Raport pierwszej listy z laboratorium Statystyki

Kamil Zdancewicz

October 23, 2025

Zadanie 1

Cel

Celem zadania było oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia za pomocą czterech estymatorów przesunięcia rozkładu normalnego $N(\theta, \sigma^2)$ dla różnych wartości θ i σ.

Stosowane metody

Dla każdego podpunktu (a) - (c) wygenerowałem 10 000 razy próbę wielkości $n = 50$. Dla każdej z tych prób obliczyłem cztery estymatory $\hat{\theta}$:

(i) $\hat{\theta}_1$ - średnia arytmetyczna próby

(ii) $\hat{\theta}_2$ - mediana próby

(iii) $\hat{\theta}_3$ - eksperymentalny nieobciążony estymator liniowy z losowymi wagami

(iv) $\hat{\theta}_4$ - potencjalnie obciążony estymator ważony sumy elementów próby, gdzie wagi są obliczane z pomocą funkcji gęstości i kwantyli

Następnie oszacowałem wariancję, błąd średniokwadratowy oraz obciążenie dla każdego z estymatorów we wszystkich podpunktach.

Wyniki

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

Estymator	Wariancja	MSE	Obciążenie
$\hat{\theta}_1$	0.019978	0.019977	0.000981
$\hat{\theta}_2$	0.030751	0.030748	0.000603
$\hat{\theta}_3$	0.026960	0.026958	0.000884
$\hat{\theta}_4$	0.009663	0.010491	-0.028793

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

Estymator	Wariancja	MSE	Obciążenie
$\hat{\theta}_1$	0.019811	0.019811	-0.001252
$\hat{\theta}_2$	0.030296	0.030301	-0.002941
$\hat{\theta}_3$	0.025945	0.025945	-0.001518
$\hat{\theta}_4$	0.009848	**9.180373**	**-3.028288**

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

Estymator	Wariancja	MSE	Obciążenie
$\hat{\theta}_1$	0.080583	0.080576	0.001082
$\hat{\theta}_2$	0.121562	0.121565	0.003831
$\hat{\theta}_3$	0.107206	0.107196	0.000555
$\hat{\theta}_4$	0.039360	**0.925652**	**0.941433**

Wnioski

- We wszystkich podpunktach estymatory (i) - (iii) mają bardzo małe obciążenie.
- Estymator (i) ma konsekwentnie najmniejszą wariancję oraz błąd średniokwadratowy.
- We wszystkich przypadkach oprócz (iv) wariancje i błędy średniokwadratowe wzrastają kwadratowo do wzrostu σ, co jest zgodne z teorią.
- Po podpunktach (b) i (c) widać, że estymator (iv) jest obciążony (bo bazuje na $N(0,1)$), co skutkuje dużymi błędami średniokwadratowymi, jednak wariancja jest najmniejsza z wszystkich estymatorów.

Załączniki

Generacja danych, obliczanie estymatorów i dalsza logika znajduje się w załączonym pliku `main.cpp`.

Zadanie 2

Cel

Celem zadania jest opisać komendę `set.seed(1)` z języka R (oraz jej odpowiedniki w innych językach programowania), oraz jej potencjalne zastosowania.

Omówienie

W komputerach poza korzystaniem z urządzeń peryferyjnych niemożliwe jest generowanie czysto losowych liczb. Komputery używają więc generatorów liczb **pseudolosowych**. Dają one iluzję generowania losowych wartości, lecz w rzeczywistości są **deterministyczne**, ponieważ są wyznaczone przez pewien algorytm. Sposobem zarządzania punktem startowym danego algorytmu generatora jest ustawienie tzw. ziarna *(seed)*. Komenda `set.seed(1)` ustawia ziarno generatora na wartość 1, choć może to być dowolna wartość przyjmowana przez tę komendę.

Najważniejszą funkcjonalnością dostarczaną przez ustawienie ziarna jest reprodukowalność wyników generatora zaczynającego od tego samego ziarna, przy każdym uruchomieniu kodu wylosowane liczby będą **takie same**. Ustawienie tego samego ziarna pozwala porównywać wyniki dwóch różnych metod na tych samych danych losowych w innych czasach (np. dwóch estymatorów MLE generujących próbki z tego samego rozkładu).

Zadanie 3

Celem zadania jest omówienie przyczyn, dla których w przypadku rozkładu logistycznego nie da się analitycznie wyznaczyć estymatora największej wiarygodności (MLE) parametru przesunięcia θ, oraz uzasadnienie potrzeby stosowania metod numerycznych w tym celu.

Omówienie

Analizę pełnego przebiegu estymacji przedstawiono szczegółowo w Stosowanych metodach zadania 5, gdzie wyprowadzono logarytm funkcji wiarygodności oraz jej pochodne względem parametru θ:

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

Równanie to nie posiada rozwiązania w postaci zamkniętej, ponieważ θ występuje w sposób nieliniowy we wszystkich składnikach sumy - nie da się go algebraicznie wyłączyć ani uprościć.

Z tego powodu estymacja MLE wymaga metod numerycznych, takich jak użyta przeze mnie metoda Brenta.

W zadaniu 5 wykorzystano tę metodę do znalezienia maksimum funkcji log-wiarygodności i wykazano, że ma ona jednoznaczne rozwiązanie, co gwarantuje istnienie i jednoznaczność estymatora.

Zadanie 4

Celem zadania jest omówienie wybranej metody numerycznej umożliwiającej wyznaczenie estymatora największej wiarogodności oraz przedstawienie zasad jej działania.

Omówienie

Głównym celem użycia metod numerycznych przy wyznaczaniu estymatora największej wiarogodności jest znalezienie **miejsc zerowych** funkcji log-wiarogodności aby znaleźć ekstrema. Nie wszystkie metody wyznaczają konkretnie maksimum/minimum oraz ustalają jednoznaczność, jednak znalezienie przybliżonego punktu znacznie ułatwia analizę.

Do omówienia wybrałem **metodę regula falsi**, znajduje ona miejsce zerowe podanej funkcji f ciągłej z jedną niewiadomą.

Główną ideą jest iteracyjne zbliżanie się do miejsca zerowego z dwóch stron (lewej a i prawej b), i założeniu, że funkcja ciągła na dostatecznie małym odcinku w przybliżeniu zmienia się w sposób liniowy oraz pierwsza i druga pochodna istnieją i mają na tym przedziale stałe znaki. Możemy wtedy na odcinku $\langle a, b\rangle$ krzywą $y = f(\theta)$ zastąpić sieczną. Za przybliżoną wartość pierwiastka przyjmujemy punkt przecięcia siecznej z osią $O\theta$.

Iteracyjnie wyznaczamy ciągi a_n, b_n takich, że $a_n \geq a_{n-1} \geq \cdots \geq a, b_n \leq b_{n-1} \leq \cdots \leq b$, co najmniej jeden z a_n i b_n jest bliżej pierwiastka po każdym kroku i dla $\forall_{n=1,2,\ldots} f(a_n)f(b_n) < 0$. Oczywiście jeśli znajdziemy pieriastek możemy przestać iterować, w przeciwnym wypadku iterujemy, aż do warunku stopu (np. a_n i b_n są blisko lub ($f(a_n)$ albo $f(b_n)) < \epsilon$)

Ciągi tworzą przedziały $[a_n, b_n]$ będące ciągiem zstępującym. A są zdefiniowane w następujący sposób:

$$\begin{cases} a_0 = a \\ b_0 = b \\ x = \frac{f(a_n)b_n - f(b_n)a_n}{f(a_n) - f(b_n)} \\ \text{Jeżeli } f(x)f(a_n) > 0 \text{ to } a_{n+1} = x, b_{n+1} = b_n \\ \text{Jeżeli } f(x)f(b_n) > 0 \text{ to } a_{n+1} = a_n, b_{n+1} = x \end{cases}$$

Polega ona na wyznaczeniu punktu przecięcia siecznej przechodzącej przez punkty $(a_n, f(a_n))$ i $(b_n, f(b_n))$ z osią $O\theta$ i podmiany punktu końcowego przedziału, którego znak funkcji jest taki sam jak w x (aby zachować $f(a_n)f(b_n) < 0$).

Zadanie 5

Cel

Celem zadania było oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia za pomocą estymatora największej wiarygodności dla rozkładu logistycznego $L(\theta, \sigma)$ dla różnych wartości θ i σ.

Stosowane metody

Dla każdego podpunktu (a) - (c) wygenerowałem 10 000 razy próbę wielkości $n = 50$. Dla każdej z tych prób obliczyłem estymator największej wiarygodności, który został wyznaczony w następujący sposób:

Ponieważ funkcja gęstości rozkładu logistycznego to:

$$f(x; \theta, \sigma) = \frac{exp\{-(x-\theta)/\sigma\}}{(1 + exp\{-(x-\theta)/\sigma\})^2}, -\infty < x < \infty, -\infty < \theta < \infty, 0 < \sigma < \infty$$

Logarytm funkcji wiarogodności dla próby wielkości n to:

$$l(\theta,\sigma)=\sum_{i=1}^{n} logf(x_i;\theta,\sigma)=n\theta/\sigma-n\overline{x}/\sigma-2\sum_{i=1}^{n} log(1+exp\{-(x_i-\theta)/\sigma\})$$

Następnie metodą Brenta znajdujemy maksimum tej funkcji. Ale aby udowodnić, że jest to jedyne rozwiązanie i maksiumum, obliczamy pochodną po θ, biorąc pochodną cząstkową:

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

Po przyrównaniu do zera i przekształceniu:

$$\sum_{i=1}^{n}\frac{exp\{-(x_i-\theta)/\sigma\}}{1+exp\{-(x_i-\theta)/\sigma\}}=\frac{n}{2\sigma} \qquad (*)$$

Biorąc pochodną lewej części równania (*):

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

więc lewa strona równania (*) jest ściśle rosnąca. Dąży do 0 gdy $\theta\to-\infty$ oraz dąży do n gdy $\theta\to\infty$. Zatem (*) ma jednoznaczne rozwiązanie. Dodatkowo możemy zauważyć, że druga pochodna $l(\theta)$ jest ujemna dla wszystkich θ, więc rozwiązanie jest maksimum.

Otrzymawszy sposób na znalezienie mle oszacowałem wariancję, błąd średniokwadratowy oraz obciążenie dla tego estymatora we wszystkich podpunktach.

Wyniki

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.060253	0.060249	-0.001438

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.059736	0.059736	0.002526

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.242051	0.242036	-0.002956

Wnioski

- We wszystkich podpunktach obciążenie jest bardzo małe, co sugeruje że estymator jest nieobciążony
- W podpunkcie (c) widać kwadratowy wzrost wariancji i błędu średniokwadratowego względem zwiększenia σ, co jest zgodne z wariancją rozkładu logistycznego

$$\text{Dla } X \sim L(\theta, \sigma), Var(\hat{\theta}) = \frac{Var(X)}{n}, \text{Gdzie } Var(X) = \frac{\sigma^2\pi^2}{3}$$

- Wariancja i MSE nie są zauważalnie zależne od θ

Załączniki

Generacja danych, obliczanie estymatora i dalsza logika znajduje się w załączonym pliku `main.cpp`.

Zadanie 6

Cel

Celem zadania było oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia za pomocą estymatora największej wiarygodności dla rozkładu Cauchy'ego $C(\theta, \sigma)$ dla różnych wartości θ i σ.

Stosowane metody

Dla każdego podpunktu (a) - (c) wygenerowałem 10 000 razy próbę wielkości $n = 50$. Dla każdej z tych prób obliczyłem estymator największej wiarygodności, który został wyznaczony w następujący sposób:

Ponieważ funkcja gęstości rozkładu Cauchy'ego to:

$$f(x; \theta, \sigma) = \frac{1}{\pi\sigma(1 + (\frac{x-\theta}{\sigma})^2)}, -\infty < x < \infty, -\infty < \theta < \infty, 0 < \sigma < \infty$$

Logarytm funkcji wiarogodności dla próby wielkości n to:

$$l(\theta, \sigma) = \sum_{i=1}^{n} logf(x_i; \theta, \sigma) = \sum_{i=1}^{n} [-ln(\pi) - ln(\sigma) - ln(1 + \left(\frac{x_i - \theta}{\sigma}\right)^2)]$$

Ponownie używamy metody Brenta, znajdujemy maksimum tej funkcji. Ponownie musimy udowodnić, że jest to jedyne rozwiązanie i maksiumum, obliczamy pochodną po θ, uznajemy σ za znaną, biorąc pochodną częściową:

$$l'(\theta) = (\text{const} - \sum_{i=1}^{n} ln(1 + \left(\frac{x_i - \theta}{\sigma}\right)^2)) = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{\sigma^2 + (x_i + \theta)}$$

Następnie druga pochodna

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

Problemem jest to, że druga pochodna log-wiarygodności może zmieniać znak w zależności od θ, więc nie możemy zagwarantować, że log-wiarygodność jest wypukła w dół. Lecz z pochodnej widać, że ekstrema istnieją, jednak metoda Brenta wokół mediany rozróżnia między maksimum i minimum, znajdując maksimum na podanym przedziale. A ponieważ mamy silne podejrzenia, co do miejsca globalnego maksimum, to w tym zadaniu umiemy stwierdzić, że ta metoda znajduje globalne maksimum.

Następnie oszacowałem wariancję, błąd średniokwadratowy oraz obciążenie dla tego estymatora we wszystkich podpunktach.

Wyniki

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.041905	0.041901	0.000785

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.042916	0.042934	0.004731

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.164951	0.164952	-0.004287

Wnioski

- We wszystkich podpunktach obciążenie jest bardzo małe, co sugeruje że estymator jest nieobciążony
- Wariancja i MSE zwiększają się około kwadratowo względem zwiększenia σ, lecz przez nieskończoność wariancji rozkładu cauchy'ego, jest to jedynie ciekawa obserwacja, nie potwierdzenie teorii.
- Wariancja i MSE nie są zauważalnie zależne od θ
- Pomimo ciężkich ogonów rozkładu cauchy'ego, MLE efektywnie estymuje wariancję, MSE i nie ma dużego obciążenia

Załączniki

Generacja danych, obliczanie estymatora i dalsza logika znajduje się w załączonym pliku `main.cpp`.

Zadanie 7

Cel

Przez zmianę wielkości próby na $n = 20$ albo $n = 100$ dla zadań 1,5 i 6 należy zauważyć różnicę między wynikami dla oryginalnej wielkości próby a zmodyfikowanej.

Wyniki dla $n = 20$

Zadanie 1

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

Estymator	Wariancja	MSE	Obciążenie
$\widehat{\theta}_1$	0.050286	0.050282	-0.001196
$\widehat{\theta}_2$	0.073266	0.073269	-0.003144
$\widehat{\theta}_3$	0.065291	0.065286	-0.001289

Estymator	Wariancja	MSE	Obciążenie
$\hat{\theta}_4$	0.023692	0.028423	-0.068799

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

Estymator	Wariancja	MSE	Obciążenie
$\hat{\theta}_1$	0.051298	0.051300	-0.002610
$\hat{\theta}_2$	0.074649	0.074646	-0.002121
$\hat{\theta}_3$	0.065986	0.065979	-0.000262
$\hat{\theta}_4$	0.023439	**9.444980**	**-3.069453**

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

Estymator	Wariancja	MSE	Obciążenie
$\hat{\theta}_1$	0.196073	0.196054	-0.000506
$\hat{\theta}_1$	0.292651	0.292623	-0.000765
$\hat{\theta}_1$	0.255682	0.255658	-0.001545
$\hat{\theta}_1$	0.092743	**0.834278**	**0.861130**

Zadanie 5

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.149437	0.149423	-0.001118

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.150906	0.150891	-0.000508

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.598632	0.598792	-0.014844

Zadanie 6

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.115662	0.115664	0.003608

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.114366	0.114425	0.008412

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.448184	0.448163	0.004836

Wyniki dla $n = 100$

Zadanie 1

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

Estymator	Wariancja	MSE	Obciążenie
$\widehat{\theta}_1$	0.010049	0.010049	0.001079
$\widehat{\theta}_2$	0.015423	0.015421	-0.000120
$\widehat{\theta}_3$	0.013253	0.013252	0.000925
$\widehat{\theta}_4$	0.004939	0.005185	-0.015676

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

Estymator	Wariancja	MSE	Obciążenie
$\widehat{\theta}_1$	0.009928	0.009929	0.001212
$\widehat{\theta}_2$	0.015462	0.015466	0.002356
$\widehat{\theta}_3$	0.013287	0.013286	0.000821
$\widehat{\theta}_4$	0.004849	**9.093935**	**-3.014811**

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

Estymator	Wariancja	MSE	Obciążenie
$\widehat{\theta}_1$	0.040376	0.040372	0.000059
$\widehat{\theta}_2$	0.061584	0.061581	-0.001761
$\widehat{\theta}_3$	0.054096	0.054091	-0.000131
$\widehat{\theta}_4$	0.019675	**0.963653**	**0.971586**

Zadanie 5

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.030107	0.030107	0.001824

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.030365	0.030363	-0.000813

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.122978	0.122998	0.005693

Zadanie 6

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.020963	0.020967	-0.002324

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.020450	0.020452	-0.001781

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

Estymator	Wariancja	MSE	Obciążenie
MLE	0.082006	0.082001	0.001775

Wnioski

Zadanie 1 We wszystkich estymatorach zauważalny jest spadek wariancji i błędu średniokwadratowego względem n.

Niezależnie od n, estymatory (i) - (iii) mają obciążenie bliskie 0, co jeszcze bardziej wskazuje na ich nieobciążoność.

Estymator (iv) nawet w próbie $n = 100$ dla podpunktów (b) i (c) utrzymuje duży błąd średniokwadratowy i obciążenie - nie jest odporny na parametryzację ani zmianę skali.

Zadanie 5 Dla $n = 100$ wariancje i błędy średniokwadratowe są około 2 razy mniejsze niż przy $n = 50$, oraz dla $n = 20$ są około 2.5 raza większe, co dobrze wskazuje na zachowanie MLE w rozkładzie logistycznym i wskazuje na liniową zależność.

Ponownie n zdaje się nie wpływać na obciążenie, więc estymator prawdopodobnie nie jest obciążony.

Zadanie 6 Tak samo jak w zadaniu 5, wielkość próby wskazuje na liniową zależność między wielkością próby a wariancją i błędem średniokwadratowym, jednak przy mniejszej próbie są one trochę większe niż wskazywałaby zależność liniowa. Może to być spowodowane ciężkimi ogonami.

Tak jak w poprzednich zadaniach n nie wpływa na obciążenie, więc estymator prawdopodobnie nie jest obciążony.

Źródła

- *Introduction to Mathematical Statistics (6th edition) - Hogg, McKean, Craig*