

# Statystyka: raport 2

Helena Sękowska-Słoka, nr indeksu 321531

2023-11-21

## SPIS TREŚCI

<b>Estymacja wielkości <math>P(3 \leq X)</math> dla rozkładu dwumianowego <math>b(5, p)</math> metodą największej wiarogodności</b>	<b>2</b>
Wyznaczanie ENW dla $P(3 \leq X)$ . . . . .	2
Wykres z oszacowaniem wariancji, błędu średniokwadratowego oraz obciążenia wyznaczonego estymatora w zależności od parametru $p$ . . . . .	2
<b>Estymacja wielkości <math>P(X = x)</math> dla rozkładu Poissona <math>P(\lambda)</math> metodą największej wiarogodności</b>	<b>4</b>
Wyznaczanie ENW dla $P(X = x)$ . . . . .	4
Wykres z oszacowaniem wariancji, błędu średniokwadratowego oraz obciążenia wyznaczonego estymatora w zależności od parametru $x$ . . . . .	4
<b>Analiza rozkładu zmiennej <math>Y = \sqrt{nI(\hat{\theta})(\hat{\theta} - \theta)}</math> wyznaczonej na podstawie estymacji informacji Fishera <math>I(\theta)</math> dla rozkładu <math>\text{beta}(\theta, 1)</math></b>	<b>7</b>
Wyznaczanie ENW informacji Fishera $I(\theta)$ dla rozkładu $\text{beta}(\theta, 1)$ . . . . .	7
Definiowanie nowej zmiennej $Y = \sqrt{nI(\hat{\theta})(\hat{\theta} - \theta)}$ i analiza jej rozkładu . . . . .	7
<b>Estymacja parametru przesunięcia dla rozkładu Laplace'a. Porównanie z estymacją średniej dla rozkładu normalnego</b>	<b>10</b>
Porównanie estymatorów dla parametru przesunięcia w rozkładzie Laplace'a . . . . .	12

## Estymacja wielkości $P(3 \leq X)$ dla rozkładu dwumianowego $b(5, p)$ metodą największej wiarogodności

### Wyznaczanie ENW dla $P(3 \leq X)$

Wiemy, że estymatorem największej wiarogodności dla parametru  $p$  z rozkładu dwumianowego o liczbie prób  $n = 5$  jest  $\frac{\bar{X}}{5}$ . Zauważmy, że chcąc wyestymować metodą największej wiarogodności wielkość  $P(3 \leq X)$  znany nam na ten moment z wykładu metodami, musimy najpierw przekształcić ją do postaci funkcji  $g(p)$ . Skorzystajmy z tego, że zachodzą następujące równości:

$$P(3 \leq X) = 1 - P(X \in \{0, 1, 2\}) = 1 - \left( \binom{5}{0} p^0 (1-p)^{5-0} + \binom{5}{1} p^1 (1-p)^{5-1} + \binom{5}{2} p^2 (1-p)^{5-2} \right)$$

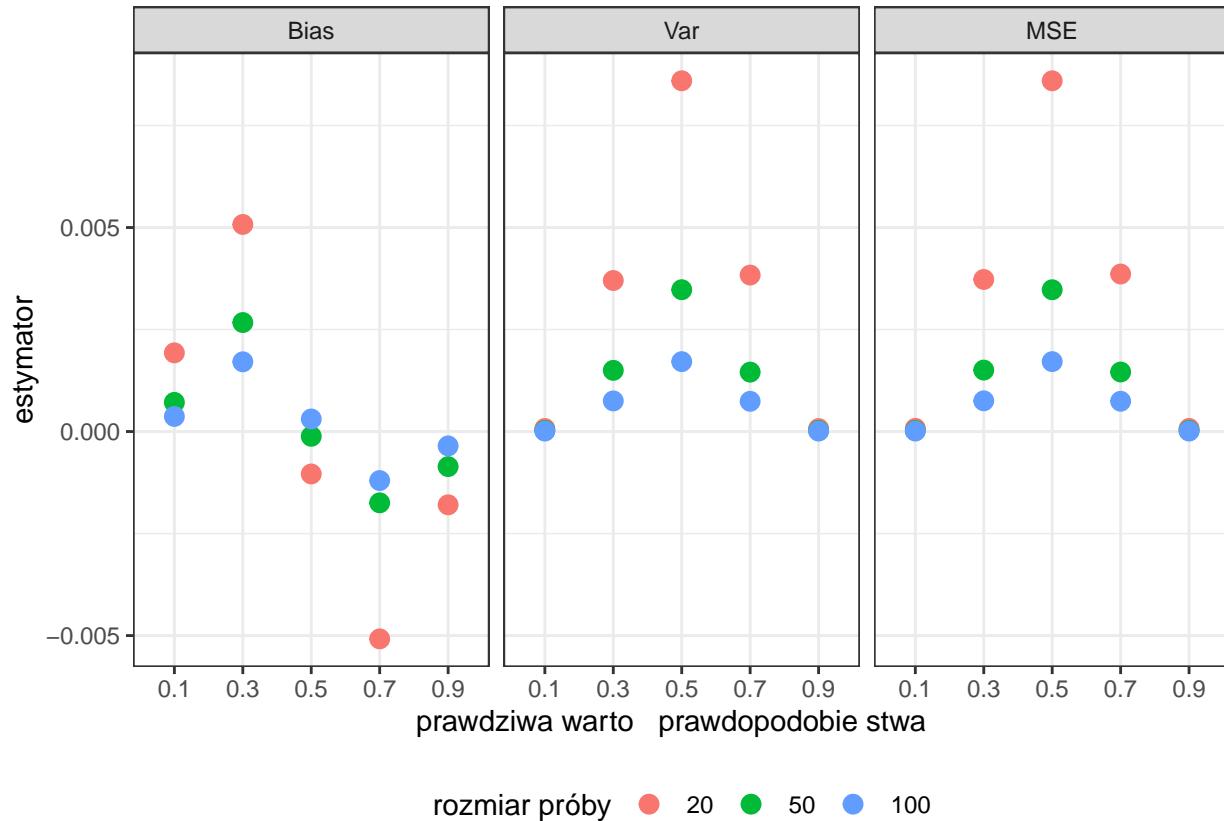
Niech zatem

$$1 - \left( \binom{5}{0} p^0 (1-p)^{5-0} + \binom{5}{1} p^1 (1-p)^{5-1} + \binom{5}{2} p^2 (1-p)^{5-2} \right) = g(p)$$

Dzięki temu przy wyznaczaniu estymatora największej wiarogodności wielkości  $P(3 \leq X)$  możemy skorzystać z twierdzenia 6.1.2 z podręcznika *Introduction to Mathematical Statistics*, którego autorami są Robert Hogg, Joseph McKean i Allen Craig. Mówią ono, że jeśli  $P(3 \leq X) = g(p)$ , zaś  $\frac{\bar{X}}{5}$  jest ENW dla  $p$ , to ENW dla  $P(3 \leq X)$  będzie  $g(\frac{\bar{X}}{5})$ .

Po wyznaczeniu tego estymatora oszacowujemy jego wariancję, błąd średniokwadratowy oraz obciążenie.

### Wykres z oszacowaniem wariancji, błędu średniokwadratowego oraz obciążenia wyznaczonego estymatora w zależności od parametru $p$



Wszystkie wartości na wykresie są stosunkowo małe, rzędu  $10^{-3}$ . Jak można było przypuszczać, największe co do modułu są one dla prób rozmiaru 20.

Obciążenie jest co do modułu symetryczne względem  $p = 0.5$ , przy czym największe jest dla  $p = 0.3$  i  $p = 0.7$ .

Symetryczne wzdłuż tej samej osi są także wariancja i błąd średniokwadratowy, przy czym tutaj największe są wartości dla  $p = 0.5$  i maleją one w kierunku do wartości skrajnych. Różnice między obciążeniami nimi są tym mniejsze, im bliżej jesteśmy skrajnych wartości  $p$  ( $p = 0.1$  i  $p = 0.9$ ).

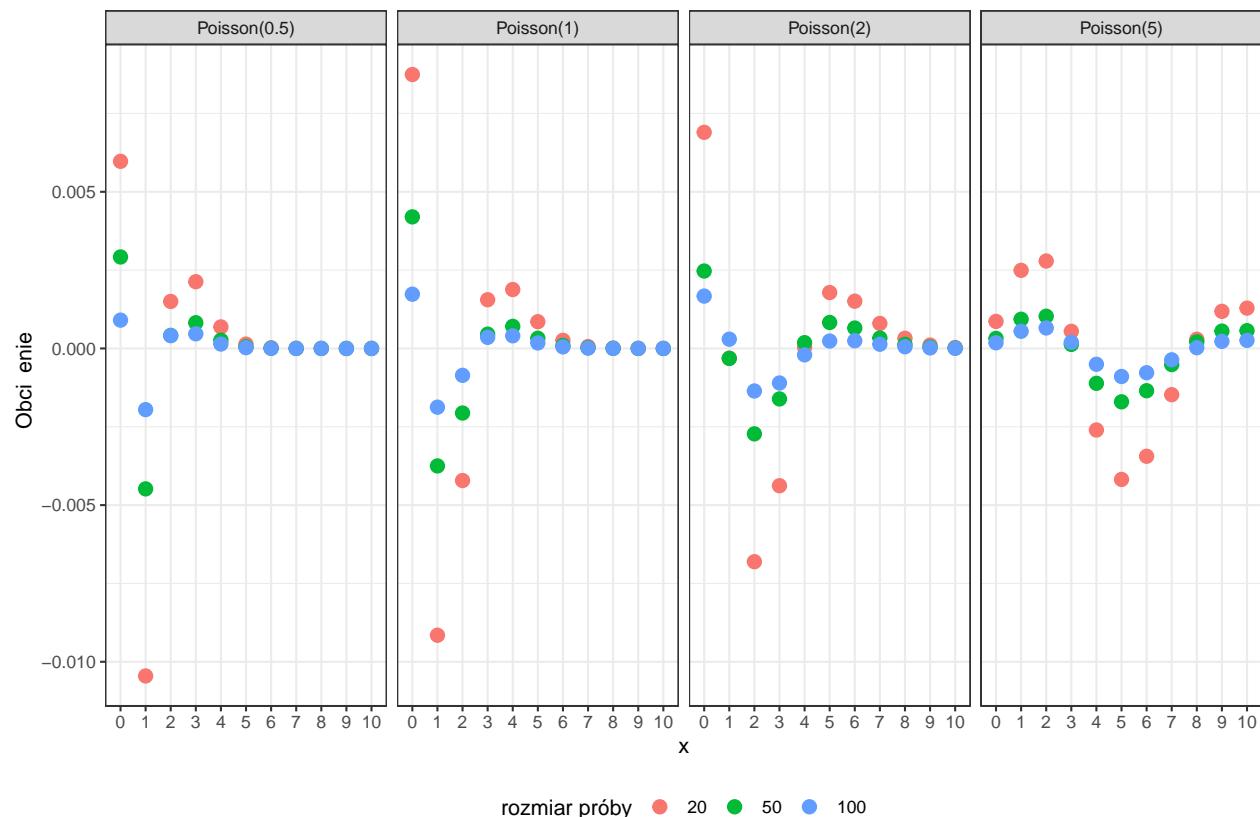
Ogólnie rzecz biorąc znaleziony estymator największej wiarygodności jest bardzo dobrym estymatorem, a najlepsze wyniki uzyskamy dla skrajnych wartości  $p$ .

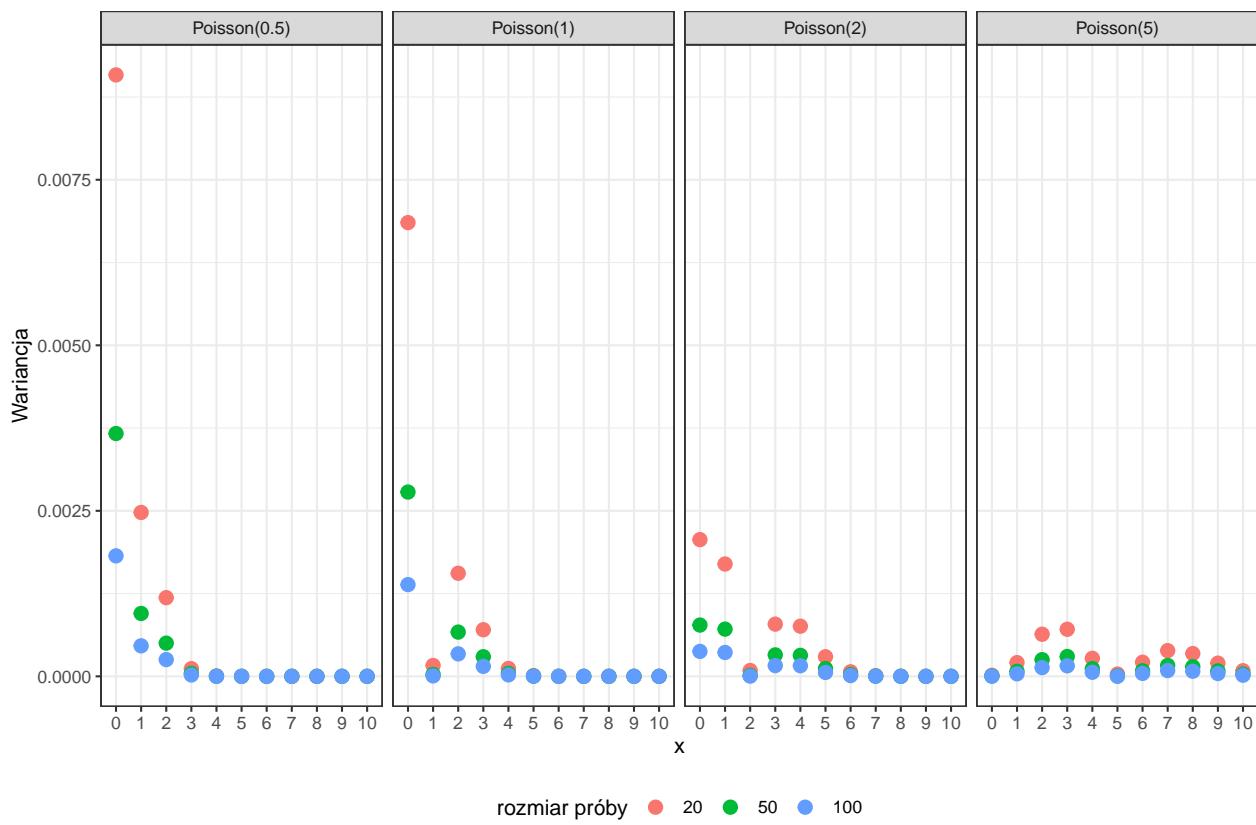
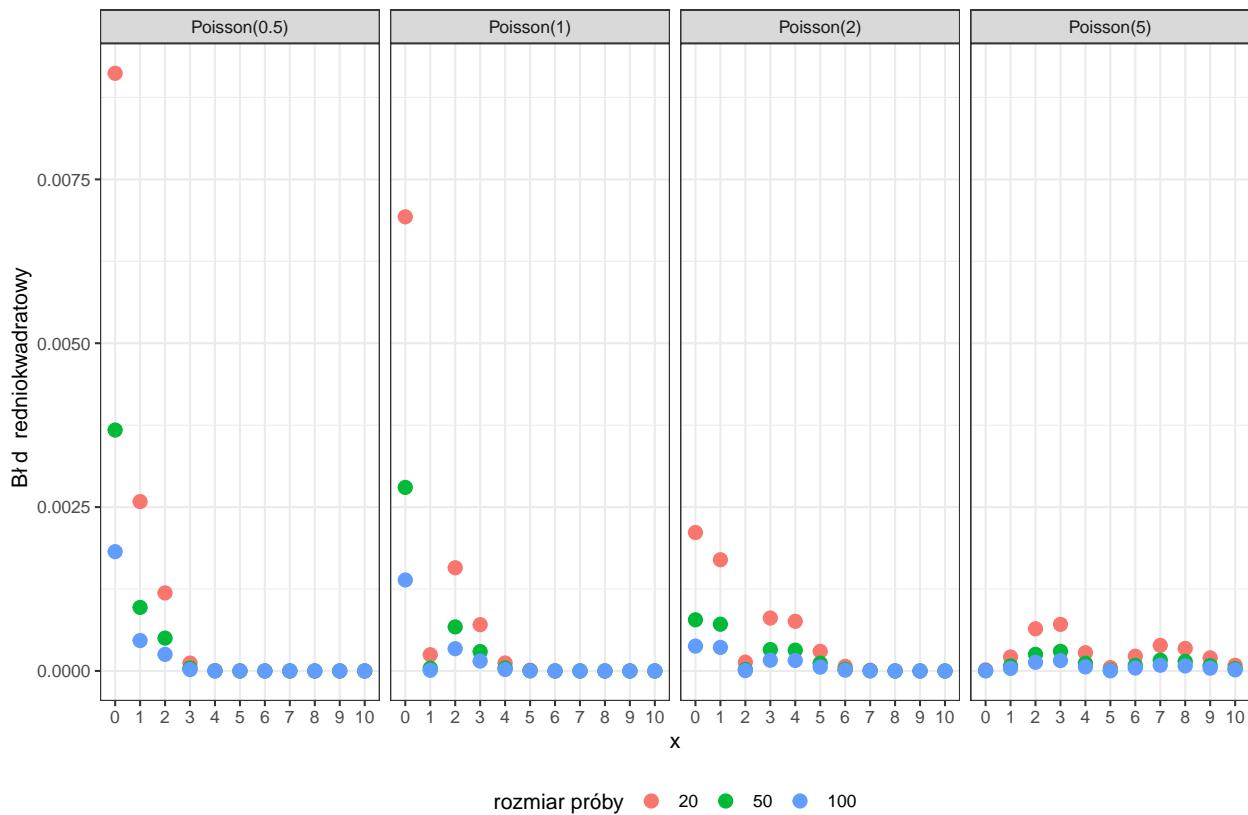
## Estymacja wielkości $P(X = x)$ dla rozkładu Poissona $P(\lambda)$ metodą największej wiarogodności

Wyznaczanie ENW dla  $P(X = x)$

Skorzystamy z tego samego twierdzenia, w którym korzystaliśmy w poprzednim zadaniu. Wiemy, że ENW parametru  $\lambda$  to  $\bar{X}$ , zaś  $P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$ , wobec tego estymatorem największej wiarogodności dla wielkości  $P(X = x)$  będzie  $\frac{e^{-\bar{X}} \bar{X}^x}{x!}$ .

**Wykres z oszacowaniem wariancji, błędu średniokwadratowego oraz obciążenia wyznaczonego estymatora w zależności od parametru  $x$**





Podobnie jak w poprzednim zadaniu, wszystkie wartości są stosunkowo małe, a nawet tego samego rzędu

$(10^{-3})$ . Również po raz kolejny obciążenie ma nieco inną strukturę niż wariancja i błąd średniokwadratowy. Dla  $\lambda \in \{0.5, 1\}$  wszystkie trzy rodzaje niedokładności estymatora maleją co do modułu wraz ze wzrostem  $x$ . Jednak dla  $\lambda = 2$  MSE i wariancja są małe w okolicach 2, a potem znowu rosną i ostatecznie maleją, natomiast przy obciążeniu nie widać takiej prawidłowości. Zaś dla  $\lambda = 5$  wariancja i błąd średniokwadratowy odnotowują najniższe wartości dla  $x \in \{1, 5, 10\}$ , natomiast obciążenie w przypadku  $x = 5$  co do modułu osiąga wartość najwyższą.

Dla mniejszych  $x$  widać również zdecydowanie większe różnice we wszystkich trzech wartościach, patrząc względem rozmiaru próby.

Podsumowując, jeśli  $x$  sporo większe lub sporo mniejsze niż  $\lambda + 3$ , wielkość rozpatrywanej przez nas próby ma drugorzędne znaczenie (im  $x$  większe, tym lepiej). Natomiast jeśli  $x$  jest bliskie  $\lambda$  (w szczególności trochę mniejsze od niej), lepiej wypadają zbiory o większych rozmiarach.

Widoczne podobieństwo między rozkładem dwumianowym a rozkładem Poissona wykorzystuje się przy oszacowywaniu tego pierwszego (łatwiej policzyć wyrażenie bez silni). Jednak przybliżenie takie stosuje się dla pojedynczej próby rozmiaru  $n > 5$ , także tutaj nie było ono możliwe do zastosowania (w tym przypadku liczebność wynosiła dokładnie 5).

## **Analiza rozkładu zmiennej $Y = \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta)$ wyznaczonej na podstawie estymacji informacji Fishera $I(\theta)$ dla rozkładu $\text{beta}(\theta, 1)$**

### **Wyznaczanie ENW informacji Fishera $I(\theta)$ dla rozkładu $\text{beta}(\theta, 1)$**

Analogicznie jak w poprzednich zadaniach, korzystamy z twierdzenia 6.1.2. Wiemy, że informacja Fishera dla parametru  $\theta$  w przypadku rozkładu beta to  $I(\theta) = \frac{1}{\theta^2}$ . Z kolei ENW parametru  $\theta$  to  $\frac{-n}{\sum_{i=1}^n \log(X_i)}$ , gdzie  $X_i$  są kolejnymi obserwacjami w próbie. W związku z tym ENW dla informacji Fishera będzie w tym przypadku  $\frac{1}{(\sum_{i=1}^n \log(X_i))^2}$ .

### **Definiowanie nowej zmiennej $Y = \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta)$ i analiza jej rozkładu**

Definiujemy na tej podstawie nową zmienną  $Y = \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta)$ . Chcąc zbadać jej rozkład, przyjrzyjmy się jej histogramom oraz wykresom kwantylowo-kwantylowym.

Ponieważ wartości  $Y$  znajdują się mniej więcej w przedziale  $(-4, 4)$ , a wykresy będądziemy umieszczać na planszy 4 na 3, zatem zbyt dużo klas na histogramie (np. domyślne 30) zaburzy odbiór wizualny, ustalmy szerokość kubelka równą 0.5 (co daje ok. 15 kubeliów).

Pozostaje jeszcze wyznaczyć kwantyle teoretyczne do wykresu kwantylowo-kwantylowego. Skorzystamy z twierdzenia 6.3.1, zgodnie z którym, ponieważ  $0 < I(\theta) < +\infty$ , zachodzi

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

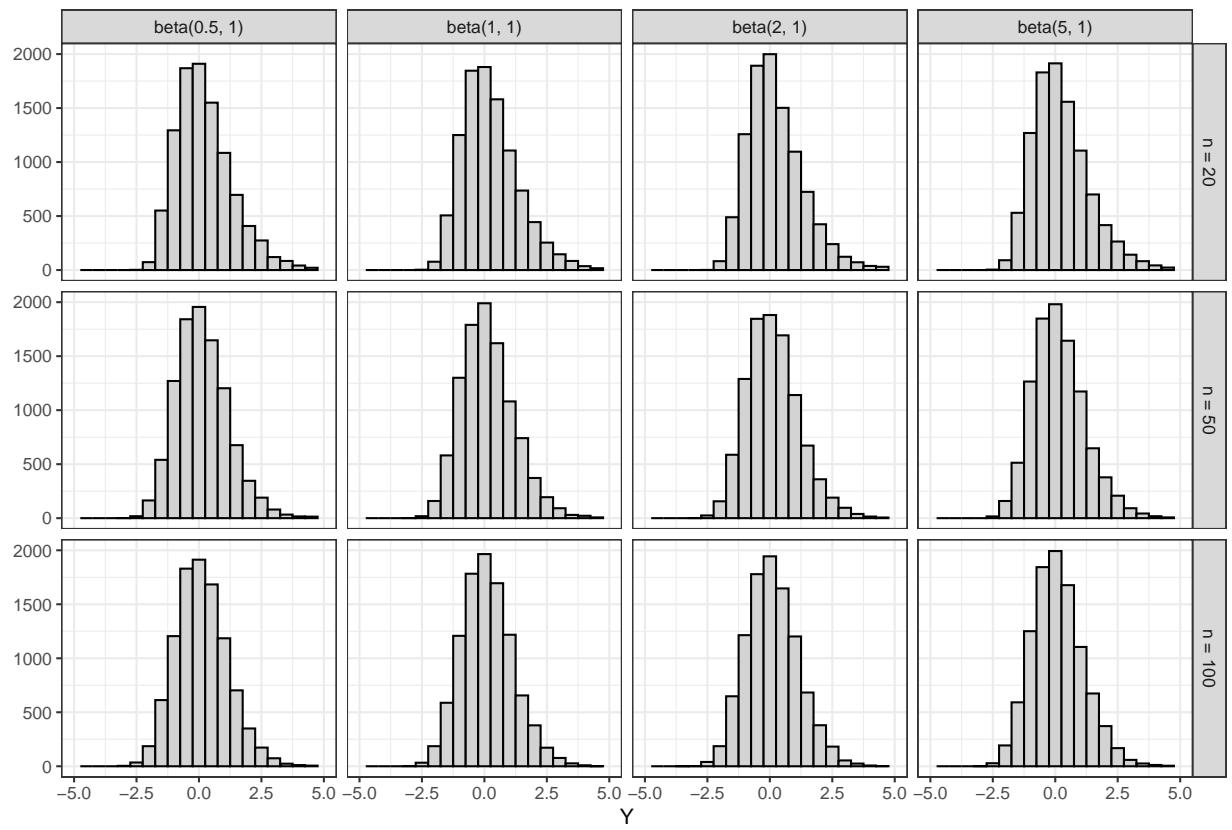
Z tego wynika, że

$$\sqrt{nI(\hat{\theta})}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, \frac{1}{I(\theta)} \cdot \sqrt{I(\theta)^2})$$

Czyli, ostatecznie

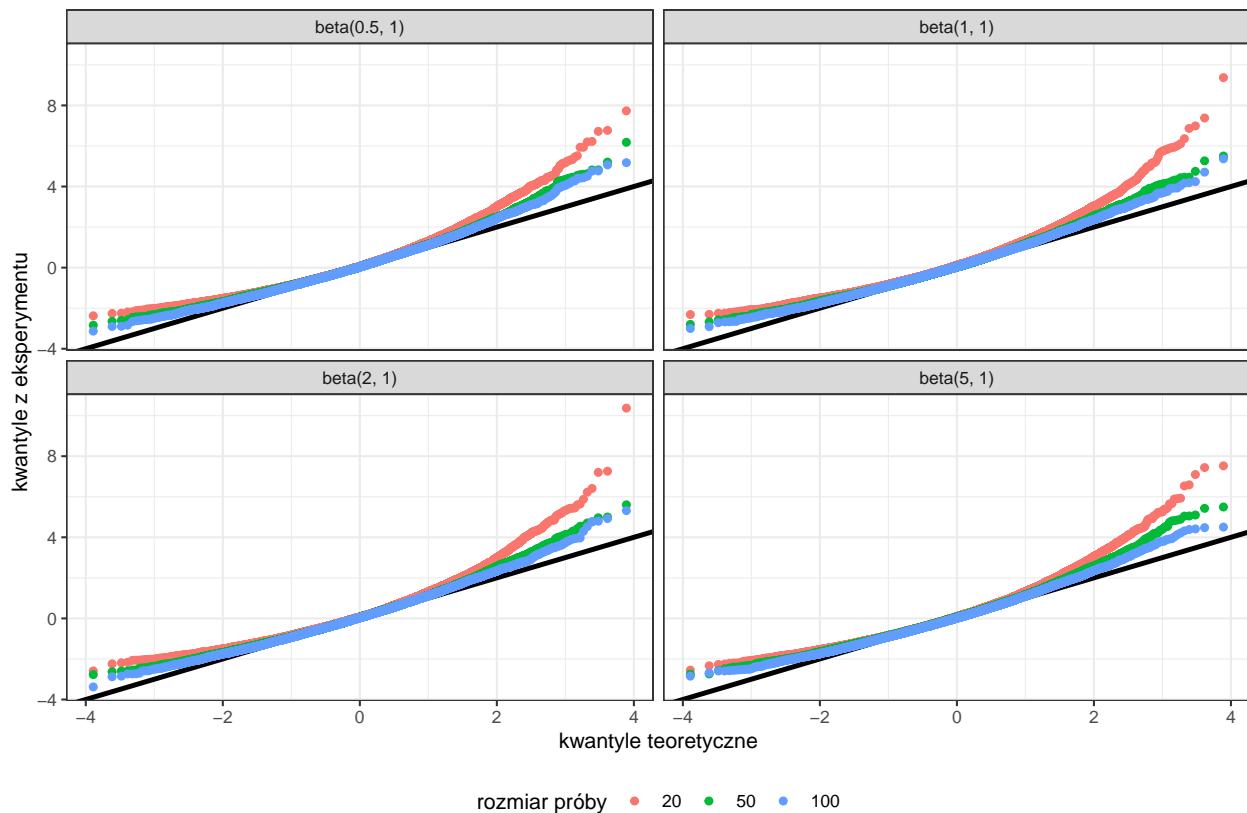
$$\sqrt{nI(\hat{\theta})}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, 1)$$

Wobec tego kwantylami teoretycznymi będą kwantyle z rozkładu normalnego  $N(0, 1)$ .



### Analiza histogramów zmiennej $Y$

Wszystkie histogramy przypominają wizualnie rozkład  $N(0, 1)$ . Dla dwóch pierwszych rzędów nieco więcej jest wartości dodatnich niż ujemnych, ale kształt poprawia się wraz ze wzrostem liczby pojedynczej próby  $n$ .



### Analiza wykresów kwantylowo-kwantylowych zmiennej $Y$

Wykresy kwantylowo-kwantylowe dodatkowo potwierdzają obserwacje poczynione przy histogramach - wartości krańcowe z eksperymetru nieco odstają od tych z rozkładu normalnego (szczególnie prawy koniec, czyli wartości dodatnie), natomiast wraz ze wzrostem  $n$  rozkład  $Y$  coraz bardziej przypomina  $N(0, 1)$ . Co więcej, analizując oba typy wykresów można dojść do wniosku, że zbieżność ta zachodzi stosunkowo szybko - widać dużą różnicę dla  $n = 20$  i  $n = 50$ .

Nie widać większych zależności między kształtem wykresów a zmianą wartości  $\theta$ , co jest uzasadnione - jak wynika twierdzenia przedstawionego na początku zadania, rozkład teoretyczny  $Y$  nie zależy od  $\theta$ .

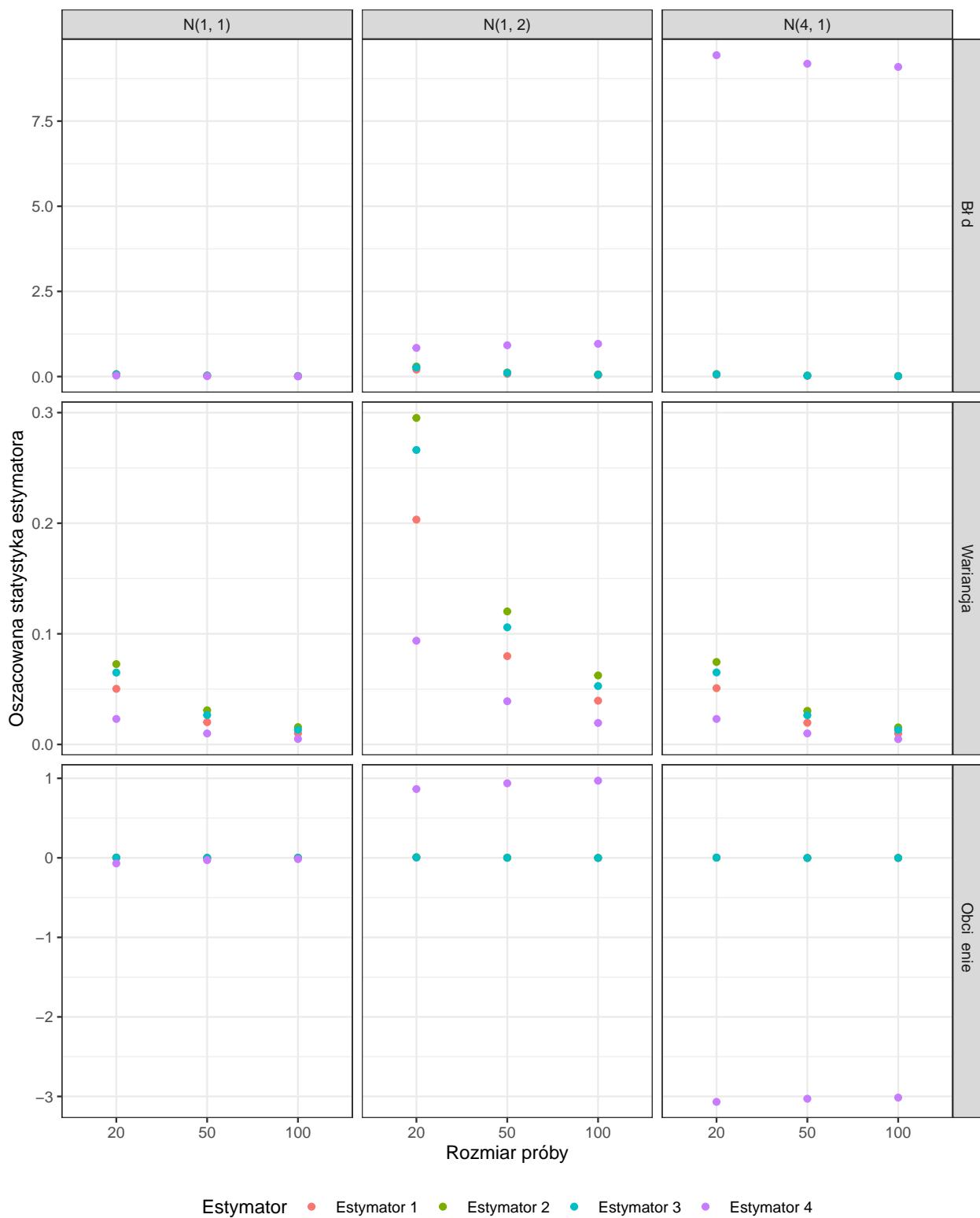
## **Estymacja parametru przesunięcia dla rozkładu Laplace'a. Porównanie z estymacją średniej dla rozkładu normalnego**

W tym zadaniu, podobnie w zadaniu 1 z listy 1, wybierać będziemy najlepszy estymator spośród podanych.  
Stosowane estymatory:

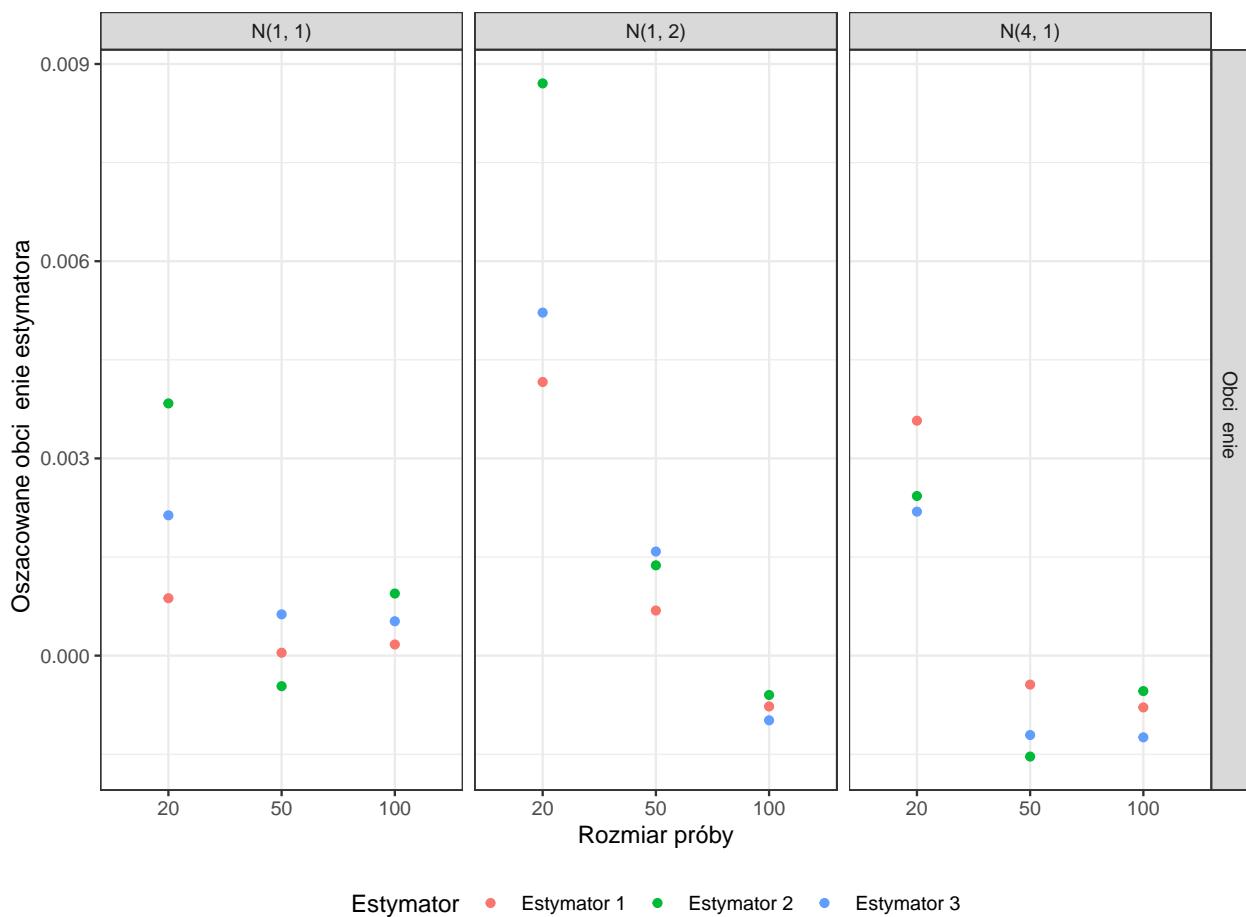
- Estymator 1: średnia arytmetyczna
- Estymator 2: mediana
- Estymator 3: średnia ważona z wybranymi wagami
- Estymator 4: średnia ważona z zadanimi wagami

Obciążenie, błąd średniokwadratowy i wariancję dla każdej z propozycji porównamy na wykresach.  
Spodziewamy się, że najlepiej wypadnie mediana, ponieważ to ona jest ENW dla tego rozkładu według przykładu z wykładu (przykład 6.1.3 we wspomnianym wyżej podręczniku).

Przypomnijmy najpierw wyniki porównania dla rozkładu normalnego:

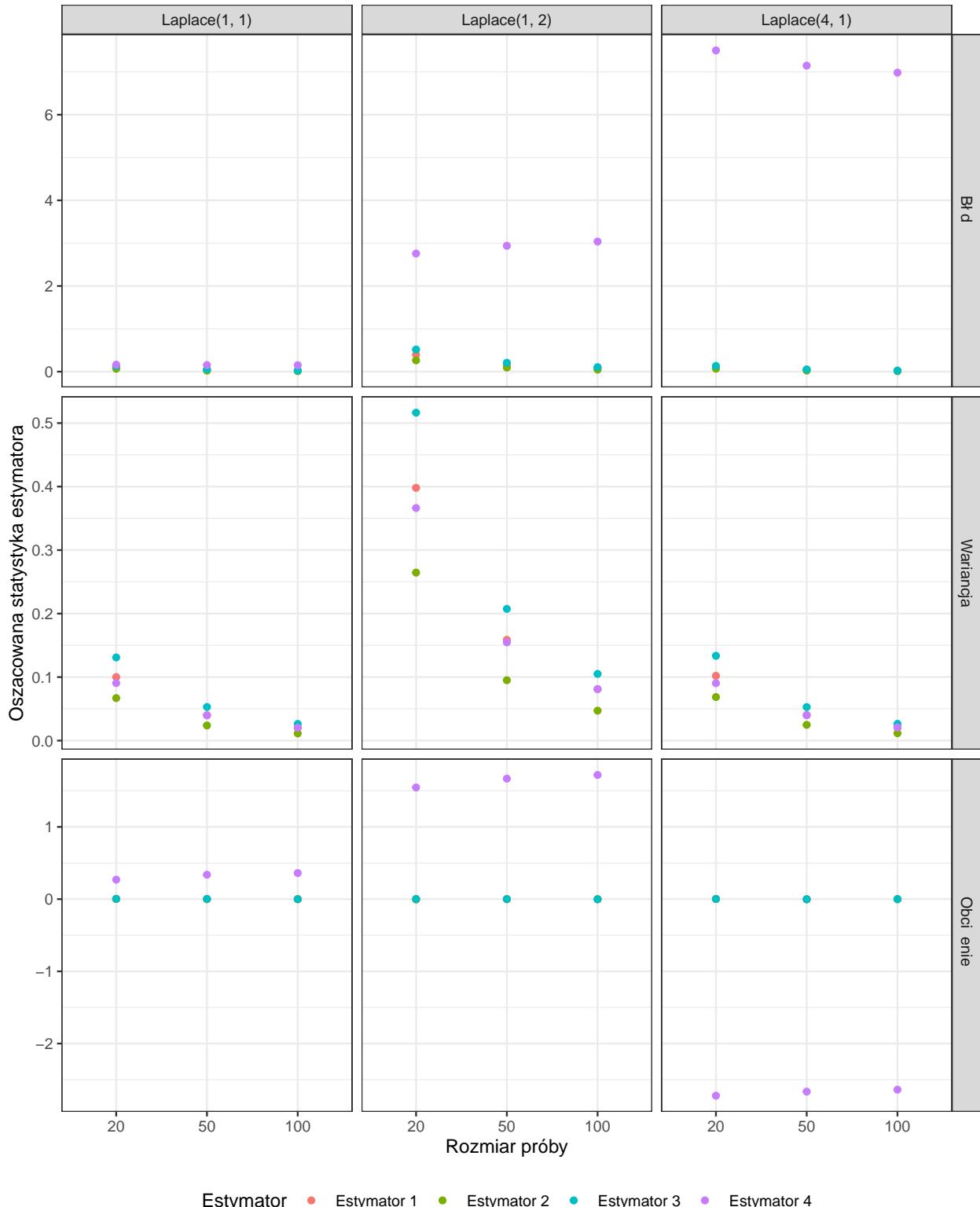


Ponieważ obciążenie dla estymatora numer 4 jest bardzo duże w porównaniu do pozostałych, zobaczymy ten wykres tylko dla propozycji 1-3 dla lepszej widoczności:

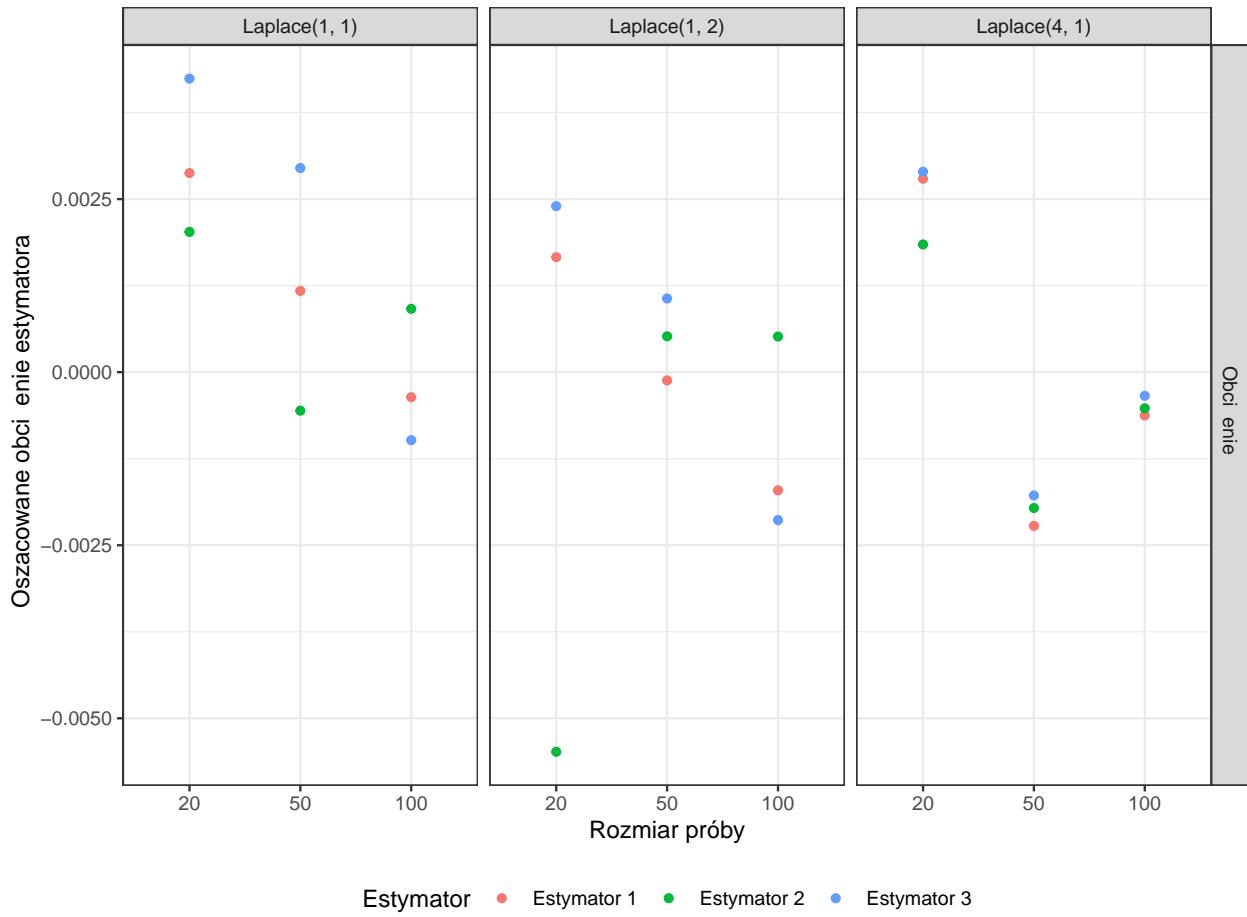


### Porównanie estymatorów dla parametru przesunięcia w rozkładzie Laplace'a

Wykresy dla rozkładu Laplace'a prezentują się następująco:

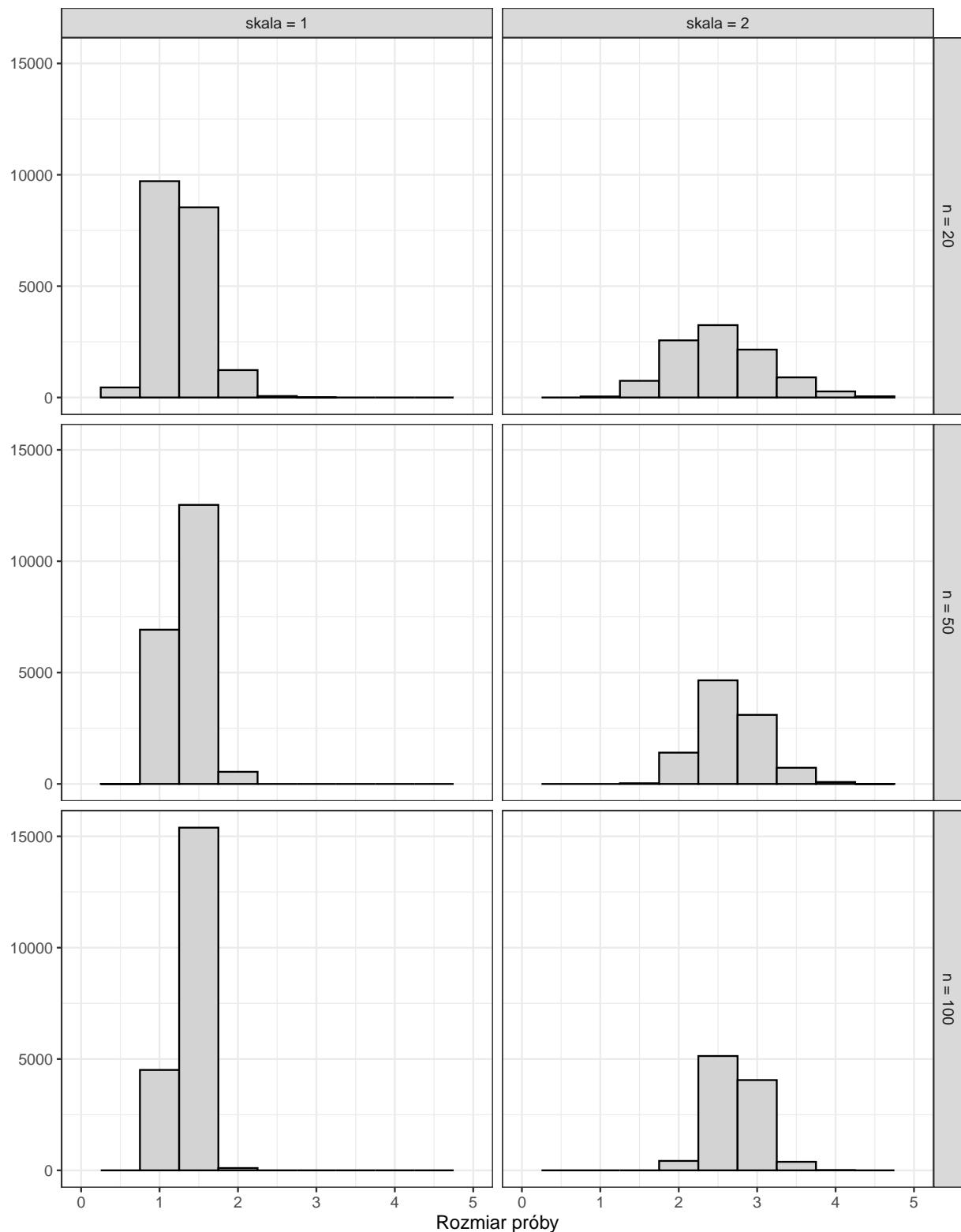


Podobnie jak w przypadku rozkładu normalnego (zadanie 1, lista 1), zdecydowanie najgorszym estymatorem parametru  $\theta$  okazał się estymator numer 4 (średnia ważona z zadanymi wagami). Jego obciążenie było tak różne od obciążenia pozostałych trzech, że zbijają się one w jeden punkt na wykresie. Zobaczmy zatem i w tym przypadku, jak wyglądałaby wizualna prezentacja obciążenia bez metody numer 4:



Zgodnie w przewidywaniami, w przeciwieństwie do rozkładu normalnego, gdzie najlepszym estymatorem okazała się średnia arytmetyczna (a drugim godnym uwagi - średnia ważona w własnymi wagami), tu stosunkowo najlepszy okazał się estymator numer 2, czyli mediana. Ma on najmniejsze obciążenie, wariancję oraz błąd średniokwadratowy. Wyniki estymatorów 1-3 poprawiają się wraz ze wzrostem liczby próby  $n$ . Natomiast estymator numer 4, czyli średnia z zadanimi wagami, tak jak w przypadku rozkładu normalnego znacząco odstaje od pozostałych.

Spójrzmy na proste histogramy, żeby zobaczyć, czy może on być estymatorem parametru skali zamiast parametru położenia:



Jak widać, estymator ten nie jest dobrym estymatorem parametru skali, zatem nie jest dobrym estymatorem żadnego z parametrów dla rozkładu Laplace'a. Za to w przypadku rozkładu normalnego był on całkiem dobrym estymatorem odchylenia standardowego.

Podsumowując, dla rozkładu Laplace'a spośród czterech rozpatrywanych opcji najlepszym estymatorem

położenia  $\theta$  okazała się mediana. Choć rozkład Laplace'a i rozkład normalny cechuje pewne podobieństwo wizualne (symetria względem pierwszego parametru, maksimum gęstości prawdopodobieństwa w  $\theta$ , lekkie ogony), to jednak estymatory dla parametrów tych rozkładów się od siebie różnią.