ADPS 2022Z — Laboratorium 2 (rozwiązania)

Adam Pruszyński

Zadanie 1 (1 pkt)

Treść zadania

Rozkład Poissona jest często używany do modelowania ruchu ulicznego (o małym natężeniu). Plik skrety.txt zawiera liczby pojazdów skręcających na pewnym skrzyżowaniu w prawo w przeciągu trzystu 3-minutowych przedziałów czasu (dane zostały zebrane o różnych porach dnia).

- Wczytaj dane za pomocą komendy scan('skrety.txt').
- Dopasuj do danych rozkład Poissona, tj. wyestymuj parametr λ rozkładu Poissona.
- Metodą bootstrapu nieparametrycznego oszacuj odchylenie standardowe estymatora parametru λ.
- Sprawdź i opisz zgodność rozkładu o wyestymowanym parametrze λ z zarejestrowanymi danymi porównując graficznie empiryczną i teoretyczną funkcję prawdopodobieństwa. Użyj funkcji table() i dpois() analogicznie jak w przykładzie 4 laboratorium 1.

Rozwiązanie

```
n = 300
turns = scan('http://elektron.elka.pw.edu.pl/~mrupniew/adps/skrety.txt')

lambda = mean(turns)
quiet = TRUE
```

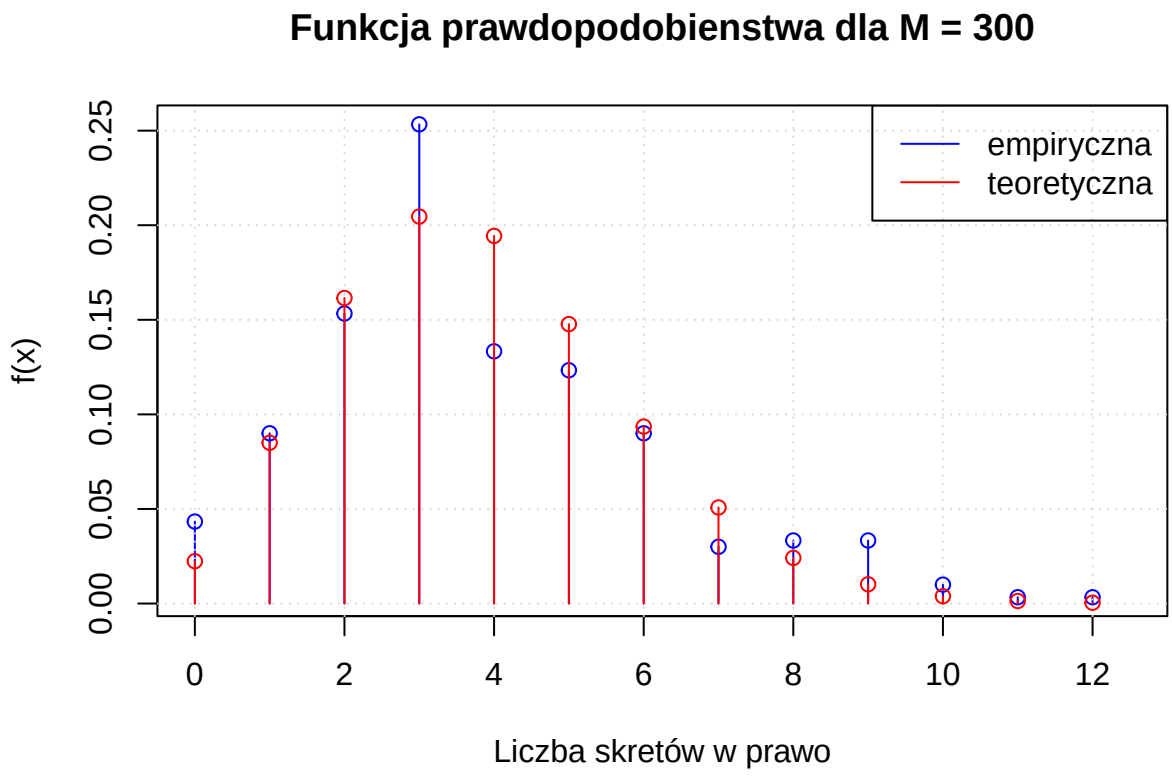
Estymata parametru lambda wynosi 3.8

```
K = 1000
boot_res = replicate(K, {
  boot_dane = sample(turns, n, replace = T)
  c(mean(boot_dane))
})
sd_lambda = sd(boot_res)
```

Odchylenie standardowe estymatora parametru lambda wynosi 0.1313761

```
Arg = 0:max(turns)
Freq = as.numeric(table(factor(turns))) / n
```

```
plot(Freq ~ Arg, type = 'h', col = 'blue',
     xlab = 'Liczba skrętów w prawo', ylab = 'f(x)',
     xlim = c(0, 12.5), main = paste0('Funkcja prawdopodobieństwa dla M = ', n))

grid()

points(Freq ~ Arg, col = 'blue')

lines(dpois(Arg, lambda) ~ Arg, type = 'h', col = 'red')

points(dpois(Arg, lambda) ~ Arg, col = 'red')

legend('topright', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



Funckja prawdopodobieństwa wyznaczona metodą teoretyczną pokrywa się z funkcją wyznaczoną metodą empriczyną. W obydwu przypadkach auta skręcały najczęściej trzykrotnie.

Zadanie 2 (1 pkt)

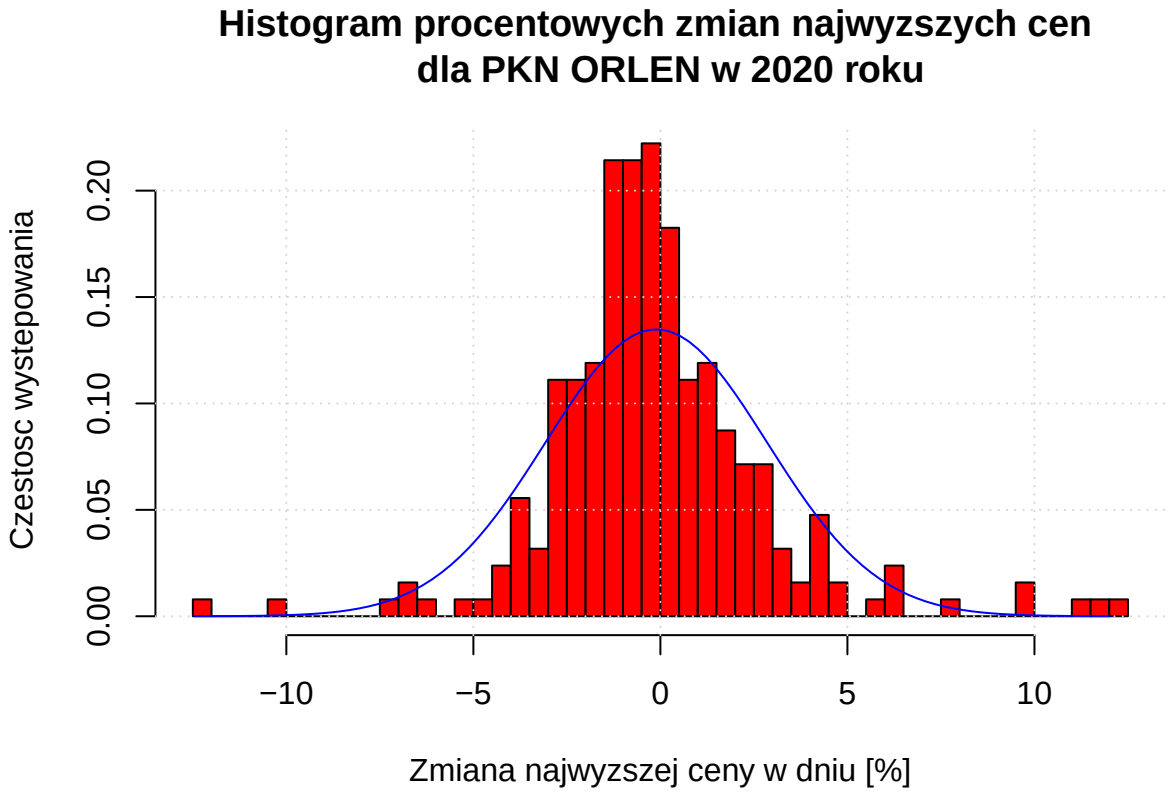
Treść zadania

- Dla wybranej jednej spółki notowanej na GPW oblicz wartości procentowych zmian najwyższych cen w dniu (high) dla roku 2020 i wykreśl ich histogram.

- Wyestymuj wartość średnią oraz wariancję procentowych zmian najwyższych cen w dniu dla wybranej spółki.
- Na podstawie histogramu i wykresu funkcji gęstości prawdopodobieństwa wyznaczonej dla wystymowanych parametrów (wartość średnia i wariancja) zweryfikuj zgrubnie, czy możemy przyjąć, że procentowe zmiany najwyższych cen w dniu mają rozkład normalny.
- Zakładając, że zmiany najwyższych cen w dniu mają rozkład normalny wyznacz 90%, 95% i 99% przedziały ufności dla wartości średniej i wariancji procentowych zmian najwyższych cen w dniu dla wybranej spółki. Przeanalizuj wyniki uzyskane dla różnych przedziałów ufności.

Rozwiązanie

```
if(!file.exists('mstall.zip')) {
download.file('https://info.bossa.pl/pub/metastock/mstock/mstall.zip','mstall.zip')
}

unzip('mstall.zip', files = c('PKNORLEN.mst'))
df_PKNORLEN = read.csv('PKNORLEN.mst')

col_names = c('ticker', 'date', 'open', 'high', 'low', 'close','vol')
names(df_PKNORLEN) = col_names

df_PKNORLEN$date = as.Date.character(df_PKNORLEN$date, format ='%Y%m%d')
df_PKNORLEN = df_PKNORLEN[which(df_PKNORLEN$date >= '2020-01-01' & df_PKNORLEN$date <= '2020-12-31'),]

df_PKNORLEN$high_ch = with(df_PKNORLEN, c(0, round(100*diff(high)/high[-length(high)], digits = 2)))

high_ch_mean = mean(df_PKNORLEN$high_ch, na.rm = T)
high_ch_sd = sd(df_PKNORLEN$high_ch, na.rm = T)
high_ch_var = var(df_PKNORLEN$high_ch, na.rm = T)

hist(df_PKNORLEN$high_ch, breaks = 50, prob = T,
     col = 'red', xlab = 'Zmiana najwyższej ceny w dniu [%]',
     ylab = 'Częstość występowania',
     main = 'Histogram procentowych zmian najwyższych cen \n  dla PKN ORLEN w 2020 roku')

grid()

min_c = min(df_PKNORLEN$high_ch, na.rm = T)
max_c = max(df_PKNORLEN$high_ch, na.rm = T)

curve(dnorm(x, mean = high_ch_mean,
sd = high_ch_sd), add = T, col = 'blue', from = min_c, to = max_c)
```



Wyestymowana wartość średniej wynosi -0.1100794, natomiast wariancji 8.7564119.

Procentowe zmiany najwyższych cen w dniu mają rozkład normalny, ponieważ naniesiona krzywa funkcji gęstości jest zbliżona do narysowanego historgramu. Maksimum występuję mniej więcej w tym samym miejscu tzn w punkcie -0.11.

```
n = length(df_PKNORLEN$high_ch)
S = high_ch_sd

lev90 = 0.9
w90 = S*qt((1+lev90)/2, n-1)/sqrt(n)
ci_mean_90 = c(high_ch_mean - w90, high_ch_mean + w90)
a = (1 - lev90)/2; b = (1 - lev90)/2
ci_var_90 = c((n-1)*S^2/qchisq(1-b,n-1), (n-1)*S^2/qchisq(a,n-1))
```

Granice 90 % przedziału ufności dla wartosci średniej wynoszą: -0.4178279, 0.1976691.

Granice 90 % przedziału ufności dla wariancji wynoszą 7.6062326, 10.2075035

```
lev95 = 0.95
w95 = S*qt((1+lev95)/2, n-1)/sqrt(n)
ci_mean_95 = c(high_ch_mean - w95, high_ch_mean + w95)
a = (1 - lev95)/2; b = (1 - lev95)/2
ci_var_95 = c((n-1)*S^2/qchisq(1-b,n-1), (n-1)*S^2/qchisq(a,n-1))
```

Granice 95 % przedziału ufności dla wartosci średniej wynoszą: -0.477201, 0.2570423.

Granice 95 % przedziału ufności dla wariancji wynoszą 7.4057791, 10.5155616.

```
lev99 = 0.99
w99 = S*qt((1+lev99)/2, n-1)/sqrt(n)
ci_mean_99 = c(high_ch_mean - w99, high_ch_mean + w99)
a = (1 - lev99)/2; b = (1 - lev99)/2
ci_var_99 = c((n-1)*S^2/qchisq(1-b,n-1), (n-1)*S^2/qchisq(a,n-1))
```

Granice 99 % przedziału ufności dla wartosci średniej wynoszą: -0.5939102, 0.3737515.

Granice 99 % przedziału ufności dla wariancji wynoszą 7.0340221, 11.154069.

Im większy poziom ufności zastosujemy, tym szerszy otrzymamy przedział zarówno dla średniej, jak i wariancji.

Zadanie 3 (1,5 pkt.)

Treść zadania

Rzucona pinezka upada ostrzem do dołu lub do góry. Doświadczenie to można opisać rozkładem Bernoulliego z parametrem p będącym prawdopodobieństwem tego, że pinezka upadnie ostrzem do góry.

Rozkład parametru p można opisać rozkładem beta o parametrach α i β. Wartość średnia i wariancja w rozkładzie beta zależą od parametrów rozkładu w następujący sposób:

$$\mathbb{E}X = \frac{\alpha}{\alpha+\beta}, \qquad \mathbb{V}X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \qquad dominanta = \frac{\alpha-1}{\alpha+\beta-2}.$$

- Na podstawie przypuszczanej (a priori) wartości oczekiwanej parametru p zaproponuj wartości parametrów α i β rozkładu a priori parametru p. Narysuj rozkład a priori parametru p (wykorzystaj funkcję dbeta()).
- Rzuć pinezką 20 razy i zanotuj wyniki kolejnych rzutów (1 - pinezka upada ostrzem do góry, 0 - pinezka upada ostrzem do dołu). Wyznacz i narysuj rozkład a posteriori parametru p oraz oblicz wartość bayesowskiego estymatora $\hat{p}$. W rozważanym przypadku rozkład aposteriori parametru p jest również rozkładem beta o parametrach:

$$\alpha_{\text{post}} = \alpha_{\text{prior}} + \sum_{i=1}^{n} x_i, \qquad \beta_{\text{post}} = \beta_{\text{prior}} + n - \sum_{i=1}^{n} x_i, \qquad x_i \in \{0,1\}.$$

- Rzuć pinezką jeszcze 20 razy i zanotuj wyniki. Wyznacz i narysuj rozkład a posteriori oparty na wszystkich 40 rzutach oraz oblicz wartość bayesowskiego estymatora $\hat{p}$ w tym przypadku. Porównaj wyniki z wynikami uzyskanymi po pierwszych 20 rzutach.
- Korzystając ze wzoru na wariancję rozkładu Beta wyznacz i porównaj wariancje rozkładu a priori, a posteriori po 20 rzutach i a posteriori po 40 rzutach.

Rozwiązanie

Flanki - gra zespołowa, polegająca na próbie trafienia czymkolwiek w pustą puszkę stojącą po 3 metry od obu drużyn. Gdy puszka zostanie przewrócona, drużyna która tego dokonała pije piwo “na hejnał”, a przeciwnik musi jak najszybciej podnieść przewróconą puszkę, wbiec z powrotem za linie, gdy przekroczymy linie krzyczymy stop, a przeciwna drużyna przestaje pić. Przyjmujemy, że strącenie puszki to 1, a chybienie - 0. Prawdopodobieństwo szacujemy na p = 0.2. Alfa i beta przyjmujemy ze wzoru powyżej.

```
p = 0.2
alpha_priori = 2
beta_priori = 8

a_priori = alpha_priori * beta_priori
b_priori = (alpha_priori + beta_priori)^2
c_priori = alpha_priori + beta_priori + 1

var_priori = a_priori / (b_priori * c_priori)
```

Wykonano 20 rzutów. Wyniki to 2 trafienia w puszkę, 18 chybień. Obliczamy a posteriori alfę i betę zgodnie z tabelą w wykładu.

```
hit_round_1 = 2
missed_round_1 = 18

alpha_round_1 = alpha_priori + hit_round_1
beta_round_1 = beta_priori + missed_round_1

p_round_1 = alpha_round_1 / (alpha_round_1 + beta_round_1)

a_round_1 = alpha_round_1 * beta_round_1
b_round_1 = (alpha_round_1 + beta_round_1)^2
c_round_1 = alpha_round_1 + beta_round_1 + 1

var_round_1 = a_round_1 / (b_round_1 * c_round_1)
```

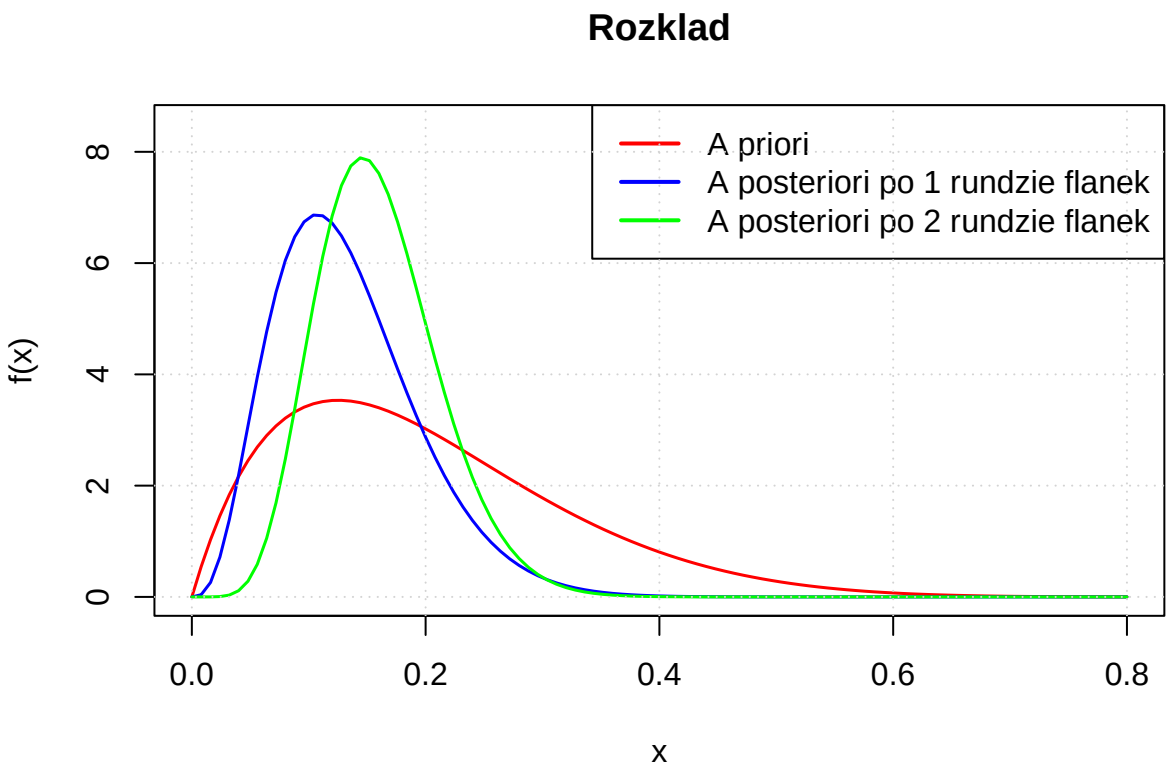
W drugiej rundzie flanek uczestnicy wykonali kolejne 20 rzutów. Puszkę strącono 4 razy.

```
hit_round_2 = 4
missed_round_2 = 16

alpha_round_2 = alpha_round_1 + hit_round_2
beta_round_2 = beta_round_1 + missed_round_2

p_round_2 = alpha_round_2 / (alpha_round_2 + beta_round_2)

a_round_2 = alpha_round_2 * beta_round_2
b_round_2 = (alpha_round_2 + beta_round_2)^2
c_round_2 = alpha_round_2 + beta_round_2 + 1

var_round_2 = a_round_2 / (b_round_2 * c_round_2)
```

Rykres rozkładów a priori oraz a posteriori po dwóch rundach flanek.

```
curve(dbeta(x, shape1 = alpha_priori, shape2 = beta_priori),
      lwd = 1.5, col = 'red', xlab = 'x', ylab = 'f(x)',
      main = 'Rozkład', xlim = c(0, 0.8), ylim = c(0, 8.5))

curve(dbeta(x, shape1 = alpha_round_1, shape2 = beta_round_1),
      add = TRUE, lwd = 1.5, col = 'blue', xlim = c(0, 0.8),
      ylim = c(0, 8.5))
```

```
curve(dbeta(x, shape1 = alpha_round_2, shape2 = beta_round_2), add = TRUE,
      lwd = 1.5, col = 'green', xlim = c(0, 0.8), ylim = c(0, 8.5))

legend('topright', c('A priori', 'A posteriori po 1 rundzie flanek',
      'A posteriori po 2 rundzie flanek'), col = c('red', 'blue', 'green'), lwd = 2)

grid()
```



Wartość bayesowskiego estymatora $\hat{p}$ po pierwszej oraz drugiej rundzie wynosi kolejno 0.1333333, 0.16.

Wariancja dla rozkładu a priori oraz a posteriori po pierwszej oraz drugiej rundzie flanek wynosi kolejno 0.0145455, 0.0037276, 0.0026353.

Wyznaczone wyniki są minimalnie mniejsze od założonego prawdopodbieństwa na początku zadania. Na wynik końcówy miało wpływ wiele czynniów takich jak oświetlenie boiska, wiatr czy ilość wypitego piwa.

Zadanie 4 (1,5 pkt.)

Treść zadania

Plik fotony.txt zawiera odstępy między chwilami rejestracji kolejnych fotonów promieniowania gamma wykonywanymi za pomocą teleskopu kosmicznego Comptona (CGRO) w roku 1991.

- Wczytaj dane za pomocą komendy scan(‘fotony.txt’)

- Metodą momentów oraz metodą największej wiarygodności wyznacz estymaty parametrów rozkładu gamma odpowiadające zarejestrowanym danym. Porównaj wyniki uzyskane dla obu metod.
- Narysuj na jednym wykresie histogram odstępów oraz funkcje gęstości rozkładu gamma o parametrach wyestymowanych za pomocą obu metod.
- Metodą bootstrapu parametrycznego wyznacz dla obu metod (momentów oraz największej wiarygodności) odchylenia standardowe estymatorów parametrów rozkładu gamma (α i β) oraz ich przedziały ufności na poziomie ufności 95%. Porównaj wyniki uzyskane dla obu metod.

```
photons = scan('http://elektron.elka.pw.edu.pl/~mrupniew/adps/fotony.txt')
n = length(photons)
quiet = TRUE
```

W celu wyznaczenia wartości parametrów rozkładu gamma α, β metodą momentów korzystamy z poniższych wzorów:

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}, \qquad \hat{\beta} = \frac{m_2 - m_1^2}{m_1}.$$

```
m1 = mean(photons)
m2 = mean(photons^2)

alpha_mom = m1^2/(m2 - m1^2)
beta_mom = (m2 - m1^2)/m1
```

Wartości estymatorów parametrów wyznaczone metodą momentów wynoszą: $\hat{\alpha} = 1.0655417$, $\hat{\beta} = 73.6240637$.

W celu wyznaczenia wartości parametru α oraz $\hat{\beta}$ metodą największej wiarygodności, korzystamy z funkcji fitdistr() z pakietu MASS:

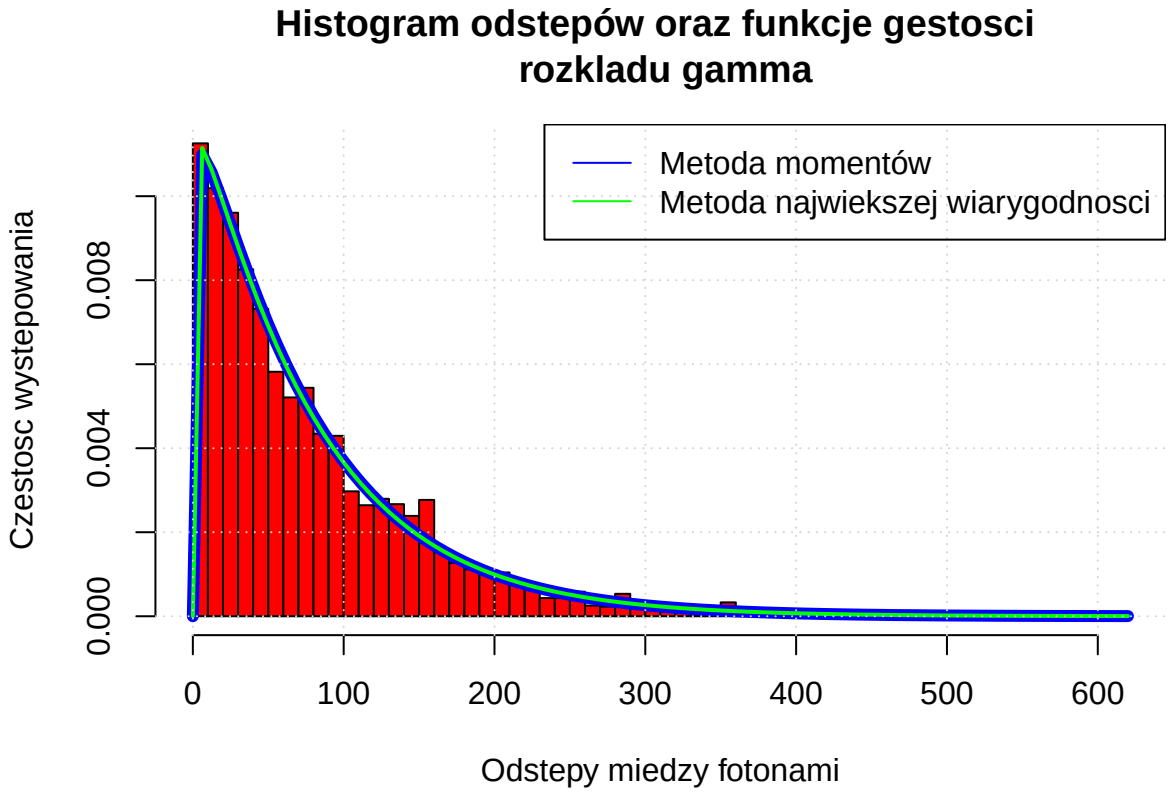
```
require(MASS)
```

```
## Loading required package: MASS
```

```
est_nw = fitdistr(photons, 'gamma', list(shape=1, scale=1), lower=0)
alpha_nw = as.numeric(est_nw$estimate[1])
beta_nw = as.numeric(est_nw$estimate[2])
```

Wartości estymatorów parametrów wyznaczone metodą największej wiarygodności z wykorzystniem funkcji fitdistr() wynoszą: $\hat{\alpha} = 1.0519734$, $\hat{\beta} = 74.5736621$.

```
hist(photons, breaks = 50, prob = T, col = 'red',
     xlab = 'Odstępy między fotonami', ylab = 'Częstość występowania',
  main = 'Histogram odstępów oraz funkcje gęstości \n rozkładu gamma')

curve(dgamma(x, shape = alpha_mom, scale = beta_mom), add = T,
      col = 'blue', lwd = 6)

curve(dgamma(x, shape = alpha_nw, scale = beta_nw), add = T,
      col = 'green', lwd = 2)

grid()

legend('topright', c('Metoda momentów', 'Metoda największej wiarygodności'), col = c('blue', 'green'),
```




```
K = 1000
boot_res_m = replicate(K, {
boot_dane = rgamma(n, shape = alpha_mom, scale = beta_mom)
c(mean(boot_dane)^2/(mean(boot_dane^2)-mean(boot_dane)^2),(mean(boot_dane^2)-mean(boot_dane)^2)/mean(boo
} )

sd_alpha_mom = sd(boot_res_m[1,])
sd_beta_mom = sd(boot_res_m[2,])
```

Odchylenie standardowe dla estymatora wartości alfy wyznaczonej metodą momentów wynosi 0.0334349.

Odchylenie standardowe dla estymatora wartości bety wyznaczonej metodą momentów wynosi 2.6257983.

```
lev = 0.95
int_alpha_mom = quantile(boot_res_m[1,], c((1-lev)/2,(1+lev)/2))
int_beta_mom = quantile(boot_res_m[2,], c((1-lev)/2,(1+lev)/2))
```

Granice 95 % przedziału ufności dla estymatora wartości alfa wyznaczonej metodą momentów wynoszą: 1.0051692, 1.135042.

Granice 95 % przedziału ufności dla estymatora wartości bety wyznaczonej metodą momentów wynoszą 68.6209183, 78.7087646.

```
K = 1000
boot_res_nw = replicate(K, {
boot_dane = rgamma(n, shape = alpha_nw, scale = beta_nw)
```

```
c(mean(boot_dane)^2/(mean(boot_dane^2)-mean(boot_dane)^2),(mean(boot_dane^2)-mean(boot_dane)^2)/mean(boo
} )

sd_alpha_nw = sd(boot_res_nw[1,])
sd_beta_nw = sd(boot_res_nw[2,])
```

Odchylenie standardowe dla estymatora wartości alfy wyznaczonej metodą największej wiarygodności wynosi 0.0314818.

Odchylenie standardowe dla estymatora wartości bety wyznaczonej metodą największej wiarygodności wynosi 2.5087404.

```
lev = 0.95
int_alpha_nw = quantile(boot_res_nw[1,], c((1-lev)/2,(1+lev)/2))
int_beta_nw = quantile(boot_res_nw[2,], c((1-lev)/2,(1+lev)/2))
```

Granice 95 % przedziału ufności dla estymatora wartości alfa wyznaczonego metodą największej wiarygodności wynoszą: 0.9910198, 1.1171353.

Granice 95 % przedziału ufności dla beta wyznaczonej metodą największej wiarygodności wynoszą 69.6589809, 79.6537506.

Obie metody wyznaczenia estymatorów parmetru α oraz $\hat{\beta}$ dały nam podobne wyniki, a narysowane funkcje gęstości pokrywają się. Świadczy to o poprawności obu metod, a tą któą wybierzemy w naszych obliczeniach, zależy wyłącznie od nas.
