Řešená cvičení: NMAI059 Pravděpodobnost a statistika 1

Karel Král, Martin Mareš, Robert Šámal

26. května 2021

Tento text není určen k šíření. Všechny chyby v tomto textu jsou samozřejmě záměrné. Reportujte je prosím na adresu `kralka@iuuk.mff.cuni...`.

Obsah

Kapitola 1

Zadání

1.1 Cvičení

1. Úvodní informace:
   (a) Slyšíte mě všichni dobře?
   (b) Literatura.
   (c) Pravidla zápočtu (domácí úkoly).

   Řešení: 1
2. Jak se generuje náhoda programem.
   - Python3
   - C++
   - R

   Řešení: 2
3. Připomeňte si definici pravděpodobnostního prostoru (Definice 2.1). Určete, co je
   - *množina elementárních jevů (sample space)*, tedy množina Ω,
   - *prostor jevů (event space)*, tedy množina $\mathcal{F} \subseteq \mathcal{P}(\Omega)$,
   - *pravděpodobnost (probability)*, tedy funkce $\Pr\colon \mathcal{F} \to [0, 1]$

   pro následující příklady:
   (a) Hod spravedlivou alkoholovou trojhrannou tužkou (není to kostka kvůli popisu prostoru jevů):

   ```
   import random
   drink = random.choice([
                   "světlé pivo",
                   "tmavé pivo",
                   "slivovice",
                   ])
   ```

   (b) Uniformně náhodné číslo z intervalu $[0, 1)$.

   ```
   import random
   print(random.random())
   ```

Řešení: 3

4. Dokažte, že $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$.

   Řešení: 4

5. Zopakujte si základní kombinatoriku:

   - Kolik je různých permutací množiny $\{A, B, C\}$?
   - Kolik různých slov skládajících se z písmen $\{A, B\}$ má délku 3?
   - Kolik různých podmnožin množiny $\{A, B, C, D, E\}$ má velikost 3?
   - Kolik různých kombinací s opakováním z množiny $\{A, B, C, D, E\}$ velikosti 3?

   Řešení: 5

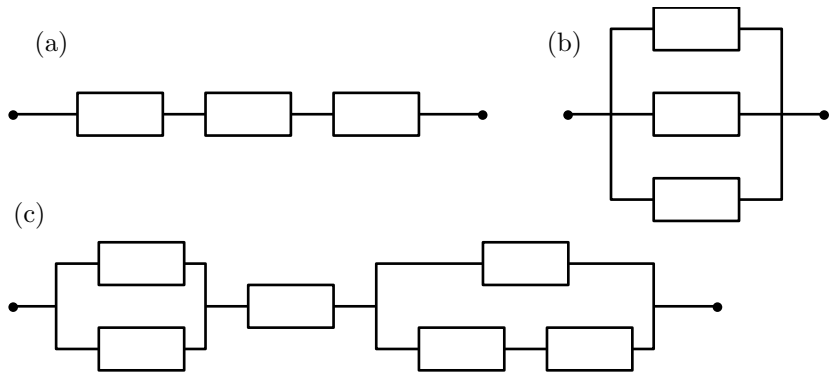
6. Jaká je pravděpodobnost, že při hodu šesti rozlišitelných spravedlivých šestistěnných kostek padnou aspoň na třech kostkách aspoň tři? Jaký je množina elementárních jevů, prostor jevů a pravděpodobnost?

   Řešení: 6

7. Nechť Ω jsou všechny permutace prvních 100 přirozených čísel, prostor jevů jsou všechny podmnožiny Ω a každý elementární jev je stejně pravděpodobný. Označme jev A_j že náhodně zvolená permutace $\pi \in \Omega$ splňuje $\pi(j) = j$ (pro $1 \leq j \leq 100$). Jsou A_1, A_2 nezávislé jevy?

   Řešení: 7

8. Každý obdélník na obrázku je součástka, která se může porouchat s pravděpodobností p. Přesněji řečeno porucha znamená, že skrz ní neteče proud. Poruchy součástek jsou na sobě nezávislé. Jaká je pravděpodobnost, že stále poteče proud mezi dvěma puntíky.



   Řešení: 8

1.2 Cvičení

1. Házíme cinknutou mincí – hlava padne s pravděpodobností $p \in [0, 1)$, orel padne s pravděpodobností $1 - p$. Házíme opakovaně dokud nepadne hlava.

   (a) Jak vypadá pravděpodobnostní prostor?

   (b) Jaká je pravděpodobnost, že hodíme právě třikrát (n-krát)?

   (c) Jaká je pravděpodobnost, že hodíme nejvýš třikrát (n-krát)?

   (d) Jaká je pravděpodobnost, že hodíme lišekrát?

(e) Simulujte předchozí.

Řešení: 1

2. Hodíme cinknutou korunou (panna s pravděpodobností $p_1 \in [0, 1]$) a cinknutou dvoukorunou (panna s pravděpodobností $p_2 \in [0, 1]$). Oba hody jsou na sobě nezávislé.

   (a) Určete pravděpodobnostní prostor.

   (b) Připoměňte si definici podmíněné pravděpodobnosti (Definice 2.2).

   (c) Spočítejte pravděpodobnost Pr[na obou padne panna | na koruně padne panna].

   (d) Spočítejte pravděpodobnost Pr[na obou padne panna | padne aspoň jedna panna].

   (e) Simulujte:

   Řešení: 2

3. Na louce rostou květiny, které mají buď bílé nebo červené květy. Náhodná květina má bílý květ s pravděpodobností $\Pr[B] = 0.6$, tedy pravděpodobnost že náhodná květina má červený květ je $\Pr[C] = 0.4$. Pravděpodobnost že červená květina je jedovatá je $\Pr[J \mid C] = 0.25$. Pravděpodobnost že bílá květina je jedovatá je $\Pr[J \mid B] = 1/12$. Snědli jsme náhodnou rostlinu a je nám zle, jaká je pravděpodobnost, že ta rostlina měla červený květ?

   Řešení: 3

4. V první krabici je b bílých míčků a c červených míčků, ve druhé krabici také. Napřed vytáhneme jeden míček z první krabice (uniformně náhodně) a dáme ho do druhé krabice. Pak vytáhneme jeden míček z druhé krabice (uniformně náhodně). Jaká je pravděpodobnost, že míček vytažený z druhé krabice je červený?

   (a) Navrhněte vhodný pravděpodobnostní prostor.

   (b) Spočítejte tu pravděpodobnost, kterou jsme chtěli.

   (c) Simulujte.

   Řešení: 4

5. Tento příklad je vymyšlený, zejména čísla nesedí a reálný svět je malinko složitější (tím se ještě budeme zabývat), ale informace v něm nejsou daleko od pravdy. V zemi nám řádí nemoc C.

   - Prostor elementárních jevů (sample space) Ω jsou všichni občané.
   - Označíme $C^+ \subseteq \Omega$ množinu všech lidí, kteří dnes mají aktivní nemoc C, označíme $C^- = \Omega \setminus C^+$ zdravé lidi.
   - Umíme uniformně náhodně samplovat lidi, tedy $\forall \omega \in \Omega\colon \Pr[\{\omega\}] = 1/|\Omega|$ (tady ω je jeden člověk).
   - Test nám pro libovolného člověka odpoví že je člověk zdravý nebo nemocný. Značme $T^+ \subseteq \Omega$ množinu lidí pro které test odpoví, že jsou nemocní. Značme $T^- = \Omega \setminus T^+$ množinu lidí pro které test odpoví, že jsou zdraví. Ale není to tak jednoduché, v příbalovém letáku testu se píše:
     - *Sensitivity:* (true positive) $\Pr[T^+ \mid C^+] = 0.9$
     - *Specificity:* (true negative) $\Pr[T^- \mid C^-] = 0.8$

     z tohoto můžeme odvodit chyby:

     - False positive = false alarm = type I error

     $$\Pr[T^+ \mid C^-] = 1 - \Pr[T^- \mid C^-] = 0.2$$

– False negative = miss = type II error

$$\Pr[T^- \mid C^+] = 1 - \Pr[T^+ \mid C^+] = 0.1$$

- Provedli jsme jeden test u každého z uniformně náhodně vybraných 50000 lidí a pozitivních testů vyšlo 1000. Tedy předpokládáme, že $\Pr[T^+] = \frac{1000}{50000} = \frac{1}{50} = 0.02$ (jak moc je tento předpoklad oprávněný budeme zkoumat nadále).
- Zajímá nás $\Pr[C^+|$ (vynásobeno 100 nám dá počet nemocných v procentech).

(a) V čem se toto liší od reality?

(b) Spočítejte $\Pr[C^+]$.

(c) Co se stalo špatně?

(d) Jak by vyšlo předchozí kdyby $\Pr[T^+ \mid C^+] = 0.99, \Pr[T^- \mid C^-] = 0.98, \Pr[T^+] = 0.2$?

(e) Simulujte předchozí.

Řešení: 5

6. V šuplíku mám $b \in \mathbb{N}$ párů bílých, $c \in \mathbb{N}$ párů černých ponožek a $s \in \mathbb{N}$ párů sepraných ponožek. Potřebuju si vytáhnout čtyři páry černých ponožek (jedu na prodloužený víkend tancovat). Když vytáhnu čtyři náhodné páry ponožek (mám je napárované v šuplíku), jaká je pravděpodobnost, že všechny budou černé?

   Řešení: 6

1.3 Cvičení

1. Rozmysleme si, proč nezávislost více jevů není to samé jako nezávislost po dvou.

   (a) Najděte jevy A, B, C takové, že jevy jsou po dvou nezávislé, ale $\Pr[A \cap B \cap C] \neq \Pr[A] \Pr[B] \Pr[C]$.

   (b) Najděte jevy A, B, C takové, že $\Pr[A \cap B \cap C] = \Pr[A] \Pr[B] \Pr[C]$, ale jevy nejsou po dvou nezávislé.

   Řešení: 1

2. Házíte dvěma rozlišitelnými kostkami.

   (a) Určete vhodný pravděpodobnostní prostor.

   (b) Spočítejte pravděpodobnost, že aspoň na jedné kostce padla šestka, když víte jaký součet padl.

   Řešení: 2

3. V truhle je sto mincí. Z nich 99 je normálních, ale jedna má na obou stranách orla.

   (a) Určete vhodný pravděpodobnostní prostor.

   (b) Vytáhneme náhodnou minci a šestkrát s ní hodíme, pokaždé padne orel. Jaká je pravděpodobnost, že jsme si vytáhli dvouorlovou minci? (Zkuste napřed odhadnout, pak spočítat.)

   (c) Simulujte.

   Řešení: 3

4. Připomeňme co je náhodná veličina a její střední hodnota.

   Co kdyby pravděpodobnost kostky nebyla uniformní?

Řešení: 4

5. Na stole jsou dvě obálky, v jedné je k korun, ve druhé ℓ korun ($k, \ell \in \mathbb{N}$). Můžete otevřít jednu obálku a na základě sumy v ní se rozhodnout jestli si necháte tu otevřenou nebo si vezmete tu druhou (nehledě na to, kolik je v té druhé). Umíte vymyslet způsob jak odejít s tou s větším obnosem s pravděpodobností ostře větší než jedna polovina? Určete střední hodnotu výhry.

   Řešení: 5

6. Spočítejte střední počet porovnání quick-sortu:

```
from random import randint


def partition(arr, begin, end):
    pivot_i = randint(begin, end - 1)
    (arr[pivot_i], arr[end-1]) = (arr[end-1], arr[pivot_i])
    pivot = arr[end - 1]
    i = begin
    for j in range(begin, end):
        if arr[j] < pivot:
            (arr[i], arr[j]) = (arr[j], arr[i])
            i += 1
    (arr[i], arr[end-1]) = (arr[end-1], arr[i])
    return i


def quick_sort(arr, begin, end):
    if end <= begin:
        return
    p = partition(arr, begin, end)
    quick_sort(arr, begin, p)
    quick_sort(arr, p+1, end)
```

   (a) Uvědomte si, že každá dvě čísla porovnáte nejvýš jednou (pokud se žádné číslo neopakuje). Pro jednoduchost budeme předpokládat, že se čísla neopakují (jinak bychom museli mluvit o jejich pozici v utříděném poli).

   (b) Vytvořte vhodný pravděpodobnostní prostor.

   (c) Definujte náhodnou proměnnou určující počet porovaných dvojic, vyjádřete ji jako součet jednodušších a použijte větu o linearitě střední hodnoty.

   (d) S jakou pravděpodobností provede quick-sort aspoň $10n \ln(n)$ porovnání?

   - Sčítáme n kladných reálných čísel $a_1, a_2, \ldots, a_n \in [0, \infty)$. Víme, že

$$\sum_{i=1}^{n} a_i = S$$

     Pro kolik z těch čísel platí $a_j \geq 5S/n$?

   - Co kdybychom ta čísla sčítali váženě? Tedy formálně: mějme náhodnou proměnnou o které víme $\Pr[X = j]$ pro $j \in \{1, 2, \ldots, m\}$ (kde pro jednoduchost předpokládáme $\sum_{j=1}^{m} \Pr[X = j] = 1$, tedy že X má hodnoty $1, 2, \ldots, m$). Víme $\mathbb{E}[X] = \sum_{j=1}^{m} j \Pr[X = j] = S$. Jaká je pravděpodobnost $\Pr[X \geq 5S]$?

   - Gratuluji, vymysleli jste Markovovu nerovnost.

- Často bývá mnohem jednodušší použít Markovovu nerovnost než přímo počítat pravděpodobnost. Občas můžeme dostat i silnější odhady pomocí Čebyševovy nebo Černovovy nerovnosti. Na to budeme potřebovat rozptyl a další znalosti o náhodných proměnných.

Řešení: 6

1.4 Cvičení

1. Házím míčem na koš. V každém pokusu mám pravděpodobnost p že se trefím (jednotlivé hody jsou nezávislé). Skončím po prvním zásahu. Označme X celkový počet hodů.
   (a) Jaké je pravděpodobnostní rozdělení X (tj. distribuce)? Jinak řečeno určete pravděpodobnostní funkci p_X (tj. pro každé x určete $p_X(x) = \Pr[X = x]$).
   (b) Jaká je $\Pr[X \geq 10 \mid X \geq 5]$?
   (c) Jaká je $\mathbb{E}[X]$?
   (d) Jaká je $\mathbb{E}[X \mid X$ je sudé$]$?
   (e) Simulujte.

   Řešení: 1

2. V testu je 20 otázek s volbami a,b,c,d. Za správnou odpověď (vždy je jen jedna odpověď správná) je 1 bod, za špatnou $-1/4$ bodu, za nevyplněnou otázku nula. Každá otázka je s pravděpodobností p jednou z těch, co se Kvído naučil a tedy zná správnou odpověď. Pokud správnou odpověď nezná, ví o tom, a může se rozhodnout, zda tipovat.
   (a) Jaká je střední hodnota počtu bodů, které Kvído získá, pokud bude odpovídat jenom otázky, u kterých zná odpověď?
   (b) A co když bude tipovat, když nezná správnou odpověď?
   (c) Jak by se musela změnit penalizace za chybnou odpověď, aby byly odpovědi v částech a, b stejné?
   (d) Simulujte.

   Řešení: 2

3. Ze standardního balíčku s 52 kartami vytáhneme dvě karty. Označíme X počet vytažených es, Y počet králů. Určete sdruženou pravděpodobnostní funkci $p_{X,Y}$ a také marginální pstní funkce p_X, p_Y.

   Řešení: 3

4. Chceme nasbírat všechny z n druhů kuponů. Můžeme si koupit jeden kupon, který má uniformně náhodný druh. Kolikrát musíme koupit kupon, než posbíráme všechny?
   (a) Jaká je střední hodnota počtu koupených kuponů, než nasbíráme všechny?
   (b) Simulujte.

   Řešení: 4

5. Připomeňte si, co je náhodná veličina, jaké máme typy náhodných veličin a jaké jsou jejich distribuce. A hlavně co vyjadřují.
   (a) Bernoulli
   (b) binomické

(c) hypergeometrické

(d) geometrické

(e) Poissonovo

```
import matplotlib.pyplot as plt
from collections import Counter
from random import randint
from random import random
from random import sample
from numpy import random as npr

p = 0.3
n = 10
k = 5
S = 20
l = 10


def bernoulli(pr: float = p) -> bool:
    return int(random() < pr)


def geometric(pr: float = p) -> int:
    """pr is success probability, return the number of tosses until
    the first success."""
    assert pr > 0
    sample = 1
    fail_pr = 1 - pr
    while random() < fail_pr:
        sample += 1
    return sample


def binomicke(n=n, pr=p):
    return sum(bernoulli(pr) for _ in range(n))


def hypergeometric(n=n, N=S, k=k):
    return sum(sample([1]*k + [0]*(N-k), k=n))


def poisson(l=l):
    return npr.poisson(lam=l, size=1)[0]


N = 100000


def expected_value(X):
    """Vraci stredni hodnotu."""
    return sum(X() for _ in range(N)) / N
```

```
def variance(X):
    """Vraci rozptyl."""
    EX = expected_value(X)
    return sum((X() - EX)**2 for _ in range(N)) / N
    # return (sum(X()**2 for _ in range(N)) / N) - EX**2


def histogram(X, strX, fig):
    cnt = Counter(X() for _ in range(N))
    distribution = {}
    for c in cnt:
        distribution[c] = cnt[c] / N

    plt.figure(fig)
    plt.bar(distribution.keys(), distribution.values())
    # plt.show()
    # plt.xlabel("")
    # plt.ylabel("")
    plt.savefig(f'{strX}.pdf')


print('Bernoulli:')
print(f'E[X] = {expected_value(bernoulli)} (= {p})')
print(f'var[X] = {variance(bernoulli)} (= {p*(1-p)})')
histogram(bernoulli, "bernoulli", 0)
print('')


print('binomické')
print(f'E[X] = {expected_value(binomicke)} (= {n*p})')
print(f'var[X] = {variance(binomicke)} (= {n*p*(1-p)})')
histogram(binomicke, "binomicke", 1)
print('')


print('hypergeometrické')
print(f'E[X] = {expected_value(hypergeometric)} (= {n*k/S})')
print(f'var[X] = {variance(hypergeometric)} (= {n*(k/S)*(1-(k/S))*(S-n)/(S-1)})')
histogram(hypergeometric, "hypergeometric", 2)
print('')


print('geometrické')
print(f'E[X] = {expected_value(geometric)} (= {1/p})')
print(f'var[X] = {variance(geometric)} (= {(1-p)/p**2})')
histogram(geometric, "geometric", 3)
print('')


print('Poissonovo')
print(f'E[X] = {expected_value(poisson)} (= {l})')
print(f'var[X] = {variance(poisson)} (= {l})')
histogram(poisson, "poisson", 4)
```

```
# Možný výstup:
# Bernoulli:
# E[X] = 0.30003 (= 0.3)
# var[X] = 0.21018846799996171 (= 0.21)

# binomické
# E[X] = 3.00221 (= 3.0)
# var[X] = 2.111969109999777 (= 2.0999999999999996)

# hypergeometrické
# E[X] = 2.49898 (= 2.5)
# var[X] = 0.9866719879994746 (= 0.9868421052631579)

# geometrické
# E[X] = 3.33374 (= 3.3333333333333335)
# var[X] = 7.851098870500136 (= 7.777777777777778)

# Poissonovo
# E[X] = 10.00683 (= 10)
# var[X] = 9.947825008000336 (= 10)
```

Řešení: 5

1.5 Cvičení

1. Hodíme třikrát mincí. Označíme X počet rubů v prvních dvou hodech a Y počet líců v posledních dvou hodech.
   (a) Určete pravděpodobnostní prostor.
   (b) Určete předpis našich náhodných veličin.
   (c) Určete sdruženou pravděpodobnostní funkci $p_{X,Y}$ a také marginální pravděpodobnostní funkce p_X, p_Y.
   (d) Určete distribuční funkce $F_X, F_Y, F_{X,Y}$ (cumulative distribution function CDF).
   (e) Jsou X a Y nezávislé?
   (f) Určete $\Pr[X < Y]$.
   (g) Určete podmíněnou pravděpodobnostní funkce $p_{X|Y}$.
   (h) Simulujte.

   Řešení: 1
2. Bonusový příklad: tady si zadefinujeme Lebesgueovu míru a integrál.
   (a) Definujte Lebesgueovu míru na $\mathbb{R}$.
   (b) Dokažte, že pro diskrétní náhodnou veličinu, takovou že $\mathrm{Im}(X) \subseteq \mathbb{N}$, platí:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \mathrm{Im}(X)} x \Pr[X = x] \\ &= \sum_{n \in \mathbb{N}} n \Pr[X = n] \\ &= \sum_{n \in \mathbb{N}} \Pr[X \geq n]\end{aligned}$$

(c) Definujte Lebesgueův integrál (pro danou míru Pr, obecně to ani nemusí být pravděpodobnostní míra, ale obecná μ).

Řešení: 2

3. Pro následující náhodné veličiny určete:

(a)

$$\begin{aligned} \Omega &= \{1,2,3\} \\ \mathcal{F} &= \mathcal{P}(\Omega) \\ \Pr[\{\omega\}] &= 1/3 \qquad \text{(pro libovolné } \omega \in \Omega\text{, zbytek určen jednoznačně)} \\ X &: \Omega \to \mathbb{R} \\ X(\omega) &= 2\omega + 1 \end{aligned}$$

- Je to náhodná veličina na daném pravděpodobnostním prostoru?
- Je to diskrétní náhodná veličina?
- Jaká je její střední hodnota?
- Jaká je její distribuční funkce?
- Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?
- Jaká je její kvantilová funkce?

(b)

$$\begin{aligned} \Omega &= \mathbb{N} = \{1,2,3,\ldots\} \\ \mathcal{F} &= \mathcal{P}(\Omega) \\ \Pr[\{n\}] &= 1/2^n \qquad \text{(pro libovolné } n \in \mathbb{N}\text{, zbytek určen jednoznačně)} \\ X &: \Omega \to \mathbb{R} \\ X(\omega) &= 2\omega + 1 \end{aligned}$$

- Je to náhodná veličina na daném pravděpodobnostním prostoru?
- Je to diskrétní náhodná veličina?
- Jaká je její střední hodnota?
- Jaká je její distribuční funkce?
- Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?
- Jaká je její kvantilová funkce?

(c)

$$\begin{aligned} \Omega &= [0,1] \\ \mathcal{F} &= \text{Lebesgueovsky měřitelné množiny} \\ \Pr[A] &= \lambda(A) \qquad \text{(pro libovolné } A \in \mathcal{F}\text{)} \\ X &: \Omega \to \mathbb{R} \\ X(\omega) &= \lceil 10x \rceil \end{aligned}$$

- Je to náhodná veličina na daném pravděpodobnostním prostoru?
- Je to diskrétní náhodná veličina?

- Jaká je její střední hodnota?
- Jaká je její distribuční funkce?
- Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?
- Jaká je její kvantilová funkce?

(d)

$$\begin{aligned} \Omega &= [0,1] \\ \mathcal{F} &= \text{Lebesgueovsky měřitelné množiny} \\ \Pr[A] &= \lambda(A) && \text{(pro libovolné } A \in \mathcal{F}) \\ X &\colon \Omega \to \mathbb{R} \\ X(\omega) &= 10x \end{aligned}$$

- Je to náhodná veličina na daném pravděpodobnostním prostoru?
- Je to diskrétní náhodná veličina?
- Jaká je její distribuční funkce?
- Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?
- Jaká je její střední hodnota?
- Jaká je její kvantilová funkce?

(e)

$$\begin{aligned} \Omega &= [0,1] \\ \mathcal{F} &= \text{Lebesgueovsky měřitelné množiny} \\ \Pr[A] &= \lambda(A)/2 + 1/2 && \text{(pro libovolné } A \in \mathcal{F} \text{ pokud } 0.1 \in A) \\ \Pr[A] &= \lambda(A)/2 && \text{(pro libovolné } A \in \mathcal{F} \text{ pokud } 0.1 \notin A) \\ X &\colon \Omega \to \mathbb{R} \\ X(\omega) &= 10\omega \end{aligned}$$

- Je to náhodná veličina na daném pravděpodobnostním prostoru?
- Je to diskrétní náhodná veličina?
- Jaká je její distribuční funkce?
- Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?
- Jaká je její střední hodnota?
- Jaká je její kvantilová funkce?

(f) Pozorujte, že kvantil je jediná funkce, pro kterou platí:

$$\forall p \in [0,1], \forall x \in \mathbb{R} \colon Q_X(p) \leq x \Leftrightarrow p \leq F_X(x)$$

Řešení: 3

1.6 Cvičení

1. - Jedno promile lidí má nemoc C, značíme $C^+ \subseteq \Omega$ množinu nemocných a $C^- = \Omega \setminus C^+$ množinu zdravých. Člověka volíme uniformně náhodně, pak $\Pr[C^+] = 0.001$.
   - Máme test, který označí množinu lidí $T^+ \subseteq \Omega$ za nemocné a $T^- = \Omega \setminus T^+$ za zdravé. Test má následující parametry:
     - *Sensitivita:* (true positive) $\Pr[T^+ \mid C^+] = 0.99$
     - *Specificita:* (true negative) $\Pr[T^- \mid C^-] = 0.98$

   (a) Uniformně náhodně jsme vybrali člověka c. Jaká je pravděpodobnost, že $c \in C^+$ (tedy že je nemocný)?

   (b) Člověku c jsme udělali jeden test a ten vyšel pozitivní. Jaká je pravděpodobnost, že $c \in C^+$ (tedy že je nemocný)?

   (c) Pro jistotu jsme člověku c udělali ještě jeden test a ten vyšel znovu pozitivní. Jaká je pravděpodobnost, že $c \in C^+$ (tedy že je nemocný)? Předpokládejte, že výsledek druhého testu je nezávislý na tom prvním.

   (d) Simulujte.

   Poznámka: Toto je případ kdy je nemoc velice zřídkavá a ty nemáš symptomy. Pokud máš symptomy a jdeš na test, tak nejsi náhodně vybraný člověk! Tento příklad je tedy spíš situace jdu darovat krev a oni musí udělat povinný test na HIV a ten vyjde pozitivní, ale přesto bych se neměl tolik strachovat, ale v klidu jít na druhý test (už to že chodím darovat krev negativně koreluje s nakažením HIV).

   Řešení: 1

2. Pojďme se podívat na další vlastnosti jednotlivých rozdělení:

   (a) Bernoulli:
   - Kde se použije?
   - Jakou má střední hodnotu a rozptyl?
   - Co se stane, když sečtu dvě nebo obecně n-náhodných veličin $X_1, X_2, \ldots, X_n$, které jsou nezávislé a pro nějaké fixní $p \in [0, 1]$ a každé $j \in [n]$ platí že $X_j \sim Bern(p)$.

   (b) geometrické:
   - Kde se použije?
   - Jakou má střední hodnotu a rozptyl?
   - Co se stane, když vezmeme minimum ze dvou nezávislých náhodných veličin X_1, X_2, které jsou kde $X_1 \sim Geom(p_1)$ a $X_2 \sim Geom(p_2)$?

   (c) Binomické:
   - Kde se použije?
   - Jakou má střední hodnotu a rozptyl?
   - Co se stane, když sečtu dvě náhodné veličiny X_1, X_2, které jsou nezávislé a pro nějaké fixní $p \in [0, 1]$ platí $X_1 \sim Bin(n, p)$ a $X_2 \sim Bin(m, p)$ (kde $n, m \in \mathbb{N}$).

   (d) Poissonovo:
   - Kde se použije?
   - Jakou má střední hodnotu a rozptyl?

- Co se stane, když sečtu dvě náhodné veličiny X_1, X_2, které jsou nezávislé a $X_1 \sim Pois(\lambda)$, a $X_2 \sim Pois(\mu)$.
- Co se stane, pokud zvolíme $\lambda = np$ a porovnáme Poissonovo rozdělení a binomiální rozdělení?
- Pro odvážné: ukažte, že pokud $X_1 \sim Pois(\lambda_1)$ a $X_2 \sim Pois(\lambda_2)$ jsou nezávislé, pak $\Pr[X_1 = k \mid X_1 + X_2 = n] = \Pr[Y = k]$ kde $Y \sim Bin(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$.

Řešení: 2

3. de Mèreho problém házeme spravedlivými šestistrannými kostkami:
   (a) Jaká je pravděpodobnost, že padne ze čtyř hodů aspoň jedna šestka?
   (b) Jaká je pravděpodobnost, že padne z 24 hodů dvojicí kostek aspoň jedna dvojitá šestka?
   (c) De Mèreho problém spočíval v tom, jestli jsou odpovědi na části (a), (b) stejné. Tak se Chevalier de Mère zeptal Blaise Pascala a ten spolu s Pierre de Fermatem položili základy teorie pravděpodobnosti.
   (d) Umíte úlohu interpretovat jako otázku o binomickém rozdělení?
   (e) Umíte úlohu interpretovat jako otázku o geometrickém rozdělení?

   Řešení: 3
4. Odhadněte π pomocí náhodných pokusů.

   Řešení: 4
5. Nechť X je náhodná veličina. Vyjádřete pomocí $F_X(t) = \Pr[X \leq t]$ pro každé $t \in \mathbb{R}$ distribuční funkci náhodných veličin:
   (a) $X^+ = \max(0, X)$
   (b) $-X$
   (c) $X^- = -\min(X, 0)$
   (d) $|X|$

   Řešení: 5
6. Mějme spojitou náhodnou veličinu X danou její pravděpodobnostní hustotou (probability density function – PDF)

$$f_X(t) = \begin{cases} 0 & t < 1 \\ 2/x^3 & t \geq 1 \end{cases}$$

   (a) Spočítejte $\Pr[X \in [5, 10]]$.
   (b) Spočítejte $\Pr[X \geq 100]$.
   (c) Spočítejte $\mathbb{E}[X]$.
   (d) Určete distribuční funkci (cumulative distribution function – CDF).

   Řešení: 6
7. Mějme spojitou nezápornou náhodnou veličinu X, která má hustotu (probability density function – PDF) f_x (a tedy comulative distribution function – CDF $F_X(t) = \int_0^t f_X(x)\ dx$). Ukažte, že $\mathbb{E}[X] = \int_0^\infty x f_X(x)\ dx = \int_\Omega F_X(\omega) d\Pr(\omega) = \int X d\Pr$.

   Pokud by X nebyla nezáporná, tak ji můžeme vyjádřit jako rozdíl dvou nezáporných náhodných veličin.

Řešení: 7

1.7 Cvičení

1. Spočítejte střední hodnotu následujících nezáporných náhodných veličin pomocí vzorce

$$\mathbb{E}[X] = \int_0^\infty 1 - F_X(t)\ dt \qquad \text{(pro nezápornou náhodnou veličinu } X\text{)}$$

   (a) $\mathrm{Im}(X) = \{2, \pi\}$, $\Pr[X = 2] = 1/3, \Pr[X = \pi] = 2/3$

   (b) X je dána hustotou (probability density function) $f_X(t) = 1/3$ pokud $t \in [0, 3]$ a jinak $f_X(t) = 0$.

   Řešení: 1

2. (a) Máme minci, kde padne hlava s pravděpodobností $p \in (0, 1)$ (ale my ani neznáme p). Jak pomocí ní vygenerujeme hod spravedlivou mincí?

   (b) Máme spravedlivou minci, jak pomocí ní vygenerujete hod mincí, kde padne hlava s pravděpodobností $p \in (0, 1)$?

   (c) Máme možnost generovat $Y \sim U(0, 1)$. Jak pomocí toho vygenerujeme hod mincí kde padne hlava s pravděpodobností p?

   (d) Máme diskrétní náhodnou veličinu $\mathrm{Im}(X) = \{x_1, x_2, x_3, x_4\}$ (nechť $x_j < x_i$ pro $j < i$). Známe $\Pr[X = x_j] = p_j$ a tím pádem známe všechny běžné parametry (p_X, F_X, Q_X). Máme možnost generovat $Y \sim U(0, 1)$. Jak pomocí výsledku Y vygenerujeme výsledek X?

   Řešení: 2

3. Autobus přijede v čas $e \doteq 2.71$. Čas mého příchodu na zastávku je náhodná veličina X s hustotou

$$f_X(t) = \begin{cases} 1/t & \text{pokud } t \in [1, e] \\ 0 & \text{jinak} \end{cases}$$

   (a) Jaká je pravděpodobnost, že přijdu v některý čas z intervalu $[1.5, 2]$?

   (b) Jaká je distribuční funkce času mého příchodu (cumulative distribution function)?

   (c) Jaká je kvantilová funkce času mého příchodu?

   (d) Jaká je střední hodnota času mého příchodu?

   (e) Jaký je rozptyl času mého příchodu?

   (f) Jaká je střední doba čekání na autobus?

   (g) Simulujte.

   Řešení: 3

4. Knihovna MFF má 1000 čtenářů – studentů informatiky – a rozhoduje se, kolik kopií nové knihy koupit. Předpokládejme, že o knihu má v daný semestr každý student zájem s pravděpodobností $p = 0.01$, nezávisle na ostatních.

   (a) Určete pravděpodobnostní funkci pro počet studentů, kteří mají o knihu zájem.

   (b) Určete pravděpodobnostní funkci pro Poissonovskou aproximaci tohoto počtu.

(c) Jaká je pravděpodobnost, že 20 kopií knihy nestačí? Vyjádřete jednak pomocí distribuční funkce, jednak pomocí sumy. A také jednak pomocí přesné formule z části (a), jednak pomocí aproximace z části (b).

(d) Je popsaný model zájmu studentů o knihy realistický?

Řešení: 4

5. *Exponenciální rozdělení* je spojitou analogií rozdělení geometrického. Vyjadřuje dobu čekání na první událost generovanou poissonovským procesem (s daným parametrem λ). Náhodná veličina $X \sim Exp(\lambda)$ má distribuční funkci

$$F_X(t) = \begin{cases} 1 - e^{-\lambda t} & \text{pro } t \geq 0, \\ 0 & \text{pro } t < 0. \end{cases}$$

Vypočítejte:

(a) hustotní funkci $f_X(t)$

(b) střední hodnotu $\mathbb{E}[X]$

(c) rozptyl $\mathrm{var}(X)$

Řešení: 5

1.8 Cvičení

1. (a) Nechť X je náhodná veličina a $X \geq 0$ skoro jistě (tzn $\Pr[X \geq 0] = 1$). Najděte nějakou takovou náhodnou veličinu, která je netriviální.

(b) Nechť X je náhodná veličina a $X \geq 0$ skoro jistě (tzn $\Pr[X \geq 0] = 1$). Dokažte, že pokud $\mathbb{E}[X]$ existuje, tak $\mathbb{E}[X] \geq 0$.

(c) Nechť Y, Z jsou náhodné veličiny a $Y \leq Z$ skoro jistě. Dokažte, že pokud $\mathbb{E}[Y]$, $\mathbb{E}[Z]$ existují, tak $\mathbb{E}[Y] \leq \mathbb{E}[Z]$.

(d) Dokažte Markovovu nerovnost $\Pr[X \geq a\mathbb{E}[X]] \leq 1/a$ pro nezápornou náhodnou veličinu X a libovolné $a \geq 1$.

Řešení: 1

2. Bublifukem vyfoukneme bublinu o poloměru $R \sim U(1, 5)$. Jaká je střední hodnota povrchu bubliny?

Řešení: 2

3. Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením se střední hodnotou μ a rozptylem σ^2. Označme $S_n = (X_1 + \cdots + X_n)/n$. To můžeme považovat za odhad střední hodnoty μ průměrem z n pokusů.

(a) Určete $\mathbb{E}[S_n]$ a $\mathrm{var}(S_n)$.

(b) Ukažte, jak lze počítat S_n z S_{n-1}, X_n a n.

(c) Použijte vhodné X_i, aby μ obsahovalo číslo π. Sestavte program v libovolném jazyce a spočítejte pomocí něj hodnotu π. (Jak velké n myslíte, že bude potřeba pro pět správných číslic?)

Řešení: 3

4. Předpokládejme, že u poštovní přepážky trvá vyřízení jednoho zákazníka čas, který má exponenciální rozdělení a střední hodnotu 4 minuty.

(a) Jaký je parametr λ, jaká je distribuční funkce?

(b) Jaká je pravděpodobnost, že budeme čekat více než 4 minuty?

(c) Jaká je pravděpodobnost, že budeme čekat něco mezi 3 a 5 minutami?

(d) Simulujte.

Řešení: 4

5. Říkáme, že náhodná veličina X (resp. její rozdělení) *nemá paměť*, pokud

$$\Pr[X > s + t \mid X > s] = \Pr[X > t]$$

pro $s, t \geq 0$. Jinými slovy, doba, kterou jsme již čekali, nemá vliv na dobu, kterou budeme ještě čekat.

(a) Ukažte, že geometrické rozdělení nemá paměť.

(b) Co z toho plyne o rozložení dalšího hodu, když už nám pětkrát v řadě padla hlava?

(c) Ukažte, že exponenciální rozdělení nemá paměť.

(d) Simulujte.

Platí dokonce, že je to jediné spojité rozdělení na kladných čísel bez paměti (a geometrické je jediné diskrétní bez paměti), ale to dokazovat nemusíte.

Řešení: 5

6. Budeme modelovat množství sněhu, který bude na Silvestra v lyžarském areálu Ještěd, pomocí normálního rozdělení se střední hodnotou 40 (centimetrů) a směrodatnou odchylkou 10.

(a) Jaká je pravděpodobnost, že nám model určí zápornou hodnotu sněhové pokrývky?

(b) Jaká je pravděpodobnost, že sněhu napadne 50–70 cm?

(c) Simulujte.

Hodnoty $\Phi(x)$ si spočítejte v Pythonu nebo v R, případně se podívejte do tabulky na `https://en.wikipedia.org/wiki/Standard_normal_table` (sekce Cummulative).

Řešení: 6

7. Plutonium-238 má poločas rozpadu 87.7 let. Jeho rozpad budeme modelovat pomocí exponenciálního rozdělení: pro každý atom budeme čas, za který se rozpadne, považovat za nezávislou náhodnou veličinu s rozdělením $Exp(\lambda)$.

(a) Jaké je λ?

(b) Jaká je střední doba života atomu ^{238}Pu?

(c) Po jaké době se rozpadne 90 % atomů?

(d) Kolik procent atomů se rozpadne po 50 letech? (Mimochodem, některé kosmické sondy a některé kardiostimulátory používají rozpad ^{238}Pu jako zdroj energie.)

(e) Simulujte.

Řešení: 7

8. Dostali jsme minci. Nevíme jestli je spravedlivá (tzn neznáme pravděpodobnost, že padne hlava). Tisíckrát jsme s ní hodili a padlo 345 hlav.

(a) Jaká je pravděpodobnost, že na spravedlivé minci padne přesně 345 hlav?

(b) Pokud $p \in (0,1)$ je proměnná vyjadřující pravděpodobnost, že na naší minci padne hlava. Vyjádřete pravděpodobnost, že padne 345 hlav jako funkci p, tedy

$$P(p) = \Pr[X = 345 \text{ kde } X \sim Bin(1000, p)]$$

(c) Pokud tedy modelujeme hod naší mincí jako $Bin(1000, p)$, pak určete p, které dává nejvyšší pravděpodobnost pozorovaného výsledku.

(d) Porovnejte s tím, jak jsme doteď simulovali.

Řešení: 8

1.9 Cvičení

1. Dostali jsme minci. Nevíme jestli je spravedlivá (tzn neznáme pravděpodobnost, že padne hlava). Tisíckrát jsme s ní hodili a padlo 345 hlav.

   (a) Jaká je pravděpodobnost, že na spravedlivé minci padne přesně 345 hlav?

   (b) Pokud $p \in (0,1)$ je proměnná vyjadřující pravděpodobnost, že na naší minci padne hlava. Vyjádřete pravděpodobnost, že padne 345 hlav jako funkci p, tedy

   $$P(p) = \Pr[X = 345 \text{ kde } X \sim Bin(1000, p)]$$

   (c) Pokud tedy modelujeme hod naší mincí jako $Bin(1000, p)$, pak určete p, které dává nejvyšší pravděpodobnost pozorovaného výsledku. Tomuto se ve statistice říká Maximum Likelihood Estimation (MLE).

   (d) Porovnejte s tím, jak jsme doteď simulovali.

   Řešení: 1

2. Za druhé světové války spojenci potřebovali vědět, kolik tanků Německo vyrobilo. Němci byli precizní a své tanky číslovali popořadě (pokud vyrobili n-tý tank, tak měl na motoru napsané číslo n).

   (a) Zajali jste uniformně náhodný tank, který měl číslo m (připomeňme, že neznáte n), jak odhadnete $\hat{n}$ (tj. odhad skutečného počtu n) pomocí MLE?

   (b) Je to, co nám vyšlo rozumný odhad?

   (c) Je možné, že se k tomuto problému ještě vrátíme s jinými statistickými odhady. Pokud jste příliš zvědaví, tak `https://en.wikipedia.org/wiki/German_tank_problem`

   Řešení: 2

3. Nechť $Z \sim N(0,1)$. Pomocí tabulky funkce Φ ověřte pravidlo 3σ, neboli spočtěte

x	-4	-3	
$\Phi(x)$	0.00003	0.00135	0.

   Další hodnoty viz `https://en.wikipedia.org/wiki/Standard_normal_table` – sekce Cumulative nebo použitím `scipy.stats.norm.cdf` `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html`

   (a) Připomeňte pravidlo 3σ.

   (b) $\Pr[|Z| \leq 1]$

   (c) $\Pr[|Z| \leq 2]$

   (d) $\Pr[|Z| \leq 3]$

   (e) Spočítejte to programem:

(f) Přepište, co to znamená pro n.v. $X \sim N(\mu, \sigma^2)$

Řešení: 3

4. Nechť X, Y mají sdruženou hustotu $f_{X,Y}(x, y) = e^{-x-y}$ pro $x, y > 0$ (a 0 jinak).

   (a) Určete marginální hustoty f_X, f_Y.

   (b) Určete také distribuční funkce F_X, F_Y, $F_{X,Y}$.

   (c) Jsou X, Y nezávislé?

   (d) Najděte $\Pr[X + Y \leq 1]$.

   (e) Najděte $\mathbb{E}[X + Y]$.

   (f) Najděte $\Pr[X > Y]$.

   (g) Simulujte předchozí tři body.

   Řešení: 4

5. Volme uniformně náhodně bod z polokruhu o poloměru 1, se středem v počátku a v horní polorovině. (Uniformně znamená, že pravděpodobnost každé podmnožiny je úměrná jejímu obsahu.) Označme X, Y souřadnice zvoleného bodu.

   (a) Najděte sdruženou hustotu $f_{X,Y}$.

   (b) Najděte marginální hustotu f_Y a spočtěte pomocí ní $\mathbb{E}[Y]$.

   (c) Pro kontrolu spočtěte $\mathbb{E}[Y]$ přímo (pomocí pravidla LOTUS).

   (d) Simulujte.

   Řešení: 5

1.10 Cvičení

1. Máme dvě mince. Jedna je spravedlivá, na druhé padá hlava s pravděpodobností $\Pr[\text{hlava}] = 1/4$. Ale nevíme která je která. Vymyslete algoritmus jak ty dvě mince rozlišit.

   - Vezmeme minci a hodíme n-krát.
   - Nechť $\hat{p}$ je pravděpodobnost, že padla hlava (počet hlav děleno n).
   - Pokud $\hat{p} \geq 3/8$ řekneme, že je férová, jinak cinklá.

   Ukažte, že pro fixní konstantu $\varepsilon \in (0, 1)$ pokud $n \geq 32 \ln(2/\varepsilon)$ náš algoritmus odpoví správně s pravděpodobností aspoň $1 - \varepsilon$.

   Řešení: 1

2. (Problém šatnářky) Náhodně přiřadíme n klobouků n lidem. Označíme X_i indikátor jevu „i-tý člověk dostal svůj klobouk“ a položíme $X = \sum_{i=1}^{n} X_i$.

   (a) Určete $\mathbb{E}[X]$.

   (b) Určete $\operatorname{var}(X)$.

   (c) Určete σ_X.

   (d) Použijte Čebyševovu nerovnost na odhad pravděpodobnosti, že $X \geq 3$ (pro $n \geq 3$).

   (e) Co by nám řekl Markov?

   (f) Co by nám řekl Černov?

(g) Simulujte.

Řešení: 2

3. Nechť X je n.v. s hustotou

$$f_X(x) = \begin{cases} x/4 & \text{pro } 1 < x \leq 3 \\ 0 & \text{jinak.} \end{cases}$$

Označme A jev $\{X \geq 2\}$.

(a) Spočtěte $\mathbb{E}[X]$.

(b) Spočtěte $\Pr[A]$.

(c) Určete $f_{X|A}$.

(d) Spočtěte $\mathbb{E}[X \mid A]$.

(e) Označme $Y = X^2$. Spočtěte $\mathbb{E}[Y]$ a $\operatorname{var}(Y)$.

(f) Simulujte.

Řešení: 3

4. Nechť X, Y mají sdruženou hustotu

$$f_{X,Y}(x,y) = \begin{cases} e^{-y} & \text{pro } 0 < x < y < \infty \\ 0 & \text{jinak.} \end{cases}$$

(a) Určete podmíněnou hustotu $f_{X|Y}$.

(b) Určete podmíněnou hustotu $f_{Y|X}$.

Řešení: 4

5. V memech na discordu se objevují pouze tyto typy memů: s opicemi – jev M, s kraby – jev C, ostatní – jev O. Některé z nich jsou ve velkém rozlišení – jev HD. Formálně Ω jsou memy a jev M je podmnožina memů na kterých je opice, ... Navíc platí že na každém memu je právě jedna z těch věcí $\Omega = M \dot\cup C \dot\cup O$ (disjunktní sjednocení – tedy $\Omega = M \cup C \cup O$ a navíc $M \cap C = M \cap O = C \cap O = \emptyset$).

- $\Pr[M] = 1/4$
- $\Pr[C] = 3/44$
- $\Pr[O] = 15/22$
- $\Pr[HD \mid M] = 1/11$
- $\Pr[HD \mid C] = 13/15$
- $\Pr[HD \mid O] = 9/15$

Napsal jsem si program, který mi jako wallpaper nastaví náhodný meme, který je ve velkém rozlišení. Jaká je pravděpodobnost, že mám jako wallpaper kraba?

Řešení: 5

6. Nechť $U \sim U(-1,1)$. Položme $V = 2|U| - 1$.

(a) Určete rozdělení V. (Tj. spočtěte distribuční funkci a případně popište, o jaké pojmenované rozdělení se jedná.)

(b) Spočtete $\operatorname{cov}(U,V)$.

(c) Jsou U, V nezávislé?

(d) Simulujte.

Řešení: 6

1.11 Cvičení

1. Označme $S = \sum_{k=0}^{30} \binom{100}{k}$. Označme dále $X = \sum_{i=1}^{100} X_i$, kde X_i je ± 1 s pravděpodobností $1/2$ a veličiny $X_1, \ldots, X_n$ jsou nezávislé.

   (a) Vyjádřete S pomocí vhodné pravděpodobnosti výroku o X.

   (b) Použijte CLV na odhad této pravděpodobnosti.

   (c) Případně vyčíslete S vhodným softwarem a srovnejte.

   Řešení: 1

2. Nechť X_j pro $1 \leq j \leq n$ jsou nezávislé náhodné veličiny, pro které platí $\Pr[X_j = 2/3] = 1/2$ a $\Pr[X_j = 0] = 1/2$ (jiných hodnot ty veličiny nenabývají). Nechť $X = \sum_{j=1}^{n} X_j$ je náhodná veličina rovná jejich součtu. V každém bodě použijte tu konkrétní verzi odhadu, kterou jsem napsal (ne že by jiné nefungovaly). Chceme shora odhadnout $\Pr[X \geq n/2]$

   (a) Určete $\mathbb{E}[X]$ (Jaký poznatek používáte? Jaké má předpoklady?)

   (b) Určete $\mathrm{var}(X)$ (Jaký poznatek používáte? Jaké má předpoklady?)

   (c) Jak $\Pr[X \geq n/2]$ odhadne Markov? (Jsou splněny předpoklady? Jaký je závěr?) Markovova nerovnost:

   - Předpoklady:
     - X je nezáporná náhodná veličina
     - $a > 0$ je reálné číslo
   - Důsledek:

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

   (d) Jak $\Pr[X \geq n/2]$ odhadne Čebyšev? (Jsou splněny předpoklady? Jaký je závěr?) Čebyševova nerovnost:

   - Předpoklady:
     - X je náhodná veličina
     - X má konečnou střední hodnotu
     - X má konečný rozptyl
   - Závěr: pro každé reálné $k > 0$ máme

$$\Pr[|X - \mu| \geq k\sigma] \leq 1/k^2$$

   (e) Jak $\Pr[X \geq n/2]$ odhadne Černov? (Jsou splněny předpoklady? Jaký je závěr?) Černovova nerovnost:

   - Předpoklady:
     - nechť $X_j \in [0, 1]$ jsou nezávislé náhodné veličiny,

  - nechť $X = \sum_{j=1}^{n} X_j$ a nechť $\mathbb{E}[X] = \sum_{j=1}^{n} \mathbb{E}[X_j] = \mu$,
  - nechť $\delta \in (0, 1)$,
- Závěr:

$$\Pr[X \geq \mu + \delta n] \leq e^{-2n\delta^2}$$
$$\Pr[X \leq \mu - \delta n] \leq e^{-2n\delta^2}$$

(f) Jak $\Pr[X \geq n/2]$ odhadne centrální limitní věta? (Jsou splněny předpoklady? Jaký je závěr?) Centrální limitní věta: Předpoklady:

- $X_1, \ldots, X_n$ jsou stejně rozdělené nezávislé náhodné veličiny se střední hodnotou $\mathbb{E}[X_j] = \mu$ a rozptylem $\mathrm{var}(X_j) = \sigma^2$ (pozor, tady mluvíme o X_j).
- Značme $Y_n = ((X_1 + \ldots + X_n) - n\mu) \,/\, (\sqrt{n}\sigma)$

Důsledek:

- $Y_n \xrightarrow{d} N(0, 1)$ tedy Y_n konverguje k normálnímu rozdělení v distribuci (pro větší a větší n)
- Ekvivalentně: pokud F_n je distribuční funkce Y_n, pak

$$\lim_{n\to\infty} F_n(x) = \Phi(x) \qquad \text{(pro každé } x \in \mathbb{R})$$

Může se hodit `scipy.stats.norm.cdf` nebo tabulky na Wikipedii. Pozor, že toto je odhad, který má platit v limitě, nikoliv pro každé n.

(g) Simulujte a porovnejte výsledek simulace s předchozími závěry. Simulujte a porovnejte s odhady pro: $n \in \{1, 2, 3, 4, 5, 10, 20, 30, 40, 50\}$.

Řešení: 2

3. Máme k dispozici samply $X_1, \ldots X_N$ kde $X_j \sim Bin(n, p)$. Ale my neznáme ani n ani p.

   Praktický příklad:

   - Na parapetu mám $N = 50$ květináčů
   - Vím, že jsem do každého květináče nasypal stejně semínek (do každého n), ale už jsem zapomněl kolik to bylo.
   - Nevím s jakou pravděpodobností p které semínko vyklíčí (předpokládám že vyklíčí nezávisle náhodně na ostatních).
   - Ale vím kolik mi v kterém květináči vyrostlo rostlinek X_j v j-tém květináči.

   Takže mám $X_1, \ldots, X_N$ výběr z modelu s parametrem ϑ (zde ϑ je (n, p)). Použijte metodu momentů k odhadu n, p.

   Řešení: 3

4. Po mírném zklamání v předchozím příkladě jsme se rozhodli aspoň zjistit klíčivost. Zasadili jsme tedy $n = 100$ semínek do takového toho platíčka 10×10 květináčků. Za pár týdnů nám některá vyklíčila, X_j je indikátor, jestli j-té semínko vyklíčilo. Pomocí maximální věrohodnosti odhadněte klíčivost p.

   Řešení: 4

5. Známe směrodatnou odchylku σ, ale neznáme střední hodnotu μ. Máme $n = 100$ samplů $X_1, \ldots, X_n$ každý nezávislý a $X_j \sim N(\mu, \sigma^2)$. Jako chybovost volme $\alpha = 0.01$. Ověřte, že intervalový odhad funguje dobře.

   Co vám na tomto příkladě bytostně vadí?

   Řešení: 5

6. William Sealy Gosset pracoval pro nejmenovaný pivovar v Dublinu a zajímaly ho malé samply (třeba tři samply), protože odhadoval kvalitu piva a samply byly drahé. Zkuste to předchozí se studentovým rozdělením. Tedy máme nezávislé náhodné proměnné $X_1, \ldots, X_n$ kde $X_j \sim N(\mu, \sigma^2)$, ale neznáme ani střední hodnotu ani rozptyl (reálná situace) a chceme intervalem odhadnout střední hodnotu (ta je často ta zajímavější).

   Řešení: 6

1.12 Cvičení

1. Po částečném úspěchu v určování klíčivosti chceme mít ještě lepší představu o skutečné hodnotě. Místo bodového odhadu (o kterém nevíme jak daleko je pravdě) chceme intervalový odhad (s pravděpodobností 99% se trefíme intervalem okolo správné hodnoty). Zasadili jsme tedy $n = 100$ semínek do takového toho platíčka 10×10 květináčků. Za pár týdnů nám některá vyklíčila, X_j je indikátor, jestli j-té semínko vyklíčilo. Pomocí intervalového odhadu odhadněte klíčivost p.

   (a) Co kdybychom chtěli odhadovat pomocí normálního rozdělení? Tedy kdybychom měli rozptyl, ale chtěli odhadnout střední hodnotu (tedy pro indikátor $\Pr[X_j = 1]$)?

   (b) Použijte centrální limitní větu a tedy studentovo rozdělení na odhad pro $\alpha = 0.01$.

   Řešení: 5

2. Podle slibu výrobce bude stroj dělat chyby nejvýše ve 3% případů. Z 600 pokusů došlo k chybě v 28 případech. Posuďte slib výrobce (coby nulovou hypotézu) na hladině významnosti 5%.

   (a) Co jsme měli udělat před pozorováním chybných strojů?

   (b) Počet chyb modelujte přesně, tj. pomocí binomického rozdělení.

   (c) Počet chyb modelujte přibližně pomocí normálního rozdělení (s vhodným μ, σ^2).

   Řešení: 2

3. Vyzkoušejte, zda Python random funguje dobře. Pomocí `random.choices([1, 2, 3, 4], k=20)` jsme vygenerovali následující hody:

$$[1, 3, 1, 1, 3, 1, 1, 2, 1, 2, 2, 2, 2, 2, 1, 1, 4, 4, 1, 4]$$

   tedy četnosti jsou:

$$\begin{aligned} X_1 &= 9 \\ X_2 &= 6 \\ X_3 &= 2 \\ X_4 &= 3 \end{aligned}$$

   Průměrně mělo vyjít $E_j = \mathbb{E}[X_j] = 5$. Otázkou je: „generuje python špatnou náhodu?“

   Řešení: 3

4. Náklon šikmé věže v Pise je měřen vzdáleností pevného bodu ve věži od jeho „správné“ polohy. V letech 1975 až 1987 tato poloha rostla následujícím způsobem: 2.9642, 2.9644, 2.9656, 2.9667, 2.9673, 2.9688, 2.9696, 2.9698, 2.9713, 2.9717, 2.9725, 2.9742, 2.9757. Proveďte lineární regresi, znázorněte i graficky.

   Řešení: 4

5. Mějme náhodné veličiny X, Y které mají sdruženou hustotu (probability density function)

$$f_{X,Y}(x,y) = \begin{cases} 8xy & x \in [0,1], 0 \leq y \leq x \\ 0 & \text{jinak} \end{cases}$$

   (všimněte si, kde je ta hustota nenulová!)

   (a) Ověřte, že je to pravděpodobnostní hustota, tedy

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy = 1$$

   (b) Spočítejte marginální hustoty

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx$$

   (c) Určete distribuční funkce

$$F_X(x) = \Pr[X \leq x] = \int_{-\infty}^{x} f_X(s)\ ds$$
$$F_Y(y) = \Pr[Y \leq y] = \int_{-\infty}^{y} f_Y(t)\ dt$$
$$F_{X,Y}(x,y) = \Pr[X \leq x \wedge Y \leq y] = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)\ dt\ ds$$

   (d) Jsou X, Y nezávislé? Tedy platí pro každé $s, t \in \mathbb{R}$

$$F_{X,Y}(s,t) = F_X(s) F_Y(t)$$

   (e) Spočítejte střední hodnotu náhodné veličiny $Z = X + 2Y$ pomocí LOTUS

$$g\colon \mathbb{R}^2 \to \mathbb{R} \qquad \text{(měřitelná)}$$
$$\mathbb{E}[Z] = \int_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y)\ dxy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\ dx\ dy$$

   (f) Nechť A je jev $X + Y \leq 1$, spočítejte

$$\Pr[A] = \int_A f_{X,Y}(x,y)\ dxy$$

   (g) Spočítejte podmíněnou hustotu $f_{X|A}$, distribuční funkci $F_{X|A}$ a střední hodnotu $\mathbb{E}[X \mid A]$:

$$F_{X|A}(x) = \Pr[X \leq x \mid A] = \frac{\Pr[X \leq x \wedge A]}{\Pr[A]}$$

$$F_{X|A}(x) = \int_{-\infty}^{x} f_{X|A}(s)\ ds$$
$$\mathbb{E}[X \mid A] = \int_{-\infty}^{\infty} x f_{X|A}(x)\ dx$$

(h) Určete podmíněnou hustotu

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad \text{(pokud } f_Y(y) > 0\text{)}$$

(i) Určete podmíněnou distribuční funkci

$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(s \mid y)\ ds$$

(j) V tomto domácím úkolu nemusíte simulovat. Nejspíš byste vymysleli sami, jak to dělat. Ale ve skutečnosti to není zas tak jednoduché.

Řešení: 5

Kapitola 2

Tahák

2.1 Pravděpodobnostní prostor

Definice. Pravděpodobnostní prostor *je trojice* $(\Omega, \mathcal{F}, \Pr)$*, kde*

1. Ω *je* množina elementárních jevů *(sample space) (je to množina)*
2. $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ *je* prostor jevů *(event space) (je to množina podmnožin* Ω*, jednotlivé prvky* $\mathcal{F}$ *nazýváme jevy)* $\mathcal{F}$ *je* σ*-algebra, tedy musí platit:*
   (a) $\emptyset \in \mathcal{F}$ *a zároveň* $\Omega \in \mathcal{F}$ *(celá množina elementárních jevů je jev)*
   (b) $A \in \mathcal{F} \Rightarrow (\Omega \setminus A) \in \mathcal{F}$ *(prostor jevů je uzavřený na doplňky)*
   (c) $(\forall n \in \mathbb{N}\colon A_n \in \mathcal{F}) \Rightarrow (\cup_{n\in\mathbb{N}} A_n) \in \mathcal{F}$ *(prostor jevů je uzavřený na spočetná sjednocení, poznámka může se stát, že* $A_1 \neq A_2 = A_3 = \cdots$*, specielně tedy je uzavřený i na všechna konečná sjednocení)*
3. $\Pr$ *je funkce* $\Pr\colon \mathcal{F} \to [0,1]$ *je* pravděpodobnost *jevu z prostoru jevů, musí splňovat:*
   (a) $\Pr[\emptyset] = 0$ *a zároveň* $\Pr[\Omega] = 1$
   (b) $\Pr$ *je spočetně aditivní, tedy pro každou* $I \subseteq \mathbb{N}$ *a každou posloupnost jevů* $(A_j)_{j\in I}$*, které jsou po dvou disjunktní (tedy* $\forall i, j \in I\colon i \neq j \Rightarrow A_i \neq A_j$*) platí:*

$$\Pr\left[\cup_{j\in I} A_j\right] = \sum_{j\in I} \Pr[A_j]$$

(tedy $\Pr$ *je pravděpodobnostní míra)*

2.2 Podmíněná pravděpodobnost

Definice. *Nechť* $(\Omega, \mathcal{F}, \Pr)$ *je pravděpodobnostní prostor. Nechť* $A, B \in \mathcal{F}$ *a navíc* $\Pr[B] > 0$*. Pak definujeme podmíněnou pravděpodobnost*

$$\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

2.3 Bayesova věta

Věta 1. *Nechť* $(\Omega, \mathcal{F}, \Pr)$ *je pravděpodobnostní prostor. Nechť* $B_1, B_2, \ldots \in \mathcal{F}$ *je rozklad* Ω *(na spočetně mnoho množin). Nech* $A \in \mathcal{F}$ *a navíc platí* $\Pr[A] > 0$ *a navíc* $\Pr[B_i] > 0$ *pro každé* i*. Pak*

$$\Pr[B_j \mid A] = \frac{\Pr[A \mid B_j] \Pr[B_j]}{\sum_i \Pr[A \mid B_i] \Pr[B_i]}$$

2.4 Nezávislé jevy

Definice. *Nechť* $(\Omega, \mathcal{F}, \Pr)$ *je pravděpodobnostní prostor. Nechť* $A, B \in \mathcal{F}$ *jsou dva jevy. Pak řekneme, že* A, B *jsou* nezávislé jevy*, pokud platí:*

$$\Pr[A \cap B] = \Pr[A] \Pr[B]$$

(také se dá říct $\Pr[A \mid B] = \Pr[A]$ *pokud* $\Pr[B] > 0$*).*

2.5 Spojité náhodné veličiny

Spousta z tohoto platí pro obecné náhodné veličiny (tedy i pro spojité i pro diskrétní).

- *Distribuční funkce* (cumulative distribution function – CDF) náhodné veličiny X:

$$F_X : \mathbb{R} \to [0, 1]$$
$$F_X(x) = \Pr[X \leq x] \qquad \text{(pro každé } x \in \mathbb{R})$$

  Tohle má smysl pro každou náhodnou veličinu.

- *Hustota* (probability density function – PDF) náhodné veličiny X:

$$f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$$
$$\Pr[X \leq x] = \int_{-\infty}^{x} f_X(t)\ dt \qquad \text{(pro každé } x \in \mathbb{R})$$

  Jen pro velice hezké spojité veličiny.

  Můžeme psát také:

$$1_A(t) = \begin{cases} 1 & t \in A \\ 0 & t \notin A \end{cases} \qquad \text{(indikátorová funkce)}$$
$$\Pr[X \in A] = \int_A f_X(t)\ dt = \int_{-\infty}^{\infty} 1_A(t) f_X(t)\ dt$$

- *Střední hodnota* náhodné veličiny X:

$$\mathbb{E}[X] = \int_\Omega X(\omega) d \Pr(\omega) = (R) \int_0^\infty \Pr\left(\{x \in \mathbb{R} \mid X(x) > t\}\right)\ dt$$

  (Pokud $X(\omega) \geq 0$ vždy, jinak vyjádříme X jako rozdíl dvou nezáporných funkcí)

$$= (R) \int_0^\infty (1 - F_X(t))\ dt$$

kde (R) značí Riemannův integrál. Tato definice funguje pro každou náhodnou veličinu (i pro diskrétní). Definice náhodné veličiny obsahuje podmínku, že $\{\omega \in \Omega \mid X(\omega) \leq t\} \in \mathcal{F}$, což je přesně podmínka, abychom mohli určit pravděpodobnost v předchozím integrálu.

Pokud známe i hustotu, tak můžeme psát:

$$\mathbb{E}[X] = (R) \int_{-\infty}^\infty t f_X(t)\ dt \qquad (X \text{ nemusí být nezáporná})$$

Případně máme pro libovolnou funkci g:

$$\mathbb{E}[g(X)] = (R) \int_{-\infty}^\infty g(t) f_X(t)\ dt$$

- *Rozptyl* náhodné veličiny:

$$\operatorname{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])$$

Tohle má smysl pro každou náhodnou veličinu (kde je pravá strana definovaná).

Kapitola 3

Řešení

3.1 Cvičení

1. **Úvodní informace:**
   (a) **Slyšíte mě všichni dobře?**
   (b) **Literatura.**
   Řešení: TODO
   (c) **Pravidla zápočtu (domácí úkoly).**
   Řešení: TODO

2. **Jak se generuje náhoda programem.**

   - **Python3** ***Řešení:***

```python
# https://docs.python.org/3/library/random.html
# Nepoužívejte pro šifrování!
import random
import scipy

# https://docs.python.org/3/library/itertools.html
import itertools

# Generuj náhodné celé číslo 1 <= x <= 10 (tedy x in range(1, 11))
x = random.randint(1, 10)
print(x)

# Generuj náhodný prvek dané neprázdné posloupnosti.
drink = random.choice([
    "světlé pivo",
    "tmavé pivo",
    "slivovice",
    "bílé víno",
    "červené víno",
    "čaj",
    ])
print(drink)

# Výsledek 'B' je třikrát pravděpodobnější než 'A'.
print(random.choice(['A', 'B', 'B', 'B']))

# Vrátí list k náhodných prvků z dané sekvence s danými váhami
# (cum_weights jsou trochu rychlejší).
# https://docs.python.org/3/library/random.html#random.choices
print(random.choices(['A', 'B', 'C'], weights=[1/6, 1/6, 2/3], k=5))

# Náhodná permutace (mění přímo daný list).
my_list = ['a', 'b', 'c', 'd']
random.shuffle(my_list)
print(my_list)

# Vybere dva prvky dané posloupnosti, ve výsledku se nebudou opakovat
# pozice. Pokud se prvky opakují, tak se mohou ve výsledku opakovat.
print(random.sample(['w', 'x', 'y', 'z'], k=2))  # two distinct letters
print(random.sample(['w', 'x', 'y', 'x'], k=2))  # 'x' can repeat

# Pro přesné počítání (vrací iterable):
print(list(itertools.permutations([1, 2, 3])))
print(list(itertools.combinations('ABCD', 2)))
print(list(itertools.combinations_with_replacement('ABCD', 2)))

# Může se hodit scipy: sudo apt-get install python3-scipy
scipy.special.comb(n=5, k=2, exact=True)  # vrátí n nad k
```

   - **C++** ***Řešení:***

```cpp
std::default_random_engine generator;
```

```
std::uniform_int_distribution<int> distribution(0, 9);
```

- **R** ***Řešení:***

```
x <- "hello"
```

3. **Připomeňte si definici pravděpodobnostního prostoru (Definice 2.1). Určete, co je**

   - ***množina elementárních jevů (sample space)*, tedy množina Ω,**
   - ***prostor jevů (event space)*, tedy množina $\mathcal{F} \subseteq \mathcal{P}(\Omega)$,**
   - ***pravděpodobnost (probability)*, tedy funkce $\Pr\colon \mathcal{F} \to [0,1]$**

   pro následující příklady:

   (a) Hod spravedlivou alkoholovou trojhrannou tužkou (není to kostka kvůli popisu prostoru jevů):

   ```
   import random
   drink = random.choice([
                   "světlé pivo",
                   "tmavé pivo",
                   "slivovice",
                   ])
   ```

   Řešení: V počítači nám stačí předchozí kód (pseudonáhodné číslo vs náhodné číslo). Fyzicky můžeme generovat pomocí hodu trojhrannou tužkou, která má značky na stěnách (a zaručeně nemůže padnout nastojato).

   - Množina elementárních jevů $\Omega = \{\text{světlé pivo, tmavé pivo, slivovice}\}$
   - Prostor jevů $\mathcal{F} = \mathcal{P}(\Omega)$, tedy $|\mathcal{F}| = 2^{|\Omega|} = 2^3$:

   $$\begin{aligned}\mathcal{F} = \{&\\ &\emptyset,\\ &\{\text{světlé pivo}\},\\ &\{\text{tmavé pivo}\},\\ &\{\text{slivovice}\},\\ &\{\text{světlé pivo, tmavé pivo}\},\\ &\{\text{světlé pivo, slivovice}\},\\ &\{\text{tmavé pivo, slivovice}\},\\ &\{\text{světlé pivo, tmavé pivo, slivovice}\}\\ \}&\end{aligned}$$

   - Pravděpodobnost

   $$\begin{aligned}\Pr[\emptyset] &= 0\\ \Pr[\{\text{světlé pivo}\}] &= 1/3\\ \Pr[\{\text{tmavé pivo}\}] &= 1/3\\ \Pr[\{\text{slivovice}\}] &= 1/3\\ \Pr[\{\text{světlé pivo, tmavé pivo}\}] &= 2/3\\ \Pr[\{\text{světlé pivo, slivovice}\}] &= 2/3\\ \Pr[\{\text{tmavé pivo, slivovice}\}] &= 2/3\\ \Pr[\{\text{světlé pivo, tmavé pivo, slivovice}\}] &= 1\end{aligned}$$

   Většinou ale nepopisujeme jev jako podmnožinu elementárních jevů, ale nějak lidsky:

   $\Pr[\text{nějaké pivo}] = \Pr[\text{světlé pivo} \vee \text{tmavé pivo}] = \Pr[\{\text{světlé pivo, tmavé pivo}\}] = 2/3$

$$\Pr[\text{ne pivo}] = \Pr[\{\text{slivovice}\}] = 1 - \Pr[\{\text{světlé pivo, tmavé pivo}\}] = 1/3$$

(b) Uniformně náhodné číslo z intervalu $[0, 1)$.

```
import random
print(random.random())
```

Řešení: V počítači nám funkce `random.random` vrátí float x, který je přesně reprezentovatelný, platí $0.0 \leq x < 1.0$ a zároveň x je celočíselný násobek 2^{-53}. To ale znamená, že nikdy nedostaneme 0.05954861408025609 i když to je číslo přesně reprezentovatelné jako python float. Můžeme generovat i vícebitová čísla, ale vždy to bude číslo! A to je dobře, většina reálných čísel nejde reprezentovat konečnou posloupností symbolů (pozor, $\sqrt{2}$ jde reprezentovat jako kořen polynomu, ale i třeba $1/\pi$ jde reprezentovat například algoritmem, který ho počítá).

Jak bychom fyzikálně generovali uniformně náhodné číslo z intervalu $[0, 1)$? Uniformně náhodné znamená, že každé číslo bude stejně pravděpodobné. Hod šipkou na interval má nevýhodu, že buď budou konce intervalu méně pravděpodobné než prostředek nebo se nám může stát že hodíme šipku mimo interval (nejspíš obojí). Můžeme ale udělat terč s jedním vyznačeným poloměrem, který bude kruh, připevnit jeho střed k vrtačce, roztočit a hodit šipku (tak abychom netrefili střed, ale určitě trefili kruh). Pak náhodné číslo $x \in [0, 1)$ bude úhel který svírá vyznačný poloměr a naše šipka dělený 2π.

- Toto je jen myšlenkový experiment, fyzická implementace je nejspíš poměrně nebezpečná. Doma to nezkoušejte!
- Můžete namítnout, že šipka nevybere přesně bod, že terč je tvořen atomy a tedy je také z nějakého pohledu diskrétní. Asi ano, ale já neříkal, že reálná čísla existují. Pro představu atom může mít poloměr okolo $10^{-10}m$, tedy na délce $1m$ jich vedle sebe vyskládáme zhruba 10^{10}. Srovnejte s přesností $2^{-53} \approx 10^{-16}$, tedy pokud bychom výsledek `random.random()` brali jako pozici v metrech, pak máme přesnost zhruba na dvě miliontiny atomu.
- *množina elementárních jevů (sample space)*, tedy množina $\Omega = [0, 1)$
- *prostor jevů (event space)*, tedy množina $\mathcal{F} \subseteq \mathcal{P}(\Omega)$

  Napřed se zeptejme, jestli by nemohlo platit, že $\mathcal{F} = \mathcal{P}(\Omega)$. Asi bychom chtěli následující vlastnosti:

  – pokud $A \in \mathcal{F}$ je interval, pak $\Pr[A]$ je rovna délce A

  – pokud $A \in \mathcal{F}$ a navíc $x + A = \{x + a \mid a \in A\} \subseteq \Omega$ pro nějaké reálné číslo x, pak $\Pr[A] = \Pr[x + A]$.

  – každá podmnožina Ω má přiřazenou nějakou pravděpodobnost

  Ale to nejde, protože ne každá množina má míru (tady Pr odpovídá takzvané pravděpodobnostní míře – míra celého prostoru je rovna jedné). Nejznámějším příkladem je `https://en.wikipedia.org/wiki/Banach%E2%80%93Tarski_paradox`. Hezké video: `https://www.youtube.com/watch?v=s86-Z-CbaHA`.

  Naše řešení: $\mathcal{F}$ je množina Lebesgueovsky měřitelných podmnožin Ω.
- *pravděpodobnost (probability)*, tedy funkce $\Pr\colon \mathcal{F} \to [0, 1]$ je Lebesgueova míra.

Pokud jste neslyšeli o tom, co je to míra, tak se nelekejte. Dobré k zapamatování:

- Pravděpodobnost je číslo mezi nulou a jedničkou (obecná míra celého prostoru nemusí být rovna jedné, ale nás zajímá pravděpodobnostní míra).

- "Nic se nestane" má pravděpodobnost nula – $\Pr[\emptyset] = 0$ (tedy specielně $\emptyset$ je měřitelná).
- "Něco se stane" má pravděpodobnost jedna – $\Pr[\Omega] = 1$ (tedy specielně Ω je měřitelná).
- "Stane se A" má pravděpodobnost 1-"Nestane se A" – $\Pr[A] = 1 - \Pr[\Omega \setminus A]$ (tedy pokud je A měřitelná, pak je i její doplněk měřitelný).
- Pr[Na kostce padne jedna tečka nebo dvě tečky] = Pr[padne jedna tečka]+Pr[padnou dvě tečky] – pravděpodobnost sjednocení disjunktních množin je rovna součtu jejich pravděpodobností (platí také pro *spočetně* mnoho disjunktních množin a navíc sjednocení spočetně mnoha disjunktních měřitelných množin je taky měřitelné)
- Lebesgueova míra jednoho bodu je nulová (tedy ze spočetného disjunktního sjednocení máme že pravděpodobnost že uniformně náhodné reálné číslo z intervalu $[0, 1)$ je racionální je nulová).

  Pozor na to, že sice platí $\Pr[\{0.1\}] = 0$, ale to neznamená, že jev že padne 0.1 je nemožný (akorát velice velice nepravděpodobný). Naopak pokud je jev nemožný Pr[na šestistěnné kostce padne sedm], pak je jeho pravděpodobnost nulová.
- Úsečka v $[0, 1] \times [0, 1] \subseteq \mathbb{R}^2$ má Lebesgueovu míru nula.

Takže ty definice, které vám přijdou divné jsou tam kvůli teorii míry (případně později kvůli teorii integrálu).

Pozor na to, že ne každá pravděpodobnost je Lebesgueova míra, můžeme uvažovat například $\Omega = [0, 1)$, $\mathcal{F}$ jsou Lebesgueovsky měřitelné, Pr která $\Pr[\{0.1\}] = 1/2$ a $\Pr[A] = \lambda(A)/2 + 1/2$ pokud $0.1 \in A$ a $\Pr[A] = \lambda(A)/2$ jinak (kde $\lambda(A)$ značí Lebesgueovu míru množiny A).

Někdy také mluvíme o pravděpodobnostní distribuci, zatím to můžete brát jako synonymum k pravděpodobnosti.

4. **Dokažte, že** $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$**.**

Řešení: Předpokládejme, že platí $A \cap B, A \cup B, A, B \in \mathcal{F}$ (jinak by výraz nahoře neměl smysl). Z definice pravděpodobnostního prostoru (Definice 2.1) máme spočetnou aditivitu pro disjunktní jevy. Rozdělíme tedy $A \cup B$ na disjunktní množiny:

$$\begin{aligned} C_1 &= A \cap B \\ C_2 &= A \setminus B \\ C_3 &= B \setminus A \end{aligned}$$

tedy máme $A \cup B = C_1 \cup C_2 \cup C_3$ a zároveň $C_i \cap C_j = \emptyset$ pro každé dvě $1 \leq i < j \leq 3$. Dle spočetné aditivity můžeme psát:

$$\begin{aligned} \Pr[A \cup B] &= \Pr[C_1 \cup C_2 \cup C_3] \\ &= \Pr[C_1] + \Pr[C_2] + \Pr[C_3] \end{aligned}$$

Máme také:

$$\begin{aligned} A &= C_2 \cup C_1 = (A \setminus B) \cup (A \cap B) \\ B &= C_3 \cup C_1 = (B \setminus A) \cup (A \cap B) \end{aligned}$$

takže můžeme psát:

$$\begin{aligned} \Pr[A \cup B] &= \Pr[C_1] + \Pr[C_2] + \Pr[C_3] \\ &= \Pr[C_1] + \Pr[C_2] + 2\Pr[C_3] - \Pr[C_3] \\ &= (\Pr[C_1] + \Pr[C_3]) + (\Pr[C_2] + \Pr[C_3]) - \Pr[C_3] \\ &= (\Pr[A]) + (\Pr[B]) - \Pr[C_3] \\ &= \Pr[A] + \Pr[B] - \Pr[A \cap B] \end{aligned}$$

což jsme chtěli dokázat.

5. **Zopakujte si základní kombinatoriku:**

   - **Kolik je různých permutací množiny $\{A, B, C\}$?**

     Řešení: $3! = 3 \cdot 2 \cdot 1$, obecně $n!$ pokud máme n rozlišitelných prvků

     ```
     import itertools

     permutations = list(itertools.permutations(['A', 'B', 'C']))
     print(f'There are {len(permutations)} permutations: {permutations}')

     # There are 6 permutations:
     #     [('A', 'B', 'C'),
     #      ('A', 'C', 'B'),
     #      ('B', 'A', 'C'),
     #      ('B', 'C', 'A'),
     #      ('C', 'A', 'B'),
     #      ('C', 'B', 'A')]
     ```

   - **Kolik různých slov skládajících se z písmen $\{A, B\}$ má délku 3?**

     Řešení: $2^3 = 8$, obecně n^r kde n je počet písmen, r délka slova

     ```
     import itertools

     all_words = list(itertools.product('AB', repeat=3))
     print(f'There are {len(all_words)} words: {all_words}')

     # There are 8 words:
     #     [('A', 'A', 'A'),
     #      ('A', 'A', 'B'),
     #      ('A', 'B', 'A'),
     #      ('A', 'B', 'B'),
     #      ('B', 'A', 'A'),
     #      ('B', 'A', 'B'),
     #      ('B', 'B', 'A'),
     #      ('B', 'B', 'B')]
     ```

   - **Kolik různých podmnožin množiny $\{A, B, C, D, E\}$ má velikost 3?**

     Řešení: $\binom{n}{r} = \frac{n!}{r!(n-r)!} = \binom{5}{3} = 10$

     ```
     import itertools

     all_subsets = list(itertools.combinations('ABCDE', 3))
     print(f'There are {len(all_subsets)} subsets: {all_subsets}')

     # There are 10 subsets:
     #     [('A', 'B', 'C'),
     #      ('A', 'B', 'D'),
     #      ('A', 'B', 'E'),
     #      ('A', 'C', 'D'),
     #      ('A', 'C', 'E'),
     #      ('A', 'D', 'E'),
     #      ('B', 'C', 'D'),
     #      ('B', 'C', 'E'),
     #      ('B', 'D', 'E'),
     #      ('C', 'D', 'E')]
     ```

- **Kolik různých kombinací s opakováním z množiny $\{A, B, C, D, E\}$ velikosti 3?**

  Řešení: $\binom{n+r-1}{r}$ kde $n = 5$, $r = 3$. Protože máme $n - 1$ svislítek a r hvězdiček a kódujeme takto:

  $$AAD = **|||*|$$

  tedy

  počet A|počet B|počet C|počet D|počet E

  kde každý počet je reprezentován počtem hvězdiček a vybíráme které z $n+r-1$ symbolů budou hvězdičky.

```python
import itertools

sorted_sequences = list(itertools.combinations_with_replacement('ABCDE', r=3))
print(f'Máme {len(sorted_sequences)} sekvencí: {sorted_sequences}')

# There are 35 sorted sequences:
#     [('A', 'A', 'A'), ('A', 'A', 'B'), ('A', 'A', 'C'),
#      ('A', 'A', 'D'), ('A', 'A', 'E'), ('A', 'B', 'B'),
#      ('A', 'B', 'C'), ('A', 'B', 'D'), ('A', 'B', 'E'),
#      ('A', 'C', 'C'), ('A', 'C', 'D'), ('A', 'C', 'E'),
#      ('A', 'D', 'D'), ('A', 'D', 'E'), ('A', 'E', 'E'),
#      ('B', 'B', 'B'), ('B', 'B', 'C'), ('B', 'B', 'D'),
#      ('B', 'B', 'E'), ('B', 'C', 'C'), ('B', 'C', 'D'),
#      ('B', 'C', 'E'), ('B', 'D', 'D'), ('B', 'D', 'E'),
#      ('B', 'E', 'E'), ('C', 'C', 'C'), ('C', 'C', 'D'),
#      ('C', 'C', 'E'), ('C', 'D', 'D'), ('C', 'D', 'E'),
#      ('C', 'E', 'E'), ('D', 'D', 'D'), ('D', 'D', 'E'),
#      ('D', 'E', 'E'), ('E', 'E', 'E')]
```

6. **Jaká je pravděpodobnost, že při hodu šesti rozlišitelných spravedlivých šestistěnných kostek padnou aspoň na třech kostkách aspoň tři? Jaký je množina elementárních jevů, prostor jevů a pravděpodobnost?**

   Řešení:

   - Množina elementárních jevů je $\Omega = \{1,2,3,4,5,6\}^6 = \{111111, 111112, \ldots, 666666\}$, tedy $|\Omega| = 6^6 = 46656$.
   - Prostor jevů je $\mathcal{F} = \mathcal{P}(\Omega)$, tedy $|\mathcal{F}| = 2^{46656}$
   - Pravděpodobnost $\Pr[\{abcdef\}] = 1/6^6$ pro libovolná $a, b, c, d, e, f \in \{1,2,3,4,5,6\}$.

   Jak takový příklad řešit? Uvědomit si přesně o čem mluvíme, pak zkusit přemýšlet o jednodušších jevech.

   - Na jedné kostce padnou aspoň tři s pravděpodobností $4/6 = 2/3$ (musí padnout $3,4,5,6$, tedy nepadne $1,2$).
   - Pokud vybereme k kostek, pak pravděpodobnost, že přesně na těchto kostkách padne aspoň tři je přesně $(2/3)^k(1/3)^{6-k}$ (kostky jsou nezávislé). Kupříkladu pokud vybereme první čtyři kostky, pak nás zajímá:

   $$\Pr\left[\{ABCDef \mid A,B,C,D \in \{3,4,5,6\}, e,f \in \{1,2\}\}\right] = \frac{4^4 \cdot 2^2}{6^6} = (2/3)^4(1/3)^{6-4}$$

   - Přesně k kostek vybereme $\binom{6}{k}$ způsoby.
   - Rozdělíme pravděpodobnostní prostor na jevy, kde přesně na k kostkách padne číslo aspoň tři (tedy máme disjunktní rozklad) a $k \geq 3$, tedy použijeme definici pravděpodobnosti a sečteme předchozí pro $k \in \{3,4,5,6\}$:

   $$\begin{aligned}\Pr[\text{aspoň na třech kostkách aspoň tři}] =& \binom{6}{3}(2/3)^3(1/3)^{6-3} \\ &+ \binom{6}{4}(2/3)^4(1/3)^{6-4} \\ &+ \binom{6}{5}(2/3)^5(1/3)^{6-5} \\ &+ \binom{6}{6}(2/3)^6(1/3)^{6-6} \\ =& 0.8998628257887514\end{aligned}$$

   Tady jsme vlastně použili větu z přednášky, že pokud $B_0, B_1, \ldots, B_{2^6-1}$ jsou rozklad Ω (tedy $B_i \neq B_j$ pro $i \neq j$ a zároveň $\cup B_i = \Omega$), pak $\Pr[A] = \sum_i \Pr[A \mid B_i]\Pr[B_i]$. Kde jev A je že na aspoň třech kostkách padne aspoň tři. Jev B_i je že na přesně určených kostkách padne aspoň tři (tedy jevy B_i, B_j jsou opravdu disjunktní). Konkrétně 22 zapsané binárně je 010110, pak jev B_{22} je jev že na druhé, čtvrté a páté kostce padlo číslo aspoň tři. Pak $\Pr[A \mid B_x] = 1$ pokud x má v binárním zápisu aspoň tři jedničky a $\Pr[A \mid B_x] = 0$ jinak. Už jsme spočítali, že $\Pr[B_x] = (2/3)^k(1/3)^{6-k}$ pokud x má v binárním zápisu k jedniček.

   Pomocí programu:

```
import itertools
import scipy.special
import random
```

```python
# Přesný výsledek pomocí kombinatoriky:

def p(k):
    """ Probability that there are exactly k out of 6 dice with at least 3. """
    return scipy.special.comb(6, k, exact=True) * ((2/3)**k) * ((1/3)**(6-k))


exact_computed = sum(p(k) for k in range(3, 7))
print(f'Přesný výsledek: {exact_computed}')


# Přesný výsledek spočítaný hrubou silou:

def indicator(dice):
    """ Return 1 if at least three dice have at least 3, otherwise
    return 0. """
    if sum(1 for x in dice if x >= 3) >= 3:
        return 1
    else:
        return 0


all_outcomes = itertools.product(range(1, 7), repeat=6)
exact_bruteforce = sum(indicator(d) for d in all_outcomes) / (6**6)
print(f'Hrubá síla: {exact_bruteforce}')


# Simulace:

N = 1000  # Number of tries
simulated = sum(indicator(random.choices(range(1, 7), k=6))
                for _ in range(N)) / N
print(f'Simulace: {simulated}')

# Možný výsledek:
# Přesný výsledek: 0.8998628257887514
# Hrubá síla: 0.8998628257887518
# Simulace: 0.903
```

Porovnejme naše metody:

- Výpočet vzorcem:
  - přesný výsledek
  - potřebovali jsme kombinatoriku a přemýšlet
  - pokud se změní zadání, tak řešení se změní celkem dost
  - velice rychlý výpočet
- Procházení všech možností:
  - jednodušší vymýšlení
  - potřebujeme programovat

  - přesný výsledek (liší se v posledních místech floatu, dáno nepřesnostmi floatové reprezentace, `exact=True` nevrací nativní float)
  - pokud se změní zadání, tak se řešení skoro nezmění
  - pokud je množina elementárních jevů velká, tak se tento postup nepoužitelný
- Simulace:
  - jednoduché vymýšlení (skoro jako předchozí případ)
  - pokud se změní zadání, tak se řešení skoro nezmění
  - časová složitost není lineární ve velikosti množiny elementárních jevů (N krát vybíráme náhodný prvek Ω, což často zvládáme v $\mathcal{O}(\log|\Omega|)$ krocích)
  - nepřesný výsledek – závisí na náhodných bitech počítače a počtu pokusů
  - budeme potřebovat trochu teorie abychom odhadli jak jistí si jsme výsledkem

7. **Nechť Ω jsou všechny permutace prvních 100 přirozených čísel, prostor jevů jsou všechny podmnožiny Ω a každý elementární jev je stejně pravděpodobný. Označme jev A_j že náhodně zvolená permutace $\pi \in \Omega$ splňuje $\pi(j) = j$ (pro $1 \leq j \leq 100$). Jsou A_1, A_2 nezávislé jevy?**

   Řešení:

   - Počet permutací v A_j je přesně 99! (jeden prvek je fixní, zbytek permutujeme), tedy

   $$\Pr[A_j] = \frac{99!}{100!} = \frac{1}{100}$$

   - Počet permutací v $A_1 \cap A_2$ je přesně 98! (dva prvky jsou fixní, zbytek permutujeme), tedy

   $$\Pr[A_1 \cap A_2] = \frac{98!}{100!} = \frac{1}{9900}$$

   - Dle definice nezávislých jevů bychom potřebovali $\Pr[A_1]\Pr[A_2] = \Pr[A_1 \cap A_2]$ (Definice 2.4), ale to neplatí:

   $$\Pr[A_1 \cap A_2] = \frac{1}{9900} \neq \frac{1}{10000} = \Pr[A_1]\Pr[A_2]$$

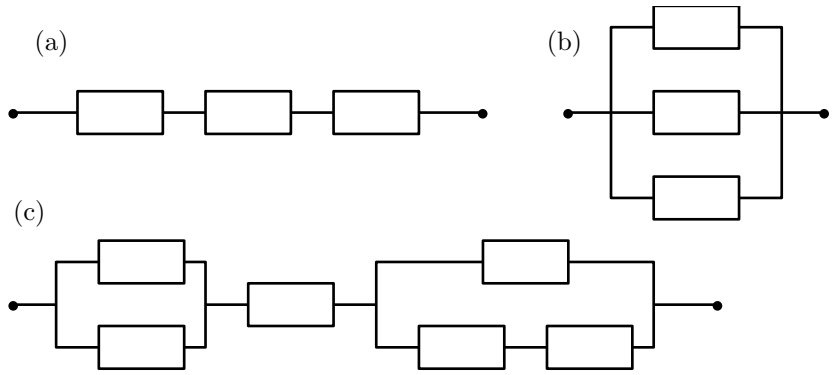
   Zamysleme se nad počítačovým řešením:

```python
import random


# indexujeme od nuly
def fixed(my_list, j):
    return my_list[j] == j


my_list = list(range(100))
N = 100000

A1 = 0
for _ in range(N):
    random.shuffle(my_list)
    A1 += 1 if fixed(my_list, 0) else 0
A1 = A1 / N
print(f'Pr[A_1] = Pr[A_2] = {A1} (= {1/100})')

A1A2 = 0
for _ in range(N):
    random.shuffle(my_list)
    A1A2 += 1 if fixed(my_list, 0) and fixed(my_list, 1) else 0
A1A2 = A1A2 / N
print(f'Pr[A_1 and A_2] = {A1A2} (= {1/9900})')

# Možný výsledek:
# Pr[A_1] = Pr[A_2] = 0.00993 (= 0.01)
# Pr[A_1 and A_2] = 0.00012 (= 0.00010101010101010101)
```

   - jednoduchá simulace

- potřebujeme více pokusů, protože potřebujeme odhadnout s větši přesností (menší pravděpodobnost, tak abychom nedostali nulu)
- vůbec nemůžeme použít hrubou sílu, neboť $100! \approx 9.33 \cdot 10^{157}$, pro představu:
  - počítač vykoná zhruba 10^9 instrukcí za sekundu
  - lineární algoritmus (další permutaci najdeme v jednotkovém čase) by trval zhruba 10^{148} sekund
  - stáří vesmíru se odhaduje na $13.787 \cdot 10^9$ let
  - jeden rok trvá zhruba $\pi \cdot 10^7$ sekund
  - stáří vesmíru je tedy zhruba $4.34 \cdot 10^{17}$ sekund
  - tedy lineární algoritmus který by prošel všechny permutace 100 prvkové množiny by běžel zhruba 10^{131} stáří vesmíru

8. **Každý obdélník na obrázku je součástka, která se může porouchat s pravděpodobností p. Přesněji řečeno porucha znamená, že skrz ní neteče proud. Poruchy součástek jsou na sobě nezávislé. Jaká je pravděpodobnost, že stále poteče proud mezi dvěma puntíky.**



Řešení:

(a) Jak postupujeme:

- Jedna součástka se neporouchá (tedy je ok, jev O, porouchá jev P) s pravděpodobností $\Pr[O] = 1 - \Pr[P] = 1 - p$.
- Aby proud tekl, tak všechny součástky musí být ok. Tedy nás zajímá

$$\Pr[O_1 \cap O_2 \cap O_3]$$

- Z nezávislosti jevů P_1, P_2 máme i nezávislost jevů O_1, O_2 (tedy jejich doplňků):
  - Chceme: $\Pr[A \cap B] = \Pr[A]\Pr[B]$ právě tehdy když $\Pr[\overline{A} \cap \overline{B}] = \Pr[\overline{A}]\Pr[\overline{B}]$ kde $\overline{A} = \Omega \setminus A$ je doplněk
  - Z minulého příkladu přeuspořádáním dostaneme (pro libovolné jevy)

$$\Pr[A \cap B] = \Pr[A] + \Pr[B] - \Pr[A \cup B]$$

  - Tedy:

$$\Pr[\overline{A} \cap \overline{B}] = \Pr[\overline{A}] + \Pr[\overline{B}] - \Pr[\overline{A} \cup \overline{B}]$$

  - Použijeme že $\overline{A} \cup \overline{B} = \overline{A \cap B}$
  - Tedy píšeme:

$$\begin{aligned}\Pr[\overline{A} \cap \overline{B}] &= \Pr[\overline{A}] + \Pr[\overline{B}] - \Pr[\overline{A} \cup \overline{B}] \\ &= \Pr[\overline{A}] + \Pr[\overline{B}] - \Pr[\overline{A \cap B}] \\ &= (1 - \Pr[A]) + (1 - \Pr[B]) - (1 - \Pr[A \cap B]) \\ &= 1 - \Pr[A] - \Pr[B] + \Pr[A]\Pr[B] \qquad \text{(z nezávislosti } A, B)\end{aligned}$$

  - Chtěli jsme $\Pr[\overline{A} \cap \overline{B}] = \Pr[\overline{A}]\Pr[\overline{B}] = (1 - \Pr[A])(1 - \Pr[B])$, což je akorát jinak napsaný předchozí řádek.

- Z nezávislosti jevů O_1, O_2, O_3 máme rovnou

$$\begin{aligned}\Pr[O_1 \cap O_2 \cap O_3] &= \Pr[O_1]\Pr[O_2]\Pr[O_3] \\ &= (1-\Pr[P_1])(1-\Pr[P_2])(1-\Pr[P_3]) \\ &= (1-p)^3\end{aligned}$$

(b) Druhý příklad je podobný, ale potřebujeme aby aspoň jedna součástka fungovala, tedy chceme

$$\Pr[O_1 \cup O_2 \cup O_3]$$

Jako nápovědu použijte pozorování že $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$.

Jednodušší postup je, uvědomit si, že se nemůžou porouchat všechny a použít nezávislost, tedy

$$\begin{aligned}\Pr[O_1 \cup O_2 \cup O_3] &= 1 - \Pr[P_1 \cap P_2 \cap P_3] \\ &= 1 - \Pr[P_1]\Pr[P_2]\Pr[P_3] \\ &= 1 - p^3\end{aligned}$$

(c) Kombinace myšlenek předchozích.

Samozřejmně můžeme simulovat (všimněte si, jak se podmínka v indikátoru mechanicky překlápí na vzorec pravděpodobnosti):

```
from random import choices
from typing import Sequence

# Pravděpodobnost chyby konkrétní součástky (chyby jsou nezávislé).
p = 0.1
# Počet pokusů.
N = 1000000


def bernoulli_list(k: int, pr: float = p) -> Sequence[bool]:
    """Vrací k samplů z Bernoulliho rozdělení s pravděpodobností pr."""
    return choices([True, False], weights=[pr, 1-pr], k=k)


# a)
#     +-----+   +-----+   +-----+
# o---|  0  |---|  1  |---|  2  |---o
#     +-----+   +-----+   +-----+

def proud_tece_a(soucastky: Sequence[bool]) -> int:
    """Indikátor, jestli proud teče.
    soucastky[i] = i-tá součástka je porouchaná."""
    assert len(soucastky) == 3
    return int(not soucastky[0] and not soucastky[1] and not soucastky[2])


# Kolikrát proud tekl z N pokusů.
proud_tekl_a = sum(proud_tece_a(bernoulli_list(3)) for _ in range(N))
```

```
simulation_a = proud_tekl_a / N
answer_a = (1 - p)**3
error_a = abs(answer_a - simulation_a)
print(f'a) Simulováno: {simulation_a} (chyba {error_a})')


# b)
#          +-----+
#      +---|  0  |---+
#      |   +-----+   |
#      |             |
#      |   +-----+   |
#  o---+---|  1  |---+---o
#      |   +-----+   |
#      |             |
#      |   +-----+   |
#      +---|  2  |---+
#          +-----+

def proud_tece_b(soucastky: Sequence[bool]) -> int:
    """Indikátor, jestli proud teče.
    soucastky[i] = i-tá součástka je porouchaná."""
    assert len(soucastky) == 3
    return int(not (soucastky[0] and soucastky[1] and soucastky[2]))


proud_tekl_b = sum(proud_tece_b(bernoulli_list(3)) for _ in range(N))

simulation_b = proud_tekl_b / N
answer_b = 1 - (p**3)
error_b = abs(answer_b - simulation_b)
print(f'b) Simulováno: {simulation_b} (chyba {error_b})')


# c)
#         +-----+                         +-----+
#     +---|  0  |---+             +-------|  3  |-------+
#     |   +-----+   |   +-----+   |       +-----+       |
# o---+             +---|  2  |---+                     +---o
#     |   +-----+   |   +-----+   |   +-----+   +-----+   |
#     +---|  1  |---+             +---|  4  |---|  5  |---+
#         +-----+                     +-----+   +-----+

def proud_tece_c(soucastky: Sequence[bool]) -> int:
    """Indikátor, jestli proud teče.
    soucastky[i] = i-tá součástka je porouchaná."""
    assert len(soucastky) == 6
    return int(not (soucastky[0] and soucastky[1])
               and not soucastky[2]
               and (not (soucastky[3] and (soucastky[4] or soucastky[5]))))


proud_tekl_c = sum(proud_tece_c(bernoulli_list(6)) for _ in range(N))
```

```
simulation_c = proud_tekl_c / N
answer_c = (1 - (p**2)) * (1 - p) * (1 - (p * (1 - ((1 - p)**2))))
error_c = abs(answer_c - simulation_c)
print(f'c) Simulováno: {simulation_c} (chyba {error_c})')

# Možný výsledek:
# a) Simulováno: 0.728696 (chyba 0.00030400000000082)
# b) Simulováno: 0.999005 (chyba 5.000000000032756e-06)
# c) Simulováno: 0.874083 (chyba 1.2000000000012001e-05)
```

3.2 Cvičení

1. **Házíme cinknutou mincí – hlava padne s pravděpodobností $p \in [0, 1)$, orel padne s pravděpodobností $1 - p$. Házíme opakovaně dokud nepadne hlava.**

   (a) Jak vypadá pravděpodobnostní prostor?

   Řešení:

   - *Množina elementárních jevů:* Série hodů, které uvažujeme můžeme kódovat pomocí H pokud padne hlava, O pokud padne orel, takže bychom mohli možné série hodů reprezentovat jako

     $$\{H, OH, OOH, OOOH, OOOOH, \ldots\}.$$

     Ale to je poněkud nepraktické, raději budeme reprezentovat počtem hodů (poslední je určitě hlava, ty před ním jsou orlové):

     $$\Omega = \{1, 2, 3, 4, \ldots\}$$

   - *Prostor jevů:* Máme spočetnou množinu elementárních jevů, takže můžeme brát jako prostor jevů celou potenční množinu množiny elementárních jevů: $\mathcal{F} = \mathcal{P}(\Omega)$.

     Lehké cvičení z matematické analýzy: dokažte, že pokud každému elementárnímu jevu přiřadíme pravděpodobnost, tedy

     $$\forall \omega \in \Omega\colon \Pr[\{\omega\}] \in [0, 1]$$

     pak víme, že

     –
     $$\Pr[\Omega] = \sum_{\omega \in \Omega} \Pr[\{\omega\}] = 1$$

     –
     $$\forall A \in \mathcal{F}\colon \Pr[A] = \sum_{\omega \in A} \Pr[\{\omega\}]$$

       a tento výraz je dobře definovaný (vzpomeňte na absolutní konvergenci, neklesající posloupnost a omezenou posloupnost).

   - *Pravděpodobnost:* Jak jsme viděli v předchozím bodě, tak stačí určit pravděpodobnost, že hodíme právě n-krát, což je

     $$\Pr[\{n\}] = (1 - p)^{n-1} p \qquad \text{(pro libovolné } n \in \mathbb{N}^+\text{)}$$

     protože napřed musí padnout $n - 1$ orlů a pak jedna hlava.

     Zkontrolujme ještě, že se vše sečte na jedničku. Pro jistotu napřed zopakujme součty geometrické řady.

     $$\begin{aligned} S &= \sum_{j=0}^{n} q^j \\ &= 1 + q + q^2 + \ldots + q^n \\ &= 1 + q\left(1 + q + q^2 + \ldots + q^{n-1}\right) \\ &= 1 + q(S - q^n) \end{aligned}$$

     tedy

     $$S = 1 + q(S - q^n)$$

$$\begin{aligned} S - qS &= 1 - q^{n+1} \\ S &= \frac{1 - q^{n+1}}{1 - q} && \text{(pokud } q \neq 1\text{)} \end{aligned}$$

a pro nekonečný případ

$$\begin{aligned} \sum_{j=0}^{\infty} q^j &= \lim_{n\to\infty} \sum_{j=0}^{n} q^j \\ &= \lim_{n\to\infty} \frac{1 - q^{n+1}}{1 - q} \\ &= \frac{1}{1 - q} && \text{(pokud } |q| < 1\text{)} \end{aligned}$$

Teď už můžeme aplikovat předchozí pro naší pravděpodobnost:

$$\begin{aligned} \sum_{n\in\Omega} \Pr[\{n\}] &= \sum_{n\in\Omega} (1-p)^{n-1} p \\ &= p\left(1 + (1-p) + (1-p)^2 + \ldots\right) \\ &= 1 \end{aligned}$$

(b) Jaká je pravděpodobnost, že hodíme právě třikrát (n-krát)?

Řešení: To už jsme určili v předchozím bodě, ale tato vlastnost je tak důležitá, že to radši zopakujeme:

$$\Pr[\{n\}] = (1-p)^{n-1} p \qquad \text{(pro libovolné } n \in \mathbb{N}^+\text{)}$$

Připomeňme, že tomuto se také někdy říká geometrické rozdělení. Dejte pozor na to, že $0 < p \leq 1$ (proč?). Hodí se, když při hodu kostkou házíme znovu a zajímá nás celkový počet hodů.

(c) Jaká je pravděpodobnost, že hodíme nejvýš třikrát (n-krát)?

Řešení: Využijeme součet geometrické posloupnosti:

$$\begin{aligned} \Pr[\{1, 2, \ldots, n\}] &= \sum_{j=1}^{n} p(1-p)^{j-1} \\ &= p \sum_{j=1}^{n} (1-p)^{j-1} \\ &= p \frac{1 - (1-p)^n}{1 - (1-p)} \\ &= 1 - (1-p)^n \end{aligned}$$

(d) Jaká je pravděpodobnost, že hodíme lišekrát?

Řešení: Opět přímý výpočet:

$$\begin{aligned} \Pr[\{1, 3, 5, 7, \ldots\}] &= \sum_{j=0}^{\infty} p(1-p)^{2j} \\ &= p \sum_{j=0}^{\infty} ((1-p)^2)^j \end{aligned}$$

$$= p \sum_{j=0}^{\infty} ((1-p)^2)^j$$
$$= p \frac{1}{1-(1-p)^2}$$

(e) Simulujte předchozí.

Řešení:

```python
from random import random


def geometric(pr: float = 0.5) -> int:
    """pr is success probability, return the number of tosses until
    the first success."""
    assert pr > 0
    sample = 1
    fail_pr = 1 - pr
    while random() < fail_pr:
        sample += 1
    return sample


N = 1000000  # Pokusů
pr = 0.3

exactly_three_sim = sum(int(geometric(pr) == 3) for _ in range(N)) / N
exactly_three = pr * (1 - pr)**2
print(f'a) Pr[tři] = {exactly_three_sim} (= {exactly_three})')

at_most_three_sim = sum(int(geometric(pr) <= 3) for _ in range(N)) / N
at_most_three = 1 - (1 - pr)**3
print(f'b) Pr[nejvýš tři] = {at_most_three_sim} (= {at_most_three})')

odd_number_sim = sum(int(geometric(pr) % 2 == 1) for _ in range(N)) / N
odd_number = pr / (1 - (1 - pr)**2)
print(f'c) Pr[lišekrát] = {odd_number_sim} (= {odd_number})')

# Možný výstup:
# a) Pr[tři] = 0.147168 (= 0.14699999999999996)
# b) Pr[nejvýš tři] = 0.65664 (= 0.657)
# c) Pr[lišekrát] = 0.58815 (= 0.5882352941176471)
```

2. **Hodíme cinknutou korunou (panna s pravděpodobností $p_1 \in [0,1]$) a cinknutou dvoukorunou (panna s pravděpodobností $p_2 \in [0,1]$). Oba hody jsou na sobě nezávislé.**

   (a) Určete pravděpodobnostní prostor.

   Řešení:

   - *Množina elementárních jevů:*

     $$\Omega = \{P_1P_2, P_1O_2, O_1P_2, O_1O_2\}$$

     kde P je panna O orel a index ukazuje na které minci to padlo.

   - *Prostor jevů:*

     $$\mathcal{F} = \mathcal{P}(\Omega)$$

   - *Pravděpodobnost:* určíme znovu jen na jednoprvkových jevech (na obecném jevu suma):

     $$\begin{aligned}\Pr[\{P_1P_2\}] &= p_1p_2\\ \Pr[\{P_1O_2\}] &= p_1(1-p_2)\\ \Pr[\{O_1P_2\}] &= (1-p_1)p_2\\ \Pr[\{O_1O_2\}] &= (1-p_1)(1-p_2)\end{aligned}$$

   (b) Připoměňte si definici podmíněné pravděpodobnosti (Definice 2.2).

   (c) Spočítejte pravděpodobnost $\Pr[$na obou padne panna $\mid$ na koruně padne panna$]$.

   Řešení: Jev A „na obou padne panna“ je formálně $A = \{P_1P_2\}$, jev B „na koruně padne panna“ je formálně $B = \{P_1P_2, P_1O_2\}$ (má pravděpodobnost ostře větší než jedna). Pak podmíněná pravděpodobnost je:

   $$\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

   tedy

   $$\begin{aligned}\Pr[\{P_1P_2\} \mid \{P_1P_2, P_1O_2\}] &= \frac{\Pr[\{P_1P_2\} \cap \{P_1P_2, P_1O_2\}]}{\Pr[\{P_1P_2, P_1O_2\}]}\\ &= \frac{\Pr[\{P_1P_2\}]}{\Pr[\{P_1P_2, P_1O_2\}]}\\ &= \frac{p_1p_2}{p_1p_2 + p_1(1-p_2)}\\ &= p_2 \quad \text{(což dává smysl, protože ty hody jsou nezávislé)}\end{aligned}$$

   (d) Spočítejte pravděpodobnost $\Pr[$na obou padne panna $\mid$ padne aspoň jedna panna$]$.

   Řešení: Jev A „na obou padne panna“ je formálně $A = \{P_1P_2\}$, jev B „aspoň jedna panna“ je formálně $C = \{P_1P_2, P_1O_2, O_1P_2\}$ (má pravděpodobnost ostře větší než jedna). Pak podmíněná pravděpodobnost je:

   $$\begin{aligned}\Pr[\{P_1P_2\} \mid \{P_1P_2, P_1O_2, O_1P_2\}] &= \frac{\Pr[\{P_1P_2\} \cap \{P_1P_2, P_1O_2, O_1P_2\}]}{\Pr[\{P_1P_2, P_1O_2, O_1P_2\}]}\\ &= \frac{\Pr[\{P_1P_2\}]}{\Pr[\{P_1P_2, P_1O_2, O_1P_2\}]}\end{aligned}$$

$$= \frac{p_1 p_2}{p_1 p_2 + p_1(1-p_2) + (1-p_1)p_2}$$

(e) Simulujte:

Řešení:

```python
from random import random


def toss(weights):
    """True = panna, False = Orel"""
    coins = [False] * len(weights)
    for i in range(len(weights)):
        coins[i] = random() < weights[i]
    return coins


N = 1000000
p1 = 0.1  # panna na první minci
p2 = 0.6  # panna na druhé minci

obe_panna = 0
prvni_je_panna = 0
aspon_jedna_panna = 0

for _ in range(N):
    coins = toss(weights=[p1, p2])
    if all(coins):
        obe_panna += 1
    if coins[0]:
        prvni_je_panna += 1
    if any(coins):
        aspon_jedna_panna += 1

pr_c_sim = obe_panna / prvni_je_panna
pr_c = p2

pr_d_sim = obe_panna / aspon_jedna_panna
pr_d = p1 * p2 / (p1 * p2 + p1 * (1 - p2) + (1 - p1) * p2)

print(f'Pr[obě panna|koruna panna] = {pr_c_sim} (={pr_c})')
print(f'Pr[obě panna|aspoň jedna panna] = {pr_d_sim} (={pr_d})')

# Možný výstup:
# Pr[obě panna|koruna panna] = 0.6017034618418556 (=0.6)
# Pr[obě panna|aspoň jedna panna] = 0.09411977858579801 (=0.09375)
```

3. **Na louce rostou květiny, které mají buď bílé nebo červené květy. Náhodná květina má bílý květ s pravděpodobností $\Pr[B] = 0.6$, tedy pravděpodobnost že náhodná květina má červený květ je $\Pr[C] = 0.4$. Pravděpodobnost že červená květina je jedovatá je $\Pr[J \mid C] = 0.25$. Pravděpodobnost že bílá květina je jedovatá je $\Pr[J \mid B] = 1/12$. Snědli jsme náhodnou rostlinu a je nám zle, jaká je pravděpodobnost, že ta rostlina měla červený květ?**

   Řešení: Řešení 1 – použití Bayesovy věty (Věta 1) jako stroj:

   - Máme Ω, což je množina všech květin.
   - Máme rozklad Ω (víme, že kvetou buď bíle nebo červeně), tedy jevy $B_1 = B, B_2 = C$.
   - Víme $\Pr[J \mid C]$ i $\Pr[J \mid B]$.
   - Zajímá nás $\Pr[C \mid J]$.
   - Dosadíme do vzorce:

$$\begin{aligned}\Pr[C \mid J] &= \frac{\Pr[J \mid C]\Pr[C]}{\Pr[J \mid C]\Pr[C] + \Pr[J \mid B]\Pr[B]} \\ &= \frac{0.25 \cdot 0.4}{0.25 \cdot 0.4 + (1/12) \cdot 0.6} \\ &= 2/3\end{aligned}$$

   Řešení 2 – použijeme představivost a kreslíme:

   - Nechť je na louce 100 kytek.
   - Takže bílých je 60.
   - Červených je 40.
   - Jedovatých červených je 10.
   - Jedovatých bílých je 5.
   - Jedovatých je 15 (to je jmenovatel v Bayesově větě).
   - Pravděpodobnost že kytka je červená když je jedovatá je $10/15 = 2/3$.

   Pár poznámek:

   - Čísla byla hezká, takže druhý postup je jednoduchý.
   - Naprostá většina lidí nezvládá tento typ úlohy. Takže si nejlépe osvojte oba postupy. První je super pro počítač, druhý pro rychlý odhad.

   Můžeme zkusit i naivní simulaci:

```
from enum import Enum
from random import random


def bernoulli(pr: float = 0.5) -> bool:
    return random() < pr


class Color(Enum):
    RED = 1
    WHITE = 2
```

```python
class Plant:
    """Random plant."""
    def __init__(self):
        if bernoulli(0.6):
            self.color = Color.WHITE
            self.is_poisonous = bernoulli(1/12)
        else:
            self.color = Color.RED
            self.is_poisonous = bernoulli(0.25)


N = 1000000  # Pokusů
poisonous = 0  # Kolik jsme viděli jedovatých rostlin.
poisonous_red = 0  # Kolik z těch jedovatých bylo červených

for _ in range(N):
    p = Plant()
    if p.is_poisonous:
        poisonous += 1
        if p.color == Color.RED:
            poisonous_red += 1

assert poisonous > 0, "Pravděpodobnost Pr[A|B] není definovaná pokud Pr[B]=0"
print(f'Viděli jsme {poisonous_red} červených jedovatých rostlin')
print(f'z celkem    {poisonous} jedovatých rostlin')
print(f'tedy Pr[červená|jedovatá] = {poisonous_red/poisonous} (= {2/3})')

# Možný výstup:
# Viděli jsme 100562 červených jedovatých rostlin
# z celkem    150617 jedovatých rostlin
# tedy Pr[červená|jedovatá] = 0.6676669964213867 (= 0.6666666666666666)
```

4. **V první krabici je b bílých míčků a c červených míčků, ve druhé krabici také. Napřed vytáhneme jeden míček z první krabice (uniformně náhodně) a dáme ho do druhé krabice. Pak vytáhneme jeden míček z druhé krabice (uniformně náhodně). Jaká je pravděpodobnost, že míček vytažený z druhé krabice je červený?**

   (a) Navrhněte vhodný pravděpodobnostní prostor.

   Řešení:

   - *Množina elementárních jevů:* $\Omega = \{B_1B_2, B_1C_2, C_1B_2, C_1C_2\}$ kde B_1B_2 značí že jsme z první krabice vytáhli bílý míček (a dali ho do druhé krabice) a pak jsme z druhé krabice vytáhli bílý míček...
   - *Prostor jevů:* $\mathcal{F} = \mathcal{P}(\Omega)$.
   - *Pravděpodobnost:* znovu jen pro jednoprvkové jevy

$$\begin{aligned}
\Pr[\{B_1B_2\}] &= \left(\frac{b}{b+c}\right)\left(\frac{b+1}{b+1+c}\right) \\
\Pr[\{B_1C_2\}] &= \left(\frac{b}{b+c}\right)\left(\frac{c}{b+1+c}\right) \\
\Pr[\{C_1B_2\}] &= \left(\frac{c}{b+c}\right)\left(\frac{b}{b+1+c}\right) \\
\Pr[\{C_1C_2\}] &= \left(\frac{c}{b+c}\right)\left(\frac{c+1}{b+1+c}\right)
\end{aligned}$$

   (b) Spočítejte tu pravděpodobnost, kterou jsme chtěli.

   Řešení: Použijeme podmíněnou pravděpodobnost (a projdeme možnosti, co se stane).

$$\begin{aligned}
\Pr[\text{2. červený}] &= \Pr[\text{2. červený} \mid \text{1. bílý}]\Pr[\text{1. bílý}] + \Pr[\text{2. červený} \mid \text{1. červený}]\Pr[\text{1. červený}] \\
&= \frac{c}{b+1+c}\frac{b}{b+c} + \frac{c+1}{b+1+c}\frac{c}{b+c} \\
&= \frac{bc}{(b+c)(b+1+c)} + \frac{c(c+1)}{(b+c)(b+1+c)} \\
&= \frac{c(b+c+1)}{(b+c)(b+1+c)} \\
&= \frac{c}{b+c}
\end{aligned}$$

   A to samé bychom dostali i přímo z rozepsání:

$$\begin{aligned}
\Pr[\{B_1C_2, C_1C_2\}] &= \left(\frac{b}{b+c}\right)\left(\frac{c}{b+1+c}\right) + \left(\frac{c}{b+c}\right)\left(\frac{c+1}{b+1+c}\right) \\
&= \dots \\
&= \frac{c}{b+c}
\end{aligned}$$

   (c) Simulujte.

   Řešení:

```
import random

bilych = 15
```

```
cervenych = 37

kbelik_1 = ['B'] * bilych + ['C'] * cervenych
kbelik_2 = ['B'] * bilych + ['C'] * cervenych

N = 1000000
cervenych_z_druheho = 0
for _ in range(N):
    druhy_kbelik = kbelik_2 + [random.choice(kbelik_1)]
    assert len(druhy_kbelik) == 1 + len(kbelik_1)
    if random.choice(druhy_kbelik) == 'C':
        cervenych_z_druheho += 1

vysledek = cervenych / (cervenych + bilych)
print(f'Nasimulovali jsme {cervenych_z_druheho/N} (={vysledek})')

# Možný výstup:
# Nasimulovali jsme 0.711212 (=0.7115384615384616)
```

5. **Tento příklad je vymyšlený, zejména čísla nesedí a reálný svět je malinko složitější (tím se ještě budeme zabývat), ale informace v něm nejsou daleko od pravdy. V zemi nám řádí nemoc C.**

   - **Prostor elementárních jevů (sample space) Ω jsou všichni občané.**
   - **Označíme $C^+ \subseteq \Omega$ množinu všech lidí, kteří dnes mají aktivní nemoc C, označíme $C^- = \Omega \setminus C^+$ zdravé lidi.**
   - **Umíme uniformně náhodně samplovat lidi, tedy $\forall \omega \in \Omega\colon \Pr[\{\omega\}] = 1/|\Omega|$ (tady ω je jeden člověk).**
   - **Test nám pro libovolného člověka odpoví že je člověk zdravý nebo nemocný. Značme $T^+ \subseteq \Omega$ množinu lidí pro které test odpoví, že jsou nemocní. Značme $T^- = \Omega \setminus T^+$ množinu lidí pro které test odpoví, že jsou zdraví. Ale není to tak jednoduché, v příbalovém letáku testu se píše:**
     - ***Sensitivity:* (true positive) $\Pr[T^+ \mid C^+] = 0.9$**
     - ***Specificity:* (true negative) $\Pr[T^- \mid C^-] = 0.8$**

     z tohoto můžeme odvodit chyby:
     - **False positive = false alarm = type I error**
     $$\Pr[T^+ \mid C^-] = 1 - \Pr[T^- \mid C^-] = 0.2$$
     - **False negative = miss = type II error**
     $$\Pr[T^- \mid C^+] = 1 - \Pr[T^+ \mid C^+] = 0.1$$
   - **Provedli jsme jeden test u každého z uniformně náhodně vybraných 50000 lidí a pozitivních testů vyšlo 1000. Tedy předpokládáme, že $\Pr[T^+] = \frac{1000}{50000} = \frac{1}{50} = 0.02$ (jak moc je tento předpoklad oprávněný budeme zkoumat nadále).**
   - **Zajímá nás $\Pr[C^+|$ (vynásobeno 100 nám dá počet nemocných v procentech).**

   (a) V čem se toto liší od reality?

   Řešení:

   - Zejména v tom náhodném testování. Uvědomte si, že trasování nevybírá lidi náhodně. Ve skutečnosti je velmi těžké vybrat náhodného člověka (více o tom později).
   - Sensitivita a specificita jsou opravdu důležité parametry testu (nezávisí na tom jaké je procento nemocných). Jejich výhoda je, že pokud známe četnost nemoci v populaci, pak můžeme pomocí Bayesovy věty spočítat pravděpodobnost, že náhodně vybraný člověk je nemocný, pokud test vyšel pozitivní. Pak se stejně dělá ještě další test, abychom si byli jistí (viz domácí úkol).
   - Sensitivita a specificita se určují experimentálně, tedy je neznáme přesně (můžete zkusit v simulaci co to udělá).
   - Milion komplikací při popisu skutečného světa, například nevíme na čem závisí pravděpodobnost chyby prvního nebo druhého druhu (třeba je pro danou krevní skupinu false positive pravděpodobnější...).

   (b) Spočítejte $\Pr[C^+]$.

   Řešení: Využijeme větu o úplné pravděpodobnosti: pokud $B_1, B_2 \in \mathcal{F}$ je rozklad Ω (tedy $B_1 \cap B_2 = \emptyset$ a navíc $B_1 \cup B_2 = \Omega$, připomeňme že věta platí i pro spočetný rozklad $B_1, B_2, \ldots$), pak pro libovolné $A \in \mathcal{F}$ máme
   $$\Pr[A] = \Pr[A \mid B_1] \Pr[B_1] + \Pr[A \mid B_2] \Pr[B_2]$$

Aplikujeme předchozí:

$$\begin{aligned}\Pr[T^+] &= \Pr[T^+ \mid C^+]\Pr[C^+] + \Pr[T^+ \mid C^-]\Pr[C^-] \\ &= \Pr[T^+ \mid C^+]\Pr[C^+] + \Pr[T^+ \mid C^-](1 - \Pr[C^+])\end{aligned}$$

přeuspořádáme

$$\begin{aligned}\Pr[T^+] &= \Pr[T^+ \mid C^+]\Pr[C^+] + \Pr[T^+ \mid C^-](1 - \Pr[C^+]) \\ \Pr[T^+] &= \Pr[T^+ \mid C^+]\Pr[C^+] + \Pr[T^+ \mid C^-] - \Pr[T^+ \mid C^-]\Pr[C^+] \\ \Pr[T^+] &= (\Pr[T^+ \mid C^+] - \Pr[T^+ \mid C^-])\Pr[C^+] + \Pr[T^+ \mid C^-] \\ \Pr[C^+] &= \frac{\Pr[T^+] - \Pr[T^+ \mid C^-]}{\Pr[T^+ \mid C^+] - \Pr[T^+ \mid C^-]} \\ \Pr[C^+] &= \frac{0.02 - 0.2}{0.9 - 0.2} \\ \Pr[C^+] &\approx -0.257\end{aligned}$$

(c) Co se stalo špatně?

Řešení: Taková data bychom nečekali ani kdyby všichni byli zdraví (vyšlo nám příliš málo pozitivních).

Erratum:

i. Původně bylo: Takže jsme rozhodně netestovali náhodný vzorek populace.

ii. Problém tohoto vysvětlení: Ani kdybychom testovali jen zdravé lidi, tak by to nebylo dobré vysvětlení.

iii. Lepší pokus o vysvětlení:

- Možná, že parametry testu byly odhadnuty chybně.
- Je možné, že laboratoř omylem poslala špatná data (například jeden laborant prohodil počet pozitivních a negativních výsledků u svých testů).
- Je možné, že náhodou testy fungovaly mnohem lépe, než měly. Třeba jsme dostali várku nečekaně přesných testů. Speciálně není nemožné, že 20 hodů spravedlivou mincí nám dá 20 hlav, je to jen extrémně nepravděpodobné. Odhadem pravděpodobnosti takovéto chyby se budeme v rámci předmětu ještě zabývat.

(d) Jak by vyšlo předchozí kdyby $\Pr[T^+ \mid C^+] = 0.99, \Pr[T^- \mid C^-] = 0.98, \Pr[T^+] = 0.2$**?**

Řešení:

$$\begin{aligned}\Pr[C^+] &= \frac{\Pr[T^+] - \Pr[T^+ \mid C^-]}{\Pr[T^+ \mid C^+] - \Pr[T^+ \mid C^-]} \\ &= \frac{0.2 - 0.02}{0.99 - 0.02} \\ &\approx 0.185\end{aligned}$$

Což dává smysl (je spíš pravděpodobné, že zdravého chybně označíme za nemocného než naopak).

(e) Simulujte předchozí.

Řešení:

```python
from random import random


def bernoulli(pr: float = 0.5) -> bool:
    return random() < pr


class Human:
    """ _illness_probability is our unknown! """
    _illness_probability = 0.185

    """Random human."""
    def __init__(self):
        self.is_ill = bernoulli(Human._illness_probability)


class IllnessTest:
    sensitivity = 0.99  # = Pr[T+|C+]
    specificity = 0.98  # = Pr[T-|C-]

    def test(h: Human) -> bool:
        """Return True if the test says h is ill."""
        if h.is_ill:
            return bernoulli(IllnessTest.sensitivity)
        else:
            return not bernoulli(IllnessTest.specificity)


N = 50000  # Number of samples
false_positives = 0
true_positives = 0
ill_humans = 0

for _ in range(N):
    h = Human()
    if h.is_ill:
        ill_humans += 1
        if IllnessTest.test(h):
            # The test is positive and the human is ill.
            true_positives += 1
    else:
        if IllnessTest.test(h):
            # The test is positive and the human is healthy.
            false_positives += 1

pr_positive_test = (true_positives + false_positives) / N
pr_true_positive = true_positives / ill_humans
pr_false_positive = false_positives / (N - ill_humans)
illness_estimate = ((pr_positive_test - pr_false_positive)
                    / (pr_true_positive - pr_false_positive))

print(f'Pr[positive test]={pr_positive_test}')
print(f'Pr[false positive]={pr_false_positive} (={1-IllnessTest.specificity})')
print(f'Pr[true positive]={pr_true_positive} (={IllnessTest.sensitivity})')
```

```
print(f'Pr[nemoc]={illness_estimate} (={Human._illness_probability})')

# Možný výstup:
# Pr[positive test]=0.1976
# Pr[false positive]=0.01945124938755512 (=0.020000000000000018)
# Pr[true positive]=0.989760348583878 (=0.99)
# Pr[nemoc]=0.1836 (=0.185)
```

6. **V šuplíku mám $b \in \mathbb{N}$ párů bílých, $c \in \mathbb{N}$ párů černých ponožek a $s \in \mathbb{N}$ párů sepraných ponožek. Potřebuju si vytáhnout čtyři páry černých ponožek (jedu na prodloužený víkend tancovat). Když vytáhnu čtyři náhodné páry ponožek (mám je napárované v šuplíku), jaká je pravděpodobnost, že všechny budou černé?**

Řešení: Dle definice podmíněné pravděpodobnosti (Definice 2.2) můžeme napsat (která pravděpodobnost musí být nenulová?):

$$\Pr[A \cap B] = \Pr[A] \Pr[B \mid A]$$
$$\Pr[A \cap B \cap C] = \Pr[A] \Pr[B \mid A] \Pr[C \mid A \cap B]$$

Tedy můžeme psát C_1, C_2, C_3, C_4 jevy, že první, druhý, třetí, čtvrtý pár vytažených ponožek jsou černé.

$$\begin{aligned}\Pr[C_1 \cap C_2 \cap C_3 \cap C_4] &= \Pr[C_1] \Pr[C_2 \mid C_1] \Pr[C_3 \mid C_1 \cap C_2] \Pr[C_4 \mid C_1 \cap C_2 \cap C_3] \\ &= \left(\frac{c}{b+c+s}\right)\left(\frac{c-1}{b+c-1+s}\right)\left(\frac{c-2}{b+c-2+s}\right)\left(\frac{c-3}{b+c-3+s}\right)\end{aligned}$$

Mohli bychom spočítat kolik je čtveřic černých ze všech čtveřic (ale předchozí postup bývá užitečný):

$$\Pr[C_1 \cap C_2 \cap C_3 \cap C_4] = \frac{\binom{c}{4}}{\binom{b+c+s}{4}}$$

```
from random import sample
from scipy.special import comb

b = 10  # bilych ponozek
c = 15  # cernych ponozek
s = 5   # sepranych ponozek

suplik = ['B'] * b + ['C'] * c + ['S'] * s

N = 1000000
cernych = 0
for _ in range(N):
    vyber = sample(suplik, k=4)
    if vyber == ['C'] * 4:
        cernych += 1

pr_vsechny_cerne = c*(c-1)*(c-2)*(c-3) / ((b+c+s)*(b+c-1+s)*(b+c-2+s)*(b+c-3+s))
print(f'Pr[4 cerne] = {cernych/N} (={pr_vsechny_cerne})')

pr_vsechny_cerne_komb_cislo = comb(c, 4, exact=True) / comb(b+c+s, 4, exact=True)
assert abs(pr_vsechny_cerne - pr_vsechny_cerne_komb_cislo) < 0.000001

# Možný výstup:
# Pr[4 cerne] = 0.049479 (=0.04980842911877394)
```

3.3 Cvičení

1. **Rozmysleme si, proč nezávislost více jevů není to samé jako nezávislost po dvou.**

   (a) Najděte jevy A, B, C takové, že jevy jsou po dvou nezávislé, ale $\Pr[A\cap B\cap C] \neq \Pr[A]\Pr[B]\Pr[C]$.

   Řešení: Mějme pravděpodobnostní prostor hod dvěma spravedlivými mincemi $\Omega = \{HH, HO, OH, OO\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\Pr[HH] = \Pr[HO] = \Pr[OH] = \Pr[OO] = 1/4$. Mějme jevy

   $$A = \{HH, HO\}$$
   $$B = \{HH, OH\}$$
   $$C = \{HO, OH\}$$

   tedy

   $$\Pr[A] = \Pr[B] = \Pr[C] = 1/2$$

   Ty jsou po dvou nezávislé:

   $$\Pr[A \cap B] = \Pr[\{HH\}] = 1/4$$
   $$\Pr[A \cap C] = \Pr[\{HO\}] = 1/4$$
   $$\Pr[B \cap C] = \Pr[\{OH\}] = 1/4$$

   Ale

   $$\Pr[A \cap B \cap C] = \Pr[\emptyset] = 0 \neq 1/8 = \Pr[A]\Pr[B]\Pr[C]$$

   (b) Najděte jevy A, B, C takové, že $\Pr[A \cap B \cap C] = \Pr[A]\Pr[B]\Pr[C]$, ale jevy nejsou po dvou nezávislé.

   Řešení: Jedno z možných řešení: Uvažme pravděpodobnostní prostor s elementárními jevy $\Omega = \{a, b, c, z\}$ a jevy $A = \{a, z\}$, $B = \{b, z\}$, $C = \{c, z\}$. Nechť $\Pr[\{a\}] = \Pr[\{b\}] = \Pr[\{c\}] = p$ a $\Pr[z] = q$. Aby bylo $\Pr[\Omega] = 1$, musí platit $3p + q = 1$, tedy $q = 1 - 3p$. Pak máme $\Pr[A] = \Pr[B] = \Pr[C] = p + q = 1 - 2p$.

   Chceme, aby jevy byly po třech nezávislé, tedy $\Pr[A \cap B \cap C] = \Pr[A]\Pr[B]\Pr[C]$. Průnik všech tří obsahuje jenom z, takže má pravděpodobnost q, všechny tři jevy na pravé straně mají pravděpodobnost $1 - 2p$. Takže musí platit $q = (1 - 2p)^3$, tedy $1 - 3p = (1 - 2p)^3$. To je kubická rovnice s jedním kořenem v intervalu $(0, 1)$, konkrétně

   $$q = \frac{3 - \sqrt{3}}{4} \doteq 0.317.$$

   Z toho dostaneme $p = 1 - 3q \doteq 0.049$. Máme tedy A, B, C po třech nezávislé.

   Jak je to s nezávislostí po dvou? Průniky dvojic obsahují zase jenom z, takže by muselo platit $q = (1 - 2p)^2$, tedy $1 - 3p = (1 - 2p)^2$. Ale jelikož $1 - 3p$ je už rovno $(1 - 2p)^3$, nemůže to být současně $(1 - 2p)^2$, leda by bylo $1 - 2p = 1$, čili $p = 0$.

2. **Házíte dvěma rozlišitelnými kostkami.**

   (a) Určete vhodný pravděpodobnostní prostor.

   Řešení:

   - Množina elementárních jevů $\Omega = \{11, 12, 13, 14, 15, 16, 21, 22, \ldots, 66\}$ ($|\Omega| = 36$).
   - Prostor jevů $\mathcal{F} = \mathcal{P}(\Omega)$.
   - Pravděpodobnost je pro každý elementární jev stejná, tedy $\Pr[\{xy\}] = 1/36$ kde $x, y \in \{1, 2, 3, 4, 5, 6\}$.

   (b) Spočítejte pravděpodobnost, že aspoň na jedné kostce padla šestka, když víte jaký součet padl.

   Řešení: Zajímá nás

   $$\Pr[\{xy\} \mid x + y = k] \qquad (\text{kde } x, y \in \{1, 2, 3, 4, 5, 6\},\ k \in \{2, 3, \ldots, 12\})$$

   Napřed zkusme jen chvíli přemýšlet, například pokud $k = 2$, tak určitě víme, že ani na jedné kostce nepadla šestka (protože bychom dostali součet aspoň sedm). Takže už víme

   $$\Pr[\{xy\} \mid x + y = 2] = \Pr[\{xy\} \mid x + y = 3] = \ldots = \Pr[\{xy\} \mid x + y = 6] = 0$$

   Pokud bychom na to šli z druhé strany, tak pokud součet je aspoň jedenáct, tak určitě aspoň na jedné kostce padla šestka (součet deset jde získat jako součet dvou pětek).

   $$\Pr[\{xy\} \mid x + y = 12] = \Pr[\{xy\} \mid x + y = 11] = 1$$

   Jak tedy dopadne pravděpodobnost, když součet je deset? Označme si jev

   $$D = \{46, 55, 64\}$$

   kdy padl součet deset. Označme si jev aspoň jedna šestka

   $$S = \{16, 26, 36, 46, 56, 66, 61, 62, 63, 64, 65\}\,.$$

   $$\begin{aligned}\Pr[S \mid D] &= \frac{\Pr[S \cap D]}{\Pr[D]} \\ &= \frac{\Pr[\{46, 64\}]}{\Pr[\{46, 55, 64\}]} \\ &= \frac{2}{3}\end{aligned}$$

   Obdobně bychom mohli spočítat zbytek. Tohle je spíš práce pro počítač (takhle malý pravděpodobnostní prostor můžeme probrat celý).

```
# soucet[k] = kolikrát jsme viděli součet k
soucet = [0] * 13
# sestek[k] = kolikrát jsme viděli aspoň jednu šestku, když součet byl k
sestek = [0] * 13

for i in range(1, 7):
    for j in range(1, 7):
        soucet[i + j] += 1
```

```
            if (i == 6) or (j == 6):
                sestek[i + j] += 1

for s in range(2, 13):
    print(f'Pr[aspoň jedna šestka | součet = {s}] = {sestek[s]}/{soucet[s]}')

# Výstup:
# Pr[aspoň jedna šestka | součet = 2] = 0/1
# Pr[aspoň jedna šestka | součet = 3] = 0/2
# Pr[aspoň jedna šestka | součet = 4] = 0/3
# Pr[aspoň jedna šestka | součet = 5] = 0/4
# Pr[aspoň jedna šestka | součet = 6] = 0/5
# Pr[aspoň jedna šestka | součet = 7] = 2/6
# Pr[aspoň jedna šestka | součet = 8] = 2/5
# Pr[aspoň jedna šestka | součet = 9] = 2/4
# Pr[aspoň jedna šestka | součet = 10] = 2/3
# Pr[aspoň jedna šestka | součet = 11] = 2/2
# Pr[aspoň jedna šestka | součet = 12] = 1/1
```

Mohli bychom i simulovat i přesně počítat pomocí jazyka R, skript napsal Robert Šámal:

```
---
title: "cvic1-simulace kostek"
output:
  pdf_document: default
  html_notebook: default
---

# Simulace

Napřed jednoduché hrátky s házením jednou kostkou.
Příkaz sample vrací datový typ vector, s tím se dá vektorově pracovat.

```{r}
N=10
kostka = sample(1:6, N, replace=TRUE)
kostka
#2*kostka
#kostka==1
#sum(kostka==1)
#kostka[c(1,2,3,4)]
#kostka[kostka<=3]
#sum(kostka==1)+sum(kostka==2)+sum(kostka==3)+sum(kostka==4)+sum(kostka==5)+sum(kostka==
```

A teď se dostáváme k simulaci domácího úkolu s kostkami.
Všimněte si, jak podmíněná pravděpodobnost znamená vlastně to,
že se omezíme (v čitateli i ve jmenovateli) na ty souřadnice, kde platí podmiňující jev.

```{r}
N = 10^4

kostka1 = sample(1:6, N, replace=TRUE)
kostka2 = sample(1:6, N, replace=TRUE)
soucet = kostka1 + kostka2
```
```

```
SD = soucet==10
PS = kostka1==6
NS = kostka1==6 | kostka2==6

cat("\nP(SD)=", sum(SD)/N)
cat("\nP(PS)=", sum(PS)/N)
cat("\nP(NS)=", sum(NS)/N)

cat("\nP(PS|SD)=", sum(PS & SD)/sum(SD))
cat("\nP(NS|SD)=", sum(NS & SD)/sum(SD))
cat("\nP(PS|NS)=", sum(PS & NS)/sum(NS))
cat("\nP(SD|NS)=", sum(SD & NS)/sum(NS))
cat("\nP(NS|PS)=", sum(NS & PS)/sum(PS))
cat("\nP(SD|PS)=", sum(SD & PS)/sum(PS))
cat("\n")
c(6/11, 1/6, 1/12, 1/3, 2/3, 2/11,11/36)
```
```
# Projití celého pravděpodobnostního prostoru

Funguje jen pro malé prostory, jinak trvá moc dlouho!

```{r}
Omega = expand.grid(k1=1:6,k2=1:6)
kostka1 = Omega$k1; kostka1
kostka2 = Omega$k2; kostka2
soucet = kostka1 + kostka2
N = length(kostka1)
SD = soucet==10

Omega$soucet = soucet
Omega$SD = SD
Omega

SD
PS = kostka1==6
NS = kostka1==6 | kostka2==6

cat("\nP(SD)=", sum(SD)/N)
cat("\nP(PS)=", sum(PS)/N)
cat("\nP(NS)=", sum(NS)/N)

cat("\nP(PS|SD)=", sum(PS & SD)/sum(SD))
cat("\nP(NS|SD)=", sum(NS & SD)/sum(SD))
cat("\nP(PS|NS)=", sum(PS & NS)/sum(NS))
cat("\nP(SD|NS)=", sum(SD & NS)/sum(NS))
cat("\nP(NS|PS)=", sum(NS & PS)/sum(PS))
cat("\nP(SD|PS)=", sum(SD & PS)/sum(PS))
cat("\n")
```
```

3. **V truhle je sto mincí. Z nich 99 je normálních, ale jedna má na obou stranách orla.**

   (a) Určete vhodný pravděpodobnostní prostor.

   Řešení:

   - Množina elementárních jevů (mince jsou očíslované):

$$\begin{aligned}\Omega = \{ & \\ & 1-HHHHHH, \\ & 1-HHHHHO, \\ & \dots \\ & 1-OOOOOO, \\ & 2-HHHHHH, \\ & \dots \\ & 99-HHHHHH, \\ & \dots \\ & 99-OOOOOO, \\ & 100-OOOOOO \\ \}&\end{aligned}$$

   - Prostor jevů $\mathcal{F} = \mathcal{P}(\Omega)$.
   - Pravděpodobnost vytažení každé mince je stejná. Pravděpodobnost že spravedlivou mincí naházíme něco je stejná jako že s ní naházíme něco jiného.

$$\Pr[j-ABCDEF] = \frac{1}{100 \cdot 2^6}$$
$$\text{(pro každé } 1 \le j \le 99,\ A,B,C,D,E,F \in \{H,O\})$$
$$\Pr[100-OOOOOO] = \frac{1}{100}$$

   (b) Vytáhneme náhodnou minci a šestkrát s ní hodíme, pokaždé padne orel. Jaká je pravděpodobnost, že jsme si vytáhli dvouorlovou minci? (Zkuste napřed odhadnout, pak spočítat.)

   Řešení: Značme

$$O_2 = \{100-OOOOOO\}$$
$$O_6 = \{1-OOOOOO, 2-OOOOOO, \dots, 100-OOOOOO\}$$

   pak píšeme

$$\begin{aligned}\Pr\left[\text{dvouorlová} \mid \text{šest orlů}\right] &= \Pr\left[O_2 \mid O_6\right] \\ &= \frac{\Pr\left[O_2 \cap O_6\right]}{\Pr\left[O_6\right]} \\ &= \frac{\Pr\left[O_2\right]}{\Pr\left[O_6\right]} \\ &= \frac{\frac{1}{100}}{99\frac{1}{100\cdot 2^6} + \frac{1}{100}} \\ &\approx 0.3826\end{aligned}$$

(c) Simulujte.

Řešení:

```
from random import choice
from random import choices


class FairCoin:
    dvou_orlova = False
    def toss(self):
        return choices([True, False], k=6)


class DvouOrlova:
    dvou_orlova = True
    def toss(self):
        return [True] * 6


mince = [FairCoin()] * 99 + [DvouOrlova()]

N = 10000000
sest_orlu = 0
dvouorlovych = 0

for _ in range(N):
    m = choice(mince)
    if all(m.toss()):
        sest_orlu += 1
        if m.dvou_orlova:
            dvouorlovych += 1

exact = 0.01 / (0.01 + 0.99*2**(-6))
print(f'Pr[dvouorlova|OOOOOO] = {dvouorlovych/sest_orlu} (={exact})')

# Možný výstup:
# Pr[dvouorlova|OOOOOO] = 0.3895867899186221 (=0.39263803680981596)
```

4. **Připomeňme co je náhodná veličina a její střední hodnota.**

   Řešení: Například objem alkoholu.

```python
from random import choice

alcohol_content = {
        "světlé pivo" : 0.045 * 0.5,  # 4,5% 0.5 litru
        "tmavé pivo" : 0.04 * 0.5,  # 4% 0.5 litru
        "slivovice" : 0.51 * 0.04,  # 51% 0.04 litru
        "bílé víno" : 0.11 * 0.2,  # 11% 0.2 litru
        "červené víno" : 0.11 * 0.2,  # 11% 0.2 litru
        "čaj" : 0.52 * 0.04,  # 52% 0.04 litru
        }


def drink():
    # Vrací jeden elementární jev z Omega.
    return choice(list(alcohol_content.keys()))


def X(drink):
    # Vrací objem vypitého alkoholu v litrech.
    # X: Omega -> R
    return alcohol_content[drink]


N = 100
alkohol = 0.0
for _ in range(N):
    alkohol += X(drink())

EX = sum(alcohol_content.values()) / len(alcohol_content)

print(f'Střední hodnota objemu alkoholu v jednom drinku je {alkohol / N} (= {EX})')

# Možný výstup:
# E[alkohol] je 0.02108899999999999 (= 0.021283333333333335)
```

$$
\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \operatorname{Im}(x)} x \Pr[X = x] \\
&= \sum_{x \in \operatorname{Im}(x)} x \Pr[\{\omega \in \Omega \mid X(\omega) = x\}] \\
&= 0.0225 \Pr[\{\text{světlé pivo}\}] \\
&\quad + 0.02 \Pr[\{\text{tmavé pivo}\}] \\
&\quad + 0.0204 \Pr[\{\text{slivovice}\}] \\
&\quad + 0.022 \Pr[\{\text{červené víno}, \text{bílé víno}\}] \\
&\quad + 0.0208 \Pr[\{\text{čaj}\}]
\end{aligned}
$$

Co kdyby pravděpodobnost kostky nebyla uniformní?

Řešení: Vzorec se nemění, akorát se mění pravděpodobnost jevů $\Pr[X = x]$.

5. **Na stole jsou dvě obálky, v jedné je k korun, ve druhé ℓ korun ($k, \ell \in \mathbb{N}$). Můžete otevřít jednu obálku a na základě sumy v ní se rozhodnout jestli si necháte tu otevřenou nebo si vezmete tu druhou (nehledě na to, kolik je v té druhé). Umíte vymyslet způsob jak odejít s tou s větším obnosem s pravděpodobností ostře větší než jedna polovina? Určete střední hodnotu výhry.**

   Řešení: Uniformně náhodně zvolíme první obálku. Pokud vidíme m korun, pak házíme spravedlivou mincí, dokud nepadne hlava. Pokud celkový počet hodů je ostře menší než m, pak si obálku necháme, jinak si vezmeme tu druhou. Když $k < \ell$ tak pravděpodobnost, že vyměníme obálku s k korunami je ostře větší než pravděpodobnost, že vyměníme obálku s ℓ korunami.

   Vzpomeňte na geometrické rozdělení z minulého cvičení. Pravděpodobnost, že odejdu s k korunami je

$$\begin{aligned}\Pr[\text{dostanu } k \text{ Kč}] &= \frac{1}{2}(1 - 0.5^{k-1}) + \frac{1}{2}0.5^{\ell-1} \\ &= \frac{1}{2} - 0.5^k + 0.5^\ell \\ &= \frac{1}{2} + (0.5^\ell - 0.5^k)\end{aligned}$$

   Střední hodnota výhry

$$\mathbb{E}[\text{win}] = k\left(\frac{1}{2} + (0.5^\ell - 0.5^k)\right) + \ell\left(\frac{1}{2} + (0.5^k - 0.5^\ell)\right)$$

```python
from random import randint
from random import random


def geometric(pr: float = 0.5) -> int:
    """pr is success probability, return the number of tosses until
    the first success."""
    assert pr > 0
    sample = 1
    fail_pr = 1 - pr
    while random() < fail_pr:
        sample += 1
    return sample


# Our unknown amounts.
envelopes = [5, 10]

N = 1000000       # Number of samples.
total_amount = 0  # Total sum that we got during all samples.
got_larger = 0    # Number of times we walked away with the larger sum.

for _ in range(N):
    # Pick the first envelope at random.
    chosen = randint(0, 1)
    if geometric() < envelopes[chosen]:
        # Keep this one.
        pass
    else:
```

```
            # Choose the other.
            chosen = 1 - chosen
        if envelopes[chosen] >= envelopes[1 - chosen]:
            got_larger += 1
        total_amount += envelopes[chosen]

print(f'Pr[selected larger] = {got_larger / N}')
print(f'E[win] = {total_amount / N}')

# Possible outcome:
# Pr[selected larger] = 0.530087
# E[win] = 7.650435
```

6. **Spočítejte střední počet porovnání quick-sortu:**

```
from random import randint


def partition(arr, begin, end):
    pivot_i = randint(begin, end - 1)
    (arr[pivot_i], arr[end-1]) = (arr[end-1], arr[pivot_i])
    pivot = arr[end - 1]
    i = begin
    for j in range(begin, end):
        if arr[j] < pivot:
            (arr[i], arr[j]) = (arr[j], arr[i])
            i += 1
    (arr[i], arr[end-1]) = (arr[end-1], arr[i])
    return i


def quick_sort(arr, begin, end):
    if end <= begin:
        return
    p = partition(arr, begin, end)
    quick_sort(arr, begin, p)
    quick_sort(arr, p+1, end)
```

(a) Uvědomte si, že každá dvě čísla porovnáte nejvýš jednou (pokud se žádné číslo neopakuje). Pro jednoduchost budeme předpokládat, že se čísla neopakují (jinak bychom museli mluvit o jejich pozici v utříděném poli).

(b) Vytvořte vhodný pravděpodobnostní prostor.

Řešení: Tohle je příklad ne příliš pěkného pravděpodobnostního prostoru. Uvidíte, že daleko lépe se bude pracovat s náhodnými proměnnými.

Ale běh algoritmu je dán permutací na vstupu a volbami pivotů. Tedy například můžeme vzít jako jeden elementární jev vstupní permutaci a stack-trace algoritmu (na které intervaly se rekurzí a jaké pivoty se volí).

Prostor jevů bude potenční množina (množina elementárních jevů je konečná).

Navrhnout pravděpodobnost není tak jednoduché (zkuste). Počítat s ní není nic moc příjemného.

I kdybychom si uvědomili, že doba běhu naší implementace nezávisí na vstupní permutaci (místo výběru pivota někde bychom ho vybrali jinde), tak tento pravděpodobnostní prostor není žádný med.

(c) Definujte náhodnou proměnnou určující počet porovaných dvojic, vyjádřete ji jako součet jednodušších a použijte větu o linearitě střední hodnoty.

Řešení: Zajímá nás střední hodnota C, což je počet porovnání. Každé dva prvky výsledku potenciálně mohou být porovnány. Tedy volme $C_{i,j}$ náhodnou proměnnou, která je rovna jedné pokud i-té nejmenší číslo je porovnáno s j-tým nejmenším číslem (kde bereme $1 \leq i < j \leq n$) a nule jinak. Takovým $C_{i,j}$ říkáme *indikátorová proměnná*.

Uvědomme si, že pokud nějaké k-té nejmenší číslo kde $i < k < j$ je zvoleno jako pivot před tím, než se pivotem stane i nebo j, tak $C_{i,j} = 0$, ale jinak je rovná jedné. Tedy

můžeme psát

$$\Pr[C_{i,j} = 1] = \frac{2}{j - i + 1}$$

Nás zajímá střední hodnota

$$\begin{aligned}\mathbb{E}[C_{i,j}] &= 1 \cdot \frac{2}{j-i+1} + 0 \cdot \left(1 - \frac{2}{j-i+1}\right) \\ &= \frac{2}{j-i+1}\end{aligned}$$

Ve skutečnosti nás ale zajímá střední hodnota počtu všech porovnání

$$\begin{aligned}\mathbb{E}[C] &= \mathbb{E}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_{i,j}\right] \\ &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbb{E}[C_{i,j}] && \text{(linearita střední hodnoty)} \\ &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{2}{j-i+1} \\ &= \sum_{i=1}^{n-1}\sum_{d=1}^{n-i} \frac{2}{d+1} && (d = j - i) \\ &= \sum_{i=1}^{n-1} 2(H_{n-i+1} - 1) \\ &\leq \sum_{i=1}^{n-1} 2\ln(n) \\ &= 2n\ln(n)\end{aligned}$$

(d) S jakou pravděpodobností provede quick-sort aspoň $10n\ln(n)$ porovnání?

- **Sčítáme n kladných reálných čísel $a_1, a_2, \ldots, a_n \in [0, \infty)$. Víme, že**

$$\sum_{i=1}^{n} a_i = S$$

Pro kolik z těch čísel platí $a_j \geq 5S/n$?

Řešení: Nejvýš pro $n/5$ z těch čísel. I kdyby ostatní byla nulová a tato velká byla přesně tolik, pak $(n/5)(5S/n) = S$.

- **Co kdybychom ta čísla sčítali váženě? Tedy formálně: mějme náhodnou proměnnou o které víme $\Pr[X = j]$ pro $j \in \{1, 2, \ldots, m\}$ (kde pro jednoduchost předpokládáme $\sum_{j=1}^{m} \Pr[X = j] = 1$, tedy že X má hodnoty $1, 2, \ldots, m$). Víme $\mathbb{E}[X] = \sum_{j=1}^{m} j \Pr[X = j] = S$. Jaká je pravděpodobnost $\Pr[X \geq 5S]$?**

Řešení: Obdobně

$$S = \mathbb{E}[X]$$

$$
\begin{aligned}
&= \sum_{j=1}^{m} j \Pr[X = j] && \text{(protože } \operatorname{Im}(X) = \{1, 2, \ldots, m\}\text{)} \\
&\geq \sum_{j=5S}^{m} j \Pr[X = j] \\
&\geq \sum_{j=5S}^{m} 5S \Pr[X = j] \\
&= 5S \Pr[X \geq 5S]
\end{aligned}
$$

tedy

$$\Pr[X \geq 5S] \leq 1/5$$

- **Gratuluji, vymysleli jste Markovovu nerovnost.**
- **Často bývá mnohem jednodušší použít Markovovu nerovnost než přímo počítat pravděpodobnost. Občas můžeme dostat i silnější odhady pomocí Čebyševovy nebo Černovovy nerovnosti. Na to budeme potřebovat rozptyl a další znalosti o náhodných proměnných.**

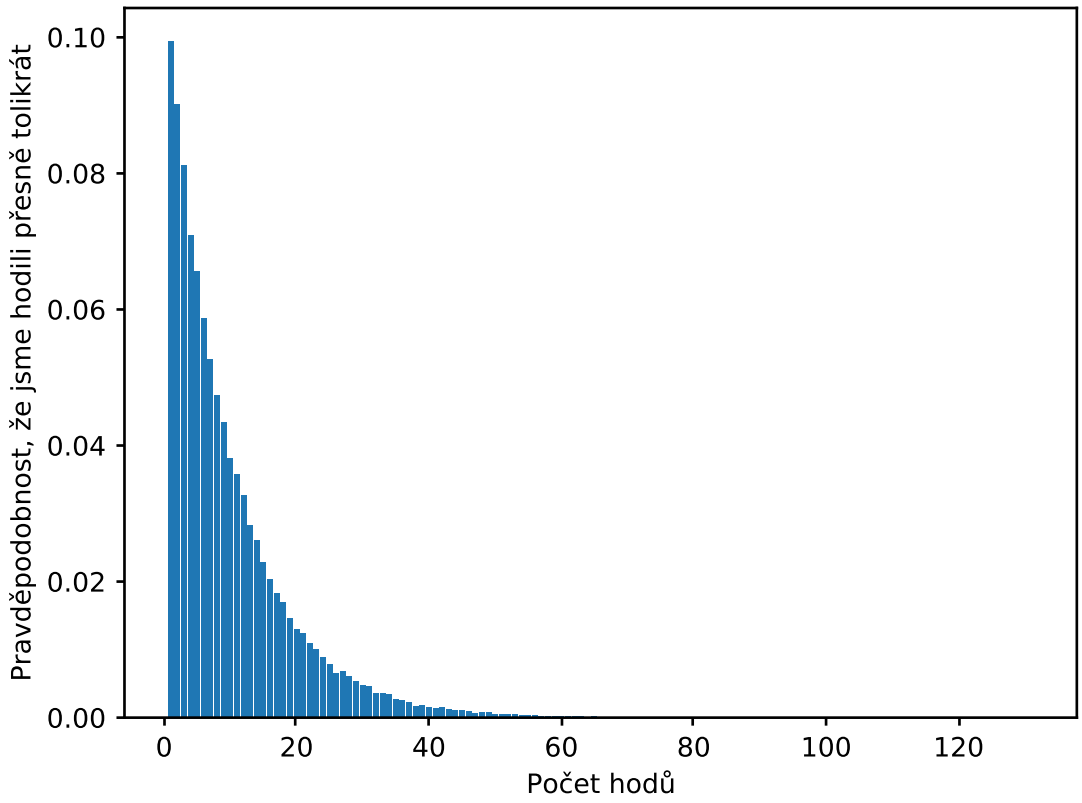
3.4 Cvičení

1. **Házím míčem na koš. V každém pokusu mám pravděpodobnost p že se trefím (jednotlivé hody jsou nezávislé). Skončím po prvním zásahu. Označme X celkový počet hodů.**

   (a) Jaké je pravděpodobnostní rozdělení X (tj. distribuce)? Jinak řečeno určete pravděpodobnostní funkci p_X (tj. pro každé x určete $p_X(x) = \Pr[X = x]$).

   Řešení: Je to geometrická distribuce, tedy $n-1$ hodů musí jít vedle a poslední se musí podařit.

   $$p_X(n) = (1-p)^{n-1}p$$



Obrázek 3.1: Histogram s jakou pravděpodobností jsme udělali právě tolik hodů.

   (b) Jaká je $\Pr[X \geq 10 \mid X \geq 5]$?

   Řešení: Můžeme si rozepsat na disjunktní jevy:

   $$\begin{aligned}\Pr[X \geq 10 \mid X \geq 5] &= \frac{\Pr[X \geq 10 \text{ a zároveň } X \geq 5]}{\Pr[X \geq 5]} \\ &= \frac{\Pr[X \geq 10]}{\Pr[X \geq 5]} \\ &= \frac{\sum_{j=10}^{\infty} p(1-p)^j}{\sum_{j=5}^{\infty} p(1-p)^j}\end{aligned}$$

$$= \frac{(1-p)^{10}/p}{(1-p)^5/p}$$
$$= (1-p)^5$$

(c) Jaká je $\mathbb{E}[X]$?

Řešení: Na přednášce bylo $\mathbb{E}[X] = 1/p$.

(d) Jaká je $\mathbb{E}[X \mid X$ je sudé$]$?

Řešení:

$$\mathbb{E}[X \mid X \text{ je sudé}] = \sum_{x \in \operatorname{Im}(X)} x \Pr[X = x \mid X \text{ je sudé}]$$
$$= \sum_{j=1}^{\infty} (2j) \Pr[X = 2j]$$

(e) Simulujte.

Řešení:

```
import matplotlib.pyplot as plt
from collections import Counter
from random import random


def geometric(pr: float = 0.5) -> int:
    """pr is success probability, return the number of tosses until
    the first success."""
    assert pr > 0
    sample = 1
    fail_pr = 1 - pr
    while random() < fail_pr:
        sample += 1
    return sample


N = 100000

# a)
cnt = Counter(geometric(0.1) for _ in range(N))
distribution = {}
for c in cnt:
    distribution[c] = cnt[c] / N

plt.bar(distribution.keys(), distribution.values())
plt.xlabel("Počet hodů")
plt.ylabel("Pravděpodobnost, že jsme hodili přesně tolikrát")
# plt.show()
plt.savefig('lemma_o_dzbanu.pdf')


# b)
```

```
p = 0.1
geq5 = 0
geq10_when_geq5 = 0
for _ in range(N):
    res = geometric(p)
    if res >= 5:
        geq5 += 1
        if res >= 10:
            geq10_when_geq5 += 1
print(f'Pr[X>=10 | X>=5] = {geq10_when_geq5/geq5} (={(1-p)**5})')


# c)
print(f'E[X] = {sum(geometric(p) for _ in range(N)) / N} (={1/p})')

# Možný výstup:
# Pr[X>=10 | X>=5] = 0.5907960009158209 (=0.5904900000000001)
# E[X] = 10.01754 (=10.0)
```

2. **V testu je 20 otázek s volbami a,b,c,d. Za správnou odpověď (vždy je jen jedna odpověď správná) je 1 bod, za špatnou $-1/4$ bodu, za nevyplněnou otázku nula. Každá otázka je s pravděpodobností p jednou z těch, co se Kvído naučil a tedy zná správnou odpověď. Pokud správnou odpověď nezná, ví o tom, a může se rozhodnout, zda tipovat.**

   (a) Jaká je střední hodnota počtu bodů, které Kvído získá, pokud bude odpovídat jenom otázky, u kterých zná odpověď?

   Řešení: Odpověď na otázku j zná s pravděpodobností p. Označme tedy náhodnou veličinu X_j počet bodů za j-tou otázku. Z linearity střední hodnoty:

$$\begin{aligned}\mathbb{E}[\text{počet bodů, netipuje}] &= \sum_{j=1}^{20} \mathbb{E}[X_j] \\ &= \sum_{j=1}^{20} p \\ &= 20p\end{aligned}$$

   (b) A co když bude tipovat, když nezná správnou odpověď?

   Řešení: S pravděpodobností p prostě zná odpověď. Pokud nezná a tipne, pak s pravděpodobností $1/4$ dostane bod, s pravděpodobností $3/4$ ztratí $1/4$ bodu. Pravděpodobnost, že dostane jeden bod tedy je

$$\Pr[Y_j = 1] = p + (1-p)/4 = (1+3p)/4$$

   a pravděpodobnost, že ztratí čtvrtbod

$$\Pr[Y_j = -1/4] = (1-p)3/4 = (3-3p)/4$$

   Označíme Y_j počet bodů za j-tou otázku, když tipuje a použijeme linearitu střední hodnoty. Tedy můžeme psát

$$\begin{aligned}\mathbb{E}[\text{počet bodů, tipuje}] &= \sum_{j=1}^{20} 1(1+3p)/4 + (-1/4)(3-3p)/4 \\ &= 20\,(1(1+3p)/4 + (-1/4)(3-3p)/4) \\ &= 20\left(\left(\frac{1}{4} - \frac{3}{16}\right) + p\left(\frac{3}{4} + \frac{3}{16}\right)\right)\end{aligned}$$

   (c) Jak by se musela změnit penalizace za chybnou odpověď, aby byly odpovědi v částech a, b stejné?

   Řešení: Potřebovali bychom, aby $\mathbb{E}[X_j] = \mathbb{E}[Y_j]$ (pro každé p). Jinak řečeno střední hodnota tipu je nulová (m je penalizace, tedy bodová ztráta):

$$\begin{aligned}0 &= 1/4 - m3/4 \\ m &= 1/3\end{aligned}$$

   (d) Simulujte.

   Řešení:

```python
from random import random


def bernoulli(pr: float = 0.5) -> bool:
    return random() < pr


def one_question(p: float, guess: bool, m: float = 1/4) -> float:
    """ Knows the answer with probability p. Return points."""
    if random() < p:
        # Knows the answer
        return 1
    else:
        if not guess:
            return 0
        if random() < 0.25:
            # Guessed the correct answer
            return 1
        else:
            return -m


def test(p: float, guess: bool, m: float = 1/4) -> float:
    return sum(one_question(p=p, guess=guess, m=m) for _ in range(20))


def student(p):
    N = 100000
    print(f'p = {p}')
    # a) netipuje
    netipuje_sim = sum(test(p, False) for _ in range(N)) / N
    print(f'E[bodů netipuje] = {netipuje_sim} (={20 * p})')
    # b) tipuje
    with_guess = 20 * ((1/4 - 3/16) + p * (3/4 + 3/16))
    tipuje_sim = sum(test(p, True) for _ in range(N)) / N
    print(f'E[bodů tipuje] = {tipuje_sim} (={with_guess})')
    # c) spravedlivý test
    tipuje_spravedlivy_sim = sum(test(p, True, m=1/3) for _ in range(N)) / N
    print(f'E[bodů ve "spravedlivém" testu] = {tipuje_spravedlivy_sim} (={20 * p})')


for p in [0.0, 0.2, 0.8, 1.0]:
    student(p)

# Možný výstup:
# p = 0.0
# E[bodů netipuje] = 0.0 (=0.0)
# E[bodů tipuje] = 1.254 (=1.25)
# E[bodů ve "spravedlivém" testu] = -0.00941333333333343 (=0.0)
# p = 0.2
# E[bodů netipuje] = 3.9956 (=4.0)
# E[bodů tipuje] = 5.0054 (=5.0)
# E[bodů ve "spravedlivém" testu] = 4.010946666666663 (=4.0)
# p = 0.8
```

```
# E[bodů netipuje] = 16.0001 (=16.0)
# E[bodů tipuje] = 16.249 (=16.25)
# E[bodů ve "spravedlivém" testu] = 16.00237333333326 (=16.0)
# p = 1.0
# E[bodů netipuje] = 20.0 (=20.0)
# E[bodů tipuje] = 20.0 (=20.0)
# E[bodů ve "spravedlivém" testu] = 20.0 (=20.0)
```

3. **Ze standardního balíčku s 52 kartami vytáhneme dvě karty. Označíme X počet vytažených es, Y počet králů. Určete sdruženou pravděpodobnostní funkci $p_{X,Y}$ a také marginální pstní funkce p_X, p_Y.**

Řešení: Z 52 karet jsou tam 4 esové karty a 4 králové (ani jeden král není eso, ani naopak). Náhodné veličiny X, Y mohou nabývat hodnot $0, 1, 2$. Ale tím, že vytáhneme jen dvě karty, tak musí platit $X + Y \leq 2$.

$$
\begin{aligned}
p_{X,Y}(0,0) &= \frac{44}{52} \cdot \frac{43}{51} \\
p_{X,Y}(1,0) &= \frac{4}{52} \cdot \frac{44}{51} + \frac{44}{52} \cdot \frac{4}{51} \text{ vytáhnu eso jako první nebo druhé} \\
p_{X,Y}(0,1) &= 2\frac{4}{52} \cdot \frac{44}{51} \\
p_{X,Y}(2,0) &= \frac{4}{52} \cdot \frac{3}{51} \\
p_{X,Y}(1,1) &= 2\frac{4}{52} \cdot \frac{4}{51} \text{ záleží na pořadí – napřed eso, pak král nebo naopak} \\
p_{X,Y}(0,2) &= \frac{4}{52} \cdot \frac{3}{51} \\
p_{X,Y}(x,y) &= 0 \text{ jinak}
\end{aligned}
$$

Z definice

$$
\begin{aligned}
p_X(x) &= \sum_{y \in \operatorname{Im}(Y)} p_{X,Y}(x,y) \\
&= p_{X,Y}(x,0) + p_{X,Y}(x,1) + p_{X,Y}(x,2)
\end{aligned}
$$

Mělo by nám vyjít:

$$
\begin{aligned}
p_X(0) &= \frac{48}{52} \cdot \frac{47}{51} \\
&= \frac{2256}{2652} \\
&= p_{X,Y}(0,0) + p_{X,Y}(0,1) + p_{X,Y}(0,2) \\
&= \frac{44}{52} \cdot \frac{43}{51} + 2\frac{4}{52} \cdot \frac{44}{51} + \frac{4}{52} \cdot \frac{3}{51} \\
&= \frac{44 \cdot 43 + 8 \cdot 44 + 12}{52 \cdot 51} \\
&= \frac{2256}{2652}
\end{aligned}
$$

$$
\begin{aligned}
p_X(1) &= \frac{4}{52} \cdot \frac{48}{51} + \frac{48}{52} \cdot \frac{4}{51} \text{ vytáhnu eso jako první nebo druhé} \\
&= \frac{384}{2652} \\
&= p_{X,Y}(1,0) + p_{X,Y}(1,1) + p_{X,Y}(1,2) \\
&= p_{X,Y}(1,0) + p_{X,Y}(1,1) + 0 \\
&= 2\frac{4}{52} \cdot \frac{44}{51} + 2\frac{4}{52} \cdot \frac{4}{51} \\
&= \frac{384}{2652}
\end{aligned}
$$

$$\begin{aligned} p_X(2) &= \frac{4}{52} \cdot \frac{3}{51} \\ &= p_{X,Y}(2,0) + p_{X,Y}(2,1) + p_{X,Y}(2,2) \\ &= p_{X,Y}(2,0) + 0 + 0 \\ &= \frac{4}{52} \cdot \frac{3}{51} \end{aligned}$$

4. **Chceme nasbírat všechny z n druhů kuponů. Můžeme si koupit jeden kupon, který má uniformně náhodný druh. Kolikrát musíme koupit kupon, než posbíráme všechny?**

   (a) Jaká je střední hodnota počtu koupených kuponů, než nasbíráme všechny?

   Řešení: Označme náhodnou veličinu t_i, která říká kolik musíme koupit kuponů, abychom získali i-tý druh když už máme $i-1$ druhů. Pravděpodobnost, že ten další kupon bude nového druhu je:

   $$\Pr[\text{dostaneme nový druh, když už máme } i-1 \text{ druhů}] = \frac{n-(i-1)}{n}$$

   Takže t_i má geometrické rozdělení (čekáme na první úspěch). Střední hodnota t_i je:

   $$\mathbb{E}[t_i] = \frac{n}{n-(i-1)}$$

   Z linearity střední hodnoty:

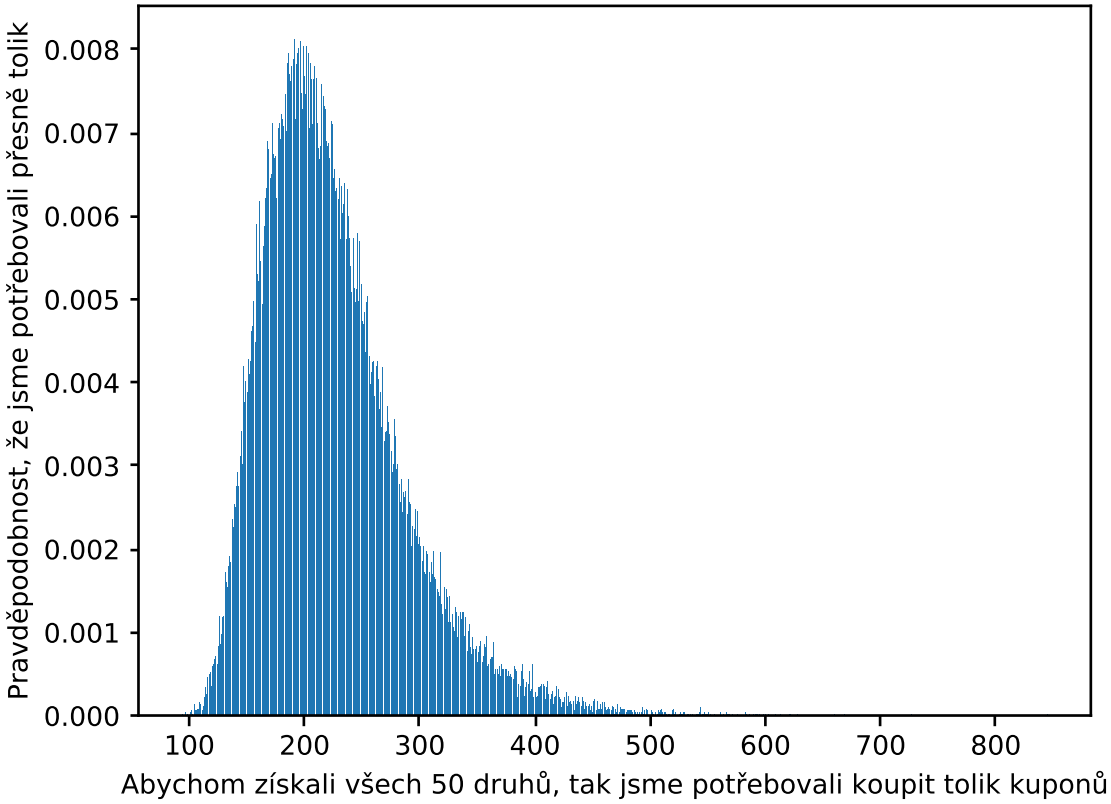
   $$\begin{aligned}
   \mathbb{E}[\text{sbírání}] &= \mathbb{E}[t_1 + t_2 + \ldots + t_n] \\
   &= \mathbb{E}[t_1] + \mathbb{E}[t_2] + \ldots + \mathbb{E}[t_n] \\
   &= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \ldots + \frac{n}{n-(n-1)} \\
   &= nH_n \\
   &= n\log(n) + n \cdot 0.577\ldots + 1/2 + \mathcal{O}(1/n) \qquad \text{(Wikipedia)}
   \end{aligned}$$

   (b) Simulujte.

   Řešení:

```python
import matplotlib.pyplot as plt
from collections import Counter
from random import randint

def catch_them_all(n: int = 50) -> int:
    coupons = [False] * n
    coupons_collected = 0
    coupons_bought = 0
    while coupons_collected < len(coupons):
        new_coupon = randint(0, len(coupons) - 1)
        coupons_bought += 1
        if not coupons[new_coupon]:
            coupons[new_coupon] = True
            coupons_collected += 1
    return coupons_bought

N = 50000
cnt = Counter(catch_them_all(50) for _ in range(N))
distribution = {}
for c in cnt:
    distribution[c] = cnt[c] / N

plt.bar(distribution.keys(), distribution.values())
plt.xlabel("Abychom získali všech 50 druhů, tak jsme potřebovali koupit tolik kuponů")
```

```
plt.ylabel("Pravděpodobnost, že jsme potřebovali přesně tolik")
# plt.show()
plt.savefig('coupon_collector.pdf')
```

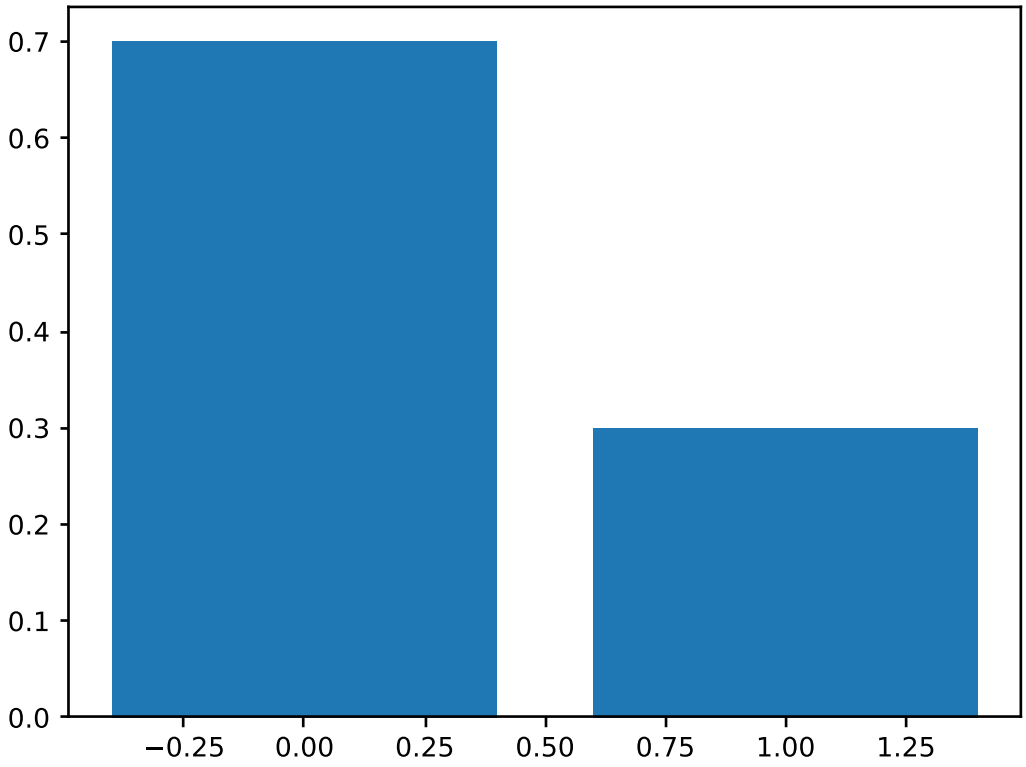


Obrázek 3.2: Histogram s jakou pravděpodobností jsme potřebovali právě tolik kuponů.

5. **Připomeňte si, co je náhodná veličina, jaké máme typy náhodných veličin a jaké jsou jejich distribuce. A hlavně co vyjadřují.**

   (a) Bernoulli

   Řešení: $X \sim Bern(p)$ pokud $\Pr[X = 1] = p$ a $\Pr[X = 0] = 1 - p$ (tedy hod mincí se stranami 0/1).



Obrázek 3.3: Bernoulli

   (b) binomické

   Řešení: Součet několika Bernoulliho.

   (c) hypergeometrické

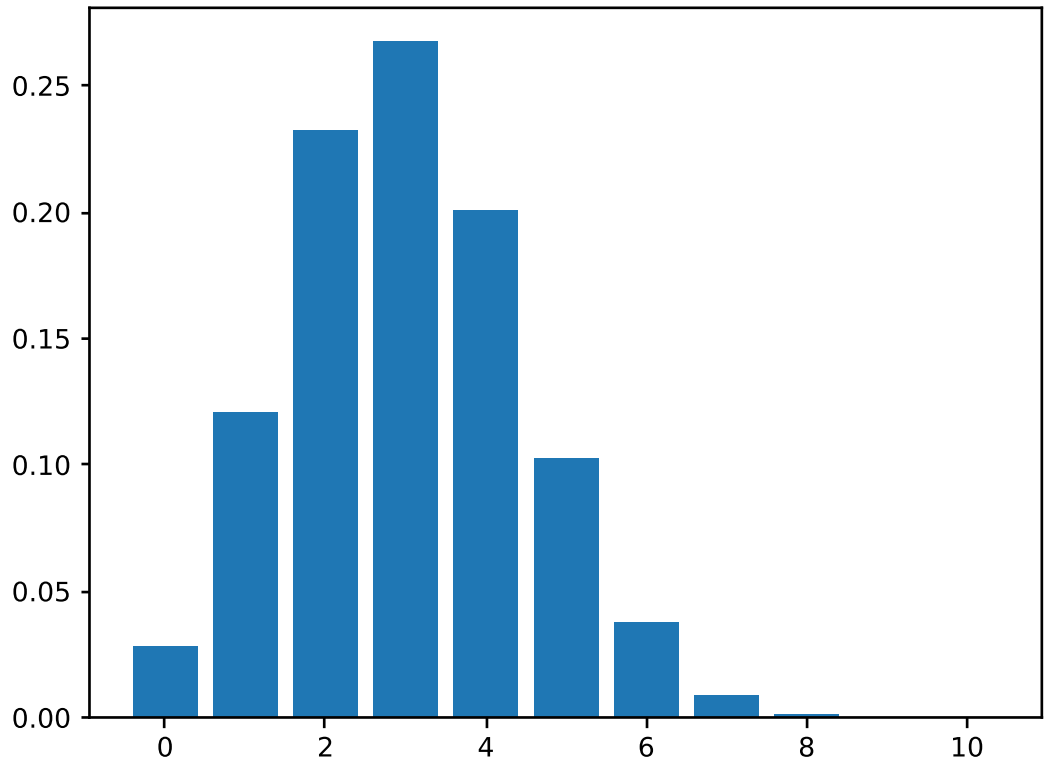
   Řešení: Když vybereme n věcí z N věcí bez vracení a k je "správných", tak tohle je počet správných.

   (d) geometrické

   Řešení: Čekání na první úspěch Bernoulliho hodu.

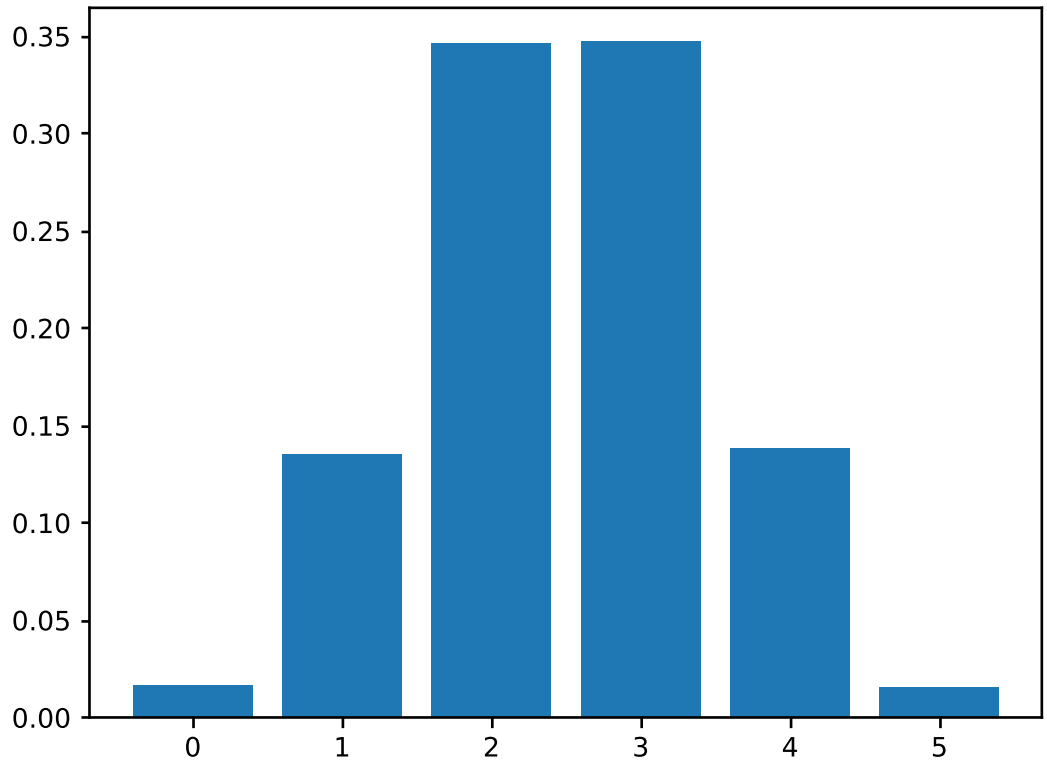
   (e) Poissonovo

   Řešení: Kolik jevů čekáme během času, když jsou všechny nezávislé a střední hodnota je λ (emaily chodí nezávisle, průměrně 10 emailů za hodinu, kolik jich přijde?).



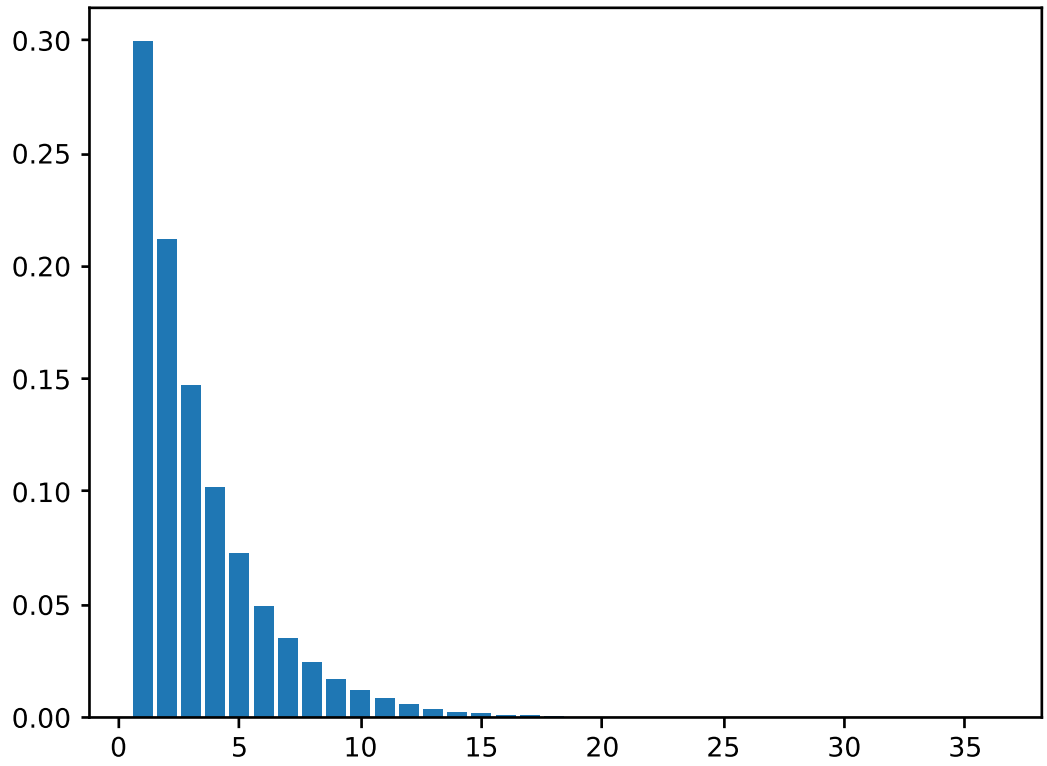
Obrázek 3.4: binomické

```
import matplotlib.pyplot as plt
from collections import Counter
from random import randint
from random import random
from random import sample
from numpy import random as npr

p = 0.3
n = 10
k = 5
S = 20
l = 10


def bernoulli(pr: float = p) -> bool:
    return int(random() < pr)


def geometric(pr: float = p) -> int:
    """pr is success probability, return the number of tosses until
    the first success."""
```



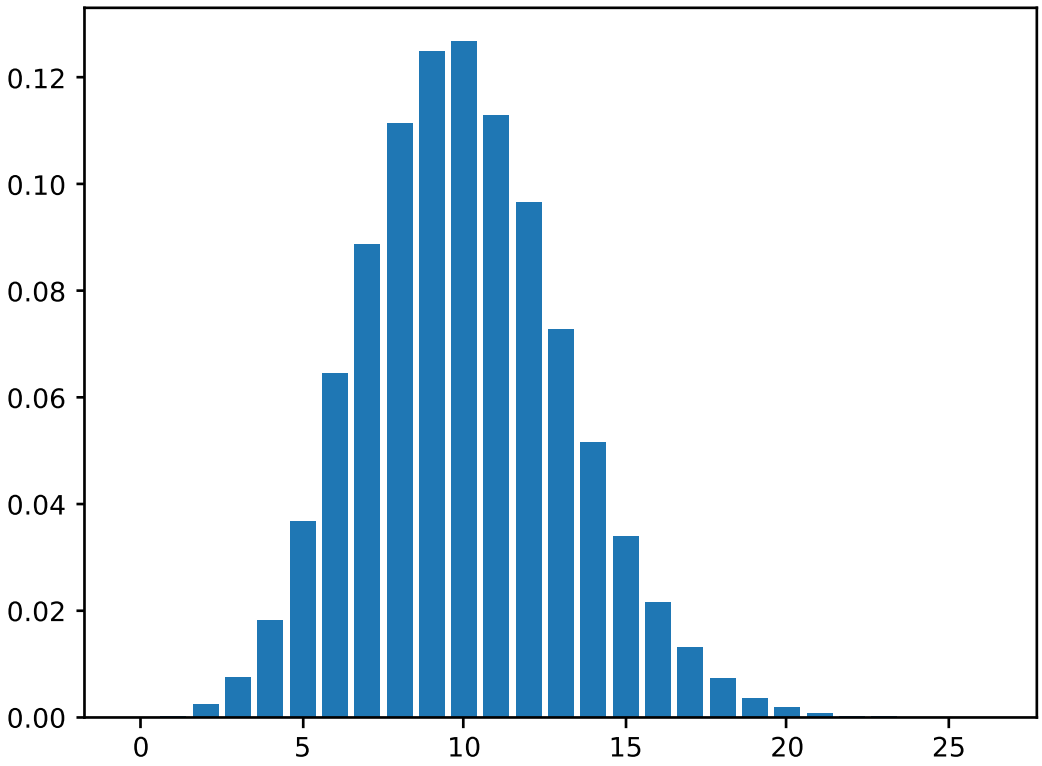
Obrázek 3.5: hypergeometrické

```
    assert pr > 0
    sample = 1
    fail_pr = 1 - pr
    while random() < fail_pr:
        sample += 1
    return sample


def binomicke(n=n, pr=p):
    return sum(bernoulli(pr) for _ in range(n))


def hypergeometric(n=n, N=S, k=k):
    return sum(sample([1]*k + [0]*(N-k), k=n))


def poisson(l=l):
    return npr.poisson(lam=l, size=1)[0]


N = 100000
```



Obrázek 3.6: geometrické

```
def expected_value(X):
    """Vraci stredni hodnotu."""
    return sum(X() for _ in range(N)) / N


def variance(X):
    """Vraci rozptyl."""
    EX = expected_value(X)
    return sum((X() - EX)**2 for _ in range(N)) / N
    # return (sum(X()**2 for _ in range(N)) / N) - EX**2


def histogram(X, strX, fig):
    cnt = Counter(X() for _ in range(N))
    distribution = {}
    for c in cnt:
        distribution[c] = cnt[c] / N

    plt.figure(fig)
```



Obrázek 3.7: Poissonovo

```
    plt.bar(distribution.keys(), distribution.values())
    # plt.show()
    # plt.xlabel("")
    # plt.ylabel("")
    plt.savefig(f'{strX}.pdf')


print('Bernoulli:')
print(f'E[X] = {expected_value(bernoulli)} (= {p})')
print(f'var[X] = {variance(bernoulli)} (= {p*(1-p)})')
histogram(bernoulli, "bernoulli", 0)
print('')


print('binomické')
print(f'E[X] = {expected_value(binomicke)} (= {n*p})')
print(f'var[X] = {variance(binomicke)} (= {n*p*(1-p)})')
histogram(binomicke, "binomicke", 1)
print('')
```

```
print('hypergeometrické')
print(f'E[X] = {expected_value(hypergeometric)} (= {n*k/S})')
print(f'var[X] = {variance(hypergeometric)} (= {n*(k/S)*(1-(k/S))*(S-n)/(S-1)})')
histogram(hypergeometric, "hypergeometric", 2)
print('')


print('geometrické')
print(f'E[X] = {expected_value(geometric)} (= {1/p})')
print(f'var[X] = {variance(geometric)} (= {(1-p)/p**2})')
histogram(geometric, "geometric", 3)
print('')


print('Poissonovo')
print(f'E[X] = {expected_value(poisson)} (= {l})')
print(f'var[X] = {variance(poisson)} (= {l})')
histogram(poisson, "poisson", 4)

# Možný výstup:
# Bernoulli:
# E[X] = 0.30003 (= 0.3)
# var[X] = 0.21018846799996171 (= 0.21)

# binomické
# E[X] = 3.00221 (= 3.0)
# var[X] = 2.111969109999777 (= 2.0999999999999996)

# hypergeometrické
# E[X] = 2.49898 (= 2.5)
# var[X] = 0.9866719879994746 (= 0.9868421052631579)

# geometrické
# E[X] = 3.33374 (= 3.3333333333333335)
# var[X] = 7.851098870500136 (= 7.777777777777778)

# Poissonovo
# E[X] = 10.00683 (= 10)
# var[X] = 9.947825008000336 (= 10)
```

3.5 Cvičení

1. **Hodíme třikrát mincí. Označíme X počet rubů v prvních dvou hodech a Y počet líců v posledních dvou hodech.**

   (a) Určete pravděpodobnostní prostor.

   Řešení:

   - Množina elementárních jevů:

   $$\Omega = \{LLL, LLR, LRL, LRR, RLL, RLR, RRL, RRR\}$$

   - Prostor jevů:

   $$\mathcal{F} = \mathcal{P}(\Omega)$$

   - Pravděpodobnost

   $$\Pr[\{\omega\}] = 1/8 \qquad \text{(pro každé } \omega \in Omega\text{)}$$

   Zbytek jednoznačně určen axiomy.

   (b) Určete předpis našich náhodných veličin.

   Řešení:

   $$X\colon \Omega \to \mathbb{R}$$
   $$Y\colon \Omega \to \mathbb{R}$$

   kde

   $$\begin{aligned} X(LLL) = X(LLR) &= 0 \\ X(LRL) = X(LRR) &= 1 \\ X(RLL) = X(RLR) &= 1 \\ X(RRL) = X(RRR) &= 2 \\ Y(LRR) = Y(RRR) &= 0 \\ Y(LRL) = Y(RRL) &= 1 \\ Y(LLR) = Y(RLR) &= 1 \\ Y(LLL) = Y(RLL) &= 2 \end{aligned}$$

   Po náhodné veličině X požadujeme, aby

   $$\forall x \in \mathbb{R}\colon\ \{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$$

   pro diskrétní jsme chtěli

   $$\forall x \in \mathbb{R}\colon\ \{\omega \in \Omega \mid X(\omega) = x\} \in \mathcal{F}$$

   ale vzhledem k tomu, že $\mathrm{Im}(X)$ je konečná množina nebo spočetná množina (dle definice diskrétní náhodné veličiny), tak to je jedno, protože $\mathcal{F}$ je σ-algebra a ta je uzavřená na spočetná sjednocení.

   Naše náhodná veličina je diskrétní a tato podmínka je celkem jednoduchá, protože naše $\mathcal{F} = \mathcal{P}(\Omega)$ (což je možné proto, že Ω je konečná a bylo by to možné i pro spočetnou množinu Ω).

(c) Určete sdruženou pravděpodobnostní funkci $p_{X,Y}$ a také marginální pravděpodobnostní funkce p_X, p_Y.

Řešení:

$$
\begin{aligned}
p_{X,Y}(0,0) &= 0 && \text{(druhý hod je buď rub nebo líc)}\\
p_{X,Y}(0,1) &= 1/8 && \text{(jev } \{LLR\})\\
p_{X,Y}(0,2) &= 1/8 && \text{(jev } \{LLL\})\\
p_{X,Y}(1,0) &= 1/8 && \text{(jev } \{LRR\})\\
p_{X,Y}(1,1) &= 1/4 && \text{(jev } \{RLR, LRL\})\\
p_{X,Y}(1,2) &= 1/8 && \text{(jev } \{RLL\})\\
p_{X,Y}(2,0) &= 1/8 && \text{(jev } \{RRR\})\\
p_{X,Y}(2,1) &= 1/8 && \text{(jev } \{RRL\})\\
p_{X,Y}(2,2) &= 0 && \text{(druhý hod je buď rub nebo líc)}
\end{aligned}
$$

Častější a přehlednější může být reprezentace tabulkou:

$p_{X,Y}$	0	1	2
0	0	1/8	1/8
1	1/8	1/4	1/8
2	1/8	1/8	0

Tabulka 3.1: Sdružená pravděpodobnostní funkce (řádky odpovídají možným hodnotám X, sloupce možným hodnotám Y.

Marginální pravděpodobnostní funkce p_X je součet řádků:

$$
\begin{aligned}
p_X(x) &= \sum_{y \in \operatorname{Im}(Y)} \Pr[X = x \wedge Y = y] = \sum_{y \in \operatorname{Im}(Y)} p_{X,Y}(x,y)\\
p_X(0) &= 1/4\\
p_X(1) &= 1/2\\
p_X(2) &= 1/4
\end{aligned}
$$

Všimněte si, že to přesně odpovídá $p_X(n) = \Pr[X = n]$ pro každé $n \in \mathbb{N}$.

Marginální pravděpodobnostní funkce p_Y je součet sloupců:

$$
\begin{aligned}
p_Y(0) &= 1/4\\
p_Y(1) &= 1/2\\
p_Y(2) &= 1/4
\end{aligned}
$$

(d) Určete distribuční funkce $F_X, F_Y, F_{X,Y}$ (cumulative distribution function CDF).

Řešení: Z definice:

$$
\begin{aligned}
F_X &: \mathbb{R} \to [0,1]\\
F_X(x) &= \Pr[X \leq x] = \Pr[\{\omega \in \Omega \mid X(\omega) \leq x\}]
\end{aligned}
$$

tedy máme:

$$
F_X(x) = 0 \qquad \text{(pro } x \in (-\infty, 0))
$$

$$
\begin{aligned}
F_X(x) &= 1/4 && (\text{pro } x \in [0,1)) \\
F_X(x) &= 3/4 && (\text{pro } x \in [1,2)) \\
F_X(x) &= 1 && (\text{pro } x \in [2,\infty))
\end{aligned}
$$

$$
\begin{aligned}
F_Y(y) &= 0 && (\text{pro } y \in (-\infty,0)) \\
F_Y(y) &= 1/4 && (\text{pro } y \in [0,1)) \\
F_Y(y) &= 3/4 && (\text{pro } y \in [1,2)) \\
F_Y(y) &= 1 && (\text{pro } y \in [2,\infty))
\end{aligned}
$$

Sdružená distribuční funkce $F_{X,Y}$ je pak:

$$
\begin{aligned}
&F_{X,Y}(x,y)\colon \mathbb{R}^2 \to [0,1] \\
&F_{X,Y}(x,y) = \Pr[X \leq x \wedge Y \leq y]
\end{aligned}
$$

tedy například: $F_{X,Y}(0.1, 1.33) = 1/8$.

(e) Jsou X a Y nezávislé?

Řešení: Nezávislost jsme měli definovanou jen pro diskrétní náhodné veličiny a to konkrétně podmínkou:

$$\forall x, y \in \mathbb{R}\colon \Pr[X = x \wedge Y = y] = \Pr[X = x] \Pr[Y = y]$$

Tato podmínka není splněná například pro $x = 0, y = 0$:

$$\Pr[X = 0 \wedge Y = 0] = 0 \neq 1/64 = \Pr[X = 0] \Pr[Y = 0]$$

Obecně je nezávislost definovaná pomocí distribuční funkce (cumulative distribution function CDF) F_X definované jako:

$$
\begin{aligned}
&F_X\colon \mathbb{R} \to [0,1] \\
&F_X(x) = \Pr[X \leq x] = \Pr[\{\omega \in \Omega \mid X(\omega) \leq x\}]
\end{aligned}
$$

Pak řekneme, že X, Y jsou nezávislé, pokud platí:

$$F_{X,Y}(x,y) = F_X(x) F_Y(y)$$

nebo když máme hustotu, pak:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

Pro diskrétní náhodné veličiny jsou ty dvě definice ekvivalentní (jednoduché cvičení na sumy).

(f) Určete $\Pr[X < Y]$.

Řešení:

$$
\begin{aligned}
\Pr[X < Y] &= p_{X,Y}(0,1) + p_{X,Y}(0,2) + p_{X,Y}(1,2) \\
&= 1/8 + 1/8 + 1/8 \\
&= 3/8
\end{aligned}
$$

(g) Určete podmíněnou pravděpodobnostní funkce $p_{X|Y}$.

Řešení: Dle definice:

$$
\begin{aligned}
p_{X|Y}(x|y) &= \Pr[X = x \mid Y = y] \\
p_{X|Y}(0|0) &= 0 \\
p_{X|Y}(0|1) &= 1/4 \\
p_{X|Y}(0|2) &= 1/2 \\
p_{X|Y}(1|0) &= 1/2 \\
p_{X|Y}(1|1) &= 1/2 \\
p_{X|Y}(1|2) &= 1/2 \\
p_{X|Y}(2|0) &= 1/2 \\
p_{X|Y}(2|1) &= 1/4 \\
p_{X|Y}(2|2) &= 0
\end{aligned}
$$

Opět je elegantnější tabulka:

$p_{X\|Y}$	0	1	2
0	0	1/4	1/2
1	1/2	1/2	1/2
2	1/2	1/4	0

Tabulka 3.2: Všimněte si, že je to ta samá tabulka jako Tabulka 3.1 akorát sloupce jsou škálované tak, aby se sečetli na jedničku.

(h) Simulujte.

Řešení:

```
from random import choices

def sample():
    """True je rub, False je líc"""
    return choices([True, False], k=3)

def X(omega):
    """X: Omega -> R"""
    return int(omega[0]) + int(omega[1])

def Y(omega):
    """Y: Omega -> R"""
    return int(not omega[1]) + int(not omega[2])

N = 1_000_000
p_XY = [[0,0,0], [0,0,0], [0,0,0]]
for _ in range(N):
    omega = sample()
    p_XY[X(omega)][Y(omega)] += 1

for i in range(3):
    for j in range(3):
        p_XY[i][j] /= N
```

```
print('Sdružená pravděpodobnostní funkce')
print(f'   0      1      2')
print(f'0  {p_XY[0][0]:.3f}  {p_XY[0][1]:.3f}  {p_XY[0][2]:.3f}')
print(f'1  {p_XY[1][0]:.3f}  {p_XY[1][1]:.3f}  {p_XY[1][2]:.3f}')
print(f'2  {p_XY[2][0]:.3f}  {p_XY[2][1]:.3f}  {p_XY[2][2]:.3f}')

print('Pravděpodobnost X<Y')
pr_X_leq_Y = 0
for _ in range(N):
    omega = sample()
    if X(omega) < Y(omega):
        pr_X_leq_Y += 1

print('')
print(f'Pr[X < Y] = {pr_X_leq_Y} (={3/8})')

p_XY = [[0,0,0], [0,0,0], [0,0,0]]
for _ in range(N):
    omega = sample()
    p_XY[X(omega)][Y(omega)] += 1

p_Y = [0,0,0]
p_Y[0] = p_XY[0][0] + p_XY[1][0] + p_XY[2][0]
p_Y[1] = p_XY[0][1] + p_XY[1][1] + p_XY[2][1]
p_Y[2] = p_XY[0][2] + p_XY[1][2] + p_XY[2][2]
for i in range(3):
    for j in range(3):
        p_XY[j][i] /= p_Y[i]

print('')
print('Podmíněná pravděpodobnostní funkce')
print(f'   0      1      2')
print(f'0  {p_XY[0][0]:.3f}  {p_XY[0][1]:.3f}  {p_XY[0][2]:.3f}')
print(f'1  {p_XY[1][0]:.3f}  {p_XY[1][1]:.3f}  {p_XY[1][2]:.3f}')
print(f'2  {p_XY[2][0]:.3f}  {p_XY[2][1]:.3f}  {p_XY[2][2]:.3f}')

# Možný výstup:
# Sdružená pravděpodobnostní funkce
   # 0      1      2
# 0  0.000  0.125  0.125
# 1  0.125  0.250  0.125
# 2  0.125  0.125  0.000
# Pravděpodobnost X<Y

# Pr[X < Y] = 375133 (=0.375)

# Podmíněná pravděpodobnostní funkce
   # 0      1      2
# 0  0.000  0.249  0.502
# 1  0.500  0.500  0.498
# 2  0.500  0.251  0.000
```

2. **Bonusový příklad: tady si zadefinujeme Lebesgueovu míru a integrál.**

   (a) Definujte Lebesgueovu míru na $\mathbb{R}$.

   Řešení:

   - Napřed definujeme délku intervalu $\ell([a,b]) = \ell((a,b)) = b - a$ pro libovolná dvě reálná čísla $a \leq b \in \mathbb{R}$.
   - Pak definujeme vnější míru pro každou $E \subseteq \mathbb{R}$:

     $$\lambda^*(E) = \inf \left\{ \sum_{k=1}^{\infty} \ell(I_k) \mid (I_k)_{k\in\mathbb{N}} \text{ je posloupnost otevřených intervalů, t.ž. } E \subseteq \cup_{k=1}^{\infty} I_k \right\}$$

   - Pak $E \in \mathcal{F}$ (což je Lebesgueova σ-algebra) právě když

     $$\forall A \subseteq \mathbb{R}\colon \lambda^*(A) = \lambda^*(A \cap E) + \lambda^*(A \cap (\mathbb{R} \setminus E))$$

   - Pro libovolné $E \in \mathcal{F}$ položíme Lebesgueovu míru

     $$\begin{aligned} \lambda\colon \mathcal{F} &\to \mathbb{R} \\ \lambda(E) &= \lambda^*(E) \qquad \text{(pro libovolné } E \in \mathcal{F}) \end{aligned}$$

   (b) Dokažte, že pro diskrétní náhodnou veličinu, takovou že $\operatorname{Im}(X) \subseteq \mathbb{N}$, platí:

   $$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \operatorname{Im}(X)} x \Pr[X = x] \\ &= \sum_{n\in\mathbb{N}} n \Pr[X = n] \\ &= \sum_{n\in\mathbb{N}} \Pr[X \geq n] \end{aligned}$$

   Řešení: Tohle je poměrně jednoduché, protože absolutně konvergentní řady můžeme přeuspořádat:

   $$\begin{aligned} \sum_{n\in\mathbb{N}} n \Pr[X = n] &= \sum_{n\in\mathbb{N}} \sum_{\ell=1}^{n} \Pr[X = n] \\ &= \sum_{n\in\mathbb{N}} \sum_{k\geq n} \Pr[X = k] \\ &= \sum_{n\in\mathbb{N}} \Pr[X \geq n] \end{aligned}$$

   (c) Definujte Lebesgueův integrál (pro danou míru $\Pr$, obecně to ani nemusí být pravděpodobnostní míra, ale obecná μ).

   Řešení: Nechť

   $$\begin{aligned} f\colon \mathbb{R} &\to \mathbb{R}^{\geq 0} \\ f(x) &\geq 0 \qquad \text{(pro každé } x \in \mathbb{R}) \end{aligned}$$

   $$\int f d\Pr = (R) \int_0^{\infty} \Pr\left(\{x \in \mathbb{R} \mid f(x) > t\}\right)\, dt$$

   Kde ten Riemannův integrál vždy existuje, protože $\Pr\left(\{x \in \mathbb{R} \mid f(x) > t\}\right)$ je neklesající funkce. Z toho taky vidíme, že $\{x \in \mathbb{R} \mid f(x) > t\}$ musí být měřitelná množina (pro náhodné veličiny je v definici $\{x \in \mathbb{R} \mid f(x) \leq t\}$, což je to samé, protože σ-algebry jsou uzavřené na doplněk).

3. **Pro následující náhodné veličiny určete:**

(a)

$$\begin{aligned} \Omega &= \{1,2,3\} \\ \mathcal{F} &= \mathcal{P}(\Omega) \\ \Pr[\{\omega\}] &= 1/3 \qquad \text{(pro libovolné } \omega \in \Omega\text{, zbytek určen jednoznačně)} \\ X &\colon \Omega \to \mathbb{R} \\ X(\omega) &= 2\omega + 1 \end{aligned}$$

- **Je to náhodná veličina na daném pravděpodobnostním prostoru?**

  Řešení: Ano, pro libovolné $x \in \mathbb{R}$ platí $\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$.

- **Je to diskrétní náhodná veličina?**

  Řešení: Ano, $\mathrm{Im}(X) = \{3,5,7\}$, což je konečná množina (tedy nejvýš spočetná).

- **Jaká je její střední hodnota?**

  Řešení:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \mathrm{Im}(X)} x \Pr[X = x] \\ &= 3\Pr[X=3] + 5\Pr[X=5] + 7\Pr[X=7] \\ &= 3\frac{1}{3} + 5\frac{1}{3} + 7\frac{1}{3} \\ &= 5 \end{aligned}$$

- **Jaká je její distribuční funkce?**

  Řešení:

$$F_X(x) = \Pr[X \leq x]$$

$$F_X(x) = \begin{cases} 0 & \text{pokud } x \in (-\infty, 3) \\ 1/3 & \text{pokud } x \in [3,5) \\ 2/3 & \text{pokud } x \in [5,7) \\ 1 & \text{pokud } x \in [7,\infty) \end{cases}$$

- **Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?**

  Řešení: Ne, protože F_X není spojitá.

- **Jaká je její kvantilová funkce?**

  Řešení:

$$\begin{aligned} Q_X &\colon [0,1] \to \mathbb{R} \\ Q_X(p) &= \inf\{x \in \mathbb{R} \mid p \leq F_X(x)\} \end{aligned}$$

$$Q_X(p) = \begin{cases} -\infty & \text{pokud } p = 0 \\ 3 & \text{pokud } p \in (0, 1/3] \\ 5 & \text{pokud } p \in (1/3, 2/3] \\ 7 & \text{pokud } p \in (2/3, 1] \end{cases}$$

Poznámka: určitě znáte percentil – měl jsem percentil 95% v testu znamená, že 95% ostatních mělo skóre menší nebo rovné mému. Obdobně medián je $Q_X(1/2)$. Dávejte pozor, jestli je percentil definovaný jako menší rovno nebo ostře menší (v literatuře se používá obojí)!

(b)

$$\begin{aligned}
\Omega &= \mathbb{N} = \{1, 2, 3, \ldots\} \\
\mathcal{F} &= \mathcal{P}(\Omega) \\
\Pr[\{n\}] &= 1/2^n && \text{(pro libovolné } n \in \mathbb{N}\text{, zbytek určen jednoznačně)} \\
X &\colon \Omega \to \mathbb{R} \\
X(\omega) &= 2\omega + 1
\end{aligned}$$

- **Je to náhodná veličina na daném pravděpodobnostním prostoru?**

  Řešení: Ano, pro libovolné $x \in \mathbb{R}$ platí $\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$.

- **Je to diskrétní náhodná veličina?**

  Řešení: Ano, $\operatorname{Im}(X) = \{3, 5, 7, 9, 11, \ldots\}$ (množina všech lichých čísel která jsou aspoň tři), což je spočetně velká množina.

- **Jaká je její střední hodnota?**

  Řešení:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \operatorname{Im}(X)} x \Pr[X = x] \\
&= \sum_{n=1}^{\infty} (2n+1) \Pr[X = (2n+1)] \\
&= \sum_{n=1}^{\infty} (2n+1) 2^{-n} \\
&= 5
\end{aligned}$$

- **Jaká je její distribuční funkce?**

  Řešení:

$$\begin{aligned}
F_X(x) &= \Pr[X \leq x] \\
F_X(x) &= 0 && \text{(pokud } x < 1\text{)} \\
F_X(x) &= \sum_{n=1}^{\lfloor (x-1)/2 \rfloor} (2n+1) 2^{-n} \\
&= 2^{-\lfloor (x-1)/2 \rfloor} \left(-2 \left(\lfloor (x-1)/2 \rfloor \right) + 5 \cdot 2^{\lfloor (x-1)/2 \rfloor} - 5 \right) && \text{(jinak)}
\end{aligned}$$

- **Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?**

  Řešení: Ne, protože F_X není spojitá.

- **Jaká je její kvantilová funkce?**

  Řešení: Pro číslo $p \in (0, 1)$ definujme $j(p)$ tak, že pokud $p = 0.p_1p_2p_3\ldots$ kde $p_i \in \{0, 1\}$ je binární zápis čísla p (tedy $p = \sum_{i=1}^{\infty} p_i 2^{-i}$), nechť $k(p) =$

$\min\{i \in \mathbb{N} \mid p_i = 0\}$, pak definujeme $j(p) = \sum_{i=1}^{k(p)} 2^{-i}$. Tedy například $j(0.11101101) = 0.1111 = 15/16$.

$$\begin{aligned} Q_X &\colon [0,1] \to \mathbb{R} \\ Q_X(p) &= \inf\{x \in \mathbb{R} \mid p \leq F_X(x)\} \\ Q_X(p) &= \begin{cases} -\infty & \text{pokud } p = 0 \\ \infty & \text{pokud } p = 1 \\ 2(k(p)-1)+1 & \text{pokud } p \in (0,1) \text{ a } j(p) < p \\ 2k(p)+1 & \text{jinak} \end{cases} \end{aligned}$$

(c)

$$\begin{aligned} \Omega &= [0,1] \\ \mathcal{F} &= \textbf{Lebesgueovsky měřitelné množiny} \\ \Pr[A] &= \lambda(A) && \text{(pro libovolné } A \in \mathcal{F}) \\ X &\colon \Omega \to \mathbb{R} \\ X(\omega) &= \lceil 10x \rceil \end{aligned}$$

- **Je to náhodná veličina na daném pravděpodobnostním prostoru?**

  Řešení: Ano, pro libovolné $x \in \mathbb{R}$ platí $\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$ (každý interval je Lebesgueovsky měřitelný).

- **Je to diskrétní náhodná veličina?**

  Řešení: Ano, $\mathrm{Im}(X) = \{0,1,2,3,4,5,6,7,8,9,10\}$ je konečná množina. Příklad s obdobnou diskrétní náhodnou veličinou, kde $\mathrm{Im}(X)$ by byl nespočetný vynecháme.

- **Jaká je její střední hodnota?**

  Řešení: Napřed určíme pravděpodobnost:

  $$\begin{aligned} \Pr[X = 0] &= 0 \\ \Pr[X = j] &= 1/10 && \text{(pro každé } j \in \{1,2,3,4,5,6,7,8,9,10\}) \end{aligned}$$

  teď můžeme spočítat střední hodnotu:

  $$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \mathrm{Im}(X)} x \Pr[X = x] \\ &= 0 \cdot 0 + \frac{1}{10}(1+2+3+4+5+6+7+8+9+10) \\ &= 5.5 \end{aligned}$$

- **Jaká je její distribuční funkce?**

  Řešení:

  $$\begin{aligned} F_X(x) &= \Pr[X \leq x] \\ F_X(x) &= 0 && \text{(pokud } x < 1) \\ F_X(x) &= 1/10 && \text{(pro } x \in [1,2)) \\ F_X(x) &= 2/10 && \text{(pro } x \in [2,3)) \end{aligned}$$

$$\begin{aligned} &\dots \\ F_X(x) &= 1 && (\text{pro } x \in [10, \infty)) \end{aligned}$$

- **Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?**

  Řešení: Ne, protože F_X není spojitá.

- **Jaká je její kvantilová funkce?**

  Řešení:

$$\begin{aligned} Q_X &: [0,1] \to \mathbb{R} \\ Q_X(p) &= \inf\{x \in \mathbb{R} \mid p \leq F_X(x)\} \\ Q_X(p) &= \begin{cases} -\infty & \text{pokud } p = 0 \\ 1 & \text{pokud } p \in (0, 1/10] \\ 2 & \text{pokud } p \in (1/10, 2/10] \\ 3 & \text{pokud } p \in (2/10, 3/10] \\ 4 & \text{pokud } p \in (3/10, 4/10] \\ 5 & \text{pokud } p \in (4/10, 5/10] \\ 6 & \text{pokud } p \in (5/10, 6/10] \\ 7 & \text{pokud } p \in (6/10, 7/10] \\ 8 & \text{pokud } p \in (7/10, 8/10] \\ 9 & \text{pokud } p \in (8/10, 9/10] \\ 10 & \text{pokud } p \in (9/10, 1] \end{cases} \end{aligned}$$

(d)

$$\begin{aligned} \Omega &= [0,1] \\ \mathcal{F} &= \textbf{Lebesgueovsky měřitelné množiny} \\ \Pr[A] &= \lambda(A) && (\text{pro libovolné } A \in \mathcal{F}) \\ X &: \Omega \to \mathbb{R} \\ X(\omega) &= 10x \end{aligned}$$

- **Je to náhodná veličina na daném pravděpodobnostním prostoru?**

  Řešení: Ano, pro libovolné $x \in \mathbb{R}$ platí $\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$. Protože každý interval je Lebesgueovsky měřitelný a my máme:

$$\{\omega \in \Omega \mid X(\omega) \leq x\} = \begin{cases} \emptyset & \text{pro } x \in (-\infty, 0) \\ [0, x/10] & \text{pro } x \in [0, 10] \\ [0,1] & \text{pro } x \in (10, \infty) \end{cases}$$

- **Je to diskrétní náhodná veličina?**

  Řešení: Ne, $\mathrm{Im}(X) = [0, 10]$ je nespočetná množina.

- **Jaká je její distribuční funkce?**

Řešení:

$$
\begin{aligned}
F_X(x) &= \Pr[X \leq x] \\
F_X(x) &= 0 && \text{(pokud } x < 0\text{)} \\
F_X(x) &= x/10 && \text{(pokud } x \in [0, 10]\text{)} \\
F_X(x) &= 1 && \text{(pro } x \in [10, \infty)\text{)}
\end{aligned}
$$

- **Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?**

  Řešení: Ano, konečně máme F_X spojitou! Hustota je:

$$
\begin{aligned}
f_X(x) &\colon \mathbb{R} \to \mathbb{R} \\
F_X(x) &= \int_{-\infty}^{x} f_X(t)\ dt \\
f_X(x) &= 1/10 && \text{pro } x \in [0, 10] \\
f_X(x) &= 0 && \text{pro } x \in (-\infty, 0) \cup (10, \infty)
\end{aligned}
$$

  (uniformní rozdělení).

- **Jaká je její střední hodnota?**

  Řešení: Dle definice:

$$
\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\ dx \\
&= \int_{0}^{10} x/10\ dx \\
&= \left[x^2/20\right]_0^{10} \\
&= 5 && \text{(což dává smysl, hodnoty jsou rovnoměrně mezi } [0, 10]\text{)}
\end{aligned}
$$

- **Jaká je její kvantilová funkce?**

  Řešení:

$$
\begin{aligned}
Q_X &\colon [0, 1] \to \mathbb{R} \\
Q_X(p) &= \inf\{x \in \mathbb{R} \mid p \leq F_X(x)\} \\
Q_X(p) &= 10p && \text{(pro } p \in (0, 1]\text{)} \\
Q_X(p) &= -\infty && \text{(pro } p = 0\text{)}
\end{aligned}
$$

(e)

$$
\begin{aligned}
\Omega &= [0, 1] \\
\mathcal{F} &= \textbf{Lebesgueovsky měřitelné množiny} \\
\Pr[A] &= \lambda(A)/2 + 1/2 && \text{(pro libovolné } A \in \mathcal{F} \text{ pokud } 0.1 \in A\text{)} \\
\Pr[A] &= \lambda(A)/2 && \text{(pro libovolné } A \in \mathcal{F} \text{ pokud } 0.1 \notin A\text{)} \\
X &\colon \Omega \to \mathbb{R} \\
X(\omega) &= 10\omega
\end{aligned}
$$

- **Je to náhodná veličina na daném pravděpodobnostním prostoru?**

  Řešení: Úplně stejně jako minule.

- **Je to diskrétní náhodná veličina?**

  Řešení: Ano, $\mathrm{Im}(X) = [0, 10]$ je nespočetná množina.

- **Jaká je její distribuční funkce?**

  Řešení:

$$\begin{aligned}
&F_X(x) = \Pr[X \leq x] \\
&F_X(x) = 0 && \text{(pokud } x < 0\text{)} \\
&F_X(x) = x/20 && \text{(pokud } x \in [0, 1)\text{)} \\
&F_X(x) = 1/2 + x/20 && \text{(pokud } x \in [1, 10)\text{)} \\
&F_X(x) = 1 && \text{(pro } x \in [10, \infty)\text{)}
\end{aligned}$$

- **Je to spojitá veličina? Jinak řečeno máme pro tuto hustotu (probability density function PDF)?**

  Řešení: A zase je F_X nespojitá, takže X není spojitá veličina.

- **Jaká je její střední hodnota?**

  Řešení: Nejobecnější definice:

$$\begin{aligned}
\mathbb{E}[X] &= \int_\Omega X(\omega) d\Pr(\omega) \\
&= (R) \int_0^\infty \Pr\left(\{x \in \mathbb{R} \mid X(x) > t\}\right)\ dt && \text{(protože } X(\omega) \geq 0\text{)} \\
&= (R) \int_0^\infty (1 - F_X(t))\ dt \\
&= (R) \int_0^1 (1 - t/20)\ dt + (R) \int_1^{10} (1 - (1/2 + t/20))\ dt \\
&= \left[x - x^2/40\right]_0^1 + \left[x/2 - x^2/40\right]_1^{10} \\
&= 39/40 + (5 - 2.5) - (1/2 - 1/40) \\
&= 3
\end{aligned}$$

- **Jaká je její kvantilová funkce?**

  Řešení:

$$\begin{aligned}
&Q_X \colon [0, 1] \to \mathbb{R} \\
&Q_X(p) = \inf\{x \in \mathbb{R} \mid p \leq F_X(x)\} \\
&Q_X(0) = -\infty \\
&Q_X(p) = 20p && \text{(pro } p \in (0, 0.05)\text{)} \\
&Q_X(p) = 1 && \text{(pro } p \in [0.05, 0.55)\text{)} \\
&Q_X(p) = 20p - 10 && \text{(pro } p \in [0.55, 1]\text{)}
\end{aligned}$$

(f) Pozorujte, že kvantil je jediná funkce, pro kterou platí:

$$\forall p \in [0,1], \forall x \in \mathbb{R}\colon Q_X(p) \leq x \Leftrightarrow p \leq F_X(x)$$

3.6 Cvičení

1.

- **Jedno promile lidí má nemoc C, značíme $C^+ \subseteq \Omega$ množinu nemocných a $C^- = \Omega \setminus C^+$ množinu zdravých. Člověka volíme uniformně náhodně, pak $\Pr[C^+] = 0.001$.**
- **Máme test, který označí množinu lidí $T^+ \subseteq \Omega$ za nemocné a $T^- = \Omega \setminus T^+$ za zdravé. Test má následující parametry:**
  - ***Sensitivita:*** **(true positive)** $\Pr[T^+ \mid C^+] = 0.99$
  - ***Specificita:*** **(true negative)** $\Pr[T^- \mid C^-] = 0.98$

(a) Uniformně náhodně jsme vybrali člověka c. Jaká je pravděpodobnost, že $c \in C^+$ (tedy že je nemocný)?

Řešení:

$$\Pr[C^+] = 0.001$$

Vybrali jsme náhodného člověka a nic jiného o tom člověku nevíme. Takže racionálně odhadujeme pravděpodobnost, že je nemocný zlomkem nemocných v populaci.

(b) Člověku c jsme udělali jeden test a ten vyšel pozitivní. Jaká je pravděpodobnost, že $c \in C^+$ (tedy že je nemocný)?

Řešení:

$$\begin{aligned}\Pr[C^+ \mid T^+] &= \frac{\Pr[C^+]\Pr[T^+ \mid C^+]}{\Pr[T^+ \mid C^+]\Pr[C^+] + \Pr[T^+ \mid C^-]\Pr[C^-]} \\ &= \frac{0.001 \cdot 0.99}{0.99 \cdot 0.001 + 0.02 \cdot 0.999} \\ &= 0.0472\end{aligned}$$

Vytáhli jsme náhodného člověka, ale pak jsme provedli test, který vyšel pozitivně. Takže je rozumné předpokládat, že pravděpodobnost, že náš člověk je nemocný je vyšší. O kolik přesně upravíme náš odhad nemocnosti nám kvantifikuje právě Bayesova věta.

(c) Pro jistotu jsme člověku c udělali ještě jeden test a ten vyšel znovu pozitivní. Jaká je pravděpodobnost, že $c \in C^+$ (tedy že je nemocný)? Předpokládejte, že výsledek druhého testu je nezávislý na tom prvním.

Řešení:

$$\begin{aligned}\Pr[C^+ \mid T^{++}] &= \frac{\Pr[C^+]\Pr[T^{++} \mid C^+]}{\Pr[T^{++} \mid C^+]\Pr[C^+] + \Pr[T^{++} \mid C^-]\Pr[C^-]} \\ &= \frac{\Pr[C^+](\Pr[T^+ \mid C^+])^2}{(\Pr[T^+ \mid C^+])^2\Pr[C^+] + (\Pr[T^+ \mid C^-])^2\Pr[C^-]} \\ &= \frac{0.001 \cdot (0.99)^2}{(0.99)^2 \cdot 0.001 + (0.02)^2 \cdot 0.999} \\ &= 0.71\end{aligned}$$

Někteří z vás se správně zamysleli, jestli nezávislost testů platí i pokud výsledek podmíníme tím, že daný člověk je nemocný nebo ne. Někteří jste argumentovali, že pokud A, B, C jsou jevy, pak následující je rovnost (což obecně neplatí)

$$\Pr[A \cap B \mid C] \neq \Pr[A \mid C]\Pr[B \mid C]$$

Jako protipříklad vezměte hod dvěma spravedlivými rozlišitelnými šestistrannými kostkami,

$$\begin{aligned} A &= \text{první kostka sudé číslo} = \{2x, 4x, 6x \mid x \in \{1,2,3,4,5,6\}\} \\ B &= \text{druhá kostka sudé číslo} = \{x2, x4, x6 \mid x \in \{1,2,3,4,5,6\}\} \\ C &= \text{součet rovný čtyřem} = \{13, 22, 31\} \\ \Pr[A \cap B \mid C] &= 1/3 \neq 1/9 = \Pr[A \mid C]\Pr[B \mid C] \end{aligned}$$

Je tedy správné předpokládat nezávislost dvou testů, zejména pak pokud podmiňujeme tím, že člověk je zdravý (nebo nemocný)? Asi to úplně správné není a nejspíš by to měl výrobce testů vyzkoušet. Prakticky však můžeme vzít testy z různých várek (různé datum výroby). A konec konců, ono nám stejně nic moc jiného nezbývá.

Řešení: Opět se ale na tento problém můžeme dívat jako na problém, kdy jsme si vytáhli člověka náhodně z nějaké podmnožiny všech lidí, kde pravděpodobnost, že je člověk nemocný je $\Pr[C^+] = \frac{0.001 \cdot 0.99}{0.99 \cdot 0.001 + 0.02 \cdot 0.999}$. Ta podmnožina jsou přesně lidi, kterým vyšel první test pozitivní, takže dle minulého bodu máme pravděpodobnost, že je nemocný danou. Opět musíme předpokládat něco o tom druhém testu – zejména to, že sensitivita a specificita jsou stejné na této podmnožině lidí. Pak můžeme počítat:

$$\begin{aligned} \Pr[C^+ \mid T^+] &= \frac{\Pr[C^+]\Pr[T^+ \mid C^+]}{\Pr[T^+ \mid C^+]\Pr[C^+] + \Pr[T^+ \mid C^-]\Pr[C^-]} \\ &= \frac{\frac{0.001 \cdot 0.99}{0.99 \cdot 0.001 + 0.02 \cdot 0.999} \cdot 0.99}{0.99 \cdot \frac{0.001 \cdot 0.99}{0.99 \cdot 0.001 + 0.02 \cdot 0.999} + 0.02 \cdot \left(1 - \frac{0.001 \cdot 0.99}{0.99 \cdot 0.001 + 0.02 \cdot 0.999}\right)} \\ &= 0.71 \end{aligned}$$

Tohle je poměrně užitečný pohled na Bayesovu větu:

- $\Pr[C^+ \mid T^+]$ je takzvaná posterior probability (to, co si myslíme, když uvidíme nějakou evidence – nějakou stopu, že tomu tak je).
- $\Pr[C^+]$ a $\Pr[T^+]$, což je prior (to, co jsme si mysleli předtím).

Místo $\Pr[T^+]$ se skoro vždy používá rozpis $\Pr[T^+ \mid C^+]\Pr[C^+] + \Pr[T^+ \mid C^-]\Pr[C^-]$ (protože to jsou věci, které v praxi známe).

Jen mimochodem poznamenejme, že tyhle dva postupy jsou ekvivaletní, což můžeme ukázat manipulací se symboly:

$$\begin{aligned} &\frac{\frac{\Pr[C^+]\Pr[T^+|C^+]}{\Pr[T^+|C^+]\Pr[C^+]+\Pr[T^+|C^-]\Pr[C^-]}\Pr[T^+ \mid C^+]}{\Pr[T^+ \mid C^+]\frac{\Pr[C^+]\Pr[T^+|C^+]}{\Pr[T^+|C^+]\Pr[C^+]+\Pr[T^+|C^-]\Pr[C^-]} + \Pr[T^+ \mid C^-]\left(1 - \frac{\Pr[C^+]\Pr[T^+|C^+]}{\Pr[T^+|C^+]\Pr[C^+]+\Pr[T^+|C^-]\Pr[C^-]}\right)} = \\ &= \frac{\Pr[C^+](\Pr[T^+ \mid C^+])^2}{(\Pr[T^+ \mid C^+])^2\Pr[C^+] + (\Pr[T^+ \mid C^-])^2\Pr[C^-]} \end{aligned}$$

(d) Simulujte.

Řešení:

```
from random import random


def bernoulli(pr: float = 0.5) -> bool:
    return random() < pr
```

```python
class Human:
    """ _illness_probability is our unknown! """
    _illness_probability = 0.001

    """Random human."""
    def __init__(self):
        self.is_ill = bernoulli(Human._illness_probability)


class IllnessTest:
    sensitivity = 0.99  # = Pr[T+|C+]
    specificity = 0.98  # = Pr[T-|C-]

    def test(h: Human) -> bool:
        """Return True if the test says h is ill."""
        if h.is_ill:
            return bernoulli(IllnessTest.sensitivity)
        else:
            return not bernoulli(IllnessTest.specificity)


N = 10000000  # Number of samples
positive_test1 = 0
positive_test2 = 0
ill_given_positive_test = 0
ill_given_2_positive_tests = 0

for _ in range(N):
    h = Human()
    if IllnessTest.test(h):
        positive_test1 += 1
        if h.is_ill:
            ill_given_positive_test += 1
        if IllnessTest.test(h):
            positive_test2 += 1
            if h.is_ill:
                ill_given_2_positive_tests += 1

pr_c_given_t = (Human._illness_probability * IllnessTest.sensitivity
                / (IllnessTest.sensitivity * Human._illness_probability
                        + (1 - IllnessTest.specificity) * (1 - Human._illness_probabilit

pr_c_given_2t = (Human._illness_probability * IllnessTest.sensitivity**2
                / (IllnessTest.sensitivity**2 * Human._illness_probability
                        + (1 - IllnessTest.specificity)**2 * (1 - Human._illness_probabi

print(f'Pr[C+|T+] = {ill_given_positive_test/positive_test1} (={pr_c_given_t})')
print(f'Pr[C+|T++] = {ill_given_2_positive_tests/positive_test2} (={pr_c_given_2t})')

# Možný výstup:
# Pr[C+|T+] = 0.047611317475626 18 (=0.047210300429184504)
# Pr[C+|T++] = 0.7123386457056321 (=0.7103718199608607)
```

Poznámka: Toto je případ kdy je nemoc velice zřídkavá a ty nemáš symptomy. Pokud máš symptomy a jdeš na test, tak nejsi náhodně vybraný člověk! Tento příklad je tedy spíš situace jdu darovat krev a oni musí udělat povinný test na HIV a ten vyjde pozitivní, ale přesto bych se neměl tolik strachovat, ale v klidu jít na druhý test (už to že chodím darovat krev negativně koreluje s nakažením HIV).

2. **Pojďme se podívat na další vlastnosti jednotlivých rozdělení:**

   (a) Bernoulli:

   - **Kde se použije?**

     Řešení: Pokus, který má jen dva výsledky $X = 0$ nebo $X = 1$ kde $X \sim Bern(p)$ a $p \in [0, 1]$. Zápis $X \sim Bern(p)$ znamená, že $\Pr[X = 1] = p$. Například hod (cinklou) mincí kde $X\colon \{hlava, orel\} \to \{0, 1\}$.

   - **Jakou má střední hodnotu a rozptyl?**

     Řešení:

$$\begin{aligned} X &\sim Bern(p) && (p \in [0,1]) \\ \Pr[X = 1] = p &= 1 - \Pr[X = 0] \\ \mathbb{E}[X] &= p \\ \mathrm{var}(E) &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 \end{aligned}$$

   - **Co se stane, když sečtu dvě nebo obecně n-náhodných veličin $X_1, X_2, \ldots, X_n$, které jsou nezávislé a pro nějaké fixní $p \in [0, 1]$ a každé $j \in [n]$ platí že $X_j \sim Bern(p)$.**

     Řešení: Dostanu $X = \sum_{j=1}^{n} X_j$ náhodnou veličinu. O té vím, že $X \sim Bin(n, p)$, tedy že je rozdělená podle binomického rozdělení (n hodů, každý vyjde s pravděpodobností p, X rovno počtu těch, které vyšly).

     Poznámka: ta nezávislost je velice důležitý předpoklad! Pokud bychom například měli $X_1 \sim Bern(p)$ a pak $X_1 = X_2 = \ldots = X_n$, pak určitě nejsou nezávislé a jejich součet není rozdělený jako binomické rozdělení.

   (b) geometrické:

   - **Kde se použije?**

     Řešení: Čekání na první úspěch v sérii Bernoulliho pokusů.

   - **Jakou má střední hodnotu a rozptyl?**

     Řešení:

$$\begin{aligned} X &\sim Geom(p) && (p \in [0,1]) \\ \Pr[X = n] &= (1-p)^{n-1}p && (n = 1, 2, \ldots) \\ \mathbb{E}[X] &= 1/p && (\text{lemma o džbánu}) \\ \mathrm{var}(E) &= \frac{1-p}{p^2} \end{aligned}$$

   - **Co se stane, když vezmeme minimum ze dvou nezávislých náhodných veličin X_1, X_2, které jsou kde $X_1 \sim Geom(p_1)$ a $X_2 \sim Geom(p_2)$?**

     Řešení: Dostaneme $X = \min(X_1, X_2)$ kde $X \sim Geom(1 - (1 - p_1)(1 - p_2))$

$$\begin{aligned} \Pr[X = k] &= \Pr[\min(X_1, X_2) = k] \\ &= \Pr[X_1 = k, X_2 < k] + \Pr[X_1 < k, X_2 = k] + \Pr[X_1 = X_2 = k] \\ &= (1-p_1)^{k-1} p_1 \left(\sum_{j=1}^{k-1} (1-p_2)^{j-1} p_2 \right) \end{aligned}$$

$$+\left(\sum_{j=1}^{k-1}(1-p_1)^{j-1}p_1\right)(1-p_2)^{k-1}p_2$$
$$+(1-p_1)^{k-1}p_1(1-p_2)^{k-1}p_2$$

zbytek výpočtu pro odvážné.

(c) Binomické:

- **Kde se použije?**

  Řešení: Počet úspěchů z n pokusů. Například pokud testovou otázku znám s pravděpodobností p a celkem je otázek n, tak mám správně $X \sim Bin(n,p)$ z nich.

- **Jakou má střední hodnotu a rozptyl?**

  Řešení:

$$X \sim Bin(n,p) \qquad (n \in \mathbb{N}, p \in [0,1])$$
$$\Pr[X=j] = \binom{n}{j}p^j(1-p)^{n-j} \qquad \text{(pro libovolné } 0 \le j \le n)$$
$$\mathbb{E}[X] = np \qquad \text{(linearita střední hodnoty)}$$
$$\operatorname{var}(E) = n(p-p^2) \qquad \text{(rozptyl nezávislých je součet rozptylů – domácí úkol)}$$

- **Co se stane, když sečtu dvě náhodné veličiny X_1, X_2, které jsou nezávislé a pro nějaké fixní $p \in [0,1]$ platí $X_1 \sim Bin(n,p)$ a $X_2 \sim Bin(m,p)$ (kde $n, m \in \mathbb{N}$).**

  Řešení: Pokud $X = X_1 + X_2$, pak $X \sim Bin(n+m,p)$ (napřed udělám n pokusů a pak m nebo rovnou udělám $n+m$ pokusů).

  Můžeme to dokázat i formálně pomocí takzvané konvoluce:

$$\begin{aligned}\Pr[X=j] &= \sum_{k=0}^{j}\Pr[X_1=k]\Pr[X_2=j-k] \\ &\qquad\qquad \text{(některé z pravděpodobností můžou být nulové)} \\ &= \sum_{k=0}^{j}\binom{n}{k}p^k(1-p)^{n-k}\binom{m}{j-k}p^{j-k}(1-p)^{m-(j-k)} \\ &= p^j(1-p)^{m+n-j}\sum_{k=0}^{j}\binom{n}{k}\binom{m}{j-k} \\ &= p^j(1-p)^{m+n-j}\binom{n+m}{j}\end{aligned}$$

(d) Poissonovo:

- **Kde se použije?**

  Řešení: Pokud známe průměrný počet výskytů za nějaký čas a každý výskyt je nezávislý na ostatních (příklady z `https://en.wikipedia.org/wiki/Poisson_distribution`):

  – Například počet atomů, které se rozpadnou za minutu v našem vzorku.

- Počet záplav na daném území ve sto letech (aby pojišťovna mohla spočítat kolik má být pojistné aby s dobrou pravděpodobností neprodělala).

- Počet gólů v jednom fotbalovém zápase.

Většinou je to jen aproximace (v našem vzorku není nekonečně mnoho atomů, záplav ve sto letech taky nebude milion, počet gólů ve fotbalovém zápase taky nebude libovolně velký).

- **Jakou má střední hodnotu a rozptyl?**

Řešení:

$$X \sim Pois(\lambda) \qquad (\lambda > 0)$$
$$\Pr[X = k] = \frac{\lambda^k e^{-\lambda}}{k!} \qquad (k \in \mathbb{N})$$
$$\mathbb{E}[X] = \lambda$$
$$\mathrm{var}(E) = \lambda$$

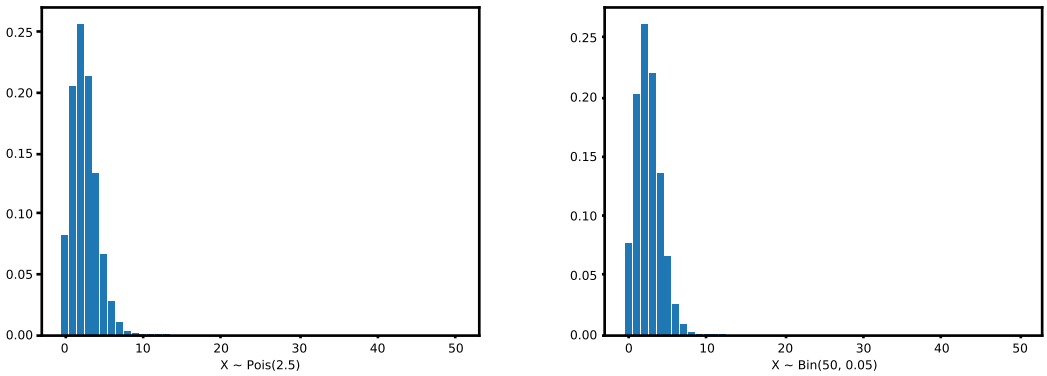
- **Co se stane, když sečtu dvě náhodné veličiny X_1, X_2, které jsou nezávislé a $X_1 \sim Pois(\lambda)$, a $X_2 \sim Pois(\mu)$.**

Řešení: Dokážeme, že $X = X_1 + X_2$ je rozdělená $X \sim Pois(\lambda + \mu)$:

$$\begin{aligned}
\Pr[X = k] &= \sum_{j=0}^{k} \Pr[X_1 = j] \Pr[X_2 = k - j] \\
&= \sum_{j=0}^{k} \frac{\lambda^j e^{-\lambda}}{j!} \frac{\mu^{k-j} e^{-\mu}}{(k-j)!} \\
&= e^{-(\lambda+\mu)} \sum_{j=0}^{k} \frac{\lambda^j}{j!} \frac{\mu^{k-j}}{(k-j)!} \\
&= e^{-(\lambda+\mu)} \frac{\sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \lambda^j \mu^{k-j}}{k!} \\
&= e^{-(\lambda+\mu)} \frac{\sum_{j=0}^{k} \binom{k}{j} \lambda^j \mu^{k-j}}{k!} \\
&= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!}
\end{aligned}$$

- **Co se stane, pokud zvolíme $\lambda = np$ a porovnáme Poissonovo rozdělení a binomiální rozdělení?**

Řešení: Pokud $n \geq 20, p \leq 0.05$ (nebo $n \geq 100, np \leq 10$), pak ta rozdělení jsou dost podobná. Porovnejte na obrázcích pro první volbu $n = 50, p = 0.05, \lambda = np = 2.5$:



(a) Poissonovo rozdělení s parametrem λ.

(b) Binomiální rozdělení

Obrázek 3.8: Aproximace binomiálního rozdělení Poissonovým.

```
import math
import matplotlib.pyplot as plt
from scipy.special import comb

n = 50
p = 0.05
l = n * p  # lambda is a python keyword

poisson = {}
for k in range(n+1):
    poisson[k] = l**k * math.exp(-l) / math.factorial(k)

plt.figure(0)
plt.bar(poisson.keys(), poisson.values())
plt.xlabel(f'X ~ Pois({l})')
# plt.ylabel("")
# plt.show()
plt.savefig(f'poisson.pdf')

binomial = {k: comb(n, k, True) * p**k * (1-p)**(n-k) for k in range(n+1)}

plt.figure(1)
plt.bar(binomial.keys(), binomial.values())
plt.xlabel(f'X ~ Bin({n}, {p})')
# plt.ylabel("")
# plt.show()
plt.savefig(f'binomial.pdf')
```

Mimochodem to, že volíme $\lambda = np$ a rozdělení jsou si podobná dává smysl, mají stejnou střední hodnotu a skoro i rozptyl (pro malé p). Pokud bychom měli konstantní $\lambda = np$ a limitili $Bin(n, \lambda/n)$ pro n větší a větší, tak bychom dostali totéž rozdělení.

- **Pro odvážné: ukažte, že pokud** $X_1 \sim Pois(\lambda_1)$ **a** $X_2 \sim Pois(\lambda_2)$ **jsou nezávislé, pak** $\Pr[X_1 = k \mid X_1 + X_2 = n] = \Pr[Y = k]$ **kde** $Y \sim Bin(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$**.**

3. **de Mèreho problém házeme spravedlivými šestistrannými kostkami:**

   (a) Jaká je pravděpodobnost, že padne ze čtyř hodů aspoň jedna šestka?

   Řešení: Můžeme spočítat pravděpodobnost, že ani v jednom ze čtyř hodů nepadne šestka a pak vzít pravděpodobnost doplňku:

$$\begin{aligned}\Pr[\text{ze čtyř hodů aspoň jedna šestka}] &= 1 - \Pr[\text{v žádném ze čtyř hodů nepadne šestka}] \\ &= 1 - (5/6)^4 \\ &= 0.5177\end{aligned}$$

   (b) Jaká je pravděpodobnost, že padne z 24 hodů dvojicí kostek aspoň jedna dvojitá šestka?

   Řešení: Obdobně:

$$1 - (35/36)^{24} = 0.4914$$

   (c) De Mèreho problém spočíval v tom, jestli jsou odpovědi na části (a), (b) stejné. Tak se Chevalier de Mère zeptal Blaise Pascala a ten spolu s Pierre de Fermatem položili základy teorie pravděpodobnosti.

   (d) Umíte úlohu interpretovat jako otázku o binomickém rozdělení?

   Řešení: V první části se ptáme na otázku $\Pr[X \geq 1]$, kde $X \sim Bin(4, 1/6)$. Ve druhé části se ptáme na otázku $\Pr[X \geq 1]$, kde $X \sim Bin(24, 1/36)$.

   (e) Umíte úlohu interpretovat jako otázku o geometrickém rozdělení?

   Řešení: V první části se ptáme na otázku $\Pr[Y \leq 4]$, kde $Y \sim Geom(1/6)$. Ve druhé části se ptáme na otázku $\Pr[Y \leq 24]$, kde $Y \sim Geom(1/36)$.

4. **Odhadněte π pomocí náhodných pokusů.**

   Řešení: Budeme generovat bod $(x, y) \in [0, 1] \times [0, 1]$ uniformně náhodně, tedy $x \sim U(0, 1)$ a $y \sim U(0, 1)$ nezávislé. Pomocí experimentu odhadneme pravděpodobnost, že (x, y) leží ve čtvrtkruhu poloměru 1, tedy jestli $x^2 + y^2 \leq 1$.

```python
from random import random

N = 1_000_000

ve_ctvrtkruhu = 0

for _ in range(N):
    x = random()
    y = random()
    if x**2 + y**2 <= 1:
        ve_ctvrtkruhu += 1

# Kruh poloměru 1 má obsah pi
print(f'pi = {4*ve_ctvrtkruhu/N}')

# Možný výstup:
# pi = 3.14286
```

5. **Nechť X je náhodná veličina. Vyjádřete pomocí $F_X(t) = \Pr[X \leq t]$ pro každé $t \in \mathbb{R}$ distribuční funkci náhodných veličin:**

   (a) $X^+ = \max(0, X)$

   Řešení:

$$\begin{aligned} \Pr[X^+ \leq t] &= \Pr[\max(0, X) \leq t] \\ &= \Pr[X \leq t] && \text{(pro } t \geq 0\text{)} \\ &= F_X(t) \end{aligned}$$

   Pro záporná t:

$$\Pr[\max(0, X) \leq t] = 0 \qquad \text{(pro } t < 0\text{)}$$

   (b) $-X$

   Řešení:

$$\begin{aligned} F_{-X}(t) &= \Pr[-X \leq t] \\ &= \Pr[X \geq -t] \\ &= 1 - \Pr[X < -t] \\ &= 1 - \lim_{x \to (-t)^-} F_X(x) && \text{(limita když } x \text{ se blíží k } -t \text{ zleva)} \end{aligned}$$

   (c) $X^- = -\min(X, 0)$

   Řešení:

$$\begin{aligned} \Pr[X^- \leq t] &= \Pr[-\min(0, X) \leq t] \\ &= \Pr[-t \leq X] && \text{(pro } t \geq 0\text{)} \\ &= 1 - \Pr[X < -t] \\ &= 1 - \lim_{x \to (-t)^-} F_X(x) \end{aligned}$$

   (d) $|X|$

   Řešení: Rovnou jako absolutní hodnota:

$$\begin{aligned} \Pr[|X| \leq t] &= \Pr[-t \leq X \leq t] \\ &= F_X(t) - \lim_{x \to (-t)^-} F_X(x) && \text{(pro } t \geq 0\text{)} \end{aligned}$$

6. **Mějme spojitou náhodnou veličinu X danou její pravděpodobnostní hustotou (probability density function – PDF)**

$$f_X(t) = \begin{cases} 0 & t < 1 \\ 2/x^3 & t \geq 1 \end{cases}$$

(a) Spočítejte $\Pr[X \in [5, 10]]$**.**

Řešení:

$$\begin{aligned} \Pr[X \in [5,10]] &= \int_5^{10} f_X(x)\ dx \\ &= \int_5^{10} 2/x^3\ dx \\ &= \left[-1/x^2\right]_5^{10} \\ &= -1/100 - (-1/25) \\ &= 0.03 \end{aligned}$$

(b) Spočítejte $\Pr[X \geq 100]$**.**

Řešení:

$$\begin{aligned} \Pr[X \geq 100] &= \int_{100}^{\infty} f_X(x)\ dx \\ &= \lim_{b\to\infty} \int_{100}^{b} f_X(x)\ dx \\ &= \lim_{b\to\infty} \left[-1/x^2\right]_{100}^{b} \\ &= \lim_{b\to\infty} -1/b^2 + 1/10000 \\ &= 1/10000 \end{aligned}$$

(c) Spočítejte $\mathbb{E}[X]$**.**

Řešení:

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\ dx \\ &= \int_1^{\infty} x 2/x^3\ dx \\ &= \int_1^{\infty} 2/x^2\ dx \\ &= \lim_{b\to\infty} \int_1^{b} 2/x^2\ dx \\ &= \lim_{b\to\infty} \left[-2/x\right]_1^{b} \\ &= \lim_{b\to\infty} -2/b + 2/1 \\ &= 2 \end{aligned}$$

(d) Určete distribuční funkci (cumulative distribution function – CDF).

Řešení:

$$\begin{aligned} F_X(t) &= \Pr[X \leq t] \\ &= \int_{-\infty}^{t} f_X(x)\ dx \\ &= \int_{1}^{t} 2/x^3\ dx \\ &= \left[-1/x^2\right]_1^t \\ &= -1/t^2 - (-1/1) \\ &= 1 - 1/t^2 \end{aligned}$$

Pro $t < 1$ máme $F_X(t) = 0$. Pozor na to, že pro distribuční funkci musí platit

$$\lim_{t\to\infty} F_X(t) = 1$$

takže integrační konstanta je jednoznačně určená.

7. **Mějme spojitou nezápornou náhodnou veličinu X, která má hustotu (probability density function – PDF) f_x (a tedy comulative distribution function – CDF $F_X(t) = \int_0^t f_X(x)\ dx$). Ukažte, že $\mathbb{E}[X] = \int_0^\infty x f_X(x)\ dx = \int_\Omega F_X(\omega) d\Pr(\omega) = \int X d\Pr$.**

   Pokud by X nebyla nezáporná, tak ji můžeme vyjádřit jako rozdíl dvou nezáporných náhodných veličin.

   Řešení: Napřed dle definice z minula upravme ten druhý integrál:

$$\begin{aligned}\int_\Omega F_X(\omega) d\Pr(\omega) &= (R)\int_0^\infty \Pr[X > t]\ dt \\ &= (R)\int_0^\infty 1 - \Pr[X \leq t]\ dt \\ &= (R)\int_0^\infty 1 - F_X(t)\ dt\end{aligned}$$

   tedy chceme ukázat rovnost:

$$\int_0^\infty x f_X(x)\ dx = \int_0^\infty 1 - F_X(t)\ dt$$

   Napřed spočítáme ten integrál jen do $b > 0$, pak vezmeme limitu $b \to \infty$:

$$\begin{aligned}\int_0^b x f_X(x)\ dx &= [x F_X(x)]_0^b - \int_0^b F_X(x)\ dx && \text{(per partes)} \\ &= b F_X(b) - \int_0^b F_X(x)\ dx \\ &= \int_0^b F_X(b) - F_X(x)\ dx\end{aligned}$$

   A pak už jen využijeme toho, že $\lim_{b\to\infty} F_X(b) = 1$.

   Přiznejme ještě, že jsme nezdůvodnili, že si můžeme sáhnout limitu $F_X(b)$ uvnitř toho integrálu.

3.7 Cvičení

1. **Spočítejte střední hodnotu následujících nezáporných náhodných veličin pomocí vzorce**

$$\mathbb{E}[X] = \int_0^\infty 1 - F_X(t)\ dt \qquad \text{(pro nezápornou náhodnou veličinu } X\text{)}$$

(a) $\mathrm{Im}(X) = \{2, \pi\}$**,** $\Pr[X = 2] = 1/3, \Pr[X = \pi] = 2/3$

Řešení: Napřed určíme distribuční funkci (cumulative distribution function)

$$F_X(t) = \begin{cases} 0 & t \in (-\infty, 2) \\ 1/3 & t \in [2, \pi) \\ 1 & t \in [\pi, \infty) \end{cases}$$

Nyní můžeme dosadit do vzorce (náhodná veličina je nezáporná) *taky ten integrál nakreslete!*

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty 1 - F_X(t)\ dt \\ &= \int_0^2 1 - 0\ dt + \int_2^\pi 1 - 1/3\ dt + \int_\pi^\infty 1 - 1\ dt \\ &= \int_0^2 1\ dt + \int_2^\pi 2/3\ dt + \int_\pi^\infty 0\ dt \\ &= 2 + 2/3(\pi - 2) \\ &= 2/3 + 2\pi/3 \end{aligned}$$

Spočítejte tu střední hodnotu tak jak jste zvyklí:

$$\mathbb{E}[X] = 2/3 + (2/3)\pi$$

(b) X **je dána hustotou (probability density function)** $f_X(t) = 1/3$ **pokud** $t \in [0, 3]$ **a jinak** $f_X(t) = 0$**.**

Řešení: Napřed určíme distribuční funkci (cumulative distribution function)

$$\begin{aligned} F_X(t) &= \int_0^t f_X(x)\ dx \\ &= t/3 \end{aligned}$$

Nyní můžeme dosadit do vzorce (náhodná veličina je nezáporná) *taky ten integrál nakreslete!*

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty 1 - F_X(t)\ dt \\ &= \int_0^3 1 - t/3\ dt + \int_3^\infty 0\ dt \\ &= \left[t - \frac{t^2}{6}\right]_0^3 \\ &= (3 - 9/6) - (0 - 0) \end{aligned}$$

$$= 3/2$$

(což odpovídá prostředku, protože máme uniformní rozdělení).

Spočítejte tu střední hodnotu tak jak jste zvyklí:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty t f_X(t)\ dt \\
&= \int_0^3 t/3\ dt \\
&= \left[t^2/6\right]_0^3 \\
&= (9/6) - (0/6) \\
&= 3/2
\end{aligned}$$

2.

(a) Máme minci, kde padne hlava s pravděpodobností $p \in (0,1)$ (ale my ani neznáme p). Jak pomocí ní vygenerujeme hod spravedlivou mincí?

Řešení: Použijeme následující trik: využijeme toho, že ve dvou hodech platí

$$p(1-p) = \Pr[HO] = \Pr[OH] = (1-p)p$$

Hodíme tedy mincí dvakrát, pokud padne HO, odpovíme "hlava," pokud padne OH odpovíme "orel," pokud padne HH nebo OO házíme znova.

Kolik je střední hodnota počtu potřebných dvoj-hodů? Čekáme na první úspěch, takže je to geometrická distribuce, o které víme střední hodnotu $1/(p(1-p))$.

(b) Máme spravedlivou minci, jak pomocí ní vygenerujete hod mincí, kde padne hlava s pravděpodobností $p \in (0,1)$?

Řešení: Kdyby $p = n2^{-m}$ pro nějaká $n, m \in \mathbb{N}$, tak nám stačí m hodů.

Ale co když $p = 1/\pi$? To nemá v dvojkové reprezentaci konečný zápis, takže žádný konečný počet hodů nám nestačí na přesné řešení (rozmyslete argument: strom hloubky T a obarvíme listy).

Představme si, že generujeme číslo $b \in [0,1]$ tak, že $b = 0.h_1h_2h_3\ldots$ kde h_j je výsledek hodu spravedlivou mincí (nula nebo jedna) a číslo b interpretujeme jako binární číslo $b = \sum_{j=1}^{\infty} h_j 2^{-j}$. Pokud $b \leq p$, pak odpovíme hlava, jinak orel. Ale to jsme si nepomohli, protože musíme hodit nekonečněkrát. Musíme? Pro spoustu hodnot můžeme určit jestli nutně $b < p$ nebo $b > p$ pomocí prvních pár hodů. Například pokud $p = 0.00101110\ldots$, a my hodíme první dvatinné místo rovné jedné, pak $b = 0.1???$ a můžeme skončit rovnou. Dá se nahlédnout, že očekávaný počet hodů je konstantní (a pokud p je $n2^{-m}$, pak stačí nejvýš m hodů).

To je mnohem lepší řešení, než to které nás napadlo jako první. V matematice a životě se občas vyplatí ptát se na obecnější otázky, protože nám mohou dát mnohem lepší řešení než jsme čekali.

(c) Máme možnost generovat $Y \sim U(0,1)$. Jak pomocí toho vygenerujeme hod mincí kde padne hlava s pravděpodobností p?

Řešení: To už jsme vlastně vyřešili v předchozím příkladu. Odpovíme hlava pokud $p \leq Y$.

(d) Máme diskrétní náhodnou veličinu $\mathrm{Im}(X) = \{x_1, x_2, x_3, x_4\}$ (nechť $x_j < x_i$ pro $j < i$). Známe $\Pr[X = x_j] = p_j$ a tím pádem známe všechny běžné parametry (p_X, F_X, Q_X). Máme možnost generovat $Y \sim U(0,1)$. Jak pomocí výsledku Y vygenerujeme výsledek X?

Řešení: Chceme zobecnit příklad z minula. Tedy chceme rozdělit interval $[0,1]$ na intervaly délek p_1, p_2, p_3, p_4. Nejjednodušší je rozdělit je podle částečných součtů.

- Odpovíme x_1 pokud $0 \leq Y < p_1$.
- Odpovíme x_2 pokud $p_1 \leq Y < p_1 + p_2$.
- Odpovíme x_3 pokud $p_1 + p_2 \leq Y < p_1 + p_2 + p_3$.
- Odpovíme x_4 pokud $p_1 + p_2 + p_3 \leq Y < p_1 + p_2 + p_3 + p_4 = 1$.

Takže

- Odpovíme x_1 pokud $0 \leq Y < F_X(x_1)$.

- Odpovíme x_2 pokud $F_X(x_1) \leq Y < F_X(x_2)$.
- Odpovíme x_3 pokud $F_X(x_2) \leq Y < F_X(x_3)$.
- Odpovíme x_4 pokud $F_X(x_3) \leq Y < F_X(x_4) = 1$.

Kdybychom tak jen měli funkci, která se co nejvíc snaží chovat jako inverzní funkce k F_X.

Odpovíme $Q_X(Y)$.

3. **Autobus přijede v čas $e \doteq 2.71$. Čas mého příchodu na zastávku je náhodná veličina X s hustotou**

$$f_X(t) = \begin{cases} 1/t & \textbf{pokud } t \in [1, e] \\ 0 & \textbf{jinak} \end{cases}$$

(a) Jaká je pravděpodobnost, že přijdu v některý čas z intervalu $[1.5, 2]$?

Řešení: Máme hustotu, tedy

$$\begin{aligned} \Pr[X \in [1.5, 2]] &= \int_{1.5}^{2} f_X(t)\ dt \\ &= \int_{1.5}^{2} 1/t\ dt \\ &= [\ln(t)]_{1.5}^{2} \\ &= \ln(2) - \ln(1.5) \\ &\doteq 0.288 \end{aligned}$$

(b) Jaká je distribuční funkce času mého příchodu (cumulative distribution function)?

Řešení: Máme hustotu, tedy

$$F_X(t) = \begin{cases} 0 & t \in (-\infty, 1) \\ \int_1^t 1/x\ dx = [\ln(x)]_1^t = \ln(t) & t \in [1, e) \\ 1 & t \in [e, \infty) \end{cases}$$

(c) Jaká je kvantilová funkce času mého příchodu?

Řešení:

$$\begin{aligned} Q_X &\colon [0, 1] \to \mathbb{R} \\ Q_X(p) &= \inf\{x \in \mathbb{R} \mid p \le F_X(x)\} \\ &= \inf\{x \in \mathbb{R} \mid p \le \ln(x)\} \\ Q_X(p) &= e^p \end{aligned}$$

My máme spojitou distribuční funkci F_X, tedy kvantilová funkce bude $Q_X(p) = (F_X)^{-1}(p)$ (inverzní funkce).

(d) Jaká je střední hodnota času mého příchodu?

Řešení: Můžeme použít třeba vzorec s hustotou

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} t f_X(t)\ dt \\ &= \int_{1}^{e} t/t\ dt \\ &= [1]_1^e \\ &= e - 1 \\ &\doteq 1.71 \end{aligned}$$

(e) Jaký je rozptyl času mého příchodu?

Řešení: Tady bude o malinko jednodušší počítat s hustotou, protože z přednášky (LOTUS) víme:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(t) f_X(t)\ dt \qquad \text{(pro rozumné funkce } g\text{)}$$

(tedy nemusíme určovat F_{X^2}.)

$$\begin{aligned}
\operatorname{var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= \int_1^e t^2/t\ dt - (e-1)^2 \\
&= \left[t^2/2\right]_1^e - (e-1)^2 \\
&= e^2/2 - 1/2 - (e-1)^2 \\
&= -e^2/2 + 2e - 3/2 \\
&\doteq 0.242
\end{aligned}$$

(f) Jaká je střední doba čekání na autobus?

Řešení: Pokud přijdeme v čas X, pak čekáme $e - X$ jednotek času:

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int_{-\infty}^{\infty} g(t) f_X(t)\ dt \qquad \text{(pro rozumné funkce } g\text{)} \\
\mathbb{E}[e-X] &= \int_1^e (e-t) 1/t\ dt \\
&= [e\ln(t) - t]_1^e \\
&= (e-e) - (e\cdot 0 - 1) \\
&= 1
\end{aligned}$$

(g) Simulujte.

Řešení: Máme problém, nevíme jak generovat náhodný float, který se bude chovat jako X (bude mít stejnou distribuci). Zachrání nás věta z přednášky (nebo bychom taky mohli říct teorie míry a integrálu, nebo věta o substituci u integrálu):

Nechť X je náhodná veličina s distribuční funkcí $F_X = F$, nechť je F spojitá a rostoucí. Pak $F(X) \sim U(0,1)$ (rovnoměrně náhodná z intervalu $[0,1]$).

Nám se hodí její varianta nazpátek:

Nechť je F funkce "typu distribuční funkce," tedy

- neklesající $\forall x < y \in \mathbb{R}\colon F(x) \le F(y)$
- zprava spojitá $\forall x \in \mathbb{R}\colon \lim_{t\to x^+} F(t) = F(x)$
- $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$

Nechť Q je příslušná kvantilová funkce k naší F. Nechť $U \sim U(0,1)$, nechť $X = Q(U)$, pak X má distribuční funkci F.

```python
import math
import random

def F_X(t):
    # F_X: R -> [0,1]
    if t <= 1:
        return 0
    if t >= math.e:
        return 1
    return math.log(t)

def Q_X(p):
    # Q_X: [0,1] -> R
    return math.exp(p)

def X():
    # U ~ U(0,1)
    # Q_X(U) ma rozdeleni F_X
    return Q_X(random.random())

N = 1_000_000

pr_a = sum(int(1.5 <= X() <= 2) for _ in range(N)) / N
print(f'a) Pr[1.5 <= X <= 2] = {pr_a} (={math.log(2) - math.log(1.5)})')

t = 2.3  # nejaky vstup
# Pr[X <= t] = F_23
F_23 = sum(int(X() <= t) for _ in range(N)) / N
print(f'b) F_X({t}) = {F_23} (={F_X(t)})')

EX = sum(X() for _ in range(N)) / N
print(f'd) E[X] = {EX} (={math.e - 1})')

varX = sum((X() - EX)**2 for _ in range(N)) / N
print(f'e) var(X) = {varX} (={-math.e**2/2 + 2*math.e - 3/2})')

E_cekani = sum(math.e - X() for _ in range(N)) / N
print(f'f) E[cekani] = {E_cekani} (={1})')

# Možný výstup:
# a) Pr[1.5 <= X <= 2] = 0.287366 (=0.2876820724517809)
# b) F_X(2.3) = 0.832592 (=0.8329091229351039)
# d) E[X] = 1.718278932279249 (=1.718281828459045)
# e) var(X) = 0.24186437641415617 (=0.24203560745276542)
# f) E[cekani] = 1.000026495456767 (=1)
```

4. **Knihovna MFF má 1000 čtenářů – studentů informatiky – a rozhoduje se, kolik kopií nové knihy koupit. Předpokládejme, že o knihu má v daný semestr každý student zájem s pravděpodobností $p = 0.01$, nezávisle na ostatních.**

   (a) Určete pravděpodobnostní funkci pro počet studentů, kteří mají o knihu zájem.

   Řešení: Jedná se o binomické rozdělení $X \sim Bin(n, p)$, konkrétně $X \sim Bin(1000, 0.01)$ víme tedy:

$$\begin{aligned} p_X(j) = \Pr[X = j] &= \binom{n}{j} p^j (1-p)^{n-j} \\ &= \binom{1000}{j} 0.01^j (1-0.01)^{1000-j} \qquad \text{(pro přirozené } 0 \leq j \leq 1000) \end{aligned}$$

   (vybereme kteří mají zájem a násobíme pravděpodobností, že mají zájem / nezájem).

   (b) Určete pravděpodobnostní funkci pro Poissonovskou aproximaci tohoto počtu.

   Řešení: Poissonovsky aproximujeme tak, aby $\lambda = np$ (aby ta rozdělení měla stejnou střední hodnotu), tedy $Y \sim Pois(np)$:

$$p_Y(k) = \Pr[Y = k] = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{(np)^k e^{-np}}{k!} = \frac{10^k e^{-10}}{k!}$$

   (c) Jaká je pravděpodobnost, že 20 kopií knihy nestačí? Vyjádřete jednak pomocí distribuční funkce, jednak pomocí sumy. A také jednak pomocí přesné formule z části (a), jednak pomocí aproximace z části (b).

   Řešení: Pomocí distribuční funkce $1 - F_X(20)$.

   Pomocí sumy:

   - Přesně pomocí wolframalpha

$$\sum_{j=21}^{1000} \binom{1000}{j} 0.01^j (1-0.01)^{1000-j} \doteq 0.00149$$

   - Pomocí aproximace

$$\sum_{j=21}^{\infty} \frac{10^k e^{-10}}{k!} \doteq 0.00159$$

   (d) Je popsaný model zájmu studentů o knihy realistický?

   Řešení: Nezávislost může být dost nerealistická. “Hele ta knížka je fakt dobrá.”

5. ***Exponenciální rozdělení*** **je spojitou analogií rozdělení geometrického. Vyjadřuje dobu čekání na první událost generovanou poissonovským procesem (s daným parametrem λ). Náhodná veličina $X \sim Exp(\lambda)$ má distribuční funkci**

$$F_X(t) = \begin{cases} 1 - e^{-\lambda t} & \textbf{pro } t \geq 0, \\ 0 & \textbf{pro } t < 0. \end{cases}$$

Vypočítejte:

(a) hustotní funkci $f_X(t)$

Řešení: Je to jen derivace $F_X(t)$ podle t:

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

(b) střední hodnotu $\mathbb{E}[X]$

Řešení: Je jednodušší napřed spočítat

$$\int xe^{-x}\ dx = -e^x(x+1) + C$$

pak použít substituci:

$$\begin{aligned} \int_a^b f(\varphi(x))\varphi'(x)\ dx &= \int_{\varphi(a)}^{\varphi(b)} f(u)\ du \\ \mathbb{E}[X] &= \int_{-\infty}^{\infty} t f_X(t)\ dt \\ &= \int_0^{\infty} t\lambda e^{-\lambda t}\ dt \\ &= \left[-e^{-u}(u+1)/\lambda\right]_0^{\infty} \\ &= 1/\lambda \end{aligned}$$

(c) rozptyl $\mathrm{var}(X)$

Řešení:

$$\mathrm{var}(X) = 1/\lambda^2$$

3.8 Cvičení

1.

(a) Nechť X je náhodná veličina a $X \geq 0$ skoro jistě (tzn $\Pr[X \geq 0] = 1$). Najděte nějakou takovou náhodnou veličinu, která je netriviální.

Řešení: Například $\mathrm{Im}(X) = (\mathbb{Q} \cap [-1, 0]) \cup [0, 1]$ kde volíme uniformně náhodně (pravděpodobnost podmnožiny je její míra).

(b) Nechť X je náhodná veličina a $X \geq 0$ skoro jistě (tzn $\Pr[X \geq 0] = 1$). Dokažte, že pokud $\mathbb{E}[X]$ existuje, tak $\mathbb{E}[X] \geq 0$.

Řešení: S výhodou využijeme naší obecnou definici:

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty 1 - F_X(t)\ dt - \int_{-\infty}^0 F_X(t)\ dt \\ &= \int_0^\infty \Pr[X > t]\ dt - \int_{-\infty}^0 \Pr[X \leq t]\ dt \\ &= \int_0^\infty \Pr[X > t]\ dt - \int_{-\infty}^0 0\ dt \\ &\geq 0\end{aligned}$$

(c) Nechť Y, Z jsou náhodné veličiny a $Y \leq Z$ skoro jistě. Dokažte, že pokud $\mathbb{E}[Y]$, $\mathbb{E}[Z]$ existují, tak $\mathbb{E}[Y] \leq \mathbb{E}[Z]$.

Řešení: Opět pouze využijeme toho, že pro každé $t \in \mathbb{R}$ platí $F_Y(t) \geq F_Z(t)$ (rozmyslete, že to je mnohem slabší podmínka než $\Pr[Y \leq Z] = 1$).

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^\infty 1 - F_Y(t)\ dt - \int_{-\infty}^0 F_Y(t)\ dt \\ &\leq \int_0^\infty 1 - F_Z(t)\ dt - \int_{-\infty}^0 F_Z(t)\ dt \\ &= \mathbb{E}[Z]\end{aligned}$$

(d) Dokažte Markovovu nerovnost $\Pr[X \geq a\mathbb{E}[X]] \leq 1/a$ pro nezápornou náhodnou veličinu X a libovolné $a \geq 1$.

Řešení: Prvně si rozmyslete následující rovnost:

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty \Pr[X \geq t]\ dt && (X \text{ je nezáporná}) \\ &= \int_0^\infty 1 - F_X(t)\ dt\end{aligned}$$

Nyní jen uděláme odhad:

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty \Pr[X \geq t]\ dt && (X \text{ je nezáporná}) \\ &\geq \int_0^A \Pr[X \geq t]\ dt && (\text{pravděpodobnost je nezáporná}) \\ &\geq \int_0^A \Pr[X \geq A]\ dt\end{aligned}$$

$$= A \Pr[X \geq A]$$

Dosadíme

$$\begin{aligned} A &= a\mathbb{E}[X] \\ \mathbb{E}[X] &\geq A \Pr[X \geq A] \\ \mathbb{E}[X] &\geq a\mathbb{E}[X] \Pr[X \geq a\mathbb{E}[X]] \\ 1/a &\geq \Pr[X \geq a\mathbb{E}[X]] \end{aligned}$$

Kdyby náhodou $\mathbb{E}[X] = 0$, pak stejně $\Pr[X = 0] = 1$.

2. **Bublifukem vyfoukneme bublinu o poloměru $R \sim U(1,5)$. Jaká je střední hodnota povrchu bubliny?**

   Řešení: Víme, že pokud poloměr je R, pak povrch bubliny bude $4\pi R^2$, pak můžeme použít větu z přednášky

   $$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(t) f_X(t)\ dt$$

   Poloměr je vybírán z uniformní distribuce, tedy $f_X(t) = 1/4$ pro libovolné $t \in [1,5]$ (nula jinak). Dosazením

   $$\begin{aligned}
   \mathbb{E}[4\pi R^2] &= \int_1^5 4\pi t^2/4\ dt \\
   &= \pi \int_1^5 t^2\ dt \\
   &= \pi \int_1^5 t^2\ dt \\
   &= \pi \left[t^3/3\right]_1^5 \\
   &= \pi \left(125 - 1\right)/3 \\
   &= \pi 124/3
   \end{aligned}$$

```python
import math
import random

def R():
    return 1 + 4 * random.random()

def Area(r):
    return 4 * math.pi * r**2

N = 1_000_000
EA = sum(Area(R()) for _ in range(N)) / N
print(f'E[povrch bubliny] = {EA} (={math.pi * 124 / 3})')

# Možný výstup:
# E[povrch bubliny] = 129.79361145575925 (=129.8524963483781)
```

3. **Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením se střední hodnotou μ a rozptylem σ^2. Označme $S_n = (X_1 + \cdots + X_n)/n$. To můžeme považovat za odhad střední hodnoty μ průměrem z n pokusů.**

   (a) Určete $\mathbb{E}[S_n]$ a $\mathrm{var}(S_n)$.

   Řešení: Dle linearity střední hodnoty máme

   $$\mathbb{E}\left[\sum_{j=1}^{n} X_j/n\right] = \mu$$

   Dle domácího úkolu z toho, že jsou proměnné nezávislé máme (matematickou indukcí)

   $$\mathrm{var}\left(\sum_{j=1}^{n} X_j/n\right) = (n\sigma^2)/n^2 = \sigma^2/n$$

   Tedy se nám nezměnila střední hodnota, ale celkem značně se zmenšil rozptyl (připomeňme, že rozptyl nám intuitivně říká jak daleko je náhodná proměnná od své střední hodnoty $\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$). Takže intuitivně více pokusů nám dá přesnější odhad.

   (b) Ukažte, jak lze počítat S_n z S_{n-1}, X_n a n.

   Řešení: Připomeňme, že počítače používají často reprezentaci čísel s omezeným maximem. Takže kdybychom jen sečetli a pak dělili, tak součet může přetéct. Navíc i desetinná čísla jsou omezená a jejich sčítáním vznikají chyby.

   - Můžeme použít algoritmus z knihy The Art of Computer Programming (Knuth):

     ```
     mean = 0.0
     n = 0

     for x in data:
         n += 1
         mean += (x - mean) / n
     ```

     Tato metoda se používá v GNU Scientific Library `https://savannah.gnu.org/git/?group=gsl` od roku 1998 (commit c91e4ff0dd04766f45cc899467b46a83ad06bd5d) neustále (kontrolováno duben 2021). Zde se využívá následujícího vztahu:

     $$mean(n) = mean(n-1) + (data[n] - mean(n-1))/(n+1)$$

     Velikou výhodou je, že tady nám nepřeteče mezivýsledek (pozor výsledný výsledek přetéct může).

   - Pokud máme čísla daná předem, tak je můžeme setřídit a sčítat od nejmenšího (to překvapivě dost pomůže).

   - Pokud nás zajímá jen součet (ne průměr), pak můžeme použít Kahan summation algorithm `https://en.wikipedia.org/wiki/Kahan_summation_algorithm` Mimochodem pokud se kompilátor může využívat asosiativitu a komutativitu pak může dost pokazit takovéto numerické algoritmy, proto když kompilujete takový software, tak volby typu `-ffast-math` člověk chce zapnout jen když ví co dělá.

   (c) Použijte vhodné X_i, aby μ obsahovalo číslo π. Sestavte program v libovolném jazyce a spočítejte pomocí něj hodnotu π. (Jak velké n myslíte, že bude potřeba pro pět správných číslic?)

   Řešení: Můžeme použít známé nerovnosti kde X je počet bodů uvnitř čtvrt-kruhu z celkem N hodů:

- Markovova nerovnost: Nechť X je nezáporná náhodná veličina, pak $\Pr[X \geq a\mathbb{E}[X]] \leq 1/a$ pro $a \geq 1$.
  i. Důkaz viz jedno z předchozích cvičení.
  ii. Použití: Například pro $a = 2$ dostáváme, že $\Pr[X > 2N\pi/4] \leq 1/2$.
- Čebyševovu nerovnost: Nechť X je nezáporná náhodná veličina s rozptylem $\mathrm{var}(X)$, pak $\Pr[|X - \mathbb{E}[X]| \geq a] \leq \mathrm{var}(X)/a^2$ pro $a > 0$.
  i. Důkaz: Použijeme Markova na náhodnou veličinu $(X - \mathbb{E}[X])^2$ (ta je z definice rozptylu):

$$\begin{aligned} \Pr\left[(X - \mathbb{E}[X])^2 \geq A^2 \,\mathrm{var}(X)\right] &\leq 1/A^2 \\ a = A\sqrt{\mathrm{var}(X)} \\ \Pr\left[(X - \mathbb{E}[X])^2 \geq a^2\right] &\leq \mathrm{var}(X)/a^2 \end{aligned}$$

  ii. Použití: Pro $a = 0.1$ ještě potřebujeme znát rozptyl, který je $\pi(1 - \pi/4)/4N$.

$$\Pr\left[|X - \mathbb{E}[X]| \geq 0.1\right] \leq 100\pi(1 - \pi/4)/4N$$

- Černovovu nerovnost: Nechť $X = \sum_{j=1}^{n} X_j$ kde $X_j \in [0, 1]$ jsou nezávislé proměnné. Nechť $\mu = \mathbb{E}[X]$ a volme $\delta \in (0, 1)$. Pak

$$\begin{aligned} \Pr[X \geq (1 + \delta)\mu] &\leq e^{-\delta^2\mu/3} \\ \Pr[X \leq (1 - \delta)\mu] &\leq e^{-\delta^2\mu/2} \end{aligned}$$

  i. Důkaz: použití Markovovy nerovnosti na $\Pr[e^{tX} \geq e^{ta}] \leq \mathbb{E}[e^{tX}]/e^{ta}$ a optimalizováním přes t. Prozatím vynecháme, možná bude na přednášce nebo na nějakém příštím cvičení.
  ii. Použití: Například $\delta = 0.1$ dostáváme

$$\begin{aligned} \Pr[X \geq 1.1N\pi/4] &\leq e^{-0.01N\pi/12} \\ \Pr[X \leq 0.9N\pi/4] &\leq e^{-0.01N\pi/8} \end{aligned}$$

4. **Předpokládejme, že u poštovní přepážky trvá vyřízení jednoho zákazníka čas, který má exponenciální rozdělení a střední hodnotu 4 minuty.**

   (a) Jaký je parametr λ, jaká je distribuční funkce?

   Řešení: Náhodná veličina $X \sim Exp(\lambda)$ má distribuční funkci

   $$F_X(t) = \begin{cases} 1 - e^{-\lambda t} & \text{pro } t \geq 0, \\ 0 & \text{pro } t < 0. \end{cases}$$

   a střední hodnotu $1/\lambda$.

   Tedy vidíme, že v našem případě aby nám seděla střední hodnota, musí být $\lambda = 1/4$.

   (b) Jaká je pravděpodobnost, že budeme čekat více než 4 minuty?

   Řešení: Známe distribuční funkci, tedy

   $$\begin{aligned} \Pr[\text{více než 4 minuty}] &= 1 - F_X(4) \\ &= 1 - \left(1 - e^{-\frac{1}{4}4}\right) \\ &= e^{-1} \\ &\doteq 0.3678 \end{aligned}$$

   (c) Jaká je pravděpodobnost, že budeme čekat něco mezi 3 a 5 minutami?

   Řešení: Opět bychom mohli integrovat hustotu, ale otázkou je proč, když můžeme dosadit do distribuční funkce:

   $$\begin{aligned} \Pr[\text{mezi 3 a 5 minutami}] &= F_X(5) - F_X(3) \\ &= \left(1 - e^{-\frac{1}{4}5}\right) - \left(1 - e^{-\frac{1}{4}3}\right) \\ &\doteq 0.1858 \end{aligned}$$

   (d) Simulujte.

   Řešení:

```
import math
from random import random
from numpy.random import exponential

def npExp(l):
    # np.random.exponential takes scale = 1 / lambda
    return exponential(1 / l)

def Q_Exp(p, l):
    # Quantile = inverse of distribution function
    return - math.log(1 - p) / l

def myExp(l):
    # my implementation of Exp using the quantile
    return Q_Exp(random(), l)

N = 1_000_000
```

```
l = 1 / 4

np_Pr_X_geq_4 = sum(int(npExp(l) >= 4) for _ in range(N)) / N
print(f'b) Pr[X>=4] = {np_Pr_X_geq_4} (={math.exp(-1)})')
my_Pr_X_geq_4 = sum(int(myExp(l) >= 4) for _ in range(N)) / N
print(f'b) Pr[X>=4] = {my_Pr_X_geq_4} (={math.exp(-1)})')

np_Pr_3_leq_X_leq_5 = sum(int(3 <= npExp(l) <= 5) for _ in range(N)) / N
print(f'c) Pr[3<=X<=5] = {np_Pr_3_leq_X_leq_5} (={math.exp(-3/4) - math.exp(-5/4)})')
my_Pr_3_leq_X_leq_5 = sum(int(3 <= myExp(l) <= 5) for _ in range(N)) / N
print(f'c) Pr[3<=X<=5] = {my_Pr_3_leq_X_leq_5} (={math.exp(-3/4) - math.exp(-5/4)})')

# Možný výstup:
# b) Pr[X>=4] = 0.368577 (=0.36787944117144233)
# b) Pr[X>=4] = 0.367661 (=0.36787944117144233)
# c) Pr[3<=X<=5] = 0.185306 (=0.1858617558808246)
# c) Pr[3<=X<=5] = 0.186097 (=0.1858617558808246)
```

5. **Říkáme, že náhodná veličina X (resp. její rozdělení) *nemá paměť*, pokud**

$$\Pr[X > s + t \mid X > s] = \Pr[X > t]$$

pro $s, t \geq 0$. Jinými slovy, doba, kterou jsme již čekali, nemá vliv na dobu, kterou budeme ještě čekat.

(a) Ukažte, že geometrické rozdělení nemá paměť.

Řešení: Mějme nějaká fixní čísla $s, t \geq 0$. Připomeňme, že pro číslo s následující je jev $X \geq s = \{\omega \in \Omega \mid X(\omega) \geq s\}$. Podmíněnou pravděpodobnost pro jevy si jistě pamatujeme

$$\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]} \qquad \text{(pokud } \Pr[B] > 0\text{)}$$

Připomeňme ještě

$$\Pr[X = n] = (1-p)^{n-1}p \qquad (n = 1, 2, \ldots, X \sim Geom(p))$$

Budeme využívat následující součet geometrické řady:

$$\sum_{n=T}^{\infty} (1-p)^{n-1}p = (1-p)^{T-1} \qquad \text{(pro } T \in \mathbb{N}\text{)}$$

Víme, že $\Pr[X > t] > 0$ pro libovolné $t \in \mathbb{R}$ pokud $X \sim Geom(p)$. Pro jednoduchost nechť $s, t \in \mathbb{N}$ (ať nemusíme brát horní celé části).

$$\begin{aligned}
\Pr[X > s + t \mid X > s] &= \frac{\Pr[X > s + t \cap X > s]}{\Pr[X > s]} \\
&= \frac{\Pr[X > s + t]}{\Pr[X > s]} \qquad (s \geq 0, t \geq 0) \\
&= \frac{\sum_{n=s+t+1}^{\infty}(1-p)^{n-1}p}{\sum_{n=s+1}^{\infty}(1-p)^{n-1}p} \\
&= \frac{(1-p)^{s+t}}{(1-p)^{s}} \\
&= (1-p)^{t} \\
&= \sum_{n=t+1}^{\infty}(1-p)^{n-1}p \\
&= \Pr[X > t]
\end{aligned}$$

(b) Co z toho plyne o rozložení dalšího hodu, když už nám pětkrát v řadě padla hlava?

Řešení: Pokud máme spravedlivou minci, tak jsou oba výsledky stejně pravděpodobné. Ta mince v sobě nemá zabudované počítátko. Přesto člověk intuitivně čeká, "že už to musí padnout." Tohle bývá velký problém pro notorické hráče.

(c) Ukažte, že exponenciální rozdělení nemá paměť.

Řešení: Budeme postupovat obdobně, jen připomeňme distribuční funkci $\Pr[X \leq t] = F_X(t) = 1 - e^{-\lambda t}$ pro libovolné $t \geq 0$ a $X \sim Exp(\lambda)$.

$$\Pr[X > s + t \mid X > s] = \frac{\Pr[X > s + t \cap X > s]}{\Pr[X > s]}$$

$$
\begin{aligned}
&= \frac{\Pr[X > s+t]}{\Pr[X > s]} && (s \geq 0, t \geq 0) \\
&= \frac{1 - F_X(s+t)}{1 - F_X(s)} \\
&= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\
&= e^{-\lambda t} \\
&= 1 - F_X(t) \\
&= \Pr[X > t]
\end{aligned}
$$

(d) Simulujte.

Řešení:

```
import math
from numpy import count_nonzero
from numpy.random import exponential
from numpy.random import geometric

l = 1 / 7
p = 0.1
s = 5
t = 20

N = 100_000_000

X = geometric(p, N)

pr_X_g_t = count_nonzero(X > t) / N
print(f'Pr[X > {t}] = {pr_X_g_t} (={(1-p)**t})')
pr_X_podminena = count_nonzero(X > s+t) / count_nonzero(X > s)
print(f'Pr[X > {s+t} | X > {s}] = {pr_X_podminena} (={(1-p)**t})')

Y = exponential(1 / l, N)

pr_Y_g_t = count_nonzero(Y > t) / N
print(f'Pr[Y > {t}] = {pr_Y_g_t} (={math.exp(-l*t)})')
pr_Y_podminena = count_nonzero(Y > s+t) / count_nonzero(Y > s)
print(f'Pr[Y > {s+t} | Y > {s}] = {pr_Y_podminena} (={math.exp(-l*t)})')

# Možný výstup:
# Pr[X > 20] = 0.12162167 (=0.12157665459056935)
# Pr[X > 25 | X > 5] = 0.12153812273722872 (=0.12157665459056935)
# Pr[Y > 20] = 0.05740661 (=0.05743261926761737)
# Pr[Y > 25 | Y > 5] = 0.05737426975802089 (=0.05743261926761737)
```

Platí dokonce, že je to jediné spojité rozdělení na kladných čísel bez paměti (a geometrické je jediné diskrétní bez paměti), ale to dokazovat nemusíte.

6. **Budeme modelovat množství sněhu, který bude na Silvestra v lyžarském areálu Ještěd, pomocí normálního rozdělení se střední hodnotou 40 (centimetrů) a směrodatnou odchylkou 10.**

   (a) Jaká je pravděpodobnost, že nám model určí zápornou hodnotu sněhové pokrývky?

   Řešení: Distribuční funkci neumíme vyjádřit pomocí jednoduchých funkcí (polynomy, sin, cos, exponenciala...). Ale umíme ji aproximovat, případně se můžeme rovnou podívat do tabulky. Tabulky jsou uvedené pro $Z \sim N(0,1)$, takže naši $X \sim N(\mu, \sigma^2)$ musíme nějak převédst:

   $$Z = \frac{X - \mu}{\sigma}$$

   kde klasicky μ je střední hodnota X ($\mathbb{E}[X] = \mu$) a σ je směrodatná odchylka, což je odmocnina rozptylu (tedy $\operatorname{var}(X) = \sigma^2$).

   Konkrétně máme $X \sim N(40, 10)$ takže $Z = \frac{X-40}{10}$.

   Otázka zní:

   $$\begin{aligned}\Pr[X < 0] &= \Pr\left[\frac{X - 40}{10} < \frac{0 - 40}{10}\right] \\ &= \Pr\left[Z < -4\right] \\ &\doteq 0.00003\end{aligned}$$

   (b) Jaká je pravděpodobnost, že sněhu napadne 50–70 cm?

   Řešení:

   $$\begin{aligned}\Pr[50 \leq X \leq 70] &= \Pr\left[\frac{50 - 40}{10} \leq \frac{X - 40}{10} \leq \frac{70 - 40}{10}\right] \\ &= \Pr\left[1 \leq Z \leq 3\right] \\ &= \Pr\left[Z \leq 3\right] - \Pr\left[Z \leq 1\right] \\ &\doteq 0.99865 - 0.84134 \\ &= 0.15731\end{aligned}$$

   (c) Simulujte.

   Řešení:

```python
from numpy import count_nonzero
from numpy.random import normal
import scipy.stats as st

mu = 40
sigma = 10

N = 100_000_000

X = normal(mu, sigma, N)

pr_negative = count_nonzero(X < 0) / N
print(f'Pr[X < 0] = {pr_negative} (={st.norm.cdf((0 - mu) / sigma)})')
```

```
pr_50_70 = count_nonzero((50 <= X) * (X <= 70)) / N
pr_50_70_exact = st.norm.cdf((70-mu)/sigma)-st.norm.cdf((50-mu)/sigma)
print(f'Pr[50 <= X <= 70] = {pr_50_70} (={pr_50_70_exact})')

# Možný výstup:
# Pr[X < 0] = 3.2e-05 (=3.167124183311986e-05)
# Pr[50 <= X <= 70] = 0.15723717 (=0.15730535589982697)
```

Hodnoty $\Phi(x)$ si spočítejte v Pythonu nebo v R, případně se podívejte do tabulky na `https://en.wikipedia.org/wiki/Standard_normal_table` **(sekce Cummulative).**

7. **Plutonium-238 má poločas rozpadu 87.7 let. Jeho rozpad budeme modelovat pomocí exponenciálního rozdělení: pro každý atom budeme čas, za který se rozpadne, považovat za nezávislou náhodnou veličinu s rozdělením $Exp(\lambda)$.**

   (a) Jaké je λ?

   Řešení: Poločas rozpadu je definován jako doba, za kdy se rozpadne polovina atomů. Atomů ve vzorku bývá opravdu hodně. Takže ekvivalentně pravděpodobnost, že se konkrétní atom za tuto dobu rozpadne je polovina. Chceme tedy:

$$\begin{aligned} 0.5 &= F_X(87.7) \\ 0.5 &= 1 - e^{-\lambda 87.7} \\ 0.5 &= e^{-\lambda 87.7} \\ \ln(0.5) &= -\lambda 87.7 \\ \ln(2)/87.7 &= \lambda \end{aligned}$$

   (b) Jaká je střední doba života atomu ^{238}Pu?

   Řešení: Přímo střední hodnota, o které už z dřívějška víme, že je $1/\lambda$. Tedy střední doba života atomu je:

$$87.7/\ln(2) \doteq 126.524 \text{ let}$$

   (c) Po jaké době se rozpadne 90 % atomů?

   Řešení: Atomů ve vzorku bývá opravdu hodně. Takže můžeme počítat za jakou dobu má atom pravděpodobnost rozpadu 90 %. Takže potřebujeme kvantil Q_X:

$$\begin{aligned} Q_X(p) &= \frac{-\ln(1-p)}{\lambda} \\ Q_X(0.9) &= \frac{-\ln(1-0.9)}{\ln(2)/87.7} \\ Q_X(0.9) &= \frac{\ln(10)}{\ln(2)/87.7} \\ Q_X(0.9) &= 291.333 \text{ let} \end{aligned}$$

   (d) Kolik procent atomů se rozpadne po 50 letech? (Mimochodem, některé kosmické sondy a některé kardiostimulátory používají rozpad ^{238}Pu jako zdroj energie.)

   Řešení: Opět předpokládáme, že podíl rozpadlých bude velmi blízký pravděpodobnosti, že se jeden konkrétní rozpadne za 50 let (atomů je hodně).

$$\begin{aligned} F_X(87.7) &= 1 - e^{-\lambda t} \\ F_X(87.7) &= 1 - e^{-50\ln(2)/87.7} \\ &\doteq 0.3264 \end{aligned}$$

   (e) Simulujte.

   Řešení:

```python
import math
import numpy as np

N = 1_000_000
l = math.log(2) / 87.7

X = np.random.exponential(1 / l, N)

EX = X.sum() / N
print(f'b) E[doba zivota] = {EX} (={87.7 / math.log(2)})')

rozpad90 = np.sort(X)[int(0.9 * N)]
print(f'c) kdy 90% rozpadlych = {rozpad90} (={87.7 * math.log(10) / math.log(2)})')

rozpadu50 = np.count_nonzero(X <= 50) / N
rozpadu50_presne = 1 - math.exp(-50 * math.log(2) / 87.7)
print(f'd) # rozpadu po 50 letech = {rozpadu50} (={rozpadu50_presne})')

# Možný výstup:
# b) E[doba zivota] = 126.61914635519292 (=126.5243550859621)
# c) kdy 90% rozpadlych = 291.4057716079031 (=291.3330939216217)
# d) # rozpadu po 50 letech = 0.326249 (=0.32644177311899625)
```

8. **Dostali jsme minci. Nevíme jestli je spravedlivá (tzn neznáme pravděpodobnost, že padne hlava). Tisíckrát jsme s ní hodili a padlo 345 hlav.**

   (a) Jaká je pravděpodobnost, že na spravedlivé minci padne přesně 345 hlav?

   Řešení: Binomické rozdělení už důvěrně známe:

$$X \sim Bin(n, p)$$
$$\Pr[X = j] = \binom{n}{j} p^j (1-p)^{n-j}$$

   Konkrétně:

$$X \sim Bin(1000, 0.5)$$
$$\Pr[X = 345] = \binom{1000}{345} 0.5^{1000}$$
$$\doteq 1.6148 \cdot 10^{-23}$$

   (b) Pokud $p \in (0, 1)$ je proměnná vyjadřující pravděpodobnost, že na naší minci padne hlava. Vyjádřete pravděpodobnost, že padne 345 hlav jako funkci p, tedy

$$P(p) = \Pr[X = 345 \textbf{ kde } X \sim Bin(1000, p)]$$

   Řešení:

$$P(p) = \binom{1000}{345} p^{345} (1-p)^{1000-345}$$
$$= \binom{1000}{345} p^{345} (1-p)^{655}$$

   (c) Pokud tedy modelujeme hod naší mincí jako $Bin(1000, p)$, pak určete p, které dává nejvyšší pravděpodobnost pozorovaného výsledku.

   Řešení: Můžeme zderivovat a položit derivaci rovnou nule (a pak zkontrolovat druhou derivaci, abychom viděli, že je funkce konkávní). Kombinační číslo je jen konstanta, tu opisovat nebudu:

$$\left(p^{345}(1-p)^{655}\right)' = 345p^{344}(1-p)^{655} - 655p^{345}(1-p)^{654}$$
(derivace součinu a derivace složené funkce)

   Tohle se nezdá být moc nadějné.

   Mohli bychom maximalizovat logaritmus toho výrazu, pak bychom neměli tak ošklivou derivaci:

$$0 = \left(\ln(p^{345}(1-p)^{655})\right)'$$
$$= (345\ln(p) + 655\ln((1-p)))'$$
$$= 345/p + 655/(1-p)$$

   Využijeme toho, že $p \in (0, 1)$, takže můžeme násobit obě strany

$$0 = 345/p + 655/(1-p)$$

$$
\begin{aligned}
0 &= 345(1-p) + 655p \\
0 &= 345 - 345p + 655p \\
0 &= 345 - 1000p \\
p &= 345/1000
\end{aligned}
$$

Pro jistotu si ještě rozmyslete, že ta funkce je konkávní.

(d) Porovnejte s tím, jak jsme doteď simulovali.

Řešení: Výsledek je stejný, pravděpodobnost hlavy bychom odhadli v simulaci jako $345/1000 = 0.345$.

3.9 Cvičení

1. **Dostali jsme minci. Nevíme jestli je spravedlivá (tzn neznáme pravděpodobnost, že padne hlava). Tisíckrát jsme s ní hodili a padlo 345 hlav.**

   (a) Jaká je pravděpodobnost, že na spravedlivé minci padne přesně 345 hlav?

   Řešení: Binomické rozdělení už důvěrně známe:

   $$X \sim Bin(n, p)$$
   $$\Pr[X = j] = \binom{n}{j} p^j (1-p)^{n-j}$$

   Konkrétně:

   $$X \sim Bin(1000, 0.5)$$
   $$\Pr[X = 345] = \binom{1000}{345} 0.5^{1000}$$
   $$\doteq 1.6148 \cdot 10^{-23}$$

   (b) Pokud $p \in (0, 1)$ je proměnná vyjadřující pravděpodobnost, že na naší minci padne hlava. Vyjádřete pravděpodobnost, že padne 345 hlav jako funkci p, tedy

   $$P(p) = \Pr[X = 345 \textbf{ kde } X \sim Bin(1000, p)]$$

   Řešení:

   $$P(p) = \binom{1000}{345} p^{345} (1-p)^{1000-345}$$
   $$= \binom{1000}{345} p^{345} (1-p)^{655}$$

   (c) Pokud tedy modelujeme hod naší mincí jako $Bin(1000, p)$, pak určete p, které dává nejvyšší pravděpodobnost pozorovaného výsledku. Tomuto se ve statistice říká Maximum Likelihood Estimation (MLE).

   Řešení: Můžeme zderivovat a položit derivaci rovnou nule (a pak zkontrolovat druhou derivaci, abychom viděli, že je funkce konkávní). Kombinační číslo je jen konstanta, tu opisovat nebudu:

   $$\left(p^{345}(1-p)^{655}\right)' = 345p^{344}(1-p)^{655} - 655p^{345}(1-p)^{654}$$
   (derivace součinu a derivace složené funkce)
   $$= (1-p)^{654} p^{344} \left(345(1-p) - 655p\right)$$

   což je nulové pokud $p = 0$ nebo $p = 1$ nebo

   $$0 = (345(1-p) - 655p)$$
   $$0 = 345 - 345p - 655p$$
   $$p = 345/1000$$

   Alternativní postup by bylo maximalizovat logaritmus toho výrazu, pak bychom neměli tak ošklivou derivaci:

   $$0 = \left(\ln(p^{345}(1-p)^{655})\right)'$$

$$
\begin{aligned}
&= (345\ln(p) + 655\ln((1-p)))' \\
&= 345/p + 655/(1-p)
\end{aligned}
$$

Využijeme toho, že $p \in (0,1)$, takže můžeme násobit obě strany

$$
\begin{aligned}
0 &= 345/p + 655/(1-p) \\
0 &= 345(1-p) + 655p \\
0 &= 345 - 345p + 655p \\
0 &= 345 - 1000p \\
p &= 345/1000
\end{aligned}
$$

Pro jistotu si ještě rozmyslete, že ta funkce je konkávní.

(d) Porovnejte s tím, jak jsme doteď simulovali.

Řešení: Výsledek je stejný, pravděpodobnost hlavy bychom odhadli v simulaci jako $345/1000 = 0.345$.

2. **Za druhé světové války spojenci potřebovali vědět, kolik tanků Německo vyrobilo. Němci byli precizní a své tanky číslovali popořadě (pokud vyrobili n-tý tank, tak měl na motoru napsané číslo n).**

   (a) **Zajali jste uniformně náhodný tank, který měl číslo m (připomeňme, že neznáte n), jak odhadnete $\hat{n}$ (tj. odhad skutečného počtu n) pomocí MLE?**

   Řešení: Chceme maximalizovat pravděpodobnost, že jsme pozorovali číslo m za podmínky, že skutečný počet tanků je n, tedy $X \sim U(n)$.

   $$\Pr[X = m] = 1/n \qquad \text{(pro } n \geq m\text{)}$$

   Takže počet tanků odhadneme jako $\hat{n} = m$.

   (b) **Je to, co nám vyšlo rozumný odhad?**

   Řešení: Intuitivně tento odhad dost podhodnocuje. Speciálně tento odhad dává vždy nejmenší možný počet, který je v souladu s naším pozorováním.

   To nám neříká, že MLE je špatná metoda! Jen to říká, že na tento problém se zas tak moc nehodí. Na tento problém se hodí lépe jiné metody (o některých se nejspíš budeme učit). A hlavně to, že nelze jen slepě dosadit do vzorečku, kterému nerozumíme.

   Další věc je, že nejsme tak daleko od skutečného odhadu (zhruba vynásobení dvěma nám dá celkem rozumný odhad).

   (c) **Je možné, že se k tomuto problému ještě vrátíme s jinými statistickými odhady. Pokud jste příliš zvědaví, tak** `https://en.wikipedia.org/wiki/German_tank_problem`

3. **Nechť $Z \sim N(0,1)$. Pomocí tabulky funkce Φ ověřte pravidlo 3σ, neboli spočtěte**

x	-4	-3	-2	-1	0	1	2	3	4
$\Phi(x)$	0.00003	0.00135	0.02275	0.15866	0.500000	0.84135	0.97725	0.99865	0.99997

Další hodnoty viz `https://en.wikipedia.org/wiki/Standard_normal_table` **– sekce Cumulative nebo použitím** `scipy.stats.norm.cdf` `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html`

(a) Připomeňte pravidlo 3σ.

Řešení: Pro normální rozdělení: `https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule` Nebo také pravidlo 68-95-99.7 (pravděpodobnost, že jsme nejvýš 1-2-3 násobek sigma od střední hodnoty).

(b) $\Pr[|Z| \leq 1]$

Řešení: Tabulka nám udává cdf tedy distribuční funkci $F_X(t) = \Pr[X \leq t]$. Můžeme tedy psát:

$$\begin{aligned}\Pr[|Z| \leq 1] &= \Pr[-1 \leq Z \leq 1] \\ &= F_X(1) - F_X(-1) \\ &= 0.6826\end{aligned}$$

(c) $\Pr[|Z| \leq 2]$

Řešení: Tabulka nám udává cdf tedy distribuční funkci $F_X(t) = \Pr[X \leq t]$. Můžeme tedy psát:

$$\begin{aligned}\Pr[|Z| \leq 2] &= \Pr[-2 \leq Z \leq 2] \\ &= F_X(2) - F_X(-2) \\ &= 0.9545\end{aligned}$$

(d) $\Pr[|Z| \leq 3]$

Řešení: Tabulka nám udává cdf tedy distribuční funkci $F_X(t) = \Pr[X \leq t]$. Můžeme tedy psát:

$$\begin{aligned}\Pr[|Z| \leq 3] &= \Pr[-3 \leq Z \leq 3] \\ &= F_X(3) - F_X(-3) \\ &= 0.9973\end{aligned}$$

(e) Spočítejte to programem:

Řešení:

```
import scipy.stats as st

print(f'Pr[|Z| <= 1] = {st.norm.cdf(1) - st.norm.cdf(-1)}')
print(f'Pr[|Z| <= 2] = {st.norm.cdf(2) - st.norm.cdf(-2)}')
print(f'Pr[|Z| <= 3] = {st.norm.cdf(3) - st.norm.cdf(-3)}')

# Výstup:
# Pr[|Z| <= 1] = 0.6826894921370859
# Pr[|Z| <= 2] = 0.9544997361036416
# Pr[|Z| <= 3] = 0.9973002039367398
```

(f) Přepište, co to znamená pro n.v. $X \sim N(\mu, \sigma^2)$

Řešení: Pokud $X \sim N(\mu, \sigma^2)$ a položíme

$$Z = \frac{X - \mu}{\sigma}$$

pak $Z \sim N(0, 1)$. Jinak řečeno

$$\begin{aligned}
\Pr[|X - \mu| \geq t\sigma] &= \Pr\left[|(\sigma Z + \mu) - \mu| \geq t\sigma\right] \\
&= \Pr\left[|\sigma Z| \geq t\sigma\right] \\
&= \Pr\left[\sigma\,|Z| \geq t\sigma\right] && (\sigma > 0) \\
&= \Pr\left[|Z| \geq t\right] && (\sigma > 0) \\
&= F_X(t) - F_X(-t)
\end{aligned}$$

4. **Nechť X, Y mají sdruženou hustotu $f_{X,Y}(x,y) = e^{-x-y}$ pro $x, y > 0$ (a 0 jinak).**

   (a) Určete marginální hustoty f_X, f_Y.

   Řešení: Dle přednášky máme:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy$$

   tedy píšeme (pro nezáporné x):

$$\begin{aligned} f_X(x) &= \int_0^{\infty} e^{-x-y}\ dy \\ &= \int_0^{\infty} e^{-x}e^{-y}\ dy \\ &= e^{-x}\int_0^{\infty} e^{-y}\ dy \\ &= e^{-x}\lim_{b\to\infty}\int_0^{b} e^{-y}\ dy \\ &= e^{-x}\lim_{b\to\infty}\left[-e^{-y}\right]_0^b \\ &= e^{-x}\lim_{b\to\infty}\left(-e^{-b} - -e^{-0}\right) \\ &= e^{-x} \end{aligned}$$

   Obdobně $f_Y(y) = e^{-y}$ (pro nezáporné y).

   (b) Určete také distribuční funkce F_X, F_Y, $F_{X,Y}$.

   Řešení:

$$\begin{aligned} F_X &: \mathbb{R} \to [0,1] \\ F_X(t) &= \Pr[X \le t] \\ F_X(t) &= 0 && \text{(pokud } t \le 0\text{)} \\ F_X(t) &= \int_0^t e^{-x}\ dx \\ &= 1 - e^{-t} \end{aligned}$$

   obdobně F_Y.

$$\begin{aligned} F_{X,Y} &: \mathbb{R}\times\mathbb{R} \to [0,1] \\ F_{X,Y}(s,t) &= \Pr[X \le s \wedge Y \le t] \\ F_{X,Y}(s,t) &= 0 && \text{(pokud } t < 0 \text{ nebo } s < 0\text{)} \\ F_{X,Y}(s,t) &= \int_0^s\int_0^t f_{X,Y}(x,y)\ dy\ dx \\ &= \int_0^s e^{-x}\int_0^t e^{-y}\ dy\ dx \\ &= \int_0^s e^{-x}\left(1 - e^{-t}\right)\ dx \\ &= \left(1 - e^{-t}\right)\int_0^s e^{-x}\ dx \end{aligned}$$

$$= \left(1 - e^{-t}\right)\left(1 - e^{-s}\right)$$

(c) Jsou X, Y nezávislé?

Řešení: Ano, protože platí

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \qquad \text{(pro každé } x, y \in \mathbb{R})$$

Ekvivalentně (pokud vůbec existuje hustota) můžeme říct, protože platí $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ pro každé $x, y \in \mathbb{R}$.

(d) Najděte $\Pr[X + Y \leq 1]$.

Řešení: Integrujeme hustotu přes trojúhelník s vrcholy $(0,0), (1,0), (0,1)$:

$$\begin{aligned}
\Pr[X + Y \leq 1] &= \int_0^1 \int_0^{1-t} f_{X,Y}(s,t)\ ds\ dt \\
&= \int_0^1 \int_0^{1-t} e^{-s-t}\ ds\ dt \\
&= \int_0^1 e^{-t} \int_0^{1-t} e^{-s}\ ds\ dt \\
&= \int_0^1 e^{-t}\left(1 - e^{t-1}\right)\ dt \\
&= \int_0^1 e^{-t} - e^{-1}\ dt \\
&= \int_0^1 e^{-t}\ dt - e^{-1} \\
&= (1 - e^{-1}) - e^{-1} \\
&= 1 - 2/e \\
&\doteq 0.2642
\end{aligned}$$

(e) Najděte $\mathbb{E}[X + Y]$.

Řešení: Použijeme LOTUS pro funkci $g(x,y) = x + y$:

$$\begin{aligned}
\mathbb{E}[g(X,Y)] &= \int_0^\infty \int_0^\infty g(x,y) f_{X,Y}(x,y)\ dy\ dx \\
&= \int_0^\infty \int_0^\infty (x+y) e^{-x-y}\ dy\ dx \\
&= 2 \qquad \text{(pár použití per-partes)}
\end{aligned}$$

(f) Najděte $\Pr[X > Y]$.

Řešení:

$$\begin{aligned}
\Pr[X > Y] &= \int_{-\infty}^\infty \int_t^\infty f_{X,Y}(s,t)\ ds\ dt \\
&= \int_0^\infty \int_t^\infty e^{-s-t}\ ds\ dt
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty e^{-t} \int_t^\infty e^{-s}\, ds\, dt \\
&= \int_0^\infty e^{-t} e^{-t}\, dt \\
&= \lim_{b\to\infty} \int_0^b e^{-t} e^{-t}\, dt \\
&= \lim_{b\to\infty} \left[-e^{-2t}/2\right]_0^b \\
&= 0.5
\end{aligned}
$$

Což je to, co bychom čekali (jsou nezávislé a stejně rozdělené).

(g) Simulujte předchozí tři body.

Řešení:

```
import math
import numpy as np
import random

def Q_X(p):
    # F_X(x) = 1 - e^(-x)
    # p = 1 - e^(-x)
    # e^(-x) = 1 - p
    # -x = ln(1 - p)
    # x = -ln(1 - p)
    return -math.log(1-p)

def X():
    return Q_X(random.random())

N = 1_000_000

d = sum(int(X() + X() <= 1) for _ in range(N)) / N
print(f'd) Pr[X+Y <= 1] = {d} (={1 - 2 / math.e})')

e = sum(X() + X() for _ in range(N)) / N
print(f'e) E[X+Y] = {e} (=2)')

f = sum(int(X() > X()) for _ in range(N)) / N
print(f'f) Pr[X>Y] = {f} (=0.5)')

# Možný výstup:
# d) Pr[X+Y <= 1] = 0.264357 (=0.26424111765711533)
# e) E[X+Y] = 2.000897213066276 (=2)
# f) Pr[X>Y] = 0.500149 (=0.5)
```

5. **Volme uniformně náhodně bod z polokruhu o poloměru 1, se středem v počátku a v horní polorovině. (Uniformně znamená, že pravděpodobnost každé podmnožiny je úměrná jejímu obsahu.) Označme X, Y souřadnice zvoleného bodu.**

   (a) Najděte sdruženou hustotu $f_{X,Y}$.

   Řešení: Hustota je všude stejná (jedná se o uniformní distribuci). Polokruh poloměru 1 má plochu $\pi/2$, tedy $f_{X,Y}(x,y) = 2/\pi$ (abychom dostali pravděpodobnost rovnou jedné).

   (b) Najděte marginální hustotu f_Y a spočtěte pomocí ní $\mathbb{E}[Y]$.

   Řešení: Víme, že (pro $y \in [0,1]$, jinak je rovnou nulová):

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx \\ &= \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} 2/\pi\ dx \\ &= 4\sqrt{1-y^2}/\pi \end{aligned}$$

   Teď můžeme spočítat střední hodnotu

$$\begin{aligned} \mathbb{E}[Y] &= \int_0^1 y f_Y(y)\ dy \\ &= \int_0^1 y 4\sqrt{1-y^2}/\pi\ dy \\ &= \left[-\frac{4}{3\pi}(1-y^2)^{3/2}\right]_0^1 && \text{(per partes)} \\ &= \frac{4}{3\pi} \\ &= 0.4244 \end{aligned}$$

   (c) Pro kontrolu spočtěte $\mathbb{E}[Y]$ přímo (pomocí pravidla LOTUS).

   Řešení: Použijeme LOTUS pro $g(x,y) = y$ a počítáme to samé.

   (d) Simulujte.

   Řešení:

```
import math
import random

def rand_point():
  # vraci nahodny bod [-1, 1) X [0, 1)
  return (2 * random.random() - 1, random.random())

def p():
  # vraci nahodny bod v terci
  x, y = rand_point()
  while x**2 + y**2 > 1:
    x, y = rand_point()
  return (x, y)
```

```python
def Y():
    return p()[1]

N = 1_000_000
EY = sum(Y() for _ in range(N)) / N
print(f'E[Y] = {EY} (={4 / (3 * math.pi)})')

# Možný výstup:
# E[Y] = 0.4241596291469348 (=0.4244131815783876)
```

3.10 Cvičení

1. **Máme dvě mince. Jedna je spravedlivá, na druhé padá hlava s pravděpodobností** $\Pr[\textbf{hlava}] = 1/4$**. Ale nevíme která je která. Vymyslete algoritmus jak ty dvě mince rozlišit.**

   - **Vezmeme minci a hodíme** n**-krát.**
   - **Nechť** $\hat{p}$ **je pravděpodobnost, že padla hlava (počet hlav děleno** n**).**
   - **Pokud** $\hat{p} \geq 3/8$ **řekneme, že je férová, jinak cinklá.**

   Ukažte, že pro fixní konstantu $\varepsilon \in (0,1)$ **pokud** $n \geq 32\ln(2/\varepsilon)$ **náš algoritmus odpoví správně s pravděpodobností aspoň** $1-\varepsilon$**.**

   Řešení: Použijeme indikátorovou veličinu ($X_j = 1$ pokud v j-tém hodě padla hlava, 0 jinak). Použijeme trochu obecnější znění Černovovy nerovnosti, než bylo na přednášce:

   - nechť $X_j \in [0,1]$ jsou nezávislé náhodné veličiny,
   - nechť $X = \sum_{j=1}^{n} X_j$ a nechť $\mathbb{E}[X] = \sum_{j=1}^{n} \mathbb{E}[X_j] = \mu$,
   - nechť $\delta \in (0,1)$,

   pak platí:

   $$\Pr[X \geq \mu + \delta n] \leq e^{-2n\delta^2}$$
   $$\Pr[X \leq \mu - \delta n] \leq e^{-2n\delta^2}$$

   - Pokud jsme házeli férovou mincí, tak pravděpodobnost chyby (řekli jsme že je cinklá) je:

   $$\begin{aligned} \mu &= n/2 \\ \delta &= 1/8 \\ \Pr[X \leq n/2 - n/8] &\leq e^{-2\cdot 32\ln(2/\varepsilon)(1/8)^2} \\ \Pr[X \leq 3n/8] &\leq e^{-\ln(2/\varepsilon)} \\ &\leq \varepsilon/2 \end{aligned}$$

   - Pokud jsme házeli cinklou mincí, tak pravděpodobnost chyby (řekli jsme že je férová) je:

   $$\begin{aligned} \mu &= n/4 \\ \delta &= 1/8 \\ \Pr[X \geq n/4 + n/8] &\leq e^{-2\cdot 32\ln(2/\varepsilon)(1/8)^2} \\ \Pr[X \geq 3n/8] &\leq e^{-\ln(2/\varepsilon)} \\ &\leq \varepsilon/2 \end{aligned}$$

   Pak už stačí jen použít union bound $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B] \leq \Pr[A] + \Pr[B]$.

2. **(Problém šatnářky) Náhodně přiřadíme n klobouků n lidem. Označíme X_i indikátor jevu „i-tý člověk dostal svůj klobouk“ a položíme $X = \sum_{i=1}^{n} X_i$.**

   (a) Určete $\mathbb{E}[X]$.

   Řešení: Indikátory bývají velice užitečné. Využijeme linearity střední hodnoty (platí i pro závislé náhodné veličiny):

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i] && \text{(linearita } \mathbb{E}\text{)} \\
&= \sum_{i=1}^{n} 1/n \\
&= 1
\end{aligned}$$

   (b) Určete $\mathrm{var}(X)$.

   Řešení: S rozptylem to není tak jednoduché, pokud jsou náhodné veličiny nezávislé, pak rozptyl součtu je součet rozptylů. Ale to, jestli i-tý člověk dostal svůj klobouk není nezávislé na ostatních. Vyzkoušejme nezávislost například pro $n = 2$:

$$\begin{aligned}
\Pr[A \cap B] &= \Pr[A]\Pr[B] && \text{(definice nezávislosti pro jevy } A, B\text{)} \\
\forall x, y \in \mathbb{R}\colon F_{X,Y}(x,y) &= F_X(x)F_Y(y) && \text{(definice nezávislosti pro náhodné veličiny } X, Y\text{)} \\
\forall x, y \in \mathbb{R}\colon \Pr[X \le x \wedge Y \le y] &= \Pr[X \le x]\Pr[Y \le y] && \text{(přepsané zhora)} \\
1/2 = \Pr[X_1 = 1 \wedge X_2 = 1] &\neq \Pr[X_1 = 1]\Pr[X_2 = 1] = 1/4 && \text{(indikátory } \mathrm{Im}(X_1) = \{0,1\}\text{)}
\end{aligned}$$

   Rozepíšeme tedy definici rozptylu (pokud $n \geq 2$, jinak je rozptyl nulový):

$$\begin{aligned}
\mathrm{var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] && \text{(definice)} \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 && \text{(věta z přednášky – snazší na počítání)} \\
&= \mathbb{E}[X^2] - 1 && \text{(minulý bod)} \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - 1 \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{n} X_j\right)\left(\sum_{i=1}^{n} X_i\right)\right] - 1 \\
&= \mathbb{E}\left[\sum_{j=1}^{n} X_j^2 + \sum_{i \neq j} X_i X_j\right] - 1 && \text{(rozepíšeme sumu)} \\
&= \mathbb{E}\left[\sum_{j=1}^{n} X_j + \sum_{i \neq j} X_i X_j\right] - 1 && (0^2 = 0,\ 1^2 = 1) \\
&= \mathbb{E}\left[\sum_{j=1}^{n} X_j\right] + \mathbb{E}\left[\sum_{i \neq j} X_i X_j\right] - 1 && \text{(linearita } \mathbb{E}\text{)}
\end{aligned}$$

$$
\begin{aligned}
&= 1 + \mathbb{E}\left[\sum_{i\neq j} X_i X_j\right] - 1 && \text{(předchozí výsledek)}\\
&= \mathbb{E}\left[\sum_{i\neq j} X_i X_j\right]\\
&= \sum_{i\neq j} \mathbb{E}\left[X_i X_j\right] && \text{(linearita } \mathbb{E}\text{)}\\
&= \sum_{i\neq j} \frac{(n-2)!}{n!} && \text{(součin indikátorů)}\\
&= \sum_{i\neq j} \frac{1}{n(n-1)}\\
&= n(n-1)\frac{1}{n(n-1)}\\
&= 1
\end{aligned}
$$

(c) Určete σ_X.

Řešení: Směrodatná odchylka je jednoduchá, protože víme, že

$$\sigma_X = \sqrt{\operatorname{var}(X)} = 1$$

(d) Použijte Čebyševovu nerovnost na odhad pravděpodobnosti, že $X \geq 3$ (pro $n \geq 3$).

Řešení: Čebyševova nerovnost:

- Předpoklady:
  - X je náhodná veličina – OK
  - X má konečnou střední hodnotu – OK
  - X má konečný rozptyl – OK
- Závěr: pro každé reálné $k > 0$ máme

$$\Pr[|X - \mu| \geq k\sigma] \leq 1/k^2$$

Dostáváme:

$$
\begin{aligned}
\Pr[X \geq 3] &= \Pr[X - 1 \geq 2]\\
&= \Pr[|X - 1| \geq 2] && (X \geq 0)\\
\Pr[|X - \mu| \geq k\sigma] &= \Pr[|X - 1| \geq 2] && (\text{volme } k = 2/\sigma = 2)\\
&\leq 1/k^2\\
&= 1/4
\end{aligned}
$$

(e) Co by nám řekl Markov?

Řešení: Markovova nerovnost:

- Předpoklady:
  - X je nezáporná náhodná veličina
  - $a > 0$
- Důsledek:

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

Tedy pro $a = 3$ bychom dostali $\Pr[X \geq 3] \leq 1/3$, což je slabší odhad než od Čebyševa.

(f) Co by nám řekl Černov?

Řešení: Černovova nerovnost:

- Předpoklady:
  - nechť $X_j \in [0, 1]$ jsou nezávislé náhodné veličiny,
  - nechť $X = \sum_{j=1}^{n} X_j$ a nechť $\mathbb{E}[X] = \sum_{j=1}^{n} \mathbb{E}[X_j] = \mu$,
  - nechť $\delta \in (0, 1)$,
- Důsledek (jedna možná forma):

$$\Pr[X \geq \mu + \delta n] \leq e^{-2n\delta^2}$$
$$\Pr[X \leq \mu - \delta n] \leq e^{-2n\delta^2}$$

Tedy zde by nám Černov neřekl nic, protože naše X_j nejsou nezávislé.

(g) Simulujte.

Řešení:

```
from random import shuffle

def X(n):
    l = list(range(n))
    shuffle(l)
    # return number of fixed points
    fp = 0
    for i in range(n):
        if l[i] == i:
            fp += 1
    return fp

N = 100_000

for n in range(1, 10):
    EX = sum(X(n) for _ in range(N)) / N
    print(f'E[X] = {EX} (=1)')

for n in range(1, 10):
    varX = sum((X(n) - 1)**2 for _ in range(N)) / N
    print(f'var(X) = {varX} (=1)')
```

```
for n in range(2, 10):
    PrXgeq3 = sum(int(X(n) >= 3) for _ in range(N)) / N
    print(f'n = {n}: Pr[X >= 3] = {PrXgeq3} <= {1 / 4} dle Čebyševa')

# Možný výstup:
# E[X] = 1.0 (=1)
# E[X] = 0.9997 (=1)
# E[X] = 1.00316 (=1)
# E[X] = 1.00257 (=1)
# E[X] = 1.00027 (=1)
# E[X] = 0.99754 (=1)
# E[X] = 1.0001 (=1)
# E[X] = 1.00024 (=1)
# E[X] = 0.99783 (=1)
# var(X) = 0.0 (=1)
# var(X) = 1.0 (=1)
# var(X) = 1.00051 (=1)
# var(X) = 1.00172 (=1)
# var(X) = 1.00731 (=1)
# var(X) = 0.99041 (=1)
# var(X) = 1.00046 (=1)
# var(X) = 0.9999 (=1)
# var(X) = 1.00116 (=1)
# n = 2: Pr[X >= 3] = 0.0 <= 0.25 dle Čebyševa
# n = 3: Pr[X >= 3] = 0.16556 <= 0.25 dle Čebyševa
# n = 4: Pr[X >= 3] = 0.0419 <= 0.25 dle Čebyševa
# n = 5: Pr[X >= 3] = 0.0931 <= 0.25 dle Čebyševa
# n = 6: Pr[X >= 3] = 0.07803 <= 0.25 dle Čebyševa
# n = 7: Pr[X >= 3] = 0.08059 <= 0.25 dle Čebyševa
# n = 8: Pr[X >= 3] = 0.08176 <= 0.25 dle Čebyševa
# n = 9: Pr[X >= 3] = 0.08019 <= 0.25 dle Čebyševa
```

3. **Nechť X je n.v. s hustotou**

$$f_X(x) = \begin{cases} x/4 & \textbf{pro } 1 < x \leq 3 \\ 0 & \textbf{jinak.} \end{cases}$$

Označme A jev $\{X \geq 2\}$.

(a) Spočtěte $\mathbb{E}[X]$.

Řešení: Máme hustotu, tedy počítáme

$$\begin{aligned} \mathbb{E}[X] &= \int_1^3 x f_X(x)\ dx \\ &= \int_1^3 x^2/4\ dx \\ &= 13/6 \end{aligned}$$

(b) Spočtěte $\Pr[A]$.

Řešení: Máme hustotu, tedy počítáme

$$\begin{aligned} \Pr[A] &= \Pr[X \geq 2] \\ &= \int_2^3 f_X(x)\ dx \\ &= \int_2^3 x/4\ dx \\ &= 5/8 \end{aligned}$$

(c) Určete $f_{X|A}$.

Řešení: Radši si připomeňme, jak se podmiňují náhodné veličiny: Nechť $(\Omega, \mathcal{F}, \Pr)$ je pravděpodobnostní prostor a $X\colon \Omega \to \mathbb{R}$ je na něm náhodná veličina a $B \in \mathcal{F}$ je jev. Pak

$$F_{X|B}(x) = \Pr[X \leq x \mid B] = \frac{\Pr[X \leq x \wedge B]}{\Pr[B]}$$
$$F_{X|B}(x) = \Pr[X \leq x \mid B] = \int_{-\infty}^{x} f_{X|B}(s)\ ds$$

Nás tedy zajímá napřed

$$\begin{aligned} F_{X|A}(x) &= \Pr[X \leq x \mid X \geq 2] \\ &= \frac{\Pr[X \leq x \wedge X \geq 2]}{\Pr[X \geq 2]} \\ &= \frac{\int_2^x f_X(s)\ ds}{5/8} \\ &= \frac{\int_2^x s/4\ ds}{5/8} \\ &= \frac{(x^2-4)/8}{5/8} \end{aligned}$$

$$= \frac{(x^2 - 4)}{5}$$

Z toho přímo derivací dostaneme (samozřejmně že jen pro $x \in [2, 3]$):

$$\begin{aligned} f_{X|A}(x) &= \big(F_{X|A}(x)\big)' \\ &= 2x/5 \end{aligned}$$

(d) Spočtěte $\mathbb{E}[X \mid A]$.

Řešení:

$$\begin{aligned} \mathbb{E}[X \mid A] &= \int_{-\infty}^{\infty} x f_{X|A}(x)\ dx \\ &= \int_{2}^{3} 2x^2/5\ dx \\ &= 38/15 \end{aligned}$$

(e) Označme $Y = X^2$. Spočtěte $\mathbb{E}[Y]$ a $\operatorname{var}(Y)$.

Řešení: Použijeme LOTUS pro $g(x) = x^2$:

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[g(X)] \\ &= \int_{-\infty}^{\infty} g(x) f_X(x)\ dx && \text{(LOTUS)} \\ &= \int_{1}^{3} x^2 \cdot x/4\ dx \\ &= 5 \end{aligned}$$

$$\begin{aligned} \operatorname{var}(Y) &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\ &= \int_{1}^{3} x^4 \cdot x/4\ dx - 25 && \text{(LOTUS)} \\ &= 91/3 - 25 \\ &= 16/3 \end{aligned}$$

(f) Simulujte.

Řešení:

```
from math import sqrt
from random import random

def Q_X(p):
    # F_X(x) = (x^2 - 1) / 8
    # p = (x^2 - 1) / 8
    # Q_X(p) = sqrt(8*p + 1)
    return sqrt(8*p + 1)
```

```python
def X():
    return Q_X(random())

N = 1_000_000

EX = sum(X() for _ in range(N)) / N
print(f'a) E[X] = {EX} (={13/6})')

PrA = sum(int(X() >= 2) for _ in range(N)) / N
print(f'b) Pr[A] = Pr[X >= 2] = {PrA} (={5/8})')

def FXA(x):
    g2 = 0
    lx = 0
    for _ in range(N):
        s = X()
        if s >= 2:
            g2 += 1
            if s <= x:
                lx += 1
    return lx / g2

steps = 10
for S in range(2 * steps, 3 * steps + 1):
    x = S / steps
    print(f'c) F_X|A({x}) = {FXA(x)} (={(x*x - 4) / 5})')

def EXA():
    g2 = 0
    mu = 0
    for _ in range(N):
        s = X()
        if s >= 2:
            g2 += 1
            mu += (s - mu) / g2
    return mu

print(f'd) E[X|A] = {EXA()} (={38/15})')

def Y():
    x = X()
    return x*x

EY = sum(Y() for _ in range(N)) / N
print(f'e) E[Y] = {EY} (=5)')

varY = sum((Y() - 5)**2 for _ in range(N)) / N
print(f'e) var[Y] = {varY} (={16/3})')

# Možný výstup:
# a) E[X] = 2.166822681118412 (=2.1666666666666665)
# b) Pr[A] = Pr[X >= 2] = 0.624733 (=0.625)
# c) F_X|A(2.0) = 0.0 (=0.0)
# c) F_X|A(2.1) = 0.08133193103106036 (=0.08200000000000003)
```

```
# c) F_X|A(2.2) = 0.16803167959738646 (=0.16800000000000015)
# c) F_X|A(2.3) = 0.25865104045919307 (=0.25799999999999984)
# c) F_X|A(2.4) = 0.3523390104992631 (=0.352)
# c) F_X|A(2.5) = 0.4499211948963705 (=0.45)
# c) F_X|A(2.6) = 0.5506763834316044 (=0.5520000000000002)
# c) F_X|A(2.7) = 0.6582086350056199 (=0.6580000000000001)
# c) F_X|A(2.8) = 0.7675547000649079 (=0.7679999999999998)
# c) F_X|A(2.9) = 0.8818763783141509 (=0.882)
# c) F_X|A(3.0) = 1.0 (=1.0)
# d) E[X|A] = 2.5336850048416495 (=2.533333333333333)
# e) E[Y] = 5.001327835687812 (=5)
# e) var[Y] = 5.332551093906959 (=5.333333333333333)
```

4. **Nechť X, Y mají sdruženou hustotu**

$$f_{X,Y}(x,y) = \begin{cases} e^{-y} & \textbf{pro } 0 < x < y < \infty \\ 0 & \textbf{jinak.} \end{cases}$$

(a) Určete podmíněnou hustotu $f_{X|Y}$.

Řešení: Radši si připomeňme, jak se podmiňují náhodné veličiny: Nechť $(\Omega, \mathcal{F}, \Pr)$ je pravděpodobnostní prostor a $X\colon \Omega \to \mathbb{R}$ je na něm náhodná veličina a $B \in \mathcal{F}$ je jev. Pak

$$F_{X|B}(x) = \Pr[X \leq x \mid B] = \frac{\Pr[X \leq x \wedge B]}{\Pr[B]}$$
$$F_{X|B}(x) = \Pr[X \leq x \mid B] = \int_{-\infty}^{x} f_{X|B}(s)\ ds$$

Taky si vzpomeňme, že $X \leq x$ je jev přímo z definice náhodné veličiny:

$$\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$$

Takže teď už snad docela dává smysl:

$$f_{X|Y}\,(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad \text{(pokud } f_Y(y) > 0\text{, jinak nedefinovaná)}$$

takže speciálně tohle bere dvě čísla a vrací jedno!

Tedy hustota je definovaná jen pro $y > 0$:

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{e^{-y}}{\int_0^y f_{X,Y}(x,y)\ dx} \\ &= \frac{e^{-y}}{\int_0^y e^{-y}\ dx} \\ &= \frac{e^{-y}}{ye^{-y}} \\ &= \frac{1}{y} \end{aligned}$$

Což dává smysl, že X je uniformní (sdružená hustota na x nezávisí).

(b) Určete podmíněnou hustotu $f_{Y|X}$.

Řešení: Tedy hustota je definovaná jen pro $x > 0$ a navíc $y > x$:

$$\begin{aligned} f_{Y|X}(y \mid x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{e^{-y}}{\int_x^\infty f_{X,Y}(x,y)\ dy} \\ &= \frac{e^{-y}}{\int_x^\infty e^{-y}\ dy} \\ &= \frac{e^{-y}}{e^{-x}} \\ &= e^{-y+x} \end{aligned}$$

Což opět dává smysl, protože pokud zafixujeme $X = 10$, pak $\mathrm{Im}(Y) = (10, \infty)$.

5. **V memech na discordu se objevují pouze tyto typy memů: s opicemi – jev M, s kraby – jev C, ostatní – jev O. Některé z nich jsou ve velkém rozlišení – jev HD. Formálně Ω jsou memy a jev M je podmnožina memů na kterých je opice, ... Navíc platí že na každém memu je právě jedna z těch věcí $\Omega = M \dot\cup C \dot\cup O$ (disjunktní sjednocení – tedy $\Omega = M \cup C \cup O$ a navíc $M \cap C = M \cap O = C \cap O = \emptyset$).**

   - $\Pr[M] = 1/4$
   - $\Pr[C] = 3/44$
   - $\Pr[O] = 15/22$
   - $\Pr[HD \mid M] = 1/11$
   - $\Pr[HD \mid C] = 13/15$
   - $\Pr[HD \mid O] = 9/15$

   Napsal jsem si program, který mi jako wallpaper nastaví náhodný meme, který je ve velkém rozlišení. Jaká je pravděpodobnost, že mám jako wallpaper kraba?

   Řešení: Protože na memu je buď opice, nebo krab, nebo něco jiného (nikdy na memu nejsou dva nebo víc z nich a na každém memu něco je), tedy tyto kategorie tvoří disjunktní rozklad pravděpodobnostního prostoru. Můžeme použít Bayesovu větu:

$$\begin{aligned}\Pr[C \mid HD] &= \frac{\Pr[C]\Pr[HD \mid C]}{\Pr[M]\Pr[HD \mid M] + \Pr[C]\Pr[HD \mid C] + \Pr[O]\Pr[HD \mid O]} \\ &= \frac{(3/44)(13/15)}{(1/4)(1/11) + (3/44)(13/15) + (15/22)(9/15)} \\ &= 13/108\end{aligned}$$

   Srovnejte úlohu s následující (čísla jsou stejná, jen přesně znáte počty): za poslední týden tam byly následující memy, všechny jsem stáhl a program vybírá wallpaper s velkým rozlišením z mého adresáře:

   - 55 memů opic, 5 z nich je ve velkém rozlišení
   - 15 memů krabů, 13 z nich je ve velkém rozlišení
   - 150 ostatních, 90 z nich je ve velkém rozlišení

   Poučení: když přemýšlíte o Bayesově větě, napište si konkrétní čísla!

6. **Nechť $U \sim U(-1,1)$. Položme $V = 2|U| - 1$.**

   (a) Určete rozdělení V. (Tj. spočtěte distribuční funkci a případně popište, o jaké pojmenované rozdělení se jedná.)

   Řešení:

   - Protože U je uniformní a symetrická okolo nuly, pak $|U|$ je taky uniformní, konkrétně $U(0,1)$.
   - Dvakrát uniformní veličina nám dá $U(0,2)$.
   - Odečtením jedné posuneme, tedy výsledek je rozdělen jako $U(-1,1)$ (tedy jako naše původní veličina).

   Mohli jsme přímo počítat distribuční funkci:

$$F_U(x) = \begin{cases} \frac{x-(-1)}{2} = (x+1)/2 & x \in (-1,1) \\ 1 & x \geq 1 \\ 0 & x \leq -1 \end{cases}$$

   (b) Spočtete $\operatorname{cov}(U,V)$.

   Řešení:

$$\begin{aligned}
\operatorname{cov}(U,V) &= \mathbb{E}[(U - \mathbb{E}[U])(V - \mathbb{E}[V])] && \text{(definice)} \\
&= \mathbb{E}[UV] - \mathbb{E}[U]\,\mathbb{E}[V] && \text{(věta)} \\
&= \mathbb{E}[UV] - 0 \\
&= \mathbb{E}[U(2|U| - 1)] \\
&= \mathbb{E}[2U|U| - U] \\
&= \mathbb{E}[g(U)] && \text{(LOTUS, } g(u) = 2u|u| - u\text{)} \\
&= \int_{-1}^{1} g(t) f_U(t)\ dt \\
&= \int_{-1}^{1} g(t)/2\ dt \\
&= \int_{-1}^{1} (2u|u| - u)/2\ dt \\
&= 0 && \text{(všimněte si, že } g(u) = -g(-u)\text{)}
\end{aligned}$$

   (c) Jsou U, V nezávislé?

   Řešení: Nikoliv, pokud $U = u$, pak $V = 2|u| - 1$, tedy jedna přesně určuje druhou.

   Pro nezávislé náhodné veličiny je jejich kovariance nulová, ale opačná implikace neplatí.

   (d) Simulujte.

   Řešení:

```
from random import random

def U(omega):
    return 2 * omega - 1
```

```python
def V(omega):
    return 2 * abs(U(omega)) - 1

def F(X, x, N = 1_000_000):
    return sum(int(X(random()) <= x) for _ in range(N)) / N

steps = 10
for S in range(-1 * steps, 1 * steps + 1):
    x = S / steps
    print(f'a) F_U({x}) = {F(U, x)} = {F(V, x)} = F_V({x}) (={(x+1)/2})')

def cov(X, Y, N = 1_000_000):
    EX = sum(X(random()) for _ in range(N)) / N
    EY = sum(Y(random()) for _ in range(N)) / N
    # cov(X, Y) = E[(X - EX)*(Y - EY)] = E[X*Y] - EX*EY
    covXY = 0.0
    for n in range(1, N+1):
        omega = random()
        covXY += ((X(omega) - EX)*(Y(omega) - EY) - covXY) / n
    return covXY

print(f'b) cov(U, V) = {cov(U, V)} (=0)')

# Možný výstup:
# a) F_U(-1.0) = 0.0 = 0.0 = F_V(-1.0) (=0.0)
# a) F_U(-0.9) = 0.049899 = 0.049737 = F_V(-0.9) (=0.04999999999999999)
# a) F_U(-0.8) = 0.100135 = 0.100227 = F_V(-0.8) (=0.09999999999999998)
# a) F_U(-0.7) = 0.149643 = 0.149651 = F_V(-0.7) (=0.15000000000000002)
# a) F_U(-0.6) = 0.200602 = 0.200123 = F_V(-0.6) (=0.2)
# a) F_U(-0.5) = 0.250112 = 0.250007 = F_V(-0.5) (=0.25)
# a) F_U(-0.4) = 0.300192 = 0.299649 = F_V(-0.4) (=0.3)
# a) F_U(-0.3) = 0.349962 = 0.349825 = F_V(-0.3) (=0.35)
# a) F_U(-0.2) = 0.399809 = 0.400493 = F_V(-0.2) (=0.4)
# a) F_U(-0.1) = 0.449575 = 0.449717 = F_V(-0.1) (=0.45)
# a) F_U(0.0) = 0.499824 = 0.499518 = F_V(0.0) (=0.5)
# a) F_U(0.1) = 0.548915 = 0.550328 = F_V(0.1) (=0.55)
# a) F_U(0.2) = 0.599676 = 0.600496 = F_V(0.2) (=0.6)
# a) F_U(0.3) = 0.649753 = 0.649714 = F_V(0.3) (=0.65)
# a) F_U(0.4) = 0.699602 = 0.699844 = F_V(0.4) (=0.7)
# a) F_U(0.5) = 0.749851 = 0.749906 = F_V(0.5) (=0.75)
# a) F_U(0.6) = 0.799615 = 0.799874 = F_V(0.6) (=0.8)
# a) F_U(0.7) = 0.850289 = 0.850031 = F_V(0.7) (=0.85)
# a) F_U(0.8) = 0.900087 = 0.899764 = F_V(0.8) (=0.9)
# a) F_U(0.9) = 0.949534 = 0.949907 = F_V(0.9) (=0.95)
# a) F_U(1.0) = 1.0 = 1.0 = F_V(1.0) (=1.0)
# b) cov(U, V) = -0.0002382489639495386 (=0)
```

3.11 Cvičení

1. **Označme $S = \sum_{k=0}^{30} \binom{100}{k}$. Označme dále $X = \sum_{i=1}^{100} X_i$, kde X_i je ± 1 s pravděpodobností $1/2$ a veličiny $X_1, \ldots, X_n$ jsou nezávislé.**

   (a) Vyjádřete S pomocí vhodné pravděpodobnosti výroku o X.

   Řešení: Pokud nejvýš 30 z veličin X_i padne kladných, pak máme jejich součet nejvýš -40. Všimneme si, že pokud bychom tu sumu vydělili 2^{100}, pak bychom dostali pravděpodobnost, že nejvýš 30 z veličin X_i padne kladných.

   $$\frac{S}{2^{100}} = \frac{\sum_{k=0}^{30} \binom{100}{k}}{2^{100}} = \Pr[X \leq -40]$$

   (b) Použijte CLV na odhad této pravděpodobnosti.

   Řešení: Napřed si připomeneme centrální limitní větu: Předpoklady:

   - $X_1, \ldots, X_n$ jsou stejně rozdělené nezávislé náhodné veličiny se střední hodnotou μ a rozptylem σ^2.
   - Značme $Y_n = ((X_1 + \ldots + X_n) - n\mu) \,/\, (\sqrt{n}\sigma)$

   Důsledek:

   - $Y_n \xrightarrow{d} N(0,1)$ tedy Y_n konverguje k normálnímu rozdělení v distribuci (pro větší a větší n)
   - Ekvivalentně: pokud F_n je distribuční funkce Y_n, pak

     $$\lim_{n\to\infty} F_n(x) = \Phi(x) \qquad \text{(pro každé } x \in \mathbb{R}\text{)}$$

   Je super, že nám CLV říká, že můžeme odhadnout distribuční funkci pomocí tabulek (softwaru). Víme, že X je součet stejně rozdělených nezávislých náhodných veličin se střední hodnotou $\mu = 0$ a rozptylem $\sigma^2 = 1$. Tedy $Y_n = X/\sqrt{n}$ je skoro normálně rozdělené. Můžeme tedy dosadit:

   $$\begin{aligned} S &= 2^{100} \Pr[X \leq -40] \\ &\doteq 2^{100} \Pr[Y \leq -4] && \text{(tady máme konvergenci pro velké } n\text{)} \\ &\doteq 2^{100} \cdot 3.167/10^5 \\ &\doteq 4.015 \cdot 10^{25} \end{aligned}$$

   (c) Případně vyčíslete S vhodným softwarem a srovnejte.

   Řešení:

```
import scipy.stats as st
import scipy.special as sp

clv = int(2**100 * st.norm.cdf(-4))
exact = sum(sp.comb(100, k, exact=True) for k in range(31))

print(clv)
print(exact)

# Výsledek:
# 40148068719727803203846144
# 49756171168061176633478360
```

2. **Nechť X_j pro $1 \leq j \leq n$ jsou nezávislé náhodné veličiny, pro které platí $\Pr[X_j = 2/3] = 1/2$ a $\Pr[X_j = 0] = 1/2$ (jiných hodnot ty veličiny nenabývají). Nechť $X = \sum_{j=1}^{n} X_j$ je náhodná veličina rovná jejich součtu. V každém bodě použijte tu konkrétní verzi odhadu, kterou jsem napsal (ne že by jiné nefungovaly). Chceme shora odhadnout $\Pr[X \geq n/2]$**

   (a) Určete $\mathbb{E}[X]$ (Jaký poznatek používáte? Jaké má předpoklady?)

   Řešení:

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\sum_{j=1}^{n} X_j\right] \\ &= \sum_{j=1}^{n} \mathbb{E}\left[X_j\right] && \text{(linearita střední hodnoty)} \\ &= n/3\end{aligned}$$

   (b) Určete $\mathrm{var}(X)$ (Jaký poznatek používáte? Jaké má předpoklady?)

   Řešení: Pro nezávislé platí $\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.

$$\begin{aligned}\mathrm{var}(X) &= \mathrm{var}\left(\sum_{j=1}^{n} X_j\right) \\ &= \sum_{j=1}^{n} \mathrm{var}(X_j) && \text{(nezávislost)} \\ &= n\left(\mathbb{E}[X_j^2] - \mathbb{E}[X_j]^2\right) \\ &= n\left(2/9 - 1/9\right) \\ &= n/9\end{aligned}$$

   (c) Jak $\Pr[X \geq n/2]$ odhadne Markov? (Jsou splněny předpoklady? Jaký je závěr?) Markovova nerovnost:

   - **Předpoklady:**
     - **X je nezáporná náhodná veličina**
     - **$a > 0$ je reálné číslo**
   - **Důsledek:**

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

   Řešení: Můžeme použít Markova, $a = n/2$.

$$\begin{aligned}\Pr[X \geq n/2] &\leq \frac{n/3}{n/2} \\ &= 2/3\end{aligned}$$

   (d) Jak $\Pr[X \geq n/2]$ odhadne Čebyšev? (Jsou splněny předpoklady? Jaký je závěr?) Čebyševova nerovnost:

- **Předpoklady:**
  - X **je náhodná veličina**
  - X **má konečnou střední hodnotu**
  - X **má konečný rozptyl**
- **Závěr: pro každé reálné** $k > 0$ **máme**

$$\Pr[|X - \mu| \geq k\sigma] \leq 1/k^2$$

Řešení: Nevím jak použít, když zvolíme $k\sigma = n/6$, tak odhadujeme

$$\Pr[X \leq n/6 \vee X \geq n/2]$$

což není to, na co jsme se ptali a museli bychom napřed ještě nějak odhadnout $\Pr[X \leq n/6]$ a použít union bound.

(e) Jak $\Pr[X \geq n/2]$ **odhadne Černov? (Jsou splněny předpoklady? Jaký je závěr?) Černovova nerovnost:**

- **Předpoklady:**
  - **nechť** $X_j \in [0, 1]$ **jsou nezávislé náhodné veličiny,**
  - **nechť** $X = \sum_{j=1}^{n} X_j$ **a nechť** $\mathbb{E}[X] = \sum_{j=1}^{n} \mathbb{E}[X_j] = \mu$**,**
  - **nechť** $\delta \in (0, 1)$**,**
- **Závěr:**

$$\Pr[X \geq \mu + \delta n] \leq e^{-2n\delta^2}$$
$$\Pr[X \leq \mu - \delta n] \leq e^{-2n\delta^2}$$

Řešení:

$$\delta = 1/6$$
$$\Pr[X \geq \mu + \delta n] \leq e^{-2n\delta^2}$$
$$\Pr[X \geq n/3 + n/6] \leq e^{-2n(1/6)^2}$$
$$\Pr[X \geq n/2] \leq e^{-n/18}$$

(f) Jak $\Pr[X \geq n/2]$ **odhadne centrální limitní věta? (Jsou splněny předpoklady? Jaký je závěr?) Centrální limitní věta: Předpoklady:**

- $X_1, \ldots, X_n$ **jsou stejně rozdělené nezávislé náhodné veličiny se střední hodnotou** $\mathbb{E}[X_j] = \mu$ **a rozptylem** $\mathrm{var}(X_j) = \sigma^2$ **(pozor, tady mluvíme o** X_j**).**
- **Značme** $Y_n = ((X_1 + \ldots + X_n) - n\mu) \,/\, (\sqrt{n}\sigma)$

Důsledek:

- $Y_n \xrightarrow{d} N(0, 1)$ **tedy** Y_n **konverguje k normálnímu rozdělení v distribuci (pro větší a větší** n**)**

- **Ekvivalentně: pokud F_n je distribuční funkce Y_n, pak**

$$\lim_{n\to\infty} F_n(x) = \Phi(x) \qquad \text{(pro každé } x \in \mathbb{R}\text{)}$$

Může se hodit `scipy.stats.norm.cdf` nebo tabulky na Wikipedii. Pozor, že toto je odhad, který má platit v limitě, nikoliv pro každé n.

Řešení: Předpoklady splněny.

$$\begin{aligned}
\Pr[X \geq n/2] &= \Pr[X - n/3 \geq n/6] \\
&= \Pr[(X - n/3)/(\sqrt{n}/9) \geq n/(6\sqrt{n}/9)] \\
&= \Pr[(X - n/3)/(\sqrt{n}/9) \geq 3\sqrt{n}/2] \\
&\doteq 1 - \Phi(3\sqrt{n}/2) \qquad \text{(v limitě pro velké } n\text{)}
\end{aligned}$$

(g) **Simulujte a porovnejte výsledek simulace s předchozími závěry. Simulujte a porovnejte s odhady pro: $n \in \{1, 2, 3, 4, 5, 10, 20, 30, 40, 50\}$.**

Řešení:

```
import math
from random import choice
import scipy.stats as st

def X(n = 1):
    return sum(choice([2/3, 0]) for _ in range(n))

def Markov(n):
    return 2/3

def Cernof(n):
    return math.exp(-n / 18)

def CLV(n):
    return 1 - st.norm.cdf(3 * math.sqrt(n) / 2)

N = 100_000

for n in [1, 2, 3, 4, 5, 10, 20, 30, 40, 50]:
    print(f'{n}:')
    P = sum(int(X(n) >= n/2) for _ in range(N)) / N
    print(f'Markov:  {P} <= {Markov(n)}')
    print(f'Černov:  {P} <= {Cernof(n)}')
    print(f'CLV:     {P} <= {CLV(n)}')

    print('')

# Možný výstup:
# 1:
# Markov:  0.50037 <= 0.6666666666666666
# Černov:  0.50037 <= 0.9459594689067654
# CLV:     0.50037 <= 0.06680720126885809
```

```
# 2:
# Markov:  0.25298 <= 0.6666666666666666
# Černov:  0.25298 <= 0.8948393168143698
# CLV:     0.25298 <= 0.01694742676234462

# 3:
# Markov:  0.12359 <= 0.6666666666666666
# Černov:  0.12359 <= 0.8464817248906141
# CLV:     0.12359 <= 0.004687384229717484

# 4:
# Markov:  0.31127 <= 0.6666666666666666
# Černov:  0.31127 <= 0.8007374029168081
# CLV:     0.31127 <= 0.0013498980316301035

# 5:
# Markov:  0.18931 <= 0.6666666666666666
# Černov:  0.18931 <= 0.7574651283969664
# CLV:     0.18931 <= 0.0003981150787953913

# 10:
# Markov:  0.05388 <= 0.6666666666666666
# Černov:  0.05388 <= 0.5737534207374327
# CLV:     0.05388 <= 1.050717977957305e-06

# 20:
# Markov:  0.00582 <= 0.6666666666666666
# Černov:  0.00582 <= 0.32919298780790557
# CLV:     0.00582 <= 9.851675031313789e-12

# 30:
# Markov:  0.00276 <= 0.6666666666666666
# Černov:  0.00276 <= 0.18887560283756183
# CLV:     0.00276 <= 1.1102230246251565e-16

# 40:
# Markov:  0.0012 <= 0.6666666666666666
# Černov:  0.0012 <= 0.10836802322189586
# CLV:     0.0012 <= 0.0

# 50:
# Markov:  0.00011 <= 0.6666666666666666
# Černov:  0.00011 <= 0.06217652402211632
# CLV:     0.00011 <= 0.0
```

3. **Máme k dispozici samply $X_1, \dots X_N$ kde $X_j \sim Bin(n, p)$. Ale my neznáme ani n ani p.**

   Praktický příklad:

   - **Na parapetu mám $N = 50$ květináčů**
   - **Vím, že jsem do každého květináče nasypal stejně semínek (do každého n), ale už jsem zapomněl kolik to bylo.**
   - **Nevím s jakou pravděpodobností p které semínko vyklíčí (předpokládám že vyklíčí nezávisle náhodně na ostatních).**
   - **Ale vím kolik mi v kterém květináči vyrostlo rostlinek X_j v j-tém květináči.**

   Takže mám $X_1, \dots, X_N$ výběr z modelu s parametrem ϑ (zde ϑ je (n, p)). Použijte metodu momentů k odhadu n, p.

   Řešení: Z přednášky víme, že pokud $X \sim Bin(n, p)$, pak platí:

$$\begin{aligned} \mathbb{E}[X] &= np \\ \mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 &= np(1-p) \end{aligned}$$

   My máme jednotlivá měření (počty rostlinek v jednotlivých květináčích). Můžeme tedy spočítat výběrové momenty (to jsou ta čísla, která skutečně spočteme z toho naměřeného):

$$\begin{aligned} m_1 &= \frac{1}{N}\sum_{j=1}^{N} X_j \\ m_2 &= \frac{1}{N}\sum_{j=1}^{N} X_j^2 \end{aligned}$$

   Z přednášky víme, že m_1 je nestranný konzistentní odhad pro $\mathbb{E}[X]$ a navíc m_2 je taky nestranný konzistentní odhad pro $\mathrm{var}(X) + \mathbb{E}[X]^2$. Takže máme soustavu rovnic:

$$\begin{aligned} m_1 &= np \\ m_2 - m_1^2 &= np(1-p) = np - np^2 \end{aligned}$$

   získáváme tedy

$$\begin{aligned} m_1 &= np \\ m_2 - m_1^2 - m_1 &= -np^2 = -pm_1 \end{aligned}$$

   z čehož

$$\begin{aligned} p &= \frac{m_1^2 + m_1 - m_2}{m_1} \\ n &= m_1/p \end{aligned}$$

```python
import numpy as np

def X(N = 50):
    # Parametry, které neznáme:
    n = 30   # semínek
    p = 0.6  # klíčivost
    return np.random.binomial(n, p, N)
```

```python
for _ in range(5):
    parapet = X()
    m_1 = np.mean(parapet)  # (sum X_j) / N
    m_2 = np.mean(parapet ** 2)  # (sum X_j^2) / N

    # Víme:
    # E[X] = np
    # var(X) = E[X^2] - (E[X])^2 = m_2 - m_1^2 = np(1-p)

    p = (m_1**2 + m_1 - m_2) / m_1
    n = m_1 / p

    print(f'Náš odhad: n = {n}, p = {p}  (m_1 = {m_1}, m_2 = {m_2})')

# Možný výstup:
# Náš odhad: n = 24.381573646533926, p = 0.7275986471251379  (m_1 = 17.74, m_2 = 319.54)
# Náš odhad: n = 28.546122448979588, p = 0.6550802139037434  (m_1 = 18.7, m_2 = 356.14)
# Náš odhad: n = 28.21872316060683, p = 0.6293693693693716  (m_1 = 17.76, m_2 = 322.0)
# Náš odhad: n = 26.157536008230462, p = 0.6896674057649664  (m_1 = 18.04, m_2 = 331.04)
# Náš odhad: n = 26.61048046462501, p = 0.6749220489977776  (m_1 = 17.96, m_2 = 328.4)
```

Co mi na tomto přístupu vadí je, že mi sice dá odhad (který je celkem dobrý), ale nedá mi žádnou záruku. Chtěl bych spíš něco jako ty parametry jsou v tomhle rozmezí s velikou pravděpodobností – intervalový odhad.

4. **Po mírném zklamání v předchozím příkladě jsme se rozhodli aspoň zjistit klíčivost. Zasadili jsme tedy $n = 100$ semínek do takového toho platíčka 10×10 květináčků. Za pár týdnů nám některá vyklíčila, X_j je indikátor, jestli j-té semínko vyklíčilo. Pomocí maximální věrohodnosti odhadněte klíčivost p.**

Řešení: Slovníček

- Budeme pozorovat náhodný výběr $X = (X_1, X_2, \ldots, X_n)$ kde $X_j \sim Bern(p)$, tedy $\vartheta = p$.
- Možný výsledek je $x = (x_1, x_2, \ldots, x_n)$ (která semínka skutečně vyklíčila a která ne)
- Věrohodnost, likelyhood $L(x; \vartheta) = \Pr[X \mid \vartheta]$ (pro fixní parametr nám řekne jak pravděpodobné je vidět náš pozorovaný výsledek)
- Metoda maximální věrohodnosti: volíme takový parametr ϑ, že pravděpodobnost pozorování x je co největší[1]

Díky nezávislosti X_j píšeme

$$L(x; \vartheta) = \Pr[x_1 \mid \vartheta] \cdots \Pr[x_n \mid \vartheta] = \Pi_{j=1}^{n} \Pr[x_j \mid \vartheta]$$
$$\ell(x; \vartheta) = \log(\Pr[x_1 \mid \vartheta]) + \ldots + \log(\Pr[x_n \mid \vartheta]) = \sum_{j=1}^{n} \log(\Pr[x_j \mid \vartheta])$$

To druhé je lepší, protože součet se může optimalizovat snáz (občas).

Protože $X_j \sim Bern(p)$, tak víme, že

$$\begin{aligned} \Pr[1 \mid p] &= p \\ \Pr[0 \mid p] &= 1 - p \\ \Pr[x_j \mid p] &= p^{x_j}(1-p)^{(1-x_j)} \qquad \text{(trik, který pomůže s optimalizací)} \end{aligned}$$

Budeme optimalizovat rovnou ten součin, protože můžeme:

$$\begin{aligned} L(x; \vartheta) &= \Pr[x_1 \mid \vartheta] \cdots \Pr[x_n \mid \vartheta] = \Pi_{j=1}^{n} \Pr[x_j \mid \vartheta] \\ &= \Pi_{j=1}^{n} p^{x_j}(1-p)^{(1-x_j)} \\ &= p^{\sum_{j=1}^{n} x_j}(1-p)^{n-\sum_{j=1}^{n} x_j} \\ &= p^S(1-p)^{n-S} \qquad (S = \textstyle\sum_{j=1}^{n} x_j) \end{aligned}$$

Funkce je hezká, koukneme kde je její derivace podle p nulová a tam může být optimum:

$$\begin{aligned} \frac{\partial}{\partial p} L(x; \vartheta) &= \frac{\partial}{\partial p} p^S(1-p)^{n-S} \qquad (S = \textstyle\sum_{j=1}^{n} x_j) \\ &= Sp^{S-1}(1-p)^{n-S} - p^S(n-S)(1-p)^{n-S-1} \\ &= (1-p)^{n-S-1}p^{S-1}\left(S(1-p) - (n-S)p\right) \end{aligned}$$

což je nulové pro $p = 0$ pro $p = 1$ nebo pro $0 = S - Sp - np + Sp$, což dává $p = S/n$. Z vlastností funkce nahlédneme, že to poslední je maximum.

Poznámka: ano, mohli jsme znova z větších květináčů odhadovat obojí n, p, ale mě se nechtělo derivovat gamma funkci.

[1] Tady nás nezajímá ta pravděpodobnost, ale ϑ. Akorát volíme tu nejpravděpodobnější ϑ.

5. **Známe směrodatnou odchylku σ, ale neznáme střední hodnotu μ. Máme $n = 100$ samplů $X_1, \ldots, X_n$ každý nezávislý a $X_j \sim N(\mu, \sigma^2)$. Jako chybovost volme $\alpha = 0.01$. Ověřte, že intervalový odhad funguje dobře.**

   Řešení: Chceme interval $[L, U]$ takový, že

   $$\Pr[L \leq \mu \leq U] \geq 1 - \alpha$$

   Dle přednášky volíme

   - $\overline{\mu} = \frac{1}{n} \sum_{j=1}^{n} X_j$
   - $z_{\alpha/2} = \text{ppf}(1 - \alpha/2)$
   - Odpovídáme intervalem $[\overline{\mu} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{\mu} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$

   Experimentálně ověříme:

```
import math
import numpy as np
import scipy.stats as st

def interval_est(x, sigma, alpha = 0.01):
    mu = np.mean(x)
    z = st.norm.ppf(1 - (alpha / 2))
    delta = z * sigma / math.sqrt(x.shape[0])
    return (mu - delta, mu + delta)

N = 10_000
alpha = 0.01
inside = 0

for i in range(N):
    mu = 180
    sigma = 7
    n = 100
    x = np.random.normal(mu, sigma, n)
    interval = interval_est(x, sigma, alpha)
    if i > N - 6:
        print(interval)
    if interval[0] <= mu <= interval[1]:
        inside += 1

print(f'Pr[odhad mu je v intervalu] = {inside / N} >= {1 - alpha}')

# Možný výstup:
# (177.3449942031492, 180.95115522811767)
# (178.75897965020843, 182.3651406751769)
# (177.37789102873765, 180.9840520537061)
# (178.85582072289046, 182.4619817478589)
# (177.22390884109217, 180.83006986606063)
# Pr[odhad mu je v intervalu] = 0.9904 >= 0.99
```

   Co vám na tomto příkladě bytostně vadí?

   Řešení: Neznáme střední hodnotu a pokoušíme se ji odhadnout, ale zato naprosto přesně známe rozptyl. To je blbost, proto potřebujeme studentovo rozdělení.

6. **William Sealy Gosset pracoval pro nejmenovaný pivovar v Dublinu a zajímaly ho malé samply (třeba tři samply), protože odhadoval kvalitu piva a samply byly drahé. Zkuste to předchozí se studentovým rozdělením. Tedy máme nezávislé náhodné proměnné $X_1, \ldots, X_n$ kde $X_j \sim N(\mu, \sigma^2)$, ale neznáme ani střední hodnotu ani rozptyl (reálná situace) a chceme intervalem odhadnout střední hodnotu (ta je často ta zajímavější).**

   Řešení: Chceme interval $[L, U]$ takový, že

   $$\Pr[L \leq \mu \leq U] \geq 1 - \alpha$$

   Dle přednášky volíme

   - $\overline{\mu} = \frac{1}{n} \sum_{j=1}^{n} X_j$ (výběrový průměr)
   - $\overline{\sigma}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{\mu})^2$ (výběrový rozptyl, všimněte si $n-1$)
   - Víme že:

     $$\frac{\overline{\mu} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

     pozor na to, co jsou čísla a co náhodné proměnné:

     - μ je číslo, parametr modelu který neznáme
     - σ je číslo, parametr modelu který neznáme
     - $\overline{\mu}$ je náhodná veličina, průměr náhodných veličin, jejichž hodnoty pozorujeme
     - $\overline{\sigma}^2$ je náhodná veličina, náš odhad rozptylu
     - n je počet samplů, ten známe

     Dále víme, že:

     $$\frac{\overline{\mu} - \mu}{\overline{\sigma} / \sqrt{n}}$$

     je rozdělena dle studentova rozdělení s $n-1$ stupni volnosti (obdobný důvod $n-1$ jako ve výběrovém rozptylu). Prakticky vypadá studentovo a normální rozdělení dost podobně, ale studentovo má o něco víc pravděpodobné extrémní hodnoty. Čím víc stupňů volnosti má studentovo, tím je podobnější normálnímu.

   - $z_{\alpha/2} = \text{t.ppf}(1 - \alpha/2)$
   - Odpovídáme intervalem $[\overline{\mu} - z_{\alpha/2} \frac{\overline{\sigma}}{\sqrt{n}}, \overline{\mu} + z_{\alpha/2} \frac{\overline{\sigma}}{\sqrt{n}}]$

   Experimentálně ověříme:

```python
import math
import numpy as np
import scipy.stats as st

def interval_est(x, alpha = 0.01):
    mu = np.mean(x)
    sigma = np.std(x, ddof=1)  # dělí n-1
    # studentovo rozdělení místo norm, n-1 stupňů volnosti
    z = st.t.ppf(1 - (alpha / 2), df=x.shape[0]-1)
    delta = z * sigma / math.sqrt(x.shape[0])
    return (mu - delta, mu + delta)
```

```
N = 10_000
alpha = 0.01
inside = 0

for i in range(N):
    mu = 180
    sigma = 7
    n = 100
    x = np.random.normal(mu, sigma, n)
    interval = interval_est(x, alpha)
    if i > N - 6:
        print(interval)
    if interval[0] <= mu <= interval[1]:
        inside += 1

print(f'Pr[odhad mu je v intervalu] = {inside / N} >= {1 - alpha}')

# Možný výstup:
# (179.56885043263986, 183.0187306219168)
# (178.78350329833617, 182.60262157104628)
# (178.31609435918605, 181.84659174973618)
# (177.56923048935607, 181.38718494483516)
# (178.1977221926624, 181.49228846237418)
# Pr[odhad mu je v intervalu] = 0.9913 >= 0.99
```

3.12 Cvičení

1. **Po částečném úspěchu v určování klíčivosti chceme mít ještě lepší představu o skutečné hodnotě. Místo bodového odhadu (o kterém nevíme jak daleko je pravdě) chceme intervalový odhad (s pravděpodobností 99% se trefíme intervalem okolo správné hodnoty). Zasadili jsme tedy $n = 100$ semínek do takového toho platíčka 10×10 květináčků. Za pár týdnů nám některá vyklíčila, X_j je indikátor, jestli j-té semínko vyklíčilo. Pomocí intervalového odhadu odhadněte klíčivost p.**

   (a) **Co kdybychom chtěli odhadovat pomocí normálního rozdělení? Tedy kdybychom měli rozptyl, ale chtěli odhadnout střední hodnotu (tedy pro indikátor $\Pr[X_j = 1]$)?**

   Řešení: Tak bychom nepotřebovali pozorování, protože $\mathrm{var}(X_j) = p(1-p)$, z čehož dopočítáme p a jdeme domů.

   (b) **Použijte centrální limitní větu a tedy studentovo rozdělení na odhad pro $\alpha = 0.01$.**

   Řešení: Podle centrální limitní věty (indikátory jsou nezávislé a stejně distribuované) sice platí, že počet semínek která vyklíčí bude zhruba rozdělený jako $N(100p, 100p(1-p))$, ale to je jen jeden sample (jedno pozorování). My ale budeme lstiví. Rozdělíme květináčky na deset řádek po deseti semínkách. Každý řádek je binomicky rozdělený podle

$$Binom(10, p),$$

   ale podle centrální limitní věty se tohle limitně blíží k

$$N(10p, 10p(1-p)),$$

   protože 10 už je poměrně velké číslo. Tedy máme deset samplů, které jsou jakž takž normálně rozdělené, takže můžeme použít studentovo rozdělení.

```
import math
import numpy as np
import scipy.stats as st

def interval_est(x, alpha = 0.01):
    mu = np.mean(x)
    sigma = np.std(x, ddof=1)  # dělí n-1
    # studentovo rozdělení místo norm, n-1 stupňů volnosti
    z = st.t.ppf(1 - (alpha / 2), df=x.shape[0]-1)
    delta = z * sigma / math.sqrt(x.shape[0])
    return (mu - delta, mu + delta)

N = 10_000
alpha = 0.01
inside = 0

for i in range(N):
    p = 0.6
    # Spočítáme kolik vyklíčilo v každém řádku, budeme doufat,
    # že n=10 bude pro centrální limitní větu stačit.
    seminek_v_radku = 10
    pocet_radku = 10
    x = np.random.binomial(seminek_v_radku, p, pocet_radku)
```

```
        interval = interval_est(x, alpha)
        if i > N - 6:
            print(interval)
        if interval[0] <= seminek_v_radku * p <= interval[1]:
            inside += 1

print(f'Pr[odhad p je v intervalu] = {inside / N} >= {1 - alpha}')

# Možný výstup (pozor, že tohle jsou odhady 10p):
# (3.6417708221235996, 7.7582291778764401)
# (4.610738262339158, 7.589261737660841)
# (5.164187640334102, 8.235812359665898)
# (5.633273333292907, 7.966726666707093)
# (5.138608155893622, 7.261391844106378)
# Pr[odhad p je v intervalu] = 0.9903 >= 0.99
```

2. **Podle slibu výrobce bude stroj dělat chyby nejvýše ve 3% případů. Z 600 pokusů došlo k chybě v 28 případech. Posuďte slib výrobce (coby nulovou hypotézu) na hladině významnosti 5%.**

   (a) Co jsme měli udělat před pozorováním chybných strojů?

   Řešení: Tohle se mělo dělat před tím než jsme pozorovali počet chyb!

   - H_0 je, že chybovost je nejvýše 0.03 (tedy 3%).
   - H_1 je, že chybovost je vyšší (takže H_0 a H_1 jsou doplňky).
   - Chyba I. druhu je, že reklamujeme stroje, které jsou dle specifikace.
   - Chyba II. druhu je, že nereklamujeme zmetky.
   - Zvolíme *hladinu významnosti* $\alpha = 0.05$. Tedy chybně reklamovat budeme jen v 5% případů.
   - Síla testu je pravděpodobnost, že zavrhneme H_0 pokud neplatí, je to $1 - \beta$.
   - Jinak řečeno:
     - $\Pr[\text{chyba I. druhu}] = \alpha = 0.05$
     - $\Pr[\text{chyba II. druhu}] = \beta$

   (b) Počet chyb modelujte přesně, tj. pomocí binomického rozdělení.

   Řešení: Ptáme se „jaká je pravděpodobnost, že uvidíme data, která jsme viděli, pokud H_0 platí?“

   Pokud chceme slovíčkařit:

   - Testovací kritérium T bude počet chybných pokusů. Tedy naše testovací kritérium se řídí rozdělením $Binom(600, 0.03)$ (povolili jsme největší možnou chybu kterou dovoluje H_0).
   - Kritické hodnoty budou ty hodnoty při kterých už zamítáme. Teda dost velké počty chyb $T \geq w$. Kde $W = \{w, w+1, \ldots, 600\}$.
   - Pro naše konkrétní hodnoty použijeme `scipy.stats.binom.cdf(w, 600, 0.03)` abychom zjistili pro které w dostaneme $w = 26$ (po chvilce hraní si). Nebo jsme mohli rovnou použít `scipy.stats.ppf(1-0.05, 600, 0.03) = 25.0`, pokud netrpíme kvantilofobií.

   Nebo rovnou můžeme spočítat pravděpodobnost, že jsme viděli 28 chyb v 600 pokusech, když pravděpodobnost chyby je nejvýš 5%:

   $$p = 1 - st.binom.cdf(27, 600, 0.03) = 0.0158 \quad (27 \text{ protože se ptáme na } \Pr[T \geq 28])$$

   (tomu se říká, že jsme spočítali p-hodnotu).

   Takže jdeme reklamovat.

   (c) Počet chyb modelujte přibližně pomocí normálního rozdělení (s vhodným μ, σ^2).

   Řešení: Počet chyb odhadneme pomocí CLV jako $N(600 \cdot 0.03, 600 \cdot 0.03 \cdot (1 - 0.03)) = N(18, 17.46)$ (aby to mělo stejnou střední hodnotu a rozptyl jako naše binomické rozdělení).

   $$p = 1 - st.norm.cdf(28, loc = 18, scale = math.sqrt(17.46)) = 0.00835$$

   (28 protože máme spojité rozdělení)

   Takže znovu jdeme reklamovat.

- Všimněte si, že kdybychom dopočítali směrodatnou odchylku $\sigma = \sqrt{17.46} = 4.1785$, což jde skoro z hlavy (je to 4 a kousek), tak pomocí pravidla 3σ můžeme odhad pravděpodobnosti udělat taky z hlavy.
- Všimněte si, že zde jsou kritické hodnoty celý interval reálných čísel.

3. **Vyzkoušejte, zda Python random funguje dobře. Pomocí** `random.choices([1, 2, 3, 4], k=20)` **jsme vygenerovali následující hody:**

$$[1,3,1,1,3,1,1,2,1,2,2,2,2,2,1,1,4,4,1,4]$$

tedy četnosti jsou:

$$\begin{aligned} X_1 &= 9 \\ X_2 &= 6 \\ X_3 &= 2 \\ X_4 &= 3 \end{aligned}$$

Průměrně mělo vyjít $E_j = \mathbb{E}[X_j] = 5$**. Otázkou je: „generuje python špatnou náhodu?“**

Řešení: Máme multinomické rozdělení (X_1, X_2, X_3, X_4) s parametry

- $n = 20$ je počet hodů
- $\vartheta = (\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4) = (0.25, 0.25, 0.25, 0.25)$

Pro připomenutí multinomické rozdělení je znamená, že n-krát opakuji pokus který může vyjít jedním z $k = 4$ možností a X_j je počet výsledků, které dopadly jako j (tedy binomické je speciální případ multinomického pro $k = 2$).

Pearsonova χ^2 statistika je

$$T = \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i} \qquad (X_i \text{ je pozorovaný})$$

$$= \frac{(9-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(3-5)^2}{5}$$

$$= 6$$

Dle věty platí že T se limitně blíží v distribuci k rozdělení χ^2_{k-1}.

Hypotéza H_0 je že generátor funguje dobře, tedy $E_i = 5$, H_1 je že je to rozdílné.

$$\gamma = F_Q^{-1}(1-\alpha) \qquad (\text{kde } Q \sim \chi^2_{k-1})$$

$$\gamma = scipy.stats.chisquare([9,6,2,3],[5,5,5,5])$$

$$= Power\_divergenceResult(statistic = 6.0, pvalue = 0.11161022509471268)$$

Takže hypotézu nezamítáme (tedy si pořád myslíme, že python generuje rozumnou náhodu).

4. **Náklon šikmé věže v Pise je měřen vzdáleností pevného bodu ve věži od jeho „správné“ polohy. V letech 1975 až 1987 tato poloha rostla následujícím způsobem: 2.9642, 2.9644, 2.9656, 2.9667, 2.9673, 2.9688, 2.9696, 2.9698, 2.9713, 2.9717, 2.9725, 2.9742, 2.9757. Proveďte lineární regresi, znázorněte i graficky.**

Řešení: Chceme „řešit“ soustavu:

$$\vartheta_1 x_1 + \vartheta_0 = y_1$$
$$\dots$$
$$\vartheta_1 x_n + \vartheta_0 = y_n$$

V našem konkrétním případě:

$$\vartheta_1 1975 + \vartheta_0 = 2.9642$$
$$\dots$$
$$\vartheta_1 1987 + \vartheta_0 = 2.9757$$

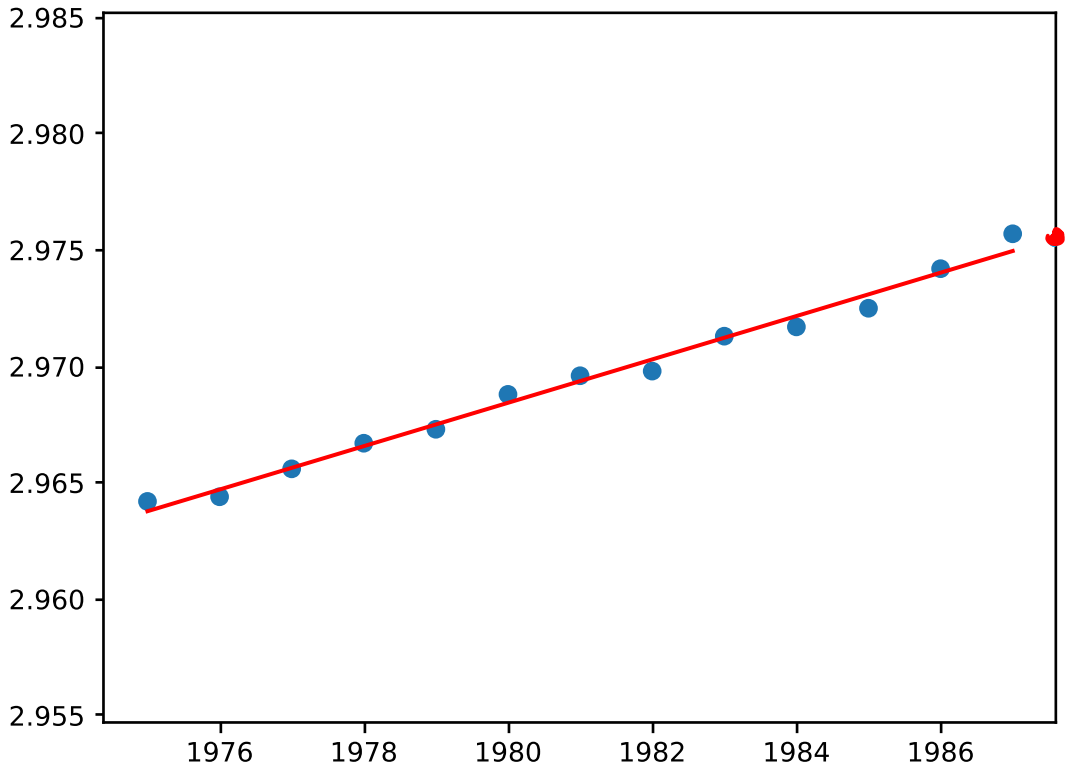
Kde „řešit“ znamená nalézt co nejlepší řešení. A „co nejlepší“ znamená minimalizovat mean square error (viz přednáška).

Máme:

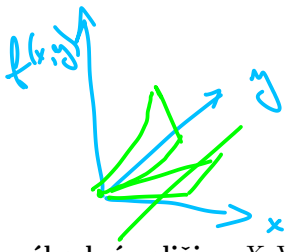
$$x = np.array(list(range(1975, 1988)))$$
$$y = np.array([2.9642, 2.9644, 2.9656, 2.9667, 2.9673, 2.9688, 2.9696, 2.9698, 2.9713, 2.9717, 2.9725, 2.9742$$
$$\overline{x} = (x_1 + \dots + x_n)/n = x.mean()$$
$$\overline{y} = (y_1 + \dots + y_n)/n = y.mean()$$
$$\hat{\vartheta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$vartheta_0 = \overline{y} - \hat{\vartheta}_1 \overline{x}$$

Pokud chcete prokládat jiným než lineárním polynomem, tak se koukněte do skript lineární algebry.

```
import matplotlib.pyplot as plt
import numpy as np

x = np.array(list(range(1975, 1988)))
y = np.array([2.9642, 2.9644, 2.9656, 2.9667, 2.9673, 2.9688, 2.9696, 2.9698, 2.9713, 2.9717

vartheta_1 = np.sum((x - x.mean()) * (y - y.mean())) / np.sum((x - x.mean())**2)
vartheta_0 = y.mean() - vartheta_1 * x.mean()

print(f'Náklon = {vartheta_0} + rok * {vartheta_1}')

# Výstup:
# Náklon = 1.1233384615384425 + rok * 0.0009318681318681415

plt.scatter(x, y, vmin=1975, vmax=1987)
plt.plot(x, [vartheta_0 + r * vartheta_1 for r in x], c='r')
# plt.show()
plt.savefig('fig/linearni_regrese.pdf')
```

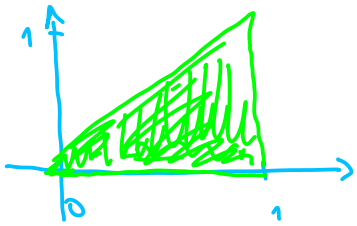


Obrázek 3.9: Lineární regrese: skutečné hodnoty jsou modré body, lineární regresí určená přímka červeně.



5. **Mějme náhodné veličiny X, Y které mají sdruženou hustotu (probability density function)**

$$f_{X,Y}(x,y) = \begin{cases} 8xy & x \in [0,1], 0 \le y \le x \\ 0 & \textbf{jinak} \end{cases}$$



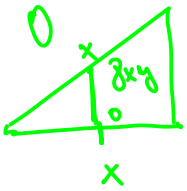
(všimněte si, kde je ta hustota nenulová!)

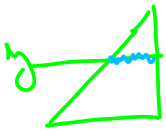
(a) Ověřte, že je to pravděpodobnostní hustota, tedy

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy = 1$$

Řešení: Přímo integrál venku přes y uvnitř přes x se některým z vás počítalo špatně, ale můžeme využít Fubiniho větu:

$$\begin{aligned}
\int_0^1\int_0^1 f_{X,Y}(x,y)\ dx\ dy &= \int_0^1\int_0^1 f_{X,Y}(x,y)\ dy\ dx && \text{(Fubini)} \\
&= \int_0^1\int_0^x 8xy\ dy\ dx && (\text{Pro } y > x \text{ je } f_{X,Y}(x,y) = 0) \\
&= \int_0^1 \left[4xy^2\right]_0^x\ dx \\
&= \int_0^1 4x^3\ dx \\
&= \left[x^4\right]_0^1 \\
&= 1 - 0 \\
&= 1
\end{aligned}$$





Nebo to spočítat přímo:

$$\begin{aligned}
\int_0^1\int_0^1 f_{X,Y}(x,y)\ dx\ dy &= \int_0^1\int_y^1 8xy\ dx\ dy \\
&= \int_0^1 \left[4x^2y\right]_y^1\ dy \\
&= \int_0^1 4y - 4y^3\ dy \\
&= \left[2y^2 - y^4\right]_0^1 \\
&= 1
\end{aligned}$$

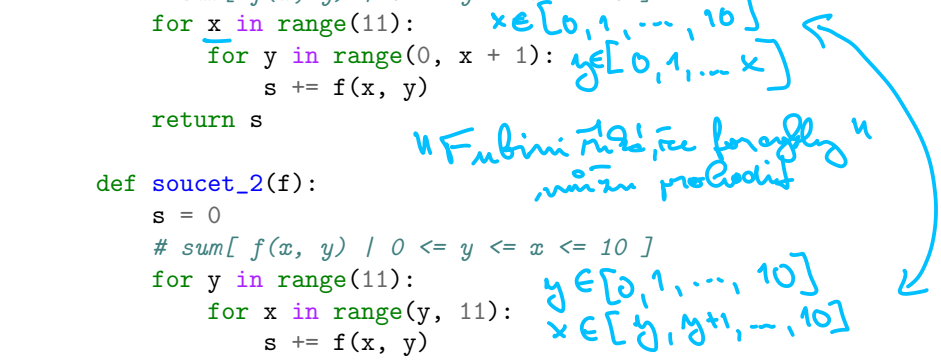
U integrálů prostě chcete "přijít na to jak to posčítat."

Je to podobné jako když chceme spočítat následující součet:

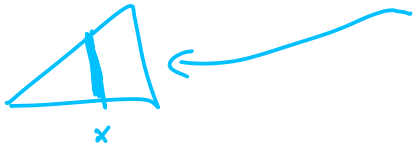
$$s = \sum \{f(x,y) \mid x, y \in \mathbb{N}, 0 \le y \le x \le 10\}$$

```
def f(x, y):
    return 8*x*y

def soucet_1(f):
    s = 0
```

```
    # sum[ f(x, y) | 0 <= y <= x <= 10 ]
    for x in range(11):
        for y in range(0, x + 1):
            s += f(x, y)
    return s

def soucet_2(f):
    s = 0
    # sum[ f(x, y) | 0 <= y <= x <= 10 ]
    for y in range(11):
        for x in range(y, 11):
            s += f(x, y)
    return s

print(soucet_1(f))
print(soucet_2(f))

# Vysledek:
# 13640
# 13640
```



(b) Spočítejte marginální hustoty



integrujeme přes y

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx$$

Řešení:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy \\ &= \int_0^x 8xy\ dy && \text{(pro } x \in [0,1]\text{)} \\ &= \left[4xy^2\right]_0^x \\ &= 4x^3 \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx \\ &= \int_y^1 8xy\ dx && \text{(jinde je hustota nulová)} \\ &= \left[4x^2y\right]_y^1 \\ &= 4y - 4y^3 \end{aligned}$$

(c) Určete distribuční funkce

$$F_X(x) = \Pr[X \le x] = \int_{-\infty}^{x} f_X(s)\ ds$$

$$F_Y(y) = \Pr[Y \leq y] = \int_{-\infty}^{y} f_Y(t)\ dt$$
$$F_{X,Y}(x,y) = \Pr[X \leq x \wedge Y \leq y] = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)\ dt\ ds$$

Řešení: Dosadíme rovnou spočítané marginální hustoty a ušetříme kousek práce.

$$\begin{aligned} F_X(x) = \Pr[X \leq x] &= \int_{-\infty}^{x} f_X(s)\ ds \\ &= \int_0^x 4s^3\ ds \\ &= \left[s^4\right]_0^x \\ &= x^4 \end{aligned}$$

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{y} f_Y(t)\ dt \\ &= \int_0^y 4t - 4t^3\ dt \\ &= \left[2t^2 - t^4\right]_0^y \\ &= 2y^2 - y^4 \end{aligned}$$

Pokud $0 \leq y \leq x$:

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)\ dt\ ds \\ &= \int_0^y\int_0^s f_{X,Y}(s,t)\ dt\ ds + \int_y^x\int_0^y f_{X,Y}(s,t)\ dt\ ds \\ &= \int_0^y\int_0^s 8st\ dt\ ds + \int_y^x\int_0^y 8st\ dt\ ds \\ &= \int_0^y \left[4st^2\right]_0^s\ ds + \int_y^x \left[4st^2\right]_0^y\ ds \\ &= \int_0^y 4s^3\ ds + \int_y^x 4sy^2\ ds \\ &= \left[s^4\right]_0^y + \left[2s^2y^2\right]_y^x \\ &= y^4 + 2(x^2 - y^2)y^2 \\ &= 2x^2 - y^4 \end{aligned}$$

Pokud $0 \leq x < y$:

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)\ dt\ ds \\ &= \int_0^x\int_0^s f_{X,Y}(s,t)\ dt\ ds \\ &= \int_0^x\int_0^s 8st\ dt\ ds \\ &= \int_0^x \left[4st^2\right]_0^s\ ds \end{aligned}$$

$$\begin{aligned}
&= \int_0^x 4s^3 \; ds \\
&= \left[s^4\right]_0^x \\
&= x^4
\end{aligned}$$

(d) Jsou X, Y nezávislé? Tedy platí pro každé $s, t \in \mathbb{R}$

$$F_{X,Y}(s,t) = F_X(s)F_Y(t)$$

Řešení: Proměnné jsou závislé.

(e) Spočítejte střední hodnotu náhodné veličiny $Z = X + 2Y$ pomocí LOTUS

$$g\colon \mathbb{R}^2 \to \mathbb{R} \qquad \text{(měřitelná)}$$

$$\mathbb{E}[Z] = \int_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y) \; dxy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) \; dx \; dy$$

Řešení:

$$\begin{aligned}
g\colon \mathbb{R}^2 \to \mathbb{R} & \qquad \text{(měřitelná)} \\
g(x,y) &= x + 2y \\
\mathbb{E}[Z] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) \; dx \; dy \\
&= \int_0^1 \int_y^1 g(x,y) f_{X,Y}(x,y) \; dx \; dy \\
&= \int_0^1 \int_y^1 (x+2y) 8xy \; dx \; dy \\
&= \int_0^1 \int_y^1 8x^2y + 16xy^2 \; dx \; dy \\
&= \int_0^1 \left[\frac{8}{3}x^3y + 8x^2y^2\right]_y^1 \; dy \\
&= \int_0^1 \frac{8}{3}y + 8y^2 - \frac{8}{3}y^4 - 8y^4 \; dy \\
&= \int_0^1 \frac{8}{3}y + 8y^2 - \frac{32}{3}y^4 \; dy \\
&= \left[\frac{4}{3}y^2 + \frac{8}{3}y^3 - \frac{32}{15}y^5\right]_0^1 \\
&= \frac{4}{3} + \frac{8}{3} - \frac{32}{15} \\
&= \frac{28}{15} \doteq 1.86666666
\end{aligned}$$

(f) Nechť A je jev $X + Y \leq 1$, spočítejte

$$\Pr[A] = \int_A f_{X,Y}(x,y) \; dxy$$

Řešení:

$$\begin{aligned}
\Pr[A] &= \int_0^{0.5}\int_0^{x} 8xy\ dy\ dx + \int_{0.5}^{1}\int_0^{1-x} 8xy\ dy\ dx \\
&= \int_0^{0.5} \left[4xy^2\right]_0^x\ dx + \int_{0.5}^{1} \left[4xy^2\right]_0^{1-x}\ dx \\
&= \int_0^{0.5} 4x^3\ dx + \int_{0.5}^{1} 4x(1-x)^2\ dx \\
&= \int_0^{0.5} 4x^3\ dx + \int_{0.5}^{1} 4x - 8x^2 + 4x^3\ dx \\
&= \left[x^4\right]_0^{0.5} + \left[2x^2 - \frac{8}{3}x^3 + x^4\right]_{0.5}^{1} \\
&= 0.5^4 + 2 - \frac{8}{3} + 1 - 0.5 + \frac{8}{3}0.5^3 - 0.5^4 \\
&= 2.5 - \frac{8}{3} + \frac{1}{3} \\
&= 1/6 \doteq 0.16666666
\end{aligned}$$

(g) Spočítejte podmíněnou hustotu $f_{X|A}$, distribuční funkci $F_{X|A}$ a střední hodnotu $\mathbb{E}[X \mid A]$:

$$\begin{aligned}
F_{X|A}(x) &= \Pr[X \leq x \mid A] = \frac{\Pr[X \leq x \wedge A]}{\Pr[A]} \\
F_{X|A}(x) &= \int_{-\infty}^{x} f_{X|A}(s)\ ds \\
\mathbb{E}[X \mid A] &= \int_{-\infty}^{\infty} x f_{X|A}(x)\ dx
\end{aligned}$$

Řešení: Pokud $x \leq 0.5$, pak $F_{X|A}(x) = F_X(x)$. Pokud $x > 0.5$, pak:

$$\begin{aligned}
F_{X|A}(x) &= \Pr[X \leq x \mid A] = \frac{\Pr[X \leq x \wedge A]}{\Pr[A]} \\
&= 6\left(\int_0^{0.5}\int_0^{s} 8st\ dt\ ds + \int_{0.5}^{x}\int_0^{1-s} 8st\ dt\ ds\right) \\
&= \ldots
\end{aligned}$$

$$F_{X|A}(x) = \int_{-\infty}^{x} f_{X|A}(s)\ ds$$

takže derivací dostaneme $f_{X|A}$.

$$\mathbb{E}[X \mid A] = \int_{-\infty}^{\infty} x f_{X|A}(x)\ dx$$

(h) Určete podmíněnou hustotu

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad \text{(pokud } f_Y(y) > 0\text{)}$$

Řešení:

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad \text{(pokud } 0 < y \leq x \leq 1\text{)} \\ &= \frac{8xy}{4y - 4y^3} \end{aligned}$$

(i) Určete podmíněnou distribuční funkci

$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(s \mid y)\ ds$$

Řešení: Zase jenom integrujeme, pokud $0 < y \leq x \leq 1$ (y je fixní):

$$\begin{aligned} F_{X|Y}(x \mid y) &= \int_{-\infty}^{x} f_{X|Y}(s \mid y)\ ds \\ &= \int_{0}^{x} \frac{8sy}{4y - 4y^3}\ ds \\ &= \left[\frac{4s^2y}{4y - 4y^3}\right]_0^x \\ &= \frac{4x^2y}{4y - 4y^3} \end{aligned}$$

(j) V tomto domácím úkolu nemusíte simulovat. Nejspíš byste vymysleli sami, jak to dělat. Ale ve skutečnosti to není zas tak jednoduché.

Řešení: Následující program by teoreticky měl fungovat, ale buď jsem někde udělal chybu nebo je numerická přesnost tohoto postupu k ničemu.

```
import math
from random import random

def XY():
    X = random() ** 0.25
    Y = math.sqrt(random() * (X**3) / 4)
    return (X, Y)

def XY2():
    # Vrací dvojici (X, Y) s hustotou 8xy na [0,1]x[0,x]
    # F_Y(y) = 2y^2 - y^4
    # Q_Y(p) = sqrt(1 - sqrt(1 - p))
    Y = math.sqrt(1 - math.sqrt(1 - random()))
    # F_X(x) = x^2 4 y / (4y - 4y^3)
    # Q_X(p) = sqrt(p (4y - 4y^3) / (4y))
    X = math.sqrt(random() * (4 * Y - 4 * (Y**3)) / (4 * Y))
    return (X, Y)
```

```python
N = 1_000_000

# E[Z] = E[X + 2Y] = 28/15
s = 0
for _ in range(N):
    xy = XY()
    s += xy[0] + 2 * xy[1]
print(f'E[Z] = {s/N} (={28/15})')

# Pr[X+Y <= 1]
pr = 0
for _ in range(N):
    xy = XY()
    pr += int(xy[0] + xy[1] <= 1)
print(f'Pr[X+Y<=1] = {pr/N} (={1/6})')
```

Na druhou stranu integrál nevidí množiny míry nula. Takže s pravděpodobností nula můžeme dostat jiné hodnoty pokud napřed samplujeme X a pak Y nebo opačně.