Domácí úkol

Dan Kostiuk, Oliver Tušla, Štefan Slavkovský

M = 5

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

V úloze budeme zkoumat účinky jodidu stříbrného při použití v oblacích na celkové srážky na plochu. Zkoumaná data se skládají ze dvou sad měření. První, kontrolní, sada (unseeded) obsahuje měření srážek ze dní, kdy nebyl použity jodid stříbrný. Druhá sada (seeded) obsahuje měření ze dní, kdy jodid stříbrný použitý byl.

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:

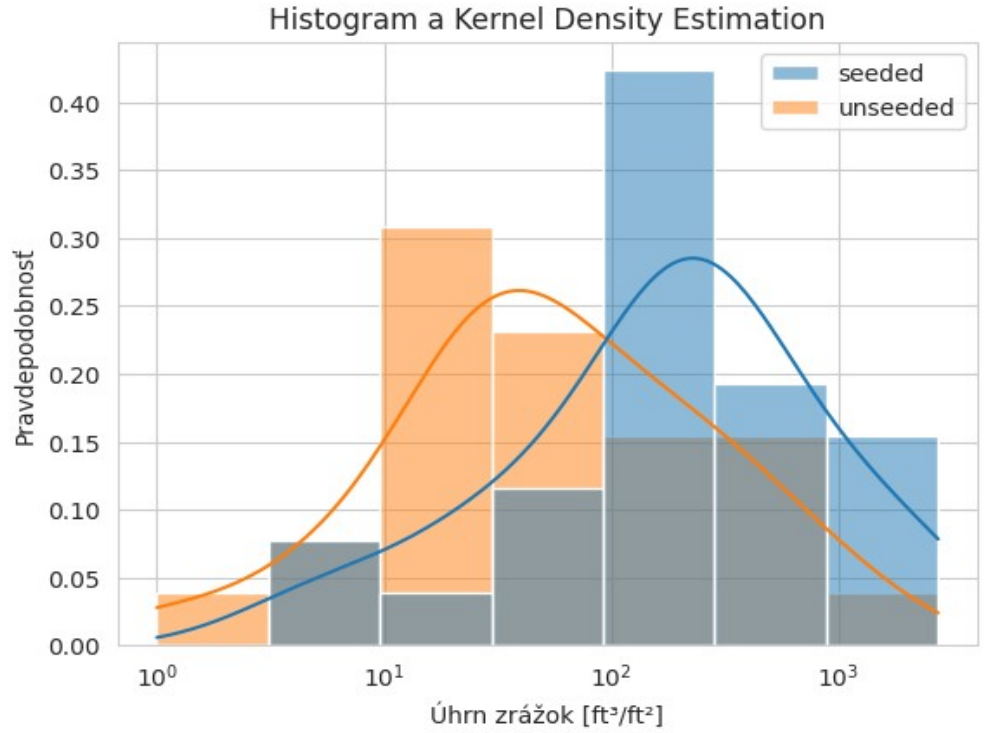
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Medián odhadneme jako prostřední hodnotu z našich seřazených dat.

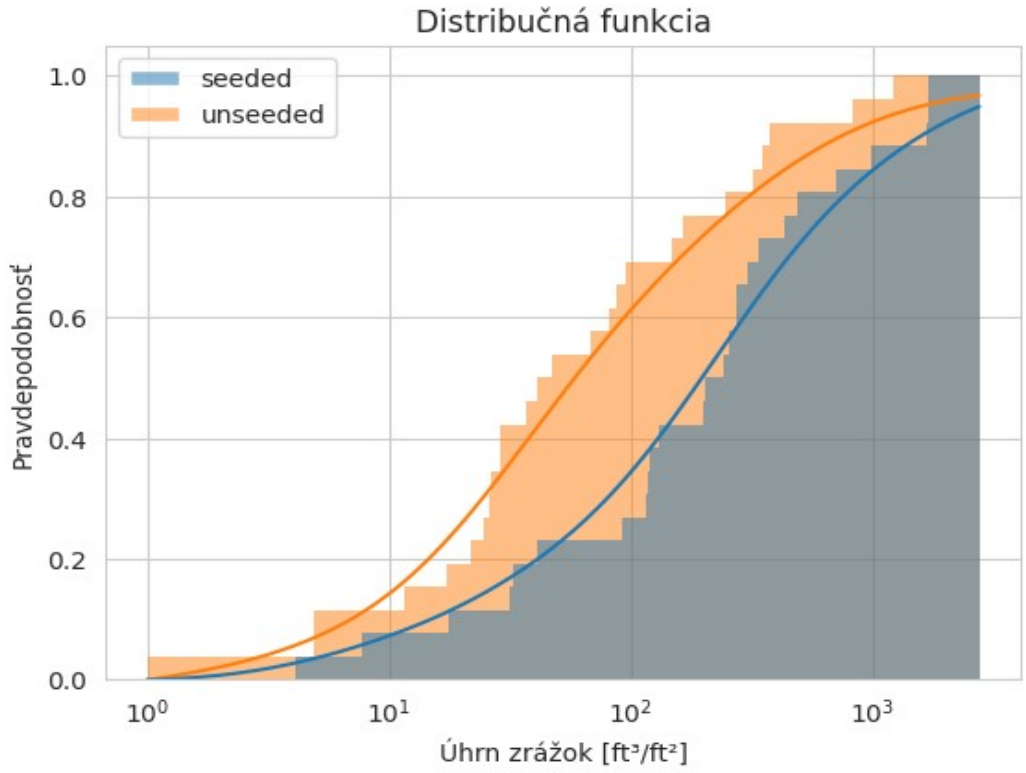
	Střední hodnota	Rozptyl	Median
Seeded	441.98	423523.96	221.60
Unseeded	164.59	77521.25	44.20

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 jsou pro přehlednost vyobrazené na logaritmické stupnici.

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

Pro získání parametrů rozdělení využijeme maximum likelihood estimation (MLE). Budeme maximalizovat pravděpodobnost, že hodnoty měření vznikly z rozdělení v závislosti na parametrech rozdělení.
Pro normální rozdělení získáme parametry pomocí 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:

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

Pro rovnoměrné rozdělení získáme parametry pomocí MLE:

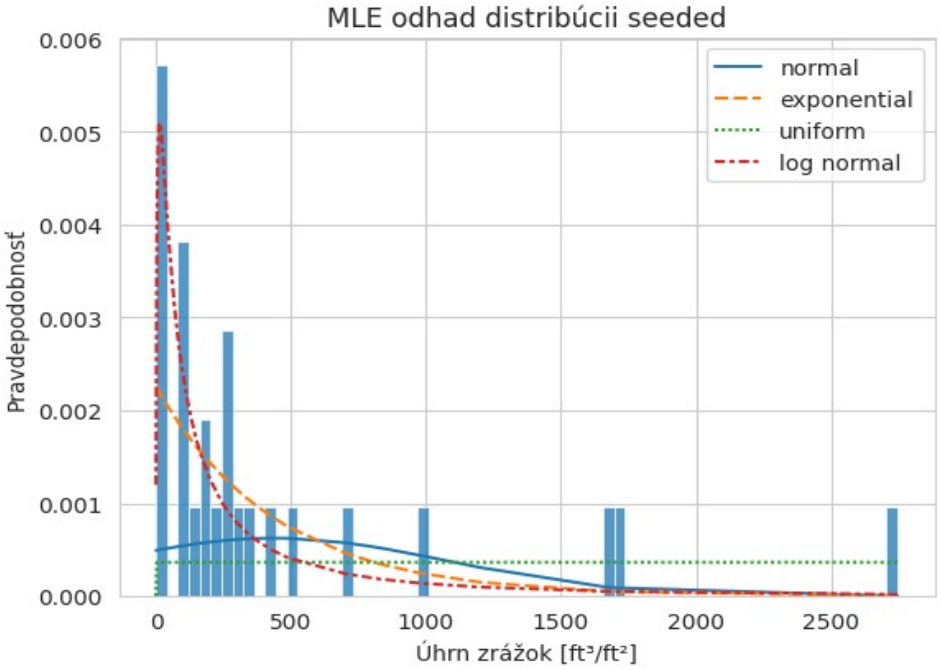
$$a = min\{x_i\}, \quad b = max\{x_i\}$$

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

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

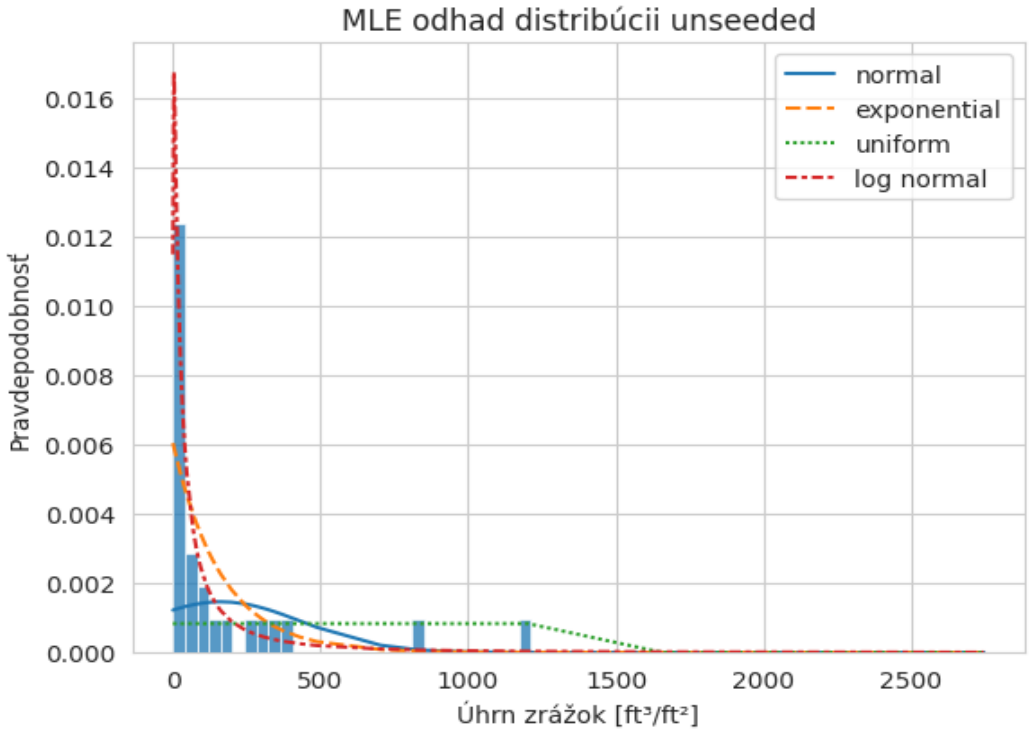
Následně vykreslíme histogram a jednotlivá rozdělení s následujícími parametry pro seeded:

Rozdělení	**Odhadnuté parametry**	
Normální	μ = 441.98	σ = 638.15
Exponenciální	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 následujícími parametry pro unseeded:

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

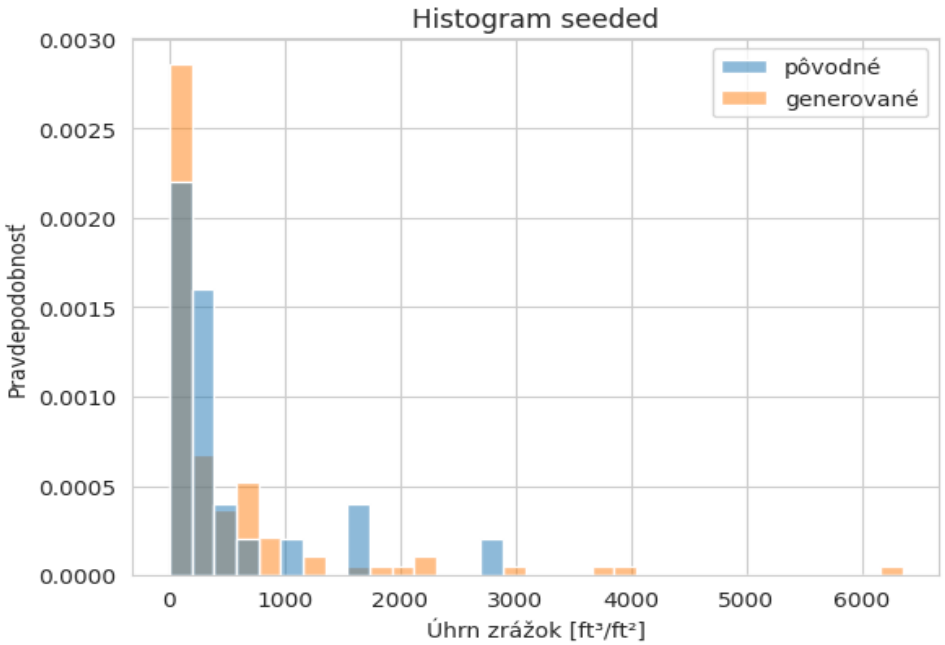


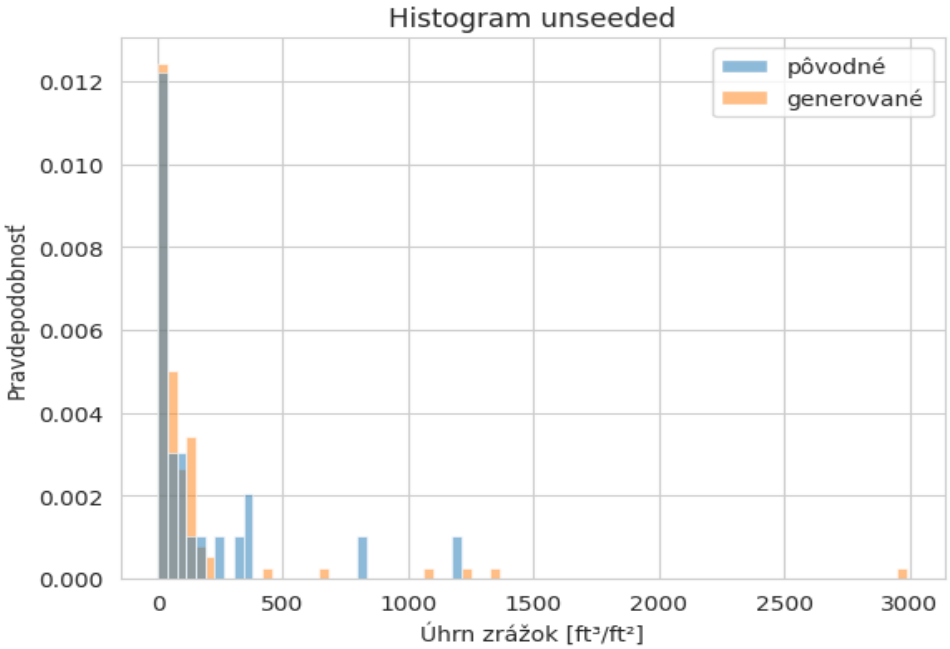
3) Diskutujte, které z rozdělení odpovídá pozorovaným datům nejlépe.

Z rozdělení, která máme na výběr, se data v obou případech nejvíce podobají logaritmicko-normálnímu rozdělení (nebo exponencionálnímu rozdělení). Stejné výsledky vychází, i když spočteme likelihood pozorovaných dat pro jednotlivé distribuce.

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-normálního rozdělení s MLE parametry a porovnáme je s původními daty:





Vidíme, že data se nejvíce podobají datům z logaritmicko-normálního rozdělení.

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

Z dat známe jen výběrovou směrodatnou odchylku a aritmetický průměr. 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 měření:

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

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.

Pro test budeme předpokládat, že hodnoty sledují logoritmicko-normální rozdělení. Uděláme statistický test pro silnější tvrzení a otestujeme, zda pozorované skupiny sledují stejné logaritmicko-normílní rozdělení. Pro test použijeme likelihood ratio test:

- Null hypotéza: obě skupiny měření se řídí stejným logaritmicko-normálním rozdělením.
- Alternatívní hypotéza: skupiny se řídí dvěma různými logaritmicko-normálními rozděleními.

Výsledná p-hodnota je 0.04, tedy můžeme null hypotézu zamítnout v prospěch alternativní hypotézy. Na hladine významnosti 5 % tedy můžeme říct, že použití jodidu stříbrného má vliv na úhrn srážek mraků.

Appendix A: Použitá data

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 výpočty

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