Estymatory największej wiarogodności i informacja Fishera

Antoni Bieniasz

2023-11-14

```
## Ładowanie wymaganego pakietu: stats4
## Ładowanie wymaganego pakietu: splines
```

Wyznaczanie estymatora największej wiarogodności dla rozkładu dwumianowego

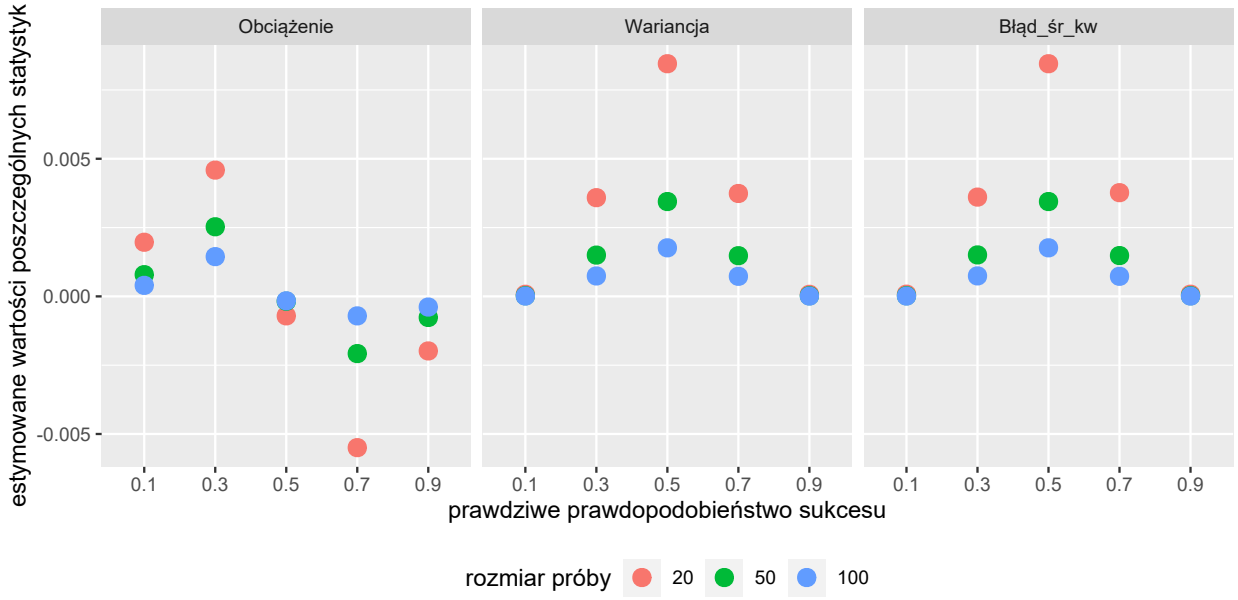
Wygenerujemy n obserwacji z rozkładu dwumianowego $b(5,p)$ dla $n \in \{20, 50, 100\}$ oraz dla każdego n dla $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Będziemy chcieli na podstawie danych prób wyznaczyć wartość estymatora największej wiarogodności (ENW) wielkości $P(X \geq 3)$, gdzie $X b(5,p)$. Powtórzymy dane doświadczenie 10000 razy. Oszacujemy wariancję, błąd średniokwadratowy oraz wariancję analizowanego estymatora. Będziemy starali się sprawdzić jak wybór parametru p, a także rozmiar próby wpływa na wyniki. Skorzystamy z twierdzenia z teorii statystyki, mówiącego, że dla próby z rozkładu o zadanej gęstości $f(x,\theta)$, dowolnej funkcji g i estymowanego parametru, takiego że $\eta = g(\theta)$, jeśli $\hat{\theta}$ jest ENW parametru θ to $g(\hat{\theta})$ jest ENW parametru $\eta = g(\theta)$. W naszym przypadku $\theta = p$ oraz $f(p) = \mathcal{P}(X \geq 3)$ Oryginalnym ENW dla parametru p z rozkładu $b(5,p)$ jest:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{5n} = \frac{\bar{X}}{5}$$

Table 1: Uśrednione wartości estymatora największej wiarogodności uzyskane na podstawie 10000 doświadczeń dla różnych wartości n oraz p

Rozmiar_próby	pstwo_0.1	pstwo_0.3	pstwo_0.5	pstwo_0.7	pstwo_0.9
20	0.0105269	0.1676683	0.4992917	0.8314240	0.9894574
50	0.0093493	0.1656095	0.4998224	0.8348414	0.9906738
100	0.0089605	0.1645244	0.4998393	0.8362114	0.9910521

Widzimy, że dla rozkładu dwumianowego o prawdopodobieństwie sukcesu równym 0,1 mamy średnią wartość ENW najbardziej zbliżoną do teoretycznej dla próby o najmniejszym rozmiarze - 20, a następnie dla prób o rozmiarach większych - 50 i 100. Dla prawdopodobieństw sukcesu równych 0,3 oraz 0,7 mamy różnice około 0,13, w możliwej skali od 0 do 1 jest to duża różnica. Dla prawdopodobieństwa 0,9 również mamy różnice, jedynie dla prawdopodobieństwa równego 0,5 obliczany estymator jest bardzo bliski tej wartości.



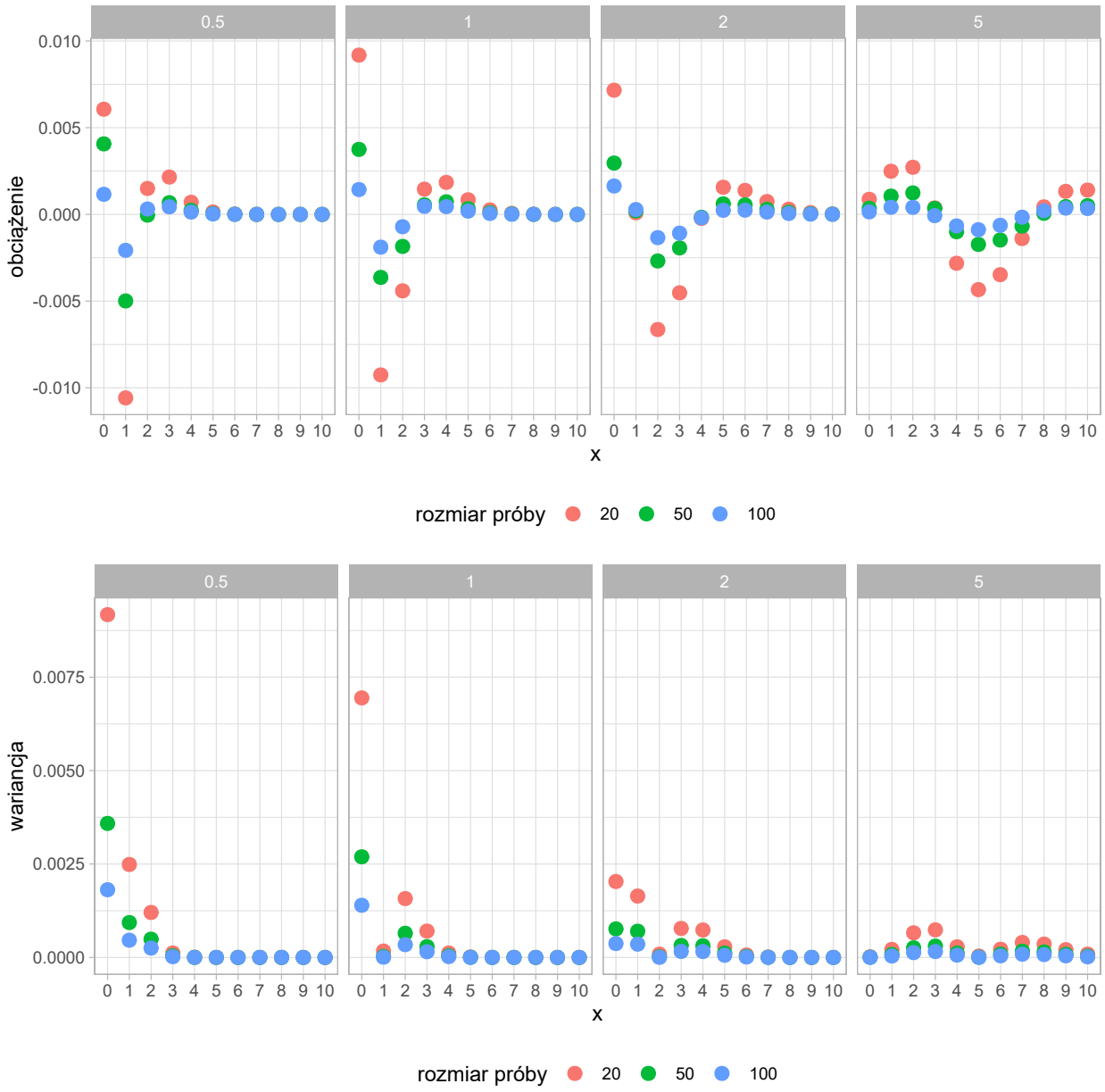
Na podstawie powyższej wizualizacji możemy stwierdzić, że im większy był rozmiar próby tym mniejsze wartości osiągały obciążenie (z dokładnością do modułu), wariancja oraz błąd średniokwadratowy danego estymatora, niezależnie od wartości parametru p określającego prawdopodobieństwo sukcesu dla rozkładu dwumianowego. Jest to wytłumaczone tym, że dla większej liczby prób, więcej wartości parametru koncentruje się około najbardziej prawdopodobnej wartości. Z kolei w przypadku parametru p, dla wariancji oraz błędu średniokwadratowego, rosły one im bardziej parametr teoretyczny, na podstawie, którego estymowano, był bardziej oddalony od wartości 0 oraz 1. Może to być wytłumaczalne tym, że, np. wokół wartości 0,5 od prawej i lewej strony mamy względnie dużo możliwych innych wartości, podczas, gdy dla wartości 0,1 jest ich dużo tylko z jednej strony.

Wyznaczanie estymatora największej wiarogodności dla rozkładu Poissona

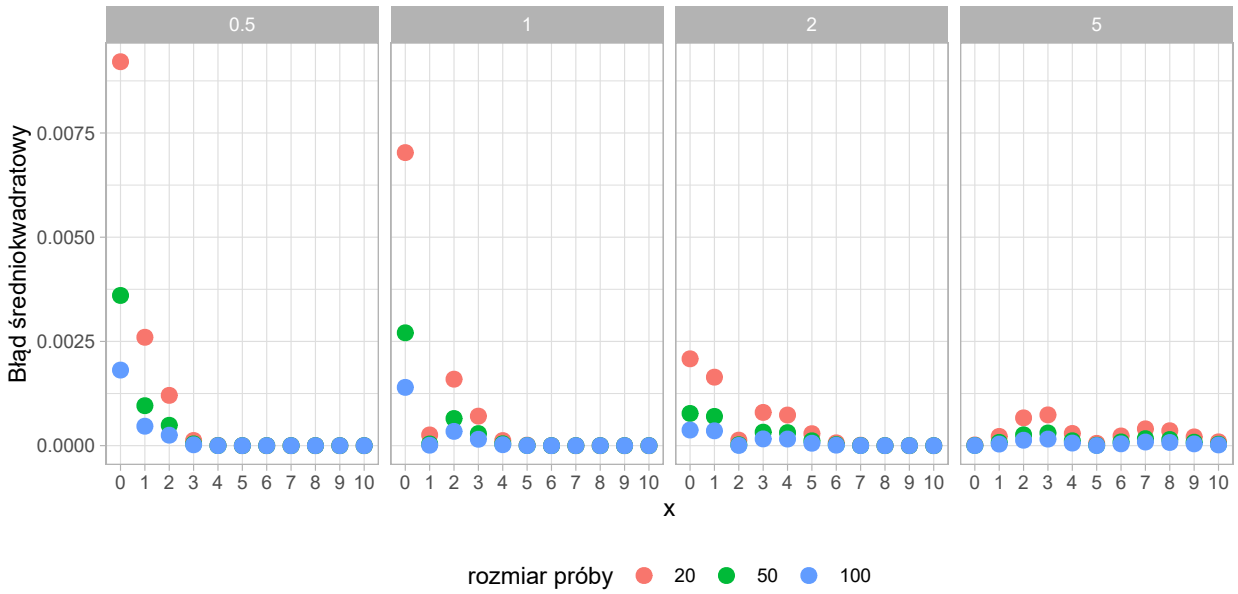
Chcemy teraz wykonać podobne czynności, co w poprzednim rozdziale, ale dla rozkładu Poissona. Wygenerujemy n obserwacji z tego rozkładu $\pi(\lambda)$ dla $n \in \{20, 50, 100\}$ oraz dla każdego n dla $\lambda \in \{0.5, 1, 2, 5\}$. Będziemy chcieli na podstawie danych prób wyznaczyć wartość estymatora największej wiarogodności (ENW) wielkości $P(X = x)$, dla $x \in \{0, 1, \ldots, 10\}$. Powtórzymy dane doświadczenie 10000 razy. Oszacujemy wariancję, błąd średniokwadratowy oraz wariancję analizowanego estymatora. Będziemy starali się sprawdzić jak wybór parametru λ, a także rozmiar próby wpływa na wyniki. Skorzystamy z tego samego twierdzenia dla ENW, jak w poprzedniej części. W naszym przypadku $\theta = \lambda$ oraz $f(\lambda) = \mathcal{P}(X = x)$ Oryginalnym ENW dla parametru λ z rozkładu $\pi(\lambda)$ jest:

$$\hat{\lambda} = \bar{X},$$

czyli średnia arytmetyczna z danej próby.


0.5
1
2
5
obciążenie
0.010
0.005
0.000
-0.005
-0.010
0 1 2 3 4 5 6 7 8 9 10
x
rozmiar próby
20
50
100

0.5
1
2
5
wariancja
0.0075
0.0050
0.0025
0.0000
0 1 2 3 4 5 6 7 8 9 10
x
rozmiar próby
20
50
100



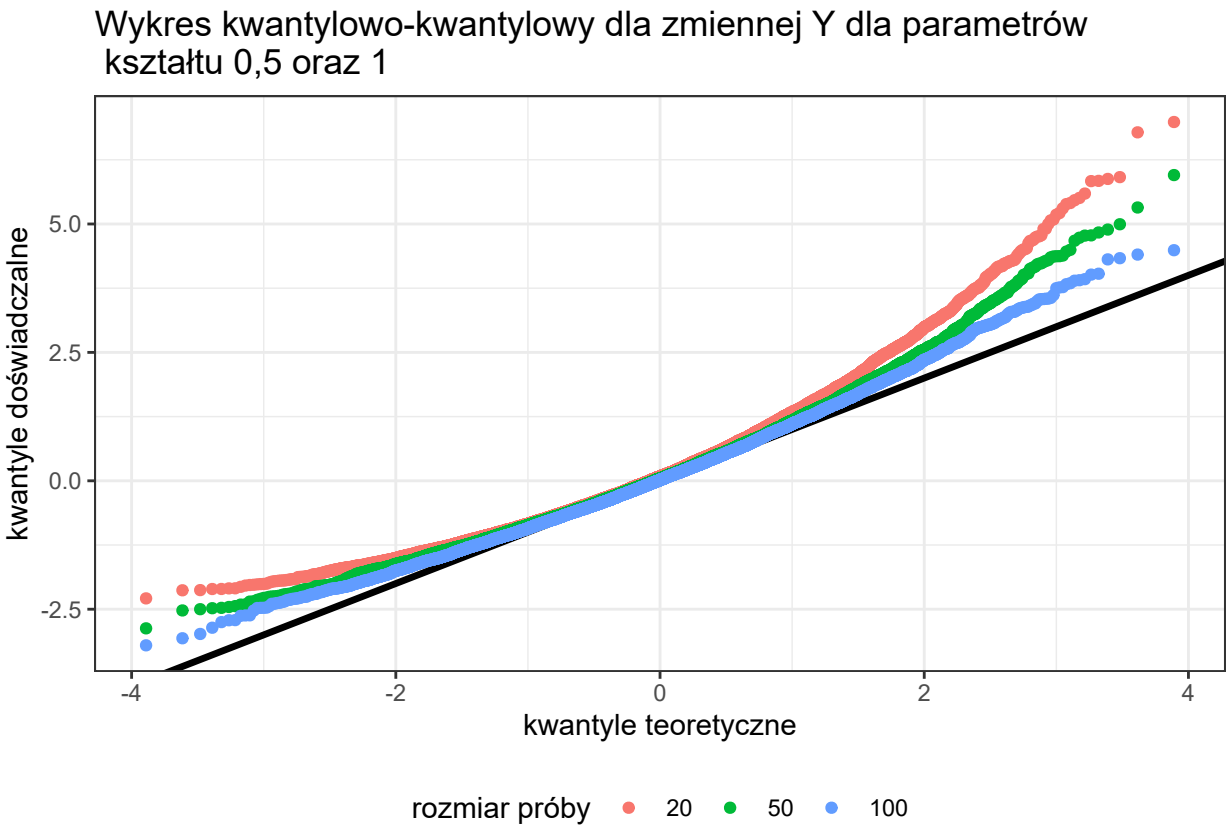
Na podstawie powyższych wykresów możemy stwierdzić, że im mniejsze x oraz rozmiar próby tym wartości obciążenia, wariancji oraz błędu średniokwadratowego są większe (z dokładnością do modułu dla obciążenia). Wpływ na wartość danych statystyk ma także zmieniająca się wartość parametru λ. Kiedy się ona zwiększa, można zauważyć, że wtedy maleją. Może mieć to związek z tym, że mniejszych λ funkcja gęstości rozkładu prawdopodobieństwa dla rozkładu Poissona (uciąglona) przypomina rozkład wykładniczy a dla większych wartości tego parametru rozkład normalny. Może to oznaczać, np. mniejszą wartość wariancji dla rozkładu normalnego a większą dla wykładniczego.

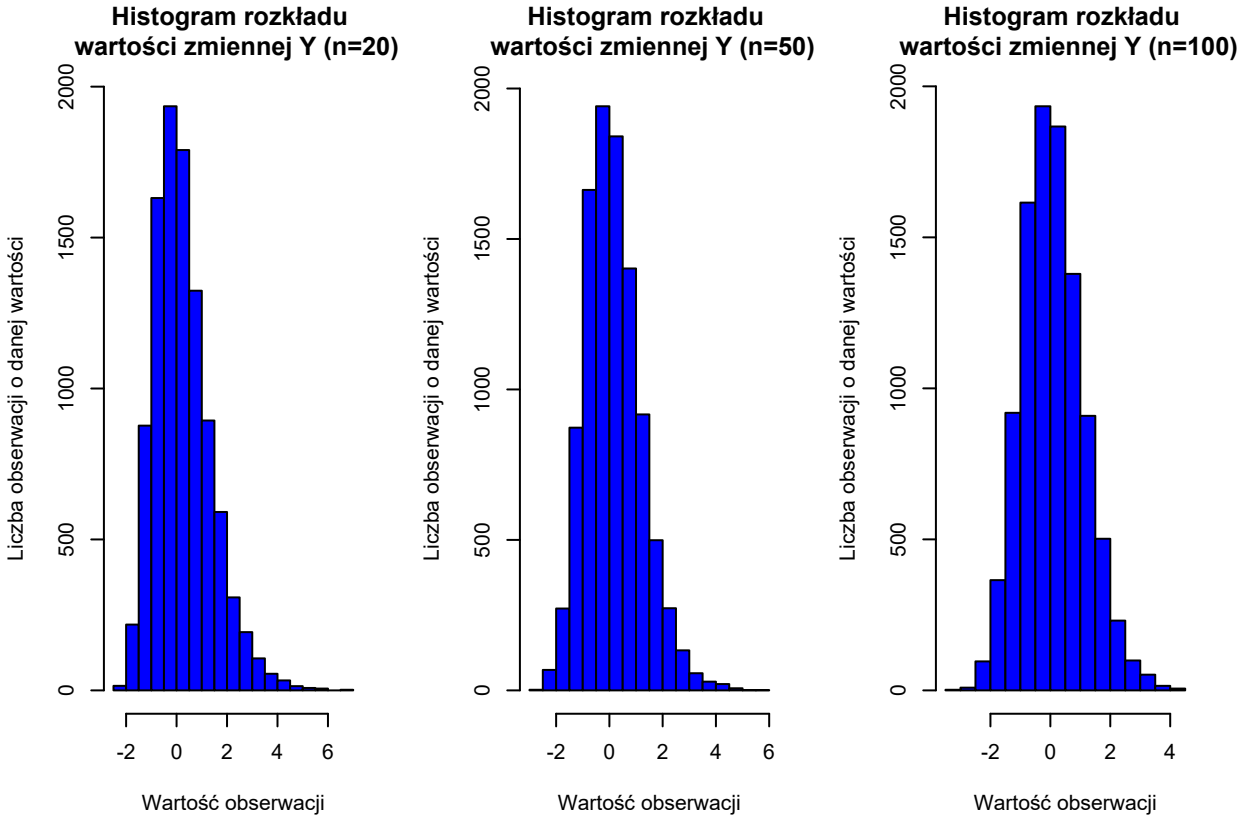
Informacja Fishera i rozkład normalny

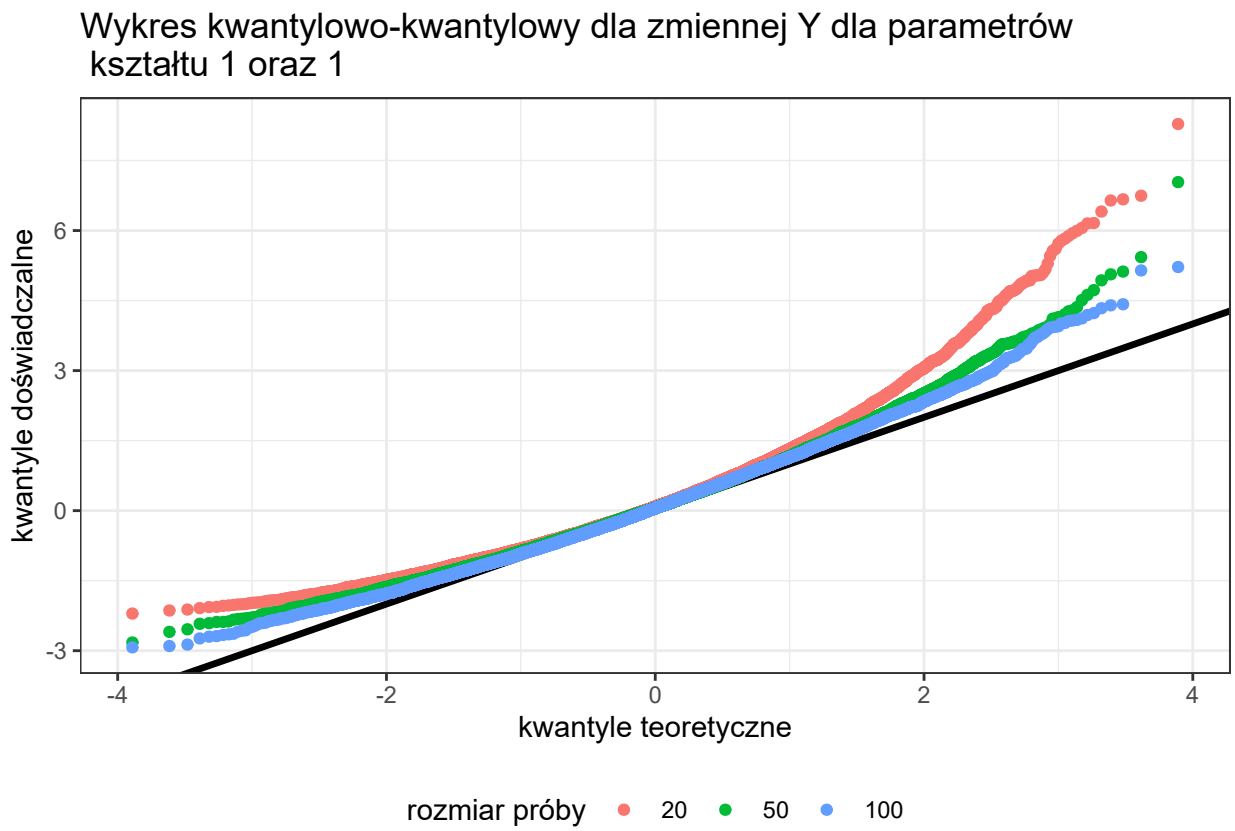
Teraz dla $n \in \{20, 50, 100\}$ i dla każdego n dla $\theta \in \{0.5, 1, 2, 5\}$ oraz drugiego parametru kształtu równego 1 wygenerujemy n obserwacji z rozkładu $b(\theta, 1)$ Powtórzymy dane doświadczenie 10 000 razy. Na podstawie uzyskanych danych obliczymy estymator $\widehat{I(\theta)}$ informacji Fishera parametru θ. Uzyskany rezultat zapamiętamy.

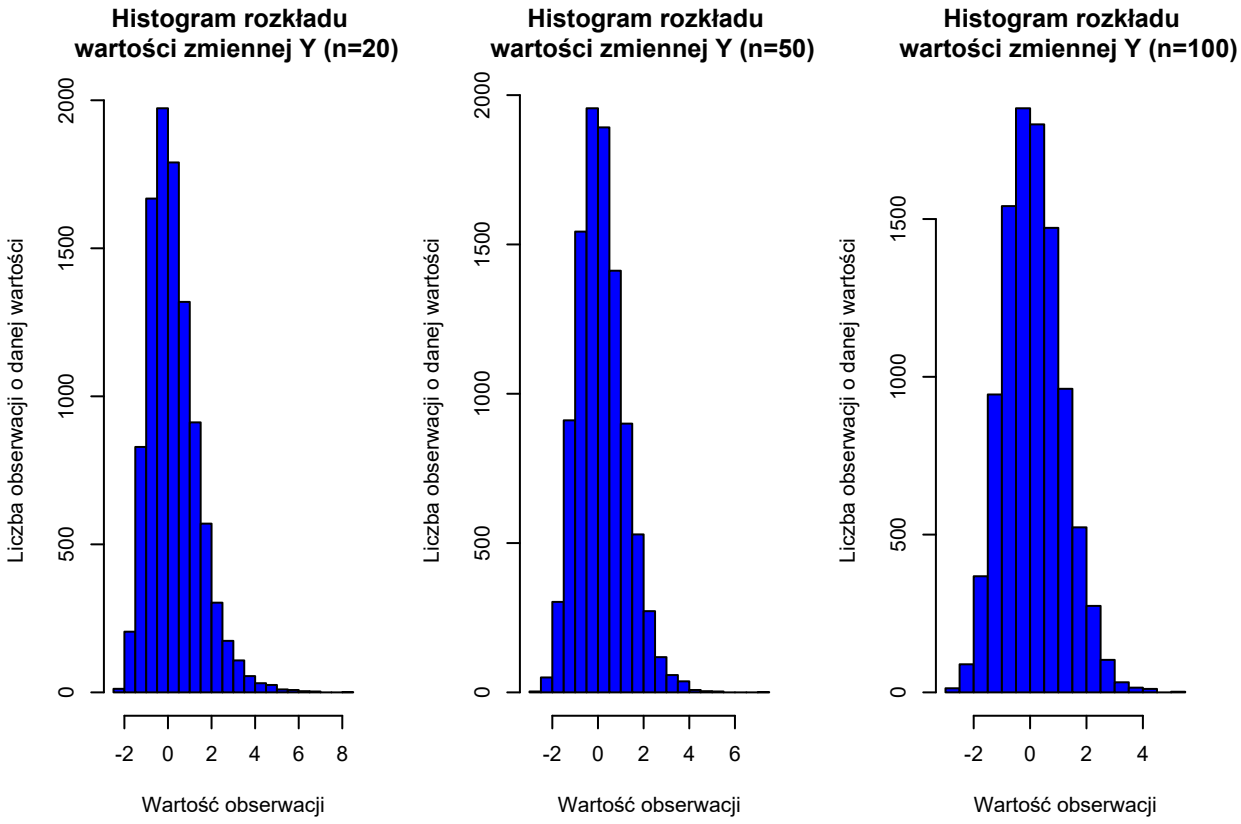
Następnie wygenerujemy niezależnie n obserwacji z rozkładu $b(\theta, 1)$. Wyznaczymy wartość estymatora największej wiarogodności parametru θ. Zdefiniujemy nową zmienną $Y = \sqrt{n\widehat{I(\theta)}}(\hat{\theta} - \theta)$ Obliczymy jej wartość na podstawie zaobserwowanej próby oraz zapamiętanego wcześniej wyniku. Powtórzymy dane doświadczenie 10 000 razy. Narysujemy histogram oraz wykres kwantylowo-kwantylowy. Wybierzemy liczbę klas na histogramie zastanowimy się nad sposobem wyznaczania kwantyli teoretycznych na wykresie kwantylowo-kwantylowym. Odpowiemy z uzasadnieniem na pytanie, czy rozkład zmiennej Y jest normalny.

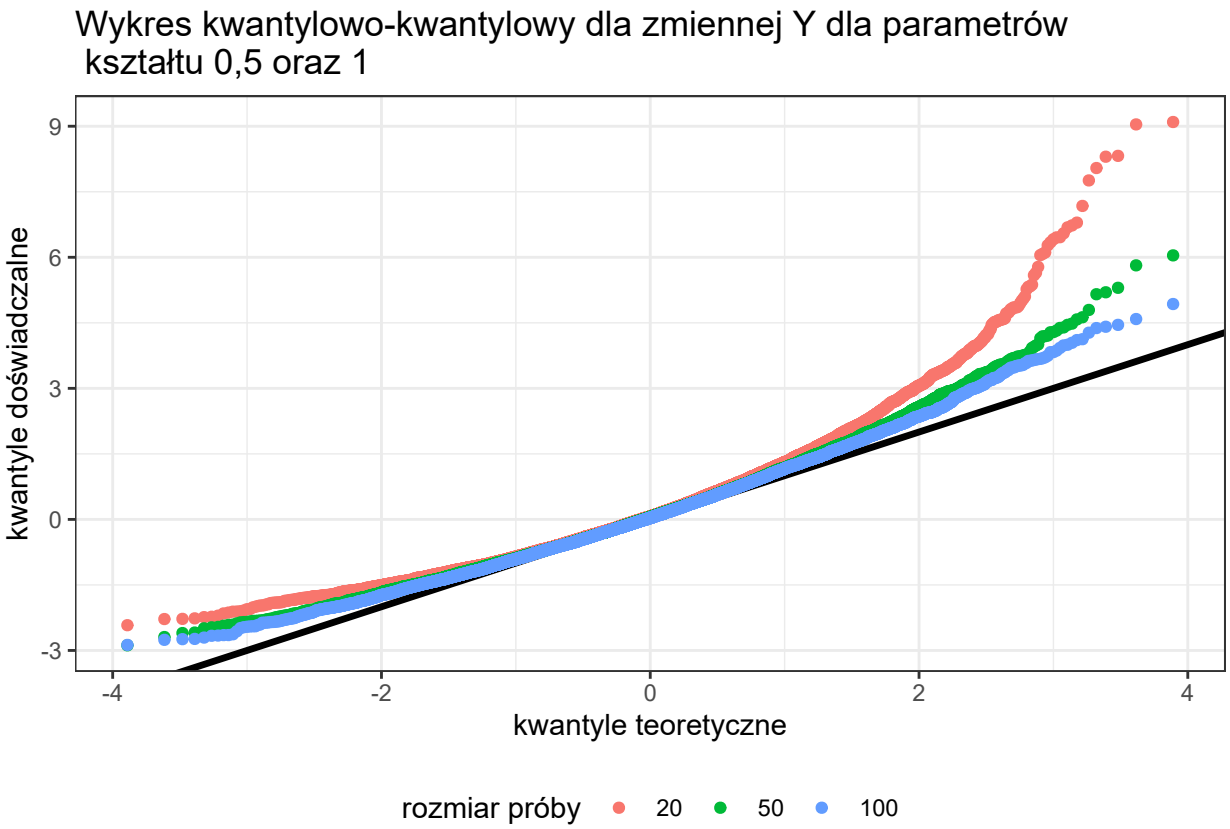
Poniższe wykresy oraz histogramy są odpowiednio dla $\theta = 0.5, 1, 2, 5$

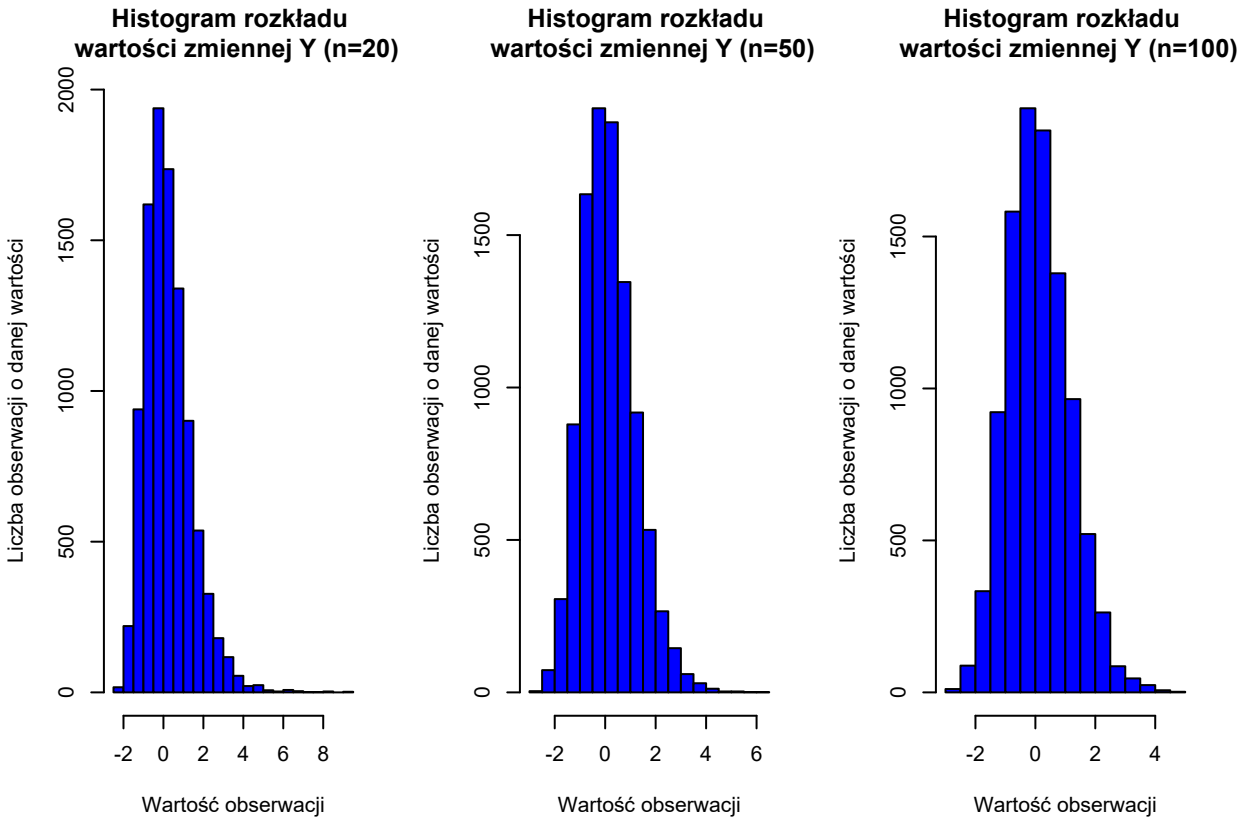

Wykres kwantylowo-kwantylowy dla zmiennej Y dla parametrów kształtu 0,5 oraz 1
kwantyle doświadczalne
5.0
2.5
0.0
-2.5
-4
-2
0
2
4
kwantyle teoretyczne
rozmiar próby
20
50
100

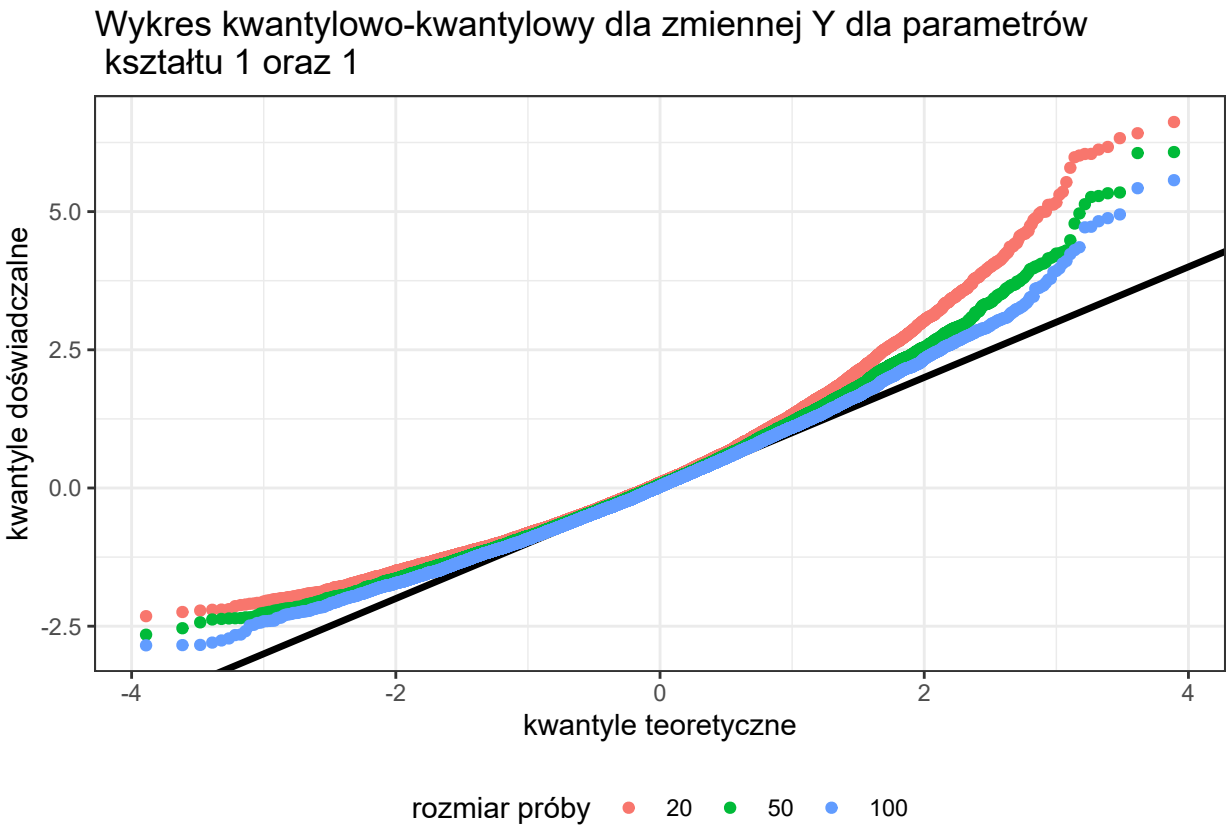

Histogram rozkładu
wartości zmiennej Y (n=20)
Liczba obserwacji o danej wartości
0
500
1000
1500
2000
-2
0
2
4
6
Wartość obserwacji
Histogram rozkładu
wartości zmiennej Y (n=50)
Liczba obserwacji o danej wartości
0
500
1000
1500
2000
-2
0
2
4
6
Wartość obserwacji
Histogram rozkładu
wartości zmiennej Y (n=100)
Liczba obserwacji o danej wartości
0
500
1000
1500
2000
-2
0
2
4
Wartość obserwacji

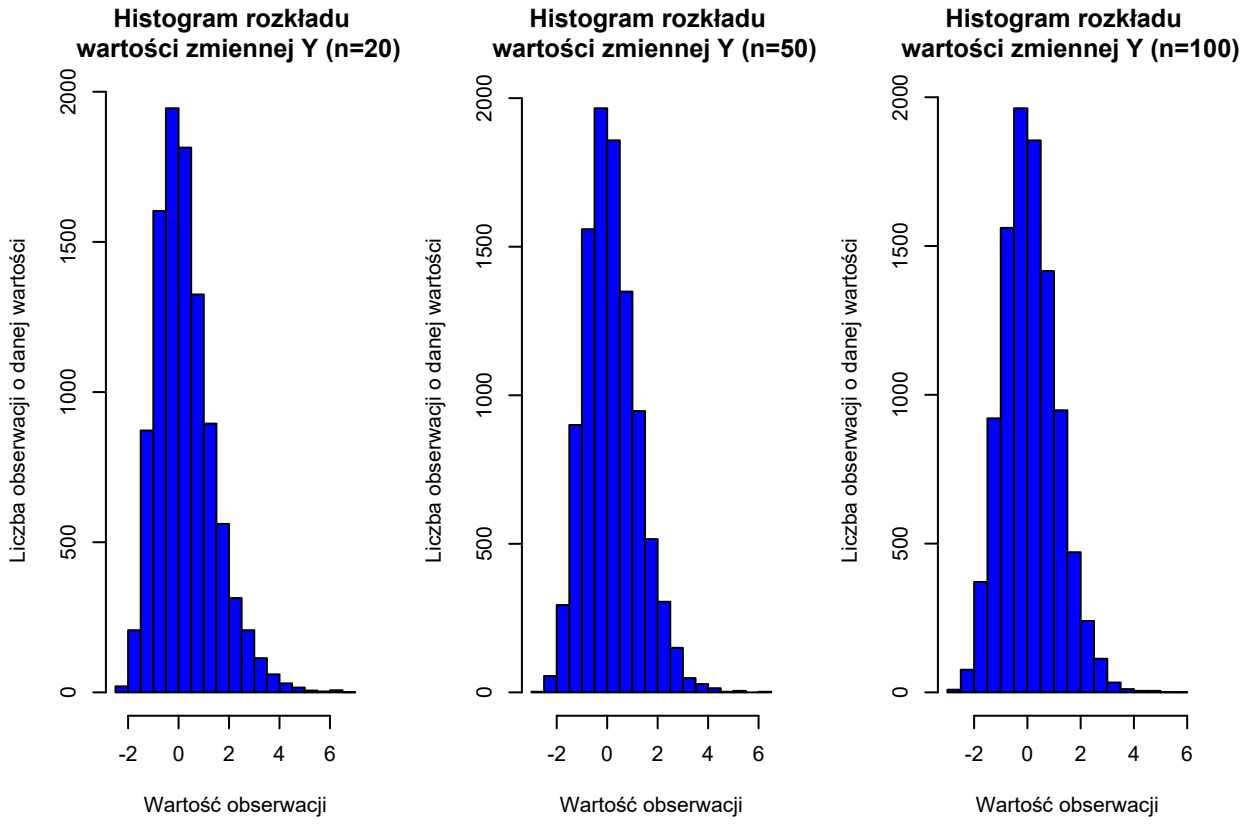

Wykres kwantylowo-kwantylowy dla zmiennej Y dla parametrów kształtu 1 oraz 1
kwantyle doświadczalne
6
3
0
-3
-4
-2
0
2
4
kwantyle teoretyczne
rozmiar próby
20
50
100


Histogram rozkładu wartości zmiennej Y (n=20)
Liczba obserwacji o danej wartości
0
500
1000
1500
2000
-2
0
2
4
6
8
Wartość obserwacji
Histogram rozkładu wartości zmiennej Y (n=50)
Liczba obserwacji o danej wartości
0
500
1000
1500
2000
-2
0
2
4
6
Wartość obserwacji
Histogram rozkładu wartości zmiennej Y (n=100)
Liczba obserwacji o danej wartości
0
500
1000
1500
-2
0
2
4
Wartość obserwacji


Wykres kwantylowo-kwantylowy dla zmiennej Y dla parametrów kształtu 0,5 oraz 1
kwantyle doświadczalne
9
6
3
0
-3
-4
-2
0
2
4
kwantyle teoretyczne
rozmiar próby
20
50
100


Histogram rozkładu wartości zmiennej Y (n=20)
Liczba obserwacji o danej wartości
0
500
1000
1500
2000
-2
0
2
4
6
8
Wartość obserwacji
Histogram rozkładu wartości zmiennej Y (n=50)
Liczba obserwacji o danej wartości
0
500
1000
1500
-2
0
2
4
6
Wartość obserwacji
Histogram rozkładu wartości zmiennej Y (n=100)
Liczba obserwacji o danej wartości
0
500
1000
1500
-2
0
2
4
Wartość obserwacji


Wykres kwantylowo-kwantylowy dla zmiennej Y dla parametrów kształtu 1 oraz 1
kwantyle doświadczalne
5.0
2.5
0.0
-2.5
-4
-2
0
2
4
kwantyle teoretyczne
rozmiar próby
20
50
100



Na podstawie powyższych wykresów oraz histogramów możemy wywnioskować, że rozkład zmiennej losowej Y jest zbliżony do normalnego. Kwantyle doświadczalne w dużym stopniu przylegają do osi wyznaczanej przez kwantyle teoretyczne. Te drugie pochodzą ze standardowego rozkładu normalnego. Dlaczego? Wiemy, że informacja Fishera w naszym przypadku jest skończona. Ciąg estymatorów największej wiarogodności spełnia odpowiednie równanie:

$$\frac{dL(\theta)}{d\theta} = 0$$

Zatem zachodzi:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, \frac{1}{I(\theta)})$$

W naszym przypadku mamy $N(0, 1)$. Histogramy przedstawiające rozkład zmiennej losowej Y przypominają rozkład zmiennej losowej o rozkładzie normalnym. Liczba klas równa jest w tym przypadku 20, dlatego, że pozwala szczegółowiej ukazać rozkład zmiennej losowej Y.

Efektywność różnych estymatorów wartości oczekiwanej dla rozkładu Laplace'a

W ostatniej części wygenerujemy kolejno n obserwacji ($n \in \{20, 50, 100\}$) z rozkładu $L(\theta, \sigma)$ (Laplace'a), dla:

(a) $\theta = 1, \sigma = 1$
(b) $\theta = 4, \sigma = 1$
(c) $\theta = 1, \sigma = 2$

Na podstawie tych danych obliczać będziemy wartość estymatora parametru θ postaci:

(i) $\hat{\theta_1} = \bar{X} = (1/n) \sum_{i=1}^{n} X_i$,

(ii) $\hat{\theta_2} = Me\{X_1, \ldots, X_n\}$,

(iii) $\hat{\theta_3} = \sum_{i=1}^{n} w_i X_i,\ \sum_{i=1}^{n} w_i = 1,\ 0 \leq w_i \leq 1,\ i = 1, \ldots, n$ (wagi własne na wykresie pudełkowym, w przypadku naszej analizy wynoszą one n/2 razy 1/2n dla pierwszej połowy wartości wektora wag oraz n/2 razy 3/2n dla kolejnych wartości wektora wag),

(iv) $\hat{\theta_4} = \sum_{i=1}^{n} w_i X_{i:n}$, gdzie $X_{i:n}, \ldots, X_{n:n}$ są uporządkowanymi obserwacjami $X_1, \ldots, X_n$,

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

gdzie ϕ jest gęstością a Φ dystrybuantą standardowego rozkładu normalnego N(0,1) (gęst. i dystr. na wykresie pudełkowym).

Powtórzymy dane doświadczenie 10 000 razy. Na jego podstawie oszacujemy wariancje, błąd średniokwadratowy oraz obciążenie każdego z estymatorów, a także zwizualizujemy wyniki naszego badania i zastanowimy się nad ich rezultatem. Przedyskutujemy, który estymator jest optymalny i odniesiemy się do zadania 1 z listy 1.

n = 20

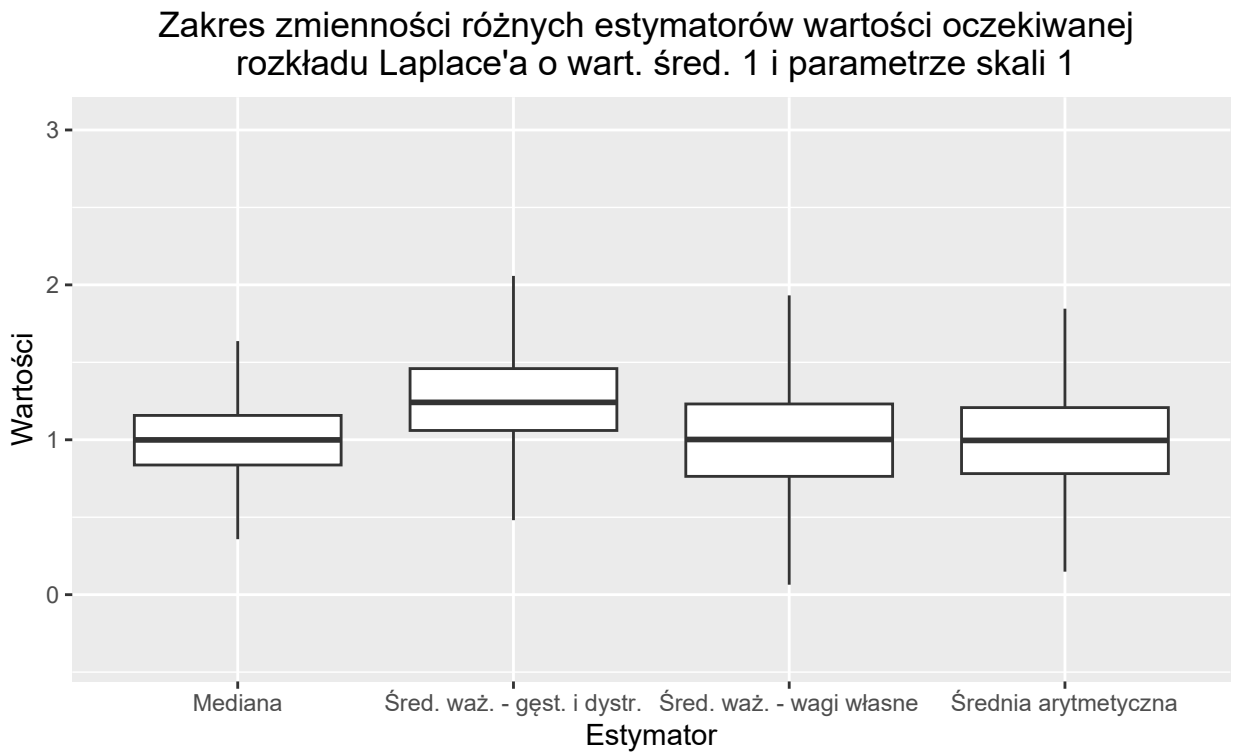


Table 2: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości średniej rozkładu Laplace'a o wart. śred. 1 i parametrze skali 1

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.1006571	0.1006659	-0.0043444
Mediana	0.0661057	0.0660994	-0.0005447
Wagi własne	0.1241785	0.1241716	-0.0023385
Gęts. i dystr.	0.0887565	0.1625412	0.2716498

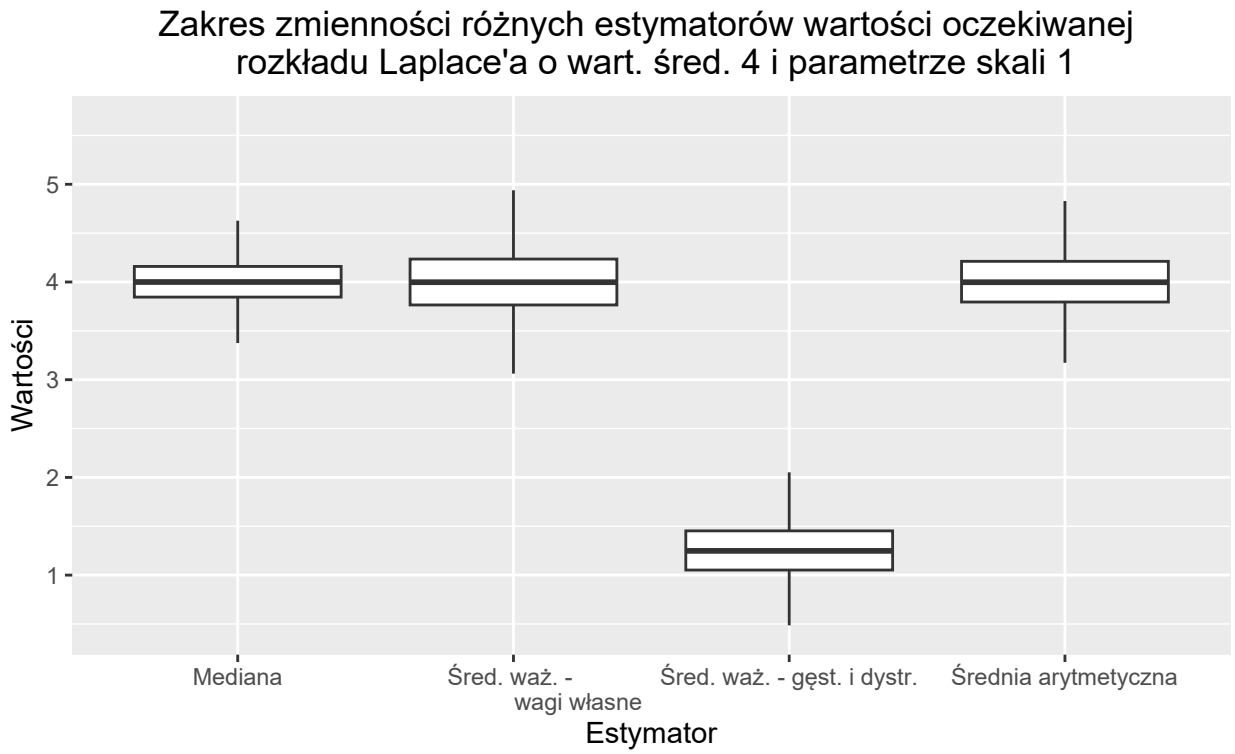


Table 3: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości średniej rozkładu Laplace'a o wart. śred. 4 i parametrze skali 1

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.1014379	0.1014278	0.0000960
Mediana	0.0669587	0.0669523	0.0005219
Wagi własne	0.1277677	0.1277564	-0.0012174
Gęts. i dystr.	0.0891415	7.5508393	-2.7316125

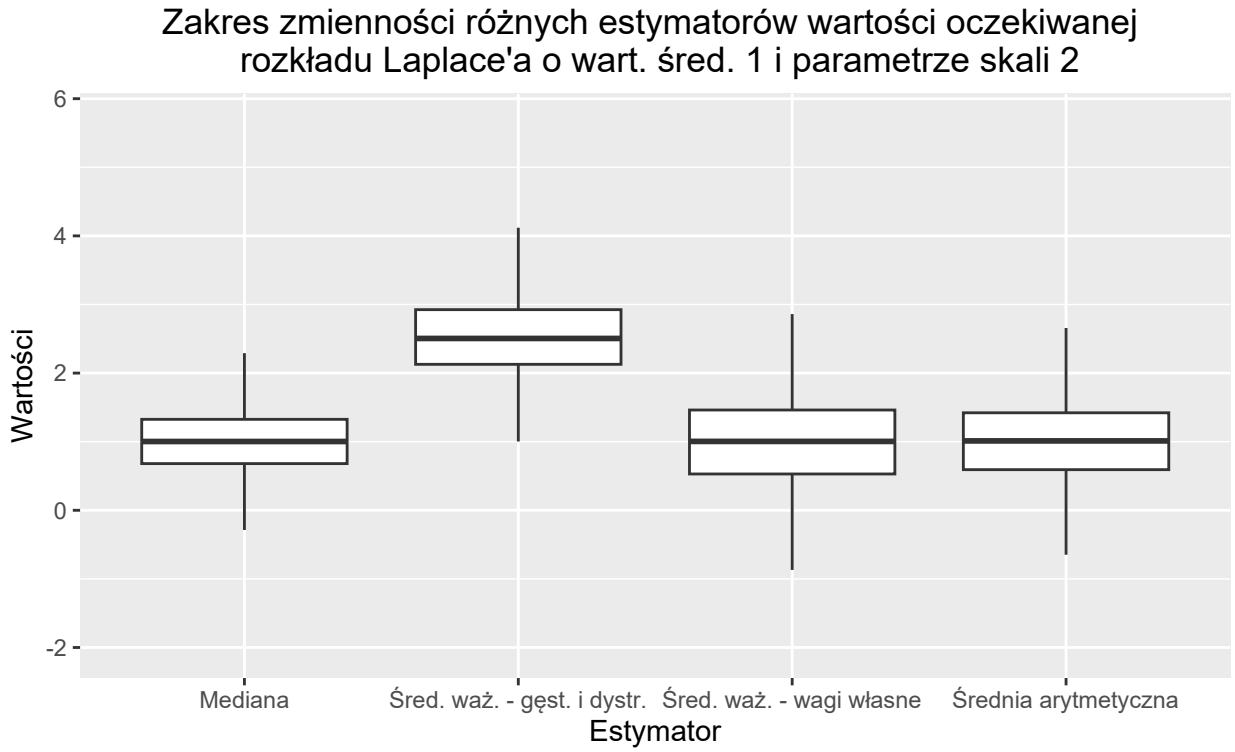


Table 4: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości średniej rozkładu Laplace'a o wart. śred. 1 i parametrze skali 2

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.4060331	0.4060229	0.0055170
Mediana	0.2657417	0.2657226	0.0027259
Wagi własne	0.5072381	0.5071881	-0.0008755
Gęts. i dystr.	0.3689960	2.7886405	1.5555325

$\mathbf{n = 50}$

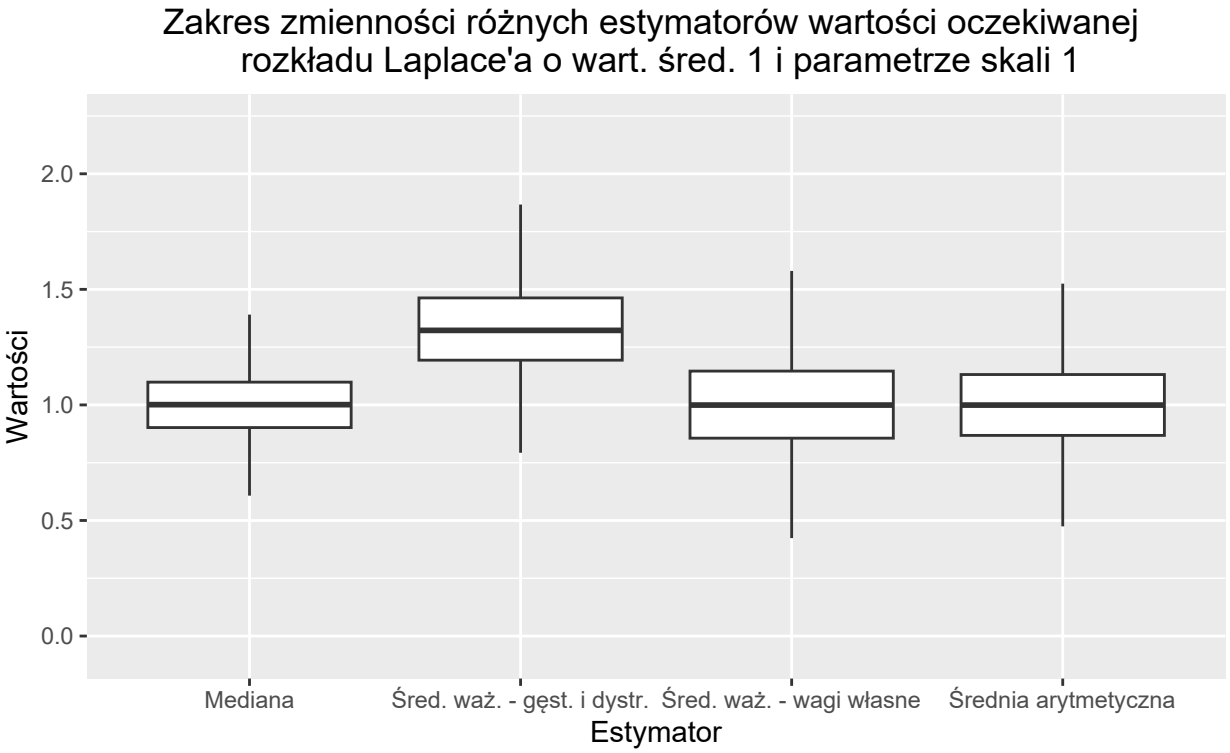


Table 5: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości średniej rozkładu Laplace'a o wart. śred. 1 i parametrze skali 1

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.0395215	0.0395189	0.0011264
Mediana	0.0243210	0.0243188	0.0004153
Wagi własne	0.0488972	0.0488967	0.0020938
Gęts. i dystr.	0.0391284	0.1508017	0.3341815

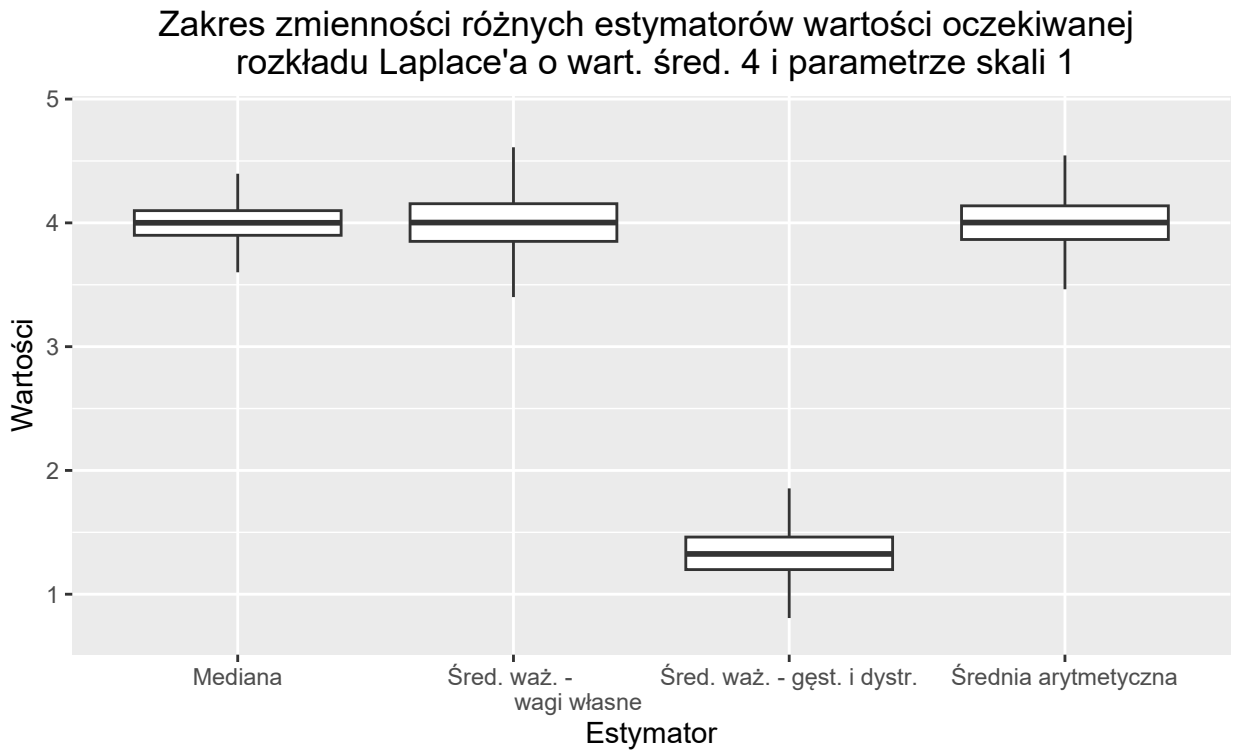


Table 6: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości średniej rozkładu Laplace'a o wart. śred. 4 i parametrze skali 1

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.0399913	0.0399926	0.0023074
Mediana	0.0253184	0.0253158	-0.0001240
Wagi własne	0.0500999	0.0501054	0.0032465
Gęts. i dystr.	0.0391885	7.1335639	-2.6635276

Table 7: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości średniej rozkładu Laplace'a o wart. śred. 1 i parametrze skali 2

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.1561789	0.1561672	0.0019957
Mediana	0.0943951	0.0944014	0.0039729
Wagi własne	0.1956953	0.1956761	-0.0005645
Gęts. i dystr.	0.1561159	2.9306649	1.6657024

$\mathbf{n = 100}$

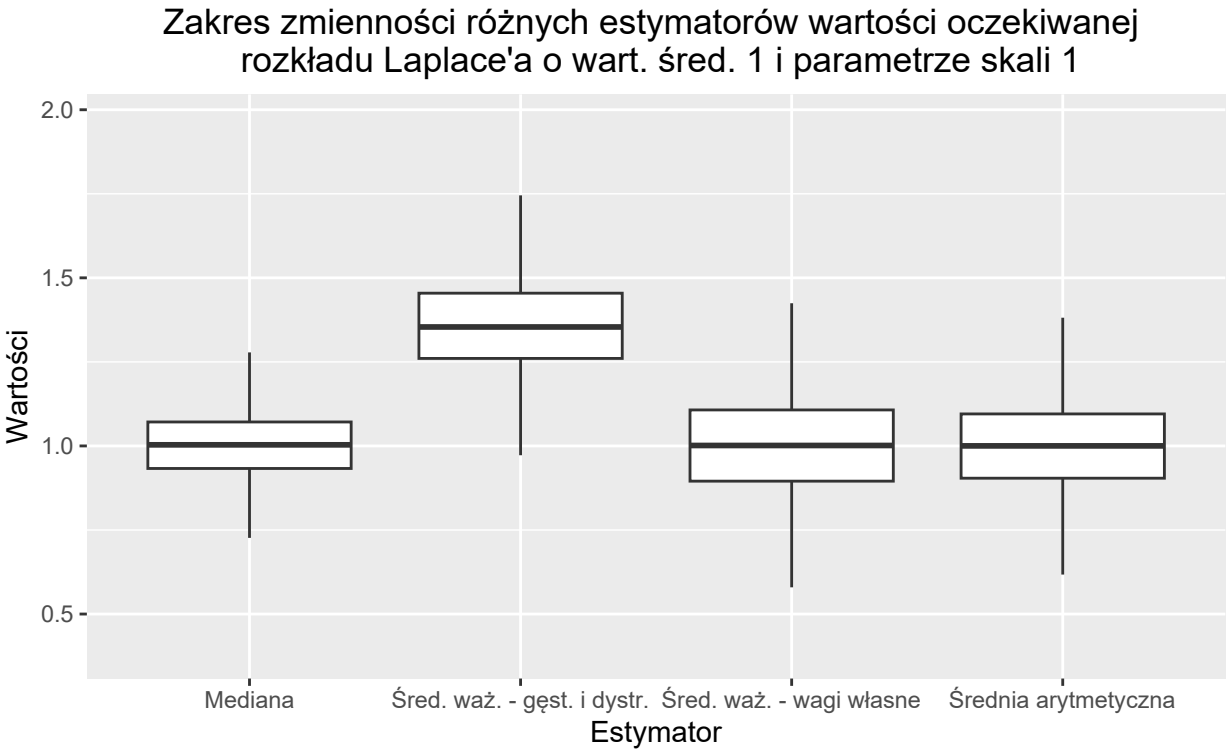


Table 8: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości średniej rozkładu Laplace'a o wart. śred. 1 i parametrze skali 1

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.0196924	0.0196911	0.0008198
Mediana	0.0113851	0.0113881	0.0020448
Wagi własne	0.0250322	0.0250324	0.0016567
Gęts. i dystr.	0.0206501	0.1503413	0.3601295

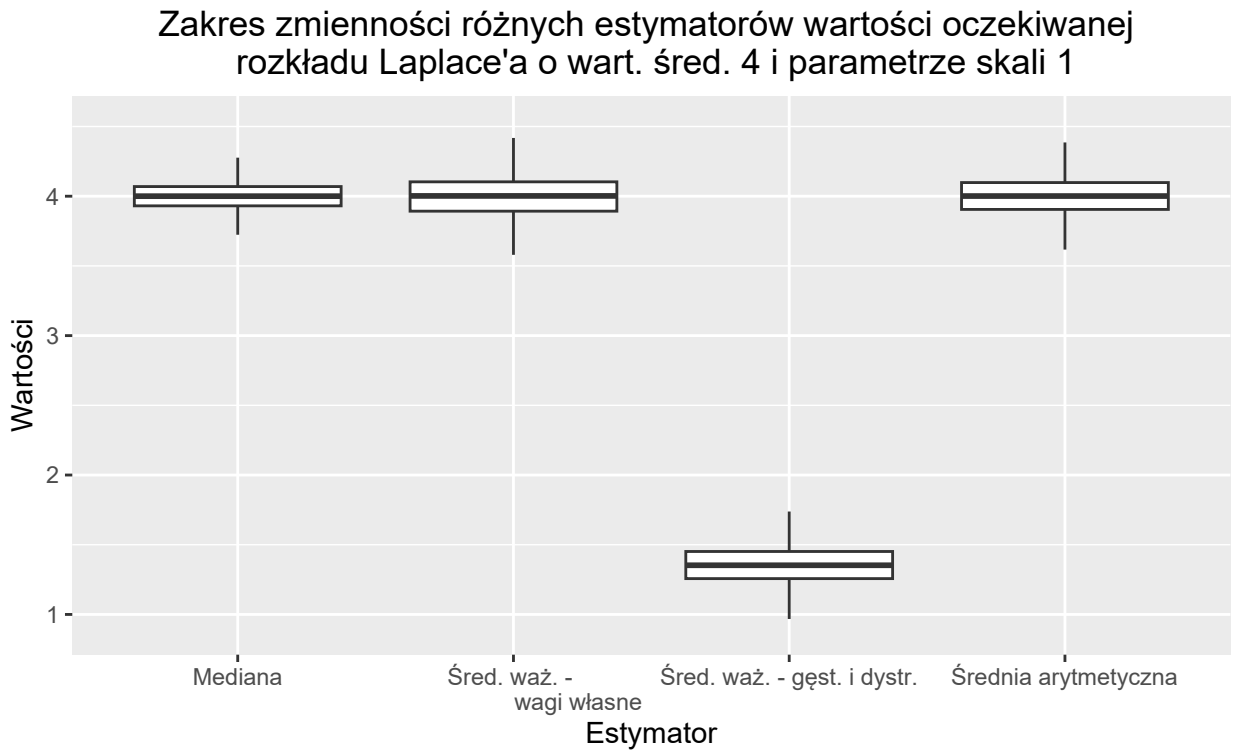


Table 9: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości średniej rozkładu Laplace'a o wart. śred. 4 i parametrze skali 1

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.0200717	0.0200701	0.0006754
Mediana	0.0115112	0.0115101	-0.0003148
Wagi własne	0.0247312	0.0247288	0.0003709
Gęts. i dystr.	0.0205516	7.0019100	-2.6422264

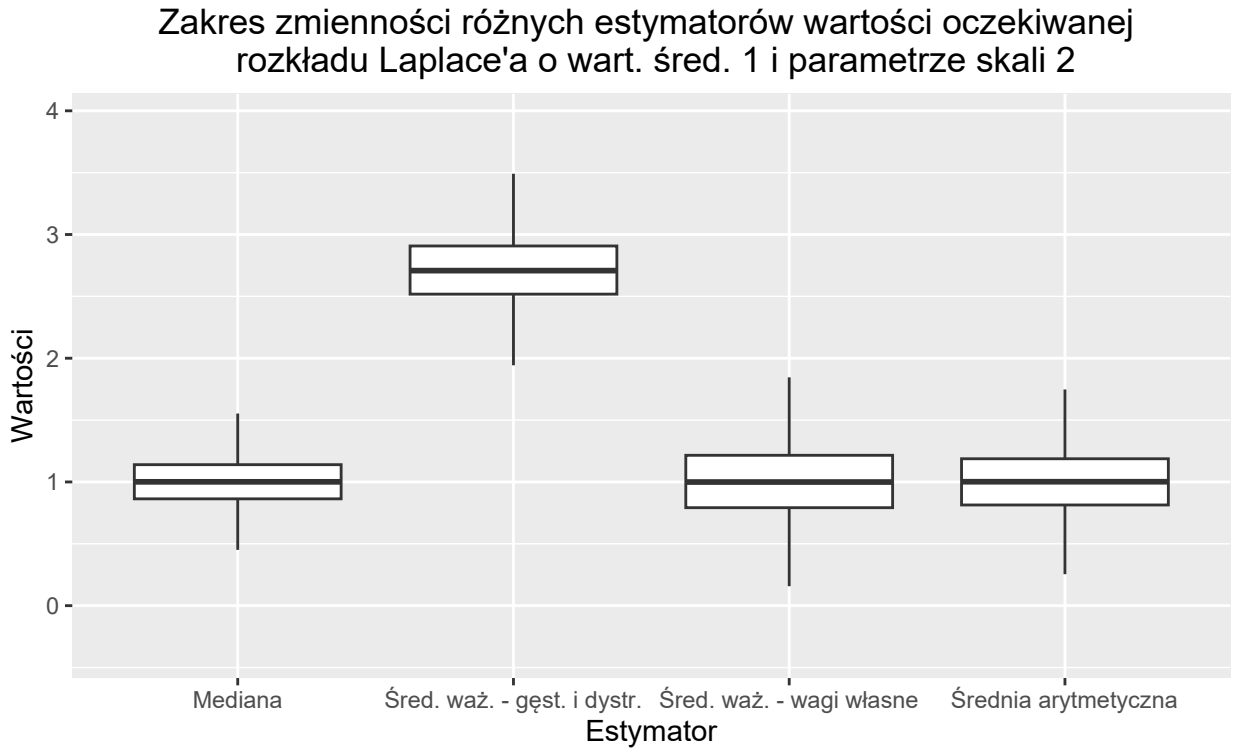


Table 10: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości średniej rozkładu Laplace'a o wart. śred. 1 i parametrze skali 2

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.0797057	0.0796978	-0.0000662
Mediana	0.0458106	0.0458065	0.0007110
Wagi własne	0.0990174	0.0990078	0.0005495
Gęts. i dystr.	0.0830560	3.0354427	1.7182535

Analizując powyższe wyniki możemy stwierdzić, że najlepszym estymatorem w naszym przypadku jest mediana. Osiąga ona generalnie najmniejsze wartości poszczególnych statystyk próby. Ma też nieduży rozrzut względem innych estymatorów. Warto zauważyć, że ENW dla rozkładu Laplace'a jest mediana a dla rozkładu normalnego średnia. W przypadku zad 1 z listy 1 to ona była optymalnym estymatorem.