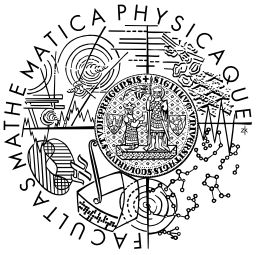


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Iveta Hrušková

Výběrové kvantily

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Arnošt Komárek, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2019

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Ráda bych na tomto místě poděkovala vedoucímu mé bakalářské práce, doc. RNDr. Arnoštovi Komárkovi, Ph.D., za jeho cenné rady, ochotu a čas, který mi věnoval.

Název práce: Výběrové kvantily

Autor: Iveta Hrušková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Jestliže je rozdělení náhodné veličiny neznámé, nejsme schopni určit hodnotu teoretického kvantilu. Jsme-li však v situaci, kdy máme náhodný výběr z onoho rozdělení, můžeme teoretický kvantil odhadovat. Takový odhad pak nazveme výběrovým kvantilem. V této práci se zaměříme na devět často používaných variant výběrového kvantilu a budeme je porovnát podle platnosti vlastností, které má smysl po výběrovém kvantilu požadovat. Pro představu si podobu těchto výběrových kvantilů budeme ilustrovat na jednoduchém příkladu. Na závěr pak ukážeme, že všechny uvedené podoby výběrového kvantilu jsou konzistentními odhady teoretického kvantilu a následně se budeme zabývat konstrukcí intervalu spolehlivosti pro teoretický kvantil.

Klíčová slova: teoretický kvantil, výběrový kvantil, odhadování

Title: Sample Quantiles

Author: Iveta Hrušková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: If the distribution of random variable is uknown, we are not able to figure out the value of theoretical quantile. In case there is a random sample from this distribution, it is possible to estimate the value of theoretical quantile. This estimation is called sample quantile. This work is focused on nine frequently used varieties of sample quantile. They will be compared by means of characteristics that can be examined when speaking about sample quantile. All these varieties will be demonstrated on simple example. Finally, there will be shown that all these versions of sample quantile are consistent estimators of theoretical quantile. The construction of confidence interval for theoretical quantile will be the topic of the final part of the work.

Keywords: theoretical quantile, sample quantile, estimating

Obsah

Úvod

Mějme reálnou náhodnou veličinu. Jestliže známe její rozdělení (distribuční funkci nebo hustotu), můžeme pro $p \in (0,1)$ určit hodnotu teoretického p-kvantilu. Například pro normované normální rozdělení jsou hodnoty teoretických kvantilů tabelovány. Pokud rozdělení náhodné veličiny neznáme, teoretický kvantil nemůžeme určit. Jsme-li však v situaci, kdy máme náhodný výběr, tedy nezávislé, stejně rozdělené náhodné veličiny, můžeme teoretický kvantil odhadovat. Tento odhad pak nazveme výběrovým kvantilem.

V praxi, tedy zejména ve statistických softwarech, se využívá více způsobů pro odhadování teoretického kvantilu. V této práci se zaměříme celkem na devět často používaných variant výběrového kvantilu, které teoretický kvantil odhadují. Těmito výběrovými kvantily se zabývali autoři Hyndman a Fan (2009) a typicky se jedná o lineární kombinaci dvou po sobě jdoucích pořádkových statistik.

Článkem autorů Hyndman a Fan (2009) se inspirovaly i statistické softwary, jako např. R (R Core Team (2016)) nebo SAS (SAS Institute, Cary NC). Zatímco v programu R (verze 3.6.0) můžeme využít všech devět podob výběrového kvantilu, v programu SAS (verze 9.3) je jich naimplementováno pouze pět.

V první kapitole se se všemi výběrovými kvantily seznámíme a uvedeme vlastnosti, které po výběrovém kvantilu požadujeme. Pokusíme se podrobněji vysvětlit, proč je vhodné, aby výběrový kvantil tyto vlastnosti splňoval. Ve druhé kapitole pak varianty výběrového kvantilu podle těchto vlastností porovnáme. Na závěr, ve třetí kapitole, se pokusíme odvodit některé statistické vlastnosti výběrových kvantilů a rozšířit tak poznatky uvedené v práci Omelka (2019).

Pro práci s výběrovými kvantily si připomeneme některé základní definice a vlastnosti.

Definice 1 (teoretický kvantil)**.** *Nechť $p \in (0,1)$ a F_X je distribuční funkce. Pak* kvantilovou funkci *rozdělení F_X definujeme jako*

$$F_X^{-1}(p) := \inf\{x \in \boldsymbol{R} : F_X(x) \geq p\}$$

a p-kvantilem *tohoto rozdělení rozumíme číslo $u_X(p) = F_X^{-1}(p)$.*

Z definice teoretického kvantilu je zřejmé, že platí

$$F_X(u_X(p)) \geq p \ , \ F_X(u_X(p) - h) < p \ \ \forall h > 0.$$

Definice 2 (uspořádaný náhodný výběr)**.** *Mějme* náhodný výběr $X_1, \ldots, X_n$*, tedy náhodné veličiny $X_1, \ldots, X_n$ jsou nezávislé a stejně rozdělené. Konstantu n nazýváme rozsah náhodného výběru. Pro $k \in \{1, \ldots, n\}$ označme symbolem $X_{(k)}$ k-tou nejmenší hodnotu náhodného výběru, nazýváme ji* k-tá pořádková statistika*. Pak $X_{(1)}, \ldots, X_{(n)}$ značí* uspořádaný náhodný výběr*, tedy platí $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$.*

Definice 3 (empirická distribuční funkce)**.** *Nechť $X_1, \ldots, X_n$ je náhodný výběr. Pak definujeme* empirickou distribuční funkci *předpisem*

$$\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{[X_k \leq x]}, \ x \in \boldsymbol{R}.$$

Věta 1. *Mějme empirickou distribuční funkci $\hat{F}_n(x)$ náhodného výběru s distribuční funkcí F_X. Pak $\forall x \in \mathbf{R}$ platí:*

$$
\begin{array}{ll}
(i) & \boldsymbol{E}\hat{F}_n(x) = F_X(x),\ \mathrm{var}(\hat{F}_n(x)) = \frac{F_X(x)[1-F_X(x)]}{n} \\
(ii) & \hat{F}_n(x) \xrightarrow[n\to\infty]{P} F_X(x) \\
(iii) & \sqrt{n}[\hat{F}_n(x) - F_X(x)] \xrightarrow[n\to\infty]{d} N\left(0,F_X(x)[1-F_X(x)]\right) \\
(iv) & n\hat{F}_n(x) \sim Bi\left(n,F_X(x)\right)
\end{array}
$$

Poznámka. Pro práci s výběrovými kvantily bude klíčová vlastnost (iv).

Důkaz. (i) První bod je vidět z toho, že můžeme psát

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \le x]} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

kde Y_i jsou nezávislé stejně rozdělené náhodné veličiny, $Y_i \sim Alt(F(x))$, tedy $\mathbf{E}Y_i = F_X(x), \mathrm{var}(Y_i) = F_X(x)\left[1 - F_X(x)\right]$.
(ii) Tato vlastnost plyne z Čebyševovy nerovnosti, kterou lze nalézt např. v textu Dupač a Hušková (2001, str. 26). Tedy $\forall \epsilon > 0$ platí

$$\mathbf{P}\left(|\hat{F}_n(x) - F_X(x)| > \epsilon\right) \le \frac{F_X(x)\left(1 - F_X(x)\right)}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0.$$

(iii) Tento bod je za použití centrální limitní věty, jejíž podobu můžeme nalézt také v textu Dupač a Hušková (2001, str. 80), zřejmý.
(iv) Stejně jako v části (i) využijeme přepisu empirické distribuční funkce pomocí součtu veličin s alternativním rozdělením a dále využijeme

$$Y_i \sim Alt(p),\ i \in \{1,\dots,n\} \Rightarrow \sum_{i=1}^{n} Y_i \sim Bi(n,p).$$

□

1. Varianty výběrového kvantilu

V celé práci budeme předpokládat, že pracujeme s reálným náhodným výběrem $X_1, \ldots, X_n$, který se řídí rozdělením s neznámou distribuční funkcí F_X. Dále budeme předpokládat, že $p \in (0,1)$.

V této kapitole uvedeme devět možných podob výběrového kvantilu, přičemž výběrový kvantil podle i-té definice budeme značit $\hat{Q}_i(p)$.

Výběrové kvantily, které si zde rozebereme, jsou všechny založeny na lineární kombinaci dvou po sobě jdoucích pořádkových statistikách (v některých případech se jedná pouze o jednu pořádkovou statistiku, neboť koeficient u zbylé pořádkové statistiky je nulový).

Definice 4 (výběrový kvantil)**.** *Hodnotou* výběrového kvantilu *v bodě p rozumíme statistiku*

$$\hat{Q}_i(p) = (1-\gamma)X_{(j)} + \gamma X_{(j+1)}, \tag{1.1}$$

kde $\frac{j-m}{n} \leq p < \frac{j-m+1}{n}$, $m \in \boldsymbol{R}$ *a* $\gamma \in [0,1]$.
Hodnota γ *pak závisí na* $j = \lfloor pn + m \rfloor$ *a na* $g = pn + m - j$.

Výběrové kvantily diskutované v této práci můžeme rozdělit do dvou skupin. První skupina se skládá z předpisů výběrového kvantilu, které nejsou spojité v p. Druhá skupina pak obsahuje výběrové kvantily založené na lineární interpolaci.

1.1 Žádoucí vlastnosti výběrového kvantilu

V této sekci si uvedeme šest vlastností, které autoři Hyndman a Fan (2009) po výběrovém kvantilu požadují a pokusíme se vysvětlit, proč je vhodné, aby výběrový kvantil tyto vlastnosti měl.

Výčet vlastností:

- $(V1)$ $\hat{Q}_i(p)$ je spojitý
- $(V2)$ $\sum_{k=1}^{n} \mathbb{1}_{[X_k \leq \hat{Q}_i(p)]} \geq \lfloor pn \rfloor$
- $(V3)$ $\sum_{k=1}^{n} \mathbb{1}_{[X_k \leq \hat{Q}_i(p)]} = \sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_i(1-p)]}$
- $(V4)$ Jestliže $\hat{Q}_i^{-1}$ je jednoznačně definována, pak pro $k \in \{1, \ldots, n\}$ platí $\hat{Q}_i^{-1}(X_{(k)}) + \hat{Q}_i^{-1}(X_{(n-k+1)}) = 1$
- $(V5)$ Jestliže $\hat{Q}_i^{-1}$ je jednoznačně definována, pak platí $\hat{Q}_i^{-1}(X_{(1)}) > 0$ a $\hat{Q}_i^{-1}(X_{(n)}) < 1$
- $(V6)$ $\hat{Q}_i(0{,}5)$ je rovno výběrovému mediánu definovanému předpisem $\frac{X_{(l)} + X_{(l+1)}}{2}$, jestliže n je sudé, tedy $n = 2l$, $X_{(l+1)}$, jestliže n je liché, tedy $n = 2l + 1$

Proč tedy po výběrovém kvantilu požadujeme tyto vlastnosti?

Vlastnost $(V1)$ vychází z toho, že chceme, aby kvantilová funkce $F_X^{-1}(p)$ byla spojitá. Tuto vlastnost má ale smysl požadovat pouze v případě, že distribuční funkce F_X náhodného výběru, se kterým pracujeme, je spojitá.

$(V2)$ je pak důsledek toho, že pro teoretický kvantil platí $F_X(u_X(p)) \geq p$.

Tato vlastnost má smysl i bez předpokladu spojitosti F_X. K jejímu ověření budeme využívat následující rozpis (převážně pro výběrové kvantily založené na lineární interpolaci).

$$\sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_i(p)]} = \sum_{k=1}^{n} \mathbb{1}_{[X_k \le (1-\gamma)X_{(j)} + \gamma X_{(j+1)}]} \ge \sum_{k=1}^{n} \mathbb{1}_{[X_k \le X_{(j)}]} \ge j = \lfloor pn + m \rfloor \quad (1.2)$$

Vlastnost $(V3)$ je rozumné požadovat pouze, je-li distribuční funkce F_X spojitá. Vychází z definice výběrového kvantilu, neboť za spojitosti F_X platí

$$\sum_{k=1}^{n} \mathbb{1}_{[X_k \le u_X(p)]} = \sum_{k=1}^{n} \mathbb{1}_{[X_k \ge u_X(1-p)]}.$$

Tuto vlastnost budeme ověřovat pomocí následujícího rozpisu (opět budeme využívat převážně pro výběrové kvantily založené na lineární interpolaci).

$$\sum_{k=1}^{n} \mathbb{1}_{[X_k \ge \hat{Q}_i(1-p)]} = n - \sum_{k=1}^{n} \mathbb{1}_{[X_k < \hat{Q}_i(1-p)]} = n - \sum_{k=1}^{n} \mathbb{1}_{[X_k < (1-\gamma')X_{(j')} + \gamma' X_{(j'+1)}]} =$$

$$= n - \sum_{k=1}^{n} \mathbb{1}_{[X_k < X_{(j')}]} = n - (j' - 1) = n - \lfloor (1-p)n + m - 1 \rfloor = \lfloor pn - m + 1 \rfloor \quad (1.3)$$

Zde $j' = \lfloor (1-p)n + m \rfloor$ a $\gamma' = (1-p)n + m - j'$.

Vlastnost $(V4)$ je vlastností symetrie. K této vlastnosti potřebujeme, aby výběrový kvantil byl spojitý a rostoucí (neboť je to jediná možnost, aby byl výběrový kvantil prostý), protože jinak inverzní funkce není jednoznačně definována.

Stejně jako v předchozím případě nemá smysl vlastnost $(V5)$ uvažovat pro výběrový kvantil, který není spojitý a rostoucí. Tato vlastnost je důsledkem toho, že pro spojitou distribuční funkci očekáváme kladnou pravděpodobnost napozorování hodnot mimo rozsah našich dat, neboli důsledkem toho, že platí $\mathbf{P}(X < X_{(1)}) > 0$ a $\mathbf{P}(X > X_{(n)}) > 0$.

1.2 Nespojité varianty výběrového kvantilu

Varianta 1. První varianta výběrového kvantilu předpokládá

$$m = 0, \quad j = \lfloor pn \rfloor, \quad g = pn - \lfloor pn \rfloor, \quad \gamma = \begin{cases} 0, & g = 0 \\ 1, & g > 0. \end{cases}$$

Všimněme si, že $g = 0$ právě tehdy, když $pn = \lfloor pn \rfloor$, neboli když $pn \in \mathbf{Z}$. Tedy můžeme psát

$$\hat{Q}_1(p) = \begin{cases} X_{(pn)}, & pn \in \mathbf{Z} \\ X_{(\lfloor pn \rfloor + 1)}, & pn \notin \mathbf{Z}, \end{cases} \quad (1.4)$$

kde $\dfrac{j}{n} \le p < \dfrac{j+1}{n}$.

Tato podoba výběrového kvantilu je vlastně zobecněnou inverzí empirické distribuční funkce. Zde je třeba chápat inverzi ve stejném smyslu, jako když jsme v úvodu definovali teoretický kvantil, tedy

$$\hat{Q}_1(p) = \hat{F}_n^{-1}(p) := \inf\{x \in \mathbf{R} : \hat{F}_n(x) \geq p\} \ .$$

Varianta 2. Druhá varianta výběrového kvantilu je podobná té první, vychází z

$$m = 0, \quad j = \lfloor pn \rfloor, \quad g = pn - \lfloor pn \rfloor, \quad \gamma = \begin{cases} \frac{1}{2}, & g = 0 \\ 1, & g > 0. \end{cases}$$

V tomto případě tedy nabývá výběrový kvantil tvaru

$$\hat{Q}_2(p) = \begin{cases} \dfrac{X_{(pn)} + X_{(pn+1)}}{2}, & pn \in \mathbf{Z} \\ \\ X_{(\lfloor pn \rfloor + 1)}, & pn \notin \mathbf{Z}, \end{cases} \tag{1.5}$$

kde $\dfrac{j}{n} \leq p < \dfrac{j+1}{n}$.

Varianta 3. Poslední z nespojitých podob výběrového kvantilu je o něco složitější. Definuje se jako k-tá pořádková statistika $X_{(k)}$, kde k je nejbližší celé číslo k np. Taková volba ale v některých případech není jednoznačná. Uvedeme si jednu z možností, která se v praxi často používá.

$$m = -\tfrac{1}{2}, \quad j = \left\lfloor pn - \tfrac{1}{2} \right\rfloor, \quad g = pn - \tfrac{1}{2} - \left\lfloor pn - \tfrac{1}{2} \right\rfloor, \quad \gamma = \begin{cases} 1, & g > 0 \\ 0, & g = 0,\ j \text{ sudé} \\ 1, & g = 0,\ j \text{ liché}. \end{cases}$$

Pak můžeme psát výběrový kvantil předpisem

$$\hat{Q}_3(p) = \begin{cases} X_{\left(pn - \frac{1}{2}\right)}, & pn - \frac{1}{2} \in \mathbf{Z},\ pn - \frac{1}{2} \text{ sudé} \\ X_{\left(\lfloor pn + \frac{1}{2} \rfloor\right)}, & pn - \frac{1}{2} \in \mathbf{Z},\ pn - \frac{1}{2} \text{ liché} \\ & \text{nebo } pn - \frac{1}{2} \notin \mathbf{Z}, \end{cases} \tag{1.6}$$

kde $\dfrac{j + \frac{1}{2}}{n} \leq p < \dfrac{j + \frac{3}{2}}{n}$.

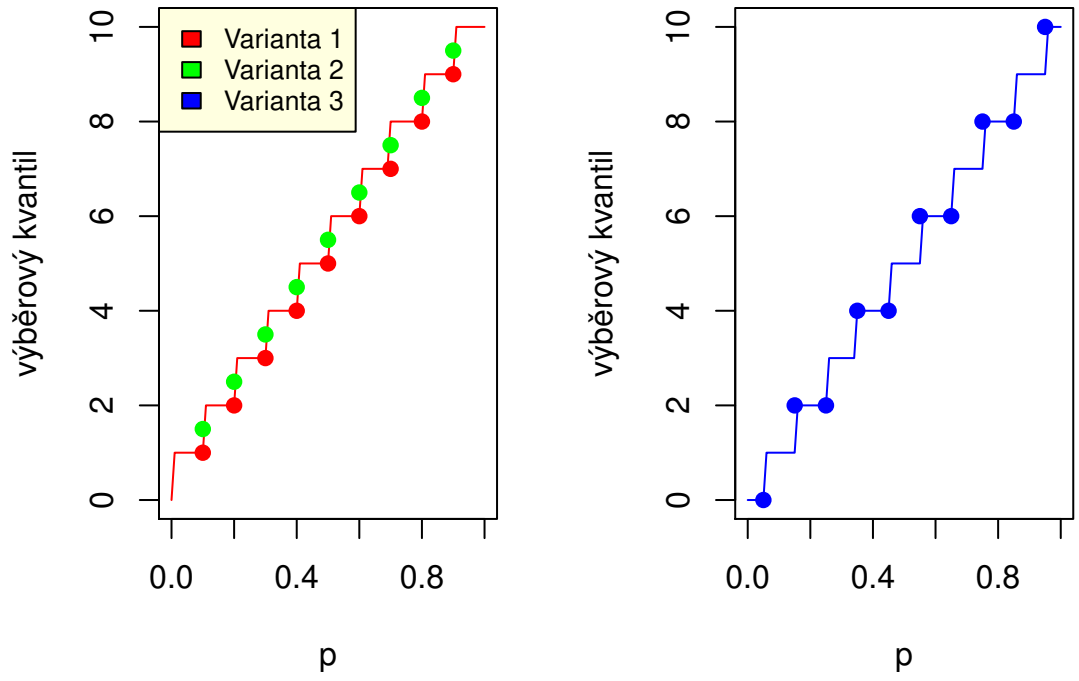
Příklad. Mějme uspořádaný náhodný výběr $X_{(1)}, X_{(2)}, \ldots, X_{(10)} = 1, 2, \ldots, 10$. Pro tento případ je na obrázku 1.1 ilustrována podoba výše uvedených výběrových kvantilů. Definice výběrových kvantilů $\hat{Q}_1$ a $\hat{Q}_2$ jsou velice podobné, a proto jsou společně vyobrazeny v levém grafu na tomto obrázku. Pro variantu 2 platí, že graf splývá s variantou 1 a liší se pouze v bodech, které jsou na obrázku zvýrazněny. V pravém grafu pak můžeme nahlédnout podobu výběrového kvantilu $\hat{Q}_3$.

1.3 Varianty založené na lineární interpolaci

Výběrové kvantily založené na lineární interpolaci jsou po částech lineární a tudíž i spojité. V tomto případě budeme při interpolaci pracovat s body $(p_k, X_{(k)})$, kde zobrazovací pozice p_k je definována jako

$$p_k = \frac{k - \alpha}{n - \alpha - \beta + 1},$$

Obrázek 1.1: Nespojité varianty výběrového kvantilu



kde α a β jsou reálné konstanty. Potom interpolací mezi body $(p_k,X_{(k)})$ dostameme výběrový kvantil ve tvaru (1.1) pro který platí

$$m = \alpha + p(1 - \alpha - \beta), \quad j = \lfloor pn + m \rfloor, \quad \gamma = g = pn + m - j.$$

Varianta 4. Pro tuto variantu výběrového kvantilu volíme

$$p_k = \frac{k}{n}, \quad \alpha = 0, \quad \beta = 1, \quad m = 0, \quad j = \lfloor pn \rfloor, \quad \gamma = pn - \lfloor pn \rfloor$$

a dostáváme předpis

$$\hat{Q}_4(p) = (1 - pn + \lfloor pn \rfloor) X_{(\lfloor pn \rfloor)} + (pn - \lfloor pn \rfloor) X_{(\lfloor pn \rfloor + 1)}, \tag{1.7}$$

kde $\frac{j}{n} \leq p < \frac{j+1}{n}$.

Varianta 5. V tomto případě budeme vycházet z

$$p_k = \frac{k - \frac{1}{2}}{n}, \quad \alpha = \tfrac{1}{2}, \quad \beta = \tfrac{1}{2}, \quad m = \tfrac{1}{2}, \quad j = \left\lfloor pn + \tfrac{1}{2} \right\rfloor, \quad \gamma = pn + \tfrac{1}{2} - \left\lfloor pn + \tfrac{1}{2} \right\rfloor.$$

Výběrový kvantil tedy nabývá tvaru

$$\hat{Q}_5(p) = \left(\frac{1}{2} - pn + \left\lfloor pn + \frac{1}{2} \right\rfloor \right) X_{(\lfloor pn + \frac{1}{2} \rfloor)} + \left(pn + \frac{1}{2} - \left\lfloor pn + \frac{1}{2} \right\rfloor \right) X_{(\lfloor pn + \frac{3}{2} \rfloor)}, \tag{1.8}$$

kde $\frac{j-\frac{1}{2}}{n} \leq p < \frac{j+\frac{1}{2}}{n}$.

Varianta 6. Šestá varianta výběrového kvantilu předpokládá

$$p_k = \frac{k}{n+1}, \quad \alpha = 0, \quad \beta = 0, \quad m = p, \quad j = \lfloor p(n+1) \rfloor,$$

$$\gamma = p(n+1) - \lfloor p(n+1) \rfloor.$$

Výběrový kvantil můžeme v tomto případě psát jako

$$\begin{aligned} \hat{Q}_6(p) = \quad & (1 - p(n+1) + \lfloor p(n+1) \rfloor) X_{(\lfloor p(n+1) \rfloor)} + \\ & + (p(n+1) - \lfloor p(n+1) \rfloor) X_{(\lfloor p(n+1) \rfloor + 1)}, \end{aligned} \tag{1.9}$$

kde $\frac{j-p}{n} \leq p < \frac{j-p+1}{n}$.

Varianta 7. Tato podoba výběrového kvantilu požaduje

$$p_k = \frac{k-1}{n-1}, \quad \alpha = 1, \quad \beta = 1, \quad m = 1-p, \quad j = \lfloor p(n-1) \rfloor + 1,$$

$$\gamma = p(n-1) + 1 - \lfloor p(n-1) + 1 \rfloor.$$

A můžeme psát

$$\begin{aligned} \hat{Q}_7(p) = \quad & (-p(n-1) + \lfloor p(n-1) + 1 \rfloor) X_{(\lfloor p(n-1) \rfloor + 1)} + \\ & + (p(n-1) + 1 - \lfloor p(n-1) + 1 \rfloor) X_{(\lfloor p(n-1) \rfloor + 2)}, \end{aligned} \tag{1.10}$$

kde $\frac{j-1+p}{n} \leq p < \frac{j+p}{n}$.

Varianta 8. Pro tuto variantu výběrového kvantilu se volí

$$p_k = \frac{k-\frac{1}{3}}{n+\frac{1}{3}}, \quad \alpha = \tfrac{1}{3}, \quad \beta = \tfrac{1}{3}, \quad m = \tfrac{1+p}{3}, \quad j = \left\lfloor p\left(n+\tfrac{1}{3}\right) + \tfrac{1}{3} \right\rfloor,$$

$$\gamma = p\left(n+\frac{1}{3}\right) + \frac{1}{3} - \left\lfloor p\left(n+\frac{1}{3}\right) + \frac{1}{3} \right\rfloor.$$

Výběrový kvantil má tedy tvar

$$\begin{aligned} \hat{Q}_8(p) = \quad & \left(\tfrac{2}{3} - p\left(n+\tfrac{1}{3}\right) + \left\lfloor p\left(n+\tfrac{1}{3}\right) + \tfrac{1}{3} \right\rfloor\right) X_{\left(\left\lfloor p\left(n+\frac{1}{3}\right) + \frac{1}{3} \right\rfloor\right)} + \\ & + \left(p\left(n+\tfrac{1}{3}\right) + \tfrac{1}{3} - \left\lfloor p\left(n+\tfrac{1}{3}\right) + \tfrac{1}{3} \right\rfloor\right) X_{\left(\left\lfloor p\left(n+\frac{1}{3}\right) + \frac{4}{3} \right\rfloor\right)}, \end{aligned} \tag{1.11}$$

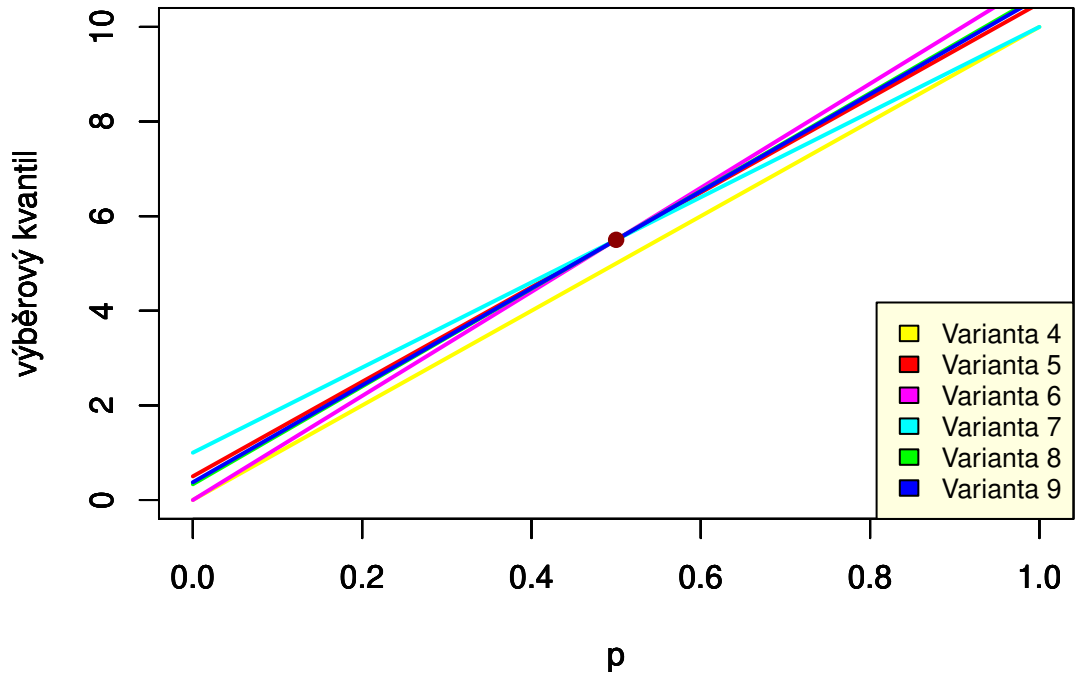
kde $\frac{j-\frac{p}{3}-\frac{1}{3}}{n} \leq p < \frac{j-\frac{p}{3}+\frac{2}{3}}{n}$.

Varianta 9. Poslední podoba výběrového kvantilu vychází z

$$p_k = \frac{k-\frac{3}{8}}{n+\frac{1}{4}}, \quad \alpha = \tfrac{3}{8}, \quad \beta = \tfrac{3}{8}, \quad m = \tfrac{3}{8} + \tfrac{p}{4}, \quad j = \left\lfloor p\left(n+\tfrac{1}{4}\right)\tfrac{3}{8} \right\rfloor,$$

$$\gamma = p\left(n+\frac{1}{4}\right) + \frac{3}{8} - \left\lfloor p\left(n+\frac{1}{4}\right) + \frac{3}{8} \right\rfloor.$$

Obrázek 1.2: Spojité varianty výběrového kvantilu



Pak dostáváme výběrový kvantil ve tvaru

$$\begin{aligned}\hat{Q}_9(p) = \quad & \left(\tfrac{5}{8} - p\left(n+\tfrac{1}{4}\right) + \left\lfloor p\left(n+\tfrac{1}{4}\right)+\tfrac{3}{8}\right\rfloor\right) X_{(\lfloor p(n+\frac{1}{4})+\frac{3}{8}\rfloor)} + \\ & + \left(p\left(n+\tfrac{1}{4}\right) + \tfrac{3}{8} - \left\lfloor p\left(n+\tfrac{1}{4}\right)+\tfrac{3}{8}\right\rfloor\right) X_{(\lfloor p(n+\frac{1}{4})+\frac{11}{8}\rfloor)},\end{aligned} \tag{1.12}$$

kde $\dfrac{j-\frac{p}{4}+\frac{3}{8}}{n} \le p < \dfrac{j-\frac{p}{4}+\frac{11}{8}}{n}$.

Příklad. Nyní se vrátíme k příkladu, který je uveden na konci předchozí sekce. Opět se budeme zabývat tvarem výše uvedených výběrových kvantilů, máme-li uspořádaný náhodný výběr $X_{(1)}, X_{(2)}, \ldots, X_{(10)} = 1, 2, \ldots, 10$. Podoba těchto výběrových kvantilů je znázorněna na obrázku 1.2.
Z obrázku je patrné, že některé varianty výběrového kvantilu jsou si velice podobné a jejich grafy téměř splývají. Například varianta 8 není na obrázku skoro vůbec vidět, protože ji překrývá varianta 9.
V obrázku je červeným bodem zakreslena hodnota výběrového mediánu pro tento případ a je vidět, že všechny výběrové kvantily až na $\hat{Q}_4$ tímto bodem prochází, tedy je pro ně v tomto případě splněna vlastnost $(V6)$. Platnost této vlastnosti pro libovolný náhodný výběr pak početně ověříme ve Větě 3 (sekce 2.2).

2. Porovnání výběrových kvantilů

V této kapitole o každé podobě výběrového kvantilu uvedené v předchozí kapitole rozhodneme, zda pro ni platí vlastnosti $(V1)-(V6)$ ze sekce 1.1. Speciálně pro ověřování vlastnosti $(V3)$ budeme předpokládat spojitou distribuční funkci F_X náhodného výběru, tedy že platí skoro jistě $X_{(1)} < \cdots < X_{(n)}$. Jinak by tato vlastnost neměla smysl. Stejně tak budeme spojitost předpokládat i při ověřování $(V4)$ a $(V5)$ u výběrových kvantilů založených na lineární interpolaci. Při ověřování platnosti ostatních vlastností si vystačíme s vlastnostmi konkrétní podoby výběrového kvantilu.

2.1 Nespojité varianty

Nejprve se podíváme na výběrové kvantily z oddílu 1.2, tedy na výběrové kvantily, které nejsou spojité v p, a tedy pro ně vlastnost $(V1)$ nemůže platit. Jelikož pro ně neplatí spojitost, nemají jednoznačně definovanou inverzní funkci, a tedy pro ně nemůžeme uvažovat platnost vlastností $(V4)$ a $(V5)$. Platnost zbylých vlastností shrneme v následující Větě, kterou následně dokážeme.

Věta 2. *Uvažujme nespojité varianty výběrového kvantilu $\hat{Q}_i, i \in \{1,2,3\}$ ze sekce 1.2. Pak platí*

- (i) *vlastnost $(V2)$ platí pro všechny tyto výběrové kvantily*
- (ii) *vlastnost $(V3)$ neplatí pro žádný z těchto výběrových kvantilů*
- (iii) *vlastnost $(V6)$ platí pouze pro $\hat{Q}_2$ a pro $\hat{Q}_1$ v případě, že n je liché*

Poznámka. Autoři Hyndman a Fan (2009) sice tvrdí, že pro druhou variantu výběrového kvantilu vlastnost $(V3)$ platí, nicméně v důkazu věty ukážeme, že tomu tak není.

Důkaz. (i) Je-li výběrový kvantil definován jako j-tá pořádková statistika, můžeme psát

$$\sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_i(p)]} = \sum_{k=1}^{n} \mathbb{1}_{[X_k \le X_{(j)}]} \ge j.$$

Tedy stačí pro jednotlivé definice dosadit:

$$\begin{array}{ll} \sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_1(p)]} & \ge pn, \qquad pn \in \mathbf{Z} \\ & \ge \lfloor pn+1 \rfloor, \quad pn \notin \mathbf{Z} \end{array}$$

$$\begin{array}{ll} \sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_3(p)]} & \ge \lfloor pn+\frac{1}{2} \rfloor, \qquad \gamma = 1 \\ & \ge pn-\frac{1}{2} = \lfloor pn \rfloor, \quad \gamma = 0 \end{array}$$

Poslední rovnost $pn - \frac{1}{2} = \lfloor pn \rfloor$ platí, neboť je-li $\gamma = 0$, pak $pn - \frac{1}{2} \in \mathbf{Z}$. Pro oba případy tedy tvrzení platí. Druhá varianta výběrového kvantilu $\hat{Q}_2(p)$ není vždy definována jako jedna pořádková statistika, ale důkaz je pro ni obdobný.

$$\begin{array}{ll} \sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_2(p)]} & \ge \lfloor pn+1 \rfloor, \qquad pn \notin \mathbf{Z} \\ & = \sum_{k=1}^{n} \mathbb{1}_{\left[X_k \le \frac{X_{(pn)}+X_{(pn+1)}}{2}\right]} = \sum_{k=1}^{n} \mathbb{1}_{[X_k \le X_{(pn)}]} \ge pn, \quad pn \in \mathbf{Z} \end{array}$$

Je-li distribuční funkce náhodného výběru spojitá, dostaneme všude místo nerovnosí rovnosti. To se nám bude hodit v následující části důkazu.
(*ii*) Podobně jako v předchozím rozepíšeme nejprve pro první variantu výběrového kvantilu

$$\begin{aligned}
\textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_1(1-p)]} &= n - \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k < X_{((1-p)n)}]}, && pn \in \mathbf{Z} \\
&= n - \lceil (1-p)n - 1 \rceil = pn + 1, && pn \in \mathbf{Z} \\
&= n - \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k < X_{(\lfloor (1-p)n+1 \rfloor)}]}, && pn \notin \mathbf{Z} \\
&= n - \lfloor (1-p)n \rfloor = \lfloor pn \rfloor, && pn \notin \mathbf{Z}
\end{aligned}$$

Podíváme-li se zpět do předchozího kroku tohoto důkazu, vidíme, že neplatí $\sum_{k=1}^{n} \mathbb{1}_{[X_k \leq \hat{Q}_1(p)]} = \sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_1(1-p)]}$.

Dále rozepíšeme stejnou vlastnost pro druhou definici výběrového kvantilu

$$\begin{aligned}
\textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_2(1-p)]} &= n - \textstyle\sum_{k=1}^{n} \mathbb{1}_{\left[X_k < \frac{X_{((1-p)n)} + X_{((1-p)n+1)}}{2}\right]}, && pn \in \mathbf{Z} \\
&= n - \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \leq X_{((1-p)n)}]}, && pn \in \mathbf{Z} \\
&= n - ((1-p)n) = pn, && pn \in \mathbf{Z} \\
&= n - \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k < X_{(\lfloor (1-p)n \rfloor + 1)}]}, && pn \notin \mathbf{Z} \\
&= n - \lfloor (1-p)n \rfloor = \lfloor pn \rfloor, && pn \notin \mathbf{Z}
\end{aligned}$$

Tedy pro druhou variantu výběrového kvantilu $(V3)$ také neplatí. Ještě ukážeme, že tato vlastnost neplatí ani pro třetí podobu výběrového kvantilu.

$$\begin{aligned}
\textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_3(1-p)]} &= n - \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k < X_{((1-p)n - \frac{1}{2})}]}, && \gamma' = 0 \\
&= n - \left((1-p)n - \tfrac{3}{2}\right) = pn + \tfrac{3}{2}, && \gamma' = 0 \\
&= n - \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k < X_{(\lfloor (1-p)n + \frac{1}{2} \rfloor)}]}, && \gamma' = 1 \\
&= n - \left(\left\lfloor (1-p)n + \tfrac{1}{2} \right\rfloor\right) = \left\lfloor pn + \tfrac{1}{2} \right\rfloor, && \gamma' = 1,
\end{aligned}$$

kde

$$\gamma' = \begin{cases} 0, & pn - \frac{1}{2} \in \mathbf{Z},\ (1-p)n - \frac{1}{2} \text{ sudé} \\ 1, & pn - \frac{1}{2} \in \mathbf{Z},\ (1-p)n - \frac{1}{2} \text{ liché} \\ & \text{nebo } pn - \frac{1}{2} \notin \mathbf{Z}. \end{cases}$$

(*iii*) Nakonec zbývá ověřit poslední vlastnost $(V6)$, tedy zjistit, zda $\hat{Q}_i(0{,}5)$ je rovno výběrovému mediánu. Opět budeme postupovat dle pořadí výběrových kvantilů.
První varianta výběrového kvantilu je vždy jedna pořádková statistika, tedy obecně vlastnost $V6$ nemůže platit. Ukážeme, že tato vlastnost platí alespoň pro n liché. Jelikož $n = 2l + 1$ je liché a $p = \frac{1}{2}$, musí být $pn \notin \mathbf{Z}$. Pak můžeme psát

$$\hat{Q}_1\left(\frac{1}{2}\right) = X_{(\lfloor pn \rfloor + 1)} = X_{(\lfloor l + \frac{1}{2} \rfloor + 1)} = X_{(l+1)}.$$

Nyní stejnou vlastnost rozepíšeme pro druhou variantu výběrového kvantilu a tím ukážeme, že je splněna.

$$\begin{aligned}
\hat{Q}_2\left(\tfrac{1}{2}\right) &= \tfrac{X_{(l)} + X_{(l+1)}}{2}, && pn \in \mathbf{Z} \Rightarrow n = 2l \\
&= X_{(\lfloor l + \frac{1}{2} \rfloor + 1)} = X_{(l+1)}, && pn \notin \mathbf{Z} \Rightarrow n = 2l + 1
\end{aligned}$$

Třetí varianta výběrového kvantilu je stejně tak jako první definice vždy jedna pořádková statistika, tedy $(V6)$ obecně nemůže platit. Podíváme se, zda platí pro n liché. Je-li n liché, můžeme psát $n = 2l + 1$. Pak $pn - \frac{1}{2} = l$, l však může být jak sudé, tak liché. Rozebereme tedy obě možnosti.

$$\begin{aligned} \hat{Q}_3\left(\tfrac{1}{2}\right) &= X_{(\lfloor \frac{2l+1}{2} + \frac{1}{2} \rfloor)} = X_{(l+1)}, \quad l \text{ liché} \\ &= X_{(\frac{2l+1}{2} - \frac{1}{2})} = X_{(l)}, \qquad l \text{ sudé} \end{aligned}$$

Tedy $(V6)$ neplatí ani pokud n je liché.

□

2.2 Varianty založené na lineární interpolaci

Nyní se budeme zabývat variantami výběrového kvantilu z oddílu 1.2, tedy výběrovými kvantily, které jsou založeny na lineární interpolaci, a tedy spojité. Vlastnost $(V1)$ je pro ně tudíž splněna. Tyto definice musí být za předpokladu spojité distribuční funkce F_X ryze monotónní, takže při ověřování vlastností $(V4)$ a $(V5)$ můžeme počítat s inverzní funkcí $\hat{Q}_i^{-1}(p)$. Kdyby F_X nebyla spojitá, nemohli bychom posoudit platnost těchto vlastností. Zda platí vlastnosti $(V2)-(V6)$ shrneme v následující Větě, kterou si následně dokážeme.

Věta 3. *Uvažujme varianty výběrového kvantilu založené na lineární interpolaci, tedy $\hat{Q}_i, i \in \{4,5,6,7,8,9\}$, ze sekce 1.3. Pak platí*

- (i) *vlastnost* $(V2)$ *platí pro všechny tyto výběrové kvantily*
- (ii) *vlastnost* $(V3)$ *platí pouze pro* $\hat{Q}_5$
- (iii) *vlastnost* $(V4)$ *platí pro všechny tyto výběrové kvantily mimo* $\hat{Q}_4$
- (iv) *vlastnost* $(V5)$ *platí pro všechny tyto výběrové kvantily mimo* $\hat{Q}_4$ *a* $\hat{Q}_7$
- (v) *vlastnost* $(V6)$ *platí pro všechny tyto výběrové kvantily mimo* $\hat{Q}_4$

Nezapomeňme, že při ověřování vlastností $(V3), (V4), (V5)$ budeme předpokládat, že distribuční funkce náhodného výběru F_X je spojitá. Pro ověření ostatních vlastností takový předpoklad nepotřebujeme.

Důkaz. (i) K důkazu využijeme přepis (1.2) a pomocí něj ověříme první tvrzení Věty postupně pro všechny varianty výběrového kvantilu.

$$\begin{aligned} \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_4(p)]} &\ge \lfloor pn \rfloor &&\ge \lfloor pn \rfloor \\ \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_5(p)]} &\ge \left\lfloor pn + \tfrac{1}{2} \right\rfloor &&\ge \lfloor pn \rfloor \\ \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_6(p)]} &\ge \lfloor p(n+1) \rfloor &&\ge \lfloor pn \rfloor \\ \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_7(p)]} &\ge \lfloor p(n-1)+1 \rfloor &&\ge \lfloor pn \rfloor \\ \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_8(p)]} &\ge \left\lfloor p\left(n + \tfrac{1}{3}\right) + \tfrac{1}{3} \right\rfloor &&\ge \lfloor pn \rfloor \\ \textstyle\sum_{k=1}^{n} \mathbb{1}_{[X_k \le \hat{Q}_9(p)]} &\ge \left\lfloor p\left(n + \tfrac{1}{4}\right) + \tfrac{3}{8} \right\rfloor &&\ge \lfloor pn \rfloor \end{aligned}$$

Je-li distribuční funkce náhodného výběru spojitá, dostaneme ve všech řádcích namísto první nerovnosti rovnost. Toho využijeme v další části důkazu.
(ii) Nyní se podíváme na vlastnost $(V3)$ a pomocí přepisu (1.3) ukážeme, že tato vlastnost platí pouze pro pátou podobu výběrového kvantilu. K tomu je potřeba nahlédnout do předchozího bodu tohoto důkazu.

$$\begin{array}{lllll}
\sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_4(1-p)]} = & \lfloor pn+1 \rfloor & \neq \lfloor pn \rfloor & = \sum_{k=1}^{n} \mathbb{1}_{[X_k \leq \hat{Q}_4(p)]} \\
\sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_5(1-p)]} = & \left\lfloor pn+\frac{1}{2} \right\rfloor & = \left\lfloor pn+\frac{1}{2} \right\rfloor & = \sum_{k=1}^{n} \mathbb{1}_{[X_k \leq \hat{Q}_5(p)]} \\
\sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_6(1-p)]} = & \lfloor p(n-1)+1 \rfloor & \neq \lfloor p(n+1) \rfloor & = \sum_{k=1}^{n} \mathbb{1}_{[X_k \leq \hat{Q}_6(p)]} \\
\sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_7(1-p)]} = & \lfloor p(n+1) \rfloor & \neq \lfloor p(n-1)+1 \rfloor & = \sum_{k=1}^{n} \mathbb{1}_{[X_k \leq \hat{Q}_7(p)]} \\
\sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_8(1-p)]} = & \left\lfloor p\left(n-\frac{1}{3}\right)+\frac{2}{3} \right\rfloor & \neq \left\lfloor p\left(n+\frac{1}{3}\right)+\frac{1}{3} \right\rfloor & = \sum_{k=1}^{n} \mathbb{1}_{[X_k \leq \hat{Q}_8(p)]} \\
\sum_{k=1}^{n} \mathbb{1}_{[X_k \geq \hat{Q}_9(1-p)]} = & \left\lfloor p\left(n-\frac{1}{4}\right)+\frac{5}{8} \right\rfloor & \neq \left\lfloor p\left(n+\frac{1}{4}\right)+\frac{3}{8} \right\rfloor & = \sum_{k=1}^{n} \mathbb{1}_{[X_k \leq \hat{Q}_9(p)]}
\end{array}$$

(*iii*) V tomto bodě budeme ověřovat vlastnost $(V4)$ a využijeme následujícího přepisu

$$\hat{Q}_i^{-1}(X_{(k)}) + \hat{Q}_i^{-1}(X_{(n-k+1)}) = 1 \Leftrightarrow p_k + p_{n-k+1} = 1.$$

Ekvivalence platí, neboť pro výběrové kvantily založené na lineární interpolaci je $\hat{Q}_i^{-1}(X_{(k)}) = p_k$.
Pomocí toho ukážeme, že vlastnost $(V4)$ platí pro všechny výběrové kvantily mimo čtvrté podoby výběrového kvantilu.

$$\begin{array}{lllll}
\hat{Q}_4(p): & \frac{k}{n} + \frac{n-k+1}{n} & = \frac{n-1}{n} & \neq 1 \\
\hat{Q}_5(p): & \frac{k-\frac{1}{2}}{n} + \frac{n-k+1-\frac{1}{2}}{n} & = \frac{n}{n} & = 1 \\
\hat{Q}_6(p): & \frac{k}{n+1} + \frac{n-k+1}{n+1} & = \frac{n+1}{n+1} & = 1 \\
\hat{Q}_7(p): & \frac{k-1}{n-1} + \frac{n-k+1-1}{n-1} & = \frac{n-1}{n-1} & = 1 \\
\hat{Q}_8(p): & \frac{k-\frac{1}{3}}{n+\frac{1}{3}} + \frac{n-k+1-\frac{1}{3}}{n+\frac{1}{3}} & = \frac{n+\frac{1}{3}}{n+\frac{1}{3}} & = 1 \\
\hat{Q}_9(p): & \frac{k-\frac{3}{8}}{n+\frac{1}{4}} + \frac{n-k+1-\frac{3}{8}}{n+\frac{1}{4}} & = \frac{n+\frac{1}{4}}{n+\frac{1}{4}} & = 1
\end{array}$$

(*iv*) Podobně jako v předchozím bodě můžeme vlastnost $(V5)$ přepsat jako

$$\hat{Q}_i^{-1}(X_{(1)}) > 0 \text{ a } \hat{Q}_i^{-1}(X_{(n)}) < 1 \Leftrightarrow p_1 > 0 \text{ a } p_n < 1.$$

Pak snadno ověříme tuto vlastnost:

$$\begin{array}{llllll}
\hat{Q}_4^{-1}(X_{(1)}) = & \frac{1}{n} & > 0, & \hat{Q}_4^{-1}(X_{(n)}) = & \frac{n}{n} & = 1 \\
\hat{Q}_5^{-1}(X_{(1)}) = & \frac{\frac{1}{2}}{n} & > 0, & \hat{Q}_5^{-1}(X_{(n)}) = & \frac{n-\frac{1}{2}}{n} & < 1 \\
\hat{Q}_6^{-1}(X_{(1)}) = & \frac{1}{n+1} & > 0, & \hat{Q}_6^{-1}(X_{(n)}) = & \frac{n}{n+1} & < 1 \\
\hat{Q}_7^{-1}(X_{(1)}) = & \frac{1-1}{n-1} & = 0, & \hat{Q}_7^{-1}(X_{(n)}) = & \frac{n-1}{n-1} & = 1 \\
\hat{Q}_8^{-1}(X_{(1)}) = & \frac{\frac{2}{3}}{n+\frac{1}{3}} & > 0, & \hat{Q}_8^{-1}(X_{(n)}) = & \frac{n-\frac{1}{3}}{n+\frac{1}{3}} & < 1 \\
\hat{Q}_9^{-1}(X_{(1)}) = & \frac{\frac{5}{8}}{n+\frac{1}{4}} & > 0, & \hat{Q}_9^{-1}(X_{(n)}) = & \frac{n-\frac{3}{8}}{n+\frac{1}{4}} & < 1
\end{array}$$

Odtud vidíme, že vlastnost $(V5)$ je splněna pouze pro výběrové kvantily $\hat{Q}_5$, $\hat{Q}_6$, $\hat{Q}_8$ a $\hat{Q}_9$.
(*v*) Zbývá už jen poslední vlastnost $(V6)$. Ukážeme, že platí pro všechny spojité definice výběrového kvantilu vyjma čtvrté definice.

$$\begin{array}{lll}
\hat{Q}_4\left(\frac{1}{2}\right) & = & \left(1-\frac{n}{2}+\left\lfloor \frac{n}{2} \right\rfloor\right) X_{(\lfloor \frac{n}{2} \rfloor)} + \left(\frac{n}{2}-\left\lfloor \frac{n}{2} \right\rfloor\right) X_{(\lfloor \frac{n}{2}+1 \rfloor)} \\
& \overset{n=2l}{=\!=} & (1-l+\lfloor l \rfloor) X_{(l)} + (l-\lfloor l \rfloor) X_{(l+1)} \\
& \overset{n=2l}{=\!=} & X_{(l)} \\
& \overset{n=2l+1}{=\!=} & \left(1-\frac{2l+1}{2}+\left\lfloor \frac{2l+1}{2} \right\rfloor\right) X_{(\lfloor \frac{2l+1}{2} \rfloor)} + \left(\frac{2l+1}{2}-\left\lfloor \frac{2l+1}{2} \right\rfloor\right) X_{(\lfloor \frac{2l+1}{2} \rfloor+1)} \\
& \overset{n=2l+1}{=\!=} & \frac{X_{(l)}+X_{(l+1)}}{2}
\end{array}$$

$$
\begin{aligned}
\hat{Q}_5\left(\tfrac{1}{2}\right) &= \left(1-\tfrac{n+1}{2}+\left\lfloor\tfrac{n+1}{2}\right\rfloor\right)X_{(\lfloor\frac{n+1}{2}\rfloor)}+\left(\tfrac{n+1}{2}+\left\lfloor\tfrac{n+1}{2}\right\rfloor\right)X_{(\lfloor\frac{n+1}{2}\rfloor+1)}\\
&\overset{n=2l}{=\!=} \left(\tfrac{1}{2}-l+\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{1}{2}\rfloor)}+\left(l+\tfrac{1}{2}-\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{3}{2}\rfloor)}\\
&\overset{n=2l}{=\!=} \tfrac{X_{(l)}+X_{(l+1)}}{2}\\
&\overset{n=2l+1}{=\!=} (-l+\lfloor l+1\rfloor)X_{(\lfloor l+1\rfloor)}+(l+1-\lfloor l+1\rfloor)X_{(\lfloor l+1\rfloor+1)}\\
&\overset{n=2l+1}{=\!=} X_{(l+1)}
\end{aligned}
$$

$$
\begin{aligned}
\hat{Q}_6\left(\tfrac{1}{2}\right) &= \left(1-\tfrac{n+1}{2}+\left\lfloor\tfrac{n+1}{2}\right\rfloor\right)X_{(\lfloor\frac{n+1}{2}\rfloor)}+\left(\tfrac{n+1}{2}-\left\lfloor\tfrac{n+1}{2}\right\rfloor\right)X_{(\lfloor\frac{n+1}{2}\rfloor+1)}\\
&\overset{n=2l}{=\!=} \left(\tfrac{1}{2}-l+\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{1}{2}\rfloor)}+\left(l+\tfrac{1}{2}-\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{1}{2}\rfloor+1)}\\
&\overset{n=2l}{=\!=} \tfrac{X_{(l)}+X_{(l+1)}}{2}\\
&\overset{n=2l+1}{=\!=} (1-(l+1)+\lfloor l+1\rfloor)X_{(\lfloor l+1\rfloor)}+(l+1-\lfloor l+1\rfloor)X_{(\lfloor l+1\rfloor+1)}\\
&\overset{n=2l+1}{=\!=} X_{(l+1)}
\end{aligned}
$$

$$
\begin{aligned}
\hat{Q}_7\left(\tfrac{1}{2}\right) &= \left(-\tfrac{n-1}{2}+\left\lfloor\tfrac{n-1}{2}+1\right\rfloor\right)X_{(\lfloor\frac{n-1}{2}+1\rfloor)}+\left(\tfrac{n-1}{2}+1-\left\lfloor\tfrac{n-1}{2}+1\right\rfloor\right)X_{(\lfloor\frac{n-1}{2}+2\rfloor)}\\
&\overset{n=2l}{=\!=} \left(-l+\tfrac{1}{2}+\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{1}{2}\rfloor)}+\left(l+\tfrac{1}{2}-\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{3}{2}\rfloor)}\\
&\overset{n=2l}{=\!=} \tfrac{X_{(l)}+X_{(l+1)}}{2}\\
&\overset{n=2l+1}{=\!=} (1-l-1+\lfloor l+1\rfloor)X_{(+1\rfloor)}+(l+1-\lfloor l+1\rfloor)X_{(\lfloor l+2\rfloor)}\\
&\overset{n=2l+1}{=\!=} X_{(l+1)}
\end{aligned}
$$

$$
\begin{aligned}
\hat{Q}_8\left(\tfrac{1}{2}\right) &= \left(\tfrac{2}{3}-\tfrac{n+\frac{1}{3}}{2}+\left\lfloor\tfrac{n+\frac{1}{3}}{2}+\tfrac{1}{3}\right\rfloor\right)X_{\left(\left\lfloor\frac{n+\frac{1}{3}}{2}+\frac{1}{3}\right\rfloor\right)}+\\
&\quad+\left(\tfrac{n+\frac{1}{3}}{2}+\tfrac{1}{3}-\left\lfloor\tfrac{2+\frac{1}{3}}{2}+\tfrac{1}{3}\right\rfloor\right)X_{\left(\left\lfloor\frac{n+\frac{1}{3}}{2}+\frac{1}{3}\right\rfloor+1\right)}\\
&\overset{n=2l}{=\!=} \left(\tfrac{1}{2}-l+\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{1}{2}\rfloor)}+\left(l+\tfrac{1}{2}-\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{3}{2}\rfloor)}\\
&\overset{n=2l}{=\!=} \tfrac{X_{(l)}+X_{(l+1)}}{2}\\
&\overset{n=2l+1}{=\!=} (-l+\lfloor l+1\rfloor)X_{(l+1)}+(l+1-\lfloor l+1\rfloor)X_{(\lfloor l+\frac{5}{3}\rfloor)}\\
&\overset{n=2l+1}{=\!=} X_{(l+1)}
\end{aligned}
$$

$$
\begin{aligned}
\hat{Q}_9(p) &= \left(\tfrac{5}{8}-\tfrac{n+\frac{1}{4}}{2}+\left\lfloor\tfrac{n+\frac{1}{4}}{2}+\tfrac{3}{8}\right\rfloor\right)X_{\left(\left\lfloor\frac{n+\frac{1}{4}}{2}+\frac{3}{8}\right\rfloor\right)}+\\
&\quad+\left(\tfrac{n+\frac{1}{4}}{2}+\tfrac{3}{8}-\left\lfloor\tfrac{n+\frac{1}{4}}{2}+\tfrac{3}{8}\right\rfloor\right)X_{\left(\left\lfloor\frac{n+\frac{1}{4}}{2}+\frac{11}{8}\right\rfloor\right)}\\
&\overset{n=2l}{=\!=} \left(\tfrac{1}{2}-l+\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{1}{2}\rfloor)}+\left(l+\tfrac{1}{2}-\left\lfloor l+\tfrac{1}{2}\right\rfloor\right)X_{(\lfloor l+\frac{3}{2}\rfloor)}\\
&\overset{n=2l}{=\!=} \tfrac{X_{(l)}+X_{(l+1)}}{2}\\
&\overset{n=2l+1}{=\!=} (-l+\lfloor l+1\rfloor)X_{(\lfloor l+1\rfloor)}+(l+1-\lfloor l+1\rfloor)X_{(\lfloor l+2\rfloor)}\\
&\overset{n=2l+1}{=\!=} X_{(l+1)}
\end{aligned}
$$

□

Nyní už o všech podobách výběrového kvantilu z kapitoly 1 víme, které vlastnosti pro ně platí a které ne. Na závěr si to pro přehlednost shrneme tabulkou 2.1.

Zjistili jsme, že jediná podoba výběrového kvantilu, která splňuje všechny po-

Tabulka 2.1: Shrnutí vlastností výběrových kvantilů

Pořadí definice	$V1$	$V2$	$V3$	$V4$	$V5$	$V6$
1		✓				✓(pro n liché)
2		✓				✓
3		✓				
4	✓	✓				
5	✓	✓	✓	✓	✓	✓
6	✓	✓		✓	✓	✓
7	✓	✓		✓		✓
8	✓	✓		✓	✓	✓
9	✓	✓		✓	✓	✓

žadované vlastnosti, je pátá podoba, tedy $\hat{Q}_5(p)$.

V úvodu bylo řečeno, že ve statistickém softwaru R můžeme využít všech devět variant výběrového kvantilu. Ačkoli $\hat{Q}_5(p)$ je jediný výběrový kvantil, který splňuje všechny požadované vlastnosti, v programu R je defaultně nastavena sedmá podoba výběrového kvantilu, tedy $\hat{Q}_7(p)$.

Ve statistickém softwaru SAS je naimplementováno pouze devět variant výběrového kvantilu. Je zajímavé že pátá varianta $\hat{Q}_5(p)$ mezi nimi není. V tomto programu se můžeme setkat s výběrovými kvantily $\hat{Q}_1(p), \hat{Q}_2(p), \hat{Q}_3(p), \hat{Q}_4(p)$ a $\hat{Q}_6(p)$.

3. Některé statistické vlastnosti

V této kapitole se podrobněji podíváme na některé statistické vlastnosti výběrových kvantilů uvedených v první kapitole, např. se budeme zabývat konstrukcí intervalu spolehlivosti.

Nejprve by bylo dobré zmínit, že všechny uvedené podoby výběrového kvantilu jsou konzistentními odhady teoretického kvantilu $u_X(p)$ za předpokladu spojité a rostoucí distribuční funkce F_X na nějakém okolí bodu $u_X(p)$.

Tvrzení 4. *Nechť náhodný výběr $X_1, \ldots, X_n$ má rozdělení dané distribuční funkcí F_X, která je spojitá a rostoucí na nějakém okolí bodu $u_X(p)$, $p \in (0,1)$. Pak $\forall i \in \{1, \ldots, 9\}$ platí*

$$\hat{Q}_i(p) \xrightarrow[n\to\infty]{P} u_X(p).$$

Poznámka. Pro $\hat{Q}_1(p)$ je tvrzení dokázáno v práci Omelka (2019, str. 53). Pomocí tohoto poznatku ukážeme konzistenci ostatních výběrových kvantilů.

Důkaz. Nechť $\epsilon > 0$. Pak k dokázání tvrzení stačí ověřit

$$\mathbf{P}\left(\hat{Q}_i(p) < u_X(p) - \epsilon\right) \xrightarrow[n\to\infty]{} 0, \tag{3.1}$$

$$\mathbf{P}\left(\hat{Q}_i(p) > u_X(p) + \epsilon\right) \xrightarrow[n\to\infty]{} 0. \tag{3.2}$$

Jelikož každá varianta výběrového kvantilu je lineární kombinací dvou po sobě jdoucích pořádkových statistik, tedy nabývá tvaru (1.1), můžeme pro nějaké $j \in \{1, \ldots, n-1\}$ psát

$$\begin{aligned}\mathbf{P}\left(\hat{Q}_i(p) < u_X(p) - \epsilon\right) &= \mathbf{P}\left((1-\gamma)X_{(j)} + \gamma X_{(j+1)} < u_X(p) - \epsilon\right) \\ &\leq \mathbf{P}\left(X_{(j)} < u_X(p) - \epsilon\right).\end{aligned}$$

Stejným způsobem jako (3.1) nyní rozepíšeme (3.2)

$$\begin{aligned}\mathbf{P}\left(\hat{Q}_i(p) > u_X(p) + \epsilon\right) &= \mathbf{P}\left((1-\gamma)X_{(j)} + \gamma X_{(j+1)} > u_X(p) + \epsilon\right) \\ &\leq \mathbf{P}\left(X_{(j+1)} > u_X(p) + \epsilon\right).\end{aligned}$$

Jelikož tvrzení platí pro $\hat{Q}_1(p)$, platí i pro ostatní varianty výběrového kvantilu, neboť za konzistence $\hat{Q}_1(p)$ vlastně platí

$$\mathbf{P}\left(X_{(j)} < u_X(p) - \epsilon\right) \xrightarrow[n\to\infty]{} 0,$$

$$\mathbf{P}\left(X_{(j+1)} > u_X(p) + \epsilon\right) \xrightarrow[n\to\infty]{} 0$$

a pravděpodobnost je vždy nezáporná.

□

Poznámka. V důkazu předchozí věty je $j = \lfloor pn + m \rfloor$, stejně tak jako jsme si zaváděli v první kapitole, a $\gamma \in [0,1]$.

Nyní si ukážeme, jak vypadá interval spolehlivosti teoretického kvantilu $u_X(p)$, založíme-li jeho konstrukci na výběrovém kvantilu $\hat{Q}_1(p)$. V práci Omelka (2019, str. 55) je konstrukce intervalu spolehlivosti ukázána za předpokladu F_X spojité a rostoucí na nějakém okolí bodu $u_X(p)$. My si zformulujeme tvrzení o konstrukci intervalu spolehlivosti bez předpokladu spojitosti distribuční funkce F_X a následně jej dokážeme.

Poznámka. Ještě než si tvrzení uvedeme, připomeneme, jak vypadá uspořádaný náhodný výběr bez předpokladu spojité distribuční funkce. V takovém případě, shodují-li se napozorované hodnoty, volíme jejich pořadí náhodně tak, aby bylo zachováno $X_{(1)} \leq \cdots \leq X_{(n)}$.

Tvrzení 5. *Nechť $X_1,\ldots,X_n$ je náhodný výběr s distribuční funkcí F_X, která je rostoucí na nějakém okolí bodu $u_X(p)$. Pak uzavřený interval $[X_{(k_L)},X_{(k_U)}]$, kde $k_L = \left\lfloor \frac{1}{2} + np - u_{1-\frac{\alpha}{2}}\sqrt{np(1-p)} \right\rfloor$ a $k_U = \left\lceil \frac{1}{2} + np + u_{1-\frac{\alpha}{2}}\sqrt{np(1-p)} \right\rceil$, je interval spolehlivosti pro $u_X(p)$, založený na odhadu $\hat{Q}_1(p)$, s asymptotickou pravděpodobností pokrytí alespoň $1-\alpha$, kde $\alpha \in (0,1)$.*

Poznámka. Zde u_α značí α-kvantil normovaného normálního rozdělení a v důkazu tvrzení budeme symbolem Φ značit distribuční funkci tohoto rozdělení.

Důkaz. Tvrzení vlastně říká, že

$$\mathbf{P}\left(\left[X_{(k_L)},X_{(k_U)}\right] \ni u_X(p)\right) \xrightarrow[n\to\infty]{} 1-\beta \geq 1-\alpha,$$

kde $\beta \in (0,1), \beta \leq \alpha$. K tomu stačí ověřit, že platí

$$\mathbf{P}\left(X_{(k_L)} > u_X(p)\right) \xrightarrow[n\to\infty]{} \frac{\beta}{2} \leq \frac{\alpha}{2}, \tag{3.3}$$

$$\mathbf{P}\left(X_{(k_U)} < u_X(p)\right) \xrightarrow[n\to\infty]{} \frac{\beta}{2} \leq \frac{\alpha}{2}. \tag{3.4}$$

Nejprve budeme řešit první výraz (3.3).

$$\mathbf{P}\left(X_{(k_L)} > u_X(p)\right) = \mathbf{P}\left(\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq u_X(p)]} \leq k_L - 1\right) = \mathbf{P}\left(Bi(n,F_X(u_X(p))) \leq k_L - 1\right),$$

kde jsme využili čtvrtou vlastnost empirické distribuční funkce, která je uvedena v úvodu ve Větě 1. V celém textu budeme výrazem $\mathbf{P}\left(Bi(n,p) \leq k\right)$ rozumět pravděpodobnost, že náhodná veličina X, která se řídí binomickým rozdělením s parametry n, p, je menší nebo rovna k, neboli $\mathbf{P}(X \leq k)$, kde $X \sim Bi(n,p)$.
Z vlastností kvantilu platí $F_X(u_X(p)) \geq p$ (přičemž rovnost nastává, pokud je F_X spojitá v $u_X(p)$). Dále víme, že pro dvě binomická rozdělení se shodným prvním parametrem (tedy se stejným počtem pokusů) platí, že distribuční funkce toho rozdělení, jehož pravděpodobnost úspěchu je větší, je menší nebo rovna distribuční funkci druhého rozdělení s menší pravděpodobností úspěchu. Tedy označíme-li $F_X(u_X(p)) = p' \geq p$, pak platí

$$\mathbf{P}\left(Bi(n,F_X(u_X(p))) \leq k_L - 1\right) = \mathbf{P}\left(Bi(n,p') \leq k_L - 1\right) \leq \mathbf{P}\left(Bi(n,p) \leq k_L - 1\right).$$

Takže nám vlastně stačí ukázat, že platí

$$\mathbf{P}\left(Bi(n,p) \leq k_L - 1\right) \xrightarrow[n\to\infty]{} \frac{\gamma}{2} \leq \frac{\alpha}{2},$$

kde $\gamma \in (0,1), \beta \leq \gamma \leq \alpha$. Nyní směřujeme k použití centrální limitní věty, a jelikož chceme binomické rozdělení, které je diskrétní, aproximovat normálním rozdělením, které je spojité, stejně tak jako v práci Omelka (2019, str. 55) použijeme korekci zvanou oprava na spojitost a dále budeme vycházet z rovnosti

$$\mathbf{P}\left(X_{k_L} > u_X(p)\right) = \mathbf{P}\left(Bi(n,p) < k_L - \frac{1}{2}\right).$$

To můžeme psát, jelikož platí následující.

$$\mathbf{P}\left(X_{(k_L)} > u_X(p)\right) = \mathbf{P}\left(\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq u_X(p)]} \leq k_L - 1\right) = \mathbf{P}\left(\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq u_X(p)]} < k_L\right)$$

Odtud pomocí centrální limitní věty a vlastností binomického rozdělení dostaneme

$$\mathbf{P}\left(Bi(n,p) < k_L - \frac{1}{2}\right) = \mathbf{P}\left(\frac{Bi(n,p) - np}{\sqrt{np(1-p)}} < \frac{k_L - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \xrightarrow[n\to\infty]{} \Phi\left(\frac{k_L - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

Řešením nerovnosti

$$\Phi\left(\frac{k_L - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \leq \frac{\alpha}{2}$$

je pak $k_L \leq \frac{1}{2} + np - u_{1-\frac{\alpha}{2}}\sqrt{np(1-p)}$. Jelikož chceme takové k_L co největší, je $k_L = \left\lfloor \frac{1}{2} + np - u_{1-\frac{\alpha}{2}}\sqrt{np(1-p)} \right\rfloor$.

Nyní podobným způsobem vyřešíme druhý výraz (3.4)

$$\mathbf{P}\left(X_{(k_U)} < u_X(p)\right) = \mathbf{P}\left(\sum_{i=1}^{n} \mathbb{1}_{[X_i < u_X(p)]} > k_U - 1\right).$$

Jelikož pracujeme s výběrem, který má konečný rozsah, tak s pravděpodobností 1 existuje $\epsilon > 0$ tak, že

$$\mathbb{1}_{[X_i < u_X(p)]} = \mathbb{1}_{[X_i \leq u_X(p) - \epsilon]} \quad \forall i \in \{1, \ldots, n\}.$$

Pak můžeme psát

$$\mathbf{P}\left(\sum_{i=1}^{n} \mathbb{1}_{[X_i < u_X(p)]} > k_U - 1\right) = \mathbf{P}\left(\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq u_X(p) - \epsilon]} > k_U - 1\right).$$

Označíme-li $F_X(u_X(p) - \epsilon) = p'' < p$ (nerovnost platí z vlastností teoretického kvantilu), pak podobně jako při řešení předchozího výrazu (3.3) využijeme čtvrté vlastnosti empirické distribuční funkce a vlastností distribuční fukce binomických rozdělení se shodným prvním parametrem a různými pravděpodobnostmi úspěchu a můžeme psát

$$\mathbf{P}\left(\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq u_X(p) - \epsilon]} > k_U - 1\right) = \mathbf{P}\left(Bi(n,p'') > k_U - 1\right) =$$

$= 1 - \mathbf{P}\left(Bi(n,p'') \leq k_U - 1\right) \leq 1 - \mathbf{P}\left(Bi(n,p) \leq k_U - 1\right) = \mathbf{P}\left(Bi(n,p) > k_U - 1\right).$

Stejně tak jako v předchozím případě použijeme korekci zvanou oprava na spojitost a dále pro použití centrální limitní věty budeme vycházet z rovnosti

$$\mathbf{P}\left(X_{(k_U)} < u_X(p)\right) = \mathbf{P}\left(Bi(n,p) > k_U - \frac{1}{2}\right).$$

To můžeme psát, neboť platí následující rovnosti.

$$\mathbf{P}\left(X_{(k_U)} < u_X(p)\right) = \mathbf{P}\left(\sum_{i=1}^{n} \mathbb{1}_{[X_i < u_X(p)]} > k_U - 1\right) = \mathbf{P}\left(\sum_{i=1}^{n} \mathbb{1}_{[X_i < u_X(p)]} \geq k_U\right)$$

Pak stejně jako v předchozím případě za využití vlastností binomického rozdělení a použitím centrální limitní věty můžeme psát

$$\mathbf{P}\left(Bi(n,p) > k_U - \frac{1}{2}\right) = \mathbf{P}\left(\frac{Bi(n,p) - np}{\sqrt{np(1-p)}} > \frac{k_U - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right),$$

$$\mathbf{P}\left(\frac{Bi(n,p) - np}{\sqrt{np(1-p)}} > \frac{k_U - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \xrightarrow[n\to\infty]{} 1 - \Phi\left(\frac{k_U - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

Jako řešení nerovnosti

$$1 - \Phi\left(\frac{k_U - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \leq \frac{\alpha}{2}$$

dostaneme $k_U \geq \frac{1}{2} + np + u_{1-\frac{\alpha}{2}}\sqrt{np(1-p)}$. Jelikož chceme takové k_U co nejmenší, je $k_U = \left\lceil \frac{1}{2} + np + u_{1-\frac{\alpha}{2}}\sqrt{np(1-p)} \right\rceil$ a tím je důkaz hotov.

□

Závěr

V této práci jsme se seznámili s devíti často používanými výběrovými kvantily, které odhadují teoretický kvantil, je-li neznámý. Uvedli jsme si šest vlastností, které po výběrovém kvantilu požadujeme a rozebrali jsme, proč je vhodné, aby výběrový kvantil tyto vlastnosti splňoval. Následně jsme všechny uvedené podoby výběrového kvantilu dle těchto vlastností porovnali a zjistili jsme, že všechny požadované vlastnosti splňuje pouze jedna z nich, a to $\hat{Q}_5(p)$. Nicméně se ukázalo, že například ve statistickém softwaru SAS není tento způsob odhadování teoretického kvantilu vůbec implementován.

V poslední kapitole jsme se zabývali některými statistickými vlastnostmi uvedených výběrových kvantilů. Ukázali jsme, že všechny varianty výběrového kvantilu jsou konzistentními odhady teoretického kvantilu. Dále jsme odvodili interval spolehlivosti pro teoretický kvantil, je-li jeho konstrukce založena na první podobě výběrového kvantilu $\hat{Q}_1(p)$, a to za obecnějších předpokladů, než je uvedeno v práci Omelka (2019).

Seznam použité literatury

Dupač, V. a Hušková, M. (2001). *Pravděpodobnost a matematická statistika*. Nakladatelství Karolinum, Praha. ISBN 80-246-0009-9.

Hyndman, R. J. a Fan, Y. (2009). Sample Quantiles in Statistical Packages. *The Americian Statistician*, **50**(4), 361–365.

Omelka, M. (2019). Poznámky k přednášce NMSA331 Matematická statistika. Naposledy navštíveno 15. 4. 2019. URL `https://www.karlin.mff.cuni.cz/~omelka/Soubory/nmsa331/ms1.pdf`.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.