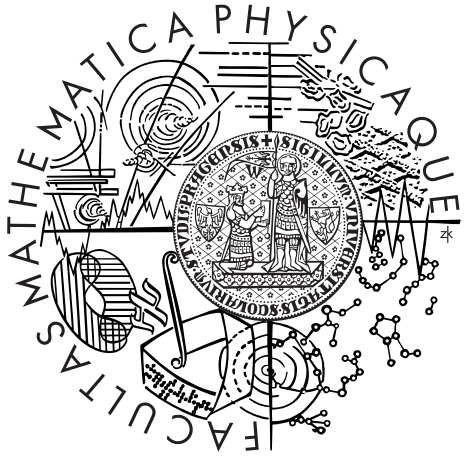
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Jakub Šimičák

Teoretické a empirické kvantily a ich využitie pri konštrukcií predikčných intervalov

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: 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 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: 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 žiadny pravdepodobnostný model a vedie na konformný predikčný interval, ktorý využíva empirické kvantily príslušného náhodného výberu. V priebehu 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: 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 selection. 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 **2**

1 Značenie, definície a vlastnosti **3**
1.1 Vlastnosti reálnej náhodnej veličiny 3
1.2 Empirické odhady 4
1.3 Predikčné intervaly 6

2 Frekventistické predikčné intervaly **8**
2.1 Konštrukcia exaktných predikčných intervalov 8
2.2 Konštrukcia asymptotických predikčných intervalov 11

3 Konformné predikčné intervaly **14**
3.1 Konštrukcia konformných predikčných intervalov 14
3.1.1 Zameniteľnosť 16

4 Simulačná štúdia **18**

Záver **22**

Literatúra **23**

Zoznam obrázkov **24**

Zoznam tabuliek **25**

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, financií, 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 väčšinou 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 prevdepodobnostný model, ktorý by im mohol zodpovedať. Vo všeobecnosti existujú dva hlavné typy 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, ako 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 predikcií. Typicky, úroveň spoľahlivosti zodpovedá hodnote $1 - \alpha$, kde α volíme z intervalu (0,1).

Cieľom bakalárskej práce je zoznámiť čitateľa s dvomi rozdielnymi typmi konštrukcie už vyššie spomenutých predikčných intervalov a postupne prejsť ich príslušné predpoklady a spôsob konštrukcie v jednoduchých príkladoch. Na záver oba typy predikčných intervalov porovnáme v simulačnej štúdií.

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 konštrukcií 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 pre všetky $u \in (0,1)$ sa funkcia definovaná predpisom*

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

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 funkcií 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, že rozdelenie náhodnej veličiny X je známe až na parameter $\boldsymbol{\theta}$, 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 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 nadchádzajúcej časti si zrekapitulujeme náhodný výber, usporiadaný náhodný 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 $X_1,\ldots,X_n$ ako celku 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 definícií.

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 sekcií 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 $x \in \mathbb{R}$, 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 $X_1,\ldots,X_n$, 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 nieje 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 $X_1,X_2,\ldots,X_n$.*

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

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

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 dispozícií a označme náhodnú veličinu $Y = X_{n+1}$, ktorej budúcu realizáciu chceme odhadnúť. Nakoľko z bodového odhadu nemáme informáciu o presnosti zostrojeného odhadu, rozhodneme sa využiť pre odhad budúcej realizácie náhodnej veličiny Y interval $D \subseteq \mathbb{R}$ s nami určenou mierou neistoty.

Interval D však nevolíme úplne náhodne. Keďže náhodné veličiny $X_1, \ldots, X_n$ a Y sú rovnako rozdelené, využijeme pri konštrukcií 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 definícií.

Definícia 10. *Nech $\alpha \in (0,1)$ a $\mathcal{X}_n$ označuje reálny náhodný výber $X_1, \ldots, X_n$ 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 $D = D_n(\alpha,\mathcal{X}) \subseteq \mathbb{R}$ 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 splňujú

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

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

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

  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 $\mathcal{X}_n$ označuje známy reálny náhodný výber $X_1,\ldots,X_n$ z rozdelenia $F_X(x,\boldsymbol{\theta})$ patriaceho do parametrickej rodiny rozdelení $\mathcal{F} = \{F_X(x,\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\, , p \in \mathbb{N}\}$. Nech náhodná veličina Y má rozdelenie $F_X(x,\boldsymbol{\theta})$ 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 $\{\mathrm{Exp}(\lambda), \lambda \in (0,\infty)\}$.

2. Frekventistické predikčné intervaly

Ako prvý, historicky starší spôsob, konštrukcie predikčných intervalov si predstavíme frekventistickú metódu a s ňou spojené frekventistické predikčné intervaly. Jedná sa 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 ďalej rozšírime aj na asymptotickú verziu.

2.1 Konštrukcia exaktných predikčných intervalov

Nech náhodný výber $\mathcal{X}_n$ a náhodná veličina Y sú definované rovnako ako v motivačnom probléme 1. Exaktný frekventistický predikčný interval môžeme skonštruovať pomocou nasledujúceho postupu.

1. Uvažujme reálnu, prostú a merateľnú funkciu $f(\mathcal{X}_n,Y) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ a zadefinujme 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 $\boldsymbol{\theta} \in \Theta$ 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 funkcií F_W naviac predpokladáme, že je absolútne spojitá a rastúca funkcia. Potom z poznámky pod definíciou 2 vieme, že kvantilová funkcia F_W^{-1} je inverzná k funkcií F_W a označíme príslušný α-kvantil náhodnej veličiny W 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

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

   Následne, exaktný obojstranný predikčný interval pre Y dostávame z rovnosti

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

aplikáciou funkcie $g(\mathcal{X}_n, \cdot\,)$

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

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

$$P[g(\mathcal{X}_n,w_{\frac{\alpha}{2}}) < Y < g(\mathcal{X}_n,w_{1-\frac{\alpha}{2}})] = 1-\alpha. \tag{2.2}$$

Z rovnice (2.2) 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 $F_X(x,\boldsymbol{\theta})$ patrí do parametrickej rodiny normálnych rozdelení s parametrami μ a σ^2.

Príklad 2. *Nech $\mathcal{X}_n$ označuje reálny náhodný výber $X_1,\ldots,X_n$ z rozdelenia $F_X(x,\boldsymbol{\theta})$ patriaceho do parametrickej rodiny normálnych rozdelení*

$$\{N(\mu,\sigma^2),(\mu,\sigma^2)^\top \in (\mathbb{R}\times(0,\infty))\}. \tag{2.3}$$

Nech náhodná veličina Y má rozdelenie $F_X(x,\boldsymbol{\theta})$ a je nezávislá na náhodnom výbere $\mathcal{X}_n$. V takom prípade je parameter $\boldsymbol{\theta} = (\mu,\sigma^2)^\top$, parametrický priestor $\Theta \subseteq (\mathbb{R}\times(0,\infty))$ a následne budeme postupovať v súlade s postupom konštrukcie exaktného predikčného intervalu .

1. *Zadefinujeme*

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

kde $\overline{X}_n$ má rozdelenie $N\left(\mu,\frac{\sigma^2}{n}\right)$ a jedná sa 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

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

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

$$W \equiv \frac{Y-\overline{X}_n}{\sqrt{\left(1+\frac{1}{n}\right)S_n^2}} \sim t_{n-1}, \tag{2.4}$$

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

$$P\left[\overline{X}_n - S_n\, t_{n-1}\left(1-\tfrac{\alpha}{2}\right)\sqrt{\left(1+\tfrac{1}{n}\right)} \le Y \le \overline{X}_n + S_n\, t_{n-1}\left(1-\tfrac{\alpha}{2}\right)\sqrt{\left(1+\tfrac{1}{n}\right)}\right] = 1-\alpha, \qquad (2.5)$$

dostávame obojstranný predikčný interval $D_n = (L_1(\mathcal{X}_n,\frac{\alpha}{2}),L_2(\mathcal{X}_n,1-\frac{\alpha}{2}))$ *s hranicami*

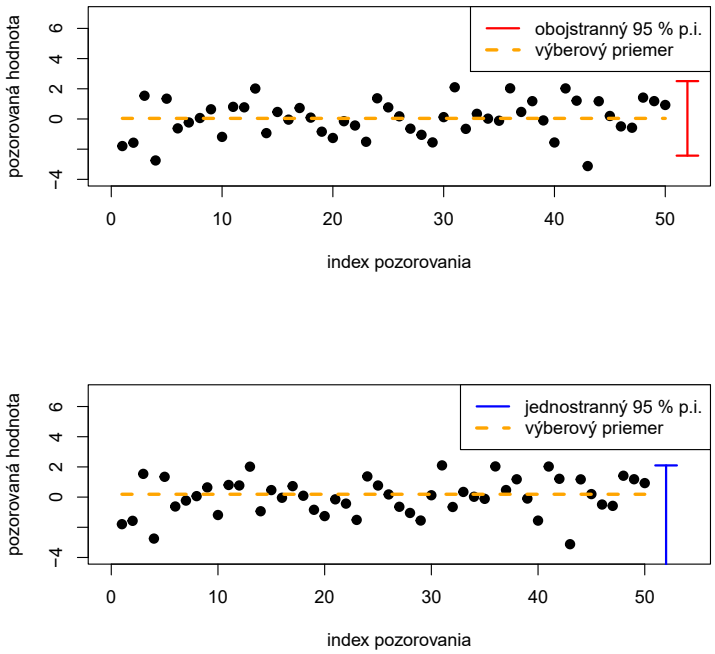
$$L_1\left(\mathcal{X}_n,\frac{\alpha}{2}\right) = \overline{X}_n - S_n\, t_{n-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\left(1+\frac{1}{n}\right)},$$
$$L_2\left(\mathcal{X}_n,1-\frac{\alpha}{2}\right) = \overline{X}_n + S_n\, t_{n-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\left(1+\frac{1}{n}\right)}.$$

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

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

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* $D_n = (-\infty,L_2(\mathcal{X}_n,1-\alpha))$ *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 predikčných intervalov

Ako si môžeme všimnúť, postup konštrukcie presných frekventistických predikčných intervalov závisí na nájdení presnej pivotálnej štatistiky W, ktorej rozdelenie nezávisí na parametri $\boldsymbol{\theta} \in \Theta$. Predpoklad, o existencií presnej pivotálnej štatistiky je v mnohých prípadoch porušený a teda je potrebné nájsť spôsob akým skonštruovať frekventistický predikčný interval aj v prípade, keď presná pivotálna štatistika neexistuje.

Uvažujme teda prípad, že rozdelenie náhodnej veličiny W závisí na parametri $\boldsymbol{\theta}$ a označme jej distribučnú funkciu $F_W(x,\boldsymbol{\theta})$ o ktorej predpokladáme, že je absolútne spojitá a rastúca. Článok od autorov Lawless a Fredette (2005) rieši tento problém definovaním náhodnej veličiny $\widehat{W}_n$ s distribučnou funkciou

$$F_{\widehat{W}_n}(x) := F_W(x,\widehat{\boldsymbol{\theta}}_n),$$

kde $\widehat{\boldsymbol{\theta}}_n$ je konzistentný odhad parametru $\boldsymbol{\theta}$, v zmysle $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\theta}$, pre $n \to \infty$. Náhodná veličina $\widehat{W}_n$, nezávisí na parametri $\boldsymbol{\theta}$, čo znamená že môže byť použitá ako pivotálna štatistika pre konštrukciu predikčného intervalu. Podľa článku od autorov Lawless a Fredette (2005), postupnosť náhodných veličín $\widehat{W}_n$ konverguje v distribúcií k náhodnej veličine W, tj.

$$\widehat{W}_n \xrightarrow{\mathrm{D}} W, \text{ pre } n \to \infty,$$

z čoho vyplýva že predikčný interval skonštruovaný na základe náhodnej veličiny

$\widehat{W}_n$ bude dosahovať pokrytie $1-\alpha$ asymptoticky. Z tohto dôvodu hovoríme že náhodná veličina $\widehat{W}_n$ je asymptoticky pivotálna štatistika.

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

Príklad 3. *Nech $\mathcal{X}_n$ je reálny náhodný výber $X_1,\ldots,X_n$ z rozdelenia $F_X(x,\boldsymbol{\theta})$ patriaceho do parametrickej rodiny exponenciálnych rozdelení*

$$\{Exp(\lambda), \lambda \in (0,\infty)\}. \tag{2.6}$$

Nech náhodná veličina Y má rozdelenie $F_X(x,\boldsymbol{\theta})$ a je nezávislá na náhodnom výbere $\mathcal{X}_n$ a distribučná funkcia exponenciálneho rozdelenia spĺňuje

$$F_X(x,\lambda) = \begin{cases} 1-e^{-\lambda x} & \text{pre } x \geq 0, \\ 0 & \text{inak .} \end{cases}$$

V tomto prípade je parameter $\boldsymbol{\theta} = \lambda$, parametrický priestor $\Theta = (0,\infty)$ a ďalej budeme budeme postupovať v súlade s postupom konštrukcie predikčných intervalov.

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

   $$W \sim Exp(\lambda).$$

   Vidíme, že rozdelenie náhodnej veličiny W závisí na parametri λ. Následne odhadneme parameter λ maximálne vierohodným odhadom $\frac{1}{\overline{X}_n}$, ktorý je konzistentný a zadefinujeme náhodnú veličinu $\widehat{W}_n$ s distribučnou funkciou $F_{\widehat{W}_n}(x) := F_W(x,\frac{1}{\overline{X}_n})$, čím dostávame

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

   Rozdelenie náhodnej veličiny $\widehat{W}_n$ nezávisí na parametri $\lambda > 0$, čo znamená že sme našli asymptoticky pivotálnu štatistiku.

2. *Pre dané $\alpha \in (0,1)$, označíme príslušné kvantily exponenciálneho rozdelenia s parametrom $\frac{1}{\overline{X}_n}$ ako $q\left(1-\frac{\alpha}{2}\right)$ a $q\left(\frac{\alpha}{2}\right)$.*

3. *Obojstranný predikčný interval následne dostávame z výrazu*

   $$P\left[q\left(\frac{\alpha}{2}\right) < Y < q\left(1-\frac{\alpha}{2}\right)\right] \to 1-\alpha, \tag{2.7}$$

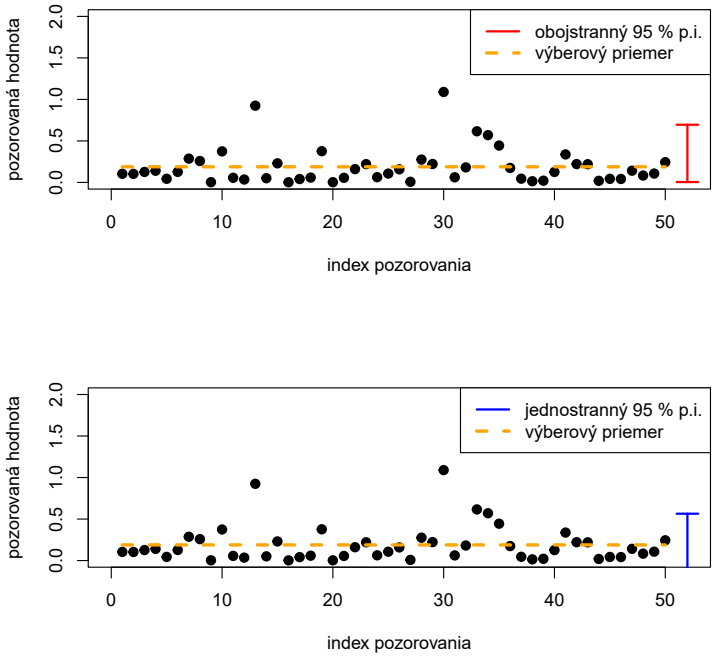
   pre $n \to \infty$. Skonštruovali sme asymptotický, obojstranný predikčný interval $D_n = (L_1(\mathcal{X}_n,\frac{\alpha}{2}), L_2(\mathcal{X}_n,1-\frac{\alpha}{2}))$ s krajnými bodmi

   $$L_1\left(\mathcal{X}_n,\frac{\alpha}{2}\right) = q\left(\frac{\alpha}{2}\right),$$
   $$L_2\left(\mathcal{X}_n,1-\frac{\alpha}{2}\right) = q\left(1-\frac{\alpha}{2}\right).$$

Podobne ľavostranný predikčný interval dostávame z výrazu

$$P\left[Y < q\left(1-\alpha\right)\right] \to 1-\alpha \text{ pre } n \to \infty,$$

v tvare $D_n(\mathcal{X}_n,\alpha) = (-\infty, q\left(1-\alpha\right))$, *kde* $q\left(1-\alpha\right)$ *je* $(1-\alpha)$ *kvantil exponenciálneho rozdelenia s parametrom* $\frac{1}{\overline{X}_n}$, *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 ľavo stranného asymptotického predikčného intervalu pre $Y = X_{51}$ a daným $\alpha = 0.05$.

Z predvedených príkladov konštrukcií frekventistických predikčných intervalov môžeme pozorovať, že:

- Pre konštrukciu predikčného intervalu je potrebné poznať parametrickú rodinu rozdelení $\mathcal{F}$ z ktorej pochádza náhodný výber $\mathcal{X}_n$, na základe ktorého konštruujeme buď pivotálnu štatistiku W alebo asymptoticky pivotálnu štatistiku $\widehat{W}_n$
- Na hraniciach frekventistických predikčných intervaloch vystupujú teoretické kvantily rozdelenia F_W v prípade pivotálnej štatistiky W, alebo teoretické kvantily rozdelenia $F_{\widehat{W}_n}$ v prípade asymptotickej pivotálnej štatistiky $\widehat{W}_n$

Problém s konštrukciou frekventistických predikčných intervalov však nastáva v momente, keď nepoznáme parametrickú rodinu $\mathcal{F}$ z ktorej pochádza rozdelenie náhodného výberu $\mathcal{X}_n$. V takom prípade nevieme nájsť žiadnu pivotálnu štatistiku, či už presnú alebo asymptotickú a teda nevieme skonštruovať ani predikčné intervaly. Ukazuje sa tak potreba pre iný postup konštrukcie predikčných intervalov, ktorý by sa dokázal vysporiadať aj s tou alternatívou, že $\mathcal{F}$ nieje známa.

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 predikčných 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 $D = (-\infty,\widehat{q}_{1-\alpha})$ 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 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 o nej uvažovať v hypotetickom zmysle a teda uvažovať všetky možné realizácie $y \in \mathbb{R}$ náhodnej veličiny Y.

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

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

Z definície 9 vieme, že empirický kvantil $\widehat{q}_\alpha$ je definovaný ako k_α-tá poriadková štatistika, čím dostávame, že náhodná veličina $\phi(\mathcal{X}_n,Y)$ predstavuje prirodzené

číslo k_α. Ďalej chceme zistiť príslušné $\alpha \in (0,1)$, ktoré dostaneme podielom $\frac{k_\alpha}{n+1}$, keďže náhodný výber $X_1,\ldots,X_n,Y$ má rozsah $n+1$. Zadefinujeme merateľné zobrazenie $\pi(\mathcal{X}_n,Y) : \mathbb{R}^n \times \mathbb{R} \to \{\frac{1}{n+1},\ldots,1\}$ s predpisom

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

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

$$P\left[\pi(\mathcal{X}_n,Y) \leq \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}\right] \geq 1-\alpha. \tag{3.1}$$

Keďže realizáciu náhodnej veličiny Y nepoznáme ale uvažujeme jej všetky možné realizácie $y \in \mathbb{R}$, využijeme nerovnosť (3.1) a ako predikčný interval budeme uvažovať všetky realizácie y náhodnej veličiny Y pre ktoré nerovnosť (3.1) platí, čím dostávame predikčný interval

$$D_n(\mathcal{X}_n,\alpha) = \left\{y \in \mathbb{R} : \pi(\mathcal{X}_n,y) \leq \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}\right\}. \tag{3.2}$$

Predikčný interval (3.2) podľa článku od autora Tibshirani (2019) splňuje požadovanú hladinu $1-\alpha$, a naviac splňuje

$$P[Y \in D_n(\mathcal{X}_n,\alpha)] = P\left[Y \leq \widehat{q}^{\,\mathrm{k}}_{1-\alpha}\right] \geq 1-\alpha, \tag{3.3}$$

kde $\widehat{q}^{\,\mathrm{k}}_{1-\alpha}$ označuje empirický $(1-\alpha)$ kvantil z náhodného výberu $X_1,\ldots,X_n,Y$. Obojstranný predikčný interval pre $\alpha \in (0,1)$ získame analogickým postupom, čím dostávame

$$D_n(\mathcal{X}_n,\alpha) = \left\{y \in \mathbb{R} : \pi(\mathcal{X}_n,y) \leq \tfrac{\lceil (n+1)(1-\frac{\alpha}{2}) \rceil}{n+1} \wedge \pi(\mathcal{X}_n,y) \geq \tfrac{\lceil (n+1)(\frac{\alpha}{2}) \rceil}{n+1}\right\}, \tag{3.4}$$

a analogicky zapíšeme v tvare

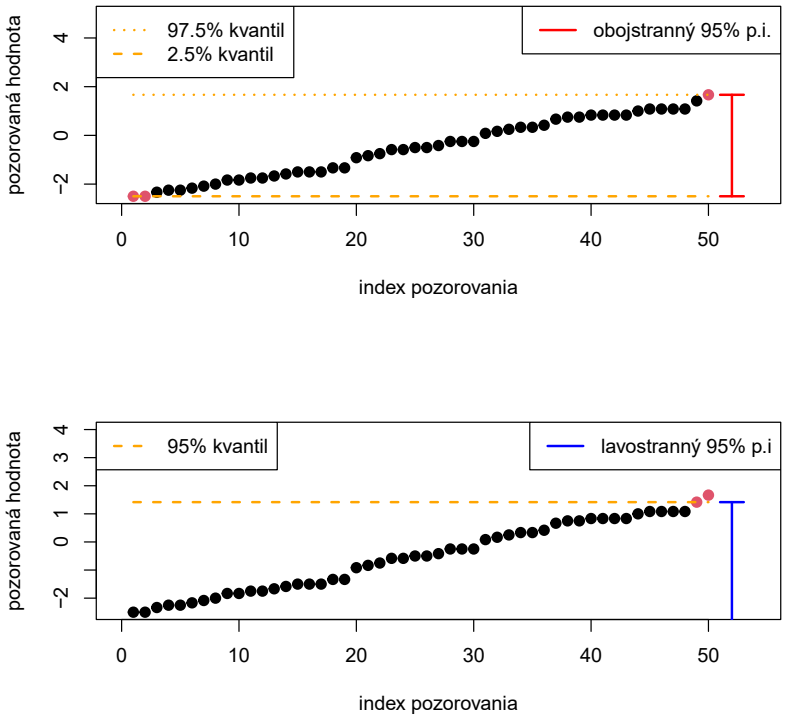
$$P[Y \in D_n(\mathcal{X}_n,\alpha)] = P\left[\widehat{q}^{\,\mathrm{k}}_{\frac{\alpha}{2}} \leq Y \leq \widehat{q}^{\,\mathrm{k}}_{1-\frac{\alpha}{2}}\right] \geq 1-\alpha,$$

kde $\widehat{q}^{\,\mathrm{k}}_{\frac{\alpha}{2}}$ a $\widehat{q}^{\,\mathrm{k}}_{1-\frac{\alpha}{2}}$ označujú empirický $(\frac{\alpha}{2})$ kvantil a $(1-\frac{\alpha}{2})$ kvantil z náhodného výberu $X_1,\ldots,X_n,Y$.

Z výpočetného hľadiska je však dôležité poznamenať, že k dispozícií máme iba pozorovaný náhodný výber $\mathcal{X}_n$ keďže o realizácií náhodnej veličiny Y uvažujeme v hypotetickom zmysle. Je preto potrebné rozšíriť pozorovaný náhodný $\mathcal{X}_n$ výber o jedno pozorovanie y, ktoré bude predstavovať pozorovanie náhodnej veličiny Y. Pre zachovanie platnosti konformných predikčných intervalov autor Tibshirani (2019) uvádza, že práve pozorovanie y je potrebné položiť nekonečnu.

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 nekonečnu.



Obr. 3.1: Pozorovaný usporiadaný náhodný výber $\mathcal{X}_{50}$ a grafické znázornenie obojstranného a ľavo stranného 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

3.1.1 Zameniteľnosť

Pri konštrukcií 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].

[1] Angl. Exchangeability.

Definícia 11 (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 11, vieme tiež ukázať, že ak $X_1, \ldots, X_n$ 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 (2019) zameniteľnosť náhodných veličín $X_1, \ldots, X_n$ a Y postačujúcim predpokladom k odvodeniu rozdelenia poradia náhodnej veličiny Y medzi náhodnými veličinami $X_1, \ldots, X_n, Y$ a následného odvodenia presného pokrytia predikčného intervalu 3.3. Autori Shafer a Vovk (2008) uvádzajú, že zameniteľnosť náhodných veličín implikuje aj rovnaké rozdelenie, čím zameniteľnosť náhodných veličín $X_1, \ldots, X_n, Y$ stáva jediným predpokladom nutným k platnosti konformných predikčných intervalov.

Na záver kapitoly je vhodné poznamenať, že empirické kvantily figurujúce v krajných bodoch predikčných intervalov môžu spôsobovať, že v simulačnej štúdií nebude empirické pokrytie konformných predikčných intervalov dosahovať odovodené teoretické pokrytie. Dôvodom je že, empirický odhad kvantilu $\widehat{q}_\alpha$ je konzistentným odhadom q_α a tým pádom môžeme očakávať asymptotické správanie predikčných intervalov.

4. Simulačná štúdia

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

1. Vygenerovanie pozorovania 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 1000 nezávislých Monte Carlo simulácií, získame empirické pokrytie podielom simulácií ktorých výsledok bol 1 a všetkých simulácií. Taktiež, počas simulácií budeme uvažovať 4 premenlivé faktory

1. Rozsah náhodného výberu $n \in \{20,100,200\}$.
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 intervaly 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 (3.4). Ďalej si zhrnieme aké typy frekventistických predikčných intervalov budeme konštruovať.

Poznamenáme, že frekventistické asymptotické predikčné intervaly dostávame analogickým postupom k príkladu 3 a preto budeme uvádzať iba predpokladanú parametrickú rodinu rozdelení, asymptoticky pivotálnu štatistiku a výsledný predikčný interval.

V prípade normovaného normálneho rozdelenia budeme predpokladať parametrickú rodinu (2.3) a presný frekventistický interval v tvare (2.5). Pri konštrukcií asymptotického predikčného intervalu opäť predpokladáme parametrickú rodinu (2.3). Asymptoticky pivotálna štatistika má v tomto prípade rozdelenie

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

kde $\overline{X}_n$ je výberový priemer spočítaný z náhodného výberu $\mathcal{X}_n$ a S_n^2 je výberový rozptyl spočítaný z náhodného výberu $\mathcal{X}_n$. Následne asymptotický predikčný interval má tvar

$$D_n(\mathcal{X}_n,\alpha) = \left(q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}\right),$$

kde $q_{\frac{\alpha}{2}}$ a $q_{1-\frac{\alpha}{2}}$ sú porade $(\frac{\alpha}{2})$ a $(1-\frac{\alpha}{2})$ kvantil normálneho rozdelenia s parametrami $\overline{\mathcal{X}}_n$ a S_n^2.

Druhým rozdelením exponenciálneho rozdelenia s parametrom $\lambda = 4$ budeme konštruovať iba asymptotický predikčný interval a budeme predpokladať parametrickú rodinu (2.6) a asymptotický predikčný interval (2.7)

Tretím rozdelením ktoré budeme uvažovať je Paretovo rozdelenie s parametrami $\alpha = 3$ a $\beta = 2$, pre ktoré budeme konštruovať iba asymptotický predikčný interval a uvažujeme parametrickú rodinu

$$\{Pareto(\alpha,\beta), (\alpha,\beta)^\top \in ((0,\infty) \times (0,\infty))\}.$$

Následne dostávame asymptoticky pivotálnu štatistiku

$$\widehat{W}_n \sim Pareto\left(\widehat{\lambda}_{MLE}, \widehat{\beta}_{MLE}\right),$$

kde

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

a jedná sa o maximálne vierohodné odhady parametrov λ a β. Výsledný predikčný interval je v tvare

$$D_n(\mathcal{X}_n,\alpha) = \left(q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}\right),$$

kde pre dané $\alpha \in (0,1)$ sú $q_{\frac{\alpha}{2}}$ a $q_{1-\frac{\alpha}{2}}$ sú porade $(\frac{\alpha}{2})$ kvantil a $(1-\frac{\alpha}{2})$ kvantil Paretovho rozdelenia s parametrami $\widehat{\lambda}_{MLE}$ a $\widehat{\beta}_{MLE}$.

Posledným rozdelením je Cauchyho rozdelenie s parametrami $a = 1$ a $b = 1$ pre ktoré budeme konštruovať asymptotický predikčný interval. O rozdelení budeme však predpokladať, že už predom poznáme parameter b. Uvažujeme teda parametrickú rodinu rozdelení

$$\{Cauchy(a,1), a \in \mathbb{R}\}.$$

Následne dostávame asymptoticky pivotálnu štatistiku

$$\widehat{W}_n \sim Cauchy(\widehat{a}_{MLE}, 1),$$

kde $\widehat{a}_{MLE}$ je maximálne vierohodný odhad parametru a, ktorý získame iteratívne, Newton-Rhapsonovým algoritmom z rovnice

$$2\sum_{i=1}^{n} \frac{X_i - \widehat{a}_{MLE}}{1 + (X_i - \widehat{a}_{MLE})^2} = 0.$$

Výsledný predikčný interval je v tvare

$$D_n(\mathcal{X}_n,\alpha) = \left(q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}\right),$$

kde pre dané $\alpha \in (0,1)$ sú $q_{\frac{\alpha}{2}}$ a $q_{1-\frac{\alpha}{2}}$ sú porade $(\frac{\alpha}{2})$ kvantil a $(1 - \frac{\alpha}{2})$ kvantil Cauchyho rozdelenia s parametrami $\widehat{a}_{MLE}$ a $b = 1$.

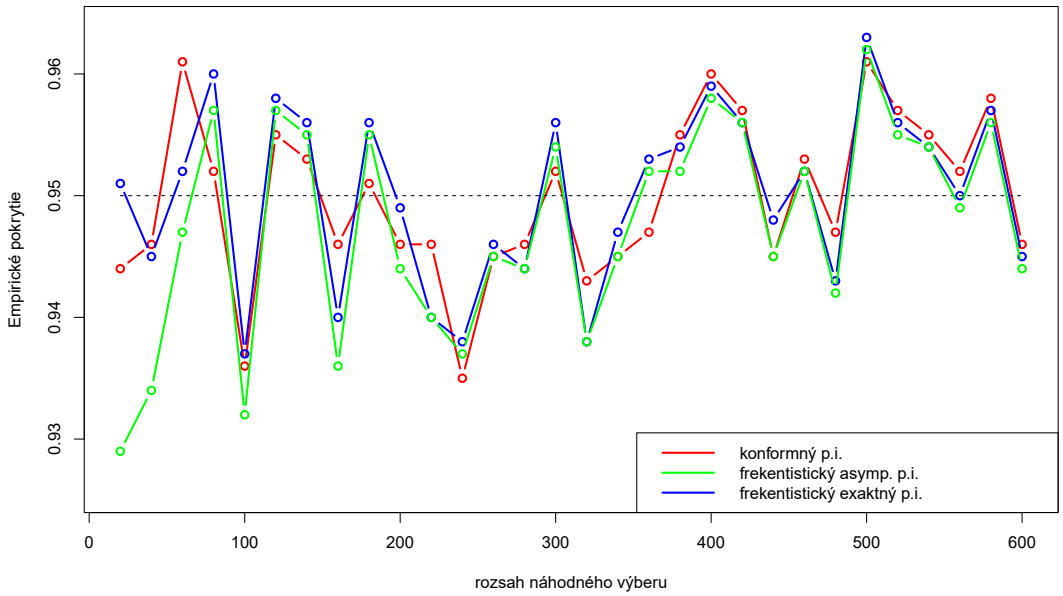
Po odvodení všetkých typov predikčných intervalov v závislosti na rozdelení zhrnieme výsledky simulačnej štúdie nasledujúcej tabuľke.

Rozdelenie	Rozsah	Frekv. ex.		Frekv. as.		Konf ex.	
		90%	95%	90%	95%	90%	95%
$N(0,1)$	$n = 20$	0.909	0.955	0.879	0.936	0.847	0.944
	$n = 100$	0.903	0.947	0.916	0.960	0.872	0.938
	$n = 200$	0.904	0.956	0.899	0.944	0.880	0.939
$Exp(4)$	$n = 20$			0.896	0.944	0.873	0.963
	$n = 100$			0.905	0.947	0.883	0.948
	$n = 200$			0.895	0.951	0.903	0.945
$Pareto(3,2)$	$n = 20$			0.845	0.892	0.861	0.954
	$n = 100$			0.885	0.943	0.898	0.958
	$n = 200$			0.906	0.955	0.896	0.951
$Cauchy(1,1)$	$n = 20$			0.897	0.949	0.844	0.953
	$n = 100$			0.900	0.939	0.898	0.958
	$n = 200$			0.880	0.940	0.890	0.949

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 stanovených úrovní až na jediný prípad, ktorý je však len 0.003% pod očakávanú úroveň. V prípade frekventistických asymptotických intervalov môžeme vidieť očakávané asymptotické správanie empirického pokrytia, ktoré je najlepšie viditeľné v prípade Paretovho rozdelenia. V prípade konformných predikčných intervalov vidíme, že v pre teoretické pokrytie 90% predikčných intervalov môžeme pozorovať predpokladané asymptotické správanie až na jedinú výnimku. V prípade 95% teoretického pokrytia vidíme, že intervaly vo väčšine prípadov spĺňajú požadované teoretické pokrytie.

Rozdielny náhľad na empirické pokrytie predikčných intervalov nám poskytuje obrázok 4.1, kde môžeme pozorovať vývoj empirického pokrytia obojstranných frekventistických exaktných, frekventistických asymptotických a konformných predikčných intervalov, kde uvažujeme normované normálne rozdelenia a rozsah náhodného výberu $n = \{20,40,60,\ldots,600\}$.



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 rozdelenia a $\alpha = 0.05$

Na obrázku 4.1 môžeme pozorovať že

- Empirické pokrytie frekventistického asymptotického predikčného intervalu pomerne jasne kopíruje empirické pokrytie frekventistického exaktného predikčného intervalu, už od rozsahu náhodného výberu $n = 100$
- Od rozsahu náhodného výberu $n = 380$ je empirické pokrytie frekventistických predikčných intervalov takmer identické a konformný predikčný interval vykazuje mierne lepšie výsledky

Záver

Cieľom bakalárskej práce bolo na začiatok pripomenúť štandardné definície a vlastnosti z oblasti pravdepodobnosti a štatistiky a taktiež zoznámiť čitateľa s predikčnými intervalmi. Ďalej sme definovali motivačný príklad na ktorom sme postupne v kapitolách 2 a 3 ilustrovali a odvádzali frekventistické a konformné predikčné intervaly, čím sme sa oboznámili aj s podkladovou teóriou.

Následne sme v poslednej kapitole uskutočnili malú simulačnú štúdiu v programovacom jazyku ***R*** (Team Development Core, 2023), ktorej účelom bolo porovnanie empirického pokrytia konformných a frekventistických predikčných intervalov a výsledky sme zhrnuli v tabuľke 4 a následne sme poskytli rozdielny náhľad na vývoj empirického pokrytia predikčných intervalov prostredníctvom obrázku 4.1.

Literatúra

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.

Zoznam obrázkov

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 $\alpha = 0.05$. 11

2.2 Pozorovaný náhodný výber $\mathcal{X}_{50}$ z exponenciálneho rozdelenia s parametrom $\lambda = 4$ a grafické znázornenie obojstranného a ľavo stranného asymptotického predikčného intervalu pre $Y = X_{51}$ a daným $\alpha = 0.05$. 13

3.1 Pozorovaný usporiadaný náhodný výber $\mathcal{X}_{50}$ a grafické znázornenie obojstranného a ľavo stranného 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 16

4.1 Porovnanie empirického pokrytia predikčných intervalov vzhľadom na meniacu sa dĺžku rozsahu náhodného výberu $n = \{20,40,60,\ldots,600\}$ pre normované normálne rozdelenia a $\alpha = 0.05$ 21

Zoznam tabuliek

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. 20