Sprawozdanie - Laboratorium 14
Generowanie ciągu liczb pseudolosowych o rozkładzie normalnym metodą eliminacji

Marta Dychała

14 czerwca 2021

1 Wstęp teoretyczny

Komputer jest maszyną deterministyczną - jeśli otrzyma ona na wejście zestaw danych i algorytm za pomocą którego ma je przetworzyć, wynik na wyjściu zawsze będzie identyczny. Oznacza to również, że w działaniu programów komputerowych nie może być żadnej losowości, a wynik działania programu jest w teorii zawsze możliwy do przewidzenia. Dlatego też pojawia się pojęcie liczb pseudolosowych, czyli liczb generowanych przez taki algorytm, który tylko stwarza pozory losowego działania - po prostu zakres otrzymywanych wyników potrafi być bardzo szeroki.

Do tworzenia liczb pseudolosowych używa się tzw. generatorów liczb pseudolosowych. Nazywane są one generatorami, ponieważ tworzą one liczbę pseudolosową na podstawie poprzedniej. Jedną z rodzin generatorów liczb pseudolosowych są generatory liniowe w ogólności określone wzorem:

$$X_{n+1} = (a_1 X_n + a_2 X_{n-1} + ... + a_k X_{n-k+1} + c) \ mod \ m \tag{1}$$

gdzie a_1, a_2, ..., a_k, c, m to pewne liczby, zwane parametrami generatora. Aby generator mógł zadziałać, wybiera się pewną liczbę bądź liczby początkowe, które zwane są ziarnem:

$$X_0, X_{-1}, ..., X_{-k}$$

W zależności od wartości parametru c generator liniowy można nazwać multiplikatywnym, jeżeli $c = 0$ bądź generatorem mieszanym w pozostałych przypadkach.

Generator mieszany określony jest wzorem:

$$X_{n+1} = (aX_n + c) \ mod \ m, \tag{2}$$

gdzie a, c, m to parametry generatora. Aby powyższy wzór był poprawny, należy określić wartość parametru początkowego X_0 zwanego ziarnem. Za pomocą generatora mieszanego można wygenerować liczby pseudolosowe, które podlegają rozkładowi jednorodnemu.

Rozkład jednorodny (jednostajny) to taki rozkład w przedziale [a,b], gdzie prawdopodobieństwo wylosowania każdego z punktów przedziału [a,b] jest jednakowe. Wielkości charakterystyczne tego rozkładu to:

- wartość średnia: $\mu = \frac{a+b}{2}$,
- odchylenie standardowe: $\sigma = \sqrt{\frac{b-a}{12}}$.

Innym rodzajem rozkładu danych jest rozkład normalny (zwany też rozkładem Gaussa). Jest to rodzaj rozkładu danych najczęściej spotykany w przyrodzie - w przedziale [a,b] najbardziej prawdopodobne staje się wylosowanie wartości średniej. Cechą charakterystyczną rozkładu normalnego jest jego funkcja gęstości

prawdopodobieństwa, która ma kształt krzywej dzwonowej (nazywanej też krzywą Gaussa). Określona jest ona wzorem:

$$f(x) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_0)^2}{2\sigma_0^2}\right). \tag{3}$$

Duży wpływ na wylosowanie danej wartości w rozkładzie normalnym ma odchylenie standardowe, bowiem prawdopodobieństwo wylosowania wartości znajdującej się w przedziale $[\mu-\sigma, \mu+\sigma]$ wynosi około 68.3%, natomiast dla przedziału $[\mu-3\sigma, \mu+3\sigma]$ prawdopodobieństwo to to aż 99.7% - jest to tak zwana reguła trzech sigm.

Ponieważ generator mieszany tworzy liczby pseudolosowe o rozkładzie jednorodnym to może być on wykorzystany do wygenerowania zmiennych podlegających rozkładowi normalnemu. W tym celu możemy posłużyć się metodą eliminacji.

Metoda eliminacji pozwala wygenerować ciąg liczb pseudolosowych o zadanej gęstości prawdopodobieństwa (w naszym przypadku to będzie wzór (3)) w przedziale $[x_{min}, x_{max}]$. Na początku losowane są liczby rzeczywiste $u_1 \in [x_{min}, x_{max}]$ oraz $u_2 \in [0, d]$ o rozkładzie jednorodnym, gdzie d to dowolna stała ograniczająca funkcję f od góry. Jeżeli $u_2 \leqslant f(u_1)$, to u_1 jest akceptowane jako x_i. W przeciwnym razie generator generuje kolejne wartości u_1, u_2 w danej iteracji dopóki dopóty warunek $u_2 \leqslant f(u_1)$ nie będzie spełniony. Całą procedurę powtarza się N razy, gdzie N to wielkość próby, a za razem ilość liczb pseudolosowych jaką trzeba znaleźć.

2 Zadanie do wykonania

2.1 Opis problemu

Celem zajęć laboratoryjnych było utworzenie generatora mieszanego liczb pseudolosowych podlegającym różnym rozkładom danym. Na początku należało zaimplementować generator mieszany

$$X_{n+1} = (aX_n + c)\ mod\ m \tag{4}$$

podlegający rozkładowi jednorodnemu. Jako parametry generatora przyjęto:

a) $a = 123$, $c = 1$, $m = 2^{15}$

b) $a = 69069$, $c = 1$, $m = 2^{32}$

W każdym z przypadków generowane liczby pseudolosowe miały się znajdować w przedziale $x_i \in [0, 1]$, innymi słowy miał być to rozkład $U(0, 1)$. Dlatego też dla wylosowanych liczb z generatora (4) należało zastosować normę:

$$x_i = \frac{X_i}{m + 1.0}. \tag{5}$$

Każdy z dwóch generatorów startował z tego samego ziarna - $X_0 = 10$. W celu porównania ich skuteczności, dla każdego z nich należało obliczyć wartość średnią:

$$\mu = \frac{1}{N}\sum_{i=0}^{N-1} x_i \tag{6}$$

oraz odchylenie standardowe:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(x_i - \mu)^2}, \tag{7}$$

gdzie N to wielkość próby, w naszym przypadku $N = 10^4$. Dokładna wartość średnia dla rozkładu U(0,1) to $\mu = 0.5$, natomiast odchylenie standardowe to $\sigma = \sqrt{\frac{1}{12}}$.

Dodatkowo dla każdego z generatorów należało wykonać histogram rozkładu gęstości prawdopodobieństwa dla $k = 12$ podprzedziałów. Proces tworzenia histogramu polegał na utworzeniu tablicy n o k

elementach. Element n_j informował ile razy trafiono w $j-ty$ podprzedział, gdzie $j = 0, 1, ..., k-1$. Każdy z elementów tej tablicy oznaczał indeks podprzedziału $[x_{j,min}, x_{j,max}]$ przedziału $[x_{min}, x_{max}]$, do którego należeć będzie dana zmienna. Ponieważ podprzedziały były równej długości ich szerokość określał wzór:

$$\Delta = \frac{x_{max} - x_{min}}{k},$$

gdzie w tym przypadku $x_{min} = 0$, $x_{max} = 1$. Po wyznaczeniu szerokości podprzedziałów wystarczyło znaleźć indeks podprzedziału, do którego należy element:

$$j = \frac{x_i - x_{min}}{\Delta}$$

Po dopasowaniu wszystkich liczb do odpowiednich podprzedziałów, należało znormalizować wartości w tablicy n dzieląc wszystkie jej elementy przez N, w celu uzyskania funkcji gęstości prawdopodobieństwa.

Kolejnym zadaniem było wygenerowanie $N = 10^4$ pseudolosowych liczb podlegających rozkładowi normalnemu. W tym celu należało wykorzystać poprzednio opisany generator mieszany o parametrach $a = 69069$, $c = 1$, $m = 2^{32}$. Utworzony miał zostać rozkład normalny o wartości średniej $\mu_0 = 0.2$, odchyleniu standardowym $\sigma_0 = 0.5$ (rozkład $N(\mu = 0.2, \sigma = 0.5)$) i zakresie danych $[x_{min}, x_{max}] = [\mu_0 - 3\sigma_0, \mu_0 + 3\sigma_0]$. Do utworzenia ciągu $N = 10^4$ liczb wykorzystano metodę eliminacji dla generatora mieszanego. Przyjęto, że $d = 1$. Otrzymany rozkład wykorzystano do utworzenia histogramu w analogiczny sposób jak to miało miejsce w przypadku rozkładu jednorodnego.

Po utworzeniu ciągu $N = 10^4$ liczb pseudolosowych podlegającym rozkładowi normalnemu $N(\mu = 0.2, \sigma = 0.5)$, obliczono wartość średnią rozkładu μ wykorzystując wzór (6), odchylenie standardowe σ zgodnie ze wzorem (7), oraz wariancję σ^2, która jest kwadratem odchylenia standardowego.

Dla otrzymanego rozkładu normalnego wyznaczono następnie statystykę testową:

$$\chi^2 = \sum_{j=0}^{k-1} \frac{(n_j - N \cdot p_j)^2}{N \cdot p_j}, \tag{8}$$

gdzie p_j to teoretyczne prawdopodobieństwo wylosowania liczby należącej do $j-tego$ podprzedziału obliczone ze wzoru:

$$p_j = P(x_{j,min} < x \leqslant x_{j,max}) = F(x_{j,max}) - F(x_{j,min}), \tag{9}$$

Funkcja $F(x)$ to dystrybuanta rozkładu:

$$F(x) = \frac{1 + erf(\frac{x-\mu_0}{\sqrt{2}\cdot\sigma_0})}{2} \tag{10}$$

z kolei $erf(x)$ jest funkcją błędu, której wartość wyznaczono za pomocą funkcji erf() z standardowej biblioteki matematycznej języka C/C++.

Po wyznaczeniu statystyki testowej kolejnym krokiem było sprawdzenie, czy hipoteza H_0 mówiąca, że „wygenerowany rozkład jest rozkładem normalnym" jest prawdziwa na poziomie istotności $\alpha = 0.05$. W tym celu należało sprawdzić nierówność $\chi^2 < \varepsilon$, dla $\varepsilon = 16.91$ (wartość odczytana z tablic statystycznych[1]) Ostatnim zadaniem było wyznaczenie poziomu ufności $P(\chi^2|\nu) = 1 - \tilde{\alpha}$, na podstawie którego można było wyznaczyć poziom istotności statystyki $\tilde{\alpha}$. Poziom ufności został wyznaczony za pomocą funkcji $gammp\left(\frac{\nu}{2}, \frac{\chi^2}{2}\right)$ pochodzącej z biblioteki numerycznej Numerical Recipes, która oblicza niekompletną funkcję Gamma-Eulera.

Parametr ν, który pojawił się w funkcji $gammp()$ to tak zwany stopień swobody. Oblicza się go ze wzoru:

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

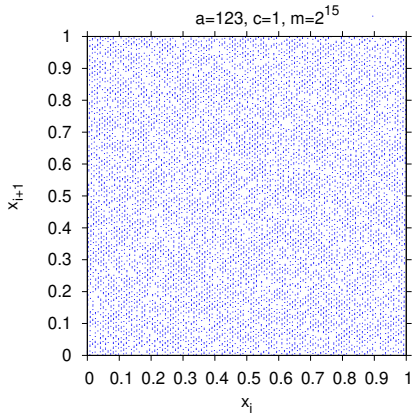
[1] http://home.agh.edu.pl/ mariuszp/wfiis_stat/tablice_ps_wir.pdf [dostęp 12.06.2021 r.]

gdzie k to liczba podprzedziałów, r to liczba parametrów rozkładu (w tym przypadku $r = 2$, ponieważ mamy dwa parametry: μ_0 oraz σ_0).

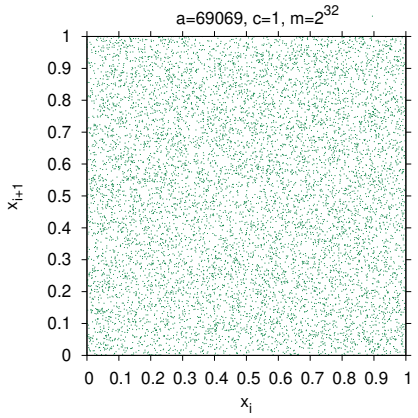
Do napisania programu używanego na zajęciach wykorzystano bibliotekę numeryczną Numerical Recipes pochodzącą z języka C, a konkretniej pliki *nrutil.h*, *nrutil.c*, *gammp.c*, *gcf.c*, *gammln.c* oraz *gser.c*. Potrzebne wykresy wygenerowano w programie Gnuplot.

2.2 Wyniki

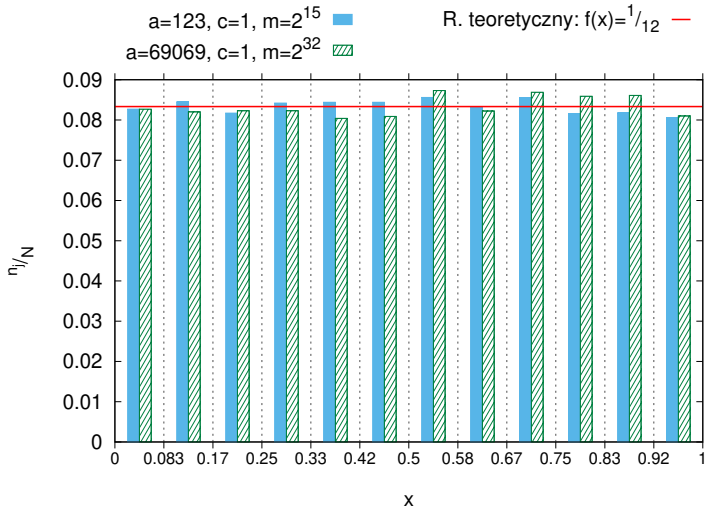
Poniżej znajduje się wykres zależności wylosowanej liczby od poprzedniej dla dwóch przypadków generatora mieszanego:



(a) $a = 123$, $c = 1$, $m = 2^{15}$



(b) $a = 69069$, $c = 1$, $m = 2^{32}$



(c) Histogram dla rozkładów pochodzących z obu przypadków generatora mieszanego

Rysunek 1: Zależności $x_{i+1}(x_i)$ dla generatorów mieszanych, gdzie ziarno to $X_0 = 10$, a wielkość próby to $N = 10^4$.

Na podstawie powyższych rysunków oraz histogramu można stwierdzić, iż generatory charakteryzują się podobną jakością, aczkolwiek na wykresie (a) można dostrzec, że wartości układają się w niektórych miejscach w linie (tzw. hiperpłaszczyzny), co nie jest cechą pożądaną generatora. Cechą wspólną dla obydwu generatorów jest jednak to, że punkty na wykresach są rozłożone w miarę równomiernie, co utrudnia znalezienie jakiejkolwiek zależności między generowanymi liczbami. Jakość generatorów można też sprawdzić,

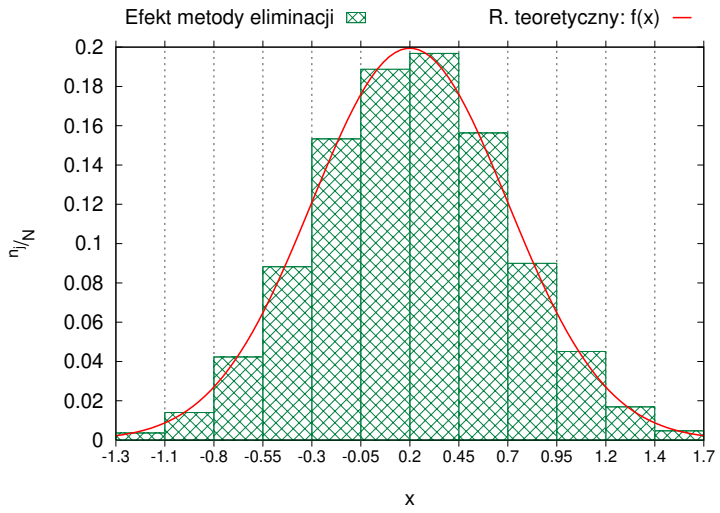
poprzez obliczenie wartości średniej i odchylenia standardowego rozkładu:

Tabela 1: Średnia μ i odchylenie standardowe σ obliczone dla rozkładów jednorodnych uzyskanych przy użyciu generatora mieszanego dla $N = 10^4$ liczb

a	c	m	Średnia μ	Odchylenie standardowe σ
123	1	2^{15}	0.498266	0.287120
69069	1	2^{32}	0.503806	0.288070

W obydwu przypadkach otrzymane wartości μ oraz σ są zbliżone do wartości teoretycznych wynoszących $\mu = 0.5$ oraz $\sigma = \sqrt{\frac{1}{12}} \approx 0.288675$, co świadczy o całkiem dobrej jakości obydwu generatorów.

Kolejne wyniki dotyczą rozkładu normalnego. Poniżej znajduje się histogram otrzymany z liczb wygenerowanych za pomocą metody eliminacji:



Rysunek 2: Histogram dla wygenerowanego rozkładu normalnego $N(\mu_0 = 0.2, \sigma_0 = 0.5)$ wraz z naniesionym rozkładem teoretycznym $f(x)$

Wartość średnia μ, wariancja σ^2 i odchylenie standardowe σ obliczone na podstawie wygenerowanych danych to kolejno: $\mu = 0.210215$, $\sigma^2 = 0.236315$ oraz $\sigma = 0.486123$. Wyniki te są zbliżone do wartości teoretycznych dla rozkładu $N(\mu = 0.2, \sigma = 0.5)$.

Dla omawianego rozkładu normalnego wyznaczono także wartości teoretyczne prawdopodobieństwa znalezienia się wylosowanej liczby w j-tym podprzedziale (p_j), którą porównano z wartością wyznaczoną numerycznie (n_j/N) co jest przedstawione w tabeli na kolejnej stronie:

Tabela 2: Teoretyczne prawdopodobieństwo wylosowania liczby z j-tego podprzedziału p_j dla rozkładu normalnego $N(\mu_0 = 0.2, \sigma_0 = 0.5)$ porównane z wartością $\frac{n_j}{N}$ obliczoną numerycznie

j	p_j	$\frac{n_j}{N}$
0	0.00485977	0.0036
1	0.0165405	0.014
2	0.0440571	0.0423
3	0.0918481	0.0882
4	0.149882	0.1533
5	0.191462	0.1888
6	0.191462	0.1968
7	0.149882	0.1563
8	0.0918481	0.09
9	0.0440571	0.0451
10	0.0165405	0.0169
11	0.00485977	0.0047

Powyższa tabela jest kolejnym dowodem na to, że wygenerowane liczby podlegają rozkładowi zbliżonemu do normalnego. Jednak najbardziej istotnym dowodem wydaje się być wynik hipotezy H_0, bowiem poniższa nierówność została spełniona:

$$\chi^2 = 15.4522 < \varepsilon = 16.91$$

co oznacza, że hipoteza H_0 została zaakceptowana na poziomie istotności $\alpha = 0.05$. Dla zaprezentowanych wyników poziom ufności statystyki χ^2 wyniósł $P(\chi^2|\nu) = 0.920758$, natomiast poziom istotności: $\tilde{\alpha} = 0.0792415$. Oznacza to, że w około 92% wygenerowane dane odpowiadają rozkładowi normalnemu, co jest wartością zadowalającą.

3 Wnioski

Generator mieszany liczb pseudolosowych jest przydatnym, szybkim i prostym w implementacji narzędziem pozwalającym na generowanie liczb sprawiających wrażenie losowych. Przy odpowiednim dobraniu parametrów zadaniem trudnym, a wręcz niewykonalnym staje się znalezienie zależności między generowanymi liczbami spoglądając jedynie na wartości liczb lub wykres zależności $x_{i+1}(x_i)$ bez znajomości wzoru. Generatory rozważane w czasie zajęć okazały się być dobrej jakości o czym świadczą zbliżone do teoretycznych wartości średnie oraz odchylenia standardowe obliczone dla tych rozkładów. Lepszej jakości okazał się być jednak generator o parametrach $a = 69069$, $c = 1$, $m = 2^{32}$, dlatego też został on wykorzystany w dalszej części zadania.

Na podstawie otrzymanych wyników dla rozkładu normalnego można dojść do wniosku, że metoda eliminacji jest wyjątkowo skuteczną metodą pozwalającą przekształcić rozkład jednorodny w normalny. Dane wygenerowane w ten sposób zachowywały się w zbliżony sposób do rozkładu $N(\mu_0 = 0.2, \sigma_0 = 0.5)$, co potwierdzały wszelkie wyniki dotyczące rozkładu normalnego. Co prawda otrzymane wyniki nieco różnią się od teoretycznych, lecz ten fakt spowodowany jest jakością generatora oraz dobraną wartością początkową dla ziarna X_0. Metoda eliminacji bowiem rozważa bieżącą wartość pseudolosową, która zależy od postaci funkcji generatora i dla określonych parametrów generatora metoda ta zachowa się zawsze tak samo. Ponieważ generator liczb pseudolosowych jest deterministyczny to zawsze jest możliwe przewidzenie wyniku metody eliminacji.