Wydział Matematyki
i Nauk Informacyjnych

POLITECHNIKA WARSZAWSKA

RPiESM, Omówienie laboratorium 1

Julian Zalewski, Antoni Zasada

21 maja 2025

Zadanie 1. Wygenerować $N = 10000$ obserwacji $X_1, X_2, \ldots, X_n$ z rozkładu

(a) dwupunktowego Binom$(1, \frac{1}{4})$

(b) wykładniczego Exp$(\frac{1}{3})$

(c) Cauchy'ego $\mathcal{C}(0, 1)$.

Dla każdego z powyższych przypadków wyznaczyć wykres ciągu $\bar{X_1}, \bar{X_2}, \ldots, \bar{X_N}$, gdzie $\bar{X_n}$ oznacza średnią z pierwszych n obserwacji, $n = 1, 2, \ldots, N$, czyli

$$\bar{X_n} = \frac{S_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Wyciągnąć wnioski dotyczące zachowania się uzyskanych ciągów średnich.

Rozwiązanie: W języku R do generowania obserwacji z zadanego rozkładu służą funkcje o nazwach postaci `r<nazwa rozkładu>`. Ich pierwszym argumentem jest liczba obserwacji, a kolejne to parametry zależne od rozkładu. W tym zadaniu korzystamy zatem odpowiednio z funkcji `rbinom(N, 1, 1/4)`, `rexp(N, 1/3)` i `rcauchy(N, 0, 1)`.

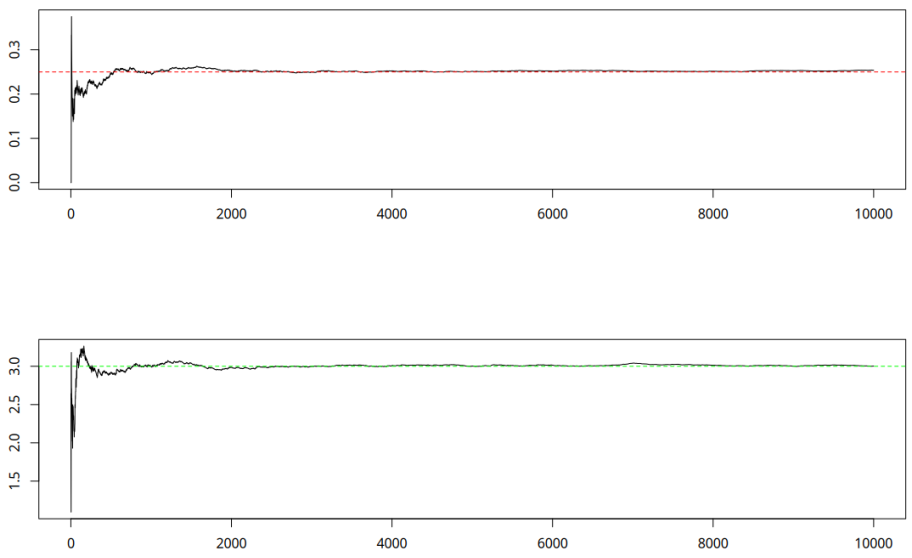
Wektor $(\bar{X_n})$ średnich z pierwszych n obserwacji otrzymamy, dzieląc (element po elemencie) wektor sum skumulowanych uzyskany funkcją `cumsum` przez wektor $[1, 2, \ldots, N]$, który w R zapiszemy jako `1:N`.

Następnie chcemy wyświetlić wykresy wyznaczonych ciągów. Wykorzystujemy do tego `plot`. Aby uzyskać wykres będący linią musimy ustawić parametr `type` na `l` (jak *line*).

Rozkłady z podpunktów (a) i (b) posiadają skończone wartości oczekiwane. Są one równe

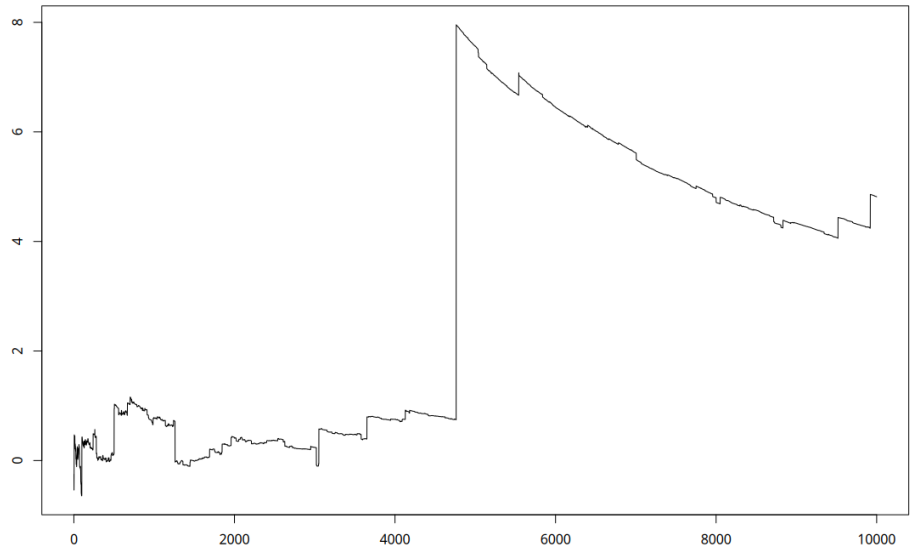
$$\mu_a = \frac{1}{4} \cdot 1 + \frac{3}{4} \cdot 0 = \frac{1}{4}$$
$$\mu_b = \frac{1}{\frac{1}{3}} = 3.$$

Spodziewamy się więc, że ich wykresy zobrazują działanie mocnego prawa wielkich liczb Kołmogorowa, tj. zauważymy zbieżność ciągu średnich do obliczonych wartości oczekiwanych. Do wykresów dodamy pomocnicze proste $y = \mu_a$ i $y = \mu_b$ korzystając z funkcji `abline`.



Rysunek 1: Wykresy (a) i (b), linie przerywane pokazują wartości oczekiwane rozkładów.

Dla rozkładu Cauchy'ego w podpunkcie (c) nie możemy spodziewać się zbieżności, ponieważ wartość oczekiwana jest w jego przypadku niezdefiniowana.



Rysunek 2: Wykres (c), rozkład nie spełnia MPWL ze względu na brak wartości oczekiwanej.

Pełny kod rozwiązania (*Zadaniem komendy* `rm(list = ls())` *w pierwszej linii jest usunięcie zalegających zmiennych z wcześniej uruchomionych skryptów w celu uniknięcia konfliktów.*):

```
rm(list = ls())

N = 1e4
n = 1:N
par(mfrow=c(3,1))

x_a = rbinom(N, size=1, prob=1/4)
S_a = cumsum(x_a)
M_a = S_a/n
plot(M_a, type="l", xlab="", ylab="")
abline(h = 1/4, lty = 2, col="red")

x_b = rexp(N, rate=1/3)
S_b = cumsum(x_b)
M_b = S_b/n
plot(M_b, type="l", xlab="", ylab="")
abline(h = 3, lty = 2, col="green")

x_c = rcauchy(N, location=0, scale=1)
S_c = cumsum(x_c)
M_c = S_c/n
plot(M_c, type="l", xlab="", ylab="")
```

Zadanie 2. Wygenerować próbę $Y = (Y_1, \ldots, Y_n),\ n = 500$ z rozkładu normalnego $N(\mu = 4, \sigma = 2)$. Utworzyć podpróby $X_i = (Y_1, \ldots, Y_i),\ i = 1, \ldots, n$ i wyznaczyć ciągi:

- średnich: $\{\bar{X}_i : i = 1, \ldots, n\}$,
- median: $\{\mathrm{Med}_i : i = 1, \ldots, n\}$,
- odchyleń standardowych: $\{S_i : i = 2, \ldots, n\}$,
- rozstępów międzykwartylowych podzielonych przez 1.35: $\{D_i = \mathrm{IQR}_i/1.35 : i = 2, \ldots, n\}$.

(a) Narysować na wspólnym wykresie ciągi średnich i median. Przeanalizować wpływ liczności próby na zachowanie się średniej i mediany z próby. Czy statystyki te wydają się być sensownymi estymatorami parametru położenia μ w tym modelu?

(b) Narysować na wspólnym wykresie ciągi odchyleń standardowych i rozstępów międzykwartylowych podzielonych przez 1.35. Przeanalizować wpływ liczności próby na zachowanie się tych statystyk. Czy wydają się one być sensownymi estymatorami parametru rozproszenia σ w tym modelu?

Rozwiązanie: Podobnie jak w zadaniu 1. generujemy próby z rozkładu normalnego $N(\mu = 4, \sigma = 2)$. Do obliczania ciągów w treści używamy funkcji R:

- `mean(wektor)` – oblicza średnią arytmetyczną:

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

- `mean(wektor, na.rm=TRUE)` – oblicza średnią, pomijając wartości `NA` (*not available* = brak danych),
- `median(wektor)` – oblicza medianę (wartość środkowa),
- `quantile(wektor, 0.25)` – dolny kwartyl (Q_1),
- `quantile(wektor, 0.75)` – górny kwartyl (Q_3),
- `quantile(wektor, c(0.1, 0.99, 0.85))` – decyle, percentyle i kwantyle (np. 10%, 99%, 85%),
- `max(wektor) - min(wektor)` – rozstęp (zakres):
- `IQR(wektor)` – rozstęp międzykwartylowy:

$$\mathrm{IQR} = Q_3 - Q_1$$

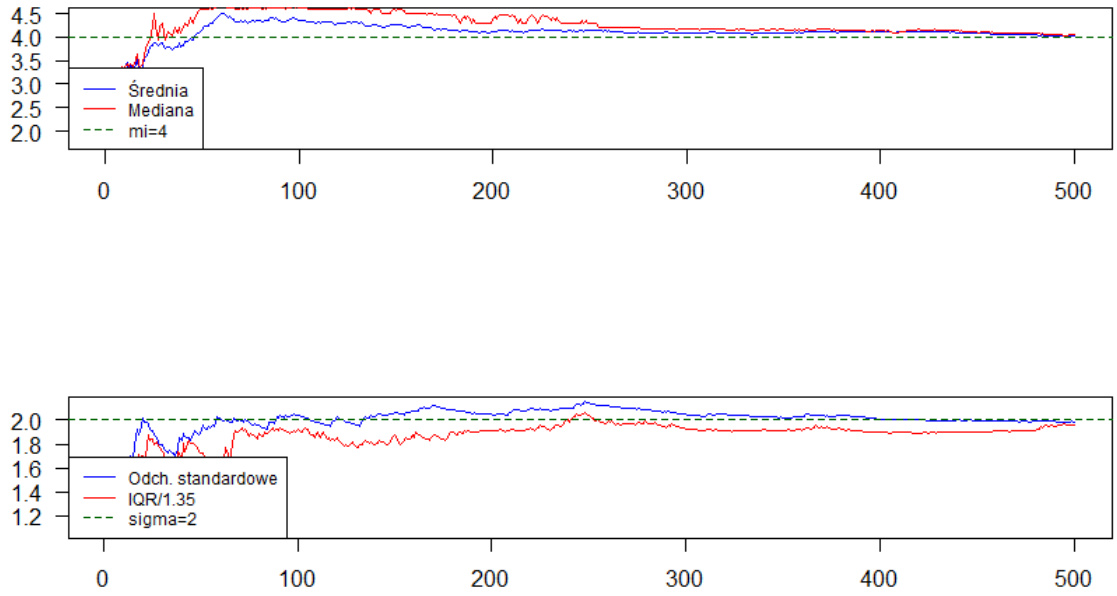
- `var(wektor)` – wariancja:

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

- `sd(wektor)` – odchylenie standardowe:

$$S = \sqrt{S^2}$$

W pętli `for` obliczamy podpróby używając `x = y[1:i]` oraz liczymy miary podane w zadaniu jednocześnie umieszczając je do wcześniej stworzonych wektorów. Do utworzenia wyżej wymienionych wektorów używamy `numeric`, która służy do tworzenia obiektów typu `numeric` o zadanej długości.



Rysunek 3: Wykresy dla podpunktów (a) i (b)

Z wykresów wynika, że wszystkie sprawdzane estymatory są sensowne.

Pełny kod rozwiązania:

```
rm(list = ls())
par(mfrow=c(2,1))
n = 500
mi = 4
sigma = 2
y = rnorm(n, mean=mi, sd=sigma)
mean_y = numeric(n)
median_y = numeric(n)
sd_y = numeric(n)
iqr_y = numeric(n)

for (i in 1:n)
{
  x = y[1:i]
  mean_y[i] = mean(x)
  median_y[i] = median(x)
  sd_y[i] = sd(x)
  iqr_y[i] = IQR(x)/1.35
}
plot(1:n, mean_y, type="l", col="blue", ylab="", las=1)
lines(1:n, median_y, col="red")
abline(h = mi, col="darkgreen", lty=2)
legend("bottomleft", legend=c("Średnia", "Mediana", "mi=4"),
       col=c("blue", "red", "darkgreen"), lty=c(1,1,2))

plot(2:n, sd_y[2:n], type="l", col="blue", ylab=" ", las=1)
lines(2:n, iqr_y[2:n], col="red")
abline(h=sigma, col="darkgreen", lty=2)
legend("bottomleft", legend=c("Odch. standardowe", "IQR/1.35", "sigma=2"),
       col=c("blue", "red", "darkgreen"), lty=c(1,1,2))
```

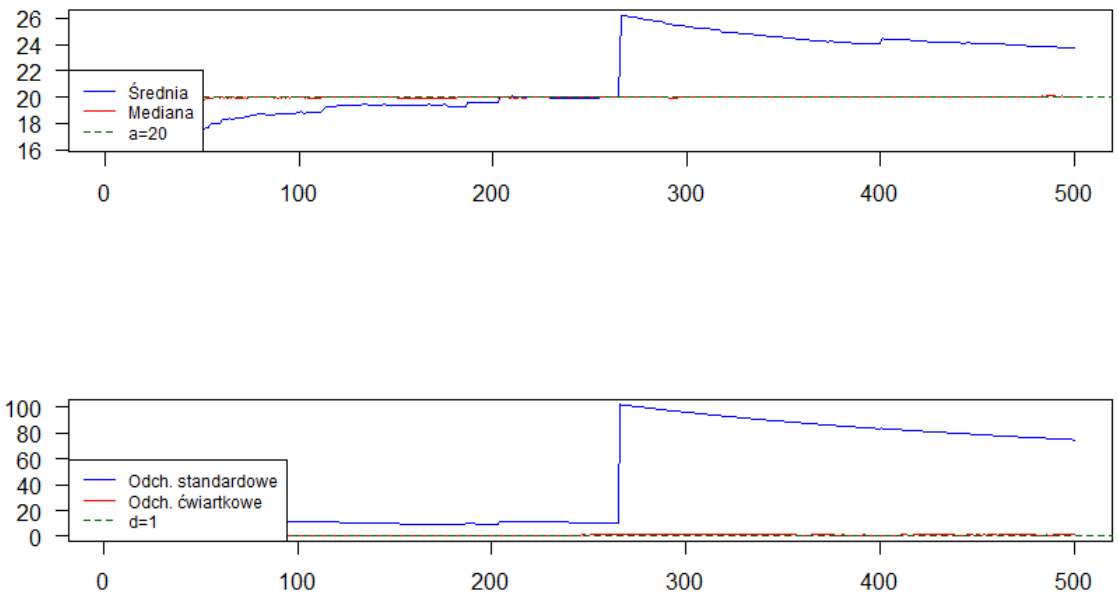
Zadanie 3. Wygenerować próbę $Y = (Y_1, \ldots, Y_n),\ n = 500$ z rozkładu Cauchy'ego $C(a = 20, d = 1)$. Utworzyć podpróby $X_i = (Y_1, \ldots, Y_i),\ i = 1, \ldots, n$ i wyznaczyć ciągi:

- średnich: $\{\bar{X}_i : i = 1, \ldots, n\}$,
- median: $\{\text{Med}_i : i = 1, \ldots, n\}$,
- odchyleń standardowych: $\{S_i : i = 2, \ldots, n\}$,
- odchyleń ćwiartkowych: $\{\text{SQR}_i = \text{IQR}_i/2 : i = 2, \ldots, n\}$.

(a) Narysować na wspólnym wykresie ciągi średnich i median. Przeanalizować wpływ liczności próby na zachowanie się średniej i mediany z próby. Czy statystyki te wydają się być sensownymi estymatorami parametru położenia a w tym modelu?

(b) Narysować na wspólnym wykresie ciągi odchyleń standardowych i ćwiartkowych. Przeanalizować wpływ liczności próby na zachowanie się tych statystyk. Czy wydają się one być sensownymi estymatorami parametru rozproszenia d w tym modelu?

Rozwiązanie: Zadanie technicznie całkowicie analogiczne do zadania 2.



Rysunek 4: Wykresy dla podpunktów (a) i (b)

Z pierwszego wykresu wynika, że mediana jest dobrym estymatorem parametru położenia a, podczas gdy średnia nie jest. Patrząc na drugi wykres wnioskujemy, że odchylenie ćwiartkowe jest sensownym estymatorem parametru rozproszenia d, w przeciwieństwie do odchylenia standardowego.

Pełny kod rozwiązania:

```
rm(list = ls())

par(mfrow=c(2,1))

n = 500
a = 20
d = 1
y = rcauchy(n, location=a, scale=d)
mean_y = numeric(n)
median_y = numeric(n)
sd_y = numeric(n)
sqr_y = numeric(n)
for (i in 1:n)
{
  x = y[1:i]
  mean_y[i] = mean(x)
  median_y[i] = median(x)
  sd_y[i] = sd(x)
  sqr_y[i] = IQR(x)/2
}

plot(1:n, mean_y, type="l", col="blue", xlab="", ylab="", las=1)
lines(1:n, median_y, col="red")
abline(h = a, col="darkgreen", lty=2)
legend("bottomleft", legend=c("Średnia", "Mediana", "a=20"),
       col=c("blue", "red", "darkgreen"), lty=c(1,1,2), cex=0.75)

plot(2:n, sd_y[2:n], type="l", col="blue", xlab="", ylab="", las=1)
lines(2:n, sqr_y[2:n], col="red")
abline(h = d, col="darkgreen", lty=2)
legend("bottomleft", legend=c("Odch. standardowe", "Odch. ćwiartkowe", "d=1"),
       col=c("blue", "red", "darkgreen"), lty=c(1,1,2), cex=0.75)
```

Zadanie 4. Niech $X_1, X_2, \ldots, X_n$ będzie prostą próbą losową z rozkładu wykładniczego $\mathrm{Exp}(\lambda)$, gdzie:

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad \text{gdzie } \lambda > 0.$$

(a) W celu oszacowania czasu działania pewnych bateryjek, dział kontroli jakości zmierzył czas pracy 8 losowo wybranych bateryjek i otrzymał następujące wyniki (w godz.):

$$483,\ 705,\ 2623,\ 347,\ 620,\ 2719,\ 1035,\ 421$$

Wiadomo, że czas pracy tych bateryjek ma rozkład wykładniczy $\mathrm{Exp}(\lambda)$ z nieznanym $\lambda > 0$. Dla danych zebranych przez dział kontroli jakości, podać wartość estymatora największej wiarygodności parametru λ.

(b) Dla danych z pkt. (a) wyznaczyć estymator największej wiarygodności dla:

- średniego czasu działania bateryjki,
- prawdopodobieństwa, że bateryjka będzie działać krócej niż 1000 godz.

Rozwiązanie: Do zapisania danych używamy funkcji `c` (combine) aby stworzyć wektor poprzez połączenie wartości pomiarów. Do obliczenia estymatora największej wiarygodności używamy `fitdistr` z wcześniej załadowanej biblioteki *MASS*. Posłużyła nam do tego funkcja `library`. `fitdistr` jako argument `densfun` może przyjąć:

- *beta*,
- *cauchy*,
- *chi-squared*,
- *exponential*,
- *geometric*,
- *log-normal*,
- *lognormal*,
- *logistic*,
- *negative binomial*,
- *normal*,
- *Poisson*,
- *weibull*

Estymator średniego czasu działania bateryjki to $\frac{1}{\hat{\lambda}}$, z własności rozkładu wykładniczego. W języku R do obliczania wartości dystrybuanty zadanego rozkładu służą funkcje o nazwach postaci `p<nazwa rozkładu>` („p" pochodzi od *probability density function* (*PDF*)). W naszym przypadku używamy `pexp`.

Pełny kod rozwiązania:

```
rm(list = ls())
library(MASS)

results = c(483, 705, 2623, 347, 620, 2719, 1035, 421)
est_lambda = fitdistr(results, "exponential")$est
est_mean = 1/est_lambda
est_pdf = pexp(1000, est_lambda)
```

Zadanie 5. Niech Gamma(a, β) oznacza rozkład gamma z parametrem kształtu a i parametrem β, tzn.

$$f(x) = \begin{cases} \dfrac{\beta^a}{\Gamma(a)} x^{a-1} e^{-\beta x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad \text{gdzie } a > 0, \beta > 0.$$

(a) Wygenerować $n = 100$ obserwacji z rozkładu Gamma$(3, 2)$.

(b) Przyjąć, że zapomniano wartości parametrów rozkładu gamma, z którego wygenerowano dane i, używając R, oszacować te parametry stosując metodę największej wiarygodności.

Rozwiązanie: Zadanie pominięte na zajęciach.

Zadanie 6.

(a) Wybrać $\theta > 0$.

(b) Wygenerować $N = 10000$ k-elementowych próbek ($k = 20$) z rozkładu jednostajnego $\mathcal{U}([0, \theta])$.

(c) Porównać empirycznie obciążenie estymatora metody momentów i ENW parametru θ.

Rozwiązanie: W kodzie ustalamy $\theta = 2$. Tworzymy wektor `estimators` używając funkcji `replicate`, która jako pierwszy argument przyjmuje liczbę powtórzeń funkcji podanej jako drugi argument której wynik zostanie połączony do wektora `estimators`. Jak widać niżej, dla rozkładu jednostajnego estymatorem metody momentów jest podwojona średnia, a estymatorem największego wiarygodności jest wartość maksymalna z próbek. Wyniki rysujemy funkcją `plot`.

Wyznaczmy estymator metodą momentów:

$$M_1 = E(X) = \int_0^\theta \frac{x}{\theta}\,\mathrm{d}x = \frac{1}{2} \cdot \frac{\theta^2}{\theta} = \frac{\theta}{2} = \frac{X_1 + \cdots + X_k}{k} = \bar{X}$$

Zatem $\hat{\theta}_{\mathrm{MM}} = 2\bar{X}$.

Teraz skorzystamy z metody największej wiarygodności. Określmy najpierw funkcję gęstości rozkładu.

$$f(x) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & \text{wpp.} \end{cases}$$

Możemy teraz wyznaczyć funkcję wiarygodności.

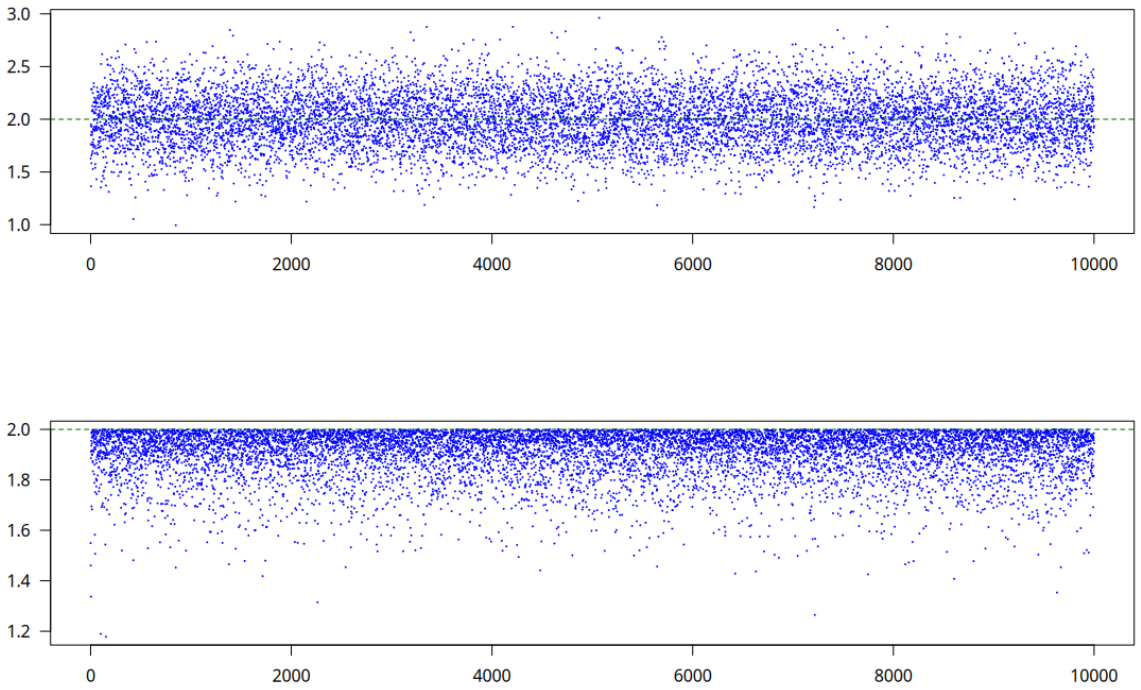
$$L(x_1, \ldots, x_k; \theta) = \prod_{i=1}^{k} f(x_i; \theta) = \prod_{i=1}^{k} \frac{1}{\theta} = \theta^{-k}, \text{ dla } x_i \in [0, \theta]\ (i = 1, 2, \ldots, k)$$

Ostatecznie chcemy znaleźć ekstremum funkcji L, więc możemy równie dobrze operować na jej logarytmie (ponieważ logarytm naturalny jest funkcją ściśle rosnącą).

$$\ln L(x_1, \ldots, x_k; \theta) = \ln\left(\theta^{-k}\right) = -k \ln(\theta)$$

Zatem pochodna funkcji wiarygodności po θ to $-k/\theta$. Jest to funkcja malejąca względem θ, więc maksimum spodziewamy się w lewym krańcu przedziału - czyli dla najmniejszej możliwej wartości $\theta = \max\{x_1, x_2, \ldots, x_k\}$. Wobec tego, $\hat{\theta}_{\mathrm{NW}} = \max\{X_1, X_2, \ldots, X_k\}$.

Na poniższych wykresach niebieskimi punktami zaznaczone są wyznaczone wartości $2\bar{X}$ oraz $\max\{X_1, \ldots, X_k\}$ dla każdej z 20-elementowych próbek.



Przypomnijmy definicję obciążenia estymatora $\hat{\theta}$: $B(\hat{\theta}) \coloneqq E(\hat{\theta}) - \theta$, gdzie $\theta \in \Theta$. Na wykresach wyraźnie widać, że estymator wyznaczony metodą momentów (górny wykres) ma wartości stosunkowo symetrycznie rozłożone wokół $\theta = 2$. Możemy się zatem spodziewać, że jest to estymator nieobciążony.

Natomiast estymator największej wiarygodności zawsze przyjmuje wartości nie większe od $\theta = 2$, więc jego wartość oczekiwana musi być mniejsza od $\theta = 2$. Jest to zatem estymator obciążony.

Pełny kod rozwiązania:

```
rm(list = ls())

par(mfrow=c(2,1))
N = 1e4
k = 20
theta = 2
estimators = replicate(N,
{
  x = runif(k, 0, theta)
  c(2 * mean(x), max(x))
})

plot(1:N, estimators[1,], cex=0.05, col="blue", ylab="", las=1)
abline(h = theta, col="darkgreen", lty=2)
plot(1:N, estimators[2,], cex=0.05, col="blue", ylab="", las=1)
abline(h = theta, col="darkgreen", lty=2)
```

Uwagi

- Do przeglądania dokumentacji używamy `?<nazwa komendy>`, a do wyszukiwania `??`. Przykładowo, aby wyszukać w dokumentacji wszystkie funkcje dotyczące rozkładu jednostajnego wpiszemy `??unif`.
- R jest „case sensitive”, tzn. funkcje lub zmienne różniące się tylko wielkością liter w nazwie zostaną uznane za różne.
- Nie należy nazywać zmiennych `c T F t dt df pt pf rt rf qt qf`, ponieważ są one zarezerwowane dla wbudowanych w R funkcji.
- W R są dwa operatory przypisania: `<-` i `=`. Jest pomiędzy nimi niewielka różnica - znak `=` powoduje przypisanie „lokalne” (dostępne tylko w ramach funkcji, w której zostało zapisane), a `<-` służy do przypisania globalnego, tj. takiego, które będzie dostępne w całej przestrzeni roboczej.
- Gdy chcemy wybrać tylko jedną kolumnę z ramki, to możemy również skorzystać z operatora `$`. Po operatorze `$` możemy podać całą nazwę zmiennej lub jej części. Jeżeli podamy część, to wynikiem będzie kolumna o nazwie rozpoczynającej się od wskazanego napisu.
- Opcja `las` w funkcji `par` kontroluje kierunek osi etykiet.
- Do wyświetlania zmiennych można:
  - otoczyć wyrażenie nawiasami np. `(x=2)`
  - użyć funkcji np. `print(x)`
  - napisać samą nazwę zmiennej w linijce np. `x`
- Przydatna „ściąga” wyjaśniająca możliwości programu RStudio