SCIS 2013 The 30th Symposium on
Cryptography and Information Security
Kyoto, Japan, Jan. 22 - 25, 2013
The Institute of Electronics,
Information and Communication Engineers

傾向性の検定における秘匿疫学調査プロトコル
Privacy-Preserving Dose-Response Relationship Test

佐藤 智貴 * 菊池 浩明 * 佐久間 淳 †
Tomoki Sato Hiroaki Kikuchi Jun Sakuma

あらまし 放射線などの危険度をプライバシーを考慮して安全に調査するために，その用量に依存して，疾病の罹患率や症状の度合いが増加するかどうかを調べることが必要である．そこで，用量を管理する組織 A とその反応を管理する組織 B が，互いの情報を秘匿したままで用量に対する傾向性を検査する暗号プロトコルを提案する．

キーワード 疫学調査, プライバシー保護, 傾向性の検定, 秘匿内積プロトコル, 秘密関数計算

1 はじめに

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]-[2]．例えば，(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 [3]”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，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，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を行うことを目指す．

2 要素技術

2.1 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 1 のデータ例を考えよう．表 1 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．傾向性の

表 1: 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	8.2	8.0	7.8	7.8

検定は，$y = \alpha + \beta \log(x)$ 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 1 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 1 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

$$\hat{\beta} = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} \tag{1}$$

となる．推定した切片 $\hat{\alpha}$ は

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \tag{2}$$

* 東海大学, 神奈川県平塚市北金目 4-1-1, Tokai University, 4-1-1, Kitakaname, Hiratsuka, Kanagawa, Japan
† 筑波大学, つくば市天王台 1-1-1 F934, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Japan

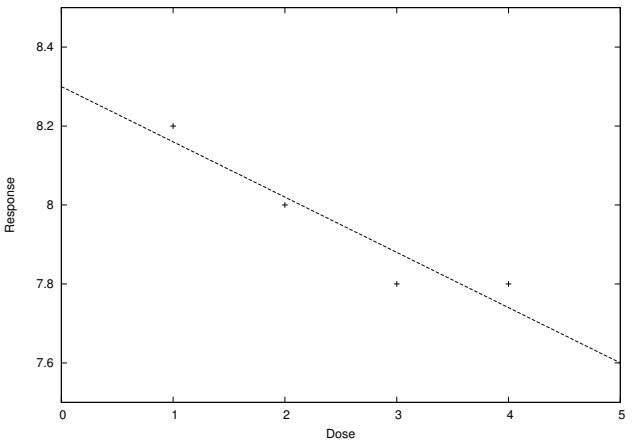

図 1: 表 1 についての回帰直線

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

$$s.e.(\hat{\beta}) = \frac{\sqrt{V_E}}{\sqrt{\sum_i^n (x_i - \overline{x})}} \tag{3}$$

を与える．ここで，統計量

$$t = \frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})} \tag{4}$$

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．

2.2 秘匿内積プロトコル [4]

本稿では，秘匿内積プロトコルを用いることで，プライバシーを保護したまま傾向性の検定を行う．秘匿内積プロトコルを Algorithm1 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．計算した結果は，$s_A + s_B = |X_A \cap X_B|$ となるような 2 つの乱数に分散され，s_A を組織 A が，s_B を組織 B が得るため，計算が終わっても秘匿されたままとなる．

Algorithm 1 秘匿内積プロトコル [4]

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．

2.3 秘密関数計算 (SFE)[5]

提案手法では，秘匿内積プロトコルを使用してそれぞれの組織が得た 2 つの分散された値を互いに秘匿したまま計算を行うために，Yao により提案された秘密関数計算 (Secure Function Evaluation(SFE)[5]) プロトコルを用いる．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．

3 提案手法

3.1 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 2 に示す．

表 2: 各組織の持つデータ

	組織 A	組織 B
ID	用量	反応
1	10ppm	8.2
2	100ppm	8.0
3	1000ppm	7.8
4	10000ppm	7.8

3.2 $\hat{\beta}$ のアプローチ

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

$$\begin{aligned}\hat{\beta} &= \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} \\ &= \frac{SS_{XY}}{SS_X} \\ &= \frac{\sum_i^n x_i y_i - (\sum_i^n x)(\sum_i^n y)/n}{\sum_i^n x^2 - (\sum_i^n x)^2/n}\end{aligned} \tag{5}$$

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．(5) 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，

$$x_{n+1} = \sum_i^n x_i, y_{n+1} = \sum_i^n y_i \tag{6}$$

とおいて，$n+1$ 次元の内積を求めれば良い．ここで，(5) 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

$$x'_i = \frac{x_i}{\sum_j^n x_j^2 - (\sum_j^n x_j)^2/n} \tag{7}$$

$(x'_1, \cdots, x'_{n+1})$ と $(y_1, \cdots, y_{n+1})$ を Algorithm 1 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．

3.3 提案プロトコル

提案するプロトコルを Algorithm 2 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．

3.4 回帰の検定

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，(4) 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．

Algorithm 2 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 1 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 1 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．

3.4.1 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．

3.4.2 方式 2

p を有意水準とする．例えば $p = 0.01$ とする．(4) 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 (3) 式と $\beta = 0$ を代入し，両辺を 2 乗すると

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

を得る．これを変形すると，

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

となる．これは，

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

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，

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

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．

4 おわりに

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．

今後の課題は，提案プロトコルについて試験実装し，その実現可能性や処理にかかる時間等を評価することが挙げられる．

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] 古川俊之，丹後俊郎，“新版 医学への統計学”, 朝倉書店, 1993.

[3] 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, 2010.

[4] 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.

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