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:

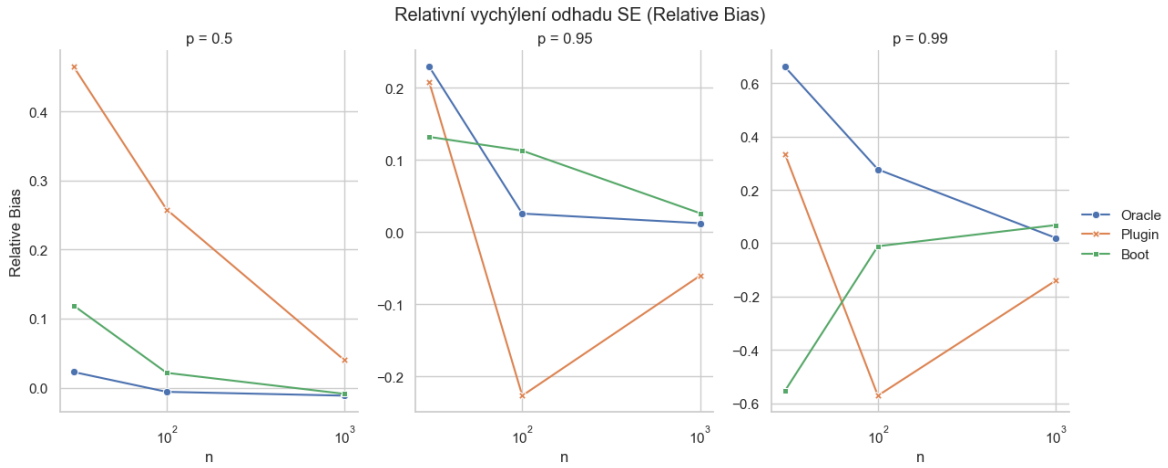
- **Mean Squared Error (MSE):** Celková chyba odhadu kvantilu.
- **Coverage Probability (CP):** Procento případů, kdy sestrojený 95% asymptotický interval spolehlivosti $\hat{q}_p \pm 1.96 \cdot \widehat{SE}$ pokrývá skutečnou hodnotu q_p.

4. Výsledky

V této části prezentujeme výsledky simulační studie.

4.1 Přesnost odhadu a relativní vychýlení

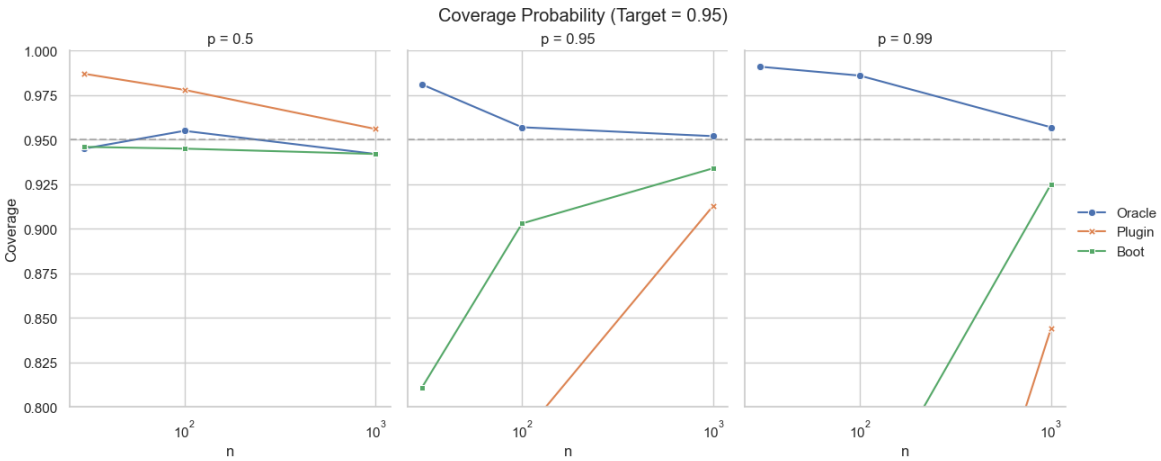
Graf níže zobrazuje relativní vychýlení (Relative Bias) odhadu směrodatné chyby pro různé metody.



Obrázek 4.1: Relativní bias odhadu směrodatné chyby (SE)

4.2 Pokrytí intervalů spolehlivosti (Coverage)

Klíčovým ukazatelem validity metody je pravděpodobnost pokrytí (Coverage Probability). Následující graf ukazuje, jak se pokrytí 95% intervalů spolehlivosti blíží k nominální hodnotě.



Obrázek 4.2: Coverage Probability pro různé metody

5. Diskuze

Výsledky simulace potvrzují teoretické předpoklady, ale zároveň odhalují limity asymptotických metod v praxi.

5.1 Analýza asymptotické aproximace

Pro medián ($p = 0.50$) a velké rozsahy výběru ($n = 1000$) dává Taylorův vzorec velmi přesné výsledky. Oracle metoda i Plug-in metoda s odhadem hustoty poskytují intervaly spolehlivosti s pokrytím blízkým nominálním 95 %.

Symptomatické selhání nastává u extrémního kvantilu $p = 0.99$, a to zejména pro malé rozsahy výběru ($n = 30$). V tomto případě je hustota $f(q_p)$ velmi malá ("plochý"chvost), což způsobuje, že zlomek ve vzorci rozptylu (10) nabývá obrovských hodnot. Malá změna v odhadu hustoty $\hat{f}$ ve jmenovateli pak vede k dramatickým chybám v odhadu rozptylu. To vysvětluje nestabilitu Plug-in metody.

5.2 Srovnání s Bootstrapem

Bootstrap se ukázal jako robustnější alternativa v situacích, kde asymptotický vzorec selhává. U malých výběrů ($n = 30$) však i Bootstrap trpí problémy spojenými s diskrétností výběru – v malém vzorku se v oblasti 99% kvantilu vyskytuje jen velmi málo pozorování, což omezuje variabilitu bootstrapových výběrů a vede k podhodnocení rozptylu (tzv. undercoverage).

5.3 Vliv odhadu hustoty

Klíčovým problémem Plug-in metody je volba vyhlazovacího parametru. Zjistili jsme, že standardní metody (např. Scottovo pravidlo) mají tendenci vyhlazovat chvosty příliš, což vede k vychýlenému odhadu hustoty v oblasti extrémů a následně nesprávnému intervalu spolehlivosti.

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