数理解析研究所講究録
第85巻 1993年 96-108

擬似乱数に対するランダムウォークによる統計的検定

大阪教育大学　高嶋恵三 (Keizo Takashima)

計算機の普及と高速化に伴い擬似乱数を使用しての数値計算やシミュレーションなども広く行なわれるようになってきた．このため，擬似乱数に対する理論的研究と共に実際の応用の観点から，統計的検定も重要になっている．本報告では離散型確率過程の基本的モデルであるランダムウォークの汎関数のシミュレーションによる検定を取り上げ，その結果と関連する諸問題について議論する．

1．　擬似乱数

擬似乱数について詳しくは，例えば Knuth[4]，伏見[2]などを参照されたい．ここでは，本報告で引用する擬似乱数生成法について簡単に述べるに止める．

1）線型合同法

$a > 0$, b, $M > 0$ を整数とし

$$x_n = (a x_{n-1} + b) \bmod M, \quad n > 0,$$

によって擬似乱数列 x_n を定める．定数 a, b, M の選び方や性質については[4],[2]を参照されたい．ここでは x_n の周期は M を越えられない（通常 M は計算機で取り扱える最大の整数または最大の素数を使用することが多いので，例えば３２ビットのパソコンなどでは 2^{32} を越えられない）ことと粗結晶構造を持つことを注意するに止める．本報告では主に，$b = 0$ の場合を考え，$M = 2^{32}$ ととることにする．a の候補としては 69069, 1664525, 1812433253, 1566083941 を考える．a がこれら値である場合，乗算合同法（$b = 0$ の場合，線型合同法を特にこう呼ぶ）はスペクトル検定によい結果が得られることが知られている．詳しくは Knuth[4] を参照されたい．

２）M系列

この方法には色々な呼び方があるがここではM系列で統一することにする．以下では簡単にM系列について述べるに止め詳しくは，伏見[2], Golomb[3] などを参照されたい．

GF(2)を 0, 1 からなる体（Galois 体と呼ばれる）とし，f(x) を GF(2) 上の原始多項式とする．一般に生成速度を重

視する場合，f(x) を３項式にとることが多い．例えば

$$f(x) = x^p + x^q + 1$$

の場合，

$$a_n = (a_{n-p} + a_{n-q}) \mod 2 \quad (n > p)$$

によって，GF(x) の元の列 a_n を定める．このような列は最大の周期 $2^p - 1$ を持つことが知られており，最大周期列（Ｍ－系列）と呼ばれる．一般に a_n の初期値 a_1，．．．，a_p は任意に与えられるが，特別な初期値を用いると擬似乱数としての性質が悪い状態が長く続くのでいろいろな初期化が考案されている．また，実際に計算機（例えば３２ビットマシン）の上でこの擬似乱数を生成するには，a_n を用いて３２ビット整数を生成するがその方法もいろいろ工夫されている．これらについては伏見[2]を参照されたい．

３）加算生成法

Ｍ系列の場合と同様に f(x) を GF(2) 上の原始多項式とする．生成速度を重視する場合，３項式にとる場合が多い．例えば f(x) が上の（1）で与えられる場合

$$X_n = (X_{n-p} + X_{n-q}) \mod 2^m \quad (n > p)$$

によって m ビット整数 X_n を定める．ここでＭ系列との違いは，Ｍ系列の場合は足算が GF(2) で行なわれるので，繰

り上がりがないのに対して，加算生成法では繰り上がりがある点に違いがある．このため，M系列の議論は GF(2) およびその拡大体の理論に依るが加算生成法の議論は体 GF(2) でなく Galois ring $GR(2^m)$ の理論に依ることになる．このため，M系列に比べ，議論が格段に難しくなるため，代数的議論が今まで余りなされていないようである．また x_n の最下位ビットはM系列であり，x_n の周期は $2^l(2^p-1)$，$l<m$ である．詳しくは Knuth[4] を参照されたい．また，Galois Ring については McDonald[8] を参照されたい．

２．ランダムウォークの汎関数

本報告では Feller[1] の第３章にあるランダムウォークのパス（道）の関数（汎関数と呼ぶことにする）について簡単に述べることにする．まず，ランダムウォークの定義を述べる．

X_n，$n=1,2,...$ を独立同分布な確率変数の列とし，

$P(X_n=1)=P(X_n=-1)=1/2$，$n=1,2,...$

とする時，ランダムウォーク S_n は以下で定義される：

$$S_n=\Sigma_{k=1}^{n} X_k,\quad n=1,2,...,\quad S_0=0.$$

n を時間パラメーターと見ることにし，$(S_0, S_1, ... S_n)$ を長さ n のパス（道）と呼ぶことにする．

1) sojourn time

長さ 2L のパスを考え，S_k の値が非負である時間の総和を時間 2L までの sojourn time T_{2L} と呼ぶことにする：

$$T_{2L} = \Sigma_{k=1}^{2L} I(S_k + S_{k-1}) ,$$

但し，$I(x) = 1$ ($x > 0$) , $= 0$ (それ以外) .

sojourn time の分布については以下の定理が有名である．

定理 (Discrete arc sine law for sojourn time)

$$P(T_{2L} = 2k) = (2k)!(2L-2k)!/(k!k!(L-k)!(L-k)!)2^{-2L} = \alpha_{2k,2L}$$

$k = 0, 1, . . . , L .$

$T = T_{2L} / 2L$ と置き，$2L \to \infty$ とすると，T の分布は漸近的にパラメーター 1/2, 1/2 のベータ分布に近付くことが知られている．その密度関数は

$$\pi^{-1} x^{-1/2} (1 - x)^{-1/2} , \quad 0 < x < 1 ,$$

で与えられる．

2) last visit time

$$V_{2L} = \max\{ k ; S_k = 0 , k = 1, . . . , 2L \}$$

を長さ 2L のパスの原点への last visit time と呼ぶことにする．last visit time に対しても arc sine law が知られている：

定理 (Discrete arc sine law for last visit time)

$$P(V2L = 2k) = \alpha_{2k,2L} , \quad k = 0, 1, \ldots, L .$$

3) change of sign

長さ 2L + 1 のパスで横軸（時間 n を横軸にとる）を横切った回数 C_{2L+1} を考える：

$$C_{2L+1} = \sum_{k=1}^{L} I(- S_{2k-1} S_{2k+1}) .$$

定理 $P(C2_{L+1} = r) = (2L+1)!/((2r+1)!(2L-2r)!)\ 2^{-2L}$

$$= \xi_{r,2L+1} , \quad r \leqq L .$$

注意：$\xi_{r,2L+1}$ は r に関して単調に減少する：

$$\xi_{0,2L+1} \geqq \xi_{1,2L+1} \geqq \xi_{2,2L+1} \geqq \ldots .$$

特に，$\xi_{L,2L+1} = 2^{-2L}$ であり，L が大きくなると共に非常に小さくなることに注意する．

その他にも，first passage time F_r

$$F_r = \inf\{ k : k > 0 , S_k = r \} ,$$

などがあるが，期待値が∞であることからも推測できるように，分布の裾が長いのでシミュレーションに必要なサンプル数が大きすぎるので本報告では省略することにする．またこの汎関数に関してはパスの長さに制限がないことにも注意が必要であろう．

３．シミュレーションとその結果

前節の汎関数のシミュレーションによる統計的検定の考え方は非常に簡単である：

１）まず，擬似乱数の例えば最上位ビットを取り出すことにより，長さ 2L のランダムウォークのパスを N 本作り，各パスに対して汎関数の値を求める．これらの N 個の値の相対頻度を求め，その相対頻度の，理論分布に対する適合度をカイ２乗検定で調べる．

２）これを M 回繰返し，M 個のカイ２乗検定値のうち，カイ２乗分布の例えば，９０％点から９５％点までに入る頻度と，９５％以上に入る頻度を求め，検定を行なう．

以下，表１から表３に検定の結果をまとめた．

表 1

sojourn time 検定				
2L	N	M	A	B
M系列， p = 31 , q = 13 , 原始3項式				
60	10,000	100	14%	34%
100	10,000	100	0%	100%
加算生成法， p = 31 , q = 13 , 原始3項式				
60	10,000	2000	4.5%	5.75%
乗算合同法， a = 69069 , b = 0 , M = 2^{32}				
60	10,000	100	2%	5%
100	10,000	100	7%	8%

A 欄は90%点から95%点， B 欄は95%以上に落ちた割合

表 2

last visit time 検定				
M系列， p = 31 , q = 13 , 原始3項式				
60	20,000	200	7%	11.5%
100	10,000	200	6.5%	10.5%
加算生成法， p = 31 , q = 13 , 原始3項式				
600	50,000	300	6.3%	4.7%
乗算合同法， a = 69069 , b = 0 , $M = 2^{32}$				
200	10,000	200	7.5%	6.5%

各欄は表1に対応する．また change of sign 検定では単にカイ2乗検定だけでなく，Kolmogorov-Smirnov 検定と組み合わせてみた．即ち，20個のカイ2乗検定値の経験分布のカイ2乗分布に対する適合度を調べた．

表 3

change of sign 検定				
2L + 1	N	M	C	D
M 系列， p = 31 , q = 3 , 原始 3 項式				
91	2,800	50	0%	0%
131	3,400	50	8%	2%
191	4,000	50	28%	16%
231	4,400	50	24%	44%
乗算合同法， a = 69069 , b = 0 , M = 2^{32}				
231	4,500	50	4%	2%
261	4,800	50	6%	2%

C 欄は Kolmogorov-Smirnov 検定で 95%点から 99%点，D 欄は 99%点以上に落ちたサンプルの割合を表わす．

これらの検定の結果より，原始３項式を特性多項式とするＭ系列は統計的に偏りが認められた．このことは，表１から表３に示したＭ系列だけでなく，検定を行なったすべてのＭ系列に関して認められた．さらに次のような問題が生ずる．

１）表１，表２から分かる様に，sojourn time 検定と last visit time 検定では，sojourn time 検定の方がより偏りがはっきりと検出されている．これは sojourn time 検定の方が last visit time 検定より「強い」のであろうか？もしそうであるとすると，それは理論的に示せるのか？change of sign 検定はさらに「弱い」のであろうか？またその理論的説明は？

２）原始３項式を特性多項式とするＭ系列の上記の統計的偏りに対する理論的説明は？本報告では触れなかったが，原始５項式を特性多項式とするＭ系列ではこの様な現象は観察されていないが，このことに対する理論的説明は？

３）原始３項式を特性多項式にもつＭ系列擬似乱数のn-個組の weight function について，Lindholm[6] や栗田[5]，Matsumoto-Kurita[7] などの研究があるが，weight functio

nと sojourn time などのランダムウォークの汎関数との関係はどうなのか？

以上の問題の他にもいろいろな問題が考えられるが，理論的に未解決な問題が多いようである．

参考文献

[1] W.Feller : An Introduction to Probability theory and its Applications, Vol.1, 3rd ed. Wiley (1968)

[2] 伏見正則：乱数，東京大学出版会，(1989)

[3] S.W.Golomb : Shift Register Sequences, Riveised ed. Aegean Park Press, (1982)

[4] D.E.Knuth ：準数値算法／乱数，サイエンス社，(1981)

[5] 栗田良春：M系列の L-tuple の weight distribution の偏りについて，数理解析研究所講究録 498, (1983)

[6] 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.

[7] M.Matsumoto - Y.Kurita : The Fixed point of an m-sequence and local non-rnadomness, Technical Report 88-027, Univ. Tokyo (1988)

[8] B.R.McDonald : Finite Rings with Identity, Dekker, (1974)