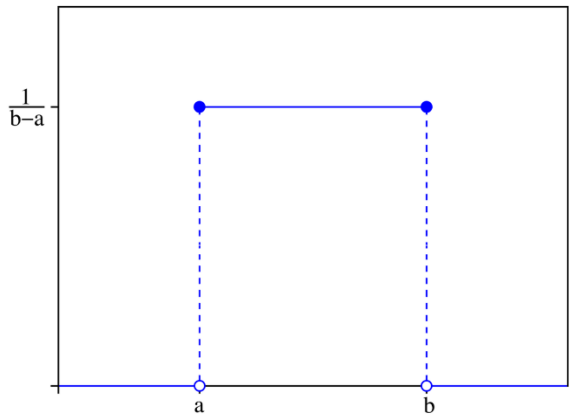
Akademia Górniczo-Hutnicza w Krakowie
Wydział Fizyki i Informatyki Stosowanej
Bartosz Rogowski, II rok, 303793
21 czerwca 2020

Metody numeryczne
Sprawozdanie nr 14 z zajęć laboratoryjnych
Generatory liczb pseudolosowych

1. Wstęp teoretyczny

Losowość oznacza brak porządku, jednoznacznego przewidywalnego rozwiązania. Zatem **liczbę losową** można zdefiniować jako taką, która jest wybrana w sposób zupełnie przypadkowy, niedający się w żaden sposób przewidzieć. Liczby takie są często potrzebne w wielu algorytmach, gdyż nie wszystkie zdarzenia są doskonale opisane przez prawa fizyki (np. kwantowej), w związku z czym niemożliwym staje się określenie rezultatu. W rzeczywistości jednak nie da się zaimplementować programu komputerowego, który zupełnie przypadkowo wybierałby liczbę z zadanego przedziału. Można co prawda powiązać losowanie z jakimś zjawiskiem (np. rozpady promieniotwórcze), jednak program nie wylosuje ich „sam z siebie". Generatory liczb (pseudo-) losowych często „wybierają" na podstawie zegara systemowego (czas zapisany w komputerze jest przecież „zmienny w każdej chwili") za pomocą ziarna startowego lub innych sposobów. Generatory te mogą tworzyć różne rozkłady.

Rozkład jednostajny (też: **jednorodny**) to taki rozkład, w którym w danym określonym przedziale gęstość prawdopodobieństwa jest stała i niezerowa (rys. 1).



Rysunek 1. Funkcja gęstości prawdopodobieństwa w **rozkładzie jednorodnym** jest stała i niezerowa w zadanym przedziale. Poza nim jest zerowa [1].

Oznaczany jest jako:

$$U(a, b), \tag{1}$$

Oznaczenia:
a, b – końce przedziału

jednak często stosuje się rozkład $U(0,1)$. Innymi słowy, każda liczba z danego przedziału ma jednakowe prawdopodobieństwo bycia wylosowaną. Jednak nie zawsze możemy chcieć, aby generator posiadał tę cechę.

Ciąg liczb z rozkładu jednorodnego można wyznaczyć poprzez użycie np. **generatora mieszanego** (uwaga: liczby są normowane do przedziału $[0,1]$):

$$x_{i+1} = \frac{(a \cdot x_i + c)\ mod\ m}{m+1}. \tag{2}$$

Oznaczenia:
a, c, m – parametry generatora

Zazwyczaj wyznaczanie liczb startuje od pewnej początkowej wartości x_0 nazywanej ziarnem.

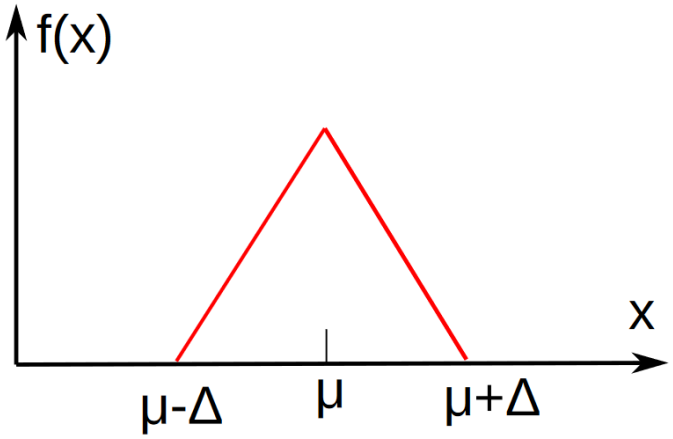
Innym generatorem, który nie wykazuje wspomnianej wyżej własności, jest ten o **rozkładzie trójkątnym**, oznaczanym jako:

$$T(\mu, \Delta). \tag{3}$$

Oznaczenia:
μ – środek rozkładu
Δ – szerokość rozkładu

Funkcja gęstości prawdopodobieństwa tego rozkładu jest opisana wzorem:

$$f(x:\ \mu, \Delta) = -\frac{|x-\mu|}{\Delta^2} + \frac{1}{\Delta}. \tag{4}$$



Rysunek 2. Wykres funkcji gęstości prawdopodobieństwa **rozkładu trójkątnego** (symetrycznego) [2].

Jego nazwa pochodzi od kształtu tej funkcji – przypomina bowiem trójkąt (w tym przypadku: równoramienny).

Dystrybuanta tego rozkładu to natomiast:

$$F(a) = P(x < a) = \int_{\mu-\Delta}^{a} f(x:\ \mu, \Delta)\, dx = \begin{cases} -\frac{1}{\Delta^2} \cdot \left(-\frac{x^2}{2} + \mu x\right) + \frac{x}{\Delta}, x \le \mu \\ -\frac{1}{\Delta^2} \cdot \left(\frac{x^2}{2} - \mu x + \mu^2\right) + \frac{x}{\Delta}, x > \mu \end{cases} \tag{5}$$

i w sposób jednoznaczny określa rozkład prawdopodobieństwa [*].

Liczbę z **rozkładu trójkątnego** można więc wyznaczyć jako:

[*] Dystrybuanta jest cechą dowolnego rozkładu, która jednoznacznie go wyznacza.

$$x = \mu + (\xi_1 + \xi_2 - 1) \cdot \Delta. \tag{6}$$

Oznaczenia:
μ – środek rozkładu
Δ – szerokość rozkładu
$\xi_1, \xi_2 \in U(0,1)$ – pseudolosowe liczby wygenerowane z rozkładu jednorodnego

Dla wylosowanych liczb można stworzyć statystyki, które dostarczają dodatkowych informacji o danym ciągu. Do podstawowych parametrów należą **średnia** [*] wszystkich elementów x_i:

$$\bar{\mu} = \frac{1}{N} \cdot \sum_{i=0}^{N-1} x_i \tag{7}$$

oraz **odchylenie standardowe**:

$$\sigma = \sqrt{\frac{1}{N} \cdot \sum_{i=0}^{N-1} (x_i - \bar{\mu})^2}\,. \tag{8}$$

Oznaczenia:
$\bar{\mu}$ – średnia

Kolejną ważną czynnością jest **testowanie hipotez**. Często wykonywanym testem jest **test χ^2** (**chi kwadrat**). Polega on na badaniu, czy wylosowane liczby z rozkładu rzeczywiście z takiego pochodzą. W tym celu konieczne jest podzielenie przedziału, w którym losowano liczby, na k podprzedziałów, a następnie „zakwalifikowanie" ciągu wygenerowanych liczb do odpowiedniego (tzn. sprawdzenie, do którego podprzedziału należą) oraz zliczenie wszystkich liczb w danym podprzedziale. Kolejnym krokiem jest obliczenie statystyki testowej:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - N \cdot p_i)^2}{N \cdot p_i}. \tag{9}$$

Oznaczenia:
n_i – liczba wygenerowanych liczb znajdująca się w i-tym podprzedziale
N – liczba wszystkich wygenerowanych liczb
p_i – teoretyczne prawdopodobieństwo znajdowania się zmiennej losowej w i-tym podprzedziale

Zmienną p_i ze wzoru (9) można wyznaczyć jako różnicę dystrybuant końców podprzedziałów:

$$p_i = F(x_{i,max}) - F(x_{i,min}). \tag{10}$$

Hipotezę H_0: *„Wygenerowany rozkład jest rozkładem …"* odrzuca się lub nie, porównując wartość statystyki χ^2 z wartością graniczną (też: krytyczną) ε danego rozkładu (odczytywaną z tablic statystycznych) dla określonej liczby stopni swobody:

$$\nu = k - r - 1 \tag{11}$$

Oznaczenia:
k – liczba podprzedziałów
r – liczba parametrów, od których zależy dany rozkład (np. dla trójkątnego $r = 2$, bo jest zależny od μ i Δ)

i założonego poziomu istotności α (często $\alpha = 0{,}05$). Jeżeli zachodzi warunek:

$$\chi^2 < \varepsilon_{\alpha,\nu}\,, \tag{12}$$

to nie ma podstaw do odrzucenia hipotezy [†], w przeciwnym wypadku hipoteza jest odrzucana.

[*] W statystyce średnia to estymator wartości oczekiwanej.
[†] Należy pamiętać, że w statystyce hipotezy nie można jednoznacznie potwierdzić. Co najwyżej można nie znaleźć podstaw do jej odrzucenia, ale i to nie zapewnia w 100%, że nie znaleziono żadnych błędów.

2. Zadanie do wykonania

2.1 Opis problemu

Naszym zadaniem było wygenerowanie kilku rozkładów:

- **jednorodnego $U(0,1)$** z pomocą dwóch generatorów mieszanych opisanych wzorem (2) o następujących parametrach:

pierwszy generator	drugi generator
$a = 123$	$a = 69069$
$c = 1$	
$m = 2^{15}$	$m = 2^{32}$
$N = 10000$ wygenerowanych liczb	
$k = 12$ podprzedziałów	

Tabela 1. Parametry podane w treści zadania dla dwóch generatorów mieszanych.

- **trójkątnego $T(\mu = 4, \Delta = 3)$** o następujących parametrach:

parametr	objaśnienie
$\mu = 4$	środek układu
$\Delta = 3$	szerokość układu
$N = 1000$	liczba wygenerowanych liczb
$k = 10$	liczba podprzedziałów

Tabela 2. Parametry podane w treści zadania dla generatora rozkładu trójkątnego.

Dla pierwszego rozkładu obliczono podstawowe **parametry statystyczne**:

- **wartość oczekiwaną** (patrz: wzór (7))
- **odchylenie standardowe** (patrz: wzór (8)),

natomiast dla drugiego – **przetestowano hipotezę** H_0: *„Wygenerowany rozkład jest trójkątny taki, że* $T(\mu = 4, \Delta = 3)$*”* za pomocą **testu χ^2**.

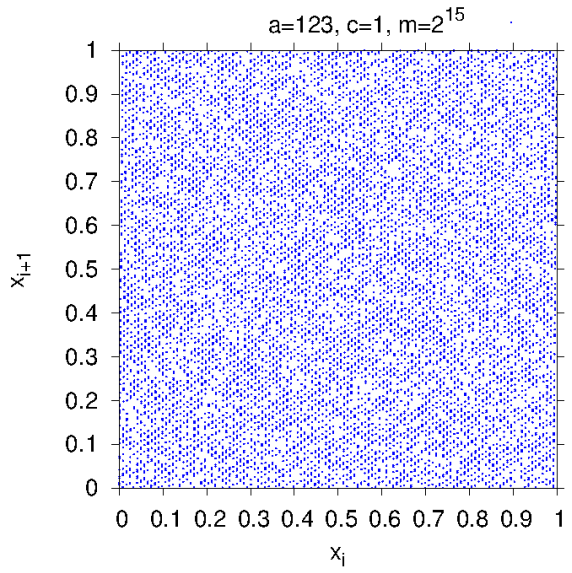
Wygenerowane liczby zostały zapisane do wektorów po to, by móc później wykonywać na nich różne operacje (np. stworzenie statystyk lub histogramów).

Do stworzenia wektora pseudolosowych liczb wykorzystano specjalną funkcję zwracającą unormowane liczby, które były do niego zapisywane. Aby zoptymalizować kod, do przechowywania x_i zastosowano zmienną statyczną, która była ustawiana na ziarno tylko podczas losowania pierwszego elementu danego ciągu.

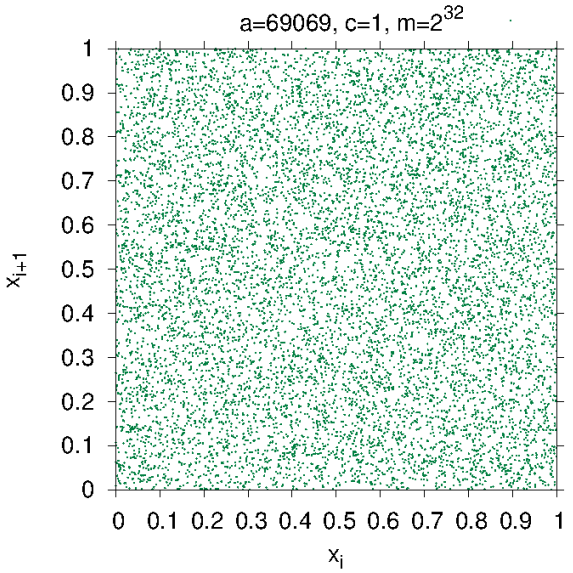
2.2 Wyniki

Łącznie wygenerowane zostały 4 pliki z różnymi danymi. Do pierwszego zostały zapisane elementy x_i oraz x_{i+1} rozkładu jednorodnego obu generatorów w celu sprawdzenia zależności między nimi. Do drugiego zapisano środek j-tego przedziału oraz iloraz $\frac{n_j}{N}$ (ozn. jak we wzorze (9)) dla obu ciągów stworzonych przez generatory mieszane (dane te są niezbędne do stworzenia histogramu). Do trzeciego natomiast zapisano środek j-tego przedziału, iloraz $\frac{n_j}{N}$ oraz wartość p_j (ozn. jak we wzorze (9)). Pozwoliło to na wygenerowanie wykresów zależności elementów od siebie oraz histogramów za pomocą specjalnego skryptu w programie Gnuplot.

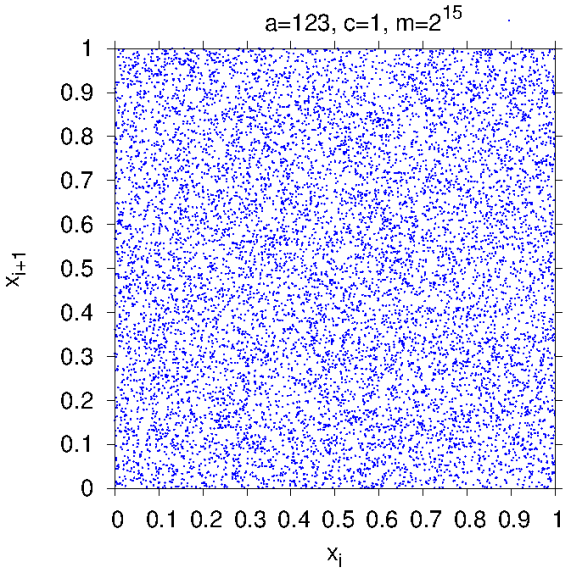
W celu porównania właściwości generatorów, do czwartego pliku zapisano liczby wygenerowane przy pomocy już zaimplementowanej funkcji *rand*, korzystającej z zegara systemowego.



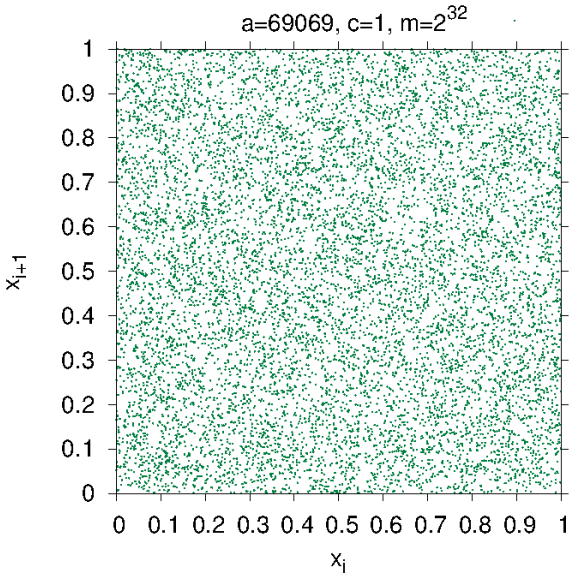
Rysunek 3. Zależność elementu od poprzedniego dla **generatora mieszanego** o podanych parametrach.



Rysunek 4. Zależność elementu od poprzedniego dla **generatora mieszanego** o podanych parametrach.



Rysunek 5. Zależność elementu od poprzedniego dla liczb wygenerowanych przez funkcję *rand*.



Rysunek 6. Zależność elementu od poprzedniego dla liczb wygenerowanych przez funkcję *rand*.

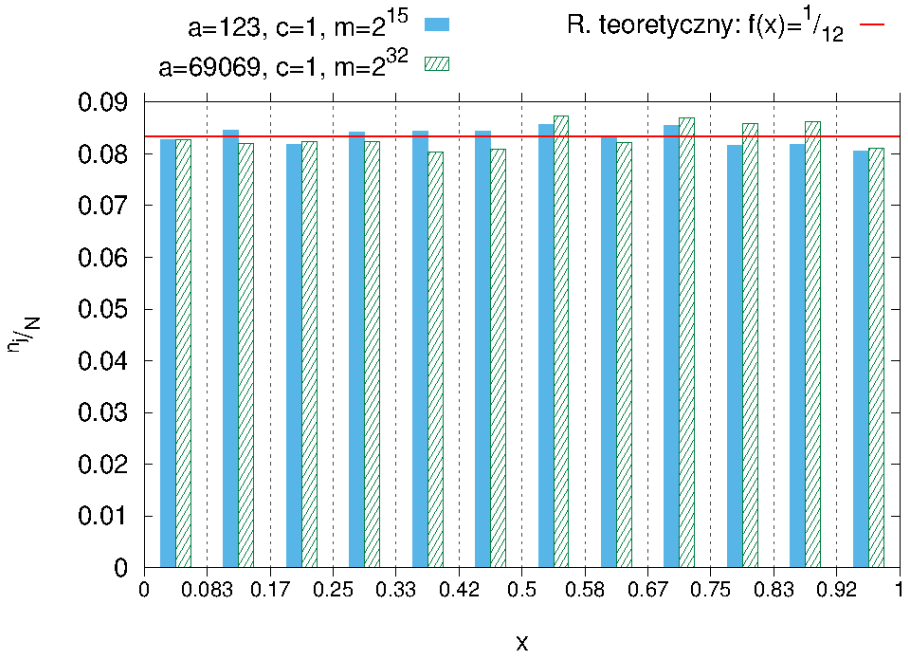
Obliczone parametry statystyczne zestawione są w poniższej tabeli.

pierwszy generator	wartość teoretyczna	drugi generator	wartość teoretyczna
$\bar{\mu} = 0{,}498266$	$\bar{\mu}_t = 0{,}5$	$\bar{\mu} = 0{,}503806$	$\bar{\mu} = 0{,}5$
$\lvert\bar{\mu} - \bar{\mu}_t\rvert = 1{,}734029 \cdot 10^{-3}$		$\lvert\bar{\mu} - \bar{\mu}_t\rvert = 3{,}805967 \cdot 10^{-3}$	
$\sigma = 0.28712$	$\sigma_t = \frac{1}{12}$	$\sigma = 0{,}28807$	$\sigma_t = \frac{1}{12}$
$\lvert\sigma - \sigma_t\rvert = 1{,}555442 \cdot 10^{-3}$		$\lvert\sigma - \sigma_t\rvert = 6{,}048147 \cdot 10^{-4}$	

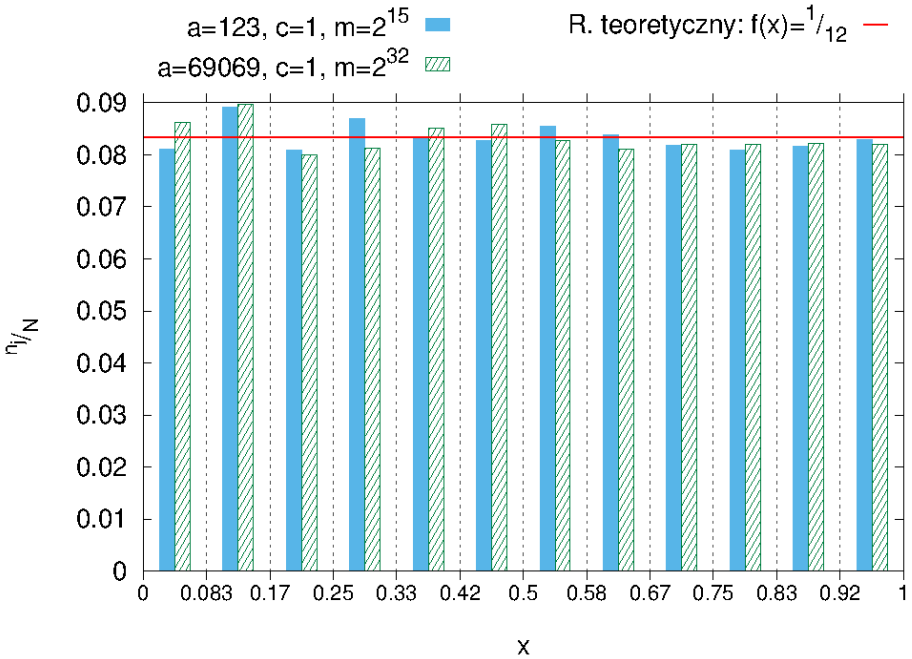
Tabela 3. Średnia oraz odchylenie standardowe dla dwóch **generatorów mieszanych o rozkładzie jednorodnym**.

pierwszy *rand*	wartość teoretyczna	drugi *rand*	wartość teoretyczna
$\bar{\mu} = 0{,}497132$	$\bar{\mu}_t = 0{,}5$	$\bar{\mu} = 0{,}495678$	$\bar{\mu} = 0{,}5$
$\lvert\bar{\mu} - \bar{\mu}_t\rvert = 2{,}867781 \cdot 10^{-3}$		$\lvert\bar{\mu} - \bar{\mu}_t\rvert = 4{,}321525 \cdot 10^{-3}$	
$\sigma = 0{,}288266$	$\sigma_t = \frac{1}{12}$	$\sigma = 0{,}28977$	$\sigma_t = \frac{1}{12}$
$\lvert\sigma - \sigma_t\rvert = 4{,}095436 \cdot 10^{-4}$		$\lvert\sigma - \sigma_t\rvert = 1{,}095353 \cdot 10^{-3}$	

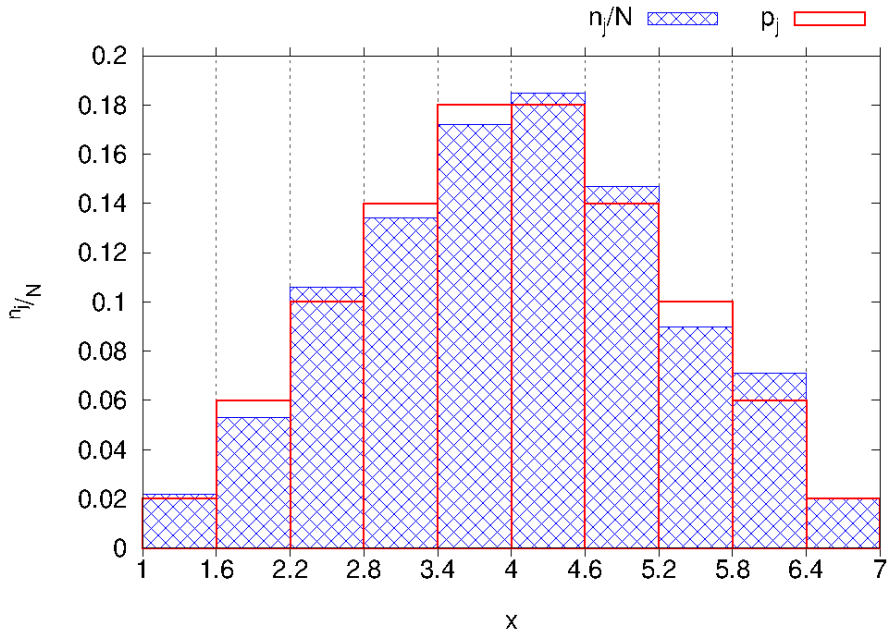
Tabela 4. Średnia oraz odchylenie standardowe dla dwóch ciągów wygenerowanych **funkcją *rand*.**



Rysunek 7. Histogram sporządzony dla liczb wygenerowanych przez dwa różne **generatory mieszane**. Zaznaczona czerwoną linią wartość oznacza teoretyczne położenie, do którego powinny dosięgać słupki.



Rysunek 8. Histogram sporządzony dla dwóch ciągów wygenerowanych **funkcją *rand***.



Rysunek 9. Histogram sporządzony dla liczb wygenerowanych przez generator **rozkładu trójkątnego $T(\mu = 4, \sigma = 3)$**. Słupki oznaczone kolorem czerwonym oznaczają teoretyczne położenia (wysokości), w jakich powinny się znajdować.

obliczona wartość		wartość tablicowa [3]
$\chi^2 = 5{,}49492$	$<$	$\varepsilon_{\alpha=0{,}05,\nu=7} = 14{,}06$

Tabela 5. Zestawienie wartości statystyki testowej z wartością graniczną na zadanym poziomie ufności $\alpha = 0{,}5$.

3. Wnioski

- Generowanie liczb prawdziwie losowych jest bardzo trudnym zadaniem dla komputera. Prawdziwie losowe liczby można uzyskać, stosując generatory fizyczne oparte na losowych zjawiskach występujących w naturze. Istnieją natomiast generatory pozwalające na wygenerowanie ciągów pseudolosowych liczb, które są w stanie dobrze imitować przypadkowość ich wybrania.
- Rysunki 3 i 4 pokazują, że pomiędzy wygenerowanymi liczbami nie da się wskazać jednoznacznej zależności (brak charakterystycznych kształtów znanych funkcji), ponadto wypełniają one niemal całą powierzchnię. W obu przypadkach można zauważyć, że uzyskane parametry statystyczne nie odbiegają zbytnio od wartości teoretycznych (dane w tabeli 3), co wskazuje na poprawność stworzonego rozkładu jednorodnego. Ponadto histogram na rys. 7 pokazuje, że wygenerowane ciągi nie odstają zbytnio od wartości teoretycznej, a zatem wygenerowane w ten sposób liczby są do zaakceptowania.
- Ciągi wygenerowane przez funkcję *rand* mają podobne właściwości do tych, otrzymanych z generatorów mieszanych, jednak rysunki 5 i 6 zdradzają, że zależności między kolejnymi elementami są „inaczej ułożone” (tendencja do „grupowania się punktów” – brak równomiernego rozłożenia w porównaniu do rozkładu jednorodnego) w porównaniu do tych z rys. 3 i 4. Ponadto analiza obu histogramów może wskazywać, że w niewielkim stopniu generatory mieszane dają lepsze rezultaty.

Jednak aby potwierdzić ten wniosek, trzeba byłoby wykonać znacznie więcej losowań. Wiele też zależy od celu, dla którego generowane są liczby pseudolosowe.

- Również w przypadku rozkładu trójkątnego można zauważyć na rys. 9, że wylosowane liczby podlegają rozkładowi trójkątnemu, co dodatkowo potwierdza wykonany test χ^2 – hipoteza, mówiąca o rozkładzie trójkątnym, nie ma podstaw do bycia odrzuconą (na zadanym poziomie istotności).

4. Bibliografia

[1] źródło: https://pl.wikipedia.org/wiki/Rozk%C5%82ad_jednostajny_ci%C4%85g%C5%82y [data dostępu: 20 czerwca 2020]
[2] źródło: http://galaxy.agh.edu.pl/~chwiej/mn/gen_trojkatny.pdf [data dostępu: 20 czerwca 2020]
[3] źródło: http://home.agh.edu.pl/~mariuszp/wfiis_stat/tablice_ps_wir.pdf (strona 4) [data dostępu: 20 czerwca 2020]