Sprawozdanie - laboratorium 14

Generowanie ciągu liczb pseudolosowych o rozkładzie normalnym metodą eliminacji

Karolina Kotłowska, 9 czerwiec 2021

1 Wstęp teoretyczny

1.1 Średnia arytmetyczna, odchylenie standardowe, rozkład normalny

Średnia arytmetyczna - suma liczb podzielona przez ich liczbę.

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

Odchylenie standardowe - klasyczna miara zmienności, obok średniej arytmetycznej najczęściej stosowane pojęcie statystyczne.

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

Rozkład normalny - rozkład prawdopodobieństwa, którego wykres funkcji prawdopodobieństwa jest krzywą w kształcie dzwonu. Gęstość prawdopodobieństwa okresla się wzorem:

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

Funkcja okreslająca dystrybuante określa się jako:

$$F(x) = \frac{1 + erf(\frac{x-\mu_0}{\sqrt{2}\sigma_0})}{2} \tag{4}$$

erf to funkcja błędu.

1.2 Generator liniowy

Generatory liniowe tworzą ciąg liczb według schematu:

$$X_{n+1} = (a_1 X_n + a_2 Xn - 1 + ... + a_k X_{n-k+1} + c) (\mod m) \tag{5}$$

gdzie $a_1, a_2, ..., a_k, c, m$ - parametry generatora (ustalone liczby)

Operację:

$$r = (a \mod n) \tag{6}$$

nazywamy dzieleniem modulo a jej wynikiem jest reszta z dzielenia liczb całkowitych a i n.

Lub inaczej: r jest kongruentne do a modulo n jeśli n jest dzielnikiem a-r.

$$a \equiv r \mod n \tag{7}$$

$$r = a - \left\lfloor \frac{a}{n} \right\rfloor n \tag{8}$$

Generatory wykorzystujące operację dzielenia modulo to generatory kongruentne lub kongruencyjne.

1.3 Metoda eliminacji

Metoda eliminacji pozwala wygenerować ciąg liczb pseudolosowych x_i o zadanej gęstości prawdopodobieństwa f(x) w przedziale [xmin,xmax] (u nas: xmin=μ_0 - 3σ_0 oraz xmax=μ_0+3σ_0) w następujących krokach:

1. Losujemy liczbę rzeczywistą $u1 \in [xmin, xmax]$ o rozkładzie jednorodnym;
2. Losujemy liczbę rzeczywistą $u2 \in [0, d]$ o rozkładzie jednorodnym;
3. Jeżeli u2$\leqslant f(u1), to$ x$_i = u_i$ (akceptacja u1). W przeciwnym razie odrzucamy u1 i u2.

1.4 Testowanie generatorów liczb pseudolosowych

Ponieważ wszystkie generatory o dowolnym rozkładzie bazują na wykorzystaniu ciągów liczb losowych o rozkładzie równomiernym więc istotne jest badanie tylko generatorów liczb o takim właśnie rozkładzie.
Testowanie generatora jest procesem złożonym:
1) dla ustalonej liczby n, generujemy n kolejnych liczb startując od losowo wybranej liczby początkowej
2) obliczamy wartość statystyki testowej (T)
3) obliczamy F(T) czyli dystrybuantę statystyki T, gdy weryfikowana hipoteza jest prawdziwa kroki 1-3 powtarzamy N-krotnie obliczając statystyki: $T_1, T_2, ..., T_N$.
Jeśli weryfikowana hipoteza jest prawdziwa to: $F(T_1), F(T_2), ..., F(T_N)$, jest ciągiem zmiennych niezależnych o rozkładzie równomiernym. Testowanie generatora kończy się sprawdzeniem tej hipotezy.

1.5 Rozkład χ^2

Rozkład chi kwadrat – rozkład zmiennej losowej, która jest sumą k kwadratów niezależnych zmiennych losowych o standardowym rozkładzie normalnym. Liczbę naturalną k nazywa się liczbą stopni swobody rozkładu zmiennej losowej.

Jest najczęściej stosowanym testem. Badamy w nim hipotezę że generowana zmienna losowa X ma rozkład prawdopodobieństwa o dystrybuancie F.

Jeżeli: $F(a) = 0, F(b) = 1$, to możemy dokonać nastepującego podziału wartości zmiennej X:

$$a < a_1 < a_2 < ... < a_k = b \tag{9}$$

$$p_i = Pa_{i-1} < X \leqslant a_i, i = 1, 2, ... \tag{10}$$

Generujemy ciąg n liczb $X_1, X_2, ..., X_n$. Sprawdzamy ile z nich spełnia warunek $a_{i-1} < X \leqslant a_i$. Ich liczbę oznaczamy n_i. Statystyką testu jest:

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \tag{11}$$

Dla dużego n statystyka ma rozkład χ^2 o (k-1) stopniach swobody. Można tak dobrać szerokość przedziałów aby otrzymać zależność: $p_i = \frac{1}{k}$. Wówczas statystyka przyjmuje prostszą postać:

$$\chi^2_{k-1} = \frac{k}{n} \sum_{i=1}^{k} n_i^2 - n \tag{12}$$

2 Zadanie do wykonania

2.1 Opis problemu

Naszym zadaniem było znaleźć rozkład jednorodny dla danych wygenerowanych przy użyciu generqatora mieszanego:

$$x_{n+1} = (ax_n + c) \mod m \tag{13}$$

Wykanano zadanie dla dwóch wersji o wspólnych parametrach $x_0 = 10$, $n = 10^4$, a dla różnych parametrów
a) a=123, c=1, m=2^{15}
b) a=69069, c=1, m=2^{32}

Dla obu przypadków sporządziliśmy rysunek z warunku normalizacji rozkłądu U(0,1). Obliczyliśmy średnią arytmetyczną i odchylenie standardowe.

Kolejnym zadaniem było znalezienie rozkładu normalnego wykorzystując generator mieszany z pierwszego podpunktu. Dla danych $n = 10^4$, $\mu = 0.2$, $\sigma = 0.5$, wykonano wykres. Liczby pseudolosowe zawierały się w przedziale x $\in [\mu - 3\sigma, \mu + 3\sigma]$. Dla tego podpunktu obliczyliśmy średnią arytmetyczną, odchylenie standarowe i wariancję. Dla ilości przedziałów równej 12, narysowaliśmy histogram (rysunek 3).

Następnie przeprowadziliśmy testowanie generatora N(μ, σ) - test χ^2. Obliczyliśmy wszytskie potrzebne wartości: średnią, odchylenie standardowe oraz wyznaczyliśmy wartość statystyki testowej. Określiliśmy funkcję błędu:

$$F(x) = \frac{1 + erf(\frac{x-\mu_0}{\sqrt{2}*\sigma_0})}{2} \tag{14}$$

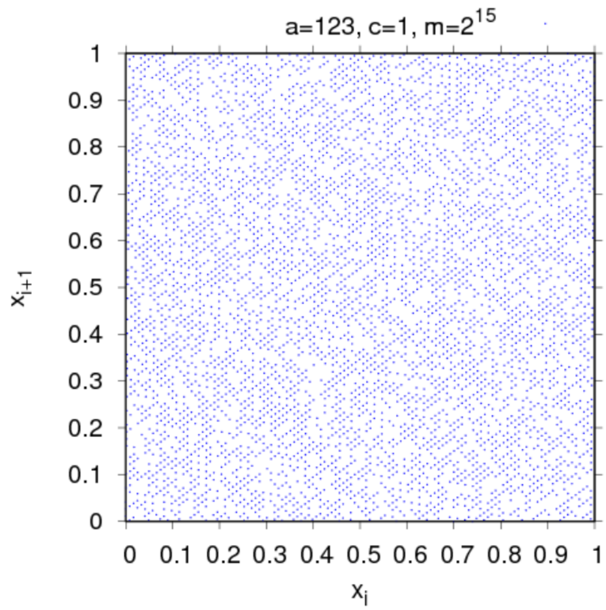
Postawiliśmy hipotezę H_0="Otrzymany rozkład jest rozkładem normalnym N($\mu_0 = 0.2$, $\sigma_0 = 0.5$)". Korzystając z tablic statystycznych potwierdziliśmy, że teza jest prawdziwa.

Ostatnim zadaniem było wyznaczenie opziomu istotności dla obliczonej statystyki $\bar{\alpha} = 1 - P(\chi^2|v)$. Do tego skorzystaliśmy z funkcji liczącej poziom ufności $P(\chi^2|v)$=gammp($\frac{v}{2}$, $\frac{\chi^2}{2}$).

Wykorzystano biblioteki: nrutil.h, nrutil.c, gammp.c, gcf.c, gammln.c, gser.c.

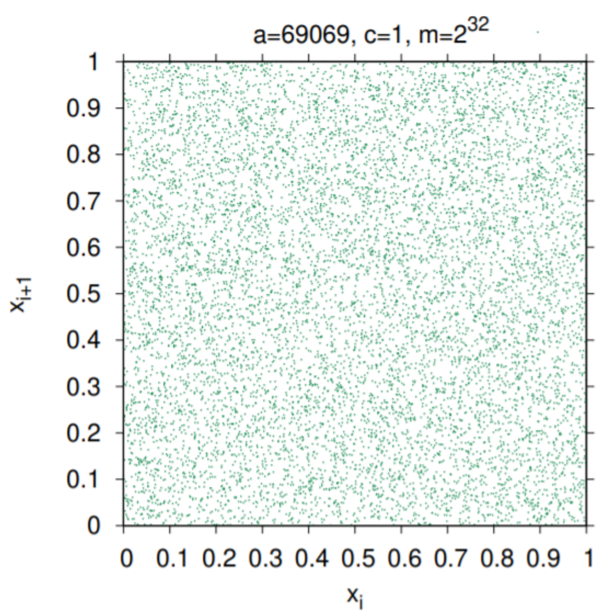
3 Wyniki

Dla części pierwszej otrzymano następujące wyniki:
a) dla a=123, m=2^{15} Średnia: 0.498266, Odchylenie standardowe: 0.28712
b) dla a=69069, m=2^{15} Średnia: 0.503806, Odchylenie standardowe: 0.28807



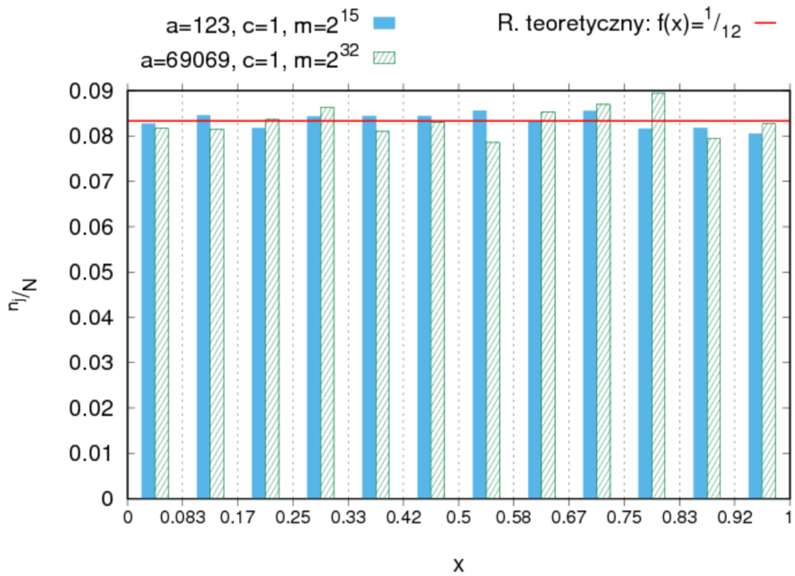
Rysunek 1: Zależność x_{i+1} (x_i) według generatora mieszanego o parametrach a = 123, c = 1, m = 2^{15}

Wykres powyżej nie prezentuje idealnych własności statystycznych, ponieważ można zauważyć, że punkty rozmieszczone na płaszczyźnie mają tendencję to układanie się skośne pasy. Wynika to z nieodpowiedniego doboru parametrów. Parametr a oraz m powinny być dużo większe.



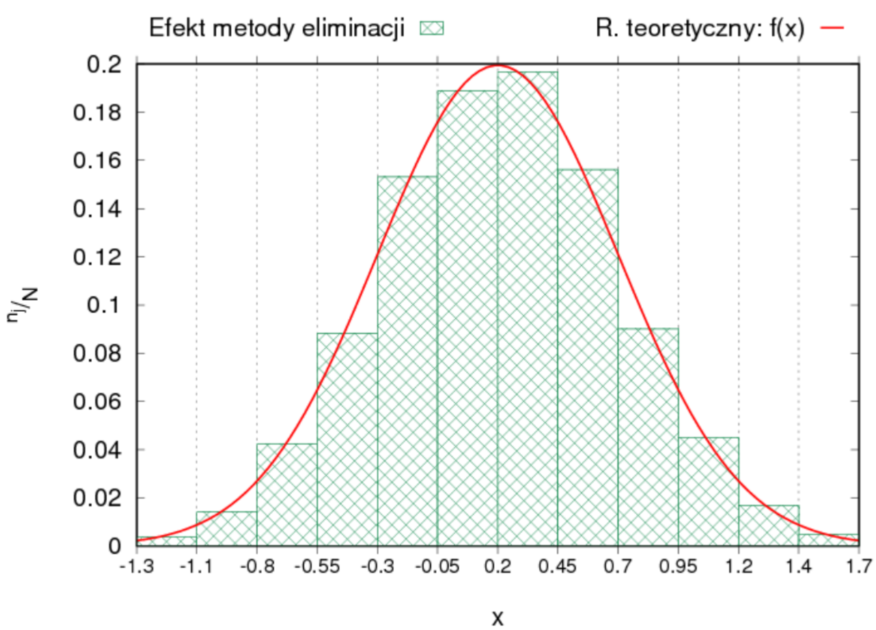
Rysunek 2: Zależność x_{i+1} (x_i) według generatora mieszanego o parametrach a = 69069, c = 1, m = 2^{32}

Wykres powyżej prezentuje dobre własności statystyczne. Punkty są losowo rozrzucone, nie widać tendencji punktów do grupowania się w określony sposób. Ten przypadek zawdzięcza to odpowiednio dobranym parametrom a oraz m.



Rysunek 3: Histogram dla rozkładów pochodzących z obu przypadków generatora mieszanego.

Na powyższym wykresie można zauważyć, że otrzymane wartości są bliskie wartości teoretycznej, zatem wyniki są zadowalające.



Rysunek 4: Histogram dla wygenerowanego rozkładu normalnego N($\mu_0 = 0.2$, $\sigma_0 = 0.5$) wraz z naniesionym rozkładem teoretycznym f(x)

Na powyższym rysunku dostaliśmy rozkład zgodny z oczekiwanym. Wykres teoretyczny przecina każdy słupek rozkładu, więc możemy stwierdzić, że wyniki są zadowalające.

```
0.00485977
0.0165405
0.0440571
0.0918481
0.149882
0.191462
0.191462
0.149882
0.0918481
0.0440571
0.0165405
0.00485977
```

Rysunek 5: Teoretyczne prawdopodobieństwo wylosowania liczby z j-tego podprzedziału dla rozkładu normalnego

Statystyka testowa χ^2, którą otrzymaliśmy wyniosła 14,982
Poziom ufności P($\chi^2|v$) wyniósł 0.908567
Poziom istotności $\bar{\alpha}$ 0.0914327

4 Wnioski

- Nie da się wygenerować liczb całkowicie losowych, można jednak wygenerować liczby pseudolosowe dzięki skorzystaniu z różnych parametrów i funkcji. Otrzymane liczby pseudolosowe, są niemal nierozróżnialne z liczbami losowymi.

- Korzystanie z liniowy generatora kongruentnego jest dobrym sposobem na wygenerowanie liczb pseudolosowych.

- Sposób działania generatorów zależy od jego postaci i doboru parametrów. Ma to kluczowe znaczenie jeśli chcemy wygenenerować losowe liczby.

- Obliczając wartość oczekiwaną i odchylenie standardowe w prosty sposób można sprawdzić poprawność wygenerowanych liczb.