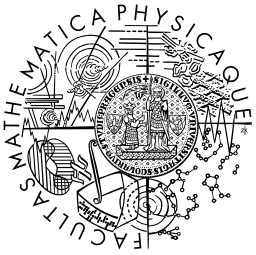


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Tereza Blatská

Koncentrační nerovnosti pro součty

Katedra pravděpodobnosti a matematické statistiky

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

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracovala 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

Ráda bych poděkovala doc. RNDr. Danielu Hlubinkovi, Ph.D. za jeho odborné rady a čas strávený nad touto prací.

Název práce: Koncentrační nerovnosti pro součty

Autor: Tereza Blatská

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Daniel Hlubinka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této bakalářské práci se zabýváme koncentračními nerovnostmi pro součty nezávislých náhodných veličin, které jsou omezené a nemusí být nutně stejně rozdělené. Hlavním pilířem práce je Hoeffdingova nerovnost, hledání jejího zpřesnění a dalších podobných nerovností. Jednotlivé nerovnosti doplňují základní příklady pro různá pravděpodobnostní rozdělení. Součástí každého příkladu je obecný teoretický výpočet, simulace pro konkrétně zvolené parametry a grafické znázornění získaných odhadů, které bylo zpracováno s pomocí programovacího jazyka R.

Klíčová slova: koncentrační nerovnosti, součty nezávislých náhodných veličin, Hoeffdingova nerovnost

Title: Concentration inequalities for sums

Author: Tereza Blatská

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Daniel Hlubinka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this bachelor thesis we focus on concentration inequalities for sums of independent random variables, which are bounded and not necessarily identically distributed. The main pillar of the thesis is Hoeffding's inequality, finding its improvement and other similar inequalities. Inequalities are completed with examples for various probability distributions. In each example there is a theoretical calculation, a simulation for specifically selected parameters and a graphical representation of all the obtained estimates, which was created using the R programming language.

Keywords: concentration inequalities, sums of independent random variables, Hoeffding's inequality

Obsah

Úvod

V pravděpodobnosti (a matematické statistice) nás často zajímá odchylka náhodné veličiny od její střední hodnoty. Odpověď na tuto otázku může dát centrální limitní věta, ovšem výsledek který dostaneme je pouze přibližný a skutečná hodnota může být větší i menší.

Ve chvíli, kdy nás zajímá, jak nejvíc se náhodná veličina může lišit od své střední hodnoty, lze využít koncentrační nerovnosti, které jasně určují horní hranici. V této práci se zaměříme na koncentrační nerovnosti pro součty, tedy snažíme se shora omezit pravděpodobnosti typu $\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq x)$, kde S_n je součet nezávislých náhodných veličin a $x > 0$ je zvolená odchylka.

V první kapitole představíme Hoeffdingovu nerovnost, která se poprvé objevila v článku, jehož autorem byl finský matematik Wassily Hoeffding (1963). Nerovnost je užitečná v tom, že nemá příliš složité předpoklady. Náhodné veličiny, které se vyskytují v součtu musí být nezávislé a shora i zdola omezené nějakými konstantami. O jejich rozdělení nemusíme vědět vůbec nic. Uvedeme dva motivační příklady.

Příklad. Pozorujme úspěšnost studentů během n let.
Označme X_k = podíl studentů, kteří úspěšně dokončili bakalářské studium v k-tém roce, přičemž úspěšnosti v jednotlivých letech jsou na sobě nezávislé. Tedy máme posloupnost vzájemně nezávislých náhodných veličin $X_1, \ldots, X_n$, n $\in \mathbb{N}$, takovou že $\forall k \in \{1, \ldots, n\}$ platí $0 \leq X_k \leq 1$, o jejich rozdělení nevíme nic. Označme střední hodnotu úspěšnosti studentů $\mathbb{E}X$. Za její odhad lze uvažovat výběrový průměr $\overline{X}_n$. Zajímá nás, jak přesný tento odhad je, respektive, jak moc se může lišit od skutečné střední hodnoty. Platí

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}X| \geq x) = \mathbb{P}\Big(\Big|\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\Big| \geq x\Big) = \mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq n\,x).$$

Příklad. Studenti se rozhodli oslavit zakončené bakalářské studium.
Celkem 50 z nich se vydalo do baru. Označme X_k = počet nápojů, které vypije k-tý student, $k \in \{1, \ldots, 50\}$. Přičemž průměrný student vypije 3 nápoje, zároveň každý student zvládne za večer vypít maximálně 10 nápojů a studenti pijí nezávisle na sobě. Máme tedy posloupnost nezávislých náhodných veličin $X_1, \ldots, X_{50}$ takových, že $\forall k \in \{1, \ldots, 50\}$ platí $0 \leq X_k \leq 50$ a očekavaná hodnota vypitých nápojů jedním studentem je $\mathbb{E}X = 3$, tedy $E[S_{50}] = 50 \cdot \mathbb{E}X = 150$.
Zajímá nás, jaká je pravděpodobnost, že celkový počet vypitých nápojů je větší než 200, neboli matematicky zapsáno

$$\mathbb{P}(S_{50} \geq 200) = \mathbb{P}(S_{50} - \mathbb{E}[S_{50}] \geq 200 - \mathbb{E}[S_{50}]) = \mathbb{P}(S_{50} - \mathbb{E}[S_{50}] \geq 50).$$

Odhady pravděpodobností v obou příkladech získáme použitím již zmíněné Hoeffdingovy nerovnosti.

V druhé kapitole se zaměříme na náhodné veličiny, které jsou omezené shora a které jsou omezené symetricky shora i zdola. Pro tyto náhodné veličiny formulujeme a dokážeme nerovnosti podobné Hoeffdingově.

Ve třetí a poslední kapitole bude naším cílem zpřesnit Hoeffdingovu nerovnost. Formulace a důkaz tohoto zpřesnění se poprvé objevil v knize od autorů Bercu a kol. (2015), jedná se tedy o poměrně nový výsledek. V celé práci vycházíme především z právě zmíněné knihy.

Předpokládá se, že čtenář má základní znalosti z teorie pravděpodobnosti.

1. Hoeffdingova nerovnost

1.1 Úvod

Cílem první kapitoly bude formulace a důkaz Hoeffdingovy nerovnosti. Dále bude následovat příklad pro použití této nerovnosti.

1.2 Připomenutí známých tvrzení

Pro připomenutí si vyslovíme věty, které budeme dále potřebovat.

Věta 1 (Markovova nerovnost)**.** *Nechť X je náhodná veličina s konečným n-tým momentem, $a > 0$. Potom platí*

$$P(|X| \geq a) \leq \frac{E|X|^n}{a^n}.$$

Věta 2 (Centrální limitní věta)**.** *Mějme $X_1, \ldots, X_n$, $n \in \mathbb{N}$ nezávislé, stejně rozdělené náhodné veličiny se střední hodnotou $\mu \in \mathbb{R}$ a rozptylem $\sigma^2 \in (0,\infty)$. Potom platí*

$$\frac{\sum_{k=1}^{n}(X_k - \mu)}{\sqrt{n\sigma^2}} \xrightarrow{D} \mathcal{N}(0,1),$$

kde $\xrightarrow{D}$ značí konvergenci v distribuci a $\mathcal{N}(0,1)$ je normované normální rozdělení.

Poznámka. V tomto textu budeme pracovat se součtem náhodných veličin, tj. s $S_n = \sum_{k=1}^{n} X_k$. Pokud náhodné veličiny X_k, $k \in \{1, \ldots, n\}$ splňují předpoklady centrální limitní věty, pak platí

$$\frac{\sum_{k=1}^{n}(X_k - \mu)}{\sqrt{n\sigma^2}} = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n\sigma^2}} = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\operatorname{var} S_n}} \xrightarrow{D} \mathcal{N}(0,1).$$

1.3 Hoeffdingova nerovnost

Nejprve si dokážeme dvě pomocná lemmata, která budeme potřebovat v důkazu samotné Hoeffdingovy nerovnosti.

Lemma 3. *Nechť X je náhodná veličina s konečným rozptylem a existují a, $b \in \mathbb{R}$ taková, že $a \leq X \leq b$ s.j. Potom platí $\operatorname{var} X \leq \frac{(b-a)^2}{4}$.*

Důkaz. Z předpokladu $a \leq X \leq b$ s.j. získáme následující nerovnosti (platí s.j.)

$$0 \leq (X-a)(b-X) = -X^2 + (a+b)X - ab \iff X^2 \leq (a+b)X - ab.$$

Potom z definice rozptylu a s využitím výše uvedeného dostáváme

$$\begin{aligned}\operatorname{var} X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 \leq \mathbb{E}[(a+b)X - ab] - (\mathbb{E}X)^2 \\ &= (a+b)\mathbb{E}X - ab - (\mathbb{E}X)^2 = (\mathbb{E}X - a)(b - \mathbb{E}X) \leq \frac{(b-a)^2}{4}.\end{aligned}$$

Poslední rovnost platí ze vztahu $xy \leq \frac{(x+y)^2}{4}$ pro $x = (\mathbb{E}X - a)$ a $y = (b - \mathbb{E}X)$. □

Lemma 4. *Nechť X je náhodná veličina s konečným rozptylem a existují a, $b \in \mathbb{R}$ taková, že $a \leq X \leq b$ s.j. Potom $\forall t \in \mathbb{R}$ platí*

$$\ln \mathbb{E}[\exp(tX)] \leq t\mathbb{E}X + \frac{t^2}{8}(b-a)^2.$$

Důkaz. Mějme $t \in \mathbb{R}$. Označme $L(t) = \mathbb{E}[\exp(tX)]$, $l(t) = \ln L(t)$ a $(\Omega, \mathcal{A})$ měřitelný prostor, na kterém je definovaná náhodná veličina X z formulace lemmatu.

Nechť $\mathbb{P}$ je pravděpodobnostní míra na $(\Omega, \mathcal{A})$, vzhledem ke které je počítána střední hodnota daná v předpisu funkce $L(t)$. Potom získáme pravděpodobnostní míru $\mathbb{Q}$ pomocí transformace míry $\mathbb{P}$ následovně

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\Big(tX - l(t)\Big) = \frac{\exp(tX)}{\exp l(t)} = \frac{\exp(tX)}{L(t)} = \frac{\exp(tX)}{\mathbb{E}[\exp(tX)]}.$$

Ověříme, že se skutečně jedná o pravděpodobností míru na prostoru $(\Omega, \mathcal{A})$. Mějme $A \in \mathcal{A}$, potom

$$\begin{aligned}\mathbb{Q}(A) &= \int_A d\mathbb{Q} = \int_A \frac{d\mathbb{Q}}{d\mathbb{P}}\, d\mathbb{P}(\omega) = \int_A \frac{\exp(tX(\omega))}{\mathbb{E}[\exp(tX)]}\, d\mathbb{P}(\omega) \\ &= \frac{1}{\mathbb{E}[\exp(tX)]} \int_A \exp(tX(\omega))\, d\mathbb{P}(\omega) \geq 0,\end{aligned}$$

dále s využitím definice střední hodnoty

$$\mathbb{Q}(\Omega) = \frac{1}{\mathbb{E}[\exp(tX)]} \int_\Omega \exp(tX(\omega))\, d\mathbb{P}(\omega) = \frac{\mathbb{E}[\exp(tX)]}{\mathbb{E}[\exp(tX)]} = 1.$$

Zřejmě platí σ-aditivita (z linearity integrálu) a $\mathbb{Q}(\emptyset) = 0$. Tedy $\mathbb{Q}$ je opravdu pravděpodobností míra na měřitelném prostoru $(\Omega, \mathcal{A})$.

Označme $\mathbb{E}_\mathbb{Q}$ střední hodnotu a $\mathrm{var}_\mathbb{Q}$ rozptyl (počítané vzhledem k pravděpodobnostní míře $\mathbb{Q}$). Můžeme spočítat

$$\mathbb{E}_\mathbb{Q}[X] = \frac{\mathbb{E}[X\exp(tX)]}{L(t)} = \frac{L'(t)}{L(t)} \qquad \text{a} \qquad \mathbb{E}_\mathbb{Q}[X^2] = \frac{\mathbb{E}[X^2\exp(tX)]}{L(t)} = \frac{L''(t)}{L(t)}.$$

Mohli jsme provést záměnu derivace a integrálu, neboť $t \in \mathbb{R}$ a tedy existuje kladná konstanta t_0 tak, že $|t| < t_0$. U první derivace a funkce $\exp(tX)$ máme integrovatelnou majorantu $\exp(t_0|X|)$, u druhé derivace a funkce $X\exp(tX)$ je integrovatelná majoranta $\exp(t_0|X|)(t_0 - |t|)^{-1}$.
Tedy z definice rozptylu a s využitím výše uvedeného dostáváme

$$\mathrm{var}_\mathbb{Q}\, X = \mathbb{E}_\mathbb{Q}[X^2] - (\mathbb{E}_\mathbb{Q}[X])^2 = \frac{L''(t)}{L(t)} - \Big(\frac{L'(t)}{L(t)}\Big)^2. \tag{1.1}$$

Všimneme si, že

$$l'(t) = \Big(\ln L(t)\Big)' = \frac{L'(t)}{L(t)} \qquad \text{a} \qquad l''(t) = \frac{L''(t)}{L(t)} - \Big(\frac{L'(t)}{L(t)}\Big)^2. \tag{1.2}$$

Celkem z (1.1) a (1.2) plyne, že $l''(t) = \mathrm{var}_\mathbb{Q}\, X$ a lze použít lemma 3, které nám dává $l''(t) \leq \frac{(b-a)^2}{4}$. Potom integrací obou stran nerovnosti získáme

$$\begin{aligned}\int_0^t l''(t)\, dt &\leq \int_0^t \frac{(b-a)^2}{4}\, dt \\ l'(t) - l'(0) &\leq t\, \frac{(b-a)^2}{4} \\ l'(t) &\leq \mathbb{E}X + t\, \frac{(b-a)^2}{4},\end{aligned}$$

neboť

$$l'(0) = \frac{L'(0)}{L(0)} = \frac{\mathbb{E}[X\exp(0X)]}{\mathbb{E}[\exp(0X)]} = \mathbb{E}X.$$

Získanou nerovnost zintegrujeme ještě jednou a dostaneme

$$\begin{aligned}\int_0^t l'(t)\,dt &\le \int_0^t \mathbb{E}X + t\,\frac{(b-a)^2}{4}\,dt\\ l(t) - l(0) &\le t\,\mathbb{E}X + \frac{t^2}{2}\,\frac{(b-a)^2}{4}\\ l(t) &\le t\,\mathbb{E}X + \frac{t^2}{8}(b-a)^2,\end{aligned}$$

neboť $l(0) = \ln L(0) = \ln 1 = 0$. □

Věta 5 (Hoeffdingova nerovnost)**.** *Mějme nezávislé náhodné veličiny $X_1, \dots, X_n$, $n \in \mathbb{N}$. Nechť $\forall k \in \{1, \dots, n\}$ existují reálné konstanty a_k, b_k takové, že $a_k < b_k$ a $a_k \le X_k \le b_k$ s.j. Označme $S_n = \sum_{k=1}^n X_k$. Potom $\forall x > 0$ platí*

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \ge x) \le 2\exp\Big(-\frac{2x^2}{\sum_{k=1}^n (b_k - a_k)^2}\Big).$$

Důkaz. K důkazu věty potřebujeme ukázat dvě nerovnosti:

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \ge x) \le \exp\Big(-\frac{2x^2}{\sum_{k=1}^n (b_k - a_k)^2}\Big) \tag{1.3}$$

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \le -x) \le \exp\Big(-\frac{2x^2}{\sum_{k=1}^n (b_k - a_k)^2}\Big). \tag{1.4}$$

Z lemmatu 4 a nezávislosti náhodných veličin platí $\forall t \in \mathbb{R}$

$$\begin{aligned}&\ln \mathbb{E}[\exp(tS_n)] = \sum_{k=1}^n \ln \mathbb{E}[\exp(tX_k)] \le \sum_{k=1}^n \Big(t\mathbb{E}[X_k] + \frac{t^2}{8}(b_k - a_k)^2\Big)\\ &= t\mathbb{E}[\sum_{k=1}^n X_k] + \frac{t^2}{8}\sum_{k=1}^n (b_k - a_k)^2 = t\mathbb{E}[S_n] + \frac{t^2}{8}\sum_{k=1}^n (b_k - a_k)^2.\end{aligned} \tag{1.5}$$

Nerovnost (1.3) lze upravit následovně

$$\begin{aligned}&\mathbb{P}(S_n - \mathbb{E}[S_n] \ge x) = \mathbb{P}(S_n \ge \mathbb{E}[S_n] + x) = \mathbb{P}(tS_n \ge t\mathbb{E}[S_n] + tx)\\ &= \mathbb{P}\Big(\exp(tS_n) \ge \exp(t\mathbb{E}[S_n])\exp(tx)\Big) \le \frac{\mathbb{E}[\exp(tS_n)]}{\exp(t\mathbb{E}[S_n])\exp(tx)}.\end{aligned} \tag{1.6}$$

V poslední úpravě jsme použili Markovovu nerovnost (věta 1). Aplikujeme logaritmus na obě strany (1.6), využijeme (1.5) a dostáváme

$$\begin{aligned}&\ln \mathbb{P}(S_n - \mathbb{E}[S_n] \ge x) \le \ln \mathbb{E}[\exp(tS_n)] - t\mathbb{E}[S_n] - tx\\ &\le t\mathbb{E}[S_n] + \frac{t^2}{8}\sum_{k=1}^n (b_k - a_k)^2 - t\mathbb{E}[S_n] - tx = \frac{t^2}{8}\sum_{k=1}^n (b_k - a_k)^2 - tx.\end{aligned} \tag{1.7}$$

Předchozí nerovnost platí $\forall t \in \mathbb{R}$ a chceme, aby byl odhad co nejlepší. Označme pravou stranu nerovnosti $f(t)$, potřebujeme najít minimum funkce $f(t)$. Derivaci funkce $f(t)$ položíme rovnou nule, tím nalezneme extrémy.

$$f'(t) = \frac{2t}{8}\sum_{k=1}^n (b_k - a_k)^2 - x = 0 \iff t = \frac{4x}{\sum_{k=1}^n (b_k - a_k)^2}$$

Protože $f''(t) \geq 0$, jedná se o konvexní funkci a tedy nalezený extrém je minimum funkce $f(t)$. Dosadíme do nerovnosti (1.7) za t.

$$\ln \mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) \leq \frac{1}{8} \frac{16x^2}{\left(\sum_{k=1}^n (b_k - a_k)^2\right)^2} \sum_{k=1}^n (b_k - a_k)^2 - \frac{4x^2}{\sum_{k=1}^n (b_k - a_k)^2}$$
$$= \frac{2x^2}{\sum_{k=1}^n (b_k - a_k)^2} - \frac{4x^2}{\sum_{k=1}^n (b_k - a_k)^2} = -\frac{2x^2}{\sum_{k=1}^n (b_k - a_k)^2}$$

Aplikujeme exponencielu na obě strany nerovnosti a dostáváme (1.3). Nerovnost (1.4) pak plyne z nerovnosti (1.3), stačí $\forall k \in \{1, \ldots, n\}$ místo X_k uvažovat $-X_k$. □

1.4 Příklad a grafické znázornění

Příklad.
Mějme náhodné veličiny $X_1, \ldots, X_n, n \in \mathbb{N}$, které jsou nezávislé a stejně rozdělené s beta rozdělením s parametry α, β. Tedy platí

$$\mathbb{E}X_1 = \frac{\alpha}{\alpha + \beta} \qquad \operatorname{var} X_1 = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}.$$

Potom pro $S_n = \sum_{k=1}^n X_k$ snadno spočteme

$$\mathbb{E}S_n = n\,\mathbb{E}X_1 = \frac{n\alpha}{\alpha + \beta} \qquad \operatorname{var} S_n = n \operatorname{var} X_1 = \frac{n\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}.$$

Z centrální limitní věty (věta 2) a poznámky za ní víme, že

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\operatorname{var} S_n}} \xrightarrow{D} \mathcal{N}(0,1).$$

Lze tedy upravit

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq x) = \mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) + \mathbb{P}(S_n - \mathbb{E}[S_n] \leq -x)$$
$$= 1 - \mathbb{P}(S_n - \mathbb{E}[S_n] \leq x) + \mathbb{P}(S_n - \mathbb{E}[S_n] \leq -x)$$
$$= 1 - \mathbb{P}\Big(\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\operatorname{var} S_n}} \leq \frac{x}{\sqrt{\operatorname{var} S_n}}\Big) + \mathbb{P}\Big(\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\operatorname{var} S_n}} \leq \frac{-x}{\sqrt{\operatorname{var} S_n}}\Big)$$
$$\doteq 1 - \Phi\Big(\frac{x}{\sqrt{\operatorname{var} S_n}}\Big) + \Phi\Big(\frac{-x}{\sqrt{\operatorname{var} S_n}}\Big) = 2 - 2\Phi\Big(\frac{x}{\sqrt{\operatorname{var} S_n}}\Big),$$

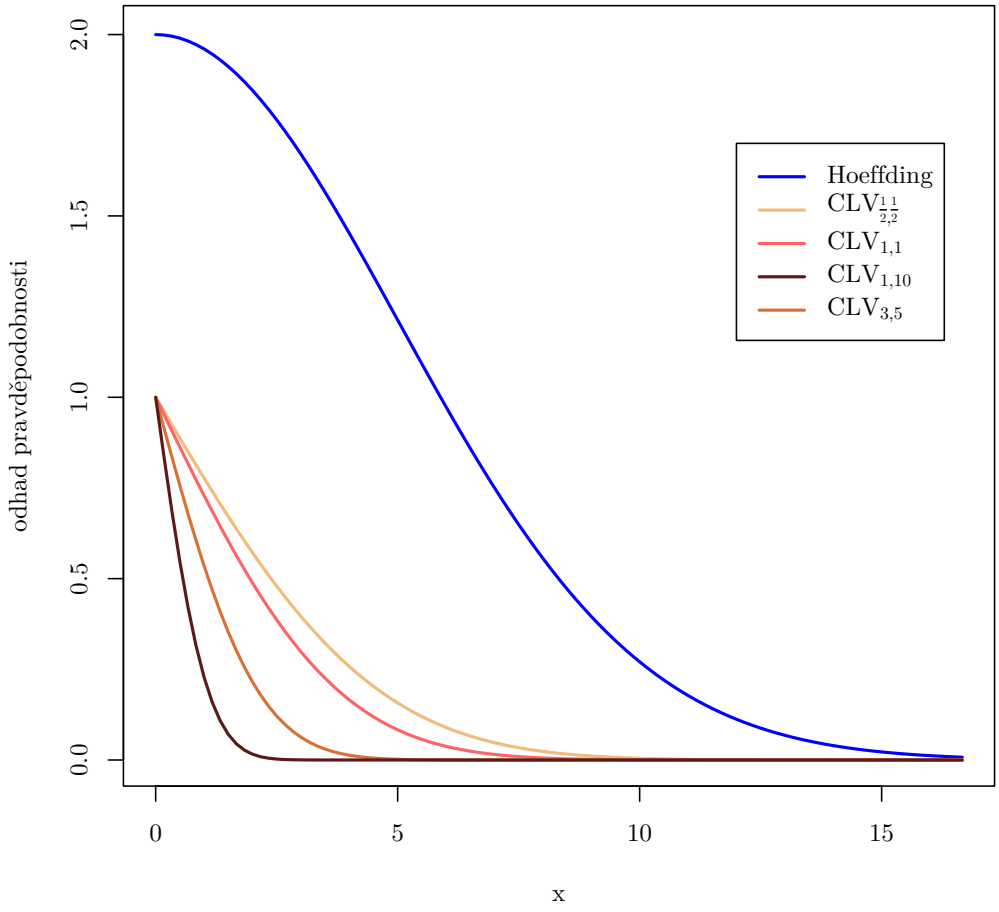
kde $\Phi(x)$ značí distribuční funkci normovaného normálního rozdělení. V poslední úpravě jsme využili rovnosti $\Phi(-x) = 1 - \Phi(x)$, která plyne ze symetrie distribuční funkce $\Phi(x)$. Výsledek je ovšem pouze přibližný, neboť máme pouze asymptotické rozdělení $\mathcal{N}(0,1)$, nikoliv přesné.

K hornímu odhadu pravděpodobnosti $\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq x)$ lze využít Hoeffdingovu větu. Její předpoklady jsou splněny, neboť pro náhodné veličiny, které mají beta rozdělení platí, že se pohybují na intervalu [0,1] s.j., to znamená $a_k = 0$ a $b_k = 1$. Tedy z věty 5 dostáváme

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq x) \leq 2\exp\Big(-\frac{2x^2}{\sum_{k=1}^n 1}\Big) = 2\exp\Big(-\frac{2x^2}{n}\Big).$$

Můžeme si všimnout, že přibližný odhad pravděpodobnosti, který jsme získali použitím centrální limitní věty se bude měnit v závislosti na tom, jaké budou parametry α a β, zatímco odhad pomocí Hoeffdingovy nerovnosti se nezmění. Navíc výpočet asymptotického rozptylu pro použití v centrální limitní větě může být často obtížný.

Nevýhodou odhadu pomocí Hoeffdingovy věty je ovšem větší nepřesnost, což lze vypozorovat z následujícího grafu. K jeho vykreslení byla použita simulace o rozsahu 100 náhodných veličin. Označme $CLV_{i,j}$ odhad pomocí centrální limitní věty s parametry $\alpha = i$ a $\beta = j$.



Obrázek 1.1: Odhady pravděpodobnosti $\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq x)$, kde $S_n = \sum_{k=1}^{n} X_k$ je součet nezávislých náhodných veličin s beta rozdělením s různými parametry. Odhady jsou získané z centrální limitní věty a Hoeffdingovy nerovnosti (věta 5).

Vidíme, že získaný odhad z Hoeffdingovy věty dokonce nabývá hodnot větších než 1. To je způsobeno tím, že uvažujeme oboustranný odhad (v pravděpodobnosti máme absolutní hodnotu). Pokud bychom uvažovali pouze odhad pravděpodobnosti $\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x)$, pak již odhad z Hoeffdingovy nerovnosti bude nabývat hodnot z intervalu (0,1). Cílem v dalších kapitolách bude najít vylepšení tohoto odhadu.

2. Nerovnosti pro součty omezených náhodných veličin

2.1 Úvod

V této kapitole se zaměříme na další nerovnosti pro součty omezených náhodných veličin. Nejprve se budeme zabývat náhodnými veličinami, které jsou omezené shora a z toho pak odvodíme další výsledky pro symetricky omezené náhodné veličiny. Tyto výsledky použijeme v následující kapitole k vylepšení Hoeffdingovy nerovnosti.

2.2 Shora omezené náhodné veličiny

2.2.1 Pomocná tvrzení

Zadefinujeme si pomocnou funkci, která se nám bude v odhadech vyskytovat a poté si dokážeme dvě pomocná lemmata, která budeme potřebovat v důkazu vět této kapitoly.

Definice 1. *Definujme funkci* $\tau : (0,\infty) \longrightarrow (0,\infty)$ *předpisem*

$$\tau(v) = \begin{cases} \frac{1-v^2}{|ln(v)|} & v < 1 \\ 2v & v \geq 1. \end{cases}$$

Lemma 6. *Nechť* $v \in (0,1]$ *a* τ *je funkce definovaná v definici 1. Potom platí*

$$\tau(v) \leq \frac{v^2 + 4v + 1}{3}.$$

Důkaz. Pro $v = 1$ zřejmě platí.
Dále mějme $v \in (0,1)$, potom platí $|\ln(v)| = -\ln(v)$ a tedy z předpisu funkce τ

$$\tau(v) = \frac{1 - v^2}{|\ln(v)|} = \frac{1 - v^2}{-\ln(v)}.$$

Dokazovanou nerovnost si lze přepsat následovně

$$\begin{aligned} \frac{1 - v^2}{-\ln(v)} &\leq \frac{v^2 + 4v + 1}{3} \\ 3(1 - v^2) &\leq -\ln(v)(v^2 + 4v + 1) \\ 0 &\leq -\ln(v)(v^2 + 4v + 1) + 3(v^2 - 1). \end{aligned} \tag{2.1}$$

Zvolme substituci $v = 1 - u$, potom $0 < u < 1$ a z Taylorova rozvoje logaritmu

$$\begin{aligned} \ln(v) = \ln(1 - u) &= \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(1 - u - 1)^k}{k} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(-u)^k}{k} \\ &= \sum_{k=1}^{\infty} (-1)^{k+1} (-1)^k \frac{u^k}{k} = -\sum_{k=1}^{\infty} \frac{u^k}{k} \end{aligned}$$

a tedy

$$
\begin{aligned}
&-\ln(1-u)\Big((1-u)^2+4(1-u)+1\Big)+3\Big((1-u)^2-1\Big)\\
&=-\ln(1-u)(u^2-6u+6)+3(u^2-2u)\\
&=-u^2\ln(1-u)+6u\ln(1-u)-6\ln(1-u)+3u^2-6u\\
&=\sum_{k=1}^{\infty}\frac{u^{k+2}}{k}-\sum_{k=1}^{\infty}\frac{6u^{k+1}}{k}+\sum_{k=1}^{\infty}\frac{6u^k}{k}+3u^2-6u\\
&=\sum_{k=1}^{\infty}\frac{u^{k+2}}{k}-\sum_{k=1}^{\infty}\frac{6u^{k+1}}{k}+6u+\frac{6u^2}{2}+\sum_{k=3}^{\infty}\frac{6u^k}{k}+3u^2-6u\\
&=\sum_{k=1}^{\infty}\frac{u^{k+2}}{k}-\sum_{k=1}^{\infty}\frac{6u^{k+1}}{k}+\sum_{k=3}^{\infty}\frac{6u^k}{k}+6u^2\\
&=\sum_{k=1}^{\infty}\frac{u^{k+2}}{k}-6u^2-\sum_{k=2}^{\infty}\frac{6u^{k+1}}{k}+\sum_{k=3}^{\infty}\frac{6u^k}{k}+6u^2\\
&=\sum_{k=1}^{\infty}\frac{u^{k+2}}{k}-\sum_{k=2}^{\infty}\frac{6u^{k+1}}{k}+\sum_{k=3}^{\infty}\frac{6u^k}{k}=\sum_{k=3}^{\infty}\frac{u^k}{k-2}-\sum_{k=3}^{\infty}\frac{6u^k}{k-1}+\sum_{k=3}^{\infty}\frac{6u^k}{k}\\
&=\sum_{k=3}^{\infty}\frac{u^k(k^2-7k+12)}{k(k-1)(k-2)}=\sum_{k=3}^{\infty}\frac{u^k(k-3)(k-4)}{k(k-1)(k-2)}=\sum_{k=5}^{\infty}\frac{u^k(k-3)(k-4)}{k(k-1)(k-2)}.
\end{aligned}
$$

V úpravách jsme si rozepsali některé členy jednotlivých sum, abychom se zbavili výrazu $3u^2-6u$. Potom jsme v sumách posunuli indexy, sumy sečetli a převedli výraz v sumě na společného jmenovatele. Poslední suma je zřejmě kladná (neboť se jedná o sumu s kladnými členy), tímto jsme dokázali nerovnost (2.1). □

Poznámka. V následujícím důkazu a několika dalších větách budeme používat značení $(x)_+^2$, čímž budeme rozumět $\Big(\max(0,x)\Big)^2$.

Lemma 7. *Nechť X je náhodná veličina taková, že $X\le 1$ s.j., $\mathbb{E}X=0$, $0<\mathrm{var}\,X\le v$ a τ je funkce definovaná v definici 1. Potom $\forall t>0$ platí*

$$\ln\mathbb{E}[\exp(tX)]\le\frac{t^2\tau(v)}{4}.$$

Důkaz. Označme Y náhodnou veličinu takovou, že Y nabývá hodnot z $\{1,-v\}$ a platí

$$
\begin{aligned}
\mathbb{P}(Y=1)&=\frac{v}{1+v}\\
\mathbb{P}(Y=-v)&=\frac{1}{1+v}.
\end{aligned}
\tag{2.2}
$$

Potom $\mathbb{E}Y=0$ a $\mathrm{var}\,Y=\mathbb{E}Y^2=\frac{v}{1+v}+\frac{v^2}{1+v}=v$.

Chceme ukázat, že $\forall t\in\mathbb{R}$ platí $\mathbb{E}[(X-t)_+^2]\le\mathbb{E}[(Y-t)_+^2]$.
Pro $t\ge 1$:

$$\mathbb{E}[(X-t)_+^2]=\mathbb{E}[0]=\mathbb{E}[(Y-t)_+^2]$$

Pro $t\le -v$:

$$
\begin{aligned}
\mathbb{E}[(X-t)_+^2]\le\mathbb{E}[(X-t)^2]&=\mathbb{E}X^2+t^2=\mathrm{var}\,X+t^2\le v+t^2=\mathrm{var}\,Y+t^2\\
&=\mathbb{E}Y^2+t^2=\mathbb{E}[(Y-t)^2]=\mathbb{E}[(Y-t)_+^2]
\end{aligned}
$$

Pro $t \in (-v,1)$:
K odhadu střední hodnoty využijeme nerovnost

$$(x-t)_+ \le \frac{1-t}{1+v}\,(x+v)_+ \qquad t,\, x \in (-v,1).$$

Ta zřejmě platí pro $x < t$, neboť potom levá strana nerovnosti je rovna 0, zatímco pravá strana nerovnosti je kladná. Pro $x \ge t$:

$$\begin{aligned}(x-t)_+ &\le \frac{1-t}{1+v}\,(x+v)_+\\ (x-t)(1+v) &\le (x+v)(1-t)\\ x+xv-t-tv &\le x-xt+v-tv\\ xv-t &\le v-xt\\ x(v+t) &\le v+t\end{aligned}$$

Poslední nerovnost platí, neboť $x \le 1$. Nyní už můžeme odhadnout střední hodnotu.

$$\begin{aligned}\mathbb{E}[(X-t)_+^2] &\le \mathbb{E}\Big[\frac{(1-t)^2}{(1+v)^2}\,(X+v)_+^2\Big] \le \frac{(1-t)^2}{(1+v)^2}\,\mathbb{E}[(X+v)^2]\\ &= \frac{(1-t)^2}{(1+v)^2}\,(\mathbb{E}X^2+v^2) = \frac{(1-t)^2}{(1+v)^2}\,(\operatorname{var}X+v^2)\\ &\le \frac{(1-t)^2}{(1+v)^2}\,(v+v^2) = \frac{(1-t)^2\,v}{1+v} = \mathbb{E}[(Y-t)_+^2]\end{aligned}$$

Tedy ukázali jsme, že $\forall t \in \mathbb{R}$ platí

$$\mathbb{E}[(X-t)_+^2] \le \mathbb{E}[(Y-t)_+^2]. \tag{2.3}$$

Označme $(\Omega, \mathcal{A}, \mathbb{P})$ pravděpodobnostní prostor, na němž je definována náhodná veličina X. Potom lze upravit

$$\begin{aligned}&\int_{\mathbb{R}} \mathbb{E}\Big[\Big(X-\frac{s}{t}\Big)_+^2\Big]\exp(s)\,ds = \int_{\mathbb{R}}\int_{\Omega}\Big(X(\omega)-\frac{s}{t}\Big)_+^2\exp(s)\,d\mathbb{P}(\omega)\,ds\\ &= \int_{\Omega}\int_{\mathbb{R}}\Big(X(\omega)-\frac{s}{t}\Big)_+^2\exp(s)\,ds\,d\mathbb{P}(\omega) = \int_{\Omega}\int_{-\infty}^{tX(\omega)}\Big(X(\omega)-\frac{s}{t}\Big)^2\exp(s)\,ds\,d\mathbb{P}(\omega)\\ &= \frac{2}{t}\int_{\Omega}\int_{-\infty}^{tX(\omega)}\Big(X(\omega)-\frac{s}{t}\Big)\exp(s)\,ds\,d\mathbb{P}(\omega) = \frac{2}{t^2}\int_{\Omega}\int_{-\infty}^{tX(\omega)}\exp(s)\,ds\,d\mathbb{P}(\omega)\\ &= \frac{2}{t^2}\int_{\Omega}\exp\Big(tX(\omega)\Big)\,d\mathbb{P}(\omega) = \frac{2}{t^2}\,\mathbb{E}[\exp(tX)].\end{aligned}$$

V úpravách jsme postupně použili Fubiniho větu a dvakrát per partes pro výpočet vnitřního integrálu. Potom z právě získaného a nerovnosti (2.3) plyne

$$\begin{aligned}\mathbb{E}[\exp(tX)] &= \frac{t^2}{2}\int_{\mathbb{R}}\mathbb{E}\Big[\Big(X-\frac{s}{t}\Big)_+^2\Big]\exp(s)\,ds \le \frac{t^2}{2}\int_{\mathbb{R}}\mathbb{E}\Big[\Big(Y-\frac{s}{t}\Big)_+^2\Big]\exp(s)\,ds\\ &= \mathbb{E}[\exp(tY)] = \frac{v}{1+v}\exp(t)+\frac{1}{1+v}\exp(-vt) = \frac{v\exp(t)+\exp(-vt)}{1+v}.\end{aligned}$$

Aplikací logaritmu na obě strany nerovnosti dostaneme

$$\ln\mathbb{E}[\exp(tX)] \le \ln\Big(v\exp(t)+\exp(-vt)\Big) - \ln(1+v). \tag{2.4}$$

Dále s využitím Legendre-Fenchelovy transformace platí, že

$$\ln\Big(v\exp(t)+\exp(-vt)\Big)-\ln(1+v)\le\frac{t^2\tau(v)}{4}. \tag{2.5}$$

Důkaz lze najít v druhé kapitole knihy od autorů Bercu a kol. (2015). Celkem z (2.4) a (2.5) dostáváme dokazovanou nerovnost. □

2.2.2 Nerovnost pro součty

Nyní už jsme schopni dokázat hlavní větu této kapitoly.

Věta 8 (Nerovnosti pro součty náhodných veličin omezených shora)**.**
Mějme nezávislé centrované náhodné veličiny $X_1,\dots,X_n, n\in\mathbb{N}$.
Nechť $\forall k\in\{1,\dots,n\}$ *existují reálné kladné* b_k, v_k *takové, že* $X_k\le b_k$ *s.j. a* $var(X_k)\le v_k$. *Označme* $S_n=\sum_{k=1}^n X_k$ *a* $V_n=\sum_{k=1}^n v_k$. *Potom* $\forall x>0$ *platí*

$$\mathbb{P}(S_n\ge x)\le\exp\Big(-\frac{x^2}{A_n}\Big)\le\exp\Big(-\frac{3x^2}{6V_n+B_n}\Big)\le\exp\Big(-\frac{3x^2}{5V_n+C_n}\Big),$$

kde $A_n=\sum_{k=1}^n b_k^2\,\tau\Big(\frac{v_k}{b_k^2}\Big)$, $B_n=\sum_{k=1}^n\Big(b_k-\frac{v_k}{b_k}\Big)_+^2$, $C_n=\sum_{k=1}^n\max(b_k^2,v_k)$ *a* τ *je funkce definovaná v definici 1.*

Důkaz. Důkaz si rozdělíme na dva kroky. V prvním kroku dokážeme první nerovnost, zbylé nerovnosti z ní pak v druhém kroku odvodíme.
1. krok: Budeme postupovat podobně jako v důkazu Hoeffdingovy nerovnosti (Věta 5). Využijeme nezávislosti náhodných veličin a lemmatu 7 pro $X:=\frac{X_k}{b_k}$ a $v:=\frac{v_k}{b_k^2}$. Předpoklady lemmatu jsou splněny, neboť z předpokladů věty $\forall k\in\{1,\dots,n\}: X_k\le b_k$ s.j. a $\operatorname{var}(X_k)\le v_k$ a tedy $X=\frac{X_k}{b_k}\le 1$, $\mathbb{E}X=0$ (neboť všechny X_k jsou centrované) a $\operatorname{var}X=\operatorname{var}\frac{X_k}{b_k}=\frac{1}{b_k^2}\operatorname{var}X_k\le\frac{v_k}{b_k^2}$. Z lemmatu $\forall t>0$ plyne

$$\begin{aligned}\ln\mathbb{E}[\exp(tS_n)]&=\sum_{k=1}^n\ln\mathbb{E}[\exp(tX_k)]=\sum_{k=1}^n\ln\mathbb{E}\Big[\exp\Big(tb_k\frac{X_k}{b_k}\Big)\Big]\\&\le\frac{t^2}{4}\sum_{k=1}^n b_k^2\tau\Big(\frac{v_k}{b_k^2}\Big)=\frac{t^2}{4}A_n.\end{aligned} \tag{2.6}$$

Dále upravíme

$$\mathbb{P}(S_n\ge x)=\mathbb{P}(tS_n\ge tx)=\mathbb{P}\Big(\exp(tS_n)\ge\exp(tx)\Big)\le\frac{\mathbb{E}[\exp(tS_n)]}{\exp(tx)}. \tag{2.7}$$

V poslední úpravě jsme použili Markovovu nerovnost (věta 1). Aplikujeme logaritmus na obě strany (2.7), využijeme (2.6) a dostáváme

$$\ln\mathbb{P}(S_n\ge x)\le\ln\mathbb{E}[\exp(tS_n)]-tx\le\frac{t^2}{4}A_n-tx. \tag{2.8}$$

Označme pravou stranu nerovnosti $f(t)$. Odhad chceme co nejpřesnější, hledejme tedy minimum funkce $f(t)$.

$$f'(t)=\frac{t}{2}A_n-x=0\iff t=\frac{2x}{A_n}$$

Protože $f''(t) \geq 0$, jedná se o konvexní funkci a nalezený bod je minimum funkce $f(t)$. V nerovnosti (2.8) položíme $t = \frac{2x}{A_n}$ a aplikujeme exponencielu na obě strany nerovnosti a dostáváme

$$\mathbb{P}(S_n \geq x) \leq \exp\left(\frac{4x^2}{4A_n^2}A_n - \frac{2x^2}{A_n}\right) = \exp\left(-\frac{x^2}{A_n}\right).$$

2. krok: Předpokládejme nejprve, že $\forall k \in \{1,\ldots,n\} : v_k < b_k^2$ a tedy $\frac{v_k}{b_k^2} \in [0,1]$ a můžeme použít lemma 6 pro $v := \frac{v_k}{b_k^2}$. Postupnými úpravami dostaneme

$$\begin{aligned}
\tau\left(\frac{v_k}{b_k^2}\right) &\leq \frac{1}{3}\left(1 + \frac{4v_k}{b_k^2} + \frac{v_k^2}{b_k^4}\right) \\
3\,b_k^2\,\tau\left(\frac{v_k}{b_k^2}\right) &\leq b_k^2 + 4v_k + \frac{v_k^2}{b_k^2} = \left(b_k - \frac{v_k}{b_k}\right)^2 + 6v_k \\
3\,b_k^2\,\tau\left(\frac{v_k}{b_k^2}\right) &\leq \left(b_k - \frac{v_k}{b_k}\right)_+^2 + 6v_k. \qquad (2.9)
\end{aligned}$$

Lze si všimnout, že poslední nerovnost platí i pro $v_k \geq b_k^2$, protože potom $\frac{v_k}{b_k^2} \geq 1$ a z předpisu funkce τ (definice 1) máme

$$3\,b_k^2\,\tau(\frac{v_k}{b_k^2}) = 3\,b_k^2\,\frac{2v_k}{b_k^2} = 6v_k.$$

Tedy nerovnost (2.9) platí pro $\forall k \in \{1,\ldots,n\}\ \forall v_k > 0$. Navíc platí

$$\left(b_k - \frac{v_k}{b_k}\right)_+^2 \leq b_k\left(b_k - \frac{v_k}{b_k}\right)_+ = (b_k^2 - v_k)_+. \qquad (2.10)$$

Dále obě strany (2.9) sčítáme přes $k \in \{1,\ldots,n\}$ a postupnými úpravami dostaneme

$$\begin{aligned}
\sum_{k=1}^{n} 3\,b_k^2\,\tau\left(\frac{v_k}{b_k^2}\right) &\leq \sum_{k=1}^{n}\left(b_k - \frac{v_k}{b_k}\right)_+^2 + \sum_{k=1}^{n} 6v_k \\
3A_n &\leq B_n + 6V_n \\
A_n &\leq \frac{1}{3}(B_n + 6V_n).
\end{aligned}$$

Odtud už nám z prvního kroku plyne první nerovnost. Pro důkaz druhé nerovnosti využijeme (2.10), platí totiž

$$\begin{aligned}
B_n + 6V_n &= \sum_{k=1}^{n}\left(b_k - \frac{v_k}{b_k}\right)_+^2 + \sum_{k=1}^{n} v_k + 5V_n \leq \sum_{k=1}^{n}\Big((b_k^2 - v_k)_+ + v_k\Big) + 5V_n \\
&= \sum_{k=1}^{n}\Big(\max(b_k^2 - v_k, 0) + v_k\Big) + 5V_n = \sum_{k=1}^{n}\max(b_k^2, v_k) + 5V_n = C_n + 5V_n.
\end{aligned}$$

□

Poznámka. Větu lze zřejmě použít i pro náhodné veličiny, které jsou omezené zdola nebo nejsou centrované. V tomto případě můžeme náhodné veličiny otočit (tzn. uvažovat je se záporným znaménkem) a odečíst od nich (po otočení přičíst) jejich střední hodnotu, tím už jsou předpoklady věty splněny. Tento případ uvidíme v následujícím příkladu.

2.2.3 Příklad a grafické znázornění

Příklad.
Mějme náhodné veličiny $X_1, \ldots, X_n, n \in \mathbb{N}$, které jsou nezávislé a stejně rozdělené s exponenciálním rozdělením s parametrem $\lambda > 0$. Tedy platí

$$\mathbb{E}X_1 = \frac{1}{\lambda} \qquad \operatorname{var} X_1 = \frac{1}{\lambda^2}.$$

Označme $Y_k = \mathbb{E}X_k - X_k$, $\forall k \in \{1, \ldots, n\}$. Potom platí $Y_k \leq \mathbb{E}X_k = \frac{1}{\lambda}$ a navíc

$$\mathbb{E}Y_1 = \mathbb{E}X_1 - \mathbb{E}X_1 = 0 \qquad \operatorname{var} Y_1 = \operatorname{var} X_1 = \frac{1}{\lambda^2}.$$

Označme $S_n = \sum_{k=1}^{n} Y_k$. Z nezávislosti a stejného rozdělení snadno spočteme

$$\mathbb{E}S_n = n\,\mathbb{E}Y_1 = 0 \quad \operatorname{var} S_n = n \operatorname{var} Y_1 = \frac{n}{\lambda^2}.$$

Nakonec z centrální limitní věty (věta 2) a poznámky za ní plyne

$$\frac{S_n}{\sqrt{\operatorname{var} S_n}} \xrightarrow{D} \mathcal{N}(0,1).$$

Můžeme upravit

$$\mathbb{P}(S_n \geq x) = 1 - \mathbb{P}\Big(\frac{S_n}{\sqrt{\operatorname{var} S_n}} \leq \frac{x}{\sqrt{\operatorname{var} S_n}}\Big) \doteq 1 - \Phi\Big(\frac{x}{\sqrt{\operatorname{var} S_n}}\Big),$$

kde $\Phi(x)$ značí distribuční funkci normovaného normálního rozdělení. Výsledek je pouze přibližný, protože nemáme přesné rozdělení $\mathcal{N}(0,1)$.

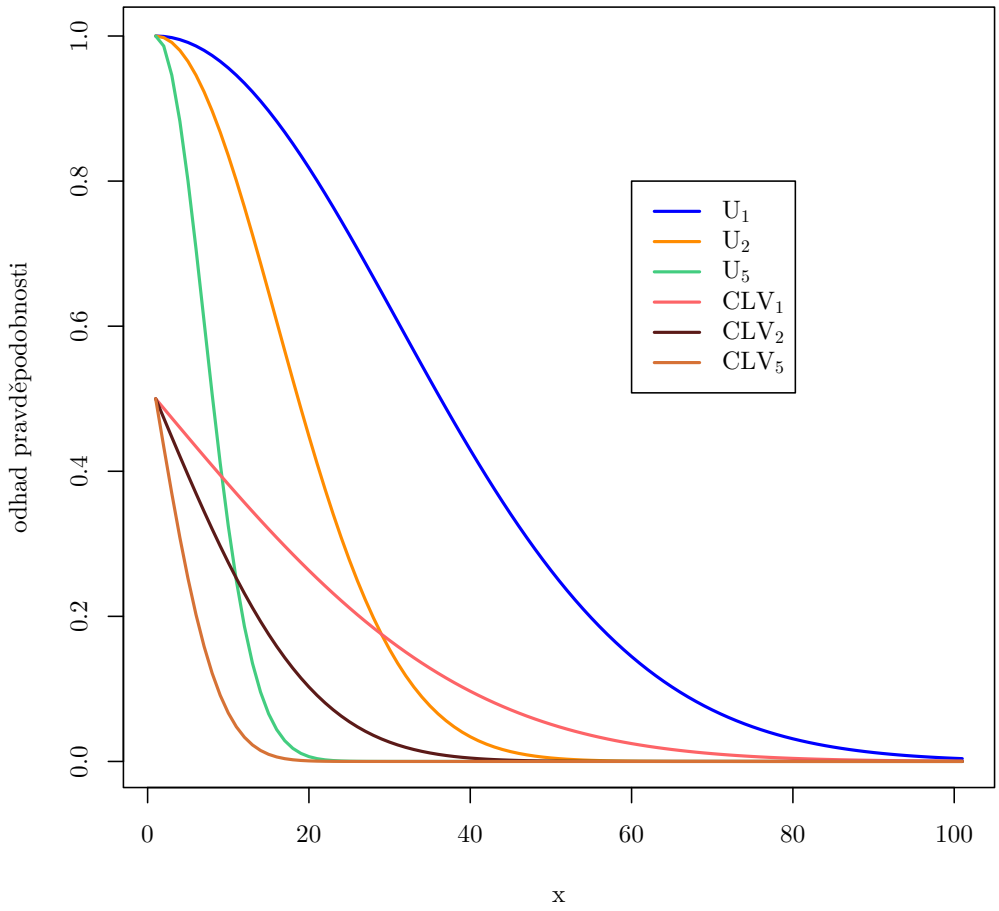
K odhadu pravděpodobnosti $\mathbb{P}(S_n \geq x)$ lze využít i větu 8, neboť náhodné veličiny $Y_1, \ldots, Y_n$, $n \in \mathbb{N}$ splňují předpoklady této věty pro $b_k = \frac{1}{\lambda}$ a $v_k = \frac{1}{\lambda^2}$. Spočteme výrazy vyskytující se ve znění věty.

$$A_n = \frac{2n}{\lambda^2} \qquad B_n = 0 \qquad C_n = V_n = \frac{n}{\lambda^2}$$

Tedy podle věty dostáváme odhad (všechny tři odhady z věty se v tomto případě rovnají)

$$\mathbb{P}(S_n \geq x) \leq \exp\Big(-\frac{x^2\,\lambda^2}{2n}\Big).$$

Můžeme si všimnout, že odhad z věty 8 se tentokrát bude měnit pro různé hodnoty parametru λ. Označme tento odhad s parametrem $\lambda = i$ jako U_i a odhad (opět pro parametr $\lambda = i$) získaný pomocí centrální věty jako CLV_i. Pomocí nasimulovaných dat pro $n = 100$ a rozdílné hodnoty parametru λ (uvažujme hodnoty 1, 2 a 5) si vykreslíme graf pro porovnání odhadů. Pro velká n nám odhad CLV spolehlivě odhaduje skutečnou hodnotu pravděpodobnosti, ovšem skutečná hodnota může být větší i menší než odhad CLV. Z grafu a odhadu U lze vyčíst nejvyšší možnou hodnotu pravděpodobnosti $\mathbb{P}(S_n \geq x)$.



Obrázek 2.1: Odhady pravděpodobnosti $\mathbb{P}(S_n \geq x)$, kde $S_n = \sum_{k=1}^{n} X_k$ je součet nezávislých náhodných veličin s exponenciálním rozdělením (postupně s parametry 1, 2 a 5). Odhady jsou získané z centrální limitní věty a z věty pro součty náhodných veličin omezených shora (věta 8).

2.3 Symetricky omezené náhodné veličiny

Věta 9 (Nerovnosti pro součty symetricky omezených náhodných veličin)**.** *Mějme nezávislé centrované náhodné veličiny* $X_1, \ldots, X_n, n \in \mathbb{N}$. *Nechť* $\forall k \in \{1, \ldots, n\}$ *existuje reálné kladné* b_k *takové, že* $|X_k| \leq b_k$ *s.j. Označme* $S_n = \sum_{k=1}^{n} X_k$, $v_k := var(X_k)$ *a* $V_n = \sum_{k=1}^{n} v_k$. *Potom* $\forall x > 0$ *platí*

$$\mathbb{P}(S_n \geq x) \leq \exp\left(-\frac{x^2}{A_n}\right) \leq \exp\left(-\frac{3x^2}{D_n + 5V_n}\right) \leq \exp\left(-\frac{x^2}{2D_n}\right),$$

kde $A_n = \sum_{k=1}^{n} b_k^2 \, \tau\left(\frac{v_k}{b_k^2}\right)$ *a* $D_n = \sum_{k=1}^{n} b_k^2$.

Důkaz. K důkazu využijeme větu o nerovnostech pro součty náhodných veličin, které jsou omezené shora (věta 8) a značení v ní zavedené. Díky omezenosti a centrovanosti náhodných veličin platí (podobně jako v důkazu lemma 3)

$$v_k = \operatorname{var} X_k \leq (b_k - \mathbb{E}X_k)(\mathbb{E}X_k + b_k) = b_k^2$$

a tedy

$$C_n = \sum_{k=1}^{n} \max(b_k^2, v_k) = \sum_{k=1}^{n} b_k^2 = D_n.$$

Potom z věty 8 máme

$$\begin{aligned}\mathbb{P}(S_n \geq x) &\leq \exp\Big(-\frac{x^2}{A_n}\Big) \leq \exp\Big(-\frac{3x^2}{5V_n + C_n}\Big) = \exp\Big(-\frac{3x^2}{5V_n + D_n}\Big) \\ &\leq \exp\Big(-\frac{3x^2}{5D_n + D_n}\Big) = \exp\Big(-\frac{x^2}{2D_n}\Big).\end{aligned}$$

V předposlední úpravě jsme využili toho, že $\exp(-x)$ je klesající funkce a

$$V_n = \sum_{k=1}^{n} v_k \leq \sum_{k=1}^{n} b_k^2 = D_n.$$

□

3. Zpřesnění Hoeffdingovy nerovnosti

3.1 Úvod

V této kapitole se budeme věnovat přesnějšímu odhadu než nám dává Hoeffdingova nerovnost. K důkazu tvrzení o tomto přesnějším odhadu využijeme poznatky získané v prvních dvou kapitolách.

3.2 Formulace a důkaz zpřesnění

Věta 10 (Zpřesnění Hoeffdingovy nerovnosti)**.**
Mějme nezávislé náhodné veličiny $X_1, \dots, X_n, n \in \mathbb{N}$.
Nechť $\forall k \in \{1, \dots, n\}$ *existují reálné* a_k, b_k *takové, že* $a_k < b_k$ *a* $a_k \leq X_k \leq b_k$ *s.j.*
Označme $S_n = \sum_{k=1}^{n} X_k$ *a* $V_n = \operatorname{var} S_n$. *Potom* $\forall x > 0$ *platí*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) \leq \exp\Big(- \frac{3x^2}{\sum_{k=1}^{n}(b_k - a_k)^2 + 2V_n}\Big).$$

Důkaz. Uvažujme $k \in \{1, \dots, n\}$. Označme $Y_k = X_k - \mathbb{E}X_k$ a $v_k = \operatorname{var} Y_k$. Potom $Y_1, \dots, Y_n$, $n \in \mathbb{N}$ je posloupnost nezávislých, centrovaných veličin splňujících $v_k = \operatorname{var} X_k$ a $\alpha_k \leq Y_k \leq \beta_k$, kde $\alpha_k = a_k - \mathbb{E}X_k$ a $\beta_k = b_k - \mathbb{E}X_k$. Z omezenosti náhodných veličin X_k plyne i omezenost středních hodnot. Tedy platí

$$a_k \leq \mathbb{E}X_k \leq b_k \iff a_k - \mathbb{E}X_k \leq 0 \leq b_k - \mathbb{E}X_k \iff \alpha_k \leq 0 \leq \beta_k,$$

přičemž rovnost nastává právě tehdy, když $Y_k = 0$. V tomto případě můžeme Y_k z posloupnosti náhodných veličin vyřadit. Dále označme $c_k = \max(|\alpha_k|, \beta_k)$, tedy $|Y_k| \leq c_k$. Celkem dostáváme, že posloupnost náhodných veličin $Y_1, \dots, Y_n$, $n \in \mathbb{N}$ splňuje předpoklady věty nerovnosti pro součty symetricky omezených náhodných veličin (věta 9). Z věty plyne

$$\begin{aligned} \mathbb{P}(\sum_{k=1}^{n} Y_k \geq x) = \mathbb{P}\Big(\sum_{k=1}^{n}(X_k - \mathbb{E}X_k) \geq x\Big) &= \mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) \\ &\leq \exp\Big(- \frac{x^2}{A_n}\Big) \qquad \text{kde } A_n = \sum_{k=1}^{n} c_k^2\, \tau\Big(\frac{v_k}{c_k^2}\Big). \end{aligned} \tag{3.1}$$

Z lemmatu 6 dostáváme

$$\tau\Big(\frac{v_k}{c_k^2}\Big) \leq \frac{1}{3}\Big(\frac{v_k^2}{c_k^4} + \frac{4v_k}{c_k^2} + 1\Big),$$

z čehož plyne nerovnost

$$\begin{aligned} A_n = \sum_{k=1}^{n} c_k^2\, \tau\Big(\frac{v_k}{c_k^2}\Big) \leq \sum_{k=1}^{n} c_k^2\, \frac{1}{3}\Big(\frac{v_k^2}{c_k^4} + \frac{4v_k}{c_k^2} + 1\Big) &= \frac{1}{3}\sum_{k=1}^{n}\Big(\frac{v_k^2}{c_k^2} + 4v_k + c_k^2\Big) \\ = \frac{1}{3}\sum_{k=1}^{n}\Big(\Big(\frac{v_k}{c_k} + c_k\Big)^2 + 2v_k\Big) &= \frac{1}{3}\sum_{k=1}^{n}\Big(\frac{v_k}{c_k} + c_k\Big)^2 + \frac{2}{3}V_n. \end{aligned} \tag{3.2}$$

Navíc stejně jako v důkazu lemmatu 3 platí

$$\begin{aligned} v_k &= \operatorname{var} Y_k = \operatorname{var} X_k \le (\mathbb{E}X_k - a_k)(b - \mathbb{E}X_k) = -\alpha_k\,\beta_k = |\alpha_k|\,\beta_k \\ &= \min(|\alpha_k|,\beta_k)\ \max(|\alpha_k|,\beta_k). \end{aligned} \tag{3.3}$$

Potom z definice c_k a nerovností (3.2) a (3.3) dostáváme

$$\begin{aligned} A_n &\le \frac{1}{3}\sum_{k=1}^{n}\Big(\frac{v_k}{c_k} + c_k\Big)^2 + \frac{2}{3}V_n \le \frac{1}{3}\sum_{k=1}^{n}\Big(\min(|\alpha_k|,\beta_k) + \max(|\alpha_k|,\beta_k)\Big)^2 + \frac{2}{3}V_n \\ &= \frac{1}{3}\sum_{k=1}^{n}\Big(|\alpha_k| + \beta_k\Big)^2 + \frac{2}{3}V_n = \frac{1}{3}\sum_{k=1}^{n}(\beta_k - \alpha_k)^2 + \frac{2}{3}V_n \\ &= \frac{1}{3}\Big(\sum_{k=1}^{n}(b_k - a_k)^2 + 2V_n\Big). \end{aligned}$$

Odtud pak s využitím toho, že $\exp(-x)$ je klesající funkce a z nerovnosti (3.1) získáváme

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \ge x) \le \exp\Big(-\frac{3x^2}{\sum_{k=1}^{n}(b_k - a_k)^2 + 2V_n}\Big),$$

což jsme chtěli dokázat. □

3.3 Příklady a grafické znázornění

3.3.1 Spojité rozdělení

Příklad. Vraťme se k příkladu z první kapitoly.
Máme náhodné veličiny $X_1,\dots,X_n, n \in \mathbb{N}$, které jsou nezávislé a stejně rozdělené s beta rozdělením s parametry α, β. Zajímá nás odhad pravděpodobnosti $\mathbb{P}(S_n - \mathbb{E}[S_n] \ge x)$. Analogicky k příkladu z první kapitoly (nyní ale nemáme v pravděpodobnosti absolutní hodnotu) z centrální limitní věty plyne

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \ge x) \doteq 1 - \Phi\Big(\frac{x}{\sqrt{\operatorname{var} S_n}}\Big).$$

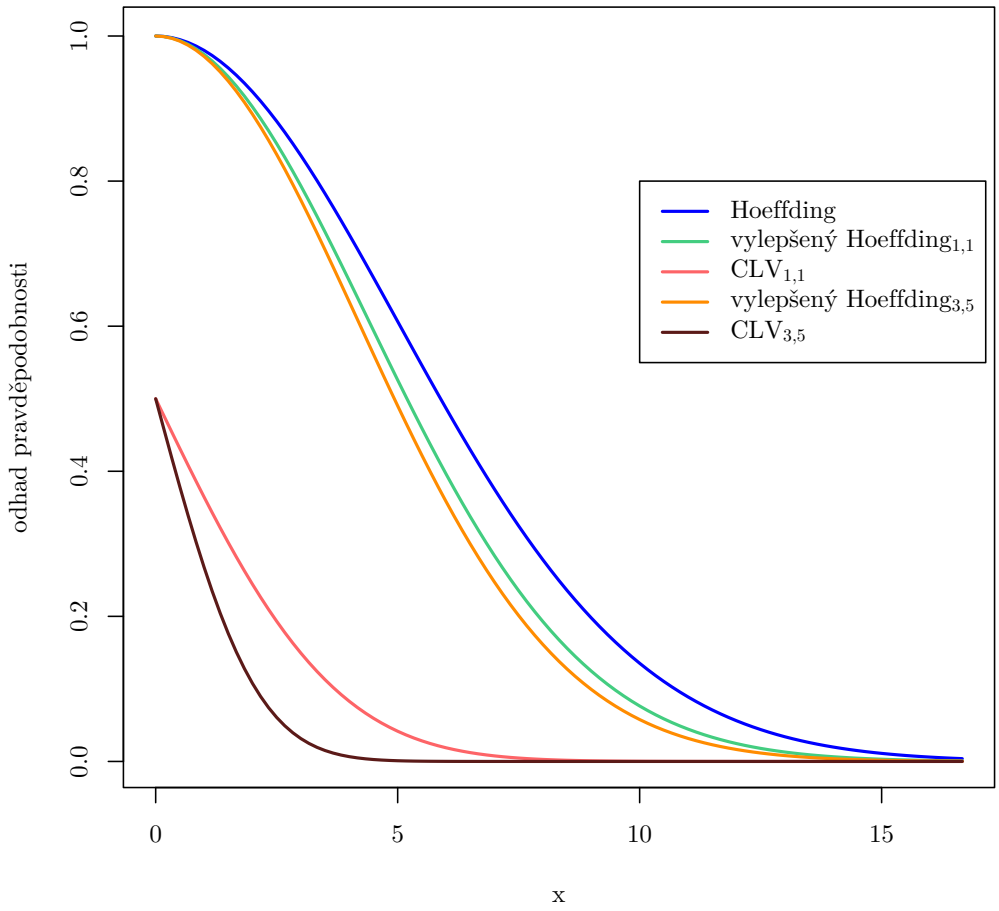
Z Hoeffdingovy věty (věta 5) dostáváme odhad

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \ge x) \le \exp\Big(-\frac{2x^2}{\sum_{k=1}^{n}1}\Big) = \exp\Big(-\frac{2x^2}{n}\Big).$$

Navíc z věty zpřesňující Hoeffdingův odhad (věta 10) získáváme

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \ge x) \le \exp\Big(-\frac{3x^2}{n + 2\operatorname{var} S_n}\Big).$$

Můžeme se všimnout, že poslední odhad (narozdíl od odhadu z klasické Hoeffdingovy nerovnosti) závisí na rozptylu S_n, tedy závisí i na parametrech α, β. Provedeme dvě simulace pro odlišné parametry, nejprve zvolme $\alpha = \beta = 1$ a v druhém případě uvažujme $\alpha = 3$, $\beta = 5$. Pro obě simulace položme $n = 100$. Opět budeme odlišovat odhady s různými parametry pomocí dolních indexů. Jednotlivé odhady můžeme porovnat ve vykresleném grafu níže.



Obrázek 3.1: Odhady pravděpodobnosti $\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x)$, kde $S_n = \sum_{k=1}^{n} X_k$ je součet nezávislých náhodných veličin s beta rozdělením s různými parametry. Odhady jsou získané z centrální limitní věty, Hoeffdingovy nerovnosti (věta 5) a zpřesněné Hoeffdingovy nerovnosti (věta 10).

3.3.2 Diskrétní rozdělení

Příklad.
Mějme náhodné veličiny $X_1, \ldots, X_n, n \in \mathbb{N}$, které jsou nezávislé a stejně rozdělené s alternativním rozdělením s parametrem $p \in (0,1)$.
Potom $S_n = \sum_{k=1}^{n} X_k$ má binomické rozdělení s parametry n, p, tedy $\mathbb{E}[S_n] = np$ a $\operatorname{var} S_n = np(1-p)$. Narozdíl od ostatních příkladů jsme schopni spočítat pravděpodobnost $\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x)$ přesně a ne pouze asymptoticky. Upravíme

$$\begin{aligned}\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) &= \mathbb{P}(S_n \geq x + n\,p) = 1 - \mathbb{P}(S_n < x + n\,p) \\ &= 1 - \sum_{0 \leq k < x + np} \binom{n}{k} p^k (1-p)^{n-k}\, 1_{\{1,\ldots,n\}}(k).\end{aligned}$$

Posloupnost náhodných veličin $X_1, \ldots, X_n, n \in \mathbb{N}$ splňuje předpoklady věty 5 a věty 10, neboť $\forall k \in \{1, \ldots, n\}$ platí $0 \leq X \leq 1$. Tedy z klasické Hoeffdingovy

nerovnosti (věta 5) máme odhad

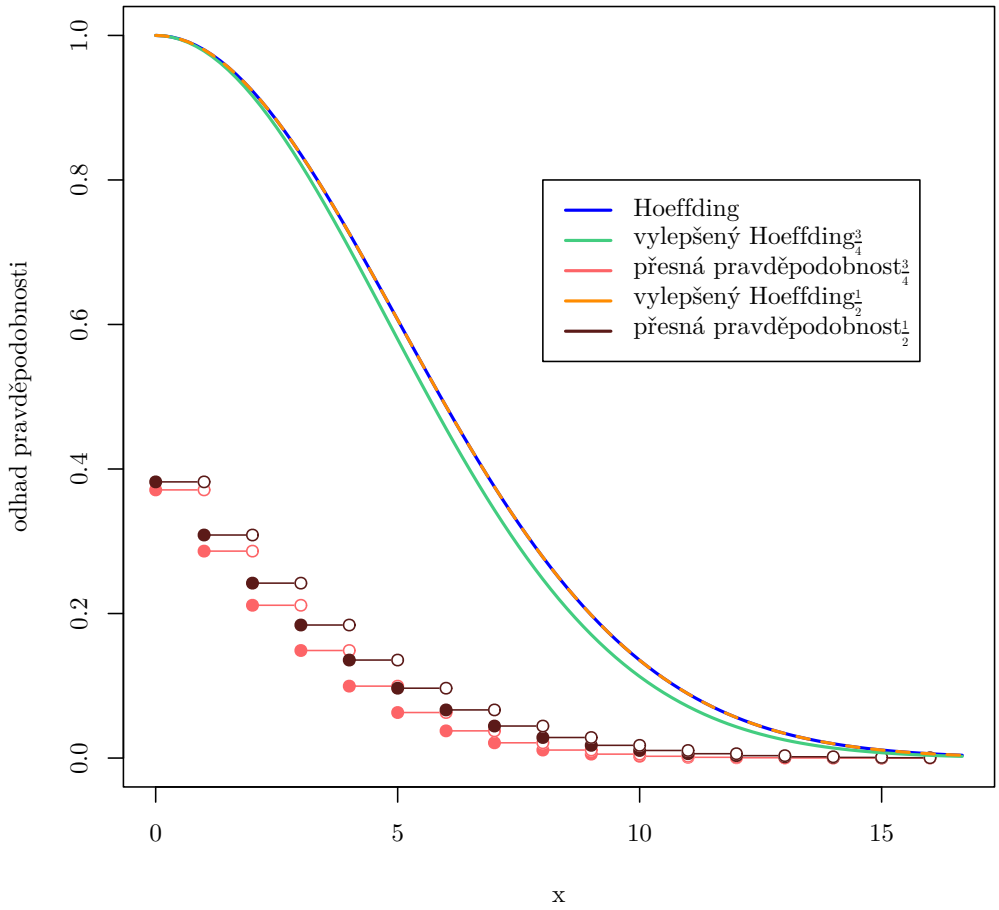
$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) \leq \exp\Big(-\frac{2x^2}{\sum_{k=1}^{n} 1}\Big) = \exp\Big(-\frac{2x^2}{n}\Big).$$

Lze si všimnout, že se jedná o stejný odhad jako v předchozím příkladu se spojitým rozdělením. V Hoeffdingově nerovnosti totiž vystupují pouze krajní nosiče rozdělení, které jsou u alternativního a beta rozdělení stejné.

Z vylepšené Hoeffdingovy nerovnosti (věta 10) máme odhad

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) \leq \exp\Big(-\frac{3x^2}{n + 2\operatorname{var} S_n}\Big) = \exp\Big(-\frac{3x^2}{n + 2\,np(1-p)}\Big).$$

Provedeme dvě simulace pro $n = 100$, pro různé hodnoty parametru p a to pro $p = \frac{1}{2}$ a $p = \frac{3}{4}$. Odhady s různými parametry opět odlišíme pomocí dolních indexů a můžeme je porovnat v následujícím grafu.



Obrázek 3.2: Odhady pravděpodobnosti $\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x)$, kde $S_n = \sum_{k=1}^{n} X_k$ je součet nezávislých náhodných veličin s alternativním rozdělením (postupně s parametry $\frac{1}{2}$ a $\frac{3}{4}$). Odhady jsou získané přímým výpočtem, použitím Hoeffdingovy nerovnosti (věta 5) a zpřesněné Hoeffdingovy nerovnosti (věta 10).

Pro odhad s parametrem $p = \frac{1}{2}$ nám odhad z Hoeffdingovy nerovnosti a odhad z vylepšené Hoeffdingovy nerovnosti splývá.

Poznámka. Z předchozích příkladů vidíme, že se odhad opravdu zlepšil (nebo zůstal stejný). Pokud bychom se v obecném případě chtěli přesvědčit, že došlo ke zpřesnění odhadu, pak postupnými úpravami dostáváme

$$\begin{aligned}
\exp\Big(-\frac{3x^2}{\sum_{k=1}^n (b_k-a_k)^2 + 2V_n}\Big) &\leq \exp\Big(-\frac{2x^2}{\sum_{k=1}^n (b_k-a_k)^2}\Big) \\
\Big(\frac{3x^2}{\sum_{k=1}^n (b_k-a_k)^2 + 2V_n}\Big) &\geq \Big(\frac{2x^2}{\sum_{k=1}^n (b_k-a_k)^2}\Big) \\
3\sum_{k=1}^n (b_k-a_k)^2 &\geq 2\sum_{k=1}^n (b_k-a_k)^2 + 4V_n \\
\frac{\sum_{k=1}^n (b_k-a_k)^2}{4} &\geq V_n.
\end{aligned}$$

Z lemmatu 3 víme, že poslední nerovnost platí.

Závěr

Nyní si shrneme získané poznatky. V této práci jsme se věnovali pravděpodobnosti $\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x)$, kde $S_n = \sum_{k=1}^{n} X_k$ je součet nezávislých náhodných veličin a $x > 0$ je zvolená odchylka.
Z centrální limitní věty (věta 2) dostáváme přibližný odhad

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) \doteq 1 - \Phi\Big(\frac{x}{\sqrt{\operatorname{var} S_n}}\Big).$$

Pracovali jsme s větami, které navíc předpokládají omezenost náhodných veličin, které se vyskytují v součtu S_n. Nechť tedy $\forall k \in \{1, \ldots, n\}$ existují reálné konstanty a_k, b_k takové, že $a_k < b_k$ a $a_k \leq X_k \leq b_k$ s.j.
Potom z Hoeffdingovy nerovnosti (věta 5) dostáváme horní odhad

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) \leq \exp\Big(-\frac{2x^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\Big).$$

Nakonec ze zpřesněné Hoeffdingovy nerovnosti (věta 10) dostáváme opět horní odhad

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x) \leq \exp\Big(-\frac{3x^2}{\sum_{k=1}^{n}(b_k - a_k)^2 + 2\operatorname{var} S_n}\Big).$$

V druhém a třetím odhadu máme horní odhad, což je jejich hlavní výhodou oproti odhadu z centrální limitní věty, který pouze přibližný a tedy nedostáváme žádnou pevnou dolní nebo horní hranici. Skutečná hodnota pravděpodobnosti $\mathbb{P}(S_n - \mathbb{E}[S_n] \geq x)$ může být větší i menší než odhad z centrální limitní věty, zatímco u odhadů z Hoeffdingovy nerovnosti a zpřesněné Hoeffdingovy nerovnosti máme jistotu, že skutečná hodnota pravděpodobnosti tyto odhady nepřesáhne.

Výhodou klasické Hoeffdingovy nerovnosti jsou její předpoklady. Jediné, co potřebujeme, je nezávislost a omezenost náhodných veličin, které se vyskytují v součtu. Nepotřebujeme znát rozdělení a především není třeba počítat rozptyl náhodné veličiny S_n, který se používá v ostatních odhadech. Ovšem zpřesněná Hoeffdingova nerovnost nám dává lepší odhad než klasická Hoeffdingova nerovnost, přestože získání odhadu je početněji náročnější nebo ho ani nelze spočítat.

Můžeme si všimnout, že druhý a třetí odhad nebudou příliš fungovat pro malé odchylky (tj. pro hodnoty x blízké nule), neboť v tomto případě se tyto odhady pohybují blízko jedné. Odhady tedy budou užitečné především pro větší odchylky, narozdíl od odhadu z centrální limitní věty, neboť ten rychle klesá k nule pro x jdoucí do nekonečna (což jsme mohli vidět v ilustrovaných příkladech v jednotlivých kapitolách) a tudíž nám o větších odchylkách nic neřekne.

Tímto jsme porovnali jednotlivé odhady a formulovali jejich výhody a nevýhody.

Seznam použité literatury

Bercu, B., Delyon, B. a Rio, E. (2015). *Concentration Inequalities for Sums and Martingales*. Springer Cham. ISBN 978-3-319-22098-7.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**(301), 13–30.

Seznam obrázků