ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1970

Seria: MATEMATYKA-FIZYKA z. 15 nr kol. 285

GERD RESZKA
Ośrodek Maszyn Matematycznych
Politechniki Śląskiej

GENEROWANIE LICZB PSEUDOLOSOWYCH NA MASZYNIE CYFROWEJ UMC-1

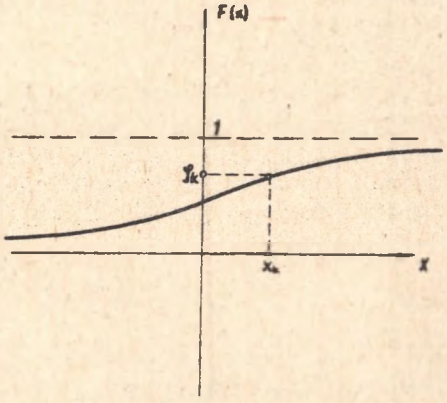
Streszczenie: W pracy omówione są wyniki badania generatora liczb pseudolosowych o rozkładzie równomiernym, opartego na wykorzystaniu zmodyfikowanego algorytmu von Neumanna.

Przy rozwiązywaniu licznych problemów przy pomocy EMC, potrzebne są często liczby losowe o określonych z góry rozkładach. Najbardziej reprezentatywnym z tych problemów jest modelowanie statystyczne zwane też metodami Monte Carlo. Na wejścia zamodelowanego przy pomocy maszyny matematycznej badanego układu podaje się wielkości przypadkowe o zadanych z góry parametrach rozkładu, a na wyjściach śledzi się pewne wielkości jak np. średnie arytmetyczne, częstość względną pewnych zdarzeń, które są estymatorami odpowiednio wartości oczekiwanej i prawdopodobieństwa wystąpienia obserwowanego zdarzenia i przy odpowiedniej ilości realizacji procesu modelowania różnica pomiędzy parametrem statystycznym wyjścia a jego estymatorem staje się dowolnie mała. Modelowanie statystyczne stosuje się zresztą nie tylko do badania procesów przypadkowych, ale także do rozwiązywania wielu problemów deterministycznych, czego najlepszym przykładem jest przybliżone obliczanie całek wielokrotnych. Powstaje teraz zagadnienie tworzenia przez maszynę cyfrową ciągów liczb przypadkowych o zadanych funkcjach rozkładu.

Przypuśćmy, że potrzebny nam jest ciąg liczb losowych o funkcji rozkładu $f(x)$ i dystrybuancie

$$F(x) = P(x \leqslant x_i) = \int_{-\infty}^{x_i} f(x)dx . \qquad (1)$$

Z rys. 1 widać, że $P(x_k \leqslant x) = F(x_k) = \xi_k$.



Rys. 1

Zakładając, że zbiór wartości ξ jest zbiorem o równomiernym rozkładzie na odcinku $[0,1]$, wobec czego wybierając losowo liczby ξ_k, możemy dla każdego ξ_k wyznaczyć odpowiednie x_k i wyznaczony w ten sposób ciąg x_k będzie ciągiem liczb losowych o dystrybuancie $F(x)$, jeżeli tylko ξ_k będzie ciągiem liczb losowych o rozkładzie równomiernym w przedziale $[0,1]$.
Analitycznie powyższe rozważania można napisać w sposób następujący:

$$\int_{-\infty}^{x_k} f(t)dt = \xi_k . \qquad (2)$$

Rozwiązując to równanie można otrzymać ciąg x_k o zadanej funkcji rozkładu f(x), dysponując ciągiem ξ_k liczb losowych o rozkładzie równomiernym w przedziale [0,1].

Przykład:

Załóżmy, że potrzebny nam jest ciąg liczb losowych o rozkładzie wykładniczym, zdefiniowanym następującym wzorem

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{dla } x \geqslant 0 \\ 0 & \text{dla } x < 0 \end{cases} \tag{3}$$

Wykonując działania określone wzorem (2) otrzymujemy

$$1 - e^{-\lambda x_k} = \xi_k \tag{4}$$

$$x_k = -\frac{1}{\lambda} \ln (1-\xi_k) \tag{5}$$

Zakładając, że ciąg $\{1 - \xi_k\}$ ma ten sam rozkład co ciąg $\{\xi_k\}$, możemy napisać wzór (5) w postaci

$$x_k = -\frac{1}{\lambda} \ln \xi_k . \tag{6}$$

Aby otrzymać ciąg liczb losowych o rozkładzie wykładniczym danym wzorem (3), wystarczy kolejne wartości liczb losowych o rozkładzie równomiernym ξ_k poddawać przekształceniu według wzoru (6).

Jak z powyższego wynika, do tworzenia ciągów liczb losowych o różnych rozkładach wystarczy umieć generować przy pomocy maszyny cyfrowej liczby losowe o rozkładzie równomiernym w przedziale [0,1].

Generatory liczb losowych w maszynach cyfrowych mogą być dwojakiego rodzaju:

1) generatory fizyczne - wykorzystujące efekt śrutowy wzmacniaczy elektronicznych, rozpad pierwiastków radioaktywnych itp. Generatory te są wykonane jako specjalne przystawki do maszyn matematycznych i nie będą tu omawiane;
2) generatory wykorzystujące do tworzenia liczb losowych algorytmy zdeterminowane, wobec czego uzyskany ciąg jest tylko z pewnym przybliżeniem losowy i dlatego nazywa się je generatorami liczb pseudolosowych.

Istnieje wiele algorytmów tworzenia ciągu liczb pseudolosowych o rozkładzie równomiernym. Jednym z nich jest zmodyfikowany algorytm von Neumanna określony w ten sposób, że dwie s-bitowe liczby dwójkowe wymnaża się przez siebie otrzymując w ten sposób liczbę podwójnie długą 2s-bitową. Z tej liczby wycinamy dowolnych s-bitów i w ten sposób otrzymujemy nową liczbę, którą zastępujemy jedną z liczb pary wyjściowej, po czym wszystkie czynności powtarzamy od nowa.

Inna wersja tego algorytmu przewiduje zamiast mnożenia dwu liczb, podnoszenie jednej do kwadratu, co jednak przy użyciu maszyny cyfrowej UMC-1 nie prowadzi do dobrych wyników ze względu na to, że maszyna ta pracuje w systemie minus minus dwójkowym, pozwalającym zapisywać liczby zarówno dodatnie jak i ujemne, a ponieważ kwadraty liczb są zawsze dodatnie prowadzi to do naruszenia losowości rozkładu zer i jedynek na poszczególnych bitach. Maszyna UMC-1 pracuje w systemie minus minus dwójkowym wobec czego i przedział liczb w niej występujących jest określony nierównością

$$-\frac{1}{3} < x < \frac{2}{3}. \qquad (7)$$

Aby więc otrzymać rozkład równomierny w przedziale [0,1] do każdej otrzymanej liczby losowej dodawana jest 1/3, algorytm generuje bowiem ciąg liczb o rozkładzie równomiernym w przedziale (-1/3, 2/3).

Celem sprawdzenia, czy liczby pseudolosowe generowane przez maszynę mają rozkład rzeczywiście równomierny w przedziale [0,1] przeprowadzono szereg prób i testów statystycznych, których wyniki przytoczymy poniżej.

Celem kontroli losowości rozkładu zastosujemy system testów Kendalla, których obejmuje:

a) test kontroli częstości,
b) test kontroli par,
c) test kontroli przedziałów,
d) test kontroli kombinacji.

Test kontroli częstości (frequency test)

Przedział [0,1] dzieli się na pewną ilość (zwykle 10-30) równych podprzedziałów i bada się ilości liczb losowych, które trafią do określonego podprzedziału. Dla dostatecznie dużej ilości liczb pseudolosowych liczebność w każdym podprzedziale powinna dążyć do N/l, gdzie l jest ilością podprzedziałów, a N liczebnością próby. Dla l=10 oraz N = 1000 otrzymano następujące rezultaty:

Tablica 1

Nr wart. pocz.	0-0,1	0,1-0,2	0,2-0,3	0,3-0,4	0,4-0,5	0,5-0,6	0,6-0,7	0,7-0,8	0,8-0,9	0,9-1,0
1	963	993	1027	991	976	1010	1024	1012	1039	965
2	1059	1020	984	985	950	1010	980	1038	969	1025
3	986	979	960	1019	1026	1008	1015	976	1011	1020
4	998	1004	999	1002	948	978	1000	1020	1052	999
5	968	1031	993	1042	979	990	1019	939	1028	1011
6	1041	966	992	983	996	1060	1034	963	995	970
7	986	980	1045	998	1005	999	976	1021	1015	975
8	1047	987	1010	976	1035	978	984	943	1033	1007
9	1000	1003	999	988	1031	970	988	1007	1013	1001
10	1018	978	949	997	1060	1004	969	993	1011	1021

Test kontroli par (serial test)

Sprawdzamy częstość występowania zer i jedynek na poszczególnych pozycjach liczby pseudolosowej - (oczywiście w postaci dwójkowej). Częstość względna powinna być w przybliżeniu równa 0,5, ponieważ prawdopodobieństwo wystąpienia zera i jedynki jest identyczne.

Zbadano częstość występowania zer w różnych miejscach (bitach) liczby dwójkowej i np. dla próby o liczebności N = 1000 oraz trzech przypadkowych wartości początkowych uzyskano dla ostatniego bitu następujące częstości względne występowania zera

$$p_1 = 0{,}506, \qquad p_2 = 0{,}496, \qquad p_3 = 0{,}515,$$

a więc niewiele odbiegające od teoretycznej wartości 0,5.

Test kontroli kombinacji (poker test)

Jest to test polegający na zbadaniu rozkładu różnych kombinacji zer i jedynek dwójkowej reprezentacji liczby losowej dla dużej objętości losowanego zbioru.

Jeżeli weźmiemy pod uwagę s pozycji liczby w układzie dwójkowym, to przy założeniu równego prawdopodobieństwa wystąpienia każdej kombinacji, prawdopodobieństwo, że w danej liczbie będzie dokładnie k-zer i s-k jedynek (lub na odwrót) jest równe

$$p_k = \frac{\binom{k}{s}}{2^s} . \tag{6}$$

Sprawdzono kombinacje dla s = 10 w zbiorze 15000 liczb pseudolosowych generowanych na maszynie UMC-1 i otrzymano następujące rezultaty:

Tablica 2

	Wyn. prób	Wyniki teoretyczne
10 zer	18	14,65
9 zer 1 jedynka	167	146,48
8 zer 2 jedynki	682	659,18
7 zer 3 jedynki	1666	1751,81
6 zer 4 jedynki	3139	3076,17 $x^2 = 6,131$
5 zer 5 jedynek	3755	3691,41 $p = 0,80$
4 zera 6 jedynek	3078	3076,17
3 zera 7 jedynek	1681	1751,81
2 zera 8 jedynek	663	659,18
1 zero 9 jedynek	139	146,48
10 jedynek	12	14,65

Zgodność rozkładu teoretycznego z empirycznym sprawdzono testem χ^2.

Test kontroli przedziałów (gap test) zastępowany bywa także testem długości serii [2].

Test serii przewiduje podział liczb pseudolosowych na dwie klasy - np. liczb większych od 0,5 i mniejszych od 0,5, który to podział został przyjęty w naszych rozważaniach.

Serią nazywa się część ciągu liczb pseudolosowych zawierającą liczby tej samej klasy występujących bezpośrednio po sobie, ilość tych liczb nazywa się długością serii.

Dla badanego ciągu, obejmującego N liczb losowych, oblicza się następujące wielkości:

s_{i1} - liczba serii pierwszej klasy o długości i

s_{i2} - " " drugiej " " "

$S_{1k} = \sum_{i=k}^{n_1} s_{i1}$ - wszystkie serie klasy pierwszej o długości równej lub większej od k,

$S_{2k} = \sum_{i=k}^{n_2} s_{i2}$ - wszystkie serie klasy drugiej o długości równej lub większej od k,

$S_1 = \sum_{k=1}^{n_1} S_{1k}$ - ogólna ilość serii pierwszej klasy,

$S_2 = \sum_{k=1}^{n_2} S_{2k}$ - ogólna ilość serii drugiej klasy,

$S = S_1 + S_2$ ogólna ilość serii.

Dla dużej liczby prób należy teraz określić empirycznie wielkości S_1, S_2 i S i porównać je z wielkościami teoretycznymi, które np. dla poziomu ufności $\alpha = 0{,}05$ oraz $N \geqslant 20$ wynoszą: dolna granica ogólnej ilości serii

$$z(S) = 0{,}5\ (N+1 - 1{,}65\ \sqrt{N-1}), \tag{9}$$

dolna granica ilości serii klasy pierwszej lub drugiej

$$z(S_1) = z(S_2) = 0{,}25\ (N-1{,}65\ \sqrt{N-1}), \tag{10}$$

górna wartość długości serii

$$z(m) = \frac{\lg\left(-\frac{N}{\ln p}\right)}{\lg 2} - 1. \tag{11}$$

Oznacza to, że wśród N liczb możemy z prawdopodobieństwem P spodziewać się serii dłuższej niż m.
Zastosowano ten test do populacji liczb pseudolosowych o liczebności $N = 634$ i otrzymano następujące wyniki:

$S_1 = 155$, $S_2 = 154$, $S = 309$ oraz 6.

Wyniki teoretyczne obliczone na podstawie wzorów (9),(10) i (11) są następujące:

$$z(S) = 297, \quad z(S_1) = z(S_2) = 148 \text{ oraz } z(m) = 13.$$

Dolna granica ilości serii obydwu klas wynosi 148. Otrzymano 155 serii klasy pierwszej i 154 serie klasy drugiej.
Ogólna ilość serii powinna być większa niż 297, otrzymano serii ogółem 309.
Z testu wynika, że z prawdopodobieństwem równym 0,05 można oczekiwać serii o długości 13, najdłuższe otrzymane serie miały długość 6.
Po sprawdzeniu losowości generowanych liczb, należy zweryfikować, czy ich rozkład rzeczywiście jest równomierny w przedziale $[0,1]$.
Można do tego użyć: testu χ^2, kryterium Kołmogorowa lub kryterium ω^2.
Test χ^2 oparty jest na statystyce

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - Np_i)^2}{N_{pi}}, \tag{12}$$

gdzie n_i jest ilością elementów i-tego przedziału, a Np_i wartością oczekiwaną tych elementów przy rozkładzie hipotetycznym Stawiamy hipotezę, że rozkład generowanych liczb pseudolosowych jest równomierny, wobec czego, jeżeli podzielimy przedział $[0,1]$ na l jednakowych podprzedziałów, to prawdopodobieństwo trafienia w i-ty podprzedział wynosi $p_i = 1/l$ dla $i = 1,2,\ldots l$, a wzór (12) przyjmuje postać

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - N\frac{1}{l})^2}{N\frac{1}{l}} = \frac{1}{lN} \sum_{i=1}^{l} (n_i l - N)^2. \tag{13}$$

W przypadku rozkładu równomiernego suma ta ma rozkład w przybliżeniu taki jak χ^2 z l-1 stopniami swobody.
Granicę górną dla danego l i poziomu ufności p odczytujemy z tablic i jeżeli obliczone χ^2 przewyższa tę granicę, to hipotezę o rozkładzie równomiernym należy odrzucić.
Jeżeli jednak χ^2 przyjmuje bardzo małe wartości, to może to świadczyć o za "dobrej" zgodności, a co za tym idzie o naruszeniu losowości generowanych liczb, dlatego też ustalamy dolną granicę dla wartości χ^2 np. w następujący sposób:

$$P\left\{\chi^2 > \chi_{l-1}(p)\right\} = P\left\{x^2 < x^2_{l-1}(1-p)\right\} \qquad (14),[1]$$

Oczywiście jednorazowe trafienie w przedział (14) jeszcze o niczym nie świadczy, podobnie jak jednorazowe wyjście poza ten przedział. Obliczenia trzeba przeprowadzać kilkakrotnie i zaobserwować zachowanie się wartości χ^2. Przy poziomie ufności p = 0,95 wyjście poza przedział występuje w wypadku słuszności hipotezy, średnio raz na 10 razy.
Przeprowadzono badania na zbiorze 10000 liczb pseudolosowych i dla 10 różnych wartości początkowych, dzieląc przedział [0,1] na 10 równych części. Z tablic dla p = 0,95 i l = 10 odczytujemy:

$$\chi^2_9(0,95) = 16,9$$

$$\chi^2_9(0,05) = 3,32$$

$$3.32 < \chi^2 < 16,9.$$

Wyniki obliczeń podano w tablicy 3.

Inną miarą niezgodności między rozkładem teoretycznym a empirycznym jest kryterium Kołmogorowa oparte na statystyce

$$D_n = \left| F_n(x) - F(x) \right|, \tag{15}$$

gdzie $F_n(x)$ jest empiryczną funkcją rozkładu równą

$$F_n(x) = \frac{m_x}{N}, \tag{16}$$

(N-liczebność badanego zbioru, m_x - ilość elementów w próbie nie większych od x).F(x) jest teoretyczną funkcją rozkładu i w naszym wypadku

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leqslant x \leqslant 1 \\ 0 & x > 1 \end{cases}. \tag{17}$$

Dla dowolnej dystrybuanty F(x) prawdopodobieństwo $P\left\{D_n < \frac{\lambda}{\sqrt{N}}\right\}$ dąży dla $N \to \infty$ do granicy

$$K(\lambda) = 1 - 2 \sum_{\nu=1}^{\infty} (-1)^{\nu-1} e^{-2\nu^2\lambda^2}. \tag{18}[2]$$

Wartość $1-K(\lambda)$ można znaleźć w tablicach.

Sprawdzenie zgodności rozkładu testem D_n przeprowadzono na 10 próbach o liczebności 1000 każda. Hipotezę o rozkładzie równomiernym odrzucano, jeżeli $D_n \sqrt{N} > 1{,}5$, odrzucano ją także,gdy $D_n \sqrt{N} < 0{,}5$ z powodów omówionych powyżej. Wyniki przytaczamy w tablicy 3.

Wraz z kryterium Kołmogorowa często stosuje się kryterium ω^2. Kryteria te mają tę przewagę nad testem χ^2, że nie dzielą przedziału co jest nieuchronnie związane z pewną stratą informacji.

Kryterium ω^2 opiera się na statystyce

$$\omega^2 = \int_{-\infty}^{\infty} \left[F_n(x)-F(x)\right]^2 dF(x) \,. \tag{19}$$

Jeżeli ułożyć losowane wartości $x_1, x_2, \ldots, x_N$ w szereg wariacyjny, otrzymamy

$$\omega^2 = \frac{1}{12N^2} + \frac{1}{N} \sum_{i=1}^{N} \left[F(x_i) - \frac{2i-1}{2N}\right]^2, \tag{20}$$

a w naszym wypadku

$$\omega^2 = \frac{1}{12N^2} + \frac{1}{N} \sum_{i=1}^{N} \left[x_i - \frac{2i-1}{2N}\right]^2 . \tag{21}$$

Odrzucamy hipotezę o równomierności rozkładu, jeżeli $\omega^2 > 0{,}0005$. Wyniki prezentujemy w poniższej tablicy.

Tablica 3

Numer wartości początkowej	χ^2 N = 10000	ω^2 N = 1000	$D_n\sqrt{N}$ N = 1000
1	6,370	0,000025	0,487
2	11,392	0,000300	0,920
3	4,660	0,000210	0,999
4	6,318	0,000420	1,246
5	9,326	0,000080	0,914
6	9,176	0,000250	1,268
7	9,406	0,000180	0,904
8	2,250	0,000710	1,6567
9	8,606	0,000150	0,996
10	12,620	0,000100	0,794
	$3{,}32 < \chi^2 < 16{,}9$	$\omega^2 < 0{,}005$	$0{,}5 < D_n\sqrt{N} < 1{,}5$

Z tablicy 3 wynika, że testy χ^2, ω^2 i D_n sprawdzają się przy założonym poziomie istotności 5%, czyli nie ma podstaw do odrzucenia hipotezy, że losowana populacja liczb pseudolosowych ma rozkład równomierny w przedziale $[0,1]$.

Celem ostatecznego sprawdzenia napisano program obliczający przy pomocy metod Monte Carlo stosunek objętości hipersfery n-wymiarowej do objętości hiperześcianu n-wymiarowego, który to stosunek znamy a priori z wzoru:

$$k_n = \frac{V_n}{E_n} = \frac{\prod^{\frac{n}{2}}}{n2^{n-1}\,\Gamma(\frac{n}{2})} . \qquad (22)$$

Wykonano obliczenia dla n=2 oraz n=4. Z wzoru (22) otrzymujemy $k_2 = \pi/4 = 0{,}785$, $k_4 = \pi^2/32 = 0{,}308$.

Dla N = 1000 realizacji i 10 różnych ciągów liczb pseudolosowych otrzymano następujące rezultaty:

Tablica 4

	Numer wartości początkowej										
	1	2	3	4	5	6	7	8	9	10	11
n=2	0,780	0,784	0,801	0,769	0,796	0,790	0,797	0,766	0,787	0,775	0,785
n=4	0,308	0,306	0,299	0,302	0,294	0,306	0,314	0,306	0,312	0,289	0,303

Uzyskane wyniki są zadowalająco zbieżne do wartości obliczonych bezpośrednio ze wzoru, co upoważnia nas do przypuszczenia, że maszyna generuje rzeczywiście ciągi liczb pseudolosowych o rozkładzie równomiernym w przedziale $[0,1]$.

LITERATURA

[1] Buslenko - Metody Monte Carlo.

[2] Smirnow N.W., Dunin-Barkowski I.W. - Kurs rachunku prawdopodobieństwa i statystyki matematycznej. Wydanie 2.Warszawa 1969, PWN, s. 316-320.

[3] Kredencer B.P. i inni - Rieszenije zadač nadiežnosti i eksploatacji na uniwersalnych ECWM. Wydanie 1. Moskwa 1967 r. Sowietskoje Radio, s. 82-88.

ВЫРАБОТКА СЛУЧАЙНЫХ ЧИСЕЛ НА ЭЛЕКТРОННОЙ ВЫЧИСЛИТЕЛЬНОЙ МАШИНЕ УМЦ-I

Р е з ю м е

В статье подаётся результаты исследования генератора случайных чисел о равномерном законе распределения, работающего на базисе модифицированного алгорифма фон Неймена.

THE COMPUTER UMC-1 AS PSEUDORANDOM-NUMBERS GENERATOR

S u m m a r y

In the paper the results of research of the pseudorandom uniformly distributed numbers generator, built on the basis the von Neumann's algorithm, are described.