# Lista 1

Helena Sękowska-Słoka, nr indeksu 321531

2023-10-28

## SPIS TREŚCI

# Zadanie 1

## Przygotowanie zbioru obserwacji oraz obliczenie wartości estymatorów

Estymator jest to szacowanie parametru dla całego zbioru na podstawie obliczenia jego wartości dla pewnej próby z tegoż zbioru.
W tym zadaniu losujemy próby z rozkładów normalnych $N(\theta, \sigma^2)$, estymujemy zaś parametr $\theta$, czyli wartość oczekiwaną.
Generujemy 9 zbiorów obserwacji z rozkładów danych w zadaniach: 10 000 prób po 50 obserwacji każda, osobno dla podpunktów a, b oraz c, z podziałem na różne wartości $n$, jak podano w poleceniu zadania 7. Następnie dla każdej z prób obliczamy wartość estymatora parametru $\theta$ w postaci kolejno średniej arytmetycznej, mediany i dwóch różnych średnich ważonych.
Przy średniej ważonej z własnym wyborem wag zastosujemy wagi wylosowane według schematu:

1. Losujemy wektor pięćdziesięciu dodatnich liczb rzeczywistych (niejednakowych).
2. Dzielimy każdą liczbę przez sumę wszystkich liczb z wektora, uzyskując tym samym wektor wag sumujących się do 1.

Losowanie poprzedzamy ustawieniem ziarna w celu zachowania powtarzalności uzyskanych wyników.

## Oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów

Otrzymawszy wartości estymatorów parametru $\theta$, oszacowano wariancję, błąd średniokwadratowy oraz obciążenie każdego z estymatorów, osobno dla każdego podpunktu, uwzględniając liczebności prób z zadania numer 7.
Wyniki podsumowano w tabelach.

Table 1: Średnia arytmetyczna

| | a | | | b | | | c | | |
|---|---|---|---|---|---|---|---|---|---|
| n | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 |
| wariancja | 0.0508 | 0.0199 | 0.0102 | 0.0494 | 0.0195 | 0.0102 | 0.2015 | 0.0792 | 0.0398 |
| MSE | 0.0508 | 0.0199 | 0.0102 | 0.0494 | 0.0195 | 0.0102 | 0.2015 | 0.0792 | 0.0398 |
| obciążenie | 0.0028 | -0.0018 | 0.0009 | -0.0024 | -0.0022 | -0.0007 | 0.0017 | 0.0049 | 0.0016 |

Table 2: Mediana

| | a | | | b | | | c | | |
|---|---|---|---|---|---|---|---|---|---|
| n | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 |
| wariancja | 0.0740 | 0.0300 | 0.0161 | 0.0731 | 0.0305 | 0.0157 | 0.2987 | 0.1235 | 0.0617 |
| MSE | 0.0740 | 0.0300 | 0.0161 | 0.0731 | 0.0305 | 0.0157 | 0.2987 | 0.1235 | 0.0617 |
| obciążenie | 0.0018 | -0.0023 | 0.0012 | -0.0009 | -0.0029 | -0.0013 | -0.0005 | 0.0038 | 0.0030 |

Table 3: Średnia ważona 1

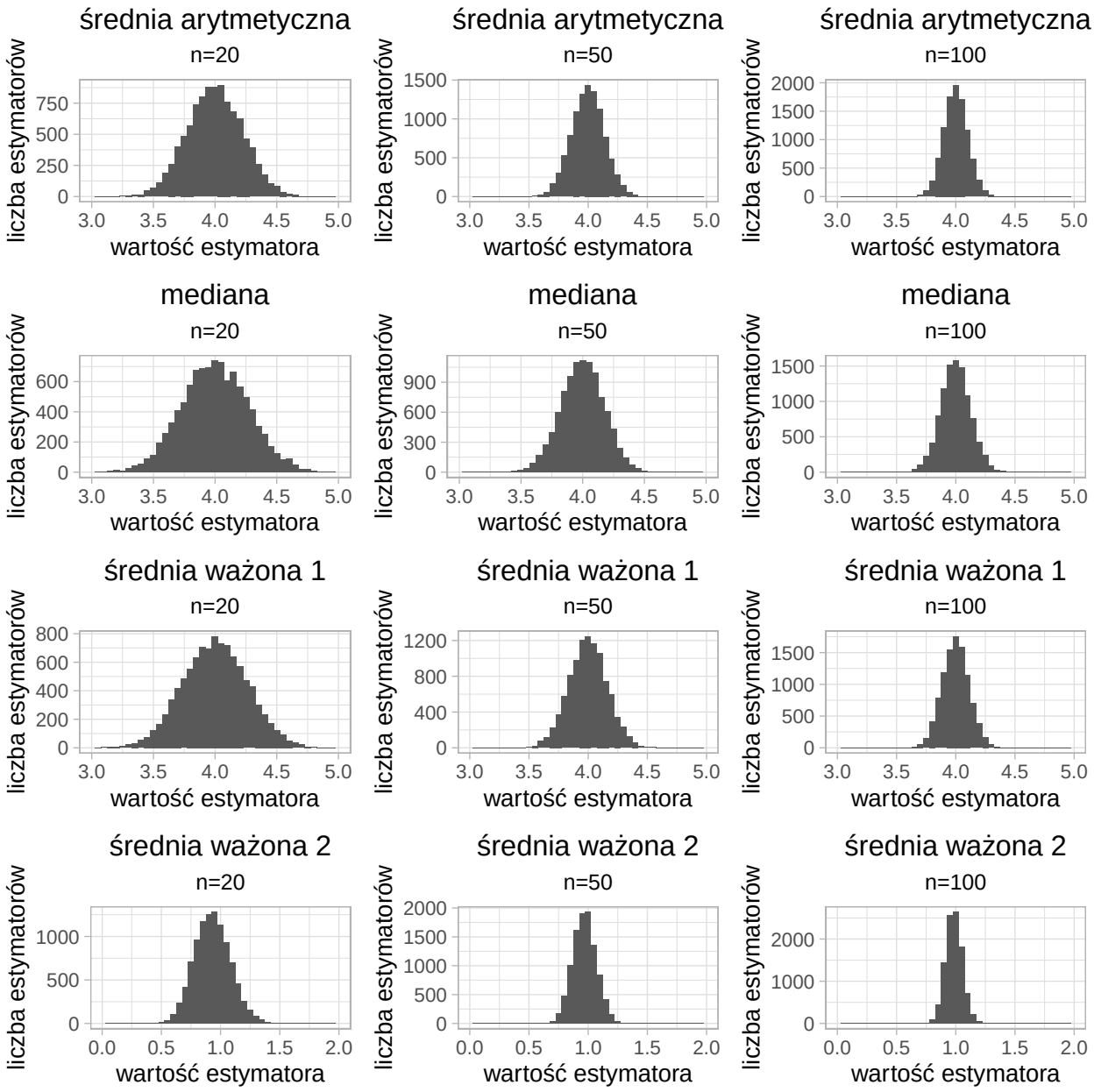
| | a | | | b | | | c | | |
|---|---|---|---|---|---|---|---|---|---|
| n | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 |
| wariancja | 0.0703 | 0.0262 | 0.0131 | 0.0687 | 0.0258 | 0.0132 | 0.2748 | 0.1038 | 0.0519 |
| MSE | 0.0703 | 0.0262 | 0.0131 | 0.0687 | 0.0258 | 0.0132 | 0.2748 | 0.1038 | 0.0519 |
| obciążenie | 0.0047 | -0.0015 | 0.0013 | -0.0005 | -0.0010 | -0.0013 | 0.0018 | 0.0037 | 0.0017 |

Table 4: Średnia ważona 2

| | a | | | b | | | c | | |
|---|---|---|---|---|---|---|---|---|---|
| n | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 |
| wariancja | 0.0236 | 0.0094 | 0.0049 | 0.0232 | 0.0098 | 0.0049 | 0.0934 | 0.0381 | 0.0193 |
| MSE | 0.0285 | 0.0102 | 0.0052 | 9.4384 | 9.1870 | 9.0991 | 0.8389 | 0.9289 | 0.9653 |
| obciążenie | -0.0696 | -0.0279 | -0.0154 | -3.0684 | -3.0294 | -3.0157 | 0.8634 | 0.9438 | 0.9726 |

## Wnioski

Estymatorem o najmniejszej wariancji dla każdego podpunktu okazała się druga średnia ważona. W każdym przypadku miała ona wariancję ponad dwukrotnie mniejszą niż następna w kolejności średnia arytmetyczna. Największą wariancję zaś uzyskała mediana, szczególnie w podpunkcie c) dla $n = 20$. Co więcej, dla podpunktu c), czyli rozkładu $N(\theta = 1, \sigma^2 = 4)$, wariancja była największa (dla każdego estymatora).
Dla podpunktu a), czyli rozkładu $N(\theta = 1, \sigma^2 = 1)$, najmniejszy błąd średniokwadratowy (MSE) miała również druga średnia ważona. Jednakże w podpunkcie b) (rozkład $N(\theta = 4, \sigma^2 = 1)$) zdecydowanie lepiej wypada średnia arytmetyczna - błąd dla drugiej średniej ważonej jest tu wyjątkowo duży, o dwa rzędy wielkości większy od pozostałych. Chcąc znaleźć potencjalną przyczynę tej anomalii, sporządźmy histogramy estymatorów dla tego podpunktu:




Zwróćmy uwagę na zakres osi $X$ w przypadku drugiej średniej ważonej - jej wartości nie oscylują, jak w pozostałych przypadkach, w okolicy 4, a w okolicy 1. Zauważmy, że dla tego podpunktu wartość $\theta$ to 4, jednak wartość $\sigma$ to 1. W podpunkcie a) zachodziło $\theta = \sigma$, zaś w podpunkcie c), gdzie $\theta$ i $\sigma$ są różne, druga średnia ważona również obarczona jest znacznie większym błędem niż pozostałe estymatory, jednak tam różnica nie jest aż tak duża, ponieważ różnica miedzy $\theta$ i $\sigma$ także jest mniejsza ($\theta = 1, \sigma = 2$).

Wobec tego można postawić tezę, że druga średnia ważona jest znacznie lepszym estymatorem parametru $\sigma$ niż parametru $\theta$.

Wartości obciążenia były stosunkowo małe dla trzech pierwszych estymatorów (odbiegające wartości dla czwartego estymatora można uzasadnić tezą postawioną powyżej). Co do modułu najmniejsze było ono dla mediany w podpunkcie c) dla $n = 20$ oraz dla pierwszej średniej ważonej w podpunkcie b) dla $n = 20$.

Patrząc na ogół danych w tabeli, najlepszym estymatorem w przypadku podpunktuów a) i b) wydaje się średnia arytmetyczna. Choć jeśli dla podpunktu b) zależy nam bardziej na małej wartości obciążenia niż na pozostałych parametrów, możemy jeszcze wziąć pod uwagę zastosowanie pierwszej średniej ważonej (szczególnie dla małej liczebności prób). Podobnie w przypadku podpunktu c), z tym, że tu najmniejsze

obciążenie dla małych wartości $n$ ma mediana (dla dużych prób zdecydowanie lepiej pod każdym względem wypada średnia arytmetyczna).
Ogólnie rzecz biorąc, wariancja i błąd średniokwadratowy maleją wraz ze wzrostem liczebności prób (za wyjątkiem drugiej średniej ważonej w podpunkcie c)), zaś obciążenie zachowuje się różnie, jednak jego wartości w obrębie różnych $n$ dla rozpatrywanych przypadków różnią się maksymalnie o jeden rząd wielkości i są stosunkowo małe.

# Zadanie 5

## Przygotowanie do obliczeń

W przeciwieństwie do zadania 1, w tym przypadku nie wyliczamy estymatora $\theta$ poprzez zastosowanie podanych wzorów, a szacujemy jego wartość za pomocą metody Newtona-Rhapsona, korzystając z pierwszej i drugiej pochodnej funkcji logwiarogodności, czyli odpowiednio:

$$l'(\theta) = \frac{n}{\sigma} - 2 \cdot \sum_{i=1}^{n} \frac{exp\frac{-(x_i-\theta)}{\sigma}}{\sigma \cdot (1+exp\frac{-(x_i-\theta)}{\sigma})}$$

$$l''(\theta) = -2 \cdot \sum_{i=1}^{n} \frac{exp\frac{-(x_i-\theta)}{\sigma}}{\sigma^2 \cdot (1+exp\frac{-(x_i-\theta)}{\sigma})^2}$$

Zauważmy, że przyrównując pierwszą pochodną do zera, otrzymujemy wyrażenie, które jest praktycznie nie do rozwiązania ręcznie w przypadku ogólnym. Z pomocą komputera można jednak obliczyć jego rozwiązanie numerycznie, oczywiście jedynie z pewnym przybliżeniem. Żeby się upewnić, że rzeczywiście znaleziony punkt jest właściwy, należy jeszcze sprawdzić wartość drugiej pochodnej - szukamy maksimum, zatem powinna być ona ujemna.
Zauważmy, że warunek ten będzie spełniony zawsze, ponieważ

$$-2 \cdot \sum_{i=1}^{n} \frac{exp\frac{-(x_i-\theta)}{\sigma}}{\sigma^2 \cdot (1+exp\frac{-(x_i-\theta)}{\sigma})^2} < 0$$

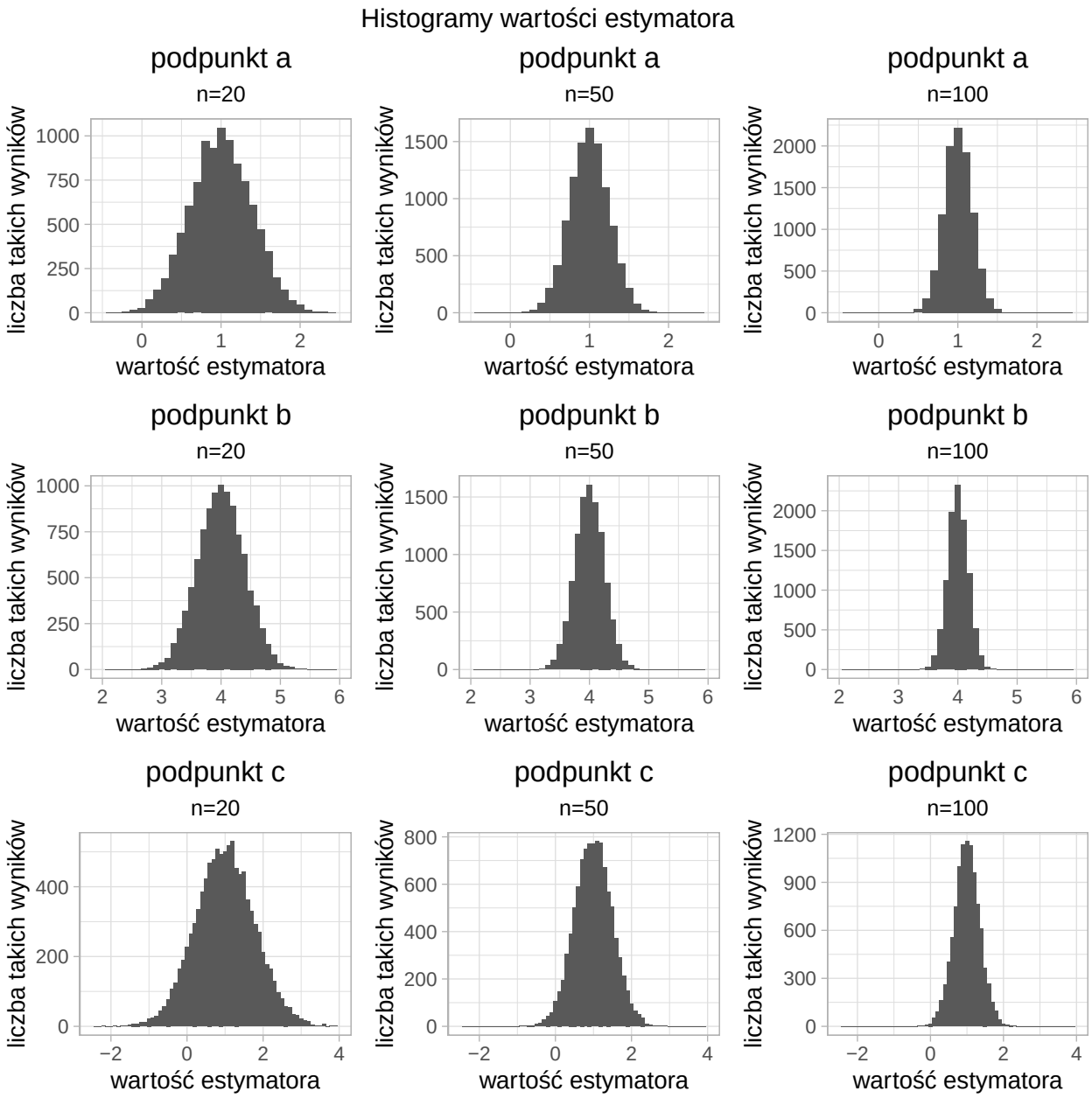
dla każdego $\theta \in \mathbb{R}$.
Ustalmy dokładność na poziomie $a = 0.001$, zaś warunkiem trwania pętli w metodzie Newtona-Rhapsona przy zadanej dokładności $a$ będzie:

$$|l'(\theta)| > a$$

Dodatkowo ograniczmy kroki od góry przez 15.
Rozkład logistyczny cechuje pewne podobieństwo do rozkładu normalnego, w szczególności - podobny wygląd wykresów gęstości, w tym symetria względem wartości oczekiwanej. Zatem, na podstawie wiedzy z wykładu oraz wyników z zadania 1, jako punkt początkowy weźmy średnią arytmetyczną z próby.

### Obliczenie wartości estymatora

Wyliczone w powyższy sposób wartości estymatora parametrów $\theta$ zaprezentujmy na histogramach:




**Wnioski**

Zauważalną różnicą jest ta w szerokości histogramów dla poszczególnych wartości $n$ - rozpiętość słupków maleje bardzo szybko wraz ze wzrostem liczebności próby. Im większe $n$, tym więcej estymatorów ma wartości bliskie prawdziwej wartości parametru $\theta$.
Warto natomiast zwrócić uwagę na fakt, że nawet dla najmniejszego $n = 20$ wartości ENW obliczonego przez algorytm układały się w przybliżeniu w kształt funkcji gęstości rozkładu normalnego wokół prawdziwej wartości $\theta$.

## Oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia estymatora

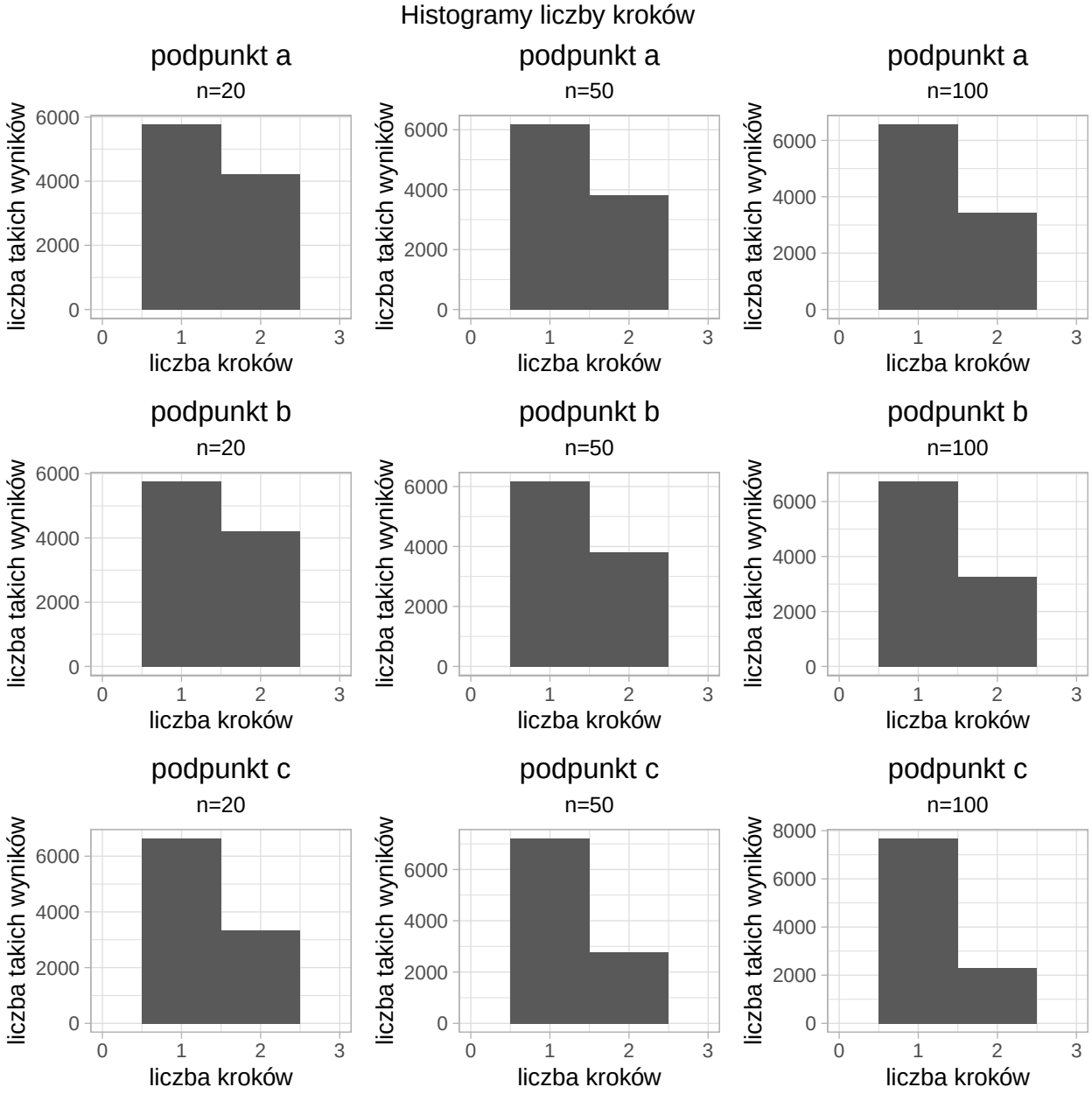
Table 5: Estymator największej wiarogodności dla rozkładu logistycznego

| | a | | | b | | | c | | |
|---|---|---|---|---|---|---|---|---|---|
| n | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 |
| wariancja | 0.1519 | 0.0608 | 0.0302 | 0.1569 | 0.0608 | 0.0297 | 0.6114 | 0.2434 | 0.0302 |
| MSE | 0.1519 | 0.0608 | 0.0302 | 0.1569 | 0.0607 | 0.0297 | 0.6114 | 0.2434 | 0.0302 |
| obciążenie | 0.0021 | -0.0032 | 0.0000 | 0.0018 | 0.0006 | 0.0008 | -0.0019 | -0.0006 | 0.0000 |

**Wnioski**

Dla $n = 20$ i $n = 50$ wariancja i błąd średniokwadratowy były zdecydowanie największe w przypadku podpunktu c), czyli rozkładu logistycznego o parametrach $\theta = 1, \sigma = 2$. Natomiast dla $n = 100$ różnice były już niemal niezauważalne, a same wartości tych statystyk - bardzo małe. Można zatem postawić tezę, że w przypadku przypuszczenia, że dane pochodzą z rozkładu o większym odchyleniu standardowym, należy wziąć do badań próbę o większej liczebności.
Obciążenie nie było duże dla żadnego z rozpatrywanych rozkładów. Największą wartość co do modułu osiągnęło dla podpunktu a) przy pięćdziesięcioelementowej próbie, ale wciąż była to wielkość rzędu zaledwie jednej tysięcznej.
Podobnie jak w zadaniu 1, i tutaj wariancja i błąd średniokwadratowy malały wraz ze wzrostem $n$, natomiast obciążenie zachowywało się różnie.
Warto wspomnieć, że na szacunkowe wartości wariancji, błędu średniokwadratowego oraz obciążenia z pewnością wpłynęła wybrana na początku dokładność.

### Analiza liczby kroków w algorytmie




Im większa wartość $n$, tym średnio mniej kroków potrzebował algorytm, żeby osiągnąć zadaną dokładność $a = 0.001$. Natomiast w każdym przypadku liczba kroków wynosiła 1 lub 2, co wskazuje na to, że wybór średniej arytmetycznej jako punktu początkowego był dobry. Stosunkowo najmniej obrotów pętli potrzebne było w podpunkcie c) (dla dowolnego $n$), jednakże warto zauważyć, że to właśnie tam dla $n = 20$ i $n = 50$ wariancja i błąd średniokwadratowy były największe.
Podsumowując, w przypadku rozkładu logistycznego liczenie ENW metodą Newtona-Rhapsona przy wyborze średniej arytmetycznej z próby jako punktu początkowego wydaje się uzasadnione i sensowne.

# Zadanie 6

## Przygotowanie do obliczeń

Analogicznie jak w zadaniu 5, użyjemy tu metody Newtona-Rhapsona, korzystając z pierwszej i drugiej pochodnej funkcji logwiarogodności, czyli odpowiednio:

$$l'(\theta) = 2 \cdot \sum_{i=1}^{n} \frac{x_i - \theta}{\sigma^2 + (x_i - \theta)^2}$$

$$l''(\theta) = 2 \cdot \sum_{i=1}^{n} \frac{-\sigma^2 + (x_i - \theta)^2}{(\sigma^2 + (x_i - \theta)^2)^2}$$

I tu przyrównując pierwszą pochodną do zera, znowu otrzymujemy wyrażenie, które jest praktycznie nie do rozwiązania ręcznie w przypadku ogólnym. Natomiast dla rozkładu Cauchy'ego nie dla każdej wartości $\theta$ druga pochodna będzie ujemna, w związku z tym może zdarzyć się sytuacja, gdzie algorytm trafi nie w minimum, którego szukamy, a w maksimum. W związku z tym musimy dokonać pewnych zmian.
Dokładność, warunek trwania pętli i ograniczenie pozostawmy bez zmian. Dodatkowo będziemy jednak musieli sprawdzać znak drugiej pochodnej. Jako porządane weźmiemy pod uwagę jedynie te wyniki, gdzie druga pochodna będzie ujemna, zaś wartość estymatora nie będzie rozbiegała.
Podobnie jak z rozkładem logistycznym i normalnym, rozkład Cauchy'ego cechuje pewne podobieństwo do rozkładu Laplace'a. Ponieważ ENW dla rozkładu Laplace'a jest mediana, weźmiemy ją w tym przypadku jako punkt wyjścia dla metody Newtona-Rhapsona.

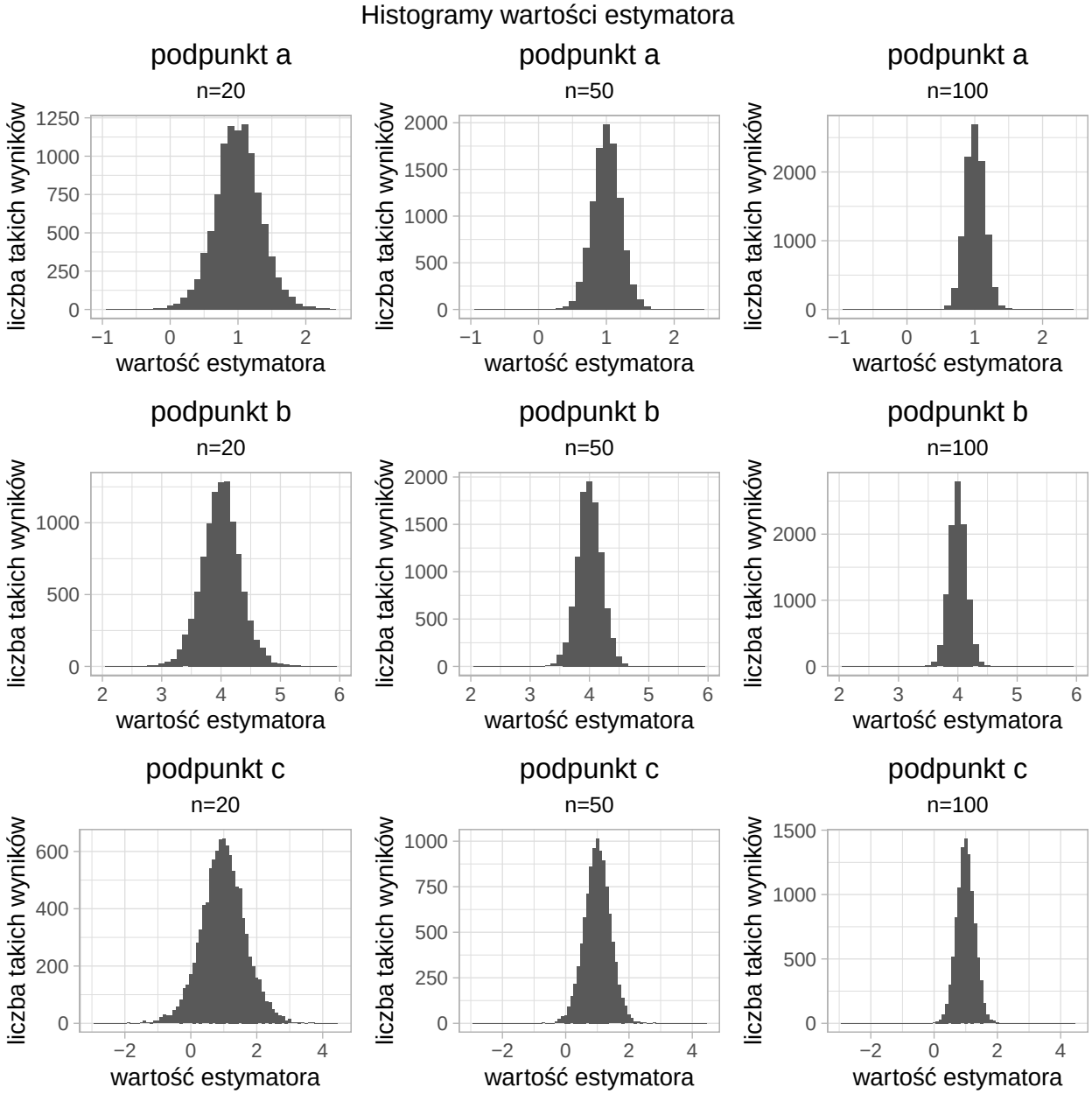
### Obliczenie wartości estymatora

W tabeli przedstawiono, ile z wyników spełniło założone wcześniej warunki:

Table 6: Liczba wartości estymatora spełniających warunki

| | a | | | b | | | c | | |
|---|---|---|---|---|---|---|---|---|---|
| n | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 |
| liczba wyników | 9972 | 9998 | 9996 | 9978 | 9998 | 9997 | 9965 | 9992 | 9993 |

Wartości zgodne z założeniami zaprezentujmy na histogramach:



**Wnioski**

Zdecydowanie najwięcej wartości było minimum zamiast maksimum lub rozbiegało dla $n = 20$, jednak i tu najgorszym wynikiem jest 35 usuniętych wartości. Dla $n = 50$ i $n = 100$ niepasujące okazały się zaledwie pojedyczne przypadki.
Podobnie jak dla rozkładu logistycznego, i tu histogramy wraz ze wzrostem $n$ robią się coraz węższe, zaś ich środki znajdują się w prawdziwych wartościach $\theta$. Warto jednak zaznaczyć, że dla tego rozkładu histogramy mają nieco lżejsze ogony, szczególnie dla $n = 20$. Może to mieć jednak związek z usunięciem rozbiegających wartości.

## Oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia estymatora

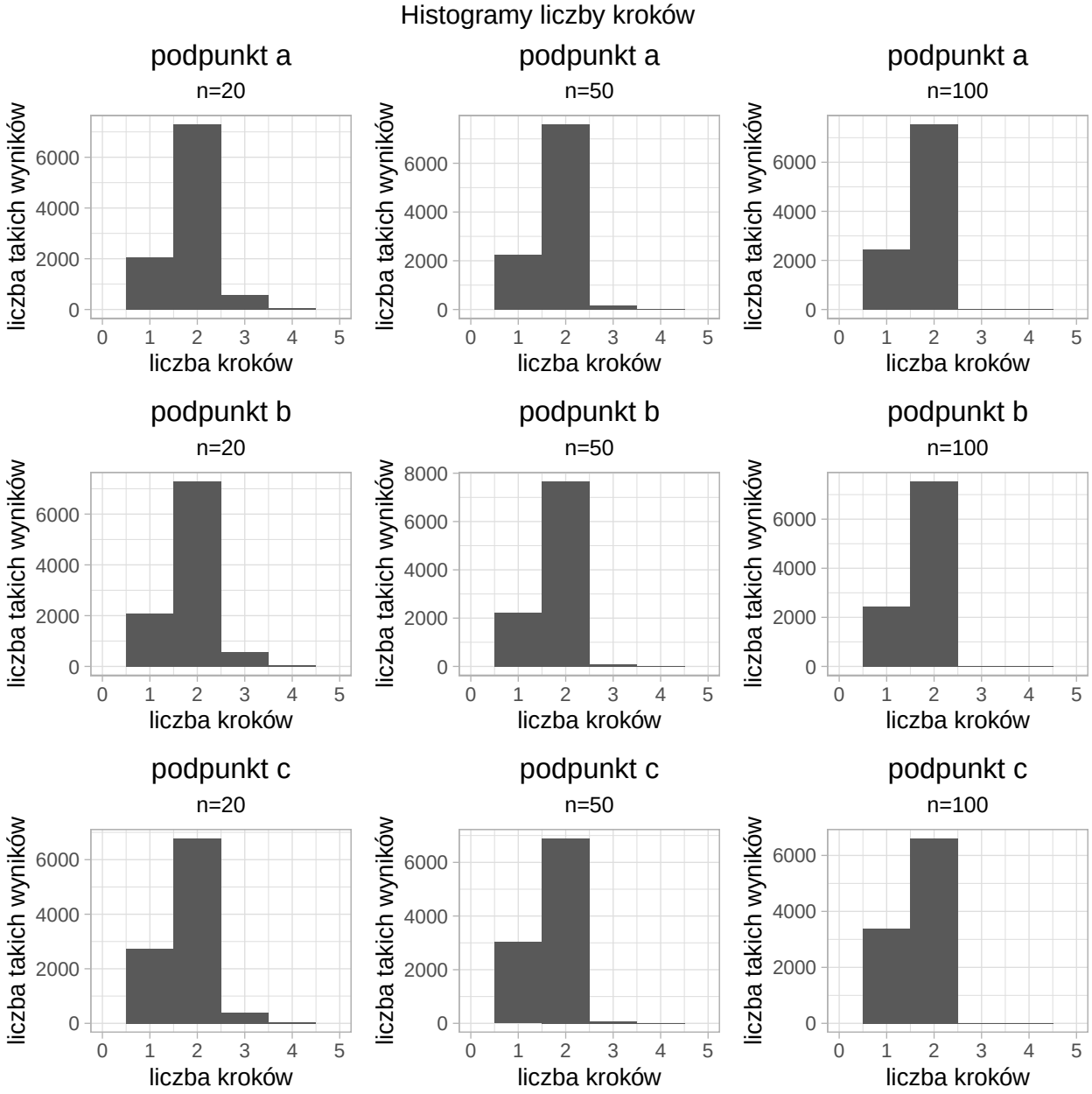
Table 7: Estymator największej wiarogodności dla rozkładu Cauchy'ego

| | a | | | b | | | c | | |
|---|---|---|---|---|---|---|---|---|---|
| n | 20 | 50 | 100 | 20 | 50 | 100 | 20 | 50 | 100 |
| wariancja | 0.1628 | 0.0421 | 0.0204 | 0.1100 | 0.0412 | 0.0207 | 0.4618 | 0.1708 | 0.0204 |
| MSE | 0.1628 | 0.0421 | 0.0204 | 0.1100 | 0.0412 | 0.0207 | 0.4618 | 0.1708 | 0.0204 |
| obciążenie | 0.0027 | 0.0001 | -0.0004 | 0.0017 | -0.0001 | -0.0009 | -0.0099 | 0.0024 | -0.0004 |

**Wnioski**

Podobnie jak w zadaniu 5, przy $n = 20$ i $n = 50$ wariancja i błąd średniokwadratowy były zdecydowanie największe w przypadku podpunktu c), czyli rozkładu o parametrach $\theta = 1, \sigma = 2$. Natomiast dla $n = 100$ różnice i wartości tych statystyk były jeszcze mniejsze niż dla rozkładu logistycznego. Zatem można tu postawić analogiczną tezę, iż w przypadku przypuszczenia, że dane pochodzą z rozkładu o większym odchyleniu standardowym, należy wziąć do badań próbę o większej liczebności.
Obciążenie nie było duże dla żadnego z rozpatrywanych rozkładów. Największą wartość co do modułu osiągnęło dla podpunktu c) przy dwudziestoelementowej próbie.
Podobnie jak w poprzednich zadaniach, i tutaj wariancja i błąd średniokwadratowy malały wraz ze wzrostem $n$, natomiast dla obciążenia nie widać takiej prawidłowości, choć warto zaznaczyć, że jest ono większe o rząd wielkości dla $n = 20$. Warto także wspomnieć, że, jak w zadaniu 5, na szacunkowe wartości wariancji, błędu średniokwadratowego oraz obciążenia z pewnością wpłynęła wybrana na początku dokładność. Byłyby też one o wiele większe, gdyby uwzględnić rozbiegające wyniki.

## Analiza liczby kroków w algorytmie




**Wnioski**

Jeśli liczony estymator był zbieżny, zbiegał stosunkowo szybko, choć nieco wolniej niż dla rozkładu logistycznego - w zdecydowanej większości przypadków liczba obrotów pętli wynosiła 2. Maksymalna wartość to 4 kroki, ale są to pojedyncze przypadki.
Liczba kroków potrzebnych do uzyskania porządanego wyniku malała wraz ze wzrostem $n$, jednak nie tak szybko jak w przypadku rozkładu logistycznego. I tu średnio najmniej kroków potrzeba było w podpunkcie c), jednak dla tego rozkładu wariancja, błąd średniokwadratowy, a także obciążenie były średnio największe (poza obciążeniem dla $n = 100$). Zatem, choć rozwiązanie da się uzyskać w tym przypadku szybciej, jest ono nieco mniej dokładne.
Podsumowując, w przypadku rozkładu Cauchy'ego liczenie ENW metodą Newtona-Rhapsona przy wyborze mediany z próby jako punktu początkowego wydaje się dobrym rozwiązaniem.