

Unseeded	Seeded
1202.60	2745.6
830.10	1697.8
372.40	1656
345.50	978
321.20	703.4
244.30	489.1
163.00	430
147.80	334.1
95.00	302.8
87.00	274.7
81.20	274.7
68.50	255
47.30	242.5
41.10	200.7
36.60	198.6
29.00	129.6
28.60	119
26.30	118.3
26.10	115.3
24.40	92.4
21.70	40.6
17.30	32.7
11.50	31.4
4.90	17.5
4.90	7.7
1.00	41

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

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:

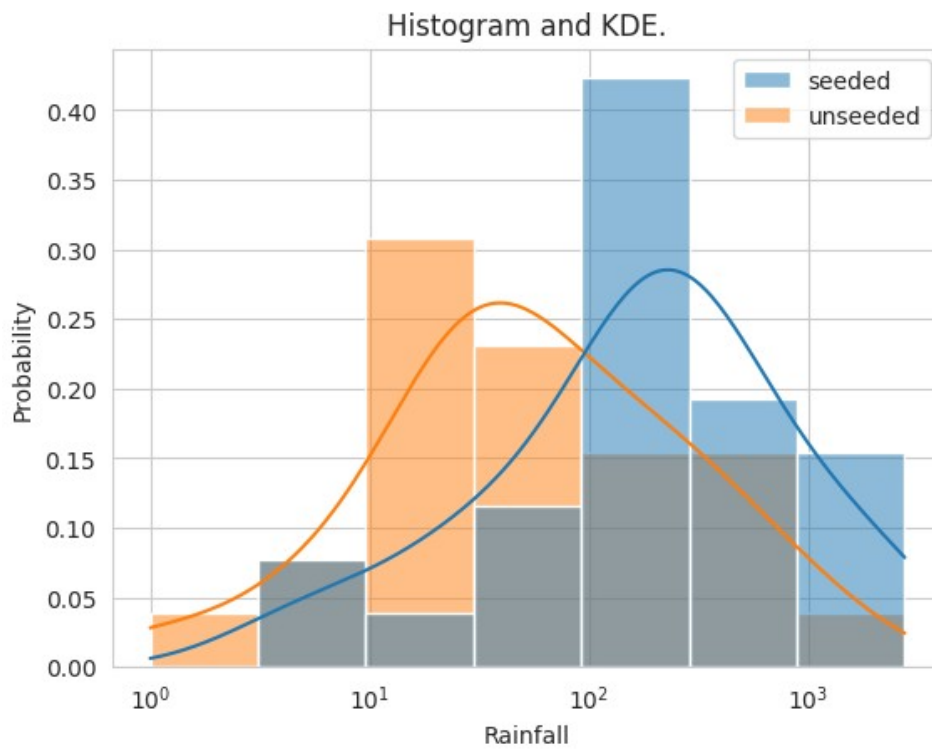
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Medián odhadneme jako prostřední hodnotu v našich seřazených datech.

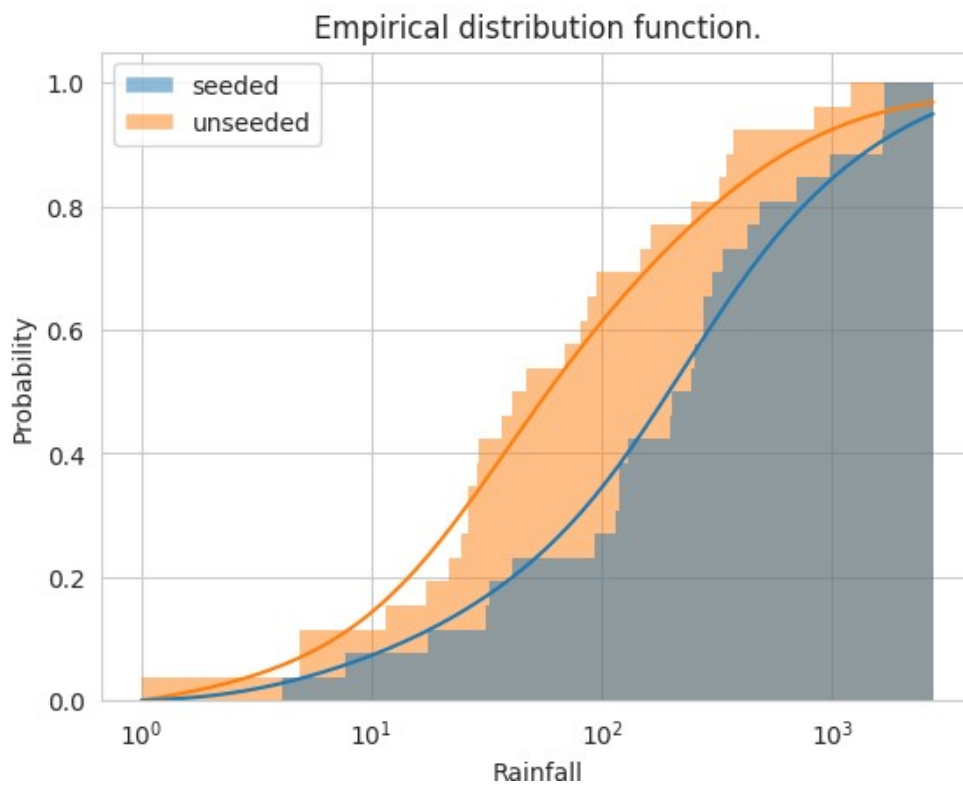
	Střední hodnota	Rozptyl	Median
Unseeded	441.98	278.43	221.6
Seeded	164.59	650.79	44.2

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

Vykreslíme histogram a distribuční funkci:



A pomocí kernel density estimation zakreslíme i odhad hustotní funkce. Stejně i pro distribuční funkci:



3) a) Odhadněte parametry normálního, exponenciálního a rovnoměrného rozdělení.

Pro normální rozdělení je potřeba znát střední hodnotu a rozptyl, což už známe z ulohy 1).

U exponenciálního rozdělení spočteme parametr λ pomocí vztahu:

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

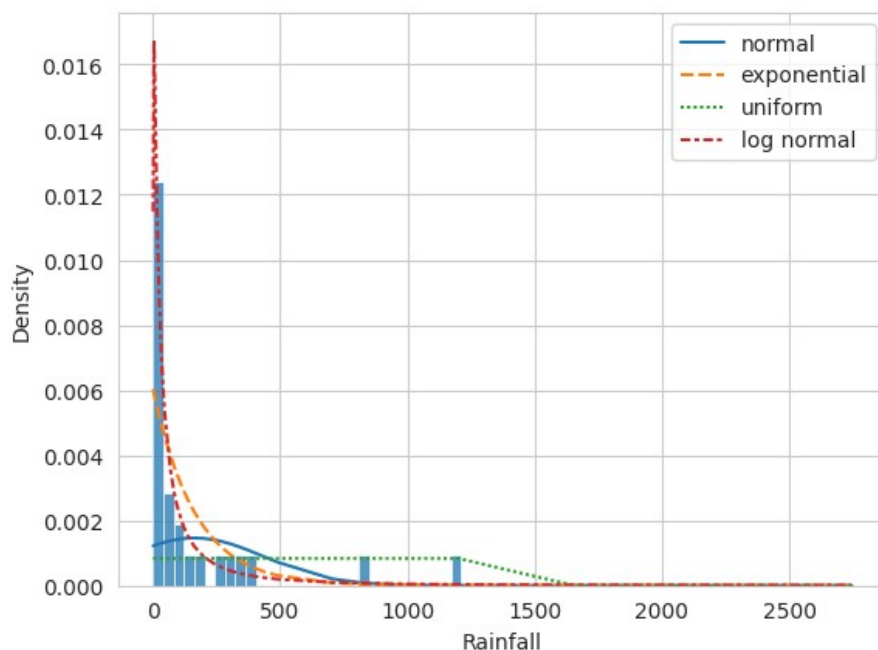
A pro rovnoměrné rozdělení zvolíme parametry a , b . Parametr a zvolíme jako minimální a b jako maximální hodnotu z naměřených dat.

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

$$\mu = \ln \bar{x} - \frac{\ln \left(1 + \frac{s^2}{(\bar{x})^2} \right)}{2}$$
$$\sigma = \sqrt{\ln \left(1 + \frac{s^2}{\bar{x}^2} \right)}$$

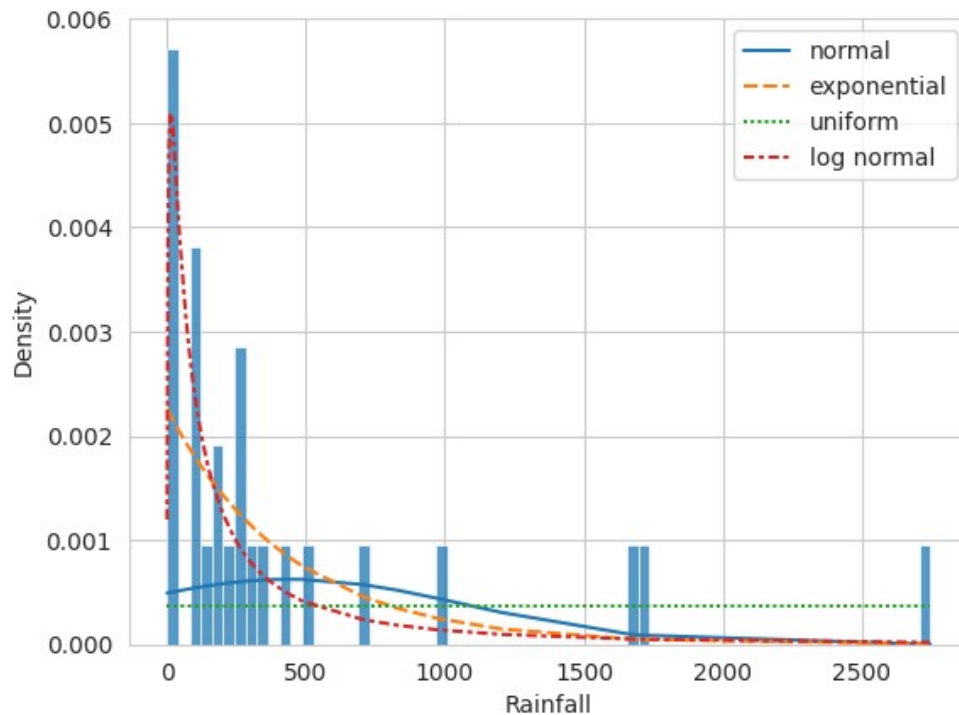
Následně vykreslíme histogram a jednotlivá rozdělení s následujícími parametry pro Unseeded:

Rozdělení	Odhadnuté parametry	
Normální	$\mu = 441.98$	$\sigma^2 = 278.43$
Exponenciální	$\lambda = 0.0023$	
Rovnoměrné	$a = 0$	$b = 1202.6$
Logaritmicko-normální	$\mu = 5.1$	$\sigma = 0.10$



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

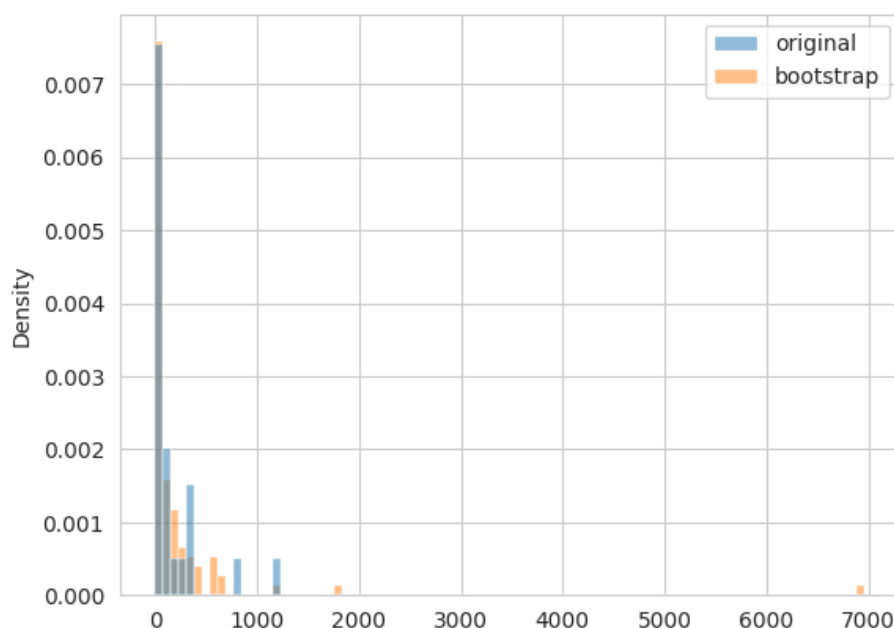
Rozdělení	Odhadnuté parametry	
Normální	$\mu = 164.59$	$\sigma^2 = 650.79$
Exponenční	$\lambda = 0.0061$	
Rovnoměrné	$a = 0$	$b = 2745.60$
Logaritmicko-normální	$\mu = 6.08$	$\sigma = 0.06$



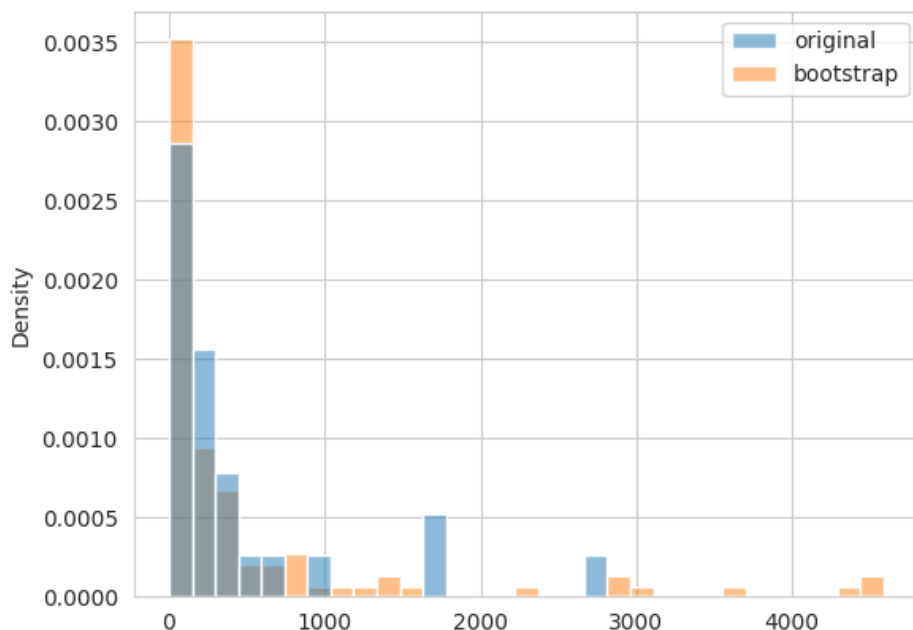
Z rozdělení které máme na výběr, tak se data v obou případech nejvíce podobají Logaritmicko-normální rozdělení (nebo exponenčnímu rozdělení).

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 parametry $\mu = 5.1$, $\sigma = 0.10$ a porovnáme je s Unseeded daty:



Vykreslíme histogram náhodných dat pocházejících z Logaritmicko-normálního rozdělení s parametry $\mu = 6.08$, $\sigma = 0.06$ a porovnáme je s Seeded daty:



A tady se můžeme utvrdit že se data 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. 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 z Unseeded dat:
(27.86, 104.96)

Výsledný oboustranný 95% konfidenční interval pro střední hodnotu z Seeded dat:
(88.95, 323.84)

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

Sestavíme nulovou a alternativní hypotézu:

H_0 : Střední hodnota je rovna 6

H_1 : Střední hodnota není rovna 6

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

Unseeded:

$6 \notin (27.86, 104.96)$ Tudíž na hladině významnosti 5% nulovou hypotézu zamítneme ve prospěch alternativy.

Seeded:

$6 \notin (88.95, 323.84)$ 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.

(μ_1 je průměr Unseeded)

(μ_2 je průměr Seeded)

$H_1: \mu_1 = \mu_2$

$H_2: \mu_1 \neq \mu_2$