Bull. of Yamagata Univ., Nat. Sci., Vol.16, No.3, Feb. 2007

M系列に対する重みディスクレパンシー検定

松本 眞*, 西村 拓士†

(2006年5月31日受理)

Abstract

In this paper, we analyze M-sequences based on trinomials and pentanomials of degree from 607 to 9689 with the weight discrepancy test proposed by Matsumoto and Nishimura.

An M-sequence based on a trinomial of degree 9689 will be rejected in average in the weight distribution test, if we consume 1.57×10^{10} samples. To reject an M-sequence based on a pentanomial of degree 9689 in the test, we need to consume 8.55×10^{16} samples in average.

1 はじめに

本論文では, M系列の重み分布検定における性質を松本と西村 [12] により提案された重みディスクレパンシー検定を用いて調べる.

重み分布検定は, 擬似乱数の出力の0と1の割合が, 理想的な分布である2項分布とみなせるかどうかを判定する統計的検定である. 重みディスクレパンシー検定とは, 有限体GF(2)上の線形漸化式に基づく擬似乱数生成法の重み分布検定における能力を測る理論的な検定である. 重みディスクレパンシー検定を用いると, 擬似乱数に対し重み分布のカイ2乗検定を行う際に, どの位のサンプル数までを用いれば検定において棄却されないか, また, どの位のサンプル数を用いると棄却されるかを計算出来る.

論文 [12] では, M系列に関しては次数が89, 218, 521, 1279であるものに対して重みディスクレパンシー検定が行われている. 本論文では、次数が607次から9689次までの3項式および5項式によるM系列に対し重みディスク

*Department of Mathematics, Faculty of Science, Hiroshima University, Hiroshima, 739–8526, Japan (e-mail address: m-mat@math.sci.hiroshima-u.ac.jp)

†Department of Mathematical Sciences, Yamagata University, Yamagata, 990–8560, Japan (e-mail address: nisimura@sci.kj.yamagata-u.ac.jp)

レパンシー検定を行う．3項式および5項式のM系列について調べる理由は，それらが生成速度が速い事などから実用上よく利用されているためである．

2　M系列

LFSR(Linear Feedback Shift Register) とは次の線形漸化式によって0と1の数列 $\{x_i \,|\, i = 1, 2, \ldots\}$ を生成する手法である．

$$x_i = a_1 x_{i-1} + a_2 x_{i-2} + \cdots + a_n x_{i-n} \pmod 2 \tag{1}$$

式(1)に対応するGF(2)上の多項式 $f(t) = t^n + \sum_{i=1}^{n} a_i t^{n-i}$ が原始的である時，数列 $\{x_i\}$ の周期は最大の $T = 2^n - 1$ となり，$\{x_i\}$ はM系列と呼ばれる．さらに n をM系列の次数と呼ぶ．式(1)において，$\#\{i \,|\, a_i \neq 0, 1 \leq i \leq n\} = 2$ である時，3項式によるM系列と呼び，$\#\{i \,|\, a_i \neq 0, 1 \leq i \leq n\} = 4$ である時，5項式によるM系列と呼ぶ事にする．

3項式によるM系列の重みの分布に関しては [8], [6], [3], [10], [11] において分析されている．3項式によるM系列は，イジングモデル検定 [5], [2], [1] やランダムウォーク検定 [4], [15] において偏った結果を示す事が報告されている．

3　カイ2乗ディスクレパンシー

まず，カイ2乗検定についてまとめる．$Z = \{0, 1, 2, \ldots, \nu\}$ を確率事象の集合とする．$\{p_k \,|\, k = 0, 1, \ldots, \nu\}$ を Z に関する確率分布とする．すなわち，$0 \leq p_k \leq 1, \sum_{k=0}^{\nu} p_k = 1$．$N$ 個のデータ $x_1, x_2, \ldots, x_N \in Z$ 与えられたとして，これらが確率分布 $\{p_k\}$ を持つ集合 Z からのランダムなサンプルであるかとみなせるかどうかを検定する．つまり，帰無仮説 H_0:“$x_1, x_2, \ldots, x_N$ は Z からのランダムなサンプルである” を検証する．まず，次の様に定義された χ^2 値を計算する

$$\chi^2 = \sum_{k=0}^{\nu} \frac{(Y_k - N p_k)^2}{N p_k} \tag{2}$$

ここで Y_k は $Y_k = \#\{i \mid x_i = k\}$ である．帰無仮説 H_0 の元で χ^2 は自由度 ν のカイ2乗分布に近似的に従う事が知られている．次にp値と呼ばれる以下の値 p を計算する

$$p = \mathrm{Prob}(X < \chi^2).$$

X は自由度 ν の χ^2 分布に従う確率変数である．このp値が大き過ぎる場合，例えば0.99より大きい場合は，帰無仮説 H_0 を有意水準0.01で棄却する．そうでなければ帰無仮説 H_0 を有意水準0.01で採択する事にする．以上がカイ2乗検定の手順である．

$\{q_k \mid k = 0, 1, \ldots, \nu\}$ をもう 1 つの確率分布とする. ここで, 2 つの確率分布 $\{p_k\}$ と $\{q_k\}$ に関するカイ 2 乗ディスクレパンシー δ を次の様に定義する.

$$\delta = \sum_{k=0}^{\nu} \frac{(q_k - p_k)^2}{p_k}$$

この δ は 2 つの確率分布 $\{p_k\}$ と $\{q_k\}$ のずれを表す指標である.

ここで, 帰無仮説 H_0 が成り立たない, すなわち N 個のデータ "$x_1, x_2, \ldots, x_N$" は $\{p_k\}$ とは異なる確率分布 $\{q_k\}$ からのランダムなサンプルであるとしたらどうなるであろうか. このとき式 (2) の χ^2 は非中心的パラメータが δ の自由度 ν の非中心的 χ^2 分布に近似的に従う事が知られている [13]. そして, この場合の式 (2) の χ^2 の期待値 $E(\chi^2)$ は

$$E(\chi^2) \sim \nu + N\delta \tag{3}$$

で近似出来る [12].

ここで, ν とカイ 2 乗ディスクレパンシー δ が与えられた時に, 有意水準 p のサンプル数 N を次の式で定義する.

$$\mathrm{Prob}(X < \nu + N\delta) = p$$

式中の X は自由度 ν の χ^2 分布に従う確率変数である.

この様に定義された有意水準 p のサンプル数 N に対し χ^2 検定を行えば, 観測される χ^2 値に対応する p 値は平均して p になる事が式 (3) より分かる. 有意水準が $p = 0.99$ の時, 対応するサンプル数を危険なサンプル数と呼び, 有意水準が $p = 0.75$ の時, 対応するサンプル数を安全なサンプル数と呼ぶ事にする.

なお, 与えられた統計的検定において, 擬似乱数生成法が棄却されてしまうようなサンプル数の近似式を求めるという考えが [7] においても提案されている.

4 重みディスクレパンシー検定

重みディスクレパンシー検定とは, 前節のカイ 2 乗ディスクレパンシーを重み分布の検定に応用したものである. 本節では 重みディスクレパンシー検定について簡単にまとめる. 詳細は論文 [12] を参照の事.

$\{x_i \mid i = 1, 2, \ldots\}$ を 0 と 1 からなる数列とする. 重み分布検定とは $\{x_i\}$ の連続する m 個からなるベクトル $\{(x_i, x_{i+1}, \ldots, x_{i+m-1}) \mid i = 1, 2, \ldots\}$ の集合のハミング重みの分布が, 理想の分布の 2 項分布とみなせるかどうかを調べる統計的検定である.

式 (1) により生成される数列 $\{x_i\}$ の連続する m 個の値によって構成されるベクトルの全体の集合とこれにゼロベクトルを加えたものを C とする．すなわち

$$C = \{(x_i, x_{i+1}, \ldots, x_{i+m-1}) \mid i = 1, 2, \ldots, T\} \cup \{(0, 0, \ldots, 0)\}$$

である．ここで T は $\{x_i\}$ の周期である．n を生成式の次数とすると，定義から C は次元 n の GF(2) 上の線形空間である．A_i を C の中のハミング重みが i のものの個数とする．カイ 2 乗ディスクレパンシー δ は

$$q_k = A_k/2^n, \quad p_k = \left(\begin{array}{c} m \\ k \end{array} \right) /2^m \quad (k = 0, 1, \ldots, m)$$

とおいて計算すれば良い．なお，m は $m > n$ となるように選ぶ．$m \leq n$ の時は M 系列の周期の最大性から $\{q_k\}$ は 2 項分布と一致する．また，この様に定義された $\{q_k\}$ が確率分布となるためには，式 (1) の初期値をランダムに選ぶという仮定が厳密に言えば必要である．

実際に $\{q_k\}$ を求めるためには C の重み分布を調べる必要がある．しかし，一般に線形空間の重み分布を調べるのは，NP 完全問題である事が知られており困難である [14]．そこで符号理論の MacWilliams の恒等式 [9] を利用する．MacWilliams の恒等式によって C の直交補空間 $C^{\perp}$ の重み分布から C の重み分布を計算する事が出来る．$C^{\perp}$ の重み分布を数え上げるのは一般的にはやはり困難であるが，$C^{\perp}$ の次元 $m - n$ が小さければ，$C^{\perp}$ の全ての元の重みを全数チェックにより計算する事が可能である．

なお，実際のカイ 2 乗ディスクレパンシー δ の計算においては，いくつかの重みを $\nu + 1$ 個のグループ $\{S_k \mid k = 0, \ldots, \nu\}$ に分けて，すなわち

$$q_k = \sum_{i \in S_k} A_i/2^n, \quad p_k = \sum_{i \in S_k} \left(\begin{array}{c} m \\ i \end{array} \right) /2^m \quad (k = 0, 1, \ldots, \nu)$$

として実験を行った．ここで $\cup_{k=0}^{\nu} S_k = \{0, 1, \ldots, \nu\}$, $S_i \cap S_j = \emptyset$ $(i \neq j)$ である．特に，各 p_k の値が大体同じになるように各 S_k を定めた．

重みディスクレパンシー検定の手順をまとめると次の様になる．まず，与えられた擬似乱数の出力から得られる線形空間 C の重み分布を求め $\{q_k\}$ を計算し，次にカイ 2 乗ディスクレパンシー δ を計算し，そして δ を用いて安全なサンプル数と危険なサンプル数を求める．

5 実験結果

表 1 および表 2 に，それぞれ 3 項式と 5 項式による M 系列に対する重みディスクレパンシー検定の結果を示す．表中の生成法 G(s_1, s_2) に対応する

生成式は $x_i = x_{i-s_1} + x_{i-s_2} (\mathrm{mod}\ 2)$ であり, G(s_1, s_2, s_3, s_4) に対応する生成式は $x_i = x_{i-s_1} + x_{i-s_2} + x_{i-s_3} + x_{i-s_4} (\mathrm{mod}\ 2)$ である. 全ての場合において, 前節の m の値として “生成式の次数”+20 となる様に選んだ. つまり直交補空間の次元が 20 になるように m を選んだ.

表 3 に表 1 と表 2 の結果を計算する時に使った, 各 m に対する重みのグループ分けを示す. 実際の実験では重みを 10 のグループ分けた. 重みのグループを $S_k (0 \leq k \leq 9)$ とすると, $S_k = [t_{k-1} + 1, t_k]$ である (ただし, 全ての m に対し $t_{-1} = -1$ とする) . 表 3 には t_k のみ表示しておく.

生成式の次数が大きくなるに従って, 安全なサンプル数および危険なサンプル数が大きくなっている事が分かる. また, 3 項式による生成法よりも, 5 項式による生成法の方がより良好な結果を示している事もわかる.

G(471, 9689) の生成速度を測ると 10^8 個の乱数を生成するのに 0.47 秒かかった. 計測環境は CPU が Pentium4 2.6Ghz のパーソナルコンピュータで, OS が Vine Linux 2.6r4 である. この生成速度から計算すると G(471, 9689) の危険なサンプル数を得るのにはわずか約 8.3 日で良い事が分かる. この G(471, 9689) の危険なサンプル数は本格的なシミュレーションを行うには小さ過ぎるため, G(471, 9689) は使用すべきではない.

G(471, 1586, 6988, 9689) の生成速度を測ると 10^8 個の乱数を生成するのに 0.93 秒かかった. この生成速度から計算すると G(471, 1586, 6988, 9689) の安全なサンプル数を得るのには約 46097 年かかる事が分かる. G(471, 1586, 6988, 9689) の安全なサンプル数は十分大きく, 重み分布の検定の観点からは, この生成法は現在 (2006 年) のパーソナルコンピュータでシミュレーションを行う限り実用上問題がないと言える. G(471, 1586, 6988, 9689) は Ziff [15] により推奨されている生成法である.

表 4 には, 表 1 で得られた結果をもとに実際に G(105, 607) にカイ 2 乗検定を行った結果を示す. 検定は異なる 5 つの初期値に対して行い, 表には p 値を示してある. サンプル数がほぼ安全なサンプル数である場合は, 得られた p 値はばらばらで, それらの平均は 0.6177 である. サンプル数がほぼ危険なサンプル数である場合は, 5 回の検定中 4 回の場合において p 値が 0.99 台と高い値を示している.

表 1: 3 項式による M 系列に対する重みディスクレパンシー検定の結果

生成法	m	安全なサンプル数	危険なサンプル数
G(105, 607)	627	7.91×10^5	4.19×10^6
G(216, 1279)	1299	7.07×10^6	3.75×10^7
G(715, 2281)	2301	3.84×10^7	2.04×10^8
G(67, 3217)	3237	1.12×10^8	5.92×10^8
G(271, 4423)	4443	2.78×10^8	1.47×10^9
G(471, 9689)	9709	2.96×10^9	1.57×10^{10}

表 2: 5 項式による M 系列に対する重みディスクレパンシー検定の結果

生成法	m	安全なサンプル数	危険なサンプル数
G(35, 70, 105, 607)	627	1.77×10^{10}	9.39×10^{10}
G(72, 144, 216, 1279)	1299	6.86×10^{11}	3.64×10^{12}
G(715, 1237, 1759, 2281)	2301	1.19×10^{13}	6.29×10^{13}
G(67, 1117, 2167, 3217)	3237	6.69×10^{13}	3.54×10^{14}
G(271, 1655, 3039, 4423)	4443	3.20×10^{14}	1.70×10^{15}
G(471, 1586, 6988, 9689)	9709	1.61×10^{16}	8.55×10^{16}

表 3: 重みのグループ分け

m	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
627	297	302	306	310	313	316	320	324	329	627
1299	626	634	640	644	649	654	658	664	672	1299
2301	1119	1130	1138	1144	1150	1156	1162	1170	1181	2301
3237	1582	1594	1603	1611	1618	1625	1633	1642	1654	3237
4443	2178	2193	2204	2213	2221	2229	2238	2249	2264	4443
9709	4791	4813	4828	4842	4854	4866	4880	4895	4917	9709

表 4: G(105, 607) に対する重み分布の検定

サンプル数	1 回目	2 回目	3 回目	4 回目	5 回目
7.9×10^5	0.8693	0.0009	0.5849	0.9983	0.6352
4.2×10^6	0.9605	0.9991	0.9999	0.9922	0.9982

References

[1] P. D. Coddington, Analysis of random number generators using Monte Carlo simulation, Int. J. Mod. Phys. C **5** (1994), 547–560.

[2] A. M. Ferrenberg, D. P. Landau, and Y. J. Wong, Monte Carlo simulations: hidden errors from 'good' random number generators, Phys. Rev. Lett. **69** (1992), 3382–3384.

[3] S. A. Fredricsson, Pseudo-randomness properties of binary shift register sequences, IEEE Trans. Inform. Theory **IT-21** (1975), 115–120.

[4] P. Grassberger, On correlations in 'good' random number generators, Phys. Lett. A **181** (1993), 43–46.

[5] A. Hoogland, J. Spaa, B. Selman, and A. Compagner, A special-purpose processor for the Monte Carlo simulation of Ising spin systems, J. Comput. Phys. **51** (1983), 250–260.

[6] H. F. Jordan and D. C. M. Wood, On the distribution of sums of successive bits of shift-register sequences, IEEE Trans. Computers **C-22** (1973), 400–408.

[7] P. L'Ecuyer and P. Hellekalek, Random number generators: selection criteria and testing, In P. Hellekalek and G. Larcher, editors, Random and Quasi-Random Point Sets, Lecture Notes in Statitics, vol. 138, Springer, New York, (1998), 223-265.

[8] J. H. Lindholm, An analysis of the pseudo-randomness properties of subsequences of long m-sequences, IEEE Trans. Inform. Theory **IT-14** (1968), 569–576.

[9] F. J. MacWilliams and N. J. A. Sloane, The Theory of Error-Correcting Codes, North-Holland, (1977).

[10] M. Matsumoto and Y. Kurita, Twisted GFSR generators II, ACM Trans. on Modeling and Computer Simulation **4** (1994), 254–266.

[11] M. Matsumoto and Y. Kurita, Strong deviations from randomness in m-sequences based on trinomials, ACM Trans. on Modeling and Computer Simulation **6** (1996), 99–106.

[12] M. Matsumoto and T. Nishimura, A nonempirical test on the weight of pseudorandom number generators, In K. T. Fang, F. J. Hickernel, and

H. Niederreiter, editors, Monte Carlo and Quasi-Monte Carlo Methods 2000, Springer, (2002), 381–395.

[13] M. Tiku, Noncentral chi-square distribution, In S. Kotz and N. L. Johnson, editors, Encyclopedia of Statistical Sciences, vol. 6, John Wiley, (1981), 276–280.

[14] A. Vardy, The intractability of computing the minimum distance of a code, IEEE Trans. Inform. Theory **IT-43** (1997), 1757–1766.

[15] R. M. Ziff, Four-tap shift-register-sequence random-number generators, Computers in Physics **12** (1998), 385–392.