Maurer's Test を基にした Highly Sensitive Universal Statistical Test における参照分布の分散の導出

◎引間 泰成　　岩﨑 淳　　梅野 健

京都大学

January 29, 2020

Today's Topic

1. Maurer's universal test
2. 参照分布の分散の導出
3. 計算機実験

1. Maurer’s universal test

2. 参照分布の分散の導出

3. 計算機実験

はじめに

乱数検定:

- （擬似）乱数生成器から生成された “0” と “1” から成る数列が応用先で求められる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
- 広く知られている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 に含まれている Maurer’s universal test について扱う

Maurer’s universal test:

- 1992 年 Ueli Maurer によって提案された手法

→ 1999 年 Coron によって修正

- 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:

$$01001111 \cdots 011110101101 \cdots 10000011$$

$$\downarrow$$

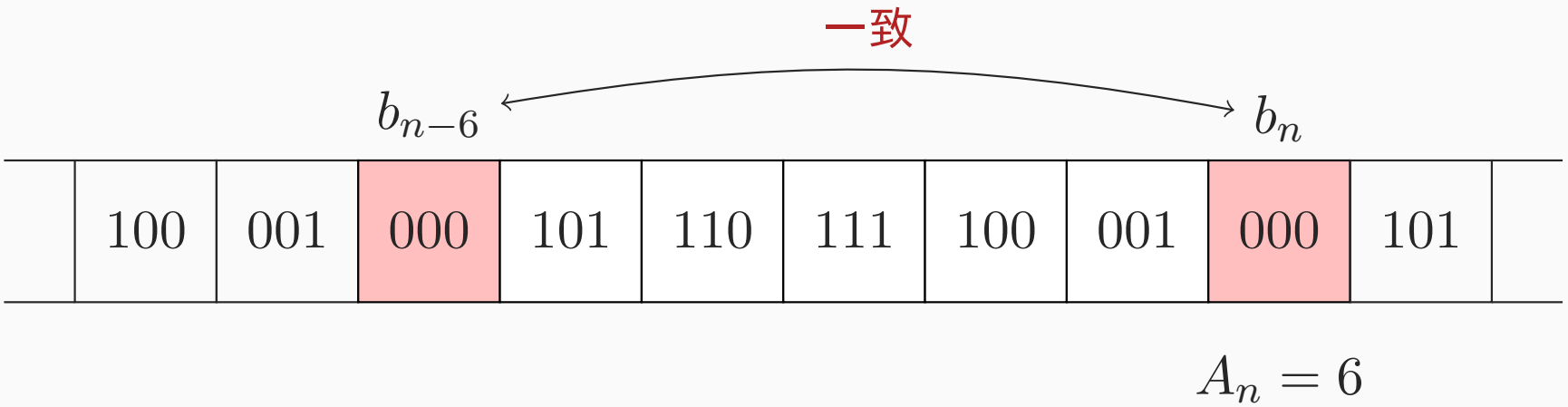
$$\overbrace{\underset{1}{0100} \mid \underset{2}{1111} \mid \cdots \mid \underset{Q}{0111}}^{Q \times L} \mid \overbrace{\underset{Q+1}{1010} \mid \underset{Q+2}{1101} \mid \cdots \mid \underset{Q+K-1}{1000} \mid \underset{Q+K}{0011}}^{K \times L}$$

※ 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:

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



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:

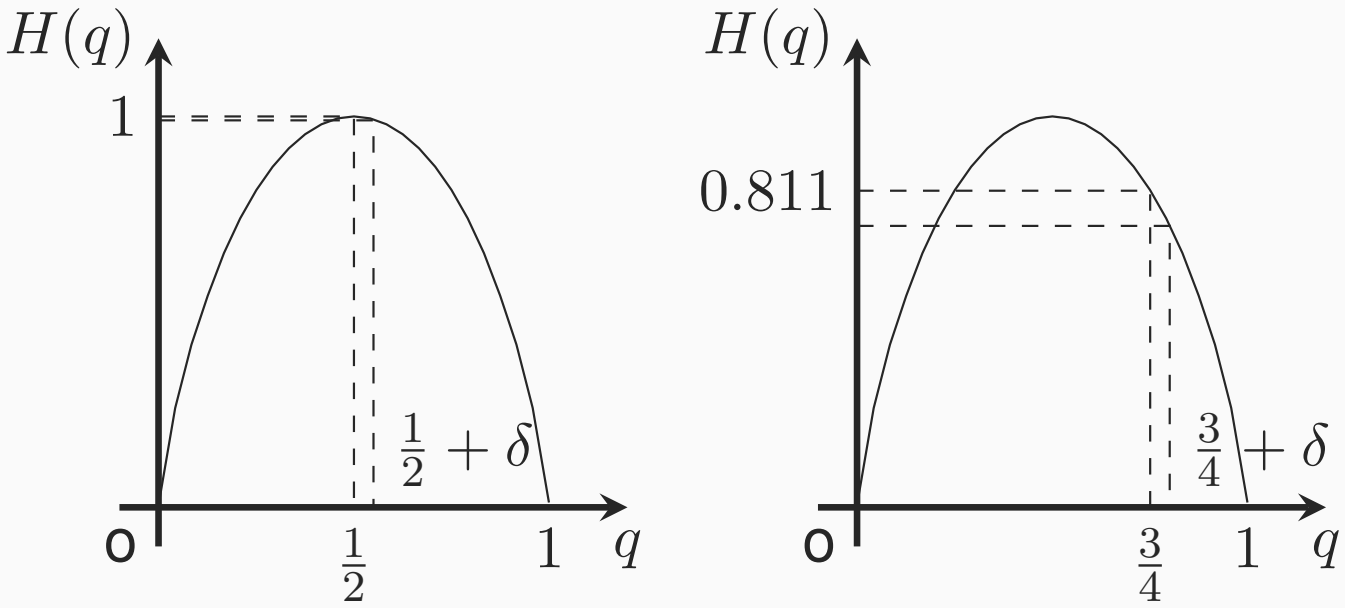
$$f_C(x^n) = \frac{1}{K} \sum_{n=Q+1}^{Q+K} g(A_n), \quad \left(g(m) = (\log_2 \mathrm{e}) \sum_{k=1}^{m-1} \frac{1}{k} \right).$$

- 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

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

→ 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:

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

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

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

[*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.

本研究の目的

- highly sensitive test は以下で定義される p 値を計算する:

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

- 理想的な乱数列 ($\hat{q} = 0.5$) における参照分布の分散 $\sigma_C(0.5)^2$ は理論的に導出されている
- 一方，$\hat{q} \neq 0.5$ の場合における参照分布の分散 $\sigma_C(0.5\alpha)^2$ に関する理論的な導出はなされていない

現状:

- 擬似乱数を用いたシミュレーションによって算出した値を使用

$\rightarrow$ パラメータが正しく与えられていないのは好ましくない

$\downarrow$

任意の $\hat{q}$ に対する参照分布の分散を理論的に導出する

1. Maurer’s universal test

2. 参照分布の分散の導出

3. 計算機実験

定常性

ブロックの添字を次のように付け替える:

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

$$\downarrow$$

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

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 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

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 $\{A_k\}_{k=1}^{K}$ の定常性を用いた
- 分散および共分散の導出が必要

分散・共分散

分散および共分散はそれぞれ次のように与えられる:

$$\mathrm{Var}[g(A_n)] = \sum_{i=1}^{\infty} \{g(i)\}^2 \Pr[A_n = i] - \{LH(\hat{q})\}^2,$$

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

※ これらの値は，n に依らない（$\because$ strictly stationary）

- 周辺分布および同時分布の導出が必要

→ 理想的な乱数列 ($\hat{q} = 0.5$) に対する結果は先行研究で与えられているが，一般の $\hat{q}$ に対する解析はなされていない

- 以下では，$w_r := \hat{q}^r(1-\hat{q})^{L-r}$ とおく

※ $\hat{q}$ は 2 値系列において各ビットが “1” である確率を表す

周辺分布の導出

事象 $\mathcal{M}$ を次のように定める:

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

このとき，各ブロックが独立同分布であるとすると

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

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:

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

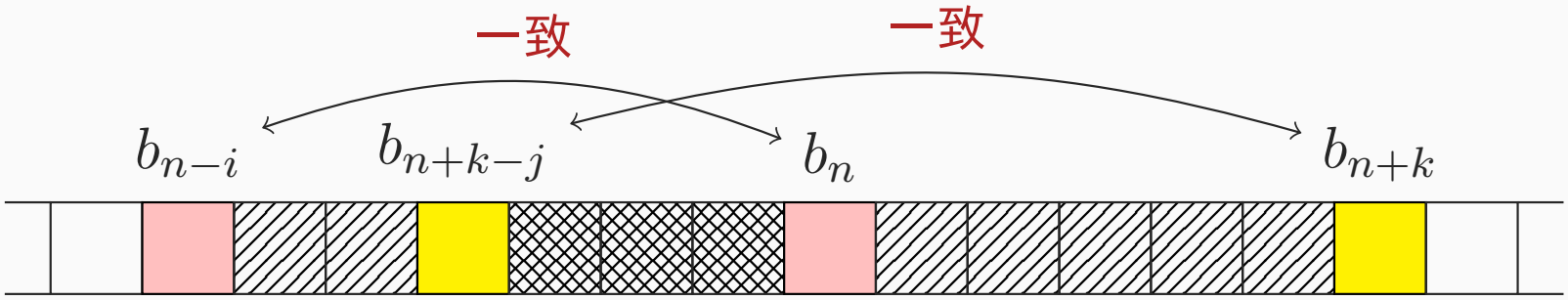
同時分布の導出 ($k+1 \leq j \leq k+i-1$ の場合)

事象 $e_3(b_1, b_2)$ を次のように定める:

$$\begin{aligned} e_3(b_1, b_2) := &\langle b_{n-i} = b_1, b_n = b_1, b_{n+k-j} = b_2, b_{n+k} = b_2 \rangle \\ &\cap \langle b_{n-i+1} \neq b_1, \ldots, b_{n+k-j-1} \neq b_1 \rangle \\ &\cap \langle b_{n+k-j+1} \neq b_1, \ldots, b_{n-1} \neq b_1 \rangle \\ &\cap \langle b_{n+k-j+1} \neq b_2, \ldots, b_{n-1} \neq b_2 \rangle \\ &\cap \langle b_{n+1} \neq b_2, \ldots, b_{n+k-1} \neq b_2 \rangle . \end{aligned}$$

事象 $e_3(b_1, b_2)$ が起こる確率は以下で与えられる:

$$\begin{aligned} &\Pr\left[e_3(b_1, b_2)\right] \\ =&w_{r_1}^2 w_{r_2}^2 (1 - w_{r_1})^{i-j+k-1} (1 - w_{r_1} - w_{r_2})^{j-k-1} (1 - w_{r_2})^{k-1}. \end{aligned}$$



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:

$$
\begin{aligned}
&\Pr[A_n = i,\, A_{n+k} = j] \\
=&\Pr\left[\bigcup_{b_1 \in B^L} \bigcup_{b_2 \in B^L \setminus \{b_1\}} e_3(b_1, b_2)\right] \\
=&\sum_{b_1 \in B^L} \sum_{b_2 \in B^L \setminus \{b_1\}} \Pr\left[e_3(b_1, b_2)\right] \\
=&\sum_{r_1=0}^{L} \sum_{r_2 \neq r_1} \binom{L}{r_1}\binom{L}{r_2} \Pr\left[e_3(b_1, b_2)\right] \\
&+ \sum_{r_1=0}^{L} \sum_{r_2 \in \{r_1\}} \binom{L}{r_1} \left\{\binom{L}{r_1} - 1\right\} \Pr\left[e_3(b_1, b_2)\right].
\end{aligned}
$$

→ 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

1. Maurer’s universal test

2. 参照分布の分散の導出

3. 計算機実験

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

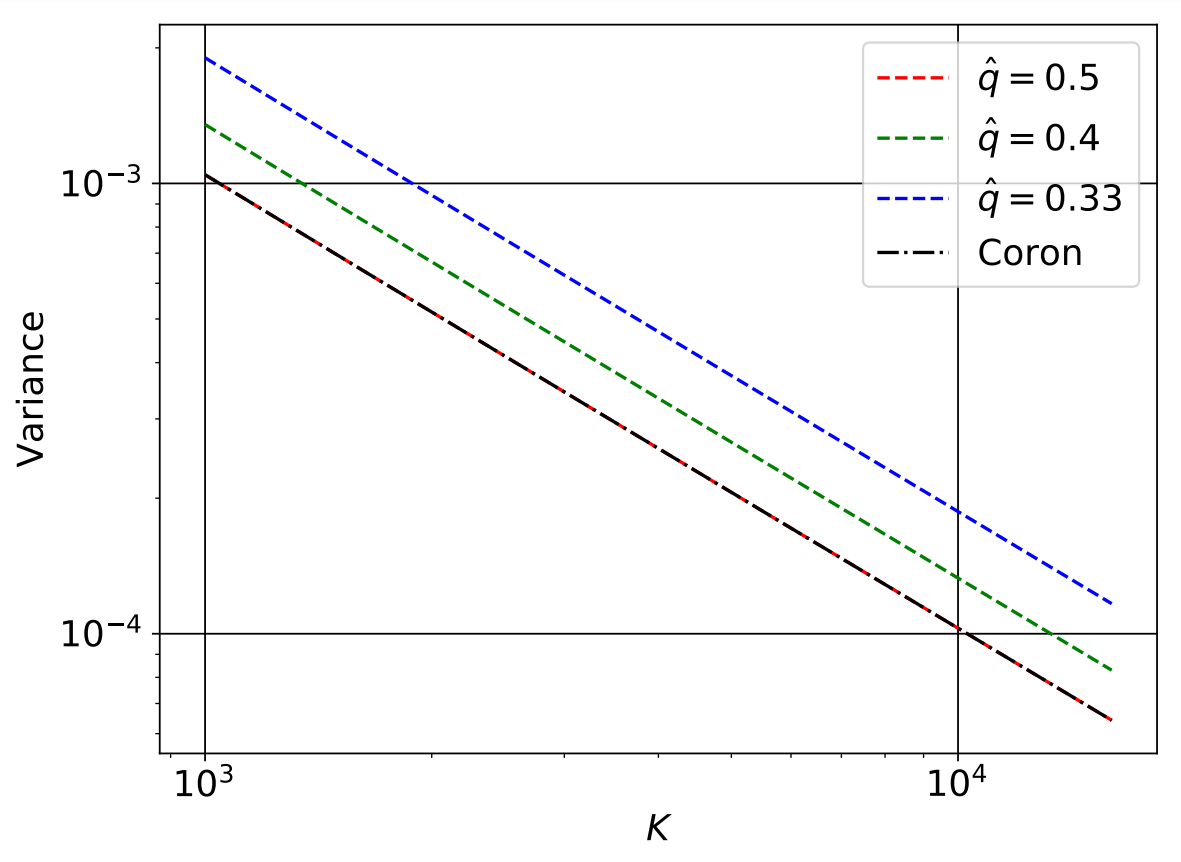
補足事項:

- 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:

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

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

実験 1: $L = 4$ の場合

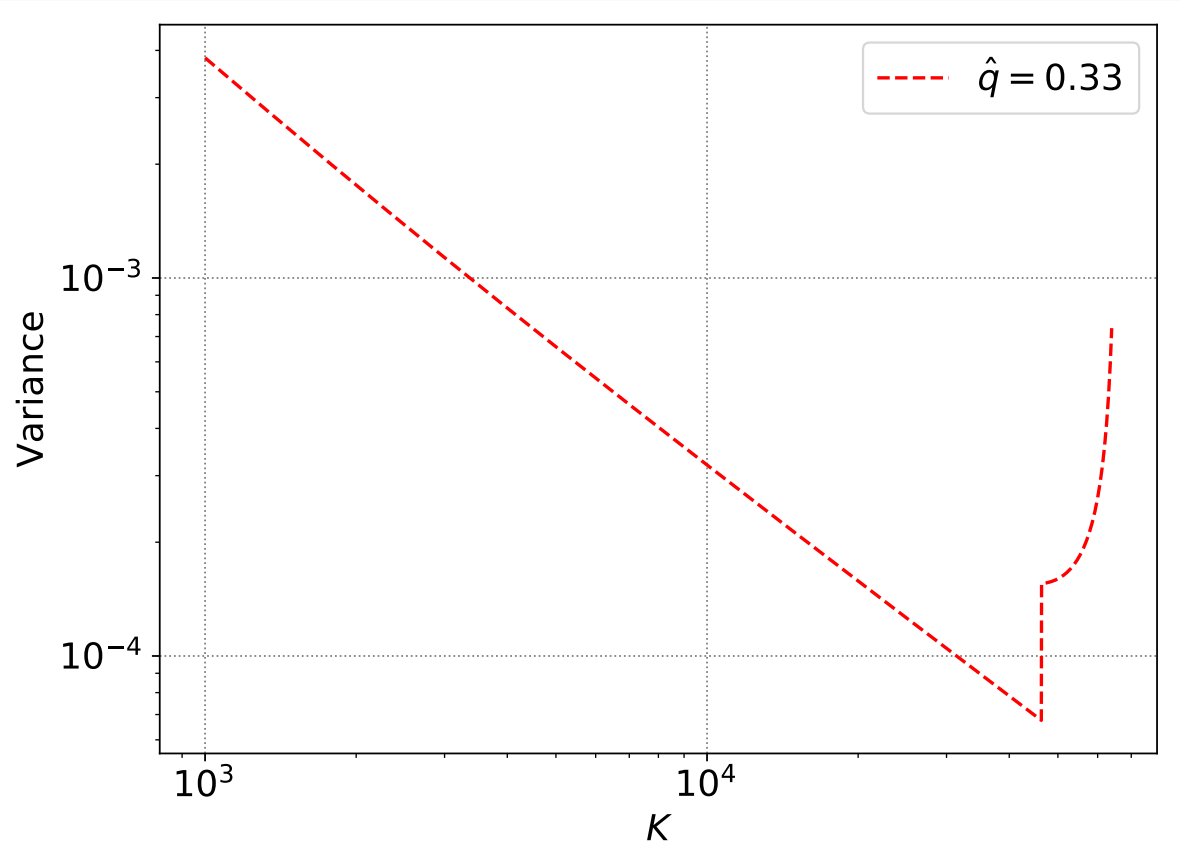


- 参照分布の分散は $\mathcal{O}(\frac{1}{K})$ で減少
- $\hat{q} = 0.5$ のとき，Coron による近似式と整合（$\hat{q} = 0.5$）
- 推奨値である $K = 1000 \times 2^4$ における値が計算可能

実験 2: $L = 8$ の場合

- NIST による推奨値は $L = 8$ であり，そのときの K の値は，$K = 1000 \times 2^8$ である

$\rightarrow$ しかしながら，$K \simeq 5 \times 10^4$ 付近で計算が破綻...

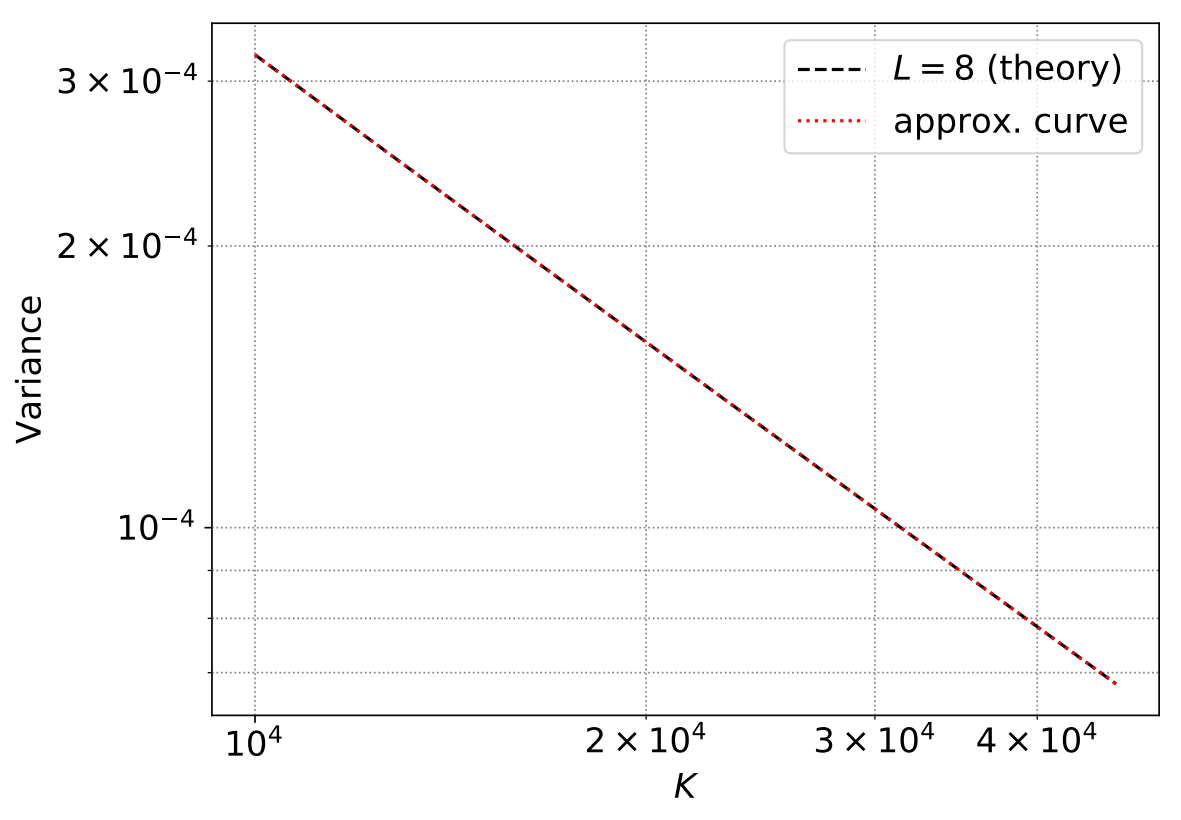


実験 2: $L = 8$ の場合

- 次式による曲線近似を考える:

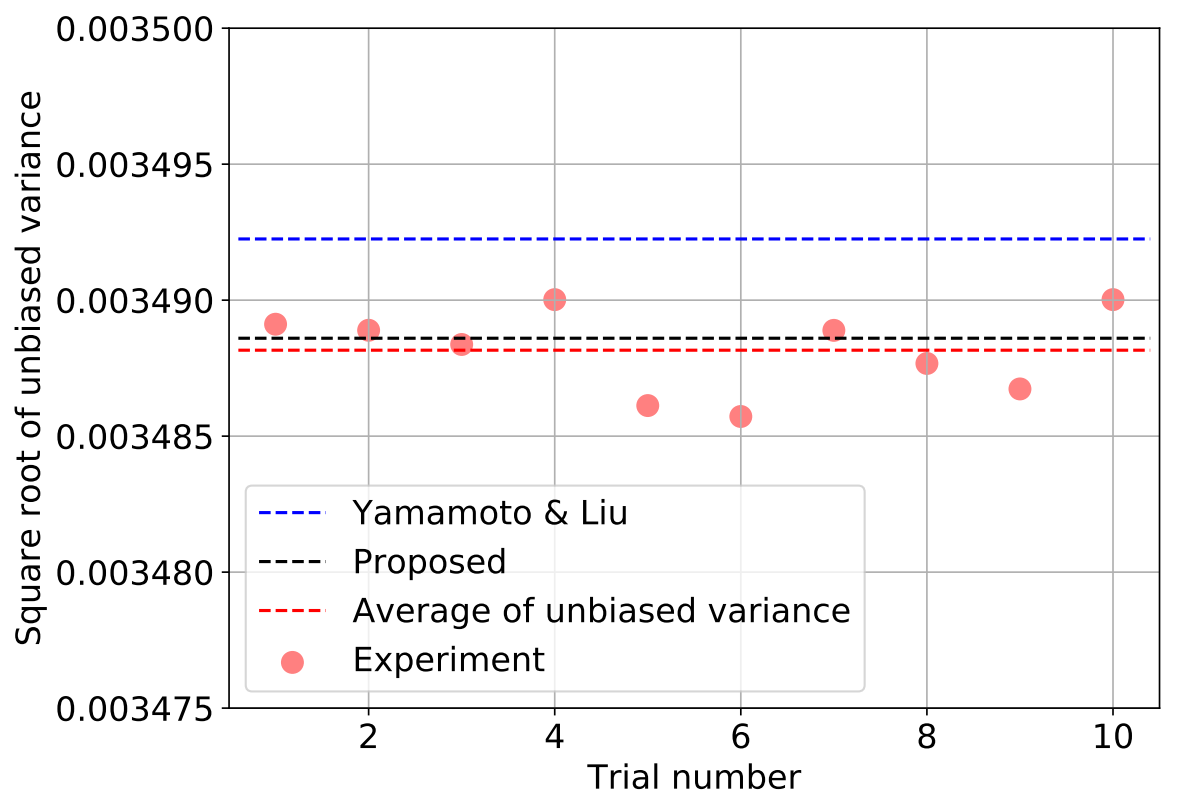
$$\sigma^2_{C,\hat{q}}(K) = \frac{1}{K}\left(a + \frac{b}{K}\right),$$

ここに，a, b は実数値定数.



実験 3: 擬似乱数を用いて算出した値との比較

	Yamamoto&Liu[6]	Experiment	Proposed
$\sigma_{C,0.33}(K)$	0.00349225	0.00348816	0.00348860



既往研究で与えられている数値よりも実験結果と整合

おわりに

- 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
- 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 [6] で与えられている値よりも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.

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:

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