Identyfikacja i Modelowanie Statystyczne - Sprawozdanie 0-3

Paweł Szczepaniak (249014)

Marzec - Czerwiec 2022

-1 Wstęp

Wszystkie ćwiczenia zrealizowano przy pomocy języka programowania Python 3 z wykorzystaniem bibliotek `numpy` i `matplotlib`.

0 Laboratorium 0 - generatory liczb pseudolosowych

W ćwiczeniu zaimplementowano generator liczb pseudolosowych z rozkładu jednostajnego.

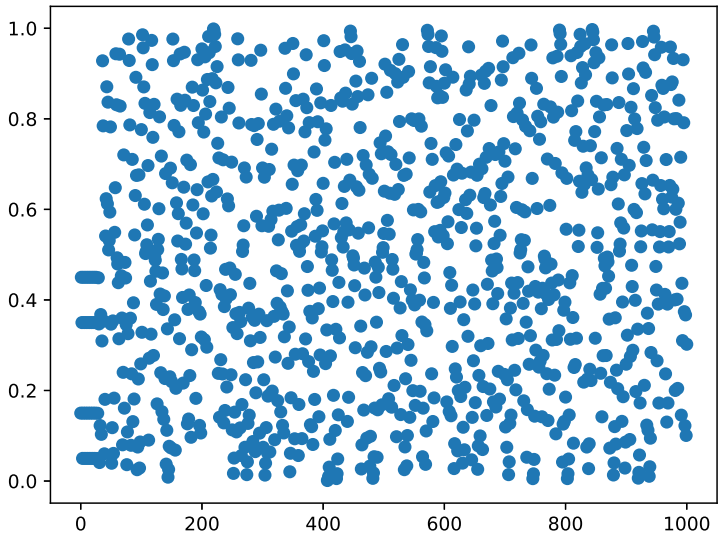
0.1 Generator liczb pseudolosowych oparty na przekształceniu piłokształtnym

Generator zaimplementowano w oparciu o przekształcenie piłokształtne

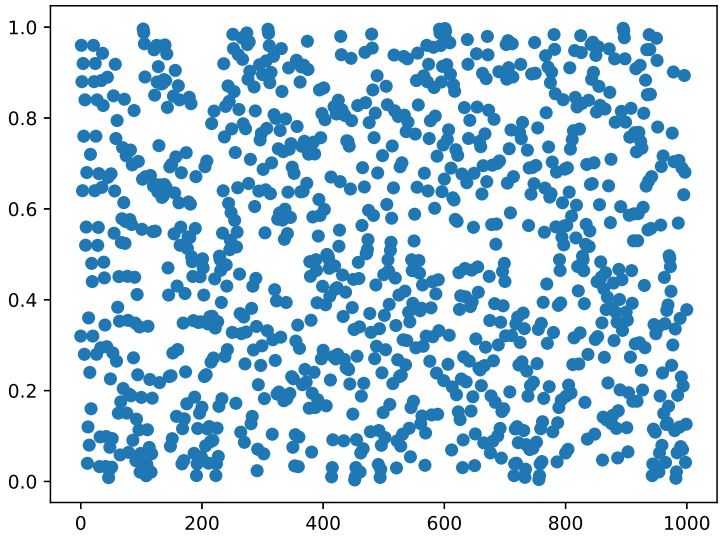
$$X_{n+1} = X_n \cdot z - \lfloor X_n \cdot z \rfloor$$

0.1.1 Badanie wpływu wartości początkowej na własności generatora

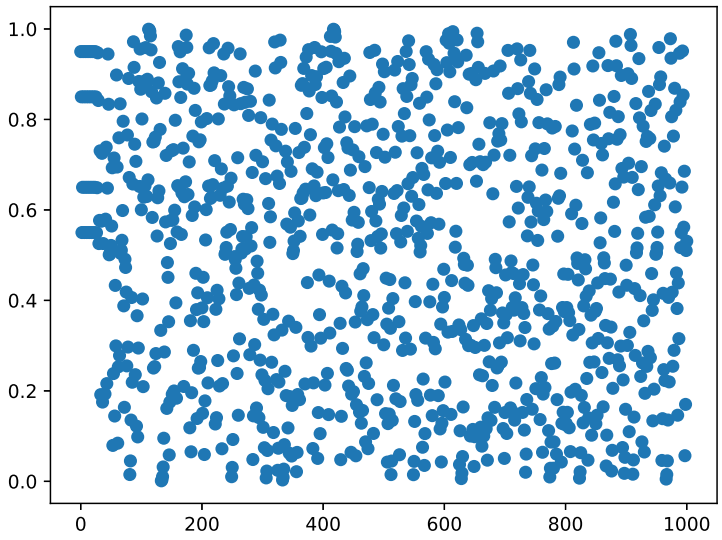
Na wykresach przedstawiono rozkłady wygenerowanych wartości dla różnych wartościach początkowych X_0 przy stałej wartości parametru $z = 3$ i stałej liczbie próbek $n = 1000$.



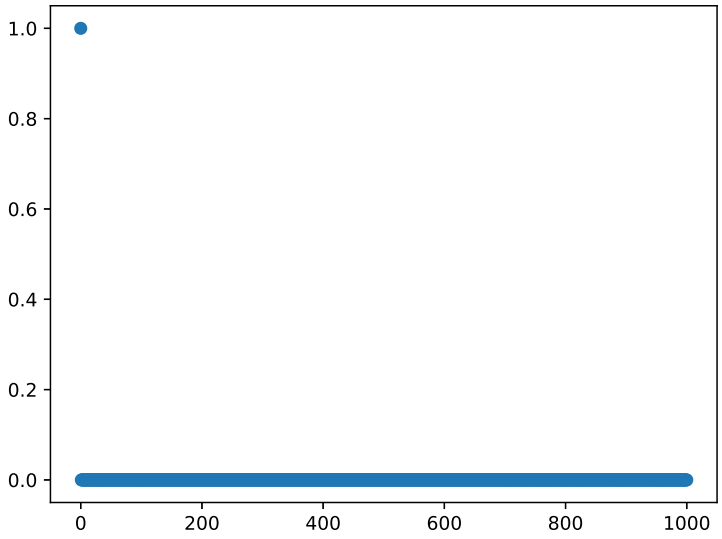
Rysunek 1: Rozkład wygenerowanych wartości dla $X_0 = 0.15$



Rysunek 2: Rozkład wygenerowanych wartości dla $X_0 = 0.32$



Rysunek 3: Rozkład wygenerowanych wartości dla $X_0 = 0.95$

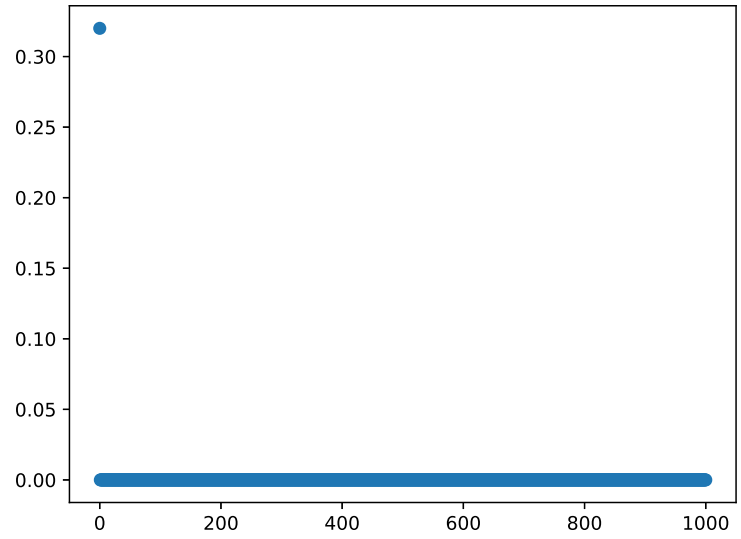


Rysunek 4: Rozkład wygenerowanych wartości dla $X_0 = 1$

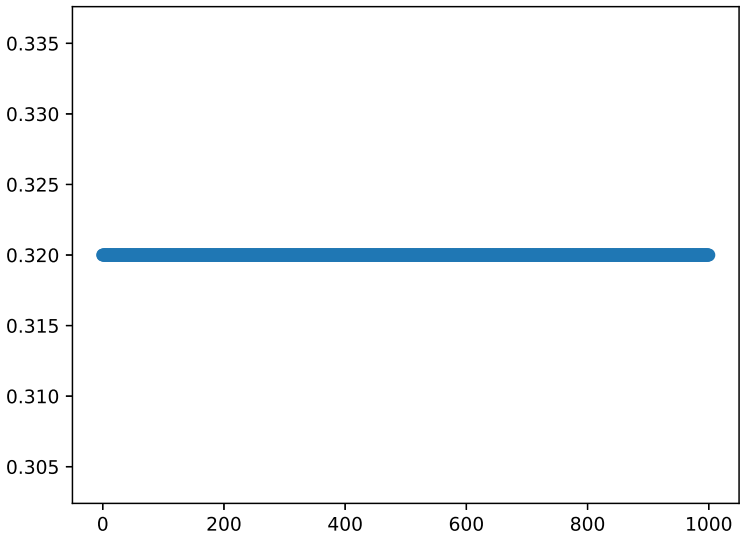
Można zaobserwować, że wartość początkowa nie ma znaczącego wpływu na działanie generatora. Wyjątek stanowią liczby całkowite, dla których generowane liczby zbiegają do wartości 0 już w drugiej próbce. Jest to spowodowane operacją odejmowania części całkowitej z liczby całkowitej, co skutkuje otrzymaniem wartości 0. Z tego powodu zalecanym przedziałem wyboru wartości początkowej jest $X_0 \in (0, 1)$.

0.1.2 Badanie wpływu wartości parametru z na własności generatora

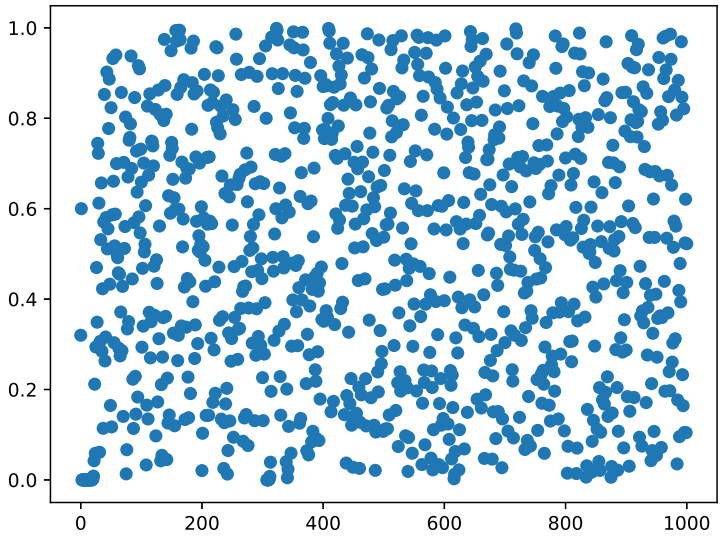
Na wykresach przedstawiono rozkłady wygenerowanych wartości dla różnych wartości parametru z przy stałej wartości początkowej $X_0 = 0.32$ i stałej liczbie próbek $n = 1000$.



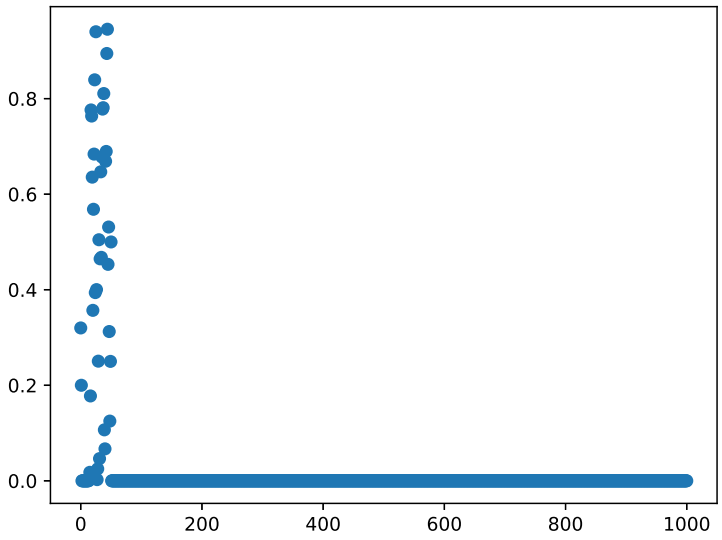
Rysunek 5: Rozkład wygenerowanych wartości dla $z = 0$



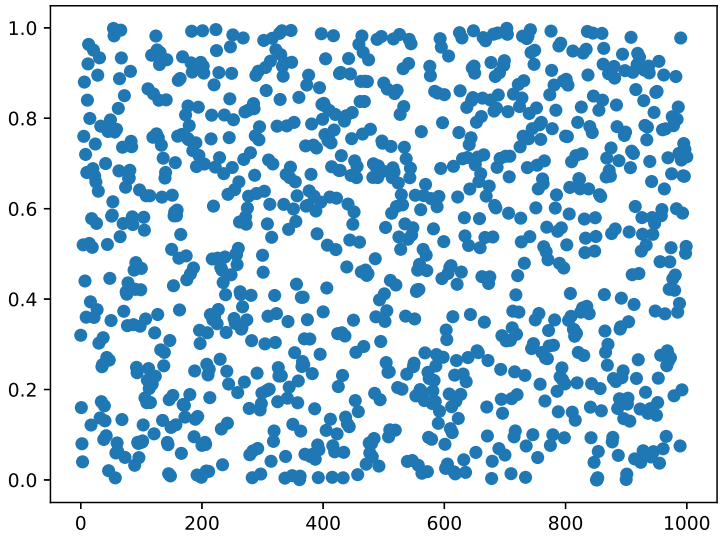
Rysunek 6: Rozkład wygenerowanych wartości dla $z = 1$



Rysunek 7: Rozkład wygenerowanych wartości dla $z = 5$



Rysunek 8: Rozkład wygenerowanych wartości dla $z = 10$



Rysunek 9: Rozkład wygenerowanych wartości dla $z = 13$

Można zauważyć, że przy $z = 0$ generowane liczby zbiegają do wartości 0 już w drugiej próbce - wynika to z mnożenia wartości przez współczynnik $z = 0$. Przy $z = 1$ wszystkie generowane liczby są równe wartości początkowej. Dodatkowo, można zaobserwować, że dla zbadanych parzystych wartościach parametru z występuje szybki spadek generowanych wartości do 0. Przy zbadanych nieparzystych wartościach parametru z generator działa prawidłowo.

0.1.3 Badanie okresu generatora

Do badania okresu generatora wykorzystano generator o wartości początkowej $X_0 = 0.32$, wartości parametru $z = 3$ oraz liczbie próbek $n = 100$. Wygenerowane wartości zaokrąglono do 2 cyfr znaczących w celu minimalizacji błędu kwantyzacji liczb zmiennoprzecinkowych.

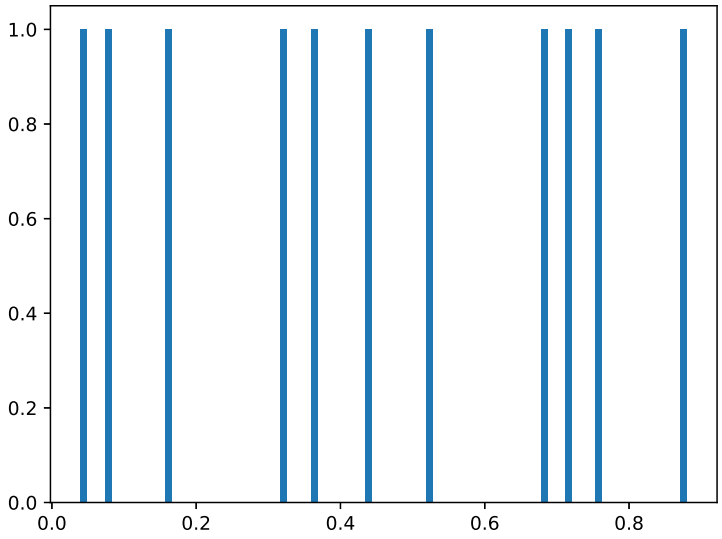
n	X	n	X	n	X	n	X
0	0.32	25	0.76	50	0.68	75	0.72
1	0.96	26	0.28	51	0.03	76	0.15
2	0.88	27	0.84	52	0.09	77	0.45
3	0.64	28	0.52	53	0.28	78	0.35
4	0.92	29	0.56	54	0.85	79	0.06
5	0.76	30	0.68	55	0.55	80	0.19
6	0.28	31	0.03	56	0.64	81	0.58
7	0.84	32	0.10	57	0.92	82	0.73
8	0.52	33	0.29	58	0.75	83	0.19
9	0.56	34	0.88	59	0.26	84	0.57
10	0.68	35	0.65	60	0.79	85	0.70
11	0.04	36	0.94	61	0.38	86	0.09
12	0.12	37	0.83	62	0.15	87	0.27
13	0.36	38	0.48	63	0.45	88	0.82
14	0.08	39	0.45	64	0.35	89	0.45
15	0.24	40	0.34	65	0.06	90	0.35
16	0.72	41	0.03	66	0.18	91	0.05
17	0.16	42	0.10	67	0.53	92	0.14
18	0.48	43	0.30	68	0.58	93	0.41
19	0.44	44	0.89	69	0.73	94	0.23
20	0.32	45	0.67	70	0.20	95	0.70
21	0.96	46	0.01	71	0.61	96	0.11
22	0.88	47	0.03	72	0.84	97	0.34
23	0.64	48	0.08	73	0.52	98	0.02
24	0.92	49	0.23	74	0.57	99	0.06

Tabela 1: Wygenerowane wartości dla $z = 3$

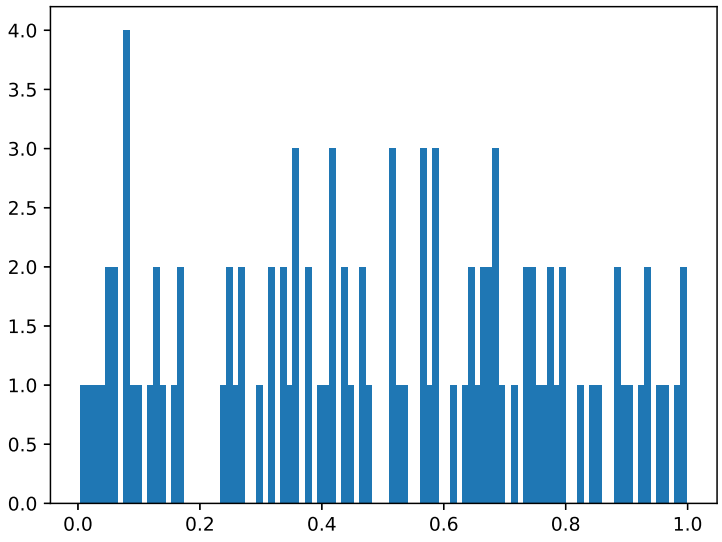
Można zauważyć, że początkowa sekwencja $0.32, 0.96, 0.88, 0.64, \ldots$ zaczyna powtarzać się w 20 próbce wygenerowanych wartości.

0.1.4 Badanie histogramu ciągu wygenerowanych liczb

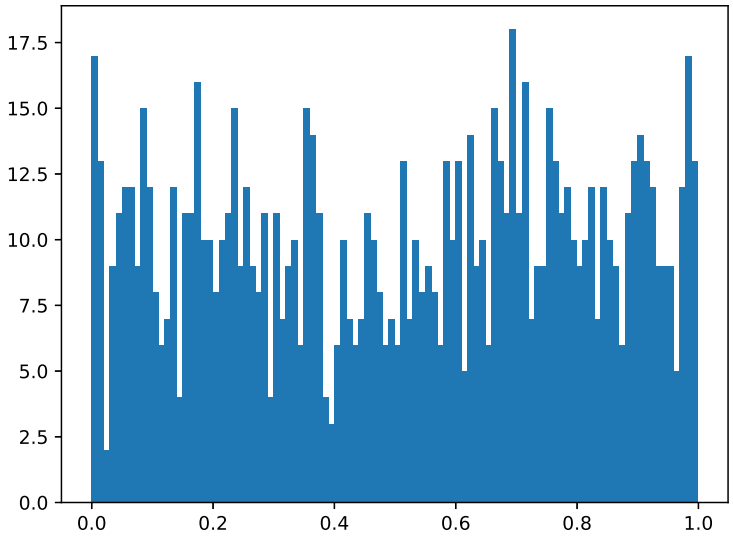
Na wykresach przedstawiono histogramy wygenerowanych wartości przy zmiennej liczbie próbek n przy stałej wartości początkowej $X_0 = 0.32$ i stałej wartości parametru $z = 13$.



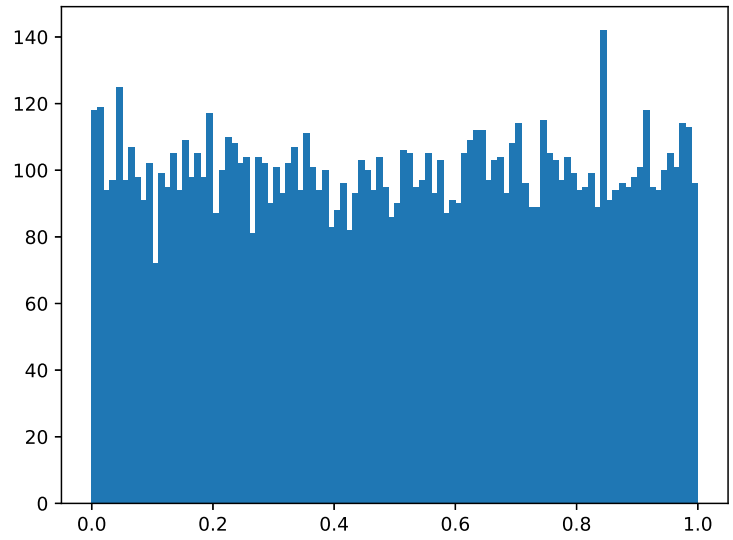
Rysunek 10: Histogram wygenerowanych wartości dla $n = 10$



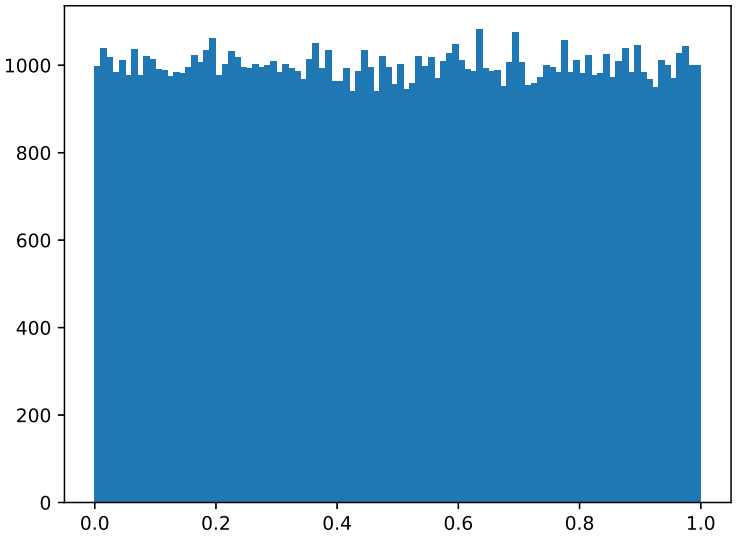
Rysunek 11: Histogram wygenerowanych wartości dla $n = 100$



Rysunek 12: Histogram wygenerowanych wartości dla $n = 1000$



Rysunek 13: Histogram wygenerowanych wartości dla $n = 10000$



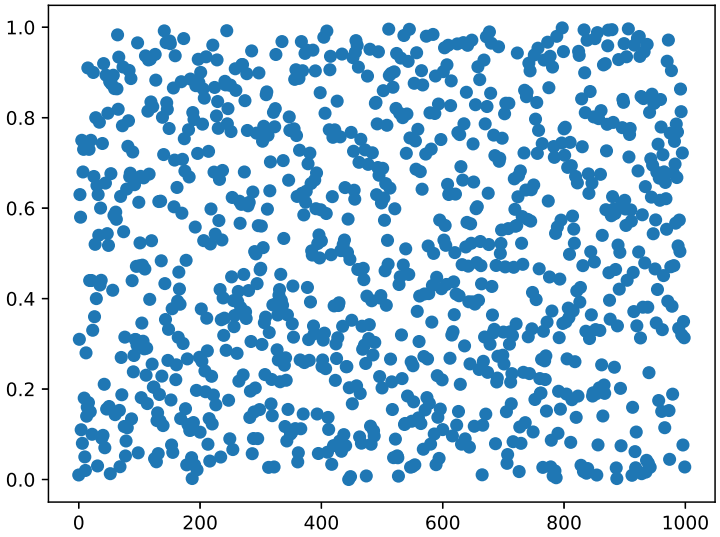
Rysunek 14: Histogram wygenerowanych wartości dla $n = 100000$

Można zauważyć, że podobieństwo histogramu dla typowego dla rozkładu jednostajnego zwiększa się wraz ze wzrostem liczby wygenerowanych próbek.

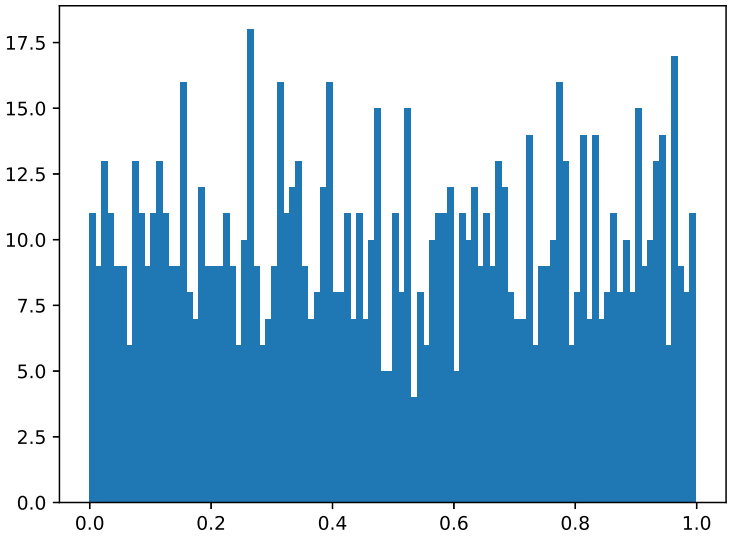
0.2 Generator liczb pseudolosowych oparty na równaniu

Generator zaimplementowano w oparciu o równanie

$$X_{n+1} = (a_0 X_n + a_1 X_{n-1} + \cdots + a_k X_{n-k} + c) \bmod m$$



Rysunek 15: Rozkład wygenerowanych wartości dla $X_0 = 0.01$, $k = 5$, $m = 1$, $c = 0.3$



Rysunek 16: Histogram wygenerowanych wartości dla $X_0 = 0.01$, $k = 5$, $m = 1$, $c = 0.3$

1 Laboratorium 1 - metoda odwracania dystrybuanty

W ćwiczeniu zaimplementowano generator liczb pseudolosowych bazujący na metodzie odwracania dystrybuanty.

1.1 Rozkład 1

Zaimplementowano generator generujący liczby pseudolosowe z rozkładu o gęstości prawdopodobieństwa

$$f(x) = \begin{cases} 2x & \text{dla} \quad x \in [0,1] \\ 0 & \text{dla} \quad x \in [-\infty, 0) \cup (1, \infty) \end{cases}$$

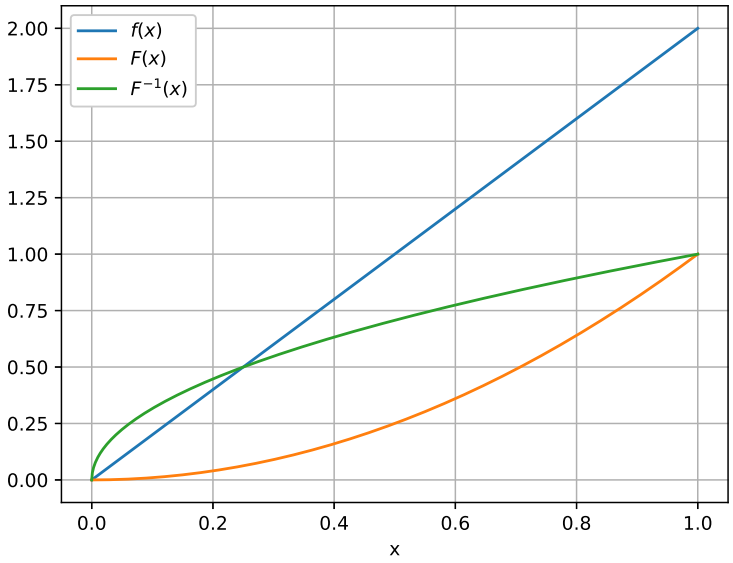
Dla rozkładu wyznaczono dystrybuantę:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = 0 + \int_{0}^{x} 2t\,dt = 2 \cdot \left[\frac{t^2}{2}\right]_0^x = x^2$$

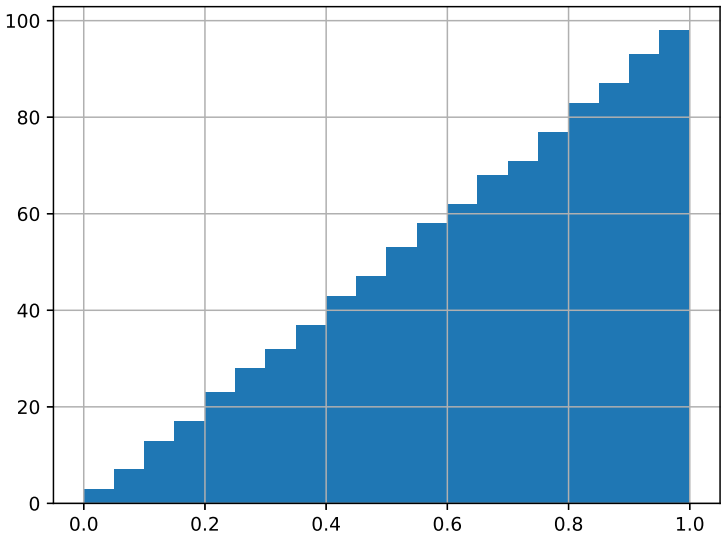
$$F(x) = \begin{cases} 0 & \text{dla} \quad x \in (-\infty, 0) \\ x^2 & \text{dla} \quad x \in [0,1] \\ 1 & \text{dla} \quad x \in (1, \infty) \end{cases}$$

Dla przedziału $x \in [0,1]$ wyznaczono odwrotną dystrybuantę:

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



Rysunek 17: Wykres gęstości prawdopodobieństwa, dystrybuanty i odwrotnej dystrybuanty rozkładu 1



Rysunek 18: Histogram wygenerowanych liczb pseudolosowych dla rozkładu 1

1.2 Rozkład 2

Zaimplementowano generator generujący liczby pseudolosowe z rozkładu o gęstości prawdopodobieństwa

$$f(x) = \begin{cases} x+1 & \text{dla} \quad x \in (-1,0) \\ -x+1 & \text{dla} \quad x \in [0,1) \\ 0 & \text{dla} \quad x \notin (-1,1) \end{cases}$$

Dla rozkładu wyznaczono dystrybuantę:
Dla przedziału $x \in (-1, 0)$

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = 0 + \int_{-1}^{x} (t+1)\,dt = \left[\frac{t^2}{2} + t\right]_{-1}^{x} = \frac{x^2}{2} + x + \frac{1}{2}$$

Dla przedziału $x \in [0, 1)$

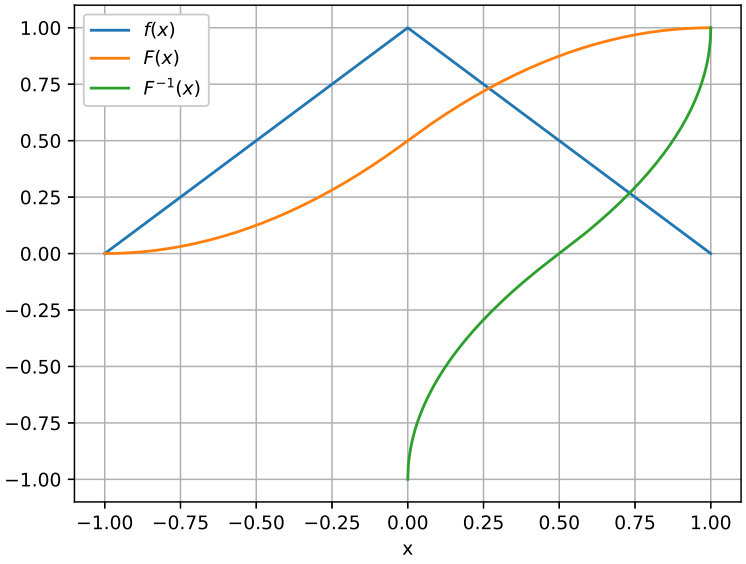
$$F(x) = \frac{1}{2} + \int_{0}^{x} f(t)\,dt = \frac{1}{2} + \int_{0}^{x} (-t+1)\,dt = \frac{1}{2} + \left[-\frac{t^2}{2} + t\right]_{0}^{x} = -\frac{x^2}{2} + x + \frac{1}{2}$$

zatem

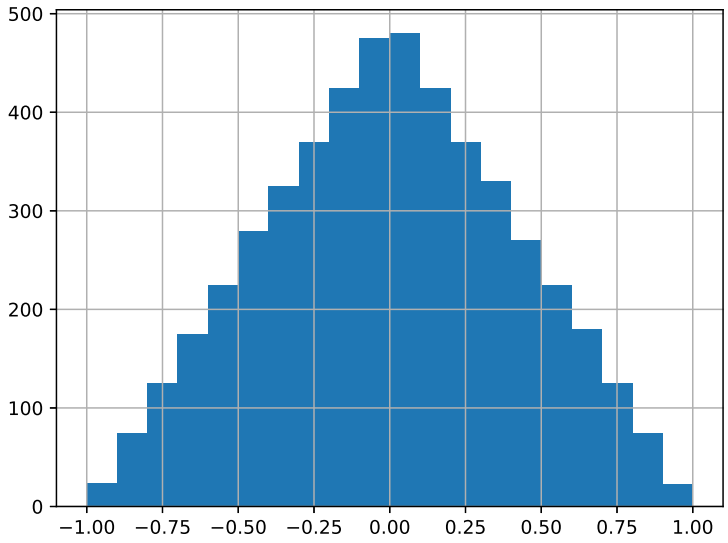
$$F(x) = \begin{cases} 0 & \text{dla} \quad x \in (-\infty, -1] \\ \dfrac{x^2}{2} + x + \dfrac{1}{2} & \text{dla} \quad x \in (-1, 0) \\ -\dfrac{x^2}{2} + x + \dfrac{1}{2} & \text{dla} \quad x \in [0, 1) \\ 1 & \text{dla} \quad x \in [1, \infty) \end{cases}$$

Dla przedziału $x \in [0, 1]$ wyznaczono odwrotną dystrybuantę:

$$F^{-1}(y) = \begin{cases} \sqrt{2y} - 1 & \text{dla} \quad x \in \left[0, \dfrac{1}{2}\right] \\ 1 - \sqrt{2 - 2y} & \text{dla} \quad x \in \left(\dfrac{1}{2}, 1\right] \end{cases}$$



Rysunek 19: Wykres gęstości prawdopodobieństwa, dystrybuanty i odwrotnej dystrybuanty rozkładu 2



Rysunek 20: Histogram wygenerowanych liczb pseudolosowych dla rozkładu 2

1.3 Rozkład wykładniczy

Zaimplementowano generator generujący liczby pseudolosowe z rozkładu wykładniczego o gęstości prawdopodobieństwa

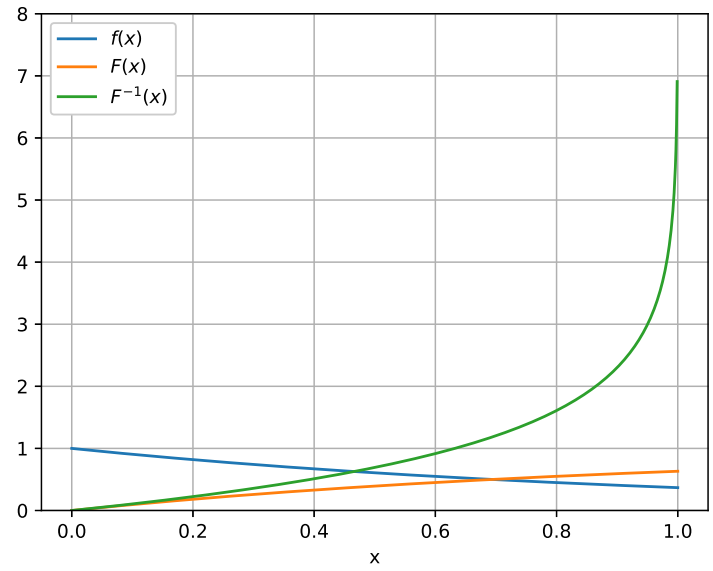
$$f(x) = e^{-x},\ x \geqslant 0$$

Dystrybuanta rozkładu wynosi:

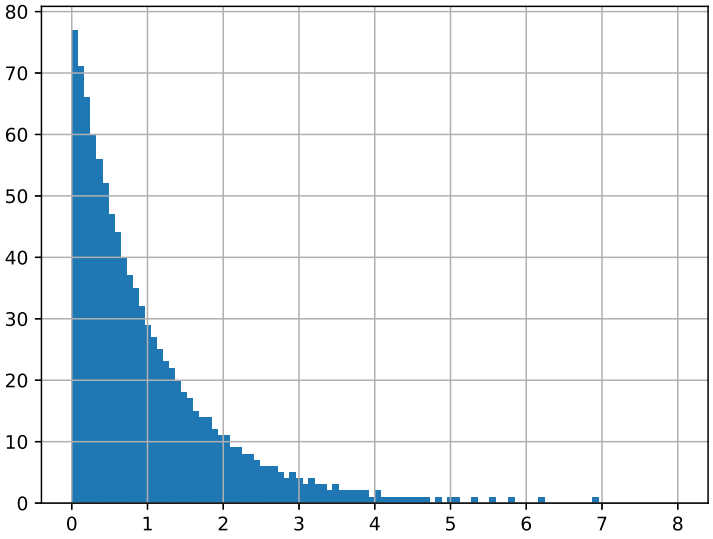
$$F(x) = 1 - e^{-x}$$

Odwrotna dystrybuanta rozkładu wynosi:

$$F^{-1}(y) = -\ln(1-y)$$



Rysunek 21: Wykres gęstości prawdopodobieństwa, dystrybuanty i odwrotnej dystrybuanty rozkładu wykładniczego



Rysunek 22: Histogram wygenerowanych liczb pseudolosowych dla rozkładu wykładniczego

1.4 Rozkład Laplace'a

Zaimplementowano generator generujący liczby pseudolosowe z rozkładu Laplace'a o gęstości prawdopodobieństwa

$$f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

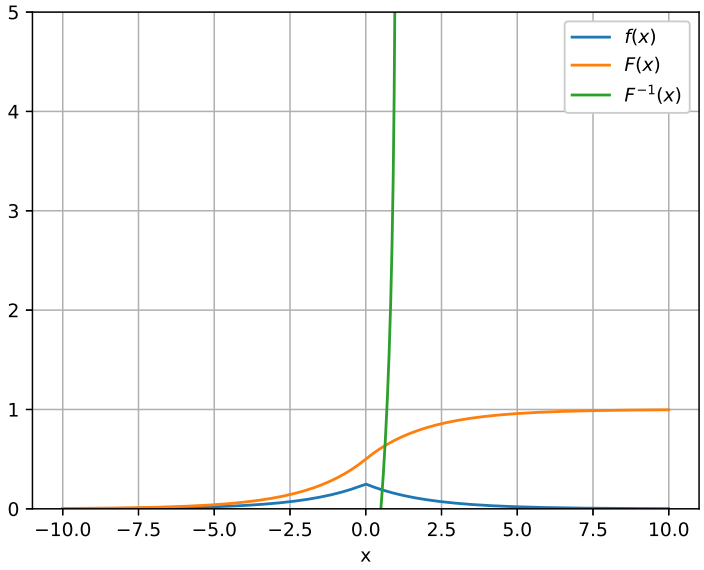
Dystrybuanta rozkładu wynosi:

$$F(x) = \begin{cases} \frac{1}{2}e^{\frac{x-\mu}{b}} & \text{dla} \quad x \leqslant \mu \\ 1 - \frac{1}{2}e^{-\frac{x-\mu}{b}} & \text{dla} \quad x \geqslant \mu \end{cases}$$

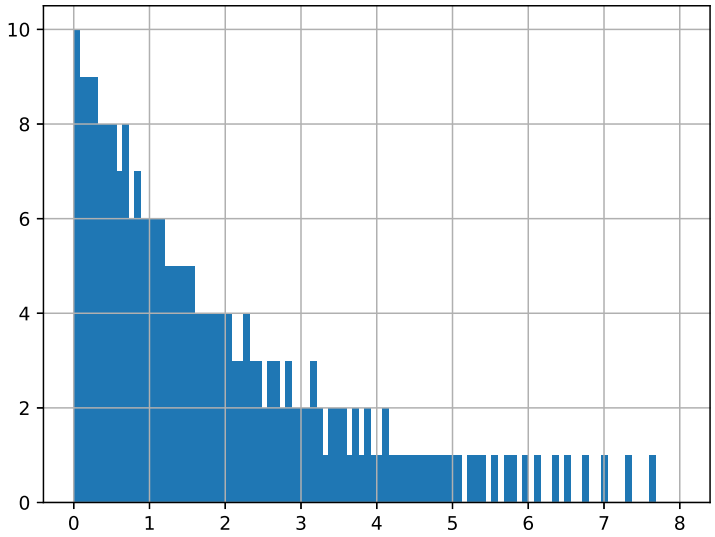
Odwrotna dystrybuanta rozkładu wynosi:

$$F^{-1}(y) = \begin{cases} \mu + b\ln(2y) & \text{dla} \quad y \leqslant \frac{1}{2} \\ \mu - b\ln(2-2y) & \text{dla} \quad y \geqslant \frac{1}{2} \end{cases}$$

Przyjęto $\mu = 0$, $b = 2$



Rysunek 23: Wykres gęstości prawdopodobieństwa, dystrybuanty i odwrotnej dystrybuanty rozkładu Laplace'a



Rysunek 24: Histogram wygenerowanych liczb pseudolosowych dla rozkładu Laplace'a

1.5 Wnioski

Metoda odwracania dystrybuanty pozwala na generowanie liczb pseudolosowych z określonego rozkładu gęstości prawdopodobieństwa. Jej główną wadą jest jednak wymaganie znajomości dystrybuanty rozkładu, oraz konieczność uprzedniego wyznaczenia odwrotnej dystrybuanty żądanego rozkładu, co w wielu przypadkach może być skomplikowane, lub nawet niemożliwe do wykonania.

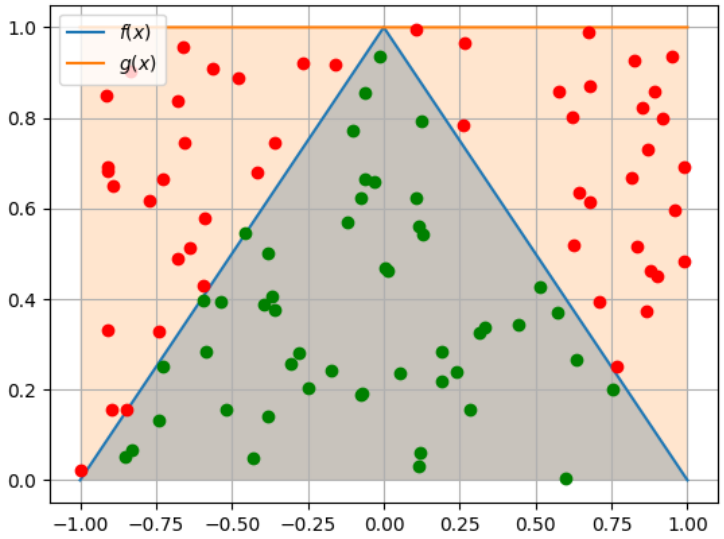
2 Laboratorium 2 - metoda odrzucania

W ćwiczeniu zaimplementowano generator liczb pseudolosowych bazujący na metodzie odrzucania.

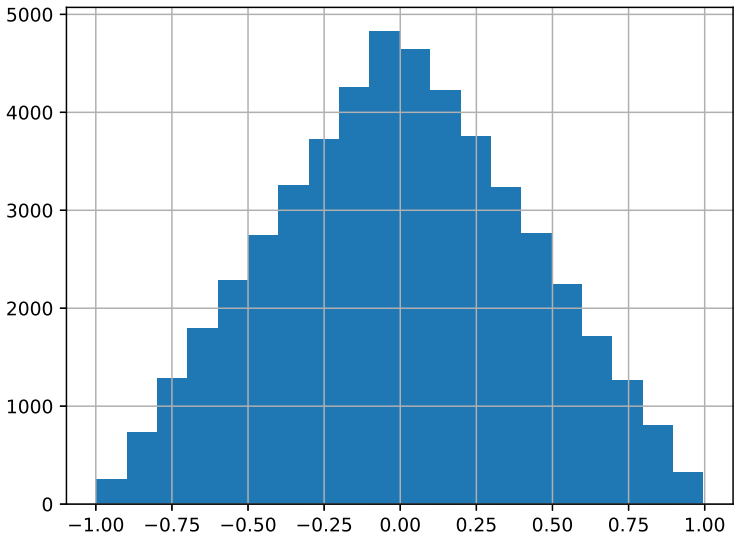
2.1 Rozkład 1

Zaimplementowano generator generujący liczby pseudolosowe z rozkładu o gęstości prawdopodobieństwa

$$f(x) = \begin{cases} x+1 & \text{dla} \quad x \in (-1, 0] \\ -x+1 & \text{dla} \quad x \in [0, 1] \\ 0 & \text{dla} \quad x \in (-\infty, -1] \cup (1, \infty) \end{cases}$$



Rysunek 25: Wykres gęstości prawdopodobieństwa $f(x)$ rozkładu 1 i funkcji aproksymacyjnej $g(x)$



Rysunek 26: Histogram wygenerowanych liczb pseudolosowych dla rozkładu 1

2.2 Rozkład 2

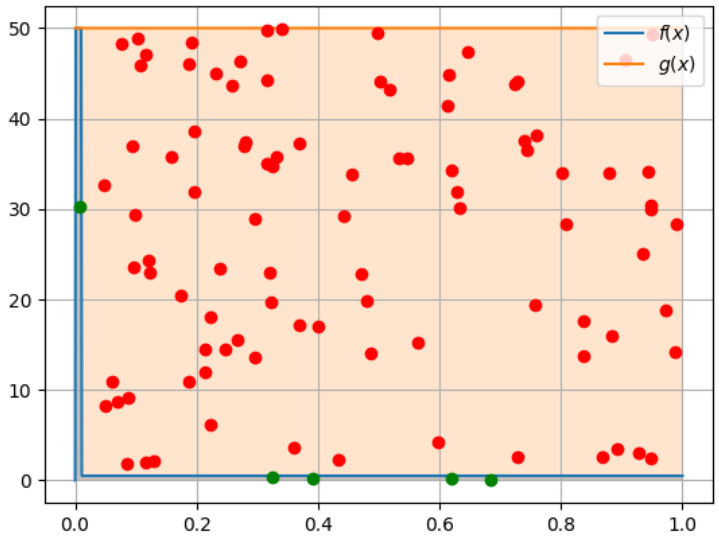
Zaimplementowano generator generujący liczby pseudolosowe z rozkładu o gęstości prawdopodobieństwa

$$f(x) = \begin{cases} 50 & \text{dla} \quad x \in \left(0, \frac{1}{100}\right] \\ c & \text{dla} \quad x \in \left(\frac{1}{100}, 1\right] \end{cases}$$

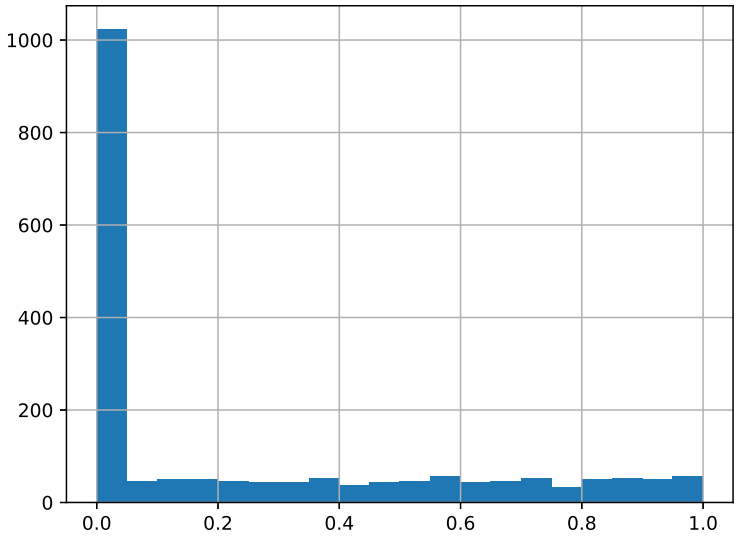
Wyznaczono wartość współczynnika c:

$$50 \cdot \frac{1}{100} + c \cdot \left(1 - \frac{1}{100}\right) = 1$$

$$c = \frac{50}{100} \cdot \frac{100}{99} = \frac{50}{99}$$



Rysunek 27: Wykres gęstości prawdopodobieństwa $f(x)$ rozkładu 2 i funkcji aproksymacyjnej $g(x)$



Rysunek 28: Histogram wygenerowanych liczb pseudolosowych dla rozkładu 2

2.3 Rozkład półokręgu

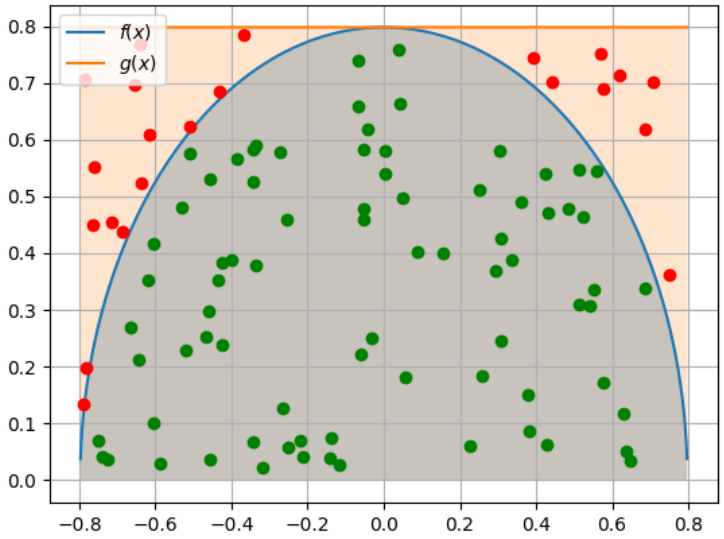
Zaimplementowano generator generujący liczby pseudolosowe z rozkładu o gęstości prawdopodobieństwa

$$f(x) = \sqrt{r^2 - x^2}$$

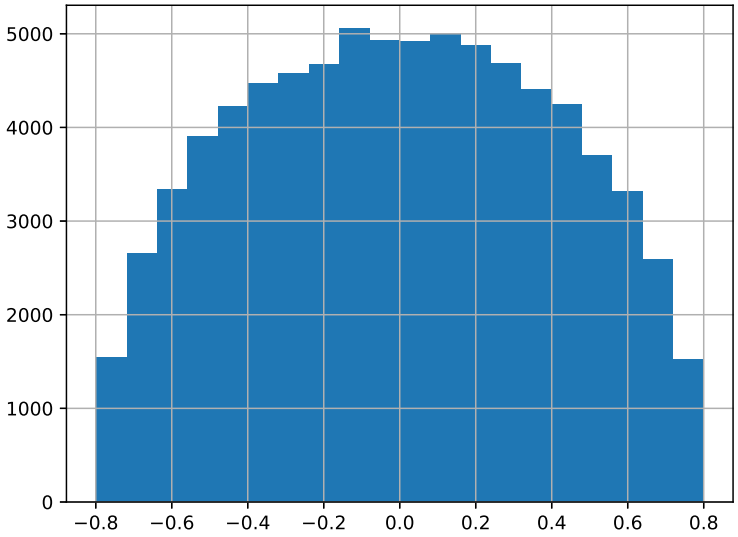
Wyznaczono wartość promienia półokręgu r:

$$\frac{\pi r^2}{2} = 1$$

$$r = \sqrt{\frac{2}{\pi}} \approx 0.79788$$



Rysunek 29: Wykres gęstości prawdopodobieństwa $f(x)$ rozkładu w kształcie półokręgu i funkcji aproksymacyjnej $g(x)$

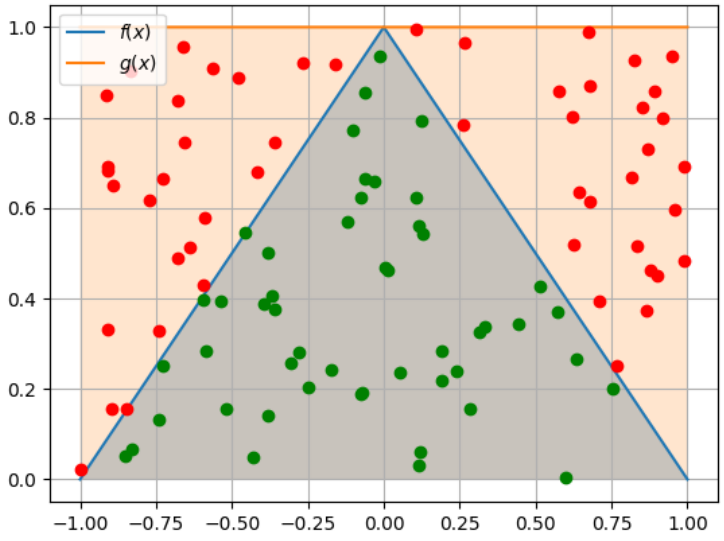


Rysunek 30: Histogram wygenerowanych liczb pseudolosowych dla rozkładu w kształcie półokręgu

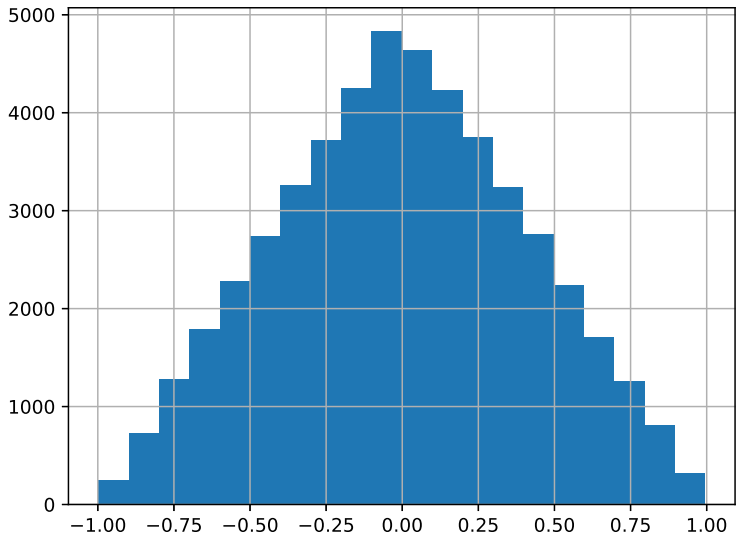
2.4 Rozkład normalny

Zaimplementowano generator generujący liczby pseudolosowe z rozkładu normalnego $\mathcal{N}(0,1)$ o gęstości prawdopodobieństwa

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$



Rysunek 31: Wykres gęstości prawdopodobieństwa $f(x)$ rozkładu normalnego $\mathcal{N}(0,1)$ i funkcji aproksymacyjnej $g(x)$



Rysunek 32: Histogram wygenerowanych liczb pseudolosowych dla rozkładu normalnego $\mathcal{N}(0,1)$

2.5 Wnioski

Metoda eliminacji jest metodą prostszą w użyciu niż metoda odwracania dystrybuanty, ponieważ nie wymaga wyznaczania odwrotności dystrybuanty. Jest jednak metodą stratną, ponieważ część losowanych wartości jest odrzucana. Skuteczność metody można poprawić poprzez zastosowanie metody hybrydowej - zamiast ograniczać funkcję gęstości prawdopodobieństwa obszarem prostokątnym, możemy ograniczyć ją inną funkcją pomocniczą, z której wartości losujemy za pomocą metody odwracania dystrybuanty. Metoda ta wymaga co prawda wyznaczenia odwrotności dystrybuanty funkcji pomocniczej, może ona jednak być dużo prostsza do wyznaczenia od odwrotności dystrybuanty funkcji gęstości prawdopodobieństwa, z której losujemy.

3 Laboratorium 3 - estymacja

W ćwiczeniu zbadano działanie estymatora wartości oczekiwanej, estymatora wariancji obciążonego i estymatora wariancji nieobciążonego dla rozkładu normalnego i rozkładu Cauchy'ego.

3.1 Estymacja rozkładu normalnego

Dla ciągu zmiennych losowych $T = X_1, X_2, \ldots, X_N$ z rozkładu normalnego $\mathcal{N}(0,16)$ zaimplementowano:
estymator wartości oczekiwanej:

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

estymator wariancji obciążony:

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

oraz estymator wariancji nieobciążony:

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

n	$\hat{\mu}_N$	$\hat{s}_N^2$	$\hat{S}_N^2$
100	0.42146669306742773	15.4706333875469	15.626902411663535
1000	0.05477120087449352	16.164785637941552	16.1809666045461
10000	-0.008719186870282721	15.820313801414944	15.821895991014046

Tabela 2: Wyniki estymacji dla n próbek

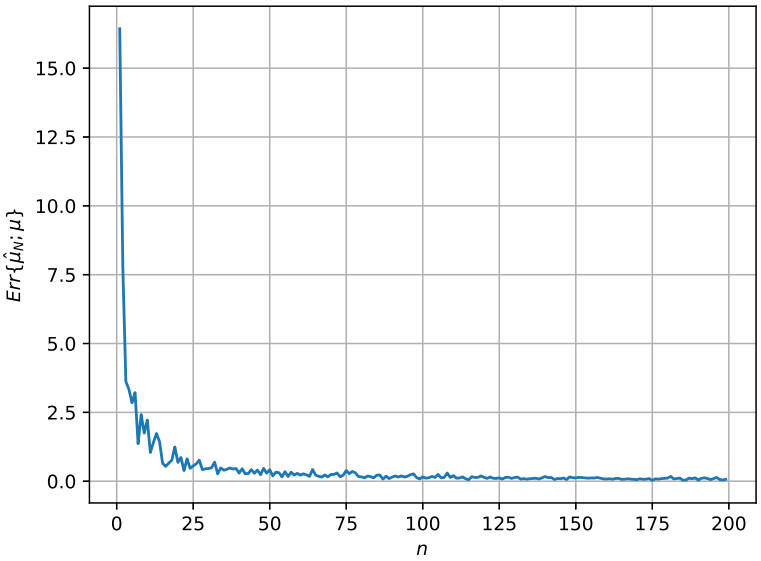
3.2 Błąd empiryczny

3.2.1 Błąd empiryczny estymatora wartości oczekiwanej

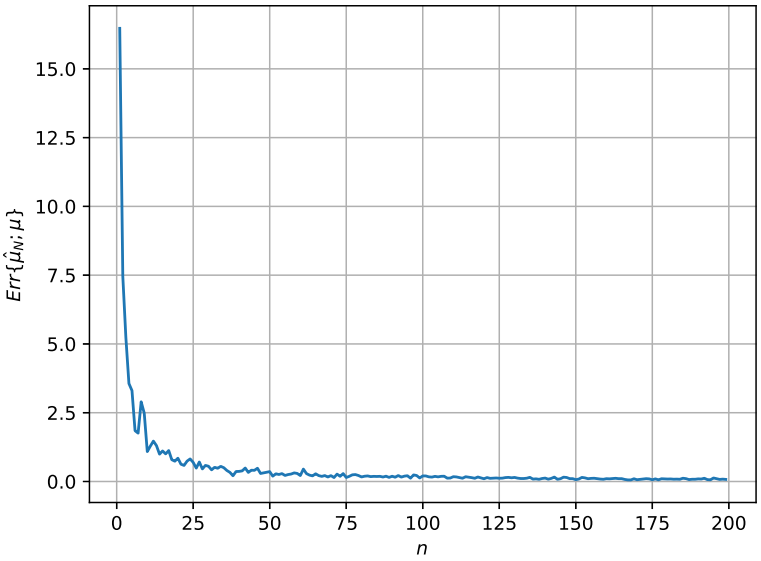
Wykreślono błąd empiryczny estymatora wartości oczekiwanej

$$Err\{\hat{\mu}_N; \mu\} = \frac{1}{L} \sum_{l=1}^{L} \left[\hat{\mu}_N^{[l]} - \mu\right]^2$$

dla rozkładu normalnego $\mathcal{N}(0, 16)$ i parametrów $L = 20$ i $L = 50$:



Rysunek 33: Wykres błędu empirycznego estymatora wartości oczekiwanej dla $L = 20$



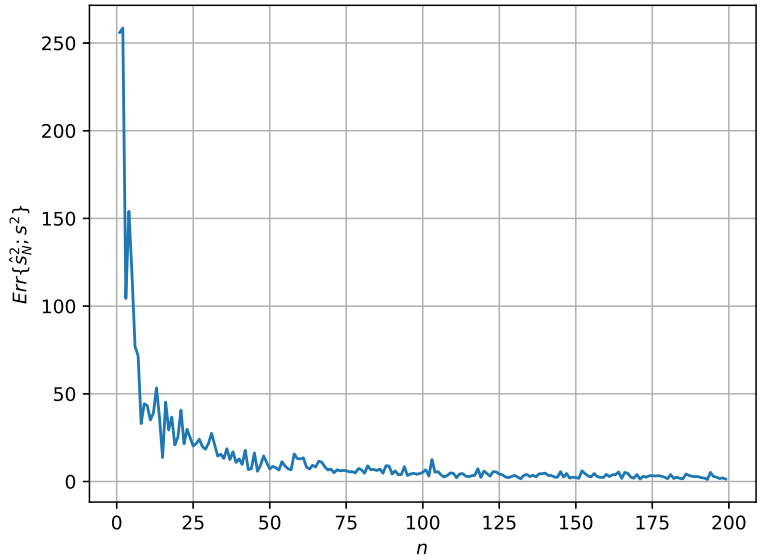
Rysunek 34: Wykres błędu empirycznego estymatora wartości oczekiwanej dla $L = 50$

3.2.2 Błąd empiryczny estymatora wariancji obciążonego

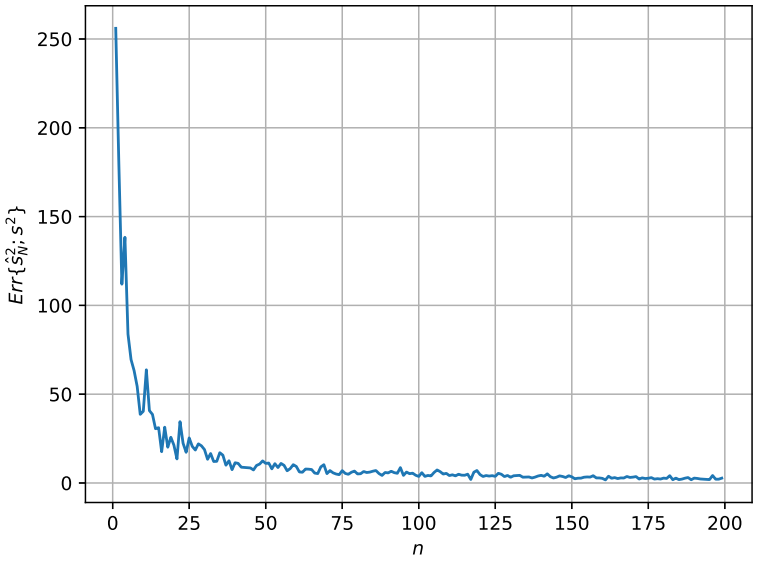
Wykreślono błąd empiryczny estymatora wariancji obciążonego

$$Err\{\hat{s}_N^2; s^2\} = \frac{1}{L}\sum_{l=1}^{L}\left[\hat{s}_N^{[l]} - \sigma\right]^2$$

dla rozkładu normalnego $\mathcal{N}(0, 16)$ i parametrów $L = 20$ i $L = 50$:



Rysunek 35: Wykres błędu empirycznego estymatora wariancji obciążonego dla $L = 20$



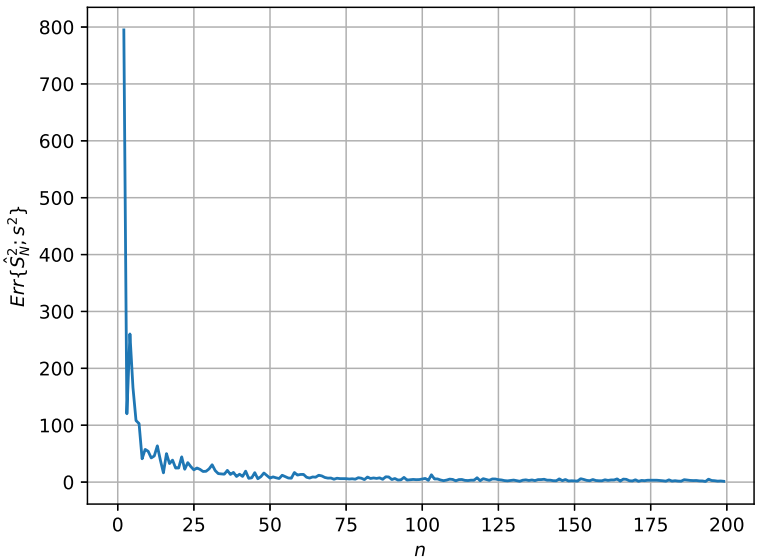
Rysunek 36: Wykres błędu empirycznego estymatora wariancji obciążonego dla $L = 50$

3.2.3 Błąd empiryczny estymatora wariancji nieobciążonego

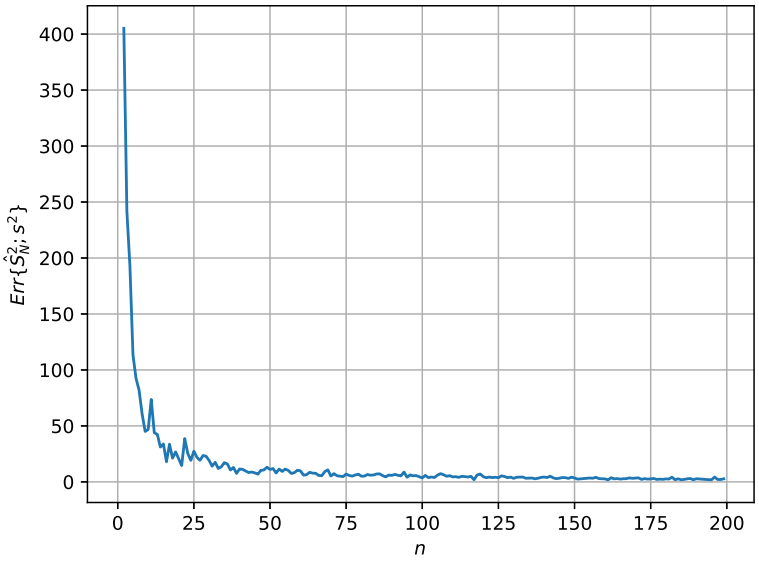
Wykreślono błąd empiryczny estymatora wariancji nieobciążonego

$$Err\{\hat{S}_N^2; s^2\} = \frac{1}{L}\sum_{l=1}^{L}\left[\hat{S}_N^{[l]} - \sigma\right]^2$$

dla rozkładu normalnego $\mathcal{N}(0, 16)$ i parametrów $L = 20$ i $L = 50$:



Rysunek 37: Wykres błędu empirycznego estymatora wariancji nieobciążonego dla $L = 20$



Rysunek 38: Wykres błędu empirycznego estymatora wariancji nieobciążonego dla $L = 50$

3.3 Estymacja rozkładu Cauchy'ego

Dla ciągu zmiennych losowych $T = X_1, X_2, \ldots, X_N$ z rozkładu Cauchy'ego zaimplementowano: estymator wartości oczekiwanej:

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

estymator wariancji obciążony:

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

oraz estymator wariancji nieobciążony:

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

n	$\hat{\mu}_N$	$\hat{s}_N^2$	$\hat{S}_N^2$
100	0.6294504530967895	42.37477922724003	42.802807300242456
1000	3.5239865670727153	14717.979817746458	14732.712530276734
10000	-1.3095096287854446	33692.450305228595	33695.819887217316

Tabela 3: Wyniki estymacji dla n próbek

3.4 Wnioski

W przypadku rozkładu normalnego wartość oczekiwana wyznaczona przez estymator $\hat{\mu}_N$ oscyluje wokół rzeczywistej wartości oczekiwanej wygenerowanego rozkładu - wartość najbliższą realnej uzyskano dla $n = 10000$. Wartości wariancji wyznaczone przez estymator wariancji obciążony i nieobciążony oscylują wokół prawidłowej wartości wariancji rozkładu, przy czym przy wyższej liczbie próbek uzyskano wynik bliższy realnej wartości.

W przypadku rozkładu Cauchy'ego wyniki estymacji odbiegają znacząco od oczekiwanych wartości. Wynika to z charakterystyki rozkładu Cauchy'ego, którego wartość oczekiwana i wariancja są niezdefiniowane. Można również zauważyć, że przy zwiększeniu liczby realizacji estymatora (zwiększeniu parametru L) wykresy błędu empirycznego estymatorów zbiegają do wartości 0 w sposób dokładniejszy i bardziej jednostajny„ co wynika ze zwiększenia dokładności estymacji wraz ze wzrostem liczby realizacji estymatora.