Domácí úkol

Dan Kostiuk, Oliver Tulša, Štefan Slavkovský

M = 5

1) Pro každou skupinu zvlášť odhadněte střední hodnotu, rozptyl a medián příslušného rozdělení.

V úlohe budeme skúmať účinky iodidu strieborného pri použitý v oblakoch na celkové zrážky na plochu. Skúmané data pozostávajú z dvoch sád meraní. Prvá kontrolná sada (unseeded) obsahuje merania zrážok z dní kedy nebol použity iodid strieborný. Druha sada (seeded) obsahuje merania z dní kedy iodid strieborný bol použitý.

Pro bodový odhad střední hodnoty použijeme výběrový průměr:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Bodový odhad výběrového rozptylu spočteme následovně:

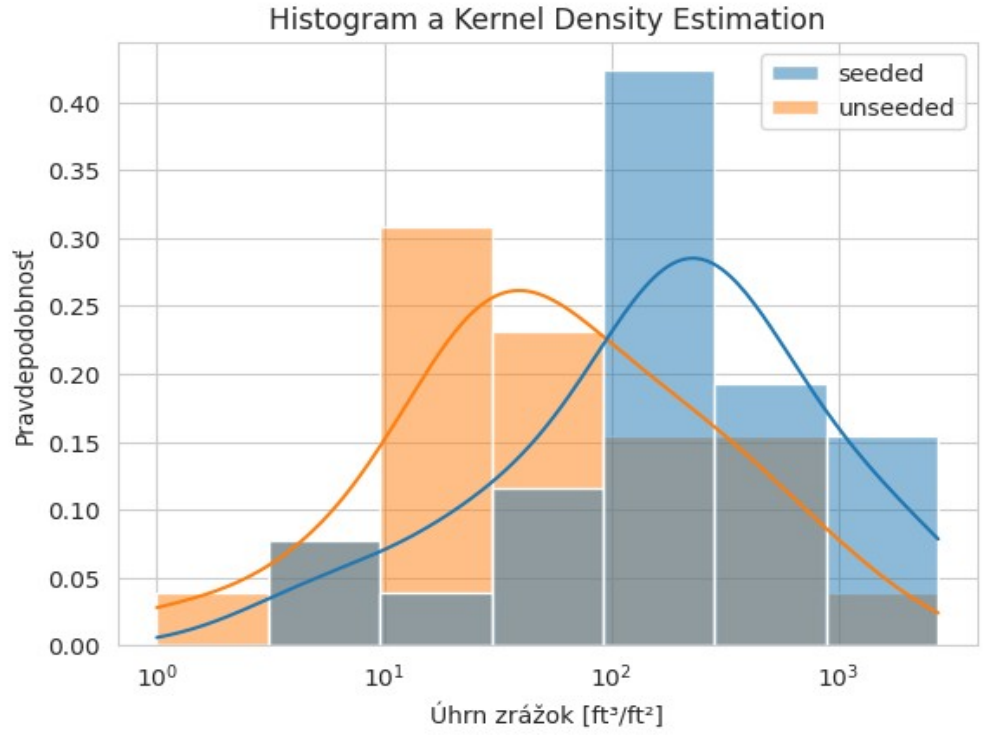
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Medián odhadneme jako prostřední hodnotu v našich seřazených datech.

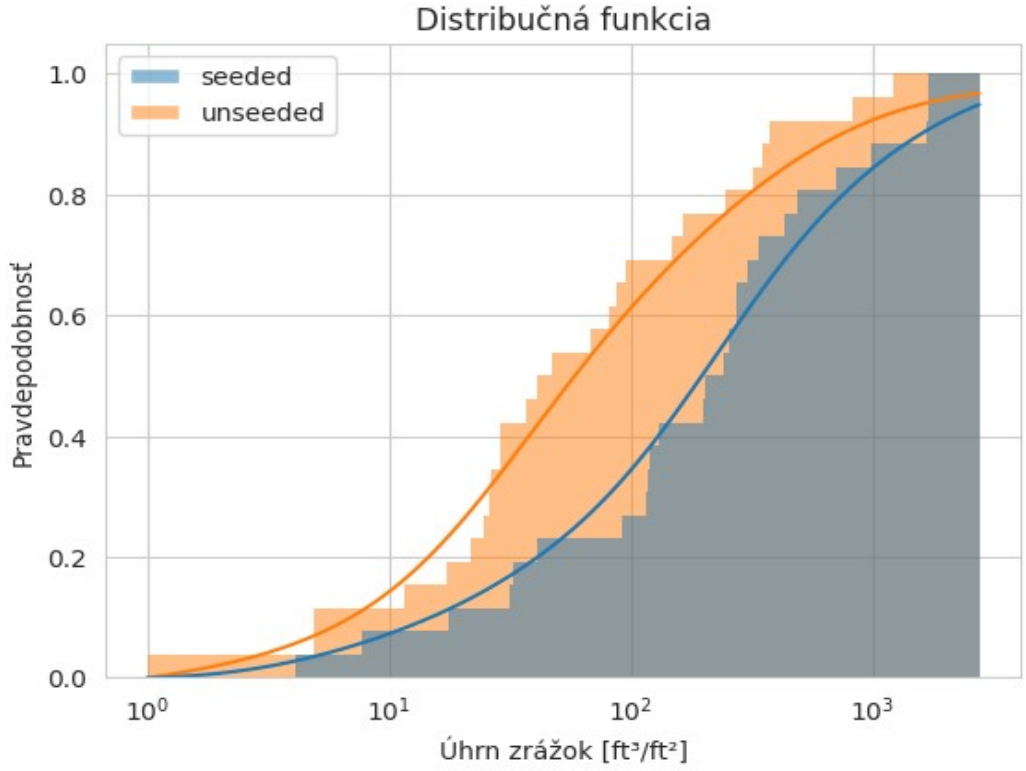
	Střední hodnota	Rozptyl	Median
Seeded	441.98	423523.96	221.60
Unseeded	164.59	77521.25	44.20

2) Pro každou skupinu zvlášť odhadněte hustotu pomocí histogramu a distribuční funkci pomocí empirické distribuční funkce.

Vykreslíme histogram, distribuční funkci a pomocí kernel density estimation zakreslíme i odhad hustotní funkce:



Stejně i pro distribuční funkci:



Oba grafy sú zakreslené v logaritmickom merítku pre prehľadnosť.

3) Odhadněte parametry normálního, exponenciálního a rovnoměrného rozdělení. Vysvětlete, jak jste odhady získali.

Pre získanie parametrov rozdelení využijeme maximum likelihood estimation (MLE). Budeme maximalizovať pravdepodobnosť že vzniknuté merania vznikli z rozdelenia v závislosti od parametrov rozdelenia.
Pre normálne rozdelenie sú parametre získane pomocou MLE:

$$\hat{\mu} = \frac{1}{n}\sum_{i=0}^{n} x_i, \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - \hat{\mu})^2}$$

U exponencionálního rozdělení spočteme parametr λ pomocí vztahu:

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

Pre rovnomerné rozdelenie sú parametre získane pomocou MLE:

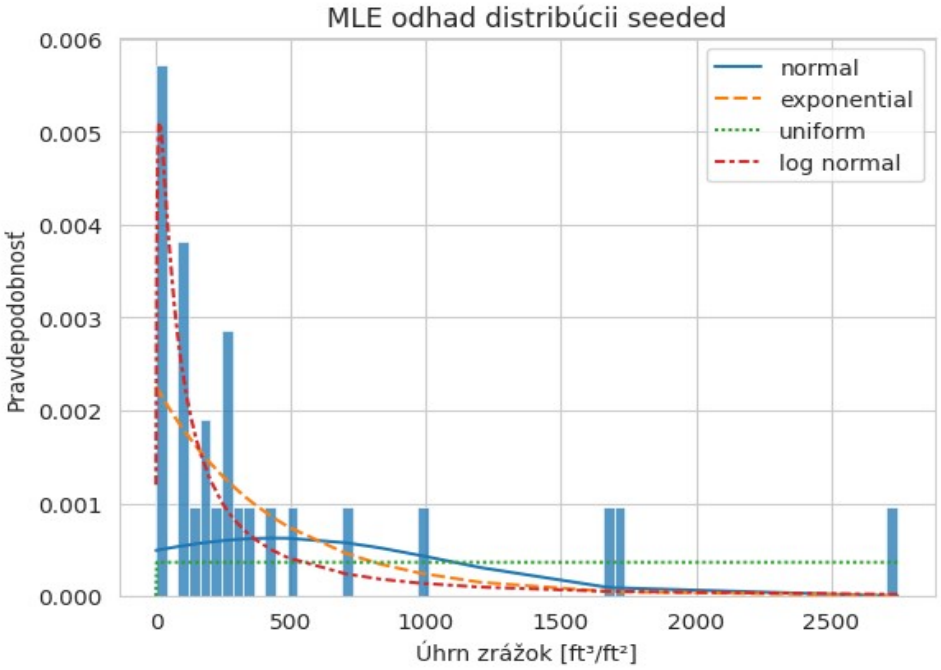
$$a = min\{x_i\}, \quad b = max\{x_i\}$$

Pro zajímavost zkusíme vykreslit také Logaritmicko-normální rozdělení jehož parametry, μ a σ spočteme následovně:

$$\hat{\mu} = \frac{1}{n}\sum_{i=0}^{n} \ln x_i, \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\ln x_i - \hat{\mu})^2}$$

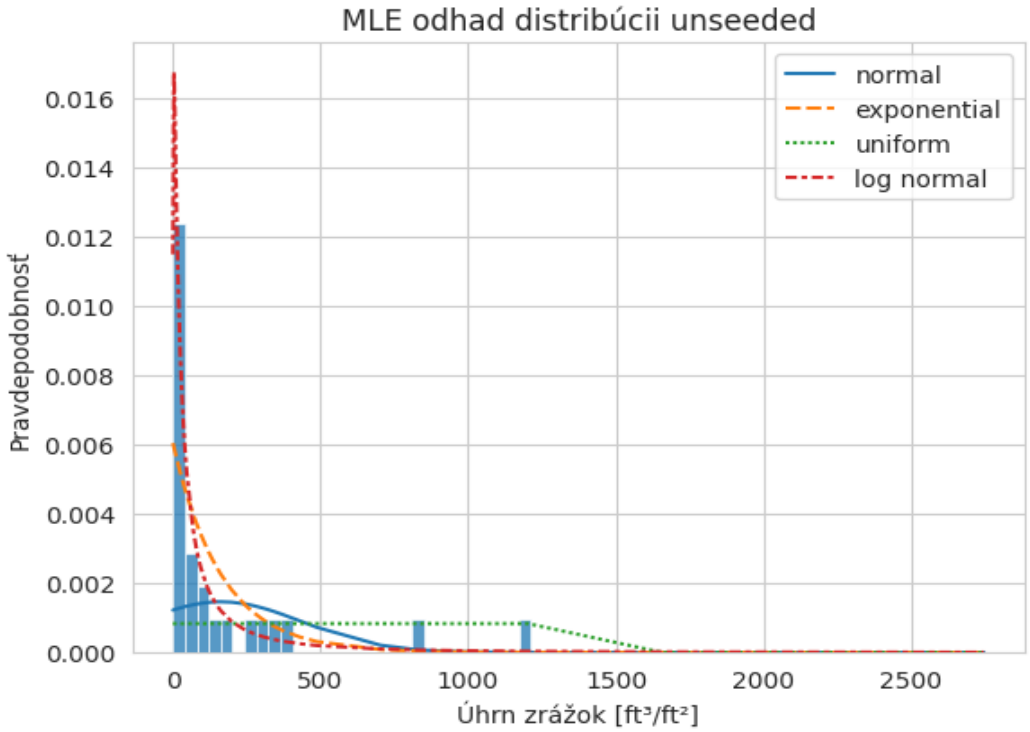
Následně vykreslíme histogram a jednotlivá rozdělení s následujícými parametry pro seeded:

Rozdělení	Odhadnuté parametry	
Normální	μ = 441.98	σ = 638.15
Exponenxionální	1 / λ = 441.98	
Rovnoměrné	a = 4.1	b = 2745.6
Logaritmicko-normální	μ = 5.13	σ = 1.23



Následně vykreslíme histogram a jednotlivá rozdělení s následujícými parametry pro unseeded:

Rozdělení	Odhadnuté parametry	
Normální	μ = 164.59	σ = 273.02
Exponenxionální	1 / λ = 164.59	
Rovnoměrné	a = 1.0	b = 1202.6
Logaritmicko-normální	μ = 4.00	σ = 1.30

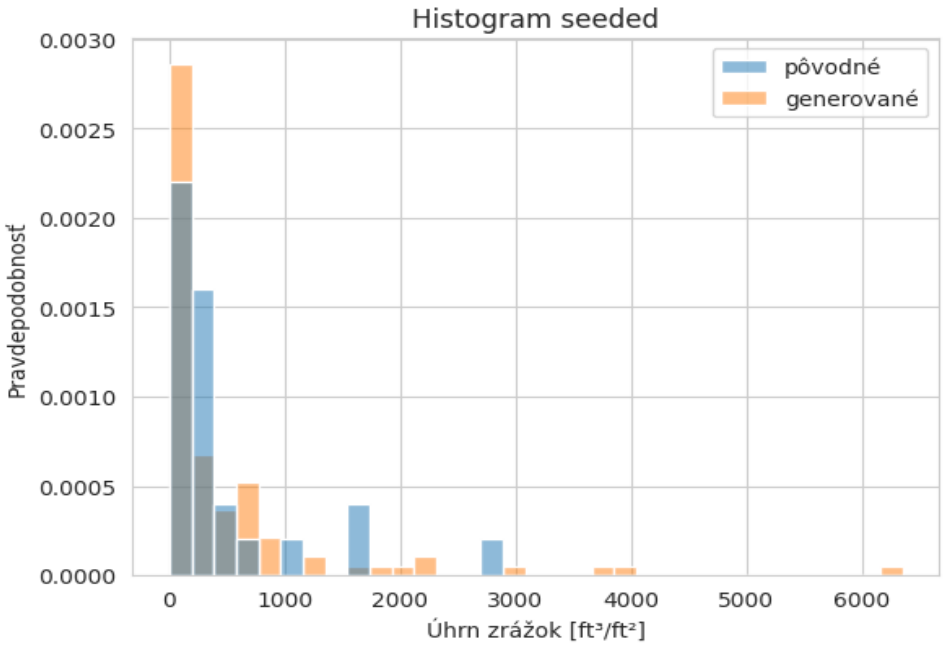


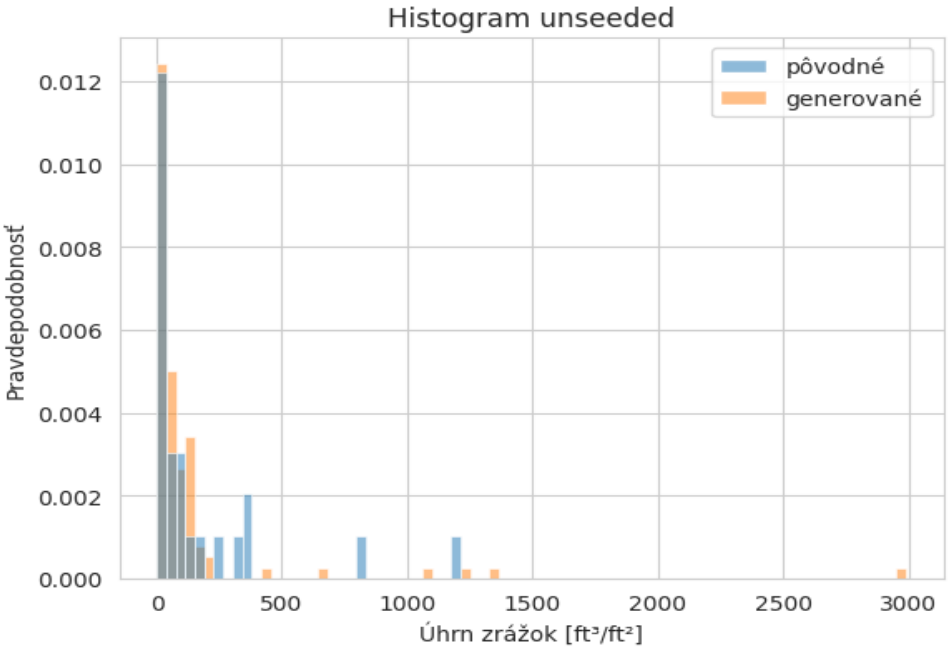
3) Diskutujte, které z rozdělení odpovídá pozorovaným datům nejlépe.

Z rozdělení které máme na výběr, tak se data v obou případech nejvíce podobají Logaritmicko-normální rozdělení (nebo exponencionálnímu rozdělení). Rovnaké výsledky vychádzajú aj keď vyrátamé likelihood pozorovaných dát pre jednotlivé distribúcie.

4) Pro každou skupinu zvlášť vygenerujte náhodný výběr o 100 hodnotách z rozdělení, které jste zvolili jako nejbližší, s parametry odhadnutými v předchozím bodě.

Vykreslíme histogram náhodných dat pocházejících z Logaritmicko-normální rozdělení s MLE parametri a porovnáme je s původními daty:





A tady se můžeme utvrdit že se data nejvíce podobají datům z Logaritmicko-normální rozdělení.

5) Pro každou skupinu zvlášť spočítejte oboustranný 95% konfidenční interval pro střední hodnotu.

Z dat známe jen výběrovou směrodatnou odchylku a aritmetický průměr. A v důsledku centrální limitní věty můžeme pro velké n stejné intervaly spolehlivosti použít přibližně i pro náhodný výběr z libovolného rozdělení. A to následovně, kde použijeme studentovo rozdělení a směrodatnou odchylku s (odmocnina z výběrového rozptylu):

$$\left(\bar{x} - \frac{t_{1-\alpha/2} \cdot s}{\sqrt{n}}, \bar{x} + \frac{t_{1-\alpha/2} \cdot s}{\sqrt{n}} \right)$$

Výsledný oboustranný 95% konfidenční interval pro střední hodnotu meření:

$$CI_{seeded} = [179.13, 704.84]$$
$$CI_{unseeded} = [52.13, 277.05]$$

6) Pro každou skupinu zvlášť otestujte na hladině významnosti 5 % hypotézu, zda je střední hodnota rovna hodnotě K, proti oboustranné alternativě.

Sestavíme nulovou a alternativní hypotézu:

- H_0 : Střední hodnota je rovna K = 6
- H_1 : Střední hodnota není rovna K = 6

Vytvoříme oboustranný 95% konfidenční interval pro střední hodnotu (viz úloha 5)):

- seeded: $6 \notin$ (179.13, 704.84) tudíž na hladině významnosti 5% nulovou hypotézu zamítneme ve prospěch alternativy.
- unseeded: $6 \notin$ (52.13, 277.05) Tudíž na hladině významnosti 5% nulovou hypotézu zamítneme ve prospěch alternativy.

7) Na hladině významnosti 5 % otestujte, jestli mají pozorované skupiny stejnou střední hodnotu. Typ testu a alternativy stanovte tak, aby vaše volba nejlépe korespondovala s povahou zkoumaného problému.

Pre test budeme predpokladať že hodnoty sledujú logoritmicko-normálne rozdelenie. Urobíme štatistický test pre silnejšie tvrdenie a otestujeme či pozorované skupiny sledujú rovnaké logaritmicko-normálne rozdelenie. Pre test použijeme likelihood ratio test:

- Null hypotéza: obe skupiny meraní sa riadia rovnakým logaritmicko-normálnym rozdelením.
- Alternatívna hypotéza: skupiny sa riadia dvomi rôznymi logaritmicko-normálnymi rozdeleniami.

Výsledná p-hodnota je 0.04, teda môžeme null hypotézu zamietnuť v prospech alternatívnej hypotézy. Na hladine významnosti 5% teda môžeme povedať že použitie iodidu strieborného má vplyv na úhrn zrážok oblakov.

Appendix A: Použité dáta

Unseeded	Seeded
1202.60	2745.60
830.10	1697.80
372.40	1656.00
345.50	978.00
321.20	703.40
244.30	489.10
163.00	430.00
147.80	334.10
95.00	302.80
87.00	274.70
81.20	274.70
68.50	255.00
47.30	242.50
41.10	200.70
36.60	198.60
29.00	129.60
28.60	119.00
26.30	118.30
26.10	115.30
24.40	92.40
21.70	40.60
17.30	32.70
11.50	31.40
4.90	17.50
4.90	7.70
1.00	4.10

Appendix B: jupyter notebook s výpočtami

Prílohy **hw.ipynb, requirements.txt, case0301.rda**