Zadanie 1. W tym zadaniu rozważać będziemy rozkład dwumianowy $b(5,p)$. Naszym celem będzie, na podstawie generowanych przez nas prób, wyznaczenie wartości estymatora największej wiarogodności wielkości $P(X \geqslant 3)$, gdzie $X \sim b(5,p)$.

```
MLE_for_binom <- function(n, p) {
  p_estim <- sum(rbinom(n, 5, p))/(n*5)
  mle <- 0
  for (i in 3:5) mle <- mle + dbinom(i, 5, p_estim)
  return(mle)
}
```

W pierwszym kroku generujemy 50 obserwacji pochodzących kolejno z $b(5,0.1)$, $b(5,0.3)$, $b(5,0.5)$, $b(5,0.7)$, $b(5,0.9)$. Estymatory $\widehat{P(X \geqslant 3)}$ obliczone w opisany powyżej sposób są wówczas równe

$b(5,0.1)$	$b(5,0.3)$	$b(5,0.5)$	$b(5,0.7)$	$b(5,0.9)$
0.011	0.163	0.507	0.815	0.988

Możemy zauważyć, że zgodnie z przypuszczeniami, wraz ze wzrostem parametru p rozkładu z którego pochodziły próby, wzrasta wartość estymatora $\widehat{P(X \geqslant 3)}$.

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, co pozwoli nam na oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów.

```
real_p <- function(p) {
  value <- 0
  for (i in 3:5) value <- value + dbinom(i, 5, p)
  return(value)
}

stats_for_binomial <- function(n, p) {
  trials <- rep(NA, 10000)
  for (i in 1:10000) {
    trials[i] = MLE_for_binom(n, p)
  }
  return(c(var(trials), sum((trials - real_p(p))^2)/10000,
           mean(trials) - real_p(p)))
}
```

W wyniku otrzymujemy następujące statystyki

	$b(5,0.1)$	$b(5,0.3)$	$b(5,0.5)$	$b(5,0.7)$	$b(5,0.9)$
wariancja	0.000025	0.001474	0.003507	0.001465	0.000025
błąd średniokwadratowy	0.000025	0.001478	0.003507	0.001469	0.000025
obciążenie	0.000757	0.002120	0.000226	-0.001996	-0.000755

Zauważmy, że na podstawie otrzymanych statystyk dla rozkładów $b(5,0.1), b(5,0.9)$ zróżnicowanie wyników jest stosunkowo niskie oraz bliskie są one wartości estymowanego parametru θ. Również obciążenie jest mniejsze niż dla $b(5,0.3), b(5,0.7)$. Jest jednak ono większe od obciążenia przyjmowanego dla rozkładu $b(5,0.5)$.

Zadanie 2. W tym zadaniu będziemy się zajmować rozkładem Poissona z parametrem λ. Naszym celem będzie oszacowanie wartości estymatora największej wiarogodności wielkości $P(X = i)$, gdzie $i = 0, 1, \ldots, 10$ oraz $X \sim \pi(\lambda)$.

```
MLE_for_pois <- function(n, lambda) {
  lambda_estim <- mean(rpois(n, lambda))
  mle <- c()
  for (i in 0:10) mle[i] <- dpois(i, lambda_estim)
  return(sum(mle))
}
```

W pierwszym kroku generujemy 50 obserwacji pochodzących kolejno z $\pi(0.5)$, $\pi(1)$, $\pi(2)$, $\pi(5)$. Estymatory $\widehat{P(X = x)}$ obliczone w opisany powyżej sposób są wówczas równe

$\pi(0.5)$	$\pi(1)$	$\pi(2)$	$\pi(5)$
0.428791	0.617107	0.859135	0.975385

Widzimy więc, że wartość estymatora $\widehat{P(X = x)}$ rośnie wzraz ze wzrostem wartości parametru λ.

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, co pozwoli nam na oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów.

```
real_p2 <- function(lambda) {
  prob <- c()
  for (i in 0:10) prob[i] <- dpois(i, lambda)
  return(sum(prob))
}

stats_for_pois <- function(n, lambda) {
  trials <- rep(NA, 10000)
  for (i in 1:10000) {
    trials[i] = MLE_for_pois(n, lambda)
  }
  return(c(var(trials), sum((trials - real_p2(lambda))^2)/10000,
           mean(trials) - real_p2(lambda)))
}
```

W wyniku otrzymujemy następujące statystyki

	$\pi(0.5)$	$\pi(1)$	$\pi(2)$	$\pi(5)$
wariancja	0.003612	0.002751	0.000754	0.000017
błąd średniokwadratowy	0.003614	0.002768	0.000761	0.000018
obciążenie	-0.001775	-0.004196	-0.002603	-0.001277

Na podstawie otrzymanych statystyk widzimy, że wraz ze wzrostem parametru λ maleje obciążenie estymatora. Dodatkowo obserwujemy, że im większa wartość parametru λ, tym mniejsze jest zróżnicowanie wyników i ich odległość od wartości estymowanego $P(X = x)$.

Zadanie 4. Generujemy 50 obserwacji z rozkładu beta z parametrami θ, 1 dla kolejno $\theta = 0.5$, $\theta = 1$, $\theta = 2$, $\theta = 5$. Doświadczenie powtarzamy 10 000 razy, a następnie na tej podstawie wyznaczamy wartość estymatora $\widehat{I(\theta)}$ informacji Fishera parametru θ.

```
MLE_for_beta <- function(x) return(-length(x)/sum(log(x)))
MLE_for_fish_beta <- function(x) return(MLE_for_beta(x)^(-2))
```

W wyniku otrzymujemy:

$\beta(0.5, 1)$	$\beta(1, 1)$	$\beta(2, 1)$	$\beta(5, 1)$
3.997	0.997	0.249	0.04

Następnie, niezależnie, generujemy 50 obserwacji z rozkładów beta z w.w. parametrami i obliczamy wartość estymatora największej wiarogodności parametru θ. Otrzymujemy następujące wyniki:

$\beta(0.5,1)$	$\beta(1,1)$	$\beta(2,1)$	$\beta(5,1)$
0.523	0.975	1.896	4.449

W dalszym kroku definiujemy nową zmienną:

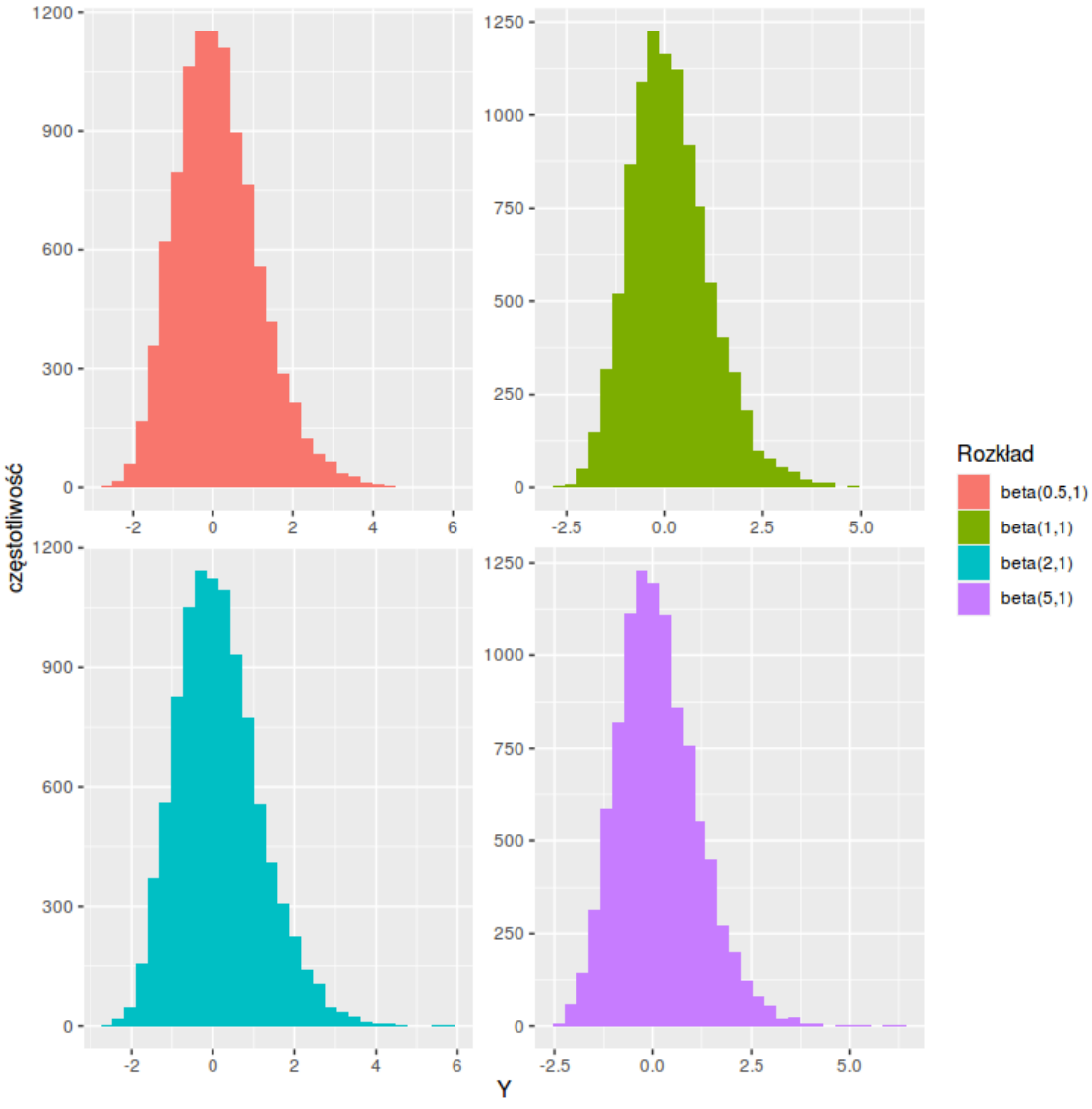
$$Y = \sqrt{n\widehat{I(\theta)}}(\widehat{\theta} - \theta),$$

```
Y_var <- function(x, theta, I) return(sqrt(length(x)*I)
                                      *(MLE_for_beta(x) - theta))
```

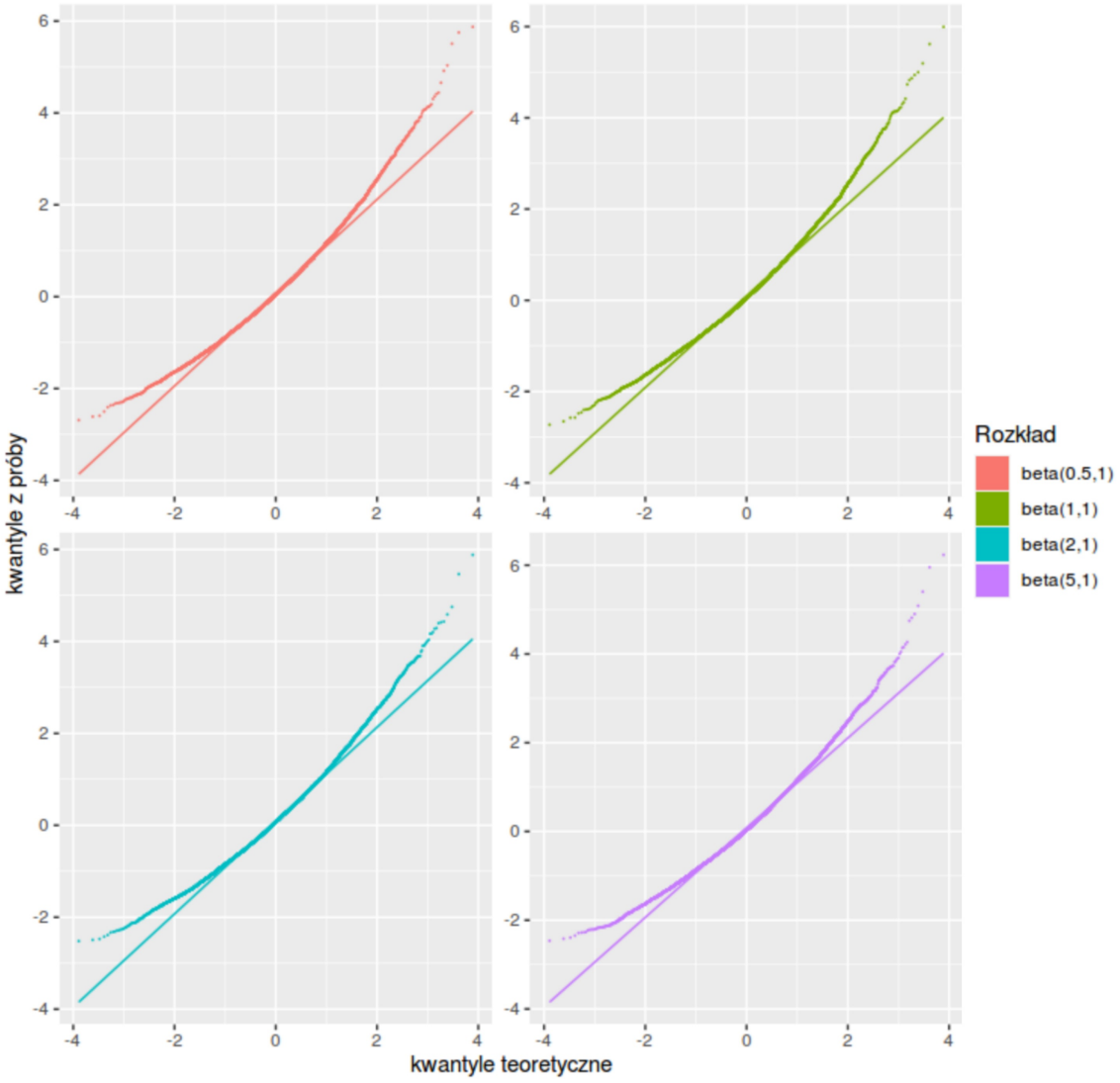
Dla wygenerowanych wcześniej prób przyjmuje ona wartości odpowiednio:

$\beta(0.5,1)$	$\beta(1,1)$	$\beta(2,1)$	$\beta(5,1)$
0.32	-0.174	-0.368	-0.78

Powyższe doświadczenie powtarzamy 10 000 razy. Uzyskane rezultaty prezentujemy na histogramie.



Patrząc na histogramy, możemy podejrzewać rozkład normalny Y. Zwerfikujemy to przy pomocy wykresów kwantylowo-kwantylowych.



Widzimy, że na wykresach kwantylowo-kwantylowych punkty układają się wzdłuż prostej. Możemy zatem wnioskować o normalności rozkładu zmiennej losowej Y.

Zadanie 5. W pierwszym kroku generujemy 50 obserwacji z rozkładu Laplace'a $L(1,1)$. Następnie na ich podstawie obliczamy wartości estymatorów parametru θ postaci

(i) $\hat{\theta}_1 = \bar{X} = (1/n)\sum_{i=1}^{n} X_i$,

```
theta_1 <- function(x) return(mean(x))
```

(ii) $\hat{\theta}_2 = \text{Me}\{X_1, \ldots, X_n\}$,

```
theta_2 <- function(x) return(median(x))
```

(iii) $\hat{\theta}_3 = \sum_{i=1}^{n} w_i X_i$, $\sum_{i=1}^{n} w_i = 1$, $0 \leqslant w_i \leqslant 1$, $i = 1, \ldots, n$, z losowym wyborem wag,

```
theta_3 <- function(x) {
    unnormed_weights <- runif(length(x), 0, 1)
    weights <- unnormed_weights/sum(unnormed_weights)
    return(sum(x * weights))
}
```

(iv) $\hat{\theta}_4 = \sum_{i=1}^{n} w_i X_{i:n}$, gdzie $X_{1:n} \leqslant \cdots \leqslant X_{n:n}$ są uporządkowanymi obserwacjami $X_1, \ldots, X_n$,

$$w_i = \varphi(\Phi^{-1}(\frac{i-1}{n})) - \varphi(\Phi^{-1}(\frac{i}{n})),$$

przy czym φ jest gęstością. a Φ dystrybuantą standardowego rozkładu normalnego $N(0,1)$

```
theta_4 <- function(x) {
    weights <- sapply(c(1:length(x)),
               function(i) dnorm(qnorm((i-1)/length(x))) -
               dnorm(qnorm(i/length(x))))
    return(sum(x * weights))
}
```

W wyniku otrzymujemy następujące wartości

$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
1.023	1.07	1.096	-0.116

Na podstawie uzyskanych wyników możemy zauważyć, że wartości estymatorów $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ są zbliżone do 1 tj. prawdziwej wartości parametru θ, natomiast wartość estymatora $\hat{\theta}_4$ znacznie odbiega od 1.

Powtarzamy powyższe doświadczenie dla rozkładu $L(4,1)$. W wyniku otrzymujemy następujące wartości

$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
3.942	3.869	4.036	0.065

Widzimy, że podobnie jak w przypadku rozkładu $L(1,1)$, dla rozkładu $L(4,1)$ najlepsze przybliżenia θ otrzymujemy dla estymatorów $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$. Dla estymatora $\hat{\theta}_4$ wyliczona wartość znacznie różni się od 4.

Powtarzamy wcześniejsze doświadczenie dla rozkładu $L(1,2)$. W wyniku otrzymujemy następujące wartości

$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
1.014	1.042	0.516	-0.409

Analogicznie do wcześniejszych doświadczeń, dla rozkładu $L(1,2)$ najlepsze przybliżenia θ są osiągane przez estymatory $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$, zaś wartość estymatora $\hat{\theta}_4$ jest znacznie odbiega od 1.

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, co pozwoli nam na oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów.

Dla estymatora $\hat{\theta}_1$ otrzymujemy następujące statystyki

	$L(1,1)$	$L(4,1)$	$L(1,2)$
wariancja	0.019833	0.020412	0.081413
błąd średniokwadratowy	0.019832	0.020410	0.081415
obciążenie	0.001176	0.000181	-0.003167

Dla estymatora $\hat{\theta}_2$ otrzymane statystyki wynoszą

	$L(1,1)$	$L(4,1)$	$L(1,2)$
wariancja	0.012052	0.012191	0.048649
błąd średniokwadratowy	0.012051	0.012193	0.048644
obciążenie	-0.000586	0.001883	0.000214

Natomiast dla estymatora $\hat{\theta}_3$ wartości statystyk są równe

	$L(1,1)$	$L(4,1)$	$L(1,2)$
wariancja	0.026373	0.026308	0.106012
błąd średniokwadratowy	0.026371	0.026305	0.106002
obciążenie	-0.000588	-0.000086	-0.001039

Widzimy zatem, że w przypadku estymatorów $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ obciążenie jest bliskie 0.

Ponadto możemy zaobserwować, że w przypadku rozkładów $L(1,1)$ oraz $L(4,1)$, dla każdego z powyższych estymatorów wariancja, podobnie jak MSE wynosi w przybliżeniu co najwyżej 0.03, a zatem zróżnicowanie wyników jest stosunkowo małe oraz są one bliskie wartości estymowanemu parametrowi θ. Jedynie dla rozkładu $L(1,2)$ wyniki są bardziej rozproszone wokół średniej, gdyż w tym przypadku wariancja osiąga wartość 0.1. Możemy również stwierdzić, że dla tego rozkładu wartości estymatora są bardziej oddalone od wartości θ, ponieważ MSE wynosi w przybliżeniu 0.11.

Przejdziemy do analizy statystyk otrzymanych dla estymatora θ_4, które wynoszą

	$L(1,1)$	$L(4,1)$	$L(1,2)$
wariancja	0.020096	0.020173	0.077875
błąd średniokwadratowy	1.012307	16.027187	1.077585
obciążenie	-0.996099	-4.000877	-0.999859

Widzimy zatem, że dla każdego z rozkładów $\hat{\theta_4}$ jest obciążony. Dodatkowo jego wartości są istotnie rozproszone oraz znacznie oddalone od prawdziwej wartości estymowanego parametru θ. Jako wniosek otrzymujemy więc, że estymator $\hat{\theta_4}$ nie stanowi dobrego przybliżenia θ.

Podsumowując możemy stwierdzić, że optymalnym estymatorem dla θ jest $\hat{\theta}_2$ ponieważ przyjmuje wartości najbliższe estymowanemu parametrowi, ponadto wariancja, MSE oraz obciążenie są w jego przypadku najmniejsze. Zauważmy, że otrzymane rezultaty pokrywają się z tymi uzyskanymi w zadaniu 1 z listy 1, jednak obciążenia estymatorów w zadaniu 1 z listy 1 były mniejsze niż tutaj.

Zadanie 6. Przejdziemy teraz do badania wpływu wielkości próby na wyniki estymacji parametrów. Zaczniemy od powtórzenia części Zadania 1., tj. wyznaczymy wartość estymatora największej wiarogodności wielkości $P(X \geqslant 3)$, gdzie $X \sim b(5,p)$. wykorzystując kolejno 20- oraz 100-elementową próbę z rozkładu $b(5, 0.1)$. W poniższej tabeli prezentujemy otrzymane wyniki.

20 obserwacji	100 obserwacji	$P(X \geqslant 3)$
0.005	0.012	0.009

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, co pozwoli nam na oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów. Dla próby 20-elementowej otrzymujemy następujące statystyki

	20 obserwacji	100 obserwacji
wariancja	0.000073	0.000012
błąd średniokwadratowy	0.000077	0.000012
obciążenie	0.001984	0.000403

Widzimy zatem, że wraz ze wzrostem rozmiaru próby, zwiększa się obciążenie i MSE. Z drugiej strony jednak dla mniejszego rozmiaru próby rozproszenie wartości estymatorów jest większe niż w przypadku próby 100-elementowej.

Przejdziemy teraz do powtórzenia powyższej procedury dla rozkładu Poissona $\pi(0.5)$ i estymatora największej wiarogodności wielkości $P(X = i)$, gdzie $i = 0, 1, \ldots, 10$. W poniższej tabeli prezentujemy otrzymane wyniki.

20 obserwacji	100 obserwacji	$P(X = i)$
0.478	0.429	0.095

	20 obserwacji	100 obserwacji
wariancja	0.008965	0.001812
błąd średniokwadratowy	0.009034	0.001813
obciążenie	-0.008350	-0.000968

Widzimy zatem, że w tym przypadku wraz ze wzrostem wiekości próby, wzrasta optymalność estymatora największej wiarogodności wiekości $P(X = i)$.

Przejdziemy teraz do powtórzenia doświadczenia omawianego w zadaniu 4 dla rozkładu beta $\beta(0.5, 1)$. Generujemy kolejno 20 oraz 100 obserwacji z rozkładu $\beta(0.5, 1)$. Doświadczenie powtarzamy 10 000 razy, a następnie na tej podstawie wyznaczamy wartość estymatora $\widehat{I(\theta)}$ informacji Fishera parametru θ. W wyniku otrzymujemy:

20 obserwacji	100 obserwacji
4.004	4.008

Następnie, niezależnie, generujemy kolejno 20 oraz 100 obserwacji z powyższego rozkładu i obliczamy wartość estymatora największej wiarogodności parametru θ. Otrzymujemy następujące wyniki:

20 obserwacji	100 obserwacji
0.451	0.564

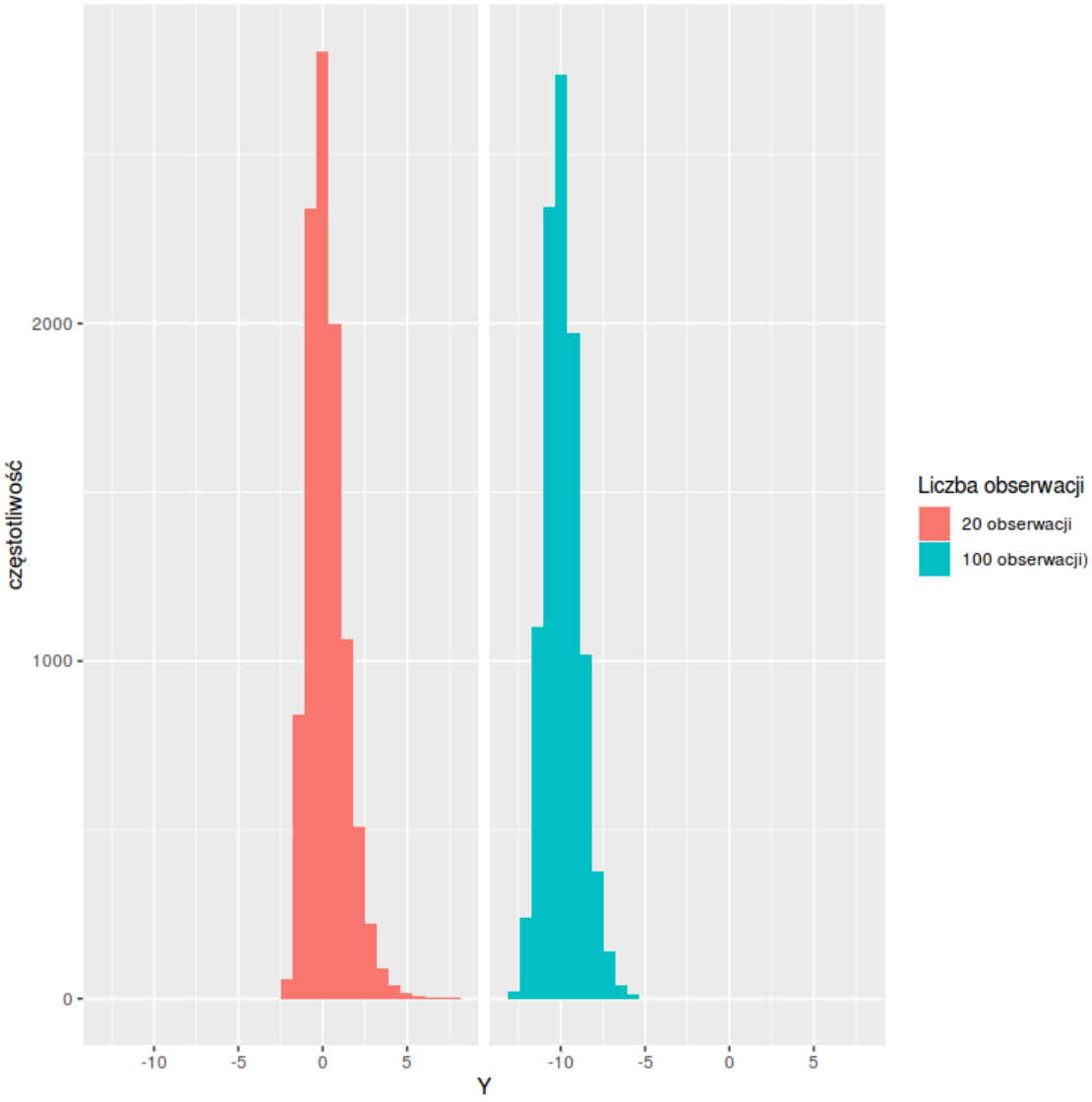
W dalszym kroku rozważamy zmienną:

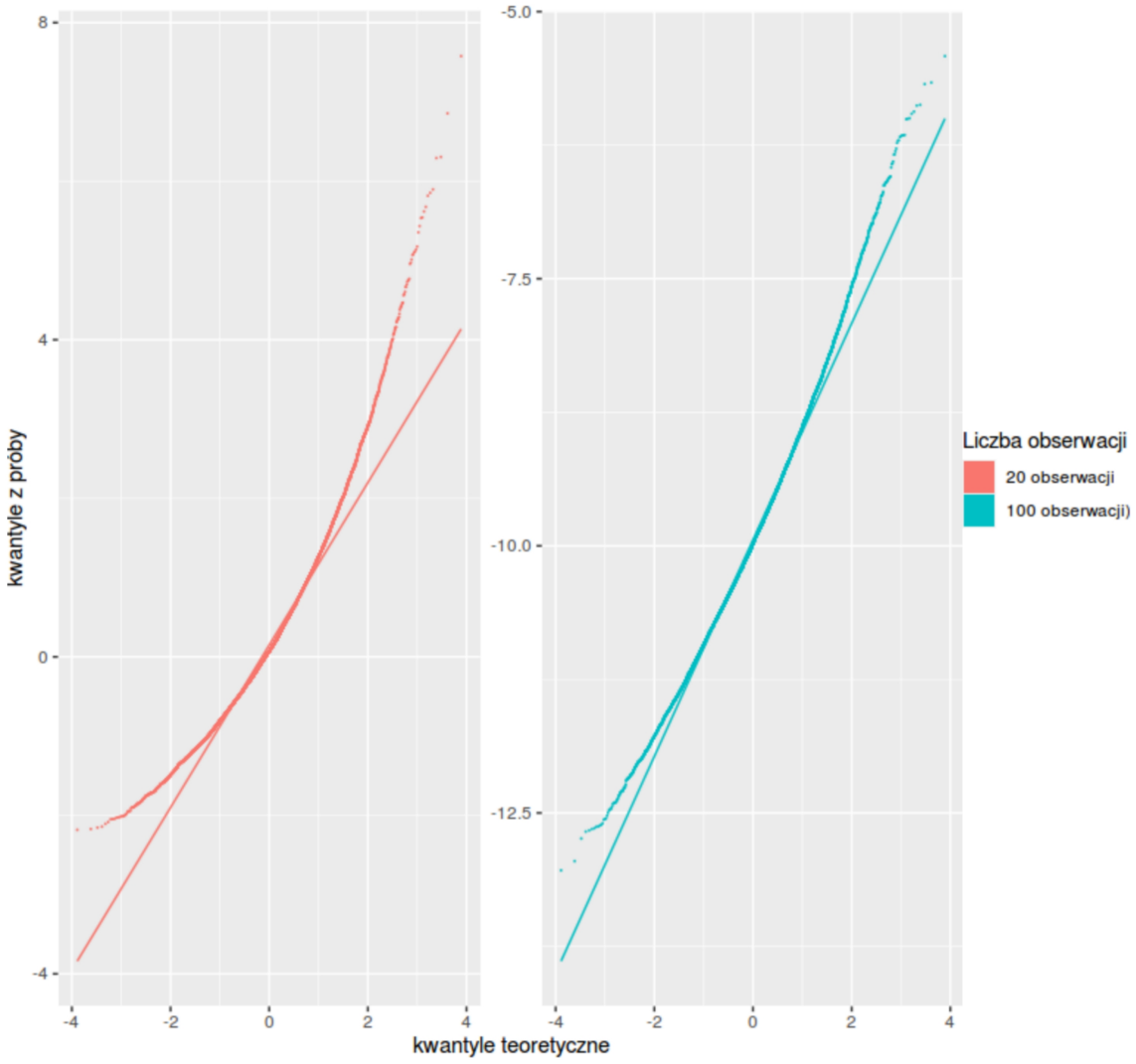
$$Y = \sqrt{n\widehat{I(\theta)}}(\widehat{\theta} - \theta),$$

Dla wygenerowanych wcześniej prób przyjmuje ona wartości odpowiednio:

20 obserwacji	100 obserwacji
-0.441	1.289

Powyższe doświadczenie powtarzamy 10 000 razy. Uzyskane rezultaty prezentujemy na histogramach oraz wykresach kwantylowo-kwantylowych.


częstotliwość
2000
1000
0
-10
-5
0
5
-10
-5
0
5
Y
Liczba obserwacji
20 obserwacji
100 obserwacji)



Widzimy zatem, że dla większego rozmiaru próby obserwacje lepiej dopasowują się do prostej niż w przypadku $n = 20$. Możemy zatem przypuszczać, że rozkład Y jest asymptotycznie normalny.

Przejdziemy teraz do powtórzenia części Zadania 5., tj. zbadamy wielkości próby na estymację parametru θ przy pomocy omawianych wcześniej estymatorów $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ oraz $\hat{\theta}_4$ wykorzystując kolejno 20- oraz 100-elementową próbę z rozkładu Laplace'a $L(1, 1)$. W poniższej tabeli prezentujemy otrzymane wyniki.

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
20 obserwacji	1.268	1.057	1.298	-0.358
100 obserwacji	0.825	0.990	1.235	-0.143

Na podstawie powyższych wyników obserwujemy, że dla większej próby estymatory $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ są bliższe 1, niż dla próby 20-elementowej. Możemy zatem przypuszczać, że poprawność estymacji wzrasta wraz ze wzrostem liczebności próby.

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, co pozwoli nam na oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów. Dla próby 20-elementowej otrzymujemy następujące statystyki

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
wariancja	0.050053	0.032889	0.065702	0.048851
błąd średniokwadratowy	0.050050	0.032887	0.065699	1.047483
obciążenie	-0.001499	0.001384	-0.001816	-0.999318

Zaś dla próby 100-elementowej otrzymane statystyki wynoszą

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
wariancja	0.010207	0.005842	0.012978	0.009855
błąd średniokwadratowy	0.010206	0.005842	0.012979	1.008195
obciążenie	-0.000498	-0.000393	-0.001747	-0.999170

Możemy zatem zauważyć, że dla większej próby wartości estymatorów są mniej rozproszone i bardziej zbliżone do rzeczywistej wartości estymowanego parametru θ. Dodatkowo obciążenie estymatorów w przypadku próby 100-elementowej jest mniejsze niż dla mniejszej próby. Możemy więc przypuszczać asymptotyczną nieobciążoność estymatorów $\hat{\theta}_1$, $\hat{\theta}_2$. Widzimy, że podobnie jak dla $n = 50$, za najoptymalniejszy przyjąć możemy estymator θ_2.