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, od ekonomie přes biomedicínu až po technické aplikace. 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 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 rozptylu pomocí Taylorova rozvoje

Pro odvození asymptotického rozdělení $\hat{q}_p$ využijeme vztah mezi empirickou distribuční funkcí a kvantily. Vycházíme z toho, že $\hat{q}_p$ konverguje v pravděpodobnosti ke q_p (konzistence). Uvažujme Taylorův rozvoj funkce distribuční funkce F v bodě výběrového kvantilu $\hat{q}_p$ v okolí skutečného kvantilu q_p (**Serfling1980**):

$$F(\hat{q}_p) \approx F(q_p) + f(q_p)(\hat{q}_p - q_p) + O_p((\hat{q}_p - q_p)^2). \tag{4}$$

Z definice výběrového kvantilu víme, že $F(\hat{q}_p) \approx \hat{F}_n(\hat{q}_p) \approx p$ (zanedbáváme diskrétnost skoků $\hat{F}_n$ řádu $1/n$). Tedy můžeme psát aproximaci:

$$F(\hat{q}_p) - F(q_p) \approx f(q_p)(\hat{q}_p - q_p). \tag{5}$$

Jelikož $F(q_p) = p$, levá strana rovnice reprezentuje odchylku empirické distribuční funkce od teoretické hodnoty. Je známo, že $\sqrt{n}(\hat{F}_n(x) - F(x))$ konverguje v distribuci k normálnímu rozdělení $N(0, F(x)(1 - F(x)))$ (Donskerova věta). Tedy pro $x = q_p$:

$$\hat{F}_n(q_p) - p \approx -(F(\hat{q}_p) - p) \approx -f(q_p)(\hat{q}_p - q_p). \tag{6}$$

Vyjádříme-li $(\hat{q}_p - q_p)$, dostáváme tzv. Bahadurovu reprezentaci kvantilu (**David2003**):

$$\hat{q}_p - q_p \approx \frac{p - \hat{F}_n(q_p)}{f(q_p)}. \tag{7}$$

Nyní aplikujeme operátor rozptylu na obě strany. Protože $f(q_p)$ je konstanta, dostáváme asymptotický rozptyl (AVar):

$$\mathrm{AVar}(\hat{q}_p) \approx \frac{1}{[f(q_p)]^2}\mathrm{Var}(\hat{F}_n(q_p)). \tag{8}$$

Víme, že $n\hat{F}_n(q_p)$ má binomické rozdělení $Bi(n, p)$, a tedy rozptyl $\hat{F}_n(q_p)$ je:

$$\mathrm{Var}(\hat{F}_n(q_p)) = \frac{p(1-p)}{n}. \tag{9}$$

Dosazením získáme finální vzorec pro asymptotický rozptyl výběrového kvantilu:

$$\mathrm{AVar}(\hat{q}_p) = \frac{p(1-p)}{n[f(q_p)]^2}. \tag{10}$$

Tento výsledek je standardní větou v asymptotické statistice (**VanDerVaart1998**; **Koenker2005**). 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 (10) 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.

3.3 Hodnotící kritéria

Pro srovnání metod sledujeme:

- **Mean Squared Error (MSE):** Celková chyba odhadu kvantilu.
- **Coverage Probability (CP):** Procento případů, kdy sestrojený 95% asymptotický interval spolehlivosti $\hat{q}_p \pm 1.96 \cdot \widehat{SE}$ pokrývá skutečnou hodnotu q_p.

4. Výsledky

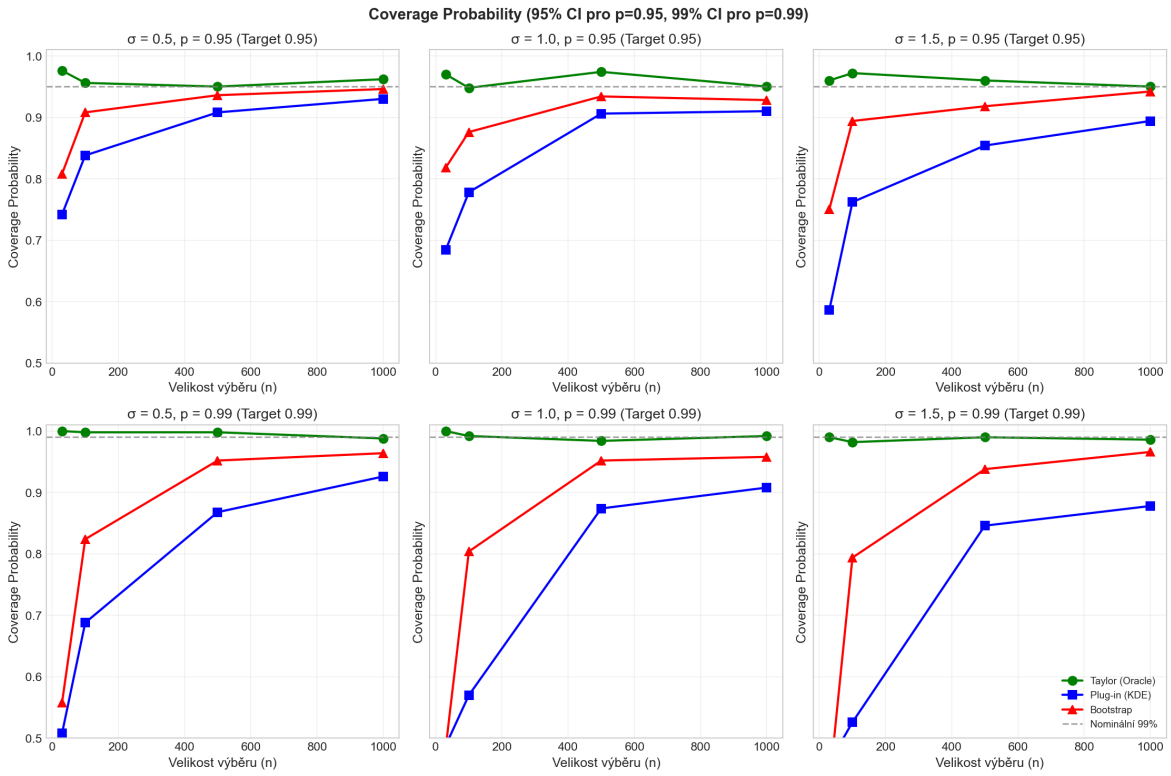
V této části prezentujeme výsledky Monte Carlo simulace. Sledujeme chování odhadů pro dvě hladiny kvantilů: $p = 0.95$ a $p = 0.99$. Abychom zachovali konzistenci mezi odhadovaným kvantilem a intervalem spolehlivosti, konstruujeme pro kvantil $p = 0.95$ interval spolehlivosti o hladině spolehlivosti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



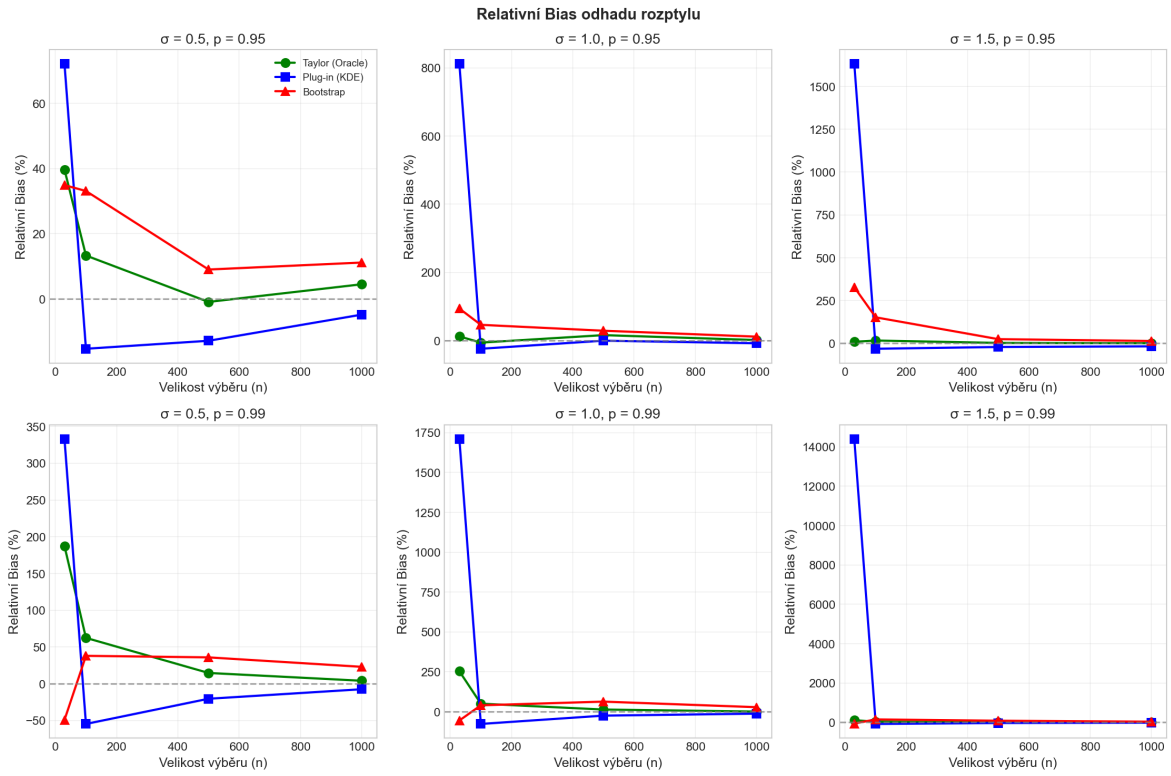
Obrázek 4.1: Coverage Probability. Horní řada: cíl 0.95 (95% CI). Dolní řada: cíl 0.99 (99% CI). Metody: Bootstrap (červená), Plug-in (modrá), Oracle (zelená).

Výsledky ukazují:

1. **Oracle metoda (zelená)** konzistentně dosahuje nominálního pokrytí, což validuje teoretický rámec.
2. **Plug-in metoda (modrá)** selhává pro extrémní kvantil $p = 0.99$ a vysokou asymetrii $\sigma = 1.5$. Coverage zde klesá výrazně pod požadovaných 0.99. Příčinou je podhodnocení rozptylu způsobené jádrovým vyhlazováním.
3. **Bootstrap (červená)** v podmínkách silné asymetrie ($\sigma = 1.5$) překonává Plug-in metodu, ale pro malé rozsahy výběru ($n = 30$) rovněž nedosahuje cílové hladiny 0.99, což je způsobeno malým počtem pozorování v chvostu.

4.2 Systematické vychýlení (Relative Bias)

Obrázek 4.2 ukazuje relativní bias odhadu směrodatné chyby.

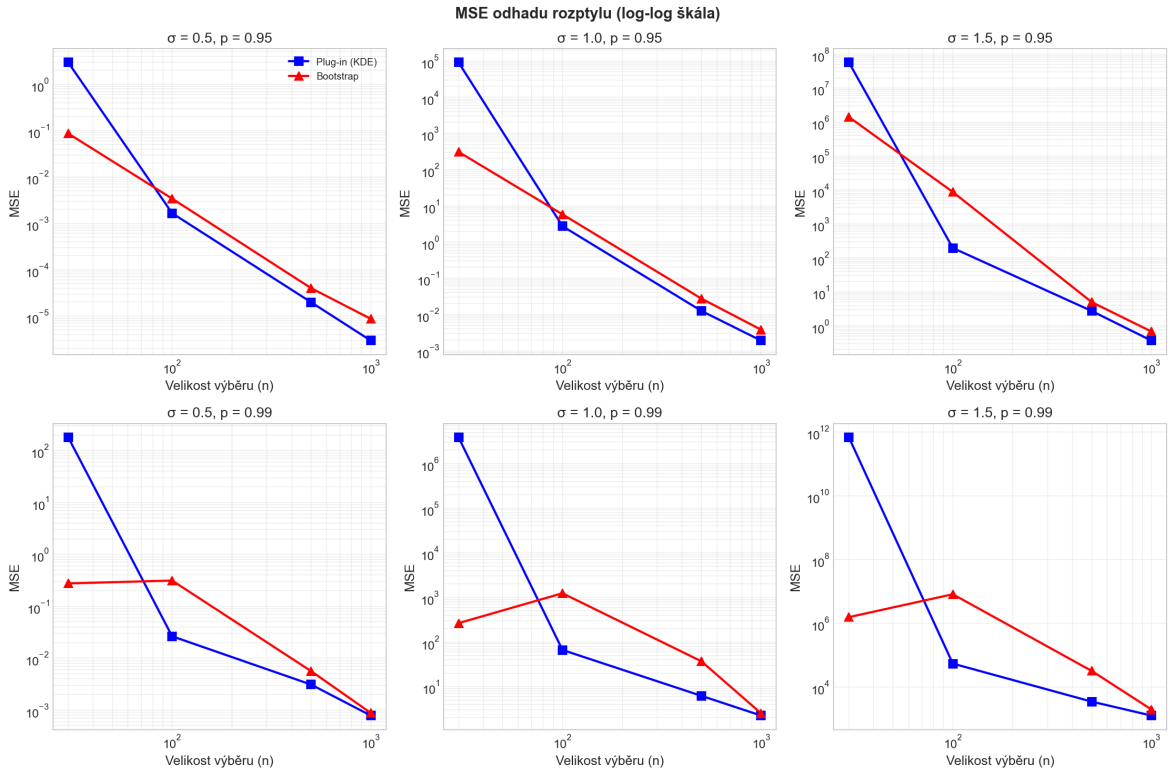


Obrázek 4.2: Relativní vychýlení odhadu směrodatné chyby (SE).

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

Zatímco pro kvantil $p = 0.95$ a 95% interval spolehlivosti metoda funguje uspokojivě, při požadavku na vysokou spolehlivost (99% interval pro kvantil $p = 0.99$) a vysoké asymetrii ($\sigma = 1.5$) dochází k selhání. Odhadnutá hustota $\hat{f}$ je systematicky vychýlená, což vede k podhodnocení směrodatné chyby až o desítky procent, jak ukazuje graf relativního vychýlení. Výsledné intervaly jsou příliš úzké, a proto nepokrývají skutečnou hodnotu s požadovanou pravděpodobností 0.99.

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