

(02) Výpočet pravdepodobností, náhodná premenná, distribučná funkcia

Mnohé veci sú opakovaním z predmetu Diskrétna matematika 1, ktorá má v sylabe aj:

- Úvod do diskkrétnej pravdepodobnosti.
- Experiment a náhodný jav.
- Bernoulliho schéma.
- Podmienená pravdepodobnosť.
- Bayesova veta.

Okrem toho budú v tejto prednáške aj nové veci.

1. Príklad 1

Príklad 1

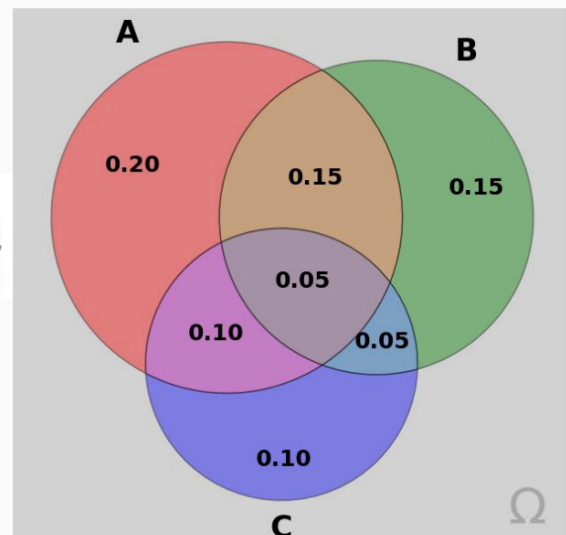
- Na konci minulej prednášky: udalosti ako množiny

A: na prvej križovatke svieti červená

B: na druhej križovatke svieti červená

C: na tretej križovatke svieti červená

- Pridáme k nim pravdepodobnosti
- Základný princíp: **Pravdepodobnosť zjednotenia disjunktných udalostí je súčet ich pravdepodobností**
- Vypočítame pp. udalostí z minulej prednášky a pp., že prejdeme všetky križovatky na zelenú



Ω : množina všetkých možných výsledkov

$$P(A \cup B) = 0,20 + 0,10 + 0,15 + 0,05 + 0,15 + 0,05 = 0,70$$

$$P(A \cap B) = 0,15 + 0,05 = 0,20$$

$$P(A^c) = 1 - P(A) = 1 - (0,20 + 0,10 + 0,15 + 0,05) = 1 - 0,5 = 0,5$$

Na prednáške minulý týždeň: červená na práve jednej križovke $\rightarrow$ pravd. = $0,20 + 0,15 + 0,10 = 0,45$

$$\begin{aligned} P(\text{všetky na zelenú}) &= 1 - P(\text{aspoň na jednej červená}) = \\ &= 1 - (0,20 + 0,10 + 0,15 + 0,05 + 0,15 + 0,05 + 0,10) = \\ &= 1 - 0,80 = 0,20 \end{aligned}$$

2. Príklad 2 (rytier de Méré)

Príklad 2 (rytier de Méré, 17. storočie)

HRA 1



Hádzame jednou kockou

Cieľ: hodiť šestku

4 pokusy

Vyhrávame, ak hodíme šestku aspoň raz

$p = 1/6$

HRA 2



Hádzame dvoma kockami

Cieľ: hodiť dve šestky

24 pokusy

Vyhrávame, ak hodíme dve šestky aspoň raz

$p = 1/36$

Rovnako pravdepodobné výsledky:

Pravdepodobnosť, že pri hádzaní kockou padne šestka, je $1/6$. Pravdepodobnosť, že pri hádzaní dvoma kockami padnú dve šestky, je $1/36$, teda šesťkrát menšia. Myšlienka je tá, že by sa to mohlo kompenzovať tým, že máme šesťkrát viac pokusov. Otázka je, či to tak naozaj výjde. K príkladu sa vrátíme, teraz sa pozrieme na časť jednej diskusie na internete.

https://www.reddit.com/r/probabilitytheory/comments/w7l0b0/chevalier_de_m%C3%A9r%C3%A9s_gambling_problem_from/

But can't you look at the problem in another way as the sum of the probability of throwing a 6 each time for up to four throws which would be:

$$4 \times 1/6 = 2/3$$

Or the sum of the probability of throwing a double six in 24 throws which is:

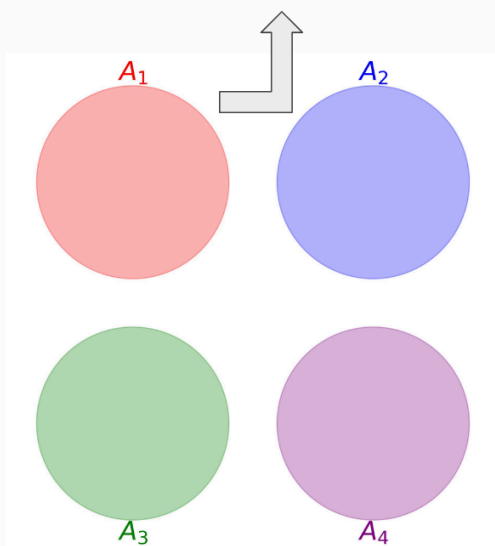
$$24 \times 1/36 = 2/3$$

Where is my thinking going wrong?

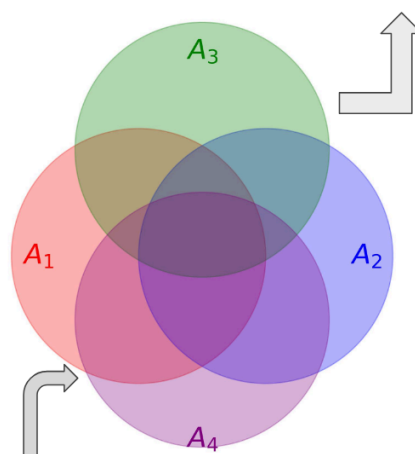


Ak by sme chceli len ukázať, že ten postup nemôže byť dobrý, mohli by sme si predstaviť, že hádžeme tou jednou kockou nie 4-krát, ale napríklad 7-krát. Podľa tejto úvahy by pravdepodobnosť bola $7/6$, čo je viac ako 1, a teda zjavne nesprávny výsledok. Ak chceme vedieť, *prečo* to nie je dobre:

$$P(A_1 \cup A_2 \cup A_3 \cup A_4) = P(A_1) + P(A_2) + P(A_3) + P(A_4)$$



$$P(A_1 \cup A_2 \cup A_3 \cup A_4) \neq P(A_1) + P(A_2) + P(A_3) + P(A_4)$$



$A_i = \{\text{na } i\text{-tej kocke padne šestka}\}$

Na redditu:



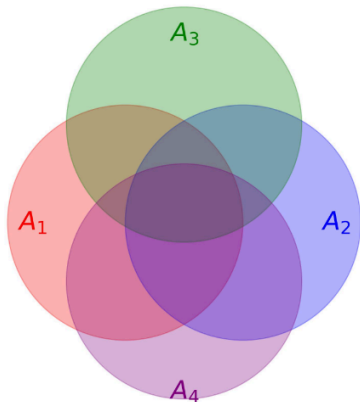
jamiesensei OP • 3y ago

Specifically, you've double-counted the chance of two sixes, triple-counted the chance of three and quadruple-counted the chance of four.

Ah! That's the piece I needed putting into words in order to understand this better.

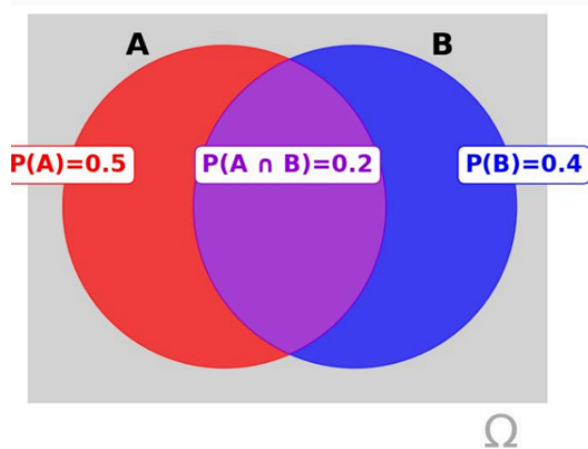
Thanks.

⊖ ↑ 1 ↓ ⌋ Reply ...



3. Príklad 3

Niekedy si treba potrebné pravdepodobnosti dopočítať.



Príklad 3

A : na prvej križovatke svieti červená
 B : na druhej križovatke svieti červená

Majme zadané:

- Pravdepodobnosť červenej na prvej križovatke je 0,5.
- Pravdepodobnosť červenej na druhej križovatke je 0,4.
- Pravdepodobnosť červenej na oboch križovatkách je 0,2.

Zaujímá nás:

- Aká je pravdepodobnosť, že aspoň na jednej križovatke bude červená?
- Aká je pravdepodobnosť, že obe križovatky prejdeme na zelenú?

$$P(\text{aspoň na jednej červena}) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0,5 + 0,4 - 0,2 = 0,7$$

$$P(\text{obe na zelenú}) = 1 - P(\text{aspoň na jednej červena}) = 1 - 0,7 = 0,3$$

4. Kombinatorická pravdepodobnosť

- Možné výsledky náhodného javu (pokusu a pod.) označme ω
- Množinu možných výsledkov označíme Ω (ako na obrázkoch)
- Náhodná udalosť A je podmnožina Ω

Klasický výpočet pravdepodobnosti:

Ak je Ω konečná a každá udalosť $\{\omega\}$ má rovnakú pravdepodobnosť, tak pravdepodobnosť udalosti A je $P(A) = |A|/|\Omega|$

Teda ak sú možné výsledky rovnako pravdepodobné, pravdepodobnosť udalosti je podiel počtu priaznivých výsledkov k počtu všetkých možných.

5. Príklad 4 (ilustrácia kombinatorickej pravdepodobnosti)



Výsledky {padla 1}, {padla 2}, {padla 3}, {padla 4}, {padla 5}, {padla 6} sú rovnako pravdepodobné

Príklad 4

Vypočítajte pravdepodobnosti udalostí:

$A = \{\text{padlo nepárne číslo}\}$

$B = \{\text{padlo číslo deliteľné tromi}\}$

Jednoducho len spočítame počet nepárnych čísel (3) a počet čísel deliteľných tromi (2).

$$P(\text{nepárne číslo}) = \frac{3}{6} = \frac{1}{2}$$

$$P(\text{číslo deliteľné tromi}) = \frac{2}{6} = \frac{1}{3}$$

6. Čo sú možné výsledky a množina všetkých výsledkov - príklad 5

Hádzeme tromi mincami, zapisujeme si, či padla hlava alebo znak. Vypíšeme si všetky možnosti a pozrieme sa na množinu, ktorá ich obsahuje,

```
import itertools
vsetky_vysledky = itertools.product(["H", "Z"], repeat=3)
for vysledok in vsetky_vysledky:
    print(vysledok)
```

Ω ω

('H', 'H', 'H')
 ('H', 'H', 'Z')
 ('H', 'Z', 'H')
 ('H', 'Z', 'Z')
 ('Z', 'H', 'H')
 ('Z', 'H', 'Z')
 ('Z', 'Z', 'H')
 ('Z', 'Z', 'Z')

každý prvok má rovnakú pravdepodobnosť

Príklad 5

Hádzeme troma mincami. Aká je pravdepodobnosť, že niekedy po sebe padnú dve hlavy?

Každý výsledok má rovnakú pravdepodobnosť, takže sa dá použiť kombinatorická pravdepodobnosť. Pretože možných výsledkov je málo, môžeme si ich vypísať a spočítať, čo potrebujeme.

$2^3 = 8$ možných výsledkov
 vyhovujú 3 (HHH, HHZ, ZHH) $\rightarrow$ pravd. = $\frac{3}{8}$

7. Príklad 5 - pokračovanie

Príklad 5 - pokračovanie

- Hádzeme desiatimi mincami. Aká je pravdepodobnosť, že niekedy po sebe padnú dve hlavy?
- Minimálne koľkými mincami musíme hádzať, aby táto pravdepodobnosť bola aspoň 0,95?
- Ako to zistiť bez vytvárania všetkých možných výsledkov?

Pre 10 mincí si situáciu znovu naprogramujeme. Výsledkov je veľa, preto ich nebudeme vypisovať, ale necháme počet priaznivých možností spočítať počítač.

Všimnime si, že na rozdiel od predchádzajúcej prednášky, kde sme videli simulácie (tie budú aj teraz), toto je presná hodnota. Nerobíme simuláciu náhodného pokusu, ale počítame počet priaznivých možností zo všetkých možných.

```

pocet_minci = 10
vsetky_vysledky = list(itertools.product(["H", "Z"], repeat=pocet_minci))

def test_HH(vysledok):
    return "HH" in "".join(vysledok)

test_HH_vsetky = [test_HH(vysledok) for vysledok in vsetky_vysledky]

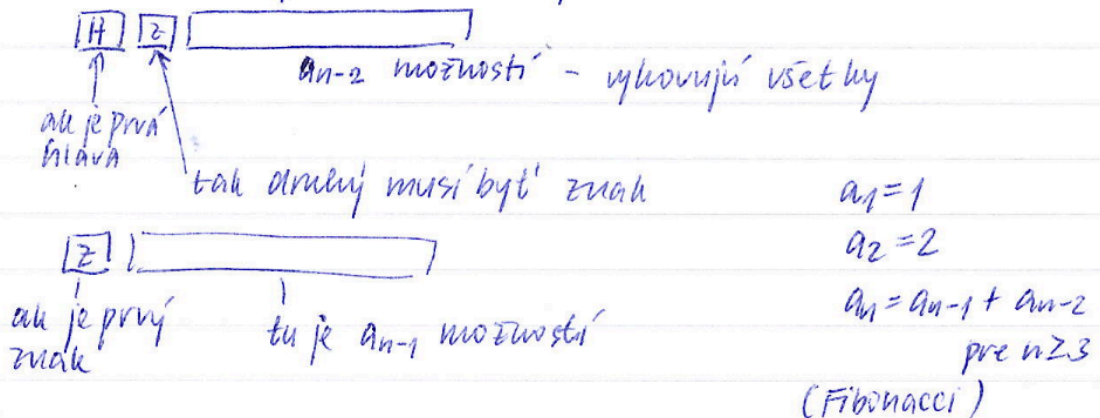
pocet_vsetkych = len(vsetky_vysledky) # |Omega|
pocet_vyhovujucich = sum(test_HH_vsetky) # |A|
pravdepodobnost = pocet_vyhovujucich/pocet_vsetkych # P(A) = |A|/|Omega|
print(pravdepodobnost)

```

0.859375

Vo všeobecnosti spravíme:

Pre všeobecný prípad budeme počítať počet možností bez HH
 n mincí: $n=1 \rightarrow$ vyhovujú: H, Z $\rightarrow$ 2 možnosti
 $n=2 \rightarrow$ vyhovujú: HZ, ZH, ZZ $\rightarrow$ 3 možnosti
 $n \geq 3 \rightarrow$ pzn. $a_n =$ počet možností pri n hodoch



Keď máme počet možností, kde nie je HH (teda počet nepriaznivých), môžeme vypočítať pravdepodobnosť tejto udalosti ako $a_n/2^n$. Tá hľadaná (pp. postupnosti bez HH) potom bude $1 - a_n/2^n$.

8. K predpokladu rovnakej pravdepodobnosti možných výsledkov

Príklad 1:

	2	3	4	5	6	7
	3	4	5	6	7	8
	4	5	6	7	8	9
	5	6	7	8	9	10
	6	7	8	9	10	11
	7	8	9	10	11	12
						

- {na prvej kocke 1, na druhej kocke 1}, {na prvej kocke 1, na druhej kocke 2}, ... **majú** rovnakú pravdepodobnosť
- {padol súčet 2}, {padol súčet 3}, {padol súčet 4}, ... **nemajú** rovnakú pravdepodobnosť

Príklad 2:

Toto poučenie je dôležité. Ak by sme ho nebrali na vedomie, mohli by sme spraviť nasledujúcu úvahu: „Sú dve možnosti. Buď dnes nájdem na ulici sto-eurovku, alebo nie. V jednom prípade z dvoch teda stoeurovku nájdem. Mal by som ju teda nájsť približne každý druhý deň.“ A ako nás presviedča každodenná prax, táto úvaha asi správna nebude.

Anino Belan: Kde sa vzala pravdepodobnosť a štatistika

9. Príklad 6 (o správnej voľbe výsledkov s rovnakou pravd.)

Hádzame dvoma kockami.

Čo je pravdepodobnejšie, že je súčet párny alebo že je nepárny?

- Nesprávna odpoveď by bola, že možných súčtov je 11, pričom párnych je 6 a nepárnych 5. Takže párny súčet má väčšiu pravdepodobnosť. To nie je správny postup, lebo vidíme, že súčty môžu nastať rôznymi spôsobmi a nie je ich rovnaký počet pre všetky súčty. Takže napríklad súčet 2 nemá rovnakú pravdepodobnosť ako súčet 7.

- Správna odpoveď je, že rovnakú pravdepodobnosť má každá usporiadaná dvojica - čo padne na prvej a čo na druhej kocke. Párny súčet dostaneme v 18 takýchto prípadoch, nepárny tiež v 18. Takže pravdepodobnosť párneho a nepárneho súčtu je rovnaká.

$$P(\text{párny súčet}) = \frac{1+3+5+5+3+1}{36} = \frac{18}{36} = \frac{1}{2}$$

$$P(\text{nepárny súčet}) = \frac{1}{2}$$

10. Príklad 7 (Bernoulliho schéma, nezávislé opakované pokusy)

- Adam, Boris, Cyril a Dominik dostali emailom newsletter z internetového obchodu. Pre každého je pravdepodobnosť, že ho otvorí rovnaká a rovná sa 0,45. Aká je pravdepodobnosť, že ho otvorí viac ako polovica z nich?
- Obchod posielal newsletter zákazníkom. Pravdepodobnosť, že ho zákazník otvorí, je 0,45. Newsletter odoslali 100 zákazníkom. Aká je pravdepodobnosť, že ho otvorí viac ako polovica z nich?
- Ako sa zmení pravdepodobnosť, ak tých zákazníkov bolo 1000?

Čo jediné potrebujeme vedieť:

n ľudí

p = pravdepodobnosť otvorenia

Ak sú otvorenia nezávislé, tak $P(\text{otvorí } k \text{ ľudí}) = \binom{n}{k} p^k (1-p)^{n-k}$

```
from scipy.special import comb # kombinacne cislo

def pravd_k_otvoreni(k, pocet_zakaznikov, p_otvori):
    return comb(pocet_zakaznikov, k, exact=True)*(p_otvori**k)*((1 - p_otvori)**(pocet_zakaznikov - k))

from math import floor # dolna cela cast

def pravd_aspon_polovica(pocet_zakaznikov, p_otvori):
    hranica = floor(pocet_zakaznikov/2) + 1 # minimalny vyhovujuci pocet
    p_moznosti = [pravd_k_otvoreni(k, pocet_zakaznikov, p_otvori) for k in range(hranica, pocet_zakaznikov + 1)]
    p_spolu = sum(p_moznosti)
    return p_spolu

print(f"Pri 4 zákazníkoch: {pravd_aspon_polovica(4, 0.45):.3g}")
print(f"Pri 100 zákazníkoch: {pravd_aspon_polovica(100, 0.45):.3g}")
print(f"Pri 1000 zákazníkoch: {pravd_aspon_polovica(1000, 0.45):.3g}")
```


11. Príklad 8 (nezávislé pokusy s rôznymi pravdepodobnosťami)

Adam, Boris, Cyril a Dominik dostali emailom newsletter z internetového obchodu. Pre každého je pravdepodobnosť, že ho otvorí iná:

- $P(\text{Adam otvorí newsletter}) = 0,45$
- $P(\text{Boris otvorí newsletter}) = 0,50$
- $P(\text{Cyril otvorí newsletter}) = 0,55$
- $P(\text{Dominik otvorí newsletter}) = 0,60$

Aká je pravdepodobnosť, že newsletter

- otvorí iba Cyril a Dominik?
- otvorí viac ako polovica z nich?

$$P(\text{iba C a D}) = P(A \text{ nie}, B \text{ nie}, C \text{ áno}, D \text{ áno}) =$$

↗
nezávislosť

$$= P(A \text{ nie}) \cdot P(B \text{ nie}) \cdot P(C \text{ áno}) \cdot P(D \text{ áno}) =$$

$$= 0,55 \cdot 0,5 \cdot 0,55 \cdot 0,6$$

$$P(\text{viac ako polovica}) = P(\text{otvorí } A, B, C, D) +$$

suštyria
→ 3 alebo 4

$$+ P(\text{otvorí } A, B, C, \text{neotvorí } D) +$$

$$+ P(\text{otvorí } A, B, D, \text{neotvorí } C) +$$

$$+ P(\text{otvorí } A, C, D, \text{neotvorí } B) +$$

$$+ P(\text{otvorí } B, C, D, \text{neotvorí } A) =$$

$$= 0,45 \cdot 0,50 \cdot 0,55 \cdot 0,60 + 0,45 \cdot 0,50 \cdot 0,55 \cdot 0,4 +$$

$$+ 0,45 \cdot 0,50 \cdot 0,45 \cdot 0,60 + 0,45 \cdot 0,50 \cdot 0,55 \cdot 0,60 +$$

$$+ 0,55 \cdot 0,50 \cdot 0,55 \cdot 0,60$$

12. Príklad 9 (podmienené pravdepodobnosti)

Adam, Boris, Cyril a Dominik dostali emailom newsletter z internetového obchodu.

- Adam, Boris, Cyril sa rozhodujú nezávisle a pre každého je pravdepodobnosť, že ho otvorí rovnaká a rovná sa 0,45.
- Dominik sa pred čítaním mailov rozprával s Cyrilom a s pravdepodobnosťou 0,9 zopakuje jeho rozhodnutie (otvoriť/neotvoriť mail s newsletterom).

Aká je pravdepodobnosť, že

- Dominik newsletter otvorí?
- newsletter otvorí viac ako polovica z nich?

Ako to funguje:

```
def rozhodnutie_dominika():
    p_cyril = 0.45
    p_rovnako = 0.9

    # rozhodnutie Cyrila
    cyril = random.choices(["otvori", "neotvori"], weights=[p_cyril, 1 - p_cyril])[0]

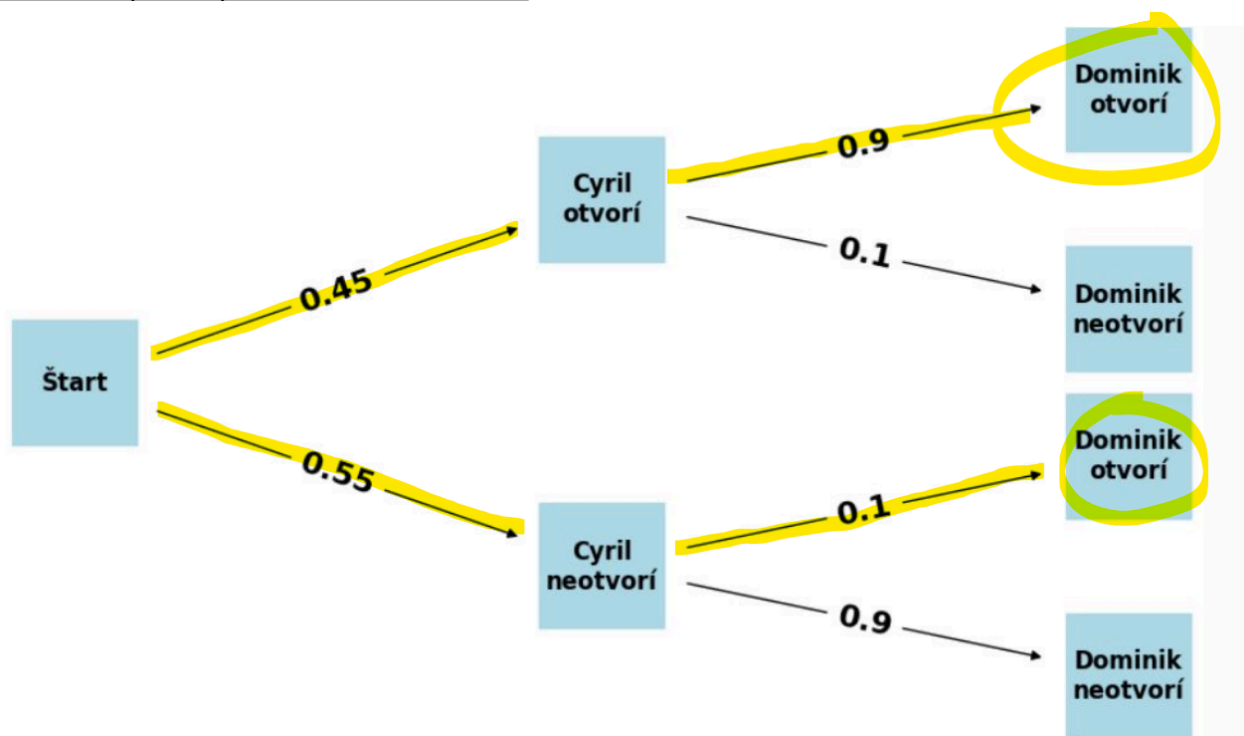
    # pravdepodobnost p_dominik, ze Dominik otvori, zavisí od Cyrila
    if cyril == "otvori":
        p_dominik = p_rovnako
    else:
        p_dominik = 1 - p_rovnako

    dominik = random.choices(["otvori", "neotvori"], weights=[p_dominik, 1 - p_dominik])[0]
    return dominik
```

Výpočet:

$$\begin{aligned} P(D \text{ otvorí}) &= \underbrace{P(D \text{ otvorí} | C \text{ otvorí})}_{0,9} \cdot \underbrace{P(C \text{ otvorí})}_{0,45} + \\ &+ \underbrace{P(D \text{ otvorí} | C \text{ neotvorí})}_{0,1} \cdot \underbrace{P(C \text{ neotvorí})}_{0,55} = \\ &= 0,9 \cdot 0,45 + 0,1 \cdot 0,55 = 0,46 \end{aligned}$$

Pomocou pravdepodobnostného stromu:



Pri otázke o pp. toho, že newsletter otvorí viac ako polovica z nich, musíme znovu spočítať pravdepodobnosti jednotlivých možností, ktoré tomu zodpovedajú.

Môžeme si to rozpísať alebo nakresliť strom a násobiť pravdepodobnosti (rozpísanie ukazuje, prečo výpočet pomocou stromu funguje):

Ozna. A : Adam otvorí
analogicky B, C, D

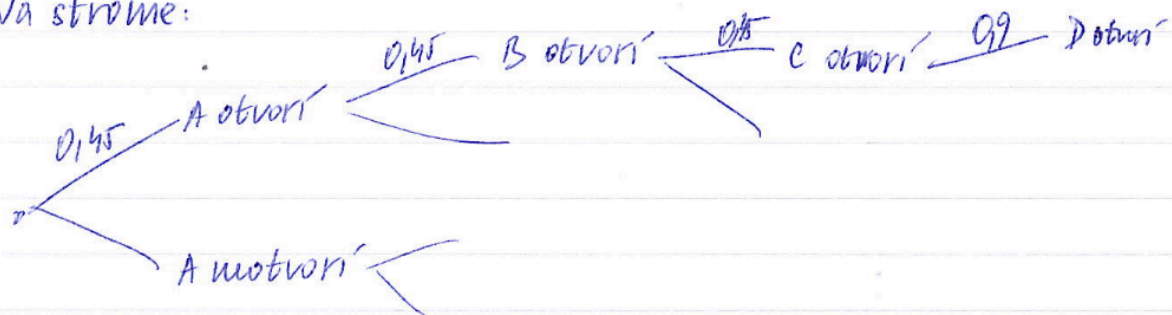
$$P(A \cap B \cap C \cap D) = \underbrace{P(A)}_{0,45} \cdot \underbrace{P(B)}_{0,45} \cdot \underbrace{P(C \cap D)}$$

nezávislé
→ násobíme
(ale C, D nie sú
nezávislé)

$$P(D|C) = \frac{P(D \cap C)}{P(C)}$$

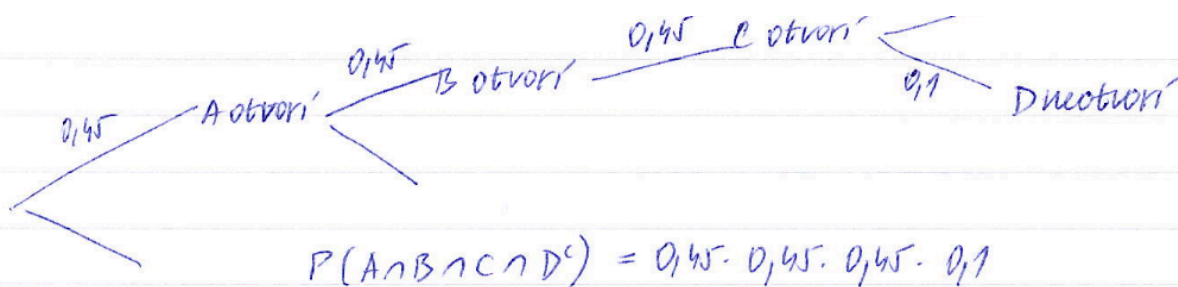
$$\Rightarrow P(C \cap D) = P(D|C) \cdot P(C) = 0,9 \cdot 0,45$$

Na strome:



$$P(A \cap B \cap C \cap D) = 0,45 \cdot 0,45 \cdot 0,45 \cdot 0,9$$

Analogicky ostatné:



$$P(A \cap B \cap C \cap D^c) = 0,45 \cdot 0,45 \cdot 0,45 \cdot 0,1$$

$$P(A \cap B \cap C^c \cap D) = 0,45 \cdot 0,45 \cdot 0,55 \cdot 0,1$$

$$P(A \cap B^c \cap C \cap D) = 0,45 \cdot 0,55 \cdot 0,45 \cdot 0,9$$

$$P(A^c \cap B \cap C \cap D) = 0,55 \cdot 0,45 \cdot 0,45 \cdot 0,9$$

13. Použitie podmienených pravdepodobností pri analýze dát: naivný Bayesov klasifikátor

Klasifikátor

- Máme dáta, chceme ich vedieť zatriediť do skupín.
- Časť dát na natréňovanie klasifikátora (napríklad odhadnutie parametrov), potom ho môžeme otestovať a používať
- Napríklad: Je tento mail spam alebo nie

Princíp Bayesovského klasifikátora:

- Triedy $C_1, C_2, \dots, C_k$
- Na dátach si všímame určité charakteristiky $X_1, \dots, X_m$
- Pre pozorovanie s hodnotami $x_1, \dots, x_m$ vypočítame pravdepodobnosti

$$P(\text{pozorovanie je z } C_k | X_1 = x_1 \cap \dots \cap X_m = x_m)$$

- Zaradíme ho do tej triedy, pre ktorú je táto pravdepodobnosť najvyššia

Všimnime si:

- Pravdepodobnosti $P(\text{pozorovanie je z } C_k | X_1 = x_1 \cap \dots \cap X_m = x_m)$ sú

$$\frac{P(\text{pozorovanie je z } C_k \cap X_1 = x_1 \cap \dots \cap X_m = x_m)}{P(X_1 = x_1 \cap \dots \cap X_m = x_m)}$$

- Menovatele sú rovnaké $\rightarrow$ na nájdenie maximálnej hodnoty stačí porovnávať čitatele
- Ak chceme pravdepodobnosti, vektor čitateľov stačí prenásobiť konštantou tak, aby bol súčet zložiek rovný 1.

Vo všeobecnosti potrebujeme súčin pravdepodobností:

- $\mathbb{P}(\text{pozorovanie je z } C_k)$
- $\mathbb{P}(X_1 = x_1 \mid \text{pozorovanie je z } C_k)$
- $\mathbb{P}(X_2 = x_2 \mid X_1 = x_1 \cap \text{pozorovanie je z } C_k)$
- $\mathbb{P}(X_3 = x_3 \mid X_1 = x_1 \cap X_2 = x_2 \cap \text{pozorovanie je z } C_k)$
- atď.

V čom spočíva naivný Bayesovský klasifikátor:

Naivný klasifikátor predpokladá:

- $\mathbb{P}(X_2 = x_2 \mid X_1 = x_1 \cap \text{pozorovanie je z } C_k)$ je rovnaká ako $\mathbb{P}(X_2 = x_2 \mid \text{pozorovanie je z } C_k)$
- $\mathbb{P}(X_3 = x_3 \mid X_1 = x_1 \cap X_2 = x_2 \cap \text{pozorovanie je z } C_k)$ je rovnaká ako $\mathbb{P}(X_3 = x_3 \mid \text{pozorovanie je z } C_k)$
- atď.

14. Naivný Bayesov klasifikátor - príklad použitia

Základná myšlienka, na webe je celý kód.

R: United States Congressional Voting Records 1984 ▾ Find in Topic

HouseVotes84 {mlbench} R Documentation

United States Congressional Voting Records 1984

Description

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes

Do Pythonu (po vynechaní NA) pomocou `rpy2.objects` → 281 riadkov

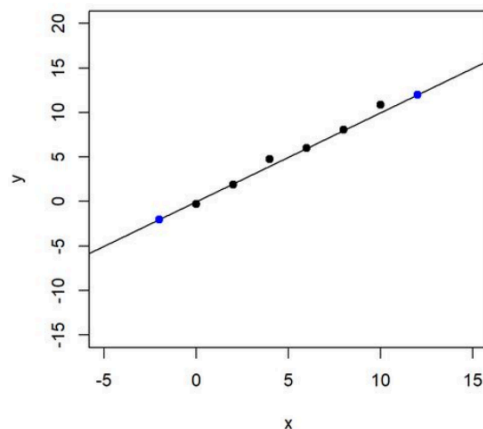
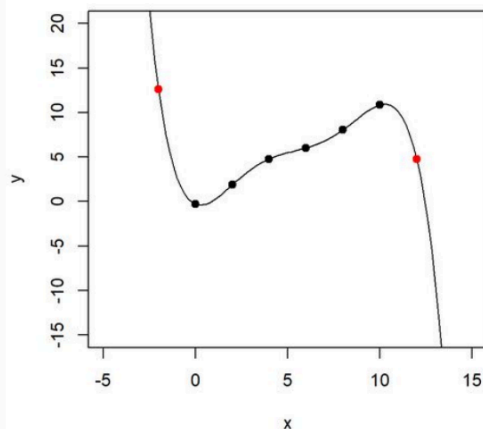
	Class	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
2	republican	n	y	n	y	y	y	n	n	n	n	n	y	y	y	n
6	democrat	n	y	y	n	y	y	n	n	n	n	n	n	y	y	y
9	republican	n	y	n	y	y	y	n	n	n	n	n	y	y	y	n
20	democrat	y	y	y	n	n	n	y	y	y	n	y	n	n	n	y
24	democrat	y	y	y	n	n	n	y	y	y	n	n	n	n	n	y

Teda vidíme, ako v týchto hlasovaniach jednotliví ľudia hlasovali a tiež to, či boli republikáni alebo demokrati. Naším cieľom bude na základe hlasovania vo vybraných prípadoch odhadnúť, či ide o republikána alebo o demokrata.

Odbočka:

Rozdelíme dáta na **trénovaciú a testovaciu časť**.

Poznámka: Prečo nie je cieľom len presne nafitovať dáta a potom to použiť na predikcie:



Dáta môžeme dokonale nafitovať, ale predikcie môžu byť úplne mimo.

Naspäť:

Vyberieme si na ukážku hlasovania 1, 8 a 15 a budeme robiť klasifikátor pomocou nich.

Funkcia **train_test_split**

- Hodnota 0.2 znamená, že testovacie dáta budú tvoriť 20 našich dát.
- Parameter **y** pri stratifikácii znamená, že chceme mať v trénovacích dáta rovnako často často zastúpené jednotlivé hodnoty tejto premennej, ako v celom súbore.
- Nastavenie **random_state** je kvôli reprodukovateľnosti (rozdelenie je náhodné)
- **X, y** je zaužívané značenie

Na trénoch dát vytvoríme model, predikujeme **Class** (republikán, demokrat) pomocou hlasovaní 1, 8, 15

```
from sklearn.model_selection import train_test_split

# X - premene ktore pouzijeme na predikovanie
# y - premenna ktoru budeme predikovat
X = votes[["V1", "V8", "V15"]]
y = votes["Class"]

# train/test
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    stratify=y,
    random_state=42
)
```

Potom predikujeme hodnoty v testovacej časti, napríklad:

```
print(X_test.head())
```

	V1	V8	V15
266	y	y	n
254	n	n	n
255	y	y	n
341	n	n	n
238	y	y	n

Z tabuliek početností v trénoch dát spočítame pravdepodobnosti ručne a overíme vo výstupe z Pythonu.

```
=== Class ===
Class
democrat      132
republican     92
Name: count, dtype: int64
```

```
=== Class vs. V1 ===
V1           n    y
Class
democrat     54   78
republican   76   16
```

```
=== Class vs. V8 ===
V8           n    y
Class
democrat     26  106
republican    79   13
```

```
=== Class vs. V15 ===
V15          n    y
Class
democrat     49   83
republican    82   10
```

Spočítajme pravdepodobnosti, že ide o republikána, resp. demokrata pre prvý údaj z testovacej časti dát:

```
print(X_test.head())
```

	V1	V8	V15
266	y	y	n
254	n	n	n

Pravdepodobnosť, že je to demokrat - budeme upravovať len čitateľ zlomku:

$$P(D | (V_1=Y) \cap (V_8=Y) \cap (V_{15}=N)) = \frac{P(D \cap (V_1=Y) \cap (V_8=Y) \cap (V_{15}=N))}{\dots}$$

$$\text{čitateľ} = P(D) \cdot P(V_1=Y | D) \cdot P(V_8=Y | D \cap (V_1=Y)) \cdot P(V_{15}=N | D \cap (V_1=Y) \cap (V_8=Y))$$

$$\Rightarrow P(D) \cdot P(V_1=Y | D) \cdot P(V_8=Y | D) \cdot P(V_{15}=N | D) =$$

predpoklad
naivného
klasifikátora

$$\Rightarrow \frac{132}{224} \cdot \frac{78}{132} \cdot \frac{106}{132} \cdot \frac{49}{132} = \frac{78 \cdot 106 \cdot 49}{224 \cdot 132 \cdot 132}$$

z tabuliek

Podobne, že je to republikán:

$$P(R | (V_1=Y) \cap (V_8=Y) \cap (V_{15}=N)) = \frac{P(R \cap (V_1=Y) \cap (V_8=Y) \cap (V_{15}=N))}{\dots}$$

$$\text{čitateľ} = \frac{92}{224} \cdot \frac{16}{92} \cdot \frac{13}{92} \cdot \frac{82}{92} = \frac{16 \cdot 13 \cdot 82}{224 \cdot 92 \cdot 92}$$

Máme čitatele zlomkov, menovateľ je spoločný a taký, že v súčte dostávame jednotku. Preto pravdepodobnosti sú:

$$P(D | (V_1=Y) \cap (V_8=Y) \cap (V_{15}=N)) = \frac{\text{čitateľ}1}{\text{čitateľ}1 + \text{čitateľ}2} = 0,920245$$

$$P(R | (V_1=Y) \cap (V_8=Y) \cap (V_{15}=N)) = \frac{\text{čitateľ}2}{\text{čitateľ}1 + \text{čitateľ}2} = 0,079755$$

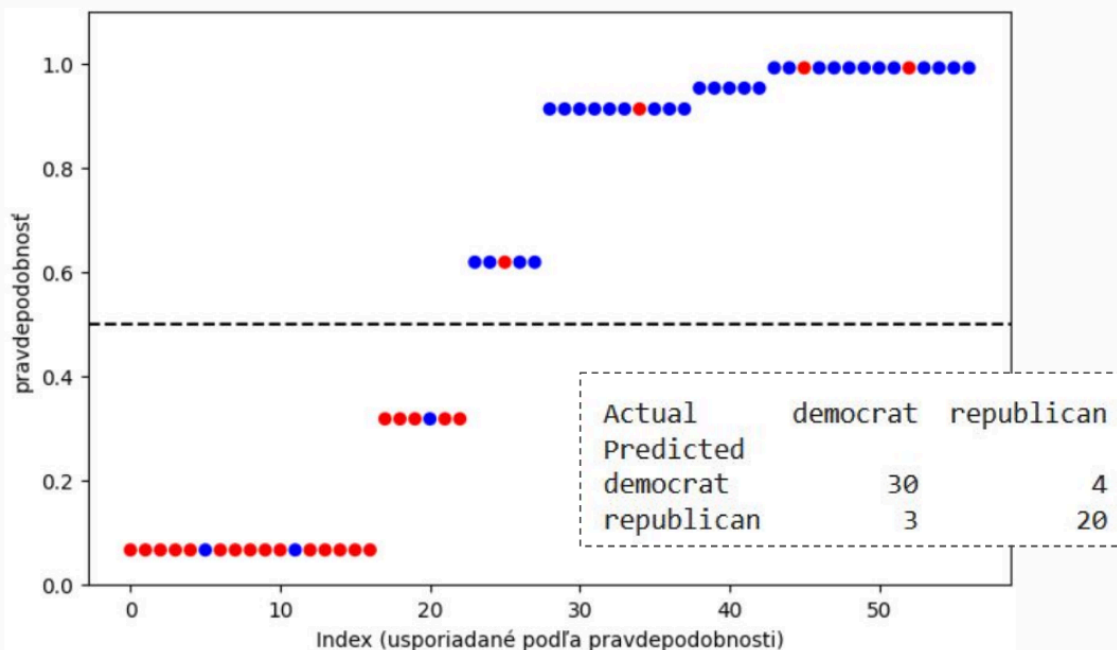
Na kontrolu výstup z Pythonu:

```

democrat republican
0 0.920245 0.079755
1 0.063564 0.936436
2 0.920245 0.079755

```

Zhodnotenie modelu:



15. Príklad 10 (Bayesova veta v medicíne)

- Pacient má niektorú z chorôb A, B, C. Na ich rozlíšenie sa používa osem testov. Testy sú nezávislé, percento pozitívnych výsledkov pri jednotlivých chorobách a testoch je nasledovné:

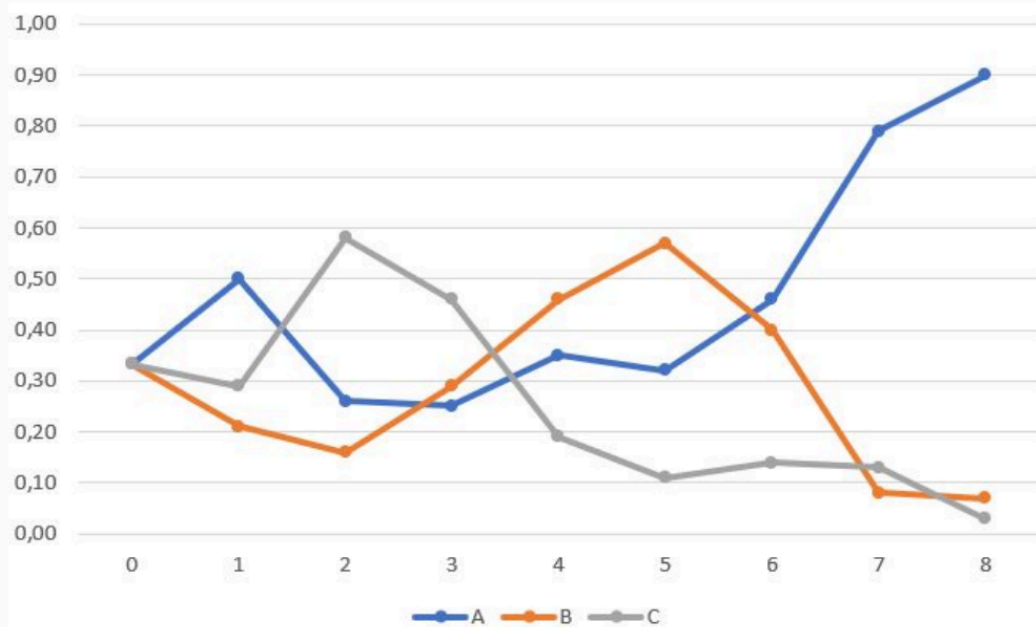
		Test							
		1	2	3	4	5	6	7	8
Choroba	A	30	20	50	70	40	80	10	60
	B	70	30	90	80	20	40	90	50
	C	60	80	40	20	60	70	50	10

Teda napríklad $P(\text{test 1 pozitívny} \mid \text{choroba A}) = 0,30$.

- Na začiatku dávame každej chorobe rovnakú pravdepodobnosť.
- Výsledky testov:

Test	1	2	3	4	5	6	7	8
Výsledok	-	+	+	+	-	+	-	+

- Pravdepodobnosti jednotlivých chorôb po každom teste:



Na ukážku si spočítame, pravdepodobnosť choroby A po prvom vykonanom teste:

$$P(A | \text{test 1-}) = \frac{P(A \cap (\text{test 1-}))}{P(\text{test 1-})} = \frac{P(\text{test 1-} | A) \cdot P(A)}{P(\text{test 1-})} =$$

$$= \frac{P(\text{test 1-} | A) \cdot P(A)}{P(\text{test 1-} | A)P(A) + P(\text{test 1-} | B)P(B) + P(\text{test 1-} | C)P(C)} =$$

$$= \frac{0.7 \cdot \frac{1}{3}}{0.7 \cdot \frac{1}{3} + 0.3 \cdot \frac{1}{3} + 0.4 \cdot \frac{1}{3}} = 0.5$$

16. Štatistická pravdepodobnosť

Nie všetkému vieme pravdepodobnosť vypočítať, niekedy sa musí odhadnúť z pozorovaných dát. Napríklad:

- V predchádzajúcom príklade citlivosť medicínskych testov.
- Pravdepodobnosť narodenia chlapca a dievčaťa.

A pod.

17. Geometrická pravdepodobnosť

- Ak má množina všetkých možných výsledkov má nekonečne veľa prvkov, nedá sa použiť klasická pravdepodobnosť.
- Pri "rovnomernosti" možných výsledkov na množine Ω sa dá použiť geometrická pravdepodobnosť: $P(A) = \text{plocha}(A)/\text{plocha}(\Omega)$

18. Geometrická pravdepodobnosť - ilustračný príklad

Príklad 11

Na Zem dopadne meteorit. Je malý, takže ho v porovnaní s povrchom Zeme budeme považovať za bod. O mieste jeho dopadu nič nevieme, pravdepodobnosť, že padne na určitú územie je teda úmerná jeho ploche. Aká je pravdepodobnosť, že dopadne na súš?

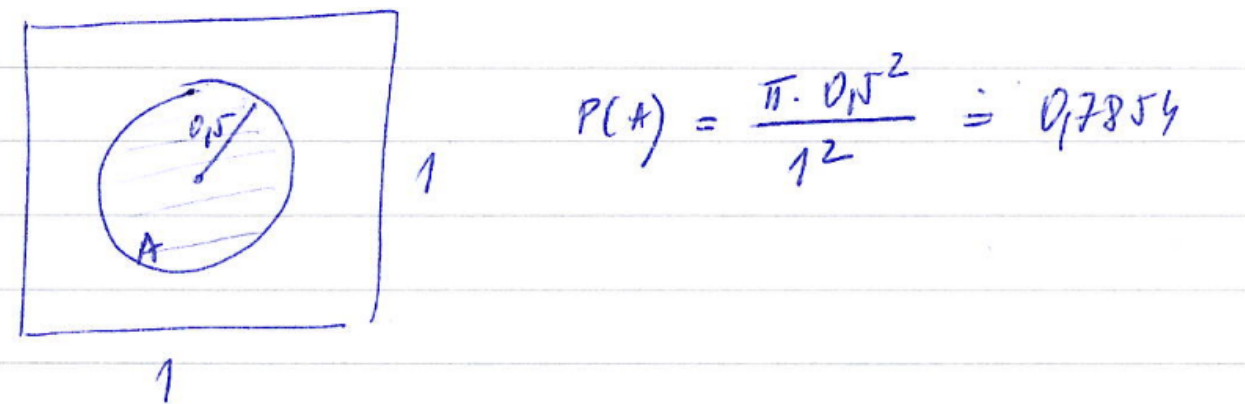
Wikipédia: Voda na Zemi zaberá asi 2/3 celkového povrchu.

Riešenie:

Súš tvorí približne $\frac{1}{3}$ povrchu, takže pravdepodobnosť je približne $\frac{1}{3}$.

19. Príklad 12

V štvorci so stranou dĺžky 1 zvolíme bod tak, že pravdepodobnosť jeho padnutia do nejakého útvaru je úmerná jeho obsahu (budeme hovoriť "bod zvolíme v tomto štvorci "rovnomerne náhodne"). Aká je pravdepodobnosť, že jeho vzdialenosť od stredu štvorca bude menej ako 0,5?



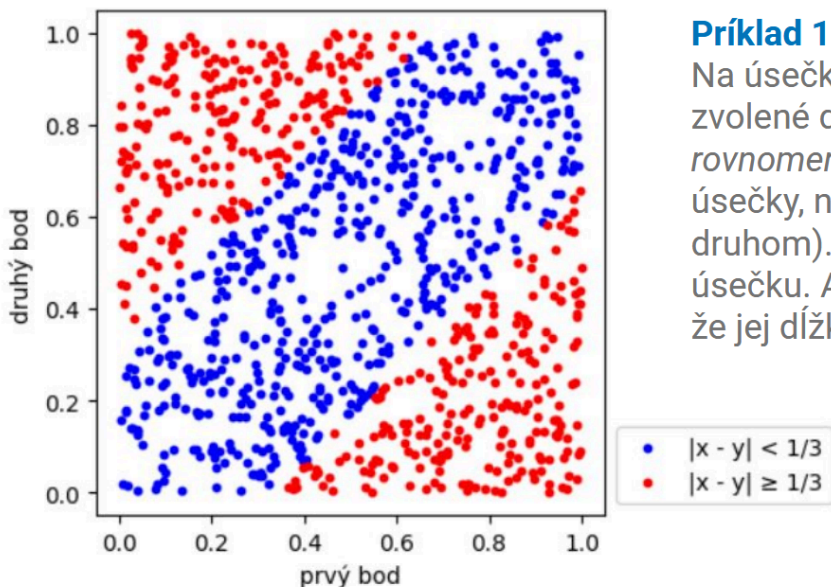
```
import numpy as np

def bod_v_stvorci():
    x = np.random.rand() # rovnomerne medzi 0 a 1
    y = np.random.rand()
    vzdialenost_od_stredu = np.sqrt((x - 0.5)**2 + (y - 0.5)**2)
    return vzdialenost_od_stredu < 0.5

pocet_pokusov = 10**5
simulacia = [bod_v_stvorci() for _ in range(pocet_pokusov)]
print(sum(simulacia)/pocet_pokusov)
```

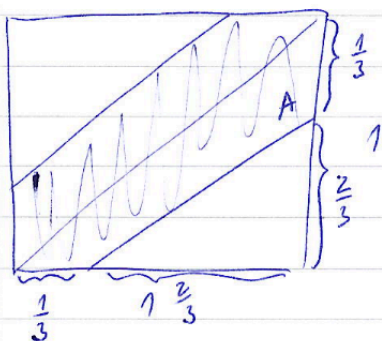
0.78548

20. Príklad 13



Príklad 13

Na úsečke dĺžky 1 sú náhodne zvolené dva body (každý rovnomerne náhodne na celej dĺžke úsečky, nezávisle jeden na druhom). Tieto dva body vytvoria úsečku. Aká je pravdepodobnosť, že jej dĺžka je menej ako $1/3$?



$$\text{obsah každého trojuholníka} \\ = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{2}{3}$$

$$\rightarrow 2 \text{ trojuholníky} : \frac{2}{3} \cdot \frac{2}{3} = \frac{4}{9}$$

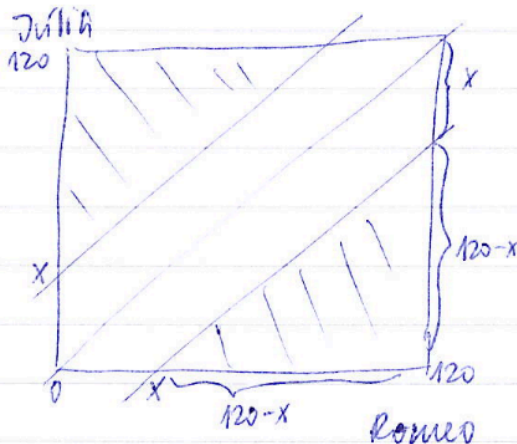
$$\rightarrow \text{plocha } A = 1 - \frac{4}{9} = \frac{5}{9}$$

$$P(A) = \frac{\frac{5}{9}}{1} = \frac{5}{9}$$

21. Príklad 14

Romeo a Júlia sa chcú stretnúť. Na miesto stretnutia prídu v náhodnom čase, ktorý si každý nezávisle zvolí *rovnomerne náhodne* medzi 18:00 a 20:00. Kto príde prvý, počká na druhého. Aká je pravdepodobnosť, že ten, kto príde prvý,

- bude čakať viac ako 15 minút
- nebude čakať viac ako x minút (pre ľubovoľné reálne číslo x)



zapisujeme čas príchodu
v minútach po 18:00

pre $x \in [0, 120]$:

$$P(\text{čakať viac ako } x \text{ min}) =$$

$$= \frac{2 \cdot \left(\frac{1}{2} \cdot (120-x)^2 \right)}{120^2} =$$

polov trojuholník

$$= \frac{(120-x)^2}{120^2}$$

$$a) P(\text{viac ako 15 min}) = (120-15)^2 \cdot \frac{1}{120^2}$$

$$b) P(\text{nebude čakať viac ako } x \text{ min}) = \begin{cases} 0 & \text{pre } x < 0 \\ 1 - \frac{(120-x)^2}{120^2} & \text{pre } x \in [0, 120] \\ 1 & \text{pre } x > 120 \end{cases}$$

22. Náhodná premenná

Náhodná premenná

- Funkcia X , ktorá prvku ω z množiny Ω priradí reálne číslo.
- Napríklad X = doba čakania z prechádzajúceho príkladu v minútach
 - Pre ω = **Romeo prišiel o 18:15:00, Júlia prišla o 18:45:30** je doba čakania 30 min. 30 s, takže **$X = 30,5$**

23. Distribučná funkcia náhodnej premennej

Distribučná funkcia náhodnej premennej

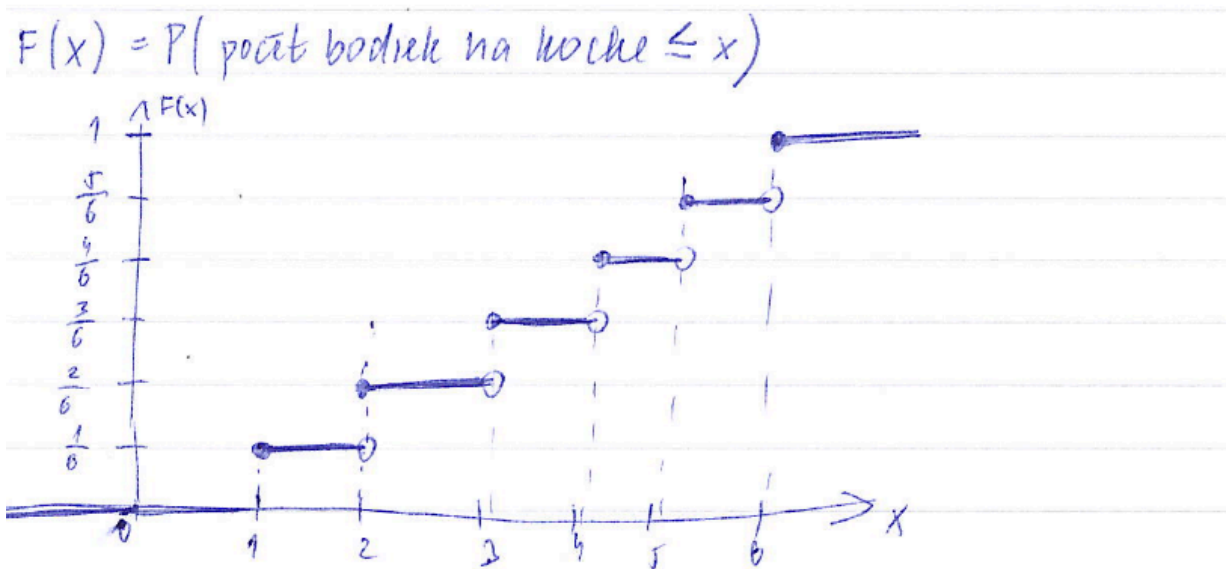
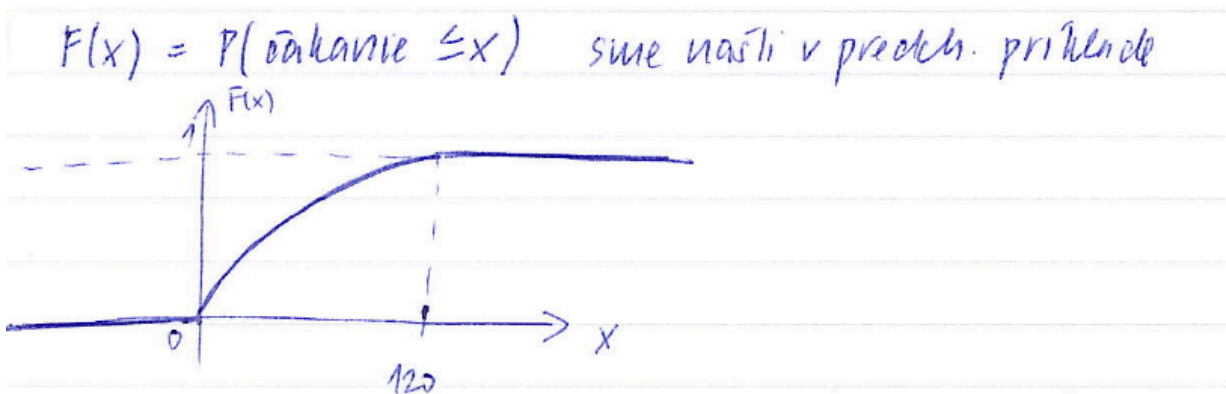
- Funkcia s definičným oborom $\mathbb{R}$, a hodnotami z $[0, 1]$.
- Pre náhodnú premennú X sa jej distribučná funkcia definuje ako

$$F(a) = P(X \leq a)$$

24. Príklad 15 - výpočet distribučnej funkcie

Nájdeme distribučnú funkciu

- doby čakania z predchádzajúceho príkladu
- náhodnej premennej, ktorá označuje počet bodov, ktoré padnú pri hodení kockou



25. Empirická distribučná funkcia

- Počíta sa pre sadu číselných dát
- Pravdepodobnosť $P(X \leq a)$ z definície distribučnej funkcie sa nahradí podielom hodnôt, ktoré tejto nerovnosti vyhovujú.
- Často sa kreslí ako schodovitá funkcia

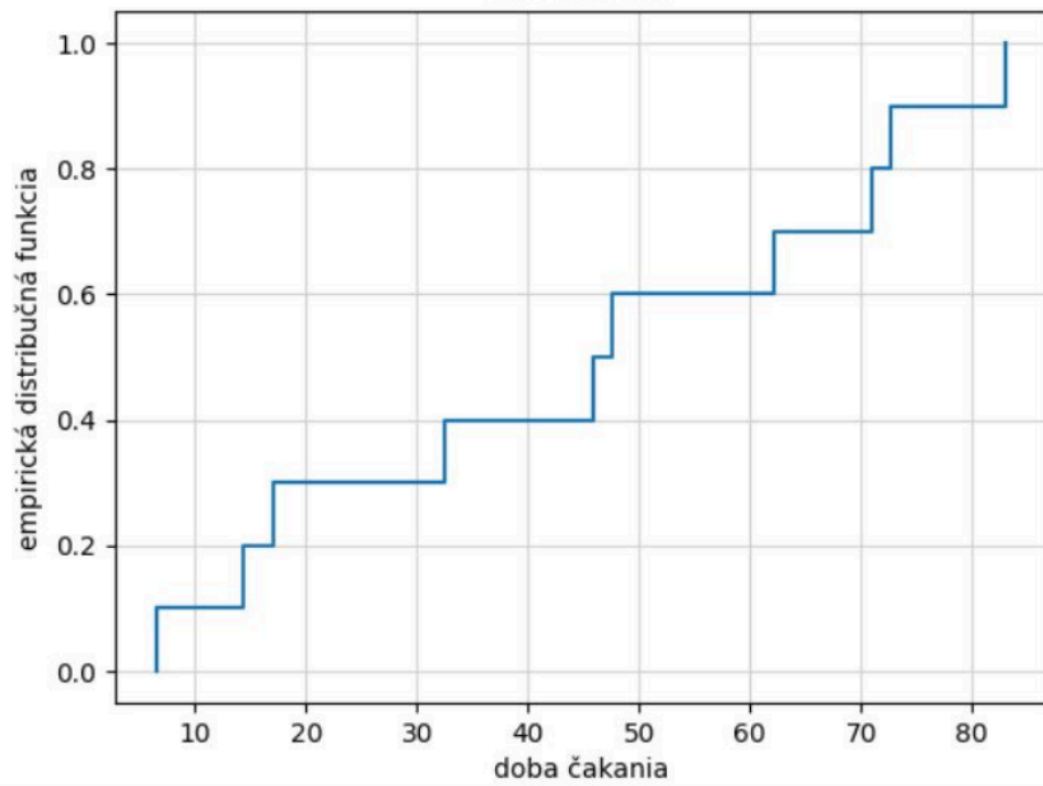
```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.distributions.empirical_distribution import ECDF

# data
pocet_simulacii = 10**2
romeo = np.random.rand(pocet_simulacii)*120 # rovnomerne medzi 0 a 120
julia = np.random.rand(pocet_simulacii)*120
cakanie = np.abs(romeo - julia)

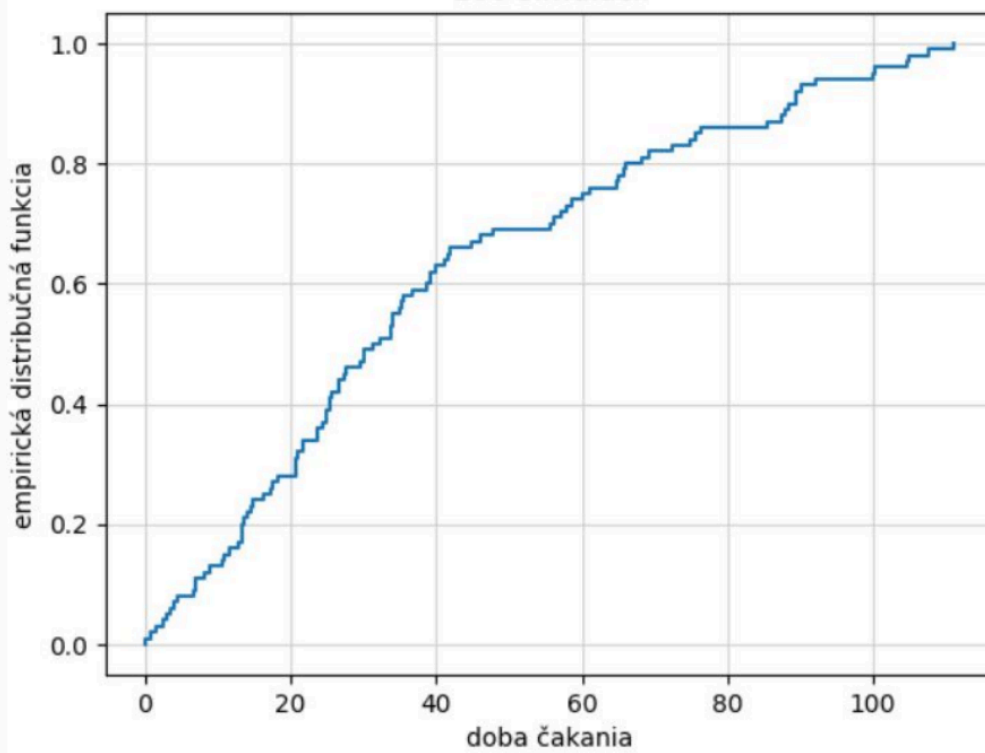
# CDF = cumulative distribution function = distribucna funkcia
# ECDF = empirical CDF
ecdf = ECDF(cakanie)

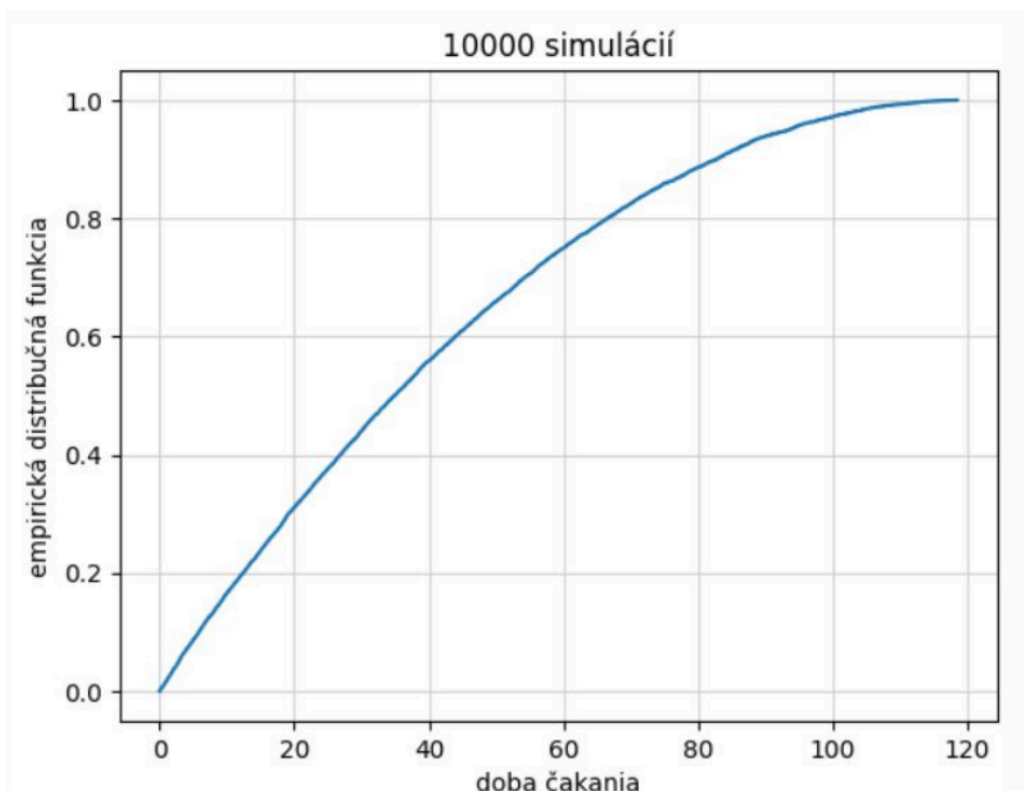
# graf
plt.step(ecdf.x, ecdf.y, where="post")
plt.xlabel("doba čakania")
plt.ylabel("empirická distribučná funkcia")
plt.title(f"{pocet_simulacii} simulácií")
plt.grid(color="lightgrey")
plt.show()
```

10 simulácií

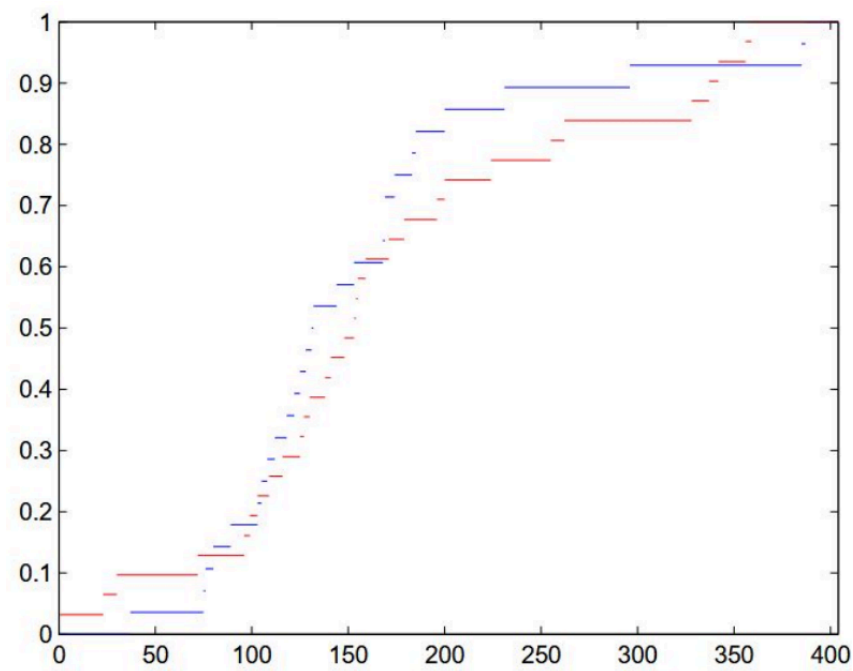


100 simulácií





26. Empirická distribučná funkcia: aplikácia



Katarína Kocsisová:
Experimenty a
vlastné dáta pri
vyučovaní
pravdepodobnosti a
štatistiky (bakalárska
práca, 2015)

- Porovnanie času čakania na obed v dvoch jedálňach na internátoch
- Pri testovaní zhody pomocou Kolmogorovovho-Smirnovovho testu je podstatný maximálny rozdiel medzi empirickými distribučnými funkciami.

Pýtame sa teda, či môžeme čas čakania na obed (ktorý je, samozrejme, náhodný) v dvoch jedálňach považovať za rovnaký. Odpoveď je *nie*, ak sa empirické distribučné funkcie “veľmi líšia”. Podrobnejšie o tom budeme hovoriť pri téme testovania štatistických hypotéz.