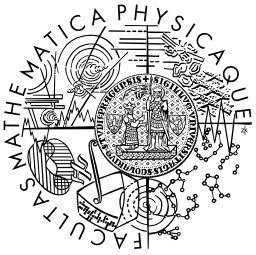


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Marek Bedřich

Toleranční intervaly

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D.
Studijní program: Obecná matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

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

V ..... dne ....

Podpis autora

Hlavní poděkování patří doc. Ing. Markovi Omelkovi, Ph.D., za pohotové reakce, pomoc při řešení všech problémů a celkově výborné vedení práce. Dále bych chtěl poděkovat Evě Březinové a Rozálii Kluvancové za příjemné hodiny strávené společným psaním.

Název práce: Toleranční intervaly

Autor: Marek Bedřich

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato bakalářská práce se zabývá tolerančními intervaly, statistickým nástrojem sloužícím ke kvantifikaci nejistoty predikce. V úvodní části textu jsou krátce připomenuty konfidenční intervaly. Následně se práce zaměřuje na predikční intervaly, které jsou mezikrokem mezi intervaly konfidenčními a tolerančními. Konkrétně je rozebrán predikční interval pro normální rozdělení a neparametrický predikční interval. Hlavní částí práce jsou pak toleranční intervaly - je rozebírána definice, konstrukce parametrických i neparametrických tolerančních intervalů, konvergence či skutečné pokrytí odvozovaných intervalů. V závěrečné části pak najdeme příklad použití tohoto nástroje v praxi.

Klíčová slova: toleranční intervaly, predikční intervaly, intervaly spolehlivosti, kvantifikace nejistoty predikce

Title: Tolerance limits

Author: Marek Bedřich

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This bachelor's thesis deals with tolerance intervals, a statistical tool used to quantify the uncertainty of statistical predictions. The introductory part of the text briefly recalls confidence intervals. The thesis then focuses on prediction intervals, which are an intermediate step between confidence intervals and tolerance intervals. Specifically, the prediction interval for normal distribution and nonparametric prediction interval are analyzed. The main part of the thesis then deals with tolerance intervals - the definition, construction of both parametric and nonparametric tolerance intervals, convergence, and actual coverage of the derived intervals. In the final part, an example of the practical application of this tool is presented.

Keywords: tolerance intervals, prediction intervals, confidence intervals, quantification of prediction uncertainty

Obsah

Úvod

Tato bakalářská práce slouží k seznámení čtenáře s tolerančními intervaly a s jejich jednodušší verzí, tj. s predikčními intervaly. Ty se staly v poslední době předmětem zájmu v oboru strojového učení, protože umožňují kvantifikovat nejistotu predikce. Přesto je problematika tolerančních intervalů spíše na okraji zájmu.

Text je určený jak pro čtenáře, kteří se s konceptem tolerančního intervalu ještě nesetkali, tak pro ty, kteří jsou již s nástrojem seznámeni a chtějí si prohloubit vědomosti. Práce je psána tak, aby byla kompletně pochopitelná pro studenty se základními znalostmi pravděpodobnosti a matematické statistiky.

Práce nejčastěji čerpá z knihy Krishnamoorthy a Mathew (2009), a jelikož je tato kniha určena spíše pro zkušenější čtenáře, práce často doplňuje informace, které v knize chybí. Zároveň se práce zaměřuje na hlubší pochopení konceptů, formálnější přístup a sjednocení značení s dalšími zdroji.

Práce je rozdělena na dvě kapitoly. V první části první kapitoly si čtenář připomene konfidenční intervaly, hlavně definici a příklad sestrojení konfidenčního intervalu pro normální rozdělení. V druhé části první kapitoly se práce věnuje predikčním intervalům - mezikrokem mezi konfidenčními a tolerančními intervaly. Kromě definice a debaty o použitelnosti se zde též odvozuje predikční interval pro normální rozdělení a neparametrický predikční interval.

Druhá kapitola patří samotným tolerančním intervalům. Důraz je už od začátku kladen na pochopení definice a na rozdíl mezi intervalem predikčním a tolerančním. Následně se práce věnuje dvěma způsobům sestrojení tolerančního intervalu pro normální rozdělení a vztahu k intervalu konfidenčnímu a predikčnímu. Dále v práci najdeme pasáž o neparametrickém tolerančním intervalu. V poslední části se pak text věnuje tolerančnímu intervalu pro exponenciální rozdělení s dvěma parametry, který se na konci využije v praktickém příkladu použití tohoto nástroje v reálném světě.

1. Konfidenční a predikční intervaly

1.1 Konfidenční interval

Konfidenční intervaly slouží ke kvantifikaci nejistoty odhadu nějakého neznámého parametru rozdělení, jako například střední hodnoty, rozptylu, či libovolného kvantilu. Cílem je najít takový interval, který s předem danou mírou spolehlivosti $1-\alpha$ pokrývá (alespoň asymptoticky) skutečnou hodnotu neznámého parametru. Pro pozdější porovnání si zopakujeme definici.

Definice 1. *Buď $X_1,...,X_n$ náhodný výběr z rozdělení F_X a $\theta_X = t(F_X) \in \mathbb{R}$ jeho neznámý parametr. Řekneme, že $I_n = I_n(X_1,...,X_n) \subset \mathbb{R}$ je přesný konfidenční interval pro parametr θ_X o spolehlivosti $1-\alpha$, pokud platí*

$$\mathrm{P}\left[I_n(X_1,...,X_n) \ni \theta_X\right] = 1-\alpha.$$

Řekneme, že $I_n = I_n(X_1,...,X_n) \subset \mathbb{R}$ je asymptotický konfidenční interval pro parametr θ_X o (asymptotické) spolehlivosti $1-\alpha$, pokud platí

$$\mathrm{P}\left[I_n(X_1,...,X_n) \ni \theta_X\right] \xrightarrow[n\to\infty]{} 1-\alpha.$$

Jako příklad si ukážeme, jak zkonstruovat přesný oboustranný konfidenční interval pro střední hodnotu normálního rozdělení.

Příklad 1. Buď $X_1,...,X_n$ náhodný výběr z rozdělení F_X z modelu

$$\mathcal{F} = \{N(\mu,\sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$$

normálních rozdělení s konečnou střední hodnotou a konečným (nenulovým) rozptylem. Parametr, který budeme odhadovat, je střední hodnota $\mu_X = \mathsf{E}X_1$. Nejprve definujme náhodnou veličinu

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu_X)}{S_n},$$

kde S_n je odmocnina z výběrového rozptylu S_n^2 a $\overline{X}_n$ je výběrový průměr z náhodného výběru $X_1,...,X_n$. Je známo, že T_n má pak Studentovo t-rozdělení s $n-1$ stupni volnosti, tedy $T_n \sim t_{n-1}$. Označíme-li nyní $t_{n-1}(\alpha)$ α-kvantil Studentova t-rozdělení s $n-1$ stupni volnosti, pak platí

$$\mathrm{P}\left[t_{n-1}\left(\tfrac{\alpha}{2}\right) < \frac{\sqrt{n}(\overline{X}_n - \mu_X)}{S_n} < t_{n-1}\left(1-\tfrac{\alpha}{2}\right)\right] = 1-\alpha.$$

Jednoduchými úpravami pak dostaneme

$$\mathrm{P}\left[\overline{X}_n - t_{n-1}\left(1-\tfrac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}} < \mu_X < \overline{X}_n + t_{n-1}\left(1-\tfrac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}}\right] = 1-\alpha$$

a získáváme přesný oboustranný konfidenční interval pro μ_X o spolehlivosti $1-\alpha$:

$$\left(\overline{X}_n - t_{n-1}\left(1-\tfrac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}},\ \overline{X}_n + t_{n-1}\left(1-\tfrac{\alpha}{2}\right)\frac{S_n}{\sqrt{n}}\right). \tag{1.1}$$

△

Konfidenční interval je silným nástrojem pro odhady parametrů rozdělení, nelze jej ale použít na odhadování budoucích napozorovaných hodnot. Pro tento specifický případ se používají tzv. predikční intervaly. Těm se budeme věnovat nyní.

1.2 Predikční interval

Buď $X_1,...,X_n$ náhodný výběr z rozdělení F_X. V této podkapitole se budeme snažit najít takový interval, který pokryje příští naměřenou hodnotu X_{n+1} s předem danou pravděpodobností $1-\alpha$. Takový interval, který v jistém smyslu predikuje příští naměřenou hodnotu, nazveme predikční.

Definice 2. *Buď $X_1,...,X_n$ a X_{n+1} náhodný výběr z rozdělení F_X. Řekneme, že $I_n = I_n(X_1,...,X_n) \subset \mathbb{R}$ je přesný predikční interval pro X_{n+1} o spolehlivosti $1-\alpha$, pokud platí*

$$\mathrm{P}\left[X_{n+1} \in I_n(X_1,...,X_n)\right] = 1-\alpha.$$

Řekneme, že $I_n = I_n(X_1,...,X_n) \subset \mathbb{R}$ je asymptotický predikční interval pro X_{n+1} o (asymptotické) spolehlivosti $1-\alpha$, pokud platí

$$\mathrm{P}\left[X_{n+1} \in I_n(X_1,...,X_n)\right] \xrightarrow[n\to\infty]{} 1-\alpha.$$

Ne vždy má sestrojení predikčního intervalu smysl - např. pro alternativní rozdělení s neznámou pravděpodobností úspěchu p_X nedává konstrukce predikčního intervalu smysl. Důvodem je, že pravděpodobnost pokrytí musí být nutně z množiny $\{0, 1-p_X, p_X, 1\}$, podle toho zda interval pokrývá hodnoty 0 a 1. My přitom skutečnou hodnotu p_X ani neznáme.

Obecně, podobně jako u konfidenčního intervalu, pro diskrétní náhodné veličiny jde málokdy sestrojit predikční interval o přesné spolehlivosti $1-\alpha$. V tom případě hledáme zpravidla nejužší interval I_n splňující

$$\mathrm{P}\left[X_{n+1} \in I_n\right] \geq 1-\alpha.$$

Extrémním případem tohoto fenoménu by mohla být např. konstrukce predikčního intervalu pro číslo, které padne při hodu spravedlivou hrací kostkou. Pokud bychom požadovali hladinu spolehlivosti $1-\alpha > 5/6$, musel by interval nutně pokrývat všechny možné hodnoty 1,...,6. Skutečná hladina spolehlivosti by pak byla nutně $1-\alpha = 1$. V takových případech, kde je skutečná hladina spolehlivosti vyšší než požadovaná, říkáme, že interval je konzervativní.

My si jako první konstrukční příklad ukážeme sestrojení predikčního intervalu pro normální rozdělení s neznámým rozptylem. K tomu budeme potřebovat následující tvrzení.

Tvrzení 1. *Buď $X_1,...,X_n$ a X_{n+1} náhodný výběr z rozdělení F_X z modelu*

$$\mathcal{F} = \{N(\mu,\sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$$

normálních rozdělení s konečnou střední hodnotou a konečným rozptylem. Pak náhodná veličina

$$T_n = \sqrt{\frac{n}{n+1}}\frac{X_{n+1} - \overline{X}_n}{S_n},$$

kde $\overline{X}_n$ a S_n počítáme pouze z $X_1,...,X_n$, má Studentovo t-rozdělení s $n-1$ stupni volnosti, tedy $T_n \sim t_{n-1}$.

Důkaz. Začneme upravením vzorce pro T_n.

$$T_n = \sqrt{\frac{n}{n+1}}\frac{X_{n+1}-\overline{X}_n}{S_n} = \sqrt{\frac{n}{n+1}}\frac{X_{n+1}-\mu+\mu-\overline{X}_n}{\frac{S_n\sqrt{n-1}}{\sqrt{n-1}}} =$$

$$= \sqrt{\frac{n}{n+1}}\frac{\frac{X_{n+1}-\mu}{\sigma}+\frac{\mu-\overline{X}_n}{\sigma}}{\frac{S_n\sqrt{n-1}}{\sigma\sqrt{n-1}}} = \frac{\sqrt{\frac{n}{n+1}}\left(\frac{X_{n+1}-\mu}{\sigma}+\frac{\mu-\overline{X}_n}{\sigma}\right)}{\sqrt{\frac{\frac{S_n^2(n-1)}{\sigma^2}}{(n-1)}}}$$

Nyní označme

$$U_1 = \frac{X_{n+1}-\mu}{\sigma},\ U_2 = \frac{\mu-\overline{X}_n}{\sigma} \text{ a } Z = \frac{S_n^2(n-1)}{\sigma^2}.$$

Protože X_{n+1} má rozdělení $N(\mu,\sigma^2)$, pak triviálně $U_1 \sim N(0,1)$. Podobně, jelikož $X_1,...,X_n$ je náhodný výběr z rozdělení $N(\mu,\sigma^2)$, tak

$$\sqrt{n}\,U_2 = \sqrt{n}\,\frac{\mu-\overline{X}_n}{\sigma} \sim N(0,1), \text{ neboli } U_2 \sim N\left(0,\tfrac{1}{n}\right).$$

Dále, protože jsou U_1 a U_2 nezávislé (z nezávislosti X_{n+1} a $X_1,...,X_n$; resp. X_{n+1} a $\overline{X}_n$), platí

$$U_1+U_2 \sim N\left(0,1+\tfrac{1}{n}\right) = N\left(0,\tfrac{n+1}{n}\right), \text{ neboli } \sqrt{\tfrac{n}{n+1}}\,(U_1+U_2) \sim N(0,1).$$

Nyní se zabývejme n.v. Z. Jelikož $X_1,...,X_n$ je náhodný výběr z normálního rozdělení, tak z vlastností výběrového rozptylu víme, že

$$Z = \frac{S_n^2(n-1)}{\sigma^2} \sim \chi^2_{n-1},$$

neboli Z má chí-kvadrát rozdělení s $n-1$ stupni volnosti. Označíme si pro přehlednost $U = \sqrt{\frac{n}{n+1}}\,(U_1+U_2)$ a můžeme psát

$$T_n = \frac{U}{\sqrt{\frac{Z}{n-1}}},$$

kde $U \sim N(0,1)$ a $Z \sim \chi^2_{n-1}$ jsou vzájemně nezávislé. Díky normalitě X totiž máme nezávislost $\overline{X}_n$ a S_n^2. Nezávislost X_{n+1} a S_n^2, resp. X_{n+1} a $\overline{X}_n$ pak plyne z nezávislosti X_{n+1} a $X_1,...,X_n$. Konečně, z definice Studentova t-rozdělení plyne $T_n \sim t_{n-1}$.

□

Nyní již máme všechny potřebné nástroje pro sestrojení predikčního intervalu normálního rozdělení s neznámými parametry.

Příklad 2. Uvažujme náhodnou veličinu $X \sim N(\mu,\sigma^2)$ s normálním rozdělením s neznámými parametry. Nechť $X_1,...,X_n$ je náhodný výběr z rozdělení F_X. Zkonstruujeme predikční interval pro X_{n+1}.

Vyjdeme z nově nabyté znalosti rozdělení n.v. T_n z minulého tvrzení. Náhodná veličina X má normální rozdělení a předpoklady tvrzení jsou tedy splněny. Tvrzení nám říká, že

$$T_n = \sqrt{\frac{n}{n+1}}\frac{X_{n+1}-\overline{X}_n}{S_n} \sim t_{n-1}.$$

Označme $t_{n-1}(\alpha)$ α-kvantil rozdělení t_{n-1}. Potom platí

$$\mathrm{P}\left[t_{n-1}\left(\tfrac{\alpha}{2}\right) < \sqrt{\frac{n}{n+1}}\frac{X_{n+1}-\overline{X}_n}{S_n} < t_{n-1}\left(1-\tfrac{\alpha}{2}\right)\right] = 1-\alpha$$

a osamocením X_{n+1} dostáváme

$$\mathrm{P}\left[\overline{X}_n - t_{n-1}\left(1-\tfrac{\alpha}{2}\right)S_n\sqrt{\tfrac{n+1}{n}} < X_{n+1} < \overline{X}_n + t_{n-1}\left(1-\tfrac{\alpha}{2}\right)S_n\sqrt{\tfrac{n+1}{n}}\right] = 1-\alpha,$$

kde jsme využili faktu, že $t_{n-1}\left(\tfrac{\alpha}{2}\right) = -t_{n-1}\left(1-\tfrac{\alpha}{2}\right)$. Odsud už získáváme oboustranný predikční interval pro X_{n+1} na hladině α:

$$\left(\overline{X}_n - t_{n-1}\left(1-\tfrac{\alpha}{2}\right)S_n\sqrt{\tfrac{n+1}{n}},\ \overline{X}_n + t_{n-1}\left(1-\tfrac{\alpha}{2}\right)S_n\sqrt{\tfrac{n+1}{n}}\right). \tag{1.2}$$

Za zmínku stojí konvergence tohoto intervalu. Víme, že platí

$$\overline{X}_n \xrightarrow[n\to\infty]{} \mu,\ t_{n-1}(1-\tfrac{\alpha}{2}) \xrightarrow[n\to\infty]{} u_{1-\frac{\alpha}{2}},\ S_n \xrightarrow[n\to\infty]{} \sigma,\ \sqrt{\tfrac{n+1}{n}} \xrightarrow[n\to\infty]{} 1,$$

kde u_α značí α-kvantil normovaného normálního rozdělení. Odtud plyne, že interval (1.2) konverguje k intervalu

$$\left(\mu - u_{1-\frac{\alpha}{2}}\,\sigma,\ \mu + u_{1-\frac{\alpha}{2}}\,\sigma\right). \tag{1.3}$$

△

Všimněme si podobnosti predikčního intervalu (1.2) s konfidenčním intervalem pro střední hodnotu normálního rozdělení (1.1):

$$\left(\overline{X}_n - t_{n-1}\left(1-\tfrac{\alpha}{2}\right)S_n\sqrt{\tfrac{1}{n}},\ \overline{X}_n + t_{n-1}\left(1-\tfrac{\alpha}{2}\right)S_n\sqrt{\tfrac{1}{n}}\right).$$

Oba intervaly jsou centrované kolem $\overline{X}_n$ a můžeme nahlédnout, že predikční interval je $\sqrt{n+1}-$krát širší než ten konfidenční. To dává smysl, neboť krajní hodnoty predikčního intervalu nekonvergují ke stejné hodnotě, nýbrž ke kvantilům skutečného rozdělení F_X.

Ukázali jsme si příklad, kde je příslušné rozdělení F_X z parametrického modelu, neboli známe jeho tvar až na konečně mnoho parametrů. Takovýto luxus si však v reálném světě nemůžeme vždy dovolit. Pokud ovšem máme k dispozici náhodný výběr dostatečně velkého rozsahu, můžeme zkonstruovat smysluplný predikční interval i pro kompletně neznámé rozdělení F_X.

Příklad 3. Buď $X_1,...,X_n$ a X_{n+1} náhodný výběr z rozdělení F_X se spojitou distribuční funkcí. Najdeme predikční interval pro X_{n+1} o spolehlivosti $1-\alpha$. Budeme k tomu používat pořádkové statistiky a jejich známé vlastnosti. Označme tedy $X_{(i)}$, $i \in \{1,...,n+1\}$, i-tou pořadovou statistiku $X_1,...,X_{n+1}$ a dodefinujme $X_{(0)} := -\infty$ a $X_{(n+2)} := \infty$. Pak (s pravděpodobností 1 - díky spojitosti) platí

$$-\infty = X_{(0)} < X_{(1)} < ... < X_{(n+1)} < X_{(n+2)} = \infty. \tag{1.4}$$

Hledáme největší k_L, resp. nejmenší $k_U \in \{1,...,n+1\}$ splňující

$$\mathrm{P}\left[X_{n+1} < X_{(k_L)}\right] \le \tfrac{\alpha}{2}\ \&\ \mathrm{P}\left[X_{(k_U)} < X_{n+1}\right] \le \tfrac{\alpha}{2},$$

pak totiž platí

$$\mathrm{P}\left[X_{(k_L-1)} < X_{n+1} < X_{(k_U+1)}\right] \geq 1-\alpha. \tag{1.5}$$

Označme R_{n+1} pořadí X_{n+1} v uspořádaném výběru (1.4). Protože je $X_1,...,X_{n+1}$ náhodný výběr, nabývá R_{n+1} každé z hodnot z množiny $\{1,...,n+1\}$ se stejnou pravděpodobností $\frac{1}{n+1}$. Odtud dostáváme, že $\forall i \in \{1,...,n+1\}$ platí

$$\mathrm{P}\left[X_{n+1} < X_{(i)}\right] = \mathrm{P}\left[\,R_{n+1} < i\,\right] = \frac{i-1}{n+1}, \text{ resp. } \mathrm{P}\left[X_{(i)} < X_{n+1}\right] = \frac{n+1-i}{n+1}.$$

Hledáme tedy největší takové k_L, resp. nejnižší takové k_U splňující

$$\frac{k_L-1}{n+1} \leq \frac{\alpha}{2}, \text{ resp. } \frac{n+1-k_U}{n+1} \leq \frac{\alpha}{2}. \tag{1.6}$$

Úpravou nerovnic získáváme

$$k_L \leq 1+\frac{\alpha(n+1)}{2}, \text{ resp. } k_U \geq n+1-\frac{\alpha(n+1)}{2}$$

a můžeme psát

$$\mathrm{P}\left[X_{n+1} < X_{\left(\left\lfloor 1+\frac{\alpha(n+1)}{2}\right\rfloor\right)}\right] \leq \frac{\alpha}{2}, \text{ resp. } \mathrm{P}\left[X_{\left(\left\lceil n+1-\frac{\alpha(n+1)}{2}\right\rceil\right)} < X_{n+1}\right] \leq \frac{\alpha}{2}.$$

Odtud využitím (1.5) získáváme

$$\mathrm{P}\left[X_{\left(\left\lfloor \frac{\alpha(n+1)}{2}\right\rfloor\right)} < X_{n+1} < X_{\left(\left\lceil n+2-\frac{\alpha(n+1)}{2}\right\rceil\right)}\right] \geq 1-\alpha.$$

Našli jsme tedy predikční interval

$$\left(X_{\left(\left\lfloor \frac{\alpha(n+1)}{2}\right\rfloor\right)},\, X_{\left(\left\lceil n+2-\frac{\alpha(n+1)}{2}\right\rceil\right)}\right), \tag{1.7}$$

jehož hladina spolehlivosti je alespoň $1-\alpha$. △

Poznámka. Všimněme si, že platí

$$\frac{1}{n}\left\lfloor\frac{\alpha(n+1)}{2}\right\rfloor \xrightarrow[n\to\infty]{} \frac{\alpha}{2}, \text{ resp. } \frac{1}{n}\left\lceil n+2-\frac{\alpha(n+1)}{2}\right\rceil \xrightarrow[n\to\infty]{} 1-\frac{\alpha}{2}.$$

Lze dokázat, podobně jako v důkazu konzistence výběrových kvantilů, že platí

$$X_{\left(\left\lfloor \frac{\alpha(n+1)}{2}\right\rfloor\right)} \xrightarrow[n\to\infty]{\mathrm{P}} F_X^{-1}\left(\tfrac{\alpha}{2}\right), \text{ resp. } X_{\left(\left\lceil n+2-\frac{\alpha(n+1)}{2}\right\rceil\right)} \xrightarrow[n\to\infty]{\mathrm{P}} F_X^{-1}\left(1-\tfrac{\alpha}{2}\right),$$

kde $F_X^{-1}(\beta)$ značí β-kvantil rozdělení F_X. Platí tedy

$$\mathrm{P}\left[X_{\left(\left\lfloor \frac{\alpha(n+1)}{2}\right\rfloor\right)} < X_{n+1} < X_{\left(\left\lceil n+2-\frac{\alpha(n+1)}{2}\right\rceil\right)}\right] \xrightarrow[n\to\infty]{} 1-\alpha$$

a predikční interval (1.7) je tím pádem asymptotický o spolehlivosti $1-\alpha$. Zároveň má pro menší hodnoty n spolehlivost alespoň požadované $1-\alpha$. Je tedy konzervativní a zároveň asymptotický.

Interval má sice asymptotickou spolehlivost $1-\alpha$, konverguje k ní však oproti parametrickým modelům velmi pomalu. Všimněme si, že už jen pro konstrukci tohoto intervalu potřebujeme velké množství pozorování - aby interval nevyšel $(-\infty, \infty)$, potřebujeme

$$X_{\left(\left\lfloor \frac{\alpha(n+1)}{2} \right\rfloor\right)} > -\infty \iff \frac{\alpha(n+1)}{2} \geq 1 \iff n \geq \frac{2}{\alpha} - 1,$$

tedy např. pro $1-\alpha = 0.95$ potřebujeme $n \geq 39$. Díky asymptotické přesnosti bude však interval pro dostatečně vysoké hodnoty n podobný intervalům vycházejícím z parametrických modelů.

Je důležité si uvědomit, že predikční interval na hladině 95 % nám neříká, že 95 % následujících pozorování padne do daného intervalu. Predikční interval lze sice rozšířit, aby pokrýval $k \in \mathbb{N}$ následujících pozorování (Otava, 2017, str. 13), pokud ale hledáme nástroj pro předem nespecifikované velké množství následujících pozorování, použijeme toleranční interval.

2. Toleranční interval

2.1 Definice

Buďte $X_1,...,X_n$ a $X_{n+1},X_{n+2},X_{n+3},..$ náhodné výběry z rozdělení F_X a buďte $\alpha, \beta \in (0,1)$. V této kapitole se budeme snažit zkonstruovat na základě náhodného výběru $X_1,...,X_n$ interval, který s pravděpodobností $1-\alpha$ pokryje alespoň $100\beta\%$ náhodných veličin z náhodného výběru $X_{n+1},X_{n+2},X_{n+3},...$ Takový interval budeme nazývat toleranční.

Definice 3. *Nechť* $\mathbf{X} = (X_1,...,X_n)^{\mathrm{T}}$ *je náhodný výběr z rozdělení* F_X *náhodné veličiny* X *a buďte* $\alpha, \beta \in (0,1)$. *Pak řekneme, že* $I_n = I_n(\mathbf{X}) \subset \mathbb{R}$ *je toleranční interval pro* X *s pokrytím* β *(resp.* $100\beta\,\%$*) o spolehlivosti* $1-\alpha$*, pokud platí*

$$\mathrm{P}_{\mathbf{X}}\left[\mathrm{P}_X\left(X \in I_n(\mathbf{X}) \mid \mathbf{X}\right) \geq \beta\right] = 1-\alpha,$$

kde $\mathrm{P}_Y[...]$ *značí, že pravděpodobnost je počítána vzhledem k rozdělení náhodné veličiny* Y.

Standardně volíme $(1-\alpha), \beta \in \{0.90, 0.95, 0.99\}$, podobně jako u predikčních a konfidenčních intervalů. Toleranční intervaly jsou zpravidla mnohem náročnější na konstrukci než predikční a zřídkakdy lze najít přesné analytické řešení.

Na místě je porovnání definic tolerančního a predikčního intervalu. Pro náh. vektor $\mathbf{X} = (X_1,...,X_n)^{\mathrm{T}}$ a náh. veličiny X_{n+1}, X, kde $X_1,...,X_n,X_{n+1},X$ jsou i.i.d., označme

$$p(\mathbf{X}) = \mathrm{P}_X\left[X \in I_n(\mathbf{X}) \mid \mathbf{X}\right] = \mathrm{P}_{X_{n+1}}\left[X_{n+1} \in I_n(\mathbf{X}) \mid \mathbf{X}\right]$$

podmíněnou pravděpodobnost, že náh. veličina X padne do intervalu I_n zkonstruovaného z náhodného výběru $\mathbf{X}$. Pro I_n predikční interval o spolehlivosti $1-\alpha$ platí

$$1-\alpha = \mathrm{P}_{\mathbf{X},X_{n+1}}\left[X_{n+1} \in I_n(\mathbf{X})\right] = \mathsf{E}_{\mathbf{X}}\left(\mathrm{P}_{X_{n+1}}\left[X_{n+1} \in I_n(\mathbf{X}) \mid \mathbf{X}\right]\right) = \mathsf{E}_{\mathbf{X}} p(\mathbf{X}).$$

Zatímco toleranční interval I_n s pokrytím β o spolehlivosti $1-\alpha$ splňuje

$$1-\alpha = \mathrm{P}_{\mathbf{X}}\left[\mathrm{P}_X\left(X \in I_n(\mathbf{X}) \mid \mathbf{X}\right) \geq \beta\right] = \mathrm{P}_{\mathbf{X}}\left[p(\mathbf{X}) \geq \beta\right] = \mathsf{E}_{\mathbf{X}} \mathbb{1}\{p(\mathbf{X}) \geq \beta\}.$$

Hladina spolehlivosti tedy u predikčního intervalu určuje průměrnou hodnotu $p(\mathbf{X})$, zatímco u tolerančního určuje s jakou pravděpodobností dosahuje $p(\mathbf{X})$ hodnoty alespoň β.

2.2 Normální rozdělení

My opět začneme tolerančním intervalem pro normální rozdělení s neznámými parametry. Odvodíme analyticky neřešitelnou integrální rovnici, jejímž numerickým vyřešením se dá získat hledaný toleranční interval, ale ukážeme si i dobrou aproximaci, pomocí které se lze numerickým výpočtům vyhnout.

2.2.1 Pomocné tvrzení

Pro odvození zmíněné integrální rovnice budeme potřebovat následující tvrzení (Krishnamoorthy a Mathew, 2009, str. 7):

Tvrzení 2. *Buďte* $U \sim N(0,\sigma^2)$, $\chi^2 \sim \frac{\chi^2_{n-1}}{n-1}$ *nezávislé náhodné veličiny a buďte* $\alpha,\beta \in (0,1)$. *Označme* Φ *distribuční funkci normovaného normálního rozdělení a* χ *odmocninu z n.v.* χ^2. *Pak konstanta* k, *která splňuje*

$$\mathrm{P}_{U,\chi}\left[\Phi(U+k\chi)-\Phi(U-k\chi)\geq\beta\right]=1-\alpha, \tag{2.1}$$

je řešením integrální rovnice

$$\sqrt{\frac{2}{\pi\sigma^2}}\int_0^\infty \mathrm{P}_{\chi^2}\left[\chi^2>\frac{\chi^2_{1,\beta}(x^2)}{k^2}\right]e^{-\frac{x^2}{2\sigma^2}}\,dx=1-\alpha,$$

kde výrazem $\chi^2_{1,\beta}(x^2)$ *značíme* β*-kvantil rozdělení náhodné veličiny* Y^2, *kde* Y *má rozdělení* $N(x,1)$.[1]

Důkaz. Tento důkaz je inspirovaný důkazem autora (Krishnamoorthy a Mathew, 2009, str. 8). Zkoumejme nejprve následující rovnici: Nechť hodnota U je daná, označme r přesné (nutně kladné) řešení rovnice

$$\Phi(U+r)-\Phi(U-r)=\beta. \tag{2.2}$$

Buď n.v. $Z \sim N(0,1)$ nezávislá s U. Provedeme úpravy:

$$\beta=\Phi(U+r)-\Phi(U-r)=\mathrm{P}_Z(Z\in(U-r,U+r)\mid U)$$

$$=\mathrm{P}_Z(Z-U\in(-r,r)\mid U)=\mathrm{P}_Z((Z-U)^2<r^2\mid U),$$

kde $(Z-U)^2$ má pro fixní U podmíněné rozdělení $(N(-U,1))^2=(N(U,1))^2$. Označíme-li tedy $\chi^2_{1,\beta}(U^2)$ β-kvantil tohoto rozdělení, hledaným řešením (2.2) je $r=\sqrt{\chi^2_{1,\beta}(U^2)}$. Díky tomu, že je funkce

$$\Phi(U+r)-\Phi(U-r)$$

roustoucí v r, už můžeme vidět, že pro fixní U je vztah

$$\Phi(U+k\chi)-\Phi(U-k\chi)\geq\beta$$

ekvivalentní vztahu

$$k\chi\geq r,\ \text{resp. }\chi^2\geq\frac{r^2}{k^2}=\frac{\chi^2_{1,\beta}(U^2)}{k^2}.$$

Nechť tedy k splňuje (2.1). Můžeme psát

$$1-\alpha=\mathrm{P}_{U,\chi}\left[\Phi(U+k\chi)-\Phi(U-k\chi)\geq\beta\right]=\mathrm{P}_{U,\chi}\left[\chi^2\geq\frac{\chi^2_{1,\beta}(U^2)}{k^2}\right]$$

[1] Jedná se o speciální případ necentrálního χ^2 rozdělení s jedním stupněm volnosti a parametrem necentrality x^2.

$$= \mathsf{E}_U \left(\mathrm{P}_\chi \left[\chi^2 \geq \frac{\chi^2_{1,\beta}(U^2)}{k^2} \right] \right) = \int_{-\infty}^{\infty} \mathrm{P}_\chi \left[\chi^2 \geq \frac{\chi^2_{1,\beta}(x^2)}{k^2} \right] f_U(x)\, dx,$$

kde f_U značí hustotu U, tedy hustotu normálního rozdělení $N(0,\sigma^2)$. Předposlední rovnost plyne z nezávislosti U a χ^2. Dosadíme za f_U a získáváme

$$\int_{-\infty}^{\infty} \mathrm{P}_\chi \left[\chi^2 \geq \frac{\chi^2_{1,\beta}(x^2)}{k^2} \right] \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\, dx = \sqrt{\frac{2}{\pi\sigma^2}} \int_0^{\infty} \mathrm{P}_{\chi^2} \left[\chi^2 > \frac{\chi^2_{1,\beta}(x^2)}{k^2} \right] e^{-\frac{x^2}{2\sigma^2}}\, dx,$$

kde jsme využili faktu, že hustota f_U je sudá z definice a pravděpodobnost $\mathrm{P}_\chi[...]$ je zde funkcí x^2, tedy je také sudá. Jejich součin je proto také sudý. Zároveň je integrál konečný, neboť jako majorantu integrandu můžeme zvolit samotnou hustotu f_U, která je z definice integrovatelná. Důkaz je tímto dokončen. □

Nyní tvrzení aplikujeme na sestrojení slíbeného tolerančního intervalu pro normální rozdělení.

2.2.2 Konstrukce tolerančního intervalu

Buď $X \sim N(\mu, \sigma^2)$ náhodná veličina s normálním rozdělením s neznámými parametry. Nechť $X_1,...,X_n$ je náhodný výběr z rozdělení F_X a buďte $\alpha, \beta \in (0,1)$. Chceme zkonstruovat toleranční interval pro X s pokrytím 100β % o spolehlivosti $1-\alpha$ ve tvaru

$$\left(\overline{X}_n - kS_n,\ \overline{X}_n + kS_n \right),$$

kde k je konstanta, kterou budeme hledat. Hledáme tedy takové k, aby bylo splněno

$$\mathrm{P}_{\mathbf{X}} \left[\mathrm{P}_X \left[\overline{X}_n - kS_n < X < \overline{X}_n + kS_n \mid \mathbf{X} \right] \geq \beta \right] = 1 - \alpha. \tag{2.3}$$

Všimněme si, že vnitřní nerovnost můžeme přepsat na

$$\mathrm{P}_X \left[\frac{\overline{X}_n - \mu - kS_n}{\sigma} < \frac{X - \mu}{\sigma} < \frac{\overline{X}_n - \mu + kS_n}{\sigma} \right] \geq \beta, \tag{2.4}$$

což se nám bude hodit, neboť platí

$$\frac{X-\mu}{\sigma} \sim N(0,1),\ \frac{\overline{X}_n - \mu}{\sigma} \sim N\left(0, \tfrac{1}{n}\right) \text{ a } \frac{S_n^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}.$$

Označíme-li si tedy náhodné veličiny

$$\chi^2 \sim \frac{\chi^2_{n-1}}{n-1}, \text{ resp. } U_n \sim N\left(0, \tfrac{1}{n}\right),$$

pak můžeme psát, že (2.4) je ekvivalentní nerovnosti

$$\Phi(U_n + k\chi) - \Phi(U_n - k\chi) \geq \beta,$$

Dosazením do (2.3) dostáváme, že hledané k musí ekvivalentně splňovat

$$\mathrm{P}_{\mathbf{X}} \left[\Phi(U_n + k\chi) - \Phi(U_n - k\chi) \geq \beta \right] = 1 - \alpha.$$

Nyní však použijeme předchozí tvrzení pro $\sigma^2 = 1/n$ a dostáváme, že k je řešením integrální rovnice

$$\sqrt{\frac{2n}{\pi}} \int_0^\infty \mathrm{P}_{\chi^2} \left[\chi^2_{n-1} > \frac{(n-1)\chi^2_{1,\beta}(x^2)}{k^2} \right] e^{-\frac{1}{2}nx^2}\, dx = 1 - \alpha.$$

Jak již bylo však zmíněno, z povahy této rovnice nelze najít její analytické řešení (Borbatc a kol., 2020, str. 2). Integrační rovnice jako tato se však dají řešit numericky s vysokou přesnotí. My místo toho využijeme aproximaci pro k, kterou odvodili Krishnamoorthy a Mathew (2009) na straně 31:

$$k \simeq \sqrt{\frac{(n-1)\chi^2_{1,\beta}\left(\frac{1}{n}\right)}{\chi^2_{n-1,\alpha}}}.$$

Tato aproximace je podle autorů velmi dobrá i pro náhodný výběr tak malého rozsahu jako $n = 3$, pokud α, β splňují $(1-\alpha), \beta \in \{0.9, 0.95, 0.99\}$. Pro $n \geq 10$ je pak rozdíl mezi aproximací a přesnou hodnotou k zpravidla v řádu tisícin. My proto aproximaci použijeme, a dosazením za k získáváme toleranční interval s pokrytím β o spolehlivosti $1-\alpha$ pro normální rozdělení s neznámými parametry:

$$\left(\overline{X}_n - \sqrt{\frac{(n-1)\chi^2_{1,\beta}\left(\frac{1}{n}\right)}{\chi^2_{n-1,\alpha}}} S_n,\ \overline{X}_n + \sqrt{\frac{(n-1)\chi^2_{1,\beta}\left(\frac{1}{n}\right)}{\chi^2_{n-1,\alpha}}} S_n \right). \tag{2.5}$$

U tolerančních intervalů jsme si nedefinovali nic jako asymptotickou přesnost. Přesto ale ukážeme, že tato aproximace konverguje k intervalu I, který splňuje $\mathrm{P}[X \in I] = \beta$. V jistém smyslu tedy aproximace konverguje k tolerančnímu intervalu pro β následujících pozorování o spolehlivosti 1. To je důležité, neboť tuto vlastnost splňuje každý přesný toleranční interval (Otava, 2017, str. 14/15). Pokud bychom tedy chtěli nějak definovat asymptoticky přesný toleranční interval, definice založená na této vlastnosti by byla solidním kandidátem. Ukážeme tedy, že interval (2.5) tuto vlastnost opravdu splňuje. Z vlastností χ^2 rozdělení víme, že platí

$$\frac{\chi^2_{n-1}}{n-1} \xrightarrow[n\to\infty]{\mathrm{P}} 1.$$

Proto i α-kvantil χ^2_{n-1} musí splňovat

$$\frac{\chi^2_{n-1,\alpha}}{n-1} \xrightarrow[n\to\infty]{} 1.$$

Dále platí

$$\chi^2_{1,\beta}\left(\tfrac{1}{n}\right) \xrightarrow[n\to\infty]{} \chi^2_{1,\beta}, \text{ neboť } \left(N\left(\sqrt{\tfrac{1}{n}}, 1\right)\right)^2 \xrightarrow[n\to\infty]{d} (N(0,1))^2 = \chi^2_1,$$

kde využíváme fakt, že konvergence v distribuci implikuje konvergenci kvantilů (Van der Vaart, 2000, Lemma 21.2). Ještě budeme potřebovat převést $\chi^2_{1,\beta}$ na kvantil normovaného normálního rozdělení. Jelikož je χ^2-rozdělení s jedním stupněm volnosti zároveň rozdělením druhé mocniny normovaného normálního rozdělení, označme náhodnou veličinu $U \sim N(0,1)$, a můžeme psát:

$$1 - \beta = \mathrm{P}\left[U^2 > \chi^2_{1,\beta}\right] = \mathrm{P}\left[U > \sqrt{\chi^2_{1,\beta}}\right] + \mathrm{P}\left[U < -\sqrt{\chi^2_{1,\beta}}\right] = 2 \cdot \mathrm{P}\left[U > \sqrt{\chi^2_{1,\beta}}\right],$$

kde jsme v poslední rovnosti využili symetrie hustoty normovaného normálního rozdělení. Vidíme tedy, že platí

$$\mathrm{P}\left[U > \sqrt{\chi^2_{1,\beta}}\right] = \frac{1-\beta}{2},$$

což lze ekvivalentně zapsat jako

$$\mathrm{P}\left[U < \sqrt{\chi^2_{1,\beta}}\right] = \frac{1+\beta}{2}.$$

Odtud už lze nahlédnout rovnost

$$\sqrt{\chi^2_{1,\beta}} = u_{\frac{1+\beta}{2}},$$

kde u_α značí α-kvantil rozdělení $N(0,1)$. Co se týče konvergence $\overline{X}_n$ a S_n, je známo, že

$$\overline{X}_n \xrightarrow[n\to\infty]{\mathrm{P}} \mathsf{E}X = \mu, \text{ a podobně } S_n \xrightarrow[n\to\infty]{\mathrm{P}} \sigma.$$

Nyní již vidíme, že interval (2.5) konverguje v pravděpodobnosti s roustoucím n k intervalu

$$I = \left(\mu - u_{\frac{1+\beta}{2}}\,\sigma,\ \mu + u_{\frac{1+\beta}{2}}\,\sigma\right). \tag{2.6}$$

Tento interval ale splňuje

$$\mathrm{P}\left[X \in I\right] = \beta,$$

a požadovaná „asymptotická přesnost“ intervalu (2.5) je tím dokázána.

Poznámka. Za zmínku stojí porovnání tolerančního intervalu (2.5) s predikčním intervalem (1.2). Porovnáme-li, k čemu tyto intervaly konvergují, zjistíme, že pro $1-\alpha = \beta$ konvergují oba ke stejnému intervalu. Interval (1.3) je v totiž tomto případě rovný intervalu (2.6). V tabulce můžeme vidět porovnání tolerančních intervalů (2.5) s predikčními intervaly (1.2). Intervaly byly konstruovány z náhodných výběrů z rozdělení $N(0,1)$ o rozsahu n pro $1-\alpha = \beta = 0.95$ generovaných v R.

$n =$	20	200	2000
Predikční	(-2.20, 1.97)	(-1.91, 2.10)	(-1.96, 1.98)
Toleranční	(-2.79, 2.57)	(-2.08, 2.27)	(-2.01, 2.03)

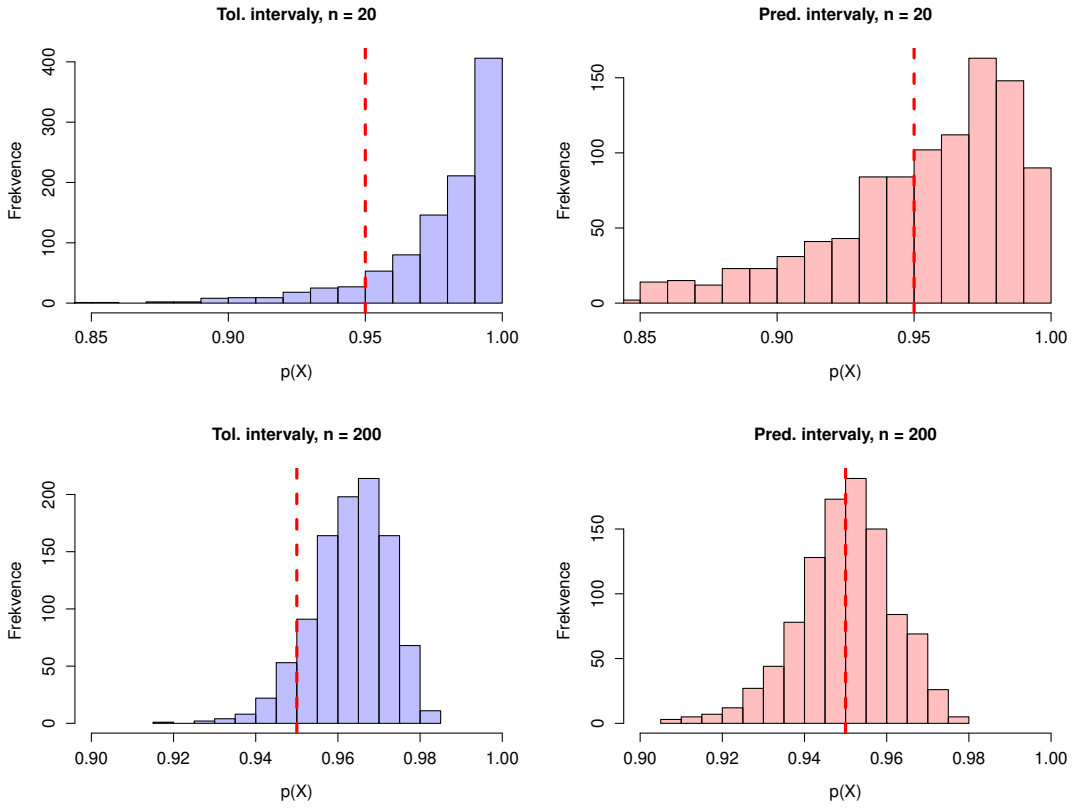
Pro srovnání, interval (2.6), ke kterému v tomto případě intervaly konvergují, je pro $1-\alpha = 0.95$ rovný (-1.96, 1.96). Z tabulky vidíme, že toleranční intervaly vyšly pro tyto hodnoty výrazně širší než predikční.

Jak jsme viděli, konstrukce tolerančního intervalu pro normální rozdělení je podstatně náročnější, než konstrukce intervalu predikčního. Jak již bylo zmíněno, predikční interval lze teoreticky rozšířit, aby pokrýval fixní množství $k \in \mathbb{N}$ následujících pozorování, tolerančním intervalům se ale v mnoha případech nevyhneme. Jedno takové praktické využití tolerančního intervalu si ukážeme na konci kapitoly.

Příklad 4. Uvedeme si nyní numerický příklad pro ilustraci role konstant α, β. Budeme pomocí výpočetního prostředí R (R Core Team, 2023) generovat náhodné výběry $\mathbf{X}$ z rozdělení $N(0,1)$. Na základě těch budeme konstruovat toleranční intervaly s pokrytím $\beta = 0.95$ o spolehlivosti $1 - \alpha = 0.90$ a predikční intervaly o spolehlivosti $\beta = 0.95$. Budeme používat toleranční interval (2.5) a predikční interval (1.2). U každého z těchto náhodných výběrů si následně spočítáme „skutečné pokrytí“ tolerančního i predikčního intervalu:

$$p(\mathbf{X}) = \mathrm{P}_X \left(X \in I(\mathbf{X}) \mid \mathbf{X} \right), \tag{2.7}$$

tj. skutečnou pravděpodobnost, že $X \sim N(0,1)$ padne při daném $\mathbf{X}$ do intervalu I. Z komentáře za definicí tolerančního intervalu pak plyne, že přibližně $1 - \alpha = 90$ % tolerančních intervalů by mělo splňovat $p(\mathbf{X}) \geq \beta = 0.95$, a že pro predikční intervaly by průměrná hodnota $p(\mathbf{X})$ měla být přibližně $\beta = 0.95$. Vygenerujeme právě popsaným procesem 1 000 náhodných výběrů nejprve pro $n = 20$, následně pro $n = 200$, a z naměřených hodnot $p(\mathbf{X})$ vykreslíme histogramy (viz Obrázek 2.1).



Obrázek 2.1: Histogramy naměřených hodnot $p(\mathbf{X})$.

Věnujme se nejprve tolerančním intervalů,. V případě $n = 20$ je vidět, že většina intervalů měla skutečné pokrytí $p(\mathbf{X})$ výrazně vyšší než $\beta = 0.95$, mnoho pokrytí mělo hodnotu dokonce vyšší než 0.99. Zároveň ale 896 z 1000 intervalů mělo skutečné pokrytí vyšší než β, což velmi přesně odpovídá požadované spolehlivosti $1 - \alpha = 0.9$. V případě $n = 200$ pak mělo 910 intervalů skutečné pokrytí vyšší než β, intervaly se skutečným pokrytím vyšším než 0.99 však zmizely. Celkově, skutečné hodnoty pokrytí byly méně rozptýlené.

Nyní k predikčním intervalům. V případech $n = 20$, resp. $n = 200$ byla průměrná hodnota skutečného pokrytí $p(\mathbf{X})$ po zaokrouhlení rovna 0.952, resp. 0.950, což též velmi přesně odpovídá spolehlivosti $\beta = 0.95$. V případě $n = 200$ však byly hodnoty pokrytí opět výrazně méně rozptýlené.

△

2.3 Neparametrický toleranční interval

V podkapitole o predikčních intervalech jsme si ukazovali neparametrický predikční interval, který nevyžadoval kromě spojitosti žádné informace o zkoumaném rozdělení. Podobně lze odvodit i neparametrický toleranční interval. Mějme tedy n pozorování z neznámého spojitého rozdělení a odpovídající uspořádaný náhodný výběr

$$-\infty < X_{(1)} < ... < X_{(n)} < \infty.$$

Toleranční interval pro β budoucích pozorování na hladině $1 - \alpha$ pak bude mít opět formu

$$\left(X_{(i)}, X_{(j)}\right).$$

Guenther (1977) se v kapitole 4 věnuje neparametrickému tolerančnímu intervalu a ukazuje, že i, j musí splňovat nerovnici

$$\sum_{k=0}^{j-i-1} \binom{n}{k} \beta^k (1-\beta)^{n-k} \geq 1 - \alpha.$$

Zároveň ale chceme minimalizovat rozdíl $j - i$, aby interval nebyl zbytečně konzervativní. Hledáme tedy nejmenší takové r splňující

$$\sum_{k=0}^{r} \binom{n}{k} \beta^k (1-\beta)^{n-k} \geq 1 - \alpha$$

a následně volíme dvojici i,j tak, aby splňovala $j - i = r + 1$, a zároveň aby byl interval symetrický, tj. volíme například

$$i = \left\lfloor \frac{n-r}{2} \right\rfloor, \text{ resp. } j = r + 1 + \left\lfloor \frac{n-r}{2} \right\rfloor,$$

kde jako v predikčním případě dodefinujeme

$$X_{(0)} = -\infty, \text{ resp. } X_{(n+1)} = \infty.$$

Podobně jako u predikční verze, toleranční intervaly vycházející z parametrických modelů jsou zde též zpravidla mnohem účinější. Narozdíl od predikčních intervalů je ale počet modelů, pro které lze odvodit parametrický toleranční interval, značně omezený. Teorie tolerančních intervalů se proto zaměřuje hlavně na konstrukci intervalů pro normální rozdělení. Strategie pro nenormální rozdělení je pak často taková, že se naměřené hodnoty transformují, aby měly (alespoň přibližně) normální rozdělení. Například pro lognormálně rozdělenou náhodnou veličinu X má $\log X$ normální rozdělení. Podobně pro n.v. X s gamma rozdělením má $\sqrt[3]{X}$ přibližně normální rozdělení (Krishnamoorthy a Mathew, 2009, kapitola 7.3).

2.4 Dvouparametrické exponenciální rozdělení

My si ukážeme ještě jeden příklad tolerančního intervalu, který se neřeší převedením na normální rozdělení. Ukážeme si jednostranný toleranční interval pro dvouparametrické exponenciální rozdělení. Nejprve si rozdělení zadefinujeme.

Definice 4. *Řekneme, že spojitá náhodná veličina X má dvouparametrické exponenciální rozdělení s parametry $\mu \in \mathbb{R}$, $\theta > 0$ právě tehdy, když má její hustota vzhledem k Lebesgueově míře následující tvar:*

$$f(x;\mu,\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{(x-\mu)}{\theta}}, & x > \mu, \\ 0, & x \leq \mu. \end{cases}$$

Jedná se vlastně o standardní exponenciální rozdělení, pouze posunuté o μ po x-ové ose v kladném směru. Střední hodnota je tudíž posunutá o μ a rozptyl zůstává stejný. Intuitivní pohled na dvouparametrické exponenciální rozdělení je rozdělení životnosti předmětu. Čas μ je „garantovaná“ životnost, a po uplynutí μ životnost podléhá exponenciálnímu rozdělení.

Příklad 5. Buď $X_1,...,X_n$ náhodný výběr z dvouparametrického exponenciálního rozdělení s neznámými parametry $\mu \in \mathbb{R}, \theta > 0$. Ukážeme si toleranční interval s pokrytím β o spolehlivosti $1-\alpha$. Budeme potřebovat odhady parametrů rozdělení. Maximálně věrohodné odhady parametrů μ, resp. θ jsou (např. Johnson a kol. (1994), str. 506):

$$\widehat{\mu} = X_{(1)}, \text{ resp. } \widehat{\theta} = \overline{X}_n - X_{(1)}.$$

Toleranční interval hledáme tentokrát ve tvaru

$$\left(\widehat{\mu} + kn\widehat{\theta},\, +\infty\right)$$

Krishnamoorthy a Mathew (2009) odvodili na straně 186 vzorec pro k, za předpokladu specifické podmínky pro n:

$$k = \frac{1}{n}\left[1 - \left(\frac{\beta^n}{\alpha}\right)^{\frac{1}{n-1}}\right], \text{ pokud } n \leq \frac{\log\alpha}{\log\beta}.$$

V případě splnění této podmínky pro n totiž vyjde $k \leq 0$ (jak můžeme vidět dosazením). Pokud tato podmínka pro n není splněna, lze k najít pouze numerickým řešením integrální rovnice (Krishnamoorthy a Mathew, 2009, str. 185). Dosazením k získáváme toleranční interval

$$\left(\widehat{\mu} + \left[1 - \left(\frac{\beta^n}{\alpha}\right)^{\frac{1}{n-1}}\right]\widehat{\theta},\, +\infty\right), \text{ za podmínky } n \leq \frac{\log\alpha}{\log\beta}. \tag{2.8}$$

Je důležité si uvědomit, že levá hranice integrálu je za této podmínky pro n větší nebo rovna $\overline{X}_n$ a s roustoucím n klesá. Podmínka pro n se může na první pohled zdát velmi restriktivní, ale např. pro jednu z nejběžnějších kombinací hodnot $\alpha = 0.05, \beta = 0.99$ dostáváme

$$n \leq \frac{\log\alpha}{\log\beta} = \frac{\log 0.05}{\log 0.99} \doteq 298.07,$$

což je poměrně vysoká maximální hodnota pro n. △

Nakonec si ukážeme možné využití tolerančního intervalu v praxi.

Příklad 6. Představme si, že jsme v pozici výrobce převodovek do automobilů. Jakožto dodavatel v tomto průmyslu musíme být schopní garantovat určitou úroveň kvality našeho produktu. Potencionální klient je s námi ochotný uzavřít dodavatelský kontrakt za následujících podmínek: Musíme být schopni garantovat, že alespoň 999 z tisíce automobilů najede 25 000 km (záruční nájezd), než dojde k poruše převodovky. V opačném případě se klientovi nevyplatí od nás převodovky za naši cenu nakupovat. Pokud nebude garance naplněna, bude naše firma penalizována. Spočítali jsme si, že pokud si budeme alespoň na 95 % jistí, že zmíněné podmínky splníme, jít do zakázky se nám vyplatí.

Jakožto dlouholetý výrobce převodovek víme, že nájezdy převodovek mají zpravidla dvouparametrické exponenciální rozdělení[2]. My jsme pro testovací účely zkonstruovali 30 prototypů tohoto modelu a provedli zátěžové testy. K dispozici tedy máme data o nájezdu těchto třiceti testovacích převodovek před první poruchou.

Nájezdy testovacích převodovek v km.									
34.93	48.80	47.47	59.32	34.82	56.32	53.66	58.06	45.78	105.58
41.54	51.09	31.52	240.58	69.74	93.89	66.28	49.94	97.54	95.89
36.66	38.06	55.35	34.30	103.01	36.72	49.97	31.07	40.85	82.25

Poznámka. Pro demonstrativní účely používáme simulovaná data z dvouparametrického exponenciálního rozdělení s parametry $\mu = 30$, $\theta = 40$ ve výpočetním prostředí R (R Core Team, 2023).

Naším cílem je zkonstruovat dolní toleranční interval pro dvouexponenciální rozdělení s neznámými parametry s pokrytím 99.9 % o spolehlivosti 95 %. Budeme k tomu chtít využít vzorec (2.8) odvozený v minulém příkladu. Nejprve ale musíme ověřit horní podmínku pro n, abychom zjistili, zda vzorec vůbec můžeme použít. Náš náhodný výběr má velikost $n = 30$, parametry spolehlivosti mají hodnoty $\alpha = 0.05$ a $\beta = 0.999$. Nerovnost

$$30 = n \leq \frac{\log \alpha}{\log \beta} = \frac{\log 0.05}{\log 0.999} \doteq 2994.23$$

je tedy s přehledem splněna. Dále si z našeho náhodného výběru spočítáme hodnoty

$$\widehat{\mu} = 31.07, \text{ resp. } \widehat{\theta} = 31.96,$$

a můžeme dosadit do vzorce. Vyjde nám interval

$$I = \left(31.07 - \left[1 - \left(\frac{0.999^{30}}{0.05}\right)^{\frac{1}{29}}\right] 31.96,\ +\infty\right) \doteq (27\,627,\ +\infty)\,.$$

Spodní hranice intervalu vyšla vyšší než požadovaná hranice 25 000. To znamená, že za platnosti předpokladů jsme schopni garantovat, že alespoň 99.9 % převodovek vydrží alespoň 25 000 km s jistotou alespoň 95 %.

△

[2]Tento předpoklad nemusí být v souladu s realitou. Krishnamoorthy a Mathew (2009) ho ale používají pro odhad nájezdu vojenských vozidel, proto ho používáme zde.

Závěr

V práci jsme začali připomenutím konfidenčních intervalů. Následně jsme se věnovali predikčnímu intervalu, kde jsme si ukázali sestrojení neparametrického pred. intervalu a pred. intervalu pro normální rozdělení. V hlavní kapitole o tolerančních intervalech jsme se nejprve zabývali prohloubením chápání definice. Zkonstruovali jsme toleranční interval pro normální, dvouparametrické exponenciální i obecné spojité rozdělení. Mnoho pozornosti bylo též věnováno roli předem daných konstant α, β, n na vlastnosti intervalu. Nakonec jsme předvedli příklad aplikace v reálném světě.

Toleranční intervaly mohou za správných podmínek sloužit jako silný nástroj pro kontrolu nejistoty predikce. Samotné téma tolerančních intervalů není touto prací zdaleka vyčerpáno. V literatuře lze najít mnoho různorodých, leč často velmi specializovaných případů, kdy lze toleranční interval (ač většinou za použití numerických metod) sestrojit. Pro zájemce o více informací o tomto tématu doporučuji knihu Krishnamoorthy a Mathew (2009).

Seznam použité literatury

Borbatc, N., Chistokletov, N. a Shkolina, T. (2020). Computing exact factors for two-sided tolerance limits in a normal distribution with unknown parameters in Matlab. *IOP Conference Series: Materials Science and Engineering*, **862**, 032033. doi: 10.1088/1757-899X/862/3/032033.

Guenther, W. C. (1977). *Sampling inspection in statistical quality control*. Macmillan.

Johnson, N. L., Kotz, S. a Balakrishnan (1994). *Continuous univariate distributions, Volume 1, Second Edition*. John Wiley & Sons, New York.

Krishnamoorthy, K. a Mathew, T. (2009). *Statistical tolerance regions: theory, applications, and computation*. John Wiley & Sons, New York.

Otava, M. (2017). Stručný průvodce statistickými intervaly. *Pokroky matematiky, fyziky a astronomie*, **62**(1), 7–16.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Van der Vaart, A. W. (2000). *Asymptotic statistics, Volume 3*. Cambridge University Press, New York.