Study on a further improvement of Maurer's universal statistical test

Y. Hikima

[1] 京都大学大学院情報学研究科修士 2 回

February 13, 2020

背景

乱数

- “0” と “1” から成る “ランダム” な数列であり，工学をはじめとする様々な分野で用いられる

$\rightarrow$ 物理乱数生成器や擬似乱数生成器によって生成

乱数検定

- (擬似) 乱数生成器から生成された数列が応用先で求められる乱数としての性質を満たしているかを，統計的仮説検定によって評価する一般的な枠組み
- 有名な乱数検定ツールとして NIST SP 800-22 がある
- 本研究では，NIST SP 800-22 に採用されている Maurer’s universal test に基づく統計検定について扱う

Maurer’s universal test

- Maurer によって提案 (1992) され，Coron が修正 (1999)
- エントロピーに基づく検定統計量を計算し検定を行う

検定の流れ

- 検定対象の系列を L ビットごとのブロックに分割し，初めの Q ブロックを初期化用，残りの K ブロックを検定用に用いる
- 第 k 番目のブロックを b_k で表し，各ブロックに対して「そのブロックと一致する直近のブロックとの長さ（何ブロック前にあるか）」を表す変数を計算する:

$$A_n := \begin{cases} n, \quad \text{if} \quad b_{n-l} \neq b_n \quad \text{for} \quad 1 \leq l \leq n-1, \\ \min\{l \in \mathbb{N} \mid l \geq 1, b_{n-l} = b_n\}, \quad \text{otherwise.} \end{cases}$$

- 検定対象の系列 x^n に対し，検定統計量を次式で計算する:

$$f_C(x^n) = \frac{1}{K} \sum_{n=Q+1}^{Q+K} g(A_n), \quad \left(g(m) = (\log_2 \mathrm{e}) \sum_{k=1}^{m-1} \frac{1}{k} \right).$$

→ 算出した検定統計量が平均 μ，分散 σ^2 の正規分布に近似的に従っているとみなして，p 値を計算する

Highly sensitive test [Yamamoto & Liu, 2016]

- Maurer's (Coron's) test を基にした統計検定
- 検定対象の系列 x^n を次の規則によってフリップする:

$$\Pr\{\hat{x}_i = 0 \mid x_i = 0\} = 1, \quad \Pr\{\hat{x}_i = 1 \mid x_i = 1\} = \alpha.$$

→ 変換後の系列 $\hat{x}^n$ は "1" をとる確率が $\hat{q} = \frac{\alpha}{2}$ となる

Highly sensitive test における帰無仮説

- $\mathcal{H}_0$:「検定対象の系列は $\{0,1\}^n$ 上の一様分布に従って生成されたとみなせる」
- $\widetilde{\mathcal{H}}_0$:「フリップに用いる乱数は理想的である」

→ $\overline{\mathcal{H}_0} := \mathcal{H}_0 \wedge \widetilde{\mathcal{H}}_0$:「変換後の系列は "1" をとる確率が $\hat{q}$ であるような $\{0,1\}^n$ 上の分布から独立に生成されたとみなせる」

本研究の目的

- Highly sensitive test では，次式で定義される p 値を計算する:

期待値

$$p = \mathrm{erfc}\left(\left|\frac{f_C(\hat{x}^n) - L \times H(\hat{q})}{\sqrt{2} \times \sigma_C(\hat{q})}\right|\right).$$

標準偏差

※ erfc は相補誤差関数，H は 2 値エントロピー関数を表す

既往研究の課題

- $\hat{q} \neq 0.5$ における参照分布の分散 $\sigma_C(\hat{q})^2$ が理論的に導出されておらず，擬似乱数を用いて算出された値が使われている

→ 定数であるパラメータが正しく与えられておらず，
検定の信頼性の観点から好ましいとは言えない

参照分布の分散を任意の $\hat{q}$ に対して理論的に導出する

参照分布の分散

参照分布の分散 $\sigma_{C,\hat{q}}(K)^2 := \sigma_C(\hat{q})^2$ は次のように与えられる:

$$
\begin{aligned}
&\sigma_{C,\hat{q}}(K)^2 \\
=&\frac{1}{K^2}\left(K \times \mathrm{Var}[g(A_n)] + 2\sum_{k=1}^{K-1}(K-k)\times \mathrm{Cov}[g(A_n), g(A_{n+k})]\right).
\end{aligned}
$$

分散および共分散はそれぞれ次のように与えられる:

$$
\begin{aligned}
\mathrm{Var}[g(A_n)] &= \sum_{i=1}^{\infty}\{g(i)\}^2 \mathrm{Pr}[A_n = i] - \{LH(\hat{q})\}^2, \\
\mathrm{Cov}[g(A_n), g(A_{n+k})] &= \sum_{i,j\geq 1} g(i)g(j)\mathrm{Pr}[A_n = i,\ A_{n+k} = j] - \{LH(\hat{q})\}^2.
\end{aligned}
$$

- 周辺分布および同時分布の導出が必要 (以下で導出)
- 以下では，$w_r := \hat{q}^r(1-\hat{q})^{L-r}$ とおく

※ $\hat{q}$ は系列において “1” をとる確率

周辺分布の導出

事象 $\mathcal{M}$ を次のように定める:

$$\mathcal{M} = \langle b_{n-i} = b_n, b_{n-i+1} \neq b_n, \ldots, b_{n-1} \neq b_n \rangle .$$

各ブロックが独立同分布に従うとき，

$$\Pr[A_n = i] = \sum_{r=0}^{L} \Pr[\mathcal{M} \mid \ell(b_n) = r] \times \Pr[\ell(b_n) = r].$$

ここに，$\ell(b)$ はブロック b における "1" の個数を表す．また，

$$\Pr[\mathcal{M} \mid \ell(b_n) = r] = w_r \times (1 - w_r)^{i-1},$$

$$\Pr[\ell(b_n) = r] = \binom{L}{r} w_r.$$

よって，周辺分布は次式で与えられる:

$$\Pr[A_n = i] = \sum_{r=0}^{L} \binom{L}{r} w_r^2 (1 - w_r)^{i-1}.$$

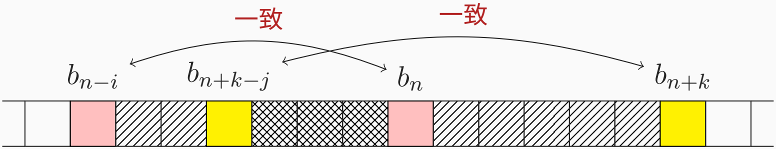
同時分布の導出 ($k+1 \le j \le k+i-1$ の場合)

事象 $e_3(b, b')$ を次のように定める:

$$
\begin{aligned}
e_3(b, b') := & \left\langle b_{n-i} = b, b_n = b, b_{n+k-j} = b', b_{n+k} = b' \right\rangle \\
& \wedge \langle b_{n-i+1} \neq b, \ldots, b_{n+k-j-1} \neq b \rangle \\
& \wedge \langle b_{n+k-j+1} \neq b, \ldots, b_{n-1} \neq b \rangle \\
& \wedge \left\langle b_{n+k-j+1} \neq b', \ldots, b_{n-1} \neq b' \right\rangle \\
& \wedge \left\langle b_{n+1} \neq b', \ldots, b_{n+k-1} \neq b' \right\rangle .
\end{aligned}
$$

事象 $e_3(b, b')$ が起こる確率は以下で与えられる:

$$
\begin{aligned}
& \Pr\left[e_3(b, b')\right] \\
= & w_{r_1}^2 w_{r_2}^2 (1 - w_{r_1})^{i-j+k-1} (1 - w_{r_1} - w_{r_2})^{j-k-1} (1 - w_{r_2})^{k-1}.
\end{aligned}
$$



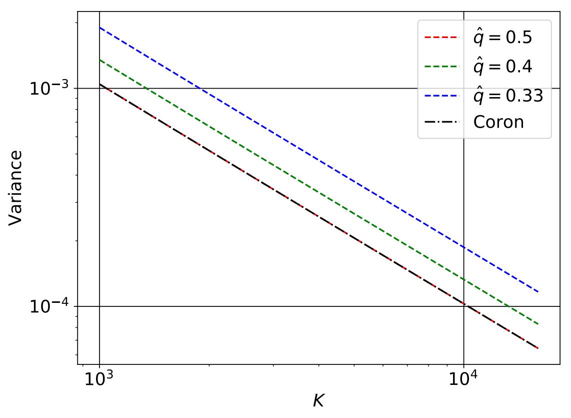
同時分布の導出

したがって，求める同時分布は次のように表される:

$$
\begin{aligned}
&\Pr[A_n = i,\, A_{n+k} = j] \\
=&\Pr\left[\bigvee_{b\in\{0,1\}^L} \bigvee_{b'\in\{0,1\}^L\setminus\{b\}} e_3(b,b')\right] \\
=& \sum_{b\in\{0,1\}^L} \sum_{b'\in\{0,1\}^L\setminus\{b\}} \Pr\left[e_3(b,b')\right] \\
=& \sum_{r_1=0}^{L} \sum_{r_2\neq r_1} \binom{L}{r_1}\binom{L}{r_2} \Pr\left[e_3(b,b')\right] \\
&+ \sum_{r_1=0}^{L} \sum_{r_2\in\{r_1\}} \binom{L}{r_1}\left\{\binom{L}{r_1} - 1\right\} \Pr\left[e_3(b,b')\right].
\end{aligned}
$$

→ 周辺分布および同時分布を参照分布の分散 $\sigma_{C,\hat{q}}(K)^2$ の式に代入することにより，求める参照分布の分散が得られる

計算機実験 1: $L = 4$ の場合



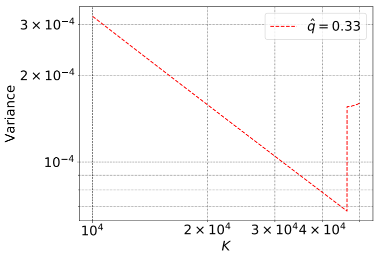
- 参照分布の分散は $\mathcal{O}(\frac{1}{K})$ で減少
- $\hat{q} = 0.5$ のとき，既往研究の結果と整合
- 推奨値である $K = 1000 \times 2^4$ における値が計算可能

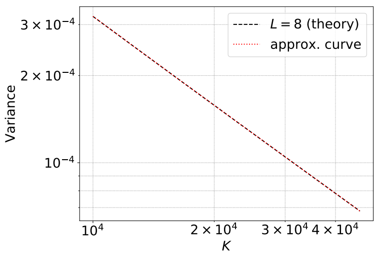
計算機実験 2: $L = 8$ の場合

- 計算が途中で破綻（左図）
- 次式による曲線近似を考える:

$$\sigma^2_{C,\hat{q}}(K) = \frac{1}{K}\left(a + \frac{b}{K}\right),$$

ここに，a, b は実数値定数.





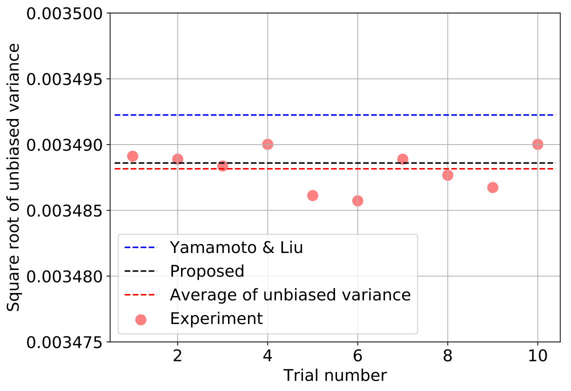
- 近似曲線が理論的な値と整合している（右図）

$\rightarrow$ 推奨値: $K = 1000 \times 2^8$ の値を算出

計算機実験 3: 擬似乱数を用いて算出した値との比較

	Yamamoto&Liu	Proposed	MT
$\sigma_{C,0.33}(K)$	0.00349225	0.00348860	0.00348816

※ MT: Mersenne Twister を用いて算出した値 (10 回の試行の平均値)



既往研究で与えられている数値よりも実験結果と整合

まとめ

- NIST SP 800-22 に含まれる乱数検定手法の一つである "Maurer's universal test" に基づいた "Highly sensitive test" における参照分布の分散を理論的に導出した

$\rightarrow$ Highly sensitive test に対する理論的な裏付けを与えた

- 導出した式を用いて $L = 4$ の場合における参照分布の分散が正しく計算できることを，計算機実験を通して確認した
- 近似曲線を求めることによって，$L = 8$ の場合における参照分布の分散を求めた
  - 推奨値である $K = 1000 \times 2^8$ における分散を計算
  - 既往研究で与えられている数値よりも擬似乱数による実験結果と整合していることを確認

参考文献

[1] Maurer, Ueli M. “A universal statistical test for random bit generators.” Journal of cryptology 5.2 (1992): 89-105.

[2] Rukhin, Andrew, et al. A statistical test suite for random and pseudorandom number generators for cryptographic applications. Booz-allen and hamilton inc mclean va, 2001.

[3] Bassham III, Lawrence E., et al. “Sp 800-22 rev. 1a. a statistical test suite for random and pseudorandom number generators for cryptographic applications”. National Institute of Standards & Technology, 2010.

[4] Coron, Jean-Sébastien, and David Naccache. “An accurate evaluation of Maurer’ s universal test.” International Workshop on Selected Areas in Cryptography. Springer, Berlin, Heidelberg, 1998.

[5] Coron, Jean-Sébastien. “On the security of random sources.” International Workshop on Public Key Cryptography. Springer, Berlin, Heidelberg, 1999.

[6] Yamamoto, Hirosuke, and Qiqiang Liu. “Highly sensitive universal statistical test.” 2016 IEEE International Symposium on Information Theory (ISIT). IEEE, 2016.

[7] Matsumoto, Makoto, and Takuji Nishimura. ”Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator.” ACM Transactions on Modeling and Computer Simulation (TOMACS) 8.1 (1998): 3-30.

検定の流れ

- 検定対象の系列を L ビットごとのブロックに分割する
- 初めの Q ブロックと残りの K ブロックに分割する
  - 初めの Q ブロック: 初期化用セグメント
  - 残りの K ブロック: 検定用セグメント

$L = 4$ の場合における例:

$$01001111 \cdots 011110101101 \cdots 10000011$$

$$\downarrow$$

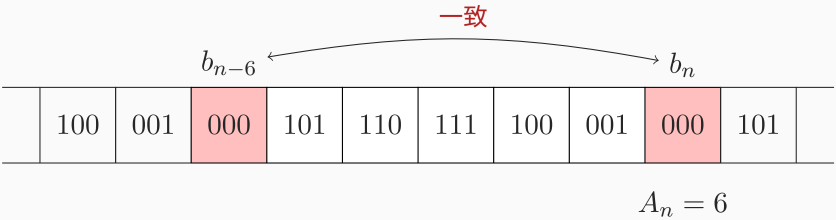
$$\overbrace{\underline{0100}_{1} \mid \underline{1111}_{2} \mid \cdots \mid \underline{0111}_{Q}}^{Q\times L} \mid \overbrace{\underline{1010}_{Q+1} \mid \underline{1101}_{Q+2} \mid \cdots \mid \underline{1000}_{Q+K-1} \mid \underline{0011}_{Q+K}}^{K\times L}$$

※ 下線の数字はブロックの番号を表す

検定統計量を算出する準備

- 系列を L ビットごとのブロックに分割し，第 k 番目のブロックを b_k で表す
- 各ブロックに対して「そのブロックと一致する直近のブロックとの長さ（何ブロック前にあるか）」を表す変数を計算する
- 式で書くと次のように表される:

$$A_n := \begin{cases} n, \quad \text{if} \quad b_{n-l} \neq b_n \quad \text{for} \quad 1 \leq l \leq n-1, \\ \min\{l \in \mathbb{N} \mid l \geq 1, b_{n-l} = b_n\}, \quad \text{otherwise.} \end{cases}$$



検定統計量

Maurer の検定統計量:

$$f_M(x^n) = \frac{1}{K} \sum_{n=Q+1}^{Q+K} \log_2 A_n.$$

Coron の検定統計量:

$$f_C(x^n) = \frac{1}{K} \sum_{n=Q+1}^{Q+K} g(A_n), \quad \left(g(m) = (\log_2 \mathrm{e}) \sum_{k=1}^{m-1} \frac{1}{k} \right).$$

- これらの検定統計量は系列のエントロピーに関係する
- これらの検定統計量が平均 μ，分散 σ^2 の正規分布に近似的に従っているとみなして，p 値を計算する

→ 平均および分散は既往研究で与えられている

参照分布の平均と分散

Maurer による検定統計量の場合:

$$\mu_M = 2^{-L} \sum_{i=1}^{\infty} (1 - 2^{-L})^{i-1} \log_2 i$$

$$\sigma_M^2 = c_M(L, K)^2 \times \frac{\mathrm{Var}[\log_2 A_n]}{K}$$

Coron による検定統計量の場合:

$$\mu_C = L \times H(p)$$

$$\sigma_C^2 = c_C(L, K)^2 \times \frac{\mathrm{Var}[g(A_n)]}{K}$$

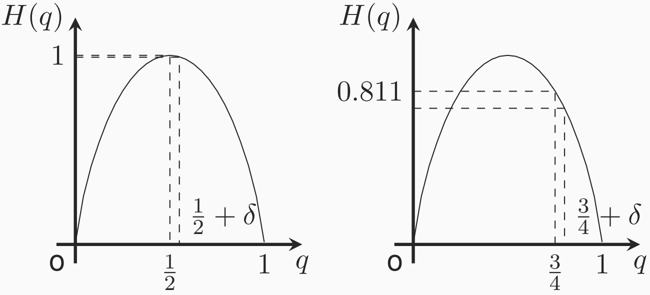
※ 関数 H は 2 値エントロピー関数を表す

※ 上式における $c_M(L, K)$, $c_C(L, K)$ は既往研究で与えられている定数である

Highly sensitive universal statistical test

- Maurer's (Coron's) test を基にした統計検定手法
- 検定対象の系列における "1" を一定の確率で "0" に変換する

$\rightarrow$ 系列において各ビットが "1" である確率を q としたとき，
$q = 0.5$ からの微妙な偏りをより検出しやすくするため



※ 関数 H は二値エントロピー関数を表す

Highly sensitive test の帰無仮説

- $\mathcal{H}_0$:「検定対象の系列は $\{0,1\}^n$ 上の一様分布に従って生成されたとみなすことができる」
- $\widetilde{\mathcal{H}}_0$:「フリップに用いる乱数は理想的である」

$$\downarrow$$

- $\overline{\mathcal{H}}_0 := \mathcal{H}_0 \wedge \widetilde{\mathcal{H}}_0$:「変換後の系列は "1" をとる確率が $\hat{q}$ であるような $\{0,1\}^n$ 上の分布から独立に生成されたとみなすことができる」

結論:

- 検定に合格 $\rightarrow \overline{\mathcal{H}}_0 := \mathcal{H}_0 \wedge \widetilde{\mathcal{H}}_0$
- 検定に不合格 $\rightarrow \neg\mathcal{H}_0$

Algorithm of highly sensitive test

1. パラメータとして，L, Q, K, α を設定する [*1)]
2. 検定対象の系列 x^n を以下の規則で $\hat{x}^n$ に変換する [*2)]

$$\Pr\{\hat{x}_i = 0 \mid x_i = 0\} = 1, \quad \Pr\{\hat{x}_i = 1 \mid x_i = 1\} = \alpha.$$

3. 変換後の系列を L ビットごとのブロックに分割し，各ブロックに対して変数 A_n を計算する．
4. 検定統計量 $f(\hat{x}^n)$ を計算し，次式により p 値を算出する:

$$p = \operatorname{erfc}\left(\left|\frac{f_C(\hat{x}^n) - L \times H(0.5\alpha)}{\sqrt{2} \times \sigma_C(0.5\alpha)}\right|\right).$$

5. 判定:
   - $p < 0.01$ ならば帰無仮説 $\mathcal{H}_0$ を棄却する

[*1)] 推奨値: $L = 8$, $Q = 10 \times 2^L$, $K = 1000 \times 2^L$, $\alpha = 0.66$

[*2)] 変換後の系列において "1" をとる確率は $\hat{q} = 0.5\alpha$ となる．

定常性

ブロックの添字を次のように付け替える:

$$\overbrace{\underset{1}{\underline{0100}} \mid \underset{2}{\underline{1111}} \mid \cdots \mid \underset{Q}{\underline{0111}}}^{Q\times L} \mid \overbrace{\underset{Q+1}{\underline{1010}} \mid \underset{Q+2}{\underline{1101}} \mid \cdots \mid \underset{Q+K-1}{\underline{1000}} \mid \underset{Q+K}{\underline{0011}}}^{K\times L}$$

$$\downarrow$$

$$\overbrace{\underset{1-Q}{\underline{0100}} \mid \underset{2-Q}{\underline{1111}} \mid \cdots \mid \underset{0}{\underline{0111}}}^{Q\times L} \mid \overbrace{\underset{1}{\underline{1010}} \mid \underset{2}{\underline{1101}} \mid \cdots \mid \underset{K-1}{\underline{1000}} \mid \underset{K}{\underline{0011}}}^{K\times L}$$

このとき，以下が成り立つ．

事実

帰無仮説の下で $Q \to \infty$ とすると，系列 $\{A_k\}_{k=1}^{K}$ は stationary ergodic (strictly stationary) である．すなわち，任意の m, n について，$\{A_k\}_{k=n}^{n+m}$ の同時分布は n に依らず，m にのみ依存する．

参照分布の分散の導出

帰無仮説の下，参照分布の分散 $\sigma_{C,\hat{q}}(K)^2 := \sigma_C(\hat{q})^2$ は次のように与えられる:

$$
\begin{aligned}
&\sigma_{C,\hat{q}}(K)^2 \\
=&\mathrm{Var}\left[\frac{1}{K}\sum_{n=Q+1}^{K+Q} g(A_n)\right] \\
=&\frac{1}{K^2}\left(\sum_{n=Q+1}^{K+Q}\mathrm{Var}[g(A_n)] + 2\sum_{1\le i<j\le K}\mathrm{Cov}[g(A_{Q+i}), g(A_{Q+j})]\right) \\
=&\frac{1}{K^2}\left(K \times \mathrm{Var}[g(A_n)] + 2\sum_{k=1}^{K-1}(K-k)\times \mathrm{Cov}[g(A_n), g(A_{n+k})]\right).
\end{aligned}
$$

- 最後の等式において系列 $\{A_k\}_{k=1}^{K}$ の定常性を適用
- 分散および共分散の導出が必要

既往研究

周辺分布

2 値系列において，各ビットが "1" である確率と "0" である確率がそれぞれ $\frac{1}{2}$ である場合，周辺分布は次のように与えられる:

$$\Pr[A_n = i] = 2^{-L}(1 - 2^{-L})^{i-1}.$$

同時分布

同様の場合において，同時分布は次のように与えられる:

$$\begin{aligned}
&\Pr[A_n = i,\, A_{n+k} = j] \\
&= \begin{cases}
2^{-2L}(1-2^{-L})^{i+j-2} & (1 \le j \le k-1) \\
2^{-2L}(1-2^{-L})^{i+k-2} & (j = k) \\
2^{-2L}(1-2^{-L})^{i-j+2k-1}\left(1-2^{-L+1}\right)^{j-k-1} & (k+1 \le j \le k+i-1) \\
0 & (j = k+i) \\
2^{-2L}(1-2^{-L})^{-i+j-1}\left(1-2^{-L+1}\right)^{i-1} & (j \ge k+i+1)
\end{cases}.
\end{aligned}$$

計算機実験

目的:

- 導出した式によって分散が正しく計算できることを確認する

補足事項:

- 無限和の計算は 10^6 で打ち切る
- $\hat{q} = 0.5$ の場合，Coron による以下の近似式が知られている:

$$\sigma_{C,\hat{q}}(K) = c(L,K)^2 \times \frac{\mathrm{Var}[g(A_n)]}{K}.$$

$\to$ 上式の $c(L,K)$ は定数であり，既往研究で与えられている．

計算機実験 3: 擬似乱数による実験（手順）

1. メルセンヌ・ツイスタ（MT）により，$n = 2,068,480$ ビットの系列を $M = 4,000,000$ 本用意する
2. 変換を施した各系列 $\hat{x}^{n,i}$ に対して，検定統計量 $f_i = f_C(\hat{x}^{n,i})$ を計算する
3. M 個の検定統計量から不偏分散を次式で計算する:

$$u^2 = \frac{1}{M}\sum_{i=1}^{M}(f_i - \overline{f}) \qquad \left(\overline{f} = \frac{1}{M}\sum_{i=1}^{M} f_i\right)$$

4. 以上を計 10 回繰り返し，不偏分散 $u_1^2, u_2^2, \ldots, u_{10}^2$ を得る
5. 不偏分散の平均値を次式で計算する:

$$\overline{u}^2 = \frac{1}{10}\sum_{i=1}^{10} u_i^2$$

計算機実験 3: 結果

Table: Value of $\sigma_{C,0.33}(1000 \times 2^8)$ computed using the MT

Trial No.	$\sigma_{C,0.33}(1000 \times 2^8)$
1	0.00348911
2	0.00348889
3	0.00348837
4	0.00349002
5	0.00348612
6	0.00348572
7	0.00348889
8	0.00348767
9	0.00348672
10	0.00349002
Total	$0.00348816 \pm 2.44 \times 10^{-6}$

NIST SP 800-22 による二段階検定:

1. (Proportion test) m 本の系列に対して，p-value $\geq \alpha$ を満たす系列の本数を m_p とする．このとき，m_p が

$$m_p \notin [m(1-\alpha) - \xi\sigma, m(1-\alpha) + \xi\sigma]$$

ならば，帰無仮説を棄却する．

2. (Uniformity test) m 個の p 値が区間 $[0, 1]$ 上一様分布しているかどうかをカイ二乗検定により検定する．カイ二乗検定による p 値 p_T が $p_T \leq \alpha_T$ となる場合，帰無仮説を棄却する．

本実験

- 10^5 本の系列 (各系列は 2068480-bit) に対して実験を施行
- 10^5 本の系列を 100 セットに分割 (1 セットは 1000 本)
- 各セットに対して，二段階検定により系列のランダム性を評価

計算機実験 4: 乱数検定

Table: Number of sets rejected by proportion test with $\xi = 3$

	Yamamoto	Proposed
MT/AES	0	0
AES/MT	0	0
AES/AES	0	0

Table: Number of sets rejected by proportion test with $\xi = 2$

	Yamamoto	Proposed
MT/AES	2	3
AES/MT	4	4
AES/AES	6	6

※ MT/AES:

- 系列→ MT により生成
- フリップ→ AES-128 CTR

計算機実験 4: 乱数検定

Table: Number of sets rejected by uniformity test with the significance level $\alpha_T = 0.01$

	Yamamoto	Proposed
MT/AES	0	0
AES/MT	0	1
AES/AES	0	0

Table: Number of sets rejected by uniformity test with the significance level $\alpha_T = 0.05$

	Yamamoto	Proposed
MT/AES	2	2
AES/MT	3	6
AES/AES	4	5