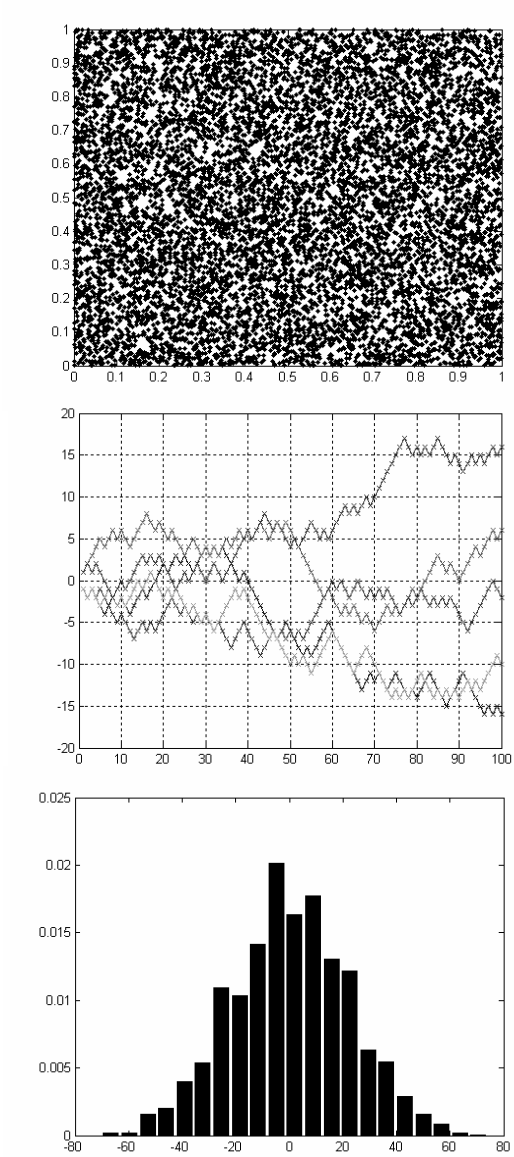
Modelowanie stochastyczne I

Generatory liczb losowych
Generatory zmiennych losowych o zadanych rozkładach



Paweł Cibis
#136008
26 X 2005

Zadanie 5

W zadaniu tym należy zbadać jakość generatora liniowego opartego na formule $X_{n+1} = (aX_n + c)\bmod m$ przy różnych parametrach a, c oraz m.

Generator mieszany (czyli generator liniowy o $c \neq 0$) osiąga pełny okres m jeżeli parametry te spełniają następujące warunki:

(1) liczby c i m nie mają wspólnych dzielników;

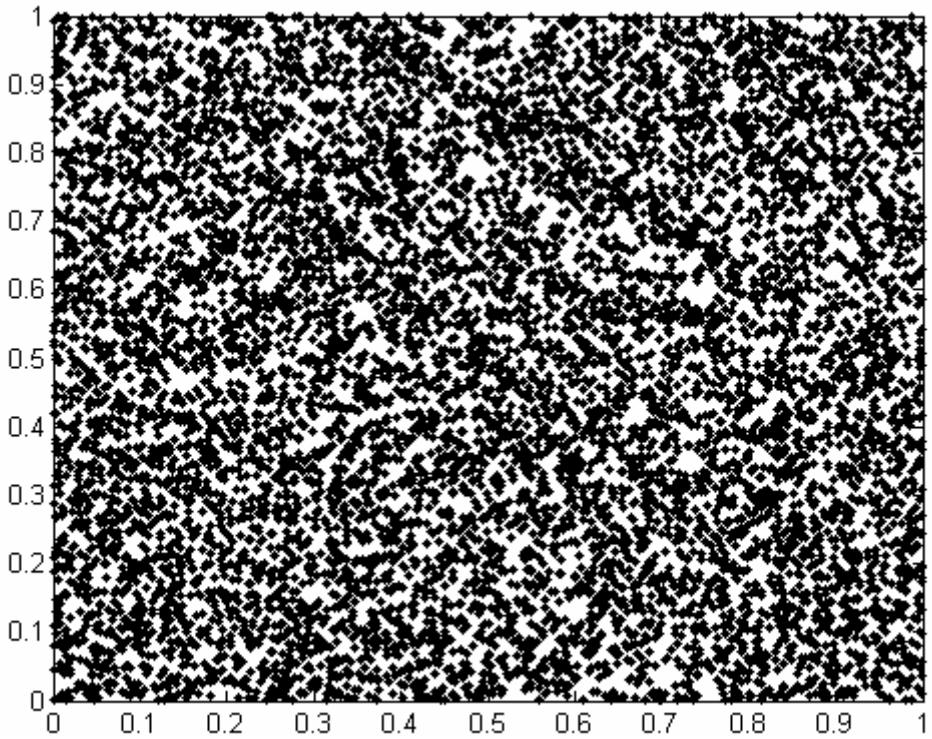
(2) $a \equiv 1 \bmod p$, dla każdego czynnika pierwszego liczby m;

(3) $a \equiv 1 \bmod 4$, jeśli 4 jest dzielnikiem liczby m.

Dla danego m jest to więc generator o najdłuższym możliwym okresie, ponieważ istnieje dokładnie tyle reszt z dzielenia przez m.

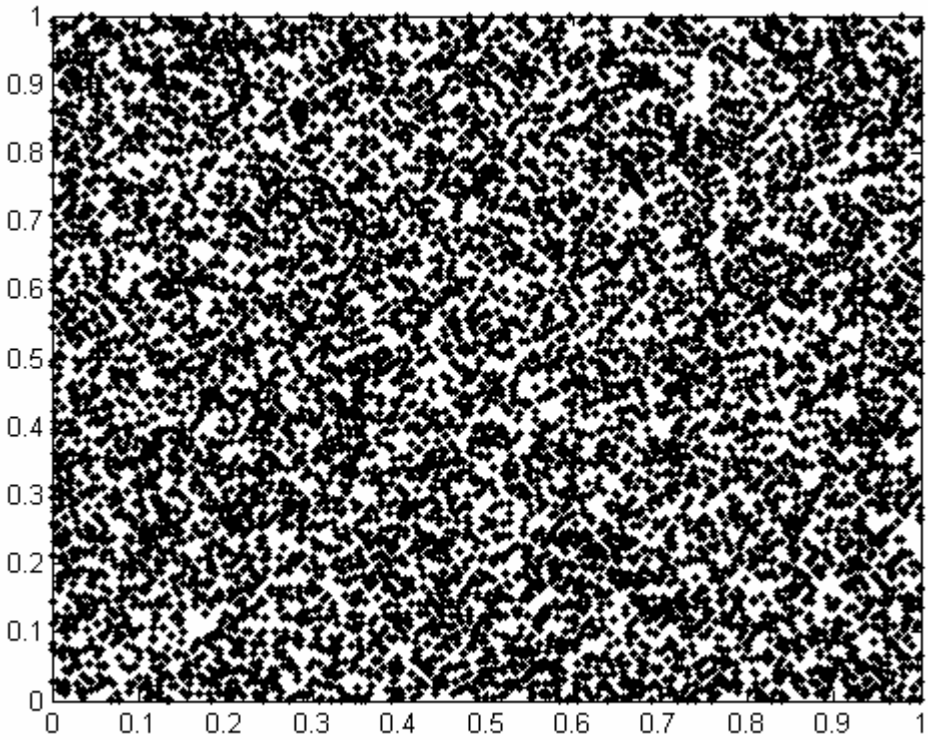
Przetestuję ten generator dla danych spełniających powyższe założenia, a także sprawdzę, co się stanie, gdy zostaną one osłabione. Będę się przy tym posługiwał prostym skryptem MATLABa, który najpierw generuje liczby z przedziału [0; m] za pomocą podanej powyżej formuły, a następnie tworzy z nich punkty postaci (X_1, X_2), (X_2, X_3), ..., (X_{n-1}, X_n) i nanosi je na wykres. Dla większej elegancji, skrypt przeskalowuje wygenerowane liczby tak, aby mieściły się w przedziale [0; 1]. Dzięki temu skala na każdym z wykresów będzie taka sama. Ocena generatora będzie polegała na wzrokowej analizie wykresu – o dobrych własnościach generatora będzie świadczyło gęste i równomierne rozłożenie punktów (brak skupisk oraz „białych plam") oraz brak regularnych kształtów utworzonych przez naniesione punkty).

1) $a = 69069$, $c = 1$, $m = 2^{32}$, $x_0 = 1$, $n = 10000$



Rys. 1. Wykres zbioru punktów dla generatora zaproponowanego przez G. Marsaglię w 1972 r.

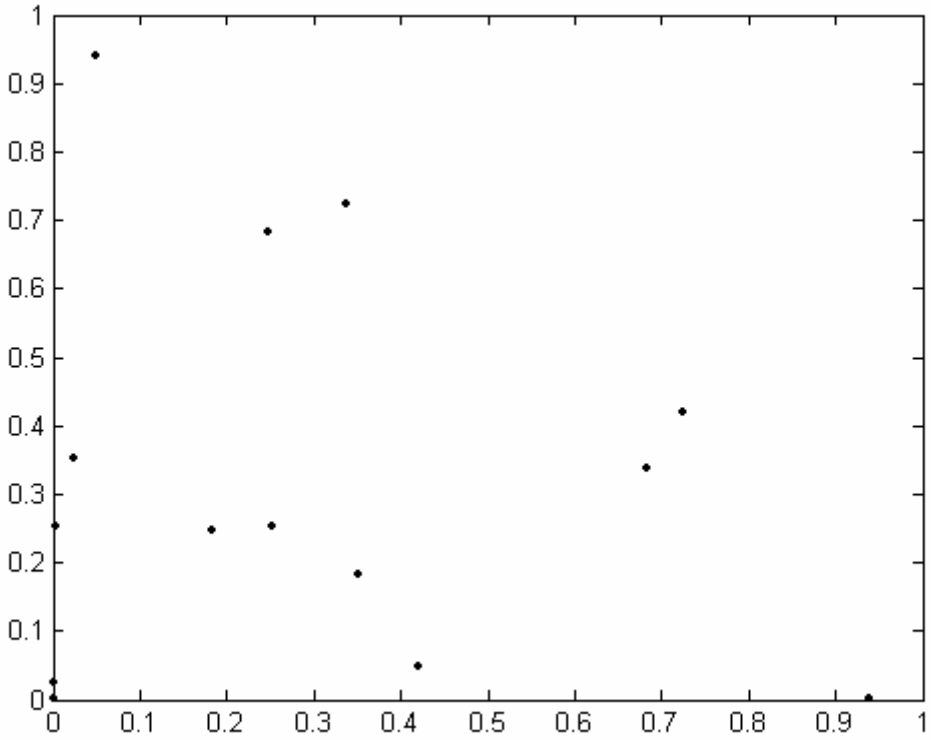
2) $a = 1103515245$, $c = 12345$, $m = 2^{31}$, $x_0 = 23$, $n = 10000$



Rys. 2. Wykres zbioru punktów dla generatora stosowanego w języku ANSI C.

Gdy parametry spełniają zadane założenia, to przy odpowiednio licznej próbie widać, iż otrzymany wykres pozbawiony jest regularnych kształtów, a punkty rozmieszczone są bardzo gęsto i nie pozostawiają większych pustych obszarów, nie widać też wyraźnych skupisk. Można więc stwierdzić, że generatory te dobrze spełniają kryteria akceptacji generatora liczb pseudolosowych.

Teraz wykonam test dla parametrów, które nie spełniają wszystkich założeń:
3) $a = 100$, $c = 1$, $m = 2^{32}$, $x_0 = 1$, $n = 10000$

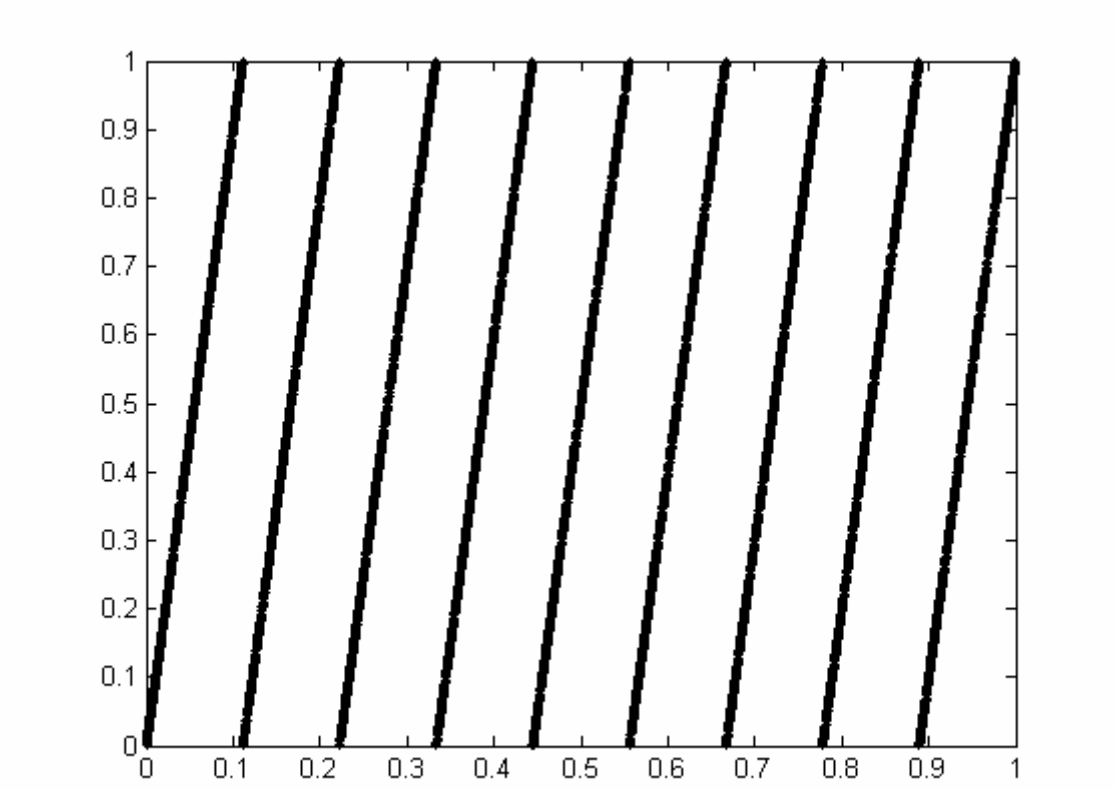


Rys. 3. Wykres zbioru punktów dla generatora nie spełniającego warunków (2) i (3).

Jest to przypadek podobny do pkt 1., ale tym razem zmodyfikowałem parametr a, tak by nie zachodził warunek (2) i (3). W istocie: 100-1 = 99. Liczba 99 nie jest podzielne ani przez 4, ani przez 2 (jedyny czynnik pierwszy liczby 2^{32}). Ilość punktów widocznych na wykresie znacznie zmalała, choć ilość generowanych liczb nie zmieniła się. Wobec tego punkty te pokrywają się, a co za tym idzie okres generatora jest mały. Ponadto, w przypadku gdy a jest liczbą parzystą, ciąg generowanych liczb jest od pewnego miejsca stały. Jeżeli zaś dodatkowo przyjmiemy jeszcze $c = 0$ – naruszając warunek (1) – dla danego x_0 będzie generowana wyłącznie jedna liczba. Dlatego generator oparty na takich parametrach jest bardzo słaby.

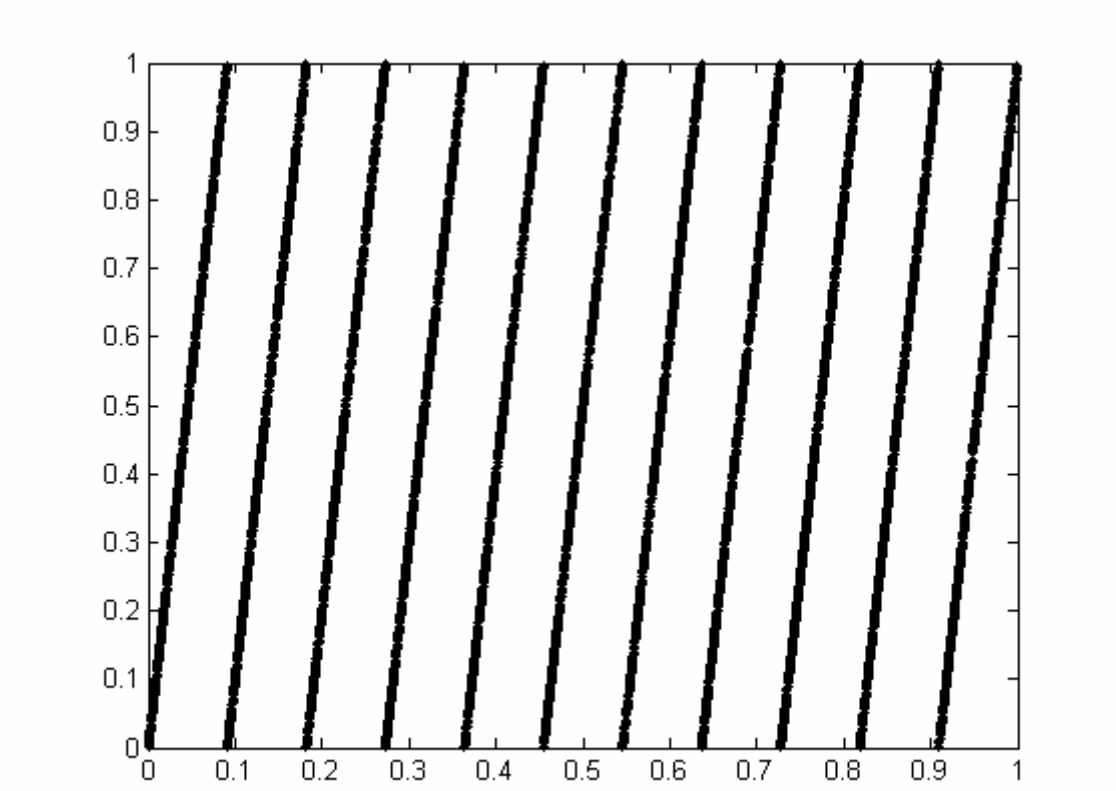
Teraz pokażę, iż nawet jeżeli spełnione są wszystkie 3 warunki, nie zawsze otrzymany generator będzie akceptowalny.

4) $a = 9$, $c = 1$, $m = 2^{32}$, $x_0 = 1$, $n = 10000$



Rys. 4. Wykres zbioru punktów dla generatora słabego pomimo spełniania wszystkich 3 warunków.

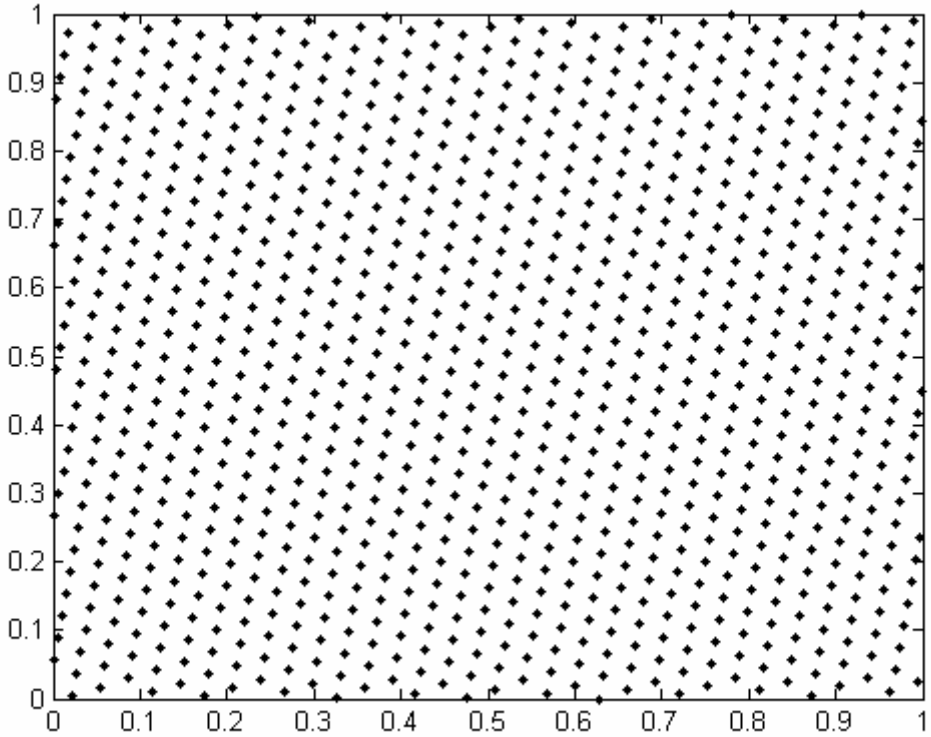
5) $a = 11$, $c = 1$, $m = 2^{32}$, $x_0 = 1$, $n = 10000$



Rys. 5. Wykres zbioru punktów dla generatora słabego pomimo spełniania wszystkich 3 warunków.

Tym razem punkty wprawdzie nie pokrywają się – ponieważ spełnione zostały 3 warunki i okres generatora jest maksymalny – ale skupiają się one bardzo gęsto, tworząc proste, których liczba odpowiada wartości parametru *a*. Stąd wniosek, iż parametr ten w dobrym generatorze musi być odpowiednio duży – inaczej liczb przez niego generowanych nie można uznać za losowe.

6) $a = 1103515245$, $c = 12345$, $m = 2^{10}$, $x_0 = 23$, $n = 10000$



Rys. 6. Wykres zbioru punktów dla generatora o zbyt małym choć maksymalnym okresie.

maksymalny możliwy okres generatora nie umożliwi nam otrzymania losowych liczb. Na rysunku punkty, z powodu małego okresu, pokrywają się oraz tworzą regularne kształty.

Warto także dodać, iż parametr m często jest postaci 2^e, gdzie e jest pewną liczbą naturalną. Dzięki temu w implementacji algorytmu niezwykle „niewygodną" dla komputera i przez to powolną operację dzielenia można zastąpić szybkim przesunięciem bitowym. Pozwoli to zwiększyć prędkość działania generatora. Dzielenia można także uniknąć przy zastosowaniu parametru m postaci $2^e \pm 1$.

Zadanie 11

Zadanie to polega na analizie trajektorii symetrycznego błądzenia losowego. Zmienną losową będzie tu pozycja końcowa trajektorii tego procesu stochastycznego. Za pomocą skryptu MATLABa generuję pewną ilość trajektorii i przedstawiam je na wykresie. Na jego podstawie mogę dokonać wstępnej analizy rozkładu pozycji końcowej trajektorii. Następnie buduję unormowany histogram pozycji końcowych, którego kształt będzie zbliżony do funkcji gęstości rozkładu analizowanej zmiennej losowej. Na końcu obliczam parametry zaobserwowanego rozkładu.

Na potrzeby zadania zdefiniuję symetryczne błądzenie losowe, jako ruch punktu materialnego na prostej, taki że:

- rozpoczyna się w punkcie 0,
- w jednostce czasu punkt materialny pokonuje odległość 1 wzdłuż osi Y (w górę lub w dół),
- w każdej jednostce czasu prawdopodobieństwo poruszenia się w dół jest równe prawdopodobieństwu poruszenia się w górę.

Ruch punktu materialnego w każdej jednostce czasu jest więc opisany zmienną losową o symetrycznym rozkładzie dwupunktowym, przyjmującą wartości 1 lub –1.

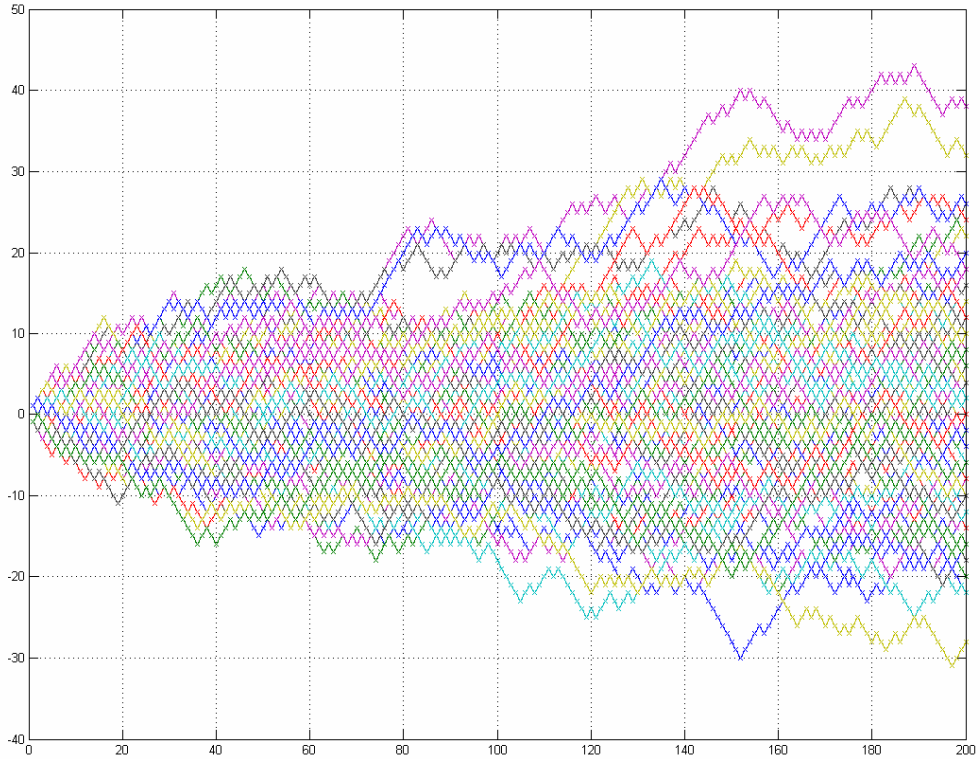
Oczywiście w symulacji przyjmuję dyskretny model czasu.

Na podstawie powyższych założeń mogę wygenerować dowolną ilość trajektorii takiego ruchu. Jak widać na rys. 7, mogą się one kształtować bardzo różnie:

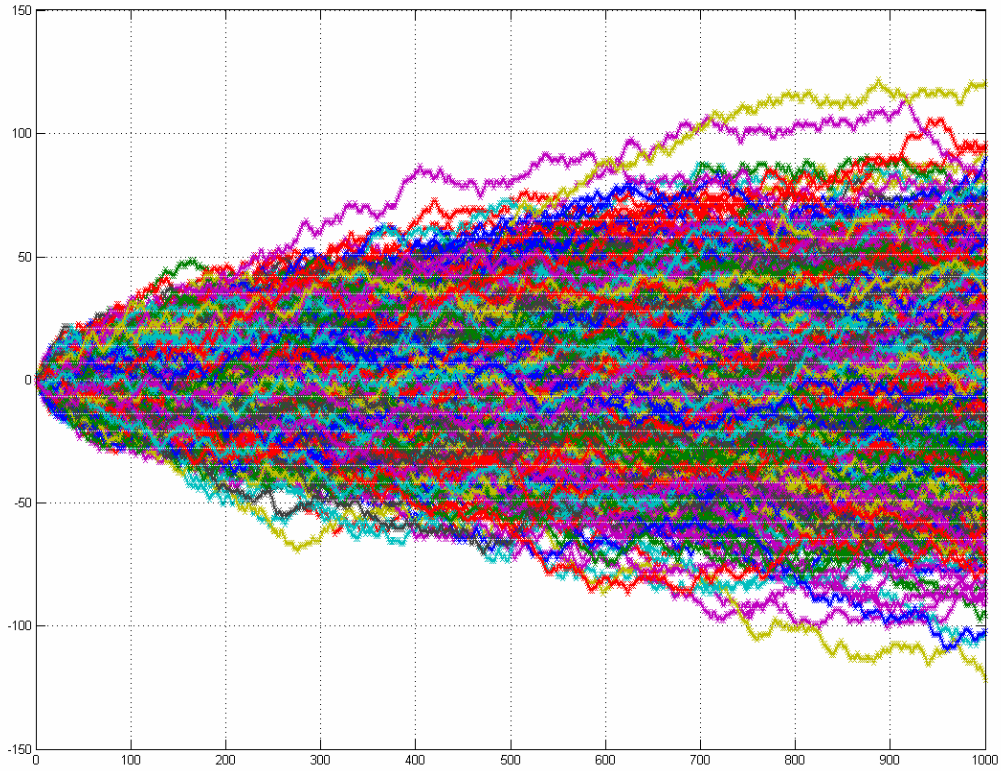
Rys. 7. Dwie trajektorie dla T = 2000.

Jednakże, gdy znacznie zwiększę liczbę trajektorii, obserwuję ciekawe zjawisko – nie rozchodzą się one bezładnie na wykresie, lecz są rozmieszczone bardzo gęsto i symetrycznie względem punktu startowego. Pozycje końcowe nie są też zbytnio oddalone od pozycji początkowej. Na rys. 8 przedstawiłem 2000 trajektorii o długości 200. Teoretycznie pozycja końcowa każdej z nich może być liczba z przedziału [–200, 200]. W rzeczywistości jednak zdecydowana większość pozycji końcowych zawiera się w przedziale [–30, 30] (a więc w jedynie 20% przedziału teoretycznego), a żadna z tych pozycji nie przekracza (z dokładnością do znaku) 60.

Po zwiększeniu długości do 1000, większość obserwacji znajduje się w przedziale [–100; 100], co stanowi 10% teoretycznego przedziału [–1000; 1000]. 5-krotne zwiększenie długości trajektorii, spowodowało niewiele ponad 3-krotny wzrost rozpiętości przedziału zawierającego większość pozycji końcowych (rys. 9). Wypełnienie przedziału teoretycznego spadło o 10%. Rozproszenie położeń końcowych następuje więc bardzo wolno wraz ze wzrostem długości trajektorii.

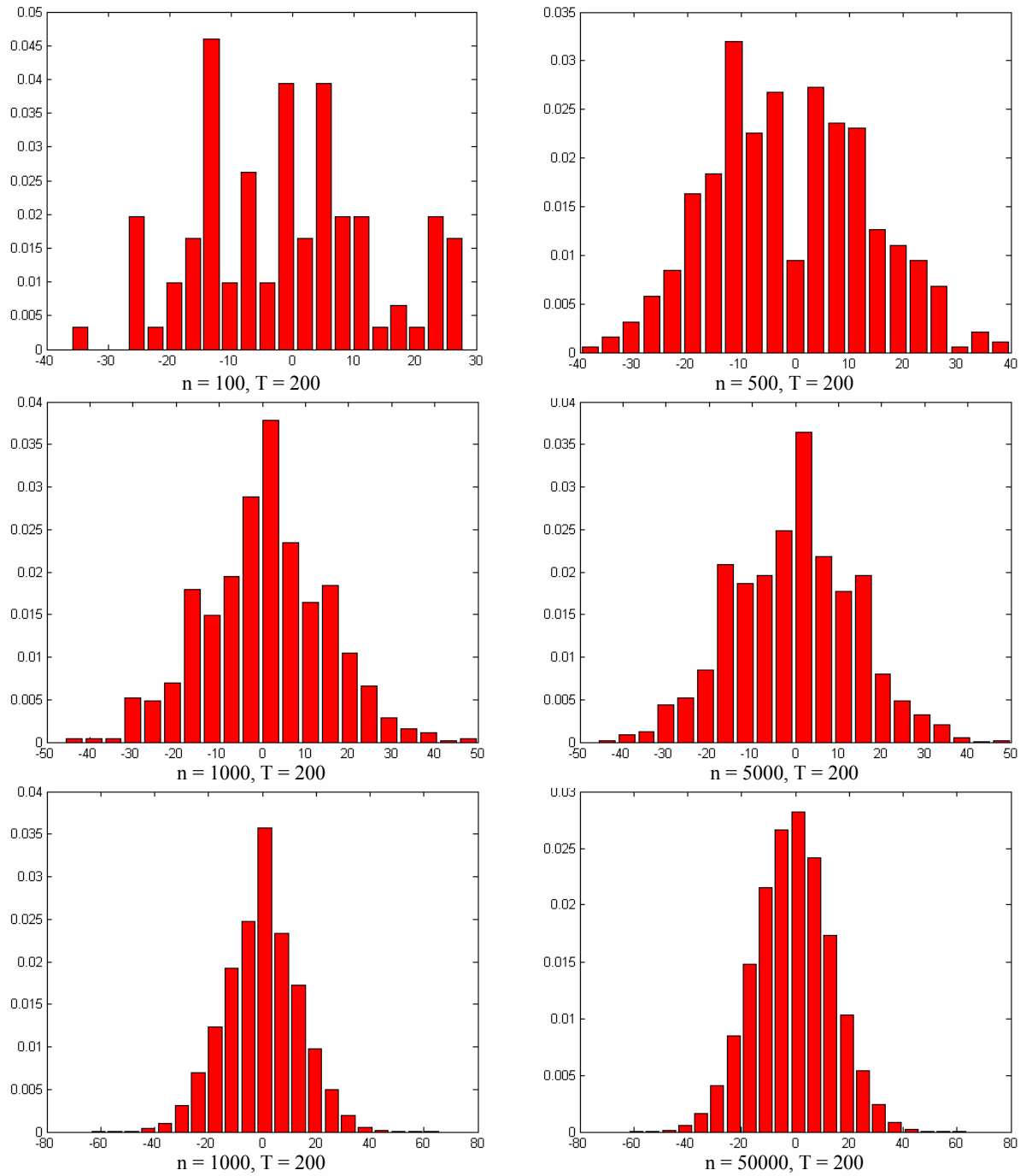


Rys. 8. Dwa tysiące trajektorii o T = 200.



Rys. 9. Dwa tysiące trajektorii o T = 1000.

Spróbuję teraz ocenić rozkład pozycji końcowej trajektorii, korzystając z unormowanego histogramu.



Rys. 10. Unormowane histogramy rozkładu pozycji końcowej trajektorii symetrycznego błądzenia losowego.

Wraz ze wzrostem liczby generowanych trajektorii unormowany histogram rozkładu pozycji końcowej trajektorii coraz bardziej przypomina wykres funkcji gęstości rozkładu normalnego. Jest to zgodne ze wcześniejszymi obserwacjami wykresów trajektorii – to właśnie w rozkładzie normalnym obserwacje skupiają się wokół wartości średniej (w naszym przypadku: punktu początkowego), a ich zdecydowana większość mieści się w pewnym przedziale (3-sigmowym). W symulacji tej można dostrzec funkcjonowanie centralnego twierdzenia granicznego Lindberga-Lèvy'ego – zmienna losowa, jaką jest pozycja końcowa

trajektorii analizowanego błądzenia losowego, jest sumą pewnej liczby zmiennych losowych o jednakowym rozkładzie. Mogę więc zapisać:

$$Z_T = X_1 + X_2 + ... + X_T.$$

Ponieważ każdy ze składników jest zmienną losową o rozkładzie dwupunktowym, o równych prawdopodobieństwach i zbiorze wartości $\{-1, 1\}$, a zmienne te są wzajemnie niezależne, mogę obliczyć:

$$E(X_i) = 0{,}5 * 1 + 0{,}5 * (-1) = 0, \text{ dla } i = 1, 2, ..., T,$$

$$Var(X_i) = 0{,}5 * 1^2 + 0{,}5 * (-1)^2 - 0^2 = 1, \text{ dla } i = 1, 2, ..., T.$$

Następnie, korzystając z własności wartości oczekiwanej i wariancji, otrzymuję:

$$E(Z_T) = T * 0 = 0,$$

$$Var(Z_T) = T * 1 = T.$$

Stąd, na mocy CTG:

$$Z_n \sim N(0, T).$$

Obliczenia wykonane z zastosowaniem CTG potwierdzają więc wnioski wyciągnięte z obserwacji wykresów i histogramów.