5 Model statystyczny i estymacja punktowa

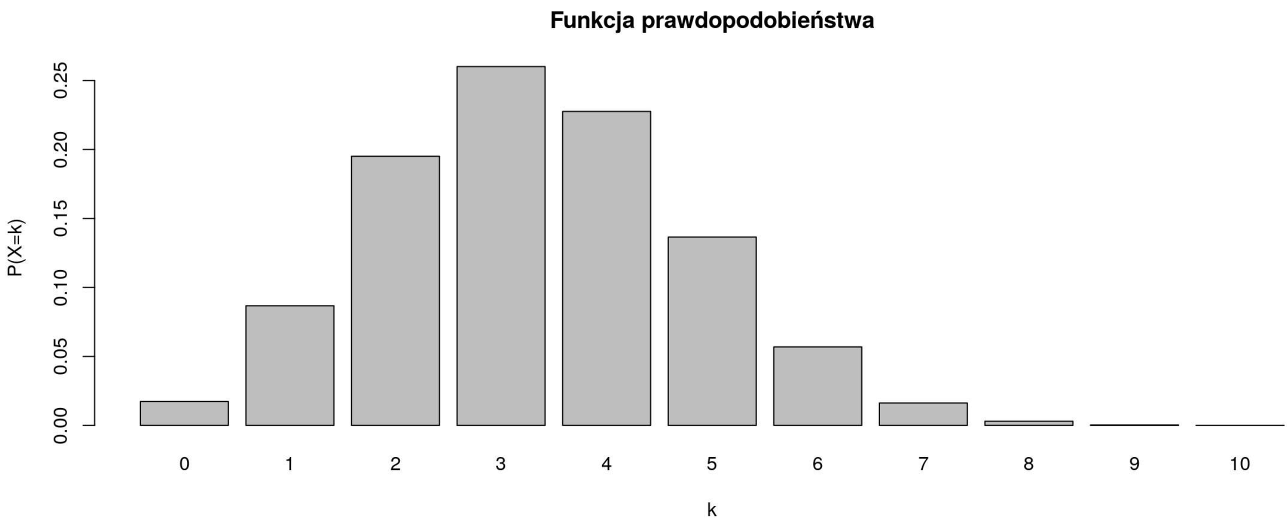
5.1 Wybrane rozkłady prawdopodobieństwa

1. rozkład dwumianowy $b(m,p)$, $m \in \mathbb{N}$, $p \in (0,1)$

$$\mathbb{P}(X=k) = \binom{m}{k} p^k (1-p)^{m-k},\; k = 0,1,\ldots,m$$

- Funkcja prawdopodobieństwa zmiennej $X \sim b(10, 1/3)$

```
barplot(dbinom(x = 0:10, size = 10, prob = 1 / 3), names.arg = 0:10,
        xlab = "k", ylab = "P(X=k)", main = "Funkcja prawdopodobieństwa")
```

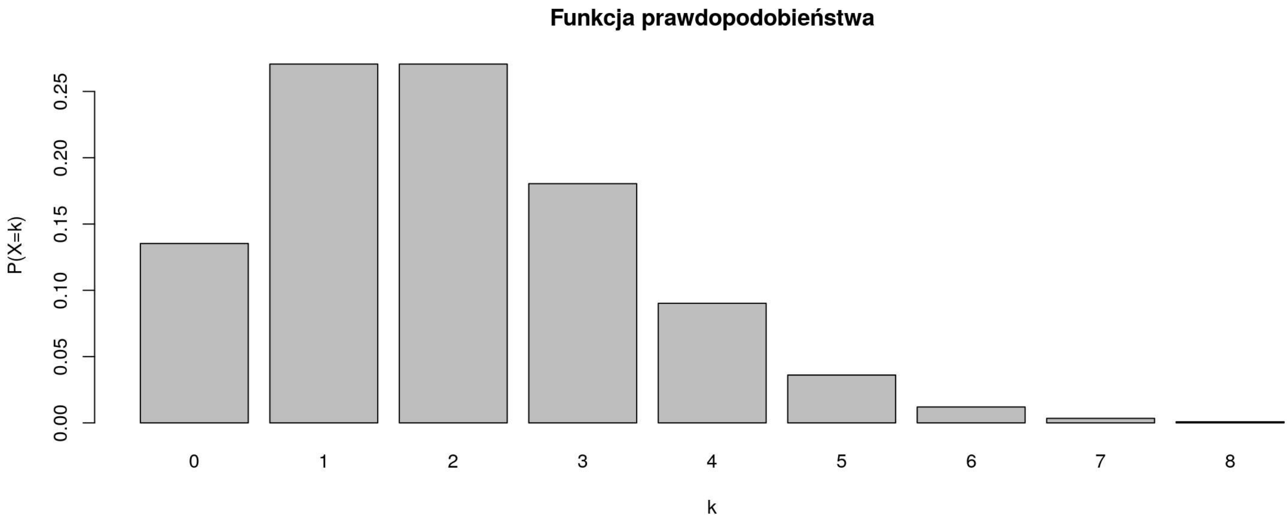


2. rozkład Poissona $\pi(\lambda)$, $\lambda > 0$

$$\mathbb{P}(X=k) = \frac{e^{-\lambda}\lambda^k}{k!},\; k = 0,1,\ldots$$

- Funkcja prawdopodobieństwa zmiennej $X \sim \pi(2)$

```
barplot(dpois(x = 0:8, lambda = 2), names.arg = 0:8,
        xlab = "k", ylab = "P(X=k)", main = "Funkcja prawdopodobieństwa")
```

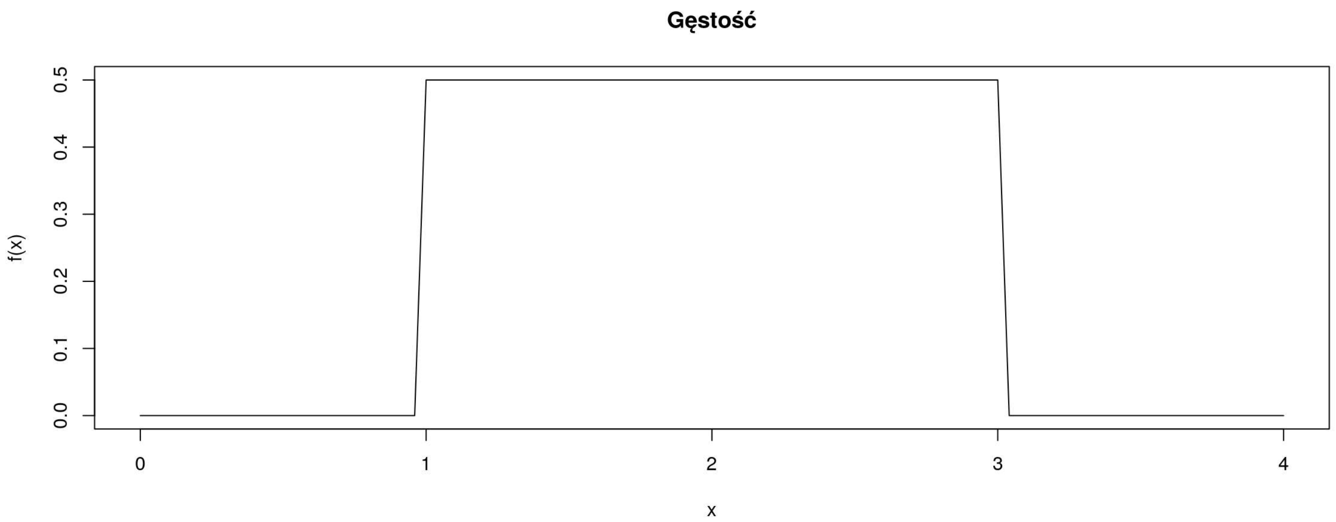


3. rozkład jednostajny $U(a,b)$, $a<b$

$$f_X(x)=\begin{cases}\frac{1}{b-a} & dla\ x\in(a,b)\\ 0 & dla\ x\notin(a,b)\end{cases}$$

- Gęstość zmiennej $X\sim U(1,3)$

```
curve(dunif(x, min = 1, max = 3), 0, 4, ylab = "f(x)", main = "Gęstość")
```

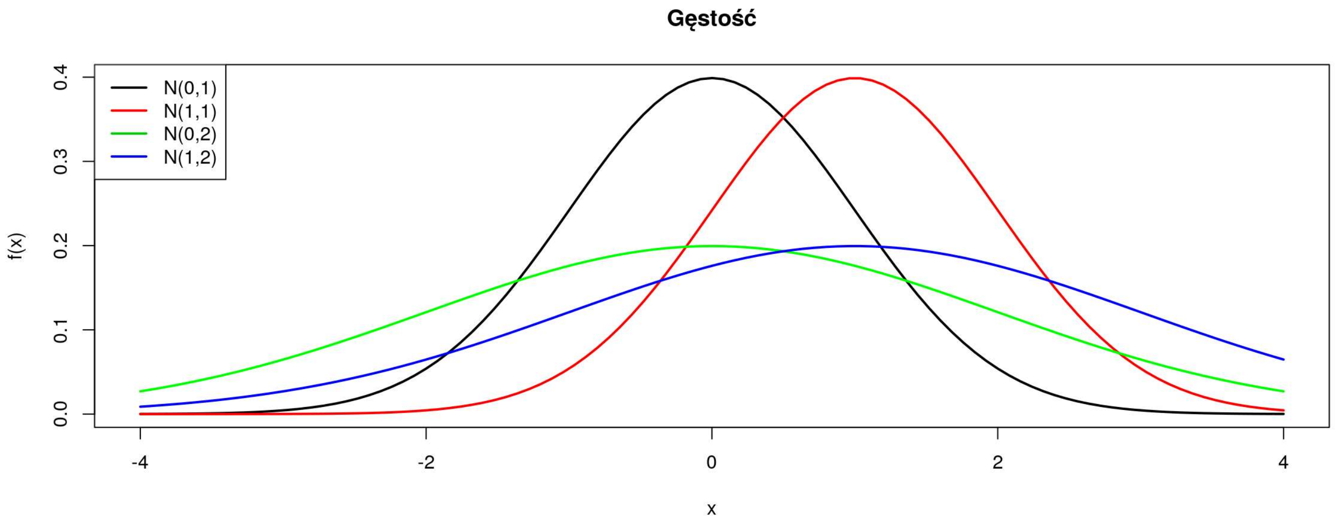


4. rozkład normalny $N(\mu,\sigma)$, $\mu\in\mathbb{R}$, $\sigma>0$

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Gęstości rozkładów normalnych

```r
curve(dnorm, -4, 4, ylab = "f(x)", main = "Gęstość", lwd = 2)
curve(dnorm(x, mean = 1), col = "red", add = TRUE, lwd = 2)
curve(dnorm(x, sd = 2), col = "green", add = TRUE, lwd = 2)
curve(dnorm(x, mean = 1, sd = 2), col = "blue", add = TRUE, lwd = 2)
legend("topleft", lwd = 2, col = 1:4, legend = c("N(0,1)", "N(1,1)", "N(0,2)", "N(1,2)"
```

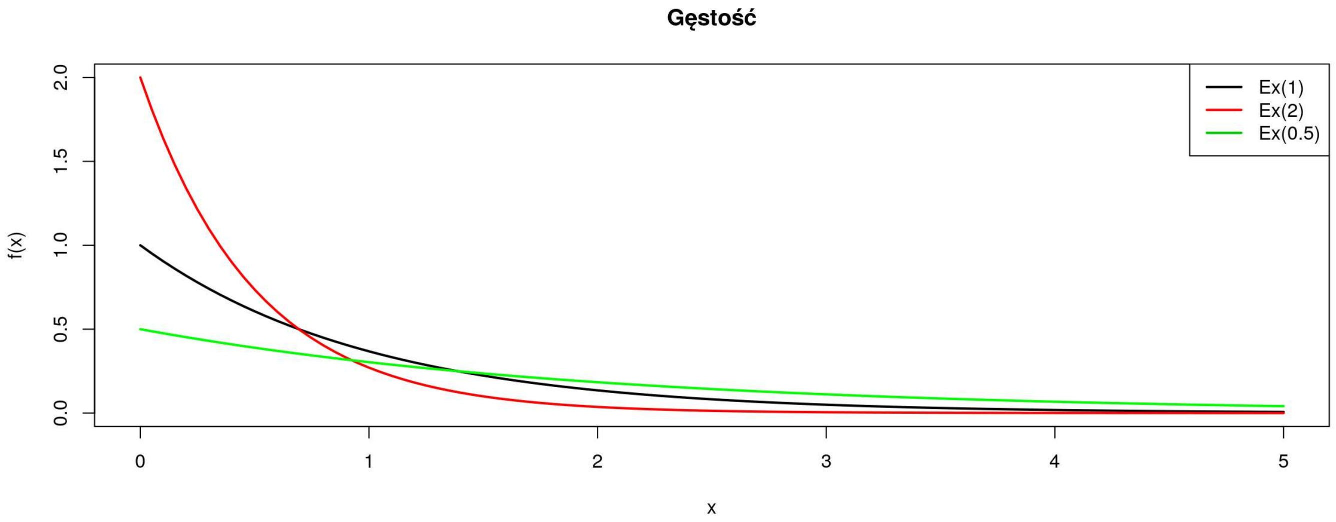


5. rozkład wykładniczy $Ex(\lambda)$, $\lambda > 0$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & dla\ x > 0 \\ 0 & dla\ x \leqslant 0 \end{cases}$$

- Gęstości rozkładów wykładniczych

```r
curve(dexp, 0, 5, ylim = c(0, 2), ylab = "f(x)", main = "Gęstość", lwd = 2)
curve(dexp(x, rate = 2), col = "red", add = TRUE, lwd = 2)
curve(dexp(x, rate = 0.5), col = "green", add = TRUE, lwd = 2)
legend("topright", lwd = 2, col = 1:3, legend = c("Ex(1)", "Ex(2)", "Ex(0.5)"))
```



6. rozkład Rayleigha $R(\lambda)$, $\lambda > 0$

$$f_\lambda(x) = \frac{2}{\lambda} x \exp\left(-\frac{x^2}{\lambda}\right) I_{(0,\infty)}(x)$$

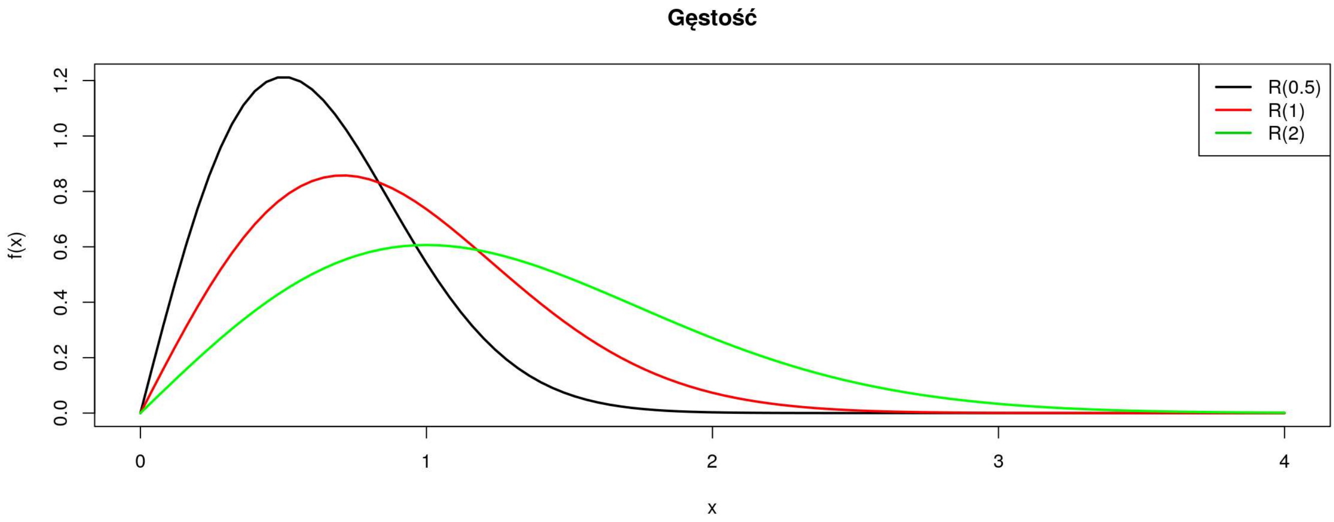
Uwaga. Rozkład Rayleigha jest zaimplementowany w pakiecie `VGAM` z następującą funkcją gęstości

$$f_\sigma(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) I_{(0,\infty)}(x),$$

więc w naszej notacji $\sigma = \sqrt{\frac{\lambda}{2}}$.

- Gęstości rozkładów Rayleigha

```
lambda <- 0.5
curve(VGAM::drayleigh(x, sqrt(lambda / 2)),
      xlim = c(0, 4), ylab = "f(x)", main = "Gęstość", lwd = 2)
lambda <- 1
curve(VGAM::drayleigh(x, sqrt(lambda / 2)),
      col = "red", add = TRUE, lwd = 2)
lambda <- 2
curve(VGAM::drayleigh(x, sqrt(lambda / 2)),
      col = "green", add = TRUE, lwd = 2)
legend("topright", lwd = 2, col = 1:3, legend = c("R(0.5)", "R(1)", "R(2)"))
```



Rozkłady prawdopodobieństwa w programie R

Rozkład	Dystrybuanta	Gęstość/Funkcja prawd.	Kwantyl	Ge
dwumianowy	pbinom	dbinom	qbinom	
Poissona	ppois	dpois	qpois	
ujemny dwumianowy	pnbinom	dnbinom	qnbinom	r
geometryczny	pgeom	dgeom	qgeom	
hipergeometryczny	phyper	dhyper	qhyper	
jednostajny	punif	dunif	qunif	
beta	pbeta	dbeta	qbeta	
wykładniczy	pexp	dexp	qexp	
gamma	pgamma	dgamma	qgamma	r
normalny	pnorm	dnorm	qnorm	
logarytmiczno-normalny	plnorm	dlnorm	qlnorm	
Weibulla	pweibull	dweibull	qweibull	
chi-kwadrat	pchisq	dchisq	qchisq	
t-Studenta	pt	dt	qt	
Cauchy'ego	pcauchy	dcauchy	qcauchy	r
F-Snedecora	pf	df	qf	
Rayleigha	VGAM::prayleigh	VGAM::drayleigh	VGAM::qrayleigh	VGAM::r

5.2 Przykłady

Przykład 1. Poniższe dane podają liczbę błędów w grupie 50 osób zdających egzamin testowy. Egzamin składał się z 18 pytań (można popełnić maksymalnie dwa błędy, aby zdać egzamin).

1	1	2	0	1	3	1	4	4	4	0	1	0	0	0	2	3
4	0	1	5	2	3	5	3	2	2	4	0	2	2	0	2	2
3	3	1	3	2	2	0	0	5	4	2	1	5	2	2	0	

Zmienna X to liczba błędów. Jest to dyskretna zmienna ilościowa.

```
liczba_bledow <- c(1, 1, 2, 0, 1, 3, 1, 4, 4, 4, 0, 1, 0, 0, 0, 2, 3,
                   4, 0, 1, 5, 2, 3, 5, 3, 2, 2, 4, 0, 2, 2, 0, 2, 2,
                   3, 3, 1, 3, 2, 2, 0, 0, 5, 4, 2, 1, 5, 2, 2, 0)
# wykres słupkowy
barplot(prop.table(table(liczba_bledow)),
        xlab = "Liczba błędów", ylab = "Prawdopodobieństwo",
        main = "Rozkład empiryczny liczby błędów")
```

- model: rozkład dwumianowy z $m = 18$
- $\mathcal{P} = \{b(18, p) : p \in (0, 1)\}$
- $\Theta = (0, 1)$ oraz $\theta = p$

```
liczba_bledow <- c(1, 1, 2, 0, 1, 3, 1, 4, 4, 4, 0, 1, 0, 0, 0, 2, 3,
                   4, 0, 1, 5, 2, 3, 5, 3, 2, 2, 4, 0, 2, 2, 0, 2, 2,
                   3, 3, 1, 3, 2, 2, 0, 0, 5, 4, 2, 1, 5, 2, 2, 0)
m <- 18
# estymator
(p_est <- mean(liczba_bledow) / m)
```

```
## [1] 0.1122222
```

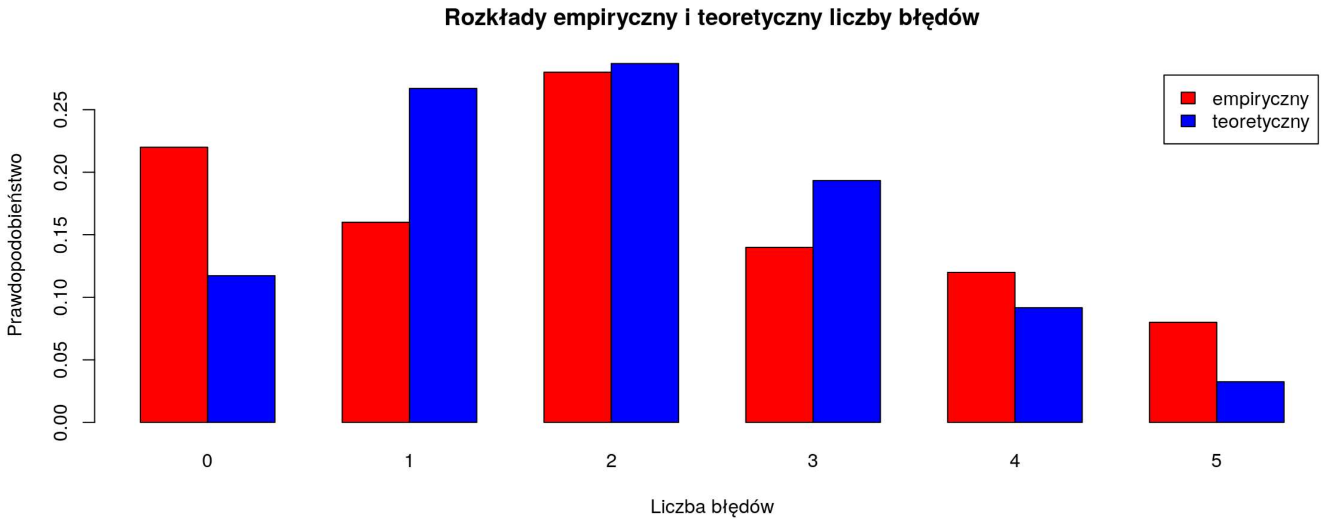
```r
probs <- dbinom(sort(unique(liczba_bledow)), size = m, prob = p_est)
sum(probs)
```

```
## [1] 0.9887985
```

```r
counts <- matrix(c(prop.table(table(liczba_bledow)), probs), nrow = 2, byrow = TRUE)
rownames(counts) <- c("empiryczny", "teoretyczny")
colnames(counts) <- sort(unique(liczba_bledow))
counts
```

```
##                     0         1         2         3          4          5
## empiryczny  0.2200000 0.1600000 0.2800000 0.1400000 0.12000000 0.08000000
## teoretyczny 0.1173483 0.2670078 0.2868914 0.1934153 0.09168466 0.03245109
```

```r
barplot(counts,
        xlab = "Liczba błędów", ylab = "Prawdopodobieństwo",
        main = "Rozkłady empiryczny i teoretyczny liczby błędów",
        col = c("red", "blue"), legend = rownames(counts), beside = TRUE)
```

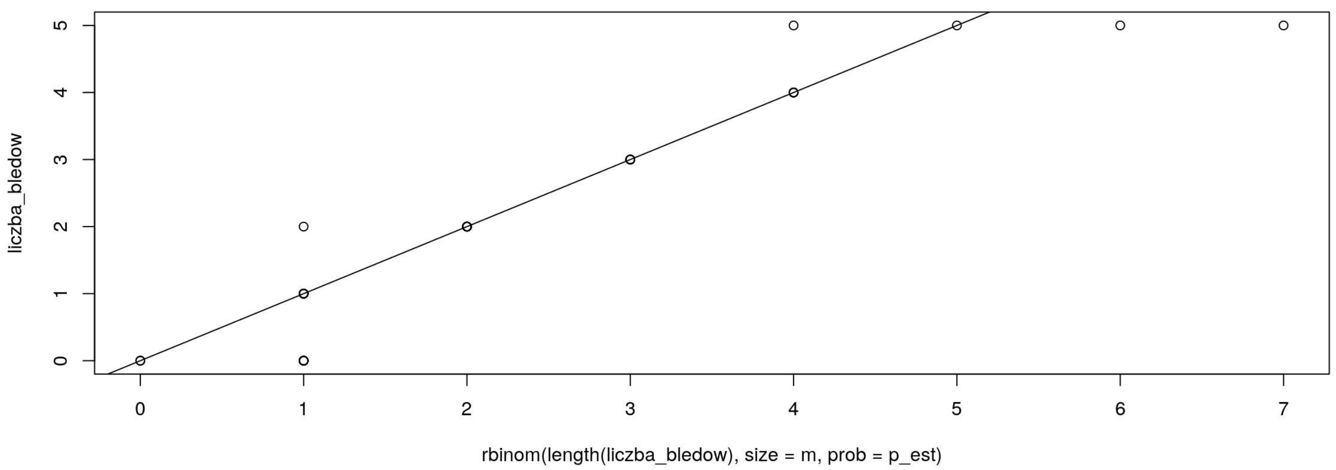


Wykres kwantyl-kwantyl

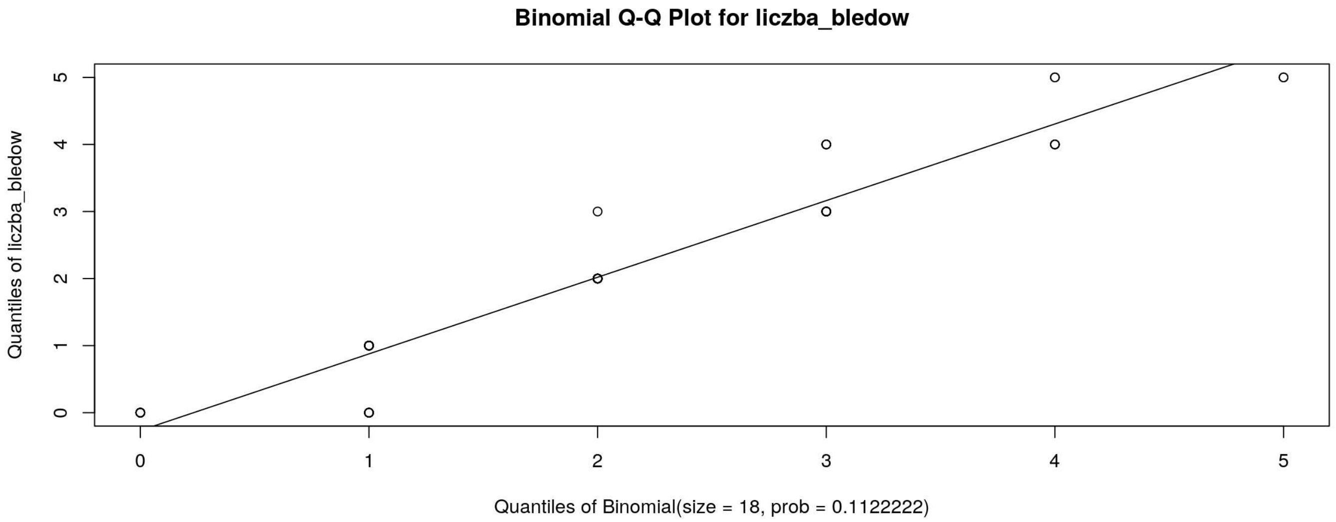
- Wykres kwantyl-kwantyl (Q-Q plot), jest wykresem zaobserwowanych statystyk porządkowych z losowej próby (kwantyle empiryczne) do odpowiadającym im (oszacowanym) wartościom średniej lub mediany w oparciu o założony rozkład lub w stosunku do empirycznych kwantyli innego zestawu danych.

- Wykresy kwantyl-kwantyl służą do oceny, czy dane pochodzą z określonego rozkładu lub czy dwa zestawy danych mają ten sam rozkład. Jeśli rozkłady mają ten sam kształt (ale niekoniecznie te same parametry położenia lub skali), wówczas wykres układa się mniej więcej na linii prostej. Jeśli rozkłady są dokładnie takie same, wówczas wykres układa się mniej więcej na linii prostej $y = x$.
- Najpierw wybiera się zbiór kwantyli pewnych rzędów. Punkt (x, y) na wykresie odpowiada jednemu z kwantyli drugiego rozkładu (współrzędna y) wykreślonemu względem kwantyla tego samego rzędu pierwszego rozkładu (współrzędna x).
- ,,qqline'' dodaje linię do ,,teoretycznego'' wykresu kwantyl-kwantyl, która przechodzi przez kwantyle rzędów `probs = c(0.25, 0.75)` , czyli domyślnie pierwszy i trzeci kwartyl.

```
# wykres kwantyl-kwantyl
qqplot(rbinom(length(liczba_bledow), size = m, prob = p_est), liczba_bledow)
qqline(liczba_bledow, distribution = function(probs) { qbinom(probs, size = m, prob = p_
```




```
# lub
library(EnvStats)
EnvStats::qqPlot(liczba_bledow,
                 distribution = "binom",
                 param.list = list(size = m, prob = p_est),
                 add.line = TRUE)
```

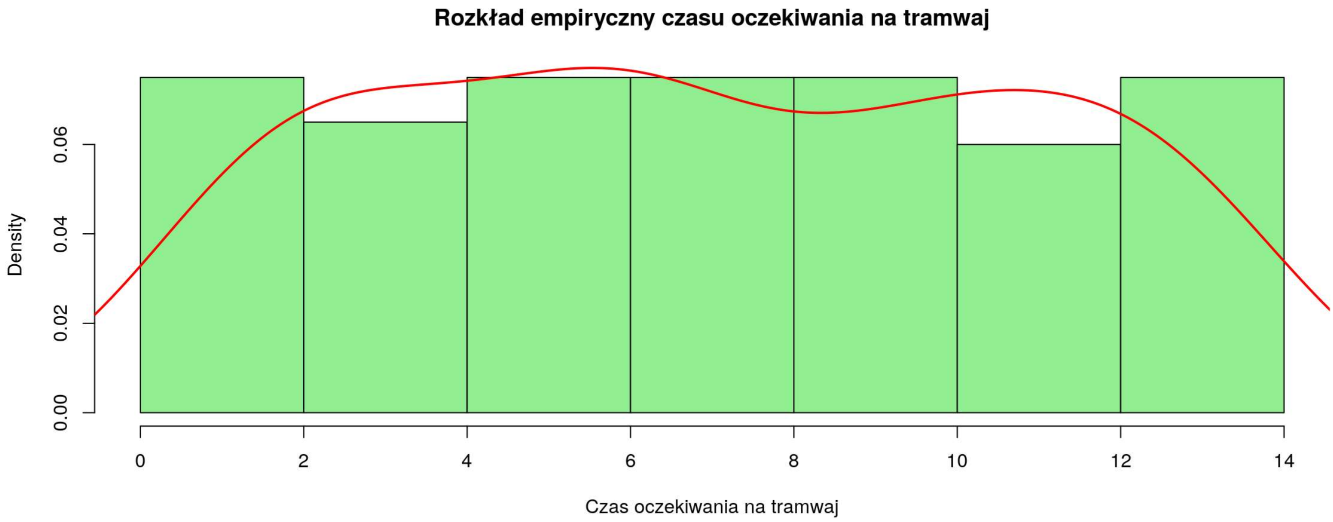


Przykład 2. Badano czas oczekiwania na tramwaj, który kursuje w jednakowych odstępach czasu. Plik czas_oczek_tramwaj.RData zawiera dane dotyczące czasu oczekiwania na tramwaj (wyrażonego w minutach) 100 osób wybranych losowo. Zmienna X to czas oczekiwania na tramwaj. Jest to zmienna ilościowa ciągła.

```
load(url("http://ls.home.amu.edu.pl/data_sets/czas_oczek_tramwaj.RData"))
head(czas_oczek_tramwaj)
```

```
## [1]  4.03 11.04  5.73 12.36 13.17  0.64
```

```
# histogram z estymatorem jądrowym gęstości
hist(czas_oczek_tramwaj,
     xlab = "Czas oczekiwania na tramwaj",
     main = "Rozkład empiryczny czasu oczekiwania na tramwaj",
     probability = TRUE,
     col = "lightgreen")
lines(density(czas_oczek_tramwaj), col = "red", lwd = 2)
```



- model: rozkład jednostajny
- $\mathcal{P} = \{U(a, b) : a, b \in \mathbb{R}, a < b\}$
- $\Theta = \{(a, b) \in \mathbb{R}^2 : a < b\}$ oraz $\theta = (a, b)$

```
load(url("http://ls.home.amu.edu.pl/data_sets/czas_oczek_tramwaj.RData"))
# estmatory
(a_est <- min(czas_oczek_tramwaj))
```

```
## [1] 0.01
```

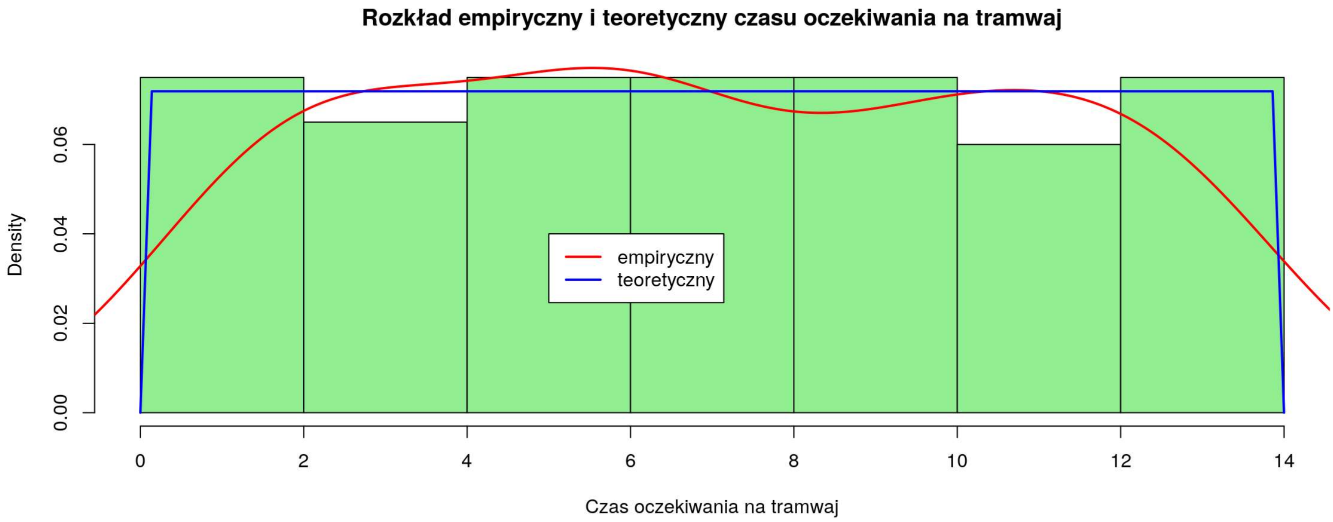
```
(b_est <- max(czas_oczek_tramwaj))
```

```
## [1] 13.92
```

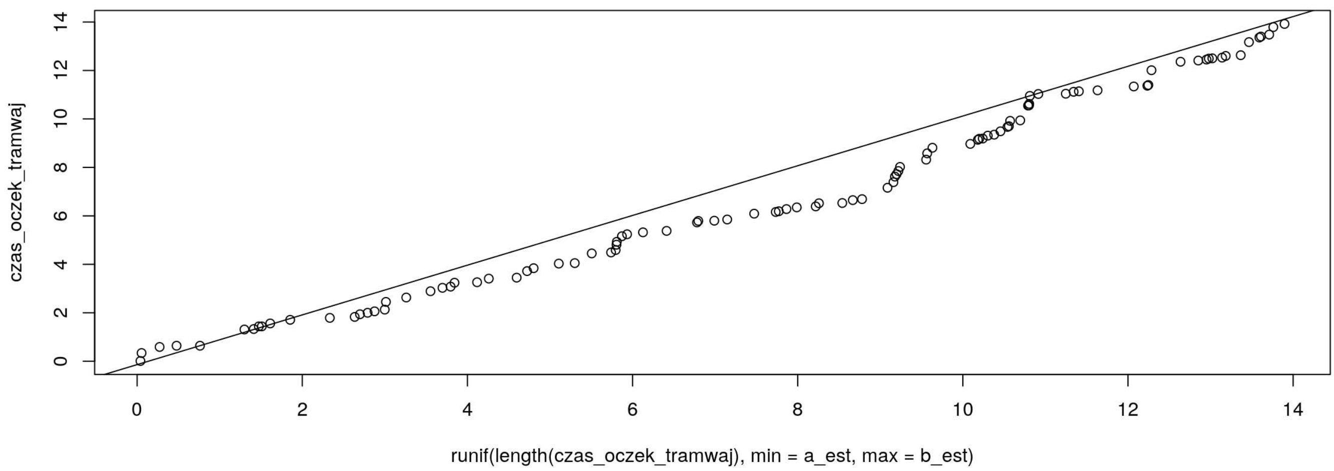
```
library(EnvStats)
EnvStats::eunif(czas_oczek_tramwaj, method = "mle")
```

```
## 
## Results of Distribution Parameter Estimation
## --------------------------------------------
## 
## Assumed Distribution:            Uniform
## 
## Estimated Parameter(s):          min =  0.01
##                                  max = 13.92
## 
## Estimation Method:               mle
## 
## Data:                            czas_oczek_tramwaj
## 
## Sample Size:                     100
```

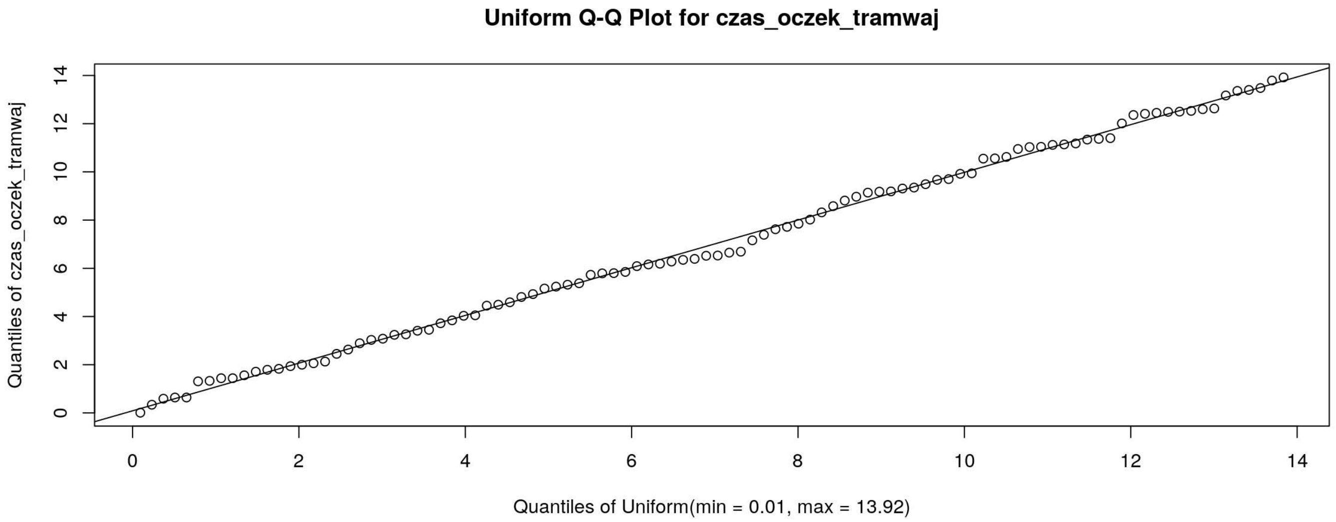
```
# histogram z estymatorem jądrowym gęstości
hist(czas_oczek_tramwaj,
     xlab = "Czas oczekiwania na tramwaj",
     main = "Rozkład empiryczny i teoretyczny czasu oczekiwania na tramwaj",
     probability = TRUE,
     col = "lightgreen")
lines(density(czas_oczek_tramwaj), col = "red", lwd = 2)
curve(dunif(x, a_est, b_est),
      add = TRUE, col = "blue", lwd = 2)
legend(x = 5, y = 0.04, legend = c("empiryczny", "teoretyczny"), col = c("red", "blue"),
```




```
# wykres kwantyl-kwantyl
qqplot(runif(length(czas_oczek_tramwaj), min = a_est, max = b_est), czas_oczek_tramwaj)
qqline(czas_oczek_tramwaj, distribution = function(probs) { qunif(probs, min = a_est, ma
```




```
# lub
library(EnvStats)
EnvStats::qqPlot(czas_oczek_tramwaj,
                 distribution = "unif",
                 param.list = list(min = a_est, max = b_est),
                 add.line = TRUE)
```



- Empiryczne i teoretyczne prawdopodobieństwo, że czas oczekiwania na tramwaj jest większy niż 10 minut, można obliczyć w następujący sposób:

```
# empirycznie
mean(czas_oczek_tramwaj > 10)
```

```
## [1] 0.27
```

```
# teoretycznie: X ~ U(a_est, b_est) oraz P(X > 10) = 1 - P(X <= 10) = 1 - F(10)
1 - punif(10, min = a_est, max = b_est)
```

```
## [1] 0.2818116
```

5.3 Zadania

Zadanie 1. Niech $\mathbf{X} = (X_1, \ldots, X_n)^\top$ będzie próbą prostą z populacji o rozkładzie jednostajnym $U(a, b)$.

1. Pokaż, że estymatory metody momentów parametrów a i b w rozkładzie jednostajnym $U(a, b)$ są postaci:

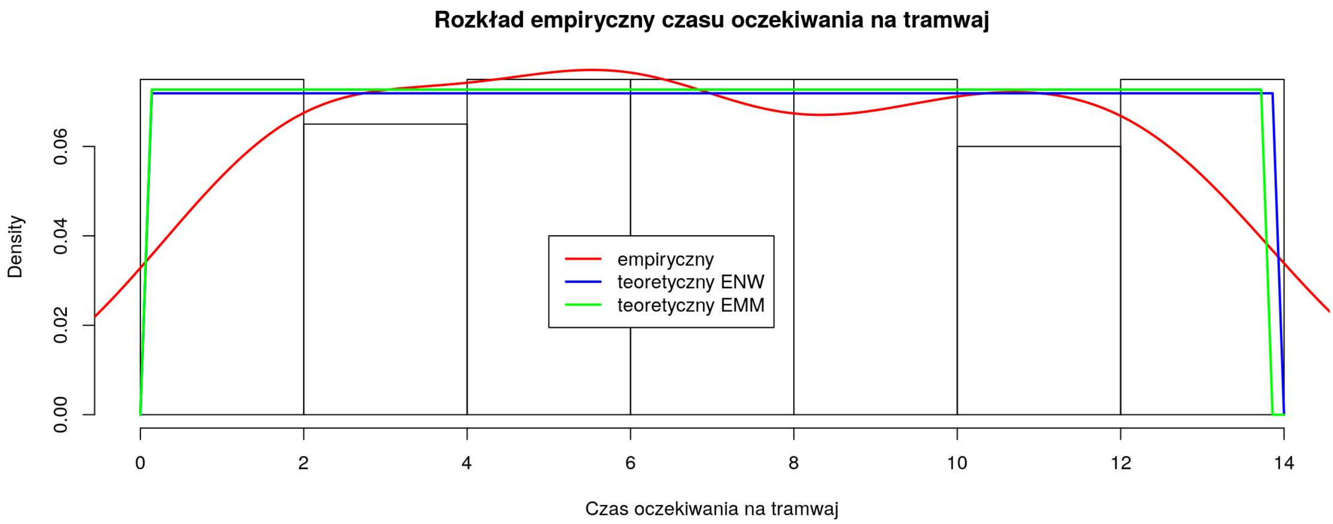
$$\hat{a} = \bar{X} - \sqrt{3}\widetilde{S},\ \hat{b} = \bar{X} + \sqrt{3}\widetilde{S},$$

gdzie $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ oraz $\widetilde{S} = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$.

2. Oblicz wartości tych estymatorów dla danych z przykładu dotyczącego czasu oczekiwania na tramwaj.

```
##
## Results of Distribution Parameter Estimation
## --------------------------------------------
##
## Assumed Distribution:            Uniform
##
## Estimated Parameter(s):          min =  0.1040974
##                                  max = 13.8551026
##
## Estimation Method:               mme
##
## Data:                            czas_oczek_tramwaj
##
## Sample Size:                     100
```

3. Zilustruj otrzymane teoretyczne funkcje gęstości korzystające z ENW i EMM na histogramie.



Zadanie 2. Przebadano 200 losowo wybranych 5-sekundowych okresów pracy centrali telefonicznej. Rejestrowano liczbę zgłoszeń. Wyniki są zawarte w pliku Centrala.RData.

1. Zasugeruj rozkład teoretyczny badanej zmiennej.

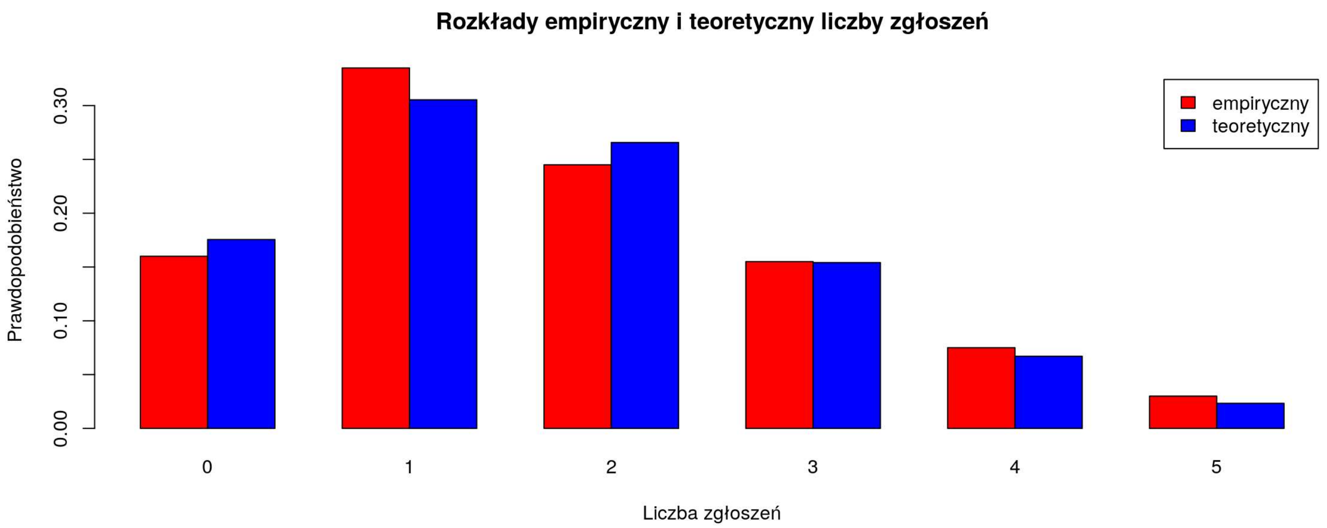
2. Oblicz wartość estymatora parametru rozkładu teoretycznego.

```
## [1] 1.74
```

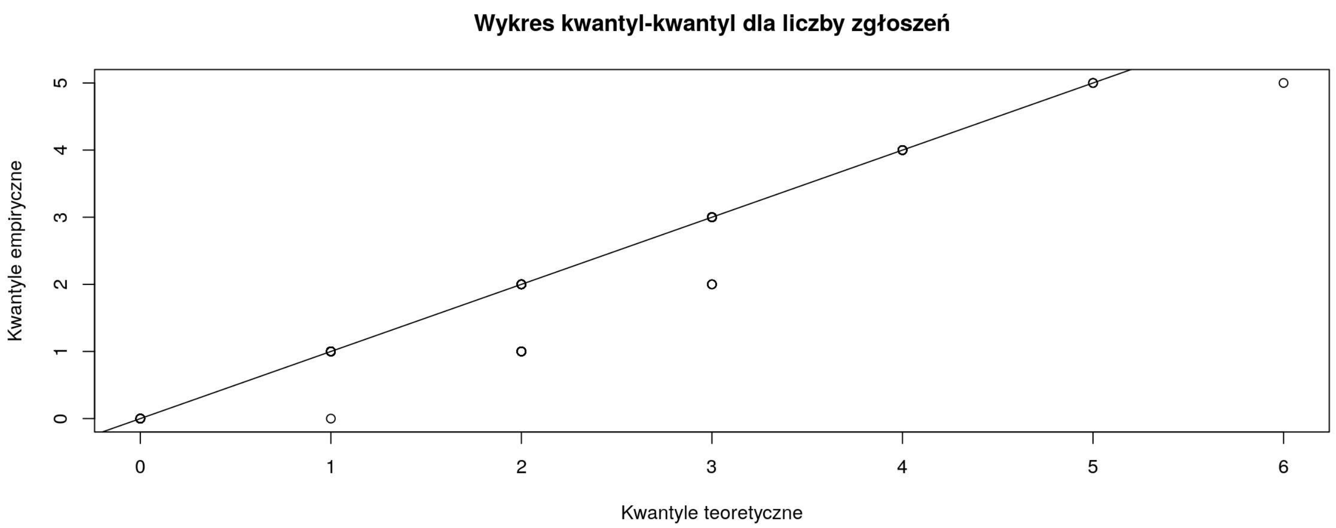
3. Porównaj empiryczne prawdopodobieństwa wystąpienia poszczególnych wartości liczby zgłoszeń w próbie z wartościami teoretycznymi uzyskanymi na podstawie rozkładu teoretycznego.

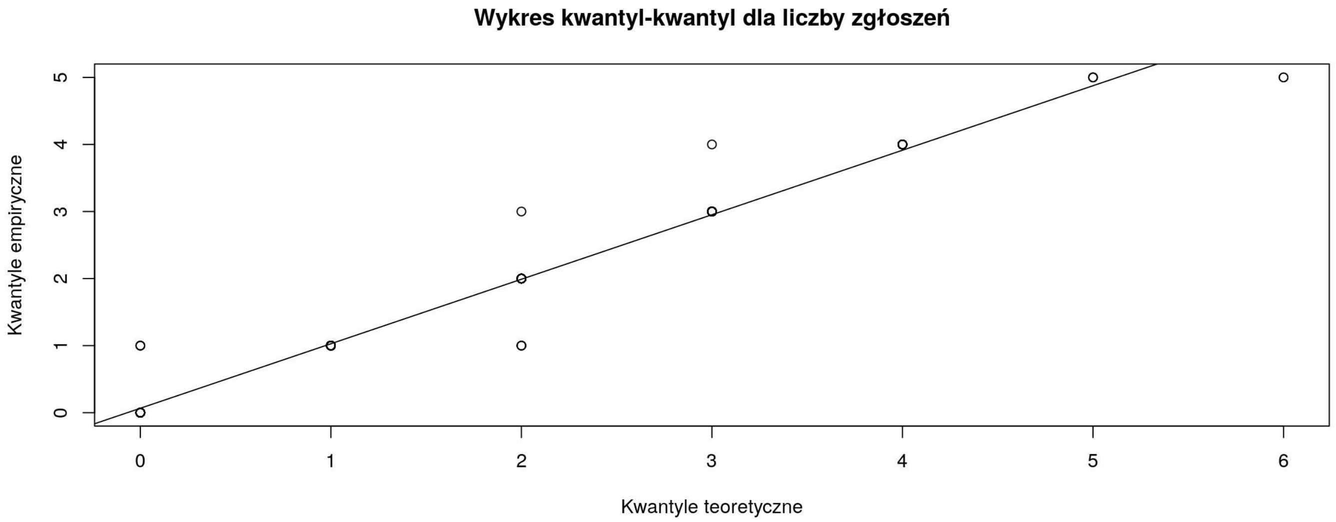
```
## [1] 0.9911019
```

```
##                     0         1         2         3          4          5
## empiryczny  0.1600000 0.3350000 0.2450000 0.1550000 0.07500000 0.03000000
## teoretyczny 0.1755204 0.3054055 0.2657028 0.1541076 0.06703681 0.02332881
```



4. Sprawdź dopasowanie rozkładu teoretycznego za pomocą wykresy kwantyl-kwantyl.





5. Czy na podstawie powyższych rozważań rozkład teoretyczny wydaje się odpowiedni?
6. Oblicz prawdopodobieństwo empiryczne i teoretyczne, że liczba zgłoszeń jest mniejsza niż 4.

```
## [1] 0.895
```

```
## [1] 0.9007363
```

Zadanie 3. Niech $\mathbf{X} = (X_1, \ldots, X_n)^\top$ będzie próbą prostą z rozkładu Rayleigha o gęstości:

$$f_\lambda(x) = \frac{2}{\lambda} x \exp\left(-\frac{x^2}{\lambda}\right) I_{(0,\infty)}(x),\ \lambda > 0.$$

Pokaż, że ENW parametru λ jest postaci:

$$\frac{1}{n}\sum_{i=1}^n X_i^2.$$

W tym celu przeprowadź następujące kroki:

1. Pokaż, że funkcja wiarogodności wynosi:

$$L(\lambda; \mathbf{x}) = f_\lambda(\mathbf{x}) = \prod_{i=1}^n f_\lambda(x_i) = \left(\frac{2}{\lambda}\right)^n \left(\prod_{i=1}^n x_i\right) \exp\left(-\frac{1}{\lambda}\sum_{i=1}^n x_i^2\right).$$

2. Wprowadź pomocniczą funkcję:

$$l = \ln L(\lambda; \mathbf{x}) = n \ln 2 - n \ln \lambda + \ln\left(\prod_{i=1}^{n} x_i\right) - \frac{1}{\lambda} \sum_{i=1}^{n} x_i^2.$$

3. Wyznacz pochodną funkcji l względem λ:

$$\frac{\partial l}{\partial \lambda} = \frac{1}{\lambda^2} \sum_{i=1}^{n} x_i^2 - \frac{n}{\lambda}.$$

4. Przyrównaj powyższą pochodną do zera i rozwiąż otrzymane równanie.

Zadanie 4. Notowano pomiary średniej szybkości wiatru w odstępach 15 minutowych wokół nowo powstającej elektrowni wiatrowej. Wyniki są następujące:

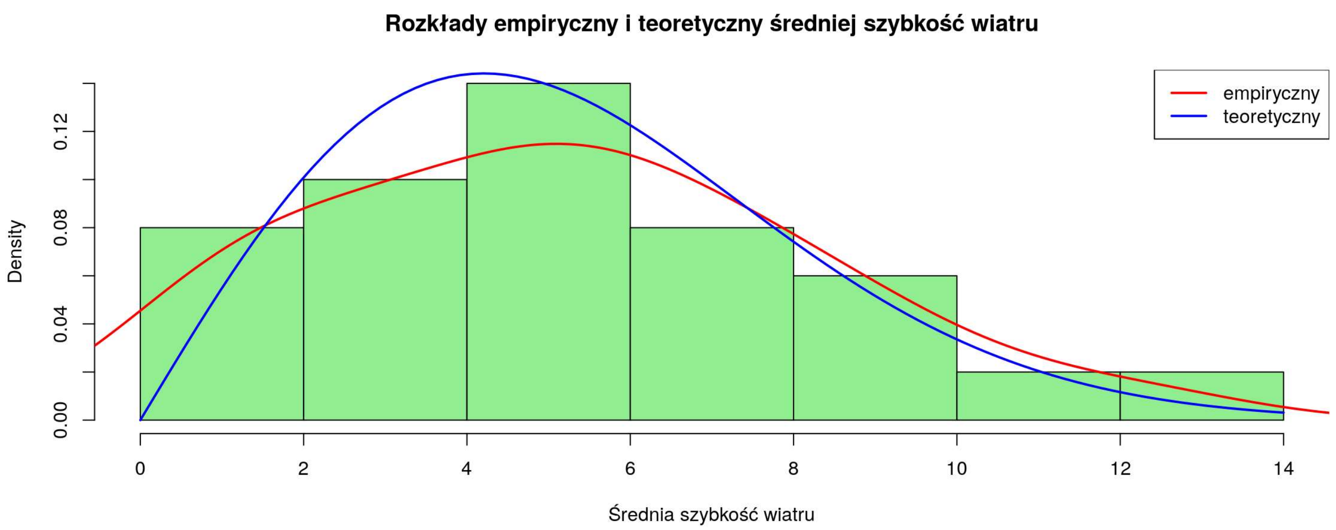
0.9	6.2	2.1	4.1	7.3
1.0	4.6	6.4	3.8	5.0
2.7	9.2	5.9	7.4	3.0
4.9	8.2	5.0	1.2	10.1
12.2	2.8	5.9	8.2	0.5

1. Zasugeruj rozkład teoretyczny badanej zmiennej.

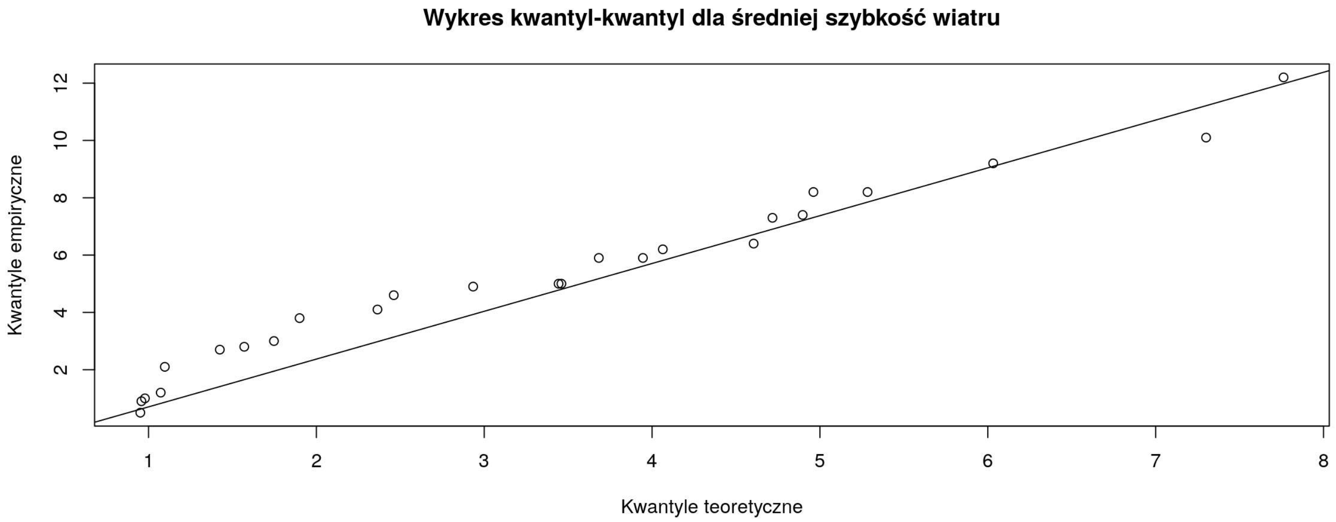
2. Oblicz wartość ENW parametru rozkładu teoretycznego.

```
## [1] 35.42
```

3. Porównaj rozkład empiryczny wystąpienia poszczególnych wartości średniej szybkości wiatru w próbie z wartościami teoretycznymi uzyskanymi na podstawie rozkładu teoretycznego.



4. Sprawdź dopasowanie rozkładu teoretycznego za pomocą wykresy kwantyl-kwantyl.



5. Czy na podstawie powyższych rozważań rozkład teoretyczny wydaje się odpowiedni?
6. Oblicz empiryczne i teoretyczne prawdopodobieństwo, że średnia szybkość wiatru jest zawarta w przedziale $[4, 8]$.

```
## [1] 0.44
```

```
## [1] 0.3782218
```

7. Oblicz wartość ENW dla wartości oczekiwanej i wariancji rozkładu teoretycznego.

```
## [1] 5.274353
```

```
## [1] 7.601197
```