ADPS 25L — Laboratorium 2 (rozwiązania)

Konrad Lis

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, zapisz jego wartość w sprawozdaniu.
- 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.
- Metodą bootstrapu nieparametrycznego oszacuj odchylenie standardowe estymatora parametru λ, zapisz jego wartość w sprawozdaniu.

Rozwiązanie

- Wczytanie danych

```
skrety = scan('skrety.txt')
```

- Dopasowanie danych

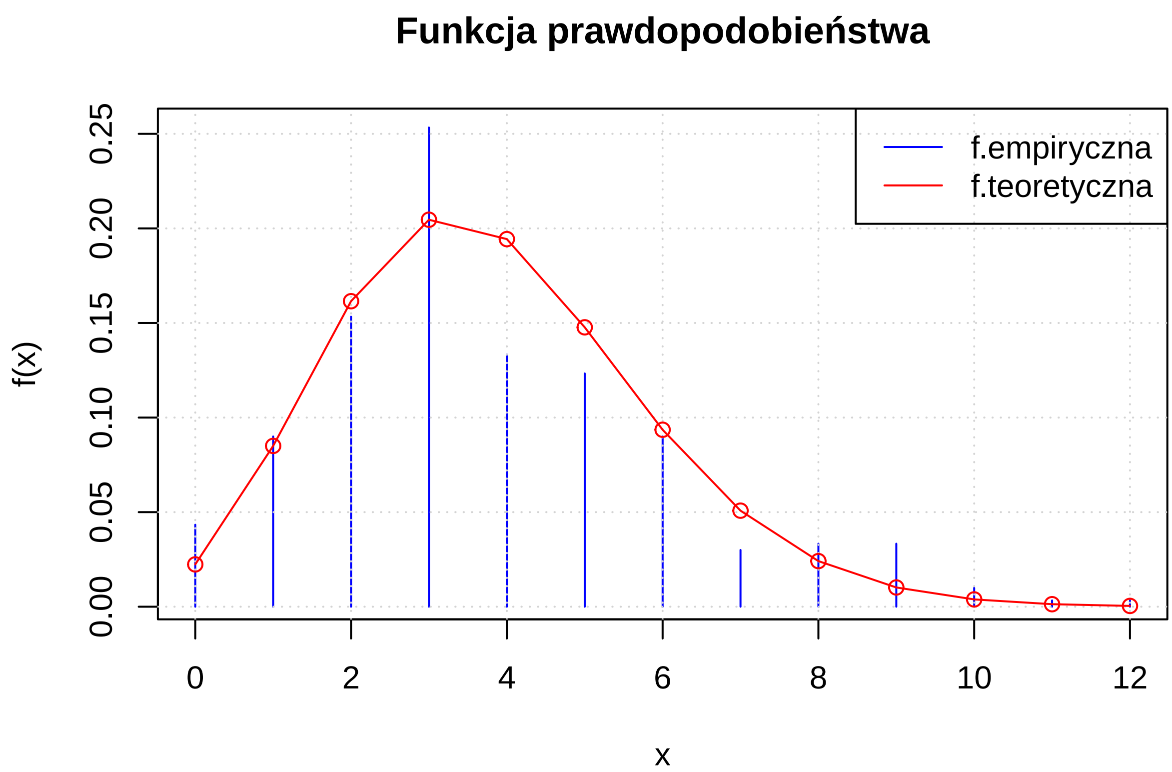
```
lambda = mean(skrety)
```

Wartość wyestymowanego parametru λ wynosi 3.8.

- Zgodność rozkładu z zarejestrowanymi danymi

```
Arg = 0:max(skrety)
Freq = as.numeric(table(factor(skrety, levels = Arg))) / length(skrety)

plot(Freq ~ Arg, type = 'h', col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = paste0('Funkcja prawdopodobieństwa'))

grid()

lines(dpois(Arg, lambda) ~ Arg, type = 'l', col = 'red',
    xlab = 'x', ylab = 'f(x)')
points(dpois(Arg, lambda) ~ Arg, col = 'red')
```

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



Dane empiryczne z są zbliżone do Rozkładu Poissona

- Bootstrap nieparametryczny

```
K = 10000

boot_lambda = replicate(K, {
    boot_dane = sample(skrety, length(skrety), replace = T)
    c(mean(boot_dane))
})
sd_lambda = sd(boot_lambda)
```

Odchylenie standardowe wynosi 0.1311

Zadanie 2 (1 pkt)

Treść zadania

- Dla wybranej jednej spółki notowanej na GPW oblicz wartości procentowych zmian najwyższych cen w dniu (high) w ciągu ostatnich dwóch lat i wykreśl ich histogram.

- Wyestymuj wartość średnią oraz wariancję procentowych zmian najwyższych cen dla wybranej spółki, zapisz te wartości w sprawozdaniu.
- Na podstawie histogramu i wykresu funkcji gęstości prawdopodobieństwa wyznaczonej dla wyestymowanych 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. Porównaj wyniki uzyskane dla różnych przedziałów ufności.

Rozwiązanie

Wartość procentowych zmian HIGH na przykładzie spółki COGNOR "COG"

- Wczytanie danych

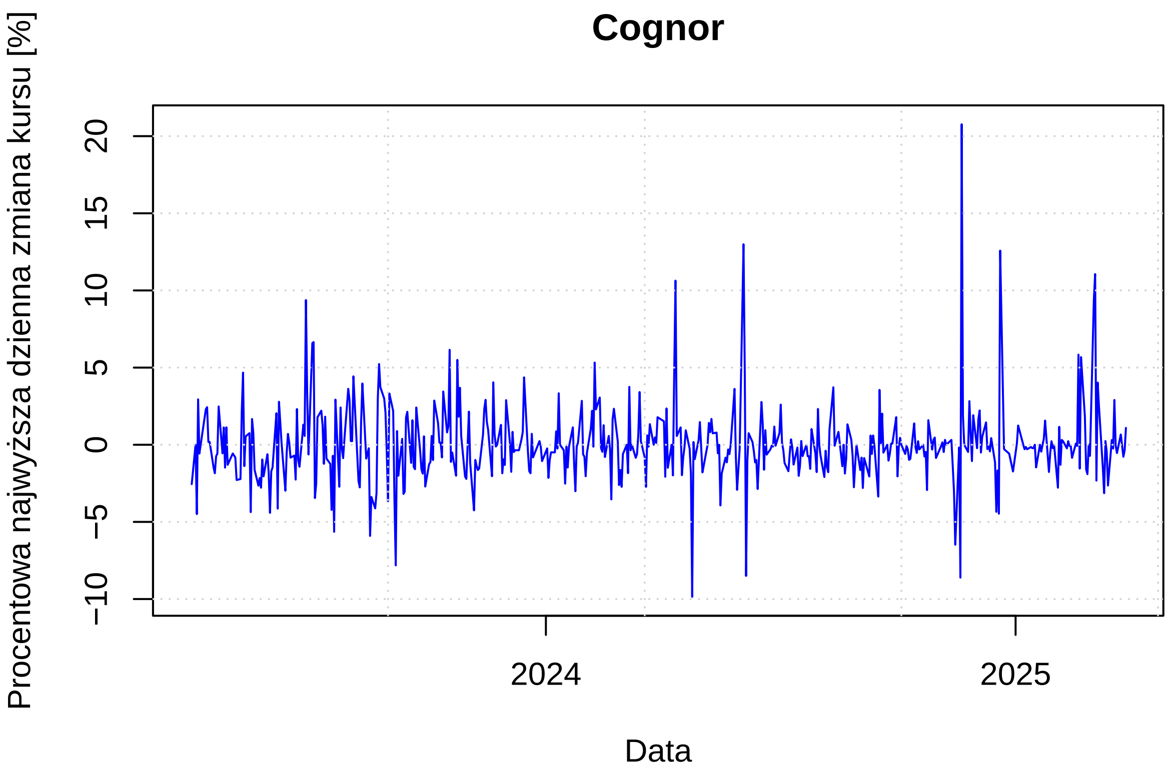
```
Ticket = 'COG'
webLink = paste0('https://stooq.pl/q/d/l/?s=', Ticket, '&i=d')
fileName = paste0(Ticket, '.csv')
# if(!file.exists(fileName)) {
  download.file(webLink, fileName)
#}

df_COG = read.csv('COG.csv')
df_COG$Data = as.Date(df_COG$Data)
```

- Ograniczenie danych do 2 lat

```
df_COG_2lata = df_COG[which(df_COG$Data >=
    '2023-03-30' & df_COG$Data <= '2025-03-30'),]
```

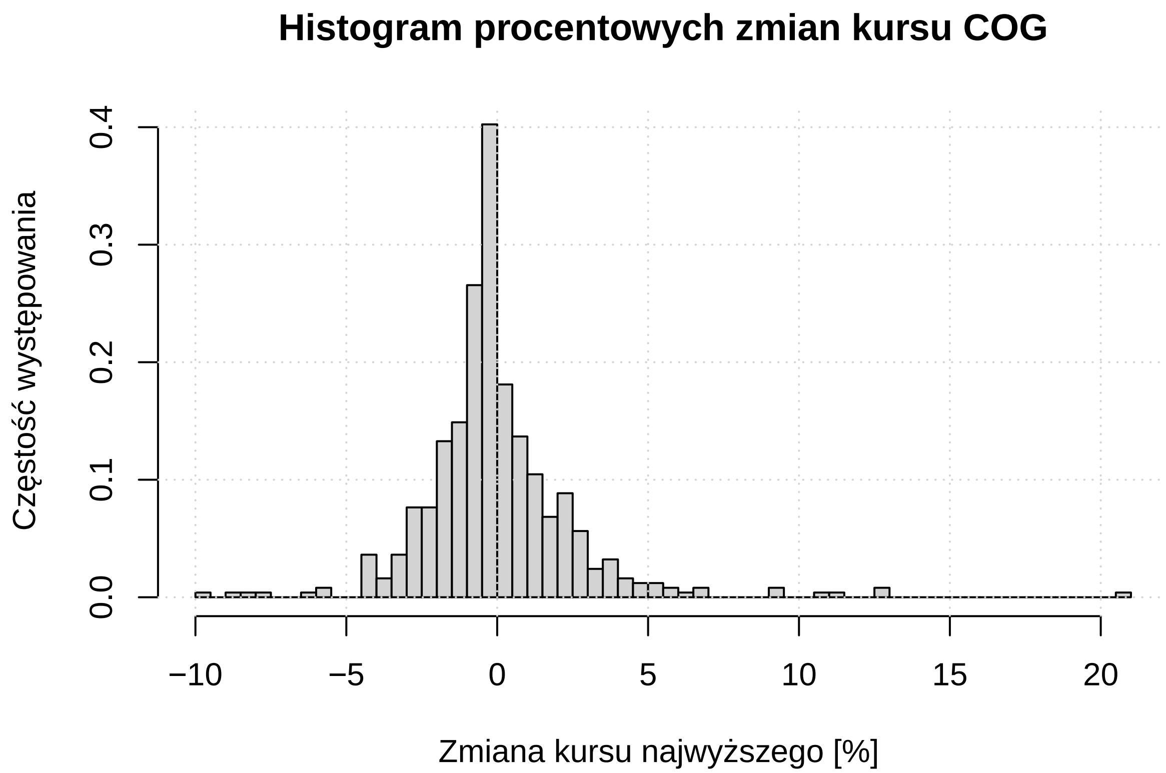
- Procentowa dzienna najwyższa zmienność

```
df_COG_2lata$Najwyższy_zm = with(df_COG_2lata, c(NA, 100*diff(Najwyzszy)/Najwyzszy[-length(Najwyzszy)]))
plot(Najwyższy_zm ~ Data, df_COG_2lata, type = 'l', col = 'blue', xlab = 'Data',
     ylab = 'Procentowa najwyższa dzienna zmiana kursu [%]', main = 'Cognor')
grid()
```



- Histogram

```
hist(df_COG_2lata$Najwyższy_zm, breaks = 50, prob = T,
  xlab = 'Zmiana kursu najwyższego [%] ',
  ylab = 'Częstość występowania',
  main = paste('Histogram procentowych zmian kursu', 'COG')  )
grid()
```



- Średnia i wariancja najwyższych dziennych zmian procentowych

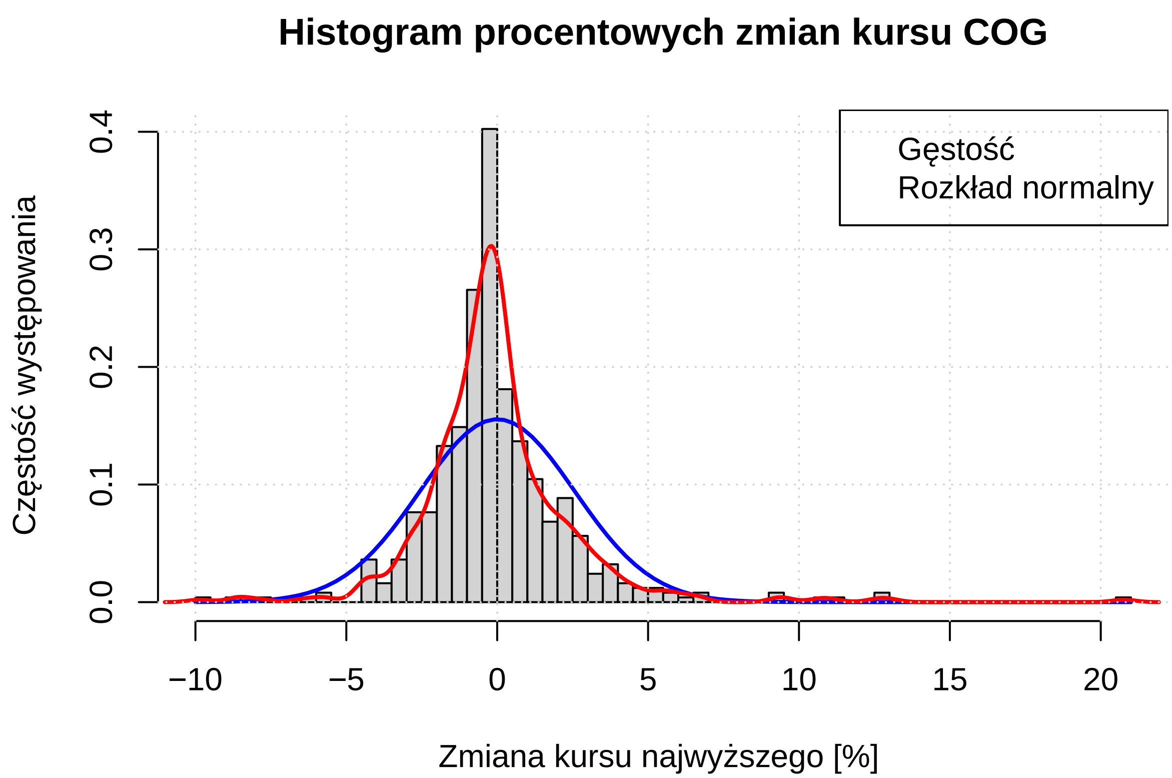
```
Cog_mean = mean(df_COG_2lata$Najwyższy_zm, na.rm = T)
Cog_var = var(df_COG_2lata$Najwyższy_zm, na.rm = T)
```

Wartość średnia 0.002

Wariancja 6.579

- Funkcja gęstości oraz histogram

```
hist(df_COG_2lata$Najwyższy_zm, breaks = 50, prob = T,
    xlab = 'Zmiana kursu najwyższego [%] ',
    ylab = 'Częstość występowania',
    main = paste('Histogram procentowych zmian kursu', 'COG'))
curve(dnorm(x,
            mean = Cog_mean,
            sd = sqrt(Cog_var)),
    add = T, col = 'blue', lwd = 2)
lines(density(na.omit(df_COG_2lata$Najwyższy_zm)), col = "red", lwd = 2)
grid()

legend("topright", legend = c("Gęstość", "Rozkład normalny"),
    col = c('red', 'blue'))
```



Zgrubnie, możemy przyjąć, że procentowe zmiany najwyższych cen w danych dniu mają rozkład zbliżony do normalnego.

- Przedziały ufności

```
n = length(df_COG_2lata$Najwyższy_zm)
```

- Przedział ufności 90%

```
lev_90 = 0.9
S = sqrt(Cog_var)
w = S * qt((1 + lev_90) / 2, df = n - 1) / sqrt(n)
mean_conf_int_90 = c(Cog_mean - w, Cog_mean + w)
a = (1 - lev_90) / 2; b = (1 - lev_90) / 2
var_conf_int_90 = c((n - 1) * S^2 / qchisq(1 - b, df = n - 1),
(n - 1) * S^2 / qchisq(a, df = n - 1))
```

- Przedział ufności 95%

```
lev_95 = 0.95
S = sqrt(Cog_var)
w = S * qt((1 + lev_95) / 2, df = n - 1) / sqrt(n)
mean_conf_int_95 = c(Cog_mean - w, Cog_mean + w)
a = (1 - lev_95) / 2; b = (1 - lev_95) / 2
var_conf_int_95 = c((n - 1) * S^2 / qchisq(1 - b, df = n - 1),
(n - 1) * S^2 / qchisq(a, df = n - 1))
```

- Przedział ufności 99%

```
lev_99 = 0.99
S = sqrt(Cog_var)
w = S * qt((1 + lev_99) / 2, df = n - 1) / sqrt(n)
mean_conf_int_99 = c(Cog_mean - w, Cog_mean + w)
a = (1 - lev_99) / 2; b = (1 - lev_99) / 2
var_conf_int_99 = c((n - 1) * S^2 / qchisq(1 - b, df = n - 1),
(n - 1) * S^2 / qchisq(a, df = n - 1))
```

Przedziały ufności dla wartości średnich wynoszą: [0.187, 0192] dla 90%, [0.223, 0.229] dla 95%, [0.294, 0.3] dla 99%.

Przedziały ufności dla wariancji wynoszą: [5.95, 7.33] dla 90%, [5.83, 7.48] dla 95%, [5.62, 7.79] dla 99%.

Porównanie: Im wyższy przedział ufności, tym szerszy zakres wartości.

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:

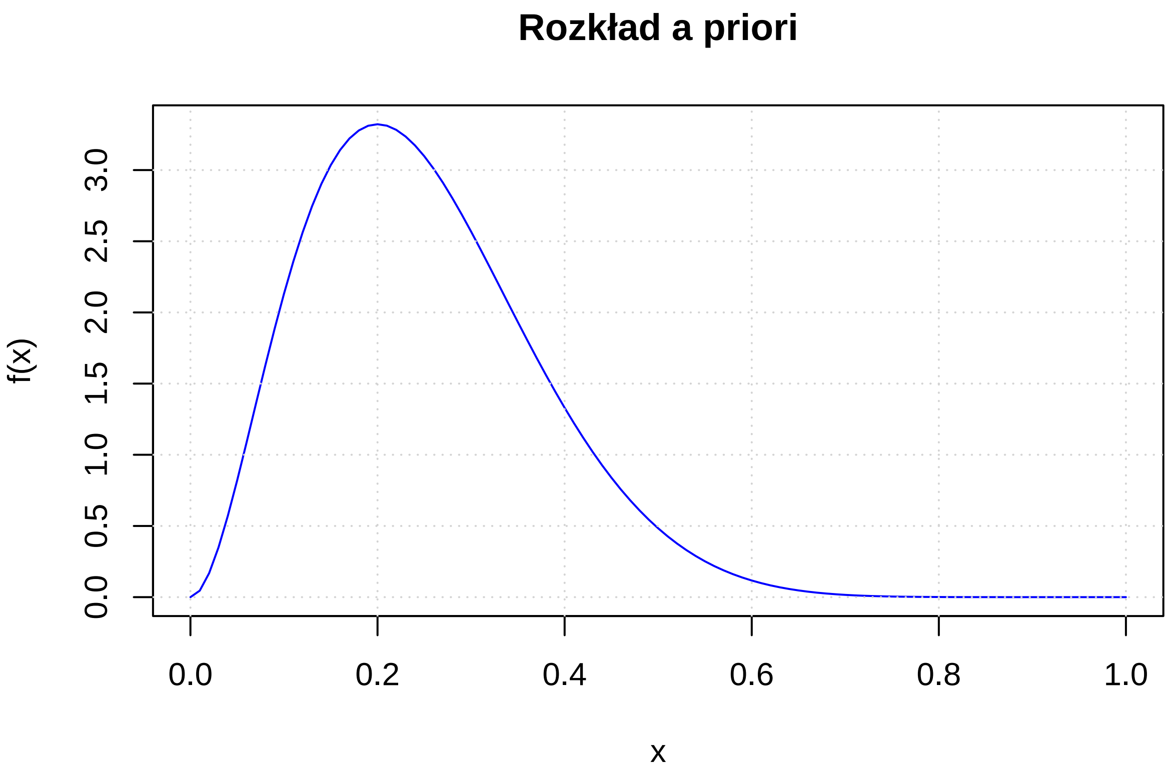
$$\alpha_{\text{post}} = \alpha_{\text{prior}} + \sum_{i=1}^{n} x_i, \qquad \beta_{\text{post}} = \beta_{\text{prior}} + n - \sum_{i=1}^{n} x_i, \qquad x_i \in \{0,1\}.$$

- 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 rozkładów a priori, a posteriori po 20 rzutach i a posteriori po 40 rzutach.

Rozwiązanie

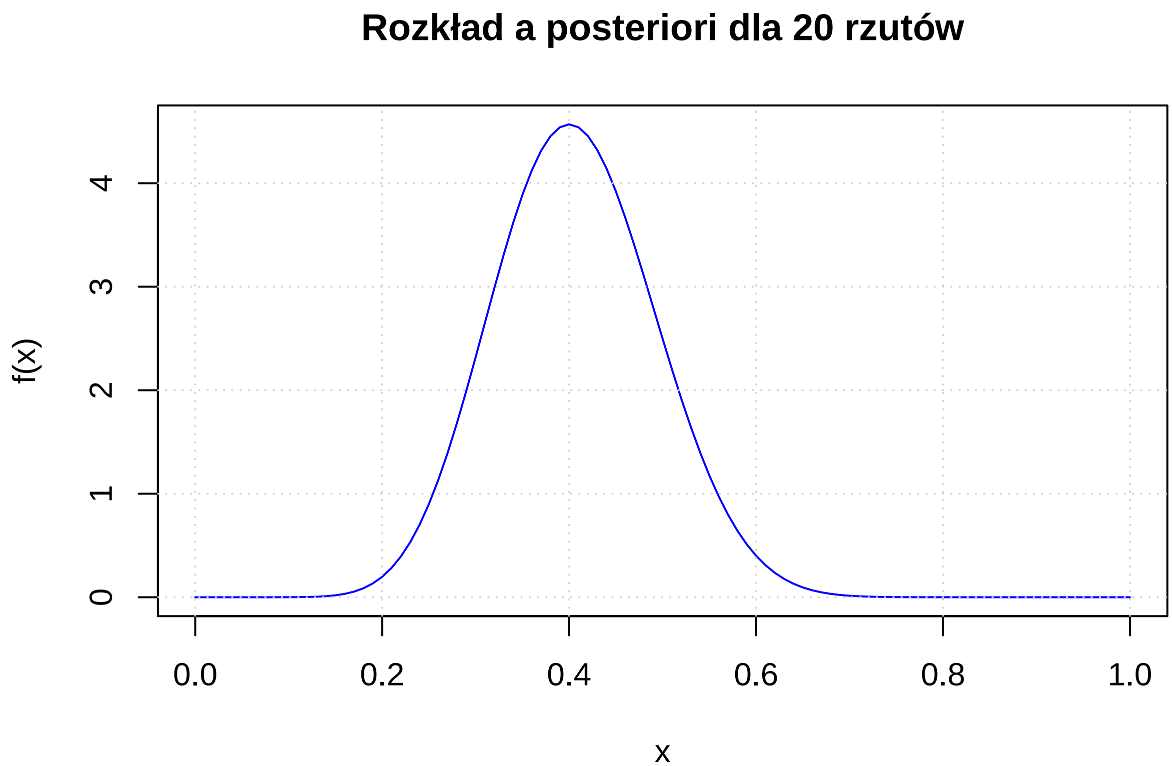
- Proponowane wartości oraz rozkład a priori

```
alpha_prior = 3
beta_prior = 9

curve(dbeta(x, alpha_prior, beta_prior),
    col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = 'Rozkład a priori')
grid()
```



- Rzuty pinezką i rozkład a posteriori

```
rzuty = c(0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0)

suma_1 = sum(rzuty)
n = length(rzuty)

alpha_post = alpha_prior + suma_1
beta_post = beta_prior + n - suma_1

curve(dbeta(x, alpha_post, beta_post),
    col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = 'Rozkład a posteriori dla 20 rzutów')
grid()
```



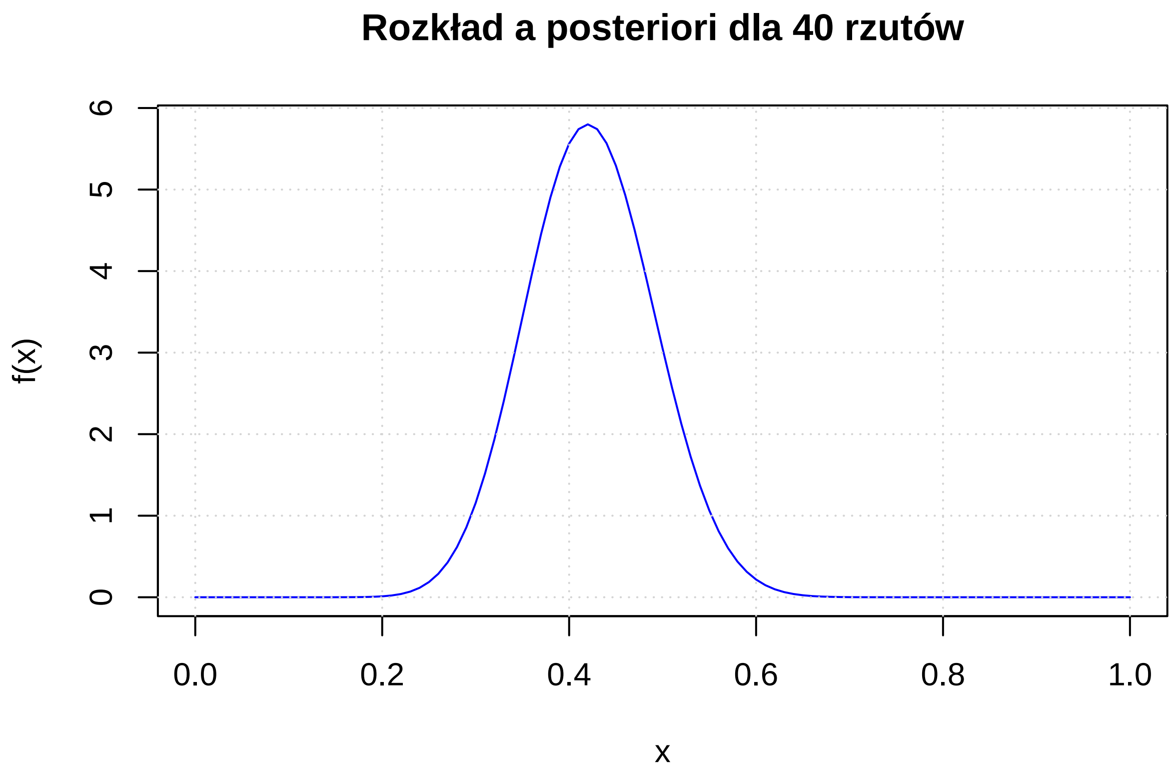
* Wartość estymatora Bayesowskiego

```
hat = alpha_post / (alpha_post + beta_post)
```

Wartość dla estymatora Bayesowskiego wynosi 0.40625.

- Ponowne rzuty pinezką i rozkład a posteriori oparty o wszystkie rzuty

```
rzuty_ponowne = c(1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1)

rzuty_wszystkie = c(rzuty, rzuty_ponowne)
suma_1_wszystkie = sum(rzuty_wszystkie)
n_wszystkie = length(rzuty_wszystkie)

alpha_post_p = alpha_prior + suma_1_wszystkie
beta_post_p = beta_prior + n_wszystkie - suma_1_wszystkie

curve(dbeta(x, alpha_post_p, beta_post_p),
    col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = 'Rozkład a posteriori dla 40 rzutów')
grid()
```



- Wartość estymatora Bayesowskiego dla wszystkich rzutów

```
hat_wszystkie = alpha_post_p / (alpha_post_p + beta_post_p)
```

Wartość dla estymatora Bayesowskiego wynosi 0.423.

- Porównanie wariancji
- Rozkład a priori

```
var_prior <- (alpha_prior * beta_prior) / ((alpha_prior + beta_prior)^2 * (alpha_prior + beta_prior + 1)
```

- Rozkład a posteriori po 20 rzutach

```
var_post1 <- (alpha_post * beta_post) / ((alpha_post + beta_post)^2 * (alpha_post + beta_post + 1))
```

- Rozkład a posteriori po 40 szutach

```
var_post2 <- (alpha_post_p * beta_post_p) / ((alpha_post_p + beta_post_p)^2 * (alpha_post_p + beta_post_
```

Wartości wariancji rozkładów: A priori: 0.014 A posteriori po 20 rzutach: 0.007 A posteriori po 40 rzutach: 0.004 Z otrzymanych wartości wynika, że im większa liczba rzutów, tym wyższa dokładność danych.

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.

Rozwiązanie

- Wczytanie danych

```
odstepy = scan('fotony.txt')
```

- Estymaty parametrów

```
mom1 = mean(odstepy)
mom2 = mean(odstepy^2)
alpha_mom = mom1^2/(mom2 - mom1^2)
beta_mom = (mom2 - mom1^2)/mom1
```

```

require(MASS)
```

Wartości estymatorów parametrów wyznaczone metodą momentów wynoszą: α 1.06 oraz β 73.62.

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

```
## Loading required package: MASS
est_nw = fitdistr(odstepy, '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: $\hat{} = 1.0519, \hat{} = 74.573$.

Metoda momentów jak i największej wairygodności dają podobne rezultaty.

- Histogram odstępów oraz funkcja gęstości

```
hist(odstepy, probability = TRUE, breaks = 100, xlab = 'Odstępy', ylab = 'Gęstość',
    main = paste('Odstępy między fotonami'))
curve(dgamma(x, shape = alpha_mom, scale = beta_mom), add = T, col = 'blue', lwd = 1)
curve(dgamma(x, shape = alpha_nw, scale = beta_nw), add = T, col = 'red', lwd = 1)
grid()
```

```r
legend('topright', c('Metoda momentów', 'Metoda największej wiarygodności'),
    lty = c(1,1), lwd = c(1.5,1.5), col = c('blue','red'))
```