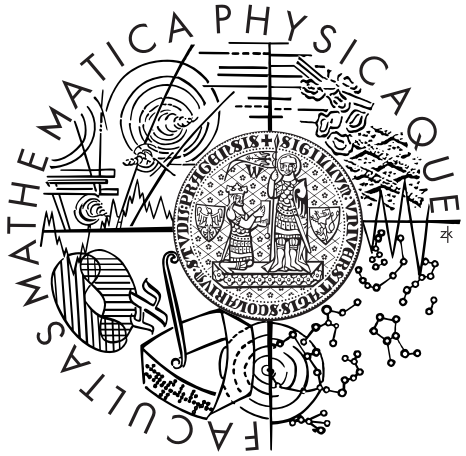
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Jakub Šimičák

Teoretické a empirické kvantily a ich využitie pri konštrukcii predikčných intervalov

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Matúš Maciak, Ph.D.
Studijní program:	Matematika
Studijní obor:	Finanční matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) 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 ..... dne ..... Podpis autora

Ďakujem vedúcemu bakalárskej práce doc. RNDr. Matúšovi Maciakovi, Ph.D. za trpezlivosť, za pomoc a rady pri spracovávaní tejto práce.

Názov práce: Teoretické a empirické kvantily a ich využitie pre konštrukciu predikčných intervalov

Autor: Jakub Šimičák

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: doc. RNDr. Matúš Maciak, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Úlohou bakalárskej práce je zoznámiť čitateľa s dvomi postupmi konštrukcie predikčných intervalov. Prvý postup predpokladá pravdepodobnostný model a vedie na frekventistický predikčný interval, ktorý využíva príslušné teoretické kvantily pravdepodobnostných rozdelení. Druhý postup nepredpokladá žiaden pravdepodobnostný model a vedie na konformný predikčný interval, ktorý využíva empirické kvantily príslušného náhodného výberu. V rámci práce budú oba prístupy všeobecne odvodené a následne ilustrované na konkrétnych príkladoch. Súčasťou práce je aj simulačná štúdia porovnávajúca empirické pokrytie frekventistických a konformných predikčných intervalov pre náhodné výbery z rôznych rozdelení.

Kľúčové slová: teoretický kvantil, empirický kvantil, predikčný interval, spoľahlivosť

Title: Theoretical and empirical quantiles and their use for prediction interval construction

Author: Jakub Šimičák

Department: Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis: doc. RNDr. Matúš Maciak, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The purpose of the bachelor thesis is to introduce the reader to two approaches to the construction of prediction intervals. The first procedure assumes a probabilistic model and leads to a frequentist prediction interval that uses the relevant theoretical quantiles of probability distributions. The second procedure assumes no probabilistic model and leads to a conformal prediction interval that uses empirical quantiles of the relevant random sample. In the course of the paper, both approaches will be derived in general terms and then illustrated with concrete examples. The thesis also includes a simulation study comparing the empirical coverage of frequentist and conformal prediction intervals for random selections from different distributions.

Keywords: theoretical quantile, empirical quantile, prediction interval, confidence

Obsah

Úvod

S narastajúcimi technologickými pokrokmi v oblasti ukladania a manipulácie dát sa stáva stále dôležitejším vedieť pracovať s obrovskými dátovými súbormi, ktoré sú charakteristické pre pojem *Big Data*. Dátová veda, ako vedecká disciplína, ktorá sa zaoberá analýzou, interpretáciou a predikciou dát, sa preto stáva nevyhnutnou súčasťou v mnohých oblastiach, vrátane obchodu, financii, marketingu, vedeckého výskumu a mnohých ďalších.

Jednou z najdôležitejších úloh v tejto oblasti je už zmienená predikcia, ktorá umožňuje odhadnúť budúce hodnoty (ako napríklad ceny aktív alebo vývoj úrokovej miery) na základe dostupných dát. Pre konštrukciu predpovedí využívame rôzne štatistické, ale aj neštatistické prístupy, ktoré sa môžu odlišovať predpokladmi, ktoré sú kladené na štruktúru a povahu dát. Často uvažujeme, že pozorované dáta sú realizácie náhodných veličín a snažíme sa nájsť pravdepodobnostný model, ktorý by im mohol zodpovedať. Vo všeobecnosti existujú dva hlavné typy predikcie, a to bodová a intervalová predikcia.

Bodová predikcia je jednoduchá a intuitívna metóda, ktorá nám poskytuje jednu konkrétnu hodnotu ako odhad budúcej hodnoty na základe dostupných dát. Tento odhad sa zvyčajne robí pomocou rôznych štatistických metód, akou je napríklad lineárna regresia. Avšak pri bodovej predikcii nemáme informáciu o tom, nakoľko je daný odhad spoľahlivý, a ako veľmi sa môže líšiť od skutočnej hodnoty.

Na druhej strane intervalová predikcia poskytuje celý interval hodnôt, ktorý bude pokrývať budúcu hodnotu na danej úrovni spoľahlivosti, čím sme schopní kvantifikovať úroveň neistoty v našej predikcii. Typicky úroveň spoľahlivosti zodpovedá hodnote $1 - \alpha$, kde α volíme z intervalu (0,1).

Cieľom bakalárskej práce je zoznámiť čitateľa s dvomi rozdielnymi spôsobmi konštrukcie predikčných intervalov. Postupne sa pozrieme na predpoklady, za ktorých ich môžeme skonštruovať a následne aj spôsoby konštrukcie. Na záver oba typy predikčných intervalov porovnáme v simulačnej štúdii.

1. Značenie, definície a vlastnosti

V tejto kapitole zavedieme jednotlivé značenia, zrekapitulujeme základné pojmy, definície a vlastnosti reálnych náhodných veličín a ich empirických náprotivkov, ktoré budeme následne používať pri konštrukcii predikčných intervalov. Taktiež sa budeme zaoberať definovaním predikčných intervalov a uvedieme motivačný problém, na ktorom budeme v priebehu práce ilustrovať konštrukcie predikčných intervalov.

1.1 Vlastnosti reálnej náhodnej veličiny

Na začiatok uvedieme niekoľko spôsobov, akými môžeme charakterizovať rozdelenie reálnej náhodnej veličiny. Prvým základným spôsobom jednoznačnej charakteristiky rozdelenia reálnej náhodnej veličiny je jej distribučná funkcia.

Definícia 1. *Nech X je reálna náhodná veličina, definovaná na pravdepodobnostnom priestore $(\Omega,\mathcal{A},P)$. Potom reálna funkcia $F_X(x) : \mathbb{R} \to [0,1]$, definovaná predpisom*

$$F_X(x) = P[X \leq x], \tag{1.1}$$

pre všetky $x \in \mathbb{R}$, sa nazýva distribučná funkcia reálnej náhodnej veličiny X.

Medzi ďalšie charakteristiky rozdelenia reálnej náhodnej veličiny X patrí kvantilová funkcia a príslušné kvantily.

Definícia 2. *Nech F_X je distribučná funkcia reálnej náhodnej veličiny X. Potom funkcia definovaná predpisom*

$$F_X^{-1}(u) = \inf\{x \in \mathbb{R} : F_X(x) \geq u\},$$

pre všetky $u \in (0,1)$, sa nazýva kvantilová funkcia reálnej náhodnej veličiny X.

Ak naviac uvažujeme distribučnú funkciu F_X, ktorá je spojitá a rastúca, potom kvantilová funkcia F_X^{-1} je inverzná k funkcii F_X.

Definícia 3. *Nech F_X je distribučná funkcia reálnej náhodnej veličiny X. Potom α-kvantil q_α rozdelenia F_X je ktorékoľvek reálne číslo spĺňujúce*

$$\lim_{h \searrow 0} F_X(q_\alpha - h) \leq \alpha \quad a \quad F_X(q_\alpha) \geq \alpha.$$

Špeciálne v druhej kapitole práce sa stretneme aj s prípadom, keď rozdelenie náhodnej veličiny X je známe až na neznámy parameter, ktorý predstavuje konštantu, respektíve vektor konštánt, všeobecne patriaci do priestoru $\Theta \subseteq \mathbb{R}^p$,

kde $p \in \mathbb{N}$. V takom prípade hovoríme, že rozdelenie náhodnej veličiny X závisí na neznámom parametri $\boldsymbol{\theta} \in \Theta$ a patrí do parametrickej rodiny rozdelení, čo budeme značiť

$$F_X(\cdot\,;\boldsymbol{\theta}) \in \mathcal{F} := \{F_X(\cdot\,;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p, p \in \mathbb{N}\},$$

kde $\mathcal{F}$ označuje parametrickú rodinu a $\Theta \subseteq \mathbb{R}^p$ nazývame parametrický priestor a predstavuje všetky možné hodnoty parametru $\boldsymbol{\theta} \in \Theta$.

1.2 Empirické odhady

V tejto časti si definujeme pojmy náhodný výber, usporiadaný ná-hodný výber a s ním súvisiace poriadkové štatistiky, ktoré budeme následne využívať pri odhadoch charakteristík reálnej náhodnej veličiny uvedených v časti 1.1.

Definícia 4. *Nech $n \in \mathbb{N}$. Postupnosť $X_1,\ldots,X_n$ nezávislých a rovnako rozdelených reálnych náhodných veličín, z ktorých má každá distribučnú funkciu F_X, nazývame reálny náhodný výber z rozdelenia F_X.*

Pre označenie náhodného výberu $(X_1,\ldots,X_n)^\top$ ako náhodného vektoru budeme používať značenie $\mathcal{X}_n$. Náhodný výber $\mathcal{X}_n$ môžeme naviac usporiadať, čím nám vznikne usporiadaný náhodný výber, ktorý si formálne zadefinujeme v nasledujúcej definícii.

Definícia 5. *Nech $n \geq 2$ a $\mathcal{X}_n = (X_1,X_2,\ldots,X_n)^\top$ je reálny náhodný výber zo spojitého rozdelenia F_X. Ak usporiadame náhodné veličiny $X_1,\ldots,X_n$ od najmenšej po najväčšiu, získame usporiadaný náhodný výber*

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)},$$

kde všetky nerovnosti platia skoro iste. Hodnota $X_{(k)}$ predstavuje k-tu najmenšiu hodnotu medzi pozorovaniami $X_1,\ldots,X_n$ a nazýva sa k-ta poriadková štatistika.

V prípade, ak by náhodný výber pochádzal z diskrétneho rozdelenia alebo by existovali rovnaké pozorovania vzniknuté vplyvom zaokrúhľovania, potom definujeme poriadkové štatistiky nasledovne

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

V usporiadaných reálnych náhodných výberoch vieme ďalej definovať aj poradie náhodnej veličiny.

Definícia 6. *Poradím náhodnej veličiny X_i v reálnom náhodnom výbere $\mathcal{X}_n$ rozumieme prirodzené číslo $R_i \in \{1,\ldots,n\}$, také že $X_i = X_{(R_i)}$.*

Dôležitou vlastnosťou poradia v náhodnom výbere je jeho diskrétne rovnomerné rozdelenie na množine $\{1,\ldots,n\}$, čo sformalizujeme v nasledujúcej vete.

Veta 7. *Nech $\mathcal{X}_n$ je reálny náhodný výber zo spojitého rozdelenia F_X. Nech R_i je poradie náhodnej veličiny X_i. Potom*

$$P[R_i = k] = \frac{1}{n}, \text{ pre } k \in \{1,\ldots,n\}.$$

Dôkaz. Kulich (2022), str. 34, Veta 2.16. ❑

Po zrekapitulovaní poriadkových štatistík ďalej pristúpime k odhadovaniu charakteristík reálnej náhodnej veličiny, ktoré boli definované v časti 1.1.

Definícia 8. *Nech $\mathcal{X}_n$ je reálny náhodný výber zo spojitého rozdelenia F_X, potom funkciu*

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(X_i),$$

definovanú pre všetky $x \in \mathbb{R}$, nazývame empirická distribučná funkcia.

Následne využijeme definíciu empirickej distribučnej funkcie, pomocou ktorej odhadneme kvantilovú funkciu ako

$$\widehat{F}_X^{-1}(u) = \inf\{x \in \mathbb{R} : \widehat{F}_n(x) \geq u\},$$

a ako empirický kvantil zvolíme odhad $\widehat{q}_\alpha := \widehat{F}_X^{-1}(\alpha)$ pre $\alpha \in (0,1)$. Z definície 8 je však jasné, že empirická distribučná funkcia $\widehat{F}_n$ je po častiach konštantná funkcia so skokmi v bodoch $X_{(1)},\ldots,X_{(n)}$, a teda empirický kvantil $\widehat{q}_\alpha$ bude vhodne vybraná poriadková štatistika z reálneho náhodného výberu $(X_1,\ldots,X_n)^\top$, kde skoro iste platí

$$\widehat{F}_n\left(X_{(k)}\right) \geq \frac{k}{n} \text{ a } \widehat{F}_n\left(X_{(k)} - h\right) < \frac{k}{n},$$

pre všetky $h > 0$ a $k \in \{1,\ldots,n\}$. Empirický kvantil bude teda spĺňovať $\widehat{q}_\alpha = X_{(k_\alpha)}$ za podmienky, že $k_\alpha = \alpha n$ je celé číslo, čo motivuje nasledujúcu definíciu.

Definícia 9. *Nech $n \in \mathbb{N}$, označme $k_\alpha = \alpha n$, ak αn je celé číslo a $k_\alpha = [\alpha n] + 1$ ak αn nie je celé číslo. Potom pre $\alpha \in (0,1)$ je empirický α-kvantil $\widehat{q}_\alpha$ definovaný ako k_α-tá poriadková štatistika reálneho náhodného výberu $(X_1,X_2,\ldots,X_n)^\top$.*

Za určitých predpokladov spojitosti rozdelenia F_X je navyše empirický kvantil $\widehat{q}_\alpha$ konzistentným odhadom teoretického kvantilu q_α v zmysle

$$\widehat{q}_\alpha \xrightarrow{\mathrm{P}} q_\alpha, \quad \text{pre } n \to \infty. \tag{1.2}$$

Celé znenie tvrdenia spolu aj s dôkazom je dostupné v skriptách Kulich (2022), str. 61, Veta 3.5.

1.3 Predikčné intervaly

Nech $X_1, \ldots, X_n, X_{n+1}$ sú nezávislé a rovnako rozdelené náhodné veličiny. Predpokladajme, že pozorovanie náhodného výberu $\mathcal{X}_n = (X_1, \ldots, X_n)^\top$ máme k dispozícii a označme náhodnú veličinu $Y = X_{n+1}$, ktorej budúcu realizáciu chceme odhadnúť. Pre odhad budúcej realizácie náhodnej veličiny Y využijeme konštrukciu intervalu $D \subseteq \mathbb{R}$ s nami určenou mierou neistoty.

Keďže náhodné veličiny $X_1, \ldots, X_n$ a Y sú rovnako rozdelené, využijeme pri konštrukcii intervalu D pre budúce pozorovanie náhodnej veličiny Y práve známy náhodný výber $\mathcal{X}_n$. Ďalším faktorom vplývajúcim na konštrukciu intervalu D je hodnota $\alpha \in (0,1)$, ktorá reprezentuje prijateľnú mieru neistoty, čo matematicky vyjadríme ako

$$P[Y \in D_n(\mathcal{X}_n,\alpha)] = 1 - \alpha,$$

kde interval $D \equiv D_n(\mathcal{X}_n,\alpha)$ nazveme predikčný interval, ktorý si sformalizujeme v nasledujúcej definícii.

Definícia 10. *Nech $\alpha \in (0,1)$ a $\mathcal{X}_n = (X_1, \ldots, X_n)^\top$ je náhodný výber z rozdelenia F_X. Nech náhodná veličina $Y = X_{n+1}$ má rozdelenie F_X a je nezávislá na náhodnom výbere $\mathcal{X}_n$. Potom náhodný interval $D \equiv D_n(\mathcal{X}_n, \alpha) \subseteq \mathbb{R}$ nazveme exaktný predikčný interval pre Y, ak platí*

$$P[Y \in D_n(\mathcal{X}_n, \alpha)] = 1 - \alpha.$$

Náhodný interval $D = D_n(\mathcal{X}_n, \alpha) \subseteq \mathbb{R}$ nazveme asymptotický predikčný interval pre Y, ak pre $n \to \infty$ platí

$$P[Y \in D_n(\mathcal{X}_n, \alpha)] \to 1 - \alpha.$$

Vo všeobecnosti rozoznávame tri typy predikčných intervalov:

- Obojstranný predikčný interval $D_n(\mathcal{X}_n,\alpha) = (L_1(\mathcal{X}_n,\frac{\alpha}{2}),L_2(\mathcal{X}_n,1-\frac{\alpha}{2}))$, kde $L_1(\mathcal{X}_n,\frac{\alpha}{2}) : \mathbb{R}^n \times (0,1) \to \mathbb{R}$ a $L_2(\mathcal{X}_n,1-\frac{\alpha}{2})) : \mathbb{R}^n \times (0,1) \to \mathbb{R}$ sú merateľné zobrazenia. Náhodné veličiny $L_1(\mathcal{X}_n,\frac{\alpha}{2})$ a $L_2(\mathcal{X}_n,1-\frac{\alpha}{2})$ naviac splňajú

$$P\left[L_1\left(\mathcal{X}_n,\frac{\alpha}{2}\right) < L_2\left(\mathcal{X}_n,1-\frac{\alpha}{2}\right)\right] = 1;$$

$$P\left[L_1\left(\mathcal{X}_n,\frac{\alpha}{2}\right) > -\infty\right] = 1;$$

$$P\left[L_2\left(\mathcal{X}_n,1-\frac{\alpha}{2}\right) < \infty\right] = 1,$$

  a predstavujú porade dolnú a hornú hranicu predikčného intervalu.

- Ľavostranný predikčný interval $D_n(\mathcal{X}_n,\alpha) = (-\infty,L_2(\mathcal{X}_n,\alpha))$, kde $L_2(\mathcal{X}_n,\alpha) : \mathbb{R}^n \times (0,1) \to \mathbb{R}$ je merateľné zobrazenie. Náhodná veličina $L_2(\mathcal{X}_n,\alpha)$ naviac splňuje $P[L_2(\mathcal{X}_n,\alpha) < \infty] = 1$ a predstavuje hornú hranicu predikčného intervalu.

- Pravostranný predikčný interval $D_n(\mathcal{X}_n,\alpha) = (L_1(\mathcal{X}_n,\alpha),\infty)$, kde $L_1(\mathcal{X}_n,\alpha) : \mathbb{R}^n \times (0,1) \to \mathbb{R}$ je merateľné zobrazenie. Náhodná veličina $L_1(\mathcal{X}_n,\alpha)$ naviac splňuje $P[L_1(\mathcal{X}_n,\alpha) > -\infty] = 1$ a predstavuje dolnú hranicu predikčného intervalu.

Ďalej si predstavíme motivačný problém, ktorý budeme rozoberať v priebehu práce.

Príklad 1 (Motivačný problém)**.** *Nech* $n \in \mathbb{N}$ *a* $\mathcal{X}_n = (X_1, \ldots, X_n)^\top$ *je známy reálny náhodný výber z rozdelenia* $F_X(\cdot, \boldsymbol{\theta})$ *patriaceho do parametrickej rodiny rozdelení* $\mathcal{F} = \{F_X(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p \ , p \in \mathbb{N}\}$. *Nech náhodná veličina* Y *má rozdelenie* $F_X(\cdot,\boldsymbol{\theta})$ *a je nezávislá na náhodnom výbere* $\mathcal{X}_n$.

Našou úlohou je skonštruovať predikčný interval $D_n(\mathcal{X}_n,\alpha) \subseteq \mathbb{R}$ *pre budúcu realizáciu náhodnej veličiny* Y *a predom dané* $\alpha \in (0,1)$.

Ako si môžeme všimnúť, motivačný problém je formulovaný všeobecne, keďže predpokladá nešpecifikovanú parametrickú rodinu rozdelení. Špeciálne v druhej kapitole sa budeme zaoberať motivačným problémom, kde budeme uvažovať konkrétnu, predom danú rodinu rozdelení, napríklad rodinu exponenciálnych rozdelení

$$\mathcal{F} = \{F_X(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in (0,\infty)\},$$

kde $F_X(\cdot,\boldsymbol{\theta})$ je distribučná funkcia exponenciálneho rozdelenia s neznámym parametrom $\boldsymbol{\theta} > 0$.

2. Frekventistické predikčné intervaly

Ako prvý a historicky starší spôsob konštrukcie predikčných intervalov si predstavíme frekventistickú metódu a s ňou spojené frekventistické predikčné intervaly. Ide o spôsob konštrukcie založený na predpoklade, že náhodný výber pochádza z rozdelenia patriaceho do parametrickej rodiny rozdelení. Na začiatok uvedieme spôsob konštrukcie presného predikčného intervalu a následne uvedieme aj jeho asymptotickú verziu.

2.1 Konštrukcia presných intervalov

Nech $\mathcal{X}_n$ je náhodný výber z rozdelenia F_X patriaceho do parametrickej rodiny rozdelení $\mathcal{F} = \{F(\cdot\,;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\ , p \in \mathbb{N}\}$ a náhodná veličina Y pochádza z rozdelenia F_X a je nezávislá na náhodnom výbere $\mathcal{X}_n$. Potom exaktný frekventistický predikčný interval môžeme skonštruovať pomocou nasledujúceho algoritmu.

Algoritmus 1. *Konštrukcia presných frekventistických predikčných intervalov*

1. *Uvažujme reálnu, prostú a merateľnú funkciu $f(\mathcal{X}_n,Y) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ a definujme reálnu náhodnú veličinu*

$$W \equiv f(\mathcal{X}_n,Y). \tag{2.1}$$

   Predpokladáme, že rozdelenie náhodnej veličiny W je známe a nezávisí na parametri $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ a ani na iných neznámych parametroch. V takom prípade sa náhodná veličina W nazýva presná pivotálna štatistika a jej distribučnú funkciu označíme ako $F_W(x) = P[W \leq x]$, pre $x \in \mathbb{R}$.

2. *O distribučnej funkcii F_W naviac predpokladáme, že je absolútne spojitá a rastúca. Z poznámky pod definíciou 2 preto vyplýva, že kvantilová funkcia F_W^{-1} je inverzná k funkcii F_W. Príslušný α-kvantil náhodnej veličiny W označíme ako $w_\alpha = F_W^{-1}(\alpha)$.*

3. *Ďalej uvažujeme merateľnú funkciu $g(\mathcal{X}_n,W) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, spĺňajúcu*

$$f(\mathcal{X}_n,g(\mathcal{X}_n,W)) = W \quad a \quad g(\mathcal{X}_n,f(\mathcal{X}_n,Y)) = Y. \tag{2.2}$$

   Následne exaktný obojstranný predikčný interval pre Y odvodíme z rovnosti

$$P\left[w_{\frac{\alpha}{2}} < W < w_{1-\frac{\alpha}{2}}\right] = 1 - \alpha,$$

ekvivalentnou úpravou, t.j. aplikáciou funkcie $g(\mathcal{X}_n, \cdot\,)$

$$P[g(\mathcal{X}_n,w_{\frac{\alpha}{2}}) < \; g(\mathcal{X}_n,W) < g(\mathcal{X}_n,w_{1-\frac{\alpha}{2}})] = 1-\alpha.$$

Následne využijeme vzťah (2.1) *a získame*

$$P[g(\mathcal{X}_n,w_{\frac{\alpha}{2}}) < g(\mathcal{X}_n,f(\mathcal{X}_n,Y)) < g(\mathcal{X}_n,w_{1-\frac{\alpha}{2}})] = 1-\alpha;$$

Požadovaný predikčný interval získame použitím (2.2) *ako*

$$P[g(\mathcal{X}_n, w_{\frac{\alpha}{2}}) < Y < g(\mathcal{X}_n,w_{1-\frac{\alpha}{2}})] = 1-\alpha. \tag{2.3}$$

Z rovnice (2.3) *už dostávame obojstranný exaktný predikčný interval v tvare*

$$D_n = \{y \in \mathbb{R} : g(\mathcal{X}_n, w_{\frac{\alpha}{2}}) < y < g(\mathcal{X}_n,w_{1-\frac{\alpha}{2}})\}$$

Z rovnice (2.3) ďalej vidíme, že pre hranice obojstranného predikčného intervalu platí $L_1(\mathcal{X}_n,\frac{\alpha}{2}) = g(\mathcal{X}_n, w_{\frac{\alpha}{2}})$ a $L_2(\mathcal{X}_n,1-\frac{\alpha}{2}) = g(\mathcal{X}_n, w_{1-\frac{\alpha}{2}})$. Pre ilustráciu postupu konštrukcie exaktného predikčného intervalu odvodíme predikčný interval pre motivačný problém 1, kde budeme predpokladať, že rozdelenie F_X patrí do parametrickej rodiny normálnych rozdelení s parametrami μ a σ^2.

Príklad 2. *Nech* $\mathcal{X}_n$ *je náhodný výber z rozdelenia* F_X*, patriaceho do parametrickej rodiny normálnych rozdelení*

$$\mathcal{F} = \{\Phi(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} = (\mu,\sigma^2)^\top, \boldsymbol{\theta} \in \mathbb{R} \times (0,\infty)\}, \tag{2.4}$$

kde $\Phi(\cdot;\boldsymbol{\theta})$ *je distribučná funkcia normálneho rozdelenia s neznámym parametrom* $\boldsymbol{\theta} = (\mu,\sigma^2)^\top$*, kde* μ *predstavuje strednú hodnotu a* $\sigma^2 > 0$ *je príslušný rozptyl. Nech náhodná veličina* Y *pochádza z rozdelenia* F_X *a je nezávislá na náhodnom výbere* $\mathcal{X}_n$*. Následne budeme postupovať v súlade s Algoritmom 1.*

1. *Definujeme*

$$W \equiv f(\mathcal{X}_n,Y) = Y - \overline{X}_n,$$

kde $\overline{X}_n$ *má rozdelenie* $N\left(\mu,\frac{\sigma^2}{n}\right)$ *a ide o výberový priemer spočítaný z náhodného výberu* $\mathcal{X}_n$*. Z nezávislosti výberového priemeru* $\overline{X}_n$ *a náhodnej veličiny* Y *a generickej vlastnosti normálneho rozdelenia dostávame rozdelenie náhodnej veličiny* W *a platí*

$$W \sim N\left(\mu + (-1)\mu, \sigma^2 + (-1)^2\frac{\sigma^2}{n}\right) \equiv N\left(0, \sigma^2\left(1+\frac{1}{n}\right)\right).$$

S využitím vlastností normálneho rozdelenia a definície t-rozdelenia ďalej dostávame

$$W \equiv \frac{Y - \overline{X}_n}{\sqrt{\left(1 + \frac{1}{n}\right) S_n^2}} \sim t_{n-1}, \tag{2.5}$$

kde S_n^2 je výberový rozptyl spočítaný z náhodného výberu $\mathcal{X}_n$. Vidíme, že rozdelenie náhodnej veličiny W nezávisí na parametri $\boldsymbol{\theta} \in \Theta$ a ani iných neznámych parametroch, čím sme našli exaktnú pivotálnu štatistiku.

2. *Pre dané $\alpha \in (0,1)$ označíme zodpovedajúce kvantily t-rozdelenia s $n-1$ stupňami voľnosti ako $t_{n-1}(\frac{\alpha}{2})$ a $t_{n-1}(1-\frac{\alpha}{2})$.*

3. *Následne využitím symetrie t-rozdelenia a ekvivalentným úpravami rovnosti*

$$P\left[t_{n-1}\left(\frac{\alpha}{2}\right) < W < t_{n-1}\left(1 - \frac{\alpha}{2}\right)\right] = 1 - \alpha,$$

do tvaru

$$P\left[L_1\left(\mathcal{X}_n, \frac{\alpha}{2}\right) \leq Y \leq L_2\left(\mathcal{X}_n, 1 - \frac{\alpha}{2}\right)\right] = 1 - \alpha, \tag{2.6}$$

kde $L_1\left(\mathcal{X}_n, \frac{\alpha}{2}\right)$ a $L_2\left(\mathcal{X}_n, 1 - \frac{\alpha}{2}\right)$ reprezentujú hornú a dolnú hranicu predikčného intervalu a platí

$$L_1\left(\mathcal{X}_n, \frac{\alpha}{2}\right) = \overline{X}_n - S_n\, t_{n-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\left(1 + \frac{1}{n}\right)},$$
$$L_2\left(\mathcal{X}_n, 1 - \frac{\alpha}{2}\right) = \overline{X}_n + S_n\, t_{n-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\left(1 + \frac{1}{n}\right)}.$$

Na základe výrazu (2.6) dostávame obojstranný predikčný interval v tvare

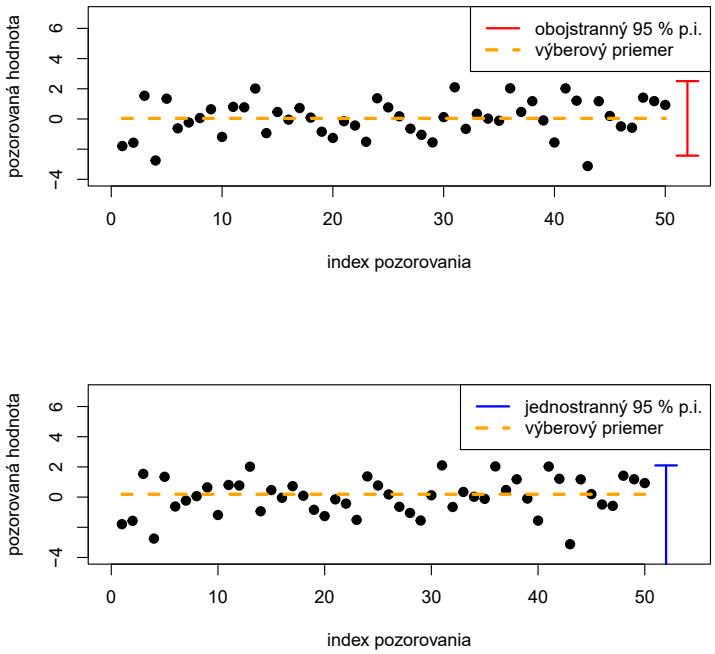
$$D_n = \left\{y \in \mathbb{R} : L_1\left(\mathcal{X}_n, \frac{\alpha}{2}\right) \leq y \leq L_2\left(\mathcal{X}_n, 1 - \frac{\alpha}{2}\right)\right\} \tag{2.7}$$

Podobne ekvivalentnými úpravami rovnice $P\left[W < t_{n-1}\left(1-\alpha\right)\right] = 1 - \alpha$ dostávame

$$P\left[Y \leq \overline{X}_n + S_n\, t_{n-1}(1-\alpha)\sqrt{\left(1 + \frac{1}{n}\right)}\right] = 1 - \alpha,$$

kde pre dané $\alpha \in (0,1)$, $t_{n-1}(1-\alpha)$ označuje $(1-\alpha)$ kvantil t-rozdelenia s $n-1$ stupňami voľnosti. Dostávame tak ľavostranný predikčný interval $D_n = \{y \in \mathbb{R} : y \leq L_2(\mathcal{X}_n, 1-\alpha)\}$, kde

$$L_2\left(\mathcal{X}_n, 1-\alpha\right) = \overline{X}_n + S_n\, t_{n-1}(1-\alpha)\sqrt{\left(1 + \frac{1}{n}\right)}.$$



Obr. 2.1: Pozorovaný náhodný výber $\mathcal{X}_{50}$ z normovaného normálneho rozdelenia a grafické znázornenie obojstranného a ľavostranného predikčného intervalu pre $Y = X_{51}$ a daným $\alpha = 0.05$.

2.2 Konštrukcia asymptotických intervalov

Ako si môžeme všimnúť, postup konštrukcie presných frekventistických predikčných intervalov závisí na nájdení presnej pivotálnej štatistiky W, ktorej rozdelenie nezávisí na parametri $\boldsymbol{\theta} \in \Theta$. V praxi sa však často stretneme aj s prípadom, keď predpoklad o existencii presnej pivotálnej štatistiky je porušený. Preto je potrebné nájsť spôsob, akým konštruovať frekventistický predikčný interval aj v prípade, keď presná pivotálna štatistika neexistuje.

V nasledujúcej časti sa preto budeme venovať spôsobu, akým môžeme aproximovať presnú pivotálnu štatistiku tak, aby rozdelenie aproximácie presnej pivotálnej štatistiky nezáviselo na parametri $\boldsymbol{\theta}$, a zároveň tak, aby sa nám podarilo dosiahnuť, že skonštruovaný predikčný interval bude dosahovať požadované pokrytie asymptoticky. Pri nasledujúcom odvodení konštrukcie predikčného intervalu sme čerpali z prác Beran (1990) a Lawless a Fredette (2005). Postup konštrukcie asymptotického predikčného intervalu zosumarizujeme prostredníctvom nasledujúceho algoritmu.

Algoritmus 2. *Konštrukcia asymptotických predikčných intervalov*

1. *Uvažujme prípad, že rozdelenie náhodnej veličiny W definovanej výrazom (2.1), závisí na parametri $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ a označme jej distribučnú funkciu $F_W(\cdot\,; \boldsymbol{\theta})$, o ktorej predpokladáme, že je absolútne spojitá a rastúca. Ďalej označme $\widehat{\boldsymbol{\theta}}_n$ ako konzistentný odhad parametru $\boldsymbol{\theta}$ v zmysle*

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}, \text{ pre } n \to \infty.$$

2. *Označíme náhodnú veličinu $\widehat{W}_n$ s distribučnou funkciou*

$$F_{\widehat{W}_n}(w;\widehat{\boldsymbol{\theta}}_n) := F_W(w;\widehat{\boldsymbol{\theta}}_n), \tag{2.8}$$

pre všetky $w \in \mathbb{R}$. Náhodná veličina $\widehat{W}_n$ má podmienene pri danom náhodnom výbere $\mathcal{X}_n$, rozdelenie, ktoré nezávisí na parametri $\boldsymbol{\theta} \in \Theta$ a platí

$$\widehat{W}_n|\mathcal{X}_n \sim F_{\widehat{W}_n}(\cdot\,;\widehat{\boldsymbol{\theta}}_n).$$

Podľa Beran (1990) jednostranný predikčný interval vieme vyjadriť v tvare

$$D_n = \left\{y \in \mathbb{R} : f(\mathcal{X}_n,y) \leq F^{-1}_{\widehat{W}_n}(1-\alpha\,;\widehat{\boldsymbol{\theta}}_n)\right\},$$

3. *Označíme príslušný α-kvantil $\widehat{w}_\alpha = F^{-1}_{\widehat{W}_n}(\alpha\,;\widehat{\boldsymbol{\theta}}_n)$, čím dostávame jednostranný predikčný interval v tvare*

$$D_n = \{y \in \mathbb{R} : f(\mathcal{X}_n,y) \leq \widehat{w}_{1-\alpha}\} \tag{2.9}$$

Z vlastnosti kvantilu rozšírime jednostranný predikčný interval (2.9) *na obojstranný predikčný interval, ktorý dostávame v tvare*

$$D_n = \left\{y \in \mathbb{R} : \widehat{w}_{\frac{\alpha}{2}} \leq f(\mathcal{X}_n,y) \leq \widehat{w}_{1-\frac{\alpha}{2}}\right\}. \tag{2.10}$$

Predikčné intervaly (2.9) *a* (2.10) *však nebudú dosahovať presné pokrytie, nakoľko sme pri konštrukcii využili konzistentný odhad $\widehat{\boldsymbol{\theta}}_n$ neznámeho parametru $\boldsymbol{\theta}$. Avšak za určitých podmienok regularity platí*

$$P[Y \in D_n|\mathcal{X}_n] \xrightarrow{P} 1-\alpha, \text{ pre } n \to \infty,$$

Dôkaz. Beran (1990), str. 718, Tvrdenie 1. ❑

4. *Analogicky k tretiemu kroku Algoritmu 1 využijeme merateľnú funkciu g, definovanú vzťahom* (2.2)*, čím dostávame obojstranný predikčný interval v tvare*

$$D_n = \left\{y \in \mathbb{R} : g\left(\mathcal{X}_n,\widehat{w}_{\frac{\alpha}{2}}\right) \leq y \leq g\left(\mathcal{X}_n,\widehat{w}_{1-\frac{\alpha}{2}}\right)\right\}. \tag{2.11}$$

Keďže predikčné intervaly (2.9) a (2.10) dosahujú pokrytie $1-\alpha$ asymptoticky, hovoríme, že ide o asymptotické predikčné intervaly a podmienenej náhodnej veličine $\widehat{W}_n|\mathcal{X}_n$ asymptoticky pivotálna štatistika.

Poznamenajme, že autori Lawless a Fredette (2005) a rovnako aj Beran (1990) pri konštrukcii predikčných intervalov ponúkajú ako vhodnú voľbu konzistentého odhadu práve maximálne vierohodný odhad neznámeho parametru $\boldsymbol{\theta} \in \Theta$.

Podobne, ako v prípade presného predikčného intervalu, aj teraz ukážeme názorný postup konštrukcie asymptotického predikčného intervalu pre motivačný problém 1, kde predpokladáme, že rozdelenie F_X patrí do parametrickej rodiny exponenciálnych rozdelení s parametrom $\boldsymbol{\theta} > 0$.

Príklad 3. *Nech $\boldsymbol{\theta} > 0$ a distribučná funkcia exponenciálneho rozdelenia spĺňa*

$$F(x;\boldsymbol{\theta}) = \begin{cases} 1 - e^{-\boldsymbol{\theta}x} & pre\ x \geq 0, \\ 0 & inak\ . \end{cases} \tag{2.12}$$

Nech $\mathcal{X}_n$ je náhodný výber z rozdelenia F_X patriaceho do parametrickej rodiny exponenciálnych rozdelení

$$\mathcal{F} = \{F(\cdot\,;\boldsymbol{\theta}), \boldsymbol{\theta} \in (0{,}\infty)\}, \tag{2.13}$$

kde $F(\cdot;\boldsymbol{\theta})$ je distribučná funkcia exponenciálneho rozdelenia definovaná v (2.12). Nech náhodná veličina Y pochádza z rozdelenia F_X a je nezávislá na náhodnom výbere $\mathcal{X}_n$. Ďalej budeme budeme postupovať v súlade s predstaveným Algoritmom 2.

1. *Definujme $W \equiv f(\mathcal{X}_n,Y) = Y$, čím dostávame*

   $$W \sim Exp(\boldsymbol{\theta}).$$

   Vidíme, že rozdelenie náhodnej veličiny W závisí na parametri $\boldsymbol{\theta}$ a označíme jej distribučnú funkciu $F_W(\,\cdot\,;\boldsymbol{\theta})$. Následne odhadneme parameter $\boldsymbol{\theta}$ maximálne vierohodným odhadom $\frac{1}{\overline{X}_n}$, kde $\overline{X}_n$ je výberový priemer spočítaný z náhodného výberu $\mathcal{X}_n$.

2. *Uvažujeme náhodnú veličinu $\widehat{W}_n$, ktorá má podmienene pri danom náhodnom výbere $\mathcal{X}_n$, rozdelenie*

   $$\widehat{W}_n|\mathcal{X}_n \sim Exp\left(\frac{1}{\overline{X}_n}\right),$$

   a označíme $F_{\widehat{W}_n}\left(\cdot\,;\frac{1}{\overline{X}_n}\right)$ ako distribučnú funkciu podmienenej náhodnej veličiny $\widehat{W}_n|\mathcal{X}_n$. Distribučná funkcia $F_{\widehat{W}_n}\left(\cdot\,;\frac{1}{\overline{X}_n}\right)$ nezávisí na neznámom parametri $\boldsymbol{\theta}$, keďže maximálne vierohodný odhad $\frac{1}{\overline{X}_n}$ podmienený daným náhodným výberom $\mathcal{X}_n$ nie je náhodný a predstavuje konkrétnu realizáciu.

3. *Pre dané $\alpha \in (0{,}1)$ označíme príslušné kvantily $\widehat{w}_{1-\frac{\alpha}{2}} = F^{-1}_{\widehat{W}_n}\left(1-\frac{\alpha}{2};\frac{1}{\overline{X}_n}\right)$ a analogicky $\widehat{w}_{\frac{\alpha}{2}}$. Asymptotický obojstranný predikčný interval následne dostávame z výrazu (2.10) v tvare*

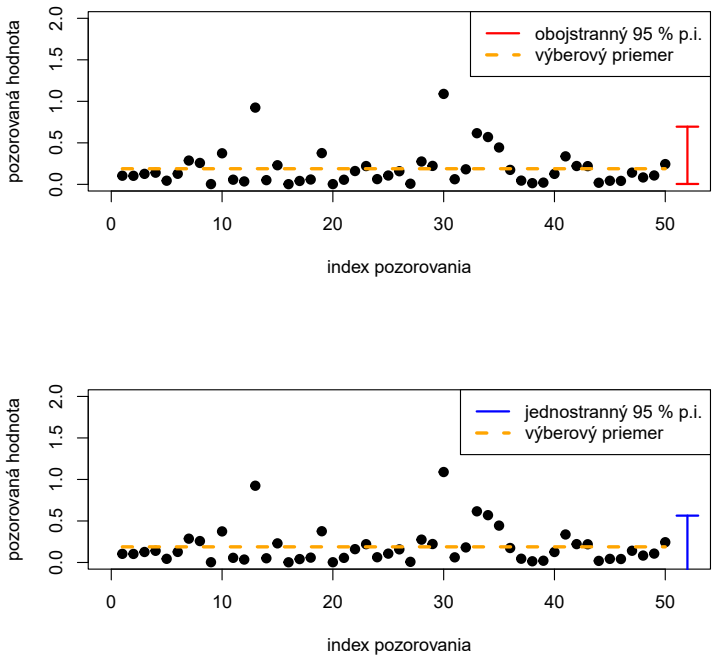
   $$D_n = \left\{y \in \mathbb{R} : \widehat{w}_{\frac{\alpha}{2}} \leq y \leq \widehat{w}_{1-\frac{\alpha}{2}}\right\}, \tag{2.14}$$

   keďže sme zvolili $f(\mathcal{X}_n,y) = y$.

Analogicky, ľavostranný predikčný interval dostávame v tvare

$$D_n = \{y \in \mathbb{R} : y \leq \widehat{w}_{1-\alpha}\},$$

kde $\widehat{w}_{1-\alpha} = F^{-1}_{\widehat{W}_n}\left(1-\alpha;\frac{1}{\overline{X}_n}\right)$, pre predom dané $\alpha \in (0{,}1)$.



Obr. 2.2: Pozorovaný náhodný výber $\mathcal{X}_{50}$ z exponenciálneho rozdelenia s parametrom $\lambda = 4$ a grafické znázornenie obojstranného a ľavostranného asymptotického predikčného intervalu pre $Y = X_{51}$ s daným asymptotickým pokrytím pre $\alpha = 0.05$.

Poznamenajme že, uvedená konštrukcia asymptotického predikčného intervalu pomocou vyššie popísanej tzv *plug-in* predstavuje základný prístup pre konštrukciu asymptotického predikčného intervalu. Pre dosiahnutie lepších vlastností asymptotických predikčných intervalov sa často používajú rôzne typy kalibrácii. Napríklad Beran (1990) využíva na kalibráciu Monte-Carlo simulácie. Inou možnosťou je využite tzv. *všeobecnej metódy*[1], ktorá je uvedená v článku Lawless a Fredette (2005).

[1]angl: A General Method

Z vyššie uvedených príkladov konštrukcii frekventistických predikčných intervalov môžeme pozorovať, že:

- Pre konštrukciu predikčného intervalu je potrebné poznať rozdelenie F_X z parametrickej rodiny rozdelení $\mathcal{F}$, z ktorej pochádza náhodný výber $\mathcal{X}_n$, na základe ktorého konštruujeme náhodnú veličinu W.

- Ak rozdelenie náhodnej veličiny W nezávisí na parametri $\boldsymbol{\theta}$, potom sme našli presnú pivotálnu štatistiku, na základe ktorej konštruujeme presný predikčný interval.

- Ak rozdelenie náhodnej veličiny W závisí na parametri $\boldsymbol{\theta}$, potom využijeme podmienenú náhodnú veličinu $\widehat{W}_n|\mathcal{X}_n$.

- Na hraniciach frekventistických predikčných intervaloch vystupujú teoretické kvantily rozdelenia F_W v prípade pivotálnej štatistiky W alebo teoretické kvantily rozdelenia $F_{\widehat{W}_n}(\cdot\,;\widehat{\boldsymbol{\theta}}_n)$.

Problém s konštrukciou frekventistických predikčných intervalov však nastáva v momente, keď nepoznáme rozdelenie F_X z parametrickej rodiny $\mathcal{F}$, z ktorej pochádza rozdelenie náhodného výberu $\mathcal{X}_n$. V takom prípade nevieme nájsť presnú pivotálnu štatistiku a nevieme ani zostrojiť *plug-in* odhad potrebný ku konštrukcii asymptotického predikčného intervalu. Ukazuje sa tak potreba pre iný postup konštrukcie predikčných intervalov, ktorý by sa dokázal vysporiadať aj s tou alternatívou, že rozdelenie F_X nie je známe.

3. Konformné predikčné intervaly

Podstatne mladšou metódou konštrukcie predikčných intervalov je konformná metóda, ktorej začiatky siahajú do roku 2002. Na rozdiel od frekventistickej metódy, v konformnej metóde neuvažujeme, že rozdelenie náhodného výberu patrí do parametrickej rodiny rozdelení. Namiesto toho chceme predikčné intervaly odhadnúť priamo na základe pozorovania náhodného výberu bez nutnosti predpokladu predom známeho rozdelenia.

3.1 Konštrukcia konformných intervalov

Nech náhodný výber $\mathcal{X}_n$ a náhodná veličina Y pochádzajú z rovnakého, ľubovoľného rozdelenia a Y je nezávislá na $\mathcal{X}_n$. Uvažujme úlohu konštrukcie ľavostranného predikčného intervalu $D_n(\mathcal{X}_n,\alpha)$ spĺňajúceho

$$P[Y \in D_n(\mathcal{X}_n,\alpha))] = P[Y \leq L_2(\mathcal{X}_n,\alpha)] = 1-\alpha,$$

pre $\alpha \in (0,1)$. Intuitívny spôsob, akým by sme mohli skonštruovať predikčný interval, je nájsť empirický $(1-\alpha)$ kvantil $\widehat{q}_{1-\alpha}$ z reálneho náhodného výberu $\mathcal{X}_n$ a položiť $L_2(\mathcal{X}_n,\alpha) = \widehat{q}_{1-\alpha}$. Z poznámky 1.2 vieme, že platí $\widehat{q}_\alpha \xrightarrow{\mathrm{P}} q_\alpha$, pre $n \to \infty$, a teda pre výsledný interval $D_n(\mathcal{X}_n,\alpha) = (-\infty,\widehat{q}_{1-\alpha})$ ďalej dostávame

$$P[Y \in D_n(\mathcal{X}_n,\alpha)] \to 1-\alpha, \text{ pre } n \to \infty.$$

Podľa Tibshirani (2019) pre dosiahnutie presného pokrytia predikčného intervalu D je možné využiť nezávislosť a rovnaké rozdelenie náhodných veličín $X_1,\ldots,X_n$ patriacich do náhodného výberu $\mathcal{X}_n$ a náhodnej veličiny Y. Je však dôležité poznamenať, že aj keď realizáciu náhodnej veličiny Y na rozdiel od náhodného výberu $\mathcal{X}_n$ nepoznáme, môžeme hypoteticky uvažovať všetky možné realizácie $y \in \mathbb{R}$ náhodnej veličiny Y.

V takom prípade poradie náhodnej veličiny Y v reálnom náhodnom výbere $(X_1,\ldots,X_n,Y)^\top$ má na základe vety 7 rovnomerné diskrétne rozdelenie na množine $\{1,\ldots,n+1\}$. Ďalej definujeme zobrazenie $\phi(\mathcal{X}_n,Y) : \mathbb{R}^n \times \mathbb{R} \to \{1,\ldots,n+1\}$, ktoré priradí náhodnej veličine Y poradie v náhodnom výbere $(X_1,\ldots,X_n,Y)^\top$ ako

$$\phi(\mathcal{X}_n,Y) = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq Y\}} + 1. \tag{3.1}$$

Z definície 9 vieme, že empirický kvantil $\widehat{q}_\alpha$ je definovaný ako k_α-tá poriadková štatistika, z čoho vyplýva že realizácia náhodnej veličiny $\phi(\mathcal{X}_n,Y)$ predstavuje prirodzené číslo k_α. Ďalej chceme zistiť príslušné $\alpha \in (0,1)$, ktoré dostaneme

podielom $\frac{k_\alpha}{n+1}$, keďže náhodný výber $(X_1,\ldots,X_n,Y)^\top$ má rozsah $n+1$. Definujeme zobrazenie $\pi(\mathcal{X}_n,Y) : \mathbb{R}^n \times \mathbb{R} \to \{\frac{1}{n+1},\ldots,1\}$ nasledovne

$$\pi(\mathcal{X}_n,Y) = \frac{1}{n+1}\left(\sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq Y\}} + 1\right). \tag{3.2}$$

Realizácia náhodnej veličiny $\pi(\mathcal{X}_n,Y)$ udáva index $\alpha \in (0,1)$ príslušný náhodnej veličine Y ako k_α-tej poriadkovej štatistike v náhodnom výbere $(X_1,\ldots,X_n,Y)^\top$. Naviac sme len vhodne prenásobili náhodnú veličinu $\phi(\mathcal{X}_n,Y)$ kladnou konštantou $\frac{1}{n+1}$, z čoho vyplýva že náhodná veličina $\pi(\mathcal{X}_n,Y)$ má diskrétne rovnomerné rozdelenie na množine $\{\frac{1}{n+1},\ldots,1\}$ a platí

$$P\left[\pi(\mathcal{X}_n,Y) \leq \frac{\lceil (n+1)(1-\alpha)\rceil}{n+1}\right] \geq 1-\alpha. \tag{3.3}$$

Keďže realizáciu náhodnej veličiny Y nepoznáme, ale uvažujeme jej všetky možné realizácie $y \in \mathbb{R}$, využijeme nerovnosť (3.3). Ľavostranný predikčný interval budeme konštruovať ako všetky hypotetické realizácie $y \in \mathbb{R}$ náhodnej veličiny Y, pre ktoré nerovnosť (3.3) platí, čím dostávame predikčný interval v tvare

$$D_n(\mathcal{X}_n,\alpha) = \left\{y \in \mathbb{R} : \pi(\mathcal{X}_n,y) \leq \frac{\lceil (n+1)(1-\alpha)\rceil}{n+1}\right\}. \tag{3.4}$$

Pre predikčný interval (3.4) bude naviac na základe nerovnosti (3.3) platiť $P[Y \in D_n(\mathcal{X}_n,\alpha)] \geq 1-\alpha$. Stanovenie predikčného intervalu však z výpočtového hľadiska predstavuje zložitú úlohu, nakoľko je potrebné určiť pre všetky $y \in \mathbb{R}$ poradia v náhodnom výbere $\mathcal{X}_n$. Ukazuje sa tak potreba pre jednoduchšiu konštrukciu predikčného intervalu, ktorá bude založená na nasledujúcich lemách.

Lema 11. *Nech $(X_1,\ldots,X_n,Y)^\top$ je náhodný výber a označme $\mathcal{X}_n = (X_1,\ldots,X_n)^\top$. Nech $\alpha \in (0,1)$ a náhodná veličina $\pi(\mathcal{X}_n,Y)$ je definovaná vzťahom (3.2). Potom*

$$P[Y \leq \widehat{q}^{\,k}_{1-\alpha}] = P\left[\pi(\mathcal{X}_n,Y) \leq \frac{\lceil (n+1)(1-\alpha)\rceil}{n+1}\right],$$

kde $\widehat{q}^{\,k}_{1-\alpha}$ je empirický kvantil náhodného výberu $X_1,\ldots,X_n,Y$.

Dôkaz. Bez újmy na všeobecnosti usporiadame náhodný výber $(X_1,\ldots,X_n,Y)^\top$, čím vznikne usporiadaný náhodný výber

$$X_{(1)}, X_{(2)}, \ldots, Y_{(k_y)}, \ldots, X_{(k_{1-\alpha})}, \ldots, X_{(n+1)}.$$

Podľa definície 9 označíme výberový kvantil $\widehat{q}^{\,\mathrm{k}}_{1-\alpha} := X_{(k_{1-\alpha})}$ a všimneme si ekvivalenciu medzi javmi

$$Y \leq \widehat{q}^{\,\mathrm{k}}_{1-\alpha} \iff k_y \leq k_{1-\alpha}.$$

Z definície 9 vieme, že $k_{1-\alpha} = (1-\alpha)(n+1)$, ak $k_{1-\alpha}$ je prirodzené číslo. Inak $k_{1-\alpha} = \lfloor (1-\alpha)(n+1)\rfloor + 1$. Uvažujme druhú možnosť a ďalej dostávame

$$k_y \leq k_{1-\alpha} \iff k_y \leq \lfloor (1-\alpha)(n+1)\rfloor + 1 \iff k_y \leq \lceil (1-\alpha)(n+1)\rceil,$$

kde druhá ekvivalencia plynie z vlastností dolnej a hornej časti kladných čísel. Vieme, že k_y je poradie náhodnej veličiny Y v náhodnom výbere $(X_1,\ldots,X_n,Y)^\top$, ktoré vieme ekvivalentne zapísať pomocou náhodnej veličiny $\phi(\mathcal{X}_n,Y)$, definovanej výrazom (3.1), čím dostávame

$$k_y \leq \lceil (1-\alpha)(n+1) \rceil \iff \phi(\mathcal{X}_n,Y) \leq \lceil (1-\alpha)(n+1) \rceil.$$

Následným vynásobením oboch strán kladnou konštantou $\frac{1}{n+1}$ dostávame

$$\phi(\mathcal{X}_n,Y) \leq \lceil (1-\alpha)(n+1) \rceil \iff \pi(\mathcal{X}_n,Y) \leq \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1},$$

čím je požadovaná rovnosť dokázaná pre $k_{1-\alpha} = \lfloor (1-\alpha)(n+1) \rfloor + 1$. V prípade, že $k_{1-\alpha} = (1-\alpha)(n+1)$ je prirodzené číslo, potom $k_{1-\alpha} = (1-\alpha)(n+1) = \lceil (1-\alpha)(n+1) \rceil$, čím je dôkaz dokončený.

❑

Z lemy 11 však nie je jasné, ktorý empirický kvantil $\widehat{q}^{\,\mathrm{k}}_{1-\alpha}$ by sme na konštrukciu predikčného intervalu mali využiť, keďže uvažujeme všetky možné realizácie $y \in \mathbb{R}$ náhodnej veličiny Y. V nasledujúcej leme preto ukážeme, akým spôsobom môžeme zvoliť jednu hodnotu y tak, aby empirický kvantil $\widehat{q}^{\,\mathrm{k}}_{1-\alpha}$ splňoval pokrytie $1-\alpha$.

Lema 12 (Tibshirani a kol. (2019a))**.** *Nech $(X_1,\ldots,X_n,Y)^\top$ je náhodný výber a $\alpha \in (0,1)$. Označme $\widehat{q}^{\,k+}_{\alpha}$ empirický α-kvantil z náhodného výberu $(X_1,\ldots,X_n,y)^\top$, pričom položíme $y = \infty$. Potom*

$$P[Y \leq \widehat{q}^{\,k+}_{\alpha}] \geq \alpha \tag{3.5}$$

Ak naviac platí $P[X_i = X_j] = 0$ pre $i \neq j$, $i,j \in \{1,\ldots,n\}$ a zároveň $P[X_i = Y] = 0$ pre všetky $i \in \{1,\ldots,n\}$, potom platí

$$P[Y \leq \widehat{q}^{\,k+}_{\alpha}] \leq \alpha + \frac{1}{n+1}$$

Dôkaz. Tibshirani a kol. (2019b), str. 1 , časť A.1.

❑

Na základe lem 11 a 12 konštruujeme ľavostranný predikčný interval

$$D_n(\mathcal{X}_n,\alpha) = \left\{ y \in \mathbb{R} : y \leq \widehat{q}^{\,\mathrm{k+}}_{1-\alpha} \right\}, \tag{3.6}$$

kde $\widehat{q}^{\,\mathrm{k+}}_{1-\alpha}$ je empirický $(1-\alpha)$ kvantil z náhodného výberu $(X_1,\ldots,X_n,y)^\top$, v ktorom položíme $y = \infty$. Na základe lemy 12 pre predikčný interval (3.6) platí

$$P[Y \in D_n(\mathcal{X}_n,\alpha)] \geq 1-\alpha.$$

Všimnime si, že ak by sme chceli konštruovať pravostranný predikčný interval, tak by sme požadovanú hladinu $1-\alpha$ nedosiahli, keďže na základe lemy 12 dostávame $P[Y > \widehat{q}^{\,\mathrm{k+}}_{\alpha}] < 1-\alpha$. Predchádzajúca úvaha o pravostranných predikčných intervaloch motivuje nadchádzajúcu lemu.

Lema 13. *Nech* $(X_1,\ldots,X_n,Y)^\top$ *je náhodný výber, ktorý splňuje* $P[X_i = X_j] = 0$, *pre* $i \neq j$, $i,j \in \{1,\ldots,n\}$ *a zároveň* $P[X_i = Y] = 0$ *pre všetky* $i \in \{1,\ldots,n\}$. *Označme symbolom* $\widehat{q}_\alpha^{k-}$ *empirický* α*-kvantil z náhodného výberu* $(X_1,\ldots,X_n,y)^\top$, *pričom položíme* $y = -\infty$. *Potom platí*

$$P[Y < \widehat{q}_\alpha^{k-}] \leq \alpha.$$

Dôkaz. Označme symbolom $\widehat{q}_\alpha^{\mathrm{k}}$ empirický kvantil z reálneho náhodného výberu $(X_1,\ldots,X_n,Y)^\top$. Všimneme si ekvivalenciu medzi nasledujúcimi javmi

$$Y < \widehat{q}_\alpha^{\mathrm{k}-} \iff Y < \widehat{q}_\alpha^{\mathrm{k}}.$$

Následne dostávame

$$\begin{aligned} P[Y \leq \widehat{q}_\alpha^{\mathrm{k}}] &= P[Y < \widehat{q}_\alpha^{\mathrm{k}}] + P[Y = \widehat{q}_\alpha^{\mathrm{k}}] \\ P[Y < \widehat{q}_\alpha^{\mathrm{k}}] &= P[Y \leq \widehat{q}_\alpha^{\mathrm{k}}] - \frac{1}{n+1} \end{aligned}$$

Z dôkazu lemy 12 vieme, že za predpokladu $P[X_i = X_j] = 0$ pre $i \neq j$, $i,j \in \{1,\ldots,n\}$ a zároveň $P[X_i = Y] = 0$ pre všetky $i \in \{1,\ldots,n\}$ platí

$$P[Y \leq \widehat{q}_\alpha^{\mathrm{k}}] \leq \alpha + \frac{1}{n+1},$$

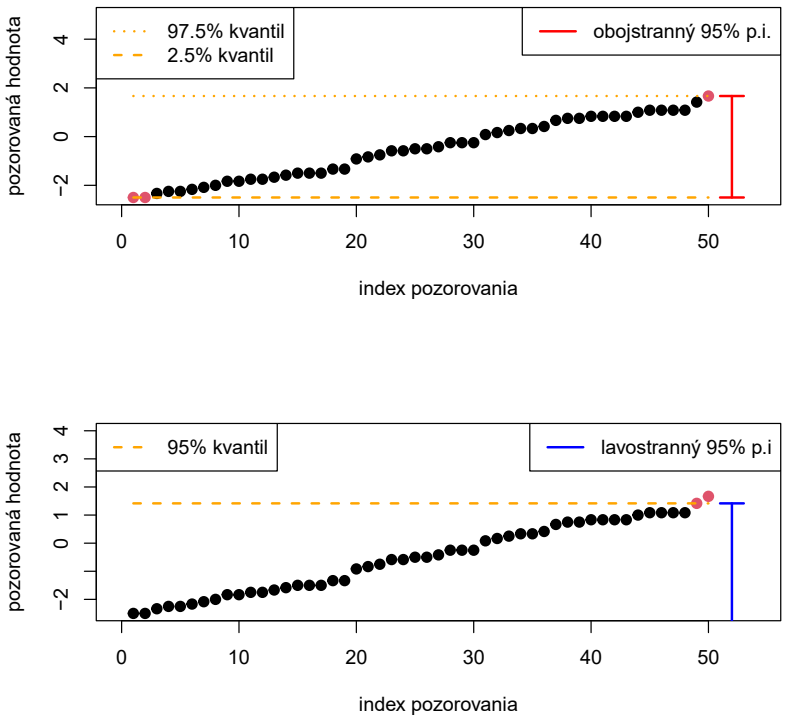
z čoho okamžite plynie požadovaná nerovnosť

$$P[Y < \widehat{q}_\alpha^{\mathrm{k}}] = P[Y \leq \widehat{q}_\alpha^{\mathrm{k}}] - \frac{1}{n+1} \leq \alpha.$$

❑

Z lemy 13 ďalej plynie $P[Y \geq \widehat{q}_\alpha^{\mathrm{k}-}] > 1 - \alpha$, čím dostávame nástroj na konštrukciu pravostranného predikčného intervalu. Obojstranný predikčný interval konštruujeme ako prienik ľavostraného a pravostranného predikčného intervalu. Takýto spôsob konštrukcie konformného obojstranného intervalu môžeme napríklad nájsť v knihe Vovk a kol. (2005) na stranách 40 a 41, kde je predvedený na príklade konformnej *ridge regresie*. Konformný obojstranný predikčný interval pre $\alpha \in (0,1)$ tak dostávame v tvare

$$\begin{aligned} D_n(\mathcal{X}_n,\alpha) &= \left\{y \in \mathbb{R} : y \geq \widehat{q}_{\frac{\alpha}{2}}^{\mathrm{k}-}\right\} \cap \left\{y \in \mathbb{R} : y \leq \widehat{q}_{1-\frac{\alpha}{2}}^{\mathrm{k}+}\right\}; \\ &= \left\{y \in \mathbb{R} : \widehat{q}_{\frac{\alpha}{2}}^{\mathrm{k}-} \leq y \leq \widehat{q}_{1-\frac{\alpha}{2}}^{\mathrm{k}+}\right\}. \end{aligned} \tag{3.7}$$



Obr. 3.1: Pozorovaný usporiadaný náhodný výber $\mathcal{X}_{50}$ a grafické znázornenie obojstranného a ľavostranného konformného predikčného intervalu pre $Y = X_{51}$ na úrovni $\alpha = 0.05$. Červená farba bodov indikuje pozorovania náhodného výberu $\mathcal{X}_{50}$, ležiace mimo skonštruovaný predikčný interval.

Z predvedeného odvodenia konformného predikčného intervalu vidíme, že

- Pre konštrukciu predikčných konformných intervalov nepredpokladáme znalosť rozdelenia náhodného výberu $\mathcal{X}_n$.
- Predpokladáme nezávislosť a rovnaké rozdelenie náhodných veličín $X_1, \ldots, X_n, Y$.
- Konformné predikčné intervaly majú vždy presné alebo väčšie pokrytie nezávisle na rozdelení.
- Na hraniciach konformných predikčných intervalov figurujú empirické kvantily z reálneho náhodného výberu, ktorý je rozšírený o ďalšie pozorovanie, ktoré pokladáme buď nekonečnu alebo mínus nekonečnu.

3.2 Zameniteľnosť

Pri konštrukcii konformných predikčných intervalov sa však v praxi stretávame s tým, že predpoklad nezávislosti náhodných veličín $X_1, \ldots, X_n$ je reštriktívny. Predpoklad nezávislosti je však možné zredukovať na predpoklad zvaný zameniteľnosť[1].

Definícia 14 (Shafer a Vovk (2008))**.** *Náhodné veličiny $X_1, \ldots, X_n$ nazveme zameniteľné, ak pre každú permutáciu $\tau : \{1, \ldots, n\} \to \{1, \ldots, n\}$ prirodzených čísel $1, \ldots, n$, má náhodný vektor $\mathbf{Z}_n = (Z_1, \ldots, Z_n)^\top$, kde $Z_i = X_{\tau(i)}$, rovnakú združenú distribučnú funkciu ako náhodný vektor $\mathbf{X}_n = (X_1, \ldots, X_n)^\top$.*

Z definície 14 vieme tiež ukázať, že ak $(X_1, \ldots, X_n)^\top$ je náhodný výber, potom sú náhodné veličiny $X_1, \ldots, X_n$ zameniteľné. Pre združenú distribučnú funkciu náhodného vektoru $\mathbf{X}_n = (X_1, \ldots, X_n)^\top$ platí z vlastností reálneho náhodného výberu

$$P[X_1 \leq x_1, \ \ldots \ , X_n \leq x_n] = P[X_1 \leq x_1] P[X_2 \leq x_2] \ldots P[X_n \leq x_n].$$

Z komutatívnej vlastnosti násobenia ďalej dostávame pre všetky permutácie $\tau : \mathbb{N} \to \mathbb{N}$ celých čísel $1, \ldots, n$

$$\begin{aligned} P[X_1 \leq x_1] P[X_2 \leq x_2] \ldots P[X_n \leq x_n] = \\ P[X_{\tau(1)} \leq x_{\tau(1)}] P[X_{\tau(2)} \leq x_{\tau(2)}] \ldots P[X_{\tau(n)} \leq x_{\tau(n)}], \end{aligned}$$

čím sme ukázali, že náhodné veličiny $X_1, \ldots, X_n$ sú zameniteľné. Naviac podľa Tibshirani a kol. (2019a), zameniteľnosť náhodných veličín $X_1, \ldots, X_n$ a Y je postačujúcim predpokladom k platnosti lemy 12. Taktiež je zameniteľnosť postačujúcim predpokladom k odvodeniu rozdelenia poradia náhodnej veličiny Y v náhodnom výbere $(X_1, \ldots, X_n, Y)^\top$, ako je uvedené v texte Tibshirani (2019). Zameniteľnosť náhodných veličín $X_1, \ldots, X_n, Y$ je teda jediným predpokladom nutným k platnosti konformných predikčných intervalov.

[1] Angl. Exchangeability.

4. Simulačná štúdia

V záverečnej kapitole budeme porovnávať empirické pokrytie predikčných intervalov prostredníctvom simulačnej štúdie. Jedna simulácia bude pozostávať z nasledujúcich krokov:

1. Vygenerovanie náhodného výberu $\mathcal{X}_n$ zo známeho rozdelenia F_X.
2. Skonštruovanie obojstranného predikčného intervalu $D_n(\mathcal{X}_n,\alpha)$ na hladine $1-\alpha$ pre $\alpha \in (0,1)$ s využitím náhodného výberu $\mathcal{X}_n$ vygenerovanom v prvom kroku.
3. Vygenerovanie pozorovania náhodnej veličiny Y z rozdelenia F_X, nezávisle na náhodnom výbere $\mathcal{X}_n$.
4. Ak $Y \in D_n(\mathcal{X}_n,\alpha)$, potom je výsledok simulácie 1, v opačnom prípade 0.

Po ukončení 100000 nezávislých Monte Carlo simulácii získame empirické pokrytie podielom počtu simulácii, ktorých výsledok bol 1 a všetkých simulácii. Taktiež počas simulácii budeme uvažovať 4 premenlivé faktory

1. Rozsah náhodného výberu $n \in \{50,100,200,400\}$.
2. Úroveň spoľahlivosti $\alpha \in \{0.1,0.05\}$.
3. Uvažujeme frekventistické/konformné a presné/asymptotické obojstranné predikčné intervaly
4. Rozdelenie F_X budeme postupne voliť z rozdelení
   - Normované normálne rozdelenie $N(0,1)$
   - Exponenciálne rozdelenie $Exp(4)$
   - Paretovo rozdelenie $Pareto(3,2)$
   - Cauchyho rozdelenie $Cauchy(1,1)$

Ďalej si uvedieme tvary predikčných intervalov pre rôzne rozdelenia uvažované v bode 4. Keďže konformné predikčné intervaly nezávisia na rozdelení náhodného výberu $\mathcal{X}_n$, budeme pre všetky rozdelenia uvažovať predikčný interval (3.7). Ďalej si zhrnieme, aké typy frekventistických predikčných intervalov budeme konštruovať. Poznamenáme, že frekventistické asymptotické predikčné intervaly budeme konštruovať vždy položením $W \equiv f(\mathcal{X}_n,Y) = Y$.

V prípade normovaného normálneho rozdelenia budeme uvažovať parametrickú rodinu (2.4) a presný frekventistický interval v tvare (2.7).

Pri konštrukcii asymptotického predikčného intervalu opäť uvažujeme parametrickú rodinu (2.4), kde máme neznámy parameter $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. Maximálne vierohodným odhadom je $\widehat{\boldsymbol{\theta}}_n = (\overline{X}_n, S_n^2)^\top$, kde $\overline{X}_n$ je výberový priemer spočítaný z náhodného výberu $\mathcal{X}_n$ a S_n^2 je výberový rozptyl spočítaný z $\mathcal{X}_n$. Ďalej definujeme náhodnú veličinu $\widehat{W}_n$, ktorá má podmienene pri danom náhodnom výbere $\mathcal{X}_n$ rozdelenie

$$\widehat{W}_n|\mathcal{X}_n \sim N(\overline{X}_n, S_n^2),$$

a označíme $F_{\widehat{W}_n}\left(\cdot;\widehat{\boldsymbol{\theta}}_n\right)$ ako distribučnú funkciu podmienenej náhodnej veličiny $\widehat{W}_n|\mathcal{X}_n$. Keďže maximálne vierohodný odhad $\widehat{\boldsymbol{\theta}}_n = \left(\overline{X}_n, S_n^2\right)^\top$ podmienený daným náhodným výberom $\mathcal{X}_n$ nie je náhodný a predstavuje konkrétnu realizáciu, distribučná funkcia $F_{\widehat{W}_n}\left(\cdot;\widehat{\boldsymbol{\theta}}_n\right)$ nezávisí na neznámom parametri $\boldsymbol{\theta} \in \Theta$. Označme kvantily $\widehat{w}_{1-\frac{\alpha}{2}} = F^{-1}_{\widehat{W}_n}\left(1-\frac{\alpha}{2};\widehat{\boldsymbol{\theta}}_n\right)$ a analogicky $\widehat{w}_{\frac{\alpha}{2}}$, pre $\alpha \in (0,1)$. Asymptotický obojstranný predikčný interval následne dostávame z výrazu (2.10) v tvare

$$D_n = \left\{y \in \mathbb{R} : \widehat{w}_{\frac{\alpha}{2}} \leq y \leq \widehat{w}_{1-\frac{\alpha}{2}}\right\}.$$

Druhým rozdelením je exponenciálne rozdelenie s parametrom $\boldsymbol{\theta}$ budeme konštruovať iba asymptotický predikčný interval a budeme predpokladať parametrickú rodinu (2.13) a asymptotický predikčný interval (2.14)

Tretím rozdelením ktoré budeme uvažovať, je Paretovo rozdelenie s neznámymi parametrami α a β, pre ktoré budeme konštruovať iba asymptotický predikčný interval. Uvažujeme teda parametrickú rodinu

$$\{F(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} = (\alpha,\beta)^\top, \boldsymbol{\theta} \in ((0,\infty) \times (0,\infty))\},$$

kde $F(\cdot;\boldsymbol{\theta})$ je distribučná funkcia Paretovho rozdelenia s neznámym parametrom $\boldsymbol{\theta} = (\alpha,\beta)^\top$. Maximálne vierohodným odhadom je $\widehat{\boldsymbol{\theta}}_n = \left(\widehat{\alpha}_{MLE}, \widehat{\beta}_{MLE}\right)^\top$, kde

$$\begin{aligned}\widehat{\beta}_{MLE} &= \min_{i\in\{1,\ldots,n\}} X_i,\\ \widehat{\alpha}_{MLE} &= \frac{n}{\sum_{i=1}^n ln(X_i) - \ n\ ln(\widehat{\beta}_{MLE})},\end{aligned}$$

Ďalej definujeme náhodnú veličinu $\widehat{W}_n$, ktorá má podmienene pri danom náhodnom výbere $\mathcal{X}_n$ rozdelenie

$$\widehat{W}_n|\mathcal{X}_n \sim Pareto(\widehat{\alpha}_{MLE}, \widehat{\beta}_{MLE}),$$

a ktorej distribučnú funkciu označíme $F_{\widehat{W}_n}\left(\cdot\,;\widehat{\boldsymbol{\theta}}_n\right)$. Distribučná funkcia náhodnej veličiny $\widehat{W}_n|\mathcal{X}_n$ nezávisí na neznámom parametri $\boldsymbol{\theta}$, nakoľko maximálne vierohodný odhad $\widehat{\boldsymbol{\theta}}_n = \left(\widehat{\alpha}_{MLE}, \widehat{\beta}_{MLE}\right)^\top$ podmienený daným náhodným výberom $\mathcal{X}_n$ nieje náhodný a predstavuje konkrétnu realizáciu. Označme kvantily $\widehat{w}_{1-\frac{\alpha}{2}} = F^{-1}_{\widehat{W}_n}\left(1-\frac{\alpha}{2};\widehat{\boldsymbol{\theta}}_n\right)$ a analogicky $\widehat{w}_{\frac{\alpha}{2}}$, pre $\alpha \in (0,1)$. Asymptotický obojstranný predikčný interval následne dostávame z výrazu (2.10) v tvare

$$D_n = \left\{y \in \mathbb{R} : \widehat{w}_{\frac{\alpha}{2}} \leq y \leq \widehat{w}_{1-\frac{\alpha}{2}}\right\}.$$

Posledným rozdelením je Cauchyho rozdelenie s neznámym parametrom $\boldsymbol{\theta}$ a predom známym parametrom $b = 1$ pre ktoré budeme konštruovať asymptotický predikčný interval. Uvažujeme teda parametrickú rodinu rozdelení

$$\{F(\cdot;(\boldsymbol{\theta},1)), \boldsymbol{\theta} \in \mathbb{R}\},$$

kde $F(\cdot;(\boldsymbol{\theta},1))$ je distribučná funkcia Cauchyho rozdelenia s neznámym parametrom $\boldsymbol{\theta}$. Následne maximálne vierohodný odhad $\widehat{\boldsymbol{\theta}}_n$, získame iteratívne Newton-Raphsonovym algoritmom z rovnice

$$2\sum_{i=1}^{n}\frac{X_i - \widehat{\boldsymbol{\theta}}_n}{1+(X_i-\widehat{\boldsymbol{\theta}}_n)^2} = 0.$$

Ďalej definujeme náhodnú veličinu $\widehat{W}_n$, ktorá má podmienene pri danom náhodnom výbere $\mathcal{X}_n$ rozdelenie

$$\widehat{W}_n|\mathcal{X}_n \sim Cauchy\left(\widehat{\boldsymbol{\theta}}_n, 1\right),$$

a ktorej distribučnú funkciu označíme $F_{\widehat{W}_n}\left(\cdot\,;\widehat{\boldsymbol{\theta}}_n\right)$. Distribučná funkcia náhodnej veličiny $\widehat{W}_n|\mathcal{X}_n$ nezávisí na neznámom parametri $\boldsymbol{\theta}$, keďže maximálne vierohodný odhad $\widehat{\boldsymbol{\theta}}_n$ podmienený daným náhodným výberom $\mathcal{X}_n$ nie je náhodný. Ako pri predchádzajúcich konštrukciách predikčných intervalov, označíme kvantily $\widehat{w}_{1-\frac{\alpha}{2}} = F^{-1}_{\widehat{W}_n}\left(1-\frac{\alpha}{2};\widehat{\boldsymbol{\theta}}_n\right)$ a analogicky $\widehat{w}_{\frac{\alpha}{2}}$, pre $\alpha \in (0,1)$. Asymptotický obojstranný predikčný interval následne dostávame z výrazu (2.10) v tvare

$$D_n = \left\{y \in \mathbb{R} : \widehat{w}_{\frac{\alpha}{2}} \leq y \leq \widehat{w}_{1-\frac{\alpha}{2}}\right\}.$$

Po odvodení všetkých typov predikčných intervalov v závislosti na rozdelení, zhrnieme výsledky simulačnej štúdie v nasledujúcej tabuľke.

Rozdelenie	Rozsah	Frekv. ex.		Frekv. as.		Konf ex.	
		90%	95%	90%	95%	90%	95%
$N(0,1)$	$n = 50$	0.8998	0.9491	0.8912	0.9428	0.9202	0.9597
	$n = 100$	0.8999	0.9497	0.8939	0.9448	0.9023	0.9605
	$n = 200$	0.9000	0.9504	0.8970	0.9478	0.8997	0.9500
	$n = 400$	0.9008	0.9504	0.8987	0.9491	0.9018	0.9508
$Exp(4)$	$n = 50$			0.8958	0.9465	0.9210	0.9604
	$n = 100$			0.8979	0.9483	0.9030	0.9609
	$n = 200$			0.8985	0.9490	0.9007	0.9509
	$n = 400$			0.9008	0.9506	0.9011	0.9495
$Pareto(3,2)$	$n = 50$			0.8765	0.9262	0.9199	0.9595
	$n = 100$			0.8883	0.9387	0.9006	0.9603
	$n = 200$			0.8950	0.9442	0.9001	0.9497
	$n = 400$			0.8960	0.9456	0.8995	0.9495
$Cauchy(1,1)$	$n = 50$			0.8997	0.9502	0.9201	0.9593
	$n = 100$			0.9009	0.9511	0.9022	0.9603
	$n = 200$			0.9010	0.9503	0.9006	0.9495
	$n = 400$			0.9012	0.9513	0.9004	0.9505

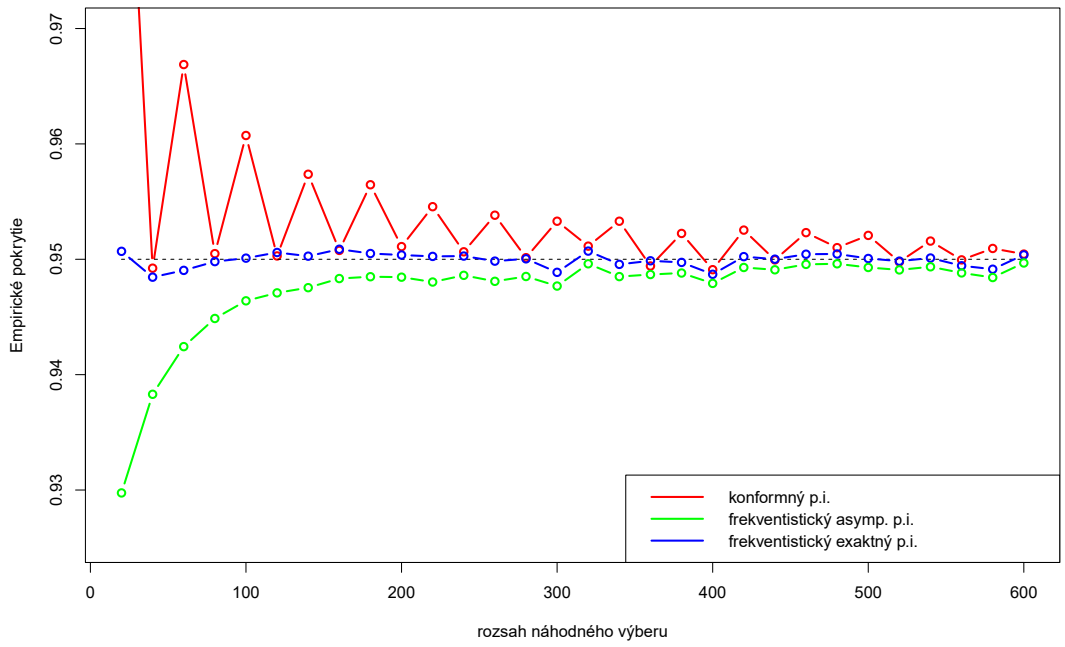
Tabuľka 4.1: Porovnanie teoretického a empirického pokrytia predikčných intervalov pre $N(0,1)$ rozdelenie, $Exp(4)$ rozdelenie, $Pareto(3,2)$ rozdelenie a $Cauchy(1,1)$ rozdelenie.

Z výslednej tabuľky 4 vidíme, že frekventistické exaktné predikčné intervaly dosahujú očakávané úrovne empirického pokrytia pre všetky rozsahy náhodného výberu, kde najvýraznejšie nedodržanie teoretického pokrytia dosahuje 0.09%.

V prípade frekventistických asymptotických intervalov môžeme pozorovať očakávané asymptotické správanie empirického pokrytia, ktoré je viditeľné v prípade všetkých preskúmaných rozdelení.

V prípade konformných predikčných intervalov pozorujeme, že empirické pokrytie predikčných intervalov dosahuje požadovaného teoretického pokrytia a najvýraznejšie nedodržanie teoretického pokrytia dosahuje 0.05%, čo je menej ,ako v prípade frekventistického exaktného predikčného intervalu.

Rozdielny náhľad na empirické pokrytie predikčných intervalov nám poskytuje Obr. 4.1, kde môžeme pozorovať vývoj empirického pokrytia obojstranných frekventistických exaktných, frekventistických asymptotických a konformných predikčných intervalov, kde uvažujeme normované normálne rozdelenie a rozsah náhodného výberu $n = \{20,40,60,\ldots,600\}$. Pre výpočet jedného bodu v grafe využijeme opäť 100000 nezávislých Monte-Carlo simulácii.

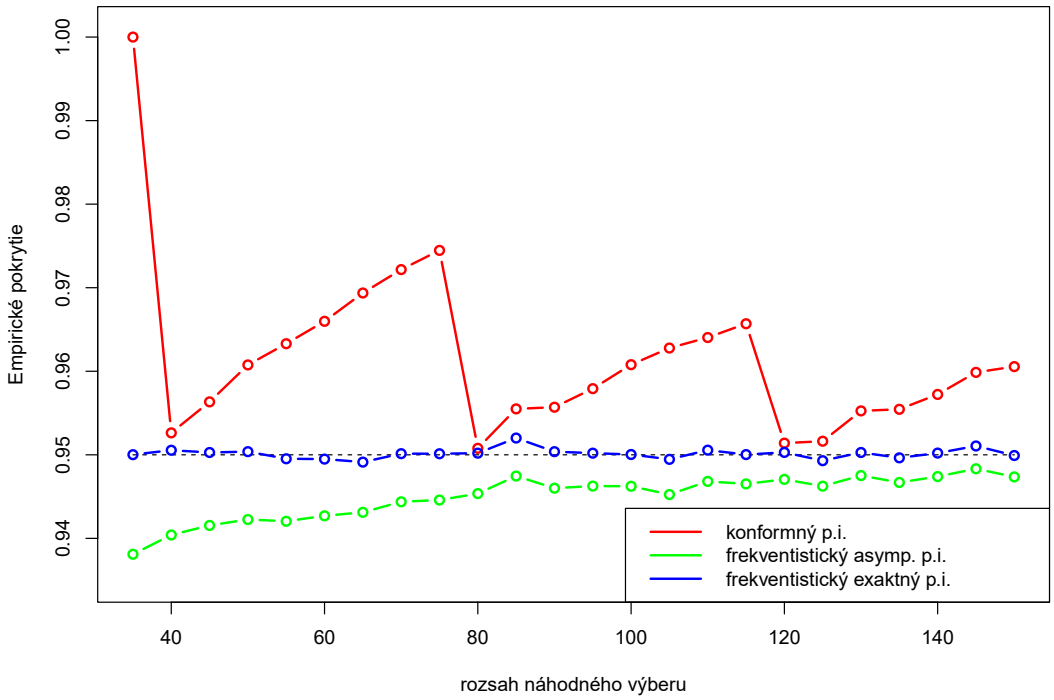


Obr. 4.1: Porovnanie empirického pokrytia predikčných intervalov vzhľadom na meniacu sa dĺžku rozsahu náhodného výberu $n = \{20,40,60,\ldots,600\}$ pre normované normálne rozdelenie a $\alpha = 0.05$

Na Obr. 4.1 môžeme pozorovať, že

- Empirické pokrytie frekventistického presného predikčného intervalu dodržuje hladinu 0.95 pre všetky uvažované rozsahy náhodného výberu s minimálnymi odchylkami ako bude bližšie pozorovateľné aj na Obr. 4.2.
- Empirické pokrytie frekventistického asymptotického predikčného intervalu sa asymptoticky približuje k požadovanému teoretickému pokrytie.
- Konformný predikčný interval dodržuje požadovanú teoretickú hladinu 0.95, avšak pozorujeme skokovitý priebeh empirického pokrytia.

Pre detailnejšie preskúmanie skokovitého priebehu empirického pokrytia konformného predikčného intervalu využijeme Obr. 4.2, kde sa bližšie pozrieme na priebeh empirických pokrytí predikčných intervalov.



Obr. 4.2: Porovnanie empirického pokrytia predikčných intervalov vzhľadom na meniacu sa dĺžku rozsahu náhodného výberu $n = \{35{,}40{,}45, \ldots, 150\}$ pre normované normálne rozdelenia a $\alpha = 0.05$

Na obrázku 4.2 pozorujeme bližšie skokovitý priebeh empirického pokrytia konformných predikčných intervalov. Ako prvé si môžeme všimnúť, že pre $n = 35$ dosahuje empirické pokrytie hodnotu 1. Dôvodom pre dosiahnutú hodnotu je fakt, že konformný predikčný interval je v tvare

$$D_{35}(\mathcal{X}_{35}) = \{y \in \mathbb{R} : -\infty \leq y \leq \infty\}.$$

Pre $n = 40$ pozorujeme skok, ktorý je dôsledkom zmien poradí udávajúcich požadované empirické kvantily. Pozorujeme, že pre $n = 35$ platí

$$\begin{aligned} \widehat{q}^{\mathrm{k}-}_{\frac{0.05}{2}} &:= X_{(k_{0.025})} = X_{(1)} = -\infty; \\ \widehat{q}^{\mathrm{k}+}_{1-\frac{0.05}{2}} &:= X_{(k_{0.975})} = X_{(36)} = \infty. \end{aligned}$$

Zatiaľ čo pre $n = 40$ platí

$$\begin{aligned} \widehat{q}^{\mathrm{k}-}_{\frac{0.05}{2}} &:= X_{(k_{0.025})} = X_{(2)}; \\ \widehat{q}^{\mathrm{k}+}_{1-\frac{0.05}{2}} &:= X_{(k_{0.975})} = X_{(40)}. \end{aligned}$$

Najbližší skok sa nachádza medzi $n = 75$ a $n = 85$, kde podobným spôsobom môžeme ukázať, že pre $n = 75$ platí

$$\begin{aligned}\widehat{q}^{\mathrm{k}-}_{\frac{0.05}{2}} &:= X_{(k_{0.025})} = X_{(2)};\\ \widehat{q}^{\mathrm{k}+}_{1-\frac{0.05}{2}} &:= X_{(k_{0.975})} = X_{(75)}.\end{aligned}$$

Zatiaľ čo pre $n = 80$ platí

$$\begin{aligned}\widehat{q}^{\mathrm{k}-}_{\frac{0.05}{2}} &:= X_{(k_{0.025})} = X_{(3)};\\ \widehat{q}^{\mathrm{k}+}_{1-\frac{0.05}{2}} &:= X_{(k_{0.975})} = X_{(79)}.\end{aligned}$$

Podobným spôsobom by sme zistili, že nasledujúci skok nastane opäť v dôsledku zmeny poradia definujúceho požadovaný empirický kvantil. Ukázali sme teda, že skokovitý priebeh empirického pokrytia konformého predikčného intervalu nastáva v dôsledku diskrétneho rozdelenia poradia náhodných veličín v náhodnom výbere $\mathcal{X}_n$, ktoré definuje empirické kvantily.

Záver

V tejto bakalárskej práci sme sa zaoberali dvomi metódami konštrukcie predikčných intervalov. Ako prvý sme predstavili historicky starší spôsob výpočtu, tzv. frekvenistické predikčné intervaly. V kapitole 2 sme uviedli všeobecný algoritmus výpočtu presných intervalov a zhrnuli teoretické predpoklady potrebné pri výpočte.

Následne sme tento postup ilustrovali na príklade, v ktorom sme predpokladali, že náhodný výber je z parametrickej rodiny normálnych rozdelení. Pre lepšiu predstavu sme výsledky graficky znázornili. V závere druhej kapitoly sme teóriu výpočtu frekvenistických predikčných intervalov rozšírili o prípad, keď neexistuje presná pivotálna štatistika W. V Algoritme 2 sme predstavili postup výpočtu asymptotických predikčných intervalov. Teoretický postup sme opäť aplikovali na príklade, kde sme uvažovali náhodný výber z parametrickej rodiny exponenciálnych rozdelení.

V tretej kapitole sme predstavili novší spôsob výpočtu predikčných intervalov, pri ktorom nie je vopred nutná informácia o tom, z akého rozdelenia pochádza náhodný výber. Zhrnuli sme základné poznatky a predpoklady, ktoré sú potrebné pri stanovení tohto typu predikčného intervalu. Čitateľovi sme poskytli grafickú predstavu výpočtu na jednoduchom príklade.

V poslednej kapitole sme predstavili výsledky vlastnej simulačnej štúdie. Využili sme programovací jazyk ***R*** (Team Development Core, 2023). Jej cieľom bolo porovnanie empirického pokrytia konformných a frekventistických predikčných intervalov. Porovnanie sme realizovali na štyroch typoch pravdepodobnostných rozdelení, dvoch úrovniach hladín významnosti a rôzne rozsahy náhodného výberu. Nakoniec sme výsledky simulačnej štúdie sme prezentovali v tabuľke 4 a obrázkoch 4.1 a 4.2.

V simulačnej štúdii sme zistili, že predikčné intervaly sa správajú v súlade s našimi očakávaniami, nakoľko pri frekventistických asymptotických predikčných intervaloch je jasne pozorovateľný asymptotický priebeh empirického pokrytia predikčného intervalu, pri frekventistických exaktných predikčných intervaloch pozorujeme minimálne odchylky od očakávaného pokrytia a po uvážení diskrétneho rozdelenia poradia je skokovitý priebeh konformných predikčných intervalov taktiež očakávaný.

Literatúra

Beran, R. (1990). Calibrating prediction regions. *J. Amer. Statist. Assoc.*, **85**(411), 715–723. ISSN 0162-1459. URL http://links.jstor.org/sici?sici=0162-1459(199009)85:411<715:CPR>2.0.CO;2-Q&origin=MSN. Accessed: 2023-03-19.

Kulich, M. (2022). *POZNÁMKY K PŘEDNÁŠCE*. Charles University, Lectures notes. URL https://www2.karlin.mff.cuni.cz/~komarek/vyuka/2022_23/nmsa331/ms1.pdf. Accessed: 2023-03-27.

Lawless, J. F. a Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, **92**(3), 529–542. ISSN 0006-3444. doi: 10.1093/biomet/92.3.529. URL https://doi.org/10.1093/biomet/92.3.529. Accessed: 2023-03-19.

Shafer, G. a Vovk, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.*, **9**, 371–421. ISSN 1532-4435. URL https://jmlr.csail.mit.edu/papers/volume9/shafer08a/shafer08a.pdf. Accessed: 2023-03-19.

Tibshirani, R. (2019). Advances and challenges in conformal inference. URL https://www.stat.cmu.edu/~ryantibs/talks/conformal-2019.pdf. Accessed: 2023-03-19.

Tibshirani, R., Barber, R., Candès, E. a Ramdas, A. (2019a). Conformal prediction under covariate shift. URL https://www.stat.cmu.edu/~ryantibs/papers/weightedcp.pdf. Accessed: 2023-03-19.

Tibshirani, R., Barber, R., Candès, E. a Ramdas, A. (2019b). Supplement to "conformal prediction under covariate shift". URL https://www.stat.cmu.edu/~ryantibs/papers/weightedcp-supp.pdf. Accessed: 2023-03-19.

Vovk, V., Gammerman, A. a Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. ISBN 978-3-031-06648-1.

Zoznam obrázkov

Zoznam tabuliek