Modelowanie i Identyfikacja

Raport z laboratoriów 3,4 oraz 5

Jan Rybarz

15 kwietnia 2025

Spis treści

1 Laboratorium 3 – Estymacja parametrów rozkładów

1.1 Cel

Celem laboratorium było porównanie estymatorów wartości oczekiwanej i wariancji dla rozkładu normalnego i Cauchy'ego oraz analiza błędu empirycznego w zależności od liczby prób i symulacji.

1.2 Estymatory i wzory

Dla zbioru $X = \{X_1, X_2, ..., X_N\}$ rozkładu normalnego $N(\mu = 0.5, \sigma = 1.5)$:

$$\hat{\mu}_N = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{1}$$

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{n=1}^{N} (X_n - \hat{\mu}_N)^2 \tag{2}$$

$$\hat{S}_N^2 = \frac{1}{N-1} \sum_{n=1}^{N} (X_n - \hat{\mu}_N)^2 \tag{3}$$

1.3 Błąd empiryczny

Dla L niezależnych powtórzeń eksperymentu definiujemy:

$$\mathrm{Err}\{\hat{\mu}_N; \mu\} = \frac{1}{L} \sum_{l=1}^{L} (\hat{\mu}_N^{[l]} - \mu)^2 \tag{4}$$

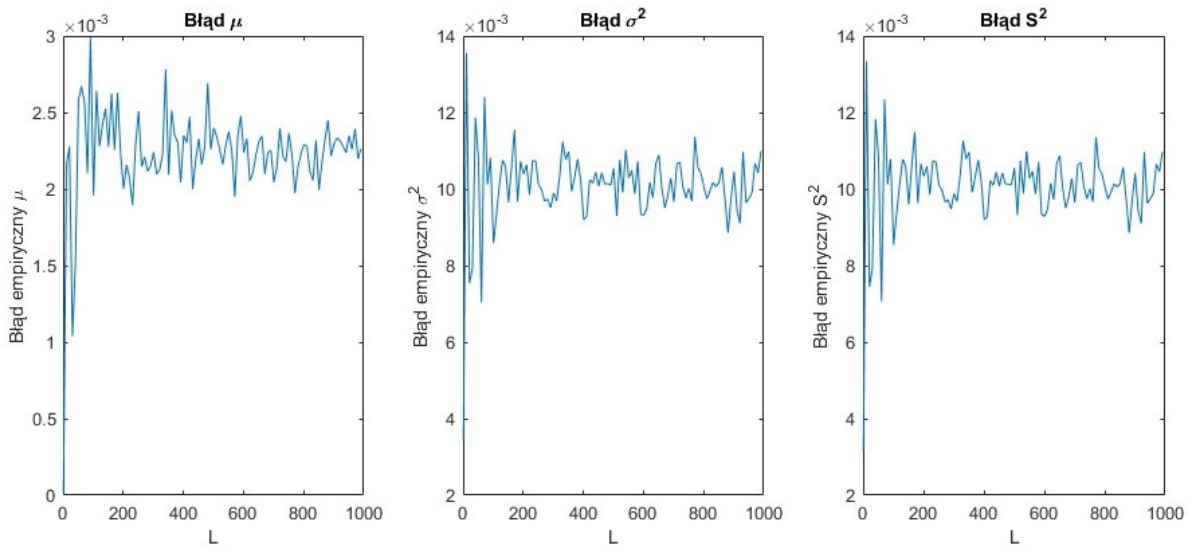
1.4 Wyniki i analiza

Parametry użyte do eksperymentu:

- Rozkład: normalny $N(0.5, 1.5^2)$
- Liczba próbek: $N = 1000$
- Liczba powtórzeń: $L = 1000$

Dla tak dobranych parametrów uzyskano następujące wartości estymatorów:

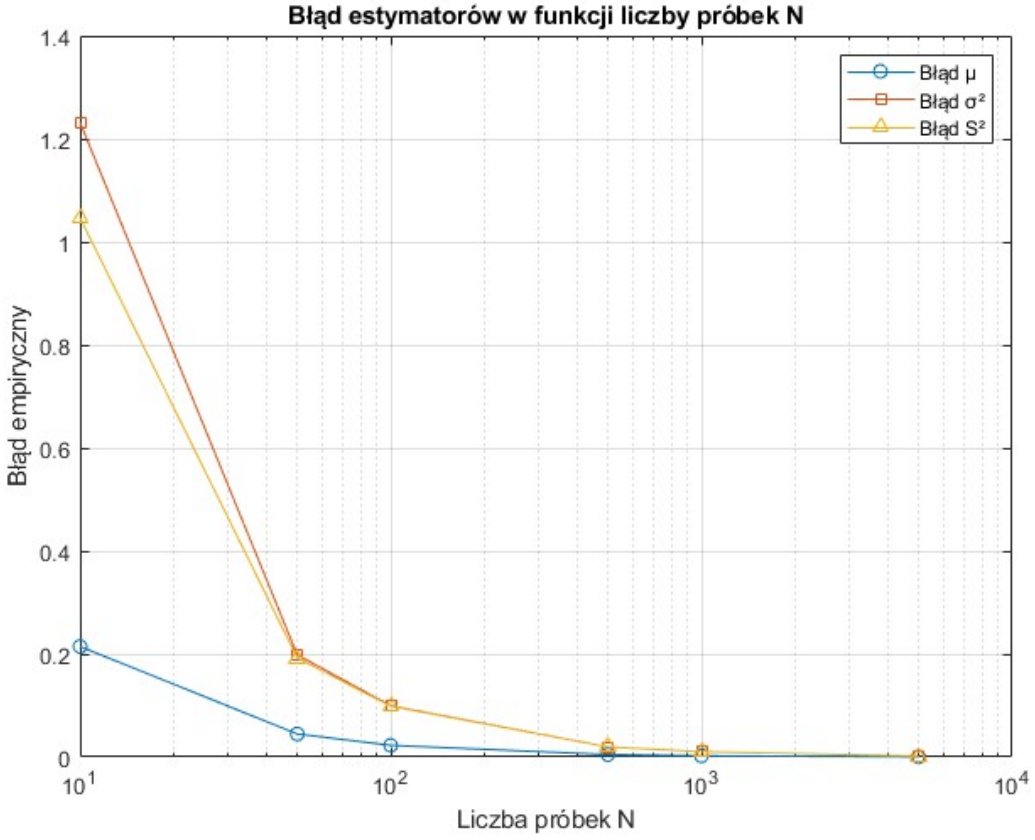
- Estymator średniej $\hat{\mu}_N = 0.4319$
- Estymator wariancji (populacyjny): 2.1728
- Estymator wariancji (próbowy): 2.1706
- Dla rozkładu Cauchy'ego: $\hat{\mu} = -1.3126$, $\hat{\sigma}^2 = 1967.7174$, $\hat{S}^2 = 1965.7497$



Rysunek 1: Błąd empiryczny estymatorów wartości oczekiwanej i wariancji w funkcji liczby powtórzeń L, dla $N = 1000$.

Wpływ liczby próbek N na błąd estymatorów

Aby zbadać wpływ liczby próbek N na dokładność estymacji parametrów, wykonano eksperyment dla różnych wartości N (od 10 do 5000) przy stałej liczbie powtórzeń $L = 1000$. Obliczono błąd empiryczny dla trzech estymatorów: wartości oczekiwanej $\hat{\mu}_N$, wariancji populacyjnej $\hat{\sigma}^2_N$ oraz wariancji próbki $\hat{S}^2_N$.



Rysunek 2: Błąd estymatorów w funkcji liczby próbek N (w skali logarytmicznej).

Z wykresu wynika, że wraz ze wzrostem liczby próbek N, błąd estymatorów gwałtownie maleje, co jest zgodne z teorią – większe próby prowadzą do dokładniejszych estymacji. Szczególnie widoczne jest to dla estymatora średniej, który szybciej stabilizuje się wokół wartości oczekiwanej. Estymatory wariancji wykazują większą niestabilność przy małych N, jednak także dążą do stabilizacji przy dużych próbach.

2 Laboratorium 4 – Dystrybuanta empiryczna

2.1 Cel

Zbadanie własności dystrybuanty empirycznej i porównanie jej z dystrybuantą teoretyczną, w tym wariancji estymacji.

2.2 Rozkład i dystrybuanta

Rozkład: $f(x) = 2x$ dla $x \in [0, 1]$

Dystrybuanta:

$$F(x) = \int_0^x 2t\,dt = x^2 \tag{5}$$

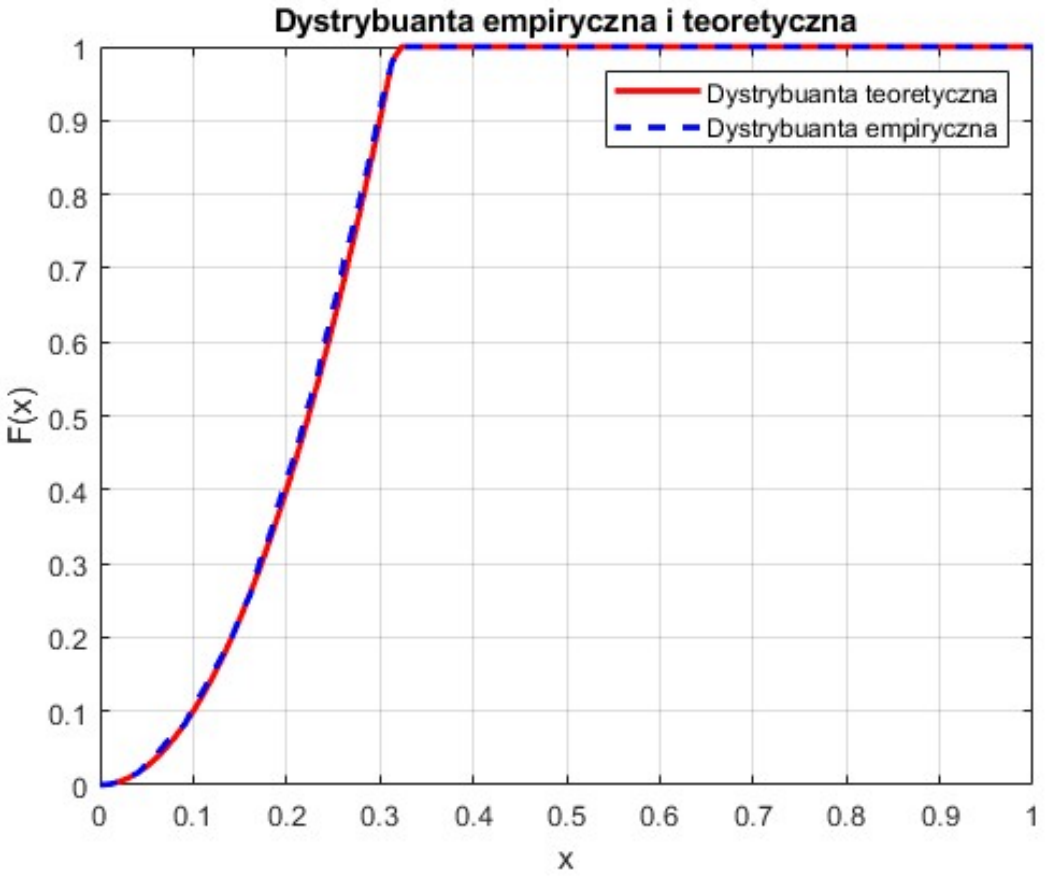
Odwrotna dystrybuanta:

$$F^{-1}(y) = \sqrt{y} \tag{6}$$

Dystrybuanta empiryczna:

$$\hat{F}_N(x) = \frac{1}{N} \sum_{n=1}^{N} I(X_n \leqslant x) \tag{7}$$

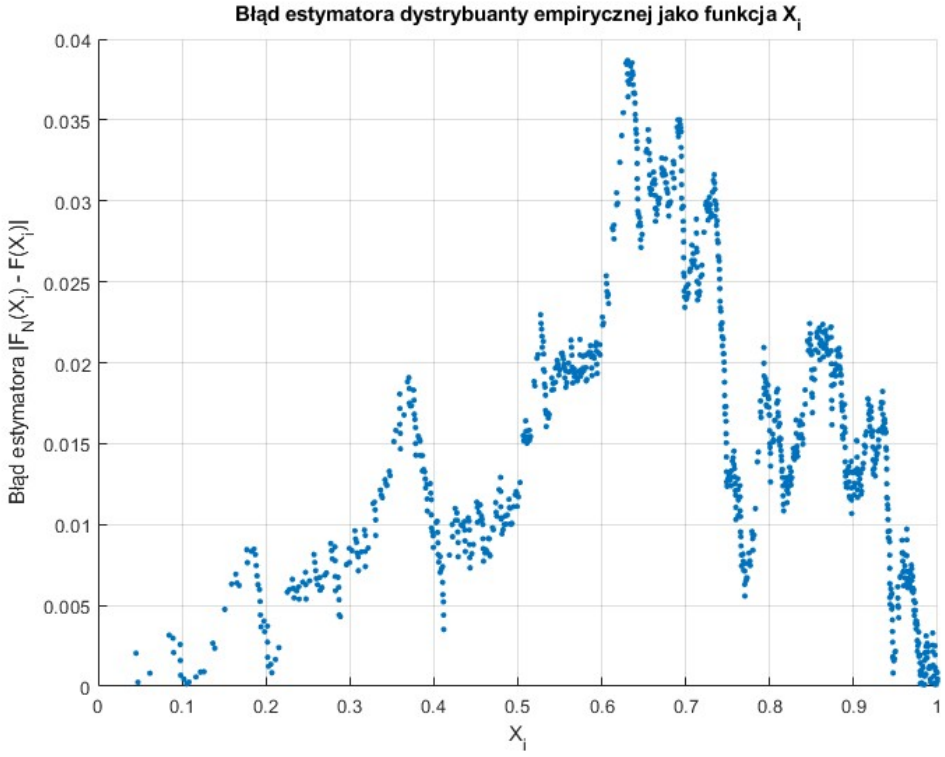
2.3 Zadanie 2 – Porównanie dystrybuanty empirycznej i teoretycznej



Rysunek 3: Dystrybuanta empiryczna i teoretyczna ($N = 1000$)

Na powyższym wykresie porównano dystrybuantę empiryczną z dystrybuantą teoretyczną $F(x) = x^2$. Dla dużej liczby próbek ($N = 1000$) estymacja empiryczna bardzo dobrze odwzorowuje rozkład teoretyczny, co potwierdza zbieżność dystrybuanty empirycznej do teoretycznej zgodnie z twierdzeniem Glivenki-Cantelliego.

2.4 Zadanie 3 – Błąd estymatora dystrybuanty

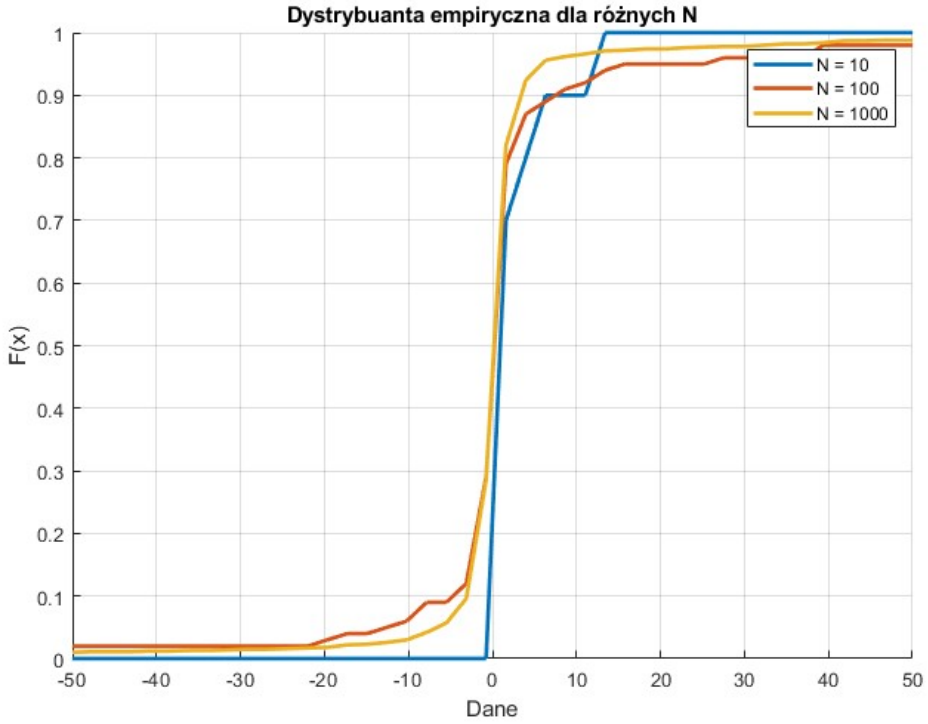


Rysunek 4: Błąd estymatora dystrybuanty empirycznej jako funkcja zmiennej losowej X_i

Na powyższym wykresie przedstawiono zależność błędu estymatora $|\hat{F}_N(X_i) - F(X_i)|$ od wartości samej zmiennej losowej X_i. Zgodnie z oczekiwaniami, największe błędy występują dla tych wartości X_i, dla których odpowiadająca wartość dystrybuanty $F(x)$ jest bliska 0.5, czyli w centrum rozkładu. Na brzegach przedziału, gdzie $F(x) \approx 0$ lub $F(x) \approx 1$, estymacja jest bardziej jednoznaczna, a błąd empiryczny – mniejszy. Wykres ten lepiej niż standardowa prezentacja względem indeksu próbki oddaje teoretyczną strukturę wariancji estymatora.

2.5 Zadanie 4 – Wpływ liczby próbek N na dystrybuantę

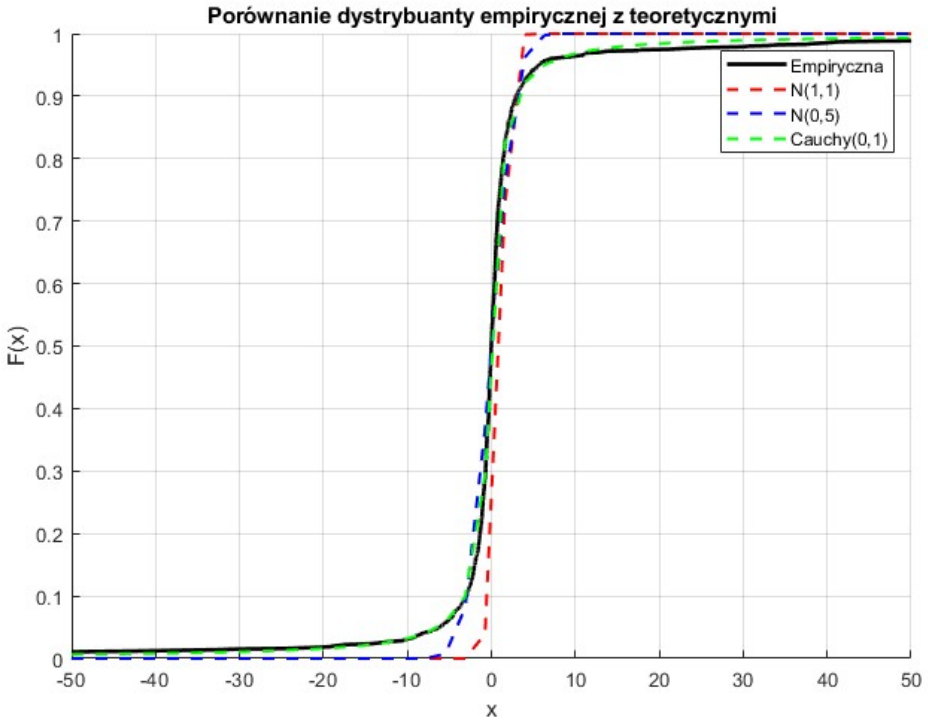
W kolejnym kroku wykorzystano dane zawarte w pliku `ModelowanieLab4Data.txt`. Dla liczności $N = 10$, 100 oraz 1000 wyznaczono empiryczne dystrybuanty $\hat{F}_N(x)$ i przedstawiono je na wykresie.



Rysunek 5: Dystrybuanta empiryczna danych z pliku dla różnych wartości N (10, 100, 1000)

Jak można zaobserwować, im większa próbka, tym dystrybuanta empiryczna jest bardziej gładka i dokładna. Dla $N = 10$ widoczne są skoki charakterystyczne dla małej liczby obserwacji. Dla $N = 1000$ przebieg jest znacznie bardziej płynny.

W celu identyfikacji potencjalnego rozkładu danych porównano dystrybuantę empiryczną z trzema teoretycznymi rozkładami: normalnym $N(1, 1)$, normalnym $N(0, 5)$ oraz rozkładem Cauchy'ego z $x_0 = 0$, $\gamma = 1$.

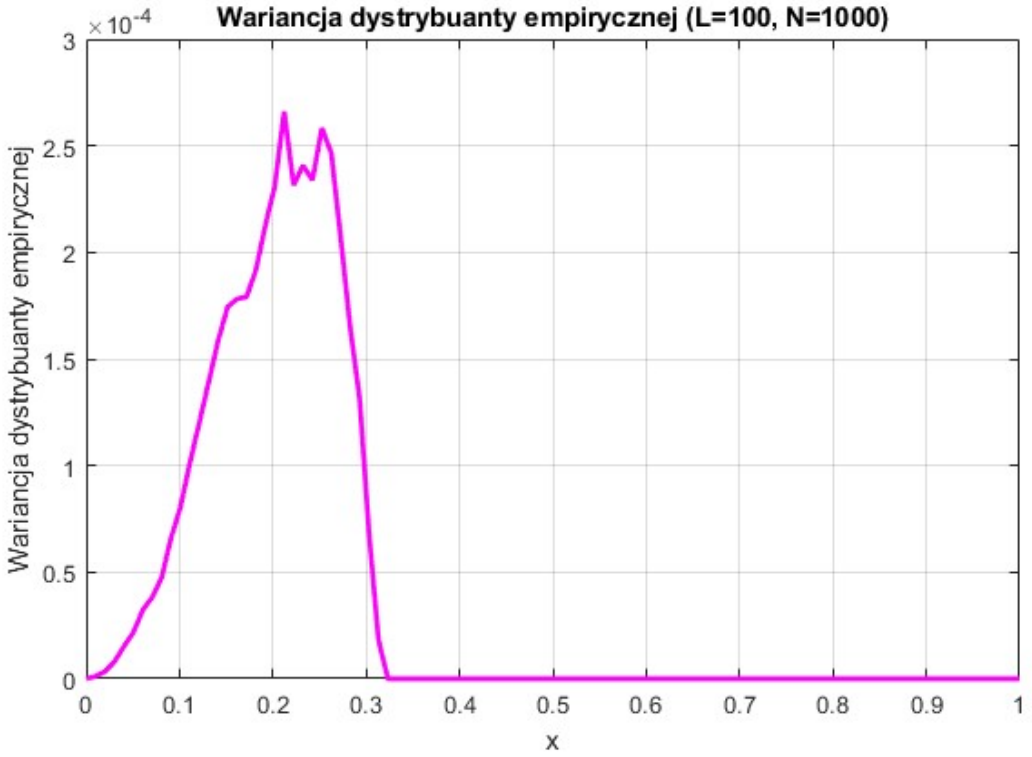


Rysunek 6: Porównanie dystrybuanty empirycznej z dystrybuantami rozkładów: normalnego $N(1,1)$, $N(0,5)$ i Cauchy'ego $(0,1)$

Analiza wykresu wskazuje, że empiryczna dystrybuanta najbliższa jest dystrybuancie rozkładu Cauchy'ego, co sugeruje, że dane mogą pochodzić właśnie z tego rozkładu. Widzimy, że dystrybuanta normalna $N(1,1)$ szybko „wyskakuje" ku 1, co nie jest zgodne z zaobserwowanym rozkładem danych, który charakteryzuje się grubszymi ogonami.

2.6 Zadanie 5 – Wariancja dystrybuanty empirycznej

$$\mathrm{Var}\left\{\hat{F}_N(x)\right\} = \frac{1}{N}F(x)(1-F(x)) \tag{8}$$



Rysunek 7: Wariancja dystrybuanty empirycznej ($L = 100$, $N = 1000$)

Zgodnie ze wzorem $\mathrm{Var}\{\hat{F}_N(x)\} = \frac{1}{N}F(x)(1-F(x))$ wariancja estymatora jest największa tam, gdzie $F(x)$ jest bliskie 0.5, czyli w środkowej części przedziału. Jest to logiczne – w centrum rozkładu losowania mają największą niepewność co do pozycji względem x, a więc i estymacja obarczona jest największą wariancją. Na brzegach wariancja maleje do zera.

3 Laboratorium 5 – Estymacja jądrowa

3.1 Cel

Zbadanie wpływu funkcji jądra oraz parametru wygładzania h_N na jakość estymacji gęstości rozkładów.

3.2 Estymator jądrowy

$$\hat{f}_N(x) = \frac{1}{Nh_N}\sum_{n=1}^{N} K\left(\frac{X_n - x}{h_N}\right) \tag{9}$$

3.3 Funkcje jądra

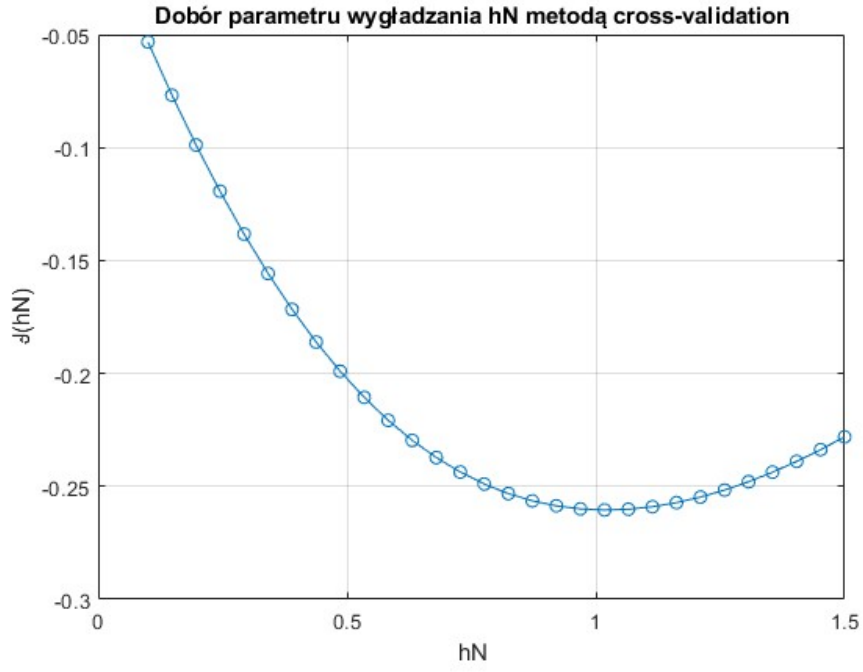
- Boxcar: $K(x) = \frac{1}{2}I(|x| \leqslant 1)$
- Gaussowskie: $K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
- Epanechnikov: $K(x) = \frac{3}{4}(1-x^2)I(|x| \leqslant 1)$
- Tricube: $K(x) = \frac{70}{81}(1-|x|^3)^3 I(|x| \leqslant 1)$

3.4 Błąd empiryczny

$$\mathrm{Err}\{\hat{f}_N\} = \frac{1}{LM}\sum_{l=1}^{L}\sum_{m=1}^{M}\left[\hat{f}_N^{[l]}(x_m) - f(x_m)\right]^2 \tag{10}$$

3.5 Dobór optymalnego parametru wygładzania h_N metodą cross-validation

W celu doboru optymalnego parametru wygładzania h_N zastosowano metodę leave-one-out cross-validation. Obliczono wartość funkcji błędu $\hat{J}(h_N)$, która jest przybliżeniem całkowego błędu estymatora jądrowego, dla szerokiego zakresu wartości h_N.



Rysunek 8: Dobór parametru wygładzania h_N metodą cross-validation.

Na wykresie widać minimum funkcji $\hat{J}(h_N)$ w okolicach $h_N \approx 0{,}75$, co sugeruje, że ta wartość zapewnia najlepszy kompromis między zbytnią wygładzoną (zbyt dużym h_N), a przetrenowaną (zbyt małym h_N) estymacją gęstości.

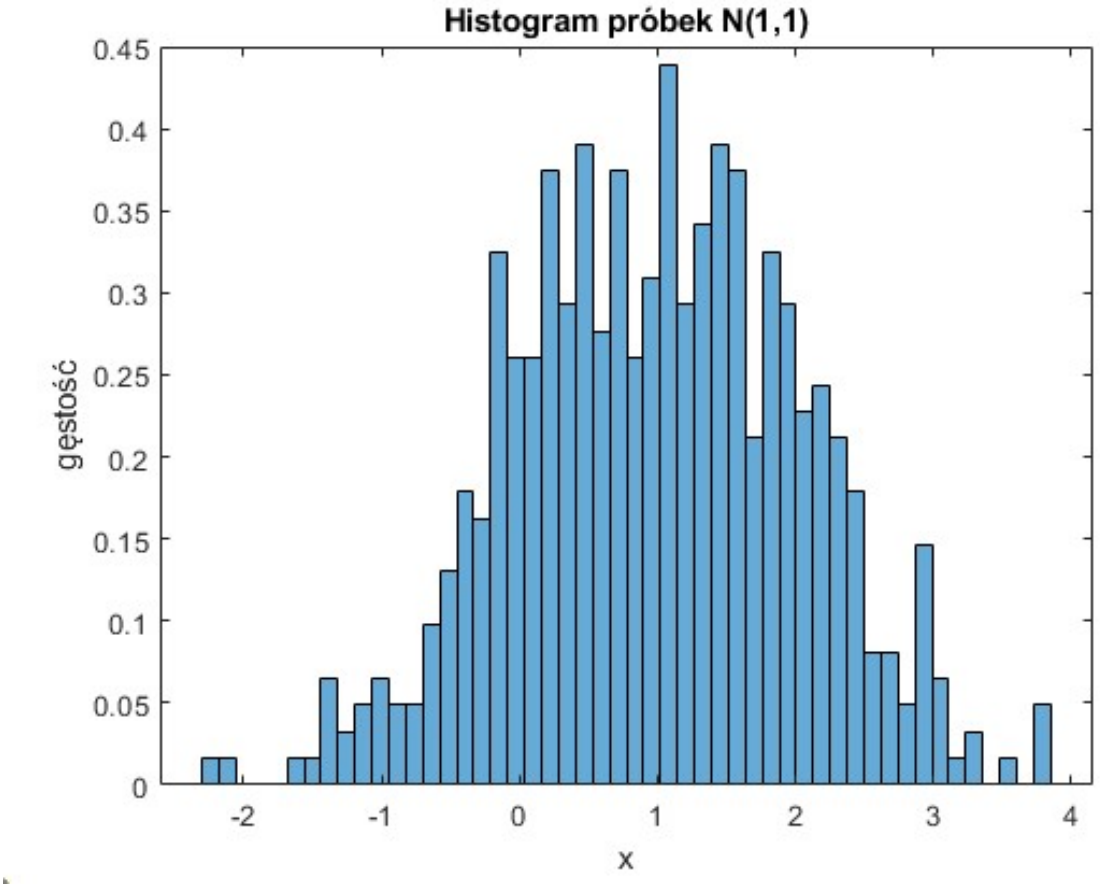
Dzięki tej metodzie możliwa jest obiektywna i niezależna od użytkownika optymalizacja parametru jądrowego estymatora gęstości, co znacząco poprawia jakość wyników.

3.6 Wnioski

Parametry: $N = 500$, $L = 10$, $M = 100$

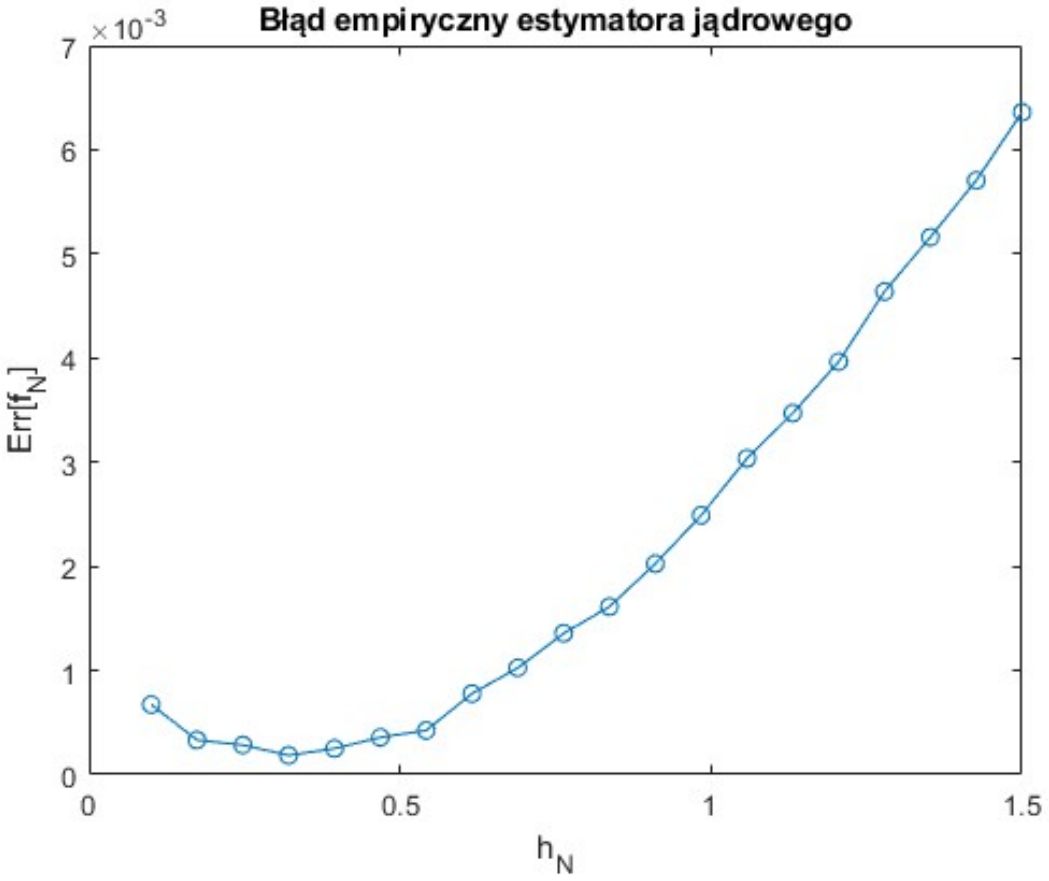
- Jądro prostokątne dawało najgorsze rezultaty.
- Jądro Tricube z $h_N \approx 0.5$ dawało najlepsze dopasowanie do $f(x) = \mathcal{N}(1, 1)$.
- Przy zbyt małym h_N estymator był zbyt niestabilny.

3.7 Wyniki i ilustracje



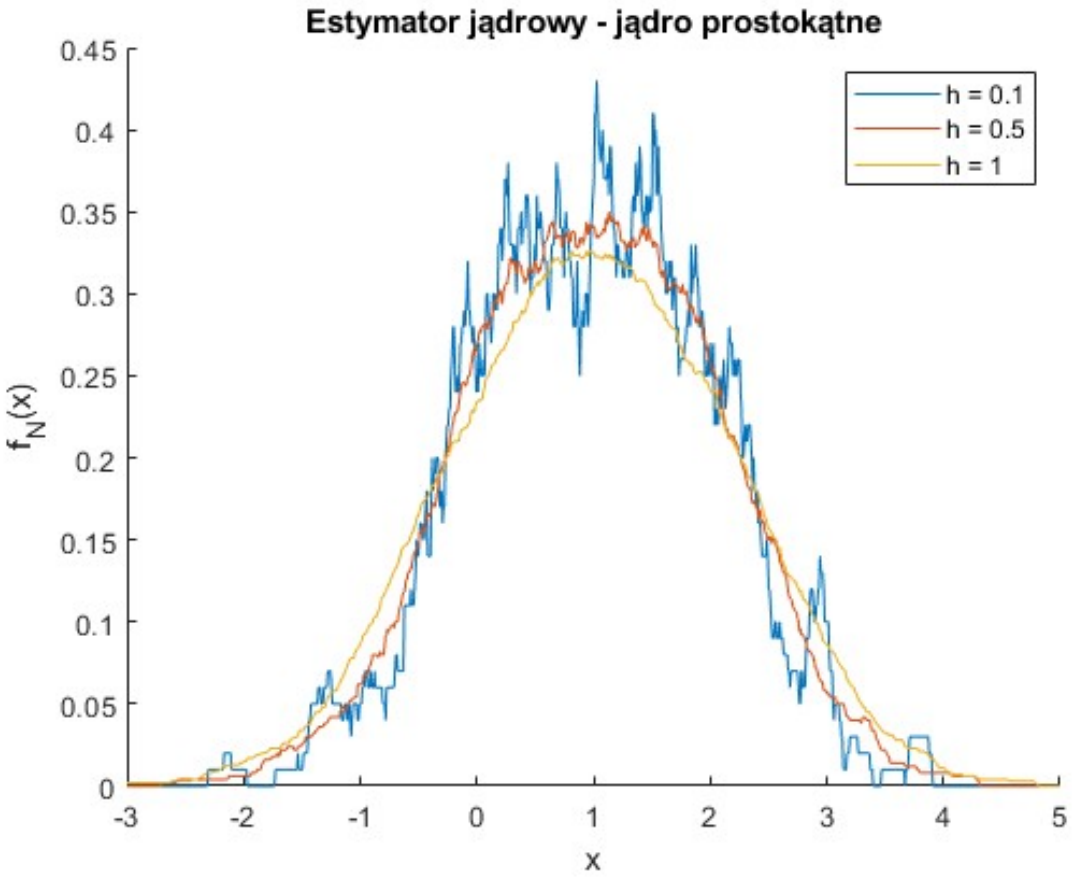
Rysunek 9: Histogram próbek z rozkładu normalnego $N(1,1)$ – $N = 1000$, 100 słupków

Histogram prezentuje próbki wygenerowane z rozkładu normalnego $N(1,1)$ i pozwala wizualnie ocenić kształt rozkładu. Próbka o wielkości $N = 1000$ pozwala na stosunkowo dokładne odwzorowanie gęstości prawdopodobieństwa, widoczne są charakterystyczne cechy rozkładu normalnego – symetria i dzwonowaty kształt.



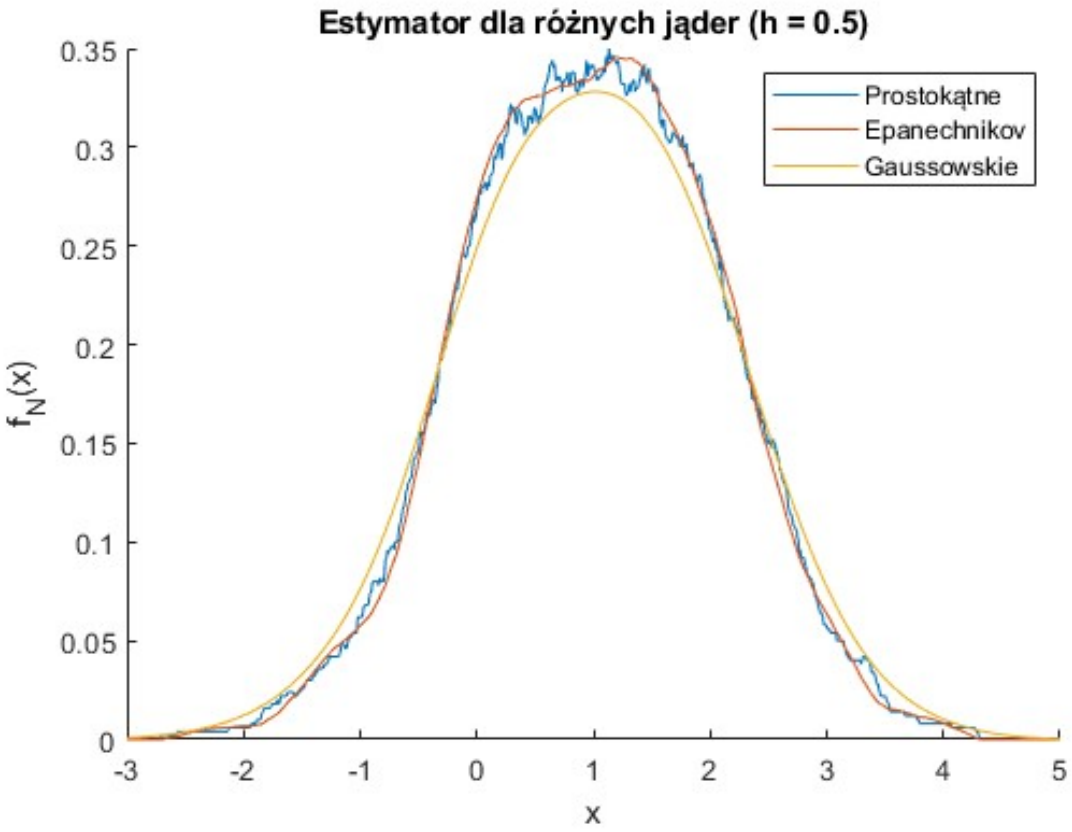
Rysunek 10: Błąd empiryczny estymatora jądrowego w funkcji h_N

Wykres prezentuje błąd empiryczny estymatora jądrowego w zależności od wartości parametru wygładzania h_N. Wartość błędu początkowo maleje – co oznacza poprawę dopasowania – a następnie rośnie, wskazując na przeuczenie modelu. Istnieje optymalne h_N, które minimalizuje błąd, co pokazuje konieczność jego odpowiedniego doboru.



Rysunek 11: Porównanie estymatora jądrowego dla różnych wartości h (jądro prostokątne)

Rysunek pokazuje wpływ parametru h_N na estymowaną gęstość. Przy zbyt małym h_N (np. 0.1) estymator jest niestabilny i przetrenowany (zbyt „szczegółowy"), natomiast zbyt duże h_N (np. 1) powoduje nadmierne wygładzenie i utratę struktury rozkładu. Optymalna wartość znajduje się pośrodku i zapewnia dobre odwzorowanie rozkładu.



Rysunek 12: Porównanie estymatorów dla różnych jąder przy $h = 0.5$

Wykres przedstawia estymację gęstości z wykorzystaniem różnych funkcji jądra przy stałym $h_N = 0.5$. Wszystkie estymatory odwzorowują kształt rozkładu normalnego, ale różnią się poziomem wygładzenia i „krawędziami". Jądro Gaussowskie zapewnia najsilniejsze wygładzenie, Epanechnikov daje estymację najbliższą teoretycznej funkcji gęstości.