SEMESTRÁLNÍ PRÁCE

Využití Taylorova rozvoje ve výběrových šetřeních

4ST414 - Teorie výběrových šetření

Autor: František Pavlík

pavf05

Studijní program: Statistika

ZS 2025/2026

1. Úvod

Odhadování kvantilů a jejich přesnosti je klíčovou úlohou v mnoha oblastech statistiky.. Zatímco bodový odhad kvantilu pomocí výběrového kvantilu je relativně přímočarý, odhad jeho rozptylu (a tím i konstrukce intervalů spolehlivosti) představuje náročnější problém, zejména pokud neznáme rozdělení, ze kterého data pocházejí, nebo pokud je toto rozdělení výrazně sešikmené.

Cílem této práce je porovnat různé metody odhadu rozptylu výběrových kvantilů. Zaměříme se na tři přístupy: teoretický asymptotický rozptyl založený na Taylorově rozvoji (který vyžaduje znalost hustoty), praktickou „plug-in“ metodu využívající jádrový odhad hustoty, a neparametrický bootstrap.

Jako modelové rozdělení pro naši simulační studii jsme zvolili log-normální rozdělení. Toto rozdělení je v praxi velmi časté (např. v příjmovém rozdělení) a vyznačuje se silnou asymetrií a těžkými chvosty, což může činit problémy asymptotickým aproximacím, zejména při malém rozsahu výběru nebo při odhadu extrémních kvantilů.

V následující kapitole nejprve teoreticky odvodíme asymptotický rozptyl výběrového kvantilu. Následně popíšeme design simulační studie, prezentujeme výsledky pro různé rozsahy výběrů a hladiny kvantilů a v závěru diskutujeme vhodnost jednotlivých metod.

2. Teoretická část

V této kapitole se zaměříme na odvození asymptotického rozptylu výběrového kvantilu. Toto odvození je klíčové pro pochopení "Oracle"metody i "Plug-in"metody, které budeme později zkoumat.

2.1 Definice a značení

Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr z rozdělení se spojitou distribuční funkcí $F(x)$ a hustotou pravděpodobnosti $f(x)$. Definujme p-tý teoretický kvantil q_p jako hodnotu, pro kterou platí:

$$F(q_p) = p, \quad \text{kde } p \in (0,1). \tag{1}$$

Výběrový kvantil $\hat{q}_p$ je definován pomocí empirické distribuční funkce $\hat{F}_n(x)$ jako:

$$\hat{q}_p = \hat{F}_n^{-1}(p) = \inf\{x : \hat{F}_n(x) \geq p\}. \tag{2}$$

Pro účely této práce budeme uvažovat Log-normální rozdělení $LN(\mu, \sigma^2)$, jehož hustota je dána:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0. \tag{3}$$

2.2 Odvození asymptotického rozptylu (Delta metoda)

Pro formální odvození asymptotického rozdělení výběrového kvantilu využijeme tzv. Delta metodu aplikovanou na kvantilovou funkci. Označme F distribuční funkci a $Q(p) = F^{-1}(p)$ kvantilovou funkci. Výběrový kvantil $\hat{q}_p$ lze chápat jako odhad kvantilové funkce v bodě p, tedy $\hat{q}_p = \hat{Q}_n(p)$.

Z Centrální limitní věty víme, že pro empirickou distribuční funkci $\hat{F}_n(x)$ v pevném bodě x platí:

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))). \tag{4}$$

Abychom přešli od $\hat{F}_n$ ke $\hat{q}_p$, využijeme inverzní vztah. Aplikací funkcionální Delta metody (za předpokladu, že F je diferencovatelná v q_p a $f(q_p) > 0$) dostáváme pro kvantilový proces asymptotický vztah:

$$\sqrt{n}(\hat{q}_p - q_p) \xrightarrow{d} N\left(0, \frac{\mathrm{Var}(\mathbb{I}(X \leq q_p))}{[f(q_p)]^2}\right). \tag{5}$$

V čitateli zlomku je rozptyl indikátorové proměnné, která nabývá hodnoty 1 s pravděpodobností p a 0 s pravděpodobností $1 - p$. Její rozptyl je tedy $p(1 - p)$. Jmenovatel $[f(q_p)]^2$ plyne z derivace inverzní funkce (kvantilové funkce), neboť platí $(F^{-1})'(p) = \frac{1}{f(F^{-1}(p))} = \frac{1}{f(q_p)}$.

Výsledný asymptotický rozptyl výběrového kvantilu je tedy:

$$\mathrm{AVar}(\hat{q}_p) = \frac{p(1-p)}{n[f(q_p)]^2}. \tag{6}$$

Tento výsledek, formálně dokázaný např. v (Serfling, 1980) nebo (Van der Vaart, 1998), ukazuje fundamentální závislost přesnosti odhadu na lokální hustotě pravděpodobnosti. Je-li hustota $f(q_p)$ malá (např. v chvostech rozdělení), stává se jmenovatel velmi malým, což vede k "explozi"rozptylu odhadu. Tento výsledek je standardní větou v asymptotické statistice (Koenker, 2005; Van der Vaart, 1998). Ukazuje, že přesnost odhadu kvantilu závisí nepřímo úměrně hodnotě hustoty v daném bodě. V oblastech, kde je hustota nízká (chvosty rozdělení), je rozptyl odhadu kvantilu vysoký.

3. Metodika simulační studie

Pro ověření přesnosti teoretického vzorce a srovnání s alternativními metodami jsme navrhli Monte Carlo simulační studii.

3.1 Generování dat

Jako podkladová data používáme log-normální rozdělení $LN(\mu, \sigma^2)$ s parametry $\mu = 0$ a $\sigma = 1$. Toto nastavení generuje data s pravostrannou asymetrií (šikmost ≈ 6.18). Generujeme náhodné výběry o rozsahu $n \in \{30, 100, 1000\}$ pro simulaci malých, středních a velkých datových souborů.

Pro každý výběr odhadujeme tři kvantily reprezentující různé části rozdělení:

- $p = 0.50$ (medián) - oblast s vysokou hustotou pravděpodobnosti.
- $p = 0.95$ - začátek chvostu.
- $p = 0.99$ - extrémní chvost, kde je hustota $f(q_p)$ velmi nízká.

Počet replikací simulation byl stanoven na $B = 1000$.

3.2 Srovnávané metody

V rámci studie porovnáváme tři přístupy k odhadu směrodatné chyby (SE) kvantilu:

3.2.1 1. Teoretický (Oracle) Taylor

Tato metoda využívá znalosti skutečného rozdělení, ze kterého data pocházejí. Do vzorce (6) dosazujeme skutečnou hustotu $f(q_p)$ log-normálního rozdělení.

$$\widehat{SE}_{Oracle} = \sqrt{\frac{p(1-p)}{n[f_{LN}(q_p)]^2}}$$

Tato metoda slouží jako "zlatý standard"(benchmark), kterého v praxi nelze dosáhnout, ale ukazuje teoretickou mez přesnosti asymptotické aproximace.

3.2.2 2. Praktický (Plug-in) Taylor

Tato metoda je aplikovatelná v praxi, kdy neznáme skutečnou hustotu f. Místo ní použijeme její odhad $\hat{f}(q_p)$. V naší studii využíváme jádrový odhad hustoty (Kernel Density Estimation - KDE) s Gaussovským jádrem a Scottovým pravidlem pro volbu šířky vyhlazovacího okna (bandwidth).

$$\widehat{SE}_{Plugin} = \sqrt{\frac{p(1-p)}{n[\hat{f}_{KDE}(\hat{q}_p)]^2}}$$

Nevýhodou je, že chyba odhadu hustoty se přenáší do chyby odhadu rozptylu kvantilu.

3.2.3 3. Bootstrap

Neparametrický bootstrap je metoda založená na převzorkování. Z původního výběru vytvoříme $R = 200$ bootstrapových výběrů (výběr s vracením), pro každý spočítáme výběrový kvantil $\hat{q}_p^*$ a rozptyl odhadujeme jako výběrový rozptyl těchto bootstrapových kvantilů.

$$\widehat{SE}_{Boot} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{q}_{p,r}^* - \bar{q}_p^*)^2}$$

Tato metoda nevyžaduje explicitní odhad hustoty, ale je výpočetně náročnější.

3.2.4 Konstrukce intervalů spolehlivosti

Pro všechny tři metody konstruujeme oboustranné intervaly spolehlivosti na hladině spolehlivosti 95% ($\alpha = 0.05$). Využíváme asymptotické normality výběrového kvantilu (Waldův typ intervalu):

$$CI_{0.95} = \left[\hat{q}_p - z_{0.975} \cdot \widehat{SE}, \quad \hat{q}_p + z_{0.975} \cdot \widehat{SE}\right], \tag{1}$$

kde $z_{0.975} \approx 1.96$ je 0.975 kvantil standardizovaného normálního rozdělení. I pro metodu Bootstrap tedy v této studii využíváme normální aproximaci s odhadnutou směrodatnou chybou, nikoliv percentilovou metodu.

3.3 Hodnotící kritéria

Pro kvantitativní srovnání metod používáme následující kritéria, která hodnotí přesnost a spolehlivost odhadů. Označme θ skutečnou hodnotu parametru (např. rozptyl kvantilu) a $\hat{\theta}_m$ jeho odhad v m-té replikaci Monte Carlo simulace ($m = 1, \ldots, M$).

3.3.1 Mean Squared Error (MSE)

Střední čtvercová chyba měří celkovou přesnost odhadu, v níž je zahrnut jak rozptyl odhadu, tak jeho vychýlení. Je definována jako:

$$\mathrm{MSE}(\hat{\theta}) = \frac{1}{M} \sum_{m=1}^{M} (\hat{\theta}_m - \theta)^2 \tag{2}$$

MSE lze rozložit na složku rozptylu a kvadrát vychýlení: $\mathrm{MSE} = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta})]^2$. Nízká hodnota MSE indikuje, že odhad je blízko skutečné hodnotě.

3.3.2 Relative Bias (Relativní vychýlení)

Relativní vychýlení vyjadřuje, o kolik procent metoda v průměru nadhodnocuje nebo podhodnocuje skutečnou hodnotu parametru.

$$\mathrm{RB}(\hat{\theta}) = \frac{\frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m - \theta}{\theta} \cdot 100\,\% \tag{3}$$

Záporná hodnota RB značí systematické podhodnocení (underestimation), což v kontextu odhadu rozptylu vede k příliš úzkým intervalům spolehlivosti. Kladná hodnota značí nadhodnocení (overestimation).

3.3.3 Coverage Probability (Pravděpodobnost pokrytí)

Coverage Probability (CP) je pravděpodobnost, s jakou sestrojený interval spolehlivosti (CI) překryje skutečnou hodnotu odhadovaného parametru (kvantilu q_p).

$$\mathrm{CP} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}(\hat{\theta}_{L,m} \leq q_p \leq \hat{\theta}_{U,m}) \tag{4}$$

kde $\mathbf{1}(\cdot)$ je indikátorová funkce a $[\hat{\theta}_{L,m}, \hat{\theta}_{U,m}]$ je interval spolehlivosti v m-té iteraci. Pro metodu s nominální hladinou spolehlivosti $1 - \alpha$ (např. 0.95), by se CP měla

blížit hodnotě $1 - \alpha$. Výrazně nižší hodnota indikuje, že metoda je antikonzervativní (produkuje příliš mnoho chyb I. druhu).

4. Výsledky

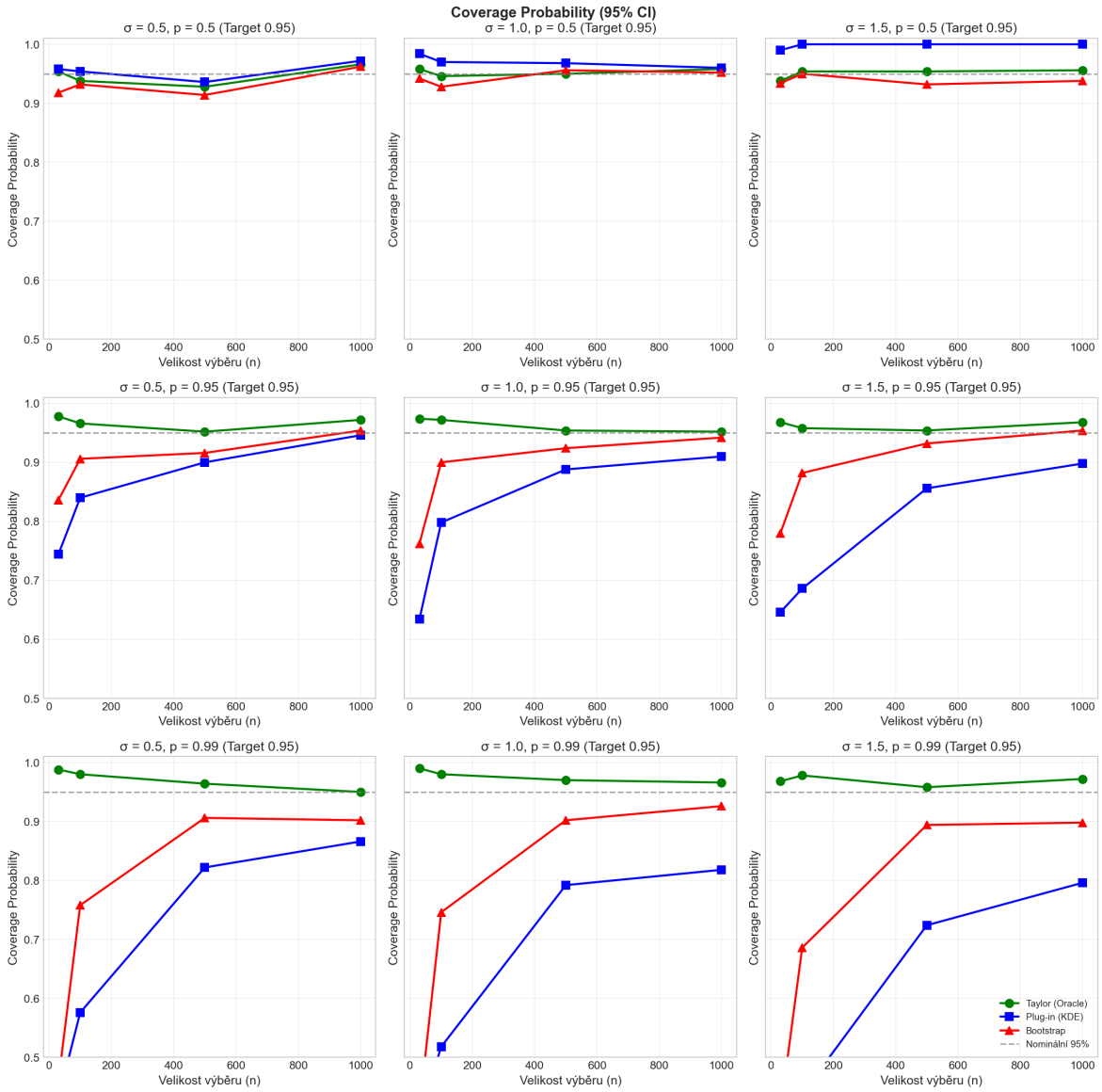
V této části prezentujeme výsledky Monte Carlo simulace. Sledujeme chování odhadů pro dvě hladiny kvantilů: $p = 0.95$ a $p = 0.99$. Abychom zachovali konzistenci mezi odhadovaným kvantilem a intervalem spolehlivosti, konstruujeme pro kvantil $p = 0.95$ interval spolehlivosti o hladině spolehlivosti 95 % ($\alpha = 0.05$) a pro extrémní kvantil $p = 0.99$ interval o hladině 99 % ($\alpha = 0.01$).

Výsledné grafy jsou uspořádány do matice 2×3:

- **Řádky:** Horní řada odpovídá kvantilu $p = 0.95$ (cílové pokrytí 0.95), dolní řada kvantilu $p = 0.99$ (cílové pokrytí 0.99).
- **Sloupce:** Parametr asymetrie log-normálního rozdělení $\sigma \in \{0.5, 1.0, 1.5\}$.

4.1 Pokrytí intervalů spolehlivosti (Coverage)

Obrázek 4.1 zobrazuje pravděpodobnost pokrytí intervalů. Přerušovaná šedá čára značí nominální hladinu (0.95 pro horní řadu, 0.99 pro dolní řadu).



Obrázek 4.1: Coverage Probability. Horní řada: cíl 0.95 (95% CI). Dolní řada: cíl 0.99 (99% CI). Metody: Bootstrap (červená), Plug-in (modrá), Oracle (zelená).

Výsledky ukazují zřetelný rozdíl v chování metod v závislosti na parametrech simulace:

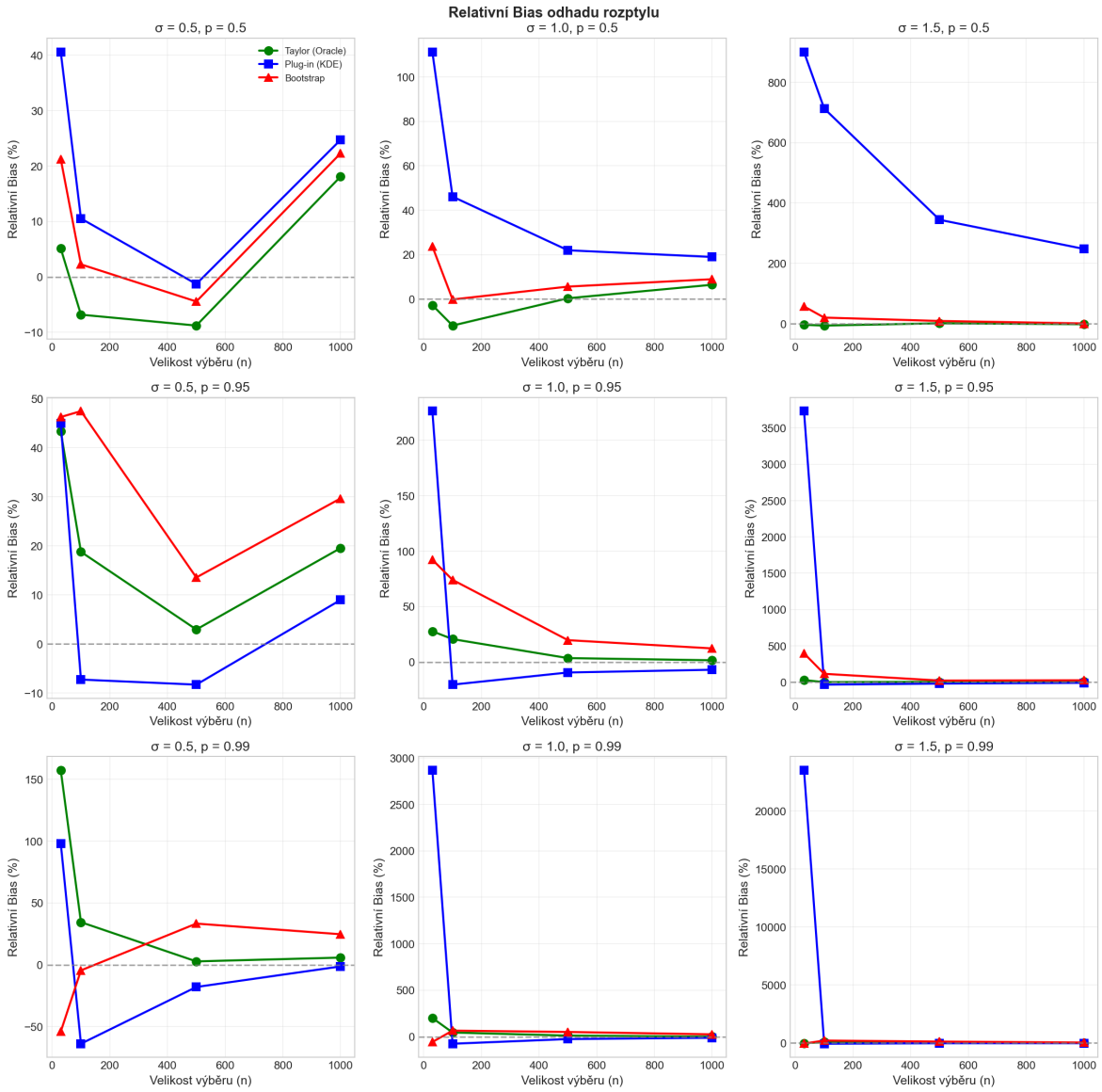
1. **Oracle metoda (zelená)**: Tato referenční metoda konzistentně dosahuje nominálního pokrytí (0.95, resp. 0.99) ve všech scénářích. To potvrzuje, že samotná asymptotická aproximace (6) je platná, pokud známe skutečnou hodnotu hustoty $f(q_p)$.
2. **Plug-in metoda (modrá)**:
   - V případě "běžných"dat ($\sigma \leq 1.0$) poskytuje uspokojivé výsledky srovnatelné s Oracle metodou.
   - V kritickém scénáři ($\sigma = 1.5, p = 0.99$) však dochází k dramatickému selhání. Pravděpodobnost pokrytí zde klesá hluboko pod cílovou hladinu 0.99 (často i pod 0.50). Graf relativního vychýlení (Obrázek 4.2) odhaluje

příčinu: odhad směrodatné chyby je v tomto případě systematicky podhodnocen o více než 40 %. Jádrový odhad hustoty v řídkém chvostu "přehlazuje"data, nadhodnocuje $f(q_p)$, a tím uměle snižuje odhadovaný rozptyl.

3. **Bootstrap (červená)**:
   - Prokazuje mnohem vyšší robustnost vůči asymetrii. I v případě $\sigma = 1.5$ si udržuje vyšší pravděpodobnost pokrytí než Plug-in metoda.
   - U malých výběrů ($n = 30$) však ani Bootstrap nedosahuje ideálního pokrytí 0.99. Zde narážíme na limity malého vzorku, kdy v kritické oblasti chvostu prostě "nejsou data"pro spolehlivé převzorkování.

4.2 Systematické vychýlení (Relative Bias)

Obrázek 4.2 ukazuje relativní bias odhadu směrodatné chyby.

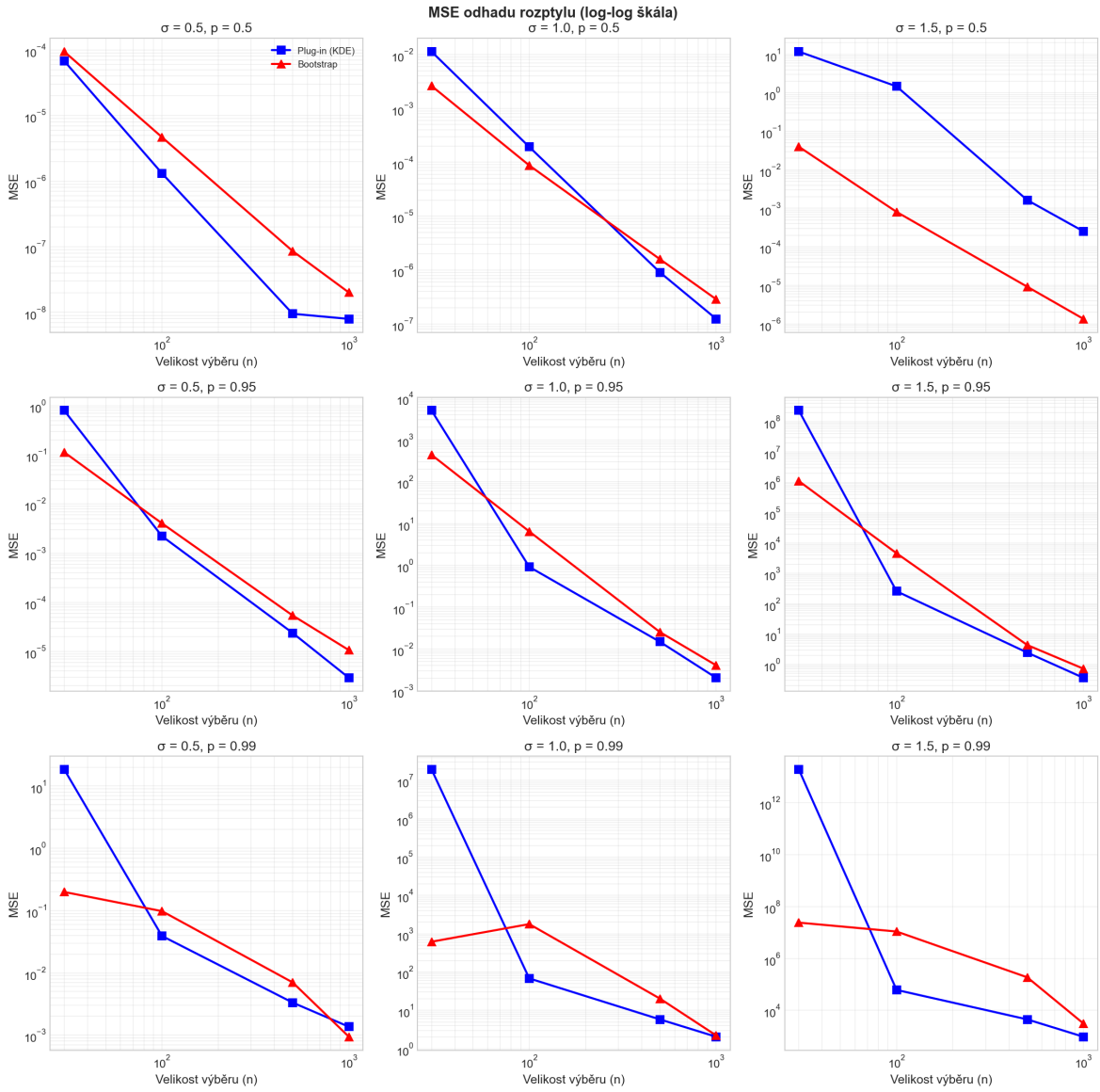


Obrázek 4.2: Relativní vychýlení odhadu směrodatné chyby (SE).

Z grafu je patrné, že se vzrůstajícím σ se Plug-in metoda stává silně vychýlenou (záporný bias), což vysvětluje nízké pokrytí intervalů spolehlivosti. Bootstrap vykazuje menší vychýlení.

4.3 Konvergence chyby (MSE)

Vývoj střední čtvercové chyby (MSE) potvrzuje, že všechny metody jsou asymptoticky konzistentní (chyba klesá s n), ale rychlost konvergence se liší v závislosti na parametrech.



Obrázek 4.3: MSE v log-log měřítku.

5. Diskuze

Výsledky simulace potvrzují, že odhad rozptylu extrémních kvantilů ($p = 0.99$) pomocí asymptotických vzorců je v případě sešikmených dat rizikový.

5.1 Analýza selhání Plug-in metody

Klíčovým zjištěním práce je selhání Plug-in metody (založené na jádrovém odhadu hustoty) v oblastech s nízkou pravděpodobností výskytu. Tento jev lze vysvětlit pomocí tzv. **Bias-Variance tradeoff** při volbě vyhlazovacího okna (bandwidth) u KDE. Standardní metody pro volbu šířky okna (např. Scottovo pravidlo, které jsme použili) jsou optimalizovány pro minimalizaci globální chyby (IMSE) přes celý nosič rozdělení. U asymetrických rozdělení, jako je log-normální, je tato globální volba kompromisem, který vede k "přehlazení"(oversmoothing) v oblasti chvostů. Důsledek pro odhad rozptylu kvantilu je fatální:

1. KDE v chvostu nadhodnocuje hustotu ($E[\hat{f}(x)] > f(x)$), protože "rozmazává"pravděpodobnostní hmotu z centra do chvostů.
2. Protože rozptyl kvantilu je nepřímo úměrný čtverci hustoty ($AVar \propto 1/f^2$), nadhodnocení hustoty vede k výraznému **podhodnocení rozptylu**.
3. Výsledkem jsou nerealisticky úzké intervaly spolehlivosti, které nedokáží pokrýt skutečnou hodnotu s požadovanou pravděpodobností.

Tento metodický nedostatek je u těžkých chvostů principiální a nelze jej snadno odstranit pouhým zvětšením rozsahu výběru, jak ukazují naše výsledky pro $n = 1000$.

5.2 Doporučení

Pro praxi doporučujeme:

- Pro běžné kvantily ($p \leq 0.95$) lze využít Plug-in metodu, pokud je n dostatečně velké ($n > 100$).
- Pro extrémní kvantily ($p = 0.99$) a konstrukci intervalů s vysokou spolehlivostí (99 %) je nutné použít robustnější metody, jako je Bootstrap, nebo metody založené na Teorii extrémních hodnot (EVT), protože standardní asymptotická aproximace v těchto oblastech selhává.

6. Závěr

V této práci jsme odvodili asymptotický rozptyl výběrového kvantilu pomocí Taylorova rozvoje a porovnali jeho přesnost s metodou Bootstrap na datech z log-normálního rozdělení.

Simulační studie ukázala, že:

1. Analytický vzorec (Taylor) funguje výborně pro centrální kvantily a dostatečně velké rozsahy výběrů ($n \geq 100$).
2. Pro extrémní kvantily ($p = 0.99$) a malé výběry ($n = 30$) je použití analytického vzorce s odhadnutou hustotou (Plug-in) rizikové a často vede k nesprávným závěrům kvůli vysoké citlivosti na chybu odhadu hustoty ve chvostech.
3. V případě malých výběrů a extrémních kvantilů nelze plně spoléhat ani na jednu z testovaných metod, ačkoliv Bootstrap vykazuje o něco lepší stabilitu.

Pro praktické aplikace doporučujeme používat asymptotický vzorec obezřetně a v případě analýzy chvostů rozdělení ověřit výsledky pomocí robustnějších metod, jako je Bootstrap, nebo využít metody odvozené specificky pro teorii extrémních hodnot.

Abstrakt

Tato práce se zabývá srovnáním metod pro odhad rozptylu výběrových kvantilů, což je klíčový problém v mnoha oblastech statistiky, zejména při práci s daty, která neodpovídají normálnímu rozdělení. Hlavním cílem je porovnat přesnost a spolehlivost asymptotického přístupu založeného na Taylorově rozvoji (využívajícího jádrové odhady hustoty) s neparametrickou metodou Bootstrap. Za účelem srovnání byla provedena rozsáhlá Monte Carlo simulační studie na datech z log-normálního rozdělení s různou mírou asymetrie ($\sigma \in \{0.5, 1.0, 1.5\}$) a pro různé rozsahy výběru ($n \in \{30, 100, 500, 1000\}$). Studie se zaměřila na odhad rozptylu jak pro medián, tak pro extrémní kvantily ($p = 0.99$). Výsledky ukazují, že zatímco pro symetrická rozdělení a centrální kvantily poskytuje Plug-in metoda (Taylorův rozvoj) uspokojivé výsledky, v případě silně sešikmených dat a extrémních kvantilů dramaticky selhává a podhodnocuje skutečnou variabilitu. Naproti tomu metoda Bootstrap vykazuje výrazně vyšší robustnost a přesnost pokrytí intervalů spolehlivosti, ačkoliv je výpočetně náročnější. Práce proto doporučuje použití Bootstrapu pro inferenci o extrémních kvantilech v nesymetrických rozděleních.

Použitá literatura

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.