ADPS 2020Z — Laboratorium 2 (rozwiązania)

Jakub Adamowicz

Zadanie 1

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 - wyestymuj parametr λ.
- Metodą bootstrapu nieparametrycznego oszacuj odchylenie standardowe estymatora parametru λ.
- Sprawdź 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

Wczytanie danych i estymacja parametru λ

```
x = scan('skrety.txt')
lambda_est = mean(x)
```

Wyestymowany parametr λ wynosi 3.8.

Oszacowanie odchylenia standardowego estymatora parametry λ metodą bootstrapu nieparametrycznego

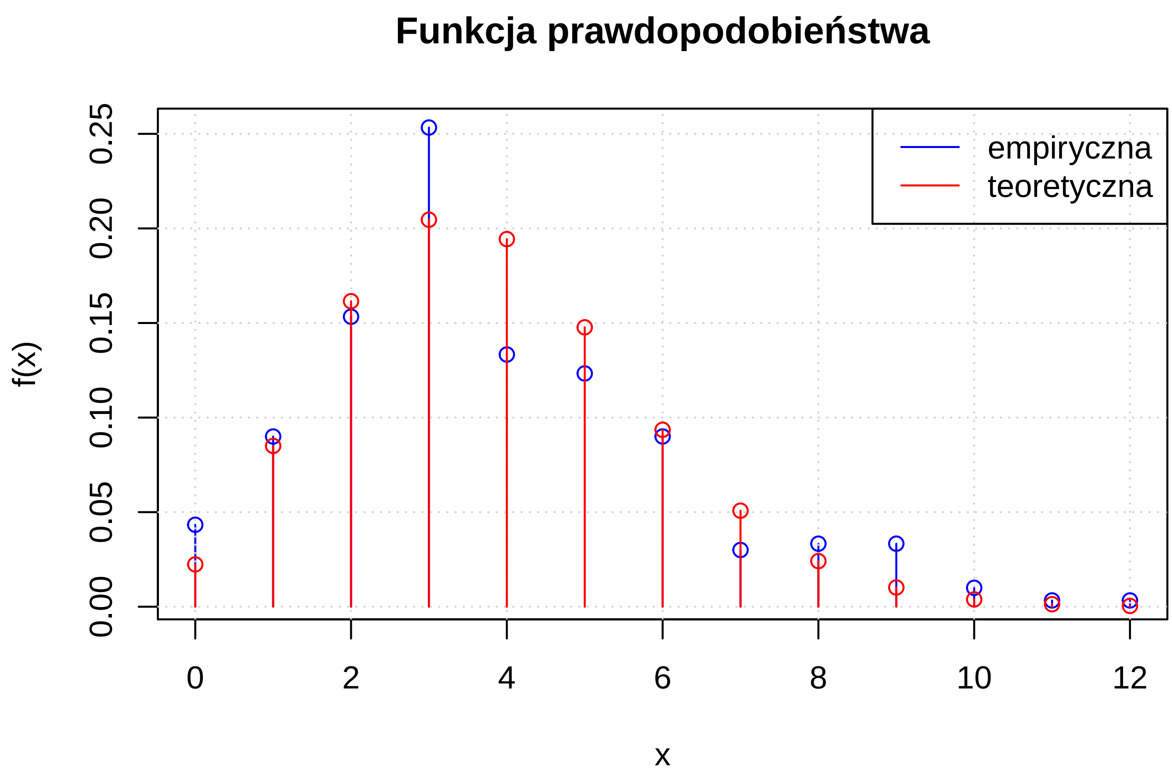
```
K = 1000
boot_res = replicate(K, {
  boot_dane = sample(x, length(x), replace = T)
  c(mean(boot_dane))
})
sd_mean = sd(boot_res)
```

Odchylenia standardowe estymatora λ wynosi 0.1367.

Stworzenie wykresów empirycznej i teoretycznej funkcji prawdopodobieństwa

```
Arg = 0:max(x)
Freq = as.numeric(table(factor(x, levels = Arg))) / length(x)
plot(Freq ~ Arg, type = 'h', col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = paste0('Funkcja prawdopodobieństwa'))
grid()
points(Freq ~ Arg, col = 'blue')

lines(dpois(Arg, lambda = 3.8) ~ Arg, type = 'h', col = 'red',
      xlab = 'x', ylab = 'f(x)')
points(dpois(Arg, lambda = 3.8) ~ Arg, col = 'red')
```

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



Zadanie 2

Treść zadania

- Dla wybranej jednej spółki notowanej na GPW oblicz wartości procentowych zmian najwyższych cen w dniu (high) i wykreśl ich histogram.
- Wyestymuj wartość średnią oraz wariancję procentowych zmian najwyższych cen w dniu dla wybranej spółki.
- Na podstawie histogramu i wykresu empirycznej funkcj gęstości prawdopodobieństwa 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.

Rozwiązanie

Załadowanie danych spółki PLAY i obliczenie wartości procentowych zmian najwyższych cen w dniu

```r
unzip('mstall.zip', 'PLAY.mst')
df_PLAY = read.csv('PLAY.mst')
names(df_PLAY) = c('ticker', 'date', 'open', 'high', 'low', 'close','vol')
```

```
df_PLAY$date = as.Date.character(df_PLAY$date, format ='%Y%m%d')
df_PLAY$high_ch= with(df_PLAY, c(NA, 100*diff(high)/high[-length(high)]))
```

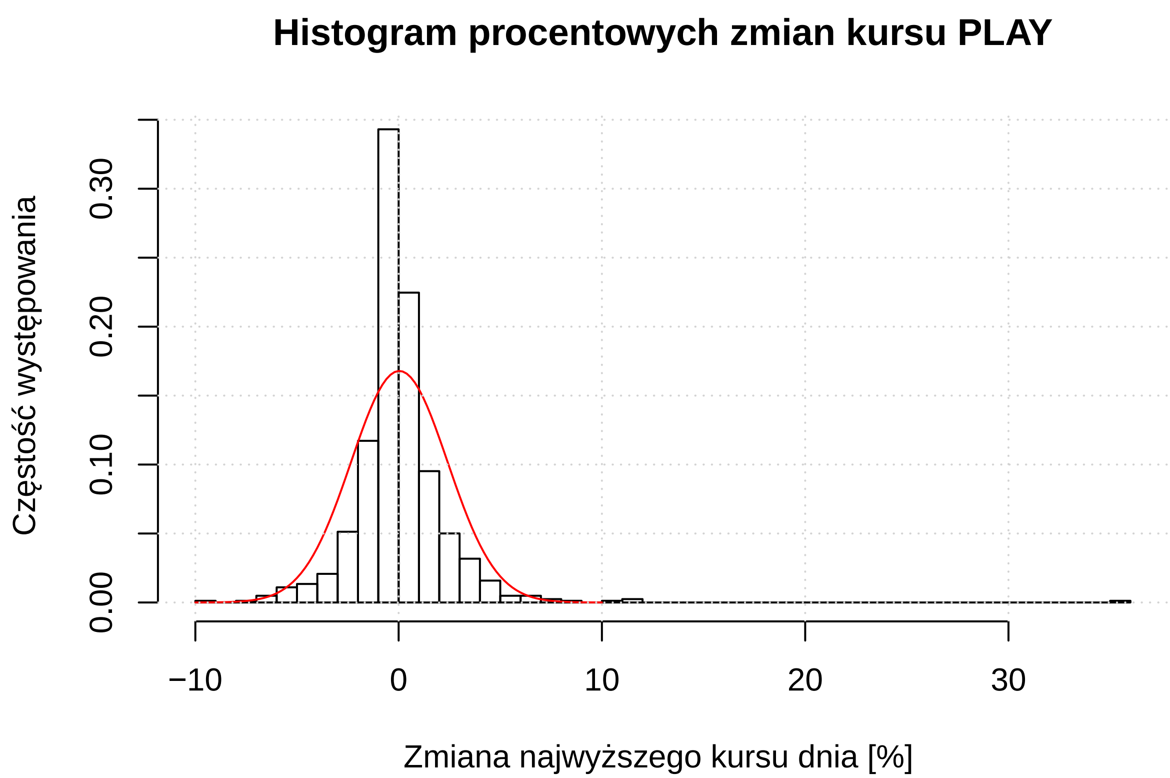
Estymacja wartości średniej i wariancji procentowych zmian najwyższych cen w dniu dla spółki PLAY

```
mean_est = mean(df_PLAY$high_ch, na.rm=T)
var_est = var(df_PLAY$high_ch, na.rm=T)
sd_est = sd(df_PLAY$high_ch, na.rm=T)
```

Wartość średnia zmian procentowych wynosi 0.03479. Wartość wariancji zmian procentowych wynosi 5.6474.

Wykreślenie histogramu zmian procentowych i empirycznej funkcji gęstości

```
hist(df_PLAY$high_ch, breaks = 50, prob = T,
xlab = 'Zmiana najwyższego kursu dnia [%] ',
ylab = 'Częstość występowania',
main = 'Histogram procentowych zmian kursu PLAY' )
curve(dnorm(x, mean = mean_est, s = sd_est), add = T, col = 'red', -10, 10)
grid()
```



Ze względu na wysoką kurtozę (K>0) nie można z dużą pewnością określić czy rozkład ma postać rozkładu normalnego.

Wyznaczenie przedziałów ufności dla wartości średniej i wariancji zmian procentowych

```
lev90= 0.9
lev95 = 0.95
lev99 = 0.99
```

```
n = length(df_PLAY$high_ch)
alfa90 = (1 - lev90)/2
beta90 = (1 - lev90)/2
w90 = sd_est*qt((1 + lev90)/2, n-1)/sqrt(n-1)
ci_mean90 = c(mean_est - w90, mean_est + w90)
ci_var90 = c((n)*sd_est^2/qchisq(1-(alfa90/2),n-1), (n)*sd_est^2/qchisq((alfa90),n-1))

alfa95 = (1 - lev95)/2
beta95 = (1 - lev95)/2
w95 = sd_est*qt((1 + lev95)/2, n-1)/sqrt(n-1)
ci_mean95 = c(mean_est - w95, mean_est + w95)
ci_var95 = c((n)*sd_est^2/qchisq(1-(alfa95/2),n-1), (n)*sd_est^2/qchisq((alfa95),n-1))

alfa99 = (1 - lev99)/2
beta99 = (1 - lev99)/2
w99 = sd_est*qt((1 + lev99)/2, n-1)/sqrt(n-1)
ci_mean99 = c(mean_est - w99, mean_est + w99)
ci_var99 = c((n)*sd_est^2/qchisq(1-(alfa99/2),n-1), (n)*sd_est^2/qchisq((alfa99),n-1))
```

Granice 90 % przedziału ufności dla wartosci średniej wynoszą: -0.102, 0.1715. Granice 90 % przedziału ufności dla wariancji wynoszą 5.1442, 6.1451.

Granice 95 % przedziału ufności dla wartosci średniej wynoszą: -0.1282, 0.1978. Granice 95 % przedziału ufności dla wariancji wynoszą 5.0756, 6.2445.

Granice 99 % przedziału ufności dla wartosci średniej wynoszą: -0.1796, 0.2492. Granice 99 % przedziału ufności dla wariancji wynoszą 4.9412, 6.4451.

Zadanie 3

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 nastepujący sposób:

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

- 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:

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

- Rzuć pinezką jeszcze 20 razy i zanotuj wyniki.

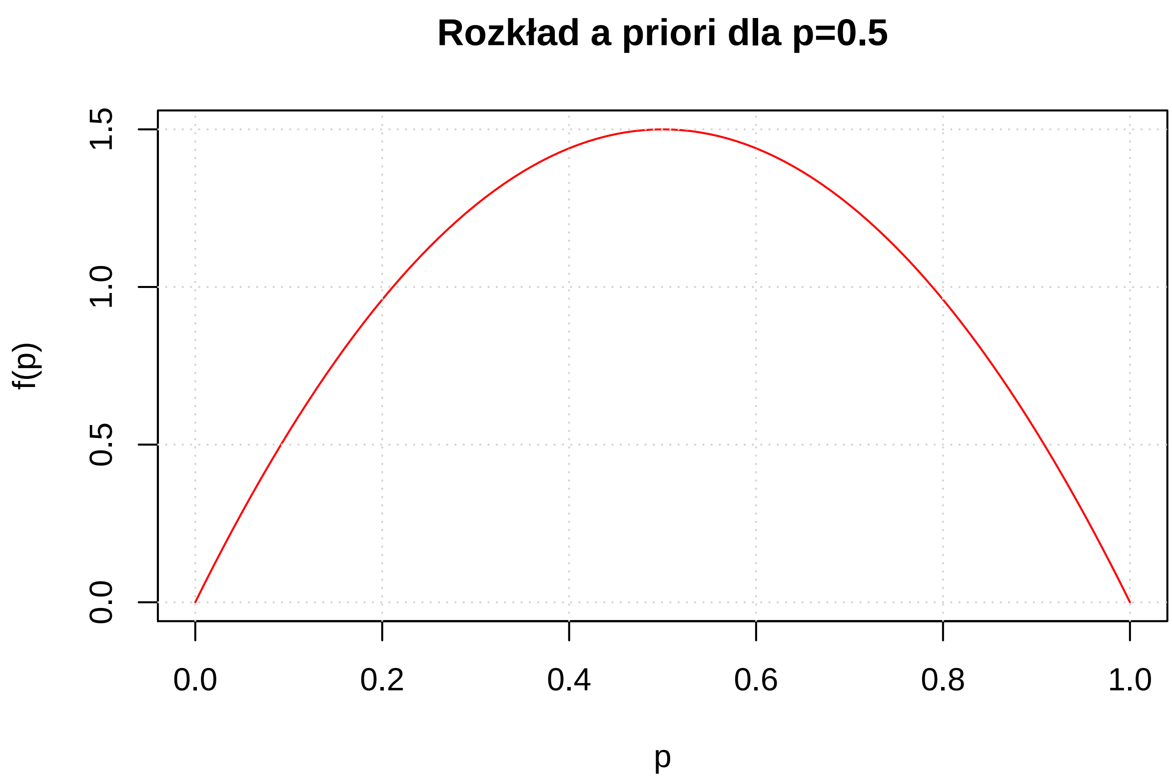
- Wyznacz i narysuj rozkład a posteriori oparty na wszystkich 40 rzutach oraz oblicz wartość bayesowskiego estymatora $\hat{p}$ w tym przypadku.
- Porównaj wyniki z wynikami uzyskanymi po pierwszych 20 rzutach.
- Korzystając ze wzoru na wariancję rozkładu Beta wyznacz i porównaj wariancje rozkładu a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

Rozwiązanie

Przypuszczalna wartość oczekiwana parametru p wynosi 0.5. Zaproponowane wartości α i β to odpowiednio $\alpha = 2$ i $\beta = 2$.

Rozkład a priori parametru p

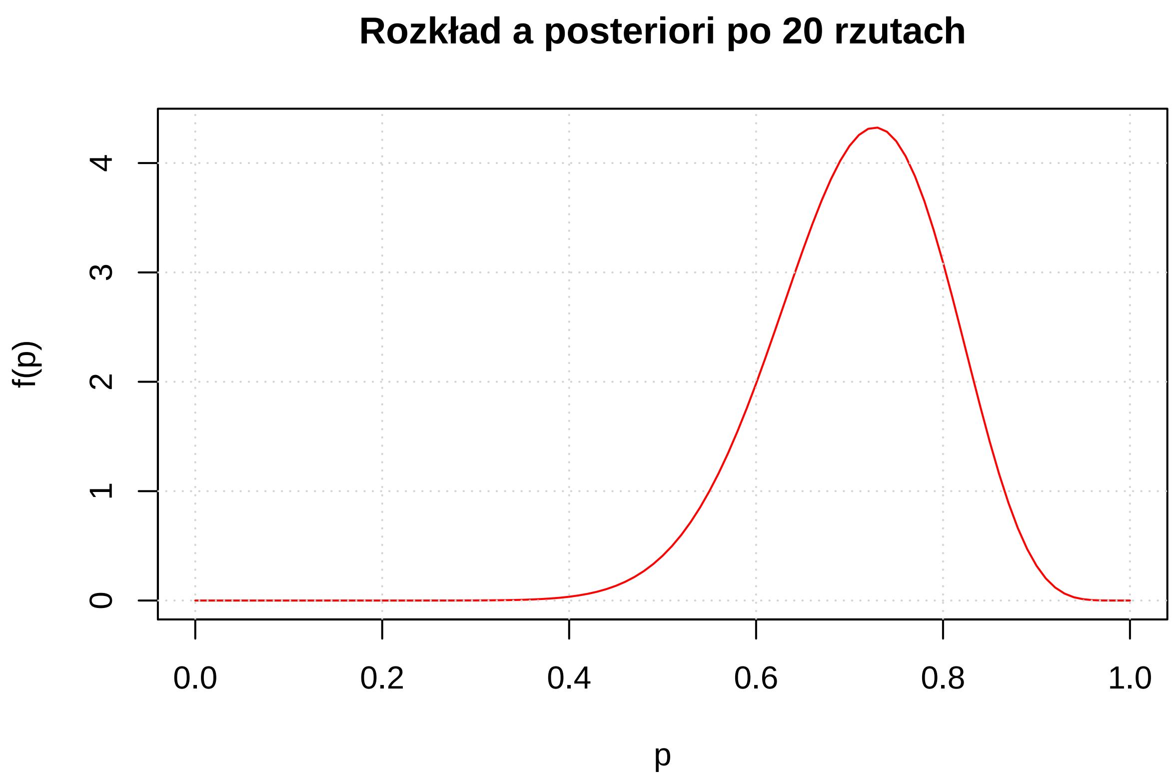
```
p = 0.5
alpha = 2
beta = 2
curve(dbeta(x, shape1 = alpha, shape2 = beta, ncp = 0, log = FALSE),
      xlab='p', ylab = 'f(p)', col = 'red', main = "Rozkład a priori dla p=0.5", 0, 1)
grid()
```



Po dwudziestu rzutach otrzymano:

```
rzut1 = c(0,1,1,0,1,1,1,1,1,1,0,0,1,1,1,1,0,1,1,1)
n1 = length(rzut1)

alpha_post1 = alpha + sum(rzut1)
```

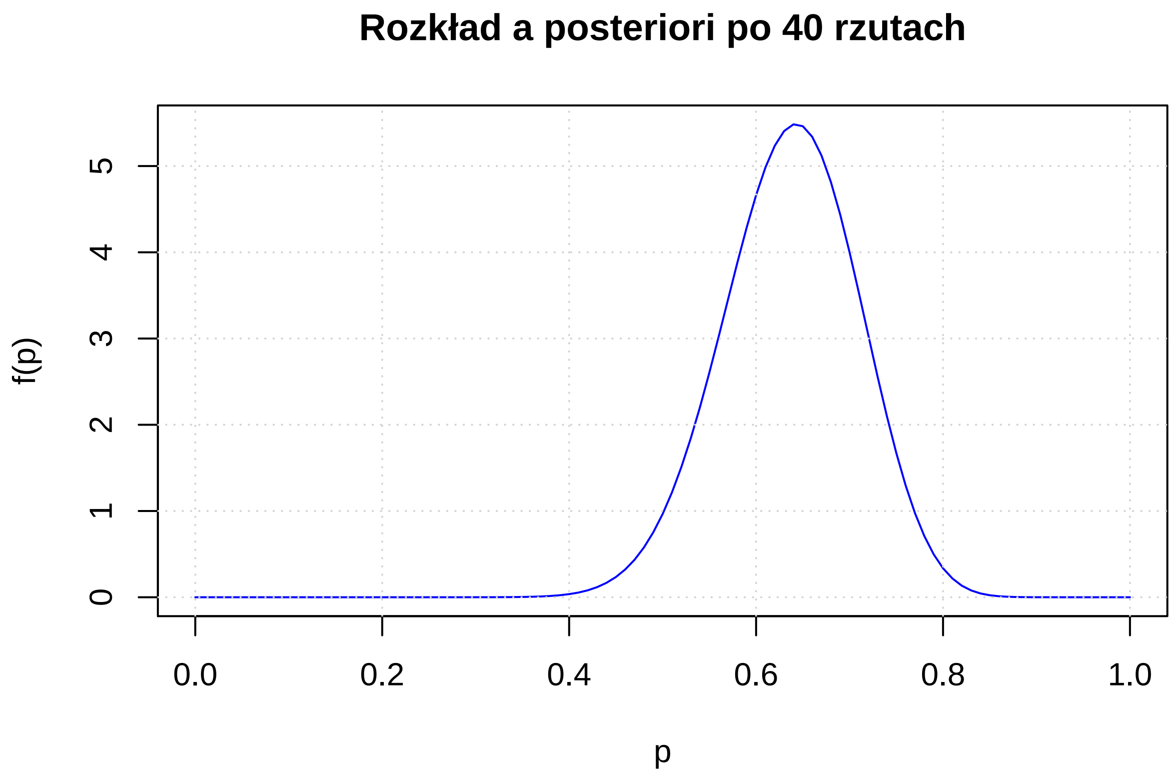
```r
beta_post1 = beta + n1 - sum(rzut1)
pp = alpha_post1/(alpha_post1 + beta_post1)

curve(dbeta(x, shape1 = alpha_post1, shape2 = beta_post1, ncp = 0, log = FALSE), xlab='p',
      ylab = 'f(p)', col = 'red', main = "Rozkład a posteriori po 20 rzutach", 0, 1)
grid()
```



Bayesowski estymator p po 20 rzutach wynosi 0.7083.

Po kolejnych dwudziestu rzutach otrzymano:

```r
rzut2 = c(0,0,1,1,0,1,1,1,0,0,0,1,1,0,1,1,0,1,0,1)
alpha_post2 = alpha_post1 + sum(rzut2)
beta_post2 = beta_post1 + n1 - sum(rzut2)
curve(dbeta(x, shape1 = alpha_post2, shape2 = beta_post2, ncp = 0, log = FALSE),
      xlab='p', ylab = 'f(p)',
      col = 'blue', main = "Rozkład a posteriori po 40 rzutach", 0, 1)
grid()
```



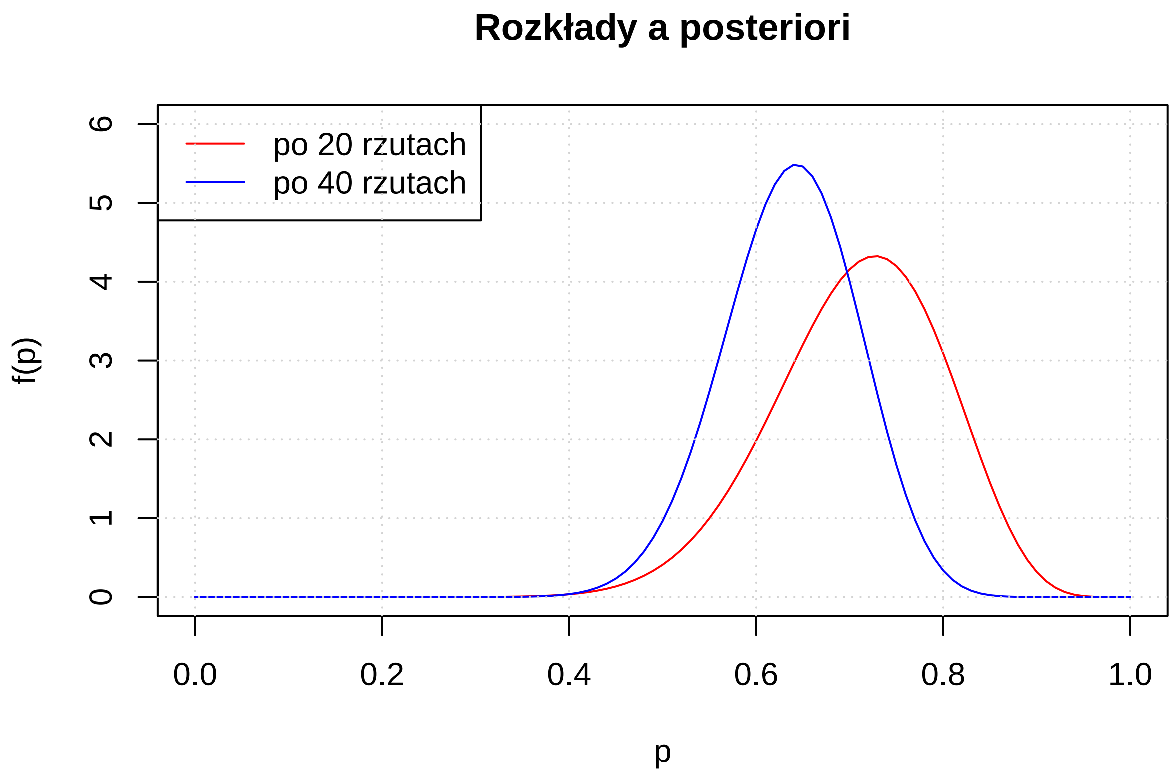
Wyznaczenie bayesowskiego estymatora p:

```
ppp = alpha_post2/(alpha_post2 + beta_post2)
por = ppp/pp
```

Estymator bayesowski p po 40 rzutach wynosi 0.6364. Stosunek estymatora bayesowskiego p po 40 rzutach do estymatora po 20 rzutach wynosi 0.8984.

Porównanie rozkładu po 20 i po 40 rzutach:

```
curve(dbeta(x, shape1 = alpha_post1, shape2 = beta_post1, ncp = 0, log = FALSE),xlab='p',
      ylab = 'f(p)',main='Rozkłady a posteriori', col = 'red',ylim = c(0, 6), 0, 1)
curve(dbeta(x, shape1 = alpha_post2, shape2 = beta_post2, ncp = 0, log = FALSE),
      add=T,col = 'blue', 0, 1)
legend('topleft', c('po 20 rzutach', 'po 40 rzutach'),
    col = c('red', 'blue'), lwd = 1)
grid()
```



Porównanie wariancji rozkładu a priori, a posteriori po 20 rzutach i a posteriori po 40 rzutach:

```
var_priori = (alpha * beta)/(((alpha + beta)^2)*(alpha + beta + 1))
var_post1 = (alpha_post1 * beta_post1)/(((alpha_post1 + beta_post1)^2)*(alpha_post1 + beta_post1 + 1))
var_post2 = (alpha_post2 * beta_post2)/(((alpha_post2 + beta_post2)^2)*(alpha_post2 + beta_post2 + 1))

por2 = var_priori/var_post2
```

Wariancja rozkładu a priori wynosi 0.05, rozkładu a posteriori po 20 rzutach wynosi 0.00826, rozkładu a posteriori po 40 rzutach wynosi 0.00514. Oznacza to, że wariancja zmalała o 9.72 porównując wariancję rozkładu a priori z wariancją rozkładu a posteriori po 40 rzutach.

Zadanie 4

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 estymatory parametrów rozkładu gamma odpowiadające zarejestrowanym danym.
- 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 odchylenia standardowe estymatorów parametrów rozkładu gamma oraz przedziały ufności na poziomie ufności 95%.

Rozwiązanie

Wczytanie danych i wyznaczenie estymatorów metodą momentów

```
g = scan('fotony.txt')
m1 = mean(g)
m2 = mean(g^2)
alpha_mom = m1^2/(m2 - m1^2)
beta_mom = (m2 - m1^2)/m1
```

Estymator parametru α wyznaczony metodą momentów wynosi 1.0655. Estymator parametru β wyznaczony metodą momentów wynosi 73.6241.

Pierwszy sposób wyznaczenia estymatorów metodą największej wiarygodności

```
fun = function(x) {digamma(x) - log(x) - mean(log(g)) + log(mean(g))}
alpha_nw_1 = uniroot(fun, lower = 0.5, upper = 4)$root
beta_nw_1 = mean(g)/alpha_nw_1
```

Estymator parametru α wyznaczony metodą największej wiarygodności (pierwszy sposób) wynosi 1.052. Estymator parametru β wyznaczony metodą największej wiarygodności (pierwszy sposób) wynosi 74.5733.

Drugi sposób wyznaczenia estymatorów metodą największej wiarygodności

```
require(MASS)
```

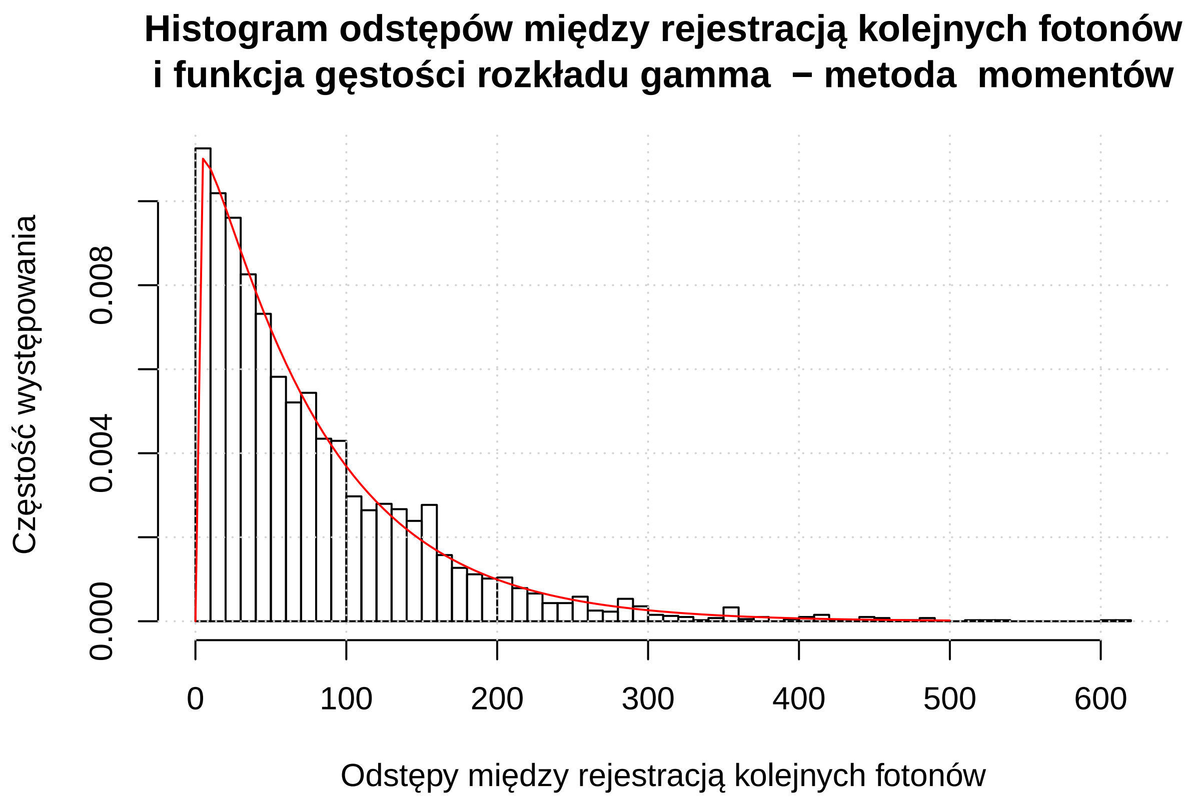
```
## Loading required package: MASS
```

```
est_nw = fitdistr(g, 'gamma', list(shape=1, scale=1), lower=0)
alpha_nw_2 = as.numeric(est_nw$estimate[1])
beta_nw_2 = as.numeric(est_nw$estimate[2])
```

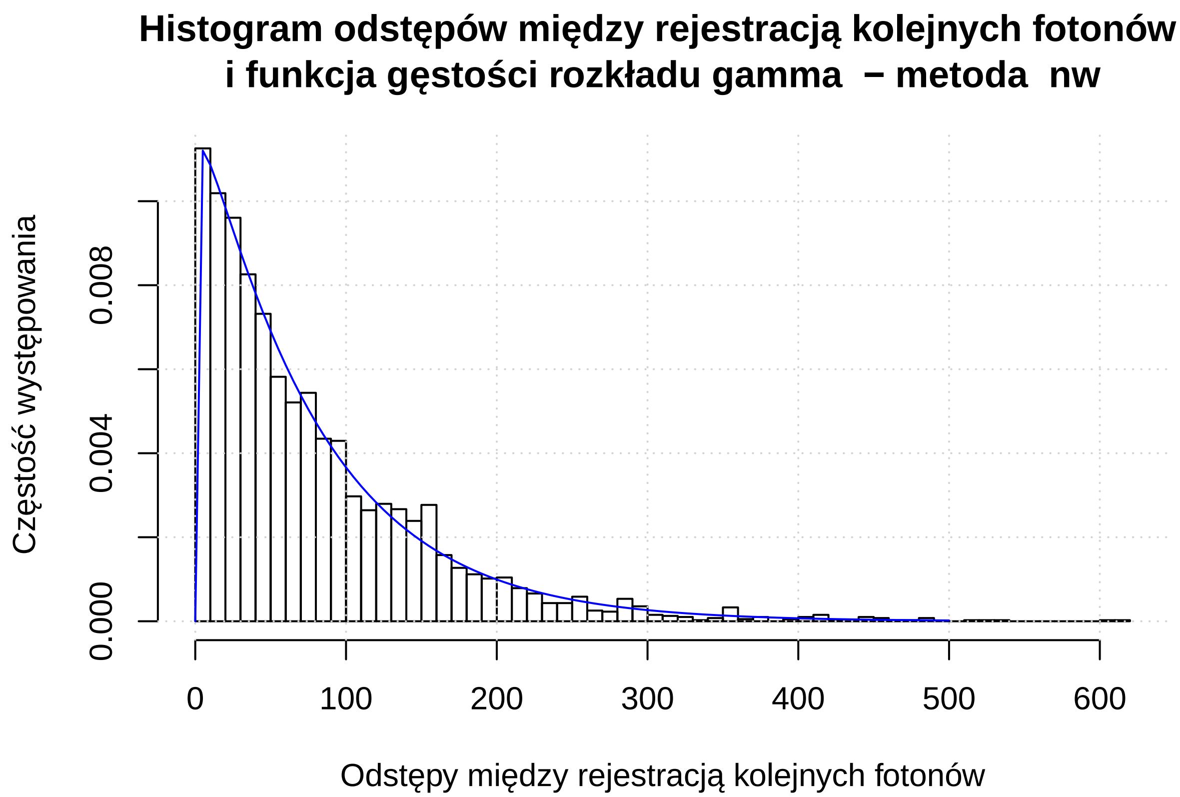
Estymator parametru α wyznaczony metodą największej wiarygodności (drugi sposób) wynosi 1.052. Estymator parametru β wyznaczony metodą największej wiarygodności (drugi sposób) wynosi 74.5737.

Histogram odstępów oraz funkcja gęstości rozkładu gamma

```
hist(g, breaks = 50, prob = T,
xlab = 'Odstępy między rejestracją kolejnych fotonów',
ylab = 'Częstość występowania',
main = 'Histogram odstępów między rejestracją kolejnych fotonów
i funkcja gęstości rozkładu gamma  - metoda  momentów' )
grid()
curve(dgamma(x, shape=alpha_mom, scale=beta_mom),add = T, col = 'red', 0, 500)
```




```
hist(g, breaks = 50, prob = T,
xlab = 'Odstępy między rejestracją kolejnych fotonów',
ylab = 'Częstość występowania',
main = 'Histogram odstępów między rejestracją kolejnych fotonów
i funkcja gęstości rozkładu gamma  - metoda  nw' )
grid()
curve(dgamma(x, shape=alpha_nw_1, scale=beta_nw_1),add = T, col = 'blue', 0, 500)
```



Wyznaczenie odchylenia standardowego estymatorów parametrów (wg metody momentów)

```
K = 1000
boot_res_g_mom = replicate(K, {
  boot = rgamma(length(g), shape=alpha_mom, scale=beta_mom)
  c((mean(boot)^2)/var(boot), var(boot)/mean(boot))
} )
sd_alpha_mom = sd(boot_res_g_mom[1,])
sd_beta_mom = sd(boot_res_g_mom[2,])
```

Odchylenie standardowe estymatora parametru α wynosi 0.03211 (wg metody momentów). Odchylenie standardowe estymatora parametru β wynosi 2.49641 (wg metody momentów).

Wyznaczenie odchylenia standardowego estymatorów parametrów (wg metody najwyższej wiarygodności)

```
K = 1000
boot_res_g_nw = replicate(K, {
  boot = rgamma(length(g), shape=alpha_nw_1, scale=beta_nw_1)
  c((mean(boot)^2)/var(boot), var(boot)/mean(boot))
} )
sd_alpha_nw = sd(boot_res_g_nw[1,])
sd_beta_nw = sd(boot_res_g_nw[2,])
```

Odchylenie standardowe estymatora parametru α wynosi 0.03327 (wg metody najwyższej wiarygodności). Odchylenie standardowe estymatora parametru β wynosi 2.63703 (wg metody najwyższej wiarygodności).

Wyznaczenie przedziałów ufności na poziomie ufności 95% (wg metody momentów)

```
int_alpha_mom = quantile(boot_res_g_mom[1,], c((1-lev95)/2,(1+lev95)/2))
int_beta_mom = quantile(boot_res_g_mom[2,], c((1-lev95)/2,(1+lev95)/2))
```

Przedział ufności na poziomie ufności 95% parametru α wynosi 1.0036, 1.1298 (wg metody momentów). Przedział ufności na poziomie ufności 95% parametru β wynosi 69.1142, 78.8585 (wg metody momentów).

Wyznaczenie przedziałów ufności na poziomie ufności 95% (wg metody najwyższej wiarygodności)

```
int_alpha_nw = quantile(boot_res_g_nw[1,], c((1-lev95)/2,(1+lev95)/2))
int_beta_nw = quantile(boot_res_g_nw[2,], c((1-lev95)/2,(1+lev95)/2))
```

Przedział ufności na poziomie ufności 95% parametru α wynosi 0.987, 1.1204 (wg metody najwyższej wiarygodności). Przedział ufności na poziomie ufności 95% parametru β wynosi 69.4956, 79.7885 (wg metody najwyższej wiarygodności).