3L3-OS-06b-3

官庁統計の情報保護基準

A confidentiality standard of Japanese official statistics

星野 伸明*1
Nobuaki Hoshino

*1金沢大学経済学類
School of Economics, Kanazawa University

Statistics Act defines Anonymized Data so that no individual shall be identified. Hence we need a technical definition of unidentifiability, but no such argument seems supported by observed facts. Therefore this article statistically models identification, by which an unidentifiable state is estimated from past experiences.

1. はじめに

日本の官庁が作成する統計は「公的統計」と呼ばれ，政策立案や研究に幅広く使われている．公的統計についてプライバシー保護の要請があることは容易に想像できると思うが，個人情報保護法は適用されない．統計法が情報保護のあり方を定めている．

個人情報保護法は個人の権利を保護する手段と考えられる．一方，統計法は統計の真実性を確保するための手段である．つまり調査客体が秘密にしたい真実も統計調査で答えてもらうことが目的である．そのためには回答が秘密のままであることを，個人だけでなく法人にも保証しなければならない．このような秘密性を “confidentiality” と呼び，プライバシーと区別する．また情報セキュリティの文脈の「機密性 (confidentiality)」とも概念的に異なる．

統計法では回答の「統計目的」以外の使用を禁止して，秘密性の実現を図る．ここで統計目的とは集計表の作成と解釈される．ただし集計表から特定個体の属性が明らかだと，脱税や独禁法違反の捜査といった統計目的外使用が可能になる．故に特定個体の属性が明らかにならないように,「秘匿処理」が施された集計表が公開されている．セルに属する個体が少数であったり，特定個体が支配的（ある地域の小売店売り上げのほとんどが巨大百貨店と分かる場合等）であったりする場合，秘匿の対象となってそのセルの値は削除される．

統計目的以外のデータ利用は，回答が秘密のままであるなら社会的に望ましい．従って現行統計法ではそのような「二次的利用」を認めている．回答が秘密のままである事は，利用者の直接行為規制とデータ情報量の管理を組み合わせて実現される．利用者の直接行為規制としては，統計利用目的の審査や守秘義務契約等が挙げられる．データ情報量の管理は，データ変換による間接的行為規制である．このような目的のデータ変換を「匿名化」と呼ぶ．軽く匿名化されたデータの二次的利用は，重く匿名化されたデータの利用より重い行為規制がかかる．

主に匿名化によって秘密性を守る二次的利用の形態は,「匿名データ」制度と呼ばれる．この匿名データの定義（統計法第２条第１２項）を引用すると,「一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む.）ができないように加工したもの」である．つまり匿名データ作成において匿名化の目的は，個体識別ができないこととなる．

回答の秘密性を具体化する文言として,「個体識別ができない」は必ずしも適切ではない．例えばある変数の条件付き分布が退化しているなら，個体識別がなくても回答の秘密は保たれない（あるマンション在住世帯の年収は全て一千万円以上など）．個体識別よりもデータの悪用が問題という主張は一理ある．しかし本報告では，統計法の理想についてこれ以上立ち入らない．

現行法の下で匿名データを作成するとして，個体識別ができないデータはできるデータといかに区別したらよいだろうか．この問題は基本的だが蔑ろにされている．匿名化についての多くの研究は，匿名化されたデータの表現だけから個体識別の危険性（開示リスク）を測る．そして統計当局にとっての開示リスクの受容域は，データの分析価値とバランスさせよとしか言わない．このような態度は要素技術の発展には好都合だが，二点問題が残る．まず開示リスクの低さと個体識別が不可能なことをつなぐ理論が別に必要となる．またデータ表現以外の情報の評価を避けている．

以下ではデータ表現以外の情報も利用し，個体識別行為を統計モデルとして表す．これにより現実に個体識別が起きていないことが，観測された開示リスクの下で個体識別が不可能なことの統計的証拠となる．

2. 個体識別の統計モデル

個体識別が可能か否かの判別モデルを観測から構成する上で，最大の困難は個体識別がほとんど起きないことである．情報が無ければ複雑なモデルは同定されない．出来るだけ特定的でないモデルを構成しよう．

2.1 個体識別の観測

個体識別が可能か否かは，明らかにデータの表現に依存する．ここで匿名化による表現の変化は滑らかだが，個体識別が不可能と可能の差は不連続である．これをモデル化する場合，データ表現の適当な実数特性値が閾値を超えれば個体識別が可能とみなすのが定石であろう．故に我々は，個体識別の難易度が閾値より高ければ個体識別が不可能とみなす．

難易度 δ を引数とする関数 f は，個体識別が可能なら 1 で不可能なら 0 を返すとする．つまり閾値が α として

$$f(\delta) = \begin{cases} 1 & \delta < \alpha \text{ の場合} \\ 0 & \delta \geq \alpha \text{ の場合} \end{cases}$$

連絡先: 〒 920-1192 石川県金沢市角間町 金沢大学 経済学類
星野伸明 (hoshino@kenroku.kanazawa-u.ac.jp)

ということになる．ここで未知の α を推定できるだろうか．

統計的に α を推定するには，観測値が必要になる．しかし個体識別が可能か否かは，観測されることではない．観測可能な事実は，個体識別が起きたか否かというになる．モデルを用いて説明しよう．確率変数 X が 1 なら個体識別が起き，0 なら起きないこととする．個体識別が不可能なら必ず $X = 0$ である．個体識別が可能の場合，難易度 δ に依存する確率 $p(\delta)$ で識別が起きると考えよう．すなわち $\Pr(X = 1; \delta < \alpha) = p(\delta), \Pr(X = 0; \delta < \alpha) = 1 - p(\delta)$ とする．危機管理を考えて，$p(\delta)$ は正と想定する．

このような状況で閾値が共通する n 件の事例が存在するとしよう．$i, i = 1, 2, \ldots, n,$ 番目について，少なくとも難易度 δ_i と識別の有無 x_i は観測できるはずだ．単純化のため $\delta_1 < \delta_2 < \cdots < \delta_n$ としよう．そして個体識別がこれまで起きていない（全ての i について $x_i = 0$）として考察を続ける．この場合モデルの尤度 ℓ は，$\delta_i < \alpha \leq \delta_{i+1}$ の時 $\ell(\alpha) = \prod_{j=1}^{i}(1 - p(\delta_j))^j$ となる．そして全ての δ について $0 < p(\delta) < 1$ なら，α の最尤推定値 $\hat{\alpha}$ は δ_1 以下である．つまり過去の事例で個体識別が起きていなければ，その最も低い難易度 δ_1 以下と閾値 α は推定される．

このように強い仮定を置かずに推測出来るのは，難易度の閾値が δ_1 以下ということまでである．当然だが，δ_1 未満の難易度について個体識別が不可能な証拠がないということだ．しかしながら匿名データを新しく作成する際，識別の難易度を δ_1 で管理することは統計的根拠を持つことになる．

2.2 個体識別の難易度測度

前節で導入した個体識別の難易度 δ をどこまで具体化出来るか考察するため，個体識別行為をまず要因分解する．Marsh 等 [Marsh 91] によると

$$\begin{aligned}\Pr(\text{識別が実際に起きる}) \quad = \quad & \Pr(\text{識別が起きる} \mid \text{識別を試みる}) \\ & \times \Pr(\text{識別を試みる}) \qquad (1)\end{aligned}$$

である．更に識別を試みた時にそれが成功する事態は，4 つの条件が成立する場合だという．すなわち

(a) 公開ファイルと個体識別のために照合するファイルのキー変数（疑似識別子）が同じ基準で記録されている．

(b) 公開ファイルに個体が含まれている．

(c) 個体が母集団一意である．

(d) 個体が母集団一意と確証出来る．

これらの条件が満たされる事象をそれぞれ a から d と書けば

$$\begin{aligned}\Pr(\text{識別が起きる} \mid \text{識別を試みる}) \quad = \quad & \Pr(a)\Pr(b|a)\Pr(c|a,b) \\ & \times \Pr(d|a,b,c) \qquad (2)\end{aligned}$$

ということになる．Marsh 等によるこれらの確率評価は説得的でないが，我々の問題でこの議論を活かすことが出来る．

(1) 式の Pr(識別が実際に起きる) は，前節の $p(\delta)$ と同じ概念である．そして (2) 式が正，つまり識別を試したときに識別が起きる確率が正ということは，個体識別が可能ということと同じである．故に (2) 式の右辺の要素のどれかが 0 なら，個体識別が不可能と言える．しかし公開される母集団一意が皆無になるのは例外的で，普通は $\Pr(a, b, c)$ は正となる．(2) 式の右辺を書き換えると

$$\Pr(\text{識別が起きる} \mid \text{識別を試みる}) = \Pr(a, b, c)\Pr(d|a, b, c)$$

であり，$\Pr(d|a, b, c)$ が 0 なら個体識別が不可能と考えられよう．つまり Marsh 等の枠組みにおいて通常の場合，個体識別が可能か否かは $\Pr(d|a, b, c)$ が 0 か否かという問題に縮退する．

従って我々は $\delta < \alpha$ なら $\Pr(d|a, b, c) > 0$ としよう．このように考えれば「個体識別の難易度」の意味が限定される．すなわち $\Pr(d|a, b, c) > 0$ と感度良く判定するには，母集団一意の確証と出来るだけ直接的に関係する量を δ として用いるべきである．

そして条件付き確率 $\Pr(d|a, b, c)$ は，正確に表現されて公開される母集団一意が所与になっている．故に δ は，正確に表現されて公開される母集団一意の割合 $\Pr(a, b, c)$ を引数とするべきだろう．それ以外に母集団一意の確証と関係する要因を $\boldsymbol{y}$ と書き，適当な関数 h について

$$\delta = h(\Pr(a, b, c), \boldsymbol{y}) < \alpha \Rightarrow \Pr(d|a, b, c) > 0$$

とすればこれまでの議論と整合する．

正確に表現されて公開される母集団一意数の増加は，母集団一意の確証をより容易にするはずだ．故に関数 h は以下の単調性を満たす．

$$q_1 \geq q_2 \Rightarrow h(q_2, \boldsymbol{y}) \geq h(q_1, \boldsymbol{y})$$

このような単調性さえ成り立てば，$\boldsymbol{y}$ が共通する複数の事例から，最も個体識別の難易度が低いものを選べる．つまり匿名データ作成において匿名化を最少にするには，$\Pr(a, b, c)$ を $\boldsymbol{y}$ が共通する過去最大の事例に合わせればよい．これは h の具体型が特定できなくても可能である．

所与のデータについて統計当局にとって最悪の場合の $\Pr(a, b, c)$ は，調査票情報を併用して推定できる．公開データの表現はその他の要因 $\boldsymbol{y}$ と無関係に変化させられるので，匿名データ作成において $\Pr(a, b, c)$ を管理するのは現実的である．

2.3 キー変数以外の個体識別要因

前節で $\boldsymbol{y}$ は母集団一意の確証と関係する要因と述べたが，Marsh 等は母集団一意の確証手法として，全数名簿と公衆の目の利用を挙げている．例えば教員という部分集団について全数名簿が利用可能なら，その中で母集団一意を確証可能である．また現職の首相のように，公衆が属性をよく知る一意な個体はあり得る．

このように母集団一意の確証に用いるのは個体情報である．故に民間データベース等が持つ変数の種類，精度，個体数は $\boldsymbol{y}$ の一部とするべきである．これらの情報から，キー変数を過去の事例と整合的に選ぶことは特に重要と思われる．

これまで述べた手法で，既存の情報による個体識別はある程度管理される．しかし既存の情報から個体識別が確証できないにせよ，可能性が高いとしよう．この場合に追加の情報を詐取などすれば，確証できるかもしれない．重要な追加情報を得るには，当該個体に接触する必要があると考えられるので，接触につながる広い意味での位置情報は精度を統制するべきである．なお識別行為の誘因の多寡は，識別を試みる確率を左右する要因と考える．

参考文献

[Marsh 91] Marsh, C., Skinner, C., Arber, S., Penhale, P., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N.: The Case for a Sample of Anonymized Records from the 1991 Census, *Journal of the Royal Statistical Society Series*, Series A, Vol. 154, pp. 305–340 (1991)