Raport 1

Magdalena Potok

2023-10-23

Zadanie 1

W tym zadaniu wygenerowałam 50 obserwacji z rozkładu $N(\theta, \sigma^2)$, gdzie
(a) $\theta = 1$, $\sigma = 1$,
(b) $\theta = 4$, $\sigma = 1$,
(c) $\theta = 1$, $\sigma = 2$.
Dla każdego z rozkładów policzyłam cztery różne estymatory parametru θ.
(i) $\hat{\theta_1} = \bar{X}$,
(ii) $\hat{\theta_2} = Me\{X_1, ..., X_n\}$,
(iii) $\hat{\theta_3} = \sum_{i=1}^{n} w_i X_i, \sum_{i=1}^{n} w_i = 1, 0 \leq w_i \leq 1$, gdzie w_i to ciąg liczb zaczynających się od 0.02 i zwiększających się o 0.02. W ten sposób cały wektor sumuje się do 1 i każda waga jest z przedziału $(0, 1)$,
(iv) $\hat{\theta_4} = \sum_{i=1}^{n} w_i X_{i:n}$, gdzie $X_{i:n}$ to uporządkowane obserwacje X_i oraz $w_i = \varphi(\Phi^{-1}(\frac{i-1}{n})) - \varphi(\Phi^{-1}(\frac{i}{n}))$, φ i Φ to kolejno gęstość oraz dystrybuanta rozkładu normalnego.
To doświadczenie zostało powtórzone R = 10 000 razy i na tej podstawie została oszacowana:
- **wariancja**: $Var_j = \frac{1}{R} \sum_{i=1}^{R} (\hat{\theta_j} - \hat{\theta_i})^2$,
- **błąd średniokwadratowy**: $MSE_j = \frac{1}{R} \sum_{i=1}^{R} (\theta - \hat{\theta_i})^2$,
- **obciążenie estymatorów**: $Bias = \frac{1}{R} \sum_{i=1}^{R} \hat{\theta_i} - \theta$.
Wyniki dla każdego z podpunktów (a)-(c) zostaną przedstawione w postaci tabeli.

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

	MSE	Var	Bias
$\hat{\theta_1}$	0.02011	0.02011	0.00040
$\hat{\theta_2}$	0.03101	0.03101	-0.00006
$\hat{\theta_3}$	0.02631	0.02631	-0.00043
$\hat{\theta_4}$	0.01063	0.00966	-0.03104

Można zawuażyć, że najniższym wynikiem w tabeli jest $Var\ \hat{\theta_4}$, ale ten parametr ma również największe (co do modułu) obciążenie estymatora. Wynik jest ujemny, co oznacza, że średnia estymacja jest wyższa od prawdziwej wartości parametru. Patrząc na pozostałę estymatory najlepszym wyborem jest $\hat{\theta_1}$, ponieważ ma najniższy błąd średniokwadratowy i wariancję, co wskazuje na dobrą jakość estymacji.

- (b) $\theta = 4$, $\sigma = 1$

	MSE	Var	Bias
$\hat{\theta_1}$	0.02000	0.02000	-0.00038
$\hat{\theta_2}$	0.03002	0.03002	-0.00083
$\hat{\theta_3}$	0.02645	0.02645	-0.00142
$\hat{\theta_4}$	9.18752	0.00981	-3.02947

Najbardziej rzucającym się w oczy wynikiem dla tych parametrów jest wartość błędu średniokwadratowego dla $\hat{\theta_4}$, co sugeruje, że ten estymator jest nieprecyzyjny i ma tendencje do dużego rozrzutu wokół prawdziwej wartości θ. Niska wariancja mówi nam, że wyniki są skoncentrowane wokół swojej średniej wartości, ale ta średnia wartość jest oddalona od prawdziwej wartości parametru.
Najniższe wyniki ponownie należą do $\hat{\theta_1}$ i to właśnie on najlepiej estymuje parametr θ dla tego rozkładu.

- (c) $\theta = 1$, $\sigma = 2$

	MSE	Var	Bias
$\hat{\theta_1}$	0.08056	0.08056	0.00062
$\hat{\theta_2}$	0.12412	0.12412	0.00170
$\hat{\theta_3}$	0.10546	0.10544	0.00426
$\hat{\theta_4}$	0.91782	0.03815	0.93791

Wartości wszystkich parametrów, dla każdejgo $\hat{\theta_i}$ co do modułu wzrosły. Jest to spowodowane większą σ tego rozkładu porównując do poprzednich tabelek. Większy rozrzut obserwacji dla rozkładu $N(\theta, \sigma^2)$ oznacza spadek dokładności podanych estymatorów θ.

Zadanie 2

W języku R komenda

```
set.seed(1)
```

służy do inicjalizacji generatora liczb pseudolosowych. Oznacza to, że ziarno generatora ustawione jest na konkretną wartość (w tym przypadku 1), co sprawia, że będziemy otrzymywać te same ciągi liczb losowych. Ta funkcja ma wiele zastosowań, przede wszystkim pozwala nam powtarzać wyniki symulacji, odtwarzać je na innych komputerach lub kontynuować obliczenia w wyniku wystąpienia jakiegoś błędu.

Zadanie 3

Estymatory największej wiarogodności (MLE) są używane do oszacowania parametrów statystycznych. W niektórych przypadkach obliczenie go analitycznie może być trudne lub niemożliwe, w takim przypadku pomagają metody numeryczne. Dobrym tego przykładem jest rozkład logistyczny z gęstością $f(x;\theta) = \frac{e^{\frac{-(x-\theta)}{\sigma}}}{(1+e^{\frac{-(x-\theta)}{\sigma}})^2}, -\infty < x, \theta < \infty$.
Chcemy znaleźć MLE tego rozkładu, liczymy funkcję logwiarogodności, która jest sumą logarytmów z gęstości prawdopodobieństwa dla wszystkich próbek

$$l(\theta) = \sum_{i=1}^{n} log f(x_i, \theta) = \frac{n\theta}{\sigma} - \frac{n\bar{x}}{\sigma} - 2\sum_{i=1}^{n} log(1 + e^{\frac{-(x_i-\theta)}{\sigma}}).$$

.
Następnie szukając największej wartości liczymy pochodną

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

przyrównując to do 0 wychodzi

$$\sum_{i=1}^{n} \frac{e^{\frac{-(x_i-\theta)}{\sigma}}}{1 + e^{\frac{-(x_i-\theta)}{\sigma}}} = \frac{n}{2}.$$

Jednak rozwiązanie tego i znalezienie takiego θ jest czasochłonne i bardzo trudne, wtedy szukamy innych metod wyznaczających estymatory największej wiarogodności, najczęściej są to metody numeryczne. Jednak

najpierw przekonajmy się, że rozwiązanie dla tego rozkładu posiada rozwiązanie jednoznaczne.

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

Druga pochodna jest ujemna, zatem rozwiązanie jest jednoznaczne. Do wyznaczenia estymatora dla tego rozkładu użyję Metody Newtona w następnych zadaniach.

Zadanie 4

Jedną z metod numerycznych wyznaczania estymatora największej wiarogodności jest Metoda Newtona. Jest to algorytm iteracyjny służący do znalezienia przybliżonego miejsca zerowego funkcji. Działa na zasadzie iteracyjnego poprawiania przybliżonego rozwiązania do momentu, aż wartość będzie nas zadowalać.
W kontekście rozkładu logistycznego chcemy znaleźć gdzie zeruje się pierwsza pochodna funkcji log-wiarogodności $l'(\theta) = \frac{n}{\sigma} - \frac{2}{\sigma}\sum_{i=1}^{n}\frac{e^{\frac{-(x_i-\theta)}{\sigma}}}{1+e^{\frac{-(x_i-\theta)}{\sigma}}}$ i potrzebna nam jest jeszcze druga pochodna, czyli $\frac{-2}{\sigma^2}\sum_{i=1}^{n} -\frac{e^{\frac{-(x_i-\theta)}{\sigma}}}{(1+e^{\frac{-(x_i-\theta)}{\sigma}})^2}$. Następnie musimy wybrać początek naszego zgadywania - θ_0, co często jest problematyczne dla tej metody. Następnie szukamy $\theta_1 = \theta_0 - \frac{l'(\theta_0)}{l''(\theta_0)}$, który jest lepszym przybliżeniem θ niż θ_0. Ten proces cały czas jest powtarzany, tzn. $\theta_n = \theta_{n-1} - \frac{l'(\theta_{n-1})}{l''(\theta_{n-1})}$, aż do momentu gdy uzyskany wynik nas zadowala i różnica między kolejnymi przybliżeniami jest wystarczająco mała, tj. $|\theta_n - \theta_{n-1}| < \epsilon$, gdzie ϵ to liczba bliska 0, która określa nam poziom tolerancji błędu. Gdy różnica między kolejnymi przybliżeniami stanie się dostatecznie mała lub gdy wartość $l'(\theta_n)$ jest już blisko zera, to θ_n jest naszym estymatorem wiarogoności dla parametru θ.
Warto zaznaczyć, że ta metoda wymaga ostrożności oraz odpowiedniej inicjalizacji (pierwszy strzał θ_0), ponieważ może doporwadzić do błędu w przypadku nieodpowiedniego początkowego przybliżenia.

Zadanie 5

W tym zadaniu zostały wygenerowanych 50 obserwacji z trzech rozkładów logistycznych $L(\theta, \sigma)$ z parametrem przesunięcia θ oraz skali σ, gdzie:
(a) $\theta = 1, \sigma = 1$,
(b) $\theta = 4, \sigma = 1$,
(c) $\theta = 1, \sigma = 2$.
Następnie wartość estymatora największej wiarogodności parametru θ został oszacowany przy pomocy Matody Newtona opisanej w poprzednich zadaniach. Jako wybór punktu początkowego wybrałam średnią próbkową rozkładów, ponieważ jest to dobry estymator tego parametru dla tego rozkładu.

Rozkład	Przyblizenie	Kroki
L(1,1)	1.1839928	2
L(4,1)	4.0171092	2
L(1,2)	0.9997143	2

Liczba kroków w algorytmie nie jest duża, mimo że tolerancję błędu wybrałam naprawdę małą, bo $\epsilon = 0.00001$. Jest to spowodowane dobrym wyborem θ_0 jako średnią całego rozkładu (mean(rlogis(50,θ,σ))).

Powyższe doświadczenie zostało powtórzone 10 000 razy, na tej podstawie oszacowałam wariancję, błąd średniokwadratowy oraz obciążenie estymatora.

Rozkład	Var	MSE	Bias
L(1,1)	0.05968	0.05969	-0.00181
L(4,1)	0.05987	0.05987	0.00059
L(1,2)	0.24113	0.24119	-0.00775

Z tabelki można zauważyć, że wartość ENW dla każdego z rozkładu jest bliska prawdziwej wartości, co świadczy o poprawności wybranej metody numerycznej. Najgorszy wynik wychodzi dla rozkładu L(1,2), jest to spowodowane tym, że w tym rozkładzie jest największa wartość wariancji, co sprawia, że wyniki są bardziej rozproszone i utrudnia to estymację.

Zadanie 6

Wygenerowałam 50 obserwacji z rozkładów Cauchy'ego $C(\theta, \sigma)$ z parametrami przesunięcia θ i skali σ, gdzie:
(a) $\theta = 1, \sigma = 1$,
(b) $\theta = 4, \sigma = 1$,
(c) $\theta = 1, \sigma = 2$.
Następnie oszacowałam wartość estymatora największej wiarogodności parametru θ na podstawie wygenerowanych prób. Do uzyskania ENW zastosowałam ponownie Metodę Newtona i jako punkt początkowy wybrałam średnią próbkową rozkładu.
Do wykorzystania Metody Newtona potrzebna nam jest gętość rozkładu Cauchy'ego:

$$f(x, \theta) = \frac{1}{\pi\sigma(1 + (\frac{x-\theta}{\sigma})^2)}$$

Następnie liczymy funkcję logwiarogodności: $l(\theta) = -n\log(\pi\sigma) - \sum_{i=1}^{n} \log\left(1 + \left(\frac{x_i-\theta}{\sigma}\right)^2\right)$ oraz liczymy pierwszą pochodną $l'(\theta) = \frac{2}{\sigma^2} \sum_{i=1}^{n} \frac{x_i-\theta}{1+\left(\frac{x_i-\theta}{\sigma}\right)^2}$ i drugą pochodną $\frac{d^2\ell(\theta)}{d\theta^2} = \frac{-2}{\sigma^2} \sum_{i=1}^{n} \frac{1-\left(\frac{x_i-\theta}{\sigma}\right)^2}{\left(1+\left(\frac{x_i-\theta}{\sigma}\right)^2\right)^2}$.

Rozkład	Przyblizenie	Kroki
C(1,1)	1.216249	2
C(4,1)	3.807656	2
C(1,2)	1.152295	2

Dla pojedynczego doświadczenia otrzymujemy wyniki bliskie prawdziwej wartości θ przy małej ilości kroków, bo tylko dla 2. Mała ilość kroków ponownie spowodowana jest dobrym doborem θ_0, czyli średniej całego rozkładu.

Następnie powyższe doświadczenie powtórzyłam 10 000 razy i oszacowałam wariancję, błąd średniokwadratowy oraz obciążenie estymatora, wyniki przedstawiłam w poniższej tabelce.

Rozkład	Var	MSE	Bias
C(1,1)	0.04309	0.04309	0.00116
C(4,1)	0.04094	0.04094	-0.00062
C(1,2)	0.16736	0.16736	-0.00114

Można zauważyć, że wyniki dla C(1,1) oraz C(4,1) są bardzo niskie, oznacza to, że otrzymaliśmy dobrze przybliżony ENW, co by oznaczało, że wielkość θ nie ma wpływu na wynik. Nie można tego samego powiedzieć o rozkładzie C(1,2), ponieważ tu widać już znacznie większą wariancję oraz MSE, co by oznaczało, że przybliżenie ENW dla tego rozkładu jest mniej dokładne. Wnioski nasuwające się po porównaniu tych wyników są takie, że niewiele większa wariancja wpływa negatywnie na poprawność estymacji tego parametru.

Zadanie 7

Powtórzę eksperymenty numeryczne z zadań 1, 5 i 6 dla n = 20 i n = 100.

Zadanie 1'

- n = 20

$N(1,1)$

	MSE	Var	Bias
$\hat{\theta_1}$	0.04968	0.04967	-0.00251
$\hat{\theta_2}$	0.07369	0.07368	-0.00214
$\hat{\theta_3}$	0.06406	0.06405	-0.00415
$\hat{\theta_4}$	0.02794	0.02317	-0.06907

$N(4,1)$

	MSE	Var	Bias
$\hat{\theta_1}$	0.05035	0.05034	0.00380
$\hat{\theta_2}$	0.07401	0.07398	0.00514
$\hat{\theta_3}$	0.06586	0.06584	0.00345
$\hat{\theta_4}$	9.43465	0.02267	-3.06790

$N(1,4)$

	MSE	Var	Bias
$\hat{\theta_1}$	0.80943	0.80943	-0.00233
$\hat{\theta_2}$	1.16954	1.16954	-0.00163
$\hat{\theta_3}$	1.05381	1.05377	0.00589
$\hat{\theta_4}$	7.80184	0.37514	2.72520

Uzyskane wyniki dla n = 20 są gorsze niż dla n = 50 z 1. zadania. Każda z wartości, co do modułu, wzrosła w każdym z przypadków. Oznacza to, że wraz z zmniejszeniem rozmiaru próby zmalała poprawność estymatorów.

- n = 100

$N(1,1)$

	MSE	Var	Bias
$\hat{\theta_1}$	0.01023	0.01023	-0.00011
$\hat{\theta_2}$	0.01540	0.01540	0.00053
$\hat{\theta_3}$	0.01357	0.01357	-0.00084

	MSE	Var	Bias
$\hat{\theta_4}$	0.00516	0.00494	-0.01468

$N(4, 1)$

	MSE	Var	Bias
$\hat{\theta_1}$	0.01013	0.01013	-0.00061
$\hat{\theta_2}$	0.01579	0.01579	-0.00033
$\hat{\theta_3}$	0.01347	0.01347	-0.00012
$\hat{\theta_4}$	9.09181	0.00501	-3.01443

$N(1, 4)$

	MSE	Var	Bias
$\hat{\theta_1}$	0.15947	0.15944	0.00534
$\hat{\theta_2}$	0.24744	0.24739	0.00685
$\hat{\theta_3}$	0.21388	0.21388	0.00243
$\hat{\theta_4}$	8.73985	0.07959	2.94283

Tym razem wyniki się poprawiły. Wartości błędu średniokwadratowego, wariancji oraz obciążenia, co do modułu, są niższe. Oznacza to, że nasza estymacja się poprawiła. Potwierdza nam ten przykład to, co napisałam wyżej, czyli im większa próba, tym bardziej dokładna estymacja. Estymator $\hat{\theta_4}$ dalej pozostaje złym wyborem estymatora parametru θ.

Zadanie 5'

- n = 20

Rozkład	Przyblizenie	Kroki
L(1,1)	0.6494422	2
L(4,1)	3.7319163	1
L(1,2)	1.3490105	2

Otrzymane estymacje są gorsze w przypadku próby n = 20. Ponownie nasuwa nam się wniosek, że im mniejsza próbka, tym dokładność przybliżenia maleje.

Rozkład	Var	MSE	Bias
L(1,1)	0.15205	0.15208	0.00543
L(4,1)	0.15352	0.15353	0.00227
L(1,2)	0.59719	0.59723	0.00657

Ponownie można zauważyć, że wyniki wariancji, błędu średniokwadratowego i obciążenia estymatora są gorsze (tj. większe co do modułu). Wraz ze zmniejszeniem próby zmalała dokładność estymacji.

- n = 100

Rozkład	Przyblizenie	Kroki
L(1,1)	0.7074250	2
L(4,1)	3.7684227	1
L(1,2)	0.6764201	2

Tym razem dla większej próbki otrzymujemy gorsze wyniki, niż dla próby n = 50. Jednak było to jedynie pojedyńcze doświadczenie, poniżej pokażę tabelkę przy 10 000 powtórzeniach.

Rozkład	Var	MSE	Bias
L(1,1)	0.03011	0.03011	0.00043
L(4,1)	0.03036	0.03036	-0.00066
L(1,2)	0.11624	0.11626	0.00499

Porównując wyniki z wartościami dla n = 50 widać, że wyniki ponownie są lepsze (tj. mniejsze co do modułu), znaczy to tyle, że większa próbka zapewnia nam lepszą estymację przy wielokrotnym powtórzeniu doświadczenia.

Zadanie 6'

- n = 20

Rozkład	Przyblizenie	Kroki
C(1,1)	0.4946939	2
C(4,1)	4.2323380	2
C(1,2)	1.6153361	2

Rzuca się w oczy wartość ENW dla C(1,2), ponieważ jest ona zdecydowanie mniej dokładna, niż dla n = 50, mimo że ilość kroków jest ta sama. W pozostałych przypadkach różnica nie jest tak duża. Jest to spowodowane większą wariancją, więc i większym rozrzutem obserwacji, ciężej estymować taki rozkład, szczególnie przy jednokrotnym próbkowaniu.

Rozkład	Var	MSE	Bias
C(1,1)	22420454427	22428585405	2851.4869
C(4,1)	24826958090	24832152588	-2279.1442
C(1,2)	47464881376	47464915910	-185.8327

Dla tego przypadku wyniki wychodzą zaskakująco ekstremalne, widać, że przybliżanie tego rozkładu Metodą Newtona zdecydowanie nie jest najlepszą opcją.

- n = 100

Rozkład	Przyblizenie	Kroki
C(1,1)	0.8042166	2
C(4,1)	3.9642819	2
C(1,2)	1.0022669	2

Już przy jednokrotnym próbkowaniu można zauważyć znaczącą poprawę ENW (tj. bliższe prawdziwej wartości) dla n = 100 niż dla n = 50.

Rozkład	Var	MSE	Bias
C(1,1)	0.02016	0.02016	0.00044
C(4,1)	0.02116	0.02116	-0.00008
C(1,2)	0.08299	0.08302	-0.00515

Porównując powyższą tabelę z tabelą z zadania 6. możemy zaobserwować zmniejszenie się wartości wariancji, błędu średniokwadratowego i obciązenia estymatowa w każdym z rozkładów. Oznacza to, że dla większej próby mamy dokładniejszą estymację.