, tj. pokud bychom data získali z experimentálního měření, pro případ lineární regrese budeme postupovat následovně:

- Mějme hodnoty nezávislé proměnné $\mathbf{X}$ a hodnoty závisle proměnné $\mathbf{Y}$. Pro tento původní soubor odhadneme regresní koeficienty $A_1, B_1, \ldots, B_k$ a dále spočteme rezidua E_i,

$$\widehat{Y}_i = A + B_1 x_{i1} + \cdots + B_k x_{ik}$$
$$E_i = Y_i - \widehat{Y}_i.$$

- Vygenerujeme B bootstrapových výběrů s opakováním $\mathbf{e}_b^*$ z reziduí E_i a z nich spočteme příslušné hodnoty $\mathbf{y}_b^*$,

$$\mathbf{e}_b^* = [E_{b1}^*, E_{b2}^*, \ldots, E_{bn}^*]^T$$
$$\mathbf{y}_b^* = [Y_{b1}^*, Y_{b2}^*, \ldots, Y_{bn}^*]^T,$$

  kde $Y_{bi}^* = \widehat{Y}_i + E_{bi}^*$ a $b = 1, \ldots, B$.
- Nyní pomocí hodnot $\mathbf{y}_b^*$ získáme bootstrap regresní koeficienty, například odhady spočteme pomocí metody nejmenších čtverců a pak

$$\mathbf{b}_b^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_b^*,$$

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

Pokud jsme zkonstruovali $\mathbf{b}_b^* = [A_b^*, B_{b1}^*, \ldots, B_{bk}^*]^T,\ b = 1, \ldots, B$, tak můžeme těchto hodnot využít pro metodu konstrukce směrodatné odchylky a pro konstrukci konfidenčních intervalů pro regresní koeficienty.

Regresní modely a podobné statistické modely mohou být bootstrapovány pomocí dvou náhledů,

1. pokud regresory jsou náhodné a máme bootstrapové výběry z pozorování $\mathbf{z} = [Y_i, X_{i1}, \ldots, X_{ik}]$

   nebo

2. pokud regresory jsou vázané a máme výběr z reziduí E_i regresního modelu. Bootstrapová pozorování se zkonstruují jako $Y_{bi}^* = \widehat{Y}_i + E_{bi}^*$, kde $\widehat{Y}_i$ jsou hodnoty z původní regrese a E_{bi}^* jsou rezidua b-tého bootstrapového výběru.

Nevýhodou vázaného X je, že procedura implicitně předpokládá funkční tvar regresního modelu, který by měl být správný a chyby rovnoměrně rozdělené.

5.6 Bootstrap metoda a regresní analýza v praxi

Uvažujme, že máme statistický soubor $(\mathbf{X}, \mathbf{Y})$ o $n = 10$ pozorováních, pro který budeme provádět regresní analýzu. Tedy uvažujme statistický soubor $(\mathbf{X}, \mathbf{Y})$, pro který sestrojíme regresní model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \text{ kde } i = 1, \ldots, n.$$

x	y
1	2,92494
2	4,68058
3	7,06106
4	11,29959
6	13,43328
7	14,45410
8	16,94145
9	19,27376
10	20,72832

Tab. 5.1: Původní soubor ($\mathbf{X}$,$\mathbf{Y}$)

Z původního souboru dle předchozí kapitolky vygenerujeme B bootstrapových souborů o velikosti n. V tomto případě se jedná o boostrapování dvojic.

V našem případě jsme vygenerovali 100 boostrapových souborů $(\mathbf{X}_i^*, \mathbf{Y}_i^*)$ o velikosti 10 vzorků, $i = 1, \ldots, B$.

x_1^*	y_1^*	x_2^*	y_2^*	x_3^*	y_3^*	x_4^*	y_4^*	x_5^*	y_5^*
10	20,728	1	2,925	4	9,319	7	14,454	3	7,061
10	20,728	7	14,454	6	13,433	2	4,681	8	16,941
8	16,941	5	11,300	9	19,274	5	11,300	2	4,681
7	14,454	6	13,433	6	13,433	1	2,925	3	7,061
5	11,300	5	11,300	4	9,319	6	13,433	10	20,728
6	13,433	8	16,941	9	19,274	9	19,274	3	7,061
7	14,454	5	11,300	2	4,681	10	20,728	9	19,274
4	9,319	1	2,925	7	14,454	3	7,061	3	7,061
6	13,433	6	13,433	9	19,274	9	19,274	2	4,681
9	19,274	1	2,925	4	9,319	8	16,941	4	9,319

$\vdots$

x_{96}^*	y_{96}^*	x_{97}^*	y_{97}^*	x_{98}^*	y_{98}^*	x_{99}^*	y_{99}^*	x_{100}^*	y_{100}^*
8	16,941	6	13,433	2	4,681	3	7,061	9	19,274
8	16,941	6	13,433	10	20,728	7	14,454	4	9,319
9	19,274	2	4,681	6	13,433	5	11,300	10	20,728
5	11,300	6	13,433	6	13,433	8	16,941	7	14,454
4	9,319	10	20,728	3	7,061	3	7,061	7	14,454
2	4,681	7	14,454	7	14,454	3	7,061	8	16,941
7	14,454	8	16,941	1	2,925	4	9,319	10	20,728
8	16,941	7	14,454	8	16,941	7	14,454	8	16,941
4	9,319	1	2,925	9	19,274	3	7,061	10	20,728
5	11,300	5	11,300	5	11,300	5	11,300	6	13,433

Regresní koeficienty původního a boostrapového souboru jsou:

Tab. 5.2: Regresní koeficienty původního souboru a boostrapového souboru

β_0	β_1
1,034	1,996

β_{0i}^*	β_{1i}^*
1,692	1,905
1,022	2,016
1,041	2,023
0,936	2,012
0,952	2,007

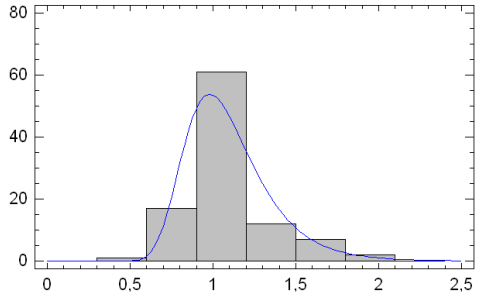
$\vdots$

β_{0i}^*	β_{1i}^*
1,131	1,986
1,068	1,984
1,001	2,004
1,453	1,906
1,468	1,928

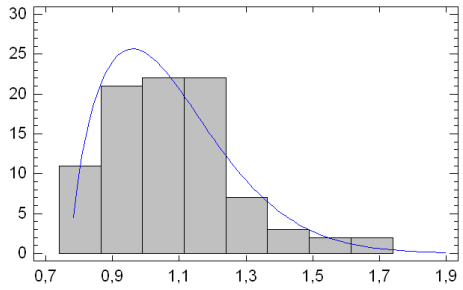
Lze metodou bootstrap odhadnout regresní koeficienty β_0^* β_1^* pomocí studentizovaného konfidenčního intervalu se spolehlivostí $1-\alpha$, $\left(\widehat{\theta} - t_{\frac{\alpha}{2}}^* \widehat{\sigma}^*; \widehat{\theta} + t_{1-\frac{\alpha}{2}}^* \widehat{\sigma}^*\right)$. Pro 90% interval spolehlivosti dostáváme:
$\beta_0^* \in \langle 0,789; 1,675 \rangle$
$\beta_1^* \in \langle 1,906; 2,047 \rangle$.
Stejným způsobem odhadneme i koeficient determinace R^{2*},
$R^{2*} \in \langle 0,992; 0,999 \rangle$.

Metodou bootstrap dostaneme i jiné parametry regresní analýzy. Nyní se podívejme na grafické znázornění našich výsledků za použití metody bootstrap pro regresní analýzu. Pro každý parametr, který jsme dostali, jsme jej vykreslili a snažili se najít jeho rozdělení pravděpodobnosti. Pro každý takto nalezený parametr jsme sestrojili 90% konfidenční interval a též jsme se dívali na rozdělení takto zkonstruované veličiny.

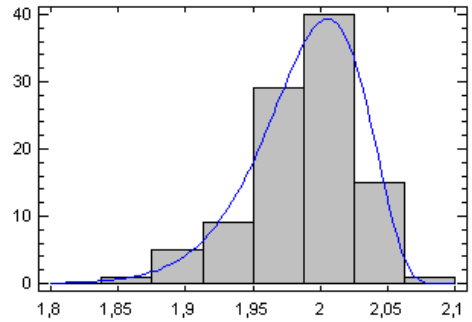
Na obrázku 5.1 vidíme vykreslené largest extreme value rozdělení pravděpodobnosti, které fituje rozdělení pravděpodobnosti odhadu parametru β_0^*. Náhodná veličina popisující sestrojený konfidenční interval pro odhad parametru β_0^*, obrázek 5.2, pochází z tří parametrického Weibullova rozdělení.



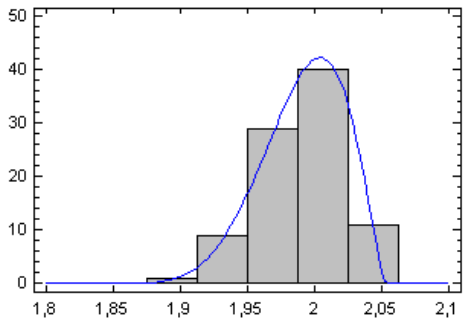
Obr. 5.1: Odhad parametru β_0^*



Obr. 5.2: Konfidenční interval pro odhad parametru β_0^*



Obr. 5.3: Odhad parametru β_1^*



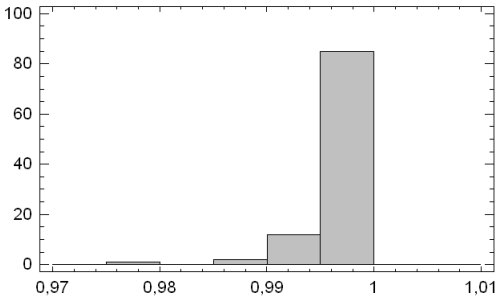
Obr. 5.4: Konfidenční interval pro odhad parametru β_1^*

Obrázek 5.3 znázorňuje čtyř parametrické beta rozdělení, které fituje rozdělení pravděpodobnosti odhadu parametru β_1^*. Náhodná veličina popisující sestrojený konfidenční interval pro odhad parametru β_1^*, obrázek 5.4, pochází také ze čtyř parametrického beta rozdělení.

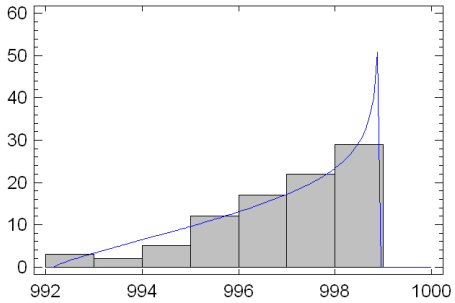
Rozdělení pravděpodobnosti, které fituje rozdělení pravděpodobnosti odhadu koeficientu determinace R^{2*}, obrázek 5.5, se nám nepodařilo najít. Ovšem náhodná veličina popisující sestrojený konfidenční interval pro odhad koeficientu determinace R^{2*}, obrázek 5.6, pochází z rozdělení čtyř parametrického beta rozdělení.

Rozdělení, která fitují rozdělení pravděpodobnosti odhadu koeficientu vícenásobné korelace r^*, obrázek 5.7,a rozdělení pravděpodobnosti náhodné veličiny popisující konfidenční interval odhadu koeficientu vícenásobné korelace r^*, obrázek 5.8, se nám nepodařilo najít.

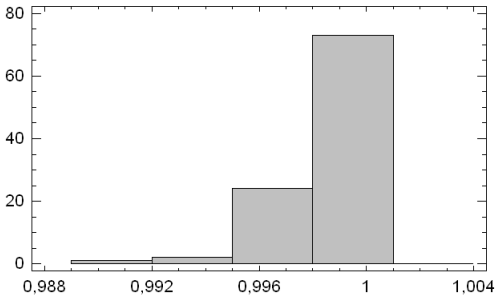
Na obrázku 5.9 je vykresleno čtyř parametrické beta rozdělení, které fituje rozdělení pravděpodobnosti odhadu chyby střední hodnoty. Náhodná veličina popisující sestrojený konfidenční interval pro odhad chyby střední hodnoty, obrázek 5.10, po-



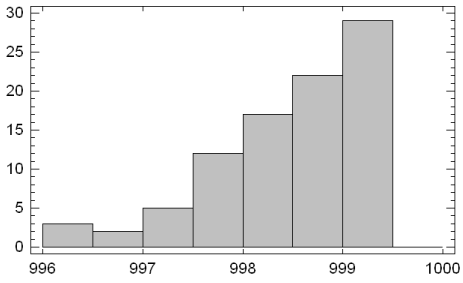
Obr. 5.5: Odhad koeficientu determinace R^{2*}



Obr. 5.6: Konfidenční interval pro odhad koeficientu determinace R^{2*}



Obr. 5.7: Odhad koeficientu vícenásobné korelace r^*



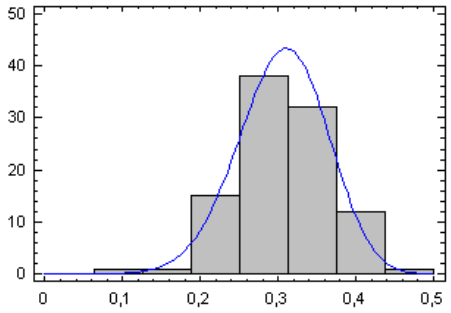
Obr. 5.8: Konfidenční interval pro odhad koeficientu vícenásobné korelace r^*

chází ze tří parametrického Weibullova rozdělení.

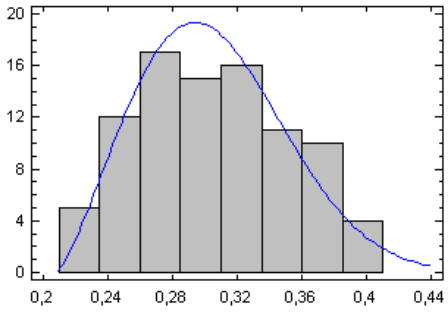
Protože jsme měli sestrojené horní a dolní meze regresních parametrů pomocí metody bootstrap, tak i pro tyto meze jsme se snažili najít rozdělení pravděpodobnosti.

Tří parametrické gama rozdělení, tří parametrické lognormální rozdělení a normální rozdělení, která fitují rozdělení pravděpodobnosti dolní meze parametru β_0^*, obrázek 5.11, se jeví téměř jako shodná na daném statistickém souboru. Náhodná veličina popisující sestrojený konfidenční interval pro dolní mez parametru β_0^*, obrázek 5.12, pochází ze čtyř parametrického beta rozdělení.

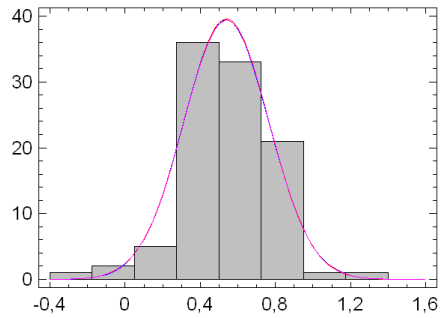
Čtyř parametrické beta rozdělení pravděpodobnosti, které fituje rozdělení pravděpodobnosti horní meze parametru β_0^*, je vykresleno na obrázku 5.13. Náhodná veličina popisující sestrojený konfidenční interval pro horní mez parametru β_0^*, obrázek 5.14, pochází také ze čtyř parametrického beta rozdělení.



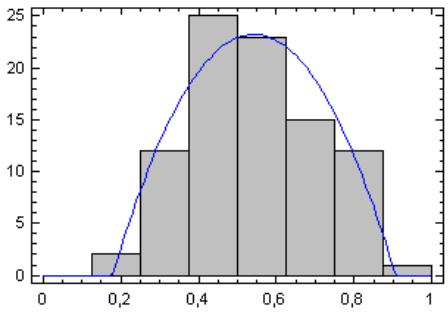
Obr. 5.9: Odhad chyby střední hodnoty



Obr. 5.10: Konfidenční interval pro odhad chyby střední hodnoty



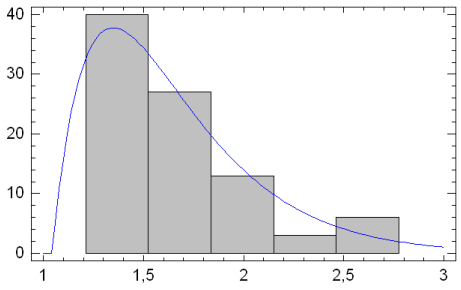
Obr. 5.11: Odhad dolní meze parametru β_0^*



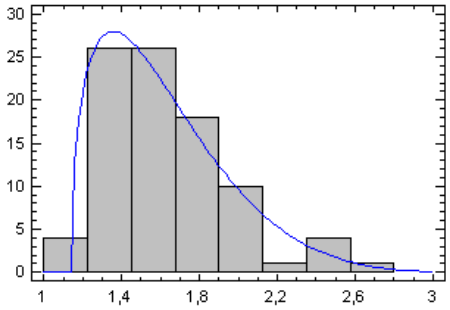
Obr. 5.12: Konfidenční interval pro odhad dolní meze parametru β_0^*

Jako nejlepší rozdělení, které fituje rozdělení pravděpodobnosti dolní meze parametru β_1^*, obrázek 5.15, se jeví tří parametrické Weibullovo rozdělení. Náhodná veličina popisující sestrojený konfidenční interval pro dolní mez parametru β_1^*, obrázek 5.16, pochází ze čtyř parametrického beta rozdělení.

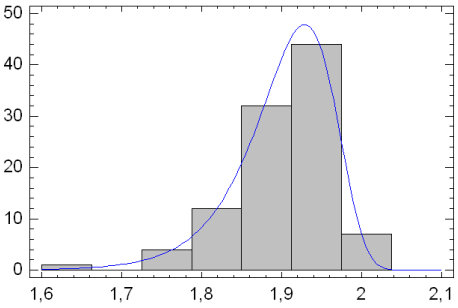
Nejlepším rozdělením, které fituje rozdělení pravděpodobnosti horní meze parametru β_1^*, obrázek 5.17, se jeví tří parametrické Weibullovo rozdělení. Náhodná veličina popisující sestrojený konfidenční interval pro horní mez parametru β_1^*, obrázek 5.18, pochází ze čtyř parametrického beta rozdělení.



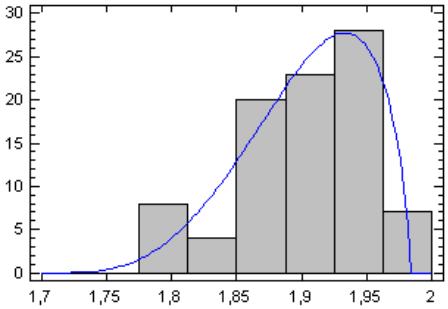
Obr. 5.13: Odhad horní meze parametru β_0^*



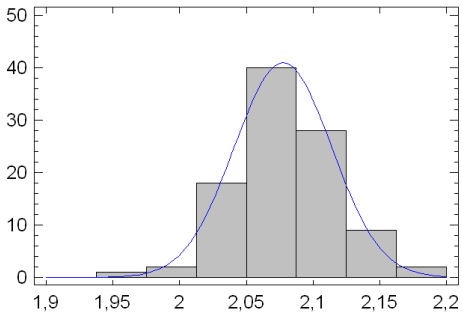
Obr. 5.14: Konfidenční interval pro odhad horní meze parametru β_0^*



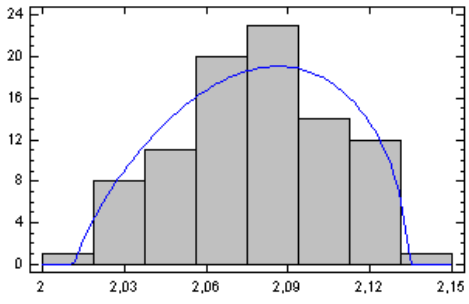
Obr. 5.15: Odhad dolní meze parametru β_1^*



Obr. 5.16: Konfidenční interval pro odhad dolní meze parametru β_1^*



Obr. 5.17: Odhad horní meze parametru β_1^*



Obr. 5.18: Konfidenční interval pro odhad horní meze parametru β_1^*

6 KONFIDENČNÍ INTERVAL PRO INDIVIDUÁLNÍ HODNOTU

6.1 Konfidenční interval pomocí regresní analýzy

Poznatky z této kapitoly jsou ze zdroje [11], [13].

Předpoklad lineární regresní analýzy je, že pozorovaná náhodná veličina Y má rozdělení pravděpodobnosti s podmíněnou střední hodnotou, která je daná lineární regresní funkcí

$$y = \sum_{1}^{m} = \beta_j f_j(\mathbf{x}),$$

kde $f_j(\mathbf{x})$ jsou známé funkce, které neobsahují regresní koeficienty $\beta_1, \ldots, \beta_m$. Při vyšetřování závislosti Y na $\mathbf{X}$ získáme realizací n experimentů vícerozměrný statistický soubor

$$((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$$

s rozsahem n, kde y_i je pozorovaná hodnota náhodné veličiny Y_i, $\mathbf{x}_i$ je pozorovaná hodnota vektoru nezávisle proměnných $\mathbf{X}$, $i = 1, \ldots, n$. Pomocí tohoto statistického souboru provádíme regresní analýzu, tedy počítáme potřebné odhady, testujeme hypotézy, verifikujeme daný model atd.

Při lineární regresní analýze, kdy hledáme lineární regresní funkci, aplikujeme tzv. *lineární regresní model*, který je založený na následujících předpokladech:

1. Vektor $\mathbf{x}$ je náhodný, tedy funkce nabývají nenáhodných hodnot $f_{ji} = f_j(\mathbf{x}_i)$ pro $j = 1, \ldots, m$ a $i = 1, \ldots, n$.
2. Matice $\mathbf{F} = \begin{pmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ n_{m1} & \cdots & n_{mn} \end{pmatrix}$ typu $(m \times n)$ s prvky f_{ji} má hodnost $m < n$.
3. Náhodná veličina Y_i má střední hodnotu $E(Y_i) = \sum_{j=1}^{m} \beta_j f_{ji}$ a konstantní rozptyl $D(Y_i) = \sigma^2 > 0$ pro $i = 1, \ldots, n$.
4. Náhodné veličiny Y_i jsou nekorelované a mají normální rozdělení pravděpodobnosti pro $i = 1, \ldots, n$.

Odhady regresních koeficientů, rozptylu, funkčních hodnot a testy statistických hypotéz o regresních koeficientech provádíme pomocí následujících vztahů, pro které si zavedeme označení matic:

$$\mathbf{H} = \mathbf{F}\mathbf{F}^T = \begin{pmatrix} \sum_{i=1}^{n} f_{1i} f_{1i} & \cdots & \sum_{i=1}^{n} f_{1i} f_{mi} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} f_{mi} f_{1i} & \cdots & \sum_{i=1}^{n} f_{mi} f_{mi} \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{g} = \mathbf{F}\mathbf{y} = \begin{pmatrix} \sum_{i=1}^n f_i y_i \\ \vdots \\ \sum_{i=1}^n f_{mi} y_i \end{pmatrix},$$

kde $\mathbf{F}^T$ označuje transponovanou matici $\mathbf{F}$. Pak platí:

1. Bodovým odhadem regresního koeficientu β_j je b_j, $j = 1, \ldots, m$ a matice $\mathbf{b}$ je řešení soustavy lineárních algebraických rovnic

$$\mathbf{H}\mathbf{b} = \mathbf{g},$$

   které se označují jako *soustavy normálních rovnic.*

2. Bodovým odhadem lineární regresní funkce je

$$y = \sum_{j=1}^m b_j f_j(\mathbf{x}).$$

3. Bodovým odhadem rozptylu σ^2 je

$$s^2 = \frac{S^*_{min}}{n-m},$$

   kde $S^*_{min} = \sum_{i=1}^n \left(y_i - \sum_{j=1}^m b_j g_j \right)^2 = \sum_{i=1}^n y_i^2 - \sum_{j=1}^m b_j g_j$ je minimální hodnota reziduálního součtu čtverců a g_j je prvek matice $\mathbf{g}$.

4. Intervalovým odhadem střední funkční hodnoty y se spolehlivostí $1-\alpha$ je

$$\left\langle \sum_{j=1}^m b_j f_j(\mathbf{x}) - t_{1-\frac{\alpha}{2}} s\sqrt{h^*}; \sum_{j=1}^m b_j f_j(\mathbf{x}) + t_{1-\frac{\alpha}{2}} s\sqrt{h^*} \right\rangle,$$

   kde $h^* = \mathbf{f}(\mathbf{x})^T \mathbf{H}^{-1} \mathbf{f}(\mathbf{x})$, kde $\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix}$ a $t_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})$-kvantil Studentova rozdělení s $n-m$ stupni volnosti. Intervalový odhad individuální funkční hodnoty y se spolehlivostí $1-\alpha$ se získá analogicky, avšak místo h^* se položí $1+h^*$.

Výše uvedené výsledky můžeme aplikovat pro odhad predikce hodnoty y pozorované náhodné veličiny Y s normálním rozdělením pravděpodobnosti $N(\mu, \sigma^2)$ a to pomocí statistického souboru $(y_1, \ldots, y_n)$, $n > 2$. My se budeme zabývat případem, kdy $m = 1$ a $f_1(\mathbf{x}) = 1$, jedná se o triviální konstantní lineární regresní funkci $y = \beta_1$. Pro tuto situaci dostaneme

$$\mathbf{H} = (n), \ \mathbf{b} = (b_1), \ \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}, \ \mathbf{g} = \begin{pmatrix} \sum_{i=1}^n y_i \end{pmatrix}.$$

Pak *bodovým odhadem individuální hodnoty* náhodné veličiny Y je

$$b_1 = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$$

a *bodovým odhadem rozptylu náhodné veličiny* Y je

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

V principu se jedná o hodnoty známých nestranných odhadů parametrů μ a σ. Po dosazení do intervalového odhadu individuální funkční hodnoty regresní funkce a příslušných úpravách dostaneme *intervalový odhad individuální hodnoty* náhodné veličiny Y se spolehlivostí $1-\alpha$,

$$\left\langle \overline{y} - t_{1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}}; \overline{y} + t_{1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}} \right\rangle,$$

kde $t_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})$-kvantil Studentova rozdělení s $n-1$ stupni volnosti.

6.2 Konfidenční interval pomocí tolerančních mezí

Druhý způsob odhadu individuální hodnoty je pomocí využití tolerančních mezí pro neznámou střední hodnotu normálního rozdělení. Nechť P je pokrytí a $1-\alpha$ spolehlivost, pak chceme nalézt takový interval, který bude se spolehlivostí $1-\alpha$ pokrývat alespoň $100P\%$ všech pozorování.

Nechť máme náhodný výběr $\mathbf{x} = (x_1, \ldots, x_n)$ z rozdělení $N(\mu, \sigma^2)$ s neznámými parametry μ, σ^2, toleranční meze volíme ve tvaru $\overline{x} \pm ks$, tj. jako funkce postačující statistiky $(\overline{x}, s^2)$.

Nejprve se podívejme na *jednostranné toleranční intervaly* $(-\infty, \overline{x} + ks)$ nebo $(\overline{x} - ks, \infty)$.

Pokud chceme nalézt konstanty k, tak uvažujme náhodné veličiny v a χ^2, které jsou nezávislé, nechť v má rozdělení $N(\delta, 1)$ a χ^2 má rozdělení $\chi^2(\nu)$. Pak náhodná veličina

$$t' = \frac{v}{\sqrt{\chi^2/\nu}}$$

má tzv. necentrální rozdělení t o ν stupních volnosti s parametrem necentrality δ a s hustotou pravděpodobnosti

$$f_\nu(t,\delta) = \frac{1}{2^{(\nu-1)/2}\Gamma(\nu/2)\sqrt{\pi}} e^{-\nu\delta^2/2(\nu+\delta^2)} \left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2} \times \int y^\nu e^{-(y-t\delta/\sqrt{\nu+t^2})^2/2} dy,$$

kde $-\infty < t < \infty$.

Necentrální rozdělení t s hustotou pravděpodobnosti $f_\nu(t,\delta)$ označíme $t'(\nu,\delta)$. Pro $\delta = 0$ přechází rozdělení $t'(\nu,\delta)$ na Studentovo rozdělení $t(v)$.

Označíme-li $100P\%$ kvantil rozdělení $t'(\nu,\delta)$ jako $t'_P(\nu,\delta)$, $0 < P < 1$, pak z hustoty $f_\nu(t,\delta)$ vyplývá, že $f_\nu(t,\delta) = f_\nu(-t,-\delta)$ pro každé reálné t, δ. Pak platí

$$t'_P(\nu,\delta) = t'_{1-P}(\nu,-\delta).$$

Pro pravostranný toleranční interval $(-\infty, \overline{x} + ks)$ je

$$z = P(x < \overline{x} + ks) = \Phi\left(\frac{\overline{x} + ks - \nu}{\sigma}\right),$$

tudíž pro dané $0 < P < 1$ je podmínka $z \geq P$ ekvivalentní podmínce

$$\frac{\overline{x} + ks - \nu}{\sigma} \geq u_P.$$

Dále chceme určit k tak, aby pro danou spolehlivost $1-\alpha$ platilo $P(z \geq P) = 1-\alpha$, tedy aby bylo splněno

$$P(\overline{x} + ks - \nu \geq \sigma u_P) = P\left(\frac{\overline{x} + ks - \sigma u_P}{s}\sqrt{n} \geq -k\sqrt{n}\right) = 1-\alpha.$$

Pak náhodné veličiny

$$v = \frac{(\overline{x} + ks - \sigma u_P)\sqrt{n}}{\sigma},$$
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

jsou nezávislé. Náhodná veličina v má rozdělení $N(-u_P\sqrt{n}, 1)$ a náhodná veličina χ^2 má rozdělení $\chi^2(n-1)$. Tudíž náhodná veličina

$$t' = \frac{\overline{x} + ks - \sigma u_P}{s}\sqrt{n}$$

má rozdělení $t'(n-1, u_P\sqrt{n})$ a z výše uvedených rovnic vyplývá, že konstanta k má tvar

$$k = -\frac{1}{\sqrt{n}}t'_\alpha(n-1, -u_P\sqrt{n}) = \frac{1}{\sqrt{n}}t'_{1-\alpha}(n-1, u_P\sqrt{n}).$$

Obdobně pro levostranný toleranční interval $(\overline{x} - ks, \infty)$ dostaneme

$$z = P(x > \overline{x} - ks) = 1 - \Phi\left(\frac{\overline{x} - ks - \nu}{\sigma}\right)$$

a pro dané P, $1-\alpha$ můžeme podmínku $P(z \geq P) = 1-\alpha$ vyjádřit ve tvaru

$$P(\overline{x} - ks - \nu \leq \sigma u_{1-P}) = P(\overline{x} - ks - \nu \leq -\sigma u_P) = P\left(\frac{\overline{x} + ks + \sigma u_P}{s}\sqrt{n} \leq k\sqrt{n}\right)$$
$$= P(t' \leq k\sqrt{n}) = 1-\alpha,$$

kde t' má rozdělení $t'(n-1, u_P\sqrt{n})$. Konstantu k dostaneme ve stejném tvaru,

$$k = -\frac{1}{\sqrt{n}} t'_\alpha(n-1, -u_P\sqrt{n}) = \frac{1}{\sqrt{n} t'_{1-\alpha}(n-1, u_P\sqrt{n})}.$$

Nyní se podívejme na *dvoustranné toleranční intervaly* $(\overline{x} - ks, \overline{x} + ks)$. Pro dvoustranné toleranční intervaly je

$$z = P(\overline{x} - ks < x < \overline{x} + ks) = \Phi\left(\frac{\overline{x} + ks - \mu}{\sigma}\right) - \Phi\left(\frac{\overline{x} - ks - \mu}{\sigma}\right).$$

Protože nás zajímá podíl rozdělení pokrytý intervalem $(\overline{x} - ks, \overline{x} + ks)$, tak můžeme volit μ, σ libovolné. Pro jednoduchost volme $\mu = 0$, $\sigma = 1$.

Pro dané $\overline{x}$ dostáváme rostoucí funkci veličiny s. Pro dané P, $0 < P < 1$, existuje jediná hodnota $r = ks$ taková, že

$$\Phi(\overline{x} + r) - \Phi(\overline{x} - r) = P,$$

přičemž podmínka $z \geq P$ je ekvivalentní podmínce $ks \geq r$, tudíž podmíněná pravděpodobnost

$$P(z \geq P|\overline{x}) = P\left\{\chi^2 = (n-1)s^2 \geq (n-1)\left(\frac{r}{k}\right)^2 |\overline{x}\right\} = P\left\{\chi^2 = \left(\frac{r}{k}\right)^2\right\},$$

protože náhodné veličiny $\overline{x}$, χ^2 jsou nezávislé.

Lze ukázat, že pro nepodmíněnou pravděpodobnost $P(z \geq P)$ platí vztah

$$P(z \geq P) = P\left(z \geq P|\frac{1}{\sqrt{n}}\right),$$

který platí již pro $n \geq 2$.

Jelikož náhodná veličina χ^2 má rozdělení $\chi^2(n-1)$, je podmínka $P(z \geq P) = 1-\alpha$ splněna pro

$$k = r\sqrt{\frac{n-1}{\chi^2_\alpha(n-1)}},$$

přičemž r určíme podle $\Phi(\overline{x} + r) - \Phi(\overline{x} - r) = P$ pro $\overline{x} = \frac{1}{\sqrt{n}}$, tedy řešíme rovnici

$$\Phi\left(\frac{1}{\sqrt{n}} + r\right) - \Phi\left(\frac{1}{\sqrt{n}} + r\right) = P.$$

V tabulkách najdeme hodnoty k pro různé hodnoty $1 - \alpha$, P, n.

7 PESIMISTICKÉ ODHADY ROZDĚLENÍ PRAVDĚPODOBNOSTI KATEGORIÁLNÍ VELIČINY

Teoretické výsledky uvedené v této kapitole jsou převzaty z článků [14],[15],[16] a zdroje [12]. Pro praktickou aplikaci ve druhém příkladu jsme pracovali s daty ze zdroje [18].

Základní praktickou úlohou při stochastickém modelování kategoriální veličiny X, která nabývá konečně mnoha různých hodnot x_j^*, $j = 1, \ldots, m$, kde $m \geq 2$, je odhad jejího rozdělení pravděpodobnosti z pozorovaných hodnot x_i, $i = 1, \ldots, n$, kde $n > m$. Zde je nutné podotknout, že označení hvězdičkou neznamená totéž označení jako při použití metody bootstrap. Proto z důvodu kolize značení budeme x_j^* značit odhad jejího rozdělení pravděpodobnosti z pozorovaných hodnot x_i a x_j^{**} značit j-tou realizaci bootstrapového výběru X^*. Nechť pozorováním $\mathbf{X}$ získáme statistický soubor $(x_1, \ldots, x_n)$ hodnot x_j^* a jeho roztříděním dostaneme roztříděný statistický soubor $\left(\left(x_1^*, \frac{f_1}{n}\right), \ldots, \left(x_m^*, \frac{f_m}{n}\right)\right)$, kde $\frac{f_j}{n} \neq 0$ je relativní četnost pozorované hodnoty x_j^*, $j = 1, \ldots, m$. Předpoklad nenulových relativních četností zajistíme, pokud vynecháme odpovídající hodnoty x_j^*. Označme odhadované rozdělení pravděpodobnosti $\mathbf{p} = (p_1, \ldots, p_m)$, kde $p_j = P(X = x_j^*)$ je pravděpodobnost, že kategoriální veličina X nabude hodnotu x_j^*, jedná se o odhad parametrů $\mathbf{p} = (p_1, \ldots, p_m)$ multinomického rozdělení pravděpodobnosti $M(n, p_1, \ldots, p_m)$ při známém n. Pokud byl statistický soubor $(x_1, \ldots, x_n)$ získán výběrem s vracením z vzájemně nezávislých pozorování $\mathbf{X}$, používá se většinou pro odhad vektor $\widehat{p} = \frac{f_1}{n}, \ldots, \frac{f_m}{n}$, který je nestranným odhadem vektoru parametrů $\mathbf{p} = (p_1, \ldots, p_m)$.

Ukážeme si odhady diskrétního rozdělení pravděpodobnosti kategoriální veličiny pomocí gradientu kvazinormy a tzv. přímkový odhad. Geometrickou interpretací přímkového odhadu rozumíme odhad ležící na úsečce, která jde z empirického rozdělení pozorovaných četností $\frac{\mathbf{f}}{n} = \frac{f_1}{n}, \ldots, \frac{f_m}{n}$ a končí v rozdělení $\mathbf{p}_0 = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$. Uvedené odhady jsou pro různé kvazinormy dostatečně vhodné pro aplikace a navíc můžeme vhodným postupem zajistit také jejich asymptotickou nestrannost.

7.1 Gradientní odhad

Nechť funkce $f : (0, \infty) \to \mathbb{R}^+$, kde $\mathbb{R}^+ = R \cup -\infty, \infty$, je konvexní na $(0, \infty)$, striktně konvexní v bodě $u = 1$ a nabývá v tomto bodě hodnoty $f(1) = 0$. Pokud $\mathbf{p} = (p_1, \ldots, p_m)$, resp. $\mathbf{q} = (q_1, \ldots, q_m)$ je diskrétní rozdělení pravděpodobnosti z pravděpodobnostního prostoru (Ω, Σ, P), resp. (Ω, Σ, Q), pak *f-divergencí* rozdě-

lení p, q rozumíme funkcionál

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

f-divergence má význam vzdálenosti daných rozdělení a platí:

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

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

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

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

Hledáme takové rozdělení pravděpodobnosti v S, které je nejblíže $\mathbf{p}_0$ a ke kterému se dostaneme od empirického rozdělení nejrychleji. Tomu odpovídá minimalizace kvazinormy $D_f(\mathbf{p},\mathbf{p}_0)$ a hledání rozdělení na křivce největšího spádu v S.

Nechť $D_f(\mathbf{p},\mathbf{p}_0)$ je kvazinorma na S. Pak *gradientním odhadem* rozdělení pravděpodobnosti $\mathbf{p} \in S$ z empirického rozdělení $\left(\frac{f_1}{n},\ldots,\frac{f_m}{n}\right)$ rozumíme rozdělení pravděpodobnosti $\mathbf{p}(t) \in S$, že

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}(t) = -\mathrm{grad}D_f(\mathbf{p}(t),\mathbf{p}_0)\ \forall t \in (0,\infty) \text{ a } \mathbf{p}(0) = \frac{\mathbf{f}}{n} = \left(\frac{f_1}{n},\ldots,\frac{f_m}{n}\right).$$

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

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

s počátečními podmínkami

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

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

Test dobré shody nám pomůže najít hodnotu $t_0 \in (0, \infty)$ jako hodnotu t, kdy ještě nezamítáme hypotézu o vhodnosti rozdělení $\mathbf{p}(t)$ na hladině významnosti α. Pro rostoucí parametr t se gradientní odhad $\mathbf{p}(t)$ vzdaluje po křivce největšího spádu S od empirického rozdělení k $\mathbf{p}_0$. Odhad $\mathbf{p}(t_0)$ je nejhorším z odhadů splňujících zvolené testové kritérium na hladině významnosti α, a proto se nazývá *pesimistický gradientní odhad*.

Nechť $f(u) = (u-1)^2$, pak

$$D_f(\mathbf{p}, \mathbf{p}_0) = \frac{1}{m} \sum_{j=1}^{m} (mp_j - 1)^2$$

je tzv. *kvadratická kvazinorma*. Složky gradientního odhadu $\mathbf{p}(t) = (p_1(t), \ldots, p_m(t))$ z empirického rozdělení $\left(\frac{f_1}{n}, \ldots, \frac{f_m}{n}\right)$ pro $\forall t \in (0, \infty)$ jsou partikulárním řešením nehomogenní lineární soustavy obyčejných diferenciálních rovnic prvního řádu s konstantními koeficienty a pravými stranami

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

a s počátečními podmínkami

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

a složka

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

Řešením této soustavy jsou složky gradientního odhadu $\mathbf{p}(t)$

kde

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

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

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

7.2 Přímkový odhad

Jiným odhadem rozdělení pravděpodobnosti $\mathbf{p}$ z pozorovaných hodnot náhodné kategoriální veličiny X v prostoru S může být přístup, kdy budeme uvažovat, že se nebudeme pohybovat po křivce největšího spádu jako u gradientního odhadu, ale po úsečce vycházející z empirického rozdělení pozorovaných četností $\frac{\mathbf{f}}{n} = \left(\frac{f_1}{n}, \ldots, \frac{f_m}{n}\right)$ a končící v rozdělení $\mathbf{p}_0 = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$. Pak odhad $\mathbf{p}(t)$ má složky

$$p_j(t) = \frac{f_j}{n} + \left(\frac{1}{m} - \frac{f_j}{n}\right) t,$$

kde $t \in \langle 0; 1\rangle$, $j = 1, \ldots, m$. Složky $p_j(t)$ odhadu $\mathbf{p}(t)$ rozdělení pravděpodobnosti $\mathbf{p}$ jsou zřejmě konvexní kombinace odpovídajících složek $\frac{\mathbf{f}}{n}$ a $\mathbf{p}_0$. Tento odhad nazveme *přímkový odhad*, který je totožný se známým diskrétním jádrovým odhadem s mocninnými jádry

$$\widehat{p_n}(x) = \frac{f_j}{n} \frac{1}{1+cm} + \frac{c}{1+cm} \text{ pro } c \in \langle 0, \infty).$$

Pokud vyjádříme složku $p_j(t)$ ve tvaru

$$p_j(t) = \frac{f_j}{n}(1-t) + \frac{1}{m} t,$$

pak platí

$$
\begin{array}{ccc}
\dfrac{1}{1+cm} = 1 - t & \Rightarrow & t = 1 - \dfrac{1}{1+cm} = \dfrac{cm}{1+cm} \\
\text{a} & & \\
\dfrac{c}{1+cm} = \dfrac{t}{m} & \Rightarrow & t = \dfrac{cm}{1+cm}.
\end{array}
$$

Gradientní odhad, resp. přímkový odhad, $\mathbf{p}(t)$ závisí na hodnotě $t \in \langle 0, \infty)$, resp. $t \in \langle 0, 1\rangle$. Hodnotu t_0 můžeme najít pomocí testu dobré shody. Pokud použijeme

Pearsonův test, pak t_0 je kořenem nelineární rovnice

$$\frac{1}{n}\sum_{j=1}^{m}\frac{f_j^2}{p_j(t)} - n = \chi^2_{1-\alpha}.$$

Při použití Pitmanova-Hellingerova testu je t_0 kořenem nelineární rovnice

$$8n\left(1 - \sum_{j=1}^{m}\sqrt{p_j(t)\frac{f_j}{n}}\right) = \chi^2_{1-\alpha}.$$

Pro oba případy $\chi^2_{1-\alpha}$ je $(1-\alpha)$-kvantil chí kvadrátu rozdělení s $m-1$ stupni volnosti, α je hladina významnosti testu dobré shody. Obě kritéria jsou asymptotická a pro praktické využití požadujeme $np_j(t_0) > 5$ pro $\forall j = 1, \ldots, m$.

Veškeré odhady $\mathbf{p}(t)$ pro $\forall t \in \langle 0; t_0\rangle$ splňují zvolené kritérium na hladině významnosti alespoň α. Odhad $\mathbf{p}(t_0)$ je "nejhorší"z těchto odhadů, tedy můžeme jej označit jako *pesimistický gradientní*, resp. *přímkový*, *odhad*.

7.3 Ukázka aplikace

Nyní si ukážeme, že pro získání intervalového odhadu můžeme použít metodu bootstrap. Praktickou aplikaci si předvedeme na dvou příkladech kategoriálních veličin, u nichž metodou bootstrap získáme intervalové odhady rozdělení pravděpodobnosti. Budeme se zabývat zejména přímkovým odhadem. K výpočtům nám poslouží MS Excel a Statgraphics Centurion.

Falešná kostka

Budeme uvažovat hrací kostku o šesti hranách, které si označíme klasickým způsobem $1, \ldots, 6$. Pokud házíme kostkou, tak pozorujeme diskrétní náhodou veličinu X, tedy číslo, které padne. Základní prostor je tvořen šesti elementárními náhodnými jevy odpovídající číslům $1, \ldots, 6$. Pokud se jedná o kostku, která není falešná, pak pravděpodobnostní funkce náhodné veličiny $\mathbf{p} = (p_1, \ldots, p_6) = (1/6, \ldots, 1/6)$. My jsme se zaměřili na případ, kdy uvažujeme falešnou kostku, tedy kostku, která má těžší stranu s číslem 6. Pravděpodobnostní funkci zvolíme

$$\mathbf{p} = (0,08; 0,13; 0,13; 0,13; 0,13; 0,4).$$

- Pozorování náhodné veličiny X nasimulujeme na počítači.
- Vygenerujeme 50 (B) náhodných bootstrapových výběrů $\mathbf{X}_i^*$, $i = 1, \ldots, B$, s opakováním o rozsahu n z původního souboru $\mathbf{X}$.
- Pro každý takto vygenerovaný bootstrapový výběr se spočítají četnosti f_i^* a relativní četnosti $\frac{f_i^*}{n}$.

- Nyní máme vše potřebné proto, abychom mohli použít Pearsonův test dobré shody.
- Tedy dostaneme B hodnot t_0^*, které je kořenem nelineární rovnice

$$\frac{1}{n}\sum_{j=1}^{m}\frac{f_j^{2*}}{p_j^*(t)} - n = \chi^2_{1-\alpha}.$$

- Pro těchto B hodnot zkonstruujeme bootstrapový t-konfidenční interval pro parametr θ na hladině významnosti $\alpha = 0,1$.

90% bootstrapové t-konfidenční intervaly pro pravděpodobnosti p_i^* a relativní četnosti $\frac{f_i}{n}^*$ příslušných stran kostky jsou:

$$\begin{array}{lcllcl}
p_1^* & \in & \langle 0,098; 0,152\rangle, & f_1 n^* & \in & \langle 0,02; 0,11\rangle,\\
p_2^* & \in & \langle 0,121; 0,166\rangle, & f_2 n^* & \in & \langle 0,07; 0,16\rangle,\\
p_3^* & \in & \langle 0,128; 0,177\rangle, & f_3 n^* & \in & \langle 0,08; 0,19\rangle,\\
p_4^* & \in & \langle 0,127; 0,175\rangle, & f_4 n^* & \in & \langle 0,08; 0,18\rangle,\\
p_5^* & \in & \langle 0,117; 0,168\rangle, & f_5 n^* & \in & \langle 0,07; 0,17\rangle,\\
p_6^* & \in & \langle 0,209; 0,358\rangle, & f_6 n^* & \in & \langle 0,33; 0,49\rangle.
\end{array}$$

Pokud bychom studovali, jak se chová náhodná veličina, která je složená z původní hodnoty a B bootstrapových hodnot, pak bychom dostali statistický soubor o velikosti $n = 51$, pro který zkonstruujeme 90% konfidenční interval pro individuální hodnotu $\theta \in \left\langle \overline{\theta} - t_{1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}}; \overline{\theta} + t_{1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}} \right\rangle$, kde θ postupně nahradíme $p_i,\ \frac{f_i}{n},\ i = 1,\ldots,6$. Takto vzniklé veličiny označíme $p_{iIH},\ \frac{f_{iIH}}{n},\ i = 1,\ldots,6$.

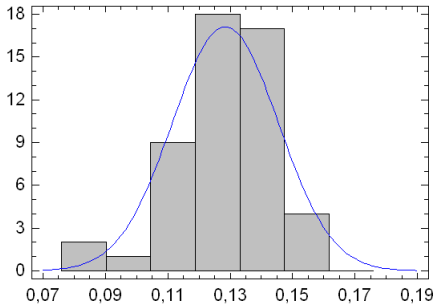
$$\begin{array}{lcllcl}
p_{1IH} & \in & \langle 0,100; 0,157\rangle, & \frac{f_{1IH}}{n} & \in & \langle 0,040; 0,118\rangle,\\
p_{2IH} & \in & \langle 0,123; 0,169\rangle, & \frac{f_{2IH}}{n} & \in & \langle 0,075; 0,164\rangle,\\
p_{3IH} & \in & \langle 0,126; 0,177\rangle, & \frac{f_{3IH}}{n} & \in & \langle 0,077; 0,190\rangle.\\
p_{4IH} & \in & \langle 0,127; 0,179\rangle, & \frac{f_{4IH}}{n} & \in & \langle 0,082; 0,191\rangle.\\
p_{5IH} & \in & \langle 0,123; 0,173\rangle, & \frac{f_{5IH}}{n} & \in & \langle 0,073; 0,175\rangle.\\
p_{6IH} & \in & \langle 0,199; 0,347\rangle, & \frac{f_{6IH}}{n} & \in & \langle 0,321; 0,495\rangle.
\end{array}$$

Sestrojíme také pro porovnání 90% konfidenční interval $\left\langle \overline{\theta} - t_{1-\frac{\alpha}{2}} s\sqrt{\frac{1}{n}}; \overline{\theta} + t_{1-\frac{\alpha}{2}} s\sqrt{\frac{1}{n}} \right\rangle$, kde θ postupně nahradíme $p_i,\ \frac{f_i}{n},\ i = 1,\ldots,6$. Takto vzniklé veličiny označíme $p_i^{'},\ \frac{f_i}{n}^{'},\ i = 1,\ldots,6$.

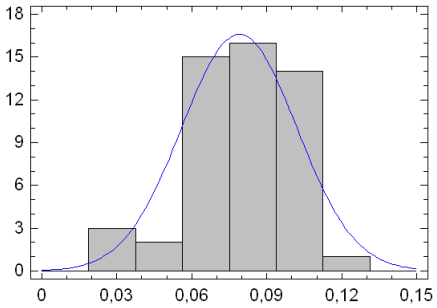
$$\begin{array}{lcllcl}
p_1^{'} & \in & \langle 0,124; 0,132\rangle, & \frac{f_1}{n}^{'} & \in & \langle 0,074; 0,084\rangle,\\
p_2^{'} & \in & \langle 0,143; 0,149\rangle, & \frac{f_2}{n}^{'} & \in & \langle 0,113; 0,126\rangle,\\
p_3^{'} & \in & \langle 0,148; 0,155\rangle, & \frac{f_3}{n}^{'} & \in & \langle 0,125; 0,141\rangle.\\
p_4^{'} & \in & \langle 0,149; 0,157\rangle, & \frac{f_4}{n}^{'} & \in & \langle 0,129; 0,144\rangle.\\
p_5^{'} & \in & \langle 0,145; 0,152\rangle, & \frac{f_5}{n}^{'} & \in & \langle 0,117; 0,131\rangle.\\
p_6^{'} & \in & \langle 0,263; 0,283\rangle, & \frac{f_6}{n}^{'} & \in & \langle 0,396; 0,420\rangle.
\end{array}$$

Při sestrojování konfidenčních intervalů v softwaru Statgraphics se objevilo podezření na normalitu dat. Podívejme se na grafické znázornění náhodných veličin. Chování náhodné veličiny je zachyceno na obrázcích 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7, 7.8, 7.9, 7.10, 7.11, 7.12.

Rozdělení, které fituje rozdělení náhodné veličiny p_1^*, je normální rozdělení, viz obrázek 7.1. Na obrázku 7.2 je zachyceno tří parametrické lognormální rozdělení pravděpodobnosti náhodné veličiny $\frac{f_1}{n}^*$.

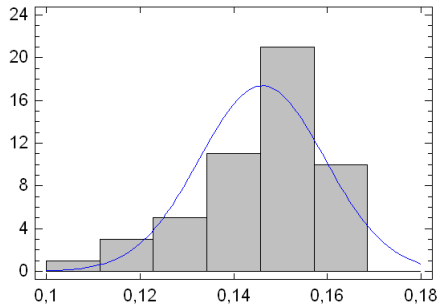


Obr. 7.1: Pravděpodobnostní funkce - strana 1

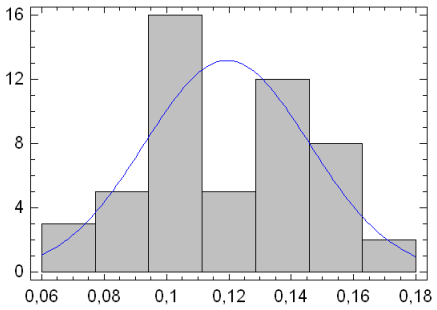


Obr. 7.2: Relativní četnosti - strana 1

Rozdělení, které fituje rozdělení náhodné veličiny p_2^*, je normální rozdělení pravděpodobnosti, viz obrázek 7.3. Na obrázku 7.4 je zachyceno normální rozdělení pravděpodobnosti náhodné veličiny $\frac{f_2}{n}^*$.



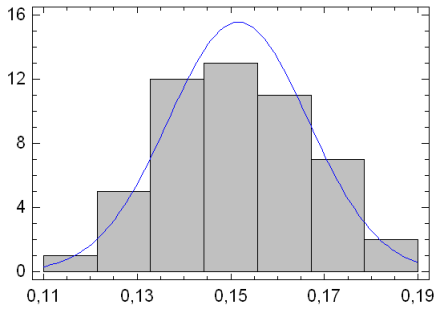
Obr. 7.3: Pravděpodobnostní funkce - strana 2



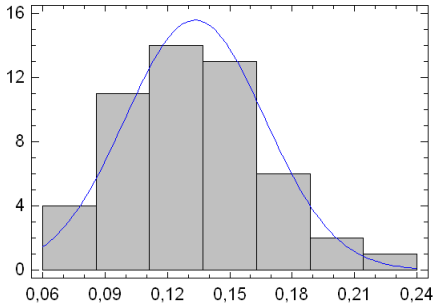
Obr. 7.4: Relativní četnosti - strana 2

Rozdělení, které fituje rozdělení náhodné veličiny p_3^*, je normální rozdělení pravděpodobnosti, viz obrázek 7.5. Na obrázku 7.6 je zachyceno normální rozdělení pravděpodobnosti náhodné veličiny $\frac{f_3}{n}^*$.

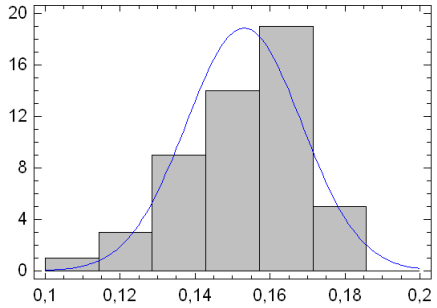
Rozdělení, které fituje rozdělení náhodné veličiny p_4^*, je normální rozdělení pravděpodobnosti, viz obrázek 7.7. Na obrázku 7.8 je zachyceno tří parametrické lognormální rozdělení pravděpodobnosti náhodné veličiny $\frac{f_4}{n}^*$.



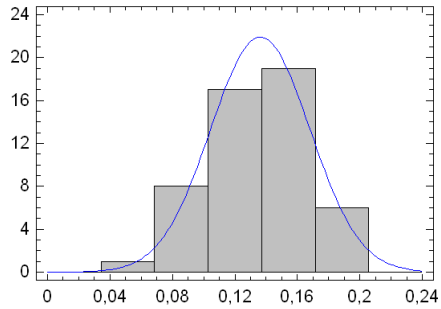
Obr. 7.5: Pravděpodobnostní funkce - strana 3



Obr. 7.6: Relativní četnosti - strana 3



Obr. 7.7: Pravděpodobnostní funkce - strana 4



Obr. 7.8: Relativní četnosti - strana 4

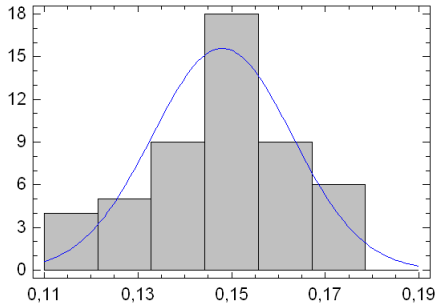
Rozdělení, které fituje rozdělení náhodné veličiny p_5^*, je normální rozdělení pravděpodobnosti, viz obrázek 7.9. Na obrázku 7.10 je zachyceno tří parametrické lognormální rozdělení pravděpodobnosti náhodné veličiny $\frac{f_5}{n}^*$.

Rozdělení, které fituje rozdělení náhodné veličiny p_6^*, je normální rozdělení pravděpodobnosti, viz obrázek 7.11. Na obrázku 7.12 je zachyceno normální rozdělení pravděpodobnosti náhodné veličiny $\frac{f_6}{n}^*$.

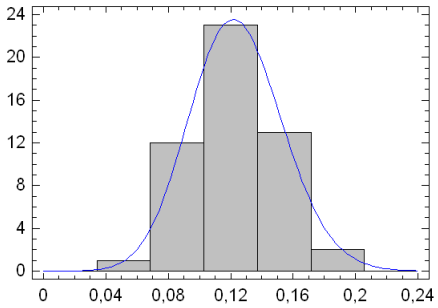
Tedy vidíme, že náhodné veličiny, které jsme dostali z původních a bootstrapových hodnot pochází z normálního nebo tří parametrického lognormálního rozdělení.

Z obrázku 7.13 je patrné, že pokud si pro ilustraci vykreslíme původní pravděpodobnosti, průměry bootstrapových výběrů a bootstrapové intervaly odhadneme průměrem jejich dolních a horních mezí, tak vidíme, že metodou bootstrap společně s přímkovými odhady se data snaží dostat k charakteru dat "obyčejné kostky". Snaží se vyrovnat k pravděpodobnosti $\frac{1}{6}$.

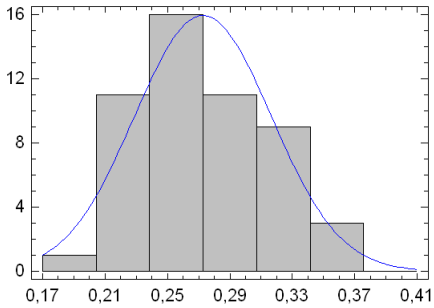
Na obrázku 7.14 vidíme vykreslené původní relativní četnosti, průměry bootstrapových výběrů a bootstrapové intervaly odhadneme průměrem jejich dolních a horních mezí.



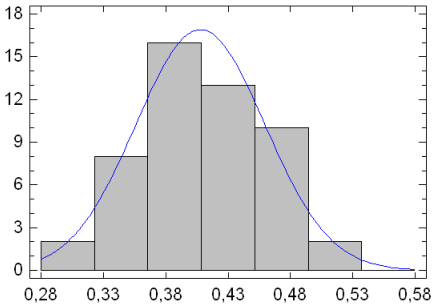
Obr. 7.9: Pravděpodobnostní funkce - strana 5



Obr. 7.10: Relativní četnosti - strana 5



Obr. 7.11: Pravděpodobnostní funkce - strana 6

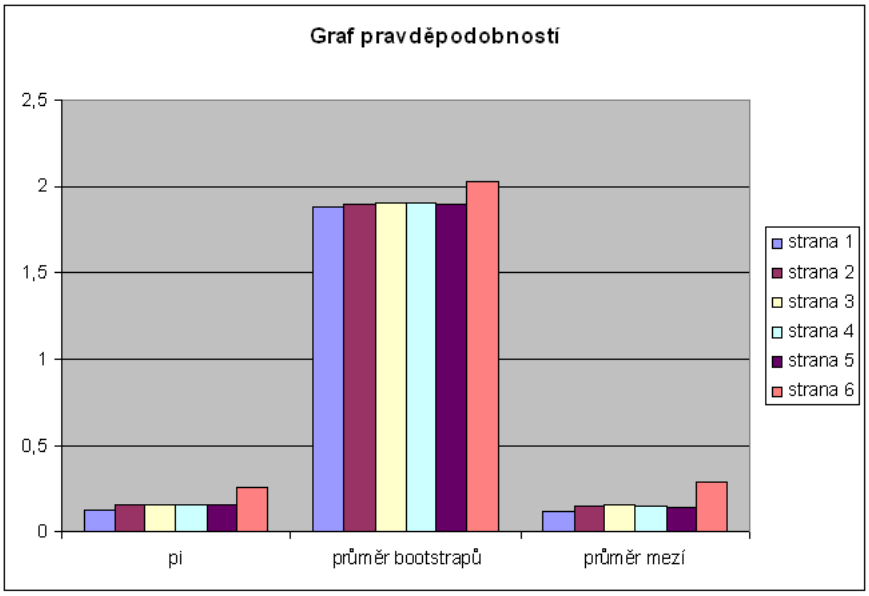


Obr. 7.12: Relativní četnosti - strana 6

Mobilita a místní přeprava v Chrudimi v roce 2011

Teď se podíváme na skutečnou situaci. V Chrudimi v roce 2011 proběhl dotazníkový průzkum na mobilitu a místní přepravu. Jednalo se o dotazníkový průzkum, kdy bylo vybráno 726 správně vyplněných dotazníků a zjistilo se, že ve vzorku respondentů převládá počet žen (55, 2 %) nad muži (44, 8 %). Vyhodnotilo se, které věkové skupiny byly nejvíce v průzkumu zastoupeny nebo jaký byl důvod cesty (z nabízených pěti možností) předchozího dne, atd. Zjišťoval se také použitý prostředek v den konání průzkumu, pěšky, kolo, autobus, MHD, vlak, motorka a osobní automobil. My jsme vycházeli z těchto výsledků, zejména jsme se zaměřili na kategorii, kdy respondenti použili osobní automobil a průzkum se dál ptal na počet osob v autě. Tedy z 278 respondentů jelo v autě 50 % jen řidič, 35 % řidič s jedním cestujícím, v 15 % jelo v autě více než 3 lidi. Grafické rozložení je zobrazeno na obrázku 7.15, 7.16.

Tedy zaplnění auta je kategoriální veličina X, která může nabývat tří hodnot, v autě byl pouze řidič, řidič s jedním cestujícím, více než 3 lidi. Pozorováním kategoriální veličiny X byl získán náhodný výběr o rozsahu 278. Na základě procentuálního



Obr. 7.13: Odhad pravděpodobností p_i^*

vyjádření si spočteme četnosti f_i a relativní četnosti $\frac{f_i}{n}$. No a nyní se podíváme, co se stane, pokud spojíme přímkový odhad a metodu bootstrap:

- Mějme náhodný výběr skládající se z pozorování kategoriální veličiny X.
- Vygenerujeme 60 (B) náhodných bootstrapových výběrů $\mathbf{X}_i^*$, $i = 1, \ldots, B$, s opakováním o rozsahu n z původního souboru $\mathbf{X}$.
- Pro každý takto vygenerovaný bootstrapový výběr se spočítají četnosti f_i^* a relativní četnosti $\frac{f_i^*}{n}$.
- Nyní máme vše potřebné proto, abychom mohli použít Pearsonův test dobré shody.
- Tedy dostaneme B hodnot t_0^*, které je kořenem nelineární rovnice

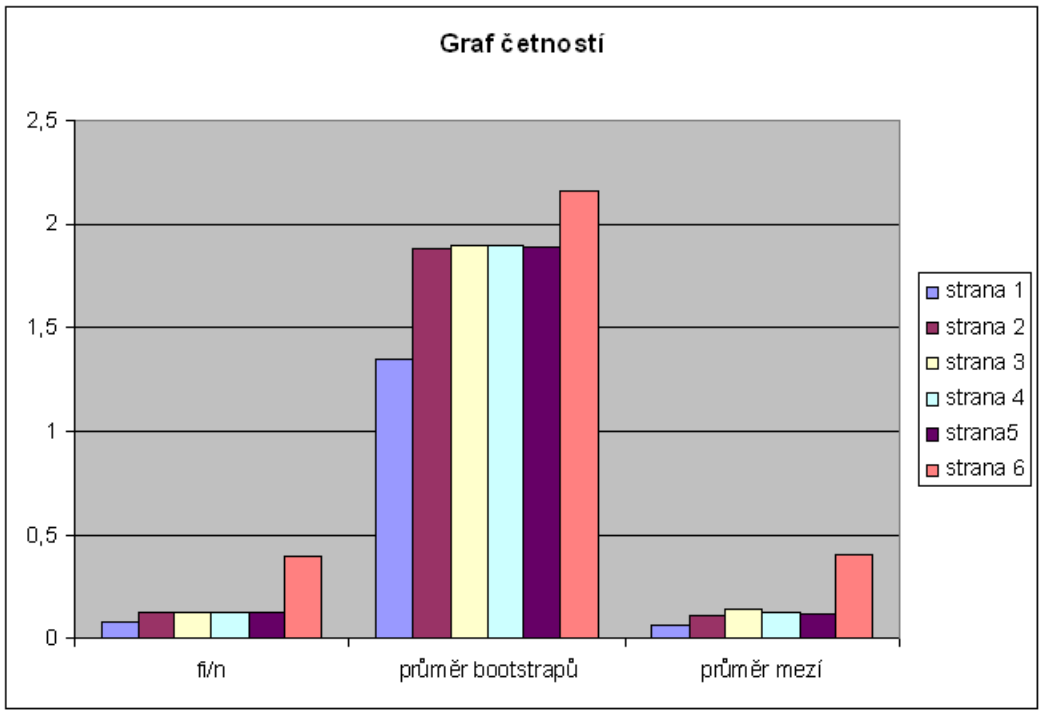
$$\frac{1}{n}\sum_{j=1}^{m}\frac{f_j^{2*}}{p_j^*(t)} - n = \chi^2_{1-\alpha}.$$

- Pro těchto B hodnot zkonstruujeme bootstrapový t-konfidenční interval pro parametr θ na hladině významnosti $\alpha = 0,1$.

90% bootstrapové t-konfidenční intervaly pro pravděpodobnosti p_i^* a relativní četnosti $\frac{f_i}{n}^*$ příslušného počtu osob v autě jsou:

$$p_1^* \in \langle 0,390; 0,451\rangle, \qquad \frac{f_1}{n}^* \in \langle 0,427; 0,529\rangle,$$
$$p_2^* \in \langle 0,320; 0,359\rangle, \qquad \frac{f_2}{n}^* \in \langle 0,306; 0,381\rangle,$$
$$p_3^* \in \langle 0,210; 0,278\rangle, \qquad \frac{f_3}{n}^* \in \langle 0,133; 0,198\rangle.$$

Pokud bychom studovali, jak se chová náhodná veličina, která je složená z původní hodnoty a B bootstrapových hodnot, pak bychom dostali statistický soubor o



Obr. 7.14: Odhad četností $\frac{f_i}{n}^*$

velikosti $n = 61$, pro který zkonstruujeme 90% konfidenční interval pro individuální hodnotu $\theta \in \left\langle \overline{\theta} - t_{1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}}; \overline{\theta} + t_{1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}} \right\rangle$, kde θ postupně nahradíme $p_{iIH},\ \frac{f_{iIH}}{n},\ i = 1,\ 2,\ 3.$

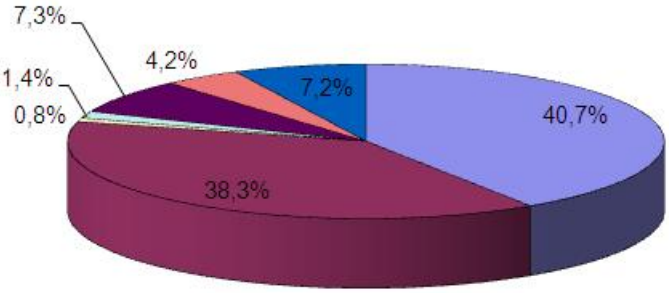
$$\begin{aligned} p_{1IH} &\in \langle 0,390; 0,458\rangle, & \frac{f_1 IH}{n} &\in \langle 0,444; 0,544\rangle, \\ p_{2IH} &\in \langle 0,317; 0,361\rangle, & \frac{f_{2IH}}{n} &\in \langle 0,303; 0,383\rangle, \\ p_{3IH} &\in \langle 0,202; 0,272\rangle, & \frac{f_{3IH}}{n} &\in \langle 0,123; 0,194\rangle. \end{aligned}$$

Sestrojíme také pro porovnání 90% konfidenční interval $\left\langle \overline{\theta} - t_{1-\frac{\alpha}{2}} s\sqrt{\frac{1}{n}}; \overline{\theta} + t_{1-\frac{\alpha}{2}} s\sqrt{\frac{1}{n}} \right\rangle$, kde θ postupně nahradíme $p_i^{'},\ \frac{f_i}{n}^{'},\ i = 1,\ 2,\ 3.$

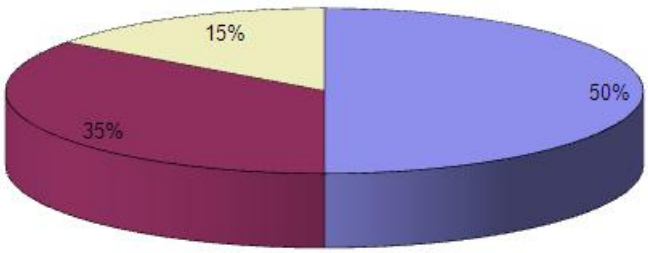
$$\begin{aligned} p_1^{'} &\in \langle 0,420; 0,429\rangle, & f_1^{'} &\in \langle 0,488; 0,500\rangle, \\ p_2^{'} &\in \langle 0,336; 0,342\rangle, & \frac{f_2}{n}^{'} &\in \langle 0,338; 0,348\rangle, \\ p_3^{'} &\in \langle 0,232; 0,241\rangle, & \frac{f_3}{n}^{'} &\in \langle 0,153; 0,163\rangle. \end{aligned}$$

Při sestrojování konfidenčních intervalů v softwaru Statgraphics se objevilo podezření na normalitu dat. Podívejme se na grafické znázornění náhodných veličin. Chování náhodné veličiny je zachyceno na obrázcích 7.17, 7.18, 7.19, 7.20, 7.21, 7.22.

Rozdělení, které fituje rozdělení náhodné veličiny p_1^*, je normální rozdělení, viz obrázek 7.17. Na obrázku 7.18 je zachyceno tří parametrické lognormální rozdělení pravděpodobnosti náhodné veličiny $\frac{f_1}{n}^*$.



Obr. 7.15: Způsob dopravy



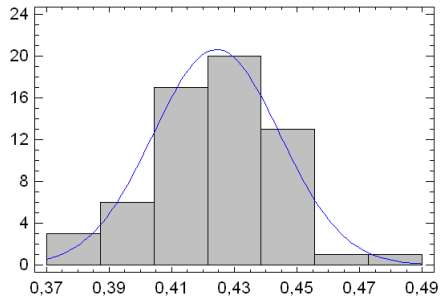
Obr. 7.16: Počet osob v autě

Rozdělení, které fituje rozdělení náhodné veličiny p_2^*, je normální rozdělení pravděpodobnosti, viz obrázek 7.19. Na obrázku 7.20 je zachyceno normální rozdělení pravděpodobnosti náhodné veličiny $\frac{f_2}{n}^*$.

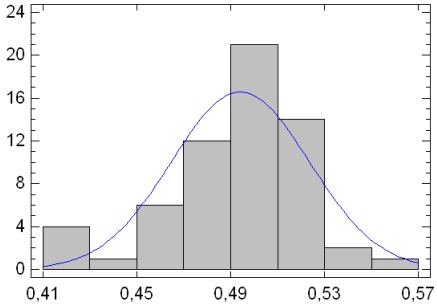
Rozdělení, které fituje rozdělení náhodné veličiny p_3^*, je tří parametrické lognormální rozdělení pravděpodobnosti, viz obrázek 7.21. Na obrázku 7.22 je zachyceno tří parametrické lognormální rozdělení pravděpodobnosti náhodné veličiny $\frac{f_3}{n}^*$.

Tedy vidíme, že náhodné veličiny, které jsme dostali z původních a bootstrapových hodnot pochází z normálního nebo tří parametrického lognormálního rozdělení.

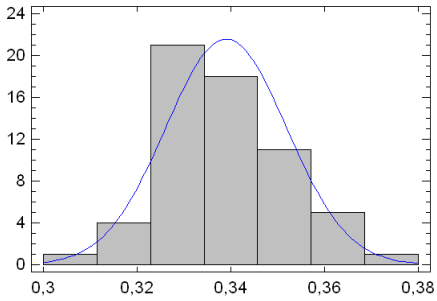
Z obrázku 7.23 je patrné, že pokud si pro ilustraci vykreslíme původní pravděpodobnosti, průměry bootstrapových výběrů a bootstrapové intervaly odhadneme průměrem jejich dolních a horních mezí, tak vidíme, že metodou bootstrap společně s přímkovými odhady si data zachovávají původní klesající charakter. Pro data s největším počtem respondentů, kteří jeli v autě sami, je sloupeček největší.



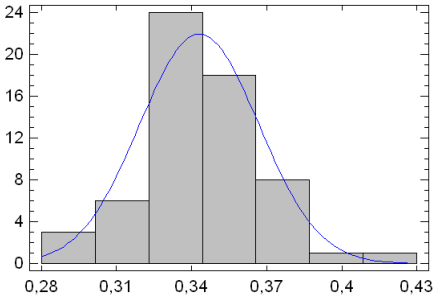
Obr. 7.17: Pravděpodobnostní funkce - pouze řidič



Obr. 7.18: Relativní četnosti - pouze řidič

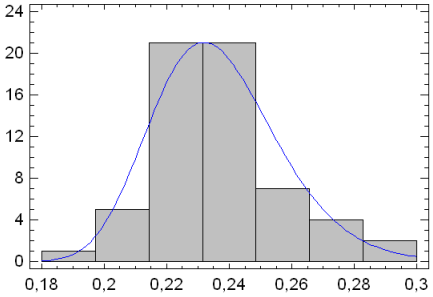


Obr. 7.19: Pravděpodobnostní funkce - řidič+1 cestující

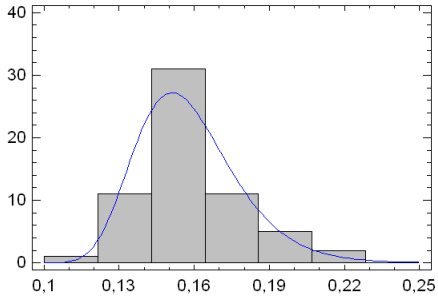


Obr. 7.20: Relativní četnosti - řidič+1 cestující

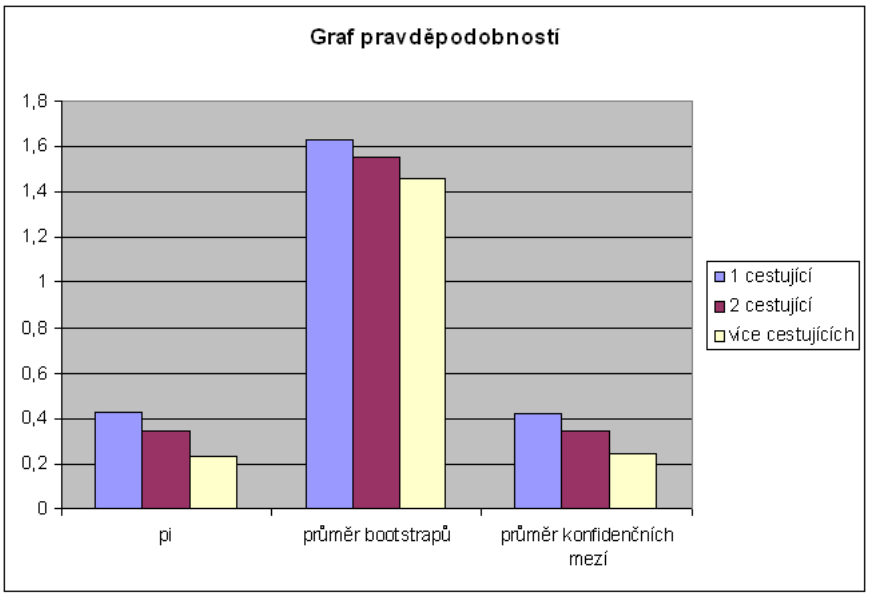
Na obrázku 7.24 vidíme vykreslené původní relativní četnosti, průměry bootstrapových výběrů a bootstrapové intervaly odhadneme průměrem jejich dolních a horních mezí.



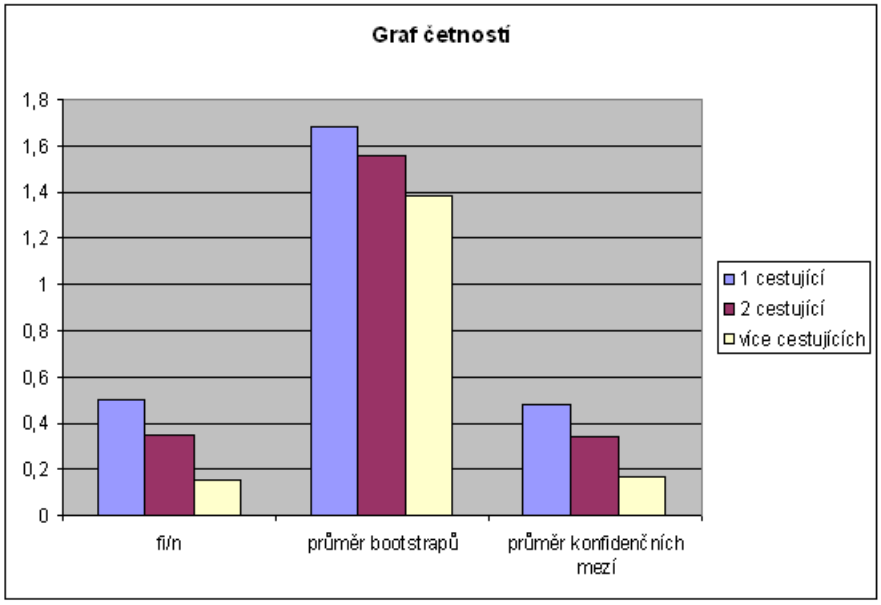
Obr. 7.21: Pravděpodobnostní funkce - více než 3 lidi v autě



Obr. 7.22: Relativní četnosti - více než 3 lidi v autě



Obr. 7.23: Odhad pravděpodobností p_i^*



Obr. 7.24: Odhad četností $\frac{f_i}{n}^*$

8 ZÁVĚR

Metoda bootstrap jistě patří mezi efektivní metody odhadů parametrů a rozdělení pravděpodobnosti. Lze ji použít při studování náhodného výběru o malém rozsahu, např.: $n = 20$. V době rozvoje počítačů není problém pomocí této metody odhadnout přesnost odhadu. Při použití metody bootstrap jsme odpoutáni od nutnosti studovat teorii do hloubky a odvozovat přesné vztahy. V případě, že analytické odvození neznáme, můžeme použít metodu bootstrap a dospějeme k řešení. Odhady odvozené neparametrickým bootstrapem jsou pro dostatečně velké výběry přesné bez přihlédnutí k pozorovanému rozdělení pravděpodobnosti. Často jsou dostatečně přesné už pro velmi malé rozsahy. I přesto zůstává, že metoda bootstrap dává jen hrubý odhad přesnosti, který by měl být použit v případě, že není možné realizovat rozsáhlejší výpočty.

Pokud spojíme pesimistické odhady s metodou bootstrap, tak získáme uspokojivé intervalové odhady pravděpodobnostní funkce. Při studiu kategoriální veličiny můžeme aplikovat gradientní odhad, který dává jen bodové odhady, ve spojení s metodou bootstrap dostaneme intervalové odhady. Pokud se budeme zabývat u kategoriální veličiny i přímkovým odhadem ve spojení s metodou bootstrap dostaneme intervalový odhad pravděpodobnostní funkce. Na datech z dotazníkového průzkumu o mobilitě ve městě Chrudim v roce 2011 jsme aplikovali přímkový odhad a metodu bootstrap, čímž jsme získali uspokojivé intervalové odhady četností a pravděpodobností.

Při použití počítačů se práce s touto metodou stává snazší. My jsme praktické výpočty a ověření prováděli v MS Excel a v softwaru Statgraphics Centurion XV. Statgraphics Centurion XV je statistický software, který má spoustu statistických nástrojů a umí spočítat i bootstrap a bootstrapové intervalové odhady střední hodnoty, směrodatné odchylky a mediánu. Nevýhodou tohoto softwaru je, že nevypisuje bootstrapové výběry. Tedy pracujeme s jinými daty než jsme si nabootstrapovali my, tudíž jsme Statgraphics Centurion XV používali pro ověření nebo v případě, kdy jsme nedělali sami bootstrapové výběry. Všechny soubory s výpočty a výstupy ze Statgraphics Centurion XV jsou na přiloženém CD.

LITERATURA

[1] MONTGOMERY, D. C., RENGER, G.: *Probability and Statistics.* New York: John Wiley & Sons, 1994.ISBN 978-047-1540-410.

[2] ANDĚL, J.: *Statistické metody.* Praha: MATFYZPRESS, 2007.ISBN 978-80-7378-003-2.

[3] ANDĚL, J.: *Základy matematické statistiky.* Praha: MATFYZPRESS, 2002.ISBN 978-80-7378-162-0.

[4] SILVERMAN, B. W.: *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall,1986.ISBN 978-0412-2246-203.

[5] VAJDA, I.: *Theory of Statistical Inference and Information.* London: Kluwer Academic Press, 1989.ISBN 978-902772-781-7.

[6] SCOTT, D.W.: *Multivariate Density Estimation. Theory, Practice and Visualization.* New York: Wiley, 1992.ISBN 978-047-1547-709.

[7] EFRON, B.:*Bootstrap Methods: Another Look at the Jackknife.* Annals of Statistics, Volume 7, Issue 1, 1979. 26 s.

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

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

[10] HALL, P.: *The Bootstrap and Edgeworth Expansion.* Springer-Verlag New York: Inc, 1992. ISBN 3-540-97720-1.

[11] KARPÍŠEK, Z.:: *Matematika IV - Statistika a pravděpodobnost.* Učební text. FSI VUT v CERM Brno, Brno 2002. ISBN 80-214-2055-3.

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

[13] ŠÁCHA, J.: *Netradiční statistické metody.* Brno:Vysoké učení technické v Brně, Fakulta strojního inženýrství, 200č. Vedoucí diplomové práce doc. RNDr. Zdeněk Karpíšek. CSc.

[14] KARPÍŠEK, Z., NERADOVÁ, V.:*Estimation of Categorical Variable Probability Distribution (Odhad rozdělení pravděpodobnosti kategoriální veličiny).* In: 7th Internati-onal Conference APLIMAT 2008. Bratislava, 5. - 8. 2. 2008, Book of abstracts p. 101, ISBN 978-80-89313-02-0, Proceedings pp. 1145-1154, ISBN 978-80-89313-03-7.

[15] LACINOVÁ, V., KARPÍŠEK, Z., SADOVSKÝ, Z.:*Pesimistické odhady rozdělení pravděpodob-nosti kategoriální veličiny.* Informační bulletin České statistické společnosti, roč. 22 (2), Praha, 2011, pp. 138-145. ISSN 1210-8022.

[16] KARPÍŠEK, Z., NERADOVÁ, V., ŽAMPACHOVÁ, E.:*A Contribution to the Estimation of Discrete Probability Distribution.* In: MENDEL 2008. 14th International Conference on Soft Computing. Brno, 18. - 20. 6. 2008, pp. 287-292, ISSN 1803-3814 (Mendel Journal Series on CD), ISBN 978-80-214-3675-6.

[17] http://www.sagepub.com/upm-data/21122_Chapter_21.pdf

[18] http://www.chrudim.eu/cs/download/zdrave-mesto/vysledky_a3_cr_2011.pdf

[19] http://fch.upol.cz/skripta/zzd/chemo/chemo.pdf

SEZNAM PŘÍLOH

Na přiloženém CD jsou následující přílohy diplomové práce:

1. bootstrap.xls
2. regrese.xls
3. KategorialniVelicina.xls
4. Složka Statgraphics obsahující reporty ze Statgraphics Centurion