プライバシー保護確率検定システムの実装と評価

佐藤智貴† 菊池浩明† 佐久間淳‡

†東海大学 ‡筑波大学

疫学調査

喫煙者

名前	年齢	喫煙
佐藤智貴	27	有
菊池浩明	32	有
佐久間淳	30	有

死亡者

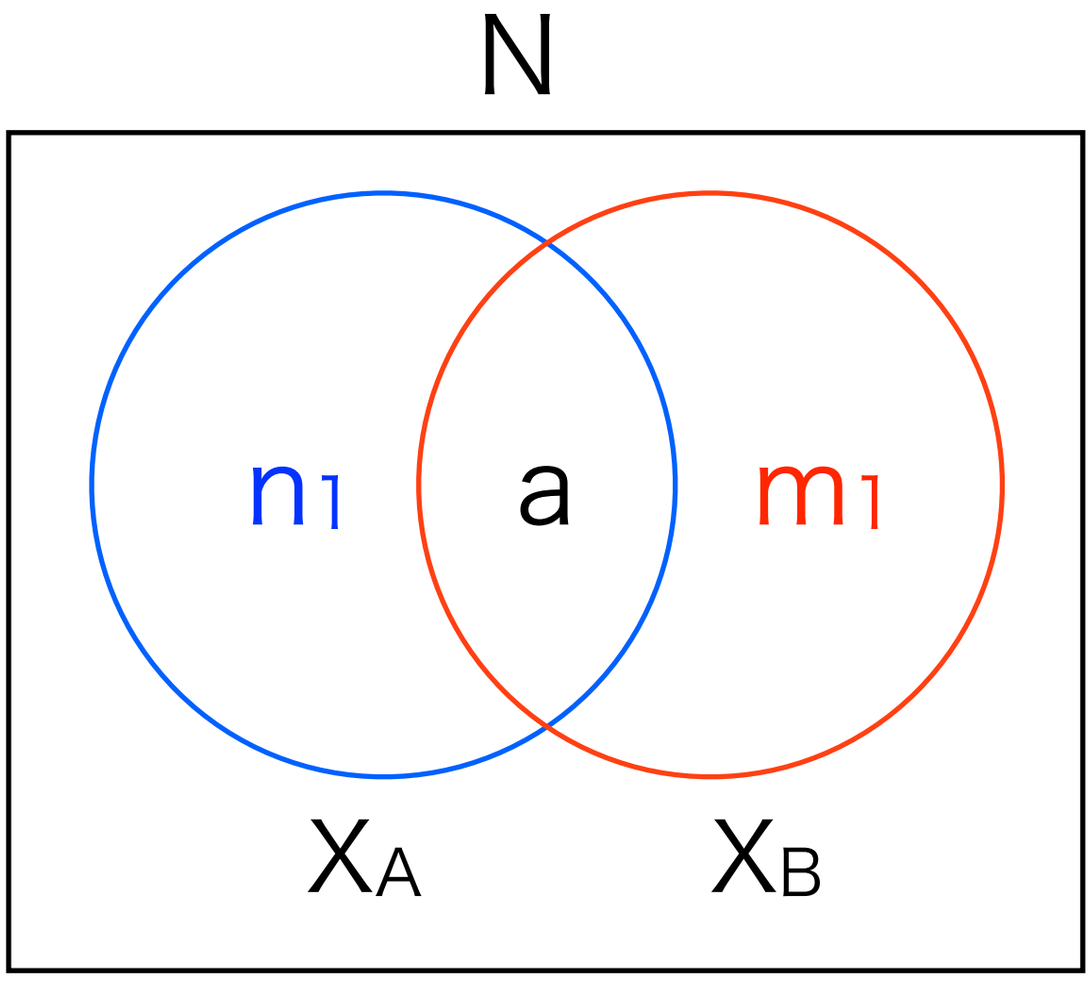
名前	死因
田中真二	胃がん
佐藤智貴	肺がん
鈴木太郎	肺がん

組織A

組織B

プライバシーを考慮する必要

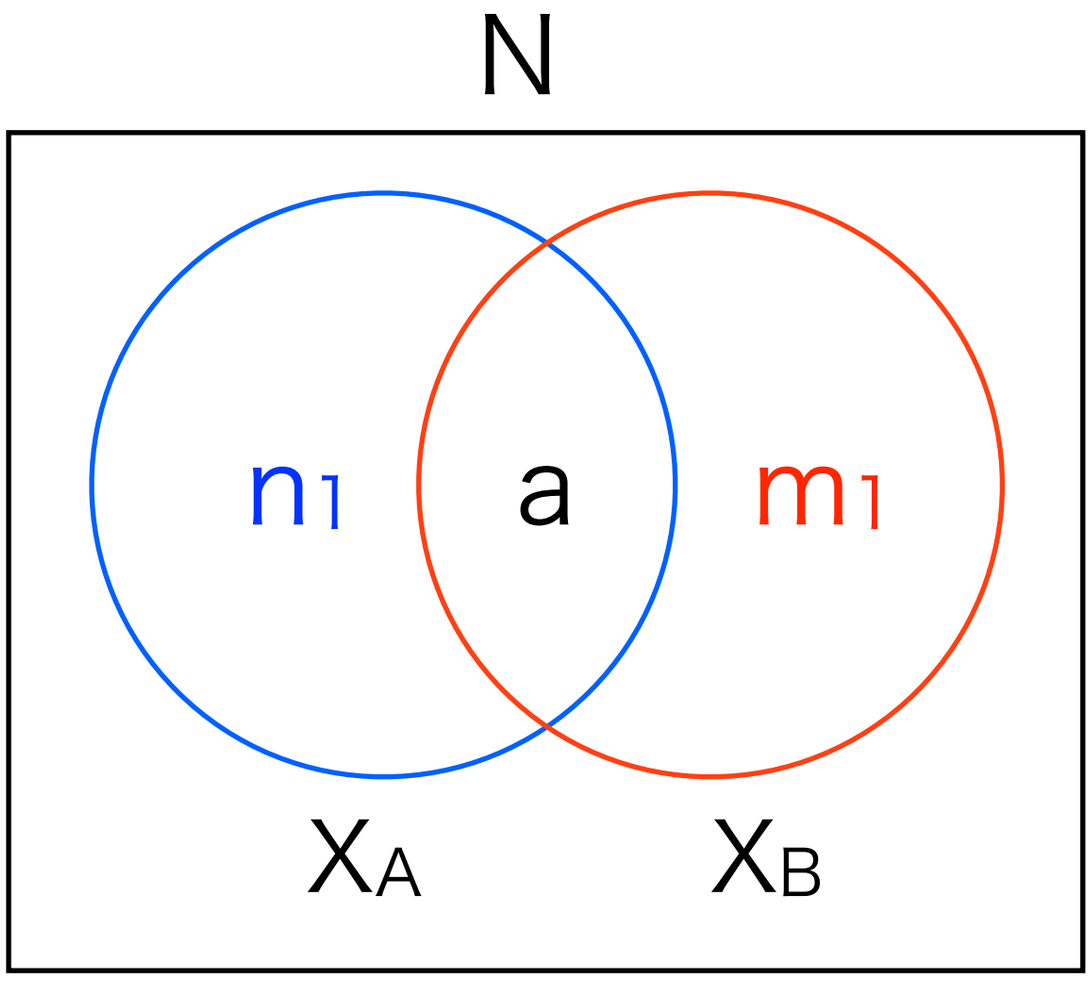
相対危険度RR



	死亡	生存	計
喫煙	a	b	n_1
非喫煙	c	d	n_2
計	m_1	m_2	N

$$RR = \frac{a}{n_1} / \frac{c}{n_2} = \frac{a(c+d)}{(a+b)c} \approx \frac{ad}{bc}$$

相対危険度RR



	死亡	生存	計
喫煙	a	b	n_1
非喫煙	c	d	n_2
計	m_1	m_2	N

$$RR = \frac{a}{n_1} / \frac{c}{n_2} = \frac{a(c+d)}{(a+b)c} \approx \frac{ad}{bc}$$

喫煙は非喫煙に比べて何倍危険か

相対危険度の有意性

- RRが1に等しいか否か
- χが両側検定の有意水準95%, Z(0.05/2)=1.960を上回っているか否かで判定

$$\chi = \frac{\sqrt{N-1}\{(ad-bc) \pm N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}$$

目的

- 暗号技術を用いて，二つの組織のデータを秘匿したままRRの有意性を検定する

秘匿内積プロトコル

組織A：$\mathbf{x} = (x_1, x_2)$
組織B：$\mathbf{y} = (y_1, y_2)$

s_B：乱数

$s_A + s_B = \mathbf{x} \cdot \mathbf{y}$

$s_A = D(c) = E(x_1)^{y1} \cdot E(x_2)^{y2} / E(s_B)$

Aがs_Aを，Bがs_Bをそれぞれ得る

互いのデータを公開せずに，積集合の大きさ$|A \cap B| = s_A + s_B$を計算できる

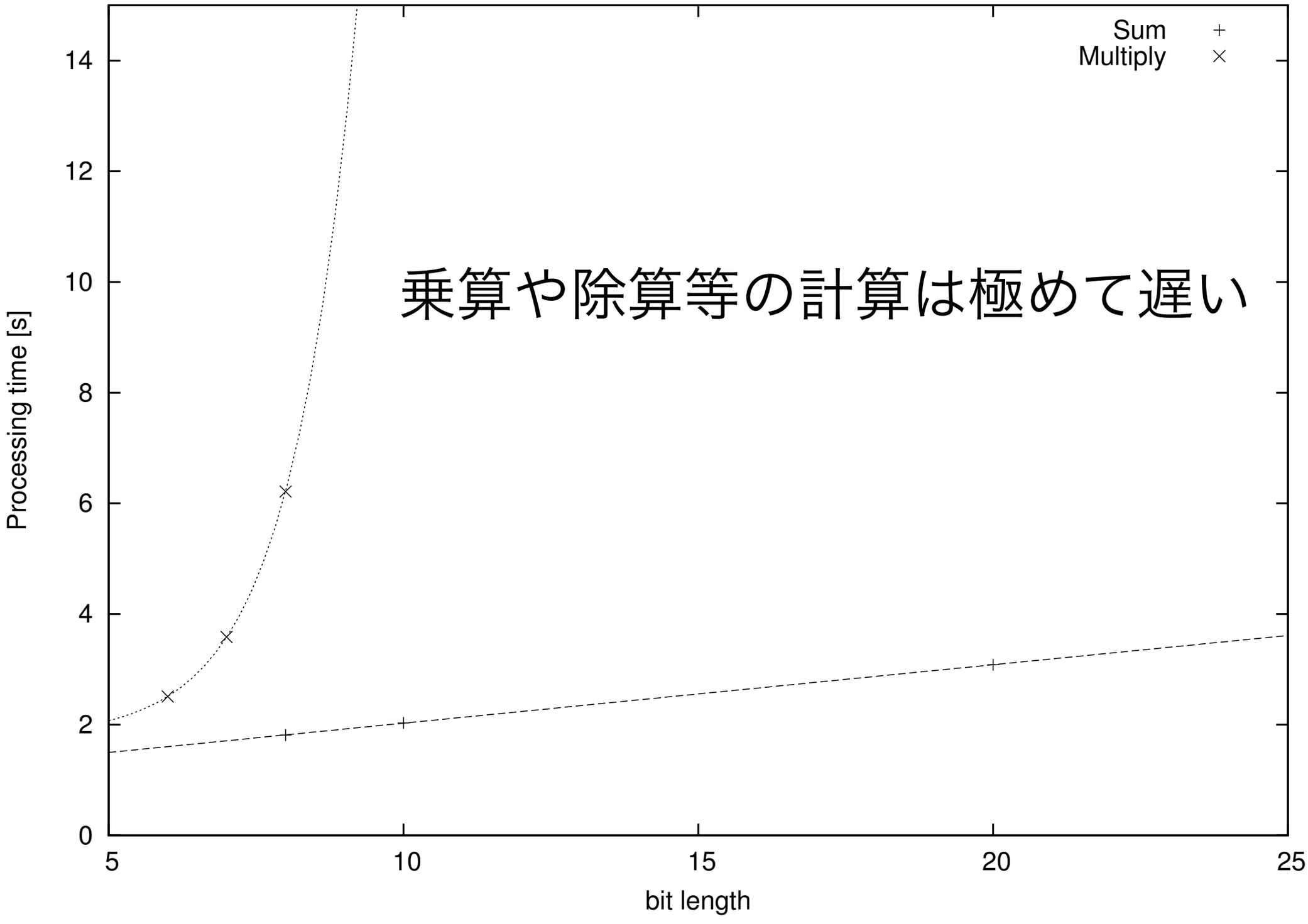
Fairplay

- Yaoの秘匿関数計算(SFE)プロトコル
- 秘匿内積プロトコルで得たs_Aとs_Bを互いに知らせることなく任意の計算を行う
- $s_A + s_B = a$

$$\chi = \frac{\sqrt{N-1}\{(ad-bc) \pm N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}$$

	死亡	生存	計
喫煙	a	b	n_1
非喫煙	c	d	n_2
計	m_1	m_2	N

Fairplayの制約



問題定義

	死亡	生存	計
喫煙	a	b	n_1
非喫煙	c	d	n_2
計	m_1	m_2	N

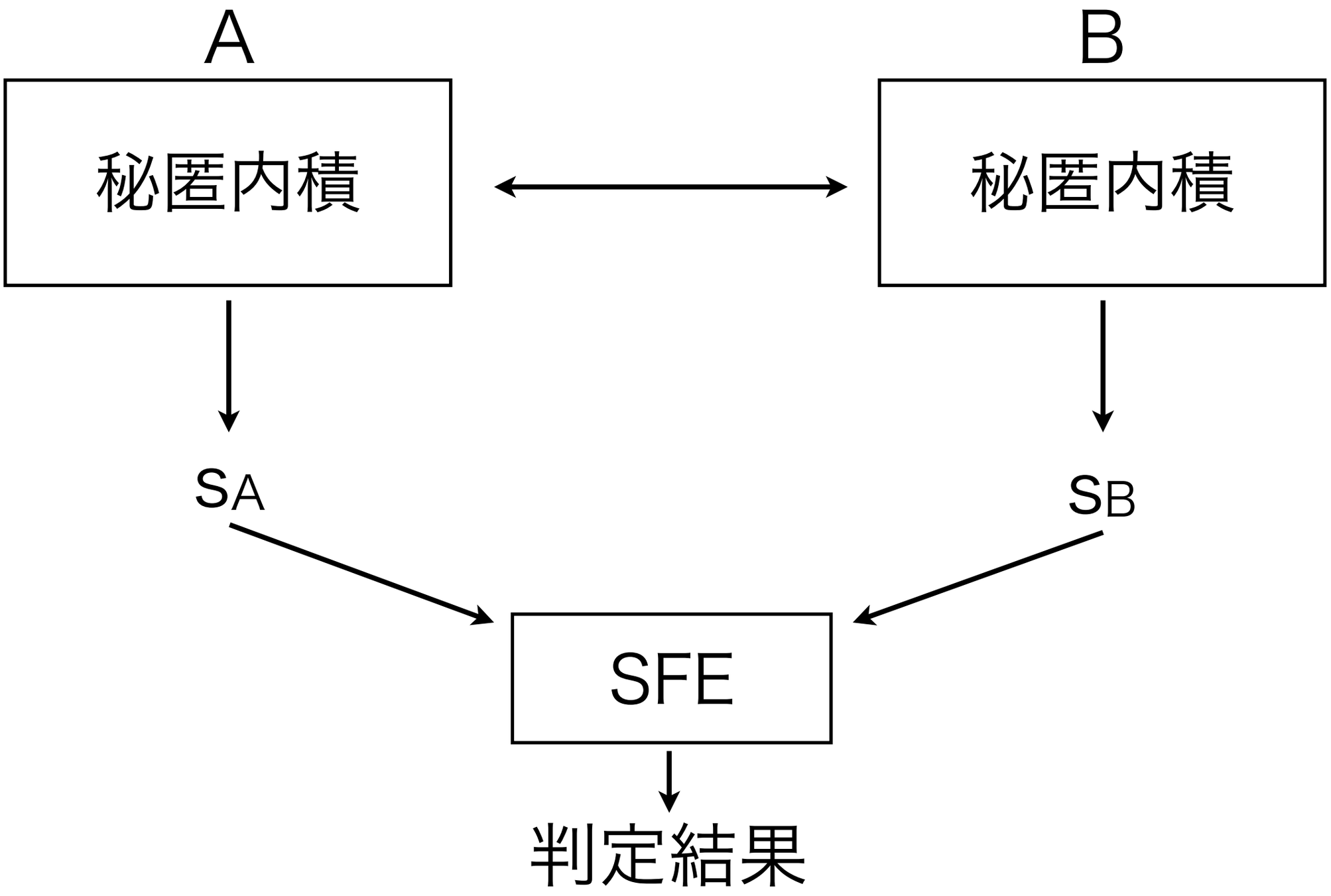
誰も知らない情報

Aのみが知っている情報

Bのみが知っている情報

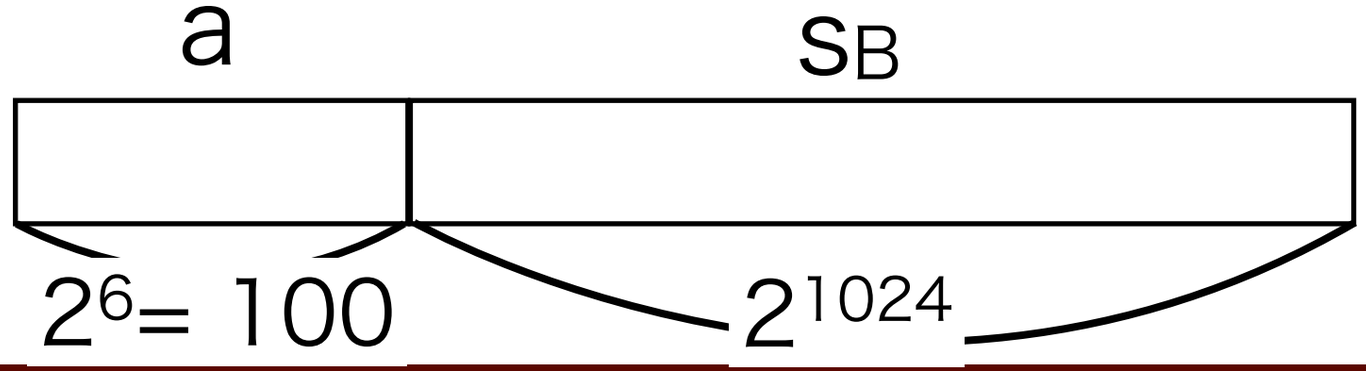
- 対象の特定要因の相対危険度が有意か否かを判定
- 出力は判定結果のみ

処理の流れ



問題点

1. 統計量χを求めるための大きな計算量
   - Fairplayでは求めることが困難 $\chi = \frac{\sqrt{N-1}\{(ad-bc) \pm N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}$
2. 分散値の定義域の大きさ
   - 乱数s_Bは安全性のため，準同型暗号の定義域Z_nから選ぶ(1024bitのような大きい値)
   - しかし，大きすぎる値はFairplayでは計算できない



(1)へのアプローチ

- Fairplayでの計算を加算，減算，比較に限定したい

$$\chi = \frac{\sqrt{N-1}\{(ad-bc) \pm N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}$$

$$= \frac{\sqrt{N-1}\{a(N-n_1-m_1+a)-(n_1-a)(m_1-a)-N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}$$

$$= \frac{\sqrt{N-1}\{aN-n_1 m_1-N/2\}}{\sqrt{n_1 n_2 m_1 m_2}} \qquad (2)$$

(1)へのアプローチ

- Fairplayでの計算を加算，減算，比較に限定したい

$$\chi = \frac{\sqrt{N-1}\{(ad-bc) \pm N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}$$

$$= \frac{\sqrt{N-1}\{a(N-n_1-m_1+a)-(n_1-a)(m_1-a)-N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}$$

$$= \frac{\sqrt{N-1}\{aN-n_1 m_1-N/2\}}{\sqrt{n_1 n_2 m_1 m_2}} \qquad (2)$$

aの一次式

図2. aを変化させた時の統計量 χ

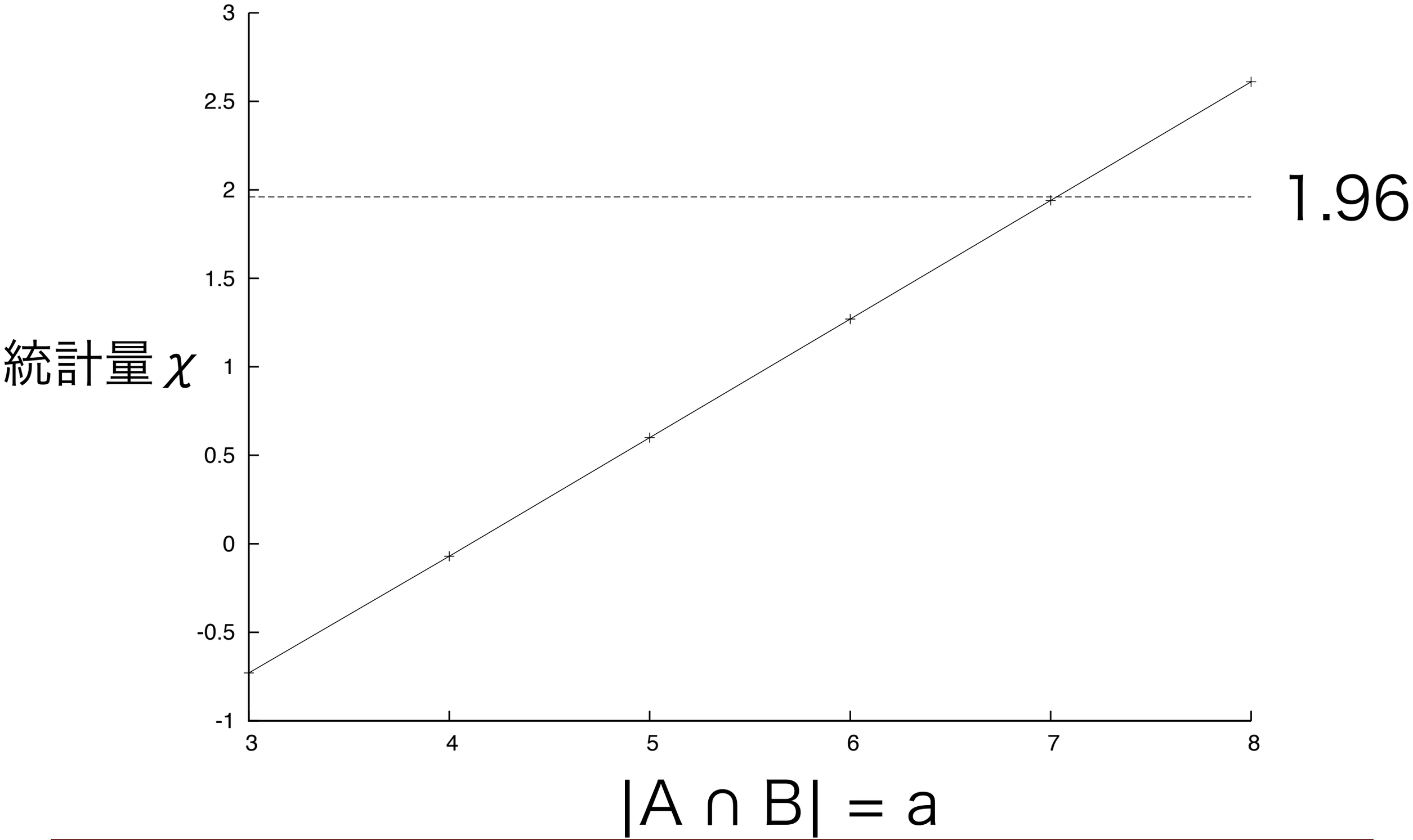
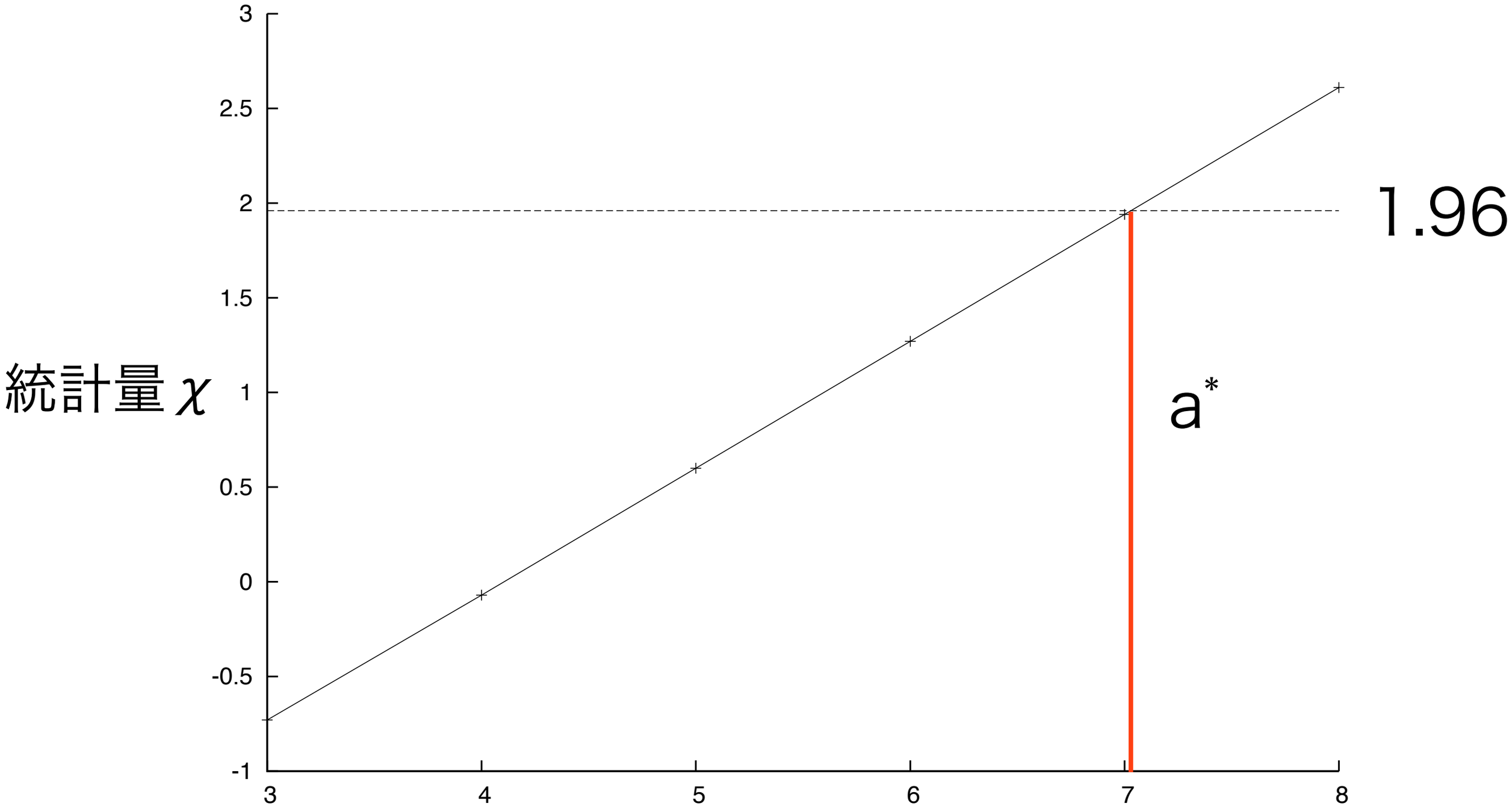


図2. aを変化させた時の統計量 χ



$|A \cap B| = a$

(1)へのアプローチ

$$a^* = \left(\frac{\chi \cdot \sqrt{n_1 n_2 m_1 m_2}}{\sqrt{N-1}} + n_1 m_2 + \frac{N}{2} \right) \cdot \frac{1}{N}$$

$$a^* N = \frac{1.960 \cdot \sqrt{n_1 n_2 m_1 m_2}}{\sqrt{N-1}} + n_1 m_2 + \frac{N}{2}$$

- χ = 1.960の時, |A ∩ B| = a がa^*を上回っているか否かで判定

(1)へのアプローチ

秘匿内積で計算，分散する

$$s_A + s_B = aN = |X_A \cap X_B|N,$$

$$t_A + t_B = \frac{\chi \cdot \sqrt{n_1 n_2 m_1 m_2}}{\sqrt{N-1}} = \left(\frac{\chi\sqrt{n_1 n_2}}{\sqrt{N-1}}\right) \cdot \sqrt{m_1 m_2},$$

$$u_A + u_B = n_1 m_1 + \frac{N}{2}$$

Fairplayで評価を行う

$$(s_A + s_B) > (t_A + t_B) + (u_A + u_B)$$

(1)へのアプローチ

秘匿内積で計算，分散する

$$s_A + s_B = aN = |X_A \cap X_B|N,$$

$$t_A + t_B = \frac{\chi \cdot \sqrt{n_1 n_2 m_1 m_2}}{\sqrt{N-1}} = \left(\frac{\chi\sqrt{n_1 n_2}}{\sqrt{N-1}}\right) \cdot \sqrt{m_1 m_2},$$

$$u_A + u_B = n_1 m_1 + \frac{N}{2}$$

Fairplayで評価を行う

$$(s_A + s_B) > (t_A + t_B) + (u_A + u_B)$$

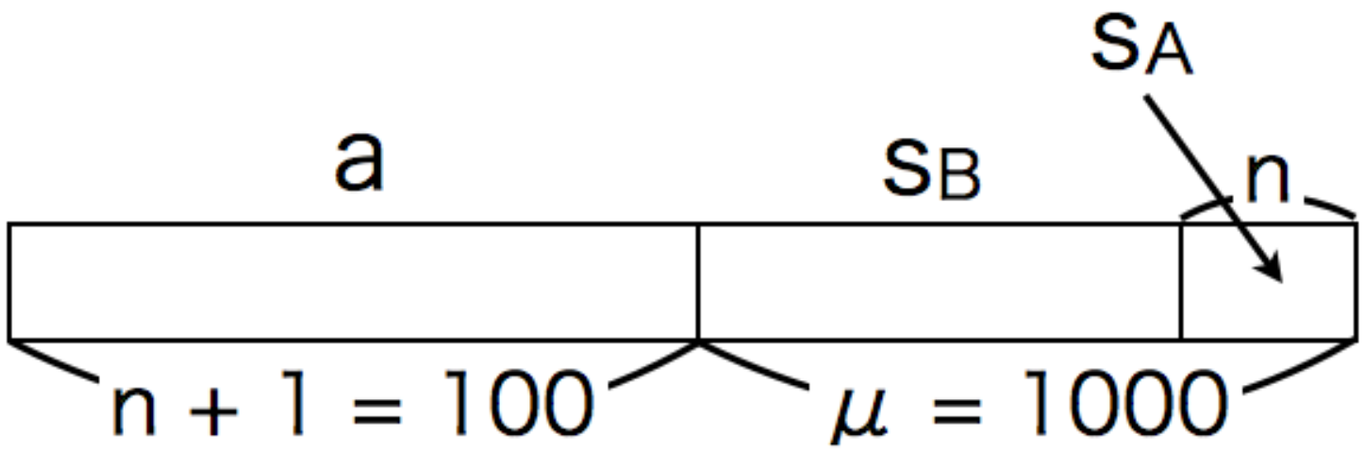
Fairplayでの計算は加算と比較のみ

(2)のための変更点

従来[3]	本提案
$s_B \in Z_n$	$s_B \in [0,\ \mu\text{-}1]$
Fairplay 68.2sec (1024bit)	Fairplay 2sec (30bit)
$c = E(x_1)^{y_1} \cdot E(x_2)^{y_2} / E(s_B)$	$c = E(x_1)^{y_1} \cdot E(x_2)^{y_2} \odot E(s_B)$
危険な領域になる確率 なし	$\frac{n-1}{2\mu}$

危険な領域

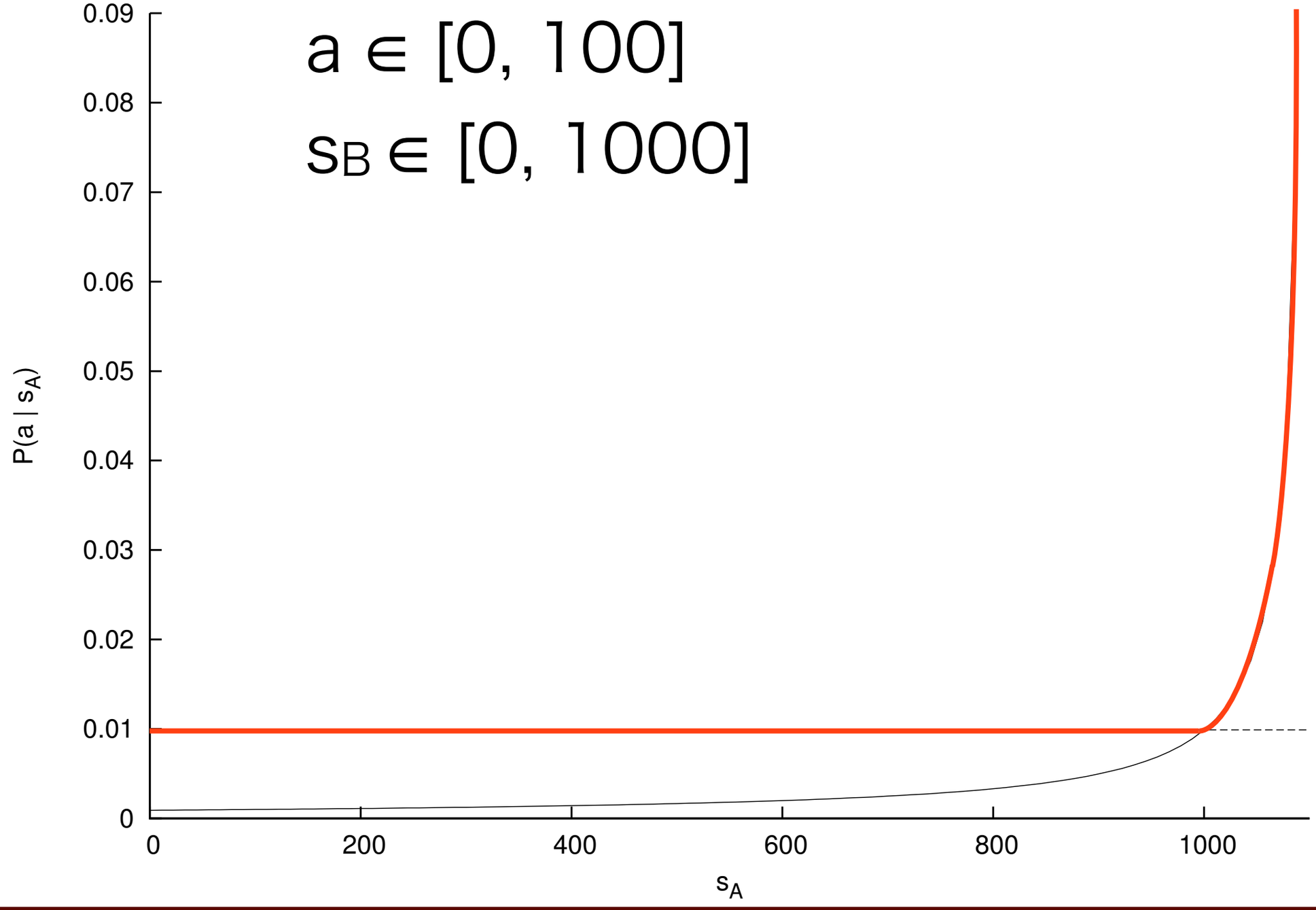
- $s_A = a + s_B$
- 要素数 $n + 1 = 100$, 乱数の最大値 $\mu = 1000$の時
- $a \in [0, 100]$, $s_B \in [0, 1000]$
- $s_A = 1050$だった場合，aは少なくとも50以上



損なわれる条件付き確率P(a|s_A)の変化

$a \in [0, 100]$

$s_B \in [0, 1000]$



定理1

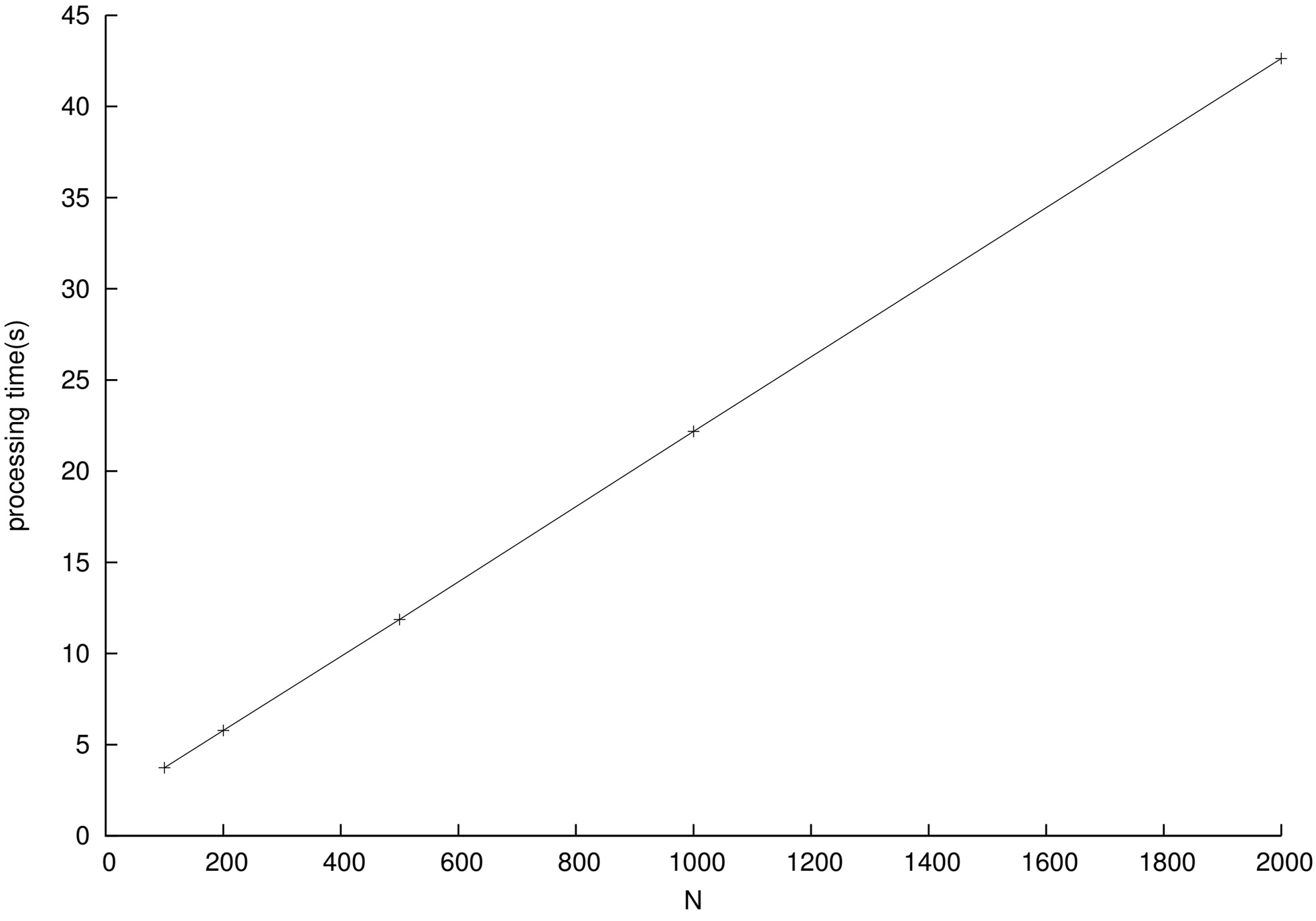
- a ∈ [0, n] , s_B ∈ [0, μ-1]の一様分布から選んだ値
- $s_A = s_B + a > \mu$となる確率は

$$P(\mu < s_A) = \frac{n-1}{2\mu}$$

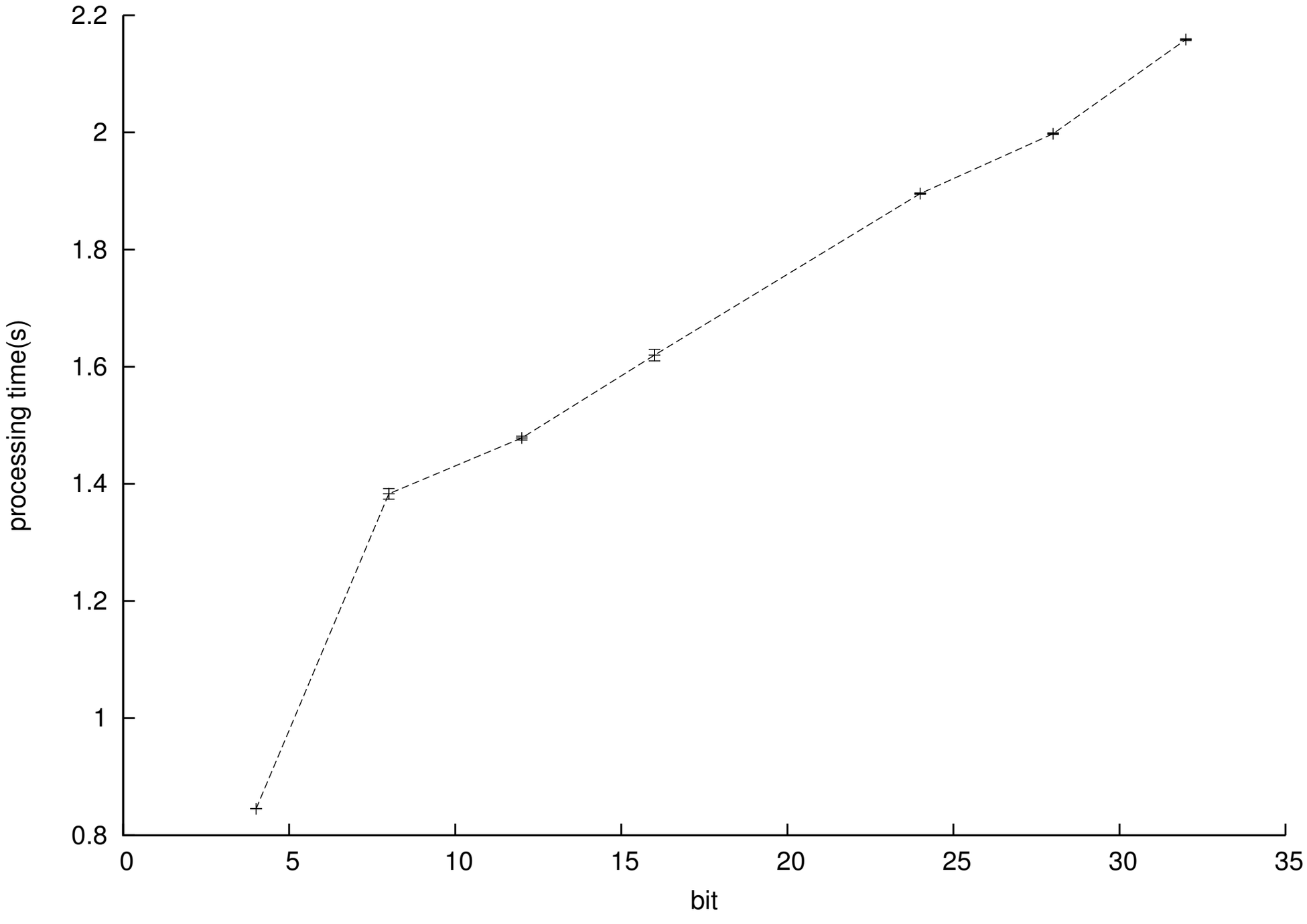
n = 140240, $\mu = 2^{30}$の時,

1/65000の確率で危険な領域に

実装したプログラムの処理時間



Fairplayによる評価の処理時間



実際の調査への適用

- 国立がん研究センターが行っている多目的コホート研究
  - 140,420名のデータを元に喫煙や飲酒のリスクなどを調査
  - 本プログラムを140,420名のデータに使用した所，約48分で処理を行えた
  - CPU 2.4GHz, メモリ4GB Java, BigIntegerクラス

おわりに

- 秘匿内積プロトコルとFairplayを用いて，２つのデータセットを秘匿した確率検定を行うシステムを実装した
- Fairplayの制約に対して，式を変形し加算,減算,比較のみを用いて処理を行った
- 乱数の大きさを抑えることで低下する安全性について評価を行った
  - $\mu = 2^{30}$, n = 140420の時，1/65000の確率で危険な領域に
- 今後の課題は多値や連続値なども扱えるシステムへの拡張

足し算から引き算への変更

従来[3]	本提案
$s_A = a - s_B$	$s_A = a + s_B$
$a = 20, s_B = 100$	$a = 20, s_B = 100$
$sA = 20 - 100 \bmod 256$ $= 48 \bmod 256$	$sA = 20 + 100 \bmod 256$ $= 120 \bmod 256$
$a = 48 + 100 \bmod 256$	$a = 120 - 100 \bmod 256$
Fairplayで正しくmodをとる必要性	