Tạp chí Tin học và Điều khiển học, T.22, S.1 (2006), 53–62

NGHIÊN CỨU THỰC NGHIỆM TÍNH PHÂN BỐ ĐỀU CỦA CÁC DÃY SỐ NGẪU NHIÊN, GIẢ NGẪU NHIÊN VÀ TỰA NGẪU NHIÊN

NGUYỄN VĂN HÙNG, BÙI VĂN THANH

Viện Công nghệ thông tin

Abstract. In numerical Monte Carlo methods it has been remarked that the randomness of sequences is unnecessary, the most importance is their uniformity. Therefore the uniformly distributed non-random sequences have get more and more important role. These are quasi-random sequences with discrepancy as a measure of uniformity. Most of the 23 sequences studied satisfy the uniformity criterion, thus they can be used effectively for calculations in Monte Carlo schema's. The algorithm proposed by the author of [12] based on the moment of sequences (criterion of uniformity of second type), however, the generated sequences V also met all the criteria of uniformity of first type (measured by discrepancy of distribution funcrions). All the tests show that these quasi-random sequences V are significantly better than random and pseudo-random sequences.

Tóm tắt. Trong phương pháp Monte Carlo số trị người ta nhận thấy rằng tính ngẫu nhiên của các dãy số là hoàn toàn không cần thiết, mà quan trọng hơn cả là tính chất phân bố đều của chúng. Vì thế các dãy số không-ngẫu-nhiên phân bố đều có vai trò ngày càng quan trọng. Đó chính là các dãy số tựa ngẫu nhiên với độ đo của tính đều là độ phân kỳ. Hầu như tất cả 23 dãy số được khảo sát đều thỏa mãn các tiêu chuẩn về tính đều, như vậy có thể sử dụng hữu hiệu chúng trong các tính toán theo sơ đồ Monte Carlo. Thuật toán do tác giả ([12]) đề xuất dựa trên các giá trị của moment (tiêu chuẩn phân bố đều loại 2), nhưng các dãy số V được tạo ra cũng đáp ứng đầy đủ cả các tiêu chuẩn phân bố đều loại 1 (về độ phân kỳ của các hàm phân phối). Qua tất cả các phép thử các dãy số tựa ngẫu nhiên V này đều vượt trội so với các dãy số ngẫu nhiên và giả ngẫu nhiên.

1. SỰ PHÂN BỐ ĐỀU CỦA CÁC DÃY SỐ

1.1. Các số ngẫu nhiên và giả ngẫu nhiên

Phương pháp Monte Carlo là phương pháp số để giải các vấn đề toán học bằng cách mô hình hóa các đại lượng ngẫu nhiên và đánh giá các đặc trưng thống kê của chúng.

Có thể nhận được đại lượng ngẫu nhiên từ ba nguồn sau đây:

a) Các bảng số ngẫu nhiên

Đúng ra phải gọi là bảng các chữ số ngẫu nhiên. Bảng số ngẫu nhiên lớn nhất đã được công bố bao gồm 1 triệu chữ số ([7]).

b) Máy phát số ngẫu nhiên

Một quá trình vật lý ngẫu nhiên như bức xạ của các chất phóng xạ hay tạp âm của đèn điện tử có thể góp phần sinh ra các số ngẫu nhiên thực sự ([1,5]). Thiết bị phụ trợ

cho máy tính để làm việc này được gọi là máy phát, bộ tạo sinh hay nguồn số ngẫu nhiên (random-number generator).

c) Các số giả ngẫu nhiên (pseudo-random numbers)

Các số ngẫu nhiên trên máy tính điện tử thực chất là các số giả ngẫu nhiên. Đó là các số gần giống với số ngẫu nhiên, được tạo ra bởi một hệ thức đệ quy hoặc một thuật toán cho trước. Chúng có thể thay thế cho các số ngẫu nhiên khi giải một số loại bài toán nào đó. Các phép thử thống kê sẽ đánh giá mức độ thích hợp đối với các yêu cầu về số ngẫu nhiên.

1.2. Các số tựa ngẫu nhiên (quasi-random numbers)

Trong một số tính toán theo sơ đồ Monte Carlo, người ta nhận thấy rằng tính ngẫu nhiên của các điểm là không cần thiết, mà quan trọng nhất là tính đều của chúng ([5, 9]). Thay cho các số ngẫu nhiên, người ta chủ định dùng tập hợp các số không-ngẫu-nhiên (non-random) phân bố đều để tăng tốc độ hội tụ (theo ý nghĩa thông thường chứ không phải hội tụ ngẫu nhiên) mà không làm hỏng cấu trúc của thuật toán Monte Carlo tương ứng. Các số giả ngẫu nhiên đơn định (deterministic pseudo-random numbers) như thế được gọi là *số tựa ngẫu nhiên*. Các phương pháp dùng đến số tựa ngẫu nhiên được gọi là *phương pháp tựa Monte Carlo* (quasi-Monte-Carlo methods).

Tuy các số tựa ngẫu nhiên có một số hạn chế, nhưng ưu điểm chính của chúng là bảo đảm được sự hội tụ của thuật toán và không cần đến các thử nghiệm thống kê. Điều quan trọng hơn cả là tăng được tốc độ hội tụ và giảm bậc sai số từ $1/N^{1/2}$ xuống $1/N^{1-\varepsilon}$, trong đó $\varepsilon > 0$ là số nhỏ tùy ý ([9]). Trong một số trường hợp sai số có bậc $\ln N/N$.

Người ta đã đề xuất những dãy số tựa ngẫu nhiên hay nói chính xác hơn là những dãy số có độ phân kỳ thấp, như dãy Van der Corput [9], dãy Faure [10], dãy Halton [1, 9], dãy loga [5], dãy LPt [9], dãy Niederreiter [10], dãy Sobol [10]. Nói chung các thuật toán tạo ra chúng đều dựa trên các tiêu chuẩn phân bố đều loại 1 (về độ phân kỳ giữa các hàm phân phối mẫu và lý thuyết).

Trong [12] đã trình bày một thuật toán đơn giản tạo dãy số V phân bố đều dựa trên các đẳng thức về moment (tiêu chuẩn phân bố đều loại 2). Trong bài này chúng tôi tiến hành kiểm định thống kê trên các dãy số tựa ngẫu nhiên V này theo cả hai tiêu chuẩn phân bố đều loại 1, loại 2 và so sánh với các dãy số ngẫu nhiên, giả ngẫu nhiên khác nhau.

1.3. Các tiêu chuẩn thống kê về phân bố đều

Nếu có một dãy số phân bố đều trong khoảng (0,1) thì bằng phép biến đổi tuyến tính đơn giản ta có thể trải đều ra trong khoảng (a, b) bất kỳ. Vì thế dù là số ngẫu nhiên hay không-ngẫu-nhiên, cũng chỉ cần khảo sát sự phân bố đều trong khoảng (0,1).

Có hai loại tiêu chuẩn về phân phối đều của một dãy số ([5]):

+ Loại 1 dựa trên độ khác biệt của các hàm phân phối mẫu và lý thuyết. Để kiểm định tiêu chuẩn loại này, thường dùng nhất là ba phép thử: χ^2, Kolmogorov và ω^2 ([1, 5, 12]). Đối với các số tựa ngẫu nhiên còn thêm độ phân kỳ ([9, 10]).

+ Loại 2 dựa trên các giá trị của moment mẫu và lý thuyết. Việc so sánh các moment sẽ được trình bày trong Mục 3.2.

1.4. Độ phân kỳ

Độ phân kỳ (discrepancy) của dãy số $X_0, X_1, ..., X_{N-1}$ được định nghĩa như sau:

$$D(X_0, X_1, ..., X_{N-1}) = D = \sup_{0<x<1} |S_N(x) - N.x|,$$

trong đó $S_N(x)$ là số điểm thuộc khoảng $[0, x)$.

Nếu X_i đã được sắp: $X_1 \leqslant X_2 \leqslant ... \leqslant X_N$ thì hệ thức trên đây được rút gọn thành công thức Niederreiter:

$$D = \frac{1}{2} + \max_{1\leqslant i\leqslant N} |i - \frac{1}{2} - N.X_i|.$$

Công thức này cho ta thấy: cực tiểu của D là $1/2$ và đạt được khi $X_i = (i-1/2)/N$ trong đó $i = 1, 2, ..., N$. Trong trường hợp N cố định, dãy số này là tối ưu theo tiêu chuẩn của độ phân kỳ. Nhưng khi chuyển từ N sang $(N+1)$ thì tất cả mọi điểm đều phải thay đổi. Ví dụ dãy số tối ưu với $N = 3$ là $(1/6, 3/6$ và $5/6)$, với $N = 4$ lại là $(1/8, 3/8, 5/8, 7/8)$. Như vậy không thể xây dựng được một dãy vô hạn các điểm X_i sao cho đoạn $X_0, X_1, X_2, ..., X_k, (k = 2, ..., N)$ nào cũng phân bố đều tối ưu ([9]).

Nhà toán học Hà Lan J. G. van der Corput đã nhận xét rằng không thể nào xây dựng được dãy số X_i sao cho $D(X_0, X_1, ..., X_{N-1})$ bị chặn với mọi giá trị N. Giả thiết của ông đã được T. van Aardenne-Ehrenfest chứng minh chặt chẽ ([4]): độ phân kỳ của tất cả các dãy số đều thỏa mãn bất đẳng thức dạng $D \geqslant C. \ln\ln\ln N/N$, và trong dãy số X_i bất kỳ, cận trên của độ phân kỳ là vô cực $\limsup D(X_0, X_1, ..., X_{N-1}) = \infty$.

Kết quả này có nghĩa là trong dãy số X_i bất kỳ có những đoạn "xấu kém" với độ dài tùy ý mà ở đó độ phân kỳ tăng vô hạn.

Sau đó K. F. Roth đã chứng minh rằng, đối với những đoạn "xấu kém" như thế $D \geqslant C_1.\sqrt{\ln N}$, trong đó C_1 là hằng số tuyệt đối (không phụ thuộc vào dãy số đang xét).

Cuối cùng W. M. Schmidt đã chính xác hóa thêm rằng $D \geqslant C_2. \ln N$, trong đó C_2 là một hằng số tuyệt đối khác. Bậc của đại lượng $\ln N$ trong đánh giá trên đây là không thể cải tiến được nữa ([9]).

2. CÁC DÃY SỐ SO SÁNH THỬ NGHIỆM

Việc thử nghiệm so sánh trong bài này được tiến hành trên 23 dãy số, mỗi dãy có 2459 phần tử (2459 là số nguyên tố).

+ 4 dãy số lấy từ các kết quả xổ số, ký hiệu là L_i (Lottery).

+ 5 dãy số do máy tính tạo ra, ký hiệu là R_i (Random).

+ 6 dãy số lấy từ các bảng số ngẫu nhiên, ký hiệu là T_i (Table).

+ 8 dãy số tựa ngẫu nhiên, tính theo thuật toán của [12], ký hiệu là V_i.

2.1. Các dãy số ngẫu nhiên và giả ngẫu nhiên

Có thể coi các kết quả rút thăm xổ số là những số ngẫu nhiên thực sự.

Các bảng số ngẫu nhiên phần lớn do máy tính tạo ra, tức là bao gồm các số giả ngẫu nhiên. Cũng có bảng số được tạo thành từ một nguồn ngẫu nhiên nào đó. Vì không biết rõ xuất xứ, cho nên không thể khẳng định là ngẫu nhiên hay giả ngẫu nhiên.

Còn các số ngẫu nhiên do máy tính tạo ra thì là số giả ngẫu nhiên theo đúng định nghĩa.

2.1.1. Các dãy số Li được tạo thành từ các kết quả xổ số

Dãy L_1 lấy từ kết quả xổ số đăng trên báo New York Times ngày 30-10-1940 và được in lại trong [2].

Dãy L_2 gồm 2 chữ số cuối các phiếu trúng thưởng hiện vật của xổ số Hungary kỳ tháng 2 năm 1987, in trên báo Népszabadság ngày 05-3-1987.

Dãy L_3 gồm 2459 kết quả trúng giải đặc biệt của xổ số kiến thiết miền Bắc từ ngày 03-11-1993 đến ngày 20-2-2002.

Dãy L_4 lấy 3 chữ số cuối các số chứng minh nhân dân của những người trúng thưởng chương trình khuyến mãi "Ngàn lộc xuân, đoạt ngàn Kim Mã", in trên báo Tuổi trẻ thành phố Hồ Chí Minh ngày 16-3-2002.

Đem các nhóm 5 chữ số liên tiếp chia cho 10^5 ta được các giá trị trong khoảng $(0, 1)$.

2.1.2. Các dãy số T_i lấy từ bảng số ngẫu nhiên

Các dãy số $T_1, ..., T_6$ lần lượt lấy từ các bảng số ngẫu nhiên của Ấn Độ ([8]), Đức ([6]), Hungary ([11]), Mỹ ([7]), Nga ([1]) và Pháp ([4]). Điều đặc biệt là trong bảng số của [8] có các chữ số ngẫu nhiên (random digits) và các hoán vị 10-chữ số (10-digit permutations). Chúng tôi thử chọn dãy số T_1 từ các hoán vị 10-chữ số và hậu quả sẽ thấy ở Mục 3.1.

2.1.3. Các dãy số giả ngẫu nhiên R_i do máy tính tạo ra

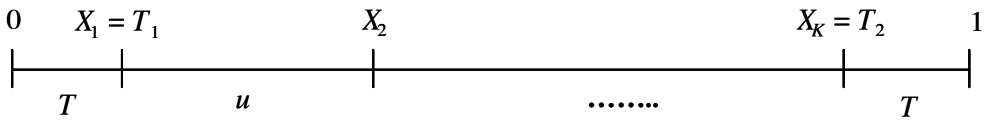
Các dãy số $R_1, ..., R_5$ gồm các số giả ngẫu nhiên do hàm Random trong Turbo Pascal tạo ra ở những thời điểm khác nhau và đã được ngẫu nhiên hóa bởi lệnh Randomize.

2.2. Một thuật toán tạo số tựa ngẫu nhiên

Thuật toán do tác giả trong [12] đề xuất dựa trên tiêu chuẩn phân bố đều loại 2: moment bậc 1 và bậc 2 của các điểm phân bố đều trong (0,1) bằng moment tương ứng của phân phối đều liên tục.

Giả sử K điểm phân bố đều trong khoảng (0,1). Các điểm cách đều nhau một khoảng u. Điểm đầu (X_1) cách điểm 0 và điểm cuối (X_k) cách điểm 1 một khoảng T (Hình 1). Như vậy

$$2T + (K-1).u = 1$$
$$u = (1-2T)/(K-1)$$
$$X_1 = T$$
$$X_2 = T + u$$
$$...$$
$$X_K = T + (K-1).u$$



Hình 1

Đặt $S_1 = \sum X_i = K.T + u.[1 + 2 + \cdots + (K-1)] = K.T + u.(K-1).K/2,$

$$S_2 = \sum X_i^2 = K.T^2 + 2T.u.[1 + 2 + \cdots + (K-1)] + u^2[1 + 4 + 9 + \cdots + (K-1)^2]$$
$$= K.T^2 + 2T.u.(K-1).K/2 + u^2.(K-1).K.(2K-1)/6.$$

Các hệ thức trên cùng các phương trình về moment bậc 1 và bậc 2:

$$M_1(K) = S_1/K = 1/2 \text{ và } \ M_2(K) = S_2/K = 1/3$$

đưa tới phương trình bậc hai đối với T

$$T^2 - T + \frac{1}{2(K+1)} = 0,$$

phương trình này có hai nghiệm:

$$T_1 = \Big(1 - \sqrt{\frac{K-1}{K+1}}\Big)/2; \ T_2 = \Big(1 + \sqrt{\frac{K-1}{K+1}}\Big)/2; \ T_1 + T_2 = 1.$$

Đây chính là hai điểm X_1 và X_K. Từ đó rút ra:

$$u = (1 - 2T_1)/(K-1) = 1/\sqrt{(K-1)(K+1)}.$$

Dễ dàng tính được các điểm $X_2, X_3, ...$ từ các giá trị T_1, T_2 và u trên đây. Xin nêu vài trị số tính toán để minh họa:
với $K = 5$ ta tính được $T_1 = 0{,}09175$; $T_2 = 0{,}90825$ và $u = 0{,}20412$;
với $K = 12$ thì $T_1 = 0{,}04007$; $T_2 = 0{,}95993$ và $u = 0{,}08362$;
với $K = 90$ thì $T_1 = 0{,}00553$; $T_2 = 0{,}99447$ và $u = 0{,}01111$.

Khi cần tạo ra N điểm phân bố đều thì ta tách N thành tổng của h số hạng:

$$N = K_1 + K_2 + ... + K_h$$

sao cho các K_i đều khác nhau.

2.3. Các cách tạo dãy số tựa ngẫu nhiên V_i

V_i $(i = 1...8)$ là ký hiệu của 8 dãy số tựa ngẫu nhiên theo thuật toán đã trình bày ở trên. Mỗi dãy số được hợp thành từ h đoạn có số phần tử khác nhau, sao cho tổng số các phần tử là 2459. Chúng tôi xét nhiều cách tạo dãy số: tùy ý, dùng các dãy truy toán và dùng các số nguyên tố.

V_1 chỉ gồm 1 đoạn duy nhất $(h = 1)$.

V_2 $(h = 24)$:

$K_1 = 29, K_2 = 49, K_3 = 14, K_4 = 36, K_5 = 183, K_6 = 113, K_7 = 178, K_8 = 206,$
$K_9 = 80, K_{10} = 15, K_{11} = 22, K_{12} = 205, K_{13} = 75, K_{14} = 7, K_{15} = 198,$
$K_{16} = 193, K_{17} = 308, K_{18} = 53, K_{19} = 39, K_{20} = 6, K_{21} = 95, K_{22} = 203,$
$K_{23} = 97, K_{24} = 55.$

V_3 $(h = 12)$:

$K_1 = 24, K_2 = 415, K_3 = 23, K_4 = 722, K_5 = 608, K_6 = 45, K_7 = 88,$
$K_8 = 92, K_9 = 18, K_{10} = 311, K_{11} = 93, K_{12} = 20.$

Hai dãy V_4 và V_5 dưới đây được hợp thành bởi các dãy truy toán (recurrent sequences) $K_{i+1} = K_i + K_{i-1}$.

V_4 $(h = 5)$: $K_1 = 125, K_2 = 262, K_3 = 387, K_4 = 649, K_5 = 1036.$

V_5 $(h = 7)$: $K_1 = 63, K_2 = 82, K_3 = 145, K_4 = 227, K_5 = 372, K_6 = 599, K_7 = 971.$

Ba dãy số $V_6 - V_7 - V_8$ bao gồm các K_i là các số nguyên tố.

V_6 $(h = 15)$:

$K_1 = 17, K_2 = 1549, K_3 = 19, K_4 = 67, K_5 = 47, K_6 = 31, K_7 = 5,$
$K_8 = 89, K_9 = 277, K_{10} = 7, K_{11} = 71, K_{12} = 3, K_{13} = 23,$

$K_{14} = 97, K_{15} = 157.$
V_7 $(h = 9)$:
$K1 = 23, K_2 = 7, K_3 = 29, K_4 = 5, K_5 = 311, K_6 = 43, K_7 = 11, K_8 = 1993, K_9 = 37.$
V_8 $(h = 24)$:
$K_1 = 5, K_2 = 43, K_3 = 283, K_4 = 7, K_5 = 29, K_6 = 277, K_7 = 61,$
$K_8 = 113, K_9 = 149, K_{10} = 2, K_{11} = 13, K_{12} = 53, K_{13} = 73,$
$K_{14} = 239, K_{15} = 3, K_{16} = 47, K_{17} = 79, K_{18} = 211, K_{19} = 11,$
$K_{20} = 41, K_{21} = 281, K_{22} = 19, K_{23} = 31, K_{24} = 389.$

3. CÁC PHÉP THỬ VÀ CÁC TÍNH TOÁN KIỂM ĐỊNH

3.1. Kiểm định về sự khác biệt giữa các hàm phân phối

Để tính giá trị χ^2_N ta chia khoảng $(0, 1)$ ra 60 đoạn bằng nhau. Số bậc tự do sẽ là 59 và các giá trị χ^2 tới hạn tương ứng là ([6]):

χ^2= 77,93 với xác suất tin cậy õ = 0,95 và mức ý nghĩa (1−õ) =0,05 = 5%,
χ^2 = 87,17 với õ = 0,99,
χ^2 = 98,32 với õ = 0,999.

Nếu dãy số phân bố đều thì trên mỗi đoạn có trung bình 2459/60 = 40,983 phần tử. Chương trình máy tính đếm số phần tử thực tế trên mỗi đoạn và tính giá trị χ^2_N theo công thức.

Bốn cột 2-5 của Bảng 1 ghi kết quả tính toán kiểm định theo 4 phép thử nêu trong Mục 1.3 và 1.4. Ba dòng cuối cùng của bảng ghi các giá trị tới hạn ở mức ý nghĩa 5%, 1% và 0,1% (tương ứng với xác suất tin cậy 95%, 99% và 99,9%).

Trong 23 dãy số được khảo sát, chỉ có một trường hợp cần bác bỏ giả thiết về phân bố đều. Giá trị χ^2_N = 164,920 của dãy T_1 tính từ bảng số ngẫu nhiên của các tác giả Ấn Độ ([8]) vượt quá cả trị số 98,32 của mức ý nghĩa 0,1%. Như đã nói ở Mục 2.1.2., dãy số T_1 được tạo thành từ các hoán vị 10-chữ số chứ không phải từ các chữ số ngẫu nhiên. Trong các hoán vị này mỗi chữ số 0,1,..., 9 xuất hiện đúng 1 lần trong bộ 10 chữ số liên tiếp. Có thể coi các chữ số ngẫu nhiên như kết cục của sự lấy mẫu có hoàn lại, và các *hoán vị 10-chữ số* như kết cục của sự lấy mẫu không hoàn lại. Vì thế tính chất thống kê của chúng khác nhau.

Bảng 2 tính các đặc trưng trung bình của các dãy số L_i, T_i, R_i, V_i và $LTR = (L_i \cup T_i \cup R_i)$. Số trong dấu ngoặc là số lượng các dãy số cùng loại.

Phần cuối của bảng ghi tỷ lệ so sánh các trị số trung bình của các dãy số L_i, T_i, R_i và LTR với các dãy số tựa ngẫu nhiên V_i và cho ta thấy:

χ^2 giảm từ 19,7 (dãy L) đến 27,6 lần (dãy T), trung bình 24,2 lần,
D_N giảm từ 7,4 (dãy T) đến 9,0 lần (dãy L), trung bình 8,3 lần,
ω^2 giảm từ 254,9 (dãy T) đến 367,1 lần (dãy L), trung bình 316,0 lần,
D giảm từ 7,6 (dãy T) đến 9,1 lần (dãy L), trung bình 8,4 lần.

3.2. So sánh moment mẫu với giá trị lý thuyết

Như đã nói trong Mục 1.3., bên cạnh các phép thử về hàm phân phối (tiêu chuẩn phân bố đều loại 1) người ta còn so sánh các moment mẫu với các giá trị lý thuyết tương ứng (tiêu chuẩn phân bố đều loại 2). Các đặc trưng thống kê sau đây của phân phối đều liên tục sẽ

Bảng 1. Đặc trưng thống kê của từng dãy số
(ba dòng cuối ghi các giá trị tới hạn ở mức ý nghĩa 5%, 1% và 0,1%,
tương ứng với xác xuất tin cậy 95%, 99% và 99,9%)

Dãy số	χ^2	$K_N = D_N\sqrt{N}$	$100.\omega^2$	Độ phân kỳ D	E_1 (sai số % moment bậc 1)	E_2 (sai số hệ số lệch và độ nhọn
L_1	54,876	1,200	32,282	60,526	1,551	2,978
L2	33,794	0,789	12,692	39,120	0,756	2,489
L3	60,195	0,749	5,827	38,142	0,369	1,856
L4	55,266	0,860	12,509	42,637	0,991	0,370
T1	**164,920**	0,738	7,453	36,580	0,960	0,498
T2	42,773	0,667	10,910	33,098	1,037	1,860
T3	45,360	0,758	11,682	37,601	1,377	2,467
T4	54,778	0,757	9,435	38,535	0,304	1,754
T5	57,120	0,631	5,280	32,283	0,419	1,612
T6	63,952	0,892	21,169	45,249	1,148	2,083
R1	66,783	1,239	33,462	62,460	4,173	2,086
R2	66,880	0,779	12,086	38,617	1,059	3,090
R3	62,879	1,015	19,258	50,326	0,896	3,875
R4	49,703	0,534	3,201	27,481	0,337	0,401
R5	60,341	0,706	7,122	34,986	1,070	0,744
V_1	0,024	0,010	0,003	0,500	0,000	0,000
V_2	5,929	0,151	0,081	7,500	0,000	0,036
V_3	3,098	0,070	0,039	3,454	0,001	0,012
V_4	1,976	0,040	0,017	1,988	0,000	0,001
V_5	1,488	0,050	0,024	2,500	0,000	0,002
V_6	4,221	0,151	0,058	7,500	0,006	0,049
V_7	1,683	0,091	0,032	4,500	0,004	0,028
V_8	2,318	0,232	0,090	11,500	0,004	0,090
Mức (1- õ)						
5%	77,93	1,358	46			
1%	87,17	1,624	74			
0,1%	98,32	1,950	117			

được dùng để so sánh:

Kỳ vọng toán học: $\mu = 1/2 = 0{,}5$.

Độ lệch chuẩn: $\sigma = 1/2\sqrt{3} = 0{,}28868$.

Độ lệch tuyệt đối trung bình: $\int abs(x-\mu).f(x).dx = 1/4 = 0{,}25$.

Hệ số lệch: $\mu_3/\sigma_3 = 0$.

Độ nhọn: $\mu_4/\sigma_4 = 9/5 = 1{,}8$.

Chúng tôi đã tính các đặc trưng trên đây của từng dãy số. Việc so sánh định lượng căn cứ vào 2 chỉ số: E_1 là tổng giá trị tuyệt đối sai số phần trăm của 3 moment bậc 1 (trung bình, độ lệch chuẩn, độ lệch tuyệt đối trung bình); E_2 là tổng giá trị tuyệt đối sai số của hệ số lệch và độ nhọn (nhân với 100). Theo dòng cuối của Bảng 2 thì khi dùng các dãy số tựa

ngẫu nhiên V_i, sai số E_1 giảm đi 585,1 lần và sai số E_2 giảm đi 68,6 lần.

Giữa các số ngẫu nhiên và giả ngẫu nhiên sự khác biệt là không đáng kể, tuy rằng các dãy số T_i (được tạo thành từ các bảng số ngẫu nhiên) có các đặc trưng tốt hơn giá trị trung bình của LTR.

Bảng 2. Các đặc trưng trung bình của các nhóm dãy số

Ký hiệu (số lượng các dãy số	χ^2	KN = $D_N\sqrt{N}$	$100\omega^2$	Độ phân kỳ D	E1 (sai số % moment bậc 1	E2 (sai sốhệ số lệch và độ nhọn
L(4)	51,0326	0,8995	15,8274	45,1064	0,9169	1,9231
T(6)	71,4840	0,7406	10,9883	37,2244	0,8742	1,7122
R(5)	61,3172	0,8545	15,0257	42,7738	1,5071	2,0390
LTR(15)	62,6413	0,8209	13,6246	41,1761	1,0965	1,8774
V(8)	2,5921	0,09943	0,04311	4,9303	0,001874	0,02738
(Tỉ số)						
L/V	19,7	9,0	367,1	9,1	489,3	70,2
T/V	27,6	7,4	254,9	7,6	466,5	62,5
R/V	23,7	8,6	348,5	8,7	804,2	74,5
LTR/V	24,2	8,3	316,0	8,4	585,1	68,6

3.3. So sánh các cực trị của các dãy số

Các trị số nhỏ nhất và lớn nhất của 6 đại lượng đặc trưng cho các dãy số V_i và các dãy số $LTR = (L_i \cup T_i \cup R_i)$ còn lại được trình bày trong Bảng 3. Ta thấy trong mọi trường hợp các giá trị kém nhất của V_i (tức là V-max) vẫn tốt hơn (nhỏ hơn) các giá trị khá nhất của LTR (tức là LTR-min). Tỷ lệ nhỏ nhất giữa 2 đại lượng này là $LTR\text{-min}/V\text{-max} = 2{,}30$ (đối với $K_N = D_N.\sqrt{N}$, giữa R_4 và V_8).

Còn tỷ lệ giữa LTR-max và V-min được ghi trong hàng cuối cùng. Giá trị lớn nhất xấp xỉ 1 triệu 391 nghìn lần (đối với E_1) là tỷ số giữa R_1 và V_1.

3.4. Tương quan giữa các chỉ số đặc trưng và số đoạn hợp thành

Hai dòng 4 và 5 của Bảng 3 cho thấy, trong số các dãy tựa ngẫu nhiên thì V_1 (gồm 1 đoạn duy nhất) có các chỉ số tốt nhất và V_8 (do 24 đoạn hợp thành) có các chỉ số kém nhất. Tất nhiên số đoạn hợp thành (h) càng tăng thì tính đều càng giảm, tức là độ phân kỳ và các sai số càng tăng. Câu hỏi đặt ra là các chỉ số tăng tỷ lệ với h hay hàm nào của h?

Chúng tôi tính các hệ số tương quan giữa các chỉ số đặc trưng với 4 biến số $h, h^2, \sqrt{h}, \ln(h)$ rồi chọn ra trị số tuyệt đối R lớn nhất của mỗi chỉ số.

Các chỉ số ω^2, D_N, độ phân kỳ D và sai số E_2 tương quan chặt chẽ với h:

$R(\omega^2, h) = 0{,}99455,$

$R(D, h) = 0{,}92776,$

$R(D_N, h) = 0{,}92776,$

$R(E_2, h) = 0{,}83142.$

Trị số χ^2 tăng theo $\sqrt{h}$, ở mức độ thấp hơn

$$R(\chi^2\sqrt{h}) = 0{,}75444.$$

Chỉ riêng sai số phần trăm moment bậc nhất (E_1) tương quan khá yếu với cả 4 biến số:
$R(E_1, \ln(h))$= 0,43285,
$R(E_1, \sqrt{h})$ = 0,41233,
$R(E_1, h)$ = 0,35487,
$R(E_1, h^2)$ = 0,23346.

Điều này dễ giải thích: ý tưởng cơ bản của thuật toán [12] là hai moment bậc 1 và bậc 2 (M_1 và M_2) của dãy số bằng đúng moment tương ứng của phân phối đều liên tục. Như vậy, bất kể h bằng bao nhiêu, nếu các đoạn hợp thành K_i (của mỗi dãy V_i) độc lập với nhau thì sai số moment bậc 1 phải bằng 0. Thực tế trong cột E_1 Bảng 1 các sai số của V_i đều rất nhỏ, không vượt quá 0,006%.

Bảng 3. Các cực trị và tỷ số

Dãy số	χ^2	$K_N = D_N\sqrt{N}$	$100.\omega^2$	Độ phân kỳ D	E_1 (sai số % moment) bậc 1	E_2 (sai số hệ số lệch và độ nhọn
LTR-min	33,793819 (L_2)	0,534012 (R_4)	3,200899 (R_4)	27,480767 (R_4)	0,304425 (T_4)	0,369813 (L_4)
LTR-max	164,92029 (T_1)	1,239398 (R_1)	33,461951 (R_1)	62,459663 (R_1)	4,172557 (R_1)	3,874531 (R_3)
V-min	0,023993 (V_1)	0,010085 (V_1)	0,003389 (V_1)	0,500102 (V_1)	0,000003 (V_1)	0,000020 (V_1)
V-max	5,928833 (V_2)	0,231910 (V_8)	0,090391 (V_8)	11,500000 (V_8)	0,006497 (V_6)	0,090418 (V_8)
(Tỉ số)						
LTR-min /V-max	5,70	**2,30**	35,41	2,39	46,86	4,09
LTR-max /V-min	6873,7	122,9	9873,7	124,9	**1390852**	193726,6

4. KẾT LUẬN

Các phép thử thống kê trên 23 dãy số ngẫu nhiên, giả ngẫu nhiên và tựa ngẫu nhiên cho thấy hầu như tất cả các dãy số đó đều thỏa mãn các tiêu chuẩn về tính đều. Như vậy có thể sử dụng hữu hiệu chúng trong các tính toán theo sơ đồ Monte Carlo. Chỉ có một trị số χ^2 của một dãy số vượt quá giá trị tới hạn. Đây là dãy số được tạo nên từ các hoán vị 10-chữ số chứ không phải từ các chữ số ngẫu nhiên.

Cả 8 dãy số tựa ngẫu nhiên V_i, được tạo thành bằng các cách khác nhau (tùy ý, truy toán, nguyên tố), đều có các độ đo về tính đều (χ^2, Kolmogorov $K_N = D_N.\sqrt{N}$, ω^2, độ phân kỳ D, sai số phần trăm moment bậc 1, sai số hệ số lệch và độ nhọn) tốt hơn hẳn các dãy số còn lại. Giữa các số ngẫu nhiên và giả ngẫu nhiên sự khác biệt là không đáng kể.

TÀI LIỆU THAM KHẢO

[1] N. P. Buslenko, D. I. Golenko, I. M. Sobol, V. G. Sragovich, Yu. A. Shreider, *Methods of Statistical Testing. (Monte Carlo method).* Elsevier Pub. Co., Amsterdam, 1964; *The*

Monte Carlo Method, Yu. A. Shreider (ed.), Pergamon Press, Oxford, 1966.

[2] D. J. Cowden, M. S. Cowden, *Practical Problems in Business Statistics*, Prentice-Hall, New York, 1948.

[3] Frey Tamás, *On the Information-Theorical Estimation of the Operation Consumption of Optimal Algorithms*, Colloquia Mathematica Societatis János Bolyai. 3. Numerical methods. Tihany (Hungary) 1968, 49–60.

[4] V. Giard, *Statistique Apliquée à la Gestion*, Economica, Paris, 1992.

[5] J. Maurin, *Simulation Déterministe du Hazard*, Masson, Paris, 1975.

[6] P. H. Muller, P. Neumann, R. Storm, *Tafeln der Mathematischen Statistik*, VEB Fachbuchverlag, Leipzig, 1979.

[7] RAND Corporation, *A Million Random Digits with 100 000 Normal Deviates*, The Free Press, Macmillan, Glencoe, Illinois, 1955.

[8] C. R. Rao, S. K. Mitra, A. Matthai, *Formulae and Tables for Statistical Work*, Statistical Publishing Society, Calcutta, 1966.

[9] I. M. Sobol, *Các phương pháp Monte Carlo số trị*, Nhà xuất bản Nauka, Moskva, 1973. (tiếng Nga).

[10] S. Tezuka, *Uniform Random Numbers: Theory and Practice*, Kluwer Academic Publishers, Boston/Dordrecht/London, 1995.

[11] Vincze István, *Matematikai Statisztika Ipari Alkalmazásokkal*, Muszaki Konyvkiadó, Budapest, 1975.

[12] Vũ Hoài Chương, Một thuật toán đơn giản tạo dãy số tựa ngẫu nhiên, *Tạp chí Khoa học và Công nghệ* **40** (số ĐB) (2002) 94–99.

Nhận bài ngày 2 - 12 - 2004
Nhận lại sau sửa ngày 11 - 5 - 2005