

(11b) Intervaly spoľahlivosti – pokračovanie, zákony veľkých čísel

Na prednáške o intervaloch spoľahlivosti sme odvodili aj interval spoľahlivosti pre pravdepodobnosť úspechu, ak v N nezávislých pokusoch nastalo k úspechov. Toto môžeme využiť, ak hľadáme pomocou simulácií pravdepodobnosť určitej udalosti. Ak v N simuláciách nastala sledovaná udalosť k -krát, okrem prirodzeného odhadu pravdepodobnosti k/N vieme spraviť aj interval spoľahlivosti. Intuitívne čakáme, že ako sa bude zväčšovať počet simulácií, odhad sa bude spresňovať. To sa naozaj stane a prejaví sa to tým, že intervaly spoľahlivosti budú pri veľkom počte pokusov užšie. Dobré to vidieť na grafe, na ktorom budú odhady a intervaly spoľahlivosti, ktoré dostaneme, keď postupne pridávame výsledky ďalších simulácií (teda hodnota n na x-ovej osi zodpovedá odhad a príslušný interval spoľahlivosti z prvých n simulácií).

Použitie tohto princípu si ukážeme na príkladoch. Na nich súčasne uvidíme generovanie hodnôt náhodných vektorov. (Doteraz sme mali generovanie hodnôt náhodných premenných, ale uvidíme, že pre náhodné vektory to pôjde priamočiaro). Ukážeme si postup, na cvičeniach sa spraví konkrétna realizácia.

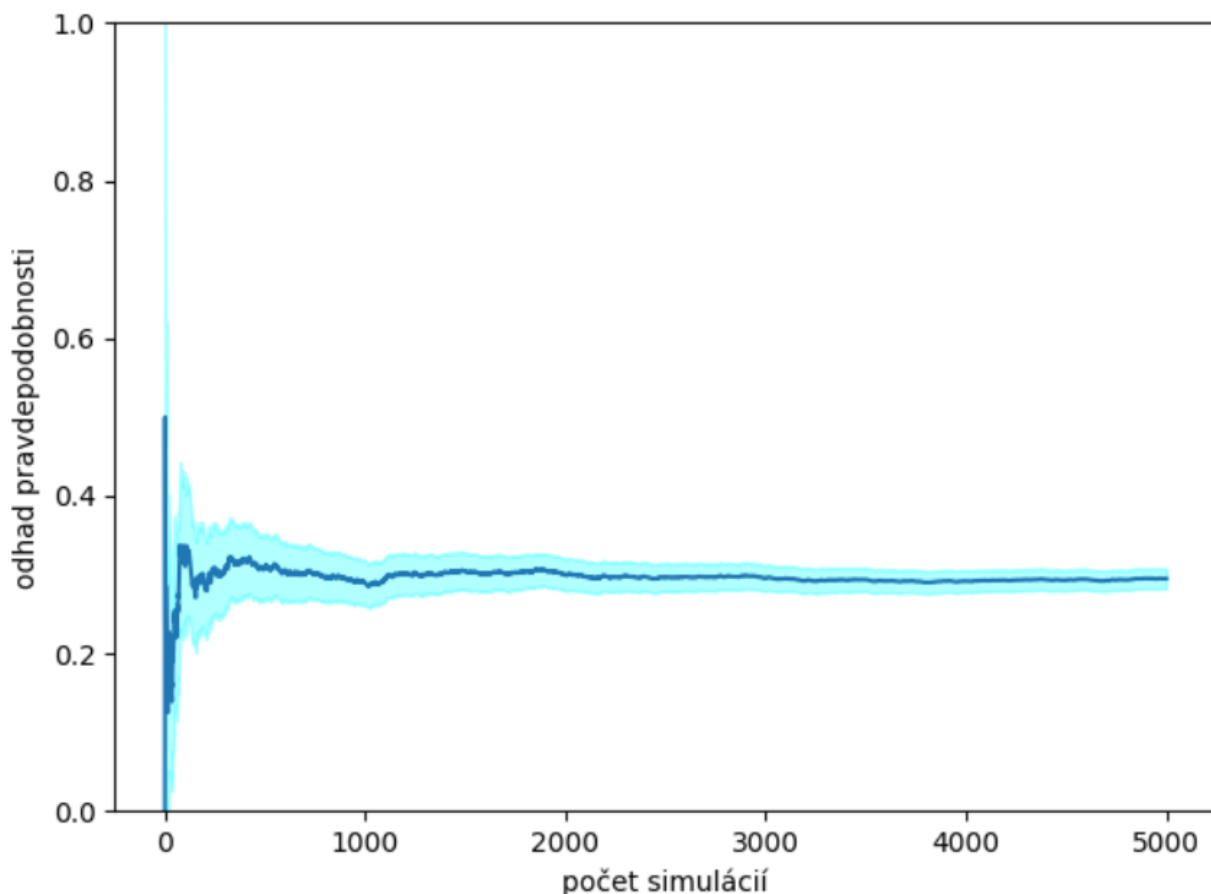
Príklad 1.

A family buys two policies from the same insurance company. Losses under the two policies are independent and have continuous uniform distributions on the interval from 0 to 10. One policy has a deductible of 1 and the other has a deductible of 2. The family experiences exactly one loss under each policy. Calculate the probability that the total benefit paid to the family does not exceed 5.

Deductible je spoluúčasť. Hodnota napr. 1 znamená, že ak je škoda menšia ako 1, tak poisťovňa nevyplatí nič. Ak je väčšia ako 1, tak poisťovňa vyplatí výšku škody zníženú o 1.

Postup:

- Máme tu náhodný vektor (X, Y) , ale jeho zložky sú nezávislé, takže ich môže generovať ako nezávislé náhodné premenné.
- Rozdelenie je rovnomerné – to vieme generovať.
- Platba, ktorá nás zaujíma, je $\max(X - 1, 0) + \max(Y - 2, 0)$
- Konkrétne nás zaujíma splnenie podmienky, že táto platba nie je vyššia ako 5.



Poznámky:

- Pre malé n nie je interval spoľahlivosti presný, to nám ale nevadí, lebo nás aj tak zaujíma to, na čom sa odhadnuté pravdepodobnosti ustália, resp. finálny interval spoľahlivosti. Graf ako taký môže slúžiť na to, aby sme sa presvedčili, či už odhad vyzerá dosť presný a či sa už pravdepodobnosti dostatočne ustálili.
- Keďže odhadujeme pravdepodobnosť, ak by nám vyšla dolná hranica intervalu spoľahlivosti záporná, zoberieme nulu. Takisto namiesto hornej hranice väčšej ako 1 zoberieme hodnotu 1.

Príklad 2.

An insurance policy is written to cover a loss X , where X has density function

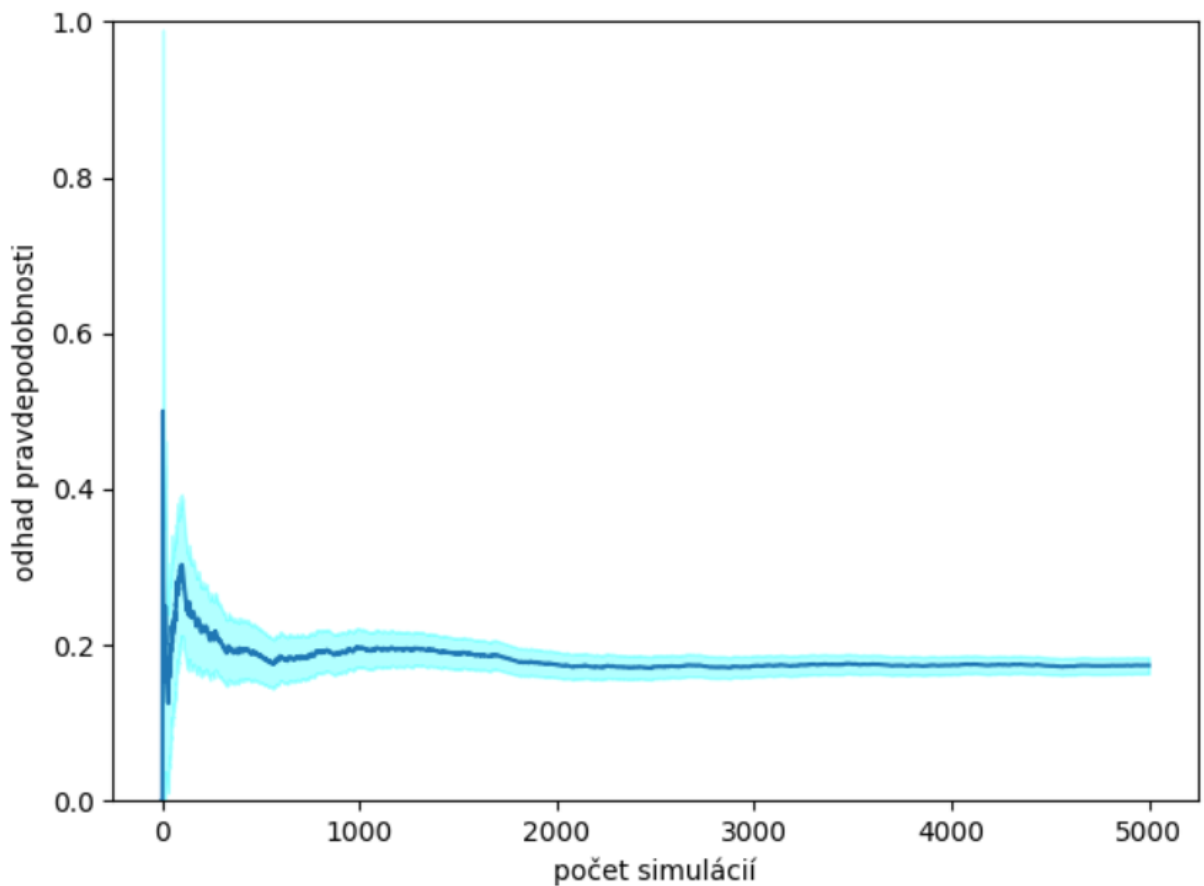
$$f(x) = \begin{cases} \frac{3}{8}x^2 & \text{for } 0 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

The time (in hours) to process a claim of size x , where $0 \leq x \leq 2$, is uniformly distributed on the interval from x to $2x$. Calculate the probability that a randomly chosen claim on this policy is processed in three hours or more.

Postup:

- Máme tu náhodný vektor $(\mathbf{X}, \mathbf{Y})$, jeho zložky nie sú nezávislé. Sú však dané vo veľmi "pohodlnom" tvare, čo sa týka simulácií:
 - Máme dané marginálne rozdelenie $\mathbf{X}$.
 - Potom máme dané podmienené rozdelenie $\mathbf{Y}$ pri danom $\mathbf{X}$.
- To znamená, že simulácia bude prebiehať nasledovne:
 - Vygenerujeme realizáciu náhodnej premennej $\mathbf{X}$ (použijeme univerzálny postup – budeme generovať náhodné čísla z rovnomerného rozdelenia na $(0,1)$ a dosadzovať ich do inverznej funkcie k distribučnej funkcii, tú vieme zo zadanej hustoty odvodiť).
 - Na základe nej vygenerujeme realizáciu náhodnej premennej $\mathbf{Y}$.
 - Overíme splnenie podmienky v zadaní.

Výsledok: (95-percentné intervaly spoľahlivosti)



Analogicky vieme **odhadovať aj strednú hodnotu**.

- Strednú hodnotu odhadneme priemerom.
- Priemer má (ak má sledovaná náhodná premenná strednú hodnotu a disperziu) podľa centrálnej limitnej vety pre veľký počet dát približne normálne rozdelenie.
- Preto môžeme spraviť interval spoľahlivosti pomocou kvantilov $N(0,1)$ rozdelenia:

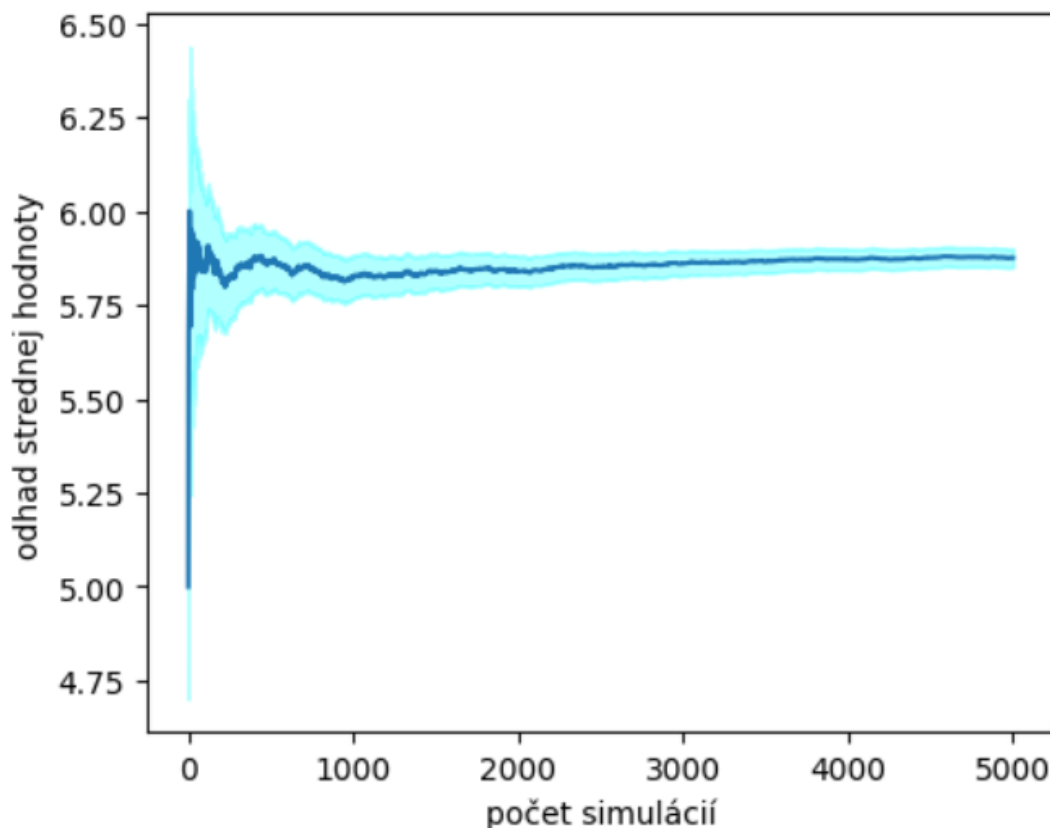
$$\left[\bar{X} - \sqrt{\frac{s^2}{n}} \cdot q_{0,975}, \bar{X} + \sqrt{\frac{s^2}{n}} \cdot q_{0,975} \right]$$

Takisto môžeme sledovať, ako sa hodnoty ustália a tiež, ako sa zužuje interval spoľahlivosti.

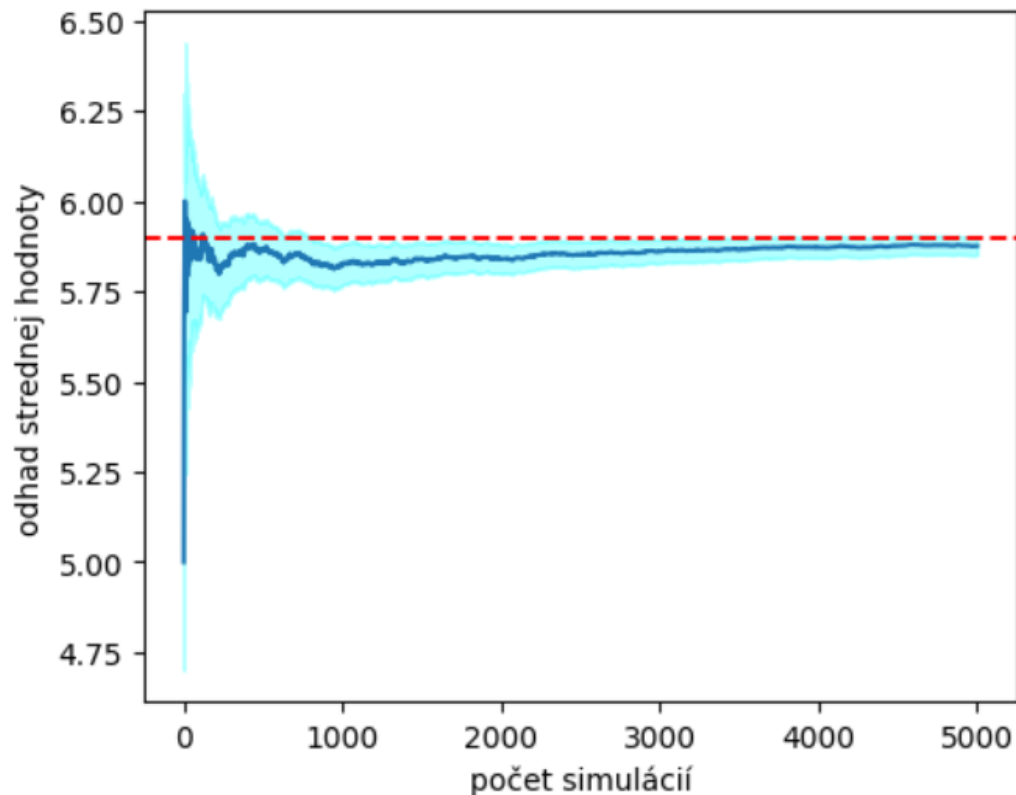
Príklad 3.

Do výťahu nastúpilo na prízemí osemposchodovej budovy 10 ľudí. Každý vystupuje s rovnakou pravdepodobnosťou na každom z poschodí, nezávisle od ostatných. Aká je stredná hodnota počtu poschodí, na ktorých výťah zastane, lebo na tomto poschodí niekto vystupuje?

Výsledok:



V tomto prípade vieme pomocou indikátorových premenných získať aj presnú hodnotu: $8 \cdot (1 - (7/8)^{10})$, čo je 5.8954.



O tejto konvergencii hovoria zákony veľkých čísel:

Nech X_i sú nezávislé rovnako rozdelené náhodné premenné so strednou hodnotou μ a $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ je priemer vypočítaný z prvých n hodnôt. Potom platí:

- **Slabý zákon veľkých čísel:** Pre ľubovoľné $\varepsilon > 0$ je

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) = 0.$$

- **Silný zákon veľkých čísel:**

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu\right) = 1.$$

Názvy “silný” a “slabý” pochádzajú z toho, že ak pre postupnosť nejakých náhodných premenných platí podmienka zo silného zákona, tak platí aj podmienka zo slabého zákona. Slovo “veľkých čísel” v názve sa vzťahuje k tomu, že ide o limity, kde máme veľa náhodných premenných.

V oboch prípadoch ide o konvergenciu priemeru k strednej hodnote, ak je priemer počítaný z veľkého počtu dát.

Ak stredná hodnota neexistuje, uvedené tvrdenia nemajú zmysel. V takom prípade sa priemer nemusí stabilizovať. Typickým príkladom je Cauchyho rozdelenie.

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.cauchy.html>

```
from scipy.stats import cauchy
E_cauchy = cauchy.moment(1)
print(E_cauchy)
```

nan

Prečo **nan**: Cauchyho rozdelenie má hustotu

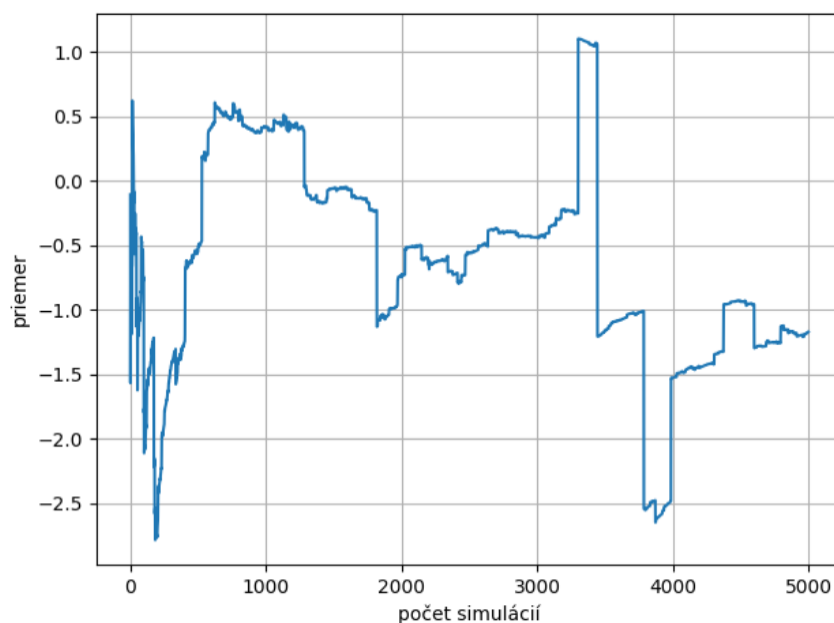
$$f(x) = \frac{1}{\pi(1+x^2)}$$

Pri výpočte strednej hodnoty dostaneme

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)} dx = \frac{1}{2\pi} [\ln(1+x^2)]_{-\infty}^{\infty}$$

Tu vzniká neurčitý výraz typu “nekonečno mínus nekonečno”, a teda stredná hodnota neexistuje.

Ako sa správa priemer:



Dá sa dokázať, že priemer z ľubovoľného počtu nezávislých náhodných premenných s Cauchyho rozdelením má znovu to isté Cauchyho rozdelenie.

Simulačne ukážka pre priemer zo 100 hodnôt:

```
import numpy as np
from scipy.stats import ks_1samp, cauchy

n_sim = 1000    # pocet simulacii
n_priemer = 100 # pocet nahodnych premennych pre vypocet priemeru

cauchy_priemer = lambda n: np.mean(cauchy.rvs(size=n))

priemery = [cauchy_priemer(n_priemer) for _ in range(n_sim)]

stat, p_hodnota = ks_1samp(priemery, cauchy.cdf)

print(f"KS štatistika: {stat:.5f}")
print(f"p-hodnota: {p_hodnota:.5f}")
```

KS štatistika: 0.03125
p-hodnota: 0.27710