東海大学大学院2012年度 修士論文

プライバシーを保護した疫学調査のための確率検定プロトコル

Privacy-Preserving Hypothesis Test Protocol for Epidemiology

指導教員 菊池 浩明 教授

東海大学大学院 工学研究科 情報理工学専攻

1BDRM015 佐藤 智貴

目 次

第1章　序論

1.1　背景

疫学調査とは，疾病の発生原因と思われる因子と疾病の因果関係を統計的に調査し，明らかにする研究である．例えば，喫煙者のがん罹患率と非喫煙者のがん罹患率を調査し，比較することで，喫煙することで変化するがん罹患に対する相対的な危険度を求めることができる [1]．また，放射線業務従事者について，累積線量群別のがん罹患率や死亡率，死因などを調べることで，低線量域での被ばくが人体に有意な健康影響を齎さないことを明らかにすることができる [2]．このように，様々な団体で因子，疾病について疫学調査が行われているが，比較するべき因子と疾病が二つの組織に分散して管理されていることが多々ある．例えば，[2] の場合，中央登録センターの持つ放射線事業従事者リストと，厚生労働省の持つ国民の死因リストを照合する必要性がある．また，調査に必要な情報は，病歴や被ばく線量，喫煙歴，身長，体重，BMI など，調査対象者のプライバシーに深く関わる情報なため，プライバシー保護に十分注意しなければならないという問題がある．さらに，調査の際には何十年という長期に渡り追跡調査を受けることもあり，プライバシー保護を理由に調査対象者が追跡を断ることも少なくない [2] という問題点がある

1.2　目的

本研究では，疫学調査におけるプライバシーに関する問題に対して，暗号技術を用いることで，調査に必要な情報を秘匿したまま疫学調査を行うことにより，調査対象者のプライバシーを守り，より頻度が高く，様々な要因との相互関係を考慮した，精度の高い疫学調査の実現を試みる．

そこで，実際に行われている疫学調査における統計的手法に対して，プライバシーを保護したまま安全に行う手法を提案する．本研究の主な成果は，1. プライバシーを保護した放射線疫学調査プロトコルの提案，2. プライバシーを保護した患者-対照調査における相対危険度の有意性検定プロトコルの提案，3. プライバシーを保護した傾向性の検定プロトコルの提案．の 3 つである．

1.3 論文構成

本論文の構成は次の通りである．第 2 章で既存の疫学調査の研究成果や手法について説明し，第 3 章で本稿で扱う要素技術について説明し，第 4 章でプライバシーを保護した放射線疫学調査プロトコルを，第 5 章でプライバシーを保護した相対危険度の検定プロトコルを，第 6 章でプライバシーを保護した傾向性の検定プロトコルを提案し，第 7 章で本論文の結論を述べる．

第2章　疫学調査

2.1　放射線事業従事者に対する疫学調査

(財) 放射線影響協会は 1990 年度から原子力発電施設等の放射線業務従事者を対象とした疫学的調査を実施しており，“原子力発電施設等放射線業務従事者等に係る疫学調査” [2] で調査報告を行っている．本稿では，まずこれらの実際に行われている疫学調査の調査報告について考察し，疫学調査におけるプライバシーの課題を挙げ，要求条件を明らかにする．

2.1.1　文献 [2] の概要 (文献 [2] より引用)

調査対象

調査の対象者数は，1999 年 3 月 31 日までに原子力事業者等から (財) 放射線影響協会放射線従事者中央登録センターへ登録され，実際に放射線業務に従事した日本人の男女，合計約 27 万 7 千人である．生死の確認は，市区町村長から調査対象者の住民票の写し等の交付を受けて確認している．調査対象者のうち，2009 年 3 月 31 日まで，男女合計約 21 万 2 千人の生死を確認できており，残りの約 6 万 5 千人については住所情報を収集できなかった等の理由で生死を確認できていない．

外部比較

「外部比較」では解析対象者の死亡率が，全日本人男性死亡率に比べて高いか否かを検討するため，標準死亡比（SMR = 観察死亡数 / 期待死亡数）を求めている．また，SMR が 1 に等しいかどうかについて両側検定を行い，p 値が 0.05 未満のときは有意であると判断している．表 2.1 は文献 [2] から引用している．

内部比較

「内部比較」では解析対象者を年度別被ばく線量の累積値により 5 群に分類し，累積線量の増加に伴って死亡率が増加する傾向があるかについて片側検定を行い，p 値が 0.05 未満のときは有意であると判断している．表 2.2 は文献 [2] から引用している．

[1)] 死因を同定できなかった 80 名を含む。

表 2.1: 表 3.3-1 死因別標準化死亡比 (SMR)(前向き観察、最短潜伏期;0 年、年齢、暦年を調整)(文献 [2] より引用)

死因	観察死亡数	期待死亡数	SMR	95%信頼区間	両側検定結果 p 値
全死因[1)]	14,224	14,086.9	1.01	(0.99 - 1.03)	0.250
食道	326	312.1	1.04	(0.93 - 1.16)	0.449
胃	1,002	989.4	1.01	(0.95 - 1.08)	0.700
肺	1,208	1,117.8	1.08	(1.02 - 1.14)	0.007

表 2.2: 表 3.4-1 死因別累積線量群別 O/E 比および傾向性の検定結果 (1) (前向き観察、最短潜伏期; 白血病 2 年 その他の新生物 10 年、年齢、暦年、地域を調整)(文献 [2] より引用)

	累積線量群 (mSv)					
	< 10	10−	20−	50−	100+	
	観察死亡数 期待死亡数 O/E 比	観察死亡数 期待死亡数 O/E 比	観察死亡数 期待死亡数 O/E 比	観察死亡数 期待死亡数 O/E 比	観察死亡数 期待死亡数 O/E 比	傾向性の 片側検定結果 p 値
死因	95%信頼区間	95%信頼区間	95%信頼区間	95%信頼区間	95%信頼区間	
全死因[2)]	10,315	1,408	1,434	639	428	
	10,515.5	1,287.5	1,343.6	652.7	424.8	
	0.98	1.09	1.07	0.98	1.01	0.136
	(0.96 - 1.00)	(1.04 - 1.15)	(1.01 - 1.12)	(0.90 - 1.06)	(0.91 - 1.11)	
全悪性新生物[3)]	3,822	494	526	245	124	
	3,902.6	475.0	488.9	225.3	119.1	
	0.98	1.04	1.08	1.09	1.04	0.032
	(0.95 - 1.01)	(0.95 - 1.14)	(0.99 - 1.17)	(0.96 - 1.23)	(0.87 - 1.24)	
食道	200	29	32	20	8	
	215.3	26.4	27.3	12.9	7.1	
	0.93	1.10	1.17	1.55	1.12	0.039
	(0.80 - 1.07)	(0.73 - 1.58)	(0.80 - 1.66)	(0.95 - 2.40)	(0.48 - 2.21)	
胃	669	85	85	41	18	
	674.4	81.3	83.8	38.2	20.3	
	0.99	1.05	1.01	1.07	0.89	0.532
	(0.92 - 1.07)	(0.84 - 1.29)	(0.81 - 1.25)	(0.77 - 1.45)	(0.53 - 1.40)	

2.1.2 既存の疫学調査における課題

既存の疫学調査には，プライバシーの関係で，

1. 特殊な法律が必要
2. 従事者の同意が必要
3. 情報の粒度や鮮度が不十分

などの問題点がある．

2.2 多目的コホート研究

独立行政法人 国立がん研究センターでは，1990 年から，日本人に適した予防医学実践のための科学的根拠の材料となるエビデンス作りを目的とし，多目的コホート研究を行っている [1]．140,420 名の対象者について，喫煙による影響，飲酒による影響，体系による影響，身体活動の影響，食事の影響等を調査している．[1] の研究成果の一部を図 2.1 に示す．図 2.1 は多目的コホート研究 [1] より引用している．

2.3 患者-対照調査

患者-対照調査とは，ある特定の疾病の患者群と非患者群について，ある要因に暴露していたか否かを調べ，因果関係を研究する調査である．表 2.4 の患者-対照調査のデータが与えられた時，

$$RR = \frac{a}{n_1} / \frac{c}{n_2} = \frac{a(c+d)}{(a+b)c} \approx \frac{ad}{bc} \tag{2.1}$$

を相対危険度 (relative risk) と呼ぶ．相対危険度とは，特定要因へ暴露した群が，暴露しなかった群に比べて，何倍の危険度を有するかを表す指標である [3]．相対危険度が大きいものほど因果関係が強い．この相対危険度 RR が 1 に等しいかどうかを検定する．推定された相対危険度の有意性は，統計量 $\chi = \frac{\sqrt{N-1}\{(ad-bc)\pm N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}$ が $RR = 1$ の仮定の元，標準正規分布 $N(0,1)$ に従うか否かで検定することができる．本稿では，両側検定の有意水準 95%，すなわち χ が $Z(0.05/2) = 1.960$ を上回っているか否かで判定を行う．

[2] 死因を同定できなかった 80 名を含む。
[3] 白血病を含め最短潜伏期 10 年とした。

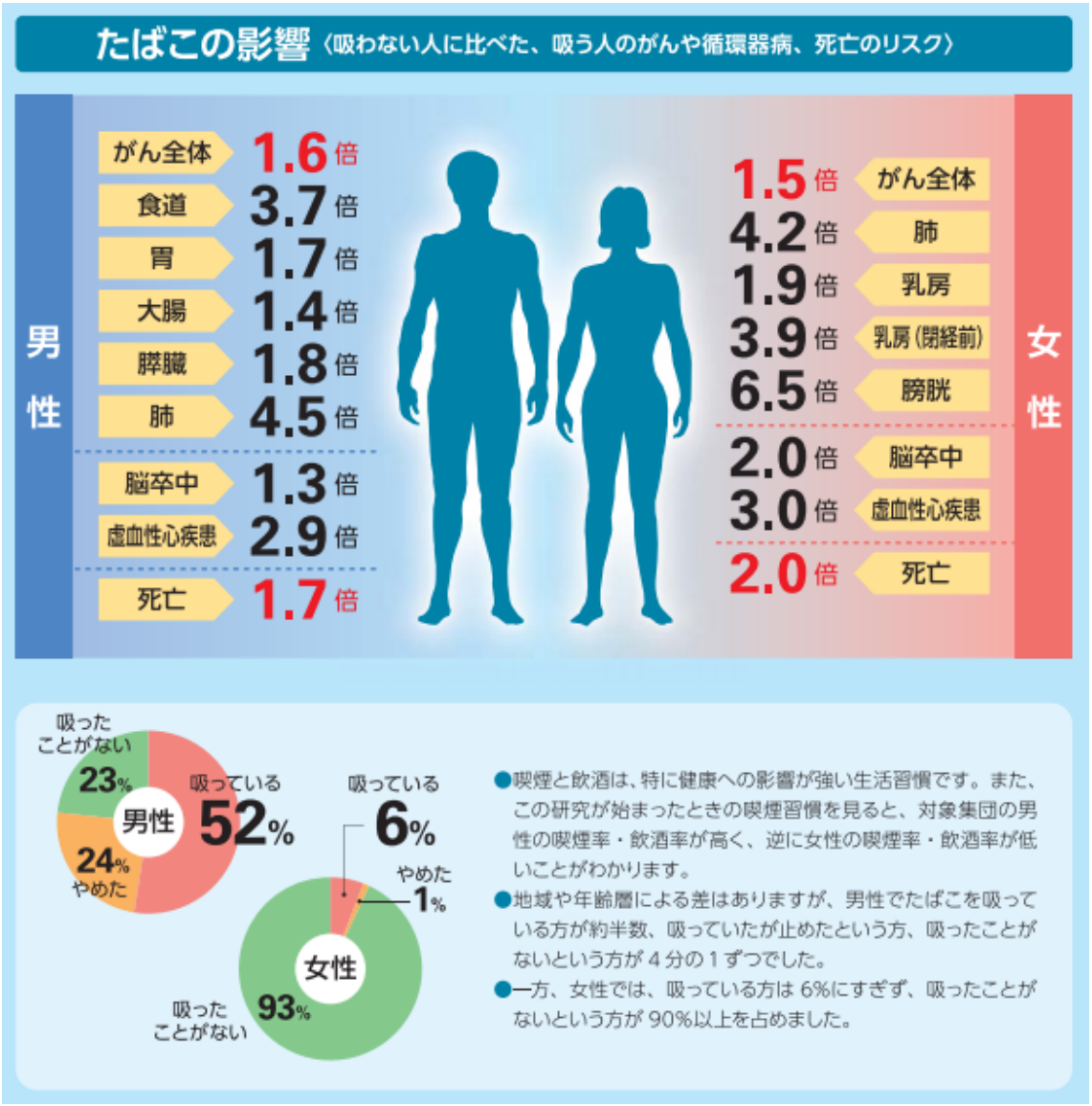

図 2.1: 多目的コホート研究の成果 ([1] より引用)

2.4 傾向性の検定 (用量-反応関係の検出)

傾向性の検定とは，臨床試験等である薬剤の効果を検討するために，その用量の大きさをいくつかの群に分けて実験を行う．その際に用量の大きさによってその反応が増加するか否かを検定することを傾向性の検定と呼ぶ．この場合の検定仮説は，反応の計量値を μ とした場合，

帰無仮説 H_0 : $\mu_1 = \mu_2 = \cdots = \mu_a$,

対立仮説 H_1 : $\mu_1 \leqq \mu_2 \leqq \cdots \leqq \mu_a$

となる (上昇傾向).

表 2.3: 患者-対照調査における母集団のデータ

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

表 2.4 のデータ例を考えよう．表 2.4 はラットを 4 群に分け，それぞれ濃度の異なる薬物を投与した後，一定期間経過した後の赤血球数を測定したものである [3]．傾向性の検定は，

表 2.4: 傾向性の検定におけるデータ例

	A 群	B 群	C 群	D 群
x_i 用量	10ppm	100ppm	1000ppm	10000ppm
y_i 反応	8.06	7.97	7.66	8.00
	8.27	7.66	7.71	7.89
	8.45	8.05	7.88	7.79
	8.51	8.30	8.05	7.91
	8.14	8.03	7.80	7.40
平均	8.286	8.002	7.820	7.798

$y = \alpha + \beta \log(x)$ の回帰分析が可能であれば，この傾き β について $H_0 : \beta = 0$, $H_1 : \beta < 0$(または$\beta > 0$) の片側検定と考えることができる．すなわち，回帰分析によって求めた，図 2.2 のような回帰直線の傾き β がその変動に対して有意であるか否かで傾向性を検定する．この例の場合の傾き β は減少傾向となる．

表 2.4 の用量を $x_i (i = 1, \cdots, n)$，それに対する反応を $y_i (i = 1, \cdots, n)$，用量の平均を $\overline{x}$，反応の平均を $\overline{y}$ とすると，β の推定値 $\hat{\beta}$ は

$$\hat{\beta} = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} \tag{2.2}$$

となる．推定した切片 $\hat{\alpha}$ は

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \tag{2.3}$$

で求めることができる．この推定した $\hat{\beta}$ の有意性を求めるために，推定値 $\hat{y_i}$ と実測値 y_i と

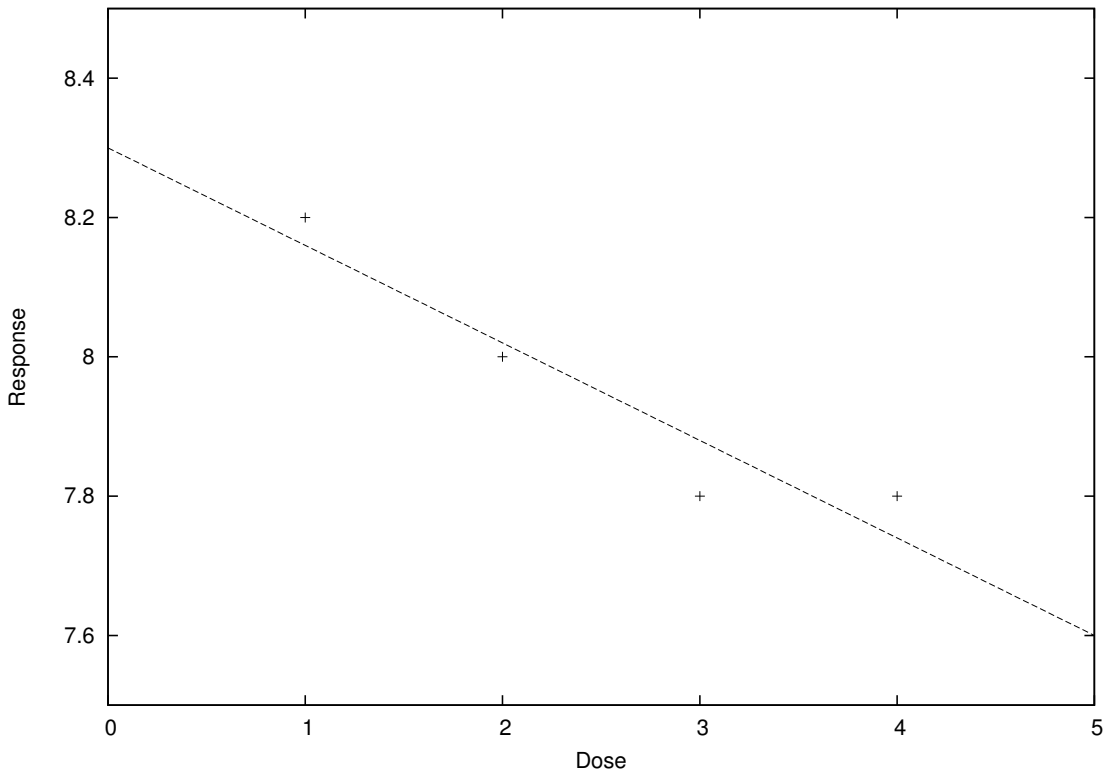

図 2.2: 表 2.4 についての回帰直線

の差の平方和 V_E

$$\begin{aligned} V_E &= \sum_i^n (y_i - \hat{y_i})^2 \cdot \frac{1}{n-2} \\ &= \sum_i^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \cdot \frac{1}{n-2} \\ &= \frac{1}{n-2} \cdot \left(SS_Y - \frac{(SS_{XY})^2}{SS_X} \right) \end{aligned}$$

を求める．ここで,

$$SS_X = \sum_i^n (x_i - \overline{x})^2,$$

$$SS_Y = \sum_i^n (y_i - \overline{y})^2,$$

$$SS_{XY} = \sum_i^n (x_i - \overline{x})(y_i - \overline{y})$$

とする．それにより $\hat{\beta}$ の標準誤差

$$s.e.(\hat{\beta}) = \frac{\sqrt{V_E}}{\sqrt{\sum_i^n (x_i - \overline{x})}} \tag{2.4}$$

を与える．ここで，統計量

$$t = \frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})} \tag{2.5}$$

が自由度 $n-2$ の t 分布に従うか否かで検定を行う．この時，β は仮説に使用する定数 0 とする．

図 2.3 と図 2.4 の例を見てみよう．図 2.3 の傾きは -0.01，図 2.4 の傾きは -0.14 である．図 2.3 のような効果のある薬の場合，薬の用量を増やしてもあまり効果が出ていないため，薬としての効果は薄い．図 2.4 の傾きは図 2.2 と同じで，用量を増やすほど効果が現れているが，分散が大きいため，次に計測した時に同じ傾きになるとは限らない．そのため，傾向性の検定を行う際には，薬の効果である傾き $\hat{\beta}$ だけでなく，その有意性の検定も行うことが重要になる．

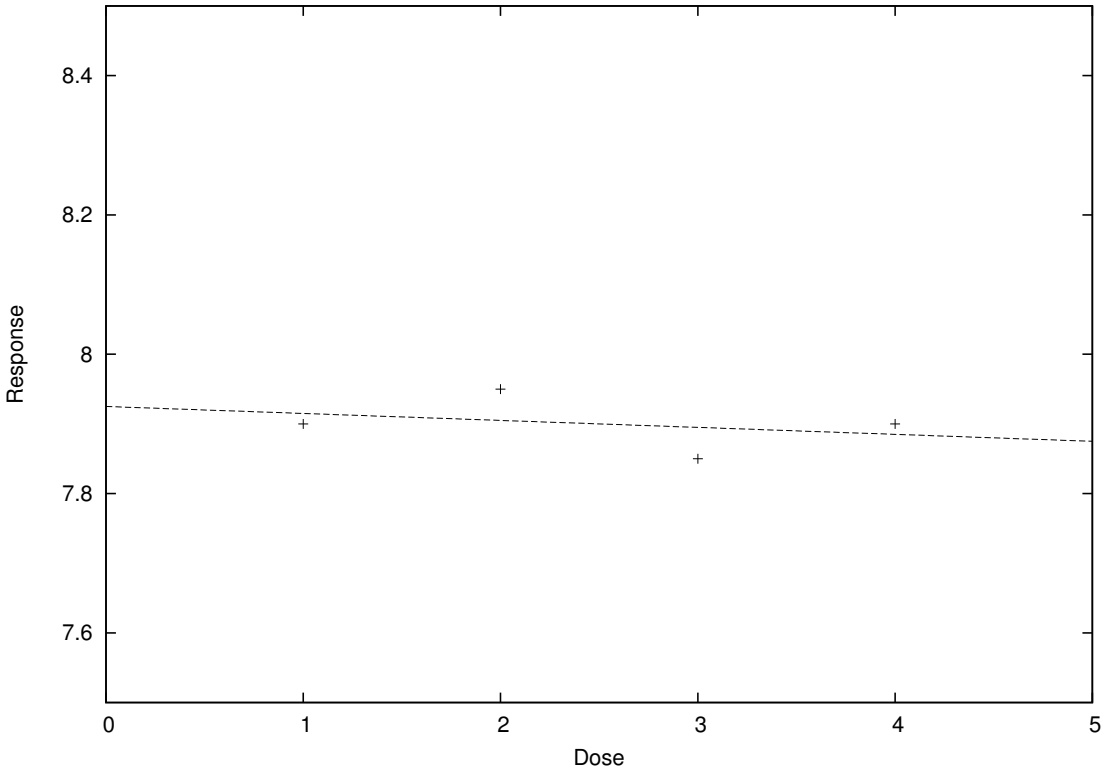

図 2.3: $\hat{\beta}$ が小さい例

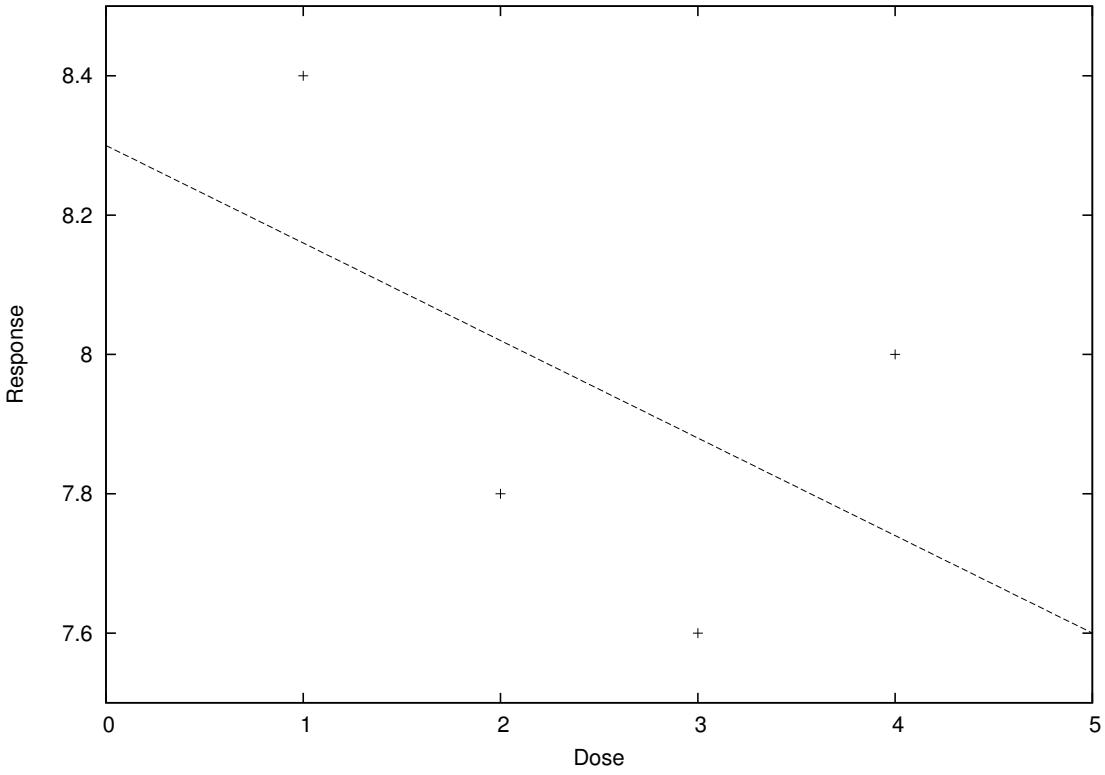

図 2.4: $\hat{\beta}$ が有意でない例

第3章　従来研究

3.1　可換な一方向性関数 (AES03)

AES03([4]) は Agrawal らによって提案された．2 つの集合 X と Y を，お互いに開示することなく，2 つの集合の共通集合 $X \cap Y$，または，共通集合の要素数 $|X \cap Y|$ を求めることができる．AES03 のアルゴリズムを Algorithm 1 に示す．

Algorithm 1 AES03[4](可換一方向性関数)

入力: 集合 $X = \{x_1, \ldots, x_{n_A}\}$ を持つ A と $Y = \{y_1, \ldots, y_{n_B}\}$ を持つ B.
出力: $|X \cap Y|$ を求める．
位数 q の巡回群 G と G を値域とするハッシュ関数 H を考える．

1. A は，乱数 $u \in Z_q$ を選び，$H(x_1)^u \ldots, H(x_{n_A})^u$ を B へ送る．
2. B は，乱数 $v \in Z_q$ を選び，$H(y_1)^v, \ldots, H(y_{n_B})^v$ と $H(x_1)^{uv}, \ldots, H(x_{n_A})^{uv}$ を求めて A へシャッフルして送る．
3. A は，$H(y_i)^{vu} = H(x_j)^{uv}$ を満たす x_j, y_i の組の個数 $(= |X \cap Y|)$ を求める．

3.2　秘匿内積プロトコル

秘匿内積プロトコル [5] を，Algorithm 2 に示す．2 つの組織がそれぞれ持つ X_A と X_B を秘匿したまま，積集合の大きさ $|X_A \cap X_B|$ のみを求める．計算結果は，s_A+s_B=$|X_A \cap X_B|$ となるような 2 つの乱数 s_A と s_B に分散されるため，計算が終わっても秘匿されている．

3.3　Fairplay

Yao により提案された秘密関数計算 (Secure Function Evaluation(SFE)[7]) プロトコルを用いることで互いの要素を秘匿したまま任意の計算を行うことができる．SFE は，AND や OR の論理ゲートレベルで，2 者間での分散評価を行うため，その回路サイズが小規模なものに制約されるが，任意の関数が秘密に評価できる．例えば，s_A と s_B を入力すると，

Algorithm 2 秘匿内積プロトコル

入力: Alice は n 次元ベクトル $\boldsymbol{x} = (x_1, \ldots, x_n)$ を持つ．Bob は n 次元の $\boldsymbol{y} = (y_1, \ldots, y_n)$ を持つ．

出力: Alice と Bob は $s_A + s_B = \boldsymbol{x} \cdot \boldsymbol{y}$ となるような s_A，s_B を得る．ここで，暗号文の定義域を Z_n とする．

1. Alice は準同型暗号の公開鍵対を作り，公開鍵を Bob に送る．
2. Alice は Bob に暗号文 $E(x_1), \ldots, E(x_n)$ を送る．
3. Bob は s_B を Z_n からランダムに選び，

   $$c = E(x_1)^{y_1} \cdots E(x_n)^{y_n} / E(s_B)$$

   を計算し，Alice に送る．
4. Alice は c を復号し，$s_A = D(c) = x_1 y_1 + \cdots + x_n y_n - s_B$ を得る．

$s_A + s_B > t$ を評価してその 1 bit の結果を出力する回路を構成することで，A と B のどちらもその和が分からないままで，有意水準を超えるか否かのみが分かる．

Fairplay は，Malkhi らによって開発された SFE の処理系である [8]．高級言語風のソースから，回路記述言語 SFDL を出力し，それに基づいて SFE を実行する．ただし，SFE の性質上，その機能には制約があり，例えば，乗算はプリミティブで用意されていないため，加算を組み合わせて書かなくてはならない．そのため，Fairplay は加算，減算，大小比較等の基本的な計算は高速にできるが，乗算，除算等の計算は極めて遅い．加算と乗算の処理時間の違いを図 3.1 に示す．加算は $t_A + t_B > \theta$ の比較，乗算は $t_A \cdot t_B > \theta$ を各々実行した時の処理時間である．それ故，Fairplay においては乗算の利用を極力避けなくてはならない．

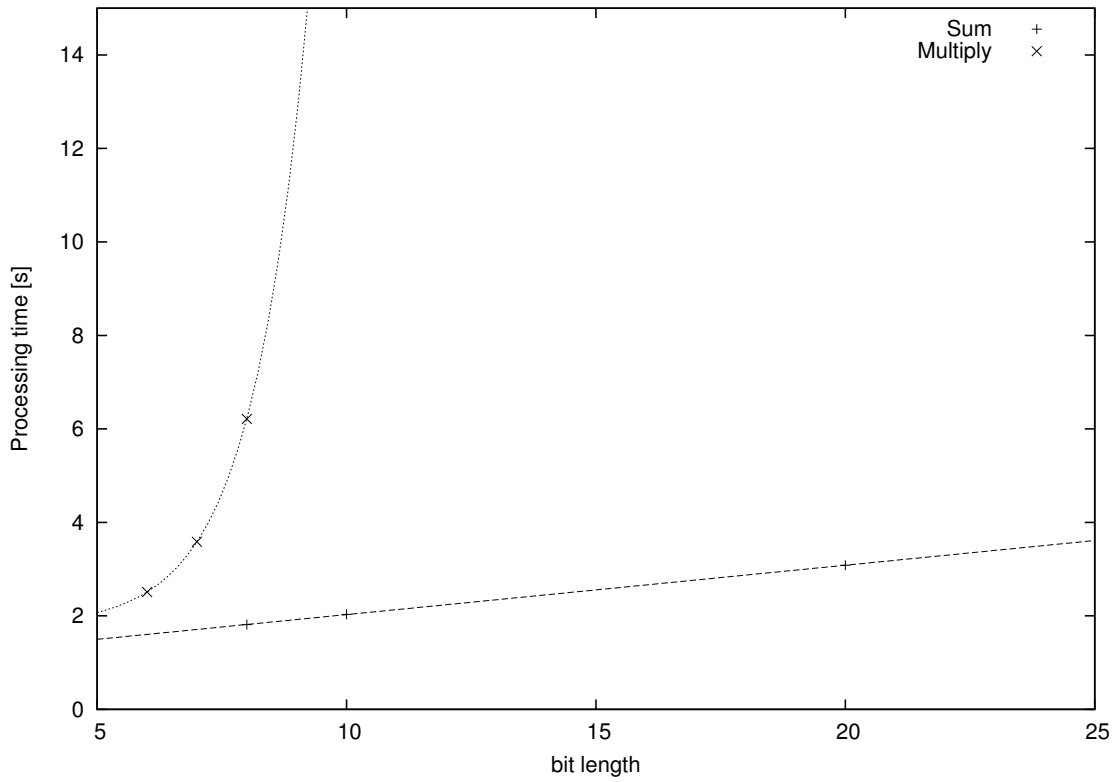

図 3.1: Fairplay における加算と乗算による分散比較の計算時間 [9]

第4章　プライバシーを保護した放射線疫学調査プロトコル

4.1　概要

放射線事業従事者に対する低線量域での放射線の健康への影響を調査することが重要となってきている．“原子力発電施設等放射線業務従事者等に係る疫学調査”[2] については，2.1 節で概要を説明したが，この調査報告によると，低線量域での健康への影響は，中央登録センターの持つ放射線事業従事者リストと，厚生労働省の持つ国民の死因リストを照合することで求められる．しかし，各リストは各組織で個別に管理されており，プライバシーの問題で互いに照合することは難しい．追跡調査される被験者の同意も必要であり，プライバシー保護を理由に長期の追跡を断ることも少なくない．現状の調査では，死亡したデータから比較調査をしているが，その前に様々な悪性新生物への発病や転移の状況も分からない．

この疫学調査の問題に対して，暗号プロトコルの適用を提案する．暗号技術により，被験者のプライバシーを守って，より頻度の高い，様々な要因との相互作用を考慮した，精度の高い疫学調査の実現を試みる．本研究では，過程に応じて 2 種類の問題設定を行い，それぞれに適したプロトコルを提案する．前者は，Agrawal らの提案した AES(Agrawal-Evfimievski-Srikant)03 プロトコル [4] を用いることで，データを秘匿したまま 2 つのリストを照合する．また，文献 [2] で行われている，内部比較と外部比較の調査結果に基づいて，提案方式の実現可能性を評価する．この評価の為に，提案方式を Java を用いて試験実装した．その性能に基づいた実現可能性評価の結果について報告する．

4.2　問題定義

秘匿の集合 $X_A \subset U$ を持つ組織 A，$X_B \subset U$ を持つ組織 B が協力して，疫学調査を行う．ここで，U は対象者の全体集合とする．例えば，放射線疫学調査の場合は，A は放射線従事者中央登録センターであり，B は (1) 人口動態調査死亡書「死亡テープ」を有する厚生労働省や，(2) 生存するがん患者のカルテを有する地域がんセンターが該当する．U は，調査全期間 (例えば 15 年間) の全人口から成る集合である．X_A には，年齢別に分割できる属性があり，例えば $X_A = X_{A,30} \cup X_{A,40} \cup \ldots \cup X_{A,80}$ と分割できるとする．この疫学調査の目的は，X_A における死亡率やがん罹患率が標準的な期待死亡率に対して，有意な差があるか判

定することにある．

4.3 検定方法

死亡や疾病などの様に，一定期間に独立に生じる事象の数は，ポアソン過程と見なせることがよく知られている．ある事象の発生数 X が期待値 λ へのポアソン分布に従う時，k 回生起する確率は

$$P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!} \tag{4.1}$$

で与えられる．

一定期間内に生じる事象の発生数，例えば死亡数の期待値 E は，

$$E=\sum_{j}^{m} d_j n_j \tag{4.2}$$

で与えられる．ここで，n_j は対象となる年令階級 j の人口であり，d_j は，j における死亡率を表す．n_j は組織 A が有しており，文献 [2] の例では表 4.1 の様になる．一方，d_j の一般

表 4.1: A の年齢分布

年齢	人数	割合 (%)
30 - 34	29,264	14.4
35 - 39	42,791	21.0
40 - 44	37,039	18.2
45 - 49	43,907	21.5
50 - 54	33,804	16.8
55 - 59	17,099	8.4

的な死亡率については，文献 [10] における，「2-26 年齢別死亡数及び死亡率」のように，公開されているものも多いが，未知の病気の患者リストを有する病院の例の様に，組織 B のみが有することとする．[2] における A の年齢階級別期待死亡率を図 4.1 に示す．A の年齢分布は 30 から 60 歳に渡っており，45 歳台が最頻値だが，加齢に応じて死亡率が高いので，図 4.1 では年齢に応じて単調に増加した分布が示されている．

なお，疫学調査は 10 年以上に渡って何度も行われるが，この年齢の分布もそれに応じて右へシフトしていく．

発生 (死亡) 数 $X=O$ が観測された時，有意性を判断する評価尺度として，標準化死亡比 : Standardized Mortality Ratio を

$$SMR=\frac{O}{E} \tag{4.3}$$

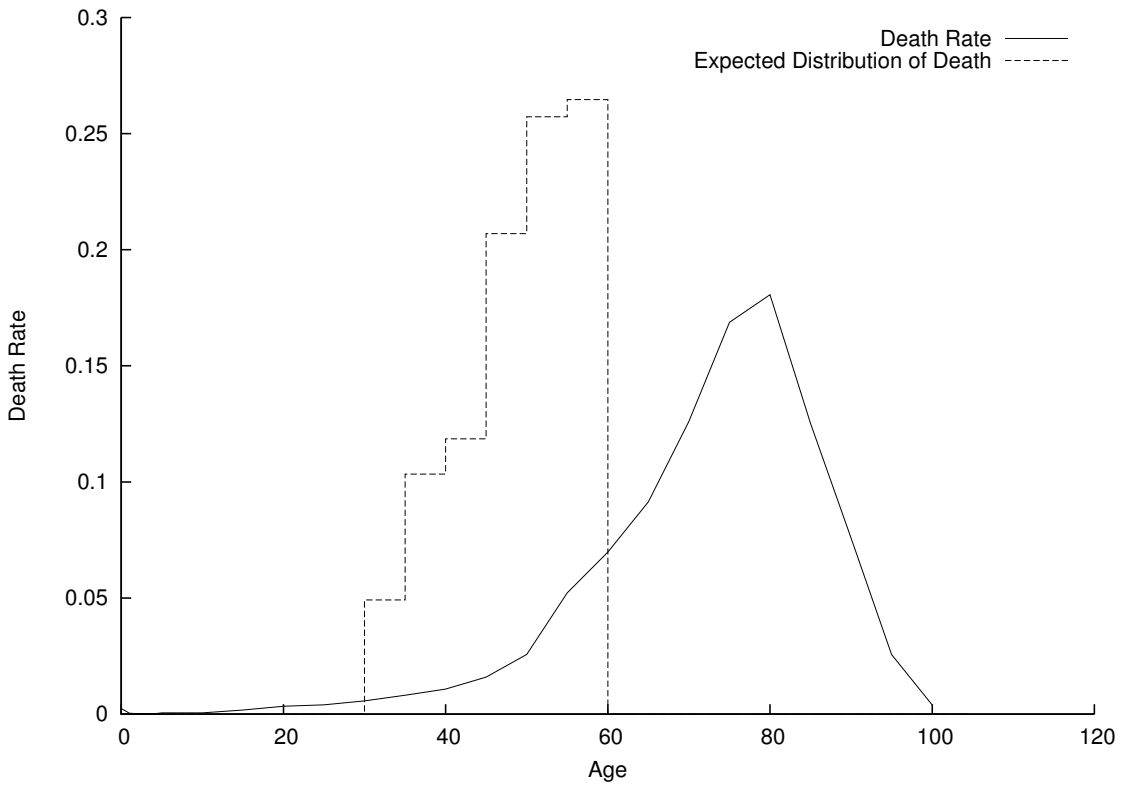

図 4.1: A における年齢階級別期待死亡率

で定義する[1)].

この SMR の期待値が 1 に等しいか否かで，有意性を判断する．すなわち，帰無仮説 H_0 : $\lambda = E$，対立仮説 $H_1 : \lambda \neq E$ を確率検定する．ポアソン分布は，E が 5 以上であれば，統計量

$$Z = \frac{O - E \pm 0.5}{\sqrt{E}} \tag{4.4}$$

により正規分布 $N(0,1)$ で近似できる．ここで，0.5 は連続修正項である．両側検定であれば，

$$Z = \frac{|O - E| - 0.5}{\sqrt{E}} > Z(\alpha/2) \tag{4.5}$$

の時に，有意水準 α で H_0 が棄却できる．

4.4 外部，内部比較

外部比較は，組織 A の死亡率が全日本人のそれと異なるかを比較し，内部比較では，A における集合を対象とする属性で分割し，それらの O/E 比の差を比較する．例えば，放射線従事者の累積放射線量 (10mSv 未満,10mSv 以上等) と死亡率の相関を検証する．

[1)] E : 年齢以外には，性別，暦年，職業，地域など問題に応じて様々な層別が行われる．

内部比較では，死亡率は累積線量に依存して増加しているという仮説の下で，スコア検定統計量を用いて傾向性の片側検定 (Breslow-Day 検定) を行う．表 2.1 に外部比較，表 2.2 に内部比較の例を示す．図 4.2 に内部比較と外部比較のイメージを示す．

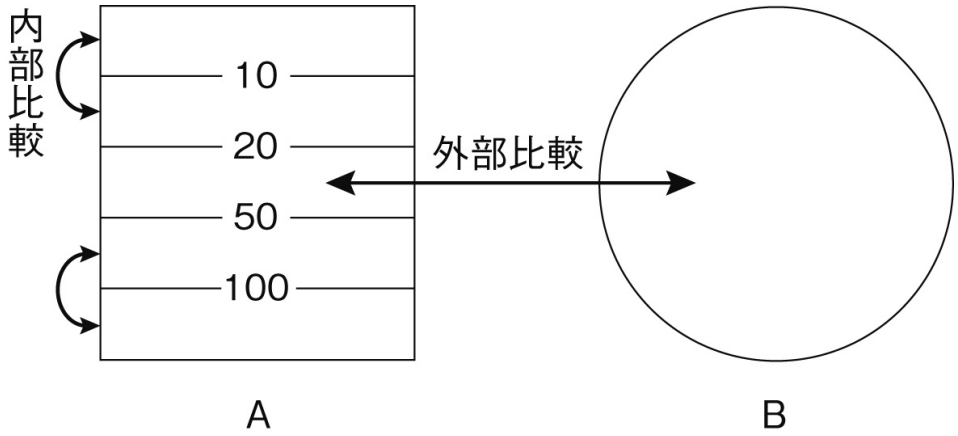

図 4.2: 内部比較と外部比較のイメージ

4.5 問題設定

次の 2 つの問題を考える．

(問題 1) 公開死亡率における仮説検定

U の部分集合 X_A と X_B を有する A と B が互いの集合を秘匿したままで，

$$O = |X_A \cap X_B| \tag{4.6}$$

$$E = \sum^{m} d_i n_i \tag{4.7}$$

を求める問題．但し，X_A は，$X_A = X_{A_1} \cup \cdots \cup X_{A_m}$ の m 層に分割され，$n_i = |X_A|$ は A が知り，層 i における X_B の割合死亡率 d_i は公開されている．例えば，表 2.1 における食道がんの観察死亡数 326 は，全て公開して明らかになる．

(問題 2) 秘密死亡率における仮説検定

未知の疾病に対して，死亡率 (罹患率) を公開できない場合がある．そこで，この問題においては層 i における X_B の割合 d_i を B のみが持ち，A に秘匿したままで，両側検定の p 値

$$p = P[Z > \alpha/2] \tag{4.8}$$

のみを求める．よって，O と E は A と B 単体には秘密とする．

ここで，問題 1 と問題 2 について整理しよう．従来の疫学調査と，それぞれの問題設定において公開される情報を表 4.2 に示す．従来の疫学調査では，全ての情報が互いの組織の間

で知られてしまう．問題 1 では，それぞれの組織の持っているデータの誰と誰が同じ人かという情報を秘密にする．問題 2 では，全ての情報を秘密にしたまま疫学調査を行う．

表 4.2: 公開される情報

	$X_A \cap X_B$	O(観察死亡数)	E(期待死亡数)	Z(統計量)
従来	公開	公開	公開	公開
問題 1	秘密	公開	公開	公開
問題 2	秘密	秘密	秘密	秘密

4.5.1 問題 1 への提案プロトコル

X_A と X_B を秘匿したまま，積集合の大きさ $|X_A \cap X_B|$ のみを求める暗号プロトコルには次の 3 つが知られている．

1. AES03(可換一方向性関数)[4]
   X_A,X_B は集合である．

2. 秘匿内積プロトコル [5]
   $\boldsymbol{X}_A$ と $\boldsymbol{X}_B$ はベクトルであり，結果は $s_A + s_B = \boldsymbol{X}_A \boldsymbol{X}_B$ となり，二つの乱数 s_A と s_B に分散して得られる．

3. FNP04(多項式評価)[6]
   X_A，X_B は集合である．加法準同型性に基づく秘匿多項式評価を応用し，X_A を根として持つ多項式 $f(x)$ を B が秘匿したままで $f(y)$ を計算する．

問題 1 は，各層 j について，これらのいずれかを適用して，O_j を求める．A は，公開 d_j を用いて，$E_j = \sum_j d_j n_j$ を公開する．4.3 節の方法で，外部比較，内部比較を実施し，p 値，信頼区間を求め，帰無仮説が有意に棄却できるか判断する．

4.5.2 問題 2 への提案プロトコル

A は，$\boldsymbol{X}_A$ と $n_1, \ldots, n_m$ を持ち，B は $\boldsymbol{X}_B$ と $d_1, \ldots, d_m$ を持つ．

1. 秘匿内積プロトコルを用いて

$$s_A + s_B = X_A X_B = O \tag{4.9}$$

となる s_A，s_B を得る．A は s_A^2 を，B は s_B^2 をそれぞれ計算する．

2. Aは加法準同型性を満たした公開鍵暗号を用いて，鍵対を作り，暗号文$E(n_1),\ldots,E(n_m)$をBへ送り，Bは乱数t_Bを作って，

$$y=\prod_i^m E(n_i)^{d_i}/E(t_B) \tag{4.10}$$

を計算し，Aは$t_A=D(y)=t_B+\sum n_i d_i$を作る．$t_A+t_B=E$である．

3. 再び，秘匿内積プロトコルを用いて，

$$w_A+w_B=s_A s_B \tag{4.11}$$

となるw_Aとw_Bを求め，AとBで分散管理する．

4. Aは，s_A,${s_A}^2$,t_A,w_Aを，Bはs_B,${s_B}^2$,t_B,w_Bを万能秘密関数計算プロトコルSFE(Secure Function Evaluation)[8]へかけて，次のp値を求める．

$$\begin{aligned} z^2&=\frac{{s_A}^2+2(w_A+w_B)+{s_B}^2}{t_A+t_B}-2(s_A+s_B)+(t_A+t_B)\\ &=\frac{O^2}{E}-2O+E \end{aligned} \tag{4.12}$$

5. z^2が自由度1，有意水準αのχ^2値未満ならば，帰無仮説H_0を棄却する．(外部比較)

6. (内部比較) 累積線量によってk個に分割された$X_{A_1},\ldots,X_{A_k}$の各々について，1から5を実行し，求めたk個の統計量の和が

$${Z_1}^2+\cdots+{Z_k}^2<\chi^2_\alpha(k-1) \tag{4.13}$$

ならば，H_0:死亡率は累積線量に依らず一定である，とする帰無仮説を棄却する．

4.5.3 評価

安全性

方式1は用いる要素技術AES03，秘匿内積プロトコル，FNP04の安全性に基づいて，各集合X_AとX_Bを秘匿する．AES03プロトコルは，入力xと$H(x)^{uv}$が一対一対応する．このため，入力が特定の分布に従う場合は，$H(x)^{uv}$の分布からxを統計的に推定する攻撃方法が存在する．しかし本研究の事例のように，入力集合の要素が取り得る値が，ユニークなIDであったり，常に一つ以下しか存在しないことが保証されている属性値であるような場合には，このような統計的攻撃は成立せず，問題にならない．

方式2は，AとBが正直に振舞うセミオネストモデルの下で，正しく検定量Z^2を計算し，ベクトル$\boldsymbol{X}_A$,$\boldsymbol{X}_B$に加えて，観察数$O=X_A,X_B$と期待値$E=\sum d_i n_i$を秘匿する．安全性は加法準同型性を満たす暗号の識別不能性に帰着する．

提案方式の比較

要素技術と提案方式との関係を表 4.3 に整理する．要素の同定とは，積集合の数だけではなく，積集合そのものを同定するプロトコルが可能なのは，AES03 と FNP04 のみである．処理性能は，リストの大きさ n と定義域の大きさ N に依存する．AES03 のパフォーマンスは，3.1 節の実装に基づいた値である．これらの要素技術は互換ではなく，本論文の問題点 2 への提案プロトコルは秘匿内積プロトコルしか適用できないことに注意が必要である．

表 4.3: 暗号プロトコルの比較

	AES03[4]	秘匿内積プロトコル [5]	FNP04[6]
要素の同定可能	yes	no	yes
入力	集合	ベクトル	集合
処理性能	$O(n)$	$O(N)$	$O(n^2)$
パフォーマンス	360 件/s	10 件/s	-
提案方式 2 への利用	no	yes	no

4.6 実装評価

4.6.1 試験実装したプログラム

問題 1 に対する提案プロトコルについて，その実現可能性を検討するために，Agrawal らの提案した AES03 プロトコル [4] を Java により実装し，その性能を評価した．BigInteger クラスとリストの照合に Collection フレームワークの Map クラスを用いた．

使用するデータの例を表 4.4 に示す[2)]．

表 4.4: 垂直分割データセットの例

A			B	
名前	年齢	累積線量 [mSv]	名前	死因
佐藤	20	12	田中	肺がん
菊池	30	51	鈴木	前立腺がん
佐久間	30	33	佐藤	外因死
鈴木	70	46	後藤	肺がん

プログラムの入力は放射線事業従事者の ID が入ったリスト X_a とがん罹患者の ID が入ったリスト X_b で，出力は標準的な日本人全体のがん罹患率と比べて，放射線事業従事者のがん罹患率が高いか低いかの判定結果を出力する．

[2)] この例では，２つのデータセットが同期されていないため，を使うことはできない．

4.6.2 試験実装 (提案方式 1) の処理性能

法のサイズ 1024 ビットの時の，試験実装したプログラムの処理時間を図 4.3 に示す．1 秒

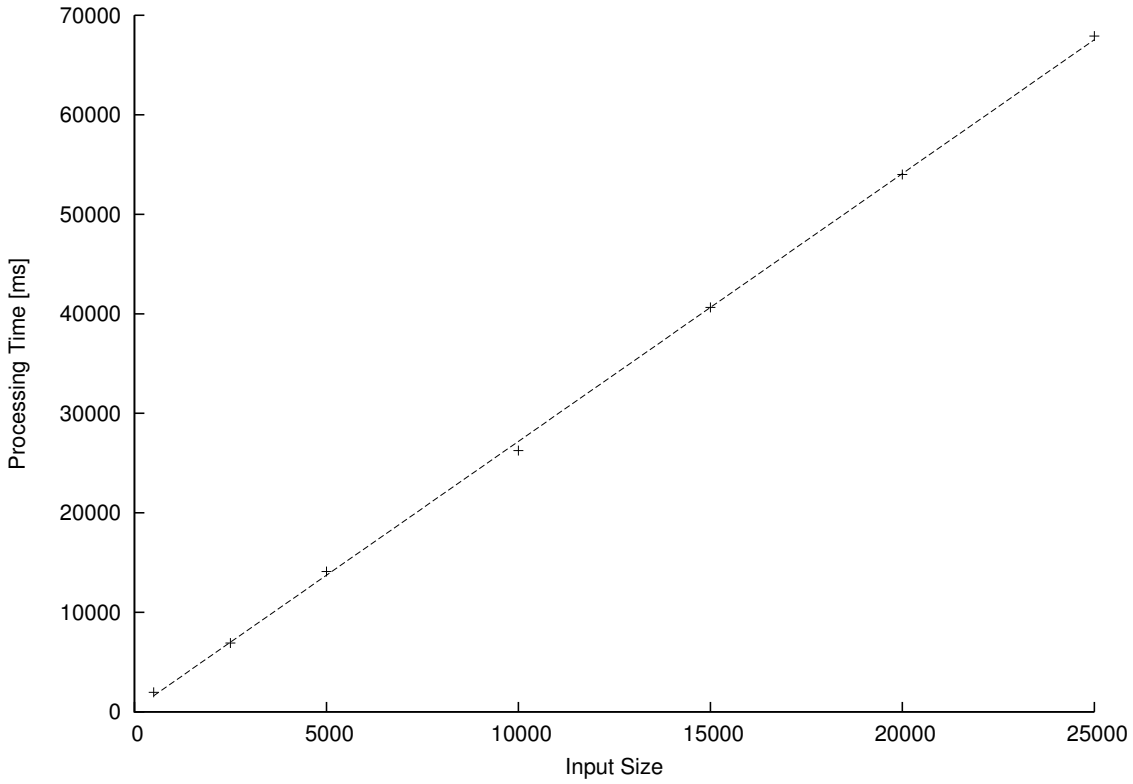

図 4.3: 試験実装 (提案方式 1) の処理時間

当たり約 360 個のデータについて照合を行っている計算になり，もう一方のリストも 20 万人と仮定した場合，文献 [2] で行っている約 20 万人のデータだと 550 秒，約 9 分かかる．

4.7 まとめ

2.1 節で，疫学におけるプライバシー問題について，既存の文献を基に検討し，課題と要求条件を定義した，被験者のプライバシーを考慮することと，より粒度の高い詳細な調査が必要であることが矛盾する要求である．

この問題に対して，本章で暗号プロトコルの適用を提案し，試験実装に基づいて十分に実現可能であることを示した．

第5章　プライバシーを保護した相対危険度の有意性検定プロトコル

5.1　概要

環境因子と疾病の因果関係を明らかにする疫学調査では，疾病の原因と考えられる因子と疾病の関係性を統計的に明らかにする [1], [3]．例えば，喫煙者のがん罹患率と非喫煙者のがん罹患率を比較する事で，喫煙による健康への影響を明らかにすることができる．しかし，これらのデータは独立した組織によって管理されている事が多く，プライバシーと個人情報保護の観点から，照合が困難であった．そこで，本章では 2.3 節の患者-対照調査における，相対危険度の有意性検定について，暗号プロトコルを適用し，プライバシーを保護したまま安全に確率検定を行うことを試みる．

秘匿内積プロトコル [5] と秘密関数計算プロトコル [7] を使用することで，理論的には安全に確率検定を行うことができる．しかし，現実的には，秘密関数計算プロトコルは処理が極めて遅く，積や平方根などの計算が困難であった．

そこで本章では，秘密関数計算プロトコルでの処理効率を考慮して効率的に評価できる確率検定プロトコルを提案する．提案方式を Java を用いて実装し，その実現可能性と安全性を評価する．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 [1] に使用されている約 14 万人のデータについて，実装したプログラムを適用した時の処理時間を見積もり，提案方式の実現可能性について検討する．

5.2　問題定義

秘匿の必要がある集合 X_A を持つ組織 A と，X_B を持つ組織 B が協力して疫学調査を行う．例えば，組織 A は特定要因である喫煙者のデータを持っている組織，組織 B は死亡者のデータを管理する組織とする．すなわち，組織 A は表 2.4 の n_1，n_2 の情報を持ち，組織 B は m_1，m_2 の情報を持つ．それぞれの持っているデータの合計 N は公開するが，属性毎の合計値，n_1，n_2，m_1，m_2 は秘密とする．これらの条件の元で，効率的に疫学調査を行い，対象とする特定要因の相対危険度が有意かどうかを検定する．出力結果は，その検定結果のみとする．

これを解くナイーブな方法は，X_A と X_B をベクトル表現し，3.2 節の秘匿内積プロトコルを適用することであるが，これには，SFE の実行に関する次の 2 つの問題点がある．

5.3 問題点

1. 統計量を求めるための大きな計算量．秘匿内積プロトコルは 2 つに分散された値を出力する．これを秘匿したまま解くには Fairplay などのシステムが必要になるが，統計量 χ を 2.3 節の式

$$\chi = \frac{\sqrt{N-1}\{(ad-bc) \pm N/2\}}{\sqrt{n_1 n_2 m_1 m_2}} \tag{5.1}$$

で求めるためには，SFE 上での平方根や積などの計算が必要になる．しかしながら，3.3 節で示した様に，Fairplay でこれらを計算するのは困難である．

2. 分散値の定義域の大きさ．秘匿内積プロトコルでの STEP(3) 計算時に乱数 s_B を生成するが，この乱数は安全性の為，準同型暗号の定義域 Z_n から選ぶ必要がある．しかし，公開鍵暗号の平文長の 2048bit の様な値は，Fairplay で計算するには大きすぎる．

5.4 アプローチ

問題点 (1) を解決するために，Fairplay での計算は加算や減算，比較のみに限定したい．そこで，次に示す様に，検定の条件を Fairplay で可能な等価な式に変形する．検定を行った際に有意になるか否かは，N，n_1，m_1 を固定した場合，a の大きさによって求めることができる．まず，式 (5.1) を a のみを用いるように変形すると，

$$\begin{aligned}
\chi &= \frac{\sqrt{N-1}\{(ad-bc) \pm N/2\}}{\sqrt{n_1 n_2 m_1 m_2}} \\
&= \frac{\sqrt{N-1}\{a(N-n_1-m_1+a) - (n_1-a)(m_1-a) - N/2\}}{\sqrt{n_1 n_2 m_1 m_2}} \\
&= \frac{\sqrt{N-1}\{aN - n_1 m_1 - N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}
\end{aligned} \tag{5.2}$$

と，a の一次式になる．こうして，a を変化させた時の統計量 χ の変化を図 5.1 に示す．この値が有意水準 $Z(0.05/2) = 1.960$ を超えたかを判断すれば良い．この境界の a を a^* とおくと，式 (5.2) より

$$\begin{aligned}
a^* &= \left(\frac{\chi \cdot \sqrt{n_1 n_2 m_1 m_2}}{\sqrt{N-1}} + n_1 m_2 + \frac{N}{2} \right) \cdot \frac{1}{N} \\
a^* N &= \frac{1.960 \cdot \sqrt{n_1 n_2 m_1 m_2}}{\sqrt{N-1}} + n_1 m_2 + \frac{N}{2}
\end{aligned} \tag{5.3}$$

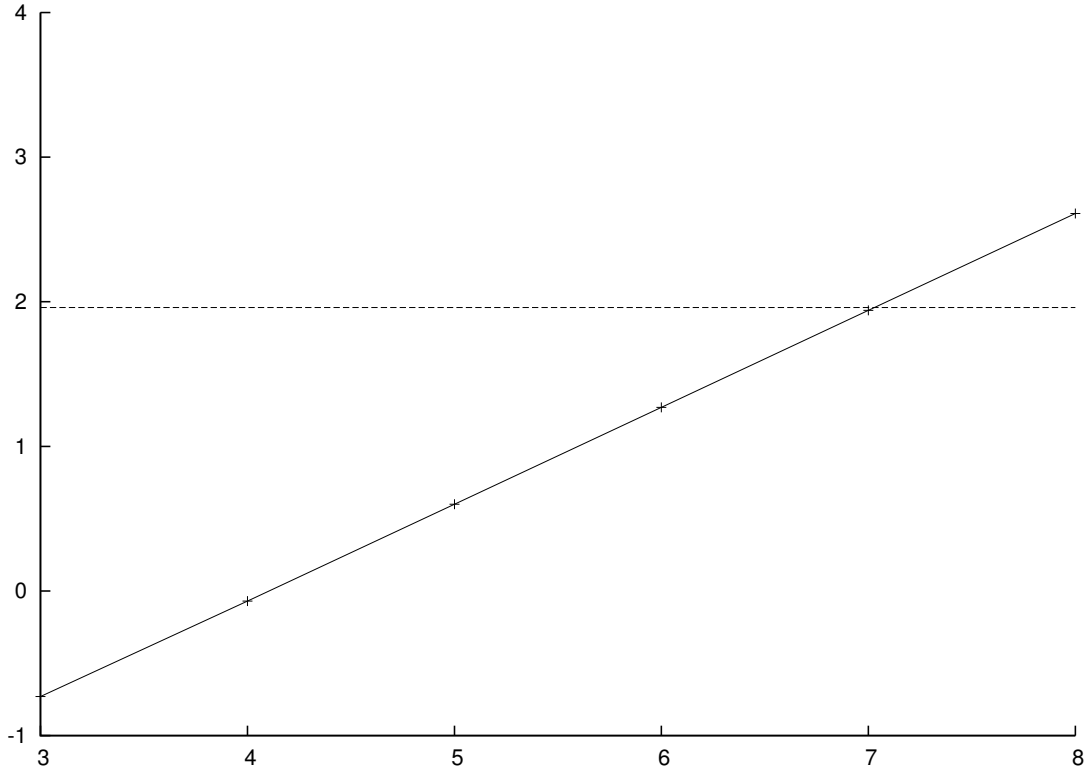

図 5.1: a を変化させた時の統計量 χ

となる．ここで，χ と N は与えられた公開情報であり，$\sqrt{n_1n_2m_1m_2}$ は n_1 と n_2 を持つ A と m_1，m_2 を持つ B が秘匿内積プロトコルを行えば分散した値が得られる．n_1m_1 も同様である．従って，秘匿内積プロトコルで X_A と X_B(喫煙者と死亡者) の積集合 a が，a^* を上回っていれば有意，下回っていれば有意ではないことを Fairplay で検査するには，

$$(s_A + s_B) > (t_A + t_B) + (u_A + u_B) \tag{5.4}$$

を評価し，判定結果のみを出力すればよい．ここで，s_A, s_B, t_A, t_B, u_A, u_B は，

$$\begin{aligned}
&s_A + s_B = aN = |X_A \cap X_B|N, \\
&t_A + t_B = \frac{\chi \cdot \sqrt{n_1n_2m_1m_2}}{\sqrt{N-1}} = \left(\frac{\chi\sqrt{n_1n_2}}{\sqrt{N-1}}\right) \cdot \sqrt{m_1m_2}, \\
&u_A + u_B = n_1m_1 + \frac{N}{2}
\end{aligned}$$

で定義される値であり，Alg. 2 により効率的に求めることができる．

5.4.1 乱数 s_B の定義域

問題 (2) で述べた通り，Alg. 2 の s_B を Z_n から選ぶと，s_A, s_B は $|Z_n|$ のサイズを持つ整数となり，SFE で求めるには大きすぎる．

そこで，Alg. 2 の STEP(3) を $E(s_B)$ で割る代わりに，小さな定義域から一様に選んだ s_B をかける，すなわち，

$$c = E(x_1)^{y_1} \times \cdots \times E(x_n)^{y_n} \times E(s_B)$$

と変更し，

$$s_A - s_B = \boldsymbol{x} \cdot \boldsymbol{y}$$

となる s_A と s_B に分散する様にする．ただし，$s_A - s_B < 0$ の時は，2 の補数表現で $|Z_n|$bit の整数が生じてしまうため，$s_A > s_B$ となる様に s_B を選ぶ．

$a = |X_A \cap X_B| \in [0, n]$ である時，(3) の乱数の定義域を μ 個の自然数，すなわち，$s_B \in [0, \mu - 1]$ とする．μ は n に対して十分大きく，かつ，SFE で処理可能な大きさに定める必要がある．なぜならば，次に挙げる安全性の問題が生じるためである．

$s_A = s_B + a < \mu$ の時に，A が s_A を知ることにより真の a の大きさについて言えることは，

$$P(a = 0|s_A) = \cdots = P(a = n|s_A) = \frac{1}{n+1}$$

であり，$[0, n]$ のどの値も一様に確からしい．しかし，$\mu < s_A$ の時，a の取り得る値は，

$$\begin{cases} P(a = 0|s_A) = \cdots = P(a = s_A - \mu - 1|s_A) = 0 \\ P(a = s_A - n|s_A) = \cdots = P(a = n + 1|s_A) = \frac{1}{n+\mu+1-s_A} \end{cases}$$

となり，偏りが生じる．図 5.2 にこの s_A の危険な領域を示す．例えば，$n+1 = 100, \mu = 1000$

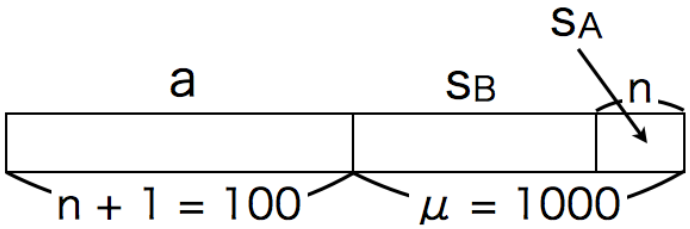

図 5.2: s_A の危険な領域

の時，$s_A = 1005$ ならば，a は少なくとも 5 以上でなくてはならない．s_A がこの「危険な領域」に落ちる確率を次で与える．

定理 1 (乱数長からの安全性) $a \in [0, n]$, $s_B \in [0, \mu - 1]$ の一様分布から選んだ値とする．$s_A = s_B + a > \mu$ となる確率は，

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

である．

証明 1 $s_A = \alpha + \beta$ となる α, β を用いると，

$$\begin{aligned} P(\mu = s_A) &= P(a = \alpha) \cdot P(s_B = \beta) \\ &= \frac{1}{n+1} \cdot \frac{1}{\mu} \end{aligned}$$

ここで，$\mu = s_A$ の時，$s_A = \alpha + \beta$ となる (α, β) 組は，$(1, \mu - 1), \cdots, (n+1, \mu - n + 1)$ の $n - 1$ 通り．一方，$n + \mu + 1 = s_A$ となるのは，$(n+1, \mu)$ の 1 通り，よって，初項 $n - 1$，公比 -1 の等比数列の和より，危険領域の条件を満たす組は，$(n-1)(n+1)/2$ 存在する．よって，

$$\begin{aligned} P(\mu < s_A) &= \sum_{\mu < s_A} P(s_A) \\ &= \sum_{\mu < s_A = \alpha + \beta} \frac{1}{n+1} \cdot \frac{1}{\mu} \\ &= \frac{(n+1)(n-1)/2}{(n+1)\mu} = \frac{n-1}{2\mu} \end{aligned}$$

で定理を得る． $(Q.E.D.)$

例えば，$n + 1 = 100$, $n = 10 \cdot n = 1000$ の乱数を選ぶと，A が a について一部の情報を得る確率は，$99/2 \cdot 1000 = 0.0495$ と十分に小さい．逆に，その確率を ϵ とすると，$\mu > \frac{n-1}{2\epsilon}$ を超える乱数を選べば良い．

最後に，a の情報が一部漏れた時の大きさを評価する．$\mu < s_A$ の時，$P(a|s_A) = \frac{1}{n+\mu+1-s_A}$ より，損なわれる条件付き確率の変化を図 5.3 に，そのエントロピーの減少を図 5.4 に示す．s_A が $\mu = 1000$ を超えてから，a について同定される程度を図示している．図 5.3 と図 5.4 では，どちらも s_A が 0 から 1000 までの間は，a が同定される確率は破線のように一様だが，s_A が 1000 を超えると，実線のように a が同定される確率が上がっていく．

5.4.2 提案方式

以上の提案方式を Algorithm 3 に示す．改良した秘匿内積プロトコルを Algorithm 4 に示す．出力される s_A と s_B が $[0, \mu - 1]$ の値域に収まることに注意されたい．

5.5 評価

5.5.1 パフォーマンス

Java BigInteger クラスを用いて実装したプログラム (Alg. 4) の処理時間を図 5.5 に示す．実装したプログラムの処理時間は，要素数 N に対して線形に増加している．また，Fairplay

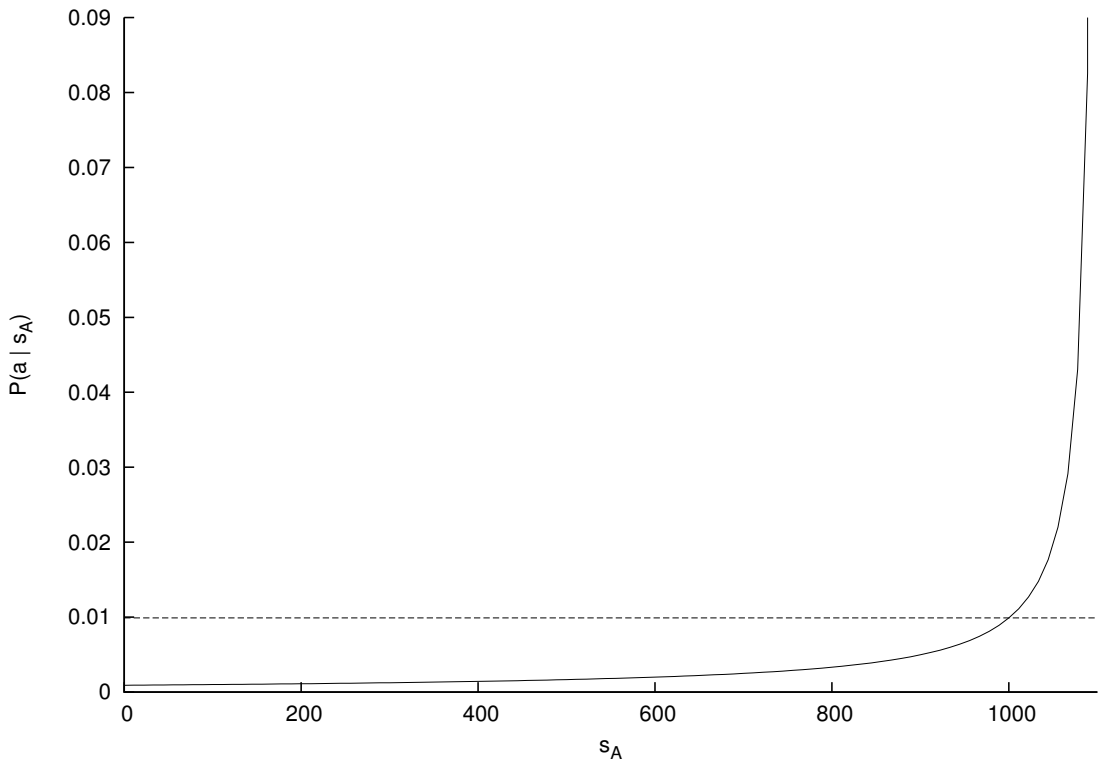

図 5.3: 損なわれる条件付き確率 $P(a|s_A)$ の変化

Algorithm 3 改良した秘匿内積プロトコル

入力: Alice は n 次元ベクトル $\boldsymbol{x} = (x_1, \ldots, x_n)$ を持つ．Bob は n 次元の $\boldsymbol{y} = (y_1, \ldots, y_n)$ を持つ．

出力: Alice と Bob は $s_A - s_B = \boldsymbol{x} \cdot \boldsymbol{y}$ となるような s_A，s_B を得る．

1. Alice は準同型暗号の公開鍵対を作り，公開鍵を Bob に送る．
2. Alice は Bob に暗号文 $E(x_1), \ldots, E(x_n)$ を送る．
3. Bob は s_B を $s_B \in [0, \mu - 1]$ をランダムに選び，

   $$c = E(x_1)^{y_1} \cdots E(x_n)^{y_n} \cdot E(s_B)$$

   を計算し，Alice に送る．
4. Alice は c を復号し，$s_A = D(c) = x_1 y_1 + \cdots + x_n y_n + s_B$ を得る．

で式 (5.4) の評価を行った時の処理時間を図 5.6 に示す．線形に増加しているが，3.1 の乗算にかかる計算時間と比較すると，十分早い．

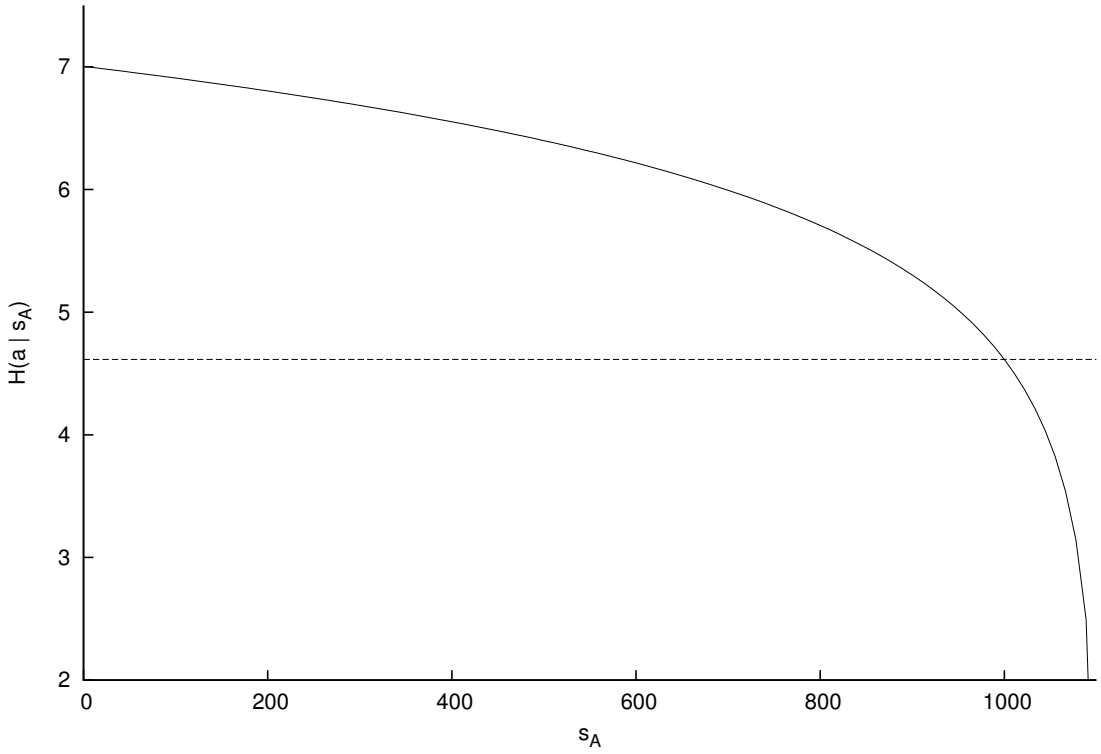

図 5.4: エントロピーの減少度合

Algorithm 4 プライバシー保護相対危険度検定

入力: $|X_A| = n_1$, $|X_B| = m_1$

出力: $|X_A \cap X_B| = a$ が 95%の水準で有意か

1. Alg. 3 を用いて，$s_A - s_B = aN$ となる s_A を A が，s_B を B が得る．
2. Alg. 3 を用いて，$t_A - t_B = \left(\frac{\chi\sqrt{n_1 n_2}}{\sqrt{N-1}}\right) \cdot \sqrt{m_1 m_2}$ となる t_A を A が，t_B を B が得る．
3. Alg. 3 を用いて，$u_A - u_B = n_1 m_1 + \frac{N}{2}$ となる w_A を A が，w_B を B が得る．
4. SFE を用いて，A は (s_A, t_A, u_A), B は (s_B, t_B, u_B) を入力し，(5.4) 式を判定する．

入力する乱数 s_B の bit 長 $(=\mu)$ を変化させ，それぞれの bit で 5 回ずつ評価した時の処理時間の平均と分散である．

Alg. 4 の通信量を図 5.7 に示す．1 要素当たり 620byte の暗号文が生成されている．

X_A と X_B の共通集合 $|X_A \cap X_B| = a$ を変化させた時の処理時間を 5.8 に示す．a が変化しても処理時間に変化が見られないため，処理時間による a の同定は行えない．

Alg. 4 の STEP(4)，Fairplay を用いて分散比較を行うプログラムを図 5.9 に示す．

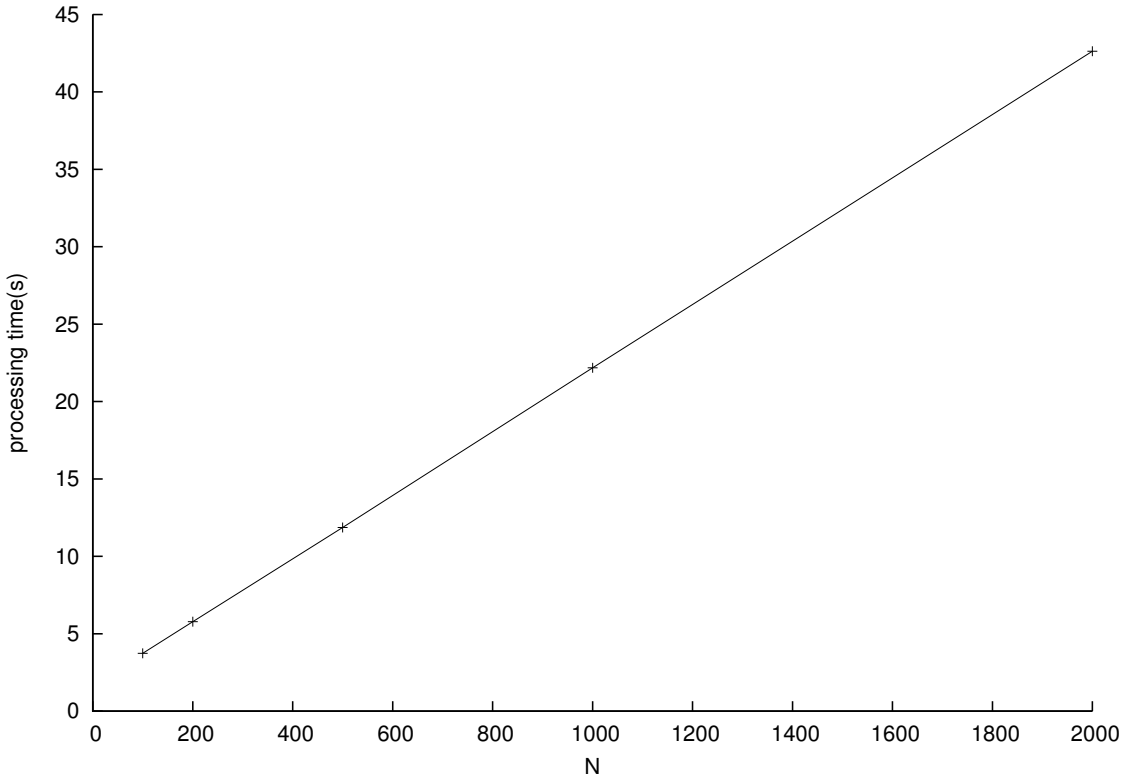

図 5.5: 実装したプログラムの処理時間

5.5.2 多目的コホート研究への適用

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 [1] では，生活習慣病の科学的な予防法を明らかにすることを目的に，140,420 名のデータを元に，喫煙や飲酒のリスクなどを調査している．

本プログラムをこの 140,420 人のデータに対して使うと，2855.5 秒，約 48 分で計算を行うことができ，十分実用的な時間と考えられる．

また，多目的コホート研究のような，ある属性に対して複数の要素の相対危険度を評価する必要がある場合，Alice は一回だけ全ての要素を暗号化するだけで済み，それに対して Bob が複数回計算を行うことで実現できる．計算にかかる時間は，ほぼ STEP(1) の暗号化処理なので，多属性について評価する場合は，より本プロトコルが有効になると考えられる．要素数に対する処理時間を表 5.1 に示す．

5.5.3 安全性

[5] では Z_n から乱数を選ぶことが提案されているが，Fairplay で計算するため，乱数の大きさを抑える必要がある．そのため，[5] の秘匿内積プロトコルよりも安全性は落ちる．し

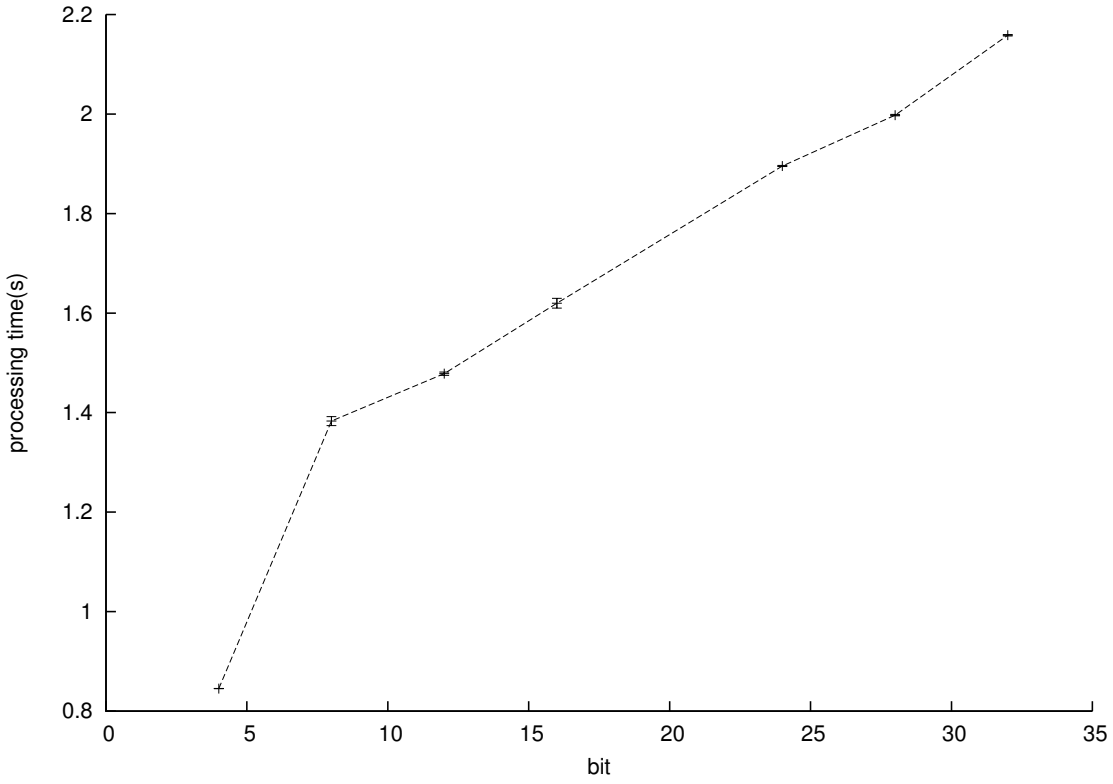

図 5.6: Fairplay を用いて式 (5.4) の評価を行った時の処理時間

表 5.1: 要素数に対する処理時間

	100	200	500	1000	2000
(1)Alice	67%	79%	90%	94%	97%
(2)Bob	17%	11%	5%	3%	2%
(3)Alice	16%	10%	5%	3%	1%

かし，5.4.1 で示したように，Fairplay で計算可能な範囲で適切な乱数を選ぶことで，リスクを抑えることができる．

提案方式の安全性は，要素技術である秘匿内積プロトコル，SFE の安全性に依存する．セミオネストモデルの仮定の下，Alg. 4 は検定結果以外の情報を漏らさない．

5.6 まとめ

本章では秘匿内積プロトコルと Fairplay を使用し，2 つのデータセットを秘匿したまま，確率検定を行うシステムを実装した．実装する際にネックとなる Fairplay の制約に対して，従来の式を変形し加算や減算，比較のみで計算を行い，乱数の大きさを抑えることで現実的な時間で処理を行えるようになった．

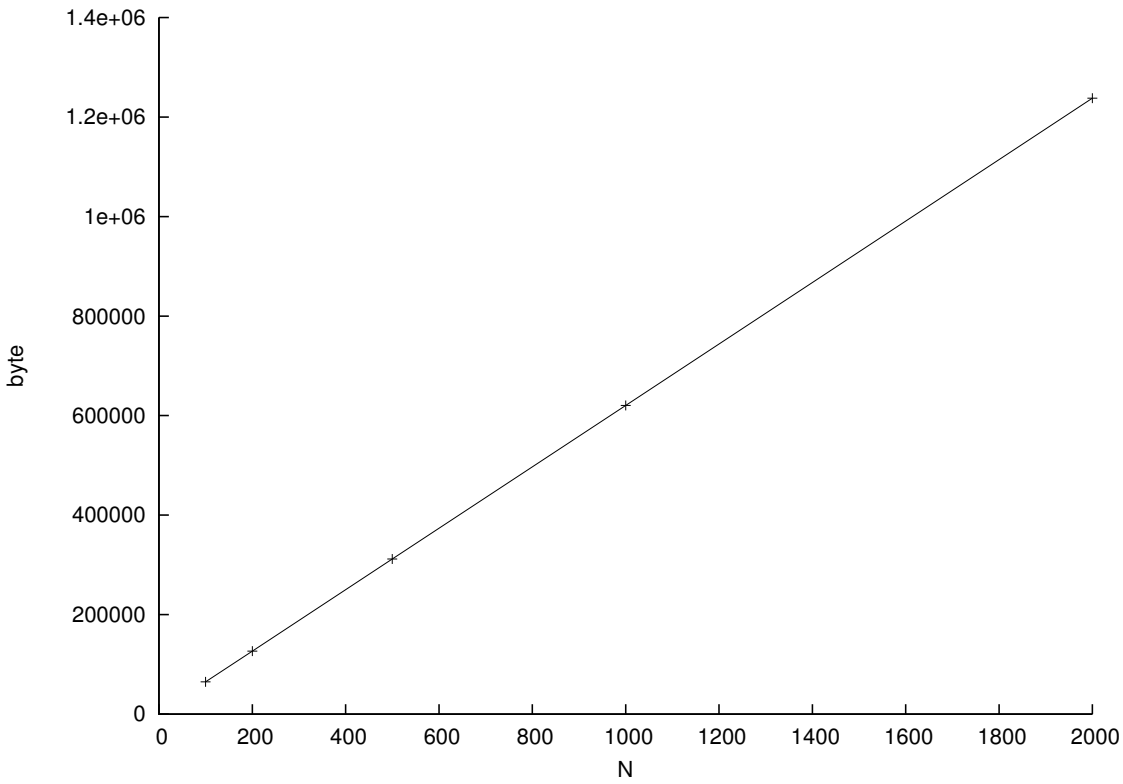

図 5.7: 要素数 N に対する Alg. 4 の通信量

本章の主要な結論は次の通りである.

1. SFE で評価可能な, 効率の良い確率検定プロトコルを提案した. その処理時間は, $N = 1000$, $\mu = 2^{31}$ の時, 約 26 秒である.
2. 秘匿内積プロトコルの結果を SFE で比較する際の乱数長の問題を指摘し, 適切な乱数長 μ と, その安全性やエントロピーの損失を明らかにした.

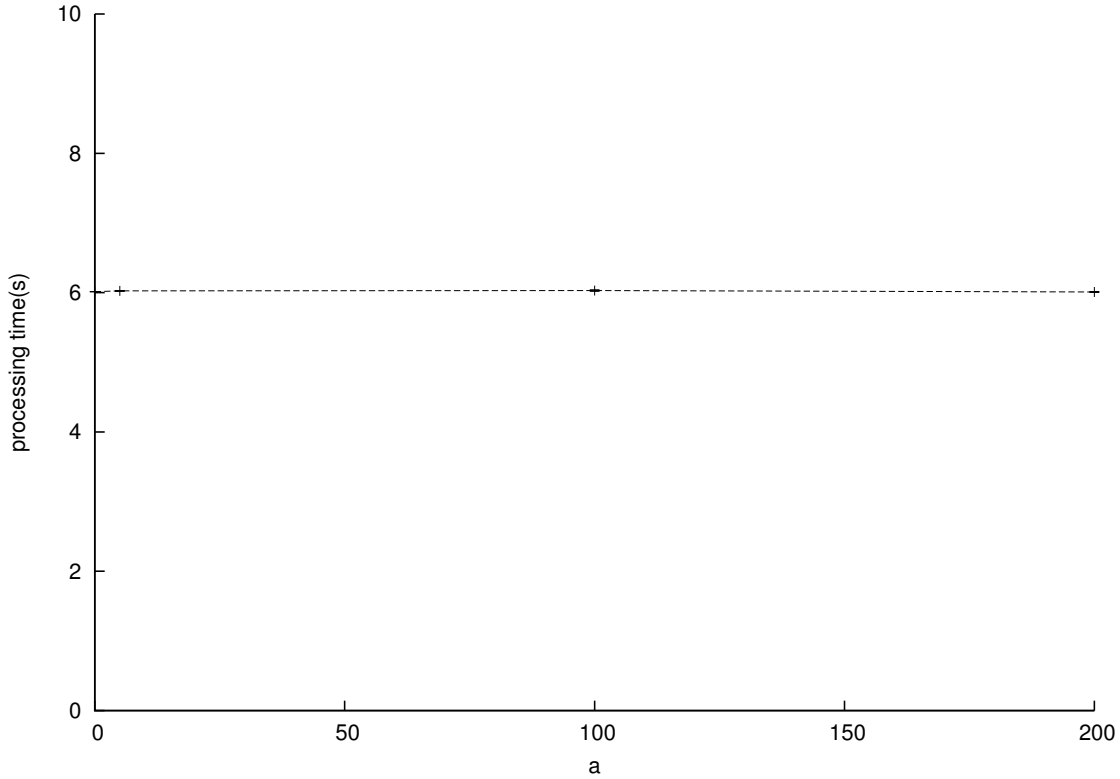

図 5.8: a を変化させた時の処理時間

```
program Comp {
   type int = Int<32>; // 32-bit integer
     type AliceInput = int[3];
     type BobInput = int[3];
     type AliceOutput = Boolean;
     type BobOutput = Boolean;
     type Output = struct {AliceOutput alice,
         BobOutput bob};
     type Input = struct {AliceInput alice,
         BobInput bob};

     function Output output(Input input) {
                  output.alice = ((input.alice[2] - input.bob[2]) > (input.alice[1] -
input.bob[1]) + (input.alice[0] - input.bob[0]));
                  output.bob = ((input.alice[2] - input.bob[2]) > (input.alice[1] -
input.bob[1]) + (input.alice[0] - input.bob[0]));
     }
}
```

図 5.9: Fairplay を用いた分散比較プログラム

第6章　プライバシーを保護した傾向性の検定プロトコル

6.1　概要

(財) 放射線影響協会の行っている “原子力発電施設等放射線業務従事者等に係る疫学的調査 [2]”では，調査対象者の累積線量群について死亡率と死因を調べ，放射線業務従事者に対する低線量域での健康への影響を明らかにしている．累積線量が増えることによる健康への影響を明らかにすることは非常に重要である．しかし，これらの情報は前述した通り累積線量を管理している放射線事業者中央登録センターと死亡者リストを持つ厚生労働省で独立して管理されており，プライバシー保護の関係で互いに照合することは難しい．

そこで，本章では暗号技術を使用することで，二つの組織のデータを秘匿したまま，累積線量が増える事によって死亡率が増加する傾向があるか，という傾向性の検定 (2.4 節) を行うことを目指す．

6.2　問題定義

秘匿する必要がある集合 X_A を持つ組織 A と X_B を持つ組織 B が協力して傾向性の検定を行う．例えば，組織 A は投与した薬物の用量と ID についてのデータを持ち，組織 B はそれに対する反応の計量値と ID を持つ組織とする．各組織の持つデータ例を表 6.1 に示す．

表 6.1: 各組織の持つデータ

	組織 A	組織 B
ID	用量 x_i	反応 y_i
1	10ppm	8.06
2	10ppm	8.27
⋮	⋮	⋮
19	10000ppm	7.91
20	10000ppm	7.40

6.3 秘匿回帰プロトコル

まず，$\hat{\beta}$，$\hat{\alpha}$ と統計量 t を求める．

$$\begin{aligned}\hat{\beta} &= \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} \\ &= \frac{SS_{XY}}{SS_X} \\ &= \frac{\sum_i^n x_i y_i - (\sum_i^n x_i)(\sum_i^n y_i)/n}{\sum_i^n x_i^2 - (\sum_i^n x_i)^2/n}\end{aligned} \tag{6.1}$$

ここで，$\sum_i^n x_i$, $\sum_i^n x_i^2$ は A のみ，$\sum_i^n y_i$ は B のみでローカルに計算できるので，$\sum_i^n x_i y_i$ のみを秘匿して求めれば良い．これは，秘匿内積プロトコルを適用すれば得られる．(6.1) 式の分子の $(\sum_i^n x_i)$ と $(\sum_i^n y_i)$ も，内積の一部で求める．すなわち，

$$x_{n+1} = -(\sum_i^n x_i), y_{n+1} = \sum_i^n y_i/n \tag{6.2}$$

とおいて，$n+1$ 次元の内積を求めれば良い．ここで，(6.1) 式の分母が A にのみ関係していることに着目すると，$i = 1, \cdots, n+1$ について x_i を次の様に置き換え

$$x_i' = \frac{x_i}{\sum_j^n x_j^2 - (\sum_j^n x_j)^2/n} \tag{6.3}$$

$(x_1', \cdots, x_{n+1}')$ と $(y_1, \cdots, y_{n+1})$ を Algorithm 2 に適用して，$\hat{\beta} = \beta_1 + \beta_2$ となる β_1 を A に，β_2 を B に分散したまま求められることが示された．

6.3.1 提案プロトコル

提案するプロトコルを Algorithm 5 に示す．α_1, α_2, β_1, β_2 を SFE に入力することで，係数を秘匿したまま任意の x についての推定値を得ることができる．

6.4 秘匿回帰検定プロトコル

傾向性を確かめるには，回帰で得られた推定値 $\hat{y} = \hat{\alpha} - \hat{\beta}x$ との残差を求め，$\hat{\beta}$ がその標準誤差 $s.e.(\hat{\beta})$ に対して有意な大きさがあるかを，(2.5) 式の検定量から判断を行う．従って，残差の平方和 $V_E = \sum_i^n (y_i - \hat{y_i})^2$ を秘匿して求めなくてはならない．

方式 1

$\alpha = \alpha_1 + \alpha_2$, $\beta = \beta_1 + \beta_2$ に分散されたままで V_E を次の様に求める．

$$
\begin{aligned}
V_E &= \sum_i^n (y_i - \hat{y_i})^2 \\
&= \sum_i^n (y_i - (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2)x_i)^2 \\
&= (\sum_i^n y_i^2 - 2(\alpha_1 + \alpha_2)\sum_i^n y_i) \\
&+ ((\beta_1 + \beta_2)^2 \sum_i^n x_i^2 - 2(\alpha_1 + \alpha_2)(\beta_1 + \beta_2)\sum_i^n x_i)) \\
&- 2(\beta_1 + \beta_2)\sum_i^n x_i y_i
\end{aligned}
$$

となるので，第 1 項を B が，第 2 項を A が，第 3 項を秘匿内積プロトコルで求める．

方式 2

p を有意水準とする．例えば $p = 0.01$ とする．(2.5) 式より，

$$t = \frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})} > t_{n-2}(p)$$

により，回帰直線 $\hat{\beta}$ の有意性を検定したい．ここで (2.4) 式と $\beta = 0$ を代入し，両辺を 2 乗すると

$$\frac{\hat{\beta}^2 \cdot SS_X}{V_E} > t_{n-2}^2(p)$$

を得る．これを変形すると，

$$
\begin{aligned}
\hat{\beta}^2 SS_X &> t_{n-2}^2(p) V_E \\
&= t_{n-2}^2(p)\left(\frac{SS_Y}{n-2} - \frac{(SS_{XY})^2}{n-2}\right) \\
&= t_{n-2}^2(p)\left(\frac{SS_Y}{n-2} - \frac{\hat{\beta}^2 \cdot SS_X}{n-2}\right)
\end{aligned}
$$

となる．これは，

$$
\begin{aligned}
\frac{t_{n-2}^2(p) SS_Y}{n-2} &< \hat{\beta}^2 SS_X \left(1 + \frac{t_{n-2}^2(p)}{n-2}\right) \\
&= (\beta_1^2 + 2\beta_1\beta_2 + \beta_2^2) SS_X \left(1 + \frac{t_{n-2}^2(p)}{n-2}\right) \\
&= (\beta_1^2 SS_X + 2SS_X\beta_1 \cdot \beta_2 + SS_X \cdot \beta_2^2) \left(1 + \frac{t_{n-2}^2(p)}{n-2}\right) \qquad (6.4)
\end{aligned}
$$

と同値である．$\left(1+\frac{t^2_{n-2}(p)}{n-2}\right)$ は定数であり，β_2 と β_2^2 は B のみで計算でき，$\beta_1^2 SS_X$ と $2SS_X\beta_1$ と SS_X は A のみで計算することができる．よって，２次元ベクトルの $(2SS_X\beta_1, SS_X)$ と (β_2, β_2^2) の秘匿内積プロトコルを実行して，A, B に分散した $\gamma_1+\gamma_2$ を求めれば与式は結局

$$\frac{t^2_{n-2}(p)SS_Y}{n-2} < \left(1+\frac{t^2_{n-2}(p)}{n-2}\right)(\beta_1^2 SS_X+\gamma_1+\gamma_2)$$

を判定する事と同値である．よって，

$$\frac{t^2_{n-2}(p)SS_Y}{n-2} - \left(1+\frac{t^2_{n-2}(p)}{n-2}\right)\gamma_2 < \left(1+\frac{t^2_{n-2}(p)}{n-2}\right)(\beta_1^2 SS_X+\gamma_1) \tag{6.5}$$

の左辺を B が，右辺を A がそれぞれで計算して，SFE に入力すれば，回帰直線 $\hat{\beta}$ の有意性のみが検定できる．

6.5 まとめ

秘匿内積プロトコルと秘密関数計算を用いる事で，2 つの組織のデータを秘匿したまま傾向性の検定を行うプロトコルを提案した．

処理にかかるコストは秘匿内積プロトコルにおける暗号化処理が支配的である．本プロトコルの暗号化の回数は，秘匿回帰プロトコルが $n+2$ 回，秘匿回帰検定プロトコルが 3 回であり，5 章のプロトコルと同程度のため，処理時間も同程度のパフォーマンスとなる．

Algorithm 5 秘匿回帰プロトコル

入力: $x_1, \cdots, x_n$ を持つ A,
$y_1, \cdots, y_n$ を持つ B.
n は A, B で共有.

出力: $\hat{\beta} = \beta_1 + \beta_2$ となる β_1 を A が，β_2 を B が得る.

1. A は $\sum_i^n x_i$, $\sum_i^n x_i^2$ を求め,

$$x_{n+1} = \sum_i^n x_i$$

とする．$i = 1, \cdots, n+1$ について,

$$x\,'_i = x_i / \sum_j^n x_j^2 - (\sum_j^n x_j)^2 / n$$

を求める.

2. B は，$\sum_i^n y_i$ を求め,

$$y_{n+1} = \sum_i^n y_i$$

とする.

3. A と B は，Algorithm 2 により，内積

$$\hat{\beta} = (x'_1, \cdots, x'_{n+1}) \cdot (y_1, \cdots, y_{n+1}) = \beta_1 + \beta_2$$

を求めて，β_1 を A が，β_2 を B が得る.

4. 同様にして，Algorithm 2 により

$$\hat{\alpha} = 1/n \sum_i^n y_i - \hat{\beta}/n \sum_i^n x_i = \alpha_1 + \alpha_2$$

となる α_1 を A が, α_2 を B が得る.

第7章　結論

7.1　結論

本稿では，既存の疫学調査におけるプライバシーの問題を明らかにし，実際に使用される統計的手法に対して，暗号技術を適用することでプライバシーを保護した疫学調査プロトコルを提案した．本稿の主な研究成果としては，4 章，5 章，6 章でそれぞれ提案したプロトコルであり，試験実装に基づき，その実現可能性を示した．

4 章の提案プロトコルは，放射線疫学調査 [2] で行われている被験者数，約 20 万人について約 9 分で処理を行うことができる．5 章の提案プロトコルは，多目的コホート研究 [1] で行われている被験者数，約 14 万人について，約 48 分で処理を行うことができる．また，6 章は 5 章のプロトコルと同程度の処理時間となる．これらの試験実装により，十分現実的な時間で処理を行えることを示した．

7.2　今後の課題

今後の課題は，更に様々な種類の疫学調査に使用される統計的手法に対してのプライバシー保護プロトコルを提案することである．

参考文献

[1] 独立行政法人 国立がん研究センター, “多目的コホート研究の成果”, pp. 1-18, 2011.

[2] 放射線影響協会, “原子力発電施設等放射線業務従事者等に係る疫学的調査”, pp. 1-120, 2010.

[3] 古川俊之，丹後俊郎, “新版 医学への統計学”, 朝倉書店, 1993.

[4] Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant, “Information shareing across private databases”, in proc. of ACM SIGMOD International Conference on Management of Data, 2003.

[5] Bart Goethals, Sven Laur, Helger Lipmaa and Taneli Mielikainen, “On Private Scalar Product Computation for Privacy-Preserving Data Mining”, The 7th Annual International Conference in Information Security and Cryptology (ICISC 2004), Vol. 3506 of LNCS, pp. 104-120, 2004.

[6] M. J. Freedman, K. Nissim, and B. Pinkas, “Efficient private matching and set intersection”, EUROCRYPT 2004, LNCS 3027, pp. 1?19, Springer-Verlag, 2004.

[7] A. C. Yao. “How to generate and exchange secrets”. In Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pp. 162-167, 1986.

[8] Dahlia Malkhi, Noam Nisan, Benny Pinkas, and Yaron Sella, “Fairplay - A Secure Two-Party Computation System”, Usenix Security Symposium, pp. 1-17, 2004.

[9] 菊池浩明, 香川大介, 石井一彦, 寺田雅之, 本郷節之, “組織間プライバシー保護データマイニングの考察”, SCIS2010.

[10] 統計局政策統括官・(統計基準担当) 統計研修所, 厚生労働省大臣官房統計情報部人口動態・保健統計課「人口動態統計」

[11] 千田浩司, 五十嵐大, 高橋克巳, “照合タグを用いた秘匿共通集合計算プロトコルとその応用”, コンピュータセキュリティシンポジウム 2009(CSS2009), pp. 1-6, 2009.

[12] 濱田浩気, 大竹茂樹, 竹之内大地, 千田浩司, 冨士仁, 高橋克巳, 村田節子, 熊田総佳, “秘匿関数計算システムによる医療データのプライバシ保護統計分析”, ライフインテリジェンスとオフィス情報システム研究会 (LOIS) 信学技報, vol. 111, no. 470, LOIS2011-102, pp. 177-181, 2012.

[13] 菊池浩明, 佐久間淳, 三上春夫, “プライバシーを保護したピロリ菌疫学調査”, 人工知能学会全国大会 (第 26 回), pp. 1-4, 2012.

[14] B. Bloom, “Space/time tradeoffs in hash coding with allowable errors”, Communications of the ACM, 13(7):422-426, 1970.

業績リスト

1. 佐藤智貴, 菊池浩明, 佐久間淳, “プライバシーを保護した放射線疫学調査システム”, CSEC54, Vol.2011-CSEC-54, No.25, pp. 1-6, 2011.

2. Tomoki Sato, Hiroaki Kikuchi, “Synthesis of Secure Password”, The 7th Asia Joint Conference on Information Security(AsiaJCIS2012), pp. 1-3, 2012.

3. 佐藤智貴, 菊池浩明, 佐久間淳, “プライバシー保護確率検定システムの実装と評価”, 第 20 回情報通信システムセキュリティ研究会 (ICSS), 信学技報 Vol. 112 No.315, pp. 61-66, 2012.

4. 佐藤智貴, 菊池浩明, 佐久間淳, “傾向性の検定における秘匿疫学調査プロトコル”, 暗号と情報セキュリティシンポジウム (SCIS2013), 3C1-4, pp. 1-4, 2013.

謝辞

本論文を執筆するにあたり多くの方々から多大なる御指導と御援助を賜りました.

特に，研究に関わらず私を導いて下さった東海大学情報通信学部通信ネットワーク工学科 菊池 浩明 教授に深く感謝を申し上げます.

また，本研究を推進するにあたって，御親切なる御教示ならびに御激励を賜りました東海大学情報理工学部情報科学科 内田 理 准教授に厚く御礼申し上げます.

さらに，本研究に多大なる御助言を賜りました，筑波大学システム情報工学研究科 佐久間 淳 准教授に深く御礼申し上げます.

そして，2 年間共に楽しみ，苦しみ，励まし合い，時には研究に対して有益な意見を与えてくれた菊池研究室，内田研究室の皆様に感謝致します.

最後に，支えてくれた家族に心より感謝の意を述べると共に，謝辞とさせて頂きます.