Charakterystyka estymatorów poprzez różne statystyki oraz numeryczne wyznaczanie estymatorów największej wiarygodności

Antoni Bieniasz

2023-10-25

Wstęp

Raport ten skupia się na analizie estymatorów. Na początku zbadane zostaną różne estymatory wartości oczekiwanej rozkładu normalnego, przede wszystkim pod względem ich wariancji, błędu średniokwadratowego oraz obciążenia. Następnie poprzez numeryczną metodę Newtona szacować będziemy wartość estymatora największej wiarygodności (będziemy pisać dalej w skrócie ENW) dla rozkładów, dla których jego dokładne wyliczenie jest trudne.

zad 1

W pierwszej kolejności wygenerujemy kolejno 50 obserwacji z rozkładu $N(\theta, \sigma^2)$, 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 1/100 dla pierwszych 25 wartości wektora wag oraz 3/100 dla kolejnych 25 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 gęstoscią 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.

W analizie poniższych wykresów oraz tabel będziemy skrótowo oznaczać estymatory takimi liczbami jakimi były one oznaczone we wstępie do zadania 1, tzn.: średnia arytmetyczna - 1, mediana - 2, wagi własne - 3, estymator obliczony przez gęstość oraz dystrybuantę rozkładu normalnego standardowego - 4.

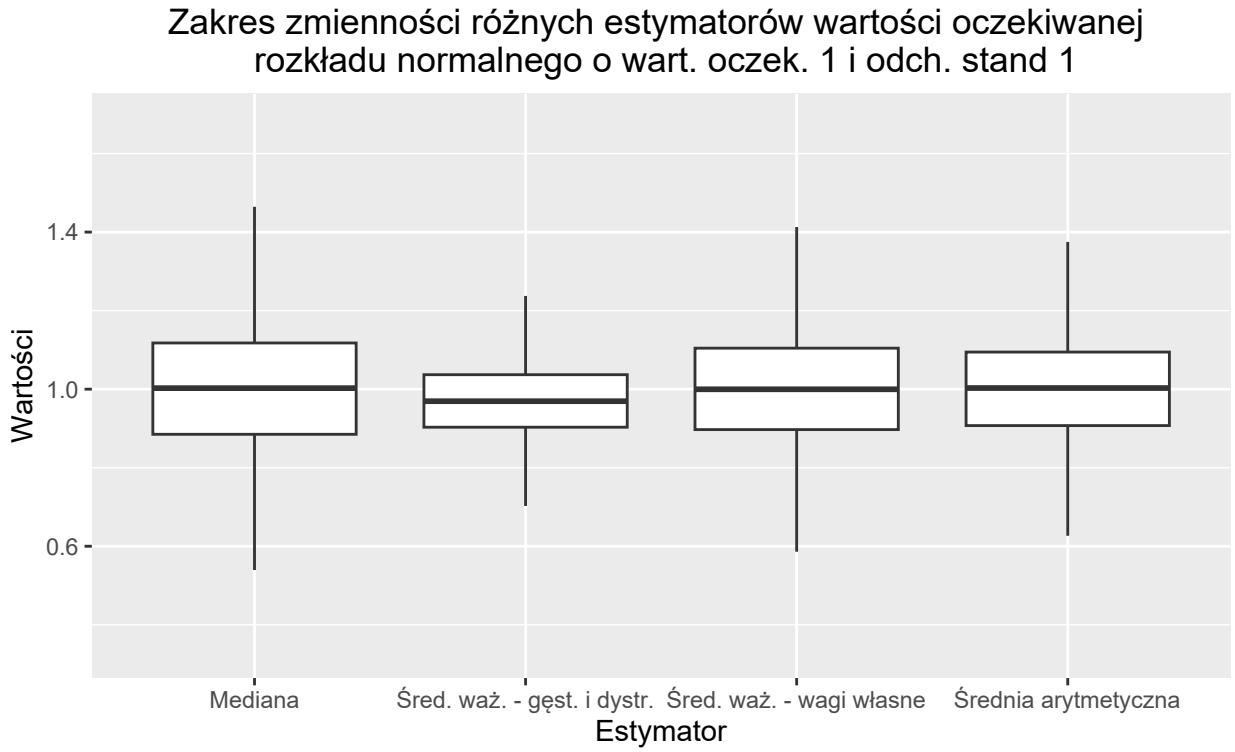

Table 1: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości oczekiwanej rozkładu normalnego o wart. oczek. 1 i odch. stand 1

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.0199310	0.0199306	0.0012630
Mediana	0.0298798	0.0298769	0.0003412
Wagi własne	0.0242825	0.0242804	0.0005332
Gęts. i dystr.	0.0097092	0.0105718	-0.0293863

Z powyższego wykresu i tabeli możemy wywnioskować, że mediana estymatora wartości oczekiwanej dla estymatorów 1, 2, 3 wyliczanych na podstawie 50 elementowej próby z rozkładu normalnego $N(1,1)$, była zbliżona do 1. Jedynie dla estymatora 4 odbiega ona od tej wartości, wynosi ona około 0,96 - 0.97. Można się zastanowić, czy jest to właściwy estymator, skoro dla tak wielu doświadczeń (10 000) daje on wynik względnie istotnie różny. W przypadku wykresu pudełkowego pominęliśmy obserwacje odstające, ponieważ skupiamy się tutaj na ogólnej analizie wartości estymatorów. Najmniejszy rozstęp ćwiartkowy obserwujemy dla estymatora 4, jednak w przypadku jego odbiegającej mediany jest to niewielka poprawa jego skuteczności. Następnie rozstęp ten zwiększa się kolejno dla estymatorów 1, 3, 2. Pod względem analizy wariancji, błędu średniokwadratowego oraz obciążenia poza obciążeniem najlepiej wypada estymator 4, jednak ponownie możemy przypomnieć sobie tutaj o jego odbiegającej medianie, co sugeruje nieco inny zakres otrzymanych wartości. Średnia arytmetyczna (1), osiąga niższe wartości otrzymanych statystyk niż mediana (2) i wagi własne (3), co przy jej medianie zbliżonej do 1 pozwala przypuszczać, że może być ona w tym przypadku najlepszym estymatorem wartości oczekiwanej.

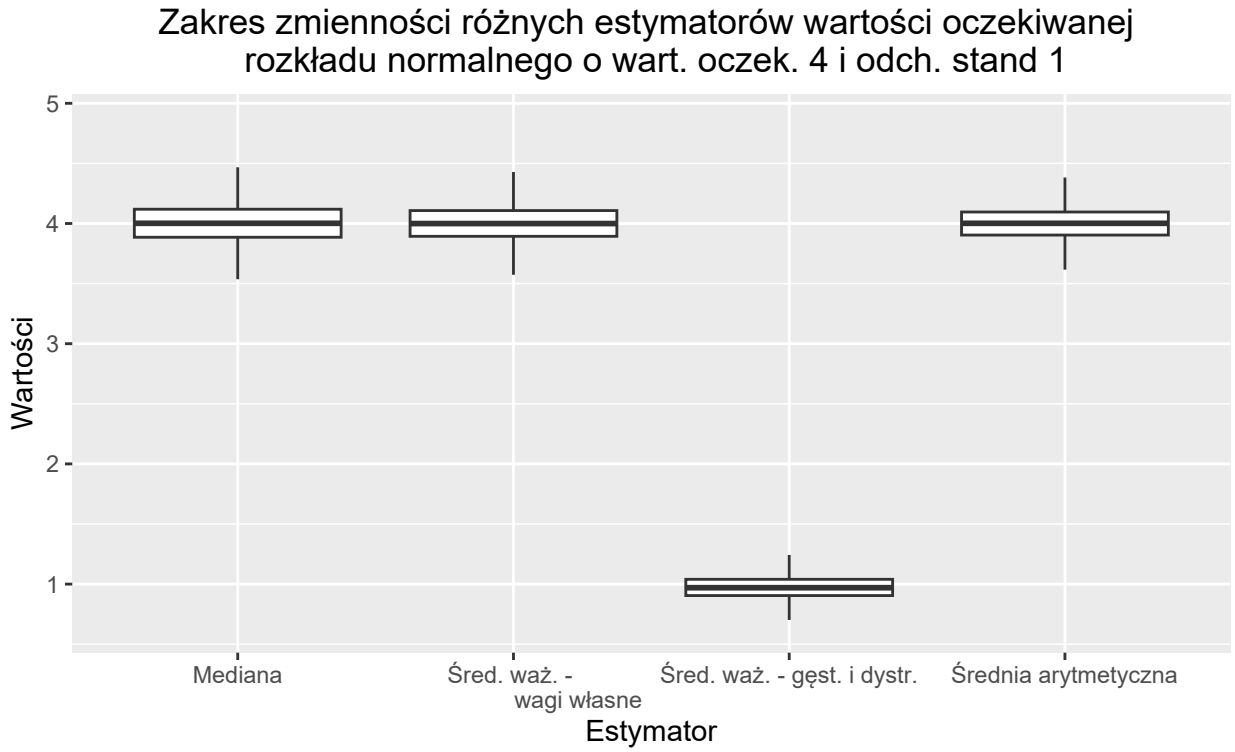

Table 2: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości oczekiwanej rozkładu normalnego o wart. oczek. 4 i odch. stand 1

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.0201061	0.0201041	-0.0001533
Mediana	0.0306121	0.0306097	0.0008062
Wagi własne	0.0253783	0.0253757	0.0001635
Gęts. i dystr.	0.0098822	9.1723076	-3.0269500

W przypadku tego wykresu i tabeli koncentrujących się na próbach z rozkładu normalnego $N(4,1)$, możemy stwierdzić, ze estymator (4) jest niewłaściwy, ponieważ jego mediana, osiągane wartości są z kompletnie innego zakresu, co znajduje potwierdzenie w względnie dużym obciążeniu oraz błędzie średniokwadratowym. Z pozostałych estymatorów Średnia arytmetyczna (1) wypada lepiej niż estymator stworzony poprzez własne wagi (3), który jest nieco lepszy od mediany (2). Widzimy to poprzez analizę rozstępu ćwiartkowego oraz analizę statystyk z tabeli (1) ma ich niższe wartości niż (3), który z kolei ma je niższe niż (2). Z wykresu pudełkowego możemy odczytać, że dla każdego z tych trzech estymatorów mediana wartości jest podobna i skupia się wokół wartości oczekiwanej 4.

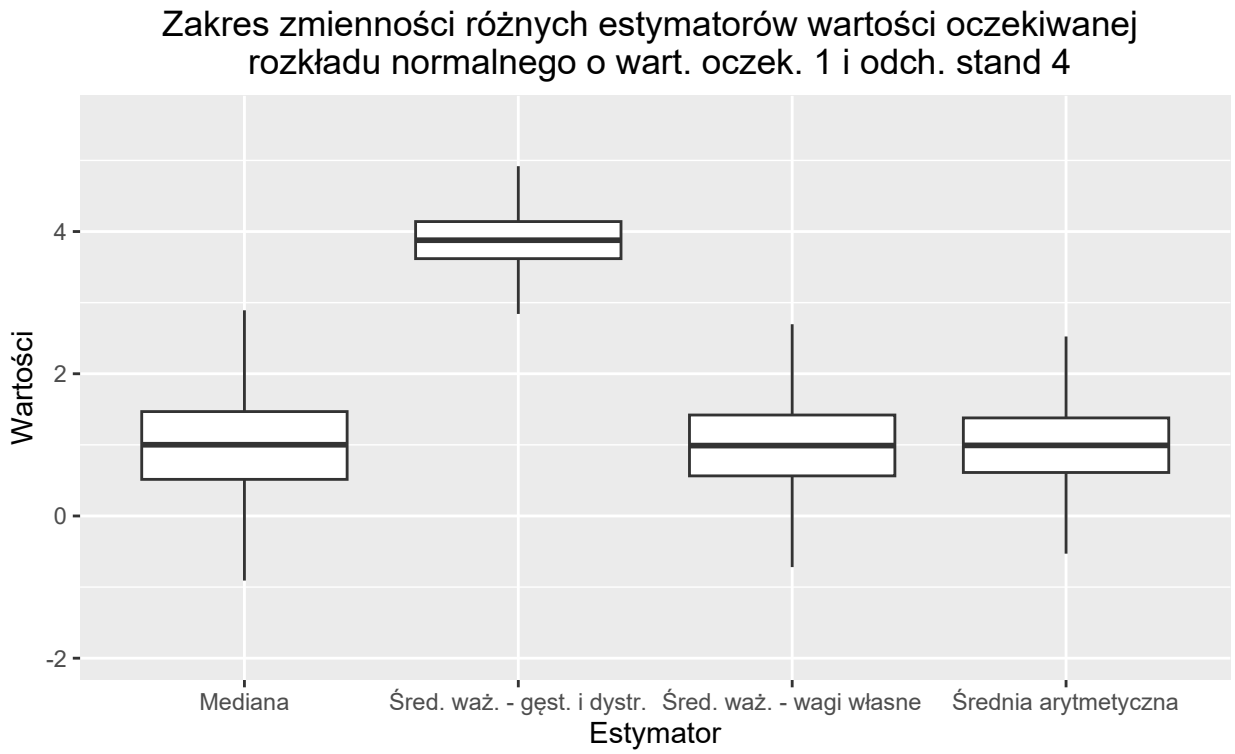

Table 3: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów wartości oczekiwanej rozkładu normalnego o wart. oczek. 1 i odch. stand 4

Estymator	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Średnia arytmetyczna	0.3266279	0.3266198	-0.0049582
Mediana	0.4980593	0.4980211	-0.0034093
Wagi własne	0.4089536	0.4089690	-0.0075039
Gęts. i dystr.	0.1524289	8.4493491	2.8804401

Ponownie analizując ostatni wykres pudełkowy oraz tabele dla rozkładu $N(1,4)$ otrzymujemy brak dopasowania estymatora (4). Najmniejszy rozstęp ćwiartkowy uzyskujemy dla średniej arytmetycznej, potem, dla wag własnych oraz mediany. Wartość mediany wartości estymatorów koncentruje się wokół 1 poza estymatorem (4). Analizując wariancje, błąd średniokwadratowy oraz obciążanie spośród estymatorów (1), (2), (3) ogólnie najmniejsze wartości osiąga estymator 1.

Podsumowując średnia arytmetyczna wydaje się być najlepszym estymatorem wartości oczekiwanej dla rozkładu normalnego dla także dla zmiennych wartości wartości oczekiwanej oraz wariancji. Konstrukcja estymatora poprzez wybór własnych wag może się okazać lepsza niż estymowanie przez medianę. Estymator, którego wagi są złożeniem gęstości i odwrotnej dystrybuanty standardowego rozkładu normalnego nie estymuje w żaden sposób wartości oczekiwanej szukanego rozkładu normalnego.

zad 5

Będziemy teraz chcieli wygenerować 50 obserwacji z rozkładu logistycznego $L(\theta, \sigma)$ z parametrem przesunięcia θ i skali σ dla:

(a) $\theta = 1, \sigma = 1$

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

Na podstawie wygenerowanej próby szacować będziemy wartość estymatora największej wiarygodności (ENW) parametru θ - obliczenie go z równania opisującego pierwszą pochodną funkcji logarytmu wiarogodności jest praktycznie niemożliwe. Zastanowimy się nad wyborem punktu początkowego oraz liczbą kroków w algorytmie - zastosujemy Metodę Newtona. Powtórzymy dane doświadczenie 10 000 razy. Na jego podstawie szacować będziemy wariancję, błąd średniokwadratowy oraz obciążenie estymatora. Przeanalizujemy uzyskane wyniki.

Jako punkt początkowy dla Metody Newtona wybieramy średnią, z każdej 50 elementowej próby - rozkład logistyczny jest podobny do normalnego, stąd średnia powinna być blisko szukanego maksimum. Nie ustalamy dokładnie liczby kroków w algorytmie, ale za to przerywamy algorytm wtedy, gdy pochodna logarytmu funkcji wiarygodności dla rozkładu logistycznego będzie mniejsza równa 10^(-6) lub liczba kroków osiągnie 10. Nie musimy sprawdzać warunku, czy osiągnięte może być maksimum - druga pochodna funkcji logarytmu wiarogodności jest zawsze ujemna.

Table 4: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów estymatora największej wiarygodności dla rozkładu logistycznego

Rodzaj_rozkładu_logistycznego	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Przesunięcie: 1, skala: 1	0.0596422	0.0596364	-0.0003535
Przesunięcie: 4, skala: 1	0.0606000	0.0605997	0.0024002
Przesunięcie: 1, skala: 2	0.2431004	0.2430761	0.0000472

W celu oszacowania (ENW) parametru θ wzięliśmy średnią z 10000 symulacji. Otrzymaliśmy dla kolejno rozkładów (a), (b), (c): 0.9996, 4.0024, 1. Analizując powyższą tabelę możemy stwierdzić, że dla rozkładów z (a) i (b) uzyskaliśmy znacznie niższą wariancję oraz błąd średniokwadratowy niż dla rozkładu (c). Rozkład (c) ma parametr skali $\sigma = 2$, co oznacza, że więcej obserwacji odstaje, są większe "ogony" stąd większe wartości tych statystyk. Nie potrzebowaliśmy wielu kroków (średnio 1.9389), więc metoda ta okazała się być skuteczna.

zad 6

Będziemy teraz chcieli wygenerować 50 obserwacji z rozkładu Cauchy'ego $C(\theta, \sigma)$ z parametrem przesunięcia θ i skali σ dla:

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

Na podstawie wygenerowanej próby szacować będziemy wartość estymatora największej wiarygodności (ENW) parametru θ. Podobnie osiągnięcie jego wartości przez równanie przyrównujące pierwszą pochodną funkcji logarytmu wiarogodności do zera, tak samo jak w przypadku rozkładu logistycznego jest bezcelowe. Zastanowimy się nad wyborem punktu początkowego oraz liczbą kroków w algorytmie - zastosujemy Metodę Newtona. Powtórzymy dane doświadczenie 10 000 razy. Na jego podstawie szacować będziemy wariancję, błąd średniokwadratowy oraz obciążenie estymatora. Przeanalizujemy uzyskane wyniki.

Jako punkt początkowy dla Metody Newtona wybieramy medianę, z każdej 50 elementowej próby - rozkład Cauchy'ego ma nieokreśloną wartość oczekiwaną oraz jest podobny do rozkładu Laplace'a, dla którego ENW jest mediana. Tak jak poprzednio nie ustalamy dokładnie liczby kroków w algorytmie, ale za to przerywamy algorytm wtedy, gdy pochodna logarytmu funkcji wiarygodności dla rozkładu Cauchy'ego będzie mniejsza równa 10^(-6) albo liczba kroków osiągnie 10. Jednak druga pochodna funkcji logarytmu wiarogodności nie zawsze jest ujemna - przy estymacji będziemy odrzucać te wyniki, które będą dodatnie.

Table 5: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów estymatora największej wiarygodności dla rozkładu Cauchy'ego

Rodzaj_rozkładu_Cauchyego	Wariancja	Błąd_Średniokwadratowy	Obciążenie	Obserwacje_zbiegające
Przesunięcie: 1, skala: 1	0.0412384	0.0412344	0.0004528	10000
Przesunięcie: 4, skala: 1	0.0436208	0.0436186	0.0014556	10000
Przesunięcie: 1, skala: 2	0.1713471	0.1713322	0.0014815	9999

Z powyższej tabeli możemy odczytać, że dla (a) oraz (b) oszacowano podobną wariancję oraz błąd średniokwadratowy. Pozostanie przy przesunięciu 1, ale zwiększenie skali do 2 (c) spowodowało największe zmiany - czterokrotnie większa wariancja oraz błąd średniokwadratowy niż w przypadku (a) i (b). Można to ponownie tłumaczyć większymi "ogonami" takiego rozkładu. Zdecydowanie najmniejsze obciążenie co do modułu osiągnięto dla rozkładu (b).

zad 7

Powtórzymy dany eksperyment numeryczny dla $n = 20, 100$. Porównamy uzyskane wyniki z poprzednimi.

Logistyczny $n = 20, 100$

Table 6: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów estymatora największej wiarygodności dla rozkładu logistycznego (n=20)

Rodzaj_rozkładu_logistycznego	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Przesunięcie: 1, skala: 1	0.1481327	0.1481202	0.0015172
Przesunięcie: 4, skala: 1	0.1561538	0.1561383	-0.0003520
Przesunięcie: 1, skala: 2	0.6109460	0.6109360	-0.0071480

Table 7: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów estymatora największej wiarygodności dla rozkładu logistycznego (n=100)

Rodzaj_rozkładu_logistycznego	Wariancja	Błąd_Średniokwadratowy	Obciążenie
Przesunięcie: 1, skala: 1	0.0299379	0.0299531	-0.0042612
Przesunięcie: 4, skala: 1	0.0301839	0.0301930	-0.0034745
Przesunięcie: 1, skala: 2	0.1160422	0.1160824	-0.0072014

Z powyższych tabel można odczytać, że zwiększenie rozmiaru próby do 100 i zmniejszenie do 20, odpowiednio zwiększa oraz zmniejsza wartości statystyk w tabeli. Analizując dokładniej wartości można przypuszczać, że jest to wzrost wprost proporcjonalny.

Table 8: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów estymatora największej wiarygodności dla rozkładu Cauchy'ego (n=20)

Rodzaj_rozkładu_Cauchyego	Wariancja	Błąd_Średniokwadratowy	Obciążenie	Obserwacje_zbiegające
Przesunięcie: 1, skala: 1	0.1151131	0.1151107	-0.0030257	9989
Przesunięcie: 4, skala: 1	0.1118864	0.1118778	0.0016218	9982
Przesunięcie: 1, skala: 2	0.4568321	0.4567867	-0.0005933	9982

Table 9: Szacowana wariancja, błąd średniokwadratowy, oraz obciążanie każdego z estymatorów estymatora największej wiarygodności dla rozkładu Cauchy'ego (n=100)

Rodzaj_rozkładu_Cauchyego	Wariancja	Błąd_Średniokwadratowy	Obciążenie	Obserwacje_zbiegające
Przesunięcie: 1, skala: 1	0.0206245	0.0206233	-0.0009498	10000
Przesunięcie: 4, skala: 1	0.0208556	0.0208544	0.0009365	10000
Przesunięcie: 1, skala: 2	0.0805437	0.0805359	0.0004991	10000

Dla prób 20 oraz 100 elementowych należących do rozkładu Cauchy'ego mamy podobnie jak dla rozkładu logistycznego, odpowiednio wzrost oraz zmniejszenie wartości statystyk: wariancji oraz błędu średniokwadratowego. Przy dokładnym przyjrzeniu się tym wartościom można spostrzec zależność wprost proporcjonalną wprost między rozmiarem próby a wielkością danych statystyk.