## Sprawozdanie z zadania 4 - Metody Probabilistyki i Statystyki

1. Celem pierwszego podpunktu pracy jest porównanie “ogonów” otrzymanych przy użyciu nierówności Markowa i Czebyszewa z dokładną wartością prawdopodobieństwa, obliczoną poprzez wykorzystanie pakietów matematycznych.

Rozważmy następującą zmienną losową X:

$$X \sim Bin(n, \frac{1}{2}).$$

Dla trzech wartości N:

$$N \in \{100,\ 1000,\ 10000\}$$

I następujących ogonów:

$$P(X \geq \frac{6}{5}E(X)) \text{ (1)}$$

$$P(|X - E(X)| \geq \frac{1}{10}E(X))\text{(2)}$$

Na początku zostaną przeprowadzone obliczenia teoretyczne dla pierwszego z prawdopodobieństw z użyciem nierówności Markowa:

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Podstawiając otrzymujemy:

$$P(X \geq \frac{6}{5}E(X)) \leq \frac{E(X)}{\frac{6}{5}E(X)}$$

Skracamy:

$$P(X \geq \frac{6}{5}E(X)) \leq \frac{5}{6}$$

Otrzymany wynik w przyszłości porównamy z dokładną wartością.

Następnie ograniczenie drugiego z “ogonów” zostanie policzone nierównością Czebyszewa. Skorzystamy z przedstawionego na wykładach wzoru:

$$P(|X - E(X)| \geq k) \leq \frac{varX}{k^2}$$

Co więcej znając rodzaj rozkładu zmiennej losowej łatwo wyprowadzić wzór na wariancje:

$$var(X) = np(1 - p) \Rightarrow var(X) = \frac{n}{4}$$

Po podstawieniu wartości i wykonaniu odpowiednich skróceń dostajemy:

$$P(|X - E(X)| \geq \frac{1}{10} E(X)) \leq \frac{100}{n}.$$

Dokładne wartości prawdopodobieństwa obliczone zostały z wykorzystaniem załączonej klasy First. Używa ona odpowiednich pakietów matematycznych języka C++. Działanie programu opiera się na prawie wielkich liczb - brana jest losowa liczba z rozkładu Bin(n,½) (dla danego n). Każdy wynik, większy od k * E(X) (gdzie k to wybrany współczyynik) jest odnotowywany. Następnie otrzymany "counter" dzielony jest przez liczbę prób. W poniższej tabeli przedstawione zostały wyniki tego eksperymentu.

| | Ograniczenie | Dokładna wartość N = 100 | Dokładna wartość N = 1000 | Dokładna wartość N = 10000 |
|---|---|---|---|---|
| podpunkt A | ⅚ | 0,02844 | 0 | 0 |
| podpunkt B | 100/n | 0,27128 | 0,00173 | 0 |

Tabela 1. Wyniki dla Zadania 1.

Na podstawie tego zadania można wyciągnąć następujące wnioski:

1. Ograniczenie Markowa nie jest efektywne - w tym przypadku mogło by się okazać przydatne tylko dla N << 100.
2. W przeciwieństwie do ograniczenia Markowa, nierówność Czebyszewa działa najlepiej dla dużych liczb - widać, że im większa wartość N tym ograniczenie zbliża się do dokładnej wartości.

Podsumowując nierówność Markowa daje lepsze oszacowanie dla małych N, podczas gdy nierówność Czebyszewa sprawdza się dla dużych wartości.

**2. Błądzenie losowe na liczbach całkowitych.**

W drugim podpunkcie celem było przeanalizowanie błądzenia losowego na liczbach całkowitych i porównanie wyników z odpowiadającym danemu przykładowi rozkładem normalnym. Mamy:

$S_n = \sum_{n=1}^{N} X_n$ gdzie, $X_n$ niezależne i każda z nich przyjmuje wartości -1 lub 1.

Najpierw obliczone zostaną wartość oczekiwana i odchylenie standardowe rozkładu normalnego przybliżającego $S_n$. Z niezależności:

$$var(S_n) = \sum_{n=1}^{N} var(X_n)$$

$$var(X_n) = E(X_n^2) - E(X_n)^2$$

$$E(X_n) = \frac{1}{2} * (-1) + \frac{1}{2} * 1 = 0$$

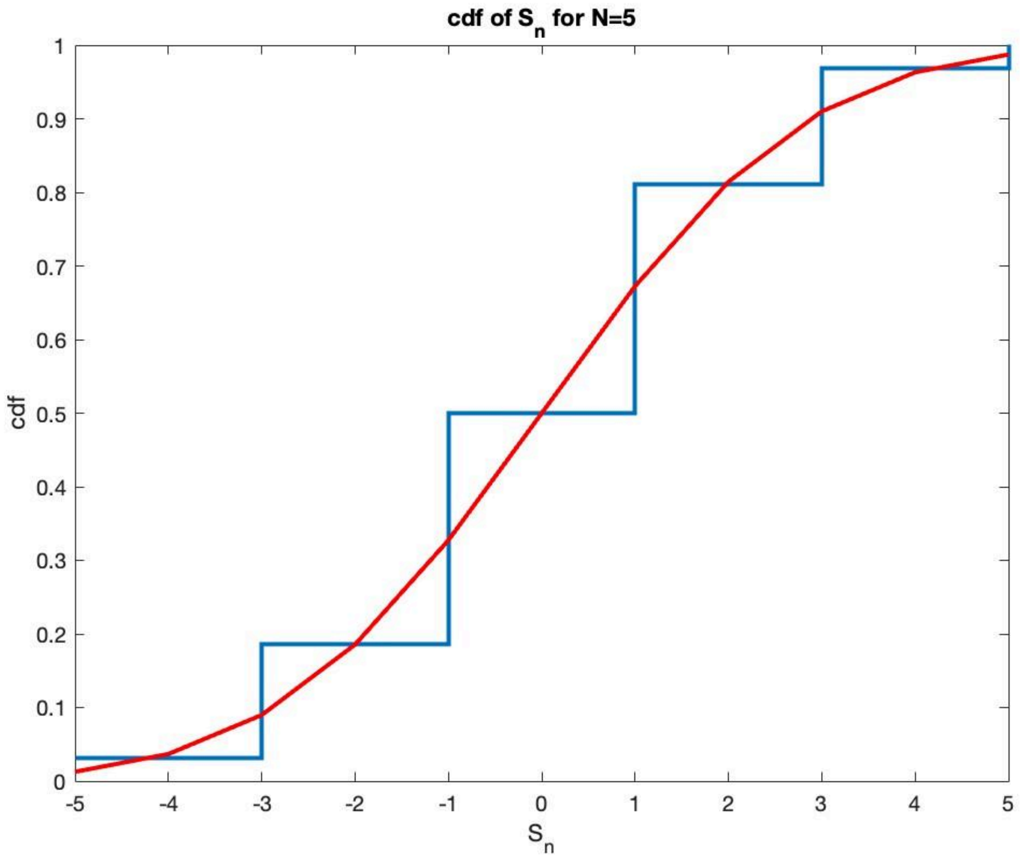
$X_n^{\ 2} = 1$ bo $X_n = \pm 1$. Więc:

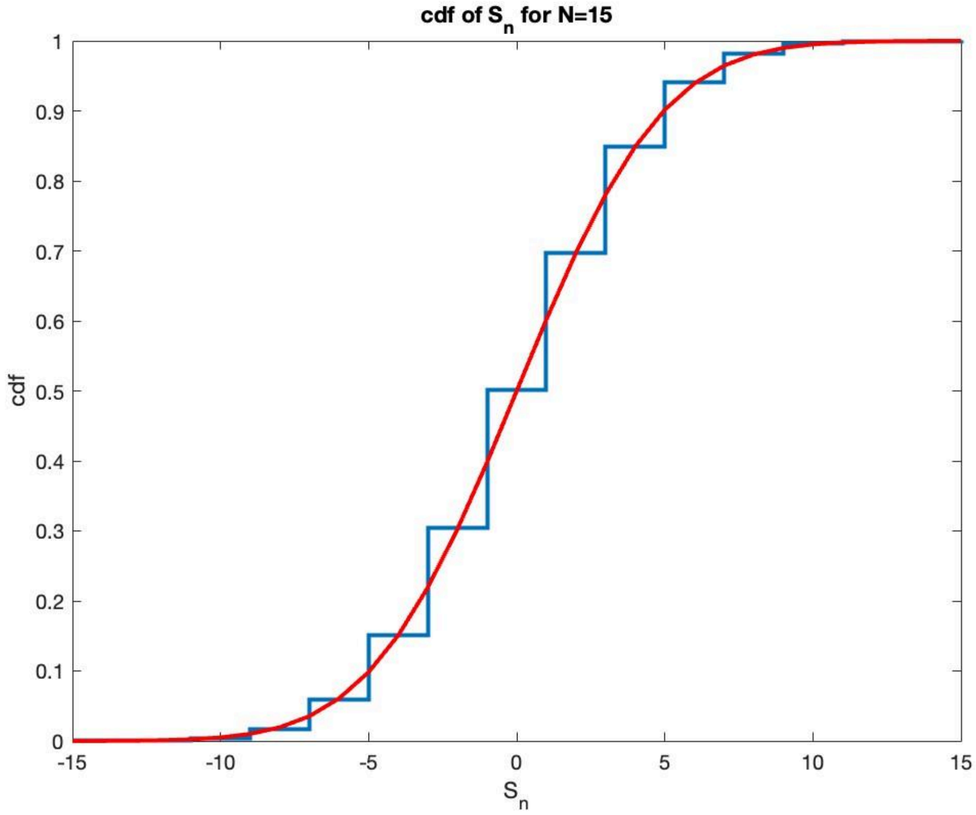
$$var(X_n) = 1$$

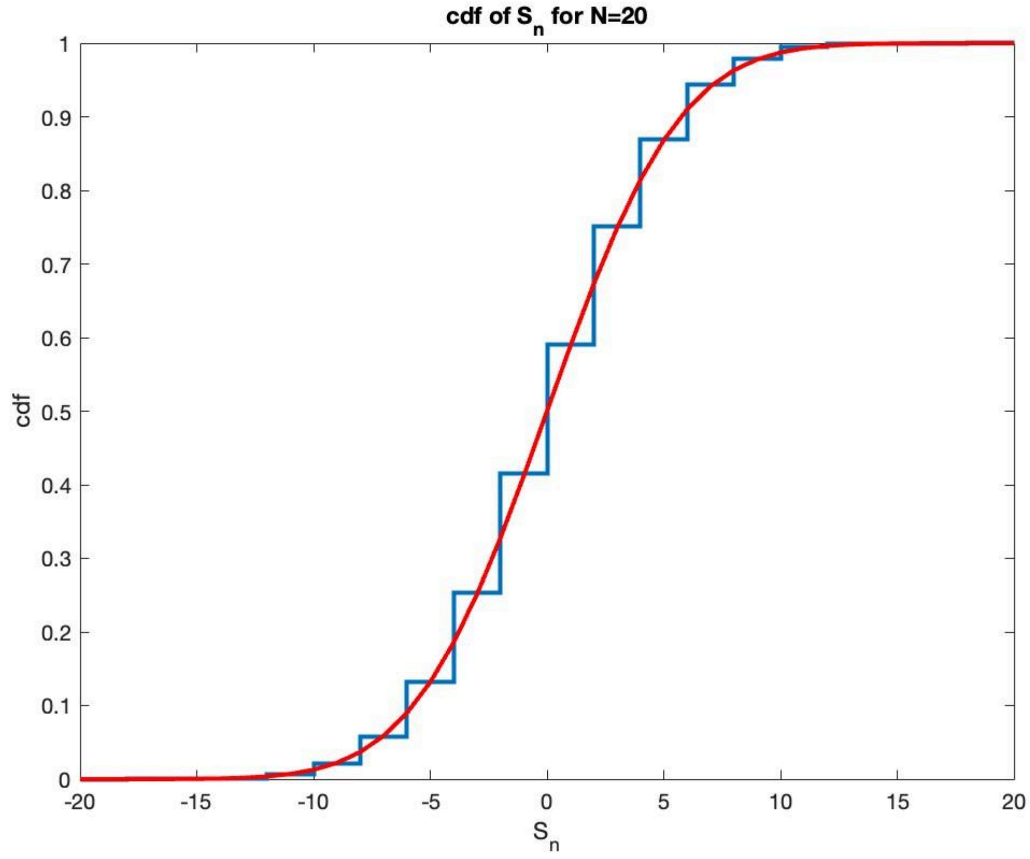
$$var(S_n) = \sum_{n=1}^{N} var(X_n) = \sum_{n=1}^{N} 1 = N$$

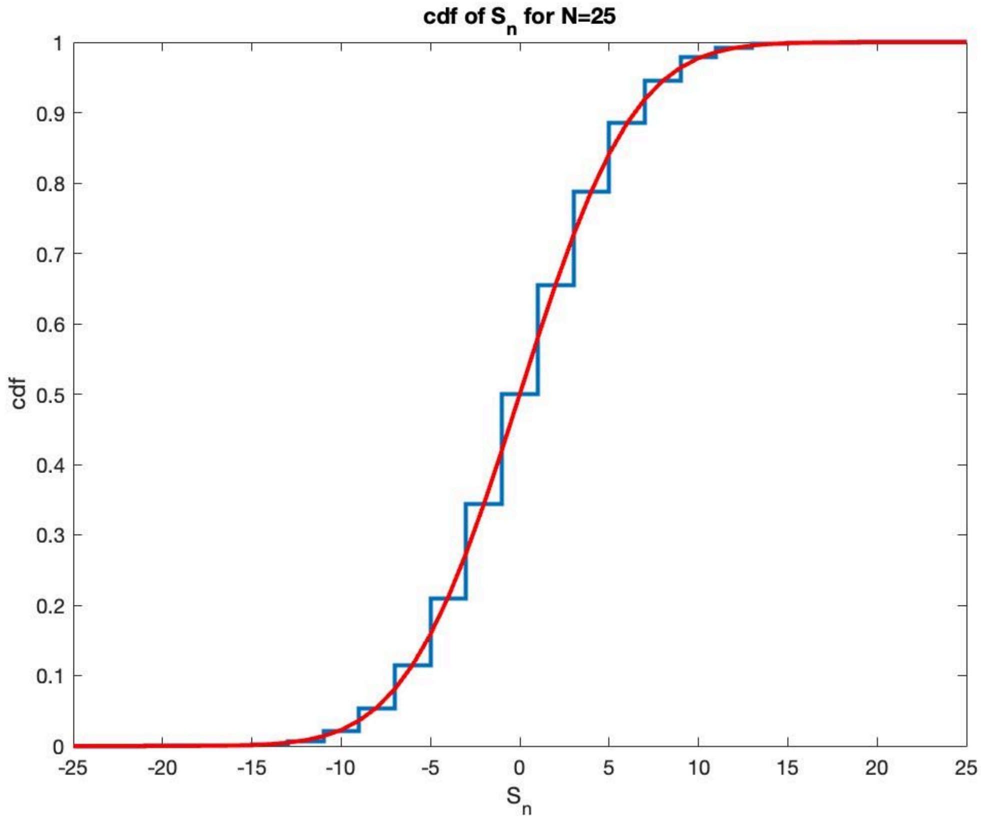
$$\sigma = \sqrt{N}$$

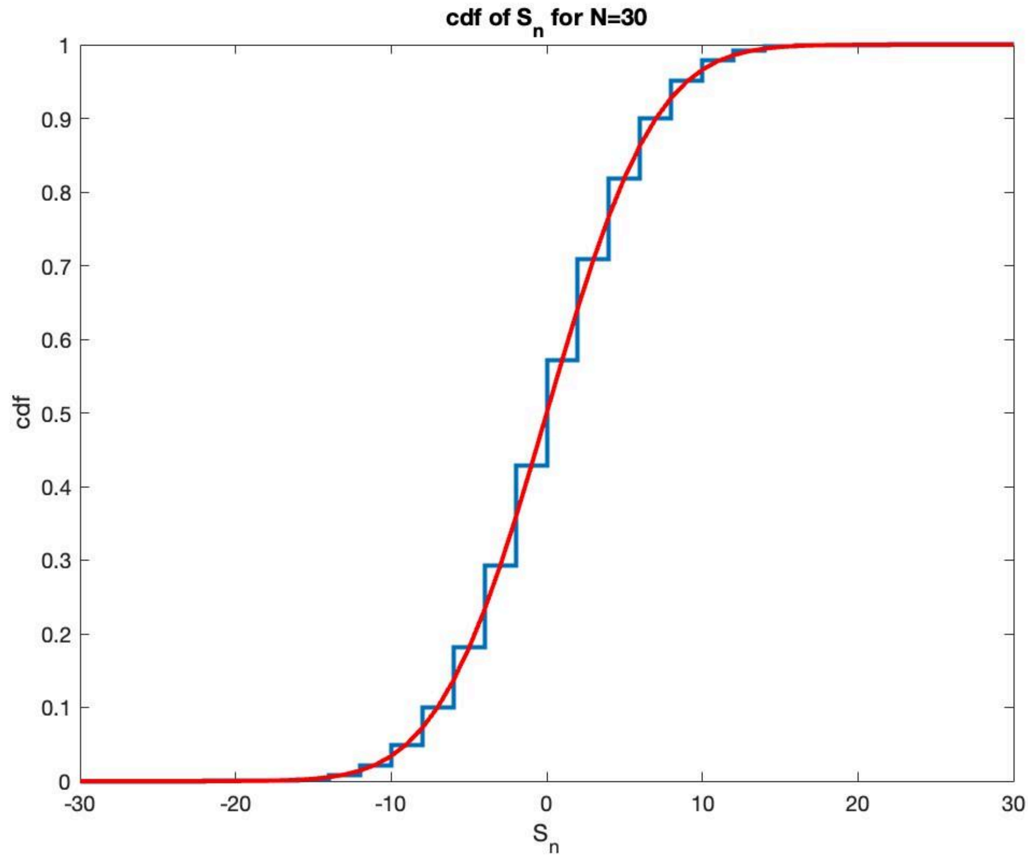
Następnie przygotowany został kod, który oblicza $S_n$. Wyniki tych numerycznych eksperymentów zostały przeniesione do Matlaba i obrobione - ostatecznie otrzymano 7 dystrybuant.

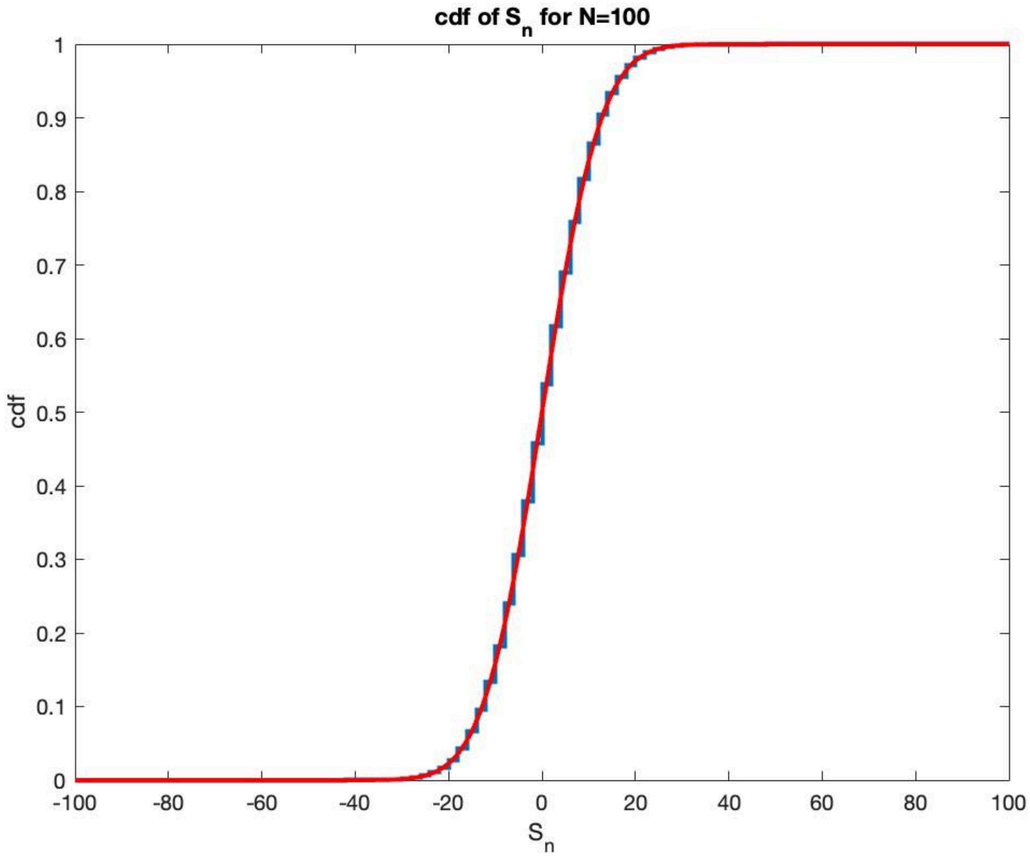



cdf of $S_n$ for N=15
cdf
$S_n$
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-15
-10
-5
0
5
10
15


cdf of $S_n$ for N=20
cdf
$S_n$
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-20
-15
-10
-5
0
5
10
15
20


cdf of $S_n$ for N=25
cdf
$S_n$
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-25
-20
-15
-10
-5
0
5
10
15
20
25


cdf of $S_n$ for N=30
cdf
$S_n$
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-30
-20
-10
0
10
20
30




Z powyższych wykresów wyciągnąć można następujący wniosek:
Ponieważ podobna część wykresu $S_n$ znajduje się zarówno nad jak i pod funkcją rozkładu normalnego (sytuacja ta przypomina w pewnym sensie aproksymacje całką Riemanna), $S_n$ jest dobrze przybliżana przez odpowiedni rozkład normalny:

$$normaldistribution(0, \sqrt{N})$$

Szczególnie ten wniosek jest prawdziwy dla dużych N - widać na wykresach, że wraz z wrostem wartości N, schodki zbliżają się do funkcji.

**3. Błądzenie losowe na liczbach całkowitych - rozkład czasu spędzonego nad osią OX**

Trzeci podpunkt wymagał lekkiej modyfikacji napisanego już kodu do poprzedniego podpunktu, tak by rejestrował on czas gdy:

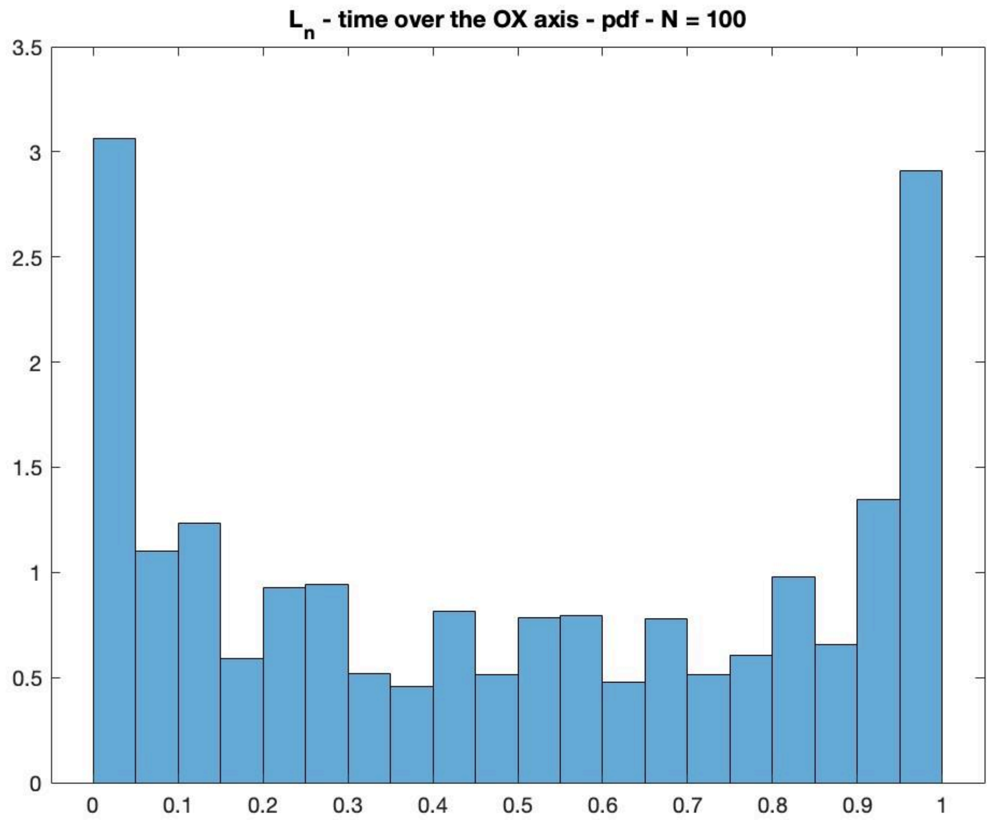
$$S_n > 0 \vee S_{n-1} > 0$$

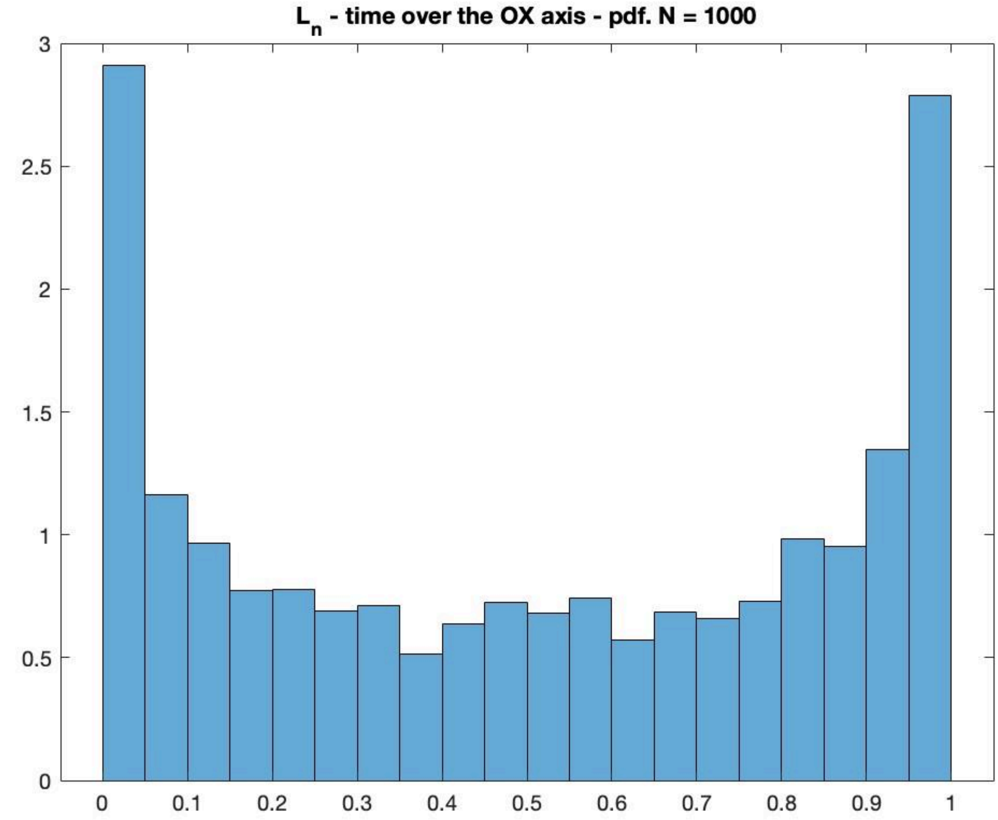
Zebrane eksperementalnie stosunki - czasu spędzonego nad osią OX z łącznym czasem - zostały następnie wprowadzone do Matlaba, gdzie z użyciem funkcji histogram zostały znormalizowane, tak by prezentowały wartość pdf.

W celu zbadania zachowania zmiennej losowej w zależności od parametru N, wykonano trzy eksperymenty dla:

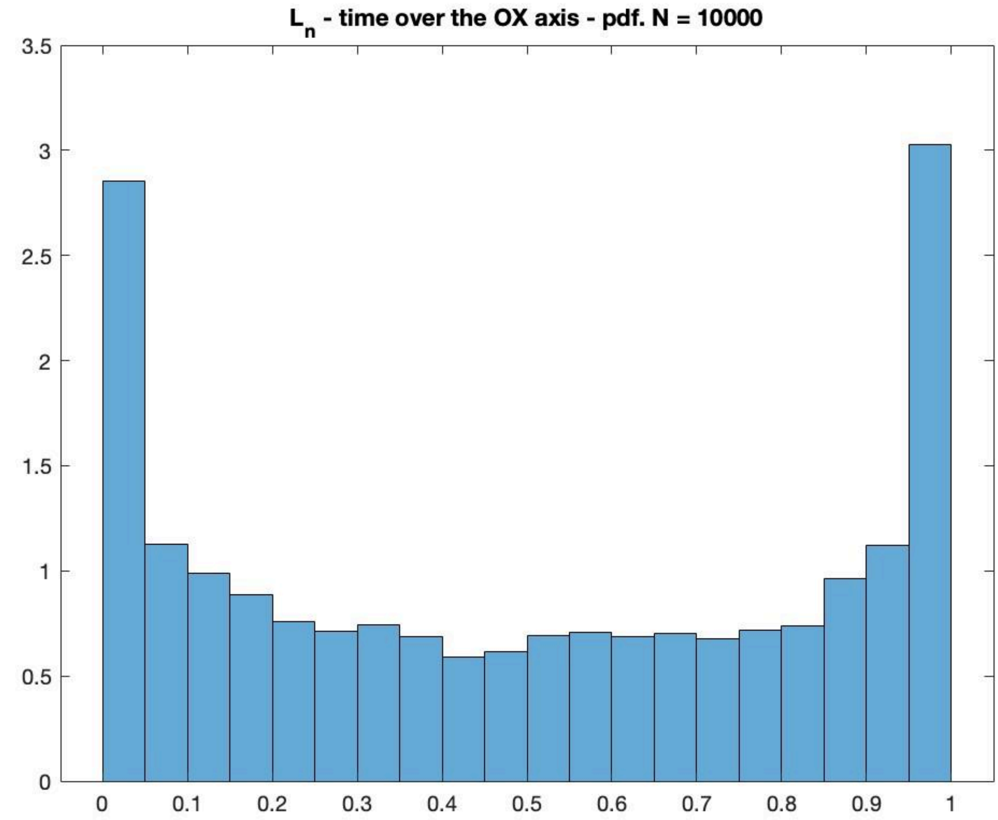
$$N \in \{100, 1000, 10000\}$$

Poniżej przedstawione zostanały wykresy:





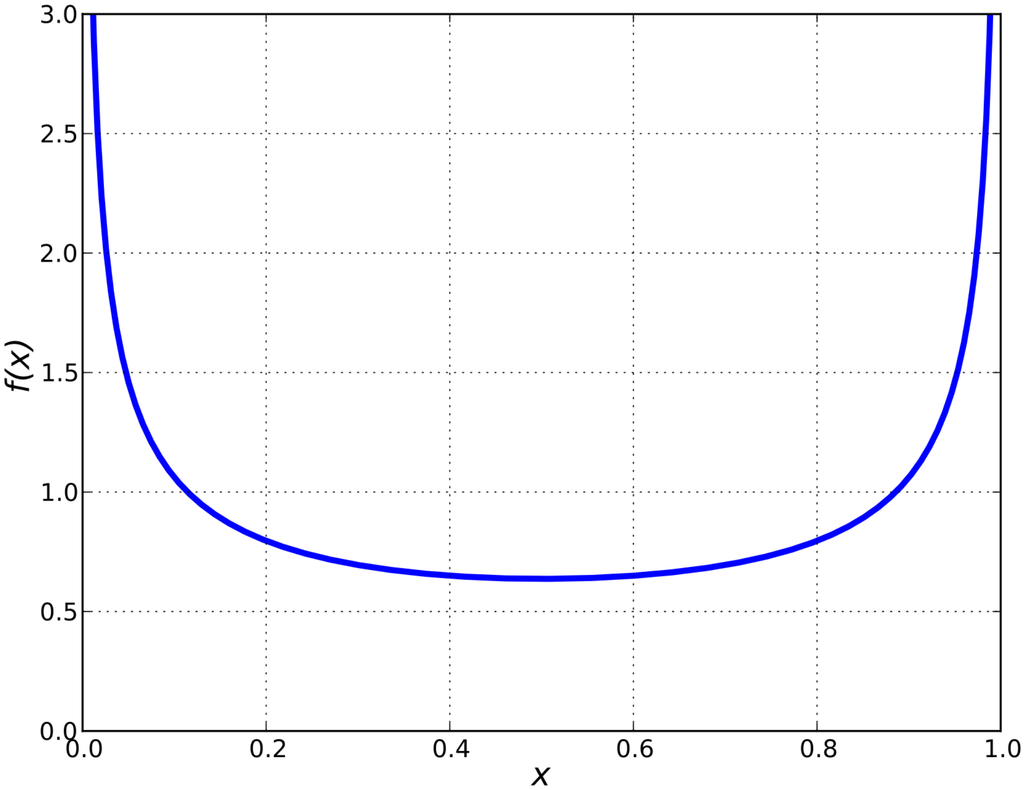
$L_n$ - time over the OX axis - pdf. N = 1000
3
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

$L_n$ - time over the OX axis - pdf. N = 10000
3.5
3
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Na wykresach widać, że mają one powtarzalny trend i wartości. Jednakże, dla większych N odczyt się stabilizuje, szczególnie w centrum wykresu. Przypomina on wykres gęstości arcusa sinusa.

Porównując poniższy wykres z wynikami eksperymentu widzimy, że tak jak w przypadku podpunktu drugiego im większe n tym lepsza aproksymacja. Ewidentnie wraz z wzrostem n wartości układają sie w łuk odwzorowujący funkcję arcus sinus..



[1]

**Kajetan Plewa**

[1] http://en.wikipedia.org/wiki/Arcsine_distribution#/media/File:Arcsin_density.svg [Accessed: 29.01.2025]