カイ二乗検定の幾何的解釈に基づく差分プライバシーの実現

柿崎 和也[1]　福地 一斗[1]　佐久間 淳[1,2]

概要：カイ二乗検定において，個人データをもとに算出された検定統計量から個人に関する機微な情報が推測される可能性がある．プライバシー保護指標の一つである差分プライバシーを保証することで，個人のプライバシーを保証し，カイ二乗検定を行うことが可能となる．本稿では，カイ二乗検定の幾何的解釈に基づいた，新しい差分プライベートなカイ二乗検定法について提案する．また，既存手法と提案手法の誤り率における理論的評価を行い，提案手法が既存手法より誤り率が小さくカイ二乗検定を行えることを示す．そして，人工データと実データを用いた実験により理論的評価の正しさ，現実的な設定において提案手法が既存手法より精度が良いことを示す．

キーワード：PWS，カイ二乗検定，差分プライバシー

1. はじめに

カイ二乗検定は仮説検定の一種であり，2 変数間の独立性を統計的保証をもって調べることができる．独立性の検定は，科学的発見のための非常に重要なタスクである．例としては，ゲノムワイド関連解析 (GWAS) において特定の疾患と強く関連する一塩基多型 (SNP) をカイ二乗検定をもとに同定している．

独立性カイ二乗検定におけるカイ二乗値は，二つの確率変数に関する分割表を用いて算出される．GWAS においては，一つの確率変数は，SNP の状態を 2 値の値でとり，もう一つの確率変数は疾患を持つ集団 (Case 群) と疾患を持たない集団 (Control 群) どちらに属するかを表す．GWAS における特定の疾患に対するデータベース集合 $\mathcal{D}$ は，N 人に対するデータベクトル x_i によって形成され，$\mathcal{D} = \{x_1, \cdots x_N\}, x_i \in \{0,1\}^M \times \{0,1\}$ と表される．ここで一つのインスタンスは各個人の情報を表し，M 個の SNP 情報と一つの疾患情報 (Case 群に属するか，Control 群に属するか) を属性として持つ．ここで $D \in \mathcal{D}$ において，ある j 番目の SNP に着目したとき，分割表は表 1 のようになる．その分割表のセルから式 (1) のようにカイ二乗値は算出される．

$$\chi^2(a,b,c,d) = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)} \qquad (1)$$

カイ二乗検定において，算出されたカイ二乗値と有意水準 α に対応するカイ二乗値のしきい値 τ と比較して，しきい値 τ より大きければ検定結果は有意に関連ありとなる．また，カイ二乗値と一対一の関係となる p 値と有意水準 α を比較することでも同様の検定結果を得ることができる．

表 1: 疾患の有無と SNP に対する分割表

	$Y=1$(Case)	$Y=0$(Control)	計
$SNP_j = 1$	a	b	n_1
$SNP_j = 0$	c	d	n_2
計	m_1	m_2	N

独立性検定において，検定統計量や分割表など検定に関する情報は論文などで公開されうる．基本的にそのような情報の公開における個人のプライバシー上の問題は考えづらい．しかしながら，特定の疾患と非常に多くの SNP の間で多数の検定を行う GWAS 設定においては，個人のプライバシー上の問題が生じることが知られている．

GWAS において，Homer らは攻撃者が被攻撃者の SNP データを保有し，一般の集団における対象疾患に関する対立遺伝子の頻度を知っている場合に，被攻撃者がその疾患をもった集団 (case) と疾患を持たない集団 (control) どちらに属するかを統計的に特定するような手法を提案している [4]．さらに，Wang らは p 値の公開においても個人のプライバシー上のリスクが存在することを示している [5]．この論文によると，GWAS において，攻撃者が被攻撃者の SNP データ，GWAS における分割表と SNP 間の相関を表す連鎖不平衡などのデータから，被攻撃者が GWAS のデータ内に含まれるかを高い信頼度で特定できる．GWAS において，相関表や連鎖不平衡は論文などで公開される情報であるため，p 値公開におけるプライバシー上のリスクは無

[1] 筑波大学 大学院システム情報工学研究科
Graduate School of SIE, University of Tsukuba
[2] 科学技術振興機構 CREST
Japan Science and Technology Agency

表 2: 各手法における仮定と β 誤り率のオーダー

各手法	分割表に対する仮定	balanced($m_1 = m_2 = \frac{N}{2}$)	unbalanced
Fienverg らの手法 [3]	m_1, m_2 が既知，$m_1 = m_2 = \frac{N}{2}$	$O(\exp(-\beta\epsilon))$	-
Yu らの手法 1[6]	m_1, m_2 が既知	$O(\exp(-\beta\epsilon))$	$O(\exp(-\frac{\beta\epsilon}{N}))$
Yu らの手法 2[6]	m_1, m_2 が既知，b, d が既知	$O(\exp(-\beta\epsilon))$	$O(\exp(-\frac{\beta\epsilon}{N}))$
提案手法	m_1, m_2 が既知	$O(\exp(-\epsilon\sqrt{\frac{\beta N}{\tau}}))$	$O(\exp(-\epsilon\sqrt{\frac{\beta}{\tau}}))$

視できない．これらの事実に基づき，NIH は GWAS 関連の統計量の一般公開を取りやめた経緯があった [7]．これは，p 値と一対一の関係にあるカイ二乗値などの検定統計量公開においても同様にプライバシー上のリスクが存在することを意味する．また，そのような p 値や検定統計量をもとに行われる仮説検定もプライバシー保護を考慮し行う必要性がある．

本稿では，カイ二乗検定における差分プライバシーの保証を目指す．差分プライバシーとは Dwork 等によって提唱されたプライバシーの定義である [1]．差分プライバシーを保証することで，統計量公開に対して個人のプライバシーを保護することができる．

1.1 関連研究

カイ二乗値の公開に対して差分プライバシー保証を行った研究は複数存在する．既存の研究では，どの手法も算出されたカイ二乗値に対して適切な分布に従ったノイズを加えることで差分プライバシーを保証する．

Fineberg らは Case 数 m_1 と Control 数 m_2 が公開情報であり，それぞれがサンプル数 N の半分である場合における，カイ二乗値公開においての差分プライバシー保証法について提案した [3]．この手法では Case 数と Control 数が同数でない場合に差分プライバシーを保証することができない．

Yu らは，Case 数 m_1，Control 数 m_2 が公開情報であるという仮定のもと，m_1，m_2 が同数でないときでも適用可能な差分プライベートなカイ二乗値の公開法を提案している [6]．また Yu らは，Case 数 m_1，Control 数 m_2 の公開に加えて，Control のセル数 b, d に対して情報が公開情報である場合に，より小さな分散となるノイズを用いて，同様の堅牢さとなる差分プライベートなカイ二乗値の公開法を提案している [6].

差分プライバシーを保証することによるカイ二乗値のエラーは，ノイズに使用する確率分布の分散によって決定される．そのような差分プライベートなカイ二乗値を用いたカイ二乗検定において，検定結果が差分プライバシー保証前の検定結果と比較して，どの程度誤るかの解析は既存研究ではなされていない．従ってどの手法がカイ二乗検定として優れているのかを比較することはできない．

1.2 貢献

本稿ではカイ二乗検定を幾何的に解釈することにより，新しい差分プライバシーを保証するカイ二乗検定の手法について提案する．既存手法ではカイ二乗値の公開に対して差分プライバシーを保証していたのに対して，提案手法では検定結果公開に対して差分プライバシーを保証する．また，提案手法では Case 数 m_1，Control 数 m_2 が公開情報であること以外の仮定を必要としない．

2 章で定めた有用性評価指標を用いて，4 章において，差分プライバシーを保証することによる，保証前の検定結果に対する誤り確率を理論的に評価し，既存手法との精度比較を行った．それぞれの手法の誤り確率のオーダーを表 2 に示す．分割表の m_1 と m_2 が等しい場合 (balanced)，提案法はサンプル数 N の増加に伴って誤り率がより小さくなり，サンプル数 N を増やすことで精度良く検定を行うことができる．これに対して，既存手法では，誤り率がサンプル数 N に依存していない．また，分割表の m_1 と m_2 に偏りが大きい場合 (unbalanced)，提案法はサンプル数 N に誤り率が依存しないのに対して，既存手法では増加に伴い，誤り率も増加することが分かる．これらのことから，balanced，unbalanced どちらの場合においても，サンプル数 N の増加に伴って既存手法が，提案手法より誤り率が小さくなることが期待される．

5 章で，これらの理論的評価の結果が正しいことを人工データを用いた実験により示し，実データを用いた現実的な設定で，提案手法が既存手法より良い精度でカイ二乗検定が行えることを示す．

2. 準備

2.1 差分プライバシー

差分プライバシーは Dwork らによって提唱されたプライバシー保護の定義である [1]．差分プライバシーは，個人情報が含まれるデータベースに対するクエリ問い合わせにおいて，データベース内の一つのインスタンスのメンバシップに対するプライバシーを保証する．データベース $D \in \mathcal{D}$ が与えられたとき，クエリはデータベースを引数とする関数 $f(D)$ で表される．$f(D)$ に対して差分プライバシーを保証しようとした場合，クエリ結果を直接返すのではなく，ランダムメカニズム $M(D)$ を介して結果を返す．このとき，差分プライバシーはメカニズム M に対して以下のように定義される．

定義 2.1 (ϵ 差分プライバシー [1]). メカニズム M が ϵ 差分プライベートであるとは，$h(D, D^{'}) = 1$ を満たす任意の $D, D^{'}$，任意のメカニズムの出力 s について，式 *(2)* を満たすことである．

$$\frac{Pr[M(D) = s]}{Pr[M(D') = s]} \leq e^{\epsilon} \tag{2}$$

ただし，$h(D, D^{'}) = 1$ はデータベース $D, D^{'}$ のハミング距離が 1 であることを表す．

ϵ はプライバシーパラメータと呼ばれ，プライバシー保護の指標となるものである．プライバシーパラメータが小さければ小さいほどより堅牢なプライバシーが保証される．

次に差分プライバシーを保証するメカニズムについて議論する．

2.2 ラプラスメカニズム

ラプラスメカニズム [1] は，ラプラス分布 $Lap(\lambda)$ に従って生成されたノイズをクエリ出力に加えることで ϵ 差分プライバシーを保証するメカニズムである．ここで分布の分散パラメータ λ は以下によって定義される Sensitivity Δ によって決定される．

定義 2.2 (Sensitivity). クエリ f の *Sensitivity* Δ は，式 *(3)* で定義される．

$$\Delta = \max_{D,D' \in \mathcal{D}: h(D,D')=1} ||f(D) - f(D')||_1 \tag{3}$$

Sensitivity Δ に基づくラプラスメカニズムを以下の定理に示す．

定理 2.1 (ラプラスメカニズム [1]). クエリ $f : \mathcal{D} \to \mathcal{R}$ に対する以下のメカニズム $M : \mathcal{D} \to \mathcal{R}$ は ϵ 差分プライベートである．

$$M(D) = f(D) + Y$$

ここで Y は $\lambda = \frac{\Delta}{\epsilon}$ のラプラス分布によって生成されるノイズである．

ラプラスメカニズムにより得られる結果と，データベースに依存しないしきい値の比較により得られる比較結果も同様に ϵ 差分プライバシーを保証する [2]．従って，ラプラスメカニズムにより得られる ϵ 差分プライベートなカイ二乗値としきい値 τ の比較結果も ϵ 差分プライバシーを保証する．差分プライバシーを保証するカイ二乗検定法を Algorithm 1 に示す．

定理 2.2. *Algorithm 1* の出力は ϵ 差分プライバシーを保証する．

証明. ラプラスメカニズムを用いているため明らかである． □

2.3 有用性評価指標

本節では，差分プライバシーを保証するカイ二乗検定手法の有用性を評価するための指標を定める．

Algorithm 1 Differential Private Chi-Squared Test

Require: Dataset D, the number of Case m_1, the number of Control m_2, threshold of chi-squared test τ ,privacy budget ϵ, sensitivity of chi-squared valueΔ
Ensure: binary
1: calculate contingency table cell (a, b) with Dataset D
2: calculate χ^2 with equation (1)
3: **if** $\chi^2 + Lap(\frac{\Delta}{\epsilon}) > \tau$ **then**
4: **return** 1(significant)
5: **else**
6: **return** 0(not significant)
7: **end if**

有用性は誤り確率によって評価する．$\chi^2(D)$ をデータベース $D \in \mathcal{D}$ を用いて算出されるカイ二乗値としたとき，ラプラスメカニズムを元にした差分プライベートなカイ二乗検定メカニズム M の誤り率を定義 2.3 のように定める．

定義 2.3 (誤り率). しきい値 $\tau > 0$，データベース $D \in \mathcal{D}$ が与えられたとき，メカニズム *M* の誤り確率を式 *(4)* とする．

$$E(M, D) = Pr[M(D) \neq I(\chi^2(D) > \tau)] \tag{4}$$

ここで I は指示関数であり，引数の命題が成り立てば 1 を，成り立たなければ *0* を返す関数である．

直感的にデータベース D から算出されるカイ二乗値 $\chi^2(D)$ がしきい値 τ に近いほどノイズの影響は大きくなり，式 4 の誤り確率は大きくなると考えられる．この近接性と誤り率の関係を明確化するために，β 誤り率を定義 2.4 のように定める．

定義 2.4 (β 誤り率). しきい値 $\tau > 0$，$\beta > 0$，サンプルサイズ N が与えられたとき，メカニズム M の β 誤り率を式 *(5)* とする．

$$\hat{E}(M, \beta) = \sup_{D \in \mathcal{D}: |\chi^2(D) - \tau| \geq \beta, |D| = N} E(M, D) \tag{5}$$

本稿では β 誤り率をもとに既存手法，提案手法の有用性評価を理論的に行う．

3. 幾何的解釈に基づくカイ二乗検定法

本章では，幾何的解釈に基づくカイ二乗検定法について説明していく．

表 1 のような分割表からカイ二乗値は式 (1) のように計算される．ここで Case m_1，Control m_2 の値が既知であると式 (1) は式 (6) のように変形され，(a, b) の関数となる．

$$\chi^2(a, b) = \frac{(am_2 - bm_1)^2 N}{(a + b)(N - a - b)m_1 m_2} \tag{6}$$

ここで，しきい値 τ におけるカイ二乗検定の幾何的解釈を補題 3.1 に示す．

補題 3.1. $\chi^2(a, b) = \tau$ を満たす実数 (a, b) の組は，式 *(7)* で表される (a, b) 平面上の楕円となる．また，$\chi^2(a, b) > \tau$ であることは，τ によって決定される式 *(7)* の楕円の外に (a, b) が位置することと等価である．

$$Aa^2 + Bb^2 + 2Cab + D(a+b) = 0 \tag{7}$$
$$\text{where} \quad A = (m_2^2 N + \tau m_1 m_2)$$
$$B = (m_1^2 N + \tau m_1 m_2)$$
$$C = m_1 m_2 (\tau - N)$$
$$D = -\tau m_1 m_2 N$$

証明. $\chi^2(a,b) = \tau$ のとき，式 (6) は式 (7) に変形され二次曲線となる．二次曲線は a^2, b^2, ab の係数を用いて，$AB - C^2 > 0$ なら楕円に決定される．今回，$AB - C^2$ は式 (8) のようになる．

$$\begin{aligned} AB - C^2 &= (m_2^2 N + \tau m_1 m_2)(m_1^2 N + \tau m_1 m_2) \\ &\quad - \{m_1 m_2 (\tau - N)\}^2 \\ &= \tau N m_1 m_2 (m_1 + m_2)^2 > 0 \end{aligned} \tag{8}$$

$\tau > 0, N > 0, m_1 > 0, m_2 > 0$ であることから，$AB - C^2 > 0$ であり，式 (7) は楕円である．

また，式 (7) の左辺を $f(a,b)$ とすると，式 (6) に対して，$\chi^2(a,b) \gtrless \tau$ が成り立つとき，式変形することにより，式 (9) のように $f(a,b) \gtrless 0$ が成り立つ．これは，$\chi^2(a,b) > \tau$ が成り立つ (a,b) は式 (7) の楕円の外に位置することを表す．また逆も同様に成り立つ．

$$\begin{aligned} &\chi^2(a,b) = \frac{(am_2 - bm_1)^2 N}{(a+b)(N-a-b)m_1 m_2} \gtrless \tau \\ \iff\ &f(a,b) \gtrless 0 \end{aligned} \tag{9}$$

□

補題 3.1 の結果により，しきい値 τ により決定される式 (7) の楕円と (a,b) の内包関係を調べることで，カイ二乗検定と同様の結果を得ることができる．楕円と点 (a,b) の内包関係は，楕円を原点を中心とした単位円へ変換するようなアフィン変換 T を用いて判断することができる．ここで，アフィン変換 T を用いた内包関係判別法を補題 3.2 に示す．

補題 3.2. 分割表セル (a,b) に対して，$||T((a,b)^t)||_2 > 1$ が成り立つならば，かつそのときに限り式 *(7)* の楕円の外に (a,b) が位置する．ここで T は式 *(10)* で定義されるアフィン変換関数であり, t は転置を表す．

$$\begin{aligned} &T((a,b)^t) \\ &= \begin{pmatrix} \sqrt{\frac{\lambda_1}{R}} & 0 \\ 0 & \sqrt{\frac{\lambda_2}{R}} \end{pmatrix} \left(\begin{pmatrix} \frac{C}{\sqrt{C^2+(\lambda_1-A)^2}} & \frac{(\lambda_1-A)}{\sqrt{C^2+(\lambda_1-A)^2}} \\ \frac{-(\lambda_1-A)}{\sqrt{C^2+(\lambda_1-A)^2}} & \frac{C}{\sqrt{C^2+(\lambda_1-A)^2}} \end{pmatrix} \right. \\ &\left. \begin{pmatrix} a \\ b \end{pmatrix} + \frac{D}{2\sqrt{C^2+(\lambda_1-A)^2}} \begin{pmatrix} \frac{C+\lambda_1-A}{\lambda_1} \\ \frac{C+\lambda_2-B}{\lambda_2} \end{pmatrix} \right) \end{aligned} \tag{10}$$

λ_1，λ_2 は行列 $\begin{pmatrix} A & C \\ C & B \end{pmatrix}$ の固有値であり，R は式 *(11)* のような値である．

$$R = \frac{D^2\Big(\lambda_2(C+\lambda_1-A)^2 + \lambda_1(C+\lambda_2-B)^2\Big)}{4\lambda_1\lambda_2\Big(C^2+(\lambda_1-A)^2\Big)} \tag{11}$$

証明は A.1 を参照.

ここで，補題 3.1，3.2 の結果より，幾何的解釈に基づくカイ二乗検定法を定理 3.1 に示す．

定理 3.1. しきい値 $\tau > 0$，*Case* 数 m_1，*Control* 数 m_2，サンプル数 N の分割表において，$||T((a,b)^t)||_2 > 1$ が成り立つならばカイ二乗検定の結果は有意に関連ありとなる．ここで T は式 *(10)* により定義されるアフィン変換関数である．

証明. 補題 3.1，3.2 の結果により，$||T((a,b)^t)|| > 1$ が成り立つならばそのときに限り，$\chi^2(a,b) > \tau$ が成り立つ．従って，$||T((a,b)^t)|| > 1$ が成り立つならばカイ二乗検定の結果は有意に関連ありとなる． □

4. 提案手法

本章では，幾何的解釈に基づいたカイ二乗検定法に対して，差分プライバシーを保証することで，新しい差分プライバシーを保証した新しいカイ二乗検定法を提案する．

4.1 幾何的解釈に基づくカイ二乗検定の差分プライバシー保証

定理 3.1 で，$||T((a,b)^t)||_2$ と 1 の比較により，カイ二乗検定と同様の結果が得られることを示した．

そこで Algorithm1 と同様に，$||T((a,b)^t)||_2$ にラプラスメカニズムを用いることによって，差分プライバシーを保証しながらカイ二乗検定を行うことができる．ラプラスメカニズムのために $||T((a,b)^t)||_2$ の Sensitivity を導出する．$||T((a,b)^t)||_2$ の Sensitivity は定理 4.1 のようになる．

定理 4.1 ($||T((a,b)^t)||_2$ の Sensitivity)**.** *Case* 数 m_1，*Control* 数 m_2 の 2×2 の分割表から算出されるカイ二乗値 $\chi^2(a,b)$ としきい値 τ のカイ二乗検定において，$||T((a,b)^t)||_2$ の *Sensitivity* Δ は式 *(12)* となる．

$$\Delta_T = 2\sqrt{\frac{(m_1^2+m_2^2)N + 2\tau m_1 m_2}{\tau m_1 m_2 N^2}} \tag{12}$$

証明は A.2 を参照.

定理 4.1 の結果とラプラスメカニズムを使用することで，差分プライバシーを保証したカイ二乗検定を行う手法を Algorithm2 に示す．Algorithm 2 は Laplace メカニズムを用いているため ϵ 差分プライバシーを保証する．

定理 4.2. *Algorithm 2* は ϵ 差分プライベートである．

証明. ラプラスメカニズムを用いているので明らかである． □

4.2 有用性評価

本節では，既存の Sensitivity を用いた手法 Algorithm1 と，提案手法である Algorithm2 の有用性評価を行い，理論的な

Algorithm 2 Differential Private Chi-Squared Test

Require: Dataset D, the number of Case m_1, the number of Control m_2, threshold of chi-squared test τ ,privacy budget ϵ
Ensure: binary
1: calculate contingency table cell $(a,b)^t$ with Dataset D
2: calculate $||T((a,b)^t)||_2$ with equation (10)
3: calculate Δ_T with equation (12)
4: **if** $||T((a,b)^t)||_2 + Lap(\frac{\Delta}{\epsilon}) > 1$ **then**
5: **return** 1
6: **else**
7: **return** 0
8: **end if**

精度比較を行う．

有用性評価の指標としては定義 2.4 の β 誤り率の上限により評価を行う．β 誤り率は，差分プライバシーを保証することによる，元の検定結果との誤り確率であり，小さな値ほど精度は良いと判断される．

カイ二乗値の Sensitivity に Δ を使用した Algorithm1 を $M_{1\Delta}$ とすると，$M_{1\Delta}$ の β 誤り率の上限は定理 4.3 のようになる．

定理 4.3. *$\tau > 0$，$\beta > 0$，$\epsilon > 0$ とする．カイ二乗値の Sensitivity を Δ，Case 数 m_1，Control 数 m_2 とするサンプル数 N の分割表に対して，Algorithm 1 の β 誤り率の上限は式 (13) のようになる．*

$$\hat{E}(M_{1\Delta}, \beta)) \leq \frac{1}{2}\exp\left(-\frac{\beta\epsilon}{\Delta}\right) \tag{13}$$

証明は A.3 を参照.

ここで Fienberg，Yu らによる既存のカイ二乗値の Sensitivity 解析結果 [3][6] を式 (14)，(15)，(16) に示す．各 Sensitivity は 1.1 節において説明した仮定をそれぞれ必要とする．

$$\Delta_F = \frac{4N}{N+2} \tag{14}$$

$$\Delta_{Y1} = \frac{N^2}{m_1 m_2}\left(\frac{\max\{m_1, m_2\}}{\max\{m_1, m_2\}+1}\right) \tag{15}$$

$$\Delta_{Y2} = \frac{N^2}{m_1 m_2}\left(\frac{\max\{b, d\}}{\max\{b, d\}+1}\right) \tag{16}$$

定理 4.3 をふまえて，Algorithm1 に Sensitivity Δ_F，Δ_{Y1}，Δ_{Y2} を用いた場合の β 誤り率の上限は命題 4.1 のようになる．

命題 4.1. *定理 4.3 と同じ条件において，カイ二乗値の Sensitivity に $\Delta_F, \Delta_{Y1}, \Delta_{Y2}$ を用いた場合，Algorithm1 の β 誤り率の上限は式 (17)，(18)，(19) のようになる．*

$$\hat{E}(M_{1\Delta_F}, \beta) \leq \frac{1}{2}\exp\left(-\frac{\beta\epsilon(N+2)}{4N}\right) \tag{17}$$

$$\hat{E}(M_{1\Delta_{Y1}}, \beta) \leq \frac{1}{2}\exp\left(-\frac{\beta\epsilon(m_1 m_2)}{N^2}\frac{\max\{m_1, m_2\}+1}{\max\{m_1, m_2\}}\right) \tag{18}$$

$$\hat{E}(M_{1\Delta_{Y2}}, \beta) \leq \frac{1}{2}\exp\left(-\frac{\beta\epsilon(m_1 m_2)}{N^2}\frac{\max\{b, d\}+1}{\max\{b, d\}}\right) \tag{19}$$

また，Sensitivity に Δ_T を用いる Algorithm 2 を $M_{2\Delta_T}$ とすると，$M_{2\Delta_T}$ の β 誤り率の上限は定理 4.4 のようになる．

定理 4.4. *$\tau > 0$，$\beta > 0$，$\epsilon > 0$，カイ二乗値の Sensitivity を Δ_T，Case 数 m_1，Control 数 m_2，サンプル数 N の分割表に対して，Algorithm 2 の β 誤り率の上限は式 (20) のようになる．*

$$\hat{E}(M_{2\Delta_T}, D) \leq \frac{1}{2}\exp\left(\frac{\epsilon N}{2}\left(1-\sqrt{1+\frac{\beta}{\tau}}\right)\sqrt{\frac{\tau m_1 m_2}{(m_1^2+m_2^2)N + 2\tau m_1 m_2}}\right) \tag{20}$$

証明は A.4 を参照.

それぞれの β 誤り率のオーダーは表 2 のようになる．Case 数と Control 数が同数，$m_1 = m_2$ (balanced) の場合，既存手法では，β 誤り率は N に依存しない．しかし，提案法ではサンプル数 N の増加に伴って，β 誤り率は小さくなっていくため有用である．また，Case 数と Control 数に偏りがある (unbalanced) な場合，例えばサンプル数 N に対して Case 数 $m_1 = 1$，Control 数 $m_2 = N-1$ の場合は提案手法の β 誤り率はサンプル数 N に依存していない．しかし，既存手法では N の増加に伴って β 誤り率が増加していくことが分かる．従って，サンプル数 N が大きくなるにつれて既存手法と比較して，提案手法の方が β 誤り率が小さくなることが期待される．

β 誤り率を用いた理論的評価の結果，Case 数と Control 数が同数，Case 数と Control 数に偏りがある場合どちらにおいても，サンプル数を増加していくことにより．提案法が既存手法よりも誤り率が小さくカイ二乗検定を行えることが期待できる．

5. 実験

人工データと実データを用いて，提案法と既存手法の精度比較を行う．

5.1 人工データによる実験

人工データを用いることにより，β 誤り率を用いた精度比較の結果の正しさを示す．人工データは Case 数と Control 数が同数，$m_1 = m_2$ なデータ (balanced) と，Case 数と Control 数に偏りがあるデータ (unbalanced) それぞれにおける各手法の誤り率を測る．

5.1.1 balanced データ

サンプル数 N に対して，Case 数 $m_1 = \frac{N}{2}$ と Control 数 $m_2 = \frac{N}{2}$ と等しいデータにおける，提案手法と既存手法の精度比較を行う．サンプル数を $N = 2^2$ から $N = 2^{25}$ まで指数的に変化させ，各サンプル数に対してカイ二乗値が $\chi^2 = \{1,2,3,4,5,6,7,8,9,10\}$ となるように 10 個の分割表を生成した．これらの分割表による 10 検定を用いて評価を行う．評価は 10 個の分割表に対する 10 検定における間違えた割合，誤り率で評価を行う．誤り率は，FP を false positive，TP を true positive，FN を false negative，FP を

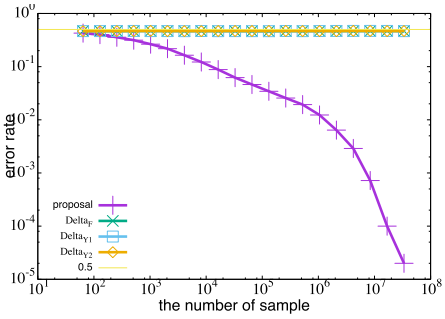

図 1: balanced データにおける誤り率

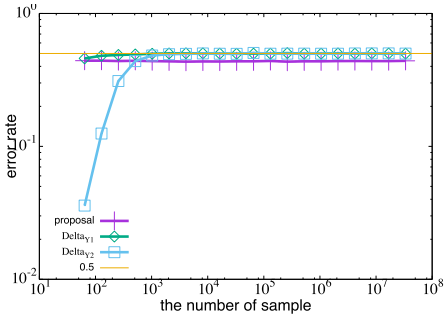

図 2: unbalanced データにおける誤り率

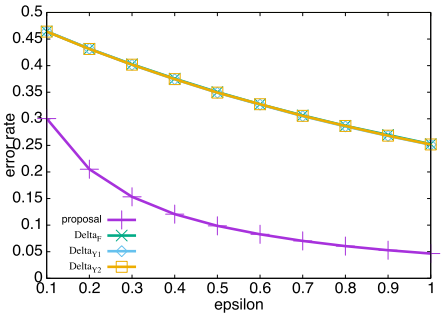

図 3: 実データにおける誤り率

false positive とすると $\frac{FP+FN}{TP+TN+FP+FN}$ となる．しきい値は有意水準 0.05 に相当する $\tau = 3.84$ を用い，プライバシーパラメータ $\epsilon = 0.1$ とした，10000 回平均を示す．

図 1 に人工データ (balanced) による実験の結果を示す．横軸はサンプル数を表し，縦軸は誤り率を表す．既存手法がサンプル数の増加に伴って誤り率に変化がないのに対して，提案手法は指数的に誤り率が小さくなっていくことが分かる．また，$N = 2^{24}, 2^{25}$ においては誤り率は 10^{-4} 以下となり非常に小さな誤り率で検定が行えていることが分かる．

5.1.2 unbalanced データ

サンプル数 N に対して，Case 数 $m_1 = 2$ と Control 数 $m_2 = N - 2$ と偏ったデータにおける，提案手法と既存手法の精度比較を行う．このデータは Case 数と Control 数が異なるデータであるため，SensitivityΔ_F は用いることができない．したがって，Δ_{Y1} と Δ_{Y2} を用いた手法と，提案手法の精度比較を行う．プライバシーパラメータが $\epsilon = 1.0$ でそれ以外の設定は，balanced データと同様の設定で評価を行った．

図 2 に人工データ (unbalanced) による実験の結果を示す．横軸はサンプル数を表し，縦軸は誤り率を表す．結果は図 2 のようになった．提案手法はサンプル数の増加に伴って誤り率に変化がないのに対して，既存手法は，誤り率が増加していくことが分かる．

5.2 実データによる実験

実データを用いて既存手法と提案手法の比較を行う．用いたデータは IDASH PRIVACY & SECULITY WORKSHOP 2015，SECURE GENOME ANALYSUS COMPETITION[*1] のデータである．このデータはある疾患と 311 の SNP に関するデータで，サンプル数 $N = 400$ Case 数 $m_1 = 200$ Control 数 $m_2 = 200$ となるデータである．また，有意水準 0.05 で 311SNP と疾患の有無に関してカイ二乗検定を行った場合に有意となる SNP は 63 個であった．

評価は誤り率 $\frac{FP+FN}{TP+TN+FP+FN}$ で行い，プライバシーパラメータ $\epsilon = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ それぞれに対して 10000 回平均を算出した．

[*1] http://www.humangenomeprivacy.org/2015/competition-tasks.html

図 3 に実データによる結果を示す．横軸はプライバシーパラメータ ϵ を表し，縦軸は誤り率を表す．いずれの ϵ においても既存手法と比較して，提案手法の誤り率が低いことが分かる．また，$\epsilon \geq 0.5$ で 311 検定に対して誤り率 10^{-1} 以下でカイ二乗検定が行えている．

6. おわりに

本研究では，幾何的解釈に基づくカイ二乗検定に対する差分プライバシー保証法を提案した．β 誤り率を定義し，既存手法と提案手法における β 誤り率の理論的な評価を行った．また，人工データを用いて理論的結果の正しさを確かめ，実データを用いて現実的な設定において提案手法が既存手法より精度良くカイ二乗検定が行えることを示した．

謝辞 本研究は，JST CREST「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域におけるプロジェクトおよび科学研究費 24680015，16H02864 の助成を受けました．

参考文献

[1] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pp. 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[2] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, Vol. 9, No. 3-4, pp. 211–407, 2014.

[3] Stephen E. Fienberg, Aleksandra Slavkovic, and Caroline Uhler. Privacy preserving gwas data sharing. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pp. 628–635, Washington, DC, USA, 2011. IEEE Computer Society.

[4] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, Vol. 4, No. 8, p. e1000167, 2008.

[5] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM conference on Computer and communications security*, pp. 534–544. ACM, 2009.

[6] Fei Yu, Stephen E Fienberg, Aleksandra B Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of biomedical informatics*, Vol. 50, pp. 133–141, 2014.

[7] Elias A Zerhouni and Elizabeth G Nabel. Protecting aggregate genomic data. *Science*, Vol. 322, No. 5898, p. 44a, 2008.

付　　録

A.1 補題 3.2 の証明

証明. アフィン変換関数 T が，式 (7) で表される楕円を原点を中心とする単位円に変換することを示す．

式 (7) は行列を用いて式 (A.1) のように書き換えることができる．

$$(a,b)\begin{pmatrix}A & C\\ C & B\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix}+D\,(1,1)\begin{pmatrix}a\\ b\end{pmatrix}=0 \qquad \text{(A.1)}$$

ここで行列 $\begin{pmatrix}A & C\\ C & B\end{pmatrix}$ の固有値は

$$\lambda_1,\lambda_2=\frac{(A+B)\pm\sqrt{(A+B)^2-4AB+4C^2}}{2}$$

となり，そのときの固有ベクトルは λ_1,λ_2 に対してそれぞれ $\begin{pmatrix}1\\ \frac{\lambda_1-A}{C}\end{pmatrix}$ と $\begin{pmatrix}\frac{-(\lambda_1-A)}{C}\\ 1\end{pmatrix}$ となる．それらを正規化して並べた正規直行行列は

$$P=\begin{pmatrix}\frac{C}{\sqrt{C^2+(\lambda_1-A)^2}} & \frac{-(\lambda_1-A)}{\sqrt{C^2+(\lambda_1-A)^2}}\\ \frac{\lambda_1-A}{\sqrt{C^2+(\lambda_1-A)^2}} & \frac{C}{\sqrt{C^2+(\lambda_2-A)^2}}\end{pmatrix}$$

となり，$\begin{pmatrix}a\\ b\end{pmatrix}=P\begin{pmatrix}a_1^{'}\\ b_2^{'}\end{pmatrix}$ とおき式 (A.1) を変形していくと，式 (A.2) のようになる．

$$\begin{aligned}&\lambda_1\left(a_1^{'}+\frac{D}{2\lambda_1\sqrt{c^2+(\lambda_1-A)^2}}(C+\lambda_1-A)\right)^2\\ &+\ \lambda_2\left(b_1^{'}+\frac{D}{2\lambda_2\sqrt{c^2+(\lambda_1-A)^2}}(C+\lambda_2-B)\right)^2\\ &=\frac{D^2\Big(\lambda_2(C+\lambda_1-A)^2+\lambda_1(C+\lambda_2-B)^2\Big)}{4\lambda_1\lambda_2\Big(C^2+(\lambda_1-A)^2\Big)}\end{aligned} \qquad \text{(A.2)}$$

(A.2) 式において，$a_1^{'}+\frac{D}{2\lambda_1\sqrt{c^2+(\lambda_1-A)^2}}(C+\lambda_1-A)=a_2^{'}$，$b_1^{'}+\frac{D}{2\lambda_2\sqrt{c^2+(\lambda_1-A)^2}}(C+\lambda_2-B)=b_2^{'}$ とすることで並行移動変換を行う．また，R を式 (11) とし，$a_2^{'}=\sqrt{\frac{R}{\lambda_1}}a_3^{'}$，$b_2^{'}=\sqrt{\frac{R}{\lambda_2}}b_3^{'}$ とするような拡大縮小変換を行う．これにより，式 (A.2) は式 $a_3^{'2}+b_3^{'2}=1$ のように変形され，$a_3^{'},b_3^{'}$ 平面上で原点を中心とし，半径を 1 とする単位円となる．

楕円から原点を中心とした単位円に変換するまでの変換式を用いて (a,b) と (a_3,b_3) の関係式を考えると式 (A.3) の関係式が成り立つ．

$$\begin{aligned}&\begin{pmatrix}a_3\\ b_3\end{pmatrix}\\ &=\begin{pmatrix}\sqrt{\frac{\lambda_1}{R}} & 0\\ 0 & \sqrt{\frac{\lambda_2}{R}}\end{pmatrix}\left(\begin{pmatrix}\frac{C}{\sqrt{C^2+(\lambda_1-A)^2}} & \frac{(\lambda_1-A)}{\sqrt{C^2+(\lambda_1-A)^2}}\\ \frac{-(\lambda_1-A)}{\sqrt{C^2+(\lambda_1-A)^2}} & \frac{C}{\sqrt{C^2+(\lambda_1-A)^2}}\end{pmatrix}\right.\\ &\left.\begin{pmatrix}a\\ b\end{pmatrix}+\frac{D}{2\sqrt{C^2+(\lambda_1-A)^2}}\begin{pmatrix}\frac{C+\lambda_1-A}{\lambda_1}\\ \frac{C+\lambda_2-B}{\lambda_2}\end{pmatrix}\right)\end{aligned} \qquad \text{(A.3)}$$

従って，アフィン変換関数 T は式 (10) のようになる． □

A.2 Sensitivity の証明

証明. Case 数 m_1，Control 数 m_2 の表 1 のような分割表について考えると，データセット D から生成される分割表を SNP=1 のセル (a,b) で表現すると，近隣データセット $D^{'}$ から生成される分割表は $(a+1,b),(a-1,b),(a,b+1),(a,b-1)$ の 4 通りが考えられる．ここで $(a+1,b)$ に変化する際の $||T((a,b)^t)||_2$ の最大変化量式 (A.4) について考える．

$$\max_{a,b}|||T((a+1,b)^t)||_2-||T((a,b)^t)||_2| \qquad \text{(A.4)}$$

ここで，まず $||T((a+1,b)^t)-T((a,b)^t)||_2$ と $|||T((a+1,b)^t)||_2-||T((a,b)^t)||_2|$ の大小比較について考えていく．

$$\begin{aligned}&||T((a+1,b)^t)-T((a,b)^t)||_2^2-|||T((a+1,b)^t)||_2\\ &-\ ||T((a,b)^t)||_2|^2\\ &=\ ||T((a+1,b)^t)-T((a,b)^t)||_2^2-||T((a+1,b)^t)||_2^2\\ &-\ ||T((a,b)^t)||_2^2+2||T((a+1,b)^t)||_2||T((a,b)^t)||_2\end{aligned} \qquad \text{(A.5)}$$

$T((a+1,b)^t)$ と $T((a,b)^t)$ が作る鋭角を θ とすると，余弦定理により，$||T((a+1,b)^t)-T((a,b)^t)||_2^2-||T((a+1,b)^t)||_2^2-||T((a,b)^t)||_2^2=-2||T((a+1,b)^t)||_2||T((a,b)^t)||_2cos\theta$ が成り立つ．そのため式 (A.5) に対して式 (A.6) が成り立つ．

$$\begin{aligned}&||T((a+1,b)^t)-T((a,b)^t)||_2^2-||T((a+1,b)^t)||_2^2\\ &-\ ||T((a,b)^t)||_2^2+2||T((a+1,b)^t)||_2||T((a,b)^t)||_2\\ &=\ 2||T((a+1,b)^t)||_2||T((a,b)^t)||_2(1-cos\theta)\geq 0\end{aligned} \qquad \text{(A.6)}$$

従って，$||T((a+1,b)^t)-T((a,b)^t)||_2\geq|||T((a+1,b)^t)||_2-||T((a,b)^t)||_2|$ が成り立つことが分かる．
また，$T((a+1,b)^t)-T((a,b)^t)$ は式 (A.7) となり，$T((a+1,b)^t)-T((a,b)^t)$ は $(a,b)^t$ に依存しない．

$$\begin{aligned}&T((a+1,b)^t)-T((a,b)^t)\\ &=\begin{pmatrix}\sqrt{\frac{\lambda_1}{R}} & 0\\ 0 & \sqrt{\frac{\lambda_2}{R}}\end{pmatrix}\begin{pmatrix}\frac{C}{\sqrt{C^2+(\lambda_1-A)^2}} & \frac{(\lambda_1-A)}{\sqrt{C^2+(\lambda_1-A)^2}}\\ \frac{-(\lambda_1-A)}{\sqrt{C^2+(\lambda_1-A)^2}} & \frac{C}{\sqrt{C^2+(\lambda_1-A)^2}}\end{pmatrix}\begin{pmatrix}1\\ 0\end{pmatrix}\\ &=\begin{pmatrix}\sqrt{\frac{\lambda_1}{R}}\frac{C}{\sqrt{C^2+(\lambda_1-A)^2}}\\ \sqrt{\frac{\lambda_2}{R}}\frac{-(\lambda_1-A)}{\sqrt{C^2+(\lambda_1-A)^2}}\end{pmatrix}\end{aligned} \qquad \text{(A.7)}$$

そのため式 (A.4) に対して，式 (A.8) が成り立つ．

$$\begin{aligned}&\max_{a,b}|||T((a+1,b)^t)||_2-||T((a,b)^t)||_2|\\ &\leq\ ||T((a+1,b)^t)-T((a,b)^t)||_2\\ &=\sqrt{\frac{1}{C^2+(\lambda_1-A)^2}\left(\left(C\sqrt{\frac{\lambda_1}{R}}\right)^2+\left(-(\lambda_1-A)\sqrt{\frac{\lambda_2}{R}}\right)^2\right)}\\ &\leq\sqrt{\left(\sqrt{\frac{\lambda_1}{R}}\right)^2+\left(\sqrt{\frac{\lambda_2}{R}}\right)^2}=\sqrt{\frac{\lambda_1+\lambda_2}{R}}\end{aligned} \qquad \text{(A.8)}$$

式 (A.8) は，他の (a,b) に対する変化 $(a-1,b),(a,b+1),(a,b-1)$ においても成り立つ．従って，$||T((a,b)^t)||_2$ の SensitivityΔ は，$\Delta=\sqrt{\frac{\lambda_1+\lambda_2}{R}}$ となる．

ここで $\sqrt{\frac{\lambda_1+\lambda_2}{R}}$ を m_1,m_2,τ,N の式に変形する．λ_1,λ_2 は行列 $\begin{pmatrix}A & C\\ C & B\end{pmatrix}$ の固有値であるので，$\lambda_1+\lambda_2=A+B$，$\lambda_1\lambda_2=AB-C^2$，固有方程式より $(\lambda_1-A)(\lambda_1-B)=C^2$ が成り立つ．これらの関係式を用いて R は式 (A.9) のように変形される．

$$R=\frac{(\tau m_1m_2N)^2\Big(\lambda_2(C+\lambda_1-A)^2+\lambda_1(C+\lambda_2-B)^2\Big)}{4\lambda_1\lambda_2\Big(C^2+(\lambda_1-A)^2\Big)}$$

$$
\begin{aligned}
&= \frac{(\tau m_1 m_2 N)^2}{4}\left(\frac{(C+\lambda_1-A)^2}{\lambda_1(C^2+(\lambda_1-A)^2)}\right.\\
&+ \left.\frac{(C-\lambda_1+A)^2}{\lambda_2(C^2+(\lambda_1-A)^2)}\right)\\
&= \frac{(\tau m_1 m_2 N)^2}{4}\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{2C(\lambda_1-A)}{\lambda_1(C^2+(\lambda_1-A)^2)}\right.\\
&+ \left.\frac{-2C(\lambda_1-A)}{\lambda_2(C^2+(\lambda_1-A)^2)}\right)\\
&= \frac{(\tau m_1 m_2 N)^2}{4}\left(\frac{\lambda_1+\lambda_2}{\lambda_1\lambda_2}+\frac{-2C(\lambda_1-A)(\lambda_1-\lambda_2)}{\lambda_1\lambda_2(C^2+(\lambda_1-A)^2)}\right)\\
&= \frac{(\tau m_1 m_2 N)^2}{4}\left(\frac{A+B}{AB-C^2}\right.\\
&+ \left.\frac{-2C(\lambda_1-A)(\lambda_1-\lambda_2)}{(AB-C^2)(\lambda_1-A)(2\lambda_1-A-B)}\right)\\
&= \frac{(\tau m_1 m_2 N)^2}{4}\left(\frac{A+B-2C}{AB-C^2}\right)\\
&= \frac{(\tau m_1 m_2 N)^2}{4}\frac{(m_1+m_2)^2 N}{\tau m_1 m_2 (m_1+m_2)^2 N}\\
&= \frac{\tau m_1 m_2 N^2}{4} \qquad \text{(A.9)}
\end{aligned}
$$

式 (A.9) の結果を用いると式 (A.8) を m_1, m_2, τ, N の式に変形でき，$||\mathbf{v}_{(a,b)}||_2$ の SensitivityΔ は式 (A.10) のようになる．

$$
\begin{aligned}
\Delta &= \sqrt{\frac{\lambda_1+\lambda_2}{R}}\\
&= 2\sqrt{\frac{(m_1^2+m_2^2)N+2\tau m_1 m_2}{\tau m_1 m_2 N^2}} \qquad \text{(A.10)}
\end{aligned}
$$

□

A.3 定理 4.3 の証明

証明. カイ二乗値の Sensitivity に Δ を使用した Algorithm1 を $M_{1\Delta}$ とする．ここで $\beta > 0$，しきい値 $\tau > 0$ に対して有意となるデータベース集合 $\{D|\chi^2(D) = \tau + \beta\}$ に対して，式 (4) の誤り率 $E(M_{1\Delta}, D)$ は式 (A.11) のようになる．

$$
\begin{aligned}
E(M_{1\Delta}, D) &= \Pr[\chi^2(D) + Lap(\frac{\Delta}{\epsilon}) \le \tau]\\
&= \Pr[Lap(\frac{\Delta}{\epsilon}) \le -\beta]\\
&= \frac{\epsilon}{2\Delta}\int_{-\infty}^{-\beta}\exp\Big(\frac{x\epsilon}{\Delta}\Big)dx\\
&= \frac{1}{2}\exp(\frac{-\beta\epsilon}{\Delta}) \qquad \text{(A.11)}
\end{aligned}
$$

式 (A.11) は β の増加に伴い小さくなっていく．従って，β 誤り率の上限は，$\hat{E}(M_{1\Delta}, \beta) \le \frac{1}{2}\exp(\frac{-\beta\epsilon}{\Delta})$ となる．

$\beta > 0$，しきい値 $\tau > 0$ に対して有意とならないデータベース集合 $\{D|\chi^2(D) = \tau - \beta\}$ についても同様の議論で，β 誤り率の上限は $\hat{E}(M(\Delta), \beta) \le \frac{1}{2}\exp(\frac{-\beta\epsilon}{\Delta})$ となる． □

A.4 定理 4.4 の証明

証明. Sensitivity に Δ_T を使用する Algorithm2 を $M_{2\Delta_T}$ とする．ここで $\beta > 0$，しきい値 $\tau > 0$ に対して有意となるデータベース集合 $\{D|\chi^2(D) \ge \tau + \beta\}$ に含まれるデータベースにより生成される分割表をセル (a_1, b_1) で表すとする．その時，式 (4) の誤り率 $E(M_{2\Delta_T}, \beta)$ は式 (A.12) のようになる．

$$
\begin{aligned}
E(M_{2\Delta_T}, D) &= \Pr[||T((a_1,b_1)^t)||_2 + Lap(\frac{\Delta_T}{\epsilon}) \le 1]\\
&= \Pr[Lap(\frac{\Delta_T}{\epsilon}) \le 1 - ||T((a_1,b_1)^t)||_2]\\
&= \frac{\epsilon}{2\Delta_T}\int_{-\infty}^{1-||T((a_1,b_1)^t)||_2} exp\Big(\frac{x\epsilon}{\Delta_T}\Big)dx\\
&= \frac{1}{2}\exp(\frac{(1-||T((a_1,b_1)^t)||_2)\epsilon}{\Delta_T}) \qquad \text{(A.12)}
\end{aligned}
$$

ここで β 誤り率の上限を求めるため，$||T((a_1,b_1)^t)||_2$ の下限について考える．

式 (7) の左辺を τ, a, b の関数 $f(\tau, a, b)$ とすると，集合 $\{(a,b)|f(\tau+\beta, a, b) \ge 0\}$ は $\chi^2(a,b) \ge \tau + \beta$ が成り立つ (a,b) の集合である．ここで式 (A.2) の左辺を τ, a, b の関数 $E(\tau, a, b)$，$g(a,b) = m_1 m_2(a^2+b^2) - m_1 m_2 N(a+b) + 2m_1 m_2 ab$ とおき，式 (11) の R をもちいて $f(\tau+\beta)$ は式 (A.13) のように書き換えられる．

$$
\begin{aligned}
f(\tau+\beta, a, b) &= f(\tau, a, b) + \beta g(a, b)\\
&= E(\tau, a, b) - R + \beta g(a, b) \qquad \text{(A.13)}
\end{aligned}
$$

ここで，$\frac{E(\tau,a,b)}{R} = ||T((a_1,b_1)^t)||_2^2$ であることから，集合 $\{(a,b)|f(\tau+\beta,a,b) \ge 0\}$ の要素 (a,b) に対して式 (A.14) の関係が成り立つ．

$$
\begin{aligned}
&E(\tau,a,b) - R + \beta g(a,b) \ge 0\\
&\frac{E(\tau,a,b)}{R} \ge 1 - \frac{\beta g(a,b)}{R}\\
&||T((a,b)^t)||_2^2 \ge 1 - \frac{\beta g(a,b)}{R} \qquad \text{(A.14)}
\end{aligned}
$$

ここで，$1 - \frac{\beta g(a,b)}{R}$ は，式 (A.15) のように変形される．

$$
\begin{aligned}
&1 - \frac{\beta g(a,b)}{R}\\
&= 1 - 4\beta m_1 m_2 \frac{a^2+b^2-Na-Nb+2ab}{\tau m_1 m_2 N^2}\\
&= 1 - 4\beta\frac{(a+b)(a+b-N)}{\tau N^2} \ge 1 + \frac{\beta}{\tau} \qquad \text{(A.15)}
\end{aligned}
$$

従って，集合 $\{(a,b)|f(\tau+\beta,a,b) \ge 0\}$ の要素 (a,b) に対して $||T((a,b)^t)||_2 \ge \sqrt{1+\frac{\beta}{\tau}}$ が成り立つ．

また，$(a_1,b_1) \in \{(a,b)|f(\tau+\beta) \ge 0\}$ であるから，$||T((a_1,b_1)^t)||_2 \ge \sqrt{1+\frac{\beta}{\tau}}$ が成り立つ．

$||T((a_1,b_1)^t)||_2$ の下限と式 (A.12) により，$M_{2\Delta_T}$ の β 誤り率 $\hat{E}(M_{2\Delta_T}, D)$ の上限は式 (A.16) のようになる．

$$
\hat{E}(M_{2\Delta_T}, D) \le \frac{1}{2}\exp(\frac{(1-\sqrt{1+\frac{\beta}{\tau}})\epsilon}{\Delta}) \qquad \text{(A.16)}
$$

ここで $||T((a_1,b_1)^t)||_2$ の SensitivityΔ は式 (12) であるため，下限は式 (A.17) のようになる

$$
\begin{aligned}
\hat{E}(M_{2\Delta_T}, D) \le \frac{1}{2}\exp\Bigg(&\frac{\epsilon N}{2}\Big(1-\sqrt{1+\frac{\beta}{\tau}}\Big)\\
&\sqrt{\frac{\tau m_1 m_2}{(m_1^2+m_2^2)N+2\tau m_1 m_2}}\Bigg) \qquad \text{(A.17)}
\end{aligned}
$$

次に，$\beta > 0$，しきい値 $\tau > 0$ に対して有意とならないデータベース集合 $\{D|\chi^2(D) \le \tau - \beta\}$ について考えると，同様の導出により β 誤り率 $\hat{E}(M_{2\Delta_T}, D)$ は式 (A.17) と同様の結果が得られる． □