Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Markéta Horejšová

Intervaly spolehlivosti pro kvantily

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. Mgr. Michal Kulich, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2016

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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 24. května 2016 Podpis autora

Název práce: Intervaly spolehlivosti pro kvantily

Autor: Markéta Horejšová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Mgr. Michal Kulich, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Náplní této práce je výklad různých metod k získání simultánních intervalů spolehlivosti jak pro jeden kvantil, tak i pro několik různých kvantilů odhadovaných z týchž dat. Větší část je zaměřena na neparametrické přístupy, mezi které patří například metoda založená na Kolmogorovově-Smirnovově statistice, výběrovém kvantilu nebo na multinomickém rozdělení. Zvláštní důraz je pak kladen na nedávno navrženou metodu založenou na multinomickém rozdělení. Dále práce vykládá parametrický přístup konstrukce simultánních intervalů spolehlivosti pro kvantily specializovaný na data z normálního rozdělení a představuje jeho různé modifikace. Popsané teoretické metody jsou následně prověřeny v simulacích na náhodně generovaných datech.

Klíčová slova: intervaly spolehlivosti, kvantily, multinomické rozdělení, normální rozdělení, Kolmogorovova-Smirnovova statistika

Title: Confidence Intervals for Quantiles

Author: Markéta Horejšová

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Mgr. Michal Kulich, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this thesis, various construction methods for simultaneous confidence intervals for quantiles are explained. Among nonparametric approaches, a special emphasis is dedicated to a recent method based on a multinomial distribution for calculating the overall confidence level of confidence intervals for all quantiles of interest using an efficient recursive algorithm, which is also described. Furthermore, a method based on Kolmogorov-Smirnov statistic or an asymptotic method using empirical distribution function and order statistics for quantile estimate are presented. A special parametric method for several quantiles of a normally distributed population is introduced along with a few of its modifications. Subsequently, a simulation is run to test the real coverage of the described theoretical methods.

Keywords: confidence intervals, quantiles, multinomial distribution, normal distribution, Kolmogorov–Smirnov statistic

Chtěla bych poděkovat svému vedoucímu doc. Mgr. Michalovi Kulichovi, Ph.D. za odborné vedení a cenné rady, které mi pomohly bakalářskou práci vypracovat.

Obsah

Úvod

V této práci si vyložíme různé metody k získání simultánních intervalů spolehlivosti jak pro jeden kvantil, tak i pro několik různých kvantilů odhadovaných z týchž dat. Mezi neparametrickými přístupy si ukážeme klasickou binomickou metodu nebo metodu založenou na Kolmogorovově-Smirnovově statistice. Odvodíme si asymptotickou metodu založenou na výběrovém kvantilu a zvláštní důraz pak budeme klást na nedávno navrženou metodu využívající multinomické rozdělení, na kterém je postaven rekurzivní algoritmus k výpočtu celkové spolehlivosti intervalů spolehlivosti pro všechny kvantily.

Dále si vyložíme parametrický přístup konstrukce simultánních intervalů spolehlivosti pro kvantily specializovaný na data z normálního rozdělení a uvedeme si jeho různé modifikace.

Popsané teoretické metody následně prověříme v simulacích na náhodně generovaných datech z různých rozdělení.

1. Neparametrický model

1.1 Úvod

Předpokládejme, že máme $X_1, \ldots, X_n$, $n \in \mathbb{N}$, náhodný výběr z neznámého spojitého rozdělení se sdruženou distribuční funkcí F. Našim cílem je najít simultánní intervaly spolehlivosti pro kvantily x_q pro všechna $q \in Q = \{q_1, \ldots, q_k, 0 < q_1 < \cdots < q_k < 1\}$, $k \in \mathbb{N}$, tak, aby celková spolehlivost těchto intervalů byla $1 - \alpha$, $\alpha \in (0,1)$.

V této kapitole si na základě základních znalostí ze statistiky (viz Anděl (2005)) ukážeme, jak sestrojit intervaly spolehlivosti pro jeden kvantil s využitím pořádkových statistik a binomického rozdělení a jak tento postup rozšířit na konstrukci intervalů spolehlivosti pro více kvantilů zároveň. Dále si předvedeme metodu založenou na Kolmogorovově-Smirnovově statistice, a to jak přesnou, tak i asymptotickou. Pro malé množství pozorování si pak ukážeme, jak lze pomocí multinomického rozdělení rekurzivně spočítat celkovou spolehlivost intervalů pro všechny kvantily a v závěru tento rekurzivní algoritmus využijeme ke konstrukci intervalů spolehlivosti.

1.2 Intervaly spolehlivosti pro jeden kvantil

1.2.1 Metoda využívající binomické rozdělení

V této části si ukážeme, jak využitím binomického rozdělení získat interval o spolehlivosti větší nebo rovné $1 - \alpha$ pro jeden kvantil x_q, $q \in (0,1)$.

Uspořádejme si náhodný výběr $X_1, \ldots, X_n$ a označme $X_{(k)}$, $k \in (1, \ldots, n)$, k-tou pořádkovou statistiku a dodefinujme si $X_{(0)} := -\infty$ a $X_{(n+1)} := \infty$. Tedy platí $X_{(0)} < X_{(1)} \leq \cdots \leq X_{(n)} < X_{(n+1)}$. Dále si označme $S(x) = \sum_{i=1}^{n} \mathbb{1}_{[X_i \leq x]}$. Všimněme si, že $S(x_q)$ má binomické rozdělení $\mathrm{Bi}(n,q)$, neboť náhodná veličina $\mathbb{1}_{[X_i \leq x_q]}$ má alternativní rozdělení $\mathrm{Alt}(q)$ a součet n náhodných veličin z alternativního rozdělení se stejným parametrem q má binomické rozdělení s parametry n a q.

Pro náhodnou veličinu $S(x_q)$ jsme schopni nalézt takové l a u, aby zároveň pro dané $\alpha_1 \in [0,\alpha]$ platilo

$$P\left(S(x_q) < l\right) = \sum_{i=0}^{l-1} \binom{n}{i} q^i (1-q)^{n-i} \leq \alpha_1,$$

$$P\left(S(x_q) > u\right) = 1 - P\left(S(x_q) \leq u\right) = 1 - \sum_{i=0}^{u} \binom{n}{i} q^i (1-q)^{n-i} \leq \alpha - \alpha_1$$

a pro každé $l' > l$ a každé $u' < u$ už tyto nerovnosti neplatily. Po nalezení takových l a u pak

$$\begin{aligned} P(l \leq S(x_q) \leq u) &= \\ &= 1 - \left[P\left(S(x_q) < l\right) + P\left(S(x_q) > u\right)\right] \geq 1 - (\alpha_1 + \alpha - \alpha_1) = 1 - \alpha. \end{aligned}$$

Pro získání přesného intervalu spolehlivosti bychom chtěli, aby platila rovnost, ale protože binomické rozdělení je diskrétní, nemusí se nám to vždy podařit.

Poznámka. Typicky za α_1 volíme $\alpha/2$, abychom dostali symetrické intervaly spolehlivosti.

Nyní je třeba si uvědomit, že $S(x_q) \geq l$ je ekvivalentní $x_q \geq X_{(l)}$ a $S(x_q) \leq u$ je ekvivalentní $x_q < X_{(u+1)}$. Potom již snadno

$$P(X_{(l)} \leq x_q < X_{(u+1)}) = P(l \leq S(x_q) \leq u) \geq 1 - \alpha.$$

A tedy interval spolehlivosti $[X_{(l)}, X_{(u+1)})$ pro kvantil x_q má spolehlivost alespoň $1 - \alpha$.

Alternativně bychom mohli využít duality intervalů spolehlivosti a testování hypotéz. Mějme hypotézu $H_0: \ x_q = x$ proti alternativě $H_1: \ x_q \neq x$ a testujme na hladině α. Pak intervalem o spolehlivosti $1 - \alpha$ pro kvantil x_q jsou všechny hodnoty x, pro které nezamítáme hypotézu H_0.

1.2.2 Metoda využívající výběrový kvantil

Připomeňme, že funkci $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq x]}$ nazýváme empirická distribuční funkce a $\hat{x}_n(q) = X_{(k_q)}$, kde $k_q = qn$, pokud je qn celé, a $k_q = \lfloor qn \rfloor + 1$, pokud qn není celé, nazýváme výběrový q-kvantil. Platí, že $\hat{x}_n(q) = \inf\{u : \hat{F}_n(u) \geq q\}$.

Je-li $X_1, \ldots, X_n$ náhodný výběr ze spojitého rozdělení, což předpokládáme, můžeme ke konstrukci asymptotických intervalů spolehlivosti pro kvantily využít následující větu:

Věta 1. *Nechť $X_1, \ldots, X_n$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F a hustotou f, která je spojitá a nenulová v okolí $\hat{x}_n(q)$, $q \in (0,1)$. Pak:*

- $\hat{x}_n(q) \xrightarrow{P} x_q, \ n \to \infty,$
- $\sqrt{n}\ [\hat{x}_n(q) - x_q] \xrightarrow{d} N(0,\ v(q)),\ v(q) = \frac{q(1-q)}{f^2(x_q)}.$

Se znalostí asymptotického rozdělení výběrového kvantilu již můžeme konstruovat asymptotické intervaly spolehlivosti. Nejprve si uvědomme, že díky spojitosti f na okolí $\hat{x}_n(q)$ z věty o spojité transformaci platí následující implikace:

$$\frac{\hat{x}_n(q)}{x_q} \xrightarrow{P} 1,\ n \to \infty \quad \Rightarrow \quad \frac{f^2(\hat{x}_n(q))}{f^2(x_q)} \xrightarrow{P} 1,\ n \to \infty.$$

Můžeme si tedy druhé tvrzení věty přepsat do následujícího tvaru:

$$\sqrt{n}\ \frac{\hat{x}_n(q) - x_q}{\sqrt{\hat{v}(q)}} \xrightarrow{d} \mathrm{N}(0,1), \text{ kde } \hat{v}(q) = \frac{q(1-q)}{f^2(\hat{x}_n(q))}.$$

Označme z_q q-kvantil normovaného normálního rozdělení. Pak platí:

$$P\left(z_{\alpha_1} \leq \sqrt{n}\ \frac{\hat{x}_n(q) - x_q}{\sqrt{\hat{v}(q)}} \leq z_{1-(\alpha-\alpha_1)}\right) =$$
$$= P\left(\hat{x}_n(q) - z_{1-(\alpha-\alpha_1)}\sqrt{\frac{\hat{v}(q)}{n}} \leq x_q \leq \hat{x}_n(q) - z_{\alpha_1}\sqrt{\frac{\hat{v}(q)}{n}}\right)$$
$$\to\ 1 - (\alpha - \alpha_1) - \alpha_1 = 1 - \alpha,\ \alpha_1 \in [0,\alpha].$$

Tedy interval

$$\left[\hat{x}_n(q) - z_{1-(\alpha-\alpha_1)}\sqrt{\frac{\hat{v}(q)}{n}},\ \hat{x}_n(q) - z_{\alpha_1}\sqrt{\frac{\hat{v}(q)}{n}}\ \right] \tag{1.1}$$

je intervalem o asymptotické spolehlivosti $1-\alpha$ pro kvantil x_q, kde i zde tradičně volíme $\alpha_1 = \alpha/2$.

Předpokládejme, že máme odhad $\hat{f}$ hustoty f. Jestliže nahradíme hustotu v intervalu (1.1) jejím odhadem $\hat{f}$, získáme interval

$$\left[\hat{x}_n(q) - z_{1-(\alpha-\alpha_1)}\sqrt{\frac{q(1-q)}{n\hat{f}^2(\hat{x}_n(q))}},\ \hat{x}_n(q) - z_{\alpha_1}\sqrt{\frac{q(1-q)}{n\hat{f}^2(\hat{x}_n(q))}}\ \right]$$

o asymptotické spolehlivosti $1-\alpha$ pro kvantil x_q.

Problémem této metody ale je, že buď potřebujeme znát hustotu f, a v tomto případě vlastně známe i kvantily, tedy je pro ně tvorba intervalů spolehlivosti zbytečná, nebo potřebujeme hustotu odhadnout, což obecně není jednoduchá záležitost.

Protože nás zajímá tvar hustoty jen na okolí jednoho bodu, přichází v úvahu možnost použít jako odhad hustoty histogram. Připomeňme, že histogram vzniká rozdělením reálné osy na stejně velké části (intervaly) a v každé části má hodnotu rovnu počtu pozorování v této části děleno celkovým počtem pozorování. Nevýhody histogramu ovšem jsou nespojitost a velká závislost na volbě dělení (malý počet intervalů dělení způsobí, že je histogram skoro konstantní, velký počet způsobuje velkou skokovitost).

Daleko sofistikovanější neparametrickou metodou odhadu pravděpodobnostní hustoty, kterou bychom v tomto případě spíš použili, je tzv. jádrová metoda (kernel density estimation), jinak také známá jako metoda Parzenových oken, viz Parzen (1962).

1.3 Intervaly spolehlivosti pro více kvantilů

1.3.1 Bonferroniho metoda

V některých případech nás zajímá jen několik kvantilů, zpravidla tři nebo pět (kvartily, medián,...), a chceme sestrojit intervaly spolehlivosti pro tyto kvantily tak, aby jejich celková spolehlivost byla $1-\alpha$, tedy aby platilo:

$$P(x_q \in C_q \quad \forall q \in Q) = 1-\alpha,$$

kde C_q je interval spolehlivosti pro kvantil x_q. Nejjednodušším (kozervativním) přístupem k tomuto problému je využití Booleovy nerovnosti, tzv. Bonferroniho metoda.

Věta 2 (Booleova nerovnost)**.** *Nechť $A_1, A_2, \dots$ je spočetné množství jevů. Pak platí*

$$P\Big(\bigcup_i A_i\Big) \le \sum_i P(A_i).$$

Označme si C_i interval spolehlivosti pro kvantil x_{q_i} a A_i jev, že $x_{q_i} \notin C_i$, $i = 1, \ldots, k$. Zvolíme-li (například pomocí některé metody pro hledání intervalů spolehlivosti pro jeden kvantil popsané výše) C_i tak, aby $\sum_{i=1}^{k} P(A_i) \leq \alpha$, typicky $P(A_i) \leq \alpha/k \ \forall i \in \{1, \ldots, k\}$, protože chceme stejnou spolehlivost pro všechny intervaly, pak

$$
\begin{aligned}
P\left(x_{q_i} \in C_i; 1 \leq i \leq k\right) &= P\Big(\bigcap_{i=1}^{k} A_i^{\mathsf{c}}\Big) = P\Big(\Big(\bigcup_{i=1}^{k} A_i\Big)^{\mathsf{c}}\Big) = 1 - P\Big(\bigcup_{i=1}^{k} A_i\Big) \\
&\geq 1 - \sum_{i=1}^{k} P(A_i) \geq 1 - \alpha,
\end{aligned}
$$

tedy celková spolehlivost intervalů spolehlivosti pro všechny kvantily je alespoň $1 - \alpha$.

Poznámka. Součet pravděpodobností jevů je hodně hrubý horní odhad pravděpodobnosti jejich sjednocení, tedy se dá očekávat, že celková spolehlivost nalezených intervalů pro kvantily bude vyšší než $1 - \alpha$, a tedy i intervaly zbytečně široké.

1.3.2 Kolmogorovova-Smirnovova statistika

Tato část se odkazuje na knihy Anděl (2005) a Hájek a Šidák (1967) ohledně Kolmogorovovy-Smirnovovy statistiky.

Předpokládejme, že náš náhodný výběr $X_1, \ldots, X_n$ pochází ze spojitého rozdělení a označme si

$$K_n = \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right|, \ K_n^+ = \sup_{x \in \mathbb{R}} (\hat{F}_n(x) - F(x)) \text{ a } K_n^- = \sup_{x \in \mathbb{R}} (F(x) - \hat{F}_n(x)).$$

Pak $K_n = \max(K_n^+, K_n^-)$. Dále víme, že si K_n^+ a K_n^- můžeme vyjádřit jako:

$$K_n^+ = \max_{1 \leq i \leq n} \left(\frac{i}{n} - F(X_{(i)}) \right) \text{ a } K_n^- = \max_{1 \leq i \leq n} \left(F(X_{(i)}) - \frac{i-1}{n} \right).$$

Rozdělení $F(X_{(i)})$ známe. Platí, že $F(X_{(i)})$ má beta rozdělení $\mathrm{B}(i, n-i+1)$. Dále

$$\sup_{q \in (0,1)} \left| \frac{S(x_q)}{n} - q \right| = \sup_{q \in (0,1)} \left| \hat{F}_n(x_q) - F(x_q) \right| = \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| = K_n. \quad (1.2)$$

Nechť G_n je rozdělení K_n a $G_n^{-1}(\alpha)$ značí α-kvantil rozdělení G_n. Pak pro dané $\alpha_1 \in [0,\alpha]$ platí

$$
\begin{aligned}
1 - \alpha &= P\left(G_n^{-1}(\alpha_1) \leq \frac{S(x_q)}{n} - q \leq G_n^{-1}(1 - (\alpha - \alpha_1)) \quad \forall q \in (0,1) \right) = \\
&= P\left(n\left[G_n^{-1}(\alpha_1) + q\right] \leq S(x_q) \leq n\left[G_n^{-1}(1 - (\alpha - \alpha_1)) + q\right] \quad \forall q \in (0,1) \right).
\end{aligned}
$$

Položíme-li $l'_q = \min\{l \in \mathbb{Z} : l \geq n\,[G_n^{-1}(\alpha_1) + q]\}$ a $u'_q = \max\{u \in \mathbb{Z} : u \leq n\,[G_n^{-1}(1 - (\alpha - \alpha_1)) + q]\}$, pak intervaly spolehlivosti pro kvantily x_q, $q \in (0,1)$,

$$[X_{(l'_q)}, X_{(u'_q+1)})$$

mají celkovou spolehlivost přesně $1 - \alpha$.

Je-li n malé, je možné, ale těžké, zjistit tvar G_n. Kvantily tohoto rozdělení lze najít v tabulkách viz Hájek a Šidák (1967).

Je-li počet pozorování velký, využijeme asymptotického rozdělení K_n, které je známé. Platí:

$$P\left(\sqrt{n}K_n \leq y\right) \to 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2y^2},\ y \in \mathbb{R}.$$

Označme si $G_d(y) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2y^2}$ a $G_d^{-1}(\alpha)$ α-kvantil G_d. Pak vynásobíme-li (1.2) $\sqrt{n}$, dostaneme

$$\sqrt{n}\sup_{q\in(0,1)}\left|\frac{S(x_q)}{n} - q\right| \leq \sqrt{n}K_n,$$

pak také

$$P\left(\sqrt{n}\left[G_d^{-1}(\alpha_1) + \sqrt{n}q\right] \leq S(x_q) \leq \sqrt{n}\left[G_d^{-1}(1-(\alpha-\alpha_1)) + \sqrt{n}q\right]\ \forall q \in (0{,}1)\right)$$
$$= P\left(G_d^{-1}(\alpha_1) \leq \sqrt{n}\left[\frac{S(x_q)}{n} - q\right] \leq G_d^{-1}(1-(\alpha-\alpha_1))\quad \forall q \in (0{,}1)\right) \to 1-\alpha.$$

Zvolíme-li obdobně jako výše $l'_q = \min\{l \in \mathbb{Z} : l \geq \sqrt{n}\left[G_d^{-1}(\alpha_1) + \sqrt{n}q\right]\}$ a $u'_q = \max\{u \in \mathbb{Z} : u \leq \sqrt{n}\left[G_d^{-1}(1-(\alpha-\alpha_1)) + \sqrt{n}q\right]\}$, pak intervaly

$$[X_{(l'_q)}, X_{(u'_q+1)})$$

mají celkovou asymptotickou spolehlivost $1-\alpha$ pro kvantily x_q, $q \in (0{,}1)$.

1.3.3 Využití multinomického rozdělení

V této části popíšeme metodu představenou v článku Hayter (2014).

Hledáme intervaly spolehlivosti pro všechny kvantily x_q, $q \in Q = \{q_1, \ldots, q_k\}$, $0 < q_1 < \cdots < q_k < 1$, kde k je přirozené, rozumně malé číslo, a chceme, aby celková spolehlivost těchto intervalů byla $1-\alpha$, $\alpha \in (0{,}1)$. Předvedeme si, jak lze s pomocí multinomického rozdělení vypočítat celkovou spolehlivost intervalů

$$[X_{(l_i)},\ X_{(u_i+1)})$$

pro x_{q_i}, $1 \leq i \leq k$, pro dané $0 \leq l_i < u_i \leq n$, $1 \leq i \leq k$.

Označme si $\boldsymbol{W} = (W_1, \ldots, W_{k+1})^T$ náhodnou veličinu z multinomického rozdělení $\mathrm{Mult}_{k+1}(n, \boldsymbol{\pi})$, kde $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{k+1})^T = (q_1, q_2 - q_1, \ldots, q_k - q_{k-1}, 1 - q_k)^T$.

Dále je potřeba si uvědomit, že rozdělení $S(x_{q_i})$ je ekvivalentní rozdělení $\sum_{j=1}^{i} W_j$. Toto můžeme získat využitím kombinatoriky a „škatulkovou" představou multinomického rozdělení. Náhodná veličina „padne" do některé škatulky W_j, $j = 1, \ldots, i$, s pravděpodobností $p = \sum_{j=1}^{i}\pi_j = q_1 + (q_2 - q_1) + \cdots + (q_i - q_{i-1}) = q_i$. Máme-li n pokusů, pak pravděpodobnost, že nejvýše m padne do některé W_j, $j = 1, \ldots, i$, je

$$P\Big(\sum_{j=1}^{i} W_j \leq m\Big) = \sum_{k=0}^{m}\binom{n}{k} q_i^k (1-q_i)^{n-k},$$

což je ekvivalentní pravděpodobnosti $P(S(x_{q_i}) \leq m)$, neboť $S(x_{q_i})$ má binomické rozdělení Bi(n,q_i), tedy $S(x_{q_i})$ a $\sum_{j=1}^{i} W_j$ mají stejné rozdělení. Pak také platí

$$P\left(l_i \leq S(x_{q_i}) \leq u_i;\ 1 \leq i \leq k\right) = P\Big(l_i \leq \sum_{j=1}^{i} W_j \leq u_i;\ 1 \leq i \leq k\Big).$$

Připomeňme, že z definice multinomického rozdělení je

$$\begin{aligned} P\left(W_1 = w_1, \ldots, W_{k+1} = w_{k+1}\right) &= \frac{n!\pi_1^{w_1} \cdots \pi_{k+1}^{w_{k+1}}}{w_1! \cdots w_{k+1}!},\ w_i \in \mathbb{N}_0,\ \sum_{i=1}^{k+1} w_i = n, \\ &= 0, \text{ jinak.} \end{aligned}$$

Dále platí

$$\begin{aligned} &P(a_1 \leq W_1 \leq b_1, \ldots, a_{k+1} \leq W_{k+1} \leq b_{k+1}) = \\ &\qquad = \sum_{w_1=a_1}^{b_1} \cdots \sum_{w_{k+1}=a_{k+1}}^{b_{k+1}} P\left(W_1 = w_1, \ldots, W_{k+1} = w_{k+1}\right). \end{aligned}$$

Nyní si ukážeme rekurzivní algoritmus k výpočtu celkové pravděpodobnosti $P\Big(l_i \leq \sum_{j=1}^{i} W_j \leq u_i;\ 1 \leq i \leq k\Big)$, který ale na rozdíl od článku Hayter (2014) podrobněji odvodíme. Předpokládejme $0 \leq l_1 \leq ... \leq l_k \leq n$, $0 \leq u_1 \leq ... \leq u_k \leq n$ a $l_i \leq u_i$, $i = 1, \ldots ,k$. Pak

$$\begin{aligned} &P\Big(l_i \leq \sum_{j=1}^{i} W_j \leq u_i; i = 1,...,k\Big) = \\ &\quad = P(l_1 \leq W_1 \leq u_1, l_2 \leq W_1 + W_2 \leq u_2, \ldots, l_{k+1} \leq W_1 + \cdots + W_{k+1} \leq u_{k+1}) \\ &\quad = P((l_1 \leq W_1 \leq u_1, \max(l_2, W_1) \leq W_1 + W_2 \leq u_2, \\ &\qquad\qquad \max(l_3, W_1 + W_2) \leq W_1 + W_2 + W_3 \leq u_3, \ldots, \\ &\qquad\qquad \max(l_{k+1}, W_1 + \cdots + W_k) \leq W_1 + \cdots + W_{k+1} \leq u_{k+1}) = \\ &\quad = \sum_{z_1=l_1}^{u_1} P(W_1 = z_1, \max(l_2, z_1) \leq z_1 + W_2 \leq u_2, \\ &\qquad\qquad \max(l_3, z_1 + W_2) \leq z_1 + W_2 + W_3 \leq u_3, \ldots, \\ &\qquad\qquad \max(l_{k+1}, z_1 + W_2 + \cdots + W_k) \leq z_1 + W_2 + \cdots + W_{k+1} \leq u_{k+1}) = \\ &\quad = \sum_{z_1=l_1}^{u_1} P(W_1 = z_1, \max(l_2 - z_1, 0) \leq W_2 \leq u_2 - z_1, \\ &\qquad\qquad \max(l_3 - z_1, W_2) \leq W_2 + W_3 \leq u_3 - z_1, \ldots, \\ &\qquad\qquad \max(l_{k+1} - z_1, W_2 + \cdots + W_k) \leq W_2 + \cdots + W_{k+1} \leq u_{k+1} - z_1) = \\ &\quad = \sum_{z_1=l_1}^{u_1} \sum_{w_2=\max(l_2-z_1,0)}^{u_2-z_1} P(W_1 = z_1, W_2 = w_2, \\ &\qquad\qquad \max(l_3 - z_1, w_2) \leq w_2 + W_3 \leq u_3 - z_1, \ldots, \\ &\quad \max(l_{k+1} - z_1, w_2 + W_3 + \cdots + W_k) \leq w_2 + W_2 + \cdots + W_{k+1} \leq u_{k+1} - z_1) = \end{aligned}$$

$$\overset{z_2:=z_1+w_1}{=} \sum_{z_1=l_1}^{u_1} \sum_{z_2=max(l_2,z_1)}^{u_2} P(W_1 = z_1, W_2 = z_2 - z_1,$$

$$\max(l_3 - z_2, 0) \le W_3 \le u_3 - z_2, \ldots,$$

$$\max(l_{k+1} - z_2, W_3 + \cdots + W_k) \le W_3 + \cdots + W_{k+1} \le u_{k+1} - z_2) =$$

$$= \sum_{z_1=l_1}^{u_1} \sum_{z_2=max(l_2,z_1)}^{u_2} \sum_{w_3=\max(l_3-z_2,0)}^{u_3-z_2} P(W_1 = z_1, W_2 = z_2, W_3 = w_3, \ldots,$$

$$\max(l_{k+1} - z_2 - w_3, 0) \le W_4 + \cdots + W_{k+1} \le u_{k+1} - z_2 - w_3) =$$

$$\overset{z_3:=z_2+w_3}{=} \sum_{z_1=l_1}^{u_1} \sum_{z_2=max(l_2,z_1)}^{u_2} \sum_{z_3=\max(l_3,z_2)}^{u_3} P(W_1 = z_1, W_2 = z_2, W_3 = z_3, \ldots,$$

$$\max(l_{k+1} - z_3, W_4 + \cdots + W_k) \le W_4 + \cdots + W_{k+1} \le u_{k+1} - z_3) =$$

$$= \sum_{z_1=l_1}^{u_1} \sum_{z_2=\max(l_2,z_2)}^{u_2} \cdots \sum_{z_k=\max(l_k,z_{k-1})}^{u_k} P(W_1 = z_1, \ldots, W_k = z_k,$$

$$\max(l_{k+1} - z_k, 0) \le W_{k+1} \le u_{k+1} - z_k) =$$

$$= \sum_{z_1=l_1}^{u_1} \sum_{z_2=\max(l_2,z_2)}^{u_2} \cdots \sum_{z_k=\max(l_k,z_{k-1})}^{u_k} P(W_1 = z_1, \ldots, W_k = z_k, W_{k+1} = n - z_k) =$$

$$= \sum_{z_1=l_1}^{u_1} \sum_{z_2=\max(l_2,z_2)}^{u_2} \cdots \sum_{z_k=\max(l_k,z_{k-1})}^{u_k} \frac{n!\pi_1^{z_1}\pi_2^{z_2-z_1}...\pi_k^{z_k-z_{k-1}}\pi_{k+1}^{n-z_k}}{z_1!(z_2-z_1)!...(z_k-z_{k-1})!(n-z_k)!} =$$

$$= n!\pi_{k+1}^n \sum_{z_1=l_1}^{u_1} \sum_{z_2=\max(l_2,z_2)}^{u_2} \cdots \sum_{z_k=\max(l_k,z_{k-1})}^{u_k}$$

$$\frac{(\pi_1/\pi_2)^{z_1}}{z_1!} \left(\prod_{i=2}^{k} \frac{(\pi_i/\pi_{i+1})^{z_i}}{(z_i - z_{i-1})!} \right) \frac{1}{(n-z_k)!} =$$

$$= n!\pi_{k+1}^n \sum_{z_1=l_1}^{u_1} \left(\sum_{z_2=\max(l_2,z_2)}^{u_2} \cdots \left(\sum_{z_k=\max(l_k,z_{k-1})}^{u_k} \right.\right.$$

$$\left.\left. \left(\frac{1}{(n-z_k)!} \right) \frac{(\pi_k/\pi_{k+1})^{z_k}}{(z_k - z_{k-1})!} \right) \cdots \frac{(\pi_2/\pi_3)^{z_2}}{(z_2 - z_1)!} \right) \frac{(\pi_1/\pi_2)^{z_1}}{z_1!}.$$

Označíme-li

$$g_i(z_{i-1}) = \sum_{z_i=\max(l_i,z_{i-1})}^{u_i} g_{i+1}(z_i) \frac{(\pi_i/\pi_{i+1})^{z_i}}{(z_i - z_{i-1})!}$$

pro $i = 2,...,k$ s $g_{k+1}(z_k) = 1/(n-z_k)!$, můžeme výraz nahoře zapsat ve zkráceném tvaru:

$$P\Big(l_i \le \sum_{j=1}^{i} W_j \le u_i; i = 1,...,k\Big) = n!\pi_{k+1}^n \sum_{z_1=l_1}^{u_1} g_2(z_1) \frac{(\pi_1/\pi_2)^{z_1}}{z_1!}. \tag{1.3}$$

Vidíme, že zpětnou rekurzí vypočítáme celkovou spolehlivost intervalů spolehlivosti pro dané l_i, u_i, $i = 1, \ldots, k$. Otázkou zůstává, jak právě tyto hodnoty l_i a u_i volit.

1.3.4 Iterovaná Bonferroniho metoda

Umíme-li zjistit celkovou spolehlivost intervalů $[X_{(l_i)},\ X_{(u_i+1)})$ pro kvantily x_{q_i}, $i = 1,\ldots,k$, viz rekurzivní algoritmus (1.3), nabízí se k získání přesnějších intervalů spolehlivosti opakovaně použít Bonferroniho metodu viz 1.3.1 v kombinaci právě s tímto algoritmem. Idea je pro dané $\alpha \in (0{,}1)$ nalézt hodnoty l_i a u_i tak, aby každý interval $[X_{(l_i)},\ X_{(u_i+1)})$ měl svoji jednotlivou (marginální) spolehlivost větší nebo ideálně rovnu $1 - \alpha/k$. Pak pro tyto nalezené hodnoty l_i a u_i pomocí rekurzivního algoritmu spočítáme celkovou spolehlivost β. Je-li vyšší než $1 - \alpha$, použijeme znovu metodu založenou na Booleově nerovnosti, ale nalezneme nové hodnoty l'_i a u'_i tak, aby tentokrát marginální spolehlivost každého intervalu byla větší nebo ideálně rovna $1 - \alpha/k - \frac{\beta-(1-\alpha)}{k}$ („přebytečnou" spolehlivost rozdělíme mezi jednotlivé intervaly). Liší-li se l_i od l'_i nebo u_i od u'_i alespoň pro jedno $i \in \{1,\ldots,k\}$, celý postup opakujeme, dokud se nedostaneme do fáze, kdy už nedochází ke změně (vylepšení). Označíme-li $BN(1-\alpha/k)$ funkci založenou na Booleově nerovnosti vracející hodnoty l_i a u_i takové, že každý interval spolehlivosti $[X_{(l_i)},\ X_{(u_i+1)})$ pro kvantil x_{q_i} má marginální spolehlivost co nejblíže a zároveň větší rovnu $1-\alpha/k$, $i = 1,\ldots,k$, a $rek.alg(lu)$ funkci vypočítávající celkovou spolehlivost intervalů pro dané hodnoty l a u - lu, mohli bychom algoritmus sepsat takto:

```
l.u <- BN(1 - alfa/k)
beta <- rek.alg(l.u)          # beta ... celková spolehlivost
zlepsuje.se <- TRUE
while ((zlepsuje.se)&&(beta >= 1 - alfa))
{
        l'.u' <- BN(1 - alfa/k - (beta - 1 + alfa)/k)
        if (l'.u' != l.u)
        {
                l.u <- l'.u'
                beta <- rek.alg(l.u)
        }
        else
                zlepsuje.se <- FALSE
}
```

Tato metoda nám zřejmě poskytne užší intervaly spolehlivosti než obyčejná Bonferroniho metoda, a to za cenu vyšší výpočetní náročnosti. Vyplatí se tedy pro malé množství pozorování a rozumně malé k.

2. Parametrický model

2.1 Normální rozdělení

2.1.1 Úvod

V této kapitole se zaměříme na metodu specializovanou na data z normálního rozdělení představenou v článku Liu a kol. (2013).

Budeme uvažovat náhodný výběr $Y_1, \ldots, Y_n$ z normálního rozdělení $N(\mu,\sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Chceme najít interval s přesnou spolehlivostí $1 - \alpha$ pro x_q, q-tý kvantil tohoto rozdělení, $q \in (0,1)$. Víme, že pro Y z $N(\mu,\sigma^2)$ má $\frac{Y-\mu}{\sigma}$ normované normální rozdělení N(0,1) a z definice kvantilu platí $P(Y \leq x_q) = q = \Phi(z_q) = P(\frac{Y-\mu}{\sigma} \leq z_q) = P(Y \leq z_q\sigma + \mu)$, kde Φ je distribuční funkce normovaného normálního rozdělení. V našem případě tedy hledáme přesné intervaly spolehlivosti pro $\mu + z_q\sigma$, kde z_q je q-tý kvantil normovaného normálního rozdělení a $q \in Q = \{q_1, \ldots, q_k : \ 0 < q_1 < \cdots < q_k < 1\}$. K tomu potřebujeme odhad $\mu + z_q\sigma$.

2.1.2 Intervaly spolehlivosti založené na nestranném odhadu

Jedním z přístupů je uvažovat nejlepší nestranný odhad. Víme, že výběrový průměr $\overline{Y_n} = \frac{1}{n}\sum_{i=1}^n Y_i$ je nejlepší nestranný odhad střední hodnoty μ. Dále S/c, kde $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y_n})^2}$ je směrodatná odchylka a $\mathsf{E}\,[S/c] = \sigma$, je nestranný odhad σ. Pro náhodný výběr z normálního rozdělení platí, že $c = c_n = \sqrt{2}\,\Gamma(n/2)/(\sqrt{n-1}\,\Gamma((n-1)/2))$, viz Liu a kol. (2013). Pak $\overline{Y_n} + z_q S/c$ je nejlepší nestranný odhad $\mu + z_q\sigma$. Dále platí, že náhodná veličina

$$V = \frac{\overline{Y}_n + z_q S/c - (\mu + z_q\sigma)}{\sqrt{\operatorname{var}(\overline{Y}_n + z_q S/c)}}$$

má rozptyl 1 a střední hodnotu 0. Označme G_n distribuční funkci náhodné veličiny V a $G_n^{-1}(\alpha)$ její α-kvantil. Platí-li pro dané $\alpha_1 \in [0,\alpha]$

$$P\left(\frac{\overline{Y}_n + z_q S/c - (\mu + z_q\sigma)}{\sqrt{\operatorname{var}(\overline{Y}_n + z_q S/c)}} > G_n^{-1}(1 - (\alpha - \alpha_1))\right) = \alpha - \alpha_1$$

a

$$P\left(\frac{\overline{Y}_n + z_q S/c - (\mu + z_q\sigma)}{\sqrt{\operatorname{var}(\overline{Y}_n + z_q S/c)}} < G_n^{-1}(\alpha_1)\right) = \alpha_1,$$

pak

$$P\left(G_n^{-1}(\alpha_1) \leq \frac{\overline{Y}_n + z_q S/c - (\mu + z_q\sigma)}{\sqrt{\operatorname{var}(\overline{Y}_n + z_q S/c)}} \leq G_n^{-1}(1 - (\alpha - \alpha_1))\right) =$$

$$= P\Bigg(\bar{Y}_n + z_q S/c - G_n^{-1}(1-(\alpha-\alpha_1))\sqrt{\operatorname{var}(\bar{Y}_n + z_q S/c)} \le \mu + z_q\sigma$$

$$\le \bar{Y}_n + z_q S/c - G_n^{-1}(\alpha_1)\sqrt{\operatorname{var}(\bar{Y}_n + z_q S)}\Bigg) = 1-\alpha.$$

Pro jakékoliv jiné rozdělení by do této chvíle vše platilo také, neboť výběrový průměr je vždy nestranný odhad, ale museli bychom uvažovat jinou konstantu c_n, která byla vypočítaná speciálně pro normální rozdělení. Pro vyjádření rozptylu $\operatorname{var}(\overline{Y_n} + z_q S/c)$ ale využijeme nezávislosti výběrového průměru a směrodatné odchylky, která platí, máme-li náhodný výběr z normálního rozdělení, ale pro jiná rozdělení už toto obecně není pravda.

$$\operatorname{var}(\overline{Y_n} + z_q S/c) \overset{\text{nez.}}{=} \operatorname{var}(\overline{Y_n}) + z_q^2 \ \operatorname{var}(S/c) = \frac{\sigma^2}{n} + z_q^2 \left[\mathsf{E}\left[(S/c)^2\right] - (\mathsf{E}\left[S/c\right])^2\right] =$$

$$= \frac{\sigma^2}{n} + z_q^2\Big(\frac{1}{c^2}\mathsf{E}\left[S^2\right] - \sigma^2\Big) = \frac{\sigma^2}{n} + z_q^2\Big(\frac{\sigma^2}{c^2} - \sigma^2\Big) = \sigma^2\left[\frac{1}{n} + z_q^2\Big(\frac{1}{c^2} - 1\Big)\right],$$

kde jsme využili toho, že výběrový rozptyl S^2 je nestranným odhadem σ^2. Celkem tedy

$$P\Bigg(\bar{Y}_n + z_q S/c - G_n^{-1}(1-(\alpha-\alpha_1))\sigma\sqrt{\frac{1}{n} + z_q^2\Big(\frac{1}{c^2} - 1\Big)} \le \mu + z_q\sigma$$

$$\le \bar{Y}_n + z_q S/c - G_n^{-1}(\alpha_1)\sigma\sqrt{\frac{1}{n} + z_q^2\Big(\frac{1}{c^2} - 1\Big)}\Bigg) = 1-\alpha$$

a interval

$$\Bigg[\bar{Y}_n + z_q S/c - G_n^{-1}(1-(\alpha-\alpha_1))\sigma\sqrt{\frac{1}{n} + z_q^2\Big(\frac{1}{c^2} - 1\Big)},$$

$$\bar{Y}_n + z_q S/c - G_n^{-1}(\alpha_1)\sigma\sqrt{\frac{1}{n} + z_q^2\Big(\frac{1}{c^2} - 1\Big)}\ \Bigg]$$

má přesnou spolehlivost $1-\alpha$.

Zde by autoři článku (Liu a kol., 2013) nahradili σ směrodatnou odchylkou S a dostali tím interval, jehož spolehlivost se blíží k $1-\alpha$. Jinou možností by bylo nahradit σ jeho nestranným odhadem S/c, jak bylo učiněno hned na začátku v odhadu kvantilu, a interval

$$\Bigg[\bar{Y}_n + z_q S/c - G_n^{-1}(1-(\alpha-\alpha_1))\frac{S}{c}\sqrt{\frac{1}{n} + z_q^2\Big(\frac{1}{c^2} - 1\Big)},$$

$$\bar{Y}_n + z_q S/c - G_n^{-1}(\alpha_1)\frac{S}{c}\sqrt{\frac{1}{n} + z_q^2\Big(\frac{1}{c^2} - 1\Big)}\ \Bigg]$$

by pak měl přibližnou spolehlivost $1-\alpha$.

K uplatnění této metody bychom potřebovali znát kvantily rozdělení G_n, které ale známy nejsou, daly by se však najít například simulačně.

2.1.3 Intervaly spolehlivosti nevyžadující nestrannost

V této části budeme za odhad $\mu + z_q\sigma$ brát $\overline{Y_n} + z_q S$, který na rozdíl od odhadu $\bar{Y_n} + z_q S/c$ z předchozí části (2.1.2) už není nestranný. Připomeňme, že $\overline{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ je výběrový průměr a $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y_n})^2}$ směrodatná odchylka.

Intervaly spolehlivosti pro 1 kvantil

Označme T náhodnou veličinu Studentova rozdělení $T_{\nu,\delta}$ s $\nu = n-1$ stupni volnosti a parametrem necentrality δ, $t_{\nu,\delta}(q)$ q-tý kvantil $T_{\nu,\delta}$, F_T její distribuční funkci a f_T její hustotu.Víme, že T získáme podílem

$$\frac{Z + \delta}{\sqrt{V/\nu}},$$

kde Z je náhodná veličina z normálního rozdělení N(0,1) a V náhodná veličina z χ^2_ν a Z, V jsou na sobě nezávislé. V našem případě tedy protože

$$\sqrt{n}\,\frac{\overline{Y_n} - \mu}{\sigma} \sim N(0{,}1) \text{ a } \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_\nu,$$

máme

$$\frac{\sqrt{n}\,\frac{(\overline{Y}_n - \mu)}{\sigma} - \sqrt{n}z_q}{S/\sigma} \sim T_{\nu,-\sqrt{n}z_q}.$$

Pro $0 \leq \alpha_1 \leq \alpha$ platí

$$P(t_{\nu,-\sqrt{n}z_q}(\alpha_1) \leq T \leq t_{\nu,-\sqrt{n}z_q}(1-(\alpha-\alpha_1))) = 1-(\alpha-\alpha_1)-\alpha_1 = 1-\alpha,$$

tedy také platí

$$1-\alpha = P\left(t_{\nu,-\sqrt{n}z_q}(\alpha_1) \leq \frac{\sqrt{n}\,\frac{(\overline{Y}_n-\mu)}{\sigma} - \sqrt{n}z_q)}{S/\sigma} \leq t_{\nu,-\sqrt{n}z_q}(1-(\alpha-\alpha_1))\right) =$$

$$= P\left(\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(1-(\alpha-\alpha_1))\frac{S}{\sqrt{n}} \leq \mu + z_q\sigma \leq \overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(\alpha_1)\frac{S}{\sqrt{n}}\right).$$

Z tohoto získáme interval o spolehlivosti $1-\alpha$

$$\left[\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(1-(\alpha-\alpha_1))\frac{S}{\sqrt{n}}, \overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(\alpha_1)\frac{S}{\sqrt{n}}\right],$$

který má délku

$$\left[t_{\nu,-\sqrt{n}z_q}(1-(\alpha-\alpha_1)) - t_{\nu,-\sqrt{n}z_q}(\alpha_1)\right]\frac{S}{\sqrt{n}}.$$

Typicky volíme symetricky $\alpha_1 = \alpha/2$, protože nám záleží na obou stranách intervalu stejně, tedy chceme, aby pravděpodobnost, že kvantil přesáhne horní hranici intervalu, byla stejná, jako že bude menší než dolní hranice, tedy aby platilo

$$P\left(\mu + z_q\sigma > \overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(\frac{\alpha}{2})\frac{S}{\sqrt{n}}\right)$$
$$= P\left(\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(1-\frac{\alpha}{2})\frac{S}{\sqrt{n}} > \mu + z_q\sigma\right) = \frac{\alpha}{2}.$$

Pravostranný interval spolehlivosti dostaneme volbou $\alpha_1 = \alpha$

$$\left(-\infty, \overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(\alpha)\frac{S}{\sqrt{n}}\right]$$

a levostranný volbou $\alpha_1 = 0$

$$\left[\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(1-\alpha)\frac{S}{\sqrt{n}}, \infty\right).$$

Pokud bychom z nějakého důvodu chtěli, aby byl interval co nejkratší, musel by být rozdíl

$$t_{\nu,-\sqrt{n}z_q}(1-(\alpha-\alpha_1)) - t_{\nu,-\sqrt{n}z_q}(\alpha_1)$$

co nejmenší. A to je pro takové $\alpha_1^* \in (0,\alpha)$, které je nulovým bodem funkce

$$K(\alpha_1; \alpha,\nu, -\sqrt{n}z_q) := [t_{\nu,-\sqrt{n}z_q}(1-(\alpha-\alpha_1)) - t_{\nu,-\sqrt{n}z_q}(\alpha_1)]'. \tag{2.1}$$

Poznámka. S rostoucím počtem pozorování/stupňů volnosti se necentrální Studentovo rozdělení velmi přibližuje rozdělení $\mathrm{N}(-\sqrt{n}z_q,1)$, které je symetrické, a α_1^* se blíží $\alpha/2$.

Více kvantilů

Nyní budeme chtít sestrojit intervaly spolehlivosti pro k kvantilů $\mu + z_q\sigma$, kde $q \in Q = \{q_1,...,q_k\}$, $0 < q_1 < ... < q_k < 1$, $k \in \mathbb{N}$. Neboli intervaly

$$\left[\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(1-(A-\alpha_1))\frac{S}{\sqrt{n}}, \overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(\alpha_1)\frac{S}{\sqrt{n}}\right] \quad \forall q \in Q,$$

takové, aby každý interval měl svou marginální spolehlivost $1 - A$ a dohromady měly spolehlivost $1 - \alpha$. Jinými slovy, chceme, aby

$$\begin{aligned}
&P\left(\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(1-(A-\alpha_1))\frac{S}{\sqrt{n}} \le \mu + z_q\sigma\right.\\
&\qquad\left.\le \overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(\alpha_1)\frac{S}{\sqrt{n}} \quad \forall q \in Q\right)\\
&= P\left(-\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(1-(A-\alpha_1))\frac{S}{\sigma} \le \sqrt{n}\ \frac{\mu - \overline{Y}_n}{\sigma}\right.\\
&\qquad\left.\le -\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(\alpha_1)\frac{S}{\sigma} \quad \forall q \in Q\right) = 1 - \alpha.
\end{aligned} \tag{2.2}$$

Protože $\sqrt{n}\ \frac{\mu - \overline{Y}_n}{\sigma} \sim \mathrm{N}(0,1)$, je

$$\begin{aligned}
H(x; A) : &= P\left(-\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(1-(A-\alpha_1))x < \sqrt{n}\ \frac{\mu - \overline{Y}_n}{\sigma}\right.\\
&\qquad\left.< -\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(\alpha_1)x \quad \forall q \in Q\right) =\\
&= P\left(\max_{q\in Q}\left(-\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(1-(A-\alpha_1))x\right) < \sqrt{n}\ \frac{\mu - \overline{Y}_n}{\sigma}\right.\\
&\qquad\left.< \min_{q\in Q}\left(-\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(\alpha_1)x\right)\right) =
\end{aligned}$$

$$= \max\left(0,\ \Phi\left(\min_{q\in Q}\left(-\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(\alpha_1)x\right)\right)\right.$$
$$\left.-\Phi\left(\max_{q\in Q}\left(-\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(1-(A-\alpha_1))x\right)\right)\right),$$

kde $\Phi(\cdot)$ je distribuční funkce normovaného normálního rozdělení.

Zafixujeme si náhodnou veličinu S/σ. Víme, že $\frac{\nu S^2}{\sigma^2}$ má rozdělení χ^2_ν, tedy $\frac{\sqrt{\nu}S}{\sigma}$ má rozdělení χ_ν a hustota rozdělení χ_ν je

$$f_{\chi_\nu}(x) = \frac{2^{1-\frac{\nu}{2}}x^{\nu-1}e^{-\frac{x^2}{2}}}{\Gamma(\frac{\nu}{2})}.$$

Chtěli bychom zjistit rozdělení (hustotu) S/σ.

$$P\left(\frac{S}{\sigma} \le z\right) = P\left(\frac{\sqrt{\nu}}{\sqrt{\nu}}\cdot\frac{S}{\sigma} \le z\right) = P\left(\frac{\sqrt{\nu}S}{\sigma} \le z\sqrt{\nu}\right) = \int_0^{z\sqrt{\nu}} f_{\chi_\nu}(x)dx =$$
$$= \int_0^{z\sqrt{\nu}} \frac{2^{1-\frac{\nu}{2}}x^{\nu-1}e^{-\frac{x^2}{2}}}{\Gamma(\frac{\nu}{2})} \overset{\text{sub.}}{=} \int_0^z \frac{2(\frac{\nu}{2})^{\frac{\nu}{2}}y^{\nu-1}e^{-\frac{\nu y^2}{2}}}{\Gamma(\frac{\nu}{2})}dy = \int_0^z f_{S/\sigma}(y)dy,\ \forall z > 0,$$

kde jsme použili substituci $y = \frac{x}{\sqrt{\nu}}$, $dy = \frac{dx}{\sqrt{\nu}}$.

Hledaná hustota S/σ je tedy

$$f_{S/\sigma} = 2(\nu/2)^{\nu/2}x^{\nu-1}\exp(-\nu x^2/2)/\Gamma(\nu/2),\ x > 0.$$

Pak se (2.2) dá vyjádřit jako

$$P\left(-\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(1-(A-\alpha_1))\frac{S}{\sigma} < \sqrt{n}\,\frac{\mu-\overline{Y}_n}{\sigma}\right.$$
$$\left. < -\sqrt{n}z_q - t_{\nu,-\sqrt{n}z_q}(\alpha_1)\frac{S}{\sigma}\quad \forall q\in Q\right) =$$
$$= \mathsf{E}_{S/\sigma}[H(x;A)] = \int_0^\infty f_{S/\sigma}(x)H(x;A)dx = 1-\alpha. \tag{2.3}$$

Po zvolení $\alpha_1 \in [0,\alpha]$ již lze dopočítat funkce $H(x,A)$ pro každý kvantil, a tím zjistit i marginální spolehlivost A jednotlivých intervalů tak, že za A vezmeme nulový bod funkce $\int_0^\infty f_{S/\sigma}(x)H(x;A)dx - 1 + \alpha$. I zde zpravidla volíme $\alpha_1 = \alpha/2$ pro symetrické intervaly spolehlivosti, případně jinak, měli-li bychom zájem o jednostranné intervaly spolehlivosti nebo jinou modifikaci.

Modifikace

Pokud bychom chtěli mít intervaly co nejkratší a nezáleželo by nám na symetrii, tak by se α_1 ve výpočtech (2.2) a (2.3) nahradilo funkcí $\alpha_1^*(A,\nu,-\sqrt{n}z_q)$, která vrací nulový bod funkce $K(A;\alpha,\nu,-\sqrt{n}z_q)$ definované v (2.1). I v tomto případě bychom snadno pomocí vyhlédávajícího algoritmu vypočítali marginální spolehlivost A^* z (2.3). Protože $A^* < \alpha$ a z Booleovy nerovnosti ($P(\cup_{i=1}^m A_i \le \sum_{i=1}^m P(A_i)$) je $A^* > \alpha/k$, můžeme hledání omezit na interval $(\alpha/k,\alpha)$. Pak by příslušné intervaly spolehlivosti vypadaly takto:

$$\Bigg[\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(1 - (A^* - \alpha_1^*(A^*,\nu, -\sqrt{n}z_q)))\frac{S}{\sqrt{n}},$$
$$\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(\alpha_1^*(A^*,\nu, -\sqrt{n}z_q))\frac{S}{\sqrt{n}}\Bigg] \quad \forall q \in Q,$$

Pro dané Q si můžeme zvolit $\Lambda = \{\lambda_1, \ldots ,\lambda_k\}$, $\lambda_i > 0$, a uvažovat intervaly spolehlivosti

$$\left[\overline{Y_n} - t_{\nu,-\sqrt{n}z_{q_i}}(1 - (\lambda_i A - \alpha_1))\frac{S}{\sqrt{n}}, \overline{Y_n} - t_{\nu,-\sqrt{n}z_{q_i}}(\alpha_1)\frac{S}{\sqrt{n}}\right] \quad \forall i = 1, \ldots , k$$

s marginální spolehlivostí $1-\lambda_i A$. Situaci popsanou výše dostaneme volbou $\lambda_1 = \cdots = \lambda_k = 1$. Pokud bychom chtěli pro nějaký kvantil relativně užší interval spolehlivosti, zvolíme $\lambda_i > 1$, nebo pokud bychom naopak interval chtěli širší, zvolíme $\lambda_i < 1$.

Stejně můžeme postupovat, i kdybychom chtěli intervaly co nejkratší, tentokrát bychom za α_1 dosazovali $\alpha_1^*(\lambda_i A,\nu, -\sqrt{n}z_q)$, tedy nulový bod funkce $K(\lambda_i A; \alpha,\nu, -\sqrt{n}z_q)$, pro každé $i \in \{1,...,k\}$. Po specifikaci Λ opět můžeme vypočítat A^*. Dostaneme intervaly

$$\Bigg[\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(1 - [\lambda_i A^* - \alpha_1^*(\lambda_i A^*,\nu, -\sqrt{n}z_q)])\frac{S}{\sqrt{n}},$$
$$\overline{Y_n} - t_{\nu,-\sqrt{n}z_q}(\alpha_1^*(\lambda_i A^*,\nu, -\sqrt{n}z_q))\frac{S}{\sqrt{n}}\Bigg] \quad \forall q \in Q.$$

3. Simulace

V této části si na vygenerovaných datech ukážeme, jak vypadají intervaly spolehlivosti získané různými metodami pro různé kvantily a porovnáme je mezi sebou.

Označme si pro přehlednost:

BM klasická Bonferroniho metoda využívající binomické rozdělení, viz 1.3.1,

BMi iterovaná metoda BM, viz 1.3.4,

NM metoda specializovaná na normální rozdělení symetrická ($\alpha_1 = 1/2$), viz 2.1.3,

VKH asymptotická metoda založená na výběrovém kvantilu s histogramem jako odhadem hustoty, viz 1.2.2,

KS Kolmogorovova-Smirnovova metoda, viz 1.3.2.

Zvolme $\alpha = 0{,}05$.

3.1 Data z normálního rozdělení N(5,2)

Nagenerujeme si data z normálního rozdělení se střední hodnotou 5 a rozptylem 2 pro postupně se zvyšující počet pozorování n a aplikujeme metody popsané v předchozích dvou kapitolách na námi vygenerovaná data. Data v tabulkách uvádějí průměrné intervaly, průměrnou délku intervalu a pokrytí na tisíci náhodných výběrech pro všechny kvantily x_q, $q \in Q = \{q_1, \dots, q_k;\ 0 < q_1 < \dots < q_k < 1\}$.

$Q = \{0{,}5\}$, $x_Q = \{5\}$

Metoda	n	**Interval**		**Délka**	**Pokrytí**
BM	20	3,83287	6,18888	2,35601	0,957
	40	4,17001	5,82282	1,65281	0,956
	80	4,40630	5,60303	1,19674	0,966
NM	20	4,06876	5,92338	1,84062	0,942
	40	4,36488	5,64131	1,27062	0,952
	80	4,56303	5,45415	0,88650	0,946
VKH	20	3,13315	6,63737	3,50422	0,960
	40	3,78687	6,10021	2,31334	0,964
	80	4,27083	5,67284	1,40201	0,970
	160	4,53767	5,43494	0,89726	0,959
KS	20	3,14378	6,83350	3,66187	0,995
	80	4,13381	5,86423	1,71349	0,996

Tabulka 3.1: Intervaly spolehlivosti pro medián rozdělení N(5,2)

V tabulce 3.1 můžeme vidět, že zajímá-li nás medián, je nepatrně lepší metoda zaměřená na normální data, ale rozdíl mezi metodami není velký, jen s výjimkou

Kolmogorovovy-Smirnovovy metody, která má téměř stoprocentní pokrytí a dává širší intervaly.

$Q = \{0{,}1\}$, $x_Q = \{2{,}43690\}$

Metoda	n	Interval		Délka	Pokrytí
BM	20	$-\infty^*$	3,82519	—	0,989
	40	0,63748	3,39729	2,75981	0,968
	80	1,30422	3,19519	1,89096	0,981
	160	1,77349	2,95897	1,18548	0,977
NM	20	0,88434	3,42932	2,54498	0,944
	40	1,44635	3,18464	1,73829	0,948
	80	1,76453	2,97037	1,20584	0,958
VKH	20	0,53534	3,90847	3,37313	0,902
	40	0,79827	3,85640	3,05813	0,933
	80	1,37079	3,38924	2,01844	0,949
	160	1,79633	3,02593	1,22960	0,947
KS	20	$-\infty^*$	4,56434	—	1
	160	$-\infty^*$	3,44864	—	1

Pozn: * místo průměru byl kvůli nekonečným hodnotám brán medián.

Tabulka 3.2: Intervaly spolehlivosti pro 0,1-kvantil rozdělení N(5,2)

$Q = \{0{,}05\}$, $x_Q = \{1{,}71029\}$

Metoda	n	Interval		Délka	Pokrytí
BM	40	$-\infty^*$	2,86434	—	0,987
	80	0,11696	2,51451	2,39755	0,969
	160	0,72368	2,31973	1,59605	0,962
NM	40	0,57210	2,55550	1,98340	0,941
	80	0,95382	2,32456	1,37074	0,956
VKH	40	0,05343	2,94202	2,88859	0,874
	160	0,85192	2,45993	1,60802	0,946
KS	40	$-\infty^*$	3,88644	—	1
	160	$-\infty^*$	3,07295	—	1

Pozn: * místo průměru byl kvůli nekonečným hodnotám brán medián.

Tabulka 3.3: Intervaly spolehlivosti pro 0,05-kvantil N(5,2)

Pokud nás ovšem zajímají kvantily více z kraje distribuční funkce viz tabulky 3.2 a 3.3, specializovaná metoda už je výrazněji přesnější. Dále asymptotická metoda už potřebuje více pozorování, aby správně fungovala, než tomu bylo u mediánu. Kolmogorovova-Smirnovova metoda vrací velmi široké intervaly se stoprocentním pokrytím, jak se dalo očekávat, neboť tato metoda odhaduje celou distribuční funkci.

$Q = \{0{,}05\ ;0{,}1;\ 0{,}5;\ 0{,}9;\ 0{,}95\}$
$x_Q = \{1{,}71029;\ 2{,}43690;\ 5;\ 7.56310;\ 8.28971\}$

Metoda	n	q	**Interval**		**Délka**	**Pokrytí**
BM	160	0,05	$-\infty^*$	1,51858	—	0,972
		0,1	0,32926	2,47588	2,14662	
		0,5	4,47098	5,52112	1,05014	
		0,9	7,53738	9,65153	2,11415	
		0,95	8,49491	∞^*	—	
NM	40	0,05	0,32774	2,69998	2,37224	0,954
		0,1	1,24058	3,31979	2,07921	
		0,5	4,23663	5,75899	1,52235	
		0,9	6,67583	8,75504	2,07921	
		0,95	7,29564	9,66787	2,37224	
	80	0,05	0,79305	2,43748	1,64444	0,951
		0,1	1,64010	3,08419	1,44408	
		0,5	4,47183	5,53562	1,06379	
		0,9	6,92327	8,36735	1,44408	
		0,95	7,56997	9,21441	1,64444	
BMi	40	0,05	$-\infty^*$	3,02657	—	0,958
		0,1	$-\infty^*$	3,56619	—	
		0,5	4,03815	5,97127	1,92579	
		0,9	6,41818	∞^*	—	
		0,95	6,93396	∞^*	—	
	80	0,05	$-\infty^*$	2,65612	—	0,971
		0,1	0,94130	3,29367	2,33926	
		0,5	4,26132	5,71288	1,45356	
		0,9	6,71260	9,09064	2,34723	
		0,95	7,34998	∞^*	—	
KS	160	0,05	$-\infty^*$	3,06467	—	0,994
		0,1	$-\infty^*$	3,42572	—	
		0,5	4,39684	5,59827	1,18285	
		0,9	6,57176	∞^*	—	
		0,95	6,94686	∞^*	—	

Pozn: * místo průměru byl kvůli nekonečným hodnotám brán medián.

Tabulka 3.4: Intervaly spolehlivosti pro 0,05-, 0,1-, 0,5-, 0,9 a 0,95-kvantil N(5,2)

Rozšíříme-li oblast našeho zájmu na více kvantilů viz tabulky 3.4 a 3.5, je patrné, že metoda specializovaná na normální rozdělení je oproti ostatním metodám výrazně lepší. Dále je vidět, že i iterační Bonferroniho metoda vrací rozumně široké intervaly s dobrým pokrytím. Kolmogorovova-Smirnovova metoda má stále až moc vysoké pokrytí a velmi široké intervaly.

$Q = \{0{,}05;\ 0{,}1;\ 0{,}5\}$, $x_Q = \{1{,}71029;\ 2{,}43690;\ 5\}$

Metoda	n	q	**Interval**		**Délka**	**Pokrytí**
BM	40	0,05	$-\infty^*$	3,01707	—	0,978
		0,1	$-\infty^*$	3,56048	—	
		0,5	4,01321	5,97345	1,96024	
	80	0,05	$-\infty^*$	2,64185	—	0,977
		0,1	0,85288	3,27749	2,42460	
		0,5	4,25828	5,73493	1,47665	
NM	40	0,05	0,39295	2,63889	2,24594	0,951
		0,1	1,30190	3,27067	1,96877	
		0,5	4,29528	5,73712	1,44184	
	80	0,05	0,84256	2,38463	1,54206	0,942
		0,1	1,68163	3,03590	1,35427	
		0,5	4,49444	5,49219	0,99775	
BMi	80	0,05	$-\infty^*$	2,67111	—	0,955
		0,1	1,29411	3,20322	1,87235	
		0,5	4,34663	5,67341	1,32461	
KS	80	0,05	$-\infty^*$	3,46990	—	0,998
		0,1	$-\infty^*$	3,77392	—	
		0,5	4,13472	5,86969	1,73731	

Pozn: * místo průměru byl kvůli nekonečným hodnotám brán medián.

Tabulka 3.5: Intervaly spolehlivosti pro 0,05-, 0,1- a 0,5-kvantil N(5,2)

3.2 Data z exponenciálního rozdělení

Pro porovnání použijeme metody na data z exponenciálního rozdělení s parametrem $\lambda = 1$. Stejně jako normální rozdělení je i exponenciální rozdělení spojité, tudíž by i zde měly fungovat všechny metody až na metodu specializovanou na normální rozdělení.

$Q = \{0{,}5\}$, $x_Q = \{0{,}69315\}$

Metoda	n	**Interval**		**Délka**	**Pokrytí**
BM	20	0,34386	1,32515	0,98129	0,960
	40	0,42144	1,09117	0,66973	0,972
	80	0,48722	0,97292	0,48571	0,974
NM	40	0,68513	1,30950	0,62437	0,532
	80	0,77772	1,21658	0,43886	0,184
VKH	20	$-0{,}81948$	2,15523	2,97471	0,994
	320	0,46589	0,91277	0,44688	0,995
KS	20	0,19192	1,70050	1,48156	0,997
	80	0,39802	1,08883	0,69077	0,997

Tabulka 3.6: Intervaly spolehlivosti pro medián rozdělení Exp(1)

Jak vidíme v tabulce 3.6 pro medián, Bonferroniho metoda funguje na datech z exponciálního rozdělení stejně dobře jako pro normální rozdělení, metoda založená na výběrovém rozptylu a Kolmogorovova-Smirnovova metoda mají obě o hodně vyšší pokrytí než bylo požadováno a metoda specializovaná na normální rozdělení zde nefunguje a se zvyšujícím se počtem pozorování její pokrytí rychle klesá, pro osmdesát pozorování nemá pokrytí ani dvacet procent.

$Q = \{0{,}95\}$, $x_Q = \{2{,}99573\}$

Metoda	n	Interval		Délka	Pokrytí
BM	160	2,39369	4,14125	1,74756	0,981
NM	80	2,31416	2,98913	0,67497	0,440
VKH	160	1,56027	4,27892	2,71865	0,977
KS	160	1,80034	$-\infty^*$	—	1

Pozn: * místo průměru byl kvůli nekonečným hodnotám brán medián.

Tabulka 3.7: Intervaly spolehlivosti pro 0,95-kvantil rozdělení Exp(1)

Zaměříme-li se na kvantil z kraje distribuce viz tabulka 3.7, jsou všechny metody až na metodu určenou pro normální rozdělení, která nemá ani padesátiprocentní pokrytí, konzervativní, tj. dosahují vyšší spolehlivosti než $1 - \alpha = 0{,}95$. Nejlépe na tom je metoda založená na binomickém rozdělení, která má oproti metodě založené na výběrovém kvantilu o třetinu kratší intervaly s relativně stejnou spolehlivostí.

$Q = \{0{,}05;\ 0{,}5;\ 0{,}95\}$, $x_Q = \{0{,}05129;\ 0{,}69315;\ 2{,}99573\}$

Metoda	n	q	Interval		Délka	Pokrytí
BM	40	0,05	$-\infty^*$	0,19128	—	0,982
		0,5	0,38852	1,19594	0,80742	
		0,95	1,84999	∞^*	—	
	80	0,05	$-\infty^*$	0,12692	—	0,975
		0,5	0,44329	1,02827	0,57357	
		0,95	2,10643	∞^*		
NM	80	0,05	−1,29147	−0,13396	1,15751	0
		0,5	0,62642	1,36926	0,74285	
		0,95	2,12964	3,28715	1,15751	
BMi	80	0,05	$-\infty^*$	0,12692	—	0,968
		0,5	0,46459	0,99667	0,52398	
		0,95	2,10643	∞^*	—	
KS	80	0,05	$-\infty^*$	0,24825	—	1
		0,5	0,40483	1,09263	0,68011	
		0,95	1,50490	∞^*	—	

Pozn: * místo průměru byl kvůli nekonečným hodnotám brán medián.

Tabulka 3.8: Intervaly spolehlivosti pro 0,05-, 0,5- a 0,95-kvantil rozdělení Exp(1)

Jak si můžeme všimnout v tabulce 3.8 pro několik kvantilů, iterovaná Bonferroniho metoda oproti klasické přináší zúžení intervalu jen u mediánu, u krajnějších kvantilů se nic nemění. Kolmogorovova-Smirnovova metoda pokryla skutečnou hodnotu kvantilů na všech testovaných datech a intervaly touto metodou získané jsou široké. Metoda určená na data z normálního rozdělení vůbec nefunguje, nepokryla ani jeden případ.

Závěr

V této bakalářské práci jsme si ukázali různé metody konstrukce intervalů spolehlivosti pro jeden či více kvantilů zároveň. Jednalo-li se o jeden kvantil, byla vysvětlena klasická metoda využívající binomické rozdělení, která ale z vlastností diskrétního rozdělení nemusí být vždy přesná. Dále byla odvozena asymptotická metoda využívající výběrový kvantil a znalosti jeho asymptotického rozdělení. U této metody ale nastal problém s odhadem hustoty, která byla do simulací odhadnuta a nahrazena histogramem, který, jak simulace ukázaly, není vždy dostatečně přesným odhadem, zvlášť zajímají-li nás kvantily z krajů distribuce, a intervaly tím buď dosahovaly menší než stanovené spolehlivosti, nebo byly příliš široké.

Pro více kvantilů zároveň byla ukázána metoda založená na Kolmogorovově-Smirnovově statistice, která je ale zaměřená na pokrytí celé distribuční funkce, pro malý počet pozorování je přesná, pro velký asymptotická. Dále byla ukázána Bonferroniho metoda a pro malý počet pozorování její iterační verze založená na výpočtu celkové pravděpodobnosti s využitím multinomického rozdělení a rekurzivního algoritmu, který jsme si také odvodili.

Kromě neparametrických metod byla ukázána speciální metoda zaměřená na data z normálního rozdělení. Využitím nezávislosti výběrového průměru a směrodatné odchylky, která u normálního rozdělení platí, jsme si odvodili výpočetní techniku a její různé modifikace.

Popsané metody byly následně aplikovány na vygenerovaná data. Simulace ukázaly, že pro náhodná data z normálního rozdělení pro malý počet pozorování dává specializovaná metoda užší a přesnější intervaly než ostatní metody, pro velký počet pozorování rozdíl mizí a vyplatí se využít jiné metody, které jsou výpočetně méně náročné. Použije-li se metoda určená pro normální rozdělení na data například z exponenciálního rozdělení, vrací nesmyslné intervaly s nízkým pokrytím. Kolmogorovova-Smirnovova metoda aplikovaná jen na několik kvantilů poskytuje intervaly, které jsou široké a s mnohem vyšším pokrytím, než bylo požadováno. Dále se ukázalo, že námi odvozená iterační Bonferroniho metoda oproti klasické Bonferroniho metodě přináší určité vylepšení.

Seznam použité literatury

ANDĚL, J. (2005). *Základy matematické statistiky.* Matfyzpress, Praha.

HAYTER, A. J. (2014). Simultaneous confidence intervals for several quantiles of an unknown distribution. *The American Statistician*, **68**(1), 56–62.

HÁJEK, J. a ŠIDÁK, Z. (1967). *Theory of Rank Tests.* Academia, Praha.

LIU, W., BRETZ, F., HAYTER, A. J. a GLIMM, E. (2013). Simultaneous inference for several quantiles of a normal population with applications. *Biometrical Journal*, **55**(3), 360–369.

PARZEN, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**(3), 1065–1076.

Seznam tabulek