ADPS 2021Z — Laboratorium 2 (rozwiązania)

Ola Jaglińska

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, 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

- Wczytaj dane za pomocą komendy scan('skrety.txt')

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

- Dopasuj do danych rozkład Poissona, tj. wyestymuj parametr λ rozkładu Poissona.

Estymata parametru lambda wynosi 'r x_lambda'

- Metodą bootstrapu nieparametrycznego oszacuj odchylenie standardowe estymatora parametru λ.

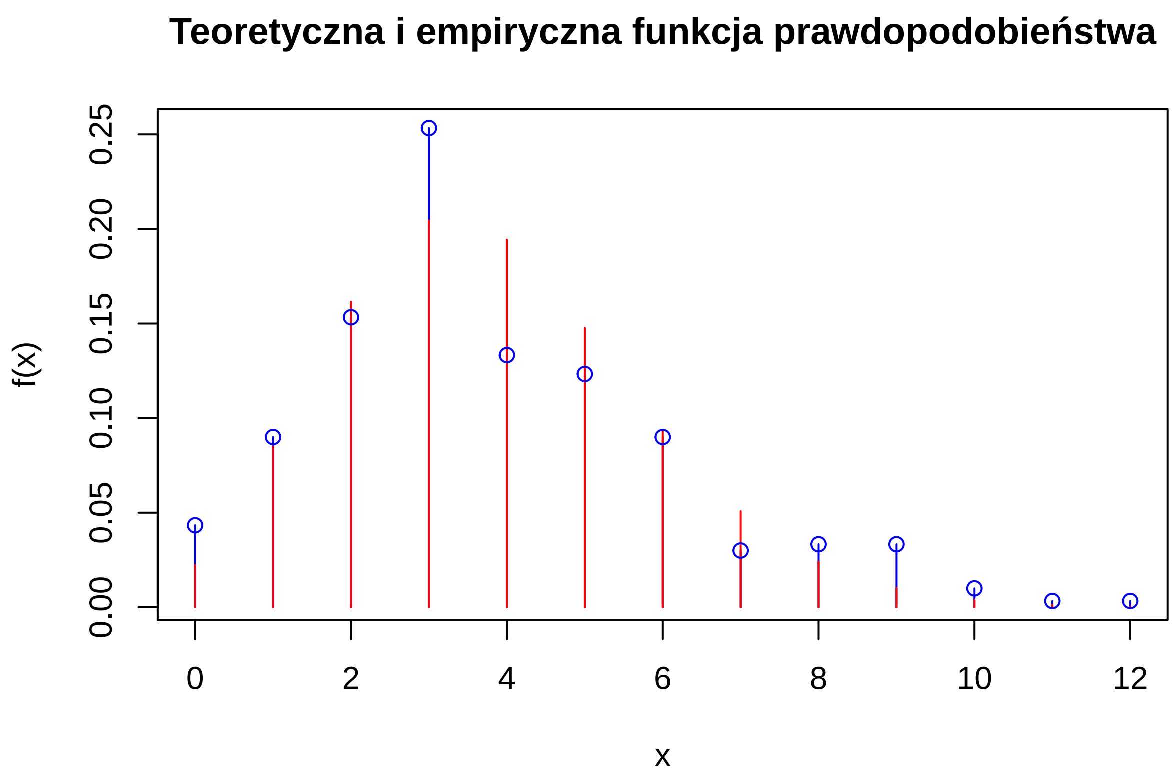
```
K = 1000
boot_res = replicate(K, {
  boot_dane = sample(x, n, replace = T)
  mean(boot_dane)
})
sd_mean = sd(boot_res)
sd_lambda = sd_mean
```

Odchylenie standardowe estymatora wartości lambda wynosi 0.1205639.

- 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.

```
Arg = min(skrety):max(skrety)
n = 300
Freq = as.numeric(table(factor(skrety, levels = Arg))) / n
```

```
plot(Freq ~ Arg, type = 'h', col = 'blue', xlab = 'x', ylab = 'f(x)', main = paste0('Teoretyczna i empi
lines(dpois(Arg, x_lambda)~ Arg, type = 'h', col = 'red')
points(Freq ~ Arg, col = 'blue')
```



Zadanie 2

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.

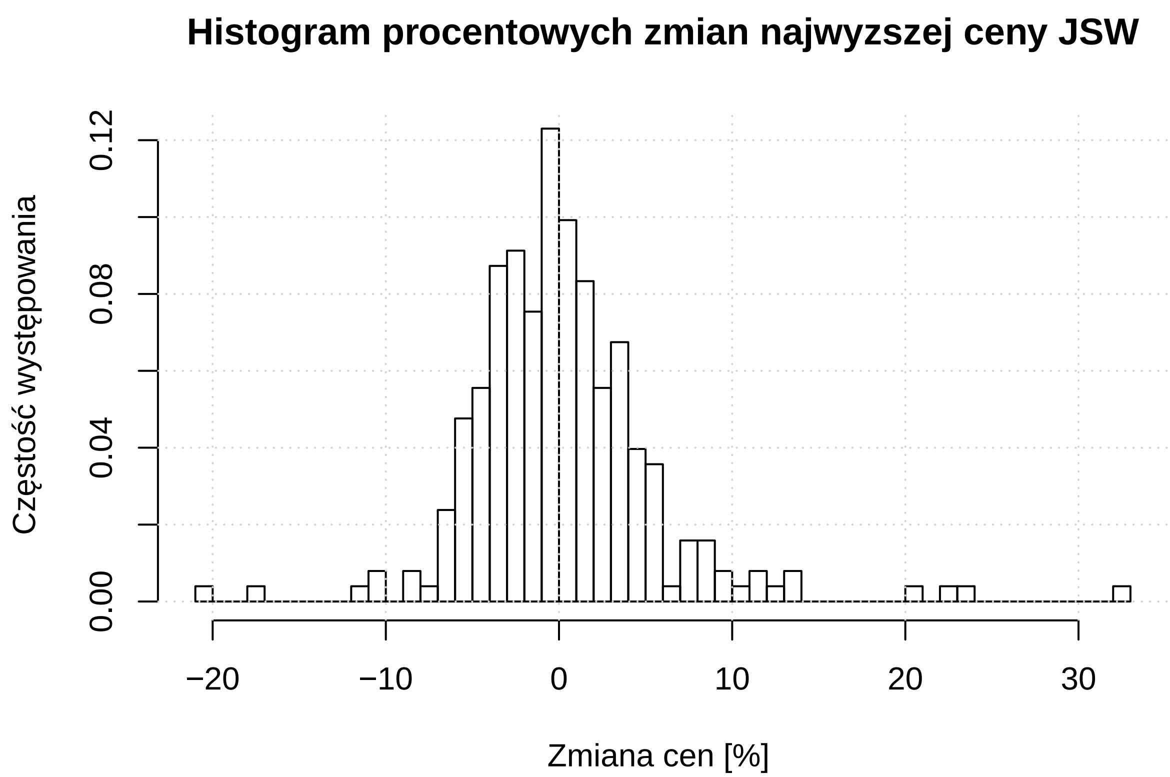
Rozwiązanie

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

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

```
unzip('mstall.zip', 'JSW.mst')
df_JSW = read.csv('JSW.mst')
```

```
names(df_JSW) = c('ticker', 'date', 'open', 'high', 'low', 'close','vol')
df_JSW$date = as.Date.character(df_JSW$date, format ='%Y%m%d')
df_JSW$high_ch = with(df_JSW, c(NA, 100*diff(high)/high[-length(high)]))
df_JSW = df_JSW[which(df_JSW$date >= '2020-01-01' & df_JSW$date <= '2020-12-31'),]
hist(df_JSW$high_ch, breaks = 50, prob = T,
xlab = 'Zmiana cen [%] ',
ylab = 'Częstość występowania',
main = paste('Histogram procentowych zmian najwyzszej ceny JSW')  )
grid()
```



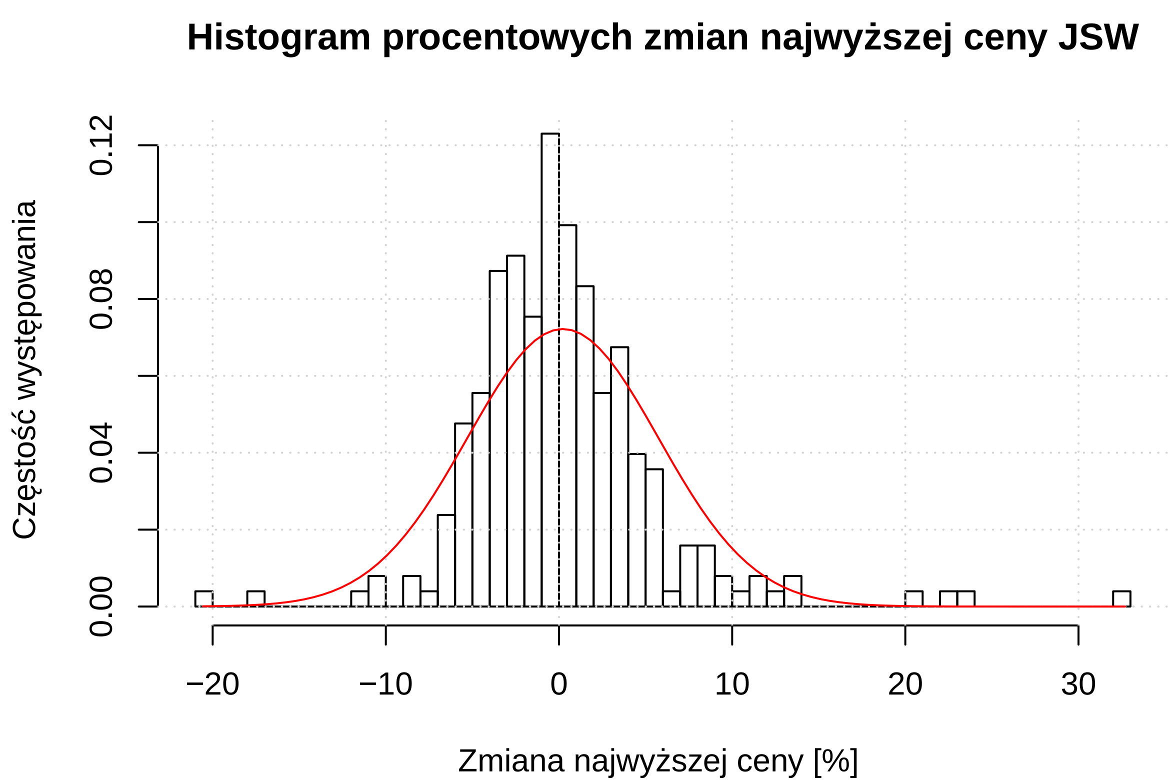
* Wyestymuj wartość średnią oraz wariancję procentowych zmian najwyższych cen w dniu dla wybranej spółki.

```
m_jsw = mean(df_JSW$high_ch, na.rm = T)
v_jsw = var(df_JSW$high_ch, na.rm = T)
s_jsw = sd(df_JSW$high_ch, na.rm = T)
```

Wartość średnia zmian najwyższych cen JSW wynosi 0.2259, a wariancja 30.5305.

- 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.

```
hist(df_JSW$high_ch, breaks = 50, prob = T,
xlab = 'Zmiana najwyższej ceny [%] ',
ylab = 'Częstość występowania',
main = paste('Histogram procentowych zmian najwyższej ceny JSW')  )
grid()
min_c = min(df_JSW$high_ch, na.rm = T)
max_c = max(df_JSW$high_ch, na.rm = T)
curve(dnorm(x, mean = m_jsw, sd = s_jsw), add = T, col = 'red', from = min_c, to = max_c)
```



Na podstawie wykresu wnioskuję, 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.

```
n = length(df_JSW$high_ch)
lev_0.90 = 0.90
w = s_jsw*qt((1+lev_0.90)/2, n-1)/sqrt(n)
ci_m_jsw_0.90 = c(m_jsw - w, m_jsw + w)
a = (1 - lev_0.90)/2; b = (1 - lev_0.90)/2
ci_v_jsw_0.90 = c((n-1)*s_jsw^2/qchisq(1-b,n-1), (n-1)*s_jsw^2/qchisq(a,n-1))
```

Granice 90 % przedziału ufności dla wartosci średniej wynoszą: -0.3487271, 0.8005642.

Granice 90 % przedziału ufności dla wariancji wynoszą 26.5202681, 35.5899881.

```
n = length(df_JSW$high_ch)
lev_0.95 = 0.95
w = s_jsw*qt((1+lev_0.95)/2, n-1)/sqrt(n)
ci_m_jsw_0.95 = c(m_jsw - w, m_jsw + w)
a = (1 - lev_0.95)/2; b = (1 - lev_0.95)/2
ci_v_jsw_0.95 = c((n-1)*s_jsw^2/qchisq(1-b,n-1), (n-1)*s_jsw^2/qchisq(a,n-1))
```

Granice 95 % przedziału ufności dla wartosci średniej wynoszą: -0.4595921, 0.9114292.

Granice 95 % przedziału ufności dla wariancji wynoszą 25.8213567, 36.664079.

```
n = length(df_JSW$high_ch)
lev_0.99 = 0.99
w = s_jsw*qt((1+lev_0.99)/2, n-1)/sqrt(n)
ci_m_jsw_0.99 = c(m_jsw - w, m_jsw + w)
a = (1 - lev_0.99)/2; b = (1 - lev_0.99)/2
ci_v_jsw_0.99 = c((n-1)*s_jsw^2/qchisq(1-b,n-1), (n-1)*s_jsw^2/qchisq(a,n-1))
```

Granice 99 % przedziału ufności dla wartosci średniej wynoszą: -0.6775182, 1.1293553.

Granice 99 % przedziału ufności dla wariancji wynoszą 24.5251703, 38.8903304.

Z powyższych danych wnioskuję, że gdy poziom ufności rośnie to przedział ufności również rośnie.

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 następujący 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:

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

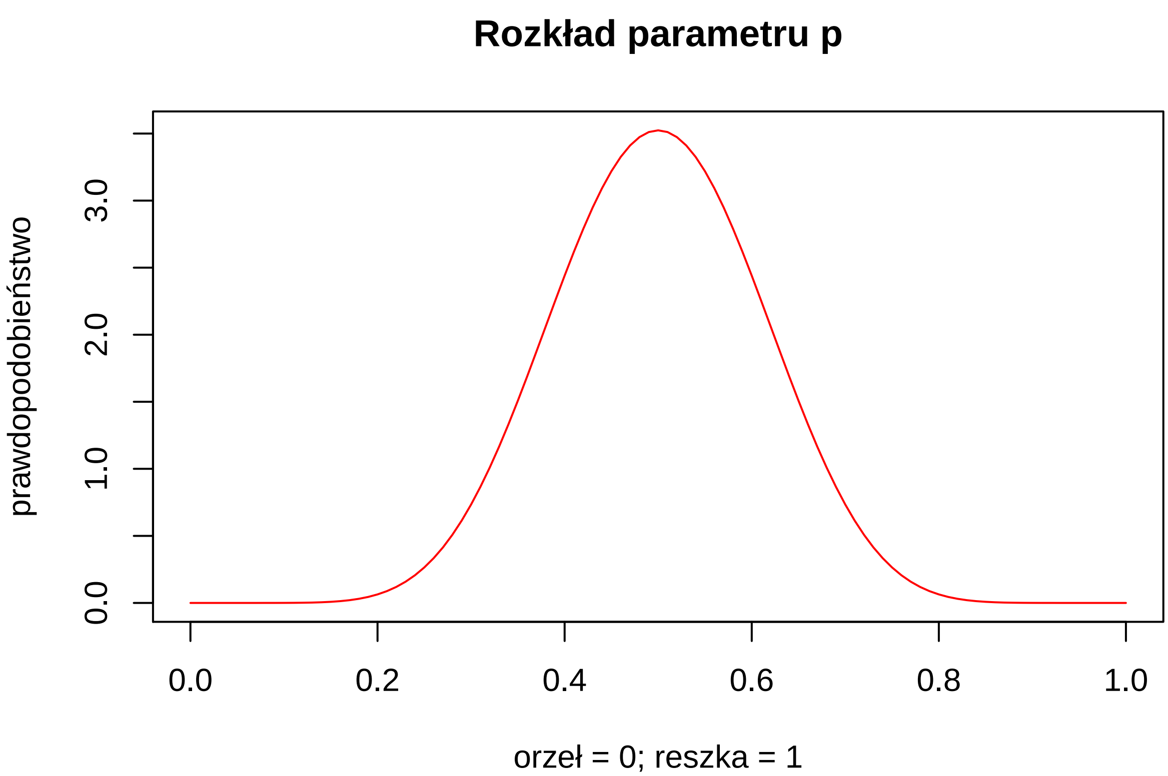
- Rzuć pinezką jeszcze 20 razy i zanotuj wyniki. Wyznacz i narysuj rozkład a posteriori oparty na wszystkich 40 rzutach oraz oblicz wartość bayesowskiego estymatora $\hat{p}$ w tym przypadku. Porównaj wyniki z wynikami uzyskanymi po pierwszych 20 rzutach.
- Korzystając ze wzoru na wariancję rozkładu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

Rozwiązanie

Eksperyment wykonałam przy użyciu monety. Reszka - 1 Orzeł - 0

1. Pierwsze Wyobrażenie -> 50% reszka i 50% orzeł. $\alpha = 10$ $\beta = 10$

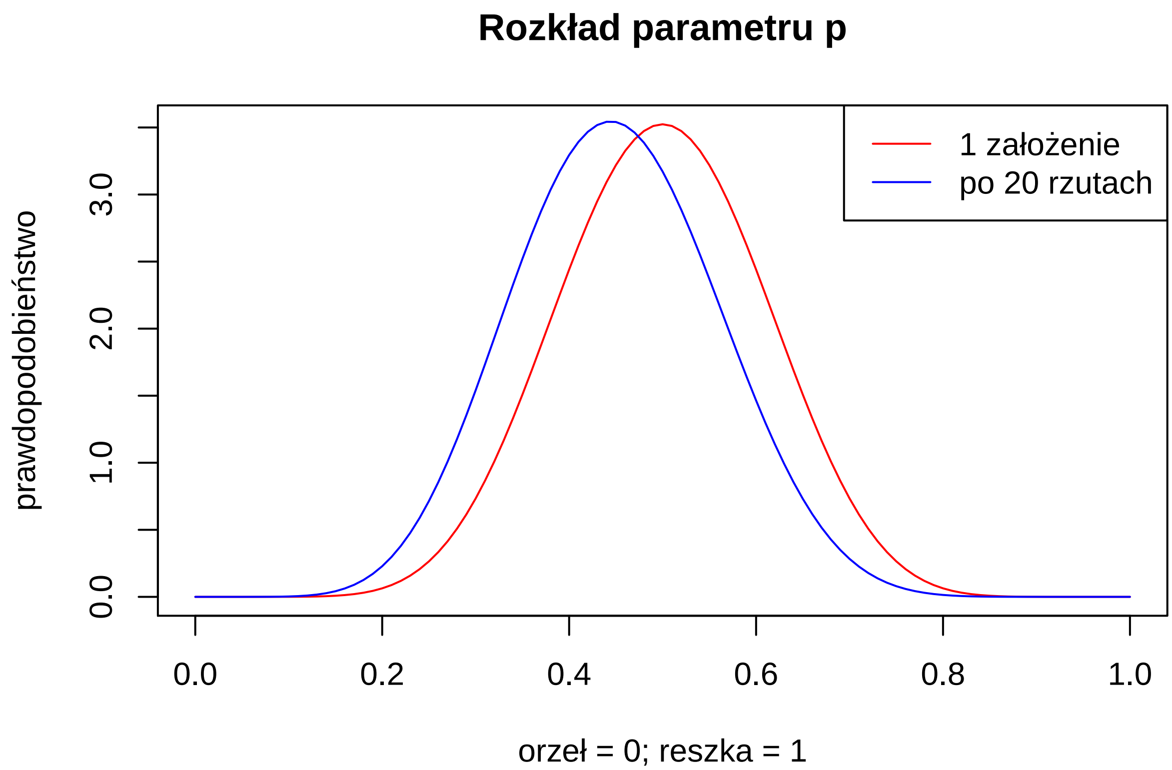
```
curve(dbeta(x, shape1 = 10, shape2 = 10), 0, 1, col='red', ylab = 'prawdopodobieństwo', xlab = 'orzeł =
      main = 'Rozkład parametru p')
```



2. Pierwszy raz rzucam 20 razy -> wyniki: orzeł (0) - 11 reszka(1) - 9

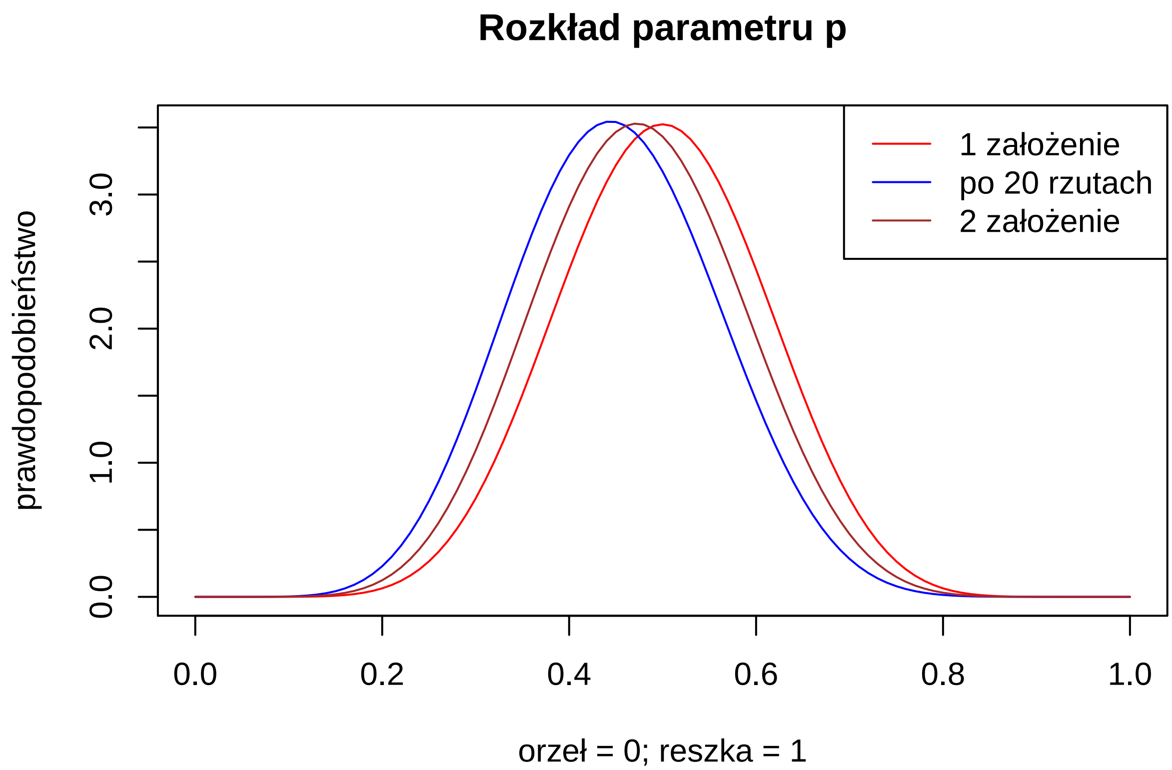
```
rzuty = data.frame(nr_rzutu = 1:20, proba1 = c(0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0,
```

```
curve(dbeta(x, shape1 = 10, shape2 = 10), 0, 1, col='red', ylab = 'prawdopodobieństwo', xlab = 'orzeł =
      main = 'Rozkład parametru p')
curve(dbeta(x, shape1 = 9, shape2 = 11), 0 , 1, col='blue', add = T)
legend('topright', c('1 założenie', 'po 20 rzutach'),
    col = c('red', 'blue'), lwd = 1)
```



3. Zmieniam wyobrażenie na: 52.5 % orzeł 47.5 % reszka $\alpha = 10.5$ $\beta = 9.5$

```
curve(dbeta(x, shape1 = 10, shape2 = 10), 0, 1, col='red', ylab = 'prawdopodobieństwo', xlab = 'orzeł =
      main = 'Rozkład parametru p')
curve(dbeta(x, shape1 = 9, shape2 = 11), 0 , 1, col='blue', add = T)
curve(dbeta(x, shape1 = 9.5, shape2 = 10.5), 0 , 1, col='brown', add = T)
legend('topright', c('1 założenie', 'po 20 rzutach', '2 założenie'),
    col = c('red', 'blue', 'brown'), lwd = 1)
```

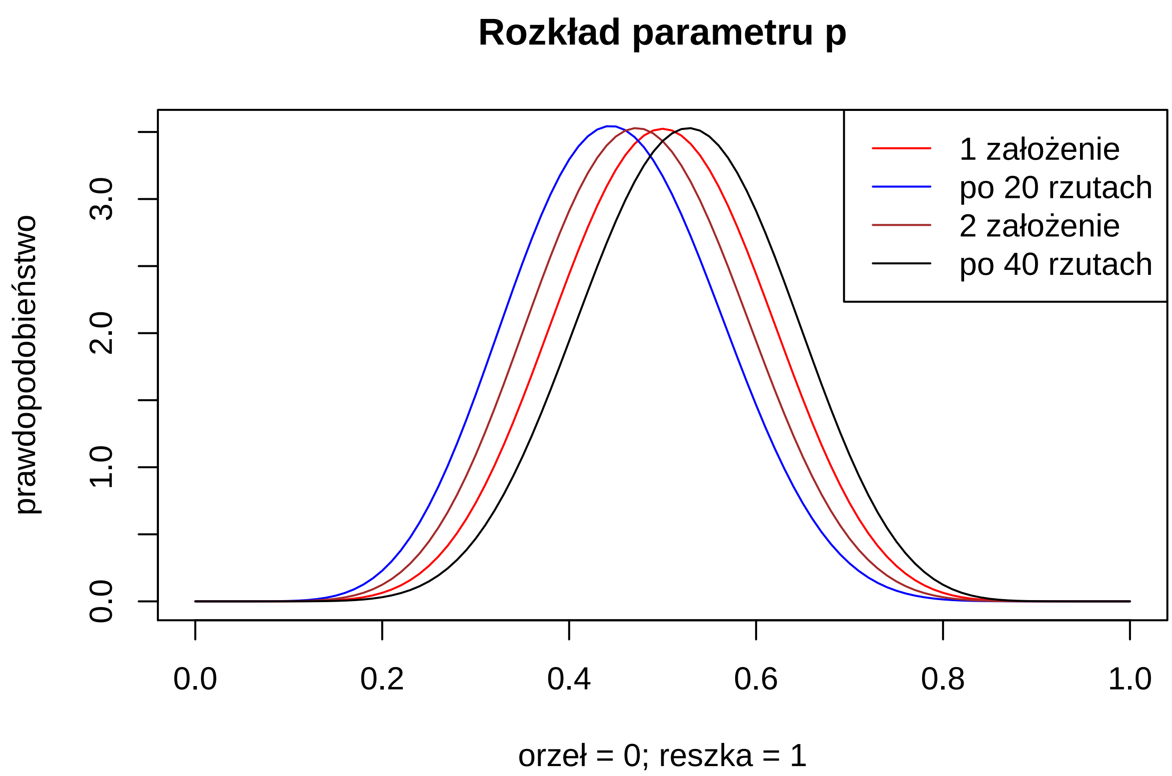


4. Drugi raz rzucam 20 razy -> wyniki: orzeł (0) - 8 reszka(1) - 12

```
rzuty = data.frame(nr_rzutu = 1:20, proba1 = c(0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0,
```

Wyniki po wszystkich 40 rzutach: orzeł (0) - 19 reszka(1) - 21 Po 20 rzutach funkcja prawdopoodobieństwa była bliżej orła, po zsumowaniu rzutów prawdopodobieństwo przesunęło się w stronę reszki. Rozbieżność jest niewielka. Wszystkie funkcje są w okolicy wartości 0.5.

```
curve(dbeta(x, shape1 = 10, shape2 = 10), 0, 1, col='red', ylab = 'prawdopodobieństwo', xlab = 'orzeł =
      main = 'Rozkład parametru p')
curve(dbeta(x, shape1 = 9, shape2 = 11), 0 , 1, col='blue', add=T)
curve(dbeta(x, shape1 = 9.5, shape2 = 10.5), 0 , 1, col='brown', add=T)
curve(dbeta(x, shape1 = 21*0.5, shape2 = 19*0.5), 0 , 1, col='black', add=T)
legend('topright', c('1 założenie', 'po 20 rzutach', '2 założenie', 'po 40 rzutach'),
    col = c('red', 'blue', 'brown', 'black'), lwd = 1)
```



***

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 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

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

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

- Metodą momentów oraz metodą największej wiarygodności wyznacz estymaty parametrów rozkładu gamma odpowiadające zarejestrowanym danym. Porównaj wyniki uzyskane dla obu metod.

```
mf1 = mean(fotony)
mf2 = mean(fotony^2)
alpha_mom_fotony = mf1^2/(mf2 - mf1^2)
beta_mom_fotony = (mf2 - mf1^2)/mf1
```

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

```
require(MASS)
```

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

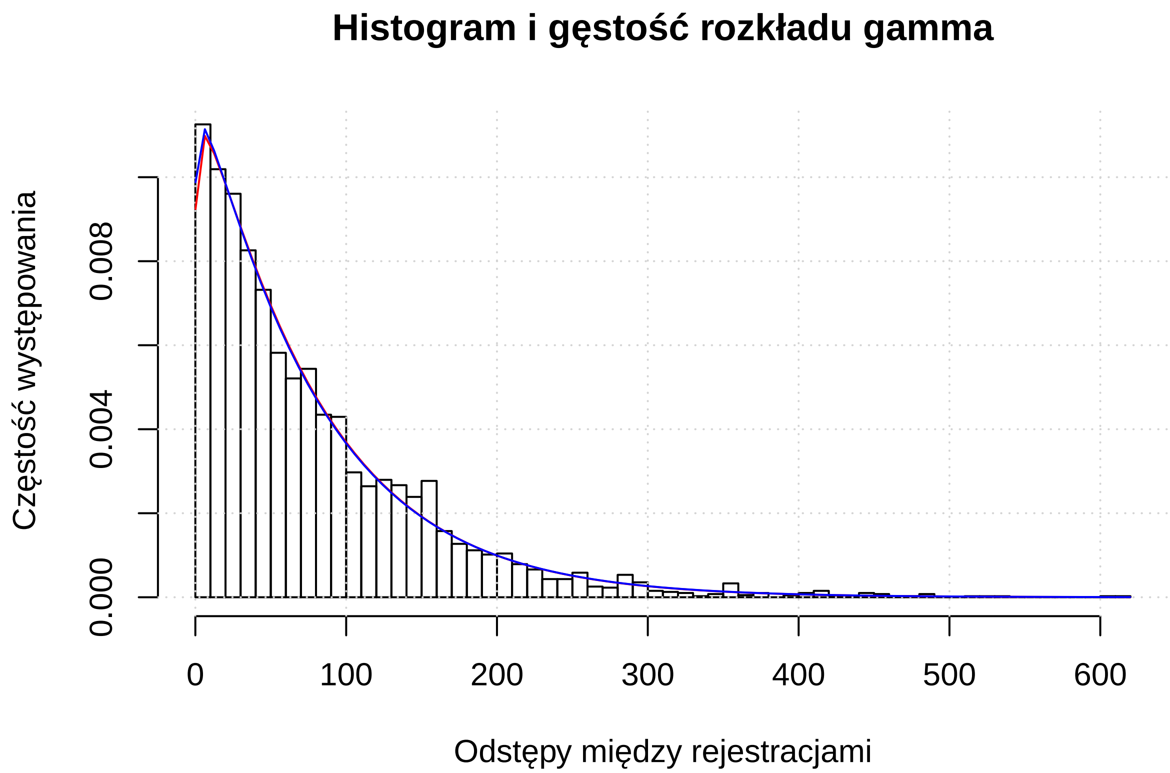
```
est_nw = fitdistr(fotony, 'gamma', list(shape=1, scale=1), lower=0)
alpha_nw_fotony = as.numeric(est_nw$estimate[1])
beta_nw_fotony = as.numeric(est_nw$estimate[2])
```

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

Wartości estymatorów parametrów rozkładu gamma wyznaczone przy użyciu metody momentów są zbliżone do wartości wyznaczonych metodą największej wiarygodności.

- Narysuj na jednym wykresie histogram odstępów oraz funkcje gęstości rozkładu gamma o parametrach wyestymowanych za pomocą obu metod.

```
hist(fotony, breaks = 50, prob = TRUE,
xlab = 'Odstępy między rejestracjami',
ylab = 'Częstość występowania',
main = paste('Histogram i gęstość rozkładu gamma'))
grid()
min_c = min(fotony, na.rm = T)
max_c = max(fotony, na.rm = T)
curve(dgamma(x, shape = alpha_mom_fotony, scale = beta_mom_fotony), add = T, col = 'red', from = min_c,
curve(dgamma(x, shape = alpha_nw_fotony, scale = beta_nw_fotony), add = T, col = 'blue', from = min_c,
```



- 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.

```
K = 1000
n = 3935

boot_res = replicate(K, {
  boot_dane = rgamma(n, shape = alpha_mom_fotony, scale = beta_mom_fotony)
  mb1 = mean(boot_dane)
  mb2 = mean(boot_dane^2)
  alpha_mom_boot_dane = mb1^2/(mb2 - mb1^2)
  beta_mom_boot_dane = (mb2 - mb1^2)/mb1
  c(alpha_mom_boot_dane, beta_mom_boot_dane)
} )
sd_alpha_mom_fotony = sd(boot_res[1,])
sd_beta_mom_fotony = sd(boot_res[2,])
```

Odchylenie standardowe estymatora wartości metody momentów wynosi $\hat{\alpha} = 0.0341373$, a $\hat{\beta} = 2.5566831$.

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